Perbandingan Model Decision Tree, Support Vector Machine dan K-Nearest Neighbors untuk Memprediksi Kualitas Air Minum

Thomas Brian[1], Alief Nur Aisyi Maulidhia[2], Evi Nafiatus Sholikhah[3], Sekarsari Wibowo[4]

[1, 2]Program Studi Teknik Kelistrikan Kapal, Jurusan Teknik Kelistrikan Kapal, Politeknik Perkapalan Negeri Surabaya

[3]Program Studi Teknik Keselamatan dan Kesehatan Kerja, Jurusan Teknik Permesinan Kapal, Politeknik Perkapalan Negeri Surabaya

[4]Program Studi Teknik Pengolahan Limbah, Jurusan Teknik Permesinan Kapal, Politeknik Perkapalan Negeri Surabaya

Email: [1]thomasbrian@ppns.ac.id, [2]aliefnur@ppns.ac.id, [3]evinafiatus@ppns.ac.id, [4]sekar@ppns.ac.id

Abstract. *The need for drinking water is increasing so that appropriate method support is needed to determine water potability. In this study, machine learning models will be implemented including Decision Tree, Support Vector Machine, and K-Nearest Neighbors to determine the best model in classifying drinking water quality from the Kaggle Water Quality dataset. The dataset consists of 3,276 data with 9 parameters consisting of ph, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes and Turbidity, and one Potability attribute as a target that indicates consumption eligibility. Based on the results of the trial using 20% and 30% testing data, the results of the confusion matrix model evaluation metrics (Accuracy, F1 Score, Precision and Recall) were shown, so it can be concluded that the Support Vector Machine classification model is the best among the three models because it has the highest value by meeting the three requirements of F1 Score, Precision and Recall values, each of which is 82.40%) of the four requirements tested.*

Keywords: *classification of drinking water, decision tree, k-nearest neighbors, machine learning, support vector machine.*

Abstrak. *Kebutuhan akan air minum semakin meningkat sehingga diperlukan dukungan metode yang sesuai untuk menentukan potabilitas air. Pada penelitian ini akan diimplementasikan model machine learning diantaranya Decision Tree, Support Vector Machine, dan K-Nearest Neighbors untuk menentukan model yang terbaik dalam pengklasifikasian kualitas air minum dari dataset Water Quality Kaggle. Dataset terdiri dari 3.276 data dengan 9 parameter terdiri dari ph, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes dan Turbidity, serta satu atribut Potability sebagai target yang menunjukkan kelayakan konsumsi. Berdasarkan hasil ujicoba menggunakan data testing 20% dan 30% menunjukkan hasil metrik evaluasi model confusion matrix (Accuracy, F1 Score, Precision dan Recall), sehingga dapat disimpulkan bahwa model klasifikasi Support Vector Machine menjadi yang terbaik diantara ketiga model tersebut dikarenakan memiliki nilai tertinggi dengan memenuhi tiga persyaratan nilai F1 Score, Precision dan Recall masing-masing sebesar 82,40%) dari empat persyaratan yang diujikan.*

Kata Kunci: *klasifikasi air minum, decision tree, k-nearest neighbors, machine learning, support vector machine.*

1. Pendahuluan

Kualitas air sangat penting untuk mendukung kehidupan manusia dan makhluk hidup lainnya. Air yang bersih dan sehat memastikan tubuh manusia berfungsi optimal dan menghindari penyakit yang disebabkan oleh kontaminasi, seperti diare atau keracunan. Selain itu, ekosistem yang bergantung pada air bersih, seperti sungai, danau, dan laut, akan terancam jika kualitas air menurun. Keberagaman hayati dan kelangsungan hidup spesies bergantung pada kualitas air yang terjaga. Air yang cukup dan bersih mendukung pertumbuhan tanaman dan hasil panen yang optimal. Untuk itu, penting bagi kita untuk menjaga kebersihan sumber air, mengurangi polusi, serta mengelola sumber daya air secara bijak. Dengan menjaga kualitas air, kita tidak hanya melindungi kesehatan dan kehidupan sekarang, tetapi juga keberlanjutan alam dan kehidupan di masa depan (Nurmalitasari,

2022).

Pengujian kualitas air melibatkan kandungan mikroorganisme patogenik, logam berat, dan parameter lainnya untuk menilai keamanan dan kelayakan air juga berfokus pada proses pengambilan dataset yang menjadi sangat penting untuk mengidentifikasi pencemaran dan merancang kebijakan pengelolaan sumber daya air yang efektif. Solusi yang ditawarkan yang akan dilakukan dalam membantu proses pengklasifikasian kualitas air bisa menggunakan beberapa algoritma klasifikasi diantaranya metode *Decision Tree, Support Vector Machine,* dan *K-Nearest Neighbors* yang akan diujicoba pada penelitian ini. Beberapa penelitian terkait adalah penelitian tentang prediksi kualitas air dengan melihat beberapa parameter yang mengandung informasi air tersebut. Dataset kualitas air berasal dari *Kaggle.* Metode data mining yang digunakan adalah algoritma *Random Forest* dan algoritma *Naïve Bayes*. Kedua algoritma ini merupakan algoritma klasifikasi yang dapat membantu kita untuk memprediksi kualitas air. Dengan menggunakan kedua algoritma tersebut didapatkan akurasi 79% untuk algoritma *Random Forest* dan akurasi 55% untuk algoritma *Naïve Bayes*. Setelah itu mengimplementasikan kedua algoritma tersebut ke *website* sederhana dengan framework flask (Christian, 2022). Penelitian lainnya dilakukan untuk mengetahui hasil evaluasi dari model yang dihasilkan untuk dapat memprediksi kualitas air yang dapat dikonsumsi atau tidaknya dengan menerapkan algoritma klasifikasi data mining yaitu adalah algoritma *K-Nearest Neighbor*. Algoritma ini diterapkan untuk menghitung probabilitas kualitas air yang aman atau tidak untuk dikonsumsi berdasarkan data rekaman yang diambil dari lingkungan sekitar terutama di daerah padat penduduk. Kumpulan data diperoleh dari website *kaggle*, hasil pemodelan diukur menggunakan tabel *Confusion Matrix* untuk menghitung akurasi. Setelah diuji, model ini memiliki tingkat akurasi tertinggi 85,52% dengan nilai *k* (tetangga terdekat) = 3 (Said, 2022). Penelitian lainnya yang telah dilakukan adalah komparasi metode pada pengklasifikasian kualitas air didapatkan kesimpulan bahwa metode SVM menghasilkan akurasi sebesar 78,70% dan *Naïve Bayes* sebesar 85,78% dengan jumlah data yang digunakan sebesar 167 data. Berdasarkan hasil yang didapatkan tersebut maka dapat ditarik sebuah kesimpulan bahwa metode yang lebih baik digunakan dalam pengklasifikasian dari kualitas air adalah metode Naïve Bayes dikarenakan memiliki rata-rata nilai yang lebih tinggi dari metode SVM (Tumangger, 2020). Dari beberapa penelitian terkait dengan pengklasifikasian kualitas air menggunakan beberapa metode dalam *machine learning*, dengan demikian masih ada peluang yang dapat dicapai untuk meningkatkan hasil akurasi dan kecepatan dari penggunaan beberapa metode *machine learning.* Oleh karena itu dalam penelitian kali ini penulis akan melakukan komparasi terhadap tiga metode *machine learning* diantaranya *Decision Tree, Support Vector Machine dan K-Nearest Neighbors.*

Metode *Decision Tree* adalah algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan regresi, yang membagi data berdasarkan fitur tertentu melalui serangkaian keputusan berbentuk pohon. Setiap pembagian data dilakukan dengan kriteria seperti *Gini Index* atau *Entropy* untuk klasifikasi, atau *Mean Squared Error* untuk regresi. Pohon ini terdiri dari *node* akar yang membagi data, *node* cabang yang mewakili keputusan, dan *node* daun yang memberikan hasil akhir. Keunggulannya adalah kemudahan interpretasi dan tidak memerlukan normalisasi data, namun rentan terhadap *overfitting* dan dapat menjadi kompleks jika tidak dipangkas dengan baik. *Decision Tree* banyak digunakan dalam berbagai aplikasi seperti prediksi pembelian produk atau klasifikasi penyakit (N. Maulidah, 2024). *Decision Tree* juga dikenal sebagai pohon keputusan yang merupakan algoritma untuk membangun struktur hierarki keputusan. Proses pembuatan *Decision Tree* dimulai dari *Root Node* hingga *Leaf Node* yang dilakukan secara rekursif. Setiap percabangan pohon menyatakan kondisi yang harus dipenuhi di setiap ujung pohon yang menyatakan nilai data (Musadi, 2023).

Metode *Support Vector Machine* (SVM) adalah metode dalam *machine learning* yang digunakan untuk klasifikasi dan regresi dengan mencari *hyperplane* terbaik yang memisahkan data ke dalam dua kelas yang berbeda, sambil memaksimalkan *margin* antara kelas-kelas tersebut. SVM berfokus pada *support vectors*, yaitu titik data yang berada paling dekat dengan *hyperplane*, yang menentukan posisi *hyperplane* (Hikmayanti, 2023). Untuk data yang tidak dapat dipisahkan secara linier, SVM menggunakan *kernel trick* untuk memetakan data ke dalam dimensi lebih tinggi. SVM dikenal dengan akurasi yang tinggi, kemampuan generalisasi yang baik, dan kemampuannya

menangani data berdimensi tinggi, meskipun dapat memiliki waktu komputasi yang lama dan memerlukan pemilihan parameter yang tepat. SVM banyak digunakan dalam aplikasi seperti klasifikasi gambar, pengklasifikasian teks, dan bioinformatika (F. Putrawansyah, 2024).

Metode *K-Nearest Neighbors* (KNN) adalah algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan regresi dengan cara mencari *k* tetangga terdekat dari data yang ingin diprediksi, berdasarkan jarak (umumnya jarak *Euclidean*) dengan data yang sudah ada. Prosesnya dimulai dengan menentukan nilai *k*, menghitung jarak antara data yang ingin diprediksi dan data lainnya, lalu memilih *k* tetangga terdekat. Untuk klasifikasi, prediksi dilakukan berdasarkan mayoritas kelas dari *k* tetangga tersebut, sementara untuk regresi, prediksi dihitung berdasarkan rata-rata nilai tetangga terdekat (T. Brian, 2025). Dalam hal ini, algoritma KNN dapat dengan mudah membantu penganalisis dalam prosedur klasifikasi titik baru yang berbeda dari himpunan data berdasarkan indeks kemiripan atau kesamaan titik dengan kedua kasus yang ada. KNN dapat digunakan saat himpunan data yang diambil berlabel, dan bebas gangguan (Bansal, 2021).

Pada penelitian ini akan membandingkan tiga metode klasifikasi (*Decision Tree*, *Support Vector Machine*, dan *K-Nearest Neighbors*) untuk mendapatkan hasil yang terbaik berdasarkan nilai *accuracy*, *F1 score*, *precision* dan *recall*. *Accuracy* mengukur seberapa akurat model dapat mengklasifikasikan data dengan benar. *F1 score* merupakan perbandingan rata-rata *precision* dan *recall* yang dibobotkan. *Precision* menggambarkan tingkat keakuratan antara data prediksi benar positif yang diminta dengan hasil prediksi yang diberikan oleh model. *Recall* menggambarkan keberhasilan model dalam menemukan kembali sebuah informasi. Dataset yang digunakan dalam penelitian ini adalah *Water Potability* dari *Kaggle* (A. Kadiwal, 2025). Sehingga diharapkan dari penelitian ini dapat menemukan metode klasifikasi yang tepat untuk memprediksi kualitas air melalui studi komparasi antar metode.

2. Tinjauan Pustaka

2.1. Data *Preprocessing*

Data *preprocessing* adalah tahap penting dalam proses analisis data atau pengembangan model *machine learning*. Ini mencakup serangkaian langkah yang bertujuan untuk membersihkan, mengubah, dan mempersiapkan data mentah agar siap untuk digunakan dalam analisis atau pelatihan model. Data *preprocessing* penting karena data dunia nyata seringkali memiliki masalah seperti nilai yang hilang, inkonsistensi, atau format yang tidak sesuai. Proses data *preprocessing* sangat penting untuk memastikan bahwa model *machine learning* yang dibangun tidak hanya akurat tetapi juga dapat digeneralisasi dengan baik pada data yang tidak terlihat sebelumnya. *Preprocessing* yang baik akan mengarah pada model yang lebih efisien dan efektif. Pada proses ini juga dilakukan proses standarisasi data untuk menyeragamkan rentang data menjadi normal menggunakan persamaan 1:

$$x' = \frac{x_i - mean(x)}{std(x)} \quad (1)$$

Keterangan:
x' adalah nilai setelah distandarisasi
x_i adalah nilai yang akan distandarisasi
$mean(x)$ adalah nilai rata-rata kolom tersebut
$std(x)$ adalah nilai standar deviasi dari kumpulan nilai pada kolom tersebut

2.2. *Decision Tree*

Decision tree adalah algoritma yang mudah dipahami dan diinterpretasikan, serta dapat menangani data numerik dan kategorikal tanpa membutuhkan asumsi distribusi tertentu. Model ini cepat dalam pelatihan dan dapat digunakan untuk masalah klasifikasi dan regresi. Namun, *decision tree* rentan terhadap *overfitting* dan sangat sensitif terhadap perubahan kecil pada data, yang dapat membuat model tidak stabil. Selain itu, ia kesulitan menangani data yang kompleks dengan banyak interaksi antar fitur dan sering kali memerlukan pemangkasan (*pruning*) untuk menghindari model yang terlalu rumit. Untuk meningkatkan performa, decision tree sering digunakan dalam *ensemble methods* seperti *Random Forest* atau *XGBoost* (Nurussakinah, 2023).

Kriteria pemilihan atribut berdasarkan *Gini Impurity* dan *Entropy*. *Gini Impurity* adalah

mengukur ketidakmurnian (*impurity*) dari sebuah *node*. Semakin kecil nilai *Gini*, semakin baik pemisahan data tersebut. *Entropy* adalah mengukur ketidakpastian dalam data. *Entropy* yang lebih rendah menunjukkan pembagian data yang lebih baik. Kedua kriteria tersebut ditunjukkan pada persamaan 2:

$$Gini(t) = 1 - \sum_{i=1}^{C} p_i^2$$
$$Entropy(t) = -\sum_{i=1}^{C} p_i log_2 p_i \quad (2)$$

Keterangan:
p_i adalah proporsi data dalam kelas i pada node t,
C adalah jumlah kelas

2.3. *Support Vector Machine*

Support Vector Machine (SVM) adalah algoritma yang efektif untuk klasifikasi, terutama pada data non-linier, dengan kemampuan untuk menangani data berdimensi tinggi menggunakan *kernel trick*. SVM memaksimalkan *margin* antara kelas, yang meningkatkan generalisasi dan mengurangi *overfitting*. Namun, SVM memerlukan pemilihan parameter yang hati-hati dan dapat membutuhkan waktu pelatihan yang lama, terutama pada dataset besar. Algoritma ini juga sensitif terhadap data yang tidak seimbang dan *noise*, serta kurang efisien pada dataset yang sangat besar karena kompleksitas komputasinya (F. Abdusyukur, 2023).

Support Vector Machine (SVM) bertujuan menemukan *hyperplane* yang memisahkan data ke dalam dua kelas dengan *margin* terbesar. Secara matematis, *hyperplane* dinyatakan pada persamaan 3:

$$w^T x + b = 0 \quad (3)$$

Keterangan:
w adalah vektor bobot dan *b* adalah bias
Margin dihitung sebagaimana persamaan 4:

$$M = \frac{2}{||w||} \quad (4)$$

SVM mengoptimalkan *w* dan *b* ntuk memaksimalkan *margin*, dengan syarat bahwa data dari masing-masing kelas terpisah dengan *margin* yang lebih besar atau sama dengan 1 seperti pada persamaan 5:

$$y_i(w^T x_i + b) \geq 1 \quad (5)$$

Untuk data yang tidak dapat dipisahkan secara linier, SVM menggunakan *kernel trick*, yang memetakan data ke ruang dimensi lebih tinggi agar dapat dipisahkan secara linier.

2.4. *K-Nearest Neighbors*

K-Nearest Neighbors (KNN) adalah algoritma sederhana dan fleksibel yang digunakan untuk klasifikasi dan regresi, tanpa memerlukan fase pelatihan dan mudah diimplementasikan. Algoritma ini hanya bergantung pada perhitungan jarak antar data dan tidak memerlukan asumsi distribusi data, namun dapat menjadi lambat dan memerlukan banyak memori pada dataset besar karena harus menyimpan seluruh data pelatihan (Vidiastanta, 2020). KNN juga sensitif terhadap fitur yang tidak relevan atau skala yang berbeda, serta kesulitan menangani data tidak seimbang. Menentukan nilai *k* yang optimal juga bisa sulit dan mempengaruhi performa model. Rumus *Euclidean Distance* antara dua titik $P(x_1, y_1)$ dan $Q(x_2, y_2)$ ditunjukkan pada persamaan 6:

$$d(P, Q) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (6)$$

Keterangan:
d adalah jarak antara dua titik *P* dan *Q*
(*x, y*) adalah titik koordinat
Menentukan nilai *k* yaitu jumlah tetangga terdekat yang akan dipertimbangkan dalam pengambilan keputusan.

2.5. Evaluasi Hasil

Hasil evaluasi dari hasil kinerja masing-masing model machine learning bisa diukur menggunakan *confusion matrix*. *Confusion matrix* adalah sebuah tabel yang digunakan untuk mengevaluasi kinerja suatu model klasifikasi, terutama dalam hal memprediksi kelas yang benar atau salah. Tabel ini memberikan gambaran yang jelas mengenai hasil prediksi model terhadap data yang telah diklasifikasikan, dengan membandingkan nilai prediksi yang dihasilkan model dengan nilai aktual.

Confusion matrix biasanya terdiri dari empat komponen utama:

1. *True Positive* (TP): Jumlah prediksi yang benar untuk kelas positif (model memprediksi positif dan data sebenarnya positif).
2. *False Positive* (FP): Jumlah prediksi yang salah untuk kelas positif (model memprediksi positif, tetapi data sebenarnya negatif).
3. *True Negative* (TN): Jumlah prediksi yang benar untuk kelas negatif (model memprediksi negatif dan data sebenarnya negatif).
4. *False Negative* (FN): Jumlah prediksi yang salah untuk kelas negatif (model memprediksi negatif, tetapi data sebenarnya positif).

Berikut adalah contoh bentuk confusion matrix untuk klasifikasi dua kelas (positif dan negatif) yang ditunjukkan pada Tabel 1.

Tabel 1. Confusion Matrix

	Predicted Positive	***Predicted Negative***
Actual Positive	*True Positive (TP)*	*False Negative (FN)*
Actual Negative	*False Positive (FP)*	*True Negative (TN)*

Dari *confusion matrix* ini, dapat dihitung beberapa metrik evaluasi model, seperti:

1. *Accuracy*: Persentase prediksi yang benar dari seluruh prediksi, dengan penerapan persamaan 7:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

2. *F1 score*: Rata-rata harmonik antara precision dan recall, yang memberikan keseimbangan antara keduanya, dengan penerapan persamaan 8:

$$F1\ Score = 2x\frac{Precision \times Recall}{Precision+Recall} \quad (8)$$

3. *Precision*: Kemampuan model untuk memberikan prediksi positif yang benar di antara semua prediksi positif, dengan penerapan persamaan 9:

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

4. *Recall*: Kemampuan model untuk menemukan semua contoh positif yang ada, dengan penerapan persamaan 10:

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

Confusion matrix sangat membantu untuk memahami tidak hanya seberapa sering model benar, tetapi juga jenis kesalahan yang dilakukan oleh model (misalnya lebih sering salah memprediksi kelas positif sebagai negatif atau sebaliknya).

3. Metode Penelitian

Pada metode penelitian akan diuraikan secara berurutan dan terstruktur melalui rangkaian langkah-langkah yang akan diimplementasikan. Rincian langkah-langkah penelitian dapat ditemukan pada Gambar 1.

 ISSN: 2579-566X (Online)
ISSN: 2477-5274 (Print)



Gambar 1. Langkah-langkah Penelitian

Metode penelitian dalam penelitian ini dimulai dengan persiapan dataset, yang mencakup pengumpulan dataset yang sesuai, dan penyusunan data dalam format yang dapat digunakan untuk analisis lebih lanjut. Setelah itu, data *preprocessing* dilakukan untuk membersihkan data dari nilai yang hilang, duplikasi, dan outlier, serta melakukan transformasi seperti normalisasi atau standarisasi fitur. Data kemudian dibagi menjadi *training set* dan *test set* untuk melatih dan menguji model. Pembagian yang tepat sangat penting untuk memastikan model tidak *overfit* dan dapat menggeneralisasi dengan baik. Langkah berikutnya adalah implementasi model *machine learning* dengan memilih algoritma yang sesuai dari perbandingan ketiga metode diantaranya *Decision Tree*, *Support Vector Machine* dan *K-Nearest Neighbors* dengan nilai k = 2 untuk melakukan pelatihan menggunakan data latih. Setelah model dilatih, dilakukan evaluasi hasil menggunakan metrik yang relevan, seperti *Accuracy, F1 score, Precision* dan *Recall*.

3.1. Persiapan Dataset

Dataset *Water Potability* yang digunakan pada penelitian ini diambil dari sumber terbuka yaitu https://www.kaggle.com/, terdiri dari 9 atribut *input* dan 1 atribut *output* dengan total 3.276 data. Deskripsi dari masing-masing atribut dijelaskan pada Tabel 2.

Tabel 2. Deskripsi Atribut

No	Atribut	Deskripsi
1	*Ph*	*pH* air (0 hingga 14).
2	*Hardness*	Kapasitas air untuk mengendapkan sabun dalam *mg/L*.
3	*Solids*	Total padatan terlarut dalam *ppm.*
4	*Chloramines*	Jumlah Kloramina dalam *ppm*
5	*Sulfate*	Jumlah Sulfat yang terlarut dalam *mg/L*.
6	*Condctivity*	Konduktivitas listrik air dalam *µS/cm*.
7	*Organic_carbon*	Jumlah karbon organik dalam *ppm.*
8	*Trihalomethanes*	Jumlah Trihalometana dalam *µg/L*.
9	*Turbidity*	Ukuran sifat air yang memancarkan cahaya dalam NTU.
10	*Potability*	Menunjukkan apakah air aman untuk dikonsumsi manusia dengan nilai 1 (potabel) dan 0 (tidak potabel).

4. Hasil dan Pembahasan

Pengujian dilakukan menggunakan *Jupyter Notebook* dengan spesifikasi CPU Intel i7 dan 16GB RAM. Dataset *Water Potability* yang digunakan pada penelitian ini diambil dari situs *Kaggle* yang menyediakan kumpulan dataset. Dataset ini terdiri dari 3.276 baris data dengan sembilan fitur dan satu output yang menentukan kualitas air layak atau tidak. Hasil tampilan lima baris pertama dari dataset ditunjukkan pada Gambar 2.

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Pot
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0

Gambar 2. Dataset *Water Potability*

Selanjutnya dataset dilakukan proses normalisasi dengan melakukan imputasi pada beberapa data yang kosong. Imputasi data adalah proses pengisian nilai yang hilang (*missing values*) dalam sebuah dataset agar data tersebut dapat digunakan dalam analisis atau model *machine learning*. Teknik imputasi yang umum digunakan meliputi penggantian nilai hilang dengan rata-rata, median, atau modus (tergantung pada jenis data. Proses ini penting untuk mengurangi bias dan memastikan integritas analisis data, karena banyak algoritma membutuhkan data lengkap untuk menghasilkan hasil yang valid. Pada Tabel 3 menunjukkan bagian kolom atribut yang masih berisi data yang kosong.

Tabel 3. Data Kosong (*Missing Value*)

	Data Kosong	**Setelah Imputasi KNN**
Ph	491	0
Hardness	0	0
Solids	0	0
Chloramines	0	0
Sulfate	781	0
Conductivity	0	0
Organic_carbon	0	0
Trihalomethanes	162	0
Turbidity	0	0
Potability	0	0

Visualisasi yang digunakan untuk menunjukkan sejauh mana hubungan linier antara berbagai variabel dalam sebuah dataset ditunjukkan menggunakan *Pearson Correlation Coefficient Heatmap*. Korelasi ini mengukur kekuatan dan arah hubungan linier antara dua variabel. Nilai koefisien ini berkisar antara -1 hingga 1. Nilai 1 menunjukkan korelasi positif sempurna (ketika satu variabel meningkat, variabel lainnya juga meningkat secara proporsional). Nilai -1 menunjukkan korelasi negatif sempurna (ketika satu variabel meningkat, variabel lainnya menurun secara proporsional). Nilai 0 menunjukkan tidak ada korelasi linier antara kedua variabel tersebut. Untuk memvisualisasikan korelasi antara banyak variabel sekaligus, kita menggunakan *heatmap*. *Heatmap* adalah representasi grafis yang menggunakan warna untuk menunjukkan nilai-nilai korelasi. Dalam konteks ini, setiap sel dalam heatmap menunjukkan nilai *Pearson Correlation Coefficient* antara dua variabel yang ditunjukkan pada Gambar 3.

Dataset dibagi menjadi dua bagian utama yaitu data pelatihan (*training*) dan data pengujian (*testing*). Data pelatihan digunakan untuk melatih model, sedangkan data pengujian digunakan untuk mengevaluasi kinerja model. Selanjutnya terdapat dua alokasi data pelatihan (*training*), dengan alokasi 80% dari total dataset berjumlah 3196 data dan 70% dari total dataset berjumlah 2797 data yang ditunjukkan pada Tabel 4. Data pengujian (*testing*) dengan alokasi 20% dari total dataset berjumlah 800 data dan 30% dari total dataset berjumlah 1199 data yang ditunjukkan pada Tabel 5.

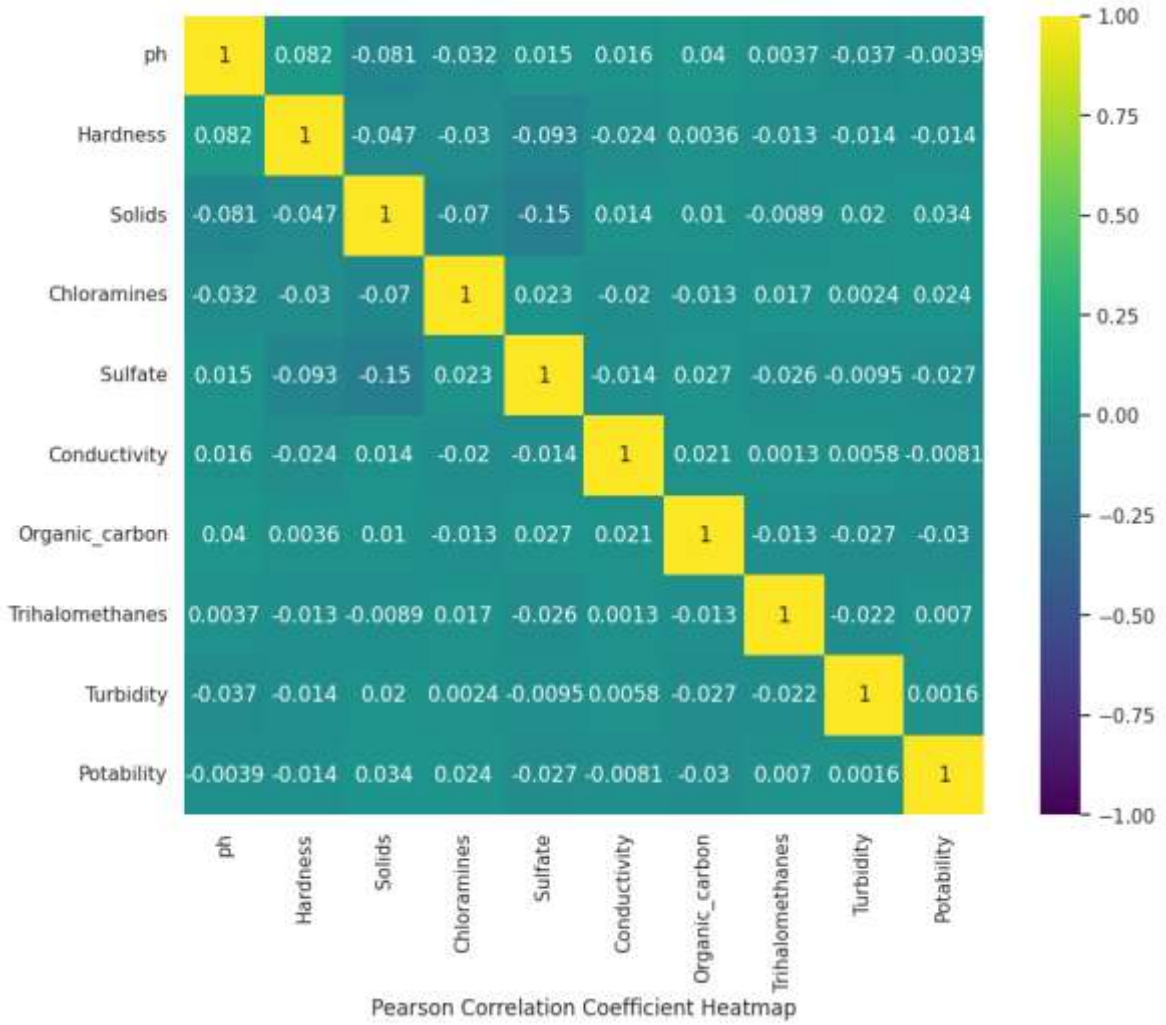
Tabel 4. Pembagian Data *Training*

	Shape of Data	
	80%	**70%**
X_train	3196	2797
y_train	3196	2797

Tabel 5. Pembagian Data *Testing*

	Shape of Data	
	20%	**30%**
X_test	800	1199
y_test	800	1199

ISSN: 2579-566X (Online)
ISSN: 2477-5274 (Print)



Gambar 3. *Pearson Correlation Coefficient Heatmap*

Berdasarkan tujuan penelitian ini adalah mencari model *machine learning* yang paling optimal, maka beberapa model yang terdiri dari *Decision Tree*, *Support Vector Machine*, dan *K-Nearest Neighbors* akan dilakukan pengujian. Hasil pengujian dengan persentase 20% data *testing* dapat dilihat pada Tabel 5.

Tabel 5. Perbandingan Model Klasifikasi dengan Data *Testing* (20%)

Model	Accuracy (%)	F1 Score (%)	Precision (%)	Recall (%)
Decision Tree	**70,50**	69,95	71,82	70,50
Support Vector Machine	52,25	**84,62**	**84,67**	**84,62**
K-Nearest Neighbors	62,12	61,05	63,99	62,12

Hasil pengujian dengan persentase 30% data testing dapat dilihat pada Tabel 6.

Tabel 6. Perbandingan Model Klasifikasi dengan Data *Testing* (30%)

Model	Accuracy (%)	F1 Score (%)	Precision (%)	Recall (%)
Decision Tree	**70,98**	70,86	71,29	70,98
Support Vector Machine	49,87	**82,40**	**82,40**	**82,40**
K-Nearest Neighbors	58,88	57,36	60,30	58,88

Dari hasil pengujian model *Support Vector Machine* mendominasi untuk hasil *F1Score*, *Precision* dan *Recall* paling unggul diantara model *Decision Tree* dan *K-Nearest Neighbors*. Tetapi hasil *Accuracy* menjadikannya paling rendah diantara lainnya. Hal ini dapat ditingkatkan dengan menggabungkan pengaturan parameter lain (*Feature Selection* dan *Hyperparameter Tuning*) yang digunakan dalam penerapan *Support Vector Machine.*

5. Kesimpulan

Tujuan dari penelitian ini adalah mendapatkan model klasifikasi yang terbaik dari hasil ujicoba menggunakan dataset kualitas air minum dengan beberapa model klasifikasi diantaranya *Decision Tree*, *Support Vector Machine*, dan *K-Nearest Neighbors*. Berdasarkan hasil ujicoba menggunakan data *testing* 20% dan 30% menunjukkan hasil yang mendekati sama untuk metrik evaluasi model *confusion matrix* (*Accuracy*, *F1 Score*, *Precision* dan *Recall*). Sehingga dapat disimpulkan bahwa model klasifikasi *Support Vector Machine* memiliki nilai tertinggi dengan memenuhi tiga persyaratan dengan nilai *F1 Score*, *Precision* dan *Recall* masing-masing sebesar 82,40% dari empat persyaratan yang diujikan. Terdapat beberapa saran yang dapat digunakan sebagai acuan bagi peneliti selanjutnya untuk meningkatkan hasil klsifikasi *Support Vector Machine* yaitu memperbanyak jumlah dataset dan adanya pengaturan parameter lain (*Feature Selection* dan *Hyperparameter Tuning*) yang digunakan dalam penerapan *Support Vector Machine.*

Referensi

Abdusyukur, F. 2023. Penerapan Algoritma Support Vector Machine (SVM) Untuk Klasifikasi Pencemaran Nama Baik di Media Sosial Twitter. Jurnal Komputa

Bansal, M., Goyal, A., Choudhary, A. 2021. A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning. Decision Analytics Journal

Brian, T., Sholikhah, E. N. 2025. Penerapan Algoritma K-Nearest Neighbor (KNN) untuk Memprediksi Kualitas Air Minum. Jurnal JTECS, vol. 5, no. 1

Christian, Y., Jacky, Winata, P. A., Ricky, dan Jeonanto, N. 2022. Prediksi Kualitas Air Menggunakan Algoritma Naïve Bayes Dan Random Forest. Komputek

Hikmayanti, H., Nurmasruriyah, A. F., Fauzi, A. 2023. Performance Comparison of Support Vector Machine Algorithm and Logistic Regression Algorithm. International Journal of Artificial Intelegence Research, Vol 7, No.1.1

Kadiwal, A. 2025. Water Potability Dataset. https://www.kaggle.com/adityakadiwal/water-potability

Maulidah, N., Maulidah, M. 2024. Prediksi Kualitas Air Menggunakan Metode Random Forest, Decision Tree, dan Gradient Boosting. Jurnal Khatulistiwa Informatika, vol. 12, no. 1, hal. 1-6

Musadi, A., Tertius, C. C., Steven, J. 2023. Comparing Artificial Neural Network and Decision Tree Algorithm to Predict Tides at Tanjung Priok Port. Procedia Computer Science

Nurmalitasari, Purwanto, E. 2022. Prediksi Performa Mahasiswa Menggunakan Model Regresi Logistik. Jurnal Derivat, vol. 9, no. 2

Nurussakinah, Faisal, M. 2023. Klasifikasi Penyakit Diabetes Menggunakan Algoritma Decision Tree. Jurnal Informatika

Putrawansyah, F., Susanti, T. 2024. Penerapan Metode Support Vector Machine Terhadap Klasifikasi Jenis Jambu Biji. Jurnal JIKO, vol. 8, no. 1, hal. 193-204

Said, H., Matondang, N., dan Irmanda, H. N. 2022. Penerapan Algoritma K-Nearest Neighbor Untuk Memprediksi Kualitas Air Yang Dapat Dikonsumsi. Techno.Com

Situngkir, R. H., Sembiring, P. 2023. Analisis Regresi Logistik Untuk Menentukan Faktor-Faktor Yang Mempengaruhi Kesejahteraan Masyarakat Kabupaten/Kota Di Pulau Nias. Jurnal Matematika dan Pendidikan Matematika

Tumangger, Sahalatua, R. M. 2020. Komparasi Metode Data Mining Support Vector Machine Dengan Naive Bayes Untuk Klasifikasi Status Kualitas Air. Universitas Brawijaya

 ISSN: 2579-566X (Online)
ISSN: 2477-5274 (Print)

Vidiastanta, Gusti, I., Hidayat, N., Dewi, R. K. 2020. Komparasi Metode K-Nearest Neighbors (K-NN) Dengan Support Vector Machine (SVM) Untuk Klasifikasi Status Kualitas Air. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer