JURNAL FASILKOM
Volume 15 No. 2 | Agustus 2025: 312-317

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Analisis dan Prediksi Kelayakan Air Minum Menggunakan Algoritma Random Forest

Ulfani Rohima Zalti[1], Dinda Rose Darmakusuma[2], Muhammad Ridwansyah[3], Edi Ismanto[4]
[1,2,3,4]Teknik Informatika, Fakultas Ilmu Komputer, Universitas Muhammadiyah Riau
[1]230401383@student.umri.ac.id* , [2]230401353@student.umri.ac.id, [3]230401140@student.umri.ac.id,
[4]edi.ismanto@umri.ac.id

Abstract

Water is a vital component for humans, with approximately 60% of the adult body consisting of water. Consuming water that does not meet standards can cause various diseases, such as diarrhea, poisoning, and Escherichia coli bacterial infections. Therefore, an accurate prediction system is needed to ensure that drinking water is suitable for consumption. This study aims to compare the performance of several machine learning classification algorithms in determining water potability based on chemical parameters contained in the Water Potability dataset. The parameters analyzed include pH, chlorine content, total dissolved solids, and other chemicals that affect water quality. The method used includes four classification algorithms, namely Logistic Regression, Decision Tree, Random Forest, and Extra Trees Classifier. Evaluation is carried out by assessing the accuracy of each model. The results show that Random Forest produces the highest accuracy, namely 66.6%, compared to other algorithms. Therefore, Random Forest is recommended as a more effective model for automatic classification of water suitability, as well as supporting data-based water quality monitoring efforts.

Keywords: Machine learning, classification algorithms, accuracy, evaluation, parameter

Abstrak

Air merupakan komponen vital bagi manusia, dengan sekitar 60% dari tubuh orang dewasa terdiri dari air. Konsumsi air yang tidak memenuhi standar dapat menyebabkan berbagai penyakit, seperti diare, keracunan, dan infeksi bakteri Escherichia coli. Oleh karena itu, diperlukan sistem prediksi yang akurat untuk memastikan bahwa air minum layak konsumsi.Penelitian ini bertujuan untuk membandingkan kinerja beberapa algoritma klasifikasi machine learning dalam menentukan potabilitas air berdasarkan parameter kimia yang terdapat dalam dataset Water Potability. Parameter yang dianalisis meliputi pH, kadar klorin, total padatan terlarut, serta zat kimia lain yang mempengaruhi kualitas air.Metode yang digunakan mencakup empat algoritma klasifikasi, yaitu Logistic Regression, Decision Tree, Random Forest, dan Extra Trees Classifier. Evaluasi dilakukan dengan menilai akurasi masing-masing model.Hasil penelitian menunjukkan bahwa Random Forest menghasilkan akurasi tertinggi, yaitu 66,6%, dibandingkan dengan algoritma lainnya. Oleh karena itu, Random Forest direkomendasikan sebagai model yang lebih efektif untuk klasifikasi otomatis kelayakan air, serta mendukung upaya pengawasan kualitas air berbasis data.

Kata kunci: *machine learning*, algoritma klasifikasi, akurasi, evaluasi, parameter

1. Pendahuluan

Air adalah salah satu kebutuhan dasar bagi seluruh makhluk hidup termasuk manusia. Walaupun pada era modern ini air bersih mudah di dapatkan, sekitar 2 milyar orang masih menggunakan sumber air minum yang terkontaminasi menurut WHO. [1] Setiap tahunnya ,di perkirakan 829.000 orang meninggal karena diare ysng disebabkan oleh sumber air yang tidak aman diminum dan sanitasi yang buruk. Akses terhadap air layak konsumsi sangat penting bagi kesehatan masyarakat. [2] Pengujian kualitas air secara tradisional memerlukan analisis kimia yang cukup mahal dan memakan waktu. Dengan adanya dataset mengenai kualitas air, machine learning memberikan metode yang cepat dan efisien untuk mengevaluasi kelayakan air untuk diminum.[10] Studi ini membandingkan empat model machine learning untuk mengklasifikasikan apakah sampel air aman untuk diminum berdasarkan berbagai parameter kimia.

Dalam penelitian ini, akan melakukan preprosessing data, membagi dataset menjadi data latih dan data uji (80:20) menggunakan python dan machine learning Metode tersebut digunakan untuk melakukan analisis dan membantu untuk melakukan pre-prosesing data dalam melakukan prediksi dan klasifikasi. Dataset ini berisi 3276 baris data dan 10 variabel yang berhubungan dengan water potability seperti ph(tingkat keasaman) atau turbidity(tingkat kekeruhan air).

Air harus dipastikan layak untuk dikonsumsi sebelum digunakan. Berdasarkan Peraturan Menkes RI No. 492/Menkes/Per/IV/2010, air yang dapat dikonsumsi harus memenuhi standar tertentu. Beberapa parameter yang harus dipenuhi antara lain tidak adanya E. Coli dan bakteri Koliform, bebas dari zat beracun, memiliki pH antara 6,5 hingga 8,5, tidak berbau, tidak berasa, jernih, dan memiliki konsentrasi zat terlarut kurang dari 500 mg/l. Parameter lainnya mencakup tidak adanya bahan kimia organik maupun

P-ISSN : 2089-3353
E-ISSN : 2808-9162

anorganik serta bebas dari kontaminasi disinfektan dan pestisida. [9].

Sumber air yang digunakan oleh rumah tangga di Indonesia bervariasi, termasuk air kemasan, PDAM, sumur, sungai, danau, serta mata air. Beberapa sumber seperti sungai, danau, dan sumur cenderung memiliki kualitas air yang tidak stabil dan sulit dikontrol, karena kualitasnya dapat berubah sewaktu-waktu akibat fenomena alam atau polusi. Masyarakat yang bergantung pada sumber air tersebut memerlukan sistem yang dapat mengklasifikasikan kelayakan konsumsi air agar terhindar dari mengonsumsi air yang tidak layak. [14].

Penelitian ini bertujuan untuk mengembangkan sistem yang dapat mengklasifikasikan kelayakan konsumsi air berdasarkan beberapa parameter sesuai dengan Peraturan Menkes RI No. 492/Menkes/Per/IV/2010, seperti pH, kekeruhan, dan konsentrasi zat terlarut. Untuk mengukur nilai pH, kekeruhan, dan konsentrasi zat terlarut, sistem ini memerlukan sensor-sensor yang dapat mengakuisisi ketiga nilai tersebut. Sensor yang dapat digunakan untuk mengukur pH antara lain adalah sensor pH-4502C. Untuk mengukur kekeruhan, salah satu sensor yang dapat digunakan adalah sensor turbidity SEN0189. Sedangkan untuk mengukur TDS air, sensor TDS SEN0244 dapat digunakan. [16]

Untuk mengklasifikasikan kelayakan konsumsi air, diperlukan algoritma klasifikasi.penelitian ini yang menggunakan metode Random Forest untuk mengklasifikasikan kualitas air sungai berhasil mencapai tingkat akurasi sebesar 66,6%.Meskipun hasil ini baik, masih ada potensi untuk meningkatkan akurasi.) menggunakan metode data mining untuk mengklasifikasikan kualitas air dengan metode klasifikasi: Support Vector Machines (SVM) **,** K-Nearest Neighbors (KNN)**.**[15]

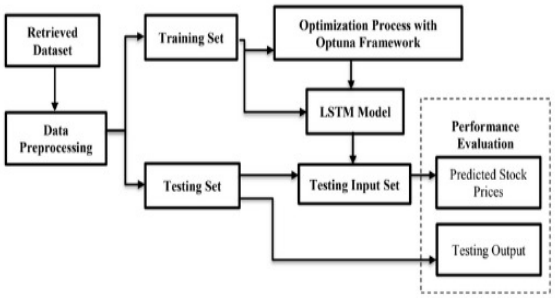
Hasil penelitian menunjukkan bahwa metode Random Forest memiliki akurasi terbaik, yaitu 66,6%. Selanjutnya, penelitian oleh mengklasifikasikan air layak minum menggunakan sembilan metode klasifikasi. Hasilnya menunjukkan bahwa Random Forest unggul dalam klasifikasi air layak minum dibandingkan metode lainnya, dengan akurasi latih 80% dan akurasi uji sebesar 20%. Penelitian ini menunjukkan hasil akurasi yang sangat baik dari metode Random Forest dalam mengklasifikasikan air.

Tabel 1. tabel Deskripsi Parameter Kualitas Air 1

Parameter	Deskripsi
Ph	Parameter yang diterapkan untuk menilai keseimbangan asam-basa dalam air.
Hardness	parameter yang diterapkan untuk mengidentifikasi kalsium dan garam magnesium.
Solids	Kemampuan air dalam melarutkan berbagai jenis mineral.
Chloramines	Kadar kolorin dan kloramin yang berfungsi sebagai disinfektan dalam air
Sulfate	Sulfat adalah zat yang terdapat dalam mineral, tanah, dan batuan.
Conductivity	Untuk mengetahui air murni tidak berfungsi sebagai pengantar arus listrik yang efektif melainkan sebagai isolator yang baik.
Organic carbon	Karbon memiliki Total Organis Carbon (TOC) yang digunakan untuk mengukur jumlah total karbon dalam air murni.
T*rihalomethanes*	Zat kimia yang dapat dijumpai dalam air yang telah diproses menggunakan klorin.
T*urbidity*	Tingkat keruh yang menunjukkn kualitas kejernihan air
potabilitas	Variabel target yang mencerminkan kelayakan air untuk diminum , dengan nilai 1 menunjukkan aman dan nilai 0 menunjukkan tidak aman

2. Metode Penelitian

Metode penelitian disusun secara sistematis melalui tahapan-tahapan tertentu yang dirancang untuk menyelesaikan permasalahan penelitian Ilustrasi pada gambar memberikan rincian mengenai setiap tahap penelitian, yang meliputi: 1. Pengumpulan data, 2. Pemrosesan data, dan 3. Pembuatan model.



Gambar 1.Struktur Dari Metode

Metode penelitian yang digunakan dalam studi Analisis dan Prediksi Kelayakan Air Minum terdiri dari beberapa tahap sebagaimana ditunjukkan pada Gambar 1. Tahap awal dimulai dengan pengambilan dataset kualitas air (*retrieved dataset*) dari sumber data yang relevan. Data mentah tersebut kemudian melalui proses pra-pemrosesan (*data preprocessing*), yang mencakup penanganan nilai hilang (*missing values*), normalisasi atau standarisasi, serta transformasi data ke dalam format yang sesuai untuk pemodelan. [1]

Dataset yang telah diproses selanjutnya dibagi menjadi data latih (*training set*) dan data uji (*testing set*). Data latih digunakan untuk membangun model, sedangkan data uji dipakai untuk mengukur kinerjanya. Proses optimisasi parameter model dilakukan dengan memanfaatkan *Optuna Framework* untuk menemukan kombinasi *hyperparameter* terbaik, seperti jumlah pohon (*n_estimators*), kedalaman maksimum (*max_depth*), dan parameter lain yang relevan dengan algoritma yang digunakan. [13]

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Model prediksi kelayakan air minum kemudian dibangun dan dilatih menggunakan data latih yang telah dioptimalkan. Selanjutnya, data uji dimasukkan ke dalam model guna menghasilkan hasil prediksi (*predicted potability*). Hasil prediksi tersebut dibandingkan dengan nilai aktual (*actual potability*) pada tahap evaluasi kinerja (*performance evaluation*). Evaluasi dilakukan menggunakan metrik pengukuran seperti *accuracy*, *precision*, *recall*, dan *f1-score*, sehingga diperoleh gambaran tingkat akurasi dan keandalan model yang dibangun dalam memprediksi kelayakan air minum. [14]

2.1 Exploratory data analysis

Langkah awal dalam penelitian ini adalah melakukan analisis data eksploratori (*exploratory data analysis*). Proses ini dilakukan untuk menelaah data kualitas serta potabilitas air secara deskriptif, dengan tujuan mengidentifikasi karakteristik ,pola distribusi, dan hubungan antar variabel yang ada. Tahap EDA membantu peneliti mendapat gambaran awal sebelum melakukan analisis lanjutan. Dataset yang dimanfaatkan memuat beragam indikator kualitas air tertentu dengan atribut seperti PH, kesadahan, jumlah padatan terlarut, dan parameter lainnya. Variabel targer "potabilitas", bersifat biner: nilai 1 berarti air layak diminum, sedangkan 0 berarti tidak layak.

2.2 Preprocessing Data

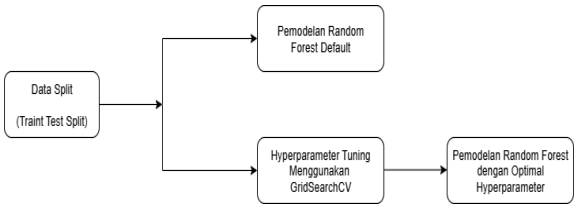
Preprocessing data merupakan serangkaian teknik yang digunakan untuk membersihkan dan menormalkan data yang tidak teratur serta tidak konsisten. Tujuan dari proses ini adalah untuk menghilangkan noise, nilai yang hilang dan ketidakseimbangan data dari basis data. Pada tahap ini dilakukan pembersihan data untuk menangani data yang tidak lengkap, data kosong serta data yang mengandung noise dan outlier. Dataset yang terdiri dari 3276 entri akan dibagi menjadi dua bagian utama, yaitu data pelatihan dan data pengujian. Pada tahap awal ini kami akan membagi proporsi 80% untuk data pelatihan dan 20% untuk data uji Selanjutnya, dalam fase ini kami akan menerapkan proses standarisasi data menggunakan rumus tertentu untuk menormalkan data ke dalam rentang yang seragam.

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

x' : nilai setelah standarisasi
x_i : nilai yang akan di standarisasi
$mean(x)$: nilai rata-rata kolom
$std(x)$: nilai standar deviasi dari kumpulan nilai dari kolom

2.3 Pembuatan model

Dalam penelitian ini, model yang dikembangkan dalam dua kondisi.Model pertama adalah random forest yang menggunakan nilai default sebelum dilakukan tuning hyperparameter. Model kedua adalah random forest yang dioptimalkan dengan parameter yang diperoleh melalui proses tuning hyperparameter menggunakan GridSearchCV. Pada model kedua, proses tuning hyperparameter dilakukan terlebih dahulu dengan GridSearchCV, kemudian Random Forest dibangun menggunakan parameter optimal yang telah ditemukan selama proses tuning tersebut.



Gambar 2. Tahapan Pembuatan Model

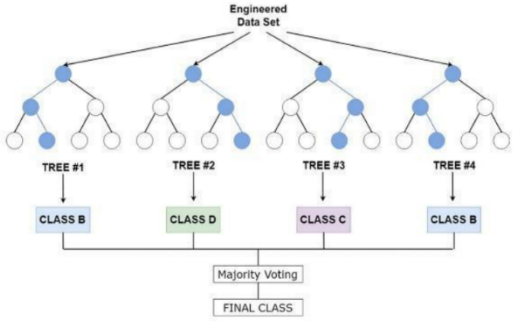
Gambar 2 menggambarkan langkah-langkah dalam pembuatan dua model, yaitu model pertama yang menggunakan parameter default dan model kedua yang menerapkan parameter optimal setelah proses tuning hyperparameter. Langkah-langkah ini dilakukan untuk membandingkan kinerja model sebelum dan sesudah tuning hyperparameter, sehingga dapat mengamati peningkatan yang terjadi. Berikut adalah penjelasan untuk setiap langkah dalam pembuatan model yang ditampilkan pada gambar.

2.3.1 Split data (*Train test split*)

Dalam tahap pembagian data, dataset dibagi menjadi data pelatihan dan data pengujian. Pembagian ini dilakukan dengan rasio 80% untuk data pelatihan dan 20% untuk data pengujian (80:20).

2.3.2 Pemodelan Random Forest

Random Forest (RF) merupakan algoritma klasifikasi ensemble yang memanfaatkan sejumlah pohon keputusan (Decision Trees). Setiap pohon keputusan memberikan prediksi, dan hasil akhir untuk suatu contoh pengujian diperoleh dari kombinasi prediksi semua pohon yang ada. Proses klasifikasi dalam Random Forest dimulai dengan membagi data, kemudian secara acak mengambil sampel untuk dianalisis di setiap pohon keputusan. Setelah pohon-pohon tersebut terbentuk, setiap kelas dari sampel data dievaluasi, dan dilakukan pemungutan suara untuk menentukan kelas yang paling sering muncul sebagai prediksi akhir. [14]



Gambar 3. Diagram Alur Random Forest

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Gambar 3 menampilkan diagram alur Random Forest yang terdiri dari beberapa pohon keputusan yang dibangun dari subset data yang berbeda. Hasil prediksi akhir ditentukan melalui metode major voting. Major voting adalah pendekatan dalam Random Forest di mana keputusan akhir untuk klasifikasi ditentukan berdasarkan suara terbanyak dari semua pohon keputusan. Menurut scikit-learn, nilai parameter default untuk Random Forest adalah sebagai berikut: n_estimators sebesar 100, max_features menggunakan nilai akar kuadrat (sqrt), max_depth diatur ke None, dan min_samples_split bernilai 2. n_estimators merujuk pada jumlah pohon keputusan yang digunakan. max_features adalah jumlah fitur acak yang diambil saat mencari simpul dalam pohon. Sementara itu, min_samples_split adalah parameter yang menentukan jumlah minimum sampel yang diperlukan dalam sebuah node agar node tersebut dapat dibagi selama proses pembangunan pohon. [1]

2.3.3. Pemodelan Random Forest menggunakan *Optimal Hyperparameter*

Hyperparameter optimal ditentukan melalui proses tuning hyperparameter. *Tuning hyperparameter* adalah metode yang digunakan untuk mencoba berbagai kombinasi parameter guna mengevaluasi kinerja masing-masing model. Grid search adalah proses yang memilih kombinasi nilai dari hyperparameter dengan mencoba berbagai kombinasi dan menilai setiap kombinasi tersebut. *Grid search* bertujuan untuk menemukan kombinasi terbaik dari hyperparameter untuk suatu model. [14]

Grid search cross-validation adalah metode yang menguji berbagai kombinasi nilai hyperparameter secara bersamaan dengan menggunakan teknik validasi silang. K-fold cross-validation merupakan salah satu metode validasi silang yang memungkinkan data pelatihan dan pengujian diulang hingga k iterasi, dengan menggunakan 1/k bagian dari dataset sebagai data pengujian. Algoritma Random Forest memiliki banyak hyperparameter yang dapat dioptimalkan. Penelitian ini berfokus pada beberapa hyperparameter dari Random Forest, seperti n_estimators, max_features, max_depth, dan min_samples_split. [12]

2.3.4. Evaluasi Model

Confusion matrix berfungsi sebagai metode evaluasi untuk menilai kinerja proses klasifikasi model dengan membandingkan prediksi yang dihasilkan oleh model dengan data aktual dari dataset yang diuji.

```
[ ] corr_matrix = df.corr()
    corr_matrix
```

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
ph	1.000000	0.075633	-0.081884	-0.031811	0.014403	0.017192	0.040061	0.002994	-0.036222	-0.003287
Hardness	0.075633	1.000000	-0.046899	-0.030054	-0.092766	-0.023915	0.003610	-0.012690	-0.014449	-0.013837
Solids	-0.081884	-0.046899	1.000000	-0.070148	-0.149840	0.013831	0.010242	-0.008875	0.019546	0.033743
Chloramines	-0.031811	-0.030054	-0.070148	1.000000	0.023791	-0.020486	-0.012653	0.016627	0.002363	0.023779
Sulfate	0.014403	-0.092766	-0.149840	0.023791	1.000000	-0.014059	0.026909	-0.025605	-0.009790	-0.020619
Conductivity	0.017192	-0.023915	0.013831	-0.020486	-0.014059	1.000000	0.020966	0.001255	0.005798	-0.008128
Organic_carbon	0.040061	0.003610	0.010242	-0.012653	0.026909	0.020966	1.000000	-0.012976	-0.027308	-0.030001
Trihalomethanes	0.002994	-0.012690	-0.008875	0.016627	-0.025605	0.001255	-0.012976	1.000000	-0.021502	0.006960
Turbidity	-0.036222	-0.014449	0.019546	0.002363	-0.009790	0.005798	-0.027308	-0.021502	1.000000	0.001581
Potability	-0.003287	-0.013837	0.033743	0.023779	-0.020619	-0.008128	-0.030001	0.006960	0.001581	1.000000

Gambar 4. Confusion Matriks

Tahap hasil melibatkan penggunaan confusion matrix, yang memberikan gambaran rinci tentang kinerja model. *Confusion matrix* adalah suatu matriks yang berisi informasi tentang keakuratan prediksi dan kesesuaian dengan keadaan sebenarnya, menunjukkan apakah prediksi tersebut benar atau salah.[1]

Dalam mengevaluasi kinerja model, performa yang digunakan diantaranya *Accuracy, Precission, Recal* dan *F1-Score*.

Tabel 5. Tabel Menetukan Nilai Prediksi

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

a. *Accuracy*

Accuracy merupakan persentase data yang berhasil diklasifikasikan dengan benar setelah dilakukan pengujian. Nilai akurasi dapat dihitung menggunakan rumus :

$$accuracy = \frac{TP+TN}{TP+FP+FN+TN} \times 100\% \quad (2)$$

b. *Precission*

Precision adalah persentase data positif yang diprediksi dengan akurat dari total data yang dikategorikan sebagai positif. Nilai precision dapat dihitung menggunakan rumus:

$$precission = \frac{TP}{TP+FP} \times 100\% \quad (3)$$

c. *Recall*

Recall adalah persentase data positif yang berhasil diprediksi dengan tepat dari total data positif yang sebenarnya. Nilai recall dapat dihitung menggunakan rumus :

$$Recal = \frac{TP}{TP+FN} \times 100\% \quad (4)$$

d. *F1-Score*.

F1-score adalah metrik yang menilai hasil dari perhitungan precision dan recall dengan menghitung rata-rata harmonik dari kedua nilai tersebut. Nilai F1-score dapat dihitung menggunakan rumus:

$$F1 - Score = \frac{2 \times Precission \times Recal}{Precission + Recal} \quad (5)$$

Tabel 6 berikut menampilkan hasil evaluasi kinerja model berdasarkan precision, recall, f1-score, dan accuracy untuk kedua kelas.

P-ISSN : 2089-3353
E-ISSN : 2808-9162

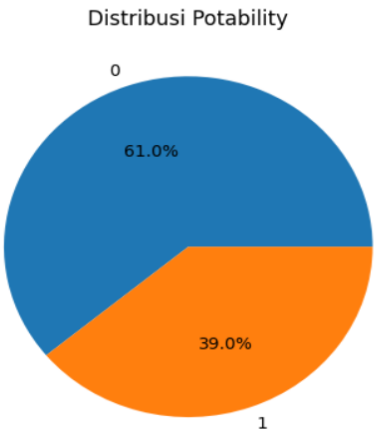
Tabel 6. Hasil evaluasi kinerja model

	precision	recall	f1-score	support
0	0.67	0.90	0.77	400
1	0.66	0.32	0.43	256
accuracy			0.67	656
macro avg	0.67	0.61	0.60	656
weighted avg	0.67	0.67	0.63	656

Model ini memiliki akurasi 0,67, artinya 67% prediksi benar. Rata-rata makro menunjukkan precision 0,67, recall 0,61, dan f1-score 0,60 tanpa mempertimbangkan proporsi data tiap kelas. Sementara itu, rata-rata tertimbang dengan precision 0,67, recall 0,67, dan f1-score 0,63 sudah memperhitungkan distribusi data di setiap kelas.

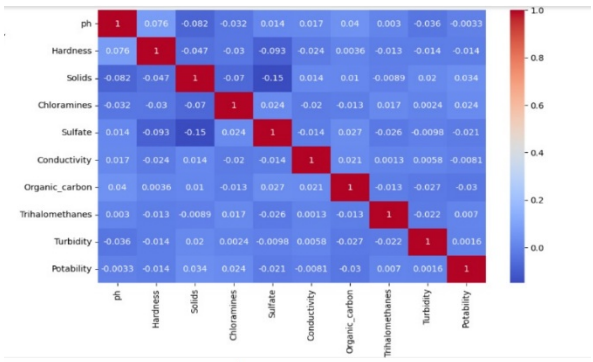
3. Hasil dan Pembahasan

Data secara keseluruhan menunjukkan bahwa 61% sumber daya air tidak memenuhi syarat untuk dikonsumsi, sementara hanya 39% yang dianggap layak. Perbedaan ini menunjukkan dengan jelas bahwa ketersediaan air minum saat ini berada dalam ancaman serius. Peningkatan polusi telah menyebabkan bakteri menyebar secara luas. Bakteri ini mencemari air tanah dan udara, yang pada gilirannya mengancam keamanan air minum. Selain itu, eksploitasi berlebihan terhadap air tanah telah menyebabkan beberapa wilayah mengalami masalah seperti penyusutan danau dan hilangnya dataran lumpur. Semua faktor ini berkontribusi pada penurunan kapasitas penyimpanan dan pemurnian air secara alami, yang akan memperburuk kondisi kelayakan air minum.



Gambar 4. Distribusi *Potability*

Gambar 4 menampilkan diagram lingkaran yang menggambarkan distribusi potabilitas air. Kategori 0 (61,0%) merepresentasikan sampel air yang tidak layak untuk dikonsumsi (non-potable), sedangkan kategori 1 (39,0%) menunjukkan sampel air yang layak diminum (potable).



Gambar 5. *Heatmap*

Gambar 5 menunjukkan heatmap dari matriks korelasi yang menggambarkan hubungan antara berbagai parameter kualitas air dan variabel *Potability* (kelayakan air minum).

Skala Warna:

1) Merah menunjukkan korelasi positif yang tinggi (dekat dengan +1).
2) Biru mencerminkan korelasi negatif yang kuat (dekat dengan -1).
3) Warna yang berada di antara putih dan biru muda menunjukkan korelasi yang lemah atau mendekati nol.

Interpretasi Nilai Korelasi:

1) Nilai pada diagonal utama (selalu 1) menunjukkan bahwa variabel tersebut berkorelasi sempurna dengan dirinya sendiri.
2) Sebagian besar hubungan antar variabel berada dalam rentang -0,15 hingga 0,1, yang menunjukkan bahwa korelasi antar parameter cenderung lemah.

Contoh:
Korelasi antara Hardness dan Solids adalah -0,067 (korelasi negatif yang lemah).
Hubungan antara Sulfate dan Hardness adalah -0,093.
Korelasi antara Chloramines dan Solids adalah -0,07.

Hasil penelitian menunjukkan bahwa nilai interaksi antara potabilitas dan variabel lainnya hampir mencapai 0. Hal ini menandakan bahwa tidak ada parameter yang menunjukkan hubungan linear yang signifikan dengan kelayakan air. Dengan kata lain, meskipun terdapat banyak faktor yang dapat memengaruhi kualitas air, tidak ada satu pun yang dapat diandalkan secara konsisten untuk menentukan apakah air tersebut aman untuk diminum.

Berbagai faktor lingkungan, seperti pH, Total Dissolved Solids (TDS), serta zat pencemar lain, dapat memengaruhi kualitas air, namun pengaruhnya belum cukup signifikan untuk memastikan air layak minum. Kondisi ini mencerminkan betapa kompleksnya sistem ekologi serta interaksi antar faktor di dalamnya.

P-ISSN : 2089-3353
E-ISSN : 2808-9162

4. Kesimpulan

Penelitian ini menunjukkan bahwa metode machine learning dapat dimanfaatkan untuk memprediksi kelayakan air minum berdasarkan parameter kimia yang terdapat dalam dataset *Water Potability*. Dari empat algoritma klasifikasi yang dibandingkan, yaitu Logistic Regression, Decision Tree, dan Random Forest, algoritma Random Forest dan Extra Trees Classifier mampu menghasilkan performa terbaik dengan tingkat akurasi sebesar 66,6%. Hasil penelitian juga menyoroti bahwa 61% sumber air dalam dataset tidak layak untuk diminum, sementara hanya 39% yang layak konsumsi. Dengan demikian, metode Random Forest dinilai cukup efektif untuk digunakan dalam tugas klasifikasi potabilitas air, meskipun akurasi model masih perlu ditingkatkan untuk hasil yang lebih optimal. Pengujian yang lebih mendalam disarankan dengan memperhatikan interaksi antara berbagai variabel dan tingkat potabilitas. Pendekatan ini dapat mencakup penerapan model statistik yang lebih canggih guna mengidentifikasi kemungkinan adanya hubungan non-linear.

Daftar Rujukan

[1] K. Kunci, "Prediksi Kualitas Air Menggunakan Metode Random Forest , Decision Tree , Dan Gradient Boosting Diterima : Diterbitkan :," vol. 12, no. 1, pp. 1–6, 2024.

[2] Pipit Muliyah, *Kualitas Air*, vol. 7, no. 2. 2020.

[3] M. Djana, "Analisis Kualitas Air Dalam Pemenuhan Kebutuhan Air Bersih Di Kecamatan Natar Hajimena Lampung Selatan," *J. Redoks*, vol. 8, no. 1, pp. 81–87, 2023, doi: 10.31851/redoks.v8i1.11853.

[4] T. Z. Jasman, M. A. Fadhlullah, A. L. Pratama, and R. Rismayani, "Analisis Algoritma Gradient Boosting, Adaboost dan Catboost dalam Klasifikasi Kualitas Air," *J. Tek. Inform. dan Sist. Inf.*, vol. 8, no. 2, pp. 392–402, 2022, doi: 10.28932/jutisi.v8i2.4906.

[5] U. Zaky, A. Naswin, and A. W. Murdiyanto, "Performance Analysis of the Decision Tree Classification Algorithm on the Water Quality and Potability Dataset," vol. 4, no. 3, pp. 145–150, 2023.

[6] H. Gao, Y. Li, H. Lu, and S. Zhu, "Water Potability Analysis and Prediction," vol. 16, pp. 70–77, 2022.

[7] P. A. Riyantoko, T. M. Fahrudin, and K. M. Hindrayani, "Analisis Sederhana Pada Kualitas Air Minum Berdasarkan Akurasi Model Klasifikasi Dengan Menggunakan Lucifer Machine Learning," *Pros. Semin. Nas. Sains Data*, vol. 1, no. 01, pp. 12–18, 2021, doi: 10.33005/senada.v1i01.20.

[8] A. Sudin, M. Salmin, M. Fhadli, and ..., "Klasifikasi Kelayakan Air Minum Bagi Tubuh Manusia Menggunakan Metode Support Vektor Machine Dengan Backward Elimination," *J. Jar. dan …*, vol. 3, no. 1, pp. 87–95, 2023, doi: 00.0000/jati.

[9] A. Kustanto, U. Sultan, and A. Tirtayasa, "Dinamika Pertumbuhan Penduduk Dan Kualitas," vol. 20, no. 1, 2020.

[10] Generosa Lukhayu Pritalia, "Analisis Komparatif Algoritme Machine Learning dan Penanganan Imbalanced Data pada Klasifikasi Kualitas Air Layak Minum," *KONSTELASI Konvergensi Teknol. dan Sist. Inf.*, vol. 2, no. 1, pp. 43–55, 2022, doi: 10.24002/konstelasi.v2i1.5630.

[11] R. G. De Luna *et al.*, "A Comparative Study of Machine Learning Techniques for Water Potability Classification," *IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON*, pp. 1345–1350, 2023, doi: 10.1109/TENCON58879.2023.10322335.

[12] A. Tangkelayuk, "The Klasifikasi Kualitas Air Menggunakan Metode KNN, Naïve Bayes, dan Decision Tree," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 9, no. 2, pp. 1109–1119, 2022, doi: 10.35957/jatisi.v9i2.2048.

[13] G. A. D. Sri Ari Ningsih and C. Pramartha, "Klasifikasi Kualitas Air Layak Minum menggunakan Algoritma Random Forest Classifier dan GridsearchCV,"*JELIKU (Jurnal Elektron. Ilmu Komput. Udayana)*, vol. 13, no. 1, p. 217, 2024, doi: 10.24843/jlk.2024.v13.i01.p22.

[14] V. Yolanda and E. Setiawan, "Klasifikasi Air Layak Konsumsi Berdasarkan pH, Kekeruhan, dan Konsentrasi Zat Terlarut Berbasis Arduino Menggunakan Random Forest," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 1, p. 3, 2017.

[15] L. Savitri and R. Nursalim, "Klasifikasi Kualitas Air Minum menggunakan Penerapan Algoritma Machine Learning dengan Pendekatan Supervised Learning," *Diophantine J. Math. Its Appl.*, vol. 2, no. 01, pp. 30–36, 2023, doi: 10.33369/diophantine.v2i01.28260.