

Analisis Metode Random Forest untuk Memprediksi Kualitas Air Berbasis Data Lingkungan

Analysis of the Random Forest Method for Predicting Water Quality Based on Environmental Data

Monica Zefanya Hendarswari

Sistem Informasi, Fakultas Teknologi Informasi, Universitas Merdeka, Malang, Indonesia

E-mail: *monica.zefanya04@gmail.com

Abstract. Water quality is an important factor that affects human health and the environment. In Indonesia, access to consumable water is still limited, so a fast and accurate water quality classification method is needed. This study aims to develop a classification model based on Machine Learning using the Random Forest algorithm to predict the feasibility of drinking water. The dataset used is Water Quality from Kaggle with 3,276 data and 9 parameters. The preprocessing stage includes data cleaning, handling missing values using IterativeImputers, data sharing, and normalization using StandardScaler. The three algorithms were compared, namely Random Forest, Decision Tree, and Naïve Bayes. The results showed that Random Forest had the highest accuracy of 69%, higher than Decision Tree (64%) and Naïve Bayes (63%). However, the model had a performance imbalance with high Recall in the non-drinkable class (89%) and low in the drinkable class (38%). The feature importance analysis shows that Sulfate, pH, and Hardness are the most influential parameters. This study contributes through the use of a more optimal imputation method in improving classification performance.

Keywords: water quality, machine learning, random forest, decision tree, iterative imputer

Abstrak. Kualitas air merupakan faktor penting yang mempengaruhi kesehatan manusia dan lingkungan. Di Indonesia, akses terhadap air layak konsumsi masih terbatas sehingga diperlukan metode klasifikasi kualitas air yang cepat dan akurat. Penelitian ini bertujuan untuk mengembangkan model klasifikasi berbasis *Machine Learning* menggunakan algoritma *Random Forest* untuk memprediksi kelayakan air minum. Dataset yang digunakan adalah *Water Quality* dari Kaggle dengan 3.276 data dan 9 parameter. Tahap *preprocessing* meliputi *data cleaning*, penanganan *missing value* menggunakan *IterativeImputer*, pembagian data, dan normalisasi menggunakan *StandardScaler*. Tiga algoritma dibandingkan yaitu *Random Forest*, *Decision Tree*, dan *Naïve Bayes*. Hasil menunjukkan bahwa *Random Forest* memiliki akurasi tertinggi sebesar 69%, lebih tinggi dibandingkan *Decision Tree* (64%) dan *Naïve Bayes* (63%). Namun model memiliki ketidakseimbangan performa dengan *Recall* tinggi pada kelas tidak layak minum (89%) dan rendah pada kelas layak minum (38%). Analisis *feature importance* menunjukkan bahwa *Sulfate*, *pH*, dan *Hardness* merupakan parameter paling berpengaruh. Penelitian ini berkontribusi melalui penggunaan metode imputasi yang lebih optimal dalam meningkatkan performa klasifikasi.

Kata kunci: kualitas air, machine learning, random forest, decision tree, iterative imputer

Submitted: 22-04-2026 | Accepted: 04-08-2026 | Published: 25-09-2026

How to Cite:

M. Z. Hendarswari, "Analisis Metode Random Forest untuk Memprediksi Kualitas Air Berbasis Data Lingkungan"
Journal of Information System and Application Development (JISAD), vol. 4, no. 2, pp. 120-129, 2026, doi:
10.26905/jisad.v4i2.17068



PENDAHULUAN

Air merupakan kebutuhan fundamental yang tidak tergantikan dalam kehidupan manusia, sehingga peruntukannya untuk konsumsi—mencakup air minum, sanitasi, dan keperluan rumah tangga—menjadi prioritas utama yang tidak dapat dikompromikan. Organisasi Kesehatan Dunia (WHO) menetapkan bahwa setiap individu memerlukan minimal 20–50 liter air bersih per hari untuk memenuhi kebutuhan minum, kebersihan diri, dan sanitasi lingkungan yang memadai [1]. Di Indonesia, permasalahan ketersediaan air bersih dan sanitasi layak terus menjadi isu yang berkembang, mengingat kualitas layanan di berbagai daerah masih jauh dari standar yang semestinya [2]. Kondisi ini diperkuat oleh data yang menunjukkan bahwa meskipun 93% masyarakat Indonesia diklaim memiliki akses air minum layak, hanya 11,9% rumah tangga yang sesungguhnya mengonsumsi air minum aman berdasarkan standar kesehatan, dengan kesenjangan signifikan antara kawasan pedesaan dan perkotaan [3]. Situasi ini semakin diperparah oleh tren peningkatan kasus pencemaran lingkungan yang terjadi setiap tahunnya, di mana pencemaran lingkungan didefinisikan sebagai kondisi yang memberikan pengaruh negatif terhadap makhluk hidup akibat aktivitas manusia [4].

Tantangan utama dalam penilaian kualitas air terletak pada kemampuan mengklasifikasikannya secara cepat, akurat, dan tepat sasaran, yang umumnya menuntut analisis mendalam terhadap beragam parameter sekaligus keahlian seorang pakar. Parameter fisika yang menjadi objek pengujian dalam penelitian ini meliputi Hardness, Turbidity, Solids, pH, Conductivity, dan Potability, di mana Hardness dan pH dianggap sebagai dua faktor paling krusial yang memengaruhi akurasi dan stabilitas pengukuran oleh produsen maupun pemasok air. Machine Learning hadir sebagai solusi efisien yang mampu menghemat waktu dan tenaga melalui kemampuannya menganalisis dan menyimpan data digital secara lebih terstruktur, dengan pendekatan yang meliputi Supervised Learning, Unsupervised Learning, dan Reinforcement Learning [5]. Supervised Learning secara khusus berpotensi meningkatkan kualitas dan keandalan data dengan meminimalkan intervensi manusia dalam identifikasi penyimpangan data, penanganan missing value, serta koreksi kekeliruan pembacaan data [6], meskipun algoritma-algoritma ini tetap memerlukan data berlabel dalam jumlah yang cukup besar sebagai prasyarat utamanya [7].

Untuk mengidentifikasi model Machine Learning yang paling sesuai, penelitian ini membandingkan tiga algoritma Supervised Learning—Random Forest, Decision Tree, dan Naïve Bayes—menggunakan dataset yang identik dalam konteks klasifikasi kualitas air. Random Forest bekerja dengan menghitung rata-rata prediksi dari sejumlah Decision Tree yang diterapkan pada berbagai subset dataset, di mana prediksi akhirnya ditentukan berdasarkan suara mayoritas seluruh pohon guna meningkatkan akurasi [8]. Decision Tree merepresentasikan prediksi dalam struktur pohon bercabang dengan node sebagai titik pemisahan menggunakan ukuran impurity seperti Gini Index, dan selain untuk keperluan klasifikasi, model ini juga dapat diterapkan pada penyelesaian masalah regresi [9]. Naïve Bayes bekerja dengan mempelajari data pelatihan untuk menghitung probabilitas setiap kelas, kemudian menetapkan kelas dengan probabilitas tertinggi—dalam penelitian ini antara Layak Minum dan Tidak Layak Minum—sebagai keluaran akhir prediksi [10].

Penelitian terdahulu yang mengklasifikasikan kualitas air menggunakan Random Forest mencatat akurasi sebesar 66,6%, dengan komposisi dataset yang menunjukkan bahwa 61% sumber air tidak layak konsumsi dan hanya 39% yang layak diminum. Ketidakseimbangan distribusi kelas ini berpotensi menciptakan kesan bahwa model berkinerja baik secara keseluruhan, padahal tingkat kesalahan pada kelas minoritas masih tergolong tinggi dan tidak dapat diabaikan. Persoalan metodologis lain yang turut ditemukan adalah tidak adanya deskripsi memadai mengenai penanganan missing value pada tahap preprocessing data, sehingga menyisakan celah yang perlu dicermati lebih lanjut. Kelemahan yang umum dijumpai pada studi-studi terdahulu adalah minimnya analisis kontribusi masing-masing parameter kualitas air, karena kebanyakan penelitian hanya menyajikan hasil klasifikasi tanpa disertai pembahasan mendalam tentang peran parameter fisika dan kimia dalam memengaruhi akurasi prediksi.

METODE

Metode yang digunakan pada penelitian ini adalah *Random Forest* karena memiliki performa yang lebih baik dibandingkan model *Decision Tree* dan *Naïve Bayes*. Tahapan penelitian proses klasifikasi ini terdiri dari pengumpulan data, *preprocessing* data, perbandingan model *Supervised Learning*, dan *model evaluation* pada model yang terpilih. *Random Forest* adalah metode yang terpilih untuk permasalahan klasifikasi kualitas air berdasarkan data input yang tersedia. *Random Forest* dapat mengurangi korelasi antar *Decision Tree* dengan menggunakan pemilihan acak sampel dan fitur. Jumlah data yang identik dipilih secara acak dari sampel *training* dalam data *training* asli [11]. *Random Forest* memiliki kelemahan salah satunya adalah kebutuhan komputasi yang relatif besar dibandingkan model pohon tunggal, khususnya saat jumlah pohon dalam model meningkat secara signifikan. Selain itu, performa *Random Forest* sangat dipengaruhi oleh pengaturan parameter seperti jumlah pohon (*n_estimators*), kedalaman maksimum pohon (*max_depth*), dan jumlah fitur yang digunakan dalam pemisahan atau split [12].

Tools

Salah satu tools *Machine Learning* yang dimanfaatkan pada penelitian ini adalah *Google Colaboratory* untuk mengeksekusi kode *Python*. *Python* kerap digunakan pada *Data Science*, *Cloud Computing*, *Data Analyst*, dan *Machine Learning* karena memiliki banyak *library* seperti *NumPy*, *Pandas*, *TensorFlow*, *Scikit-learn*, *SciPy*, dan *Matplotlib* [13].

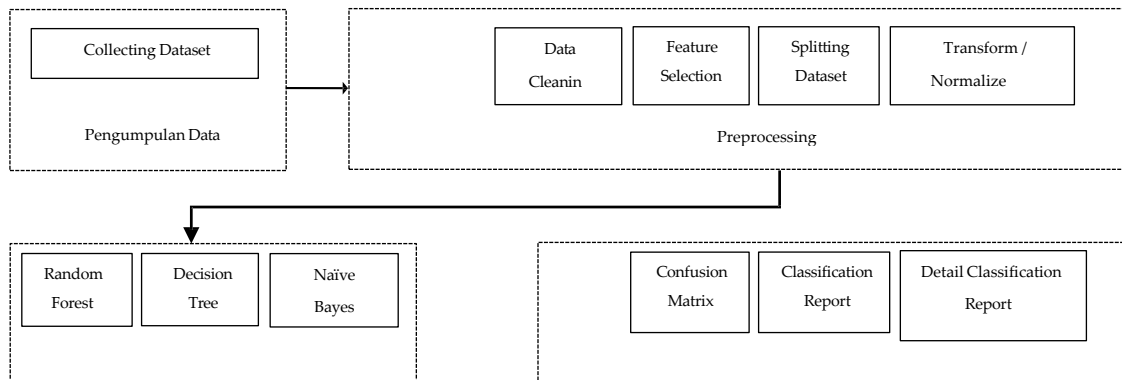
Pengumpulan Data

Tahap pengumpulan data adalah proses pengambilan data dari dataset kualitas air yang bertujuan untuk memperoleh data mentah yang akan digunakan sebagai dasar dalam seluruh proses penelitian. Dalam penelitian ini akan digunakan dataset *Water Quality* dari *Kaggle* (<https://www.kaggle.com/datasets/adityakadiwal/water-potability>). Dataset yang tersedia di *Kaggle* sering kali sudah dibersihkan dan diformat dengan baik, membuatnya lebih mudah untuk digunakan dalam analisis. Selain itu, *Kaggle* menyediakan deskripsi dan metadata yang sangat berguna untuk memahami konteks dan struktur data [14]. Sebelum digunakan dalam proses pemodelan, dataset terlebih dahulu diperiksa untuk memastikan kelengkapan atribut dan konsistensi format data.

Preprocessing Data

Dataset yang telah dikumpulkan kemudian masuk ke tahapan penting yaitu *preprocessing* data, yang merupakan proses pengolahan data dalam *Machine Learning* dimana data diubah/dienkripsi sehingga mesin dapat dengan cepat memproses atau menganalisis data tersebut [15]. Alur penelitian yang digambarkan pada gambar dibawah memiliki tujuan untuk memberikan gambaran secara menyeluruh mengenai proses penelitian. Proses utama pada tahap *preprocessing* data yaitu *cleaning*, *feature selection/feature extraction*, pembagian dataset (*splitting dataset*), serta transformasi atau normalisasi data. Permasalahan *missing value* dan inkonsistensi data ditangani dengan *data cleaning*. Proses *data cleaning* memeriksa nilai yang termasuk dalam kategori nilai yang hilang untuk setiap atribut dalam kumpulan dataset yang ada. Kemudian ditemukan *missing* data pada kolom *pH*, *Sulfate*, dan *Trihalomethanes*. *Missing values* tersebut berpotensi mengganggu proses pembelajaran model apabila tidak ditangani dengan baik.

Oleh karena itu, dilakukan penanganan *missing values* menggunakan metode *IterativeImputer*, yang mampu memperkirakan nilai hilang berdasarkan hubungan antar atribut dalam dataset. *IterativeImputer* memiliki strategi untuk mengisi *missing values* dengan memodelkan setiap fitur yang memiliki *missing values* sebagai fungsi dari fitur-fitur lain secara bergantian. Gambaran alur ini membantu memperjelas hubungan antar tahapan, sekaligus memastikan bahwa tidak ada langkah krusial yang terlewat dalam proses pemodelan dan evaluasi model klasifikasi.

**Gambar 1.** Alur Penelitian

Dataset kemudian dibagi menjadi dua bagian melalui proses *splitting* data, yaitu data *training* dan data *testing*. Pembagian dilakukan dengan perbandingan 8:2 (80% data *training* dan 20% data *testing*). Proses ini juga dilanjutkan dengan menentukan objek data dengan memisahkan fitur data seperti *pH*, *Hardness*, *Sulfate* dan *Potability* (data target). Setelah dilakukan pemisahan data tersebut, dari total 3.276 data, didapatkan 2.620 data sebagai data *train* dan 656 data sebagai data *test*. Data *training* digunakan untuk melatih model, sedangkan data *testing* digunakan untuk mengevaluasi kinerja model terhadap data yang belum pernah dilihat sebelumnya. Model *Random Forest* digunakan pada data *training* untuk menghasilkan algoritma. Kemudian algoritma yang terbentuk dievaluasi menggunakan *confusion matrix* dengan memakai data *testing* [16].

Selanjutnya adalah tahap transformasi data yang dimaksudkan untuk mengubah skala pengukuran dari nilai asli menjadi bentuk nilai yang terstandarisasi. Mengingat nilai dari semua atribut pada dataset kualitas air yang digunakan adalah data numerik (tipe `Numerik64`), proses transformasi data dilakukan menggunakan teknik Standarisasi melalui *StandardScaler*. *StandardScaler* adalah metode untuk menyamakan skala antar fitur numerik pada dataset kualitas air agar tidak terjadi dominasi fitur tertentu dalam proses pembelajaran model.

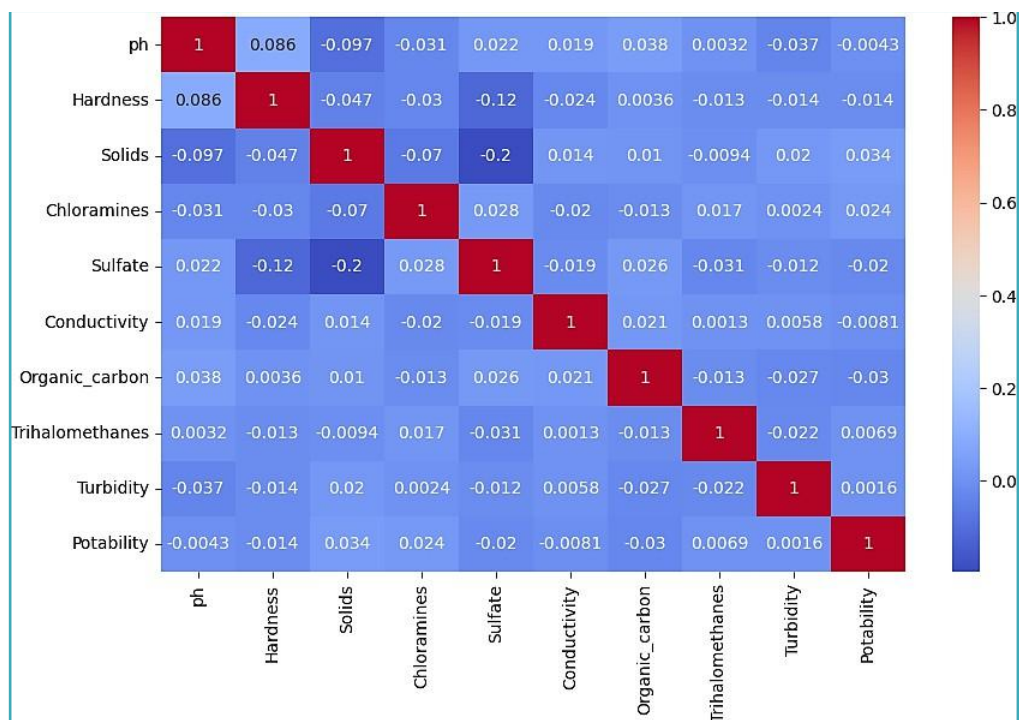
Modelling & Model Evaluation

Setelah melalui tahap *preprocessing*, data memasuki tahap *modelling*, yaitu perbandingan pada tiga model *Supervised Learning* yang menggunakan data *training* dan data *testing* untuk melatih dan menguji performa setiap model. Ke-3 model ini yaitu *Random Forest*, *Decision Tree*, dan *Naïve Bayes* yang dianalisa dengan metrik-metrik tertentu dan menghasilkan model terpilih. Analisa ini tidak hanya mengukur tingkat ketepatan model dalam memberikan prediksi yang benar, tetapi juga menilai keseimbangan antara kemampuan model dalam mendeteksi kasus positif secara benar (*Recall*), serta ketepatan prediksi positif (*Precision*) [17]. Oleh karena itu, metrik-metrik penting yang digunakan untuk menganalisis 3 model tersebut adalah *Accuracy*, *Precision*, *Recall*, dan *F1-Score*. Perbandingan model bertujuan untuk mengetahui model *Machine Learning* yang memiliki performa terbaik untuk melakukan klasifikasi kualitas air (model final).

Pada tahap *Model Evaluation* menghasilkan *Confusion Matrix* dan detail klasifikasi. *Confusion Matrix* adalah sebuah tabel yang digunakan untuk mengevaluasi kinerja model atau sistem klasifikasi pada data yang telah diketahui label kelasnya [18]. *Confusion Matrix* memberikan informasi rinci mengenai jumlah data yang diklasifikasikan dengan benar maupun salah pada masing-masing kelas, sehingga memudahkan analisis kesalahan klasifikasi yang terjadi pada model.

HASIL DAN PEMBAHASAN

Analisa distribusi jumlah data telah dilakukan terhadap variabel target yaitu *Potability* dan membuktikan bahwa jumlah data dengan label Tidak Layak Minum lebih dominan dibandingkan data dengan label Layak Minum. Kondisi ini menunjukkan adanya ketidakseimbangan kelas (*class imbalance*), di mana kelas dengan jumlah data lebih besar memiliki kemungkinan lebih mudah diidentifikasi oleh model. Berdasarkan *heatmap*, terdapat korelasi relatif lebih tinggi antara parameter *Hardness* dan *pH* dibandingkan pasangan parameter lainnya. Keterkaitan ini sejalan dengan teori kualitas air, di mana kandungan mineral tertentu dapat memengaruhi tingkat keasaman atau kebasaan air. Parameter *Solids* dan *Sulfate* merupakan parameter yang paling sedikit memiliki korelasi dengan parameter lainnya. Hal ini mengindikasikan bahwa parameter tersebut cenderung berdiri sendiri dan memiliki variasi yang tidak terlalu dipengaruhi oleh parameter kualitas air lainnya. Analisis korelasi ini tidak digunakan sebagai dasar penghapusan fitur, melainkan sebagai alat bantu untuk memahami hubungan antarparameter. Seluruh fitur tetap digunakan dalam proses pemodelan karena masing-masing parameter merepresentasikan aspek penting dari kualitas air.



Gambar 2. Heatmap Correlation

Berdasarkan hasil pemeriksaan data awal, ditemukan *missing value* pada tiga parameter yaitu *pH* sebanyak 491 data, *Sulfate* sebanyak 781 data, dan *Trihalomethanes* sebanyak 162 data. Nilai-nilai yang hilang tersebut diisi menggunakan *IterativeImputer* yang bekerja secara iteratif dengan memodelkan setiap fitur yang memiliki *missing value* sebagai fungsi dari fitur-fitur lain. Setelah proses imputasi selesai dilakukan, seluruh *missing value* berhasil terisi sehingga tidak ada lagi nilai kosong pada ketiga parameter tersebut. Hal ini memastikan dataset siap digunakan secara optimal pada tahap pemodelan berikutnya. Proses terakhir adalah transformasi data menggunakan *StandardScaler*. Tujuan transformasi adalah menghubungkan dan menyesuaikan fitur yang relevan sehingga siap digunakan dalam proses klasifikasi [19]. Metode ini mengubah setiap nilai fitur sehingga memiliki rata-rata 0 dan standar deviasi 1. Meskipun *Random Forest* relatif tidak sensitif terhadap skala data, standarisasi tetap dilakukan untuk menjaga konsistensi tahapan *preprocessing*.

	0
ph	491
Hardness	0
Solids	0
Chloramines	0
Sulfate	781
Conductivity	0
Organic_carbon	0
Trihalomethanes	162
Turbidity	0
Potability	0

→

	0
ph	0
Hardness	0
Solids	0
Chloramines	0
Sulfate	0
Conductivity	0
Organic_carbon	0
Trihalomethanes	0
Turbidity	0
Potability	0

Gambar 3. Hasil *IterativeImputer*

Penulis menampilkan 5 baris data pertama yang berisi kolom label aktual dan label prediksi berdasarkan prediksi gabungan masing-masing parameter. Terdapat pula kolom perbedaan untuk melihat kesesuaian prediksi model *Random Forest*. Dapat dilihat pada baris data 2 dan 5 bahwa air layak minum (1) diprediksi tidak layak minum (0) oleh model.

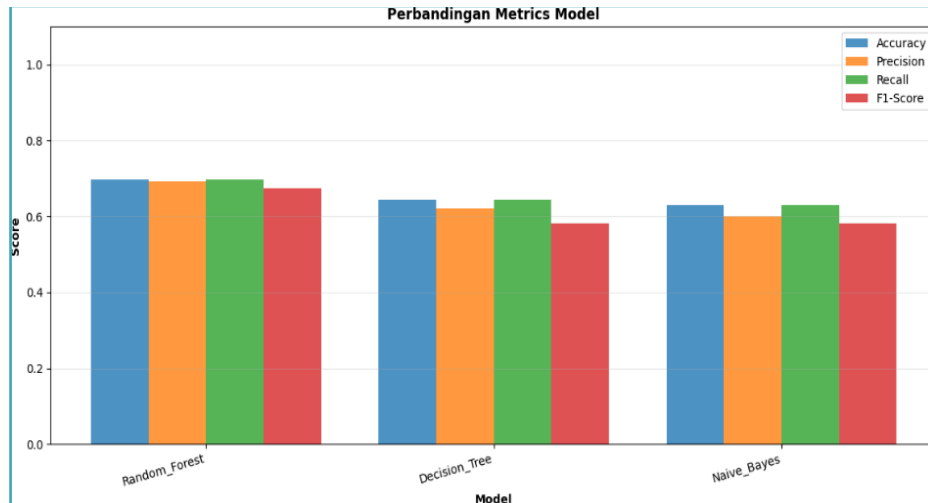
Tabel 1. Baris Data Teratas Hasil Prediksi Model *Random Forest*

No	Label Aktual	Label Prediksi	Perbedaan	Status
1	0,00	0,00	0	Benar
2	1,00	0,00	-1	Salah
3	0,00	0,00	0	Benar
4	0,00	0,00	0	Benar
5	1,00	0,00	-1	Salah

Selanjutnya adalah tahap pemodelan yang dilakukan pada 3 algoritma yaitu *Decision Tree*, *Naïve Bayes*, dan *Random Forest*, hasil evaluasi dapat dilihat pada tabel 2 dan gambar 4. Berdasarkan hasil perbandingan, *Random Forest* memperoleh akurasi tertinggi sebesar 69% diikuti dengan *Decision Tree* sebesar 64% dan *Naïve Bayes* sebesar 63%. Demikian pula dengan metrik *Precision*, *Recall*, dan *F1-Score* yang menunjukkan hasil tertinggi daripada 2 model lainnya. *Random Forest* memiliki keunggulan untuk dapat mengurangi *overfitting* dan lebih baik menangkap pola yang kompleks. Sementara itu, *Decision Tree* termasuk algoritma yang kuat namun tergolong rentan *overfitting*. *Naïve Bayes* memiliki hasil paling rendah karena independensi fitur yang tidak sepenuhnya terpenuhi di dalam data. Ditemukan korelasi yang relatif tinggi antara parameter *Hardness* dan pH, yang menunjukkan bahwa independensi fitur pada *Naïve Bayes* tidak sepenuhnya terpenuhi pada dataset ini. Kondisi ini menyebabkan model kesulitan menangkap hubungan antar parameter secara akurat, sehingga berdampak pada rendahnya performa klasifikasi dibandingkan *Random Forest* dan *Decision Tree*.

Tabel 2. Hasil Perbandingan Model

Model	Accuracy	Precision	Recall	F1-Score
<i>Random Forest</i>	0,698171	0,691724	0,698171	0,673495
<i>Decision Tree</i>	0,643293	0,620289	0,643293	0,582620
<i>Naïve Bayes</i>	0,631098	0,600008	0,631098	0,582512

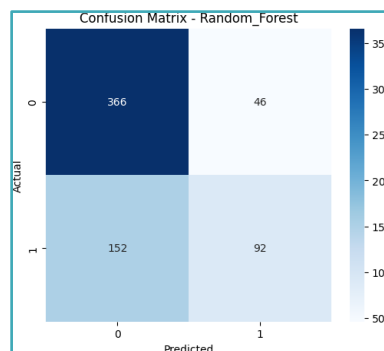


Gambar 4. Grafik Perbandingan Model

Dapat disimpulkan bahwa model *Random Forest* lebih baik dalam meminimalkan kesalahan klasifikasi serta mampu menjaga keseimbangan antara *Precision* dan *Recall*. Oleh karena itu, *Random Forest* dipilih sebagai model final karena memiliki performa yang paling stabil. Dalam penelitian ini, meskipun hasil performa *Random Forest* lebih unggul dibandingkan penelitian sebelumnya, model *Random Forest* masih melakukan banyak kesalahan dalam memprediksi dengan data *testing*. Hal ini ditandai dengan nilai *Recall* yang rendah. Analisis lebih lanjut terhadap nilai *Precision*, *Recall*, dan *F1-Score* menjadi sangat penting untuk memahami performa model secara lebih mendalam. Penelitian ini memiliki perbedaan dengan penelitian sebelumnya yaitu dalam menangani *missing values*, menggunakan metode yang lebih canggih yaitu *IterativeImputer*, sementara pada penelitian sebelumnya tidak dijelaskan. Dapat disimpulkan bahwa penelitian ini unggul dalam persiapan data dan pengaturan model sehingga mendapat akurasi yang lebih tinggi. Penelitian ini tidak hanya menghasilkan peningkatan nilai performa, tetapi juga memberikan pendekatan evaluasi yang lebih matang dibandingkan penelitian sebelumnya.

Evaluasi Model *Random Forest*

Hasil evaluasi model *Random Forest* menghasilkan *confusion matrix* dan classification report yang dapat dilihat di bawah ini. *Confusion matrix* memberikan gambaran yang lebih rinci mengenai jenis kesalahan klasifikasi yang dilakukan oleh model. Kesalahan berupa False Negative, yaitu air layak minum yang diprediksi sebagai tidak layak minum, jumlahnya lebih besar dibandingkan False Positive. Dari 412 data Tidak Layak Minum, sebanyak 366 diprediksi benar dan sebanyak 46 salah diprediksi sebagai Layak Minum, sementara dari 244 data Layak Minum, sebanyak 92 diprediksi benar dan 152 salah diprediksi sebagai Tidak Layak Minum.



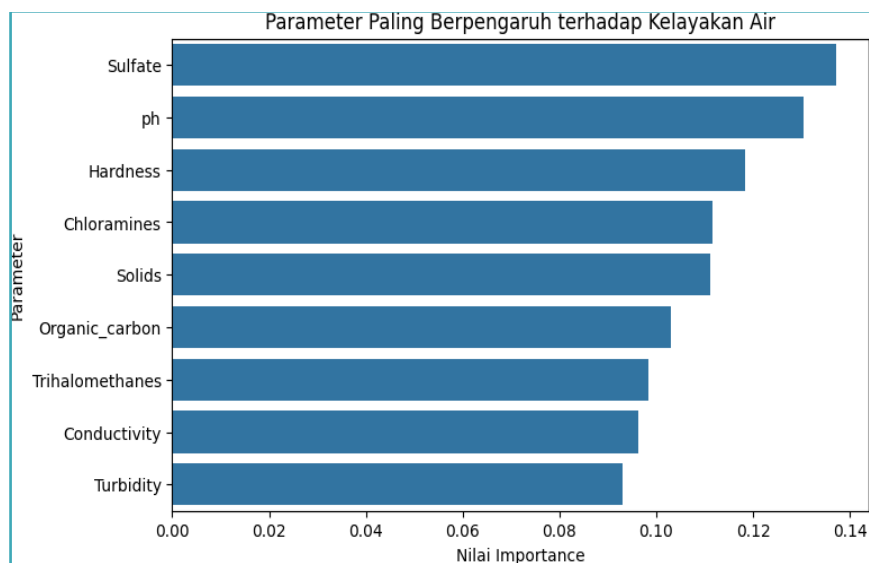
Gambar 5. Confusion Matrix Random Forest

Tabel 3. *Classification Report Random Forest*

	<i>Precision</i>	<i>Recall</i>	<i>F1 Score</i>	<i>Support</i>
0,0	0,71	0,89	0,79	412
1 .0	0,67	0,38	0,48	244
<i>Accuracy</i>			0,70	656
<i>Macro avg</i>	0,69	0,63	0,63	656
<i>Weighted avg</i>	0,69	0,70	0,67	656

Analisis Feature Importance

Dari total 3.276 baris data, 20% atau 656 baris digunakan sebagai data *testing* (x_{test} dan y_{test}), dan 2620 baris sisanya digunakan sebagai data *training* (x_{train} dan y_{train}). Grafik di bawah adalah *feature importance* yang berasal dari proses *training* model *Random Forest*. *Feature importance* berguna untuk mengidentifikasi kontribusi setiap parameter terhadap proses klasifikasi. Hasil *feature importance* menunjukkan bahwa parameter yang mempengaruhi peningkatan akurasi adalah *Sulfate* (0,14), *pH* (0,13), dan *Hardness* (0,12). Selain menjadi parameter yang paling berpengaruh pada penelitian ini, *Sulfate* dan *pH* juga menjadi parameter utama menurut standar WHO. Sedangkan parameter yang memiliki nilai *importance* paling rendah adalah *Turbidity* (0,09) yang disebabkan oleh kecilnya kontribusi fitur terhadap penurunan *Gini Impurity* / *Information Gain*.



Gambar 6. *Feature Importance*

SIMPULAN DAN SARAN

Penelitian ini menunjukkan bahwa prediksi model *Random Forest* terhadap air layak minum dan tidak layak minum mendapatkan performa lebih baik dari 2 model lainnya yaitu *Decision Tree* dan *Naïve Bayes*. Hasil penelitian ini menegaskan bahwa kualitas data memiliki peran yang sangat penting dalam menentukan performa model *Machine Learning*. Tahapan *preprocessing* data, khususnya penanganan *missing values* dan transformasi data, terbukti memberikan kontribusi terhadap peningkatan kinerja model. Dengan penanganan data yang lebih sistematis, model dapat mempelajari pola data secara lebih optimal dibandingkan pendekatan yang kurang memperhatikan kualitas data. Selain itu, hasil evaluasi menggunakan berbagai metrik performa menunjukkan bahwa distribusi data yang tidak seimbang memberikan pengaruh signifikan terhadap hasil klasifikasi.

Model *Random Forest* cenderung memiliki performa yang lebih baik pada kelas mayoritas dibandingkan kelas minoritas. Dengan demikian, penelitian ini tidak hanya menghasilkan model klasifikasi kualitas air, tetapi juga memberikan pemahaman mengenai tantangan yang muncul dalam penerapan *Machine Learning* pada data lingkungan. Hasil yang diperoleh dapat dijadikan sebagai dasar pengembangan penelitian lanjutan maupun implementasi sistem pemantauan kualitas air berbasis data monitoring lingkungan. Peneliti yang akan datang disarankan untuk melakukan eksperimen yang lebih mendalam terhadap pengaruh pengaturan parameter model terhadap hasil klasifikasi dan dapat diarahkan pada penerapan model klasifikasi kualitas air pada sistem monitoring berbasis perangkat nyata. Dengan demikian hasil penelitian memiliki potensi untuk diterapkan secara langsung dalam mendukung pengambilan keputusan di bidang kesehatan dan lingkungan.

Hasil ini juga dapat memberikan dasar yang kuat untuk lebih dikembangkan oleh peneliti selanjutnya dengan menerapkan penambahan fitur atau penyeimbangan data. Selain itu, meskipun dalam hal penanganan *missing value* sudah diterapkan teknik *IterativeImputer*, peneliti selanjutnya diharapkan untuk dapat menjelaskan secara lebih rinci tahapan *preprocessing* data yang digunakan. Kurangnya penjelasan mengenai pengolahan data pada penelitian terdahulu dapat menimbulkan kesulitan dalam mereplikasi hasil penelitian serta berpotensi menimbulkan bias interpretasi terhadap performa model. Oleh karena itu, transparansi metodologi menjadi aspek penting yang perlu diperhatikan dalam penelitian *Machine Learning*.

DAFTAR PUSTAKA

- [1] V. B. Silahooy *et al.*, *Alam dan Perkembangannya*. TOHAR MEDIA, 2024.
- [2] A. R. Alfian, *Krisis Air di Indonesia: Fenomena, Dampak & Solusi*. Suluah Kato Khatulistiwa, 2023.
- [3] A. P. Ekowati and M. F. D. Lusno, "Analisis Capaian dan Tantangan Akses Air Minum Aman di Indonesia Menuju SDGS 6.1.1," *J. Penelit. Inov.*, vol. 5, no. 2, pp. 1707–1714, Apr. 2025, doi: 10.54082/jupin.1538.
- [4] I. Dewata and Y. H. Danhas, *Pencemaran Lingkungan*. PT. RajaGrafindo Persada - Rajawali Pers, 2023.
- [5] C. Ihsan, *Dasar-Dasar Machine Learning: Teori, Algoritma, dan Implementasi*. Greenbook Publisher, 2025.
- [6] D. K. Phasinam, D. A. K. Singh, M. T. Singh, and M. K. Sharma, *Fundamental of Machine Learning*. Blue Rose Publishers, 2022.
- [7] T. V. Geetha and S. Sendhilkumar, *Machine Learning: Concepts, Techniques and Applications*. CRC Press, 2023.
- [8] A. B. Adetunji, O. N. Akande, F. A. Ajala, O. Oyewo, Y. F. Akande, and G. Oluwadara, "House Price Prediction using Random Forest Machine Learning Technique," *Procedia Comput. Sci.*, vol. 199, pp. 806–813, Jan. 2022, doi: 10.1016/j.procs.2022.01.100.
- [9] D. Safitri, "Prediksi Tingkat Stres Mahasiswa Menggunakan Algoritma Decision Tree," *J. Inf. Syst. Appl. Dev.*, vol. 4, no. 1, pp. 188–194, Mar. 2026, doi: 10.26905/jisad.v4i1.16616.
- [10] L. A. Supriyono *et al.*, *Buku Ajar Big Data dan Data Mining: Konsep, Metodologi, dan Aplikasi*. PT. Sonpedia Publishing Indonesia, 2025.
- [11] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylon. J. Mach. Learn.*, vol. 2024, pp. 69–79, Jun. 2024, doi: 10.58496/BJML/2024/007.
- [12] M. W. Nugroho, "Analisis Performa Algoritma Random Forest dalam Mengatasi Overfitting pada Model Prediksi," *J. Teknol. Inf. Dan Komun.*, vol. 9, no. 4, pp. 1562–1571, Dec. 2025, doi: 10.35870/jtik.v9i4.4236.
- [13] O. Castro, P. Bruneau, J.-S. Sottet, and D. Torregrossa, "Landscape of High-Performance Python to Develop Data Science and Machine Learning Applications," *ACM Comput. Surv.*, vol. 56, no. 3, pp. 65:1–65:30, Oct. 2023, doi: 10.1145/3617588.

- [14] L. F. M. Horhoruw, A. Pratama, Y. Asri, and N. P., *Eksplorasi Machine Learning dengan Scikit-Learn: Strategi Belajar Machine Learning*. Uwais Inspirasi Indonesia, 2024.
- [15] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Glob. Transit. Proc.*, vol. 3, no. 1, pp. 91–99, Jun. 2022, doi: 10.1016/j.gltp.2022.04.020.
- [16] S. Amaliah, M. Nusrang, and A. Aswi, "Penerapan Metode Random Forest untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng," *Variansi J. Stat. Its Appl. Teach. Res.*, vol. 4, no. 3, pp. 121–127, Dec. 2022, doi: 10.35580/variansiunm31.
- [17] M. A. Ridla and A. A. Bawazin, "Klasifikasi Gangguan Tidur Berdasarkan Gaya Hidup Menggunakan Algoritma Decision Tree," *J. Inf. Syst. Appl. Dev.*, vol. 3, no. 2, pp. 119–126, Sep. 2025, doi: 10.26905/jisad.v3i2.16038.
- [18] J. R. Aryo Michael, "Klasifikasi Cacat Biji Kopi Menggunakan Metode Transfer Learning dengan Hyperparameter Tuning Gridsearch," *J. Teknol. Dan Manaj. Inform.*, vol. 9, no. 1, pp. 37–45, 2023.
- [19] N. Prayoga and M. A. Ridla, "Analisis Prediksi Diabetes Menggunakan Decision Tree pada Dataset Diabetes Pima Indians," *J. Inf. Syst. Appl. Dev.*, vol. 4, no. 1, pp. 146–153, Mar. 2026, doi: 10.26905/jisad.v4i1.16494.