p-ISSN: 2301-5373
e-ISSN: 2654-5101

Optimasi *Hyperparameter Random Forest* untuk Prediksi Kualitas Air

Lutfi Julpian[1], Alam Rahmatulloh[2*]

[a]Informatika Departement, Faculty of Engineering, Siliwangi University
Jl. Siliwangi No.24, Tasikmalaya, West Java, Indonesia 46115
217006518@student.unisl.ac.id, alam@unsil.ac.id

Abstrak

Water is an essential source of life for all living beings, making the preservation of water quality crucial for human health and the sustainability of ecosystems. However, population growth, industrial development, and other economic activities have led to a decline in water quality and an increase in pollution. To address this issue, early detection through the application of data mining techniques, particularly classification, is necessary to predict the quality of consumable water. This study aims to enhance the accuracy of water quality prediction by applying hyperparameter optimization techniques using the Grid Search method on the Random Forest algorithm. The results show that hyperparameter optimization improved water quality prediction accuracy from 88.33% to 91.32%. This improvement underscores the effectiveness of machine learning techniques in monitoring water quality and contributes to better decision-making to safeguard water resources and protect public health. Furthermore, this research provides insights into the importance of data mining techniques in identifying relevant patterns in water quality data, thereby helping to prevent health risks associated with contaminated water.

Keywords: *Water quality, Data mining, Hyperparameter optimization, Random Forest, Machine learning*

1. Pendahuluan

Air merupakan sumber kehidupan yang tak tergantikan bagi seluruh makhluk hidup[1]. Sekitar 60% berat tubuh manusia terdiri dari air, yang menunjukkan betapa pentingnya air bagi kelangsungan hidup[2]. Kualitas air yang baik tidak hanya menjamin kesehatan manusia, tetapi juga keberlangsungan ekosistem serta mendukung berbagai aktivitas ekonomi seperti pertanian dan industri. Untuk menjaga kualitas air agar tetap bersih dan bebas dari zat-zat pencemar seperti zat asam dan basa, kita perlu melakukan upaya perlindungan yang serius[3]. Kesehatan masyarakat di suatu wilayah sangat dipengaruhi oleh kualitas air yang dikonsumsi[4]. Upaya perlindungan terhadap sumber daya air diperlukan agar masyarakat dapat mengakses air yang bersih dan layak konsumsi untuk mendukung kehidupan sehari-hari.

Dengan peningkatan jumlah penduduk, pertumbuhan industri, peningkatan ekonomi, dan peningkatan standar hidup telah memberikan tekanan yang besar pada sumber daya air. Akibatnya, terjadi ketidakseimbangan antara pemanfaatan dan ketersediaan air, baik dalam jumlah maupun kualitas[5]. Hal ini menyebabkan penurunan kualitas air secara signifikan dan berkurangnya ketersediaan air untuk memenuhi kebutuhan yang terus meningkat. Pencemaran air menjadi penyebab utama berbagai penyakit dan masalah kesehatan masyarakat. Pencemaran air mengancam kesehatan manusia melalui berbagai penyakit yang disebabkan oleh konsumsi air yang tercemar dan paparan terhadap lingkungan yang terkontaminasi [6].

Untuk mengetahui kualitas air yang layak dikonsumsi oleh masyarakat luas, diperlukan deteksi dini melalui penerapan teknik data mining. Salah satu teknik yang dapat digunakan adalah klasifikasi, yaitu proses untuk menemukan model atau fungsi yang dapat menjelaskan dan membedakan konsep atau kelas data[7]. Data mining menyediakan alat untuk menggali informasi yang kompleks dan tersembunyi dalam data yang dihasilkan dari simulasi komputer dan eksperimen daring[8]. Data mining adalah proses penggalian informasi dan pola yang berharga dari kumpulan data yang sangat besar. Proses ini mencakup beberapa tahap, yaitu pengumpulan data, ekstraksi data, analisis data, serta penggunaan statistik [9].

Sebelumnya, telah banyak penelitian yang mengkaji deteksi dini kualitas air melalui penerapan data mining. Salah satu penelitian tersebut adalah tentang prediksi kualitas air menggunakan teknik *machine learning* [1]. Penelitian ini membandingkan tiga algoritma, yaitu *Random Forest, Decision Tree,* dan *Gradient Boosting*. Hasil pengujian menunjukkan bahwa algoritma *Random Forest* memberikan akurasi terbaik, yaitu sebesar 88,33%.

Penelitian selanjutnya adalah penerapan sistem prediksi kualitas air yang dapat dikonsumsi dengan menerapkan teknik *machine learning* [10]. Dalam penelitian ini, digunakan algoritma *K-Nearest Neighbor* (KNN) untuk menganalisis data kualitas air. Hasil dari data pengujian menunjukkan bahwa algoritma KNN mencapai akurasi sebesar 85,24%. Akurasi yang cukup tinggi ini menunjukkan bahwa teknik *machine learning* dapat diandalkan dalam memprediksi kualitas.

Dalam penelitian sebelumnya, teknik *machine learning* untuk predisksi kualitas air yang paling bagus yaitu *Random Forest* menunjukkan akurasi tertinggi sebesar 88,33%. Namun, hal ini masih dapat ditingkatkan melalui optimasi *hyperparameter*. Optimasi *hyperparameter* adalah proses menemukan kombinasi optimal *hyperparameter* yang meminimalkan kesalahan prediksi atau memaksimalkan akurasi prediksi[11]. Algoritma optimasi *hyperparameter* secara otomatis menemukan konfigurasi *hyperparameter* yang memberikan kinerja baik untuk algoritma *machine learning*[12]. Metode optimasi yang diterapkan dalam penelitian ini adalah metode *Grid Search*. Metode *Grid Search* dipilih karena sifatnya yang sangat teliti dan kemampuannya untuk menjamin hasil yang optimal[13]. Meskipun metode ini membutuhkan waktu komputasi yang lebih lama dibandingkan metode lain, namun ketelitiannya membuat metode ini sangat cocok digunakan ketika akurasi hasil sangat penting. Dalam metode ini, sejumlah kombinasi nilai hyperparameter yang telah ditentukan disusun dalam bentuk grid. Sehingga memungkinkan eksplorasi yang lebih efektif terhadap ruang pencarian *hyperparameter*. Dengan demikian, penelitian ini bertujuan untuk meningkatkan akurasi prediksi kualitas air dengan memanfaatkan teknik optimasi yang tepat.

Penelitian ini memberikan manfaat dalam meningkatkan kualitas prediksi air yang layak dikonsumsi oleh masyarakat. Dengan memanfaatkan teknik data mining dan machine learning, khususnya melalui optimasi *hyperparameter* menggunakan metode *Grid Search*, penelitian ini berhasil menunjukkan peningkatan akurasi yang signifikan sebesar 91,32%. Penelitian lain juga menunjukkan bahwa penerapan *Grid Search* CV dapat meningkatkan hasil klasifikasi pada pemantauan kualitas udara [9].

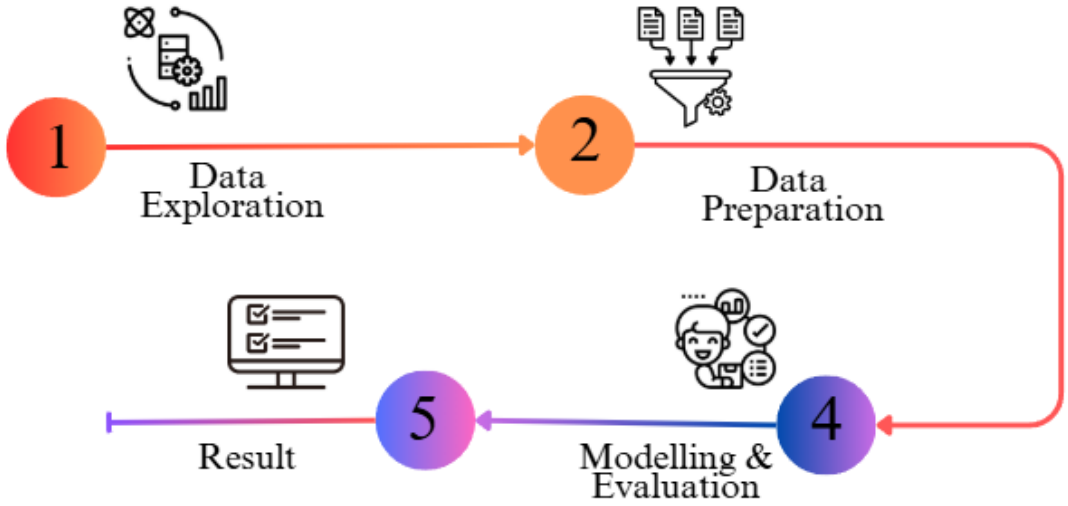
Teknik optimasi hyperparameter ini secara signifikan meningkatkan akurasi model prediksi kualitas air, yang sebelumnya telah menunjukkan kinerja memadai dengan algoritma Random Forest sebesar 88,33% [1]. Peningkatan akurasi ini memberikan manfaat yang sangat besar, terutama dalam mendukung pengambilan keputusan yang lebih tepat dan berbasis data untuk menjaga kualitas sumber daya air. Dengan prediksi yang lebih akurat, pihak berwenang dapat mengidentifikasi risiko kontaminasi secara dini, menerapkan langkah-langkah pencegahan yang efektif, serta merumuskan kebijakan yang relevan untuk melindungi kesehatan masyarakat. Selain itu, kemampuan model yang lebih unggul ini juga berperan penting dalam mencegah penyebaran penyakit yang disebabkan oleh air tercemar, sehingga memastikan ketersediaan air bersih yang aman bagi kebutuhan sehari-hari. Hal ini menunjukkan betapa pentingnya inovasi dan pengembangan metode analitik dalam manajemen kualitas air secara berkelanjutan.

2. Metode

Metodologi yang digunakan dalam penelitian ini ditampilkan pada Gambar 1. Kebaruan dari penelitian ini terletak pada penerapan *hyperparameter* tuning menggunakan metode *Grid Search* pada algoritma *Random Forest Classifier*. Skenario yang digunakan untuk mencapai akurasi

p-ISSN: 2301-5373
e-ISSN: 2654-5101

terbaik mengikuti metode yang diterapkan dalam penelitian sebelumnya[1], dengan membagi kumpulan data menjadi 90% untuk data latih dan 10% untuk data uji.



Gambar 1. Tahapan penelitian

2.1. Data *Exploration*

Pada penelitian ini, data yang digunakan merupakan data sekunder yang diperoleh peneliti melalui sumber tidak langsung atau pihak ketiga. Data tersebut diambil dari Dataset *Water-Potability-Datasets* yang tersedia di platform Kaggle (www.kaggle.com/datasets/ibrahimzaitoun/water-potability-datasets/code). Dataset ini berisi berbagai informasi yang relevan dengan kualitas air, mencakup parameter-parameter penting yang digunakan untuk menilai potabilitas air, yaitu apakah air tersebut layak dikonsumsi oleh manusia atau tidak. Parameter-parameter ini meliputi berbagai aspek fisik, kimia, dan biologi, yang memengaruhi kualitas air secara keseluruhan. Informasi dalam dataset ini menjadi dasar untuk memahami berbagai faktor yang menentukan potabilitas air dan membantu dalam pengambilan keputusan yang berbasis data.

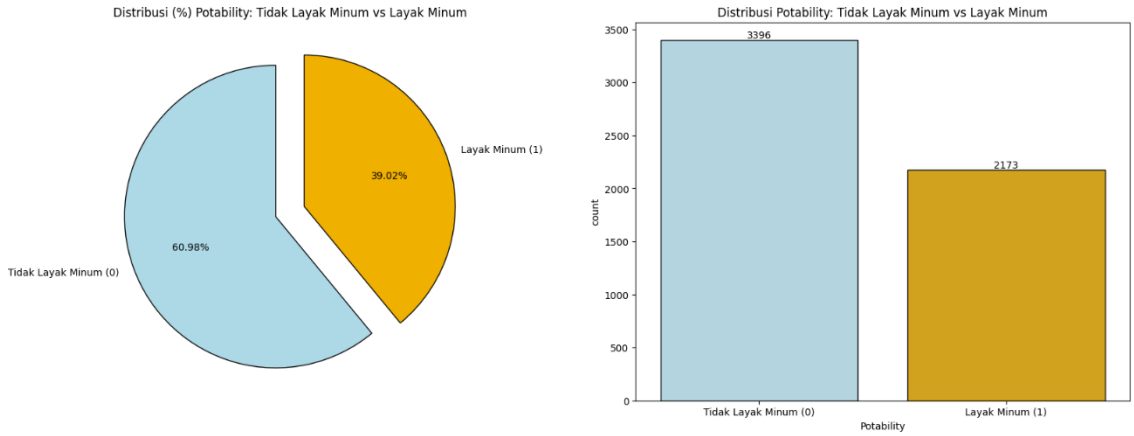
Tabel 1. Deskripsi informasi data

Nama Atribut Kolom	Deskripsi Data	Kategori Tipe Data
ph	Air(0 sampai 14)	Numerik (float)
Hardness	Kekerasan	Numerik (float)
Solids	Padatan	Numerik (float)
Chloramines	Kloraamin	Numerik (float)
Sulfate	Sulfat	Numerik (float)
Conductivity	Konduktivitas	Numerik (float)
Organic Carbon	Karbon Organik	Numerik (float)
Trihalomethanes	*Trihalomethanes*	Numerik (float)
Turbidity	Kekeruhan	Numerik (float)
potability	Sifat dapat diminum	Numerik(int)

Dataset ini memiliki struktur data yang terdiri dari 5569 baris dan 10 kolom, yang mencakup variabel-variabel penting yang digunakan dalam analisis dan prediksi. Setiap baris mewakili sampel data, sementara setiap kolom berisi informasi terkait parameter kualitas air seperti tingkat

pH, konsentrasi zat kimia tertentu, dan tingkat kontaminasi. Dengan ukuran dataset yang cukup besar dan variasi data yang beragam, penelitian ini dapat memberikan hasil analisis yang lebih akurat. Informasi lebih rinci mengenai variabel-variabel tersebut dapat ditemukan pada Tabel 1 dalam laporan penelitian. Dataset ini menjadi dasar yang signifikan dalam penerapan teknik data mining, khususnya untuk melakukan prediksi kualitas air yang lebih komprehensif dan mendukung pencapaian tujuan penelitian.

Pada tahap *data exploration* bertujuan untuk memahami karakteristik dataset secara menyeluruh sebelum proses *data preparation* dilakukan. Langkah-langkah yang dilakukan meliputi membaca dataset dengan mengimpor data, menampilkan struktur data untuk memeriksa jenis variabel, serta menganalisis karakteristik setiap fitur dan mengidentifikasi variabel yang relevan. Selain itu, dilakukan pemeriksaan *missing value* untuk menangani data yang hilang. Distribusi variabel target juga dianalisis guna memastikan keseimbangan data yang bisa dilihat pada Gambar 2, karena distribusi yang tidak seimbang dapat mempengaruhi akurasi model. Eksplorasi ini memberikan dasar yang kuat dalam memahami data dan merancang strategi pemodelan yang optimal.



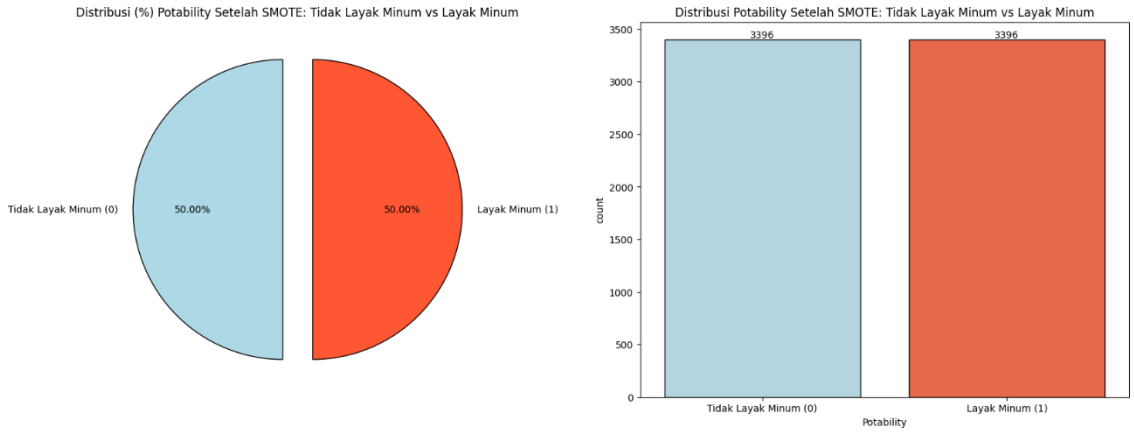
Gambar 2. Distribusi variabel

Selain analisis awal, dilakukan pula visualisasi data untuk mengidentifikasi pola, anomali, dan hubungan antar variabel yang dapat memengaruhi hasil prediksi. Teknik seperti histogram, *scatter plot*, dan heatmap digunakan untuk mengevaluasi korelasi antar fitur, sehingga dapat membantu dalam proses *feature selection*. Pemahaman mendalam terhadap dataset ini memungkinkan identifikasi fitur-fitur yang paling signifikan, serta memberikan gambaran mengenai distribusi data yang *outlier* atau *noise*, yang nantinya akan ditangani pada tahap data *preparation*. Langkah eksplorasi yang menyeluruh ini tidak hanya membantu meminimalkan kesalahan pada tahap pemodelan tetapi juga memastikan bahwa algoritma yang digunakan dapat bekerja dengan data yang lebih bersih dan terstruktur.

2.2. Data *Preparation*

Tahapan Data *Preparation* merupakan bagian penting dalam proses analisis data yang bertujuan mempersiapkan dataset sebelum memasuki proses pemodelan. Pada tahap ini, dilakukan berbagai kegiatan seperti pembersihan data, pengisian *missing value*, penerapan teknik *Synthetic Minority Over-sampling Technique* (SMOTE*)* dengan tujuan mengontrol pemerataan distribusi data pada suatu kelas dalam dataset dengan membuat sampel buatan dari kelas minoritas[14]. Hasil Teknik SMOTE bisa dilihat pada Gambar 3, serta pembagian data menjadi data latih (training set) dan data uji (test set). Pada penelitian ini setelah dilakukan *preparation* data maka data yang sebelumnya 5569 data mejadi 4317 data. Hal ini dilakukan untuk memastikan bahwa dataset siap digunakan dalam pemodelan dengan kualitas data yang terjaga.

p-ISSN: 2301-5373
e-ISSN: 2654-5101



Gambar 3. Penerapan SMOTE pada variabel target

2.3. *Modelling & Evaluation*

Tahapan dalam proses analisis data ini bertujuan untuk membangun model prediksi serta mengevaluasi kinerjanya. Pada tahap ini, peneliti menggunakan algoritma *Random Forest Classifier* yang dioptimalkan dengan *hyperparameter* melalui metode *Grid Search*. Setelah model berhasil dibangun, langkah selanjutnya adalah melakukan evaluasi terhadap performa model menggunakan data uji. Evaluasi ini menghasilkan ukuran akurasi sebagai indikator kinerja model.

Untuk mengukurkinerja classifier, penelitian ini menggunakan *confusion matrix*, dengan perhitungannya tercantum pada Tabel 2, sebagaimana dilakukan dalam penelitian sebelumnya[1]. *Confusion matrix* didefinisikan sebagai matriks yang menyajikan perbandingan antara instans kelas yang diprediksi dan instans kelas yang sebenarnya[15].

Tabel 2. *Confusion Matrix*

		Predicted Label	
		0 (Sehat)	1 (Terinfeksi)
Actual Label	0 (Sehat)	*True Positive (TP)*	*False Positive (FP)*
	1 (terinfeksi)	*False Negative (FN)*	*True Negative (TN)*

Perhitungan menggunakan confusion matrix digunakan untuk menentukan nilai akurasi, presisi, *recall*, dan *F1-score*, yang semuanya merupakan metrik penting untuk mengevaluasi efektivitas model dalam menangani data dengan distribusi kelas tertentu[14][1][15]. Dengan menggunakan metrik ini, penelitian dapat mengevaluasi seberapa baik model dalam mengklasifikasikan data, termasuk dalam mendeteksi kelas minoritas yang sering kali lebih sulit untuk diprediksi.

Dalam konteks ini, terdapat dua kelas: 0 (Sehat) yang mewakili kelas negatif atau kondisi tidak terinfeksi, dan 1 (Terinfeksi) yang mewakili kelas positif atau kondisi terinfeksi. Berdasarkan hasil prediksi model, terdapat empat istilah utama yang menggambarkan performa prediksi:

a. *True Positive* (TP) adalah situasi di mana label aktual adalah positif (Terinfeksi) dan model berhasil memprediksi dengan benar sebagai positif
b. *True Negative* (TN) adalah situasi di mana label aktual adalah negatif (Sehat) dan model memprediksi dengan benar sebagai negatif.
c. *False Positive* (FP) terjadi ketika label aktual adalah negatif (Sehat), tetapi model salah memprediksi sebagai positif (Terinfeksi). Kesalahan ini dapat berimplikasi pada tindakan yang tidak perlu bagi individu yang sehat.

d. *False Negative* (FN) terjadi ketika label aktual adalah positif (Terinfeksi), tetapi model salah memprediksi sebagai negatif (Sehat). Kesalahan ini berisiko karena individu yang terinfeksi dapat tidak terdeteksi dan tidak mendapatkan perawatan yang diperlukan

Confusion matrix membantu memberikan gambaran lengkap tentang kinerja model dalam memisahkan kedua kelas, baik dalam hal keberhasilan prediksi maupun kesalahan yang dibuat. Dengan menggunakan *confusion matrix*, kita dapat lebih mudah mengidentifikasi di mana model melakukan kesalahan dan berapa besar dampaknya terhadap hasil akhir. Berikut formula untuk menghitung akurasi:

$$Akurasi\ (\%) = \frac{TP + TN}{TP + TN + FP + FN}\ x\ 100 \qquad 1$$

Akurasi mengukur seberapa baik model dalam memprediksi dengan benar, baik untuk kelas positif maupun negatif. Namun, dalam kasus data yang tidak seimbang, misalnya jumlah data sehat jauh lebih banyak daripada yang terinfeksi, akurasi saja mungkin tidak memberikan gambaran yang memadai tentang kinerja model. Oleh karena itu, penting untuk juga mempertimbangkan metrik lain seperti *precision, recall, dan F1-score* untuk evaluasi yang lebih komprehensif.

3. Hasil dan Pembahasan

Dataset Water Quality digunakan dalam penelitian ini untuk memprediksi potabilitas air atau kelayakan air minum. Data ini diperoleh dari situs *Kaggle* dan menggunakan dataset "*Water-Potability-Datasets*" dalam format CSV. Dataset ini berisi beberapa fitur kimia dan fisika yang digunakan untuk menentukan apakah air layak diminum atau tidak. Fitur-fitur tersebut mencakup pH, *Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes*, dan *Turbidity*, dengan target variabel berupa *Potability*, yang memiliki dua nilai: layak minum (1) dan tidak layak minum (0).

Dataset ini memberikan gambaran mendetail tentang kualitas air melalui pengukuran paramert-parameter yang relavan. Setiap fitur dalam dataset mencerminkan karakteristik spesifik air yang dapat digunakan untuk menilai dampaknya terhadap kesehatan manusia, sementara *Hardness* mengukur kadar mineral seperti kalsium dan megnesium. Dengan, memanfaatkan fitur-fitur dan kimia air terhadap kelayakan konsumsinya, sehingga mendukung pengambilan keputusan data dalam pengelolaan.

Pada penelitian ini, proses optimasi *hyperparameter* dilakukan pada beberapa parameter kunci dari *Random Forest,* termasuk:

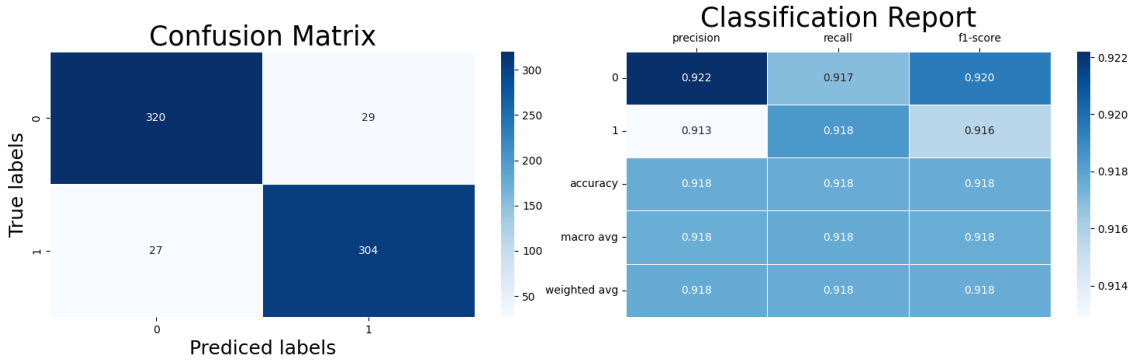
a. *n_estimator*s: [50, 100, 200] Jumlah pohon dalam hutan.
b. *max_depth*: [*None,* 10, 20] Kedalaman maksimal setiap pohon.
c. *min_samples_split*: [2, 5, 10] Jumlah minimal sampel yang dibutuhkan untuk memisahkan node.
d. *min_samples_leaf*: [1, 2, 4] Jumlah minimal sampel yang dibutuhkan pada setiap leaf node.

Hasil dari optimasi menunjukkan bahwa kombinasi *hyperparameter* terbaik menghasilkan akurasi sebesar 91.80% selama 5-*fold cross-validation*. Ini berarti model memiliki kemampuan yang cukup baik untuk memprediksi kelayakan air minum, dengan performa yang cukup konsisten di seluruh pembagian data dalam proses validasi silang.

3.1. Implementasi dengan Evaluasi *Confosion Matrix*

Pada tahap ini, evaluasi dilakukan menggunakan algoritma *Random Forest*. Setelah melakukan optimasi *hyperparameter*, hasil evaluasi dilakukan dengan menggunakan confusion matrix untuk menilai performa model bisa dilihat pada Gambar 4. Tujuannya adalah untuk melihat apakah akurasi model meningkat setelah optimasi dibandingkan dengan akurasi sebelum optimasi. Proses optimasi dan evaluasi ini dilakukan dengan menggunakan tools data mining seperti *Google Colab* dan *Python*, yang menyediakan *library* untuk melakukan optimasi, evaluasi, serta visualisasi hasil secara efektif dan efisien.

p-ISSN: 2301-5373
e-ISSN: 2654-5101



Gambar 4. *Confusion Matrix Random Forest*

Model ini mencapai akurasi 91.3%, yang dihitung dengan membagi total prediksi benar dengan total keseluruhan prediksi. Perhitungannya adalah:

$$Akurasi\ (\%) = \frac{316 + 305}{316 + 305 + 26 + 33}\ x\ 100 \quad (2)$$

$$Akurasi\ (\%) = \frac{621}{680}\ x\ 100 = 91{,}3\% \quad (3)$$

Precision mengukur seberapa banyak prediksi positif yang benar. Untuk kelas negatif (0), precision-nya adalah 92.4%, artinya dari semua prediksi negatif, 92.4% benar-benar negatif. Untuk kelas positif (1), *precision*-nya mencapai 90.2%, menunjukkan bahwa dari semua prediksi positif, 90.2% adalah positif sebenarnya.

3.2. Perbandingan Akurasi

Dalam upaya untuk mengevaluasi dan membandingkan efektivitas berbagai algoritma dalam memprediksi kualitas air, Tabel 3 menyajikan perbandingan akurasi dan *Precision* dari beberapa penelitian yang telah dilakukan. Nilai pada baris 1 diperoleh dari penelitian[10], sedangkan nilai pada baris 2 berasal dari penelitian [1]. Penelitian ini mencakup algoritma *K-Nearest Neighbor* (KNN), *Random Forest*, serta optimasi hyperparameter yang diterapkan pada model *Random Forest*.

Tabel 3. *Perbandingan Akurasi dan Precision Algoritma Prediksi Kualitas Air*

NO	Penelitian	Akurasi(%)	*Precision* Negatif(%)	*Precision* Positif(%)
1	KNN (2022) [10]	85,24	-	-
2	*Random Forest* (2024) [1]	88,33	90,0	88,0
3	***Optimasi Hyperparameter Random Forest***	**91,80**	**92,2**	**91,2**

Tabel 3 menunjukkan bahwa optimasi *hyperparameter* memberikan peningkatan signifikan pada akurasi dan *precision* jika dibandingkan dengan algoritma KNN dan *Random Forest* tanpa optimasi. Hal ini menunjukkan pentingnya pengembangan dan penerapan teknik optimasi dalam meningkatkan performa model prediksi kualitas air, serta membuka peluang untuk penelitian lebih lanjut dalam sistem prediksi yang lebih dinamis dan akurat.

4. Kesimpulan

Penelitian ini menegaskan pentingnya pengelolaan kualitas air yang baik untuk kesehatan masyarakat dan keberlanjutan lingkungan. Melalui penerapan teknik data mining dan *machine learning*, khususnya dengan *optimasi hyperparameter* menggunakan metode *Grid Search* pada algoritma *Random Forest*, akurasi prediksi kualitas air dapat ditingkatkan dari 88,33% menjadi

91,80%. Peningkatan ini menunjukkan bahwa teknik-teknik canggih dalam analisis data mampu memberikan kontribusi signifikan dalam pemantauan dan pengelolaan kualitas air. Dengan pemahaman yang lebih baik mengenai kualitas air, diharapkan dapat mendorong tindakan *preventif* yang lebih efektif dalam melindungi sumber daya air dan kesehatan masyarakat.

Untuk penelitian selanjutnya, disarankan agar peneliti mengeksplorasi lebih banyak algoritma *machine learning* dan metode *optimasi hyperparameter* lainnya seperti *random search, bayesian optimization* dan *genetic algorthm* untuk memperbandingkan kinerjanya dalam prediksi kualitas air. Penelitian selanjutnya juga dapat berfokus pada pengembangan model yang lebih dinamis dan responsif terhadap perubahan kualitas air secara *real-time*.

Daftar Pustaka

[1] N. Maulidah, M. Maulidah, R. Supriyadi, H. Nalatissifa, S. Diantika, and A. Fauzi, “Prediksi Kulitas Air menggunakan Metode Random Forest, Decision Tree, dan Gradient Boosting,” *JURNAL KHATULISTIWA INFORMATIKA*, vol. 12, pp. 1–6, Jun. 2024.

[2] F. Kordbacheh and & Golnaz Heidari, “Water Pollutants and Approaches for Their Removal,” *Mater. Chem. Horizons*, vol. 2023, no. 2, pp. 139–153, Jul. 2023, doi: 10.22128/MCH.2023.684.1039.

[3] S. A. Akbar, D. B. Kalbuadi, and A. Yudhana, “Online Monitoring Kualitas Air Waduk Berbasis Thingspeak,” *Transmisi*, vol. 21, no. 4, pp. 109–115, Oct. 2019, doi: 10.14710/transmisi.21.4.109-115.

[4] M. Fida, P. Li, Y. Wang, S. M. K. Alam, and A. Nsabimana, “Water Contamination and Human Health Risks in Pakistan: A Review,” Sep. 01, 2023, *Springer Science and Business Media B.V.* doi: 10.1007/s12403-022-00512-1.

[5] I. Desti and A. Ula, “Analisis Sumber Daya Alam Air,” *Jurnal Sains Edukatika Indonesia (JSEI)*, vol. 3, no. 2, pp. 17–24, 2021.

[6] Z. KILIÇ, “Water Pollution: Causes, Negative Effects and Prevention Methods,” *İstanbul Sabahattin Zaim Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, vol. 3, no. 2, pp. 129–132, Aug. 2021, doi: 10.47769/izufbed.862679.

[7] I. Budiman, Mauliadi, and R. Ramadina, “Penerapan Fungsi Data Mining Klasifikasi untuk Prediksi Masa Studi Mahasiswa Tepat Waktu pada Sistem Informasi Akademik Perguruan Tinggi,” *IJCCS*, vol. 7, no. 1, pp. 1–5, Apr. 2015.

[8] X. Shu and Y. Ye, “Knowledge Discovery: Methods from data mining and machine learning,” *Soc Sci Res*, vol. 110, Feb. 2023, doi: 10.1016/j.ssresearch.2022.102817.

[9] I. Gede Iwan Sudipa *et al.*, *Data Mining*. Padang Sumatera barat: PT GLOBAL EKSEKUTIF TEKNOLOGI, 2023. [Online]. Available: www.globaleksekutifteknologi.co.id

[10] H. Said, N. Hafifah Matondang, and H. N. Irmanda, “Sistem Prediksi Kualitas Air Yang Dapat Dikonsumsi Dengan Menerapkan Algoritma K-Nearest Neighbor,” Jakarta, Apr. 2022.

[11] Y. Zhao, W. Zhang, and X. Liu, “Grid search with a weighted error function: Hyper-parameter optimization for financial time series forecasting,” *Appl Soft Comput*, vol. 154, Mar. 2024, doi: 10.1016/j.asoc.2024.111362.

p-ISSN: 2301-5373
e-ISSN: 2654-5101

[12] B. Bischl *et al.*, “Hyperparameter Optimization: Foundations, Algorithms, Best Practices and Open Challenges.”

[13] M. Ogunsanya, J. Isichei, and S. Desai, “Grid Search Hyperparameter Tuning in Additive Manufacturing Processes,” 2023, [Online]. Available: www.sciencedirect.com

[14] S. Keputusan Direktur Jenderal Pendidikan Tinggi, dan Teknologi Nomor, N. Sharfina, and N. Ghaniaviyanto Ramadhan, “Terakreditasi SINTA Peringkat 3 Analisis SMOTE Pada Klasifikasi Hepatitis C Berbasis Random Forest dan Naïve Bayes,” Purwokerto, Jun. 2023.

[15] I. Markoulidakis, I. Rallis, I. Georgoulas, G. Kopsiaftis, A. Doulamis, and N. Doulamis, “Multiclass Confusion Matrix Reduction Method and Its Application on Net Promoter Score Classification Problem,” *Technologies (Basel)*, vol. 9, no. 4, Dec. 2021, doi: 10.3390/technologies9040081.

This page is intentionally left blank.