Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2647−2658
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7055

Klasifikasi Kelayakan Air Minum Mengkombinasikan Algoritma Random Forest dengan Teknik Optimasi Bayesian

Aditya Aqil Darmawan[1,*],Ishak Bintang D[1], Yani Parti Astuti[2], Agus Winarno[2]

[1] Fakultas Ilmu Komputer, Teknik Informatika,Universitas Dian Nuswantoro, Semarang, Indonesia
[2] Fakultas Ilmu Komputer, Sistem Informasi,Universitas Dian Nuswantoro, Semarang, Indonesia
Email: [1,*]111202113390@mhs.dinus.ac.id, [2]111202113763@mhs.dinus.ac.id, [3] yanipartiastuti@dsn.dinus.ac.id , [4]agus.winarno@dsn.dinus.ac.id , [5]
Email Penulis Korespondensi: 111202113390@mhs.dinus.ac.id
Submitted: **23/02/2025**; Accepted: **23/03/2025**; Published**: 24/03/2025**

Abstrak−Kualitas air minum yang bersih dan aman sangat penting untuk kesehatan masyarakat, namun pencemaran lingkungan akibat limbah industri, domestik, dan urbanisasi telah menyebabkan penurunan kualitas air secara signifikan. Metode manual dalam analisis kualitas air, seperti Water Quality Index (WQI) dan STORET, memiliki keterbatasan dalam efisiensi dan ketepatan hasil. Oleh karena itu, penelitian ini mengusulkan sistem klasifikasi berbasis pembelajaran mesin untuk menentukan kelayakan air minum secara lebih akurat dan efisien. Dataset Water Potability dari Kaggle, yang terdiri dari 3.276 sampel dengan sembilan parameter utama, digunakan dalam penelitian ini. Analisis awal menunjukkan sebagian besar fitur memiliki distribusi hampir normal, meskipun beberapa variabel seperti Solids dan Conductivity mengalami skewness ke kanan akibat nilai ekstrem. Korelasi antar variabel menunjukkan tidak adanya hubungan linear signifikan antara parameter kualitas air. Tahap preprocessing mencakup imputasi data yang hilang menggunakan metode mean, normalisasi, rekayasa fitur, serta oversampling dengan SMOTE untuk mengatasi ketidakseimbangan kelas. Model pembelajaran mesin yang digunakan meliputi LightGBM, Random Forest, XGBoost, dan CatBoost, dengan optimasi model menggunakan Bayesian Search CV yang meningkatkan performa, terutama pada Random Forest. Hasil eksperimen menunjukkan bahwa model Random Forest setelah optimasi memiliki performa terbaik dengan akurasi 85,38%, precision 85,86%, recall 85,38%, dan F1-score 85,37%. Meskipun demikian, masih terdapat beberapa kesalahan klasifikasi, terutama dalam mendeteksi sampel air yang layak konsumsi, yang menunjukkan bahwa metode ensemble learning dapat digunakan secara efektif untuk mengevaluasi kelayakan air minum.

Kata Kunci: Kualitas Air; Machine Learning; Random Forest; XGBoost; SMOTE; Bayesian Search CV

Abstract−The quality of clean and safe drinking water is crucial for public health; however, environmental pollution from industrial waste, domestic waste, and urbanization has significantly deteriorated water quality. Manual methods for water quality analysis, such as the Water Quality Index (WQI) and STORET, have limitations in efficiency and accuracy. Therefore, this study proposes a machine learning-based classification system to determine the potability of drinking water more accurately and efficiently. The Water Potability dataset from Kaggle, consisting of 3,276 samples with nine key parameters, was used in this research. Initial analysis showed that most features had a nearly normal distribution, although some variables, such as Solids and Conductivity, exhibited right-skewness due to extreme values. Correlation analysis revealed no significant linear relationships between water quality parameters. The preprocessing stage included missing data imputation using the mean method, normalization, feature engineering, and oversampling with SMOTE to address class imbalance. The machine learning models used in this study include LightGBM, Random Forest, XGBoost, and CatBoost, with model optimization performed using Bayesian Search CV, which improved performance, particularly for Random Forest. Experimental results showed that the optimized Random Forest model achieved the best performance with an accuracy of 85.38%, precision of 85.86%, recall of 85.38%, and an F1-score of 85.37%. However, some misclassifications remained, especially in detecting potable water samples, indicating that ensemble learning methods can be effectively used to evaluate drinking water potability.

Keywords: Water Quality; Machine Learning; Random Forest; XGBoost; SMOTE; Bayesian Search CV

1. PENDAHULUAN

Hak atas air bersih dan aman merupakan kebutuhan mendasar manusia, namun kualitas air semakin terancam akibat pencemaran, aktivitas industri, serta kurangnya kesadaran masyarakat dalam menjaga lingkungan. *WHO* mencatat bahwa sekitar 2,2 juta orang meninggal setiap tahun akibat penyakit yang ditularkan melalui air yang terkontaminasi, sementara 800 juta orang masih kesulitan memperoleh air layak minum. *Sustainable Development Goals (SDGs)* menekankan pentingnya akses terhadap air dan sanitasi yang berkelanjutan, tetapi tantangan utama yang dihadapi adalah meningkatnya polusi air akibat urbanisasi dan limbah industri[1].

Penilaian kualitas air umumnya dilakukan dengan metode manual seperti *Water Quality Index (WQI)* dan STORET[2], yang membutuhkan waktu lama serta analisis yang mendalam. Sayangnya, metode ini kurang efisien dalam situasi yang memerlukan keputusan cepat terkait kelayakan air. Oleh karena itu, diperlukan sistem otomatis berbasis pembelajaran mesin yang dapat mengklasifikasikan kualitas air secara lebih akurat dan efisien. Dengan pemanfaatan kecerdasan buatan dalam pemantauan kualitas air, pencemaran dapat dideteksi lebih cepat, sehingga upaya pencegahan dan penanganan dapat dilakukan secara lebih efektif.

Berbagai penelitian telah dilakukan untuk mengklasifikasikan kualitas air menggunakan metode pembelajaran mesin dengan tingkat akurasi yang bervariasi. Riyantoko dkk. (2023) menggunakan metode Lucifer Machine Learning Technique yang mencakup exploratory data analysis, skewness correction, dan model klasifikasi. Dari hasil eksperimen, model Random Forest Classifier mencapai akurasi tertinggi sebesar 72,81%[3], penelitian oleh Malagi

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2647−2658
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7055

(2023) yang membandingkan Random Forest dan Support Vector Machine (SVM) dalam memprediksi potabilitas air menunjukkan bahwa Random Forest mencapai akurasi 69,20%[4].

Elmeftahi dkk. (2024) membandingkan tujuh algoritma pembelajaran mesin dan menemukan bahwa Random Forest memiliki akurasi tertinggi sebesar 84,8%, diikuti oleh XGBoost (82,9%) dan CatBoost (80,2%)[5]. Sementara itu, Mukati dkk. (2024) menggunakan dataset Water Potability yang terdiri dari 3.276 entri dan menerapkan Random Forest, Decision Tree, serta XGBoost, dengan hasil bahwa Random Forest mencapai akurasi 70%[6]. Maulana dkk. (2024) mengimplementasikan algoritma XGBoost pada dataset dengan 2.400 sampel dan memperoleh akurasi sebesar 82,29%, menunjukkan keunggulan XGBoost dibandingkan algoritma lainnya[7].

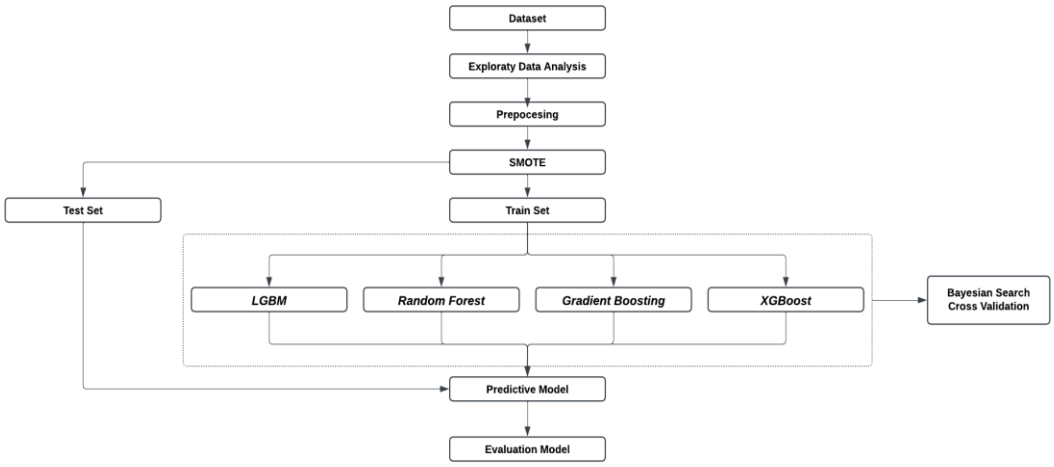
Penelitian-penelitian sebelumnya menunjukkan variasi dalam penggunaan algoritma pembelajaran mesin untuk klasifikasi kualitas air. Lucifer Machine Learning Technique yang digunakan oleh Riyantoko dkk. menerapkan pendekatan semi-supervised yang meningkatkan interpretabilitas model. Sementara itu, pendekatan supervised learning yang diterapkan oleh Malagi dan Elmeftahi dkk. membuktikan efektivitas Random Forest dalam menganalisis data kualitas air. Mukati dkk. dan Maulana dkk. berfokus pada perbandingan berbagai model dengan dataset yang berbeda, mengonfirmasi bahwa metode ensemble seperti XGBoost dan Random Forest mampu meningkatkan akurasi dalam prediksi potabilitas air.

Berdasarkan studi-studi tersebut, penelitian ini bertujuan untuk meningkatkan akurasi model klasifikasi kualitas air dengan menerapkan teknik preprocessing, seperti imputasi nilai hilang, normalisasi data, dan feature engineering. Selain itu, untuk mengatasi ketidakseimbangan data, digunakan teknik oversampling dengan Synthetic Minority Over-sampling Technique (SMOTE). Model yang digunakan dalam penelitian ini adalah LGBMClassifier, RandomForestClassifier, CatboostClasifier, dan XGBClassifier

Dari hasil tersebut, Random Forest terbukti menjadi model dengan performa terbaik. Dengan demikian, penelitian ini berkontribusi dalam pengembangan sistem klasifikasi kualitas air yang lebih efisien, mendukung pemantauan kualitas air secara real-time, serta membantu dalam perumusan kebijakan kesehatan masyarakat untuk memastikan ketersediaan air bersih.

2. METODOLOGI PENELITIAN

Metode penelitian ini menggunakan kerangka kerja empat tahap untuk menganalisis kelayakan air minum berdasarkan parameter kualitas air, seperti yang ditunjukkan pada Gambar 1. Pendekatan ini bertujuan untuk memahami faktor-faktor yang memengaruhi kualitas air serta mengembangkan model prediksi yang akurat. Tahapan penelitian mencakup pengumpulan data dari Kaggle, preprocessing melalui analisis eksplorasi data, normalisasi, dan penyeimbangan kelas dengan SMOTE, serta pelatihan model menggunakan algoritma seperti LGBM, Random Forest, Gradient Boosting, dan XGBoost. Model kemudian dioptimalkan dengan Bayesian Search Cross-Validation, dan dilakukan *stacking ensemble* dengan Random Forest sebagai *meta-learner* untuk meningkatkan performa klasifikasi. Evaluasi dilakukan berdasarkan metrik akurasi, presisi, recall, dan F1-score guna memastikan sistem dapat memprediksi kelayakan air minum secara lebih efisien dan akurat.



Gambar 1. Alur Penelitian

2.1 Data Collection

Pada tahap ini, dilakukan pengumpulan data yang akan digunakan dalam penelitian. Data diperoleh dari situs Kaggle melalui tautan https://www.kaggle.com/datasets/adityakadiwal/water-potability. Dataset tersedia dalam format CSV (*Comma Separated Values*) dan berisi informasi mengenai kualitas air berdasarkan berbagai parameter fisik dan kimia.

Dataset ini mencakup 3.276 entri dengan 9 fitur utama serta satu atribut label bernama *Potability*, yang menunjukkan apakah air tersebut layak minum (1) atau tidak layak minum (0). Data ini menjadi dasar dalam proses

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2647−2658
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7055

analisis dan pemodelan untuk menilai kualitas air berdasarkan parameter yang telah ditentukan. Sebelum digunakan dalam analisis lebih lanjut, dataset akan melalui tahap *preprocessing* guna memastikan kualitas data tetap optimal.

2.2 Prepocesing

Tahap *preprocessing* dilakukan untuk memastikan bahwa data yang digunakan dalam penelitian memiliki kualitas yang optimal sebelum dianalisis lebih lanjut. Proses ini mencakup beberapa langkah utama, yaitu *Exploratory Data Analysis* (EDA), penanganan nilai yang hilang (*missing values*), rekayasa fitur (*feature engineering*), penyeimbangan kelas data (*resampling*), Normalisasi dataset *(min max),*serta pembagian dataset (*data splitting*) dengan rasio 80:20[8][9][10].

2.3 Pengembangan Model

Penelitian ini mengembangkan model klasifikasi berbasis pembelajaran mesin untuk menilai kelayakan air minum berdasarkan parameter kualitas air. Proses pengembangan model melibatkan penerapan beberapa algoritma pembelajaran mesin, yaitu Light Gradient Boosting Machine (LGBM), Random Forest (RF), Gradient Boosting Machine (GBM), dan Extreme Gradient Boosting (XGBoost). Setiap algoritma memiliki keunggulan tersendiri dalam menangani data, terutama dalam klasifikasi berdasarkan parameter kualitas air. Setelah model dasar dibuat, dilakukan optimasi *hyperparameter* menggunakan Bayesian Optimization dengan Cross-Validation (CV) untuk meningkatkan performa model secara optimal.

2.3.1 Light Gradient Boosting Machine Learning(LGBM)

Light Gradient Boosting Machine (LGBM) merupakan algoritma *boosting* yang dikembangkan oleh Guolin Ke untuk meningkatkan efisiensi serta skalabilitas dalam pembelajaran mesin[11]. Algoritma ini memanfaatkan *Gradient-based One-Side Sampling* (GOSS) dan *Exclusive Feature Bundling* (EFB) guna mempercepat pemrosesan, terutama saat menangani dataset berukuran besar. Berbeda dengan GBM konvensional, LGBM menerapkan strategi pertumbuhan pohon secara *leaf-wise*, yang memungkinkan pohon berkembang lebih dalam pada cabang dengan potensi prediksi terbaik. Pendekatan ini meningkatkan akurasi tanpa mengorbankan efisiensi komputasi. Selain itu, LGBM memiliki keunggulan dalam menangani dataset dengan distribusi fitur yang tidak seimbang dan dapat beradaptasi dengan baik pada berbagai skala data, baik kecil maupun besar.

2.3.2 Random Forest (RF)

Random Forest (RF) adalah algoritma ensemble berbasis pohon keputusan yang dikembangkan oleh Breiman dengan menggunakan metode bagging untuk meningkatkan akurasi dan mengurangi overfitting[12]. Algoritma ini membangun sejumlah pohon keputusan dari subset data dan fitur yang dipilih secara acak, kemudian mengombinasikan prediksinya agar hasilnya lebih stabil. Dengan menerapkan teknik bootstrapping, RF mampu meningkatkan kemampuan generalisasi terhadap data baru. Keunggulan RF terletak pada kemampuannya menangani data berdimensi tinggi, ketahanannya terhadap noise, serta kebutuhan tuning hyperparameter yang lebih minimal dibandingkan metode boosting. Selain itu, RF juga menyediakan fitur **feature importance**, yang memungkinkan identifikasi faktor utama dalam klasifikasi kualitas air.

2.3.3 Cat Boost(CatBoost)

CatBoost (Categorical Boosting) adalah algoritma boosting yang dirancang untuk meningkatkan akurasi secara bertahap dengan membangun pohon keputusan yang memperbaiki kesalahan dari iterasi sebelumnya[13]. Berbeda dari teknik boosting lainnya, CatBoost dioptimalkan untuk menangani fitur kategorikal secara lebih efisien tanpa perlu encoding tambahan seperti one-hot encoding. Algoritma ini menggunakan skema pemrosesan berbasis permutasi untuk mengurangi bias dan meningkatkan generalisasi model. Selain itu, CatBoost menerapkan pendekatan khusus dalam penyesuaian bobot pada setiap iterasi, menjadikannya lebih tahan terhadap overfitting dibandingkan metode boosting lainnya. Meskipun demikian, optimasi hyperparameter tetap diperlukan untuk mencapai keseimbangan antara bias dan varians guna memastikan kinerja yang optimal.

2.3.4 Extreme Gradien Boosting(XGBoost)

Extreme Gradient Boosting (XGBoost) adalah varian dari *gradient boosting* yang dikembangkan oleh Chen dan Guestrin (2016) untuk meningkatkan efisiensi komputasi serta mengurangi *overfitting* melalui regularisasi L1 dan L2[14]. Algoritma ini memiliki keunggulan dalam menangani nilai yang hilang menggunakan *sparsity-aware split finding* serta mempercepat proses pencarian pemisahan terbaik melalui *histogram-based optimization*. Dengan kemampuannya dalam menangani data berdimensi tinggi serta fleksibilitas dalam penyetelan *hyperparameter*, XGBoost menjadi salah satu algoritma yang sering digunakan dalam kompetisi pembelajaran mesin.

2.3.5 Bayesian Optimazation dengan Cross-Validation (CV)

Setelah mengembangkan model, optimasi hyperparameter dilakukan dengan Bayesian Optimization dan Cross-Validation (CV)[15]. Bayesian Optimization menggunakan pendekatan probabilistik untuk mencari kombinasi hyperparameter terbaik secara efisien, berbeda dari grid search atau random search yang mencoba kombinasi secara

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2647−2658
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7055

eksplisit. Dalam penelitian ini, Bayesian Optimization dikombinasikan dengan CV untuk mencegah overfitting dan meningkatkan generalisasi model. CV membagi dataset menjadi beberapa subset (folds), di mana model dilatih dan diuji secara bergantian, lalu hasilnya dirata-ratakan untuk estimasi performa yang lebih akurat[16].

2.4 Metriks Performa

Setelah model klasifikasi dikembangkan, tahap selanjutnya adalah mengujinya pada data uji untuk mengevaluasi kinerjanya. Kinerja model klasifikasi mencerminkan seberapa baik model dalam mengidentifikasi dan mengelompokkan sampel dengan benar ke dalam kategori yang sesuai. Evaluasi performa model dalam penelitian ini dilakukan menggunakan empat metrik utama, yaitu Akurasi, Presisi, Recall, dan F1-score.Metrik-metrik evaluasi yang digunakan dijelaskan sebagai berikut:

a. Akurasi menunjukkan persentase keseluruhan sampel yang diklasifikasikan dengan benar, baik sebagai air yang layak minum maupun yang tidak layak minum. Semakin tinggi nilai akurasi, semakin baik kemampuan model dalam mengklasifikasikan data secara keseluruhan. Akurasi dapat dihitung menggunakan rumus berikut:

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

b. **Presisi** (*Precision*) mengukur seberapa andal model dalam memprediksi kategori positif (air layak minum)[17]. Metrik ini menunjukkan proporsi sampel yang diklasifikasikan sebagai layak minum yang benar-benar layak minum. Rumus presisi adalah:

$$Presisi = \frac{TP}{TP+FP} \quad (2)$$

c. **Recall** mengukur sejauh mana model dapat mendeteksi sampel yang benar-benar termasuk dalam kategori positif (air layak minum)[18]. Semakin tinggi recall, semakin sedikit sampel positif yang terlewatkan oleh model. Recall dirumuskan sebagai:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

d. **F1-score** merupakan rata-rata harmonik antara presisi dan recall. Metrik ini berguna untuk menyeimbangkan trade-off antara kesalahan False Positive dan False Negative, terutama dalam kondisi di mana distribusi kelas tidak seimbang[11]. F1-score dirumuskan sebagai:

$$F1-score = 2 \times \frac{Presisi \times Recall}{Presisi+Recall} \quad (4)$$

Dalam proses klasifikasi, hasil prediksi dapat dikategorikan ke dalam empat kemungkinan:

a. **True Positive (TP):** Sampel air yang benar-benar layak minum dan diklasifikasikan dengan benar sebagai layak minum.
b. **True Negative (TN):** Sampel air yang benar-benar tidak layak minum dan diklasifikasikan dengan benar sebagai tidak layak minum.
c. **False Positive (FP):** Sampel air yang sebenarnya layak minum tetapi salah diklasifikasikan sebagai tidak layak minum.
d. **False Negative (FN):** Sampel air yang sebenarnya tidak layak minum tetapi salah diklasifikasikan sebagai layak minum.

Penggunaan keempat metrik ini memberikan gambaran yang lebih komprehensif mengenai kinerja model. Selain akurasi, metrik presisi, recall, dan F1-score membantu dalam memahami sejauh mana model dapat menangani kesalahan prediksi, terutama ketika terdapat ketidakseimbangan antara jumlah sampel dalam kategori yang berbeda.

3. HASIL DAN PEMBAHASAN

3.1 Overview

Pada penelitian ini, dataset yang digunakan diperoleh dari website Kaggle dengan format CSV, berisi 3.276 baris data dan 10 atribut. Dataset ini digunakan untuk menganalisis kualitas air dan menentukan potabilitasnya. Atribut yang terdapat dalam dataset ini meliputi pH, Hardness (kesadahan), Solids (total padatan terlarut), Chloramines, Sulfate, Conductivity (konduktivitas), Organic_carbon (karbon organik), Trihalomethanes, Turbidity (kekeruhan), dan Potability (kelayakan air untuk dikonsumsi).

Tabel 1. Deskriptif Dataset

Variable	Min	Mean	Max
Ph	0.000000	7.080795	7.080795
Hardness	47.432000	196.369496	323.124000
Solids	320.942611	22014.092526	61227.196008
Choloramines	0.352000	7.122277	13.127000

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2647−2658
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7055

Sulfate	129.000000	333.775777	481.030642
Conductivty	181.483754	426.205111	753.342620
Organic Carbon	2.200000	14.284970	28.300000
Trihalomathenes	0.738000	66.396293	124.000000
Turbidit	1.450000	3.966786	6.739000
Potabolty	0.000000	0.390110	1.000000

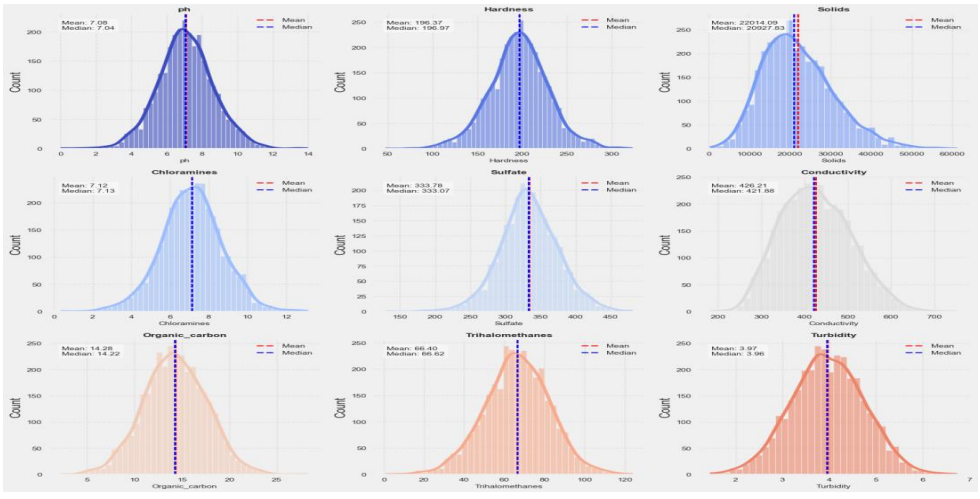
Pada Tabel 1 menunjukkan bahwa rata-rata pH dalam dataset ini adalah 7,08 dengan standar deviasi 1,59, sesuai dengan kisaran yang direkomendasikan WHO (6,5–8,5). Hardness memiliki rata-rata 196,37, dipengaruhi oleh kandungan garam kalsium dan magnesium. TDS rata-rata 22.014,09 menunjukkan variasi mineralisasi air. Chloramines, dengan rata-rata 7,12, berfungsi sebagai disinfektan, namun kadar yang aman adalah di bawah 4 mg/L. Sulfate rata-rata 333,77, dengan nilai maksimum 481,03. Konduktivitas listrik rata-rata 426,20, melebihi batas WHO (400 µS/cm) untuk beberapa sampel. TOC rata-rata 14,28, sedangkan WHO merekomendasikan di bawah 2 mg/L untuk air minum. THM memiliki rata-rata 66,39, dengan batas WHO 80 ppm. Kekeruhan rata-rata 3,96, sesuai dengan batas WHO di bawah 5 NTU. Hanya sekitar 39% dari sampel yang dapat diminum. Sebagian besar parameter kualitas air sesuai standar WHO, meskipun beberapa seperti konduktivitas dan THM melebihi batas yang direkomendasikan. Analisis lebih lanjut diperlukan untuk memahami faktor-faktor yang mempengaruhi potabilitas air.

3.2 Exploratory Data Analysis

Distribusi label dalam dataset menunjukkan bahwa 61% data tergolong dalam kelas 0, sedangkan 39% berada dalam kelas 1. Meskipun ketidakseimbangan ini tidak terlalu signifikan, tetap dapat memengaruhi kinerja model klasifikasi, karena model cenderung lebih sering memprediksi kelas yang lebih dominan. Oleh karena itu, diperlukan teknik smote menggunakan memastikan model tetap bekerja secara optimal.

3.2.1 Distribusi Dataset

Pada gambar 2 menunnjukan distribusi setiap fitur dalam dataset umumnya mendekati pola distribusi normal, dengan nilai rata-rata dan median yang hampir berimpit pada sebagian besar variabel. Nilai pH memiliki rata-rata 7,08 dan median 7,04, menunjukkan distribusi yang simetris tanpa skewness yang signifikan. Pola serupa terlihat pada Hardness, dengan rata-rata dan median sebesar 196,97, yang menandakan distribusi seimbang. Chloramines memiliki rata-rata 7,12 dan median 7,13, sementara Sulfate menunjukkan nilai rata-rata 333,78 dan median 333,07, keduanya mengindikasikan distribusi normal. Organic Carbon memiliki rata-rata 14,28 dan median 14,22, sedangkan Trihalomethanes dengan rata-rata 66,40 dan median 66,62 juga mencerminkan distribusi simetris tanpa pencilan yang mencolok. Selain itu, Turbidity, dengan rata-rata 3,97 dan median 3,96, menunjukkan pola distribusi yang hampir normal.



Gambar 2. Distribusi Dataset

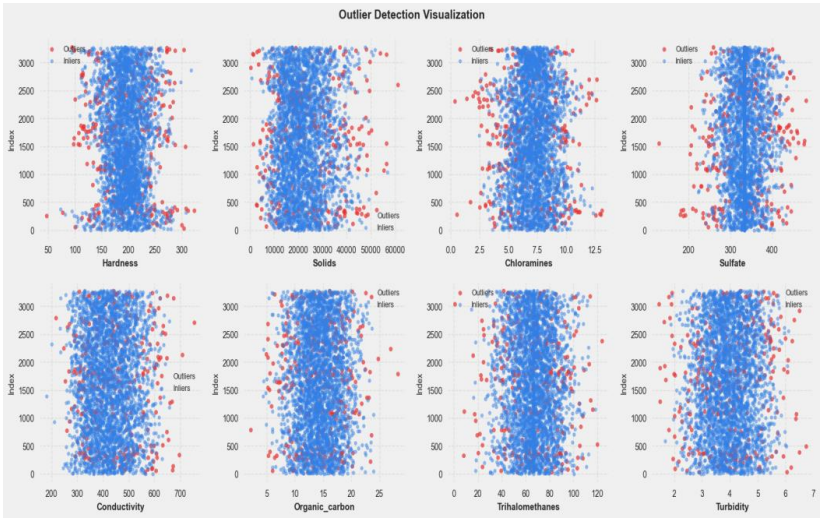
Namun, beberapa fitur dalam dataset menunjukkan sedikit skewness ke kanan, yang mengindikasikan adanya sejumlah nilai ekstrem yang lebih tinggi dibandingkan sebagian besar data. Sebagai contoh, fitur Solids memiliki rata-rata 22.014,09 dan median 20.927,26, dengan selisih yang cukup besar di antara keduanya, yang mengisyaratkan kemungkinan keberadaan outlier yang memengaruhi distribusi. Kondisi serupa juga terjadi pada Conductivity, dengan rata-rata 426,21 dan median 421,88, yang menunjukkan skewness ke kanan meskipun dalam tingkat yang lebih rendah. Perbedaan antara rata-rata dan median ini mengindikasikan bahwa beberapa sampel memiliki nilai yang jauh lebih besar dibandingkan nilai tipikal dalam dataset. Secara keseluruhan, distribusi dataset relatif seimbang, tetapi fitur dengan skewness yang lebih tinggi memerlukan analisis lebih lanjut untuk menentukan apakah nilai ekstrem

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2647−2658
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7055

tersebut perlu ditangani, misalnya melalui normalisasi atau metode penanganan outlier lainnya, sebelum digunakan dalam pemodelan.

3.2.2 Deteksi Outlier

Deteksi outlier penting dalam eksplorasi data untuk mengidentifikasi nilai ekstrem yang dapat mempengaruhi analisis. Berdasarkan scatterplot, beberapa fitur dalam dataset menunjukkan outlier yang tersebar di luar distribusi utama, mencerminkan variasi ekstrim dalam kualitas air akibat faktor alami atau proses pengolahan.



Gambar 3. Deteksi Outlier

Pada Gambar 3 menunnjukan bahwaFitur Hardness terkonsentrasi antara 150–250, dengan outlier di bawah 100 dan di atas 300, menunjukkan perbedaan komposisi mineral. Solids memiliki distribusi luas (20.000–40.000), dengan beberapa outlier di bawah 10.000 dan di atas 50.000, mengindikasikan variasi zat terlarut. Chloramines sebagian besar berkisar 4–10, tetapi terdapat outlier di bawah 2 dan di atas 12, menunjukkan fluktuasi kadar disinfektan.

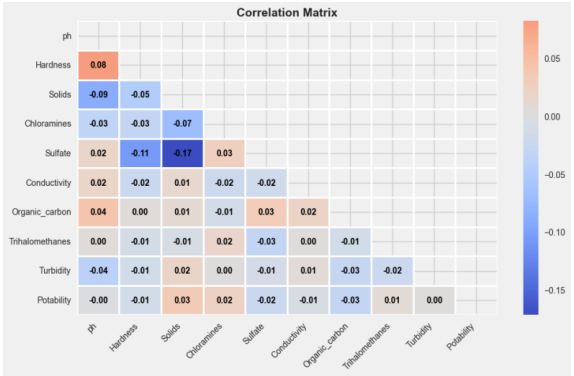
Sulfate umumnya berada dalam 250–400, dengan outlier di bawah 200 dan di atas 450, dipengaruhi oleh sumber air dan tingkat kontaminasi. Conductivity mayoritas 200–600, namun beberapa outlier melampaui rentang ini, menunjukkan variasi ion terlarut. Organic Carbon memiliki distribusi 5–20 dengan beberapa titik ekstrem, menunjukkan polutan organik.

Trihalomethanes berkisar 40–80, tetapi terdapat outlier di bawah 20 dan di atas 100, mencerminkan variasi akibat proses desinfeksi. Turbidity mayoritas 2–5, dengan beberapa outlier di atas 6, mengindikasikan fluktuasi partikel tersuspensi.

Secara keseluruhan, Solids, Sulfate, dan Trihalomethanes memiliki jumlah outlier yang cukup signifikan. Analisis lebih lanjut diperlukan untuk menentukan apakah outlier ini akan dihapus guna meningkatkan akurasi .

3.2.3 Analisis Korelasi Antar Parameter Kualitas Kualitas Air

Hubungan antar parameter kualitas air dianalisis menggunakan matriks korelasi Pearson. Koefisien korelasi Pearson digunakan untuk mengukur keterkaitan linear antara dua variabel, dengan rentang nilai antara -1 hingga 1. Nilai yang mendekati 1 menandakan korelasi positif yang kuat, sedangkan nilai yang mendekati -1 menunjukkan korelasi negatif yang kuat. Sementara itu, jika nilainya mendekati 0, maka tidak terdapat hubungan linear yang signifikan antara kedua variabel tersebut.



Gambar 4. Korelasi Matriks

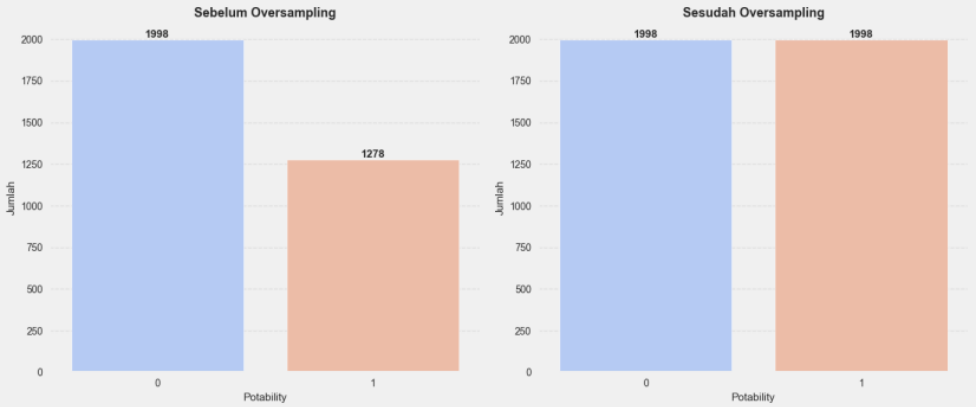
Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2647−2658
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7055

Pada Gambar 4 meliharkan bahwa beberapa korelasi yang lebih menonjol dalam dataset ini adalah antara Solids dan Sulfate (-0.17), menunjukkan bahwa peningkatan kadar sulfat dapat mengurangi jumlah padatan terlarut dalam air. Korelasi negatif juga ditemukan antara Chloramines dengan Sulfate (-0.07) dan Hardness (-0.03), namun pengaruhnya sangat kecil. Conductivity tidak memiliki hubungan yang kuat dengan parameter lainnya, dengan korelasi tertinggi hanya 0.03 terhadap Organic Carbon. Begitu pula dengan Trihalomethanes, Turbidity, dan Potability, yang semuanya memiliki korelasi yang sangat kecil dengan variabel lain. Potability sendiri menunjukkan korelasi tertinggi dengan Solids (0.03) dan Chloramines (0.02), tetapi nilai ini terlalu kecil untuk menunjukkan hubungan linear yang berarti.

Berdasarkan hasil analisis, tidak ditemukan hubungan linear yang kuat antara parameter kualitas air dalam dataset ini. Mayoritas nilai korelasi yang rendah menunjukkan bahwa perubahan dalam satu variabel tidak memiliki dampak besar terhadap variabel lainnya. Korelasi negatif tertinggi terjadi antara Sulfate dan Solids (-0.17), sementara korelasi positif tertinggi adalah antara pH dan Hardness (0.08). Dengan demikian, analisis regresi non-linear atau metode machine learning mungkin lebih efektif dalam memahami pola keterkaitan antar parameter dalam menentukan kualitas air minum. Selain itu, faktor eksternal seperti sumber air, metode pengolahan, serta kontaminan lain yang tidak termasuk dalam dataset juga perlu dipertimbangkan dalam evaluasi kualitas air secara menyeluruh.

3.3 Prepocesing

Dalam tahap preprocessing data sebelum pemodelan, berbagai teknik diterapkan untuk meningkatkan kualitas data dan memastikan model bekerja secara optimal. Salah satu tantangan utama yang dihadapi adalah ketidakseimbangan kelas pada variabel target *Potability*, di mana jumlah sampel air yang tidak layak konsumsi (kelas 0) mencapai 2.000 sampel, sementara sampel air yang layak konsumsi (kelas 1) hanya 1.250 sampel. Ketidakseimbangan ini berisiko membuat model lebih cenderung memprediksi kelas mayoritas. Untuk mengatasi hal tersebut, dilakukan oversampling menggunakan teknik SMOTE (*Synthetic Minority Over-sampling Technique*). Setelah penerapan SMOTE, jumlah sampel pada kedua kelas menjadi seimbang, masing-masing sebanyak 2.000 sampel, sehingga model dapat mempelajari karakteristik air layak dan tidak layak konsumsi secara lebih adil seperti yang ditunjjukan pada Gambar 5[20].



Gambar 5. Distribusi Label Sebelum dan Sesudah SMOTE

Selain menangani ketidakseimbangan data, preprocessing juga mencakup pemilihan dan rekayasa fitur untuk meningkatkan efektivitas model. Beberapa fitur yang kurang relevan, seperti "Turbidity" dan "Organic_carbon," dihapus, karena berdasarkan analisis awal, korelasinya terhadap potabilitas air sangat rendah. Selanjutnya, dilakukan penambahan dua fitur baru yang diyakini dapat memberikan informasi tambahan bagi model. Fitur pertama adalah "pH_Difference," yang dihitung dengan rumus |pH - 7.0|, yang menunjukkan seberapa jauh nilai pH dari kondisi ideal netral. Sebagai contoh, jika suatu sampel memiliki pH sebesar 8.2, maka nilai "pH_Difference"-nya adalah 1.2. Fitur kedua adalah "TDS_to_Hardness_Ratio," yang dihitung dengan rumus Solids / Hardness, di mana nilai ini memberikan informasi mengenai rasio total zat terlarut terhadap tingkat kekerasan air. Misalnya, jika suatu sampel memiliki Solids sebesar 20.000 mg/L dan Hardness sebesar 250 mg/L, maka nilai "TDS_to_Hardness_Ratio" adalah 80.

Terakhir, proses preprocessing juga mencakup penanganan nilai yang hilang (missing values) untuk memastikan data tidak mengandung informasi yang tidak lengkap. Beberapa fitur yang memiliki missing values adalah "Sulfate" dengan 490 nilai kosong, "pH" dengan 250 nilai kosong, dan "Trihalomethanes" dengan 350 nilai kosong. Untuk mengatasi masalah ini, dilakukan imputasi dengan menggantikan nilai yang hilang menggunakan mean (rata-rata) dari masing-masing fitur. Sebagai contoh, jika nilai rata-rata "Sulfate" dalam dataset adalah 333 mg/L, maka nilai yang hilang akan diisi dengan angka tersebut. Dengan metode ini, distribusi data tetap terjaga tanpa menghilangkan sampel, sehingga model tetap dapat belajar dari keseluruhan dataset tanpa bias akibat data yang tidak lengkap.Dan melakukan normalisai data menggunakan minmax untuk menangani data outlier[8].

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2647−2658
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7055

3.4 Model Developmet

Setelah proses oversampling, jumlah data untuk kedua kelas menjadi seimbang, yaitu 2.000 sampel untuk masing-masing kelas, sehingga model memiliki kesempatan yang lebih adil dalam mempelajari karakteristik air layak dan tidak layak konsumsi.

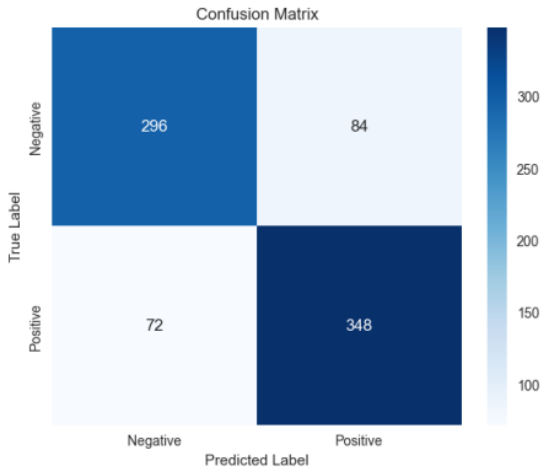
Tabel 2. Model Development

Classifier	Akurasi	Precision	Recall	F1-Score
LGBM	0.80625	0.8065	0.8062	0.8063
Random Forest	0.85	0.8585	0.85	0.8498
XGBoost	0.8275	0.8284	0.8275	0.8276
CatBoost	0.8	0.8019	0.8	0.8001

Pada Tabel 2 menunnjukkan analisis lebih lanjut terhadap Random Forest, yang mencapai akurasi tertinggi sebesar 85%, menunjukkan bahwa model ini memiliki performa terbaik dibandingkan model lainnya. Dengan precision 85.85% dan recall 85%, model ini mampu mengklasifikasikan kedua kelas dengan tingkat kesalahan yang lebih rendah. Sementara itu, XGBoost menunjukkan akurasi 82.75%, menjadikannya model dengan kinerja terbaik kedua setelah Random Forest, sedangkan LGBM dan CatBoost memiliki performa yang sedikit lebih rendah dengan akurasi 80.63% dan 80%.Meskipun LGBM memiliki akurasi di bawah Random Forest dan XGBoost, model ini tetap menjadi alternatif yang menarik karena efisiensi komputasinya yang lebih baik. Untuk meningkatkan performa lebih lanjut, langkah selanjutnya adalah hyperparameter tuning menggunakan Bayesian Search CV, yang diharapkan dapat mengoptimalkan parameter model guna memperoleh hasil yang lebih baik.

3.5 LGBM

Berdasarkan hasil optimasi hyperparameter menggunakan Bayesian Search CV, model LightGBM (LGBM) menunjukkan peningkatan performa dengan akurasi sebesar 80.5%. Model ini memiliki precision sebesar 80.49% dan recall sebesar 80.5%, menunjukkan keseimbangan yang baik dalam mengklasifikasikan air layak dan tidak layak minum. Meskipun demikian, masih terdapat beberapa kesalahan dalam prediksi, terutama pada kelas air layak minum.



Gambar 5. Confusion Matriks LGBM setelah Bayesian Search CV

Dari Confusion Matrix pada gambar 5, model berhasil mengklasifikasikan 296 sampel negatif dengan benar (True Negative) dan 348 sampel positif dengan benar (True Positive). Namun, masih terdapat 84 False Positive, yaitu sampel air tidak layak minum yang salah diklasifikasikan sebagai layak, serta 72 False Negative, di mana air layak minum salah diprediksi sebagai tidak layak. Kesalahan ini menunjukkan bahwa meskipun model cukup akurat, masih ada ruang untuk perbaikan, terutama dalam meningkatkan recall untuk kelas positif.

Tabel 6. Hasil LGBM dengan Bayesian Search CV

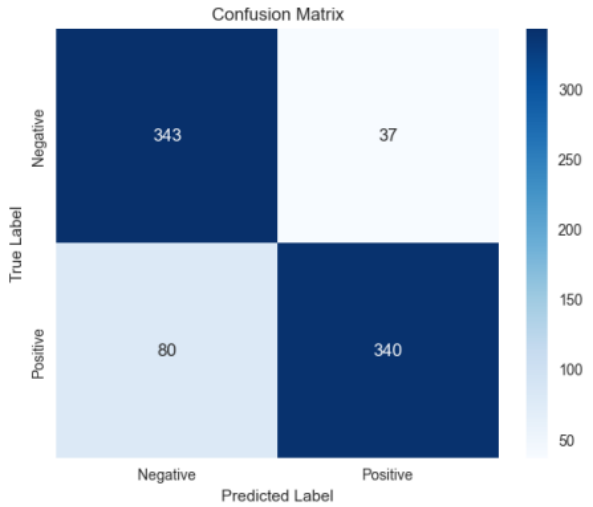
Algoritma Klasifikasi	LGBM
Akurasi	0.805
Presisi	0.8049
Recall	0.805
F1-Score	0.804
Parameter	OrderedDict([('colsample_bytree', 1.0), ('lambda_l1', 2.4764233822962796), ('lambda_l2', 7.5993974483241855), ('learning_rate', 0.2024976829302216), ('max_depth', 10), ('min_child_weight', 1), ('n_estimators', 284), ('subsample', 0.776717068436972)])

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2647−2658
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7055

Seperti Tabel 6 untuk meningkatkan performa lebih lanjut, beberapa hyperparameter optimal diterapkan, termasuk colsample_bytree sebesar 1.0, lambda_l1 sebesar 2.476, dan lambda_l2 sebesar 7.599, yang membantu mengurangi overfitting. Selain itu, learning_rate sebesar 0.202 memungkinkan model untuk belajar secara bertahap dengan keseimbangan yang baik antara bias dan varians. Dengan n_estimators sebanyak 284 dan max_depth sebesar 10, model mampu menangkap pola yang lebih kompleks dalam data tanpa kehilangan efisiensi[11].

3.6 Random Forest

Setelah dilakukan tuning hyperparameter menggunakan Bayesian Search CV, performa model Random Forest menunjukkan peningkatan yang signifikan. Model yang telah dioptimalkan mencapai akurasi sebesar 85.38%, lebih tinggi dibandingkan versi sebelumnya. Precision model mencapai 85.86%, sedangkan recall berada pada angka 85.38%, menghasilkan f1-score sebesar 85.37%. Hasil ini menunjukkan bahwa model dapat mengklasifikasikan sampel dengan keseimbangan yang cukup baik antara presisi dan sensitivitas.



Gambar 6. Confusion Matriks Random Forest setelah Bayesian Search CV

Berdasarkan Confusion Matrix pada Gambar 6, model berhasil mengklasifikasikan 343 sampel negatif dengan benar (True Negative) dan 340 sampel positif dengan benar (True Positive). Namun, masih terdapat 37 False Positive, yaitu sampel air yang tidak layak konsumsi tetapi diklasifikasikan sebagai layak, serta 80 False Negative, di mana sampel air layak minum salah dikategorikan sebagai tidak layak. Dengan recall kelas 0 (tidak layak konsumsi) sebesar 90.27%, model mampu mengidentifikasi sebagian besar air yang tidak layak minum dengan baik. Sebaliknya, recall kelas 1 (layak konsumsi) sebesar 80.95% menunjukkan bahwa masih ada beberapa kesalahan dalam mendeteksi air yang seharusnya layak dikonsumsi[20].

Tabel 6. Hasil Random Forest dengan Bayesian Search CV

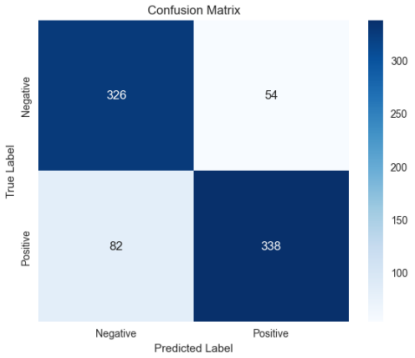
Algoritma Klasifikasi	Random Forest
Akurasi	0.85375
Presisi	0.85864
Recall	0.85375
F1-Score	0.85372
Parameter	OrderedDict([('max_depth', 20), ('max_features', 0.5787121462720156), ('min_samples_leaf', 1), ('min_samples_split', 2), ('n_estimators', 1000)])

Untuk meningkatkan performa model, dilakukan tuning terhadap beberapa hyperparameter utama, seperti yang terdapat dalam Tabel 6 yaitu max_depth sebesar 20, max_features sebesar 0.58, min_samples_leaf sebesar 1, min_samples_split sebesar 2, dan n_estimators sebanyak 1000. Nilai max_depth 20 memungkinkan model menangkap pola yang lebih kompleks tanpa overfitting, sementara n_estimators 1000 memastikan hasil prediksi yang lebih stabil. Dengan kombinasi parameter ini, Random Forest mampu memberikan prediksi yang lebih akurat dan optimal dalam menentukan kelayakan air minum.

3.7 XGBoost

Setelah dilakukan optimasi menggunakan Bayesian Search CV, model XGBoost menunjukkan peningkatan performa yang cukup signifikan. Model ini mencapai akurasi sebesar 83%, yang lebih stabil dibandingkan sebelum dilakukan tuning. Precision model berada pada 83.22%, sementara recall mencapai 83%, menghasilkan f1-score sebesar 83%. Hasil ini menunjukkan bahwa model dapat mengenali pola dalam data dengan cukup baik.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2647−2658
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7055



Gambar 7. Confusion Matriks XGBoost setelah Bayesian Search CV

Berdasarkan Confusion Matrix pada gambar 7, model berhasil mengklasifikasikan 326 sampel negatif dengan benar (True Negative) dan 338 sampel positif dengan benar (True Positive). Namun, masih terdapat 54 False Positive, yaitu sampel air yang tidak layak konsumsi tetapi diklasifikasikan sebagai layak, serta 82 False Negative, di mana sampel air layak konsumsi salah dikategorikan sebagai tidak layak. Recall untuk kelas 0 (tidak layak konsumsi) sebesar 85.8%, menunjukkan bahwa model cukup baik dalam mendeteksi air yang tidak layak minum. Sementara itu, recall untuk kelas 1 (layak konsumsi) sebesar 80.5%, yang menandakan masih ada beberapa kesalahan dalam mendeteksi air yang seharusnya layak konsumsi.

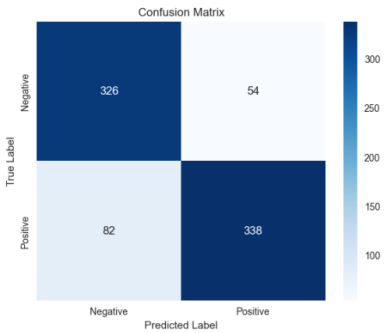
Tabel 5. Hasil XGBoost dengan Bayesian Search CV

Algoritma Klasifikasi	XGBoost
Akurasi	0.83
Presisi	0.83221
Recall	0.83
F1-Score	0.83
Parameter	OrderedDict([('colsample_bytree', 0.5929667531263008), ('gamma', 0.1), ('learning_rate', 0.01), ('max_depth', 10), ('min_child_weight', 1), ('n_estimators', 1000), ('subsample', 1.0)])

Untuk meningkatkan performa model, dilakukan tuning terhadap beberapa hyperparameter utama. Model menggunakan colsample_bytree sebesar 0.59, sehingga model hanya mempertimbangkan sekitar 59% fitur dalam setiap pohon keputusan. Nilai gamma sebesar 0.1 memberikan batasan pada pemisahan cabang untuk mencegah overfitting. Learning rate sebesar 0.01 membuat model belajar secara bertahap untuk meningkatkan stabilitas. Selain itu, max_depth 10 memungkinkan model menangkap pola yang lebih kompleks dalam data, sementara min_child_weight sebesar 1 memastikan bahwa setiap pemisahan dilakukan dengan data yang cukup. Model menggunakan 1000 estimators, yang membantu meningkatkan kinerja model secara keseluruhan. Dengan subsample sebesar 1.0, model memanfaatkan seluruh data dalam setiap iterasi untuk meningkatkan akurasi. Dengan kombinasi hyperparameter ini, XGBoost mampu memberikan keseimbangan yang baik antara bias dan varians, menjadikannya lebih andal dalam menentukan kelayakan air minum.

3.8 CatBoost

Setelah dilakukan optimasi hyperparameter menggunakan Bayesian Search CV, model CatBoost menunjukkan peningkatan performa yang cukup signifikan dalam mengklasifikasikan kelayakan air. Model ini mencapai akurasi sebesar 82.13%, yang lebih stabil dibandingkan sebelum dilakukan tuning. Precision model berada pada 82.18%, sementara recall mencapai 82.13%, menghasilkan f1-score sebesar 82.13%. Hasil ini menunjukkan bahwa model memiliki keseimbangan yang baik dalam mendeteksi sampel yang layak dan tidak layak konsumsi.



Gambar 7. Confusion Matriks Gradient Boosting setelah Bayesian Search CV

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2647−2658
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7055

Berdasarkan Confusion Matrix pada Gambar 7, model berhasil mengklasifikasikan 337 sampel negatif dengan benar (True Negative) dan 320 sampel positif dengan benar (True Positive). Namun, masih terdapat 78 False Positive, yaitu sampel air yang tidak layak konsumsi tetapi diklasifikasikan sebagai layak, serta 65 False Negative, di mana sampel air layak konsumsi salah dikategorikan sebagai tidak layak. Recall untuk kelas 0 (tidak layak konsumsi) sebesar 81.2%, menunjukkan bahwa model cukup baik dalam mendeteksi air yang tidak layak minum. Sementara itu, recall untuk kelas 1 (layak konsumsi) sebesar 83.1%, yang menandakan bahwa model masih mengalami sedikit kesalahan dalam mengklasifikasikan air layak konsumsi.

Tabel 6. Hasil Gradien Boosting dengan Bayesian Search CV

Algoritma Klasifikasi	Cat Boost
Akurasi	0.82125
Presisi	0.821807
Recall	0.82125
F1-Score	0.82131
Parameter	OrderedDict([('depth', 10), ('iterations', 574), ('l2_leaf_reg', 2.0537498490775676), ('learning_rate', 0.012417471789894306)])

Berdasarkan Tabel 6 untuk meningkatkan performa model, dilakukan tuning terhadap beberapa hyperparameter utama. Parameter depth ditetapkan sebesar 10, yang memberikan keseimbangan antara kompleksitas model dan risiko overfitting. Model menggunakan 574 estimators, yang berarti model menggabungkan total 574 pohon keputusan dalam proses boosting untuk meningkatkan akurasi prediksi. Parameter l2_leaf_reg sebesar 2.05 membantu mengontrol regularisasi dan mencegah overfitting, sedangkan learning rate sebesar 0.0124 memastikan bahwa pembaruan bobot dilakukan secara bertahap untuk meningkatkan stabilitas model. Dengan kombinasi hyperparameter ini, model CatBoost menunjukkan performa yang lebih optimal dan akurat dalam menentukan kelayakan air minum, menjadikannya model yang andal dalam sistem klasifikasi potabilitas air.

3.9 Komperasi Hasil Cross Validation

Cross-validation digunakan untuk mengevaluasi performa model secara lebih objektif dengan membagi dataset menjadi beberapa subset, sehingga setiap model diuji pada bagian data yang berbeda. Teknik ini membantu mengurangi bias akibat pemisahan dataset yang tidak merata dan memberikan gambaran lebih akurat tentang generalisasi model.Berdasarkan hasil cross-validation,Seperti yang ditunjukkan tabel 7 Random Forest memiliki performa terbaik dengan akurasi tertinggi sebesar 85.38% setelah tuning hyperparameter menggunakan Bayesian Search CV. Model ini juga memiliki precision tertinggi, yaitu 85.86%, serta recall sebesar 85.38%, menunjukkan keseimbangan yang baik dalam mengklasifikasikan kelas mayoritas dan minoritas.Sementara itu, XGBoost menunjukkan performa yang kompetitif dengan akurasi sebesar 83.0%, diikuti oleh CatBoost dengan 82.13% dan LightGBM dengan 80.5%.

Tabel 7. Cross Validasi

Classifier	Akurasi	Precision	Recall	F1-Score
LGBM	0.805	0.8049	0.805	0.8048
Random Forest	0.85375	0.85864	0.85375	0.85372
XGBoost	0.83	0.83221	0.83	0.83
CatBoost	0.82125	0.82180	0.82125	0.8213

Dibandingkan model lainnya, XGBoost menunjukkan keseimbangan yang baik antara precision dan recall, menjadikannya alternatif yang layak dalam mengklasifikasikan sampel dengan benar. CatBoost juga memiliki performa yang cukup stabil dengan precision sebesar 82.18% dan recall 82.13%.Di sisi lain, LightGBM memiliki performa paling rendah dengan akurasi 80.5%, meskipun precision dan recall yang dihasilkan tetap cukup kompetitif, masing-masing sebesar 80.49% dan 80.5%.Secara keseluruhan, Random Forest menjadi pilihan terbaik untuk klasifikasi potabilitas air berdasarkan hasil cross-validation. XGBoost dan CatBoost dapat menjadi alternatif yang layak dengan performa yang kompetitif, sementara LightGBM dapat dipilih jika efisiensi komputasi menjadi pertimbangan utama. Pemilihan model yang optimal bergantung pada kebutuhan spesifik dalam implementasi sistem klasifikasi air bersih.

4. KESIMPULAN

Penelitian ini mengusulkan sistem klasifikasi berbasis pembelajaran mesin untuk menentukan kelayakan air minum secara lebih akurat dan efisien dibandingkan metode manual seperti Water Quality Index (WQI) dan STORET. Dengan menggunakan dataset *Water Potability* dari Kaggle, penelitian ini menerapkan berbagai teknik *preprocessing*, termasuk imputasi data yang hilang, normalisasi, rekayasa fitur, serta oversampling dengan SMOTE untuk mengatasi ketidakseimbangan kelas. Hasil analisis awal menunjukkan bahwa sebagian besar parameter kualitas air tidak memiliki hubungan linear yang signifikan, sehingga pendekatan berbasis pembelajaran mesin menjadi solusi yang

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2647−2658
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7055

lebih adaptif dalam klasifikasi kelayakan air. Hasil eksperimen menunjukkan bahwa model *ensemble learning*, khususnya Random Forest, memberikan performa terbaik dengan akurasi 85,38%, precision 85,86%, recall 85,38%, dan F1-score 85,37%. Optimasi lebih lanjut menggunakan Bayesian Search CV meningkatkan performa model, terutama pada Random Forest dan XGBoost. Meskipun demikian, masih terdapat beberapa kesalahan klasifikasi, terutama dalam mendeteksi sampel air layak konsumsi. Oleh karena itu, penelitian lanjutan dapat difokuskan pada peningkatan akurasi melalui teknik *feature selection*, *deep learning*, atau kombinasi model berbasis *hybrid learning*. Kesimpulannya, metode pembelajaran mesin, khususnya *ensemble learning*, dapat menjadi alternatif yang efektif dalam menilai kelayakan air minum secara lebih cepat dan akurat.

REFERENCES

[1] C. Allen, G. Metternicht, and T. Wiedmann, “Initial progress in implementing the Sustainable Development Goals (SDGs): a review of evidence from countries,” *Sustain. Sci.*, vol. 13, no. 5, pp. 1453–1467, 2018, doi: 10.1007/s11625-018-0572-3.

[2] S. Tyagi, B. Sharma, P. Singh, and R. Dobhal, “Water Quality Assessment in Terms of Water Quality Index,” *Am. J. Water Resour.*, vol. 1, no. 3, pp. 34–38, 2020, doi: 10.12691/ajwr-1-3-3.

[3] 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.

[4] N. Malagi, “Water Potability Prediction using Machine Learning,” *Int. Res. J. Mod. Eng. Technol. Sci.*, no. 08, pp. 2779–2782, 2023, doi: 10.56726/irjmets44413.

[5] C. N. Ihsan *et al.*, “Comparison of Machine Learning Algorithms in Detecting Tea Leaf Diseases,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 8, no. 1, pp. 135–141, 2024, doi: 10.29207/resti.v8i1.5587.

[6] L. Díaz-González, R.A. Aguilar-Rodríguez, J.C. Pérez-Sansalvador, N. Lakouari, “AQuA-P: A machine learning-based tool for water quality assessment”, J Contam Hydrol, vol 269, no 104498, 2025, doi:10.1016/j.jconhyd.2025.104498

[7] Malik, N., Kalonia, A., Dalal, S. *et al.* "Optimized XGBoost Hyper-Parameter Tuned Model with Krill Herd Algorithm (KHA) for Accurate Drinking Water Quality Prediction," *SN COMPUT. SCI.* **6**, 263, 2025, https://doi.org/10.1007/s42979-025-03813-9

[8] Y. Cai and C. Daskalakis, “On minmax theorems for multiplayer games,” *Proc. Annu. ACM-SIAM Symp. Discret. Algorithms*, pp. 217–234, 2011, doi: 10.1137/1.9781611973082.20.

[9] R. R. R. Arisandi, B. Warsito, and A. R. Hakim, “Aplikasi Naïve Bayes Classifier (Nbc) Pada Klasifikasi Status Gizi Balita Stunting Dengan Pengujian K-Fold Cross Validation,” *J. Gaussian*, vol. 11, no. 1, pp. 130–139, 2022, doi: 10.14710/j.gauss.v11i1.33991.

[10] Zichong Wang, Zhipeng Yin, Yuying Zhang, Liping Yang, Tingting Zhang, Niki Pissinou, Yu Cai, Shu Hu, Yun Li, Liang Zhao, and Wenbin Zhang, " FG-SMOTE: Towards Fair Node Classification with Graph Neural Network," SIGKDD Explor. Newsl. 26, 2 (December 2024), 99–108. https://doi.org/10.1145/3715073.3715082

[11] H. Los *et al.*, “Evaluation of Xgboost and Lgbm Performance in Tree Species Classification With Sentinel-2 Data,” *Int. Geosci. Remote Sens. Symp.*, vol. 2021-July, pp. 5803–5806, 2021, doi: 10.1109/IGARSS47720.2021.9553031.

[12] J. Hu and S. Szymczak, “A review on longitudinal data analysis with random forest,” *Brief. Bioinform.*, vol. 24, no. 2, pp. 1–11, 2023, doi: 10.1093/bib/bbad002.

[13] Wowon Priatna, "Dampak Pengambilan Sampel Data untuk Optimalisasi Data tidak seimbang pada Klasifikasi Penipuan Transaksi E-Commerce " The Indonesian Journal of Computer Science ,Vol. 13, No.2, 2024, doi:10.33022/ijcs.v13i2.2698.

[14] Xiaowei Li and Lanxin Shi and Yang Shi and Junqing Tang and Pengjun Zhao and Yuting Wang and Jun Chen," Exploring interactive and nonlinear effects of key factors on intercity travel mode choice using XGBoost " Applied Geography,Vol. 166, No. 103264, Doi:10.1016/ j.apgeog.2024.103264

[15] F. Aziz, P. Ishak, and S. Abasa, “Klasifikasi Depresi Menggunakan Support Vector Machine: Pendekatan Berbasis Data Text Mining,” *J. Pharm. Appl. Comput. Sci.*, vol. 2, no. 2, pp. 33–38, 2024, doi: 10.59823/jopacs.v2i2.53.

[16] L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, “Cross validation for model selection: A review with examples from ecology,” *Ecol. Monogr.*, vol. 93, no. 1, pp. 1–24, 2023, doi: 10.1002/ecm.1557.

[17] M. I. K. Saraan and R. F. A. K. Rambe, “Kebijakan Pengembangan Inovasi Teknologi Pertanian Presisi di Provinsi Sumatera Utara,” *J. Kaji. Agrar. dan Kedaulatan Pangan*, vol. 2, no. 1, pp. 1–5, 2023, doi: 10.32734/jkakp.v2i1.13319.

[18] M. E. Lestari, I. Asror, and I. L. Sardi, “Penerapan PCA (Principal Component Analysis) pada Deteksi Outlier untuk Data Text,” *e-Proceeding Eng.*, vol. 10, no. 3, p. 3549, 2023.

[19] Pedro Lucas Negromonte Guerra, Inaê Carolline Silveira da Silva, Deoclides Lima Bezerra Júnior, Anderson Albert Primo Lopes, Geraldo de Sá Carneiro Filho, Eduardo Vieira de Carvalho Júnior,"Epidemiological and clinical characteristics of primary spinal cord glioblastomas ", Journal of Clinical Neuroscience,Vol. 130, No. 110862, doi : 10.1016/j.jocn.2024.110862

[20] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, “AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes,” *J. Supercomput.*, vol. 77, no. 5, pp. 5198–5219, 2021, doi: 10.1007/s11227-020-03481-x.