cBuilding of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2546−2558
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7014

Pendekatan Machine Learning dengan Teknik Stacking untuk Memprediksi Kualitas Air Minum

Ishak Bintang D[1,*], Pulung Nurtantio Andono[1], Ricardus Anggi Pramunendar[1] , Agus Winarno[2], Aditya Aqil Darmawan[1]

[1] Fakultas Ilmu Komputer, Teknik Informatika,Universitas Dian Nuswantoro, Semarang, Indonesia
[2] Fakultas Ilmu Komputer, Sistem Informasi,Universitas Dian Nuswantoro, Semarang, Indonesia
Email: [1,*]111202113763@mhs.dinus.ac.id, [2]pulung@dsn.dinus.ac.id, [3]ricardus.anggi@dsn.dinus.ac.id,
[4]agus.winarno@dsn.dinus.ac.id , [5]111202113390@mhs.dinus
Email Penulis Korespondensi: 111202113763@mhs.dinus.ac.id
Submitted: **19/02/2025**; Accepted: **10/03/2025**; Published**: 13/03/2025**

Abstrak−Kualitas air minum yang aman penting bagi kesehatan, namun pencemaran lingkungan menurunkan kualitasnya secara signifikan. Metode manual seperti WQI dan STORET kurang efisien, sehingga penelitian ini mengusulkan sistem klasifikasi berbasis pembelajaran mesin untuk menentukan kelayakan air minum dengan lebih akurat. Dataset *Water Potability* dari Kaggle digunakan dengan 3.276 sampel dan sembilan parameter utama. Tahap preprocessing mencakup imputasi data, normalisasi, *feature engineering*, serta oversampling dengan SMOTE. Model yang diterapkan meliputi LGBM, Random Forest, GBM, dan XGBoost, dengan optimasi *Bayesian* dan *stacking ensemble* untuk meningkatkan akurasi. Hasil menunjukkan *stacking ensemble* mencapai akurasi 85,38%, precision 88,02%, recall 85,38%, dan F1-score 85,23%, lebih unggul dibandingkan model individu. Sistem ini memungkinkan pemantauan kualitas air secara real-time dengan hasil lebih cepat dan akurat, mendukung pengambilan keputusan terkait kebijakan sanitasi dan ketersediaan air bersih.

Kata Kunci: Kualitas Air; Pembelajaran Mesin; Klasifikasi; Stacking Ensemble; Water Potability.

Abstract-Safe drinking water quality is essential for public health, yet environmental pollution has significantly degraded its quality. Manual methods such as WQI and STORET are inefficient, prompting this study to propose a machine learning-based classification system for more accurate water potability assessment. The *Water Potability* dataset from Kaggle is used, consisting of 3,276 samples with nine key parameters. The preprocessing stage includes data imputation, normalization, feature engineering, and oversampling with SMOTE. The applied models include LGBM, Random Forest, GBM, and XGBoost, optimized using Bayesian techniques and stacking ensemble to enhance accuracy. Results show that the stacking ensemble achieves an accuracy of 85.38%, precision of 88.02%, recall of 85.38%, and F1-score of 85.23%, outperforming individual models. This system enables real-time water quality monitoring with faster and more accurate results, supporting decision-making in sanitation policies and clean water availability.

Keywords: Water Quality; Machine Learning; Classification; Stacking Ensemble; Water Potability

1. PENDAHULUAN

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

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

Berbagai penelitian telah dilakukan untuk klasifikasi 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]. Malagi (2023) meneliti prediksi potabilitas air dengan machine learning menggunakan Random Forest dan Support Vector Machine (SVM), dengan hasil menunjukkan bahwa Random Forest memiliki 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 dengan 3.276 entri dan menerapkan model Random Forest, Decision Tree, serta XGBoost, dengan hasil bahwa Random Forest mencapai akurasi 70% [6]. Maulana dkk. (2024) mengimplementasikan algoritma XGBoost dengan dataset 2.400 sampel dan memperoleh akurasi 82,29%, menunjukkan bahwa XGBoost unggul dibandingkan beberapa algoritma lain [7].

Penelitian-penelitian sebelumnya menunjukkan variasi dalam penggunaan algoritma pembelajaran mesin untuk klasifikasi kualitas air. Lucifer Machine Learning Technique yang digunakan oleh Riyantoko dkk.

Copyright © 2025 **Author**, Page 2546
This Journal is licensed under a Creative Commons Attribution 4.0 International License

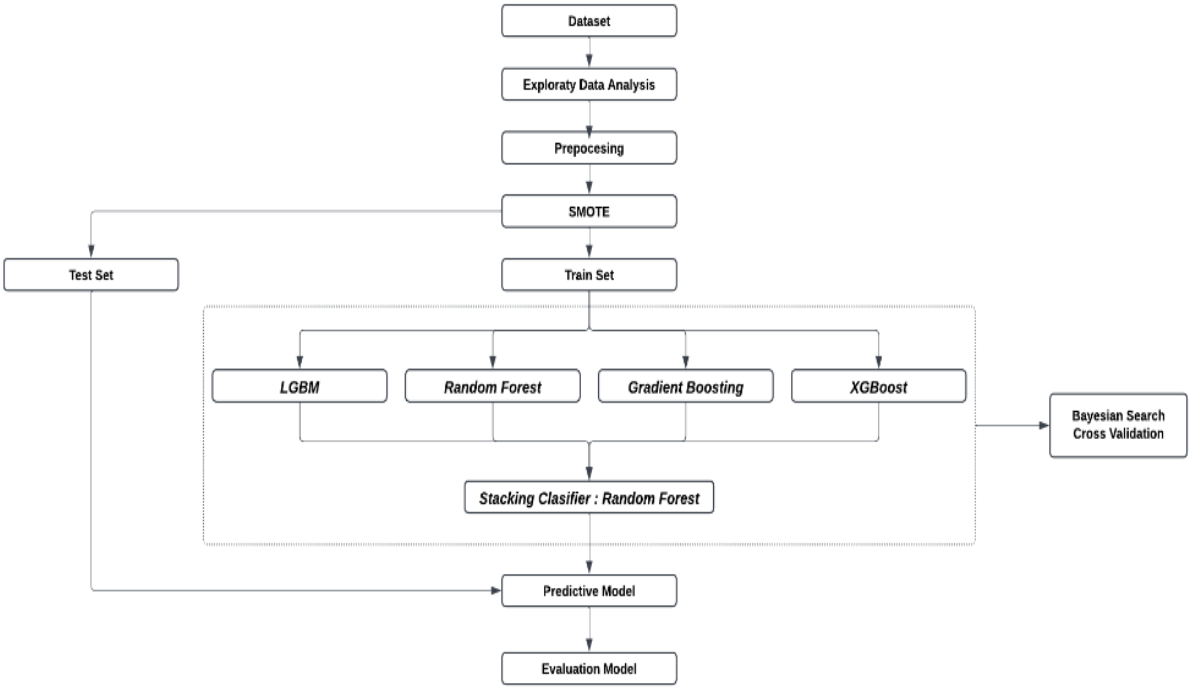
cBuilding of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2546−2558
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7014

menunjukkan pendekatan semi-supervised yang dapat meningkatkan interpretabilitas model, sementara pendekatan supervised learning yang digunakan oleh Malagi dan Elmeftahi dkk. membuktikan efektivitas Random Forest dalam analisis data air. Mukati dkk. dan Maulana dkk. berfokus pada perbandingan berbagai model dengan dataset berbeda, yang mengonfirmasi keunggulan metode ensemble seperti XGBoost dan Random Forest dalam meningkatkan akurasi 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, teknik oversampling dengan Synthetic Minority Over-sampling Technique (SMOTE) diterapkan untuk mengatasi ketidakseimbangan data[8]. Model yang digunakan dalam penelitian ini adalah kombinasi metode ensemble stacking yang menggabungkan LGBMClassifier, RandomForestClassifier, GradientBoostingClassifier, dan XGBClassifier, dengan meta model RandomForestClassifier. Hasil eksperimen menunjukkan peningkatan performa dengan akurasi 85,38%, precision 88,02%, recall 85,38%, dan F1-score 85,23%. Dengan demikian, penelitian ini berkontribusi dalam pengembangan sistem klasifikasi kualitas air yang lebih efisien untuk pemantauan kualitas air secara real-time dan mendukung kebijakan kesehatan masyarakat dalam 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

Berisi Pada tahap ini, dilakukan proses perolehan data yang akan digunakan dalam penelitian. Data yang digunakan ini diperoleh dari situs Kaggle dengan tautan https://www.kaggle.com/datasets/adityakadiwal/water-potability. Dataset tersebut memiliki format CSV (Comma Separated Values) dan berisi informasi mengenai kualitas air berdasarkan beberapa parameter fisik dan kimia.

Dataset ini terdiri dari 3.276 baris data dan memiliki 9 fitur atau parameter, serta 1 atribut label bernama Potability yang menunjukkan apakah air tersebut layak minum (1) atau tidak layak minum (0)

Copyright © 2025 **Author**, Page 2547
This Journal is licensed under a Creative Commons Attribution 4.0 International License

cBuilding of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2546−2558
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7014

Dataset ini digunakan sebagai dasar dalam proses analisis dan pemodelan untuk menentukan kualitas air berdasarkan parameter-parameter yang telah disebutkan. Data yang diperoleh akan melalui tahap preprocessing untuk memastikan kualitas data sebelum digunakan dalam analisis lebih lanjut.

2.2 Prepocesing

Tahap preprocessing dilakukan untuk memastikan bahwa data yang digunakan dalam penelitian memiliki kualitas yang baik sebelum digunakan dalam analisis lebih lanjut. Proses preprocessing yang dilakukan dalam penelitian ini terdiri dari beberapa langkah utama, yaitu Exploratory Data Analysis (EDA), penanganan data hilang (missing values), rekayasa fitur (feature engineering), penyeimbangan kelas data (resampling), dan pembagian dataset (data splitting) dengan komposisi 80:20[9].

2.3 Pengembangan Model

Dalam penelitian ini, dikembangkan suatu model klasifikasi berbasis pembelajaran mesin untuk menentukan kelayakan air minum berdasarkan parameter kualitas air. Pengembangan model dilakukan dengan menerapkan berbagai algoritma pembelajaran mesin, yaitu Light Gradient Boosting Machine (LGBM), Random Forest (RF), Gradient Boosting Machine (GBM), dan Extreme Gradient Boosting (XGBoost). Setiap algoritma memiliki karakteristik dan keunggulan masing-masing dalam menangani data, khususnya dalam konteks klasifikasi berbasis parameter kualitas air. Setelah model dasar dikembangkan, dilakukan optimasi hyperparameter menggunakan Bayesian Optimization dengan Cross-Validation (CV) guna meningkatkan performa model secara signifikan. Langkah terakhir dalam pengembangan model adalah penerapan teknik stacking, yang bertujuan untuk menggabungkan keunggulan dari berbagai model guna meningkatkan akurasi dan generalisasi model terhadap data baru.

2.3.1 Light Gradient Boosting Machine Learning(LGBM)

Light Gradient Boosting Machine (LGBM) adalah algoritma boosting yang dikembangkan oleh Guolin Ke untuk meningkatkan efisiensi dan skalabilitas dalam pembelajaran mesin[10]. LGBM menggunakan Gradient-based One-Side Sampling (GOSS) dan Exclusive Feature Bundling (EFB) untuk mempercepat pemrosesan, terutama pada dataset besar. Berbeda dari GBM konvensional, LGBM menerapkan pertumbuhan pohon leaf-wise, yang memungkinkan pembangunan pohon lebih dalam pada cabang dengan potensi perbaikan prediksi terbesar. Strategi ini meningkatkan akurasi tanpa mengorbankan efisiensi komputasi. Selain itu, LGBM unggul dalam menangani dataset dengan distribusi fitur tidak seimbang dan beradaptasi dengan baik untuk skala kecil maupun besar.

2.3.2 Random Forest (RF)

Random Forest (RF) adalah algoritma ensemble berbasis pohon keputusan yang dikembangkan oleh Breiman dengan pendekatan bagging untuk meningkatkan akurasi dan mengurangi overfitting[11]. RF membangun banyak pohon keputusan dari subset data dan fitur yang dipilih secara acak, lalu menggabungkan prediksinya untuk hasil yang lebih stabil. Dengan bootstrapping, RF meningkatkan generalisasi terhadap data baru. Algoritma ini unggul dalam menangani data berdimensi tinggi, robust terhadap noise, serta minim tuning hyperparameter dibandingkan boosting. Selain itu, RF menyediakan feature importance, yang membantu memahami faktor utama dalam klasifikasi kualitas air.

2.3.3 Gradient Boosting Machine (GBM)

Gradient Boosting Machine (GBM) adalah teknik boosting yang meningkatkan akurasi secara bertahap dengan membangun pohon keputusan yang mengoreksi kesalahan sebelumnya[12]. GBM mengoptimalkan fungsi loss melalui gradient descent, memungkinkan prediksi lebih efisien. Dengan penyesuaian bobot di setiap iterasi, model menjadi lebih responsif terhadap kesalahan. Namun, GBM rentan terhadap overfitting jika jumlah pohon berlebihan atau parameter tidak diatur dengan baik, sehingga optimasi hyperparameter diperlukan untuk keseimbangan bias dan varians.

2.3.4 Extreme Gradien Boosting(XGBoost)

Extreme Gradient Boosting (XGBoost) adalah varian gradient boosting yang dikembangkan oleh Chen dan Guestrin [13]untuk meningkatkan efisiensi komputasi dan mengatasi overfitting melalui regularisasi L1 dan L2[5]. Algoritma ini unggul dalam menangani missing values dengan sparsity-aware split finding dan mempercepat pencarian split terbaik melalui histogram-based optimization. Berkat kemampuannya menangani data berdimensi tinggi dan fleksibilitas dalam optimasi hyperparameter, XGBoost 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)[14]. Bayesian Optimization menggunakan pendekatan probabilistik untuk mencari kombinasi hyperparameter terbaik secara efisien, berbeda dari grid search atau random search yang mencoba kombinasi secara eksplisit. Dalam penelitian ini, Bayesian Optimization dikombinasikan dengan CV untuk mencegah overfitting dan

This Journal is licensed under a Creative Commons Attribution 4.0 International License

cBuilding of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2546−2558
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7014

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[15].

2.3.6 Stacking

Stacking adalah teknik ensemble learning yang meningkatkan akurasi dengan menggabungkan prediksi dari beberapa base models ke dalam meta-learner. Dalam penelitian ini, base models terdiri dari LGBM, RF, GBM, dan XGBoost, dengan Random Forest sebagai meta-learner. Prosesnya meliputi dua tahap: pertama, base models dilatih secara independen dan menghasilkan prediksi; kedua, prediksi tersebut digunakan sebagai fitur tambahan untuk melatih meta-learner. Stacking mengombinasikan keunggulan berbagai algoritma, meningkatkan generalisasi model terhadap data baru.

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. 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.

Metrik-metrik evaluasi yang digunakan dijelaskan sebagai berikut:

a. Akurasi mengukur persentase total sampel yang berhasil diklasifikasikan dengan benar, baik sebagai air layak minum maupun tidak layak minum[17]. Semakin tinggi akurasi, semakin baik model dalam mengklasifikasikan data secara Kesselring. Akurasi 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)[18]. 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)[19]. 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[13]. F1-score dirumuskan sebagai:

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

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).

Copyright © 2025 **Author**, Page 2549
This Journal is licensed under a Creative Commons Attribution 4.0 International License

cBuilding of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2546−2558
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7014

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
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 kelas 0 mencakup 61%, sedangkan kelas 1 mencakup 39% dari total data. Perbedaan proporsi ini menunjukkan adanya sedikit ketidakseimbangan dalam dataset, meskipun tidak dalam tingkat yang ekstrem. Dalam konteks klasifikasi, kondisi ini dapat menyebabkan model lebih cenderung memprediksi kelas mayoritas, sehingga perlu dilakukan analisis lebih lanjut terhadap metrik evaluasi seperti precision, recall, dan F1-score untuk memastikan performa model tetap optimal.

Untuk menangani potensi bias dalam klasifikasi, beberapa strategi dapat dipertimbangkan, seperti resampling data (oversampling kelas minoritas atau undersampling kelas mayoritas) atau penggunaan metode berbasis pembobotan dalam model. Selain itu, evaluasi model sebaiknya tidak hanya bergantung pada akurasi, tetapi juga menggunakan Precision-Recall dan f1 score untuk memahami keseimbangan prediksi antar kelas dengan lebih baik. Dengan memahami distribusi ini, langkah-langkah yang lebih tepat dapat diambil untuk meningkatkan kualitas prediksi model.

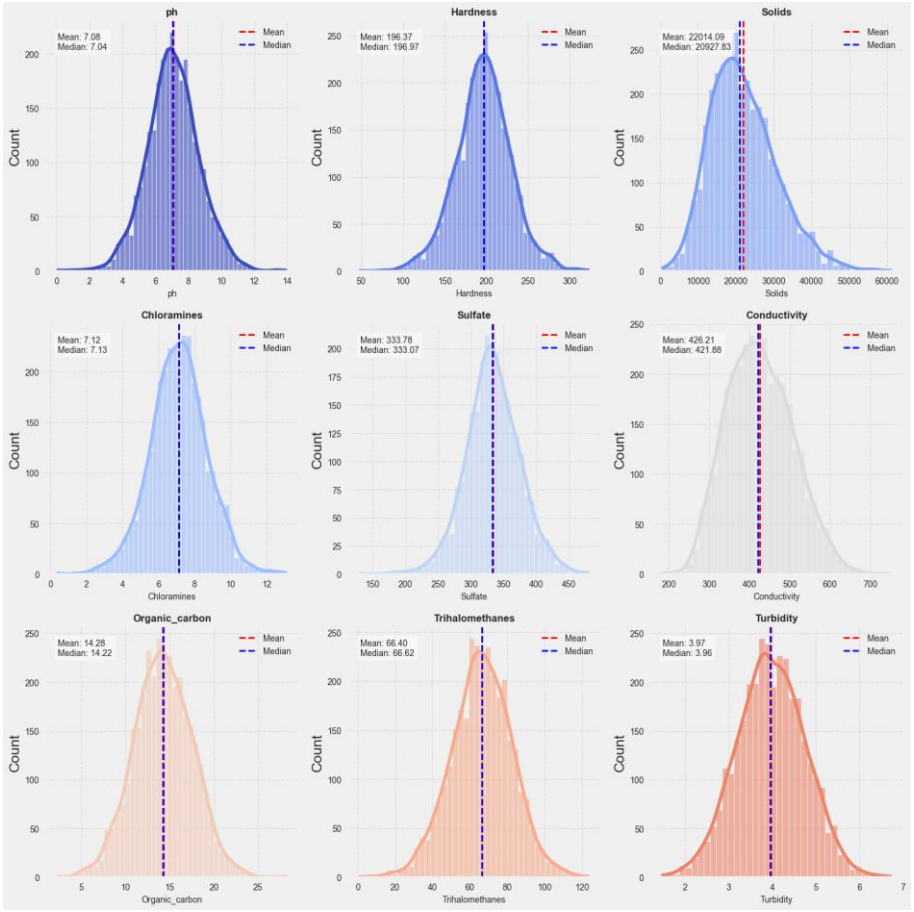
3.2.1 Distribusi Dataset

Distribusi setiap fitur dalam dataset ini menunjukkan karakteristik yang umumnya mendekati normal, dengan nilai mean dan median yang hampir berimpit pada sebagian besar variabel seperti yang ditunjukkan pada gambar 2.Distribusi Dataset [19]. pH memiliki rata-rata 7.08 dan median 7.04, menunjukkan distribusi yang simetris tanpa indikasi skewness yang signifikan. Hal serupa terjadi pada Hardness, dengan mean dan median yang sama, yaitu 196.97, yang mengindikasikan distribusi yang seimbang. Chloramines juga memiliki distribusi normal dengan rata-rata 7.12 dan median 7.13, serta Sulfate dengan rata-rata 333.78 dan median 333.07. Variabel Organic Carbon memiliki mean 14.28 dan median 14.22, sementara Trihalomethanes menunjukkan nilai rata-rata 66.40 dan median 66.62, yang keduanya mencerminkan distribusi simetris tanpa adanya pencilan yang signifikan. Turbidity, dengan mean 3.97 dan median 3.96, juga menunjukkan pola distribusi yang hampir normal.

Namun, beberapa fitur menunjukkan sedikit skewness ke kanan, yang mengindikasikan keberadaan beberapa nilai ekstrem yang lebih tinggi dibandingkan mayoritas data. Solids memiliki mean 22014.09 dan median 20927.26, dengan selisih yang cukup besar antara keduanya, menunjukkan kemungkinan adanya nilai outlier yang mempengaruhi distribusi. Hal yang serupa terjadi pada Conductivity, dengan rata-rata 426.21 dan median 421.88, yang mengindikasikan skewness ke arah kanan meskipun dalam tingkat yang lebih kecil. Perbedaan antara mean dan median ini menunjukkan bahwa beberapa pengamatan memiliki nilai yang jauh lebih tinggi dibandingkan nilai tipikal dalam dataset. Secara keseluruhan, dataset ini cenderung memiliki distribusi yang seimbang, tetapi pada fitur dengan skewness yang lebih tinggi, diperlukan analisis lebih lanjut untuk memastikan apakah nilai ekstrem tersebut perlu ditangani melalui normalisasi atau metode penanganan outlier lainnya sebelum digunakan dalam pemodelan lebih lanjut.

Copyright © 2025 **Author**, Page 2550
This Journal is licensed under a Creative Commons Attribution 4.0 International License

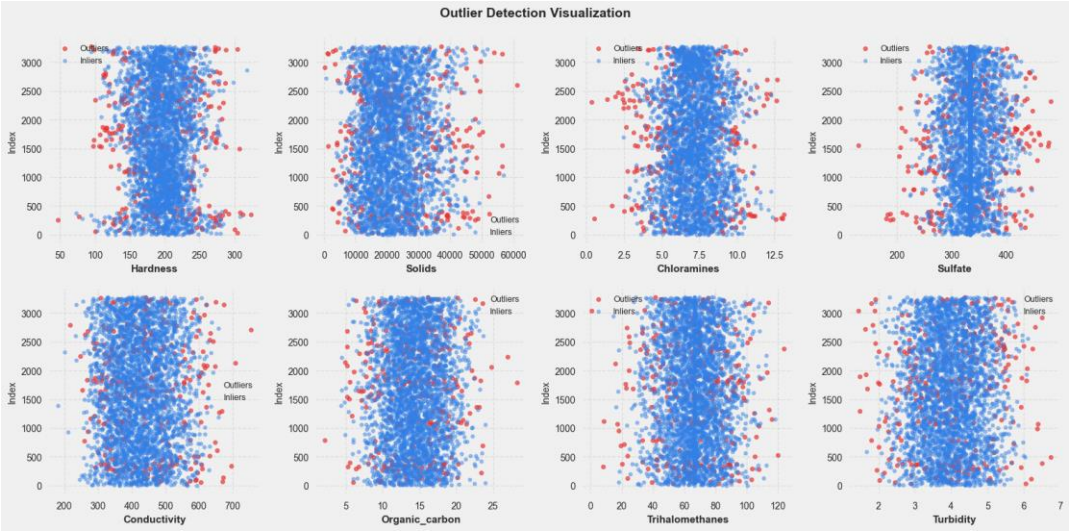
cBuilding of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2546−2558
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7014



Gambar 2. Distribusi Dataset

3.2.2 Deteksi Outlier

Deteksi outlier penting dalam eksplorasi data untuk mengidentifikasi nilai ekstrem yang dapat mempengaruhi analisis[19]. Berdasarkan scatterplot yang ditunjjukkan pada gambar 3.Deteksi Outlier, 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

Fitur 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.

Copyright © 2025 **Author**, Page 2551
This Journal is licensed under a Creative Commons Attribution 4.0 International License

cBuilding of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2546−2558
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7014

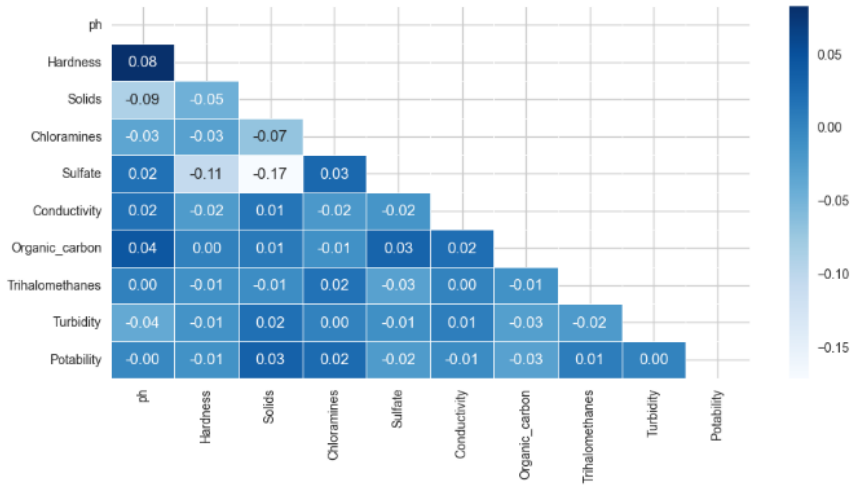
Sulfate umumnya berada dalam 250–400, dengan outlier di bawah 200 dan di atas 450, dipengaruhi oleh sumber air dan tingkat kontaminasi. Conductivity mayoritas 200–600, namun beberapa outlier melampaui rentang ini, menunjukkan variasi ion terlarut. Organic Carbon memiliki distribusi 5–20 dengan beberapa 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, dikoreksi, atau dipertahankan guna memastikan akurasi dan representasi kondisi nyata kualitas air.

3.2.3 Analisis Korelasi Antar Parameter Kualitas Kualitas Air

Analisis korelasi antar parameter kualitas air menggunakan matriks korelasi Pearson seperti yang ditunjukkan pada gambar 4.Korelasi Matriks menunjukkan bahwa hubungan antar variabel cenderung lemah, dengan sebagian besar nilai korelasi berada dalam rentang -0.17 hingga 0.08[20]. pH memiliki korelasi tertinggi dengan Hardness (0.08), yang menunjukkan bahwa air dengan tingkat kesadahan tinggi cenderung memiliki pH yang sedikit lebih tinggi, meskipun hubungannya sangat lemah. Selain itu, pH memiliki korelasi negatif dengan Solids (-0.09), sementara Chloramines (-0.03) tidak memiliki pengaruh signifikan terhadap pH. Hardness sendiri menunjukkan korelasi negatif dengan Sulfate (-0.11) dan Solids (-0.05), mengindikasikan bahwa peningkatan kadar sulfat atau padatan terlarut dapat sedikit menurunkan tingkat kesadahan. Namun, hubungan ini tetap lemah dan tidak cukup signifikan secara linear.



Gambar 4. Korelasi Matriks

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.

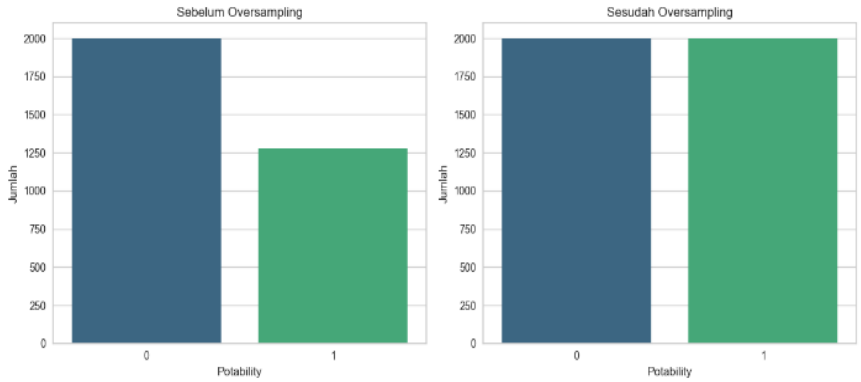
3.3 Prepocesing

Dalam tahap preprocessing data sebelum pemodelan, dilakukan berbagai teknik untuk meningkatkan kualitas data dan memastikan model dapat bekerja secara optimal. Salah satu tantangan utama yang dihadapi adalah ketidakseimbangan kelas pada variabel target"Potability, sebagai mana yang ditunjjukan pada gambar 5.Distribusi Label di mana jumlah sampel air tidak layak konsumsi (kelas 0) sebanyak 2.000 sampel, sedangkan sampel air layak konsumsi (kelas 1) hanya 1.250 sampel. Ketidakseimbangan ini dapat menyebabkan model lebih condong untuk memprediksi kelas mayoritas, sehingga dilakukan teknik oversampling menggunakan SMOTE (Synthetic Minority Over-sampling Technique)[8][21]. Setelah proses oversampling, jumlah data untuk kedua kelas menjadi seimbang, yaitu 2.000

Copyright © 2025 **Author**, Page 2552
This Journal is licensed under a Creative Commons Attribution 4.0 International License

cBuilding of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2546−2558
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7014

sampel untuk masing-masing kelas, sehingga model memiliki kesempatan yang lebih adil dalam mempelajari karakteristik air layak dan tidak layak konsumsi.



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.

3.4 Model Developmet

Dalam tahap preprocessing data sebelum pemodelan, dilakukan berbagai teknik untuk meningkatkan kualitas data dan memastikan model dapat bekerja secara optimal. Salah satu tantangan utama yang dihadapi adalah ketidakseimbangan kelas pada variabel target "Potability," di mana jumlah sampel air tidak layak konsumsi (kelas 0) sebanyak 2.000 sampel, sedangkan sampel air layak konsumsi (kelas 1) hanya 1.250 sampel. Ketidakseimbangan ini dapat menyebabkan model lebih condong untuk memprediksi kelas mayoritas, sehingga dilakukan teknik oversampling menggunakan SMOTE (Synthetic Minority Over-sampling Technique). 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
Random Forest	0.845	0.8529	0.845	0.8448
Gradient Boosting	0.6888	0.6949	0.6888	0.6883
XGBoost	0.8275	0.8284	0.8275	0.8274
LGBM	0.795	0.7952	0.795	0.7951

Pada tabel 2 bisa dilakukan analisis lebih lanjut terhadap LightGBM (LGBM), yang memiliki akurasi 79.5%, menunjukkan bahwa model ini cukup kompetitif dengan precision 79.52% dan recall 79.5%, yang berarti model mampu mengenali kelas dengan tingkat kesalahan yang relatif rendah. Berdasarkan classification report untuk LGBM, kelas 0 (tidak layak konsumsi) memiliki recall 79%, sedangkan kelas 1 (layak konsumsi) memiliki recall 80%, yang menunjukkan keseimbangan deteksi antara kedua kelas. Meskipun performanya masih di bawah Random Forest dan XGBoost, LGBM tetap menjadi pilihan yang menarik karena efisiensi komputasi yang lebih baik. Untuk meningkatkan performa lebih lanjut, langkah selanjutnya adalah hyperparameter tuning menggunakan Bayesian Search CV, yang diharapkan dapat mengoptimalkan parameter model untuk mencapai hasil yang lebih baik.

Copyright © 2025 **Author**, Page 2553
This Journal is licensed under a Creative Commons Attribution 4.0 International License

cBuilding of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2546−2558
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7014

3.5 Random Forest

Setelah dilakukan tuning hyperparameter menggunakan Bayesian Search CV, pada tabel 3menunjukkan performa model Random Forest mengalami peningkatan yang signifikan dibandingkan dengan sebelumnya. Dari hasil evaluasi, model yang telah dioptimalkan mencapai akurasi sebesar 85.38%, yang menunjukkan peningkatan dibandingkan versi sebelumnya yang memiliki akurasi 84.5%. Selain itu, metrik lain seperti precision, recall, dan f1-score juga mengalami peningkatan, dengan precision mencapai 86.78%, recall 85.38%, dan f1-score 85.31%. Hasil classification report menunjukkan bahwa model mampu mendeteksi kelas 0 (tidak layak konsumsi) dengan recall 94%, yang berarti sebagian besar sampel air yang tidak layak konsumsi dapat teridentifikasi dengan baik oleh model. Sebaliknya, kelas 1 (layak konsumsi) memiliki recall 77%, yang berarti masih terdapat beberapa sampel yang tidak terklasifikasi dengan sempurna sebagai layak konsumsi. Namun, dengan precision kelas 1 yang tinggi, yaitu 94%, dapat disimpulkan bahwa ketika model memprediksi bahwa air layak konsumsi, kemungkinan besar prediksi tersebut benar

Tabel 3. Hasil Random Forest dengan Bayesian Search CV

Algoritma Klasifikasi	Random Forest
Akurasi	0.85375
Presisi	0.86782
Recall	0.85375
F1-Score	0.85314
Parameter	OrderedDict([('max_depth', 20), ('max_features', 0.17454578258063885), ('min_samples_leaf', 1), ('min_samples_split', 3), ('n_estimators', 780)])

Optimasi model dilakukan dengan menyesuaikan lima hyperparameter utama, yaitu max_depth, max_features, min_samples_leaf, min_samples_split, dan n_estimators. Hasil tuning menunjukkan bahwa parameter terbaik untuk max_depth adalah 20, yang memberikan keseimbangan antara kompleksitas model dan risiko overfitting. Untuk max_features, nilai optimal yang ditemukan adalah 0.1745, yang berarti model hanya menggunakan sekitar 17.45% dari total fitur pada setiap pohon keputusan, meningkatkan efisiensi sambil tetap mempertahankan akurasi. Parameter min_samples_leaf ditetapkan pada 1, yang memungkinkan model menangkap pola yang lebih kompleks dalam data. Sementara itu, min_samples_split diatur menjadi 3, yang memastikan bahwa pembagian pada setiap node tidak terlalu cepat sehingga dapat meningkatkan generalisasi model. Terakhir, n_estimators dioptimalkan menjadi 780, yang berarti model membangun 780 pohon keputusan untuk meningkatkan prediksi secara keseluruhan. Dengan kombinasi parameter ini, Random Forest dapat mencapai keseimbangan antara bias dan varians, menghasilkan model yang lebih akurat dan dapat diandalkan untuk klasifikasi potabilitas air.

3.6 Gradient Boosting

Setelah dilakukan optimasi hyperparameter menggunakan Bayesian Search CV, model Gradient Boosting menunjukkan peningkatan performa yang signifikan dalam mengklasifikasikan potabilitas air. Pada tabel 4 menunjukan hasil model gradient boosting meningkat hingga 84.38%, dibandingkan dengan sebelumnya yang hanya 68.88% sebelum tuning. Peningkatan ini juga tercermin dalam metrik evaluasi lainnya, di mana nilai precision mencapai 84.99%, recall sebesar 84.38%, dan f1-score sebesar 84.36%. Berdasarkan classification report, model menunjukkan kecenderungan yang lebih baik dalam mengklasifikasikan air yang tidak layak konsumsi (kelas 0), dengan recall sebesar 90%, yang berarti hampir semua sampel air yang tidak layak berhasil diidentifikasi dengan benar. Sementara itu, kelas 1 (air layak konsumsi) memiliki recall 79%, yang mengindikasikan bahwa masih terdapat beberapa kasus air layak yang salah diklasifikasikan sebagai tidak layak. Namun, precision yang tinggi pada kelas ini, yakni 90%, menunjukkan bahwa prediksi air layak konsumsi memiliki tingkat kepercayaan yang tinggi.

Tabel 4. Hasil Gradien Boosting dengan Bayesian Search CV

Algoritma Klasifikasi	Gradient Boosting
Akurasi	0.84375
Presisi	0.84989
Recall	0.84375
F1-Score	0.84364
Parameter	OrderedDict([('learning_rate', 0.10824270964420458), ('max_depth', 10), ('min_samples_leaf', 1), ('min_samples_split', 2), ('n_estimators', 287), ('subsample', 1.0)])

Peningkatan performa model diperoleh melalui pemilihan hyperparameter yang optimal. Learning rate terbaik yang ditemukan adalah 0.1082, yang menentukan seberapa besar perubahan bobot pada setiap iterasi pembelajaran. Kedalaman maksimum pohon keputusan (max_depth) ditetapkan pada 10, memberikan keseimbangan antara kompleksitas model dan risiko overfitting. Parameter min_samples_leaf diatur pada 1, memungkinkan model menangkap pola dari data dengan ukuran minimum, sedangkan min_samples_split bernilai 2, memastikan bahwa pemisahan cabang dalam pohon keputusan terjadi hanya jika terdapat minimal dua sampel. Model menggunakan 287

Copyright © 2025 **Author**, Page 2554
This Journal is licensed under a Creative Commons Attribution 4.0 International License

cBuilding of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2546−2558
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7014

estimators, yang berarti jumlah total pohon keputusan yang digabungkan dalam proses boosting, memungkinkan model belajar lebih dalam terhadap pola dalam data. Selain itu, subsample ditetapkan sebesar 1.0, yang berarti seluruh dataset digunakan dalam setiap iterasi tanpa pengambilan sampel secara acak. Dengan kombinasi parameter ini, model Gradient Boosting menunjukkan performa yang lebih optimal dan akurat dalam menentukan potabilitas air, sehingga lebih layak digunakan dalam sistem klasifikasi air bersih.

3.7 XGBoost

Setelah dilakukan optimasi menggunakan Bayesian Search CV,pada tabel 5 hasil XGBoost setelah Bauesian Search CV menunjukkan peningkatan performa yang signifikan dalam mengklasifikasikan kelayakan air. Akurasi model mencapai 82.75%, yang mencerminkan peningkatan stabilitas dibandingkan dengan hasil sebelum tuning. Precision model berada di angka 82.88%, sementara recall tercatat sebesar 82.75%, dan f1-score mencapai 82.76%, menunjukkan keseimbangan yang baik antara mendeteksi sampel positif dan negatif. Dari classification report, kelas 0 (air tidak layak minum) memiliki recall 84%, menunjukkan bahwa model mampu mengidentifikasi mayoritas sampel dengan baik. Di sisi lain, kelas 1 (air layak konsumsi) memiliki recall 81%, menandakan adanya beberapa sampel yang kurang tepat diklasifikasikan. Namun, dengan precision kelas 1 sebesar 85%, model memberikan tingkat keakuratan yang cukup tinggi dalam memprediksi air yang dapat dikonsumsi.

Tabel 5. Hasil XGBoost dengan Bayesian Search CV

Algoritma Klasifikasi	XGBoost
Akurasi	0.8275
Presisi	0.8287
Recall	0.8275
F1-Score	0.8276
Parameter	OrderedDict([('colsample_bytree', 1.0), ('gamma', 0.1), ('learning_rate', 0.01), ('max_depth', 10), ('min_child_weight', 1), ('n_estimators', 1000), ('subsample', 0.5489614597924242)])

Proses tuning dilakukan dengan menentukan kombinasi hyperparameter yang optimal. Model menggunakan colsample_bytree sebesar 1.0, yang berarti setiap pohon dalam model mempertimbangkan seluruh fitur yang tersedia. Parameter gamma sebesar 0.1 memberikan batasan pada pemisahan cabang guna menghindari overfitting. Nilai learning rate sebesar 0.01 memastikan model belajar secara bertahap agar lebih stabil. Selain itu, max_depth ditetapkan pada 10, memungkinkan model menangkap pola yang lebih kompleks dalam data. Parameter min_child_weight sebesar 1 memastikan setiap pemisahan dilakukan dengan data yang cukup. Model menggunakan 1000 estimators, menunjukkan bahwa jumlah pohon keputusan yang digunakan cukup besar untuk meningkatkan kinerja model. Dengan nilai subsample sebesar 0.5489, model hanya menggunakan sekitar 55% data dalam setiap iterasi untuk mengurangi overfitting. Dengan kombinasi hyperparameter ini, model XGBoost mampu memberikan keseimbangan yang lebih baik antara bias dan varians, menjadikannya lebih andal dalam menentukan kelayakan air minum.

3.8 LGBM

Setelah dilakukan optimasi hyperparameter menggunakan Bayesian Search CV, model LightGBM (LGBM) menunjukkan performa yang cukup baik berdasarkan tabel 6 menunjukkan peningkatan sebesar 81.88%. Precision model mencapai 81.95%, sementara recall berada di angka 81.88%, menghasilkan f1-score sebesar 81.89%. Pada klasifikasi kelas 0 (air tidak layak minum), model memiliki recall 83%, menunjukkan kemampuannya dalam mengidentifikasi sebagian besar sampel yang tidak layak. Sementara itu, untuk kelas 1 (air layak minum), recall tercatat sebesar 81%, mengindikasikan bahwa model masih memiliki beberapa kesalahan dalam mengenali sampel yang seharusnya dikategorikan sebagai layak minum. Dengan precision kelas 1 yang mencapai 84%, model memiliki keakuratan yang baik dalam memprediksi air yang layak konsumsi.

Tabel 6. Hasil LGBM dengan Bayesian Search CV

Algoritma Klasifikasi	LGBM
Akurasi	0.81875
Presisi	0.81953
Recall	0.81875
F1-Score	0.81885
Parameter	OrderedDict([('colsample_bytree', 0.9132400290623335), ('lambda_l1', 0.1), ('lambda_l2', 0.1), ('learning_rate', 0.04889038160528904), ('max_depth', 10), ('min_child_weight', 1), ('n_estimators', 1000), ('subsample', 0.5)])

Hyperparameter yang diperoleh dari tuning menunjukkan kombinasi yang optimal untuk meningkatkan performa LGBM. Parameter colsample_bytree sebesar 0.91 memungkinkan model untuk mempertimbangkan

This Journal is licensed under a Creative Commons Attribution 4.0 International License

cBuilding of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2546−2558
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7014

sebagian besar fitur dalam setiap pohon keputusan. Nilai lambda_l1 dan lambda_l2 masing-masing ditetapkan sebesar 0.1, yang membantu mengurangi overfitting melalui regularisasi. Learning rate sebesar 0.0489 memastikan model belajar secara bertahap untuk meningkatkan generalisasi. Dengan max_depth sebesar 10, model dapat menangkap hubungan yang lebih kompleks dalam data tanpa kehilangan efisiensi. Parameter min_child_weight bernilai 1, mencegah pembentukan pohon yang terlalu dalam dan tidak signifikan. Selain itu, model menggunakan 1000 estimators, memungkinkan pembelajaran lebih mendalam, dan subsample sebesar 0.5, yang mengindikasikan bahwa model hanya menggunakan 50% data pada setiap iterasi untuk meningkatkan generalisasi. Dengan kombinasi ini, LGBM mampu memberikan prediksi yang lebih stabil dan akurat dalam menentukan kelayakan air minum.

3.9 Stacking

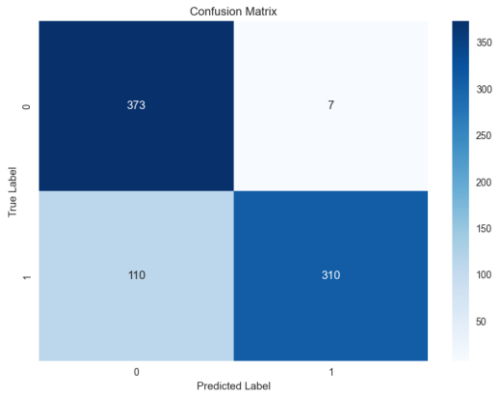
Stacking merupakan teknik ensemble learning yang digunakan untuk meningkatkan performa model dengan mengombinasikan prediksi dari beberapa model dasar (base models) menggunakan model tambahan yang disebut meta model. Dalam kasus ini, stacking digunakan untuk meningkatkan akurasi dalam masalah klasifikasi biner dengan menggabungkan keunggulan dari berbagai algoritma machine learning. Model-model dasar yang digunakan adalah LightGBM (LGBM), Random Forest (RF), Gradient Boosting (GB), dan XGBoost (XGB), sementara meta model yang digunakan adalah Random Forest dengan 500 estimators dan max_depth = 10.

Pada implementasi ini, setiap model dasar memiliki karakteristik yang berbeda dalam menangani data. LightGBM dikenal karena kecepatannya dan kemampuannya dalam menangani dataset besar dengan fitur numerik. Random Forest sangat baik dalam mengurangi varians melalui metode bootstrap aggregation (bagging) dan memberikan hasil yang lebih stabil dibandingkan model individual. Gradient Boosting dan XGBoost, sebagai model boosting, secara iteratif memperbaiki kesalahan dari prediksi sebelumnya, sehingga mampu menangani pola yang kompleks dalam data. Dengan memanfaatkan keunggulan dari setiap model dasar, stacking memungkinkan kombinasi prediksi yang lebih akurat dibandingkan hanya menggunakan satu model saja.

Tabel 7. Hasil Stacking

Classifier	Akurasi	Precision	Recall	F1-Score
Random Forest + Bayes Search CV	0.85375	0.86782	0.85375	0.85214
Gradient Boosting + Bayes Search CV	0.84375	0.84989	0.84375	0.84364
XGBoost + Bayes Search CV	0.8275	0.8287	0.8275	0.8276
LGBM + Bayes Search CV	0.81875	0.818953	0.81857	0.81885
Stacking*	0.8538	0.8802	0.8538	0.8523

Proses stacking dilakukan dalam dua tahap utama. Pada tahap pertama, setiap model dasar dilatih menggunakan dataset pelatihan, kemudian masing-masing model menghasilkan probabilitas prediksi untuk setiap sampel dalam dataset validasi. Probabilitas ini kemudian digunakan sebagai fitur tambahan untuk pelatihan meta model. Dengan kata lain, output dari model dasar menjadi input bagi meta model. Pada tahap kedua, meta model menerima probabilitas prediksi dari keempat model dasar dan belajar dari pola kombinasi prediksi tersebut untuk menghasilkan keputusan akhir yang lebih akurat. Meta model yang digunakan dalam penelitian ini adalah Random Forest dengan 500 estimators dan max_depth = 10, yang dipilih karena kemampuannya dalam menangkap pola yang kompleks tanpa overfitting secara berlebihan.



Gambar 6. Confusion Matriks Stacking

Hasil evaluasi yang ditunjukkan pada Gambar 6.Confusion Matriks Stacking dan tabel 7.Hasil Stacking menunjukkan bahwa stacking model memiliki performa lebih baik dibandingkan dengan masing-masing model dasar secara individual. Dari confusion matrix yang ditampilkan, model memiliki 373 true negatives (TN), 7 false positives (FP), 110 false negatives (FN), dan 310 true positives (TP). Dari hasil ini, dapat dihitung metrik evaluasi utama: akurasi 85.38%, precision 88.02%, recall 85.38%, dan F1-score 85.23%.

Copyright © 2025 **Author**, Page 2556
This Journal is licensed under a Creative Commons Attribution 4.0 International License

cBuilding of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2546−2558
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7014

Jika dibandingkan dengan model individu, Random Forest memiliki akurasi tertinggi di antara base models, yaitu 84.5%, sementara Gradient Boosting memiliki keseimbangan yang cukup baik dengan akurasi 84.38%, precision 84.99%, recall 84.38%, dan F1-score 84.36%. XGBoost mencapai akurasi 82.75%, sementara LightGBM sedikit lebih rendah dengan akurasi 81.88%. Stacking model berhasil meningkatkan akurasi menjadi 85.38%, yang menunjukkan peningkatan dari model individu terbaik. Precision meningkat hingga 88.02%, mengindikasikan bahwa model lebih akurat dalam mengidentifikasi kelas positif, sementara recall 85.38% menunjukkan bahwa model mampu menangkap lebih banyak sampel positif dibandingkan dengan model dasar.

Dari confusion matrix, terlihat bahwa model memiliki jumlah false negatives (FN) sebesar 110, yang berarti model masih memiliki kelemahan dalam mengidentifikasi beberapa sampel positif. Namun, jumlah false positives (FP) yang hanya 7 menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam menghindari kesalahan klasifikasi untuk kelas negatif. Hal ini menjadi penting dalam kasus di mana false positives harus diminimalkan, seperti dalam deteksi penyakit atau penipuan keuangan.

Secara keseluruhan, peningkatan performa ini terjadi karena stacking mampu mengombinasikan keunggulan masing-masing model dasar dan mengurangi bias serta varians dalam model individu. Dalam kasus ini, model boosting seperti XGBoost dan Gradient Boosting memberikan daya prediksi yang kuat terhadap pola yang kompleks, sementara Random Forest dan LightGBM memberikan stabilitas dan fleksibilitas yang lebih baik. Meta model kemudian mengolah kombinasi prediksi dari base models untuk menghasilkan keputusan akhir yang lebih akurat dan generalisasi yang lebih baik terhadap data baru. Oleh karena itu, teknik stacking dapat menjadi pendekatan yang efektif dalam berbagai permasalahan klasifikasi yang membutuhkan akurasi tinggi dan stabilitas model

4. KESIMPULAN

Penelitian ini berhasil mengembangkan sistem klasifikasi kualitas air berbasis pembelajaran mesin dengan pendekatan stacking ensemble, yang menggabungkan LGBM, Random Forest, Gradient Boosting, dan XGBoost, serta dioptimalkan menggunakan Bayesian Optimization. Dengan preprocessing data seperti imputasi, normalisasi, feature engineering, dan oversampling SMOTE, sistem ini mencapai akurasi terbaik sebesar 85,38%, mengungguli model individu lainnya. Metode ini lebih efisien dibandingkan pendekatan manual seperti Water Quality Index (WQI) dan STORET. Namun, penelitian ini masih memiliki keterbatasan, seperti kurangnya parameter biologis dalam dataset dan distribusi kelas yang belum sepenuhnya seimbang. Untuk pengembangan lebih lanjut, model ini dapat diperluas dengan dataset yang lebih beragam serta diimplementasikan dalam sistem pemantauan kualitas air secara real-time guna meningkatkan keandalan dalam memastikan ketersediaan air bersih bagi masyarakat.

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] A. Mukati, R. Rathore, G. Patidar, and M. Patel, “Memahami Konsep Potensi Air Melalui Pembelajaran Mesin,” *J Curr Trends Comp Sci Res*, vol. 3, no. 2, 2024
[7] A. P. Sari *et al.*, “Classification of Drinking Water Source Suitability in West Java Using XGBoost and Cluster Analysis Based on SHAP Values,” Indonesian Journal of Statistics and its Applications, vol. 8, no. 2, pp. 202–214, 2024.
[8] H. Hairani, T. Widiyaningtyas, and D. Dwi Prasetya, “Addressing Class Imbalance of Health Data: A Systematic Literature Review on Modified Synthetic Minority Oversampling Technique (SMOTE) Strategies,” *JOIV Int. J. Informatics Vis.*, vol. 8, no. 3, pp. 1310–1318, 2024.
[9] K. Yu, S. Xia, Y. Zhang, and S. Wang, “Loan Approval Prediction Improved by XGBoost Model Based on Four-Vector Optimization Algorithm,” Preprints, vol. 0, pp. 35–44, doi: 10.20944/preprints202410.0783.v1.
[10] J. Yan *et al.*, “LightGBM: accelerated genomically designed crop breeding through ensemble learning,” *Genome Biol.*, vol. 22, no. 1, pp. 1–24, 2021, doi: 10.1186/s13059-021-02492-y.
[11] 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.
[12] R. Soelistijadi *et al.*, “Pemodelan Prediktif Menggunakan Metode Ensemble Learning XGBoost dalam Peningkatan Akurasi Klasifikasi Penyakit Ginjal,” Kesatria, vol. 5, no. 4, pp. 1866–1875, 2024.
[13] 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.
[14] 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.
[15] L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, “Cross validation for model selection: A review with examples

This Journal is licensed under a Creative Commons Attribution 4.0 International License

cBuilding of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2546−2558
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7014

from ecology," *Ecol. Monogr.*, vol. 93, no. 1, pp. 1–24, 2023, doi: 10.1002/ecm.1557.
[16] W. Musu, A. Ibrahim, and Heriadi, "Pengaruh Komposisi Data Training dan Testing terhadap Akurasi Algoritma C4.5," *Pros. Semin. Ilm. Sist. Inf. Dan Teknol. Inf.*, vol. X, no. 1, pp. 186–195, 2021.
[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] F. Jabnabillah and N. Margina, "Analisis Korelasi Pearson Dalam Menentukan Hubungan Antara Motivasi Belajar Dengan Kemandirian Belajar Pada Pembelajaran Daring," *J. Sintak*, vol. 1, no. 1, pp. 14–18, 2022.
[20] J. Brandt and E. Lanzén, "A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification," *2021*, p. 42, 2020, [Online]. Available: https://www.diva-portal.org/smash/record.jsf?pid=diva2:1519153

Copyright © 2025 **Author**, Page 2558
This Journal is licensed under a Creative Commons Attribution 4.0 International License