PERBANDINGAN NAIVE BAYES, RANDOM FOREST, XGBOOST UNTUK KLASIFIKASI MUTU AIR

Gavriel Joseph Lim [1] Jason Chainara Putra [2] Paulina Agusia [3]
Marcia Yanprincessa Utama [4]
[1)2)3)4)] Teknik Informatika, FTI Universitas Tarumanagara
Jl. Letjen S. Parman St No.1, Jakarta 11440 Indonesia
[1)]email : gavriel.535220049@stu.untar.ac.id [2)]email : jason.535220045@stu.untar.ac.id
[3)]email : paulina.535220048@stu.untar.ac.id [4)]email : marcia.535220044@stu.untar.ac.id

ABSTRAK

Penelitian ini bertujuan untuk membandingkan kinerja algoritma Naive Bayes, Random Forest, dan XGBoost dalam mengklasifikasikan kualitas air budidaya perikanan. Kualitas air merupakan faktor penting dalam menjaga kesehatan organisme akuatik serta mengoptimalkan produktivitas akuakultur. Penelitian ini menggunakan *dataset* yang mencakup beberapa parameter, seperti pH, suhu, oksigen terlarut, salinitas, dan kekeruhan, yang digunakan sebagai fitur masukan untuk memprediksi kelas kualitas air secara keseluruhan. Data dibagi menjadi dua bagian, yaitu data pelatihan dan data pengujian. Hasil eksperimen menunjukkan bahwa algoritma Random Forest mencapai akurasi klasifikasi tertinggi sekaligus waktu pemrosesan tercepat dibandingkan dengan XGBoost dan Naive Bayes. Algoritma XGBoost menghasilkan performa yang kompetitif dengan sensitivitas yang sedikit lebih tinggi terhadap variasi data, sedangkan Naive Bayes menunjukkan efisiensi komputasi yang tinggi namun dengan akurasi yang lebih rendah. Secara keseluruhan, Random Forest memberikan kinerja yang paling konsisten dan akurat dalam mengklasifikasikan kualitas air budidaya, sehingga dinilai paling cocok untuk diterapkan pada sistem pemantauan kualitas air secara *real-time.*

Key words

Klasifikasi, Naive Bayes, Random Forest, XGBoost, Akuakultur

1. Pendahuluan

Kualitas air merupakan faktor penting yang menentukan keberhasilan dalam sistem akuakultur karena faktor air sangat berpengaruh langsung terhadap kesehatan, pertumbuhan, dan produktivitas organisme air seperti ikan dan udang [1]. Dalam sistem budidaya intensif, perubahan kecil dalam komposisi kimia atau parameter fisika air dapat menimbulkan stres fisiologis yang berujung pada penurunan pertumbuhan, efisiensi pakan, serta tingkat kelangsungan hidup organisme budidaya. Oleh karena itu, menjaga keseimbangan kualitas air menjadi aspek fundamental dalam pengelolaan lingkungan tambak yang berkelanjutan dan produktif.

Parameter fisika dan kimia air seperti oksigen terlarut (DO), Biochemical Oxygen Demand (BOD), pH, suhu, amonia, dan kekeruhan digunakan sebagai indikator utama mutu perairan [2]. DO berperan penting dalam respirasi organisme, sementara BOD menunjukkan tingkat kebutuhan oksigen untuk menguraikan bahan organik dalam air. Nilai pH memengaruhi proses biokimia dan toksisitas amonia, sedangkan suhu berpengaruh terhadap metabolisme dan pertumbuhan organisme akuatik. Kombinasi dari parameter-parameter ini memberikan gambaran menyeluruh terhadap kondisi ekosistem akuatik dan memungkinkan dilakukan klasifikasi mutu perairan untuk menentukan apakah lingkungan tersebut mendukung kehidupan organisme budidaya. Perubahan kecil pada salah satu parameter dapat berdampak signifikan terhadap kelangsungan hidup organisme budidaya, terutama pada sistem tertutup atau semi-tertutup yang umum digunakan dalam tambak modern. Oleh karena itu, pemantauan kualitas air yang akurat, cepat, dan adaptif menjadi kebutuhan utama dalam sistem manajemen tambak berbasis data [3].

Pendekatan konvensional dalam pengelolaan kualitas air yang masih mengandalkan pengambilan sampel manual dan analisis laboratorium sering kali memakan waktu lama, membutuhkan biaya tinggi, serta kurang responsif terhadap perubahan kondisi lingkungan [4]. Selain itu, metode konvensional tidak mampu menangkap dinamika perubahan yang cepat terjadi di lingkungan tambak, terutama pada sistem intensif dengan kepadatan tebar tinggi dan siklus air yang cepat. Hal ini menyebabkan keterlambatan dalam tindakan korektif dan dapat berujung pada kerugian ekonomi yang signifikan. Perkembangan teknologi sensor dan sistem Internet of

Things (IoT) kini memungkinkan pemantauan parameter air secara waktu nyata dan berkelanjutan, membuka peluang integrasi dengan teknik *machine learning* (ML) untuk melakukan klasifikasi dan prediksi mutu air [5].

Teknologi sensor modern mampu menghasilkan data dalam jumlah besar (big data) dengan resolusi temporal tinggi. Data tersebut mencerminkan dinamika kualitas air secara berkelanjutan dan kompleks, yang sulit diinterpretasikan hanya dengan metode statistik konvensional. Di sinilah machine learning berperan penting. Pendekatan berbasis machine learning mampu mengenali pola kompleks dan hubungan non-linear antarparameter kualitas air yang sulit diidentifikasi secara manual [6]. Dengan kemampuan untuk belajar dari data historis, model ML dapat mengidentifikasi tren, menganalisis faktor-faktor penyebab perubahan kualitas air, dan memprediksi kondisi masa depan. Implementasi model prediktif ini dapat digunakan untuk memberikan peringatan dini terhadap potensi penurunan mutu air serta membantu pengambilan keputusan berbasis data dalam pengelolaan tambak yang efisien dan berkelanjutan.

Beberapa penelitian terdahulu telah menerapkan algoritma pembelajaran mesin seperti Support Vector Machine (SVM), K-Nearest Neighbors (KNN), dan Random Forest untuk melakukan analisis mutu air di lingkungan sungai, danau, dan waduk [7]. Namun, penerapan dalam konteks akuakultur masih terbatas meskipun data tambak memiliki karakteristik berbeda, seperti adanya pengaruh pakan, aktivitas plankton, serta sirkulasi air yang lebih dinamis dan fluktuatif [8]. Model yang efektif di perairan umum belum tentu memberikan hasil serupa pada sistem akuakultur karena kompleksitas dan variabilitas data yang lebih tinggi. Selain itu, sebagian besar penelitian terdahulu berfokus pada peningkatan akurasi tanpa mempertimbangkan interpretabilitas model dan efisiensi komputasi dalam konteks penerapan lapangan [9]. Dalam pengelolaan akuakultur modern, model yang baik tidak hanya akurat tetapi juga harus dapat dijelaskan agar mudah diimplementasikan oleh operator tambak dalam pengambilan keputusan rutin.

Sejumlah penelitian terkait juga menunjukkan adanya potensi besar penggunaan algoritma machine learning dalam klasifikasi mutu air. Danuri dan Pozi [10] membandingkan algoritma Random Forest, Gaussian Naive Bayes, serta Decision Tree untuk klasifikasi kualitas air kolam ikan dan menyimpulkan bahwa Gaussian Naive Bayes memberikan akurasi tertinggi. Temuan ini menunjukkan bahwa algoritma sederhana seperti Naive Bayes tetap kompetitif dalam kasus data dengan distribusi yang teratur. Zhang et al. [11] memanfaatkan multispektral UAV dengan menggunakan algoritma Random Forest, XGBoost, dan CatBoost untuk memprediksi parameter kekeruhan dan klorofil-a. Studi ini memperlihatkan bahwa integrasi data citra multispektral dengan pembelajaran mesin mampu meningkatkan resolusi spasial pemantauan kualitas air secara signifikan. Sementara itu, Liu et al. [12] menggunakan kombinasi fitur spektral dan tekstur dari citra satelit untuk mengidentifikasi kolam akuakultur dan memperoleh hasil terbaik dengan model XGBoost (akurasi 96,15%). Temuan tersebut menegaskan bahwa metode *gradient boosting* seperti XGBoost unggul dalam menangkap hubungan non-linear dan kompleksitas tinggi antarvariabel lingkungan.

Studi lain oleh Elmotawakkil et al. [13] membandingkan beberapa model seperti Artificial Neural Network (ANN), SVM, Random Forest, dan XGBoost untuk memprediksi *Water Quality Index* dan menemukan bahwa XGBoost memiliki performa paling baik dan stabil. Hasil ini menunjukkan bahwa model boosting lebih adaptif terhadap data dengan noise tinggi dan hubungan antarfitur yang kompleks. Remya et al. [14] juga melaporkan bahwa Random Forest menunjukkan kinerja paling baik dalam klasifikasi mutu air di beberapa sumber perairan di India dibandingkan dengan Naive Bayes dan XGBoost. Hal ini mengindikasikan bahwa performa algoritma sangat bergantung pada karakteristik dataset dan kondisi lingkungan tempat penelitian dilakukan.

Penelitian ini berfokus pada pengembangan model klasifikasi mutu air akuakultur dengan membandingkan tiga paradigma machine learning, yaitu Naive Bayes, Random Forest, dan XGBoost [15]. Ketiga model ini dipilih karena mewakili tiga pendekatan utama dalam pembelajaran mesin yang saling melengkapi Naive Bayes sebagai model probabilistik yang sederhana dan efisien; Random Forest sebagai model ansambel berbasis pohon dengan kemampuan generalisasi yang kuat; serta XGBoost sebagai model *boosting* modern dengan kemampuan mengatasi data besar dan hubungan non-linear yang kompleks. Dataset yang digunakan adalah *Aquaculture – Water Quality Dataset* dari Mendeley Data yang berisi 4300 observasi dengan 14 parameter [16], mencakup berbagai kondisi fisik dan kimia air tambak yang representatif terhadap sistem akuakultur tropis.

Penelitian ini bertujuan untuk mengembangkan model klasifikasi mutu air akuakultur berbasis machine learning serta membandingkan kinerja tiga algoritma utama, yaitu Naive Bayes, Random Forest, dan XGBoost, berdasarkan metrik akurasi, presisi, recall, F1-score, dan waktu pelatihan. Selain itu, penelitian ini juga mengevaluasi tingkat kepentingan fitur (*feature importance*) dari masing-masing parameter air terhadap hasil klasifikasi guna memahami faktor dominan yang memengaruhi mutu air. Dengan pendekatan ini, penelitian diharapkan dapat memberikan kontribusi berupa analisis komparatif lintas paradigma pembelajaran, pemetaan faktor dominan mutu air, dan rekomendasi model yang dapat diterapkan pada sistem pemantauan mutu air akuakultur secara otomatis [17]. Lebih jauh, hasil penelitian ini diharapkan menjadi dasar bagi pengembangan sistem pengawasan mutu air berbasis kecerdasan buatan (*AI-driven aquaculture monitoring system*) yang mendukung efisiensi produksi, keberlanjutan lingkungan, dan peningkatan ketahanan sistem akuakultur nasional.

2. Landasan Teori

2.1 Machine Learning

Machine learning merupakan salah satu bidang dalam ilmu komputer yang berfokus pada pengembangan sistem yang mampu melakukan pembelajaran melalui data. Sistem ini tidak hanya mengandalkan aturan eksplisit, tetapi membangun pola dari data yang tersedia untuk menghasilkan prediksi atau keputusan secara otomatis [18]. Teknologi ini telah banyak diterapkan di berbagai bidang seperti kesehatan, pertanian, transportasi, dan lingkungan [19].

Menurut Boyd, penerapan machine learning pada lingkungan sangat penting karena data yang dikumpulkan dari sensor biasanya bersifat besar dan tidak linier [20]. Dengan algoritma machine learning, hubungan antar parameter lingkungan seperti suhu, pH, dan oksigen terlarut dapat dianalisis dengan lebih efektif. Pendekatan ini memungkinkan peneliti memprediksi mutu air secara cepat dan akurat tanpa perlu melakukan pengujian manual di laboratorium [21].

2.2 Klasifikasi

Klasifikasi merupakan salah satu teknik dalam supervised learning yang digunakan untuk mengelompokkan data ke dalam kelas tertentu berdasarkan variabel masukan yang ada. Proses ini dilakukan dengan melatih model pada data berlabel agar model dapat mengenali pola hubungan antara fitur serta label kelasnya [22].

Dalam konteks mutu air akuakultur, klasifikasi digunakan untuk menentukan apakah suatu air tergolong baik, sedang, atau buruk berdasarkan parameter seperti suhu, oksigen, pH, dan masih banyak lagi [23]. Model klasifikasi yang baik mampu membedakan setiap kondisi mutu air secara konsisten walaupun terdapat variasi nilai pada data input.

2.3 Naive Bayes

Naive Bayes merupakan algoritma klasifikasi yang didasarkan pada Teorema Bayes, yaitu pendekatan probabilistik untuk menghitung kemungkinan suatu data masuk ke dalam kelas tertentu [24] Persamaan dasar Teorema Bayes ditunjukkan pada Persamaan (1):

$$P(X) = \frac{P(C) \times P(C)}{P(X)} \quad (1)$$

Keterangan:
$P(X)$: Probabilitas kelas C jika diketahui fitur X
$P(C)$: Probabilitas fitur X muncul pada kelas C
$P(C)$: Probabilitas awal kelas C
$P(X)$: Probabilitas total dari fitur X $P(X)$

2.4 Random Forest

Random Forest merupakan metode ensemble learning yang dikembangkan oleh Breiman. Algoritma ini menghasilkan sejumlah pohon keputusan secara acak dan menggabungkan hasilnya untuk memperoleh hasil akhir klasifikasi yang lebih akurat dan stabil.

Prinsip kerja dari Random Forest adalah melakukan proses *bootstrap aggregating*, yaitu membuat beberapa subset data latih secara acak, kemudian melatih setiap subset menggunakan model decision tree yang berbeda. Hasil akhir ditentukan dengan penentuan hasil melalui suara terbanyak seperti ditunjukkan pada persamaan (2) [25]:

$$\hat{y} = mode\{h_1(x), h_2(x), \dots, h_T(x)\} \quad (2)$$

Keterangan:
$h_1(x)$: Hasil prediksi dari pohon ke-t
T: Jumlah total pohon
$\hat{y}$: Hasil prediksi akhir

2.5 XGBoost

*Extreme Gradient Boosting (*XGBoost*)* merupakan pengembangan dari metode *gradient boosting decision tree* yang dioptimalkan untuk efisiensi komputasi dan akurasi. XGBoost membangun pohon keputusan secara bertahap, di mana setiap pohon baru fokus memperbaiki kesalahan dari model sebelumnya. Fungsi objektif XGBoost dapat dituliskan pada persamaan (3) [26]:

$$Obj = \sum_{i=1}^{n} \quad l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \quad \Omega(f_k) \quad (3)$$

Dengan fungsi regularisasi seperti pada persamaan (4):

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\left|\left|w\right|\right|^2 \quad (4)$$

Keterangan:
$l(y_i, \hat{y}_i)$: Loss function antara nilai aktual dan prediksi
$\Omega(f_k)$: Regularisasi kompleksitas model
γ: Parameter penalti kompleksitas
λ: Parameter regularisasi bobot
T: Jumlah leaf node
w: Bobot leaf node

2.6 Ensemble Learning

Ensemble Learning adalah pendekatan dalam machine learning yang memadukan beberapa model (*base learners*) untuk meningkatkan performa prediksi dibandingkan dengan hanya menggunakan model tunggal [27]

Terdapat dua metode utama, yaitu bagging (bootstrap aggregating) dan boosting. Bagging digunakan untuk meminimalisasi variance dengan melatih model pada subset data yang diambil acak dari dataset asli, contohnya pada algoritma Random Forest. Sedangkan boosting, seperti pada algoritma XGBoost, difokuskan pada peningkatan akurasi dengan membangun model secara urut, dimana setiap model baru berupaya untuk memperbaiki kesalahan dari model sebelumnya [28].

Kelebihan utama dari ensemble learning ini adalah kemampuan dalam mengurangi *overfitting* serta meningkatkan generalisasi model pada data baru. Untuk konteks klasifikasi mutu air, metode ini digunakan untuk membantu menghasilkan prediksi yang lebih akurat meskipun data bervariasi tinggi atau terdapat noise.

3. Hasil Percobaan

Proses klasifikasi data menggunakan aplikasi berbasis web. Pada bagian ini menjelaskan hasil percobaan terhadap tiga algortima klasifikasi, yaitu Naive Bayes, Random Forest, dan XGBoost. Evaluasi dilakukan dengan menggunakan metode K-Fold Validation serta pembagian data 70:30 dan 80:20 untuk melihat seberapa konsisten performa model terhadap variasi proporsi data latih dan data uji. Data yang digunakan berasal dari repositori Mendeley Data dengan link https://data.mendeley.com/datasets/y78ty2g293/1 dengan jumlah 4300 sampel data. Berikut merupakan langkah-langkah yang dilakukan pada pengujian klasifikasi, yaitu sebagai berikut:

1. Menyiapkan dataset dalam format dan memisahkannya menjadi data latih dan data uji dengan rasio 70:30 dan 80:20.
2. Melakukan preprocessing terhadap data dalam bentuk excel.
3. Menerapkan tiga algoritma klasifikasi menggunakan platform Python (Google Colab).
4. Mengukur performa tiap algoritma berdasarkan akurasi, presisi, recall, dan F1-score.
5. Melakukan analisis hasil klasifikasi dan membandingkan performa dari masing-masing model berdasarlan evaluasi.

3.1 K-Fold Validation

Hasil validasi model dengan metode K-Fold Validation yang dilihat dari nilai akurasi, presisi, recall, dan F1-macro menunjukkan bahwa Random Forest memperoleh performa paling tinggi, diikuti oleh XGBoost, sementara Naive Bayes memiliki performa cenderung rendah. Dapat dikatakan bahwa model ensemble seperti Random Forest dan XGBoost memiliki kemampuan generalisasi yang lebih baik dibandingkan model pendekatan statistik sederhana seperti Naive Bayes.

Tabel 1 Perbandingan algoritma dengan K-Fold Validation

	Naive Bayes	Random Forest	XGBoost
Accuracy	0.9021 ± 0.0125	0.9960 ± 0.0033	0.9928 ± 0.0045
Precision	0.9149 ± 0.0098	0.9960 ± 0.0033	0.9928 ± 0.0045
Recall	0.9064 ± 0.0119	0.9962 ± 0.0032	0.9930 ± 0.0044
F-1 macro	0.8994 ± 0.0136	0.9961 ± 0.0033	0.9929 ± 0.0044

3.2 Perbandingan 70:30

Pengujian model dengan pembagian data 70% untuk data latih dan 30% untuk data uji pada Tabel 2 menunjukan akurasi tertinggi oleh Random Forest, diikuti XGBoost , kemudian Naive Bayes.

Tabel 2 Perbandingan algortima dengan Split Evaluasi 70 30

	Naive Bayes	Random Forest	XGBoost
Accuracy	0.892	0.994	0.992
Precision	0.907	0.994	0.992
Recall	0.897	0.994	0.993
F-1 macro	0.889	0.994	0.992

3.3 Perbandingan 80:20

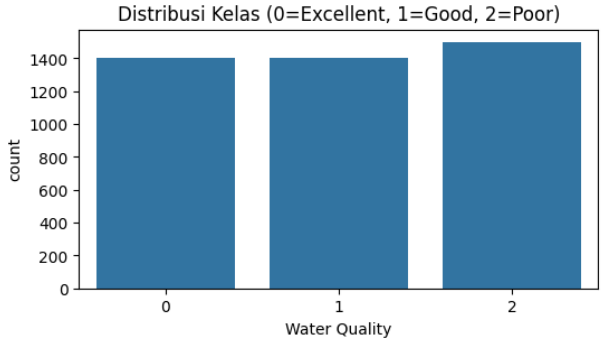
Hasil pada Tabel 3 menunjukan bahwa perubahan proporsi pada data latih yang menjadi 80% tidak mempengaruhi pola performa secara signifikan dimana Random Forest tetap menjadi model algoritma terbaik, kemudian diikuti oleh XGBoost dan Naive Bayes. Hal ini menunjukkan kestabilan model ensemble terhadap perubahan proporsi data.

Tabel 3 Perbandingan algoritma dengan Split Evaluasi 80 20

	Naive Bayes	Random Forest	XGBoost
Accuracy	0.893	0.993	0.992
Precision	0.900	0.993	0.992
Recall	0.888	0.993	0.993
F-1 macro	0.878	0.993	0.992

3.4 Distribusi Kelas Kualitas Air

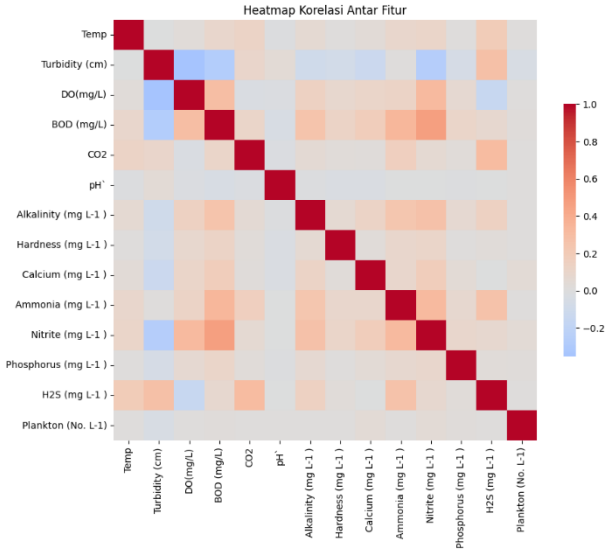
Pada Gambar 1 menunjukkan distribusi data berdasarkan tiga kategori kualitas air yaitu *Excellent*, *Good*, dan *Poor*. Terlihat bahwa jumlah data pada ketiga kelas relatif seimbang, dimana masing-masing memiliki kisaran sejumlah 1.400 hingga 1.500 sampel. Kondisi distribusi yang merata penting untuk membantu model klasifikasi dalam melakukan pembelajaran secara adil terhadap seluruh kelas, sehingga risiko bias terhadap banyak kelas mayoritas dapat diminimalisir.



Gambar 1 Distribusi Kelas Kualitas Ai

3.5 Heatmap Korelasi Antar Fitur Parameter Kualitas Air

Pada Gambar 2 menggambarkan hubungan korelasi antar parameter fisika dan kimia pada data kualitas air. Nilai korelasi positif ditunjukkan oleh warna merah, sedangkan nilai korelasi negatif oleh warna biru. Terlihat bahwa beberapa parameter memiliki hubungan yang cukup kuat, seperti BOD, dan DO yang menunjukkan korelasi negatif, hal tersebut menandakan bahwa peningkatan kadar BOD cenderung menurunkan kadar oksigen terlarut ketika di dalam air. Selain itu, parameter lain seperti pH, suhu (*Temperature*), dan amonia (*Ammonia*) menunjukkan korelasi yang lebih lemah dibandingkan sebagian besar variabel yang lain, menandakan bahwa parameter tersebut memiliki kontribusi independen terhadap kualitas air secara keseluruhan.

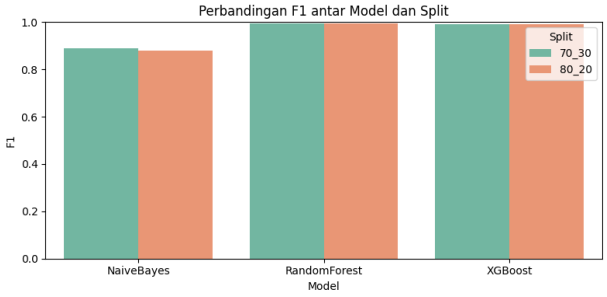


Gambar 2 Heatmap Korelasi Antar Fitur Parameter Kualitas Air

3.6 Perbandingan F1 antar Model dan Split

Dari ketiga algoritma pada tiap data yang terbagi, terlihat pada Gambar 3 yang menunjukkan Random Forest sebagai algoritma dengan performa paling konsisten dan tertinggi, diikuti oleh XGBoost, sedangkan pada Naive Bayes menunjukkan penurunan yang signifikan. Hasil tersebut dapat menjadi indikasi bahwa metode ensemble lebih unggul dalam mengenali pola yang kompleks antar parameter kualitas air.

Berdasarkan hasil keseluruhan pengujian pada perbandingan F1 antar model dan split, terlihat bahwa performa antar algoritma cenderung konsisten pada semua skenario evaluasi. Model *ensemble* seperti Random Forest dan XGBoost dapat dikatakan mampu dalam beradaptasi dengan variasi data serta mempertahankan akurasi yang tinggi, sedangkan Naive Bayes cenderung mengalami penurunan kinerja pada kelas dengan distribusi nilai yang berdekatan. Performa yang stabil dari algoritma Random Forest menunjukkan efektivitas pendekatan bootstrap aggregating dalam mengurangi overfitting serta meningkatkan generalisasi. Selain itu, kecepatan pelatihan Naive Bayes menjadi salah satu keunggulan tersendiri jika digunakan pada sistem dengan keterbatasan sumber daya komputasi.

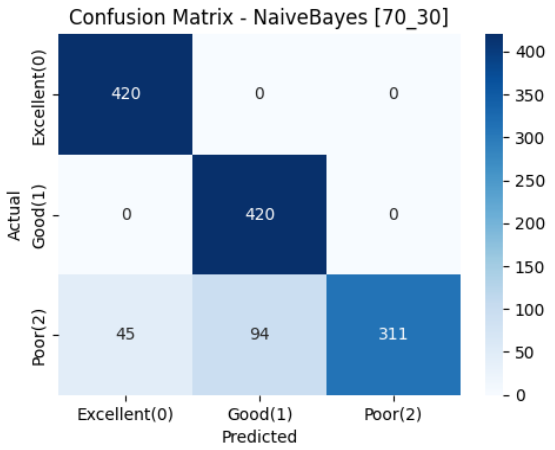


Gambar 3 Grafik perbandingan F1 antar model dan split

3.7 Confusion Matrix 70:30

3.7.1 Naive Bayes

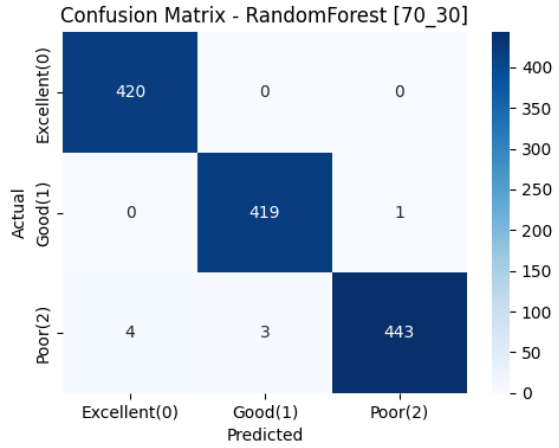
Pada hasil dengan menggunakan metode Naive Bayes, masih terdapat kesalahan klasifikasi pada kelas sedang (*Good*) dan buruk (*poor*), yang menandakan keterbatasan model tersebut dalam menangani data non-linear. Kesalahan tersebut disebabkan oleh metode tersebut yang cenderung mengasumsikan independensi antar fitur, padahal beberapa parameter seperti suhu, pH, dan oksigen memiliki keterkaitan yang cukup kuat. Hal ini menyebabkan model tersebut tidak mampu menangkap hubungan yang kompleks antar variabel sehingga membuat performa menurun pada kelas yang memiliki distribusi nilai yang saling beririsan.



Gambar 4 Grafik confusion matrix Naive Bayes 70 30

3.7.2 Random Forest

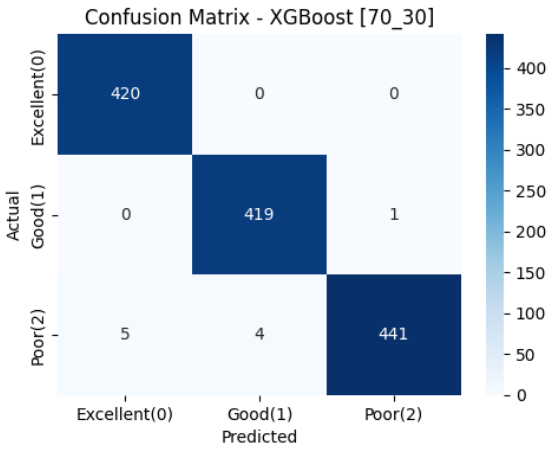
Metode algoritma Random Forest menunjukkan hampir seluruh prediksi berada pada diagonal utama, dimana tingkat akurasi tinggi dan mampu membedakan kelas dengan baik. Hasil pada Gambar 5 menunjukkan bahwa model tersebut mampu mempelajari pola dari data dengan sangat baik karena proses *ensemble* dari banyaknya decission tree membuat prediksi menjadi lebih stabil sehingga dapat mengurangi risiko overfitting. Model Random Forest ini juga efektif dalam menangani interaksi antar fitur yang bersifat non-linear, sehingga memberikan klasifikasi yang lebih presisi.



Gambar 5 Grafik confusion matrix Random Forest 70 30

3.7.3 XGBoost

Bagian algoritma XGBoost memiliki hasil serupa dengan Random Forest, namun terdapat sedikit perbedaan pada kelas minoritas yang menandakan sensitivitas model terhadap variasi kecil pada data. Hal tersebut disebabkan oleh mekanisme boosting yang menekankan pada perbaikan kesalahan dari model sebelumnya, sehingga model ini cenderung lebih reaktif terhadap data dengan jumlah sampel yang tidak seimbang. Meskipun demikian, XGBoost memiliki performa yang sangat baik dan mendekati Random Forest dalam hal akurasi dan kecepatan klasifikasi.

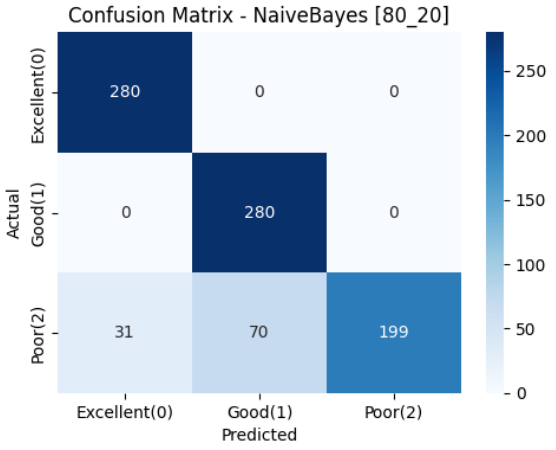


Gambar 6 Grafik confusion matrix XGBoost 70 30

3.8 Confusion Matrix 80:20

3.8.1 Naive Bayes

Metode ini memiliki pola kesalahan yang seupa dengan perbandingan 70:30, dimana peningkatan data latih belum secara signifikan memperbaiki hasil model. Kondisi tersebut dapat menjadi indikasi bahwa keterbatasan pada Naive Bayes tidak terletak pada jumlah data latih, melainkan pada asumsi distribusi data yang digunakan, sehingga model ini cenderung sulit untuk beradaptasi dengan karakteristik data yang kompleks, terutama pada saat adanya korelasi antar parameter mutu air yang kuat.

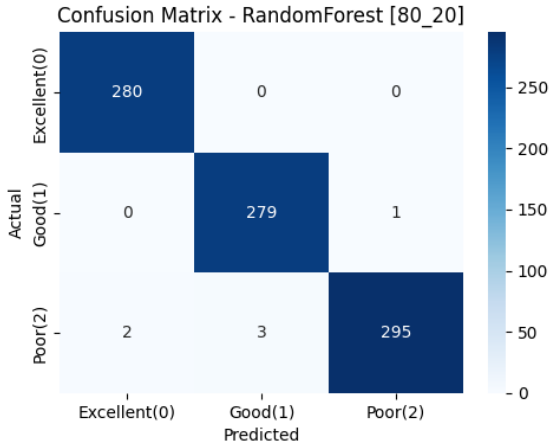


Gambar 7 Grafik confusion matrix Naive Bayes 80 20

3.8.2 Random Forest

Pada algoritma Random Forest dengan perbandingan 80% data latih dan 20% data uji, menunjukkan bahwa metode tersebut menghasilkan klasifikasi secara konsisten yang hampir sempurna dalam memperkuat stabilitas model ensemble ini. Konsistensi

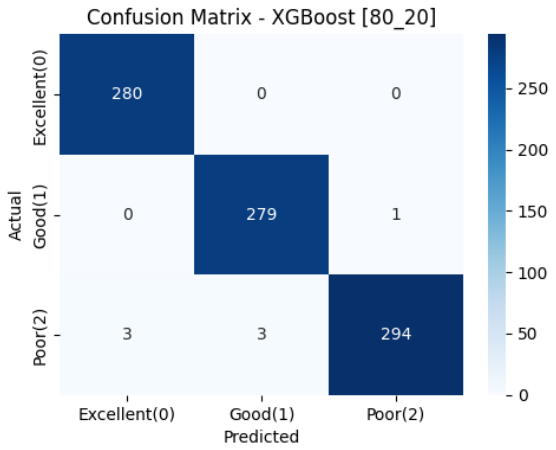
ini menegaskan bahwa peningkatan jumlah data latih tidak menimbulkan perubahan secara signifikan dikarenakan model telah memiliki kemampuan generalisasi yang tinggi. Disamping itu, Random Forest tetap mampu meminimalkan kesalahan klasifikasi dengan baik meskipun terdapat variasi dalam pembagian data.



Gambar 8 Grafik confusion matrix Random Forest 80 20

3.8.3 XGBoost

Percobaan terakhir dengan menggunakan metode algoritma XGBoost pada perbandingan 80:20, menunjukkan hasil yang mendekati metode Random Forest namun terdapat sedikit kesalahan minor yang memperlihatkan kemampuan adaptasi yang baik terhadap variasi data. Performa yang tetap tinggi menunjukkan XGBoost mampu melakukan pembelajaran bertahap secara efisien, dimana setiap iterasi perbaikan kesalahan dapat memperkuat kemampuan model dalam mengenali pola data. Hal ini membuktikan bahwa XGBoost merupakan model algoritma yang efektif untuk digunakan pada data yang memiliki variasi tinggi seperti parameter kualitas air akuakultur.



Gambar 9 Grafik confusion matrix XGBoost 80 20

Secara keseluruhan, hasil percobaan diatas menunjukkan bahwa masing-masing metode algoritma memiliki karakteristik yang saling melengkapi. Pada Random Forest memiliki keunggulan dalam akurasi dan kestabilan, lalu XGBoost unggul dalam memberikan keseimbangan antara efisiensi dan performa, kemudian Naive Bayes tetap relevan untuk kasus dengan kebutuhan prediksi cepat dan data yang sederhana. Hasik ini memperlihatkan bahwa pemilihan algoritma sebaiknya mempertimbangkan kebutuhan praktis sitem pemantauan mutu air, baik dari segi ketepatan dalam klasifikasi maupun efisiensi sumber daya.

4. Kesimpulan

1. Pengujian yang dilakukan menggunakan metode K-Fold Validation dengan pembagian data 70:30 dan 80:20 menunjukkan bahwa algoritma Random Forest memiliki performa terbaik di antara algoritma lain yang diuji. Algoritma ini consistently menghasilkan nilai akurasi, presisi, recall, dan F1-score tertinggi pada seluruh skenario pengujian, yang menunjukkan kemampuannya dalam melakukan generalisasi dan klasifikasi mutu air secara andal. Keunggulan ini menunjukkan bahwa Random Forest mampu menangani kompleksitas data dan variasi antar fitur dengan tingkat stabilitas yang tinggi. Hal ini menjadi indikasi bahwa algoritma ini sangat sesuai digunakan untuk aplikasi pemantauan kualitas air yang membutuhkan prediksi yang konsisten dan handal, terutama dalam kondisi data yang bervariasi. Selain itu, sifat Random Forest yang menggunakan ensemble learning dari banyak decision tree membuat model ini mampu mengurangi risiko overfitting, sehingga prediksi yang dihasilkan lebih robust terhadap perubahan distribusi data.
2. Di posisi kedua, algoritma XGBoost menempati peringkat yang cukup tinggi dengan performa yang hampir setara dengan Random Forest. Model ini mampu menghasilkan akurasi yang tinggi dan performa yang stabil, meskipun menunjukkan sedikit sensitivitas terhadap variasi data, terutama pada kondisi ketidakseimbangan kelas. Misalnya, kelas minoritas pada dataset mutu air cenderung memiliki prediksi yang sedikit lebih rendah dibanding kelas mayoritas. Namun, keunggulan XGBoost terletak pada kemampuannya dalam menangani data yang kompleks melalui mekanisme gradient boosting, regularisasi, dan teknik tree pruning, sehingga model ini tetap efektif dalam menangkap pola non-linear dan interaksi antar fitur. Dengan kata lain, meskipun XGBoost sedikit lebih sensitif terhadap variasi distribusi data dibanding Random Forest, algoritma ini tetap memberikan performa prediksi yang baik dan bisa menjadi alternatif yang kuat ketika Random Forest tidak dapat diterapkan secara optimal.
3. Sementara itu, algoritma Naive Bayes menunjukkan performa terendah di antara algoritma yang diuji. Hal ini terutama disebabkan oleh keterbatasannya dalam asumsi independensi antar fitur, yang dalam kenyataannya jarang terpenuhi dalam data mutu air. Parameter-parameter kualitas air seperti suhu, pH,

oksigen terlarut, dan kekeruhan memiliki hubungan non-linear dan interaksi yang kompleks, sehingga Naive Bayes kurang mampu menangkap pola tersebut dengan akurat. Akibatnya, prediksi yang dihasilkan cenderung lebih kasar dan memiliki tingkat kesalahan yang lebih tinggi dibandingkan algoritma ensemble. Meski demikian, Naive Bayes tetap memiliki keunggulan dalam hal kecepatan komputasi dan kesederhanaan model, sehingga masih dapat digunakan pada kasus dengan keterbatasan sumber daya atau ketika interpretabilitas model lebih diutamakan daripada performa prediksi.

4. Secara keseluruhan, hasil pengujian menegaskan bahwa metode ensemble learning, khususnya Random Forest dan XGBoost, memiliki keunggulan yang signifikan dibandingkan pendekatan statistik sederhana seperti Naive Bayes. Keunggulan tersebut meliputi stabilitas prediksi, akurasi tinggi, dan kemampuan menangani data yang kompleks atau variatif. Hal ini menunjukkan bahwa algoritma *ensemble* lebih adaptif dalam menghadapi ketidakseimbangan data, outlier, serta interaksi non-linear antar fitur, yang merupakan karakteristik umum pada dataset mutu air.
5. Sebagai arah pengembangan penelitian selanjutnya, beberapa strategi dapat diterapkan untuk meningkatkan efektivitas klasifikasi mutu air. Pertama, penerapan model ensemble dengan sistem pemantauan kualitas air berbasis *Internet of Things* (IoT) secara real-time akan memungkinkan hasil klasifikasi dapat langsung diterapkan di lapangan, sehingga pengambilan keputusan berbasis data dapat dilakukan lebih cepat dan akurat. Kedua, optimasi hyperparameter pada algoritma ensemble dapat meningkatkan performa model secara signifikan, termasuk dalam menangani kelas minoritas dan variasi data yang tidak seimbang. Ketiga, integrasi algoritma ensemble dengan deep learning dapat membuka peluang pengembangan model hybrid yang mampu menangkap pola non-linear lebih kompleks dan memanfaatkan informasi temporal jika data kualitas air dikumpulkan secara berurutan. Terakhir, penelitian lanjutan juga dapat mempertimbangkan penerapan data augmentation, teknik feature selection, dan strategi ensemble stacking untuk meningkatkan generalisasi model, sehingga sistem klasifikasi mutu air menjadi lebih andal, adaptif, dan siap diterapkan pada skala besar di lapangan.
6. Dengan demikian, kesimpulan ini tidak hanya menekankan superioritas Random Forest sebagai algoritma dengan performa terbaik, tetapi juga menyoroti potensi algoritma ensemble secara umum dan arahan strategis untuk pengembangan penelitian yang lebih lanjut di bidang klasifikasi kualitas air.

REFERENSI

[1] T. Li, J. Lu, J. Wu, Z. Zhang, and L. Chen, "Predicting Aquaculture Water Quality Using Machine Learning Approaches," *Water (Switzerland)*, vol. 14, no. 18, Sep. 2022, doi: 10.3390/w14182836.

[2] M. Miller, A. Kisiel, D. Cembrowska-Lech, I. Durlik, and T. Miller, "IoT in Water Quality Monitoring—Are We Really Here?," *Sensors*, vol. 23, no. 2, Jan. 2023, doi: 10.3390/s23020960.

[3] S. Mamatha, P. Aakash, and N. Sravya, "Water Quality Analysis and Prediction Using Random Forest and Naive Bayes."

[4] P. Jayaraman, K. K. Nagarajan, P. Partheeban, and V. Krishnamurthy, "Critical review on water quality analysis using IoT and machine learning models," *International Journal of Information Management Data Insights*, vol. 4, no. 1, Apr. 2024, doi: 10.1016/j.jjimei.2023.100210.

[5] M. Y. Shams, A. M. Elshewey, E. S. M. El-kenawy, A. Ibrahim, F. M. Talaat, and Z. Tarek, "Water quality prediction using machine learning models based on grid search method," *Multimed Tools Appl*, vol. 83, no. 12, pp. 35307–35334, Apr. 2024, doi: 10.1007/s11042-023-16737-4.

[6] W. Chen, D. Xu, B. Pan, Y. Zhao, and Y. Song, "Machine Learning-Based Water Quality Classification Assessment," *Water (Switzerland)*, vol. 16, no. 20, Oct. 2024, doi: 10.3390/w16202951.

[7] N. Radhakrishnan and A. S. Pillai, "Comparison of Water Quality Classification Models using Machine Learning," Institute of Electrical and Electronics Engineers (IEEE), Jul. 2020, pp. 1183–1188. doi: 10.1109/icces48766.2020.9137903.

[8] M. S. Islam Khan, N. Islam, J. Uddin, S. Islam, and M. K. Nasir, "Water quality prediction and classification based on principal component regression and gradient boosting classifier approach," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 4773–4781, Sep. 2022, doi: 10.1016/j.jksuci.2021.06.003.

[9] M. H. Al-Adhaileh and F. W. Alsaade, "Modelling and prediction of water quality by using artificial intelligence," *Sustainability (Switzerland)*, vol. 13, no. 8, Apr. 2021, doi: 10.3390/su13084259.

[10] D. Danuri and M. Mohd Pozi, "Machine Learning Approaches for Fish Pond Water Quality Classification: Random Forest, Gaussian Naive Bayes, and Decision Tree Comparison," European Alliance for Innovation n.o., Feb. 2024. doi: 10.4108/eai.21-9-2023.2342964.

[11] Zhang, X., Li, Y., Chen, Z., & Wu, H. (2024). Monitoring water quality parameters of freshwater aquaculture ponds using UAV-based multispectral images. *Ecological Indicators, 162*, 111234. https://doi.org/10.1016/j.ecolind.2024.111234

[12] X. Liu *et al.*, "Monitoring water quality parameters of freshwater aquaculture ponds using UAV-based multispectral images," *Ecol Indic*, vol. 167, Oct. 2024, doi: 10.1016/j.ecolind.2024.112644.

[13] A. Elmotawakkil, N. Enneya, S. K. Bhagat, M. M. Ouda, and V. Kumar, "Advanced machine learning models for robust prediction of water quality index and classification," *Journal of Hydroinformatics*, vol. 27, no. 2, pp. 299–319, Feb. 2025, doi: 10.2166/hydro.2025.290.

[14] R. R S and E. Antony, "Water Quality Analysis of Different Water Sources in Kerala, India," *International Journal of Agriculture & Environmental Science*, vol. 9, no. 4, pp. 6–11, Aug. 2022, doi: 10.14445/23942568/ijaes-v9i4p102.

[15] Dewi, D. A., Wei, A. S., Lin, L. C., & Heng, C. D. (2024). Water quality prediction using Random Forest algorithm and optimization. *Journal of Data Science and Environmental Informatics*.

[16] "Aquaculture - Water Quality Dataset." Accessed: Oct. 08, 2025. [Online]. Available: https://data.mendeley.com/datasets/y78ty2g293/1

[17] X. Yan, T. Zhang, W. Du, Q. Meng, X. Xu, and X. Zhao, "A Comprehensive Review of Machine Learning for Water Quality Prediction over the Past Five Years," Jan. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/jmse12010159.

[18] P. Patel, "Machine Learning Applications in Agriculture: A Software Engineering Perspective," Jul. 13, 2025. doi: 10.31224/4824.

[19] D. B. Olawade, O. Z. Wada, A. O. Ige, B. I. Egbewole, A. Olojo, and B. I. Oladapo, "Artificial intelligence in environmental monitoring: Advancements, challenges, and future directions," Dec. 01, 2024, *Elsevier B.V.* doi: 10.1016/j.heha.2024.100114.

[20] M. Jemeļjanova, A. Kmoch, and E. Uuemaa, "Adapting machine learning for environmental spatial data - A review," Jul. 01, 2024, *Elsevier B.V.* doi: 10.1016/j.ecoinf.2024.102634.

[21] Z. Zhou, C. Qiu, and Y. Zhang, "A comparative analysis of linear regression, neural networks and random forest regression for predicting air ozone employing soft sensor models," *Sci Rep*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-49899-0.

[22] M. A. Pienaar and K. Naidoo, "Classification and predictive models using supervised machine learning: A conceptual review," *Southern African Journal of Critical Care*, p. e2937, May 2025, doi: 10.7196/sajcc.2025.v411.2937.

[23] Chen, W., Xu, D., Pan, B., Zhao, Y., & Song, Y. (2024). Machine learning-based water quality classification assessment. *Environmental Informatics Letters, 5*(1), 23–32.*

[24] D. Berrar, "Bayes' theorem and naive bayes classifier," in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, vol. 1–3, Elsevier, 2018, pp. 403–412. doi: 10.1016/B978-0-12-809633-8.20473-1.

[25] A. A. Khan, O. Chaudhari, and R. Chandra, "A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation," Jun. 15, 2024, *Elsevier Ltd*. doi: 10.1016/j.eswa.2023.122778.

[26] M. Chen, Q. Liu, S. Chen, Y. Liu, C. H. Zhang, and R. Liu, "XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System," *IEEE Access*, vol. 7, pp. 13149–13158, 2019, doi: 10.1109/ACCESS.2019.2893448.

[27] I. Domor Mienye and Y. Sun, "Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000. A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects", doi: 10.1109/ACCESS.2017.DOI.

[28] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," Oct. 01, 2022, *Elsevier Ltd*. doi: 10.1016/j.engappai.2022.105151.

Gavriel Joseph Lim, mahasiswa S1, program studi Teknik Informatika Universitas Tarumanagara

Jason Chainara Putra, mahasiswa S1, program studi Teknik Informatika Universitas Tarumanagara

Paulina Agusia, mahasiswa S1, program studi Teknik Informatika Universitas Tarumanagara

Marcia Yanprincessa Utama, mahasiswa S1, program studi Teknik Informatika Universitas Tarumanagara