Jurnal Pendidikan Perikanan Kelautan
Vol. 6, No. 1, Juni 2026, Hal. 55-68

Analysis of the Performance of the XGBoost Algorithm in the Feasibility Classification of Vaname Shrimp Pond Water Quality Using Data Before and After Scaling

Ittaqi Tafuzi*[1], Anisa Putri Wibowo[2], Nia Afrilia[3]
[1,2,3]Universitas Pendidikan Indonesia Kampus Serang; Jl. Ciracas No. 38, Serang, Banten 42116, Indonesia.
[1,2,3]Program Studi Sistem Informasi Kelautan, Banten.
e-mail: *[1]Ittaqitafuzi@upi.edu, [2]Anisaputriwibowo@upi.edu, [3]Niaafrilia@upi.edu

ABSTRACT

Water quality is a critical factor influencing the success of Pacific white shrimp (*Litopenaeus vannamei*) aquaculture. Fluctuations in water quality parameters such as temperature, pH, salinity, dissolved oxygen (DO), and ammonia can significantly affect the growth, health, and survival rate of the shrimp. With the increasing volume of water quality monitoring data, an efficient method is required to classify aquatic conditions rapidly and objectively. This study aims to evaluate the performance of the *Extreme Gradient Boosting* (XGBoost) algorithm in classifying the water quality of vaname shrimp ponds, utilizing data profiles before and after the *scaling* process. The dataset was sourced from Figshare and subsequently underwent Exploratory Data Analysis (EDA), missing value imputation, class labeling, and data splitting into training and testing sets. Experiments were conducted under two scenarios: without data *scaling* and with the implementation of StandardScaler. To provide a comprehensive evaluation of model performance, the assessment focused not only on accuracy but also incorporated log loss and *overfitting* gap analysis. The results demonstrated that the model without data *scaling* achieved an absolute accuracy of 100%, accompanied by a minimal *log loss* of 0.014 and an *overfitting* gap of 0%. Conversely, the model utilizing feature *scaling* experienced a significant performance degradation, yielding an accuracy of only 38.24% and a *log loss* reaching 1.239. Furthermore, *feature importance* analysis revealed that ammonia contributed the highest impact at 44%, followed by pH at 40%, temperature at 7.8%, salinity at 4.3%, and dissolved oxygen at 3.64%.

Keywords: *XGBoost, water quality, vannamei shrimp pond, feature importance, scaling*

ABSTRAK

Kualitas air adalah faktor penting yang memengaruhi keberhasilan dalam budidaya udang vaname (*Litopenaeus vannamei*). Perubahan dalam kualitas air, seperti suhu, pH, salinitas, oksigen terlarut (DO), dan amonia, bisa memengaruhi pertumbuhan, kesehatan, serta tingkat kelangsungan hidup udang. Dengan semakin banyaknya data hasil pemantauan kualitas air, diperlukan cara yang bisa mengklasifikasikan kondisi perairan secara cepat dan tidak memihak. Penelitian ini bertujuan untuk mempelajari bagaimana algoritma *Extreme Gradient Boosting* (XGBoost) bekerja dalam mengklasifikasikan kualitas air tambak udang vaname, dengan menggunakan data yang telah diubah sebelum dan setelah dilakukan proses *scaling*. Dataset yang digunakan berasal dari Figshare, kemudian melalui proses analisis data secara eksploratif (EDA), mengatasi nilai-nilai yang hilang, memberi label pada kelas, serta membagi data menjadi bagian latihan dan bagian uji. Uji dilakukan dalam dua skenario, yaitu tanpa melakukan

scaling dan dengan menggunakan *StandardScaler*. Evaluasi tidak hanya melihat akurasi saja, tetapi juga memperhatikan *log loss* dan selisih *overfitting* agar menggambarkan kinerja model secara lebih lengkap. Penelitian menunjukkan bahwa model tanpa penyesuaian skala memberikan akurasi 100% dengan *log loss* yang sangat kecil yaitu 0,014 dan gap *overfitting* sebesar 0%. Sebaliknya, model yang menggunakan penyesuaian skala mengalami penurunan kinerja dengan akurasi hanya 38,24% dan *log loss* mencapai 1,239. Analisis pentingnya fitur menunjukkan bahwa parameter amonia memiliki kontribusi terbesar sebesar 44%, disusul oleh pH dengan 40%, suhu sebesar 7,8%, salinitas 4,3%, dan dissolved oxygen sebesar 3,64%.

***Kata kunci** : XGBoost, kualitas air, tambak udang vaname, feature importance, scaling*

PENDAHULUAN

Budidaya udang vaname (Litopenaeus vannamei) merupakan salah satu sektor akuakultur yang memiliki nilai ekonomi tinggi dan berkontribusi terhadap peningkatan produksi perikanan budidaya di Indonesia. Tingginya permintaan pasar mendorong penerapan sistem budidaya yang semakin intensif untuk meningkatkan produktivitas tambak. Namun, sistem budidaya yang intensif juga meningkatkan risiko terjadinya perubahan kualitas lingkungan perairan yang dapat memengaruhi pertumbuhan dan kelangsungan hidup udang. Oleh karena itu, kualitas air menjadi salah satu faktor penting yang harus diperhatikan dalam keberhasilan budidaya udang vaname (Ritonga et al., 2021).

Kualitas air berperan langsung terhadap kondisi fisiologis, kesehatan, serta produktivitas udang selama masa pemeliharaan. Beberapa parameter yang umum digunakan dalam pemantauan kualitas air tambak meliputi suhu, derajat keasaman (pH), salinitas, *dissolved oxygen* (DO), dan amonia. Ketidaksesuaian nilai parameter tersebut dengan kisaran optimal dapat menyebabkan stres, menghambat pertumbuhan, menurunkan daya tahan tubuh, hingga meningkatkan risiko serangan penyakit pada udang. Oleh karena itu, pemantauan kualitas air secara berkala diperlukan untuk menjaga kondisi lingkungan budidaya tetap stabil dan mendukung hasil produksi yang optimal (Bambang et al., 2024; Ritonga et al., 2021).

Perkembangan teknologi monitoring dan pencatatan data menghasilkan informasi kualitas air dalam jumlah yang semakin besar. Data tersebut memiliki potensi untuk dimanfaatkan sebagai dasar pengambilan keputusan dalam pengelolaan tambak secara lebih cepat dan objektif. Akan tetapi, proses analisis secara manual sering kali membutuhkan waktu yang relatif lama serta berpotensi menghasilkan penilaian yang kurang konsisten, terutama ketika jumlah data yang tersedia terus meningkat. Selain itu, hubungan antarparameter kualitas air tidak selalu bersifat sederhana sehingga diperlukan pendekatan yang mampu mengidentifikasi pola yang terdapat dalam data secara lebih efektif (Singh et al., 2025).

Selain jumlah data yang besar, data kualitas air yang diperoleh dari berbagai sumber pengamatan sering kali memiliki nilai yang tidak lengkap (*missing value*). Kondisi tersebut dapat terjadi akibat kesalahan pencatatan, keterbatasan alat ukur, maupun perbedaan metode pengumpulan data. Keberadaan *missing value* dapat memengaruhi kualitas model yang dibangun karena sebagian algoritma *machine learning* memerlukan data yang lengkap pada setiap atribut. Oleh sebab itu, diperlukan proses penanganan *missing value* sebelum data digunakan dalam proses klasifikasi agar informasi yang tersedia tetap dapat dimanfaatkan secara optimal dan menghasilkan model yang lebih representatif.

Perkembangan teknologi informasi mendorong pemanfaatan teknik data mining dan machine learning dalam berbagai bidang, termasuk sektor akuakultur. Machine learning memungkinkan sistem komputer mempelajari pola dari data historis untuk menghasilkan model yang mampu melakukan prediksi maupun klasifikasi terhadap data baru. Dalam bidang kualitas air, pendekatan machine learning semakin banyak digunakan karena mampu mengidentifikasi hubungan kompleks antarparameter lingkungan dan meningkatkan akurasi klasifikasi maupun prediksi kualitas perairan (Ab Karim et al., 2026). Dalam konteks kualitas air, penerapan *machine learning* dapat membantu mengidentifikasi kondisi perairan berdasarkan kombinasi berbagai parameter yang diamati sehingga proses evaluasi dapat dilakukan secara lebih cepat dan objektif dibandingkan pendekatan konvensional (Babshette & S, 2025; Singh et al., 2025)

Salah satu algoritma machine learning yang banyak digunakan dalam permasalahan klasifikasi adalah *Extreme Gradient Boosting* (XGBoost). Algoritma ini banyak diterapkan pada studi kualitas air karena mampu menangani data tabular, menghasilkan akurasi tinggi, dan menyediakan interpretasi melalui *feature importance* maupun explainable AI (Alnemari et al., 2025; Nallakaruppan et al., 2024). Algoritma ini merupakan pengembangan dari metode *gradient boosting* yang memanfaatkan kombinasi beberapa pohon keputusan untuk menghasilkan model dengan performa yang tinggi. XGBoost memiliki berbagai keunggulan, seperti kemampuan menangani data tabular secara efektif, efisiensi komputasi yang baik, serta mekanisme regularisasi yang dapat membantu mengurangi risiko *overfitting*. Selain itu, XGBoost juga mampu menunjukkan tingkat kontribusi masing-masing variabel terhadap hasil klasifikasi melalui analisis *feature importance* (Babshette & S, 2025; Naim et al., 2025).

Penentuan tingkat kelayakan kualitas air tambak merupakan salah satu aspek penting dalam pengelolaan budidaya udang vaname. Informasi mengenai kondisi air yang layak, cukup layak, atau tidak layak dapat membantu pembudidaya dalam mengambil tindakan yang tepat untuk menjaga stabilitas lingkungan tambak. Dengan adanya sistem klasifikasi berbasis data, proses penilaian kondisi air dapat dilakukan secara lebih konsisten dan objektif dibandingkan metode yang hanya mengandalkan pengamatan manual.

Sebelum proses pemodelan dilakukan, data umumnya melalui tahapan *preprocessing* untuk meningkatkan kualitas data yang akan digunakan dalam proses pelatihan model. Salah satu teknik yang umum digunakan adalah *scaling*, yaitu proses transformasi nilai atribut ke dalam rentang tertentu agar memiliki skala yang lebih seragam. Meskipun algoritma berbasis pohon keputusan seperti XGBoost dikenal relatif tidak sensitif terhadap perbedaan skala data, pengaruh penggunaan *scaling* terhadap performa model masih perlu dianalisis pada berbagai kasus, termasuk klasifikasi kualitas air tambak udang vaname. Pengujian ini diperlukan untuk mengetahui sejauh mana proses *scaling* memengaruhi kemampuan model dalam membedakan kategori kelayakan kualitas air (Babshette & S, 2025; Naim et al., 2025)

Sejumlah penelitian telah memanfaatkan algoritma *machine learning* untuk memprediksi maupun mengklasifikasikan kondisi kualitas air. Namun sebagian besar penelitian berfokus pada peningkatan akurasi model dan integrasi sistem monitoring, sementara evaluasi pengaruh *preprocessing* khususnya *scaling* terhadap performa XGBoost masih relatif terbatas (Baena-Navarro et al., 2025; Nuangpirom et al., 2025). Padahal, karakteristik data kualitas air yang memiliki rentang nilai berbeda serta keberadaan *missing value* berpotensi memengaruhi proses pembentukan model. Oleh karena itu, penelitian mengenai pengaruh *scaling* terhadap performa XGBoost menjadi penting untuk dilakukan. Selain itu, sebagian besar penelitian

terkini lebih berfokus pada peningkatan akurasi model dan interpretasi hasil prediksi, sementara kajian mengenai dampak *preprocessing* terhadap stabilitas performa XGBoost pada data kualitas air masih relatif terbatas (Ab Karim et al., 2026; Nallakaruppan et al., 2024).

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk menganalisis performa algoritma XGBoost dalam mengklasifikasikan kelayakan kualitas air tambak udang vaname menggunakan data sebelum dan sesudah proses *scaling*. Klasifikasi dilakukan berdasarkan parameter suhu, pH, salinitas, *dissolved oxygen* (DO), dan amonia yang dikelompokkan ke dalam kategori layak, cukup layak, dan tidak layak. Performa model dievaluasi menggunakan metrik *accuracy*, *log loss*, dan *overfitting gap* untuk memperoleh gambaran performa yang lebih komprehensif, tidak hanya dari sisi ketepatan prediksi tetapi juga dari sisi kalibrasi probabilistik dan risiko *overfitting*. Selain itu, penelitian ini juga bertujuan untuk mengidentifikasi parameter kualitas air yang memiliki pengaruh terbesar terhadap hasil klasifikasi melalui analisis *feature importance*. Hasil penelitian ini diharapkan dapat memberikan gambaran komparatif mengenai pengaruh proses *scaling* terhadap stabilitas kinerja algoritma XGBoost, sekaligus memvalidasi akurasi model secara objektif.

METODE PENELITIAN

2.1 Studi Literatur dan Landasan Penelitian

2.1.1 Budidaya Udang Vaname

Udang vaname (*Litopenaeus vannamei*) merupakan salah satu komoditas budidaya yang banyak dikembangkan karena memiliki pertumbuhan yang relatif cepat, tingkat adaptasi yang baik terhadap lingkungan, serta nilai ekonomi yang tinggi. Permintaan pasar yang terus meningkat menjadikan budidaya udang vaname sebagai salah satu sektor penting dalam kegiatan akuakultur. Namun, keberhasilan budidaya tidak hanya ditentukan oleh kualitas benih dan manajemen pakan, tetapi juga dipengaruhi oleh kondisi lingkungan tambak, khususnya kualitas air yang menjadi media utama kehidupan udang (Ananta et al., 2024).

Kualitas air yang tidak sesuai dengan kebutuhan biologis udang dapat menyebabkan gangguan pertumbuhan, menurunkan daya tahan tubuh, dan meningkatkan risiko serangan penyakit. Oleh karena itu, pemantauan kualitas air secara berkala diperlukan untuk menjaga kondisi tambak tetap stabil dan mendukung produktivitas budidaya. Seiring perkembangan teknologi, pemanfaatan metode berbasis *machine learning* mulai diterapkan untuk membantu proses pemantauan dan pengambilan keputusan dalam kegiatan budidaya secara lebih cepat dan objektif (Ananta et al., 2024; Babshette & S, 2025).

2.1.2 Parameter Kualitas Air Tambak

Dalam budidaya udang vaname, kualitas air ditentukan oleh sejumlah parameter fisika dan kimia yang memengaruhi kondisi lingkungan tambak. Parameter yang digunakan dalam penelitian ini meliputi suhu, derajat keasaman (*potential hydrogen* atau pH), salinitas, *dissolved oxygen* (DO), dan amonia. Kelima parameter tersebut dipilih karena memiliki keterkaitan langsung dengan pertumbuhan, kesehatan, serta tingkat kelangsungan hidup udang vaname (Ananta et al., 2024; Himawan et al., 2025). Suhu merupakan parameter yang mempengaruhi laju metabolisme dan aktivitas fisiologis udang. Nilai suhu yang berada di luar kisaran optimal dapat menyebabkan gangguan pertumbuhan dan menurunkan produktivitas budidaya. pH

menunjukkan tingkat keasaman atau kebasaan air tambak yang berpengaruh terhadap proses fisiologis dan keseimbangan lingkungan perairan. Salinitas menggambarkan konsentrasi garam terlarut yang berperan dalam proses osmoregulasi udang. Sementara itu, DO merupakan indikator ketersediaan oksigen yang dibutuhkan dalam proses respirasi. Adapun amonia merupakan senyawa hasil sisa metabolisme yang bersifat toksik pada konsentrasi tertentu sehingga perlu dikendalikan agar tidak membahayakan udang (Ananta et al., 2024; Himawan et al., 2025).

Tabel 2.1 Standar Parameter Kualitas Air Tambak Udang Vaname

Parameter	Layak (Baik)	Cukup Layak (Sedang)	Tidak Layak (Buruk)
pH	7,5 - 8,5	7,0 - 7,4 atau 8,6 - 9,0	< 7 atau > 9
Suhu (°C)	28 - 32	26 - 27	< 26 atau > 32
Salinitas (ppt)	15 - 30	10 - 14	< 10 atau > 35
DO (mg/L)	> 4	3 - 4	< 3
Amonia (mg/L)	< 0,1	0,1 - 0,5	> 0,5

Suhu merupakan parameter fisika yang menentukan laju metabolisme udang. Suhu optimal untuk pertumbuhan udang vaname berkisar antara 26--30°C. Di luar kisaran ini, proses fisiologis udang akan terganggu secara signifikan (Ananta et al., 2024).

pH mencerminkan derajat keasaman air tambak. Nilai pH yang ideal berada pada kisaran 7,5--8,5. Nilai yang terlalu rendah atau terlalu tinggi dapat mengganggu sistem pernapasan dan osmoregulasi udang (Himawan et al., 2025).

Salinitas mengacu pada kadar garam terlarut dalam air. Udang vaname bersifat euryhaline sehingga dapat beradaptasi pada berbagai tingkat salinitas, namun pertumbuhan optimal tercapai pada kisaran 15--25 ppt (Ananta et al., 2024; Himawan et al., 2025).

Dissolved oxygen (DO) atau oksigen terlarut merupakan parameter vital untuk respirasi aerobik udang. Kadar DO yang direkomendasikan adalah lebih dari 4 mg/L. Penurunan DO di bawah nilai minimum dapat menyebabkan stres hingga kematian massal (Ananta et al., 2024).

Amonia (NH_3) merupakan senyawa hasil metabolisme nitrogen yang bersifat toksik bagi udang apabila konsentrasinya melebihi batas yang direkomendasikan. Peningkatan kadar amonia dapat mengganggu pertumbuhan, menurunkan daya tahan tubuh, serta meningkatkan risiko kematian pada udang vaname. Oleh karena itu, parameter amonia menjadi salah satu indikator penting dalam pemantauan kualitas air tambak (Ananta et al., 2024; Himawan et al., 2025).

2.2 Klasifikasi Kelayakan Kualitas Air

Klasifikasi kelayakan kualitas air merupakan pendekatan untuk mengkategorikan kondisi perairan tambak berdasarkan kombinasi nilai parameter ke dalam kelas-kelas tertentu. Dalam penelitian ini, kelayakan kualitas air diklasifikasikan ke dalam tiga kelas, yaitu: (1) Layak, (2) Cukup Layak, dan (3) Tidak Layak, berdasarkan kesesuaian nilai parameter dengan standar budidaya udang vaname (Ananta et al., 2024).

Penerapan *machine learning* untuk klasifikasi kualitas air telah dibuktikan efektif dalam berbagai studi. Salah satunya adalah penelitian pemantauan kelayakan air rumah tangga berbasis IoT dan XGBoost, di mana model mampu mengklasifikasikan kondisi air secara akurat berdasarkan parameter pH, turbidity, dan konduktivitas yang dikumpulkan secara *real-*

time (Ananta et al., 2024). Pendekatan serupa relevan diterapkan pada konteks tambak udang vaname dengan parameter yang disesuaikan.

2.3 Algoritma XGBoost (*eXtreme Gradient Boosting*)

eXtreme Gradient Boosting (XGBoost) merupakan algoritma *machine learning* berbasis ensemble yang menggunakan pendekatan *gradient boosting* dengan optimasi bertahap. Algoritma ini membangun serangkaian pohon keputusan (*decision tree*) secara sekuensial, di mana setiap pohon baru berusaha memperbaiki kesalahan prediksi pohon sebelumnya (Babshette & S, 2025; Himawan et al., 2025).

XGBoost memiliki beberapa keunggulan yang menjadikannya pilihan populer dalam tugas klasifikasi, di antaranya:

(1) Efisiensi komputasi yang tinggi berkat implementasi paralel dan optimasi memori.

(2) Mekanisme regularisasi (L1 dan L2) yang mengurangi risiko *overfitting*.

(3) Kemampuan menangani data tabular secara efektif, termasuk data dengan distribusi yang tidak seragam.

(4) Relatif tidak sensitif terhadap perbedaan skala antar fitur karena berbasis pohon keputusan.

(5) Dukungan analisis *feature importance* yang dapat mengidentifikasi parameter paling berpengaruh (Babshette & S, 2025; Himawan et al., 2025).

Penelitian sebelumnya menunjukkan bahwa XGBoost mampu menghasilkan akurasi yang kompetitif dalam klasifikasi kualitas air dibandingkan algoritma lain seperti SVM, Random Forest, maupun Neural Network (Alnemari et al., 2025; Babshette & S, 2025; L. Li & Wei, 2024; Naim et al., 2025; Singh et al., 2025).

2.4 *Feature Importance*

Feature importance merupakan salah satu mekanisme yang tersedia pada algoritma XGBoost untuk mengukur tingkat kontribusi masing-masing fitur terhadap hasil prediksi model. Dalam penelitian kualitas air modern, analisis *feature importance* sering dikombinasikan dengan pendekatan *Explainable Artificial Intelligence* (XAI) untuk meningkatkan transparansi dan interpretabilitas model machine learning (Ab Karim et al., 2026; Aderemi et al., 2025; Nallakaruppan et al., 2024). Nilai *feature importance* diperoleh berdasarkan frekuensi penggunaan fitur dalam proses pembentukan pohon keputusan serta kontribusinya dalam mengurangi kesalahan prediksi model (Naim et al., 2025).

Pada penelitian ini, analisis *feature importance* digunakan untuk mengetahui tingkat pengaruh masing-masing parameter kualitas air terhadap hasil klasifikasi kelayakan tambak udang vaname. Melalui analisis tersebut dapat diketahui parameter mana yang memiliki kontribusi paling besar dalam proses pengambilan keputusan oleh model sehingga dapat menjadi informasi tambahan dalam pengelolaan kualitas air tambak.

2.5 Data *Scaling* (*StandardScaler*)

Data scaling merupakan salah satu tahapan *preprocessing* yang bertujuan untuk menyamakan skala antar fitur sebelum digunakan dalam proses pelatihan model. Pada penelitian ini, metode *StandardScaler* digunakan untuk melakukan transformasi data berdasarkan nilai rata-rata dan standar deviasi masing-masing fitur. Transformasi tersebut menghasilkan data dengan rata-rata mendekati nol dan standar deviasi mendekati satu.

Secara matematis, proses transformasi menggunakan *StandardScaler* dinyatakan dengan persamaan berikut:

$$z_i = \frac{x_i - \mu}{s}$$

Keterangan :

- z_i = nilai hasil normalisasi,
- x_i = nilai asli data,
- μ = rata-rata data,
- s = standar deviasi data.

Pada penelitian ini, proses *scaling* dilakukan setelah data dibagi menjadi data latih dan data uji untuk menghindari terjadinya *data leakage*. Selanjutnya, performa model XGBoost yang menggunakan data hasil *scaling* dibandingkan dengan model yang menggunakan data asli. Perbandingan tersebut dilakukan untuk mengetahui pengaruh proses *scaling* terhadap hasil klasifikasi kelayakan kualitas air tambak udang vaname.

2.6 Evaluasi Model

Evaluasi model dilakukan agar mengetahui sejauh mana kemampuan algoritma XGBoost dalam mengklasifikasikan tingkat kelayakan kualitas air. Dalam penelitian ini, penilaian dilakukan dengan menggunakan tiga metrik utama:

(1) Akurasi: proporsi prediksi yang benar terhadap seluruh data pengujian. Meskipun merupakan metrik yang mudah dipahami, akurasi saja tidak cukup untuk menilai kualitas model secara menyeluruh, terutama pada dataset berukuran kecil.

(2) *Overfitting* Gap: selisih antara akurasi data latih (*train accuracy*) dan akurasi data uji (*test accuracy*), dinyatakan dalam persen. Nilai yang mendekati nol menunjukkan model mampu menggeneralisasi dengan baik. Nilai positif besar mengindikasikan *overfitting*, sementara nilai negatif mengindikasikan gangguan dalam proses pembelajaran model.

(3) *Log Loss* (*Cross-Entropy Loss*): mengukur seberapa jauh prediksi probabilistik model dari label sebenarnya. Metrik ini penting untuk menilai kalibrasi model: nilai *log loss* yang rendah menunjukkan model memberikan prediksi probabilistik yang yakin dan tepat, sedangkan nilai yang tinggi menunjukkan model sering salah dengan tingkat keyakinan yang besar. Secara matematis:

$$LogLoss = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

Keterangan Rumus *Log Loss*:

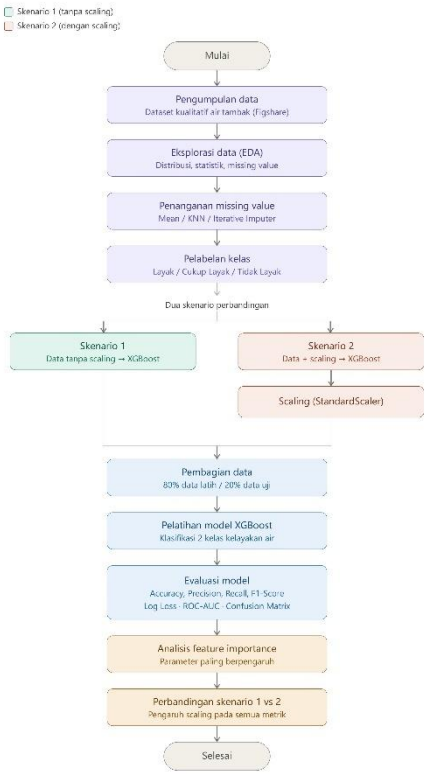
- LogLoss = Nama metrik evaluasi model
- – = Tanda negatif (karena nilai log dari probabilitas selalu negatif)
- 1/N = Rata-rata dari seluruh data
- Σ = Penjumlahan seluruh data dari i=1 sampai N
- y_i = Label sebenarnya (bernilai 0 atau 1)
- log = Logaritma natural
- ŷ_i = Probabilitas prediksi model (nilai antara 0 sampai 1)
- (1 – y_i) = Kebalikan dari label sebenarnya
- (1 – ŷ_i) = Probabilitas prediksi untuk kelas negatif

Penggunaan ketiga metrik ini secara bersamaan memungkinkan evaluasi yang lebih jujur dan komprehensif dibandingkan hanya mengandalkan nilai akurasi, khususnya dalam mendeteksi potensi *overfitting* laten yang tidak terlihat dari angka akurasi saja.

2.7 Alur Penelitian

Penelitian ini dilaksanakan melalui beberapa tahapan sistematis: (1) pengumpulan dataset dari Figshare; (2) EDA dan penanganan *missing value* melalui imputasi; (3) pelabelan kelas berdasarkan standar kualitas air; (4) pembagian data latih (80%) dan data uji (20%); (5) pembangunan model XGBoost pada dua skenario (*tanpa scaling* dan *dengan scaling*); (6) evaluasi menggunakan akurasi, *overfitting gap*, dan *log loss*; serta (7) analisis *feature importance*.

Tahap berikutnya adalah pembangunan model menggunakan algoritma XGBoost melalui dua skenario pengujian. Skenario pertama menggunakan data tanpa proses *scaling*, sedangkan skenario kedua menggunakan data yang telah ditransformasi dengan metode *StandardScaler*. Kedua model kemudian dievaluasi menggunakan nilai accuracy untuk mengetahui performa klasifikasi yang dihasilkan. Selain evaluasi performa model, dilakukan pula analisis *feature importance* untuk mengidentifikasi parameter kualitas air yang memberikan kontribusi terbesar terhadap hasil klasifikasi. Selanjutnya, hasil dari kedua skenario dibandingkan untuk mengetahui pengaruh proses *scaling* terhadap performa algoritma XGBoost pada klasifikasi kelayakan kualitas air tambak udang vaname. Alur penelitian secara lengkap ditunjukkan pada **Gambar 2.1**



Gambar 2.1. Alur Penelitian Klasifikasi Kualitas Air Tambak Udang Vaname Menggunakan XGBoost

HASIL DAN PEMBAHASAN

3.1 Hasil Eksplorasi dan Pra-pemrosesan Data

Tahap awal penelitian dilakukan melalui *Exploratory Data Analysis* (EDA) untuk memahami karakteristik dataset kualitas air tambak udang vaname. Dataset yang digunakan terdiri atas lima parameter utama, yaitu suhu, pH, salinitas, *dissolved oxygen* (DO), dan amonia. Kelima parameter tersebut dipilih karena merupakan indikator yang umum digunakan dalam pemantauan kualitas air tambak.

Hasil eksplorasi menunjukkan bahwa terdapat beberapa atribut yang memiliki nilai kosong (*missing value*). Kondisi ini dapat terjadi karena keterbatasan alat ukur, kesalahan pencatatan, maupun perbedaan waktu pengambilan data. Keberadaan *missing value* berpotensi menurunkan kualitas model karena sebagian besar algoritma *machine learning* memerlukan data yang lengkap pada setiap atribut. Oleh karena itu, dilakukan proses imputasi untuk mengisi nilai yang hilang sehingga seluruh data dapat digunakan dalam proses klasifikasi.

Setelah proses imputasi selesai, dilakukan pelabelan data berdasarkan standar kualitas air tambak udang vaname. Setiap data diklasifikasikan ke dalam tiga kategori, yaitu layak, cukup layak, dan tidak layak. Pelabelan dilakukan berdasarkan kombinasi nilai parameter kualitas air yang mengacu pada standar budidaya udang vaname.

3.2 Perbandingan Performa Sebelum dan Sesudah *Scaling*

Perbandingan hasil klasifikasi dilakukan untuk mengetahui pengaruh penggunaan *scaling* terhadap performa algoritma XGBoost. Evaluasi mencakup akurasi, *train accuracy*, *test accuracy*, *overfitting gap*, dan *log loss*.

Tabel 3.1 Perbandingan Performa Model XGBoost

Metode	Akurasi (%)	Train Accuracy (%)	Test Accuracy (%)	Overfitting Gap (%)	Log Loss
Tanpa *Scaling*	100,00	100,00	100,00	0,00	0,014
Dengan *Scaling*	38,24	31,82	38,24	−6,42	1,239

3.2.1 Analisis Kondisi Tanpa *Scaling*

Model tanpa *scaling* menghasilkan akurasi 100% pada data latih maupun data uji. Hasil ini secara wajar menimbulkan kecurigaan: apakah model benar-benar belajar, atau hanya "menghafal" data? Untuk menjawab pertanyaan tersebut, dua metrik tambahan digunakan. *Overfitting gap* bernilai 0,00%, yang berarti tidak ada perbedaan antara performa pada data latih dan data uji model tidak menunjukkan tanda-tanda *overfitting* konvensional. *Log loss* yang sangat rendah sebesar 0,014 mengindikasikan bahwa model memberikan prediksi probabilistik yang sangat yakin dan tepat pada hampir seluruh sampel. Kedua metrik ini secara bersama memberikan dukungan empiris bahwa performa model konsisten dan tidak bersifat kebetulan semata.

Meski demikian, nilai sempurna pada dataset berukuran kecil (~100 sampel) tetap harus diinterpretasikan dengan hati-hati. Kemungkinan penjelasan yang perlu dipertimbangkan meliputi: (1) data yang mudah dipisahkan (*linearly separable*) karena pola hubungan antar parameter kualitas air dan label kelas bersifat tegas; (2) ukuran dataset yang terbatas sehingga belum tentu mencerminkan keragaman kondisi tambak secara menyeluruh. Oleh karena itu, diperlukan validasi lebih lanjut menggunakan data yang lebih besar dan beragam untuk mengonfirmasi kemampuan generalisasi model.

3.2.2 Analisis Kondisi Dengan Scaling

Penerapan *feature scaling* menghasilkan penurunan performa yang signifikan: akurasi turun menjadi 38,24%, *log loss* melonjak ke 1,239, dan *overfitting gap* bernilai −6,42%. Beberapa faktor yang dapat menjelaskan hasil ini:

Inkompatibilitas algoritma berbasis pohon dengan *scaling*. XGBoost bekerja berdasarkan pemisahan nilai fitur secara ambang batas (*threshold-based splitting*), bukan berdasarkan jarak antar sampel. Secara teoritis, transformasi linier monoton seperti normalisasi tidak mengubah urutan relatif nilai sehingga tidak seharusnya memengaruhi hasil. Namun pada dataset kecil, transformasi dapat memperbesar pengaruh *outlier* atau mengubah distribusi nilai secara tidak proporsional.

Overfitting gap negatif. Nilai *overfitting gap* sebesar −6,42% (akurasi uji lebih tinggi dari akurasi latih) merupakan kondisi tidak lazim yang mengindikasikan model gagal mempelajari pola dari data latih secara efektif setelah transformasi. Hal ini dapat disebabkan

oleh perubahan distribusi fitur yang tidak sesuai dengan karakteristik data, sehingga model kehilangan kemampuan membedakan kelas.

Log loss sebesar 1,239 menunjukkan bahwa model tidak hanya menghasilkan prediksi yang salah, tetapi juga memberikan keyakinan tinggi pada kelas yang keliru. Ini merupakan indikasi kegagalan pembelajaran yang lebih serius dibandingkan sekadar akurasi rendah.

Secara keseluruhan, temuan ini menegaskan bahwa penerapan *scaling* pada algoritma berbasis pohon seperti XGBoost tidak hanya tidak memberikan manfaat, tetapi dapat secara aktif merusak kemampuan model, terutama pada dataset berukuran kecil.

3.3 *Feature Importance*

Selain mengevaluasi performa model, penelitian ini juga melakukan analisis *feature importance* untuk mengetahui tingkat kontribusi masing-masing parameter kualitas air terhadap hasil klasifikasi.

Tabel 3.2 Hasil Fe*ature Importance*

Parameter	Importance (%)
Amonia	44,00
pH	40,00
Suhu	7,80
Salinitas	4,30
DO	3,64

Hasil analisis menunjukkan bahwa parameter amonia memiliki tingkat kontribusi tertinggi (44%), diikuti pH (40%), suhu (7,8%), salinitas (4,3%), dan *dissolved oxygen* (3,64%). Temuan ini sejalan dengan berbagai penelitian terkini yang menunjukkan bahwa parameter kimia perairan seperti amonia, pH, nitrogen, dan oksigen terlarut sering menjadi faktor dominan dalam model klasifikasi maupun prediksi kualitas air berbasis machine learning (W. Li et al., 2025; Nallakaruppan et al., 2024). Dominasi amonia dan pH dalam menentukan kelayakan kualitas air sesuai dengan karakteristik biologis udang vaname: amonia bersifat toksik pada konsentrasi tinggi (Alwateer, 2026) sementara pH memengaruhi proses fisiologis dan keseimbangan lingkungan perairan secara langsung (Himawan et al., 2025).

Parameter suhu, salinitas, dan DO memiliki kontribusi lebih rendah, namun tetap berperan sebagai faktor pendukung dalam pengambilan keputusan model. Hasil ini menunjukkan bahwa klasifikasi kualitas air tidak hanya dipengaruhi satu parameter, melainkan kombinasi beberapa parameter yang saling berkaitan.

3.4 Diskusi dengan Penelitian Sebelumnya

Hasil penelitian ini menunjukkan bahwa algoritma XGBoost mampu menghasilkan akurasi sangat tinggi (100%) pada kondisi tanpa *scaling*. Temuan ini konsisten dengan beberapa penelitian terkini yang melaporkan performa unggul XGBoost dalam pemodelan kualitas air dan pengambilan keputusan berbasis data lingkungan (Alnemari et al., 2025; W. Li et al., 2025). Temuan ini sejalan dengan penelitian sebelumnya yang menerapkan XGBoost untuk klasifikasi kualitas air akuakultur dan memperoleh performa tinggi (Naim et al., 2025) serta penelitian yang menyatakan bahwa algoritma berbasis pohon keputusan cenderung tidak sensitif terhadap perbedaan skala data (W. Li et al., 2025; Lokman et al., 2025).

Namun, penelitian ini memberikan kontribusi lebih lanjut dengan menunjukkan bahwa *scaling* tidak hanya netral, tetapi dapat bersifat destruktif terhadap performa XGBoost pada dataset kecil. Penurunan drastis akurasi dari 100% menjadi 38,24% setelah *scaling* merupakan temuan yang penting secara praktis: pemilihan teknik *preprocessing* harus disesuaikan dengan jenis algoritma dan karakteristik dataset, bukan diterapkan secara default. Hasil tersebut menunjukkan bahwa pemilihan teknik *preprocessing* perlu mempertimbangkan karakteristik algoritma yang digunakan dan tidak selalu memberikan dampak positif terhadap performa model (Ab Karim et al., 2026; Nallakaruppan et al., 2024).

Penggunaan metrik *log loss* dan *overfitting gap* dalam penelitian ini juga memberikan gambaran yang lebih lengkap dibandingkan hanya menggunakan akurasi. Pendekatan evaluasi yang komprehensif semakin direkomendasikan dalam penelitian machine learning karena mampu menggambarkan reliabilitas, interpretabilitas, dan kemampuan generalisasi model secara lebih objektif (Ab Karim et al., 2026; Aderemi et al., 2025). Nilai *log loss* yang sangat rendah (0,014) pada kondisi tanpa *scaling* mengonfirmasi bahwa model tidak sekadar "benar", tetapi juga "yakin dengan benar". Sebaliknya, *log loss* tinggi pada kondisi *scaling* (1,239) mengungkapkan kegagalan kalibrasi yang tidak terdeteksi dari akurasi saja.

KESIMPULAN

Berdasarkan temuan penelitian yang telah dilakukan, algoritma XGBoost berhasil mengklasifikasikan kelayakan kualitas air tambak udang vaname berdasarkan parameter suhu, pH, salinitas, *dissolved oxygen* (DO), dan amonia dengan akurasi 100% pada kondisi tanpa *scaling*. Nilai *overfitting gap* sebesar 0% dan *log loss* yang sangat rendah sebesar 0,014 memberikan konfirmasi tambahan bahwa performa model bersifat konsisten dan tidak hanya mencerminkan hafalan data, melainkan mencerminkan kemampuan generalisasi yang baik pada data yang tersedia (Babshette & S, 2025; Naim et al., 2025).

Sementara itu, penerapan *feature scaling* justru menurunkan performa model secara signifikan, dengan akurasi turun ke 38,24%, *log loss* meningkat ke 1,239, dan *overfitting gap* bernilai negatif sebesar −6,42%. Temuan ini konsisten dengan sifat algoritmik XGBoost yang berbasis pohon keputusan, di mana proses pemisahan nilai fitur dilakukan secara ambang batas sehingga tidak bergantung pada keseragaman skala antarfitur (Babshette & S, 2025; Lokman et al., 2025). Dengan demikian, penggunaan *StandardScaler* tidak hanya tidak memberikan manfaat, tetapi secara aktif merusak kemampuan model pada dataset berukuran kecil.

Evaluasi komprehensif menggunakan akurasi, *log loss*, dan *overfitting gap* secara bersamaan terbukti lebih informatif dibandingkan hanya mengandalkan nilai akurasi. Pendekatan ini sejalan dengan perkembangan penelitian machine learning terkini yang menekankan pentingnya interpretabilitas, transparansi, dan kemampuan generalisasi model dalam mendukung pengambilan keputusan berbasis data (Ab Karim et al., 2026; Aderemi et al., 2025). Ketiga metrik ini memberikan gambaran yang lebih jujur tentang kualitas model, terutama dalam mendeteksi potensi *overfitting* laten yang tidak terlihat dari angka akurasi saja (Naim et al., 2025; Singh et al., 2025). Selain itu, analisis *feature importance* mengidentifikasi amonia (44%) dan pH (40%) sebagai parameter paling berpengaruh dalam menentukan kelayakan kualitas air tambak, yang sesuai dengan karakteristik biologis udang vaname di mana kadar amonia yang tinggi bersifat toksik dan perubahan pH secara langsung memengaruhi proses fisiologis udang (Bambang et al., 2024; Himawan et al., 2025; Ritonga et al., 2021).

UCAPAN TERIMA KASIH

Penulis menyampaikan rasa syukur kepada Program Studi Sistem Informasi Kelautan di Universitas Pendidikan Indonesia Kampus Serang atas bantuan akademik yang diberikan selama penelitian ini dilakukan. Terima kasih juga ditujukan kepada seluruh pengajar pembimbing serta semua pihak yang telah memberikan saran, petunjuk, dan dukungan selama proses penyusunan penelitian hingga selesai penyusunan artikel ini. Di samping itu, penulis menghaturkan terima kasih kepada penyedia dataset melalui platform Figshare yang telah memberikan data tentang kualitas air tambak udang vaname sehingga penelitian ini bisa berjalan dengan lancar.

DAFTAR PUSTAKA

Ab Karim, M. A., Wan Ismail, W. Z., Mohd Shuib, F. M., Ab Aziz, N. A., & Ghazali, A. K. (2026). Water Quality Monitoring and Assessment Using Machine Learning: A Review of Formulation, Modeling Approaches, and Explainable Artificial Intelligence. *Environments*, *13*(5), 1–20. https://doi.org/10.3390/environments13050267

Aderemi, I. A., Kehinde, T. O., Daniel Okwor, U., Ahmad, K. H., Adjei, K. Y., & Cyriacus Ekechi, C. (2025). Explainable AI for Water Quality Monitoring: A Systematic Review of Transparency, Interpretability, and Trust. *IEEE Sensors Reviews*, *2*(June), 419–443. https://doi.org/10.1109/sr.2025.3595500

Alnemari, A. M., Elmessery, W. M., Qazaq, A. S., Moustapha, M. E., Rakhimgaliyeva, S., Abuhussein, M. F. A., Alhag, S. K., Al-Shuraym, L. A., Moghanm, F. S., Szűcs, P., Eid, M. H., & Elwakeel, A. E. (2025). Developing highly accurate machine learning models for optimizing water quality management decisions in tilapia aquaculture. *Scientific Reports*, *15*(1), 1–19. https://doi.org/10.1038/s41598-025-16939-w

Alwateer, M. (2026). Water quality prediction using Machine Learning Models. *Sustainability (Switzerland)*, *18*. https://doi.org/https://doi.org/10.3390/su18041721

Ananta, M. fauzi, Nariswana, R., Supriyadi, davin F., Fauzan, H. T. G. Al, & Nugroho, R. S. (2024). Smart water quality monitoring with IoT and AI: XG approach in household water feasibility. *Journal of Scientech Research and Development*, *6*(2), 1092–1099. https://idm.or.id/JSCR/index.php/JSCR/article/view/14

Babshette, N., & S, V. (2025). Water Quality Classification Using SVM And XGboost Method Machine Learning Naveen. *JOURNAL OF SCIENTIFIC RESEARCH AND TECHNOLOGY PAGES:*, *3*(8), 51–54. https://doi.org/10.1109/ICSGRC55096.2022.9845143

Baena-Navarro, R., Carriazo-Regino, Y., Torres-Hoyos, F., & Pinedo-López, J. (2025). Intelligent Prediction and Continuous Monitoring of Water Quality in Aquaculture: Integration of Machine Learning and Internet of Things for Sustainable Management. *Water (Switzerland)*, *17*(1). https://doi.org/10.3390/w17010082

Bambang, S., Bintari, yunia karisma, Aulia, D., Rizky, putri nurhanida, & Watina, S. (2024). KELIMPAHAN PLANKTON DAN PROFIL KUALITAS AIR BUDIDAYA UDANG VANNAMEI (*Litopenaeus vannamei*) SISTEM INTENSIF. *Aurelia Journal*, *6*(75), 227–236.

Himawan, S. N., Arif, W., & Nugraha, N. B. (2025). Prediction of Nile Tilapia (*Oreochromis niloticus*) Harvest Yield in Brackishwater Pond Aquaculture Using XGBoost. *Journal of*

Applied Informatics and Computing (JAIC), *10*(1).

Li, L., & Wei, J. (2024). Evaluation of Tree-Based Voting Algorithms in Water Quality Classification Prediction. *Sustainability (Switzerland)*, *16*(23). https://doi.org/10.3390/su162310634

Li, W., Deng, M., Liu, C., & Cao, Q. (2025). Analysis of Key Influencing Factors of Water Quality in Tai Lake Basin Based on XGBoost-SHAP. *Water (Switzerland)*, *17*(11), 1–19. https://doi.org/10.3390/w17111619

Lokman, A., Ismail, W. Z. W., & Aziz, N. A. A. (2025). A Review of Water Quality Forecasting and Classification Using Machine Learning Models and Statistical Analysis. *Water (Switzerland)*, *17*(15), 1–31. https://doi.org/10.3390/w17152243

Naim, S. M., Das, P., Tiang, J. J., & Nahid, A. Al. (2025). Aquaculture Water Quality Classification Using XGBoost Classifier Model Optimized by the Honey Badger Algorithm with SHAP and DiCE-Based Explanations. *Water (Switzerland)*, *17*(20), 1–18. https://doi.org/10.3390/w17202993

Nallakaruppan, M. K., Gangadevi, E., Shri, M. L., Balusamy, B., Bhattacharya, S., & Selvarajan, S. (2024). Reliable water quality prediction and parametric analysis using explainable AI models. *Scientific Reports*, *14*(1), 1–24. https://doi.org/10.1038/s41598-024-56775-y

Nuangpirom, P., Pitjamit, S., Jaikampan, V., Peerakam, C., Nakkiew, W., & Jewpanya, P. (2025). Machine Learning on Low-Cost Edge Devices for *Real-Time* Water Quality Prediction in Tilapia Aquaculture. *Sensors*, *25*(19), 1–22. https://doi.org/10.3390/s25196159

Ritonga, L. B., Asmarany, A., & Aritmatika, P. E. (2021). Management of Water Quality in Intensive Enlargement of Vannamei Shrimp (*Litopenaeus vannamei*) in PT. Andulang Shrimp Farm. *Journal of Aquaculture Development and Environment*, *4*(1), 218. https://doi.org/10.31002/jade.v4i1.3739

Singh, P., Hasija, T., Bharany, S., Naeem, H. N. T., Rao, B. C., Hussen, S., & Rehman, A. U. (2025). An ensemble-driven machine learning framework for enhanced water quality classification. *Discover Sustainability*, *6*(1). https://doi.org/10.1007/s43621-025-01467-4