KLASIFIKASI KUALITAS AIR BERSIH DI KABUPATEN BOJONEGORO MENGGUNAKAN *XGBOOST*

SKRIPSI

Oleh:
NAILA NABILA ILIANA
NIM. 200605110110

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2026

KLASIFIKASI KUALITAS AIR BERSIH DI KABUPATEN BOJONEGORO MENGGUNAKAN *XGBOOST*

SKRIPSI

Diajukan Kepada:
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk Memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Sarjana Komputer (S.Kom)

Oleh :
NAILA NABILA ILIANA
NIM. 200605110110

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2026

HALAMAN PERSETUJUAN

KLASIFIKASI KUALITAS AIR BERSIH DI KABUPATEN BOJONEGORO MENGGUNAKAN *XGBOOST*

SKRIPSI

Oleh :
NAILA NABILA ILIANA
NIM. 200605110110

Telah Diperiksa dan Disetujui untuk Diuji:
Tanggal: 23 Juni 2026

Pembimbing I,	Pembimbing II,
Dr. Ir. Fachrul Kurniawan, M.MT., IPU	Ashri Shabrina Afrah, M.T
NIP. 197710202009121001	NIP. 199004302020122003

Mengetahui,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Supriyono, M. Kom
NIP. 19841010 201903 1 012

HALAMAN PENGESAHAN

KLASIFIKASI KUALITAS AIR BERSIH DI KABUPATEN BOJONEGORO MENGGUNAKAN *XGBOOST*

SKRIPSI

Oleh :
NAILA NABILA ILIANA
NIM. 200605110110

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal: 24 Juni 2026

Susunan Dewan Penguji

Ketua Penguji	: Dr. Yunifa Miftachul Arif, M.T NIP. 19830616 201101 1 004	( )
Anggota Penguji I	: Ahmad Fahmi Karami, M.Kom NIP. 19870909 202012 1 001	( )
Anggota Penguji II	: Dr. Ir. Fachrul Kurniawan, M.MT., IPU NIP. 19771020 200912 1 001	( )
Anggota Penguji III	: Ashri Shabrina Afrah, M.T NIP. 199004302020122003	( )

Mengetahui dan Mengesahkan,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Supriyono, M.Kom
NIP. 19841010 201903 1 012

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama	: Naila Nabila Iliana
NIM	: 200605110110
Fakultas / Prodi	: Sains dan Teknologi / Teknik Informatika
Judul Skripsi	: KLASIFIKASI KUALITAS AIR BERSIH DI KABUPATEN BOJONEGORO MENGGUNAKAN *XGBOOST*.

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambilalihan data, tulisan, atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan skripsi ini merupakan hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 23 Juni 2026
Yang membuat pernyataan,

Naila Nabila Iliana
NIM.200605110110

MOTTO

"Sesungguhnya Allah tidak akan mengubah keadaan suatu kaum sebelum mereka mengubah keadaan diri mereka sendiri."
(Q.S Ar-Ra'd:11)

"Kekhawatiran adalah separuh penyakit, ketenangan adalah separuh kesembuhan, dan kesabaran adalah langkah awal menuju jalan keluar."
(Ali bin Abi Thalib)

HALAMAN PERSEMBAHAN

Alhamdulillahi rabbil 'alamin segala puji bagi Allah SWT yang telah melimpahkan rahmat, hidayah, serta kemudahan sehingga penulis dapat menyelesaikan skripsi ini dengan baik. Shalawat dan salam senantiasa tercurah kepada baginda Rasulullah SAW yang telah menuntun umat manusia keluar dari zaman jahiliyah menuju cahaya Islam.

Dengan rasa hormat dan terima kasih, penulis mempersembahkan skripsi tugas akhir ini kepada kedua orang tua, diri sendiri, kakak dan adik-adik tersayang, para dosen yang telah membimbing, teman-teman, serta orang-orang yang telah membantu, mendoakan, serta menyemangati penulis hingga selesainya skripsi ini.

KATA PENGANTAR

Assalamu'alaikum warahmatullahi wabarakatuh

Puji syukur penulis panjatkan kepada Allah Subhanahu wa Ta'ala yang senantiasa memberikan rahmat serta kesehatan, sehingga penulis mampu menyelesaikan penulisan skripsi yang berjudul "Klasifikasi Kualitas Air Bersih di Kabupaten Bojonegoro Menggunakan *XGBoost*".

Penulis menyampaikan ucapan terima kasih kepada semua pihak yang pernah terlibat dalam menyelesaikan penelitian ini, bukan hanya karena usaha keras dari penulis sendiri, akan tetapi karena adanya dukungan dari berbagai pihak. Oleh karena itu penulis berterima kasih kepada::

1. Prof. Dr. Hj. Ilfi Nur Diana, M.Si, selaku Rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
2. Dr. H. Agus Mulyono, M.Si, selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
3. Supriyono, M.Kom, selaku Ketua Program Studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Dr. Ir. Fachrul Kurniawan, ST.,M.MT.,IPU, selaku dosen pembimbing 1 yang dengan penuh kesabaran senantiasa membimbingan, memberi masukan, dan memberikan dukungan yang sangat berarti bagi penulis selama proses penelitian.
5. Ashri Shabrina Afrah, M.T, selaku dosen pembimbing 2, atas arahan dan masukan yang sangat membantu untuk menyempurnakan karya ini

6. Dr. Ir. Yunifa Miftachul Arif, S.ST., M.T., SMIEEE selaku dosen penguji I dan Ahmad Fahmi Karami, M.Kom selaku dosen penguji II yang telah memberikan kritik dan saran yang membangun demi kesempurnaan skripsi ini.
7. Segenap civitas akademik Program Studi Teknik Informatika yang telah memberikan ilmu serta arahan semasa kuliah..
8. Nia Faricha, S.Si selaku admin program Studi Teknik Informatika yang sangat baik hati dan sabar untuk membantu memberikan arahan, informasi terkait perkuliahan.
9. Kedua orang tua penulis, Bapak Ali Ma'zum dan Ibu Siti Junaidah yang selalu memberikan banyak dukungan dan doa serta selalu menjadi semangat sehingga penulis mampu menyelesaikan masa studi hingga mencapai gelar sarjana. Semoga Allah Ta'ala melimpahkan kesehatan, keberkahan, kebahagiaan, perlindungan, serta membalas seluruh kebaikan dengan pahala yang berlipat ganda.
10. Keluarga besar penulis terkhusus Kakak Durrotun Nafisah, Adik Azka Salsabila, Adik Ahmad Hannan Amrulloh, Paman Huda dan Adik Shofia yang memberikan dorongan semangat, nasihat, doa, motivasi sehingga penulis mampu menyelesaikan masa studi.
11. Sahabat penulis penghuni grup "calon juara"; yaitu Ayum, Tasa, Bayu, Baiq, Fakhar dan Luthfi terima kasih sudah membersamai penulis dalam proses perkuliahan dari awal hingga akhir, terima kasih atas segala bantuan, dukungan, semangat yang diberikan, tampungan-tampungan keluh kesah

penulis hingga terselesaikan skripsi ini.

12. Sahabat penulis yaitu Poty, Shiyu dan Putri, terima kasih telah menjadi manusia yang sabar dan memberikan dukungan, waktu, selalu peduli, dan selalu merayakan serta memotivasi penulis hingga terselesaikannya skripsi ini.
13. Seluruh keluarga besar Angkatan ”INTEGER”; terkhusus teman teman dekat penulis yang tidak dapat disebutkan satu persatu, terima kasih telah memberikan bantuan, motivasi disaat hilangnya harapan harapan serta telah bersama-sama membuat banyak pengalaman berharga yang tidak pernah penulis dapatkan dimanapun.
14. Seluruh pihak yang turut berkontribusi, baik secara langsung maupun tidak langsung.

Penulis berharap semoga skripsi ini dapat memberikan kontribusi dan manfaat di masa yang akan datang. Penulis menyadari bahwa karya ini masih jauh dari kata sempurna. Oleh karena itu, penulis sangat menghargai dan senang jika terdapat kritik dan saran yang diberikan.

Wassalamu’alaikum Warahmatullahi Wabaraktuh

Malang, 24 Juni 2026

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Iliana, Naila Nabila. 2026. **Klasifikasi Kualitas Air Bersih di Kabupaten Bojonegoro Menggunakan *XGBoost***. Skripsi. Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Dr. Ir. Fachrul Kurniawan, M.MT., IPU. (II) Ashri Shabrina Afrah, M.T

Kata Kunci: Kualitas Air Bersih, Klasifikasi, *XGBoost*, *Hyperparameter Tuning*, Kabupaten Bojonegoro

Kualitas air bersih merupakan kebutuhan dasar yang sangat vital bagi kesehatan, namun beberapa sumber air sumur warga di Kabupaten Bojonegoro terindikasi memiliki kadar parameter fisik yang melebihi ambang batas aman. Penelitian ini bertujuan untuk mengembangkan arsitektur model klasifikasi kelayakan kualitas air bersih yang paling andal menggunakan algoritma *Extreme Gradient Boosting* (*XGBoost*) melalui pengujian variasi rasio pembagian data dan optimasi *hyperparameter*. Pengumpulan data primer dilakukan secara langsung menggunakan perangkat *Internet of Things* (IoT) pada tiga lokasi sumur, menghasilkan 3.511 baris data observasi yang terdiri dari parameter pH, suhu, kekeruhan, dan *Total Dissolved Solids* (TDS). Pengelompokan status kelayakan air dilakukan berdasarkan aturan standar baku mutu kesehatan lingkungan yang berlaku. Proses pelatihan dan pengujian model menerapkan teknik optimasi *hyperparameter* menggunakan *GridSearchCV* dan divalidasi dengan *5-Fold Cross Validation* pada empat skenario pembagian dataset. Hasil penelitian menyimpulkan bahwa arsitektur model terbaik dicapai pada skenario rasio data latih dan uji 70:30, menggunakan konfigurasi max_depth = 7, learning_rate = 0.01, dan n_estimators = 300. Arsitektur tersebut menghasilkan performa klasifikasi yang superior dengan tingkat Akurasi 98,77%, Precision 98,88%, *Recall* 99,50%, dan F1-Score 99,19%. Model ini juga terbukti sangat presisi dan aman karena berhasil menekan tingkat kesalahan klasifikasi fatal (*False Positive*) menjadi hanya 9 kasus dari total 1.054 sampel uji. Secara keseluruhan, algoritma *XGBoost* terbukti sangat andal dan stabil dalam mengklasifikasikan kualitas air, sehingga layak diimplementasikan sebagai instrumen cerdas untuk mendukung pemantauan kelayakan air bersih di masyarakat.

ABSTRACT

Iliana, Naila Nabila. 2026. **Classification of Drinking Water Quality in Bojonegoro Regency Using XGBoost**. Thesis. Informatics Engineering Study Program, Faculty of Science and Technology, Maulana Malik Ibrahim State Islamic University Malang. Advisors: (I) Dr. Ir. Fachrul Kurniawan, M.MT., IPU. (II) Ashri Shabrina Afrah, M.T.

Clean water quality is a highly vital basic need for health, but several residential well water sources in Bojonegoro Regency indicate physical parameter levels that exceed safe thresholds. This study aims to develop the most reliable clean water quality classification model architecture using the Extreme Gradient Boosting (*XGBoost*) algorithm through testing variations in data split ratios and *hyperparameter* optimization. Primary data collection was conducted directly using Internet of Things (IoT) devices at three well locations, generating 3,511 rows of observation data consisting of pH, temperature, turbidity, and Total Dissolved Solids (TDS) parameters. The grouping of water feasibility status was carried out based on applicable environmental health quality standard rules. The model training and testing process applied *hyperparameter* optimization techniques using GridSearchCV and was validated with 5-*Fold* Cross Validation across four dataset split scenarios. The research results conclude that the best model architecture (Best Estimator) was achieved in the 70:30 train-test data ratio scenario, using the configuration of max_depth = 7, learning_rate = 0.01, and n_estimators = 300. This architecture produced superior classification performance with an Accuracy of 98.77%, Precision of 98.88%, Recall of 99.50%, and an F1-Score of 99.19%. This model also proved to be highly precise and safe because it successfully suppressed the fatal classification error rate (False Positives) to only 9 cases out of a total of 1,054 test samples. Overall, the *XGBoost* algorithm proved to be highly reliable and stable in classifying water quality, making it feasible to be implemented as an intelligent instrument to support the monitoring of clean water feasibility in the community.

Keywords: Clean Water Quality*, Classification, XGBoost, Hyperparameter Tuning,* Bojonegoro Regency.

الملخص

إيليانا، نايلة نبيلة. 2026. تصنيف جودة مياه الشرب في مقاطعة بوجونيجورو باستخدام *XGBoost*. أطروحة تخرج. برنامج هندسة المعلوماتية، كلية العلوم والتكنولوجيا، جامعة مولانا مالك إبراهيم الإسلامية الحكومية في مالانج. المشرفان: (I) د. إير. فاخرول كورنياوان، ماجستير في الهندسة الميكانيكية، عضو في الاتحاد الدولي للمهندسين (II) .(IPU) أشري شابرينا أفراح، ماجستير في الهندسة

الكلمات المفتاحية: جودة المياه النظيفة، التصنيف، *XGBoost*، ضبط المعلمات الفائقة، مقاطعة بوجونيجورو

تعد جودة المياه النظيفة حاجة أساسية وحيوية للغاية للصحة، إلا أن بعض مصادر مياه الآبار الخاصة بالسكان في مقاطعة بوجونيجورو تشير إلى وجود مستويات من المعلمات الفيزيائية تتجاوز الحدود الآمنة. تهدف هذه الدراسة إلى تطوير بنية نموذج تصنيف صلاحية جودة المياه النظيفة الأكثر موثوقية باستخدام خوارزمية *Extreme Gradient Boosting* (*XGBoost*) من خلال اختبار تباين نسب تقسيم البيانات وتحسين المعلمات الفائقة. تم جمع البيانات الأولية مباشرةً باستخدام أجهزة *إنترنت الأشياء* (IoT) في ثلاثة مواقع للآبار، مما أسفر عن 3,511 صفًا من بيانات المراقبة التي تتألف من معلمات الرقم الهيدروجيني (pH)، ودرجة الحرارة، والعكارة، *وإجمالي المواد الصلبة الذائبة* (TDS). وتم تصنيف حالة صلاحية المياه بناءً على القواعد القياسية المعمول بها لجودة الصحة البيئية. استخدمت عملية تدريب النموذج واختباره تقنية تحسين *المعلمات الفائقة* باستخدام *GridSearchCV*، وتم التحقق من صحتها من خلال *«التحقق المتقاطع الخماسي» (5-Fold Cross Validation)* في أربعة سيناريوهات لتقسيم مجموعة البيانات. وخلصت نتائج البحث إلى أن أفضل بنية للنموذج تم تحقيقها في السيناريو الذي تبلغ فيه نسبة البيانات التدريبية إلى البيانات الاختبارية 70:30، باستخدام التكوين التالي: max_depth = 7، و learning_rate = 0.01، و n_estimators = 300. وقد حققت هذه البنية أداءً تصنيفيًّا متميزًا بمعدل دقة (*Accuracy*) بلغ 98,77%، ودقة تحديد (*Precision*) بلغت 98,88%، ومعدل *استرجاع (Recall)* بلغ 99,50%، ومؤشر F1-Score بلغ 99,19%. كما أثبت هذا النموذج دقة وأمانًا فائقين، حيث نجح في خفض معدل الأخطاء الفادحة في التصنيف (*النتائج الإيجابية الخاطئة*) إلى 9 حالات فقط من إجمالي 1,054 عينة اختبار. وبشكل عام، أثبتت خوارزمية *XGBoost* موثوقيتها واستقرارها الفائقين في تصنيف جودة المياه، مما يجعلها جديرة بالتطبيق كأداة ذكية لدعم مراقبة صلاحية المياه النظيفة في المجتمعات المحلية.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Air bersih merupakan kebutuhan dasar manusia yang sangat krusial bagi kesehatan dan kehidupan sehari-hari. Ketersediaan air bersih yang aman untuk dikonsumsi menjadi salah satu faktor utama dalam meningkatkan kualitas hidup masyarakat. Dalam perspektif Islam, air tidak hanya dipandang sebagai unsur fisik, melainkan sebagai anugerah *Ilahi* dan tanda kebesaran Allah SWT. Al-Qur'an secara eksplisit menyebutkan pentingnya air sebagai sumber kehidupan dan kebersihan, sebagaimana firman Allah dalam Surah Al-Anbiya' ayat 30 yang berbunyi:

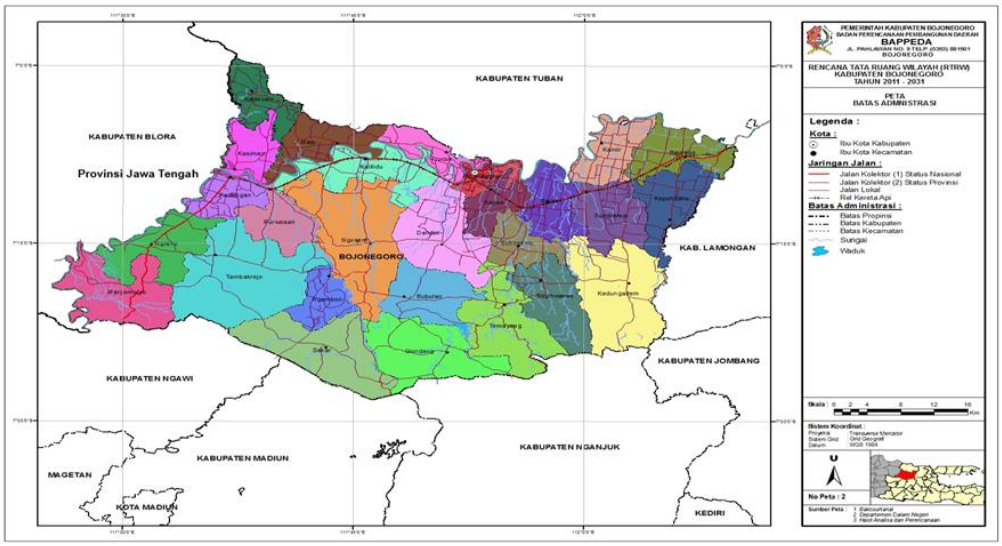
وَلَمْ يَرَ الَّذِيْنَ كَفَرُوْٓا اَنَّ السَّمٰوٰتِ وَالْاَرْضَ كَانَتَا رَتْقًا فَفَتَقْنٰهُمَاۗ وَجَعَلْنَا مِنَ الْمَاۤءِ كُلَّ شَيْءٍ حَيٍّۗ اَفَلَا يُؤْمِنُوْنَ ﴿٣٠﴾

"Dan dari air Kami jadikan segala sesuatu yang hidup." (QS Al-Anbiya:30)

Ayat ini menegaskan peran sentral air dalam keberlangsungan hidup, sekaligus mengisyaratkan pentingnya menjaga kualitas dan ketersediaan air sebagai bagian dari amanah *khalifatullah fil ardh*. Ilmu pengetahuan modern, khususnya di bidang Teknik Informatika, dapat menjadi alat untuk menjalankan amanah tersebut dengan mengembangkan sistem yang mampu memantau dan mengklasifikasikan kualitas air demi kemaslahatan umat.

Meskipun memiliki peran fundamental, kualitas air bersih masih menjadi permasalahan yang signifikan di berbagai daerah. Banyak wilayah menghadapi

tantangan dalam penyediaan air layak konsumsi, baik akibat pencemaran, kandungan zat berbahaya, maupun ketidakseimbangan akses. Permasalahan ini selaras dengan salah satu target dalam *Sustainable Development Goals* (SDG) poin 6, yaitu *Clean Water and Sanitation*, yang bertujuan untuk menjamin akses air bersih dan sanitasi yang layak bagi seluruh masyarakat (United Nations, 2016). Untuk mencapai target tersebut, diperlukan langkah-langkah strategis, termasuk pemantauan kualitas air berbasis data serta pemanfaatan teknologi dalam proses klasifikasi dan prediksi kualitas air.



Gambar 1. 1 Peta Bojonegoro

Penelitian ini memfokuskan studi kasus di Kabupaten Bojonegoro, sebuah daerah dengan sektor pertanian dan industri yang berkembang, di mana kualitas dan ketersediaan air sangat dipengaruhi oleh faktor lingkungan. Berdasarkan data hasil pemantauan yang dirilis oleh Dinas Lingkungan Hidup Kabupaten Bojonegoro (2023), terungkap adanya temuan signifikan terkait kondisi air tanah di daerah tersebut. Secara spesifik, ditemukan bahwa beberapa sumber air sumur yang digunakan oleh warga di Kecamatan Tambakrejo dan Kedungadem memiliki

parameter kualitas yang tidak sesuai standar. Kadar *Total Dissolved Solids* (TDS) atau jumlah total padatan terlarut serta tingkat kekeruhannya tercatat telah melebihi ambang batas baku mutu yang ditetapkan dalam Peraturan Menteri Kesehatan. Kondisi ini mengindikasikan bahwa air tersebut secara teknis tidak layak untuk dikonsumsi secara langsung dan memerlukan pengolahan lebih lanjut untuk memastikan keamanannya bagi kesehatan.

Berbagai faktor seperti pencemaran limbah domestik, pertanian, serta aktivitas industri dapat menurunkan kualitas air dan meningkatkan risiko penyakit berbasis air (*waterborne diseases*), seperti diare dan infeksi saluran pencernaan (*World Health Organization*, 2022). Kandungan *Total Dissolved Solids* (TDS), pH, suhu, dan kekeruhan air menjadi parameter penting dalam menentukan kelayakan air bersih untuk dikonsumsi. Jika kualitas air tidak memenuhi standar, maka dampaknya bisa sangat serius bagi kesehatan masyarakat yang bergantung pada air dengan kualitas yang baik.

Dalam beberapa penelitian, pendekatan berbasis kecerdasan buatan (*artificial intelligence*) dan pembelajaran mesin (*machine learning*) telah digunakan untuk mengklasifikasikan kualitas air (Grbčić et al., 2022; Khoi et al., 2022). Selain itu, metode *Naïve Bayes*, *Random Forest*, dan *XGBoost* juga telah dibandingkan dalam penelitian sebelumnya, dan *XGBoost* terbukti lebih unggul dalam menangani klasifikasi data kualitas air dengan tingkat akurasi yang lebih tinggi (Darshan & Subburaj, 2023).

XGBoost merupakan salah satu algoritma *gradient boosting* yang efektif dalam menangani data kompleks, terutama setelah dilakukan optimasi

hyperparameter. Keunggulan *XGBoost* dalam menangani dataset yang besar, mengenali pola dalam data, serta memberikan hasil klasifikasi yang lebih akurat menjadikannya pilihan utama dalam penelitian ini. Selain itu, optimasi *hyperparameter* memungkinkan model *XGBoost* untuk bekerja lebih optimal, mengurangi risiko *overfitting*, dan meningkatkan performa klasifikasi (Elvin & Wibowo, 2024).

Berdasarkan permasalahan yang telah diuraikan, penelitian ini bertujuan untuk melakukan klasifikasi kualitas air bersih di Kabupaten Bojonegoro menggunakan metode *XGBoost* dengan optimasi *hyperparameter*. Dengan pendekatan ini, diharapkan hasil klasifikasi yang dihasilkan dapat lebih akurat dan menjadi dasar bagi pemerintah serta pemangku kebijakan dalam merancang strategi pengelolaan sumber daya air yang lebih baik. Selain itu, penelitian ini diharapkan dapat berkontribusi dalam mendukung target SDG poin 6 dengan menyediakan informasi berbasis data yang lebih akurat terkait kualitas air, sehingga kebijakan yang diambil dapat lebih tepat sasaran dalam meningkatkan akses terhadap air bersih, baik di Bojonegoro maupun di daerah lain yang menghadapi permasalahan serupa.

1.2 Rumusan Masalah

Bagaimana tingkat performa terbaik model klasifikasi kelayakan kualitas air menggunakan algoritma *Extreme Gradient Boosting* (*XGBoost)* berdasarkan pengujian skenario komposisi data latih dan data uji serta optimasi *hyperparameter*?

1.3 Batasan Masalah

Agar penelitian lebih terarah dan fokus, beberapa batasan masalah yang diterapkan dalam penelitian ini adalah sebagai berikut:

1. Penelitian ini hanya berfokus pada klasifikasi kualitas air bersih di Kabupaten Bojonegoro.
2. Penelitian ini terbatas pada analisis empat parameter fisis air, yaitu pH, suhu, kekeruhan, dan *Total Dissolved Solids* (TDS).
3. Dataset yang digunakan dalam penelitian ini berasal dari hasil pengukuran di lapangan. Sumber air yang diuji terbatas pada air sumur gali milik warga di tiga desa yang telah ditentukan.
4. Penelitian tidak menggunakan perhitungan Indeks Kualitas Air (IKA) penuh, mengingat keterbatasan parameter yang tersedia. Labeling kualitas air dilakukan berdasarkan standar baku mutu kualitas air bersih yang ditetapkan oleh Permenkes No. 2 Tahun 2023 dan *World Health Organization* (2022).

1.4 Tujuan Penelitian

Tujuan utama dari penelitian ini adalah untuk mengembangkan arsitektur model klasifikasi kelayakan air yang paling andal menggunakan algoritma *XGBoost* melalui pengujian variasi rasio pembagian data dan optimasi *hyperparameter*.

1.5 Manfaat Penelitian

Manfaat yang diperoleh dari penelitian klasifikasi kualitas air bersih di kabupaten Bojonegoro menggunakan *XGBoost* dengan optimasi *hyperparameter*, yaitu:

1. Memberikan wawasan mendalam mengenai penerapan algoritma *XGBoost* dan optimasi *hyperparameter*-nya dalam analisis klasifikasi kualitas air berbasis *machine learning*.
2. Menyediakan informasi relevan mengenai kualitas air bersih yang dapat dimanfaatkan oleh masyarakat untuk kebutuhan sehari-hari.
3. Mendukung pengambilan keputusan berbasis data bagi pemerintah daerah dalam pengelolaan sumber daya air serta berkontribusi pada pencapaian target SDG 6 terkait ketersediaan air bersih dan sanitasi layak.
4. Menjadi referensi akademis bagi penelitian selanjutnya serta memberikan landasan bagi pengembangan sistem pemantauan kualitas air yang lebih canggih dan otomatis di masa depan.

BAB II

STUDI PUSTAKA

2.1 Penelitian Terdahulu

Penelitian terkait didefinisikan sebagai penelitian yang telah selesai dilakukan sebelum investigasi saat ini dimulai. Teori-teori yang relevan dari penelitian terdahulu dapat digunakan dalam penelitian ini, yang menggunakan penelitian terdahulu sebagai referensi atau panduan.

"Penerapan Algoritma *Random Forest* dalam Prediksi Kelayakan Air Minum" oleh Khairul Abdi et al.(2024) bertujuan untuk mengimplementasikan algoritma *Random Forest* dalam klasifikasi kelayakan air minum. Dataset yang digunakan berasal dari Kaggle dan mencakup berbagai parameter kualitas air. Hasil penelitian menunjukkan bahwa model *Random Forest* menghasilkan akurasi sebesar 69%. Analisis *Feature Importance Score* mengidentifikasi fitur yang berkontribusi signifikan terhadap prediksi, sementara kurva ROC dan *Confusion Matrix* menunjukkan kinerja model yang masih dapat ditingkatkan (Abdi et al., 2024).

"Analisis Sederhana Pada Kualitas Air Minum Berdasarkan Akurasi Model Klasifikasi Dengan Menggunakan *Lucifer Machine Learning*" oleh Prismahardi Aji Riyantoko et al. membahas penerapan teknik *Lucifer Machine Learning* dalam klasifikasi kualitas air minum. Studi ini menggunakan database terbuka dan menerapkan berbagai metode klasifikasi. Model dengan akurasi tertinggi adalah *Random Forest* Classifier dengan tingkat akurasi sebesar 72,81%. Metode ini mempermudah eksplorasi data dan analisis klasifikasi dengan pendekatan semi-

supervised learning (Riyantoko et al., 2021).

"Klasifikasi Indeks Standar Pencemaran Udara (ISPU) Menggunakan Algoritma *XGBoost* dengan Teknik *Imbalanced* Data (SMOTE)" oleh Achmad Fauzihan Bagus Sajiwo et al. mengusulkan penggunaan algoritma *XGBoost* dalam klasifikasi ISPU di DKI Jakarta. Data yang digunakan berasal dari pengukuran polusi udara tahun 2022–2023. Untuk mengatasi ketidakseimbangan data, teknik SMOTE diterapkan. Hasil pengujian menunjukkan bahwa model *XGBoost* dengan SMOTE menghasilkan tingkat akurasi sebesar 99,63%, menunjukkan efektivitas metode ini dalam mengklasifikasikan kualitas udara berdasarkan standar pencemaran (Sajiwo et al., 2024).

"Klasifikasi Kelayakan Sumber Air Minum di Jawa Barat Menggunakan *XGBoost* dan Analisis Klasterisasi Berdasarkan Nilai SHAP" oleh Annisa Permata Sari et al. membahas penggunaan model *XGBoost* untuk mengklasifikasikan kelayakan sumber air minum di Jawa Barat. Data berasal dari survei rumah tangga tahun 2023 dengan 4187 sampel. Model *XGBoost* yang diterapkan dengan fitur terpilih menghasilkan akurasi 77,43% dan *F1-Score* 80,17%. Analisis nilai SHAP mengidentifikasi lokasi sumber air, waktu pengambilan air, dan pengeluaran per kapita sebagai faktor utama yang memengaruhi kelayakan air minum (Sari et al., 2024).

"Klasifikasi Kualitas Air Bersih Menggunakan Metode *Naïve Bayes*" oleh Sutisna dan Mirsandi Nazar Yuniar mengusulkan penggunaan metode *Naïve Bayes* dalam klasifikasi kualitas air bersih. Studi ini menggunakan dataset yang dianalisis dengan perangkat lunak *Rapid Miner*. Hasil eksperimen menunjukkan bahwa

model *Naïve Bayes* mampu mencapai tingkat akurasi sebesar 97,35%, membuktikan keandalannya dalam mengklasifikasikan kualitas air dengan cepat dan efisien (Sutisna & Yuniar, 2023).

Dari tinjauan penelitian-penelitian di atas, terlihat bahwa penggunaan *machine learning* untuk klasifikasi dan prediksi kualitas air merupakan bidang yang aktif diteliti. Berbagai algoritma seperti *Random Forest, Naïve Bayes*, hingga SVM telah diterapkan dengan tingkat keberhasilan yang bervariasi. *XGBoost* secara konsisten menunjukkan performa yang unggul dalam banyak studi perbandingan, baik untuk prediksi kualitas air pesisir, estimasi *Water Quality Index* (*WQI*) di sungai, maupun klasifikasi umum. Beberapa studi bahkan melaporkan *XGBoost* mencapai akurasi di atas 90%.

Tabel 2. 1 Penelitian Terdahulu

No	**Peneliti**	**Judul**	**Metode Penelitian**	**Hasil Penelitian**	**Perbedaan**
1	(Abdi et al., 2024)	Penerapan Algoritma *Random Forest* dalam Prediksi Kelayakan Air Minum	*Random Forest*	Hasil yang diperoleh menunjukkan akurasi sebesar 69%, dengan analisis *feature importance* yang digunakan untuk mengetahui kontribusi masing-masing parameter terhadap prediksi.	Penelitian ini menggunakan *Random Forest* tanpa optimasi *hyperparameter*, sedangkan penelitian saya menggunakan *XGBoost* dengan optimasi *hyperparameter*.
2	(Riyantoko et al., 2021)	Analisis Sederhana Pada Kualitas Air Minum Berdasarkan Akurasi Model Klasifikasi Dengan Menggunakan *Lucifer Machine*	*Lucifer Machine Learning*	Dari beberapa model yang diuji, hasil akurasi tertinggi 72,81% memakai *Random Forest*	Penelitian ini menggunakan semi-*supervised learning* dengan *library* otomatisasi dan tidak fokus pada satu model tertentu, sementara penelitian saya menggunakan

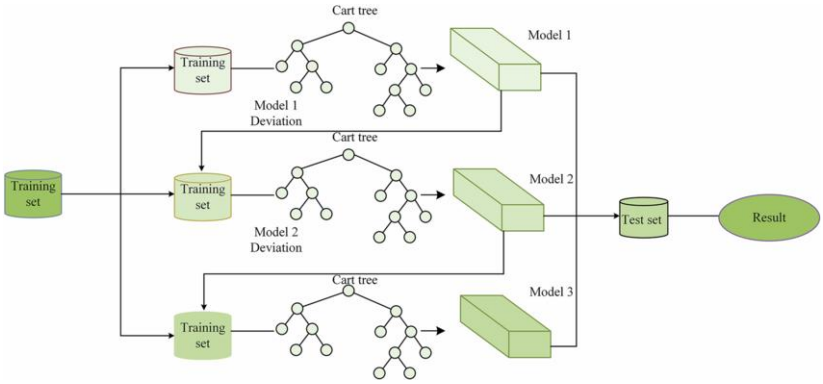
No	Peneliti	Judul	Metode Penelitian	Hasil Penelitian	Perbedaan
		Learning			*XGBoost* secara spesifik
3	(Sajiwo et al., 2024)	Klasifikasi Indeks Standar Pencemaran Udara (ISPU) Menggunakan Algoritma *XGBoost* dengan Teknik *Imbalanced* Data (SMOTE)	Algoritma *XGBoost* dengan penerapan teknik SMOTE	Hasil yang diperoleh menunjukkan akurasi sebesar 99,63%	Fokus penelitian ini adalah pada kualitas udara sedangkan penelitian saya pada kualitas air bersih. Selain itu, penelitian ini menggunakan SMOTE untuk mengatasi *imbalance* data, sedangkan penelitian saya fokus pada klasifikasi kualitas air bersih tanpa menggunakan teknik *balancing* data
4	(Sari et al., 2024)	Klasifikasi Kelayakan Sumber Air Minum di Jawa Barat Menggunakan *XGBoost* dan Analisis Klasterisasi Berdasarkan Nilai SHAP	*XGBoost*	Hasil penelitian menunjukkan akurasi sebesar 77,43% dan *F1-Score* sebesar 80,17%	Penelitian ini lebih menitikberatkan pada faktor sosio-ekonomi (seperti lokasi rumah, pengeluaran bulanan) sebagai variabel input, sementara penelitian saya fokus pada parameter fisik-kimia air untuk klasifikasi kualitas air bersih.
5	(Sutisna & Yuniar, 2023)	Klasifikasi Kualitas Air Bersih Menggunakan Metode *Naïve Bayes*	Metode *Naïve Bayes*	Hasil penelitian menunjukkan tingkat akurasi sebesar 97,35%.	Penelitian ini menggunakan dataset kecil dan metode *Naïve Bayes* tanpa optimasi *hyperparameter*, sementara penelitian saya menggunakan *XGBoost*

2.2 Klasifikasi dengan *XGBoost*

Klasifikasi merupakan salah satu teknik dalam *supervised learning* yang bertujuan untuk mengelompokkan data ke dalam kelas tertentu berdasarkan fitur-fitur yang ada. *Extreme Gradient Boosting* (*XGBoost*) merupakan salah satu algoritma ensemble berbasis pohon keputusan yang dikenal efisien dan akurat dalam menyelesaikan permasalahan klasifikasi. Chen & Guestrin (2016) memperkenalkan *XGBoost* sebagai sistem boosting pohon yang sangat efisien dan digunakan secara luas dalam kompetisi data *science* karena kecepatannya dan regularisasi internalnya yang mencegah *overfitting*.

XGBoost merupakan pengembangan dari algoritma *Gradient Boosting Decision Tree* (GBDT) dengan berbagai optimasi, termasuk regulasi tambahan untuk mencegah *overfitting*, pemrosesan paralel untuk meningkatkan efisiensi, serta kemampuan menangani data dengan jumlah besar secara lebih cepat dibandingkan metode lainnya seperti *Random Forest* atau *Support Vector Machine* (Chen & Guestrin, 2016).

Secara umum, algoritma *XGBoost* bekerja dengan membangun sejumlah pohon keputusan (*decision trees*) secara bertahap dalam proses *boosting*, di mana setiap pohon baru dibuat untuk memperbaiki kesalahan prediksi dari pohon sebelumnya. Proses ini dilakukan dengan pendekatan *gradient boosting*, yaitu menggunakan gradien dari fungsi *loss* untuk mengoptimalkan bobot model pada setiap iterasi.



Gambar 2. 1 Arsitektur Konsep *XGBoost*

Menurut Chen & Guestrin (2016), *XGBoost* memiliki beberapa keunggulan dibandingkan algoritma klasifikasi lainnya, antara lain:

a. Efisiensi Komputasi

XGBoost mampu melakukan komputasi secara paralel dan memanfaatkan alokasi memori yang lebih efisien, sehingga lebih cepat dibandingkan metode lain seperti GBDT biasa.

b. Regularisasi L1 dan L2

Algoritma ini menerapkan regulasi untuk menghindari *overfitting* dengan mengontrol kompleksitas model.

c. Kemampuan Menangani Data yang Hilang

XGBoost dapat menangani nilai yang hilang (*missing values*) dengan menentukan jalur optimal dalam pohon berdasarkan distribusi data.

d. Dukungan untuk *Hyperparameter* Tuning

Algoritma ini mendukung berbagai teknik pencarian *hyperparameter*, seperti *grid search* dan *random search*, untuk meningkatkan akurasi prediksi.

Fungsi *Loss* dan Regularisasi dalam *XGBoost* digunakan untuk mengukur kesalahan prediksi dan mengoptimalkan model. Salah satu fungsi *loss* yang digunakan dalam klasifikasi adalah *log loss* untuk klasifikasi biner:

$$\mathcal{L} = -\sum_{\{i=1\}}^{n} [\, y_i log(\widehat{y_i}) + \, (1 \, - \, y_i) log(1 \, - \, \widehat{y_i})] \tag{2.1}$$

Keterangan :
$\mathcal{L}$: Nilai error keseluruhan dari model alam memprediksi status kelayakan air
n : Jumlah total sampel data sensor air yang dievaluasi
y_i : Label aktual kualitas air (1 = "Layak", 0 = "Tidak Layak")
$\widehat{y_i}$: Probabilitas prediksi model terhadap kelayakan sampel air ke- i

Untuk mencegah *overfitting*, *XGBoost* menerapkan regulasi L1 (*Lasso Regularization*) dan L2 (*Ridge Regularization*) yang dirumuskan sebagai:

$$\Omega(f) = \, \gamma T \, + \frac{1}{2} \lambda \sum_{j} w_j^2 \tag{2.2}$$

Keterangan :
$\Omega(f)$: Nilai penalti untuk membatasi kompleksitas pohon dan mencegah overfitting
γ: Batas penalti untuk mengontrol pembentukan cabang keputusan baru
T : Jumlah daun keputusan (*leaf nodes*) pada sebuah pohon *XGBoost*
λ : Parameter regularisasi L2 (*Ridge regression*) pada bobot daun
$\sum_j w$: Penjumlahan untuk semua daun dari daun pertama hingga daun ke-T
w_j^2 : Nilai kuadrat dari bobot skor pada daun keputusan ke -j

di mana T adalah jumlah *leaf nodes* dalam pohon, w_j adalah bobot pada *leaf node* ke-j, dan γ serta λ adalah parameter regulasi.

Regularisasi ini membantu membatasi kompleksitas model dengan menambahkan penalti pada bobot yang besar, sehingga mengurangi risiko *overfitting* dan meningkatkan generalisasi model terhadap data baru. Penerapan regulasi ini sangat penting untuk memastikan model tidak hanya berkinerja baik pada data pelatihan, tetapi juga mampu menggeneralisasi dengan baik pada data baru yang belum pernah dilihat sebelumnya. Dalam konteks prediksi kualitas air,

XGBoost terbukti unggul. Misalnya, Elvin & Wibowo (2024) berhasil mencapai akurasi sebesar 97,06% pada prediksi kualitas air dengan optimasi *hyperparameter* yang tepat. Studi lain oleh Darshan & Subburaj (2023) membandingkan *XGBoost* dengan SVM dan menemukan bahwa *XGBoost* memiliki performa klasifikasi lebih tinggi terhadap kualitas air.

Algoritma *XGBoost* bekerja secara iteratif dengan membangun pohon keputusan secara bertahap. Setiap pohon bertujuan untuk memperbaiki kesalahan prediksi dari pohon sebelumnya menggunakan pendekatan *gradient boosting*. Berikut adalah tahapan utama dalam proses klasifikasi menggunakan *XGBoost* :

1. Inisialisasi Mode

   Model pertama dibangun dengan prediksi awal berdasarkan rata-rata target atau nilai awal tertentu.

2. Perhitungan Gradient dan Hessian

   Menghitung gradien dan *Hessian* dari fungsi *loss* untuk menentukan arah optimasi.

   Gradien dihitung menggunakan rumus:

$$g_i = \frac{\partial \mathcal{L}}{\partial \hat{y}_i} \tag{2.3}$$

Keterangan :
g_i: Nilai gradien untuk sampel data ke-*i*
$\hat{y}_i$: Nilai prediksi dari model untuk data ke-*i*
∂ : Simbol turunan parsial
$\mathcal{L}$: Fungsi kerugian

Hessian digunakan untuk mempercepat proses optimasi:

$$h_i = \frac{\partial^2 \mathcal{L}}{\partial \hat{y}_i^{\,2}} \tag{2.4}$$

Keterangan :
h_i : Nilai *Hessian* untuk sampel data ke
∂^2 : Simbol turunan parsial tingkat dua
$\mathcal{L}$: Fungsi kerugian
$\hat{y}_i$: Nilai prediksi dari model untuk data ke

3. Pembangunan Pohon Keputusan Baru

Setiap pohon baru dibuat dengan mencari pemisahan terbaik (*best split*) berdasarkan *Gain Function*:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \tag{2.5}$$

Keterangan :
$Gain$: Nilai evaluasi kelayakan pembelahan cabang berdasarkan fitur air
G_L: Total gradien g_i dari semua sampel air yang masuk ke cabang Kiri
G_R: Total gradien g_i dari semua sampel air yang masuk ke cabang Kanan
H_L: Total *hessian* h_i dari semua sampel air yang masuk ke cabang Kiri
H_R: Total *hessian* h_i dari semua sampel air yang masuk ke cabang Kanan
λ : Parameter regularisasi L2
γ : Nilai penalti minimal untuk membuat cabang baru (batas ambang)

4. Perhitungan Nilai *Leaf Node*

Setelah pohon selesai dibuat, nilai untuk setiap *leaf node* dihitung dengan:

$$w_j = -\frac{G_j}{H_j + \lambda} \tag{2.6}$$

Keterangan :
w_j : Bobot atau nilai prediksi optimal untuk daun ke
G_j : Total gradien dari seluruh sampel air yang masuk ke dalam daun ke-*j*
H_j : Total *Hessian* dari seluruh sampel air yang masuk ke dalam daun ke-*j*
λ : Parameter regularisasi L2 (mengontrol ukuran bobot agar tidak terlalu besar)

5. Pembaruan Model

Hasil dari pohon keputusan baru ditambahkan ke model sebelumnya menggunakan *learning rate* (η):

$$\hat{y}_i^{new} = \hat{y}_i^{\{old\}} + \eta w_j \quad (2.7)$$

Keterangan :
$\hat{y}_i^{new}$: Nilai prediksi baru sampel data ke-*i* setelah ditambahkan pohon baru
$\hat{y}_i^{\{old\}}$: Nilai prediksi lama
η : *Learning rate*
w_j : Bobot daun optimal dari pohon baru tempat sampel air ke-*i* mendarat

6. Iterasi hingga Model Optimal

Proses ini diulang hingga model mencapai jumlah pohon maksimum atau konvergen.

7. Prediksi Akhir

Hasil akhir diperoleh dengan menjumlahkan kontribusi dari semua pohon:

$$\hat{y} = \sum_{t=1}^{T} f_t(x) \quad (2.8)$$

Keterangan :
$\hat{y}$: Hasil prediksi akhir berupa nilai mentah kelayakan air
$\sum_{t=1}^{T}$. : Penjumlahan total prediksi pohon pertama hingga pohon terakhir
T : Jumlah total pohon keputusan yang dibangun oleh algoritma
t : Indeks untuk pohon keputusan yang sedang dihitung
$f_t(x)$: Hasil tebakan pohon ke-*t* berdasarkan input parameter sensor *x*

Untuk klasifikasi biner, probabilitas dihitung dengan fungsi sigmoid:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-\hat{y}}} \quad (2.9)$$

Keterangan :
e : Konstanta Euler atau bilangan basis logaritma alami (2.718)
$\hat{y}$: Nilai prediksi mentah hasil jumlah seluruh pohon keputusan

$P(y = 1 \mid x)$: Probabilitas target *y* diprediksi "Layak"(1) berdasarkan kondisi parameter fisik *x* yang terukur oleh sensor

Untuk klasifikasi multikelas, digunakan fungsi *softmax*:

$$P(y = k \mid x) = \frac{e^{\hat{y}_k}}{\sum_j e^{\hat{y}_j}} \quad (2.10)$$

Keterangan :
$P(y = k \mid x)$: Probabilitas bahwa input fitur *x* termasuk ke dalam kelas *k*
k : Indeks kelas spesifik yang sedang dihitung probabilitasnya
e : Konstanta *Euler* atau bilangan basis logaritma alami
$\hat{y}_k$: Nilai prediksi mentah untuk kelas spesifik *k*
$\hat{y}_j$: Nilai prediksi mentah untuk kelas ke-*j*

2.3 Evaluasi Model Klasifikasi

Dalam algoritma *machine learning*, kinerja sebuah model klasifikasi perlu diukur menggunakan metrik evaluasi yang tepat untuk mengetahui seberapa baik model tersebut membedakan antar kelas. Evaluasi model klasifikasi umumnya didasarkan pada *Confusion Matrix*, yaitu sebuah tabel representasi yang membandingkan hasil prediksi model dengan nilai aktual dari data uji (Sajiwo et al., 2024).

Untuk mengukur performa model setelah proses *training*, digunakan beberapa metrik evaluasi yang diturunkan berdasarkan nilai dari *Confusion Matrix* guna mengukur tingkat keandalan algoritma (Sari et al., 2024), yaitu sebagai berikut:

Akurasi presentase prediksi yang benar dibanding jumlah total datas

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.1)$$

Keterangan :
TP : Jumlah sampel air berkategori "Layak" yang diprediksi benar
TN : Jumlah sampel air berkategori "Tidak Layak" yang diprediksi benar
FP : Jumlah air berkategori "Tidak Layak" yang keliru diprediksi sebagai "Layak"
FN : Jumlah air berkategori "Layak" yang keliru diprediksi sebagai "Tidak Layak"

Precision kemampuan model dalam memprediksi kelas positif dengan benar

$$Precision = \frac{\text{TP}}{TP + FP} \quad (3.\ 2)$$

Keterangan :
$Precision$: Presisi
TP : Jumlah sampel air berkategori "Layak" yang diprediksi benar
FP : Jumlah air berkategori "Tidak Layak" yang keliru diprediksi sebagai "Layak"

Recall seberapa baik model mengenali seluruh data positif

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

Keterangan :
TP : Jumlah sampel air berkategori "Layak" yang diprediksi benar
FN : Jumlah air berkategori "Layak" yang keliru diprediksi sebagai "Tidak Layak"
$Recall$: Rasio seluruh sampel air "Layak" aktual yang berhasil dideteksi dengan benar

F1-Score yaitu rata-rata harmonik dari *Precision* dan *Recall*

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3.\ 4)$$

Keterangan :
$F1$: *F1-Score* nilai rata-rata harmonik yang menyeimbangkan Presisi dan Recall
$Precision$: Persentase tebakan air "Layak" yang memang benar-benar layak
$Recall$: Nilai *Recall* atau sensitivitas dari hasil model

Confusion Matrix yaitu untuk melihat kesalahan klasifikasi pada tiap kelas.

2.4 Kualitas Air

Kualitas air merupakan salah satu faktor penting dalam berbagai aspek kehidupan, termasuk kesehatan manusia, pertanian, dan ekosistem perairan. Kualitas air dapat dinilai berdasarkan beberapa parameter fisik, kimia, dan biologi yang mencerminkan kondisi suatu sumber air dan kelayakannya untuk digunakan dalam berbagai keperluan (*World Health Organization*, 2022). Penilaian kualitas

air yang akurat menjadi dasar dalam pengelolaan sumber daya air berkelanjutan dan mitigasi risiko kesehatan masyarakat.

Kualitas air dipengaruhi oleh berbagai parameter seperti pH, suhu, kekeruhan, dan TDS. Mengingat pentingnya sumber air yang bersih untuk keberlangsungan hidup dan pertanian, pemantauan dan prediksi kualitas air menjadi prioritas. Grbčić et al. (2022) menggabungkan model *machine learning* termasuk *XGBoost* dengan analisis spasial dan temporal untuk memprediksi kualitas air pesisir secara akurat. Selain itu, Khoi et al. (2022) memanfaatkan model ini untuk memprediksi *Water Quality Index* (*WQI*) pada sungai di Vietnam, dengan hasil yang sangat akurat dan efisien.

2.4.1 Parameter kualitas air

Beberapa parameter utama dalam menilai kualitas air meliputi:

a. PH (Derajat Keasaman)

PH menunjukkan tingkat keasaman atau kebasaan suatu larutan dan mempengaruhi ketersediaan unsur hara dalam air serta tanah. Nilai pH yang ideal untuk konsumsi manusia dan pertumbuhan tanaman berkisar antara 6,5 hingga 8,5 (Kementerian Kesehatan RI, 2022). pH yang terlalu rendah dapat menyebabkan korosi pada pipa distribusi air, sedangkan pH yang terlalu tinggi dapat mengurangi efektivitas desinfektan dalam air minum (*World Health Organization*, 2022).

b. Suhu (°C)

Suhu air berpengaruh terhadap keseimbangan ekosistem akuatik dan metabolisme organisme di dalamnya. Air dengan suhu tinggi cenderung memiliki kadar oksigen terlarut yang lebih rendah, yang dapat menyebabkan stres pada biota

perairan. Selain itu, dalam sektor pertanian, suhu air yang terlalu tinggi dapat menghambat proses penyerapan nutrisi oleh tanaman, sementara suhu yang terlalu rendah dapat memperlambat pertumbuhan tanaman. Kisaran suhu optimal untuk air pertanian dan ekosistem akuatik berkisar antara 20-30°C (Kementerian Lingkungan Hidup, 2022).

c. Kekeruhan (NTU - Nephelometric Turbidity Units)

Kekeruhan menunjukkan jumlah partikel tersuspensi dalam air, yang sering kali berasal dari sedimen, limbah organik, atau mikroorganisme. Air dengan tingkat kekeruhan tinggi dapat mengindikasikan adanya pencemaran yang berpotensi mengganggu kesehatan manusia. Menurut standar WHO, batas maksimal kekeruhan untuk air minum adalah 5 NTU (Kemenkes, 2023).

d. Total Dissolved Solids (TDS) - mg/L

TDS mengukur jumlah zat padat terlarut dalam air, termasuk mineral, garam, dan logam berat. Nilai TDS yang tinggi dapat mempengaruhi rasa air dan berkontribusi terhadap kerak pada peralatan rumah tangga. WHO merekomendasikan batas aman TDS untuk air minum adalah kurang dari 500 mg/L (*World Health Organization*, 2022).

2.4.2 Standar Kualitas Air Bersih

Standar kualitas air bersih umumnya ditetapkan berdasarkan baku mutu parameter fisis dan kimia yang diatur dalam peraturan Kementerian Kesehatan Republik Indonesia maupun rekomendasi *World Health Organization* (WHO). Parameter yang paling sering digunakan dalam penilaian kualitas air bersih meliputi pH, suhu, kekeruhan, dan *Total Dissolved Solids* (TDS).

Baku mutu tersebut antara lain:

- pH: 6,5 – 8,5 (Kemenkes, 2022; WHO, 2022)
- Suhu: 20 – 30 °C (Kementerian Lingkungan Hidup, 2022)
- Kekeruhan: ≤ 3 NTU (Kemenkes, 2023; WHO, 2022)
- TDS: ≤ 300 mg/L (WHO, 2022)

Dalam penelitian ini, penilaian kualitas air bersih tidak menggunakan perhitungan Indeks Kualitas Air (IKA) penuh, mengingat parameter yang tersedia terbatas hanya pada pH, suhu, kekeruhan, dan TDS. Oleh karena itu, klasifikasi kualitas air bersih dilakukan dengan membandingkan nilai parameter hasil pengukuran terhadap standar baku mutu yang telah ditetapkan

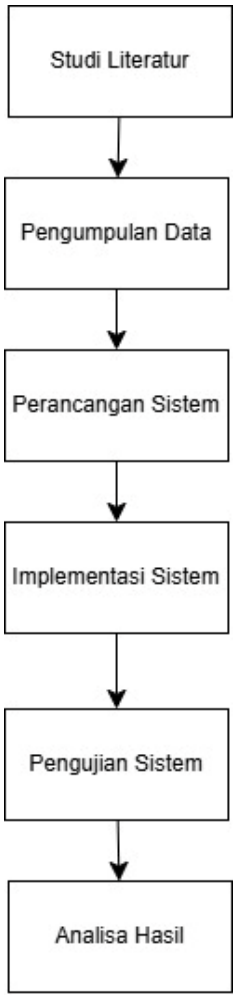
.

BAB III

METODE PENELITIAN

3.1 Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan penerapan algoritma *Extreme Gradient Boosting (XGBoost*) untuk melakukan klasifikasi kualitas air bersih di Kabupaten Bojonegoro. Desain penelitian disusun secara sistematis melalui serangkaian tahapan yang mencakup pengumpulan data, pengolahan data, pelatihan model, evaluasi, serta analisis hasil klasifikasi.



Gambar 3. 1 Alur Penelitian

Tahap awal dilakukan melalui studi literatur guna mengidentifikasi teori, metode, dan penelitian terdahulu yang relevan dengan klasifikasi kualitas air dan penerapan *XGBoost*. Selanjutnya dilakukan pengambilan data primer di lapangan untuk memperoleh parameter kualitas air berdasarkan pengukuran langsung. Data tersebut kemudian diproses melalui tahapan *preprocessing*, meliputi pembersihan data dari nilai kosong (*missing values*), normalisasi, dan pembagian dataset.

Setelah data siap, dilakukan pelatihan model (*training*) menggunakan algoritma *XGBoost* dengan sejumlah parameter dasar (`learning_rate`, `max_depth`, `n_estimators`, dan `subsample`). Proses validasi model menggunakan metode 5-*Fold Cross Validation* untuk mengukur kemampuan generalisasi model terhadap data baru.

Tahap akhir adalah evaluasi performa model menggunakan *confusion matrix* serta metrik pengukuran performa seperti *Accuracy, Precision, Recall,* dan *F1-Score*, guna menilai tingkat keberhasilan model dalam mengklasifikasikan kualitas air bersih.

3.2 Pengumpulan Data

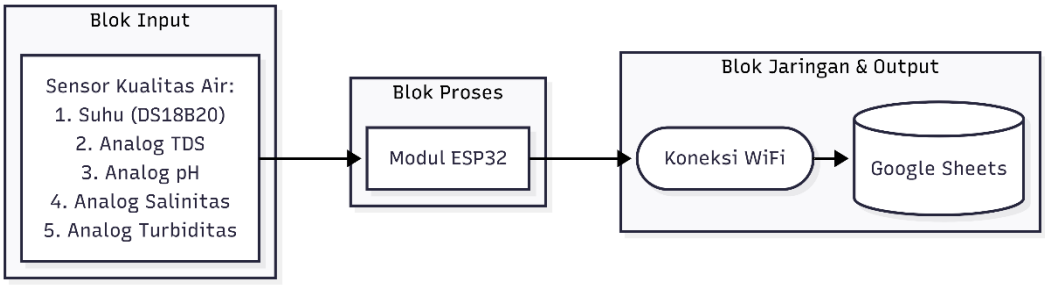
Penelitian ini menggunakan data primer yang dikumpulkan melalui metode observasi dan pengukuran langsung di lapangan berbasis runtun waktu (*time-series*). Pendekatan ini dipilih untuk mendapatkan gambaran karakteristik air yang komprehensif di setiap titik sampel, dengan menangkap nilai fluktuasi mikro dalam interval waktu yang singkat. Proses pengumpulan data ini dibagi menjadi spesifikasi instrumen yang digunakan dan protokol pengambilannya.

3.2.1 Spesifikasi Instrumen Pengambilan Data

Proses akuisisi data primer pada penelitian ini dilakukan menggunakan instrumen pengukur kualitas air terintegrasi *Internet of Things* (IoT). Perangkat keras ini difasilitasi oleh pihak pembimbing atau laboratorium yang dirancang khusus untuk memantau kelayakan air secara *real-time*. Arsitektur instrumen dan sistem transmisi data nirkabel pada perangkat ini mengadaptasi teknologi yang telah divalidasi keandalannya dalam penelitian Kurniawan et al. (2026). Instrumen ini dikendalikan oleh mikrokontroler ESP32 sebagai pusat pemroses data, serta memanfaatkan panel surya sebagai sumber daya mandiri saat beroperasi di lapangan.

Untuk mengukur parameter fisis dan kimia air, instrumen ini dilengkapi dengan empat sensor utama, yaitu:

1. Sensor Suhu Digital (DS18B20): Untuk mendeteksi temperatur air secara presisi.
2. Sensor pH Analog: Untuk mengukur tingkat derajat keasaman atau kebasaan air.
3. Sensor *Total Dissolved Solids* (TDS): Untuk mengukur jumlah zat padat terlarut dalam satuan mg/L.
4. Sensor Turbiditas (Kekeruhan): Untuk mengukur tingkat kejernihan air dalam satuan NTU.



Gambar 3. 2 Blok Diagram Sistem IoT

3.2.2 Skenario dan Protokol Pengambilan Data Lapangan

Proses pengambilan sampel dilakukan di tiga lokasi sumber air sumur bor milik warga yang tersebar di Desa Baureno, Desa Pasinan, dan Desa Seratu (Sratu Rejo). Pemilihan lokasi menggunakan metode *purposive* sampling berdasarkan temuan laporan Dinas Lingkungan Hidup Kabupaten Bojonegoro (2023) yang mengindikasikan adanya permasalahan kelayakan air di wilayah tersebut.

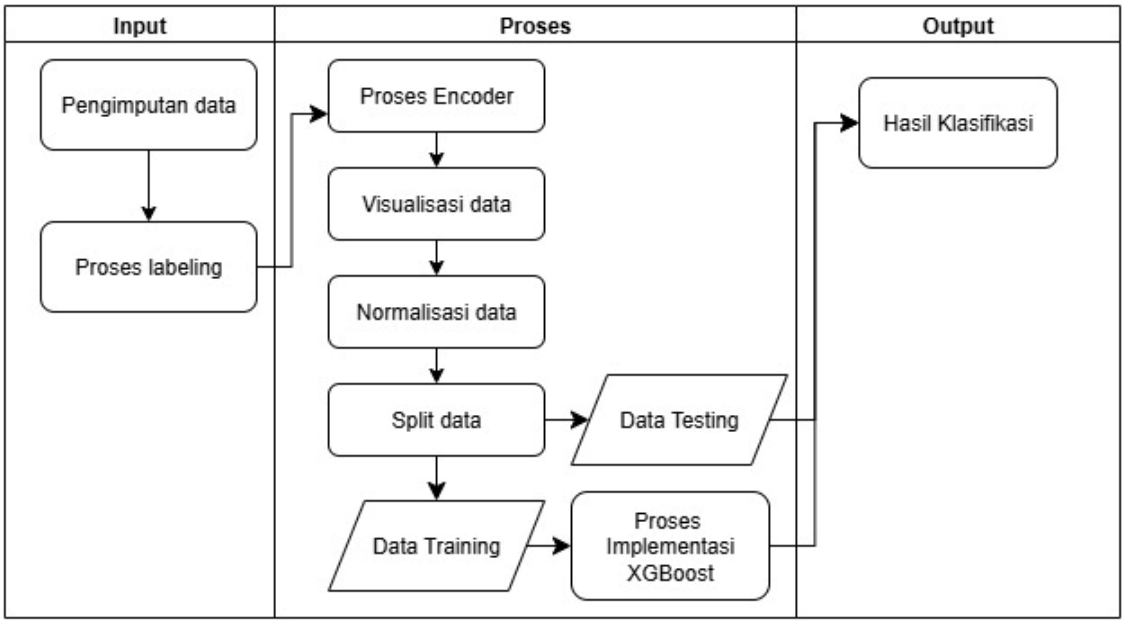
Karena karakteristik sumur bor memiliki lubang pipa yang sempit, instrumen tidak memungkinkan untuk dicelupkan langsung ke dalam sumber mata air. Oleh karena itu, prosedur pengambilan data diadaptasi menggunakan metode on-site sampling, di mana air sumur ditarik dan ditampung ke dalam wadah ukur dengan volume representatif sebanyak 2 hingga 3 liter. Ujung probe sensor kemudian dicelupkan ke dalam sampel air tersebut. Instrumen secara otomatis membaca dan merekam parameter air setiap 4 detik. Proses perekaman ini dibiarkan berjalan hingga berhasil mengumpulkan akumulasi total 3.511 baris data observasi. Seluruh data hasil pembacaan dikirimkan secara nirkabel oleh mikrokontroler menuju basis data *Google Sheets* menggunakan koneksi WiFi, untuk kemudian diekspor ke dalam format CSV agar siap digunakan pada tahap komputasi machine learning.

Selain prosedur teknis, perancangan skenario akuisisi data ini juga sangat mempertimbangkan kondisi lingkungan saat pengukuran. Pelaksanaan pengambilan data lapangan dilakukan tepat pada rentang tanggal 4 hingga 7 Mei 2026 di Kabupaten Bojonegoro. Berdasarkan prakiraan BMKG dan laporan cuaca lokal, periode tersebut merupakan fase awal musim kemarau yang secara umum berlangsung dari April hingga Oktober, dengan puncak kemarau diperkirakan terjadi pada bulan Agustus (BMKG Jawa Timur, 2025).

Kondisi iklim ini dicatat secara khusus karena transisi musim sangat memengaruhi parameter kualitas air. Penurunan debit air sumur pada musim kemarau berpotensi meningkatkan konsentrasi padatan terlarut (TDS). Di sisi lain, fenomena anomali cuaca berupa hujan lokal sesekali masih terjadi di wilayah Bojonegoro pada periode ini akibat variasi atmosfer, yang mana limpasan airnya juga berpotensi memengaruhi kondisi sumber air tanah (Shaqia, 2025). Dengan demikian, pencatatan dinamika cuaca ini dirancang untuk memberikan gambaran konteks lingkungan yang lebih komprehensif dalam proses penelitian.

3.3 Implementasi *XGBoost*

Pada tahap perancangan sistem, proses klasifikasi kualitas air bersih dilakukan melalui beberapa langkah utama, dimulai dari input data, *preprocessing* data, pembagian dataset, *training* model, hingga klasifikasi akhir menggunakan metode *XGBoost* dengan optimasi *hyperparameter*.



Gambar 3. 3 Desain Sistem

3.3.1 Input Data

Tahapan paling awal dalam alur desain sistem klasifikasi adalah penerimaan input data. Data masukan yang digunakan oleh algoritma berupa dataset kualitas air berformat CSV (*Comma Separated Values*). Dataset ini merupakan hasil ekspor (keluaran) dari proses akuisisi data lapangan menggunakan instrumen IoT yang telah diuraikan pada tahap pengumpulan data (Sub-bab 3.2).

Sebagai bahan baku *machine learning*, dataset ini murni berisi rekaman angka dari sensor tanpa ada campur tangan perhitungan manual. Dataset masukan ini terdiri atas 3.511 baris data yang memuat empat atribut parameter fisik dan kimiawi air, yaitu nilai derajat keasaman (pH), suhu air (°C), tingkat kekeruhan (Turbiditas/NTU), dan jumlah padatan terlarut (TDS/mg/L).

Tabel 3.1 memperlihatkan cuplikan struktur dari dataset kualitas air yang bertindak sebagai input .

Tabel 3. 1 Dataset Air

Indeks	**pH**	**Suhu(°C)**	**Kekeruhan (NTU)**	***Total Dissolved Solids*(mg/L)**
1	6.8	25.3	1.5	150
2	7.2	26.1	2.1	180
3	6.5	24.8	3.0	210
4	7.0	25.5	1.8	160
5	7.5	26.7	0.9	140
6	6.9	24.5	2.5	190
7	7.1	25.9	1.2	170

3.3.2 Preprosesing Data

Preprocessing data bertujuan untuk memastikan bahwa data yang digunakan dalam model *XGBoost* telah bersih, terstruktur, dan siap untuk dianalisis. Proses *preprocessing* mencakup penanganan nilai kosong (*missing values*), normalisasi, serta *encoding* variabel kategori jika diperlukan. (Hidayaturrohman & Hanada, 2024) menunjukkan bahwa teknik *preprocessing* seperti imputasi dan standarisasi dapat secara signifikan meningkatkan performa model *XGBoost* dalam prediksi data kesehatan, yang prinsipnya juga dapat diterapkan dalam kasus klasifikasi kualitas air.

1. Pembersihan Data

Langkah pertama dalam *preprocessing* adalah pembersihan data, di mana data yang tidak lengkap (*missing values*) atau mengandung *outlier* yang tidak wajar akan dihapus atau ditangani agar tidak mengganggu hasil analisis.

2. Labeling data

Proses *labeling* data dilakukan dengan memberikan kategori kualitas air bersih berdasarkan standar baku mutu yang ditetapkan oleh Kementerian Kesehatan (2022) dan *World Health Organization* (2022). Empat parameter yang digunakan adalah pH, suhu, kekeruhan, dan TDS.

Aturan kategorisasi kualitas air ditetapkan menjadi dua, yakni kategori layak apabila seluruh parameter memenuhi standar baku mutu, dan kategori tidak layak jika terdapat minimal satu parameter yang melebihi batas standar tersebut. Melalui pendekatan ini, setiap data hasil pengukuran dapat diberikan label kategori kualitas air bersih secara langsung tanpa perlu menggunakan perhitungan indeks gabungan. Label inilah yang selanjutnya akan digunakan sebagai variabel target dalam proses klasifikasi menggunakan algoritma *XGBoost*.

3. Encoding

Selanjutnya, dilakukan *encoding* untuk mengubah data kategorikal menjadi format numerik yang dapat dipahami oleh model. Misalnya, jika terdapat atribut kondisi air dalam bentuk kategori seperti “Layak”, atau “Tidak Layak”, maka data ini dikonversi menggunakan metode label *encoding* atau *one-hot encoding* agar dapat diproses lebih lanjut dalam model.

4. Normalisasi atau standarisasi fitur

Tahap terakhir dalam *preprocessing* adalah normalisasi fitur. Parameter numerik seperti pH, Suhu, TDS, dan Kekeruhan memiliki rentang nilai yang berbeda. Untuk memastikan tidak ada satu fitur pun yang mendominasi proses pelatihan model hanya karena skalanya yang lebih besar, akan diterapkan normalisasi menggunakan metode *Min-Max Scaling*. Metode ini akan mengubah skala setiap fitur ke dalam rentang [0, 1] sambil mempertahankan distribusi asli dari data.

5. Pembagian Dataset

Setelah tahap pra-pemrosesan, dataset dibagi menjadi data latih dan data uji secara acak untuk menghindari bias. Penelitian ini menggunakan empat skenario rasio pembagian data, yaitu 90:10, 80:20, 70:30, dan 60:40. Variasi rasio ini bertujuan untuk menguji pengaruh volume data latih terhadap performa model *XGBoost*. Data latih digunakan untuk pembelajaran algoritma dan optimasi *hyperparameter*, sedangkan data uji digunakan murni untuk mengukur akurasi akhir model.

3.3.3 *Training* Model

Pada tahap ini, model klasifikasi *Extreme Gradient Boosting* (*XGBoost*) akan dibangun dan dilatih menggunakan data latih (*training set*) yang telah melalui tahap *preprocessing XGBoost* dipilih karena kemampuannya dalam menangani data dengan jumlah fitur yang beragam serta kemampuannya dalam menangani *overfitting* melalui mekanisme regularisasi L1 dan L2. Selain itu, model *XGBoost* mampu memberikan hasil klasifikasi yang optimal dengan memanfaatkan teknik *boosting*, di mana setiap pohon keputusan dalam model diperbaiki secara iteratif berdasarkan kesalahan sebelumnya. Implementasi akan dilakukan menggunakan bahasa pemrograman *Python* dengan pustaka utama *Scikit-learn* dan *XGBoost* .

Untuk mengevaluasi performa model klasifikasi dan menguji kemampuan generalisasi algoritma, penelitian ini menerapkan skenario pengujian yang difokuskan pada variasi pembagian rasio data latih dan data uji. Skenario pengujian yang dilakukan meliputi empat rasio pembagian dataset, yaitu::

1. Skenario Pembagian Data 90:10 (90% Data Latih : 10% Data Uji)

2. Skenario Pembagian Data 80:20 (80% Data Latih : 20% Data Uji)
3. Skenario Pembagian Data 70:30 (70% Data Latih : 30% Data Uji)
4. Skenario Pembagian Data 60:40 (60% Data Latih : 40% Data Uji)

Pada masing-masing skenario pembagian data tersebut, model bertujuan untuk mencapai performa klasifikasi setinggi mungkin. Untuk mencapai hal ini, pada tiap skenario akan dilakukan proses optimasi *hyperparameter* menggunakan teknik *GridSearchCV* dari pustaka *Scikit-learn*. *GridSearchCV* dipilih karena kemampuannya untuk secara sistematis menguji dan mengevaluasi semua kemungkinan kombinasi dari rentang nilai *hyperparameter* yang telah didefinisikan.

Proses pencarian ini divalidasi secara internal menggunakan metode 5-*Fold Cross-Validation* pada subset data latih untuk memastikan bahwa kombinasi *hyperparameter* terbaik yang ditemukan bersifat tangguh dan terhindar dari *overfitting*. Tabel 3.2 berikut merinci *hyperparameter* yang akan dioptimalkan beserta rentang nilai yang akan diuji.

Tabel 3. 2 Desain Eksperimen Optimasi *Hyperparameter*

Hyperparameter	**Deskripsi**	**Rentang Nilai yang Diuji**
`n_estimators`	Jumlah pohon keputusan dalam model.	[100, 200, 300]
`learning_rate`	Laju pembelajaran yang mengontrol kontribusi setiap pohon baru.	[0.01, 0.1, 0.2]
`max_depth`	Kedalaman maksimum dari setiap pohon keputusan untuk mengontrol kompleksitas model.	[3, 5, 7]
`subsample`	Proporsi sampel data latih yang digunakan untuk membangun setiap pohon.	[0.8, 0.9]
`colsample_bytree`	Proporsi fitur yang digunakan saat membangun setiap pohon.	[0.8, 0.9]

Setelah proses *GridSearchCV* selesai pada masing-masing skenario pembagian data, kombinasi *hyperparameter* yang menghasilkan skor akurasi *cross-*

validation tertinggi akan dipilih sebagai arsitektur terbaik. Kombinasi optimal inilah yang kemudian digunakan untuk melatih model *XGBoost* final pada keseluruhan porsi data latih di tiap rasio. Model final dari masing-masing rasio ini yang kemudian akan dievaluasi kinerjanya menggunakan data uji (testing set) yang belum pernah dilihat sebelumnya, lalu dibandingkan untuk menentukan skenario pembagian data mana yang menghasilkan model paling akurat dan stabil.

3.3.4 Evaluasi Model

Tahap terakhir dalam perancangan sistem ini adalah mengevaluasi performa model klasifikasi *XGBoost* yang telah dilatih. Evaluasi dilakukan dengan membandingkan hasil prediksi kelayakan air dari model terhadap label aktual pada data uji. Pengukuran kinerja model ini dianalisis menggunakan instrumen *Confusion Matrix* guna mendapatkan nilai metrik *Accuracy, Precision, Recall*, dan *F1-Score*. Penjelasan teoretis serta formulasi matematis dari masing-masing metrik evaluasi tersebut telah diuraikan sebelumnya pada Bab 2.

3.4 Perhitungan Manual metode *XGBoost*

Perhitungan manual pada metode *XGBoost* dimulai dengan inisialisasi model awal, di mana prediksi pertama ditentukan dari nilai rata-rata target dalam data.

Tabel 3. 3 Contoh Dataset Kualitas Air

ID	pH	Suhu (°C)	Kekeruhan (NTU)	TDS (mg/L)	Label (y)
1	6.8	25.3	1.5	150	1
2	7.2	26.1	2.1	180	1
3	6.5	24.8	3.0	210	0
4	7.0	25.5	1.8	160	1
5	7.5	26.7	0.9	140	0

Sebagai contoh, untuk data yang ada , maka prediksi awal model adalah:

$$\hat{y} = \frac{1+1+0+1+0}{5} = 0.6$$

Keterangan:
$\hat{y}$: prediksi awal

Rata-rata ini digunakan untuk memulai proses boosting. Setelah prediksi awal ditetapkan, langkah berikutnya adalah menghitung gradien dan Hessian. Dalam *XGBoost*, turunan pertama (gradien) dan turunan kedua (Hessian) dari fungsi loss log-loss digunakan untuk klasifikasi biner.

Gradien (g_i) dihitung dengan formula:

$$g_i = \hat{y}_i - y_i \quad (3.5)$$

Keterangan:
g_i : gradient untuk data ke-i
$\hat{y}_i$: nilai prediksi model
y_i : label sebenarnya

Hessian (h_i) dihitung dengan:

$$h_i = \hat{y}_i(1 - \hat{y}_i) \quad (3.6)$$

Keterangan:
h_i : *Hessian* untuk data ke-i
Hessian digunakan dalam perhitungan regulasi pohon

Hasil perhitungan:

Tabel 3. 4 Hasil Perhitungan Manual *XGBoost*

ID	y (Label)	Prediksi y_i	*Gradient* g_i	*Hessian* h_i
1	1	0.6	-0.4	0.24
2	1	0.6	-0.4	0.24
3	0	0.6	0.6	0.24
4	1	0.6	-0.4	0.24
5	0	0.6	0.6	0.24

Nilai-nilai ini menjadi dasar untuk pembangunan pohon keputusan, di mana pencarian pemisahan terbaik dilakukan berdasarkan *Gain Function*.

$$Gain = \frac{1}{2}\left[\frac{(G_L)^2}{H_L + \lambda} + \frac{(G_R)^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \quad (3.7)$$

Keterangan:
G_L, H_L : total *gradient* dan *hessian* untuk bagian kiri setelah *split*
G_R, H_R : total *gradient* dan *hessian* untuk bagian kanan setelah *split*
G_{Total}, H_{Total} : total sebelum *split*
λ : parameter regulasi (*default* 1)
γ : penalti *split* untuk mengakhiri *overfitting* (asumsi 0.1)

Berikut contoh perhitungan jika *split* pada pH = 7.0:

GL (pH < 7.0) = -0.4 + 0.6 = 0.2
HL = 0.24 + 0.24 = 0.48
GR (pH ≥7.0) = -0.4 + (-0.4) + 0.6 = -0.2
HR = 0.24 + 0.24 + 0.24 = 0.72
GTotal = 0.2 + (-0.2) = 0
Htotal = 0.48 + 0.72 = 1.2

Jika Gain > 0, split dilakukan jika tidak, maka split lain diuji.

Setelah pemisahan optimal ditemukan, proses dilanjutkan dengan menghitung nilai untuk setiap leaf node. Bobot (w_j) untuk setiap *leaf* dihitung berdasarkan *gradien* dan *hessian*.

$$w_j = -\frac{G_j}{H_j + \lambda} \quad (3.8)$$

Keterangan:
w_j adalah bobot untuk leaf ke-*j*

Perhitungan untuk *leaf* kiri (pH < 7.0):

$$w_L = -\frac{0.2}{0.48 + 1} = -0.135$$

Perhitungan untuk *leaf* kanan (pH ≥ 7.0):

$$w_R = -\frac{-0.2}{0.72 + 1} = 0.116$$

Terakhir, model diperbarui dengan menambahkan pohon baru yang telah dibuat. Prediksi baru dihitung menggunakan learning rate(η).

$$\hat{y}_{new} = \hat{y}_{old} + \eta . w_j \quad (3.\ 9)$$

Keterangan:
η adalah learning rate (misalkan 0.3)

Jika data masuk ke *leaf* kiri:

$$\hat{y}_{new} = 0.6 + (0.3 \times -0.135) = 0.56$$

Jika data masuk ke *leaf* kanan:

$$\hat{y}_{new} = 0.6 + (0.3 \times 0.116) = 0.635$$

Proses iteratif ini terus diulang hingga jumlah maksimum pohon tercapai atau model mencapai konvergen.

BAB IV

UJI COBA DAN PEMBAHASAN

4.1 Implementasi Penelitian

4.1.1 Perangkat Lunak

Implementasi sistem klasifikasi kualitas air bersih ini dibangun menggunakan bahasa pemrograman *Python* versi 3.12. Proses penulisan kode, pengolahan data, hingga pelatihan model dilakukan pada lingkungan berbasis *cloud* yaitu *Google Colab*. Beberapa pustaka (*library*) utama *Python* yang digunakan dalam penelitian ini meliputi:

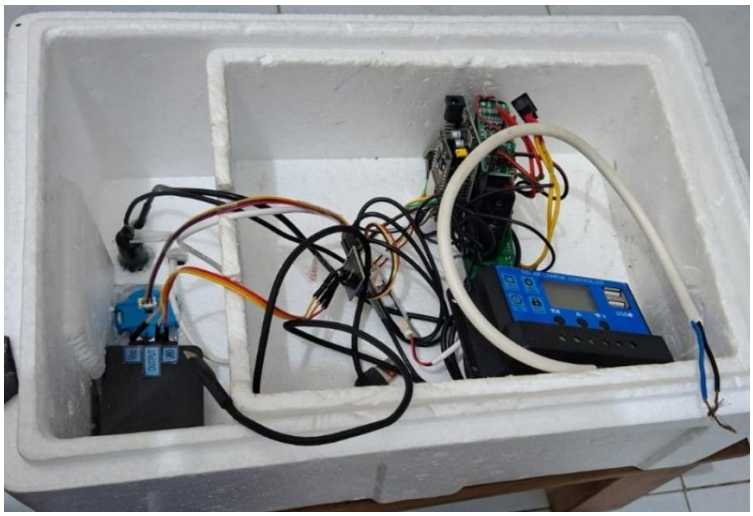
- *Pandas* dan *NumPy*: Untuk pembersihan, dan analisis struktur data numerik.
- *Scikit-Learn*: Untuk proses pembagian dataset (data *splitting*) dan perhitungan metrik evaluasi seperti confusion matrix.
- *XGBoost* (*Extreme Gradient Boosting*): Sebagai algoritma utama dalam pembentukan model klasifikasi kualitas air.

4.1.2 Implementasi Perangkat Keras Instrumen IoT

Implementasi perangkat keras bertujuan untuk memvalidasi instrumen yang digunakan dalam mengakuisisi data primer di lapangan. Walaupun instrumen ini merupakan fasilitas yang disediakan oleh pihak pembimbing, dokumentasi fisik dan uji fungsi tetap disertakan untuk memperlihatkan kesiapan operasional perangkat sebelum diterjunkan ke lokasi observasi.

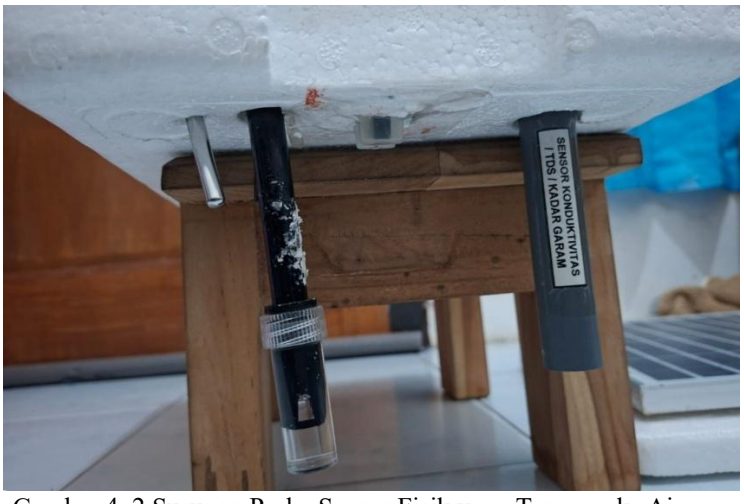
Secara fisik, keseluruhan komponen elektronik mikrokontroler dan modul transmisi dikemas di dalam kotak pelindung berbahan *styrofoam*. Penggunaan

material ini berfungsi ganda, yakni sebagai pelindung kedap cipratan air (*water-resistant*) sekaligus memberikan isolasi suhu dari lingkungan luar.



Gambar 4. 1 Rangkaian Komponen Internal Instrumen IoT

Berdasarkan Gambar 4.1, dapat dilihat konfigurasi internal dari perangkat pengukur. Mikrokontroler ESP32 diletakkan bersama dengan modul *Solar Charge Controller* berwarna biru dan baterai. Modul-modul transmitter pembaca sensor (berlabel *Vcc*, *OUTPUT*, GND) juga diletakkan di bagian dalam agar aman dari kelembapan.



Gambar 4. 2 Susunan Probe Sensor Fisik yang Terpapar ke Air

Sementara itu, komponen yang bersentuhan langsung dengan objek air ditempatkan menembus bagian bawah kotak (Gambar 4.2). Terdapat *probe* berbahan logam (sensor Suhu DS18B20), *probe* hitam transparan (sensor pH), dan *probe* berbentuk pipa PVC abu-abu yang merupakan modul sensor konduktivitas untuk membaca nilai *Total Dissolved Solids* (TDS).

Sebelum instrumen ini diterjunkan ke lokasi observasi, tahapan kalibrasi dilakukan secara mandiri untuk menjamin tingkat akurasi pembacaan masing-masing sensor. Proses kalibrasi untuk sensor suhu dilakukan dengan menggunakan media air mendidih guna menguji respons dan presisi sensor pada titik referensi suhu atas. Pada sensor pH, pengujian dilakukan menggunakan larutan sampel yang tingkat keasamannya telah divalidasi terlebih dahulu secara manual menggunakan alat ukur pH meter konvensional sebagai nilai rujukan (*ground truth*). Selanjutnya, untuk sensor konduktivitas (TDS) beserta sensor pendukung lainnya, kalibrasi dilakukan menggunakan medium larutan air garam dengan konsentrasi tertentu guna memastikan sensitivitas rentang pembacaan mineral terlarut beroperasi secara normal.

Instrumen pengukur kualitas air berbasis IoT yang difasilitasi dalam penelitian ini telah dilengkapi dengan mikrokontroler ESP32 yang terprogram untuk beroperasi secara otomatis di lapangan. Untuk memastikan proses perekaman dan pengiriman data berjalan secara presisi, sistem perangkat lunak (*firmware*) pada instrumen tersebut bekerja berdasarkan serangkaian tahapan logika algoritma. Tahapan ini mencakup mekanisme sinkronisasi waktu, pemfilteran nilai sensor, hingga transmisi data menuju pangkalan data *Google Sheets*.

1. Algoritma Sinkronisasi Waktu (*Timestamp*)

Mengingat penelitian ini berbasis *time-series*, setiap data yang direkam wajib memiliki stempel waktu yang akurat. Karena mikrokontroler tidak memiliki jam internal (*Real Time Clock*), sinkronisasi dilakukan melalui jaringan internet. Alur logika sinkronisasi waktu ini disajikan pada Gambar 4.3.

```
INPUT: WiFi Credential (SSID, Password)
OUTPUT: Formatted Timestamp (String)
INITIALIZE WiFiManager
INITIALIZE NTPClient (server: "pool.ntp.org", offset: +7)
WHILE TRUE DO
    IF WiFi.status() == CONNECTED THEN
        NTP_UPDATE()
        IF NTP_TIME == FAILED THEN
            formattedTimestamp = "N/A"
        ELSE
            formattedTimestamp = FORMAT_DATE("DD/MM/YYYY
HH:MM:SS")
        END IF
    END IF
END WHILE
```

Gambar 4. 3 *Pseudocode* Sinkronisasi Waktu via NTP

Algoritma ini memerlukan masukan berupa kredensial koneksi *WiFi* yang aktif. Setelah terhubung, sistem menjalankan proses inisialisasi layanan *NTPClient* untuk menarik data waktu dari *server* global. Sebagai keluaran, algoritma menghasilkan *formatted timestamp* yang akurat. Jika koneksi atau sinkronisasi gagal, sistem akan memberikan label "N/A" agar data yang tercatat di *server* tidak memberikan informasi waktu yang menyesatkan.

2. Algoritma Pembacaan dan Pemfilteran Sensor

Karakteristik sensor analog yang bersentuhan langsung dengan air sangat rentan terhadap gangguan kelistrikan (*noise*) yang memicu fluktuasi nilai fisis. Untuk menjaga objektivitas data, diterapkan algoritma pemfilteran dengan mengambil nilai tengah (*median*) dari sepuluh sampel pembacaan.

```
INPUT: Pin sensor (analog_pin)
OUTPUT: Nilai ukur (float)
CONSTANT SCOUNT = 10
ARRAY sensor_buffer[SCOUNT]
FOR i = 0 TO SCOUNT - 1 DO
    sensor_buffer[i] = ANALOG_READ(analog_pin)
    DELAY(10)
END FOR
SORT(sensor_buffer)
median_value = sensor_buffer[SCOUNT / 2]
final_value = CONVERT_TO_UNIT(median_value)
RETURN final_value
```

Gambar 4. 4 *Pseudocode* Pembacaan Sensor dengan Filter *Median*

Algoritma ini memiliki masukan berupa sinyal kelistrikan analog yang diterima dari pin sensor. Proses intinya adalah melakukan perulangan pengambilan data sebanyak sepuluh kali (*SCOUNT*), mengurutkan nilai tersebut, dan memilih nilai tengahnya. Sebagai keluaran, algoritma menghasilkan nilai parameter (TDS, pH, dan turbiditas) yang lebih stabil dan terbebas dari *noise* listrik sesaat, sehingga data yang dikirim ke pangkalan data lebih merepresentasikan kondisi air yang sebenarnya.

3. Algoritma Transmisi Data ke Pangkalan Data

Setelah parameter fisis air yang valid dan *timestamp* didapatkan, data tersebut kemudian dikemas dan dikirimkan secara nirkabel menuju *server* pangkalan data.

```
INPUT: Suhu, pH, TDS, Turbiditas, Timestamp, API_Key
OUTPUT: Data terekam di Google Sheets
CONSTANT interval = 4000 // dalam milidetik
waktu_sekarang = GET_SYSTEM_TIME()
IF (waktu_sekarang - waktu_terakhir_kirim) >= interval THEN
    IF WiFi.status() == CONNECTED AND Timestamp != "N/A" THEN
        // Konstruksi URL transmisi data
        URL = BASE_URL + "?api_key=" + API_Key + "&suhu=" + Suhu
              + "&ph=" + pH + "&tds=" + TDS + "&turb=" +
Turbiditas
              + "&time=" + Timestamp
        // Eksekusi pengiriman
        respon = HTTP_GET(URL)
```

```
        IF respon.status_code == 200 THEN
            LOG("Pengiriman Berhasil")
            waktu_terakhir_kirim = waktu_sekarang
        END IF
    END IF
END IF
```

Gambar 4. 5 *Pseudocode* Transmisi Data via HTTP *GET*

Algoritma ini menggunakan masukan berupa nilai suhu, pH, TDS, turbiditas, *timestamp*, serta *API Key* sebagai kunci autentikasi. Logikanya adalah membatasi pengiriman data dengan interval 4 detik guna menjaga kestabilan *bandwidth*. Sebagai keluaran, data tersebut terekam secara terstruktur ke dalam baris dan kolom pada *Google Sheets*. Proses ini dijalankan secara otomatis setelah sistem memastikan bahwa koneksi internet dalam kondisi stabil dan stempel waktu tersedia.

4.1.3 Pelaksanaan Pengambilan Data Lapangan

Tahap pengambilan data dilaksanakan secara langsung pada tiga lokasi sumber air sumur bor milik warga yang tersebar di Desa Baureno, Desa Pasinan, dan Desa Seratu (Sratu Rejo). Secara kasat mata, sampel air dari ketiga sumur bor tersebut terlihat cukup jernih. Namun, pengujian secara fisis dan kimiawi menggunakan sensor tetap diperlukan untuk memastikan parameter yang tidak kasat mata (seperti pH dan TDS) memenuhi standar kelayakan.

Karena karakteristik sumur bor memiliki lubang pipa yang sempit, instrumen tidak memungkinkan untuk dicelupkan langsung ke dalam sumber mata air. Oleh karena itu, prosedur pengambilan data diadaptasi menggunakan metode on-site sampling, di mana air sumur ditarik dan ditampung ke dalam wadah ukur dengan volume representatif sebanyak 2 hingga 3 liter. Probe sensor kemudian

dicelupkan ke dalam wadah penampungan tersebut dan dibiarkan merekam secara otomatis.

Berikut adalah rincian rentang waktu pelaksanaan perekaman data (*time-series*) di ketiga titik lokasi:

a. Sumur Desa Baureno: Dilaksanakan pada tanggal 4 Mei 2026, dengan perekaman data yang berlangsung mulai pukul 10:04:07 hingga pukul 11:46:35 WIB.
b. Sumur Desa Pasinan: Dilaksanakan pada hari yang sama, 4 Mei 2026, dengan sesi perekaman mulai pukul 16:06:20 hingga pukul 17:53:17 WIB.
c. Sumur Desa Seratu (Sratu Rejo): Dilaksanakan pada tanggal 7 Mei 2026, dengan waktu perekaman yang berlangsung dari pukul 14:03:18 hingga pukul 16:26:27 WIB.

Selama proses pengujian, instrumen berhasil mengirimkan *log* data ke *server Google Sheets* setiap 4 detik. Dalam pelaksanaannya, terdapat dinamika terkait manajemen daya perangkat. Meskipun kotak instrumen telah dilengkapi dengan baterai internal dan modul kontroler panel surya, operasional di lapangan menunjukkan bahwa perekaman *time-series* yang memakan waktu hingga hampir dua jam secara terus-menerus membutuhkan suplai arus yang sangat konstan. Untuk mencegah kegagalan transmisi akibat inefisiensi daya baterai bawaan, sumber daya utama dialihkan menggunakan *port* USB dari perangkat laptop.

ServerTimeStam	OriginalTimeStar	apiKey	Suhu	TDS	Salinitas	pH	Turbiditas
05/04/2026 10:	2026-05-04 10:	AKfycbz9YYBL	26,81	251,05	309	7,04	0,99
05/04/2026 10:	2026-05-04 10:	AKfycbz9YYBL	26,81	251,4	293	7,04	1,09
05/04/2026 10:	2026-05-04 10:	AKfycbz9YYBL	26,81	251,31	289	7,02	0,9
05/04/2026 10:	2026-05-04 10:	AKfycbz9YYBL	26,81	251,05	286	6,99	1,09
05/04/2026 10:	2026-05-04 10:	AKfycbz9YYBL	26,81	251,14	284	7	1,09
05/04/2026 10:	2026-05-04 10:	AKfycbz9YYBL	26,87	250,96	284	6,99	0,99
05/04/2026 10:	2026-05-04 10:	AKfycbz9YYBL	26,81	251,66	284	6,99	1,14
05/04/2026 10:	2026-05-04 10:	AKfycbz9YYBL	26,81	251,31	283	7	1,12
05/04/2026 10:	2026-05-04 10:	AKfycbz9YYBL	26,81	251,14	282	7	1,19
05/04/2026 10:	2026-05-04 10:	AKfycbz9YYBL	26,81	251,23	283	7	1,09

Gambar 4. 6 Tangkapan Layar *log* data *time-series* pada *Google Sheets*

Secara keseluruhan, pelaksanaan observasi *on-site* di ketiga sumur bor ini mengumpulkan akumulasi 3.511 baris data mentah yang valid. Keberhasilan akuisisi ini memastikan bahwa sistem memiliki volume data primer yang cukup untuk diekspor menjadi format CSV.

4.2 Deskripsi & Persiapan Data

Data primer yang digunakan dalam penelitian ini merupakan himpunan data yang dikumpulkan secara langsung dari hasil observasi lapangan di wilayah Kabupaten Bojonegoro. Pada tahap paling awal sebelum dilakukan pemrosesan apapun, data mentah (*raw* data) hasil rekaman sensor *IoT* yang masuk ke pangkalan data memiliki beberapa atribut tambahan yang merupakan bawaan dari sistem transmisi *server*.

Tabel 4.1 berikut memperlihatkan cuplikan struktur data mentah (raw data) sesaat setelah diakuisisi, di mana masih terdapat atribut pencatatan waktu (*ServerTimeStamp* dan *OriginalTimeStamp*) serta kunci autentikasi (*apiKey*) bersamaan dengan nilai pembacaan parameter fisik air.

Tabel 4. 1 Cuplikan Data Mentah Hasil Akuisisi Sensor IoT

***Server* TimeStamp**	**Original TimeStamp**	***apiKey***	**Suhu**	**TDS**	**Salinitas**	**pH**	**Turbiditas**
05/04/2026 10:04:11	2026-05-04 10:04:07	AKfycbz9YYBLFRGe IZCMs…	26,81	251,05	309	7,04	0,99
05/04/2026 10:04:16	2026-05-04 10:04:11	AKfycbz9YYBLFRGe IZCMs…	26,81	251,4	293	7,04	1,09
05/04/2026 10:04:20	2026-05-04 10:04:17	AKfycbz9YYBLFRGe IZCMs…	26,81	251,31	289	7,02	0,9
05/04/2026 10:04:24	2026-05-04 10:04:22	AKfycbz9YYBLFRGe IZCMs…	26,81	251,05	286	6,99	1,09
05/04/2026 10:04:29	2026-05-04 10:04:26	AKfycbz9YYBLFRGe IZCMs…	26,81	251,14	284	7	1,09

Kehadiran kolom atribut seperti waktu transmisi dan *API Key* pada Tabel 4.1 sangat penting sebagai bukti otentikasi pengiriman data secara *real-time* dari perangkat keras ke *server*. Namun, karena algoritma klasifikasi kualitas air hanya membutuhkan parameter fisik dan kimiawi, data mentah tersebut harus melalui serangkaian tahapan persiapan dan pembersihan sebelum dimasukkan ke dalam model *machine learning*.

Dataset akhir pasca-pembersihan menyisakan total 3.511 baris sampel data yang direpresentasikan murni oleh 4 fitur parameter fisis air, yaitu: tingkat keasaman (pH), tingkat kekeruhan (turbiditas), jumlah zat padat terlarut (*Total Dissolved Solids*/TDS), dan suhu air. Untuk memberikan gambaran umum secara kuantitatif mengenai karakteristik sebaran data yang telah bersih, ringkasan statistik deskriptif dari seluruh parameter masukan disajikan pada Tabel 4.2.

Tabel 4. 2 Ringkasan Statistik Deskriptif Dataset Kualitas Air

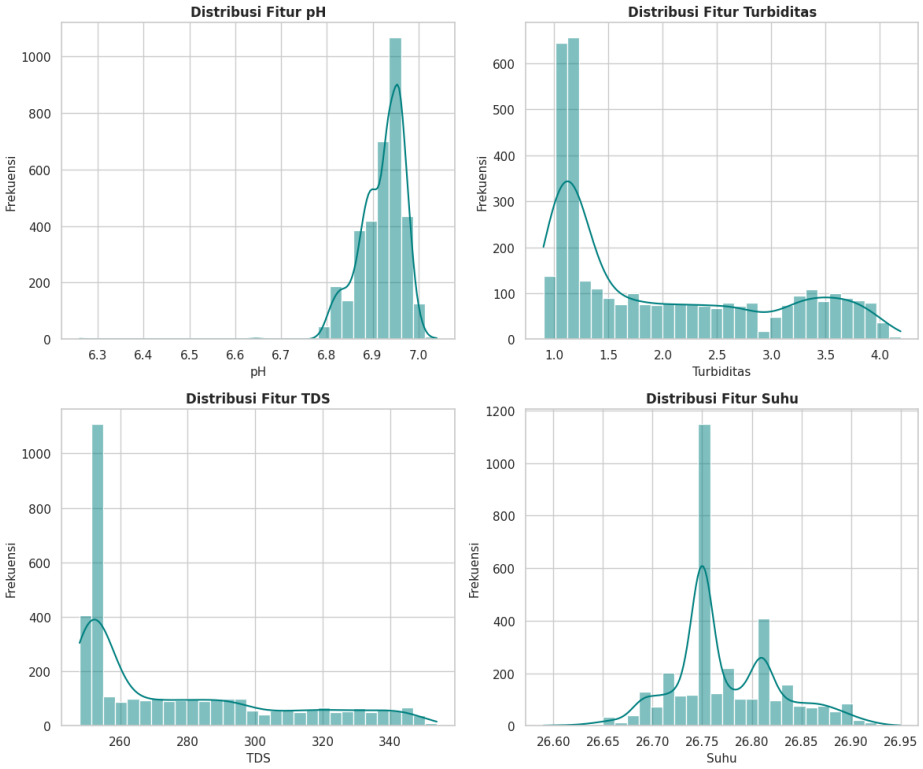
Parameter	**Count**	***Mean***	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
Suhu (°C)	3511	26.77	0.05	26.59	26.75	26.75	26.81	26.95
TDS (mg/L)	3511	277.14	30.24	247.9	251.83	263.64	295.7	353.85
pH	3511	6.92	0.82	6.26	6.89	6.93	6.96	7.04
Turbiditas (NTU)	3511	1.98	0.99	0.9	1.14	1.54	2.78	4.19

Dari ringkasan tersebut, parameter Suhu terlihat paling stabil karena memiliki standar deviasi yang kecil (0,05) dengan rentang nilai yang wajar untuk air, yaitu antara 26,59°C hingga 26,95°C. Untuk parameter pH, nilai rata-ratanya berada di angka netral (6,92), dengan rentang nilai terendah 6,26 dan tertinggi 7,04. Hal ini menunjukkan bahwa pembacaan sensor berjalan dengan baik tanpa menangkap adanya nilai ekstrem atau *outlier* yang tidak wajar.

Di sisi lain, parameter TDS dan Turbiditas memiliki rentang nilai yang sangat jauh berbeda dibandingkan parameter lainnya. Rata-rata TDS berada di angka 277,14 mg/L dengan nilai maksimum 353,85 mg/L dengan standar deviasi sebesar 30,24. Sementara itu, parameter Turbiditas berada pada skala yang sangat kecil dengan nilai rata-rata 1,98 NTU dan maksimum 4,19 NTU. Perbedaan skala yang sangat timpang ini di mana TDS berada di angka ratusan sementara pH dan Turbiditas hanya bernilai satuan tunggal menjadi landasan utama mengapa tahap pra-pemrosesan sangat dibutuhkan. Kondisi data mentah ini mewajibkan diterapkannya normalisasi menggunakan *Min-Max Scaling* agar semua nilai fitur berubah menjadi rentang yang seragam (0 sampai 1), sehingga tidak ada fitur bernilai besar yang mendominasi perhitungan saat algoritma *XGBoost* dilatih.

Visualisasi histogram (Gambar 4.7) memetakan distribusi frekuensi dari 3.511 sampel data kualitas air. Berdasarkan grafik yang terbentuk, parameter pH menunjukkan distribusi data yang cenderung memusat pada rentang nilai 6,8 hingga 7,0, dengan lonjakan frekuensi tertinggi berada di sekitar 6,95 dan perlahan melandai ke arah nilai yang lebih rendah. Sementara itu, parameter Suhu menunjukkan pola sebaran yang cukup unik dengan beberapa titik puncak, di mana

pemusatan data paling tinggi terjadi tepat pada suhu 26,75°C, diikuti dengan puncak sekunder yang lebih kecil pada kisaran 26,80°C.



Gambar 4. 7 Grafik Distribusi Kelas

Sebaliknya, parameter Turbiditas dan TDS menunjukkan kecenderungan penumpukan data pada nilai-nilai yang rendah sebelum akhirnya menyebar ke arah nilai yang lebih besar. Mayoritas sampel Turbiditas menumpuk pada rentang nilai 1,0 hingga 1,2 NTU, sebelum akhirnya menurun perlahan dan membentuk sedikit gelombang data di rentang 3,0 hingga 4,0 NTU. Pada parameter TDS, observasi menunjukkan penumpukan data yang sangat drastis pada batas bawah sekitar 250 mg/L, yang kemudian perlahan melandai dan menyebar hingga menyentuh kisaran 350 mg/L.

Secara keseluruhan, visualisasi ini memberikan bukti nyata bahwa meskipun data berada dalam rentang baku mutu yang wajar, tetap terdapat perbedaan skala angka yang sangat mencolok antar parameter. Parameter TDS yang bergerak pada skala ratusan (250–350 mg/L) dan Suhu pada skala puluhan (26°C) berpotensi mendominasi perhitungan model secara tidak seimbang jika disandingkan dengan parameter pH dan Turbiditas yang hanya berskala satuan. Kondisi ini memperkuat alasan mengapa penerapan *Min-Max Scaling* pada tahap pra-pemrosesan data sangat dibutuhkan. Penyeragaman skala ini mutlak diperlukan untuk memastikan algoritma *XGBoost* dapat memproses dan mempelajari setiap parameter secara adil tanpa terpengaruh oleh fitur yang sekadar memiliki angka bawaan lebih besar.

4.2.1 Analisis *Class Imbalance*

Penentuan kelas target (*ground truth*) didasarkan pada standar baku mutu kesehatan lingkungan. Aturan ambang batas parameter merujuk secara spesifik pada Peraturan Menteri Kesehatan Republik Indonesia (Permenkes RI) Nomor 2 Tahun 2023 tentang Peraturan Pelaksanaan Peraturan Pemerintah Nomor 66 Tahun 2014 tentang Kesehatan Lingkungan.

Tabel 4. 3 Parameter Air Minum Permenkes Nomor 2 Tahun 2023

No	Jenis Parameter	Kadar Maksimum yang diperbolehkan	Satuan	Metode Pengujian
	FISIKA			
1	Suhu	± 3 suhu udara	°C	SNI/APHA
2	*Total Dissolve Solids*	< 300	mg/L	SNI/APHA
3	Kekeruhan	< 3	NTU	SNI atau yang setara
	KIMIA			
1	pH	6,5 – 8,5	-	SNI/APHA

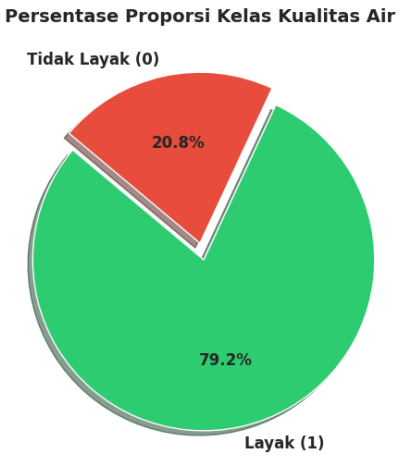
Proses pelabelan pada 3.511 data dilakukan secara otomatis memanfaatkan operasi logika kondisional melalui pustaka *Pandas Python*.

Tabel 4. 4 Logika Pelabelan Kualitas Air berdasarkan Aturan dari Permenkes

Status Kelayakan (Y)	**Kondisi Parameter**
Layak	pH antara 6,5 – 8,5 Turbiditas ≤ 3 TDS ≤ 300 Suhu antara 20°C – 30°C
Tidak Layak	1. Melanggar satu atau lebih ambang batas mutlak Permenkes. 2. Atau terdeteksi Risiko Biologis (Suhu ≥ 28,0°C dan Turbiditas ≥ 2,0 NTU). 3. Atau terdeteksi Risiko Kimiawi (pH ≤ 6,8 dan TDS ≥ 250 mg/L)

Air dinyatakan tidak memenuhi syarat jika ada satu atau lebih parameter wajib yang melewati batas aman yang ditetapkan pemerintah. Suhu air ≥ 28,0 dan Turbiditas (kekeruhan) ≥ 2,0 NTU mengindikasikan adanya potensi pertumbuhan mikroorganisme, bakteri, dan alga yang cepat (PT Bhumi Jati Power, 2022). Keasaman pH ≤ 6,8 menunjukkan air bersifat terlalu asam sehingga dapat memicu korosi dan melarutkan logam beracun. Sementara TDS (Total Padatan Terlarut) ≥ 250 mg/L menandakan tingginya konsentrasi ion atau mineral terlarut yang dapat mengubah rasa air dan membebani ginjal (Asmadi et al., 2020).

Sistem mengevaluasi setiap baris data menggunakan logika kondisional tersebut. Jika parameter memenuhi standar Permenkes dan terbebas dari risiko interaksi laten (biologis dan kimiawi), maka data dilabeli “Layak” (1). Namun, jika satu saja parameter menyimpang atau terdeteksi adanya risiko interaksi kritis, data seketika dilabeli “Tidak Layak” (0). Label ini disimpan dalam kolom Target sebagai acuan kebenaran (*ground truth*) bagi algoritma *XGBoost*.



Gambar 4. 8 Grafik Distribusi Kelas Layak dan Tidak Layak

Terdapat ketidakseimbangan kelas (*imbalanced dataset*), di mana kelas mayoritas (“Layak”) mendominasi distribusi data. Hal ini diatasi pada tahap pemodelan menggunakan parameter `scale_pos_weight` dalam algoritma *XGBoost* untuk menyeimbangkan bobot pembelajaran pada kelas minoritas.

```
FOR EACH baris i IN D DO
    IF D[i].memenuhi_baku_mutu == TRUE THEN
        D[i].label = "air layak"
    ELSE
        D[i].label = "air tidak layak"
    END IF
END FOR
D_label = rebalance_data(D, target_kelas="air sanitasi",
min_sampel=550)
```

Gambar 4. 9 *Pseudocode* Pelabelan Kelas Biner dan Penyesuaian Distribusi

Proses pelabelan mengonversi data hasil pengukuran sensor menjadi dua kelas klasifikasi biner, yaitu "air layak minum" dan "air sanitasi". Untuk memastikan algoritma *XGBoost* dapat mengenali pola air yang tercemar dengan baik, dilakukan intervensi pada distribusi kelas target sehingga secara spesifik terdapat minimal 550 baris data yang dialokasikan pada kelas "air sanitasi".

Penyesuaian ini mencegah model mengalami bias akibat dominasi data kelas mayoritas.

4.2.2 Pra-pemrosesan

Sistem ini akan mengimplementasi beberapa langkah *preprocessing* data. Tujuan dari *preprocessing* data ini adalah untuk menghindari *overfitting* pada sistem dan agar mendapat hasil yang maksimal. Beberapa tahapan yang dilakukan dalam *preprocessing* adalah:

1. Pembersihan Data (*Data Cleaning*)

Kolom atribut yang tidak relevan untuk proses klasifikasi prediktif, seperti *ServerTimeStamp*, *OriginalTimeStamp*, dan *apiKey*, dihapus dari dataset agar tidak mengganggu proses komputasi model. Proses ini menyisakan 4 fitur parameter kualitas air utama, yaitu Suhu, TDS, pH, dan Turbiditas.

2. Normalisasi Fitur (*Feature Scaling*)

Mengingat setiap parameter sensor memiliki rentang nilai asli yang sangat berbeda (seperti parameter TDS yang bernilai ratusan disandingkan dengan parameter Turbiditas yang bernilai desimal kecil), seluruh fitur numerik diseragamkan menggunakan metode *MinMaxScaler*. Metode ini mentransformasi seluruh nilai data asli ke dalam skala interval yang sama, yaitu rentang [0, 1], untuk mempermudah proses komputasi algoritma.

```
scaler = MinMaxScaler(feature_range=(0, 1))
X_norm = scaler.fit_transform(X_raw)
RETURN X_norm
```

Gambar 4. 10 *Pseudocode* Normalisasi Fitur menggunakan *MinMaxScaler*

Algoritma pada Gambar 4.10 menerima masukan berupa matriks kolom parameter air yang masih menggunakan satuan aslinya. Dengan memanggil fungsi *MinMaxScaler*, setiap nilai di dalam kolom secara matematis ditransformasikan sedemikian rupa sehingga menyusut atau merenggang ke dalam rentang nilai mutlak antara 0 hingga 1. Transformasi ini sangat krusial untuk mempercepat proses konvergensi komputasi pada model klasifikasi.

4.3 Hasil Skenario Pengujian

Pengujian Tahap pelatihan model *XGBoost* dilakukan secara komprehensif menggunakan tiga skenario pembagian rasio data (Data Latih : Data Uji), yakni 60:40, 70:30, 80:20, dan 90:10. Pada masing-masing skenario, dilakukan pencarian parameter arsitektur paling optimal (*Hyperparameter Tuning*) memanfaatkan metode *GridSearchCV* dengan *5-Fold Cross Validation*.

4.3.1 Hasil dengan *Hyperparameter Tuning (GridSearchCV)*

Untuk meminimalkan nilai kerugian (*log loss*) serta menemukan struktur pohon keputusan yang paling efisien dan memiliki generalisasi tertinggi, dilakukan proses *hyperparameter* tuning pada Skenario B. Proses pencarian dilakukan menggunakan metode *GridSearchCV* yang dikombinasikan dengan validasi silang 5-lipatan (5-*Fold Cross Validation*) pada subset data latih masing-masing rasio. Eksperimen ini mengeksplorasi 108 kombinasi parameter unik yang mencakup variasi `max_depth`, `learning_rate`, `n_estimators`, `subsample`, dan `colsample_bytree`.

4.3.1.1 Hasil Skenario 90:10

Pada eksperimen skenario pertama, dataset kualitas air dibagi menggunakan rasio 90:10. Proses pemisahan data (*train-test split*) ini mengalokasikan porsi data latih yang sangat masif, yakni 3.159 baris data (90%) untuk proses pembelajaran algoritma, dan menyisihkan 352 baris data uji (10%) sebagai instrumen evaluasi akhir yang benar-benar asing bagi model.

Pemilihan arsitektur algoritma pada 3.159 data latih ini dieksekusi melalui komputasi *GridSearchCV* dengan skema 5-*Fold* Cross Validation. Tabel 4.5 menampilkan rincian peringkat 1 hingga 10 dari total 108 kombinasi *hyperparameter* yang dieksplorasi oleh sistem, diurutkan secara menurun berdasarkan pencapaian nilai rata-rata validasi tertinggi (*Mean Test Score*).

Tabel 4. 5 Kombinasi Parameter Terbaik pada Skenario 90:10

Rank Test Score	***Max Depth***	***Learning Rate***	***N Estimators***	***Subsample***	***Colsample Bytree***	***Mean Test Score***	***Std Test Score***
1	3	0.1	100	0.8	0.8	98.32%	0.0037
2	3	0.1	100	0.8	0.9	98.32%	0.0037
3	3	0.01	200	0.9	0.8	98.29%	0.006
4	3	0.01	200	0.9	0.9	98.29%	0.006
5	5	0.1	100	0.9	0.8	98.29%	0.0039
6	5	0.1	100	0.9	0.9	98.29%	0.0039
7	3	0.01	300	0.8	0.8	98.26%	0.0054
8	3	0.01	300	0.8	0.9	98.26%	0.0054
9	3	0.01	200	0.8	0.8	98.26%	0.0057
10	3	0.01	200	0.8	0.9	98.26%	0.0057

Berdasarkan Tabel 4.5, arsitektur *Best Estimator* jatuh pada kombinasi kedalaman pohon (`max_depth`) = 3, laju pembelajaran (`learning_rate`) = 0.1, jumlah iterasi pohon (`n_estimators`) = 100, serta pengaktifan parameter stokastik (`subsample`) = 0.8 dan (`colsample_bytree`) = 0.8. Arsitektur ini mencatatkan rata-rata akurasi latih tertinggi sebesar 98.32%.

Ketersediaan volume data latih yang lebih masif pada rasio 90% ini memungkinkan algoritma *XGBoost* untuk belajar dengan lebih efisien. Hal ini dibuktikan dengan terpilihnya *learning rate* yang lebih besar (0.1) dan jumlah pohon yang lebih sedikit (100) dibandingkan skenario rasio lainnya. Kombinasi ini memberikan keseimbangan optimal antara kecepatan komputasi dan ketajaman deteksi, tanpa mengorbankan kesederhanaan model.

Tabel 4. 6 Rincian Akurasi 5-*Fold Cross Validation* Skenario 90:10

Iterasi Pengujian	**Akurasi Validasi**
Lipatan (*Fold*) 1	98.26%
Lipatan (*Fold*) 2	97.78%
Lipatan (*Fold*) 3	98.10%
Lipatan (*Fold*) 4	98.73%
Lipatan (*Fold*) 5	98.73%
Rata-rata Akurasi (*Mean*)	98.32% (± 0.37%)

Rincian pada Tabel 4.6 mengonfirmasi bahwa penambahan volume data latih berbanding lurus dengan peningkatan kestabilan model. Toleransi simpangan baku yang tercatat sangat sempit, yakni hanya ± 0.37%. Angka fluktuasi yang mendekati nol ini menjadi jaminan statistik bahwa model peringkat 1 tidak mengalami *overfitting* dan sangat kebal terhadap perubahan variasi sampel data lingkungan yang diujikan.

4.3.1.2 Hasil Skenario 80:20

Pada eksperimen skenario kedua ini, dataset kualitas air dibagi menggunakan rasio 80:20. Proporsi ini menghasilkan himpunan data latih sebanyak 2808 baris data (80%) untuk membangun pengetahuan algoritma, dan menyisihkan 703 baris data uji (20%) sebagai instrumen evaluasi akhir.

Sebanyak 2808 data latih dievaluasi menggunakan metode *GridSearchCV* yang dipadukan dengan 5-*Fold Cross Validation*. Evaluasi ini bertujuan untuk mencari arsitektur model *XGBoost* yang paling optimal. Tabel 4.7 menampilkan rincian peringkat 1 hingga 10 dari keseluruhan 108 kombinasi parameter yang dihasilkan dari proses optimasi, diurutkan berdasarkan nilai rata-rata akurasi validasi terbaik.

Tabel 4. 7 Kombinasi Parameter Terbaik pada Skenario 80:20

Rank Test Score	***Max Depth***	***Learning Rate***	***N Estimators***	***Subsample***	***Colsample Bytree***	***Mean Test Score***	***Std Test Score***
1	3	0.01	300	0.9	0.8	98.15%	0.0061
2	3	0.01	300	0.9	0.9	98.15%	0.0061
3	3	0.01	100	0.8	0.8	98.11%	0.006
4	5	0.01	100	0.8	0.8	98.11%	0.006
5	7	0.01	100	0.8	0.8	98.11%	0.006
6	3	0.01	100	0.8	0.9	98.11%	0.006
7	5	0.01	100	0.8	0.9	98.11%	0.006
8	7	0.01	100	0.8	0.9	98.11%	0.006
9	3	0.01	200	0.9	0.8	98.08%	0.0066
10	3	0.01	200	0.9	0.9	98.08%	0.0066

Berdasarkan Tabel 4.7, terdapat fenomena di mana kombinasi kedalaman pohon (`max_depth`) 3 dan 5 menghasilkan rata-rata akurasi validasi yang identik, yakni sebesar 98.15%. Dalam kaidah pembelajaran mesin, arsitektur yang lebih sederhana (`max_depth` = 3) dipilih secara definitif sebagai Best Estimator. Pemilihan pohon keputusan yang lebih dangkal ini bertujuan untuk membatasi kompleksitas model, sehingga algoritma berfokus mempelajari pola fundamental dari parameter air Kemenkes dan terhindar dari risiko penghafalan data (*overfitting*).

Selain itu, penggunaan laju pembelajaran (`learning_rate`) yang sangat konservatif di angka 0.01 dipadukan dengan jumlah pohon yang banyak

(`n_estimators` = 300) membuktikan bahwa model *XGBoost* pada skenario ini melakukan proses perbaikan kesalahan (*Boosting*) secara perlahan, hati-hati, dan sangat terukur.

Untuk menguji kekebalan arsitektur Rank 1 tersebut terhadap variasi porsi data, berikut dijabarkan rincian skor akurasi pada kelima iterasi pengujian silang internal.

Tabel 4. 8 Rincian Akurasi 5-*Fold* Cross Validation Skenario 80:20

Iterasi Pengujian	**Akurasi Validasi**
Lipatan (*Fold*) 1	97.86%
Lipatan (*Fold*) 2	97.15%
Lipatan (*Fold*) 3	98.22%
Lipatan (*Fold*) 4	98.93%
Lipatan (*Fold*) 5	98.57%
Rata-rata Akurasi (*Mean*)	98.15% (± 0.61%)

Data pada Tabel 4.8 memberikan bukti empiris bahwa model yang dibangun sangat stabil. Pada lima kali perputaran blok data latih yang berbeda-beda, model tidak pernah mencatatkan akurasi di bawah 97%. Nilai toleransi simpangan yang sangat sempit, yakni hanya ± 0.61%, dibandingkan skenario 90:10, angka ini sangat wajar mengingat porsi data latih yang diberikan lebih sedikit.

4.3.1.3 Hasil Skenario 70:30

Pada eksperimen ketiga, dataset kualitas air didistribusikan menggunakan rasio 70:30. Pemisahan ini menyisakan 2457 baris data latih (70%) sebagai bahan pembelajaran algoritma, sementara porsi data uji diperbesar menjadi 1054 baris data (30%) untuk menguji daya generalisasi model secara lebih ketat.

Optimasi arsitektur pada 2457 data latih ini dilakukan menggunakan algoritma *GridSearchCV* dan divalidasi melalui 5-*Fold Cross Validation*.

Pemeringkatan konfigurasi *hyperparameter* terbaik disajikan pada Tabel 4.9, yang merangkum peringkat 1 hingga 10 teratas dari total 108 kombinasi yang diujikan pada skenario ini.

Tabel 4. 9 Kombinasi Parameter Terbaik pada Skenario 70:30

Rank Test Score	***Max Depth***	***Learning Rate***	***N Estimators***	***Subsample***	***Colsample Bytree***	***Mean Test Score***	***Std Test Score***
1	7	0.01	300	0.9	0.8	98.13%	0.0039
2	7	0.01	300	0.9	0.9	98.13%	0.0039
3	5	0.01	300	0.9	0.8	98.05%	0.0047
4	5	0.01	300	0.9	0.9	98.05%	0.0047
5	3	0.01	200	0.9	0.8	98.01%	0.0052
6	3	0.01	200	0.9	0.9	98.01%	0.0052
7	5	0.01	200	0.9	0.8	97.97%	0.0054
8	5	0.01	200	0.9	0.9	97.97%	0.0054
9	7	0.01	300	0.8	0.8	97.97%	0.0045
10	7	0.01	300	0.8	0.9	97.97%	0.0045

Berdasarkan Tabel 4.9, terdapat pergeseran kelakuan pembelajaran model yang sangat rasional pada skenario ini. Arsitektur *Best Estimator* menuntut kedalaman pohon yang jauh lebih kompleks, yakni `max_depth` = 7, dengan laju pembelajaran `learning_rate` = 0.01, `n_estimators` = 300, serta `subsample` = 0.9 dan `colsample_bytree` = 0.8.

Secara teoritis, peningkatan kedalaman pohon ini mengindikasikan bahwa pengurangan volume data latih memaksa algoritma untuk membangun batasan keputusan yang lebih rumit guna menangkap pola kelayakan air. Meski menuntut kompleksitas struktur yang lebih tinggi, kombinasi ini tetap berhasil mengamankan nilai rata-rata validasi yang sangat baik di angka 98.13%.

Untuk mengevaluasi dampak dari peningkatan kompleksitas pohon tersebut terhadap stabilitas model, rincian skor dari kelima pengujian internal disajikan pada Tabel 4.10

Tabel 4. 10 Rincian Akurasi 5-*Fold* Cross Validation Skenario 70:30

Iterasi Pengujian	**Akurasi Validasi**
Lipatan (*Fold*) 1	97.97%
Lipatan (*Fold*) 2	97.76%
Lipatan (*Fold*) 3	98.37%
Lipatan (*Fold*) 4	98.78%
Lipatan (*Fold*) 5	97.76%
Rata-rata Akurasi (*Mean*)	98.13% (± 0.39%)

Data pada Tabel 4.10 memperlihatkan bahwa meskipun algoritma menggunakan arsitektur pohon yang cukup dalam, proses validasi internal tidak menunjukkan gejala *overfitting* yang fatal. Akurasi pada setiap lipatan data tetap konsisten berada di atas rentang 97%, dengan toleransi simpangan baku sebesar ± 0.39%. Hal ini membuktikan bahwa mekanisme perlindungan bawaan *XGBoost* mampu menjaga kestabilan matematis meskipun model dipaksa beroperasi dengan suplai data latih yang lebih terbatas.

4.3.1.4 Hasil Skenario 60:40

Pada eksperimen skenario terakhir ini, dataset kualitas air dibagi menggunakan rasio 60:40. Proporsi ini merupakan skenario dengan alokasi data latih paling minimal, di mana 2.106 baris data (60%) digunakan untuk melatih algoritma, sedangkan porsi data uji dimaksimalkan menjadi 1.405 baris data (40%). Skenario ini bertujuan untuk menguji batas kemampuan *XGBoost* dalam membentuk model klasifikasi yang akurat dengan ketersediaan informasi latih yang terbatas.

Proses pencarian arsitektur optimal pada 2.106 data latih tersebut dieksekusi melalui *GridSearchCV* dan divalidasi dengan 5-*Fold Cross Validation*. Tabel 4.11 menampilkan peringkat 1 hingga 10 kombinasi *hyperparameter* teratas dari total

108 skema yang direkomendasikan oleh sistem, diurutkan berdasarkan nilai *Mean Test Score* tertinggi.

Tabel 4. 11 Kombinasi Parameter Terbaik pada Skenario 60:40

Rank Test Score	***Max Depth***	***Learning Rate***	***N Estimators***	***Subsample***	***Colsample Bytree***	***Mean Test Score***	***Std Test Score***
1	5	0.01	200	0.8	0.8	98.20%	0.006469
2	7	0.01	200	0.8	0.8	98.20%	0.006469
3	5	0.01	200	0.8	0.9	98.20%	0.006469
4	7	0.01	200	0.8	0.9	98.20%	0.006469
5	3	0.01	200	0.8	0.8	98.15%	0.006936
6	3	0.01	200	0.9	0.8	98.15%	0.006936
7	3	0.01	200	0.8	0.9	98.15%	0.006936
8	3	0.01	200	0.9	0.9	98.15%	0.006936
9	5	0.01	200	0.9	0.8	98.15%	0.006608
10	7	0.01	200	0.9	0.8	98.15%	0.006608

Tabel 4.11, hasil evaluasi kembali menunjukkan fenomena nilai yang identik antara kedalaman pohon (`max_depth`) 5 dan 7 yang sama-sama mencatatkan akurasi validasi sebesar 98.20%. Sesuai dengan prinsip efisiensi komputasi, arsitektur dengan `max_depth` = 5 ditetapkan sebagai Best Estimator.

Kombinasi ini menggunakan laju pembelajaran yang lambat (`learning_rate` = 0.01) dan jumlah iterasi sedang (`n_estimators` = 200). Terpilihnya parameter ini menunjukkan bahwa dengan proporsi data latih 60%, model membutuhkan kedalaman pohon tingkat menengah untuk bisa mengenali pola kelayakan air dengan baik, tanpa harus mengambil risiko membentuk pohon yang terlalu rumit (`max_depth` = 7) yang rentan terhadap penghafalan data. Oleh karena itu, model menggunakan kedalaman pohon 5 dan memperketat pengambilan sampel acak menjadi hanya 80% (`subsample` = 0.8).

Untuk melihat efek dari minimnya rasio data latih terhadap fluktuasi performa model, rincian skor akurasi pada tiap lipatan pengujian internal dapat dilihat pada Tabel 4.12.

Tabel 4. 12 Rincian Akurasi 5-*Fold* Cross Validation Skenario 60:40

Iterasi Pengujian	**Akurasi Validasi**
Lipatan (*Fold*) 1	97.63%
Lipatan (*Fold*) 2	98.10%
Lipatan (*Fold*) 3	99.05%
Lipatan (*Fold*) 4	97.39%
Lipatan (*Fold*) 5	98.81%
Rata-rata Akurasi (*Mean*)	98.20% (± 0.65%)

Data pada Tabel 4.12 menunjukkan tren yang sangat rasional. Dengan semakin mengecilnya porsi data latih, tingkat toleransi fluktuasi menjadi sedikit lebih lebar, yakni sebesar ± 0.65% (tertinggi di antara semua skenario). Terdapat lonjakan akurasi hingga 99.05% pada *Fold* 3, namun sempat turun ke 97.39% pada *Fold* 4.

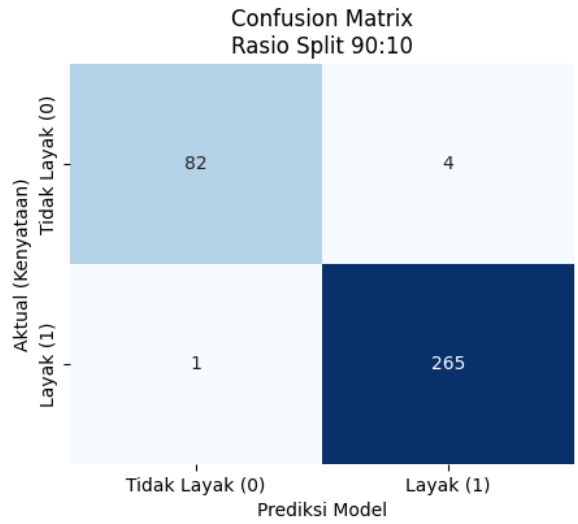
Meskipun menunjukkan rentang fluktuasi yang sedikit lebih besar, nilai akurasi terendah yang dicapai tetap berada pada taraf yang sangat memuaskan (di atas 97%) serta dengan rata-rata akurasi 98,20%. Hal ini menegaskan bahwa meskipun *XGBoost* dihadapkan pada skenario pembatasan data latih yang ekstrem (60:40), model *Best Estimator* yang dihasilkan tetap mampu menjaga keandalan klasifikasinya.

4.3.2 Pengujian Model Klasifikasi Akhir

Setelah arsitektur terbaik untuk masing-masing skenario pembagian data berhasil diidentifikasi pada fase pelatihan, tahap selanjutnya adalah mengukur performa generalisasi model tersebut pada lingkungan yang sesungguhnya. Pada

tahap ini, model *XGBoost* yang telah dikunci parameternya dihadapkan pada porsi Data Uji yang sejak awal diisolasi dan belum pernah dipelajari oleh algoritma.

4.3.2.1 Evaluasi Data Uji Skenario 90:10



Gambar 4. 11 *Confusion Matrix* Skenario 90:10

Berdasarkan matriks tersebut, dari total 352 data uji, model berhasil menebak benar 265 kelas "Layak" (*True Positive*) dan 82 kelas "Tidak Layak" (*True Negative*). Performa metrik evaluasi akhir dari prediksi ini mencatatkan Akurasi 98,58%, *Precision* 98,51%, *Recall* 99,62%, dan *F1-Score* 99,07%.

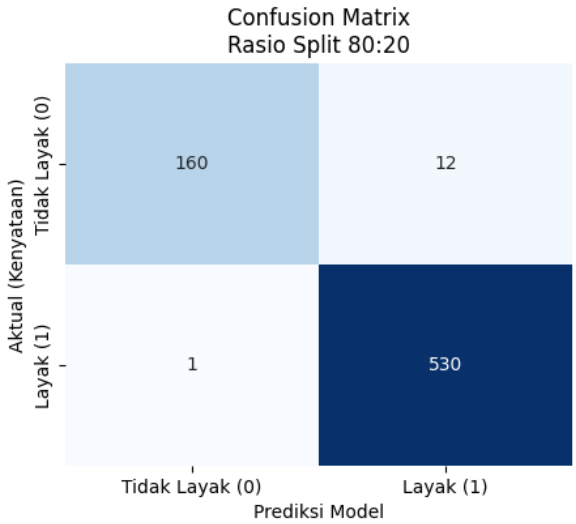
Secara analitis, tingginya nilai *Recall* (99,62%) sejalan dengan temuan pada matriks di mana model hanya menghasilkan 1 kasus *False Negative*. Hal ini menunjukkan bahwa sensitivitas model sangat tajam dalam mengenali air yang benar-benar layak. Jika terjadi 1 kasus *False Negative* ini di lapangan, dampaknya hanyalah warga membuang atau menggunakan 1 sampel air minum tersebut untuk kebutuhan mencuci/sanitasi, sehingga tidak mengancam nyawa. Namun, evaluasi kritis harus dijatuhkan pada adanya 4 kasus *False Positive*. Meskipun secara

persentase terlihat sangat kecil dan Akurasi model menyentuh 98,58%, dalam implementasi *Internet of Things* (IoT) di masyarakat, 4 kasus ini merupakan celah berbahaya di mana air yang berisiko (misalnya memiliki kadar TDS atau kekeruhan di atas batas aman) justru diloloskan dan diberi label "Layak Minum" oleh sistem.

Tingginya asupan porsi data latih pada skenario ini (90%) ditengarai membuat algoritma cenderung *over-optimistic* dalam menebak kelas mayoritas (kelas "Layak" yang mendominasi 79,2% dataset). Model menjadi kurang waspada terhadap sampel-sampel air sanitasi yang parameter fisisnya berada sangat tipis di batas aman standar Permenkes. Oleh karena itu, meskipun skenario 90:10 memiliki akurasi latih tertinggi, adanya 4 kesalahan fatal ini membuktikan bahwa model masih memiliki celah keamanan jika diimplementasikan langsung sebagai sistem peringatan dini kualitas air.

4.3.2.2 Evaluasi Data Uji Skenario 80:20

Model final pada skenario 80:20 diujikan pada 703 baris data uji. Hasil evaluasi matriks disajikan pada Gambar 4.12.



Gambar 4. 12 *Confusion Matrix* Skenario 80:20

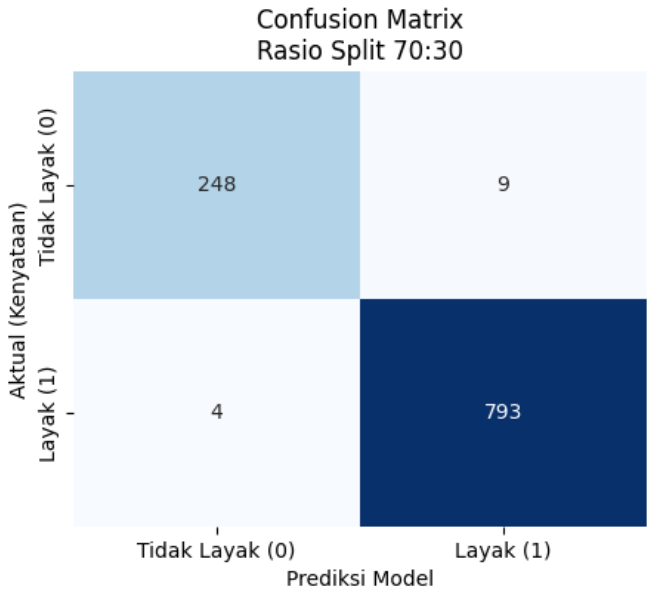
Dari 703 data uji, algoritma mencatatkan 530 tebakan benar untuk kelas "Layak" dan 160 tebakan benar untuk "Tidak Layak". Skor metrik evaluasi akhirnya menunjukkan Akurasi 98,15%, *Precision* 97,79%, *Recall* 99,81%, dan *F1-Score* 98,79%.

Anomali performa mulai terlihat pada skenario ini. Di satu sisi, model sangat konservatif dengan hanya menghasilkan 1 kasus *false negative*. Namun di sisi lain, terjadi lonjakan kesalahan fatal pada *false positive* yang cukup tajam menjadi 12 kasus. Ada 12 sampel air sanitasi yang gagal dibendung oleh sistem dan dilabeli sebagai air layak minum.

Penurunan nilai *Precision* menjadi 97,79% dan tingginya angka *False Positive* ini merupakan kelemahan bawaan dari arsitektur pohon yang terlalu dangkal (`max_depth` = 3) yang terpilih pada tahap pelatihan 80:20. Struktur pohon yang dangkal terbukti tidak memiliki ketajaman analitis yang cukup untuk memilah sampel-sampel batas (misalnya air dengan kekeruhan yang sedikit melewati ambang batas), sehingga sistem terlalu mudah menyimpulkan air tersebut sebagai air layak minum.

4.3.2.3 Evaluasi Data Uji Skenario 70:30

Pengujian pada skenario 70:30 dilakukan terhadap 1.054 baris data uji. Hasil evaluasinya menunjukkan tingkat ketepatan klasifikasi yang baik. Meskipun dihadapkan pada lebih dari seribu soal ujian, matriks membuktikan model mampu memprediksi 793 data "Layak" dan 248 data "Tidak Layak" secara sangat presisi. Hasil evaluasi metriknya melonjak drastis dengan Akurasi 98,77%, *Precision* 98,88%, *Recall* 99,50%, dan *F1-Score* 99,19%.

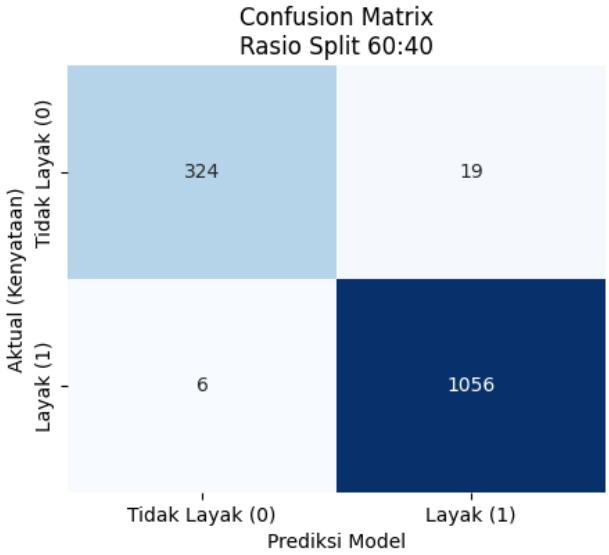


Gambar 4. 13 *Confusion Matrix* Skenario 70:30

Hasil uji Skenario 70:30 ini merupakan titik puncak performa algoritma (*Best Estimator*). Secara matematis, jumlah soal ujian (data uji) jauh lebih banyak daripada skenario 80:20, namun model secara brilian berhasil menekan jumlah *False Positive* turun dari 12 menjadi hanya 9 kasus.

Penurunan jumlah kesalahan fatal ini membuktikan bahwa keputusan algoritma saat proses tuning untuk memperdalam struktur pohon (`max_depth` = 7) terbukti sangat jitu. Arsitektur pohon keputusan yang lebih dalam dan kompleks ini berhasil mengekstraksi batas-batas parameter fisis yang samar seperti perbedaan tipis TDS atau pH air yang sebelumnya gagal ditangkap oleh skenario 80:20. Hasilnya, metrik melonjak drastis dengan Akurasi 98,77%, *Precision* 98,88%, *Recall* 99,50%, dan *F1-Score* 99,19%, menjadikan arsitektur ini paling andal dan aman untuk diimplementasikan di lapangan.

4.3.2.4 Evaluasi Data Uji Skenario 60:40



Gambar 4. 14 *Confusion Matrix* Skenario 60:40

Gambar 4.14 diatas menunjukkan hasil klasifikasi pada skenario dengan porsi data uji terbesar yaitu 1.405 baris data uji dianalisis menggunakan arsitektur model finalnya. Pada matriks tersebut, model berhasil menebak 1.056 kelas "Layak" dan 324 kelas "Tidak Layak". Evaluasi metriknya mencatatkan Akurasi 98,22%, *Precision* 98,23%, *Recall* 99,44%, dan *F1-Score* 98,83%.

Munculnya 19 kasus *False Positive* pada skenario ini merupakan angka kesalahan fatal tertinggi di antara semua percobaan. Hal ini sangat logis karena porsi data latih dipangkas menjadi hanya 60%, sehingga referensi model dalam mengenali pola air sanitasi menjadi paling minim.

Meski demikian, jika dikomparasikan dengan besarnya volume data yang harus ditebak (1.405 data uji), pencapaian Akurasi 98,22% membuktikan bahwa model *XGBoost* memiliki tingkat ketahanan yang luar biasa. Algoritma ini tidak

mengalami keruntuhan performa secara drastis meskipun dihadapkan pada skenario kelangkaan suplai data.

4.4 Pembahasan

Setelah seluruh tahapan pelatihan dan pengujian dieksekusi secara terpisah pada keempat skenario rasio pembagian data, tahap selanjutnya adalah melakukan pembahasan dan komparasi secara komprehensif. Pembahasan ini bertujuan untuk menganalisis sejauh mana pengaruh variasi volume data latih terhadap kemampuan generalisasi algoritma *XGBoost* dalam mengklasifikasikan kelayakan air bersih, serta menentukan model arsitektur mana yang paling optimal untuk diimplementasikan.

4.4.1 Analisis Perbandingan dan Penentuan Model Terbaik

Evaluasi perbandingan difokuskan pada dua metrik utama, yaitu performa saat model dievaluasi secara internal pada fase pelatihan (*Mean Test Score* / Akurasi Latih) dan performa akhir saat model dihadapkan pada Data Uji yang sesungguhnya (*Test Score* / Akurasi Uji Akhir). Ringkasan komparasi performa beserta arsitektur dari keempat skenario disajikan pada Tabel 4.13.

Tabel 4. 13 Hasil Komparasi Akurasi dan Arsitektur Algoritma *XGBoost*

Rasio (Latih:Uji)	**Arsitektur Terbaik (**`max_depth, learning_rate, n_estimators, subsample, colsample`**)**	**Akurasi Latih (%)**	**Akurasi Uji Akhir (%)**
90 : 10	3, 0.10, 100, 0.8, 0.8	98.32	98.58
80 : 20	3, 0.01, 300, 0.9, 0.8	98.15	98.15
70 : 30	7, 0.01, 300, 0.9, 0.8	98.13	98.77
60 : 40	5, 0.01, 200, 0.8, 0.8	98.2	98.22

Berdasarkan hasil rekapitulasi pada Tabel 4.13, terlihat dinamika komputasi yang sangat menarik terkait kemampuan adaptasi algoritma *XGBoost* terhadap perubahan volume informasi. Pada fase pelatihan, skenario dengan asupan data latih terbesar yakni 90:10 berhasil mendominasi evaluasi internal dengan capaian Akurasi Latih tertinggi (98,32%). Keunggulan pada fase ini wajar terjadi karena algoritma memiliki referensi data historis yang sangat melimpah untuk dikenali.

Meskipun demikian, tingginya metrik pada fase latih terbukti tidak linier dengan jaminan dominasi performa di lapangan. Ketika keempat model diuji secara definitif menggunakan data uji, performa klasifikasi puncak justru direbut oleh Skenario 70:30, yang mencatatkan lonjakan Akurasi Uji Akhir tertinggi di angka 98,77%.

Kemenangan performa pada rasio 70:30 ini memberikan justifikasi empiris yang sangat kuat bahwa proporsi 70% data latih dan 30% data uji merupakan titik keseimbangan yang paling ideal pada dataset kualitas air ini. Ketangguhan skenario 70:30 tidak lepas dari penyesuaian arsitektur parameternya yang sangat presisi; di mana pengurangan data latih direspon oleh algoritma dengan membangun struktur pohon keputusan yang jauh lebih dalam (`max_depth` = 7) dan jumlah iterasi maksimal (`n_estimators` = 300). Arsitektur yang lebih kompleks ini secara sukses mengekstraksi batasan fitur yang lebih tajam tanpa terjebak pada *overfitting* (berkat laju pembelajaran lambat 0.01 dan parameter stokastik). Hasilnya, Skenario 70:30 dinobatkan sebagai model *Best Estimator* mutlak dalam penelitian ini karena terbukti paling tangguh, presisi, dan aman dalam meminimalkan tingkat kesalahan fatal (*False Positive*).

4.4.2 Jawaban atas Rumusan Masalah

Berdasarkan rangkaian eksperimen, pengujian, dan analisis evaluasi metrik yang telah dilakukan, rumusan masalah dalam penelitian ini dapat terjawab secara komprehensif. Algoritma *XGBoost* terbukti sangat andal, stabil, dan efektif dalam mengklasifikasikan kelayakan kualitas air bersih di Kabupaten Bojonegoro.

Kemampuan performa model klasifikasi ini ditunjukkan oleh pencapaian metrik evaluasi akhir yang superior pada arsitektur terbaiknya (skenario rasio 70:30), di mana model berhasil menyentuh tingkat Akurasi sebesar 98,77%, *Precision* 98,88%, *Recall* 99,50%, dan *F1-Score* 99,19%. Model ini juga menunjukkan tingkat sensitivitas dan keamanan yang sangat tinggi, dibuktikan dengan kemampuannya menekan tingkat Kesalahan Tipe I (*False Positive*) menjadi hanya 9 kasus dari total 1.054 sampel pengujian.

Keberhasilan algoritma *XGBoost* dalam mengadaptasi arsitektur parameternya secara dinamis terutama fleksibilitas penyesuaian kedalaman pohon (`max_depth`) dan laju pembelajaran (`learning_rate`) sebagai respon terhadap kelangkaan maupun kelimpahan volume data latih memvalidasi tingkat ketahanan (robustness) model yang sangat tinggi. Hal ini membuktikan secara matematis bahwa metode *XGBoost* yang dioptimasi mampu mengekstraksi pola parameter fisik air secara optimal, presisi, dan terhindar dari bias penghafalan data, sehingga sangat layak untuk diintegrasikan sebagai instrumen cerdas pendukung keputusan kesehatan lingkungan di masyarakat.

4.5 Integrasi Penelitian

Penelitian ini tidak hanya dikembangkan sebagai sebuah karya keilmuan di bidang Teknik Informatika, tetapi juga didudukkan dalam dua dimensi fundamental ajaran Islam, yaitu *Muamalah Ma'allah dan Muamalah Ma'annas*. Dalam kerangka *Muamalah Ma'allah*, penelitian ini diposisikan sebagai wujud syukur dan upaya nyata seorang hamba dalam *menadabburi* ayat-ayat *kauniyah* Allah SWT melalui pengamatan terhadap fenomena alam. Sementara itu, dalam dimensi Mu*amalah Ma'annas*, pengembangan sistem klasifikasi kualitas air ini menjadi kontribusi untuk melindungi masyarakat dari risiko kesehatan akibat konsumsi air yang tidak layak, sekaligus merefleksikan nilai-nilai kepedulian sosial sebagai upaya memelihara keselamatan jiwa sesama manusia.

4.5.1 *Muamalah Ma'allah*

Keberhasilan pemodelan menggunakan *Extreme Gradient Boosting (XGBoost)* dalam membedakan kelas kelayakan air berdasarkan keragaman parameter fisisnya pada hakikatnya selaras dengan perintah pengamatan ciptaan Allah sebagaimana dijelaskan dalam Al-Qur'an Surah An-Nahl ayat 13:

وَمَا ذَرَاَ لَكُمْ فِى الْاَرْضِ مُخْتَلِفًا اَلْوَانُهٗۗ اِنَّ فِيْ ذٰلِكَ لَاٰيَةً لِّقَوْمٍ يَّذَّكَّرُوْنَ

"Dan (Dia menundukkan pula) apa yang Dia ciptakan untukmu di bumi ini dengan berlain-lainan macamnya. Sesungguhnya pada yang demikian itu benar-benar terdapat tanda (kekuasaan Allah) bagi kaum yang mengambil pelajaran." (QS. An-Nahl: 13)

Menurut tafsir Ibnu Katsir (2008), Allah menciptakan berbagai entitas di bumi dengan wujud, sifat, dan karakteristik (berlain-lainan macamnya) yang menakjubkan. Tafsir ini mengisyaratkan bahwa keanekaragaman sifat alam seperti

variasi nilai derajat keasaman (pH), fluktuasi suhu, dan perbedaan tingkat kepadatan zat terlarut (TDS) pada air merupakan tanda-tanda kebesaran yang patut diamati dan dikelompokkan oleh manusia. Penggunaan algoritma *machine learning* pada penelitian ini adalah bentuk implementasi akal (mengambil pelajaran) dalam mengenali pola-pola keragaman tersebut.

4.5.2 *Muamalah Ma'annas*

Lebih lanjut, temuan dari klasifikasi ini memberikan manfaat langsung terhadap upaya pelestarian salah satu elemen hidup paling vital. Hal ini sejalan

dengan firman Allah dalam Surah Al-Anbiya' ayat 30:

وَجَعَلْنَا مِنَ الْمَآءِ كُلَّ شَيْءٍ حَيٍّ اَفَلَا يُؤْمِنُوْنَ

"...Dan dari air Kami jadikan segala sesuatu yang hidup. Maka mengapakah mereka tiada juga beriman?" (QS. Al-Anbiya: 30)

Tafsir Ibnu Katsir (2008) menjelaskan bahwa ayat ini menegaskan air sebagai penopang utama kehidupan. Karena air adalah urat nadi kehidupan, maka memastikan kualitasnya aman untuk dikonsumsi merupakan sebuah tanggung jawab sosial yang besar. Di sinilah letak urgensi dari capaian algoritma *XGBoost* dalam penelitian ini yang difokuskan untuk menekan tingkat kesalahan *False Positive* (bahaya meloloskan air sanitasi tercemar menjadi seolah-olah air layak minum).

Keberhasilan teknologi dalam mencegah risiko konsumsi air kotor ini sejalan dengan prinsip *Hifdzun Nafs* (memelihara jiwa manusia), sebagaimana ditegaskan dalam firman Allah Surah Al-Maidah ayat 32:

...وَمَنْ اَحْيَاهَا فَكَاَنَّمَآ اَحْيَا النَّاسَ جَمِيْعًا...

"...Dan barangsiapa memelihara kehidupan seorang manusia, maka seakan-akan dia telah memelihara kehidupan semua manusia..." (QS. Al-Maidah: 32)

Menurut Tafsir Ibnu Katsir Jilid 3 (2008), makna "memelihara kehidupan" dalam ayat ini mencakup tindakan menyelamatkan jiwa seseorang dari kebinasaan dan malapetaka yang mengancam nyawa. Berdasarkan tafsir tersebut, upaya menyelamatkan masyarakat dari penyakit mematikan akibat meminum air tercemar melalui sistem deteksi dini bernilai sangat mulia, seakan-akan telah memelihara kehidupan umat manusia secara keseluruhan.

Oleh karena itu, hasil dari penelitian ini diharapkan tidak hanya menjadi kontribusi di bidang Teknik Informatika, tetapi juga diimplementasikan sebagai instrumen preventif bagi masyarakat Kabupaten Bojonegoro. Melalui pemanfaatan teknologi, masyarakat dapat menghindari konsumsi air tercemar, yang sekaligus merupakan bentuk *Muamalah Ma'annas* (menjaga keselamatan sesama) dan wujud nyata menjaga amanah Allah *Subhanahu wa Ta'ala* atas kelestarian sumber kehidupan.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian dan pengujian klasifikasi kualitas air di Kabupaten Bojonegoro, algoritma *XGBoost* terbukti sangat tangguh, terutama saat menggunakan rasio pembagian data 70:30. Dengan menerapkan arsitektur optimal yang terdiri dari `subsample`, dan `colsample_bytree`

`max_depth` = 7, `learning_rate` = 0.01, dan `n_estimators` = 300, model berhasil mencapai performa yang sangat tinggi, yaitu Akurasi 98,77%, *Precision* 98,88%, *Recall* 99,50%, dan *F1-Score* 99,19%. Selain itu, model ini sangat aman untuk diimplementasikan karena terbukti mampu menekan Kesalahan Tipe I (*False Positive*) hingga hanya menyisakan 9 kasus dari total 1.054 data uji.

Tingkat presisi model juga tetap terjaga berkat penyesuaian parameter scale_pos_weight yang terbukti sukses menangani ketidakseimbangan dataset (79,2% kelas "Layak" vs 20,8% kelas "Tidak Layak"), sehingga model tidak terbias dan tetap akurat mengenali kelas minoritas. Selain itu, transformasi *Min-Max Scaling* mutlak diperlukan untuk menormalisasi ketimpangan magnitudo parameter seperti nilai TDS yang berskala ratusan dan pH yang berskala satuan sehingga mencegah terjadinya dominasi komputasi oleh fitur yang bernilai besar.stabil

5.2 Saran

Untuk pengembangan penelitian selanjutnya dan implementasi sistem yang lebih baik, disarankan untuk menggunakan metode pelabelan dataset berdasarkan

hasil uji laboratorium riil, seperti penambahan parameter biologi (misalnya bakteri *E. coli*). Langkah ini diperlukan agar pola data tidak sebatas deterministik, sehingga pengujian model dapat menjadi lebih kompleks dan representatif. Selain itu, peneliti selanjutnya juga perlu memperbanyak sampel data minoritas, khususnya pada titik-titik air tercemar, dari berbagai lokasi tambahan di wilayah Bojonegoro guna memperkaya karakteristik pembelajaran model saat menghadapi kondisi lingkungan yang lebih ekstrem.

Dari segi implementasi teknis, model *XGBoost* yang telah selesai dievaluasi disarankan untuk diintegrasikan secara langsung ke dalam mikrokontroler pada perangkat keras IoT yang tersedia. Melalui integrasi tersebut, sistem klasifikasi ini dapat dioperasikan langsung sebagai sistem peringatan dini (*early warning system*) kualitas air secara *real-time* di lapangan.

DAFTAR PUSTAKA

Abdi, K., Warjaya, A., Muthmainnah, I., & Pahutar, P. H. (2024). Penerapan Algoritma Random Forest dalam Prediksi Kelayakan Air Minum. *Jurnal Ilmu Komputer Dan Informatika*, *3*(2), 81–88. https://doi.org/10.54082/jiki.81

Asmadi, Susilawati, & Salbiah. (2020). *Aplikasi Kalsium Laktat (C6 H10 Cao6) Dan Ultrafiltrasi Pada Proses Pengolahan Air Sungai Menjadi Air Minum Di Kabupaten Sambas*.

BMKG Jawa Timur. (2025). *Curah Hujan di Kabupaten Bojonegoro untuk 6 bulan ke Depan*. https://staklim-jatim.bmkg.go.id/index.php/prakiraan-iklim/prakiraan-bulanan/prakiraan-curah-hujan-bulanan/6-bulan-ke-depan/555561117-prakiraan-bulanan-curah-hujan-di-kabupaten-bojonegoro-untuk-6-bulan-ke-depan?utm_source=chatgpt.com

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 785–794. https://doi.org/10.1145/2939672.2939785

Darshan, P., & Subburaj, T. (2023). Evaluation of the Quality of Water Considering SVM and the XGBoost Technique. *Shanlax International Journal of Arts, Science and Humaniteis*, *11*(S1), 65–70.

Dinas Lingkungan Hidup Kabupaten Bojonegoro. (2023). *DLH Bojonegoro: Air Berbau Belerang di Desa Jari Tak Layak Konsumsi Tapi Aman Bagi Tumbuhan*. Situs Resmi Pemerintah Kabupaten Bojonegoro. https://bojonegorokab.go.id/berita/7257/dlh-bojonegoro-air-berbau-belerang-di-desa-jari-tak-layak-konsumsi-tapi-aman-bagi-tumbuhan

Elvin, & Wibowo, A. (2024). Forecasting water quality through machine learning and *hyperparameter* optimization. *Indonesian Journal of Electrical Engineering and Computer Science*, *33*(1), 496–506. https://doi.org/10.11591/ijeecs.v33.i1.pp496-506

Grbčić, L., Družeta, S., Mauša, G., Lipić, T., Lušić, D. V., Alvir, M., Lučin, I., Sikirica, A., Davidović, D., Travaš, V., Kalafatovic, D., Pikelj, K., Fajković, H., Holjević, T., & Kranjčević, L. (2022). Coastal water quality prediction based on machine learning with feature interpretation and spatio-temporal analysis. *Environmental Modelling and Software*, *155*. https://doi.org/10.1016/j.envsoft.2022.105458

Hidayaturrohman, Q. A., & Hanada, E. (2024). Impact of Data Pre-Processing Techniques on XGBoost Model Performance for Predicting All-Cause Readmission and Mortality Among Patients with Heart Failure.

BioMedInformatics, *4*(4), 2201–2212. https://doi.org/10.3390/biomedinformatics4040118

Ibnu Katsir. (2008a). *Tafsir Ibnu Katsir Jilid 5* (Terj. M. A). Pustaka Imam Asy-Syafi'i.

Ibnu Katsir. (2008b). *Tafsir Ibnu Katsir Jilid 6* (Terj. M. A). Pustaka Imam Asy-Syafi'i.

Kemenkes. (2023). *Laporan Tahunan : Pengamanan Kualitas Air Minum Tahun 2022*.

Kementerian Lingkungan Hidup. (2022). *Metode Pengukuran Indeks Kualitas Air*. https://doi.org/10.1787/ffa8d501-id

Khoi, D. N., Quan, N. T., Linh, D. Q., Thi, P., & Nhi, T. (2022). Using Machine Learning Models for Predicting the Water. *Mdpi*, 1–12.

Kurniawan, F., Aziza, M. R., Hasanah, N. A., Putri, F. I., & Prasetya, A. (2026). *IoT-Based Water Quality Monitoring and Suitability Modeling for Smart Campuses*. *10*(2), 398–406.

PT Bhumi Jati Power. (2022). Rencana Pengelolaan Lingkungan Hidup dan Rencana Pemantauan Lingkungan Hidup (RKL-RPL) PLTU Tanjung Jati B Unit 5 & 6. In *Bhumi Jati Power Official Website*. https://www.jbic.go.jp/en/business-areas/environment/projects/image/48649_21.pdf

Riyantoko, P. A., Fahrudin, T. M., & Hindrayani, K. M. (2021). Analisis Sederhana Pada Kualitas Air Minum Berdasarkan Akurasi Model Klasifikasi Dengan Menggunakan Lucifer Machine Learning. *Prosiding Seminar Nasional Sains Data*, *1*(01), 12–18. https://doi.org/10.33005/senada.v1i01.20

Sajiwo, A. F. B., Rahmat, B., & Junaidi, A. (2024). Klasifikasi Indeks Standar Pencemaran Udaran (Ispu) Menggunakan Algoritma Xgboost Dengan Teknik Imbalanced Data (Smote). *Jurnal Informatika Dan Teknik Elektro Terapan*, *12*(3). https://doi.org/10.23960/jitet.v12i3.4699

Sari, A. P., Tsaqif, D. A., & Firdawanti, A. R. (2024). *Classification of Drinking Water Source Suitability in West Java Using XGBoost and Cluster Analysis Based on SHAP Values *. 8*(2), 202–214.

Shaqia, H. (2025). *Fenomena Hujan di Musim Kemarau, Ahli Klimatologi Unigoro Peringatkan Warga untuk Waspada Potensi Bencana*. https://www.bojonegorotv.com/berita/971404855/fenomena-hujan-di-musim-kemarau-ahli-klimatologi-unigoro-peringatkan-warga-untuk-waspada-potensi-bencana?utm_source=chatgpt.com

Sutisna, & Yuniar, N. M. (2023). Klasifikasi Kualitas Air Bersih Menggunakan Metode Naïve baiyes. *Jurnal Sains Dan Teknologi*, *5*(1), 243–246. https://doi.org/10.55338/saintek.v5i1.1383

United Nations. (2016). Integrated Monitoring Guide for SDG 6 Targets and global indicators Cross-cutting and fragmented , at the core of sustainable development. *The Water Cycle in the Sustainable Development Goals*, *July 2016*, 1–26.

World Health Organization. (2022). Guidelines for drinking-water quality. In *ReVision* (Vol. 21, Issue 6, pp. 3–6). WHO. https://www.who.int/publications/i/item/9789241549950

LAMPIRAN

Tabel Kombinasi Parameter Terbaik pada Skenario 90:10

Rank Test Score	Max Depth	Learning Rate	N Estimators	Subsample	Colsample Bytree	*Mean* test score	Std test score
1	3	0,1	100	0,8	0,8	0,983223836	0,003686009
2	3	0,1	100	0,8	0,9	0,983223836	0,003686009
3	3	0,01	200	0,9	0,8	0,982907882	0,006034097
4	3	0,01	200	0,9	0,9	0,982907882	0,006034097
5	5	0,1	100	0,9	0,8	0,98290738	0,003922591
6	5	0,1	100	0,9	0,9	0,98290738	0,003922591
7	3	0,01	300	0,8	0,8	0,982591426	0,005384926
8	3	0,01	300	0,8	0,9	0,982591426	0,005384926
9	3	0,01	200	0,8	0,8	0,982591426	0,00565701
10	3	0,01	200	0,8	0,9	0,982591426	0,00565701
11	3	0,1	100	0,9	0,8	0,982590925	0,003998131
12	3	0,1	100	0,9	0,9	0,982590925	0,003998131
13	5	0,01	300	0,8	0,8	0,98227497	0,004287181
14	5	0,01	300	0,8	0,9	0,98227497	0,004287181
15	3	0,1	200	0,9	0,8	0,982274469	0,003921944
16	3	0,1	200	0,9	0,9	0,982274469	0,003921944
17	7	0,01	300	0,8	0,8	0,981958515	0,004645677
18	7	0,01	300	0,9	0,8	0,981958515	0,004645677
19	7	0,01	300	0,8	0,9	0,981958515	0,004645677
20	7	0,01	300	0,9	0,9	0,981958515	0,004645677
21	3	0,01	300	0,9	0,8	0,981958515	0,00605012
22	3	0,01	300	0,9	0,9	0,981958515	0,00605012
23	5	0,01	200	0,8	0,8	0,981642059	0,005439819
24	5	0,01	200	0,8	0,9	0,981642059	0,005439819
25	3	0,2	100	0,8	0,8	0,981641558	0,00464602
26	3	0,2	100	0,8	0,9	0,981641558	0,00464602
27	5	0,01	200	0,9	0,8	0,981325603	0,005778831
28	7	0,01	200	0,9	0,8	0,981325603	0,005778831
29	5	0,01	200	0,9	0,9	0,981325603	0,005778831
30	7	0,01	200	0,9	0,9	0,981325603	0,005778831
31	5	0,01	300	0,9	0,8	0,981325102	0,004402249
32	5	0,1	200	0,8	0,8	0,981325102	0,004402249
33	3	0,2	100	0,9	0,8	0,981325102	0,004402249
34	5	0,01	300	0,9	0,9	0,981325102	0,004402249
35	5	0,1	200	0,8	0,9	0,981325102	0,004402249
36	3	0,2	100	0,9	0,9	0,981325102	0,004402249

Rank Test Score	Max Depth	Learning Rate	N Estimators	Subsample	Colsample Bytree	*Mean* test score	Std test score
37	5	0,1	100	0,8	0,8	0,981325102	0,003791119
38	5	0,1	100	0,8	0,9	0,981325102	0,003791119
39	7	0,1	100	0,9	0,8	0,981325102	0,004046661
40	7	0,1	100	0,9	0,9	0,981325102	0,004046661
41	7	0,01	200	0,8	0,8	0,981009148	0,006164357
42	7	0,01	200	0,8	0,9	0,981009148	0,006164357
43	7	0,1	100	0,8	0,8	0,981008646	0,00468839
44	7	0,1	100	0,8	0,9	0,981008646	0,00468839
45	7	0,01	100	0,9	0,8	0,981008646	0,006856898
46	7	0,01	100	0,9	0,9	0,981008646	0,006856898
47	7	0,2	100	0,9	0,8	0,981007643	0,003603763
48	7	0,2	100	0,9	0,9	0,981007643	0,003603763
49	7	0,1	300	0,8	0,8	0,981007643	0,004358412
50	7	0,1	300	0,8	0,9	0,981007643	0,004358412
51	5	0,01	100	0,8	0,8	0,98069219	0,006738858
52	5	0,01	100	0,9	0,8	0,98069219	0,006738858
53	7	0,01	100	0,8	0,8	0,98069219	0,006738858
54	5	0,01	100	0,8	0,9	0,98069219	0,006738858
55	5	0,01	100	0,9	0,9	0,98069219	0,006738858
56	7	0,01	100	0,8	0,9	0,98069219	0,006738858
57	3	0,1	300	0,8	0,8	0,980691689	0,004168869
58	3	0,1	300	0,8	0,9	0,980691689	0,004168869
59	3	0,1	200	0,8	0,8	0,980691689	0,004624417
60	3	0,1	200	0,8	0,9	0,980691689	0,004624417
61	5	0,1	200	0,9	0,8	0,980691187	0,004288557
62	5	0,1	200	0,9	0,9	0,980691187	0,004288557
63	7	0,1	200	0,8	0,8	0,980691187	0,004403767
64	7	0,1	200	0,8	0,9	0,980691187	0,004403767
65	5	0,1	300	0,8	0,8	0,980690686	0,004171842
66	5	0,1	300	0,8	0,9	0,980690686	0,004171842
67	7	0,2	100	0,8	0,8	0,980690184	0,004052158
68	7	0,2	100	0,8	0,9	0,980690184	0,004052158
69	7	0,1	300	0,9	0,8	0,980690184	0,004173898
70	7	0,1	300	0,9	0,9	0,980690184	0,004173898
71	3	0,01	100	0,8	0,8	0,980375735	0,006678961
72	3	0,01	100	0,8	0,9	0,980375735	0,006678961
73	3	0,1	300	0,9	0,8	0,980375233	0,004645715
74	3	0,1	300	0,9	0,9	0,980375233	0,004645715
75	5	0,2	100	0,8	0,8	0,980374732	0,004311477
76	5	0,2	100	0,9	0,8	0,980374732	0,004311477

Rank Test Score	Max Depth	Learning Rate	N Estimators	Subsample	Colsample Bytree	*Mean* test score	Std test score
77	5	0,2	100	0,8	0,9	0,980374732	0,004311477
78	5	0,2	100	0,9	0,9	0,980374732	0,004311477
79	5	0,1	300	0,9	0,8	0,98037423	0,004195362
80	5	0,1	300	0,9	0,9	0,98037423	0,004195362
81	3	0,01	100	0,9	0,8	0,980059279	0,006900022
82	3	0,01	100	0,9	0,9	0,980059279	0,006900022
83	7	0,1	200	0,9	0,8	0,979740817	0,003224889
84	7	0,1	200	0,9	0,9	0,979740817	0,003224889
85	5	0,2	200	0,8	0,8	0,979740817	0,003661179
86	5	0,2	200	0,8	0,9	0,979740817	0,003661179
87	3	0,2	200	0,8	0,8	0,979425365	0,005195327
88	3	0,2	200	0,8	0,9	0,979425365	0,005195327
89	7	0,2	300	0,9	0,8	0,97942386	0,003875797
90	7	0,2	300	0,9	0,9	0,97942386	0,003875797
91	7	0,2	200	0,9	0,8	0,979423359	0,003611255
92	7	0,2	200	0,9	0,9	0,979423359	0,003611255
93	3	0,2	300	0,8	0,8	0,979107906	0,00392422
94	7	0,2	200	0,8	0,8	0,979107906	0,00392422
95	3	0,2	300	0,8	0,9	0,979107906	0,00392422
96	7	0,2	200	0,8	0,9	0,979107906	0,00392422
97	3	0,2	200	0,9	0,8	0,978158539	0,004939742
98	3	0,2	200	0,9	0,9	0,978158539	0,004939742
99	5	0,2	200	0,9	0,8	0,978158037	0,003661009
100	5	0,2	200	0,9	0,9	0,978158037	0,003661009
101	5	0,2	300	0,8	0,8	0,978158037	0,003925032
102	5	0,2	300	0,8	0,9	0,978158037	0,003925032
103	7	0,2	300	0,8	0,8	0,978158037	0,004405864
104	7	0,2	300	0,8	0,9	0,978158037	0,004405864
105	5	0,2	300	0,9	0,8	0,977842083	0,00458174
106	5	0,2	300	0,9	0,9	0,977842083	0,00458174
107	3	0,2	300	0,9	0,8	0,976891713	0,004428897
108	3	0,2	300	0,9	0,9	0,976891713	0,004428897

Tabel Kombinasi Parameter Terbaik pada Skenario 80:20

Rank Test Score	Max Depth	Learning Rate	N Estimators	Subsample	Colsample Bytree	*Mean* test score	Std test score
1	3	0,01	300	0,9	0,8	0,981486	0,006114
2	3	0,01	300	0,9	0,9	0,981486	0,006114

Rank Test Score	Max Depth	Learning Rate	N Estimators	Subsample	Colsample Bytree	*Mean* test score	Std test score
3	3	0,01	100	0,8	0,8	0,98113	0,006009
4	5	0,01	100	0,8	0,8	0,98113	0,006009
5	7	0,01	100	0,8	0,8	0,98113	0,006009
6	3	0,01	100	0,8	0,9	0,98113	0,006009
7	5	0,01	100	0,8	0,9	0,98113	0,006009
8	7	0,01	100	0,8	0,9	0,98113	0,006009
9	3	0,01	200	0,9	0,8	0,980774	0,006592
10	3	0,01	200	0,9	0,9	0,980774	0,006592
11	3	0,01	100	0,9	0,8	0,980773	0,005662
12	5	0,01	100	0,9	0,8	0,980773	0,005662
13	7	0,01	100	0,9	0,8	0,980773	0,005662
14	3	0,01	100	0,9	0,9	0,980773	0,005662
15	5	0,01	100	0,9	0,9	0,980773	0,005662
16	7	0,01	100	0,9	0,9	0,980773	0,005662
17	3	0,01	200	0,8	0,8	0,980773	0,005881
18	3	0,01	200	0,8	0,9	0,980773	0,005881
19	3	0,01	300	0,8	0,8	0,980418	0,007109
20	3	0,01	300	0,8	0,9	0,980418	0,007109
21	5	0,01	200	0,8	0,8	0,980062	0,007405
22	5	0,01	200	0,8	0,9	0,980062	0,007405
23	5	0,01	300	0,8	0,8	0,979706	0,007756
24	5	0,01	300	0,8	0,9	0,979706	0,007756
25	5	0,01	200	0,9	0,8	0,978994	0,007405
26	7	0,01	300	0,9	0,8	0,978994	0,007405
27	5	0,01	200	0,9	0,9	0,978994	0,007405
28	7	0,01	300	0,9	0,9	0,978994	0,007405
29	7	0,01	200	0,8	0,8	0,978638	0,007788
30	7	0,01	200	0,9	0,8	0,978638	0,007788
31	7	0,01	200	0,8	0,9	0,978638	0,007788
32	7	0,01	200	0,9	0,9	0,978638	0,007788
33	7	0,01	300	0,8	0,8	0,978638	0,00737
34	7	0,01	300	0,8	0,9	0,978638	0,00737
35	3	0,1	100	0,8	0,8	0,978636	0,005616
36	3	0,1	100	0,8	0,9	0,978636	0,005616
37	5	0,01	300	0,9	0,8	0,978282	0,006872
38	5	0,01	300	0,9	0,9	0,978282	0,006872
39	7	0,1	100	0,9	0,8	0,978282	0,006873
40	7	0,1	100	0,9	0,9	0,978282	0,006873
41	7	0,1	100	0,8	0,8	0,978282	0,007144
42	7	0,1	100	0,8	0,9	0,978282	0,007144

Rank Test Score	Max Depth	Learning Rate	N Estimators	Subsample	Colsample Bytree	*Mean* test score	Std test score
43	7	0,2	100	0,8	0,8	0,978281	0,007575
44	7	0,2	100	0,8	0,9	0,978281	0,007575
45	7	0,1	300	0,9	0,8	0,977925	0,007507
46	7	0,1	300	0,9	0,9	0,977925	0,007507
47	3	0,1	200	0,8	0,8	0,977924	0,005456
48	3	0,1	200	0,8	0,9	0,977924	0,005456
49	5	0,1	100	0,8	0,8	0,977569	0,007422
50	5	0,1	100	0,8	0,9	0,977569	0,007422
51	5	0,1	300	0,8	0,8	0,977569	0,006415
52	5	0,1	300	0,8	0,9	0,977569	0,006415
53	7	0,1	200	0,8	0,8	0,977569	0,006983
54	7	0,1	200	0,8	0,9	0,977569	0,006983
55	3	0,2	100	0,9	0,8	0,977568	0,006007
56	3	0,2	100	0,9	0,9	0,977568	0,006007
57	3	0,1	300	0,9	0,8	0,977568	0,007336
58	3	0,1	300	0,9	0,9	0,977568	0,007336
59	7	0,2	300	0,8	0,8	0,977567	0,005097
60	7	0,2	300	0,8	0,9	0,977567	0,005097
61	3	0,1	100	0,9	0,8	0,977213	0,00609
62	3	0,1	100	0,9	0,9	0,977213	0,00609
63	5	0,1	300	0,9	0,8	0,977213	0,006872
64	5	0,1	300	0,9	0,9	0,977213	0,006872
65	5	0,2	200	0,8	0,8	0,977213	0,007319
66	7	0,2	200	0,9	0,8	0,977213	0,007319
67	5	0,2	200	0,8	0,9	0,977213	0,007319
68	7	0,2	200	0,9	0,9	0,977213	0,007319
69	7	0,2	100	0,9	0,8	0,977213	0,007901
70	7	0,2	100	0,9	0,9	0,977213	0,007901
71	7	0,2	300	0,9	0,8	0,977212	0,005659
72	7	0,2	300	0,9	0,9	0,977212	0,005659
73	7	0,1	300	0,8	0,8	0,977212	0,00588
74	7	0,1	300	0,8	0,9	0,977212	0,00588
75	3	0,2	200	0,9	0,8	0,977212	0,00678
76	3	0,2	200	0,9	0,9	0,977212	0,00678
77	7	0,1	200	0,9	0,8	0,976858	0,007868
78	7	0,1	200	0,9	0,9	0,976858	0,007868
79	5	0,2	100	0,8	0,8	0,976857	0,007196
80	5	0,2	100	0,8	0,9	0,976857	0,007196
81	7	0,2	200	0,8	0,8	0,976856	0,006255
82	7	0,2	200	0,8	0,9	0,976856	0,006255

Rank Test Score	Max Depth	Learning Rate	N Estimators	Subsample	Colsample Bytree	*Mean* test score	Std test score
83	3	0,2	300	0,9	0,8	0,976856	0,006355
84	3	0,2	300	0,9	0,9	0,976856	0,006355
85	3	0,1	200	0,9	0,8	0,976856	0,006834
86	5	0,2	300	0,9	0,8	0,976856	0,006834
87	3	0,1	200	0,9	0,9	0,976856	0,006834
88	5	0,2	300	0,9	0,9	0,976856	0,006834
89	3	0,2	100	0,8	0,8	0,976856	0,007109
90	3	0,2	100	0,8	0,9	0,976856	0,007109
91	3	0,2	300	0,8	0,8	0,976856	0,005385
92	3	0,2	300	0,8	0,9	0,976856	0,005385
93	5	0,2	200	0,9	0,8	0,976856	0,006836
94	5	0,2	200	0,9	0,9	0,976856	0,006836
95	5	0,1	200	0,8	0,8	0,976501	0,007142
96	5	0,1	200	0,8	0,9	0,976501	0,007142
97	3	0,2	200	0,8	0,8	0,9765	0,006685
98	3	0,2	200	0,8	0,9	0,9765	0,006685
99	5	0,2	100	0,9	0,8	0,9765	0,008059
100	5	0,2	100	0,9	0,9	0,9765	0,008059
101	3	0,1	300	0,8	0,8	0,976499	0,00566
102	3	0,1	300	0,8	0,9	0,976499	0,00566
103	5	0,1	200	0,9	0,8	0,976144	0,007592
104	5	0,1	200	0,9	0,9	0,976144	0,007592
105	5	0,2	300	0,8	0,8	0,976143	0,005455
106	5	0,2	300	0,8	0,9	0,976143	0,005455
107	5	0,1	100	0,9	0,8	0,975789	0,007671
108	5	0,1	100	0,9	0,9	0,975789	0,007671

Tabel Kombinasi Parameter Terbaik pada Skenario 70:30

Rank Test Score	Max Depth	Learning Rate	N Estimators	Subsample	Colsample Bytree	*Mean* test score	Std test score
1	7	0,01	300	0,9	0,8	0,98128	0,003937
2	7	0,01	300	0,9	0,9	0,98128	0,003937
3	5	0,01	300	0,9	0,8	0,980466	0,004739
4	5	0,01	300	0,9	0,9	0,980466	0,004739
5	3	0,01	200	0,9	0,8	0,980061	0,005197
6	3	0,01	200	0,9	0,9	0,980061	0,005197
7	5	0,01	200	0,9	0,8	0,979653	0,00545
8	5	0,01	200	0,9	0,9	0,979653	0,00545

Rank Test Score	Max Depth	Learning Rate	N Estimators	Subsample	Colsample Bytree	*Mean* test score	Std test score
9	7	0,01	300	0,8	0,8	0,979652	0,004453
10	7	0,01	300	0,8	0,9	0,979652	0,004453
11	3	0,01	300	0,9	0,8	0,979246	0,004699
12	5	0,01	200	0,8	0,8	0,979246	0,004699
13	3	0,01	300	0,9	0,9	0,979246	0,004699
14	5	0,01	200	0,8	0,9	0,979246	0,004699
15	5	0,1	100	0,8	0,8	0,979243	0,003257
16	3	0,2	100	0,9	0,8	0,979243	0,003257
17	5	0,1	100	0,8	0,9	0,979243	0,003257
18	3	0,2	100	0,9	0,9	0,979243	0,003257
19	5	0,1	100	0,9	0,8	0,979242	0,002707
20	5	0,1	100	0,9	0,9	0,979242	0,002707
21	3	0,01	300	0,8	0,8	0,978839	0,004904
22	3	0,01	300	0,8	0,9	0,978839	0,004904
23	3	0,01	200	0,8	0,8	0,978839	0,00523
24	3	0,01	200	0,8	0,9	0,978839	0,00523
25	7	0,01	200	0,8	0,8	0,978839	0,003974
26	7	0,01	200	0,8	0,9	0,978839	0,003974
27	5	0,01	300	0,8	0,8	0,978838	0,003766
28	5	0,01	300	0,8	0,9	0,978838	0,003766
29	3	0,1	100	0,9	0,8	0,978834	0,002091
30	3	0,1	100	0,9	0,9	0,978834	0,002091
31	5	0,2	100	0,8	0,8	0,978834	0,002773
32	5	0,2	100	0,8	0,9	0,978834	0,002773
33	5	0,01	100	0,9	0,8	0,978432	0,005388
34	7	0,01	100	0,9	0,8	0,978432	0,005388
35	5	0,01	100	0,9	0,9	0,978432	0,005388
36	7	0,01	100	0,9	0,9	0,978432	0,005388
37	7	0,01	200	0,9	0,8	0,978431	0,00491
38	7	0,01	200	0,9	0,9	0,978431	0,00491
39	3	0,1	100	0,8	0,8	0,978429	0,001631
40	3	0,1	100	0,8	0,9	0,978429	0,001631
41	5	0,01	100	0,8	0,8	0,978026	0,00594
42	7	0,01	100	0,8	0,8	0,978026	0,00594
43	5	0,01	100	0,8	0,9	0,978026	0,00594
44	7	0,01	100	0,8	0,9	0,978026	0,00594
45	3	0,01	100	0,8	0,8	0,978026	0,006345
46	3	0,01	100	0,8	0,9	0,978026	0,006345
47	5	0,1	200	0,8	0,8	0,978021	0,004352
48	5	0,1	200	0,8	0,9	0,978021	0,004352

Rank Test Score	Max Depth	Learning Rate	N Estimators	Subsample	Colsample Bytree	*Mean* test score	Std test score
49	7	0,1	100	0,9	0,8	0,978021	0,003508
50	7	0,1	100	0,9	0,9	0,978021	0,003508
51	7	0,1	300	0,9	0,8	0,97802	0,004892
52	7	0,1	300	0,9	0,9	0,97802	0,004892
53	3	0,01	100	0,9	0,8	0,977618	0,006161
54	3	0,01	100	0,9	0,9	0,977618	0,006161
55	3	0,1	200	0,9	0,8	0,977615	0,002575
56	3	0,1	200	0,9	0,9	0,977615	0,002575
57	7	0,1	100	0,8	0,8	0,977614	0,003863
58	7	0,1	100	0,8	0,9	0,977614	0,003863
59	7	0,1	200	0,9	0,8	0,977612	0,005316
60	7	0,1	200	0,9	0,9	0,977612	0,005316
61	3	0,1	200	0,8	0,8	0,977208	0,001527
62	3	0,1	200	0,8	0,9	0,977208	0,001527
63	3	0,2	100	0,8	0,8	0,977208	0,001998
64	3	0,2	100	0,8	0,9	0,977208	0,001998
65	5	0,1	300	0,9	0,8	0,977206	0,003956
66	5	0,1	300	0,9	0,9	0,977206	0,003956
67	5	0,2	200	0,8	0,8	0,9768	0,003313
68	5	0,2	200	0,8	0,9	0,9768	0,003313
69	5	0,2	100	0,9	0,8	0,976799	0,002772
70	5	0,2	100	0,9	0,9	0,976799	0,002772
71	3	0,2	200	0,9	0,8	0,976799	0,00378
72	3	0,2	200	0,9	0,9	0,976799	0,00378
73	5	0,1	300	0,8	0,8	0,976394	0,002761
74	5	0,1	300	0,8	0,9	0,976394	0,002761
75	5	0,1	200	0,9	0,8	0,976394	0,003308
76	5	0,1	200	0,9	0,9	0,976394	0,003308
77	7	0,2	100	0,9	0,8	0,976393	0,002086
78	7	0,2	100	0,9	0,9	0,976393	0,002086
79	7	0,2	200	0,9	0,8	0,976393	0,00277
80	7	0,2	200	0,9	0,9	0,976393	0,00277
81	3	0,1	300	0,8	0,8	0,976393	0,003051
82	3	0,1	300	0,8	0,9	0,976393	0,003051
83	7	0,2	300	0,9	0,8	0,976393	0,00378
84	7	0,2	300	0,9	0,9	0,976393	0,00378
85	7	0,1	300	0,8	0,8	0,976392	0,003996
86	7	0,1	300	0,8	0,9	0,976392	0,003996
87	7	0,1	200	0,8	0,8	0,976392	0,003786
88	7	0,1	200	0,8	0,9	0,976392	0,003786

Rank Test Score	Max Depth	Learning Rate	N Estimators	Subsample	Colsample Bytree	*Mean* test score	Std test score
89	3	0,2	300	0,9	0,8	0,976391	0,005092
90	3	0,2	300	0,9	0,9	0,976391	0,005092
91	7	0,2	300	0,8	0,8	0,975987	0,00237
92	7	0,2	300	0,8	0,9	0,975987	0,00237
93	5	0,2	300	0,8	0,8	0,975986	0,00238
94	5	0,2	300	0,8	0,9	0,975986	0,00238
95	7	0,2	200	0,8	0,8	0,975986	0,002995
96	7	0,2	200	0,8	0,9	0,975986	0,002995
97	3	0,1	300	0,9	0,8	0,975985	0,003956
98	5	0,2	300	0,9	0,8	0,975985	0,003956
99	3	0,1	300	0,9	0,9	0,975985	0,003956
100	5	0,2	300	0,9	0,9	0,975985	0,003956
101	5	0,2	200	0,9	0,8	0,975984	0,004894
102	5	0,2	200	0,9	0,9	0,975984	0,004894
103	7	0,2	100	0,8	0,8	0,975581	0,003402
104	7	0,2	100	0,8	0,9	0,975581	0,003402
105	3	0,2	300	0,8	0,8	0,974766	0,002071
106	3	0,2	300	0,8	0,9	0,974766	0,002071
107	3	0,2	200	0,8	0,8	0,974358	0,003052
108	3	0,2	200	0,8	0,9	0,974358	0,003052

Tabel Kombinasi Parameter Terbaik pada Skenario 60:40

Rank Test Score	Max Depth	Learning Rate	N Estimators	Subsample	Colsample Bytree	*Mean* test score	Std test score
1	5	0,01	200	0,8	0,8	0,981959	0,006469
2	7	0,01	200	0,8	0,8	0,981959	0,006469
3	5	0,01	200	0,8	0,9	0,981959	0,006469
4	7	0,01	200	0,8	0,9	0,981959	0,006469
5	3	0,01	200	0,8	0,8	0,981485	0,006936
6	3	0,01	200	0,9	0,8	0,981485	0,006936
7	3	0,01	200	0,8	0,9	0,981485	0,006936
8	3	0,01	200	0,9	0,9	0,981485	0,006936
9	5	0,01	200	0,9	0,8	0,981484	0,006608
10	7	0,01	200	0,9	0,8	0,981484	0,006608
11	5	0,01	200	0,9	0,9	0,981484	0,006608
12	7	0,01	200	0,9	0,9	0,981484	0,006608
13	3	0,01	300	0,9	0,8	0,981009	0,006
14	3	0,01	300	0,9	0,9	0,981009	0,006

Rank Test Score	Max Depth	Learning Rate	N Estimators	Subsample	Colsample Bytree	*Mean* test score	Std test score
15	3	0,01	300	0,8	0,8	0,980535	0,006431
16	3	0,01	300	0,8	0,9	0,980535	0,006431
17	3	0,1	100	0,8	0,8	0,980533	0,005496
18	3	0,1	100	0,8	0,9	0,980533	0,005496
19	5	0,01	100	0,9	0,8	0,980061	0,006965
20	7	0,01	100	0,9	0,8	0,980061	0,006965
21	5	0,01	100	0,9	0,9	0,980061	0,006965
22	7	0,01	100	0,9	0,9	0,980061	0,006965
23	7	0,01	300	0,8	0,8	0,980059	0,005108
24	7	0,01	300	0,8	0,9	0,980059	0,005108
25	3	0,01	100	0,9	0,8	0,979586	0,006462
26	3	0,01	100	0,9	0,9	0,979586	0,006462
27	5	0,01	300	0,8	0,8	0,979584	0,005326
28	5	0,01	300	0,9	0,8	0,979584	0,005326
29	7	0,01	300	0,9	0,8	0,979584	0,005326
30	5	0,01	300	0,8	0,9	0,979584	0,005326
31	5	0,01	300	0,9	0,9	0,979584	0,005326
32	7	0,01	300	0,9	0,9	0,979584	0,005326
33	3	0,01	100	0,8	0,8	0,979111	0,005878
34	5	0,01	100	0,8	0,8	0,979111	0,005878
35	7	0,01	100	0,8	0,8	0,979111	0,005878
36	3	0,01	100	0,8	0,9	0,979111	0,005878
37	5	0,01	100	0,8	0,9	0,979111	0,005878
38	7	0,01	100	0,8	0,9	0,979111	0,005878
39	3	0,1	100	0,9	0,8	0,978632	0,004507
40	3	0,1	100	0,9	0,9	0,978632	0,004507
41	5	0,1	100	0,8	0,8	0,978159	0,006614
42	5	0,1	100	0,8	0,9	0,978159	0,006614
43	7	0,1	100	0,8	0,8	0,977685	0,006812
44	7	0,1	100	0,8	0,9	0,977685	0,006812
45	3	0,2	100	0,9	0,8	0,977683	0,003855
46	3	0,2	100	0,9	0,9	0,977683	0,003855
47	3	0,1	200	0,9	0,8	0,97721	0,004396
48	3	0,1	300	0,9	0,8	0,97721	0,004396
49	3	0,1	200	0,9	0,9	0,97721	0,004396
50	3	0,1	300	0,9	0,9	0,97721	0,004396
51	3	0,2	100	0,8	0,8	0,97721	0,005325
52	3	0,2	100	0,8	0,9	0,97721	0,005325
53	3	0,1	300	0,8	0,8	0,977208	0,004889
54	3	0,1	300	0,8	0,9	0,977208	0,004889

Rank Test Score	Max Depth	Learning Rate	N Estimators	Subsample	Colsample Bytree	*Mean* test score	Std test score
55	7	0,2	300	0,9	0,8	0,976736	0,005058
56	7	0,2	300	0,9	0,9	0,976736	0,005058
57	5	0,1	200	0,8	0,8	0,976736	0,005486
58	5	0,1	200	0,8	0,9	0,976736	0,005486
59	7	0,1	100	0,9	0,8	0,976736	0,006939
60	7	0,1	100	0,9	0,9	0,976736	0,006939
61	7	0,2	100	0,9	0,8	0,976734	0,003791
62	7	0,2	100	0,9	0,9	0,976734	0,003791
63	7	0,1	200	0,9	0,8	0,976734	0,004598
64	7	0,1	200	0,9	0,9	0,976734	0,004598
65	5	0,2	200	0,9	0,8	0,976259	0,003967
66	5	0,2	200	0,9	0,9	0,976259	0,003967
67	5	0,2	100	0,8	0,8	0,976259	0,005411
68	5	0,2	100	0,8	0,9	0,976259	0,005411
69	5	0,1	100	0,9	0,8	0,976259	0,006714
70	5	0,1	100	0,9	0,9	0,976259	0,006714
71	7	0,2	200	0,9	0,8	0,975785	0,003474
72	7	0,2	200	0,9	0,9	0,975785	0,003474
73	3	0,2	200	0,9	0,8	0,975785	0,004341
74	5	0,2	300	0,9	0,8	0,975785	0,004341
75	3	0,2	200	0,9	0,9	0,975785	0,004341
76	5	0,2	300	0,9	0,9	0,975785	0,004341
77	7	0,1	300	0,8	0,8	0,975785	0,005279
78	7	0,1	300	0,9	0,8	0,975785	0,005279
79	7	0,1	300	0,8	0,9	0,975785	0,005279
80	7	0,1	300	0,9	0,9	0,975785	0,005279
81	3	0,1	200	0,8	0,8	0,975785	0,006257
82	3	0,1	200	0,8	0,9	0,975785	0,006257
83	7	0,1	200	0,8	0,8	0,975784	0,004349
84	7	0,2	100	0,8	0,8	0,975784	0,004349
85	7	0,1	200	0,8	0,9	0,975784	0,004349
86	7	0,2	100	0,8	0,9	0,975784	0,004349
87	5	0,1	200	0,9	0,8	0,97531	0,003846
88	5	0,1	200	0,9	0,9	0,97531	0,003846
89	5	0,1	300	0,9	0,8	0,97531	0,003846
90	5	0,1	300	0,9	0,9	0,97531	0,003846
91	5	0,1	300	0,8	0,8	0,975309	0,003855
92	5	0,1	300	0,8	0,9	0,975309	0,003855
93	7	0,2	200	0,8	0,8	0,974836	0,0051
94	7	0,2	200	0,8	0,9	0,974836	0,0051

Rank Test Score	Max Depth	Learning Rate	N Estimators	Subsample	Colsample Bytree	*Mean* test score	Std test score
95	3	0,2	300	0,9	0,8	0,974835	0,003208
96	3	0,2	300	0,9	0,9	0,974835	0,003208
97	3	0,2	300	0,8	0,8	0,974835	0,004395
98	3	0,2	300	0,8	0,9	0,974835	0,004395
99	3	0,2	200	0,8	0,8	0,974361	0,006605
100	3	0,2	200	0,8	0,9	0,974361	0,006605
101	5	0,2	200	0,8	0,8	0,97436	0,00379
102	5	0,2	200	0,8	0,9	0,97436	0,00379
103	5	0,2	100	0,9	0,8	0,97436	0,004598
104	5	0,2	100	0,9	0,9	0,97436	0,004598
105	7	0,2	300	0,8	0,8	0,973886	0,004969
106	7	0,2	300	0,8	0,9	0,973886	0,004969
107	5	0,2	300	0,8	0,8	0,973411	0,004592
108	5	0,2	300	0,8	0,9	0,973411	0,004592

Surat Keterangan Telah Melakukan Penelitian

1. Desa Baureno

PEMERINTAH KABUPATEN BOJONEGORO
KECAMATAN BAURENO
KANTOR DESA BAURENO
Jl. Raya Baureno Nomor 316 Desa Baureno Kodepos 62192
BAURENO

SURAT KETERANGAN
Nomor : 470 / 458 / 412.406.2006 / 2026

Yang bertanda tangan di bawah ini Kepala Desa Baureno, Kecamatan Baureno, Kabupaten Bojonegoro, menerangkan bahwa :

1.	Nama	: NAILA NABILA ILIANA
2.	NIM	: 200605110110
3.	Jurusan/Prodi	: Teknik Informatika
4.	Fakultas	: Sains dan Teknologi
5.	Perguruan Tinggi	: Universitas Islam Negeri Maulana Malik Ibrahim Malang
8.	**Keterangan**	: Sampai dengan dibuatnya Surat Keterangan ini, yang bersangkutan benar-benar telah melakukan penelitian kualitas air bersih di Desa Baureno Kecamatan Baureno pada tanggal 4 Mei 2026.

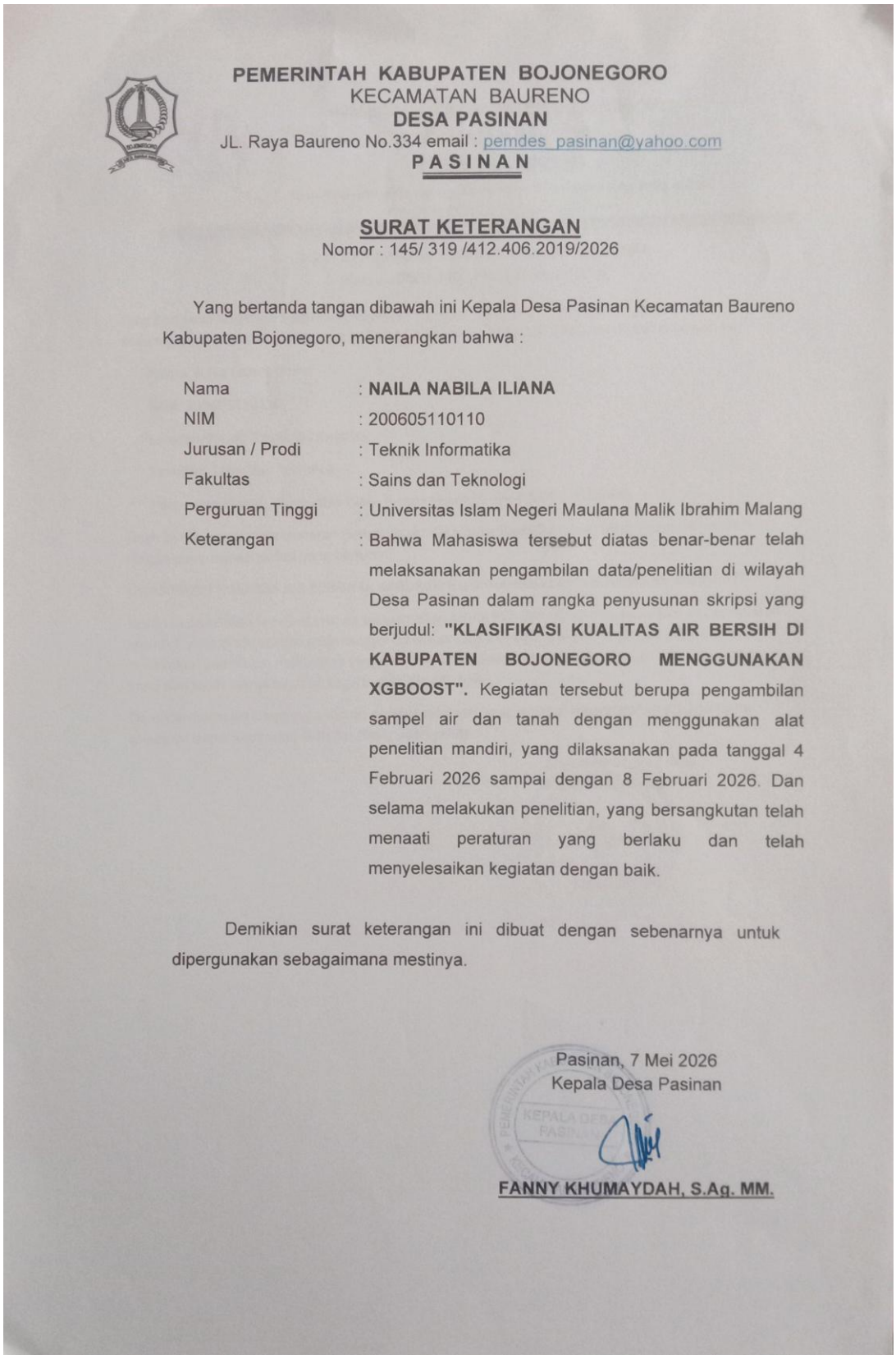
Demikian Surat Keterangan ini dibuat dengan sebenar-benarnya dan dapat dipergunakan sebagaimana mestinya.

Baureno, 7 Mei 2026
AN. KEPALA DESA BAURENO
SEKRETARIS DESA

PEMERINTAH KABUPATEN BOJONEGORO KECAMATAN BAURENO KEPALA DESA BAURENO

SYAIFUL ARIF

2. Desa Pasinan



PEMERINTAH KABUPATEN BOJONEGORO
KECAMATAN BAURENO
DESA PASINAN
JL. Raya Baureno No.334 email : pemdes_pasinan@yahoo.com
PASINAN

SURAT KETERANGAN
Nomor : 145/ 319 /412.406.2019/2026

Yang bertanda tangan dibawah ini Kepala Desa Pasinan Kecamatan Baureno Kabupaten Bojonegoro, menerangkan bahwa :

Nama	: **NAILA NABILA ILIANA**
NIM	: 200605110110
Jurusan / Prodi	: Teknik Informatika
Fakultas	: Sains dan Teknologi
Perguruan Tinggi	: Universitas Islam Negeri Maulana Malik Ibrahim Malang
Keterangan	: Bahwa Mahasiswa tersebut diatas benar-benar telah melaksanakan pengambilan data/penelitian di wilayah Desa Pasinan dalam rangka penyusunan skripsi yang berjudul: **"KLASIFIKASI KUALITAS AIR BERSIH DI KABUPATEN BOJONEGORO MENGGUNAKAN XGBOOST".** Kegiatan tersebut berupa pengambilan sampel air dan tanah dengan menggunakan alat penelitian mandiri, yang dilaksanakan pada tanggal 4 Februari 2026 sampai dengan 8 Februari 2026. Dan selama melakukan penelitian, yang bersangkutan telah menaati peraturan yang berlaku dan telah menyelesaikan kegiatan dengan baik.

Demikian surat keterangan ini dibuat dengan sebenarnya untuk dipergunakan sebagaimana mestinya.

Pasinan, 7 Mei 2026
Kepala Desa Pasinan

FANNY KHUMAYDAH, S.Ag. MM.

3. Desa Sraturejo

PEMERINTAH KABUPATEN BOJONEGORO
KECAMATAN BAURENO
DESA SRATUREJO

Jalan Raya Sraturejo Nomor 76 Baureno Bojonegoro Telp. 0322 456202
e-mail : pemdes.sraturejo@gmail.com

SURAT KETERANGAN TELAH MELAKUKAN PENELITIAN
Nomor : 140 / 364 / 412.406.2004 / 2026

Yang bertanda tangan di bawah ini, Kepala Desa **Sraturejo**, Kecamatan Baureno, Kabupaten Bojonegoro, dengan ini menerangkan bahwa mahasiswa yang namanya tercantum di bawah ini:

Nama: Naila Nabila Iliana

NIM: 200605110110

Jurusan/Prodi: Teknik Informatika

Fakultas: Sains dan Teknologi

Perguruan Tinggi: Universitas islam Negeri Maulana Malik Ibrahim Malang

Telah benar-benar melaksanakan pengambilan data/penelitian di wilayah Desa Sraturejo dalam rangka penyusunan skripsi yang berjudul:

"KLASIFIKASI KUALITAS AIR BERSIH DI KABUPATEN BOJONEGORO MENGGUNAKAN *XGBOOST*"

Kegiatan penelitian tersebut berupa pengambilan sampel air dan tanah menggunakan alat penelitian mandiri, yang dilaksanakan pada tanggal 8 Februari 2026 sampai dengan 9Februari 2026. Selama melakukan penelitian, mahasiswa yang bersangkutan telah menaati peraturan yang berlaku di desa kami dan telah menyelesaikan kegiatannya dengan baik.

Demikian surat keterangan ini dibuat dengan sebenarnya agar dapat dipergunakan sebagaimana mestinya untuk keperluan Seminar Hasil/Sidang Skripsi.

Sraturejo, 07 Mei 2026
Kepala Desa Sraturejo

(Hj.KUSRINI)