BAB II
TINJAUAN PUSTAKA

1.1 Penelitian Relevan

Kajian kualitas dan pencemaran air laut telah dilakukan di berbagai negara dengan metode manual maupun *machine learning*, mengingat keterbatasan metode konvensional yang cenderung lambat, mahal, dan sering menghasilkan ketidakpastian.

Penelitian kualitas air laut di Cork Harbour, Irlandia, dilakukan pada 29 titik pemantauan dengan menggunakan 11 parameter kualitas air, yaitu suhu, TON, amonia, DO, pH, salinitas, fosfat, BOD, transparansi, dan klorofil-a, yang dipilih karena merepresentasikan indikator utama kondisi perairan. Penelitian ini menguji delapan algoritma pembelajaran mesin, termasuk XGBoost, *Random Forest*, dan *Decision Tree*, dengan tujuan membandingkan performa model prediksi kualitas air. Hasil penelitian menunjukkan bahwa XGBoost memberikan performa paling unggul dengan nilai R^2 sebesar 1,0, RMSE 0,0, dan PREI ±0,16%, serta menghasilkan klasifikasi WQI yang terdiri atas 44,82% lokasi pada kategori *fair* dan 55,18% pada kategori *marginal*. Temuan ini menegaskan keunggulan XGBoost dalam menekan tingkat ketidakpastian prediksi. Namun, penelitian tersebut masih berfokus pada konteks ekosistem dan regulasi wilayah Eropa. Oleh karena itu, penelitian ini mengembangkan pendekatan tersebut dengan memanfaatkan XGBoost untuk klasifikasi multi-kelas secara langsung guna memprediksi kategori status mutu air laut sesuai standar nasional, sehingga lebih adaptif dalam mendukung upaya konservasi biota laut di Teluk Jakarta (Uddin, Nash, Mahammad Diganta, et al., 2022).

Penelitian kualitas air laut di pesisir Kota Makassar dilakukan dengan menggunakan delapan variabel, yaitu BOD, DO, fosfat, nitrat, pH, fenol, sulfida, dan salinitas, yang ditetapkan berdasarkan Kepmen LH Nomor 51 Tahun 2004. Analisis status mutu air dilakukan menggunakan metode Indeks Pencemaran (PI) sesuai Kepmen LH Nomor 115 Tahun 2003 sebagai standar nasional dalam klasifikasi mutu air. Hasil penelitian menunjukkan bahwa seluruh lokasi pengamatan berada pada kategori tercemar ringan, dengan nilai Indeks Pencemaran tertinggi sebesar 3,2 yang ditemukan pada kawasan pelabuhan saat musim hujan. Penelitian ini relevan sebagai rujukan penerapan

metode Indeks Pencemaran, namun masih memiliki keterbatasan pada jumlah parameter yang digunakan serta pendekatan analisis yang bersifat manual. Oleh karena itu, penelitian ini memperluas cakupan parameter kualitas air yang mencakup aspek fisika, kimia, biologi, dan logam berat, serta mengintegrasikannya dengan algoritma XGBoost untuk klasifikasi multi-kelas, sehingga diharapkan mampu menghasilkan klasifikasi status pencemaran yang lebih adaptif dan akurat (Suharto et al., 2019).

Penelitian mengenai *Assessment of Marine Water Quality in the Bay of Bengal Using Improved WQI Models* dilakukan dengan menggunakan variabel fisika-kimia yang meliputi salinitas, suhu, pH, transparansi, DO, total nitrogen teroksidasi, dan *molybdate reactive phosphorus*, yang dipilih berdasarkan standar EPA, Uni Eropa, dan ATSEBI. Penelitian ini mengembangkan dua model matematis indeks kualitas air, yaitu WQM-WQI dan RMS-WQI, yang kemudian divalidasi menggunakan algoritma XGBoost untuk mengurangi tingkat ketidakpastian dalam penilaian kualitas air laut. Hasil penelitian menunjukkan nilai WQI rata-rata berada pada kisaran 72–77, dengan 85–90% lokasi pengamatan termasuk dalam kategori *fair* dan 10–15% dalam kategori *good*. Penelitian ini relevan karena menerapkan kerangka indeks kualitas air yang terintegrasi dengan *machine learning*, namun masih memiliki keterbatasan pada cakupan variabel yang hanya mencakup parameter fisika-kimia serta penggunaan acuan standar internasional. Dengan demikian, terdapat *gap* penelitian pada aspek kelengkapan parameter dan kesesuaian regulasi, karena penelitian tersebut hanya menguji model global, sedangkan penelitian ini mengusulkan pendekatan multi-parameter dan klasifikasi multi-kelas yang lebih aplikatif untuk mendukung konservasi perairan laut di Teluk Jakarta (Uddin, Rahman, et al., 2023).

Penelitian kualitas air di wilayah Telangana, India, dilakukan dengan menggunakan 958 sampel data kualitas air yang dikumpulkan selama Periode II018–2020. Parameter fisika-kimia yang digunakan meliputi TDS, *fluoride*, nitrat, pH, kalsium, kalium, bikarbonat, karbonat, konduktivitas listrik, sulfat, dan klorida, yang dipilih karena kesesuaiannya dengan standar WHO. Penelitian ini menerapkan metode *ensemble learning* dengan mengombinasikan beberapa algoritma, yaitu *Decision Tree*, *Logistic Regression*, dan *Support Vector Machine*, melalui skema *hard voting* dan *soft voting* untuk mengurangi bias model tunggal serta meningkatkan kemampuan generalisasi. Hasil

penelitian menunjukkan bahwa metode *soft voting* memberikan performa terbaik dengan nilai akurasi sebesar 96,39%, precision 96,49%, recall 96,39%, dan F1-score 96,41%, yang lebih tinggi sebesar 1,46% dibandingkan *base learner* terbaik serta mampu menurunkan tingkat kesalahan sebesar 27,8%. Meskipun relevan karena sama-sama memanfaatkan pendekatan *ensemble learning*, penelitian ini masih terbatas pada konteks perairan darat dan parameter fisika-kimia tertentu. Penelitian ini selanjutnya berfokus pada perairan laut dengan pendekatan klasifikasi multi-kelas serta menggunakan parameter kualitas air yang lebih komprehensif sesuai dengan regulasi nasional (Singh et al., 2025).

Penelitian lain menilai kualitas air dengan menggunakan enam parameter utama, yaitu BOD, DO, pH, fosfat, amonia, dan nitrat, yang dipilih berdasarkan *DENR Administrative Order 2016-08*. Indeks kualitas air dihitung menggunakan metode *Weighted Arithmetic Water Quality Index* (WAWQI), kemudian diklasifikasikan ke dalam kategori *Excellent* hingga *Unsuitable*. Untuk meningkatkan akurasi klasifikasi, penelitian ini membandingkan performa XGBoost dengan *Decision Tree* (DT) dan *Random Forest* (RF). Hasil evaluasi menunjukkan bahwa XGBoost memberikan kinerja paling unggul dengan akurasi sebesar 96,72%, *precision* 96,4%, *recall* 97%, dan *F1-score* 96,6%, melampaui DT (93,45%) dan RF (93,35%). Selain itu, hasil perhitungan WQI mengindikasikan bahwa hampir seluruh stasiun pemantauan berada pada kategori *Unsuitable* atau *Very Poor*, dengan nilai ekstrem tercatat pada *Station* V tahun 2017 (WQI = 170,80) dan *Station* I tahun 2022 (WQI = 205,93). Temuan ini memperkuat relevansi penggunaan XGBoost untuk klasifikasi multi-kelas berbasis regulasi resmi. Namun demikian, penelitian tersebut masih terbatas pada ekosistem air tawar dan jumlah variabel yang relatif sempit. Berbeda dengan penelitian ini, pendekatan yang diusulkan mengintegrasikan parameter multi-aspek serta berfokus pada pencemaran perairan laut, sehingga lebih komprehensif dan aplikatif dalam mendukung perumusan kebijakan konservasi lingkungan pesisir (Limbago, 2023).

Penelitian yang dilakukan di Cork Harbour, Irlandia, menganalisis kualitas air laut menggunakan sepuluh parameter utama, yaitu suhu, pH, DO, total oxidized nitrogen (TON), amonia, *molybdate reactive phosphorus* (MRP), BOD_5, transparansi, klorofil-a, dan *dissolved inorganic nitrogen* (DIN). Untuk mengurangi bias yang berpotensi muncul

dari penggunaan satu model, penelitian ini menguji empat algoritma klasifikasi, yaitu *Naïve Bayes, Support Vector Machine* (SVM), *K-Nearest Neighbor* (KNN), dan XGBoost. Hasil evaluasi menunjukkan bahwa XGBoost memberikan kinerja paling optimal dengan tingkat akurasi mencapai 100% (AUC = 1,0; F1-score = 0,99), melampaui KNN (94%), SVM (92%), dan Naïve Bayes (91%). Analisis kepentingan fitur mengidentifikasi transparansi, klorofil-a, amonia, dan MRP sebagai variabel yang paling dominan dalam menentukan status mutu perairan. Relevansi penelitian ini terletak pada penerapan XGBoost untuk klasifikasi multi-kelas berbasis *Water Quality Index* (WQI) yang efektif dalam meningkatkan akurasi prediksi. Namun demikian, *gap* penelitian masih terlihat pada keterbatasan konteks wilayah Eropa serta cakupan parameter yang belum mencakup logam berat dan indikator biologi. Berbeda dengan penelitian tersebut, penelitian ini mengembangkan pendekatan yang lebih komprehensif dengan menambahkan parameter logam berat dan biologi serta mengacu pada regulasi nasional, sehingga lebih representatif untuk analisis pencemaran perairan laut di Teluk Jakarta (Uddin, Nash, et al., 2023).

Penelitian *Coastal Water Quality Prediction* di Rijeka, Kroasia, berfokus pada pemodelan kualitas perairan pesisir dengan menitikberatkan pada bakteri fekal *Escherichia coli* dan enterokokus sebagai indikator utama kesehatan publik. Dataset yang digunakan mencakup 33 parameter fisika-kimia dan meteorologi, seperti salinitas, suhu, kelembapan, curah hujan, tekanan udara, dan kecepatan angin, yang dipilih karena berpengaruh terhadap dinamika dan sebaran bakteri di perairan pesisir. Pendekatan yang digunakan melibatkan berbagai metode regresi berbasis *machine learning*, yaitu CatBoost, XGBoost, *Random Forest* (RF), *Support Vector Regression* (SVR), dan *Artificial Neural Network* (ANN), guna membandingkan kinerja prediktif masing-masing model. Hasil penelitian menunjukkan bahwa CatBoost memberikan performa terbaik dengan nilai koefisien determinasi sebesar $R^2 = 0,71$ untuk *E. coli* dan $R^2 = 0,68$ untuk enterokokus, serta peningkatan kinerja pada validasi spasial dengan R^2 masing-masing sebesar 0,85 dan 0,83. Meskipun demikian, penelitian ini masih berfokus pada aspek kesehatan publik di kawasan pantai wisata dan menggunakan indikator biologis terbatas pada bakteri fekal. Berbeda dengan penelitian tersebut, penelitian ini mengadopsi pendekatan yang lebih komprehensif dengan memasukkan parameter fisika, kimia,

biologi, dan logam terlarut untuk menilai status mutu air laut secara menyeluruh, serta diarahkan untuk mendukung konservasi ekosistem laut dan pengelolaan lingkungan pesisir yang berkelanjutan sesuai dengan regulasi nasional. (Grbčić et al., 2022).

Penelitian yang mengkaji pencemaran logam berat di perairan pesisir Kepulauan Sangihe, Sulawesi Utara, menitikberatkan pada dampak aktivitas penambangan emas dengan menganalisis parameter tembaga (Cu), besi (Fe), merkuri (Hg), timbal (Pb), dan seng (Zn). Analisis konsentrasi logam dilakukan menggunakan metode *Atomic Absorption Spectrophotometry* (AAS), sedangkan penentuan status mutu air mengacu pada metode Storet sebagai regulasi nasional. Hasil penelitian menunjukkan bahwa konsentrasi Pb (0,6242–0,7318 mg/L) dan Cu (0,039 mg/L) telah melampaui baku mutu sebesar 0,008 mg/L, sehingga kualitas perairan dikategorikan sebagai tercemar berat dengan skor Storet berkisar antara −36 hingga −40. Relevansi penelitian ini terletak pada penerapan regulasi resmi nasional dalam penilaian mutu air laut. Namun demikian, *gap* penelitian masih terlihat pada keterbatasan parameter yang hanya mencakup logam berat serta penggunaan metode evaluasi yang bersifat manual. Berbeda dengan penelitian tersebut, penelitian ini mengembangkan pendekatan yang lebih komprehensif melalui integrasi parameter fisika, kimia, biologi, dan logam terlarut, serta menerapkan klasifikasi multi-kelas berbasis *machine learning* untuk mendukung pengelolaan dan konservasi ekosistem laut secara lebih adaptif dan berkelanjutan (Nyoman, 2021).

Secara keseluruhan, penelitian terdahulu menunjukkan bahwa metode manual seperti Indeks Pencemaran memiliki peran penting sebagai dasar penilaian mutu air, namun masih terbatas pada analisis yang bersifat deskriptif dan kurang adaptif terhadap kompleksitas data multi-parameter. Di sisi lain, penelitian berbasis *ensemble learning*, khususnya algoritma XGBoost, secara konsisten menunjukkan kinerja unggul dengan tingkat akurasi berkisar antara 96–100% dalam tugas klasifikasi kualitas air. Meskipun demikian, gap penelitian yang berulang masih terlihat, terutama pada keterbatasan konteks ekosistem, cakupan parameter yang belum komprehensif, serta ketergantungan pada kerangka *Water Quality Index* (WQI) berbasis standar internasional yang belum sepenuhnya selaras dengan regulasi nasional. Oleh karena itu, penelitian ini menempati posisi strategis dengan mengintegrasikan kerangka regulasi nasional, yaitu Keputusan Menteri Lingkungan Hidup Nomor 115 Tahun 2003 dan Keputusan Menteri Lingkungan

Hidup Nomor 51 Tahun 2004, dengan pendekatan *smart ensemble learning* berbasis XGBoost untuk klasifikasi multi-kelas. Integrasi ini memungkinkan analisis multi-parameter yang mencakup aspek fisika, kimia, biologi, dan logam terlarut guna membangun model prediktif yang mampu mengklasifikasikan status mutu air laut ke dalam empat kategori resmi, yaitu Baik, Cemar Ringan, Cemar Sedang, dan Cemar Berat, secara akurat, adaptif, dan aplikatif dalam mendukung upaya konservasi biota laut di Teluk Jakarta.

1.2 Landasan Teori

Secara teoretis, landasan teori merupakan struktur konseptual yang berperan penting dalam mengidentifikasi, mengorganisasi, dan menyusun konsep-konsep penelitian secara sistematis, sehingga memberikan kerangka berpikir yang jelas bagi peneliti dalam memahami fenomena yang dikaji. Landasan teori berfungsi sebagai penghubung antara teori yang telah ada dengan metodologi penelitian yang digunakan serta hasil penelitian yang diperoleh, sehingga setiap tahapan penelitian memiliki pijakan ilmiah yang kuat dan terarah. Melalui landasan teori, konsep dan variabel penelitian dapat didefinisikan secara jelas, pendekatan dan metode penelitian dapat ditentukan secara tepat, serta data dan hasil penelitian dapat diinterpretasikan secara kritis dalam konteks keilmuan yang relevan. Dengan demikian, landasan teori tidak hanya memperkuat validitas ilmiah penelitian, tetapi juga memastikan bahwa analisis dan pembahasan dilakukan secara logis, konsisten, dan dapat dipertanggungjawabkan secara akademik (Luft et al., 2022).

1.2.1 Kerangka Regulasi Pengelolaan Kualitas Air Laut di Indonesia

Kerangka regulasi pengelolaan kualitas air laut di Indonesia didasarkan pada Keputusan Menteri Negara Lingkungan Hidup Nomor 51 Tahun 2004 tentang Baku Mutu Air Laut, yang menjadi acuan utama dalam menjaga kelestarian fungsi ekologis laut sekaligus mendukung pemanfaatan ekonominya. Regulasi ini bertujuan untuk mengendalikan kegiatan manusia yang berpotensi mencemari atau merusak lingkungan laut, sehingga kualitas perairan tetap berada dalam batas yang aman bagi kehidupan biota laut serta aktivitas manusia seperti pariwisata dan pelabuhan (Suharto et al., 2019).

Dalam Pasal 1 peraturan ini dijelaskan beberapa istilah penting, antara lain, laut yang diartikan sebagai ruang wilayah lautan dengan batas dan sistem yang ditentukan berdasarkan aspek fungsional, baku mutu air laut sebagai ukuran batas atau kadar makhluk hidup, zat, energi, atau unsur pencemar yang ditenggang keberadaannya di dalam air laut, serta biota laut yang mencakup seluruh organisme hidup di perairan laut.

Lebih lanjut, Pasal 2 menetapkan bahwa baku mutu air laut diberlakukan berdasarkan tiga kategori peruntukan utama, yaitu Perairan Pelabuhan, Wisata Bahari, dan Biota Laut, di mana masing-masing memiliki standar kualitas air yang spesifik. Pemerintah daerah, baik di tingkat provinsi maupun kabupaten/kota, diwajibkan untuk melakukan pemantauan kualitas air laut secara berkala minimal dua kali dalam setahun serta menindaklanjuti hasil pemantauan tersebut dengan langkah-langkah pengendalian pencemaran. Ketentuan ini menegaskan tanggung jawab pemerintah daerah dalam menjaga keseimbangan ekosistem laut di wilayah administratifnya.

Di antara ketiga kategori tersebut, peruntukan Biota Laut memiliki posisi paling fundamental karena menjadi dasar baku mutu untuk sebagian besar wilayah perairan Indonesia. Pasal 7 Keputusan Menteri LH No. 51 Tahun 2004 menegaskan bahwa kawasan laut di luar peruntukan Pelabuhan dan Wisata Bahari harus mengikuti baku mutu untuk Biota Laut, yang dirancang agar kondisi perairan mampu mendukung kesehatan, pertumbuhan, dan reproduksi organisme laut. Parameter dan nilai ambang batas untuk kategori ini tercantum dalam Lampiran III keputusan tersebut, mencakup aspek fisika, kimia, logam berat, dan mikrobiologi. Tabel 2.1 menampilkan standar Baku Mutu Air Laut untuk biota laut sebagaimana tercantum dalam regulasi Keputusan Menteri Negara Lingkungan Hidup Nomor 51 Tahun 2004, Lampiran III (Patimah et al., 2022).

Tabel 2. 1 Baku Mutu Air Laut untuk Biota Laut

No.	Parameter	Satuan	Baku Mutu
Fisika			
1	Kecerahan	m	coral: >5; mangrove: -; lamun: >3; alami
2	Kebauan	-	alami
3	Kekeruhan	NTU	<5
4	Padatan tersuspensi total	mg/L	coral: 20; mangrove: 80; lamun: 20

No.	Parameter	Satuan	Baku Mutu
5	Sampah	-	nihil
6	Suhu	°C	alami ±3; coral: 28–30; mangrove: 28–32; lamun: 28–30
7	Lapisan minyak	-	nihil
Kimia			
1	pH	-	7–8,5
2	Salinitas	‰	alami ±3; coral: 33–34; mangrove: ≤34; lamun: 33–34
3	Oksigen terlarut (DO)	mg/L	>5
4	BOD5	mg/L	20
5	Ammonia total (NH_3–N)	mg/L	0,3
6	Fosfat (PO_4–P)	mg/L	0,015
7	Nitrat (NO_3–N)	mg/L	0,008
8	Sianida (CN^-)	mg/L	0,5
9	Sulfida (H_2S)	mg/L	0,01
10	PAH (Poliaromatik hidrokarbon)	mg/L	0,003
11	Senyawa Fenol total	mg/L	0,002
12	PCB total (poliklor bifenil)	µg/L	0,01
13	Surfaktan (deterjen)	mg/L MBAS	1
14	Minyak & lemak	mg/L	1
15	Pestisida	µg/L	0,01
16	TBT (tributil tin)	µg/L	0,01
Logam Berat			
1	Raksa (Hg)	mg/L	0,001
2	Kromium heksavalen (Cr(VI))	mg/L	0,005
3	Arsen (As)	mg/L	0,012
4	Kadmium (Cd)	mg/L	0,001
5	Tembaga (Cu)	mg/L	0,008
6	Timbal (Pb)	mg/L	0,008
7	Seng (Zn)	mg/L	0,05
8	Nikel (Ni)	mg/L	0,05
Biologi			
1	Coliform (total)	MPN/100 ml	1000
2	Patogen	sel/100 ml	nihil
3	Plankton	sel/100 ml	tidak bloom

No.	Parameter	Satuan	Baku Mutu
Radio Nuklida 1	Komposisi yang tidak diketahui	Bq/L	4

Tabel 2.1 tersebut menjadi referensi utama dalam penelitian dan evaluasi kualitas air laut di Indonesia. Setiap hasil pengukuran di lapangan dibandingkan dengan nilai ambang batas dalam baku mutu ini untuk menentukan tingkat kesesuaian perairan terhadap fungsi ekologisnya. Dengan demikian, regulasi ini tidak hanya memberikan dasar hukum bagi perlindungan ekosistem laut, tetapi juga menyediakan instrumen ilmiah yang konkret untuk pemantauan, evaluasi, dan pengambilan kebijakan lingkungan berbasis data.

1.2.2 *Pollution Index* (PI)

Selain menetapkan baku mutu air laut, kerangka regulasi nasional juga menyediakan pedoman teknis untuk menentukan status mutu air secara komprehensif berdasarkan hasil pemantauan berbagai parameter kualitas air. Pedoman tersebut diatur dalam Keputusan Menteri Negara Lingkungan Hidup Nomor 115 Tahun 2003 tentang Pedoman Penentuan Status Mutu Air, yang menjadi dasar hukum bagi instansi pemerintah maupun peneliti dalam mengevaluasi kondisi badan air di Indonesia (Keputusan Menteri Negara Lingkungan Hidup No 115, 2003).

Peraturan ini memperkenalkan Metode Indeks Pencemaran atau *Pollution Index* (PI) sebagai pendekatan kuantitatif untuk menilai tingkat pencemaran air secara menyeluruh terhadap baku mutu yang berlaku. Metode Indeks Pencemaran dirancang untuk menghasilkan satu nilai indeks tunggal yang merepresentasikan tingkat pencemaran relatif suatu badan air terhadap peruntukan tertentu. Prinsip dasar metode ini adalah membandingkan konsentrasi terukur suatu parameter kualitas air (C_i) dengan nilai baku mutu yang ditetapkan (L_{ij}), kemudian menggabungkan hasil perbandingan tersebut melalui formula matematis. Konsep dasar metode ini pertama kali dikembangkan oleh Sumitomo dan Nemerow (1970) dan selanjutnya diadaptasi dalam konteks pengelolaan lingkungan di Indonesia melalui Kepmen LH No. 115 Tahun 2003.

Secara metodologis, perhitungan Indeks Pencemaran dilakukan melalui beberapa tahapan sistematis. Tahap awal adalah menghitung rasio antara konsentrasi parameter kualitas air hasil pengukuran (C_i) dan nilai baku mutu untuk peruntukan tertentu (L_{ij}), yang dinyatakan dalam Persamaan 2.1.

$$Rasio = \frac{C_i}{L_{ij}} \tag{2.1}$$

Pada Persamaan 2.1, nilai rasio menggambarkan tingkat pencemaran relatif masing-masing parameter. Nilai rasio sama dengan satu menunjukkan bahwa konsentrasi parameter berada tepat pada ambang baku mutu. Nilai rasio lebih besar dari satu menunjukkan bahwa konsentrasi parameter telah melampaui baku mutu dan berpotensi mencemari, sedangkan nilai rasio kurang dari satu menunjukkan bahwa konsentrasi parameter masih berada dalam batas aman.

Namun demikian, tidak semua parameter kualitas air memiliki hubungan linier terhadap tingkat pencemaran. Beberapa parameter, seperti oksigen terlarut atau *Dissolved Oxygen* (DO), memiliki hubungan berbanding terbalik dengan kualitas perairan, di mana penurunan nilai DO justru mencerminkan kondisi perairan yang semakin tercemar. Untuk menjaga konsistensi interpretasi nilai rasio, parameter-parameter tersebut memerlukan penyesuaian rasio. Penyesuaian rasio untuk parameter DO dilakukan menggunakan Persamaan 2.2.

$$\left(\frac{C_i}{L_{ij}}\right)_{\text{baru}} = \frac{C_{im} - C_i}{C_{im} - L_{ij}} \tag{2.2}$$

Pada Persamaan 2.2, nilai $\left(\frac{C_i}{L_{ij}}\right)_{\text{baru}}$ merepresentasikan tingkat penyimpangan konsentrasi parameter hasil pengukuran terhadap kondisi idealnya. C_{im} merupakan nilai jenuh atau kondisi ideal parameter yang bersangkutan. Khusus untuk parameter DO, nilai C_{im} dihitung menggunakan Persamaan Weiss (1970) yang dikembangkan berdasarkan data eksperimen kelarutan oksigen dalam air laut. Persamaan ini dinyatakan dalam bentuk logaritma alami untuk merepresentasikan hubungan eksponensial antara suhu,

salinitas, dan kelarutan gas (Weiss, 1970). Bentuk umum persamaan Weiss dituliskan pada Persamaan 2.3.

$$\ln C_{im} = A1 + A2(\frac{100}{T}) + A3\ln(\frac{T}{100}) + A4(\frac{T}{100}) + S(B1 + B2(\frac{T}{100}) + B3(\frac{T}{100})^2) \quad (2.3)$$

Persamaan 2.3 merupakan persamaan yang digunakan untuk menghitung konsentrasi maksimum oksigen terlarut atau konsentrasi jenuh dalam air laut. Besaran C_{im} menyatakan kelarutan oksigen pada kondisi kesetimbangan dengan atmosfer, sedangkan bentuk $ln\, C_{im}$ digunakan untuk memodelkan hubungan nonlinier antara kelarutan oksigen dengan suhu dan salinitas.

Dalam persamaan tersebut, T adalah suhu air dalam satuan Kelvin (K) dan S adalah salinitas dalam satuan per mil (‰). Koefisien A_1 hingga A_4 merepresentasikan pengaruh suhu terhadap kelarutan oksigen, sementara koefisien B_1 hingga B_3 menggambarkan pengaruh salinitas yang menurunkan kelarutan oksigen melalui efek salting-out. Nilai konstanta yang digunakan mengacu pada Weiss (1970) sebagaimana ditunjukkan pada Tabel 2.2.

Tabel 2. 2 Nilai konstanta (Weiss, 1970)

Konstanta	Nilai
A1	-173.4292
A2	249.6339
A3	143.3483
A4	-21.8492
B1	-0.033096
B2	0.014259
B3	-0.0017000

Sebelum Persamaan 2.3 digunakan, suhu air yang diukur dalam derajat Celsius (*t*) harus dikonversi ke satuan Kelvin karena perhitungan kelarutan gas didasarkan pada skala suhu absolut. Konversi suhu dilakukan menggunakan Persamaan 2.4.

$$T = t\ (°\mathrm{C}) + 273.15 \quad (2.4)$$

Pada Persamaan 2.4, nilai ln C_{im} yang diperoleh dari Persamaan 2.3 selanjutnya ditransformasikan ke bentuk konsentrasi aktual melalui fungsi eksponensial, sebagaimana ditunjukkan pada Persamaan 2.5.

$$C_{im} = e^{\ln C_{im}} \tag{2.5}$$

Pada Persamaan 2.5, nilai C_{im} dinyatakan dalam satuan mL O_2/l, yang merepresentasikan volume oksigen sebagai gas nyata pada kondisi tekanan satu atmosfer. Agar sesuai dengan satuan baku mutu kualitas air dan hasil pengukuran lapangan, nilai tersebut dikonversi ke satuan mg/L menggunakan faktor konversi 1,42905 mg/mL, sebagaimana dijelaskan oleh Benson dan Krause 1980 (USGS, 2011). Proses konversi ini dinyatakan dalam Persamaan 2.6.

$$DO^* = C_{im} \times 1.42905 \tag{2.6}$$

Persamaan 2.6 digunakan untuk mengonversi nilai konsentrasi oksigen jenuh C_{im} yang diperoleh dari Persamaan 2.3 ke dalam satuan mg/L. Faktor konversi 1,42905 merupakan konstanta yang menyatakan massa oksigen (dalam miligram) per satuan volume gas pada kondisi standar. Hasil perhitungan DO^* merepresentasikan konsentrasi oksigen terlarut jenuh dalam air laut pada kondisi suhu dan salinitas tertentu. Nilai DO^* ini selanjutnya digunakan sebagai nilai acuan dalam perhitungan rasio oksigen terlarut (DO) pada metode Indeks Pencemaran, untuk menilai tingkat deviasi kondisi aktual terhadap kondisi jenuh.

Untuk parameter kualitas air yang memiliki baku mutu dalam bentuk rentang nilai $((L_{\text{ij minimum}}) - (L_{\text{ij maksimum}}))$, digunakan nilai rata-rata baku mutu $(L_{\text{ij rata-rata}})$ sebagai titik referensi. Penyesuaian rasio dilakukan berdasarkan posisi nilai pengukuran terhadap nilai rata-rata tersebut. Jika nilai pengukuran lebih kecil atau sama dengan nilai rata-rata baku mutu, penyesuaian rasio dilakukan menggunakan Persamaan 2.7.

$$\left(\frac{C_i}{L_{ij}}\right)_{\text{baru}} = \frac{[C_i - (L_{\text{ij rata-rata}})]}{\{(L_{\text{ij minimum}}) - (L_{\text{ij rata-rata}})\}} \tag{2.7}$$

Sebaliknya, jika nilai pengukuran lebih besar dari nilai rata-rata baku mutu, penyesuaian rasio dilakukan menggunakan Persamaan 2.8.

$$\left(\frac{C_i}{L_{ij}}\right)_{baru} = \frac{[C_i - (L_{ij\ rata\text{-}rata})]}{\{(L_{ij\ maksimum}) - (L_{ij\ rata\text{-}rata})\}} \tag{2.8}$$

Untuk mencegah dominasi berlebihan dari parameter dengan nilai rasio yang sangat besar, diterapkan penyesuaian logaritmik apabila nilai rasio lebih besar dari satu $(C_i/L_{ij} \geq 1{,}0)$. Penyesuaian ini dinyatakan dalam Persamaan 2.9.

$$\left(\frac{C_i}{L_{ij}}\right)_{baru} = 1 + P \cdot \log\left(\frac{C_i}{L_{ij}}\right) \text{hasil pengukuran} \tag{2.9}$$

Pada Persamaan 2.9, $P = 5$ adalah konstanta yang digunakan untuk mengatur sensitivitas perubahan rasio, yang berfungsi untuk menghindari perubahan rasio yang terlalu besar. Jika $C_i/L_{ij} \leq 1{,}0$, maka nilai rasio digunakan tanpa penyesuaian tambahan.

Setelah seluruh rasio parameter dihitung, diperoleh nilai rasio maksimum $(C_i/L_{ij})_M$ dan nilai rata-rata rasio $(C_i/L_{ij})_R$. Kedua nilai tersebut digunakan untuk menghitung Indeks Pencemaran pada lokasi ke-j menggunakan Persamaan 2.10.

$$PI_j = \sqrt{\frac{(Ci/L_{ij})_M^2 + (Ci/L_{ij})_R^2}{2}} \tag{2.10}$$

Persamaan 2.10 menggabungkan kondisi parameter terburuk dan kondisi keseluruhan secara seimbang, memberikan gambaran komprehensif mengenai tingkat pencemaran pada lokasi tersebut. Berdasarkan nilai PI yang diperoleh, status mutu air dapat diklasifikasikan dalam empat kategori, seperti yang tercantum dalam Tabel 2.3.

Tabel 2. 3 Klasifikasi Status Mutu Air Berdasarkan Nilai PI

Rentang Nilai IP	Status Mutu Air	Keterangan
$0 \leq PI \leq 1,0$	Memenuhi Baku Mutu	Kondisi Baik
$1,0 < PI \leq 5,0$	Cemar Ringan	Tercemar Ringan
$5,0 < PI \leq 10,0$	Cemar Sedang	Tercemar Sedang
$PI > 10,0$	Cemar Berat	Tercemar Berat

Tabel 2.3 menunjukkan rentang nilai PI dan status mutu air yang sesuai dengan setiap kategori, dari kondisi baik hingga kondisi tercemar berat.

1.2.3 *Machine learning*

Machine learning merupakan cabang dari kecerdasan buatan atau *artificial intelligence* yang mempelajari pengembangan algoritma dan model komputasi yang memungkinkan sistem komputer untuk belajar dari data dan melakukan prediksi atau pengambilan keputusan tanpa diprogram secara eksplisit. Pembelajaran dilakukan dengan mengidentifikasi pola, hubungan, dan struktur dalam data sehingga model dapat melakukan generalisasi terhadap data baru.

Berdasarkan jenis pembelajarannya, *machine learning* dibedakan menjadi *supervised learning*, *unsupervised learning*, dan *reinforcement learning*. Penelitian ini termasuk dalam kategori *supervised learning*, yaitu pembelajaran menggunakan data berlabel, di mana model dilatih untuk memetakan hubungan antara variabel independen atau fitur dan variabel dependen atau target. Salah satu algoritma *supervised learning* yang banyak digunakan karena sifatnya yang sederhana dan interpretatif adalah *Decision Tree* (Alshboul et al., 2022).

1.2.4 *Ensemble Learning*

Ensemble learning merupakan pendekatan dalam *machine learning* yang bertujuan untuk meningkatkan kinerja prediksi dengan mengombinasikan beberapa model dasar atau *base learners* menjadi satu model kolektif. Prinsip utama *ensemble learning* didasarkan pada asumsi bahwa penggabungan sejumlah model yang relatif lemah atau *weak learners* dapat menghasilkan model yang lebih kuat atau *strong learner*

dibandingkan model tunggal. Pendekatan ini efektif dalam mengurangi kesalahan prediksi yang disebabkan oleh bias, *varians*, maupun *noise* pada data (Alshboul et al., 2022).

Dalam praktiknya, *Decision Tree* sering digunakan sebagai *base learner* dalam *ensemble learning* karena strukturnya yang fleksibel, kemampuannya memodelkan hubungan non-linear, serta efisiensinya dalam proses pelatihan. Meskipun demikian, *Decision Tree* tunggal memiliki kelemahan utama berupa *varians* yang tinggi dan kecenderungan mengalami *overfitting*. *Ensemble learning* dikembangkan untuk mengatasi keterbatasan tersebut dengan menggabungkan banyak pohon keputusan melalui mekanisme tertentu sehingga prediksi menjadi lebih stabil dan akurat (Alshboul et al., 2022).

Salah satu pendekatan utama dalam *ensemble learning* adalah *bagging* atau *bootstrap aggregating*, yang membangun beberapa model secara paralel dengan sampel data yang diambil secara acak dengan pengembalian atau *bootstrap sampling*. Pada setiap model, subset fitur juga dipilih secara acak untuk memperluas variasi pohon dan menurunkan korelasi antarmodel. *Random Forest* adalah contoh utama metode *bagging* berbasis *Decision Tree. Random Forest* membangun banyak pohon keputusan dari sampel *bootstrap* dan menggabungkan prediksi pohon-pohon tersebut melalui *majority voting* (klasifikasi) atau rata-rata (regresi), sehingga *varians* model berkurang dibandingkan *Decision Tree* tunggal dan model menjadi lebih *robust* terhadap *noise* dalam data (GeeksforGeeks, 2025).

Meskipun *Random Forest* lebih stabil dibandingkan *Decision Tree*, metode ini tetap memiliki keterbatasan dalam menangkap pola kompleks secara adaptif karena setiap pohon dibangun secara independen tanpa memanfaatkan informasi kesalahan prediksi dari model sebelumnya. Selain itu, *Random Forest* cenderung membutuhkan jumlah pohon yang besar untuk mencapai performa optimal, yang dapat mempengaruhi efisiensi komputasi pada dataset berdimensi tinggi (GeeksforGeeks, 2025).

Pendekatan boosting dikembangkan sebagai strategi *ensemble* yang bersifat sekuensial, di mana setiap model baru dibangun untuk memperbaiki kesalahan prediksi yang dibuat oleh model sebelumnya. Dalam *boosting*, model-model dasar seperti *Decision Tree* dangkal ditambahkan satu per satu, dan setiap model baru berfokus pada

sampel yang sulit diprediksi oleh model sebelumnya. Salah satu representasi dari metode ini adalah *Gradient Boosting*, yang membangun model aditif dengan meminimalkan fungsi kerugian atau *loss function* menggunakan teknik gradien. model *Gradient Boosting* dapat dituliskan menggunakan Persamaan 2.11.

$$\hat{y}_i = \sum_{t=0}^{T} f_t\,(x_i) \tag{2.11}$$

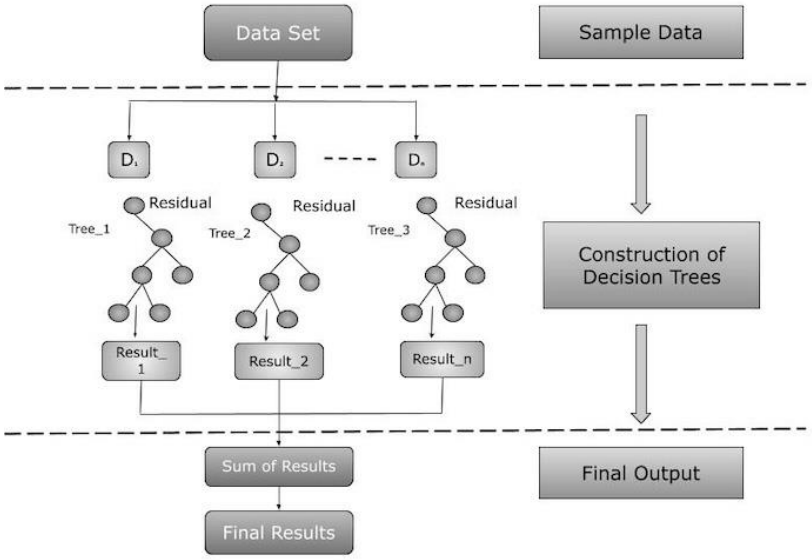
Pada Persamaan 2.11, $\hat{y}_i$ menyatakan nilai prediksi untuk observasi ke-i, yang diperoleh dari hasil akumulasi kontribusi seluruh pohon keputusan dalam model. Fungsi $f_t(x_i)$ merepresentasikan keluaran atau prediksi dari pohon keputusan ke-t terhadap data masukan x_i, sedangkan T menunjukkan jumlah total pohon keputusan yang dibangun dalam proses pembelajaran. *Gradient Boosting* mampu memodelkan hubungan non-linear dan interaksi kompleks antarvariabel secara efektif, tetapi cenderung rentan terhadap *overfitting* serta memiliki kompleksitas komputasi yang lebih tinggi jika parameter tidak diatur dengan tepat (Nopebrian et al., 2025).

Extreme Gradient Boosting (XGBoost) merupakan pengembangan lanjutan dari *Gradient Boosting* yang dirancang untuk meningkatkan efisiensi komputasi, stabilitas model, dan akurasi prediksi. XGBoost tidak hanya menerapkan prinsip boosting, tetapi juga menyertakan regularisasi eksplisit terhadap kompleksitas model melalui penalti atas jumlah node daun dan bobot pada setiap daun, serta menggunakan optimisasi fungsi objektif berbasis turunan orde kedua untuk meningkatkan kecepatan konvergensi. Selain itu, XGBoost mendukung komputasi paralel dan mekanisme penanganan nilai hilang atau *missing values*, sehingga lebih efisien dibandingkan implementasi *Gradient Boosting* konvensional (Nopebrian et al., 2025).

1.2.5 *Extreme Gradient Boosting* (XGBoost)

Extreme Gradient Boosting (XGBoost) merupakan pengembangan dari algoritma *gradient boosting* yang dirancang untuk memberikan performa prediksi tinggi dengan efisiensi komputasi yang optimal. XGBoost bekerja dengan menggabungkan sejumlah model *weak learner* berbasis *Decision Tree* secara bertahap atau *sequential*, di mana setiap model baru dibangun untuk memperbaiki kesalahan dari model sebelumnya

melalui optimisasi fungsi kehilangan atau *loss function* dengan metode *gradient descent* sehingga mencegah *overfitting* (Chen & Guestrin, 2016b)



Gambar 2. 1 *Architecture* XGBoost (tutorialspoint, 2025)

XGBoost memiliki beberapa keunggulan utama yang menjadikannya sangat efektif dalam pemodelan data kompleks. Algoritma ini menggunakan optimasi berbasis gradien dan Hessian sehingga proses pembelajaran model menjadi lebih akurat dan terarah dalam meminimalkan kesalahan prediksi. Selain itu, XGBoost dilengkapi dengan mekanisme regularisasi yang berfungsi untuk mengontrol kompleksitas model dan mengurangi risiko *overfitting*. XGBoost juga mampu menangani data berdimensi tinggi serta kondisi multikolinearitas antar variabel dengan baik, sehingga tetap stabil meskipun jumlah fitur cukup banyak. Dari sisi komputasi, algoritma ini dirancang efisien dan teroptimasi, sehingga mampu memproses data dalam skala besar dengan waktu pelatihan yang relatif cepat.

1.2.5.1 Inisialisasi Prediksi Awal pada XGBoost

Proses pelatihan XGBoost diawali dengan penentuan prediksi awal yang dinotasikan sebagai $\hat{y}^{(0)}$. Pada iterasi ke-0, model belum membangun pohon keputusan dan belum memanfaatkan fitur masukan x_i. Oleh karena itu, XGBoost hanya mencari satu

nilai konstan yang meminimalkan fungsi kerugian secara keseluruhan (Chen & Guestrin, 2016). Secara matematis, prediksi awal ditentukan dengan menyelesaikan permasalahan optimasi yang dirumuskan menggunakan Persamaan 2.12.

$$\hat{y}^{(0)} = \arg\min_{c} \sum_{i=1}^{n} L(y_i, c) \tag{2.12}$$

Berdasarkan Persamaan 2.12, c merupakan suatu konstanta tunggal yang digunakan sebagai prediksi untuk seluruh data, sedangkan $L(y_i, c)$ menyatakan fungsi kerugian antara nilai aktual y_i dan prediksi konstan tersebut. Nilai konstanta c yang meminimalkan fungsi kerugian inilah yang digunakan sebagai prediksi awal $\hat{y}^{(0)}$.

Fungsi kerugian yang umum digunakan adalah *Mean Squared Error* (MSE), yang dirumuskan menggunakan Persamaan 2.13.

$$L(y_i, c) = \frac{1}{2}(y_i - c)^2 \tag{2.13}$$

Dengan mensubstitusikan Persamaan 2.13 ke dalam Persamaan 2.12, fungsi objektif awal menjadi Persamaan 2.14.

$$\sum_{i=1}^{n} \frac{1}{2}(y_i - c)^2 \tag{2.14}$$

Untuk memperoleh nilai c yang optimal, dilakukan diferensiasi terhadap c, sehingga diperoleh Persamaan 2.15.

$$\frac{\partial}{\partial c}\sum_{i=1}^{n} \frac{1}{2}(y_i - c)^2 = \sum_{i=1}^{n}(c - y_i) \tag{2.15}$$

Dengan mensyaratkan turunan pertama sama dengan nol, diperoleh Persamaan 2.16.

$$\sum_{i=1}^{n}(c - y_i) = 0 \tag{2.16}$$

yang selanjutnya menghasilkan Persamaan 2.17.

$$c = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2.17}$$

Pada Persamaan 2.17, prediksi awal XGBoost diinisialisasi sebagai nilai rata-rata dari seluruh data target. Oleh karena itu, prediksi awal dapat dinyatakan menggunakan Persamaan 2.18.

$$\hat{y}^{(0)} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2.18}$$

1.2.5.2 Model Prediksi Aditif pada XGBoost

Setelah prediksi awal ditentukan, XGBoost membangun model prediksi secara bertahap dengan menambahkan pohon keputusan pada setiap iterasi. Prediksi pada iterasi ke-t dirumuskan menggunakan Persamaan 2.19.

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i) \tag{2.19}$$

Pada Persamaan 2.19, $\hat{y}_i^{(t)}$ menyatakan prediksi untuk observasi ke-i pada iterasi ke-t, η adalah *learning rate* yang mengontrol besarnya kontribusi setiap pohon keputusan, dan $f_t(x_i)$ merupakan fungsi pohon keputusan yang ditambahkan pada iterasi ke-t. Jika seluruh pembaruan dijumlahkan, maka prediksi akhir setelah t iterasi dapat dirumuskan menggunakan Persamaan 2.20.

$$\hat{y}_i^{(t)} = \hat{y}^{(0)} + \eta \sum_{k=1}^{t} f_k\,(x_i) \tag{2.20}$$

1.2.5.3 Fungsi Objektif XGBoost

Tujuan utama XGBoost adalah meminimalkan fungsi objektif yang terdiri atas fungsi kerugian dan fungsi regularisasi (Chen & Guestrin, 2016). Fungsi objektif pada iterasi ke-t dirumuskan menggunakan Persamaan 2.21.

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{2.21}$$

Pada Persamaan 2.21, $\mathcal{L}^{(t)}$ merupakan fungsi objektif total yang diminimalkan oleh algoritma. Simbol n menunjukkan jumlah observasi, y_i adalah nilai aktual dari observasi ke-i, sedangkan $\hat{y}_i$ merupakan nilai prediksi model. Fungsi $l(y_i, \hat{y}_i)$ menggambarkan kerugian antara nilai aktual dan prediksi. Sementara itu, $\Omega(f_t)$ adalah fungsi regularisasi yang diterapkan pada pohon ke-t, berfungsi membatasi kompleksitas pohon berdasarkan jumlah daun dan besarnya bobot pada masing-masing daun. Dengan formulasi ini, XGBoost tidak hanya berfokus pada akurasi prediksi, tetapi juga memastikan bahwa model yang dihasilkan tidak terlalu kompleks dan lebih mampu melakukan generalisasi terhadap data baru.

Fungsi regularisasi pada XGBoost didefinisikan untuk mengontrol kompleksitas model dengan memberikan penalti terhadap struktur pohon dan besarnya bobot pada setiap daun (Chen & Guestrin, 2016). Secara matematis, fungsi regularisasi dinyatakan pada Persamaan 2.22.

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{2.22}$$

Pada Persamaan 2.22, T menyatakan jumlah *leaf* pada pohon, w_j adalah bobot *leaf* ke-j, serta γ dan λ merupakan parameter regularisasi.

1.2.5.4 Pendekatan Taylor Expansion Orde Kedua

Karena fungsi pohon keputusan bersifat diskrit, fungsi objektif tidak dapat diminimalkan secara langsung. Oleh karena itu, XGBoost menggunakan aproksimasi

Taylor Expansion orde kedua terhadap fungsi kerugian di sekitar prediksi sebelumnya $\hat{y}_i^{(t-1)}$. Aproksimasi ini dinyatakan menggunakan Persamaan 2.23 (Chen & Guestrin, 2016).

$$l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) \approx l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \quad (2.23)$$

dengan:

$$g_i = \frac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}_i} \quad (2.24)$$

$$h_i = \frac{\partial^2 l(y_i, \hat{y}_i)}{\partial \hat{y}_i^2} \quad (2.25)$$

Karena suku pertama bersifat konstan, maka dapat diabaikan dalam proses optimasi. *Mean Squared Error* (MSE), nilai gradien dan Hessian dapat disederhanakan sebagaimana ditunjukkan pada Persamaan 2.26.

$$g_i = \hat{y}_i - y_i, \;\; h_i = 1 \quad (2.26)$$

1.2.5.5 Fungsi Objektif yang Disederhanakan

Untuk mengoptimalkan fungsi objektif secara efisien, XGBoost menggunakan aproksimasi Taylor Expansion orde kedua terhadap fungsi kerugian pada setiap iterasi. Pendekatan ini bertujuan untuk memperkirakan perubahan fungsi kerugian akibat penambahan sebuah pohon keputusan baru. Dengan memanfaatkan informasi turunan pertama yaitu gradien dan turunan kedua yaitu Hessian, XGBoost mampu memperoleh estimasi arah dan besaran pembaruan model secara lebih akurat dibandingkan metode *boosting* konvensional yang hanya menggunakan gradien orde pertama (Chen & Guestrin, 2016). Dengan mengabaikan konstanta, fungsi objektif hasil aproksimasi Taylor dapat dituliskan menggunakan Persamaan 2.27.

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t) \quad (2.27)$$

Persamaan 2.27 digunakan sebagai dasar dalam pembangunan pohon keputusan pada setiap iterasi.

1.2.5.6 Perhitungan Bobot Leaf dan Gain

Jika suatu data berada pada *leaf* yang sama, maka fungsi pohon dapat dinyatakan menggunakan Persamaan 2.28.

$$f_t(x_i) = w_j \text{ jika } x_i \in I_j \quad (2.28)$$

Fungsi objektif pada *leaf* ke-j menjadi Persamaan 2.29.

$$\mathcal{L}_j = G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2 \quad (2.29)$$

dengan:

$$G_j = \sum_{i \in I_j} g_i \, , H_j = \sum_{i \in I_j} h_i$$

Bobot *leaf* optimal diperoleh menggunakan Persamaan 2.30.

$$w_j^* = -\frac{G_j}{H_j + \lambda} \quad (2.30)$$

Untuk menentukan *split* terbaik, digunakan nilai *gain* yang dirumuskan menggunakan Persamaan 2.31.

$$\text{Gain} = \frac{1}{2}\left[\frac{\left(\sum g_L\right)^2}{\sum h_L + \lambda} + \frac{\left(\sum g_R\right)^2}{\sum h_R + \lambda} - \frac{\left(\sum g\right)^2}{\sum h + \lambda} \right] - \gamma \quad (2.31)$$

Dalam penelitian ini, XGBoost digunakan untuk memprediksi nilai Indeks Pencemaran berdasarkan parameter kualitas air laut. Karakteristik data yang bersifat multivariat dan nonlinier menjadikan XGBoost sebagai metode yang sesuai karena kemampuannya dalam menangkap hubungan kompleks antar parameter serta ketahanannya terhadap noise dan multikolinearitas (Arif Ali et al., 2023).

2.2.6 *Koefisien Determinasi*

Koefisien determinasi atau R^2 merupakan ukuran statistik yang digunakan untuk mengetahui seberapa besar variasi nilai observasi yang dapat dijelaskan oleh model prediksi (Jumiati & Yuliansyah, 2020). Dalam penelitian ini, nilai R^2 digunakan untuk mengevaluasi kinerja model XGBoost dalam memprediksi kualitas air berdasarkan parameter fisika, kimia, biologi dan logam terlarut. Secara matematis, nilai R^2 dihitung menggunakan Persamaan 2.32.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{2.32}$$

Berdasarkan Persamaan 2.32, SS_{res} adalah jumlah kuadrat residual dihitung dengan $\sum(y_i - \hat{y}_i)^2$, yaitu jumlah kuadrat selisih antara nilai aktual y_i dan nilai prediksi $\hat{y}_i$ dari model. Nilai ini menunjukkan seberapa besar kesalahan model dalam memprediksi data.

Sementara itu, SS_{tot} adalah jumlah kuadrat total dihitung dengan $\sum(y_i - \bar{y}_a)^2$, yaitu jumlah kuadrat selisih antara nilai aktual y_i dengan rata-rata nilai aktual $\bar{y}_a$. Nilai ini mencerminkan total variasi yang ada dalam data.

Dengan rumus ini, R^2 mengukur proporsi variasi total yang dapat dijelaskan oleh model. Jika $R^2 = 1$, berarti model memprediksi data dengan sempurna tanpa kesalahan, namun jika $R^2 = 0$, model tidak lebih baik daripada rata-rata sederhana dalam memprediksi nilai aktual (Chicco et al., 2021).

2.2.7 Evaluasi Model

Evaluasi kinerja model prediksi berbasis *Extreme Gradient Boosting* (XGBoost) dilakukan menggunakan kombinasi metrik kesalahan numerik dan metrik klasifikasi. Metrik *Mean Squared Error* (MSE), *Root Mean Squared Error* (RMSE), dan *Mean*

Absolute Error (MAE) digunakan untuk menilai tingkat kesalahan prediksi nilai Indeks Pencemaran, sedangkan *confusion matrix* digunakan untuk mengevaluasi kemampuan model dalam mengklasifikasikan status mutu air laut ke dalam kategori yang sesuai. Penggunaan kombinasi metrik tersebut bertujuan untuk memperoleh gambaran kinerja model secara menyeluruh.

2.2.7.1 *Mean Squared Error* (MSE)

Mean Squared Error (MSE) merupakan metrik evaluasi yang mengukur rata-rata kuadrat selisih antara nilai prediksi dan nilai observasi. Dengan menggunakan kuadrat selisih, kesalahan dengan deviasi yang lebih besar akan memberikan kontribusi yang lebih besar terhadap nilai MSE, sehingga metrik ini bersifat sensitif terhadap keberadaan *outlier* dalam data (Kusuma et al., 2023). Secara matematis, perumusan MSE ditunjukkan pada Persamaan 2.33.

$$\text{MSE} = \frac{1}{m} \sum_{i=1}^{m} (Y_i - \hat{Y}_i)^2 \tag{2.33}$$

Pada Persamaan 2.33, m adalah jumlah data observasi. Nilai MSE yang semakin mendekati nol menunjukkan bahwa kesalahan prediksi model semakin kecil, sehingga performa model dapat dikatakan semakin baik (Chicco et al., 2021).

2.2.7.2 *Root Mean Squared Error* (RMSE)

Untuk memberikan interpretasi kesalahan dalam satuan yang sama dengan variabel yang diprediksi, digunakan *Root Mean Squared Error* (RMSE). RMSE merupakan akar kuadrat dari MSE dan sering digunakan untuk membandingkan performa beberapa model prediksi. Secara matematis, perhitungan RMSE ditunjukkan pada Persamaan 2.34 (Kusuma et al., 2023).

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (Y_i - \hat{Y}_i)^2} \tag{2.34}$$

Berdasarkan Persamaan 2.34, nilai RMSE yang kecil menunjukkan bahwa prediksi model memiliki deviasi yang rendah terhadap nilai observasi. Karena berada pada skala asli data, RMSE memberikan gambaran yang lebih intuitif mengenai besar kesalahan prediksi yang dihasilkan oleh model (Chicco et al., 2021).

2.2.7.3 *Mean Absolute Error* (MAE)

Metrik evaluasi lainnya yang digunakan adalah *Mean Absolute Error* (MAE), yaitu rata-rata nilai absolut dari selisih antara nilai prediksi dan nilai observasi (Rifa'i et al., 2019). MAE lebih tangguh terhadap pengaruh outlier dibandingkan MSE dan RMSE karena tidak melibatkan pengkuadratan selisih. Secara matematis, perhitungan MAE ditunjukkan pada Persamaan 2.35.

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m} | Y_i - \hat{Y}_i | \tag{2.35}$$

Berdasarkan Persamaan 2.35, nilai MAE yang rendah menunjukkan bahwa secara rata-rata kesalahan prediksi model relatif kecil (Chicco et al., 2021).

2.2.7.4 *Confusion Matrix*

Confusion matrix merupakan alat evaluasi yang digunakan untuk menilai kinerja model dengan membandingkan antara kelas aktual dan kelas hasil prediksi (Ummah et al., 2022). Pada permasalahan klasifikasi empat kelas, *confusion matrix* berbentuk matriks berukuran 4×4, di mana setiap baris merepresentasikan kelas aktual dan setiap kolom menunjukkan kelas hasil prediksi model (Tharwat, 2026).

Tabel 2. 4 Struktur Confusion Matrix Empat Kelas

Kelas Aktual \ Prediksi	Kelas 1	Kelas 2	Kelas 3	Kelas 4
Kelas 1	n_{11}	n_{12}	n_{13}	n_{14}
Kelas 2	n_{21}	n_{22}	n_{23}	n_{24}
Kelas 3	n_{31}	n_{32}	n_{33}	n_{34}
Kelas 4	n_{41}	n_{42}	n_{43}	n_{44}

Elemen diagonal utama $(n_{11}, n_{22}, n_{33}, n_{44})$ menunjukkan jumlah data yang diklasifikasikan dengan benar untuk masing-masing kelas, sedangkan elemen di luar diagonal menunjukkan kesalahan klasifikasi antar kelas.

Untuk menghitung metrik evaluasi pada klasifikasi multikelas, digunakan pendekatan *one-vs-rest*, yaitu setiap kelas diperlakukan sebagai kelas positif secara bergantian, sedangkan kelas lainnya dianggap sebagai kelas negatif. Pendekatan ini memungkinkan perhitungan komponen evaluasi pada masing-masing kelas secara terpisah sehingga performa model dapat dianalisis secara lebih rinci. Pada klasifikasi empat kelas, setiap kelas ke-i $(\mathrm{i} = 1,2,3,4)$ memiliki komponen *True Positive* (TP), *False Negative* (FN), *False Positive* (FP), dan *True Negative* (TN) yang didefinisikan berdasarkan posisi data pada *confusion matrix* (Zeng, 2025).

True Positive ($\mathrm{TP_i}$) menyatakan jumlah data dari kelas ke-i yang berhasil diprediksi dengan benar sebagai kelas ke-i. Nilai ini diperoleh dari elemen diagonal utama *confusion matrix* yang bersesuaian dengan kelas tersebut (Zeng, 2025). Secara matematis, nilai *True Positive* dirumuskan menggunakan Persamaan 2.36.

$$\mathrm{TP_i} = \mathrm{n_{ii}} \tag{2.36}$$

Berdasarkan Persamaan 2.36, nilai $\mathrm{n_{ii}}$ merupakan elemen diagonal pada *confusion matrix* yang menunjukkan jumlah data dari kelas ke-i yang diprediksi secara tepat sebagai kelas ke-i.

False Negative ($\mathrm{FN_i}$) merupakan jumlah data yang berasal dari kelas ke-i tetapi salah diprediksi sebagai kelas lain. Nilai ini menunjukkan kegagalan model dalam mengenali data yang seharusnya termasuk ke dalam kelas ke-i (Zeng, 2025). Secara matematis, *False Negative* dirumuskan menggunakan Persamaan 2.37.

$$\mathrm{FN_i} = \sum_{\mathrm{j=1,j\neq i}}^{4} \mathrm{n_{ij}} \tag{2.37}$$

Berdasarkan Persamaan 2.37, nilai $\mathrm{n_{ij}}$ merepresentasikan jumlah data dari kelas ke-i yang diprediksi sebagai kelas ke-j, dengan $\mathrm{j} \neq \mathrm{i}$. Penjumlahan terhadap seluruh kelas

selain kelas ke-i menunjukkan total kesalahan prediksi model dalam mengenali data yang seharusnya termasuk ke dalam kelas ke-i.

False Positive (FP_i) menunjukkan jumlah data dari kelas lain yang secara keliru diprediksi sebagai kelas ke-i. Nilai ini mencerminkan kesalahan model dalam memberikan prediksi positif palsu terhadap kelas ke-i (Zeng, 2025). Secara matematis, *False Positive* dirumuskan menggunakan Persamaan 2.38.

$$FP_i = \sum_{j=1,j\neq i}^{4} n_{ji} \tag{2.38}$$

Berdasarkan Persamaan 2.38, nilai n_{ji} merepresentasikan jumlah data dari kelas ke-j yang diprediksi sebagai kelas ke-i, dengan $j \neq i$. Penjumlahan seluruh nilai tersebut menunjukkan tingkat kesalahan model dalam memberikan prediksi positif palsu terhadap kelas ke-i.

True Negative (TN_i) menyatakan jumlah data dari kelas lain yang diprediksi dengan benar sebagai bukan kelas ke-i (Zeng, 2025). Secara matematis, *True Negative* dirumuskan menggunakan Persamaan 2.39.

$$TN_i = \sum_{j=1}^{4} \sum_{k=1}^{4} n_{jk} - (TP_i + FP_i + FN_i) \tag{2.39}$$

Berdasarkan Persamaan 2.39, nilai TN_i diperoleh dari total seluruh data pada *confusion matrix* yang dikurangi dengan jumlah *True Positive* (TP_i), *False Positive* (FP_i), dan *False Negative* (FN_i) untuk kelas ke-i. Dengan demikian, TN_i merepresentasikan kemampuan model dalam mengidentifikasi data negatif secara benar terhadap kelas ke-i.

Setelah nilai TP_i, FP_i, FN_i, dan TN_i diperoleh, metrik evaluasi kinerja model dapat dihitung untuk masing-masing kelas. *Accuracy* digunakan untuk menggambarkan proporsi prediksi yang benar terhadap seluruh data yang diuji. Pada klasifikasi empat kelas, akurasi dihitung berdasarkan total prediksi benar dari seluruh kelas terhadap jumlah keseluruhan data menggunakan Persamaan 2.40.

$$\text{Accuracy} = \frac{\sum_{i=1}^{4} \text{TP}_i}{\sum_{j=1}^{4} \sum_{k=1}^{4} n_{jk}} \quad (2.40)$$

Berdasarkan Persamaan 2.40, pembilang pada persamaan tersebut, $\sum_{i=1}^{4} T\,P_i$, menunjukkan jumlah total prediksi yang benar dari seluruh kelas, sedangkan penyebutnya, $\sum_{j=1}^{4} \sum_{k=1}^{4} n_{jk}$, merepresentasikan jumlah keseluruhan data pengujian. Dengan demikian, nilai akurasi memberikan gambaran umum mengenai kemampuan model dalam mengklasifikasikan data secara tepat pada seluruh kelas.

Precision untuk kelas ke-i mengukur tingkat ketepatan prediksi model terhadap kelas tersebut, yaitu seberapa besar proporsi data yang diprediksi sebagai kelas ke-i benar-benar berasal dari kelas ke-i. Secara matematis, *Precision* dirumuskan menggunakan Persamaan 2.41 (Yosrita et al., 2021).

$$\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \quad (2.41)$$

Berdasarkan Persamaan 2.41, Perbandingan antara TP_i dan jumlah $\text{TP}_i + \text{FP}_i$ menghasilkan nilai *Precision*, yang merepresentasikan tingkat ketepatan model dalam memberikan prediksi positif pada kelas ke-i.

Recall atau *Sensitivity* untuk kelas ke-i mengukur kemampuan model dalam mendeteksi seluruh data yang benar-benar termasuk ke dalam kelas ke-i. Nilai *Recall* yang tinggi menunjukkan bahwa model memiliki tingkat kegagalan yang rendah dalam mengenali data kelas tersebut. Secara matematis, *Recall* dirumuskan menggunakan Persamaan 2.42 (Yosrita et al., 2021).

$$\text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \quad (2.42)$$

Berdasarkan Persamaan 2.42, perbandingan antara TP_i dan jumlah $TP_i + FN_i$ menghasilkan nilai Recall, yang merepresentasikan kemampuan model dalam mendeteksi seluruh data yang sebenarnya termasuk ke dalam kelas ke-i.

F1-Score ($F1_i$) merupakan metrik yang menggabungkan *Precision* dan *Recall* dalam bentuk rata-rata harmonik (Zeng, 2025). Secara matematis, *F1-Score* untuk kelas ke-i dirumuskan menggunakan Persamaan 2.43.

$$F1_i = \frac{2TP_i}{2TP_i + FP_i + FN_i} \tag{2.43}$$

Berdasarkan Persamaan 2.43, Nilai *F1-Score* memberikan ukuran kinerja model yang seimbang dengan mempertimbangkan ketepatan dan kelengkapan prediksi pada kelas ke-i.

Untuk memperoleh satu nilai evaluasi yang mewakili kinerja model secara keseluruhan pada klasifikasi empat kelas, nilai metrik dari masing-masing kelas dapat dirata-ratakan menggunakan beberapa pendekatan. *Macro-average* menghitung rata-rata sederhana dari metrik setiap kelas tanpa mempertimbangkan jumlah data pada masing-masing kelas, sehingga setiap kelas memiliki bobot yang sama (Farhadpour et al., 2024). *Macro-average* dirumuskan menggunakan Persamaan 2.44.

$$\text{Metric}_{\text{macro}} = \frac{1}{4}\sum_{i=1}^{4} \text{Metric}_i \tag{2.44}$$

Selain itu, digunakan pula *weighted-average*, yaitu pendekatan perataan metrik dengan mempertimbangkan proporsi jumlah data pada setiap kelas. Pendekatan ini lebih representatif ketika distribusi data antar kelas tidak seimbang (Farhadpour et al., 2024). *Weighted-average* dirumuskan menggunakan Persamaan 2.45.

$$\text{Metric}_{\text{weighted}} = \sum_{i=1}^{4} w_i \cdot \text{Metric}_i \tag{2.45}$$

dengan w_i menyatakan proporsi jumlah data pada kelas ke-i.

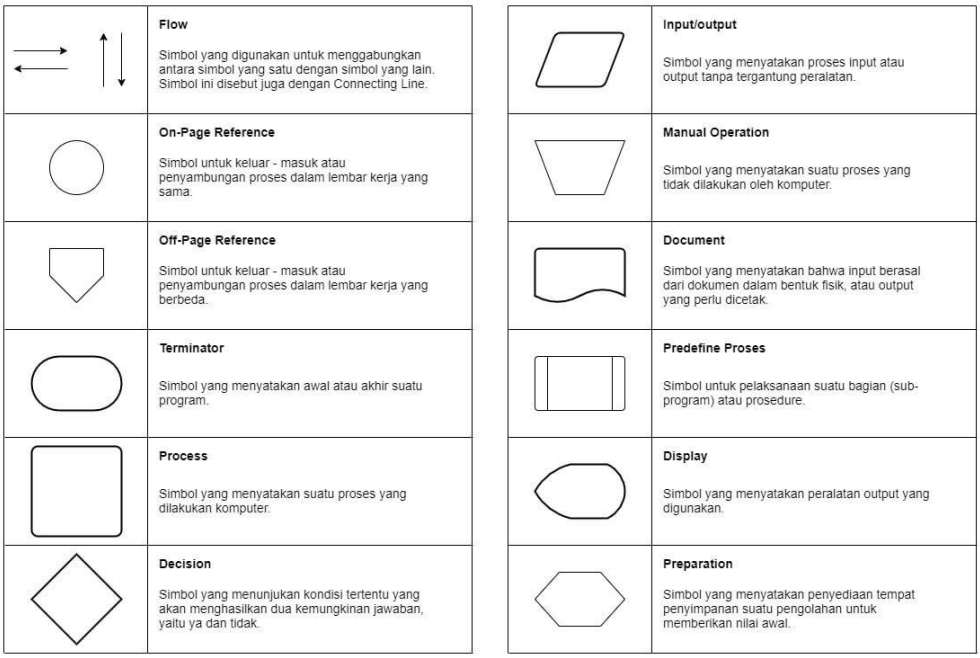
Dengan demikian, *confusion matrix* pada klasifikasi empat kelas tidak hanya berfungsi untuk menunjukkan jumlah prediksi yang benar dan salah, tetapi juga menjadi dasar utama dalam perhitungan berbagai metrik evaluasi kinerja model. Pendekatan *one-vs-rest* memungkinkan evaluasi performa model dilakukan secara terperinci pada setiap kelas, sehingga sangat sesuai digunakan dalam penelitian klasifikasi multikelas yang membutuhkan analisis kinerja model secara komprehensif dan sistematis.

2.2.8 *Flow Chart*

Flowchart merupakan sebuah representasi visual yang memetakan langkah-langkah, keputusan, dan alur kerja dalam sebuah proses atau program secara sistematis. Dengan menggunakan serangkaian simbol standar yang dihubungkan oleh garis panah, *flowchart* mampu menyajikan urutan logika yang kompleks menjadi format yang lebih sederhana dan mudah dicerna.

Fungsi utama dari *flowchart* adalah untuk memberikan gambaran menyeluruh mengenai jalannya sebuah program atau prosedur, sehingga dapat dipahami dengan mudah oleh berbagai pihak, terlepas dari latar belakang teknis mereka. Dengan menyederhanakan rangkaian prosedur yang rumit, *flowchart* secara efektif mengurangi ambiguitas dan potensi terjadinya salah penafsiran terhadap alur kerja. Hal ini menjadikannya alat yang sangat berharga dalam tahap perencanaan dan dokumentasi sebuah proyek.

Selain sebagai alat visualisasi, *flowchart* juga berfungsi sebagai media komunikasi yang efektif antara pihak teknis dan non-teknis. Melalui visualisasi alur yang jelas, seluruh pihak yang terlibat dapat memperoleh pemahaman yang seragam mengenai logika, fungsi, dan tujuan dari sistem atau metode yang diterapkan. Oleh karena itu, flowchart menjadi instrumen yang penting dalam perencanaan, dokumentasi, dan penyampaian informasi pada suatu penelitian atau proyek pengembangan sistem (Setiawan, 2021)



Gambar 2. 2 *Flowchart* (dicoding, 2021)

2.2.9 *Python*

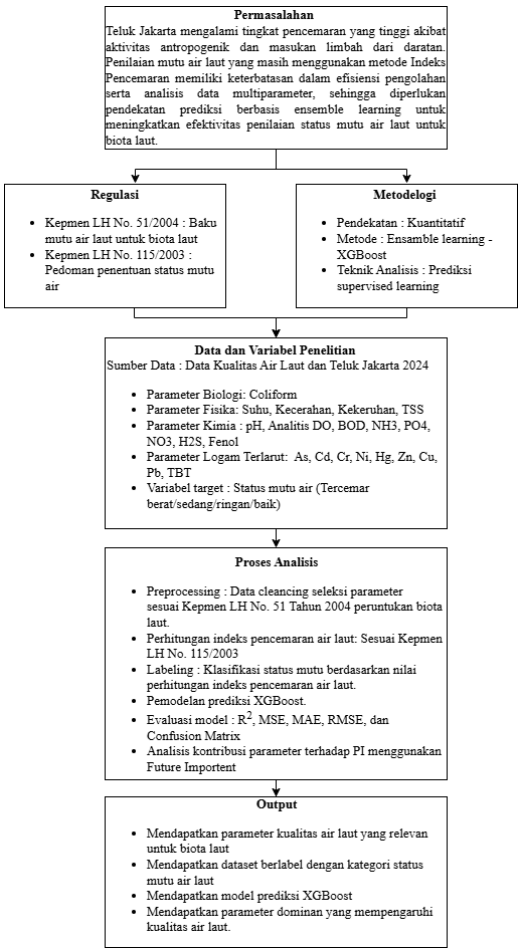
Python adalah sebuah bahasa pemrograman tingkat tinggi yang bersifat interpretatif dan serbaguna atau *general-purpose*, yang pertama kali diperkenalkan oleh Guido van Rossum pada 1991. Bahasa ini, yang namanya terinspirasi dari sebuah tayangan komedi, dikenal karena keterbacaan sintaksnya yang tinggi, menjadikannya dapat diakses oleh berbagai kalangan pengembang. Fleksibilitasnya didukung oleh kemampuannya mengakomodasi berbagai paradigma pemrograman, termasuk prosedural, berorientasi objek (OOP), dan fungsional. Selain itu, sifatnya yang portabel memungkinkan kode dijalankan di berbagai sistem operasi tanpa modifikasi signifikan. Ekosistem Python yang ekstensif, didukung oleh *Python Package Index* (PyPI), menawarkan akses ke ribuan pustaka untuk beragam domain aplikasi, seperti pengembangan web, analisis data, dan *machine learning* (Hanif, 2024).

2.2.10 Kerangka Penelitian

Kerangka penelitian ini disusun untuk menggambarkan alur berpikir dan tahapan penelitian secara sistematis, mulai dari identifikasi permasalahan hingga diperolehnya

luaran penelitian. Penelitian ini berangkat dari permasalahan tingginya tekanan pencemaran di perairan Teluk Jakarta akibat aktivitas antropogenik dan masukan limbah dari daratan, serta karakteristik data kualitas air laut yang bersifat multiparameter. Di sisi lain, metode Indeks Pencemaran yang selama ini digunakan secara konvensional masih memiliki keterbatasan karena bersifat manual, deskriptif, dan belum mampu mengolah data dalam jumlah besar maupun memberikan prediksi status mutu air secara otomatis dan akurat.

Tingginya kompleksitas permasalahan tersebut menuntut adanya metode penilaian mutu air laut yang tidak hanya sesuai dengan regulasi nasional, tetapi juga mampu mengolah data kualitas air secara efisien dan konsisten. Oleh karena itu, penelitian ini mengintegrasikan pendekatan konvensional Indeks Pencemaran dengan metode prediksi berbasis machine learning untuk meningkatkan efektivitas penilaian dan penentuan status mutu air laut di Teluk Jakarta.



Gambar 2. 3 Kerangka Penelitian

Berdasarkan Gambar 2.3, penelitian ini berlandaskan pada kerangka regulasi nasional, yaitu Keputusan Menteri Lingkungan Hidup Nomor 51 Tahun 2004 tentang baku mutu air laut untuk biota laut dan Keputusan Menteri Lingkungan Hidup Nomor 115 Tahun 2003 tentang pedoman penentuan status mutu air. Regulasi ini digunakan sebagai acuan dalam menentukan ambang batas parameter kualitas air laut serta sebagai dasar perhitungan Indeks Pencemaran yang selanjutnya dimanfaatkan sebagai label kelas pada proses prediksi.

Data yang digunakan dalam penelitian ini berupa data kualitas air laut Teluk Jakarta tahun 2024 yang mencakup parameter fisika, kimia, biologi, dan logam terlarut. Parameter fisika meliputi suhu, kecerahan, kekeruhan, dan total padatan tersuspensi (TSS), sementara parameter kimia mencakup pH, oksigen terlarut (DO), kebutuhan oksigen biokimia (BOD), nutrien seperti nitrat, nitrit, fosfat, amonia, serta senyawa pencemar lainnya. Parameter biologi direpresentasikan oleh koliform total, sedangkan parameter logam terlarut meliputi As, Cd, Cr, Ni, Hg, Zn, Cu, Pb, dan TBT. Status mutu air laut yang dihasilkan dari perhitungan Indeks Pencemaran digunakan sebagai variabel target dalam proses prediksi, dengan kategori baik, tercemar ringan, tercemar sedang, dan tercemar berat.

Tahapan analisis diawali dengan proses *preprocessing,* yang mencakup pembersihan data, serta seleksi parameter sesuai dengan peruntukan biota laut berdasarkan Keputusan Menteri Lingkungan Hidup Nomor 51 Tahun 2004. Selanjutnya dilakukan perhitungan Indeks Pencemaran sesuai Keputusan Menteri Lingkungan Hidup Nomor 115 Tahun 2003 untuk menentukan status mutu air laut yang kemudian dijadikan sebagai label kelas pada dataset.

Proses prediksi dilakukan menggunakan algoritma *Extreme Gradient Boosting* (XGBoost) sebagai bagian dari pendekatan *ensemble learning* dengan teknik supervised learning. Model dibangun untuk mempelajari hubungan nonlinier antarparameter kualitas air laut dan memprediksi status mutu air berdasarkan data latih. Kinerja model dievaluasi menggunakan beberapa metrik evaluasi, antara lain *koefisien determinasi* (R^2), *Mean Squared Error* (MSE), *Mean Absolute Error* (MAE), *Root Mean Squared Error* (RMSE), serta *confusion matrix* untuk menilai kemampuan klasifikasi model secara komprehensif.

Sebagai tahap akhir, dilakukan analisis kontribusi parameter kualitas air laut terhadap hasil prediksi menggunakan *fitur importance*. Analisis ini bertujuan untuk mengidentifikasi parameter dominan yang berpengaruh terhadap status mutu air laut, sehingga hasil penelitian tidak hanya menghasilkan model prediksi yang akurat, tetapi juga memberikan informasi yang relevan bagi pengelolaan kualitas lingkungan laut dan perlindungan biota laut secara berkelanjutan.