BAB I
PENDAHULUAN

1.1 Latar Belakang

Pencemaran air merupakan permasalahan lingkungan global yang berdampak langsung terhadap kesehatan manusia, keberlanjutan ekosistem, dan pembangunan sosial ekonomi. Aktivitas industri, pertanian, serta limbah domestik yang tidak terkelola dengan baik telah menyebabkan penurunan kualitas air permukaan dan air tanah di berbagai wilayah. Kondisi ini menuntut adanya metode penilaian kualitas air yang akurat dan efisien untuk mendukung pengambilan keputusan lingkungan [1], [2]. Penilaian kualitas air yang tepat menjadi fondasi penting dalam upaya pengendalian pencemaran dan perlindungan sumber daya air.

Metode konvensional dalam analisis kualitas air umumnya mengandalkan pengujian laboratorium dan pendekatan statistik klasik. Meskipun metode tersebut memiliki tingkat akurasi yang baik, prosesnya sering kali memerlukan waktu lama dan kurang adaptif terhadap data berskala besar serta berpola nonlinier. Seiring meningkatnya ketersediaan data kualitas air publik, pendekatan berbasis data menjadi semakin relevan. Penelitian sebelumnya menunjukkan bahwa metode statistik tradisional memiliki keterbatasan dalam menangani kompleksitas data kualitas air modern [3]. Oleh karena itu, dibutuhkan pendekatan yang lebih fleksibel dan adaptif.

Machine learning telah banyak diterapkan dalam analisis kualitas air karena kemampuannya memodelkan hubungan nonlinier dan menangani data multidimensi. Berbagai algoritma klasifikasi seperti *Decision Tree*, *Random Forest*, *Support Vector Machine* , *Naïve Bayes*, dan *Logistic Regression* telah digunakan untuk menentukan tingkat kualitas maupun potabilitas air berdasarkan parameter fisikokimia. Hasil penelitian menunjukkan bahwa model berbasis ensemble sering memberikan performa yang lebih baik dibandingkan model tunggal [4], [5]. Selain itu, studi tinjauan sistematis juga menegaskan keunggulan *machine learning* dalam meningkatkan akurasi dan robustitas klasifikasi kualitas

air.

Meskipun banyak penelitian telah mengkaji penerapan *machine learning* dalam klasifikasi kualitas air, perbedaan metode evaluasi dan pemilihan metrik kinerja masih menjadi permasalahan. Beberapa penelitian hanya berfokus pada nilai akurasi tanpa mempertimbangkan metrik lain seperti *precision*, *Recall*, *F1-Score*, dan *ROC-AUC*. Padahal, evaluasi multi-metrik sangat penting terutama pada data yang tidak seimbang [6]. Oleh karena itu, diperlukan penelitian yang melakukan perbandingan performa berbagai metode klasifikasi secara sistematis menggunakan metrik evaluasi yang komprehensif.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Kondisi dataset kualitas air yang memiliki nilai kosong (*missing values*) yang signifikan pada parameter krusial seperti *pH*, *Sulfate*, dan *Trihalomethanes* berpotensi menurunkan performa model jika tidak ditangani dengan teknik Mean Imputation yang tepat.
2. Adanya ketidakseimbangan kelas (*imbalanced data*) pada label potabilitas, di mana sampel air tidak layak minum lebih dominan dibandingkan air layak minum, sehingga diperlukan penerapan teknik Manual *Oversampling* untuk mencegah model mengalami bias klasifikasi.
3. Belum diketahuinya tingkat signifikansi masing-masing dari sembilan parameter fisikokimia terhadap prediksi potabilitas, sehingga diperlukan identifikasi fitur kunci menggunakan metode seleksi fitur *ANOVA F-value* (*SelectKBest*).
4. Ketidakpastian mengenai algoritma klasifikasi yang paling reliabel di antara *Decision Tree*, *Random Forest*, *Naive Bayes*, *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), *Logistic Regression*, dan *XGBoost* dalam memberikan akurasi tertinggi dan nilai *ROC-AUC* yang optimal untuk data kualitas air minum beserta implementasi.

1.3 Batasan Masalah

Batasan masalah dalam penelitian ini ditetapkan untuk memfokuskan ruang lingkup penelitian, yaitu sebagai berikut:

1. Dataset: Penelitian ini hanya menggunakan dataset sekunder *Water Potability* dari Kaggle dengan total 3.276 observasi.
2. Metodologi: Pra-pemrosesan data difokuskan pada Penanganan *Mising Values*, *Standard Scaling*, dan penyeimbangan data secara manual (tanpa membandingkan dengan teknik *Random Oversampling*).
3. Metode klasifikasi yang digunakan meliputi *Decision Tree*, *Random Forest*, *Support Vector Machine* , *Naïve Bayes*, *K-Nearest Neighbor*, *Logistic Regression*, dan *XGBoost*.
4. Evaluasi model dilakukan menggunakan skema pembagian data latih dan data uji dengan rasio 90:10. Dengan metrik evaluasi yang digunakan meliputi *accuracy*, *precision*, *Recall*, *F1-Score*, dan *ROC-AUC*.
5. Komputasi: Implementasi model dilakukan menggunakan bahasa pemrograman Python dengan pustaka utama *Scikit-Learn* dan *XGBoost*.

1.4 Tujuan Penelitian

Tujuan yang akan dicapai oleh dalam penelitian ini adalah menghasilkan sebuah sistem klasifikasi cerdas yang mampu menganalisis efektivitas teknik *Mean Imputation* dan *Standard Scaling* dalam menangani ketidaklengkapan data serta menstandarisasi distribusi parameter fisikokimia air agar siap diolah oleh algoritma. Selain itu, penelitian ini bertujuan untuk memvalidasi penerapan teknik *Manual Oversampling* sebagai strategi optimal dalam mengatasi ketidakseimbangan kelas (*imbalanced data*) guna meningkatkan sensitivitas model terhadap sampel air layak minum. Lebih lanjut, penelitian ini berupaya menghitung skor signifikansi fitur melalui metode seleksi *ANOVA F-value* untuk mengidentifikasi parameter kualitas air yang paling berpengaruh terhadap label potabilitas. Pada akhirnya, penelitian ini ditujukan untuk membandingkan performa tujuh algoritma *machine learning*, yaitu *Decision Tree, Random Forest, Naive Bayes, SVM, KNN, Logistic Regression,* dan

XGBoost, guna menentukan model dengan perolehan akurasi, *F1-Score*, dan *ROC-AUC* terbaik yang paling reliabel untuk direkomendasikan dalam sistem pengujian kualitas air otomatis.

1.5 Manfaat Penelitian

Manfaat penelitian ini secara teoritis diharapkan dapat memberikan kontribusi ilmiah bagi perkembangan ilmu *Data Science* di bidang lingkungan, khususnya dalam memberikan bukti empiris mengenai efektivitas integrasi teknik pra-pemrosesan data seperti *Mean Imputation* dan *Manual Oversampling* pada dataset yang memiliki karakteristik nilai hilang dan ketidakseimbangan kelas. Penelitian ini juga memperkaya literatur mengenai perbandingan performa algoritma klasifikasi, mulai dari model linier hingga algoritma *ensemble* tingkat tinggi seperti *XGBoost* dan *Random Forest* dalam konteks prediksi potabilitas air. Secara praktis, hasil penelitian ini dapat dimanfaatkan oleh instansi pengelola sumber daya air atau laboratorium kesehatan sebagai referensi dalam membangun sistem pemantauan kualitas air otomatis yang lebih cepat dan efisien dibandingkan metode pengujian manual. Bagi pengembang teknologi, kerangka kerja (*framework*) pengolahan data yang dihasilkan dapat digunakan sebagai dasar algoritma untuk aplikasi pendeteksi kelayakan air minum berbasis sensor, sehingga masyarakat dapat memperoleh informasi mengenai keamanan air konsumsi secara lebih akurat dan objektif.

1.6 Sistematika Penulisan

Sistematika penulisan skripsi ini disusun sebagai berikut:

BAB I PENDAHULUAN, berisi latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA, berisi kajian literatur dan teori-teori yang mendukung penelitian terkait kualitas air dan metode *machine learning*.

BAB III METODE PENELITIAN, menjelaskan dataset, tahapan praproses data, metode klasifikasi, serta skema evaluasi yang digunakan.

BAB IV HASIL DAN PEMBAHASAN, menyajikan hasil pengujian model,

analisis performa, dan pembahasan hasil penelitian.

BAB V PENUTUP, berisi kesimpulan yang diperoleh dari penelitian serta saran untuk penelitian selanjutnya.