

BAB I

PENDAHULUAN

1.1 Latar Belakang

Air merupakan sumber daya alam yang sangat berharga dan penting untuk keperluan keberlangsungan makhluk hidup di bumi. Air minum yang tidak terkontaminasi dan aman merupakan keutamaan yang sangat penting untuk kesehatan manusia dikarenakan mengkonsumsi air yang tidak aman dapat membahayakan ginjal dan hati [1]. Kualitas air sangat penting bagi kehidupan manusia dan merupakan komponen penting dari berbagai kebutuhan manusia, baik untuk dikonsumsi maupun kegiatan industri. Peningkatan populasi dan aktivitas manusia dapat memengaruhi berbagai faktor kualitas air, termasuk pencemaran industri, kontaminasi bahan kimia, dan masalah alami seperti kandungan mineral yang berlebihan [2].

Selain itu, kualitas air dinilai dalam empat kategori: air minum yang aman untuk dikonsumsi dan digunakan rumah tangga, air enak yang mengandung unsur kimia yang mungkin berbahaya bagi kesehatan manusia, air terkontaminasi yang mengandung zat kimia, biologi, dan bahkan radiologi, dan tidak cocok untuk keperluan rumah tangga, dan air terinfeksi yang terkontaminasi oleh bakteri berbahaya [3]. Selanjutnya, untuk menentukan apakah air aman untuk dikonsumsi manusia dan tidak akan menyebabkan bahaya atau penyakit, penting untuk mengevaluasi apakah air tersebut memenuhi standar yang ditetapkan dan tidak menimbulkan risiko terhadap kesehatan manusia. Untuk menilai dan memantau kualitas air secara efektif, berbagai faktor dapat dipertimbangkan, termasuk tingkat pH, kekerasan, padatan, konduktivitas, karbon organik, dan kekeruhan [4].

Kombinasi perubahan iklim dan pertumbuhan penduduk dapat memperburuk kondisi kualitas air, sehingga menimbulkan potensi risiko terhadap kesehatan manusia [5]. Di era modern, teknologi berkembang dengan pesat sehingga tidak diragukan lagi bahwa para peneliti menggunakan *machine learning* dalam berbagai bidang penelitian. Jenis-jenis utama *machine learning* terbagi

menjadi tiga yaitu: *supervised learning*, *unsupervised learning*, dan *reinforcement learning* [6]. Oleh sebab itu, peneliti kali ini bertujuan untuk melakukan penelitian lebih lanjut menggunakan dataset kualitas air serta algoritma *Random Forest*, dengan menerapkan optimasi *hyperparameter* pada metode *grid search* serta penggunaan teknik *oversampling* untuk mengatasi ketidakseimbangan kelas. Untuk mengatasi masalah ketidakseimbangan kelas, pendekatan yang efektif adalah dengan menerapkan teknik *oversampling* untuk kelas-kelas yang kurang terwakili. Di mana jumlah sampel dalam kelas positif potable (1) lebih sedikit daripada kelas not potable (0). Dengan menerapkan *oversampling*, performa model pada kelas minoritas dapat ditingkatkan tanpa perlu menghapus sampel dari kelas mayoritas [7].

Pemanfaatan pendekatan tuning melibatkan penerapan objek *GridSearchCV* dengan model. Pemilihan metode *grid search* dilakukan untuk mengidentifikasi konfigurasi optimasi terbaik untuk model pengklasifikasi random forest, sehingga dapat meningkatkan kinerja klasifikasi pada dataset yang disediakan dan mempengaruhi tingkat akurasi untuk meningkatkan efektivitas model [8]. Penelitian ini mengimplementasikan teknik *oversampling* untuk mencapai kelas data yang seimbang beserta dengan normalisasi data menggunakan *standartscaler*. Selain itu, pembagian data dilakukan melalui *K-fold Cross Validation* untuk memastikan pengukuran kinerja model yang lebih tepat. Selanjutnya, menggunakan evaluasi model *confusion matrix* serta *classification matrix*.

Berdasarkan penelitian ini, terdapat permasalahan penelitian yang dimana terletak pada pentingnya kualitas air bagi kesehatan tubuh manusia dan lingkungan. Dengan pertumbuhan populasi yang pesat, banyak daerah yang mengalami kekurangan air bersih. Sehingga, tujuan dari penelitian ini ialah mengembangkan model klasifikasi yg mampu memprediksi potensi kualitas air. Prediksi ini didasarkan pada banyak fitur termasuk pH, kesadahan air, padatan terlarut, kloramin, sulfat, konduktivitas, karbon organik, trihalometana, serta kekeruhan. Selain itu, hasil penelitian ini pula diharapkan bisa menyampaikan kontribusi dalam

meningkatkan pemahaman mengenai kinerja algoritma random forest dalam bidang klasifikasi.

1.2 Rumusan Masalah

Adapun rumusan masalah pada penelitian ini sebagai berikut:

1. Bagaimana cara mengoptimalkan penggunaan model *Random Forest* dengan metode *Grid Search* untuk meningkatkan prediksi kualitas air?
2. Bagaimana kinerja model *Random Forest* yang dioptimalkan dengan metode *Grid Search* dibandingkan tanpa metode *Grid Search* dalam hal akurasi, presisi, recall dan *f1-score*?

1.3 Batasan Masalah

Batasan masalah dalam penelitian ini dapat mencakup beberapa batasan sebagai berikut:

1. Penelitian ini difokuskan pada algoritma *Random Forest* dan penggunaan metode *grid search*.
2. Penelitian ini mengambil data dari dataset publik pada platform website *kaggle*.
3. Penelitian ini menggunakan bahasa pemrograman python
4. Penelitian ini melakukan evaluasi model menggunakan *confusion matrix* dengan matriks seperti *accuracy*, *precision*, *recall* dan *f1-score*.

1.4 Tujuan Penelitian

Tujuan penelitian berfokus untuk mengoptimalkan penggunaan model *Random Forest* dengan optimasi metode *grid search* yang berguna untuk meningkatkan akurasi serta untuk mengetahui hasil prediksi menggunakan matriks evaluasi seperti *accuracy*, *precision*, *recall* dan *f1-score*.

1.5 Manfaat Penelitian

Manfaat pada penelitian ini akan dilampirkan dibawah ini:

1. Secara praktis, hasil penelitian ini dapat membantu mengevaluasi kinerja model klasifikasi dalam memprediksi kualitas air yang layak

dikonsumsi.

2. Secara teoritis, hasil penelitian ini dapat digunakan sebagai referensi untuk penelitian selanjutnya dalam memprediksi kualitas air menggunakan algoritma *random forest* dengan metode *grid search*.

1.6 Sistematika Penulisan

Sistematika penulisan digunakan sebagai penyusunan penulisan agar mudah dibaca, berikut dibawah ini penyusunan penulisan dalam penelitian ini:

BAB I PENDAHULUAN

Pada bagian bab ini berisikan latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian dan sistematika penulisan dalam penelitian ini.

BAB II TINJAUAN PUSTAKA

Pada bab ini membahas mengenai teori-teori yang berkaitan dengan topik penelitian seperti *grid search*, algoritma *random forest*, kualitas air, klasifikasi, teknik *oversampling*, serta *confusion matrix*.

BAB III METODE PENELITIAN

Bab ini membahas mengenai alur penelitian yang berupa *flowchart* dengan dimulai dari pengumpulan data, *preprocessing data*, implementasi algoritma *random forest* serta penggunaan metode *grid search* hingga ke tahap akhir yaitu evaluasi.

BAB IV HASIL DAN PEMBAHASAN

Bab ini berikan hasil implementasi algoritma *random forest* serta penggunaan metode *grid search* dan hasil dari evaluasi model menggunakan *confusion matrix* dan *classification report*.

BAB V PENUTUP

Bab terakhir berisi kesimpulan dan saran dari hasil penelitian menggunakan algoritma *random forest* dan metode *grid search* akan dijabarkan secara singkat dan mudah dipahami.