

BAB I PENDAHULUAN

1.1 Latar Belakang Masalah

Kualitas air merupakan faktor fundamental dalam menjaga kesejahteraan manusia dan kelangsungan ekosistem, mengingat ketersediaan air bersih yang aman berperan penting dalam mendukung aktivitas sosial dan ekonomi. Kualitas air ditentukan oleh berbagai parameter fisika, kimia, dan biologi yang mencerminkan kondisi suatu sumber air serta kelayakannya untuk digunakan sesuai peruntukan tertentu [1]. Selain itu parameter lain yang berkaitan juga akan menjadi pembahasan dalam penelitian ini. Kompleksitas dan keterkaitan antar parameter tersebut menyebabkan proses evaluasi kualitas air menjadi tidak sederhana, terutama ketika melibatkan volume data yang besar dan beragam. Di Indonesia, hasil pemantauan mutu air menunjukkan adanya upaya perbaikan, seperti peningkatan mutu pada sungai, namun permasalahan fragmentasi data dan ketimpangan pengelolaan kualitas air antarwilayah masih menjadi tantangan yang signifikan [2]. Dengan demikian, diperlukan peningkatan sistem pemantauan dan klasifikasi kualitas air yang mampu merespon kondisi secara lebih akurat dan efisien.

Seiring perkembangan teknologi, penerapan *machine learning* telah membuka peluang untuk mengevaluasi serta mengklasifikasikan kualitas air dengan cara yang lebih cepat dan efisien [3]. Namun, penerapan model ML tunggal dalam studi klasifikasi kualitas air seringkali belum mencapai performa optimal. Sebuah studi yang membandingkan beberapa algoritma untuk kelayakan air minum mencatat akurasi 66,6% untuk Random Forest [4]. Penelitian lain yang spesifik membandingkan algoritma *boosting* pada data kualitas air juga melaporkan akurasi yang masih perlu ditingkatkan, yaitu 68% untuk CatBoost, 60% untuk Gradient Boosting dan Ada Boost diangka 58% [5]. Upaya lain menggunakan model tunggal yang lebih canggih, seperti XGBoost, juga menunjukkan hasil yang lebih baik namun masih belum optimal. Sebuah penelitian yang menguji XGBoost melaporkan bahwa model tersebut hanya mencapai akurasi 82,29% [6].

Menyadari keterbatasan performa model tunggal tersebut, pendekatan *ensemble stacking* menawarkan alternatif yang menjanjikan untuk meningkatkan akurasi klasifikasi. Sebuah studi yang secara spesifik membandingkan model tunggal dengan model *ensemble* untuk identifikasi baku mutu air menyimpulkan bahwa metode *ensemble* mampu memberikan performa yang lebih unggul [7]. Secara khusus, metode *ensemble stacking* telah terbukti efektif dalam meningkatkan kinerja secara signifikan meliputi akurasi, presisi, dan recall dibandingkan dengan penggunaan *classifier* tunggal [8]. Arsitektur *stacking* bekerja dengan mengkombinasikan prediksi dari beberapa model *base-learner*, yang hasilnya kemudian digunakan sebagai fitur untuk melatih sebuah *meta-learner* [9]. Pendekatan ini bertujuan memanfaatkan kelebihan dari masing-masing algoritma dasar, sehingga menghasilkan keputusan klasifikasi akhir yang lebih robust dan akurat [10].

Pemilihan *base-learner* yang kuat menjadi krusial dalam arsitektur *stacking*. Algoritma *gradient boosting* modern, khususnya XGBoost, LightGBM, dan CatBoost, dikenal karena performa dan efisiensinya pada data yang kompleks. XGBoost (*Extreme Gradient Boosting*) menawarkan keunggulan kinerja, efisiensi waktu, dan memori, yang dioptimalkan 10 kali lebih cepat untuk *machine learning* skala besar [11]. Sementara itu, LightGBM (*Light Gradient Boost Machine*), metode *decision tree* yang lebih cepat dan terdistribusi, memiliki performa tinggi berkat strategi *gradient-based one-side sampling* (GOSS) dan *leaf-wise growth* [12]. Melengkapi ini, CatBoost unggul dalam menangani fitur kategorikal secara langsung tanpa memerlukan *prapemrosesan* tambahan [13]. Sementara itu yang berperan sebagai *meta-learner* adalah Random Forest yang unggul dalam mencegah *overfitting* dan menjaga stabilitas model melalui penggabungan banyak pohon keputusan [14]. Berdasarkan potensi superioritas tersebut, penulis akan menerapkan metode *ensemble stacking* untuk klasifikasi kualitas air.

1.2 Rumusan Masalah

Berdasarkan uraian latar belakang di atas, maka rumusan masalah dalam penelitian ini adalah sebagai berikut.

1. Bagaimana performa dasar dari masing-masing algoritma tunggal (XGBoost, LightGBM, dan CatBoost) sebelum diintegrasikan?
2. Bagaimana penerapan metode *Ensemble Stacking* dapat meningkatkan akurasi dan kinerja klasifikasi kualitas air dibandingkan dengan penggunaan model tunggal?
3. Seberapa banyak peningkatan metrik klasifikasi yang didapat dengan melakukan optimasi?

1.3 Batasan Masalah

Agar penelitian ini terarah dan tidak meluas dari fokus utama, maka batasan masalah dalam penelitian ini ditetapkan sebagai berikut.

1. Data yang digunakan diperoleh dari platform kaggle .
2. Algoritma yang digunakan dalam penelitian ini hanya XGBoost, LightGBM, dan CatBoost, yang berperan sebagai *base-learner*.
3. Algoritma Random Forest berperan sebagai *meta-learner*.
4. Proses penelitian ini didasarkan murni pada 21 fitur yang tersedia dalam dataset penelitian yang berisi parameter terkait pengukuran kualitas air.

1.4 Tujuan Penelitian

Tujuan yang akan dicapai dalam penelitian ini yaitu:

1. Mengetahui performa dasar dari masing-masing algoritma tunggal (XGBoost, LightGBM, dan CatBoost) sebelum diintegrasikan ke metode *Stacking*.
2. Menerapkan metode ensemble stacking terhadap XGBoost, LightGBM, dan CatBoost agar dapat meningkatkan akurasi dan kinerja klasifikasi kualitas air dibandingkan dengan penggunaan model tunggal.

3. Mengukur seberapa banyak peningkatan metrik klasifikasi yang didapat dengan melakukan optimasi.

1.5 Manfaat Penelitian

Berikut adalah penjabaran manfaat penelitian yang dapat dicapai.

1. Manfaat Teoritis

Penelitian ini memberikan kontribusi pada bidang *machine learning*, khususnya dalam studi komparatif *ensemble stacking*. Hasil penelitian akan menyajikan bukti keunggulan metode *ensemble stacking* dibandingkan model tunggal (XGBoost, LightGBM, CatBoost) pada kasus data klasifikasi kualitas air berskala besar.

2. Manfaat Praktis

Menyediakan model yang akurat untuk mengklasifikasikan kualitas air, yang berpotensi mengurangi biaya pengujian laboratorium konvensional yang mahal dan memakan waktu. Ini memungkinkan pemantauan yang lebih luas dan respon yang lebih cepat terhadap indikasi pencemaran.

1.6 Sistematika Penulisan

BAB I PENDAHULUAN

Bab ini berisi Latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan yang menggambarkan arah penulis dalam melakukan penelitian ini.

BAB II TINJAUAN PUSTAKA

Bab ini berisi tinjauan pustaka seperti penelitian terdahulu yang terkait dengan topik pembahasan penelitian ini, dasar-dasar teori yang digunakan dalam menyusun penelitian ini baik dari jurnal terkait maupun dari sumber internet.

BAB III METODE PENELITIAN

Bab ini berisi pembahasan mengenai metode penelitian yang dilakukan dalam penelitian ini. Meliputi alur penelitian, objek penelitian, preprocessing data, konfigurasi model, training model, optimasi metode, evaluasi akhir, hasil, alat, dan

bahan.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menjelaskan hasil dan tahapan yang penulis lakukan dalam mengembangkan model klasifikasi dengan metode ensemble stacking yang telah dibangun. Untuk itu, dilakukan analisis data, preprocessing, *base-learners* training, *meta-learner* training, optimasi akhir, evaluasi, dan visualisasi hasil.

BAB V PENUTUP

Bab ini menyajikan kesimpulan dari hasil optimasi metode dalam penelitian yang dilakukan, sekaligus memberikan saran untuk penelitian selanjutnya.

