

BAB I

PENDAHULUAN

1.1 Latar Belakang

Kualitas air merupakan salah satu aspek penting yang mempengaruhi kesehatan manusia dan keberlanjutan ekosistem. Air bersih dan higienis sangat diperlukan dalam kebutuhan sehari-hari, termasuk untuk konsumsi, pertanian, dan industri. Kondisi air yang baik untuk dikonsumsi harus memenuhi beberapa kriteria penting, baik secara fisik, kimia, dan mikrobiologi. Air bersih harus memiliki *pH* seimbang, tidak berwarna, tidak berbau, tidak berasa, serta terbebas dari mikrobiologi dan zat kimia berbahaya (Sahabuddin, 2010). Kriteria ini menjadi indikator utama untuk menilai kebersihan air.

Kualitas air di Indonesia sering kali terancam oleh pencemaran limbah domestik dan industri, serta perubahan iklim yang berdampak pada ketersediaan sumber daya air. Menurut Kementerian Lingkungan Hidup dan Kehutanan Republik Indonesia, Indeks Kualitas Air (IKA) Indonesia tahun 2023 sebesar 54,59 poin. Angka ini mengalami peningkatan dari tahun sebelumnya yaitu 53,88 poin, namun masih di bawah target nasional yang ditetapkan sebesar 55,03 poin. Pemantauan kualitas air di Indonesia dilakukan pada 812 titik oleh KLHK dan 5.157 titik oleh pemerintah daerah dengan hasil menunjukkan bahwa 18% titik mengalami perbaikan kualitas air, 67% tidak mengalami perubahan, dan 15% mengalami penurunan.

Penggunaan air tanah sebagai sumber utama air bersih di Provinsi DKI Jakarta semakin meningkat seiring dengan pertumbuhan populasi dan urbanisasi yang pesat. Namun, kualitas air sumur di Provinsi DKI Jakarta telah menjadi perhatian serius

karena tingginya tingkat pencemaran. Berdasarkan data Dinas Lingkungan Hidup Provinsi DKI Jakarta pada tahun 2023, menunjukkan bahwa 52% dari 266 titik pemantauan menunjukkan status indeks cemar berat, sementara 27% lainnya tercemar ringan (DLH, 2024). Masyarakat menghadapi risiko kesehatan yang signifikan akibat mengonsumsi air sumur yang tercemar. Pencemaran ini disebabkan oleh berbagai faktor, termasuk limbah domestik maupun industri, serta bakteri *E.coli* dari tangki septik.

Kondisi pencemaran air di Provinsi DKI Jakarta semakin memprihatinkan. Hampir seluruh sungai di wilayah ini, yaitu sekitar 72,7% mengalami pencemaran berat akibat limbah rumah tangga, sehingga klasifikasi data menjadi langkah penting untuk menganalisis dan memahami kualitas air. Klasifikasi adalah proses pengelompokan data ke dalam kategori tertentu berdasarkan karakteristik yang dimilikinya (Prasetyo, 2012). Klasifikasi dapat dilakukan menggunakan berbagai algoritma *machine learning*, seperti *Random Forest* dan *K-Nearest Neighbor*. Keduanya merupakan algoritma yang populer digunakan untuk proses klasifikasi dengan keunggulan dan cara kerja yang berbeda.

Random Forest adalah metode *ensemble* yang menggabungkan beberapa pohon keputusan (*Decision Tree*) untuk meningkatkan akurasi dan stabilitas prediksi (Breiman & Cutler, 2001). Setiap pohon dalam *Random Forest* dibangun menggunakan sampel acak dari dataset pelatihan dan subset acak dari fitur, sehingga menghasilkan model yang tahan terhadap *overfitting* dan dapat menangani data tidak seimbang dengan baik. Algoritma *K-Nearest Neighbor* merupakan algoritma yang berbasis pada kedekatan, yakni kelas dari data baru ditentukan berdasarkan mayoritas kelas dari k jumlah tetangga terdekatnya (Hancock dkk., 2014). KNN mudah

diimplementasikan dan efektif untuk dataset kecil, namun menjadi lambat pada dataset besar dikarenakan memerlukan perhitungan jarak ke semua titik data.

Penggunaan algoritma *Random Forest* untuk klasifikasi kualitas air sumur telah menunjukkan hasil yang signifikan. Penelitian oleh Mutoffar dan Fadillah (2022) bertujuan untuk menganalisis dan mengklasifikasikan kualitas air sumur di Jakarta menggunakan algoritma *Random Forest* dengan presisi sebesar 82,3% dan sensitivitas sebesar 83% ini menunjukkan bahwa algoritma dapat memprediksi kondisi air sumur dengan baik. Model berhasil mengklasifikasikan 40 data sebagai air yang dapat dikonsumsi dan 4 data sebagai air yang tidak dapat dikonsumsi, sesuai dengan target klasifikasi.

Algoritma *K-Nearest Neighbor* terbukti lebih unggul dibandingkan *Support Vector Machine* dalam hal akurasi untuk klasifikasi kelayakan air minum berdasarkan penelitian yang dilakukan oleh Ulum, dkk (2023). Pengujian parameter optimal KNN dilakukan menggunakan metode *k-fold* menghasilkan akurasi rata-rata sebesar 88,94%, sedangkan pada algoritma *SVM* memiliki akurasi rata-rata sebesar 87,71%. Optimasi parameter yang dilakukan tersebut tidak hanya meningkatkan kinerja model tetapi juga memberikan pemahaman yang lebih baik tentang struktur dan pola dalam dataset yang digunakan.

Kondisi ketidakseimbangan kelas dalam dataset sering kali menjadi tantangan dalam analisis data, terutama pada proses klasifikasi. Jika satu kelas memiliki jumlah data yang lebih dominan dibandingkan kelas lainnya, maka model klasifikasi yang dihasilkan akan mengalami kesulitan dalam memprediksi kelas minoritas. Teknik *Synthetic Minority Over-sampling Technique (SMOTE)* menjadi solusi untuk mengatasi masalah ini. *SMOTE* bekerja dengan cara menciptakan sampel sintetis

untuk kelas minoritas, sehingga meningkatkan jumlah data dan membantu model belajar lebih baik dari variasi yang ada (Chawla dkk., 2002).

Siringoringo (2018) dalam penelitiannya menerapkan algoritma *SMOTE* untuk menangani ketidakseimbangan pada dataset *credit card fraud*. Optimasi parameter menggunakan *10-fold cross-validation* menunjukkan bahwa penerapan *SMOTE* berhasil meningkatkan nilai *G-Mean* menjadi 81,0% dan *F-Measure* menjadi 81,8%. *SMOTE* efektif dalam meningkatkan performa klasifikasi dengan algoritma KNN. Penelitian lain dilakukan oleh Erlin, dkk (2022) menunjukkan bahwa penerapan *SMOTE* dapat mengurangi *overfitting* dan meningkatkan kinerja model *Random Forest*. Nilai *G-Mean* dan *F1-score* mengalami peningkatan menjadi 80% dan 85%. Penelitian ini menegaskan bahwa *SMOTE* dapat meningkatkan kinerja model dalam memprediksi penyakit jantung, serta dapat diterapkan pada dataset lain dengan kondisi ketidakseimbangan data.

Penulis akan melakukan penelitian klasifikasi terhadap kualitas air sumur di Provinsi DKI Jakarta dengan menerapkan metode *Random Forest* dan KNN serta penanganan data tidak seimbang menggunakan teknik *SMOTE*, dengan judul “Penerapan *Synthetic Minority Over-Sampling Technique* pada Klasifikasi Kualitas Air Sumur di Provinsi DKI Jakarta Menggunakan Metode *Random Forest* dan *K-Nearest Neighbor*”. Penelitian ini diharapkan mampu menghasilkan metode yang tepat dalam mengklasifikasikan kualitas air sumur dengan tingkat akurasi yang tinggi, sehingga mampu membantu mengidentifikasi dan penanganan masalah kualitas air bagi kesehatan masyarakat.

1.2 Rumusan Masalah

Rumusan masalah berdasarkan latar belakang tersebut adalah sebagai berikut:

1. Bagaimana proses klasifikasi kualitas air sumur di Provinsi DKI Jakarta menggunakan metode *Random Forest* dan KNN?
2. Bagaimana proses penanganan data tidak seimbang dengan teknik SMOTE pada klasifikasi kualitas air sumur di Provinsi DKI Jakarta menggunakan metode *Random Forest* dan KNN?
3. Bagaimana perbandingan kebaikan model klasifikasi antara metode *Random Forest* dan KNN untuk klasifikasi kualitas air sumur di Provinsi DKI Jakarta?

1.3 Batasan Masalah

Batasan masalah yang digunakan pada penelitian tugas akhir ini adalah data sekunder yang diperoleh dari publikasi kualitas air sumur di Provinsi DKI Jakarta pada tahun 2023. Variabel yang digunakan meliputi nitrat (kandungan nitrogen dalam NO_3), nitrit (NO_2), krom heksavalen ($Cr(VI)$), besi (Fe), mangan (Mn), *total coliform* dan *E.Coli*. Label kategori kualitas air sumur yang digunakan sebagai kelas klasifikasi yaitu kondisi baik, tercemar ringan, sedang, dan berat.

1.4 Tujuan Penelitian

Tujuan dilakukan penelitian ini adalah sebagai berikut:

1. Menerapkan klasifikasi pada kualitas air sumur di Provinsi DKI Jakarta menggunakan metode *Random Forest* dan KNN kemudian menganalisis berdasarkan nilai *accuracy*, *precision*, *recall*, *F1-Score*, dan *balanced accuracy*.

2. Menerapkan penanganan data tidak seimbang dengan teknik SMOTE pada klasifikasi kualitas air sumur di Provinsi DKI Jakarta menggunakan metode *Random Forest* dan KNN kemudian menganalisis berdasarkan nilai *accuracy*, *precision*, *recall*, *F1-Score*, dan *balanced accuracy*.
3. Membandingkan kebaikan model klasifikasi antara metode *Random Forest* dan KNN yang digunakan dalam mengklasifikasikan kualitas air sumur di Provinsi DKI Jakarta.