# BAB I
# PENDAHULUAN

## 1.1 Latar Belakang

Saat ini, penggunaan kendaraan bermotor, baik roda dua maupun roda empat, terus meningkat seiring dengan pesatnya pertumbuhan ekonomi di Jakarta. Keadaan ini berdampak serius terhadap kualitas udara, karena emisi kendaraan bermotor merupakan salah satu penyebab utama pencemaran udara. Selain berdampak terhadap lingkungan, polusi udara juga memberikan dampak negatif terhadap kesehatan masyarakat, seperti meningkatkan risiko penyakit pernafasan [1].

Sebuah studi mendalam yang dilakukan oleh *United Nations Environment Programme* (UNEP) pada tahun 2021 menemukan masalah serius terkait kualitas udara di kota-kota besar di seluruh dunia. Sekitar 90% populasi kehidupan di kota-kota besar di Amerika, Eropa, Asia, dan Afrika terkena dampak dari kualitas udara yang tidak memenuhi standar kesehatan. *World Health Organization* (WHO) telah menetapkan standar yang aman untuk kualitas udara berdasarkan konsentrasi partikel halus (PM2.5), partikel kasar (PM10), oksida nitrogen, oksida sulfur, ozon troposfer, dan berbagai polutan lainnya. Sayangnya, hampir sembilan dari sepuluh kota besar tersebut, setidaknya satu jenis polutan melewati batas aman yang ditetapkan oleh WHO [2].

Berdasarkan data yang dikumpulkan oleh *IQAir*, pada bulan Oktober 2021, Jakarta menempati posisi kesembilan dalam hal kualitas udara dan polusi. Selain itu, dalam daftar 106 negara dengan tingkat polusi tertinggi di dunia pada tahun 2020, Indonesia menduduki peringkat kesembilan dalam konsentrasi PM2.5 [3]. Oleh karena itu, penting untuk melakukan analisis dan mengklasifikasikan kualitas udara di DKI Jakarta secara efisien untuk mengurangi dampak negatif terhadap kesehatan masyarakat dan lingkungan. Dalam analisis data, terdapat berbagai jenis algoritma yang dapat digunakan, masing-masing algoritma memiliki keunggulan tersendiri ketika melakukan analisis.

Dalam analisis kualitas udara, penggunaan teknologi *machine learning* sangat penting untuk memberikan pemahaman yang lebih baik tentang pola pencemaran dan potensi risikonya. Algoritma seperti *machine learning* dapat memprediksi kategori kualitas udara berdasarkan jumlah polutan dan waktu pengukuran dengan menggunakan data yang tersedia. Hal ini memungkinkan masyarakat dan pemerintah mengambil tindakan yang lebih cepat dan tepat untuk mengurangi efek negatif polusi udara [4].

Klasifikasi memberikan gambaran umum tentang kualitas udara dan dapat menjadi alat penting dalam perencanaan kebijakan lingkungan. Misalnya, data yang dihasilkan dapat digunakan untuk menentukan daerah dengan tingkat polusi tinggi dan merencanakan strategi untuk mengurangi polusi, seperti membatasi kendaraan bermotor atau menambah ruang hijau [4][5]. Selain itu, pengumpulan data kualitas udara menjadi lebih mudah dan akurat berkat peningkatan teknologi sensor. Sepanjang hari, atau bahkan sepanjang tahun, data *real-time* tentang fluktuasi kualitas udara dapat dikumpulkan. Masyarakat dapat menggunakan informasi ini untuk mengubah hal-hal, seperti mengurangi waktu di luar saat kondisi sedang buruk [4].

Sayangnya di dalam proses pengolahan data terdapat *imbalance data* dan *missing value*. Masalah *imbalance data* bisa berdampak buruk pada proses klasifikasi karena model akan cenderung memihak pada kelas yang lebih banyak. Untuk mengatasi masalah ini, model *resampling* diimplementasikan dengan menggunakan teknik SMOTE (*Synthetic Minority Oversampling Technique*). *SMOTE* digunakan untuk membuat sampel sintesis dari kelas minoritas dalam dataset [6] Sedangkan masalah *Missing Value* dalam dataset dapat mengurangi jumlah data yang digunakan untuk proses prediksi, sehingga dapat menyebabkan hasil prediksi menjadi kurang akurat [7].

Memahami kelebihan dan kekurangan masing-masing algoritma yang dipilih dalam penelitian ini sangat penting. Dengan demikian, algoritma yang dipilih tidak hanya memberikan hasil yang akurat, tetapi juga efisien dalam hal waktu komputasi dan penggunaan sumber daya. Ini sangat relevan untuk situasi

kehidupan nyata di mana kecepatan dan akurasi sering kali lebih penting daripada yang lain.

*Random Forest* adalah sebuah algoritma machine learning secara umum digunakan untuk mengklasifikasi kumpulan data yang besar memiliki fungsi yang dapat digunakan dalam banyak dimensi dengan skala yang berbeda dan memiliki kinerja tinggi [8]. Sedangkan *Support Vector Machine* (SVM) adalah sistem pembelajaran yang menggunakan algoritma pembelajaran berbasis hipotesis yang didasarkan pada teori optimasi [9]. *K-Nearest Neighbor* merupakan bagian dari algoritma supervised learning yang menggunakan data dengan jarak terdekat untuk mengklasifikasikan item data baru [29]. *Naïve Bayes* merupakah salah satu algoritma data mining yang menerapkan teori bayes dalam proses klasifikasi [30].

Penelitian ini akan berfokus pada perbandingan algoritma *K-Nearest Neighbors, Random Forest, Support Vector Machine* dan *Naïve Bayes*. Tujuan utama dari penelitian ini adalah untuk melakukan perbandingan algoritma yang paling efektif dan untuk mendapatkan kinerja yang optimal.

## 1.2 Rumusan Masalah

1. Apakah penggunaan metode imputasi Fillna dalam mengatasi *missing value* data kualitas udara Jakarta dapat meningkatkan kinerja algoritma *machine learning klasifikasi*?
2. Apakah penggunaan metode *resampling* SMOTE dalam mengatasi ketidakseimbangan data kualitas udara Jakarta dapat meningkatkan kinerja algoritma *machine learning* klasifikasi?
3. Algoritma *machine learning* klasifikasi mana yang memiliki performa terbaik pada klasifikasi kualitas udara DKI Jakarta?

## 1.3 Batasan Masalah

1. Penelitian ini hanya menggunakan data kualitas udara Jakarta yang tersedia secara online.
2. Algoritma yang dibandingkan terbatas pada *K-Nearest Neighbors, Random Forest, Support Vector Machine,* dan *Naïve Bayes.*

3. Penelitian menggunakan teknik SMOTE untuk menangani *imbalance data* dan metode Fillna untuk menangani *missing value*.

## 1.4 Tujuan Penelitian

1. Mengetahui permasalahan dan memberikan solusi mengenai *missing value*.
2. Mengetahui permasalahan dan memberikan solusi mengenai *imbalance data*.
3. Mengetahui algoritma machine learning mana yang memiliki performa terbaik pada klasifikasi kualitas udara DKI Jakarta.

## 1.5 Manfaat Penelitian

Penelitian ini dapat memberikan pemahaman yang lebih baik tentang efektivitas algoritma *K-Nearest Neighbors*, *Random Forest*, SVM, dan *Naïve Bayes* dalam klasifikasi kualitas udara perkotaan.

1. Sebagai referensi dan sumber pembelajaran untuk memahami klasifikasi kualitas udara dan dampaknya terhadap kesehatan.
2. Menjadi dasar untuk penelitian lanjutan yang membahas klasifikasi kualitas udara dengan menggunakan metode seperti *Regresi Logistik*.
3. Berdasarkan hasil klasifikasi kualitas udara, dapat memberikan masukan bagi kebijakan pengendali polusi udara di Jakarta.

## 1.6 Sistematika Penulisan

BAB I PENDAHULUAN: Menguraikan latar belakang masalah yang menjelaskan alasan dan pentingnya penelitian ini dilakukan, rumusan masalah yang memfokuskan pada pertanyaan utama yang ingin dijawab, batasan masalah yang menentukan ruang lingkup penelitian, tujuan penelitian yang menggambarkan hasil yang ingin dicapai, manfaat penelitian yang menjelaskan kontribusi penelitian ini bagi berbagai pihak, dan sistematika penulisan yang memberikan Gambaran singkat tentang isi setiap bab.

BAB II TINJAUAN PUSTAKA: Menyajikan literatur dan teori yang relevan dengan penelitian ini, mencakup hasil penelitian terdahulu dan konsep-

konsep dasar yang digunakan, seperti big data, data mining, serta algoritma *Random Forest, Support Vector Machine, K-Nearest Neighbor, dan Naïve Bayes*.

BAB III METODE PENELITIAN, Menjelaskan metodologi yang diterapkan dalam penelitian, mulai dari objek penelitian, alur penelitian yang meliputi tahap-tahapan yang dilalui, teknik pengumpulan data, *preprocessing* data yang mencakup penanganan *missing value* dan *imbalance data*, hingga pemodelan algoritma dan evaluasi menggunakan *matrix* seperti *accuracy*, *precision*, *recall*, dan *f1-score*.

BAB IV HASIL DAN PEMBAHASAN: Memaparkan hasil penelitian secara rinci, termasuk deskripsi dataset, proses *preprocessing* data, pemodelan algoritma, dan hasil evaluasi model.

BAB V PENUTUP: Berisi Kesimpulan yang menyajikan temuan utama dari penelitian dan saran untuk penelitian selanjutnya, yang mungkin melibatkan eksplorasi lebih lanjut tentang algoritma atau dataset yang beragam.