

BAB I

PENDAHULUAN

1.1 Latar Belakang

Udara adalah gabungan beberapa gas yang mengelilingi atmosfer bumi (Norrahan dkk., 2023). Kualitas udara sangatlah perlu diperhatikan karena semua makhluk hidup, termasuk manusia, membutuhkan udara untuk hidup. Hal ini membuat penting bagi manusia untuk memiliki udara yang bersih dan bebas dari polusi agar kegiatan sehari-hari dapat berjalan dengan efisien (Sartika, E., 2023). Berbagai masalah dalam kehidupan sehari-hari yang membuat aktivitas menjadi tidak nyaman dapat ditimbulkan oleh kualitas udara yang buruk.

Berbagai aktivitas manusia dapat menyebabkan kualitas udara di bumi menjadi tercemar. Sektor transportasi adalah sektor yang menyumbang polusi udara terbesar di Indonesia, yakni mencapai 60%-70% kontribusinya. Alasan sektor transportasi berkontribusi besar pada polusi udara adalah karena penggunaan transportasi yang tak terhindarkan untuk mobilisasi penduduk di Indonesia yang jumlahnya sekitar 278 juta jiwa; ketiga terbanyak di Asia. Selain sektor transportasi, terdapat sektor pabrik yang juga menyumbang polusi di Indonesia, yaitu sekitar 25% kontribusinya (Saputra dkk., 2023). Hasil pemantauan IQAir sebagai platform informasi kualitas udara *real-time* terbesar di dunia, Jakarta menempati peringkat ke-12 sebagai ibu kota paling berpolusi di dunia (IQAir, 2021). Melalui aktivitas sehari-hari, manusia sering mencemari udara melalui aktivitasnya padahal dampak dari pencemaran udara sangat merugikan manusia itu sendiri.

Manusia, hewan, tumbuhan, bahkan material yang ada di alam dapat terkena dampak dari udara yang tercemar (Yasir, 2021). Salah satu contohnya adalah

dampak pada manusia, dengan berbagai penyakit dapat ditimbulkan ketika terpapar udara yang sudah tercemar. Penyakit yang disebabkan oleh pencemaran udara di antaranya adalah gangguan pernapasan seperti asma, bronkitis, dan pneumonia; meningkatnya risiko terkena penyakit jantung dan stroke; gangguan perkembangan janin; penyakit kronis; dan penurunan fungsi paru-paru (Maharani dan Aryanta, 2023). Hal itu membuktikan bahwa dampak negatif dari pencemaran udara akan merugikan manusia dan makhluk hidup lainnya.

Saat ini, pemerintah melaksanakan program sesuai Peraturan Menteri Lingkungan Hidup dan Kehutanan Republik Indonesia Nomor 14 Tahun 2020, yang mengamanatkan menteri, gubernur, dan bupati/wali kota wajib menyediakan informasi publik mengenai hasil penentuan indeks standar pencemar udara (ISPU). Informasi yang disediakan oleh pemerintah, yaitu partikulat PM_{10} dan $PM_{2.5}$, NO_2 , SO_2 , CO , O_3 , serta kategori kualitas udara. Berhubungan dengan amanat yang diberikan pada wakil rakyat tersebut, masyarakat Indonesia dapat mencari informasi tersebut untuk mengurangi risiko menghirup udara berkualitas rendah yang berdampak buruk bagi kesehatan. Penentuan kategori kualitas udara ini dapat dilakukan dengan mengklasifikasikan parameter-parameter yang telah diukur sebelumnya.

Zaman yang canggih ini menawarkan satu metode untuk mengklasifikasikan parameter-parameter kualitas udara, yakni dengan menggunakan algoritma *classification machine learning*. Klasifikasi merupakan suatu proses evaluasi data yang bertujuan untuk mengkategorikannya ke dalam kelas yang telah terdefinisi sebelumnya (Paramitha dkk., 2023). Klasifikasi parameter-parameter kualitas udara dapat dilakukan dengan menerapkan algoritma *machine learning*. *Machine*

learning sendiri dapat diartikan sebagai metode yang bertujuan untuk meningkatkan hasil prediksi berdasarkan data yang dimiliki dengan menggunakan komputer (Hardadi dan Bunyamin, 2023).

Model yang tersedia untuk melakukan klasifikasi menggunakan *machine learning* sangatlah beragam dengan kelebihan dan kekurangan yang dimilikinya. Pada *classification machine learning*, terdapat *base learners* dan *ensemble learners* yang merupakan jenis model klasifikasi *machine learning*. *Base learners* dapat didefinisikan sebagai *learning* yang kinerjanya lebih baik daripada hanya menebak-nebak secara acak (Tugay dan Oguducu, 2020). *Ensemble learning* merupakan gabungan dari beberapa model *base learner* guna memperkuat akurasi yang akan dihasilkan (Cendani dan Wibowo, 2022). Contoh model klasifikasi *machine learning base learner* adalah *decision tree*, *naive bayes classifier*, dan *k-nearest neighbor*, sedangkan contoh dari *ensemble model* adalah *random forest*, *adaboost*, dan *gradient boosting*.

Decision tree merupakan *base learner* yang memiliki beberapa algoritma, di antaranya adalah CART, ID3, dan C4.5. Algoritma CART merupakan metode non-parametrik yang berguna untuk mendapatkan suatu kelompok data yang akurat sebagai penciri dari suatu pengklasifikasian. CART memiliki kelebihan untuk mencakup penyeimbangan kelas otomatis dan penanganan nilai hilang otomatis (Wu dan Kumar, 2009). Kelebihan lainnya dari metode dan algoritma ini juga dianggap mudah untuk diinterpretasikan, cepat, dan akurat (Pratiwi dan Zain, 2014).

Random forest merupakan salah satu *ensemble model* yang merupakan *bagging* dari model *decision tree* khususnya CART. Metode ini memiliki

keunggulan yang menarik dibandingkan dengan *decision tree* dengan algoritma CART, yaitu mampu mengurangi risiko *overfitting*. *Random forest* juga mampu menghasilkan prediksi yang akurat dengan memperkirakan nilai yang hilang (*missing value*). Kemampuan *random forest* tersebut diakibatkan oleh struktur metode yang terdiri dari sejumlah *decision tree* yang tentunya akan sangat bervariasi.

Nugroho (2023) melakukan penelitian terhadap tingkat kualitas udara DKI Jakarta menggunakan algoritma *random forest* dan mendapatkan *accuracy* tertinggi sebesar 90% (Nugroho dkk., 2023). Penelitian tersebut tidak menggunakan GridSearchCV yang dapat mencegah *overfitting*. Penelitian tersebut menggunakan data dari Februari 2021 hingga Oktober 2021 dan data yang tersedia hanya menggunakan dua kategori, yaitu kategori sedang dan kategori tidak sehat.

Widartama dan Fansyuri (2023) melakukan penelitian perbandingan antara algoritma CART dengan *support vector machine* pada klasifikasi kualitas udara di DKI Jakarta (Widartama dan Fansyuri, 2023). Data yang digunakan pada penelitian tersebut menggunakan data sejak 1 Januari 2021 hingga 31 Januari 2021. Pada penelitian tersebut, didapatkan akurasi pada CART sebesar 99.67% dan 95.05% pada SVM. Pada penelitian tersebut tidak menggunakan sesuatu yang memastikan bahwa model tersebut tidak terdapat *overfitting*.

Sang dkk. (2021) juga melakukan penelitian terhadap indeks standar pencemar udara di DKI Jakarta dengan menggunakan algoritma *decision tree* dan *support vector machine*. Data yang digunakan pada penelitian tersebut menggunakan data pada tahun 2020. Variabel yang digunakan pada penelitian tersebut adalah PM_{10} dan $PM_{2.5}$, NO_2 , SO_2 , CO , O_3 . Pada penelitian tersebut,

didapatkan bahwa *decision tree* mendapatkan hasil yang paling baik, yaitu mendapatkan *accuracy* sebesar 99.40% sedangkan SVM mendapatkan 94.93% (Sang dkk., 2021).

Penelitian ini membahas mengenai perbandingan antara algoritma CART dan algoritma *random forest* dalam konteks klasifikasi kategori indeks standar pencemar udara di DKI Jakarta dengan kategori kualitas udara yang lebih dari dua kategori atau disebut juga dengan multikelas. Variabel yang digunakan pada penelitian ini adalah PM₁₀, NO₂, O₃, SO₂, dan CO. Pada penelitian kali ini juga digunakan GridSearchCV guna mengurangi *overfitting* yang kemungkinan terjadi dan mencari parameter yang optimal sehingga dapat menghasilkan *accuracy* yang sebaik mungkin. Penelitian ini diharapkan dapat memberikan metode alternatif yang akurat dalam mengklasifikasikan indeks standar pencemar udara (ISPU) dengan menggunakan *random forest* dan algoritma *classification and regression tree* yang merupakan salah satu algoritma dari *decision tree* sehingga dapat membantu pihak terkait dalam memantau dan mengantisipasi tingkat polusi udara secara lebih efektif.

1.2 Rumusan Masalah

Dengan merujuk pada konteks yang telah dijelaskan sebelumnya, perumusan permasalahan dalam penelitian ini adalah sebagai berikut:

1. Bagaimana kinerja dari *classification and regression tree* dengan GridSearchCV berdasarkan nilai *accuracy* pada kasus klasifikasi kategori indeks standar pencemar udara (ISPU) di Provinsi DKI Jakarta?

2. Bagaimana kinerja dari *random forest* dengan GridSearchCV berdasarkan nilai *accuracy* pada kasus klasifikasi kategori indeks standar pencemar udara (ISPU) di Provinsi DKI Jakarta?
3. Bagaimana perbandingan kinerja antara *classification and regression tree* dan *random forest* dengan masing-masing terdapat GridSearchCV berdasarkan nilai *accuracy* dalam mengklasifikasi kategori indeks standar pencemar udara (ISPU) di Provinsi DKI Jakarta?

1.3 Batasan Masalah

Pembahasan pada Subbab 1.1 sebelumnya mengarah pada munculnya beberapa masalah yang dirumuskan sebagai berikut:

1. Data yang digunakan terdiri dari data indeks standar pencemar udara (ISPU) di Provinsi DKI Jakarta pada rentang waktu Januari 2010 hingga Desember 2021.
2. Variabel prediktor yang digunakan di penelitian ini merupakan jenis polutan dan lokasi stasiun itu sendiri, yaitu karbon monoksida (CO), sulfur dioksida (SO₂), nitrogen dioksida (NO₂), ozon permukaan (O₃), dan partikel debu (PM₁₀). Variabel target yang digunakan adalah kategori kualitas udara.
3. Ketepatan klasifikasi terbatas oleh hasil kinerja algoritma yang sudah ditentukan, yaitu nilai *accuracy*.

1.4 Tujuan Penelitian

Tujuan yang ingin dicapai pada penelitian ini adalah:

1. Mengimplementasikan algoritma CART dan *random forest* dengan menggunakan GridSearchCV pada kasus klasifikasi kategori indeks standar pencemar udara di Provinsi DKI Jakarta.

2. Membandingkan evaluasi kinerja *accuracy* antara model CART dan *random forest* dengan masing-masing disertai penambahan GridSearchCV.