BAB I

PENDAHULUAN

1.1 Latar Belakang

Udara merupakan campuran gas yang terdapat pada lapisan yang mengelilingi bumi (atmosfer) yang terdiri dari banyak komponen seperti gas, partikel padat, partikel cair, energi, ion, dan zat organik yang terdistribusi secara bebas mengikuti volume bentuk ruang (Cahyono, 2017). Udara termasuk ke dalam hal vital yang berperan penting bagi kelangsungan hidup makhluk hidup. Udara bersih yang dibutuhkan untuk kehidupan di bumi harus berupa gas yang tidak berbau, terasa segar, sejuk, dan ringan saat dihirup (Siburian, 2020).

Manusia saat ini semakin sulit mendapatkan udara bersih, khususnya di daerah yang memiliki banyak industri. Padahal, kebutuhan udara bersih semakin diperlukan seiring meningkatnya jumlah penduduk di dunia. Namun, saat ini kualitas udara semakin menurun akibat aktivitas manusia. Berdasarkan data Kementerian Lingkungan Hidup dan Kehutanan Republik Indonesia, penyumbang terbesar pencemaran udara di Indonesia berasal dari sektor transportasi sebesar 44% disusul 34% disebabkan oleh Pembangkit Listrik Tenaga Uap (PLTU).

IQAir sebagai *platform* penyedia informasi kualitas udara *real-time* terbesar dunia melaporkan pada Maret 2022, Indonesia menduduki peringkat ke-17 sebagai negara dengan tingkat pencemar udara tertinggi di dunia dengan konsentrasi *Particulate Matter* (PM) 2,5 mencapai angka 34,3μg per meter kubik. Hal itu menandakan bahwa Indonesia menduduki peringkat teratas negara penghasil polusi udara di kawasan Asia Tenggara. Laporan kualitas udara dunia yang dirilis IQAir

menempatkan Jakarta sebagai peringkat ke 7 dari 10 ibu kota dengan tingkat pencemaran polusi udara tertinggi di dunia pada 2023. Dalam data terbaru pada September 2024, indeks standar pencemar udara di wilayah Jakarta mencapai angka 160 dan termasuk ke kategori tidak sehat dengan konsentrasi partikel halus PM 2,5 mencapai 13,7 kali dari nilai panduan kualitas udara tahunan dari *World Health Organization* (WHO).

Kondisi tersebut tentunya akan menimbulkan risiko besar bagi kesehatan manusia. Badan Riset dan Inovasi Nasional (BRIN) menyebutkan bahwa terdapat lima besar penyakit yang terjadi akibat polusi udara, diantaranya yaitu stroke, penyakit jantung iskemik (*ischemic hearth disease*), diabetes melitus, penyakit paru obstruktif kronis (*chronic obstructive pulmonary disease*/COPD), dan *neonatal disorders*. Selain berdampak terhadap kesehatan manusia, polusi udara juga akan berdampak buruk terhadap lingkungan. Polutan yang terlepas ke atmosfer seperti Karbon dioksida (CO_2), Metana (CH_4), dan Dinitrogen oksida (N_2O) menjadi kontributor terhadap pemanasan global dan perubahan iklim yang dapat menganggu ekosistem dan keanekaragaman hayati.

Pemerintah melalui Pasal 20 Peraturan Pemerintah RI Nomor 41 Tahun 1999 tentang pengendalian pencemaran udara melakukan upaya pencegahan pencemaran udara dengan kebijakan menetapkan baku mutu udara ambien, baku mutu emisi sumber tidak bergerak, baku tingkat gangguan, ambang batas emisi gas buang dan kebisingan kendaraan bermotor. Dalam Peraturan Menteri Lingkungan Hidup dan Kehutanan RI Nomor 14 Tahun 2020 tentang Indeks Standar Pencemar Udara disebutkan bahwa pemerintah daerah provinsi maupun kabupaten/kota wajib menyediakan informasi publik mengenai hasil penentuan Indeks Standar Pencemar

Udara (ISPU) kepada masyarakat, dengan parameter ISPU yaitu partikulat PM_{10}, partikulat $PM_{2,5}$, Nitrogen dioksida (NO_2), Ozon (O_3), Sulfur dioksida (SO_2), Hidrokarbon (HC), dan Karbon monoksida (CO). Penentuan kategori kualitas udara dapat dilakukan dengan mengklasifikasikan parameter-parameter yang telah diukur sebelumnya.

Banyak riset telah dilakukan untuk melakukan pengklasifikasian terkait parameter kualitas udara, salah satunya dengan metode yang popular yaitu *Machine Learning. Machine Learning* (ML) adalah cabang dari kecerdasan buatan atau *Artificial Intelligence* (AI) yang membantu komputer atau mesin pengajaran belajar dari semua data sebelumnya dan membuat keputusan yang cerdas (Fathurohman *et al*., 2021). *Machine learning* juga memungkinkan komputer untuk melakukan klasifikasi atau prediksi berdasarkan informasi yang diperoleh (Russel & Norvig, 2010).

Klasifikasi merupakan metode statistika yang digunakan untuk mengetahui atau memperkirakan kelas dari suatu objek berdasarkan atribut yang ada. Dalam *Machine Learning,* klasifikasi termasuk ke dalam teknik dasar belajar *supervised learning* (pembelajaran terarah) yang dapat memperoleh informasi atau prediksi yang membutuhkan data berlabel untuk melakukan pembelajaran pada mesin sehingga mesin mampu mengidentifikasi label masukan (*input*) untuk dilanjutkan prediksi maupun klasifikasi. Keakuratan prediksi yang dihasilkan akan menunjukkan performa dari setiap metode klasifikasi yang digunakan. Permasalahan yang terkadang ditemui dalam metode klasifikasi yaitu masalah ketidakseimbangan data (*imbalance* data) (Fitriani *et al.,* 2021).

Data *imbalance* merupakan kondisi suatu kelas memiliki cacah data yang jauh berbeda dibandingkan dengan kelas lainnya. Kelas dengan cacah data lebih banyak disebut dengan *majority class* dan kelas yang memiliki cacah data lebih sedikit disebut *minority class* (Barro *et al*., 2013). Kondisi tersebut akan berpengaruh terhadap klasifikasi data yang digunakan untuk menentukan kelas suatu data. Jika kondisi data tidak seimbang, maka kelas data cenderung tidak stabil karena data akan lebih condong ke kelas data yang memiliki cacah data lebih besar. Ketidakseimbangan data tersebut akan menghasilkan nilai keakuratan yang sangat tinggi untuk kelas mayoritas dan nilai keakuratan sangat rendah pada kelas minoritas.

Penanganan data *imbalance* perlu dilakukan agar model yang dihasilkan dapat bekerja dengan baik dan lebih akurat dalam melakukan prediksi dari data yang digunakan. Upaya yang dapat dilakukan ialah melakukan pendekatan level data dengan cara melakukan *resampling* data. Terdapat dua jenis teknik *resampling*, yaitu menaikkan jumlah sampel pada kelas minoritas (*oversampling*) dan menurunkan jumlah sampel pada kelas mayoritas (*undersampling*). *Oversampling* lebih banyak diterapkan karena jumlah data yang digunakan tidak terlalu banyak. Penerapan metode *oversampling* yang berlebihan dapat menyebabkan terjadinya *overfitting* (Kaur & Gosain, 2018).

Penanganan ketidakseimbangan kelas dapat menggunakan teknik *resampling*, salah satunya adalah *Random Oversampling* (ROS). Teknik *oversampling* dipilih karena tidak mengurangi jumlah data, akan tetapi menambah dataset yang kurang pada kelas minoritas. ROS bekerja dengan menggandakan atau menambahkan salinan dari sampel-sampel pada kelas minoritas hingga proporsinya

seimbang dengan kelas mayoritas. Penelitian yang dilakukan Aryanti *et al.,* (2023) menunjukkan peningkatan akurasi yang dihasilkan sebesar 81,86% dengan penggunaan ROS yang lebih tinggi dibandingkan akurasi tanpa penggunaan ROS sebesar 80,77%.

Pemilihan metode klasifikasi perlu disesuaikan dengan jenis data yang digunakan. Data yang digunakan dalam penelitian ini bersifat multikelas atau lebih dari dua kategori pada variabel terikatnya yang dinotasikan dalam angka 0, 1, 2, dan seterusnya. Algoritma yang digunakan dalam melakukan klasifikasi dalam penelitian tugas akhir ini adalah *Gradient Boosting* dan *Adaptive Boosting (AdaBoost).*

Gradient boosting merupakan algoritma pembelajaran mesin yang dapat menangani data kompleks dan tidak seimbang (Friedman, 2001). *Gradient boosting* juga diartikan sebagai klasifikasi *machine learning* dengan menggunakan teknik *ensemble* dari *decision tree* yang dapat menyelesaikan persoalan klasifikasi dan prediksi data (Ismanto & Novalia, 2021). Cara kerja *gradient boosting* dimulai dengan membangun pohon klasifikasi awal dilanjutkan iterasi secara berulang dan memperbaiki pohon klasifikasi sebelumnya dengan membuat model baru.

Gradient boosting telah banyak digunakan sebagai algoritma dalam sebuah penelitian. Penelitian yang dilakukan Suryana (2021) menggunakan metode *gradient boosting* dengan optimasi *hyperparameter* untuk memprediksi keberhasilan telemarketing bank menghasilkan akurasi yang tinggi mencapai angka 90,39%. Penelitian Abarna *et al* (2020) dengan *gradient boosting* juga menghasilkan hasil akurasi yang tinggi saat melakukan klasifikasi penyakit alzheimer pada dataset OASIS-1 dengan akurasi yang dihasilkan sebesar 97,22%.

Adaptive boosting (*AdaBoost*) merupakan algoritma *boosting* yang dikemukakan oleh Yoav Freund & Robert Schapire pada tahun 1995 (Bauer & Kohavi, 1999). *AdaBoost* adalah algoritma pembelajaran terintegrasi yang bertujuan untuk memperoleh dan menyesuaikan bobot semua sampel penelitian melalui pelatihan model secara berulang. *AdaBoost* dapat meningkatkan kelompok klasifikasi yang salah diprediksi dengan memberikan bobot sampel sesuai dengan hasil klasifikasi dasar (Sanjaya *et al.,* 2020).

Penelitian Putri (2022) untuk dataset Faisalabad *hearth disease institute* menggunakan hyperparameter *RandomSearchCV* dengan algoritma *AdaBoost* menghasilkan nilai akurasi sebesar 85%. Penelitian lain yang dilakukan Paput *et al.,* (2023) membandingkan metode *random forest* dan *adaptive boosting* untuk data indeks pembangunan manusia (IPM) menghasilkan akurasi dengan *adaptive boosting* sebesar 96,08% yang lebih tinggi dari akurasi dengan *random forest* sebesar 95,10%.

Penelitian yang pernah dilakukan sebelumnya menunjukkan metode *Gradient Boosting* dan *AdaBoost* diketahui dapat menghasilkan evaluasi kinerja klasifikasi yang cukup baik. Dengan mempertimbangkan keunggulan dari kedua metode yang telah dijelaskan sebelumnya, mendorong peneliti untuk melakukan penelitian yang berjudul "Perbandingan Algoritma *Gradient Boosting* dan *Adaptive Boosting* untuk Klasifikasi Indeks Standar Pencemar Udara". Penelitian ini berfokus pada pembahasan terkait klasifikasi kategori Indeks Standar Pencemar Udara (ISPU) di Provinsi DKI Jakarta menggunakan algoritma *Gradient Boosting* dan *Adaptive Boosting* untuk mengetahui metode yang lebih baik dalam melakukan perhitungan akurasi. Variabel bebas (independen) yang digunakan dalam penelitian yaitu PM_{10},

Nitrogen dioksida (NO_2), Ozon (O_3), Sulfur dioksida (SO_2), dan Karbon monoksida (CO). Penelitian ini diharapkan dapat memberikan metode alternatif yang akurat dalam melakukan klasifikasi sehingga dapat membantu pihak terkait dalam proses pemantauan tingkat polusi udara secara lebih efektif.

1.2 Rumusan Masalah

Perumusan masalah dalam penelitian ini berdasarkan latar belakang yang telah dijelaskan adalah sebagai berikut:

1. Bagaimana hasil klasifikasi kategori Indeks Standar Pencemar Udara (ISPU) di DKI Jakarta berdasarkan nilai *accuracy* menggunakan *Gradient Boosting*?
2. Bagaimana hasil klasifikasi kategori Indeks Standar Pencemar Udara (ISPU) di DKI Jakarta berdasarkan nilai *accuracy* menggunakan *Adaptive Boosting* (*AdaBoost*)?
3. Bagaimana perbandingan ketepatan hasil klasifikasi kategori Indeks Standar Pencemar Udara (ISPU) di DKI Jakarta berdasarkan nilai *accuracy* menggunakan *Gradient Boosting* dan *Adaptive Boosting*?

1.3 Batasan Masalah

Batasan masalah yang digunakan pada penelitian tugas akhir ini adalah:

1. Data sekunder yang diperoleh dari *website* Open Data Jakarta terkait data Indeks Standar Pencemar Udara (ISPU) provinsi DKI Jakarta rentang waktu Desember 2022 hingga November 2023.
2. Pengolahan data menggunakan algoritma *Gradient Boosting* dan *Adaptive Boosting* dengan pengolahan menggunakan *software* Python.

3. Variabel prediktor yang digunakan dalam penelitian ini ialah polutan utama, yaitu partikulat PM_{10}, Nitrogen dioksida (NO_2), Ozon (O_3), Sulfur dioksida (SO_2), dan Karbon monoksida (CO).

1.4 Tujuan Penelitian

Tujuan dari dilakukannya penelitian ini adalah:

1. Mengimplementasikan algoritma *Gradient Boosting* dan *Adaptive Boosting* dalam kasus klasifikasi kategori Indeks Standar Pencemar Udara (ISPU) di DKI Jakarta.
2. Membandingkan ketepatan hasil klasifikasi algoritma *Gradient Boosting* dan *Adaptive Boosting* berdasarkan hasil *accuracy* untuk mengetahui algoritma yang lebih baik dalam kasus klasifikasi kategori Indeks Standar Pencemar Udara (ISPU) di DKI Jakarta.