BAB I
PENDAHULUAN

1.1 Latar Belakang

Menurut Organisasi Kesehatan Dunia (WHO), polusi udara adalah campuran berbagai bahan seperti partikel padat, tetesan cairan, dan gas. Polusi udara ini telah menjadi masalah besar dan sangat mendesak di seluruh dunia karena bisa menyebabkan berbagai penyakit seperti gangguan pernapasan, penyakit jantung, dan kanker paru-paru (Mujtaba et al., 2025).WHO juga menyatakan bahwa polusi udara menyebabkan sekitar tujuh juta kematian dini setiap tahunnya, baik akibat polusi udara di luar ruangan maupun di dalam rumah, dan jutaan orang lain menderita sakit karena menghirup udara yang tercemar. Lebih dari separuh kematian tersebut terjadi di negara-negara berkembang (WHO, 2021). Ada berbagai sumber yang menyebabkan polusi udara, seperti kegiatan industri, kebakaran hutan, asap rokok, serta transportasi (Syukur & Chamidy, 2025). Beberapa polutan yang sering diukur termasuk $PM_{2.5}$, PM_{10} (partikel dengan ukuran aerodinamis kurang dari atau sama dengan 2,5 dan 10 mikrometer), ozon (O_3), nitrogen dioksida (NO_2), karbon monoksida (CO), dan sulfur dioksida (SO_2) (WHO, 2021).

Di Indonesia, khususnya ibu kota Jakarta, masalah polusi udara telah mencapai tingkat yang mengkhawatirkan. Menurut laporan *real-time* dari *IQAir*, sebuah platform informasi kualitas udara global, Jakarta seringkali menempati peringkat teratas sebagai kota dengan kualitas udara terburuk di dunia. Sebagai contoh, data yang diambil pada 17 September 2025 menunjukkan bahwa Jakarta berada di peringkat satu kota besar paling berpolusi dengan nilai Indeks Standar Pencemar Udara (ISPU) mencapai 172, yang masuk dalam kategori "Tidak Sehat". Kondisi ini mendorong kebutuhan mendesak akan sistem prediksi yang akurat untuk memberikan informasi dini dan mendukung pengambilan keputusan yang efektif dalam mitigasi polusi.

Di Indonesia, kualitas udara dinilai melalui Indeks Standar Pencemar Udara (ISPU), yaitu angka yang tidak memiliki satuan dan menunjukkan kondisi udara di

suatu tempat. Penilaian ini didasarkan pada dampak terhadap kesehatan manusia, nilai estetika, serta kehidupan makhluk lainnya. Peraturan Menteri Lingkungan Hidup dan Kehutanan Nomor P.14/MENLHK/SETJEN/KUM.1/7/2020 menetapkan beberapa kategori ISPU, yaitu:

- Baik (0-50)
- Sedang (51-100)
- Tidak Sehat bagi Kelompok Sensitif (101-150)
- Tidak Sehat (151-200)
- Sangat Tidak Sehat (201-300)
- Berbahaya (>300)

Setiap kategori menunjukkan tingkat risiko kesehatan yang berbeda. Jika ISPU mencapai 101 atau lebih, kondisi udara dianggap tidak aman, terutama bagi kelompok yang lebih rentan terhadap dampak negatif udara tercemar.

Machine learning telah menjadi pendekatan yang efektif dalam memodelkan fenomena kompleks seperti polusi udara. Berbagai penelitian telah menerapkan algoritma *machine learning* untuk memodelkan ISPU, dengan mayoritas menggunakan pendekatan klasifikasi untuk mengelompokkan kualitas udara ke dalam kategori yang ditetapkan. Penelitian-penelitian di Jakarta menunjukkan hasil yang baik dalam tugas klasifikasi, seperti yang dilakukan oleh (Amalia et al., 2022) menggunakan algoritma *K-Nearest Neighbor* (KNN), (Jayadi et al., 2023) yang membandingkan KNN dan *Support Vector Machine* (SVM), serta (Luthfi & Fauzi, 2024) yang membandingkan *Random Forest*, SVM, dan LGBM untuk klasifikasi kualitas udara di Jakarta.

Meskipun pendekatan klasifikasi memberikan gambaran umum mengenai kualitas udara dengan membaginya ke dalam kategori tertentu seperti "Baik" atau "Tidak Sehat" (Kannajmi & Saputra, 2025) pendekatan ini tidak dapat menangkap detail dan variasi nilai ISPU secara spesifik. Pendekatan regresi sangat penting karena dapat memprediksi nilai ISPU yang lebih akurat dan memberikan informasi yang lebih presisi untuk mengetahui kondisi udara secara detail (Gupta et al., 2023).

Informasi nilai yang tepat ini sangat diperlukan untuk pengambilan keputusan yang lebih akurat dalam upaya mitigasi polusi udara. Penelitian ini akan menggunakan metode regresi dengan membandingkan dua algoritma *machine learning* yang sangat kuat, yaitu *Random Forest* dan *XGBoost.* Kedua algoritma tersebut dipilih berdasarkan beberapa keunggulan: pertama, keduanya termasuk dalam kategori *ensemble learning* yang terbukti akurat dan mampu menangani data yang kompleks (Sabri et al., 2025). Kedua, *Random Forest* bekerja dengan cara membuat banyak pohon keputusan secara paralel dan menggabungkan hasilnya, yang efektif dalam mengurangi *overfitting* dan memberikan tingkat akurasi yang tinggi (Yenkikar et al., 2025). Ketiga, *XGBoost* (*Extreme Gradient Boosting*) adalah algoritma *boosting* yang membangun pohon keputusan secara bertahap untuk memperbaiki kesalahan model sebelumnya, dan dikenal memiliki kecepatan tinggi serta performa yang sangat baik dalam berbagai tugas prediksi (Alfasanah et al., 2025).

Penelitian yang menggunakan pendekatan regresi untuk memodelkan kualitas udara sudah banyak dilakukan. Sebagai contoh, (Gupta et al., 2023) secara khusus menggunakan metode regresi seperti *Random Forest Regression* (RFR), *Support Vector Regression* (SVR), dan *CatBoost Regression* (CR) untuk memperkirakan Indeks Kualitas Udara (AQI) di beberapa kota di India. Studi ini menunjukkan bahwa baik RFR maupun *CatBoost* menghasilkan akurasi yang tinggi, namun kinerja model bervariasi antar kota. Pada dataset yang telah diseimbangkan, RFR menunjukkan performa terbaik di Kolkata (93.74%) dan Hyderabad (97.61%), sementara *CatBoost* unggul di New Delhi (85.08%) dan Bangalore (90.31%). Di sisi lain, (Singh et al., 2025) juga menggunakan pendekatan regresi dengan membandingkan beberapa model *supervised machine learning*, termasuk *XGBoost* dan *Random Forest*, untuk memodelkan Indeks Kualitas Udara (AQI) di India. Penelitian ini berhasil menemukan bahwa model *XGBoost* memberikan performa terbaik, bahkan mencapai akurasi sempurna 1.0 setelah penyesuaian *hyperparameter*, sedangkan model *Random Forest* mencapai MAE 0.97 dan R-squared 0.998.

Namun, penelitian-penelitian tersebut dilakukan di India dengan karakteristik polusi yang berbeda dari Jakarta. Sementara itu, (Syahreza et al., 2024) memang membandingkan *Random Forest* dan *XGBoost*, namun dalam konteks prediksi cuaca, bukan polusi udara, dan menemukan bahwa *XGBoost* menunjukkan kinerja terbaik dengan nilai R^2 sebesar 0,8183. Oleh karena itu, belum ada penelitian yang secara khusus membandingkan performa kedua algoritma regresi ini untuk memodelkan nilai ISPU di Jakarta. Secara teoretis, penelitian ini akan memberikan kontribusi dalam bidang *machine learning* untuk pemodelan kualitas udara dengan membandingkan secara detail kinerja kedua algoritma *ensemble learning* terkuat. Secara praktis, model terbaik yang ditemukan dapat dimanfaatkan sebagai alat bantu yang lebih akurat oleh instansi terkait, seperti Dinas Lingkungan Hidup DKI Jakarta, untuk memberikan informasi yang lebih presisi, sehingga mendukung upaya strategis dalam menjaga kesehatan warga dan mengurangi dampak buruk polusi udara di Jakarta.

Pencemaran udara tidak hanya dipengaruhi oleh tingginya konsentrasi polutan, tetapi juga oleh kondisi meteorologi yang berperan dalam dinamika polutan di atmosfer. Faktor-faktor seperti suhu udara, kelembapan, curah hujan, serta kecepatan dan arah angin memengaruhi proses penyebaran, pengendapan, hingga pencucian polutan melalui mekanisme *wet deposition* (Mujtaba et al., 2025). Oleh karena itu, untuk memperoleh hasil prediksi yang lebih menyeluruh, penelitian ini mengintegrasikan enam parameter polutan utama dengan enam variabel meteorologi, yaitu radiasi matahari (GRAD W/m2), suhu udara (TempAir degrees C), kelembapan udara (HumAir %), kecepatan angin (FF m/s), arah angin (DD degrees), dan curah hujan (RAIN mm).

Data yang digunakan dalam penelitian ini merupakan data historis periode tahun 2020 hingga 2022 yang telah diagregasi ke dalam bentuk data bulanan. Pendekatan ini diharapkan dapat memberikan gambaran yang lebih stabil terhadap pola kualitas udara di Jakarta. Dengan demikian, penelitian ini diharapkan mampu memberikan kontribusi secara ilmiah dalam menentukan model prediksi yang paling sesuai untuk memprediksi nilai ISPU. Selain itu, secara praktis, hasil penelitian ini

dapat dimanfaatkan sebagai bahan pertimbangan bagi instansi terkait, khususnya Dinas Lingkungan Hidup DKI Jakarta, dalam merumuskan strategi mitigasi pencemaran udara yang lebih tepat sasaran dan berbasis pada integrasi data meteorologi serta parameter polutan.

1.2 Rumusan Masalah

1. Bagaimana kinerja model *Random Forest Regressor* dan *XGBoost Regressor* dalam memprediksi nilai Indeks Standar Pencemar Udara (ISPU) maksimum bulanan di Jakarta?
2. Model mana dari kedua model tersebut yang memberikan hasil terbaik berdasarkan metrik evaluasi regresi RMSE, MAE, dan R^2?

1.3 Tujuan

1. Menganalisis kinerja model *Random Forest Regressor* dan *XGBoost Regressor* dalam memprediksi nilai Indeks Standar Pencemar Udara (ISPU) maksimum bulanan di Jakarta.
2. Membandingkan performa kedua model tersebut berdasarkan metrik evaluasi regresi, yaitu *Root Mean Squared Error* (RMSE), *Mean Absolute Error* (MAE), dan *R-squared* (R^2), untuk menentukan model yang paling efektif dalam studi kasus ini.

1.4 Manfaat

1. Menambah pengetahuan di bidang *machine learning*, khususnya penerapan algoritma regresi (*Random Forest* dan *XGBoost*) untuk pemodelan ISPU sebagai referensi pengembangan metode prediksi kualitas udara selanjutnya.
2. Memberikan model prediksi kualitas udara yang akurat sehingga dapat dimanfaatkan oleh instansi terkait, seperti Dinas Lingkungan Hidup DKI Jakarta, dalam menyusun strategi mitigasi polusi udara serta mendukung perumusan kebijakan berbasis data dalam perencanaan jangka menengah.
3. Memberikan informasi yang lebih akurat mengenai kondisi kualitas udara, sehingga masyarakat dapat meningkatkan kesadaran lingkungan, mengambil

langkah preventif terhadap dampak kesehatan, serta mendukung partisipasi dalam upaya menjaga kualitas udara.

1.5 Ruang Lingkup

Ruang lingkup penelitian ini difokuskan pada perbandingan kinerja algoritma *machine learning* dalam memprediksi Indeks Standar Pencemar Udara (ISPU). Pembatasan ini dilakukan untuk memastikan penelitian dapat dilaksanakan secara mendalam dan komprehensif, dengan rincian sebagai berikut:

1. Penelitian ini menggunakan data kualitas udara yang berasal dari Provinsi DKI Jakarta. Data yang digunakan mencakup periode pengamatan tahun 2020 hingga 2022 dengan total 36 bulan data pada setiap stasiun pemantauan kualitas udara. Seluruh data diperoleh dari Dinas Lingkungan Hidup DKI Jakarta.
2. Penelitian ini menerapkan pendekatan regresi dengan membandingkan dua algoritma *ensemble learning*, yaitu *Random Forest Regressor* dan *XGBoost Regressor*, dalam memprediksi nilai maksimum Indeks Standar Pencemar Udara (ISPU) secara kontinu.
3. Variabel terikat (*dependent variable*) dalam penelitian ini adalah nilai maksimum (Max) Indeks Standar Pencemar Udara (ISPU). Sementara itu, variabel bebas (*independent variables*) terdiri dari 12 parameter yang mencakup enam parameter polutan, yaitu PM2.5, PM10, SO_2, CO, O_3, dan NO_2, serta enam parameter meteorologi, yaitu radiasi matahari (GRAD W/m2), suhu udara (TempAir degrees C), kelembapan udara (HumAir %), kecepatan angin (FF m/s), arah angin (DD degrees"), dan curah hujan (RAIN mm).
4. Penelitian ini menggunakan data kualitas udara dan meteorologi yang pada awalnya tersedia dalam bentuk data harian, kemudian diagregasi menjadi data bulanan untuk mengurangi fluktuasi harian dan menangkap tren kualitas udara jangka menengah.
5. Penelitian ini memodelkan hubungan antara konsentrasi polutan dan kondisi meteorologi terhadap nilai Indeks Standar Pencemar Udara (ISPU). Model yang dihasilkan selanjutnya digunakan untuk memprediksi nilai ISPU.

6. Evaluasi kinerja kedua model dilakukan menggunakan metrik regresi standar, yaitu *Root Mean Squared Error* (RMSE), *Mean Absolute Error* (MAE), dan *R-squared* (R^2).

1.6 Kebaruan Penelitian

Kebaruan dalam penelitian ini difokuskan pada penerapan model prediksi dengan karakteristik sebagai berikut:

1. Penerapan Pendekatan Regresi pada Data ISPU: Sebagian besar penelitian terdahulu mengenai prediksi kualitas udara di Jakarta menerapkan pendekatan klasifikasi untuk mengelompokkan kondisi udara ke dalam kategori diskret seperti Baik, Sedang, atau Tidak Sehat. Hal ini tercermin pada studi (Amalia et al., 2022) yang menggunakan algoritma K-Nearest Neighbor untuk klasifikasi kualitas udara, (Jayadi et al., 2023) yang membandingkan KNN dan SVM dalam konteks klasifikasi serupa, serta (Kannajmi & Saputra, 2025) yang berfokus pada penentuan model klasifikasi terbaik untuk kualitas udara Jakarta. Berbeda dari kecenderungan tersebut, penelitian ini menerapkan pendekatan regresi untuk memprediksi nilai numerik ISPU maksimum secara langsung dan kontinu. Pendekatan ini menghasilkan informasi tingkat pencemaran yang lebih terperinci dan presisi, sehingga lebih relevan untuk keperluan analisis tren jangka menengah dan perumusan strategi mitigasi pencemaran udara berbasis data kuantitatif. Keunggulan pendekatan regresi ini juga sejalan dengan temuan (Gupta et al., 2023) yang menegaskan bahwa prediksi berbasis nilai kontinu memberikan wawasan yang lebih komprehensif dibandingkan model kategorikal semata.
2. Penambahan Variabel Meteorologi: Penelitian ini menyertakan enam parameter polutan utama, yakni PM10, PM2.5, SO_2, CO, O_3, dan NO_2, yang dilengkapi dengan enam variabel meteorologi meliputi radiasi matahari (GRAD, W/m^2), suhu udara (TempAir, °C), kelembapan udara (HumAir, %), kecepatan angin (FF, m/s), arah angin (DD, °), dan curah hujan (RAIN, mm). Sebagian penelitian terdahulu, seperti (Jayadi et al., 2023) dan (Luthfi & Fauzi, 2024), tidak menyertakan parameter meteorologi secara menyeluruh dalam model

prediksinya. Penambahan variabel meteorologi secara komprehensif dalam penelitian ini dilandasi oleh fakta bahwa kondisi cuaca berpengaruh signifikan terhadap dispersi dan konsentrasi polutan di atmosfer. Mujtaba et al. (2025) dalam kajiannya mengenai prediksi kualitas udara berbasis *machine learning* untuk perencanaan perkotaan berkelanjutan menegaskan bahwa penyertaan parameter lingkungan dan meteorologi yang lengkap merupakan faktor determinan dalam meningkatkan akurasi dan keandalan model prediksi.

3. Pendekatan Agregasi Data Bulanan untuk Analisis Tren Jangka Menengah: Penelitian-penelitian terdahulu terkait prediksi kualitas udara di Jakarta, seperti (Amalia et al., 2022), serta (Kannajmi & Saputra, 2025), umumnya beroperasi pada resolusi temporal harian yang rentan terhadap fluktuasi jangka pendek dan derau data (*noise*). Penelitian ini menerapkan agregasi data bulanan untuk menganalisis tren kualitas udara jangka menengah pada periode 2020–2022, menghasilkan representasi pola yang lebih stabil, khususnya dalam menangkap variasi musiman yang terkait dengan dinamika iklim tropis Jakarta. Pendekatan ini juga relevan secara kebijakan, sebagaimana ditegaskan dalam regulasi KLHK (2020) yang menekankan pentingnya pemantauan kualitas udara untuk mendukung perumusan kebijakan lingkungan. Lebih lanjut, urgensi pendekatan agregasi periodik ini diperkuat oleh fakta lapangan yang diperoleh dari keterangan Muhammad Ade Firmansyah, ASN Bidang Kemitraan Data dan Informasi Dinas Lingkungan Hidup (DLH) Provinsi DKI Jakarta. Beliau menyatakan bahwa hingga saat penelitian ini dilaksanakan, DLH Jakarta belum memiliki sistem pemantauan atau pelaporan kualitas udara secara bulanan maupun periodik yang terstruktur. Data kualitas udara yang tersedia selama ini hanya bersifat harian. Dengan demikian, penelitian ini mengisi celah nyata dalam praktik pemantauan lingkungan di tingkat daerah, sekaligus memberikan kontribusi metodologis yang dapat dimanfaatkan oleh DLH Jakarta dalam pengembangan sistem pelaporan kualitas udara ke depan.
4. Studi Komparasi Algoritma *Ensemble Learning*: Penelitian ini membandingkan kinerja dua algoritma *ensemble learning*, yaitu *Random Forest Regressor* dan *XGBoost Regressor*, secara spesifik pada dataset kualitas udara Jakarta untuk

menentukan algoritma mana yang paling efisien untuk karakteristik data tersebut. Hasil perbandingan ini diharapkan dapat menjadi referensi bagi penelitian selanjutnya dalam pengembangan model prediksi ISPU yang lebih akurat.

1.7 Hipotesis

- H_0: Tidak terdapat perbedaan yang signifikan antara kinerja *Random Forest Regressor* dan *XGBoost Regressor* dalam memprediksi nilai ISPU bulanan di Jakarta berdasarkan metrik evaluasi RMSE, MAE, dan R^2.
- H_1: Terdapat perbedaan yang signifikan antara kinerja *Random Forest Regressor* dan *XGBoost Regressor* dalam memprediksi nilai ISPU bulanan di Jakarta berdasarkan metrik evaluasi RMSE, MAE, dan R^2.