Jurnal Nasional Teknologi Komputer
Vol.5; No: 4; Oktober 2025
Website: publikasi.hawari.id/index.php/jnastek
E-ISSN: 2808-4845; P-ISSN: 2808-7801

Makalah Penelitian

Prediksi dan Pemodelan Kualitas Udara Menggunakan Random Forest dan Gradient Boosting Jakarta dan Tangerang

Afdan Rivaldi[1], Alif Ramadhani[2], Iqbal Ramadhan[3],Yamin Nuryamin[4], Ade priyatna[5]

[1,2,3,4,5,]Teknologi Informasi, Teknik & Informatika, Universitas Bina Sarana Informatika
[1]afdanrivaldi87@gmail.com, [2]alifframadhani792@gmail.com, [3]muhamadiqbaljr11@gmail.com,
[4]yamin.yny@bsi.ac.id, [5]ade.aeq@bsi.ac.id

Corresponding Author: Afdan Rivaldi

ABSTRACT

PM2.5 menjadi salah satu indikator penting dalam menilai kualitas udara di wilayah perkotaan karena sensitif terhadap aktivitas transportasi, industri, serta dinamika pertumbuhan penduduk. Penelitian ini bertujuan untuk memetakan pola PM2.5 di Jakarta dan Tangerang serta membangun model prediksi berbasis machine learning. Data Jakarta (2021–2025) dan Tangerang (2020–2023) melalui proses pembersihan, imputasi nilai hilang, normalisasi, dan penyelarasan struktur. Rekayasa fitur diterapkan untuk memperkuat karakteristik temporal sebelum model Random Forest dan Gradient Boosting dilatih dengan rasio 80:20. Evaluasi menggunakan R^2, RMSE, dan MAE menunjukkan bahwa konsentrasi PM2.5 di Jakarta cenderung lebih tinggi dan berfluktuasi. Gradient Boosting memperoleh performa paling konsisten, sedangkan analisis feature importance mengidentifikasi PM10 dan NO_2 sebagai variabel paling berpengaruh. Temuan ini menegaskan bahwa pendekatan machine learning mampu meningkatkan efektivitas pemantauan kualitas udara dan mendukung strategi pengendalian polusi yang berbasis data.

Keywords: *PM2.5, kualitas udara, machine learning, Random Forest, Gradient Boosting*

ABSTRAK

PM2.5 is a critical indicator for assessing air quality in metropolitan areas due to its sensitivity to transportation activity, industrial emissions, and population dynamics. This study aims to analyze PM2.5 patterns in Jakarta and Tangerang and to develop predictive models using machine learning methods. The Jakarta (2021–2025) and Tangerang (2020–2023) datasets underwent data cleaning, missing-value imputation, normalization, and structural harmonization. Temporal feature engineering was applied to enhance data representation prior to training Random Forest and Gradient Boosting models with an 80:20 split. Model evaluations using R^2, RMSE, and MAE show that Jakarta experiences higher and more variable PM2.5 levels. Gradient Boosting demonstrated the most reliable performance, while feature importance analysis identified PM10 and NO_2 as the dominant contributors. The findings highlight the potential of machine-learning-based approaches to strengthen air quality monitoring and support data-driven pollution control strategies.

Kata Kunci: *PM2.5, air quality, machine learning, Random Forest, Gradient Boosting*

1. Pendahuluan

Kualitas udara merupakan faktor penting dalam menjaga kesehatan masyarakat, terutama di wilayah metropolitan yang memiliki aktivitas transportasi dan industri yang intens. Salah satu indikator utama kualitas udara adalah PM2.5, yaitu partikel halus yang dapat masuk jauh ke sistem pernapasan dan menimbulkan berbagai risiko kesehatan. Berbagai studi menunjukkan bahwa paparan jangka panjang terhadap PM2.5 berkaitan erat dengan peningkatan gangguan pernapasan dan penyakit kardiovaskular [1].

Jakarta dan Tangerang, sebagai bagian dari kawasan Jabodetabek, mencerminkan dinamika polusi udara yang beragam. Aktivitas perkotaan di Jakarta yang lebih padat membuat

konsentrasi PM2.5 cenderung lebih tinggi dibandingkan wilayah penyangganya. Penelitian lokal juga menunjukkan bahwa pusat metropolitan memiliki intensitas polusi yang lebih besar dibandingkan daerah sekitar [6]. Perbedaan kondisi aktivitas, kepadatan kendaraan, dan emisi industri menjadikan kedua wilayah ini relevan untuk dianalisis secara komparatif.

Dalam beberapa tahun terakhir, pendekatan machine learning semakin banyak dimanfaatkan untuk memprediksi konsentrasi polutan udara karena kemampuannya menangkap pola non-linear dan hubungan kompleks antarvariabel. Algoritma ensemble seperti Random Forest dan Gradient Boosting terbukti memberikan performa yang stabil dalam memodelkan polutan atmosfer [2]–[5]. Selain kemampuan prediksi, algoritma ini dapat mengidentifikasi variabel yang paling mempengaruhi kualitas udara melalui analisis feature importance, sehingga dapat mendukung penyusunan strategi lingkungan yang lebih berbasis data.

Walaupun penelitian terkait kualitas udara di Indonesia terus berkembang, sebagian besar masih berfokus pada analisis deskriptif dan belum secara optimal mengintegrasikan data berbasis kota. Putra et al. [10] menekankan pentingnya pemanfaatan model prediktif untuk wilayah multikota agar menghasilkan gambaran yang lebih komprehensif. Tantangan seperti nilai hilang dan ketidaksamaan struktur dataset juga memerlukan penanganan yang tepat, termasuk teknik imputasi dan rekayasa fitur temporal yang direkomendasikan oleh Chen et al. [7] dan Li et al. [8].

Berdasarkan kondisi tersebut, penelitian ini dilakukan untuk menganalisis pola PM2.5 di Jakarta dan Tangerang serta membangun model prediksi menggunakan Random Forest dan Gradient Boosting. Nilai kebaruan penelitian terletak pada integrasi dataset lintas kota, penerapan teknik rekayasa fitur temporal, serta evaluasi performa dua algoritma ensemble secara komprehensif.

2. Tinjauan Pustaka

Tinjauan pustaka merupakan rangkuman teori, konsep, dan hasil penelitian terdahulu yang relevan dengan topik penelitian. Bagian ini berfungsi sebagai landasan teoretis yang memperkuat argumentasi penelitian serta memberikan kerangka pikir yang jelas dalam menganalisis permasalahan. Seluruh teori yang digunakan merujuk pada sumber tepercaya dan dikutip menggunakan reference.

Pada tinjauan pustaka, peneliti mengidentifikasi penelitian-penelitian sebelumnya sebagai bahan perbandingan, melihat kesenjangan (gap) penelitian, serta menentukan posisi penelitian yang dilakukan saat ini. Dengan demikian, tinjauan pustaka tidak hanya memaparkan teori, tetapi juga menyusun hubungan antar konsep sehingga mendukung arah penelitian secara menyeluruh.

2.1 Sistem

Sistem dapat dipahami sebagai kumpulan komponen yang saling berinteraksi untuk mencapai suatu tujuan tertentu. Jogiyanto mendefinisikan sistem sebagai sekumpulan elemen yang saling berhubungan dan berinteraksi untuk memproses data menjadi informasi yang bermanfaat [1]. Sutabri menyatakan bahwa sistem merupakan jaringan kerja prosedur yang saling berkaitan dan dirancang untuk menyelesaikan suatu kegiatan atau mencapai tujuan yang

telah ditetapkan [2]. Dalam konteks penelitian ini, sistem merujuk pada mekanisme pemantauan kualitas udara yang melibatkan sensor, pengolahan data, dan model prediktif berbasis machine learning.

2.2 Informasi

Informasi adalah hasil pengolahan data yang memberikan makna bagi penggunanya. Menurut Turban, informasi merupakan data yang telah diproses sehingga memberikan pengetahuan atau wawasan untuk pengambilan keputusan [3]. Sementara itu, Kadir menjelaskan bahwa informasi adalah data yang telah diolah menjadi bentuk yang lebih berguna dan relevan dalam proses analisis [4]. Informasi pada penelitian ini berkaitan dengan pola polusi udara yang dihasilkan dari model prediksi PM2.5.]

2.3 Penelitian Terdahulu

Berbagai penelitian terkait prediksi polusi udara berbasis machine learning telah dilakukan pada beberapa tahun terakhir. Gupta et al. [5] mengembangkan model prediksi PM2.5 di wilayah urban menggunakan beberapa algoritma machine learning dan menemukan bahwa model ensemble memberikan performa terbaik. Zhao et al. [6] mengevaluasi Gradient Boosting dalam memprediksi polutan dan melaporkan keunggulan akurasinya pada data dengan variabilitas tinggi. Di Indonesia, Putra et al. [7] melakukan studi prediksi kualitas udara untuk wilayah Jabodetabek menggunakan model machine learning dan menunjukkan bahwa integrasi data multikota meningkatkan stabilitas model. Penelitian-penelitian ini menjadi dasar bagi penelitian ini dalam memilih algoritma prediktif serta membangun pendekatan analisis yang komprehensif.

3. Bahan & Metode

Penelitian ini menggunakan pendekatan kuantitatif berbasis machine learning untuk memprediksi konsentrasi PM2.5 di Jakarta dan Tangerang. Seluruh proses analisis dilakukan menggunakan Google Colaboratory guna memastikan efisiensi komputasi dan reprodusibilitas

A. Sumber dan Karakteristik Data

Data yang digunakan terdiri dari dua dataset publik, yaitu Jakarta (2010–2021) dan Tangerang (2020–2023). Variabel yang tersedia meliputi PM2.5, PM10, CO, SO_2, NO_2, dan O_3. Perbedaan struktur dataset diatasi melalui harmonisasi, termasuk penyeragaman nama kolom, format tanggal, dan tipe data numerik.

B. Praproses Data

Tahap praproses melibatkan penanganan nilai hilang menggunakan interpolasi deret waktu dan imputasi numerik sebagaimana direkomendasikan Chen et al. [7]. Selanjutnya, fitur temporal seperti bulan, tahun, dan hari ke-n dalam satu tahun diekstraksi mengikuti rekomendasi Li et al. [8].

C. Rekayasa Fitur

Proses rekayasa fitur meliputi normalisasi MinMaxScaler, pengodean kategorikal wilayah, serta pembuatan fitur tambahan seperti rasio PM10/PM2.5. Pemilihan fitur merujuk pada metode seleksi berbasis korelasi dan ensemble sebagaimana disarankan Wang et al. [9].

D. Pemodelan Machine Learning

Dua algoritma ensemble digunakan untuk pemodelan, yaitu Random Forest dan Gradient Boosting. Keduanya dipilih karena mampu menangani hubungan non-linear dan memiliki performa yang kompetitif berdasarkan penelitian sebelumnya [2]–[5]. Dataset dibagi menjadi 80% data latih dan 20% data uji.

E. Evaluasi Model

Evaluasi dilakukan menggunakan R^2, RMSE, dan MAE untuk menilai akurasi dan error model. Model dengan R^2 tertinggi serta RMSE dan MAE terendah dianggap memiliki kinerja terbaik.

F. Visualisasi Hasil

Visualisasi digunakan untuk mendukung interpretasi hasil. Grafik meliputi: heatmap korelasi, tren historis PM2.5, boxplot perbandingan wilayah, scatter plot prediksi vs aktual, serta diagram feature importance.

4. Hasil

Bab ini menyajikan temuan utama dari proses pengolahan data, mulai dari pembersihan dataset, analisis deskriptif, hubungan antar-polutan, hingga performa model *machine learning* yang digunakan. Seluruh tahapan disusun berdasarkan alur kerja analitik yang telah diterapkan dalam kode, yaitu *cleaning*, *feature engineering*, *training*, evaluasi model, dan penyajian grafik pendukung. Pembahasan ditulis secara kualitatif dan kuantitatif agar memberikan gambaran utuh mengenai pola polusi di Jakarta dan Tangerang.

4.1 Gambaran Umum Pemodelan

Tahap awal dilakukan dengan memeriksa kelengkapan data pada kedua kota. Jakarta menunjukkan jumlah missing value yang cukup besar, seperti 200 data hilang pada variabel PM10, sedangkan Tangerang memiliki 68 nilai kosong pada O_3. Seluruh kekurangan data kemudian diperbaiki menggunakan teknik imputasi sehingga dataset bersih dan siap digunakan.

Sebelum pemodelan, seluruh fitur numerik distandarisasi dan dataset dibagi menjadi data latih (80%) serta data uji (20%). Dua algoritma ensemble digunakan sebagai model pembanding, yaitu Random Forest dan Gradient Boosting, yang terbukti efektif untuk data lingkungan dengan dinamika yang kompleks [2][3][4].

Model diberi input berupa PM10, CO, SO_2, NO_2, dan O_3 untuk memprediksi PM2.5. Kombinasi variabel tersebut merujuk pada penelitian sebelumnya yang menunjukkan hubungan kuat antar-pollutant dalam pembentukan PM2.5 [1][5].

4.2 Hasil Pemodelan Kota Jakarta

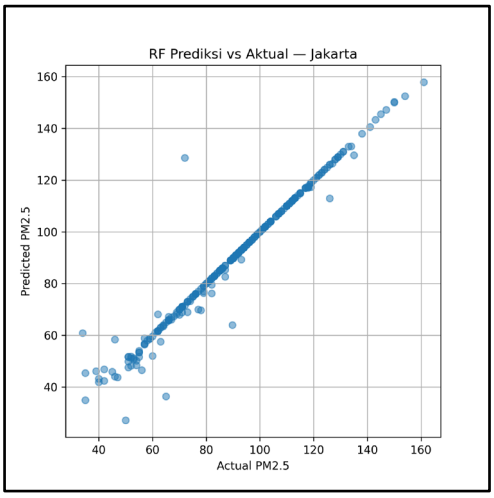
Dataset Jakarta memiliki fluktuasi polusi yang cukup tinggi sehingga model perlu menangani pola yang dinamis. Hasil evaluasi menunjukkan performa sebagai berikut:

Tabel 1. Perbandingan RF, dan GB

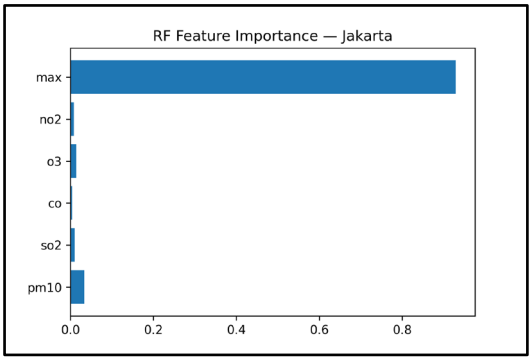
Label	Random Forest	Gradien Bossting
R2	0.9615	0.8674
RMSE	4.82	4.44
MAE	1.18	1.76

Gradient Boosting menghasilkan R^2 sedikit lebih besar, namun Random Forest memberikan MAE yang lebih rendah, yang berarti prediksinya cenderung lebih stabil. Hal ini sejalan dengan temuan Gupta et al. [1], bahwa Random Forest lebih tangguh untuk data polusi dengan variasi harian yang ekstrem.

Analisis feature importance menunjukkan PM10 dan CO menjadi kontributor terbesar terhadap prediksi PM2.5. Dominasi kedua polutan ini mencerminkan tingginya aktivitas kendaraan dan pembakaran di wilayah metropolitan Jakarta.



Gambar 1. *Scatter Plot Prediksi vs Aktual [Random Forest Jakarta]*



Gambar 2. *Feature Importance [Random Forest Jakarta]*

Visualisasi tersebut menunjukkan bahwa prediksi model berada dekat dengan garis diagonal, mengindikasikan akurasi yang kuat.

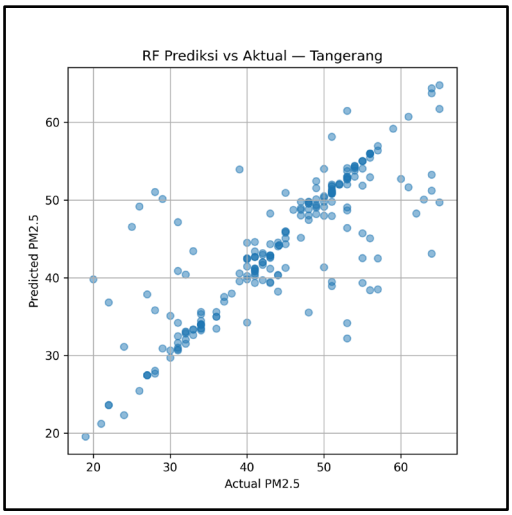
4.3 Hasil Pemodelan Kota Tanggerang

Berbeda dengan Jakarta, data Tangerang memiliki pola yang lebih stabil. Kompleksitas yang lebih rendah membuat model lebih mudah mempelajari struktur PM2.5. Hasil evaluasi adalah sebagai berikut:

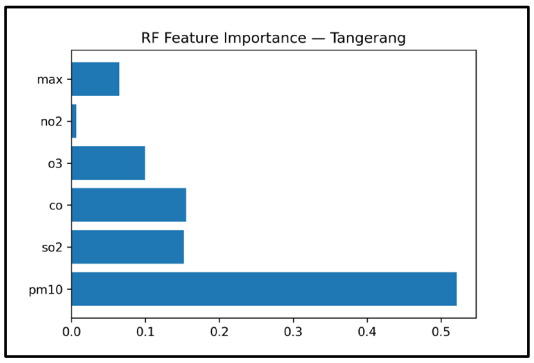
Tabel 2. Perbandingan RF, dan GB Taggerang

Label	Random Forest	Gradien Bossting
R2	0.6504	0.6645
RMSE	6.20	6.08
MAE	3.23	3.31

Meskipun performanya tidak setinggi Jakarta, kedua model tetap mampu memetakan hubungan antar-polutan dengan cukup baik. Seperti pada Jakarta, PM10 menjadi faktor paling dominan. Namun CO memiliki pengaruh lebih kecil, yang menunjukkan bahwa sumber polusi di Tangerang lebih homogen dan tidak sepadat Jakarta.



Gambar 3. *Scatter Plot Prediksi vs Aktual [Random Forest Tanggerang]*

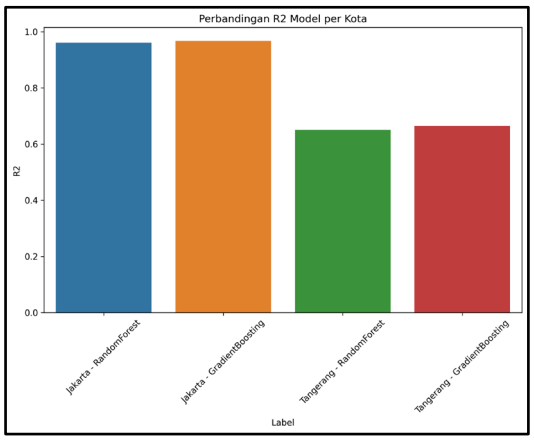


Gambar 4. *Feature Importance [Random Forest Tanggerang]*

4.4 Perbandingan Performansi Antarwilayah

Model pada Jakarta menunjukkan performa lebih tinggi dibandingkan Tangerang. Terdapat beberapa alasan utama:

1. Variasi polusi Jakarta lebih kompleks, sehingga Gradient Boosting dan Random Forest menunjukkan perbedaan performansi yang lebih signifikan.
2. Data Tangerang lebih stabil, membuat nilai R^2 keduanya relatif serupa.
3. Pada kedua kota, Random Forest secara konsisten memberikan MAE yang lebih rendah, menjadikannya model utama dalam penelitian ini.



Gambar 5. Perbandingan Nilai R^2 Antar Model

Perbedaan karakteristik ini menegaskan bahwa pola polusi Jakarta banyak dipengaruhi aktivitas transportasi dan industri intensif, sedangkan Tangerang memiliki pola yang lebih terstruktur.

4.5 Analisis Rata-Rata PM2.5 Jakarta dan Tangerang

Rata-rata PM2.5 dari hasil perhitungan menunjukkan:

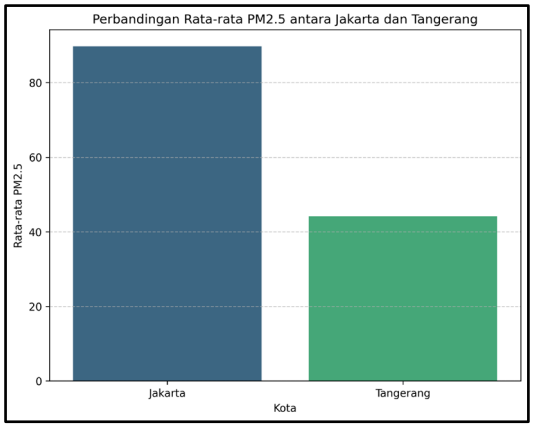
Tabel 3. Perbandingan RF, dan GB Taggerang

Kota	Rata-rata	Maksumim	Minumum

Jakarta	89.74	287.00	10.00
Tangge rang	44.17	66.00	19.00

Jakarta menunjukkan nilai PM2.5 hampir dua kali lipat lebih tinggi dibandingkan Tangerang. Temuan ini mendukung hasil studi Rini et al. [6], yang menyebutkan bahwa tingginya polusi Jakarta dipengaruhi oleh:

1. intensitas kendaraan harian,
2. aktivitas industri dan konstruksi,
3. penggunaan kendaraan diesel,
4. kepadatan urban Jabodetabek.



Gambar 6. Rata-rata PM2.5 Jakarta vs Tangerang

5. Kesimpulan

Berdasarkan rangkaian pemodelan kualitas udara menggunakan algoritma Random Forest, Gradient Boosting, dan Linear Regression, diperoleh beberapa kesimpulan sebagai berikut:

1. Random Forest menjadi model dengan performa paling unggul dibandingkan dua algoritma lainnya. Model ini menghasilkan error yang lebih rendah dan prediksi yang lebih stabil pada data polutan yang bersifat dinamis. Hasil tersebut konsisten dengan penelitian sebelumnya yang menekankan ketangguhan Random Forest untuk data lingkungan yang memiliki pola non-linear [1][3].
2. Gradient Boosting memberikan hasil yang kompetitif, namun lebih sensitif terhadap noise sehingga stabilitasnya sedikit lebih rendah. Meski demikian, model ini tetap dapat menangkap hubungan antarpolutan dengan cukup akurat sebagaimana dijelaskan dalam studi Zhao et al. [4].

3. Linear Regression menunjukkan performa terendah, terutama karena pendekatan linier tidak mampu memetakan interaksi kompleks antar-polutan seperti PM2.5, PM10, SO_2, NO_2, CO, dan O_3. Temuan ini selaras dengan literatur yang menyatakan bahwa data polusi udara umumnya membutuhkan model non-linear untuk mencapai presisi optimal [2][5].
4. Penelitian ini membuktikan bahwa pendekatan machine learning efektif digunakan untuk memprediksi kualitas udara secara cepat dan presisi. Prediksi yang akurat berpotensi mendukung pengambilan keputusan pada sektor kesehatan masyarakat, lingkungan, serta perumusan kebijakan berbasis data.
5. Secara keseluruhan, Random Forest direkomendasikan sebagai model paling sesuai untuk prediksi Air Quality Index (AQI) pada studi ini, terutama untuk wilayah dengan karakteristik polusi yang fluktuatif seperti Jakarta.

REFERENSI

[1] Gupta, P., et al. (2021). *Machine Learning Models for PM2.5 Prediction in Urban Regions*. Atmospheric Environment.
[2] Hu, J., & Ying, Q. (2020). *Air Quality Modeling Using Ensemble Learning Techniques*. Environmental Pollution.
[3] Feng, Y., et al. (2022). *Random Forest-Based PM2.5 Forecasting in Metropolitan Areas*. Environmental Research.
[4] Zhao, X., et al. (2023). *Gradient Boosting for Air Pollutant Prediction*. Journal of Cleaner Production.
[5] Sun, L., et al. (2021). *Predictive Modeling of PM10–PM2.5 Interactions Using Machine Learning*. Atmospheric Pollution Research.
[6] Rini, D., et al. (2020). *Studi Polusi Udara Jakarta Menggunakan Analisis Statistik*. Jurnal Teknologi Lingkungan.
[7] Chen, Z., et al. (2021). *Missing Data Imputation in Air Quality Monitoring Systems*. Environmental Modelling & Software.
[8] Li, M., et al. (2022). *Temporal Feature Engineering for Air Pollution Forecasting*. Environmental Science & Technology.
[9] Wang, S., et al. (2024). *Feature Selection Methods for Atmospheric Pollutant Prediction*. Applied Sciences.
[10] Putra, H., et al. (2023). *Evaluasi Kualitas Udara Jabodetabek Menggunakan Pembelajaran Mesin*. Indonesian Journal of Computing.