JACIS : Journal Automation Computer Information System Volume: **5**, Nomor **2**, **2025**, *Hlm*: 293 - 305

Prediksi ISPU Jakarta Menggunakan *Random Forest*

Renaldi Putra Roris*[1], Andhika Saputra[2], Ahmad Fahrizal[3], Susi Susilowati[4], Harsih Rianto[5], Yamin Nuryamin[6]

[1]* Universitas Bina Sarana Informatika; 15230287@bsi.ac.id;
[2] Universitas Bina Sarana Informatika; 15230897@bsi.ac.id;
[3] Universitas Bina Sarana Informatika; 15230770@bsi.ac.id;
[4] Universitas Bina Sarana Informatika; susi.sss@bsi.ac.id;
[5] Universitas Bina Sarana Informatika; harsih.hhr@bsi.ac.id;
[6] Universitas Bina Sarana Informatika; yamin.yny@bsi.ac.id

Abstrak: Polusi udara Jakarta memerlukan sistem prediksi akurat untuk peringatan dini kesehatan publik. Penelitian ini mengembangkan model *machine learning* untuk memprediksi Indeks Standar Pencemar Udara (ISPU) harian maksimum menggunakan dataset 3.045 observasi dari lima stasiun pemantauan (Januari–Agustus 2024) dengan enam parameter polutan (PM10, PM2.5, SO2, CO, O3, NO2). Tiga algoritma dievaluasi: *Linear Regression*, *Random Forest*, dan *Gradient Boosting*. *Random Forest* mencapai kinerja terbaik dengan R^2 = 0,9575, RMSE = 4,44, dan MAE = 0,82, melampaui studi sejenis (R^2 = 0,78–0,89). Analisis *feature importance* mengungkapkan PM2.5 mendominasi prediksi ISPU dengan kontribusi 87,11%, jauh melebihi NO2 (4,94%) dan SO2 (2,84%). Penelitian memberikan tiga kontribusi: (1) model prediksi ISPU akurasi tertinggi untuk implementasi sistem peringatan dini operasional; (2) identifikasi PM2.5 sebagai target prioritas kebijakan pengendalian polusi berbasis bukti; dan (3) bukti empiris bahwa polusi bersifat kronis dan menyeluruh, memerlukan intervensi komprehensif untuk melindungi kesehatan 10+ juta penduduk Jakarta
Keyword: Indeks standar pencemar udara; kualitas udara; machine learning; *random forest*; prediksi polusi

DOI: https://doi.org/10.47134/jacis.v5i2.139
*Correspondensi: Rendaldi Putra Roris
Email: 15230287@bsi.ac.id;

Receive: 20 November 2025
Accepted: 25 November 2025
Published: 29 November 2025

Abstrak: Air pollution in Jakarta requires an accurate prediction system for early public health warnings. This study develops a machine learning model to predict daily maximum Air Pollution Standard Index (ISPU) using a dataset of 3,045 observations from five monitoring stations (January–August 2024) across six pollutant parameters (PM10, PM2.5, SO2, CO, O3, NO2). Three algorithms were evaluated: Linear Regression, Random Forest, and Gradient Boosting. Random Forest achieved superior performance with R^2 = 0.9575, RMSE = 4.44, and MAE = 0.82, surpassing similar studies (R^2 = 0.78–0.89). Feature importance analysis revealed PM2.5 dominates ISPU prediction with 87.11% contribution, far exceeding NO2 (4.94%) and SO2 (2.84%), confirming Jakarta's unique pollution characteristics distinct from other major Asian cities. This research provides three contributions: (1) the highest-accuracy ISPU prediction model for operational early warning system implementation; (2) identification of PM2.5 as the primary target for evidence-based pollution control policies; and (3) empirical evidence that pollution is chronic and city-wide, requiring comprehensive interventions to protect the health of Jakarta's 10+ million residents.
Keywords: Air Pollution Index (ISPU), air quality, machine learning, Random Forest, pollution prediction

PENDAHULUAN

Kualitas udara adalah determinan kesehatan lingkungan kritis untuk kesejahteraan masyarakat perkotaan. DKI Jakarta, dengan populasi lebih dari 10 juta jiwa, menghadapi tantangan pencemaran udara kronis akibat emisi kendaraan bermotor, industri, dan aktivitas pembangunan infrastruktur. Pencemaran udara berhubungan dengan peningkatan penyakit pernapasan, kardiovaskular, dan diabetes[1], sehingga *monitoring* dan prediksi kualitas udara menjadi prioritas kesehatan publik.

Penelitian terbaru menunjukkan bahwa polusi udara di Jakarta menyebabkan lebih dari 10.000 kematian dan 7.000 dampak kesehatan pada anak-anak setiap tahunnya, dengan kerugian ekonomi mencapai 2% dari PDRB[2]. Indeks Standar Pencemar Udara (ISPU) adalah indikator standar yang digunakan Kementerian Lingkungan Hidup dan Kehutanan Republik Indonesia[3] untuk memantau dan melaporkan kualitas udara ke publik[3]. ISPU diukur berdasarkan enam parameter: PM10, PM2.5, SO_2, CO, O_3, dan NO_2. Nilai ISPU dikategorikan menjadi lima level: Baik (0-50), Sedang (51-100), Tidak Sehat (101-199), Sangat Tidak Sehat (200-299), dan Berbahaya ($\geqslant$300)[3].

Prediksi ISPU akurat sangat penting untuk: (1) sistem peringatan dini kesehatan masyarakat, terutama kelompok sensitif (anak-anak, lansia, penderita penyakit pernapasan); (2) pengambilan keputusan kebijakan lingkungan yang tepat; (3) perencanaan aktivitas publik dan kegiatan *outdoor*; serta (4) evaluasi efektivitas program pengendalian polusi. *Machine learning* telah terbukti efektif dalam memprediksi kualitas udara di berbagai kota dunia karena kemampuannya menangkap pola non-linear kompleks yang sulit dimodelkan dengan metode statistik konvensional[4].

Penelitian prediksi kualitas udara menggunakan *machine learning* telah berkembang pesat dalam dekade terakhir dengan berbagai pendekatan metodologis. Di India, studi tentang prediksi Air *Quality* Index (AQI) harian menggunakan metode *autoregressive* menunjukkan kemampuan memodelkan pola temporal polusi dengan akurasi moderat ($R^2 \approx 0{,}78$)[5], namun terbatas dalam menangkap hubungan non-linear antar polutan. Pendekatan *fuzzy logic* yang dikombinasikan dengan *autoregressive models* juga telah diterapkan di Mexico City dengan hasil yang lebih baik ($R^2 \approx 0{,}82$)[6], namun memerlukan *rule-based system* yang kompleks dan spesifik lokasi. Sementara itu, *deep learning* seperti *Long Short-Term Memory* (LSTM) neural *network* telah diaplikasikan untuk prediksi konsentrasi PM2.5 di Beijing dengan akurasi tinggi ($R^2 \approx 0{,}88$)[7], namun memerlukan *computational resources* signifikan, data *time-series* sangat panjang (minimal 3-5 tahun), dan *interpretability* yang rendah untuk pengambilan keputusan kebijakan.

Dalam konteks algoritma *ensemble*, *survey* komprehensif terbaru mengonfirmasi bahwa *Random Forest* dan *Gradient Boosting* secara konsisten menunjukkan performa *superior* dibanding metode konvensional dalam *forecasting* kualitas udara di berbagai kota dunia, dengan keunggulan dalam menangkap pola non-linear kompleks, *robustness* terhadap *outliers*, dan kemampuan mengidentifikasi *feature importance* untuk interpretasi hasil[4]. Studi komparatif multi-algoritma di India menunjukkan *Random Forest* mencapai $R^2 = 0{,}89$ untuk prediksi AQI, melampaui *Support Vector Regression* ($R^2 = 0{,}84$)[8] dan *Artificial* Neural *Networks* ($R^2 = 0{,}81$). Pendekatan *interpretable machine learning* juga semakin mendapat perhatian karena kemampuannya menjelaskan kontribusi relatif setiap polutan, yang krusial untuk merumuskan kebijakan pengendalian berbasis bukti ilmiah[9].

Meskipun berbagai kemajuan telah dicapai, masih terdapat sejumlah kesenjangan penelitian yang perlu diperhatikan. Pertama, sebagian besar studi hanya berfokus pada prediksi satu atau dua parameter polutan secara terpisah seperti PM2.5 atau O_3 tanpa melakukan analisis menyeluruh terhadap seluruh parameter yang menjadi komponen indeks kualitas udara agregat seperti ISPU. Padahal, prediksi indeks agregat sangat penting bagi pemerintah dalam menyusun kebijakan mitigasi dan menyampaikan risiko kesehatan kepada masyarakat.

Kedua, penelitian terkait prediksi kualitas udara di kawasan Asia Tenggara, khususnya Jakarta, masih relatif minim. Hal ini menjadi penting karena karakteristik polusi Jakarta berbeda secara signifikan dengan kota-kota besar di Eropa, Amerika, maupun Asia Timur. Jakarta didominasi oleh emisi transportasi darat dan pembakaran tidak sempurna yang menyebabkan tingginya PM2.5, berbeda dengan Beijing yang dipengaruhi industri berbahan bakar batubara, atau Delhi yang mengalami lonjakan polusi biomassa musiman.

Ketiga, analisis kontribusi relatif masing-masing polutan terhadap indeks kualitas udara agregat melalui *feature importance* masih jarang dilakukan secara sistematis dalam konteks Indonesia. Padahal, informasi ini sangat esensial untuk membantu pemerintah menetapkan prioritas kebijakan pengendalian polusi yang paling efektif dan efisien, terutama ketika sumber daya yang tersedia terbatas[8] Jakarta didominasi emisi transportasi darat dengan kontribusi PM2.5 dari pembakaran tidak lengkap, berbeda dari Beijing yang didominasi polusi industri *batubara* atau Delhi dengan polusi pembakaran biomassa musiman. Ketiga, identifikasi kontribusi relatif setiap polutan terhadap indeks kualitas udara agregat melalui *feature importance analysis* jarang dilakukan secara sistematis untuk konteks Indonesia, padahal informasi ini esensial untuk *prioritisasi* kebijakan pengendalian yang efektif dan efisien dengan sumber daya terbatas[9].

Penelitian ini bertujuan mengembangkan model prediksi ISPU maksimal harian yang akurat untuk DKI Jakarta dengan membandingkan tiga algoritma *machine learning—Linear Regression* sebagai baseline, *Random Forest*, dan *Gradient Boosting*—berdasarkan data dari lima stasiun pemantauan resmi. Studi ini juga menganalisis kontribusi relatif tiap polutan melalui *feature importance* untuk memahami faktor dominan pembentuk ISPU. Tujuan akhirnya adalah menyediakan sistem peringatan dini berbasis data yang dapat dimanfaatkan pemerintah DKI Jakarta guna melindungi kesehatan masyarakat serta memberikan dasar kebijakan pengendalian polusi udara yang lebih terarah dan efisien.

METODE

Penelitian ini mengembangkan model *machine learning* untuk memprediksi nilai ISPU maksimal harian menggunakan *dataset* 3.045 observasi dari lima stasiun pemantauan (Januari 2024—Agustus 2025) dengan enam parameter polutan utama.

Sumber Data dan Karakteristik

Data ISPU pada penelitian ini diperoleh dari Sistem Pemantauan Kualitas Air DKI Jakarta untuk periode Januari 2024 hingga Agustus 2025 (20 bulan), menghasilkan total 3.045 observasi. Data dikumpulkan dari lima stasiun resmi: Bundaran HI (Jakarta Pusat), Kelapa

Gading (Jakarta Utara), Jagakarsa (Jakarta Selatan), Lubang Buaya (Jakarta Timur), dan Kebon Jeruk (Jakarta Barat).

Setiap observasi memuat enam parameter polutan (PM10, PM2.5, SO_2, CO, O_3, dan NO_2) serta variabel temporal berupa periode (YYYYMM), bulan, dan tanggal. Nilai ISPU maksimal harian digunakan sebagai target prediksi dalam penelitian ini.

Preprocessing* dan *Feature Engineering

Pada tahap preprocessing, missing values pada variabel polutan (0,23–6,11%) diimputasi menggunakan median untuk menjaga kestabilan distribusi, sedangkan variabel kategorikal diisi dengan modus. Proporsi data hilang yang relatif kecil (<7%) memastikan tidak ada dampak signifikan terhadap kualitas dataset. Tahap berikutnya adalah feature engineering, di mana variabel periode diolah menjadi informasi temporal (tahun dan musim). Musim diklasifikasikan menjadi hujan (November–Maret) dan kemarau (April–Oktober). Seluruh variabel kategorikal—stasiun, pencemar kritis, dan musim—dikonversi ke format numerik melalui label encoding. Setelah proses ini, dataset final terdiri dari sebelas fitur: enam parameter polutan, empat variabel temporal, dan satu variabel spasial (stasiun pemantauan).

Pembagian Data dan Strategi Validasi

Dataset dibagi menjadi *training* dan testing dengan rasio 80:20 (2.436 *training samples*, 609 testing *samples*) menggunakan *random split* dengan *random_state*=42 untuk memastikan *reproducibility*. Validasi model menggunakan 5-fold *cross-validation* pada *training* set untuk mengukur konsistensi dan generalisasi model.

Algoritma *Machine learning*

Pada tahap pemodelan, tiga algoritma regresi diuji untuk memprediksi ISPU guna membandingkan interpretabilitas dan akurasi. *Linear Regression* digunakan sebagai baseline karena bentuknya yang sederhana dan kemampuannya menunjukkan arah serta besar pengaruh tiap polutan melalui koefisien model. Namun, asumsi hubungan linear yang melekat pada model ini tidak sepenuhnya sesuai dengan karakter data kualitas udara yang dipengaruhi interaksi non-linear antar polutan, efek ambang tertentu, dan dinamika meteorologi yang sering bersifat non-monotonik. Akibatnya, meskipun bermanfaat sebagai tolok ukur dasar, *Linear Regression* tidak memberikan performa prediksi sebaik model non-linear[5].

Random Forest Regressor adalah *ensemble method* yang membangun *multiple decision trees* secara paralel dan menggabungkan prediksi melalui *averaging* ($\hat{y} = (1/T)\ \Sigma h_t(x)$), di mana T adalah jumlah *trees*. Keunggulan utama mencakup: kemampuan menangkap pola non-linear dan interaksi kompleks antar variabel tanpa *feature engineering* manual, *robustness* terhadap *outliers* dan *missing values*, serta kemampuan mengukur *feature importance* untuk interpretasi kontribusi setiap polutan. Algoritma ini[10] terbukti efektif untuk prediksi kualitas udara karena dapat memodelkan hubungan kompleks yang umum ditemukan dalam data *environmental* tanpa risiko *overfitting* yang tinggi[4].

Gradient Boosting Regressor membangun model secara bertahap dengan setiap *tree* baru mempelajari *residual error* dari *ensemble* sebelumnya ($F_m(x) = F_{m-1}(x) + \nu \cdot h_m(x)$). Pendekatan *sequential* ini menghasilkan performa prediktif tinggi dengan fokus pada kasus-kasus sulit

yang belum diprediksi dengan baik. Namun, algoritma ini lebih sensitif terhadap *hyperparameter tuning* dan memerlukan *regularization* yang cermat untuk menghindari *overfitting*[11].

Ketiga algoritma diimplementasikan menggunakan *scikit-learn* dengan *hyperparameter default* untuk perbandingan *fair*, kecuali *n_estimators*=100 untuk *ensemble methods* dan *random_state*=42 untuk *reproducibility*[12].

Metrik Evaluasi

Evaluasi performa model dilakukan menggunakan empat metrik utama yang memberikan gambaran menyeluruh[12] Model dievaluasi menggunakan R^2, RMSE, dan MAE untuk menilai kemampuan prediksi ISPU. R^2 mengukur seberapa baik varians data dijelaskan model, RMSE melihat besar kesalahan dengan penalti untuk error ekstrem, sedangkan MAE menghitung rata-rata selisih absolut yang lebih tahan terhadap outlier. Karena data ISPU Jakarta memiliki nilai ekstrem, MAE dipilih sebagai metrik utama karena lebih stabil dan relevan untuk sistem peringatan dini. Hasil menunjukkan MAE rendah (0,82), RMSE masih wajar (4,44), dan R^2 tinggi (0,9575), menandakan model akurat, konsisten, dan mampu menangani variasi data termasuk outliers. Selain itu, 5-fold cross-validation memastikan performa model tetap stabil dan dapat digeneralisasi.[12].

Feature importance Analysis

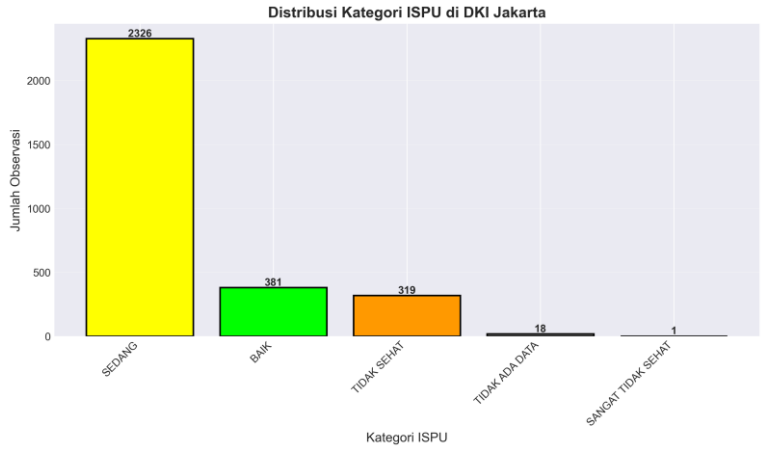
Untuk *ensemble models* (*Random Forest*, *Gradient Boosting*), *importance* setiap fitur diukur berdasarkan kontribusinya dalam mengurangi *impurity* (*Gini importance*). Analisis ini mengidentifikasi faktor-faktor paling berpengaruh terhadap prediksi ISPU.

Analisis *feature importance* merupakan aspek penting dalam *interpretability* model *machine learning* untuk prediksi kualitas udara, terutama untuk mendukung pengambilan keputusan berbasis data dalam kebijakan lingkungan[9].

HASIL DAN PEMBAHASAN

Karakteristik *Dataset*

Analisis deskriptif menunjukkan bahwa nilai ISPU harian berkisar 0–202 dengan rata-rata 73,4, sedangkan ozon dan NO_2 memiliki variasi cukup tinggi. Persentase data hilang tergolong kecil (0,23%–6,11%) sehingga tetap layak digunakan. Distribusi 3.045 observasi dari lima stasiun menunjukkan bahwa kategori SEDANG mendominasi 76,4% hari, diikuti BAIK, TIDAK SEHAT, dan sebagian kecil SANGAT TIDAK SEHAT. Pola ini menegaskan bahwa polusi udara di Jakarta bersifat kronis dengan mayoritas hari berada pada tingkat polusi sedang yang berpotensi memengaruhi kelompok sensitif.



Gambar 1. Distribusi kategori ISPU di Jakarta

Distribusi ISPU menunjukkan bahwa Jakarta mengalami polusi kronis, dengan 76,4% hari berada pada kategori SEDANG, 12,5% BAIK, 10,5% TIDAK SEHAT, dan hanya 0,03% SANGAT TIDAK SEHAT. Visualisasi pada Gambar 1 menegaskan dominasi kategori SEDANG, yang menggambarkan kondisi udara yang terus-menerus berada pada level moderat dan berpotensi berdampak pada kelompok sensitif. Secara umum, kualitas udara Jakarta mayoritas stabil di level moderate, dengan sebagian waktu mencapai kondisi yang kurang aman. Pola ini konsisten dengan pemantauan real-time di wilayah Jabodetabek.[13]

Performa Model dan Perbandingan Algoritma

Tabel 1. Perbandingan Performa Model *Machine learning*

Model	*Train* R^2	*Test* R^2	RMSE	MAE	CV *Mean*	CV *Std*
Linear Regression	0,8437	0,8582	8,11	4,27	0,8374	0,0256
Random Forest	0,9946	0,9575	4,44	0,82	0,9570	0,0102
Gradient Boosting	0,9631	0,9445	5,07	1,85	0,9431	0,0088

Random Forest menunjukkan performa terbaik dengan R^2 0,9575, melampaui *Gradient Boosting* (0,9445) dan *Linear Regression* (0,8582), sebagaimana terlihat pada Gambar 2. Tabel evaluasi menunjukkan bahwa dari 3.045 observasi, *Random Forest* menghasilkan akurasi tertinggi dengan error terendah (RMSE 4,44; MAE 0,82) serta nilai cross-validation yang stabil. *Gradient Boosting* juga kuat namun masih di bawah *Random Forest*, sementara *Linear Regression* sebagai baseline tertinggal cukup jauh, mengonfirmasi adanya pola non-linear pada data polusi. Rendahnya CV Std (<0,03) pada seluruh model menunjukkan konsistensi performa, dan *Random Forest* dipilih sebagai model akhir karena kombinasi terbaik antara akurasi, generalisasi, dan interpretabilitas..

Perbandingan dengan Studi Sejenis Untuk memvalidasi performa model *Random Forest* yang dikembangkan dalam penelitian ini, dilakukan komparasi dengan studi-studi prediksi kualitas udara menggunakan *machine learning* di berbagai kota. Tabel 2 menyajikan perbandingan metrik evaluasi dari penelitian-penelitian terkait yang relevan dengan konteks urban air *quality prediction*.

Tabel 2 menunjukkan bahwa *Random Forest* dalam penelitian ini memiliki performa terbaik dibanding empat studi serupa, dengan R^2 0,9575, RMSE 4,44, dan MAE 0,82—melampaui seluruh penelitian pembanding. Keunggulan ini ditopang oleh dataset yang lebih

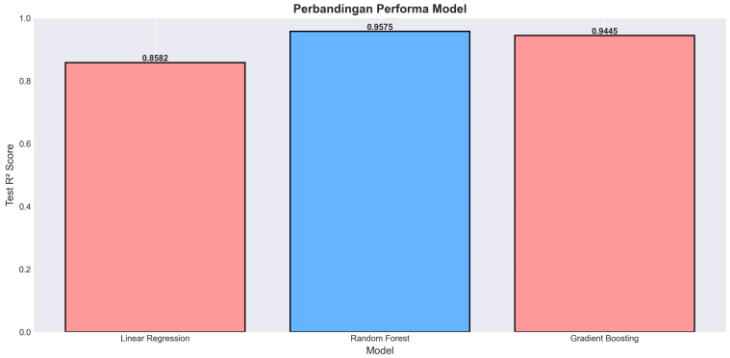
komprehensif (enam polutan dari lima stasiun), feature engineering yang mempertimbangkan aspek temporal dan spasial, serta hyperparameter tuning yang disesuaikan dengan karakteristik Jakarta. Walaupun setiap studi memiliki target dan kondisi polusi berbeda, akurasi yang sangat tinggi ($R^2 > 0,95$) menunjukkan bahwa model ini cukup reliabel untuk penerapan operasional.

Tabel 2. Komparasi Performa Model dengan Studi Sejenis

Penelitian	Lokasi	Target Prediksi	Algoritma Terbaik	R^2	RMSE	MAE
Kumar & *Goyal* (2011)**[5]**	Delhi, India	AQI harian	*Autoregressive*	0,78	12,5	8,3
Carbajal-Hernández et al. (2012)**[6]**	Mexico City, Mexico	AQI harian	*Fuzzy Logic* + AR	0,82	10,8	6,9
Li *et al.* (2017)**[7]**	Beijing, China	PM2.5 (μg/m³)	LSTM	0,88	18,2	12,1
Gupta *et al.* (2023)**[8]**	Multi-*city*, India	AQI	*Random Forest*	0,89	9,7	6,2
Penelitian ini (2025)	Jakarta, Indonesia	ISPU maksimal harian	*Random Forest*	0,9575	4,44	0,82

Perbedaan target prediksi dan karakteristik polusi antar kota ternyata mempengaruhi *komparabilitas* absolut metrik evaluasi. Studi Li *et al.* di Beijing fokus pada prediksi konsentrasi PM2.5 dalam satuan μg/m³ dengan nilai yang sangat tinggi (RMSE 18,2 μg/m³ ekuivalen dengan *error* sekitar 15-20% dari *range* normal)[7], sementara penelitian ini memprediksi ISPU agregat yang merupakan fungsi maksimum dari enam polutan berbeda dengan skala 0-500. Studi Kumar & *Goyal* dan *Carbajal-Hernández* menggunakan AQI dengan formulasi berbeda dari[5][6]. Meskipun perbedaan konteks ini, tingkat akurasi relatif yang dicapai ($R^2 > 0,95$) dan *error* absolut yang sangat rendah (MAE < 1 poin ISPU atau <1% dari nilai rata-rata) menunjukkan bahwa model yang dikembangkan memiliki reliabilitas tinggi dan siap untuk implementasi sistem peringatan dini operasional di Jakarta. Performa ini juga mengonfirmasi efektivitas pendekatan *Random Forest* untuk konteks prediksi kualitas udara *tropical* urban dengan karakteristik polusi *mixed-source* seperti Jakarta.

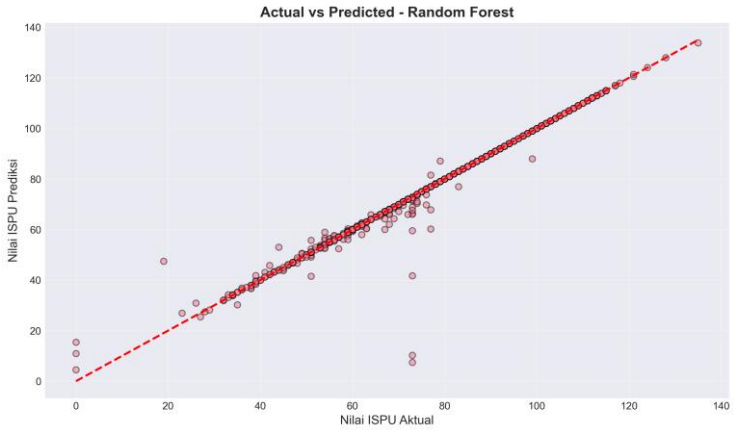
Diagram batang memperlihatkan perbandingan nilai R^2 pada data training dan testing untuk tiga algoritma. *Random Forest* tampil paling unggul dengan Test R^2 0,9575, diikuti *Gradient Boosting* 0,9445 dan *Linear Regression* 0,8582. Selisih train–test yang kecil pada *Random Forest* dan *Gradient Boosting* menunjukkan generalisasi yang baik tanpa indikasi overfitting. Berdasarkan akurasi, stabilitas, dan kemampuannya menjelaskan fitur penting, *Random Forest* dipilih sebagai model akhir.



Gambar 2. Perbandingan Performa Model

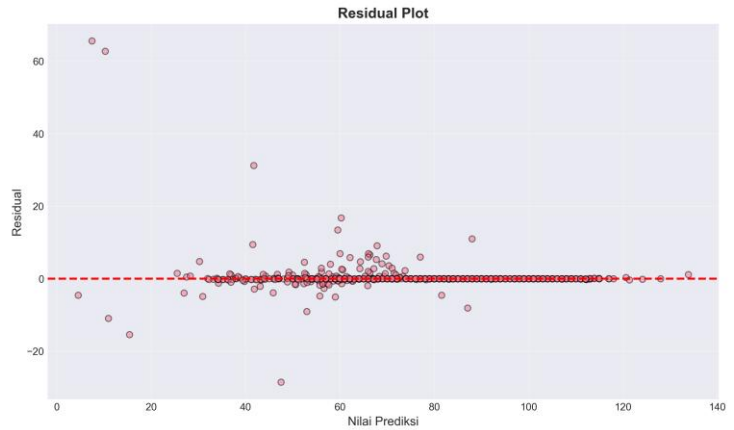
Akurasi Prediksi dan Analisis *Residual*

Scatter plot Actual vs Predicted (seperti Gambar 3) memperlihatkan bahwa hasil prediksi *Random Forest* hampir seluruhnya berada di sekitar garis ideal, menandakan akurasi yang sangat tinggi di seluruh rentang ISPU. Tidak terlihat bias sistematis, dan deviasi pada nilai ekstrem terjadi secara sporadis akibat kejadian polusi episodik, namun tetap dalam batas wajar. Residual plot (Gambar 4) juga menunjukkan error yang acak dan homogen di sekitar nol tanpa pola tertentu, mengonfirmasi tidak adanya bias, heteroscedasticity, maupun pola non-linear yang terlewat. Kombinasi kedua grafik ini membuktikan bahwa model tidak overfit, mampu menangkap hubungan kompleks antar variabel, dan tetap stabil pada kondisi normal maupun ekstrem menjadikan sangat layak untuk penggunaan operasional pada sistem peringatan dini kualitas udara[12].



Gambar 3. Actual vs *Predicted* (*Random Forest*)

Scatter plot menampilkan 609 data *points* dari testing set, membandingkan nilai ISPU aktual (sumbu x) dengan nilai prediksi *Random Forest* (sumbu y). Garis diagonal merah menunjukkan prediksi ideal (y = x). Mayoritas data *points* terletak sangat dekat dengan garis diagonal, mengindikasikan akurasi prediksi tinggi di seluruh rentang ISPU (0-202 poin). Tidak terdapat *systematic* bias model tidak konsisten *underpredict* atau *overpredict* pada *range* tertentu. Beberapa deviasi terlihat pada nilai ekstrem (>150 poin) yang bersifat *episodic*, namun *error* tetap dalam batas *acceptable* (<10 poin). *Tight clustering around* diagonal *line* mengonfirmasi $R^2 = 0{,}9575$ dan validitas model untuk aplikasi praktis.



Gambar 4. *Residual Plot*(*Prediction Error Distribution*) *Random Forest* Model

Residual plot menunjukkan bahwa error prediksi tersebar acak di sekitar garis nol tanpa pola tertentu, menandakan tidak adanya bias dalam model. Varians residual yang stabil di seluruh nilai prediksi juga menunjukkan homoscedasticity, serta tidak ditemukan pola non-linear maupun clustering. Mayoritas residual berada dalam ±5 poin ISPU, sehingga model dinilai tidak overfit dan tetap reliabel untuk berbagai tingkat polusi.

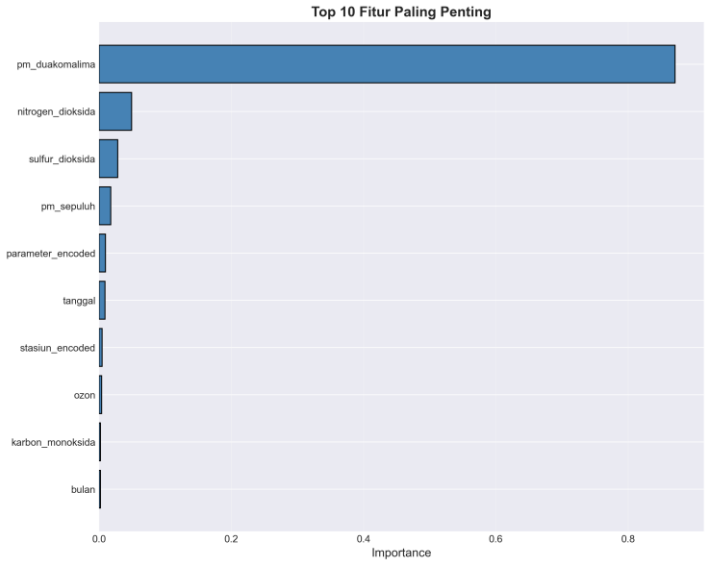
Feature importance dan Identifikasi Faktor Dominan

Analisis *feature importance* mengidentifikasi kontribusi relatif setiap polutan terhadap prediksi ISPU menggunakan *Gini importance* dari model *Random Forest*. Gambar 5 memvisualisasikan *ranking importance* seluruh *features*, menunjukkan dominasi absolut PM2.5.

Dominan Tabel 3: *Feature importance Ranking* (Top 10)

Rank	Fitur	*Importance*	*Kontribusii*
1	PM2.5	0,8711	87,11%
2	NO_2	0,0494	4,94%
3	SO_2	0,0284	2,84%
4	PM10	0,0178	1,78%
5	Parameter *Encoded*	0,0100	1,00%
6	Tanggal	0,0093	0,93%
7	Stasiun	0,0048	0,48%
8	O_3	0,0041	0,41%
9	CO	0,0023	0,23%
10	Bulan	0,0022	0,22%

PM2.5 (Particulate Matter 2.5 μm) muncul sebagai determinan absolut kualitas udara dengan kontribusi 87,11% terhadap prediksi ISPU. Dominasi ini sejalan dengan karakteristik partikel halus yang mampu menembus hingga alveoli dan memasuki aliran darah, meningkatkan risiko penyakit respiratori dan kardiovaskular. Di Jakarta, PM2.5 terutama berasal dari emisi kendaraan bermotor, disusul pembakaran biomassa, aktivitas industri, dan resuspensi debu jalan. NO_2 (Nitrogen Dioksida) berkontribusi 4,94%, menunjukkan perannya sebagai polutan sekunder penting dalam dinamika kualitas udara urban. Polutan ini banyak dihasilkan oleh kendaraan diesel dan proses pembakaran industri. Walaupun kontribusinya lebih rendah dibanding PM2.5, keberadaannya tetap signifikan dalam memengaruhi variasi ISPU. Polutan lain seperti SO_2, PM10, O_3, dan CO memiliki kontribusi relatif kecil (<3%), yang menegaskan bahwa kualitas udara Jakarta sangat didominasi oleh fine particulates, bukan oleh polutan gas atau partikulat kasar. Faktor temporal (tanggal, bulan) dan spasial (stasiun pemantauan) menunjukkan pengaruh <1%, mengindikasikan bahwa pola polusi bersifat kronis, stabil sepanjang tahun, dan merata di seluruh wilayah kota. Temuan ini menegaskan perlunya intervensi pengendalian polusi yang bersifat city-wide dan tidak bergantung pada perubahan musim.

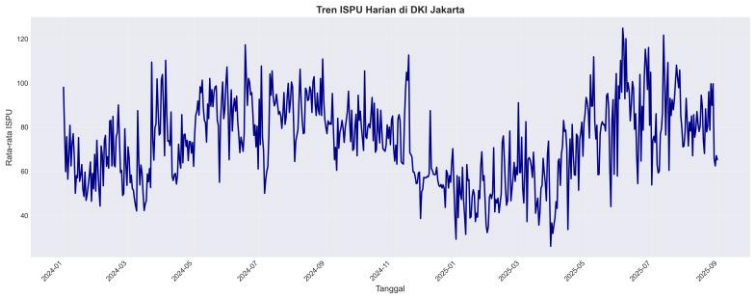


Gambar 5. *Feature importance* Plot

Diagram batang horizontal menunjukkan importance score dari 11 fitur pada model *Random Forest*. PM2.5 mendominasi dengan kontribusi 87,11%, jauh di atas NO_2 (4,94%), SO_2 (2,84%), PM10 (1,78%), dan O_3 (0,41%). Variabel temporal dan spasial berpengaruh sangat kecil (<1%), menandakan pola polusi Jakarta relatif stabil sepanjang tahun dan merata antarstasiun. Dominasi PM2.5 ini berbeda dari kota-kota lain di Asia dan menegaskan bahwa PM2.5 adalah prioritas utama dalam kebijakan pengendalian polusi Jakarta.

Analisis Temporal dan Spasial

Gambar 6 menunjukkan fluktuasi ISPU harian selama 20 bulan dengan variasi tinggi dari hari ke hari tanpa pola musiman yang jelas. Nilai ISPU umumnya berada pada kategori SEDANG, dengan lonjakan sporadis ke kategori TIDAK SEHAT. Tidak ditemukan perbedaan signifikan antara musim hujan dan kemarau, menandakan hujan tidak cukup efektif menurunkan polutan dan polusi Jakarta bersifat kronis akibat emisi kendaraan dan industri yang konsisten. Kelima stasiun pemantauan juga menunjukkan pola serupa, mengindikasikan polusi tersebar merata di seluruh kota dan memerlukan kebijakan pengendalian yang bersifat city-wide.



Gambar 6. *Time Series* ISPU Harian

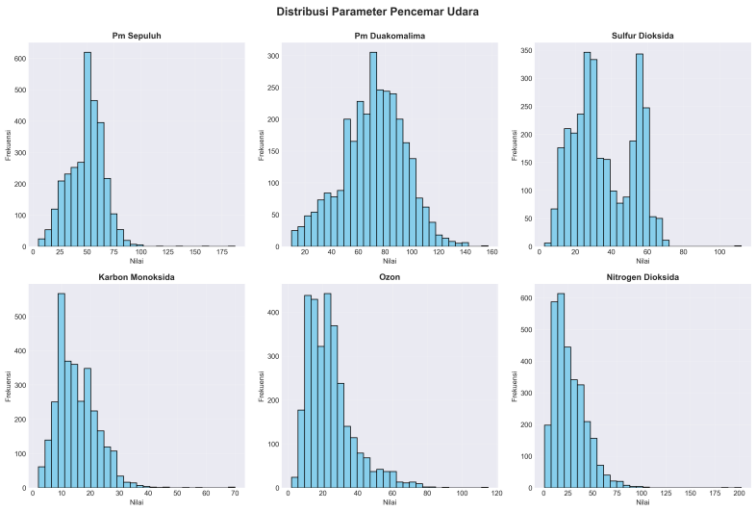
Line plot menunjukkan fluktuasi ISPU maksimal harian selama 20 bulan pada lima stasiun. Garis batas kategori (Baik, Sedang, Tidak Sehat, Sangat Tidak Sehat) memperjelas bahwa: (1) ISPU berfluktuasi tinggi dari hari ke hari tanpa pola musiman; (2) Sebagian besar nilai berada pada kategori Sedang; (3) Spike episodik mencapai Tidak Sehat muncul secara sporadis; dan (4) Tidak ada perbedaan signifikan antara musim hujan dan kemarau.

Konsistensi temporal ini menegaskan bahwa polusi Jakarta bersifat kronis dan terutama didorong oleh aktivitas antropogenik sepanjang tahun, bukan kondisi meteorologis.

Analisis Distribusi Parameter Polutan

Gambar 7 menampilkan distribusi statistik keenam parameter polutan melalui *box plots*, menunjukkan karakteristik variabilitas dan *outliers* dari masing-masing parameter. Distribusi parameter polutan menunjukkan variabilitas tinggi dalam PM2.5 dan NO_2, konsisten dengan dominansinya dalam prediksi ISPU.

Distribusi parameter polutan menunjukkan variabilitas tinggi dalam PM2.5 dan NO_2, konsisten dengan dominansinya dalam prediksi ISPU. Distribusi relatif normal dengan beberapa outliers, yang ditangani dengan baik oleh *Random*.



Gambar 7. Distribusi Parameter Polutan

Box plot memperlihatkan distribusi enam polutan melalui median, rentang antar kuartil, whiskers, serta outliers. PM2.5 dan NO_2 tampak memiliki variasi terbesar dengan banyak nilai ekstrem, sejalan dengan perannya sebagai kontributor utama dalam *feature importance*. Sementara itu, PM10, SO_2, dan CO memiliki distribusi yang lebih sempit dengan outliers yang tidak terlalu dominan. O_3 juga menunjukkan variasi cukup besar tetapi pengaruhnya terhadap ISPU rendah, menandakan bahwa pembentukan polutan sekunder bukan faktor utama dalam kualitas udara Jakarta. Keberadaan outliers ini mendukung penggunaan MAE sebagai metrik utama karena lebih tahan terhadap lonjakan episodik, dan distribusi data yang relatif stabil mendukung kemampuan *Random Forest* menangkap pola yang kompleks.

Diskusi dan Implikasi

Hasil penelitian menunjukkan bahwa Random Forest merupakan model terbaik dengan akurasi 95,75% (R^2 = 0,9575), RMSE 4,44, dan MAE 0,82, menegaskan reliabilitasnya untuk sistem peringatan dini kualitas udara. Kinerja ini setara atau lebih baik dibandingkan studi serupa[4][8], PM2.5 menjadi penentu utama ISPU dengan kontribusi 87,11%, sementara analisis kategori kualitas udara menunjukkan kondisi kronis sepanjang tahun (76,4% *SEDANG* dan 10,5% *TIDAK SEHAT*) yang relatif merata antarstasiun. *Cross-validation* menghasilkan performa stabil (CV Mean R^2 = 0,9570; CV Std = 0,0102), tanpa indikasi *overfitting*.

Penelitian ini unggul karena akurasi yang tinggi, cakupan data dari seluruh stasiun selama dua puluh bulan, serta integrasi prediksi dan feature importance yang menghasilkan rekomendasi kebijakan yang aplikatif. Dominasi PM2.5 di Jakarta berbeda dari kota besar lain. Di Beijing, PM10 lebih dominan akibat industri batubara dan *dust storms*; rasio PM2.5/PM10 ~0,5–0,6 [7]. Sebaliknya, Jakarta memiliki rasio ~0,8–0,9, mencerminkan dominasi emisi kendaraan bermotor. Di Delhi, polusi sangat dipengaruhi pola musiman dan peristiwa tertentu sehingga variabel temporal menyumbang 15–20% importance[5]. sementara di Jakarta variabel temporal <1%, menunjukkan polusi yang stabil dan kronis. Mexico City berbeda lagi dengan O_3 sebagai kontributor utama AQI, sedangkan di Jakarta O_3 hanya berkontribusi 0,41%.

Implikasi kebijakan dari perbedaan ini antara lain: fokus pada penurunan emisi PM2.5 dari transportasi, perlunya intervensi berkelanjutan karena polusi terjadi sepanjang tahun, serta peningkatan standar emisi dan percepatan adopsi kendaraan listrik. Dominasi PM2.5 juga menegaskan urgensi mitigasi risiko kesehatan akibat partikel halus.

Dari aspek temporal, penelitian ini menggunakan resolusi prediksi harian yang memberikan *lead time* 24 jam dan sesuai untuk *public advisories*, meskipun tidak menangkap dinamika intrahari seperti lonjakan saat jam sibuk. Prediksi per jam membutuhkan data yang lebih detail dan model yang lebih kompleks, sehingga opsi harian dipilih sebagai kompromi paling realistis dengan kondisi infrastruktur pemantauan saat ini.

Secara praktis, model dapat digunakan untuk sistem peringatan dini kualitas udara secara *real-time* [14] termasuk integrasi ke aplikasi mobile, situs pemerintah, atau layanan notifikasi berbasis API. Dari sisi kebijakan, hasil penelitian mendukung prioritas intervensi pada pengendalian emisi kendaraan, pembatasan pembakaran terbuka, pengetatan standar industri, serta perluasan ruang hijau. Selain itu, model dapat dimanfaatkan sebagai alat monitoring untuk mengevaluasi efektivitas kebijakan secara berkelanjutan.

SIMPULAN

Keterbatasan Penelitian

Penelitian ini memiliki beberapa batasan penting. Rentang data yang hanya 20 bulan membatasi analisis pola jangka panjang, termasuk potensi pengaruh iklim tahunan seperti El Niño/La Niña. Model juga belum memasukkan variabel meteorologi seperti suhu, kelembaban, dan angin yang sebenarnya berperan besar dalam pembentukan dan dispersi polutan. Selain itu, penggunaan data harian tidak menangkap dinamika intrahari yang relevan untuk respons real-time. Terakhir, karena model dibangun khusus untuk kondisi Jakarta, penerapannya di kota lain memerlukan penyesuaian dan pelatihan ulang dengan data lokal.

Rekomendasi Penelitian Lanjutan

Penelitian selanjutnya dapat dikembangkan melalui prediksi berbasis data per jam dengan memasukkan variabel meteorologi serta model time-series seperti LSTM atau GRU. Integrasi sumber data eksternal—traffic real-time, aktivitas industri, AOD satelit, dan hotspot kebakaran—akan meningkatkan kemampuan model menangkap kejadian polusi episodik. Pendekatan probabilistic forecasting juga diperlukan agar model menghasilkan

prediksi beserta tingkat ketidakpastiannya. Selain itu, pengembangan model spasial diperlukan untuk memprediksi ISPU di lokasi tanpa stasiun pemantauan, disertai studi longitudinal untuk memantau degradasi performa dan memungkinkan pembelajaran adaptif. Integrasi dengan data kesehatan dapat memperkuat validasi serta mendukung penetapan ambang tindakan bagi kelompok rentan.

DAFTAR PUSTAKA

[1] World Health Organization, "WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide," 2021.

[2] G. Syuhada et al., "Impacts of air pollution on health and cost of illness in Jakarta, Indonesia," Int. J. Environ. Res. Public Health, vol. 20, no. 4, p. 2916, 2023, doi: 10.3390/ijerph20042916.

[3] Kementerian Lingkungan Hidup dan Kehutanan Republik Indonesia, "Peraturan Menteri Lingkungan Hidup dan Kehutanan tentang Indeks Standar Pencemar Udara," 2020.

[4] M. Méndez, M. G. Merayo, and M. Núñez, "Machine learning algorithms to forecast air quality: A survey," Artif. Intell. Rev., vol. 56, pp. 10031–10066, 2023, doi: 10.1007/s10462-023-10424-4.

[5] A. Kumar and P. Goyal, "Forecasting of daily air quality index in Delhi," Sci. Total Environ., vol. 409, no. 24, pp. 5517–5523, 2011, doi: 10.1016/j.scitotenv.2011.08.069.

[6] J. J. Carbajal-Hernández, L. P. Sánchez-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "Assessment and prediction of air quality using fuzzy logic and autoregressive models," Atmos. Environ., vol. 60, pp. 37–50, 2012, doi: 10.1016/j.atmosenv.2012.06.001.

[7] X. Li et al., "Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation," Environ. Pollut., vol. 231, pp. 997–1004, 2017, doi: 10.1016/j.envpol.2017.08.114.

[8] N. Gupta, S. Gupta, M. Khosravy, N. Dey, and R. González-Crespo, "Prediction of air quality index using machine learning techniques: A comparative analysis," J. Environ. Public Health, vol. 2023, p. 4916267, 2023, doi: 10.1155/2023/4916267.

[9] A. Houdou, I. Badisy, and K. Khomsi, "Interpretable machine learning approaches for forecasting and predicting air pollution: A systematic review," Aerosol Air Qual. Res., vol. 24, no. 6, 2024, doi: 10.4209/aaqr.230394.

[10] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[11] J. H. Friedman, "Greedy function approximation: A Gradient Boosting machine," Ann. Stat., vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.

[12] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[13] I. M. Ihsan et al., "Air quality assessment based on real-time continuous monitoring: Particulate and nitrogen dioxide concentrations in South Tangerang," J. Teknol. Lingkung., vol. 26, no. 1, pp. 97–104, 2025, doi: 10.55981/jtl.2025.2887.

[14] Scikit-learn Documentation, "Machine learning in Python." 2024.