

PREDIKSI AQI HARIAN YOGYAKARTA MENGGUNAKAN RANDOM FOREST DAN XGBOOST

Andita Nurrizki Rahmani, Riya Widayanti, Hani Dewi Ariessanti, Bayu Sulistiyanto Ipung Sutejo

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Esa Unggul

Jalan Arjuna Utara No. 9, Kebon Jeruk, Jakarta Barat, Indonesia

1riya.widayanti@esaunggul.ac.id

ABSTRAK

Kualitas udara merupakan faktor penting yang memengaruhi kesehatan masyarakat, sementara sistem pemantauan yang tersedia umumnya hanya memberikan informasi kondisi saat ini tanpa kemampuan prediksi. Penelitian ini bertujuan membandingkan performa algoritma Random Forest dan XGBoost dalam peramalan AQI harian di Kota Yogyakarta. Penelitian ini menggunakan data polutan dan meteorologi harian periode 1 Januari 2021 hingga 19 November 2024. Setelah tahap prapemrosesan dan pembentukan fitur, diperoleh 1.411 observasi yang digunakan dalam pemodelan. Nilai AQI dihitung berdasarkan PermenLHK No. 14 Tahun 2020 sebagai target prediksi, kemudian model dibangun menggunakan fitur lag, *rolling statistics*, dan fitur temporal. Evaluasi model menggunakan MAE, RMSE, dan R^2 , serta uji Diebold–Mariano. Hasil penelitian menunjukkan bahwa Random Forest menghasilkan MAE 4,3095, RMSE 6,3869, dan R^2 0,7756, sedangkan XGBoost menghasilkan MAE 5,0629, RMSE 7,6771, dan R^2 0,6757. Uji Diebold–Mariano menunjukkan bahwa perbedaan performa kedua model signifikan secara statistik ($p = 0,0087$). Sebagai pembandingan, model persistensi menghasilkan MAE 3,1476, RMSE 5,9883, dan R^2 0,8027, yang menunjukkan adanya autokorelasi temporal yang kuat pada data AQI Kota Yogyakarta. Analisis *Feature Importance* dan SHAP menunjukkan bahwa AQI_Hari_Ini_Feature, AQI_Lag1, dan PM2.5_Lag1 merupakan fitur yang paling berpengaruh terhadap hasil prediksi.

Kata kunci : aqi, kualitas udara, peramalan deret waktu, random forest, xgboost, yogyakarta

1. PENDAHULUAN

Kualitas udara merupakan faktor penting yang memengaruhi kesehatan masyarakat dan lingkungan. Polusi udara menjadi salah satu tantangan kesehatan lingkungan terbesar di dunia akibat meningkatnya urbanisasi dan aktivitas industri [1].

World Health Organization (WHO) menyebutkan bahwa polusi udara menyebabkan sekitar 7 juta kematian dini setiap tahun [2].

Selain itu, peningkatan konsentrasi polutan seperti PM_{2.5}, PM₁₀, O₃, dan NO₂ diketahui berkaitan erat dengan meningkatnya risiko gangguan pernapasan dan penyakit kardiovaskular [3].

Hal tersebut menunjukkan bahwa kualitas udara yang buruk dapat memberikan dampak serius terhadap kesehatan manusia.

Permasalahan kualitas udara juga menjadi perhatian di Daerah Istimewa Yogyakarta. Penelitian di wilayah Bantul menunjukkan bahwa konsentrasi PM_{2.5} berada pada rentang 50,7–61,9 $\mu\text{g}/\text{m}^3$, melebihi ambang batas yang direkomendasikan [4].

Meskipun penelitian tersebut tidak dilakukan di Kota Yogyakarta, hasilnya menunjukkan bahwa polusi partikulat masih menjadi tantangan kualitas udara di wilayah perkotaan DIY. Selain itu, Dinas Kesehatan Kota Yogyakarta mencatat sebanyak 59.092 kasus ISPA sepanjang tahun 2024 [5].

Walaupun penelitian ini tidak mengkaji hubungan kausal antara kualitas udara dan ISPA, kondisi tersebut menunjukkan pentingnya pemantauan kualitas udara yang lebih baik.

Dalam pemantauan kualitas udara, Indonesia menggunakan Indeks Standar Pencemar Udara (ISPU)

atau Air Quality Index (AQI) sebagai indikator tingkat pencemaran udara [6].

Namun, sistem pemantauan yang tersedia umumnya hanya menyajikan kondisi aktual dan belum mampu memberikan informasi mengenai kualitas udara pada hari berikutnya. Di sisi lain, jumlah stasiun pemantauan kualitas udara di Yogyakarta masih terbatas sehingga representasi kondisi kualitas udara secara spasial maupun temporal belum optimal [7].

Data kualitas udara memiliki karakteristik deret waktu yang kompleks karena dipengaruhi pola musiman, tren, dan hubungan non-linear antara polutan dengan faktor meteorologi [8].

Oleh karena itu, diperlukan metode peramalan yang mampu menangkap pola tersebut secara efektif. Pendekatan machine learning seperti Random Forest dan XGBoost telah banyak digunakan dalam penelitian kualitas udara karena mampu menangani hubungan non-linear dengan baik [9].

Namun, hasil penelitian terdahulu menunjukkan temuan yang beragam mengenai performa kedua algoritma tersebut. Beberapa penelitian melaporkan bahwa Random Forest memberikan hasil yang lebih baik pada prediksi kualitas udara dengan validasi kronologis [10], sementara penelitian lain menunjukkan bahwa Random Forest dan XGBoost memiliki performa yang relatif sebanding [8].

Berdasarkan penelitian terdahulu, penelitian mengenai kualitas udara di Yogyakarta masih lebih banyak berfokus pada metode K-Nearest Neighbor (KNN) untuk klasifikasi kualitas udara atau LSTM dan GRU untuk prediksi konsentrasi polutan tertentu. Berdasarkan studi literatur yang dilakukan, belum

ditemukan penelitian yang membandingkan algoritma Random Forest dan XGBoost untuk peramalan AQI harian di Kota Yogyakarta. Selain itu, sebagian besar penelitian masih mengevaluasi performa model berdasarkan metrik seperti MAE, RMSE, atau R^2 tanpa membandingkannya dengan model persistensi maupun menguji signifikansi perbedaan performa menggunakan uji Diebold–Mariano. Penelitian yang menggabungkan analisis interpretabilitas menggunakan SHAP dengan perbandingan langsung antara Random Forest dan XGBoost pada data AQI harian juga masih terbatas, karena penelitian sebelumnya yang menerapkan SHAP umumnya menggunakan pendekatan model hybrid [9] dan bukan pada konteks Kota Yogyakarta. Oleh karena itu, penelitian ini membandingkan performa Random Forest dan XGBoost untuk peramalan AQI harian di Kota Yogyakarta, menggunakan model persistensi sebagai baseline, uji Diebold–Mariano untuk mengevaluasi signifikansi perbedaan performa, serta analisis SHAP untuk menjelaskan kontribusi setiap fitur terhadap hasil prediksi.

2. TINJAUAN PUSTAKA

2.1. Indeks Kualitas Udara (AQI/ISPU)

Kualitas udara ditentukan oleh konsentrasi berbagai polutan di atmosfer. Namun data konsentrasi polutan mentah seperti PM2.5 (ug/m3) atau O3 (ppb) terlalu teknis untuk dipahami oleh masyarakat umum maupun pengambil kebijakan. Untuk itulah Indeks Kualitas Udara atau Air Quality Index (AQI) dikembangkan. AQI berfungsi sebagai alat penilaian kuantitatif yang menyederhanakan data polutan yang kompleks menjadi satu angka tunggal yang mudah dimengerti [6].

Di Indonesia, instrumen dengan fungsi serupa dikenal sebagai Indeks Standar Pencemar Udara (ISPU) yang diatur melalui Peraturan Menteri Lingkungan Hidup dan Kehutanan Nomor 14 Tahun 2020. Nilai AQI komposit harian ditentukan dari nilai maksimum sub-indeks polutan (PM2.5, PM10, SO2, CO, O3, dan NO2).

2.2. Peramalan Deret Waktu

Data AQI harian merupakan contoh klasik data deret waktu yang memiliki beberapa karakteristik utama, yaitu ketergantungan temporal, tren jangka panjang, dan pola musiman. Nilai AQI pada suatu waktu sangat berkaitan dengan nilai pada waktu-waktu sebelumnya karena polusi udara memiliki sifat inersia - polutan tidak hilang seketika [8].

Hal ini membuat nilai AQI historis menjadi prediktor yang sangat kuat untuk nilai di masa depan. Pendekatan machine learning menangani masalah ini melalui rekayasa fitur dengan memanfaatkan nilai-nilai dari masa lalu sebagai variabel prediktor, sebuah teknik yang dikenal sebagai pembentukan variabel lag [9].

2.3. Random Forest

Random Forest (RF) adalah algoritma ensemble learning yang dikembangkan oleh Leo Breiman. Algoritma ini menggabungkan prediksi dari sejumlah pohon keputusan (decision tree) yang dibangun secara paralel menggunakan teknik Bootstrap Aggregating (bagging) [11].

Setiap pohon dilatih menggunakan sampel data acak yang berbeda, dan pada setiap percabangan hanya sebagian fitur yang dipertimbangkan secara acak.

$$\hat{y} = \frac{1}{N} \sum_{i=1}^N T_i(x) \quad (1)$$

Keterangan:

$\hat{y}$ = hasil prediksi akhir

N = jumlah decision tree

$T_i(x)$ = prediksi dari tree ke-i

Pada kasus regresi, prediksi akhir merupakan rata-rata dari seluruh pohon. Kelebihan RF terletak pada kemampuannya menangani data yang mengandung banyak noise serta menyediakan mekanisme internal untuk mengukur tingkat kepentingan fitur (feature importance) [12].

2.4. XGBoost

eXtreme Gradient Boosting (XGBoost) merupakan implementasi algoritma gradient boosting yang efisien. Berbeda dengan Random Forest yang membangun setiap pohon secara independen dan paralel, XGBoost menerapkan prinsip boosting di mana serangkaian pohon keputusan dibangun secara berurutan [13].

Setiap pohon baru dilatih secara spesifik untuk memperbaiki kesalahan atau residual yang dihasilkan oleh pohon-pohon sebelumnya. XGBoost memiliki mekanisme regularisasi untuk mencegah overfitting dan kemampuan bawaan untuk menangani nilai yang hilang (missing values), sehingga menyederhanakan tahap pra-pemrosesan data [14].

Pada proses pelatihannya, XGBoost meminimalkan fungsi objektif yang terdiri dari fungsi loss dan regularisasi, sebagaimana ditunjukkan pada Persamaan (2).

$$Obj = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (2)$$

dengan l menyatakan fungsi loss, sedangkan $\Omega(f_k)$ merupakan fungsi regularisasi untuk mengontrol kompleksitas model.

2.5. Penelitian Terdahulu

Beberapa penelitian telah menggunakan machine learning untuk prediksi kualitas udara dengan hasil yang bervariasi tergantung konteks dan karakteristik data yang digunakan. Tabel 1 merangkum penelitian-penelitian yang relevan beserta posisi penelitian ini dalam konteks studi terdahulu.

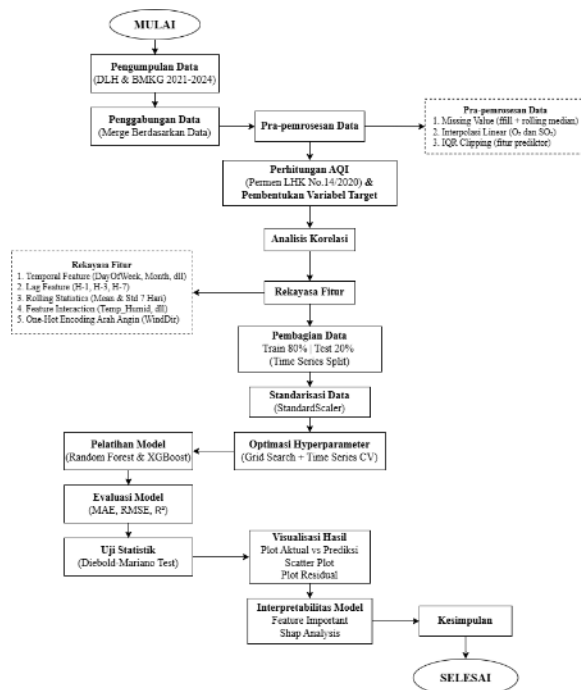
Tabel 1. Ringkasan penelitian terdahulu yang relevan

Peneliti	Metode	Fitur dan Preprocessing	Hasil Utama
Celikten (2026)	RF, XGBoost, MLR	Lag features, Rolling Average, Validasi Kronologis	RF terbaik untuk PM10 dengan R2: 0,64
Juarez & Petersen (2022)	XGBoost, RF, KNN, LSTM, SVR	Seasonal training data, RobustScaler	XGBoost dan RF terbaik dengan R2: 0,61
Syahira & Arianto (2024)	KNN	Data polutan Yogyakarta	Akurasi klasifikasi 99% untuk kategori Baik dan Sedang
Shinzi et al. (2025)	LSTM, GRU	Data polutan Yogyakarta	LSTM unggul untuk CO, GRU untuk NO2
Natarajan et al. (2024)	Decision Tree + GWO	Data kualitas udara India	Akurasi meningkat signifikan setelah optimasi

Berdasarkan Tabel 1 dapat dilihat bahwa belum ada penelitian yang secara khusus membandingkan Random Forest dan XGBoost untuk peramalan AQI harian di Yogyakarta dengan analisis SHAP. Hal ini menjadi justifikasi utama dilaksanakannya penelitian ini.

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan desain eksperimen komparatif untuk membandingkan performa Random Forest dan XGBoost dalam peramalan deret waktu AQI harian di Kota Yogyakarta menggunakan data historis polutan dan meteorologi. Seluruh tahapan penelitian diilustrasikan pada Gambar 1.



Gambar 1. Alur Penelitian

4.1. Sumber dan Pengumpulan Data

Data yang digunakan merupakan data sekunder dari dua sumber: (1) Data kualitas udara harian dari Stasiun Pemantauan Kualitas Udara (AQMS) milik DLH Kota Yogyakarta, berupa konsentrasi PM_{2.5}, PM₁₀, SO₂, CO, O₃, dan NO₂; (2) Data meteorologi harian dari BMKG Yogyakarta meliputi suhu, kelembapan, curah hujan, kecepatan, dan arah angin.

Penelitian menggunakan data deret waktu harian periode 1 Januari 2021 hingga 19 November 2024 dengan total 1.419 observasi.

4.2. Pra-pemrosesan Data

Tahap pra-pemrosesan meliputi: (1) Penggabungan dataset DLH dan BMKG berdasarkan kolom tanggal; (2) Penanganan nilai hilang menggunakan interpolasi linear dengan batas tujuh hari untuk polutan dan forward fill untuk variabel meteorologi; (3) Penanganan pencilaan menggunakan metode IQR dengan teknik capping; (4) Konversi satuan CO dari µg/m³ ke mg/m³ sesuai standar perhitungan ISPU.

4.3. Konstruksi Variabel Target (AQI)

Nilai AQI harian dihitung berdasarkan PermenLHK No. 14 Tahun 2020 menggunakan formula interpolasi linear. Sub-indeks setiap polutan dihitung dengan Persamaan (3), kemudian nilai AQI harian ditetapkan sebagai nilai maksimum dari seluruh sub-indeks:

$$I = \frac{I_a - I_b}{X_a - X_b} (X_x - X_b) + I_b \quad (3)$$

Di mana I = nilai ISPU, I_a = batas atas indeks, I_b = batas bawah indeks, X_a = batas atas konsentrasi ambien, X_b = batas bawah konsentrasi ambien, dan X_x = konsentrasi hasil pengukuran.

4.4. Rekayasa Fitur

Rekayasa fitur menghasilkan 39 fitur baru yang terdiri dari: (1) Fitur temporal (DayOfWeek, Month, month_sin, is_holiday); (2) Fitur lag AQI, PM_{2.5}, PM₁₀, TAVG, RH_AVG pada H-1, H-3, dan H-7; (3) Fitur statistik bergerak (rolling mean dan std 7 hari untuk PM_{2.5}, PM₁₀, dan AQI); (4) Fitur interaksi suhu-kelembapan (temp_humid); (5) One-Hot Encoding arah angin menjadi 13 kolom biner. Total fitur prediktor yang digunakan adalah 51 variabel.

4.5. Pembagian Data dan Pemodelan

Dataset dibagi secara kronologis dengan rasio 80:20, menghasilkan 1.128 data latih (8 Januari 2021–9 Februari 2024) dan 283 data uji (10 Februari–18 November 2024). Rasio 80:20 dipilih untuk memberikan keseimbangan antara jumlah data latih yang cukup untuk mempelajari pola deret waktu dan jumlah data uji yang memadai untuk mengevaluasi

performa model. Seluruh proses optimasi hyperparameter menggunakan Time Series Cross-Validation dilakukan hanya pada data latih, sedangkan data uji digunakan untuk evaluasi akhir sehingga tetap independen dari proses pelatihan dan tuning model. Standardisasi fitur menggunakan StandardScaler yang difit hanya pada data latih untuk mencegah data leakage.

Dua algoritma yang dibandingkan adalah: (1) Random Forest (RF) menggunakan prinsip bagging dengan menggabungkan prediksi dari sejumlah decision tree; (2) XGBoost yang menggunakan prinsip boosting dengan membangun pohon secara sekuensial untuk memperbaiki kesalahan model sebelumnya [11],[12].

Optimasi hyperparameter dilakukan menggunakan GridSearchCV yang dikombinasikan dengan Time Series Cross-Validation 5-fold. Sebagai pembanding, digunakan model persistensi (naïve forecast) yang memprediksi nilai AQI pada hari berikutnya sama dengan nilai AQI pada hari sebelumnya. Model ini tidak memerlukan proses pelatihan dan digunakan sebagai baseline untuk mengevaluasi nilai tambah yang diberikan oleh model machine learning.

Evaluasi kinerja dilakukan menggunakan Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), dan koefisien determinasi (R^2). Selain itu, dilakukan uji Diebold–Mariano untuk menguji signifikansi perbedaan performa prediksi antara Random Forest dan XGBoost. Sebagai pelengkap, dihitung bootstrap confidence interval 95% terhadap selisih RMSE menggunakan 1.000 pengambilan sampel ulang.

Untuk menginterpretasikan kontribusi masing-masing variabel terhadap hasil prediksi, digunakan analisis feature importance dan SHAP (SHapley Additive exPlanations).

4. HASIL DAN PEMBAHASAN

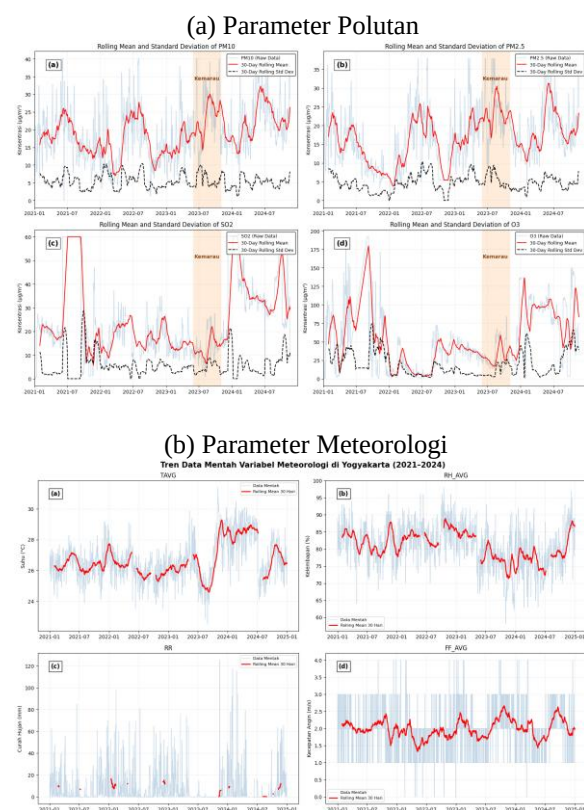
4.1. Eksplorasi dan Karakteristik Data

Tabel 2. Statistik Deskriptif Parameter Polutan (2021-2024)

Parameter	Mean	Std	Min	Max
PM ₁₀	19.75	9.12	0.00	66.70
PM _{2.5}	18.15	8.79	2.00	71.33
SO ₂	22.44	14.08	0.00	93.00
CO	1.07	0.61	0.00	3.04
O ₃	49.66	39.79	0.00	193.67
NO ₂	33.89	36.97	1.00	189.63
TAVG	26.66	1.33	22.50	31.40
RH_AVG	80.97	6.08	58.00	98.00
RR	7.77	15.66	0.00	125.30
FF_AVG	1.99	0.69	0.00	4.00

Berdasarkan Tabel 2, parameter polutan menunjukkan tingkat variasi yang berbeda selama periode pengamatan. O₃ memiliki rata-rata sebesar 49,66 $\mu\text{g}/\text{m}^3$ dengan standar deviasi 39,79 $\mu\text{g}/\text{m}^3$, sedangkan NO₂ memiliki rata-rata 33,89 $\mu\text{g}/\text{m}^3$ dengan

standar deviasi 36,97 $\mu\text{g}/\text{m}^3$, yang menunjukkan variasi konsentrasi harian yang relatif tinggi. Sebaliknya, PM₁₀ dan PM_{2.5} memiliki standar deviasi yang lebih kecil sehingga cenderung lebih stabil. Pada variabel meteorologi, suhu rata-rata (TAVG) sebesar 26,66°C dengan fluktuasi yang relatif rendah, sedangkan curah hujan (RR) memiliki standar deviasi 15,66 mm yang lebih besar daripada nilai rata-ratanya, menunjukkan variasi curah hujan yang tinggi selama periode pengamatan. Variabel kelembapan relatif (RH_AVG) dan kecepatan angin rata-rata (FF_AVG) menunjukkan variasi yang lebih moderat. Karakteristik ini mengindikasikan bahwa data polutan dan meteorologi memiliki tingkat variabilitas yang berbeda, sehingga berpotensi memberikan informasi yang saling melengkapi dalam pemodelan AQI.



Gambar 2. Rolling Mean dan Rolling Standard Deviation Variabel Penelitian di Yogyakarta Tahun 2021–2024.

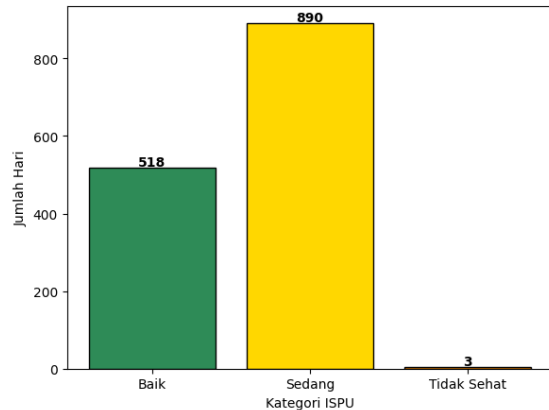
Berdasarkan Gambar 2(a), parameter polutan menunjukkan karakteristik temporal yang berbeda selama periode pengamatan. PM₁₀ dan PM_{2.5} memperlihatkan pola fluktuasi yang relatif berulang, sedangkan O₃ dan NO₂ memiliki variabilitas yang lebih tinggi yang ditunjukkan oleh perubahan *rolling standard deviation* yang lebih besar pada beberapa periode. Hal ini menunjukkan bahwa konsentrasi O₃ dan NO₂ cenderung mengalami perubahan yang lebih dinamis dibandingkan parameter polutan lainnya.

Sementara itu, Gambar 2(b) menunjukkan bahwa variabel meteorologi juga memiliki karakteristik yang berbeda. Suhu rata-rata (TAVG) dan kecepatan angin

rata-rata (FF_AVG) cenderung stabil sepanjang periode pengamatan, sedangkan curah hujan (RR) menunjukkan fluktuasi paling tinggi dengan beberapa lonjakan pada periode tertentu. Kelembapan relatif (RH_AVG) berfluktuasi pada rentang yang lebih moderat. Perbedaan karakteristik temporal antara variabel polutan dan meteorologi menunjukkan bahwa masing-masing variabel memiliki pola dinamika yang berbeda, sehingga berpotensi memberikan informasi yang saling melengkapi dalam proses pemodelan AQI.

4.2. Konstruksi Target AQI

Distribusi Hari Berdasarkan Kategori Kualitas Udara (ISPU) 2021-2024

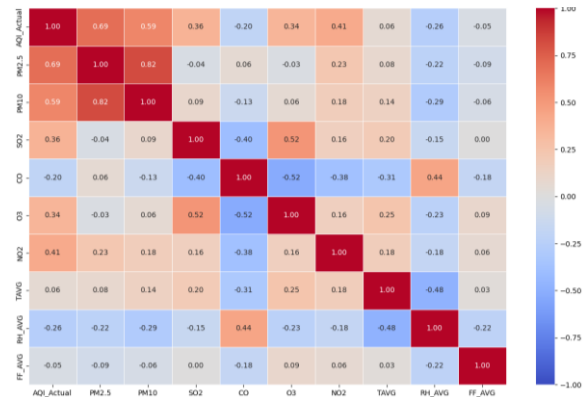


Gambar 3. Distribusi Nilai AQI Aktual di Yogyakarta (2021-2024)

Berdasarkan Gambar 3, setelah tahap pra-pemrosesan dan perhitungan AQI harian berdasarkan Permen LHK Nomor 14 Tahun 2020, diperoleh distribusi kategori kualitas udara yang didominasi oleh kategori Sedang sebanyak 890 hari (63,1%), diikuti kategori Baik sebanyak 518 hari (36,7%), sedangkan kategori Tidak Sehat hanya tercatat sebanyak 3 hari (0,2%), yaitu pada tanggal 18–20 Juni 2023. Nilai AQI rata-rata sebesar 52,01 sehingga secara umum kualitas udara di Yogyakarta selama periode penelitian berada pada kategori Sedang berdasarkan standar ISPU. Meskipun demikian, keberadaan tiga hari dengan kategori Tidak Sehat menunjukkan bahwa pada periode tertentu terjadi peningkatan konsentrasi polutan yang menyebabkan penurunan kualitas udara. Secara keseluruhan, distribusi ini menunjukkan bahwa sebagian besar kondisi kualitas udara berada pada kategori Baik hingga Sedang, dengan kejadian kualitas udara yang buruk relatif jarang terjadi selama periode penelitian.

4.3. Analisis Korelasi Variabel

Berdasarkan Gambar 4, PM_{2.5} (0,69) dan PM₁₀ (0,59) memiliki korelasi positif paling kuat terhadap AQI_Actual dibandingkan variabel lainnya. Selain itu, PM_{2.5} dan PM₁₀ juga memiliki korelasi yang kuat satu sama lain (0,82), yang menunjukkan bahwa kedua polutan cenderung mengalami perubahan searah selama periode pengamatan.



Gambar 4. Matriks Korelasi Polutan dan Meteorologi

Sementara itu, NO₂ (0,41), SO₂ (0,36), dan O₃ (0,34) menunjukkan korelasi positif sedang terhadap AQI_Actual, sedangkan CO memiliki korelasi negatif yang relatif lemah (-0,20). Variabel meteorologi seperti TAVG (0,06), RH_AVG (-0,26), dan FF_AVG (-0,05) juga menunjukkan korelasi yang relatif lemah terhadap AQI_Actual.

Hasil analisis korelasi ini memberikan gambaran awal mengenai hubungan linear antarvariabel. Namun, seluruh variabel tetap dipertimbangkan dalam tahap pemodelan karena algoritma Random Forest dan XGBoost mampu mempelajari hubungan nonlinier serta interaksi antarfitur yang tidak sepenuhnya tercermin dari nilai korelasi Pearson.

4.4. Hasil Optimasi Hyperparameter

Tabel 3. Hasil Optimalisasi Parameter Terbaik (*Best Parameters*)

Model	Parameter	Nilai Optimal
Random Forest	n_estimators	300
	max_depth	20
	min_samples_leaf	4
XGBoost	n_estimators	200
	learning_rate	0.01
	max_depth	3
	subsample	1.0

Berdasarkan hasil optimasi menggunakan GridSearchCV dengan TimeSeriesSplit 5-fold, diperoleh konfigurasi hyperparameter terbaik sebagaimana ditunjukkan pada Tabel 3. Berdasarkan tabel tersebut, konfigurasi parameter terbaik menunjukkan perbedaan karakteristik antara algoritma XGBoost dan Random Forest. XGBoost menggunakan nilai *max_depth* yang rendah dan *learning_rate* kecil sebesar 0,01 sehingga proses pembelajaran berlangsung secara bertahap untuk mengurangi risiko *overfitting*, sedangkan parameter *subsample* sebesar 1,0 membuat model menggunakan seluruh data latih pada setiap iterasi. Di sisi lain, Random Forest menggunakan kedalaman pohon yang lebih besar (*max_depth* = 20) dan *min_samples_leaf* sebesar 4 sehingga mampu menangkap pola data yang lebih kompleks dengan tetap menjaga stabilitas prediksi melalui mekanisme rata-rata dari 300 pohon.

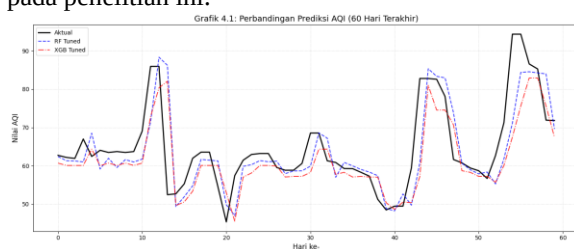
Parameter terbaik tersebut kemudian digunakan pada proses pelatihan akhir dan evaluasi model.

4.5. Evaluasi dan Perbandingan Kinerja Model

Table 4. Rekapitulasi Perbandingan Kinerja Model

Model	Skenario	MAE	RMSE	R ²
Persistensi	Baseline	3,1476	5,9883	0,8027
Random Forest	Baseline	5,1259	7,0419	0,7272
	Tuned	4,3095	6,3869	0,7756
XGBoost	Baseline	5,2408	7,4416	0,6953
	Tuned	5,0629	7,6771	0,6757

Berdasarkan hasil evaluasi pada Tabel 4, Random Forest Tuned menghasilkan MAE sebesar 4,3095, RMSE sebesar 6,3869, dan R² sebesar 0,7756, sedangkan XGBoost Tuned menghasilkan MAE sebesar 5,0629, RMSE sebesar 7,6771, dan R² sebesar 0,6757. Random Forest Tuned menunjukkan performa yang lebih baik dibandingkan XGBoost Tuned pada seluruh metrik evaluasi. Selain itu, proses optimasi hyperparameter juga meningkatkan performa Random Forest dibandingkan model baseline, sedangkan pada XGBoost optimasi yang dilakukan belum menghasilkan peningkatan performa. Hal ini menunjukkan bahwa konfigurasi hyperparameter yang diperoleh lebih sesuai untuk karakteristik data AQI Yogyakarta pada algoritma Random Forest. Untuk menilai signifikansi perbedaan tersebut, dilakukan uji Diebold–Mariano yang menghasilkan nilai statistik DM sebesar $-2,6252$ dengan p -value sebesar 0,0087. Selain itu, *bootstrap confidence interval* 95% untuk selisih RMSE berada pada rentang $[-1,9196; -0,7628]$. Karena p -value lebih kecil dari 0,05 dan interval kepercayaan tidak mencakup nilai nol, maka terdapat perbedaan performa yang signifikan antara kedua model. Dengan demikian, Random Forest Tuned dipilih sebagai model *machine learning* terbaik pada penelitian ini.

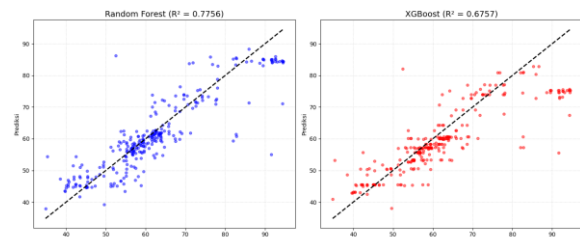


Gambar 5. Perbandingan Prediksi AQI (Aktual vs Model Terbaik)

Model persistensi menghasilkan MAE sebesar 3,1476, RMSE sebesar 5,9883, dan R² sebesar 0,8027, yang lebih baik dibandingkan kedua model *machine learning* yang diuji. Hasil ini menunjukkan bahwa data AQI Yogyakarta memiliki autokorelasi temporal yang kuat sehingga nilai AQI pada hari sebelumnya sudah mampu memberikan prediksi yang kompetitif untuk hari berikutnya. Meskipun demikian, model persistensi hanya memanfaatkan informasi dari nilai AQI sebelumnya, sedangkan pendekatan *machine learning* memanfaatkan informasi dari berbagai

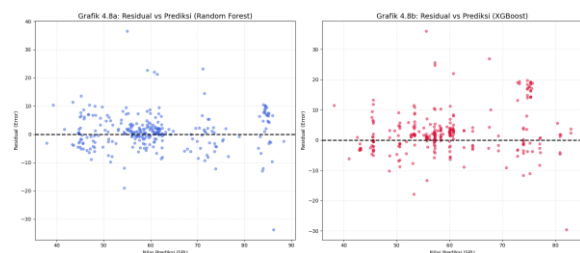
variabel polutan, meteorologi, dan fitur hasil rekayasa sehingga tetap memiliki potensi untuk dikembangkan pada skenario prediksi yang lebih kompleks.

Berdasarkan Gambar 7, kedua model mampu mengikuti pola perubahan AQI aktual dengan cukup baik, termasuk pada penurunan sekitar hari ke-20 dan peningkatan pada hari ke-43 hingga ke-58. Namun, pada sekitar hari ke-55 hingga ke-57 ketika AQI aktual meningkat mendekati 95, kedua model hanya memprediksi pada kisaran 84–85, menunjukkan kecenderungan under-prediction pada kondisi AQI tinggi. Kurva prediksi RF terlihat sedikit lebih dekat dengan kurva aktual dibandingkan XGBoost, sejalan dengan hasil evaluasi kuantitatif.



Gambar 6. Perbandingan Scatter Plot Prediksi Aktual Model Random Forest dan XGBoost

Berdasarkan Gambar 6, scatter plot memperlihatkan bahwa titik-titik prediksi Random Forest terkonsentrasi lebih dekat ke garis diagonal ideal dibandingkan XGBoost, yang menunjukkan tingkat kesesuaian prediksi terhadap nilai AQI aktual yang lebih baik. Temuan ini konsisten dengan hasil evaluasi, di mana Random Forest memperoleh nilai RMSE yang lebih rendah dan nilai R² yang lebih tinggi dibandingkan XGBoost. Selain itu, pada rentang AQI aktual sekitar 80–95, kedua model cenderung menghasilkan nilai prediksi yang lebih rendah daripada nilai aktual (*under-prediction*), sedangkan pada beberapa pengamatan dengan nilai AQI yang lebih rendah terlihat kecenderungan prediksi sedikit lebih tinggi daripada nilai aktual (*over-prediction*).

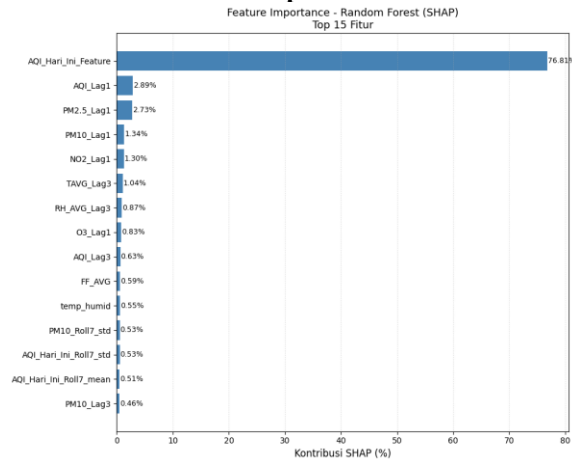


Gambar 7. Plot Residual Model Random Forest dan XGBoost

Berdasarkan Gambar diatas , residual kedua model tersebar di sekitar garis nol tanpa menunjukkan pola tertentu, yang mengindikasikan bahwa tidak terdapat bias sistematis yang jelas pada hasil prediksi. Pada Random Forest, sebagian besar residual berada di sekitar nol, meskipun pada nilai prediksi AQI yang lebih tinggi masih terlihat beberapa residual positif

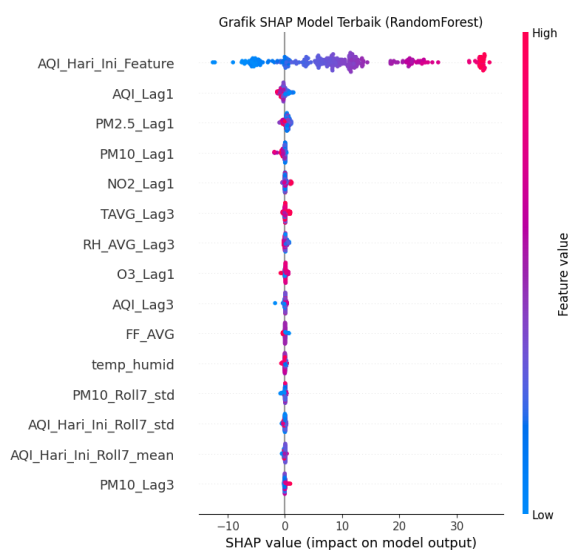
yang mengindikasikan kecenderungan *under-prediction*. Sementara itu, XGBoost juga menunjukkan pola residual yang relatif acak, namun dengan penyebaran residual yang lebih lebar pada beberapa pengamatan. Temuan ini sejalan dengan hasil evaluasi sebelumnya yang menunjukkan bahwa Random Forest menghasilkan nilai RMSE yang lebih rendah dibandingkan XGBoost.

4.6. Analisis Feature Importance dan SHAP



Gambar 8. Tingkat Kepentingan Fitur pada Model Random Forest

Berdasarkan Gambar 8, *AQI_Hari_Ini_Feature* mendominasi kontribusi fitur dengan nilai SHAP sebesar 76,81%, jauh lebih tinggi dibandingkan variabel lainnya. Temuan ini menunjukkan bahwa informasi AQI saat ini memiliki pengaruh yang sangat besar terhadap prediksi AQI hari berikutnya. Kondisi tersebut mencerminkan adanya autokorelasi temporal yang kuat pada data kualitas udara di Yogyakarta, di mana perubahan kualitas udara cenderung berlangsung secara bertahap dari hari ke hari.



Gambar 9. SHAP Beeswarm Plot

Di bawah *AQI_Hari_Ini_Feature*, *AQI_Lag1* serta beberapa variabel polutan seperti *PM2.5_Lag1*, *PM10_Lag1*, *NO2_Lag1*, dan *O3_Lag1* memberikan kontribusi terhadap prediksi, meskipun nilainya jauh lebih kecil. Hal ini menunjukkan bahwa konsentrasi polutan memberikan kontribusi yang lebih besar terhadap prediksi AQI dibandingkan sebagian besar variabel meteorologi. Sementara itu, variabel cuaca seperti suhu, kelembapan, dan arah angin memiliki kontribusi yang relatif kecil, sehingga fluktuasi AQI pada data penelitian ini lebih banyak dipengaruhi oleh kondisi polutan dan informasi historis AQI dibandingkan perubahan cuaca harian.

Berdasarkan Gambar 9, *AQI_Hari_Ini_Feature* merupakan variabel yang paling berpengaruh terhadap prediksi model Random Forest. Temuan ini menunjukkan bahwa informasi AQI pada hari ini memiliki kontribusi yang sangat besar terhadap prediksi AQI hari berikutnya, yang mengindikasikan adanya autokorelasi temporal yang kuat pada data AQI Yogyakarta.

Selain itu, fitur polutan seperti *PM2.5_Lag1*, *PM10_Lag1*, *NO2_Lag1*, dan *O3_Lag1* juga memberikan kontribusi terhadap prediksi, meskipun pengaruhnya lebih kecil dibandingkan fitur historis AQI. Sementara itu, variabel meteorologi seperti *RH_AVG_Lag3* dan *TAVG_Lag3* memiliki kontribusi yang relatif lebih rendah dibandingkan variabel polutan.

Secara keseluruhan, hasil analisis SHAP menunjukkan bahwa prediksi model Random Forest lebih banyak dipengaruhi oleh informasi historis AQI dan konsentrasi polutan dibandingkan variabel meteorologi.

5. KESIMPULAN DAN SARAN

Penelitian ini membandingkan performa Random Forest dan XGBoost dalam peramalan AQI harian di Kota Yogyakarta menggunakan data polutan dan meteorologi periode 2021–2024. Random Forest Tuned menghasilkan performa terbaik di antara model *machine learning* dengan MAE sebesar 3,67, RMSE sebesar 5,00, dan R^2 sebesar 0,7722, sedangkan XGBoost Tuned menghasilkan MAE sebesar 3,79, RMSE sebesar 5,13, dan R^2 sebesar 0,7601. Namun, uji Diebold–Mariano menunjukkan bahwa perbedaan performa kedua model tidak signifikan secara statistik ($p\text{-value} = 0,1395$), sehingga keduanya dapat dianggap memiliki performa yang relatif setara. Model persistensi menghasilkan RMSE sebesar 4,8895 dan R^2 sebesar 0,7823, yang lebih baik dibandingkan kedua model *machine learning*. Temuan ini menunjukkan bahwa data AQI Yogyakarta memiliki autokorelasi temporal yang kuat. Analisis feature importance dan SHAP menunjukkan bahwa *AQI_Hari_Ini_Feature* merupakan fitur yang memberikan kontribusi terbesar terhadap prediksi, diikuti oleh *AQI_Lag1* dan beberapa fitur polutan seperti *PM2.5_Lag1*, sehingga informasi historis AQI dan konsentrasi polutan menjadi faktor utama dalam

prediksi AQI harian. Penelitian selanjutnya dapat mempertimbangkan penggunaan data yang lebih lengkap, pengembangan prediksi multi-hari, serta penambahan variabel eksternal untuk meningkatkan kemampuan model dalam menangkap variasi kualitas udara.

DAFTAR PUSTAKA

- [1] I. E. Agbehadji and I. C. Obagbuwa, "Systematic Review of Machine Learning and Deep Learning Techniques for Spatiotemporal Air Quality Prediction," *Atmosphere (Basel)*, vol. 15, no. 11, 2024, doi: 10.3390/atmos15111352.
- [2] W. H. Organization, "Ambient (outdoor) air pollution," WHO Fact Sheets. Accessed: Nov. 05, 2025. [Online]. Available: [https://www.who.int/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health#](https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health#)
- [3] S. K. Natarajan, P. Shanmurthy, D. Arockiam, B. Balusamy, and S. Selvarajan, "Optimized machine learning model for air quality index prediction in major cities in India," *Sci. Rep.*, pp. 1–18, 2024, doi: 10.1038/s41598-024-54807-1.
- [4] A. U. Abidin, A. L. Munawaroh, A. Rosinta, A. T. Sulistiyani, I. Ardianta, and F. M. Iresha, "Environmental health risks and impacts of PM2.5 exposure on human health in residential areas, Bantul, Yogyakarta, Indonesia," *Toxicol. Rep.*, vol. 14, no. February, p. 101949, 2025, doi: 10.1016/j.toxrep.2025.101949.
- [5] Adminwarta, "Musim Pancaroba Warga Diminta Waspada Penyakit ISPA," Portal Berita Pemerintah Kota Yogyakarta. Accessed: Nov. 16, 2025. [Online]. Available: <https://warta.jogjakota.go.id/detail/index/36237>
- [6] N. Syahira and D. B. Arianto, "Prediksi Tingkat Kualitas Udara Dengan Pendekatan Algoritma K-Nearest Neighbor," *Jurnal Ilmiah Informatika Komputer*, vol. 29, no. 1, pp. 45–59, 2024, doi: 10.35760/ik.2024.v29i1.10069.
- [7] A. U. Hasanah, K. I. Lutfia, N. I. Wantoputri, and R. Dewi, "Analisis Dispersi SO₂ dan NO₂ di Kawasan Budaya Kota Yogyakarta Menggunakan Model AERMOD," vol. 9, no. September, pp. 153–168, 2025, doi: 10.20885/ajie.vol9.iss3.art1.
- [8] E. K. Juarez and M. R. Petersen, "A Comparison of Machine Learning Methods to Forecast Tropospheric Ozone Levels in Delhi," *Atmosphere (Basel)*, vol. 13, no. 1, pp. 1–26, 2022, doi: 10.3390/atmos13010046.
- [9] A. Yenikar, V. P. Mishra, M. Bali, and T. Ara, "Explainable forecasting of air quality index using a hybrid random forest and ARIMA model," *MethodsX*, vol. 15, no. May, p. 103517, 2025, doi: 10.1016/j.mex.2025.103517.
- [10] H. Çelikten, "Evaluating machine learning (RF , XGBoost) and statistical model (MLR) for PM 10 and air quality prediction : A case from Kars , Türkiye," *Atmos. Pollut. Res.*, no. December 2025, p. 102975, 2026, doi: 10.1016/j.apr.2026.102975.
- [11] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow*. 2022.
- [12] T. Nguyen, C. Truong, and T. Quang, "Development of an Air Quality Forecasting System Based on Random Forest and XGBoost Algorithms," vol. 9, no. 10, pp. 128–136, 2025.
- [13] S. Tırnık, "Machine learning-based forecasting of air quality index under long-term environmental patterns : A comparative approach with XGBoost , LightGBM , and SVM," pp. 1–21, 2025, doi: 10.1371/journal.pone.0334252.
- [14] E. Koçak, "Comprehensive evaluation of machine learning models for real - world air quality prediction and health risk assessment by AirQ +," *Earth Sci. Inform.*, 2025, doi: 10.1007/s12145-025-01941-7