Journal of Computer Science and Computational Intelligence
May 2026 | Vol. 2 | No. 1 | Page 31-46
E-ISSN: 3109-7383 P-ISSN: 3110-8393
DOI: https://doi.org/10.30598/algorithm.v2i1.31-46

Prediksi Konsentrasi Karbon Monoksida (CO) pada Stasiun Kualitas Udara DKI1 Jakarta Menggunakan *Random Forest*

(Prediction of Carbon Monoxide (CO) Concentration at the DKI1 Jakarta Air Quality Station Using Random Forest)

Shania Hery Wattimury[1], Emanuella M C Wattimena[2], Helda Yunita Taihuttu[3*]

[1,2,3] Program Studi Ilmu Komputer, Fakultas Sains dan Teknologi, Universitas Pattimura
Jl.Ir.M.Putuhena, Ambon, 97233, Indonesia
Corresponding Author: **yunitahelda24@gmail.com**

Manuscript submitted: ***15th April 2026***
Manuscript revision: ***20th April 2026***
Accepted for publication: ***5th May 2026***

Abstract

This study develops a prediction model for carbon monoxide (CO) concentration in Jakarta using the Random Forest Regressor algorithm with Grid Search-based parameter optimization. The dataset consists of 1,540 daily observations from the DKI1 Air Quality Monitoring Station (January 2017 - March 2021) including meteorological variables and lagged CO values. Feature engineering produces 12 predictors through lag features, rolling mean, and rolling standard deviation. The optimal model with the configuration n_estimators=300, max_depth=10, min_samples_leaf=4, min_samples_split=10, and max_features='sqrt' achieves a testing RMSE of 4.3216 μg/m³ with a coefficient of determination R^2 = 0.3741. Feature importance analysis revealed that temporal features dominated (52.83% cumulative contribution), with the 3-day rolling mean (17.87%), lag 1 (17.62%), and 7-day rolling mean (17.34%) as the top 3 predictors. Although the model captured the overall trend well, systematic underprediction occurred at extreme values (errors up to -25 μg/m³), indicating the need for a hybrid approach with quantile regression or gradient boosting for improved tail risk capture. The findings support the use of temporal features as the primary anchor in short-term CO forecasting.

Keywords: *Air Quality; Carbon Monoxide; Feature Engineering; Grid Search; Machine Learning; Pollution Prediction, Random Forest, Time Series Forecasting.*

How to cite this article:
S. H. Wattimury, E. M. C. Wattimena, and H. Y. Taihuttu, "Prediksi Konsentrasi Karbon Monoksida (CO) pada Stasiun Kualitas Udara DKI1 Jakarta Menggunakan Random Forest", *algorithm*, vol. 2, no. 1, pp. 31-46 May 2026.

Journal homepage: https://ojs3.unpatti.ac.id/index.php/algorithm
Research Article **Open Access**

1. PENDAHULUAN

Polusi udara merupakan masalah lingkungan kritis yang mengancam kesehatan publik di wilayah metropolitan dunia, termasuk Jakarta [1]. Karbon monoksida (CO), senyawa tak berwarna dan berbau yang dihasilkan dari pembakaran bahan bakar fosil, merupakan polutan utama yang terutama berasal dari sektor transportasi (>80%) dan aktivitas industri [2]. Paparan CO dalam konsentrasi tinggi dapat mengganggu kemampuan darah dalam mengikat oksigen, sehingga berpotensi menimbulkan gangguan kesehatan, terutama pada kelompok rentan seperti anak-anak, lansia, dan penderita penyakit pernapasan. Di kawasan perkotaan dengan tingkat mobilitas tinggi, fluktuasi konsentrasi CO dapat terjadi secara dinamis akibat perubahan volume kendaraan, kondisi meteorologi, serta intensitas aktivitas manusia. Di Jakarta dengan populasi >10 juta jiwa dan volume lalu lintas ekstensif, konsentrasi CO sering melampaui standar baku mutu lingkungan yang ditetapkan Kementerian Lingkungan Hidup dan Kehutanan (KLHK) sebesar 10 mg/m^3 untuk 8 jam pemaparan [3].

Pemantauan dan peramalan kualitas udara menjadi komponen esensial dalam strategi manajemen lingkungan dan intervensi kesehatan masyarakat. Sistem prediksi akurat memungkinkan pengambilan keputusan proaktif untuk mitigasi polusi dan imbauan kesehatan publik [4]. Dalam konteks wilayah perkotaan, kemampuan memprediksi perubahan konsentrasi polutan seperti CO sangat penting karena kualitas udara dapat berubah secara cepat akibat aktivitas transportasi, kondisi cuaca, dan pola mobilitas harian. Informasi prediktif juga dapat membantu pemerintah dan pemangku kepentingan dalam menyusun kebijakan pengendalian emisi, pengaturan lalu lintas, serta peringatan dini kepada masyarakat pada periode risiko tinggiMetode tradisional berbasis statistical time series (ARIMA, *exponential smoothing*) seringkali tidak memadai dalam menangkap hubungan non linear kompleks dan perubahan struktural yang melekat dalam data polusi atmosfer [5]. Oleh karena itu, pendekatan berbasis *machine learning* menjadi alternatif yang relevan karena mampu mempelajari pola data historis secara lebih fleksibel dan menghasilkan prediksi yang lebih adaptif terhadap dinamika lingkungan.

Algoritma *machine learning*, khususnya metode *ensemble* seperti *Random Forest*, telah menunjukkan kinerja unggul dalam tugas peramalan lingkungan [6]. *Random Forest* menggabungkan beberapa *decision trees* dengan agregrasi *bootstrap (bagging)*, menghasilkan ketahanan terhadap *overfitting* dan kemampuan menangkap hubungan nonliniear yang kompleks tanpa asumsi distribusi ketat [7]. Keunggulan ini menjadikan *Random Forest* relevan untuk digunakan pada data kualitas udara yang umumnya memiliki pola fluktuatif, dipengaruhi oleh banyak faktor, serta tidak selalu mengikuti hubungan linear sederhana. Dalam konteks prediksi konsentrasi CO, *Random Forest* mampu mempelajari pola historis berdasarkan perubahan waktu dan variasi nilai polutan, sehingga dapat memberikan estimasi yang lebih stabil dibandingkan pendekatan statistik konvensional. Selain itu, sifat algoritma yang berbasis banyak pohon keputusan memungkinkan model mengenali pola lokal maupun perubahan ekstrem yang mungkin muncul pada periode tertentu. Optimasi *hyperparameter* melalui *Grid Search* memungkinkan eksplorasi sistematis dari ruang parameter yang luas untuk mengidentifikasi konfigurasi model yang optimal. Meskipun model *machine learning* telah banyak diaplikasikan, masih terbatas studi komprehensif yang memodelkan konsentrasi CO secara spesifik di Stasiun DKI1 pada rentang waktu kritis 2017-2021. Penelitian ini menawarkan kebaruan dengan menguji kemampuan optimasi *Random Forest* dalam menangkap pergeseran struktural data akibat dinamika polusi pra dan selama pandemi COVID-19 di titik pantau sentral Jakarta tersebut. Proses ini penting karena pemilihan parameter seperti jumlah pohon, kedalaman maksimum, dan jumlah fitur yang dipertimbangkan pada setiap pemisahan dapat memengaruhi akurasi serta kemampuan generalisasi model.

Meskipun model *machine learning* telah banyak diaplikasikan, masih terbatas studi komprehensif yang memodelkan konsentrasi CO secara spesifik di Stasiun DKI1 pada rentang waktu

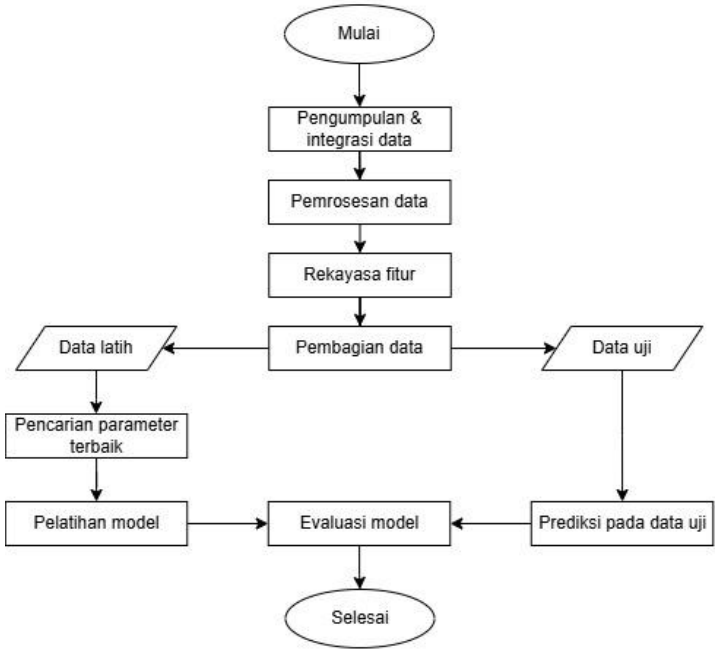
kritis 2017–2021. Rentang waktu tersebut penting karena mencakup kondisi sebelum, selama, dan setelah fase awal pandemi COVID-19 yang berpotensi menyebabkan perubahan pola aktivitas masyarakat, mobilitas transportasi, dan tingkat emisi di wilayah perkotaan. Penelitian ini menawarkan kebaruan dengan menguji kemampuan optimasi *Random Forest* dalam menangkap pergeseran struktural data akibat dinamika polusi pra dan selama pandemi COVID-19 di titik pantau sentral Jakarta tersebut.

2. METODE PENELITIAN

2.1 Alur Penelitian

Penelitian ini dilaksanakan melalui serangkaian tahapan sistematis yang dirancang untuk membangun model prediksi konsentrasi gas Karbon Monoksida (CO) yang akurat. Secara garis besar, kerangka kerja penelitian ini terdiri dari lima fase utama, yaitu: pengumpulan dan integrasi data, prapemrosesan data (*data preprocessing*), rekayasa fitur (*feature engineering*), pelatihan dan optimisasi model menggunakan algoritma *Random Forest*, serta diakhiri dengan evaluasi performa prediksi. Seluruh tahapan metodologi tersebut diilustrasikan secara komprehensif pada Gambar 1.

Gambar 1: diagram alir tahapan penelitian pemodelan *Random Forest*



2.2 Sumber Data dan Deskripsi Dataset

Data pencemaran udara DKI Jakarta yang digunakan di dalam penelitian ini disediakan oleh "Dinas Lingkungan Hidup" Provinsi DKI Jakarta pada situs web https://data.jakarta.go.id/. Data meteorologi diunduh dari https://dataonline.bmkg.go.id/).

Data dalam penelitian ini mencakup periode 1 Januari 2017 hingga 31 Maret 2021, yang terdiri dari 1.551 observasi mentah dengan lima variabel, yaitu: (1) Tavg (suhu harian rata-rata, °C), (2) RH_avg (kelembapan relatif rata-rata, %), (3) ss (lama penyinaran matahari, jam), (4) ff_avg (kecepatan angin rata-rata, m/s), dan (5) co_dki1 (konsentrasi CO harian rata-rata, μg/m^3). Hasil evaluasi kualitas data menunjukkan tidak terdapat data yang kosong (*missing values*) pada seluruh variabel. Hal ini mengindikasikan bahwa proses pencatatan telah dilakukan secara lengkap dengan protokol pemantauan yang konsisten.

2.3 Data Pre-processing dan Cleaning

Tahapan *pre-processing* meliputi tiga tahapan, yaitu:

1) Penguraian Indeks Tanggal

   Data mentah yang digunakan memiliki rentang waktu dari 1 Januari 2017 hingga 31 Maret 2021. Proses penguraian (*parsing*) indeks tanggal dilakukan untuk memastikan format waktu terbaca dengan benar oleh sistem. Hasil validasi menunjukkan tidak terdapat kegagalan pembacaan format tanggal (0 *error*), sehingga seluruh rentang observasi dapat diproses ke tahap selanjutnya.

2) Penanganan Data Duplikat

   Pemeriksaan integritas waktu mengidentifikasi adanya 11 tanggal yang terduplikasi, merepresentasikan sekitar 1,1% dari total data harian. Penanganan anomali ini dilakukan dengan menghitung nilai rata-rata (*averaging*) pada baris data yang memiliki tanggal identik. Pendekatan ini dipilih untuk menjaga konsistensi deret waktu. Setelah penghapusan duplikasi, dimensi data menyusut menjadi 1.540 observasi harian unik.

3) Verifikasi Konversi Numerik

   Langkah terakhir dalam prapemrosesan adalah memastikan seluruh fitur bertipe data numerik. Hasil verifikasi mengonfirmasi bahwa tidak ada nilai yang kosong (*zero missing values*) pada kelima variabel utama (Tavg, RH_avg, ss, ff_avg, dan co_dki1). Dataset akhir yang dihasilkan telah bersih dan terstruktur untuk diumpankan ke proses rekayasa fitur.

2.4 Rekayasa Fitur dan Fitur Temporal

Rekayasa fitur (*feature engineering*) menerapkan transformasi khusus deret waktu (*time series*) untuk menangkap ketergantungan temporal yang melekat pada data polusi udara. Fitur *lag* (jeda waktu) dibuat pada beberapa langkah waktu (*time steps*), yaitu: CO_lag_1 (t-1), CO_lag_2 (t-2), CO_lag_3 (t-3), dan CO_lag_7 (t-7), yang merepresentasikan nilai konsentrasi CO pada 1 hari, 2 hari, 3 hari, dan 7 hari sebelumnya.

Secara keseluruhan, himpunan fitur (*feature set*) terdiri dari 12 prediktor, dengan rincian: 4 fitur meteorologi + 4 fitur *lag* + 2 fitur rata-rata bergulir + 2 fitur simpangan baku bergulir. Baris-baris awal data yang mengandung nilai kosong (*NaN*) akibat pembuatan fitur *lag* (karena membutuhkan data 7 hari sebelumnya) dihapus setelah proses rekayasa fitur selesai. Eliminasi nilai *NaN* ini menghasilkan 1.533 observasi yang efektif untuk digunakan.

Tabel 1. *deskripsi fitur prediktor setelah rekayasa fitur*

Nama Fitur	Kategori	Deskripsi	Notasi Matematis
Tavg, RH_avg, ss, ff_avg	Meteorologi	Data cuaca harian (Suhu, Kelembapan, Penyinaran Matahari, dan Kecepatan Angin)	X_t
co_lag_k	Temporal (Lag)	Konsentrasi CO pada k hari sebelumnya dengan nilai k ∈ {1, 2, 3, 7}	CO_{t-k}
co_roll_mean_w	Temporal (Rolling)	Rata-rata pergerakan konsentrasi CO selama w hari	$\frac{1}{w}\sum_{i=1}^{w} CO_{t-i}$

Nama Fitur	Kategori	Deskripsi	Notasi Matematis
		sebelumnya dengan nilai w ∈ {3, 7}	
co_roll_std_w	Temporal (*Rolling*)	Simpangan baku pergerakan konsentrasi CO selama w hari sebelumnya dengan nilai w ∈ {3, 7}	$\sqrt{\frac{1}{w-1}\sum_{i=1}^{w}(CO_{t-i}-\mu)^2}$

2.5 Pembagian Data dan Strategi Cross-Validation

Untuk menjaga integritas temporal dan mencegah kebocoran data (*data leakage*) dalam konteks deret waktu (*time series*), diterapkan pembagian data secara kronologis dengan rasio 80:20. Pembagian ini terdiri dari data latih (*training set*) sebanyak 1.226 sampel (8 Januari 2017 – 17 Mei 2020) dan data uji (*testing set*) sebanyak 307 sampel (18 Mei 2020 – 31 Maret 2021). Cross validation menggunakan TimeSeriesSplit dengan 5 lipatan (*folds*) yang secara inheren mempertahankan urutan waktu. Sebagai contoh, pada lipatan pertama, 20% data awal digunakan untuk pelatihan, sedangkan pada lipatan terakhir, 20% data terbaru digunakan untuk validasi. Pendekatan ini berfungsi untuk mencegah kebocoran informasi dari data masa depan, sebuah masalah yang umum terjadi apabila menggunakan validasi silang naif (*naive cross-validation*) pada pemodelan deret waktu [8].

2.6 Pelatihan Model dengan *Random Forest*

Tahap pelatihan dilakukan menggunakan algoritma dasar *Random Forest Regressor*. *Random Forest* merupakan metode pembelajaran *ensemble* berbasis pohon keputusan (*decision trees*) yang digabungkan melalui prinsip *Bootstrap Aggregating* (*Bagging*)[7]. Algoritma ini bekerja dengan melatih banyak pohon keputusan secara paralel menggunakan sampel acak dengan pengembalian (*with replacement*) dari subset data latih.

Untuk semakin meminimalkan korelasi antar-pohon tunggal dan mencegah *overfitting*, pemisahan percabangan (*node split*) pada setiap pohon hanya mempertimbangkan sebagian kecil fitur input acak yang ditentukan secara ketat[9]. Karena penelitian ini merupakan kasus regresi, hasil prediksi akhir dari arsitektur *Random Forest* diperoleh dengan menghitung nilai rata-rata (*mean*) dari seluruh estimasi yang dikeluarkan oleh pohon individu di dalam model *ensemble* tersebut[8]. Pendekatan matematika agregasi ini dirumuskan sebagai berikut:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b\,(x) \qquad (1)$$

di mana B merepresentasikan jumlah total pohon keputusan (n_estimators), dan T_b(x) adalah hasil estimasi prediksi dari pohon keputusan ke-b untuk input data x. Melalui mekanisme ini, varians kesalahan model dapat ditekan secara signifikan tanpa mengorbankan bias secara berlebih.

2.7 Optimisasi Hiperparameter dengan *Grid Search*

Metode *Grid Search* secara menyeluruh mengeksplorasi 324 kombinasi parameter. Setiap kombinasi dievaluasi melalui 5 lipatan TimeSeriesSplit menggunakan *negative RMSE* sebagai metrik penilaian (*scoring metric*). Proses ini secara total menghasilkan 1.620 iterasi pelatihan model (*model fits*). Himpunan parameter terbaik (*best parameter set*) kemudian diidentifikasi berdasarkan nilai rata-rata dari skor validasi silangnya.

2.8 Model Evaluation Metrics

Kinerja model dievaluasi menggunakan empat metrik yang saling melengkapi [9]:

1) Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

Metrik ini memberikan interpretasi kesalahan (*error*) dalam satuan aslinya (μg/m^3) dengan memberikan bobot yang sama pada semua besaran nilai.

2) Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (3)$$

Metrik ini menghitung rata-rata dari selisih kuadrat antara nilai aktual dan nilai prediksi. MSE sangat berguna untuk mengukur stabilitas varians eror model.

3) Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (4)$$

Metrik ini memberikan penalti kuadratik pada kesalahan prediksi yang lebih besar, sehingga metrik ini sangat penting untuk menangkap kejadian polusi yang ekstrem.

4) Skor R^2 (R^2 *Score*)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y})^2} \quad (5)$$

Skor ini mengindikasikan proporsi varians yang dapat dijelaskan oleh model dengan rentang nilai antara 0 hingga 1.

5) Analisis Residu (*Residual Analysis*)

Kombinasi dari metrik-metrik tersebut memberikan penilaian yang komprehensif terhadap kemampuan prediktif sekaligus batasan dari model.

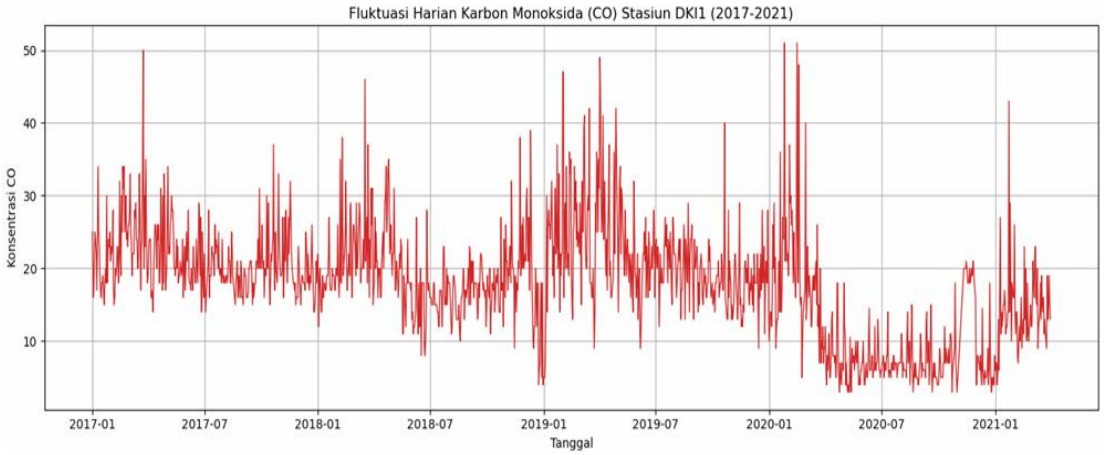
3. HASIL DAN PEMBAHASAN

3.1 Eksplorasi dan Visualisasi Data

Gambar 2 memvisualisasikan fluktuasi harian konsentrasi CO di Jakarta selama periode observasi (2017–2021). Data tersebut menunjukkan pola musiman (*seasonal patterns*) yang jelas. Konsentrasi tertinggi terjadi pada musim hujan (Desember–Februari) yang mencapai lebih dari 50 μg/m^3, sedangkan konsentrasi terendah terjadi pada musim kemarau (Juni–Agustus) pada kisaran 5–15 μg/m^3.

Pola ini konsisten dengan karakteristik meteorologis di wilayah tersebut. Musim hujan ditandai dengan penurunan ketinggian percampuran atmosfer (*atmospheric mixing height*) dan peningkatan inversi suhu yang menjebak polutan di dekat permukaan tanah. Sebaliknya, musim kemarau yang disertai kecepatan angin lebih tinggi dan lapisan percampuran (*mixing layer*) yang lebih dalam akan memfasilitasi proses dispersi polutan. Tren jangka panjang (*long-term*) menunjukkan adanya penurunan secara bertahap dari tahun 2017 hingga 2019, yang kemudian diikuti oleh fase stabilisasi pada periode 2020–2021. Hal ini kemungkinan besar dipengaruhi oleh implementasi kebijakan lingkungan dan menurunnya emisi lalu lintas selama pandemi COVID-19 [10]

Gambar 2. Fluktuasi harian karbon monoksida (CO) Stasiun DKI1 (2017-2021)



Tabel 2 merangkum hasil eksplorasi data dengan rincian jumlah data sebanyak 1.540 observasi, tidak terdapat data yang kosong (*missing values*) pada seluruh variabel, dan cakupan rentang waktu pemantauan penuh selama 4,25 tahun. Statistik deskriptif menunjukkan bahwa suhu rata-rata (Tavg) berada pada rentang 24,2–30,9°C (rata-rata 28,9°C), kelembapan (RH_avg) 70,6–81,2% (rata-rata 74,9%), durasi sinar matahari (ss) 1,1–8,2 jam (rata-rata 5,1 jam), kecepatan angin (ff_avg) 1–2 m/s (rata-rata 1,4 m/s), dan konsentrasi CO (co_dki1) pada rentang 3–50 µg/m^3 (rata-rata 15,2 µg/m^3). Koefisien variasi tertinggi terdapat pada variabel co_dki1 (1,24), yang mengindikasikan adanya volatilitas tinggi pada variabel target tersebut, sementara variabel meteorologi lainnya cenderung relatif stabil.

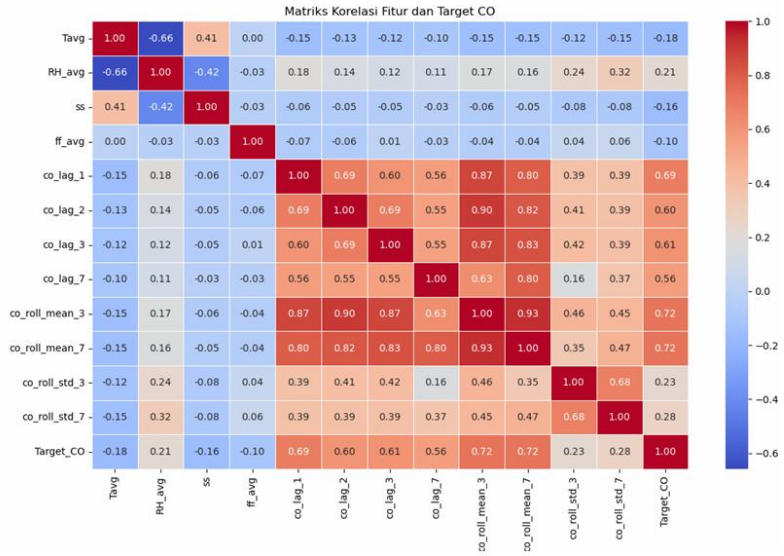
Tabel 2. *Statistik Deskriptif dan Ringkasan Data Eksplorasi Stasiun DKI1 Jakarta (2017–2021)*

Variabel	Nilai (Range / Mean)	Satuan	Keterangan
Jumlah observasi	1.540	-	Data harian unik setelah proses pembersihan (*cleaning*) dan penghapusan duplikasi.
Missing values	0	-	Tidak terdapat data yang hilang pada semua variabel.
Temporal range	Jan 2017 – Mar 2021	-	4,25 tahun data pemantauan lengkap.
Tavg	24,2 – 30,9 (mean 28,9)	°C	Suhu harian rata-rata relatif stabil.
RH_avg	70,6 – 81,2 (mean 74,9)	%	Kelembapan relatif harian rata-rata.
ss	1,1 – 8,2 (mean 5,1)	jam	Durasi penyinaran matahari per hari.
ff_avg	1 – 2 (mean 1,4)	m/s	Kecepatan angin harian rata-rata.
co_dki1	3 – 50 (mean 15,2)	µg/m^3	Koefisien variasi tertinggi (CV = 1,24), volatilitas tinggi.

3.2 Korelasi Fitur dan Analisis Target

Berdasarkan matriks korelasi Pearson, suhu dan kelembapan memiliki korelasi negatif yang kuat, mencerminkan karakteristik iklim tropis Jakarta. Dalam memprediksi konsentrasi CO, data historis (*lag*) menunjukkan autokorelasi positif yang kuat, di mana rata-rata konsentrasi polusi dalam beberapa hari terakhir (*rolling statistics*) terbukti menjadi prediktor terbaik dibandingkan data jeda waktu harian tunggal. Sebaliknya, variabel meteorologi hanya berperan sebagai prediktor sekunder karena korelasinya terhadap CO tergolong lemah hingga sedang. Meskipun terdapat multikolinearitas yang tinggi antarfitur temporal, hal tersebut bukan merupakan masalah karena pemodelan menggunakan algoritma *Random Forest* yang memiliki kemampuan seleksi fitur bawaan.

Gambar 3. Matriks korelasi fitur dan target



3.3 Optimisasi *Grid Search*

Evaluasi *Grid Search* terhadap 324 kombinasi parameter melalui 5 lipatan (*folds*) TimeSeriesSplit berhasil mengidentifikasi konfigurasi yang paling optimal dengan nilai RMSE validasi silang (*cross-validation*) terbaik sebesar 5,329 μg/m^3. Parameter optimal dari hasil identifikasi tersebut disajikan pada Tabel 3.

Tabel 3. Hasil Optimisasi *Grid Search*

Parameter	Optimal Value
n_estimators	300
max_depth	10
min_samples_split	10
min_samples_leaf	4
max_features	'sqrt'

3.4 Performa Model pada Training dan Testing Data

Tabel 4. Performa Metrik Optimisasi Model *Random Forest*

Metric	*Training Set*	*Testing Set*
MAE (μg/m^3)	2.7383	3.4505
MSE	14.8290	18.6761
RMSE (μg/m^3)	3.8508	4.3216
R^2 Score	0.6545	0.3741

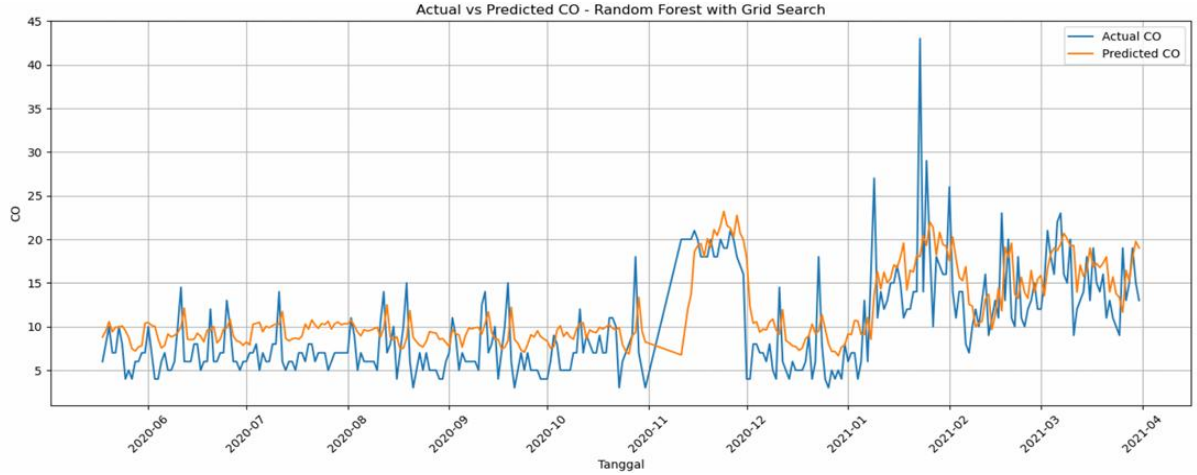
Nilai R^2 pada data latih sebesar 0,6545 menunjukkan bahwa model mampu menjelaskan 65,45% varians pada data latih. Hal ini mengindikasikan tingkat kesesuaian (*fit*) yang memadai terhadap pola historis. Penurunan nilai R^2 pada data uji menjadi 0,3741 (turun sebesar 0,2804 atau 42,8%) mengindikasikan adanya celah generalisasi (*generalization gap*) tingkat sedang. Kondisi ini wajar terjadi dalam peramalan deret waktu (*time series forecasting*) ketika melakukan ekstrapolasi pada periode masa depan yang belum dikenali oleh model (*unseen future periods*).

Nilai MAE pengujian sebesar 3,45 μg/m^3 dan RMSE sebesar 4,32 μg/m^3 memberikan besaran galat (*error*) yang dapat diinterpretasikan secara langsung melalui satuan aslinya. Artinya, prediksi model rata-rata menyimpang sebesar 3,45 satuan dari nilai aktualnya. Peningkatan metrik galat dari tahap pelatihan ke pengujian (MAE naik 26%, RMSE naik 12,2%) mengonfirmasi terjadinya *overfitting* ringan. Meskipun demikian, besaran yang tergolong sedang tersebut menunjukkan bahwa model tidak mengalami *overfitting* yang parah (*severely overfit*). Nilai RMSE validasi silang (*cross-validation*) sebesar 5,329, apabila disandingkan dengan nilai RMSE pelatihan (3,85) dan pengujian (4,32), turut mengonfirmasi konsistensi serta keandalan dari proses pemilihan hiperparameter.

3.5 Analisis Prediksi Temporal dan Visualisasi Aktual vs Prediks

Visualisasi aktual versus prediksi seperti yang ditunjukkan oleh Gambar 4 digunakan untuk mengevaluasi kemampuan model dalam mengikuti pola temporal konsentrasi CO sepanjang periode pengamatan. Melalui grafik ini, dapat diamati sejauh mana hasil prediksi *Random Forest* mampu merepresentasikan fluktuasi nilai aktual CO, termasuk pada periode dengan perubahan konsentrasi yang relatif rendah maupun saat terjadi peningkatan nilai secara tiba-tiba.

Gambar 4. Plot Aktual vs Prediksi CO



Periode 1 (Mei–Juli 2020): Model menunjukkan kemampuan pelacakan yang sangat baik terhadap nilai aktual, dimana hasil prediksi mengikuti fluktuasi CO yang diobservasi secara ketat. Prediksi model mampu menangkap tren peningkatan (lonjakan pada bulan Mei) maupun tren

penurunan (penurunan selama Juni–Juli), dengan galat prediksi (*prediction errors*) yang minimal (±2–3 μg/m^3).

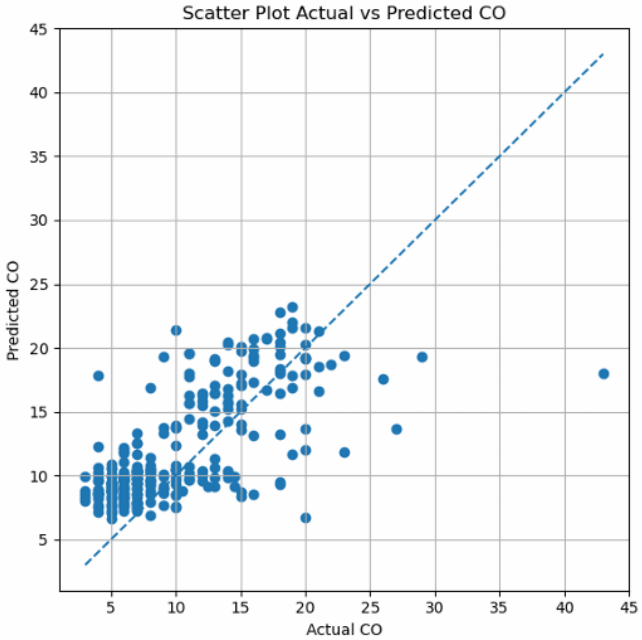
Selanjutnya, Periode 2 (Agustus–Desember 2020): Model terus memberikan prediksi yang akurat dengan pelacakan yang cukup baik, meskipun terlihat adanya sedikit keterlambatan (*lagging*) dalam merespons perubahan secara cepat. Hasil prediksi mempertahankan rentang yang relatif stabil pada angka 10–15 μg/m^3, yang mana sejalan dengan pola musiman di wilayah tersebut (musim kemarau umumnya mencatatkan konsentrasi polusi yang lebih rendah).

Sementara itu, Periode 3 (Januari–Maret 2021): Pada periode ini, terjadi prediksi rendah yang sistematis (*systematic underprediction*). Nilai aktual CO melonjak drastis hingga 42–43 μg/m^3 (bertepatan dengan puncak episode polusi di musim hujan), sementara prediksi model tertahan pada angka 20–22 μg/m^3. Kesenjangan (*discrepancy*) yang mencapai lebih dari 20 μg/m^3 ini merepresentasikan galat prediksi terbesar di sepanjang periode pengujian. Pola ini mencerminkan bias bawaan dari algoritma *Random Forest* terhadap pembalikan ke nilai rata-rata (*mean reversion*)—kejadian ekstrem yang berada di luar rentang distribusi data latih akan selalu diprediksi lebih rendah dari nilai aslinya.

3.6 Analisis Diagram Pencar dan Distribusi Kesesuaian Model

Berdasarkan Gambar 5, diagram pencar (*scatter plot*), nilai CO aktual dan prediksi menunjukkan kesesuaian yang sangat baik pada rentang konsentrasi rendah hingga normal (5–15 μg/m^3), di mana titik-titik data mengelompok secara rapat di sepanjang garis referensi diagonal. Namun, penyimpangan sistematis berupa prediksi rendah (*underprediction*) mulai terlihat jelas saat konsentrasi polusi meningkat (>20 μg/m^3), dengan kesenjangan nilai prediksi yang terpangkas hingga mencapai selisih ~20 μg/m^3 pada kondisi ekstrem (42–43 μg/m^3). Analisis sebaran ini memberikan wawasan penting bahwa kinerja model tidak seragam dan sangat bergantung pada tingkat polusi, di mana nilai R^2 diperkirakan berada di atas 0,70 pada kondisi normal, tetapi anjlok hingga di bawah 0,30 saat menghadapi episode polusi ekstrem.

Gambar 5. Scatter Plot Aktual vs Prediksi CO

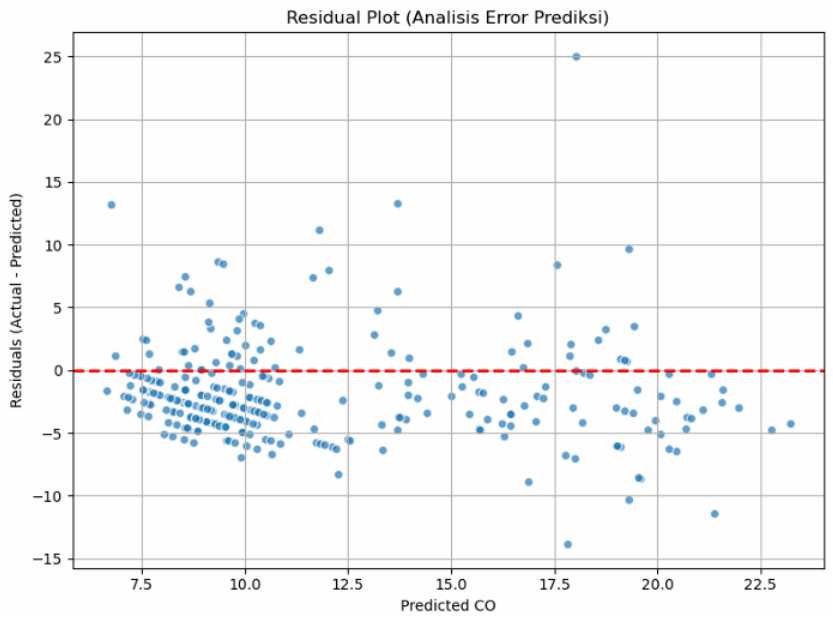


3.7 Pelatihan Model Akhir dan Evaluasi Data Uji

Plot residual pada Gambar 6 memvisualisasikan selisih antara nilai aktual dan prediksi (*residual = aktual - prediksi*) terhadap besaran nilai prediksi untuk mengidentifikasi perilaku galat model. Hasil plot menunjukkan adanya pola heteroskedastisitas yang jelas, di mana sebaran residual melebar secara sistematis membentuk pola terompet atau kipas (*trumpet/fan shape*). Pada prediksi rendah (5–8 µg/m^3), dispersi residual cenderung rapat (±2–3 µg/m^3), sedangkan pada prediksi tinggi (15–25 µg/m^3), rentang residual membentang lebar hingga ±15 µg/m^3. Pola ketidaksetaraan varians galat ini melanggar asumsi homoskedastisitas, yang mengindikasikan perlunya perbaikan model di masa mendatang melalui transformasi penstabil varians (*variance-stabilizing transformations*) atau penggunaan regresi kuantil (*quantile regression*).

Selain masalah varians, plot ini juga mengonfirmasi adanya bias sistematis pada nilai-nilai ekstrem. Ketika nilai prediksi melebihi 15 µg/m^3, kumpulan titik residual secara dominan berada di bawah garis referensi galat-nol (y=0), dengan rata-rata residual yang semakin negatif seiring meningkatnya nilai prediksi. Kecenderungan ini mencerminkan bias prediksi rendah yang sistematis (*systematic underprediction*), di mana model secara konsisten menghasilkan estimasi yang lebih rendah dari nilai aktual pada saat terjadi polusi tinggi, dengan besaran bias mencapai -10 hingga -15 µg/m^3. Adapun beberapa pencilan (*outliers*) ekstrem—seperti titik dengan residual ~ +25 µg/m^3 pada prediksi 8 µg/m^3 —merepresentasikan lonjakan polusi mendadak yang gagal diantisipasi oleh model akibat perubahan atmosfer yang cepat atau episode meteorologi tidak biasa yang kurang terwakili di dalam data latih.

Gambar 6. Plot Residual



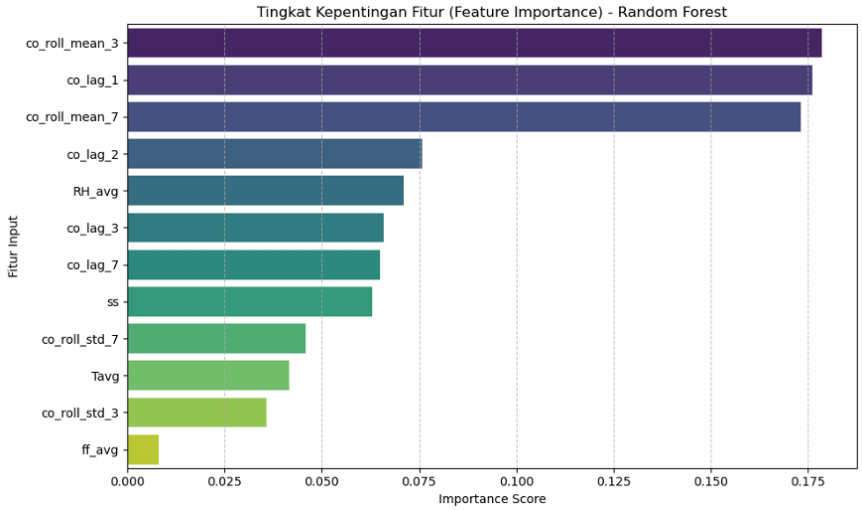
3.8 Analisis Tingkat Kepentingan Fitur dan Dominasi Prediktor

Skor tingkat kepentingan fitur (*feature importance scores*) yang dihitung melalui model *Random Forest* berfungsi untuk mengidentifikasi kontribusi relatif dari setiap prediktor terhadap kemampuan prediktif model. Tabel 5 menyajikan peringkat lengkap dari 12 fitur yang digunakan, sementara Gambar 7 memvisualisasikannya ke dalam bentuk diagram batang horizontal (*horizontal bar chart*).

Tabel 5. Ranking 12 Fitur Lengkap

Rank	*Feature*	*Importance*	*% Cumulative*	*Type*
1	CO_roll_mean_3	0.1787	17.87%	Temporal
2	CO_lag_1	0.1762	35.49%	Temporal
3	CO_roll_mean_7	0.1734	52.83%	Temporal
4	CO_lag_2	0.0758	60.41%	Temporal
5	RH_avg	0.0711	67.52%	Meteorological
6	CO_lag_3	0.0659	74.11%	Temporal
7	CO_lag_7	0.0650	80.61%	Temporal
8	ss (sunshine)	0.0628	86.89%	Meteorological
9	co_roll_std_7	0.0458	91.47%	Temporal
10	Tavg	0.0415	95.62%	Meteorological
11	co_roll_std_3	0.0356	99.18%	Temporal
12	ff_avg	0.0081	100.00%	Meteorological

Gambar 7. Plot tingkat kepentingan fitur



Analisis tingkat kepentingan fitur menunjukkan bahwa fitur temporal mendominasi daya prediktif model dengan total kontribusi mencapai 52,83%. Rata-rata bergulir 3 hari dan *lag*-1 terbukti menjadi prediktor terkuat karena kemampuannya dalam menangkap momentum serta memperhalus fluktuasi polusi harian. Sebaliknya, variabel meteorologi secara keseluruhan hanya memberikan kontribusi sekunder sekitar 18%. Secara spesifik, kecepatan angin rata-rata justru menjadi prediktor terlemah dengan skor 0,008; sebuah anomali yang kemungkinan disebabkan oleh hilangnya variabilitas sub-harian pada metrik rata-rata harian, rendahnya profil angin secara umum di Jakarta, atau karena efek angin tersebut sebenarnya telah tertangkap secara tidak langsung di dalam persistensi temporal data polusi itu sendiri.

4. Pembahasan

4.1 Interpretasi Hasil Model dan Penilaian Kinerja

Model *Random Forest* yang telah dioptimisasi mencapai nilai RMSE pengujian sebesar 4,3216 $\mu g/m^3$ dan R^2 sebesar 0,3741. Angka ini merepresentasikan kemampuan prediktif tingkat sedang,

yang mengindikasikan adanya kompromi (*trade-off*) antara utilitas model dan kebutuhan akurasi. Sebagai konteks, standar baku mutu CO di Jakarta (10 mg/m^3 untuk paparan 8 jam [3]) setara dengan 10.000 µg/m^3, sedangkan rata-rata konsentrasi harian yang diteliti dalam studi ini hanya berada pada rentang 3–50 µg/m^3. Nilai RMSE model sebesar 4,32 µg/m^3 merepresentasikan sekitar 29% galat relatif terhadap rata-rata tingkat CO (14,8 µg/m^3). Tingkat galat ini masih dapat diterima dalam aplikasi peramalan lingkungan, di mana ketepatan angka prediksi tidak lebih krusial dibandingkan kemampuan model dalam menangkap tren dan fluktuasi relati [11].

Kesenjangan kinerja antara tahap pelatihan dan pengujian (selisih R^2 sebesar 0,2804) merupakan fenomena yang wajar terjadi dalam peramalan deret waktu (*time series forecasting*). Hal ini dikarenakan model dilatih menggunakan pola historis yang mungkin tidak bertahan secara identik pada periode masa depan [12]. Celah generalisasi (*generalization gap*) yang tergolong sedang (sekitar 43%) ini menunjukkan bahwa model tidak mengalami *overfitting* yang parah. Meskipun demikian, celah tersebut berpotensi untuk ditekan lebih lanjut melalui peningkatan regularisasi pada optimisasi hiperparameter (seperti memperkecil nilai max_depth dan memperbesar min_samples_leaf) atau melalui penyempurnaan teknik validasi silang. Konfigurasi model saat ini telah merepresentasikan keseimbangan praktis antara kemampuan menangkap pola pada data latih dan kemampuan mempertahankan generalisasi prediksi.

4.2 Analisis Pola Temporal dan Dinamika Musiman

Dominasi fitur temporal (52,83%) mengungkapkan adanya dinamika autoregresif yang kuat pada konsentrasi CO di Jakarta. Autokorelasi yang signifikan (korelasi *lag*-1 sebesar 0,69) mengindikasikan bahwa nilai polusi CO sangat persisten—konsentrasi pada hari sebelumnya sangat kuat dalam memprediksi tingkat polusi hari ini. Hal ini mencerminkan lambatnya proses percampuran atmosfer dan tingginya akumulasi polutan. Kinerja fitur rata-rata bergulir (*rolling mean*) yang mengungguli nilai *lag* tunggal (17,87% berbanding 17,62%) turut mengonfirmasi bahwa tren polusi yang mendasarinya jauh lebih informatif dibandingkan sekadar melihat fluktuasi dari hari ke hari [13].

Tingkat kepentingan *lag* 7 hari (0,0650) merepresentasikan siklus mingguan dalam pola CO di Jakarta. Kondisi ini kemungkinan besar didorong oleh siklus lalu lintas mingguan (volume kendaraan yang lebih tinggi pada hari kerja berbanding dengan penurunannya pada akhir pekan) yang dikombinasikan dengan variasi pola angin mingguan. Kehadiran fitur *lag* jangka pendek (3 hari) maupun jangka panjang (7 hari) sebagai prediktor teratas mengonfirmasi adanya ketergantungan temporal multiskala (*multi-scale temporal dependencies*). Algoritma *Random Forest* terbukti mampu menangkap ketergantungan ini secara efektif melalui teknik ansambel yang menggabungkan berbagai perspektif dari banyak pohon keputusan (*ensembling multiple tree perspectives*).

4.3 Variabel Meteorologi: Faktor Penggerak Sekunder vs Persistensi Temporal

Rendahnya tingkat kepentingan variabel meteorologi (18% secara kumulatif) tidak mengindikasikan bahwa faktor cuaca tidak relevan. Kondisi ini menunjukkan bahwa untuk peramalan skala harian, faktor persistensi temporal jauh lebih mendominasi dibandingkan dorongan meteorologis [14]. Korelasi yang lemah antara suhu dan konsentrasi CO (-0,18, negatif lemah) mengungkapkan adanya hubungan yang kompleks. Di satu sisi, suhu yang lebih tinggi umumnya memfasilitasi proses percampuran atmosfer dan dispersi polutan (efek negatif). Di sisi lain, peningkatan suhu berpotensi meningkatkan volume lalu lintas harian yang pada akhirnya menaikkan emisi gas buang.

Korelasi positif pada kelembapan relatif (0,21) merepresentasikan kondisi bulan-bulan basah atau musim hujan, yang secara simultan menunjukkan kelembapan tinggi dan konsentrasi CO yang meningkat akibat penekanan lapisan percampuran atmosfer (*mixing layer suppression*) selama

musim tersebut. Sementara itu, sangat kecilnya pengaruh kecepatan angin (0,008) mencerminkan adanya keterbatasan dalam pengukuran data. Penggunaan nilai rata-rata kecepatan angin harian cenderung menyamarkan variasi hembusan angin sub-harian (*sub-daily gust variations*) yang sebenarnya mengontrol proses percampuran polutan secara riil. Ditambah lagi, karakteristik wilayah Jakarta yang umumnya didominasi oleh regime angin tenang (*calm wind*) turut membatasi daya penjelas dari variabel tersebut.

4.4 Prediksi Rendah Nilai Ekstrem dan Keterbatasan Model

Arsitektur *Random Forest* memiliki keterbatasan fundamental berupa prediksi rendah yang sistematis (*systematic underprediction*) pada kejadian polusi ekstrem akibat mekanisme rata-rata ansambel yang cenderung meredam nilai-nilai prediksi tertinggi. Model ini juga menunjukkan pola heteroskedastisitas di mana varians ketidakpastian tidak seragam, sehingga penggunaan interval prediksi standar menjadi tidak akurat saat terjadi lonjakan emisi. Keterbatasan ini sangat problematis dalam manajemen kualitas udara karena berpotensi menunda sistem peringatan dini kesehatan masyarakat yang sangat bergantung pada deteksi episode ekstrem. Oleh karena itu, diperlukan pendekatan hibrida—seperti integrasi dengan *quantile regression*, *mixture models*, atau *gradient boosting*—guna menyempurnakan keandalan prediksi *Random Forest* dalam menangkap risiko polusi ekstrem secara presisi.

4.5 Perbandingan dengan Metode Peramalan Kualitas Udara yang Ada

Kinerja *Random Forest* dalam penelitian ini (R^2 0,3741; RMSE 4,32 μg/m^3) terbukti sangat kompetitif dan sebanding dengan metode statistik tradisional seperti ARIMA. Dibandingkan ARIMA, *Random Forest* menawarkan keunggulan komparatif karena tidak memerlukan asumsi stasioneritas maupun transformasi data nonlinear, serta mampu memberikan interpretasi fitur yang lebih baik, terlepas dari kelemahannya dalam memprediksi kejadian ekstrem dan beban komputasi yang lebih tinggi. Di sisi lain, meskipun pendekatan *deep learning* (seperti LSTM) menjanjikan akurasi yang lebih superio, metode tersebut menuntut penggunaan data yang sangat masif (>5.000 observasi). Oleh karena itu, penggunaan *Random Forest* tetap menjadi pilihan pemodelan yang paling praktis dan rasional untuk diimplementasikan pada *dataset* penelitian ini yang hanya berjumlah 1.540 observasi efektif.

4.6 Pertimbangan Implementasi untuk Peramalan Operasional

Untuk menerapkan (*deployment*) model ini ke dalam sistem peramalan kualitas udara yang beroperasi secara riil, terdapat beberapa pertimbangan praktis yang penting untuk diperhatikan :

1) Pelatihan Ulang Model (*Model Retraining*)

   Pelatihan ulang model secara berkala sangat disarankan (baik secara kuartalan maupun semesteran) untuk mengakomodasi pola data terbaru dan menangkap pergeseran jangka panjang (*long-term shifts*) pada karakteristik kualitas udara di Jakarta. Langkah ini sangat penting untuk mencegah degradasi kinerja model yang diakibatkan oleh pergeseran distribusi temporal (*temporal distribution shift*).

2) Pendekatan Ansambel Hibrida (*Ensemble Hybrid Approach*)

   Pendekatan ini mengombinasikan hasil prediksi *Random Forest* dengan regresi kuantil (*quantile regression*) atau batas kepercayaan (*confidence bounds*) dari model ARIMA untuk menyediakan peramalan yang bersifat probabilistik. Penggunaan ansambel hibrida pada umumnya mampu meningkatkan keandalan model, baik pada hasil prediksi titik (*point forecasts*) maupun estimasi interval (*interval estimates*), jika dibandingkan dengan pendekatan model tunggal.

3) Kualitas Data Masukan Waktu Nyata (*Real-time Input Quality*)

Keandalan model sangat bergantung pada kualitas pengukuran historis gas CO dan data meteorologi. Oleh karena itu, pemeriksaan kualitas data secara otomatis (*automated data quality checks*)—yang mencakup deteksi pencilan (*outlier detection*) dan penanganan data yang kosong (*missing value handling*)—menjadi sangat esensial untuk mencegah keluaran prediksi yang buruk akibat data masukan yang tidak valid (prinsip *garbage-in, garbage-out*).

4) Pemantauan Kinerja (*Performance Monitoring*)

Pemantauan yang berkelanjutan terhadap metrik akurasi peramalan (seperti RMSE, MAE, dan R^2) harus dilakukan secara saksama dengan membandingkannya terhadap nilai observasi aktual. Apabila kinerja model mengalami penurunan hingga berada di bawah ambang batas yang telah ditetapkan (sebagai contoh, nilai RMSE pengujian >5.5 μg/m^3), kondisi tersebut harus memicu (*trigger*) proses pelatihan ulang model atau investigasi lebih lanjut mengenai potensi perubahan distribusi data.

5. Kesimpulan

Penelitian ini mengoptimasi model *Random Forest* untuk memprediksi konsentrasi karbon monoksida (CO) di Jakarta dengan cukup baik untuk memantau rentang polusi normal (5–20 μg/m^3) meski memiliki keterbatasan bawaan berupa *underprediction* pada tingkat polusi ekstrem. Hasil analisis secara jelas menegaskan bahwa riwayat historis polusi CO (fitur temporal) jauh lebih krusial dan prediktif dalam menentukan kualitas udara harian dibandingkan dengan faktor meteorologi. Secara praktis, kerangka model ini aplikatif untuk dimanfaatkan dalam prospek kualitas udara harian, manajemen emisi industri dan lalu lintas, serta sistem peringatan kesehatan masyarakat.

REFERENSI

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

[2] J. H. Seinfeld and S. N. Pandis, *Atmospheric chemistry and physics: from air pollution to climate change*. John Wiley & Sons, 2016.

[3] Pemerintah Republik Indonesia, *Peraturan Pemerintah Nomor 22 Tahun 2021 tentang Penyelenggaraan Perlindungan*. Indonesia: Sekretariat Negara, 2021.

[4] P. Holnicki, A. Kałuszko, and Z. Nahorski, "A statistical approach to air pollution forecasting and visualization.," *Environ. Monit. Assess.*, vol. 190, no. 5, p. 290, 2018.

[5] M. Zhu, Z. Huang, and Y. Wu, ""A review of the application of machine learning in air quality estimation.," *Aerosol Air Qual. Res.*, vol. 18, pp. 1067–1083, May 2018.

[6] D. Stathakis, "How many hidden layers and nodes?," *Int. J. Remote Sens.*, vol. 30, no. 8, pp. 2133–2147, 2009.

[7] L. Breiman, "*Random Forest*s," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[8] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[9] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning," 2009, *Springer series in statistics New-York*.

[10] J. P. Putaud and et al, "Changes in air quality related to the COVID-19 lockdown in Europe and China. Atmosphere," 2020.

[11] X. Chen and H. Ishwaran, "*Random Forest*s for time series forecasting," *J. Forecast.*, 2018.

[12] J. S. Armstrong, *Principles of forecasting: a handbook for researchers and practitioners*, vol. 30. Springer, 2001.

[13] C. Chatfield and H. Xing, *The analysis of time series: an introduction with R*. Chapman and hall/CRC, 2019.

[14] D. J. Jacob, "Introduction to atmospheric chemistry," 1999.