Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 3, June 2026, Page. 2154-2164
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5412

Comparative Analysis of the Performance of *Random Forest* and *CatBoost* for Air Quality Prediction Based on Meteorological Factor

Nirsal*[1], Nurchaerani Kadir[2]

[1,2]Informatika, Universitas Cokroaminoto Palopo, Indonesia

Email: [1]nirsal@uncp.ac.id

Received : Oct 30, 2025; Revised : Dec 19, 2025; Accepted : Dec 30, 2025; Published : Jun 15, 2026

Abstract

Air quality in urban centers such as Tangerang City has become an increasingly urgent issue due to the expansion of industrial activities, rapid population growth, and rising vehicle emissions. As a key city within the Greater Jakarta metropolitan area, Tangerang is highly vulnerable to air pollution caused by human activities and varying meteorological conditions. This study aims to assess the performance of two machine learning algorithms, Random Forest and CatBoost, in predicting air quality in Tangerang under two scenarios: models that incorporate meteorological factors and models that exclude them. The dataset includes concentrations of key air pollutants alongside meteorological variables such as temperature, humidity, and wind speed. Model performance was evaluated using MAE, MSE, RMSE, and R^2. The findings indicate that both algorithms perform excellently when meteorological variables are included. Random Forest achieved an MAE of 0.0099, MSE of 0.000309, RMSE of 0.0152, and an R^2 of 0.9931, slightly outperforming CatBoost, which recorded an MAE of 0.0135, MSE of 0.000419, RMSE of 0.0170, and an R^2 of 0.9907. Excluding meteorological variables decreased accuracy for both models, with Random Forest reaching an R^2 of 0.9519 and CatBoost 0.9487. These results underscore the importance of temperature, humidity, and wind speed in enhancing predictive accuracy. Notably, this study introduces a comparative evaluation of machine learning models in a unique urban context, providing new insights into how meteorological factors influence air quality predictions. The study contributes to the development of adaptive air quality prediction models, supporting sustainable environmental management planning in Tangerang City.

Keywords : *Air Quality, Meteorological Variables, Random Forest, CatBoost, Machine Learning*

1. INTRODUCTION

Kualitas udara telah menjadi isu yang semakin krusial dalam masyarakat modern karena berpengaruh besar terhadap kesehatan masyarakat maupun stabilitas lingkungan [1][2]. Kualitas udara yang buruk dapat menyebabkan beragam masalah kesehatan, seperti gangguan pada sistem pernapasan, penyakit kardiovaskular, bahkan meningkatkan risiko kematian dini [3]. Organisasi Kesehatan Dunia (WHO) melaporkan bahwa polusi udara berkontribusi terhadap sekitar 4,2 juta kematian dini setiap tahun, terutama akibat paparan partikel halus (PM2.5) dan polutan berbahaya lain yang mampu menembus sistem pernapasan dan kardiovaskular secara mendalam [4]. Selain itu, polusi udara menimbulkan risiko signifikan terhadap kesehatan masyarakat, ekosistem, dan iklim, khususnya di wilayah perkotaan [5][6]. Oleh sebab itu, model prediksi kualitas udara yang akurat menjadi sangat penting sebagai alat bantu bagi pembuat kebijakan dan pengelola lingkungan dalam menerapkan intervensi tepat waktu [7].

Di Indonesia, terutama di Kota Tangerang, upaya menjaga kualitas udara menghadapi kendala besar yang dipicu oleh laju industrialisasi, peningkatan jumlah penduduk, serta tingginya emisi dari kendaraan bermotor [8]. Sebagai kota penyangga Jabodetabek, Tangerang memiliki pola polusi udara yang fluktuatif dan kompleks [9]. Kondisi ini menuntut penggunaan model prediksi yang mampu memanfaatkan variabel lingkungan secara efektif, termasuk karakteristik meteorologi yang memengaruhi dinamika polutan [10].

Faktor meteorologi berperan penting dalam menentukan konsentrasi, dispersi, dan deposisi polutan [11][12]. Parameter seperti suhu, kelembapan, kecepatan angin, dan curah hujan memiliki

keterkaitan erat dengan dinamika polutan [13]. Kondisi ini menuntut penggunaan model prediksi yang mampu memanfaatkan variabel lingkungan secara efektif, termasuk karakteristik meteorologi yang memengaruhi dinamika polutan [14]. Integrasi data meteorologi dalam model prediksi terbukti meningkatkan akurasi prakiraan kualitas udara di berbagai wilayah [15].

Dalam beberapa tahun terakhir, pemanfaatan machine learning (ML) untuk pemodelan kualitas udara berkembang pesat [16][17]. Metode tradisional dinilai kurang mampu menangkap hubungan non-linear dalam data polusi [18], sedangkan algoritma ML seperti *Random Forest* (RF), *CatBoost*, dan model berbasis deep learning (misalnya LSTM) lebih efektif dalam mengatasi kompleksitas tersebut. *Random Forest* sebagai metode ensemble mampu mengolah dataset besar, memberikan analisis pentingnya variabel, serta mengurangi risiko overfitting [19]. Beberapa studi menunjukkan bahwa RF efektif mengidentifikasi variabel dominan dalam prediksi kualitas udara [12][20].

Penelitian sebelumnya oleh [17] menunjukkan bahwa *Random Forest* dalam pemodelan prediksi kualitas udara menunjukkan tingkat akurasi yang sangat tinggi serta performa yang lebih unggul dibandingkan dengan *artificial neural networks.* Sementara itu, studi lain [10] melakukan evaluasi terhadap beberapa algoritma *machine learning*, termasuk *Random Forest, decision tree,* dan *deep backpropagation neural network*, untuk menyelesaikan permasalahan prediksi yang sejenis. Temuan penelitian tersebut menegaskan bahwa *Random Forest* memberikan performa prediksi paling optimal di antara metode yang dibandingkan. Hasil ini sejalan dengan temuan dalam penelitian [21] , yang mengonfirmasi keunggulan Random Forest dalam menangani data dengan kompleksitas tinggi, terutama dalam masalah prediksi yang melibatkan banyak variabel non-linier seperti kualitas udara. Namun, *Random Forest* juga memiliki keterbatasan yaitu tingginya kebutuhan komputasi dan terkadang kinerjanya kalah dibanding metode boosting seperti *CatBoost* [17].

CatBoost, sebagai algoritma *gradient boosting*, unggul dalam menangani variabel kategorikal dan data dengan nilai hilang [17]. Penelitian di Visakhapatnam (India) menunjukkan bahwa *CatBoost* menghasilkan akurasi sangat tinggi ($R^2 = 0.9998$) dan RMSE rendah (0.76), melampaui RF pada dataset yang sama [17]. Mekanisme ordered boosting pada *CatBoost* membantu mengurangi overfitting dan meningkatkan generalisasi model [7].

Keunggulan masing-masing algoritma sangat dipengaruhi oleh karakteristik dataset yang dianalisis [7]. Efektivitas metode pembelajaran mesin pada akhirnya ditentukan oleh struktur data serta kondisi lingkungan yang diteliti [22]. Dengan mempertimbangkan karakteristik khas Kota Tangerang, seperti kepadatan penduduk yang tinggi, aktivitas industri yang padat, serta pengaruh faktor meteorologi karena kedekatannya dengan wilayah pesisir, diperlukan penelitian yang secara spesifik membandingkan performa algoritma *Random Forest* dan *CatBoost* guna mengkaji kemampuan prediksi kualitas udara [23]. Penelitian ini diperlukan secara mendesak mengingat keterbatasan model prediksi kualitas udara yang ada, sehingga pengembangan model dengan akurasi yang lebih tinggi menjadi krusial sebagai landasan ilmiah dalam perumusan kebijakan lingkungan berkelanjutan di wilayah Tangerang.

2. METHOD

Kota Tangerang merupakan salah satu kawasan penyangga Jakarta yang mengalami urbanisasi dan industrialisasi pesat, disertai peningkatan jumlah penduduk yang signifikan dalam dua dekade terakhir [24]. Kondisi ini memicu tingginya emisi kendaraan bermotor dan aktivitas industri sehingga berdampak pada penurunan kualitas udara di wilayah tersebut. Kompleksitas permasalahan polusi udara semakin diperparah oleh pengaruh faktor meteorologi yang menentukan penyebaran, pengendapan, dan konsentrasi polutan [24]. Pemahaman mengenai keterkaitan meteorologi dan kualitas udara menjadi penting karena dapat membantu pengembangan model prediksi yang lebih akurat.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 3, June 2026, Page. 2154-2164
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5412

Penelitian ini disusun melalui beberapa tahapan utama, meliputi proses pengumpulan data, praproses data, pembagian dataset, serta penerapan model prediksi berbasis *Machine Learning*. Pada tahap implementasi, digunakan metode *Random Forest* dan *CatBoost* karena keduanya terbukti efektif dalam mengolah data non-linear dan mampu memberikan kinerja prediksi yang tinggi pada berbagai studi terkait kualitas udara [17][25]. Berikut alur pelaksanaan penelitian ini disajikan pada Figure 1.

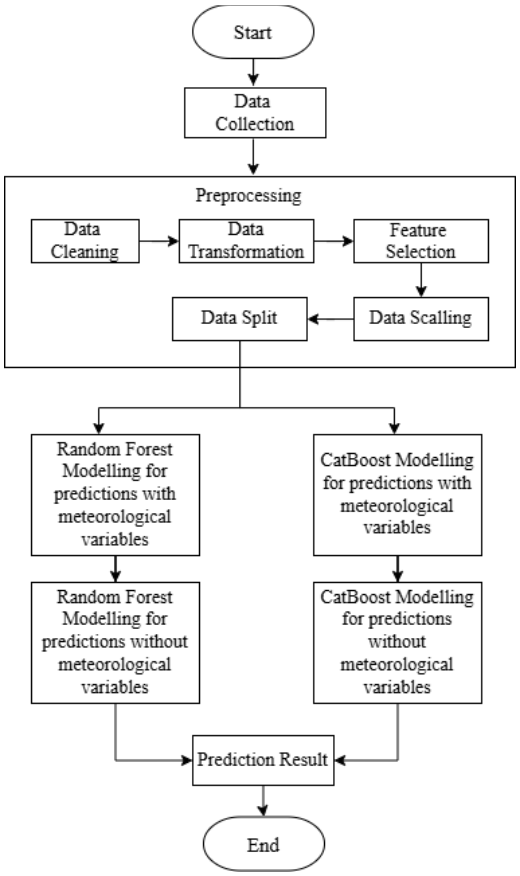

Figure 1. Desain Metode Penelitian

2.1. Data Collection

Data polusi udara Kota Tangerang yang digunakan dalam penelitian ini diperoleh dari situs resmi Kementerian Lingkungan Hidup dan Kehutanan (https://ispu.menlhk.go.id). Dataset yang digunakan dipantau setiap jam selama 24 jam penuh mulai 02 Januari hingga 29 Februari 2024 dengan variabel utama meliputi PM_{25}, PM_{10}, SO_2, NO_2, O_3, CO, dan HC. Sementara itu, data meteorologi dikumpulkan melalui situs resmi (www.visualcrossing.com), yang meliputi variabel suhu, arah angin, kecepatan angin, serta kelembapan udara. Visual Crossing Weather merupakan perusahaan penyedia data meteorologi dan alat analisis terkemuka yang ditujukan bagi ilmuwan data, analis bisnis, profesional, serta kalangan akademisi. Sejak didirikan pada tahun 2003, perusahaan ini berkomitmen untuk memberdayakan pengguna dan analis data dalam pengambilan keputusan yang lebih tepat melalui penyediaan data berkualitas tinggi yang mudah diakses.

2.2. Preprocessing

Tahap *preprocessing* merupakan prosedur penting yang bertujuan untuk mengubah data mentah yang awalnya kurang optimal menjadi bentuk yang lebih terstruktur dan berkualitas, sehingga dapat dimanfaatkan secara efektif sebagai masukan dalam pengujian metode yang diusulkan serta

meningkatkan tingkat akurasi prediksi. Berikut ini disajikan tahapan-tahapan yang dilakukan dalam proses pra-pemrosesan data.

2.2.1. Data Cleaning

Tahapan pembersihan data merupakan bagian krusial dalam proses persiapan dataset. Pada fase awal preprocessing, langkah ini mencakup penghapusan data yang tidak relevan atau tidak konsisten dengan format yang ditetapkan dalam dataset. Tujuan utama dari proses ini adalah meningkatkan keandalan serta akurasi model prediksi yang dihasilkan.

2.2.2. Data Transformation

Pada tahap transformasi, struktur seluruh variabel data diseragamkan, misalnya dengan menyesuaikan format data agar sesuai dengan operasi analisis yang akan dilakukan. Pada langkah ini, variabel temporal seperti tanggal dan waktu dikonversi ke dalam format Datetime untuk menjamin konsistensi serta kompatibilitas dengan analisis selanjutnya.

2.2.3. Feature Selection

Mengingat tujuan utama penelitian ini adalah mengevaluasi pengaruh faktor meteorologi terhadap kualitas udara, analisis awal dilakukan dengan memprediksi kualitas udara hanya berdasarkan variabel polutan tanpa melibatkan masukan meteorologi. Sebaliknya, analisis lanjutan mengintegrasikan variabel meteorologi dan polutan untuk menghasilkan evaluasi yang lebih komprehensif. Dengan demikian, pada tahap awal, variabel polutan yang dipilih sebagai fitur input meliputi PM_{25}, PM_{10}, SO_2, NO_2, O_3, CO, dan HC.

2.2.4. Data Scaling

Sebelum memulai proses pelatihan model yang diusulkan, dataset terlebih dahulu melalui tahap normalisasi data. Tujuan utama dari prosedur ini adalah untuk menyelaraskan skala data ke dalam rentang standar, yaitu antara 0 dan 1. Transformasi tersebut dilakukan dengan menggunakan persamaan berikut:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

2.2.5. Data Splitting

Setelah tahap pra-pemrosesan data selesai dilakukan, dataset dibagi menjadi dua subset utama, yaitu data pelatihan dan data pengujian. Data pelatihan berperan dalam proses pembentukan model yang diusulkan, sedangkan data pengujian digunakan untuk mengevaluasi serta memvalidasi kemampuan prediktif dari model yang dihasilkan. Proses pembagian dataset dilakukan dengan menggunakan rasio 80:20, di mana 80% data digunakan untuk pelatihan model, sementara 20% sisanya dialokasikan untuk pengujian guna memastikan akurasi dan keandalan hasil prediksi.

2.3. Implementation Method

Pada tahap implementasi metode, dua algoritma *Machine Learning*, yaitu *Random Forest* dan *CatBoost*, diterapkan untuk memprediksi tingkat kualitas udara di Kota Tangerang. Kedua model ini dirancang dan dievaluasi melalui dua skenario pengujian guna menganalisis pengaruh variabel meteorologi terhadap performa prediksi, yaitu prediksi kualitas udara dengan variabel meteorologi, dan prediksi kualitas udara tanpa variabel meteorologi.

Model prediksi dengan variabel meteorologi menggunakan suhu, arah angin, kecepatan angin, kelembapan, PM25, PM10, SO_2, NO_2, O_3, CO, dan HC sebagai fitur input. Sedangkan model prediksi tanpa variabel meteorologi hanya menggunakan PM25, PM10, SO_2, NO_2, O_3, CO, dan HC sebagai input.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 3, June 2026, Page. 2154-2164
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5412

Dengan menerapkan dua skenario pemodelan yang serupa pada algoritma *Random Forest* dan *CatBoost*, penelitian ini bertujuan untuk memperoleh pemahaman yang mendalam mengenai pengaruh variabel meteorologi terhadap kualitas udara, serta membandingkan kinerja kedua algoritma berdasarkan tingkat akurasi prediksi yang dihasilkan. Melalui perbandingan tersebut, diharapkan dapat diketahui algoritma yang memiliki performa paling akurat dalam memprediksi kualitas udara di Kota Tangerang.

2.4. Evaluation Method

Kinerja prediksi dari kedua model dievaluasi menggunakan sejumlah metrik statistik kuantitatif yang berfungsi untuk mengukur perbedaan antara nilai hasil prediksi dengan nilai observasi aktual. Beberapa metrik yang digunakan meliputi:

- *Mean Absolute Error* (MAE): menggambarkan rata-rata besarnya kesalahan absolut antara nilai yang diprediksi dengan nilai sebenarnya.
- *Mean Squared Error* (MSE): menghitung nilai rata-rata dari kuadrat selisih antara hasil prediksi dan data observasi, sehingga memberikan bobot penalti yang lebih besar terhadap kesalahan dengan nilai ekstrem
- *Root Mean Squared Error* (RMSE): merupakan akar kuadrat dari nilai MSE yang menggambarkan besarnya kesalahan dalam satuan yang sama dengan variabel target.
- *Koefisien Determinasi* (R^2): menunjukkan proporsi variasi data yang dapat dijelaskan oleh model, sehingga menjadi indikator kemampuan model dalam merepresentasikan hubungan antara variabel input dan output.

Persamaan matematis dari masing-masing metrik ditunjukkan sebagai berikut [26][27]:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|x_{observed} - x_{predicted}\right| \quad (2)$$

$$\mathrm{MSE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(x_{observed} - x_{predicted}\right)^2 \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{observed} - x_{predicted})^2} \quad (4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_{observed} - x_{predicted})^2}{\sum_{i=1}^{n}(x_{observed} - \bar{x})^2} \quad (5)$$

3. RESULT

Penelitian ini menggunakan bahasa pemrograman Python untuk membangun model *Random Forest* dan *CatBoost* untuk menganalisis kemampuan prediksi kualitas udara di Kota Tangerang, baik dengan melibatkan variabel meteorologi maupun tanpa variabel meteorologi. Data variabel polusi udara (PM2.5, PM10, SO2, NO2, O3, CO, HC) dan variabel meteorologi seperti suhu, arah dan kecepatan angin, serta kelembapan dipantau setiap jam dari 2 Januari hingga 29 Februari 2024. Hasil perbandingan kinerja kedua metode tersebut dalam memprediksi kualitas udara di Kota Tangerang disajikan pada bagian berikut.

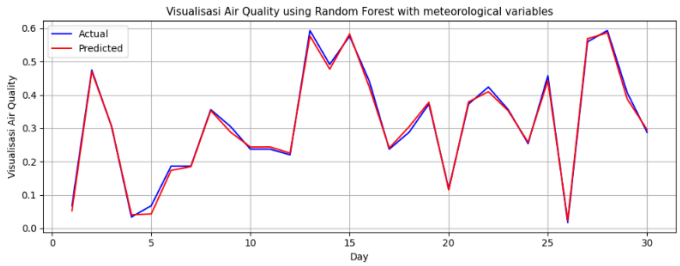

Figure 2. Air Quality Prediction using *Random Forest* with Meteorological Variables

Vol. 7, No. 3, June 2026, Page. 2154-2164
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5412

Berdasarkan Figure 2, hasil prediksi kualitas udara menggunakan model *Random Forest* dengan memasukkan variabel meteorologi menunjukkan tingkat kesesuaian yang sangat kuat antara data aktual dan hasil prediksi. Pola fluktuasi kedua garis tampak hampir identik, menandakan bahwa model mampu merepresentasikan perubahan kualitas udara yang dipengaruhi oleh faktor meteorologi seperti suhu, kelembapan, dan kecepatan angin. Temuan ini mengindikasikan bahwa penambahan variabel meteorologi mampu meningkatkan akurasi model, karena kondisi cuaca memiliki pengaruh langsung terhadap penyebaran dan konsentrasi partikel pencemar di udara.

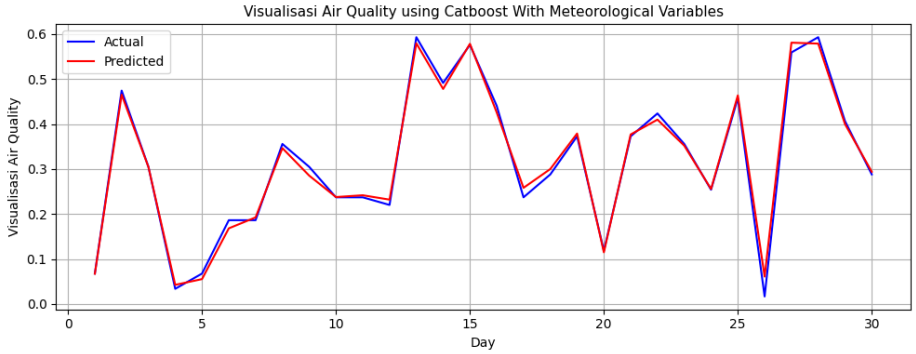

Figure 3. Air Quality Prediction using *CatBoost* with Meteorological Variables

Sebagaimana terlihat pada Figure 3, model *CatBoost* dengan variabel meteorologi juga menunjukkan hubungan yang kuat antara nilai aktual dan nilai prediksi kualitas udara. Garis merah (prediksi) dan garis biru (aktual) hampir saling berimpit, yang menunjukkan bahwa *CatBoost* mampu mempelajari hubungan kompleks dan nonlinier antar variabel meteorologi dengan baik. Dibandingkan dengan *Random Forest*, model *CatBoost* menghasilkan pola prediksi yang lebih halus dengan deviasi yang lebih kecil, sehingga menandakan keunggulannya dalam menangani data campuran (numerik dan kategorik) ketika variabel meteorologi diikutsertakan.

Berdasarkan Figure 4, terlihat bahwa ketika variabel meteorologi tidak dilibatkan, model *Random Forest* memperlihatkan perbedaan yang lebih jelas antara nilai aktual dan nilai prediksi. Garis hasil prediksi tampak menyimpang, khususnya pada beberapa titik dengan nilai ekstrem (puncak dan lembah).

Kondisi ini menunjukkan bahwa tanpa memasukkan variabel meteorologi, kemampuan model untuk menangkap pengaruh lingkungan terhadap kualitas udara menjadi terbatas. Akibatnya, tingkat akurasi prediksi menurun karena model hanya mengandalkan variabel polutan tanpa memperhitungkan faktor cuaca yang turut memengaruhi distribusi polusi.

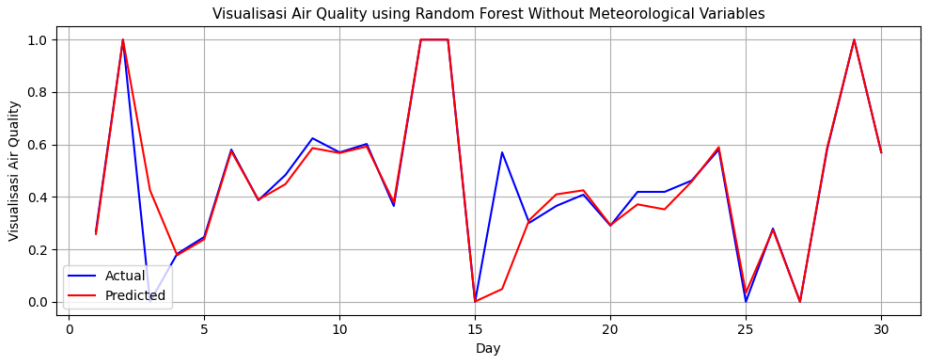

Figure 4. Air Quality Prediction using *Random Forest* without Meteorological Variables

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 3, June 2026, Page. 2154-2164
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5412

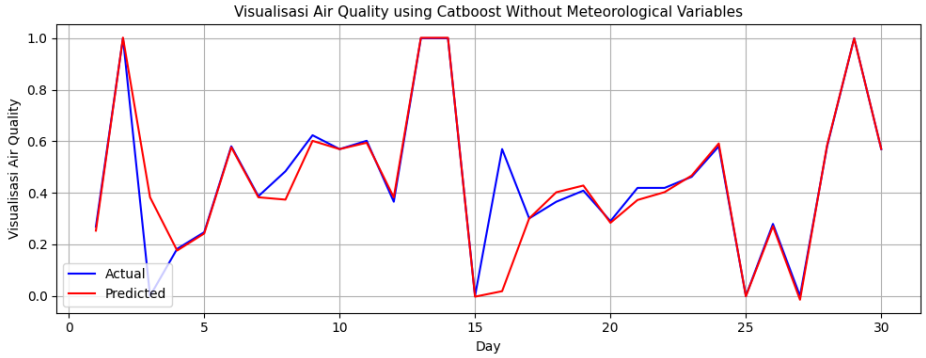

Figure 5. Air Quality Prediction using *CatBoost* without Meteorological Variables

Sebagaimana ditunjukkan pada Figure 5, model *CatBoost* tanpa variabel meteorologi masih mampu mempertahankan pola tren yang relatif konsisten antara nilai aktual dan nilai prediksi. Namun, terdapat beberapa penyimpangan pada titik-titik ekstrem perubahan kualitas udara. Hal ini menunjukkan bahwa meskipun *CatBoost* memiliki kemampuan generalisasi yang baik meski dengan fitur terbatas, keberadaan variabel meteorologi tetap berperan penting dalam meningkatkan akurasi dan kestabilan hasil prediksi kualitas udara.

Hasil pengujian model prediksi kualitas udara, baik dengan maupun tanpa variabel meteorologi, disajikan pada Table 1. Evaluasi dilakukan terhadap dua algoritma pembelajaran mesin, yaitu *Random Forest* dan *CatBoost*, dengan memperhatikan nilai metrik evaluasi MAE, MSE, RMSE, serta R^2. Metrik-metrik tersebut digunakan untuk menilai tingkat akurasi model dalam memprediksi kualitas udara berdasarkan perbandingan antara data aktual dan hasil prediksi.

Berdasarkan hasil evaluasi yang ditunjukkan pada Tabel 1, dapat dilihat bahwa baik *Random Forest* maupun *CatBoost* menunjukkan performa yang sangat baik pada skenario prediksi dengan menggunakan variabel meteorologi. Nilai *mean absolute error* (MAE), *mean squared error* (MSE), dan *root mean squared error* (RMSE) yang rendah menunjukkan bahwa selisih antara nilai aktual dan nilai prediksi relatif kecil, sedangkan nilai koefisien determinasi (R^2) yang mendekati 1 menandakan tingkat akurasi yang tinggi.

Table 1. Model Performance Evaluation

Evaluation parameters	Prediction with meteorological variables		Prediction without meteorological variables	
	Random Forest	*CatBoost*	*Random Forest*	*CatBoost*
MAE	0.0099	0.0135	0.0237	0.0232
MSE	0.000309	0.000419	0.003132	0.003342
RMSE	0.017583	0.020458	0.055964	0.057808
R^2	0.9931	0.9907	0.9519	0.9487

Pada kondisi dengan variabel meteorologi, model *Random Forest* memiliki nilai MAE sebesar 0.0099 dan R^2 sebesar 0.9931, sedikit lebih baik dibandingkan *CatBoost* dengan MAE 0.0135 dan R^2 0.9907. Hal ini menunjukkan bahwa *Random Forest* mampu mempelajari hubungan antara variabel input dan kualitas udara dengan sangat baik ketika informasi meteorologi tersedia.

Sebaliknya, pada kondisi tanpa variabel meteorologi, kinerja kedua model mengalami penurunan. Nilai MAE dan RMSE meningkat, sementara nilai R^2 menurun menjadi 0.9519 untuk *Random Forest* dan 0.9487 untuk *CatBoost*.

Vol. 7, No. 3, June 2026, Page. 2154-2164
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5412

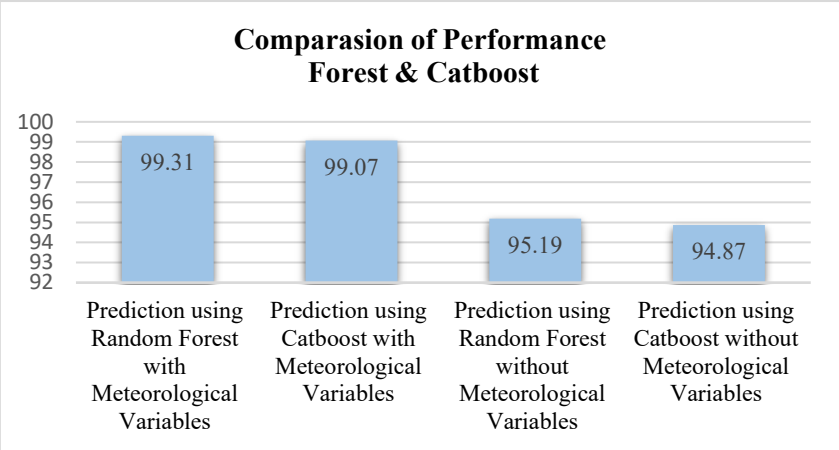

Figure 6. Comparasion of Performance *Random Forest* & *CatBoost*

Analisis perbandingan kinerja kedua metode yang digunakan ditampilkan pada Figure 6. Hasil perbandingan menunjukkan bahwa prediksi dengan variabel meteorologi memberikan hasil yang lebih baik dibandingkan prediksi tanpa variabel meteorologi. Akurasi yang dicapai dalam pengujian dengan variabel meteorologi menggunakan *Random Forest* adalah 99,31%, sedangkan dengan *CatBoost* mencapai 99,07%. Sementara itu, ketika variabel meteorologi tidak digunakan, akurasi yang tercatat untuk *Random Forest* adalah 95,19%, dan untuk *CatBoost* adalah 94,87%.

Faktor meteorologi ternyata memiliki pengaruh yang signifikan terhadap prediksi kualitas udara. Berdasarkan hasil perbandingan kinerja yang ditunjukkan pada Figure 6, terbukti bahwa kinerja *Random Forest* sedikit lebih unggul dibandingkan *CatBoost* ketika melibatkan faktor cuaca dalam memprediksi kualitas udara. Kedua model menunjukkan penurunan performa ketika variabel meteorologi dihilangkan, tetapi *Random Forest* tetap menunjukkan sedikit keunggulan dibandingkan *CatBoost*.

Analisis terhadap *Random Forest* dan *CatBoost* dalam memprediksi kualitas udara menunjukkan bahwa kedua metode ini memiliki kemampuan prediksi polusi udara yang kuat, baik di Kota Tangerang maupun kota lainnya. Metode *Random Forest* cenderung memberikan performa yang sedikit lebih stabil dan akurat. Sementara itu, *CatBoost* tetap menunjukkan kemampuan generalisasi yang baik, terutama dalam menghadapi data dengan fitur yang terbatas. Keterbatasan jumlah data yang digunakan dalam proses pemodelan prediksi kualitas udara memengaruhi tingkat akurasi prediksi.

4. DISCUSSIONS

Hasil penelitian ini menunjukkan bahwa model *Random Forest* yang melibatkan variabel meteorologi memiliki akurasi yang lebih tinggi dibandingkan dengan model *CatBoost*, baik pada kondisi dengan maupun tanpa variabel meteorologi. Berdasarkan Tabel 1 dan Figure 6, model *Random Forest* yang menggunakan variabel meteorologi memperoleh akurasi 99,31%, sedikit lebih tinggi dibandingkan dengan *CatBoost* yang mencatatkan akurasi 99,07%. Namun, ketika variabel meteorologi tidak dimasukkan, performa kedua model mengalami penurunan, dengan *Random Forest* memperoleh akurasi 95,19% dan *CatBoost* mencapai 94,87%.

Penurunan kinerja yang signifikan pada kedua model ketika variabel meteorologi dihilangkan mengindikasikan bahwa faktor meteorologi seperti suhu, kelembapan, dan kecepatan angin memainkan peranan penting dalam memprediksi kualitas udara. Nilai *Mean Absolute Error* (MAE) dan *Root Mean Square Error* (RMSE) yang lebih tinggi pada model tanpa variabel meteorologi menunjukkan bahwa tanpa informasi meteorologi, model kesulitan dalam menghasilkan prediksi yang akurat. Hal ini sejalan dengan temuan dari penelitian [28] yang mengungkapkan bahwa variabel meteorologi memiliki dampak besar terhadap kualitas udara, terutama dalam menangkap fluktuasi musiman dan perubahan cuaca.

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5412

Selain itu, penelitian [29] juga menunjukkan bahwa kondisi meteorologi, seperti suhu tinggi, kecepatan angin rendah, dan kelembapan rendah, cenderung meningkatkan konsentrasi polutan udara. Penurunan kinerja yang tercatat pada model yang tidak mempertimbangkan variabel meteorologi memberikan bukti lebih lanjut bahwa informasi meteorologi sangat penting untuk meningkatkan akurasi prediksi kualitas udara. Kedua model menunjukkan penurunan yang cukup tajam pada MAE dan RMSE, serta penurunan nilai R^2, yang mengindikasikan berkurangnya kemampuan model dalam menjelaskan variabilitas kualitas udara tanpa mempertimbangkan faktor cuaca.

Secara keseluruhan, hasil penelitian ini menunjukkan bahwa variabel meteorologi berpengaruh signifikan terhadap akurasi prediksi kualitas udara, dengan *Random Forest* sedikit lebih unggul dibandingkan *CatBoost*. Namun, meskipun kedua model memberikan hasil yang baik, penelitian ini juga memiliki beberapa keterbatasan yang perlu diperhatikan. Pertama, data yang digunakan hanya mencakup periode observasi selama dua bulan, yang mungkin tidak cukup representatif untuk menangkap variasi kualitas udara dalam jangka waktu yang lebih panjang. Kedua, faktor eksternal lainnya, seperti polusi lintas batas atau kegiatan antropogenik (misalnya, emisi dari kendaraan atau aktivitas industri), yang dapat mempengaruhi kualitas udara, tidak dipertimbangkan dalam model ini. Oleh karena itu, penelitian lanjutan dengan cakupan data yang lebih besar dan mempertimbangkan faktor eksternal lainnya sangat diperlukan untuk memperkuat temuan yang ada dan meningkatkan generalisasi model.

5. CONCLUSION

Hasil penelitian dapat disimpulkan bahwa prediksi kualitas udara di Kota Tangerang, Indonesia, yang dianalisis menggunakan data polutan udara dan meteorologi per jam yang dipantau dari 2 Januari hingga 29 Februari 2024, memberikan wawasan yang signifikan terkait pengaruh variabel meteorologi terhadap akurasi model prediksi. Penelitian ini menerapkan metode Random Forest dan CatBoost untuk memodelkan kualitas udara di Kota Tangerang, dengan fokus pada dampak variabel meteorologi. Kinerja prediktif dievaluasi menggunakan MAE, MSE, RMSE, R-squared, dan akurasi. Hasil penelitian menunjukkan bahwa prediksi dengan variabel meteorologi memberikan hasil yang lebih baik, dengan Random Forest mencapai akurasi 99,31% dan CatBoost 99,07%. Sebaliknya, ketika variabel meteorologi tidak digunakan, kedua model mengalami penurunan kinerja, dengan Random Forest memperoleh akurasi 95,19% dan CatBoost 94,87%.

Nilai kesalahan terbesar untuk kedua model tanpa variabel meteorologi adalah 0,0237 MAE, 0,003132 MSE, dan 0,055964 RMSE untuk Random Forest, dan 0,0232 MAE, 0,003342 MSE, dan 0,057808 RMSE untuk CatBoost. Sebaliknya, nilai kesalahan terkecil diperoleh ketika data meteorologi dimasukkan, dengan 0,0099 MAE, 0,000309 MSE, dan 0,017583 RMSE untuk Random Forest, dan 0,0135 MAE, 0,000419 MSE, dan 0,020458 RMSE untuk CatBoost.

Temuan ini mengonfirmasi bahwa variabel meteorologi, seperti suhu, kelembapan, dan kecepatan angin, berpengaruh signifikan terhadap akurasi prediksi kualitas udara. Selain kontribusinya pada bidang lingkungan, penelitian ini juga memberikan kontribusi penting terhadap pengembangan ilmu komputer dan machine learning, khususnya dalam pemahaman kinerja algoritma ensemble dan boosting pada data lingkungan yang bersifat nonlinier dan dipengaruhi oleh faktor temporal. Studi ini menunjukkan bagaimana integrasi fitur kontekstual, seperti variabel meteorologi, dapat meningkatkan performa model prediktif serta memberikan dasar empiris bagi pemilihan algoritma yang tepat dalam aplikasi prediksi berbasis data nyata.

Berdasarkan hasil tersebut, disarankan agar penelitian selanjutnya mengeksplorasi model prediksi kualitas udara di daerah perkotaan dengan kepadatan populasi dan aktivitas industri yang lebih tinggi, serta mengintegrasikan pendekatan pembelajaran mesin yang lebih lanjut (misalnya model temporal

atau hybrid) guna meningkatkan kemampuan generalisasi dan keandalan model prediksi di berbagai konteks perkotaan.

REFERENCES

[1] A. H. Khoshakhlagh, M. Mohammadzadeh, A. Gruszecka-Kosowska, and E. Oikonomou, "Burden of cardiovascular disease attributed to air pollution: a systematic review," *Global. Health*, vol. 20, no. 1, pp. 1–24, 2024, doi: 10.1186/s12992-024-01040-0.

[2] N. S. Gupta, Y. Mohta, K. Heda, R. Armaan, B. Valarmathi, and G. Arulkumaran, "Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis," *J. Environ. Public Health*, vol. 2023, pp. 1–26, 2023, doi: 10.1155/2023/4916267.

[3] Z. Zhang, S. Zhang, C. Chen, and J. Yuan, "A systematic survey of air quality prediction based on deep learning," *Alexandria Eng. J.*, vol. 93, pp. 128–141, 2024, doi: 10.1016/j.aej.2024.03.031.

[4] World Health Organization (WHO), "Polusi udara ambien (luar ruangan)," 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

[5] G. Ravindiran *et al.*, "Impact of air pollutants on climate change and prediction of air quality index using machine learning models," *Environ. Res.*, vol. 239, p. 117354, 2023, doi: https://doi.org/10.1016/j.envres.2023.117354.

[6] Afifa, K. Arshad, N. Hussain, M. H. Ashraf, and M. Z. Saleem, "Air pollution and climate change as grand challenges to sustainability," *Sci. Total Environ.*, vol. 928, p. 172370, 2024, doi: 10.1016/j.scitotenv.2024.172370.

[7] Y. Özüpak, F. Alpsalaz, and E. Aslan, "Air Quality Forecasting Using Machine Learning: Comparative Analysis and Ensemble Strategies for Enhanced Prediction," *Water. Air. Soil Pollut.*, vol. 236, no. 7, pp. 1–17, 2025, doi: 10.1007/s11270-025-08122-8.

[8] H. Chen, G. Deng, and Y. Liu, "Monitoring the Influence of Industrialization and Urbanization on Spatiotemporal Variations of AQI and PM2.5 in Three Provinces, China," *Atmosphere (Basel).*, vol. 13, no. 9, 2022, doi: 10.3390/atmos13091377.

[9] S. G. Bontong, D. A. Permadi, and P. Benjamin, "Determination of Air Quality Protection and Management Strategic Area : Case Study of Tangerang City," *J. Presipitasi Media Komun. dan Pengemb. Tek. Lingkung.*, vol. 21, no. 3, pp. 852–868, 2024, doi: 10.14710/presipitasi.v21i3.852-868.

[10] Y. Liu, P. Wang, Y. Li, L. Wen, and X. Deng, "Air quality prediction models based on meteorological factors and real-time data of industrial waste gas," *Sci. Rep.*, vol. 12, no. 1, pp. 1–15, 2022, doi: 10.1038/s41598-022-13579-2.

[11] C. Girotti *et al.*, "Air pollution Dynamics: The role of meteorological factors in PM10 concentration patterns across urban areas," *City Environ. Interact.*, vol. 25, 2024, doi: 10.1016/j.cacint.2024.100184.

[12] R. Liu *et al.*, "Air Quality—Meteorology Correlation Modeling Using Random Forest and Neural Network," *Sustain.*, vol. 15, no. 5, 2023, doi: 10.3390/su15054531.

[13] X. Que, "Analysis of the Influence of Meteorological Factors on Air Pollutants in Nanning from 2018 to 2020," *Highlights Sci. Eng. Technol.*, vol. 9, pp. 148–155, 2022, doi: 10.54097/hset.v9i.1734.

[14] X. Tian *et al.*, "Research on Air Quality in Response to Meteorological Factors Based on the Informer Model," *Sustainability*, vol. 16, no. 16, 2024, doi: 10.3390/su16166794.

[15] M. T. Udristioiu, Y. EL Mghouchi, and H. Yildizhan, "Prediction, modelling, and forecasting of PM and AQI using hybrid machine learning," *J. Clean. Prod.*, vol. 421, p. 138496, 2023, doi: https://doi.org/10.1016/j.jclepro.2023.138496.

[16] L. Gao, C. Cai, and X.-M. Hu, "Air Quality Prediction Using Machine Learning," in *Machine Learning in Chemical Safety and Health*, 2022, pp. 267–288. doi: https://doi.org/10.1002/9781119817512.ch11.

[17] G. Ravindiran, G. Hayder, K. Kanagarathinam, A. Alagumalai, and C. Sonne, "Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam," *Chemosphere*, vol. 338, no. May, 2023, doi:

Vol. 7, No. 3, June 2026, Page. 2154-2164
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5412

10.1016/j.chemosphere.2023.139518.

[18] R. Fang, S. Collingwood, Y. Zhang, J. B. Stanford, C. Porucznik, and D. Sleeth, "Optimizing Air Quality Monitoring: Comparative Analysis of Linear Regression and Machine Learning in Low-Cost Sensor Calibration," *Aerosol Air Qual. Res.*, vol. 25, no. 1, pp. 1–17, 2025, doi: 10.1007/s44408-025-00009-x.

[19] M. Mihirani, L. Yasakethu, and S. Balasooriya, "Machine Learning-based Air Pollution Prediction Model," *2023 IEEE IAS Glob. Conf. Emerg. Technol. GlobConET 2023*, no. 2, pp. 1–6, 2023, doi: 10.1109/GlobConET56651.2023.10150203.

[20] S. Li, X. Deng, and B. Tang, "Using Machine Learning Methods for Prediction of Air Quality in Wuling Mountain Area in China," in *2021 International Conference on Electronic Information Technology and Smart Agriculture (ICEITSA)*, 2021, pp. 426–430. doi: 10.1109/ICEITSA54226.2021.00087.

[21] R. E. Saputro and G. Karyono, "Comparative Analysis of Decision Tree , Random Forest , Svm , and Neural Network Models for Predicting Earthquake Magnitude," vol. 6, no. 2, pp. 755–774, 2025.

[22] 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.

[23] L. Mampitiya *et al.*, "Machine Learning Techniques to Predict the Air Quality Using Meteorological Data in Two Urban Areas in Sri Lanka," *Environ. - MDPI*, vol. 10, no. 8, pp. 1–18, 2023, doi: 10.3390/environments10080141.

[24] N. Cholianawati *et al.*, "Diurnal and Daily Variations of PM2.5 and its Multiple-Wavelet Coherence with Meteorological Variables in Indonesia," *Aerosol Air Qual. Res.*, vol. 24, no. 3, pp. 1–18, 2024, doi: 10.4209/aaqr.230158.

[25] M. Madhuri, G. H. Samyama Gunjal, and S. Kamalapurkar, "Air pollution prediction using machine learning supervised learning approach," *Int. J. Sci. Technol. Res.*, vol. 9, no. 4, pp. 118–123, 2020.

[26] C. M. Ellis, *The Orange Book of Machine Learning: The essentials of making predictions using supervised regression and classification for tabular data*. 2024.

[27] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, pp. 1–24, 2021, doi: 10.7717/PEERJ-CS.623.

[28] T. Wang *et al.*, "Prediction of the Impact of Meteorological Conditions on Air Quality during the 2022 Beijing Winter Olympics," 2022. doi: 10.3390/su14084574.

[29] R. Janarthanan, P. Partheeban, K. Somasundaram, and P. Navin Elamparithi, "A deep learning approach for prediction of air quality index in a metropolitan city," *Sustain. Cities Soc.*, vol. 67, 2021, doi: 10.1016/j.scs.2021.102720.