

ANALISIS PREDIKSI KUALITAS UDARA DENGAN METODE RANDOM FOREST BERDASARKAN DATA CUACA

Ryan Dwi Aprilyanto¹⁾, Riansyah Gustian²⁾, Muhammad Hendra Hernawan³⁾,
Sumanto⁴⁾, Ade Surya Budiman⁵⁾

^{1,2,3,4,5}Ilmu Komputer, Fakultas Teknik Informatika, Universitas bina Sarana Informatika

Jl. Kramat Raya No.98, Kwitang, Senen, Jakarta Pusat, 10450, Indonesia

Corresponding Author: ¹ 15220702@bsi.ac.id

Article Info

Article history:

Received: Juny 29, 2025

Revised: July, 20, 2025

Accepted: sept, 24, 2025

Published: Okt, 30, 2025

Keywords:

Prediksi cuaca

Random Forest

Machine Learning

ABSTRACT

Penelitian ini bertujuan untuk memprediksi suhu rata-rata harian (TAVG) di Kota Jakarta menggunakan algoritma *Machine Learning Random Forest*. Data yang digunakan berasal dari Stasiun Meteorologi Kemayoran dengan periode pengamatan 1 Januari 2024 hingga 30 Mei 2025 dan mencakup tujuh atribut cuaca seperti suhu, kelembapan, curah hujan, dan kecepatan angin. Proses penelitian melibatkan tahapan pengumpulan data, persiapan, pemodelan, dan evaluasi menggunakan metrik regresi seperti *Mean Squared Error* (MSE), *Root Mean Squared Error* (RMSE), *Mean Absolute Error* (MAE), *Mean Absolute Percentage Error* (MAPE), dan koefisien determinasi (R^2). Hasil evaluasi menunjukkan performa model yang cukup baik, dengan nilai MSE sebesar 0.398, RMSE 0.631, MAE 0.488, MAPE 1.7%, dan R^2 sebesar 0.653. Visualisasi distribusi data menunjukkan kecocokan antara prediksi dan data aktual, yang menandakan model mampu menangkap pola cuaca secara akurat. Dengan demikian, model *Random Forest* dinilai efektif dan andal untuk digunakan dalam prediksi cuaca serta memiliki potensi untuk dikembangkan lebih lanjut dalam mendukung pengambilan keputusan di berbagai sektor.



This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY SA 4.0)

1. INTRODUCTION

Cuaca merupakan peristiwa alam yang terjadi di atmosfer Bumi dan berkaitan dengan kondisi udara seperti suhu, kelembapan, waktu, serta lokasi tertentu[1]. Perubahan kondisi cuaca dapat berdampak besar pada berbagai sektor kehidupan manusia, seperti pertanian, transportasi, pariwisata, hingga penanggulangan bencana. Oleh karena itu, kemampuan untuk memprediksi kondisi cuaca secara akurat dan tepat waktu sangatlah penting dalam mendukung pengambilan keputusan dalam aktivitas sehari-hari.

Di tengah kemajuan era digital, teknologi dan metode komputasi telah memungkinkan pengembangan model prediksi cuaca yang lebih canggih dan presisi. Salah satu pendekatan yang populer dalam memprediksi cuaca adalah dengan menggunakan metode *Machine Learning*. *Machine Learning* atau Pembelajaran Mesin merupakan salah satu pendekatan dalam kecerdasan buatan AI

(*Artificial Intelligence*) yang dirancang untuk meniru kemampuan manusia dalam menyelesaikan berbagai tugas, sehingga dapat menggantikan peran manusia dalam proses pemecahan masalah. [2]

Dalam konteks prediksi cuaca, metode *Machine Learning* dimanfaatkan untuk memahami pola serta hubungan yang kompleks antara berbagai variabel cuaca, seperti suhu, kelembapan, waktu, dan lokasi. Teknik ini menggunakan algoritma dan pendekatan statistik untuk membangun model prediktif yang mampu memperkirakan kondisi cuaca di masa mendatang berdasarkan data historis.[3]

Salah satu algoritma *Machine Learning* yang sering digunakan dalam prediksi cuaca adalah *Random Forest*. *Random Forest* adalah sebuah algoritma ensemble yang terdiri dari kumpulan pohon keputusan (decision trees)[4]. Masih di cari, Kelebihan dari Random Forest adalah kemampuannya dalam mengatasi masalah prediksi cuaca yang kompleks, termasuk non-linearitas, interaksi antar variabel, dan keberadaan outliers dalam data.[5]

Penelitian ini bertujuan untuk mengeksplorasi dan menganalisis pemanfaatan algoritma Machine Learning, khususnya *Random Forest* dalam memprediksi kondisi cuaca secara akurat dan efisien. Dengan memanfaatkan data historis cuaca dan teknik analisis statistik, jurnal ini bertujuan untuk menunjukkan bagaimana *Random Forest* dapat mengidentifikasi pola-pola kompleks antar variabel cuaca seperti suhu, kelembaban, waktu, dan lokasi, sehingga dapat memberikan hasil prediksi yang bermanfaat dalam mendukung pengambilan keputusan di berbagai sektor kehidupan.

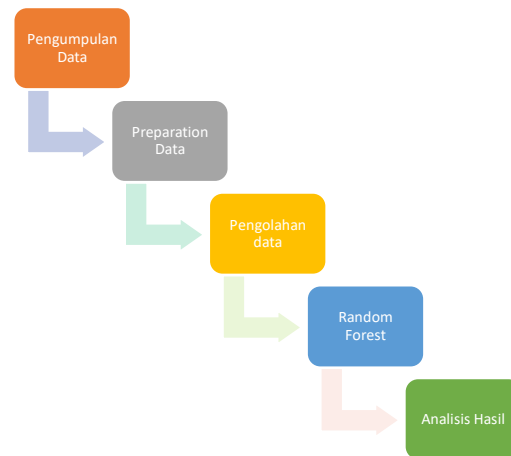
Pada akhir penelitian, akan dilakukan analisis hasil evaluasi kinerja model *Random Forest* dalam memprediksi cuaca di Kota Jakarta. Evaluasi dilakukan dengan menggunakan metrik regresi seperti Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), dan koefisien determinasi (R^2). Penggunaan metrik-metrik ini bertujuan untuk memberikan gambaran yang komprehensif mengenai tingkat akurasi, kesalahan, dan kemampuan model dalam menjelaskan variasi data, sehingga dapat diketahui sejauh mana model dapat diandalkan dalam menghasilkan prediksi cuaca yang akurat.

2. MATERIALS AND METHODS

Jenis penelitian ini adalah penelitian kuantitatif (*quantitative research*). Penelitian ini menggunakan teknik data mining untuk mengolah series data pengamatan cuaca dan menganalisis hasilnya untuk mendapatkan serta menentukan algoritma apa yang paling baik digunakan untuk prediksi cuaca. [6]

2.1 Tahapan Penelitian

Tahapan penelitian merupakan sebuah metode yang digunakan untuk mendeskripsikan langkah-langkah yang dilakukan mulai dari identifikasi masalah sampai mendapatkan sebuah solusi dari permasalahan yang ada[7]. Metodologi penelitian ini mengadaptasi sebagian tahapan dari CRISP-DM, dengan fokus pada tahapan pemrosesan data, modeling (menggunakan *Random Forest*), dan evaluasi hasil. Tahapan *business understanding* dan *deployment* tidak dibahas karena ruang lingkup penelitian ini hanya pada aspek teknis. Tahapan penelitian yang dilakukan dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Penelitian

2.2 Pengumpulan Data

Tahapan awal dalam penelitian ini adalah pengumpulan data berbentuk dataset. Dataset yang digunakan merupakan data cuaca harian yang tercatat di Stasiun Meteorologi Kemayoran dan diunduh dari Database Online BMKG [8]. Terdiri atas total 7 atribut dari 1 Januari 2024 hingga 30 Mei 2025.

Tabel 1. Detail Atribut Dataset

No	Nama Atribut	Keterangan
1.	Tanggal	Tanggal data diambil
2.	TAVG	Temperatur rata-rata harian (TAVG = Temperature Average), dalam °C.
3.	RH_AVG	Kelembapan relatif rata-rata (Relative Humidity Average), dalam persen (%).
4.	RR	Curah hujan (Rainfall/Precipitation), dalam milimeter (mm).
5.	SS	Lama penyinaran matahari (Sunshine Duration), dalam jam.
6.	DDD_X	Arah angin maksimum (Direction of the strongest wind), dalam derajat (°).
7.	FF_AVG	Kecepatan angin rata-rata (Wind Speed Average), dalam m/s atau knot.

Selanjutnya adalah sebagian Preview Dataset yang digunakan dalam penelitian ini.

Tabel 2. Sampel Dataset

TANGG AL	TAV G	RH_A VG	RR	SS	DDD _X	FF_A VG
01-01- 2024	29,6	77	1,8	5,6	290	1
02-01- 2024	29,7	78	1,5	0,7	300	1
03-01- 2024	28	84	2,5	0,3	260	1
04-01- 2024	27,1	87	8,5	3,1	330	0

2.4 Preparation Data

Pada tahap ini data disiapkan untuk dilakukan proses pelatihan. Data. Kemudian tahap pengolahan data yang bertujuan untuk membangun dataset akhir yang akan diproses pada tahap pemodelan. Tahapan pengolahan data mencakup pemilihan kelas, atribut-atribut data serta transformasi data kemudian dilakukan proses pembersihan dan regulasi data, pada proses ini mencoba untuk menghilangkan missing value dan lain sebagainya. [9]

2.5 Pemmodelan Random Forest

Langkah berikutnya adalah menerapkan model machine learning dengan algoritma Random Forest bekerja dengan cara membuat banyak Decision Tree. Kemudian dari seluruh hasil prediksi yang dibuat, dilakukan majority voting untuk menentukan hasil prediksi akhir. Hal ini secara langsung dapat mengatasi masalah ketika melakukan klasifikasi menggunakan satu pohon keputusan saja sering kali tidak optimal [10].

2.6 Analisa Model

Tahap ini bertujuan untuk mengevaluasi kinerja model *Random Forest* yang telah dibangun pada tahap pemodelan. Analisis dilakukan dengan membandingkan nilai hasil prediksi terhadap nilai aktual menggunakan beberapa metrik evaluasi. Metrik yang digunakan dalam penelitian ini antara lain adalah, yaitu:

- *Mean Squared Error* (MSE): rata-rata kuadrat selisih antara nilai aktual dan prediksi
- *Root Mean Squared Error* (RMSE): akar dari MSE, untuk mengetahui rata-rata kesalahan dalam satuan yang sama dengan data asli
- *Mean Absolute Error* (MAE): rata-rata nilai absolut dari kesalahan prediksi
- *Mean Absolute Percentage Error* (MAPE): persentase kesalahan rata-rata dibandingkan dengan nilai aktual
- Koefisien Determinasi (R^2): mengukur seberapa besar variasi data yang dapat dijelaskan oleh model

3. RESULTS AND DISCUSSION

Pada penelitian ini, kinerja model *Random Forest* dalam memprediksi cuaca di wilayah Jakarta dievaluasi melalui analisis sejumlah metrik yang telah dijelaskan sebelumnya. Metrik tersebut mencakup MSE, RMSE, MAE, MAPE, dan R^2 , yang digunakan untuk memberikan gambaran yang lebih komprehensif terkait akurasi dan efektivitas model.

Pada tahap evaluasi, performa model Random Forest diuji dengan menggunakan dataset pengujian yang dipisahkan dari dataset pelatihan, dengan perbandingan 80:20 antara data latih dan data tes. Hasil evaluasi ini memberikan gambaran objektif mengenai kemampuan model untuk menggeneralisasi pola yang diperoleh dari data pelatihan ke data yang belum pernah ditinjau sebelumnya.

Tabel di bawah ini menampilkan nilai rata-rata dari berbagai metrik evaluasi yang diperoleh melalui pengujian model Random Forest pada dataset pengujian.

Tabel 3. Hasil Evaluasi

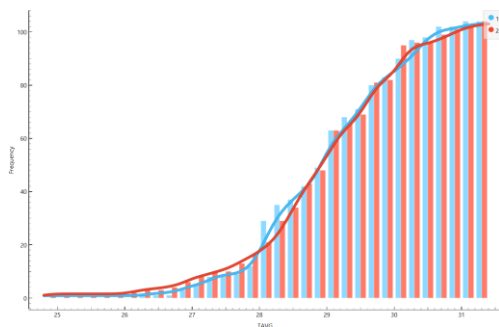
No	Metrik	Hasil
1.	MSE	0.398
2.	RMSE	0.631
3.	MAE	0.488
4.	MAPE	0.017
5.	R^2	0.653

Tabel yang disajikan di atas menggambarkan hasil evaluasi menyeluruh terhadap kinerja model Random Forest dalam memprediksi suhu rata-rata TAVG di Kota Jakarta. Berdasarkan nilai-nilai metrik yang dihasilkan, model menunjukkan performa yang cukup baik. Nilai MSE sebesar 0.398 dan RMSE sebesar 0.631 mengindikasikan bahwa kesalahan prediksi model relatif kecil dan berada dalam kisaran yang dapat diterima. Selain itu, nilai MAE sebesar 0.488 menunjukkan bahwa secara rata-rata, selisih antara prediksi dan nilai aktual suhu hanya sekitar setengah derajat Celcius, yang menandakan ketepatan prediksi yang tinggi.

Lebih lanjut, nilai *Mean Absolute Percentage Error* (MAPE) yang sangat rendah, yaitu sebesar 0.017 atau 1.7%, memperkuat bukti bahwa model memiliki tingkat ketelitian yang sangat baik dalam melakukan prediksi. Sementara itu, nilai *koefisien determinasi* (R^2) sebesar 0.653 menunjukkan bahwa sekitar 65.3% variasi suhu dapat dijelaskan oleh model berdasarkan fitur-fitur cuaca yang digunakan. Walaupun nilai R^2 ini belum mencapai angka maksimal, performanya sudah menunjukkan bahwa model cukup handal dalam mengidentifikasi pola dalam data cuaca. Secara keseluruhan, hasil evaluasi ini menunjukkan bahwa

model Random Forest dapat digunakan secara efektif untuk mendukung prediksi suhu di Jakarta dengan tingkat akurasi yang memadai.

Dengan menggabungkan hasil dari berbagai metrik evaluasi dan visualisasi distribusi data, dapat disimpulkan bahwa model Random Forest menunjukkan kinerja yang kuat dalam memprediksi kondisi cuaca, khususnya di lingkungan yang dinamis seperti Kota Jakarta. Hasil ini memberikan dasar yang meyakinkan bagi kesimpulan mengenai efektivitas model dalam memberikan solusi yang handal dan akurat dalam prediksi cuaca.



Gambar 2. Grafik Distribusi

Berdasarkan grafik Distribusi yang tersaji dapat diamati bahwa kurva distribusi dari kedua kelompok data—baik hasil prediksi maupun data aktual—menunjukkan pola kenaikan yang sangat serupa dan konsisten. Kedua kurva menjulur secara bertahap hingga mencapai frekuensi maksimum, dengan bentuk yang hampir tumpang tindih, khususnya pada rentang nilai suhu antara 27 hingga 31 derajat. Hal ini mengindikasikan bahwa model Random Forest berhasil menangkap karakteristik distribusi data dengan sangat baik.

Ketepatan ini tercermin dari kecocokan distribusi kumulatif antara data prediksi dan data aktual, yang menunjukkan bahwa model tidak hanya mampu menghasilkan prediksi mendekati nilai sebenarnya, tetapi juga mempertahankan struktur distribusi data secara keseluruhan. Dengan distribusi prediksi yang sejalan dengan pola aktual, dapat disimpulkan bahwa model memiliki akurasi dan konsistensi yang tinggi dalam menyajikan hasil prediksi.

Keselarasan kurva CDF dari kedua kelompok memberikan indikasi kuat bahwa model mampu memetakan suhu rata-rata secara tepat dalam berbagai kondisi, termasuk dinamika cuaca yang kompleks di wilayah perkotaan seperti Kota Jakarta. Hal ini memperkuat kepercayaan terhadap keandalan model Random Forest dalam memberikan informasi cuaca yang presisi.

Secara keseluruhan, grafik ini melengkapi hasil evaluasi kuantitatif lainnya (seperti MSE, RMSE, MAE, dan R^2), dan menegaskan bahwa model Random Forest tidak hanya efektif secara numerik, tetapi juga stabil dan terpercaya secara visual dalam menggambarkan distribusi cuaca. Dengan demikian, model ini sangat layak digunakan untuk berbagai keperluan praktis, seperti perencanaan sektor pertanian, transportasi, serta mendukung upaya mitigasi risiko iklim di wilayah urban seperti Jakarta.

4. CONCLUSION

Penelitian ini menggunakan metode Random Forest untuk melakukan prediksi suhu rata-rata harian (TAVG) di Kota Jakarta. Berdasarkan evaluasi menggunakan metrik MSE, RMSE, MAE, MAPE, dan R^2 , model menunjukkan performa yang baik dengan tingkat kesalahan yang rendah dan akurasi yang memadai. Nilai MSE sebesar 0.398, RMSE sebesar 0.631, MAE sebesar 0.488, dan MAPE sebesar 1.7% menunjukkan ketelitian yang tinggi dalam prediksi suhu. Selain itu, nilai R^2 sebesar 0.653 mengindikasikan bahwa model mampu menjelaskan sebagian besar variasi dalam data cuaca. Visualisasi distribusi data dan kurva kumulatif memperkuat temuan ini, dengan pola prediksi yang sejalan dengan data aktual.

Dengan hasil ini, model Random Forest memiliki potensi besar untuk dikembangkan sebagai alat prediksi cuaca yang akurat dan andal di wilayah perkotaan seperti Jakarta. Keakuratan dan kestabilan model menjadikannya relevan untuk digunakan dalam berbagai sektor seperti pertanian, transportasi, pariwisata, dan mitigasi risiko bencana.

Meskipun kinerja model cukup baik, masih terdapat ruang untuk pengembangan lebih lanjut, seperti dengan menambahkan variabel cuaca lainnya atau mencoba algoritma pembelajaran mesin lain untuk meningkatkan akurasi dan generalisasi model. Dengan demikian, model ini dapat semakin optimal dalam mendukung pengambilan keputusan berbasis prediksi cuaca di masa mendatang.

ACKNOWLEDGEMENTS

Terima kasih kepada Bapak Sumanto dan Ade Surya Budiman selaku pembimbing pada penelitian ini dan kepada teman-teman S1 Teknik Informatika Angkatan 2022 yang telah memberikan support. Tak lupa untuk mengucapkan terima kasih kepada kedua orang tua yang telah mendukung baik itu materil maupun non materil.

REFERENCES

- [1] G. Zittis *et al.*, "Climate Change and Weather Extremes in the Eastern Mediterranean and Middle East," *Rev. Geophys.*, vol. 60, no. 3, 2022, doi: 10.1029/2021RG000762.

- [2] Wijoyo A, Saputra A, Ristanti S, Sya'ban S, Amalia M, and Febriansyah R, "Pembelajaran Machine Learning," *OKTAL (Jurnal Ilmu Komput. dan Sci.*, vol. 3, no. 2, pp. 375–380, 2024, [Online]. Available: <https://journal.mediapublikasi.id/index.php/oktal/article/view/2305>
- [3] I. Nurmaulida, A. S. Sunge, and A. T. Zy, "Penggunaan Naïve Bayes dalam Implementasi Prediksi Tingkat Curah Hujan," *J. Ilm. Mhs. Pendidik. Sej.*, vol. 8, no. 3, pp. 3149–3157, 2023, [Online]. Available: <https://jim.usk.ac.id/sejarah>
- [4] U. Indahyanti, N. L. Azizah, and H. Setiawan, "Pendekatan Ensemble Learning Untuk Meningkatkan Akurasi Prediksi Kinerja Akademik Mahasiswa," *J. Sains dan Inform.*, vol. 8, no. 2, pp. 160–169, 2022, doi: 10.34128/jsi.v8i2.459.
- [5] Z. A. Dwiyantri and C. Prianto, "Prediksi Cuaca Kota Jakarta Menggunakan Metode Random Forest," *J. Tekno Insentif*, vol. 17, no. 2, pp. 127–137, 2023, doi: 10.36787/jti.v17i2.1136.
- [6] R. Prasetya, "Penerapan Teknik Data Mining Dengan Algoritma Classification Tree untuk Prediksi Hujan," *J. Widya Climago*, vol. 2, no. 2, pp. 13–23, 2020.
- [7] A. Yudhistira and R. Andika, "Pengelompokan Data Nilai Siswa Menggunakan Metode K-Means Clustering," *J. Artif. Intell. Technol. Inf.*, vol. 1, no. 1, pp. 20–28, 2023, doi: 10.58602/jaiti.v1i1.22.
- [8] A. S. Agung, A. A. Fauzi, A. A. Nur Risal, and F. Adiba, "Implementasi Teknik Data Mining terhadap Klasifikasi Data Prediksi Curah Hujan BMKG Di Sulawesi Selatan," *J. Tekno Insentif*, vol. 17, no. 1, pp. 22–23, 2023, doi: 10.36787/jti.v17i1.955.
- [9] R. Supriyadi, W. Gata, N. Maulidah, and A. Fauzi, "Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah," *E-Bisnis J. Ilm. Ekon. dan Bisnis*, vol. 13, no. 2, pp. 67–75, 2020, doi: 10.51903/e-bisnis.v13i2.247.
- [10] L. Sari, A. Romadloni, and R. Listyaningrum, "Penerapan Data Mining dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest," *Infotekmesin*, vol. 14, no. 1, pp. 155–162, 2023, doi: 10.35970/infotekmesin.v14i1.1751.