



Komparasi Metode Gradient Boosting dan Metode Random Forest Untuk Prediksi Curah Hujan

Kintan Febiola¹, Savitri²

¹²Teknik Informatika, Fakultas Ilmu Komputer, Universitas Pamulang

¹Kintanfebiola04@gmail.com, ²dosen02410@unpam.ac.id

Abstrak

Curah hujan di wilayah Tangerang Selatan memiliki pola yang berubah-ubah sehingga sulit diprediksi secara tepat. Penelitian ini bertujuan membandingkan metode Gradient Boosting dan Random Forest untuk memprediksi curah hujan berdasarkan data meteorologi wilayah Tangerang Selatan periode 2024–2026. Data diperoleh dari BMKG Wilayah II dan terdiri atas 707 record harian. Variabel yang digunakan meliputi rata-rata suhu udara (TAVG), rata-rata kelembapan udara (RH_AVG), lama penyinaran matahari (SS), kecepatan angin maksimum (FF_X), kecepatan angin rata-rata (FF_AVG), tekanan udara (Pressure), dan curah hujan sebagai target. Penelitian menggunakan pendekatan kuantitatif dan eksperimen. Data dibagi menjadi 565 data training dan 142 data testing dengan rasio 80:20. Evaluasi menggunakan Mean Absolute Error (MAE), Root Mean Square Error (RMSE), dan R^2 . Gradient Boosting memperoleh MAE 7,956, RMSE 18,175, dan R^2 0,032, sedangkan Random Forest memperoleh MAE 7,886, RMSE 18,021, dan R^2 0,048. Hasil menunjukkan bahwa Random Forest memiliki performa komparatif lebih baik, meskipun nilai R^2 kedua model masih rendah sehingga kemampuan dalam menjelaskan variasi curah hujan masih terbatas.

Kata Kunci: Gradient Boosting, Random Forest, Prediksi Curah Hujan, Machine Learning, Data Meteorologi

1. Latar Belakang

Curah hujan merupakan salah satu unsur meteorologi yang memiliki pola berubah-ubah. Perubahan curah hujan berkaitan dengan kondisi beberapa unsur cuaca, seperti suhu udara, kelembapan, tekanan udara, kecepatan angin, dan lama penyinaran matahari. Hubungan antarvariabel tersebut dapat membentuk pola yang berbeda dari waktu ke waktu sehingga prediksi curah hujan menjadi suatu permasalahan yang cukup kompleks. Penggunaan data meteorologi sebagai dasar pemodelan prediksi telah banyak diterapkan dalam penelitian berbasis *Machine Learning* [1].

Kondisi perubahan curah hujan juga terlihat pada wilayah Tangerang Selatan. Berdasarkan data meteorologi yang digunakan dalam penelitian, curah hujan pada tahun 2024 tercatat sebesar 601,1 mm pada Januari, kemudian turun menjadi 37,2 mm pada September dan kembali meningkat menjadi 379,1 mm pada November. Pada Maret 2025, curah hujan tercatat sebesar 573,7 mm. Variasi juga terlihat pada variabel meteorologi lainnya. Kelembapan udara rata-rata berada pada kisaran 69,9%–83,1%, tekanan udara berkisar 1004,1–1007,1 mb, sedangkan kecepatan angin rata-rata berkisar 0,6–8,5 knot. Kondisi tersebut menunjukkan bahwa karakteristik meteorologi wilayah Tangerang Selatan bersifat dinamis.

Perubahan curah hujan yang tidak menentu dapat meningkatkan risiko berbagai permasalahan lingkungan, seperti banjir dan kekeringan. Kondisi tersebut menunjukkan pentingnya pendekatan yang dapat memanfaatkan data historis untuk mengenali pola dan menghasilkan prediksi. *Machine Learning* dapat digunakan untuk mempelajari hubungan antara variabel meteorologi dan curah hujan berdasarkan data yang tersedia. Penerapan *Machine Learning* pada data lingkungan juga telah digunakan untuk membantu proses prediksi dan mendukung pengambilan keputusan berbasis data [2].

Berbagai penelitian telah menerapkan metode *Machine Learning* untuk memprediksi curah hujan maupun fenomena lingkungan. Gradient Boosting dan Random Forest merupakan beberapa metode yang dapat digunakan untuk membangun model prediksi berdasarkan data meteorologi. Penggunaan kedua metode tersebut pada permasalahan lingkungan telah dilakukan dalam penelitian sebelumnya, meskipun objek dan karakteristik data yang digunakan berbeda [3][4].

Penelitian mengenai prediksi curah hujan juga telah dilakukan dengan membandingkan beberapa algoritma. Random Forest dan Gradient Boosting digunakan dalam penelitian prediksi curah hujan harian di Sumatra Selatan dengan pendekatan *time series cross-validation* [5]. Random Forest juga digunakan untuk prediksi curah hujan

dalam mendukung mitigasi kebakaran hutan dan lahan [6]. Selain itu, penelitian lain menggunakan Random Forest bersama beberapa algoritma *Machine Learning* untuk memprediksi curah hujan sekaligus menganalisis parameter iklim yang berpengaruh [7].

Penelitian mengenai prediksi curah hujan juga dilakukan menggunakan pendekatan regresi berdasarkan data historis cuaca. Hasil penelitian tersebut menunjukkan bahwa data meteorologi dapat digunakan sebagai dasar untuk membangun model prediksi nilai curah hujan [8]. Evaluasi model prediksi curah hujan berbasis *Machine Learning* pada wilayah lain juga menunjukkan adanya perbedaan performa model berdasarkan karakteristik wilayah dan data yang digunakan [9].

Penggunaan algoritma *ensemble* lainnya juga telah diterapkan dalam penelitian curah hujan. XGBoost dan Random Forest digunakan untuk prediksi curah hujan dalam mendukung mitigasi kebakaran hutan dan lahan, sedangkan XGBoost dan LightGBM digunakan untuk klasifikasi curah hujan di Jakarta [6][10].

Penelitian lain membandingkan beberapa algoritma *Machine Learning*, seperti Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Decision Tree, Random Forest, dan Gradient Boosting untuk prediksi hujan harian. Perbandingan tersebut menunjukkan bahwa berbagai algoritma dapat memberikan hasil yang berbeda terhadap data dan permasalahan yang digunakan [11]. Selain itu, metode CatBoost juga telah digunakan untuk prediksi curah hujan dengan memanfaatkan variabel meteorologi dan tahapan *preprocessing* sebelum pemodelan[12].

Evaluasi kinerja algoritma *Machine Learning* pada dataset dengan karakteristik dan ukuran yang berbeda juga menjadi bagian penting dalam menentukan model yang sesuai[13]. Pendekatan berbasis deret waktu seperti Long Short-Term Memory juga telah digunakan untuk memodelkan jumlah hari hujan sehingga menunjukkan adanya berbagai pendekatan yang dapat diterapkan pada data curah hujan [14]. Penelitian mengenai Random Forest dengan pendekatan optimasi juga telah diterapkan pada permasalahan lingkungan lain, yang menunjukkan bahwa pengembangan model dapat disesuaikan dengan karakteristik data dan tujuan penelitian [15].

Meskipun penelitian mengenai prediksi curah hujan menggunakan *Machine Learning* telah dilakukan pada berbagai wilayah, penelitian dengan karakteristik data dan wilayah yang berbeda dapat menghasilkan performa model yang berbeda. Penelitian yang secara khusus membandingkan Gradient Boosting dan Random Forest menggunakan data meteorologi harian wilayah Tangerang Selatan masih terbatas. Oleh karena itu, penelitian ini melakukan komparasi metode Gradient Boosting dan Random Forest untuk prediksi curah hujan berdasarkan data meteorologi wilayah Tangerang Selatan periode 2024–2026. Perbandingan dilakukan menggunakan MAE, RMSE, dan R^2 untuk mengetahui metode yang memberikan performa komparatif lebih baik.

2. Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen. Penelitian dilakukan dengan menerapkan metode Gradient Boosting dan Random Forest pada data meteorologi wilayah Tangerang Selatan. Kedua metode menggunakan dataset dan proses evaluasi yang sama sehingga hasil performanya dapat dibandingkan secara objektif. Tahapan penelitian meliputi pengumpulan data, *preprocessing* data, pembangunan model, pelatihan dan pengujian model, evaluasi, analisis komparatif, serta implementasi sistem. Tahapan tersebut ditunjukkan pada Gambar 1.



Gambar. 1 Tahapan Penelitian

Data penelitian diperoleh dari BMKG Wilayah II untuk wilayah Tangerang Selatan dengan periode pengamatan 2024–2026. Dataset terdiri atas 707 record harian. Data tersebut digunakan untuk membangun model prediksi curah hujan dengan memanfaatkan variabel meteorologi sebagai variabel masukan.

Variabel yang digunakan terdiri atas TAVG, RH_AVG, SS, FF_X, FF_AVG, Pressure, dan RR. TAVG merupakan rata-rata suhu udara, RH_AVG merupakan rata-rata kelembapan udara, SS merupakan lama penyinaran matahari, FF_X merupakan kecepatan angin maksimum, FF_AVG merupakan kecepatan angin rata-rata, dan Pressure merupakan tekanan udara. Sementara itu, RR digunakan sebagai variabel target berupa curah hujan.

Tabel.1 Variabel Penelitian

Variabel	Keterangan	Peran
TAVG	Rata-rata suhu udara	Variabel input
RH_AVG	Rata-rata kelembapan udara	Variabel input
SS	Lama penyinaran matahari	Variabel input
FF_X	Kecepatan angin maksimum	Variabel input
FF_AVG	Kecepatan angin rata-rata	Variabel input
Pressure	Tekanan udara	Variabel input
RR	Curah hujan	Variabel target

Tahap *preprocessing* dilakukan untuk mempersiapkan data sebelum proses pemodelan. Proses yang dilakukan meliputi pembersihan data, penanganan *missing value*, penentuan variabel penelitian, dan transformasi variabel target menggunakan metode log1p. Transformasi tersebut diterapkan untuk memperoleh distribusi data yang lebih baik sebelum digunakan dalam proses pelatihan model.

Setelah *preprocessing*, dataset dibagi menjadi data training dan testing menggunakan metode *train-test split*. Dari 707 record, sebanyak 565 data digunakan sebagai data training dan 142 data digunakan sebagai data testing dengan rasio 80:20. Pembagian yang sama digunakan untuk kedua algoritma sehingga hasil evaluasi dapat dibandingkan secara langsung.

Pemodelan dilakukan menggunakan metode Gradient Boosting dan Random Forest. Gradient Boosting membangun model secara bertahap dengan memperbaiki kesalahan prediksi dari model sebelumnya, sedangkan Random Forest menggunakan sejumlah pohon keputusan untuk menghasilkan prediksi. Kedua metode diterapkan pada variabel meteorologi yang telah melalui proses *preprocessing*. Penggunaan Gradient Boosting dan Random Forest dalam pemodelan data lingkungan telah diterapkan pada beberapa penelitian sebelumnya [3][5].

Selanjutnya, kedua model dilatih menggunakan data training dan diuji menggunakan data testing. Performa model dievaluasi menggunakan Mean Absolute Error (MAE), Root Mean Square Error (RMSE), dan Coefficient of Determination (R^2). MAE dan RMSE digunakan untuk mengukur tingkat kesalahan prediksi, sedangkan R^2 digunakan untuk mengetahui kemampuan model dalam menjelaskan variasi data curah hujan. Penggunaan metrik tersebut juga diterapkan dalam penelitian prediksi berbasis *Machine Learning* pada data cuaca dan lingkungan [8][9].

Tahap terakhir adalah analisis komparatif dan implementasi. Nilai MAE, RMSE, dan R^2 dari kedua model dibandingkan untuk mengetahui metode dengan performa yang lebih baik. Model yang telah dievaluasi kemudian diimplementasikan menggunakan Gradio pada Google Colab. Antarmuka tersebut digunakan untuk melakukan pengujian prediksi berdasarkan nilai variabel meteorologi yang dimasukkan oleh pengguna.

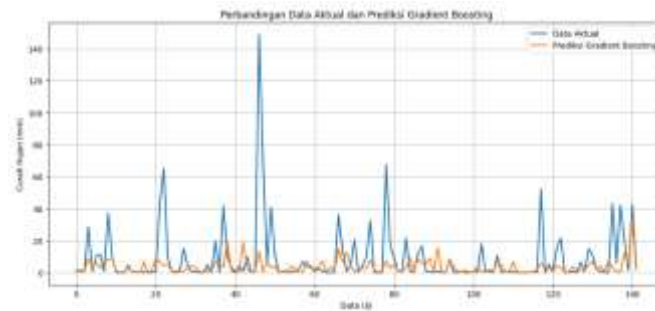
3. Hasil dan Diskusi

Pengujian dilakukan menggunakan data testing yang sama pada kedua model. Penggunaan dataset pengujian yang sama bertujuan agar hasil Gradient Boosting dan Random Forest dapat dibandingkan secara objektif. Hasil evaluasi dari kedua model disajikan pada Tabel 2.

Tabel. 2 Hasil Evaluasi Gradient Boosting dan Random Forest

Model	MAE	RMSE	R^2
Gradient Boosting	7,956	18,175	0,032
Random Forest	7,886	18,021	0,048

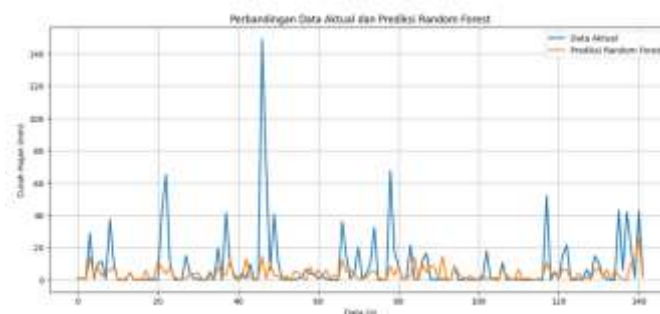
Berdasarkan Tabel 2, Gradient Boosting memperoleh nilai MAE sebesar 7,956, RMSE sebesar 18,175, dan R^2 sebesar 0,032. Nilai tersebut menunjukkan tingkat kesalahan prediksi yang dihasilkan model pada data testing. Nilai R^2 yang diperoleh juga menunjukkan bahwa kemampuan Gradient Boosting dalam menjelaskan variasi curah hujan pada dataset penelitian masih terbatas.



Gambar.2 Perbandingan Data Aktual dan Prediksi Gradient Boosting

Visualisasi pada Gambar 2 menunjukkan perbandingan antara data aktual dan hasil prediksi Gradient Boosting. Grafik digunakan untuk melihat perbedaan antara nilai curah hujan aktual dengan nilai yang dihasilkan oleh model pada data pengujian. Gambar tersebut merupakan visualisasi hasil pengujian model Gradient Boosting yang terdapat dalam skripsi.

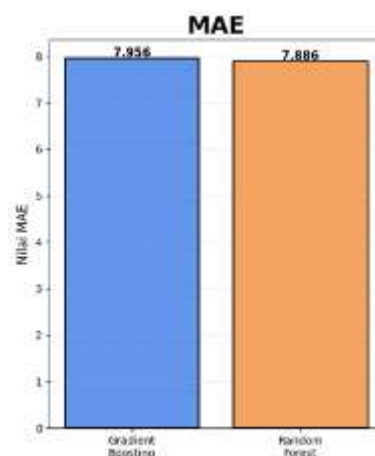
Pada model Random Forest, hasil pengujian memperoleh MAE sebesar 7,886, RMSE sebesar 18,021, dan R^2 sebesar 0,048. Nilai MAE dan RMSE tersebut lebih rendah dibandingkan hasil Gradient Boosting. Sementara itu, nilai R^2 Random Forest lebih tinggi dibandingkan Gradient Boosting.



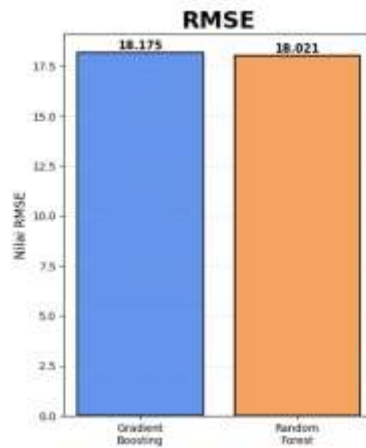
Gambar. 3 Perbandingan Data Aktual dan Prediksi Random Forest

Gambar 3 memperlihatkan perbandingan antara data aktual dan hasil prediksi Random Forest. Visualisasi tersebut digunakan untuk melihat hasil prediksi model terhadap data aktual pada data testing

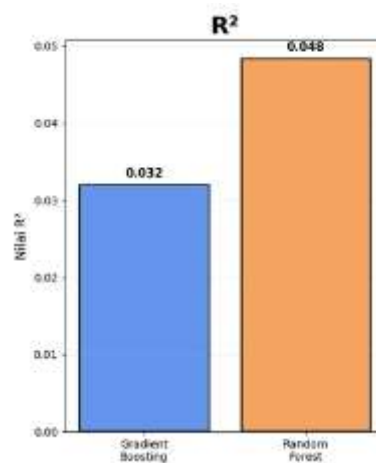
Perbandingan ketiga metrik menunjukkan bahwa Random Forest memiliki performa yang lebih baik pada dataset penelitian. Nilai MAE Random Forest sebesar 7,886 lebih rendah daripada Gradient Boosting sebesar 7,956. Demikian pula nilai RMSE Random Forest sebesar 18,021 lebih rendah daripada Gradient Boosting sebesar 18,175. Pada metrik R^2 , Random Forest memperoleh nilai 0,048, lebih tinggi dibandingkan Gradient Boosting sebesar 0,032.



Gambar. 4 Perbandingan Nilai MAE



Gambar. 5 Perbandingan Nilai RMSE



Gambar. 6 Perbandingan Nilai R2

Berdasarkan hasil tersebut, Random Forest memiliki performa komparatif yang lebih baik dibandingkan Gradient Boosting pada dataset penelitian. Random Forest menghasilkan nilai MAE dan RMSE yang lebih rendah serta R^2 yang lebih tinggi. Hasil ini menunjukkan bahwa Random Forest menghasilkan kesalahan prediksi yang lebih rendah pada data yang digunakan dalam penelitian.

Hasil penelitian ini dapat dibandingkan dengan penelitian sebelumnya. Penelitian sebelumnya menggunakan Random Forest dan Gradient Boosting untuk prediksi curah hujan harian di Sumatra Selatan dengan pendekatan *time series cross-validation* [5]. Perbedaan wilayah dan pendekatan evaluasi menunjukkan bahwa hasil penelitian tersebut tidak dapat langsung disamakan dengan penelitian ini. Penelitian ini menggunakan data meteorologi Tangerang Selatan periode 2024–2026 dengan pembagian data *training* dan *testing* sebesar 80:20.

Penelitian sebelumnya juga menggunakan data cuaca BMKG serta Gradient Boosting dan Random Forest. Namun, penelitian tersebut berfokus pada klasifikasi kondisi cuaca, sedangkan penelitian ini menggunakan pendekatan regresi untuk menghasilkan nilai curah hujan. Perbedaan tujuan pemodelan tersebut menjadi salah satu faktor yang membedakan hasil penelitian [1].

Penelitian sebelumnya menggunakan Random Forest bersama beberapa algoritma lainnya untuk prediksi curah hujan serta menganalisis variabel iklim yang berpengaruh. Penelitian tersebut memiliki kesamaan pada penggunaan data meteorologi dan Random Forest, tetapi fokus penelitian ini adalah membandingkan performa Gradient Boosting dan Random Forest tanpa melakukan analisis interpretasi fitur secara mendalam [7].

Hasil penelitian sebelumnya menunjukkan bahwa Gradient Boosting dapat memberikan hasil yang lebih stabil dibandingkan Random Forest pada permasalahan prediksi banjir. Hasil tersebut berbeda dengan penelitian ini karena Random Forest justru memberikan performa komparatif yang lebih baik. Perbedaan tersebut menunjukkan bahwa performa algoritma dapat dipengaruhi oleh karakteristik dataset dan permasalahan yang digunakan [3].

Meskipun Random Forest memberikan hasil yang lebih baik, nilai R^2 kedua model masih rendah. Gradient Boosting memperoleh R^2 sebesar 0,032, sedangkan Random Forest memperoleh 0,048. Hal ini menunjukkan bahwa kemampuan kedua model dalam menjelaskan variasi curah hujan pada dataset penelitian masih terbatas. Oleh karena itu, Random Forest dapat dinyatakan sebagai metode yang lebih baik secara komparatif pada dataset penelitian ini, bukan sebagai metode yang selalu lebih unggul pada semua kondisi.

Model yang telah dievaluasi kemudian diimplementasikan menggunakan Gradio pada Google Colab. Implementasi ini digunakan sebagai antarmuka sederhana untuk melakukan prediksi berdasarkan nilai variabel meteorologi yang dimasukkan pengguna.

Gambar. 7 Implementasi Antarmuka Prediksi Curah Hujan

Implementasi tersebut menunjukkan penerapan model hasil penelitian ke dalam sebuah antarmuka. Pengguna dapat memasukkan nilai variabel meteorologi yang tersedia, memilih model yang digunakan, kemudian menjalankan proses prediksi curah hujan.

4. Kesimpulan

Penelitian ini melakukan komparasi metode Gradient Boosting dan Random Forest untuk prediksi curah hujan berdasarkan data meteorologi wilayah Tangerang Selatan periode 2024–2026. Data yang digunakan berasal dari BMKG Wilayah II dan terdiri atas 707 record harian, dengan TAVG, RH_AVG, SS, FF_X, FF_AVG, dan Pressure sebagai variabel input serta RR sebagai variabel target. Dataset dibagi menjadi 565 data training dan 142 data testing dengan rasio 80:20. Hasil evaluasi menunjukkan bahwa Gradient Boosting memperoleh MAE 7,956, RMSE 18,175, dan R^2 0,032, sedangkan Random Forest memperoleh MAE 7,886, RMSE 18,021, dan R^2 0,048. Berdasarkan ketiga metrik tersebut, Random Forest memberikan performa komparatif yang lebih baik karena memiliki MAE dan RMSE lebih rendah serta R^2 lebih tinggi dibandingkan Gradient Boosting. Meskipun Random Forest memberikan hasil yang lebih baik, nilai R^2 kedua model masih rendah sehingga kemampuan model dalam menjelaskan variasi curah hujan masih terbatas. Dengan demikian, Random Forest merupakan metode dengan performa komparatif lebih baik pada dataset yang digunakan dalam penelitian ini. Pengembangan selanjutnya dapat dilakukan dengan menggunakan data meteorologi dalam periode yang lebih panjang, wilayah berbeda, atau metode *Machine Learning* lainnya untuk memperoleh perbandingan yang lebih luas.

Referensi

- [1] W. Indrasari and H. Suhendar, "ANALISIS MODEL PREDIKSI CUACA MENGGUNAKAN SUPPORT VECTOR MACHINE, GRADIENT BOOSTING, RANDOM FOREST, DAN," vol. XII, pp. 119–128, 2024.
- [2] E. Pitaloka *et al.*, "Penerapan Machine Learning untuk Prediksi Bencana Banjir," vol. 01, 2024, doi: 10.21456/vol14iss1pp62-76.
- [3] I. Maulita *et al.*, "Analisis Komparatif Linear Regression, Random Forest, dan Gradient Boosting untuk Prediksi Banjir Comparative Analysis of Linear Regression, Random Forest, and Gradient Boosting Algorithms in Flood Probability Prediction," vol. 4, no. 8, pp. 369–379, 2024.

- [4] P. A. Saputra, S. S. Irawan, R. Prianto, and T. Hidayat, "PREDIKSI DAN ANALISIS POLA PERUBAHAN IKLIM," vol. 13, no. 2, 2025.
- [5] M. Alvines *et al.*, "KOMPARASI RIDGE REGRESSION , RANDOM FOREST , DAN GRADIENT BOOSTING UNTUK PREDIKSI CURAH HUJAN HARIAN DI SUMATRA SELATAN BERBASIS TIME SERIES CROSS-VALIDATION," vol. 9, no. 4, pp. 5742–5748, 2025.
- [6] D. Sangaji and T. Sutabri, "E ISSN : 2809-4069 Analisis XGBoost dan Random Forest untuk Prediksi Curah Hujan dalam Mendukung Mitigasi Karhutla," vol. 5, no. 1, pp. 13–18, 2025.
- [7] H. Gani, A. D. Damayanti, M. M. Mubarak, and H. Gani, "Model Pembelajaran Mesin Yang Dapat Dijelaskan Untuk Menjelaskan Parameter Iklim Yang Berpengaruh Berdasarkan Prediksi Curah Hujan," vol. 15, no. 2, pp. 98–110, 2024.
- [8] D. P. Diyani, E. R. Sriwijayanti, and R. Gustrianysah, "KOMPUTA : Jurnal Ilmiah Komputer dan Informatika Prediksi Curah Hujan Berbasis Regresi Polinomial Menggunakan Data Historis Cuaca Polynomial Regression Based Rainfall Prediction Using Historical Weather Data KOMPUTA : Jurnal Ilmiah Komputer dan Informatika," vol. 14, no. 2, pp. 6–13, 2025, doi: 10.34010/komputa.v14i2.
- [9] J. Nasional, S. Informasi, I. Hapsari, and S. Pandya, "Evaluasi Model Prediksi Curah Hujan Berbasis Machine Learning di Kota Bandung," vol. 02, pp. 136–143, 2025.
- [10] C. Gde and L. Pringandana, "A Comparative Analysis of Hyperparameter-Tuned XGBoost and LightGBM for Multiclass Rainfall Classification in Jakarta," vol. 6, no. 4, pp. 2467–2483, 2025.
- [11] J. Ilmiah, S. Informasi, A. Aprilia, A. B. Wahidin, and P. K. Karo, "Perbandingan Algoritma Machine Learning (Logistic Regression , SVM , KNN , Decision Tree , Random Forest , dan Gradient Boosting) dalam Prediksi Hujan Harian di Provinsi Lampung Ketersediaan data meteorologi yang berkembang seiring dengan berkembangnya Teknik Machine Learning (ML). Machine Learning menjadi salah satu pendekatan yang baik dalam memodelkan fenomena lingkungan khususnya atmosfer . Algoritma ML memberikan berbagai model Machine Learning yaitu Logistic Regression , Support Vector Machine , K-Nearest," vol. 5, no. November, 2025.
- [12] A. F. Istianto, A. I. Hadiana, and F. R. Umbara, "PREDIKSI CURAH HUJAN MENGGUNAKAN METODE CATEGORICAL BOOSTING (CATBOOST)," vol. 7, no. 4, pp. 2930–2937, 2023.
- [13] D. Lusiyantri, S. Musdalifah, A. Sahari, and I. Al Fajri, "EVALUASI KINERJA ALGORITMA MACHINE LEARNING PADA DATASET SKALA BESAR," vol. 07, no. 01, pp. 84–92, 2025.
- [14] N. H. Setyawan and N. Wakhidah, "Analisis Perbandingan Metode Logistic Regression, Random Forest, Gradient Boosting Untuk Prediksi Diabetes," *JUPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 10, no. 1, pp. 150–162, 2025, doi: 10.29100/jupi.v10i1.5743.
- [15] I. Taufiq, T. Azhima, Y. Siswa, and W. J. Pranoto, "Model Optimasi Random Forest dengan PSO-CHI-SM dalam Mengatasi High Dimensional dan Imbalanced Data Banjir Kota Samarinda," vol. 7, no. 3, 2024, doi: 10.32493/jtsi.v7i3.41632.