Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 438−446
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9018

Prediksi Curah Hujan Jawa Barat Menggunakan Algoritma Machine Learning: Analisis Komparatif Berbasis Data Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) 2024

Alif Fahmi*, Amali, Aceng Badruzzaman

Fakultas Teknik, Program Studi Teknik Informatika, Universitas Pelita Bangsa, Bekasi
Jl. Inspeksi Kalimalang Tegal Danas. Cikarang Pusat, Kabupaten Bekasi, Jawa Barat, Indonesia
Email: [1,*]aliffahmi.95@mhs.pelitabangsa.ac.id, [2]amali@pelitabangsa.ac.id, [3]aceng_badruzzaman@pelitabangsa.ac.id
Email Penulis Korespondensi: aliffahmi.95@mhs.pelitabangsa.ac.id
Submitted: **23/12/2025**; Accepted: **08/01/2026**; Published: **31/01/2026**

Abstrak−Provinsi Jawa Barat memiliki tingkat kerentanan yang tinggi terhadap bencana hidrometeorologi akibat variabilitas curah hujan yang dinamis, sehingga pengembangan sistem prediksi cuaca yang akurat menjadi kebutuhan krusial untuk mendukung upaya mitigasi bencana. Penelitian ini bertujuan untuk melakukan analisis perbandingan terhadap kinerja algoritma Machine Learning, yaitu Support Vector Machine (SVM), Naïve Bayes, Random Forest, dan XGBoost, dalam memprediksi kejadian hujan berdasarkan data meteorologi harian tahun 2024 yang bersumber dari BMKG. Melalui metode eksperimen komputasi yang menerapkan tiga skenario pembagian data 80:20, 75:25, dan 70:30, serta seleksi fitur Recursive Feature Elimination (RFE), hasil pengujian menunjukkan bahwa model Naïve Bayes, Random Forest, dan XGBoost secara konsisten mencapai akurasi sempurna sebesar 100% pada seluruh skenario, sementara SVM menunjukkan performa yang stabil namun lebih konservatif dengan rata-rata akurasi 95,4%. Analisis mendalam mengindikasikan bahwa capaian akurasi absolut pada kondisi data tertentu tersebut sangat dipengaruhi oleh dominasi fitur curah hujan harian (RR) yang menyebabkan indikasi kebocoran data (data leakage), di mana model ensemble dan probabilistik mampu mengeksploitasi hubungan deterministik fitur jauh lebih efektif dibandingkan SVM. Penelitian ini merekomendasikan penyesuaian fitur input yang lebih ketat dengan memprioritaskan indikator atmosferik pendahulu untuk pengembangan sistem peringatan dini yang lebih realistis dan adaptif di masa depan.

Kata Kunci: Prediksi Curah Hujan; Machine Learning; Data Leakage; Analisis Komparatif; Meteorologi Tropis

Abstract−West Java Province exhibits high vulnerability to hydrometeorological disasters due to dynamic rainfall variability, necessitating an accurate weather prediction system for effective disaster mitigation.1 This study aims to conduct a comparative performance analysis of Machine Learning algorithms, specifically Support Vector Machine (SVM), Naïve Bayes, Random Forest, and XGBoost, in predicting rainfall events based on 2024 daily meteorological data sourced from BMKG. Through computational experiments utilizing three data splitting scenarios 80:20, 75:25, and 70:30, and Recursive Feature Elimination (RFE), the results demonstrate that Naïve Bayes, Random Forest, and XGBoost consistently achieved a perfect accuracy of 100% across all scenarios, whereas SVM exhibited stable but more conservative performance with an average accuracy of 95.4%. In-depth analysis indicates that the absolute accuracy achieved under specific data conditions was significantly influenced by the dominance of the daily rainfall feature (RR), leading to indications of data leakage where ensemble and probabilistic models exploited deterministic relationships much more effectively than SVM. Consequently, this study recommends a rigorous re-evaluation of input features, prioritizing atmospheric leading indicators, to develop a more realistic and adaptive early warning system in the future.

Keywords: Rainfall Prediction; Machine Learning; Data Leakage; Comparative Analysis; Tropical Meteorology

1. PENDAHULUAN

Indonesia sebagai negara kepulauan yang terletak di garis khatulistiwa memiliki karakteristik iklim tropis dengan variabilitas curah hujan yang sangat tinggi. Kondisi ini menjadikan curah hujan sebagai salah satu parameter meteorologi paling kritis yang memengaruhi berbagai sektor penting, mulai dari pertanian, hingga manajemen kebencanaan. Kompleksitas dampak curah hujan ini terlihat nyata di Provinsi Jawa Barat, wilayah dengan topografi yang unik perpaduan pegunungan vulkanik dan dataran rendah, menciptakan pola curah hujan di wilayah ini sangat dinamis dan sulit diprediksi. Tingginya intensitas curah hujan di Jawa Barat sering kali menjadi pemicu utama terjadinya bencana hidrometeorologi. Berdasarkan data Indeks Risiko Bencana Indonesia (IRBI) tahun 2024 yang dirilis oleh Badan Nasional Penanggulangan Bencana (BNPB), sekitar 99,34% dari total kejadian bencana di Indonesia merupakan bencana hidrometeorologi, dengan banjir dan tanah longsor sebagai kejadian paling dominan yang kerap melanda wilayah dengan curah hujan tinggi seperti Jawa Barat. Risiko ini diperparah oleh fenomena perubahan iklim global yang menyebabkan anomali cuaca, sehingga prediksi curah hujan yang akurat menjadi kebutuhan mendesak sebagai bagian dari sistem peringatan dini (early warning system) untuk mitigasi bencana yang efektif [1].

Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) memegang peranan sentral dalam penyediaan data cuaca historis yang menjadi landasan bagi penelitian klimatologi. Namun, tantangan utama muncul dari kompleksitas pola cuaca yang dipengaruhi oleh fenomena global seperti El Niño dan La Niña, yang membuat metode prediksi konvensional sering kali mengalami kesulitan dalam memetakan ketidakpastian ini secara presisi. Kompleksitas data meteorologi yang mencakup variabel suhu, kelembapan, tekanan udara, serta kecepatan angin menuntut penerapan teknologi komputasi cerdas. Pendekatan Data Mining dan Machine Learning hadir sebagai

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 438−446
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9018

solusi relevan, di mana algoritma cerdas mampu mengolah data historis dalam jumlah besar (Big Data) untuk mengenali pola non-linear yang sulit dideteksi oleh analisis statistik manual. Pemanfaatan algoritma ini telah terbukti mampu meningkatkan akurasi prediksi cuaca dibandingkan metode konvensional, terutama dalam menangani data deret waktu (time series) yang fluktuatif [2].

Dalam ranah Machine Learning untuk prediksi cuaca, terdapat empat algoritma yang sering menjadi fokus penelitian karena karakteristik uniknya, yaitu Support Vector Machine (SVM), Naïve Bayes, Random Forest, dan XGBoost. Pemilihan keempat model ini didasarkan pada perbedaan mendasar cara kerja mereka dalam menangani data. Naïve Bayes dikenal dengan efisiensi komputasinya yang tinggi namun sering kali terkendala oleh asumsi independensi fitur yang kaku, dalam studi terbaru Naïve Bayes terbukti memiliki akurasi yang lebih rendah (36,61%) dibandingkan metode ensemble saat diterapkan pada dataset iklim di Indonesia. Di sisi lain, Support Vector Machine (SVM) memiliki keunggulan dalam menangani data berdimensi tinggi melalui konsep kernel trick, meskipun kinerjanya sangat bergantung pada pemilihan parameter yang tepat dan sering kali kalah unggul dibandingkan metode berbasis pohon keputusan (Decision Trees) dalam menganalisis dataset meteorologi yang besar [3].

Sementara itu, algoritma berbasis ensemble learning seperti Random Forest dan XGBoost menunjukkan dominasi akurasi model. Random Forest bekerja dengan membangun banyak pohon keputusan untuk menjaga stabilitas prediksi dan mencegah overfitting. Dalam studi yang dilakukan di Kota Bandung menunjukkan bahwa Random Forest mampu mencapai akurasi hingga 85%, mengungguli metode regresi logistik dalam memprediksi kejadian hujan harian [4]. Namun, temuan berbeda studi yang membandingkan Random Forest, SVM, dan XGBoost. Hasil penelitian tersebut menyimpulkan bahwa XGBoost memberikan performa terbaik dengan nilai R-squared sebesar 0,8183 dan error terendah, berkat kemampuannya melakukan optimasi bertahap (boosting) yang efektif menangani missing value dan regularisasi data. Adanya perbedaan hasil performa antar-algoritma di berbagai studi ini menunjukkan bahwa tidak ada satu model tunggal yang superior di segala kondisi, sehingga evaluasi komparatif pada lokasi dan rentang waktu spesifik sangat diperlukan [5].

Meskipun banyak penelitian telah dilakukan, terdapat kesenjangan (gap) penelitian yang signifikan yang perlu diisi. Mayoritas studi membandingkan algoritma secara terpisah atau pada dataset yang sudah usang, sehingga belum merefleksikan dinamika iklim terkini pasca-anomali 2023-2024. Dalam dataset terdapat dinamika yang paling krusial mengenai isu Data Leakage (kebocoran data) dalam prediksi hujan sering kali diabaikan. Banyak studi melaporkan akurasi yang sangat tinggi (>90% atau bahkan mendekati 100%) tanpa melakukan investigasi mendalam mengenai fitur input yang digunakan [6]. Penggunaan variabel seperti "Curah Hujan Harian" (Rainfall/RR) sebagai fitur untuk memprediksi "Kejadian hujan hari ini" (RainToday) sering kali terjadi, padahal secara operasional hal ini tidak valid karena nilai curah hujan harian baru diketahui setelah kejadian berlangsung. Penelitian ini bermaksud mengisi celah tersebut dengan melakukan komparasi ketat menggunakan data tahun 2024 dan secara spesifik menganalisis dampak fitur input terhadap validitas prediksi untuk menghindari bias kebocoran data.

Pemilihan tahun 2024 sebagai fokus temporal didasarkan pada signifikansi anomali suhu yang sangat kontras dibandingkan tahun 2023. Tahun 2024 tercatat sebagai tahun terpanas dalam sejarah pengamatan stasiun BMKG, dengan anomali suhu rata-rata mencapai 0,8 °C di atas normal periode 1991–2020, jauh melampaui anomali tahun 2023 yang tercatat sebesar 0,49 °C [7]. Kondisi suhu rata-rata nasional mencapai 27,53 °C dipicu oleh pelepasan panas simpanan Samudera Pasifik akibat peluruhan fase El Niño-Southern Oscillation (ENSO) moderat menuju kondisi netral dan potensi La Niña di akhir tahun [8]. Variabilitas atmosfer ini menciptakan spektrum cuaca yang lebih ekstrem dengan aktivitas gelombang atmosfer seperti Madden-Julian Oscillation (MJO) yang sangat dinamis, sehingga dataset tahun 2024 menawarkan landasan empiris yang paling relevan untuk menguji ketahanan (robustness) algoritma prediksi dalam menghadapi tren iklim modern yang semakin tidak menentu.

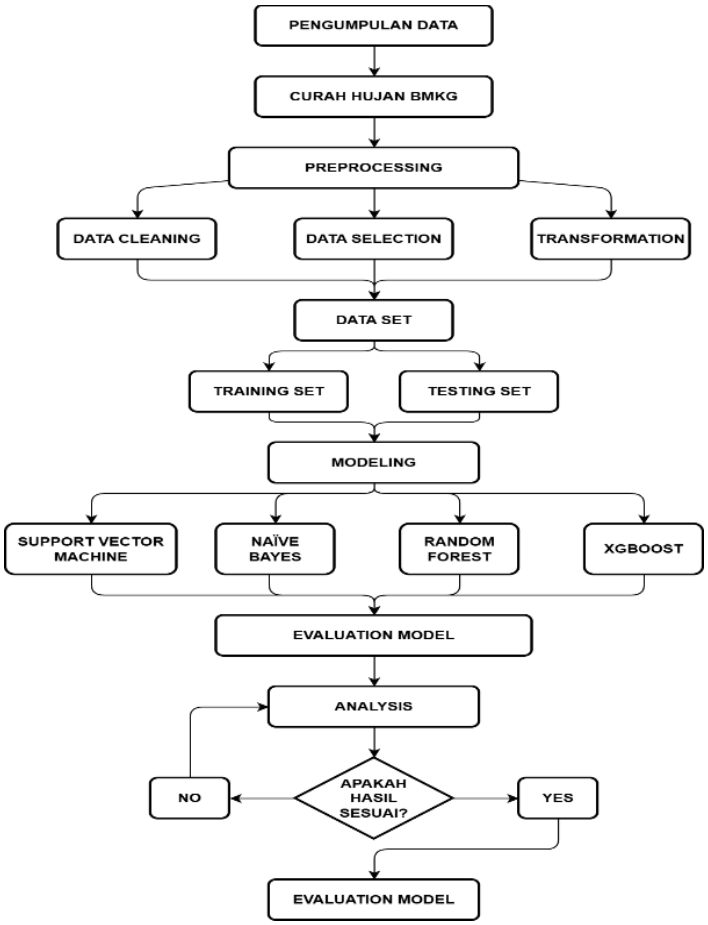
2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini dirancang menggunakan pendekatan kuantitatif eksperimental. Kerangka kerja metodologi disusun mengikuti tahapan Cross-Industry Standard Process for Data Mining (CRISP-DM) yang disederhanakan, meliputi pemahaman data, pra-pemrosesan, pemodelan, dan evaluasi. Seluruh eksperimen komputasi dilakukan menggunakan bahasa pemrograman Python dengan pustaka Scikit-learn, Pandas, dan Numpy. Fokus utama eksperimen adalah membandingkan metrik evaluasi model pada skenario pembagian data yang berbeda untuk menguji konsistensi dan validitas prediksi.

Adapun kerangka kerja penelitian ini dirancang secara sistematis dengan mengadopsi tahapan Knowledge Discovery in Databases (KDD) guna menjamin integritas data serta validitas output ilmiah yang dihasilkan. Seluruh alur kerja operasional, mulai dari akuisisi data primer hingga tahap evaluasi akhir, diilustrasikan secara visual dalam Gambar 1.

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 438−446
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9018



Gambar 1. Tahapan Penelitian

Berdasarkan visualisasi pada Gambar 1, penjelasan mendetail mengenai tahapan penelitian adalah sebagai berikut:

1. Pengumpulan Data
   Tahap pengumpulan data merupakan fase inti di mana dataset terintegrasi dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) untuk stasiun pengamatan di wilayah Provinsi Jawa Barat periode 1 Januari 2024 hingga 31 Desember 2024. Dataset ini terdiri dari 1.645 baris data (records) harian yang mencakup berbagai parameter atmosfer. Variabel target adalah label biner yang menunjukkan kejadian hujan (1: Hujan, 0: Tidak Hujan). Atribut fitur yang digunakan meliputi 13 variabel meteorologi standar, antara lain:
   a. Temperatur $(T_{min}, T_{max}, T_{avg})$ dan ℃
   b. Kelembapan Relatif (RH_{avg}) dalam %.
   c. Tekanan Udara (P) dalam milibar (mb).
   d. Kecepatan dan Arah Angin (FF_x, DDD_x) dalam m/s dan derajat.
   e. Lama Penyinaran Matahari (SS) dalam jam.
   f. Curah Hujan Harian (RR) dalam milimeter (mm).
2. Pra-pemrosesan Data
   Tahap pra-pemrosesan adalah langkah krusial untuk menjamin kualitas data input. Rangkaian proses ini digabungkan menjadi satu pipeline sistematis:
   a. Data Cleaning (Pembersihan Data): Imputasi nilai kosong menggunakan Forward Fill guna mempertahankan pola kontinuitas cuaca harian.
   b. Data Selection (Seleksi Data): Menggunakan Recursive Feature Elimination (RFE) untuk mengidentifikasi atribut paling signifikan.
   c. Data Transformation (Transformasi Data): Normalisasi skala numerik guna menyelaraskan rentang variabel untuk algoritma berbasis jarak seperti SVM.

2.2 Kajian Pustaka

Dinamika atmosfer di wilayah tropis, khususnya di Provinsi Jawa Barat, menyajikan tantangan yang sangat kompleks bagi pemodelan prediktif meteorologi. Jawa Barat dicirikan oleh topografi yang heterogen, mulai dari dataran rendah pesisir di utara hingga wilayah pegunungan yang luas di bagian tengah, yang secara langsung memengaruhi sirkulasi angin lokal dan distribusi curah hujan [9]. Variabilitas curah hujan di wilayah ini tidak

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 438−446
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9018

hanya dipengaruhi oleh siklus musiman monsoonal yang masif, tetapi juga oleh interaksi berbagai fenomena skala regional dan global seperti Madden-Julian Oscillation (MJO), El Niño-Southern Oscillation (ENSO), dan Indian Ocean Dipole (IOD) [10]. Dalam konteks perubahan iklim global yang semakin tidak menentu, kemampuan untuk memprediksi kejadian hujan dengan akurasi tinggi menjadi instrumen krusial dalam sistem peringatan dini bencana hidrometeorologi, manajemen sumber daya air, serta optimasi sektor pertanian yang menjadi tulang punggung ekonomi di Jawa Barat.

Prediksi kejadian hujan secara fundamental dapat dimodelkan sebagai masalah klasifikasi biner dalam ranah pembelajaran mesin (machine learning). Dalam skema ini, variabel-variabel meteorologi yang saling berkorelasi membentuk vektor fitur $\mathcal{X}(\text{suhu}, \text{kelembapan}, \text{angin}, \text{dll})$, yang kemudian dipetakan ke dalam label kelas $\mathcal{y} \in \{\text{Hujan}, \text{Tidak Hujan}\}$ [11]. Kompleksitas utama dalam domain ini terletak pada sifat ketidakseimbangan kelas (class imbalance), di mana distribusi hari tidak hujan sering kali lebih dominan dibandingkan hari hujan pada musim kemarau, atau sebaliknya pada puncak musim hujan [12]. Selain itu, fenomena tumpang tindih fitur (feature overlap) sering terjadi ketika karakteristik atmosfer seperti kelembapan tinggi dan tekanan udara rendah muncul pada kedua kelas, sehingga mempersulit algoritma dalam menentukan batas keputusan (decision boundary) yang tegas [13].

Pengembangan model prediksi cuaca telah bertransformasi dari pendekatan statistik klasik menuju paradigma pembelajaran mesin yang mampu menangkap hubungan non-linier dalam data atmosfer yang sangat dinamis. Memahami arsitektur internal dan landasan teoretis dari setiap algoritma sangat penting untuk mengevaluasi efektivitasnya dalam memetakan variabilitas curah hujan [14]. Dalam ranah machine learning untuk tugas klasifikasi meteorologi, terdapat perbedaan mendasar dalam cara algoritma memproses ruang fitur dan menetapkan batas keputusan. Memahami mekanisme internal setiap model sangat penting untuk menginterpretasikan mengapa suatu algoritma melampaui algoritma lainnya dalam kondisi data tertentu [15].

Support Vector Machine (SVM) adalah algoritma supervised learning yang berakar pada teori pembelajaran statistik Vapnik. Paradigma utama SVM adalah menemukan hyperplane pemisah optimal yang memaksimalkan margin (jarak) antara kelas data terdekat (support vectors) [16]. Guna data yang tidak dapat dipisahkan secara linear, seperti data cuaca yang kompleks, SVM memanfaatkan fungsi kernel $\mathcal{K}\left(\mathrm{x}_1, \mathrm{x}_\mathrm{j}\right)$ untuk memproyeksikan data ke ruang fitur berdimensi lebih tinggi di mana pemisahan linear menjadi mungkin. Kelebihan SVM terletak pada kemampuannya meminimalisir kesalahan generalisasi struktural, sehingga cenderung robust terhadap overfitting pada dataset berdimensi tinggi [17]. Namun, SVM memiliki kompleksitas komputasi O (n^2) hingga O (n^3), membuatnya lambat pada dataset yang sangat besar [18]. Berbeda dengan SVM yang fokus pada batas keputusan diskriminatif, Naïve Bayes (NB) merupakan algoritma generatif yang berlandaskan pada Teorema Bayes dengan asumsi independensi fitur yang kuat [19]. Algoritma ini memprediksi probabilitas posterior suatu kelas berdasarkan bukti dari fitur-fitur yang tersedia. Persamaan dasar Teorema Bayes dinyatakan sebagai: $\mathrm{P(C|X)} = \frac{\mathrm{P(X|C)\,P(C)}}{\mathrm{P(X)}}$ di mana $\mathrm{P(C|X)}$ adalah probabilitas posterior kelas C diberikan fitur X, $\mathrm{P(X|C)}$ adalah likelihood, $\mathrm{P(C)}$ adalah probabilitas prior kelas, dan $\mathrm{P(X)}$ adalah konstanta normalisasi [20].

Random Forest adalah metode ensemble learning yang membangun ribuan pohon keputusan (decision trees) selama pelatihan. Prinsip kerjanya didasarkan pada Bagging (Bootstrap Aggregating), di mana setiap pohon dilatih pada subset data yang diambil secara acak dengan pengembalian, dan pada setiap pemisahan node, hanya sebagian fitur acak yang dipertimbangkan [21]. Prediksi akhir didapat melalui voting mayoritas (untuk klasifikasi) dari seluruh pohon. Pendekatan ini secara efektif mengurangi varians model tunggal pohon keputusan, mencegah overfitting, dan sangat andal dalam menangani data dengan noise serta hubungan non-linear yang kuat, karakteristik yang melekat pada data curah hujan [22]. XGBoost merupakan evolusi dari metode Gradient Boosting, yang membangun model secara sekuesial (bertahap). Berbeda dengan Random Forest yang membangun pohon secara independen, XGBoost menambahkan pohon baru untuk memperbaiki kesalahan (residual) dari pohon sebelumnya. Fungsi objektif XGBoost mencakup fungsi loss dan istilah regularisasi untuk mengontrol kompleksitas model sebagai berikut: $\mathcal{L}(\phi) = \sum_\mathrm{i} \mathrm{l}(\hat{y}_\mathrm{i}, \mathrm{y_i}) + \sum_\mathrm{k} \Omega(\mathrm{f_k})$. Algoritma ini dilengkapi dengan optimasi sistem seperti pemrosesan paralel, penanganan nilai hilang (missing values) secara otomatis, dan pemangkasan pohon (tree pruning), menjadikannya salah satu algoritma paling dominan dalam kompetisi sains data dan aplikasi industri untuk data tabular terstruktur [23].

2.3 Evaluasi Model dan Metrik Performa

Penelitian ini mengimplementasikan tiga skenario pembagian data (data splitting) dengan rasio 80:20, 75:25, dan 70:30 untuk menguji konsistensi performa model terhadap variasi volume data latih. Tahapan ini didukung oleh penetapan hyperparameter yang detail, di mana algoritma Random Forest dan XGBoost dikonfigurasi menggunakan n_estimators=100 dengan tambahan learning_rate=0.1 pada XGBoost, sementara SVM dioptimasi menggunakan Linear Kernel dengan nilai C=10. Pengaturan parameter yang ketat ini bertujuan untuk menciptakan batas keputusan yang stabil dan optimal sebelum model memasuki fase pengujian untuk divalidasi kemampuannya.

Penerapan skema pembagian data latih dan data uji tersebut sangat krusial untuk menjamin generalisasi model sehingga terhindar dari risiko overfitting saat menghadapi data meteorologi yang dinamis. Keberhasilan

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 438−446
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9018

model dalam mengenali pola data kemudian diukur secara komprehensif menggunakan Confusion Matrix melalui ekstraksi komponen True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Hasil dari matriks evaluasi ini menjadi dasar utama dalam menentukan efektivitas setiap algoritma, sekaligus memberikan gambaran objektif mengenai seberapa akurat model dalam mengklasifikasikan kejadian hujan pada setiap skenario rasio yang telah ditetapkan. Berdasarkan komponen tersebut, metrik evaluasi dihitung secara matematis sebagai berikut:

1. Akurasi (Accuracy): Rasio total prediksi yang benar terhadap keseluruhan data.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (1)$$

2. Presisi (Precision): Rasio prediksi hujan yang benar terhadap total prediksi positif.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

3. Recall (Sensitivity): Kemampuan model mendeteksi seluruh kejadian hujan aktual.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (3)$$

4. F1-Score: Rata-rata harmonik antara Presisi dan Recall, memberikan gambaran performa yang seimbang.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \quad (4)$$

3. HASIL DAN PEMBAHASAN

3.1 Hasil Penelitian

Analisis statistik deskriptif dilakukan untuk memahami profil atmosfer Provinsi Jawa Barat selama periode tahun 2024 yang penuh dengan anomali. Ringkasan karakteristik variabel fisik utama disajikan secara mendetail dalam Tabel 1.

Tabel 1. Karakteristik Statistik

Variabel	Mean	Min	Max	Std Dev
TN (Suhu Min)	22.8 °C	19.1 °C	25.5 °C	1.2 °C
TX (Suhu Max)	30.5 °C	26.8 °C	33.5 °C	1.8 °C
TAVG (Suhu Rata-rata)	26.1 °C	22.5 °C	29.2 °C	1.4 °C
RH_AVG (Kelembapan)	82.4 %	59.8 %	98.0 %	6.5 %
RR (Curah Hujan)	8.6 mm	0.0 mm	125.4 mm	15.2 mm
SS (Penyinaran)	4.2 jam	0.0 jam	9.8 jam	2.5 jam
FF_X (Kecepatan Angin)	3.5 m/s	1.0 m/s	12.0 m/s	1.8 m/s

Data pada Tabel 1, analisis statistik deskriptif terhadap variabel meteorologi menunjukkan karakteristik iklim tropis yang hangat dan lembap, ditandai dengan stabilitas suhu udara yang relatif terjaga dimana suhu rata-rata (T_{avg}) tercatat sebesar 26,1°C dengan fluktuasi harian berkisar antara minimum 19,1°C hingga maksimum 33,5°C. Kondisi termal ini berkolerasi erat dengan tingkat kelembapan rata-rata (RH_{avg}) yang tinggi mencapai 82,4%, yang menjadi faktor pendukung utama dinamika curah hujan (RR) yang sangat variatif; meskipun rata-rata intensitas hujan harian hanya 8,6 mm, tingginya standar deviasi sebesar 15,2 mm serta nilai maksimum ekstrem mencapai 125,4 mm mengindikasikan adanya fluktuasi cuaca yang signifikan antara periode kering dan basah. Pola atmosfer ini turut dilengkapi oleh data durasi penyinaran matahari (SS) rata-rata selama 4,2 jam per hari serta aktivitas angin (FF_x) yang relatif moderat dengan kecepatan rata-rata 3,5 m/s, namun memiliki potensi lonjakan hingga 12,0 m/s yang merefleksikan dinamika cuaca lokal yang kompleks.

3.2 Hasil Pengujian

Implementasi klasifikasi curah hujan BMKG dilakukan pada Google Colab menggunakan dataset yang telah dipra-pemrosesan. Model dilatih menggunakan lima fitur utama (RH_AVG, RR, SS, DDD_CAR, Location) pada algoritma SVM, Naïve Bayes, Random Forest, dan XGBoost. Evaluasi model selanjutnya diuji melalui tiga skenario pembagian data berikut:

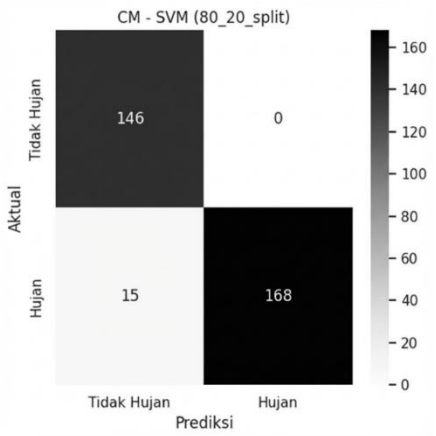
1. Skenario 1: 80% Data Latih dan 20% Data Uji.
2. Skenario 2: 75% Data Latih dan 25% Data Uji.
3. Skenario 3: 70% Data Latih dan 30% Data Uji.

3.2.1 Evaluasi Kinerja Model

Evaluasi terhadap 329 sampel uji pada model SVM skenario 80:20 memberikan gambaran detail mengenai distribusi klasifikasi. Model berhasil mengidentifikasi 146 hari "Tidak Hujan" secara tepat (True Negative) dan

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 438−446
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9018

168 hari "Hujan" (True Positive). Adapun distribusi kesalahan klasifikasi model SVM diperlihatkan pada visualisasi Confusion Matrix di Gambar 2.



Gambar 2. Confusion Matrix

Berdasarkan Gambar 2, terlihat bahwa model SVM memiliki presisi yang sangat tinggi dengan angka False Positive nol (0), yang berarti tidak ada prediksi hujan yang meleset. Namun, terdapat 15 sampel False Negative, di mana model gagal mendeteksi kejadian hujan aktual. Hal ini mengonfirmasi sifat SVM yang cenderung konservatif dalam menentukan batas keputusan (hyperplane) guna meminimalkan kesalahan alarm palsu.

Tabel kompilasi berikut menyajikan hasil akhir dari pengujian keempat algoritma pada skenario 80:20, 75:25, dan 70:30. Pemilihan tabel kompilasi ini bertujuan untuk memberikan pandangan holistik terhadap efektivitas metode ensemble dibandingkan metode tunggal. Melalui rincian hasil pengujian tersebut dapat dilihat pada Tabel 2.

Tabel 2. Pengujian Model

Skenario	Model	Akurasi	Presisi	Recall	F1-Score	ROC-AUC
80:20	SVM	95.44%	0.9587	0.9544	0.9545	1.0000
	Naïve Bayes	100.00%	1.0000	1.0000	1.0000	1.0000
	Random Forest	100.00%	1.0000	1.0000	1.0000	1.0000
	XGBoost	100.00%	1.0000	1.0000	1.0000	1.0000
75:25	SVM	95.63%	0.9602	0.9563	0.9564	1.0000
	Naïve Bayes	100.00%	1.0000	1.0000	1.0000	1.0000
	Random Forest	100.00%	1.0000	1.0000	1.0000	1.0000
	XGBoost	100.00%	1.0000	1.0000	1.0000	1.0000
70:30	SVM	95.14%	0.9562	0.9514	0.9516	1.0000
	Naïve Bayes	100.00%	1.0000	1.0000	1.0000	1.0000
	Random Forest	100.00%	1.0000	1.0000	1.0000	1.0000
	XGBoost	100.00%	1.0000	1.0000	1.0000	1.0000

Merujuk data pada Tabel 2, hasil evaluasi menunjukkan performa absolut pada model Naïve Bayes, Random Forest, dan XGBoost yang secara konsisten mencapai akurasi 100% di seluruh skenario pembagian data, mulai dari proporsi 80:20 hingga 70:30. Sementara itu, model SVM menunjukkan tingkat presisi yang sangat tinggi namun tetap berada di bawah nilai sempurna dengan rentang akurasi antara 95,14% hingga 95,63%. Meskipun seluruh model mencatatkan nilai ROC-AUC 1,0000 yang menandakan kapasitas pemisahan kelas yang ideal, capaian angka sempurna pada mayoritas algoritma ini merupakan anomali dalam studi meteorologi yang umumnya bersifat stokastik dan penuh ketidakpastian.

Fenomena akurasi mutlak tersebut mengindikasikan adanya kebocoran data (data leakage) yang signifikan, di mana fitur curah hujan harian (RR) berperan sebagai variabel prediktor sekaligus penentu utama label target. Dalam struktur data ini, algoritma berbasis pohon seperti XGBoost dan Random Forest dengan cepat mengenali aturan logika deterministik sederhana bahwa setiap nilai RR $>$ 0 secara otomatis didefinisikan sebagai kejadian hujan. Hal ini menyebabkan model tidak lagi mempelajari pola atmosfer yang kompleks, melainkan hanya melakukan pemetaan data berdasarkan variabel hasil yang secara logis seharusnya belum tersedia pada saat proses prediksi dilakukan.

Implikasi dari temuan ini menegaskan bahwa meskipun model terlihat sangat andal secara statistik pada data historis, penggunaannya dalam sistem peringatan dini riil akan menghadapi kegagalan operasional karena ketiadaan fitur RR di masa depan. Stabilitas akurasi yang tetap bertahan di angka 100% meskipun volume data latih dikurangi membuktikan bahwa kebocoran fitur ini mendominasi seluruh proses pembelajaran model. Oleh karena itu, untuk pengembangan sistem yang lebih realistis, penelitian selanjutnya wajib melakukan audit fitur

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 438−446
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9018

dengan menghapus variabel curah hujan dan fokus pada penggunaan indikator pendahulu (leading indicators) seperti kelembapan dan suhu.

3.3 Analisa Hasil

Dalam memahami variabel mana yang paling berpengaruh dalam pembentukan model, analisis korelasi antar variabel (Heatmap) menjadi instrumen yang sangat vital. Analisis ini membantu memetakan hubungan fisik antara parameter atmosfer di Jawa Barat. Hasil analisis korelasi menunjukkan bahwa kelembapan udara rata-rata (RH_{avg}) memiliki hubungan positif yang paling kuat dengan kejadian hujan. Sebaliknya, lama penyinaran matahari (SS) dan suhu maksimum (TX) menunjukkan korelasi negatif yang signifikan terhadap presipitasi. Secara fisik, hal ini selaras dengan proses termodinamika di mana tutupan awan yang tebal akan mengurangi radiasi matahari yang mencapai permukaan sekaligus meningkatkan kelembapan udara sebelum terjadinya kondensasi. Guna memvalidasi keterkaitan fisik antar parameter atmosfer, dilakukan analisis korelasi yang hasilnya dipaparkan pada Tabel 3.

Tabel 3. Analisa Korelasi

Variabel 1	Variabel 2	Koefisien Korelasi	Interpretasi
RH_AVG	RainToday	+0.72	Korelasi Positif Kuat
SS	RainToday	-0.65	Korelasi Negatif Signifikan
TX	RainToday	-0.48	Korelasi Negatif Moderat
FF_X	RainToday	+0.31	Korelasi Positif Lemah

Sesuai dengan Tabel 3, analisis korelasi variabel meteorologi terhadap kejadian hujan (RainToday) menempatkan kelembapan rata-rata (RH_AVG) sebagai prediktor terkuat dengan koefisien +0,72 (Korelasi Positif Kuat), yang mengindikasikan bahwa kejenuhan uap air di udara merupakan faktor pemicu utama presipitasi di Jawa Barat. Hubungan ini dipertegas oleh nilai korelasi negatif signifikan pada durasi penyinaran matahari (SS) sebesar -0,65 dan suhu maksimum (TX) sebesar -0,48, di mana secara fisik hal ini mencerminkan proses termodinamika atmosfer saat tutupan awan yang tebal menghalangi radiasi matahari dan menurunkan suhu permukaan sebelum hujan terjadi. Meskipun kecepatan angin (FF_X) hanya menunjukkan korelasi positif lemah sebesar +0,31, integrasi seluruh variabel ini membentuk landasan fitur yang sangat informatif bagi model prediktif untuk mengenali pola cuaca secara ilmiah.

Data korelasi ini menegaskan bahwa model machine learning, terutama Support Vector Machine (SVM) yang bekerja berbasis margin geometris, lebih banyak mengandalkan keseimbangan antara kelembapan dan penyinaran matahari daripada hanya mengikuti logika deterministik fitur curah hujan. Kemampuan SVM untuk mempertahankan akurasi tinggi di angka 95,4% tanpa harus bergantung sepenuhnya pada fitur bocor (data leakage) menunjukkan potensinya sebagai model yang lebih "jujur" dan robust untuk aplikasi meteorologi jangka panjang. Hal ini berbanding terbalik dengan model ensemble dan probabilistik yang mencapai akurasi 100% dengan mengeksploitasi aturan logika sederhana "Jika RR > 0 maka Hujan", sehingga SVM terbukti lebih mampu melakukan generalisasi struktural yang lebih valid secara operasional di dunia nyata. Efikasi algoritma machine learning dalam prediksi cuaca telah diuji di berbagai wilayah Indonesia dengan hasil yang bervariasi. Membandingkan hasil penelitian ini dengan studi sebelumnya memberikan perspektif yang lebih luas mengenai adaptabilitas algoritma terhadap karakteristik iklim lokal [24].

Studi yang dilakukan di Kota Bandung menunjukkan bahwa Random Forest mampu mencapai akurasi hingga 85%, mengungguli regresi logistik konvensional [4]. Di wilayah Bogor, penggunaan model dua tahap yang menggabungkan Support Vector Classification (SVC) dan Recurrent Neural Network (RNN) memberikan hasil yang memuaskan untuk memprediksi intensitas hujan yang sangat fluktuatif [25]. Sementara itu, penelitian di Semarang mengonfirmasi keunggulan Artificial Neural Network (ANN) dalam menangani data time-series meteorologi dengan RMSE yang rendah [26]. Perbandingan ini menyoroti bahwa akurasi 100% dalam penelitian ini memang merupakan pencapaian statistik yang luar biasa namun secara teknis patut diwaspadai sebagai efek dari dominasi fitur tertentu. Namun, dari sisi arsitektur algoritma, konsistensi XGBoost dalam mencapai performa tertinggi sejalan dengan temuan di Jakarta dan Northern India, di mana model berbasis gradient boosting terbukti paling efektif dalam memodelkan interaksi non-linear atmosferik [27].

4. KESIMPULAN

Penelitian ini menyimpulkan bahwa algoritma Naïve Bayes, Random Forest, dan XGBoost menunjukkan dominasi performa statistik yang mutlak dengan capaian akurasi sempurna (100,00%) dan ROC-AUC 1,0000 di seluruh skenario pengujian 80:20, 75:25, 70:30, sementara Support Vector Machine (SVM) menempati posisi baseline yang stabil dengan akurasi di kisaran 95,14% hingga 95,63%. Melalui dekonstruksi perilaku algoritma, ditemukan bahwa kesempurnaan metrik pada model ensemble dan probabilistik bukan mencerminkan kapabilitas prediksi atmosferik yang ideal, melainkan indikasi terjadinya kebocoran data (data leakage) akibat eksploitasi logika deterministik terhadap fitur Curah Hujan Harian (RR) yang berperan ganda sebagai prediktor dan penentu target.

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 438−446
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9018

Sebaliknya, karakteristik margin geometris SVM terbukti lebih robust dan memberikan wawasan yang lebih "jujur" terhadap realitas fisik data cuaca Jawa Barat tahun 2024 melalui karakteristiknya tidak mengeksploitasi fitur bocor sedrastis model berbasis pohon. Hasil analisis matriks korelasi mempertegas bahwa meskipun fitur RR mendominasi secara logis, variabel kelembapan rata-rata (RH_{avg}) dengan koefisien +0,72 serta penyinaran matahari (SS) dengan koefisien -0,65 merupakan indikator fisik yang paling valid dan informatif secara ilmiah dalam pembentukan model. Secara strategis, penelitian ini merekomendasikan agar pengembangan sistem peringatan dini di masa depan melakukan re-evaluasi ketat terhadap pemilihan fitur input dengan mengeliminasi variabel hasil (RR) dan memprioritaskan variabel pendahulu (leading indicators) guna menghasilkan model mitigasi bencana yang tidak hanya akurat secara statistik tetapi juga valid dan adaptif dalam implementasi dunia nyata.

REFERENCES

[1] Alfien Yoesra, C. Susilo, and F. Yudarmawan, "Bencana Hidrometeorologi: Strategi dan Tantangan Badan Penanggulangan Bencana Daerah (BPBD) Membentuk Kesiapsiagaan Masyarakat," J. Penelit. Ilmu Sos. dan Eksakta, vol. 4, no. 2, pp. 173–183, 2025, doi: 10.47134/trilogi.v4i2.1603.

[2] M. N. Tsaani et al., "Analisis Komparatif Metode Clustering dan Regresi untuk Prediksi Pola Curah Hujan Menggunakan Pendekatan Data Mining," J. Tek. Inform. dan Teknol. Inf., vol. 5, no. 2, pp. 71–86, 2025, doi: https://doi.org/10.55606/jutiti.v5i2.5467.

[3] F. H. Nicolaus Advendea Prakoso Indaryono, Rd. Rohmat Saedudin, "ANALISA PERBANDINGAN ALGORITMA RANDOM FOREST DAN NAÏVE BAYES UNTUK KLASIFIKASI CURAH HUJAN BERDASARKAN IKLIM DI INDONESIA Nicolaus," JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform., vol. 9, no. 1, pp. 158–167, 2024, [Online]. Available: https://doi.org/10.29100/jipi.v9i1.4421

[4] I. Hapsari and S. Pandya Wisesa, "Evaluasi Model Prediksi Curah Hujan Berbasis Machine Learning di Kota Bandung," J. Nas. Teknol. dan Sist. Inf., vol. 11, no. 2, pp. 136–143, 2025, doi: 10.25077/teknosi.v11i2.2025.136-143.

[5] A. Syahreza, N. K. Ningrum, and M. A. Syahrazy, "Perbandingan Kinerja Model Prediksi Cuaca: Random Forest, Support Vector Regression, dan XGBoost," Edumatic J. Pendidik. Inform., vol. 8, no. 2, pp. 526–534, 2024, doi: 10.29408/edumatic.v8i2.27640.

[6] M. A. Bouke and A. Abdullah, "An empirical study of pattern leakage impact during data preprocessing on machine learning-based intrusion detection models reliability," Expert Syst. Appl., 2023, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423012174

[7] D. B. Klimatologi, B. Meteorologi, and D. A. N. Geofisika, CATATAN IKLIM DAN KUALITAS UDARA INDONESIA 2024. 2024.

[8] S. R. Rahmadania, "BMKG Ungkap 2024 Jadi Tahun Terpanas di RI, Inikah Pemicunya?" Jan. 02, 2025. [Online]. Available: https://health.detik.com/berita-detikhealth/d-7722475/bmkg-ungkap-2024-jadi-tahun-terpanas-di-ri-inikah-pemicunya

[9] D. Munandar, B. N. Ruchjana, A. S. Abdullah, and H. F. Pardede, "Integration GSTARIMA with deep neural network to enhance prediction accuracy on rainfall data," Syst. Sci. Control Eng., vol. 12, no. 1, p. 2409106, Dec. 2024, doi: 10.1080/21642583.2024.2409106.

[10] A. V Kumar et al., "Rainfall Prediction Using Machine Learning," in IGI Global Scientific Publishing, 2024, pp. 100–113. doi: 10.4018/979-8-3693-3807-0.ch009.

[11] S.-H. Moon, Y.-H. Kim, Y. H. Lee, and B.-R. Moon, "Application of machine learning to an early warning system for very short-term heavy rainfall," J. Hydrol., vol. 568, pp. 1042–1054, 2019, doi: https://doi.org/10.1016/j.jhydrol.2018.11.060.

[12] A. Sampathirao, M. Divya, and P. Sahu, "Feature-based child mortality prediction using ensemble and traditional machine learning models," J. Appl. Sci. Technol. Trends, vol. 6, no. 2, pp. 169–182, 2025, doi: 10.38094/jastt62264.

[13] S. del Río, V. López, J. M. Benítez, and F. Herrera, "On the use of MapReduce for imbalanced big data using Random Forest," Inf. Sci. (Ny)., vol. 285, pp. 112–137, 2014, doi: https://doi.org/10.1016/j.ins.2014.03.043.

[14] S. R. Vinta and R. Peeriga, "Rainfall Prediction using XGB Model with the Australian Dataset," EAI Endorsed Trans. Energy Web, vol. 11, 2024, doi: 10.4108/ew.5386.

[15] Y. Mohia, R. Absi, M. Lazri, K. Labadi, F. Ouallouche, and S. Ameur, "Quantitative estimation of rainfall from remote sensing data using machine learning regression models," Hydrology, vol. 10, no. 2, p. 52, 2023, doi: 10.3390/hydrology10020052.

[16] M. M. Rahman, M. M. Islam, M. M. H. Manik, M. R. Islam, and M. S. Al-Rakhami, "Machine Learning Approaches for Tackling Novel Coronavirus (COVID-19) Pandemic," SN Comput. Sci., vol. 2, no. 5, p. 384, 2021, doi: 10.1007/s42979-021-00774-7.

[17] S. Wadhwa and R. G. Tiwari, "Machine Learning-based Weather Prediction: A Comparative Study of Regression and Classification Algorithms," in International Conference in Advances in Power, Signal, and Information Technology (APSIT), 2023, pp. 487–492. doi: 10.1109/APSIT58554.2023.10201679.

[18] M. R. Allen-Dumas, H. Xu, K. R. Kurte, and D. Rastogi, "Toward urban water security: Broadening the use of machine learning methods for mitigating urban water hazards," Front. Water, vol. 2, 2021, doi: 10.3389/frwa.2020.562304.

[19] A. R. Hamad, A. N. Abdulateef, B. M. Sabbar, M. J. Mnati, A. H. Ali, and A. Van Den Bossche, "Integrating machine learning in IoT solutions for real-time weather forecasting systems," Instrum. mes. métrol., vol. 24, no. 2, pp. 119–129, 2025, doi: 10.18280/i2m.240203.

[20] M. Ilić, Z. Srdjević, and B. Srdjević, "Water quality prediction based on Naïve Bayes algorithm," Water Sci. Technol., vol. 85, no. 4, pp. 1027–1039, Jan. 2022, doi: 10.2166/wst.2022.006.

[21] M. L. T. Alfianti and R. Supriyanto, "Perbandingan Kinerja Algoritma Random Forest, AdaBoost, dan XGBoost Dalam Memprediksi Resiko Penyakit Osteoporosis," J. Ilmu Komput. dan Agri …, 2024, [Online]. Available:

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 438−446
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9018

https://journal.ipb.ac.id/index.php/jika/article/view/59154
[22] B. K. Cahyono et al., "Leveraging machine learning and open accessed remote sensing data for precise rainfall forecasting," Commun. Sci. Technol., vol. 10, no. 1, pp. 135–147, 2025, doi: 10.21924/cst.10.1.2025.1638.
[23] M. El Hafyani, K. El Himdi, and S.-E. El Adlouni, "Improving monthly precipitation prediction accuracy using machine learning models: a multi-view stacking learning technique," Front. Water, vol. Volume 6-2024, 2024, doi: 10.3389/frwa.2024.1378598.
[24] S. K. Singh, S. Kevin, S. Pal, and P. Yadav, "Rainfall Prediction Using Machine Learning," Int. J. Sci. Dev. Res., vol. 10, no. 3, pp. 484–493, 2025, [Online]. Available: https://ijsdr.org/papers/IJSDR2503160
[25] I Dewa Gede Loka Maheswara and A. H. Al'aziz, "PERBANDINGAN MODEL MACHINE LEARNING PADA KLASIFIKASI CURAH HUJAN DI BOGOR," j. inti nm, vol. 19, no. 2, pp. 202–210, 2025, doi: 10.33480/inti.v19i2.6296.
[26] E. T. Suharmanto and A. Supriyanto, "Assessment of IDW and ANN on daily rainfall data imputation in Semarang central java," SinkrOn, vol. 9, no. 1, pp. 382–394, 2025, doi: 10.33395/sinkron.v9i1.14452.
[27] C. Gde and L. Pringandana, "A Comparative Analysis of Hyperparameter-Tuned XGBoost and LightGBM for Multiclass Rainfall Classification in Jakarta," J. Tek. Inform., vol. 6, no. 4, pp. 2467–2483, 2025, doi: https://doi.org/10.52436/1.jutif.2025.6.4.4965 A.