Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9475

Perbandingan Performa Multi-Algoritma Machine Learning dengan Dua Strategi Validasi pada Klasifikasi Curah Hujan

I Dewa Gede Loka Maheswara[1*], Arya Zaki Ramadhan[1], Rica Azzura Maldina[1], Muhammad Fany Nurwibowo[1], Yosik Norman[2]
[1]Program Studi Meteorologi, Sekolah Tinggi Meteorologi Klimatologi dan Geofisika, Kota Tangerang, Indonesia
[2]Stasiun Klimatologi Banten, Kota Tangerang Selatan, Indonesia
Email: [1*]maheswaradewaloka@gmail.com, [2]aryazakiramadhan@gmail.com, [3]ricamaldina@gmail.com, [4]mfanynurwibowo62@gmail.com, [5]yosikbrebes@gmail.com
Email Penulis Korespondensi: maheswaradewaloka@gmail.com[*]
Submitted: **04/01/2026**; Accepted:**28/02/2026**; Published: **31/03/2026**

Abstrak−Prediksi curah hujan yang akurat masih menjadi tantangan karena kompleksitas proses atmosfer serta dampaknya terhadap berbagai sektor. Performa algoritma machine learning dalam klasifikasi curah hujan sangat dipengaruhi oleh karakteristik data dan metode validasi, sehingga diperlukan evaluasi komparatif untuk menentukan model yang paling sesuai pada konteks lokal. Penelitian ini bertujuan membandingkan performa lima model machine learning, yaitu Random Forest, XGBoost, Support Vector Machine, K-Nearest Neighbor, dan Decision Tree dalam klasifikasi curah hujan di Kabupaten Tapanuli Tengah menggunakan data observasi harian periode 2015–2024 sebanyak 32.796 data yang diperoleh dari Stasiun Meteorologi FL Tobing. Evaluasi dilakukan melalui skema pembagian data dan *10-cross fold validation* dengan metrik *precision*, *recall*, dan *f1-score*. Hasil penelitian menunjukkan bahwa Random Forest secara konsisten memberikan performa terbaik pada kedua skema validasi dengan *f1-score* sebesar 62% dan 63%, lebih stabil dibandingkan model lainnya pada kondisi distribusi kelas yang tidak seimbang. Temuan ini menunjukkan bahwa pendekatan *ensemble* lebih adaptif dalam menangkap hubungan nonlinier parameter meteorologi serta memberikan dasar metodologis dalam pemilihan model klasifikasi curah hujan untuk mendukung mitigasi bencana hidrometeorologi.

Kata Kunci: Perbandingan Model; Klasifikasi Hujan; Machine Learning; Random Forest; Mitigasi Bencana

Abstract−Accurate rainfall prediction remains a major challenge due to the complexity of atmospheric processes and their significant impacts across various sectors. The performance of machine learning algorithms in rainfall classification is strongly influenced by data characteristics and validation strategies, therefore a comparative evaluation is necessary to determine the most appropriate model within a local context. This study aims to compare the performance of five machine learning models, such as Random Forest, XGBoost, Support Vector Machine, K-Nearest Neighbor, and Decision Tree for rainfall classification in Central Tapanuli Regency using 32,796 daily observational records from 2015 to 2024 obtained from FL Tobing Meteorological Station. Model evaluation was conducted using data splitting scheme and 10-fold cross-validation with precision, recall, and F1-score as performance metrics. The results indicate that Random Forest consistently achieved the best performance under both validation schemes attaining F1-scores of 62% and 63%. It also demonstrated greater stability compared to the other models under imbalanced class distribution conditions. These findings suggest that ensemble based approaches are more adaptive in capturing nonlinear relationships among meteorological parameters and provide a methodological basis for selecting rainfall classification models to support hydrometeorological disaster mitigation efforts.

Keywords: Model Comparison; Rainfall Classification; Machine Learning; Random Forest; Disaster Mitigation

1. PENDAHULUAN

Bencana hidrometeorologi, khususnya tanah longsor dan banjir, merupakan dampak langsung dari curah hujan berintensitas tinggi yang berlangsung dalam durasi tertentu [1], [2], [3]. Curah hujan yang terjadi secara terus-menerus dengan intensitas tinggi dapat meningkatkan kejenuhan tanah, memperbesar limpasan permukaan, serta menurunkan stabilitas lereng, sehingga meningkatkan potensi terjadinya bencana di wilayah rawan. Kabupaten Tapanuli Tengah menjadi salah satu daerah dengan tingkat curah hujan tahunan yang tinggi di Indonesia, dengan rata-rata mencapai 4.760,10 mm per tahun [4]. Kondisi klimatologis tersebut menunjukkan bahwa wilayah ini memiliki kerentanan yang cukup tinggi terhadap kejadian hidrometeorologi ekstrem. Oleh karena itu, pemahaman yang komprehensif terhadap karakteristik, variabilitas, dan pola curah hujan menjadi aspek penting dalam mendukung perencanaan mitigasi risiko bencana yang lebih efektif dan berbasis data.

Prediksi curah hujan merupakan permasalahan yang kompleks karena melibatkan interaksi nonlinier antarparameter atmosfer yang saling berkaitan [5], [6], [7]. Proses pembentukan hujan dipengaruhi oleh berbagai faktor meteorologi seperti suhu udara, kelembapan, angin, dan radiasi matahari, yang berinteraksi dalam berbagai skala waktu dan ruang. Kompleksitas ini menyebabkan pendekatan statistik konvensional sering kali mengalami keterbatasan dalam menangkap dinamika atmosfer secara menyeluruh. Dalam beberapa tahun terakhir, pendekatan machine learning banyak digunakan untuk memodelkan hubungan nonlinier tersebut serta meningkatkan akurasi prediksi berbasis data historis [8], [9], [10]. Keunggulan utama machine learning terletak pada kemampuannya mempelajari pola implisit dalam data tanpa harus mendefinisikan hubungan fisis secara eksplisit.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9475

Berbagai algoritma machine learning, seperti Random Forest, XGBoost, Support Vector Machine, K-Nearest Neighbor, dan Decision Tree telah diterapkan dalam studi klasifikasi maupun prediksi curah hujan [11], [12], [13]. Setiap algoritma memiliki karakteristik dan mekanisme pembelajaran yang berbeda, sehingga performanya sangat bergantung pada sifat data yang digunakan. Selain pemilihan algoritma, skema validasi juga berperan penting dalam menilai kestabilan dan kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya [14], [15], [16]. Skema validasi yang berbeda dapat menghasilkan evaluasi performa yang bervariasi, sehingga pemilihan strategi evaluasi menjadi faktor krusial dalam analisis kinerja model machine learning, terutama pada aplikasi operasional dan pengambilan keputusan berbasis risiko.

Namun demikian, sebagian besar penelitian sebelumnya masih menggunakan rentang data yang relatif pendek, terbatas pada satu atau dua algoritma, serta menerapkan satu skema validasi tanpa menguji konsistensi performa terhadap metode evaluasi yang berbeda [17], [18]. Kondisi ini menyebabkan hasil penelitian cenderung spesifik terhadap konfigurasi tertentu dan sulit digeneralisasikan. Selain itu, kajian komparatif multi-algoritma dengan skema validasi ganda pada wilayah tropis bercurah hujan tinggi masih terbatas. Padahal, wilayah tropis memiliki karakteristik curah hujan yang sangat variatif dan sering kali didominasi oleh kejadian hujan lebat hingga ekstrem dengan distribusi kelas yang tidak seimbang.

Berdasarkan kesenjangan tersebut, penelitian ini memanfaatkan data observasi curah hujan dan parameter meteorologi selama sepuluh tahun untuk menangkap variabilitas curah hujan yang lebih representatif. Penggunaan data jangka menengah diharapkan mampu menggambarkan pola curah hujan yang lebih stabil dibandingkan penelitian yang menggunakan data dalam rentang waktu lebih pendek. Pendekatan klasifikasi yang digunakan adalah *supervised multi-class classification*, dengan mengelompokkan curah hujan ke dalam enam kategori intensitas berdasarkan peraturan Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) untuk curah hujan 24 jam. Pengelompokan ini mencerminkan kebutuhan operasional dalam menyajikan informasi cuaca dalam bentuk kategori yang mudah dipahami dan relevan untuk pengambilan keputusan.

Pendekatan klasifikasi dipilih karena informasi intensitas hujan dalam bentuk kategori lebih representatif untuk keperluan operasional dan mitigasi bencana dibandingkan nilai curah hujan kontinu semata. Klasifikasi memungkinkan penekanan pada kejadian hujan berdampak tinggi, seperti hujan lebat hingga ekstrem, yang meskipun jarang terjadi namun memiliki konsekuensi yang besar. Selain itu, pendekatan ini juga lebih sesuai untuk menangani kondisi distribusi data yang tidak seimbang, di mana jumlah kejadian hujan ekstrem jauh lebih sedikit dibandingkan hujan ringan atau sedang.

Penelitian ini bertujuan mengidentifikasi model machine learning yang paling stabil dalam mengklasifikasikan curah hujan di Tapanuli Tengah serta mengevaluasi pengaruh perbedaan skema validasi terhadap kinerja model. Untuk itu, dilakukan studi komparatif terhadap lima model klasifikasi, yaitu Random Forest, XGBoost, Support Vector Machine, K-Nearest Neighbor, dan Decision Tree dengan menggunakan dua skema validasi, yaitu pembagian data dan *10-Cross Fold Validation*. Kontribusi penelitian ini terletak pada evaluasi komparatif *multi-algoritma* dengan validasi ganda pada dataset observasi yang lebih banyak di wilayah tropis. Hasil penelitian ini diharapkan dapat memberikan dasar metodologis yang lebih kuat dalam pemilihan model klasifikasi curah hujan yang stabil dan robust, serta mendukung pengembangan sistem mitigasi bencana hidrometeorologi yang lebih efektif dan berbasis data.

2. METODOLOGI PENELITIAN

2.1 Wilayah Studi dan Data

Wilayah yang digunakan pada penelitian ini adalah Kabupaten Tapanuli Tengah yang terletak pada koordinat 1,60° LU – 2,30° LU dan 98,50° BT – 99,50° BT. Wilayah ini termasuk daerah dengan karakteristik curah hujan tinggi dan variabilitas temporal yang signifikan, sehingga relevan untuk kajian klasifikasi curah hujan berbasis data observasi jangka panjang. Penelitian ini memanfaatkan data harian parameter meteorologi yang meliputi suhu minimum, suhu rata-rata, suhu maksimum, kelembapan rata-rata, durasi penyinaran matahari, arah kecepatan angin maksimum, kecepatan angin rata-rata, serta curah hujan. Data tersebut mencakup rentang waktu 1 Januari 2015 hingga 31 Desember 2024 selama sepuluh tahun dan diperoleh dari Stasiun Meteorologi FL Tobing melalui portal resmi Badan Meteorologi, Klimatologi, dan Geofisika (BMKG), yang dapat diakses melalui laman https://dataonline.bmkg.go.id/ menggunakan akun terdaftar. Ringkasan variabel data yang digunakan ditampilkan pada Tabel 1.

Tabel 1. Parameter Meteorologi

Parameter	Tipe Data	Satuan	Valid Records	Missing Values
Suhu Minimum (Tn)	float64	°C	3488	156
Suhu Maximum (Tx)	float64	°C	3560	84

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9475

Suhu Rata-rata (Tavg)	float64	°C	3635	9
Kelembaban Rata-rata (RH_avg)	float64	%	3636	8
Lama Penyinaran Matahari (ss)	float64	Jam	3539	105
Kecepatan Angin Rata-rata (ff_avg)	float64	m/s	3643	1
Kecepatan Angin Maximum (ff_x)	float64	m/s	3643	1
Arah Angin (ddd_x)	float64	°	3643	1
Curah Hujan (RR)	float64	mm	3590	54

2.2 Tahap Prapemrosesan Data

Tahap prapemrosesan data dilakukan untuk memastikan kualitas dan kesiapan data sebelum digunakan dalam proses pemodelan. Tahapan ini dimulai dengan pengumpulan data parameter meteorologi seperti yang terlihat pada Tabel 1. Selanjutnya, dilakukan transformasi tipe data menjadi tipe data numerik untuk seluruh dataset agar berada dalam format yang sesuai untuk proses komputasi. Kemudian, dilakukan penanganan *missing value* pada dataset. Nilai kosong pada setiap kolom parameter meteorologi diisi menggunakan nilai rata-rata dari kolom terkait, merujuk pada penelitian sebelumnya yang dilakukan oleh [15] dan [16]. Pendekatan ini dipilih karena proporsi data hilang relatif kecil dibandingkan total dataset, serta bertujuan untuk mempertahankan jumlah data agar tetap representatif dan konsisten antar model yang diuji. Selain itu, *missing value* yang tidak ditangani dapat menurunkan akurasi dan stabilitas hasil klasifikasi [19]. Pada Tabel 2 ditampilkan contoh *missing value* pada kolom parameter meteorologi yang ditandai dengan simbol "-", kemudian dilakukan pengisian menggunakan nilai rata-rata dari masing-masing kolom.

Tabel 2. Ilustrasi Nilai Hilang (*Missing Value)* pada Parameter Meteorologi

Tanggal	(Tn)	(Tx)	(Tavg)	(RH_avg)	(ss)	(ff_avg)	(ff_x)	(ddd_x)	(RR)
01-01-2015	-	32.6	26.6	78	5	0	0	0	1.6
02-01-2015	23	-	26.2	87	3.2	0	-	90	0
03-01-2015	22	31.6	26.5	-	6.1	-	5	-	5
04-01-2015	-	31.8	26.4	83	4.1	1	3	-	0

Selanjutnya, data curah hujan dikelompokkan ke dalam enam kelas berdasarkan peraturan Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) terkait probabilistik curah hujan 24 jam. Kategori tersebut terdiri atas tidak hujan dengan curah hujan 0 mm yang diberi label "1", hujan ringan dengan curah hujan lebih dari 0 mm hingga 20 mm sebagai label "2", hujan sedang dengan rentang 20–50 mm sebagai label "3", hujan lebat dengan rentang 50–100 mm sebagai label "4", hujan sangat lebat dengan rentang 100–150 mm sebagai label "5", serta hujan ekstrem dengan curah hujan lebih dari 150 mm yang diberi label "6". Klasifikasi ini digunakan agar hasil penelitian tetap relevan secara operasional, meskipun distribusi kelas berpotensi tidak seimbang.

Tahap selanjutnya adalah normalisasi data menggunakan metode Min–Max Scaling. Metode ini mentransformasikan nilai pada setiap kolom ke dalam rentang 0 hingga 1, dengan tujuan membatasi skala data agar berada dalam interval yang seragam sehingga mempermudah proses prediksi, sebagaimana merujuk pada penelitian [15] dan [18]. Secara matematis, proses normalisasi ini dinyatakan pada Persamaan (1). Contoh data sebelum dan sesudah normalisasi masing-masing disajikan pada Tabel 3 dan Tabel 4.

$$X_{SC} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

Tabel 3. Ilustrasi Data Sebelum Tahap Normalisasi Data

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9475

Tanggal	(Tn)	(Tx)	(Tavg)	(RH_avg)	(ss)	(ff_avg)	(ff_x)	(ddd_x)	(RR)
27-12-2024	22.2	32.8	27.3	83.0	5.7	1.0	6.0	230.0	0.0
28-12-2024	23.3	33.4	28.2	81.0	6.1	0.0	4.0	260.0	0.0
29-12-2024	22.3	33.4	27.4	83.0	6.3	1.0	5.0	240.0	1.0
30-12-2024	23.4	33.0	27.6	86.0	7.3	1.0	3.0	270.0	0.0

Tabel 4. Ilustrasi Data Setelah Tahap Normalisasi Data

Tanggal	(Tn)	(Tx)	(Tavg)	(RH_avg)	(ss)	(ff_avg)	(ff_x)	(ddd_x)	(RR)
27-12-2024	0.604651	0.671756	0.533333	0.578947	0.491379	0.333333	0.400000	0.638889	0.000000
28-12-2024	0.732558	0.717557	0.633333	0.526316	0.525862	0.000000	0.266667	0.722222	0.000000
29-12-2024	0.616279	0.717557	0.544444	0.578947	0.543103	0.333333	0.333333	0.666667	0.004125
30-12-2024	0.744186	0.687023	0.566667	0.657895	0.629310	0.333333	0.200000	0.750000	0.000000

2.3 Tahap Pemrosesan Data

Setelah dilakukan tahap prapemrosesan data, dilakukan tahap pemrosesan data dengan menggunakan lima model klasifikasi, diantaranya Random Forest, XGBoost, Support Vector Machine, K-Nearest Neighbor, dan Decision Tree. Pemilihan kelima algoritma tersebut bertujuan untuk merepresentasikan berbagai pendekatan klasifikasi, mulai dari metode berbasis pohon keputusan, *distance-based*, hingga *margin-based*, sehingga memungkinkan evaluasi performa model yang lebih komprehensif terhadap karakteristik data curah hujan.

Random Forest digunakan karena keunggulannya dalam menangani dataset berukuran besar dan data tidak seimbang melalui mekanisme *ensemble* [20]. XGBoost dipilih karena efisiensi komputasionalnya serta kemampuannya dalam meningkatkan kinerja prediksi melalui teknik *boosting* [21], dan dapat memberikan kinerja yang baik pada dataset yang tidak seimbang. Support Vector Machine dipilih karena kemampuannya dalam memodelkan hubungan nonlinier yang kompleks antara variabel *input* dan *output* [22], [10]. K-Nearest Neighbor digunakan karena kesederhanaannya serta kemampuannya dalam menangani batas keputusan yang tidak beraturan. Algoritma Decision Tree dipilih karena kemampuannya yang fleksibel dalam merepresentasikan pola serta hubungan antarvariabel pada data [23]. Variabel bebas yang digunakan dalam pemodelan meliputi suhu maksimum, suhu minimum, suhu rata-rata, kelembapan rata-rata, lamanya penyinaran matahari, kecepatan angin maksimum, kecepatan angin rata-rata, dan arah kecepatan angin maksimum. Variabel terikat pada penelitian ini adalah kategori curah hujan hasil klasifikasi. Untuk mengevaluasi kinerja model secara komprehensif digunakan dua skema validasi data, yaitu skema pembagian data latih dan data uji dengan rasio tertentu serta metode *10-Fold Cross Validation*. Penggunaan dua skema validasi ini bertujuan untuk mengevaluasi sensitivitas dan kestabilan performa model terhadap perbedaan skema evaluasi data, sehingga model yang dihasilkan tidak hanya unggul secara numerik tetapi juga *robust* terhadap variasi metode validasi.

2.4 Tahap Pascapemrosesan Data

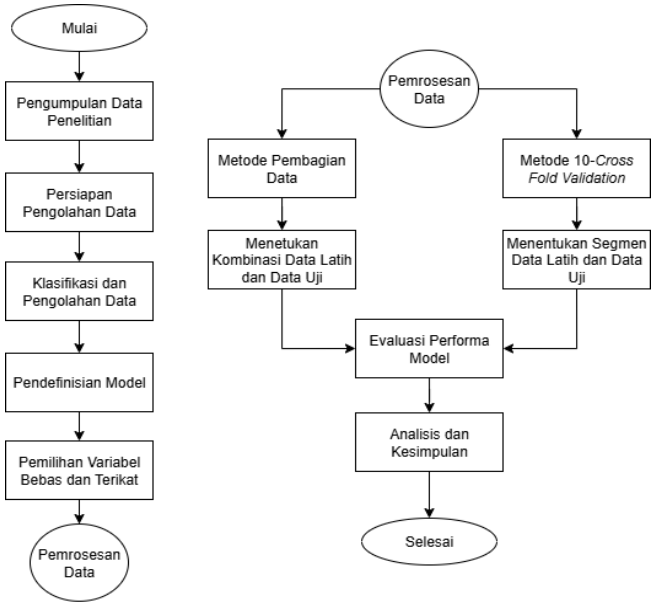
Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9475

Setelah dilakukan tahap pemrosesan data, dilakukan tahap pascapemrosesan data dengan mengevaluasi kinerja kelima model klasifikasi menggunakan metrik *precision*, *recall*, dan *f1-score* untuk setiap skema validasi data. Pemilihan metrik ini dilakukan karena lebih representatif dalam mengevaluasi performa klasifikasi multi-kelas, khususnya pada kondisi distribusi kelas yang tidak seimbang. Nilai *precision*, *recall*, dan *f1-score* secara matematis dapat ditentukan menggunakan persamaan (2), (3), dan (4). Hasil evaluasi kinerja model selanjutnya disajikan dalam bentuk tabel untuk memudahkan perbandingan antar model klasifikasi dan skema validasi. Seluruh proses pengolahan dan analisis data dilakukan menggunakan bahasa pemrograman Python pada lingkungan komputasi berbasis *cloud* untuk memastikan konsistensi, efisiensi, dan reprodusibilitas penelitian [24].

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{3}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$



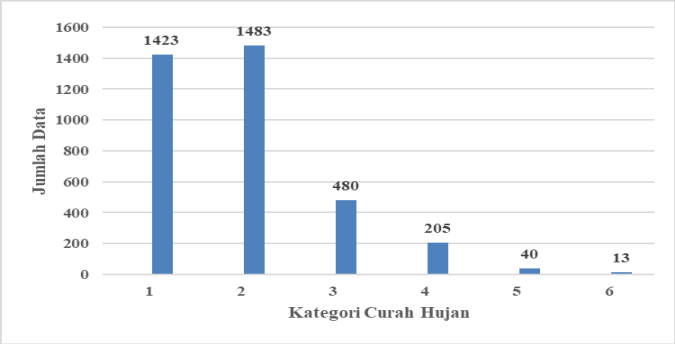
Gambar 1. Diagram Alir Penelitian

3. HASIL DAN PEMBAHASAN

3.1 Distribusi dan Karakteristik Data

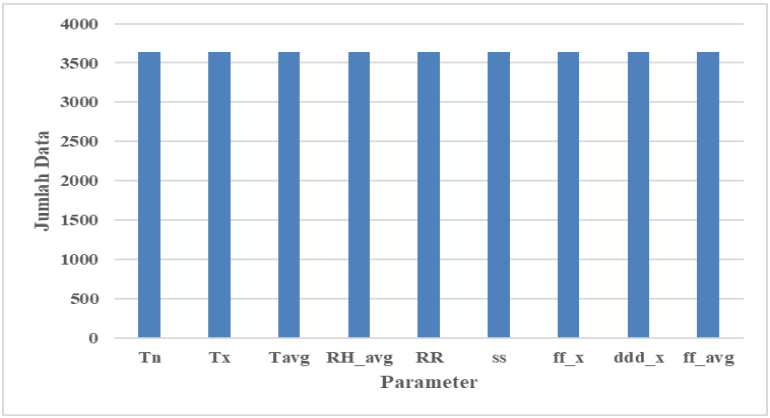
Hasil klasifikasi curah hujan berdasarkan peraturan Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) untuk curah hujan 24 jam ditunjukkan pada Gambar 2. Distribusi data menunjukkan bahwa kategori tidak hujan dan hujan ringan mendominasi dataset, masing-masing dengan 1.423 dan 1.483 data. Sementara itu, jumlah data pada kategori hujan sedang, hujan lebat, hujan sangat lebat, dan hujan ekstrem relatif lebih sedikit, dengan jumlah berturut-turut 480, 205, 40, dan 13 data. Distribusi tersebut mengindikasikan adanya ketidakseimbangan kelas (*class imbalance*) yang cukup signifikan, khususnya pada kategori hujan intensitas tinggi. Kondisi ini merupakan karakteristik umum pada data curah hujan observasi dan berpotensi memengaruhi performa model klasifikasi, terutama dalam mendeteksi kejadian hujan lebat hingga ekstrem. Model cenderung lebih teroptimasi pada kelas mayoritas sehingga berisiko menghasilkan false negative pada kelas ekstrem yang justru memiliki dampak kebencanaan tinggi. Oleh karena itu, pada penelitian ini evaluasi performa model difokuskan pada metrik *precision*, *recall*, dan *f1-score* yang representatif untuk permasalahan klasifikasi multi-kelas dengan distribusi data tidak seimbang serta relevan untuk konteks mitigasi bencana.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9475



Gambar 2. Histogram Klasifikasi Curah Hujan

Pada penelitian ini digunakan 3.644 baris data untuk setiap parameter meteorologi. Tabel 1 menunjukkan jumlah data awal pada masing-masing parameter berbeda-beda akibat adanya data hilang pada beberapa parameter. Setelah melalui tahap prapemrosesan, seluruh parameter diselaraskan sehingga memiliki jumlah observasi yang sama, sebagaimana ditunjukkan pada Gambar 3. Penyelarasan ini menyebabkan total keseluruhan data menjadi 32.796, meningkat sebesar 1,28% dibandingkan data awal. Peningkatan tersebut terjadi akibat proses pengisian *missing value* untuk menjaga konsistensi panjang data antarparameter, sehingga seluruh variabel dapat diproses secara seragam oleh model klasifikasi. Proses prapemrosesan ini merupakan tahap krusial karena ketidakselarasan jumlah data dapat menyebabkan bias pada proses pelatihan model serta mengurangi reliabilitas evaluasi kinerja. Dengan data yang telah tersinkronisasi, setiap observasi merepresentasikan kondisi atmosfer yang utuh pada waktu yang sama, sehingga hubungan antarparameter meteorologi dapat dipelajari secara lebih akurat oleh algoritma machine learning yang digunakan.



Gambar 3. Jumlah Data pada Setiap Parameter Meteorologi

3.2 Kinerja Model dengan Skema Rasio Pembagian Data

Performa kelima model pada skema rasio pembagian data dievaluasi menggunakan metrik *precision*, *recall*, dan *f1-score*. Hasil evaluasi menunjukkan adanya perbedaan kinerja model pada setiap rasio pembagian data latih dan data uji yang digunakan. Perbedaan ini mencerminkan sensitivitas masing-masing algoritma terhadap proporsi data latih, terutama dalam mempelajari pola dari kelas minoritas yang jumlahnya terbatas. Pada rasio data latih yang lebih kecil, beberapa model cenderung mengalami penurunan *recall* pada kelas hujan berintensitas tinggi, yang mengindikasikan keterbatasan model dalam mengenali kejadian ekstrem. Sebaliknya, peningkatan porsi data latih umumnya memperbaiki stabilitas performa, meskipun tidak selalu diikuti oleh peningkatan *f1-score* yang signifikan. Hal ini menunjukkan adanya *trade-off* antara kompleksitas model, ukuran data latih, dan kemampuan generalisasi pada data uji.

Tabel 5. Hasil Kinerja Model Random Forest

Rasio Data Latih – Rasio Data Uji	Precision	Recall	F1-Score

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9475

10-90	0.59424	0.62480	0.59498
20-80	0.60322	0.63786	0.60875
30-70	0.60784	0.63505	0.60492
40-60	0.61723	0.64380	0.61387
50-50	0.62111	0.64618	0.61470
60-40	0.62355	0.64060	0.60957
70-30	0.63450	0.65082	0.62086
80-20	0.61344	0.64426	0.61437
90-10	0.58228	0.62740	0.59537

Tabel 5 menunjukkan model Random Forest memiliki kinerja terbaik ketika menggunakan rasio data latih 70% dan data uji 30%, dengan nilai *f1-score* mencapai 62%. Nilai ini lebih tinggi dibandingkan rasio pembagian data lainnya, yang mengindikasikan bahwa Random Forest memerlukan proporsi data latih yang cukup besar untuk menangkap pola kompleks pada data meteorologi. Selain itu, rasio ini memberikan keseimbangan yang lebih baik antara proses pembelajaran dan evaluasi, sehingga mengurangi risiko overfitting maupun underfitting pada data uji.

Tabel 6. Hasil Kinerja Model XGBoost

Rasio Data Latih – Rasio Data Uji	Precision	Recall	F1-Score
10-90	0.55698	0.58364	0.56744
20-80	0.57036	0.59648	0.57948
30-70	0.56972	0.59506	0.57870
40-60	0.57826	0.60646	0.58800
50-50	0.57642	0.60538	0.58552
60-40	0.58179	0.60768	0.58815
70-30	0.58411	0.61365	0.59413
80-20	0.58232	0.61271	0.59105
90-10	0.56213	0.60274	0.57776

Berdasarkan Tabel 6, serupa dengan model Random Forest, model XGBoost mencapai performa terbaik pada rasio data latih 70% dan data uji 30% dengan *f1-score* sebesar 59,4%. Hal ini menunjukkan bahwa model berbasis *boosting* cenderung membutuhkan jumlah data latih yang memadai untuk meningkatkan kualitas pembelajaran secara iteratif. Dengan data latih yang cukup, XGBoost mampu memperbaiki kesalahan pada iterasi sebelumnya dan membangun model yang lebih adaptif terhadap pola nonlinier dalam data curah hujan.

Tabel 7. Hasil Kinerja Model Support Vector Machine

Rasio Data Latih – Rasio Data Uji	Precision	Recall	F1-Score
10-90	0.57893	0.50244	0.52781
20-80	0.58686	0.48091	0.51038
30-70	0.58884	0.47341	0.50505
40-60	0.58166	0.45603	0.48684

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9475

50-50	0.57948	0.44823	0.48176
60-40	0.59660	0.45496	0.49145
70-30	0.60374	0.45521	0.49211
80-20	0.59524	0.45039	0.48863
90-10	0.57667	0.44201	0.47859

Berbeda dengan model Random Forest dan XGBoost, Support Vector Machine (Tabel 7) menunjukkan performa terbaik pada rasio data latih 10% dan data uji 90%, dengan *f1-score* sebesar 52,8%. Temuan ini mengindikasikan bahwa performa Support Vector Machine sangat sensitif terhadap komposisi data latih, di mana penggunaan data latih yang terlalu besar tidak selalu menghasilkan generalisasi yang lebih baik pada data uji. Kondisi ini diduga berkaitan dengan keterbatasan kernel dalam memisahkan kelas pada data yang tidak seimbang serta meningkatnya kompleksitas batas keputusan.

Tabel 8. Hasil Kinerja Model K-Nearest Neighbor

Rasio Data Latih – Rasio Data Uji	Precision	Recall	F1-Score
10-90	0.54607	0.58283	0.55965
20-80	0.54844	0.58562	0.56415
30-70	0.55131	0.58735	0.56687
40-60	0.55302	0.59259	0.57054
50-50	0.56926	0.59385	0.57026
60-40	0.55506	0.59602	0.57236
70-30	0.55410	0.59781	0.57347
80-20	0.55564	0.59762	0.57467
90-10	0.56314	0.60639	0.58180

Model K-Nearest Neighbor (Tabel 8) menunjukkan performa terbaik ketika menggunakan rasio data latih 90% dan data uji 10%, dengan *f1-score* sebesar 58,2%. Hasil ini sesuai dengan karakteristik K-Nearest Neighbor sebagai metode berbasis jarak yang memerlukan banyak contoh data latih untuk menentukan kedekatan antar data pengamatan secara akurat. Sementara itu, Decision Tree (Tabel 9) mencapai performa terbaik pada rasio data latih dan data uji yang seimbang, yaitu 50%:50%, dengan *f1-score* sebesar 54,2%. Kondisi ini mengindikasikan bahwa Decision Tree relatif sensitif terhadap variasi data latih dan berpotensi mengalami overfitting pada proporsi data latih yang terlalu besar.

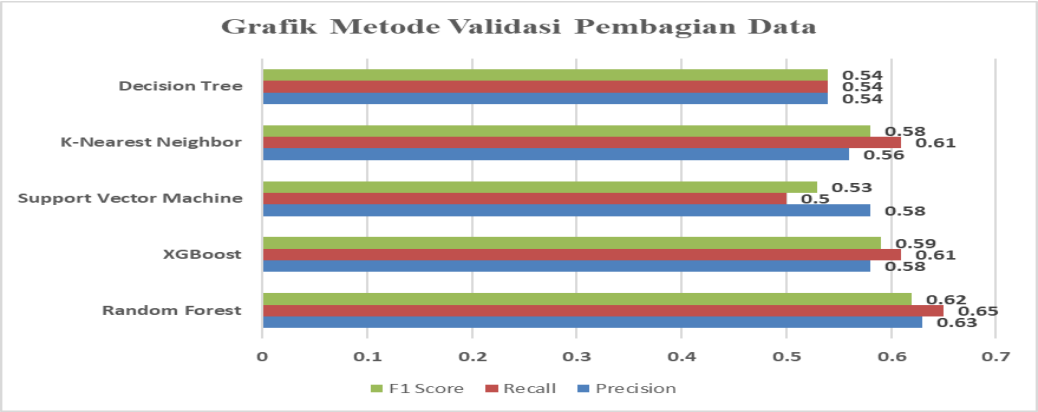
Tabel 9. Hasil Kinerja Model Decision Tree

Rasio Data Latih – Rasio Data Uji	Precision	Recall	F1-Score
10-90	0.51470	0.51006	0.51186
20-80	0.52382	0.52172	0.52267
30-70	0.53323	0.52933	0.53112
40-60	0.53305	0.52675	0.52969
50-50	0.54496	0.53970	0.54220
60-40	0.53690	0.53109	0.53375
70-30	0.52906	0.52163	0.52511

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9475

80-20	0.54003	0.53864	0.53914
90-10	0.53211	0.52603	0.52843

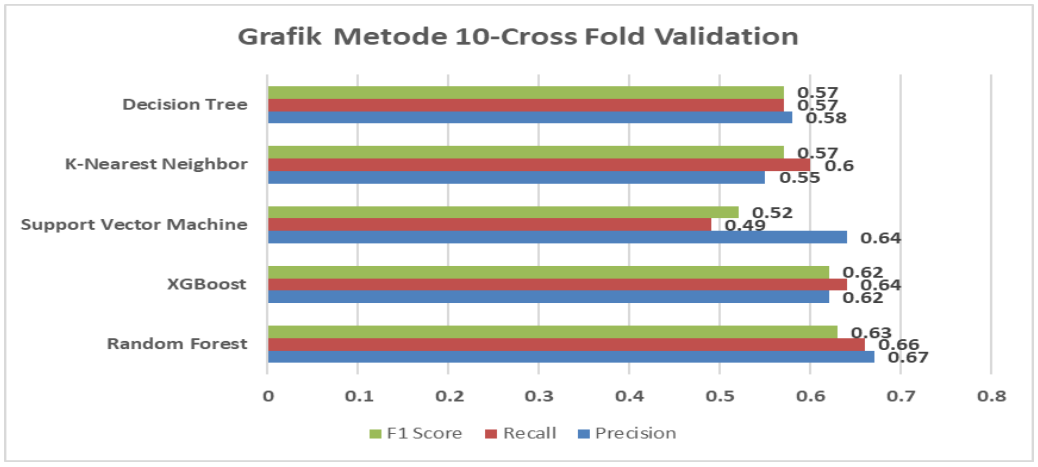
Perbandingan nilai *precision*, *recall*, dan *f1-score* dari performa terbaik masing-masing model ditampilkan pada Gambar 4.



Gambar 4. Perbandingan Performa Model Random Forest, XGBoost, Support Vector Machine, K-Nearest Neighbor, dan Decision Tree dengan Metode Validasi Pembagian Data

Secara keseluruhan, Random Forest menunjukkan performa paling baik dibandingkan model lainnya pada skema validasi pembagian data yang dibuktikan dengan nilai *f1-score* mencapai 62% ketika Random Forest menggunakan pembagian data dengan rasio data latih 70% dan rasio data uji 30%. Keunggulan ini menunjukkan kemampuan Random Forest dalam menyeimbangkan precision dan recall pada dataset dengan distribusi kelas yang tidak seimbang, khususnya pada kategori hujan berintensitas tinggi. Selain itu, mekanisme ensemble yang mengombinasikan banyak pohon keputusan memungkinkan model ini mereduksi varians dan meningkatkan stabilitas prediksi dibandingkan model tunggal. Hasil ini juga mengindikasikan bahwa Random Forest lebih adaptif terhadap kompleksitas data meteorologi tropis yang bersifat nonlinier.

3.3 Performa Model dengan Metode *10-Fold Cross Validation*



Gambar 5. Perbandingan Performa Model Random Forest, XGBoost, Support Vector Machine, K-Nearest Neighbor, dan Decision Tree dengan Metode Validasi *10-Cross Fold Validation*

Evaluasi performa model menggunakan metode *10-Fold Cross Validation* ditampilkan pada Gambar 5. Berbeda dengan metode pembagian data tunggal, pendekatan ini memungkinkan seluruh data digunakan secara bergantian sebagai data latih dan data uji, sehingga menghasilkan estimasi performa yang lebih stabil dan mengurangi bias akibat pemilihan sampel tertentu. Hasil menunjukkan bahwa Random Forest kembali menghasilkan performa terbaik dibandingkan empat model lainnya, dengan nilai *precision* sebesar 67%, *recall* sebesar 66%, dan *f1-score* sebesar 63%. Keunggulan ini mengindikasikan Random Forest tidak hanya unggul pada

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9475

skema pembagian data dengan rasio tertentu, tetapi juga konsisten ketika diuji menggunakan skema validasi *10-Fold Cross Validation*.

Model-model lainnya menunjukkan performa yang relatif lebih rendah, dengan nilai *f1-score* berada pada kisaran 52% hingga 62%. Support Vector Machine dan Decision Tree cenderung menunjukkan fluktuasi performa antar-*fold*, yang mengindikasikan sensitivitas terhadap variasi data latih, terutama pada kelas hujan berintensitas tinggi yang jumlah datanya terbatas. Sementara itu, K-Nearest Neighbor dan XGBoost menunjukkan performa yang lebih stabil dibandingkan SVM dan Decision Tree, namun masih mengalami penurunan recall pada kelas hujan ekstrem. Perbedaan hasil antara metode pembagian data dan 10-Fold Cross Validation menegaskan bahwa strategi validasi data memiliki pengaruh signifikan terhadap evaluasi performa model klasifikasi curah hujan, terutama pada dataset dengan ketidakseimbangan kelas yang tinggi dan variabilitas atmosfer yang kompleks.

3.4 Diskusi Perbandingan dan Implikasi

Secara keseluruhan, Random Forest menunjukkan kinerja yang paling konsisten dan *robust* pada kedua skema validasi yang digunakan. Keunggulan Random Forest dalam penelitian ini berkaitan dengan kemampuannya dalam menangani hubungan nonlinier antarparameter meteorologi serta mereduksi varians model melalui mekanisme *ensemble* dan *bagging*. Dengan jumlah variabel input yang relatif banyak, Random Forest mampu mengekstraksi pola yang lebih stabil dibandingkan model klasifikasi tunggal sebagaimana juga dilaporkan pada penelitian sebelumnya [25]. Selain itu, proses pemilihan fitur secara acak pada setiap pohon keputusan membuat model ini lebih tahan terhadap noise dan fluktuasi data, yang umum dijumpai pada data curah hujan tropis. Kombinasi tersebut menjadikan Random Forest lebih andal dalam menghasilkan prediksi yang konsisten pada berbagai skema validasi.

Sebaliknya, model seperti Support Vector Machine menunjukkan sensitivitas terhadap distribusi data latih dan ketidakseimbangan kelas. Hal ini dapat dijelaskan karena pendekatan berbasis *hyperplane* pada Support Vector Machine cenderung mengalami kesulitan ketika ruang fitur antar kelas saling tumpah tindih (*overlapping*), yang umum terjadi pada klasifikasi curah hujan tropis [26]. Pada kondisi tersebut, batas keputusan menjadi kurang tegas, sehingga meningkatkan kemungkinan salah klasifikasi terutama pada kelas dengan jumlah sampel lebih sedikit. Selain itu, pemilihan kernel dan parameter regularisasi yang kurang optimal juga dapat memperbesar penurunan performa pada data yang tidak seimbang. Sementara itu, model K-Nearest Neighbor membutuhkan proporsi data latih yang besar untuk membentuk representasi jarak yang memadai [27], karena kepadatan data yang rendah pada kelas minoritas dapat menyebabkan kesalahan identifikasi tetangga terdekat. Model Decision Tree cenderung rentan terhadap *overfitting* ketika dihadapkan pada distribusi kelas yang tidak seimbang [28], karena pembentukan aturan keputusan lebih didominasi oleh kelas mayoritas.

Analisis kesalahan menunjukkan bahwa sebagian besar mis-klasifikasi terjadi pada kelas hujan ekstrem. Model cenderung mengklasifikasikan kejadian hujan ekstrem sebagai hujan sedang atau lebat, yang mengindikasikan adanya keterbatasan dalam membedakan pola fitur antara kedua kategori tersebut. Kondisi ini terutama dipengaruhi oleh distribusi kelas yang tidak seimbang, di mana jumlah kejadian hujan ekstrem jauh lebih sedikit dibandingkan kelas lainnya. Ketidakseimbangan ini menyebabkan model lebih dominan mempelajari karakteristik kelas mayoritas. Selain itu, sifat hujan konvektif di wilayah tropis yang sporadis, berdurasi singkat, serta dipengaruhi dinamika mesoskal dan vertikal atmosfer yang tidak sepenuhnya tercermin pada parameter observasi permukaan turut memperbesar tingkat kesalahan prediksi pada kelas ekstrem. Keterbatasan resolusi temporal data harian juga berkontribusi, karena intensitas hujan ekstrem sering terjadi dalam interval waktu yang singkat sehingga sinyalnya tereduksi pada agregasi harian.

Nilai *f1-score* yang berada pada kisaran 52% hingga 63% menunjukkan bahwa kemampuan model dalam menyeimbangkan *precision* dan *recall* masih tergolong moderat. F1-score moderat pada kelas kejadian hujan ekstrem mengindikasikan bahwa sebagian kejadian ekstrem masih berpotensi tidak terdeteksi (*false negative*). Dalam sistem peringatan dini, kesalahan jenis ini lebih kritis dibandingkan *false positive* karena dapat menyebabkan keterlambatan respons terhadap potensi banjir atau tanah longsor. Dengan demikian, meskipun model menunjukkan performa relatif baik secara komparatif, tingkat akurasi tersebut belum sepenuhnya ideal untuk aplikasi operasional tanpa dukungan sistem pendukung lainnya. Oleh karena itu, hasil klasifikasi sebaiknya diintegrasikan dengan informasi meteorologis tambahan, seperti analisis sinoptik atau produk radar cuaca, agar pengambilan keputusan dapat dilakukan secara lebih komprehensif dan andal.

Temuan ini juga menegaskan bahwa tantangan utama dalam klasifikasi curah hujan berbasis data observasi jangka menengah di wilayah tropis yang memiliki variabilitas tinggi dan distribusi kelas yang tidak seimbang. Selain itu, keterbatasan panjang data observasi dibandingkan dengan penelitian sebelumnya oleh [18], [16] yang menggunakan data hingga 30 tahun atau lebih serta adanya ketidakseimbangan kelas pada kategori hujan ekstrem turut memengaruhi kemampuan model dalam melakukan prediksi secara akurat. Variabilitas antar-musim dan antar-tahun yang kuat pada wilayah tropis menyebabkan pola hujan sulit direpresentasikan secara konsisten

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9475

hanya dari data jangka menengah. Kondisi ini memperkuat kebutuhan akan data yang lebih panjang dan beragam untuk meningkatkan kemampuan generalisasi model.

Meskipun memiliki keterbatasan, hasil penelitian ini menunjukkan bahwa pemilihan algoritma dan skema validasi data memiliki peran signifikan dalam evaluasi performa klasifikasi curah hujan. Model dengan performa numerik tertinggi belum tentu merupakan model yang paling stabil, sehingga pendekatan evaluasi yang mempertimbangkan *robustnes* model menjadi aspek penting dalam penerapan machine learning untuk prediksi curah hujan operasional. Evaluasi yang komprehensif memungkinkan identifikasi model yang tidak hanya unggul pada satu skenario pengujian, tetapi juga konsisten pada berbagai konfigurasi data. Pendekatan ini penting untuk memastikan bahwa model yang diimplementasikan memiliki reliabilitas yang memadai dalam mendukung pengambilan keputusan terkait mitigasi bencana hidrometeorologi.

4. KESIMPULAN

Hasil penelitian menegaskan bahwa perbedaan performa setiap model tidak hanya disebabkan oleh kemampuan komputasional model, tetapi juga oleh karakteristik data hujan tropis yang kompleks, nonlinier, dan tidak seimbang. Kompleksitas ini muncul akibat keterlibatan berbagai proses atmosfer, mulai dari skala lokal hingga regional, yang berinteraksi secara dinamis dan sulit direpresentasikan secara utuh hanya melalui parameter observasi permukaan. Ketidakpastian tersebut menyebabkan hubungan antara variabel prediktor dan kelas intensitas hujan menjadi tidak konsisten, sehingga membatasi kemampuan model dalam membentuk pola klasifikasi yang tegas, khususnya pada kejadian hujan berintensitas tinggi. Oleh karena itu, peningkatan performa model ke depan dapat diarahkan pada beberapa strategi. Penerapan teknik penyeimbangan kelas, seperti *oversampling*, *undersampling*, atau *Synthetic Minority Over-sampling Technique* (SMOTE), berpotensi meningkatkan kemampuan model dalam mempelajari karakteristik kelas hujan ekstrem yang jumlah datanya terbatas. Dengan distribusi kelas yang lebih seimbang, model diharapkan mampu mengurangi bias terhadap kelas mayoritas dan meningkatkan recall pada kelas minoritas yang memiliki dampak paling besar dalam konteks mitigasi bencana. Selain itu, penambahan variabel prediktor yang merepresentasikan kondisi atmosfer lapisan atas, seperti indeks kestabilan atmosfer, kelembapan vertikal, atau parameter dinamika angin, dapat memperkaya informasi yang relevan dengan proses pembentukan hujan ekstrem. Variabel-variabel tersebut berpotensi menangkap sinyal fisis yang tidak sepenuhnya tercermin pada data permukaan, sehingga membantu model membedakan lebih baik antara hujan lebat dan hujan ekstrem. Integrasi data dari radar cuaca atau satelit juga dapat menjadi alternatif untuk meningkatkan resolusi spasial dan temporal informasi hujan. Eksplorasi pendekatan *ensemble hybrid* yang menggabungkan kekuatan beberapa model sekaligus juga menjadi arah pengembangan yang menjanjikan. Pendekatan ini memungkinkan pemanfaatan keunggulan masing-masing algoritma dalam menangkap pola tertentu, sehingga menghasilkan prediksi yang lebih stabil dan adaptif terhadap variasi data. Dengan strategi-strategi tersebut, penelitian lanjutan diharapkan dapat meningkatkan kinerja klasifikasi curah hujan secara signifikan serta memperkuat peran machine learning sebagai alat pendukung yang andal dalam sistem mitigasi bencana hidrometeorologi di wilayah tropis.

REFERENCES

[1] M. T. Chaudhary dan A. Piracha, “Natural Disasters—Origins, Impacts, Management,” *Encyclopedia*, vol. 1, no. 4, hlm. 1101–1131, Des 2021, doi: 10.3390/encyclopedia1040084.

[2] J. A. Marengo *dkk.*, “Flash floods and landslides in the city of Recife, Northeast Brazil after heavy rain on May 25–28, 2022: Causes, impacts, and disaster preparedness,” *Weather Clim. Extrem.*, vol. 39, Mar 2023, doi: 10.1016/j.wace.2022.100545.

[3] E. Alcantara *dkk.*, “Deadly disasters in southeastern South America: flash floods and landslides of February 2022 in Petrópolis, Rio de Janeiro,” *Natural Hazards and Earth System Sciences*, vol. 23, no. 3, hlm. 1157–1175, Mar 2023, doi: 10.5194/nhess-23-1157-2023.

[4] P. Ismartini, “Statistik Indonesia 2025,” 2025.

[5] R. Al Fauzi, “Analisis Tingkat Kerawanan Banjir Kota Bogor Menggunakan Metode Overlay dan Scoring Berbasis Sistem Informasi Geografis,” *Geomedia*, vol. 20, no. 2, hlm. 96–107, 2022, doi: https://journal.uny.ac.id/index.php/geomedia/index.

[6] S. Q. Dotse, I. Larbi, A. M. Limantol, dan L. C. De Silva, “A review of the application of hybrid machine learning models to improve rainfall prediction,” 1 Februari 2024, *Springer Science and Business Media Deutschland GmbH*. doi: 10.1007/s40808-023-01835-x.

[7] H. Jamaludin dan E. S. Wijaya, “Analisis Korelasi Curah Hujan dan Tinggi Muka Air Sungai Menggunakan Metode Regresi Linear,” *Jurnal Media Pratama*, vol. 17, no. 2, hlm. 141–147, 2023.

[8] C. Ley, R. K. Martin, A. Pareek, A. Groll, R. Seil, dan T. Tischer, “Machine learning and conventional statistics: making sense of the differences,” 1 Maret 2022, *Springer Science and Business Media Deutschland GmbH*. doi: 10.1007/s00167-022-06896-6.

[9] A. Wicaksono, “Prediksi dan Deteksi Bug Pada Software Menggunakan Pendekatan Machine Learning,” *Journal Signin*, hlm. 14–17, 2023.

[10] F. Sulianta, *Dasar dan Konsep Machine Learning*. Bandung: Feri Sulianta, 2025.
[11] R. F. Putra *dkk.*, *Data Mining: Algoritma dan Penerapannya*. Jakarta: Sonpedia, 2023.
[12] P. W. Rahayu *dkk.*, *Buku Ajar Data Mining*. Jakarta: Sonpedia, 2024.
[13] Z. Setiawan *dkk.*, *Buku Ajar Data Mining*. Jakarta: Sonpedia, 2023.
[14] G. Ashari Rakhmat dan W. Mutohar, "Prakiraan Hujan menggunakan Metode Random Forest dan Cross Validation," *MIND Journal*, vol. 8, no. 2, hlm. 173–187, 2023, doi: 10.26760/mindjournal.v8i2.173-187.
[15] I. D. G. L. Maheswara dan A. H. Al'aziz, "PERBANDINGAN MODEL MACHINE LEARNING PADA KLASIFIKASI CURAH HUJAN DI BOGOR," *INTI Nusa Mandiri*, vol. 19, no. 2, hlm. 202–210, Feb 2025, doi: 10.33480/inti.v19i2.6296.
[16] I. Saputra dan D. Ajeng Kristiyanti, "Application of Data Mining for Rainfall Prediction Classification in Australia with Decision Tree Algorithm and C5.0 Algorithm," hlm. 13–2021, [Daring]. Tersedia pada: https://www.kaggle.com/jsphyg/weather-dataset-rattle-
[17] S. S. P. Rachmawati, K. V. Prakusa, dan S. Rihastuti, "Penerapan Data Mining dengan Metode Decision Tree untuk Prediksi Cuaca di Kota Seattle menggunakan Aplikasi Weka," dalam *Seminar Nasional AMIKOM Surakarta (SEMNASA)*, Surakarta, Nov 2023, hlm. 93–100.
[18] L. Mdegela, E. Municio, Y. De Bock, E. Luhanga, J. Leo, dan E. Mannens, "Extreme Rainfall Event Classification Using Machine Learning for Kikuletwa River Floods," *Water (Switzerland)*, vol. 15, no. 6, Mar 2023, doi: 10.3390/w15061021.
[19] W. Sudrajat dan I. Cholid, "K-NEAREST NEIGHBOR (K-NN) UNTUK PENANGANAN MISSING VALUE PADA DATA UMKM," 2023.
[20] E. Erlin, Y. Desnelita, N. Nasution, L. Suryati, dan F. Zoromi, "Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, hlm. 677–690, Jul 2022, doi: 10.30812/matrik.v21i3.1726.
[21] M. A. Hasanah, S. Soim, dan A. S. Handayani, "Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir," 2021. [Daring]. Tersedia pada: http://jurnal.polibatam.ac.id/index.php/JAIC
[22] F. R. Rustan, H. W. Tanje, A. S. Sukri, M. K. Amir, M. Sriwati, dan R. M. Rachman, *Hidrologi*. Makassar: Tohar Media, 2024.
[23] A. P. Permana, A. Kurniyatul, dan F. H. H. Khadijah, "Analisis Perbandingan Algoritma Decision Tree, kNN, dan Naive Bayes untuk Prediksi Kesuksesan Start-up," *JISKa*, vol. 6, no. 3, hlm. 178–188, 2021, [Daring]. Tersedia pada: https://www.kaggle.com/manishkc06/startup-success-prediction.
[24] A. Pannadhitthana Candra, "Analisis Data Menggunakan Python: Memperkenalkan Pandas dan NumPy," vol. 3, no. 1, hlm. 11–16, 2025.
[25] N. A. P. Indaryono, R. R. Saedudin, dan F. Hamami, "Comparison Analysis of Random Forest and Naive Bayes Algorithms for Rainfall Classification based on Climate in Indonesia," *SITEKNIK*, vol. 1, no. 2, hlm. 102–109, 2024.
[26] Z. Mahmood, L. Jamel, D. A. Salem, dan I. Ashraf, "Improving learning from the complex multi-class imbalanced and overlapped data by mapping into higher dimension using SVM++," *Sci. Rep.*, vol. 15, no. 1, Des 2025, doi: 10.1038/s41598-025-13929-w.
[27] H. Vega-Huerta *dkk.*, "K-Nearest Neighbors Model to Optimize Data Classification According to the Water Quality Index of the Upper Basin of the City of Huarmey," *Applied Sciences (Switzerland)*, vol. 15, no. 18, Sep 2025, doi: 10.3390/app151810202.
[28] E. Halabaku dan E. Bytyçi, "Overfitting in Machine Learning: A Comparative Analysis of Decision Trees and Random Forests," *Intelligent Automation and Soft Computing*, vol. 39, no. 6, hlm. 987–1006, 2024, doi: 10.32604/iasc.2024.059429.