VOL. 19. NO. 2 FEBRUARI 2025
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri

INTI NUSA MANDIRI
DOI: https://doi.org/10.33480/inti.v19i2.6296

PERBANDINGAN MODEL *MACHINE LEARNING* PADA KLASIFIKASI CURAH HUJAN DI BOGOR

I Dewa Gede Loka Maheswara[1*]; Amad Hanif Al'aziz[2]

Program Studi Meteorologi[1, 2]
Sekolah Tinggi Meteorologi Klimatologi dan Geofisika, Kota Tangerang, Indonesia[1, 2]
www.stmkg.ac.id[1,2]
maheswaradewaloka@gmail.com[1*], ahmd.hanif.a19@gmail.com[2]
(*) Corresponding Author

Abstract— *Accurate rainfall prediction remains a significant challenge due to the involvement of complex physical processes and its substantial impact on various sectors of society. Rainfall prediction can be performed using classification techniques in Data Mining. Each algorithm employed for rainfall prediction may yield different performance outcomes, depending on factors such as the size of the dataset, the number of missing values, and the meteorological parameters utilized in the study. Selecting the appropriate algorithm for rainfall prediction continues to pose a challenge. This study aims to compare the performance of Naïve Bayes, Decision Tree, and Random Forest in order to identify the best model for classifying rainfall in Bogor Regency. The data utilized in this study includes maximum temperature, minimum temperature, average temperature, average humidity, duration of sunlight exposure, maximum wind speed, average wind speed, maximum wind direction, and rainfall. The dataset spans five years comprising a total 1.825 of data obtained from the Class III Citeko Meteorological Station. The results indicate that Random Forest, when trained with a smaller proportion of data compared to the proportion of test data to be predicted, achieves the best performance, with a precision of 59.1%, recall of 64.3%, and f1-score of 65.5%. This performance is attributed to the ensemble principle employed by Random Forest, which combines multiple weak learner trees to produce a robust learner tree.*

Keywords*: bogor, data mining, rainfall prediction, random forest.*

Abstrak—Prediksi curah hujan yang akurat sampai saat ini masih menjadi salah satu tantangan signifikan karena melibatkan proses fisis yang kompleks dan memiliki dampak yang substansial bagi masyarakat di berbagai sektor. Prediksi curah hujan dapat dilakukan dengan menggunakan teknik klasifikasi dalam Data Mining. Setiap algoritma yang digunakan untuk melakukan prediksi curah hujan dapat menghasilkan performa yang berbeda, tergantung pada ukuran dataset, jumlah nilai yang hilang, serta parameter meteorologi yang digunakan dalam penelitian. Memilih algoritma yang tepat untuk prediksi curah hujan merupakan hal yang masih menjadi tantangan hingga saat ini. Penelitian ini bertujuan membandingkan performa Naïve Bayes, Decision Tree, dan Random Forest, sehingga didapatkan model terbaik untuk melakukan klasifikasi curah hujan di Kabupaten Bogor. Data yang digunakan berupa suhu maksimum, suhu minimum, suhu rata-rata, kelembapan rata-rata, lamanya penyinaran matahari, kecepatan angin maksimum, kecepatan angin rata-rata, arah kecepatan angin maksimum, dan curah hujan. Ukuran dataset lima tahun dengan total 1.825 data yang diperoleh dari Stasiun Meteorologi Kelas III Citeko. Hasil penelitian menunjukkan Random Forest yang dilatih dengan proporsi data lebih sedikit dibandingkan proporsi data uji yang harus diprediksi mampu menghasilkan performa terbaik dengan nilai precision 59.1%, recall 64.3%, dan f1-score 65.5%. Ini dikarenakan Random Forest menggunakan prinsip ensemble, menggabungkan beberapa pohon pembelajar lemah sehingga dapat menghasilkan pohon pembelajar yang kuat.

Kata kunci: *bogor, data mining, prediksi curah hujan, random forest .*

VOL. 19. NO. 2 FEBRUARI 2025
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri

PENDAHULUAN

Kabupaten Bogor adalah salah satu wilayah di Indonesia yang mengalami curah hujan tahunan tinggi mencapai 5.000 mm per tahun (Mahfudz, Riadi, & Rifaldi, 2022; Mukti, 2023; H. Setiawan, Wibowo, & Supriatna, 2021). Peningkatan curah hujan di daerah ini berpotensi menyebabkan banjir, yang dapat menimbulkan kerugian signifikan bagi masyarakat di wilayah terdampak (Khoirunnisa et al., 2024; Ramadhani, 2022; Syukur, 2021). Oleh karena itu, ketepatan prediksi pola hujan penting untuk keperluan mitigasi bencana alam yang lebih efisien. Meskipun demikian, prediksi hujan yang tepat dan akurat masih menjadi tantangan yang signifikan karena melibatkan proses fisis kompleks dan mengingat dampaknya yang signifikan pada berbagai sektor (Al Fauzi, 2022; Dotse, Larbi, Limantol, & De Silva, 2024; Jamaludin & Wijaya, 2023). Prediksi hujan umumnya dibuat untuk periode waktu setiap jam, harian, hingga mingguan (Wijanarko, 2021; Yusuf, Setyanto, & Aryasa, 2022). Optimalisasi prediksi hujan dapat dicapai dengan mengatasi kompleksitas statistik data parameter meteorologi, seperti arah dan kecepatan angin, tekanan, kelembaban, radiasi matahari, dan parameter lainnya (Hasanah, Soim, Handayani, & others, 2021; Herdhyanti, Muflikhah, & Cholissodin, 2022; Sukman et al., 2024). Salah satu metode yang banyak digunakan untuk prediksi hujan adalah metode *data mining* (Hasanah et al., 2021).

Metode *data mining* memungkinkan pembuatan prediksi pola hujan dengan tingkat akurasi yang tinggi. Data mining merupakan metode pemecahan suatu permasalahan dengan menggunakan sekumpulan algoritma, seperti *Random Forest*, *Decision Tree*, *Support Vector Machine*, *K-Nearest Neighbor*, *Naïve Bayes*, dan lainnya (Putra et al., 2023; Rahayu et al., 2024; Z. Setiawan et al., 2023). Proses pemecahan masalah dilakukan dengan mencari hubungan antar data yang belum diketahui, mengumpulkan pola-pola implisit dalam data, serta membangun model baru untuk memperoleh informasi dari dalam basis data (Putra et al., 2023).

Dalam menyelesaikan permasalahan, data mining umumnya menggunakan fungsi *clustering*, asosiasi, prediksi, dan klasifikasi (Rahayu et al., 2024; Sa'adah, Rochayani, Lestari, & Lusia, 2021). Untuk melakukan klasifikasi, terdapat dua tahapan yang harus dilakukan. Tahapan dimulai dengan proses belajar, yaitu data latih akan dipelajari oleh algoritma klasifikasi. Tahapan kedua adalah proses klasifikasi, yaitu melakukan evaluasi kinerja algoritma klasifikasi menggunakan data uji (Al Arif, Firdaus, & Maruhawa, 2022; Syamsurizal, Cumel, Zamri, & Rahmaddeni, 2022).

Penelitian mengenai penerapan teknik data mining untuk klasifikasi prediksi curah hujan telah dilakukan sebelumnya di berbagai wilayah. Di Kota Seattle, Rachmawati, Prakusa, & Rihastuti (2023) berhasil memperoleh tingkat akurasi prediksi cuaca yang tinggi dengan menggunakan *Decision Tree*. Di sisi lain, Indaryono, Saedudin, & Hamami (2024) menunjukkan bahwa meskipun performa model *Random Forest* lebih baik dalam klasifikasi curah hujan dibandingkan *Naïve Bayes* dalam hal akurasi, presisi, *recall*, dan *F1-Score*, model *Naïve Bayes* memiliki skor AUC yang jauh lebih rendah. Penelitian oleh Mdegela et al. (2023) di Tanzania menunjukkan bahwa *Random Forest* dan XGBoost memiliki performa yang lebih baik secara keseluruhan dibandingkan *Support Vector Machine*, *K-Nearest Neighbour*, dan *Multilayer Perceptron* dalam memprediksi kejadian hujan lebat pada dataset yang tidak seimbang. Pada wilayah yang berbeda, Rizqi & Kusumaningsih (2022) di Bogor menemukan bahwa *Naive Bayes* efektif dalam menangani data tidak seimbang dengan akurasi mencapai 92%.

Pada wilayah berbeda, Rakhmat & Mutohar (2023) menunjukkan bahwa model *Random Forest* dengan teknik *cross-validation* memiliki performa lebih optimal dibandingkan tanpa *cross-validation* untuk prediksi curah hujan. Saputra & Kristiyanti (2021) menunjukkan bahwa metode validasi data memengaruhi akurasi model, dengan *Decision Tree* mencapai akurasi paling baik menggunakan *10-Cross Fold Validation.* Hasil-hasil ini menunjukkan bahwa performa model dalam klasifikasi prediksi curah hujan dipengaruhi oleh metode validasi dan jumlah data yang digunakan.

Penelitian ini akan berfokus pada perbandingan performa tiga model klasifikasi, yaitu *Random Forest*, *Decision Tree*, dan *Naïve Bayes* dengan menggunakan metode validasi pembagian data dan *10-Cross Fold Validation*. Diharapkan, penelitian ini dapat menghasilkan model klasifikasi terbaik untuk melakukan klasifikasi hujan di Kabupaten Bogor serta dapat memberikan kontribusi signifikan terhadap upaya mitigasi bencana alam yang lebih efisien.

BAHAN DAN METODE

Wilayah yang menjadi fokus kajian dalam penelitian ini adalah Kabupaten Bogor, yang terletak pada koordinat 6.19°LS hingga 6.47°LS dan 106°BT hingga 107°BT. Penelitian ini menggunakan data dengan rentang waktu Lima tahun, dimulai dari tanggal 1 Januari 2019 hingga 31 Desember 2023. Parameter meteorologi yang digunakan dalam penelitian ini meliputi suhu maksimum, suhu minimum, suhu rata-rata, kelembapan rata-rata,

lamanya penyinaran matahari, kecepatan angin maksimum, kecepatan angin rata-rata, arah kecepatan angin maksimum, dan curah hujan. Data tersebut diperoleh dari Stasiun Meteorologi Kelas III Citeko melalui portal resmi BMKG di https://dataonline.bmkg.go.id/ dengan menggunakan akun yang sudah terdaftar sebelumnya, sebagaimana ditampilkan pada Tabel 1.

Penelitian ini memanfaatkan aplikasi pengolah data angka untuk melihat kelengkapan data serta Google Colab sebagai *platform* untuk melakukan proses pengisian data kosong, normalisasi data, validasi data, dan perhitungan evaluasi performa model. Semua proses dilakukan menggunakan bahasa pemrograman *Python*, yang memungkinkan pengolahan dan analisis data secara efektif (Priyatno et al., 2023).

Tabel 1. Parameter Meteorologi

Parameter	Tipe Data	Satuan	*Valid Records*	*Missing Values*
Suhu Maksimum (Tx)	float64	°C	1821	4
Suhu Minimum (Tn)	float64	°C	1806	19
Suhu Rata-Rata (Tavg)	float64	°C	1820	5
Kelembapan Rata-Rata (RH_avg)	float64	%	1820	5
Lama Penyinaran Matahari (ss)	float64	jam	1817	8
Kecepatan Angin Maksimum (ff_x)	int64	m/s	1823	2
Kecepatan Angin Rata-Rata (ff_avg)	int64	m/s	1823	2
Arah Kecepatan Angin Maksimum (ddd_x)	int64	°	1822	3
Curah Hujan (RR)	float64	mm	1702	123

Sumber: (Hasil Penelitian, 2025)

Tahap *pre-processing* data dilakukan dengan tujuan mempersiapkan data agar siap digunakan dalam penelitian. Tahap *pre-processing* dimulai dengan pengumpulan data parameter meteorologi sebagaimana ditampilkan pada Tabel 1. Selanjutnya, dilakukan transformasi data dengan mengubah seluruh dataset menjadi tipe data numerik. Proses pembersihan *missing value* dilakukan dengan mengisi nilai rata-rata dari keseluruhan data pada setiap kolom yang memiliki nilai kosong, merujuk pada penelitian yang dilakukan oleh Saputra & Kristiyanti (2021). *Missing value* dapat mengurangi akurasi hasil prediksi (Sudrajat & Cholid, 2023). Tabel 2 menunjukkan contoh *missing value* pada kolom parameter meteorology yang disimbolkan dengan "*" yang selanjutnya akan diisi dengan nilai rata-rata dari keseluruhan data setiap kolom pada kolom data yang kosong.

Tabel 2. Contoh Missing Value Pada Parameter Meteorologi

Tanggal	05/03/2019	06/03/2019	27/04/2020	18/01/2021	04/08/2022
Tx	23.8	*	*	22	24.4
Tn	19.2	*	20	*	*
T_avg	21.5	*	*	19.7	20.9
RH_avg	90	94	*	95	90
ss	2.4	0.2	3.5	0.5	4.5
ff_x	3	2	0	4	3
ddd_x	360	310	0	320	330
Ff_avg	1	0	0	2	1
RR	*	15.8	6.6	8	*

Sumber: (Hasil Penelitian, 2025)

Kemudian, curah hujan diklasifikasikan menjadi enam kategori, yaitu tidak hujan (curah hujan 0 mm) dikategorikan sebagai "1", hujan ringan (curah hujan 0 mm sampai 20 mm) dikategorikan sebagai "2", hujan sedang (curah hujan 20 mm sampai 50 mm) dikategorikan sebagai "3", hujan lebat (curah hujan 50 mm sampai 100 mm) dikategorikan sebagai "4", hujan sangat lebat (curah hujan 100 mm sampai 150 mm) dikategorikan sebagai "5", dan hujan ekstrem (curah hujan melebihi 150 mm) dikategorikan sebagai "6", merujuk pada ketentuan Badan Meteorologi Klimatologi dan Geofisika (BMKG) mengenai probabilistik curah hujan 24 jam.

Selanjutnya, dilakukan normalisasi data menggunakan metode *Min-Max Scaling*, yang mengubah nilai setiap kolom menjadi rentang 0 hingga 1 untuk membatasi nilai dalam interval tertentu agar memudahkan proses prediksi, merujuk pada penelitian yang dilakukan oleh Mdegela et al. (2023). Secara matematis, normalisasi data dapat dilakukan dengan persamaan 1.

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

Tabel 3 merupakan contoh data pada kolom parameter meteorologi yang belum melalui tahap normalisasi data dan Tabel 4 menunjukkan contoh data pada kolom parameter meteorologi yang sudah melalui tahap normalisasi data dengan menggunakan persamaan 1.

VOL. 19. NO. 2 FEBRUARI 2025
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri

Tabel 3. Contoh Data Sebelum Tahap Normalisasi Data.

Tanggal	01/01/2019	02/01/2019	03/01/2019	30/12/2023	31/12/2023
Tx	24.5	26.3	26	25.8	25.6
Tn	19	19.5	19.6	20.6	19.7
T_avg	20.2	21.2	21.8	21.7	21.6
RH_avg	94	89	82	92	92
ss	0.5	0.3	1.5	2.8	0
ff_x	3	2	3	1	3
ddd_x	280	280	270	320	360
ff_avg	1	1	1	0	0
RR	23.5	23.3	8.8	3	13.3

Sumber: (Hasil Penelitian, 2025)

Tabel 4. Contoh Data Sesudah Tahap Normalisasi Data.

Tanggal	01/01/2019	02/01/2019	03/01/2019	30/12/2023	31/12/2023
Tx	0.42	0.57	0.54	0.52	0.51
Tn	0.66	0.73	0.75	0.91	0.77
T_avg	0.26	0.4	0.48	0.47	0.45
RH_avg	0.88	0.78	0.65	0.84	0.84
ss	0.05	0.03	0.15	0.28	0
ff_x	0.25	0.17	0.25	0.08	0.25
ddd_x	0.78	0.78	0.75	0.89	1
ff_avg	0.5	0.5	0.5	0	0
RR	0.19	0.19	0.07	0.02	0.11

Sumber: (Hasil Penelitian, 2025)

Setelah data melalui tahap *pre-processing*, dilakukan tahap *processing* data dengan menggunakan tiga model klasifikasi, yaitu *Naïve Bayes*, *Decision Tree*, dan *Random Forest*. *Naïve Bayes* digunakan karena kemampuannya menghasilkan prediksi akurat jika data set yang digunakan sangat besar (Husin, 2023). *Decision Tree* dipilih karena fleksibilitasnya yang memungkinkan untuk menangkap berbagai pola dan hubungan dalam data (Permana, Ainiyah, & Holle, 2021). *Random Forest* dipilih karena kemampuannya menangani data set yang besar dan data set yang tidak seimbang (Erlin, Desnelita, Nasution, Suryati, & Zoromi, 2022). Kemudian, dilakukan pemilihan variabel bebas yang akan digunakan untuk memengaruhi hasil prediksi kategori curah hujan, di antaranya 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, serta variabel terikat berupa kategori curah hujan sebagai target prediksi. Metode validasi data yang digunakan adalah pembagian data dengan rasio tertentu dan *10-cross fold validation*.

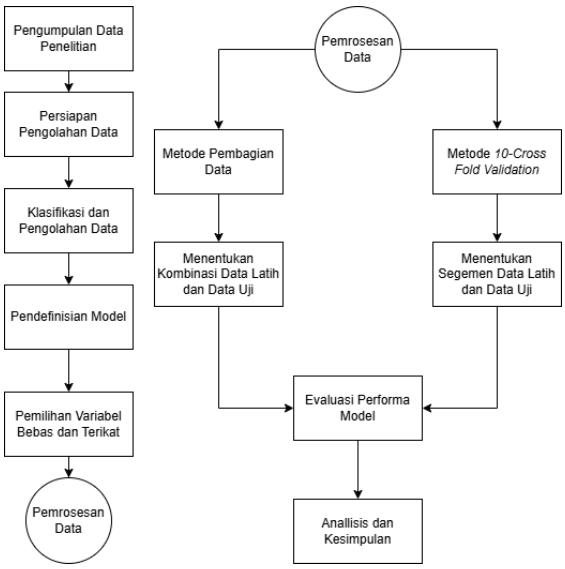
Setelah melalui tahap *processing* data, dilakukan tahap *post-processing* data dengan melakukan evaluasi performa model menggunakan nilai *precision*, *recall,* dan *f1-score* untuk setiap metode validasi data. Secara matematis, nilai *precision*, *recall,* dan *f1-score* dapat ditentukan dengan persamaan 2, 3, dan 4. Hasil evaluasi performa model akan ditampilkan dalam bentuk tabel.

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (3)$$

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

Langkah kerja penelitian disajikan dalam bentuk diagram alir pada Gambar 2.



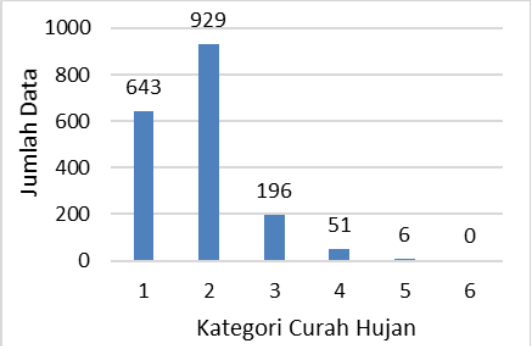
Sumber: (Hasil Penelitian, 2025)

Gambar 1. Diagram alir penelitian

HASIL DAN PEMBAHASAN

Gambar 2 merupakan hasil pengkategorian dari data curah hujan yang diperoleh dari Stasiun Meteorologi Kelas III Citeko dengan ketentuannya merujuk pada ketentuan Badan Meteorologi Klimatologi dan Geofisika (BMKG) mengenai probabilistik curah hujan 24 jam. Pada Gambar 2, dapat dilihat kategori 1 (tidak hujan) terdapat 643 data, kategori 2 (hujan ringan) terdapat 929 data, kategori 3 (hujan sedang) terdapat 196 data, kategori 4 (hujan lebat) terdapat 51 data, kategori 5

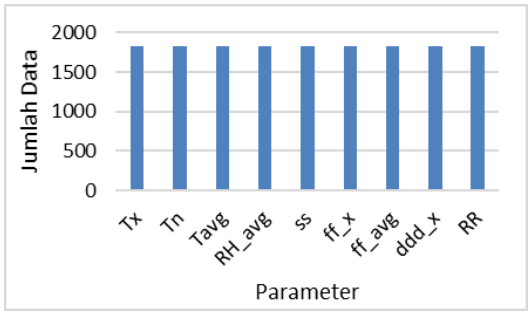
(hujan sangat lebat) terdapat 6 data, dan tidak terdapat data untuk kategori 6 (hujan ekstrem).



Sumber: (Hasil Penelitian, 2025)

Gambar 2. Histogram kategori curah hujan

Penelitian ini menggunakan 1.825 data curah hujan, yang masing-masing memiliki 9 parameter. Berdasarkan Tabel 1, jumlah data pada setiap parameter meteorologi awalnya berbeda-beda. Setelah tahap *preprocessing*, setiap parameter memiliki 1.825 data, seperti yang terlihat pada Gambar 3, sehingga total keseluruhan mencapai 16.425 parameter. Total baris data/parameter ini mengalami peningkatan sebesar 1.05% dibandingkan dengan jumlah baris data awal yang diperoleh dari Stasiun Meteorologi Kelas III Citeko.



Sumber: (Hasil Penelitian, 2025)

Gambar 3. Grafik Baris Data Setiap Parameter

Hasil performa ketiga model dengan metode pembagian data yang dievaluasi menggunakan *precision*, *recall,* dan *f1-score*. Hasil penelitian menunjukkan variasi performa model pada setiap rasio pembagian data yang diterapkan.

Tabel 5. Hasil Performa Model *Naïve Bayes*

Pembagian Data (Data Latih-Data Uji)	*Precision*	*Recall*	*F1-Score*
90-10	0.622070	0.644809	0.631841
80-20	0.604477	0.616438	0.603717
70-30	0.596584	0.602190	0.597434
60-40	0.590344	0.591781	0.589643
50-50	0.577734	0.567360	0.571686
40-60	0.602968	0.610959	0.602920
30-70	0.609281	0.617371	0.608949
20-80	0.591112	0.615753	0.595034
10-90	0.599284	0.597079	0.591551

Sumber: (Hasil Penelitian, 2025)

Tabel 5 merupakan Hasil performa model *Naïve Bayes* menggunakan metode validasi pembagian data latih dan data uji dengan rasio tertentu. Berdasarkan hasil tersebut, model *Naïve Bayes* menunjukkan performa terbaik ketika menggunakan metode pembagian data dengan rasio data latih 90% dan rasio data uji 10%. Hal ini terlihat dari nilai *f1-score* yang mencapai 63.2%, yang lebih tinggi dibandingkan nilai *f1-score* pada rasio pembagian data lainnya.

Tabel 6. Hasil Performa Model *Decision Tree*

Pembagian Data (Data Latih-Data Uji)	*Precision*	*Recall*	*F1-Score*
90-10	0.547180	0.546448	0.545369
80-20	0.515723	0.504110	0.509538
70-30	0.536276	0.525547	0.530069
60-40	0.517262	0.508219	0.512251
50-50	0.512584	0.507119	0.509274
40-60	0.514341	0.507763	0.510078
30-70	0.541842	0.532864	0.536857
20-80	0.521453	0.525342	0.522669
10-90	0.547185	0.545344	0.545217

Sumber: (Hasil Penelitian, 2025)

Tabel 6 merupakan hasil performa model *Decision Tree* menggunakan metode validasi pembagian data latih dan data uji dengan rasio tertentu. Berdasarkan Tabel 6, model *Decision Tree* menunjukkan performa terbaik ketika menggunakan metode pembagian data dengan rasio data latih 90% dan rasio data uji 10%. Ini terlihat dari nilai *f1-score* yang mencapai 54.5%, yang lebih tinggi dibandingkan nilai *f1-score* dengan rasio pembagian data lainnya.

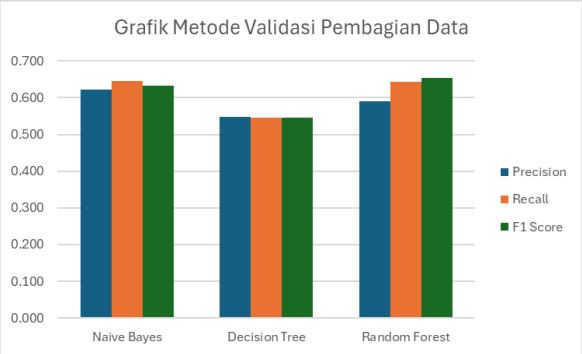
Tabel 7. Hasil Performa Model *Random Forest*

Pembagian Data (Data Latih-Data Uji)	*Precision*	*Recall*	*F1-Score*
90-10	0.645956	0.704918	0.597283
80-20	0.637622	0.657534	0.607494
70-30	0.602234	0.655109	0.610902
60-40	0.624099	0.658904	0.608767
50-50	0.578078	0.648412	0.601630
40-60	0.607992	0.647489	0.615642
30-70	0.606947	0.650235	0.610274
20-80	0.589306	0.652740	0.610898
10-90	0.590872	0.643335	0.654803

Sumber: (Hasil Penelitian, 2025)

VOL. 19. NO. 2 FEBRUARI 2025
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri

Tabel 7 menjabarkan hasil performa model *Random Forest* menggunakan metode validasi pembagian data latih dan data uji dengan rasio tertentu. Berdasarkan Tabel 7, model *Random Forest* menunjukkan performa terbaik ketika menggunakan metode pembagian data dengan rasio data latih 10% dan rasio data uji 90%. Ini terlihat dari nilai *f1-score* yang mencapai 65.5%, yang lebih tinggi dibandingkan nilai *f1-score* dengan rasio pembagian data lainnya.

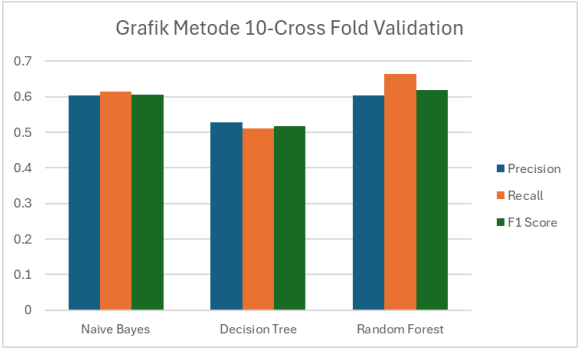


Sumber: (Hasil Penelitian, 2025)

Gambar 4. Perbandingan Performa Model *Naïve Bayes*, *Decision Tree*, dan *Random Forest* Dengan Metode Validasi Pembagian Data.

Gambar 4 menyajikan grafik bar yang menunjukkan nilai *precision*, *recall* dan *f1-score* yang digunakan untuk mengevaluasi performa masing-masing model, dimana nilai yang disajikan diambil dari hasil performa tertinggi pada Tabel 5, Tabel 6 dan Tabel 7. Hasil tersebut menunjukkan bahwa metode pembagian data dengan rasio tertentu memberikan performa yang berbeda untuk ketiga model yang diuji. Dapat dilihat, model *Random Forest* menunjukkan performa paling baik yang dibuktikan dengan nilai *f1-score* mencapai 65.5% ketika menggunakan pembagian data dengan rasio data latih 10% dan rasio data uji 90%.

Hasil ini menunjukkan model Random Forest memerlukan lebih banyak data uji untuk menghasilkan performa model yang baik dalam melakukan prediksi klasifikasi curah hujan. Sebaliknya, *Naïve Bayes* dan *Decision Tree* menunjukkan performa yang baik menggunakan pembagian data dengan rasio data latih 90% dan rasio data uji 10%. Hasil ini menunjukkan kedua model tersebut lebih efektif ketika dilatih dengan proporsi data yang lebih banyak, yang memungkinkan kedua model belajar dari lebih banyak informasi dan menghasilkan prediksi yang lebih akurat.



Sumber: (Hasil Penelitian, 2025)

Gambar 5. Perbandingan Performa Model Naïve Bayes, Decision Tree, dan Random Forest Dengan Metode Validasi *10-Cross Fold Validation*.

Gambar 5 menunjukkan metode *10-cross fold validation* memberikan performa yang berbeda untuk ketiga model. Grafik bar menunjukkan nilai *precision*, *recall,* dan *f1-score* yang digunakan untuk mengevaluasi performa model, semakin tinggi nilainya, maka semakin baik performa model dalam melakukan prediksi. Dapat dilihat model *Random Forest* menunjukkan performa lebih baik dibandingkan *Naïve Bayes* dan *Decision Tree* dalam melakukan prediksi klasifikasi curah hujan yang dibuktikan dengan nilai *precision* sebesar 60.4%, *recall* sebesar 66.3%, dan *f1-score* sebesar 61.9%.

Random Forest menunjukkan performa lebih baik dibandingkan dengan Naïve Bayes dan *Decision Tree* pada metode pembagian data maupun *10-cross fold validation* yang dibuktikan dengan nilai *f1-score* yang lebih tinggi. Hal ini dapat disebabkan oleh banyaknya parameter meteorologi yang digunakan pada penelitian ini. Random Forest sebagai model yang lebih kompleks dibandingkan *Naïve Bayes* dan *Decision Tree* dapat menjelaskan hubungan antara variabel yang lebih banyak (Indaryono et al., 2024).

Random Forest merupakan metode pembelajaran *ensemble* yang menggabungkan beberapa *Decision Tree* untuk menghasilkan keputusan yang lebih kuat. Dalam penelitian ini, *Random Forest* menunjukkan performa terbaik dalam memprediksi klasifikasi curah hujan di Kabupaten Bogor, dengan *f1-score* sebesar 65.5% meskipun menggunakan proporsi data latih yang lebih kecil (10%) dibandingkan data uji (90%). *Random Forest* memberikan nilai *f1-score* lebih baik dibandingkan dengan *Naïve Bayes* dan *Decision Tree* yang menggunakan lebih banyak data latih.

Ketiga model yang diuji masih menunjukkan performa prediksi yang rendah, dengan *f1-score* berkisar 59% hingga 65%. Performa yang rendah dapat disebabkan oleh ukuran *dataset* yang digunakan pada penelitian ini antara 1 Januari 2019 hingga 31 Desember 2023 (lima tahun), sedangkan penelitian sebelumnya (Fazil, Hakimi, Akbari, Quchi,

& Khaliqyar, 2023; Mdegela et al., 2023; Saputra & Kristiyanti, 2021) menggunakan ukuran *dataset* antara 10 tahun sampai 35 tahun . Selain itu, adanya 123 *missing value* pada data curah hujan serta penggunaan banyak parameter meteorologi yang kurang representatif juga dapat memengaruhi akurasi model. Hal ini disebabkan oleh adanya parameter yang tidak relevan dapat mengalihkan perhatian model dari parameter yang benar-benar penting, sehingga mengurangi akurasi prediksi.

KESIMPULAN

Penelitian ini telah berhasil mencapai tujuan untuk membandingkan performa model *Naïve Bayes, Decision Tree*, dan *Random Forest* dalam melakukan klasifikasi prediksi curah hujan di Kabupaten Bogor. Hasil penelitian menunjukkan *Random Forest* menghasilkan performa terbaik yang dibuktikan dengan nilai *f1-score* tertinggi, baik pada metode pembagian data dengan rasio tertentu maupun dengan metode *10-cross fold validation. Random Forest* menghasilkan performa yang baik karena kemampuannya menggunakan sedikit proporsi data latih untuk memprediksi proporsi data uji yang lebih banyak. Meskipun *Naïve Bayes* dan *Decision Tree* dilatih dengan proporsi data yang lebih banyak dibandingkan *Random Forest*, keduanya tidak menunjukkan performa yang lebih baik dibandingkan *Random Forest*. Kompleksitas *Random Forest* juga membantu dalam menjelaskan hubungan variabel yang lebih banyak.

Untuk penelitian selanjutnya, disarankan menggunakan ukuran *dataset* yang lebih panjang antara rentang 10 tahun sampai 50 tahun, memilih parameter meteorologi yang lebih relevan, serta membandingkan lebih banyak model dan lokasi penelitian untuk memperluas pemahaman mengenai efektivitas teknik *data mining* dalam klasifikasi prediksi curah hujan.

REFERENSI

Al Arif, A., Firdaus, M., & Maruhawa, Y. (2022). Perbandingan Metode Data Mining untuk Prediksi Curah Hujan dengan Algoritma C4. 5, Naive Bayes, dan KNN: Comparison of Data Mining Methods for Predition of Rainfall with C4. 5, Naive Bayes, and KNN Algorithm. *SENTIMAS: Seminar Nasional Penelitian Dan Pengabdian Masyarakat*, 187–197.

Al Fauzi, R. (2022). Analisis tingkat kerawanan banjir Kota Bogor menggunakan metode overlay dan scoring berbasis sistem informasi geografis. *Geomedia: Majalah Ilmiah Dan Informasi Kegeografian*, *20*(2), 96–107. http://dx.doi.org/10.21831/gm.v20i2.48017

Dotse, S., Larbi, L., Limantol, A., & De Silva, L. (2024). A Review of The Application of Hybrid Machine Learning Models to Improve Rainfall Prediction. *Modeling Earth Systems and Environment*, *10*, 19. https://doi.org/10.1007/s40808-023-01835-x

Erlin, E., Desnelita, Y., Nasution, N., Suryati, L., & Zoromi, F. (2022). Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *21*(3). https://doi.org/https://doi.org/10.30812/matrik.v21i3.1726

Fazil, A. W., Hakimi, M., Akbari, R., Quchi, M. M., & Khaliqyar, K. Q. (2023). Comparative Analysis of Machine Learning Models for Data Classification: An In-Depth Exploration. *Journal of Computer Science and Technology Studies*, *5*(4), 160–168. https://doi.org/10.32996/jcsts.2023.5.4.16

Hasanah, M. A., Soim, S., Handayani, A. S., & others. (2021). Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir. *Journal of Applied Informatics and Computing*, *5*(2), 103–108. https://doi.org/10.30871/jaic.v5i2.3200

Herdhyanti, A., Muflikhah, L., & Cholissodin, I. (2022). Prediksi Curah Hujan dengan Empat Parameter menggunakan Backpropagation (Studi Kasus: Stasiun Meteorologi Ahmad Yani). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *6*(12), 5862–5870. Diambil dari https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/12022

Husin, N. (2023). Komparasi Algoritma Random Forest, Naïve Bayes, dan Bert Untuk Multi-Class Classification Pada Artikel Cable News Network (CNN). *Jurnal Esensi Infokom : Jurnal Esensi Sistem Informasi Dan Sistem Komputer*. Retrieved from https://api.semanticscholar.org/CorpusID:260081418

Indaryono, N. A. P., Saedudin, R. R., & Hamami, F. (2024). Comparison Analysis of Random Forest and Naive Bayes Algorithms for Rainfall Classification Based on Climate in Indonesia. *SITEKNIK: Information Systems, Engineering and Applied Technology*, *1*(2), 102–109.

VOL. 19. NO. 2 FEBRUARI 2025
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri

Jamaludin, H., & Wijaya, E. S. (2023). Analisis Korelasi Curah Hujan dan Tinggi Muka Air Sungai Menggunakan Metode Regresi Linear. *Jurnal Media Pratama*, *17*(2), 141–147.

Khoirunnisa, A. P., Andini, A. P., Keriswali, M. G., Husein, R. K., Setiyoko, R., Kusuma, W. A., & Situmorang, M. T. N. (2024). Kebijakan Pemerintah Kota Bogor Dalam Penanggulangan Gempa Bumi. *Iuris Studia: Jurnal Kajian Hukum*, *5*(3), 789–793.

Mahfudz, M., Riadi, B., & Rifaldi, I. (2022). Pemetaan Area Potensi Banjir Berdasarkan Topographic Wetness Index (TWI) di Kecamatan Cigudeg Kabupaten Bogor. *Geomatika*, *28*(1), 13–20.

Mdegela, L., Municio, E., Bock, Y. D., Luhanga, E., & Leo, J. (2023). Extreme Rainfall Event Classification Using Machine Learning for Kikuletwa River Floods. *Water*, *15*, 1–14. https://doi.org/10.3390/w15061021

Mukti, A. (2023). Penggunaan lahan dan deforestasi di Kabupaten Bogor. *Jurnal Bisnis Kehutanan Dan Lingkungan*, *1*(1). https://doi.org/10.61511/jbkl.v1i1.2023.206

Permana, A. P., Ainiyah, K., & Holle, K. F. H. (2021). Analisis Perbandingan Algoritma Decision Tree, kNN, dan Naive Bayes untuk Prediksi Kesuksesan Start-up. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *6*(3), 178–188. https://doi.org/10.14421/jiska.2021.6.3.178-188

Priyatno, A., Firmanda, F., Farhas, R., Amalia, F., Febri, W., & Sudirman, R. (2023). Pelatihan Data Science menggunakan Bahasa Pemrograman Python di PT Ilmu Data Indonesia. *Dedikasi: Jurnal Pengabdian Pendidikan Dan Teknologi Masyarakat*, *1*, 31–36. https://doi.org/10.31004/dedikasi.v1i1.12

Putra, R. F., Zebua, R. S. Y., Budiman, B., Rahayu, P. W., Bangsa, M. T. A., Zulfadhilah, M., … Andiyan, A. (2023). *Data Mining: Algoritma dan Penerapannya*. PT. Sonpedia Publishing Indonesia.

Rachmawati, S. S. P., Prakusa, K. V., & Rihastuti, S. (2023). Penerapan Data Mining dengan Metode Decision Tree untuk Prediksi Cuaca di Kota Seattle Menggunakan Aplikasi Weka. *Prosiding Seminar Nasional Amikom Surakarta*, 93–100. Surakarta: STMIK Amikom Surakarta.

Rahayu, P. W., Sudipa, I. G. I., Suryani, S., Surachman, A., Ridwan, A., Darmawiguna, I. G. M., … Maysanjaya, I. M. D. (2024). *Buku Ajar Data Mining*. PT. Sonpedia Publishing Indonesia.

Rakhmat, G. A., & Mutohar, W. (2023). Prakiraan Hujan Menggunakan Metode Random Forest dan Cross Validation. *MIND Journal*, *8*(2), 173–187. https://doi.org/10.26760/mindjournal.v8i2.173-187

Ramadhani, S. S. (2022). *Stategi Bpbd Kabupaten Pacitan Dalam Upaya Penanggulangan Bencana Banjir Dan Tanah Longsor*. Universitas Muhammadiyah Ponorogo.

Rizqi, A. A., & Kusumaningsih, D. (2022). Klasifikasi Curah Hujan di Kota Bogor Provinsi Jawa Barat dengan Menggunakan Metode Naive Bayes. *Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)*. Retrieved from https://senafti.budiluhur.ac.id/index.php/senafti/article/view/410

Sa'adah, U., Rochayani, M. Y., Lestari, D. W., & Lusia, D. A. (2021). *Kupas Tuntas Algoritma Data Mining dan Implementasinya Menggunakan R*. Universitas Brawijaya Press.

Saputra, I., & Kristiyanti, D. A. (2021). Application of Data Mining for Rainfall Prediction Classification in Australia with Decision Tree Algorithm and C5.0 Algorithm. *Seminar Nasional Informatika (SEMNASIF)*. Retrieved from http://jurnal.upnyk.ac.id/index.php/semnasif/article/view/6060

Setiawan, H., Wibowo, A., & Supriatna, S. (2021). Pembuatan peta curah hujan untuk evaluasi kesesuaian rencana tata ruang kawasan hutan Kabupaten Bogor. *Geomedia: Majalah Ilmiah Dan Informasi Kegeografian*, *19*(2), 113–121. Diambil dari https://journal.uny.ac.id/index.php/geomedia/article/view/43227/16848

Setiawan, Z., Fajar, M., Priyatno, A. M., Putri, A. Y. P., Aryuni, M., Yuliyanti, S., … others. (2023). *Buku Ajar Data Mining*. PT. Sonpedia Publishing Indonesia.

Sudrajat, W., & Cholid, I. (2023). K-Nearest Neighbor (K-NN) untuk Penanganan Missing Value pada Data UMKM. *Jurnal Rekayasa Sistem Informasi Dan Teknologi*, *1*(2), 54–63. https://doi.org/10.59407/jrsit.v1i2.77

Sukman, Rustan, F. R., Yusman, Tanje, H. W., Sukri, A. S., Amir, M. K., … Rachman, R. M. (2024). *Hidrologi*. TOHAR MEDIA.

Syamsurizal, Cumel, Zamri, D., & Rahmaddeni. (2022). Perbandingan Metode Data Mining untuk Prediksi Banjir Dengan Algoritma Naive Bayes dan KNN: Comparison of Data Mining Methods for Predition of Floods with Naive Bayes and KNN Algorithm. *SENTIMAS: Seminar Nasional Penelitian Dan Pengabdian Masyarakat*, 40–48.

Syukur, A. (2021). *Buku Pintar Penanggulangan Banjir*. DIVA PRESS.

Wijanarko, H. (2021). *Validasi Data Curah Hujan Pos Penakar Hujan dengan Data Curah Hujan TRMM*. Universitas Lampung.

Yusuf, M., Setyanto, A., & Aryasa, K. (2022). Analisis Prediksi Curah Hujan Bulanan Wilayah Kota Sorong Menggunakan Metode Multiple Regression. *Jurnal Sains Komputer Dan Informatika*, *6*(1), 405–417. http://dx.doi.org/10.30645/j-sakti.v6i1.455