Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.5, October 2025, pp. 2364~2370
e-ISSN: 2548-6861

Application of Feature Selection and Comparative Analysis of Machine Learning Models for Rainfall Prediction in Jakarta

Indah Dwi Sulistyowati [1]*, Sunarno [2]*, Iqbal [3], KGS M Nurs Syamsuri [4]****
* Program Studi Fisika, Universitas Negeri Semarang
** Badan Meteorologi Klimatologi dan Geofisika
ndahe1549@students.unnes.ac.id [1], narnophysics@mail.unnes.ac.id [2], iqbal@bmkg.go.id [3], nur.syamsuri@bmkg.go.id [4]

Article Info

Article history:

Received 2025-09-01
Revised 2025-09-15
Accepted 2025-09-19

Keyword:

Classification,
Machine learning,
Prediction,
Feature selection

ABSTRACT

Accurate rainfall prediction plays a vital role in reducing disaster risks and supporting public preparedness, particularly in Jakarta where dense population and frequent floods cause serious economic and social impacts. In this study, weather data from the Kemayoran Meteorological Station covering 2004–2023 were analyzed to build rainfall prediction models using machine learning. Three classification algorithms were compared: Logistic Regression, Decision Tree, and Random Forest, selected to represent linear, non-linear, and ensemble approaches. Feature selection was applied using Recursive Feature Elimination (RFE) to identify the most relevant predictors. The models were evaluated using 5-fold cross-validation with metrics including Accuracy, Precision, Recall, F1 Score, ROC AUC, and Cohen's Kappa. The results indicate that Random Forest achieved the best overall performance with Accuracy of 0.7622, Precision around 0.70, Recall up to 0.63, F1 Score about 0.65, ROC AUC ranging from 0.8044 to 0.8171, and Cohen's Kappa near 0.48. Logistic Regression also performed competitively with Accuracy of 0.7648, ROC AUC of 0.829, and Kappa of 0.49, while Decision Tree showed lower results with Accuracy of 0.6890 and ROC AUC of 0.6636. The RFE process successfully reduced 18 meteorological attributes to 5 influential features, mainly temperature and relative humidity, which were dominant in distinguishing rainfall events. These findings demonstrate that both Random Forest and Logistic Regression outperform Decision Tree, and Random Forest with RFE can be recommended as the most robust model for rainfall prediction in Jakarta.

I. PENDAHULUAN

Prakiraan cuaca adalah salah satu tantangan utama dalam ilmu meteorologi yang sering menjadi fokus berbagai penelitian [1]. Prakiraan cuaca yang tepat sangat membantu dalam pengelolaan kualitas hidup peradaban dengan cara yang lebih efektif dan efisien. Prakiraan cuaca merupakan topik yang menarik untuk dieksplorasi, terutama dengan perkembangan kecerdasan buatan yang semakin banyak diterapkan dalam tengah masyarakat global [2]. Menentukan kondisi cuaca memiliki peran penting karena melibatkan kerja sama antara ilmu pengetahuan dan teknologi dalam menganalisis atmosfer bumi [3].

Hujan adalah salah satu unsur meteorologi yang berpengaruh besar terhadap berbagai aspek kehidupan kita sehari-hari. Banjir merupakan salah satu dampak yang terjadi salah satunya karena tingginya curah hujan yang dapat menyebabkan kerusakan infrastruktur yang perlu diperhatikan [4]. Proses memprediksi curah hujan memiliki kerumitan tersendiri yang diperburuk juga dengan variasi iklim yang ekstrem [5]. Kondisi ini menjadi semakin penting ketika dikaitkan dengan kondisi Jakarta sebagai ibu kota negara dengan kepadatan penduduk yang sangat tinggi. Berdasarkan data Badan Pusat Statistik, Jakarta dihuni oleh lebih dari 10 juta jiwa dengan tingkat urbanisasi yang pesat [6]. Hal ini menjadikan Jakarta sangat rentan terhadap

bencana hidrometeorologi, khususnya banjir, yang hampir setiap tahun terjadi pada musim hujan [7]. Banjir di Jakarta tidak hanya menimbulkan kerugian ekonomi berupa kerusakan infrastruktur, gangguan transportasi, dan penurunan produktivitas [8], tetapi juga dampak sosial yang signifikan, seperti relokasi penduduk, gangguan layanan publik, serta meningkatnya risiko penyakit berbasis lingkungan [9]. Oleh karena itu, upaya meningkatkan akurasi prediksi hujan di Jakarta memiliki relevansi praktis dalam mendukung mitigasi bencana, perencanaan tata kota, serta perlindungan masyarakat secara luas.

Kemajuan teknologi saat ini mendorong metode data mining dapat diterapkan untuk memprediksi data cuaca sehingga menghasilkan tingkat akurasi yang tinggi [10]. *Data mining* merupakan proses penggabungan berbagai teknik untuk menganalisis pola-pola yang penting. Selain itu, dapat juga diartikan sebagai proses memilah data, melakukan observasi dan menghasilkan versi tertentu dari sejumlah data guna menemukan pola-pola yang sering kali tidak disadari keberadaannya. Algoritma *machine learning* secara luas digunakan untuk prediksi, dalam beberapa tahun terakhir. *Machine learning* dapat diartikan sebagai sebuah algoritma yang dirancang untuk mengidentifikasi pola-pola dalam data [11].

Model *machine learning* menawarkan pendekatan yang lebih fleksibel dan adaptif dibandingkan statistik tradisional [12]. Metode statistik tradisional, seperti regresi linier atau model deret waktu (misalnya ARIMA), umumnya efektif untuk data yang bersifat linier dan stasioner. Namun, data curah hujan di wilayah tropis seperti Jakarta cenderung non-linier, bersifat musiman, serta dipengaruhi banyak variabel atmosfer yang saling berinteraksi. Hal ini membuat pendekatan tradisional sering mengalami keterbatasan dalam menangkap pola yang kompleks [13]. Sebaliknya, machine learning memiliki kemampuan mengolah data berukuran besar, menangkap hubungan non-linier, serta memanfaatkan banyak variabel prediktor secara simultan. Algoritma seperti *Random Forest* atau *Support Vector Machine* dapat mengidentifikasi pola tersembunyi yang sulit dijelaskan secara matematis oleh model klasik [14]. Dengan demikian, *machine learning* lebih cocok untuk prediksi curah hujan karena dapat menghasilkan akurasi lebih tinggi sekaligus fleksibilitas dalam mengakomodasi dinamika iklim yang berubah-ubah.

Penelitian yang berkaitan dengan prediksi hujan dengan menggunakan *machine learning* telah banyak diterapkan seperti *Comparative analysis of different rainfall prediction models: A case study of Aligarh City, India* [15], Comparative analysis of machine learning models for rainfall prediction [16], Prediksi Curah Hujan di Kabupaten Rembang dengan Model *Random Forest*[17] , dan Optimasi Klasifikasi Curah Hujan Menggunakan Support Vector Machine (SVM) dan Recursive Feature Elimination (RFE) [18].

Dengan banyaknya fitur meteorologi yang digunakan maka akan dilakukan seleksi fitur untuk mengatasi fitur yang tidak relevan. Seleksi fitur dilakukan dengan cara mengurangi jumlah fitur yang ada dan hanya memilih fitur-fitur yang benar-benar memberikan kontribusi. Pengurangan jumlah fitur tidak akan mengurangi kemampuan diskriminatif, bahkan justru memberikan berbagai keuntungan, seperti mencegah *overfitting*, mengurangi kompleksitas dalam analisis data, serta meningkatkan kinerja analisis data [19]. Terdapat banyak metode dalam seleksi fitur, salah satunya dalah RFE (*Recursive Feature Elimination*). RFE adalah metode yang digunakan untuk memilih fitur-fitur yang paling relevan dari data yang tersedia. Dengan kata lain, RFE mebantu kita mengidentifikasi faktor-faktor yang paling berpengaruh terhadap hasil prediksi. Dengan mengombinasikan kedua metode ini, kita dapat membangun model prediksi yang lebih tepat dan efisien [20], *Decision Tree* mewakili model non-linier yang intuitif namun rentan terhadap *overfitting* [21], sedangkan *Random Forest* mewakili pendekatan ensemble non-linier yang lebih kompleks, stabil, dan mampu mengurangi varians [22]. Dengan demikian, ketiga algoritma ini dipandang cukup representatif untuk membandingkan kinerja model linier, non-linier tunggal, dan non-linier berbasis ensemble dalam prediksi hujan.

Pada penelitian ini dipilih tiga algoritma klasifikasi, yaitu Logistic Regression, Decision Tree, dan Random Forest. Pemilihan ketiga algoritma tersebut didasarkan pada pertimbangan bahwa *Logistic Regression* mewakili model linier yang sederhana dan mudah diinterpretasi [23].

Oleh karena itu, pada penelitian ini dilakukan analisis prediksi curah hujan di Kota Jakarta dengan memanfaatkan machine learning. Beberapa algoritma dicoba untuk mendapatkan nilai akurasi terbaik pada suatu algoritma dalam memperkirakan curah hujan. Algoritma yang digunakan meliputi *Logistic Regression, Random Forest, dan Decision Tree* dengan menggunakan metode RFE sebagai seleksi fitur untuk mengetahui fitur yang berkontribusi dalam meningkatkan hasil akurasi prediksi curah hujan.

II. METODE

Metode Penelitian adalah tahapan sistematis yang dimanfaatkan dalam merencanakan, melaksanakan, dan menganalisa sebuah penelitian. Tahapan penelitian yang dilakukan secara beurutan seperti terlihat pada Gambar 1.

Gambar 1 Alur Metode Penelitian

A. Pengumpulan Data

Data penelitian diambil dari F.Klim71 di Stasiun Meteorologi Kemayoran Provinsi DKI Jakarta tahun 2004-2023. Data tersebut diproses menggunakan 3 (tiga) algoritma dan dibantu menggunakan tools *Google Colab* untuk proses koding. Data yang dikumpulkan mempunyai 18 atribut.

Adapun rincian atribut beserta jenis datanya sebelum pra pemrosesan terlihat pada Tabel 1.

TABEL 1
DAFTAR ATRIBUT DATA SEBELUM DILAKUKAN PEMROSESAN

No	Nama Atribut	Jenis Data
1	TEMPERATURE_07LT_C	float64
2	TEMPERATURE_13LT_C	float64
3	TEMPERATURE_18LT_C	float64
4	TEMPERATURE_AVG_C	float64
5	TEMP_24H_MIN_C	float64
6	TEMP_24H_MAX_C	float64
7	RAINFALL_24H_MM	float64
8	SUNSHINE_24H_H	float64
9	WEATHER_SPECIFIC	object
10	QFF_24H_MEAN_MB	float64
11	REL_HUMIDITY_07LT_PC	float64
12	REL_HUMIDITY_13LT_PC	float64
13	REL_HUMIDITY_18LT_PC	float64
14	REL_HUMIDITY_AVG_PC	float64
15	WIND_SPEED_24H_MAX_MS	float64
16	WIND_DIR_24H_MAX_DEG	float64
17	WIND_SPEED_24H_MEAN_MS	float64
18	WIND_DIR_24H_CARDINAL	object

B. Pre Processing Data

Adapun *Preprocessing* data adalah langkah penting yang harus dijalankan sebelum data dimasukkan untuk klasifikasi. Proses ini adalah tahap untuk mengubah data mentah menjadi data yang bersih dari *noise* sehingga dapat digunakan sebagai input ke dalam model *machine learning*. *Preprocessing* memastikan data yang akan diklasifikasikan bebas dari atribut kosong. Atribut kosong dapat berpengaruh pada hasil klasifikasi dan akurasi [24]. Pada penelitian ini, *preprocessing* data yang dilakukan diantaranya:

- Memeriksa *time series* berdasarkan index timestamp dan menangani Null, Spesial Value dan Categorical.
- Membuat kategori “hujan hari ini” dengan 0 (tidak hujan) dan 1 (Hujan). Variabel kategori “Hujan Hari ini” dan “Hujan Esok” apabila “Tidak Hujan” dinyatakan dengan kode 0 dan “Ya, Hujan” dinyatakan dengan kode 1
- Melakukan penghapusan (*drop*) atribut kosong dan outlier agar tidak menghambat proses prediksi nantinya dengan menggunakan metode IQR.
- Melakukan pemisahan fitur dan target

C. Seleksi dan Eliminasi Fitur

Seiring dengan banyaknya fitur yang digunakan dalam aplikasi *machine learning* saat ini, para peneliti sering kali kesulitan untuk memahami bagaimana kontribusi masing-masing fitur terhadap hasil klasifikasi. Namun, saat memproses data dengan berbagai fitur yang digunakan, ada peluang bahwa beberapa fitur tidak relevan atau redundan. Seiring dengan munculnya berbagai metode untuk menyelesaikan permasalahan tersebut maka diterapkan seleksi dan eliminasi fitur guna memahami data, mengoptimalkan kebutuhan komputasi dan meningkatkan akurasi prediksi [25]. Penelitian ini memilih metode seleksi RFE (*Recursive Feature Eliminiation*).

Metode *Recursive Feature Elimination* (RFE) bekerja dengan melakukan pemrosesan berulang, di mana fitur diberi peringkat berdasarkan tingkat kontribusinya terhadap hasil prediksi. Setiap iterasi, tingkat kepentingan masing-masing fitur akan dievaluasi, kemudian fitur yang dianggap tidak relevan akan dihapus. Model klasifikasi baru kemudian dilatih menggunakan fitur yang tersisa [26]. Metode RFE adalah salah satu teknik untuk mengeliminasi fitur secara otomatis. Metode ini sangat bermanfaat ketika jumlah fitur yang digunakan untuk melatih model sangat besar, sehingga sulit untuk eliminasi fitur secara manual. Pada penelitian ini, metode RFE yang digunakan meliputi algoritma:

a. *Logistic Regression*
b. *SVM Classifier*
c. *Random Forest Calssifier*

D. Pemodelan Machine Learning

Pemodelan *Machine Learning* pada penelitian ini menggunakan 3 (tiga) model klasifikasi yaitu *Logistic Regression, Decision Tree,* dan *Random Forest.*

Logistic Regression

Logistic Regression merupakan suatu metode analisis statistika yang bertujuan untuk mengetahui pengaruh suatu variabel respon (*Y*) terhadap variabel penduga (*X*). Regresi logistik diterapkan khusus untuk kasus di mana variabel respon (Y) merupakan variabel kualitatif yang bersifat biner atau dikotom, yang hanya memiliki dua kemungkinan nilai, yaitu "ketika hasilnya terjadi" (Y=1) atau "ketika hasilnya tidak terjadi" (Y=0) [23]. Adapun persamaan regresi logistik adalah:

$$\pi(x_i) = \frac{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)} \quad (1)$$

dengan k = jumlah parameter (variabel bebas)

Decision tree

Decision Tree merupakan salah satu metode yang relatif mudah dipahami dan diinterpretasikan oleh manusia [21]. *Decision Tree* adalah suatu metode untuk mengidentifikasi pola-pola atau fungsi-fungsi yang menggambarkan dan memisahkan berbagai kelas data, yang kemudian dapat digunakan untuk memprediksi kelas data pada data yang belum terklasifikasi [27].

Random Forest

Random Forest adalah kumpulan pohon keputusan yang dibangun dengan menggunakan sampel acak, namun memiliki aturan pemisahan simpul yang berbeda-beda. Model ini bekerja dengan memilih subset fitur untuk setiap pohon,

kemudian mencari batas ambang terbaik untuk memisahkan data. Akibatnya, model ini menghasilkan sejumlah pohon yang dilatih dengan pendekatan yang lebih sederhana, dan setiap pohon memberikan prediksi yang berbeda. Hasil dari seluruh pohon ini dapat dianalisis dengan dua cara, dimana cara yang paling sering digunakan adalah berdasarkan suara terbanyak, yang kemudian dianggap sebagai kelas yang benar. Namun, dalam implementasi algoritma di *scikit-learn*, hasil prediksi dihitung berdasarkan rata-rata dari seluruh prediksi tersebut, sehingga menghasilkan hasil yang sangat akurat [22].

E. Evaluasi Kinerja Model

Pada penelitian ini model dilatih dengan metode *cross-validation* dengan 5fold, bukan dengan pembagian sederhana *train-test split*. Teknik ini dipilih karena lebih stabil, mampu memanfaatkan seluruh data untuk pelatihan dan validasi secara bergantian, serta mengurangi risiko *overfitting*. *Cross validation* merupakan teknik validasi dalam machine learning yang digunakan untuk mengevaluasi performa model dengan lebih akurat, menghindari *overfitting*, serta memastikan model dapat melakukan generalisasi dengan baik. Dalam metode *cross-validation* dengan 5fold ini, data latih dibagi menjadi 5 subset (biasanya disebut "fold") yang masing-masing memiliki ukuran yang sama. Pada setiap iterasi validasi, satu fold digunakan sebagai data validasi, sementara sisanya digunakan sebagai data latih. Proses ini diulang sebanyak 5 kali, sehingga setiap fold menjadi data validasi satu kali sehingga pada setiap pelatihan, 80% data digunakan sebagai data training dan 20% sisanya digunakan validasi.

Selanjutnya, untuk menemukan algoritma terbaik, model klasifikasi ini akan dievaluasi menggunakan matrik kinerja berikut:

a. *Accuracy*: perbandingan antara prediksi *True* (baik positif maupun negatif) terhadap keseluruhan data.
b. *Recall*: perbandingan prediksi *True* positif terhadap keseluruhan data yang benar positif.
c. *Precission*: perbandingan prediksi *True* positif terhadap keseluruhan hasil yang diprediksi positif.
d. F1 *Score*: perbandingan antara rata-rata *precission* dan *recall* yang telah diberi bobot.
e. ROC: Grafik yang menggambarkan kemampuan diagnostik dari sistem pengklasifikasi biner ini dapat dilihat melalui variasi ambang diskriminasi. ROC juga bisa dianggap sebagai sebuah kurva yang menunjukkan kekuatan sistem sebagai fungsi dari Kesalahan Tipe I, berdasarkan aturan keputusan (ketika kinerja dihitung hanya dari sampel populasi, yang dapat dianggap sebagai perkiraan dari jumlah tersebut). Kinerja akurasi AUC dapat dikategorikan dalam beberapa kelompok yaitu [28]:
    1. 0,90 – 1,00 = Kinerja Sangat Baik
    2. 0,80 – 0,90 = Kinerja Baik
    3. 0,70 – 0,80 = Kinerja Cukup Baik
    4. 0,60 – 0,70 = Kinerja Buruk
    5. 0,50 – 0,60 = Kinerja Gagal
f. Cohen's Kappa: sebuah uji statistik yang digunakan untuk menentukan tingkat kesepakatan antara dua perkiraan yang berbeda dari variabel respon. Koefisien Kappa Cohen memiliki rentang nilai antara -1 hingga 1, di mana nilai 1 menandakan kesepakatan sempurna, nilai 0 menunjukkan tingkat kesepakatan yang terjadi secara kebetulan, dan nilai negatif menunjukkan tingkat kesepakatan yang lebih rendah dari yang diharapkan secara kebetulan. Nilai Cohen Kappa dihitung menggunakan Persamaan (2).

$$K = \frac{N \sum_{i=1}^{k} m_{ij} - \sum_{i=1}^{n} G_i C_i}{N^2 - \sum_{i=1}^{n} (G_i C_i)} \quad (2)$$

III. HASIL DAN PEMBAHASAN

Berikut ini adalah hasil dari setiap tahapan yang dilakukan dalam penelitian ini dimulai dari Pengumpulan data, *Pre Processing Data,* Seleksi dan Eliminasi Fitur, Pemodelan *machine Learning*, dan Evaluasi Kinerja model. Data kosong dan *special value* pada masing-masing atribut yang diperoleh dapat dilihat pada Tabel 2.

TABEL 2
BANYAKNYA DATA KOSONG DAN SPECIAL VALUE

No	Nama Atribut	Banyaknya Data Kosong dan Spesial Value
1	TEMPERATURE_07LT_C	3
2	TEMPERATURE_13LT_C	8
3	TEMPERATURE_18LT_C	9
4	TEMPERATURE_AVG_C	14
5	TEMP_24H_MIN_C	2
6	TEMP_24H_MAX_C	2
7	RAINFALL_24H_MM	422
8	SUNSHINE_24H_H	98
9	WEATHER_SPECIFIC	5
10	QFF_24H_MEAN_MB	2
11	REL_HUMIDITY_07LT_PC	3
12	REL_HUMIDITY_13LT_PC	8
13	REL_HUMIDITY_18LT_PC	10
14	REL_HUMIDITY_AVG_PC	15
15	WIND_SPEED_24H_MAX_MS	2
16	WIND_DIR_24H_MAX_DEG	2
17	WIND_SPEED_24H_MEAN_MS	2
18	WIND_DIR_24H_CARDINAL	2

Selanjutnya, untuk menangani data kosong dan *special value* tersebut maka dilakukan:

a. Penghapusan data *time series* yang hilang
b. Untuk data RAINFALL_24_MM : mengganti 9999 dengan 0; 8888 dengan 0,1; Null dengan 0
c. Untuk nilai Null pada *categorical_features* diisi dengan nilai modus
d. Untuk nilai Null pada *numerical_features* disi dengan *Multivariate Imputation by Chained Equations* (MICE), dimana MICE digunakan dalam acuan

statistika sebagai salah satu metode untuk menangani *missing data*.

Setelah tidak ada data kosong, yaitu dengan membuat kategori RAINFALL_TODAY dari RAINFALL_24H_MM dengan 0 jika tidak hujan dan 1 jika terjadi hujan pada kolom baru.

Langkah selanjutnya yaitu menangani data *outlier* dengan metode IQR. Hasil distribusi plot yang diperoleh terhadap analisis ditemukannya data outlier yang digambarkan dalam tabel 3.

TABEL 3
BANYAKNYA OUTLIER

No	Atribut	Banyaknya *outlier*
1	TEMPERATURE_07LT_C	23
2	TEMPERATURE_13LT_C	129
3	TEMPERATURE_18LT_C	218
4	TEMPERATURE_AVG_C	16
5	TEMP_24H_MIN_C	1
6	TEMP_24H_MAX_C	50
7	RAINFALL_24H_MM	-
8	SUNSHINE_24H_H	0
9	WEATHER_SPECIFIC	-
10	QFF_24H_MEAN_MB	22
11	REL_HUMIDITY_07LT_PC	16
12	REL_HUMIDITY_13LT_PC	62
13	REL_HUMIDITY_18LT_PC	67
14	REL_HUMIDITY_AVG_PC	5
15	WIND_SPEED_24H_MAX_MS	140
16	WIND_DIR_24H_MAX_DEG	0
17	WIND_SPEED_24H_MEAN_MS	66
18	WIND_DIR_24H_CARDINAL	-

Setelah melalui tahap pembersihan data kosong, nilai hilang, serta *outlier* dengan metode IQR, diperoleh total 2.836 *record* data bersih yang siap digunakan pada tahap pemodelan. Jumlah ini merupakan hasil akhir dari keseluruhan proses preprocessing dan menjadi basis input bagi algoritma machine learning.

Setelah dilakukan pembersihan, kemudian dilakukan seleksi dan eliminasi fitur dengan menggunakan metode RFE. Algoritma RFE yang diterapkan dalam penelitian ini adalah *Logistic Regression, SVM Classifier,* dan *Random Forest Classifier*. Adapun 5 (lima) atribut yang paling berpengaruh dalam prediksi hujan pada masing-masing algoritma RFE dapat dilihat pada Tabel 4.

TABEL 4
HASIL DARI 3 (TIGA) ALGORITMA RFE

No	*Logistic Regression*	*SVC*	*Random Forest Classifier*
1	TEMPERATURE_07LT_C	TEMPERATURE_07LT_C	TEMPERATURE_AVG_C
2	TEMPERATURE_13LT_C	TEMPERATURE_13LT_C	TEMP_24H_MIN_C
3	TEMP_24H_MIN_C	TEMP_24H_MIN_C	REL_HUMIDITY_13LT_PC
4	REL_HUMIDITY_18LT_PC	REL_HUMIDITY_18LT_PC	REL_HUMIDITY_18LT_PC
5	REL_HUMIDITY_AVG_PC	REL_HUMIDITY_AVG_PC	REL_HUMIDITY_AVG_PC

Dari 3 (tiga) algoritma RFE tersebut, terlihat bahwa pada algoritma RFE *Logistic Regression* dan SVC memiliki 5 (lima) atribut berpengaruh yang sama, sedangkan untuk algoritma *Random Forest Classifier* menghasilkan 5 (lima) atribut yang berbeda. Dengan demikian, dapat disimpulkan bahwa metode RFE mampu mereduksi jumlah fitur dari total 18 atribut awal menjadi 5 atribut utama yang paling relevan terhadap prediksi curah hujan.

Secara lebih rinci, fitur dominan yang dipilih oleh *Logistic Regression* dan *SVC* meliputi suhu udara jam 07 (TEMPERATURE_07LT_C), suhu udara jam 13 (TEMPERATURE_13LT_C), suhu minimum harian (TEMP_24H_MIN_C), kelembapan udara jam 18 (REL_HUMIDITY_18LT_PC), dan kelembapan relatif rata-rata (REL_HUMIDITY_AVG_PC). Sementara itu, pada *Random Forest Classifier*, fitur utama yang terpilih adalah suhu udara rata-rata harian (TEMPERATURE_AVG_C), suhu udara minimum (TEMP_24H_MIN_C), kelembapan relatif jam 13 (REL_HUMIDITY_13LT_PC), kelembapan relatif udara (REL_HUMIDITY_18LT_PC), dan kelembapan relatif rata-rata (REL_HUMIDITY_AVG_PC).

Hasil ini menunjukkan bahwa variabel suhu udara dan kelembapan relatif merupakan faktor paling berpengaruh dalam membedakan kondisi hujan dan tidak hujan di Jakarta. Dominasi kedua faktor ini sejalan dengan sifat termodinamika atmosfer, di mana fluktuasi suhu dan kelembapan sangat menentukan proses kondensasi awan hingga terjadinya hujan. Reduksi fitur dari 18 menjadi 5 tidak hanya menyederhanakan model, tetapi juga meningkatkan efisiensi komputasi, mencegah *overfitting*, serta tetap mempertahankan kemampuan prediksi yang baik.

Analisis Performa Model

Proses selanjutnya adalah pengolahan pemodelan klasifikasi *machine learning* untuk menentukan kinerja algoritma terbaik yang ditunjukkan pada Tabel 5.

TABEL 5
NILAI MATRIKS EVALUASI SETIAP ALGORITMA

Algoritma	*Matrik Evaluasi*	*Raw Data*	*RFE Logistic Regression*	*RFE SVC*	*RFE Random Forest*
Logistic Regression	*Accuracy*	0.7627	0.7684	0.7684	0.7648
	Recall	0.6259	0.6239	0.6239	0.6200
	Precision	0.7066	0.7200	0.7200	0.7163
	F1 *Score*	0.6576	0.6620	0.6620	0.6573
	ROC AUC	0.8299	0.8294	0.8294	0.8260
	Cohen's Kappa	0.4775	0.4875	0.4875	0.4802
Decision Tree	*Accuracy*	0.6890	0.6788	0.6843	0.6791
	Recall	0.5746	0.5775	0.5835	0.5943
	Precision	0.5693	0.5574	0.5648	0.5516
	F1 *Score*	0.5693	0.5641	0.5706	0.5702

	ROC AUC	0.6636	0.6563	0.6612	0.6603
	Cohen's Kappa	0.3266	0.3114	0.3214	0.3156
Random Forest	*Accuracy*	0.7574	0.7662	0.7648	0.7528
	Recall	0.5875	0.6299	0.6269	0.6161
	Precision	0.7043	0.7051	0.7030	0.6859
	F1 *Score*	0.6358	0.6592	0.6574	0.6418
	ROC AUC	0.8171	0.8108	0.8098	0.8044
	Cohen's Kappa	0.4560	0.4829	0.4798	0.4548

Perbedaan performa antar algoritma dapat dipahami dari karakteristik masing-masing model serta nilai evaluasi yang diperoleh. *Random Forest* menunjukkan performa paling unggul dengan akurasi tertinggi sebesar 0.7622, *F1 score* sekitar 0.65, *ROC AUC* antara 0.8044–0.8171, dan *Cohen's Kappa* mendekati 0.48. Nilai ini menandakan kemampuan model yang baik dalam membedakan kelas serta konsistensi prediksi. Keunggulan *Random Forest* berasal dari sifatnya sebagai metode *ensemble* yang menggabungkan banyak pohon keputusan, sehingga mampu menangkap interaksi non-linear antar variabel meteorologi dan mengurangi risiko overfitting.

Sementara itu, *Logistic Regression* juga menunjukkan hasil kompetitif dengan akurasi 0.7648, *ROC AUC* mencapai 0.829, dan *Cohen's Kappa* sekitar 0.49. Meskipun berbasis model linier, *Logistic Regression* tetap mampu memberikan prediksi yang cukup baik karena variabel suhu dan kelembapan yang dominan masih memiliki hubungan linier dengan peluang terjadinya hujan. Namun, keterbatasannya dalam menangkap pola non-linear membuat kinerjanya sedikit di bawah *Random Forest*.

Berbeda dengan keduanya, *Decision Tree* menghasilkan nilai yang relatif lebih rendah, dengan akurasi 0.6890, F1 *score* sekitar 0.57, dan *ROC AUC* sebesar 0.6636. Hal ini menunjukkan bahwa *Decision Tree* cenderung kurang stabil dan lebih rentan terhadap *overfitting*, sehingga tidak sehandal *Random Forest* maupun *Logistic Regression* dalam memprediksi hujan.

Dengan demikian, keunggulan *Random Forest* dibanding dua model lainnya bukan hanya terlihat dari nilai akurasi dan ROC AUC yang lebih tinggi, tetapi juga dari kemampuannya menangkap hubungan non-linear kompleks antar fitur serta menyeimbangkan bias dan varians. Faktor-faktor ini menjadikan Random Forest sebagai model yang paling stabil dan akurat untuk prediksi hujan di Jakarta setelah penerapan RFE.

.

IV. KESIMPULAN

Penelitian ini membandingkan kinerja tiga algoritma machine learning *Logistic Regression*, *Decision Tree*, dan *Random Forest* dalam memprediksi hujan di Jakarta dengan menggunakan data cuaca dari Stasiun Meteorologi Kemayoran periode 2004–2023. Metode *Recursive Feature Elimination* (RFE) diterapkan untuk menyeleksi variabel paling relevan, sehingga jumlah atribut berkurang dari 18 menjadi 5 fitur utama yang didominasi oleh variabel suhu udara dan kelembapanudara.

Hasil evaluasi menunjukkan bahwa *Random Forest* memberikan performa terbaik dengan akurasi 0.7622, *ROC AUC* 0.8044–0.8171, *F1 score* sekitar 0.65, dan C*ohen's Kappa* mendekati 0.48. *Logistic Regression* juga tampil kompetitif dengan akurasi 0.7648 dan *ROC AUC* 0.829, sedangkan *Decision Tree* menunjukkan performa lebih rendah (akurasi 0.6890; *ROC AUC* 0.6636). Analisis lebih lanjut menegaskan bahwa keunggulan *Random Forest* berasal dari kemampuannya menangkap hubungan non-linear kompleks antar fitur dan menyeimbangkan bias serta varians, sehingga lebih stabil dibandingkan *Logistic Regression* yang terbatas pada pola linier maupun *Decision Tree* yang rentan overfitting.

Temuan ini membuktikan bahwa penerapan RFE dapat meningkatkan efisiensi dan akurasi prediksi hujan, serta memperjelas fitur meteorologi paling berpengaruh. Secara praktis, model *Random Forest* dengan RFE dapat direkomendasikan sebagai pendekatan yang lebih andal untuk mendukung sistem peringatan dini dan mitigasi risiko banjir di Jakarta. Penelitian selanjutnya dapat memperluas pendekatan dengan menambahkan algoritma lain, menguji data multivariat dari beberapa stasiun, serta mengintegrasikan faktor eksternal seperti topografi dan tata guna lahan untuk meningkatkan robustnes prediksi.

DAFTAR PUSTAKA

[1] R. Prasetya, "Penerapan Teknik Data Mining Dengan Algoritma Classification Tree untuk Prediksi Hujan," *J. Widya Climago*, vol. 2, no. 2, pp. 13–23, 2020.

[2] I. P. Putri, T. Terttiaavini, and N. Arminarahmah, "Analisis Perbandingan Algoritma Machine Learning untuk Prediksi Stunting pada Anak," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 1, pp. 257–265, 2024, doi: 10.57152/malcom.v4i1.1078.

[3] R. Herwanto, R. F. Purnomo, and S. Sriyanto, "Rainfall Prediction Using Data Mining Techniques - A Survey," *3rd Int. Conf. Inf. Technol. Bus.*, pp. 188–193, 2017, doi: 10.5121/csit.2013.3903.

[4] T. T. Le, B. T. Pham, H. B. Ly, A. Shirzadi, and L. M. Le, "Development of 48-hour precipitation forecasting model using nonlinear autoregressive neural network," *Lect. Notes Civ. Eng.*, vol. 54, pp. 1191–1196, 2020, doi: 10.1007/978-981-15-0802-8_191.

[5] G. de Colombia, "Smart Cities - SMART CITIES," *Res. Gate*, pp. 1–51, 2017, [Online]. Available: https://bibliotecadigital.fgv.br/dspace/handle/10438/18386

[6] Badan Pusat Statistik, "Statistik Daerah Provinsi DKI Jakarta 2023," BPS Provinsi DKI Jakarta. [Online]. Available: https://jakarta.bps.go.id/publication

[7] BNPB, "Laporan Tahunan Penanggulangan Bencana 2023," Badan Nasional Penanggulangan Bencana. [Online]. Available: https://bnpb.go.id/publikasi

[8] Y. O. Izadkhah and L. Gibbs, "A study of preschoolers' perceptions of earthquakes through drawing," *Int. J. Disaster Risk Reduct.*, vol. 14, pp. 132–139, 2015, doi: 10.1016/j.ijdrr.2015.06.002.

[9] L. Mastronardi and A. Cavallo, "The spatial dimension of income inequality: An analysis at municipal level," *Sustain.*, vol. 12, no. 4,

pp. 1–18, 2020, doi: 10.3390/su12041622.
[10] I. D. Sulistyowati, S. Sunarno, and D. Djuniadi, "Penerapan Machine Learning Dengan Algoritma Support Vector Machine Untuk Prediksi Kelembapan Udara Rata-Rata," *Just IT J. Sist. Informasi, Teknol. Inf. dan Komput.*, vol. 15, no. 1, pp. 284–290, 2024.
[11] S. Chen *et al.*, "Automated poisoning attacks and defenses in malware detection systems: An adversarial machine learning approach," *Comput. Secur.*, vol. 73, no. Bo Li, pp. 326–344, 2018, doi: 10.1016/j.cose.2017.11.007.
[12] A. Sutaryani, S. Sunarno, and D. Djuniadi, "Perbandingan Performa Model Machine Learning dalam Prediksi Suhu di Semarang," *JITET (Jurnal Inform. dan Tek. Elektro Ter.*, vol. 12, no. 3, pp. 2770–2775, 2024.
[13] P. G. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, 2003, doi: 10.1016/S0925-2312(01)00702-0.
[14] A. Y. Barrera-Animas, L. O. Oyedele, M. Bilal, T. D. Akinosho, J. M. D. Delgado, and L. A. Akanbi, "Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting," *Mach. Learn. with Appl.*, vol. 7, no. August 2021, p. 100204, 2022, doi: 10.1016/j.mlwa.2021.100204.
[15] M. Usman Saeed Khan, K. Mohammad Saifullah, A. Hussain, and H. Mohammad Azamathulla, "Comparative analysis of different rainfall prediction models: A case study of Aligarh City, India," *Results Eng.*, vol. 22, no. January, p. 102093, 2024, doi: 10.1016/j.rineng.2024.102093.
[16] P. K. Das, R. L. Sahu, and P. C. Swain, "Comparative analysis of machine learning models for rainfall prediction," *J. Atmos. Solar-Terrestrial Phys.*, vol. 264, no. August, p. 106340, 2024, doi: 10.1016/j.jastp.2024.106340.
[17] G. Fibarkah, M. A. Tondang, N. W. Yulistyaningrum, and M. Afrad, "Prediksi Curah Hujan di Kabupaten Rembang dengan Model Random Forest," no. Ml, pp. 863–871, 2024.
[18] A. R. I. Pratama, S. A. Latipah, and B. N. Sari, "Optimasi Klasifikasi Curah Hujan Menggunakan Support Vector Machine (Svm) Dan Recursive Feature Elimination (Rfe)," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 7, no. 2, pp. 314–324, 2022, doi: 10.29100/jipi.v7i2.2675.
[19] E. S. Wahyuni, "Penerapan Metode Seleksi Fitur Untuk Meningkatkan Hasil Diagnosis Kanker Payudara," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 7, no. 1, p. 283, 2016, doi: 10.24176/simet.v7i1.516.
[20] F. M. Herza, B. Rahmat, M. Muharrom, and A. L. Haromainy, "Pengaruh Rfe Terhadap Logistic Regression Dan Support Vector Machine Pada Analisis Sentimen Hotel Shangri-La Surabaya," vol. 8, no. 6, pp. 11612–11619, 2024.
[21] I. Sutoyo, "Implementasi Algoritma Decision Tree Untuk Klasifikasi Data Peserta Didik," *J. Pilar Nusa Mandiri*, vol. 14, no. 2, p. 217, 2018, doi: 10.33480/pilar.v14i2.926.
[22] D. P. Sinambela, H. Naparin, M. Zulfadhilah, and N. Hidayah, "Implementasi Algoritma Decision Tree dan Random Forest dalam Prediksi Perdarahan Pascasalin," *J. Inf. dan Teknol.*, vol. 5, no. 3, pp. 58–64, 2023, doi: 10.60083/jidt.v5i3.393.
[23] D. Kusrini, Dwi Endah; Puspitasari, "Penggunaan Analisis Regresi Logistik Untuk Menganalisis Perilaku Dan Faktor-Faktor Yang Mempengaruhi Minat Baca Pengunjung Badan Perpustakaan Propinsi Jawa Timur," *J. Mat.*, vol. Vol.9 No.1, pp. 149–155, 2006.
[24] Bertalya, Prihandoko, L. Setyowati, F. I. Irawan, and S. R. Irlianti, "Formulation of city health development index using data mining," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 23, no. 1, pp. 362–369, 2021, doi: 10.11591/ijeecs.v23.i1.pp362-369.
[25] M. S. Wibawa and K. D. P. Novianti, "Reduksi Fitur untuk Optimalisasi Klasifikasi Tumor Payudara Berdasarkan Data Citra FNA," *Konf. Nas. Sist. Inform.*, pp. 73–78, 2017.
[26] M. S. Wibawa, H. A. Nugroho, and N. A. Setiawan, "Performance evaluation of combined feature selection and classification methods in diagnosing Parkinson disease based on voice feature," *Proc. - 2015 Int. Conf. Sci. Inf. Technol. Big Data Spectr. Futur. Inf. Econ. ICSITech 2015*, no. July, pp. 126–131, 2016, doi: 10.1109/ICSITech.2015.7407790.
[27] A. Shiddiq, R. K. Niswatin, and I. N. Farida, "Ahmad Shiddiq Analisa Kepuasan Konsumen Menggunakan Klasifikasi Decision Tree Di Restoran Dapur Solo (Cabang Kediri)," *Gener. J.*, vol. 2, no. 1, p. 9, 2018, doi: 10.29407/gj.v2i1.12051.
[28] A. Purwanto, "Jurnal Teknoinfo," *Tong Sampah Pint. Dengan Perintah Suara Guna Menghilangkan Perilaku Siswa Membuang Sampah Sembarangan Di Sekol.*, vol. 14, pp. 48–58, 2020, [Online]. Available: https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/article/view/336/329