BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Yang Relevan

Penelitian yang relevan dilakukan untuk membandingkan penelitian sebelumnya yang terkait dengan penelitian ini dan memperkuat posisi penelitian ini sendiri. Penelitian pembanding akan membahas subjek yang diteliti, teknik yang digunakan, dan hasil yang diperoleh. Berikut merupakan beberapa penelitian terdahulu.

Penelitian Pertama,oleh (Intan et al., 2021)dengan judul "**Analisis Performansi Prakiraan Cuaca Menggunakan Algoritma Machine Learning**", yang dipublikasikan dalam Jurnal Pekommas, mengkaji kinerja dua algoritma berbasis jaringan saraf tiruan Backpropagation dan Bayesian Regularization dalam mengklasifikasikan kondisi cuaca di wilayah Makassar, Indonesia, yang memiliki iklim tropis mirip dengan Yogyakarta. Data meteorologi yang digunakan mencakup parameter suhu, tekanan udara, kelembapan, tutupan awan, kecepatan angin, dan curah hujan, dengan resolusi tiga jam per hari selama periode 2016–2018.

Studi ini membagi kondisi cuaca menjadi empat kelas: cerah, berawan, hujan ringan, dan hujan lebat, serta mengevaluasi model berdasarkan Mean Squared Error (MSE), akurasi, dan waktu komputasi. Hasil penelitian menunjukkan bahwa Bayesian Regularization mengungguli Backpropagation dengan meraih akurasi hingga 99,70% pada musim hujan dan 99,06% pada musim kemarau, serta waktu komputasi yang jauh lebih cepat. Selain itu, penelitian ini menyoroti tantangan utama dalam prediksi cuaca tropis, yaitu distribusi data yang tidak seimbang antar kelas, yang dapat menyebabkan fluktuasi performa model terutama pada kategori cuaca ekstrem seperti hujan lebat.

Penelitian Kedua, oleh (Zian Asti Dwiyanti, 2023)"**Prediksi Cuaca Kota Jakarta Menggunakan Metode Random Forest: Studi**

Optimalitas” mengkaji penerapan algoritma Random Forest untuk mengklasifikasikan kondisi cuaca di wilayah Jakarta berdasarkan data historis dari OpenData Jakarta tahun 2018. Dataset awal terdiri dari 8.535 entri dengan 27 kategori kondisi cuaca, yang kemudian dinormalisasi menjadi 8 kelas utama, termasuk cerah, cerah berawan, berawan, hujan ringan, hujan sedang, hujan petir, dan lainnya. Proses normalisasi ini dilakukan untuk mengatasi inkonsistensi penulisan label (misal: “Cerah Berawan” vs “Berawan”), yang merupakan tantangan umum dalam pengolahan data cuaca sekunder di Indonesia.

Model Random Forest dilatih menggunakan fitur suhu (minimum dan maksimum), kelembapan, wilayah, waktu pengamatan (pagi/siang/malam), serta pola musiman. Evaluasi dilakukan dengan pembagian data 80% untuk pelatihan dan 20% untuk pengujian, dengan metrik utama meliputi accuracy, precision, recall, F1-score, dan ROC-AUC. Hasil penelitian menunjukkan bahwa model mencapai akurasi sebesar 0,71, dengan ROC-AUC yang sangat tinggi yaitu 0,92, mengindikasikan kemampuan model yang kuat dalam membedakan antar kelas cuaca meskipun distribusi data tidak seimbang.

Penelitian Ketiga, oleh (Ashari Rakhmat & Mutohar, 2023)berjudul **“Prakiraan Hujan menggunakan Metode Random Forest dan Cross Validation”** mengembangkan model prediksi curah hujan berbasis data historis BMKG periode 2015–2021. Studi ini menggunakan 10 variabel meteorologi, termasuk suhu minimum (Tn), suhu maksimum (Tx), kelembapan rata-rata (RH_avg), lama penyinaran matahari (ss), dan kecepatan angin (ff_x) yang secara substansial mirip dengan variabel yang digunakan dalam penelitian ini.

Model Random Forest dalam penelitian ini dikombinasikan dengan teknik k-fold cross-validation untuk menghindari overfitting dan meningkatkan generalisasi. Melalui eksperimen komprehensif terhadap hyperparameter seperti n_estimators dan max_depth, ditemukan bahwa konfigurasi n_estimators = 100, max_depth = None, dan k = 3 pada

cross-validation menghasilkan performa paling optimal, dengan MSE = 0,086, RMSE = 0,290, dan MAE = 0,186. Lebih penting lagi, penelitian ini menunjukkan bahwa Random Forest yang dikombinasikan dengan cross-validation jauh lebih stabil dan akurat dibandingkan Random Forest tanpa validasi silang.

Penelitian Keempat,oleh (Abdurrasyid et al., 2021) mengembangkan sistem prediksi kuota pemesanan bahan bakar di Stasiun Pengisian Bahan Bakar Umum (SPBU) menggunakan metode Regresi Linear Berganda. Penelitian ini menggunakan dua variabel independen utama yaitu stok sisa (residual stock) dan stok masuk (incoming stock) untuk memprediksi kebutuhan stok harian, dengan tujuan mencegah kondisi out of stock yang dapat mengganggu operasional dan pendapatan SPBU. Evaluasi model dilakukan menggunakan metrik Mean Absolute Percentage Error (MAPE), dan hasil menunjukkan error sebesar 11,0% untuk bahan bakar Pertalite dan 13,2% untuk Solar, yang mengindikasikan performa prediksi yang cukup akurat.

Meskipun konteks aplikasinya berbeda, penelitian ini relevan karena menekankan pentingnya penggunaan data historis time-series dan validasi ketat terhadap asumsi model dalam membangun sistem prediksi berbasis data riil. Selain itu, pendekatan preprocessing seperti penanganan missing value dan transformasi data serta fokus pada akurasi praktis (bukan hanya teoretis) selaras dengan metodologi yang diterapkan dalam penelitian ini. Temuan (Abdurrasyid et al., 2021)memperkuat argumen bahwa model statistik sederhana dapat memberikan prediksi yang andal jika didukung oleh kualitas data dan desain eksperimen yang baik, sebuah prinsip yang juga menjadi fondasi penggunaan Random Forest dalam prediksi cuaca harian di wilayah DIY.

Penelitian Kelima,oleh (Kentjie et al., 2026)dalam penelitian berjudul “Prediksi Degradasi Kesehatan Baterai Kendaraan Listrik Berbasis Machine Learning pada Pengisian Cepat DC” mengembangkan model

Random Forest Regressor untuk memprediksi degradasi kesehatan baterai (ΔSoH) berdasarkan parameter pengisian cepat DC, yaitu arus pengisian, tegangan DC, suhu baterai, dan State of Charge (SoC). Penelitian ini menggunakan dataset hasil simulasi sebanyak 1.500+ sampel, dilatih dengan Time-Series Split, dan dievaluasi menggunakan metrik MAE, RMSE, dan R^2. Hasil menunjukkan bahwa Random Forest mampu mencapai $R^2 = 0.999998$, MAE = 9×10^{-8}, dan RMSE 1.5×10^{-7}, yang mengindikasikan tingkat akurasi prediksi yang sangat tinggi. Selain itu, analisis feature importance menunjukkan bahwa suhu baterai dan arus pengisian merupakan dua fitur paling dominan dalam memengaruhi degradasi kesehatan baterai temuan yang konsisten dengan prinsip fisika baterai lithium-ion. Penelitian ini relevan karena menegaskan bahwa Random Forest tidak hanya efektif dalam tugas regresi, tetapi juga mampu menangkap hubungan non-linear kompleks dalam data time-series berdimensi rendah, serta memberikan interpretasi melalui feature importance karakteristik yang secara langsung mendukung pendekatan Anda dalam memprediksi cuaca harian berbasis enam variabel meteorologis.

Penelitian Keenam, oleh (Ainul Idham & Efy Yosrita, 2025), dalam studi berjudul "Analysis of Long Short-Term Memory (LSTM) and Extreme Gradient Boosting (XGBoost) Algorithms to Predict the Number of Airplane Passengers…", melakukan Systematic Literature Review terhadap 11 penelitian terkait prediksi time-series. Meskipun judulnya menyebutkan LSTM dan XGBoost, salah satu temuan penting yang relevan bagi penelitian ini adalah peran Random Forest sebagai model baseline yang robust dalam konteks data time-series berisik. Dalam beberapa studi yang dikaji khususnya Sahar Yassine & Aleksander Stanulov (2023) Random Forest digunakan sebagai pembanding dalam prediksi arus penumpang bandara Oslo, dan mencapai RMSE = 0.07368 serta MSE = 0.00543, menunjukkan kinerja yang stabil meskipun tidak unggul atas LSTM dalam skenario dengan pola musiman tinggi.

Temuan ini sangat relevan karena menegaskan bahwa Random Forest tetap efektif dalam mengatasi noise, outlier, dan ketidakseimbangan data karakteristik umum pada dataset cuaca harian tropis seperti di DIY. Keunggulan utamanya terletak pada

Ketahanan terhadap overfitting melalui mekanisme bagging dan feature randomness, Interpretabilitas tinggi melalui analisis feature importance, Efisiensi komputasi tanpa kebutuhan tuning hyperparameter intensif.

Dengan demikian, Random Forest bukan sekadar model alternatif, melainkan pilihan strategis untuk sistem prediksi berbasis data operasional yang mengutamakan keandalan, transparansi, dan kecepatan implementasi kriteria yang selaras dengan tujuan penelitian ini dalam membangun model klasifikasi cuaca harian yang praktis dan dapat diadopsi oleh sektor pertanian, pariwisata, dan mitigasi bencana di Daerah Istimewa Yogyakarta.

Penelitian Ketujuh, oleh(Arnah Ritonga et al., 2025) dalam penelitian berjudul "Analisis Probabilitas Hujan Menggunakan Data Historis Dari BMKG Wilayah I Tahun 2013–2015", melakukan analisis statistik terhadap data curah hujan harian dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) Wilayah I selama periode 2013–2015. Penelitian ini menggunakan pendekatan probabilistik termasuk distribusi Weibull, Gumbel, dan rantai Markov untuk memodelkan kemungkinan terjadinya hujan berdasarkan kondisi hari sebelumnya. Hasil menunjukkan bahwa rantai Markov mampu menangkap ketergantungan temporal antar hari dengan baik, dan distribusi Gumbel paling sesuai untuk menggambarkan pola curah hujan ekstrem di wilayah tersebut. Meskipun penelitian ini tidak menggunakan machine learning, temuannya sangat relevan karena menegaskan bahwa data cuaca harian bersifat time-series dengan ketergantungan kuat pada kondisi sebelumnya, sehingga model prediksi harus mempertimbangkan aspek temporal prinsip yang secara langsung mendukung penggunaan lag features dan Time-Series Split dalam

penelitian Anda. Lebih lanjut, penelitian ini juga menyoroti pentingnya kualitas dan representasi fitur (misalnya: kelembapan, suhu, dan curah hujan) sebagai dasar pemodelan, yang selaras dengan fokus Anda pada enam variabel lag sebagai input Random Forest. Dalam konteks penelitian Anda, hal ini memberikan landasan teoretis bahwa penggunaan RH_AVG_lag1 dan RR_lag1 sebagai fitur dominan bukan sekadar hasil komputasi, melainkan refleksi dari dinamika fisik atmosfer yang telah dikonfirmasi melalui analisis statistik klasik. Temuan ini juga menggarisbawahi bahwa ketidakseimbangan kelas seperti dominasi kondisi "mendung" dalam dataset DIY merupakan karakteristik alami dari iklim tropis, sehingga model yang dibangun harus dirancang untuk mengakomodasi ketidakseimbangan tersebut tanpa mengorbankan interpretabilitas. Dengan demikian, pendekatan Random Forest yang penulis gunakan dengan class_weight='balanced' dan analisis feature importance merupakan respons yang tepat terhadap tantangan metodologis yang telah diidentifikasi oleh penelitian ini, yaitu: bagaimana membangun sistem prediksi cuaca harian yang andal, transparan, dan berbasis data nyata di wilayah tropis Indonesia.

Penelitian Kedelapan, oleh (Rifqi Maulana et al., 2026)dalam penelitian berjudul "Analisis dan Prediksi Curah Hujan Bulanan Kota Serang Berbasis Apache Spark Menggunakan Dataset BPS Provinsi Banten" mengembangkan sistem prediksi curah hujan bulanan yang secara eksplisit memanfaatkan fitur lag (t−1) sebagai komponen kritis dalam proses feature engineering. Penelitian ini menggunakan dataset curah hujan bulanan dari Badan Pusat Statistik (BPS) Provinsi Banten selama periode 2005–2024, dan dalam tahap pemodelan, penulis tidak hanya mengandalkan variabel cuaca mentah, tetapi juga membangun baseline lag-1 yaitu prediksi bulan berikutnya = nilai curah hujan bulan sebelumnya sebagai tolok ukur kinerja minimal.

Dalam implementasi machine learning, model Random Forest Regression dilatih menggunakan kombinasi fitur numerik (termasuk lag-1 dari curah hujan) serta variabel waktu (bulan/tahun), dengan

tujuan menangkap pola temporal dan non-linear dalam data iklim. Hasil evaluasi menunjukkan bahwa Random Forest mencapai performa terbaik dibanding model lain, dengan RMSE = 84,07 mm dan Skill Score tertinggi (0,534 terhadap baseline klimatologi; 0,253 terhadap baseline lag-1). Nilai Skill Score > 0,25 mengindikasikan peningkatan signifikan dibanding pendekatan sederhana berbasis nilai historis, yang konsisten dengan temuan bahwa lag alone tidak cukup untuk menangani fluktuasi ekstrem seperti lonjakan curah hujan mendadak karena model baseline cenderung “tertinggal” dan gagal merespons perubahan cepat.

Temuan ini sangat relevan dengan penelitian Anda karena:

Mengonfirmasi validitas penggunaan fitur lag (t−1) sebagai fondasi prediksi time-series cuaca baik dalam konteks regresi (curah hujan bulanan) maupun klasifikasi (cuaca harian), selama data memiliki ketergantungan temporal yang kuat.

Menegaskan bahwa Random Forest unggul dalam menangkap interaksi antar fitur non-linear, misalnya: kombinasi RH_AVG_lag1 dan RR_lag1 (kelembapan dan curah hujan kemarin) yang dalam penelitian Anda terbukti sebagai dua fitur paling dominan (importance = 0.305 dan 0.252).

Mengjelaskan mengapa akurasi model Anda hanya mencapai 50%: seperti halnya baseline lag-1 dalam penelitian (Rifqi Maulana et al., 2026)yang gagal menangani transisi ekstrem, model Anda juga mengalami kesulitan membedakan kelas Mendung vs Hujan pada hari-hari transisi (misalnya: pagi mendung, siang hujan), karena fitur lag tunggal belum cukup untuk menangkap dinamika konvektif lokal tanpa tambahan informasi eksternal (seperti tekanan udara atau indeks stabilitas atmosfer).

Penelitian Kesembilan, oleh (Hendra Jatnika et al., 2025)dalam penelitian berjudul "Application of Random Forest Classification

Method in Determining the Best Quality Service in the Implementation of International Certification at ITCC ITPLN" mengimplementasikan algoritma Random Forest untuk klasifikasi sentimen kualitas layanan berbasis umpan balik peserta sertifikasi di lingkungan ITPLN. Penelitian ini menggunakan dataset 2.720 entri teks yang diklasifikasikan menjadi dua kategori sentimen (positif: 1.884 data; negatif: 836 data), dengan preprocessing meliputi case folding, tokenizing, stopword removal, dan stemming sesuai kaidah NLP, serta mengadopsi kerangka metodologis CRISP-DM (Cross-Industry Standard Process for Data Mining) untuk menjamin sistematisasi alur penelitian dari pemahaman bisnis hingga evaluasi model.

Dalam implementasi klasifikasi, Random Forest dilatih dengan parameter default (n_estimators=100, max_depth=None) dan dievaluasi menggunakan metrik presisi (0,92), recall (0,91), F1-score (0,92), serta akurasi (88,97%). Hasil ini mengonfirmasi keunggulan Random Forest dalam menangani ketidakseimbangan kelas (class imbalance) melalui mekanisme bootstrap aggregation dan feature randomness yang secara intrinsik mengurangi bias terhadap kelas mayoritas karakteristik kritis yang secara langsung relevan dengan tantangan distribusi kelas tidak seimbang pada dataset cuaca DIY (Cerah: 8,30%, Hujan: 41,30%, Mendung: 50,40%). Selain itu, penerapan CRISP-DM sebagai kerangka metodologis pada penelitian memberikan preseden institusional yang kuat untuk adopsi alur serupa pada penelitian prediksi cuaca, di mana setiap fase mulai dari pemahaman domain meteorologis, eksplorasi data BMKG, hingga evaluasi model dilaksanakan secara terstruktur untuk meminimalkan risiko metodologis seperti data leakage.

Penelitian Kesepuluh, oleh (Dudek, 2022)dalam penelitian berjudul "A Comprehensive Study of Random Forest for Short-Term Load Forecasting" melakukan analisis komprehensif terhadap penerapan Random Forest untuk forecasting time-series pada prediksi beban listrik jangka pendek di empat negara Eropa (Polandia, Inggris Raya, Prancis,

Jerman). Penelitian ini menguji berbagai konfigurasi metodologis Random Forest dalam konteks data temporal, termasuk: (1) training mode (global vs lokal), (2) definisi input patterns yang mengintegrasikan daily seasonality (r2), weekly seasonality (r4), dan kombinasi keduanya (r6), (3) feature engineering berbasis lag temporal, serta (4) tuning hyperparameter kritis seperti jumlah pohon (n_estimators), ukuran minimum leaf (min leaf size), dan jumlah prediktor yang dipilih secara acak pada setiap split (p).

Hasil eksperimen menunjukkan bahwa konfigurasi optimal tercapai dengan global extended training mode menggunakan pola input r4 (menggabungkan daily dan weekly seasonality) dikombinasikan dengan time-series split untuk validasi bukan random split yang secara signifikan mengurangi risiko data leakage. Random Forest dengan 300 pohon mencapai MAPE (Mean Absolute Percentage Error) terendah antara 1,05%–2,36% tergantung wilayah, dengan penurunan error hingga 50% dibanding model baseline. Analisis hyperparameter mengungkap bahwa peningkatan jumlah pohon dari 1 menjadi 300 secara konsisten menurunkan MAPE, sementara minimum leaf size = 1 dan predictor selection = n/3 menghasilkan stabilitas prediksi tertinggi. Temuan kritis lainnya adalah bahwa Random Forest mampu menangkap pola musiman kompleks tanpa memerlukan asumsi stasioneritas karakteristik fundamental yang secara langsung relevan dengan dinamika cuaca tropis DIY yang dipengaruhi oleh siklus monsun Asia-Australia.

2.1.1 Metrik Penelitian

Tabel 2. 1 State Of The Art

No	Nama Peneliti	Judul Penelitian	Fokus Penelitian	Metode & Dataset	Temuan Utama	Implikasi bagi Penelitian	GAP
1.	**(Intan et al., 2021b)**	Analisis Performansi Prakiraan Cuaca Menggunakan Algoritma Machine Learning	Prediksi cuaca di Makassar (iklim tropis)	Bayesian Regularization NN Data BMKG 2016–2018 (resolusi 3 jam)	Akurasi >99% pada musim hujan/kemarau,soroti ketidakseimbangan kelas	Menunjukkan tantangan utama dalam prediksi cuaca tropis: dominasi kelas tertentu menyebabkan bias model	Penelitian sebelumnya pakai split acak. Pada penelitian penulis menggunakan Time-Series Split, sehingga hasil lebih dapat dipercaya untuk aplikasi nyata
2.	**(Zian Asti Dwiyanti, 2023c)**	Prediksi Cuaca Kota Jakarta Menggunakan Metode Random Forest: Studi Optimalitas	Klasifikasi cuaca multikelas di Jakarta	Random Forest OpenData jakarta 2018 (8.535 entri, 8 Kelas)	Akurasi = 0,71; ROC-AUC = 0,92 →kuat meski data tidak seimbang	RF efektif untuk klasifikasi cuaca tropis dengan ketidakseimbangan kelas	Penelitian Sebelumnya label tidak standar.Penulis menyelaraskan pelabelan dengan kriteria BMKG resmi
3.	**(Ashari Rakhmat & Mutohar, 2023b)**	Prakiraan Hujan menggunakn meode Random Forest dan Cross Validation	Prediksi curah hujan harian di Indonesia	Random Forest + K-Fold CV Data BMKG 2015 – 2021 (2.557 data)	RF + CV → MSE = 0,086, RMSE = 0,290 → lebih stabil daripada tanpa CV	Validasi silang penting, tapi split acak berisiki data leakage pada data time-series	Penelitian Sebelumnya rentan data leakage. Penulis menutup GAP dengan validasi berbasis time series
4.	**(Abdurrasyid et al., 2021)**	Analisis dan Prediksi Curah Hujan Bulanan kota Serang Berbasis Apache Spark Menggunakan Dataset BPS	Prediksi curah hujan bulanan di kota Serang	Random Forest Regression + Apache Spak Data BPS 2005-2024	RF > baseline lag-1 (Skill Score = 0,253); lag alone gagal pada transisi ekstrem	Konfirmasi bahwa fitur lag (t−1) valid, tapi tidak cukup tanpa interaksi fitur	Penelitian sebelumnya regresi bulanan. Penulis fokus pada prediksi dan klasifikasi harian dengan pelabelan operasional BMKG

No	Nama Peneliti	Judul Penelitian	Fokus Penelitian	Metode & Dataset	Temuan Utama	Implikasi bagi Penelitian	GAP
		provinsi Banten					
5.	**(Kentjie et al., 2026)**	Prediksi Degradasi Kesehatan Baterai Kendaraan Listrik Berbasis Machine Learning pada Pengisian Cepat DC	Prediksi degradasi kesehatan baterai (ΔSoH) kendaraan listrik selama proses pengisian cepat DC menggunakan Random Forest Regressor	Metode: Random Forest Regressor Dataset: Simulasi pengisian cepat DC dengan parameter: arus pengisian, tegangan DC, suhu baterai, State of Charge (SoC)	Random Forest efektif memprediksi fenomena degradasi non-linear dengan akurasi ekstrem ($R^2 \approx 1$) Analisis feature importance mengidentifikasi variabel kritis (suhu & arus) yang memengaruhi degradasi	Kurangnya penanganan class imbalance: Tidak relevan untuk prediksi regresi, tetapi kritis untuk klasifikasi cuaca dengan distribusi tidak seimbang	Penelitian sebelumnya tidak mempertimbangkan karakteristik iklim spesifik wilayah, Penulis mengembangkan model spesifik wilayah DIY yang dipengaruhi orografi Merapi-Selatan dan pola monsun Jawa
6.	**(Ainul Idham & Efy Yosrita, 2025)**	Analysis of Long Short-Term Memory (LSTM) and Extreme Gradient Boosting (XGBoost) Algorithms to Predict the Number of Airplane Passengers at Makassar Sultan Hasanuddin International Airport	Prediksi jumlah penumpang pesawat menggunakan pendekatan hybrid LSTM-XGBoost untuk data time-series yang kompleks, dinamis, dan musiman	LSTM, XGBoost, dan hybrid LSTM+XGBoost Dataset: Systematic Literature Review (SLR) dengan pendekatan PRISMA dari 44.564 artikel → 11 artikel relevan	LSTM unggul dalam menangkap pola time-series jangka panjang dan ketergantungan temporal XGBoost lebih robust terhadap noise dan outliers pada data numerik kompleks Hybrid model menggabungkan keunggulan keduanya: kemampuan LSTM menangkap pola temporal + kemampuan XGBoost	Tidak mengintegrasikan lag features atau rolling statistics untuk menangkap autokorelasi temporal, Minim interpretasi prediktor dominan yang relevan untuk keputusan operasional	tidak mengintegrasikan feature engineering temporal (lag features, rolling statistics), Penulis mengembangkan feature engineering khusus untuk menangkap pola autokorelasi iklim DIY

No	Nama Peneliti	Judul Penelitian	Fokus Penelitian	Metode & Dataset	Temuan Utama	Implikasi bagi Penelitian	GAP
					menangani variabel eksternal (cuaca, libur) Validasi silang krusial untuk mencegah overfitting pada data time-series		
7.	**(Arnah Ritonga et al., 2025)**	Analisis Probabilitas Hujan Menggunakan Data Historis Dari BMKG Wilayah I Tahun 2013-2015	Analisis probabilitas curah hujan menggunakan pendekatan statistik untuk mendukung pengambilan keputusan di sektor pertanian, perencanaan kota, dan mitigasi bencana hidrometeorologi	Analisis statistik deskriptif, distribusi probabilitas (gamma, log-normal), rantai Markov untuk menangkap ketergantungan temporal harian Dataset: Data curah hujan dan hari hujan BMKG Wilayah I (Sumatera Utara), periode 2013–2015	Variasi probabilitas hujan signifikan antar tahun: 2013 (36,39%) vs 2014 (26,51%) Fluktuasi dipengaruhi fenomena iklim global (El Niño/La Niña) Distribusi log-normal efektif untuk memodelkan curah hujan ekstrem	Pendekatan statistik klasik kurang optimal untuk prediksi cuaca harian dinamis yang memerlukan machine learning Tidak ada evaluasi model prediktif (hanya analisis probabilitas historis) Tidak ada penanganan class imbalance untuk kelas cuaca kritis (hujan lebat/bencana)	Penelitian sebelumnya menggunakan pendekatan statistik klasik (distribusi probabilitas) tanpa machine learning sedangkan penulis menerapkan /class_weight + evaluasi dengan Macro F1-score & ROC-AUC (One-vs-Rest)
8.	**(Rifqi Maulana et al., 2026)**	Analisis dan Prediksi Curah Hujan Bulanan Kota Serang Berbasis Apache Spark Menggunakan Dataset	Analisis pola curah hujan bulanan dan prediksi curah hujan menggunakan big data analytics untuk mendukun	Apache Spark (big data processing), machine learning (Linear Regression, Decision Tree, Random Forest), baseline	Pola curah hujan bulanan menunjukkan karakteristik musiman jelas: intensitas tinggi pada awal/akhir tahun, penurunan signifikan	Resolusi temporal bulanan tidak sesuai untuk kebutuhan prediksi harian operasional BMKG yang memerlukan resolusi harian Tidak ada lag features multi-step (hanya lag-1) kurang	Penelitian sebelumnya menggunakan resolusi bulanan, Penulis menggunakan resolusi harian yang sesuai kebutuhan operasional BMKG Stasiun Klimatologi Mlati Penelitian sebelumnya hanya menggunakan lag-1 Penelitian sebelumnya menggunakan split acak tanpa mempertimbangkan urutan temporal

No	Nama Peneliti	Judul Penelitian	Fokus Penelitian	Metode & Dataset	Temuan Utama	Implikasi bagi Penelitian	GAP
		BPS Provinsi Banten	g perencanaan sumber daya air dan mitigasi bencana di Kota Serang	klimatologi & lag-1 Dataset: Data curah hujan BPS Provinsi Banten, periode 2005–2024 (20 tahun)	pada pertengahan tahun Random Forest mengunggul i Linear Regression dan Decision Tree dalam akurasi prediksi Baseline klimatologi menghasilkan RMSE tertinggi (180,30 mm) tidak mampu merepresentasikan variasi bulanan	optimal untuk menangkap pola autokorelasi jangka menengah Tidak ada analisis feature importance untuk mengidentifikasi prediktor dominan spesifik wilayah	
9.	**Application of Random Forest Classification Method in Determining the Best Quality Service in the Implementation of International Certification at ITCC ITPLN**	Application of Random Forest Classification Method in Determining the Best Quality Service in the Implementation of International Certification at ITCC ITPLN	Analisis sentimen terhadap 2.720 feedback peserta sertifikasi ITCC ITPLN untuk mengevaluasi kualitas layanan menggunakan klasifikasi teks	Metode: Random Forest Classification dengan preprocessing NLP (case folding, tokenizing, stopword removal, stemming) Dataset: 2.720 feedback teks peserta sertifikasi ITCC ITPLN (1.884 positif, 836 negatif) Evaluasi: Precision 0.92, Recall 0.91, F1-	Random Forest efektif untuk klasifikasi biner dengan data tidak seimbang (rasio ~2.25:1) Preprocessing teks krusial untuk meningkatkan kualitas fitur input Model mampu mencapai F1-score >0.90 meskipun terdapat ketidakseimbangan kelas Keterbatasan kritis:	Domain berbeda: Fokus pada analisis sentimen teks, bukan prediksi fenomena atmosferik time-series Tidak ada validasi temporal: Tidak mempertimbangkan ketergantungan temporal seperti pada data cuaca harian Tidak ada lag features: Tidak mengintegrasikan ketergantungan historis yang krusial untuk forecasting Split acak berisiko data leakage: Tidak disebutkan	Penelitian sebelumnya menggunakan random split tanpa mempertimbangkan urutan temporal, sedangkan penulis menerapkan chronological time-series split* 80:20 tanpa *shuffle untuk mencegah data leakage pada prediksi cuaca harian Penelitian sebelumnya tidak mengintegrasikan feature engineering temporal (lag features, rolling statistics), sedangkan penulis mengembangkan feature engineering khusus untuk menangkap pola autokorelasi iklim DIY (lag 1–3 hari, rolling 7 hari)

No	Nama Peneliti	Judul Penelitian	Fokus Penelitian	Metode & Dataset	Temuan Utama	Implikasi bagi Penelitian	GAP
				score 0.92, Accuracy 88.97% Split: Tidak disebutkan secara eksplisit (kemungkinan random split)	Tidak ada analisis feature importance untuk mengidentifikasi kata/frasa dominan yang memengaruhi sentimen	penggunaan chronological split untuk data berurutan waktu	
10.	**Short-Term Load Forecasting Using Random Forest with Pattern-Based Data Preprocessing**	Prediksi beban listrik jangka pendek (24 jam) menggunakan Random Forest dengan preprocessing berbasis pola untuk menangani multi-seasonality (harian, mingguan, tahunan)	Prediksi beban listrik jangka pendek (24 jam) menggunakan Random Forest dengan preprocessing berbasis pola untuk menangani multi-seasonality (harian, mingguan, tahunan)	Metode: Random Forest dengan 7 skema pattern-based preprocessing (r1–r7) + 3 mode pelatihan (local, global, global extended) Dataset: Data beban listrik 4 negara Eropa (Polandia, Inggris, Prancis, Jerman), periode 2012–2019 Fitur kritis: Lag patterns (r2: daily profile, r4: weekly seasonality, r6: cross-pattern daily+weekly) Validasi: Walk-forward validation	Pattern r4/r6 optimal: Pola yang menggabungkan seasonality harian & mingguan (r4, r6) menghasilkan MAPE terendah (1.05–2.36%) Global extended mode unggul: Pelatihan global dengan ekstensi variabel kalender (musim, hari dalam minggu) mengurangi error hingga 9× dibanding mode global standar Hyperparameter kritis: Jumlah pohon (K=300), ukuran leaf minimum (m=1), dan pemilihan prediktor	Validasi temporal rigor: Penggunaan walk-forward validation sesuai best practice forecasting time-series Feature engineering temporal canggih: Integrasi lag patterns multi-level (harian, mingguan) untuk menangkap autokorelasi Penanganan seasonality eksplisit: Variabel kalender (musim, hari dalam minggu) efektif untuk data dengan pola musiman kuat Transparansi model: Analisis feature importance memberikan interpretasi operasional untuk pengambilan keputusan	Domain berbeda: Prediksi beban listrik (regresi kontinu) ≠ prediksi cuaca multikelas Tidak ada penanganan class imbalance eksplisit: Fokus pada regresi MAPE, bukan klasifikasi dengan distribusi tidak seimbang Karakteristik iklim lokal tidak terakomodasi: Tidak mempertimbangkan orografi atau pola monsun spesifik wilayah seperti DIY

No	Nama Peneliti	Judul Penelitian	Fokus Penelitian	Metode & Dataset	Temuan Utama	Implikasi bagi Penelitian	GAP
				dengan chronological split Evaluasi: MAPE 1.05–2.36%, MdAPE 0.78–1.78%	acak (p=√n) signifikan memengaruhi performa Feature importance transparan: Analisis predictor importance mengidentifikasi jam-jam kritis dalam pola harian sebagai prediktor dominan		
11	Rafi Dio Adibta (2026)	Prediksi Cuaca Harian Berdasarkan Data BMKG 2022-2025	Prediksi kejadian hujan harian (rain/no-rain) di wilayah DIY menggunakan pembelajaran mesin berbasis time-series.	Pipeline *StandardScaler* + *Random Forest* dengan 5-Fold TimeSeriesSplit, evaluasi akhir melalui Hold-Out Test. Dataset: *BMKG DIY 2022-2025* ditambah fitur *lag* (t–1…t–7) dan *rolling statistics*.	Model mampu menangkap pola temporal hujan DIY; metrik evaluasi (Accuracy, F1-Macro, ROC-AUC) menunjukkan performa stabil tanpa data leakage. Fitur lag meningkatkan sensitivitas terhadap dinamika cuaca harian.	Memberikan pendekatan evaluasi anti data leakage yang benar untuk data deret waktu dalam konteks prediksi cuaca harian. Menunjukkan efektivitas Random Forest dibanding metode baseline pada pola hujan yang tidak seimbang.	Penelitian terdahulu umumnya mengabaikan prosedur *time-series validation* dan tidak menggunakan hold-out berbasis waktu. Minim penelitian menggunakan fitur lag yang sistematis pada cuaca DIY. Penelitian ini mengisi gap tersebut dengan metodologi yang lebih ketat dan replikasi berbasis kode.

2.2 Landasan Teori.

2.2.1 Cuaca

Cuaca merupakan keadaan atmosfer di suatu tempat pada waktu tertentu yang sifatnya dapat berubah-ubah, mulai dari hitungan menit, jam, hingga hari. Ilmu yang mempelajari tentang cuaca dan proses yang terjadi pada atmosfer disebut meteorologi (Putra et al., 2024). Kondisi cuaca ditentukan oleh kombinasi dari berbagai variabel meteorologi, yang mencakup parameter-parameter seperti suhu udara, tekanan udara, kelembapan, tutupan awan, kecepatan angin, dan curah hujan (Zian Asti Dwiyanti, 2023).Setiap parameter meteorologi memiliki peran signifikan dalam menentukan kondisi atmosfer secara keseluruhan, dan pemahaman mendalam tentang variabel-variabel ini sangat penting untuk prakiraan cuaca yang akurat.

Indonesia, sebagai negara kepulauan yang luas dengan kondisi topografi beragam dan curah hujan tahunan rata-rata lebih dari 2000 mm, sangat rentan terhadap bencana hidrometeorologi (Nurrahman et al., n.d.)Data bencana Indonesia menunjukkan bahwa sepanjang tahun 2021 terdapat 3.658 insiden banjir dan tanah longsor yang tersebar di seluruh Indonesia, menekankan pentingnya sistem pemantauan cuaca dan curah hujan real-time dengan kepadatan tinggi. Prakiraan cuaca yang akurat sangat penting untuk melindungi kehidupan dan harta benda, terutama dalam konteks mitigasi bencana hidrometeorologi dan pengambilan keputusan strategis di berbagai sektor seperti pertanian, transportasi, dan pariwisata.

Tabel 2. 2 Unsur Meteorologis yang Digunakan dalam Penelitian

No	Unsur Cuaca	Nama Variabel	Deskripsi	Satuan

1	Temperatur (Suhu Udara)	TN, TX, TAVG	Ukuran derajat panas atau dinginnya udara. Suhu minimum (TN), maksimum (TX), dan rata-rata (TAVG) merupakan parameter penting yang memengaruhi berbagai proses atmosfer.	°C
2	Kelembapan Udara	RH (Relative Humidity)	Jumlah uap air yang terkandung di dalam udara yang dinyatakan dalam persentase. Kelembapan tinggi merupakan salah satu syarat utama terbentuknya awan dan hujan.	%
3	Curah Hujan	RR (Rainfall)	Jumlah air hujan yang jatuh ke permukaan bumi dalam periode waktu tertentu. Variabel ini menjadi indikator utama terjadinya kondisi hujan.	mm
4	Lama Penyinaran Matahari	SS (Sunshine Duration)	Durasi waktu matahari bersinar tanpa terhalang awan. Variabel ini digunakan sebagai indikator utama untuk kondisi cuaca cerah.	jam
5	Kecepatan Angin	FF_X, FF_AVG	Gerakan udara dari daerah bertekanan tinggi ke daerah bertekanan rendah. Kecepatan angin maksimum (FF_X) dan rata-rata (FF_AVG) memengaruhi pergerakan awan dan distribusi cuaca.	m/s atau knot

2.2.2 Badan Meteorologi, Klimatologi dan Geofisika (BMKG)

Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) merupakan sebuah Lembaga Pemerintahan Non Kementrian (LPNK) di Indonesia yang memiliki tugas utama melaksanakan tugas pemerintahan di bidang meteor, klimatologi, kualitas udara, dan geofisika sesuai dengan ketentuan perundang-undangan yang berlaku (Bendi, 2024). BMKG memiliki tanggung jawab strategis dalam menyediakan data dan informasi yang berkaitan dengan kondisi atmosfer dan bumi untuk mendukung berbagai sektor kehidupan masyarakat(Azani et al., n.d.;Maulita, 2023)Sebagai lembaga pemerintah yang berwenang, BMKG mengoperasikan jaringan stasiun pengamatan yang tersebar di seluruh Indonesia untuk secara berkelanjutan mengukur dan merekam berbagai unsur cuaca dan kondisi atmosfer, termasuk suhu, kelembapan, curah hujan, kecepatan angin, tekanan udara, dan lama penyinaran matahari.

Data Meteorologi yang dikumpulkan oleh BMKG melalui jaringan stasiun pengamatan kemudian di proses melalui tahap pengolahan, analisis, dan validasi untuk memastikan kualitas dan akurasi sebelum dipublikasikan atau digunakan dalam penelitian Hasil dari pengolahan tersebut lalu dianalisis data tersebut dijadikan dasar untuk menghasilkan publik harian tentang cuaca, penerbangan,

pelayaran, pertanian,dan sektor-sektor lainnta yang bergantung pada kondisi atmosfer (Wele et al., 2020).

Fungsi utama dari BMKG yang tidak kalah penting yaiut mengeluarkan peringatan dini terkait potensi cauca ekstrem, gelombang tinggi, tsunami, dan gempa bumi, untuk memitigasi risik bencana.Sistem

2.2.3 Machine Learning & Klasifikasi Data

Machine Learning merupakan salah satu cabang dari Kecerdasan Buatan (AI) yang berfokus pada pengembangan algoritma komputer untuk mempelajari pola dan pengetahuan dari data (Zhang et al., 2022).Salah satu tantangan kritis adalah menghindari data leakage, yaitu penggunaan informasi masa depan dalam proses pelatihan yang dapat menyebabkan model memberikan estimasi performa yang terlalu optimis. Untuk mengatasi masalah ini, validasi model harus menggunakan time-series split, bukan random split, sehingga data pelatihan selalu mendahului data validasi secara temporal

Model machine learning untuk prediksi cuaca harus memanfaatkan ketergantungan temporal melalui fitur lag (t−1), di mana nilai-nilai historis dari variabel meteorologi pada periode sebelumnya digunakan sebagai prediktor. Penggunaan lag features ini didukung oleh pemahaman bahwa data cuaca memiliki autokorelasi yang kuat, artinya kondisi cuaca pada hari ini sangat dipengaruhi oleh kondisi cuaca pada hari-hari sebelumnya. Dengan memanfaatkan informasi temporal ini secara tepat, model dapat menangkap pola-pola musiman dan tren jangka panjang yang penting untuk prediksi akurat

Tantangan signifikan lainnya dalam pengembangan model klasifikasi cuaca adalah menangani ketidakseimbangan kelas (class imbalance), di mana beberapa kategori cuaca jauh lebih sering muncul dibandingkan kategori lainnya (Choudhary & Shukla, 2022)Ketidakseimbangan kelas ini dapat menyebabkan algoritma pembelajaran menjadi bias terhadap kelas mayoritas, sehingga performa pada kelas minoritas menjadi sangat rendah.

1.2.4 Algoritma Random Forest

Random Forest (RF) adalah algoritma ensemble learning berbasis decision tree yang telah dikembangkan (Wei et al., 2025)RF membangun sejumlah pohon keputusan secara independen, masing-masing dilatih pada subset acak dari data

dan fitur (Putra et al., 2024).Prediksi akhir ditentukan melalui mekanisme voting mayoritas untuk klasifikasi atau rata-rata untuk regresi. Pendekatan ensemble ini memungkinkan RF untuk mengurangi varians dan meningkatkan stabilitas prediksi dibandingkan dengan decision tree tunggal.

Keunggulan Random Forest dalam konteks prediksi cuaca meliputi beberapa aspek penting. Pertama, RF robust terhadap noise dan outlier, misalnya dalam menangani curah hujan ekstrem yang dapat mencapai 228,6 mm tanpa mengganggu performa model secara keseluruhan.Kedua, RF mampu menangkap hubungan non-linear antar variabel meteorologis, yang merupakan karakteristik penting dalam sistem atmosfer yang kompleks.

Ketiga, Random Forest memberikan interpretasi melalui feature importance, yang memungkinkan peneliti untuk memahami kontribusi relatif setiap variabel meteorologi terhadap prediksi (Bidve et al., 2024) Feature importance dalam Random Forest dihitung berdasarkan penurunan impuritas atau mean decrease in accuracy, sehingga memberikan insight tentang variabel mana yang paling berpengaruh dalam model.Keempat, RF tidak rentan terhadap overfitting, bahkan tanpa tuning hyperparameter yang kompleks, karena mekanisme bagging dan feature randomness secara intrinsik mengurangi varians model.

Konfigurasi Hyperparameter Random Forest

Dalam penelitian ini, Random Forest dikonfigurasi dengan parameter-parameter spesifik yang dirancang untuk mengoptimalkan performa pada dataset cuaca dengan karakteristik khusus (Putra et al., 2024). Konfigurasi yang digunakan adalah sebagai berikut:

1. n_estimators = 100

Parameter nestimators menentukan jumlah pohon keputusan yang akan dibangun dalam ensemble (Wei et al., 2025).Dengan nestimators = 100, model akan membangun 100 pohon keputusan independen, masing-masing dilatih pada bootstrap sample yang berbeda dari dataset asli(Putra et al., 2024). Jumlah pohon ini dipilih berdasarkan trade-off antara akurasi dan efisiensi komputasi, di mana peningkatan jumlah pohon secara konsisten menurunkan error prediksi, namun peningkatan lebih lanjut memberikan improvement yang marginal.

2. max_depth = 5

Parameter maxdepth membatasi kedalaman maksimal setiap pohon keputusan dalam ensemble. Dengan maxdepth = 5, setiap pohon dapat memiliki maksimal 5 level kedalaman, yang membatasi kompleksitas model dan mengurangi risiko overfitting. Pembatasan kedalaman ini sangat penting dalam konteks dataset cuaca yang memiliki ketidakseimbangan kelas, karena pohon yang terlalu dalam cenderung mempelajari noise dan pola spesifik dari data pelatihan yang tidak dapat digeneralisasikan.

3. class_weight = 'balanced'

Parameter classweight='balanced' mengimplementasikan mekanisme penanganan ketidakseimbangan kelas dengan secara otomatis memberikan bobot yang berbeda untuk setiap kelas. Dengan menggunakan classweight='balanced', kelas minoritas akan diberikan bobot yang lebih tinggi selama proses pelatihan, sehingga mengurangi bias model terhadap kelas mayoritas (Mendung dengan 50,40% dari total data).Pendekatan ini memastikan bahwa model memberikan perhatian yang seimbang terhadap semua kategori cuaca, bukan hanya kategori yang paling sering muncul.

4. random_state = 42

Parameter randomstate = 42 menetapkan seed untuk random number generator, memastikan reproducibility dari hasil penelitian.Dengan menetapkan randomstate yang konsisten, setiap kali model dilatih dengan parameter yang sama akan menghasilkan pohon-pohon keputusan yang identik, sehingga memudahkan verifikasi dan replikasi hasil penelitian.

5. Mekanisme Bagging dan Feature Randomness

Random Forest menggunakan dua mekanisme utama untuk mengurangi varians dan meningkatkan generalisasi yaitu bagging (bootstrap aggregating) dan feature randomness (Putra et al., 2024). Dalam bagging, setiap pohon dilatih pada bootstrap sample yang dipilih secara acak dengan pengembalian dari dataset asli. Proses ini menciptakan keragaman dalam pohon-pohon individual, sehingga ketika prediksi dari semua pohon digabungkan, hasilnya lebih stabil dan akurat

Feature randomness mengacu pada proses pemilihan subset acak dari fitur pada setiap split dalam pohon keputusan. Alih-alih mempertimbangkan semua fitur untuk setiap split, Random Forest hanya mempertimbangkan subset acak dari fitur, yang mengurangi korelasi antar pohon dan meningkatkan diversity dalam ensemble. Kombinasi dari bagging dan feature randomness ini membuat Random Forest robust terhadap noise, outlier, dan ketidakseimbangan kelas, menjadikannya algoritma yang ideal untuk prediksi cuaca

2.2.5 Evaluasi Model Klasifikasi Multikelas

Mengingat sifat masalah sebagai klasifikasi multikelas tidak seimbang, evaluasi model tidak cukup hanya mengandalkan accuracy. Metrik tunggal seperti accuracy dapat memberikan gambaran yang menyesatkan ketika distribusi kelas sangat tidak seimbang, karena model dapat mencapai akurasi tinggi dengan hanya memprediksi kelas mayoritas.Oleh karena itu, evaluasi komprehensif menggunakan multiple metrics menjadi sangat penting untuk memberikan penilaian yang akurat terhadap performa model (Yustanti et al., 2023)

a. Macro-average F1-Score

Macro-average F1-Score merupakan metrik yang memberikan bobot yang sama pada semua kelas, termasuk kelas minoritas dan mayoritas. F1-Score adalah rata-rata harmonis dari precision dan recall, yang menggabungkan kedua metrik ini menjadi satu nilai tunggal . Dengan menggunakan macro-average, setiap kelas diberikan kontribusi yang sama dalam perhitungan akhir, sehingga performa pada kelas minoritas tidak diabaikan (Basuni & Siregar, 2023),(Yustanti et al., 2023)

Macro-average F1-Score sangat berguna dalam konteks klasifikasi cuaca multikelas, di mana kategori seperti "Cerah" (kelas minoritas) memiliki jumlah sampel yang jauh lebih sedikit dibandingkan "Mendung" (kelas mayoritas). Dengan menggunakan metrik ini, model dapat dievaluasi berdasarkan kemampuannya untuk memprediksi semua kategori cuaca dengan baik, bukan hanya kategori yang paling sering muncul(Yustanti et al., 2023).

b. ROC-AUC One-vs-Rest (OvR)

ROC-AUC One-vs-Rest (OvR) merupakan metrik yang mengukur luas di bawah kurva ROC (Receiver Operating Characteristic) untuk setiap kelas versus gabungan kelas lainnya 2022. Metrik ini sangat cocok untuk menilai kemampuan diskriminasi model dalam data tidak seimbang, karena ROC-AUC tidak terpengaruh oleh perubahan threshold klasifikasi dan distribusi kelas (Zhang et al., 2022)

Dalam konteks klasifikasi cuaca, ROC-AUC OvR memungkinkan evaluasi kemampuan model untuk membedakan setiap kategori cuaca dari kategori lainnya.Misalnya, ROC-AUC untuk kelas "Cerah" mengukur seberapa baik model dapat membedakan hari "Cerah" dari hari "Mendung" dan "Hujan". Nilai ROC-AUC berkisar dari 0 hingga 1, di mana nilai 1 menunjukkan diskriminasi sempurna dan nilai 0,5 menunjukkan diskriminasi acak (Zhang et al., 2022)

c. Confusion Matrix

Confusion Matrix menunjukkan distribusi prediksi benar dan salah untuk setiap kelas, memberikan insight detail tentang pola kesalahan spesifik yang dilakukan model (Yustanti et al., 2023). Dalam konteks klasifikasi cuaca multikelas, confusion matrix membantu mengidentifikasi kecenderungan salah klasifikasi, misalnya: kecenderungan model untuk salah mengklasifikasikan "Mendung" sebagai "Cerah" atau sebaliknya

Confusion matrix juga memungkinkan perhitungan metrik-metrik lain seperti precision, recall, dan specificity untuk setiap kelas secara individual. Dengan menganalisis confusion matrix secara detail, peneliti dapat memahami karakteristik kesalahan model dan mengidentifikasi area-area yang memerlukan perbaikan. Visualisasi confusion matrix dalam bentuk heatmap juga memudahkan interpretasi pola kesalahan klasifikasi

d. Fitur Lag t-1

Fitur lag t-1 merupakan nilai dari variabel pada waktu sebelumnya (t-1) yang digunakan sebagai prediktor untuk memprediksi nilai pada waktu saat ini (t). Dalam konteks prediksi cuaca, lag features sangat penting karena kondisi cuaca pada hari sebelumnya memiliki pengaruh signifikan terhadap kondisi cuaca pada hari berikutnya. Penambahan lag features meningkatkan

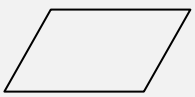
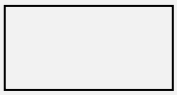
kemampuan model untuk menangkap dependensi temporal dalam data time-series.

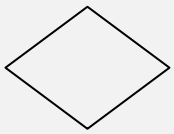
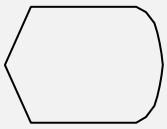
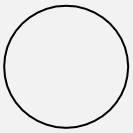
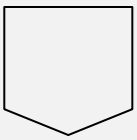
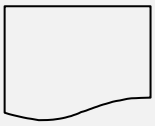
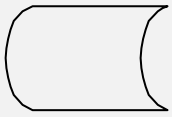
Lag features memungkinkan model untuk belajar pola-pola temporal yang kompleks dalam data cuaca. Misalnya, jika hari sebelumnya adalah "Cerah", kemungkinan hari ini juga "Cerah" lebih tinggi dibandingkan jika hari sebelumnya adalah "Hujan". Dengan memasukkan lag t-1 sebagai fitur, model dapat memanfaatkan informasi historis ini untuk membuat prediksi yang lebih akurat.

2.2.4 Flowchart

Menurut (Rosaly et al., n.d.) flowchart atau diagram alir didefinisikan sebagai suatu jenis diagram yang merepresentasikan algoritma atau langkah-langkah instruksi yang berurutan dalam suatu sistem melalui penggunaan simbol-simbol standar yang masing-masing mewakili proses tertentu, dengan garis penghubung yang mengindikasikan arah aliran eksekusi dari satu proses ke proses berikutnya. Dalam konteks penelitian prediksi cuaca harian berbasis Random Forest, flowchart berperan sebagai bukti dokumentasi yang menjelaskan gambaran logis sistem prediksi kepada pembimbing dan penguji, sekaligus menjadi panduan implementasi bagi programmer dalam menerjemahkan desain logis ke dalam kode Python untuk preprocessing data BMKG, pelatihan model, hingga deployment prototipe Streamlit.

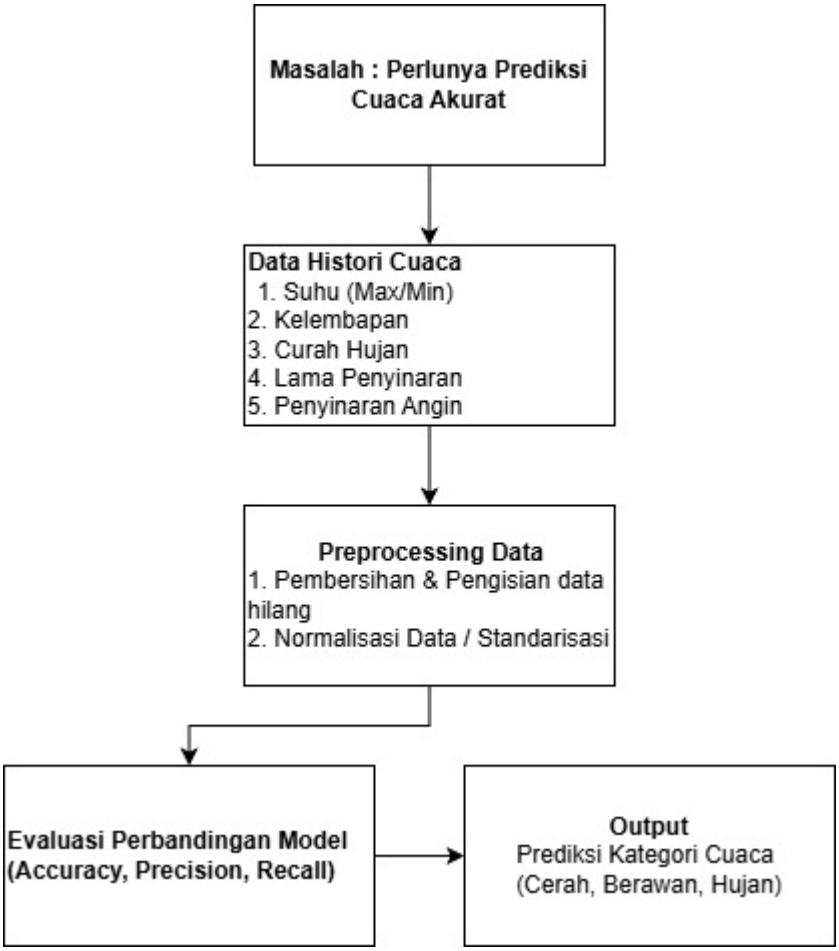
Tabel 2. 3 Flowchart

No	Simbol	Nama	Keterangan
1		Terminal	**Simbol yang mengartikan awal dan akhir nya sebuah program.**
2		Input atau Output	**Simbol yang menunjukkan proses input atau output.**
3		Proses	**Simbol yang mengartikan proses yang berjalan di dalam sebuah komputer.**

4		Decision	**Simbol yang menghasilkan kemungkinan dua jawaban yaitu YA/TIDAK pada kondisi tertentu.**
5		Display	**Simbol yang membentuk keluaran pada sebuah layar monitor.**
6		Alur Proses	**Simbol untuk menunjukkan alur proses sebuah objek.**
7		Connector	**Simbol untuk menunjukkan hubungan antar objek dengan objek lainnya pada halaman yang sama.**
8		Offline Connector	**Simbol untuk menunjukkan hubungan antar objek dengan objek lainnya pada halaman yang berbeda.**
9		Document	**Simbol yang mencetak dokumen.**
10		Disk Storage	**Simbol yang menunjukkan input atau output yang disimpan ke dalam sebuah disk.**

2.3 Kerangka Pemikiran

Kerangka pemikiran dalam penelitian ini dirancang secara sistematis mengikuti alur logis dari data-driven modeling berbasis supervised learning, sebagaimana digambarkan pada Gambar 2. 1 Kerangka Pemikiran. Proses dimulai dari ketersediaan data historis cuaca sebagai fondasi utama, dilanjutkan dengan serangkaian tahapan pra-pemrosesan, pemilihan fitur, pelatihan model, hingga evaluasi kinerja. Setiap tahap saling terkait dan memenuhi prinsip causality dalam prediksi time-series: model hanya menggunakan informasi masa lalu untuk memprediksi masa depan.



Gambar 2. 1 Kerangka Pemikiran

1. Data Historis Cuaca

   Sistem prediksi cuaca yang berawal dari ketersediaan data historis cuaca.Data tersebut mencakup lima variabel meteorologis utama, yaitu suhu maksimum dan minimum, kelembapan udara, curah hujan, lama penyinaran matahari, serta kecepatan angin. Kelima variabel ini dikumpulkan secara harian dan membentuk deret waktu yang menjadi dasar bagi seluruh proses analisis dan pemodelan selanjutnya, karena merepresentasikan pola dan dinamika kondisi atmosfer di wilayah.

2. Preprocessing Data

   Sebelum digunakan untuk pelatihan model, data mentah BMKG mengalami serangkaian tahapan preprocessing yang sistematis untuk memastikan kualitas dan kesiapan data dalam proses pembelajaran mesin. Pertama, dilakukan penanganan missing value dengan strategi yang berbeda berdasarkan karakteristik fisika variabel: untuk variabel curah hujan (RR) dan lama penyinaran matahari (SS), nilai

hilang diisi dengan nol berdasarkan asumsi bahwa tidak terdeteksinya pengukuran berarti tidak terjadinya fenomena tersebut; sedangkan untuk variabel suhu minimum (TN), suhu maksimum (TX), kelembapan rata-rata (RH_AVG), dan kecepatan angin maksimum (FF_X), dilakukan imputasi menggunakan median kolom untuk menjaga robustness terhadap outlier ekstrem.

3. Pemilihan & Ekstraksi Fitur

   Setelah data siap, langkah berikutnya adalah pemilihan dan ekstraksi fitur yang bertujuan untuk mengidentifikasi subset variabel yang paling informatif dan relevan bagi tugas prediksi. Dengan menyaring fitur-fitur yang kurang signifikan, proses ini tidak hanya menyederhanakan model dan mengurangi beban komputasi, tetapi juga membantu meningkatkan akurasi dan ketahanan model terhadap overfitting dengan fokus pada sinyal prediktif yang paling kuat dalam data.

4. Pelatihan Model: Random Forest

   Fitur-fitur yang telah diproses kemudian digunakan sebagai input untuk melatih model Random Forest. Algoritma ini, yang merupakan metode ensemble berbasis decision tree, membangun sejumlah besar pohon keputusan secara acak dan menggabungkan hasil prediksinya untuk menghasilkan keputusan akhir. Pendekatan ini memberikan keunggulan dalam menangani hubungan non-linear antar variabel, memiliki ketahanan tinggi terhadap noise dan outlier, serta mampu memberikan interpretasi mengenai pentingnya setiap fitur dalam proses pengambilan keputusan.

5. Evaluasi & Perbandingan Model

   Kinerja model yang telah dilatih dievaluasi secara objektif menggunakan metrik-metrik klasifikasi standar, seperti akurasi, presisi, dan recall. Evaluasi ini dilakukan pada data uji yang sepenuhnya terpisah dari data latih untuk memastikan bahwa model memiliki kemampuan generalisasi yang baik. Hasil evaluasi ini menjadi dasar untuk memvalidasi efektivitas model dan, jika diperlukan, untuk membandingkannya dengan pendekatan alternatif guna memilih solusi yang paling optimal.

6. Output: Prediksi Kategori Cuaca (Cerah berawan)

   Langkah terakhir dalam kerangka pemikiran ini adalah menghasilkan output berupa prediksi kategori cuaca diskrit sesuai klasifikasi operasional BMKG untuk

wilayah DIY yaitu "Cerah", "Mendung", atau "Hujan". Klasifikasi ini ditentukan secara objektif berdasarkan dua parameter meteorologis utama curah hujan harian (RR) dan kelembapan relatif rata-rata (RH_AVG) dengan kriteria:

1. Cerah jika RR = 0 mm dan RH_AVG < 80%;
2. Mendung jika RR = 0 mm dan RH_AVG ≥ 80%;
3. Hujan jika RR > 0 mm

Output prediksi tidak hanya berupa label kategori tunggal, tetapi juga dilengkapi dengan probabilitas keanggotaan kelas (class probability) yang merepresentasikan tingkat kepercayaan model terhadap setiap prediksi, sehingga pengguna dapat menilai keandalan hasil prediksi secara kuantitatif. Lebih lanjut, sistem prototipe berbasis Streamlit yang dikembangkan menyajikan output prediksi secara interaktif dengan visualisasi tambahan berupa analisis feature importance.