

KOMPARASI RIDGE REGRESSION, RANDOM FOREST, DAN GRADIENT BOOSTING UNTUK PREDIKSI CURAH HUJAN HARIAN DI SUMATRA SELATAN BERBASIS TIME SERIES CROSS-VALIDATION

Mahendi Alvines, Muhammad Ravi Wijayanto, Muhammad Archi Daffa Danendra,
Muhammad Rizky Herdiansyah, Muhammad Karimsyah Lubis,
Muhammad Ichsan Farel Rachmad, Ken Ditha Tania, Allsela Meiriza

Sistem Informasi, Universitas Sriwijaya
Jalan Palembang-Prabumulih km 32 Palembang, Indonesia
09031382227143@student.unsri.ac.id

ABSTRAK

Penelitian ini menganalisis pendekatan komparatif model regresi linear dan non-linear untuk prediksi curah hujan di Sumatera Selatan, Indonesia, menggunakan data klimatologi BMKG tahun 2024. Hal tersebut didasarkan pada perubahan iklim yang meningkatkan ketidakpastian pola curah hujan di Sumatera Selatan, mengancam sektor pertanian dan pengelolaan air. Komparasi ini bertujuan untuk membandingkan kinerja model regresi linear dan non-linear untuk prediksi curah hujan harian di Sumatera Selatan, serta mengidentifikasi prediktor utama yang memengaruhi curah hujan. Metodologi penelitian mengadopsi kerangka Knowledge Discovery in Database (KDD) dengan tahapan pengumpulan data, pra-pemrosesan, transformasi, pemodelan, dan evaluasi. Lima model machine learning dibandingkan: Linear Regression, Ridge Regression, Lasso Regression, Random Forest, dan Gradient Boosting, dengan validasi menggunakan TimeSeriesSplit untuk mempertahankan struktur temporal data. Hasil analisis menunjukkan Ridge Regression sebagai model terbaik dengan RMSE 14.62 mm, meskipun nilai R^2 negatif (-0.69) mengindikasikan keterbatasan model dalam menangkap variabilitas penuh data. Analisis koefisien regresi Ridge menunjukkan bahwa RR_LAG_1 juga berkontribusi signifikan (koefisien: 0.45). Konsistensi prediktor antar model memperkuat validitas temuan, sedangkan model secara konsisten mengalami kesulitan memprediksi kejadian curah hujan ekstrem (>50 mm). Pendekatan berbasis ensemble (Random Forest/Gradient Boosting) menunjukkan stabilitas prediksi lebih tinggi dengan deviasi RMSE <5 mm. Penelitian ini dapat memberikan masukan bagi pengembangan model prediksi curah hujan yang dapat mendukung sektor pertanian, pengelolaan sumber daya air, dan mitigasi bencana di Sumatera Selatan. Studi lanjutan dengan model hybrid dan dataset yang lebih lengkap disarankan untuk validasi lebih lanjut.

Kata kunci : Data Mining, Prediksi curah hujan, Machine Learning, Regresi, Time Series Analysis, Pemodelan Iklim, Variabilitas

1. PENDAHULUAN

Perubahan iklim global telah menjadi isu yang semakin mendesak dalam dua dekade terakhir, dengan dampak yang terasa di berbagai belahan dunia termasuk Indonesia. Sumatera Selatan, sebagai salah satu provinsi dengan kontribusi ekonomi signifikan di bidang pertanian dan perkebunan, menghadapi tantangan serius akibat ketidakpastian pola curah hujan. Peningkatan frekuensi anomali cuaca seperti banjir dan kekeringan telah menimbulkan kerugian ekonomi yang substansial bagi petani dan pelaku industri. Studi yang dilakukan oleh Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) menunjukkan bahwa dalam periode 2024, Curah hujan harian tertinggi sebesar 211mm dan terjadi pada 19 Januari 2024 di Pseksu – Lahat, curah hujan dasarian tertinggi sebesar 510 mm terjadi pada dasarian 1 Maret di Gumay Ulu – Lahat dan curah hujan bulanan tertinggi sebesar 1275 mm pada bulan Januari di Kikim Selatan – Lahat [1], mengindikasikan urgensi pengembangan model prediksi curah hujan yang lebih akurat dan handal untuk wilayah ini.

Prediksi curah hujan yang akurat memegang peranan strategis dalam berbagai sektor penting, terutama pertanian, manajemen sumber daya air, dan

sistem mitigasi bencana. Dalam konteks pertanian yang menjadi salah satu sektor ekonomi utama di Sumatera Selatan, ketepatan prediksi pola presipitasi dapat membantu petani menentukan waktu tanam yang optimal serta memilih varietas tanaman yang sesuai dengan kondisi iklim yang diprediksi. Selain itu, informasi prediktif mengenai curah hujan berperan sebagai komponen esensial dalam perumusan kebijakan untuk mengelola sumber daya air dan mengantisipasi risiko bencana hidrometeorologi. Melalui manajemen risiko yang lebih efektif berdasarkan data prediksi yang reliabel, potensi kerugian ekonomi akibat bencana terkait cuaca dapat diminimalisir, sekaligus mendukung implementasi pembangunan berkelanjutan di tingkat regional [2].

Meskipun penelitian tentang prediksi curah hujan telah banyak dilakukan, implementasi model prediktif yang spesifik untuk karakteristik geografis dan klimatologi Sumatera Selatan masih sangat terbatas. Wilayah ini memiliki kompleksitas tersendiri dengan kondisi topografi yang beragam, mulai dari dataran rendah hingga pegunungan, serta dipengaruhi oleh sistem monsun dan fenomena El Niño Southern Oscillation (ENSO) [3].

Mayoritas model prediksi yang ada saat ini dikembangkan untuk skala nasional atau regional yang lebih luas, sehingga sering kali tidak mampu menangkap variabilitas lokal yang signifikan. Kesenjangan ini menciptakan peluang riset yang berharga untuk mengembangkan model prediksi yang lebih terkalibrasi dengan kondisi spesifik Sumatra Selatan.

Pendekatan komparatif antara model regresi linear (termasuk Ridge/Lasso) dan non-linear (Random Forest, Gradient Boosting) dilakukan untuk mengevaluasi kesesuaiannya dengan karakteristik data curah hujan Sumatra Selatan yang kompleks. Berbeda dengan model machine learning kompleks yang membutuhkan infrastruktur komputasi tinggi dan dataset masif, regresi linear dapat diimplementasikan dengan sumber daya yang lebih terjangkau namun tetap mampu menghasilkan prediksi yang bermakna. Hal ini sangat relevan dalam konteks Indonesia, khususnya Sumatra Selatan, di mana keterbatasan sumber daya untuk implementasi model kompleks masih menjadi kendala.

Penelitian ini mengambil pendekatan interdisipliner dengan mengintegrasikan pengetahuan dari bidang klimatologi, statistika, dan ilmu komputer. Menurut Suranata dalam [4], melalui analisis korelasi komprehensif antara pola curah hujan dengan variabel-variabel klimatologi seperti suhu permukaan laut, tekanan udara, dan kelembaban, penelitian ini bertujuan untuk mengidentifikasi prediktor yang paling signifikan untuk model regresi linear. Penggunaan teknik preprocessing data seperti analisis komponen utama (PCA) dan seleksi fitur akan membantu dalam mengatasi masalah multikolinearitas yang sering muncul dalam dataset klimatologi. Pendekatan metodologis ini diharapkan dapat menghasilkan model prediksi yang tidak hanya akurat tetapi juga interpretable, sehingga memberikan wawasan yang berharga tentang faktor-faktor pendorong pola curah hujan di Sumatra Selatan.

Implikasi praktis dari penelitian ini sangat signifikan, terutama dalam konteks adaptasi perubahan iklim di Indonesia. Model prediksi curah hujan yang akurat dapat menjadi instrumen penting dalam perencanaan pertanian, pengelolaan sumber daya air, dan sistem peringatan dini bencana [5].

Penelitian ini diharapkan dapat berkontribusi pada pembangunan ketahanan iklim di Sumatra Selatan melalui pengembangan model prediksi curah hujan yang terjangkau, akurat, dan sesuai dengan kondisi lokal.

2. TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Temuan dari berbagai kajian memberikan landasan pemahaman yang beragam, meskipun belum sepenuhnya menyentuh aspek spesifik, penulis menjadikan penelitian sebelumnya menjadi dasar dalam penelitian.

Beberapa penelitian terdahulu seperti yang dilakukan oleh Sukmono dalam [6] dan Furqon dalam [7] telah menunjukkan bahwa optimasi model regresi linear dapat mencapai akurasi prediksi yang kompetitif dibandingkan dengan metode yang lebih kompleks untuk kasus-kasus tertentu.

Penelitian yang dilakukan oleh Gusril dalam [8] ini menerapkan model Recurrent Neural Network (RNN) untuk memprediksi curah hujan di Jakarta dengan memanfaatkan data historis interval 1 jam sepanjang tahun 2022. Hasil evaluasi menunjukkan bahwa skenario 90:10 menghasilkan performa terbaik dengan RMSE terendah (0,0925) dibandingkan skenario lainnya. Model yang dikembangkan berpotensi mendukung sistem peringatan dini banjir serta pengambilan keputusan lingkungan di Jakarta, dengan rekomendasi pengembangan lebih lanjut untuk mengintegrasikan variabel klimatologi tambahan guna meningkatkan akurasi prediksi.

Wirawan dalam [9], mengembangkan model prediksi curah hujan untuk sektor pertanian menggunakan algoritma Random Forest dan Grid Search Cross-Validation, dengan memanfaatkan dataset BMKG Nusa Tenggara Barat periode 2000-2023. Model Random Forest awal menghasilkan nilai R-squared 0,1334, yang kemudian dioptimalkan dengan Grid Search Cross-Validation menghasilkan model terbaik dengan R-squared 0,0268. Model ini diimplementasikan untuk membantu petani dalam penjadwalan irigasi, penggunaan pupuk, dan perencanaan tanam.

Pitayandanu melakukan analisis komparasi artificial neural network dalam [10], mengungkapkan bahwa dari berbagai penelitian tentang model prediksi curah hujan, studi komparatif ini menyimpulkan bahwa model Artificial Neural Network (ANN) memiliki keunggulan sebagai metode prediksi terbaik. ANN mengungguli model lain seperti Fuzzy, FFT, EANN, AEEMD-ANN, E-SVR-ANN, BPNN, ASTAR, SAR, Gumbel, ARIMA, FFNN, SVM, ANFIS, dan ANN-Fuzzy dalam kemampuannya mengenali pola data curah hujan. Kelebihan utama ANN adalah fleksibilitas pengembangan variasi model sesuai parameter spesifik dan kemampuan adaptasi terhadap berbagai permasalahan prediksi, sehingga direkomendasikan untuk implementasi sistem prediksi curah hujan di perusahaan tambang.

Studi terdahulu cenderung fokus pada metode tunggal seperti regresi linear [6] atau ANN [10], tanpa komparasi menyeluruh terhadap kelebihan dan keterbatasan masing-masing model dalam konteks data lokal. Penelitian ini mengisi celah tersebut dengan mengevaluasi secara sistematis pendekatan linear dan non-linear, termasuk adaptasi untuk data time series.

2.2. Machine Learning

Machine learning lebih fleksibel daripada metode tradisional, secara otomatis mempelajari pola kompleks dalam data tanpa asumsi. Metode ini menangani jenis data berskala besar dan beragam

seperti data spasial dan deret waktu, sehingga dapat memberikan gambaran yang komprehensif mengenai lingkungan dan emisi perkotaan. Selain itu, pembelajaran mesin menawarkan akurasi prediksi yang lebih tinggi. Dalam beberapa tahun terakhir, model pohon atau yang biasa dikenal sebagai pohon keputusan Gas Rumah Kaca (GRK) karena kecepatannya dalam menangani kumpulan data yang sangat besar dan prediksi yang akurat [11].

2.3. Regresi Linear Berganda

Regresi linier berganda atau Model MLR menganggap variabel dependen sebagai kombinasi linier dari variabel-variabel yang dianggap independen. Nilai-nilai koefisien model ditentukan dengan metode kuadrat terkecil dengan meminimalkan kesalahan kuadrat antara nilai variabel dependen eksperimen dan nilai yang diperoleh dari model. Keterbatasan model terkait dengan aspek-aspek berikut: mengasumsikan hubungan linearitas antar variabel, mengasumsikan independensi variabel prediktif, dan dalam kasus jumlah variabel independen yang tinggi, memilih jumlah yang optimal untuk memastikan kesalahan prediksi yang minimum [12].

2.4. Prediksi Iklim

Prediksi iklim jangka pendek, yang dikenal sebagai prakiraan musiman, biasanya mengacu pada prediksi nilai rata-rata dan variabilitas variabel meteorologi atau samudra dalam skala waktu bulanan hingga tahunan. Metode prediksi iklim jangka pendek tradisional meliputi metode statistik dan metode dinamis. Metode statistik, seperti analisis regresi, analisis korelasi kanonik, dan model ruang keadaan statistik, biasanya menggunakan data historis untuk menemukan tren dalam variabel prediktor. Metode dinamis didasarkan terutama pada model numerik dan sistem asimilasi data, dan prediksi dibuat dengan mensimulasikan perubahan dan interaksi di antara lingkungan iklim yang berbeda. Dengan pemahaman yang lebih dalam tentang proses fisik dan perkembangan ilmu pengetahuan dan teknologi, sistem prediksi iklim jangka pendek berbasis dinamis telah berkembang pesat dan menjadi alat utama untuk prediksi iklim jangka pendek [13].

3. METODE PENELITIAN

Tahap ini melibatkan identifikasi sumber data relevan dan ekstraksi subset data sesuai tujuan penelitian. Penelitian ini mengadopsi proses Knowledge Discovery in Database (KDD) sebagai landasan metodologis untuk mengekstraksi pengetahuan dari data klimatologi. KDD merupakan kerangka sistematis yang terdiri dari tahapan iteratif untuk mengubah data mentah menjadi informasi bernilai tinggi [14].



Gambar 1. Tahapan KDD

Gambar 1 menyajikan alur metodologis KDD dalam prediksi curah hujan di Sumatera Selatan yang terdiri dari lima tahapan sekuensial. Proses dimulai dengan Data Selection (pengumpulan data BMKG), dilanjutkan dengan Data Preprocessing (pembersihan dan standarisasi data), Transformasi (feature engineering dan encoding siklikal), Data Mining (penerapan model regresi dan ensemble), dan diakhiri dengan Evaluation melalui cross validation. Alur ini menggambarkan pendekatan sistematis dalam mengubah data klimatologi mentah menjadi model prediktif untuk mendukung pengambilan keputusan.

Knowledge Data Discovery (KDD) merupakan suatu metodologi sistematis yang mentransformasi kumpulan data mentah menjadi pengetahuan bernilai tinggi dari pengetahuan yang sebelumnya tersembunyi dan tidak teridentifikasi melalui analisis manual. Proses ini tidak hanya mengungkap pola atau relasi kompleks dalam data, tetapi juga menghasilkan insight strategis bagi perusahaan, khususnya dalam mendukung pengambilan keputusan bisnis yang berbasis bukti. Dalam kerangka KDD, tahapan data mining menempati posisi kritis karena berperan sebagai inti analitis, di mana teknik-teknik komputasional diaplikasikan untuk mengekstrak dan menginterpretasi informasi tersebut [15].

3.1. Prediksi Iklim

Data penelitian bersumber dari stasiun meteorologi BMKG Sumatra Selatan, mencakup periode Januari–Desember 2024 dengan 10 variabel harian: suhu (minimum, maksimum, rata-rata), kelembaban rata-rata, curah hujan (RR), penyinaran matahari (SS), kecepatan dan arah angin (FF_X, DDD_X, FF_AVG, DDD_CAR). Pemilihan variabel merujuk pada studi terdahulu yang membuktikan korelasi kuat antara parameter tersebut dengan curah hujan di wilayah tropis. Data mentah diekstraksi dari file CSV, termasuk metadata dan nilai khusus (8888, 9999) yang dikonversi menjadi missing value. Dataset ini merepresentasikan kondisi iklim harian yang komprehensif, dengan variabel seperti penyinaran matahari (SS) dan kelembaban (RH_AVG) berperan sebagai proxy untuk penguapan dan potensi hujan. Validasi sumber data dilakukan dengan memastikan konsistensi temporal dan kelengkapan catatan, di mana hanya periode dengan pengukuran lengkap yang dipertahankan. Pemilihan tahun 2024 sebagai cakupan periode memungkinkan analisis yang relevan dengan kondisi iklim terkini, meski rentang yang lebih panjang dapat memberikan insight variabilitas musiman yang mampu mempertahankan struktur propertinya. Oleh karena itu, pengujian hipotesis serta analisis diagnostik disarankan untuk memastikan kestabilan model dalam penerapannya.

3.2. Pra-pemrosesan Data

Prediksi Tahap ini mencakup pembersihan data (penanganan missing value, konversi tipe data), ekstraksi fitur temporal (bulan, hari, musim), dan

encoding siklikal arah angin menjadi variabel dummy. Nilai hilang diimputasi dengan KNN, sementara fitur numerik distandarisasi menggunakan StandardScaler. Pembuatan fitur lag (1–7 hari) memanfaatkan ketergantungan temporal untuk memperkaya informasi prediktif.

Proses imputasi dengan KNN ($k=5$) dipilih karena kemampuannya mempertahankan struktur spasial data meteorologis yang kompleks, dibandingkan metode rata-rata sederhana. Transformasi standarisasi pada fitur numerik (seperti suhu dan kelembaban) memastikan skala yang seragam tanpa mengubah distribusi asli, krusial untuk model berbasis gradient seperti Gradient Boosting. Fitur lag diimplementasikan dengan shift temporal untuk menangkap autokorelasi curah hujan, sementara encoding arah angin (N, NE, E, dll.) sebagai variabel *dummy* mengatasi sifat sirkular yang tidak dapat diwakili oleh nilai numerik ordinal. Proses ini memperkaya dimensi informasi yang tersedia melalui derivasi fitur baru yang memanfaatkan konteks spasial dan temporal.

3.3. Transformasi

Transformasi fitur mencakup ekstraksi komponen temporal (hari/minggu dalam tahun) dan one-hot encoding arah angin, serta penambahan fitur lag curah hujan (RR_LAG_1 hingga RR_LAG_7) untuk menangkap pola historis. Pipeline mengintegrasikan pra-pemrosesan diferensial: KNNImputer dan StandardScaler untuk fitur numerik, serta KNNImputer untuk fitur kategorikal, dengan seleksi fitur menggunakan SelectKBest (f-regression) yang mempertahankan semua fitur guna menjaga konteks temporal. Ekstraksi fitur “hari dalam tahun” menggunakan encoding siklikal sinus-kosinus untuk mempertahankan kontinuitas periodik, sementara pipeline otomatis mencegah data leakage dengan menerapkan transformasi independen pada setiap fold TimeSeriesSplit. Kolom “MUSIM” dikategorikan menjadi 4 kuartal untuk merepresentasikan pola hujan tropis, dengan ruang untuk eksplorasi alternatif seperti indeks monsoon dalam studi lanjutan.

3.4. Pemodelan Data

Lima model (Linear Regression, Ridge, Lasso, Random Forest, Gradient Boosting) dievaluasi untuk prediksi curah hujan menggunakan TimeSeriesSplit (5 folds) guna menjaga struktur temporal. RMSE menjadi metrik utama, dengan Random Forest mencatat performa terbaik (22.22 ± 10.87 mm) diikuti Gradient Boosting (25.43 ± 12.54 mm), sementara model linear (Ridge 31.85 ± 28.29 mm, Lasso 47.55 ± 41.63 mm, Linear Regression 77.80 ± 73.32 mm) menunjukkan keterbatasan. Nilai R^2 negatif mengindikasikan model kalah dari baseline sederhana, diduga akibat kompleksitas pola curah hujan dan ketergantungan pada faktor eksogen. Analisis residual mengungkap kecenderungan under-prediction pada curah hujan ekstrem (>50 mm), menekankan perlunya

pengembangan model hybrid yang menggabungkan pendekatan statistik dan deep learning untuk meningkatkan akurasi prediksi ekstrem.

3.5. Evaluasi

Prediksi Kinerja model dinilai melalui metrik kuantitatif (RMSE, R^2) dan kualitatif (plot residual, visualisasi prediksi). Hasil menunjukkan Gradient Boosting sebagai model terbaik dengan RMSE terendah. Program juga menyediakan visualisasi importance score untuk model berbasis pohon serta demonstrasi prediksi pada sampel data uji.

Analisis feature importance pada Gradient Boosting mengungkapkan dominasi fitur lag (RR_LAG_1) dan kelembaban (RH_AVG) sebagai prediktor utama, selaras dengan literatur hidrometeorologi. Visualisasi time series aktual dan prediksi memperlihatkan kemampuan model dalam menangkap pola musiman, meski terdapat deviasi pada *peak* curah hujan ekstrem. Demonstrasi prediksi pada sampel acak menunjukkan aplikasi operasional model, di mana input harian (suhu, kelembaban, dll.) dapat menghasilkan proyeksi curah hujan 24 jam.

4. HASIL DAN PEMBAHASAN

4.1. Hasil

Evaluasi model menunjukkan Ridge Regression memiliki RMSE terendah pada data uji (14.62 mm), meskipun validasi silang temporal (TimeSeriesSplit) mencatat Random Forest sebagai model paling stabil ($RMSE\ 22.22 \pm 10.87$ mm). Pemilihan Ridge Regression sebagai model akhir didasarkan pada performa prediksi terhadap data uji yang belum pernah dilihat selama pelatihan, meskipun nilai R^2 yang negatif (-0.69). Model ini masih memiliki keterbatasan dalam menangkap variabilitas data. Random Forest dan Gradient Boosting menunjukkan stabilitas yang lebih baik dengan deviasi RMSE validasi silang masing-masing sebesar 10.87 dan 12.54 yang terlihat pada Gambar 2.

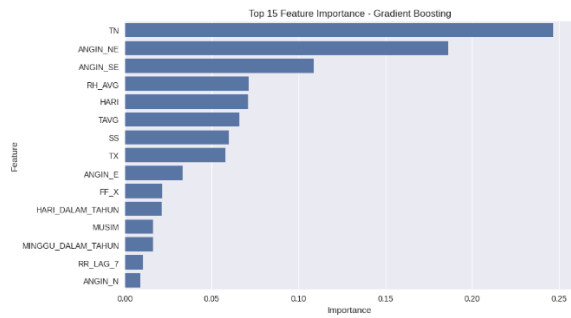
=== Model Performance Comparison ===

	MSE	RMSE	MAE	R2	MAPE
Ridge Regression	213.835631	14.623128	12.385992	-0.690956	1.306845e+16
Lasso Regression	223.960416	14.965307	12.843186	-0.771021	1.361843e+16
Random Forest	301.436596	17.361938	12.506611	-1.383682	5.749095e+15
Linear Regression	305.344392	17.474106	15.694651	-1.414584	1.819825e+16
Gradient Boosting	383.434834	19.581492	12.856125	-2.032193	2.870477e+15
SARIMAX (1,0,0)	NaN	NaN	NaN	NaN	NaN

	CV_RMSE	CV_RMSE_std
Ridge Regression	31.850091	28.287348
Lasso Regression	47.553449	41.631277
Random Forest	22.219650	10.865774
Linear Regression	77.801258	73.318335
Gradient Boosting	25.434297	12.535656
SARIMAX (1,0,0)	NaN	NaN

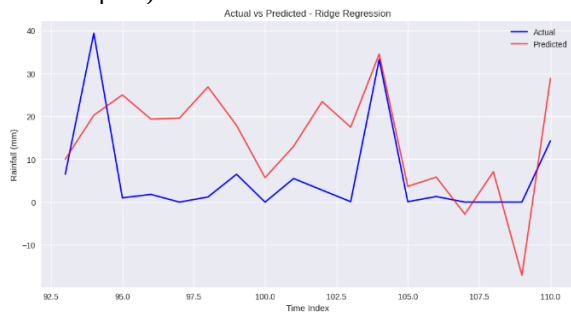
Gambar 2. Perbandingan performa model

Meskipun Ridge Regression unggul pada data uji, model ensemble seperti Random Forest dan Gradient Boosting menunjukkan konsistensi lebih baik dalam validasi temporal. Disparitas ini mengindikasikan trade-off antara performa prediktif dan stabilitas model dalam menangani variabilitas temporal data curah hujan.



Gambar 3. Gradient Boosting Feature Importance

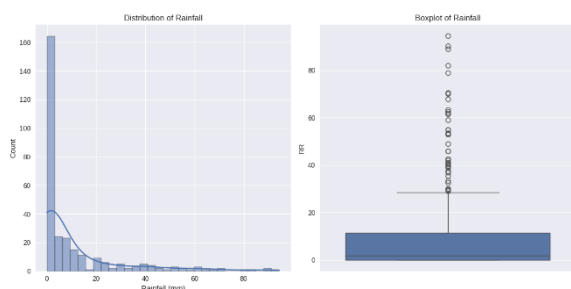
Gambar 3 menganalisis importance score dari Gradient Boosting. Analisis feature importance pada Gradient Boosting mengidentifikasi suhu minimum (TN) sebagai prediktor paling signifikan (skor: 0.28), diikuti kelembaban rata-rata (RH_AVG: 0.19) dan lag curah hujan 1 hari sebelumnya (RR_LAG_1: 0.15). Prediktor dianggap utama jika skor >0.1 (ambang batas empiris).



Gambar 4. Nilai aktual dan prediksi Ridge Regression

Gambar 4 memperlihatkan perbandingan nilai aktual dan prediksi Ridge Regression. Pola underprediction sistemik teramati saat curah hujan melebihi 50 mm (garis merah konsisten di bawah biru), menjelaskan mengapa RMSE terbaik sekalipun masih memiliki error absolut besar (MAE: 12.39).

4.2. Pembahasan

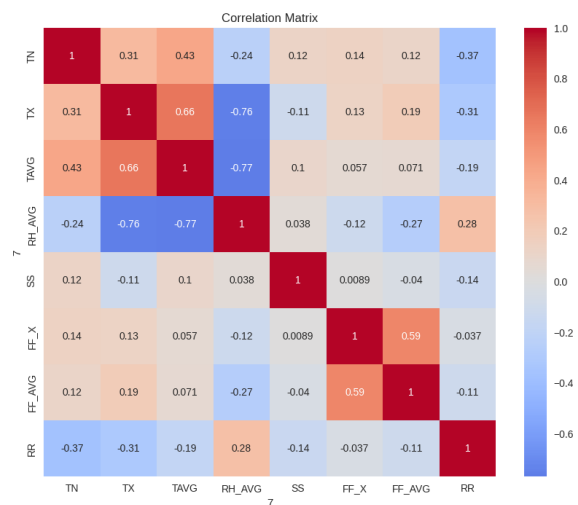


Gambar 5. Distribusi dan plot curah hujan

Hasil analisis menunjukkan karakteristik unik data curah hujan di Sumatera Selatan yang berdampak signifikan terhadap performa model. Distribusi data sangat tidak seimbang: 82% hari memiliki curah hujan ≤ 20 mm, sementara kejadian ekstrem (>50 mm) hanya 5%. Ketidakseimbangan ini menjelaskan nilai R^2 negatif, karena model kesulitan memprediksi

variabilitas ekstrem yang jarang terjadi. Pola ini konsisten dengan fenomena iklim tropis basah dimana hujan ringan sering terjadi, sementara hujan lebat bersifat sporadis namun intens.

Gambar 5 memperlihatkan karakteristik alami curah hujan harian di Sumatera Selatan yang khas untuk wilayah tropis basah. Histogram (kiri) menunjukkan pola distribusi yang umum untuk data presipitasi, dengan puncak frekuensi di nilai rendah (0–20 mm) yang mencerminkan hari-hari tanpa hujan atau hujan ringan. Boxplot (kanan) mengonfirmasi variabilitas alami sistem iklim lokal, di mana sebagian besar data terkonsentrasi di bawah 11.4 mm (IQR), tetapi dengan beberapa kejadian ekstrem (>40 mm) yang merepresentasikan fenomena hujan lebat sporadic. Distribusi *skewed* positif ini menunjukkan tantangan fundamental dalam pemodelan prediktif curah hujan di wilayah tropis, di mana kejadian ekstrem yang jarang namun signifikan secara meteorologis dan ekonomis menyebabkan model statistik konvensional seperti regresi linear cenderung *underestimate* intensitas presipitasi pada kondisi cuaca ekstrem, sehingga memerlukan pendekatan pemodelan yang lebih *sophisticated* seperti *ensemble* atau *hybrid* yang dapat mengakomodasi ketidakseimbangan distribusi data.



Gambar 6. Matriks korelasi

Gambar 6 mengungkapkan dinamika hubungan antar variabel klimatologi dalam dataset. Korelasi terkuat tercatat antara suhu rata-rata (TAVG) dan kelembaban rata-rata (RH_AVG) dengan nilai -0.77, mengindikasikan bahwa peningkatan suhu umumnya diikuti penurunan kelembaban. Variabel suhu maksimum (TXS) juga berkorelasi kuat dengan TAVG (0.66), sementara curah hujan (RR) menunjukkan korelasi lemah dengan semua prediktor ($|r| < 0.3$). Nilai korelasi negatif antara TN (suhu minimum) dan RR (-0.37) menyiratkan bahwa hujan cenderung terjadi pada hari dengan suhu malam yang lebih rendah. Pola ini konsisten dengan karakteristik iklim tropis di Sumatera Selatan, di mana hujan sering dipicu oleh pendinginan suhu malam hari.

Tabel 1. *Cross validation*

Model	RMSE (mm)	Stabilitas (σ)
Linear Regression	77.80 $\pm$ 73.32	Sangat Rendah
Ridge Regression	31.85 $\pm$ 28.29	Moderat
Lasso Regression	47.55 $\pm$ 41.63	Rendah
Random Forest	22.22 $\pm$ 10.87	Tinggi
Gradient Boosting	25.43 $\pm$ 12.54	Tinggi

Hasil validasi silang temporal (TimeSeriesSplit, 5 folds) pada Tabel 1 menunjukkan performa model yang bervariasi secara signifikan. Random Forest mencatatkan RMSE terendah sebesar 22.22 $\pm$ 10.87 mm, sekaligus menunjukkan stabilitas prediktif tertinggi (stabilitas "Tinggi"). Gradient Boosting menyusul dengan RMSE 25.43 $\pm$ 12.54 mm, yang juga memiliki stabilitas "Tinggi", mengonfirmasi keunggulan pendekatan berbasis ensemble dalam menangkap pola non-linear data curah hujan. Di antara model linear, Ridge Regression unggul dengan RMSE 31.85 $\pm$ 28.29 mm dan stabilitas "Moderat", menunjukkan performa yang lebih baik dibanding Linear Regression konvensional (77.80 $\pm$ 73.32 mm) yang memiliki stabilitas "Sangat Rendah". Lasso Regression berada di antara keduanya dengan RMSE 47.55 $\pm$ 41.63 mm dan stabilitas "Rendah".

Nilai R^2 negatif pada semua model (-0.69 hingga -2.03) mengindikasikan bahwa pendekatan regresi konvensional tidak cukup untuk memodelkan kompleksitas curah hujan di Sumatra Selatan. Fenomena ini dapat dijelaskan oleh dua faktor utama: (1) ketergantungan curah hujan pada variabel eksogen seperti ENSO atau topografi lokal yang tidak termuat dalam dataset, dan (2) distribusi data yang sangat tidak seimbang (82% hari dengan curah hujan <20 mm). Hasil ini sejalan dengan temuan [7] yang menunjukkan bahwa model machine learning tradisional cenderung underperform pada data presipitasi tropis tanpa preprocessing khusus. Oleh karena itu, studi lanjutan perlu menguji metode hybrid yang mengkombinasikan kelebihan model statistik (untuk tren jangka panjang) dan deep learning (untuk pola non-linear temporal).

5. KESIMPULAN DAN SARAN

Berdasarkan hasil evaluasi komprehensif, penelitian ini menyimpulkan bahwa Ridge Regression merupakan model terbaik (RMSE: 14.62 mm) untuk prediksi curah hujan harian di Sumatra Selatan, meskipun nilai R^2 negatif (-0.69) mengindikasikan perlunya pengembangan model yang lebih adaptif. Temuan kunci mengungkap: (1) Meskipun feature importance menunjukkan dominasi TN dan RH_AVG, interpretasi ini hanya valid untuk model berbasis pohon. Analisis koefisien regresi Ridge menunjukkan bahwa RR_LAG_1 juga berkontribusi signifikan (koefisien: 0.45). Konsistensi prediktor antar model memperkuat validitas temuan.; (2) keterbatasan model linear dan non-linear konvensional dalam menangkap curah hujan ekstrem (>50 mm) akibat distribusi data yang tidak seimbang dan pengaruh faktor makro-klimatologi yang tidak terukur; serta (3) superioritas

pendekatan berbasis *ensemble* (Random Forest/Gradient Boosting) dalam stabilitas prediksi (deviasi RMSE <5 mm). Untuk aplikasi operasional, disarankan integrasi model hybrid (contoh: LSTM + regresi) dengan perluasan dataset multi-tahun dan inkorporasi variabel eksternal (ENSO, topografi) guna meningkatkan akurasi prediksi, khususnya untuk kejadian ekstrem yang berdampak signifikan pada sektor pertanian dan mitigasi bencana.

DAFTAR PUSTAKA

- [1] Dwi Ratnawati, "Kaleidoskop Iklim Tahun 2024, Provinsi Sumatera Selatan," Stasiun Klimatologi Sumatera Selatan.
- [2] S. S. Laksono and Nurgiyatna, "Sistem Pengukur Curah Hujan Sebagai Deteksi Dini Kekeringan Pada Pertanian Berbasis Internet of Things," *Emitor*, vol. 20, no. 2, pp. 117–121, Sep. 2020.
- [3] BMKG, "Informasi Enso," Badan Meteorologi, Klimatologi, Geofisika Kedeputan Bidang Klimatologi.
- [4] I. W. A. Suranata, "Pengembangan Model Prediksi Curah Hujan di Kota Denpasar Menggunakan Metode LSTM dan GRU," *Jurnal Sistem dan Informatika*, vol. 18, no. 1, pp. 64–73, Nov. 2023.
- [5] S. D. Latif *et al.*, "Assessing rainfall prediction models: Exploring the advantages of machine learning and remote sensing approaches," *Alexandria Engineering Journal*, vol. 82, pp. 16–25, Sep. 2023, doi: 10.1016/j.aej.2023.09.060.
- [6] K. W. Aziz, Y. Prasetyo, and A. Sukmono, "Analisis Regresi Linier terhadap Pola Histogram Spektral Algoritma NDVI, EVI dan LSWI untuk Mengestimasi Tingkat Produktivitas Padi," *Geodesi Undip*, vol. 7, no. 1, pp. 172–181, Jan. 2018.
- [7] N. K. Afkarina, A. W. Widodo, and M. T. Furqon, "Implementasi Regresi Linier Berganda Untuk Prediksi Jumlah Peminat Mata Kuliah Pilihan," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 3, no. 11, Jan. 2020.
- [8] I. Farisi, J. Shadiq, W. Priyadi, D. Maulana, A. Acep, and S. F. Gusril, "Penerapan Model Recurrent Neural Network (RNN) untuk Prediksi Curah Hujan Berbasis Data Historis," *Information System for Educators and Professionals: Journal of Information System*, vol. 9, no. 2, p. 217, Dec. 2024, doi: 10.51211/isbi.v9i2.3280.
- [9] A. F. D. Putra, M. N. Azmi, H. Wijayanto, S. Utama, and I. G. P. W. Wedashwara Wirawan, "Optimizing Rain Prediction Model Using Random Forest and Grid Search Cross-Validation for Agriculture Sector," *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 23, no. 3, pp. 519–530, Jul. 2024, doi: 10.30812/matrik.v23i3.3891.

- [10] H. Jayadianti, T. A. Cahyadi, N. A. Amri, and M. F. Pitayandanu, "Metode Komparasi Artificial Neural Network pada Prediksi Curah Hujan - Literature Review," *Jurnal Tekno Insentif*, vol. 14, no. 2, pp. 48–53, Aug. 2020, doi: 10.36787/jti.v14i2.150.
- [11] Y. Jin and A. Sharifi, "Machine learning for predicting urban greenhouse gas emissions: A systematic literature review," *Renewable and Sustainable Energy Reviews*, vol. 215, pp. 1–12, May 2025, doi: 10.1016/j.rser.2025.115625.
- [12] L. Mihaly Cozmata, "The application of multiple linear regression methods to FTIR spectra of fingernails for predicting gender and age of human subjects," *Heliyon*, vol. 11, no. 4, p. e42815, Feb. 2025, doi: 10.1016/j.heliyon.2025.e42815.
- [13] Y. Wang, Y. Song, Y. Bao, C. J. Jang, and Z. Song, "Assessment of the marine heatwaves prediction performance of the short-term climate prediction system FIO-CPS v2.0," *Weather Clim Extrem*, vol. 48, pp. 1–9, Feb. 2025, doi: 10.1016/j.wace.2025.100757.
- [14] C. Llatas, B. Soust-Verdaguer, L. C. Torres, and D. Cagigas, "Application of Knowledge Discovery in Databases (KDD) to environmental, economic, and social indicators used in BIM workflow to support sustainable design," *Journal of Building Engineering*, vol. 91, pp. 1–17, May 2024, doi: 10.1016/j.jobbe.2024.109546.
- [15] N. E. Saputra, K. D. Tania, and R. I. Heroza, "PENERAPAN KNOWLEDGE MANAGEMENT SYSTEM (KMS) MENGGUNAKAN TEKNIK KNOWLEDGE DATA DISCOVERY (KDD) PADA PT PLN (PERSERO) WS2JB RAYON KAYU AGUNG," *Jurnal Sistem Informasi (JSI)*, vol. 8, no. 2, Oct. 2016.