



PENERAPAN *BAYESIAN RIDGE REGRESSION* UNTUK PREDIKSI CURAH HUJAN JANGKA PENDEK

Sri Ningrum¹, Farid Akbar Siregar²

Jurusan Teknologi Informasi, Ilmu Komputer Dan Teknologi Informasi,
Universitas Muhammadiyah Sumatera Utara
Email: sningrum814@gmail.com

Abstrak

Penelitian ini bertujuan untuk menerapkan metode *Bayesian Ridge regression* dalam memprediksi curah hujan jangka pendek berdasarkan data historis curah hujan dan variabel meteorologi pendukung. Prediksi curah hujan merupakan salah satu aspek penting dalam bidang klimatologi dan mitigasi bencana karena memiliki pengaruh terhadap sektor pertanian, pengelolaan sumber daya air, dan aktivitas masyarakat. Metode *Bayesian Ridge regression* dipilih karena mampu mengatasi permasalahan multikolinearitas antar variabel prediktor serta mengurangi resiko overfitting melalui pendekatan probabilistik Bayesian. Data penelitian diperoleh dari Kaggle yang terdiri atas 516 data observasi dan ditingkatkan melalui proses augmentasi hingga mencapai 1000 dataset. Variabel yang digunakan meliputi curah hujan periode sebelumnya, suhu udara, kelembapan udara, tekanan udara, dan kecepatan angin. Tahapan penelitian meliputi preprocessing data, standardisasi menggunakan StandardScaler, pembagian data training dan testing dengan rasio 80:20, pelatihan model, serta evaluasi menggunakan *Mean Squared Error (MSE)*, *Mean Absolute Error (MAE)*, dan *R-Squared (R^2)*. Hasil penelitian Menunjukkan bahwa model Bayesian Ridge Regression mampu menghasilkan prediksi curah hujan dengan nilai MAE sebesar 15,70 mm, MSE sebesar 459.09, dan R^2 sebesar -0,0127. Meskipun model belum mampu menangkap seluruh pola non-linear dan lonjakan curah hujan ekstream, metode ini tetap menunjukkan kestabilan dalam memprediksi tren umum curah hujan harian. Dengan demikian, Bayesian Ridge regression dapat digunakan sebagai pendekatan alternatif dalam sistem prediksi curah hujan jangka pendek berbasis machine learning.

Kata Kunci : *Bayesian Ridge Regression*, prediksi curah hujan, machine learning, data meteorologi, *Mean Squared Error (MSE)*.

Article History

Received: September 2026
Reviewed: September 2026
Published: September 2026

Plagiarism Checker
No 237.u873247.fh8374

Copyright : Author
Publish by : Kohesi



This work is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/)

PENDAHULUAN

Prediksi curah hujan jangka pendek memiliki peran vital dalam mendukung manajemen sumber daya air, sektor pertanian, dan mitigasi risiko bencana hidrometeorologi. Namun, pemodelan curah hujan harian kerap dihadapkan pada tantangan karakteristik data deret waktu yang fluktuatif serta interelasi kompleks antar-variabel meteorologi. Metode regresi klasik, seperti *subset selection*, rentan terhadap masalah multikolinearitas yang menyebabkan ketidakstabilan estimasi parameter dan penurunan akurasi prediksi (Schreiber-Gregory, 2018). Kondisi multikolinearitas dapat memicu estimasi koefisien menjadi tidak stabil, bernilai besar secara tidak wajar, bahkan mengalami pergeseran tanda positif-negatif secara ekstrem (Kruschke & Liddell, 2018).

Sebagai upaya mengatasi keterbatasan metode klasik, pendekatan regularisasi seperti *Ridge Regression* banyak diterapkan. *Ridge Regression* efektif menangani multikolinearitas dengan menyeimbangkan *bias* dan *varians* melalui penambahan penalti kuadrat pada parameter



model (Hodson et al., 2021). Beberapa penelitian terdahulu membuktikan efektivitas pendekatan ini; Siswanto dkk. (2025) menerapkan *Linear Ridge Regression* dengan *MM-Estimator* (LRMM) untuk *statistical downscaling* dan menghasilkan korelasi tinggi (0,94) serta R^2 sebesar 86,57%. Meski demikian, studi Alvines dkk. (2025) menunjukkan bahwa *Ridge Regression* standar masih memiliki keterbatasan dalam menangkap variabilitas penuh data curah hujan ekstrem, ditandai dengan nilai R^2 negatif pada data uji yang tidak seimbang, meskipun menghasilkan *Root Mean Squared Error* (RMSE) terendah.

Untuk menyempurnakan pendekatan regularisasi tersebut, *Bayesian Ridge Regression* (BRR) menawarkan modifikasi strategis dengan mengintegrasikan prinsip Bayesian dalam estimasi parameter. Berbeda dengan regresi linier biasa atau *Ridge Regression* konvensional yang hanya menghasilkan estimasi poin tunggal (*point estimate*), BRR menyajikan distribusi probabilitas untuk setiap koefisien. Pendekatan ini memberikan ketahanan lebih terhadap *overfitting* serta mampu mengkuantifikasi ketidakpastian dalam data cuaca (Kruschke & Liddell, 2018). Mekanisme ini sangat adaptif untuk meramalkan curah hujan jangka pendek (1 hingga 7 hari) yang sarat akan ketidakpastian.

Penelitian ini bertujuan untuk mengimplementasikan dan menguji performa *Bayesian Ridge Regression* dalam memprediksi curah hujan harian hingga mingguan berbasis data meteorologi (suhu, kelembapan, tekanan udara, dan curah hujan historis). Kinerja model dievaluasi menggunakan metrik *Mean Squared Error* (MSE), yang dipilih karena ketegasannya dalam memberikan penalti terhadap kesalahan prediksi ekstrem (Samadi-Koucheksaraee & Chu, 2024). Sebagai wujud kontribusi praktis, pemodelan berbasis Python ini diintegrasikan ke dalam antarmuka berbasis *web* guna memudahkan visualisasi dan pengambilan keputusan. Melalui penelitian ini, diharapkan diperoleh gambaran komprehensif mengenai tingkat akurasi dan efektivitas BRR dalam mengelola fluktuasi data iklim lokal.

METODE

1. Data Penelitian dan Variabel

Penelitian ini menggunakan dataset sekunder berupa data meteorologi *time-series* harian yang bersumber dari stasiun pengamatan cuaca (diakses melalui Kaggle). Dataset awal terdiri dari 516 observasi yang kemudian diekstensi melalui teknik augmentasi data deret waktu menjadi 1.000 data observasi untuk memperluas variasi pemodelan.

Variabel yang digunakan dalam pemodelan dikategorikan menjadi dua jenis:

- **Variabel Terikat (Y):** Curah hujan harian (mm).
- **Variabel Bebas (X):** Meliputi variabel meteorologi pendukung yaitu suhu udara ($^{\circ}\text{C}$), kelembapan udara (%), tekanan udara (hPa), arah angin ($^{\circ}$), dan *lag* curah hujan periode sebelumnya (Y_{t-1}).

2. Pra-Pemrosesan dan Pembagian Data

Untuk menjaga performa dan integritas pemodelan deret waktu, dilakukan beberapa tahapan pra-pemrosesan data:

1. **Pembersihan Data:** Memeriksa dan menangani nilai yang hilang (*missing values*) serta penyesuaian format data.
2. **Pembentukan *Lag Feature*:** Membuat variabel *lag* curah hujan (Y_{t-1}) guna menangkap ketergantungan temporal (*autocorrelation*).
3. **Standardisasi Data:** Melakukan normalisasi skala fitur menggunakan *Z-score standardization* agar setiap variabel memiliki mean = 0 dan varians = 1. Langkah ini krusial untuk mencegah dominasi fitur ber-skala besar terhadap proses penalti regularisasi (Samadi-Koucheksaraee & Chu, 2024).
4. **Pembagian Data (*Data Splitting*):** Data dibagi secara kronologis (tanpa *random shuffling* guna mencegah *data leakage*) dengan proporsi 80% sebagai data latih (*training set*) untuk pembentukan model dan 20% sebagai data uji (*testing set*) untuk evaluasi.



3. Pemodelan Bayesian Ridge Regression (BRR)

Metode *Bayesian Ridge Regression* memadukan prinsip regularisasi “L2” (Ridge) dengan pendekatan inferensi probabilistik Bayes (Assaf et al., 2019). Secara matematis, hubungan antara variabel respons “Y” dan variabel prediktor “X” dirumuskan sebagai:

$$Y = Xw + \epsilon$$

di mana “w” adalah vektor bobot koefisien dan epsilon, $\text{sim, } \mathcal{N}(0, \alpha^{-1} \mathbf{I})$ mewakili *noise* Gaussian dengan presisi alpha.

Dalam pendekatan Bayesian, koefisien “w” diasumsikan memiliki distribusi *prior* Gaussian nol-rata-rata yang dikontrol oleh hiperparameter presisi “lambda”:

$$p(w|\lambda) = \mathcal{N}(w|0, \lambda^{-1} \mathbf{I})$$

Estimasi parameter dilakukan secara probabilistik dengan memperbarui distribusi *posterior* melalui aturan Bayes $p(w \text{ vert } Y) \propto p(Y \text{ vert } w)p(w)$. Berbeda dengan *Ridge Regression* standar yang menghasilkan nilai estimasi tunggal, BRR secara otomatis memperbarui estimasi hiperparameter (alpha dan lambda) dari data secara terintegrasi (*empirical Bayes*) sehingga memberikan distribusi prediksi lengkap beserta estimasi ketidakpastiannya (*uncertainty estimate*).

4. Evaluasi Performa Model

Performa estimasi model BRR pada data uji dievaluasi menggunakan metrik *Mean Squared Error* (MSE) untuk mengukur tingkat dispersi dan kesalahan prediksi:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

di mana “ y_i ” adalah nilai curah hujan aktual, hat “ y_i ” adalah nilai prediksi, dan “n” adalah jumlah sampel data uji. Nilai MSE yang mendekati nol mengindikasikan tingkat akurasi prediksi yang semakin tinggi.

5. Implementasi Antarmuka Sistem

Sebagai bentuk penerapan praktis, model BRR yang telah dilatih diintegrasikan ke dalam sistem aplikasi *web* berbasis Python. Antarmuka sistem memfasilitasi pengguna untuk melakukan pengunggahan data CSV, pra-pemrosesan otomatis, pengaturan hiperparameter model (alpha dan lambda), visualisasi grafik perbandingan data aktual vs prediksi, serta ekspor hasil prediksi.

HASIL DAN PEMBAHASAN

1. Karakteristik Dataset dan Hasil Pra-Pemrosesan

Dataset awal yang bersumber dari Kaggle terdiri dari 516 sampel data cuaca harian. Setelah dilakukan pembersihan data (*missing values handling*), rekayasa fitur tunda (*lag feature*: Y_{t-1} dan Y_{t-3}), serta augmentasi data deret waktu, total sampel yang diolah menjadi 1.000 entri data. Dataset ini mencakup 7 fitur meteorologi utama sebagai prediktor: Kelembapan (07.00), Suhu Minimum, Curah Hujan H-1, Curah Hujan H-3, Kecepatan Angin Rata-rata, Tekanan Udara, dan Suhu Rata-rata 24 Jam.



Sebelum pemodelan, fitur-fitur tersebut ditransformasikan menggunakan *StandardScaler* agar memiliki skala seragam ($\text{mean}=0$, $\text{std}=1$). Dataset kemudian dibagi secara kronologis menjadi 80% data latih (800 sampel) dan 20% data uji (200 sampel). Klasifikasi intensitas curah hujan dikelompokkan berdasarkan standar meteorologi: Rendah ($<5\{\text{ mm}\}$), Sedang ($5\{\text{--}\}20\{\text{mm}\}$), Lebat ($20\{\text{--}\}50\{\text{ mm}\}$), dan Sangat Lebat ($>50\{\text{ mm}\}$).

2. Kinerja dan Evaluasi Model Bayesian Ridge Regression

Pengujian model dilakukan dalam dua tahap: evaluasi stabilitas pelatihan menggunakan *5-Fold Cross-Validation* (CV) pada data latih, dan pengujian kemampuan generalisasi pada data uji (*testing set*).

Metrik Evaluasi	Nilai Hasil Pengujian
Rata-rata MSE (5-Fold CV)	239.2603
MSE (Test Set)	459.0906
MAE (Test Set)	15.7052 mm
$\$R^2\$$ Score	-0.0127

Berdasarkan hasil pada tabel di atas, model menghasilkan *Mean Absolute Error* (MAE) sebesar 15,70 mm pada data uji. Nilai ini menunjukkan bahwa secara rata-rata, estimasi prediksi model memiliki deviasi sebesar 15,70 mm dari nilai aktual harian. Sebaliknya, metrik *Mean Squared Error* (MSE) meningkat signifikan dari tahap pelatihan (239.26 pada CV) menjadi 459.09 pada data uji. Peningkatan MSE ini dipicu oleh sifat kuadratik dari metrik MSE yang memberikan penalti sangat tinggi terhadap presipitasi ekstrem (*peak rainfall values*). Sementara itu, perolehan R^2 score sebesar -0.0127 mengindikasikan bahwa model regresi linier probabilistik berbasis BRR kesulitan menjelaskan variabilitas non-linear ekstrem pada data uji dibandingkan dengan sekadar menggunakan nilai rata-rata (*baseline mean model*).

3. Analisis Sensitivity Skenario Input dan Transisi Curah Hujan

Untuk memahami perilaku prediksi model terhadap kombinasi parameter meteorologi, dilakukan pengujian skenario interaktif berbasis 10 kondisi cuaca yang bervariasi dari kondisi cerah hingga ekstrem.

No	Kelembapan (%)	Suhu Min (°C)	Hujan H-1 (mm)	Hujan H-3 (mm)	Kecepatan Angin (m/s)	Tekanan Udara (hPa)	Suhu Avg (°C)	Hasil Prediksi (mm)	Kategori Prediksi
1	65	27	0	0	2	1012	30	0.00	Tidak Hujan
2	70	28	0	0	2	1013	31	0.00	Tidak Hujan
3	82	25	2	1	3	1010	28	4.90	Hujan Ringan
4	82	26	3	2	3	1010	29	7.41	Hujan Ringan
5	92	24	26	20	4	1009	27	21.76	Hujan Sedang
6	95	24	55	40	10	1010	28	26.92	Hujan Sedang
7	98	20	200	150	13	998	24	50.18	Hujan Lebat
8	100	20	200	180	12	995	23	55.83	Hujan Lebat



9	100	20	250	200	25	900	20	168.79	Hujan Sangat Lebat
10	100	18	300	250	25	900	19	175.29	Hujan Sangat Lebat

Hasil eksperimen pada tabel di atas mengonfirmasi bahwa model BRR berhasil mempelajari hubungan dinamis atmosferik secara logis:

1. **Pengaruh Kelembapan dan Tekanan:** Kenaikan kelembapan udara di atas 90% yang bersumber bersamaan dengan penurunan tekanan udara di bawah 1000 hPa secara konsisten mendorong estimasi prediktif ke kategori hujan lebat hingga sangat lebat.
2. **Korelasi Temporal (*Lag Features*):** Variabel Y_{t-1} (*rr_lag1*) dan Y_{t-3} (*rr_lag3*) memberikan kontribusi signifikan dalam mempertahankan tren durasi hujan berkepanjangan.
3. **Kontrol Penalti Pasca-Pemrosesan:** Penerapan *post-processing non-negative constraint* terbukti efektif mencegah munculnya estimasi matematis bernilai negatif pada hari-hari tanpa hujan.

4. Pembahasan Ilmiah (Discussion)

Temuan dalam penelitian ini menyoroti keunggulan sekaligus limitasi dari *Bayesian Ridge Regression* dalam ranah peramalan hidrometeorologi. Dari aspek keunggulan, penerapan regularisasi L_2 probabilistik terbukti menjaga ketahanan parameter model terhadap multikolinearitas antar-variabel cuaca (misalnya antara suhu minimum, suhu rata-rata, dan kelembapan). Nilai MAE sebesar 15,70 mm membuktikan bahwa untuk tren umum harian, BRR cukup stabil dan tidak mengalami *extreme overfitting*, didukung oleh estimasi *5-Fold Cross-Validation* yang konsisten.

Namun, keterbatasan utama BRR terletak pada **asumsi hubungan linier** antara variabel prediktor dan variabel terikat. Fenomena presipitasi atmosferik memiliki sifat non-linear stokastik yang sangat kompleks (Alvines et al., 2025). Ketika terjadi hujan lebat yang dipicu oleh lonjakan mendadak (*skewed data distribution*), pendekatan linier pada BRR cenderung menghasilkan *under-prediction* pada nilai-nilai puncak (*peaks*). Hal ini menjelaskan mengapa metrik MSE pada data uji membengkak (459.09) dan R^2 bernilai negatif (-0.0127). Hasil ini sejalan dengan temuan Alvines dkk. (2025), yang juga menemukan bahwa model regresi terregularisasi (*Ridge*) mengalami keterbatasan dalam merespons ketidakseimbangan distribusi data curah hujan harian.

KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan mengenai penerapan metode *Bayesian Ridge Regression* untuk prediksi intensitas curah hujan, dapat disimpulkan bahwa:

1. Tahap pra-pemrosesan data seperti pembersihan *missing values* dan penggunaan *Standard Scaling* terbukti krusial dalam menjaga kestabilan model, di mana dataset berhasil diseleksi menjadi 174 baris data berkualitas untuk proses pelatihan.
2. Berdasarkan hasil pengujian, model menunjukkan performa yang cukup stabil dengan nilai Mean Absolute Error (MAE) sebesar 15.70 mm, yang berarti rata-rata selisih prediksi sistem terhadap data aktual masih berada dalam rentang toleransi untuk estimasi cuaca harian.
3. Meskipun model mampu mengikuti tren umum, nilai MSE (459.09) dan R^2 Score (-0.0127) yang rendah menunjukkan bahwa model masih kesulitan dalam memprediksi lonjakan curah hujan ekstrem yang bersifat tiba-tiba dan non-linear.



4. Kesimpulan dari penelitian ini menunjukkan bahwa fitur yang paling berpengaruh terhadap prediksi curah hujan adalah kelembaban udara, suhu rata-rata 24 jam, suhu minimum, dan tekanan udara. Variabel lain seperti rr_lag dan kecepatan angin juga berpengaruh, tetapi lebih kecil.

DAFTAR PUSTAKA

- Afham, M., Nur, I. M., & Utami, W. (2017). Pemodelan Regresi Ridge Pada Kasus Curah Hujan Di Kota Semarang. *Prosiding Seminar Nasional Dan International*, 165-176.
- Alvines, M., Ravi Wijayanto, M., Archi Daffa Danendra, M., Rizky Herdiansyah, M., Karimsyah Lubis, M., Ichsan Farel Rachmad, M., Ditha Tania, K., & Meiriza, A. (2025). Komparasi Ridge Regression, Random Forest, Dan Gradient Boosting Untuk Prediksi Curah Hujan Harian Di Sumatra Selatan Berbasis Time Series Cross-Validation. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 9(4), 5742-5748. <https://doi.org/10.36040/jati.v9i4.13915>
- Assaf, A. G., Tsionas, M., & Tasiopoulos, A. (2019). Diagnosing and correcting the effects of multicollinearity: Bayesian implications of ridge regression. *Tourism Management*, 71, 1-8. <https://doi.org/10.1016/j.tourman.2018.09.008>
- Astarini, A., & Adriat, R. (2022). *Studi Perbandingan Metode Penentuan Intensitas Curah Hujan Berdasarkan Karakteristik Curah Hujan Kalimantan Barat*. 10(1), 1-7.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). Bayesian Data Analysis. In *Bayesian Data Analysis*. Chapman and Hall/CRC. <https://doi.org/10.1201/9780429258411>
- Golam Kibria, B. M., & Lukman, A. F. (2020). A new ridge-type estimator for the linear regression model: Simulations and applications. *Scientifica*, 2020. <https://doi.org/10.1155/2020/9758378>
- Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin and Review*, 25(1), 155-177. <https://doi.org/10.3758/s13423-017-1272-1>
- Muharsyah, R., Ratri, D. N., Informasi, P., Iklim, P., Meteorologi, B., & No, J. A. (2021). Meningkatkan Kualitas Prediksi Curah Hujan Musiman saat Fase ENSO Menggunakan Metode Bayesian Model Averaging (BMA), Studi Kasus : Pulau Jawa Improving Quality of Seasonal Rainfall Prediction during ENSO Phase using Bayesian Model Averaging (BMA) Meth. 45(1), 13-25
- Sahrman, S., & Upa, D. (2024). Statistical Downscaling Model with Jackknife Ridge Regression and Modified Jackknife Ridge Regression to Forecast Rainfall. *Indonesian Journal of Statistics and Its Applications*, 8, 155-165. <https://doi.org/10.29244/ijsa.v8i2p155-165>
- Samadi-koucheksaraee, A., & Chu, X. (2024). Development of a novel modeling framework based on weighted kernel extreme learning machine and ridge regression for streamflow forecasting. *Dept 2470*, 1-26.
- Saputra Saragih, M. R., & Akbar Siregar, F. (2024). APLIKASI DETEKSI GAMBAR HASIL SCAN MENGGUNAKAN METODE COSINE SIMILARITY BERBASIS MOBILE. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 8, 11598-11604. <https://doi.org/10.36040/jati.v8i6.1154>
- Schreiber-Gregory, D. N. (2018). Ridge Regression and multicollinearity: An in-depth review. *Model Assisted Statistics and Applications*, 13, 359-365. <https://doi.org/10.3233/MAS-180446>
- SISWANTO, S., Gozhi, M. Z. H., Taufik, M. I., Kalondeng, A., & Sahrman, S. (2025). Linearized Ridge Regression Modeling with MM-Estimator in Statistical Downscaling for Rainfall Forecasting. *Jurnal Matematika, Statistika Dan Komputasi*, 21(3), 796-812. <https://doi.org/10.20956/j.v21i3.4320>
- Snaan, R. A., & Shukur, O. B. (2025). Using Bayesian Ridge Regression Model and ESN for Climatic Time Series Forecasting. 14(September), 1226-1243. <https://doi.org/10.19139/soic-2310-5070-2637>
- Tew, S. Y., Boley, M., & Schmidt, D. F. (2023). Bayes beats Cross Validation: Efficient and Accurate Ridge Regression via Expectation Maximization.



Wang, Y., Pei, L., & Wang, J. (2024). Precipitation prediction in several Chinese regions using machine learning methods. *International Journal of Dynamics and Control*, 12, 1180-1196. <https://doi.org/10.1007/s40435-023-01250-1>