Seminar Nasional Sains Data 2024 (SENADA 2024)
UPN "Veteran" Jawa Timur
E-ISSN 2808-5841
P-ISSN 2808-7283

Metode Seleksi Variabel dalam Pemodelan Regresi Linear Data Curah Hujan Provinsi Lampung

Elok Fiola[1], Feryadi Yulius[2], Presilia[3], Dea Mutia Risani[4], Mika Alvionita[5], Febri Dwi Irawati[6]

[1, 2, 3, 4, 5, 6]*Program Studi Sains Data, Fakultas Sains, Institut Teknologi Sumatera*

[1]elok.122450051@student.itera.ac.id, [2]feryadi.122450087@student.itera.ac.id, [3] presilia.122450081@student.itera .ac.id, [4]dea.122450099@student.itera.ac.id, [5]mika.alvionita@sd.itera.ac.id, [6]febri.dwi@sd.itera.ac.id

Corresponding author email: mika.alvionita@sd.itera.ac.id

Abstract: *Understanding the amount of rainfall is crucial for engineering planning, especially for water-related infrastructure such as irrigation systems, dams, urban drainage, and other hydraulic structures. Consequently, accurate modeling and identification of optimal predictor variables are essential to support effective rainfall prediction for design and decision-making purposes. This research aims to identify the optimal number of variables in a linear regression model using best subset, forward stepwise, and backward stepwise selection methods. The evaluation is based on the highest Adjusted R^2 value. The analysis results indicate that the optimal number of variables in the linear regression model is five, namely the number of rainy days, average wind speed, average air humidity, average air temperature, and average minimum air temperature. The highest Adjusted R^2 value obtained is 67.1%, and the Bayesian Information Criterion (BIC) value is -5.715773, the smallest compared to other models. Additionally, the Residual Sum of Squares (RSS) value is 10383.326, which, although not the smallest, is sufficiently optimal. The Cp value is 4.380471, indicating a good fit. With the smallest BIC value, and optimal RSS and Cp values, the model with five variables is validated as the best model. The methods of best subset, forward stepwise, and backward stepwise selection also show consistency in selecting the predictor variables. Conversely, variables such as average air pressure and average solar irradiation do not exhibit a significant influence in the linear regression model.*

Keywords: *Adjusted R square, Backward Stepwise, Best Subset, Forward Stepwise, Linear Regression.*

Abstrak: Pemahaman mengenai jumlah curah hujan sangat penting untuk perancanaan teknik khususnya untuk pembangungan air misalnya irigasi, bendungan, drainase perkotaan, dan lain-lain. Oleh karena itu, pemodelan yang akurat dan identifikasi variabel prediktor yang optimal sangat diperlukan untuk mendukung prediksi curah hujan demi perancangan dan pengambilan keputusan yang efektif nantinya. Penelitian ini dilakukan dengan tujuan untuk mengidentifikasi jumlah variabel optimal dalam model regresi linear dengan menggunakan metode *best subset*, *forward stepwise*, dan *backward stepwise*. Evaluasi dilakukan berdasarkan nilai *Adjusted* R^2 tertinggi. Hasil analisis menunjukkan bahwa jumlah variabel optimal pada model regresi linear adalah lima, yaitu jumlah hari hujan, rata-rata kecepatan angin, rata-rata kelembaban udara, rata-rata suhu udara, dan rata-rata suhu udara minimum. Kemudian didapat nilai *adjusted* R^2 tertinggi yang diperoleh adalah 67,1%, serta nilai *Bayesian Information Criterion (BIC),* yaitu senilai -5,715773 merupakan nilai terkecil dibanding model lain, diikuti nilai *RSS* yaitu senilai 10383,326 yang meskipun bukan nilai *RSS* terkecil, tapi sudah cukup optimal. Sedangkan untuk nilai *Cp* berada di angka 4,380471 yang cukup baik. Dengan nilai *BIC* terkecil, nilai *RSS* dan *Cp* yang optimal membuktikan model dengan lima variabel layak untuk dijadikan model terbaik. Ketiga metode *best subset, forward stepwise, dan backward stepwise* juga menunjukkan konsistensi dalam memilih variabel prediktor yang dimasukkan. Adapun variabel rata-rata tekanan udara dan rata-rata penyinaran matahari tidak menunjukkan pengaruh signifikan dalam model regresi linear.

Kata kunci: Adjusted R2, Backward Stepwise, Best Subset, Forward Stepwise, Regresi Linear.

I. PENDAHULUAN

Curah hujan merupakan fenomena alam yang mempengaruhi banyak aspek kehidupan, termasuk pertanian, infrastruktur, dan pengelolaan air. Di wilayah Lampung, penting untuk memiliki pemahaman yang lebih baik tentang pola curah hujan guna mendukung perencanaan dan pengambilan keputusan yang lebih baik di berbagai sektor. Pada sektor pertanian misalnya, perubahan curah hujan sangat berpengaruh terhadap pola penanam bibit dan hasil panen. Selain itu, pembangunan drainase di daerah perkotaan di Provinsi Lampung memerlukan perkiraan jumlah curah hujan yang biasanya turun di daerah sana. Namun, perubahan iklim yang kerap terjadi dan tidak menentu mengakibatkan perubahan

dan harus senantiasa diperbaharui setiap waktu. Sedangkan curah hujan sendiri cukup sulit untuk diprediksi.

Salah satu cara untuk memprediksi curah hujan adalah dengan menggunakan model regresi linier. Regresi linear adalah model matematika untuk memprediksi nilai suatu variabel dependen berdasarkan satu atau lebih variabel independen. Model ini memungkinkan kita untuk mengidentifikasi dan mengukur pengaruh beberapa variabel independen terhadap variabel dependen yaitu curah hujan. Namun, tantangan utama dalam pemodelan regresi linier adalah menentukan variabel independen mana yang mempunyai pengaruh paling besar terhadap curah hujan sehingga model dapat menghasilkan prediksi yang akurat dan andal.

Banyak penelitian sebelumnya yang mencoba memprediksi curah hujan menggunakan satu atau lebih variabel independen. Pada tahun 2011, Yunus S [1] mengembangkan model regresi yang hanya menggunakan dua variabel, yaitu suhu udara dan kelembaban udara, untuk memprediksi curah hujan. Namun pemilihan variabel yang optimal seringkali dilakukan secara subjektif atau *trial and error* yang mana mungkin tidak memberikan hasil yang optimal. Oleh karena itu, penelitian ini menggunakan metode sistematis pemilihan dan validasi variabel lainnya seperti pemilihan komponen optimal, seleksi maju, dan seleksi mundur. Metode ini memungkinkan Anda menentukan secara akurat kumpulan variabel independen terpenting dalam model regresi linier dengan mempertimbangkan berbagai kriteria evaluasi, seperti *R-squared* yang disesuaikan, *Bayesian Information Criterion (BIC), Radial Sum of Squares (RSS) dan Mallows' Cp*. Dengan metode seleksi variabel, maka pemilihan model dengan jumlah variabel optimal akan lebih baik demi mendapatkan hasil prediksi yang lebih akurat.

Penelitian ini memiliki tujuan untuk mengidentifikasi jumlah variabel yang optimal dalam model regresi linier untuk memprediksi curah hujan di wilayah Lampung. Dengan menggunakan ketiga metode seleksi yang berbeda ini, kami ingin mencari sampel yang tidak hanya memiliki nilai *Adjusted R-square* paling optimal, tetapi juga memiliki nilai *BIC* dan *RSS* yang optimal serta nilai *Mallows Cp* yang baik. Kami berharap hasil penelitian ini dapat sangat membantu dalam pengelolaan curah hujan di wilayah Lampung, dan dapat digunakan dalam penelitian serupa di wilayah lain.

II. METODE PENELITIAN

Penelitian ini menggunakan analisis regresi linear untuk mengetahui hubungan antara variabel cuaca dan curah hujan di Provinsi Lampung Tahun 2022. Dengan begitu dapat mengidentifikasi faktor-faktor yang paling berpengaruh terhadap curah hujan.

2.1. Regresi Linear

Regresi linear sederhana adalah metode statistik yang digunakan untuk mengukur sejauh mana keterkaitan antara satu variabel independen *X* dan satu variabel dependen *Y*. Dalam regresi linear sederhana, variabel *X* berfungsi sebagai prediktor atau penyebab potensial, sementara variabel *Y* merupakan hasil atau respon yang diamati [2].

Pada regresi linear sederhana menggunakan (1), sebagai berikut:

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (1)$$

dimana:
Y = Variabel dependen
β_0 = Populasi Y intersep
β_1 = Koefisien kemiringan populasi

X = Variabel independen
ε = Komponen kesalahan acak

Besarnya konstanta β_1 dan β_0 dapat diduga menggunakan (2) dan (3). Pendugaan bagi koefisien kemiringan garis β_1 ialah:

$$b_1 = (n\sum X_i Y_i - \sum X_i \sum Y_i)/\left(n\sum X_i^2 - (\sum X_i)^2\right) \qquad (2)$$

Pendugaan bagi intersep β_0 ialah:

$$b_0 = (\sum Y_i - b_1 \sum X_i)/n \qquad (3)$$

b_1 = Dugaan bagi intersep β_1
b_0 = Dugaan bagi intersep β_1
X_i = Variabel independen ke-i
Y_i = Variabel dependen ke-i
n = Jumlah data [2].

1. Regresi Linear Berganda

Regresi linear berganda adalah teknik statistik yang digunakan untuk mengidentifikasi arah dan seberapa besar pengaruh dari satu variabel dependen terhadap dua atau lebih variabel independen [3]. Variabel dependen sering kali disimbolkan sebagai Y, sementara variabel independen disimbolkan sebagai X. Ketika terdapat lebih dari satu variabel independen, biasanya mereka disimbolkan dengan X_1, X_2, dan seterusnya. Model regresi antara Y dan beberapa variabel independen dapat dijelaskan dengan (4) sebagai berikut:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i \qquad (4)$$

dimana:
i = 1, 2, ..., n dan $\varepsilon_i \sim N(0, \sigma^2)$
β_0 = Intersep
β_k = Koefisien kemiringan parsial ke – k
ε_i = Error ke – i
n = Banyaknya observasi

Perbedaan dalam regresi linear berganda adalah adanya dua atau lebih variabel X. Proses penentuan koefisien dalam regresi linear berganda melibatkan pertimbangan pengaruh dari variabel independen tambahan, yang membuat perhitungannya menjadi lebih rumit [4].

$$\beta_1 = \left((\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)\right) / \left((\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2\right) \qquad (5)$$

$$\beta_2 = \left((\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)\right) / \left((\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2\right) \qquad (6)$$

$$\beta_0 = (\sum y/n) - \beta_1(\sum x_1/n) - \beta_2(\sum x_2/n) \qquad (7)$$

dimana:
β_0 = Populasi variabel dependen Y intersep
β_1 = Koefisien kemiringan populasi variabel independen kedua (X_1)
β_2 = Koefisien kemiringan populasi variabel independen kedua (X_2)
x_i = Variabel independen ke-i
y_i = Variabel dependen ke-i
n = Jumlah data

2.2. *Variabel Selection*

1. *Best Subset Selection*

Best Subset Selection merupakan metode statistik yang digunakan untuk memilih sekumpulan variabel yang paling signifikan dalam model regresi. Pendekatan ini melibatkan proses penghapusan variabel yang tidak relevan terhadap model, dengan mengeluarkan variabel yang memiliki nilai *p-value* lebih tinggi daripada ambang toleransi yang telah ditetapkan. Proses tersebut berlangsung secara berulang hingga tidak ada lagi variabel yang memenuhi syarat untuk dimasukkan atau dihapus dari model.

Dalam analisis regresi ganda, Metode *Best Subset Selection* digunakan untuk menguji hipotesis bahwa semua koefisien βk (k=1,2,…,p-1) tidak sama dengan nol. Hipotesis ini akan ditolak jika nilai F-parsial dari variabel yang dimasukkan lebih kecil daripada nilai ambang F-tabel yang telah ditentukan. Proses ini berlanjut secara iteratif hingga tidak ada lagi variabel yang memenuhi kriteria untuk dimasukkan atau dihilangkan dari model [5].

2. *Stepwise Selection*

a. *Forward Stepwise Selection*

Dalam metode *forward selection*, pembentukan model optimal dilakukan secara bertahap dengan penambahan variabel satu per satu. Regresi linier sederhana dimulai dengan memasukkan satu variabel prediktor. Langkah berikutnya melibatkan penambahan variabel prediktor baru sehingga jumlah variabel prediktor dalam model menjadi dua. Prosedur analisis menggunakan metode *forward selection* ini melibatkan evaluasi semua kemungkinan model yang dibentuk, kemudian memilih model terbaik berdasarkan kriteria R^2 terbesar. Metode *forward selection* ini mengutamakan hubungan antara variabel respons Y dengan variabel prediktor X_i yang memiliki nilai R^2 terbesar. Setiap langkah selanjutnya melibatkan penambahan variabel X berikutnya yang memiliki korelasi parsial terbesar, dan proses akan berhenti saat penambahan variabel X lain tidak memberikan peningkatan signifikan terhadap nilai R^2 [6].

b. *Backward Stepwise Selection*

Backward Stepwise Selection merupakan pendekatan statistik yang digunakan dalam pemilihan variabel yang signifikan dalam model regresi. Metode ini melibatkan langkah-langkah penghapusan variabel yang tidak berpengaruh signifikan pada model, dengan mengeliminasi variabel yang memiliki nilai *p-value* lebih tinggi daripada ambang toleransi yang ditetapkan. Proses ini berlangsung secara berulang hingga tidak ada lagi variabel yang memenuhi kriteria untuk dimasukkan atau dihilangkan dari model.

Dalam analisis regresi ganda, Metode *Backward Stepwise Selection* digunakan untuk menguji hipotesis bahwa semua koefisien βk ($k=1,2,\ldots,p-1$) tidak sama dengan nol. Hipotesis ini ditolak jika nilai F-parsial dari variabel yang disertakan lebih kecil daripada nilai ambang F-tabel yang telah ditentukan. Proses ini berlanjut secara iteratif hingga tidak ada lagi variabel yang memenuhi syarat untuk dimasukkan atau dihilangkan dari model [7].

2.3. *Pemilihan Model Optimal*

1. *Residual Sum of Square (RSS)*

Residual Sum of Square adalah hasil penjumlahan kuadrat sisa antara nilai prediksi variabel dan nilai aktual variabel. *RSS* merupakan komponen dalam analisis regresi yang berguna untuk

mengevaluasi seberapa baik model dalam menjelaskan variabel dependen (*Y*) dengan menggunakan variabel independen (*X*) [8].

$$RSS = \sum(e_i)^2 = \sum\left(Y_i - \widehat{Y_i}\right)^2 = \sum\left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right)^2 \quad (8)$$

dimana:

RSS = Jumlah kuadrat residu
e_i = residual ke-*i*
Y_i = nilai aktual dari variabel respons ke-*i*
$\widehat{Y_i}$ = nilai yang diprediksi oleh model regresi untuk observasi ke-*i*
$\widehat{\beta_0}$ = Estimasi koefisien regresi untuk intersep
$\widehat{\beta_1}$ = Estimasi koefisien regresi untuk koefisien regresi
x_i = Nilai dari variabel prediktor ke-*i* [9]

2. Koefisien Determinasi (R^2)

Menurut Ghozali[10], koefisien determinasi digunakan untuk mengukur sejauh mana model dapat menjelaskan variasi dari variabel dependen. Rentang nilai koefisien determinasi adalah antara nol dan satu. Nilai *R* yang rendah menunjukkan bahwa kemampuan variabel independen dalam menjelaskan variasi dari variabel dependen terbatas. Sebaliknya nilai yang mendekati satu menandakan bahwa variabel independen memberikan sebagian besar informasi yang diperlukan untuk memprediksi variabel dependen.

Nilai koefisien determinasi (R^2) dapat dihitung menggunakan (8):

$$R^2 = RSS_p/SS_{total} = 1 - SSE_p/SS_{total} \quad (9)$$

dimana:

R^2 = Koefisien determinasi
RSS_p = Jumlah kuadrat residu
SS_{total} = Jumlah kuadrat total
SSE = Jumlah kuadrat galat
p = Jumlah peubah bebas
n = Jumlah sampel [11]

3. *Adjusted* R^2

Menambahkan lebih banyak variabel bebas ke dalam model selalu meningkatkan nilai R^2 dan tidak pernah menurunkannya. Karena R^2 sering dapat diperbesar dengan menambahkan banyak variabel bebas, ada yang menyarankan untuk memodifikasi metrik ini dengan memperhitungkan jumlah variabel bebas dalam model. Koefisien determinasi yang disesuaikan (*adjusted coefficient multiple determination*), yang dilambangkan dengan R_{adj}, mengoreksi R^2 dengan membagi setiap jumlah kuadrat dalam rumus R^2 dengan derajat kebebasannya masing-masing.

$$adj\ R^2 = 1 - (n-1)/(n-p) = SSE_p/SS_{total} \quad (10)$$

dimana:

R^2 = Koefisien determinasi
n = Jumlah sampel
p = Jumlah peubah bebas
SSE = Jumlah kuadrat regresi
SS_{total} = Jumlah kuadrat total

Nilai R^2_{adj} hanya akan naik jika nilai $(n-p)SSE_p$ turun, karena $(n-1)SS_{total}$ tetap. Model yang baik memiliki R^2_{adj} yang besar [11].

4. *Cp-Mallow*

Estimasi yang diperoleh dari model regresi berdasarkan sebagian variabel bebas sering kali memiliki bias. Untuk mengevaluasi kebaikan model, kita menggunakan mean square error (*MSE*) yang mencakup varian dan bias. C.L. Mallow merekomendasikan statistik.

$$C - p = (SSE_p/\widehat{\alpha^2}) - (n - 2p) \qquad (11)$$

dimana:

C = Biaya
p = Jumlah peubah bebas
SSE = Jumlah kuadrat regresi
$\widehat{\alpha^2}$ = Perkiraan dari varians kesalahan

Deviasi C - p dari p digunakan sebagai indikator bias. Model optimal berdasarkan C - p adalah model yang memiliki deviasi C - p yang paling mendekati jumlah variabel dalam model [11].

5. *Bayesian Information Criterion (BIC)*

Bayesian Information Criterion adalah suatu kriteria dalam analisis statistik yang digunakan untuk memilih model yang paling cocok dengan data yang ada. *BIC* menggabungkan kriteria *likelihood*, penalti parameter, dan diterapkan ketika terdapat beberapa model yang dapat menjelaskan data. Tujuannya untuk menemukan model yang paling sesuai dengan data yang tersedia.

BIC dapat dihitung menggunakan (12):

$$BIC = k \cdot ln(n) - 2 \cdot ln(L) \qquad (12)$$

dimana:

k = Jumlah parameter dalam model
ln = Logaritma natural dari jumlah pengamatan n
n = Jumlah data yang tersedia
L = Nilai maksimum fungsi *likelihood* dari model.

Model dengan nilai *BIC* yang lebih rendah dianggap lebih cocok dengan data. *BIC* memperhitungkan baik kriteria *likelihood* maupun jumlah parameter, yang membantu menghindari *overfitting* dan memungkinkan pemilihan model yang lebih sederhana dengan parameter yang lebih sedikit. Dikembangkan oleh Gideon E. Schwarz pada tahun 1978, *BIC* mirip dengan *Akaike's Information Criterion (AIC),* tetapi memberikan penalti yang lebih besar terhadap jumlah parameter, membuatnya lebih efektif dalam mengatasi *overfitting* pada dataset yang lebih besar [12].

2.4. Studi Kasus

1. Deskripsi Data

Data yang dianalisis merupakan kumpulan data cuaca Provinsi Lampung selama tahun 2022. Dalam data ini mencangkup data Rata-rata Hari Hujan, Rata-rata Kecepatan Angin, Rata-rata Kelembapan Udara, Rata-rata Tekanan Udara, Rata-rata Penyinaran Udara, Rata-rata Suhu Udara, Rata-rata Udara Minimum, dan Jumlah Curah Hujan. Sumber data ini berasal dari Diskominfotik Provinsi Lampung melalui situs resmi Dashboard Lampung. Variabel-variabel ini digunakan sebagai prediktor untuk mengidentifikasi jumlah variabel optimal dalam model regresi linear dengan menggunakan metode *best subset*, *forward stepwise*, dan *backward stepwise*.

E-ISSN 2808-5841
P-ISSN 2808-7283

Tabel 1. Deskripsi Operasional Variabel

Variabel	Definisi Operasional	Deskripsi	Satuan
Y	Jumlah Curah Hujan	Jumlah turunnya curah hujan yang tercatat dalam milimeter kubik	mm
X_1	Rata-rata Hari Hujan	Jumlah rata-rata suhu udara Provinsi Lampung dalam hari	hari
X_2	Rata-rata Kecepatan Angin	Jumlah rata-rata kecepatan angin Provinsi Lampung dalam kilometer per jam (km/jam)	(km/jam)
X_3	Rata-rata Kelembaban Udara	Jumlah rata-rata kelembapan udara Provinsi Lampung dalam persentase kelembapan relatif	%RH
X_4	Rata-rata Tekanan Udara	Jumlah rata-rata tekanan udara Provinsi Lampung dalam pascal	(hPa)
X_5	Rata-rata Penyinaran Matahari	Jumlah rata-rata penyinaran matahari Provinsi Lampung dalam joule per meter persegi per detik	(jam/hari)
X_6	Rata-rata Suhu Udara	Jumlah rata-rata suhu udara Provinsi Lampung dalam derajat Celcius	°C
X_7	Rata-rata Suhu Udara Minimum	Jumlah rata-rata suhu udara minimum Provinsi Lampung dalam derajat celcius	°C

Dengan menggunakan data ini, pola tren dan hubungan antara variabel prediktor dan respon dapat diprediksi. Model prediktif yang dihasilkan kemudian divalidasi dengan membandingkan hasil prediksi prediksi curah hujan dengan data observasi curah hujan yang sesungguhnya. Hasil dari validasi ini akan membantu dalam mengukur akurasi model dan menilai kegunaannya dalam prakiraan curah hujan di masa mendatang.

Tabel 2. Data yang digunakan

Rata-rata Hari Hujan	Rata-rata Kecepatan Angin	Rata-rata Kelembaban Udara	Rata-rata Tekanan Udara	Rata-rata Penyinaran Matahari	Rata-rata Suhu Udara	Rata-rata Suhu Udara Minimum	Jumlah Curah Hujan
X_1	X_2	X_3	X_4	X_5	X_6	X_7	Y
19	13	100	1003,7	3,5	34,2	22,4	317,3
15	19	100	1007,7	4,7	34,4	22,0	183,2
14	22	100	1008,0	4,3	34,6	23,4	111,1
7	16	98	1009,2	5,5	35,2	23,0	120,5
16	14	100	1008,2	4,4	35,0	22,8	199,5
16	16	98	1008,9	3,1	34,2	22,0	113,6

Rata-rata Hari Hujan	Rata-rata Kecepatan Angin	Rata-rata Kelembaban Udara	Rata-rata Tekanan Udara	Rata-rata Penyinaran Matahari	Rata-rata Suhu Udara	Rata-rata Suhu Udara Minimum	Jumlah Curah Hujan
X_1	X_2	X_3	X_4	X_5	X_6	X_7	Y
5	13	98	1010,8	5,2	33,4	21,2	77,40
11	17	98	1006,5	5,2	34,2	21,9	117,4
15	19	96	1005,2	4,7	34,0	22,2	65,60
15	13	98	1010,1	4,2	34,3	22,0	211,9
16	14	99	1003,6	4,3	35,0	21,1	114,8
19	20	100	1004,3	1,4	33,6	22,4	215,8

Sumber: Diskominfotik Provinsi Lampung [13]

Tabel 2 menyajikan data bulanan curah hujan di Provinsi Lampung selama tahun 2022, dari bulan Januari hingga Desember. Data ini mencangkup variabel prediktor seperti X_1, X_2, X_3, X_4, X_5, X_6, dan X_7 serta variabel respon yaitu Y. Sebagai contoh, pada bulan dengan rata-rata 5 hari hujan, kecepatan angin rata-rata adalah 13 km/jam, kelembaban udara mencapai 98% RH, dan tekanan udara berada pada 1010,8 hPa. Penyinaran matahari rata-rata tercatat selama 5,2 jam per hari, dengan suhu udara rata-rata sebesar 33,4°C dan suhu udara minimum rata-rata 21,2°C. Pada bulan tersebut, jumlah curah hujan yang tercatat adalah 77,4 mm. Data ini memberikan gambaran menyeluruh tentang kondisi cuaca bulanan di Provinsi Lampung dan digunakan untuk mengidentifikasi pola serta tren yang mempengaruhi curah hujan.

2. Statistika Deskriptif

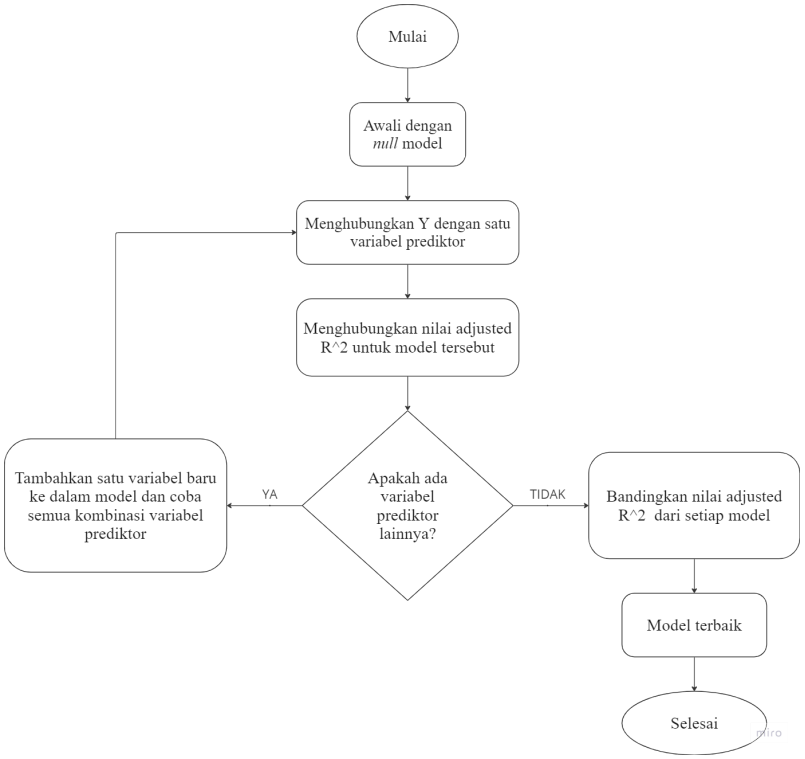
Tabel 3. Statistika deskriptif

	Rata-rata	Simpangan Baku	N
Rata-rata Hari Hujan	14	4,31	12
Rata-rata Kecepatan Angin	16,33	3,01	12
Rata-rata Kelembaban Udara	98,75	1,23	12
Rata-rata Tekanan Udara	1007,18	2,49	12
Rata-rata Penyinaran Matahari	4,21	1,12	12
Rata-rata Suhu Udara	34,34	0,55	12
Rata-rata Suhu Udara Minimum	22,2	0,67	12
Jumlah Curah Hujan	154,01	72,56	12

Berdasarkan analisis statistika deskriptif, data yang digunakan berjumlah tujuh variabel prediktor dan satu variabel respon. Dari data yang dihasilkan diketahui bahwa nilai rata-rata suhu udara memiliki nilai simpangan baku yang lebih signifikan di antara variabel-variabel prediktor yang diamati lainnya dengan nilai sebesar 0,55. Hal ini berbanding terbalik dengan rata-rata hari hujan yang menghasilkan nilai simpangan baku sebesar 4,31 dan rata-rata sebesar 14.

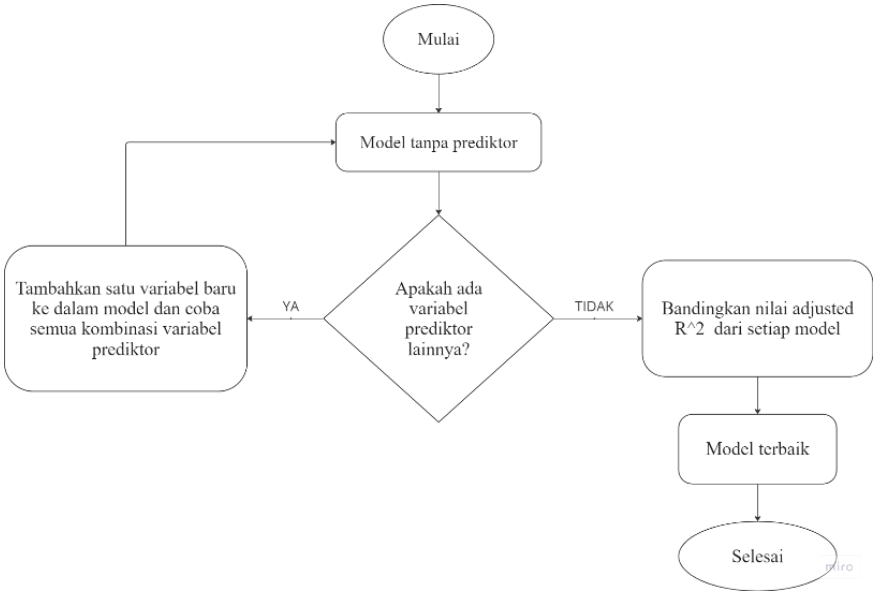
3. Perancangan Sistem Seleksi Variabel

a. Best Subset Selection



Gambar 1. Diagram alir metode best subset selection

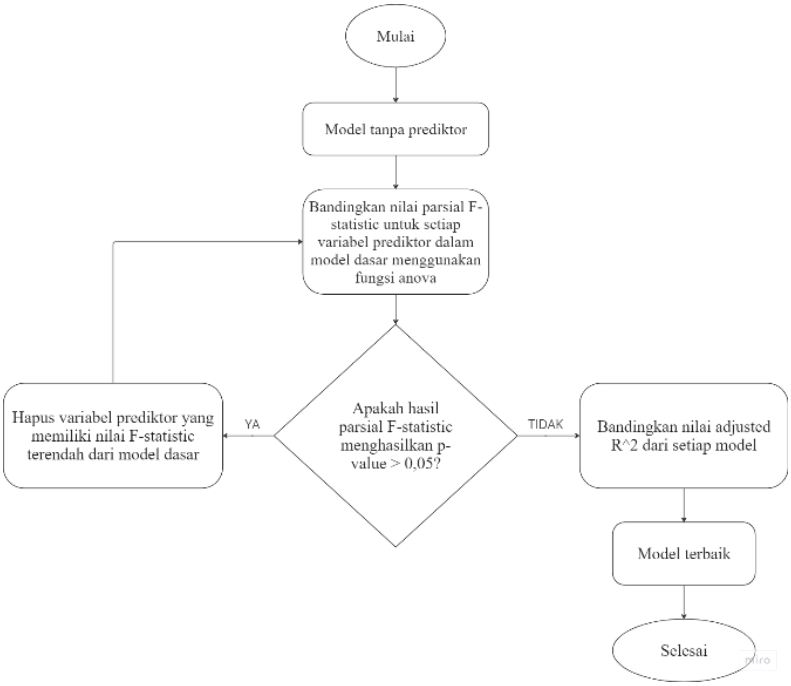
b. Forward Stepwise Selection



Gambar 2. *Diagram alir metode forward selection*

E-ISSN 2808-5841
P-ISSN 2808-7283

c. Backward Stepwise Selection



Gambar 3. *diagram alir metode backward selection*

III. HASIL DAN PEMBAHASAN

Dalam penelitian ini, pengujian dilakukan menggunakan data yang terdiri dari tujuh variabel prediktor, dengan masing-masing variabel memiliki 12 baris data pada rentang tahun 2022. Pengujian dilakukan dengan mempertimbangkan variabel prediktor terbaik yang memberikan kontribusi signifikan terhadap model, menggunakan metode *best subset selection, forward stepwise selection,* dan *backward stepwise selection*. Setiap metode dilakukan perhitungan dengan menggunakan bahasa pemrograman R untuk menentukan nilai *RSS, Adjusted* R^2*, BIC, dan Cp*, yang kemudian direpresentasikan dalam bentuk grafik. Model dengan nilai *Adjusted* R^2 tertinggi atau nilai *RSS, BIC dan Cp* terendah dianggap sebagai model yang memiliki variabel prediktor terbaik. Sebagai perbandingan awal, hasil ringkasan model untuk regresi berganda dengan semua prediktor digunakan tanpa penyeleksian variabel juga disajikan.

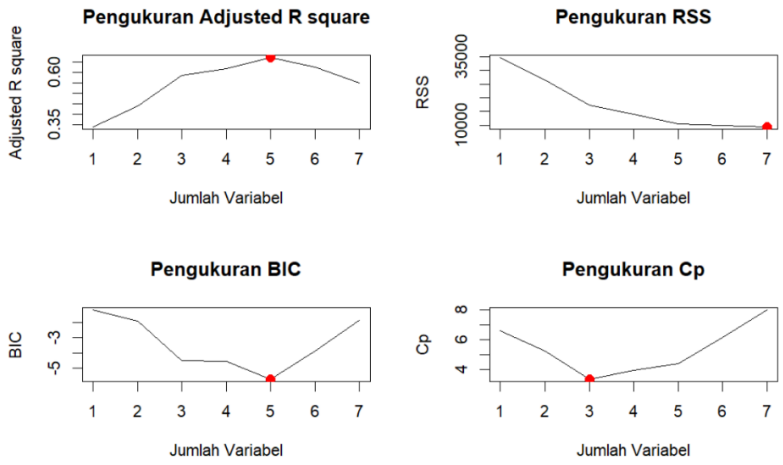
Pertama diperlukan perancangan model regresi linear menggunakan ketujuh variabel prediktor yang akan digunakan sebagai pembanding dengan model yang dihasilkan oleh metode *best subset selection, forward stepwise selection,* dan *backward stepwise selection*. Didapatlah nilai R^2 dan *Adjusted* R^2 diangka 0,8362852 dan 0,5497884.

1. Metode Best Subset Selection

Pengujian pertama menggunakan metode *best subset selection* atau metode seleksi terbaik, didapat tujuah model regresi linear dengan jumlah variabel predikor optimal berjumlah lima, yaitu variabel Jumlah hari hujan, Rata-rata kecepatan angin, Rata-rata kelembaban udara, Rata-rata suhu udara, dan Rata-rata suhu udara minimum. Nilai R^2 yang didapat adalah 0,8207131 dan nilai *Adjusted* R^2 di angka 0,6713073. Nilai *Adjusted* R^2 di angka 0,6713073 artinya variabel prediktor secara signifikan bisa menjelaskan model regresi yang didapat sekitar 67,13%. Dapat dilihat juga nilai R^2 lebih rendah dari nilai R^2 ketika seluruh variabel prediktor dimasukan ke dalam model,

dengan selisih 0,0155721 dan *Adjusted* R^2 yang lebih besar dari pada ketika seluruh variabel dimasukan, yaitu dengan selisih 0,1215229. Hal ini terbilang wajar dan cukup baik dengan pengingkatan nilai *Adjusted* R^2 yang signifikan.

Dalam menggunakan metode *best subset selection* dengan tujuh variabel prediktor, didapat tujuh model dengan nilai *Adjusted* R^2 berturut-turut sebagai sebesar 0.3415069, 0.4389509, 0,5872222, 0,6186270, 0,6713073, 0,6251874, dan 0,5497844. Model kelima, ketika jumlah variabelnya berjumlah lima, memiliki nilai *Adjusted* R^2 tertinggi pada angka 67,13%, sehingga dipilih sebagai model terbaik. Kemudian terjadi penurunan seiring berkurangnya nilai *RSS* dengan nilai terkecil ada ketika variabel prediktor berjumlah tujuh menunjukkan peningkatan kualitas model (lihat Gambar (5)). Selain itu, nilai *BIC* terkecil yang didapat ada pada angka -5,715773 ketika model memiliki lima variabel prediktor, sama seperti nilai *Adjusted* R^2. Demikian juga nilai *Cp* terkecil berada di angka 3,334758, ketika variabel prediktor berjumlah tiga. Grafik dapat dilihat pada Gambar (5) berikut:



Gambar 4. *Grafik Adjusted R2, RSS, BIC, dan CP metode Best subset selection*

Dari hasil pemodelan regresi linear menggunakan metode *backward selection*, didapat model regresi yang dapat ditulis menjadi (13) sebagai berikut:

$$Y = -1802{,}7 + 8{,}05X_1 - 15{,}18X_2 + 23{,}86X_3 - 38{,}51X_6 + 47{,}68X_7 \qquad (13)$$

keterangan:
Y = Jumlah curah hujan
X_1 = Rata-rata hari hujan
X_2 = Rata-rata kecepatan angin
X_3 = Rata-rata kelembaban udara
X_6 = Rata-rata suhu udara
X_7 = Rata-rata suhu udara minimum

Model regresi linear yang diberikan menunjukkan bahwa beberapa variabel cuaca mempengaruhi jumlah curah hujan (Y). Intersep sebesar -1802,7 menunjukkan nilai dasar curah hujan jika semua variabel lainnya bernilai nol, meskipun ini mungkin tidak realistis dalam konteks nyata. Curah hujan akan meningkat sebesar 8,05 mm^3 setiap terjadi kenaikan satu unit pada rata-rata hari hujan (X_1). Sebaliknya, setiap kenaikan satu unit pada rata-rata kecepatan angin (X_2) akan mengurangi curah hujan sebesar 15,18 mm^3. Setiap kenaikan satu unit pada rata-rata kelembaban

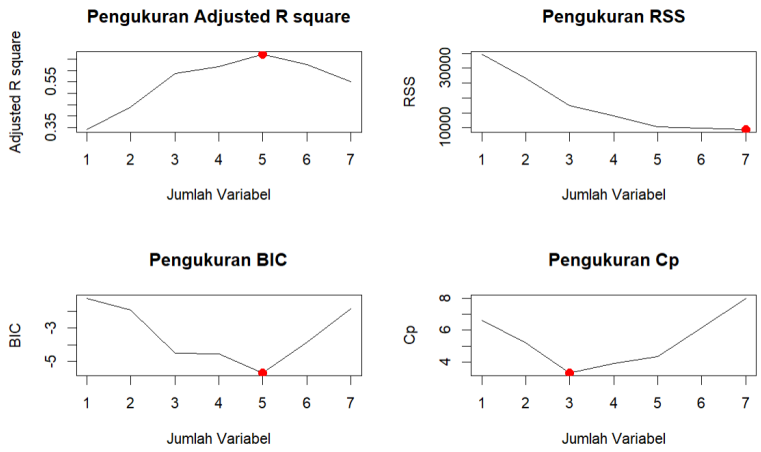
udara (X_3) akan meningkatkan curah hujan sebesar 23,86 mm^3. Namun, setiap kenaikan satu unit pada rata-rata suhu udara (X_6) akan mengurangi curah hujan sebesar 38,51 mm^3. Di sisi lain, setiap kenaikan satu unit pada rata-rata suhu udara minimum (X_7) akan meningkatkan curah hujan sebesar 47,68 mm^3. Secara keseluruhan, rata-rata hari hujan, kelembaban udara, dan suhu udara minimum memiliki dampak positif terhadap curah hujan, sementara kecepatan angin dan suhu udara memiliki dampak negatif.

Berdasarkan interpretasi model optimum menggunakan metode *best subset selection* maka diperoleh faktor yang mempengaruhi jumlah curah hujan di Provinsi Lampung periode 2022 secara signifikan adalah rata-rata suhu udara minimum dan rata-rata suhu udara.

2. *Metode Forward Stepwise Selection*

Metode berikutnya yang digunakan adalah *foward stepwise selection*, yang kemudian didapatkan lima variabel prediktor terbaik, yaitu Jumlah hari hujan, Rata-rata kecepatan angin, Rata-rata kelembaban udara, Rata-rata suhu udara, dan Rata-rata suhu udara minimum. Model yang didapat ketika menggunakan metode *foward stepwise selection* serupa dengan ketika menggunakan metode *Best Subset Selection* dengan nilai R^2 yang didapat adalah 0,8207131 dan nilai *Adjusted* R^2 di angka 0,6713073. Nilai *Adjusted* R^2 di angka 0,6713073. Hasil menunjukkan bahwa terdapat konsistensi dalam nilai R^2 dan *Adjusted* R^2 ketika melakukan seleksi variabel terbaik menggunakan metode *forward stepwise selection* dan metode *best subset selection*. Hal ini terjadi karena penggunaan prediktor yang serupa dalam kedua metode tersebut. Dari hasil pemodelan regresi linear menggunakan metode *best subset selection*, didapat nilai R^2 lebih rendah dari nilai R^2 ketika seluruh variabel prediktor dimasukan ke dalam model, dengan selisih 0,0155721 dan *adjusted* R^2 yang lebih besar dari pada ketika seluruh variabel dimasukan, yaitu dengan selisih 0,1215229, nilai serupa yang didapat ketika menggunakan metode *best subset selection*.

Selain itu, analisis grafik pada Gambar 6 mengungkapkan hasil serupa dalam pemilihan subset terbaik, dengan nilai-nilai *RSS, Adjusted R Square, BIC, dan Cp* yang sama. Fenomena ini cenderung terjadi ketika hanya sedikit variabel prediktor yang digunakan, sehingga variabel tersebut sering kali terpilih secara berulang sebagai variabel terbaik. Namun, jika jumlah variabel prediktor yang digunakan lebih banyak, kemungkinan terjadinya variasi dalam variabel terbaik antar metode akan meningkat. Grafik dapat dilihat pada Gambar (6) berikut:



Gambar 1. Grafik Adjusted R^2, RSS, BIC, dan CP metode Forward Stepwise Selection

Dari hasil pemodelan regresi linear menggunakan metode *backward selection*, didapat model regresi yang dapat ditulis menjadi (14) sebagai berikut:

$$Y = -1802{,}7 + 8{,}05X_1 - 15{,}18X_2 + 23{,}86X_3 - 38{,}51X_6 + 47{,}68X_7 \quad (14)$$

keterangan:

Y = Jumlah curah hujan
X_1 = Rata-rata hari hujan
X_2 = Rata-rata kecepatan angin
X_3 = Rata-rata kelembaban udara
X_6 = Rata-rata suhu udara
X_7 = Rata-rata suhu udara minimum

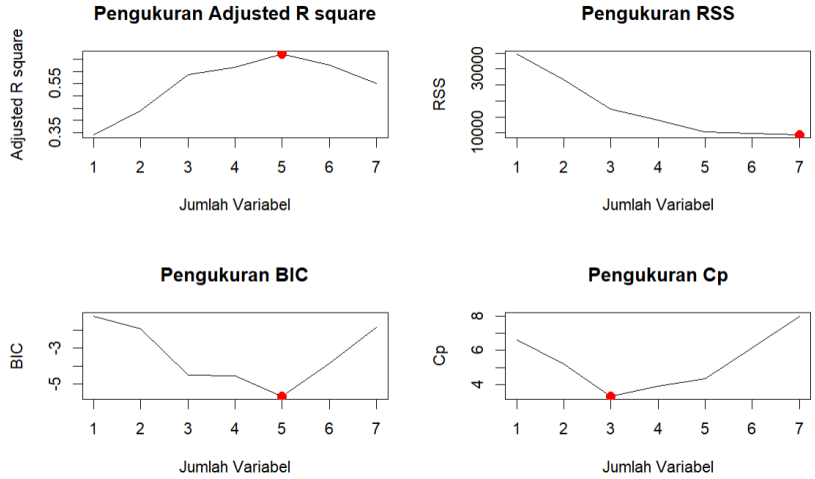
Model regresi linear yang diberikan menunjukkan bahwa beberapa variabel cuaca mempengaruhi jumlah curah hujan (Y). Intersep sebesar -1802,7 menunjukkan jumlah curah hujan jika semua variabel lainnya bernilai nol, meskipun ini mungkin tidak realistis dalam konteks nyata. Curah hujan akan meningkat sebesar 8,05 mm^3 setiap terjadi kenaikan satu unit pada rata-rata hari hujan (X_1).Setiap peningkatan satu unit pada rata-rata kecepatan angin (X_2) akan mengurangi curah hujan sebesar 15,18 mm^3. Setiap kenaikan satu unit pada rata-rata kelembaban udara (X_3) akan meningkatkan curah hujan sebesar 23,86 mm^3. Kenaikan satu unit pada rata-rata suhu udara (X_6) akan mengurangi curah hujan sebesar 38,51 mm^3. Di sisi lain, setiap kenaikan satu unit pada rata-rata suhu udara minimum (X_7) akan meningkatkan curah hujan sebesar 47,68 mm^3.

Berdasarkan interpretasi model optimum menggunakan metode *foward stepwise selection* maka diperoleh faktor yang mempengaruhi jumlah curah hujan di Provinsi Lampung periode 2022 secara signifikan adalah rata-rata suhu udara minimum dan rata-rata suhu udara.

3. *Metode Backward Stepwise Selection*

Metode ketiga yang digunakan adalah *backward stepwise selection*, sehingga nilai yang serupa dengan hasil dari metode *best subset selection* dan *foward stepwise selection*. Didapati tujuh model dengan nilai *Adjusted* R^2 berturut-turut sebesar 0,3415069, 0.4389509, 0,5872222, 0,6186270, 0,6713073, 0,6251874, dan 0,5497844. Model kelima, ketika jumlah variabelnya berjumlah lima, memiliki nilai *Adjusted* R^2 tertinggi pada angka 67,13%. Sedangkan untuk nilai-nilai *RSS, Adjusted R Square, BIC, dan Cp* juga didapati hasil yang sama, yaitu nilai *RSS* secara konsisten turun ketika bertambahnya variabel, nilai *BIC* terkecil yang didapat ada pada angka -5,715773 ketika model memiliki lima variabel prediktor, sama seperti nilai *Adjusted* R^2. Demikian juga nilai *Cp* terkecil berada di angka 3,334758, ketika variabel prediktor berjumlah tiga. Grafik dapat dilihat pada Gambar (7) berikut:

E-ISSN 2808-5841
P-ISSN 2808-7283



Gambar 2. Grafik Adjusted R^2, RSS, BIC, dan CP metode backward stepwise selection

Dari hasil pemodelan regresi linear menggunakan metode *backward selection*, didapat model regresi yang dapat ditulis menjadi (15) sebagai berikut:

$$Y = -1802{,}7 + 8{,}05X_1 - 15{,}18X_2 + 23{,}86X_3 - 38{,}51X_6 + 47{,}68X_7 \quad (15)$$

keterangan:

Y = Jumlah curah hujan
X_1 = Rata-rata hari hujan
X_2 = Rata-rata kecepatan angin
X_3 = Rata-rata kelembaban udara
X_6 = Rata-rata suhu udara
X_7 = Rata-rata suhu udara minimum

Model regresi linear yang diberikan menunjukkan bahwa beberapa variabel cuaca mempengaruhi jumlah curah hujan (Y). Intersep sebesar -1802,7 menunjukkan nilai dasar curah hujan jika semua variabel lainnya bernilai nol, meskipun ini mungkin tidak realistis dalam konteks nyata. Curah hujan akan meningkat sebesar 8,05 mm^3 setiap terjadi kenaikan satu unit pada rata-rata hari hujan (X_1). Sebaliknya, setiap kenaikan satu unit pada rata-rata kecepatan angin (X_2) akan mengurangi curah hujan sebesar 15,18 mm^3. Setiap kenaikan satu unit pada rata-rata kelembaban udara (X_3) akan meningkatkan curah hujan sebesar 23,86 mm^3. Namun, setiap kenaikan satu unit pada rata-rata suhu udara (X_6) akan mengurangi curah hujan sebesar 38,51 mm^3. Di sisi lain, setiap kenaikan satu unit pada rata-rata suhu udara minimum (X_7) akan meningkatkan curah hujan sebesar 47,68 mm^3.

Berdasarkan interpretasi model optimum menggunakan metode *backward stepwise selection* maka diperoleh faktor yang mempengaruhi jumlah curah hujan di Provinsi Lampung periode 2022 secara signifikan adalah rata-rata suhu udara minimum dan rata-rata suhu udara.

4. Pemilihan Model Optimal

Setelah dilakukan pencarian model paling optimal menggunakan ketiga metode *Best Subset, Foward Stepwise,* dan *Backward Stepwise*, didapat model regresi paling optimal sebagai berikut:

Tabel 4. Perbandingan model

Metode	Model Regresi	*Adj* R^2	*RSS*	*BIC*	*Cp*
Best Subset	$Y = -1802{,}7 + 8{,}05X_1 - 15{,}18X_2 + 23{,}86X_3 - 38{,}51X_6 + 47{,}68X_7$	0,671	10383.326	-5,71	4,380471
Forward	$Y = -1802{,}7 + 8{,}05X_1 - 15{,}18X_2 + 23{,}86X_3 - 38{,}51X_6 + 47{,}68X_7$	0,671	10383.326	-5,71	4,380471
Backward	$Y = -1802{,}7 + 8{,}05X_1 - 15{,}18X_2 + 23{,}86X_3 - 38{,}51X_6 + 47{,}68X_7$	0,671	10383.326	-5,71	4,380471

Diperoleh Tabel 4 model di atas menggunakan metode *Best Subset, Foward Stepwise, dan Backward Stepwise*. Terlihat bahwa dengan ketiga metode yang digunakan mendapatkan model regresi linear berganda yang sama dengan lima variabel prediktor, yaitu variabel Jumlah hari hujan, Rata-rata kecepatan angin, Rata-rata kelembaban udara, Rata-rata suhu udara, dan Rata-rata suhu udara minimum. Dengan model regresi yang dapat dituliskan menjadi (16) sebagai berikut:

$$Y = -1802{,}7 + 8{,}05X_1 - 15{,}18X_2 + 23{,}86X_3 - 38{,}51X_6 + 47{,}68X_7 \quad (16)$$

keterangan:
Y = Jumlah curah hujan
X_1 = Rata-rata hari hujan
X_2 = Rata-rata kecepatan angin
X_3 = Rata-rata kelembaban udara
X_6 = Rata-rata suhu udara
X_7 = Rata-rata suhu udara minimum

Nilai *Adjusted* R^2 sama, yaitu berada di angka 0,671 atau 67,1% dan begitu juga dengan nilai *RSS* dan *BIC* yang masing-masing model dari ketiga metode yang digunakan mendapat nilai *RSS* di angka 10383,326 dan nilai *BIC* di angka -5,715773. Sedangkan nilai Cp yang didapat berada pada angka 4,380471.

VI. KESIMPULAN

Analisis regresi linear dengan lima variabel independen terpilih sebagai model optimal berdasarkan metode *Best Subset, Forward Stepwise,* dan *Backward Stepwise*. Model ini menghasilkan nilai *adjusted R square* sebesar 67,1%, dimana angka tersebut cukup tinggi dan layak untuk memvalidasi model dan menunjukkan bahwa variabel terpilih menjelaskan sebagian besar variansi variabel dependen. Juga dapat dilihat berdasarkan nilai *Bayesian Information Criterion (BIC)* yaitu senilai -5,7109908 yang merupakan nilai terkecil dibandingkan dengan model lain, begitu pula dengan didapatnya nilai *Cp* senilai 4,380471 yang merupakan nilai terkecil dibanding model lainnya diikuti nilai *RSS* yang didapat yaitu senilai 10383,326 dimana angka tersebut mungkin bukan nilai *RSS* yang tertinggi jika dibandingkan dengan model bervariabel enam atau tujuh tetapi nilai tersebut sudah menunjukkan bahwa model dengan lima variabel cukup optimal. Konsistensi ketiga metode dalam memilih variabel yang sama memperkuat validitas dan reliabilitas model. Hal ini menunjukkan bahwa model regresi linear menggunakan lima variabel independen ini dapat digunakan untuk memprediksi nilai variabel dependen dengan cukup akurat.

UCAPAN TERIMA KASIH

Kami mengucapkan terima kasih yang mendalam kepada Tim SEDANA atas kesempatan dan dukungan yang telah diberikan dalam penyelenggaraan acara ini. Kami juga ingin menyampaikan penghargaan dan rasa terima kasih kepada para dosen pembimbing yang telah memberikan bimbingan, saran, dan masukan yang sangat berharga selama proses penyusunan artikel ini. Ucapan terima kasih juga kami tujukan kepada seluruh anggota tim yang telah bekerja keras dan berkontribusi dalam penelitian ini. Terima kasih kepada semua pihak yang telah terlibat dan membantu, baik secara langsung maupun tidak langsung, dalam menyukseskan penelitian ini. Semoga artikel ini dapat memberikan manfaat dan kontribusi positif bagi perkembangan ilmu pengetahuan.

REFERENSI

1. T. S. Swarionoto and S. , "Pemanfaatan Suhu Udara dan Kelembapan Udara dalam Persamaan Regresi untuk Simulasi Prediksi Total Hujan Bulanan di Bandar Lampung," *Meteorologi dan Geofisika,* vol. 12, pp. 271-281, 2011.
2. P. Bruce, A. Bruce and P. Gedeck, Practical Statistics for Data Scientiests: 50+ Essential Concepts Using R and Python, United States of America: O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, 2020.
3. P. U. Gio and E. Rosmaini, Belajar Olah Data dengan SPSS, Minitab, R, Microsoft Excel, Eviews, Lisrel, Amos, dan Smartpls (disertai beberapa contoh perhitungan manual), Medan: USU press, 2016.
4. R. Prasetyo and Helma, "Analisis Regresi Linear Berganda Untuk Melihat Faktor Yang Berpengaruh Terhadap Kemiskinan di Provinsi Sumatera Barat," *Journal of Mathematics UNP,* pp. 62-68, 2022.
5. P. Subekti, "Perbandingan Metode Best Subset Dan Stepwise Untuk," *Jurnal Ilmiah Teknologi dan Informasia ASIA (JITIKA),* pp. 6-14, 2015.
6. P. Subekti and M. Islamiyah, "Penentuan Model Hubungan Kepadatan Penduduk dan Faktornya Menggunakan Metode Forward Selection," *Jurnal Matematika dan Pendidikan Matematika Vol. 2 No. 1 Maret 2017,* pp. 48-57, 2017.
7. A. Yanke, N. E. Zendrato and A. M. Soleh, "Penanganan Masalah Multikolinieritas pada Pemodelan Pertumbuhan," *Indonesian Journal of Statistics and Its Applications,* pp. 228-244, 2021.
8. A. I. Rahutami, "Konsep Ekonometri dan Regresi berganda," pp. 1-20, 2011.
9. J. Roy, "Ekonometrika (Pemodelan dan Analisis Regresi)," Samarinda.
10. I. Ghozali, Aplikasi analisis multivariate dengan program IBM SPSS 25 edisi ke-9, Semarang: Universitas Diponegoro, 2018.
11. H. Hanum, "Perbandingan Metode Stepwise, Best Subset Regression, dan," *Jurnal Penelitian Sains,* vol. 14, pp. 14201-14201-6, 2011.
12. N. Kustinah, "Pemilihan Model Regresi Terbaik dengan Bayesian Information Criterion (Bic)," *Institutional Repository,* 2011.
13. D. P. Lampung, "Dashboard Lampung," Diskominfotik Provinsi Lampung, [Online]. Available: https://dashboard.lampungprov.go.id/dashboard?topik=n8K6qdmGPw. [Accessed 15 Februari 2024].