

BAB II

LANDASAN TEORI

2.1. Model Linear

Model linear merupakan salah satu alat analisis yang paling fundamental dalam statistik dan pembelajaran mesin. Model ini telah digunakan secara luas dalam berbagai bidang penelitian selama lebih dari dua abad karena kemampuannya dalam menjelaskan hubungan antar variabel dengan cara yang sederhana namun kuat.

Model linear mengasumsikan adanya hubungan yang berbentuk garis lurus antara variabel prediktor (variabel yang digunakan untuk memprediksi) dengan variabel respons (variabel yang ingin diprediksi). Dengan kata lain, model linear mencoba menjelaskan bagaimana perubahan pada variabel prediktor akan mempengaruhi variabel respons secara proporsional. Secara matematis, model linear dapat ditulis sebagai (Hastie et al., 2009):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \quad (2.1)$$

Keterangan:

β_0 = Intercept

$\beta_1, \beta_2, \dots, \beta_p$ = Koefisien regresi untuk masing-masing variabel prediktor

ε = Error term yang diasumsikan berdistribusi normal dengan mean nol dan varians konstan σ^2

Dalam notasi matriks, model ini dapat ditulis sebagai $y = X\beta + \varepsilon$.

Agar model linear dapat bekerja dengan baik dan menghasilkan estimasi yang valid, terdapat beberapa asumsi yang perlu dipenuhi. Asumsi-asumsi tersebut meliputi hubungan antara variabel prediktor dan respons harus berbentuk linear, setiap observasi harus independen satu sama lain, varians dari error harus konstan di semua level prediktor (homoskedastisitas), residual harus mengikuti distribusi normal, dan tidak boleh ada korelasi sempurna antar variabel prediktor (tidak ada multikolinearitas sempurna) (Greene, 2020; James et al., 2021).

Model linear umum dapat diekspresikan sebagai (McCullagh & Nelder, 1989):

$$E(Y|X) = X\beta \quad (2.2)$$

Keterangan:

$E(Y|X)$ = Nilai ekspektasi respons

X = Matriks desain

β = Vektor parameter regresi

2.2. Multikolinearitas

Multikolinearitas adalah kondisi yang terjadi ketika dua atau lebih variabel prediktor dalam model regresi memiliki korelasi yang sangat tinggi satu sama lain. Dengan kata lain, variabel-variabel prediktor tersebut saling terkait erat sehingga memberikan informasi yang hampir sama atau redundant.

Keberadaan multikolinearitas menimbulkan masalah serius dalam analisis regresi. Multikolinearitas menyebabkan estimasi koefisien regresi menjadi sangat tidak stabil, sehingga perubahan kecil pada data dapat mengubah nilai koefisien secara drastis yang mengakibatkan interpretasi mengenai pengaruh masing-masing

variabel prediktor terhadap variabel respons menjadi tidak dapat diandalkan. Menurut Greene (2020) dan Gujarati & Porter (2009), multikolinearitas juga dapat membuat koefisien yang secara teoretis penting menjadi tidak signifikan secara statistik, meskipun model secara keseluruhan menunjukkan *goodness of fit* yang baik.

Untuk mendeteksi ada tidaknya multikolinearitas, peneliti dapat menggunakan beberapa metode seperti analisis korelasi berpasangan, indeks kondisi, aturan Klein, dan *Redefined Variance Inflation Factor* (RVIF). Metode yang paling umum adalah *Variance Inflation Factor* (VIF), yang mengukur seberapa besar varians dari koefisien regresi meningkat akibat adanya korelasi dengan prediktor lain. Berikut adalah rumus dari *Variance Inflation Factor* (VIF):

$$VIF = \frac{1}{1 - R_j^2} \quad (2.3)$$

Keterangan:

R_j^2 = koefisien determinasi dari regresi variabel prediktor ke- j terhadap variabel prediktor lainnya

Sebagai aturan praktis, nilai VIF di atas 10 umumnya digunakan sebagai indikator bahwa multikolinearitas sudah cukup tinggi dan perlu ditangani (Montgomery et al., 2012). Selain VIF, pemeriksaan korelasi Pearson antar variabel prediktor juga dapat memberikan indikasi awal adanya multikolinearitas, terutama jika terdapat pasangan variabel dengan korelasi lebih dari 0,8. Pendekatan lain adalah memeriksa *condition number* dari matriks desain, di mana nilai lebih dari 30 menunjukkan adanya potensi multikolinearitas yang serius (Belsley et al., 1980).

2.3. Regresi Logistik

Regresi logistik adalah metode statistik yang digunakan untuk memodelkan hubungan antara variabel respons yang bersifat biner (dikotomis) dengan satu atau

lebih variabel prediktor. Misalkan Y adalah variabel respons biner dengan nilai 0 dan 1, serta $\mathbf{X} = (X_1, X_2, \dots, X_p)$ adalah vektor variabel prediktor. Tujuan utama regresi logistik adalah memodelkan probabilitas $p(\mathbf{X}) = P(Y = 1|\mathbf{X})$ sebagai fungsi dari $\mathbf{X}$.

2.3.1. Model Regresi Logistik

Model regresi logistik menggunakan transformasi logit untuk menghubungkan probabilitas kejadian dengan variabel prediktor secara linear:

$$\ell\left(\frac{p(\mathbf{X})}{1-p(\mathbf{X})}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p \quad (2.4)$$

Keterangan:

$\ell\left(\frac{p}{1-p}\right)$ = Logit (log-odds)

$p(\mathbf{X})$ = Probabilitas kejadian $Y = 1$

β_j = Koefisien regresi untuk variabel prediktor ke- j

Dengan menyelesaikan Persamaan (2.4) untuk $p(\mathbf{X})$, diperoleh fungsi logistik:

$$p(\mathbf{X}) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p))} \quad (2.5)$$

Keterangan:

Fungsi logistik menghasilkan nilai probabilitas yang selalu berada pada interval (0,1)

Fungsi logistik ini memiliki bentuk kurva sigmoid (*S-shaped curve*) yang memastikan bahwa nilai probabilitas yang diprediksi selalu berada dalam rentang yang valid antara 0 dan 1, terlepas dari nilai *linear predictor*.

2.3.2. Estimasi Parameter dengan Maximum Likelihood Estimation (MLE)

Parameter $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ dalam regresi logistik tidak dapat diestimasi menggunakan metode *Ordinary Least Squares* (OLS) seperti pada regresi linear. Sebagai gantinya, digunakan metode *Maximum Likelihood Estimation* (MLE).

Misalkan diberikan sampel independen $(Y_i, \mathbf{X}_i)$ untuk $i = 1, 2, \dots, n$. Fungsi probabilitas untuk setiap observasi adalah:

$$P(Y_i = y_i | \mathbf{X}_i) = p_i^{y_i} (1 - p_i)^{1-y_i}, \quad y_i \in \{0, 1\} \quad (2.6)$$

Keterangan:

p_i merupakan probabilitas $Y_i = 1$ yang didefinisikan pada persamaan (2.5).

Fungsi *likelihood* untuk seluruh sampel adalah perkalian dari fungsi probabilitas masing-masing observasi:

$$L(\beta) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i} \quad (2.7)$$

Untuk mempermudah perhitungan, digunakan fungsi *log-likelihood*:

$$\ell(\beta) = \sum_{i=1}^n [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (2.8)$$

Pada Persamaan (2.8), y_i dan p_i merupakan besaran skalar untuk setiap observasi ke- i , dengan $y_i \in \{0, 1\}$ dan $0 < p_i < 1$. Operasi $\log(p_i)$ dan $\log(1 - p_i)$ dilakukan pada masing-masing nilai p_i . Jika $y_i = 1$, maka suku yang berkontribusi terhadap fungsi *log-likelihood* adalah $\log(p_i)$, sedangkan jika $y_i = 0$, maka suku yang berkontribusi adalah $\log(1 - p_i)$. Selanjutnya, seluruh nilai tersebut dijumlahkan untuk $i = 1, 2, \dots, n$ sehingga menghasilkan nilai skalar $\ell(\beta)$. Sementara itu, $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ merupakan vektor parameter yang menjadi argumen fungsi *log-likelihood*.

Estimator MLE $\hat{\beta}_{MLE}$ diperoleh dengan memaksimalkan fungsi *log-likelihood*:

$$\hat{\beta}_{MLE} = \arg \max_{\beta} \ell(\beta) \quad (2.9)$$

2.3.3. Algoritma *Iteratively Reweighted Least Squares* (IRLS)

Algoritma *Iteratively Reweighted Least Squares* (IRLS) merupakan metode numerik iteratif yang digunakan untuk mengestimasi parameter pada regresi logistik melalui proses maksimisasi fungsi *log-likelihood* (Nelder & Wedderburn, 1972). Dalam kerangka *Generalized Linear Model* (GLM), regresi logistik memodelkan hubungan antara nilai harapan respons dan prediktor linear melalui fungsi penghubung logit, sehingga estimasi parameter tidak dapat diperoleh secara analitik dan memerlukan pendekatan numerik (McCullagh & Nelder, 1989).

Sebagaimana telah dijelaskan pada subbab sebelumnya, persamaan yang dihasilkan dari proses maksimisasi tersebut tidak memiliki solusi analitik dalam bentuk tertutup, sehingga diperlukan metode iteratif untuk memperoleh estimasi parameter.

Secara umum, IRLS bekerja dengan mengubah permasalahan optimasi *log-likelihood* yang bersifat nonlinear menjadi serangkaian permasalahan *Weighted Least Squares* (WLS) melalui pendekatan linearisasi lokal. Pada setiap iterasi, model diperbarui berdasarkan estimasi parameter saat ini, kemudian digunakan untuk membentuk masalah regresi linear berbobot yang ekuivalen secara lokal. Proses ini diulang hingga diperoleh konvergensi (McCullagh & Nelder, 1989).

Algoritma IRLS untuk regresi logistik terdiri dari langkah-langkah berikut.

1. Inisialisasi

Sebelum iterasi dimulai, tetapkan nilai awal seluruh koefisien sebagai vektor

nol dan atur penghitung iterasi $t = 0$:

$$\boldsymbol{\beta}^{(0)} = \left(\beta_0^{(0)}, \beta_1^{(0)}, \dots, \beta_p^{(0)} \right)^T = \mathbf{0}, \quad t = 0 \quad (2.10)$$

Keterangan:

$\boldsymbol{\beta}^{(0)}$ = Nilai awal koefisien, ditetapkan sebagai vektor nol

t = Indeks iterasi, dimulai dari $t = 0$

Penetapan $\boldsymbol{\beta}^{(0)} = \mathbf{0}$ merupakan konvensi yang paling umum digunakan. Nilai awal ini berarti pada iterasi pertama diasumsikan tidak ada variabel prediktor yang berpengaruh terhadap variabel respons, sehingga probabilitas prediksi untuk setiap observasi dimulai dari nilai netral $p_i^{(0)} = 0,5$. Meskipun nilai awal lain juga dapat digunakan, titik awal yang netral ini umumnya menghasilkan konvergensi yang baik pada model regresi logistik yang teridentifikasi dengan baik (Agresti, 2013).

2. Menghitung *Linear Predictor*

Pada setiap iterasi ke- t , langkah pertama adalah menghitung nilai *linear predictor* $\eta_i^{(t)}$ untuk setiap observasi $i = 1, 2, \dots, n$. *Linear predictor* merupakan kombinasi linear dari nilai-nilai variabel prediktor menggunakan koefisien yang dimiliki model pada iterasi saat ini:

$$\eta_i^{(t)} = \beta_0^{(t)} + \beta_1^{(t)} x_{i1} + \beta_2^{(t)} x_{i2} + \dots + \beta_p^{(t)} x_{ip} = \mathbf{x}_i^T \boldsymbol{\beta}^{(t)} \quad (2.11)$$

Keterangan:

$\eta_i^{(t)}$ = Nilai *linear predictor* untuk observasi ke- i pada iterasi ke- t , dapat bernilai sembarang bilangan real dari $-\infty$ hingga $+\infty$

$\mathbf{x}_i = (1, x_{i1}, x_{i2}, \dots, x_{ip})^T =$ Vektor variabel prediktor untuk observasi ke- i

$\beta^{(t)}$ = Vektor koefisien yang dimiliki model pada iterasi ke- t

Nilai $\eta_i^{(t)}$ tidak terbatas dan dapat bernilai positif maupun negatif, sehingga belum dapat langsung digunakan sebagai prediksi probabilitas karena probabilitas harus berada dalam interval $(0, 1)$. Oleh karena itu, diperlukan transformasi lebih lanjut pada langkah berikutnya.

3. Menghitung Probabilitas Prediksi

Nilai *linear predictor* $\eta_i^{(t)}$ kemudian ditransformasikan menjadi nilai probabilitas menggunakan fungsi logistik:

$$p_i^{(t)} = \frac{1}{1 + \exp(-\eta_i^{(t)})}, \quad p_i^{(t)} \in (0, 1) \quad (2.12)$$

Keterangan:

$p_i^{(t)}$ = Probabilitas prediksi bahwa $Y_i = 1$ pada iterasi ke- t

Fungsi logistik memastikan nilai yang dihasilkan selalu berada dalam interval $(0, 1)$, sehingga konsisten dengan sifat probabilitas. Secara geometris, $p_i^{(t)}$ merepresentasikan posisi observasi ke- i pada kurva sigmoid yang ditentukan oleh koefisien $\beta^{(t)}$ saat ini. Perilaku fungsi ini dapat dipahami sebagai berikut:

- Jika $\eta_i^{(t)}$ sangat besar positif, maka $p_i^{(t)}$ mendekati 1, artinya model sangat yakin bahwa $Y_i = 1$.
- Jika $\eta_i^{(t)}$ sangat besar negatif, maka $p_i^{(t)}$ mendekati 0, artinya model sangat yakin bahwa $Y_i = 0$.
- Jika $\eta_i^{(t)} = 0$, maka $p_i^{(t)} = 0,5$, artinya model tidak dapat membedakan

kedua kelas.

4. Menghitung Matriks Bobot

Berdasarkan probabilitas yang telah dihitung, dibentuk matriks bobot diagonal $\mathbf{W}^{(t)} \in \mathbb{R}^{n \times n}$ dengan elemen diagonal ke- i :

$$w_i^{(t)} = p_i^{(t)} (1 - p_i^{(t)}) \quad (2.13)$$

Keterangan:

$w_i^{(t)}$ = Bobot untuk observasi ke- i pada iterasi ke- t , selalu bernilai positif dalam interval $(0, \frac{1}{4}]$

$\mathbf{W}^{(t)} = \text{diag}(w_1^{(t)}, w_2^{(t)}, \dots, w_n^{(t)})$ = Matriks bobot diagonal berukuran $n \times n$

Bobot $w_i^{(t)}$ sesungguhnya merupakan varians dari distribusi Bernoulli untuk observasi ke- i , yaitu $\text{Var}(Y_i) = p_i(1 - p_i)$. Nilai bobot ini memiliki dua sifat penting yang membuat IRLS bekerja secara adaptif:

- (a) Bobot mencapai nilai maksimum $\frac{1}{4}$ ketika $p_i^{(t)} = 0,5$. Artinya, observasi yang paling "tidak pasti" atau yang belum dapat diklasifikasikan dengan yakin mendapat bobot terbesar dan memberikan kontribusi paling besar terhadap pembaruan koefisien. Dengan demikian, algoritma secara otomatis memberi perhatian lebih pada observasi yang masih sulit diklasifikasikan.
- (b) Bobot mendekati nol ketika $p_i^{(t)}$ mendekati 0 atau 1. Artinya, observasi yang sudah diprediksi dengan sangat yakin hampir tidak memengaruhi pembaruan koefisien, karena model sudah benar dalam mengklasifikasikannya (McCullagh & Nelder, 1989).

Mekanisme pembobotan adaptif inilah yang menjadi keunggulan utama IRLS: algoritma secara cerdas memfokuskan perhatiannya pada observasi yang paling memerlukan perbaikan pada setiap iterasi.

5. Menghitung *Working Response*

Selanjutnya, hitung variabel kerja (*working response*) $z_i^{(t)}$ untuk setiap observasi:

$$z_i^{(t)} = \eta_i^{(t)} + \frac{y_i - p_i^{(t)}}{w_i^{(t)}} = \eta_i^{(t)} + \frac{y_i - p_i^{(t)}}{p_i^{(t)}(1 - p_i^{(t)})} \quad (2.14)$$

Keterangan:

$z_i^{(t)}$ = *Working response* untuk observasi ke- i pada iterasi ke- t

y_i = Nilai respons aktual untuk observasi ke- i , bernilai 0 atau 1

$y_i - p_i^{(t)}$ = Residual pada iterasi ke- t , yaitu selisih antara nilai aktual dan probabilitas prediksi saat ini

Secara matematis, *working response* $z_i^{(t)}$ merupakan hasil dari linearisasi pertama (ekspansi Taylor orde pertama) fungsi *logit* di sekitar nilai saat ini $\eta_i^{(t)}$.

Besaran ini dapat dipahami melalui dua komponen penyusunnya:

(a) Komponen $\eta_i^{(t)}$: merepresentasikan posisi prediksi model saat ini pada skala *linear predictor*.

(b) Komponen $\frac{y_i - p_i^{(t)}}{w_i^{(t)}}$: merupakan koreksi *Newton* yang menyesuaikan prediksi berdasarkan seberapa jauh prediksi saat ini menyimpang dari nilai aktual, kemudian diskalakan oleh bobot $w_i^{(t)}$.

Bersama dengan bobot $w_i^{(t)}$, variabel $z_i^{(t)}$ berperan mengubah masalah regresi logistik nonlinear menjadi masalah regresi linear berbobot yang ekuivalen

secara lokal (*local linear approximation*). Inilah inti dari transformasi IRLS: setiap iterasi pada dasarnya “melinearkan” masalah nonlinear di sekitar titik estimasi saat ini (Hastie et al., 2009).

6. Memperbarui Estimasi Koefisien

Dengan $z_i^{(t)}$ dan $w_i^{(t)}$ yang telah dihitung, estimasi koefisien yang baru $\beta^{(t+1)}$ diperoleh dengan menyelesaikan permasalahan *Weighted Least Squares* (WLS):

$$\beta^{(t+1)} = \arg \min_{\beta} \sum_{i=1}^n w_i^{(t)} \left(z_i^{(t)} - \mathbf{x}_i^T \beta \right)^2 \quad (2.15)$$

Permasalahan WLS ini memiliki solusi *closed-form* sebagai berikut:

$$\beta^{(t+1)} = (\mathbf{X}^T \mathbf{W}^{(t)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(t)} \mathbf{z}^{(t)} \quad (2.16)$$

Keterangan:

$\mathbf{X}$ = Matriks desain berukuran $n \times (p + 1)$

$\mathbf{W}^{(t)} = \text{diag}(w_1^{(t)}, w_2^{(t)}, \dots, w_n^{(t)})$ = Matriks bobot diagonal

$\mathbf{z}^{(t)} = (z_1^{(t)}, z_2^{(t)}, \dots, z_n^{(t)})^T$ = Vektor *working response*

$\mathbf{X}^T \mathbf{W}^{(t)} \mathbf{X}$ = Matriks informasi Fisher yang diamati, yaitu matriks yang mengukur seberapa banyak informasi yang dikandung data tentang parameter β

$\mathbf{X}^T \mathbf{W}^{(t)} \mathbf{z}^{(t)}$ = Vektor skor Fisher berbobot

Persamaan (2.16) merupakan rumus standar WLS dan merupakan solusi *closed-form* yang dapat dihitung langsung tanpa iterasi tambahan. Dalam perspektif optimasi, langkah ini setara dengan satu langkah metode Newton-Raphson pada fungsi *log-likelihood* regresi logistik, karena matriks $\mathbf{X}^T \mathbf{W}^{(t)} \mathbf{X}$ merupakan negatif dari Hessian *log-likelihood* (Agresti, 2013).

Syarat agar solusi Persamaan (2.16) ada dan unik adalah matriks desain X harus memiliki *full column rank* dan tidak terjadi *complete separation* dalam data. *Complete separation* terjadi ketika ada hyperplane yang memisahkan sempurna kelas $Y = 0$ dan $Y = 1$, yang menyebabkan estimasi MLE tidak ada (koefisien menuju tak hingga). Kondisi inilah yang sering muncul pada data dengan n kecil dan p besar, dan menjadi salah satu alasan penggunaan regularisasi LASSO pada penelitian ini.

7. Memeriksa Konvergensi

Iterasi dihentikan apabila perubahan relatif koefisien antara dua iterasi berturutan sudah sangat kecil, yaitu ketika kondisi berikut terpenuhi:

$$\max_j \left| \frac{\beta_j^{(t+1)} - \beta_j^{(t)}}{\beta_j^{(t)}} \right| < \varepsilon \quad (2.17)$$

Keterangan:

$\beta_j^{(t)}$ = Estimasi koefisien regresi ke- j pada iterasi ke- t

ε = Nilai toleransi konvergensi yang ditetapkan sebelum iterasi dimulai, biasanya bernilai antara 10^{-6} hingga 10^{-8} .

Jika kondisi Persamaan (2.17) terpenuhi, atau jika jumlah iterasi telah mencapai batas maksimum yang ditetapkan (misalnya $t_{\max} = 100$), maka algoritma dinyatakan konvergen dan menghasilkan estimator:

$$\hat{\beta}_{MLE} = \beta^{(t+1)} \quad (2.18)$$

Jika konvergensi belum tercapai, tetapkan $t \leftarrow t + 1$ dan kembali ke Langkah

2.

Sifat Konvergensi dan Hubungan dengan Newton-Raphson

Konvergensi algoritma IRLS pada regresi logistik terjamin karena fungsi *log-likelihood* binomial bersifat konkaf (*strictly concave*), yang berarti hanya terdapat satu titik maksimum global dan tidak ada maksimum lokal. Ini berbeda dengan banyak masalah optimasi lain yang memiliki banyak titik ekstrem lokal.

Hubungan antara IRLS dan algoritma Newton-Raphson dapat dilihat secara eksplisit dari persamaan pembaruan Newton-Raphson berikut (Agresti, 2013):

$$\beta^{(t+1)} = \beta^{(t)} - [H(\beta^{(t)})]^{-1} \nabla \ell(\beta^{(t)}) \quad (2.19)$$

Keterangan:

$H(\beta) = -\mathbf{X}^T \mathbf{W} \mathbf{X}$ = Matriks Hessian dari *log-likelihood*, yaitu matriks turunan kedua yang menggambarkan kelengkungan fungsi *log-likelihood*. Tanda negatifnya menunjukkan bahwa fungsi tersebut bersifat konkaf

$\nabla \ell(\beta) = \mathbf{X}^T (\mathbf{y} - \mathbf{p})$ = Vektor gradien dari *log-likelihood*, yaitu vektor turunan pertama yang menunjukkan arah peningkatan terbesar fungsi *log-likelihood*

Dengan mensubstitusikan H dan $\nabla \ell$ ke Persamaan (2.19), dapat ditunjukkan bahwa pembaruan Newton-Raphson tersebut ekuivalen secara aljabar dengan Persamaan (2.16), sehingga IRLS sesungguhnya adalah implementasi Newton-Raphson yang diformulasikan ulang sebagai masalah WLS. Karena menggunakan informasi Fisher yang diharapkan sebagai pengganti Hessian yang dievaluasi pada data sampel, IRLS juga sering disebut sebagai metode *Fisher Scoring* (Agresti, 2013).

Sifat konvergensi Newton-Raphson yang terkenal adalah *quadratic convergence* di sekitar titik optimal, artinya jumlah digit presisi yang benar kira-kira berlipat ganda pada setiap iterasi, inilah yang membuat IRLS jauh lebih efisien di-

bandingkan metode gradien orde pertama seperti *gradient descent*. Kompleksitas komputasi setiap iterasi IRLS adalah $\mathcal{O}(np^2 + p^3)$, di mana n adalah jumlah observasi dan p adalah jumlah variabel prediktor, dan dalam praktiknya IRLS pada regresi logistik umumnya konvergen dalam 10 hingga 20 iterasi (McCullagh & Nelder, 1989).

Dalam penelitian ini, algoritma IRLS untuk regresi logistik standar (tanpa regularisasi) diimplementasikan menggunakan fungsi `glm()` pada perangkat lunak R dengan argumen `family = binomial(link = "logit")`. Pada kasus regresi logistik LASSO, IRLS dikombinasikan dengan algoritma *Coordinate Descent* sebagaimana dijelaskan pada subbab berikutnya.

Interpretasi Koefisien

Interpretasi koefisien dalam regresi logistik dilakukan melalui konsep *odds ratio*. *Odds* didefinisikan sebagai rasio antara probabilitas kejadian dan probabilitas tidak terjadinya kejadian:

$$\text{Odds} = \frac{p(\mathbf{X})}{1 - p(\mathbf{X})} \quad (2.20)$$

Dari Persamaan (2.4), diketahui bahwa $\log(\text{Odds}) = \mathbf{X}^T \boldsymbol{\beta}$, sehingga untuk satu unit peningkatan pada variabel prediktor X_j dengan variabel lain dianggap tetap, *odds ratio* didefinisikan sebagai (Hosmer et al., 2013):

$$OR_j = \exp(\beta_j) \quad (2.21)$$

Keterangan:

OR_j = *Odds ratio* untuk variabel prediktor ke- j

β_j = Koefisien regresi logistik untuk variabel ke- j

Interpretasinya adalah: jika $OR_j > 1$ maka peningkatan X_j meningkatkan odds terjadinya $Y = 1$; jika $OR_j = 1$ maka X_j tidak berpengaruh; dan jika $OR_j < 1$ maka peningkatan X_j menurunkan odds terjadinya $Y = 1$.

2.4. LASSO

LASSO (*Least Absolute Shrinkage and Selection Operator*) adalah metode regularisasi yang diperkenalkan oleh Tibshirani (1996) dengan tujuan mengatasi keterbatasan metode regresi konvensional, terutama dalam situasi di mana terdapat banyak variabel prediktor dan potensi multikolinearitas yang tinggi. Keunikan LASSO terletak pada kemampuannya untuk melakukan *shrinkage* (pengecilan) koefisien dan seleksi variabel secara simultan dalam satu proses.

LASSO menggunakan penalti norma- L_1 yang memiliki karakteristik khusus, yaitu metode ini dapat menghasilkan beberapa koefisien yang bernilai tepat nol. Karakteristik ini menjadikan LASSO lebih efisien karena secara otomatis dapat mengidentifikasi variabel mana yang penting dan mana yang tidak, sehingga menghasilkan model yang lebih sederhana dan lebih mudah diinterpretasikan.

2.4.1. Formulasi LASSO

Secara umum, untuk kasus regresi linear, estimator LASSO didefinisikan sebagai berikut:

$$\hat{\beta}_{LASSO} = \arg \min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \beta)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\} \quad (2.22)$$

Keterangan:

λ = Parameter penalti

$\sum |\beta_j|$ = Penalti L1 yang menghasilkan shrinkage dan seleksi variabel

Persamaan di atas terdiri dari dua komponen yaitu komponen pertama adalah fungsi loss (*sum of squared errors*) yang mengukur seberapa baik model fit terhadap data, dan komponen kedua adalah penalti L1 yang mengontrol kompleksitas model. Parameter λ mengatur *trade-off* antara kedua komponen ini. Ketika $\lambda = 0$, tidak ada penalti dan LASSO menjadi identik dengan OLS. Sebaliknya, ketika λ sangat besar, semua koefisien akan dikecilkan menuju nol.

Penalti L1 memiliki karakteristik geometris yang unik. Bentuk *constraint region* dari penalti L1 adalah *diamond-shaped* dengan sudut-sudut tajam pada sumbu koordinat. Ketika solusi optimal dicari, solusi cenderung berada di sudut-sudut ini, yang berarti beberapa koefisien akan bernilai tepat nol. Inilah yang memungkinkan LASSO untuk melakukan seleksi variabel secara otomatis.

2.4.2. Regresi Logistik LASSO

Regresi logistik LASSO merupakan penggabungan antara metode regularisasi LASSO dengan model regresi logistik. Formulasi matematis dari regresi logistik LASSO adalah:

$$\hat{\beta}_{LASSO} = \arg \min_{\beta} \left\{ -\ell(\beta) + \lambda \sum_{j=1}^p |\beta_j| \right\} \quad (2.23)$$

Keterangan:

$\ell(\beta)$ = Log-likelihood regresi logistik pada Persamaan (2.8)

λ = Parameter penalti yang mengontrol tingkat regularisasi ($\lambda \geq 0$)

$\sum_{j=1}^p |\beta_j|$ = Penalti norma-L1

Untuk menyelesaikan masalah optimasi ini, digunakan algoritma iteratif yang mengkombinasikan *Iteratively Reweighted Least Squares* (IRLS) dengan *Coordi-*

nate Descent. Pada setiap iterasi IRLS, masalah regresi logistik ditransformasikan menjadi masalah *weighted least squares*:

$$\mathbf{z}^{(t)} = \mathbf{X}\boldsymbol{\beta}^{(t)} + \mathbf{W}^{(t)-1}(\mathbf{y} - \mathbf{p}^{(t)}) \quad (2.24)$$

Keterangan:

$\mathbf{W}^{(t)}$ = Matriks bobot diagonal dengan elemen $p_i^{(t)}(1 - p_i^{(t)})$

$\mathbf{p}^{(t)}$ = Vektor probabilitas prediksi pada iterasi ke- t

$\mathbf{z}^{(t)}$ = Variabel kerja (*working response*) pada iterasi ke- t

Selanjutnya, masalah *weighted least squares* diselesaikan dengan *coordinate descent* yang mengakomodasi penalti LASSO.

2.4.3. Algoritma IRLS dan *Coordinate Descent* untuk Regresi Logistik LASSO

Estimasi parameter pada regresi logistik LASSO tidak dapat diselesaikan secara langsung karena adanya penalti norma- L_1 yang bersifat tidak diferensiabel di titik nol. Oleh karena itu, digunakan pendekatan iteratif yang mengkombinasikan algoritma *Iteratively Reweighted Least Squares* (IRLS) dengan *Coordinate Descent* (CD) (Friedman et al., 2010; Hastie et al., 2009).

Pada setiap iterasi IRLS ke- t , masalah regresi logistik ditransformasikan menjadi masalah *weighted least squares* dengan *working response* $\mathbf{z}^{(t)}$ dan matriks bobot $\mathbf{W}^{(t)}$ sebagaimana pada Persamaan (2.24). Permasalahan optimasi kemudian menjadi:

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \arg \min_{\boldsymbol{\beta}} \frac{1}{2n} \sum_{i=1}^n w_i^{(t)} \left(z_i^{(t)} - \mathbf{x}_i^T \boldsymbol{\beta} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (2.25)$$

Masalah di atas merupakan bentuk *weighted LASSO* yang dapat diselesaikan

menggunakan algoritma *Coordinate Descent*. Metode ini bekerja dengan memperbarui satu koefisien pada satu waktu, sementara koefisien lainnya dianggap tetap.

Untuk setiap variabel prediktor X_j , dilakukan pembaruan koefisien dengan menghitung residual parsial:

$$r_{ij}^{(t)} = z_i^{(t)} - \sum_{k \neq j} x_{ik} \beta_k \quad (2.26)$$

Selanjutnya, dihitung korelasi berbobot antara prediktor X_j dan residual parsial:

$$c_j = \frac{1}{n} \sum_{i=1}^n w_i^{(t)} x_{ij} r_{ij}^{(t)} \quad (2.27)$$

Koefisien β_j kemudian diperbarui menggunakan operator *soft-thresholding*:

$$\beta_j^{(t+1)} = \frac{S(c_j, \lambda)}{\frac{1}{n} \sum_{i=1}^n w_i^{(t)} x_{ij}^2} \quad (2.28)$$

dengan fungsi *soft-thresholding* didefinisikan sebagai:

$$S(z, \lambda) = \text{sign}(z)(|z| - \lambda)_+ \quad (2.29)$$

Keterangan: $(a)_+ = \max(a, 0)$

Proses pembaruan koefisien dilakukan secara siklik untuk seluruh variabel prediktor hingga konvergensi tercapai, yaitu ketika perubahan nilai koefisien antar iterasi lebih kecil dari nilai toleransi yang telah ditentukan.

Kombinasi IRLS dan *Coordinate Descent* ini memungkinkan penyelesaian masalah regresi logistik LASSO secara efisien, bahkan pada kasus dengan jumlah

variabel prediktor yang relatif banyak (Friedman et al., 2010).

2.5. Akaike Information Criterion (AIC)

Akaike Information Criterion (AIC) merupakan salah satu kriteria pemilihan model yang paling banyak digunakan dalam statistik. Kriteria ini dikembangkan oleh Hirotugu Akaike, seorang statistikawan Jepang, pada tahun 1970-an. AIC didasarkan pada ide fundamental tentang *trade-off* antara *goodness of fit* (seberapa baik model cocok dengan data) dan kompleksitas model (seberapa banyak parameter yang digunakan) (Burnham & Anderson, 2002).

Pada AIC, model yang baik bukan hanya model yang fit dengan data training, tetapi juga model yang tidak terlalu kompleks sehingga dapat generalisasi dengan baik pada data baru. Model yang terlalu kompleks cenderung mengalami *overfitting*, sedangkan model yang terlalu sederhana mungkin tidak mampu menangkap pola penting dalam data (*underfitting*). AIC memberikan cara yang sistematis dan berdasarkan teori untuk menyeimbangkan kedua aspek ini.

2.5.1. Definisi dan Formulasi AIC

$$AIC = -2\ell(\hat{\theta}_{MLE}) + 2k \quad (2.30)$$

Keterangan:

$\ell(\hat{\theta})$ = Log-likelihood maksimum

k = Jumlah parameter dalam model

Komponen pertama, $-2\ell(\hat{\theta}_{MLE})$, mengukur seberapa buruk model fit terhadap data (semakin kecil log-likelihood maksimum, semakin buruk fit-nya). Tanda negatif digunakan sehingga nilai yang lebih kecil menunjukkan model yang lebih

baik. Komponen kedua, $2k$, adalah penalti untuk kompleksitas model, di mana k adalah jumlah parameter dalam model. Setiap parameter tambahan akan meningkatkan nilai AIC sebesar 2.

2.5.2. Interpretasi AIC

Prinsip dasar dalam menggunakan AIC adalah memilih model dengan nilai AIC yang paling kecil. Namun, perbedaan nilai AIC antara model-model kandidat juga perlu diinterpretasikan dengan hati-hati untuk memahami seberapa besar perbedaan dukungan empiris antar model.

Secara umum, pedoman interpretasi yang sering digunakan adalah sebagai berikut. Jika perbedaan AIC antara dua model ($\Delta AIC = AIC_i - AIC_{min}$) kurang dari 2, maka kedua model dapat dianggap kompetitif dan memiliki dukungan empiris yang sebanding. Dalam situasi ini, peneliti dapat mempertimbangkan faktor lain seperti interpretabilitas atau kesederhanaan model dalam memilih model final.

Jika ΔAIC berada dalam rentang 4 hingga 7, model dengan AIC lebih tinggi masih memiliki dukungan empiris dari data meskipun dukungannya lebih lemah dibandingkan model terbaik. Namun, jika ΔAIC lebih dari 10, model dengan AIC lebih tinggi praktis tidak memiliki dukungan empiris dari data dan sebaiknya tidak dipertimbangkan lebih lanjut.

2.5.3. AIC dalam LASSO untuk Regresi Logistik

Dalam konteks regresi logistik LASSO, AIC dapat digunakan untuk memilih parameter penalti λ yang optimal. Untuk setiap nilai λ , AIC dihitung sebagai:

$$AIC(\lambda) = -2\ell(\hat{\beta}(\lambda)) + 2 \cdot df(\lambda) \quad (2.31)$$

Keterangan:

$\ell(\hat{\beta}(\lambda))$ = Log-likelihood maksimum pada nilai λ tertentu

$df(\lambda)$ = Derajat bebas efektif (*effective degrees of freedom*) dari model LASSO

Berdasarkan Zou et al. (2007), derajat bebas efektif LASSO dapat diestimasi sebagai jumlah koefisien yang tidak bernilai nol:

$$df(\lambda) = \#\{j : \hat{\beta}_j(\lambda) \neq 0\} \quad (2.32)$$

Keterangan:

$\#\{j : \hat{\beta}_j(\lambda) \neq 0\}$ = Banyaknya koefisien yang tidak bernilai nol pada nilai λ tertentu.

Perlu dicatat bahwa hasil Zou et al. (2007) secara formal dibuktikan untuk kasus regresi linear. Dalam konteks regresi logistik, pendekatan ini diadopsi secara luas dalam praktik karena jumlah koefisien non-nol terbukti merupakan estimasi yang konsisten untuk derajat bebas efektif secara asimtotik. Implementasi paket *glmnet* (Friedman et al., 2010) menggunakan pendekatan yang sama, sehingga persamaan (2.31) tetap valid digunakan dalam penelitian ini sebagai pendekatan praktis yang didukung literatur.

Selanjutnya, pemilihan nilai λ yang optimal dalam penelitian ini dilakukan menggunakan AIC. Dibandingkan dengan metode pemilihan λ lainnya seperti *cross-validation*, AIC memiliki keunggulan komputasi yang signifikan karena tidak memerlukan pembagian data secara berulang. Hal ini menjadi pertimbangan penting mengingat jumlah observasi yang terbatas ($n = 38$), di mana *cross-validation* dapat menghasilkan estimasi yang tidak stabil akibat ukuran *subset* yang terlalu kecil. Al-khateeb (2019) menunjukkan bahwa AIC memiliki kinerja yang kompetitif dengan *cross-validation* dalam hal akurasi estimasi koefisien dan seleksi variabel,

khususnya pada kasus dengan jumlah observasi yang terbatas, sehingga menjadi pilihan yang tepat untuk konteks penelitian ini (Burnham & Anderson, 2002).

Pendekatan ini memberikan cara sistematis untuk memilih tingkat regulasi yang optimal dengan mempertimbangkan trade-off antara kecocokan model (*log-likelihood*) dan kompleksitas model (jumlah koefisien non-nol).

2.6. Stunting dan Klasifikasi Status Stunting

Stunting merupakan kondisi gagal tumbuh pada anak balita akibat kekurangan gizi kronis. Secara medis, seorang anak dikategorikan stunting jika tinggi badannya berada di bawah minus dua standar deviasi (-2 SD) dari median standar pertumbuhan anak menurut usia yang ditetapkan oleh *World Health Organization* (WHO) (WHO, 2014).

2.6.1. Prevalensi Stunting dan Target Nasional

Prevalensi stunting didefinisikan sebagai proporsi anak balita yang mengalami stunting dalam suatu populasi. Pemerintah Indonesia melalui Rencana Pembangunan Jangka Menengah Nasional (RPJMN) 2020-2024 menetapkan target penurunan prevalensi stunting nasional hingga 14% pada tahun 2024. Hasil Survei Status Gizi Indonesia (SSGI) 2024 menunjukkan prevalensi stunting nasional sebesar 19,8%, dengan disparitas antar provinsi berkisar antara 8,7% hingga 37,0% (Kemenkes, 2024).

2.6.2. Klasifikasi Status Stunting Tinggi dan Rendah

Dalam penelitian ini, untuk keperluan analisis klasifikasi menggunakan regresi logistik biner, status stunting pada tingkat provinsi dikategorikan menjadi dua kelas:

1. **Stunting Rendah (kode 0):** provinsi dengan prevalensi stunting kurang dari 20%.
2. **Stunting Tinggi (kode 1):** provinsi dengan prevalensi stunting sebesar 20% atau lebih.

Pemilihan batas threshold 20% didasarkan pada target nasional RPJMN dan rekomendasi WHO (WHO, 2014). Berdasarkan kriteria tersebut, dari 38 provinsi di Indonesia, terdapat 12 provinsi dalam kategori stunting rendah dan 26 provinsi dalam kategori stunting tinggi.

2.6.3. Faktor-Faktor Risiko Stunting

Berdasarkan kerangka konsep SSGI 2024 (Kemenkes, 2024), faktor-faktor yang mempengaruhi stunting meliputi:

1. Faktor ibu (jarak kehamilan, usia ibu, ANC, TTD, BBLR)
2. Faktor pemberian ASI dan MPASI (IMD, ASI eksklusif, MDD)
3. Faktor kesehatan (imunisasi, ISPA, diare)
4. Faktor lingkungan (air minum layak, sanitasi layak)
5. Faktor sosial ekonomi (JKN, pendidikan ibu, perkotaan, status ekonomi)

Keberadaan banyak faktor yang diduga berpengaruh ini menuntut pendekatan analisis yang mampu melakukan seleksi variabel secara otomatis. Regresi logistik LASSO dengan optimasi AIC dipilih sebagai metode yang tepat untuk mengidentifikasi faktor-faktor dominan di antara banyak variabel prediktor tersebut.

BAB III

METODE PENELITIAN

3.1. Jenis Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode inferensial statistik yang bertujuan untuk membangun dan mengoptimalkan model regresi logistik Lasso dalam mengklasifikasikan faktor risiko stunting pada anak berdasarkan data Survei Status Gizi Indonesia (SSGI) tahun 2024. Penelitian ini termasuk dalam penelitian terapan karena menerapkan metode statistika untuk menyelesaikan permasalahan pada data nyata. Analisis dilakukan melalui beberapa tahapan, yaitu penyusunan model regresi logistik biner sebagai model awal menggunakan metode *Maximum Likelihood Estimation* (MLE), pengujian multikolinearitas pada variabel prediktor, penerapan standarisasi data, pembentukan model regresi logistik LASSO, serta penentuan parameter penalti (λ) optimum menggunakan kriteria *Akaike Information Criterion* (AIC). Model yang diperoleh selanjutnya dievaluasi dan diinterpretasikan untuk mengidentifikasi faktor-faktor risiko yang berpengaruh terhadap klasifikasi stunting. Seluruh proses analisis dilakukan menggunakan perangkat lunak R.

3.2. Sumber Data

Data bersumber dari SSGI 2024, yaitu survei nasional yang dirancang untuk memperoleh informasi terkini mengenai status gizi masyarakat Indonesia. Survei dilaksanakan oleh Badan Kebijakan Pembangunan Kesehatan, Kementerian Kesehatan Republik Indonesia. Data yang digunakan dalam penelitian ini terdiri dari

38 observasi, yang masing-masing merepresentasikan 38 provinsi di Indonesia dan setiap observasi memuat 21 variabel.

Meskipun jumlah observasi dalam penelitian ini relatif kecil ($n = 38$ provinsi) dibandingkan jumlah variabel prediktor ($p = 19$), kondisi ini justru merupakan situasi yang tepat untuk penerapan metode regularisasi LASSO. Tibshirani (1996) menjelaskan bahwa LASSO dirancang untuk menangani kondisi di mana p mendekati n , karena penalti L_1 secara otomatis melakukan seleksi variabel sehingga model yang dihasilkan tetap parsimoni. Dalam kondisi n kecil, penggunaan *maximum likelihood estimation* tanpa regularisasi akan menghasilkan estimasi yang sangat tidak stabil akibat masalah *complete separation* dan multikolinearitas (Hosmer et al., 2013). Dengan demikian, penerapan LASSO pada data ini bukan sekadar pilihan, melainkan merupakan kebutuhan metodologis.

3.2.1. Populasi dan Sampel

Populasi target: Seluruh balita usia 0–59 bulan yang tercatat dalam SSGI 2024.

Teknik sampling: SSGI 2024 menggunakan desain multistage probability sampling yang representatif nasional dan provinsi. Penelitian ini menggunakan seluruh data yang tersedia setelah data cleaning.

Kriteria inklusi:

- Balita usia 0–59 bulan dalam SSGI 2024.
- Memiliki data lengkap untuk variabel respons (status gizi).
- Memiliki data lengkap minimal 80% variabel prediktor.

Kriteria eksklusi:

- Provinsi dengan data missing pada variabel respons.
- Nilai ekstrem yang tidak masuk akal (biological implausible values).
- Data duplikat.

3.3. Variabel Penelitian

3.3.1. Variabel Respons

Variabel respons dalam penelitian ini adalah status stunting tingkat provinsi yang dikategorikan menjadi dua kelas berdasarkan target nasional Rencana Pembangunan Jangka Menengah Nasional (RPJMN) 2020-2024:

1. **Stunting Rendah (kode 0):** provinsi dengan prevalensi stunting kurang dari 20%.
2. **Stunting Tinggi (kode 1):** provinsi dengan prevalensi stunting sebesar 20% atau lebih.

Pemilihan threshold 20% didasarkan pada rekomendasi WHO yang menetapkan ambang batas masalah kesehatan masyarakat untuk stunting adalah 20/

3.3.2. Variabel Prediktor

Variabel prediktor dalam penelitian ini dikelompokkan ke dalam lima domain utama berdasarkan kerangka konsep SSGI 2024, yaitu faktor ibu, faktor pemberian ASI dan MPASI, faktor kesehatan dan morbiditas, faktor lingkungan dan akses, serta faktor sosial ekonomi.

Variabel prediktor dalam penelitian ini terdiri dari 19 variabel yang dikelompokkan ke dalam lima domain berdasarkan kerangka konsep SSGI 2024, seluruhnya diukur dalam satuan persen (0–100%) pada tingkat provinsi.

Domain pertama adalah faktor ibu, yang terdiri dari sembilan variabel: proporsi ibu dengan jarak kehamilan kurang dari dua tahun (Jarak_Kehamilan), proporsi ibu hamil di usia kurang dari 21 tahun (Umur_Ibu), proporsi ibu dengan pemeriksaan kehamilan minimal empat kali (ANC_K4) dan minimal enam kali (ANC_K6), proporsi ibu yang menerima Tablet Tambah Darah (TTD_Diterima), proporsi ibu yang mengonsumsi minimal 90 tablet TTD (TTD_Diminum_90), dan proporsi balita dengan berat lahir kurang dari 2.500 gram (BBLR).

Domain kedua adalah faktor pemberian ASI dan MPASI, terdiri dari tiga variabel: proporsi balita yang mendapat Inisiasi Menyusu Dini (IMD), proporsi bayi usia 0–5 bulan yang hanya menerima ASI dalam 24 jam terakhir (ASI_Eksklusif), dan proporsi balita usia 6–23 bulan yang memenuhi keragaman pangan minimal atau *Minimum Dietary Diversity* (MDD).

Domain ketiga adalah faktor kesehatan dan morbiditas, terdiri dari tiga variabel: proporsi anak usia 12–23 bulan dengan imunisasi dasar lengkap (Imunisasi_Lengkap), proporsi balita yang mengalami ISPA dalam satu bulan terakhir (ISPA), dan proporsi balita yang mengalami diare dalam satu bulan terakhir (Diare).

Domain keempat adalah faktor lingkungan dan akses, terdiri dari dua variabel: proporsi rumah tangga dengan akses air minum layak (Air_Minum_Layak) dan proporsi rumah tangga dengan akses sanitasi layak (Sanitasi_Layak).

Domain kelima adalah faktor sosial ekonomi, terdiri dari empat variabel: proporsi balita dengan kepemilikan Jaminan Kesehatan Nasional (JKN_Kepemilikan), proporsi ibu berpendidikan diploma atau sarjana (Pendidikan_Ibu_Tinggi), proporsi balita yang tinggal di wilayah perkotaan (Perkotaan), dan proporsi balita dari rumah tangga dengan status ekonomi kuintil teratas (Status_Ekonomi_Teratas).

Secara keseluruhan, terdapat 19 variabel prediktor yang digunakan dalam

penelitian ini. Seluruh variabel diukur dalam satuan persen (0–100%) pada tingkat provinsi, sehingga setiap nilai merepresentasikan proporsi atau prevalensi karakteristik tertentu dalam suatu provinsi.

Tabel 3.1 Variabel Penelitian

Variabel	Keterangan	Kode
Y	Status stunting balita	Stunting
X1	Jarak kehamilan dengan anak sebelumnya	Jarak_Kehamilan
X2	Usia ibu saat memulai kehamilan	Umur_Ibu
X3	Pemeriksaan kehamilan ANC K4	ANC_K4
X4	Pemeriksaan kehamilan ANC K6	ANC_K6
X5	Penerimaan Tablet Tambah Darah	TTD_Diterima
X6	Konsumsi Tablet Tambah Darah minimal 90 tablet	TTD_Diminum_90
X7	Berat badan lahir rendah	BBLR
X8	Inisiasi Menyusu Dini	IMD
X9	ASI eksklusif	ASI_Eksklusif
X10	Keragaman pangan minimal (MDD)	MDD
X11	Imunisasi dasar lengkap	Imunisasi_Lengkap
X12	Kejadian ISPA	ISPA
X13	Kejadian diare	Diare
X14	Akses air minum layak	Air_Minum_Layak
X15	Akses sanitasi layak	Sanitasi_Layak
X16	Kepemilikan Jaminan Kesehatan Nasional	JKN_Kepemilikan
X17	Pendidikan ibu tingkat tinggi	Pendidikan_Ibu_Tinggi
X18	Wilayah perkotaan	Perkotaan
X19	Status ekonomi teratas	Status_Ekonomi_Teratas

3.4. Metode Pengumpulan Data

Data yang digunakan merupakan data sekunder yang diperoleh melalui prosedur resmi Kementerian Kesehatan RI. Data telah melalui quality control internal SSGI. Penelitian menjaga aspek etika dan kerahasiaan data responden.

3.5. Software

Penelitian ini menggunakan perangkat lunak R versi 4.5.0 untuk pengolahan dan analisis data. Bahasa pemrograman R dipilih karena menyediakan berbagai paket statistik yang mendukung analisis regresi logistik LASSO serta visualisasi hasil analisis. Paket (*library*) yang digunakan dalam penelitian ini meliputi:

- **car**, digunakan untuk mendeteksi multikolinearitas antarvariabel prediktor melalui nilai *Variance Inflation Factor* (VIF).
- **caret**, digunakan untuk pembagian data (*data splitting*) serta evaluasi kinerja model melalui *confusion matrix*.
- **glmnet**, digunakan untuk membangun model regresi logistik LASSO dan memperoleh estimasi koefisien pada berbagai nilai parameter penalti (λ).
- **pROC**, digunakan untuk mengevaluasi kemampuan klasifikasi model melalui kurva *Receiver Operating Characteristic* (ROC) dan nilai *Area Under the Curve* (AUC).

3.6. Source code

Kode sumber utama berasal dari paket R glmnet (Friedman et al., 2010) dan tutorial terbuka di platform seperti RPubs. Kode dimodifikasi untuk mengimplementasikan pemilihan parameter penalti menggunakan *Akaike Information Criterion* (AIC) dengan derajat bebas efektif diestimasi sebagai jumlah koefisien non-nol (Zou et al., 2007).

3.7. Tahapan Analisis

Tahapan analisis dalam penelitian ini dilakukan melalui beberapa langkah sebagai berikut:

1. Import Data

Data hasil Survei Status Gizi Indonesia (SSGI) 2024 diimpor ke dalam perangkat lunak R menggunakan fungsi `read.csv()` untuk membaca data dalam format `.csv` sehingga dapat dilakukan proses pengolahan dan analisis lebih lanjut.

2. Pembersihan Data (Data Cleaning)

Pada tahap ini dilakukan:

- Pemeriksaan nilai hilang (*missing values*) menggunakan fungsi `is.na()` dan `colSums()`,
- Pemeriksaan data duplikat menggunakan fungsi `duplicated()`,
- Pemeriksaan validitas rentang data dengan fungsi (*range checking*),
- Pemeriksaan struktur variabel melalui fungsi `str()` dan `as.numeric()`.

3. Eksplorasi Data (Descriptive Analysis)

Analisis deskriptif dilakukan menggunakan statistik ringkas untuk menggambarkan karakteristik masing-masing variabel, serta penyajian grafik untuk pola awal data.

4. Pemeriksaan Multikolinearitas

Untuk memeriksa adanya hubungan linear antarvariabel prediktor digunakan *Variance Inflation Factor* (VIF). Multikolinearitas dapat menyebabkan estimasi koefisien menjadi tidak stabil sehingga diperlukan metode regularisasi.

5. Standardisasi Variabel Prediktor

Variabel prediktor distandardisasi menggunakan fungsi `scale()` agar seluruh variabel berada pada skala yang sama sebelum diterapkan regresi logistik LASSO.

6. Penyusunan Model Dasar: Regresi Logistik

Model regresi logistik biner digunakan untuk memodelkan probabilitas kejadian stunting berdasarkan variabel-variabel prediktor yang diamati. Parameter model diestimasi menggunakan metode *Maximum Likelihood Estimation* (MLE). Proses optimasi fungsi likelihood dilakukan melalui algoritma *Iteratively Reweighted Least Squares* (IRLS) yang tersedia pada fungsi `glm()` dalam perangkat lunak R. Model regresi logistik penuh digunakan sebagai model pembanding sebelum diterapkan metode regularisasi Lasso.

7. Penerapan Regresi Lasso

Regresi Lasso diterapkan untuk melakukan:

- Seleksi variabel prediktor yang paling relevan,
- Penyusutan koefisien (*shrinkage*) guna mengatasi multikolinearitas,
- Regularisasi model agar lebih stabil.

8. Pemilihan Parameter Penalti Menggunakan AIC

Parameter regularisasi λ dipilih menggunakan *Akaike Information Criterion* (AIC) untuk membandingkan kompleksitas dan *goodness-of-fit*. Nilai λ yang menghasilkan AIC terkecil dipilih sebagai parameter penalti optimal.

9. Interpretasi Model Terpilih

Koefisien model dengan variabel aktif (koefisien tidak bernilai nol) diinterpre-

tasikan melalui *odds ratio* untuk menjelaskan faktor-faktor yang berpengaruh terhadap stunting.

10. Evaluasi Kinerja Model

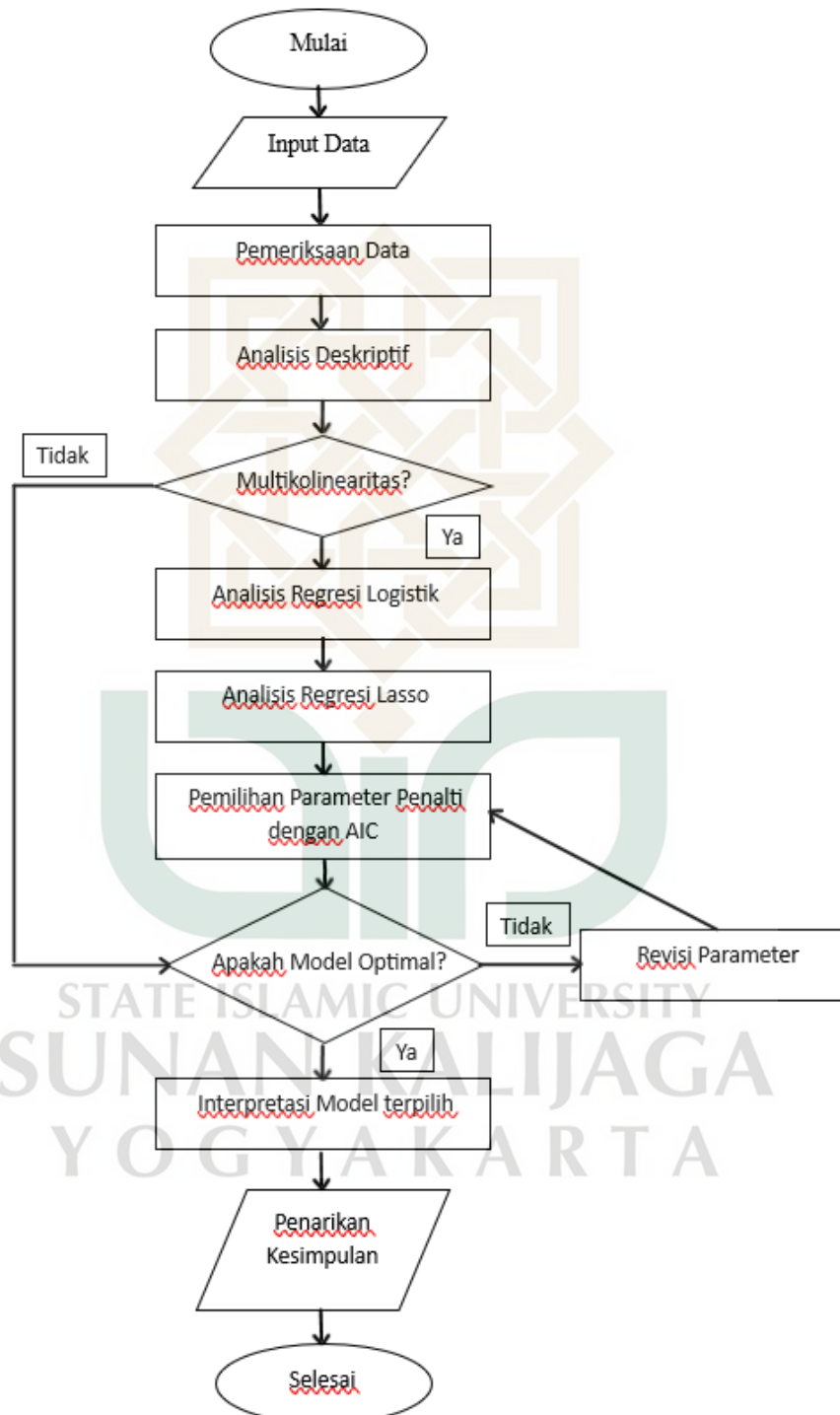
Kinerja model dievaluasi menggunakan *confusion matrix* dari paket `caret`.

Kemampuan diskriminasi model dianalisis menggunakan kurva ROC dan nilai AUC melalui fungsi `roc()` dan `auc()` dari paket `pROC`.

11. Penarikan Kesimpulan

Berdasarkan hasil model akhir, ditarik kesimpulan mengenai variabel yang paling berpengaruh terhadap prevalensi stunting pada tingkat provinsi.

3.8. Diagram Alur Penelitian



BAB IV

HASIL DAN PEMBAHASAN

4.1. Deskripsi Data Penelitian

Data yang digunakan dalam penelitian ini merupakan data sekunder yang bersumber dari Survei Status Gizi Indonesia (SSGI) Tahun 2024. Unit analisis dalam penelitian ini adalah 38 provinsi di Indonesia, sehingga jumlah observasi yang digunakan sebanyak 38 observasi. Variabel yang digunakan terdiri atas satu variabel respons dan 19 variabel prediktor yang merepresentasikan faktor-faktor yang diduga berpengaruh terhadap kejadian stunting. Variabel prediktor tersebut meliputi faktor kesehatan ibu, pelayanan kesehatan, status gizi anak, penyakit infeksi, kondisi lingkungan, kepemilikan jaminan kesehatan, tingkat pendidikan ibu, karakteristik wilayah, serta kondisi sosial ekonomi masyarakat.

Variabel respons dalam penelitian ini dinyatakan dalam bentuk biner, yaitu kategori stunting rendah dan kategori stunting tinggi. Pembentukan kategori dilakukan berdasarkan target prevalensi stunting nasional tahun 2024 sebesar 20%. Penggunaan target nasional sebagai batas klasifikasi bertujuan agar pengelompokan provinsi mengacu pada indikator yang telah ditetapkan pemerintah dalam upaya percepatan penurunan stunting. Dengan demikian, provinsi yang memiliki prevalensi stunting melebihi target nasional dikategorikan sebagai provinsi dengan tingkat stunting tinggi, sedangkan provinsi yang telah mencapai atau berada di bawah target nasional dikategorikan sebagai provinsi dengan tingkat stunting rendah.

Secara matematis, pembentukan variabel respon dapat dinyatakan sebagai

berikut:

$$Y = \begin{cases} 1, & \text{jika prevalensi stunting} \geq 20\%, \\ 0, & \text{jika prevalensi stunting} < 20\%. \end{cases} \quad (4.1)$$

dengan

Y kategori prevalensi stunting pada suatu provinsi;

$Y = 1$ provinsi dengan prevalensi stunting di atas target nasional (stunting tinggi);

$Y = 0$ provinsi dengan prevalensi stunting di bawah atau sama dengan target nasional (stunting rendah).

Tabel 4.1 Pengelompokan Provinsi Berdasarkan Kategori Stunting

Stunting Rendah (Y=0)		Stunting Tinggi (Y=1)	
Provinsi	Prevalensi (%)	Provinsi	Prevalensi (%)
Bali	8,7	Aceh	28,6
Bengkulu	18,8	Bangka Belitung	20,1
DI Yogyakarta	17,4	Banten	21,1
DKI Jakarta	17,3	Gorontalo	23,8
Jambi	17,1	Kalimantan Barat	26,8
Jawa Barat	15,9	Kalimantan Selatan	22,9
Jawa Tengah	17,1	Kalimantan Tengah	22,1
Jawa Timur	14,7	Kalimantan Timur	22,2
Kalimantan Utara	17,6	Maluku	28,4
Kepulauan Riau	15,0	Maluku Utara	23,2
Lampung	15,9	Nusa Tenggara Barat	29,8
Sumatera Selatan	15,9	Nusa Tenggara Timur	37,0
		Papua	24,7
		Papua Barat	24,6
		Papua Barat Daya	30,5
		Papua Pegunungan	24,8
		Papua Selatan	24,8
		Papua Tengah	24,7
		Riau	20,1
		Sulawesi Barat	35,4
		Sulawesi Selatan	23,3
		Sulawesi Tengah	26,1
		Sulawesi Tenggara	26,1
		Sulawesi Utara	20,8
		Sumatera Barat	24,9
		Sumatera Utara	22,0

Berdasarkan Tabel 4.1, diperoleh 26 provinsi (68,42%) yang termasuk dalam kategori stunting tinggi ($Y = 1$) dan 12 provinsi (31,58%) yang termasuk dalam kategori stunting rendah ($Y = 0$). Distribusi tersebut menunjukkan bahwa sebagian besar provinsi di Indonesia masih memiliki prevalensi stunting yang melebihi target nasional sebesar 20%, sehingga diperlukan identifikasi faktor-faktor yang berkontribusi terhadap tingginya prevalensi stunting sebagai dasar penyusunan model klasifikasi menggunakan regresi logistik LASSO.

Selanjutnya, penelitian ini menggunakan 19 variabel prediktor yang terdiri atas jarak kehamilan berisiko, umur ibu berisiko, cakupan ANC K4, cakupan ANC K6, penerimaan tablet tambah darah (TTD), konsumsi TTD minimal 90 tablet, kejadian bayi berat lahir rendah (BBLR), inisiasi menyusui dini (IMD), ASI eksklusif, minimum dietary diversity (MDD), imunisasi lengkap, kejadian ISPA, kejadian diare, akses air minum layak, sanitasi layak, kepemilikan Jaminan Kesehatan Nasional (JKN), pendidikan ibu, persentase wilayah perkotaan, dan status ekonomi teratas. Seluruh variabel prediktor berbentuk numerik dengan skala persentase dan tidak ditemukan nilai hilang (missing value), sehingga seluruh data dapat digunakan pada tahapan analisis tanpa memerlukan proses imputasi atau penghapusan observasi.

4.2. Analisis Deskriptif

Sebelum dilakukan pemodelan menggunakan regresi logistik LASSO, terlebih dahulu dilakukan analisis deskriptif untuk memperoleh gambaran umum mengenai karakteristik data penelitian. Analisis deskriptif bertujuan untuk mengetahui penyebaran nilai setiap variabel prediktor melalui ukuran pemusatan dan penyebaran data, sehingga dapat memberikan informasi awal mengenai kondisi data yang akan digunakan dalam proses pemodelan.

Variabel prediktor yang digunakan dalam penelitian ini terdiri atas 19 indikator yang merepresentasikan faktor-faktor yang diduga berhubungan dengan kejadian stunting pada anak, yaitu *Jarak Kehamilan Risiko*, *Umur Ibu Risiko*, *ANC K4*, *ANC K6*, *TTD Diterima*, *TTD Diminum 90*, *BBLR*, *IMD*, *ASI Eksklusif*, *MDD*, *Imunisasi Lengkap*, *ISPA*, *Diare*, *Air Minum Layak*, *Sanitasi Layak*, *JKN Kepemilikan*, *Pendidikan Ibu Tinggi*, *Perkotaan*, dan *Status Ekonomi Teratas*. Ringkasan statistik deskriptif dari seluruh variabel penelitian disajikan pada Tabel 4.2.

Tabel 4.2 Statistik Deskriptif Variabel Penelitian

Variabel	Min	Q1	Median	Mean	Max	SD
Jarak Kehamilan Risiko	5,10	9,28	11,10	11,88	20,60	3,65
Umur Ibu Risiko	4,10	5,63	6,45	7,29	14,20	2,77
ANC K4	40,30	58,77	70,00	68,51	92,00	12,90
ANC K6	5,80	18,32	22,90	25,07	67,60	11,98
TTD Diterima	65,10	73,42	83,00	81,69	96,40	9,37
TTD Diminum 90 Hari	18,40	32,12	48,15	47,83	78,50	16,78
BBLR	3,30	5,55	6,55	6,56	9,30	1,50
IMD	27,70	50,70	58,85	57,37	73,10	10,10
ASI Eksklusif	41,40	57,70	63,55	63,80	85,20	10,77
MDD	18,40	36,15	42,30	42,31	67,90	13,26
Imunisasi Lengkap	5,40	38,58	48,75	48,96	89,50	17,35
ISPA	16,80	23,70	29,25	30,15	49,10	8,60
Diare	0,90	2,00	2,85	2,90	5,80	1,26
Air Minum Layak	67,20	82,12	89,05	86,41	99,40	8,80
Sanitasi Layak	62,00	74,60	79,00	78,53	89,10	6,43
JKN Kepemilikan	19,50	38,75	49,15	51,34	89,20	17,53
Pendidikan Ibu Tinggi	1,20	4,23	5,70	6,03	13,10	2,81
Perkotaan	28,20	33,20	38,35	41,60	100,00	13,45
Status Ekonomi Teratas	6,50	9,80	11,85	12,92	32,50	5,20

Berdasarkan Tabel 4.2, terlihat bahwa setiap variabel memiliki rentang nilai yang berbeda sehingga menunjukkan adanya keragaman karakteristik antarprovinsi. Nilai standar deviasi pada tabel ini memperkuat gambaran tersebut. Variabel dengan standar deviasi besar, seperti Imunisasi Lengkap (17,35) dan JKN Kepemilikan (17,53), menunjukkan bahwa capaian program pada indikator ini bervariasi cukup lebar antarprovinsi. Sebaliknya, variabel dengan standar deviasi kecil, seperti BBLR (1,50) dan Diare (1,26), menunjukkan bahwa nilai antarprovinsi pada

indikator ini relatif seragam.

Secara keseluruhan, statistik deskriptif menunjukkan bahwa seluruh variabel memiliki variasi nilai yang cukup beragam. Variasi tersebut mengindikasikan bahwa setiap indikator memiliki karakteristik yang berbeda pada masing-masing provinsi, sehingga layak digunakan sebagai variabel prediktor dalam pembentukan model regresi logistik LASSO untuk mengidentifikasi faktor-faktor yang berhubungan dengan klasifikasi stunting di Indonesia.

4.3. Uji Multikolinearitas

Uji multikolinearitas dilakukan untuk mengetahui ada atau tidaknya hubungan yang kuat antarvariabel prediktor pada model regresi logistik awal. Keberadaan multikolinearitas dapat menyebabkan ketidakstabilan dalam pendugaan parameter sehingga koefisien regresi menjadi kurang reliabel untuk diinterpretasikan. Deteksi multikolinearitas pada penelitian ini dilakukan menggunakan nilai *Variance Inflation Factor* (VIF). Suatu variabel dikatakan mengalami multikolinearitas apabila memiliki nilai VIF lebih dari 10. Hasil pengujian multikolinearitas disajikan pada Tabel 4.3.

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

Tabel 4.3 Hasil Uji *Variance Inflation Factor* (VIF)

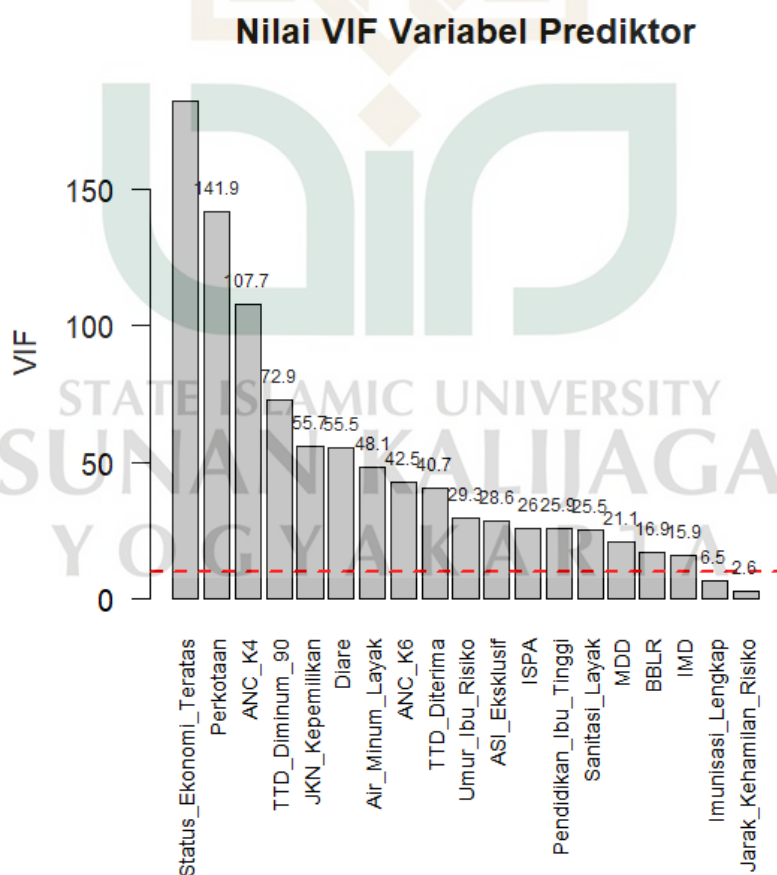
No.	Variabel	Nilai VIF	Kesimpulan
1	Jarak Kehamilan Risiko	2,6367	Tidak terjadi multikolinearitas
2	Umur Ibu Risiko	29,3489	Terjadi multikolinearitas
3	ANC K4	107,7212	Terjadi multikolinearitas
4	ANC K6	42,5331	Terjadi multikolinearitas
5	TTD Diterima	40,6825	Terjadi multikolinearitas
6	TTD Diminum 90 Hari	72,8843	Terjadi multikolinearitas
7	BBLR	16,9376	Terjadi multikolinearitas
8	IMD	15,9242	Terjadi multikolinearitas
9	ASI Eksklusif	28,6172	Terjadi multikolinearitas
10	MDD	21,0826	Terjadi multikolinearitas
11	Imunisasi Lengkap	6,5208	Tidak terjadi multikolinearitas
12	ISPA	25,9795	Terjadi multikolinearitas
13	Diare	55,4728	Terjadi multikolinearitas
14	Air Minum Layak	48,0654	Terjadi multikolinearitas
15	Sanitasi Layak	25,5165	Terjadi multikolinearitas
16	JKN Kepemilikan	55,6565	Terjadi multikolinearitas
17	Pendidikan Ibu Tinggi	25,8718	Terjadi multikolinearitas
18	Perkotaan	141,8663	Terjadi multikolinearitas
19	Status Ekonomi Teratas	182,4764	Terjadi multikolinearitas

Berdasarkan Tabel 4.3, diperoleh bahwa sebagian besar variabel prediktor memiliki nilai VIF lebih besar dari 10, sehingga terindikasi mengalami multikolinearitas. Dari 19 variabel prediktor yang digunakan, hanya variabel *Jarak Kehamilan Risiko* dan *Imunisasi Lengkap* yang memiliki nilai VIF masing-masing sebesar 2,6367 dan 6,5208, sehingga tidak menunjukkan adanya gejala multikolinearitas. Sementara itu, sebanyak 17 variabel lainnya memiliki nilai VIF di atas 10.

Variabel *Status Ekonomi Teratas* memiliki nilai VIF tertinggi, yaitu sebesar 182,4764, diikuti oleh variabel *Perkotaan* sebesar 141,8663 dan *ANC K4* sebesar 107,7212. Nilai VIF yang sangat tinggi tersebut menunjukkan adanya hubungan

linear yang kuat antarvariabel prediktor, sehingga dapat menyebabkan pendugaan koefisien regresi menjadi tidak stabil dan memiliki galat baku yang besar.

Keberadaan multikolinearitas menyebabkan interpretasi pengaruh masing-masing variabel prediktor terhadap status stunting menjadi kurang akurat karena informasi yang dikandung oleh beberapa variabel saling tumpang tindih. Oleh karena itu, diperlukan metode yang mampu mengatasi permasalahan multikolinearitas sekaligus melakukan seleksi variabel secara otomatis. Pada penelitian ini digunakan regresi logistik LASSO, yang menerapkan penalti L1 sehingga koefisien variabel yang kurang berkontribusi dapat diperkecil hingga bernilai nol, sedangkan variabel yang memiliki kontribusi penting tetap dipertahankan dalam model.



Gambar 4.1 Nilai VIF Variabel Prediktor

Gambar 4.1 menampilkan nilai VIF seluruh variabel prediktor dalam bentuk diagram batang yang diurutkan dari nilai tertinggi ke terendah, lengkap dengan angka VIF di atas setiap batang. Garis putus-putus merah menandai ambang batas VIF sebesar 10. Sebanyak 17 batang berada di atas garis batas ini, mulai dari Status Ekonomi Teratas di ujung kiri hingga BBLR yang nilainya masih sedikit di atas 10. Hanya dua batang di ujung kanan, yaitu IMD dan Jarak Kehamilan Risiko, yang berada di bawah garis batas. Pola batang yang menurun tajam ini memperjelas bahwa multikolinearitas pada data penelitian bukan hanya terjadi pada satu atau dua variabel, tetapi menyebar hampir ke seluruh variabel prediktor.

4.4. Persiapan Data dan Standardisasi

Sebelum data digunakan pada tahap estimasi regresi logistik LASSO, terlebih dahulu dilakukan persiapan data dan standardisasi variabel prediktor. Tahap ini dilakukan setelah hasil uji VIF pada subbab sebelumnya menunjukkan bahwa sebagian besar variabel prediktor memiliki skala nilai yang jauh berbeda satu sama lain.

Variabel prediktor pada penelitian ini memang seluruhnya berbentuk persentase, tetapi rentang nilainya tidak seragam. Sebagai contoh, variabel Diare hanya berkisar antara 0,90 hingga 5,80, sedangkan variabel Perkotaan berkisar antara 28,20 hingga 100,00, sebagaimana telah ditunjukkan pada Tabel 4.2. Apabila data dengan skala yang berbeda ini langsung digunakan pada regresi LASSO, penalti L1 akan menekan variabel dengan skala nilai besar secara tidak proporsional. Akibatnya, proses seleksi variabel akan lebih dipengaruhi oleh perbedaan skala pengukuran, bukan oleh kekuatan hubungan variabel tersebut dengan status stunting. Oleh sebab itu, standardisasi menjadi tahap yang wajib dilakukan sebelum estimasi LASSO, bukan sekadar pilihan teknis.

Proses standardisasi dilakukan dengan mengubah setiap nilai variabel menjadi skor standar (*z-score*), mengikuti persamaan berikut.

$$Z_i = \frac{X_i - \bar{X}}{S_X} \quad (4.2)$$

Keterangan:

Z_i = nilai variabel setelah distandardisasi;

X_i = nilai asli variabel pada observasi ke- i ;

$\bar{X}$ = rata-rata variabel sebelum standardisasi;

S_X = standar deviasi variabel sebelum standardisasi.

Melalui persamaan ini, setiap variabel prediktor diubah sehingga memiliki rata-rata sama dengan 0 dan standar deviasi sama dengan 1. Dengan demikian, seluruh variabel berada pada skala yang setara dan dapat dibandingkan secara adil dalam proses penalisasi LASSO, terlepas dari satuan atau rentang nilai aslinya.

Tabel 4.4 Rata-rata dan Standar Deviasi Variabel Prediktor Sebelum Standardisasi

Variabel	Rata-rata	Standar Deviasi
Jarak Kehamilan Risiko	11.882	3.653
Umur Ibu Risiko	7.287	2.774
ANC K4	68.505	12.903
ANC K6	25.068	11.984
TTD Diterima	81.692	9.372
TTD Diminum 90 Hari	47.832	16.779
BBLR	6.558	1.498
IMD	57.368	10.096
ASI Eksklusif	63.800	10.765
MDD	42.308	13.263
Imunisasi Lengkap	48.961	17.349
ISPA	30.147	8.604
Diare	2.903	1.263
Air Minum Layak	86.408	8.795
Sanitasi Layak	78.532	6.426
JKN Kepemilikan	51.339	17.532
Pendidikan Ibu Tinggi	6.029	2.808
Perkotaan	41.597	13.445
Status Ekonomi Teratas	12.924	5.198

Verifikasi terhadap hasil standardisasi dilakukan untuk memastikan bahwa proses transformasi data telah berjalan dengan benar. Hasil verifikasi menunjukkan bahwa seluruh variabel prediktor telah memiliki rata-rata mendekati 0 dan standar deviasi sebesar 1. Ringkasan hasil verifikasi disajikan pada Tabel 4.5.

Tabel 4.5 Hasil Verifikasi Standardisasi Variabel Prediktor

Indikator	Hasil
Rata-rata seluruh variabel setelah standardisasi	0
Standar deviasi seluruh variabel setelah standardisasi	1

Nilai rata-rata dan standar deviasi sebelum standardisasi tetap disimpan se-

bagaimana disajikan pada Tabel 4.4. Penyimpanan nilai tersebut diperlukan agar hasil estimasi model masih dapat ditelusuri kembali ke skala aslinya apabila diperlukan pada tahap interpretasi.

Data yang telah distandardisasi selanjutnya digunakan sebagai input dalam pembentukan model regresi logistik LASSO, termasuk pada proses penyusunan grid nilai λ , estimasi koefisien model, serta pemilihan parameter penalti optimum menggunakan kriteria Akaike Information Criterion (AIC).

4.5. Pembentukan Model Regresi Logistik Lasso

Sebelum dilakukan seleksi variabel dengan LASSO, terlebih dahulu dibentuk model regresi logistik biasa yang melibatkan seluruh 19 variabel prediktor. Model regresi logistik ini menyatakan hubungan antara peluang kejadian stunting kategori 1 dengan seluruh variabel prediktor.

Hasil estimasi model regresi logistik penuh menghasilkan nilai AIC sebesar 40,00 dengan 19 variabel prediktor. Namun demikian, proses estimasi model ini menunjukkan peringatan bahwa algoritma tidak konvergen dan beberapa nilai peluang prediksi bernilai numerik 0 atau 1 secara sempurna. Kondisi ini menandakan terjadinya pemisahan sempurna, yaitu situasi ketika kombinasi variabel prediktor mampu memisahkan kedua kategori respons secara sempurna tanpa tumpang tindih. Pemisahan sempurna menyebabkan estimasi koefisien menjadi tidak terhingga secara teoritis dan tidak stabil secara numerik.

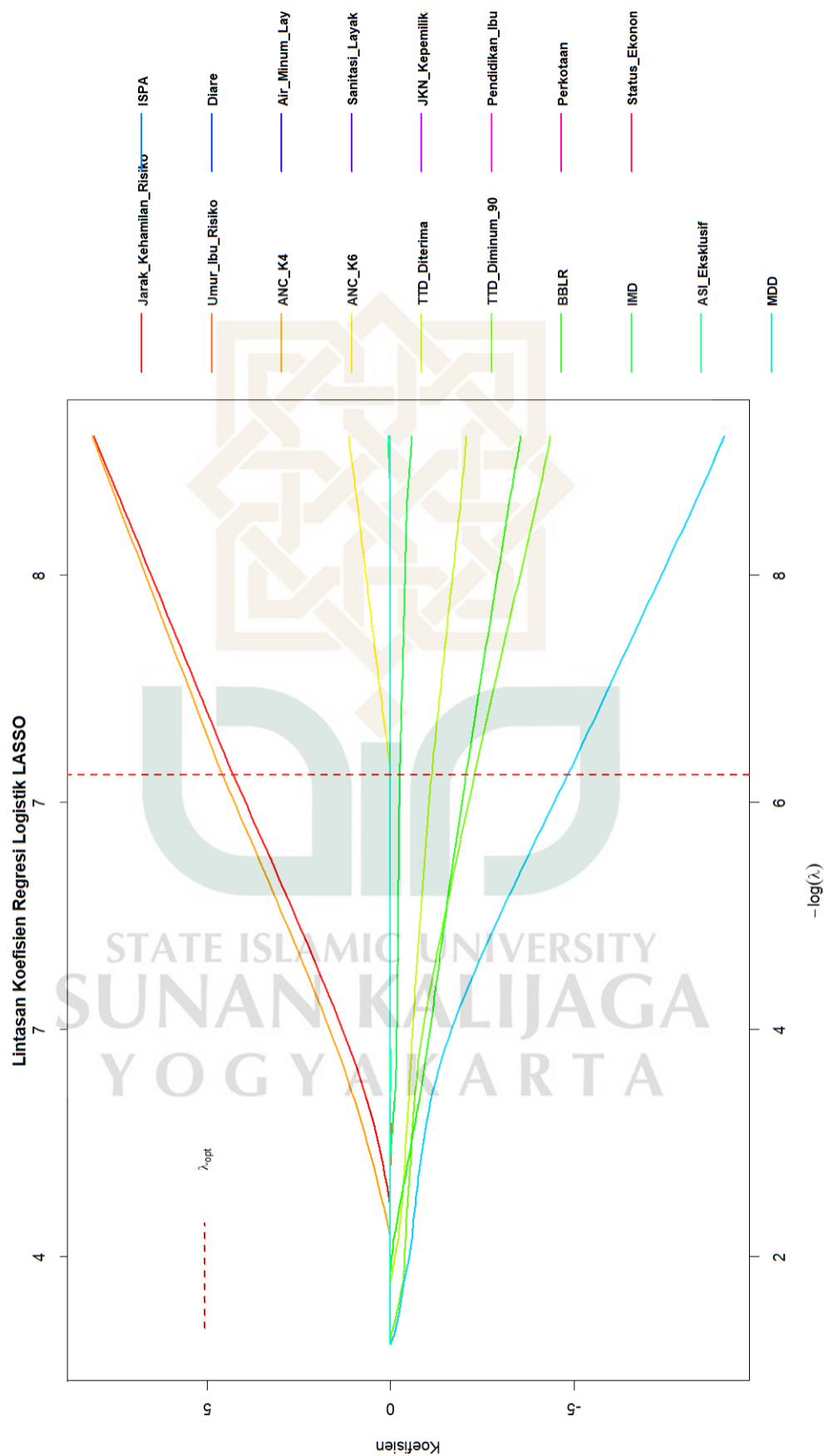
Sebelum dilakukan penentuan parameter penalti optimum menggunakan AIC, terlebih dahulu dibentuk model regresi logistik LASSO menggunakan grid nilai λ yang telah ditentukan. Pada tahap ini, model LASSO menghasilkan 13 variabel prediktor dengan koefisien tidak sama dengan nol, yaitu Jarak Kehamilan Risiko,

Umur Ibu Risiko, ANC K4, IMD, ASI Eksklusif, MDD, Imunisasi Lengkap, ISPA, Diare, Air Minum Layak, JKN Kepemilikan, Pendidikan Ibu Tinggi, dan Perkotaan.

Tabel 4.6 Koefisien Regresi Logistik LASSO

Variabel	Koefisien
Jarak Kehamilan Risiko	0,7611
Umur Ibu Risiko	5,9229
ANC K4	-1,4720
IMD	-0,8506
ASI Eksklusif	0,3477
MDD	-1,3682
Imunisasi Lengkap	-0,6076
ISPA	-2,0306
Diare	4,0135
Air Minum Layak	-2,5497
JKN Kepemilikan	2,0821
Pendidikan Ibu Tinggi	-2,1487
Perkotaan	2,6023

Sementara itu, enam variabel lainnya dieliminasi oleh metode LASSO karena koefisiennya menyusut menjadi nol, yaitu ANC K6, TTD Diterima, TTD Di-minum 90 Hari, BBLR, Sanitasi Layak, dan Status Ekonomi Teratas. Model ini merupakan hasil seleksi awal dan belum menjadi model akhir karena nilai parameter penalti optimum belum ditentukan.



Gambar 4.2 Lintasan Koefisien Regresi Logistik LASSO

Gambar 4.2 memperlihatkan perubahan nilai koefisien setiap variabel prediktor terhadap perubahan nilai $-\log(\lambda)$ pada regresi logistik LASSO. Setiap garis berwarna pada grafik menunjukkan lintasan koefisien dari masing-masing variabel prediktor selama perubahan nilai $-\log(\lambda)$. Garis yang berada di atas nilai nol menunjukkan koefisien positif, sedangkan garis yang berada di bawah nilai nol menunjukkan koefisien negatif. Pada bagian kiri grafik, ketika nilai $-\log(\lambda)$ masih kecil, nilai λ relatif besar sehingga penalti LASSO kuat dan sebagian besar koefisien menyusut mendekati nol. Seiring meningkatnya nilai $-\log(\lambda)$, nilai λ semakin kecil sehingga pengaruh penalti berkurang dan koefisien beberapa variabel mulai menjauh dari nol. Variabel yang koefisiennya tetap berada pada nilai nol tidak masuk ke dalam model pada nilai penalti tersebut, sedangkan variabel dengan koefisien tidak nol dipertahankan dalam model.

Garis vertikal putus-putus berwarna merah menunjukkan posisi λ optimum yang diperoleh berdasarkan nilai AIC minimum. Pada titik tersebut, beberapa variabel telah memiliki koefisien tidak nol sehingga terpilih sebagai prediktor dalam model LASSO-AIC, sedangkan variabel lainnya memiliki koefisien nol dan tidak terpilih. Nilai koefisien yang berada di atas nol menunjukkan hubungan positif terhadap log-odds kategori stunting tinggi, sedangkan koefisien yang berada di bawah nol menunjukkan hubungan negatif terhadap log-odds kategori tersebut. Dengan demikian, lintasan koefisien menunjukkan proses penyusutan dan seleksi variabel yang dilakukan oleh metode LASSO seiring perubahan tingkat penalti.

4.6. Penentuan Parameter Optimum Menggunakan AIC

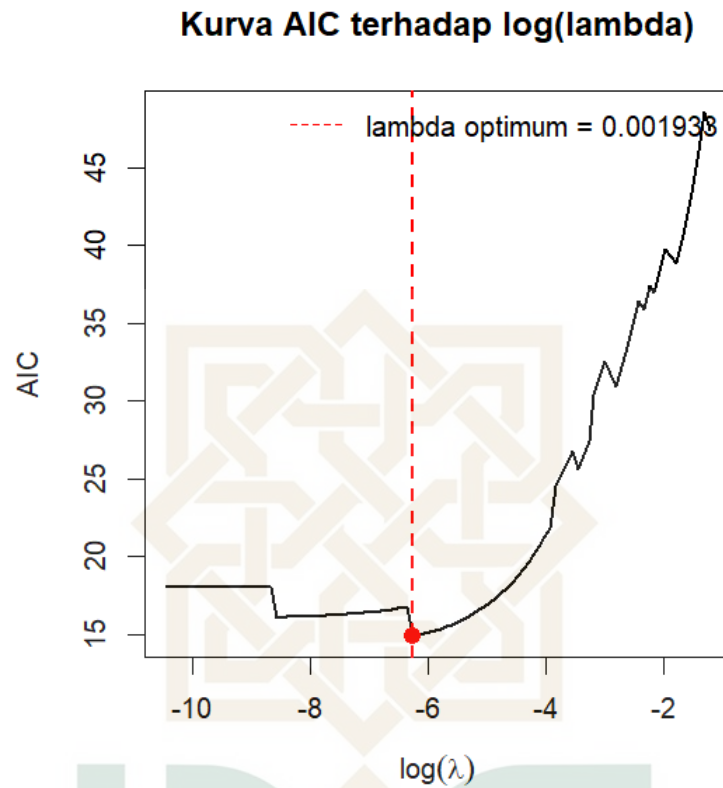
Metode LASSO memerlukan penentuan nilai parameter penalti λ yang optimal. Nilai λ yang terlalu kecil menghasilkan model yang mendekati model penuh, sedangkan nilai λ yang terlalu besar menyebabkan semakin banyak koefisien me-

nyusut menjadi nol sehingga model menjadi terlalu sederhana. Pada penelitian ini, pemilihan nilai λ optimum dilakukan dengan meminimumkan nilai *Akaike Information Criterion* (AIC) pada setiap nilai λ dalam grid pencarian.

Tabel 4.7 Hasil Penentuan Parameter Optimum

Keterangan	Nilai
Lambda optimum	0,001933
AIC minimum	14,8712
Jumlah variabel terpilih	7

Nilai AIC minimum sebesar 14,8712 diperoleh pada nilai $\lambda = 0,001933$, sehingga nilai tersebut dipilih sebagai parameter penalti optimum pada model LAS-SO. Pada nilai λ tersebut, model mempertahankan 7 variabel prediktor dengan koefisien tidak sama dengan nol. Model tersebut dipilih sebagai model terbaik karena menghasilkan nilai AIC minimum, sehingga memberikan keseimbangan yang optimal antara kecocokan model terhadap data dan kompleksitas model di antara seluruh kandidat nilai λ yang dievaluasi.



Gambar 4.3 Kurva AIC terhadap Nilai Log Lambda

Gambar 4.3 menunjukkan perubahan nilai AIC pada setiap nilai parameter penalti λ dalam grid pencarian. Nilai AIC menurun seiring perubahan nilai λ hingga mencapai titik minimum pada $\lambda = 0,001933$, yang ditunjukkan oleh garis vertikal putus-putus berwarna merah dan titik merah pada kurva. Nilai AIC minimum tersebut selanjutnya digunakan sebagai acuan dalam menghitung perbedaan nilai AIC atau ΔAIC pada setiap nilai λ . Nilai ΔAIC dihitung menggunakan persamaan

$$\Delta AIC_i = AIC_i - AIC_{\min} \quad (4.3)$$

Keterangan:

ΔAIC_i = perbedaan nilai AIC pada model ke- i terhadap nilai AIC minimum;

AIC_i = nilai AIC pada nilai λ_i ;

$AIC_{\min}$ = nilai AIC minimum dari seluruh kandidat λ .

Nilai $\Delta AIC = 0$ menunjukkan bahwa nilai λ tersebut menghasilkan AIC minimum, sedangkan nilai ΔAIC yang lebih besar menunjukkan bahwa nilai AIC pada λ tersebut semakin jauh dari nilai AIC minimum. Berdasarkan hasil tersebut, diperoleh $\lambda = 0,001933$ sebagai parameter penalti optimum karena menghasilkan nilai AIC terendah dan $\Delta AIC = 0$.

4.7. Hasil Regresi Logistik LASSO dengan AIC

Berdasarkan nilai λ optimum yang telah ditentukan, dibentuk model regresi logistik LASSO final. Hasil estimasi koefisien model ini disajikan pada tabel berikut.

Tabel 4.8 Koefisien Model Regresi Logistik LASSO

Variabel	Koefisien
Jarak Kehamilan Risiko	4,3076
BBLR	4,6009
ASI Eksklusif	-1,1181
MDD	-2,2914
Imunisasi Lengkap	-2,0567
ISPA	-0,2550
Pendidikan Ibu Tinggi	-4,8505

Dari 19 variabel prediktor, metode LASSO mempertahankan tujuh variabel prediktor dengan koefisien tidak sama dengan nol. Ketujuh variabel tersebut disajikan pada Tabel 4.8. Model juga memiliki satu koefisien intersep, yaitu $\beta_0 = 5,8323$. Intersep tersebut merupakan konstanta dalam model regresi logistik. Dengan demikian, model akhir memiliki delapan parameter, yang terdiri atas satu intersep dan tujuh koefisien variabel prediktor. Sebanyak 12 variabel prediktor lainnya dieliminasi karena koefisiennya menyusut menjadi nol akibat penalti LASSO. Variabel tersebut meliputi Umur Ibu Risiko, ANC K4, ANC K6, TTD Diterima, TTD Diminum 90,

IMD, Diare, Air Minum Layak, Sanitasi Layak, JKN Kepemilikan, Perkotaan, dan Status Ekonomi Teratas.

Ketujuh variabel prediktor yang bertahan berasal dari kelompok faktor ibu, status gizi anak, pelayanan kesehatan, dan penyakit infeksi. Hal ini menunjukkan bahwa faktor yang berkaitan dengan status stunting provinsi bersifat multidimensi dan tidak hanya berasal dari satu aspek tertentu.

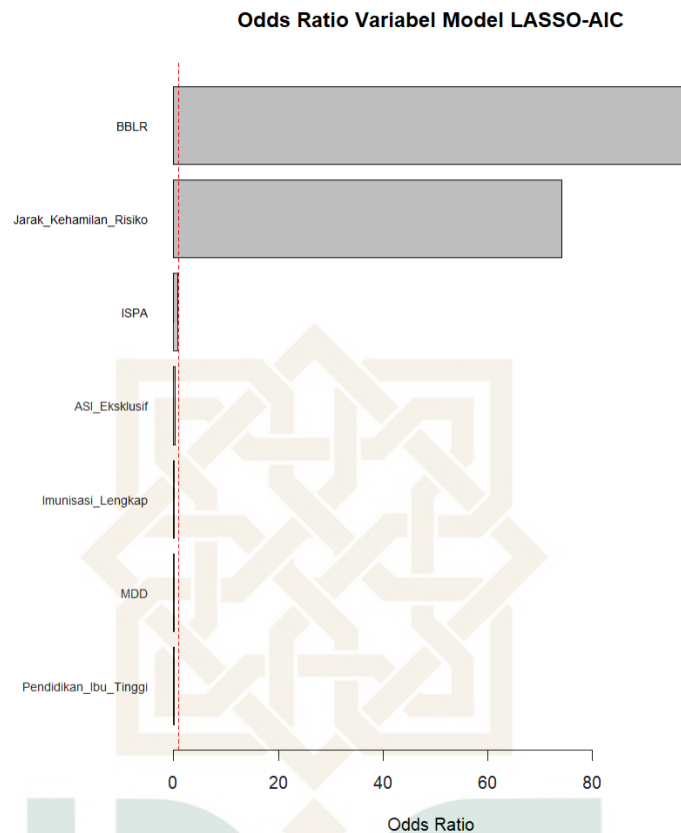
4.8. Interpretasi Odds Ratio

Interpretasi koefisien regresi logistik lebih mudah dipahami dalam bentuk odds ratio (OR). Odds ratio diperoleh dengan mengeksponenkan nilai koefisien regresi.

Tabel 4.9 Odds Ratio Variabel Terpilih

Variabel	Koefisien	Odds Ratio
Jarak Kehamilan Risiko	4,3076	74,2610
BBLR	4,6009	99,5784
ASI Eksklusif	-1,1181	0,3269
MDD	-2,2914	0,1011
Imunisasi Lengkap	-2,0567	0,1279
ISPA	-0,2550	0,7749
Pendidikan Ibu Tinggi	-4,8505	0,0078

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA



Gambar 4.4 Odds Ratio Variabel Model LASSO-AIC

Gambar 4.4 menampilkan odds ratio ketujuh variabel model LASSO dalam bentuk diagram batang. Garis putus-putus di dekat titik nol menandai posisi odds ratio sama dengan satu. Dua batang paling panjang, yaitu BBLR dan Jarak Kehamilan Risiko, terlihat jauh lebih dominan dibandingkan lima batang lainnya, yang panjangnya hampir tidak terlihat pada skala grafik ini. Perbedaan panjang batang yang sangat mencolok ini menunjukkan bahwa kedua variabel tersebut memberi kontribusi jauh lebih besar terhadap peningkatan odds status stunting tinggi dibandingkan variabel lainnya, sementara lima variabel sisanya memang bekerja pada skala odds ratio yang jauh lebih kecil, sebagaimana juga terlihat pada Tabel di atas.

Variabel Jarak Kehamilan Risiko memiliki odds ratio sebesar 74,26. Nilai ini menunjukkan bahwa setiap kenaikan satu satuan standar pada persentase keha-

milan berisiko dikaitkan dengan peningkatan *odds* suatu provinsi untuk berada pada kategori stunting tinggi. Variabel BBLR memiliki odds ratio sebesar 99,58. Nilai ini menunjukkan bahwa setiap kenaikan satu satuan standar pada persentase bayi dengan berat badan lahir rendah dikaitkan dengan peningkatan odds suatu provinsi untuk berada pada kategori stunting tinggi.

Variabel ASI Eksklusif memiliki odds ratio sebesar 0,3269, MDD sebesar 0,1011, dan Imunisasi Lengkap sebesar 0,1279. Ketiga nilai odds ratio ini kurang dari satu, yang berarti peningkatan pada ketiga variabel tersebut dikaitkan dengan penurunan odds status stunting tinggi. Temuan ini sejalan dengan teori bahwa pemberian ASI eksklusif, keragaman pangan yang cukup, dan cakupan imunisasi lengkap merupakan faktor protektif terhadap stunting.

Variabel Pendidikan Ibu Tinggi memiliki odds ratio terkecil, yaitu 0,0078. Nilai ini menunjukkan efek protektif yang sangat kuat, yaitu provinsi dengan persentase ibu berpendidikan tinggi yang lebih besar cenderung memiliki odds status stunting tinggi yang jauh lebih rendah. Temuan ini mendukung teori bahwa pendidikan ibu berperan penting dalam pengasuhan anak, termasuk pengetahuan gizi dan pola asuh yang lebih baik.

Variabel ISPA memiliki odds ratio sebesar 0,7749, yang menunjukkan arah hubungan negatif dengan status stunting tinggi pada tingkat provinsi. Arah hubungan ini perlu dimaknai secara hati-hati. Data pada penelitian ini bersifat agregat tingkat provinsi, sehingga hubungan yang teramati pada tingkat wilayah tidak dapat langsung disamakan dengan hubungan pada tingkat individu. Kemungkinan penjelasan lain adalah adanya korelasi antara ISPA dengan variabel lain yang telah tereliminasi dari model, sehingga efek murni ISPA sulit dipisahkan pada data agregat ini.

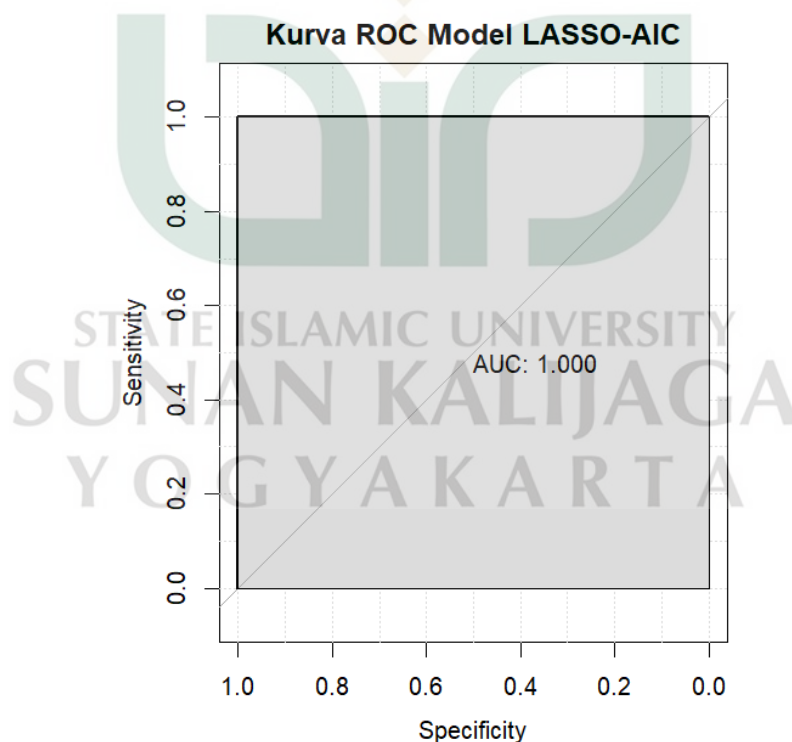
4.9. Evaluasi Model

Evaluasi kinerja model LASSO dilakukan menggunakan confusion matrix serta beberapa ukuran turunannya, yaitu accuracy, sensitivity, specificity, precision, dan area under the ROC curve (AUC).

Tabel 4.10 Confusion Matrix Model LASSO

	Aktual 0	Aktual 1
Prediksi 0	12	0
Prediksi 1	0	26

Berdasarkan hasil confusion matrix, model LASSO menghasilkan accuracy, sensitivity, specificity, dan precision masing-masing sebesar 100 persen. Nilai AUC yang diperoleh dari kurva ROC juga sebesar 1,00, yang berarti model mampu membedakan kedua kategori status stunting secara sempurna pada data yang digunakan.



Gambar 4.5 Kurva ROC Model LASSO-AIC

Gambar 4.5 menampilkan kurva ROC yang membentuk sudut siku sem-

purna, menyentuh sensitivitas 1,0 pada seluruh nilai specificity, dengan luas area di bawah kurva atau AUC sebesar 1,000 sebagaimana tertulis langsung pada grafik. Secara visual, bentuk kurva ini menegaskan bahwa model mampu memisahkan provinsi berkategori stunting tinggi dan rendah tanpa satu pun kesalahan klasifikasi pada data yang digunakan.

Nilai evaluasi yang sempurna ini perlu dimaknai dengan hati-hati. Evaluasi pada penelitian ini dilakukan pada data yang sama dengan data yang digunakan untuk membentuk model, tanpa pemisahan data latih dan data uji maupun validasi silang. Jumlah observasi pada penelitian ini juga tergolong kecil, yaitu hanya 38 provinsi, sedangkan model tetap mempertahankan 7 variabel prediktor. Kombinasi antara ukuran sampel yang kecil dan evaluasi pada data yang sama berpotensi menghasilkan nilai kinerja yang terlalu optimis. Kurva ROC yang sempurna pada Gambar 4.5 sebaiknya dibaca dalam konteks ini, bukan sebagai bukti bahwa model akan berkinerja sama baiknya pada data provinsi lain di luar data penelitian. Oleh karena itu, hasil evaluasi ini lebih tepat dimaknai sebagai kesesuaian model terhadap data yang tersedia, bukan sebagai jaminan kinerja model pada data provinsi lain di luar data penelitian ini.

4.10. Perbandingan Regresi Logistik Penuh dan Regresi Logistik LASSO AIC

Perbandingan dilakukan antara model regresi logistik penuh dengan model regresi logistik LASSO berbasis AIC, untuk menjawab tujuan penelitian mengenai efektivitas metode LASSO dalam melakukan seleksi variabel.

Tabel 4.11 Perbandingan Model Regresi Logistik Penuh dan Regresi Logistik LASSO AIC

Model	Jumlah Variabel	λ	AIC	Kondisi Model	AUC
Regresi Logistik Penuh	19	-	40,00	Tidak konvergen	-
LASSO + AIC	7	0,001933	14,87	Konvergen	1,00

Model regresi logistik penuh menggunakan seluruh 19 variabel prediktor dengan nilai AIC sebesar 40,00. Namun, proses estimasi menunjukkan bahwa model tidak konvergen akibat terjadinya pemisahan sempurna, sehingga estimasi koefisien yang dihasilkan tidak stabil untuk diinterpretasikan. Kondisi tersebut juga berkaitan dengan adanya multikolinearitas pada data, sebagaimana ditunjukkan oleh hasil uji *Variance Inflation Factor* (VIF) pada subbab sebelumnya, di mana sebagian besar variabel prediktor memiliki nilai VIF yang tinggi.

Sebaliknya, regresi logistik LASSO dengan optimasi AIC menghasilkan model yang mempertahankan 7 dari 19 variabel prediktor. Berdasarkan proses optimasi, diperoleh nilai parameter penalti optimum sebesar $\lambda = 0,001933$ yang menghasilkan nilai AIC minimum sebesar 14,87. Pada nilai λ tersebut, penalti LASSO menyusutkan sebagian koefisien menjadi nol sehingga 12 variabel dieliminasi dari model. Dengan demikian, model LASSO + AIC menghasilkan model yang lebih sederhana dibandingkan model regresi logistik penuh.

Selain menghasilkan jumlah variabel yang lebih sedikit, model LASSO + AIC juga menunjukkan kondisi konvergen dengan nilai AUC sebesar 1,00. Hasil ini menunjukkan bahwa model LASSO + AIC mampu membentuk model klasifikasi yang stabil pada data penelitian, sekaligus melakukan seleksi variabel melalui mekanisme penyusutan koefisien. Penyusutan tersebut membantu mengurangi dampak multikolinearitas dan permasalahan estimasi yang muncul pada model regresi logistik penuh.

Berdasarkan perbandingan tersebut, LASSO dengan optimasi AIC menghasilkan model yang lebih sederhana dan konvergen, dengan prediktor berkurang dari 19 menjadi 7 variabel. Oleh karena itu, LASSO + AIC digunakan sebagai model akhir penelitian klasifikasi status stunting pada data SSGI 2024.