RESEARCH ARTICLE • OPEN ACCESS

Analisis Regresi Logistik Biner dan *Random Forest* untuk Prediksi Faktor-Faktor *Stunting* di Pulau Jawa

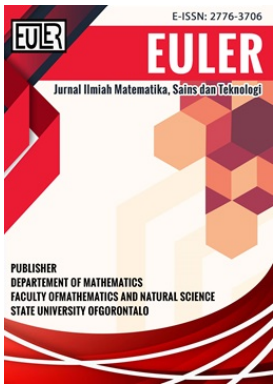
Rizqi Dwi Yuniarsyih R.A. dkk.

Volume 13, Issue 2, Pages 147–156, Aug. 2025

Diterima 17 April 2025, Direvisi 21 Juni 2025, Disetujui 26 Juni 2025, Diterbitkan 1 Juli 2025

To Cite this Article : R. D. Yuniarsyih R.A. dkk.,"Analisis Regresi Logistik Biner dan *Random Forest* untuk Prediksi Faktor-Faktor *Stunting* di Pulau Jawa", *Euler J. Ilm. Mat. Sains dan Teknol.*, vol. 13, no. 2, pp. 147–156, 2025, *https://doi.org/10.37905/euler.v13i2.31680*

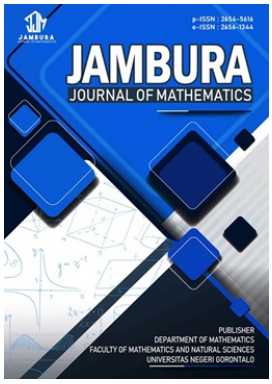
JOURNAL INFO • EULER : JURNAL ILMIAH MATEMATIKA, SAINS DAN TEKNOLOGI



Homepage	:	http://ejurnal.ung.ac.id/index.php/euler/index
Journal Abbreviation	:	Euler J. Ilm. Mat. Sains dan Teknol.
Frequency	:	Three times a year
Publication Language	:	English (preferable), Indonesia
DOI	:	https://doi.org/10.37905/euler
Online ISSN	:	2776-3706
License	:	Creative Commons Attribution-NonCommercial 4.0 International License
Publisher	:	Department of Mathematics, Universitas Negeri Gorontalo
Country	:	Indonesia
OAI Address	:	http://ejurnal.ung.ac.id/index.php/euler/oai
Google Scholar ID	:	QF_r-_gAAAAJ
Email	:	euler@ung.ac.id

JAMBURA JOURNAL • FIND OUR OTHER JOURNALS

Jambura Journal of Biomathematics



Jambura Journal of Mathematics

Jambura Journal of Mathematics Education

Jambura Journal of Probability and Statistics

Euler : Jurnal Ilmiah Matematika, Sains dan Teknologi, Volume 13, Issue 2, Pages 147–156, Aug. 2025 *https://doi.org/10.37905/euler.v13i2.31680*

Research Article

Analisis Regresi Logistik Biner dan *Random Forest* untuk Prediksi Faktor-Faktor *Stunting* di Pulau Jawa

Rizqi Dwi Yuniarsyih R.A[1], **Rizqi Annafi Muhadi**[1,*], **Anwar Fitrianto**[1], **Pika Silvianti**[1]

[1]*Program Studi Statistika dan Sains Data, IPB University, Bogor 16680, Indonesia*

ARTICLE HISTORY
Diterima 17 April 2025
Direvisi 21 Juni 2025
Disetujui 26 Juni 2025
Diterbitkan 1 Juli 2025

KATA KUNCI
Stunting
Regresi Logistik Biner
Random Forest

KEYWORDS
Stunting
Binary Logistic Regression
Random Forest

ABSTRAK. *Penelitian ini bertujuan untuk membandingkan performa dan kemampuan identifikasi variabel penting antara model Regresi Logistik Biner dan Random Forest dalam analisis klasifikasi. Berdasarkan hasil analisis, kedua metode menunjukkan kesamaan dalam mengidentifikasi variabel X1, X3, dan X4 sebagai faktor yang paling berpengaruh terhadap hasil prediksi. Namun, Regresi Logistik Biner juga mengidentifikasi variabel X6 sebagai signifikan secara statistik, yang tidak ditunjukkan oleh Random Forest. Dari segi performa model, Random Forest menunjukkan hasil yang lebih unggul pada seluruh metrik evaluasi seperti akurasi, presisi, sensitivitas, spesifisitas, dan balanced accuracy. Hasil ini menunjukkan bahwa Random Forest lebih andal dalam menangani kompleksitas data dan memberikan hasil klasifikasi yang optimal, sedangkan Regresi Logistik Biner unggul dalam memberikan interpretasi yang lebih mendalam terhadap hubungan antar variabel. Oleh karena itu, pemilihan metode sebaiknya disesuaikan dengan tujuan analisis, di mana kombinasi kedua pendekatan dapat memberikan hasil yang lebih komprehensif.*

ABSTRACT. *This study aimed to compare the performance and variable identification capabilities of Binary Logistic Regression and Random Forest models in classification analysis. The results showed that both methods consistently identified variables X1, X3, and X4 as the most influential factors in predicting outcomes. However, Binary Logistic Regression also identified variable X6 as statistically significant, which was not reflected in the Random Forest model. In terms of model performance, Random Forest outperformed Binary Logistic Regression across all evaluation metrics, including accuracy, precision, sensitivity, specificity, and balanced accuracy. These findings suggested that Random Forest was more effective in handling complex data structures and delivering optimal classification results, while Binary Logistic Regression excelled in providing deeper interpretability of variable relationships. Therefore, the choice of method should have aligned with the analytical objectives, and combining both approaches could have yielded more comprehensive insights.*

1. Pendahuluan

Statistika merupakan alat penting dalam berbagai bidang, termasuk kesehatan, karena kemampuannya dalam menganalisis data, memahami hubungan antar variabel, serta mendukung pengambilan keputusan berbasis bukti [1]. Salah satu metode statistik yang banyak digunakan dalam analisis prediktif untuk variabel kategorik adalah regresi logistik biner. Metode ini sederhana, mudah diinterpretasikan, dan efektif dalam mengidentifikasi pengaruh variabel independen terhadap probabilitas suatu peristiwa [2].

Namun, dengan semakin kompleksnya data, pendekatan tradisional seperti regresi logistik mulai dilengkapi oleh metode *machine learning*. Salah satu metode yang banyak digunakan adalah *Random Forest*, yang mampu menghasilkan prediksi yang lebih akurat dan stabil melalui pendekatan *ensemble learning* [3]. Dibandingkan regresi logistik, *Random Forest* unggul dalam menangani data berdimensi tinggi dan mengurangi risiko *overfitting* [4]. Beberapa studi menunjukkan bahwa algoritma ini lebih unggul dalam memprediksi status kesehatan, termasuk *stunting* [5].

Stunting adalah kondisi gagal tumbuh akibat kekurangan gizi kronis yang berdampak jangka panjang terhadap kesehatan dan produktivitas manusia [6]. Di Indonesia, prevalensi *stunting* masih menjadi perhatian utama, dengan angka nasional sebesar 21,60% pada tahun 2022 menurut SSGI [7]. Pulau Jawa sebagai wilayah berpenduduk terbesar turut menyumbang angka prevalensi tinggi, meskipun memiliki fasilitas kesehatan yang relatif memadai. Laporan Indeks Khusus Penanganan *Stunting* (IKPS) 2022 juga menempatkan provinsi di Pulau Jawa seperti Jawa Barat, Jawa Tengah, dan Jawa Timur sebagai wilayah prioritas [8].

Penelitian sebelumnya menunjukkan bahwa pendekatan statistik dan *machine learning* dapat digunakan untuk memetakan faktor risiko dan mengembangkan strategi intervensi dalam menangani *stunting*. Studi oleh Wicaksono dan Harsanti [9] serta Sari dan Syafiq [10] mengidentifikasi variabel-variabel sosial ekonomi seperti IPM, kemiskinan, dan sanitasi sebagai faktor signifikan dalam penurunan *stunting*. Sementara itu, pendekatan *Random Forest* terbukti efektif dalam klasifikasi status *stunting* berdasark-

*Penulis Korespondensi.

Email : *rizqiannafi@apps.ipb.ac.id* (R. A. Muhadi)
Homepage : http://ejurnal.ung.ac.id/index.php/euler/index / E-ISSN : 2776-3706

an data sosiodemografis [11].

Penelitian ini bertujuan untuk membandingkan kinerja Regresi Logistik Biner dan *Random Forest* dalam memprediksi risiko *stunting* di tingkat kabupaten/kota di Pulau Jawa. Regresi Logistik Biner dipilih karena merupakan metode statistik klasik yang umum digunakan untuk analisis klasifikasi dua kelas, serta memberikan interpretasi yang jelas terhadap pengaruh variabel prediktor. Sementara itu, *Random Forest* digunakan sebagai representasi metode *machine learning* yang mampu menangani kompleksitas data dan interaksi *non-linier* antar variabel. Perbandingan kedua model ini penting untuk mengetahui pendekatan mana yang lebih efektif dalam konteks data *stunting* di Indonesia. Selain itu, penelitian ini juga bertujuan untuk mengidentifikasi variabel-variabel yang paling berkontribusi terhadap status risiko *stunting*. Hasil yang diperoleh diharapkan dapat menjadi dasar bagi pengambilan kebijakan berbasis data dalam rangka percepatan penurunan *stunting*, terutama di wilayah prioritas nasional.

2. Metode

2.1. Data dan Variabel Penelitian

Penelitian ini menggunakan data sekunder pada tahun 2023 yang diperoleh dari Kementerian Kesehatan RI, Badan Pusat Statistik, dan Badan Pangan Nasional. Banyaknya observasi dalam penelitian ini mencakup 85 Kabupaten dan 34 Kota yang berasal dari 6 Provinsi di Pulau Jawa. Variabel yang digunakan pada penelitian ini terdiri dari variabel respon (Y) dan variabel prediktor (X). Variabel respon biner yang digunakan adalah status *stunting*, yaitu kategorisasi dari prevalensi *stunting* kedalam dua kelas dimana kelas 1 untuk wilayah dengan kasus *stunting* tinggi ($>$20%) dan kelas 0 untuk wilayah dengan kasus *stunting* rendah ($<$20%) [12]. Variabel yang digunakan dalam penelitian ini disajikan pada Tabel 1.

Tabel 1. Variabel penelitian

Variabel	Keterangan	Skala
Y	Risiko Kasus *Stunting*	Ordinal
X1	Indeks Ketahanan Pangan	Rasio
X2	Persentase Rumah Tangga yang Memiliki Sanitasi Layak	Rasio
X3	Persentase Bayi Mendapat Imunisasi Dasar Lengkap	Rasio
X4	Rasio Posyandu terhadap Desa/ Kelurahan	Rasio
X5	Persentase Bayi $<$6 Bulan yang Diberi ASI Eksklusif	Rasio
X6	Persentase Tempat Pengelolaan Makanan yang Memenuhi Standar Kesehatan	Rasio
X7	Harapan Lama Sekolah	Rasio

2.2. Prosedur Penelitian

Penelitian ini menggunakan dua metode analisis yaitu regresi logistik biner dan *random forest*. Tahapan analisis data pada penelitian ini adalah sebagai berikut:

1. Meng*input* data prediksi prevalensi *stunting* dengan total 119 data dan 8 variabel.
2. Tahap *preprocessing* data yaitu melakukan pengecekan data *cleaning* dan mengonversi beberapa variabel sesuai tipe data (transformasi data). Kemudian melihat gambaran umum data dengan statistika deskriptif
3. Melakukan analisis menggunakan regresi logistik biner dengan *Cross-Validation*:
   (a) Menentukan metode validasi model menggunakan *k-fold cross-validation* (10-*fold*), untuk memaksimalkan pemanfaatan seluruh data secara efisien tanpa kehilangan data untuk pengujian.
   (b) Membangun model regresi logistik menggunakan seluruh data.
   (c) Melakukan Estimasi parameter dengan metode *Maximum Likelihood Estimation* (MLE) untuk memperoleh nilai koefisien terbaik.
   (d) Melakukan pengujian parameter dengan uji serentak dan uji parsial untuk melihat pengaruh dari setiap variabel prediktor terhadap variabel respon.
   (e) Melakukan Uji *Goodness of Fit* menggunakan *Hosmer-Lemeshow Test* untuk mengevaluasi sejauh mana model sesuai dengan data aktual.
   (f) Menginterpretasikan nilai koefisien melalui perhitungan *odds ratio*, untuk mengetahui seberapa besar pengaruh masing-masing variabel prediktor terhadap peluang kejadian.
   (g) Mengevaluasi kinerja model regresi logistik menggunakan metrik akurasi dan *balanced accuracy* dari hasil *cross-validation* sebagai indikator ketepatan klasifikasi model.
4. Melakukan analisis menggunakan *random forest* dengan *Cross-Validation*:
   (a) Menentukan metode validasi model menggunakan *k-fold cross-validation* (10-*fold*), dengan memanfaatkan seluruh data secara efisien untuk pelatihan dan evaluasi.
   (b) Membangun model *Random Forest* menggunakan seluruh data.
   (c) Melakukan penyesuaian parameter penting seperti jumlah pohon (*ntree*), untuk meningkatkan akurasi dan menghindari *overfitting*.
   (d) Mengidentifikasi variabel berpengaruh dengan melihat *feature importance*, yaitu kontribusi relatif tiap variabel terhadap hasil klasifikasi, yang ditampilkan dalam bentuk tabel dan visualisasi.
   (e) Mengevaluasi performa prediksi model *Random Forest* dengan membandingkan hasil klasifikasi terhadap label aktual dan menghitung metrik akurasi dan *balanced accuracy*.
5. Membandingkan ketepatan klasifikasi dengan melihat nilai akurasi dan *balanced accuracy* pada masing-masing model untuk mengambil kesimpulan terbaik.

2.3. Regresi Logistik Biner

Regresi logistik digunakan untuk menganalisis hubungan antara variabel respon dan variabel prediktor. Perbedaan utama antara regresi logistik dan regresi linear terletak pada tipe variabel respon yang dianalisis. Regresi logistik digunakan ketika variabel respon bersifat biner atau dikotomik [2]. Sekumpulan variabel prediktor sebanyak p dapat dianalisis menggunakan model regresi logistik berganda, dengan bentuk fungsi logit:

$$g(x) = \ln\left(\frac{\pi(x)}{1-\mu(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p. \quad (1)$$

Model regresi logistik berganda dapat dinyatakan dalam bentuk probabilitas pada pers. (2):

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}. \tag{2}$$

Dalam hal ini, $\pi(x)$ menunjukkan probabilitas bahwa suatu pengamatan termasuk ke dalam kategori Y. Nilai β_0 merupakan *intersep* dari model regresi logistik, sedangkan β_j adalah koefisien untuk prediktor ke-j. Variabel x menggambarkan nilai dari variabel prediktor, dan p adalah jumlah keseluruhan prediktor yang digunakan dalam model.

2.4. Pengujian Parameter

Pengujian parameter dilakukan untuk menilai hubungan antara variabel prediktor dan variabel respon. Pengujian ini bisa dilakukan secara bersamaan untuk keseluruhan model atau secara parsial untuk masing-masing variabel prediktor.

2.4.1. Uji Simultan

Uji simultan bertujuan menilai apakah seluruh variabel prediktor secara bersama-sama memiliki pengaruh signifikan terhadap variabel respon pada tingkat signifikansi tertentu (taraf alpha). Tujuan uji ini mengetahui apakah model secara keseluruhan sudah layak. Dalam pengujian ini, digunakan statistik G sebagai alat ukur dengan rumus diformulasikan pada pers. (3):

$$G = -2ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\widehat{\pi}_i^{y_i}(1-\widehat{\pi}_i)^{1-y_i}}\right], \tag{3}$$

dimana $n_1 = \sum_{i=1}^{n} y_i$ adalah total observasi dan $n_0 = \sum_{i=1}^{n}(1-y_i)$ adalah probabilitas dari model regresi logistik untuk observasi.

2.4.2. Uji Parsial

Uji parsial bertujuan untuk mengevaluasi pengaruh masing-masing variabel prediktor secara individual terhadap variabel respon. Melalui uji ini, dapat diketahui apakah suatu variabel prediktor secara signifikan mempengaruhi respon. Uji statistik yang digunakan adalah statistik *Wald*, yang dinyatakan pada pers. (4):

$$W_j = \frac{\widehat{\beta}_j}{\sqrt{var(\widehat{\beta}_j)}}. \tag{4}$$

2.4.3. Uji Kelayakan Model

Uji kelayakan model dilakukan untuk menilai sejauh mana model regresi logistik cocok atau layak digunakan dalam menggambarkan data yang diamati. Digunakan statistik *Hosmer-Lemeshow*, yang dirumuskan pada pers. (5):

$$\hat{C} = \sum_{k=1}^{g}\frac{\left(O_k - n_k^{'}\overline{\pi}_k\right)^2}{n_k^{'}\overline{\pi}_k\left(1-\overline{\pi}_k\right)}. \tag{5}$$

2.5. Random Forest

Random Forest adalah metode klasifikasi berbasis *ensemble* yang terdiri dari sejumlah pohon keputusan (*decision tree*) yang dibentuk secara acak dari data pelatihan. Pada setiap pohon, pemilihan fitur dilakukan secara acak, dan hasil prediksi ditentukan melalui mekanisme voting mayoritas dari seluruh pohon. Metode ini mengandalkan dua parameter utama: jumlah pohon yang dibangun dan jumlah fitur yang dipilih secara acak pada setiap pemisahan. Semakin banyak pohon yang dibentuk, semakin stabil dan akurat prediksinya. Algoritma *random forest* berkaitan erat dengan *decision tree*, karena pada dasarnya *random forest* terdiri dari kumpulan *decision tree* yang digabungkan melalui proses voting mayoritas. *Decision tree* merupakan metode yang menggunakan pemisahan (*splitting*) bersyarat secara bertingkat untuk menentukan pemisahan terbaik berdasarkan nilai *entropy*, yaitu ukuran ketidakpastian dalam suatu kondisi yang dinyatakan pada pers. (6). Tujuannya adalah memaksimalkan nilai *information gain* yang dirumuskan pada pers. (7).

$$\text{Entropy} = \sum p_i \log(p_i), \tag{6}$$

$$\text{IG} = E(\text{parent}) = \sum w_i E(\text{child}_i). \tag{7}$$

Dalam rumus *entropy*, *pi* merepresentasikan probabilitas kemunculan kelas ke-*i*. Sedangkan *information gain (IG)* diperoleh dari selisih antara *entropy* dari node induk (*parent*) dan *entropy* gabungan dari node anak (*child*) yang telah diberi bobot. Pada algoritma *random forest*, sejumlah *decision tree* dibentuk, dan hasil prediksi dari masing-masing pohon akan digabungkan. Kelas yang paling sering diprediksi (mayoritas) kemudian dijadikan hasil akhir dari algoritma *random forest*.

Salah satu keunggulan utama dari *Random Forest* adalah kemampuannya dalam mengukur *feature importance*, yaitu kontribusi masing-masing variabel terhadap akurasi prediksi model. *Feature importance* dihitung berdasarkan seberapa besar suatu fitur menurunkan nilai *entropy* secara keseluruhan pada semua pohon dalam model. Variabel dengan nilai *importance* tinggi dianggap memiliki pengaruh besar dalam menentukan hasil prediksi dan dapat digunakan untuk mengidentifikasi faktor-faktor kunci dalam klasifikasi [13].

2.6. Evaluasi Kinerja Model

Prediksi dalam model klasifikasi tidak selalu sempurna, sehingga evaluasi kinerja diperlukan. Salah satu metode evaluasi yang umum digunakan adalah *confusion matrix*, yaitu tabulasi antara kelas aktual dan kelas prediksi. Untuk data dua kelas, evaluasi mencakup empat komponen: *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN), digunakan untuk menilai akurasi dan ketepatan klasifikasi [14]. Beberapa metrik evaluasi kinerja model klasifikasi meliputi akurasi, presisi, *recall*, spesifisitas, dan *balanced accuracy* [14, 15].

2.6.1. Akurasi

Akurasi mengukur proporsi total prediksi yang benar dibandingkan dengan keseluruhan data, diperoleh dengan menggunakan pers. (8):

$$\text{Akurasi} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{FP} + \text{FN} + \text{TP}}. \tag{8}$$

2.6.2. Presisi

Presisi mengukur proporsi prediksi positif yang benar, diperoleh dengan menggunakan pers. (9):

$$\text{Presisi} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{9}$$

2.6.3. *Sensitivitas (Recall)*

Sensitivitas mengukur seberapa baik model mengidentifikasi kejadian positif dari seluruh kasus yang sebenarnya positif, diperoleh dengan menggunakan pers. (10)

$$\text{Sensitivitas} = \frac{\text{TP}}{\text{FN} + \text{TP}}. \tag{10}$$

2.6.4. Spesifisitas

Spesifisitas mengukur seberapa baik model mengidentifikasi kejadian negatif dari seluruh kasus yang sebenarnya negatif, diperoleh dengan menggunakan pers. (11)

$$\text{Spesifisitas} = \frac{\text{TN}}{\text{FN} + \text{FP}}. \tag{11}$$

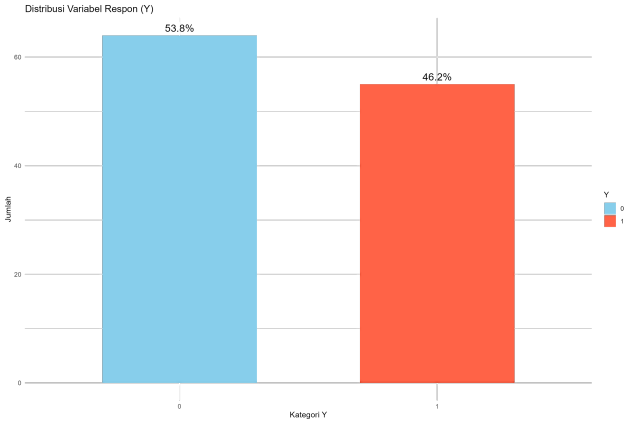
2.6.5. *Balanced Accuracy*

Balanced Accuracy mengukur rata-rata antara sensitivitas dan spesifisitas.

3. Hasil dan Pembahasan

3.1. Eksplorasi Data

Eksplorasi data merupakan tahap awal untuk memahami karakteristik dataset sebelum pemodelan dilakukan [16]. Proses ini mencakup pemeriksaan struktur data, distribusi nilai, hubungan antar variabel, serta deteksi *outlier* dan *missing value*. Dataset terdiri dari 119 observasi dengan tujuh variabel prediktor dan satu variabel respon biner, yaitu status *stunting*. Variabel ini diklasifikasikan ke dalam dua kelas: kelas 1 untuk wilayah dengan risiko *stunting* tinggi (>20%) dan kelas 0 untuk wilayah dengan risiko rendah (<20%) [12]. Proporsi masing-masing kelas ditampilkan pada Gambar 1.



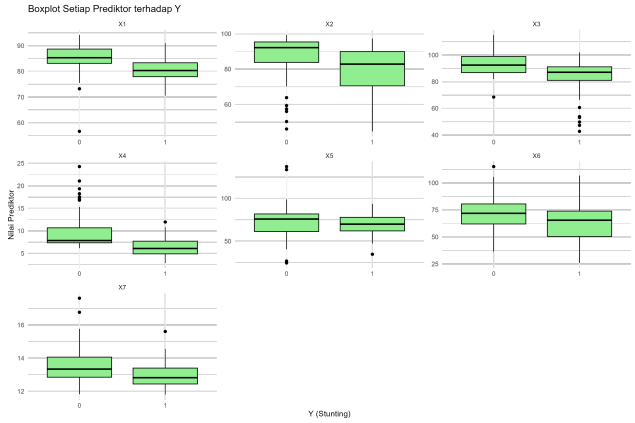
Gambar 1. Distribusi variabel respon

Gambar 1 menunjukkan bahwa pada variabel respon terdiri dari 53,80% kelompok 0 (64 observasi) dan 46,20% kelompok 1 (55 observasi). Selisih antara kedua kelas tidak terlalu besar, menunjukkan bahwa data relatif seimbang secara proporsi kelas. Selanjutnya dilakukan ekspolarasi data untuk setiap variabel prediktor. Hasil ekspolarasi data untuk setiap variabel prediktor dapat dilihat pada pada Tabel 2.

Tabel 2. Statistik deskriptif variabel prediktor

Variabel	*Minimum*	*Median*	*Mean*	*Maximum*	*Standar Deviasi*
X1	56,63	83,57	82,93	94,20	5,71
X2	44,76	87,83	83,57	99,29	13,60
X3	42,81	88,44	88,19	115,10	12,09
X4	2,874	7,439	8,297	24,268	3,71
X5	24,50	73,20	70,99	137,20	16,18
X6	25,90	69,10	68,35	115,16	18,30
X7	11,80	13,17	13,29	17,62	1,03

Berdasarkan Tabel 2, diketahui bahwa sebagian besar variabel prediktor memiliki nilai rata-rata dan median yang relatif berdekatan, menunjukkan distribusi yang cenderung simetris. Namun, terdapat beberapa variabel dengan standar deviasi tinggi dan selisih antara nilai maksimum dan minimum yang besar, yang mengindikasikan adanya sebaran yang luas atau potensi outlier. Selanjutnya, hubungan antara masing-masing variabel prediktor dengan variabel respon akan ditelaah lebih lanjut dengan visualisasi menggunakan *boxplot*. Hal untuk menilai potensi kontribusinya dalam membentuk model prediktif. Visualisasi *boxplot* hubungan variabel prediktor dengan variabel respon dapat dilihat pada Gambar 2.

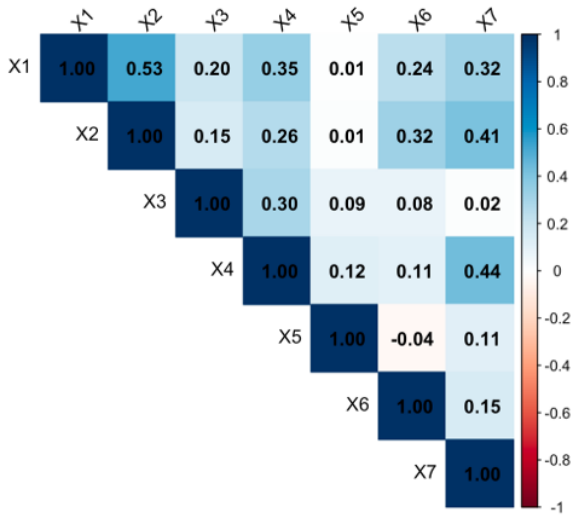


Gambar 2. *Boxplot* hubungan variabel prediktor dengan variabel respon

Gambar 2 menunjukkan bahwa pada variabel *X1*, *X2*, *X3*, *X4*, dan *X6* memiliki nilai *median* lebih rendah pada kelompok risiko *stunting* dibandingkan dengan kelompok tidak *stunting*. Hal ini berarti bahwa semakin rendah nilai setiap variabel prediktor, kemungkinan risiko *stunting* cenderung meningkat yang artinya kelima variabel tersebut berkontribusi negatif terhadap kejadian risiko *stunting*. Sedangkan variabel *X5* dan *X7* menunjukkan distribusi nilai yang cukup mirip antara kelompok *stunting* dan tidak *stunting*. Tidak terdapat perbedaan median yang mencolok, sehingga kemungkinan pengaruh kedua variabel ini terhadap *stunting* relatif kecil atau tidak signifikan.

Analisis korelasi dilakukan untuk memeriksa korelasi antar variabel *independent*. Tujuannya adalah untuk memastikan bahwa tidak terjadi multikolinearitas dalam pemodelan regresi logistik biner. Multikolinearitas adalah kejadian ketika dua atau lebih va-

riabel prediktor dalam model sangat berkorelasi, sehingga menyulitkan model untuk mengestimasi koefisien secara akurat [17]. Deteksi awal terhadap multikolinearitas penting untuk menjaga validitas model regresi dan interpretasi hasil analisis [18]. Hasil *plot* korelasi antar variabel prediktor dapat dilihat pada Gambar 3.



Gambar 3. Matriks korelasi variabel prediktor

Gambar 3 menampilkan matriks korelasi antar variabel prediktor, di mana warna biru lebih gelap menunjukkan korelasi positif yang kuat, sedangkan warna lebih terang menunjukkan korelasi yang lebih lemah. Korelasi tertinggi terdapat antara *X1* dan *X2* $(r = 0,53)$ sedangkan korelasi antara variabel lainnya relatif lebih rendah karena nilainya $< 0,5$. Meskipun korelasi antar variabel prediktor tidak menunjukkan hubungan yang sangat kuat, penting untuk menghitung *Variance Inflation Factor* (VIF) untuk memastikan bahwa tidak ada multikolinearitas yang merugikan dalam model regresi logistik. Hal ini karena plot korelasi hanya menunjukkan hubungan dua variabel secara langsung, namun tidak mendeteksi multikolinearitas kompleks antar beberapa prediktor.

VIF adalah ukuran statistik yang digunakan untuk mendeteksi multikolinearitas antar variabel prediktor dalam model regresi. Perhitungan VIF lebih komprehensif karena menguji pengaruh gabungan dari semua prediktor terhadap masing-masing prediktor lainnya, sehingga lebih andal dalam mendeteksi potensi multikolinearitas dalam model regresi. Sebuah variabel dikatakan memiliki multikolinearitas tinggi apabila nilai VIF-nya melebihi 10, karena ini menunjukkan bahwa varians estimasi koefisiennya meningkat drastis akibat korelasi dengan variabel lain [17]. Nilai VIF dari tiap variabel prediktor dapat dilihat pada Tabel 3.

Tabel 3. Nilai VIF model penuh

Variabel	*X1*	*X2*	*X3*	*X4*	*X5*	*X6*	*X7*
VIF	1,29	1,60	1,17	1,19	1,05	1,11	1,53

Nilai VIF pada Tabel 3 menunjukkan bahwa tidak terdapat satupun variabel yang memiliki nilai VIF lebih dari 10. Hal ini menunjukkan bahwa tidak terdapat indikasi multikolinearitas yang tinggi antar variabel prediktor. Dengan demikian, model bebas dari masalah multikolinearitas yang kuat antar variabel bebas. Ketiadaan multikolinearitas ini juga memastikan bahwa estimasi koefisien regresi yang dihasilkan akan stabil dan dapat diinterpretasikan secara akurat sehingga pemodelan dapat dilanjutkan dengan analisis regresi logistik biner secara valid.

3.2. *Regresi Logistik Biner*

Regresi logistik biner merupakan metode analisis yang digunakan untuk memprediksi probabilitas dari suatu kejadian yang memiliki dua kemungkinan hasil, seperti kejadian *stunting* atau tidak *stunting*. Model ini menghubungkan variabel dependen kategorik dengan satu atau lebih variabel independen melalui fungsi logit [2].

Dalam regresi logistik biner, salah satu asumsi penting yang perlu dipenuhi adalah adanya hubungan linear antara logit (*log-odds*) dari probabilitas kejadian dan variabel kontinu [2]. Pelanggaran terhadap asumsi ini dapat menghasilkan estimasi parameter yang bias serta menurunkan akurasi prediksi model [19]. Untuk menguji asumsi ini, metode *Box-Tidwell test* dapat digunakan, di mana interaksi antara masing-masing variabel prediktor dan logaritma alaminya dimasukkan ke dalam model. Jika koefisien interaksi tersebut signifikan secara statistik (p-*value* $< 0,05$), maka terdapat bukti bahwa hubungan antara prediktor dan logit tidak bersifat linear, dan asumsi linearitas tidak terpenuhi [20]. Hasil *Box-Tidwell test* dapat dilihat pada Tabel 4.

Tabel 4. Hasil *Box-Tidwell test*

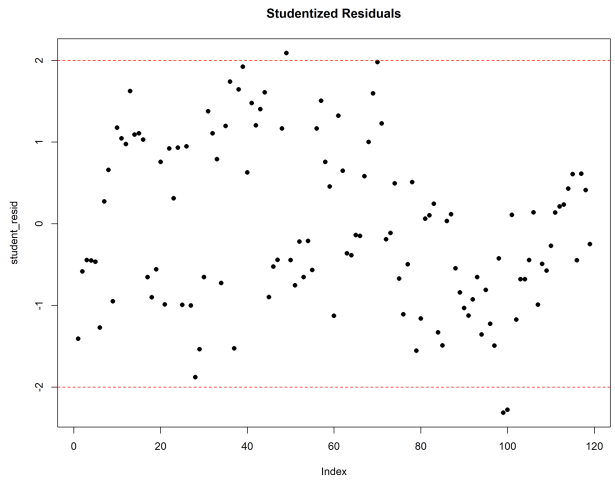
Variabel	*p-value*
X1	0,2098
X2	0,3515
X3	0,5589
X4	0,0784
X5	0,3803
X6	0,1213
X7	0,6722

Tabel 4 menunjukkan bahwa seluruh variabel memiliki nilai p-value $> 0,05$. Artinya, tidak cukup bukti untuk menyatakan bahwa hubungan antara masing-masing variabel prediktor dan logit tidak linier. Dengan kata lain, hasil ini menunjukkan bahwa asumsi linearitas antara *logit* dan seluruh prediktor numerik telah terpenuhi, sehingga model regresi logistik yang dibangun dapat dianggap sesuai secara statistik.

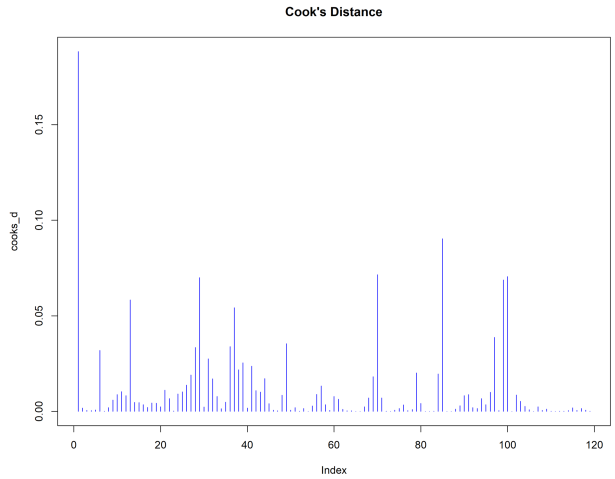
Setelah asumsi linearitas diperiksa, dilakukan analisis residual untuk mengevaluasi kecocokan model dan mendeteksi observasi yang mungkin menyimpang secara signifikan atau memiliki pengaruh besar terhadap hasil estimasi. Beberapa metrik residual yang digunakan antara lain *studentized residual*, *Cook's distance*, dan residual terhadap nilai prediksi [2].

Gambar 4 menunjukkan bahwa sejumlah observasi memiliki nilai *studentized residual* melebihi ambang ± 2, yang secara statistik dapat diindikasikan sebagai pencilan. Namun, berdasarkan Gambar 5, nilai *Cook's distance* untuk seluruh observasi tercatat di bawah ambang batas 1, yang menunjukkan bahwa tidak terdapat pengamatan yang memiliki pengaruh besar terhadap estimasi parameter model. Artinya, meskipun terdapat pencilan, tidak ada satupun observasi yang dikategorikan sebagai pengamatan berpengaruh secara statistik (*influential points*) [16].

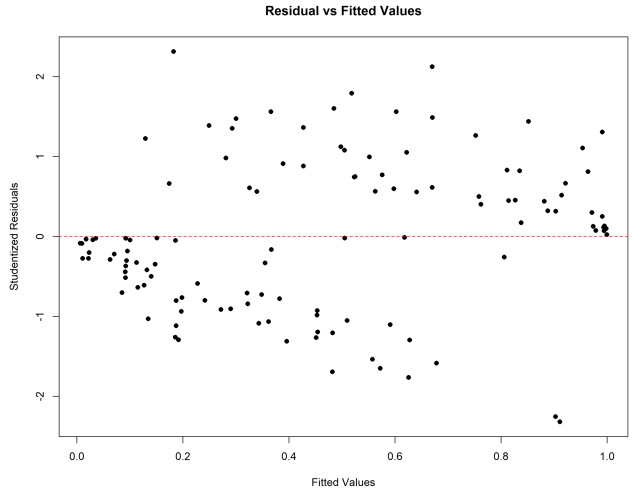
Gambar 6 menunjukkan visualisasi antara residual terstandarisasi dan nilai prediksi memiliki pola sebaran yang acak di se-



Gambar 4. *Plot Studentized Residual*



Gambar 5. *Plot cook's distance*



Gambar 6. Plot residual terhadap nilai prediksi

kitar garis nol tanpa kecenderungan sistematis. Hal ini mengindikasikan tidak adanya pelanggaran spesifikasi model, serta mendukung bahwa struktur model telah mewakili hubungan yang ada dalam data dengan cukup baik [19]. Berdasarkan hasil pemeriksaan asumsi dan analisis residual, model regresi logistik yang dikembangkan dinilai telah memenuhi syarat dasar pemodelan yang baik. Oleh karena itu, model dinilai layak untuk digunakan dalam analisis inferensial selanjutnya. Hasil analisis regresi logistik biner dapat dilihat pada Tabel 5.

Tabel 5. Hasil analisis regresi logistik biner

Variabel	Estimasi Koefisien	*Z-hitung*	*p-value*
Intercept	27,908	3,690	0,000
X1	-0,106	-2,023	0,043
X2	-0,007	-0,323	0,746
X3	-0,104	-2,932	0,003
X4	-0,517	-3,243	0,001
X5	-0,011	-0,683	0,494
X6	-0,032	-2,161	0,031
X7	-0,191	-0,513	0,608

Hasil analisis regresi logistik biner Tabel 5 memberikan informasi estimasi koefisien, nilai *Z*-hitung, dan signifikansi (*p-value*) untuk masing-masing variabel independen. Estimasi koefisien menunjukkan arah dan besarnya pengaruh suatu variabel terhadap kemungkinan terjadinya risiko *stunting*. Nilai *Z*-hitung dan *p-value* digunakan untuk mengevaluasi apakah pengaruh tersebut signifikan secara statistik terhadap variabel dependen. Informasi ini menjadi dasar dalam melakukan uji parameter, baik secara simultan maupun parsial untuk menilai signifikansi model dan kontribusi masing-masing variabel prediktor dalam memengaruhi kejadian *stunting* [2].

3.2.1. Uji Parameter

Uji parameter mencakup dua pendekatan, yaitu uji simultan (uji penuh) dan uji individual (uji parsial). Uji simultan atau uji penuh dilakukan untuk mengetahui apakah model regresi logistik biner secara keseluruhan signifikan dalam memprediksi variabel dependen. Uji ini dilakukan dengan membandingkan model yang memasukkan semua variabel independen (model penuh) dengan model tanpa variabel independen (model nol) menggunakan statistik *likelihood ratio test*. Statistik uji simultan mengikuti distribusi *Chi-square* [2]. Hasil Uji Simultan berdasarkan pers. (3) dapat dilihat pada Tabel 6.

Tabel 6. Hasil uji simultan

χ^2	*df*	*p-value*
65,578	7	1,151 x 10^{-11}

Berdasarkan hasil pada Tabel 6 diperoleh nilai statistika pengujian *chi-square* $G^2\,(65,678) > \chi^2_{(6;0,05)}\,(12,592)$, dengan nilai *p-value* sebesar 0,000 $<$ 0,05 sehingga tolak H_0. Hal ini menunjukkan bahwa model regresi logistik biner secara simultan signifikan, sehingga terdapat setidaknya satu variabel prediktor yang berpengaruh terhadap variabel respon. Dengan hasil ini, analisis dapat dilanjutkan ke tahap uji parsial.

Uji parsial bertujuan untuk mengevaluasi pengaruh masing-masing variabel prediktor (independen) terhadap variabel respon secara individual, dengan melihat apakah koefisien regresi masing-masing variabel signifikan secara statistik. Dalam regresi logistik, uji parsial biasanya dilakukan menggunakan *Wald test* [2]. Uji parsial dilakukan menggunakan nilai *Z*-hitung berdasarkan pers. (4) dan *p-value* yang dihasilkan tiap variabel prediktor pada Tabel 5.

Tabel 5 menunjukkan bahwa terdapatkan variabel yang nilai *p-value*nya $< 0{,}05$ yaitu *X1*, *X3*, *X4*, dan *X6* sehingga tolak H_0. Hal ini berarti variabel tersebut berpengaruh signifikan terhadap variabel respon. Sedangkan variabel lainnya seperti *X2*, *X5*, dan *X7* mempunyai nilai *p-value* $> 0{,}05$ sehingga gagal tolak H_0. Hal ini berarti variabel tersebut tidak berpengaruh signifikan terhadap variabel respon.

3.2.2. *Uji Goodness Of Fit (Kecocokan Model)*

Uji kecocokan model (*goodness of fit*) merupakan uji selanjutnya dalam analisis regresi logistik biner. Uji ini bertujuan untuk mengetahui seberapa baik model yang dibentuk dapat menjelaskan data, atau dengan kata lain, apakah model sesuai (*fit*) dengan data yang diamati. Salah satu cara untuk mengukur ini adalah melalui uji *Hosmer dan Lemeshow* [2]. Hasil uji kecocokan model menggunakan pers. (5) dapat dilihat pada Tabel 7.

Tabel 7. Hasil uji kecocokan model

χ^2	*df*	*p-value*
4,258	7	0,83311

Tabel 7 menunjukkan bahwa hasil pengujian *chi-square* $G^2\,(4,258) < \chi^2_{(6;0,05)}\,(12,592)$, dengan nilai *p-value* sebesar $0{,}8331 > 0{,}05$ sehingga gagal tolak H_0. Hal ini menunjukkan bahwa model regresi logistik biner dianggap sesuai dengan data. Artinya, tidak terdapat perbedaan yang signifikan antara nilai yang diamati dan nilai yang diprediksi oleh model. Oleh karena itu, model dapat digunakan untuk keperluan prediksi atau penarikan kesimpulan lebih lanjut.

3.2.3. Model Logistik Biner

Berdasarkan uji parsial yang dilakukan sebelumnya, terdapat 4 variabel prediktor yang berpengaruh signifikan terhadap variabel respon. Dengan demikian, persamaan regresi logistik sesuai pers. (1) dapat dinyatakan sebagai:

$$\begin{aligned}\text{Logit}\left[P\left(Y=1\right)\right] = 27,908 - 0,106X_1 - 0,104X_3 \\ - 0,517X_4 - 0,032X_6.\end{aligned}$$

Persamaan regresi logistik dapat pula dinyatakan dalam bentuk probabilitas sesuai pers. (2):

$$\pi(x) = \frac{e^{27,908-0,106X_1-0,104X_3-0,517X_4-0,032X_6}}{1+e^{27,908-0,106X_1-0,104X_3-0,517X_4-0,032X_6}}.$$

Dari variabel hasil pemodelan, variabel yang menunjukkan faktor memengaruhi risiko *stunting* adalah variabel signifikan pada taraf nyata 5%. Selanjutnya variabel signifikan tersebut di interpretasikan berdasarkan *odds ratio*. *Odds rasio* adalah suatu peluang yang menggambarkan kemungkinan terjadinya suatu kejadian dibandingkan dengan kemungkinan kejadian tersebut tidak terjadi [21]. *Odds Ratio* dari setiap variabel signifikan dapat dilihat pada Tabel 8.

Nilai odds yang dihasilkan dari model logistik menunjukkan beberapa hubungan penting antara variabel independen dengan kemungkinan terjadinya risiko *stunting*. Variabel *X1* memiliki *odds ratio* 0,899 yang berarti bahwa setiap peningkatan 1 nilai Indeks Ketahanan Pangan akan menurunkan 0,899 kali risiko *stunting* dengan variabel bebas lainnya konstan. Variabel *X3* memiliki *odds ratio* 0,901 yang berarti bahwa setiap peningkatan 1% bayi yang mendapat imunisasi dasar lengkap akan menurunkan 0,901 kali risiko *stunting* dengan variabel bebas lainnya konstan.

Tabel 8. Rasio odds variabel signifikan

Variabel	*Exp(β)*
X1	0,899
X3	0,901
X4	0,596
X6	0,967

Variabel *X4* memiliki *odds ratio* 0,596 yang berarti bahwa setiap peningkatan 1 nilai rasio posyandu terhadap desa/kelurahan akan menurunkan 0,596 kali risiko *stunting* dengan variabel bebas lainnya konstan. Variabel *X6* memiliki *odds ratio* 0,967 yang berarti bahwa setiap peningkatan 1% tempat pengelolaan makanan yang memenuhi standar kesehatan akan menurunkan 0,967 kali risiko *stunting* dengan variabel bebas lainnya konstan.

Model yang telah dibentuk kemudian dievaluasi untuk menilai sejauh mana model mampu melakukan prediksi secara akurat dan andal terhadap variabel respon. Evaluasi ini bertujuan untuk memastikan bahwa model tidak hanya signifikan secara statistik, tetapi juga memiliki kemampuan klasifikasi yang baik dalam praktik, terutama saat diterapkan pada data baru [22, 23]. Evaluasi dilakukan dengan menggunakan beberapa metrik performa, yaitu akurasi sesuai pers. (8), presisi sesuai pers. (9), sensitivitas (*recall*) sesuai pers. (10), spesifisitas sesuai pers. (11), dan *balanced accuracy*. Hasil evaluasi performa model dapat dilihat pada Tabel 9.

Tabel 9. Evaluasi performa model

Metrik	Nilai
Akurasi	0,765
Presisi	0,755
Sensitivitas	0,727
Spesifisitas	0,797
Balanced Accuracy	0,761

Hasil evaluasi pada Tabel 9 menunjukkan bahwa model memiliki akurasi sebesar 76,50%, yang menggambarkan proporsi klasifikasi yang benar secara keseluruhan. Nilai presisi sebesar 75,50% menunjukkan bahwa dari seluruh prediksi positif, sebanyak 75,50% merupakan kasus positif yang benar. Sensitivitas model sebesar 72,70% mencerminkan kemampuannya dalam mendeteksi kasus positif secara tepat, sementara spesifisitas sebesar 79,70% menunjukkan keakuratan model dalam mengenali kasus negatif. Nilai *balanced accuracy* sebesar 76,10%, sebagai rata-rata dari sensitivitas dan spesifisitas, menunjukkan bahwa model memiliki kinerja yang cukup seimbang dalam mengklasifikasikan kedua kategori respon. Berdasarkan hasil evaluasi ini, dapat disimpulkan bahwa model memiliki performa yang cukup baik dan dapat digunakan untuk keperluan prediksi dan pengambilan keputusan, meskipun masih terdapat ruang untuk peningkatan.

3.3. *Random Forest*

Random Forest adalah metode *machine learning* berbasis *ensemble decision tree* yang digunakan untuk klasifikasi maupun re-

gresi. Setiap *decision tree* dalam algoritma ini membagi data berdasarkan atribut yang mampu meminimalkan heterogenitas kelas. Heterogenitas tersebut diukur menggunakan nilai *entropy* sesuai pers. (6), sedangkan efektivitas pemisahan data diukur menggunakan *information gain* sesuai pers. (7).

Salah satu fitur penting dari *Random Forest* adalah kemampuannya untuk menghitung *feature importance*, yaitu ukuran seberapa besar kontribusi masing-masing variabel dalam meningkatkan akurasi model. Dalam penelitian ini, *Random Forest* digunakan untuk mengidentifikasi variabel-variabel yang paling berkontribusi terhadap prediksi dengan mengacu pada nilai feature importance. Variabel dengan nilai *feature importance* yang tinggi dianggap memiliki pengaruh yang lebih besar dalam menentukan hasil klasifikasi, sehingga dapat diprioritaskan dalam interpretasi hasil maupun dalam perancangan kebijakan atau intervensi yang relevan [3]. Hasil *feature importance random forest* dapat dilihat pada Tabel 10.

Tabel 10. Hasil *feature importance random forest*

Variabel	Nilai
X4	100
X1	33,780
X3	18,800
X6	8,807
X2	3,711
X5	1,245
X7	0,000

Hasil analisis *feature importance* menggunakan metode *Random Forest* pada Tabel 10 menunjukkan bahwa variabel *X4* merupakan variabel yang paling berpengaruh terhadap hasil prediksi model. Hal ini terlihat dari nilai *importance*-nya yang paling tinggi dibandingkan variabel lainnya. Disusul oleh variabel *X1* dan *X3*, yang juga memberikan kontribusi signifikan meskipun lebih rendah dari *X4*. Sementara itu, variabel *X6*, *X2*, *X5*, dan *X7* memiliki nilai *importance* yang relatif rendah, menunjukkan kontribusi yang lebih kecil dalam membentuk keputusan model.

Hasil ini menunjukkan bahwa *X4* memiliki peran dominan dalam proses klasifikasi oleh *Random Forest*, sehingga dapat menjadi fokus utama dalam interpretasi maupun pengambilan keputusan berbasis model. Sebaliknya, variabel dengan importance rendah mungkin dapat dipertimbangkan untuk dikeluarkan atau dipantau lebih lanjut jika diperlukan proses seleksi fitur.

Setelah model *Random Forest* dibentuk, dilakukan evaluasi performa untuk menilai sejauh mana model mampu melakukan klasifikasi dengan baik. Tujuan dari evaluasi ini adalah memastikan model tidak hanya menghasilkan hasil yang stabil, tetapi juga andal saat diterapkan pada data baru. Hasil evaluasi performa random forest dapat dilihat pada Tabel 11.

Tabel 11. Evaluasi performa model *random forest*

Metrik	*Mtry = 2*	*Mtry = 4*	*Mtry = 7*
Akurasi	0,781	0,781	0,765
Presisi	0,784	0,774	0,746
Sensitivitas	0,727	0,746	0,746
Spesifisitas	0,828	0,813	0,781
Balanced Accuracy	0,778	0,779	0,763

Berdasarkan hasil evaluasi performa model *Random Forest* dengan variasi parameter *mtry* (jumlah variabel acak yang dipertimbangkan pada setiap split pohon), diperoleh bahwa model dengan $mtry = 2$ dan $mtry = 4$ memberikan hasil yang serupa dan lebih optimal dibandingkan $mtry = 7$. Nilai akurasi pada $mtry = 2$ dan 4 sama, yaitu sebesar 0,781, yang menunjukkan bahwa sekitar 78,1% dari total observasi berhasil diklasifikasikan dengan benar. Dari sisi presisi, model dengan $mtry = 2$ mencatat nilai tertinggi yaitu 0,784, menunjukkan bahwa prediksi positif model paling tepat pada konfigurasi ini. Sementara itu, sensitivitas (*recall*) tertinggi dicapai pada $mtry = 4$ dan 7 sebesar 0,746, menunjukkan kemampuan model dalam menangkap kasus positif dengan baik. Namun, spesifisitas tertinggi masih terdapat pada $mtry = 2$ yaitu 0,828, mengindikasikan model paling akurat dalam mengenali kasus negatif. Untuk metrik *balanced accuracy*, yang merupakan rata-rata dari sensitivitas dan spesifisitas, nilai terbaik juga ditunjukkan oleh $mtry = 4$ (0,779), sedikit lebih tinggi dari $mtry = 2$ (0,778). Model dengan $mtry = 7$ memiliki performa paling rendah di semua metrik. Secara keseluruhan, $mtry = 2$ dan $mtry = 4$ dapat dianggap sebagai pilihan parameter terbaik, dengan keseimbangan performa yang baik antara presisi, *sensitivitas*, dan spesifisitas.

3.4. Perbandingan Regresi Logistik Biner dan Random Forest

3.4.1. Identifikasi Faktor Berpengaruh

Berdasarkan hasil analisis, terdapat kesamaan dan perbedaan dalam identifikasi variabel penting antara model Regresi Logistik Biner dan *Random Forest*. Regresi Logistik Biner menunjukkan bahwa variabel *X1, X3, X4*, dan *X6* berpengaruh secara signifikan terhadap variabel respon, berdasarkan nilai koefisien dan uji signifikansi statistik. Sementara itu, *Random Forest* melalui analisis *feature importance* mengidentifikasi variabel *X4*, *X1*, dan *X3* sebagai variabel yang paling berkontribusi terhadap proses klasifikasi.

Dari hasil ini, dapat disimpulkan bahwa kedua metode sepakat terhadap pentingnya variabel *X1* (Indeks Ketahanan Pangan), *X3* (Persentase Bayi Mendapat Imunisasi Dasar Lengkap), dan *X4* (Rasio Posyandu terhadap Desa/ Kelurahan) yang memperkuat keandalan ketiga variabel tersebut dalam mempengaruhi hasil prediksi. Namun, perbedaan muncul pada variabel *X6* (Persentase Tempat Pengelolaan Makanan yang Memenuhi Standar Kesehatan), yang dianggap signifikan oleh regresi logistik tetapi tidak termasuk dalam variabel paling penting menurut *Random Forest*. Hal ini bisa terjadi karena regresi logistik mempertimbangkan hubungan linier dan signifikansi statistik [24], sedangkan *Random Forest* lebih fokus pada kontribusi terhadap performa model secara keseluruhan, termasuk hubungan *non-linier* atau interaksi antar variabel [25]. Dengan demikian, kedua pendekatan dapat saling melengkapi. *Regresi logistik* mempunyai keunggulan dalam kemudahan interpretasi dimana pengguna dapat mengetahui arah (positif atau negatif) serta kekuatan hubungan antara variabel prediktor terhadap variabel respon [24].

Temuan ini dapat menjadi dasar rekomendasi kebijakan. Peningkatan ketahanan pangan dapat dilakukan melalui lumbung pangan desa dan edukasi gizi. Cakupan imunisasi perlu diperluas lewat optimalisasi layanan posyandu. Peningkatan jumlah dan kualitas posyandu penting untuk memperluas akses layanan dasar. Terakhir, pembinaan tempat pengolahan makanan mendukung pencegahan penyakit yang berdampak pada pertumbuhan anak. Dengan mengacu pada hasil kedua model, strategi penu-

runan *stunting* dapat dirancang secara lebih efektif, berbasis data, dan lintas sektor.

3.4.2. Hasil Evaluasi Performa

Hasil perbandingan performa antara model regresi logistik biner dan algoritma *Random Forest* berdasarkan beberapa metrik evaluasi seperti akurasi, presisi, sensitivitas, spesifisitas, dan *balanced accuracy* disajikan pada Tabel 12.

Tabel 12. Perbandingan evaluasi performa

Metrik	Logistik Biner	*Random Forest*
Akurasi	0,765	0,781
Presisi	0,755	0,784
Sensitivitas	0,727	0,727
Spesifisitas	0,797	0,828
Balanced Accuracy	0,761	0,778

Berdasarkan perbandingan evaluasi performa pada Tabel 12, model *Random Forest* dan regresi logistik biner menunjukkan hasil yang cukup kompetitif namun dengan karakteristik performa yang berbeda. Model *Random Forest* memiliki nilai akurasi sebesar 0,781, sedikit lebih tinggi dibandingkan regresi logistik biner yang bernilai 0,765. Nilai presisi *Random Forest* juga mencapai 0,784, yang menunjukkan bahwa model cukup baik dalam menghasilkan prediksi positif yang tepat dibandingkan regresi logistik biner yang bernilai 0,755. Sementara itu, sensitivitas *Random Forest* berada pada angka 0,746, sedikit lebih tinggi dibandingkan regresi logistik dengan nilai 0,727, yang mengindikasikan kemampuannya lebih baik dalam menangkap kasus positif. Dari sisi spesifisitas, *Random Forest* mencatat nilai 0,828, yang menunjukkan performa kuat dalam mendeteksi kasus negatif dibandingkan regresi logistik biner yang bernilai 0,797. Nilai *balanced accuracy* yang dihasilkan *random forest* sebesar 0,778 sedangkan regresi logistik biner sebesar 0,761, ini mengindikasikan bahwa keseimbangan yang baik antara sensitivitas dan spesifisitas *random forest* lebih baik dibandingkan regresi logistik biner.

Jika dibandingkan, *Random Forest* menunjukkan keunggulan pada seluruh metrik evaluasi, menandakan bahwa model ini lebih fleksibel dan mampu menangkap kompleksitas data dibandingkan regresi logistik yang bersifat linier. Temuan ini mengindikasikan bahwa *Random Forest* lebih optimal dalam mengklasifikasikan risiko *stunting*. Oleh karena itu, hasil dari model ini dapat dimanfaatkan untuk mengidentifikasi wilayah dan faktor prioritas yang perlu mendapatkan intervensi lebih lanjut dalam program percepatan penurunan *stunting* di Pulau Jawa.

4. Kesimpulan

Berdasarkan hasil analisis, Regresi Logistik Biner dan *Random Forest* memiliki keunggulan masing-masing dalam identifikasi faktor dan performa prediksi. Kedua model secara konsisten mengidentifikasi *X1* (Indeks Ketahanan Pangan), *X3* (Persentase Bayi Mendapat Imunisasi Dasar Lengkap), dan *X4* (Rasio Posyandu terhadap Desa/ Kelurahan) sebagai faktor penting dalam menentukan risiko *stunting*, dengan Regresi Logistik juga menyoroti *X6* (Persentase Tempat Pengelolaan Makanan yang Memenuhi Standar Kesehatan) sebagai variabel signifikan secara statistik. *Random Forest* menunjukkan keunggulan pada hampir semua metrik evaluasi, menandakan kemampuannya menangkap pola *non-linier* dan kompleksitas data secara lebih baik. Hasil *feature importance* dari *Random Forest* menempatkan *X4* sebagai variabel paling dominan, diikuti oleh *X1* dan *X3* yang mengindikasikan bahwa aspek layanan kesehatan dasar, akses pangan bergizi, dan cakupan imunisasi perlu menjadi prioritas dalam perumusan kebijakan penurunan *stunting* di Pulau Jawa. Oleh karena itu, saran kebijakan yang dapat diajukan meliputi: peningkatan jumlah dan kualitas posyandu di desa-desa yang masih tertinggal, penguatan program ketahanan pangan lokal untuk menjamin ketersediaan makanan bergizi bagi rumah tangga miskin, serta optimalisasi program imunisasi dasar melalui layanan posyandu dan edukasi masyarakat. Dengan demikian, kedua metode dapat digunakan secara komplementer: *Random Forest* untuk prediksi akurat dan Regresi Logistik untuk interpretasi arah pengaruh, guna mendukung kebijakan penurunan *stunting* yang lebih efektif dan berbasis data.

Kontribusi Penulis. **Rizqi Dwi Yuniarsyih R.A:** Konseptualisasi, metodologi, perangkat lunak, validasi, analisis formal, investigasi, sumber daya, kurasi data, penulisan–persiapan draf asli, visualisasi. **Rizqi Annafi Muhadi:** Konseptualisasi, metodologi, perangkat lunak, validasi, investigasi, penulisan–persiapan draf asli. **Anwar Fitrianto:** penulisan–tinjauan dan penyuntingan, supervisi. **Pika Silvianti:** penulisan–tinjauan dan penyuntingan, supervisi. Semua penulis telah membaca dan menyetujui versi manuskrip yang diterbitkan.

Ucapan Terima Kasih. Para penulis mengucapkan terima kasih kepada semua pihak yang telah berkontribusi dalam penelitian ini dan dalam penyusunan manuskrip. Kami sangat menghargai editor dan reviewer atas masukan serta dukungannya dalam menyempurnakan karya ini.

Pembiayaan. Penelitian ini tidak menerima pendanaan eksternal.

Konflik Kepentingan. Para penulis menyatakan tidak ada konflik kepentingan yang terkait dengan artikel ini.

Referensi

[1] D. A. Freedman, R. Pisani, and R. A. Purves, *Statistics*, 4th ed. New York, NY: W. W. Norton & Company, 2007.

[2] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ: Wiley, 2013, doi: 10.1002/9781118548387.

[3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY: Springer, 2009, doi: 10.1007/978-0-387-84858-7.

[5] R. Alfonsus and A. Rofiq, "Predictive analysis of stunting using random forest algorithm," *J. Teknol. dan Sist. Komput.*, vol. 10, no. 3, pp. 421–427, 2022, doi: 10.14710/jtsiskom.2022.421-427.

[6] UNICEF, *Improving Child Nutrition: The achievable imperative for global progress*. United Nations Children's Fund, 2013.

[7] Kementerian Kesehatan Republik Indonesia, *Hasil Survei Status Gizi Indonesia (SSGI) Tahun 2022*. Jakarta: Kemenkes RI, 2022.

[8] Badan Kependudukan dan Keluarga Berencana Nasional (BKKBN), *Laporan Indeks Khusus Penanganan Stunting 2023*. Jakarta: BKKBN, 2023.

[9] F. Wicaksono and T. Harsanti, "Determinants of stunting in Indonesia: A spatial analysis at the district level," *Kesmas: J. Kesehat. Masyarakat Nasional*, vol. 16, no. 2, pp. 72–79, 2021, doi: 10.21109/kesmas.v16i2.4730.

[10] N. Sari and A. Syafiq, "Analisis faktor-faktor sosial ekonomi terhadap stunting di Indonesia," *J. Gizi dan Pembangunan*, vol. 18, no. 1, pp. 10–18, 2023.

[11] D. Rahmawati, A. Firmansyah, and D. Lestari, "Implementasi random forest untuk prediksi stunting pada balita menggunakan data sosial ekonomi," *J. Teknol. Inf. dan Ilmu Komput. (JTIIK)*, vol. 10, no. 2, pp. 180–188, 2023.

[12] Food and Agriculture Organization of the United Nations, *Prevalence thresholds for wasting, overweight and stunting in children under 5 years*, 2019.

[13] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," in *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, 2013, pp. 431–439, doi: 10.48550/arXiv.1311.0456.

[14] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. San Francisco, CA: Morgan Kaufmann, 2012, doi: 10.1016/C2009-0-61819-5.

[15] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, NY: Springer, 2013, doi: 10.1007/978-1-4614-6849-3.

[16] A. Field, *Discovering Statistics Using IBM SPSS Statistics*, 4th ed. London: Sage Publications, 2013. (6th ed., 2024) doi: 10.1177/21677026241240456.

[17] D. N. Gujarati and D. C. Porter, *Basic Econometrics*, 5th ed. New York, NY: McGraw-Hill Education, 2009.

[18] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th ed. Hoboken, NJ: Wiley, 2012, doi: 10.1111/biom.12129.

[19] S. Menard, *Logistic Regression: From Introductory to Advanced Concepts and Applications*, 2nd ed. Thousand Oaks, CA: SAGE Publications, 2010.

[20] M. García, C. Fernández, and M. Rodríguez, "Logistic regression based in-service assessment of mobile web browsing service quality acceptability," *EURASIP J. Wireless Commun. Netw.*, vol. 2020, no. 1, Article 43, 2020, doi: 10.1186/s13638-020-01680-7.

[21] S. Tenny and M. R. Hoffman, "Odds ratio," in *StatPearls*, StatPearls Publishing, 2025.

[22] J. S. Aguilar-Ruiz, "The certainty ratio Cp: A novel metric for assessing the reliability of classifier predictions," *arXiv preprint*, arXiv:2411.01973, 2024, doi: 10.48550/arXiv.2411.01973.

[23] A. Asro, J. Chaidir, Chairuddin, and J. Friadi, "Evaluasi kinerja algoritma klasifikasi dalam studi kasus prediksi kelulusan di Universitas XYZ," *Zona Teknik: Jurnal Ilmiah*, vol. 19, no. 1, pp. 15-22, Feb. 2025, doi: 10.37776/zt.v19i1.1674.

[24] P. Biecek and T. Burzykowski, *Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models*. Boca Raton, FL: CRC Press, 2021.

[25] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017, doi: 10.5555/3295222.3295408.