BAB I
PENDAHULUAN

1.1 Latar Belakang

Klasifikasi merupakan metode yang digunakan untuk memetakan suatu objek ke dalam kelas tertentu berdasarkan sekumpulan atribut atau fitur yang dimilikinya. Metode ini banyak diterapkan di berbagai bidang, seperti kesehatan, keuangan, manufaktur, hingga perdagangan sebagai alat bantu pengambilan keputusan pada permasalahan yang kompleks berbasis data (Maulani dkk., 2023). Beragam metode dapat digunakan untuk tugas klasifikasi, antara lain pendekatan statistika klasik seperti Analisis Diskriman dan Regresi Logistik serta pendekatan *machine learning* seperti *Artificial Neural Network (ANN), Naive Bayes, Classification and Regression Tree* (*CART*) dan *Support Vector Machine (SVM)* (Han dkk., 2012). Salah satu metode statistika klasik yang sering digunakan, selain Analisis Diskriminan adalah Analisis Regresi Logistik.

Regresi logistik digunakan untuk menganalisis hubungan antara satu atau lebih variabel prediktor dengan variabel respon kategorik. Metode ini dibedakan menjadi tiga jenis, yaitu regresi logistik biner yang digunakan ketika variabel respon memiliki dua kategori, regresi logistik multinomial yang digunakan untuk variabel respon dengan lebih dari dua kategori, dan regresi logistik ordinal yang diterapkan ketika variabel respon memiliki urutan peringkat (Hosmer dkk., 2013). Keunggulan utamanya adalah tidak mensyaratkan asumsi normalitas pada prediktor, sehingga model lebih fleksibel untuk berbagai tipe data, termasuk gabungan prediktor numerik dan kategorik. Selain itu, regresi logistik menghasilkan probabilitas kejadian suatu kelas, sehingga hasilnya mudah dipakai dalam pengambilan keputusan berbasis risiko (Bewick dkk., 2005). Meskipun demikian, regresi logistik memiliki keterbatasan apabila salah satu kelas memiliki jumlah sampel lebih besar dari kelas lainnya yang disebut data tidak seimbang.

King dan Zeng (2001) menyatakan bahwa ketika regresi logistik digunakan pada kasus data tidak seimbang, maka *classifier* cenderung menihilkan peluang dari kelompok minoritas karena nilai prediksi akan cenderung kepada kategori data yang mayoritas, sehingga tingkat ketepatan yang dihasilkan menjadi kurang baik. Terdapat banyak kasus permasalahan data tidak seimbang di berbagai bidang, antara lain deteksi penipuan kartu kredit dengan kategori transaksi *fraud* jauh lebih sedikit daripada transaksi normal (Gupta dkk*.,* 2022); inspeksi cacat manufaktur, dengan kategori produk cacat jauh lebih sedikit daripada produk baik (Bai dkk*.,* 2024); serta dalam bidang medis, seperti pemodelan luaran kritis di perawatan intensif (Atallah dkk*.,* 2023) dan prediksi mortalitas ICU pada pasien sepsis, dengan kejadian kematian lebih sedikit daripada non-kematian (Al-Ansari dkk*.,* 2025). Hal tersebut menegaskan bahwa perlunya penanganan ketidakseimbangan kelas pada data.

Berbagai pendekatan telah dikembangkan untuk menangani data tidak seimbang. Salah satu pendekatan yang paling banyak digunakan pada regresi logistik adalah teknik *resampling*, yaitu suatu teknik untuk membentuk sampel baru dari data yang tersedia sehingga proporsi kelas menjadi seimbang (Erlangga dkk., 2024). Teknik ini dikelompokkan menjadi tiga kategori, yaitu *undersampling* yang mengurangi observasi pada kelas mayoritas, *oversampling* yang menambah representasi kelas minoritas serta pendekatan *over–under sampling* yang menggabungkan keduanya (Sir & Soepranoto,

2022). Namun, teknik *resampling* memiliki beberapa kelemahan, yaitu *undersampling* berisiko membuang pengamatan penting yang justru membedakan kedua kelas, sehingga informasi pemodelan berkurang. Sementara itu, *oversampling* rentan menimbulkan *overfitting* pada kelas minoritas karena penambahan data sering dilakukan melalui duplikasi sampel yang jumlahnya sedikit, sehingga model cenderung menghafal pola spesifik dan kemampuan generalisasinya menurun (Jonathan dkk*.,* 2020).

Penanganan data tidak seimbang pada regresi logistik terus dikembangkan untuk meningkatkan ketepatan klasifikasi pada kelas minoritas. King dan Zeng (2001) mengusulkan koreksi bias untuk data tidak seimbang agar estimasi probabilitas kelas minoritas menjadi lebih akurat. Maalouf dan Trafalis (2011) juga mengusulkan metode *Rare Event Weighted Kernel Logistic Regression* (RE-WKLR) yang mengintegrasikan antara *weighting*, regularisasi, kernelisasi. Kombinasi ini menjadikan RE-WKLR efektif untuk klasifikasi pada data tidak seimbang. Namun, keterbatasan dari metode ini adalah hanya disarankan untuk data yang berukuran kecil sampai menengah. Untuk mengatasi hal tersebut Maalouf dan Siddiqi (2014) kemudian memperkenalkan metode *Rare Event Weighted Logistic Regression* (RE-WLR), yakni regresi logistik linier berbobot yang dirancang untuk data besar melalui kombinasi regularisasi, pembobotan, dan koreksi bias sehingga dapat dihasilkan model yang baik untuk data yang tidak seimbang. Metode ini diuji pada sejumlah himpunan data yang bervariasi dalam jumlah sampel, banyaknya variabel, dan tingkat ketidakseimbangan kelas. RE-WLR secara konsisten memberikan akurasi lebih tinggi sekaligus waktu komputasi lebih cepat dibandingkan *Truncated Regularized Iteratively Reweighted Least Squares* (TR-IRLS), serta mencapai akurasi setara tetapi dengan komputasi jauh lebih cepat daripada *Support Vector Machine* (SVM) pada skenario data besar yang tidak seimbang (Maalouf & Siddiqi, 2014).

Penelitian terdahulu mengenai penerapan RE-WLR telah dilakukan oleh Triasmoro dkk. (2018) yang membandingkan RE-WLR dengan *Truncated Regularized Prior Correction* (TR-PC) pada klasifikasi kesejahteraan rumah tangga di Bali. Hasilnya menunjukkan bahwa RE-WLR memberikan kinerja yang lebih baik dibanding TR-PC dengan nilai AUC sebesar 83.5% serta menunjukkan RE-WLR secara deskriptif lebih baik dengan rata-rata sensitivitas lebih tinggi, sehingga lebih mampu memprediksi kelas minoritas. Merdiko & Ratnasari (2022) juga menerapkan RE-WLR pada data jenis perceraian di Kabupaten Bojonegoro yang tidak seimbang dan memperoleh hasil kinerja dengan kategori *good classification*. Selanjutnya, Naba dkk*.* (2023) yang menerapkan RE-WLR untuk memodelkan status kemiskinan rumah tangga pada data SUSENAS Provinsi Jawa Timur yang tidak seimbang, hasilnya menunjukkan bahwa RE-WLR mampu menangani data tersebut dengan cukup baik. Ketiga penelitian tersebut berfokus pada isu sosial-kependudukan dan kesejahteraan, sehingga pada penelitian ini RE-WLR diterapkan pada bidang kesehatan masyarakat dengan studi kasus data Keluarga Berisiko Stunting di Kabupaten Jeneponto.

Stunting merupakan masalah gizi kronis yang disebabkan oleh asupan nutrisi yang tidak memadai dalam jangka waktu panjang, sering kali karena pola pemberian makan yang tidak sesuai kebutuhan. Kondisi ini dapat mulai sejak masa kehamilan dan umumnya baru tampak ketika anak berusia dua tahun serta berdampak jangka panjang pada kesehatan, perkembangan kognitif, serta produktivitas di masa depan (Kemenkes, 2016). Masalah kesehatan ini sering kali melibatkan banyak faktor yang saling

berhubungan, seperti keterbatasan akses layanan kesehatan, malnutrisi kronis, serta sanitasi yang buruk (Unicef, 2023). Berdasarkan hasil Surveri Status Gizi Indonesia (SSGI) prevalensi stunting pada tingkat nasional mengalami penurunan dari 21.5% di tahun 2023 menjadi 19,8% di tahun 2024 (Kemenkes BKPK, 2025). Angka ini masih menjadi tantangan karena target Rencana Pembangunan Jangka Menengah Nasional (RPJMN) adalah 14,2% pada tahun 2025-2029 (Bappenas, 2025). Meskipun prevalensi stunting pada tingkat nasional mengalami penurunan, capaian tersebut masih terdapat kesenjangan antarwilayah sehingga penting untuk meninjau kondisi di tingkat provinsi hingga kabupaten/kota. Di Sulawesi Selatan, SSGI 2024 menunjukkan prevalensi sekitar 23,3%, lebih tinggi dari rata-rata nasional. Salah satu kabupaten yang menyumbang prevalensi tertinggi adalah Kabupaten Jeneponto, sebesar 37% jauh diatas rata-rata provinsi (Kemenkes BKPK, 2025). Hal ini menunjukkan urgensi penelitian yang lebih mendalam untuk memahami dan menganalisis faktor risiko utama stunting.

Berdasarkan urain tersebut, penulis tertarik untuk melakukan analisis faktor risiko stunting di Kabupaten Jeneponto menggunakan metode *Rare Event Weighted Logistic Regression*. Variabel respon yang digunakan adalah Keluarga Berisiko Stunting memiliki skala biner (“Ya” atau “Tidak”). Data menunjukkan ketidakseimbangan kelas, karena jumlah keluarga berisiko lebih sedikit daripada yang tidak berisiko. Oleh karena itu, RE-WLR dipilih agar estimasi probabilitas tidak bias terhadap kelas minoritas sehingga dapat memberikan prediksi yang lebih akurat terkait risiko stunting di Kabupaten Jeneponto.

1.2 Tujuan Penelitian

Berdasarkan latar belakang, maka diperoleh tujuan penelitian sebagai berikut:

1. Memperoleh estimasi parameter model *Rare Event Weighted Logistic Regression* menggunakan metode *Maximum Likelihood Estimation*.
2. Memperoleh model dan informasi mengenai faktor yang berpengaruh signifikan terhadap keluarga berisiko stunting di Kabupaten Jeneponto tahun 2024 menggunakan metode *Rare Event Weighted Logistic Regression*.
3. Memperoleh performa model hasil klasifikasi keluarga berisiko stunting di Kabupaten Jeneponto tahun 2024 menggunakan metode *Rare Event Weighted Logistic Regression*.

1.3 Manfaat Penelitian

Adapun penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Memberikan wawasan mengenai penerapan metode *Rare Event Weighted Logistic Regression* dalam menangani permasalahan klasifikasi pada data tidak seimbang.
2. Memberikan wawasan mengenai faktor risiko yang berkontribusi terhadap keluarga berisiko stunting dan dapat digunakan sebagai dasar dalam perumusan kebijakan kesehatan di Kabupaten Jeneponto.

1.4 Batasan Masalah

Untuk membatasi ruang lingkup permasalahan pada penelitian ini, maka diberikan batasan bahwa:

1. Data yang digunakan adalah data keluarga berisiko stunting beserta sepuluh variabel yang diduga memengaruhinya di Kabupaten Jeneponto tahun 2024.

2. Pemodelan menggunakan pembagian data 80% training dan 20% testing menggunakan teknik sampling stratifikasi.
3. Nilai λ yang digunakan sebagai parameter regularisasi adalah 1-10 dengan *increment* sebesar 1.

1.5 Teori

1.5.1 Regresi Logistik Biner

Regresi logistik biner merupakan metode yang digunakan untuk menggambarkan hubungan antara satu atau lebih variabel prediktor dan variabel respons bersifat dikotomik (Tampil dkk., 2017). Sifat tersebut berarti bahwa variabel responnya terdiri dari dua kategori, misalkan $y = 1$ menyatakan hasil yang diperoleh 'sukses' dan $y = 0$ menyatakan hasil yang diperoleh 'gagal'. Model regresi logistik biner termasuk ke dalam distribusi Bernoulli, yaitu distribusi dari variabel respon yang hanya mempunyai dua kategori yaitu 0 dan 1. Adapun fungsi kepadatan peluang distribusi bernoulli pada Persamaan (1) (Utami dkk., 2024).

$$f(y_i) = \pi(x_i)^{y_i}\big(1 - \pi(x_i)\big)^{1-y_i} \quad ; y_i = 0{,}1 \tag{1}$$

dengan $\pi(x_i)$ merupakan probabilitas yang bergantung pada x_i, i merupakan indeks yang mengacu pada data ke$-i$, y_i merupakan variabel respon ke$-i$ yang menunjukkan hasil 0 atau 1.

Hosmer dkk. (2013) menyatakan bahwa peluang kejadian $\pi(x_i)$ dapat dinyatakan seperti pada Persamaan (2)

$$\pi(x_i) = \frac{\exp\big(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\big)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)} = \frac{\exp\big(\mathbf{X}_i^T\boldsymbol{\beta}\big)}{1 + \exp(\mathbf{X}_i^T\boldsymbol{\beta})} \tag{2}$$

dengan $\mathbf{X}_i^T = \big(1, x_{i1}, x_{i2}, \dots, x_{ip}\big)$ merupakan vektor prediktor untuk observasi ke$-i$, dan $\boldsymbol{\beta} = \big(\beta_0, \beta_1, \dots, \beta_p\big)^T$ merupakan vektor parameter.

Hubungan variabel respon dan prediktor dapat diketahui dengan memperoleh fungsi linier terlebih dahulu. Fungsi linier diperoleh dengan melakukan transformasi $\pi(x_i)$ ke dalam bentuk logit. Bentuk logit dari $\pi(x_i)$ adalah:

$$g(x_i) = \ln\left(\frac{\pi(x_i)}{1 - \pi(x_i)}\right) = \beta_0 + \sum\nolimits_{j=1}^{p}\beta_j x_{ij} \tag{3}$$

atau dapat dijabarkan menjadi $g(x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$.

Metode yang digunakan dalam mengestimasi parameter model regresi logistik adalah *Maksimum Likelihood Estimation* (MLE) yang dilakukan dengan memaksimumkan fungsi *likelihood* dimana mensyaratkan data mengikuti suatu distribusi tertentu (Agresti,2002). Sehingga dengan mengasumsikan antar observasi saling bebas, dan y_i untuk $i = 1{,}2, \dots, n$ merupakan sampel acak yang telah terambil, maka fungsi likelihoodnya dapat dituliskan pada persamaan 4.

$$L(\beta) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \pi(x_i)^{y_i}\big(1 - \pi(x_i)\big)^{1-y_i} \tag{4}$$

Dalam perhitungannya, fungsi *likelihood* akan lebih mudah dimaksimumkan dalam bentuk *log-likelihood* yang dituliskan pada persamaan 5.

$$\ln L(\beta) = \ln\left(\prod_{i=1}^{n} \pi(x_i)^{y_i}\left(1-\pi(x_i)\right)^{1-y_i}\right)$$
$$= \sum_{i=1}^{n}\left[y_i \ln\left(\pi(x_i)\right) + (1-y_i)\ln\left(1-\pi(x_i)\right)\right] \quad (5)$$

dengan mensubsitusikan Persamaan (2) ke dalam Persamaan (5), maka log likelihood dapat dinyatakan secara eksplisit dalam $\boldsymbol{\beta}$ sebagai berikut:

$$\ln \mathrm{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left[y_i \ln\left(\frac{\exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)}{1+\exp(\mathbf{X}_i^T\boldsymbol{\beta})}\right) + (1-y_i)\ln\left(\frac{1}{1+\exp(\mathbf{X}_i^T\boldsymbol{\beta})}\right)\right] \quad (6)$$

Estimasi parameter dengan metode MLE tidak menghasilkan solusi yang berbentuk *closed form*, sehingga estimasi parameter pada regresi logistik biner dengan metode MLE ini diperoleh menggunakan optimasi numerik. Salah satu metode numerik yang biasa digunakan adalah menggunakan metode iterasi Newton-Raphson.

1.5.2 Data Tidak Seimbang

Data tidak seimbang merupakan kondisi ketika jumlah observasi pada satu kelas jauh lebih banyak daripada kelas lainnya, sehingga distribusi kelas condong ke mayoritas dan kelas minoritas hanya mewakili proporsi kecil dari keseluruhan data (Ridwan dkk., 2024). Kelas yang memiliki data sampel sangat besar disebut kelas mayoritas, sedangkan kelas yang memiliki data sampel sedikit disebut kelas minoritas (Sabilla & Vista, 2021). Ketimpangan ini lazim dijumpai pada berbagai tugas klasifikasi di dunia nyata dan sering dinyatakan dengan *imbalance ratio* (IR) yang menggambarkan perbandingan ukuran kelas mayoritas terhadap minoritas (Gumelar dkk., 2023). Ketidakseimbangan data berdampak pada kinerja pemodelan karena model cenderung berpihak pada kelas mayoritas sehingga peluang kelas minoritas kerap terestimasi terlalu rendah dan banyak kasus minoritas salah klasifikasi (Fitriani dkk., 2021). Di sisi lain, penggunaan ambang keputusan standar 0,5 sering tidak memadai karena membuat sensitivitas terhadap kelas minoritas rendah (Khairunnisa, 2023). Oleh karena itu, evaluasi kinerja tidak cukup mengandalkan akurasi keseluruhan, tetapi perlu menekankan metrik yang peka terhadap ketidakseimbangan kelas seperti *sensitivity*, *specificity*, G-mean, serta *Area Under Curve* (AUC), karena metrik-metrik tersebut lebih representatif untuk menilai kemampuan model pada kedua kelas (Saifudin dkk., 2015).

1.5.3 *Regularization*

Pada regresi logistik untuk klasifikasi, umumnya menggunakan data *training* dan *testing* untuk mengevaluasi ketepatan klasifikasi dari model. Namun, proses ini permasalahan yang dapat muncul diantaranya adalah adanya overfitting pada data training, yaitu suatu kondisi ketika model yang dihasilkan sangat baik untuk data training tetapi mempunyai performa yang buruk untuk data *testing* (Naba dkk., 2023). Selain itu, model regresi logistik sangat rentan terhadap multikolinearitas, yaitu adanya korelasi yang tinggi antar variabel prediktor (Merdiko & Ratnasari, 2022). Oleh karena itu, salah satu solusi yang

dapat mengatasi kedua masalah tersebut adalah dengan menggunakan *regularization*. Fungsi *regularized log likelihood* yang menerapkan regularisasi ridge yang didefinisikan pada Persamaan 7.

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left[y_i \ln\left(\frac{\exp(\mathbf{X}_i^T\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i^T\boldsymbol{\beta})}\right) + (1-y_i)\ln\left(\frac{1}{1+\exp(\mathbf{X}_i^T\boldsymbol{\beta})}\right)\right] - \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2$$

$$= \sum_{i=1}^{n}\ln\left(\frac{\exp(y_i\mathbf{X}_i^T\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i^T\boldsymbol{\beta})}\right) - \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2 \quad (7)$$

dengan nilai $\|\boldsymbol{\beta}\|^2 = \sum_{j=1}^{p}\beta_j^2$ adalah penalty dan $\lambda > 0$ adalah parameter regularisasi. Penambahan unsur $\frac{\lambda}{2}\|\boldsymbol{\beta}\|^2$ berfungsi sebagai regularisasi yang digunakan untuk meningkatkan kinerja generalisasi yang lebih baik dan dapat mengatasi masalah *overfitting* (Maalouf & Siddiqi, 2014).

1.5.4 *Rare Event Weighted Logistic Regression*

Rare Event Weighted Logistic Regression (RE-WLR) pertama kali diperkenalkan oleh Maalouf dan Siddiqi (2014) merupakan pengembangan dari *Rare Event Weighted Kernel Logistic Regression* (RE-WKLR) guna mengatasi permasalahan data tidak seimbang pada dataset yang berukuran besar. Pada penerapannya RE-WLR mengintegrasikan *regularization* dan *weighting* pada regresi logistik. Secara definisi, *rare events* merupakan suatu kejadian yang frekuensi terjadinya secara substansial lebih kecil daripada frekuensi kejadian pada umumnya (Merdiko & Ratnasari, 2022).

Menurut King dan Zeng (2001) untuk mengatasi kasus data tidak seimbang atau *rare events data* maka dapat dilakukan prosedur alternatif yaitu dengan cara *weighting* atau pembobotan untuk mengoreksi estimasi pemilihan sampel y. Prosedur tersebut diterapkan karena proporsi salah satu kelas jauh lebih kecil dibandingkan kelas yang lain sehingga diperlukan suatu *treatment* pada data training yang memungkinkan suatu model dapat menangkap informasi lebih tepat. Pembobotan ini dilakukan saat memodelkan dengan menggunakan data *training*, sehingga fungsi *log-likelihood* untuk *weighted logistic regression* adalah dengan menambahkan bobot w_i pada *log-likelihood* regresi logistik biasa (Sulasih dkk., 2015).

$$\ln \mathrm{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i \ln\left(\frac{\exp(y_i\mathbf{X}_i^T\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i^T\boldsymbol{\beta})}\right) \quad (8)$$

dengan $w_i = \left(\frac{\tau}{\bar{y}}\right)y_i + \left(\frac{1-\tau}{1-\bar{y}}\right)(1-y_i)$.

Keterangan:
$\tau =$ proporsi kejadian sukses (*event*) dalam seluruh data
$\bar{y} =$ proporsi kejadian sukses (*event*) dalam data *training*

Pada WLR, vektor gradien dan matriks Hessian dapat diperoleh dengan cara menurunkan *regularized weighted log-likelihood* terhadap masing-masing β.

$$\ln \mathrm{L}_W(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i \ln\left(\frac{\exp(y_i \mathbf{X}_i^T \boldsymbol{\beta})}{1+\exp(\mathbf{X}_i^T \boldsymbol{\beta})}\right) - \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2 \tag{9}$$

sehingga apabila dituliskan dalam bentuk matriks, didapatkan persamaan gradien sebagai berikut:

$$\frac{\partial}{\partial \boldsymbol{\beta}} \ln \mathrm{L}_W(\boldsymbol{\beta}) = \mathbf{X}^{\mathrm{T}}\mathbf{W}(\mathbf{y}-\boldsymbol{\pi}) - \boldsymbol{\lambda}\boldsymbol{\beta} \tag{10}$$

dengan $\mathbf{W} = diag(w_i)$ dan $\boldsymbol{\pi} = \big(\pi(x_1), \dots, \pi(x_n)\big)^T$ adalah vektor peluang seperti pada Persamaan (2), sedangkan Matriks Hessiannya adalah

$$H(\boldsymbol{\beta}) = -\mathbf{X}^T\mathbf{D}\mathbf{X} - \lambda\mathbf{I} \tag{11}$$

dengan $\mathbf{D} = diag(v_i w_i); v_i = \pi(x_i)\big(1-\pi(x_i)\big)$.

Estimasi parameter *regularized weighted log likelihood* menggunakan metode iterasi Newton-Raphson dengan rumus update Newton-Raphson untuk $\boldsymbol{\beta}$ pada iterasi ke$-(c+1)$ adalah

$$\widehat{\boldsymbol{\beta}}^{(c+1)} = (\mathbf{X}^T\mathbf{D}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{D}\mathbf{z}^{(c)}, \tag{12}$$

dengan

$$\mathbf{z}^{(\mathrm{c})} = \mathbf{X}\widehat{\boldsymbol{\beta}}^{(\mathrm{c})} + \mathbf{V}^{-\mathbf{1}}(\mathbf{y}-\boldsymbol{\pi})$$

sebagai variabel respon yang disesuaikan $\mathbf{V} = diag(v_i)$. Dengan demikian, pada setiap iterasi diperoleh Persamaan (13).

$$(\mathbf{X}^T\mathbf{D}\mathbf{X} + \lambda\mathbf{I})\widehat{\boldsymbol{\beta}}^{(c+1)} = \mathbf{X}^T\mathbf{D}\mathbf{z}^{(c)}. \tag{13}$$

Permasalahan proses iteratif pada Persamaan (13) dapat diatasi dengan menggunakan metode *Conjugate Gradient* (CG) dengan memberikan batas tertentu pada jumlah iterasi CG untuk menghindari waktu komputasi yang terlalu lama. Setelah diperoleh estimasi parameternya menggunakan algoritma CG, maka dapat dilanjutkan dengan memasukkan estimasi parameter tersebut ke dalam persamaan logit untuk memperoleh prediksi probabilitas. Persamaan RE-WLR untuk peluang kejadian pada observasi ke$-i$ dapat dituliskan pada Persamaan (14).

$$\tilde{\pi}(x_i) = \frac{\exp\big(\mathbf{X}_i^T\widehat{\boldsymbol{\beta}}\big)}{1+\exp\big(\mathbf{X}_i^T\widehat{\boldsymbol{\beta}}\big)} \tag{14}$$

1.5.5 Pengujian Signifikansi Parameter

Pengujian signifikansi parameter merupakan pengujian yang digunakan untuk menguji signifikansi koefisien β dari model. Pengujian ini dapat menggunakan uji secara parsial dan secara simultan.

1.5.6.1 Pengujian Signifikansi Parameter secara Simultan

Uji simultan dalam regresi logistik menggunakan uji *Likelihood Ratio* (LR) untuk menguji signifikansi keseluruhan model. Uji ini membandingkan model penuh (dengan variabel

prediktor) terhadap model terbatas (tanpa variabel prediktor). Hipotesis yang digunakan adalah sebagai berikut:

$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$
H_1: Minimal terdapat satu $\beta_j \neq 0$, untuk $j = 1,2, \ldots p$

Statistik uji LR didasarkan pada nilai G yang merupakan perbandingan antara *log-likelihood* dari model penuh dan model terbatas sesuai pada Persamaan (15).

$$G = -2 \times \left(\ln L_{terbatas} - \ln L_{penuh}\right) \tag{15}$$

dengan $\ln L_{terbatas}$ adalah *log-likelihood* dari model terbatas dan $\ln L_{penuh}$ adalah *log-likelihood* dari model penuh. Statistik uji ini mengikuti distribusi chi-kuadrat (χ^2) dengan derajat kebebasan (db) sama dengan jumlah variabel prediktor dalam model (Hidayat & Hajarisman, 2023). Sementara itu, kriteria penolakannya adalah ketika nilai $G > \chi^2_{p;\alpha}$ atau $p - value < \alpha$, maka H_0 ditolak yang berarti bahwa model dengan variabel prediktor lebih baik dalam menjelaskan data.

1.5.6.2 Pengujian Signifikansi Parameter secara Parsial

Pengujian signifikansi parameter β_j dalam model regresi logistik dilakukan dengan menggunakan Uji Wald. Uji ini bertujuan untuk mengevaluasi signifikansi pengaruh tiap parameter β_j terhadap variabel respon secara parsial dengan mengikuti hipotesis berikut:

$H_0: \beta_j = 0$
H_1: $\beta_j \neq 0$ untuk $j = 1,2, \ldots p$
Adapun statistik uji Wald dinyatakan pada Persamaan (16).

$$W_j = \left(\frac{\hat{\beta}_j}{SE\left(\hat{\beta}_j\right)}\right)^2 \tag{16}$$

dengan $\hat{\beta}_j$ merupakan estimasi koefisien regresi dan $SE\left(\hat{\beta}_j^2\right)$ merupakan *standard error* dari estimasi koefisien tersebut. Nilai W_j mengikuti distribusi *chi-square* dengan $df = 1$. Sementara itu, kriteria uji yang digunakan adalah ketika $\left|W_j\right| > \chi^2_{\alpha;1}$ atau $p - value < \alpha$, maka keputusan adalah menolak H_0 yang berarti X_j berpengaruh signifikan terhadap variabel respon secara parsial (Hosmer dkk., 2013).

1.5.6 Interpretasi Koefisien Parameter

Salah satu ukuran yang dapat digunakan untuk menginterpretasikan koefisien variabel prediktor disebut *Odds ratio*. *Odds ratio* merupakan perbandingan peluang munculnya suatu kejadian dengan peluang tidak munculnya kejadian tersebut. *Odds ratio* memudahkan interpretasi model regresi logistik yang diperoleh. Interpretasi parameter bertujuan untuk memahami makna nilai taksiran parameter pada variabel prediktor (Hosmer dkk., 2013). Pada regresi logistik dengan variabel prediktor biner, nilai X dikategorikan 0 atau 1, sehingga terdapat dua nilai $\pi(x)$ dan dua nilai $1 - \pi(x)$ seperti yang ditunjukkan pada Tabel 1.

Tabel 1. Probabilitas respon biner Y berdasarkan variabel prediktor biner

Variabel Respon	Variabel Prediktor	
	$X = 1$	$X = 0$
$Y = 1$	$\pi(1) = \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}$	$\pi(0) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$
$Y = 0$	$1 - \pi(1) = \frac{1}{1 + \exp(\beta_0 + \beta_1)}$	$1 - \pi(0) = \frac{1}{1 + \exp(\beta_0)}$
Total	1	1

Nilai *odds* dari variabel respon di antara observasi dengan $X = 1$ adalah $\frac{\pi(1)}{1-\pi(1)}$, sedangkan jika $X = 0$ maka nilai odds $\frac{\pi(0)}{1-\pi(0)}$. Odds ratio dituliskan dengan OR yang didefinisikan sebagai rasio odds untuk $X = 1$ terhadap odds untuk $X = 0$, yang dapat dituliskan pada Persamaan (17).

$$OR = \frac{\frac{\pi(1)}{[1-\pi(1)]}}{\frac{\pi(0)}{[1-\pi(0)]}}$$

berdasarkan Tabel 1, maka diperoleh nilai odds ratio sebagai berikut

$$\begin{aligned} OR &= \frac{\left(\frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}\right)}{\left(\frac{1}{1 + \exp(\beta_0 + \beta_1)}\right)} \Big/ \frac{\left(\frac{\exp(\beta_0)}{1 + \exp(\beta_0)}\right)}{\left(\frac{1}{1 + \exp(\beta_0)}\right)} \\ &= \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} \\ &= \exp(\beta_1) \end{aligned} \tag{17}$$

Jika nilai $OR = 1$, maka tidak ada hubungan antara variabel prediktor dengan variabel respon. Jika $OR < 1$, maka ada hubungan negatif antara variabel prediktor dan variabel respon pada setiap perubahan nilai X. Jika $OR > 1$, maka ada hubungan positif antara variabel prediktor dengan variabel respon pada setiap perubahan nilai X (Hosmer dkk., 2013).

1.5.7 Uji Kesesuaian Model

Uji kesesuaian model atau *goodness-of-fit test* bertujuan untuk mengetahui apakah tidak terdapat perbedaan antara hasil observasi dengan kemungkinan hasil prediksi. Pada regresi logistik digunakan nilai *deviance* untuk mengetahui kesesuaian model (Hosmer dkk., 2013). Hipotesis pada uji kesesuaian model adalah sebagai berikut:

H_0: Model sesuai (tidak terdapat perbedaan yang signifikan antara hasil observasi dengan kemungkinan hasil prediksi model)
H_1: Model tidak sesuai (terdapat perbedaan yang signifikan antara hasil observasi dengan kemungkinan hasil prediksi model)

Statistik uji yang digunakan untuk menghitung nilai deviance dinyatakan pada Persamaan (18)

$$D = -2\sum_{i=1}^{n}\left(y_i \ln\left(\frac{\tilde{\pi}(x_i)}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\tilde{\pi}(x_i)}{1-y_i}\right)\right) \tag{18}$$

dengan daerah penolakan $D > \chi^2_{(db,\alpha)}$ atau $p - value < \alpha$.

1.5.8 Ketepatan Klasifikasi Model

Ketepatan klasifikasi model merupakan ukuran sejauh mana hasil prediksi yang dihasilkan oleh suatu model klasifikasi sesuai dengan kelas sebenarnya dari data yang diuji. Evaluasi ketepatan ini sangat penting dalam menentukan seberapa baik model dapat mengklasifikasikan data baru secara benar. Pada regresi logistik, ketepatan klasifikasi dapat dievaluasi dengan menggunakan *confusion matrix* (Susetyoko dkk., 2022). Pada kejadian biner, metode ini menganalisis hasil klasifikasi model dengan merangkum prediksi ke dalam empat kategori sesuai pada Tabel 2 (Sathyanarayanan & Tantri, 2024).

Tabel 2. *Confusion Matrix*

Hasil Observasi	Hasil Prediksi	
	Positive	*Negative*
Positive	*True Positive*	*False Negative*
Negative	*False Positive*	*True Negative*

dengan

True Positive : Prediksi positif yang benar (TP).
False Positive : Prediksi positif yang salah (FP).
True Negative : Prediksi negatif yang benar (TN).
False Negative : Prediksi negatif yang salah (FN).

Berdasarkan Tabel 2, maka metrik akurasi dapat dihitung sebagai proporsi dari jumlah prediksi yang benar (TP dan TN) dari total prediksi, baik untuk prediksi positif maupun negatif. Adapun perhitungannya mengikuti Persamaan (19).

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \tag{19}$$

Metrik akurasi pada Persamaan (19) tidak mampu menunjukkan frekuensi dari FN dan FP. Karena pada kasus data tidak seimbang, akurasi dapat menjadi bias. Hal ini terjadi ketika distribusi kelas tidak merata, sehingga model dapat menghasilkan akurasi tinggi hanya dengan memprediksi kelas mayoritas (Fitriani dkk., 2021). Sebagai indikator yang dapat mengakomodir akurasi pada kelas positif (*sensitivity*) dan kelas negatif (*specificity*) dalam kasus data tidak seimbang, maka kinerja klasifikasi dihitung juga menggunakan indikator G-Mean, yang menunjukkan keseimbangan kinerja model pada kedua kelas (Sathyanarayanan & Tantri, 2024) . Adapun perhitungannya mengikuti Persamaan (20)

$$G - mean = \sqrt{sensitivity \times specificity} \tag{20}$$

dengan nilai *sensitivity* menunjukkan seberapa besar kemampuan model dalam mengenali atau mengidentifikasi dengan benar kejadian-kejadian yang termasuk dalam kelas positif. Nilai *sensitivity* dapat dihitung menggunakan Persamaan (21).

$$Sensitivity = \frac{TP}{TP + FN} \tag{21}$$

sedangkan untuk nilai *specificity* menunjukkan sejauh mana model mampu mengenali atau mengidentifikasi dengan benar kejadian-kejadian yang termasuk dalam kelas negative. Nilai ini dapat dihitung menggunakan Persamaan (22).

$$Specificity = \frac{TN}{TN + FP} \tag{22}$$

Selain menggunakan G-Mean, ketepatan suatu model dalam melakukan prediksi juga dapat dievaluasi melalui AUC (*Area Under the Curve*). Nilai AUC menunjukkan luas area di bawah kurva ROC yang merepresentasikan kemampuan model dalam membedakan antara kelas positif dan negatif (Saraswati dkk., 2019). Semakin besar nilai AUC, maka semakin baik pula kinerja model dalam melakukan klasifikasi, karena model mampu mengidentifikasi kelas secara lebih akurat dan konsisten. Nilai AUC dapat dihitung menggunakan Persamaan (23).

$$AUC = \frac{1}{2}(Sensitivity + Specificity) \tag{23}$$

Nilai AUC yang semakin mendekati 1 menunjukkan bahwa model klasifikasi yang terbentuk semakin akurat. Terdapat beberapa kriteria keakuratan hasil klasifikasi berdasarkan AUC yang ditampilkan pada Tabel 3 (Merdiko & Ratnasari, 2022).

Tabel 3. Kriteria Nilai AUC

Nilai AUC	Kriteria
$0.9 - 1.0$	*Excellent Classification*
$0.8 - 0.9$	*Good Classification*
$0.7 - 0.8$	*Fair Classification*
$0.6 - 0.7$	*Poor Classification*
$0.5 - 0.6$	*Failed*

1.5.9 Keluarga Berisiko Stunting

Keluarga Berisiko Stunting (KRS) adalah keluarga yang memiliki satu atau lebih faktor yang meningkatkan kemungkinan terjadinya stunting pada anak. Penapisan KRS digunakan pemerintah sebagai dasar penentuan sasaran pendampingan di lapangan dalam kerangka percepatan penurunan stunting (Permatasari dkk., 2022). Dalam praktik program, penapisan KRS diarahkan pada keluarga dengan pasangan usia subur, ibu hamil, anak 0–23 bulan, dan anak 24–59 bulan, dengan indikator yang mudah diamati di rumah tangga seperti akses air minum, sanitasi, serta kondisi reproduksi dan kesejahteraan keluarga (BKKBN, 2023). Secara umum faktor-faktor yang dapat memengaruhi status Keluarga Berisiko Stunting diantaranya adalah periode 1.000 Hari Pertama Kehidupan (HPK), kualitas lingkungan, karakteristik reproduksi pasangan usia subur, kondisi sosial ekonomi, dan dukungan sosial.

Masa 1.000 HPK penting karena sejak kehamilan hingga anak berusia dua tahun, merupakan periode emas untuk tumbuh kembang anak. Jika selama waktu ini ibu kekurangan gizi atau anak tidak mendapatkan asupan yang cukup, pertumbuhan fisik

dan kemampuan kognitif anak bisa terganggu. Misalnya, ibu hamil yang anemia atau kurang energi kronis dapat melahirkan bayi dengan berat badan rendah yang kemudian lebih rentan stunting. Oleh Karena itu, penting bagi keluarga dengan ibu hamil atau memiliki anak di bawah dua tahun untuk diperhatikan status gizinya sejak dini (Gunardi, 2021; Hasna dkk., 2024).

Kualitas lingkungan juga turut memengaruhi risiko, sebab lingkungan yang bersih dan sehat membantu mencegah anak dari penyakit infeksi seperti diare dan cacingan yang dapat menghambat penyerapan gizi (Kanan dkk., 2024). Akses terhadap air minum yang aman dan fasilitas sanitasi yang layak, seperti jamban sehat, sangat berpengaruh terhadap kesehatan keluarga. Rumah tangga yang masih menggunakan air dari sumber tidak terlindungi atau buang air sembarangan lebih berisiko mengalami stunting karena anak lebih sering sakit (Nasyidah dkk., 2023).

Faktor reproduksi, seperti usia dan jarak kehamilan, juga penting. Salah satu pendekatan untuk memaham faktor risiko ini adalah melalui identifikasi Pasangan Usia Subur (PUS). PUS didefinisikan sebagai pasangan suami istri dengan istri berusia 15–49 tahun yang masih mengalami haid, atau istri berusia kurang dari 15 tahun tetapi sudah haid. Keluarga dengan PUS yang berusia terlalu muda atau terlalu tua, memiliki jumlah anak banyak, atau jarak kelahiran yang terlalu dekat lebih berisiko memiliki anak stunting karena sumber daya keluarga terbagi dan terbatas (Permatasari dkk., 2022).

Bantuan sosial pemerintah seperti Program Keluarga Harapan (PKH) berfungsi sebagai jaring pengaman bagi keluarga rentan. Melalui bantuan ini, keluarga didorong untuk secara rutin memeriksakan kehamilan, menimbang anak, serta mengikuti penyuluhan gizi dan kesehatan (Hidayat dkk., 2011). Selain itu keluarga yang mendapatkan pendampingan Komunikasi, Informasi, dan Edukasi (KIE) menunjukkan bahwa kegiatan KIE mampu meningkatkan pengetahuan dan praktik pencegahan stunting di tingkat keluarga (Haryani dkk., 2021), Oleh karena itu, Pendampingan yang baik dapat meningkatkan kesadaran dan perilaku sehat keluarga, sehingga membantu menurunkan angka stunting (Munawaroh dkk., 2024).

BAB II
METODE PENELITIAN

2.1 Jenis dan Sumber Data

Data yang digunakan pada penelitian ini merupakan data sekunder, yaitu data Keluarga Berisiko Stunting di Kabupaten Jeneponto pada tahun 2024. Data tersebut diperoleh dari Badan Kependudukan dan Keluarga Berencana Nasional (BKKBN) Perwakilan Sulawesi Selatan. Adapun surat pengajuan data dapat dilihat pada Lampiran 1. Data Keluarga Berisiko Stunting yang digunakan terdiri atas 42.688 observasi yang terdapat pada Lampiran 2.

2.2 Variabel Penelitian

Variabel yang digunakan pada penelitian ini terdiri atas satu variabel respon dan sepuluh variabel prediktor. Variabel respon yang digunakan merupakan variabel kategorik biner, yaitu Keluarga Berisiko Stunting yang terdiri atas dua kategori “Ya” dan “Tidak”. Adapun variabel prediktornya merupakan faktor-faktor yang diduga memengaruhi Keluarga Berisiko Stunting di Kabupaten Jeneponto tahun 2024 yang dijelaskan pada Tabel 4.

Tabel 4. Variabel Penelitian

Jenis Variabel	Kode	Nama Variabel	Kategori
Respon	y	Keluarga Berisiko Stunting	1=Ya 0=Tidak
Prediktor	x_1	Status Hamil	1=Ya 2=Tidak
	x_2	Keluarga Memiliki Baduta	1=Ya 2=Tidak
	x_3	Kondisi Sumber Air Minum Utama	1=Layak 2=Tidak Layak
	x_4	Kondisi Tempat Buang Air Besar	1=Layak 2=Tidak Layak
	x_5	Keluarga memiliki PUS Terlalu Muda	1=Ya 2=Tidak
	x_6	Keluarga memiliki PUS Terlalu Tua	1=Ya 2=Tidak
	x_7	Keluarga memiliki PUS Terlalu Dekat	1=Ya 2=Tidak
	x_8	Keluarga memiliki PUS Terlalu Banyak	1=Ya 2=Tidak
	x_9	Keluarga Mendapatkan Bansos	1=Ya 2=Tidak
	x_{10}	Keluarga Mendapatkan Pendampingan KIE	1=Ya 2=Tidak

Sumber: BKKBN Perwakilan Provinsi Sulawesi Selatan (2024)

2.3 Metode Analisis Data

Tahapan metode analisis data yang dilakukan pada penelitian ini terdiri atas tiga bagian untuk mencapai atau menjawab tujuan penelitian, yaitu:

1. Melakukan estimasi parameter model *rare event weighted logistic regression,* dengan tahapan sebagai berikut
   a. Mendefinisikan model dasar regresi logistic seperti pada Persamaan (3).
   b. Membentuk fungsi *log-likelihood* untuk model regresi logistik sebagaimana didefinisikan pada Persamaan (6).
   c. Membentuk fungsi *regularized log-likelihood* dengan mengambil bentuk *log-likelihood* pada tahap b dan menambahkan komponen parameter regularisasi sebagaimana didefinisikan pada Persamaan (7).
   d. Menentukan pembobot pada model RE-WLR berdasarkan rasio kejadian dalam sampel dan populasi.
   e. Membentuk fungsi *regularized weighted log-likelihood* untuk model RE-WLR dengan memasukkan bobot w_i ke dalam fungsi *regularized log-likelihood* pada Persamaan (7), sebagaimana didefinisikan pada Persamaan (8).
   f. Memperoleh vektor gradien dengan cara mendapatkan turunan parsial pertama dari *regularized weighted log-likelihood function*.
   g. Memperoleh matriks Hessian dengan cara mendapatkan turunan parsial kedua dari *regularized weighted log-likelihood function*.
   h. Memperoleh iterasi *Newton-Raphson* untuk $\widehat{\boldsymbol{\beta}}$ dengan menggunakan vektor gradien dan matriks Hessian pada hasil turunan parsial pertama dan kedua yaitu tahap (f) dan (g).
   i. Memperoleh penaksir parameter $\widehat{\boldsymbol{\beta}}$ secara numerik menggunakan *Truncated Newton* dengan algoritma *conjugate gradient* (CG) linier.

2. Melakukan pemodelan dengan mengunakan *rare event weighted logistic regression* pada data tidak seimbang menggunakan data Keluarga Berisiko Stunting di Kabupaten Jeneponto pada Tahun 2024. Adapun tahapan pemodelan ini sebagai berikut:
   a. Melakukan analisis statistik deskriptif pada data berdasarkan faktor-faktor yang memengaruhinya.
   b. Melakukan pembagian data *training* dan data *testing* dengan menggunakan sampling stratifikasi.
   c. Melakukan pemodelan RE-WLR sesuai pembagian data yang ditentukan.
   d. Mendapatkan estimasi parameter model RE-WLR berdasarkan prosedur estimasi pada Bagian 1.
   e. Menghitung probabilitas $(\tilde{\pi})$ untuk masing-masing observasi menggunakan Persamaan (15) sehingga diperoleh nilai prediksi kelas dari model RE-WLR.
   f. Memilih parameter regularisasi terbaik dengan membandingkan nilai AUC yang diperoleh dari beberapa kandidat nilai λ yaitu 1-10 dengan *increment* sebesar 1, nilai dengan AUC tertinggi ditetapkan sebagai parameter regularisasi optimum, dan model yang bersesuaian dijadikan model RE-WLR terbaik yang digunakan dalam analisis selanjutnya.

g. Menetapkan model RE-WLR terbaik berdasarkan parameter regularisasi terpilih.
h. Melakukan uji signifikansi parameter model RE-WLR terbaik secara simultan menggunakan Persamaan (16) dan secara parsial menggunakan Persamaan (17).
i. Melakukan uji Kesesuaian model RE-WLR terbaik menggunakan Persamaan (19).
j. Menginterpretasikan hasil estimasi model RE-WLR terbaik.

3. Menghitung ketepatan klasifikasi model *rare event weighted logistic regression* terbaik dari data Keluarga Berisiko Stunting di Kabupaten Jeneponto Tahun 2024. Adapun tahapan untuk memperoleh ketepatan klasifikasi model RE-WLR sebagai berikut:
   a. Membuat confusion matrix dari model RE-WLR terbaik berdasarkan data testing.
   b. Menghitung sensitivity dari hasil klasifikasi model RE-WLR menggunakan Persamaan (22).
   c. Menghitung specificity dari hasil klasifikasi model RE-WLR menggunakan Persamaan (23).
   d. Menghitung nilai G-Mean dari hasil klasifikasi model RE-WLR menggunakan Persamaan (21).
   e. Menghitung nilai AUC dari hasil klasifikasi model RE-WLR menggunakan Persamaan (24).
   f. Menarik Kesimpulan.