BAB I
PENDAHULUAN

1.1 Latar Belakang

Regresi logistik digunakan untuk menganalisis hubungan antara satu atau lebih variabel independen dengan variabel dependen kategorik. Terdapat tiga jenis regresi logistik, yaitu regresi logistik biner yang digunakan ketika variabel dependen memiliki dua kategori, regresi logistik multinomial yang digunakan untuk variabel dependen dengan lebih dari dua kategori, dan regresi logistik ordinal yang diterapkan ketika variabel dependen memiliki urutan peringkat (Hosmer dkk., 2013). Penelitian kesehatan sering menggunakan regresi logistik biner untuk memprediksi hasil biner, seperti “ya” atau “tidak” dengan variabel independen dapat berupa data kontinu maupun kategorik (Ranganathan dkk., 2017). Regresi logistik biner dianggap efektif dalam menganalisis faktor risiko dan hasil pengobatan, serta menghasilkan estimasi yang mudah diinterpretasikan oleh peneliti (Kleinbaum & Klein, 2010). Walaupun demikian, regresi logistik biner memiliki keterbatasan ketika menghadapi masalah multikolinieritas.

Multikolinieritas terjadi ketika variabel independen memiliki korelasi yang tinggi satu sama lain. Kondisi ini menyebabkan varians estimasi parameter meningkat, sehingga hasil regresi menjadi tidak stabil dan sulit diinterpretasikan (Senaviratna & A. Cooray, 2019). Oleh karena itu, diperlukan metode untuk mengatasi masalah ini, salah satunya adalah *Principal Component Analysis* (PCA). Namun, ketika multikolinieritas terjadi pada variabel kategorik, PCA tidak dapat diterapkan secara langsung (Gujarati & Porter, 2008). Sebagai solusinya, digunakan *Categorical Principal Component Analysis* (CATPCA) yang merupakan varian dari PCA yang dapat mereduksi dimensi data kategorik dengan mengubah variabel independen menjadi komponen utama yang tidak berkorelasi. CATPCA dianggap efektif dalam mengatasi multikolinieritas karena tidak sampai menghilangkan variabel, sehingga informasi penting dari tiap variabel tetap terjaga (Khikmah dkk., 2017).

Regresi logistik biner dapat dioptimalkan lebih lanjut melalui pendekatan nonparametrik. Pendekatan ini memungkinkan model untuk menangkap hubungan yang lebih kompleks antar variabel. Salah satu estimator yang umum digunakan dalam pendekatan ini adalah estimator *spline truncated*. Metode ini bekerja dengan memecah data menjadi beberapa segmen dan menyesuaikan kurva untuk menangkap variasi data (Sari dkk., 2023). Penelitian oleh Huang (2020) mengungkapkan bahwa penggunaan estimator *spline truncated* dalam regresi logistik mampu menggantikan fungsi linier yang umumnya digunakan dengan fungsi yang lebih fleksibel. Oleh karena itu, dengan penempatan titik knot yang tepat, maka penggunaan estimator *spline truncated* dapat meningkatkan akurasi model secara signifikan (Rifada dkk., 2023). Dengan demikian, kombinasi CATPCA dan estimator *spline truncated* dalam regresi logistik biner dipandang efektif untuk menangani data kompleks.

Masalah kesehatan seperti *stunting* sering kali melibatkan data dengan banyak faktor yang saling berhubungan, seperti keterbatasan akses layanan kesehatan, malnutrisi kronis, dan sanitasi yang buruk (UNICEF, 2024). Kompleksitas hubungan antara faktor-faktor ini menghadirkan tantangan dalam analisis, karena hubungan antar faktor-faktor yang memengaruhi *stunting* menjadi sulit dipahami (Danso & Appiah, 2023). Akibatnya, penanganan *stunting* masih belum sepenuhnya efektif, termasuk di Provinsi Sulawesi Selatan. Prevalensi *stunting* di Provinsi Sulawesi Selatan mencapai 27,4% pada 2023, meningkat dari 24,9% di tahun sebelumnya (Kemenkes BKPK, 2023). Berdasarkan kajian lebih lanjut, Kabupaten Tana Toraja menyumbang prevalensi tertinggi sebesar 36,9% yang jauh di atas rata-rata provinsi. Angka ini menegaskan perlunya penelitian yang lebih mendalam untuk memahami dan menganalisis faktor risiko utama *stunting*.

Masa 1.000 hari pertama kehidupan anak dapat menjadi salah satu fokus utama dalam penelitian ini, terlebih pada bayi di bawah usia dua tahun (baduta) yang mengalami periode perkembangan kritis. Anak dalam kategori usia ini rentan terhadap risiko *stunting* yang sulit diperbaiki jika terjadi kekurangan gizi (El Taguri dkk., 2009). Di sisi lain, faktor lingkungan seperti akses air minum dan sanitasi yang layak juga memainkan peran penting, karena kondisi lingkungan yang tidak sehat meningkatkan adanya risiko infeksi dan malnutrisi (Cumming & Cairncross, 2016). Adapun faktor seperti usia ibu saat melahirkan dan jarak antar kelahiran turut memberikan pengaruh pada risiko *stunting*, terkait dengan kondisi fisik ibu serta kemampuannya memenuhi nutrisi keluarga (Kurniawati dkk., 2022). Di samping itu, status sosial-ekonomi keluarga dan intervensi pemerintah melalui bantuan sosial tetap relevan dalam mitigasi risiko *stunting*. Khususnya bagi keluarga prasejahtera yang memiliki keterbatasan akses layanan kesehatan dan gizi (Munawaroh dkk., 2024; Titaley dkk., 2008).

Berdasarkan uraian tersebut, penelitian ini menganalisis faktor risiko *stunting* di Kabupaten Tana Toraja menggunakan model regresi logistik biner dengan pendekatan *spline truncated* dan CATPCA. Variabel dependen yang digunakan adalah Keluarga Berisiko *Stunting*, yang berskala biner ("Ya" atau "Tidak"). Penelitian sebelumnya, seperti yang dilakukan oleh Auliyah (2021) telah menggunakan regresi logistik CATPCA dalam mengatasi masalah multikolinieritas pada data kategorik, namun belum mengintegrasikan pendekatan nonparametrik untuk meningkatkan kinerja model. Penelitian lain oleh Arifin dkk. (2023) dan Naim (2020) mengembangkan regresi logistik dengan estimator *spline truncated* pada kasus status gizi balita, tetapi belum sepenuhnya menangani potensi multikolinieritas pada data. Oleh karena itu, penelitian ini diharapkan dapat mengisi kesenjangan tersebut dengan mengembangkan model regresi logistik biner menggunakan CATPCA untuk mengatasi multikolinieritas dan estimator *spline truncated*, guna memberikan prediksi yang lebih akurat terkait risiko *stunting* di Kabupaten Tana Toraja.

1.2 Tujuan dan Manfaat

Berdasarkan latar belakang, maka diperoleh tujuan penelitian sebagai berikut:

1. Mengestimasi parameter model regresi logistik biner CATPCA dengan estimator *spline truncated* pada data yang mengandung multikolinieritas.
2. Memodelkan hubungan antara keluarga berisiko *stunting* dengan faktor-faktor yang memengaruhinya di Kabupaten Tana Toraja tahun 2023 menggunakan model regresi logistik biner CATPCA dengan estimator *spline truncated*.

Adapun penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:
1. Memberikan wawasan baru dan memperkaya literatur ilmiah dalam penerapan model regresi logistik biner CATPCA dengan estimator *spline truncated*.
2. Menyediakan model prediksi yang lebih akurat dan tahan terhadap pelanggaran multikolinieritas, sehingga lebih stabil untuk menganalisis faktor risiko *stunting* dan dapat digunakan sebagai dasar dalam perumusan kebijakan kesehatan di Kabupaten Tana Toraja.

1.3 Batasan Masalah

Untuk membatasi ruang lingkup permasalahan pada penelitian ini, maka diberikan beberapa batasan bahwa:
1. Data yang digunakan adalah data keluarga berisiko *stunting* beserta sembilan variabel yang diduga memengaruhinya di Kabupaten Tana Toraja pada tahun 2023.
2. Metode penaksiran parameter yang digunakan adalah *Maximum Likelihood Estimation* (MLE).
3. Pemodelan *spline truncated* dalam regresi logistik biner hanya dibatasi pada orde satu (linier) dengan satu hingga tiga titik knot.
4. Pemilihan titik knot optimal dipilih berdasarkan metode *Generalized Cross-Validation* (GCV).

1.4 Teori

1.4.1 Multikolinieritas

Multikolinieritas dalam analisis regresi merupakan kondisi ketika dua atau lebih variabel independen memiliki hubungan linier yang kuat. Kondisi ini dapat memengaruhi akurasi model regresi dan hasil estimasi koefisien yang tidak tepat. O'Brien (2007) menjelaskan bahwa multikolinieritas dapat terjadi akibat pemilihan variabel yang tidak tepat atau pengukuran yang tidak akurat. Dampak dari multikolinieritas adalah meningkatnya varians dari koefisien regresi, sehingga mengurangi presisi dan mengaburkan signifikansi uji (Gujarati & Porter, 2008).

Salah satu cara yang umum digunakan untuk mendeteksi multikolinieritas adalah dengan menggunakan matriks korelasi. Metode ini mengukur tingkat hubungan antar variabel menggunakan koefisien korelasi Pearson yang nilainya berkisar antara -1 dan 1. Nilai yang mendekati 1 menunjukkan adanya hubungan positif linier yang sangat kuat, sedangkan nilai yang mendekati -1 menunjukkan hubungan negatif linier yang sangat kuat. Sementara itu, ketika nilainya mendekati 0, maka menunjukkan hubungan linier yang lemah (Jeng, 2023). Nilai koefisien korelasi Pearson (r) dapat diperoleh melalui Persamaan (1).

$$r_{jj^*} = \frac{n\left(\sum x_j x_{j^*}\right) - \left(\sum x_j\right)\left(\sum x_{j^*}\right)}{\sqrt{\left[n\sum x_j^2 - \left(\sum x_j\right)^2\right]\left[n\sum x_{j^*}^2 - \left(\sum x_{j^*}\right)^2\right]}};\ j = 1,2,\dots,p \text{ dan } j^* = 1,2,\dots,p \quad (1)$$

dengan r_{jj^*} adalah koefisien korelasi Pearson antara variabel independen x_j dan x_{j^*}, n adalah jumlah amatan data, x_j adalah variabel independen ke-j, dan x_{j^*} adalah variabel independen ke-j^*, serta p adalah jumlah variabel independen dalam model.

Setelah menghitung nilai koefisien korelasi untuk setiap pasangan variabel independen, nilai-nilai tersebut dapat direpresentasikan dalam bentuk matriks korelasi $\boldsymbol{R}$ seperti berikut:

$$\boldsymbol{R} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$$

Matriks korelasi $\boldsymbol{R}$ menunjukkan korelasi antara setiap pasangan variabel independen dalam model. Sebagai contoh, elemen r_{12} mewakili koefisien korelasi antara variabel x_1 dan x_2, sementara diagonal utama matriks ini selalu bernilai 1 karena variabel yang sama memiliki korelasi sempurna, yaitu ketika $j = j^*$, maka $r_{jj^*} = 1$. Interpretasi dari nilai r mengikuti kriteria yang terdapat dalam Tabel 1.

Tabel 1. Interpretasi Nilai Koefisien Korelasi Pearson

Interval Koefisien	Tingkat Hubungan
0,0000 – 0,1999	Sangat Lemah
0,2000 – 0,3999	Lemah
0,4000 – 0,5999	Sedang
0,6000 – 0,7999	Kuat
0,8000 – 1,0000	Sangat Kuat

Secara umum, jika nilai absolut dari koefisien korelasi Pearson tergolong kuat atau bahkan sangat kuat, maka potensi terjadinya multikolinieritas juga semakin besar (Kim, 2019). Namun, untuk memastikan bahwa korelasi spesifik antar variabel independen bersifat signifikan secara statistik, maka perlu dilakukan uji signifikansi terhadap koefisien korelasi Pearson. Hipotesis yang digunakan dalam uji ini adalah sebagai berikut:

H_0: $\rho_{jj^*} = 0$, untuk $j \neq j^*$

H_1: $\rho_{jj^*} \neq 0$, untuk $j \neq j^*$

Uji signifikansi dilakukan dengan menghitung nilai statistik uji t yang diperoleh melalui Persamaan (2):

$$t_{jj^*} = \frac{r_{jj^*}\sqrt{n-2}}{\sqrt{1 - r_{jj^*}^2}} \quad (2)$$

Nilai statistik uji t ini kemudian dibandingkan dengan nilai kritis dari distribusi t dengan $df = n - 2$ dan α yang telah ditetapkan. Jika nilai statistik uji $|t_{jj^*}| \geq t_{tabel\left(\frac{\alpha}{2};df\right)}$ atau $p - value \leq \alpha$, maka H_0 ditolak. Hal ini menandakan bahwa korelasi antar variabel independen tersebut signifikan (Sanny & Dewi, 2020).

1.4.2 Nilai Eigen dan Vektor Eigen

Analisis regresi memanfaatkan nilai eigen serta vektor eigen dari matriks kovarians atau matriks korelasi. Misalkan matriks $\boldsymbol{R}$ berukuran $p \times p$, kemudian terdapat skalar λ (nilai eigen), dan vektor tak nol $\boldsymbol{v}$ (vektor eigen) yang memenuhi Persamaan (3).

$$\begin{aligned} \boldsymbol{R}\boldsymbol{v} &= \lambda\boldsymbol{v} \\ \boldsymbol{R}\boldsymbol{v} - \lambda\boldsymbol{v} &= 0 \\ (\boldsymbol{R} - \lambda\boldsymbol{I})\boldsymbol{v} &= 0 \end{aligned} \tag{3}$$

Persamaan (3) menunjukkan bahwa vektor eigen $\boldsymbol{v}$ adalah vektor yang arahnya tidak berubah ketika dikalikan dengan matriks $\boldsymbol{R}$, meskipun panjangnya dapat berubah sesuai dengan nilai eigen λ. Agar Persamaan (3) memiliki solusi tak nol untuk $\boldsymbol{v}$, maka determinan dari $(\boldsymbol{R} - \lambda\boldsymbol{I})$ harus sama dengan nol, sehingga diperoleh persamaan karakteristik seperti pada Persamaan (4) (Abdi & Williams, 2010).

$$\det(\boldsymbol{R} - \lambda\boldsymbol{I}) = 0 \tag{4}$$

Solusi dari Persamaan (4) akan menghasilkan akar-akar karakteristik yang merupakan nilai eigen $\lambda_1, \lambda_2, \ldots, \lambda_p$ dengan $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$. Setelah semua nilai eigen diperoleh, vektor eigen $\boldsymbol{v}_j$ untuk masing-masing komponen utama yang sesuai dapat diperoleh dengan menyelesaikan sistem persamaan linier $\left(\boldsymbol{R} - \lambda_j\boldsymbol{I}\right)\boldsymbol{v}_j = 0$ (Hidayatullah dkk., 2024).

1.4.3 *Categorical Principal Component Analysis*

Categorical Principal Component Analysis (CATPCA) merupakan bentuk nonlinier dari *Principal Component Analysis* (PCA) yang digunakan pada data kategorik. Metode ini bekerja dengan mentransformasi variabel kategorik menjadi variabel numerik melalui proses *optimal scaling* (kuantifikasi optimal). Transformasi ini memungkinkan analisis komponen utama dilakukan bahkan pada data kategorik dengan tujuan memaksimalkan varians antar variabel (Linting dkk., 2007).

1.5.3.1 Kuantifikasi Optimal

Proses kuantifikasi optimal dalam CATPCA dilakukan dengan menggunakan algoritma *Alternating Least Squares* (ALS). Algoritma ini bekerja dengan memperbarui nilai parameter secara bergantian pada setiap iterasi hingga mencapai solusi yang meminimalkan *loss function* (fungsi kerugian). Misalkan (θ_1, θ_2) adalah parameter yang digunakan dalam fungsi kerugian $\sigma(\theta_1, \theta_2)$. Pada iterasi ke-$t$, estimasi parameter tersebut dinyatakan sebagai $\theta^{(t)}$. Parameter $\theta_1^{(t+1)}$ dan $\theta_2^{(t+1)}$ akan diperbarui dengan memperhitungkan informasi dari iterasi sebelumnya, sebagaimana dijelaskan dalam Persamaan (5).

$$\theta_1^{(t+1)} = \underset{\theta_1}{\operatorname{argmin}} \sigma\left(\theta_1, \theta_2^{(t)}\right)$$
$$\theta_2^{(t+1)} = \underset{\theta_2}{\operatorname{argmin}} \sigma\left(\theta_1^{(t+1)}, \theta_2\right) \quad (5)$$

Persamaan (5) menunjukkan bahwa pada setiap iterasi, algoritma ALS akan memperbarui kedua parameter secara bergantian. Langkah pertama adalah memperbarui θ_1 yang dilakukan dengan cara mencari nilai $\theta_1^{(t+1)}$ yang meminimalkan $\sigma\left(\theta_1, \theta_2^{(t)}\right)$, dengan $\theta_2^{(t)}$ adalah nilai parameter θ_2 yang digunakan pada iterasi sebelumnya. Setelah memperbarui θ_1, proses dilanjutkan dengan pembaruan θ_2 yang dilakukan dengan meminimalkan fungsi kerugian $\sigma\left(\theta_1^{(t+1)}, \theta_2\right)$, dengan $\theta_1^{(t+1)}$ merupakan nilai parameter yang telah diperbarui pada langkah sebelumnya.

Proses ini akan berulang hingga perubahan antara dua iterasi berturut-turut menjadi lebih kecil atau sama dengan ambang toleransi yang telah ditentukan, seperti $\varepsilon = 10^{-5}$, yang menandakan bahwa konvergensi telah tercapai. Melalui penggunaan algoritma ALS, nilai kuantifikasi optimal yang diperoleh memungkinkan analisis komponen utama dilakukan pada data hasil transformasi. Analisis ini bertujuan untuk mereduksi dimensi data dengan tetap mempertahankan sebanyak mungkin informasi varians antar kategori variabel. (Mori dkk., 2016).

1.5.3.2 Pembentukan Komponen Utama

Misalkan terdapat n amatan dan p variabel kategorik yang direpresentasikan oleh $\boldsymbol{X}$ berukuran $n \times p$. Setiap variabel kategorik didefinisikan sebagai $\boldsymbol{X}_j = \boldsymbol{G}_j \boldsymbol{C}_j$ untuk $j = 1,2,\dots,p$. Adapun fungsi kerugian yang digunakan sesuai dengan Persamaan (6).

$$\sigma\left(\boldsymbol{X}, \boldsymbol{C}_j\right) = \frac{1}{n} \sum_{j=1}^{p} \frac{1}{j} \left\{ tr\left(\left(\boldsymbol{X} - \boldsymbol{G}_j \boldsymbol{C}_j\right)' \left(\boldsymbol{X} - \boldsymbol{G}_j \boldsymbol{C}_j\right) \right) \right\} \quad (6)$$

dengan $\boldsymbol{G}_j = [\boldsymbol{g}_{j1} \quad \boldsymbol{g}_{j2} \quad \cdots \quad \boldsymbol{g}_{jv_j}]$ merupakan matriks indikator berukuran $n \times v_j$ (v_j adalah jumlah kategori variabel $\boldsymbol{X}_j$) dan $\boldsymbol{C}_j$ merupakan matriks hasil kuantifikasi kategorik setiap $\boldsymbol{X}_j$ berukuran $v_j \times p$.

Setelah melalui proses kuantifikasi dengan algoritma ALS sesuai pada Persamaan (5), maka kuantifikasi optimal $\boldsymbol{X}_j^*$ dihitung melalui Persamaan (7).

$$\boldsymbol{X}_j^* = \boldsymbol{G}_j \boldsymbol{C}_j^* \quad (7)$$

Selanjutnya, dibentuk matriks korelasi $\boldsymbol{R}$ yang mewakili hubungan antara variabel-variabel hasil kuantifikasi. Jika $\boldsymbol{X}^*$ telah distandarisasi, maka $\boldsymbol{R}$ dapat dinyatakan dengan Persamaan (8).

$$\boldsymbol{R} = \frac{1}{n-1} \boldsymbol{X}^{*\prime} \boldsymbol{X}^* \quad (8)$$

dengan $\boldsymbol{X}^*$ adalah matriks hasil kuantifikasi. Matriks $\boldsymbol{R}$ digunakan sebagai dasar untuk menentukan komponen utama dalam CATPCA. Komponen utama ini diperoleh melalui nilai eigen λ dan vektor eigen $\boldsymbol{v}$ dari matriks korelasi tersebut. Pasangan nilai eigen dan vektor eigen, yaitu $(\lambda_1, \boldsymbol{v}_1), (\lambda_2, \boldsymbol{v}_2), \dots, (\lambda_p, \boldsymbol{v}_p)$ membentuk komponen utama yang diperoleh dari variabel hasil kuantifikasi.

Setiap komponen utama $\boldsymbol{Z}_k^*$ merupakan kombinasi linier dari variabel hasil kuantifikasi yang didefinisikan oleh Persamaan (9).

$$\boldsymbol{Z}_k^* = \boldsymbol{v}_k' \boldsymbol{X}^* = v_{1k}\boldsymbol{X}_1^* + v_{2k}\boldsymbol{X}_2^* + \cdots + v_{pk}\boldsymbol{X}_p^*, \qquad k = 1,2,\ldots,p \tag{9}$$

atau dalam bentuk matriks:

$$\boldsymbol{v}_k' \boldsymbol{X}^* = [v_{1k} \quad v_{2k} \quad \cdots \quad v_{pk}] \begin{bmatrix} \boldsymbol{X}_1^* \\ \boldsymbol{X}_2^* \\ \vdots \\ \boldsymbol{X}_p^* \end{bmatrix}$$

Komponen utama ini sering disebut sebagai variabel komposit yang berfungsi memaksimalkan varians dalam data. Misalnya, komponen utama pertama $\boldsymbol{Z}_1^*$ memaksimalkan $\boldsymbol{v}_1' \boldsymbol{R} \boldsymbol{v}_1 = \lambda_1$, lalu komponen utama berikutnya $\boldsymbol{Z}_2^*$ maksimalkan varians setelah $\boldsymbol{Z}_1^*$, dan seterusnya hingga $\boldsymbol{Z}_p^*$ tercapai (Yamagishi dkk., 2019).

Pemilihan jumlah komponen optimal pada CATPCA didasarkan pada proporsi kumulatif varians yang dijelaskan oleh komponen tersebut. Komponen optimal dipilih apabila proporsi kumulatif varians yang terjelaskan mencapai atau melebihi batas yang ditentukan. Proporsi kumulatif varians dihitung dengan menggunakan Persamaan (10), yaitu dengan menjumlahkan varians yang dijelaskan oleh tiap komponen utama terpilih dibandingkan dengan total varians keseluruhan.

$$Persentase\ varians\ kumulatif\ ke-m = \frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \times 100\% \tag{10}$$

dengan

λ_j : nilai eigen komponen utama ke-j,

m : komponen utama yang dipilih (dengan $m \leq p$),

p : total komponen utama yang mungkin.

Jika persentase kumulatif ini telah mencapai atau melebihi 80%, maka komponen utama tersebut sudah dianggap cukup memadai untuk menjelaskan variabilitas utama dalam data (Jolliffe, 2002). Meskipun beberapa peneliti menggunakan kriteria bahwa hanya komponen utama dengan $\lambda_j > 1$ yang dipilih, tidak ada pedoman teoretis yang secara jelas menentukan kriteria mana yang paling tepat. Oleh karena itu, pemilihan kriteria didasarkan pada pertimbangan praktis masing-masing, seperti hasil analisis, kemudahan interpretasi, dan relevansi dalam bidang yang dikaji (Astutik dkk., 2018).

1.5.3.3 Nilai *Loading*

Nilai *loading* dalam CATPCA menunjukkan besarnya korelasi antara variabel asli dan komponen utama yang terbentuk, sehingga memberikan informasi tentang hubungan linier antara keduanya. Semakin besar nilai *loading*, maka semakin besar kontribusi variabel tersebut dalam membentuk komponen utama. Variabel dengan nilai *loading* absolut $\geq 0{,}5000$ dianggap berpengaruh signifikan terhadap komponen utama yang terkait (Mayapada dkk., 2019).

Perhitungan nilai *loading* dapat dilakukan dengan menggunakan vektor eigen dan nilai eigen yang dapat dituliskan dalam bentuk representasi matriks, seperti pada Persamaan (11) (Enzellina & Suhaedi, 2022).

$$\boldsymbol{L} = \boldsymbol{V}\boldsymbol{\Lambda}^{\frac{1}{2}} \tag{11}$$

dengan

$\boldsymbol{L}$: matriks *loading* berukuran $p \times m$

$\boldsymbol{V}$: matriks berisi vektor eigen berukuran $p \times m$

$\boldsymbol{\Lambda}^{\frac{1}{2}}$: matriks diagonal berisi akar kuadrat dari nilai eigen.

Matriks $\boldsymbol{L}$ menunjukkan kontribusi masing-masing variabel pada setiap komponen utama yang dibentuk.

1.4.4 Regresi Logistik Biner

Regresi logistik biner merupakan suatu model statistik yang digunakan untuk menganalisis hubungan antara satu atau lebih variabel independen dengan variabel dependen yang bersifat biner atau dikotomus, seperti 0 dan 1, atau "ya" dan "tidak". Model ini bertujuan untuk memprediksi peluang terjadinya suatu peristiwa berdasarkan kombinasi linier dari variabel-variabel independen yang dapat berupa variabel numerik maupun kategorikal (Hosmer dkk., 2013).

Misalkan terdapat $x_{i1}, x_{i2}, \dots, x_{ip}$ yang merupakan nilai dari p variabel independen pada amatan ke-i. Sementara nilai variabel dependen untuk setiap pengamatan, yaitu $y_1, y_2, \dots, y_n$, dengan n mewakili jumlah amatan. Dalam keadaan demikian, maka variabel Y mengikuti distribusi Bernoulli untuk setiap amatan tunggal dengan fungsi peluang sesuai pada Persamaan (12).

$$f(y_i) = \pi(x_i)^{y_i}\big(1 - \pi(x_i)\big)^{1-y_i} \tag{12}$$

dengan $\pi(x_i)$ merupakan peluang sukses $(Y_i = 1|x_i)$ untuk amatan ke-i. Peluang $\pi(x_i)$ bergantung pada kombinasi linier dari variabel independen $x_{i1}, x_{i2}, \dots x_{ip}$, yang dinyatakan pada Persamaan (13).

$$\pi(x_i) = P\big(Y_i = 1\big|x_{i1}, x_{i2}, \dots, x_{ip}\big) = \frac{e^{(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip})}}{1 + e^{(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip})}}$$

$$\pi(x_i) = \frac{e^{\left(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right)}}{1 + e^{\left(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right)}} \tag{13}$$

Persamaan (13) merupakan model yang digunakan untuk memprediksi kejadian biner berdasarkan variabel independen (Agresti, 2012). Selanjutnya, untuk memperoleh model yang linier terhadap parameter, dilakukan transformasi menggunakan fungsi *logit* yang merupakan logaritma dari rasio peluang (*log-odds*), sehingga diperoleh Persamaan (14).

$$g(x_i) = logit[\pi(x_i)] = \ln\left[\frac{\pi(x_i)}{1 - \pi(x_i)}\right] \tag{14}$$

Selanjutnya, model pada Persamaan (14) disubtitusikan ke dalam Persamaan (13), sehingga diperoleh Persamaan (15).

$$g(x_i) = \ln\left[\frac{\dfrac{e^{\left(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right)}}{1+e^{\left(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right)}}}{1-\dfrac{e^{\left(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right)}}{1+e^{\left(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right)}}}\right] = \ln\left[e^{\left(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right)}\right]$$

$$g(x_i) = \beta_0 + \sum_{j=1}^{p}\beta_j x_{ij} \tag{15}$$

Dengan demikian, transformasi *logit* ini memungkinkan hubungan linier antara peluang sukses $\pi(x_i)$ dan variabel independen $x_1, x_2, \dots x_p$, dengan parameter β_0 sebagai intersep, serta $\beta_1, \beta_2, \dots, \beta_p$ sebagai koefisien regresi (Hosmer dkk., 2013). Melalui pendekatan ini, regresi logistik biner tidak hanya memberikan estimasi terjadi atau tidaknya suatu kejadian, tetapi juga mengukur seberapa besar pengaruh variabel independen terhadap peluang tersebut (Agresti, 2012).

1.4.5 *Maximum Likelihood Estimation*

Maximum Likelihood Estimation (MLE) digunakan dalam regresi logistik biner untuk mengestimasi parameter model $\boldsymbol{\beta}$ dengan cara memaksimalkan kemungkinan (*likelihood)* bahwa model yang dibangun sesuai dengan data yang diamati (Dobson & Barnett, 2018). Jika x_i dan y_i adalah pasangan variabel independen dan dependen pada amatan ke-i, serta setiap pasangan pengamatan diasumsikan saling bebas dengan $i = 1,2, \dots, n$, maka distribusinya mengikuti distribusi Bernoulli dengan fungsi peluang untuk setiap pasangannya dapat ditunjukkan pada Persamaan (12). Maka, diperoleh fungsi *likelihood* $L(\boldsymbol{\beta})$ yang ditunjukkan dalam Persamaan (16).

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} P(Y = y_i) = \prod_{i=1}^{n} \pi(x_i)^{y_i}\left(1-\pi(x_i)\right)^{1-y_i} \tag{16}$$

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n}\left[\left(1+e^{\left(\sum_{j=0}^{p}\beta_j x_{ij}\right)}\right)^{-1} e^{\left(\sum_{j=0}^{p}\left(\sum_{i=1}^{n} y_i x_{ij}\right)\beta_j\right)}\right]$$

Fungsi *likelihood* pada Persamaan (16) akan lebih mudah untuk dimaksimumkan apabila diubah ke dalam bentuk fungsi *log-likelihood* yang dinyatakan dengan $\ell(\boldsymbol{\beta})$.

$$\ell(\boldsymbol{\beta}) = \ln\left(L(\boldsymbol{\beta})\right) = \ln\left[\prod_{i=1}^{n} \pi(x_i)^{y_i}\left(1-\pi(x_i)\right)^{1-y_i}\right]$$

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left[y_i \ln \pi(x_i) + (1-y_i)\ln\left(1-\pi(x_i)\right)\right] \tag{17}$$

Langkah selanjutnya adalah menurunkan Persamaan (17) terhadap parameter $\boldsymbol{\beta}$ dan menyetarakannya dengan nol untuk menemukan estimasi parameter $\boldsymbol{\beta}$ yang memaksimalkan fungsi *likelihood*. Turunan pertama dari Persamaan (17) ditunjukkan pada Persamaan (18).

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n}\left[\frac{y_i}{\pi(x_i)}\frac{\partial \pi(x_i)}{\partial \beta_j} - \frac{1-y_i}{1-\pi(x_i)}\frac{\partial\big(1-\pi(x_i)\big)}{\partial \beta_j}\right] \tag{18}$$

Turunan dari $\pi(x_i)$ terhadap β_j dapat dihitung menggunakan rumus dari fungsi sigmoid $\frac{\partial \pi(x_i)}{\partial \beta_j} = \pi(x_i)\big(1-\pi(x_i)\big)x_{ij}$. Sehingga, diperoleh turunan lengkapnya seperti pada Persamaan (19).

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n}\big(y_i - \pi(x_i)\big)x_{ij} = 0 \tag{19}$$

Persamaan (19) menunjukkan bahwa bentuk fungsi *log-likelihood* ini cukup kompleks untuk diselesaikan secara analitik, sehingga digunakan metode numerik seperti *Newton-Raphson* untuk mencari nilai $\boldsymbol{\beta}$ yang memaksimalkan *log-likelihood* (Salam dkk., 2021). Metode ini memperbarui parameter secara iteratif dengan menggunakan turunan pertama (*gradient*) dan turunan kedua (*Hessian matrix*) dari fungsi *log-likelihood*.

Proses iteratif ini dimulai dengan tebakan awal untuk parameter $\boldsymbol{\beta}$, kemudian menghitung *gradient* dan *Hessian matrix* untuk memperbarui estimasi parameter dengan menggunakan Persamaan (20).

$$\widehat{\boldsymbol{\beta}}^{(t+1)} = \widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{H}^{-1}\boldsymbol{g} \tag{20}$$

dengan

$\boldsymbol{H}$: *Hessian matrix*,

$\boldsymbol{g}$: *gradient* dari *log-likelihood*.

Proses ini berlanjut hingga konvergensi tercapai, yaitu ketika $\left|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\right| \leq 10^{-5}$. Dengan demikian, nilai β yang memaksimalkan *log-likelihood* memberikan estimasi terbaik untuk parameter model. Namun, perlu diketahui bahwa metode tersebut digunakan untuk mendekati solusi maksimum *likelihood*, bukan solusi eksak secara langsung (Wooldridge, 2010).

1.4.6 Pengujian Signifikansi Parameter

Pengujian signifikansi parameter penting untuk menentukan signifikansi pengaruh dari variabel independen terhadap peluang terjadinya suatu kejadian biner. Pengujian ini dilakukan untuk mengevaluasi signifikansi pengaruh parameter β_j dari variabel independen X_j terhadap variabel dependen biner.

1.5.6.1 Pengujian Signifikansi Parameter secara Simultan

Uji simultan dalam regresi logistik menggunakan uji *Likelihood Ratio* (LR) untuk menguji siginifikansi keseluruhan model. Uji ini membandingkan model penuh (dengan variabel independen) terhadap model terbatas (tanpa variabel independen). 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 D yang merupakan perbandingan antara *log-likelihood* dari model penuh dan model terbatas sesuai pada Persamaan (21).

$$D = -2 \times \left(\ln L_{terbatas} - \ln L_{penuh}\right) \tag{21}$$

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 (df) sama dengan jumlah variabel independen dalam model (Hidayat & Hajarisman, 2023). Sementara itu, kriteria penolakannya adalah ketika nilai $D > \chi^2_{p;\alpha}$ atau $p - value < \alpha$, maka H_0 ditolak yang berarti bahwa model dengan variabel independen lebih baik dalam menjelaskan data.

1.5.6.2 Pengujian Signifikansi Parameter secara Parsial

Pengujian signifikansi tiap 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 dependen secara parsial dengan mengikuti hipotesis berikut:

$H_0: \beta_j = 0$

$H_1: \beta_j \neq 0$ untuk $j = 1,2, \dots p$

Adapun statistik uji Wald dinyatakan pada Persamaan (22).

$$W_j = \frac{\hat{\beta}_j^2}{SE\left(\hat{\beta}_j^2\right)} \tag{22}$$

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 adalh menolak H_0 yang berarti X_j berpengaruh signifikan terhadap variabel dependen secara parsial (Hosmer dkk., 2013).

1.4.7 Regresi Nonparametrik

Regresi nonparametrik digunakan untuk memodelkan hubungan antara variabel dependen dan independen ketika pola hubungan tersebut tidak diketahui secara jelas. Berbeda dengan regresi parametrik yang memerlukan asumsi tertentu mengenai bentuk fungsi hubungan, regresi nonparametrik bersifat fleksibel karena bentuk kurva regresinya ditentukan oleh data, bukan berdasarkan asumsi bentuk tertentu (Sifriyani dkk., 2023). Hal ini membuat metode regresi nonparametrik menjadi berguna dalam kasus dengan pola hubungan kompleks. Secara umum, model regresi nonparametrik dapat dituliskan seperti pada Persamaan (23).

$$y_i = f(x_i) + \varepsilon_i, \;\; i = 1,2, \dots, n \tag{23}$$

dengan

y_i : variabel dependen pada amatan ke-i,

$f(x_i)$: fungsi regresi yang dihampiri,

x_i : variabel independen pada amatan ke-i,

ε_i : *error* pada amatan ke-i

1.5.7.1 Estimator *Spline truncated*

Spline truncated merupakan salah satu metode regresi nonparametrik yang paling umum digunakan karena kemampuannya dalam menangani perubahan perilaku data pada sub-interval tertentu. Metode ini memanfaatkan titik-titik *knot* untuk mendeteksi perubahan pola dan menyesuaikan kurva regresi secara lokal (Dani dkk., 2021). Fungsi kurva regresi $f(x_i)$ dalam model *spline truncated* untuk satu variabel independen dengan titik *knot* $K_1, K_2, \dots, K_r$ dapat dinyatakan melalui Persamaan (24).

$$f(x_i) = \sum_{k=0}^{q} \beta_k x_i^k + \sum_{h=1}^{r} \beta_{q+h} (x_i - K_h)_+^q \tag{24}$$

Model regresi *spline truncated* yang dinyatakan melalui Persamaan (25) dapat diperoleh melalui subtitusi Persamaan (24) ke dalam Persamaan (23).

$$y_i = \sum_{k=0}^{q} \beta_k x_i^k + \sum_{h=1}^{r} \beta_{q+h} (x_i - K_h)_+^q + \varepsilon_i, \qquad i = 1,2,\dots,n \tag{25}$$

dengan $(x_i - K_h)_+^q$ disebut sebagai fungsi *truncated* yang didefinisikan sebagai:

$$(x_i - K_h)_+^q = \begin{cases} (x_i - K_h)_+^q, & x \geq K_h \\ 0 \quad , & x < K_h \end{cases}$$

Parameter β_j adalah koefisien polinomial, sedangkan β_{q+h} adalah koefisien pada komponen *truncated*. Sementara itu, r menunjukkan jumlah titik *knot*, K_h menunjukkan posisi titik *knot* ke-h, dan nilai q adalah derajat polinomial.

Model regresi *spline truncated* untuk n data pengamatan sesuai pada Persamaan (26) dapat dinyatakan dalam notasi matriks yang ditunjukkan sebagai berikut:

$$\boldsymbol{y} = \boldsymbol{X}[K]\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{26}$$

dengan bentuk eksplisit:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21}^2 & \cdots & x_{p1}^q & (x_{11} - K_{11})_+^q & \cdots & (x_{p1} - K_{pr})_+^q \\ 1 & x_{12} & x_{22}^2 & \cdots & x_{p2}^q & (x_{12} - K_{11})_+^q & \cdots & (x_{p2} - K_{pr})_+^q \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n}^2 & \cdots & x_{pn}^q & (x_{1n} - K_{11})_+^q & \cdots & (x_{pn} - K_{pr})_+^q \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_q \\ \beta_{(q+1)} \\ \vdots \\ \beta_{q+r} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Estimasi parameter $\boldsymbol{\beta}$ pada Persamaan (26) dapat dilakukan dengan menggunakan metode *Least Square* ataupun MLE.

1.5.7.2 Pemilihan Titik Knot Optimal

Pemilihan model terbaik merupakan tahapan penting dalam regresi nonaprametrik. Untuk regresi *spline truncated*, kriteria yang sering digunakan adalah nilai *Generalized Cross Validation* (GCV). Metode GCV bersifat optimal asimtotik, tidak bergantung pada varians populasi ($\boldsymbol{\sigma^2}$) yang tidak diketahui, dan *invariance* terhadap

transformasi (Xiang & Wahba, 1996). Pemilihan model terbaik dilakukan dengan nilai GCV yang paling minimum. Fungsi GCV untuk pemilihan titik *knot* optimal pada model regresi *spline truncated* ditunjukkan dalam Persamaan (27).

$$GCV(K) = \frac{MSE}{\left[n^{-1}tr\left(\boldsymbol{I} - \boldsymbol{A}(K)\right)\right]^2} \tag{27}$$

dengan

n : jumlah amatan,

K : titik knot,

$\boldsymbol{I}$: matriks identitas,

Adapun $\boldsymbol{A}(K) = \boldsymbol{X}[K](\boldsymbol{X}'[K]\boldsymbol{X}[K])^{-1}\boldsymbol{X}'[K]$ merupakan matriks estimasi dan *Mean Squared Error* (MSE) diperoleh melalui $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Pemilihan titik *knot* optimal menggunakan GCV telah terbukti efektif dalam berbagai model *spline* untuk meminimalkan MSE, sehingga menghasilkan model yang lebih akurat (Pramudita dkk., 2024).

1.4.8 Ketepatan Klasifikasi Model

Dalam regresi logistik, ketepatan klasifikasi dapat dievaluasi dengan menggunakan *confusion matrix*. Pada kejadian biner, metode ini menganalisis hasil klasifikasi model dengan merangkum prediksi ke dalam empat kategori sesuai pada Tabel 2.

Tabel 2. *Confusion Matrix*

Aktual	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).

Confusion matrix memberikan dasar bagi perhitungan metrik performa model, seperti akurasi. Metrik ini merupakan proporsi prediksi yang benar terhadap total prediksi (Obi, 2023). 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 (28).

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \tag{28}$$

Akurasi yang tinggi mengindikasikan bahwa model bekerja dengan baik secara keseluruhan dalam mengklasifikasikan data. Namun, dalam kasus *dataset* yang 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 (Handoyo dkk., 2021).

1.4.9 Keluarga Berisiko *Stunting*

Stunting merupakan kondisi gagal tumbuh pada anak balita akibat kekurangan gizi kronis selama periode seribu Hari Pertama Kehidupan (1000 HPK). Kondisi ini tidak hanya menyebabkan anak memiliki tinggi badan di bawah rata-rata usianya, tetapi juga menghambat perkembangan kemampuan motorik dan kognitif mereka. Selain itu, anak yang *stunting* berisiko lebih tinggi mengalami masalah kesehatan di masa dewasa, seperti penyakit jantung dan diabetes, yang berdampak negatif pada produktivitas individu serta kesehatan masyarakat secara luas (World Health Organization, 2015) (Black dkk., 2013).

Faktor lingkungan menjadi salah satu penyebab utama risiko *stunting*. Lingkungan yang tidak higienis serta keterbatasan akses terhadap air minum aman dan fasilitas sanitasi yang memadai meningkatkan paparan anak terhadap infeksi. Infeksi pada balita, terutama yang disebabkan oleh lingkungan buruk, dapat mengganggu penyerapan nutrisi dalam tubuh. Kondisi ini menyebabkan anak kesulitan mencapai status gizi yang ideal, sehingga meningkatkan kemungkinan terjadinya *stunting* (Victora dkk., 2021).

Karakteristik demografis keluarga juga memengaruhi risiko *stunting*. Salah satu pendekatan untuk memahami faktor risiko ini adalah melalui identifikasi Pasangan Usia Subur (PUS). Menurut BKKBN (2019), PUS didefinisikan sebagai pasangan suami-istri dengan istri berusia 15–49 tahun yang masih haid atau di bawah 15 tahun tetapi sudah haid. Keluarga dengan PUS yang terlalu muda atau tua, memiliki banyak anak, atau jarak kelahiran yang terlalu dekat lebih rentan terhadap *stunting* karena keterbatasan sumber daya. Namun, keluarga tanpa PUS juga dapat menghadapi risiko *stunting*, terutama jika faktor lain seperti kesejahteraan ekonomi dan akses terhadap layanan kesehatan kurang memadai (Mulyaningsih dkk., 2021).

Faktor ekonomi keluarga menjadi aspek penting dalam menentukan risiko *stunting*. Keluarga dengan tingkat kesejahteraan rendah sering kali kesulitan memenuhi kebutuhan dasar anak, terutama asupan gizi dan layanan kesehatan. Untuk mengurangi risiko ini, berbagai program pendampingan seperti bantuan sosial dirancang untuk meningkatkan kesejahteraan keluarga. Pendekatan ini bertujuan membantu keluarga yang kurang mampu memenuhi kebutuhan gizi anak sehingga dapat mencegah terjadinya *stunting* (Titaley dkk., 2008).

BAB II
METODE PENELITIAN

2.1 Jenis dan Sumber Data

Data yang digunakan dalam penelitian ini merupakan data sekunder, yaitu data Keluarga Berisiko *Stunting* di Kabupaten Tana Toraja pada tahun 2023. Data ini diperoleh dari Badan Kependudukan dan Keluarga Berencana Nasional (BKKBN) Perwakilan Sulawesi Selatan. Adapun data lengkapnya terdiri dari 23150 amatan yang dapat dilihat pada Lampiran 1.

2.2 Variabel Penelitian

Variabel yang digunakan dalam penelitian ini terdiri dari satu variabel dependen (y) dan sembilan variabel independen (x) yang dijelaskan pada Tabel 3.

Tabel 3. Variabel Penelitian

Variabel Penelitian	Nama Variabel	Kategori
y	Keluarga Berisiko *Stunting*	1 = Ya
		0 = Tidak
x_1	Keluarga Mempunyai Baduta	1 = Ya
		2 = Tidak
x_2	Kondisi Sumber Air Minum Utama	1 = Layak
		2 = Tidak Layak
x_3	Kondisi Sumber Fasilitas Buang Air Besar	1 = Layak
		2 = Tidak Layak
x_4	Keluarga Mempunyai PUS Terlalu Muda	1 = Ya
		2 = Tidak
		3 = Tidak Berlaku
x_5	Keluarga Mempunyai PUS Terlalu Tua	1 = Ya
		2 = Tidak
		3 = Tidak Berlaku
x_6	Keluarga Mempunyai PUS Terlalu Dekat	1 = Ya
		2 = Tidak
		3 = Tidak Berlaku
x_7	Keluarga Mempunyai PUS Terlalu Banyak	1 = Ya
		2 = Tidak
		3 = Tidak Berlaku
x_8	Peringkat Kesejahteraan Keluarga	0 = Peringkat Kesejahteraan > 4
		1 = Peringkat Kesejahteraan 1
		2 = Peringkat Kesejahteraan 2
		3 = Peringkat Kesejahteraan 3
		4 = Peringkat Kesejahteraan 4
		5 = Keluarga belum teridentifikasi tingkat kesejahteraannya
x_9	Keluarga Mendapatkan Pendampingan Bansos	1 = Ya
		2 = Tidak

Sumber: BKKBN (2023)

Sesuai dengan Tabel 3, variabel dependen yang digunakan berupa variabel kategorik biner, yaitu Keluarga Berisiko *Stunting* yang hanya terdiri atas dua kategori "Ya" dan "Tidak". Adapun sembilan variabel independennya merupakan faktor-faktor yang diduga memengaruhi Keluarga Berisiko *Stunting*.

2.3 Metode Analisis Data

Langkah-langkah analisis data yang dilakukan dalam penelitian ini melalui adalah sebagai berikut:

1. Mengestimasi parameter model regresi logistik biner CATPCA dengan estimator *spline truncated* dengan tahapan sebagai berikut:
   a. Mengonversi variabel independen kategorik menjadi numerik kontinu $\boldsymbol{X}^*$ melalui *optimal scaling* dengan meminimumkan fungsi kerugian sesuai Persamaan (6).
   b. Membentuk matriks korelasi dari variabel $\boldsymbol{X}^*$ yang telah dikuantifikasi sesuai Persamaan (7).
   c. Menghitung nilai eigen dan vektor eigen dari matriks korelasi yang terbentuk.
   d. Memilih sejumlah m komponen utama optimal berdasarkan persentase varians kumulatif $\geq 80\%$ dari variabel $\boldsymbol{X}^*$ sesuai pada Persamaan (10).
   e. Membentuk komponen utama $\boldsymbol{Z}_k^*$ menggunakan Persamaan (9).
   f. Menyatakan komponen utama $\boldsymbol{Z}_k^*$ ke dalam model regresi logistik biner dengan estimator *spline truncated*. Maka dari itu, diperoleh persamaan sebagai berikut:
   $$g(Z_i) = logit[\pi(Z_i^*)] = \beta_0 + \sum_{k=1}^{m} \beta_{k1} Z_{ki}^* + \sum_{k=1}^{m} \sum_{h=1}^{r} \beta_{k(1+h)} (Z_{ki}^* - K_h)_+$$
   g. Mengestimasi parameter β_0, β_{k1}, hingga $\beta_{k(1+h)}$ menggunakan metode MLE.

2. Memodelkan keluarga berisiko *stunting* di Kabupaten Tana Toraja pada tahun 2023 yang mengandung multikolinieritas dengan menggunakan regresi logistik biner CATPCA dengan estimator *spline truncated*.
   a. Menyusun statistik deskriptif dari setiap variabel independen yang berpengaruh terhadap keluarga berisiko *stunting*.
   b. Melakukan kuantifikasi optimal terhadap variabel independen kategorik menggunakan algoritma ALS sesuai pada Persamaan (5), sehingga diperoleh $\boldsymbol{X}^*$.
   c. Menguji keberadaan multikolinieritas dengan membentuk matriks korelasi dari $\boldsymbol{X}^*$ mengikuti Persamaan (8), lalu menghitung nilai eigen dan vektor eigennya.
   d. Membentuk komponen utama $\boldsymbol{Z}_k^*$ untuk mengatasi multikolinieritas pada $\boldsymbol{X}^*$ dan memilih sejumlah m komponen utama optimal berdasarkan persentase varians kumulatif $\geq 80\%$.

e. Melakukan pemodelan regresi logistik biner dengan estimator *spline truncated* orde satu (linier) pada 1,2, hingga 3 titik knot menggunakan komponen $\boldsymbol{Z}_k^*$ yang telah dibentuk.
f. Memilih model terbaik berdasarkan titik knot dengan nilai *Generalized Cross Validation* (GCV) terendah yang diperoleh melalui Persamaan (27).
g. Melakukan uji signifikansi parameter model terbaik baik secara simultan dengan menggunakan Persamaan (21), maupun secara parsial menggunakan Persamaan (22).
h. Menginterpretasi hasil estimasi model terbaik dan mengukur akurasi klasifikasinya menggunakan Persamaan (28).
i. Menarik kesimpulan.