Jurnal Riset Mahasiswa Matematika
Volume 4 (5) (2025), Pages 169-180
p-ISSN: 2808-1552; e-ISSN: 2808-4926

Reduksi *Imbalanced Data* Diagnosa Hipertensi dengan *Tomek Links* pada Regresi Logistik

Putri Aulia Fachreza, Ria Dhea Layla Nur Karisma*, and Erna Herawati

Program Studi Matematika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia

Abstrak

Masalah *imbalanced data* seringkali menghambat akurasi dalam proses klasifikasi, terutama dalam kasus diagnosis hipertensi, di mana jumlah kelas minoritas jauh lebih sedikit dibandingkan kelas mayoritas. Penelitian ini bertujuan untuk membangun model regresi logistik yang akurat dengan mengatasi ketidakseimbangan data menggunakan metode *Tomek Links*. Metode ini bekerja dengan menghapus pasangan data terdekat dari kelas berbeda untuk mereduksi *noise* dan memperbaiki distribusi data. Setelah dilakukan *undersampling* dengan *Tomek Links*, model regresi logistik dibentuk dengan pendekatan *Maximum Likelihood Estimation* melalui metode iteratif *Newton-Raphson*. Evaluasi model dilakukan melalui pengujian multikolinearitas, uji signifikansi parameter, uji kesesuaian model, dan pengukuran ketepatan klasifikasi berdasarkan nilai *Apparent Error Rate* (APER). Hasil penelitian menunjukkan bahwa variabel jenis kelamin, konsumsi gula berlebih, lemak berlebih, dan usia secara signifikan mempengaruhi kemungkinan seseorang menderita hipertensi. Model akhir menghasilkan tingkat akurasi sebesar 89,5%. Penelitian ini menunjukkan bahwa kombinasi metode *Tomek Links* dan regresi logistik dapat menjadi pendekatan efektif dalam menangani *imbalanced data* pada diagnosa hipertensi.

Kata Kunci: Klasifikasi; *Imbalanced Data*; *Tomek Links*; Regresi Logistik; Hipertensi

Abstract

The problem of imbalanced data often hampers accuracy in the classification process, especially in the case of hypertension diagnosis, where the amount of minority class data is much less than the majority class. This study aims to build an accurate logistic regression model by overcoming data imbalance using the Tomek Links method. This method works by removing the closest pair of data from different classes to reduce noise and improve data distribution. After undersampling with Tomek Links, a logistic regression model is formed using the Maximum Likelihood Estimation approach through the Newton-Raphson iterative method. Model evaluation was conducted through multicollinearity testing, parameter significance testing, model fit testing, and measurement of classification accurancy based on Apparent Error Rate (APER) values. The result showed that the variables of gender, excess sugar consumption, excess fat, and age significantly influenced the likelihood of a person suffering from hypertension. The final model produced an accuracy rate of 89,5%. This study shows that the combination of Tomek Links method and logistic regression can be an effective approach in handling imbalanced data in hypertension diagnosis.

Keywords: Classification; Imbalanced Data; Tomek Links, Logistic Regression; Hypertension

*Corresponding author. E-mail: riadhea@uin-malang.ac.id

Submitted: ... Reviewed: ... Accepted: ...
DOI: https://dx.doi.org/10.18860/jrmm.v4i5.34567

1 Pendahuluan

Perkembangan teknologi dan informasi di era modern telah memperkuat peran penting Matematika, khususnya dalam pengolahan dan analisis data. Disiplin seperti data sains, *big data*, pembelajaran mesin, dan kecerdasan buatan memerlukan pendekatan matematis untuk mengelola data yang kompleks agar dapat menghasilkan informasi yang bermanfaat. salah satu teknik yang digunakan dalam analisis data adalah klasifikasi, yang bertujuan mengelompokkan data ke suatu kelas tertentu. Di antara berbagai metode klasifikasi, regresi logistik merupakan metode yang terkenal karena kesederhanaan dan kemampuannya dalam menangani variabel independen berskala kategorik ataupun kontinu [1].

Beberapa penelitian telah memanfaatkan regresi logistik, seperti Ramandhani dkk., (2017) yang memperoleh akurasi 79,87% [2]. Penelitian Wan dkk., (2019) yang mencapai akurasi 75,2% [3]. Akan tetapi, terkadang metode ini cenderung sensitif apabila data yang digunakan tidak seimbang [1]. Kondisi ini dapat mengurangi akurasi dan efektivitas model prediksi [4]. Salah satu solusi yang dapat diterapkan adalah teknik *undersampling*, seperti *Tomek Links* yang bertujuan mereduksi data kelas mayoritas dengan menghilangkan *noise* [5]. Beberapa penelitian menggunakan *Tomek Links* sebelumnya menunjukkan bahwa penggunaan metode ini dapat meningkatkan tingkat akurasi klasifikasi.

Penelitian Dewi dkk., (2023) menghasilkan peningkatan akurasi sebesar 0,489% [6]. Sementara itu, Kumalasanti dan Aprilianti (2024) juga menerapkan metode serupa, dengan peningkatan akurasi sebesar 1,2% pada data sebelum pandemi dan 0,6% pada data setelah pandemi [7]. Masalah ketidakseimbangan data juga ditemukan dalam klasifikasi diagnosa hipertensi, yang merupakan salah satu penyakit tidak menular dengan prevalensi tinggi secara global. Berdasarkan data WHO jumlah penderita hipertensi terus meningkat, namun kesadaran masyarakat terhadap kondisi ini masih rendah[1]. Sebanyak 47,6% penduduk Indonesia mengetahui bahwa telah menderita hipertensi berdasarkan hasil skrining *May Measurement Month* (MMM) tahun 2018 [8].

Hipertensi terjadi saat tekanan darah sistolik $\geq$ 140 mmHg dan tekanan darah diastolik $\geq$ 90 mmHg. Faktor risiko hipertensi meliputi jenis kelamin, usia, makanan minuman yang dikonsumsi atau pola makan tinggi garam serta gula, kurangnya aktivitas fisik, merokok, dan konsumsi alkohol. Beberapa penelitian sebelumnya menunjukkan bahwa regresi logistik efektif dalam mengklasifikasi hipertensi. Misna dkk., (2018) menggunakan metode ini dan memperoleh akurasi sebesar 77,6% [9]. Sementara itu, Sahila dkk., (2024) membandingkan regresi logistik dengan algoritma naïve bayes, dan hasilnya regresi logistik memberikan akurasi yang lebih tinggi, yaitu 93,3% dibandingkan 63,6% untuk naïve bayes [10].

Penelitian ini bertujuan untuk mengetahui bagaimana model dan tingkat akurasi model regresi logistik pada reduksi *imbalanced data* diagnosa hipertensi dengan pendekatan *Tomek Links* untuk mengatasi ketidakseimbangan data. Dengan demikian, penelitian ini diharapkan mampu meningkatkan akurasi model klasifikasi dan mendukung upaya yang menjadi langkah penting untuk deteksi dini hipertensi dan mencegah komplikasi yang lebih serius.

2 Metode

Penelitian ini menggunakan data sekunder yang diperoleh dari Puskesmas Mojolangu, Kecamatan Lowokwaru, Kota Malang. Data yang digunakan mencakup variabel dependen dan independen. Variabel dependen (Y) adalah diagnosa hipertensi yang berskala nominal, dengan dua kategori yaitu 1 untuk menderita hipertensi dan 0 untuk tidak menderita hipertensi. Variabel independen (X) terdiri dari sembilan variabel yang seluruhnya berskala nominal, yaitu yaitu jenis kelamin (x_1), konsumsi rokok (x_2), aktivitas fisik (x_3), konsumsi gula berlebih (x_4), konsumsi garam berlebih (x_5), lemak berlebih (x_6), konsumsi buah dan sayur (x_7), konsumsi alkohol (x_8), dan usia (x_9).

Teknik analisis data dalam penelitian ini dilakukan melalui beberapa tahapan dengan bantuan *software* R Studio, dimulai dari pengumpulan data diagnosa hipertensi yang mencakup variabel dependen dan independen, dilanjutkan dengan analisis statistik deskriptif untuk melihat karakteristik data. Data

[1] https://www.who.int

kemudian dibagi secara acak menjadi data *training* dan data *testing* dengan perbandingan 70:30. Selanjutnya, dilakukan reduksi data tidak seimbang pada data *training* menggunakan metode *Tomek Links* dengan menghapus data dari kelas mayoritas berdasarkan jarak *Euclidean*. Kemudian data yang telah direduksi selanjutnya dilakukan pengolahan dan analisis data menggunakan regresi logistik dengan beberapa tahapan dan asumsi yang harus terpenuhi.

2.1 Regresi Logistik

Tahapan pertama pada analisis regresi logistik adalah dengan membentuk model awal. Model tersebut dibangun dengan melibatkan semua variabel independen berdasarkan Persamaan (1) dengan keterangan nilai i = 1, 2,. . . , N [11].

$$\pi(X_i) = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})} \tag{1}$$

2.2 Estimasi Parameter

Estimasi parameter dilakukan menggunakan metode *Maximum Likelihood Estimation* (MLE) dengan pendekatan numerik *Newton-Raphson* dengan berdasarkan Persamaan (2) [12]. Proses iterasi *Newton-Raphson* akan terhenti apabila telah konvergen dengan t adalah banyaknya iterasi.

$$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + \left(X^T V^{(t)} X\right)^{-1} X^T W^{(t)} \tag{2}$$

2.3 Uji Multikolinearitas

Pengujian multikolinearitas dilakukan untuk memastikan tidak terjadi pelanggaran multikolinearitas antar variabel penelitian. Apabila terjadi pelanggaran multikolinear, estimasi parameter dari model dapat menjadi bias [13]. Multikolinear terjadi apabila nilai VIF $\geq$ 10. Pengujian ini dilakukan dengan nilai *Variance Inflation Factor* (VIF) berdasarkan Persamaan (3) [14].

$$\text{VIF} = \frac{1}{\text{Tolerance}} \tag{3}$$

2.4 Pengujian Parameter

Uji signifikansi parameter dilakukan secara simultan dengan uji G dan secara parsial dengan uji *Wald* [11]. Pengujian dilakukan untuk mengetahui bagaimana variabel independen berpengaruh terhadap variabel dependen baik secara simultan ataupun parsial.

Hipotesis untuk uji G untuk uji signifikansi secara simultan sebagai berikut:

$$H_0 : \beta_0 = \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_1 : \text{Minimal terdapat satu } \beta_j \neq 0 \text{ dengan } j = 0, 1, 2, \ldots, p$$

Statistika uji ditulis pada Persamaan (4):

$$G = D_{\text{dengan variabel}} - D_{\text{tanpa variabel}} \tag{4}$$

Kriteria Penolakan: Tolak H_0 apabila *p-value* < 0,05.

Hipotesis untuk uji *Wald* untuk uji signifikansi secara parsial sebagai berikut:

$$H_0 : \beta_j = 0 \quad \text{(Variabel ke-}j\text{ tidak memiliki pengaruh signifikan)}$$

$$H_1 : \beta_j \neq 0 \quad \text{dengan } j = 0, 1, 2, \ldots, p \text{ (Variabel ke-}j\text{ memiliki pengaruh signifikan)}$$

Statistika uji ditulis pada Persamaan (5):

$$W = \left(\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \right)^2 \tag{5}$$

Kriteria Penolakan: Tolak H_0 apabila *p-value* < 0,05.

2.5 Uji Kesesuaian Model

Uji keseuaian model dilakukan setelah membentuk model akhir regresi logistik dengan berdasarkan variabel signifikan yang diketahui dalam pengujian parameter. Model regresi logistik akhir dibentuk berdasarkan Persamaan (1) kemudian diuji kesesuaiannya dengan melihat nilai *Deviance* [11]. Pengujian kesesuaian model dilakukan untuk mengevaluasi apakah terdapat perbedaan antara data aktual yang diamati dengan nilai prediksi yang dihasilkan oleh model.

Hipotesis uji kesesuaian model sebagai berikut:

H_0 : Model sudah sesuai

H_1 : Model belum sesuai

Statistik uji ditulis pada Persamaan (6):

$$D = -2 \sum_{i=1}^{N} \left[y_i \ln \left(\frac{\hat{\pi}(X_i)}{y_i} \right) + (1 - y_i) \ln \left(\frac{1 - \hat{\pi}(X_i)}{1 - y_i} \right) \right] \tag{6}$$

Kriteria Penolakan: Tolak H_0 apabila *p-value* < 0,05.

2.6 Interpretasi Koefisien Model

Koefisien parameter dalam model kemudian diinterpretasikan melalui nilai *Odds Ratio* [12]. Tujuan dari interpretasi koefisien model regresi logistik adalah untuk mengetahui dampak variabel independen terhadap variabel dependen. Nilai *Odds Ratio* diperoleh berdasarkan Persamaan (7) sebagai berikut:

$$\psi = \exp(\hat{\beta}_j) \tag{7}$$

2.7 Ketepatan Klasifikasi

Evaluasi performa model dalam klasifikasi dilakukan menggunakan *Confusion Matrix*. Tingkat akurasi kesalahan model dalam klasifikasi dihitung dengan nilai APER berdasarkan Persamaaan (8) sedangkan untuk menghitung ketepatan klasifikasi berdasarkan Persamaan (9) sebagai berikut [15]:

$$\text{APER} = \left(\frac{n_{12} + n_{21}}{n} \right) \times 100 \tag{8}$$

$$\text{1 - APER} = \left(\frac{n_{11} + n_{22}}{n} \right) \times 100 \tag{9}$$

- n_{11} : Banyak subjek dari y_1 yang benar diklasifikasikan sebagai y_1
- n_{12} : Banyak subjek dari y_1 yang salah diklasifikasikan sebagai y_2
- n_{21} : Banyak subjek dari y_2 yang salah diklasifikasikan sebagai y_1
- n_{22} : Banyak subjek dari y_2 yang benar diklasifikasikan sebagai y_2

3 Hasil Dan Pembahasan

3.1 Statistika Deskriptif

Analisis statistik deskriptif yang dilakukan menunjukkan gambaran umum karekteristik 2.387 pasien dengan informasi yang dikumpulkan meliputi 9 variabel independen yaitu, jenis kelami, konsumsi rokok, aktivitas fisik, konsumsi gula berlebih, konsumsi garam berlebih, lemak berlebih, konsumsi buah dan sayur, konsumsi alkohol, serta usia. Statistika deskriptif setiap variabel penelitian disajikan pada Tabel 1.

Tabel 1: Statistika deskriptif variabel penelitian

Variabel	Hipertensi	Tidak Hipertensi	Total	Persentase Hipertensi (%)
Jenis Kelamin - Laki-laki	89	290	379	23,5
Jenis Kelamin - Perempuan	636	1372	2008	31,7
Konsumsi Rokok - Ya	103	242	345	29,9
Konsumsi Rokok - Tidak	622	1420	2042	30,5
Aktivitas Fisik - Ya	311	622	973	32,0
Aktivitas Fisik - Tidak	414	1000	1414	29,3
Konsumsi Gula - Ya	229	365	594	38,6
Konsumsi Gula - Tidak	496	1297	1793	27,7
Konsumsi Garam - Ya	179	255	434	41,2
Konsumsi Garam - Tidak	546	1407	1953	28,0
Lemak Berlebih - Ya	218	380	598	36,5
Lemak Berlebih - Tidak	507	1282	1789	28,3
Konsumsi Buah dan Sayur - Ya	278	580	858	32,4
Konsumsi Buah dan Sayur - Tidak	447	1082	1529	29,2
Konsumsi Alkohol - Ya	4	5	9	44,4
Konsumsi Alkohol - Tidak	721	1657	2378	30,3
Usia Rentan - Ya	693	1308	2001	34,6
Usia Rentan - Tidak	32	354	386	8,3

Tabel 1 menunjukkan deskripsi dari jumlah pasien pada setiap variabel independen berdasarkan kategori diagnosa hipertensi.

3.2 Pembagian Data *Training* dan *Testing*

Data dibagi menjadi dua bagian, yaitu 70% untuk data pelatihan dan 30% untuk data pengujian. Rasio tersebut dipilih karena menghasilkan akurasi yang lebih optimal dibandingkan dengan rasio pembagian lainnya. Hasil dari pembagian data disajikan pada Tabel 2.

Tabel 2: Pembagian data *training* dan *testing*

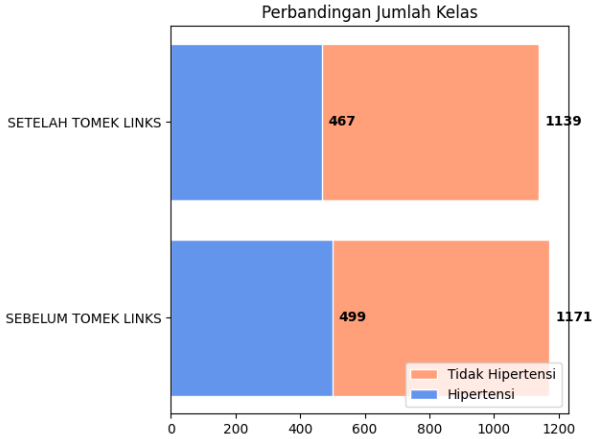
Pembagian Data	Testing (30%)	Training (70%)	Total
Diagnosa Hipertensi	226	499	725
Diagnosa Tidak Hipertensi	491	1171	1662
Total	717	1670	2387

Tabel 2 menyajikan hasil pembagian data, dari 725 sampel kategori hipertensi, 499 termasuk dalam data pelatihan dan 226 termasuk ke dalam data pengujian. Sedangkan dari 1.662 sampel tidak hipertensi, 1.171 digunakan untuk pelatihan dan 491 untuk pengujian.

3.3 Penerapan *Tomek Links*

Metode *Tomek Links* diterapkan pada data *training* yang sebelumnya terdiri dari 499 data dalam kategori hipertensi dan 1.171 data kategori tidak hipertensi. Hal tersebut menunjukkan bahwa data tidak seimbang karena antara kelas satu dengan kelas lainnya memiliki jumlah yang sangat berbeda jauh. *Tomek Link*

digunakan untuk mengatasi ketidakseimbangan data tersebut dengan cara mereduksi jumlah data pada kelas mayoritas. Perbedaan data sebelum dan setelah dilakukan *Tomek Links* disajikan pada Gambar 1.



Gambar 1: Perbedaan data sebelum dan setelah dilakukan *Tomek Links*

Gambar 1 memperlihatkan adanya pengurangan jumlah data pada kedua kategori setelah dilakukan proses *Tomek Links*. Pada kategori tidak hipertensi (kelas mayoritas), jumlah data menurun dari 1.171 menjadi 1.139. Sementara itu, kategori hipertensi (Kelas Mayoritas) juga mengalami pengurangan dari 499 menjadi 467 data. Dengan demikian, total data *training* setelah penerapan Tomek Links adalah 1.606.

3.4 Model Awal Regresi Logistik

Model regresi logistik awal dibangun dengan memasukkan semua variabel independen yang digunakan dalam penelitian ini. Pada tahap ini, pemilihan variabel belum dilakukan, sehingga seluruh variabel independen langsung dimasukkan ke dalam model. Berdasarkan Persamaan (1), diperoleh bentuk model regresi logistik sebagai berikut:

$$\pi(X) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots + \beta_9 x_9)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots + \beta_9 x_9)}$$

3.5 *Maximum Likelihood Estimation*

Metode MLE digunakan sebagai pendekatan untuk mengestimasi parameter dalam regresi logistik. Namun, karena bentuk fungsi *likelihood* dalam regresi logistik cukup kompleks dan tidak dapat diselesaikan secara langsung, maka digunakan metode numerik *Newton-Raphson* untuk memperoleh nilai estimasi parameter β secara bertahap melalui proses iterasi. Nilai estimasi parameter pada iterasi ke - 0 dihitung dengan menggunakan metode OLS. Kemudian nilai estimasi tersebut digunakan untuk menghitung nilai estimasi parameter pada iterasi ke - 1. Hasil estimasi parameter disajikan pada Tabel 3.

Persamaan iterasi berdasarkan Persamaan (2).:

$$\beta^{(1)} = \beta^{(0)} + \left(X^T V^{(0)} X\right)^{-1} X^T W^{(0)}$$

Perhitungan nilai estimasi parameter pada iterasi ke - 1:

$$\beta^{(1)} = \begin{bmatrix} 0{,}11284 \\ 0{,}10299 \\ \vdots \\ 0{,}81396 \end{bmatrix} + \left[\begin{bmatrix} 0{,}00788 & -0{,}00242 & \cdots & -0{,}02850 \\ -0{,}00242 & 0{,}02282 & \cdots & 0{,}00062 \\ \vdots & \vdots & \ddots & \vdots \\ -0{,}00285 & 0{,}00062 & \cdots & 0{,}01865 \end{bmatrix} \right] \cdot \begin{bmatrix} -447{,}344 \\ -91{,}741 \\ \vdots \\ 73{,}573 \end{bmatrix}$$

$$\beta^{(1)} = \begin{bmatrix} 0{,}11284 \\ 0{,}10299 \\ \vdots \\ 0{,}81396 \end{bmatrix} + \begin{bmatrix} -1{,}67284 \\ -0{,}32549 \\ \vdots \\ 2{,}73744 \end{bmatrix} = \begin{bmatrix} -1{,}56000 \\ -0{,}42848 \\ \vdots \\ 3{,}55140 \end{bmatrix}$$

Tabel 3: Estimasi parameter

Variabel	β_i	Estimasi Parameter
Intercept	β_0	$-2{,}143$
x_1	β_1	$-1{,}576$
x_2	β_2	$-0{,}150$
x_3	β_3	$-0{,}131$
x_4	β_4	0,543
x_5	β_5	0,395
x_6	β_6	0,485
x_7	β_7	0,080
x_8	β_8	1,334
x_9	β_9	5,210

Tabel 3 menyajikan hasil estimasi parameter pada model regresi logistik setelah melalui proses iteratif menggunakan Newton-Raphson. Dengan tingkat konvergensi sebesar $\varepsilon = 0{,}000001$, seluruh parameter telah mencapai konvergensi pada iterasi ke - 7 dengan nilai konvergensi yaitu sebesar $\varepsilon = 0{,}0000000007$, yang menunjukkan bahwa proses estimasi telah stabil.

3.6 Uji Multikolinearitas

Salah satu cara yang umum digunakan untuk mendeteksi multikolinearitas adalah dengan melihat nilai VIF. Multikolinear terjadi apabila nilai VIF ≥ 10. Hasil uji multikolinearitas disajikan pada Tabel 4.

Perhitungan VIF untuk variabel x_1 berdasarkan Persamaan (3):

$$\text{VIF} = \frac{1}{\text{Tolerance}}$$

$$\text{VIF} = \frac{1}{0{,}858}$$

$$\text{VIF} = 1{,}166$$

Tabel 4: Nilai VIF setiap variabel independen

Variabel	VIF	Keputusan
x_1	1,166	Terima H_0
x_2	1,139	Terima H_0
x_3	1,037	Terima H_0
x_4	1,043	Terima H_0
x_5	1,087	Terima H_0
x_6	1,047	Terima H_0
x_7	1,053	Terima H_0
x_8	1,017	Terima H_0
x_9	1,012	Terima H_0

Tabel 4 menunjukkan nilai VIF untuk masing-masing variabel < 10. Hal ini menunjukkan bahwa tidak terdapat indikasi multikolinearitas antara variabel.

3.7 Uji Signifikansi Parameter

Pengujian signifikansi parameter dalam regresi logistik terdiri dari dua bentuk yaitu, uji simultan dan uji parsial. Uji simultan dimaksudkan untuk mengkaji secara bersama-sama pengaruh seluruh variabel independen terhadap variabel diagnosa hipertensi. Sedangkan uji parsial bertujuan untuk mengkaji kontribusi masing-masing variabel independen terhadap variabel diagnosa hipertensi. Hasil pengujian signifikansi parameter melalui *software* R Studio secara simultan disajikan pada Tabel 5 dan secara parsial disajikan pada Tabel 6.

Perhitungan statistik G berdasarkan Persamaan (4):

$$G = 2\left(L_{\text{full}} - L_{\text{null}}\right)$$

$$G = 2\left(1936{,}363 - 1034{,}961\right)$$

$$G = 901{,}402$$

Tabel 5: Uji signifikansi simultan

	G	**Df**	**P-Value**	χ^2
Model	901,402	9	0,000	16,919

Tabel 5 menunjukkan nilai G sebesar 901,402 dengan derajat kebebasan 9 dan *p-value* sebesar 0,000. Karena *p-value* < 0,05, maka keputusan yang diambil adalah tolak H_0 sehingga mengindikasi bahwa seluruh variabel independen secara bersama-sama memiliki pengaruh terhadap diagnosa hipertensi.

Perhitungan statistik *Wald* berdasarkan Persamaan (5):

$$W_0 = \left(\frac{\hat{\beta}_0}{SE(\hat{\beta}_0)}\right)^2$$

$$W_0 = \left(\frac{-2{,}143}{0{,}144}\right)^2$$

$$W_0 = 220{,}352$$

Tabel 6: Uji signifikansi parsial

Parameter	**Estimasi**	**Std. Error**	**P-Value**	**Wald**	**Keputusan**
β_0	−2,143	0,144	0,000	220,352	Tolak H_0
β_1 (1)	−1,576	0,341	0,000	21,347	Tolak H_0
β_2 (1)	−0,150	0,286	0,601	0,273	Terima H_0
β_3 (1)	−0,131	0,172	0,447	0,579	Terima H_0
β_4 (1)	0,543	0,181	0,003	9,019	Tolak H_0
β_5 (1)	0,395	0,211	0,061	3,500	Terima H_0
β_6 (1)	0,485	0,180	0,007	7,279	Tolak H_0
β_7 (1)	0,080	0,175	0,646	0,210	Terima H_0
β_8 (1)	1,334	1,113	0,231	1,435	Terima H_0
β_9 (1)	5,210	0,298	0,000	305,397	Tolak H_0

Tabel 6 menunjukkan hasil uji signifikansi secara parsial diketahui terdapat 4 variabel yang signifikan dimana keempat variabel tersebut memiliki *p-value* < 0,05, sehingga keputusannya adalah tolak H_0. Keempat variabel tersebut adalah x_1 (jenis kelamin), x_4 (konsumsi gula berlebih), x_6 (lemak berlebih), dan x_9 (usia). Keempat variabel tersebut secara parsial memiliki pengaruh terhadap diagnosa hipertensi. Kemudian pengujian signifikansi parameter dengan hanya menggunakan keempat variabel

yang signifikan dilakukan dan hasil dari uji signifikansi kedua secara parsial disajikan pada Tabel 7 dan secara simultan pada Tabel 8.

Tabel 7: Uji signifikansi parsial kedua

Parameter	Estimasi	Std. Error	P-Value	Wald	Keputusan
β_0	$-2{,}130$	0,119	0,000	319,636	Tolak H_0
β_1 (1)	$-1{,}608$	0,323	0,000	24,775	Tolak H_0
β_4 (1)	0,594	0,178	0,001	11,085	Tolak H_0
β_6 (1)	0,552	0,175	0,002	9,904	Tolak H_0
β_9 (1)	5,189	0,296	0,000	308,141	Tolak H_0

Tabel 7 menunjukkan hasil uji signifikansi secara parsial kedua keempat variabel tersebut memiliki *p-value* < 0,05, sehingga keputusannya adalah tolak H_0, yang berarti keempat variabel tersebut secara parsial memiliki pengaruh terhadap diagnosa hipertensi.

Tabel 8: Uji signifikansi simultan kedua

	G	Df	P-Value	χ^2
Model	895,760	4	0,000	9,487

Tabel 8 menunjukkan bahwa nilai *p-value* < 0,05, sehingga keputusannya adalah tolak H_0, yang berarti keempat variabel tersebut secara bersama-sama memiliki pengaruh terhadap diagnosa hipertensi.

3.8 Model Regresi Logistik

Model regresi logistik disusun berdasarkan hasil uji signifikansi parameter, baik secara parsial maupun secara simultan. Berdasarkan pengujian diketahui variabel independen yang memiliki pengaruh signifikan terhadap diagnosa hipertensi adalah x_1 (jenis kelamin), x_4 (konsumsi gula berlebih), x_6 (lemak berlebih), dan x_9 (usia). Dengan demikian, model regresi logistik dirumuskan berdasarkan hasil estimasi parameter regresi pada Tabel 7 dan berdasarkan Persamaan (1) diperoleh model sebagai berikut:

$$\pi(X) = \frac{\exp(-2{,}130 - 1{,}608x_1(1) + 0{,}594x_4(1) + 0{,}552x_6(1) + 5{,}189x_9(1))}{1 + \exp(-2{,}130 - 1{,}608x_1(1) + 0{,}594x_4(1) + 0{,}552x_6(1) + 5{,}189x_9(1))}$$

3.9 Uji Kesesuaian Model

Pengujian kesesuaian model dilakukan berdasarkan nilai *Deviance* pada Persamaan (6). Hasil uji kesesuaian model disajikan pada Tabel 9.

Tabel 9: Uji Kesesuaian Model

	Deviance	DF	P-Value	χ^2
Model	1040,603	1601	1,000	1695,2

Tabel 9 menyajikan hasil pengujian kesesuaian model yaitu, nilai *p-value* sebesar 1,000 yang > dari taraf signifikansi $\alpha = 0{,}05$. Berdasarkan hasil tersebut, keputusan yang diambil adalah terima H_0, yang menunjukkan bahwa model regresi yang digunakana sudah sesuai dengan data yang diamati.

3.10 Interpretasi Koefisien Model

Koefisien pada model regresi logistik dapat diinterpretasikan melalui nilai *Odds Ratio* berdasarkan Persamaan (7). Nilai*Odds Ratio* yang diperoleh dari bantuan *software* R Studio disajikan pada Tabel 10.

Tabel 10: Nilai *Odds Ratio*

Variabel	Estimasi	*Odds Ratio*
Intercept	−2,130	0,1188
x_1	−1,608	0,2004
x_4	0,594	1,8117
x_5	0,552	1,7372
x_9	5,189	179,28

Variabel jenis kelamin (laki-laki) dengan koefisien negatif memiliki *Odds Ratio* sebesar 0,2004, yang berarti laki-laki memiliki risiko rendah dibandingkan perempuan untuk mengalami hipertensi. Variabel konsumsi gula berlebih memiliki *Odds Ratio* sebesar 1,8117, menunjukkan bahwa pasien dengan konsumsi gula berlebih berisiko 1,81 kali lebih tinggi terkena hipertensi. Sementara itu, variabel lemak berlebih menunjukkan *Odds Ratio* sebesar 1,7372, yang mengindikasi risiko hipertensi 1,74 kali lebih tinggi. Variabel usia memiliki pengaruh paling besar, dengan *Odds Ratio* sebesar 179,28, yang menunjukkan peningkatan risiko hipertensi seiring bertambahnya usia.

3.11 Ketepatan Klasifikasi Model

Ketepatan klasifikasi model bertujuan untuk menilai kemampuan model dalam mengklasifikasikan data ke dalam kategori yang benar. Hasil perhitungan akurasi klasifikasi disajikan dalam bentuk *Confusion Matrix* pada Tabel 11.

Tabel 11: *Confusion Matrix* Hasil Klasifikasi Model

Hasil Observasi	**Prediksi** y_0	**Prediksi** y_1	**Total**
y_0	483	67	550
y_1	8	159	167
Total	491	226	717

Tabel 11 menunjukkan bahwa sebanyak 483 pasien yang tidak hipertensi dan 159 pasien penderita hipertensi berhasil diklasifikasikan dengan benar. Namun, terdapat kesalahan klasifikasi pada 8 kasus hipertensi yang terdeteksi sebagai tidak hipertensi, serta 67 kasus tidak hipertensi yang terklasifikasikan sebagai hipertensi. Kesalahan pada kasus tersebut kemungkinan disebabkan oleh faktor-faktor lain, seperti pasien yang tengah menjalani pengobatan, sehingga tekanan darahnya normal saat pemeriksaan.

Tingkat kesalahan model dalam mengklasifikasi data dihitung menggunakan rumus APER berdasarkan Persamaan (8), yaitu:

$$\begin{aligned}\text{APER} &= \frac{n_{01}+n_{10}}{n} \\ &= \frac{67+8}{717} \\ &= 0{,}105 \\ &= 10{,}5\%\end{aligned}$$

Ketepatan klasifikasi dihitung berdasarkan Persamaan (9) sebagai berikut:

$$\begin{aligned}\text{Tingkat ketepatan klasifikasi} &= 1-\text{APER} = \frac{n_{00}+n_{11}}{n} \\ &= \frac{483+159}{717} \\ &= 0{,}895 \\ &= 89{,}5\%\end{aligned}$$

Berdasarkan nilai APER sebesar 10,5%, diketahui bahwa tingkat kesalahan klasifikasi model cukup rendah. Artinya, hanya sebagian kecil data yang salah klasifikasi, sementara tingkat ketepatan model mencapai 89,5%. Dengan tingkat akurasi tinggi dan kesalahan minimal, model regresi logistik menunjukkan kinerja yang baik dalam memprediksi diagnosa hipertensi berdasarkan variabel yang digunakan.

4 Kesimpulan

Berdasarkan hasil penelitian, model regresi logistik yang dibangun untuk menganalisis data diagnosa hipertensi dengan metode reduksi imbalanced data menggunakan *Tomek Links* menunjukkan bahwa variabel jenis kelamin, usia, serta konsumsi gula dan lemak berlebih berpengaruh signifikan terhadap risiko hipertensi. Model ini memiliki tingkat akurasi klasifikasi sebesar 89,5% dan tingkat kesalahan model sebesar 10,5%, yang menunjukkan performa yang cukup baik dalam mengklasifikasikan data. Tingkat ketepatan yang tinggi ini menunjukkan bahwa model yang dibangun tidak hanya memiliki performa statistik yang baik, tetapi juga berpotensi memberikan manfaat praktis untuk penelitian selanjutnya, Dinas Kesehatan terkait, maupun masyarakat umum.

Pernyataan Kontribusi Penulis

Putri Aulia Fachreza: Bertanggung jawab atas perencanaan konsep penelitian, penyusunan metodologi, serta penulisan draf awal. **Ria Dhea Layla Nur Karisma:** Memberikan rekomendasi sumber data dan berperan penting dalam peninjauan dan penyuntingan naskah agar sesuai dengan standar akademik. **Erna Herawati:** Memberikan peran dalam menyusun interpretasi dan penyampaian hasil penelitian.

Deklarasi Penggunaan AI atau Teknologi Berbasis AI

Model ChatGPT versi 4 dimanfaatkan dalam proses memperbaiki struktur kalimat dan melakukan parafrase guna meningkatkan kejelasan dan kesesuaian bahasa akademik dalam penulisan.

Deklarasi Konflik Kepentingan

Penulis menegaskan bahwa tidak terdapat konflik kepentingan yang mempengaruhi proses perancangan, pelaksanaan, atau pelaporan penelitian ini.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih yang sedalam-dalamnya kepada semua pihak yang telah memberikan dukungan dan semangat selama proses penelitian. Terima kasih kepada dosen pembimbing I, Ibu Ria Dhea Layla Nur Karisma, M.Si., dan dosen pembimbing II, Ibu Erna Herawati, M.Pd. Terima Kasih penulis sampaikan juga kepada keluarga dan teman-teman yang selalu memberikan doa dan motivasi kepada penulis, sehingga penulis dapat menyelesaikan penelitian ini.

Ketersediaan Data

Data yang mendukung temuan dalam penelitian ini dapat diperoleh dengan mengajukan permintaan kepada instansi terkait. Akses terhadap data tersebut akan diberikan dengan mempertimbangkan ketentuan kerahasiaan dan hanya untuk tujuan akademik atau penelitian lebih lanjut, sesuai denfan perjanjian yang disepakati bersama.

Daftar Pustaka

[1] E. Antipov and E. Pokryshevskaya, "Applying chaid for logistic regression diagnostics and classification accuracy improvement," *The State University Higher School of Economics*, 2009, Munich Personal RePEc Archive.

[2] R. Ramandhani, Sudarno, and D. Safitri, "Metode bootstrap aggregating regresi logistik biner untuk ketepatan klasifikasi kesejahteraan rumah tangga di kota pati," *Jurnal Gaussian*, vol. 6, pp. 121–130, 2017. DOI: 10.14710/j.gauss.6.1.121-130

[3] C. M. Wan, A. Nosedal-Sanchez, J. Nosedal-Sanchez, A. Asgary, and B. Pantin, "Modeling provision of disaster mutual assistance by electricity utilities using logistic regression," *International Journal of Disaster Risk Reduction*, pp. 8–9, 2019. DOI: 10.1016/j.ijdrr.2019.101110

[4] G. Qiong, X.-M. Wang, Z. Wu, B. Ning, and C.-S. Xin, "An improved SMOTE algorithm based on genetic algorithm for imbalanced data classification," *Journal of Digital Information Management*, vol. 14, no. 2, pp. 92–103, 2016. Available online.

[5] I. Tomek, "Two modifications of cnn," *IEEE Transactions of Systems, Man, and Communications*, vol. 6, pp. 769–772, 1997. DOI: 10.1109/TSMC.1976.4309452

[6] I. A. M. C. Dewi, I. K. Dharmendra, and N. W. Setiasih, "Analisis sentimen review aplikasi satu sehat mobile menggunakan model sampling tomek links," *Jurnal Informatika dan Komputer*, vol. 12, no. 3, pp. 45–55, 2023. DOI: 10.36002/jutik.v9i5.2644

[7] R. Kumalasanti and N. M. D. Aprilianti, "Sentiment analysis of bali calendar application reviews using k-nearest neighbour," *International Journal of Engineering Technology and Natural Sciences*, vol. 6, no. 1, pp. 70–73, 2024. DOI: 10.46923/ijets.v6i1.339

[8] Y. Turana et al., "May measurement month 2018: An analysis of blood pressure screening results from indonesia," *European Heart Journal Supplements*, vol. 22, no. Suppl H, H66–H69, 2020. DOI: 10.1093/eurheartj/suaa031

[9] Misna, Rais, and I. T. Utami, "Analisis regresi logistik biner untuk mengklasifikasi penderita hipertensi berdasarkan kebiasaan merokok di rsu mokopido toli-toli," *Natural Science: Journal of Science and Technology*, vol. 7, no. 3, pp. 341–348, 2018. Available online.

[10] R. Sahila, T. Widiharih, and I. T. Utami, "Analisis klasifikasi menggunakan regresi logistik biner dan algoritma naïve bayes classifier pada penyakit hipertensi," *Jurnal Gaussian*, vol. 13, no. 2, pp. 319–327, 2024. DOI: 10.14710/j.gauss.13.2.319-327

[11] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd ed. New York: John Wiley & Sons, 2000.

[12] A. Agresti, *Categorical Data Analysis*. New York: John Wiley & Sons, 1990.

[13] J. Sungkono and T. K. Nugrahaningsih, "Simulasi dampak multikolinearitas pada kondisi penyimpangan asumsi normalitas," *Magistra*, vol. 29, no. 101, pp. 45–50, 2017. Available online.

[14] R. R. Hocking, *Methods and Applications of Linear Models*, 2nd ed. New Jersey: John Wiley & Sons, 2003.

[15] R. A. Johnson and D. Wichern, *Applied Multivariate Statistical Analysis*, 6th ed. New Jersey: Pearson Education, Inc., 2007.