MATHunesa
Jurnal Ilmiah Matematika
e-ISSN : 2716-506X | p-ISSN : 2301-9115
Volume 13 No 03
Tahun 2025

EVALUASI KINERJA PENANGANAN DATA TIDAK SEIMBANG DALAM MEMPREDIKSI LAJU PERTUMBUHAN PENDUDUK DI KALIMANTAN

Khusnia Nurul Khikmah
Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Palangka Raya, Kota Palangka Raya, Kalimantan Tengah, 74874 Indonesia
Email: khusnia.nurulkhikmah@mipa.upr.ac.id*

A'yunin Sofro
Program Studi Sains Aktuaria, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Surabaya, Kota Surabaya, Jawa Timur, 60231 Indonesia
Email: ayuninsofro@unesa.ac.id

Abstrak

Laju pertumbuhan penduduk merupakan indikator demografi krusial yang memengaruhi berbagai aspek kabupaten dan kota di Kalimantan, sehingga membutuhkan analisis komprehensif dalam memodelkan data. Fakta lapangannya masalah ini salah satunya dipengaruhi oleh fenomena ketidakseimbangan data atau satu kategori laju pertumbuhan penduduk lebih dominan. Oleh karena itu, penelitian ini mengusulkan kajian perbandingan metode penanganan data tidak seimbang dengan model analisis regresi logistik. Empat metode penanganan data tidak seimbang ini adalah tanpa penanganan atau *baseline*, kedua dengan metode *random over sampling* (ROS), *random undersampling* (RUS), dan *synthetic minority oversampling technique* (SMOTE). Dimana data yang digunakan adalah data sekunder yang diambil dari Badan Pusat Statistik (BPS) lima provinsi di Kalimantan. Hasil analisis laju pertumbuhan penduduk kabupaten dan kota di Kalimantan menunjukkan bahwa model regresi logistic biner dengan tanpa penanganan atau baseline memberikan akurasi hasil prediksi terbaik. Berdasarkan nilai akurasi, *balanced accuracy*, dan ROC menunjukkan nilai tertinggi dibandingkan metode penanganan data tidak seimbang lainnya, yaitu untuk data latih sebesar 66.7%, 53.1%, dan 76.51%. Sedangkan untuk data uji sebesar 72.7%, 62.5%, dan 82.14%.
Kata Kunci: Laju Pertumbuhan Penduduk, *Random Over Sampling*, *Random Undersampling*, Regresi Logistik Biner, SMOTE.

Abstract

Population growth rate is a crucial demographic indicator that affects various aspects of districts and cities in Kalimantan, thus requiring comprehensive analysis in modelling data. One of the factors influencing this issue is the phenomenon of data imbalance, where one category of population growth rate is more dominant. Therefore, this study proposes a comparative analysis of methods for handling imbalanced data using a logistic regression analysis model. The four methods for handling imbalanced data are: no handling or baseline, random over-sampling (ROS), random under-sampling (RUS), and synthetic minority oversampling technique (SMOTE). The data used are secondary data obtained from the Central Statistics Agency (BPS) of five provinces in Kalimantan. The analysis of population growth rates in districts and cities in Kalimantan shows that the binary logistic regression model without handling or baseline provides the best prediction accuracy. Based on accuracy, balanced accuracy, and ROC values, it achieved the highest values compared to other imbalanced data handling methods, namely 66.7%, 53.1%, and 76.51% for training data, and 72.7%, 62.5%, and 82.14% for test data.
Keywords: *Population Growth Rate, Random Over Sampling, Random Under Sampling, Binary Logistic Regression, SMOTE.*

PENDAHULUAN

Indikator demografi fundamental yang merefleksikan fluktuasi populasi dari suatu wilayah adalah melalui laju pertumbuhan penduduk. Fluktuasi ini berimplikasi pada berbagai sektor lainnya yang signifikan bagi suatu wilayah seperti sosial dan ekonomi (Sadigov, 2022) termasuk Kalimantan. Secara demografis Kalimantan memiliki wilayah yang luas dan kekayaan sumber daya alam sehingga perkembangan ekonomi dan pemahaman terhadap faktor-faktor yang memengaruhi laju penduduk di tingkat kabupaten dan kotanya menjadi hal yang krusial. Akibatnya laju pertumbuhan penduduk menjadi salah satu faktor penentu yang signifikan jika terjadi

perubahan dan distribusi penduduknya terhadap kebutuhan infrastruktur (O'Sullivan, 2023), lapangan pekerjaan (Pandey et al., 2022), hingga pada layanan public seperti pendidikan dan kesehatan (Chen et al., 2023). Oleh karena itu, kebutuhan terhadap pendekatan untuk memodelkan dan memahami masalah laju pertumbuhan penduduk ini menjadi syarat penting dalam merumuskan kebijakan pembangunan yang adaptif terhadap tantangan demografi.

Pendekatan yang mampu memodelkan sekaligus memahami determinan terhadap laju pertumbuhan penduduk ini adalah dengan regresi logistik biner (Joshi & Dhakal, 2021; Khikmah et al., 2022). Model ini memungkinkan untuk mengestimasi peluang suatu kejadian berdasarkan berbagai variabel prediktor dan mampu menginterpretasikan pengaruh dari masing-masing variabel prediktor tersebut. Selain itu, relevansi dan fleksibilitasnya dalam menangani berbagai jenis variabel prediktor menjadi pilihan alat pemodelan untuk digunakan dalam analisis. Penelitian sebelumnya pada (Joshi & Dhakal, 2021) menunjukkan bahwa regresi logistik biner memberikan hasil akurasi yang baik dibandingkan metode lainnya dengan akurasi terbaik sebesar 78.26%. Namun, penggunaan model regresi logistik biner ini dalam praktiknya seringkali berbenturan dengan kondisi data aktual yang mana pada masalah ketidakseimbangan data. Oleh karena itu, penelitian ini menawarkan perbandingan pendekatan penanganan data tidak seimbang.

Empat pendekatan penanganan data tidak seimbang yang diajukan oleh penelitian ini adalah dengan tanpa penanganan atau *baseline*, kedua dengan metode *random over sampling* (ROS) (Ependi et al., 2023), *random undersampling* (RUS), dan *synthetic minority oversampling technique* (SMOTE) (Mishra, 2017). Metode *baseline* dipilih sebagai titik referensi atau pembanding dasar untuk mengevaluasi efektivitas metode penanganan lainnya. Selanjutnya metode ROS dipilih karena mampu menduplikasi secara acak observasi dari kelas minoritas yang menurut penelitian sebelum pada (Khushi et al., 2021; Wongvorachan et al., 2023) merupakan metode penanganan data tidak seimbang terbaik. Ketiga, metode RUS dipilih dengan pertimbangan kelebihannya dalam pengurangan sampel acak dari kelas mayoritas hingga seimbang dengan kelas minoritas yang mana menurut penelitian sebelumnya pada (Bagui & Li, 2021) merupakan metode penanganan data tidak seimbang terbaik. Terakhir, SMOTE metode ini dipilih karena memiliki kemampuan untuk menghasilkan sampel sintesis baru dengan menginterpolasi antara sampel minoritas yang ada dengan tetangga terdekatnya (Wang et al., 2021).

Urgensi terkait pemahaman fluktuasi laju pertumbuhan penduduk dan tantangan masalah ketidakseimbangan data, penelitian ini bertujuan untuk mengidentifikasi skenario penanganan data yang paling efektif dalam meningkatkan kinerja prediksi regresi logistik biner dalam masalah laju pertumbuhan penduduk kabupaten dan kota di Kalimantan. Penelitian ini juga akan memberikan wawasan tentang faktor-faktor yang memengaruhi laju pertumbuhan penduduk di Kalimantan. Selain itu, wawasan terkait panduan praktis bagi peneliti dan pembuat kebijakan dalam menemukan pendekatan terbaik dan perbandingannya untuk analisis data demografi dengan masalah data tidak seimbang juga tercakup dalam penelitian ini.

KAJIAN TEORI

REGRESI LOGISTIK BINER

Regresi logistik biner adalah model yang secara umum bertujuan untuk menjelaskan hubungan antara variabel respon (biner) dengan variabel bebasnya. Jika model regresi logistik biner dengan koefisien kemiringan atau *slope* β_0 dan fungsi $f(x)$ dapat diestimasi dengan maksimum likelihood untuk nilai x. Maka persamaan umum regresi logistik dengan menggunakan fungsi logistik kumulatif dan transformasi $logit\pi(x)$ dengan $i = 1,2, \dots, j$ dapat didefinisikan sebagai berikut (Fatimah et al., 2023; Khikmah et al., 2022).

$$\begin{aligned} f(x) &= ln\left[\frac{\pi(x)}{1-\pi(x)}\right] \\ &= ln\left[\frac{e^{\beta_0+\sum_{i=1}^{j}\beta_i x_i}}{\left(1-\frac{e^{\beta_0+\sum_{i=1}^{j}\beta_i x_i}}{1+e^{\beta_0+\sum_{i=1}^{j}\beta_i x_i}}\right)}\right] \\ &= \beta_0 + \sum_{i=1}^{j}\beta_i x_i \end{aligned} \quad (1)$$

Sehingga interpretasi atas makna model logit yang diperoleh dapat melalui nilai *odds ratio*. *Odds ratio* adalah ukuran asosiasi yang didefinisikan sebagai perbandingan peluang terjadi atau tidaknya suatu kejadian. Secara matematis nilai *odds ratio* dapat diperoleh sebagai berikut.

$$odds\ ratio = \frac{odds_1}{odds_2} = \frac{\left(\frac{\pi(1)}{1-\pi(1)}\right)}{\left(\frac{\pi(2)}{1-\pi(2)}\right)} \quad (2)$$

Selanjutnya model dilakukan pengujian hipotesis dengan tiga penguji, pertama uji kesesuaian model, uji simultan, dan uji parsial. Uji kesesuaian model yang digunakan dalam penelitian ini adalah uji Hosmer-Lemeshow (χ^2). Uji ini bertujuan untuk mengukur kelayakan model dimana hipotesis awalnya H_0 afalah model regresi logistik cocok dengan data dan hipotesis alternatifnya H_1 adalah model regresi tidak cocok dengan data. Dengan keputusan, jika nilai dari $p - value \leq \alpha$ maka akan menolak H_0 (Hosmer Jr et al., 2013).

Kedua uji simultan (*likelihood ratio test*) yang bertujuan untuk mengetahui signifikansi seluruh variabel prediktornya secara simultan terhadap variabel responnya, dengan statistik uji yang digunakan adalah uji G. Secara matematis nilai dari statistic uji G ini dapat dihitung melalui persamaan (Harris, 2021; Zaidi & Al Luhayb, 2023).

$$G = -2\ ln\left[\frac{lik\ tanpa\ variabel\ prediktor}{lik\ dengan\ variabel\ prediktor}\right] \quad (3)$$

Hipotesis awal dari uji G yang diajukan ini adalah $H_0: \beta_0 = \beta_1 = \cdots, \beta_j = 0$ atau tidak ada pengaruh yang signifikan secara simultan dari variabel prediktornya terhadap variabel responnya dan hipotesis alternatifnya adalah H_1: minimal ada satu $\beta_i \neq 0$, dengan $i = 1,2, \ldots, j$ atau setidaknya ada satu pengaruh yang signifikan dari variabel prediktornya terhadap variabel responnya. Dengan keputusan Dengan keputusan, jika nilai dari $p - value \leq \alpha$ maka akan menolak H_0.

Terakhir uji parsial *Wald*, uji ini bertujuan untuk mengetahui signifikansi masing-masing variable prediktor terhadap variabel responnya. Dimana hipotesis awal uji *Wald* ini adalah $H_0: \beta_i = 0$ dan hipotesis alternatifnya $H_1: \beta_i \neq 0$, dengan $i = 1,2, \ldots, j$. Dengan keputusan, jika nilai dari $p - value \leq \alpha$ maka akan menolak H_0. Secara matematis uji *Wald* ini dapat ditulis sebagai berikut (Buya et al., 2020).

$$W^2 = \left[\frac{\hat{\beta}}{SE_{\hat{\beta}}}\right]^2 \quad (4)$$

RANDOM OVER SAMPLING (ROS)

Metode *random over samping* (ROS) adalah teknik penyeimbangna data yang sedrhana dimana metode ini bekerja dengan cara menduplikasi sampel dari kelas minoritas secara acak hingga jumlahnya sama dengan kelas mayoritas. Proses acak duplikasi ini dilakukan dengan penggantian yang mana bertujuan untuk meningkatkan representasi kelas minoritas dalam data. Selain itu, metode ini mudah diimplementasikan dan sederhana dalam penerapannya (Ependi et al., 2023).

RANDOM UNDER SAMPLING (RUS)

Metode *random under sampling* (RUS) ini adalah Teknik menyeimbangkan data dengan menyeimbangkan distribusi kelas dengan mengurangi jumlah sampel dari kelas mayoritas secara acak hingga jumlah kelas tersebut sama dengan atau sebanding dengan kelas minoritasnya (Iori et al., 2022). Tujuan utama metode ini adalah menciptakan distribusi kelas yang seimbang. Selain itu, metode ini memiliki kelebihan terkait penerapannya yang mudah dan sederhana (Kumar et al., 2021).

SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE)

Metode synthetic minority over-sampling technique (SMOTE) adalah salah satu teknik penanganan data tidak seimbang dengan menghasilkan sampel sintesis baru dari kelas minoritasnya (Pradipta et al., 2021). Proses sintesis ini dilakukan dengan memilih sampel minoritas secara acak berdasarkan k tetangga terdekatnya hingga terbentuk sample sintesis baru. Sehingga menurut penelitian sebelunya pada (Matharaarachchi et al., 2024) metode ini mengurangi risiko *overfitting* dan *robust* sebagai metode penanganan data tidak seimbang.

EVALUASI MODEL

Evaluasi model yang digunakan dalam penelitian ini adalah dengan Akaike's information criterion (AIC) dimana metode ini mengukur kebaikan model

(Cavanaugh & Neath, 2019). Secara umum model terbaik yang didapatkan diperoleh berdasarkan nilai AIC terkecilnya. Untuk model dengan estimasi parameter k, dimana nilai dari maximum likelihoodnya adalah $ln\ ln\ (L)$ maka secara matematis nilai AIC nya dapat ditulis dengan persamaan sebagai berikut (De & Acquah, 2010; Portet, 2020).

$$AIC = 2k - 2\ ln\ (L) \quad (5)$$

MATRIKS EVALUASI

Evaluasi hasil prediksi model yang dipilih dalam penelitian ini degan mempertibangkan masala data tidak seimbang, matriks evaluasi yang digunakan dalam penelitian ini adalah akurasi, *balanced accuracy*, dan *area under the receiver operating characteristic* (AUC-ROC). Secara umum akurasi, memiliki definisi sebagai proporsi preidksi benar dari total observasi dan merupakan matriks evaluasi dasar yang digunakan sebagai pembanding. JIka suatu data memiliki confuse matriks yang tersaji pada Tabel 1, maka nilai akurasinya secara matematis dapat ditulis sebagai berikut (Van den Goorbergh et al., 2022).

$$Akurasi = \frac{Tp + TN}{TP + TN + FP + FN} \quad (6)$$

Matriks evaluasi kedua adalah *balanced accuracy*, matriks ini didefinisikan sebagai rata-rata *recall* (sensitivitas dan spesifisitas) yang bertujuan untuk mengukur kemampuan prediksi model kedua kelas secara seimbang tanpa mempertimbangkan proporsi sampelnya. Secara matematis nilai *balanced accuracy* dapat dirumuskan sebagai berikut (Thölke et al., 2023).

$$BA = \frac{Sensitivitas + Spesifisitas}{2} \quad (7)$$

Terakhir nilai ROC, dimana matrik evaluasi ini menjelaskan *trade-off* antara sensitivitas dan spesifisitas pada berbagai kondisi atau ambang batas klasifikasi dengan $ROC \in [0,1]$ (Carrington et al., 2022). Nilai ROC ini mengukur kemampuan prediksi model secara keseluruhan tanpa terpengaruh oleh distribusi kelas dan ambang batas klasifikasi sehingga dapat membedakan prediksi kelas positif dan negatif model dan menjadikannya indikator kinerja model yang stabil dan komprehensif.

Tabel 1. Konfusi Matriks

Prediksi	**Empiris**	
	True	False
True	True Positive (TP)	False Positive (FP)
False	False Negative (FN)	True Negative (TN)

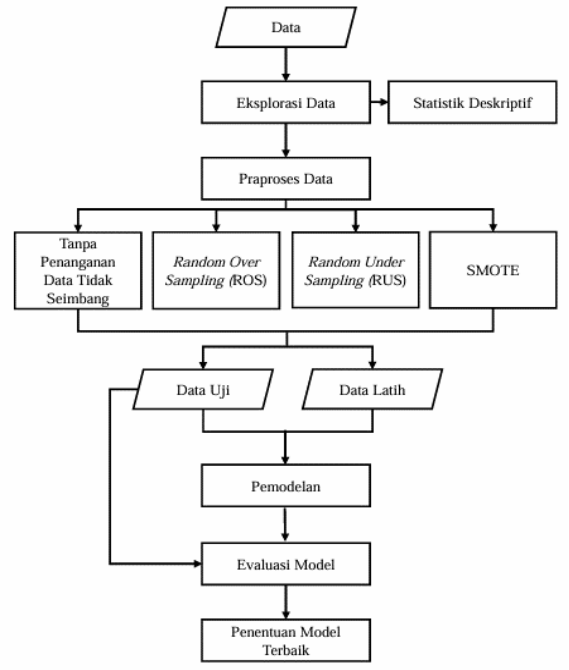
METODE

Data yang digunakan pada penelitian ini adalah data sekunder yang diambil dari website resmi Badan Pusat Statistik (BPS) lima provinsi di Kalimantan yang secara jelas masing-masing keterangan di setiap variabelnya tersaji pada Tabel 1.

Tabel 1. Statistik deskriptif variabel respon

Simbol	**Data**	**Penjelasan**
Y	Laju pertumbuhan penduduk	0: Lambat [0%, 2%)
		1: Cepat ≥ 2%
X_1	Jumlah penduduk	Dalam ribu
X_2	Persentase penduduk miskin	%
X_3	Persentase jumlah penduduk	%
X_4	Kepadatan penduduk	/km²
X_5	Lama sekolah	Pertahun
X_6	Pengerluaran perkapita untuk makanan	%
X_7	Penduduk yang bekerja	Dalam ribu

Selanjutnya tahapan analisis yang dilakukan pada penelitian ini dimulai dengan statistik deskriptif, kemudian melakukan *preprocessing* data dengan penanganan data tidak seimbang, kemudian data yang telah dilakukan penanganan dilakukan pemodelan dengan menggunakan data hasil pembagian data yaitu data latih. Pemodelan dengan regresi logistik biner dilakukan pada tahap ini kemudian dilakukan perbandingan kebaikan model yang selanjutnya dilakukan pengujian hipotesis. Hasil uji hipotesis model yang memenuhi asumsi akan digunakan sebagai acuan model terbaik. Terakhir dilakukan validasi dengan data uji dan penentuan model terbaik berdasarkan matriks evaluasi yang diajukan. Secara umum tahapan analisis ini terangkum pada *flowchart* berikut.



Gambar 1. *Flowchart* analisis

HASIL DAN PEMBAHASAN

Hasil analisis yang akan dijelaskan pada bagian ini membahas terkait hasil analisis regresi logistik biner dalam memodelkan dan memprediksi laju pertumbuhan penduduk kabupaten/kota di Kalimantan, yang mana berfokus pada perbandingan empat skenario penanganan data tidak seimbang. Empat skenario yang diajukan oleh penelitian ini adalah tanpa penanganan atau *baseline*, ROS, RUS, dan SMOTE. Keempat skenario ini diajukan untuk mengidentifikasi metode penanganan data tidak seimbang yang paling efektif dalam meningkatkan kinerja model regresi logistik biner dalam memodelkan, memprediksi, dan mengidentifikasi faktor-faktor yang signifikan mempengaruhi laju pertumbuhan penduduk kabupaten/kota di Kalimantan.

Data yang digunakan dalam penelitian ini adalah sebanyak 56 kabupaten dan kota di Kalimantan dengan tujuh variabel prediktor dan laju pertumbuhan penduduk sebagai variabel respon. Secara umum ukuran kelas variable respon dan statistic deskriptif variable prediktor yang digunakan dalam penelitian disajikan pada Tabel 2 dan 3.

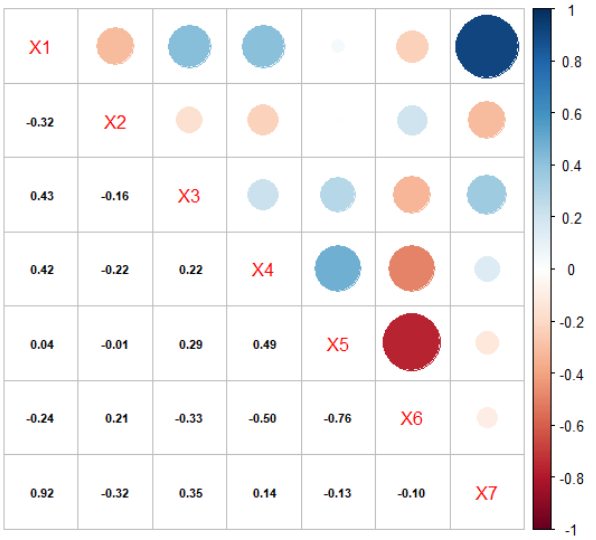
Tabel 2. Ukuran kelas variable respon

Kategori		Ukuran Kelas
0	Laju pertumbuhan penduduk lambat	36
1	Laju pertumbuhan penduduk cepat	20

Tabel 3. Statistik deskriptif variabel respon

Ukuran Statistik	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Min	28.8	2.23	0.81	2.0	6.54	41.43	14640
Q_1	171.9	4.52	4.67	16.0	7.94	50.12	86679
Median	269.7	5.96	6.36	37.0	8.39	53.31	130072
Rata-rata	320.6	8.54	8.93	382.4	8.66	52.51	150735
Q_3	427.1	9.13	11.15	128.5	9.24	55.72	213925
Maks	865.3	27.57	34.52	6819.0	11.71	61.98	403138

Secara umum tahapan analisis dengan regresi logistik biner dengan penanganan data tidak seimbang dimulai melakukan pembagian data. Pembagian data yang dilakukan oleh penelitian ini adalah 80% dari total data digunakan sebagai data latih dan 20% sisanya digunakan sebagai data uji. Selanjutnya, hasil pembagian data ini dilakukan pengujian korelasi, dimana korelasi Pearson (r) digunakan dalam penelitian ini. Hasil uji korelasi yang dilakukan memberikan informasi bahwa variabel prediktor dengan nilai $|r| \in (0.7,1]$ memiliki korelasi yang sangat kuat, untuk $|r| \in [0.5,0.7]$ memiliki korelasi yang kuat, dengan $|r| \in [0.3,0.5]$ memiliki korelasi sedang, dan $|r| \in [0.0,0.3]$ memiliki korelasi lemah. Secara jelas korelasi masing-masing variabel prediktor tersaji dengan jelas pada Gambar 2.



Gambar 2. Plot nilai uji korelasi (r)

Selanjutnya dilakukan pemodelan dengan regresi logistik biner dimana data yang digunakan dalam pemodelan adalah data latih. Keempat skenario penanganan data tidak seimbang dilakukan pada bagian ini. Pertama dengan tanpa penanganan data tidak seimbang atau *baseline*, kedua dengan ROS dimana data dengan jumlah kelas minoritas dilakukan pemilihan secara acak untuk dilakukan salinan yang selanjutnya ditambahkan ke dataset pelatihan, sehingga diperoleh jumlah kelas minoritas sama dengan jumlah kelas data mayoritas atau masing-masing sebesar 36. Ketiga dilakukan penanganan dengan RUS dimana pada penanganan ini secara umum menyaman jumlah observasi kelas mayoritas dengan kelas minoritas dengan cara mengatasi secara acak data dari kelas mayoritas sehingga diperoleh masing-masing kelas sebanyak 20. Keempat, penanganan dengan SMOTE dimana pada penanganan ini dilakukan sampel sintesis baru berdasarkan hasil interpolasi antara data kelas minoritas yang ada dengan tetangga terdekatnya dimana nilai tetangga terdekatnya $(k = 5)$.

Empat data yang selanjutnya dilakukan pemodelan berdasarkan hasil penanganan data tidak seimbang ini dilakukan perbandingan model terbaik berdasarkan nilai AICnya. Hasil perbandingan model regresi logistik biner yang dilakukan disajikan pada Tabel 4.

Tabel 4. Model Regresi Logistik Biner

Penanganan Data Tidak Seimbang	**Model**	**Nilai AIC**
Baseline	$Y = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7$	65.67
ROS	$Y = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7$	77.59
RUS	$Y = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7$	52.03
SMOTE	$Y = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7$	79.19

Tabel 4 menunjukkan bahwa data yang dilakukan penanganan data tidak seimbang dengan RUS merupakan model dengan nilai AIC terkecil kemudian yang kedua model tanpa penanganan data tidak seimbang yang didapatkan untuk memodelkan laju pertumbuhan penduduk kabupaten dan kota di Kalimantan yang akan dibuktikan selanjutnya dengan pengujian hipotesis. Keempat model yang diperoleh ini selanjutnya dilakukan pendugaan parameter yang secara lengkap tersaji pada Tabel 5.

Tabel 5. Nilai Dugaan Model Regresi Logistik Biner

Penanganan Data Tidak Seimbang	**Nilai Dugaan**		**Wald**	**p-val**	**Odds rat.**
Baseline	β_0	13.98	1.29	0.19	1.18×10^6
	X_1	1.85×10^{-2}	1.81	0.07	1.02
	X_2	-3.81×10^{-2}	-0.53	0.59	0.96
	X_3	-6.13×10^{-3}	-0.10	0.92	0.99
	X_4	-1.65×10^{-3}	-1.77	0.07	0.98
	X_5	-6.07×10^{-1}	-1.09	0.27	0.54
	X_6	-1.51×10^{-1}	-1.16	0.24	0.86
	X_7	-4.33×10^{-5}	-1.88	0.06	0.99
ROS	β_0	6.63	0.66	0.51	762.29
	X_1	2.72×10^{-2}	2.51	0.01*	1.03
	X_2	-3.71×10^{-2}	-0.61	0.54	0.96
	X_3	-2.71×10^{-1}	-1.48	0.13	0.76

	X_4	-1.80×10^{-3}	-2.33	0.02*	0.99
	X_5	-1.26×10^{-1}	-0.25	0.81	0.88
	X_6	-6.77×10^{-2}	-0.56	0.57	0.93
	X_7	-5.32×10^{-5}	-2.48	0.01*	0.99
RUS	β_0	6.61	0.59	0.55	742.14
	X_1	1.65×10^{-2}	1.53	0.12	1.02
	X_2	-7.98×10^{-2}	-0.97	0.33	0.92
	X_3	-2.80×10^{-2}	-0.37	0.71	0.97
	X_4	-2.03×10^{-3}	-1.08	0.28	0.99
	X_5	-1.67×10^{-1}	-0.28	0.78	0.85
	X_6	-4.49×10^{-2}	-0.32	0.75	0.96
	X_7	-4.43×10^{-5}	-1.77	0.08	0.99
SMOTE	β_0	22.248	2.19	0..03 *	4.59×10^9
	X_1	0.024	2.12	0.03*	1.02
	X_2	0.022	0.37	0.71	1.02
	X_3	-0.019	-0.31	0.76	0.98
	X_4	-0.002	-2.04	0.04*	0.99
	X_5	-1.121	-2.07	0.03*	0.33
	X_6	-0.206	-1.77	0.07	0.81
	X_7	-0.00006	-2.43	0.01*	0.99

Note: signifikan pada $\alpha = 5\%$.

Berdasarkan Tabel 5, model regresi logistik biner dengan empat penanganan data tidak seimbang diperoleh model yang secara matematis dapat ditulis sebagai berikut.

$$Y_{baseline} = 13.98 + 1.85 \times 10^{-2}X_1 - 3.81 \times 10^{-2}X_2 - 6.13 \times 10^{-3}X_3 - 1.65 \times 10^{-3}X_4 - 6.07 \times 10^{-1}X_5 - 1.51 \times 10^{-1}X_6 - 4.33 \times 10^{-5}X_7$$

$$Y_{ROS} = 6.63 + 2.72 \times 10^{-2}X_1 - 3.71 \times 10^{-2}X_2 - 2.71 \times 10^{-1}X_3 - 1.80 \times 10^{-3}X_4 - 1.26 \times 10^{-1}X_5 - 6.77 \times 10^{-2}X_6 - 5.32 \times 10^{-5}X_7$$

$$Y_{RUS} = 6.61 + 1.65 \times 10^{-2}X_1 - 7.98 \times 10^{-2}X_2 - 2.80 \times 10^{-2}X_3 - 2.03 \times 10^{-3}X_4 - 1.67 \times 10^{-1}X_5 + -4.49 \times 10^{-2}X_6 - 4.43 \times 10^{-5}X_7$$

$$Y_{SMOTE} = 22.248 + 0.024X_1 + 0.022X_2 - 0.019X_3 - 0.002X_4 - 1.121X_5 - 0.206X_6 - 0.00006X_7$$

Ukuran hubungan antara peluang terjadinya suatu peristiwa atau *odds ratio*. Dengan mengambil contoh pada interpretasi nilai *odds ratio* model regresi logistik biner dengan penanganan data tidak seimbang RUS pada Tabel 6. Sehingga, untuk variable X_1 atau jumlah penduduk di kabupaten/ kota di Kalimantan memiliki risiko sebesar 1.02 kali lebih tinggi dalam mempengaruhi laju pertumbuhan penduduk.

Selanjutnya, dilakukan pengujian hipotesis yang pertama adalah dengan uji parsial dengan Wald dimana hipotesis awal $H_0 = \beta_j = 0$ atau variael prediktor ke-j untuk $j = 1,2,\dots,7$ tidak memiliki pengaruh signifikan terhadap variabel respon dan hipotesis tandingan $H_1 = \beta_j \neq 0$ atau variabel prediktor ke-j untuk $j = 1,2,\dots,7$ memiliki pengaruh signifikan terhadap variabel respon. Pengambilan keputusan ini juga didasarkan pada nilai p-value dimana jika nilai $p - value \leq \alpha$ maka akan menolak H_0. Dimana hasil uji ini tersaji pada Tabel 5 dengan variabel yang signifikan mempengaruhi variabel respon ditunjukkan pada variabel prediktor dengan nilai p-value dengan *asterisk*.

Pengujian kedua yaitu masing-masing model yang diperoleh ini dilakukan pengujian hipotesis terhadap kesesuaian model atau *goodness-of-fit* dengan menggunakan uji Hosmer Lemeshow dimana hipotesis awal H_0 adalah model regresi logistik cocok dengan data dan hipotesis tandingan H_1 adalah model regresi tidak cocok dengan data. Pengambilan keputusan ini didasarkan pada nilai p-value jika nilai $p - value \leq \alpha$ maka akan menolak H_0. Uji hipotesis kedua adalah uji simultan dengan uji rasio likelihood atau uji G dimana hipotesis awal H_0 adalah semua koefisien regresi variabel prediktor dalam model adalah nol dan hipotesis tandingan H_1 adalah setidaknya ada satu koefisien regresi variabel prediktor dalam model adalah tidak nol. Pengambilan keputusan ini juga didasarkan pada nilai p-value dimana jika nilai $p - value \leq \alpha$ maka akan menolak H_0. Secara umum hasil uji kedua pengujian ini disajikan pada Tabel 6.

Tabel 6. Uji Hipotesis Model

Penanganan Data Tidak Seimbang	Uji Hosmer Lemeshow		Uji G	
	p-value	Keputusan	p-value	Keputusan
Baseline	0.4218	Gagal Tolak H_0	0.2597	Gagal Tolak H_0

ROS	0.3389	Gagal Tolak H_0	0.0089	Tolak H_0
RUS	0.0883	Gagal Tolak H_0	0.3040	Gagal Tolak H_0
SMOTE	0.2881	Gagal Tolak H_0	0.0034	Tolak H_0

Model dengan keempat penanganan data tidak seimbang yang telah dilakukan pengujian hipotesis menunjukkan bahwa model yang memenuhi asumsi adalah model regresi logistik biner dengan tanpa penanganan atau *baseline* dan penanganan data tidak seimbang dengan RUS. Pada uji Hosmer Lemeshow menunjukkan bahwa tidak memberikan bukti statistik yang signifikan untuk menunjukkan bahwa model terbaik yang diperoleh tidak sesuai dengan data atau bahwa model yang diperoleh sesuai atau konsisten dengan hasil aktual yang diamati. Sedangkan pada uji G kedua model tersebut menunjukkan tidak terdapat bukti statistik yang cukup untuk menyatakan bahwa model terbaik yang diperoleh memiliki variabel prediktor yang secara tidak bersamaan mempengaruhi variabel respons.

Sehingga, kedua model terbaik yang diperoleh selanjutnya diterapkan ada data uji untuk dilakukan validasi prediksi. Hasil perbandingan akurasi prediksi model baik pada data latih dan data uji ini tersaji secara lengkap pada Tabel 7.

Tabel 7. Perbandingan Hasil Prediksi

Penanganan Data Tidak Seimbang	**Data**	**Matriks evaluasi** (%)		
		Akurasi	***Balanced accuracy***	**ROC**
Baseline	Latih	66.7	53.1	76.51
	Uji	72.7	62.5	82.14
ROS	Latih	41.4	41.4	19.20
	Uji	35.7	35.7	95.92
RUS	Latih	37.5	37.5	24.02
	Uji	37.5	37.5	59.38
SMOTE	Latih	36.1	34.4	18.05
	Uji	46.7	43.8	58.93

Matriks evaluasi *balanced accuracy* dan nilai ROC (*receiver operating characteristic curve*) dipilih pada penelitian ini denga mmepertimbangkan ke-*robust*-an matriks evaluasi tersebut terhadap prediksi data tidak seimbang. Sehingga berdasarkan hasil perbandingan evaluasi pada Tabel 7 menunjukkan bahwa model regresi logistik biner dengan tanpa penanganan data tidak seimbang adalah yang terbaik. Hal ini ditunjukkan oleh nilai akurasi, *balanced accuracy*, dan ROC yang seimbang baik pada data latih maupun pada data uji. Hal ini, mengindikasikan bahwa dalam kasus laju pertumbuhan penduduk kabupaten / kota di Kalimantan menunjukkan karakteristik asli data yang baik, dimana intervensi sampling justru mengganggu pola data asli yang merupakan representative untuk model. Temuan ini menunjukkan bahwa evaluasi komprehensif pada data tidak seimbang dan membuktikan bahwa asumsi *resampling* tidak secara otomatis akan meningkatkan kinerja model.

Hasil penelitian terhadap perbandingan penanganan data tidak seimbang yang diusulkan oleh penelitian ini secara umum memberikan wawasan baru terhadap dalam membandingkan kinerja klasifikasi yang lebih lanjut terkait adanya penanganan data tidak seimbang. Selain itu, juga memberikan wawasan baru terhadap faktor-faktor penyebab laju pertumbuhan penduduk kabupaten dan kota di Kalimantan yang dipengaruhi oleh jumlah penduduk, persentase penduduk miskin, persentase jumlah penduduk, kepadatan penduduk, lama sekolah dari penduduknya, pengeluaran perkapita untuk makanan, dan jumlah penduduk bekerja. Lebih lanjut hasil penelitian ini diharapkan mampu memberikan alternative pengetahuan terhadap hal yang menyebabkan laju pertumbuhan penduduk kabupaten dan kota di Kalimantan sehingga dapat dilakukan upaya dan pemilihan penanganan terbaik oleh pemangku kebijakan.

PENUTUP

SIMPULAN

Penelitian ini menganalisis laju pertumbuhan penduduk kabupaten / kota di Kalimantan menggunakan model regresi logistik biner dengan masalah data tidak seimbang. Empat skenario penanganan data tidak seimbang diusulkan dalam penelitian ini dengan tujuan untuk memberikan perbandingan kinerja terbaik, yaitu tanpa

penanganan atau baseline, kedua dengan metode random over sampling (ROS), random undersampling (RUS), dan synthetic minority oversampling technique (SMOTE). Secara umum hasil penelitian menunjukkan bahwa model regresi logistik dengan tanpa penanganan data tidak seimbang memberikan akurasi, *balanced accuracy*, dan nilai ROC terbaik baik pada data latih sebesar 66.7%, 53.1%, dan 76.61% maupun data uji sebesar 72.7%, 62.5%, dan 82.14% mengungguli metode ROS, RUS, dan SMOTE. Temuan ini mengindikasikan bahwa pada data laju pertumbuhan penduduk di Kalimantan memiliki karakteristik intrinsik dan relevansi variabel prediktor yang lebih efektif pada distribusi data asli dibandingkan dengan *re-sampling*.

SARAN

Penelitian lanjutan yang mampu dikembangkan berdasarkan penelitian yang telah dilakukan ini bisa dengan mempertimbangkan variabel prediktor lain yang relevan dengan laju pertumbuhan penduduk sehingga dapat memperkaya model dan memberikan wawasan yang lebih komprehensif. Selain itu, bisa melalui uji generalisasi dimana model pada data laju pertumbuhan penduduk dari periode waktu atau wilayah di luar Kalimantan.

DAFTAR PUSTAKA

Bagui, S., & Li, K. (2021). Resampling imbalanced data for network intrusion detection datasets. *Journal of Big Data*, *8*(1), 6.

Buya, S., Tongkumchum, P., & Owusu, B. (2020). Modelling of land-use change in Thailand using binary logistic regression and multinomial logistic regression. *Arabian Journal of Geosciences*, *13*, 437. https://doi.org/10.1007/s12517-020-05451-2

Carrington, A. M., Manuel, D. G., Fieguth, P. W., Ramsay, T., Osmani, V., Wernly, B., Bennett, C., Hawken, S., Magwood, O., & Sheikh, Y. (2022). Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 329–341.

Cavanaugh, J. E., & Neath, A. A. (2019). The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *Wiley Interdisciplinary Reviews: Computational Statistics*, *11*(3), e1460.

Chen, L., Chen, T., Lan, T., Chen, C., & Pan, J. (2023). The contributions of population distribution, healthcare resourcing, and transportation infrastructure to spatial accessibility of health care. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, *60*, 00469580221146041.

De, H., & Acquah, G. (2010). Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship. *Journal of Development and Agricultural Economics*, *2*, 1–6.

Ependi, U., Rochim, A. F., & Wibowo, A. (2023). A hybrid sampling approach for improving the classification of imbalanced data using ROS and NCL methods. *International Journal of Intelligent Engineering and Systems*, *16*(3), 345–361.

Fatimah, F., Fitrianto, A., Indahwati, I., Erfiani, E., & Khikmah, K. N. (2023). Synthetic Minority Oversampling Technique Pada Model Logit dan Probit Status Pengangguran Terdidik. *Jambura Journal of Mathematics*, *5*(1), 166–178.

Harris, J. K. (2021). Primer on binary logistic regression. *Family Medicine and Community Health*, *9*(Suppl 1), e001290.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Iori, M., Di Castelnuovo, C., Verzellesi, L., Meglioli, G., Lippolis, D. G., Nitrosi, A., Monelli, F., Besutti, G., Trojani, V., & Bertolini, M. (2022). Mortality prediction of COVID-19 patients using radiomic and neural network features extracted from a wide chest X-ray sample size: A robust approach for different medical imbalanced scenarios. *Applied Sciences*, *12*(8), 3903.

Joshi, R. D., & Dhakal, C. K. (2021). Predicting type 2 diabetes using logistic regression and machine learning approaches. *International Journal of Environmental Research and Public Health*, *18*(14), 7346.

Khikmah, K. N., Indahwati, I., Fitrianto, A., Erfiani, E., & Amelia, R. (2022). Backwards stepwise binary logistic regression for determination population growth rate factor in Java Island. *Jambura Journal of Mathematics*, *4*(2), 177–187.

Khushi, M., Shaukat, K., Alam, T. M., Hameed, I. A., Uddin, S., Luo, S., Yang, X., & Reyes, M. C. (2021). A comparative performance analysis of data resampling methods on imbalance medical data. *Ieee Access*, *9*, 109960–109975.

Kumar, P., Bhatnagar, R., Gaur, K., & Bhatnagar, A. (2021). Classification of imbalanced data: review of methods and applications. *IOP Conference Series: Materials Science and Engineering*, *1099*(1), 012077.

Matharaarachchi, S., Domaratzki, M., & Muthukumarana, S. (2024). Enhancing SMOTE for imbalanced data with abnormal minority instances. *Machine Learning with Applications*, *18*, 100597.

Mishra, S. (2017). Handling imbalanced data: SMOTE vs. random undersampling. *Int. Res. J. Eng. Technol*, *4*(8), 317–320.

O'Sullivan, J. N. (2023). Demographic delusions: World population growth is exceeding most projections and jeopardising scenarios for sustainable futures. *World*, *4*(3), 545–568.

Pandey, B., Brelsford, C., & Seto, K. C. (2022). Infrastructure inequality is a characteristic of urbanization. *Proceedings of the National Academy of Sciences*, *119*(15), e2119890119.

Portet, S. (2020). A primer on model selection using the Akaike Information Criterion. *Infectious Disease Modelling*, *5*, 111–128.

Pradipta, G. A., Wardoyo, R., Musdholifah, A., Sanjaya, I. N. H., & Ismail, M. (2021). SMOTE for handling imbalanced data problem: A review. *2021 Sixth International Conference on Informatics and Computing (ICIC)*, 1–8.

Sadigov, R. (2022). Rapid growth of the world population and its socioeconomic results. *The Scientific World Journal*, *2022*(1), 8110229.

Thölke, P., Mantilla-Ramos, Y.-J., Abdelhedi, H., Maschke, C., Dehgan, A., Harel, Y., Kemtur, A., Berrada, L. M., Sahraoui, M., & Young, T. (2023). Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *NeuroImage*, *277*, 120253.

Van den Goorbergh, R., van Smeden, M., Timmerman, D., & Van Calster, B. (2022). The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *Journal of the American Medical Informatics Association*, *29*(9), 1525–1534.

Wongvorachan, T., He, S., & Bulut, O. (2023). A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. *Information*, *14*(1), 54.

Zaidi, A., & Al Luhayb, A. S. M. (2023). Two statistical approaches to justify the use of the logistic function in binary logistic regression. *Mathematical Problems in Engineering*, *2023*(1), 5525675.