Xplore (2022), Vol. 11, No. 2, 157-167
DOI: https://doi.org/10.29244/xplore.v11i2.922

Xplore: Journal of Statistics
Available at https://journal.stats.id/index.php/xplore
e-ISSN: 2655-2744

Kajian Metode Pohon Model Logistik (*Logistic Model Tree*) dengan Penanganan Ketakseimbangan Data*

Akmala Firdausi[1], Aam Alamudi[2‡], and Kusman Sadik[3]

[1,2,3]Department of Statistics, IPB University, Indonesia
‡corresponding author: aamalamudi@gmail.com

Abstract

Logistic model tree is a nonparametric modelling method that combines decision tree with linear logistic regression. Logistic model tree handles multicollinearity well, but is not immune to problems that arise due to data imbalance. This study was carried to compare the performance of undersampling, SMOTE, and ROSE in handling imbalanced data when used in tandem with logistic model tree. The data used in the simulation was obtained by generating random numbers following the Bernoulli distribution as the response variable and the Bivariate Normal distribution as the explanatory variables, based on five different imbalance levels. Comparisons done on the AUC value showed that logistic model trees built with methods to handle imbalanced data performed better than logistic model trees built without applying any such method on every level of tested data imbalance in classifying objects. Among those, logistic model trees built with ROSE performed better than logistic model trees built with other methods. On datasets with low level of imbalance, the performance of logistic model trees built with ROSE and undersampling do not significantly differ.

Keywords: imbalanced data handling, logistic model tree, ROSE, SMOTE, undersampling

* Received: Jan 2022; Reviewed: Jan 2022; Published: May 2022

1. Pendahuluan

Klasifikasi adalah kegiatan mengidentifikasi label atau kelas dari suatu amatan dengan mempertimbangkan karakteristik atribut-atribut (peubah) yang melekat pada amatan tersebut. Model klasifikasi yang ideal adalah model yang mampu mengklasifikasi seluruh amatan ke dalam kelas yang tepat. Berbagai jenis model klasifikasi terus dikembangkan untuk mencapai kondisi ideal tersebut. Beberapa di antara model tersebut adalah model pohon keputusan (*decision tree*) dan model regresi logistik linier.

Pohon keputusan adalah model sekuensial yang menggabungkan serangkaian tes sederhana logis. Setiap tes membandingkan sebuah peubah numerik dengan sebuah nilai batas atau peubah nominal terhadap serangkaian kemungkinan nilai. Pohon keputusan dikenal mampu menghasilkan model dengan bias yang rendah, namun memiliki keragaman yang tinggi. Sebaliknya, regresi logistik linier adalah metode klasifikasi yang menghasilkan model dengan ragam yang rendah, tetapi biasnya cukup tinggi. Regresi logistik linier secara umum memiliki dikembangkan menggunakan aturan yang sama dengan regresi linier biasa, hanya saja keluarannya berbentuk biner, sehingga model parametrik serta asumsi-asumsinya pun ikut berbeda (Hosmer dan Lemeshow 2013).

Logistic model tree (pohon model logistik) adalah metode pemodelan nonparametrik yang mengombinasikan metode pohon keputusan dengan regresi logistik linier. Pohon model logistik dibangun dengan membangun pohon keputusan C4.5, lalu membangun model regresi logistik pada masing-masing nodus terminalnya. Landwehr *et al.* (2005) mengembangkan pohon model logistik dengan tujuan mendapatkan kelebihan-kelebihan dari kedua metode tersebut. Pohon model logistik dapat menangani multikolinearitas dengan baik, mengingat multikolinearitas bukanlah masalah bagi metode pohon keputusan (Kotsiantis 2013). Pohon model logistik juga mampu memberikan informasi peluang sebuah masukan merupakan anggota masing-masing kelas. Sayangnya, sama seperti kedua metode tersebut, pohon model logistik tidak kebal terhadap masalah yang diakibatkan oleh ketakseimbangan data.

Kegagalan dalam mengklasifikasikan data ketika estimasi model didasarkan pada data latih yang sebarannya condong ke satu sisi (tidak simetris) merupakan salah satu masalah yang paling banyak didokumentasikan pada literatur. Ketika kondisi tersebut terjadi, metode klasifikasi standar umumnya kewalahan menangani kelas yang sering muncul dan cenderung mengabaikan kelas yang jarang muncul. Hal ini tentu dapat menjadi masalah yang besar, terutama ketika kepentingan utama peneliti adalah mendeteksi kasus yang jarang muncul tersebut, misalnya pada kasus deteksi penyakit kanker atau kredit macet. King dan Zen (2001, dalam Menardi dan Torelli 2014) menyatakan bahwa metode regresi logistik tidak disarankan ketika kelas data tidak seimbang, karena peluang bersyarat dari kelas minoritas cenderung ter-*underestimate*. Pada kondisi ekstrem, algoritma pohon keputusan cenderung memangkas pohon ke akar sehingga hampir seluruh contoh terklasifikasikan sebagai anggota kelas mayoritas (Kotsiantis 2013). Karenanya, sangat penting bagi masalah ini untuk ditangani.

Penanganan ketakseimbangan data secara umum terbagi ke dalam tiga pendekatan, yaitu pendekatan *undersampling*, *oversampling*, dan pembangkitan data

sintetis. Pendekatan dengan pembangkitan data sintetis dipelopori oleh SMOTE yang dicetuskan oleh Chawla *et al.* (2002), dengan klaim bahwa metode tersebut bekerja lebih baik daripada kombinasi dari dua pendekatan lainnya. Namun, studi yang dilakukan oleh Van Hulsen dan Khoshgoftaar (2009) menunjukkan bahwa *undersampling* acak kelas minoritas secara umum menghasilkan performa klasifikasi yang lebih baik dibandingkan kedua metode tersebut. Meski begitu, metode dengan pendekatan pembangkitan data sintetis terus dikembangkan. Salah satu metode penanganan data tak seimbang dengan pendekatan pembangkitan data sintetis yang dikenal memiliki performa cukup baik adalah ROSE (Lunardon et al. 2014). ROSE mengombinasikan *undersampling*, *oversampling*, dan pembangkitan data sintetik, sehingga mampu menangani masalah ketidakseimbangan data lebih baik dari *undersampling*.

Selain penerapan metode penanganan ketakseimbangan data, penggunaan ukuran evaluasi model yang tepat juga sangat penting ketika pemodelan dilakukan dengan data yang tidak seimbang. Ukuran evaluasi yang tidak tepat akan mengarah pada misrepresentasi performa model. Akurasi dan sensitivitas, misalnya, tidak mampu menunjukkan kegagalan model dalam mengklasifikasikan amatan anggota kelas minoritas pada kasus data tak seimbang. Meski model tidak mampu mengklasifikasikan amatan tersebut, ia akan tetap terlihat memiliki performa yang baik karena jumlah amatan anggota kelas mayoritas yang terklasifikasikan dengan benar jauh melampaui jumlah amatan anggota kelas minoritas yang terklasifikasikan salah. Karenanya, kedua ukuran evaluasi ini biasa digunakan bersama-sama dengan spesifisitas yang dapat menunjukkan kebaikan model dalam mengklasifikasikan amatan anggota kelas minoritas. Ukuran evaluasi lain yang dapat merepresentasikan kemampuan klasifikasi model yang dibangun dengan data latih tak seimbang dengan baik dan singkat adalah luasan di bawah kurva (*Area Under Curve*, AUC) dari *Receiving Operating Characteristic* (ROC).

Penelitian ini dilakukan dengan menggunakan data simulasi untuk membandingkan kemampuan pohon model logistik dalam mengklasifikasikan amatan dengan dan tanpa penanganan ketakseimbangan data pada tingkat ketakseimbangan dan jumlah amatan tertentu. Metode penanganan ketakseimbangan data yang digunakan adalah *undersampling*, SMOTE, dan ROSE. Performa dari model yang dihasilkan kemudian dibandingkan menggunakan AUC. AUC diketahui mampu mengukur performa model dengan lebih akurat pada kasus data dengan kelas tak seimbang dibandingkan ukuran evaluasi lain yang umum digunakan.

Penelitian ini dilakukan untuk membandingkan performa pohon model logistik dengan dan tanpa penerapan metode penanganan ketakseimbangan data bagi data dengan distribusi kelas yang tak seimbang, serta untuk m embandingkan performa metode penanganan ketakseimbangan data *undersampling*, SMOTE, dan ROSE pada pohon model logistik.

2. Metodologi

2.1 Bahan dan Data

Data yang digunakan pada penelitian ini adalah beberapa gugus data bangkitan dengan tingkat ketakseimbangan yang berbeda-beda. Pembangkitan gugus data

disadur dari skenario Lee (2000) yang telah disempurnakan oleh Menardi dan Torelli (2014). Vektor label kelas y untuk masing-masing gugus dibangkitkan mengikuti sebaran Bernoulli dengan nilai dan berturut-turut merupakan proporsi kelas dengan label = 1 dan = 0.

Tabel 1 Distribusi kelas gugus data simulasi

Skenario	Kelas mayoritas (y = 0)	Kelas mayoritas (y = 1)
P1	99%	1%
P2	95%	5%
P3	90%	10%
P4	75%	25%
P5	60%	40%

Tabel 1 menunjukkan distribusi kelas gugus data simulasi bagi lima jenis populasi yang dibangkitkan. Vektor atribut yang mengikuti sebaran normal ganda berdimensi 2 kemudian dibangkitkan mengikuti sebaran normal dengan persamaan berikut.

$$(x,y)\ s.t.\begin{cases} \boldsymbol{x}\sim\boldsymbol{N_2}\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right), & y=0\\ \boldsymbol{x}\sim\boldsymbol{N_2}\left(\begin{pmatrix}1\\1\end{pmatrix},\begin{pmatrix}1 & -0.5\\-0.5 & 1\end{pmatrix}\right), & y=1\end{cases}$$

Setiap gugus data dibangkitkan dengan = 10000 dengan ulangan = 1000. Antara peubah-peubah dan peubah tidak ada hubungan fungsional, mengingat pada kasus klasifikasi tidak diperlukan adanya hubungan fungsional antara peubah penjelas dan peubah respons.

2.2 Prosedur Analisis Data

Penelitian dilakukan dengan bantuan perangkat lunak R versi 4.1.1 (*Kick Things*) dengan environment RStudio versi 1.4.1106 dan package MASS, ggplot2, caret, DMwR, ROSE, xlsx, dan Rmisc.

Tahapan analisis data bangkitan adalah sebagai berikut.

1. Membangkitkan peubah acak Y sebagai peubah respon sebanyak n. Peubah acak dibangkitkan mengikuti sebaran Bernoulli dengan nilai p dan q berturut-turut merupakan proporsi kelas mayoritas dan kelas minoritas yang telah ditentukan.
2. Membangkitkan peubah penjelas X sebanyak n. Peubah acak X dibangkitkan mengikuti sebaran normal ganda berdimensi dua dengan parameter vektor $\mu = [0 \quad 0]'$ dan $\sigma^2 = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$ untuk kelas mayoritas dan vektor $\mu = [0 \quad 0]'$ dan $\sigma^2 = \begin{bmatrix}1 & -0.5\\-0.5 & 1\end{bmatrix}$ untuk kelas minoritas.
3. Menggabungkan peubah penjelas dan peubah respon.
4. Membangun pohon model logistik

a. Membagi data ke dalam 10 lipat terstratifikasi
b. Untuk model tanpa penanganan ketakseimbangan data:
    i. Melakukan pemodelan dengan validasi silang 10 lipat terstratifikasi, di mana satu lipat berlaku sebagai data uji dan sembilan lipat lainnya berlaku sebagai data latih
    ii. Menghitung nilai akurasi, sensitivitas, spesifitas, dan AUC model
c. Untuk model tanpa penanganan ketakseimbangan data:
    i. Menangani masalah ketakseimbangan data pada data latih dengan metode *undersampling*
    ii. Melakukan pemodelan dengan validasi silang 10 lipat terstratifikasi, di mana satu lipat berlaku sebagai data uji dan sembilan lipat lainnya berlaku sebagai data latih
    iii. Menghitung nilai akurasi, sensitivitas, spesifisitas, dan AUC model
    iv. Mengulang langkah i–iii sebanyak 2 kali, masing-masing dengan menerapkan metode penanganan ketakseimbangan data SMOTE dan ROSE pada langkah i
d. Mengulang langkah i–iv hingga seluruh bin pernah digunakan sebagai data uji

5. Mengulang langkah 1–4 untuk masing-masing gugus data pada tiap skenario.
6. Membandingkan performa model dengan dan tanpa penanganan ketakseimbangan data lewat analisis eksplorasi. Performa model dengan penanganan ketakseimbangan data pada penelitian ini direpresentasikan oleh rata-rata nilai ukuran kebaikan model dari seluruh model yang dibangun dengan penanganan ketakseimbangan data (*undersampling*, SMOTE, dan ROSE) bagi masing-masing gugus data. Rataan nilai AUC digunakan sebagai tolok ukur utama kebaikan model. Apabila sebaran kedua model terlihat mirip, uji-t berpasangan dilakukan terhadap nilai AUC dari kedua model tersebut untuk menentukan model dengan performa terbaik di antara keduanya.
7. Membandingkan performa model dengan masing-masing metode penanganan ketakseimbangan data lewat analisis eksplorasi. Rataan nilai AUC digunakan sebagai tolok ukur utama kebaikan model. Apabila terdapat dua model dengan sebaran yang terlihat mirip dan memiliki nilai rataan AUC yang tidak jauh berbeda, uji-t berpasangan dilakukan terhadap nilai AUC dari kedua model tersebut untuk menentukan model dengan performa terbaik di antara keduanya.

3. Hasil dan Pembahasan

3.1 Deskripsi Data Bangkitan

Data bangkitan terdiri atas lima skenario yang dibangkitkan dengan sebaran yang sama tetapi dengan tingkat ketakseimbangan data yang berbeda-beda. Data bangkitan ini kemudian diseimbangkan dengan metode *undersampling,* SMOTE, dan

ROSE, sebelum dilakukan pemodelan. Tabel 2 menunjukkan rata-rata ukuran contoh yang digunakan untuk membangun model pada masing-masing skenario dengan masing-masing metode penanganan ketakseimbangan data berdasarkan kelasnya.

Tabel 2 Ukuran contoh

Model	Kelas Mayoritas		Kelas Minoritas	
	Rata-rata	Simpangan Baku	Rata-rata	Simpangan Baku
		Skenario P1		
Tanpa penanganan	9900	10,1023	100	10,1023
Undersampling	90	9,1041	90	9,1041
SMOTE	181	18,1842	181	18,1842
ROSE	4492	0	4508	0
		Skenario P2		
Tanpa penanganan	9501	21,8598	499	21,8598
Undersampling	450	19,6594	450	19,6594
SMOTE	900	39,3476	900	39,3476
ROSE	4492	0	4508	0
		Skenario P3		
Tanpa penanganan	9000	30,1670	1000	30,1670
Undersampling	899	27,1390	899	27,1390
SMOTE	1799	54,3006	1799	54,3006
ROSE	4492	0	4508	0
		Skenario P4		
Tanpa penanganan	7501	43,9496	2499	43,9496
Undersampling	2249	39,5631	2249	39,5631
SMOTE	4499	79,1093	4499	79,1093
ROSE	4492	0	4508	0,1044
		Skenario P5		
Tanpa penanganan	5999	49,5486	4001	49,5486
Undersampling	3601	44,5939	3601	44,5938
SMOTE	7203	89,1875	7203	89,1875
ROSE	4492	0	4508	0,1531

Model yang datanya diseimbangkan dengan metode *undersampling* (selanjutnya disebut model *undersampling*) secara konsisten memiliki rata-rata ukuran contoh yang paling kecil di seluruh skenario. Model yang dibangun dengan metode ROSE (selanjutnya disebut model ROSE) memiliki ukuran contoh dan distribusi kelas yang cenderung konstan pada seluruh skenario. Pada skenario P1, P2, dan P3, urutan model dengan rata-rata ukuran contoh terkecil hingga terbesar secara berturut-turut adalah model *undersampling*, model yang dibangun dengan metode SMOTE (selanjutnya disebut model SMOTE), kemudian disusul oleh model ROSE dan model tanpa penanganan ketakseimbangan data (selanjutnya disebut model tanpa penanganan) yang memiliki rata-rata ukuran contoh yang sama. Pada skenario P4, model *undersampling* memiliki rata-rata ukuran contoh terkecil, disusul oleh ketiga

model lain yang rata-rata ukuran contohnya sama. Pada skenario P5, model *undersampling* memiliki rata-rata ukuran contoh terkecil, disusul oleh model ROSE dan model tanpa penanganan yang memiliki rata-rata ukuran contoh sama, dan model SMOTE dengan rata-rata ukuran contoh terbesar.

Selain pada model ROSE, ukuran contoh yang digunakan untuk membangun model berbanding lurus dengan proporsi jumlah contoh anggota kelas minoritas. Semakin besar proporsi jumlah contoh anggota kelas minoritas, semakin besar pula ukuran contoh setelah diseimbangkan. Jumlah data pada masing-masing kelas setelah diseimbangkan umumnya sama. Ukuran data latih akhir dari data yang diseimbangkan oleh metode ROSE secara *default* disamakan dengan jumlah ukuran data latih asli. Metode ROSE membangkitkan data sintetis dengan pendekatan *smoothed bootstrap*, di mana data tersebut dibangkitkan berdasarkan fungsi sebaran peluang yang dibangun berdasarkan data yang hendak diseimbangkan. Hal ini berakibat pada jumlah contoh yang sama pada masing-masing kelas di seluruh skenario setelah diseimbangkan, karena peubah penjelas pada seluruh skenario dibangkitkan dengan sebaran yang sama.

3.2 Evaluasi Kebaikan Model

Secara keseluruhan, sebanyak 40.000 model dibangun pada penelitian ini. Evaluasi kebaikan model dilakukan dengan dua tahap. Tahap pertama dilakukan dengan membandingkan nilai akurasi, sensitivitas, spesifisitas, dan AUC dari model yang dibangun tanpa dan dengan metode penanganan ketakseimbangan data dari masing-masing skenario simulasi, dengan nilai AUC sebagai tolok ukur utama. Nilai kebaikan dari model yang dibangun dengan penanganan ketakseimbangan data pada penelitian ini direpresentasikan oleh rataan dari nilai-nilai kebaikan dari model *undersampling,* SMOTE, dan ROSE. Apabila nilai AUC dari model yang dibangun dengan penanganan ketakseimbangan data lebih tinggi daripada nilai AUC dari model yang dibangun tanpa penanganan ketakseimbangan data, maka berikutnya dilakukanlah perbandingan terhadap model-model yang dibangun dengan penanganan ketakseimbangan data untuk menentukan pasangan model dengan metode penanganan ketakseimbangan data yang memiliki performa terbaik.

Sebaliknya, pada model yang dibangun dengan penanganan ketakseimbangan data, nilai akurasi, sensitivitas, spesifisitas, dan AUC pada setiap skenario cenderung seimbang. Model yang dibangun dengan penanganan ketakseimbangan data secara konsisten memiliki performa secara signifikan lebih baik dibandingkan dengan model yang dibangun tanpa penanganan ketakseimbangan data pada seluruh skenario, sehingga selanjutnya dicek performa model dengan tiap metode penanganan ketakseimbangan data pada masing-masing skenario.

Tabel 3 Nilai kebaikan model tanpa dan dengan penanganan ketakseimbangan data

Model	Ukuran Kebaikan Model			
	Akurasi	**Sensitivitas**	**Spesifisitas**	**AUC**
		P1		
Tanpa penanganan	0,98998	0,99999	0,00004	0,50003
Dengan penanganan	0,76279	0,76216	0,82559	**0,79388**
		P2		
Tanpa penanganan	0,94985	0,99938	0,00826	0,50399
Dengan penanganan	0,76452	0,76077	0,83573	**0,79825**
		P3		
Tanpa penanganan	0,90130	0,98712	0,12804	0,55759
Dengan penanganan	0,76704	0,75898	0,83954	**0,79926**
		P4		
Tanpa penanganan	0,82026	0,90304	0,57165	0,73735
Dengan penanganan	0,77793	0,75578	0,84438	**0,80008**
		P5		
Tanpa penanganan	0,79782	0,81746	0,76827	0,79287
Dengan penanganan	0,77245	0,75336	0,88171	**0,79968**

Tabel 4 Nilai kebaikan model dengan penanganan ketakseimbangan data

Model	Ukuran Kebaikan Model			
	Akurasi	**Sensitivitas**	**Spesifisitas**	**AUC**
		P1		
Undersampling	0,76721	0,76667	0,82088	0,79378
SMOTE	0,76313	0,76258	0,81693	0,78976
ROSE	0,75804	0,75722	0,83897	**0,79810**
		P2		
Undersampling	0,77105	0,76807	0,82784	0,79795
SMOTE	0,76215	0,75842	0,83301	0,79572
ROSE	0,76036	0,75583	0,84634	**0,80108**
		P3		
Undersampling	0,77304	0,76652	0,83171	0,79911
SMOTE	0,76315	0,75462	0,83998	0,79730
ROSE	0,76492	0,75581	0,84692	**0,80136**
		P4		
Undersampling	0,78140	0,76217	0,83912	0,80065
SMOTE	0,77343	0,74894	0,84693	0,79794
ROSE	0,77895	0,75622	0,84709	**0,80165**
		P5		
Undersampling	0,79323	0,75971	0,84349	**0,80160**
SMOTE	0,78550	0,74445	0,84704	0,79574
ROSE	0,79258	0,75593	0,84745	**0,80169**

Tabel 4 menunjukkan rataan nilai kebaikan model yang dibangun dengan metode penanganan ketakseimbangan data *undersampling,* SMOTE, dan ROSE. Berdasarkan rataan nilai AUC, model ROSE memiliki performa paling baik pada skenario P1, P2, P3, dan P4, disusul oleh model *undersampling* dan model SMOTE. Pada skenario P5, rata-rata performa model ROSE dan model undersampling tidak berbeda signifikan pada taraf nyata $\alpha = 0.05$ (p-value = 0.1521277), sehingga kedua model tersebut dapat dikatakan sama baiknya dalam menangani ketakseimbangan data.

3.3 Waktu Komputasi

Faktor lain yang dapat diperhatikan dalam pembangunan model, terutama ketika performanya tidak jauh berbeda, adalah waktu komputasi. Tabel 5 menunjukkan waktu yang dibutuhkan untuk membangun masing-masing model, baik yang dibangun dengan menggunakan metode penanganan ketakseimbangan data *undersampling*, SMOTE, dan ROSE, maupun yang dibangun tanpa menggunakan metode tersebut.

Tabel 5 Waktu komputasi model

Model	Rata-rata (s)	Simpangan Baku (s)
	P1	
Tanpa penanganan	91,8667	12,2716
Undersampling	**0,3756**	0,0387
SMOTE	0,7924	0,1060
ROSE	10,5662	0,6332
	P2	
Tanpa penanganan	102,2853	40,2170
Undersampling	**1,0862**	0,1901
SMOTE	3,8962	0,4478
ROSE	10,7048	1,2301
	P3	
Tanpa penanganan	18,2451	22,6018
Undersampling	**2,0230**	0,1901
SMOTE	5,7627	0,4478
ROSE	10,5372	1,2301
	P4	
Tanpa penanganan	12,3575	0,7880
Undersampling	**5,0622**	0,3248
SMOTE	19,9875	1,8797
ROSE	10,5748	0,8082
	P5	
Tanpa penanganan	11,9350	0,8276
Undersampling	**8,4640**	0,6572
SMOTE	36,6293	7,3997
ROSE	10,7212	0,5231

Semakin tak seimbang data, semakin besar waktu komputasi yang dibutuhkan oleh model tanpa penanganan ketakseimbangan data. Sebaliknya, pada model yang dibangun dengan penanganan ketakseimbangan data, waktu komputasi akan bertambah seiring dengan semakin seimbangnya data yang digunakan. Hal ini terjadi karena semakin seimbang data maka semakin banyak pula jumlah contoh yang digunakan untuk membangun model.

Waktu komputasi secara konsisten berbanding lurus dengan ukuran contoh pada setiap skenario. Hal ini terlihat dari waktu komputasi model *undersampling* yang selalu lebih rendah dari waktu komputasi model lain pada seluruh skenario. Waktu komputasi yang dibutuhkan oleh model ROSE yang ukuran sampel serta distribusi kelasnya cenderung konstan juga kurang lebih sama, dengan nilai simpangan baku yang cukup rendah. Pada skenario P1, P2, dan P3, urutan waktu komputasi yang dibutuhkan model dari terkecil hingga terbesar sama dengan urutan ukuran contohnya.

Skenario P4 menunjukkan bahwa ketika ukuran sampel yang digunakan untuk membangun model sama, model SMOTE membutuhkan waktu komputasi yang lebih besar daripada model ROSE dan model tanpa penanganan ketakseimbangan data. Skenario P4 dan P5 menunjukkan bahwa model ROSE, meskipun harus melalui proses penanganan ketakseimbangan data terlebih dahulu sebelum membangun model, membutuhkan waktu komputasi yang paling kecil dibandingkan dengan model yang dibangun tanpa penanganan ketakseimbangan data.

4. Simpulan dan Saran

4.1 Simpulan

Semakin ekstrem kasus ketakseimbangan data, semakin penting bagi masalah tersebut untuk ditangani sebelum membangun model klasifikasi. Hasil penelitian ini menunjukkan bahwa pada kasus ekstrem, pohon model logistik yang dibangun dengan data tak seimbang tanpa dilakukan penanganan sebelumnya memiliki kemampuan klasifikasi yang setara dengan tebakan asal, sehingga metode penanganan ketakseimbangan data sangat penting untuk diterapkan.

Pohon model logistik dengan pengaplikasian metode penanganan ketakseimbangan data ROSE bekerja paling baik pada seluruh tingkat ketakseimbangan data. Hal ini mungkin terjadi karena ROSE menyeimbangkan data dengan membangkitkan data latih baru yang seluruh amatannya merupakan amatan sintetis, sehingga ia tidak terpengaruh gangguan yang ada pada data latih asli. Pada tingkat ketakseimbangan data sangat rendah, performa klasifikasi model yang dibangun dengan metode ROSE dan model yang dibangun dengan metode *undersampling* dalam tidak berbeda signifikan.

Waktu komputasi bagi pohon model logistik sangat dipengaruhi oleh jumlah amatan pada gugus data latih. Semakin banyak jumlah amatan, semakin banyak waktu yang dibutuhkan komputer untuk membangun model. Pohon model logistik yang dibangun dengan metode penanganan ketakseimbangan data *undersampling* secara konsisten memiliki jumlah amatan dan waktu komputasi yang terendah pada seluruh skenario. Pohon model logistik yang dibangun dengan penanganan ketakseimbangan data ROSE memiliki jumlah amatan dan waktu komputasi yang paling konsisten pada seluruh skenario.

4.2 Saran

Pengaturan *hyperparameters* pada pohon model logistik maupun SMOTE dan ROSE dapat berpengaruh pada performa klasifikasi, sehingga perlu dilakukan kajian lebih lanjut untuk pemodelan dengan pengaturan *hyperparameter* tersebut. Faktor kombinasi lain, seperti jumlah amatan, dapat ditambahkan pada skenario simulasi untuk memberi hasil studi yang lebih konkret. Selain metode undersampling, SMOTE, dan ROSE, masih banyak metode penanganan ketakseimbangan data lain yang dapat digunakan, seperti *oversampling*, WF-SMOTE, dan Borderline SMOTE.

Daftar Pustaka

Chawla NV, Hall LO, Bowyer KW, Kegelmeyer WP. 2002. SMOTE: Synthetic minority oversampling technique. Journal of Artificial Intelligence Research. 16: 321–357.

Chawla NV. 2003. C4.5 and imbalanced data sets: Investigating the effect of sampling method, probabilistic estimate, and decision tree structure. Di dalam: Chawla NV, editor. Proceedings of ICML 2003 Workshop on Learning from Imbalanced Data Sets (II); 2003 Agu 21. Washington DC (US): hlm 9-17.

Hosmer DW, Lemeshow S, Sturivant RX. 2013. Applied Logistic Regression. Hoboken (US): John Wiley & Sons Inc.

Kotsiantis SB. 2013. Decision trees: a recent overview. Artificial Intelligence Review. 39: 261-283.

Landwehr N, Hall M, Frank E. 2005. Logistic model tree. Machine Learning. 59: 161-205.

Lee SS. 2000. Noisy replication in skewed binary classification. Computational Statistics & Data Analysis. 34(2): 165-191.

Lunardon N, Menardi G, Torelli N. 2014. ROSE: A package for binary imbalanced learning. The R Journal. 6(1): 79-89.

Menardi G, Torelli N. 2014. Training and assessing classification rules with unbalanced data. Working Paper Series. 28(1): 92-122.

Van Hulse J, Khoshgoftaar T. 2009. Knowledge discovery from imbalanced and noisy data. Data & Knowledge Engineering. 68(12): 1513-1542.