Xplore, 2013, *Vol. 1(1):e9(1-6)* © 2013 Departemen Statistika FMIPA IPB

PENERAPAN SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE) TERHADAP DATA TIDAK SEIMBANG PADA PEMBUATAN MODEL KOMPOSISI JAMU

Rossi Azmatul Barro*†, Itasia Dina Sulvianti*, Farit Mochamad Afendi*
*Departemen Statistika, Institut Pertanian Bogor
†E-mail: roshi.azmatul@gmail.com

***Ringkasan*—As the times many people use herbal remedies (jamu) to address health issues. Herbal medicines are made from plants with a specific composition to produce certain properties, so a model is needed to be made in order to find the right formula to make herbal medicine with certain properties. In this study, the response being investigated is a potent herbal medicine in treating mood and behavior disorder. In this analysis, the model is developed using logistic regression. The accuracy of the model can be seen from the Area Under Curve (AUC). Imbalanced data on the response variable can cause the value of AUC become low. One of the ways to solve it is using Synthetic Minority Oversampling Technique (SMOTE). From this analysis, Nagelkerke R2 values generated by the model with SMOTE 3.2% lower than model without SMOTE. Nonetheless, the model with SMOTE is more accurate than model without SMOTE because has higher AUC value. The resulting AUC is equal to 0.976 for the model with SMOTE and 0.908 for model without SMOTE. The results show that SMOTE can increase the accuracy of the model for imbalanced data.**

***Keywords*-imbalance data, logistic regression, SMOTE**

I. PENDAHULUAN

A. Latar Belakang

Peningkatan aktivitas yang tidak diimbangi dengan asupan gizi yang cukup akan menyebabkan tubuh lebih mudah terserang penyakit. Perkembangan zaman juga membuat banyak orang mudah mengalami stres atau suasana hati yang tidak baik. Untuk mengurangi risiko terserang penyakit maupun suasana hati yang tidak baik, beberapa orang memilih untuk mengkonsumsi obat dan tidak sedikit yang memilih jamu. Jamu dipilih karena dianggap alami dan tidak memiliki efek samping yang berbahaya.

Badan Pengawas Obat dan Makanan (BPOM) menyatakan bahwa jamu adalah obat tradisional Indonesia. Obat tradisional adalah bahan atau ramuan yang berupa bahan tumbuhan, bahan hewan, bahan mineral, sediaan sarian (galenik) atau campuran dari bahan tersebut, yang secara turun-temurun telah digunakan untuk pengobatan berdasarkan pengalaman. Hal tersebut tercantum pada Pasal 1 Peraturan Kepala Badan POM No. HK.00.05.4.1384 Tahun 2005. Sebagian besar jamu dibuat menggunakan berbagai macam tanaman dengan khasiat yang bermacam-macam. Oleh karena itu, diperlukan model agar ditemukan formulasi yang pas untuk membuat jamu dengan khasiat tertentu. Pada penelitian ini khasiat yang diteliti sebagai respon model adalah adanya khasiat dalam mengatasi gangguan suasana hati dan perilaku.

Penelitian ini menggunakan model yang dibangun dengan regresi logistik. Metode tersebut cocok digunakan karena respon yang diamati berskala kategorik. Salah satu hal yang perlu diperhatikan dalam evaluasi model adalah tingkat akurasi sebuah model dalam memprediksi respon dengan benar. Kebaikan model dipengaruhi salah satunya oleh adanya keseimbangan antara kelas mayor dengan kelas minor. Kelas mayor adalah data yang ukuran kelasnya (jumlah amatan) lebih besar dari kelas minor berdasarkan peubah respon. Jika data yang digunakan untuk membuat model tidak seimbang maka akan meningkatkan salah klasifikasi kelas minor. Oleh karena itu, salah satu alternatif untuk meningkatkan akurasi model adalah melakukan *Synthetic Minority Oversampling Technique* (SMOTE) pada praposes.

B. Tujuan Penelitian

Penelitian ini bertujuan membandingkan model komposisi jamu yang berkhasiat dalam mengatasi gangguan suasana hati dan perilaku yang dihasilkan menggunakan regresi logistik melalui tahap SMOTE dengan model tanpa tahap SMOTE.

II. TINJAUAN PUSTAKA

A. SMOTE (Synthetic Minority Oversampling Technique)

Ketidakseimbangan data terjadi jika jumlah objek suatu kelas data lebih banyak dibandingkan dengan kelas lain. Kelas data yang objeknya lebih banyak disebut kelas mayor sedangkan lainnya disebut kelas minor. Pengaruh penggunaan data tidak seimbang untuk membuat model sangat besar pada hasil model yang diperoleh. Pengolahan algoritma yang tidak menghiraukan ketidakseimbangan data akan cenderung diliputi oleh kelas mayor dan mengacuhkan kelas minor [1].

Metode SMOTE diusulkan oleh [1] sebagai salah satu solusi dalam menangani data tidak seimbang dengan prinsip yang berbeda dengan Metode *oversampling* yang telah diusulkan sebelumnya. Bila Metode *oversampling* berprinsip memperbanyak pengamatan secara acak, Metode SMOTE menambah jumlah data kelas minor agar setara dengan kelas mayor dengan cara membangkitkan data buatan. Data

buatan atau sintesis tersebut dibuat berdasarkan *k*-tetangga terdekat (*k*-nearest neighbor). Jumlah *k*-tetangga terdekat ditentukan dengan mempertimbangkan kemudahan dalam melaksanakannya. Pembangkitan data buatan yang berskala numerik berbeda dengan kategorik. Data numerik diukur jarak kedekatannya dengan jarak Euclidean sedangkan data kategorik lebih sederhana yaitu dengan nilai modus. Perhitungan jarak antar contoh kelas minor yang peubahnya berskala kategorik dilakukan dengan rumus *Value Difference Metric* (VDM) yaitu [2]:

$$\boldsymbol{\Delta}(\mathbf{X},\mathbf{Y}) = \mathbf{w_x w_y} \sum_{\mathbf{i=1}}^{\mathbf{N}} \delta(\mathbf{x_i},\mathbf{y_i})^{\mathbf{r}} \tag{1}$$

dengan:

Δ(X, Y)	:	jarak antara amatan X dengan Y
w_x, w_y	:	bobot amatan (dapat diabaikan)
N	:	banyaknya peubah penjelas
R	:	bernilai 1 (jarak Manhattan) atau 2 (jarak Euclidean)
$\delta(x_i, y_i)^r$	:	jarak antar kategori, dengan rumus:

$$\delta(\mathbf{V_1},\mathbf{V_2}) = \sum_{\mathbf{i=1}}^{\mathbf{n}} \left| \frac{\mathbf{C_{1i}}}{\mathbf{C_1}} - \frac{\mathbf{C_{2i}}}{\mathbf{C_2}} \right|^{\mathbf{k}} \tag{2}$$

dengan:

$\delta(V_1, V_2)$	:	jarak antara nilai V_1 dan V_2
C_{1i}	:	banyaknya V_1 yang termasuk kelas i
C_{2i}	:	banyaknya V_2 yang termasuk kelas i
I	:	banyaknya kelas; i = 1,2,...,m
C_1	:	banyaknya nilai 1 terjadi
C_2	:	banyaknya nilai 2 terjadi
N	:	banyaknya kategori
K	:	konstanta (biasanya 1)

Prosedur pembangkitan data buatan untuk :

1) Data Numerik
   a) Hitung perbedaan antar vektor utama dengan *k*-tetangga terdekatnya.
   b) Kalikan perbedaan dengan angka yang diacak diantara 0 dan 1.
   c) Tambahkan perbedaan tersebut ke dalam nilai utama pada vektor utama asal sehingga diperoleh vektor utama baru.
2) Data Kategorik
   a) Pilih mayoritas antara vektor utama yang dipertimbangkan dengan k-tetangga terdekatnya untuk nilai nominal. Jika terjadi nilai samamaka pilih secara acak.
   b) Jadikan nilai tersebut data contoh kelas buatan baru.

B. Analisis Regresi Logistik

Analisis ini dapat mengetahui hubungan antar respon dengan satu atau lebih peubah penjelas. Tujuan penggunaan regresi logistik sama halnya dengan teknik membangun model dalam statistika [3]. Regresi logistik dapat juga disebut model logit karena fungsi transformasinya menggunakan logit. Untuk respon biner peubah Y dan peubah penjelas X, maka $\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$. Model regresi logistik adalah

$$\pi(\mathbf{x}) = \frac{\mathbf{exp}(\alpha + \beta\mathbf{x})}{\mathbf{1} + \mathbf{exp}(\alpha + \beta\mathbf{x})} \tag{3}$$

yang setara dengan log odd, disebut logit, yaitu

$$\mathbf{logit}[\pi(\mathbf{x})] = \mathbf{ln}\left[\frac{\pi(\mathbf{x})}{\mathbf{1} - \pi(\mathbf{x})}\right] = \mathbf{g}(\mathbf{x}) = \alpha + \beta\mathbf{x} \tag{4}$$

Metode umum pendugaan parameter regresi logistik adalah metode kemungkinan maksimum [3]. Untuk menerapkan metode ini, yang pertama harus dilakukan adalah membentuk fungsi kemungkinan:

$$\ell(\beta) = \prod_{\mathbf{i=1}}^{\mathbf{n}} \pi(\mathbf{x_i})^{\mathbf{y_i}} \left[\mathbf{1} - \pi(\mathbf{x_i})\right]^{\mathbf{1-y_i}} \tag{5}$$

Prinsip dari metode kemungkinan maksimum adalah dengan memaksimumkan fungsi kemungkinan yang secara matematis lebih mudah dengan memaksimumkan logaritma fungsi kemungkinan:

$$\begin{aligned}\mathbf{L}(\beta) &= \mathbf{ln}[\ell(\beta)] \\ &= \sum_{\mathbf{i=1}}^{\mathbf{n}} \{\mathbf{y_i ln}[\pi(\mathbf{x_i})] + (\mathbf{1} - \mathbf{y_i})\,\mathbf{ln}[\mathbf{1} - \pi(\mathbf{x_i})]\}\end{aligned} \tag{6}$$

Untuk mendapatkan nilai dugaan koefisien regresi logistik ($\widehat{\beta}$) dilakukan dengan penurunan L(β) terhadap β dan disamakan dengan nol.

Kesesuaian model digunakan untuk mengetahui peubah penjelas yang berpengaruh nyata terhadap respon. Pengujian parameter β secara bersama dengan uji-G yaitu uji nisbah kemungkinan. Uji-G untuk pengujian parameter β_j dengan hipotesis :

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$

H_1 : minimal salah satu $\beta_i \neq 0$, dengan $i = 1, 2, \ldots, p$

Statistik uji untuk uji G adalah :

$$G = -2ln\left(\frac{\text{fungsi kemungkinan tanpa peubah bebas}}{\text{fungsi kemungkinan dengan peubah bebas}}\right) \tag{7}$$

Jika H_0 benar, statistik G akan berdistribusi χ^2 dengan derajat bebas p. Oleh karena itu, jika H_0 ditolak, maka selanjutnya dilakukan uji Wald untuk menguji parameter β secara parsial. Hipotesis yang diujikan adalah :

$H_0 : \beta_j = 0$
$H_1 : \beta_j \neq 0$, dengan $i = 1, 2, \ldots, p$

Uji Wald dihitung dengan membandingkan pendugaan slope parameter maksimum kemungkinan dengan dugaan standar erornya, sebagai berikut:

$$W = \frac{\beta_j}{Se\left(\widehat{\beta_j}\right)} \tag{8}$$

C. Evaluasi Model

1) Area Under Curve (AUC): AUC adalah luas di bawah kurva yang dalam hal ini merupakan kurva *Receiver Operating Characteristic* (ROC). Menurut [4], kurva ROC menggambarkan performa pengklasifikasi secara dua dimensi. Kurva tersebut adalah plot peluang salah negatif (1-spesifitas) dengan prediksi benar positif (sensitifitas). Nilai sensitifitas dan spesifitas dapat dilihat pada Tabel I. Jika ingin membandingkan beberapa performa pengklasifikasi maka ROC dapat diubah ke dalam bentuk skalar salah satunya menjadi AUC. AUC adalah suatu bagian dari daerah satuan persegi yang nilainya antara 0 hingga 1. Nilai AUC semakin mendekati satu maka akurasi model atau klasifikasi semakin tinggi. AUC dapat dihubungkan dengan koefisien Gini dengan persamaan Gini + 1 = 2 x AUC.

Tabel I
KESESUAIAN KLASIFIKASI

Aktual	Prediksi Model	
	0	1
0	Benar (-) Spesifisitas	Salah (+)
1	Salah (-)	Benar (+) Sensitivitas

2) R^2 Nagelkerke: R^2 Nagelkerke mengukur tingkat keragaman respon yang dapat dijelaskan model dalam regresi logistik [5]. Rumus yang digunakan yaitu

$$R^2 = 1 - \left(\frac{\text{fungsi kemungkinan dengan hanya intersep}}{\text{fungsi kemungkinan dugaan model}}\right)^{\frac{2}{n}} \tag{9}$$

III. METODOLOGI

A. Data

Data yang digunakan dalam penelitian ini adalah data mengenai status penggunaan tanaman pada komposisi jamu untuk khasiat tertentu. Data ini terdiri dari 1002 jenis jamu di Indonesia yang terdaftar pada Badan Pengawas Obat dan Makanan. Tiap jenis jamu terdiri dari 294 peubah penjelas berupa tanaman. Peubah respon yang digunakan adalah khasiat jamu dengan kategori (1) jamu berkhasiat dalam mengatasi gangguan suasana hati dan perilaku dan kategori (0) jamu berkhasiat dalam mengatasi gangguan pencernaan. Seluruh peubah penjelas bersifat kategorik dengan dua kategori yaitu tanaman komposisi jamu dan bukan tanaman komposisi jamu. Terdapat 22 jamu atau sekitar 2.2% jamu yang memiliki khasiat dalam mengatasi gangguan suasana hati dan perilaku sedangkan 980 jamu sisanya tidak memiliki khasiat tersebut. Penelitian ini lebih memfokuskan pada model dalam memprediksi khasiat jamu untuk mengatasi gangguan suasana hati dan perilaku.

B. Metode Penelitian

Tahapan metode yang dilakukan adalah:

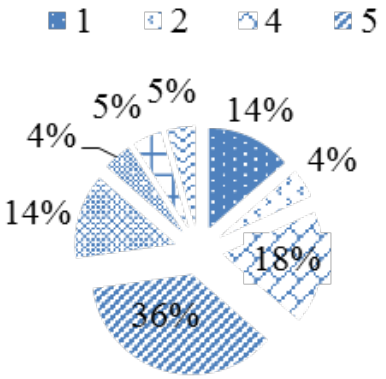
1) Melakukan deskripsi data untuk mengetahui gambaran umum data jamu yang diperoleh melalui diagram lingkaran dan batang.
2) Membangun model dengan regresi logistik dengan mencari nilai dugaan parameternya.
3) Melakukan pengujian parameter.
4) Mengevaluasi model dengan melihat nilai AUC dan R^2 Nagelkerke.
5) Melakukan SMOTE pada tahap praproses data jamu, yaitu:
   a) Menghitung jarak antar amatan pada kelas minor menggunakan rumus VDM.
   b) Menentukan nilai k yaitu 5 dan persentase oversampling sebesar 4200
   c) Dipilih satu contoh dari kelas minor secara acak.
   d) Menentukan amatan k tetangga terdekat dengan mengurut jarak contoh terpilih dengan semua amatan pada kelas minor.
   e) Data sintesis dibuat dengan menentukan nilai per peubah penjelasnya. Nilai tersebut diperoleh dari mayoritas nilai pada k tetangga terdekat. Jika semua peubah telah dibuat maka diperoleh satu amatan baru.
   f) Langkah c hingga e dilakukan berulang hingga banyaknya oversampling yang diinginkan telah tercapai.
6) Membangun model dengan data yang telah melalui tahap SMOTE.
7) Menguji parameternya.
8) Mengevaluasi tingkat akurasi model.
9) Membandingkan hasil model yang dihasilkan tanpa SMOTE dan dengan SMOTE dari AUC dan R^2 Nagelkerke masing-masing model.

IV. HASIL DAN PEMBAHASAN

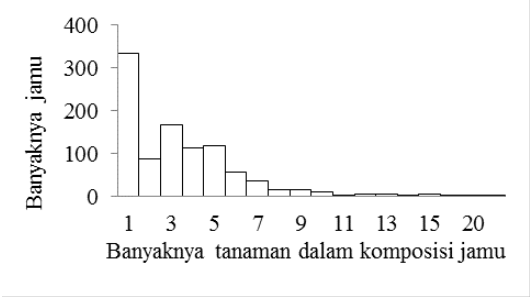
A. Deskripsi Tanaman Komposisi Jamu dan Khasiatnya

Berdasarkan data, terdapat 22 jamu atau sebanyak 2% dari total jamu yang berkhasiat untuk mengatasi gangguan suasana hati dan perilaku, sedangkan sisanya 980 jamu berkhasiat selain mengatasi gangguan tersebut. Setiap jamu memiliki komposisi tanaman yang berbeda baik dari segi jumlah maupun jenis. Mayoritas jamu yang berkhasiat dalam mengatasi gangguan suasana hati menggunakan lima

tanaman yang berbeda, yaitu sebanyak 36% dari total 22 jamu. Semua jumlah tanaman yang digunakan dalam membuat jamu berkhasiat gangguan suasana hati dan perilaku disajikan lebih lengkap pada Gambar 1 sedangkan jamu yang berkhasiat dalam mengatasi gangguan pencernaan sebagian besar menggunakan satu tanaman seperti yang terlihat pada Gambar 2.



Gambar 1. Persentase jamu berdasarkan banyaknya tanaman pada jamu yang berkhasiat dalam mengatasi gangguan suasana hati dan perilaku



Gambar 2. Banyaknya tanaman pada jamu yang berkhasiat dalam mengatasi gangguan pencernaan

Tanaman yang paling banyak digunakan sebagai komposisi jamu adalah *Curcuma xanthorrhiza* yaitu sebanyak 257 jamu atau sekitar 25% dari seluruh jamu pada data. Tanaman lain yang juga banyak digunakan adalah *Foeniculum vulgare* yang dijadikan salah satu komposisi oleh 152 jamu. Sebanyak 147 jamu menggunakan *Curcuma longa* sebagai salah satu komponen penyusunnya. Penyebab tanaman tersebut banyak digunakan dalam pembuatan jamu tidak dibahas dalam penelitian ini.

B. Model Tanpa SMOTE

Regresi logistik dapat digunakan untuk membuat model menurut Hosmer dan Lemeshow (2002). Metode tersebut diterapkan pada data jamu dengan respon biner yaitu khasiat jamu untuk mengatasi gangguan pencernaan (0) dan untuk mengatasi gangguan suasana hati dan perilaku (1). Data dibagi menjadi dua yaitu data pemodelan sebesar 70% (701 amatan) dan data prediksi sebesar 30% (301 amatan). Dugaan parameter yang dihasilkan dengan memaksimumkan fungsi kemungkinan diuji secara bersama menggunakan uji G. Hasil uji G menunjukkan *p-value* bernilai 0.0 yang artinya ada parameter beta yang berpengaruh nyata terhadap model.

Pengujian dilanjutkan dengan uji parsial menggunakan uji Wald yang sebelumnya dilakukan pereduksian peubah menggunakan *forward stepwise*. Hasil yang diperoleh dapat dilihat pada Tabel II yang menunjukkan tidak ada tanaman yang berpengaruh nyata pada model. Hal tersebut terlihat pada *p-value* peubah penjelas yang lebih dari taraf nyata yaitu 5%. Pada penelitian ini terjadi ketidakkonsistenan antara uji bersama dan uji parsial. Uji bersama menunjukkan bahwa terdapat peubah yang signifikan pada taraf nyata 5%. Akan tetapi, setelah dilakukan uji parsial tidak ditemukan tanaman yang berpengaruh pada taraf nyata 5%. Hal ini disebabkan galat baku yang dihasilkan pada dugaan parameter sangat tinggi yang dapat dilihat pada Tabel II. Peubah yang tercantum pada Tabel II merupakan kode peubah penjelas tanaman.

Tabel II
HASIL SIGNIFIKANSI MODEL TANPA SMOTE

Peubah	Koefisien	Galat Baku	*p-value*
P0029	-67.591	2951.993	0.982
P0175	-71.925	40243.654	0.999
P0188	-31.712	1753.011	0.986
P0214	32.591	2550.177	0.990
P0236	67.109	3717.102	0.986
P0255	-34.276	1999.518	0.986
P0256	-67.144	3621.970	0.985
P0325	-34.192	1859.028	0.985
P0340	49.661	5692.464	0.993
P0345	-33.300	3151.496	0.992
P0452	-83.313	3530.229	0.981

Persentase ketepatan model dalam mengklasifikan khasiat jamu dengan benar sebesar 99.9% pada data pemodelan dan 98.3% pada data prediksi dengan batas peluang sebesar 0.5 yang ditunjukkan Tabel III. Akan tetapi, pada data prediksi terdapat ketidakseimbangan dalam memprediksi khasiat mengatasi gangguan suasana hati dan perilaku dengan khasiat mengatasi gangguan pencernaan dengan perbandingan 99.3% dan 57.1%. Ketidakseimbangan tersebut menyebabkan prediksi model lebih mengarah pada khasiat dalam mengatasi gangguan pencernaan sehingga perlu digunakan SMOTE. Kebaikan model yang ditunjukkan oleh nilai R^2 Nagelkerke pada model tanpa SMOTE adalah sebesar 98.3%. Ketepatan prediksi berimplikasi pada akurasi model yang ditunjukkan oleh nilai AUC. Nilai AUC model tanpa SMOTE adalah 0.908.

C. Model dengan SMOTE

Persentase awal jumlah amatan pada kelas minor sebesar 2% ditambahkan data buatan melalui tahap SMOTE

Tabel III
KETEPATAN KLASIFIKASI MODEL TANPA SMOTE

Observasi	Prediksi					
	Data pemodelan			Data prediksi		
	0	1	% ketepatan	0	1	% ketepatan
0	686	0	100.0	292	2	99.3
1	1	14	93.3	3	4	57.1
persentase keseluruhan			99.9			98.3

sehingga persentasenya menjadi sekitar 50% jumlah amatan. Hal tersebut diperoleh dari *oversampling* sebanyak 4200% sehingga jumlah kelas minor menjadi 946 amatan. Jumlah amatan menjadi 1926 setelah melalui tahap SMOTE. Data tersebut kemudian dibagi menjadi data pemodelan dan data prediksi. Data pemodelan yang digunakan merupakan data asli dan data buatan hasil SMOTE sedangkan data prediksi pada model tanpa SMOTE sama dengan data prediksi pada model dengan SMOTE.

Reduksi peubah penjelas dengan *forward stepwise* dilakukan sebelum menduga parameter. Kemudian dugaan parameter dilakukan dengan memaksimumkan fungsi kemungkinan. Pengujian dugaan parameter secara simultan menggunakan uji G diperoleh *p-value* 0.0 yang artinya ada tanaman yang berpengaruh nyata terhadap model. Setelah itu, dilakukan uji parsial (uji Wald) yang menghasilkan peubah-peubah yang berpengaruh nyata terhadap model. Terdapat 17 tanaman yang berpengaruh nyata pada model komposisi jamu yang berkhasiat dalam mengatasi gangguan suasana hati dan perilaku karena memiliki *p-value* kurang dari taraf nyata (0.05). Tanaman tersebut terdapat pada peubah dalam Tabel IV.

Kebaikan model yang ditunjukkan oleh nilai R^2 Nagelkerke pada model yang telah melaui tahap SMOTE adalah 95.1%. Prediksi jamu yang berkhasiat dengan yang tidak berkhasiat dalam mengatasi gangguan suasana hati dan perilaku disajikan dalam Tabel V. Ketepatan prediksi berimplikasi pada akurasi model yang ditunjukkan oleh nilai AUC. Nilai AUC model yang telah melalui tahap SMOTE adalah 0.976.

D. Perbandingan Model

Kedua model yang telah diperoleh dibandingkan tingkat akurasinya dengan nilai AUC atau luas di bawah kurva ROC dan kebaikan model dengan R^2 Nagelkerke (Tabel VI). Nilai R^2 Nagelkerke pada model SMOTE lebih rendah 3.2% dibandingkan pada model tanpa SMOTE. Meski demikian, nilai AUC pada model dengan SMOTE lebih tinggi 0.68 dibandingkan dengan nilai AUC yang dihasilkan model tanpa SMOTE. Hal tersebut menunjukkan model dengan SMOTE lebih akurat dibandingkan dengan model tanpa SMOTE. Nilai AUC disusun oleh spesifitas dan sensitifitas. Sensitifitas atau ketepatan model dalam memprediksi jamu

Tabel IV
HASIL SIGNIFIKANSI MODEL DENGAN SMOTE

Peubah	Koefisien	Galat Baku	*p-value*
P0001	-2.289	0.549	0.000
P0006	-3.110	0.635	0.000
P0029	-5.361	0.922	0.000
P0040	-3.964	1.007	0.000
P0142	-2.463	1.080	0.023
P0144	-2.948	0.487	0.000
P0171	-2.595	1.006	0.010
P0186	-3.623	0.646	0.000
P0233	3.371	1.160	0.004
P0236	5.378	1.021	0.000
P0255	-5.046	1.074	0.000
P0281	4.497	1.080	0.000
P0308	-3.816	1.434	0.008
P0311	-3.896	0.811	0.000
P0325	-2.529	0.834	0.002
P0339	-2.023	0.628	0.001
P0452	-8.478	2.178	0.000

Tabel V
KETEPATAN KLASIFIKASI MODEL DENGAN SMOTE

Observasi	Prediksi					
	Data pemodelan			Data prediksi		
	0	1	% ketepatan	0	1	% ketepatan
0	675	11	98.4	285	9	96.9
1	26	913	97.2	1	6	85.7
persentase keseluruhan			97.7			96.7

berkhasiat mengatasi gangguan suasana hati dan perilaku pada model dengan SMOTE (85.7%) lebih tinggi dibandingkan dengan model tanpa SMOTE (57.1%). Spesifitas atau ketepatan prediksi model dalam mengklasifikan jamu yang berkhasiat untuk mengatasi gangguan pencernaan ada model dengan SMOTE sedikit lebih rendah dibandingkan dengan spesifitas pada model tanpa SMOTE. Besarnya spesifitas pada model dengan SMOTE adalah sebesar 96.9% sedangkan pada model tanpa SMOTE adalah sebesar 99.3%.

Tabel VI
PERBANDINGAN MODEL TANPA SMOTE DAN DENGAN SMOTE

Kriteria	Model	
	tanpa SMOTE	dengan SMOTE
R^2 Nagelkerke	98.3%	95.1%
Sensitifitas (*true positive rate*)	57.1%	85.7%
Spesifisitas (*true negative rate*)	99.3%	96.9%
AUC	0.908	0.976

V. SIMPULAN

Ukuran kebaikan model ditunjukkan oleh nilai R^2 Nagelkerke. Nilai R^2 Nagelkerke yang dihasilkan model dengan SMOTE lebih rendah 3.2% dibandingkan dengan R^2 Nagelkerke yang dihasilkan model tanpa SMOTE. Meskipun

demikian, model dengan SMOTE lebih akurat karena nilai AUC yang dihasilkan lebih tinggi daripada model tanpa SMOTE. Model dengan SMOTE memiliki nilai AUC sebesar 0.976 sedangkan model tanpa SMOTE memiliki nilai AUC sebesar 0.908. Hasil tersebut menunjukkan bahwa SMOTE dapat menaikkan tingkat akurasi model pada data tidak seimbang.

PUSTAKA

[1] V. N. Chawla, K. W. Bowyer, L. O. Hall dan W. P. Kegelmeyer, *Journal of Artificial Intelligence Research* [Internet], SMOTE: Synthetic Minority Over-Sampling Technique. [diunduh pada 2013 Mei 31] 16:321-357.Tersedia pada: http://arxiv.org/pdf/1106.1813.pdf, 2002.

[2] S. Cost dan S. Salzberg, *Machine Learning* [Internet], A weighted Nearest Neighbor Algorithm for Learning with Symbolic Features. [diunduh pada 2013 Juli 17] 10:57-58. Boston (US): Kluwer Academic Publisher. Tersedia pada: http://parati.dca.fee.unicamp.br/media/Attachments/course IA368Q1S2012/Monografia/cost_1993.pdf, 1993.

[3] D. W. Hosmer dan S. Lemeshow, *Applied Logistic Regression Second Edition*, New Jersey (US): John Wiley dan Sons, 2000.

[4] T. Fawcett, *Pattern Recognition Letter* [internet]. An introduction to ROC analysis.[diunduh pada 2013 September 6] 27:861-874. Tersedia pada:https://ccrma.stanford.edu/workshops/mir2009 /references/ROCintro.pdf, 2006.

[5] N. J. D. Nagelkerke, *Maximum Likelihood Estimation of Functional Relationship, Pays-Bas, Lecture Notes in Statistics*, Springer Verlag. Volume 69, 1992.