Penerapan Metode Resampling dalam Mengatasi Imbalanced Data Pada Determinan Kasus Diare Pada Balita di Indonesia

Andriansyah Muqiit WS
Politeknik Statistika STIS, 211709554@stis.ac.id

Intan Putri Ananda
Politeknik Statistika STIS

M. Alfa Rizki
Politeknik Statistika STIS

Zahrotin Dwi Hapsari
Politeknik Statistika STIS

Rani Nooraeni
Politeknik Statistika STIS, raninoor@stis.ac.id

ABSTRAK, Data yang memiliki rasio yang tidak berimbang antara data satu dengan data lainnya dapat dikatakan sebagai imbalanced. Metode Synthetic Minority Oversampling Technique (SMOTE) merupakan metode yang populer diterapkan dalam rangka menangani imbalanced. Pada penelitian ini penulis ingin membandingkan model data sebelum dan data sesudah dilakukan oversampling menggunakan analisis regresi logistik berganda. Data yang digunakan adalah data sekunder dari hasil Survei Demografi dan Kesehatan (SDKI) tahun 2017. Variabel respon (dependent) yang digunakan adalah balita mengalami diare dalam dua minggu terakhir sebelum pencacahan. Variabel penjelas dikelompokkan menjadi dua yaitu faktor sosio-demografi dan faktor lingkungan. Variabel penjelas yang termasuk ke dalam faktor sosio-demografi antara lain jenis kelamin balita, umur balita, umur ibu, dan tingkat pendidikan ibu. Sedangkan variabel penjelas yang merupakan faktor lingkungan antara lain sumber air minum, jenis fasilitas toilet, jenis lantai rumah utama, dan daerah tempat tinggal. Berdasarkan penelitian yang telah dilakukan bahwa penerapan metode SMOTE sangat tepat digunakan untuk meningkatkan keakurasian analisis regresi logistik berganda serta dapat menghindari terjadinya overfitting pada data diare balita di Indonesia tahun 2017 yang memiliki karakteristik imbalanced (rasio tidak berimbang).

***Kata Kunci**: Imbalanced, Oversampling, SMOTE, Diare*

1. PENDAHULUAN

Data yang memiliki rasio yang tidak berimbang antara data satu dengan data lainnya dapat dikatakan sebagai *imbalanced*. *Data mining* mengartikan *imbalanced* dengan jumlah data kelas mayoritas lebih banyak dibandingkan dengan kelas minoritas[1]. Penurunan akurasi pada *imbalanced* disebabkan banyak ditemukan *noise* atau outlier pada dataset uji yang berasal dari kelas minoritas[2]. *Imbalanced* pada data yang digunakan akan menghasilkan model yang tidak cocok. Model yang tidak cocok sehubungan dengan terjadinya *overfitting* dan tidak dapat mengklasifikasikan data dengan baik. *Overfitting* dapat terjadi dikarenakan kurva logistik yang cenderung mengarah pada salah satu kategori. Hal ini juga disebabkan oleh ketidakseimbangan data yang digunakan dalam pemodelan.

Klasifikasi pada data dengan kelas tidak seimbang *(imbalanced)* merupakan masalah utama pada bidang *machine learning* dan *data mining*, misalnya pada masalah kesehatan. Jika bekerja pada data *imbalanced*, hampir semua algoritma klasifikasi akan menghasilkan akurasi yang jauh lebih tinggi untuk kelas mayoritas daripada kelas minoritas. Namun pada beberapa kasus, kelas minoritas justru lebih penting untuk diidentifikasi daripada kelas mayoritas[3]. Misalnya kasus diare pada balita, kebanyakan kejadian yang terjadi adalah balita tidak mengalami diare dalam dua minggu sebelum pencacahan, hanya sedikit kasus yang dapat ditemukan dimana balita mengalami kejadian diare.

Logistic Regression merupakan klasifikasi linier yang telah terbukti menghasilkan klasifikasi yang powerful dengan statistik probabilitas dan menangani masalah klasifikasi *multiclass*[4]. Metode tersebut cocok digunakan karena respon yang diamati berskala

kategorik. Namun, *Logistic Regression* masih memiliki kelemahan yaitu rentan terhadap *underfitting/overfitting* dan memiliki akurasi yang cenderung rendah[5]. Salah satu hal yang perlu diperhatikan dalam evaluasi model adalah tingkat akurasi sebuah model dalam memprediksi respon dengan benar [6]. Pemodelan dengan *Logistic Regression* menghasilkan model yang tepat dan tingkat akurasi yang lebih baik apabila *imbalanced* telah diatasi.

Imbalanced dapat diselesaikan dengan metode *resampling*, yang merupakan cara paling populer untuk mengatasi masalah *imbalanced*. *Resampling* sebagai sarana mengubah distribusi kelas minoritas sehingga tidak kurang terwakili ketika *training* data pada algoritma *machine learning*. Metode *resampling* terbagi menjadi dua yaitu metode *oversampling* dan *undersampling*. Metode *oversampling* adalah metode yang paling sederhana untuk menangani kelas minoritas dengan melakukan random kelas selama proses pengambilan sampel. Proses pengambilan sampel dengan teknik *oversampling* ini adalah dengan menduplikasi kelas positif dan dilakukan penyeimbangkan kelas secara acak[7]. Namun, karena metode ini menduplikasi kelas positif yang ada dikelas minoritas, kemungkinan terjadi *overfitting* pasti akan terjadi.

Adapun metode *undersampling*, metode ini hampir sama dengan metode *oversampling* yaitu dengan menghitung selisih kelas mayoritas dan kelas minoritas. Selanjutnya dilakukan perulangan sebanyak selisih kelas mayoritas dengan kelas minoritas. Selama proses perulangan dilakukan penghapusan terhadap kelas mayoritas sehingga didapatkan jumlah yang sama dengan kelas minoritas. Metode *Synthetic Minority Oversampling Technique* (SMOTE) merupakan metode yang populer diterapkan dalam rangka menangani *imbalanced*. Metode SMOTE ini merupakan pengembangan dari metode *oversampling*, dimana teknik ini mensintesis sampel baru dari kelas minoritas untuk menyeimbangkan dataset dengan cara sampling ulang sampel kelas minoritas[3].

Kasus kesehatan yang dibahas pada penelitian ini yaitu tentang kasus diare pada balita di Indonesia tahun 2017 (SDKI 2017). *World Health Organization* (WHO) tahun 2017 mendefiniskan diare adalah kondisi di mana individu mengalami buang air besar dengan frekuensi sebanyak tiga kali atau lebih dengan konsistensi tinja cair, dapat disertai dengan darah dan atau lendir[8]. Biasanya merupakan gejala infeksi pada saluran usus yang dapat disebabkan oleh berbagai organisme bakteri, virus, dan parasit yang menyebar melalui makanan atau air minum yang terkontaminasi atau dari orang ke orang sebagai akibat dari kebersihan yang buruk.

Diare merupakan penyakit endemis di Indonesia dan juga merupakan penyakit potensial Kejadian Luar Biasa (KLB) yang sering disertai dengan kematian. Diare membawa kematian lebih cepat pada anak-anak dibanding orang dewasa karena terjadinya dehidrasi dan malnutrisi. Angka kematian (CFR) di Indonesia saat KLB Diare masih berada diatas 1 persen kecuali pada tahun 2011 sebesar 0,40 persen. Sedangkan tahun 2018 CFR Diare saat KLB mengalami peningkatan di banding tahun 2017 yaitu menjadi 4,76 persen.

Dari data kasus diare pada balita di SDKI 2017, dimana dari sampel terpilih sebanyak 49.250 rumah tangga, tercatat ada sekitar 11.134 rumah tangga yang memiliki balita, dan hanya ada sekitar 1.560 balita yang mengalami diare dalam dua minggu terakhir pada saat pencacahan. Hal ini menunjukkan bahwa kasus kejadian diare pada balita hanya berkisar 14 persen. Sedangkan sisanya 86 persen balita tidak mengalami diare. Kejadian diare pada balita di Indonesia tahun 2017 ini menggambarkan keadaan *imbalanced*, dimana kasus balita diare dalam dua minggu terakhir masuk ke dalam kelas minoritas. Sehingga untuk menemukan model yang cocok, keadaan *imbalanced* harus diatasi.

Penelitian ini bertujuan untuk mengatasi masalah *imbalanced* dari data diare pada balita di Indonesia tahun 2017 menggunakan metode *oversampling* dengan metode SMOTE. Kemudian peneliti akan membandingkan model data sebelum dan data sesudah dilakukan *oversampling* menggunakan analisis regresi logistik berganda untuk melihat model mana yang cocok digunakan dalam mendeterminasikan kasus diare balita di Indonesia tahun 2017.

2. TINJAUAN TEORI

Analisis Regresi Logistik Berganda

Teknik analisis inferensia yang digunakan dalam penelitian ini adalah model regresi logistik berganda atau *multiple logistic regression* dengan variabel tak bebas biner di mana variabel bebasnya kontinu dan kategorik. Regresi logistik berganda (*multiple logistic regression*) adalah suatu analisis regresi yang digunakan untuk menggambarkan hubungan antara variabel bebas dengan sekumpulan variabel terikat, dimana variabel terikat bersifat nominal lebih dari 2 kategorik[9].

Bentuk umum model peluang regersi logistik berganda (*multiple logistic* regression) dengan p variabel penjelas (in*dependent)*, diformulasikan sebagai berikut:

$$\pi(x) = \frac{\exp(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip})}{1+\exp(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip})} \quad (2.1)$$

Nilai π(x) adalah peluang kejadian sukses dengan nilai probabilita $0 \leq \pi(x) \leq 1$ dan β_k adalah nilai parameter dengan k=1, 2, …, p. π(x) merupakan fungsi yang non linier, sehingga perlu dilakukan transformasi ke bentuk logit untuk memperoleh fungsi yang linier agar dapat dilihat hubungan antara variabel respon (*dependent*) dan variabel penjelas (*independent*). Dengan melakukan transformasi dari logit π(x), maka didapat persamaan:

$$g(x) = ln\frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \quad (2.2)$$

Jika dari beberapa variabel penjelas (*dependent*) ada yang berskala nominal atau ordinal maka diperlukan variabel *dummy*. Untuk variabel penjelas (*dependent*) dengan skala ordinal atau nominal dengan k kategori diperlukan k-1 variabel *dummy*.

Metode Synthetic Minority Oversampling Technique (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) pertama kali diperkenalkan oleh Nithes V. Chawla sebagai salah satu solusi dalam menangani data tidak seimbang dengan prinsip yang berbeda dengan metode oversampling yang telah diusulkan sebelumnya[10]. 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[6].

Metode ini bekerja dengan mengelompokkan data terdekat yang dipilih berdasarkan jarak Euclidean antara kedua data. Penentuan jumlah replikasi yang dilakukan disesuaikan dengan jumlah anggota pada kelas mayor. Jumlah replikasi harus sesuai dengan jumlah k pada *nearest neighbour*, jika jumlah replikasi sebanyak n maka jumlah k harus sebanyak n-1.

Misalkan terdapat dua struktur data dengan p dimensi yaitu $x^t = [x_1, x_2, \ldots x_p]$ dan $y^t = [y_1, y_2, \ldots y_p]$, maka jarak Euclidean d(x,y) yang dihasilkan antara kedua data ditunjukkan pada persamaan sebagai berikut :

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2} \quad (2.3)$$

"*Synthetic*" atau replikasi data dilakukan dengan menggunakan persamaan sebagai berikut:

$$x_{syn} = x_i + (x_{knn} - x_i) \times \tau \quad (2.4)$$

Dengan:

x_{syn} = data hasil replikasi
x_i = data yang akan direplikasi
x_{knn} = data yang memiliki jarak terdekat dari data yang akan direplikasi
τ = bilangan random 0 sampai 1

3. METODOLOGI

Data yang digunakan adalah data sekunder dari hasil Survei Demografi dan Kesehatan (SDKI) tahun 2017 yang dilaksanakan oleh BPS yang berkolaborasi dengan Badan Kependudukan dan Keluarga Berencana Nasional (BKKBN) dan Kementerian Kesehatan. Pada penelitian ini ada dua jenis

variabel yang digunakan, yaitu variabel penjelas (*independent*) dan variabel respon (*dependent*). Variabel respon (*dependent*) yang digunakan adalah balita mengalami diare dalam dua minggu terakhir sebelum pencacahan. Variabel penjelas dikelompokkan menjadi dua yaitu faktor sosio-demografi dan faktor lingkungan. Variabel penjelas yang termasuk ke dalam faktor sosio-demografi antara lain jenis kelamin balita, umur balita, umur ibu, dan tingkat pendidikan ibu. Sedangkan variabel penjelas yang merupakan faktor lingkungan antara lain sumber air minum, jenis fasilitas toilet, jenis lantai rumah utama, dan daerah tempat tinggal.

Definisi Operasional Variabel

1. Kejadian diare pada balita terhitung dalam dua minggu sebelum pencacahan. Data yang dihasilkan adalah apakah balita mengalami diare pada sekurang-kurangnya dua minggu yang lalu atau tidak.
2. Umur balita adalah umur balita yang dihitung sejak lahir sampai waktu pencacahan yang dinyatakan dalam tahun.
3. Umur ibu adalah umur ibu yang dihitung sejak lahir sampai waktu pencacahan yang dinyatakan dalam tahun.
4. Jenis kelamin balita adalah perbedaan biologis dan fisiologis yang dapat membedakan laki-laki dan perempuan diukur dengan variabel dummy dan digolongkan dalam dua kategori, yakni diberi skor 0 apabila balita berjenis kelamin perempuan dan diberi skor 1 apabila balita berjenis kelamin laki-laki.
5. Daerah tempat tinggal menunjukkan wilayah dimana balita sehari-hari berdomisili, digolongkan dengan variabel dummy dan diberi skor 0 apabila balita berdomisili di pedesaan dan skor 1 apabila balita berdomisili di perkotaan.
6. Tingkat pendidikan tertinggi menunjukkan pendidikan terakhir yang ditamatkan oleh ibu yang digolongkan menjadi 4 kategori, yaitu tidak bersekolah, SD/sederajat, SMP/sederajat, dan SMA/sederajat.
7. Jenis lantai rumah utama menunjukkan jenis lantai rumah tinggal yang digolongkan menjadi 3 kategori, yaitu lantai alami, lantai bahan, dan lantai jadi.
8. Jenis fasilitas toilet menunjukkan jenis jamban yang digunakan pada rumah tinggal balita yang digolongkan menjadi 4 kategori, yaitu jamban layak, jamban bersama, jamban tidak layak, dan tidak memiliki jamban.
9. Sumber air minum utama menunjukkan kelayakan sumber air minum utama balita yang diukur dengan variabel dummy dan digolongkan dalam 2 kategori, yaitu diberi skor 0 apabila sumber air minum utama tidak layak dan diberi skor 1 apabila sumber air minum utama layak.

4. PEMBAHASAN

Gambaran Umum *Oversampling*

Perbandingan antara balita yang mengalami diare dua minggu yang lalu dan yang tidak, berdasarkan SDKI 2017 memiliki perbedaan yang cukup besar yakni dari 11.134 balita hanya 1.560 saja yang mengalami diare. Dapat dilihat dari gambar 1 bahwa persentase balita yang mengalami diare dua minggu terakhir dengan yang tidak mengalami sangat berbeda. Persentase balita yang mengalami diare dua minggu terakhir sebesar 14 persen sedangkan yang tidak mengalami diare sebesar 86 persen.

Gambar 1. Persentase kejadian diare pada balita di Indonesia tahun 2017

Data seperti ini apabila dilakukan pemodelan akan berakibat fatal dan hasil estimasinya tidak representatif. Oleh karena itu, dengan dilakukannya *oversampling* dengan metode SMOTE data akan lebih proporsional.

Tabel 1. Perbandingan Jumlah Kejadian Diare Balita Pada Data Sebelum dan Sesudah *Oversampling.*

Kejadian Diare Balita	Data	
	Sebelum	Hasil *Oversampling*
Ya, 2 minggu yang lalu	1560	4680
Tidak	9574	6240
Jumlah	11134	10920

Sumber: SDKI 2017 (data diolah)

Dari Tabel 1 terlihat data sebelum dan sesudah dilakukan *oversampling* perbedaannya sangat jauh. Data sebelum dilakukan *oversampling* tidak proporsional, dimana balita yang tidak mengalami diare jumlah kejadiannya lebih banyak daripada yang mengalami diare. Tetapi setelah dilakukan *oversampling* perbandingan data kejadian diare pada balita menjadi lebih proporsional.

Pemodelan Data Dengan Regresi Logistik Berganda

Pemodelan dilakukan menggunakan analisis regresi logistik berganda dengan data sebelum dan sesudah dilakukan *oversampling*. Sehingga diperoleh hasil pengujian secara simultan dengan menggunakan statistic uji G sebagai berikut.

Tabel 2. *Omnibus Test of Model Coefficients*

Model	*Chi-square*	*Df*	*p-value*
Tanpa *oversampling*	115.680	13	0.000*
Dengan *oversampling*	1879.906	13	0.000*

Sumber: SDKI 2017 (data diolah)

Keterangan: *) = signifikan pada taraf signifikansi 5 persen

Berdasarkan hasil uji simultan pada Tabel 2 diperoleh hasil *p-value* untuk model tanpa *oversampling* sebesar 0.000. Dikarenakan *p-value* tersebut nilainya lebih kecil dari 5 persen, maka dengan tingkat signifikansi 5 persen dapat disimpulkan bahwa terdapat cukup bukti untuk menyatakan bahwa minimal terdapat satu variabel bebas yang berpengaruh signifikan secara statistik terhadap kejadian diare pada balita di Indonesia tahun 2017.

Pada model dengan *oversampling* diperoleh hasil *p-value* sebesar 0.000. Dikarenakan *p-value* tersebut nilainya lebih kecil dari 5 persen, maka dengan tingkat signifikansi 5 persen dapat disimpulkan bahwa terdapat cukup bukti untuk menyatakan bahwa minimal terdapat satu variabel bebas yang berpengaruh signifikan secara statistik terhadap kejadian diare pada balita di Indonesia tahun 2017 pada data *oversampling*. Selanjutnya adalah melakukan estimasi dengan regresi logistik berganda. Diperoleh hasil estimasi pada Tabel 3 sebagai berikut.

Tabel 3. Estimasi Parameter Tanpa dan Dengan *Oversampling* (SPSS)

Nama Variabel	Variabel	Kategori	Estimasi β	
			Tanpa oversampling	Dengan oversampling
Umur balita	X_1		-0.006	-0.029
Umur ibu	X_2		-0.032	-0.037
Jenis kelamin balita	D_1	Laki-laki Perempuan*	-0.135	-0.080
Daerah tempat tinggal	D_2	Perkotaan Pedesaan*	-0.125	-0.162
Tingkat pendidikan tertinggi ibu	D_{31}	Tidak bersekolah SMA/sederajat*	0.484	0.883
	D_{32}	SD/sederajat	0.372	0.411

Nama Variabel	Variabel	Kategori	Estimasi β	
			Tanpa oversampling	Dengan oversampling
		SMA/sederajat*		
	D_{33}	SMP/sederajat SMA/sederajat*	0.223	0.129
Jenis lantai rumah utama	D_{41}	Lantai alami Lantai jadi*	-0.318	0.331
	D_{42}	Lantai bahan Lantai jadi*	0.077	0.881
Jenis fasilitas toilet	D_{51}	Jamban layak Tidak ada*	-0.180	-0.848
	D_{52}	Jamban bersama Tidak ada*	-0.015	-0.170
	D_{53}	Jamban tidak layak Tidak ada*	-0.338	-0.810
Sumber air minum utama	D_6	Layak Tidak layak*	-0.113	1.190
Konstan			-0.897	1.009

Sumber: SDKI 2017 (data diolah)
Keterangan: *) = kategori referensi

Tabel 4. Hasil Uji Parsial Tanpa dan Dengan *Oversampling* (SPSS)

Nama Variabel	Variabel	Kategori	Sig.	
			Tanpa oversampling	Dengan oversampling
Umur balita	X_1		0.738	0.048**
Umur ibu	X_2		0.000**	0.000**
Jenis kelamin balita	D_1	Laki-laki Perempuan*	0.014**	0.060
Daerah tempat tinggal	D_2	Perkotaan Pedesaan*	0.039**	0.000**
Tingkat pendidikan tertinggi ibu	D_{31}	Tidak bersekolah SMA/sederajat*	0.020**	0.000**
	D_{32}	SD/sederajat SMA/sederajat*	0.000**	0.000**
	D_{33}	SMP/sederajat SMA/sederajat*	0.010**	0.053
Jenis lantai rumah utama	D_{41}	Lantai alami Lantai jadi*	0.034**	0.000**
	D_{42}	Lantai bahan Lantai jadi*	0.324	0.000**
Jenis fasilitas toilet	D_{51}	Jamban layak Tidak ada*	0.045**	0.000**

Nama Variabel	Variabel	Kategori	Sig.	
			Tanpa oversampling	Dengan oversampling
	D_{52}	Jamban bersama Tidak ada*	0.892	0.030**
	D_{53}	Jamban tidak layak Tidak ada*	0.036**	0.000**
Sumber air minum utama	D_6	Layak Tidak layak*	0.191	0.000**
Konstan			0.014**	0.000**

Sumber: SDKI 2017 (data diolah)
Keterangan: *) = kategori referensi **) = signifikan pada taraf signifikansi 5 persen

Berdasarkan hasil pengujian parsial pada Tabel 4, hasil dari model yang tanpa *oversampling* diperoleh 9 variabel yang berpengaruh terhadap kejadian diare pada balita di Indonesia yakni dilihat dari *p-value* yang nilainya lebih kecil dari tingkat signifikansi 5 persen. Oleh karena itu, dengan tingkat signifikansi 5 persen dapat disimpulkan bahwa terdapat cukup bukti untuk menyatakan bahwa umur ibu dari balita, jenis kelamin balita, daerah tempat tinggal, tingkat pendidikan tertinggi ibu kategori tidak bersekolah, SD/sederajat, dan SMP/sederajat, jenis lantai rumah kategori lantai alami, jenis fasilitas toilet kategori jamban layak dan jamban tidak layak.

Sedangkan untuk hasil dari model dengan *oversampling* diperoleh 11 variabel yang berpengaruh terhadap kejadian diare pada balita di Indonesia yakni dilihat dari *p-value* yang nilainya lebih kecil dari tingkat signifikansi 5 persen. Oleh karena itu, dengan tingkat signifikansi 5 persen dapat disimpulkan bahwa terdapat cukup bukti untuk menyatakan bahwa umur balita, umur ibu dari balita, daerah tempat tinggal, tingkat pendidikan tertinggi ibu kategori tidak bersekolah dan SD/sederajat, jenis lantai rumah kategori lantai alami dan lantai bahan, jenis fasilitas toilet kategori jamban layak, jamban bersama, dan jamban tidak layak, dan sumber air minum utama.

Tabel 5. Prediksi dan Klasifikasi (tanpa *oversampling*)

		Prediksi		
		Tidak	Ya, 2 minggu terakhir	Persentase
Observasi	Tidak	9574	0	100
	Ya, 2 minggu terakhir	1560	0	0
Persentase				86.0

The cut value is 0.5
Sumber: SDKI 2017 (data diolah).

Tabel 6. Prediksi dan Klasifikasi (*oversampling*)

		Prediksi		
		Tidak	Ya, 2 minggu terakhir	Persentase
Observasi	Tidak	5110	1130	81.9
	Ya, 2 minggu terakhir	2231	2449	52.3
Persentase				69.2

The cut value is 0.5

Sumber: SDKI 2017 (data diolah)

Berdasarkan Tabel 5, dapat diketahui bahwa model regresi logistik berganda pada data sebelum dilakukan over*sampling* terjadi *overfitting*. Salah satu indikasi terjadinya *overfitting* adalah dari hasil tabel prediksi dan klasifikasinya ternyata model regresi logistik tidak mampu memprediksi kejadian diare pada balita berdasarkan variabel independen yang ada.

Sedangkan pada Tabel 6, dapat diketahui bahwa dengan menggunakan data hasil *oversampling*, model regresi logistik berganda mampu memprediksi kejadian diare pada balita berdasarkan variabel independen yang ada. Persentase hasil prediksinya adalah sebesar 69.2 persen. Dikarenakan *oversampling* dapat mengatasi terjadinya *overfitting* pada data yang tidak berimbang (*imbalanced data*), maka model regresi logistik berganda dengan *oversampling* lebih cocok untuk memodelkan kejadian diare pada balita.

Interpretasi Model Terbaik

Berdasarkan hasil estimasi pada Tabel 3, maka dengan menggunakan persamaan (2) model regresi logistik berganda untuk kejadian diare pada balita adalah sebagai berikut:

$$\hat{g}(x) = 1.009 - 0.029X_1 - 0.037X_2 - 0.080D_1 - 0.162D_2 + 0.883D_{31} + 0.411D_{32} + 0.129D_{33} + 0.331D_{41} + 0.881D_{42} - 0.848D_{51} - 0.170D_{52} - 0.810D_{53} + 1.190D_6$$

Dalam menginterpretasikan hasil estimasi model regresi logistik berganda adalah dengan melihat rasio kecenderungan dari masing-masing variabel. Nilai rasio kecenderungan untuk setiap variabel adalah sebesar $\exp(\hat{\beta})$ berdasarkan nilai dari masing-masing koefisien $\hat{\beta}$. Sehingga interpretasinya adalah sebagai berikut:

a. Umur balita
   Setiap bertambahnya umur balita sebesar satu tahun, maka kecenderungan seorang balita untuk mengalami diare adalah sebesar $\exp(-0.029) = 0.971$ kali lebih kecil.
b. Umur ibu
   Setiap bertambahnya umur ibu sebesar satu tahun, maka kecenderungan seorang balita untuk mengalami diare adalah sebesar $\exp(-0.037) = 0.964$ kali lebih kecil.
c. Jenis kelamin balita
   Balita berjenis kelamin laki-laki memiliki kecenderungan untuk mengalami diare sebesar $\exp(-0.080) = 0.923$ kali lebih kecil dibandingkan dengan balita berjenis kelamin perempuan.
d. Daerah tempat tinggal
   Balita yang bertempat tinggal di daerah perkotaan memiliki kecenderungan untuk mengalami diare sebesar $\exp(-0.162) = 0.85$ kali lebih kecil dibandingkan dengan yang bertempat tinggal di daerah perdesaan.
e. Tingkat pendidikan tertinggi ibu
   Balita yang ibunya tidak bersekolah, berpendidikan SD/sederajat, dan berpendidikan SMP/sederajat memiliki kecenderungan untuk mengalami diare masing-masing sebesar 2.418, 1.554, dan 1.138 kali lebih besar dibandingkan dengan yang berpendidikan SMA/sederajat.
f. Jenis lantai rumah utama
   Balita yang jenis lantai rumah utamanya berupa lantai alami dan lantai bahan memiliki kecenderungan untuk mengalami diare sebesar 1.392 dan 2.413 kali lebih besar dibandingkan dengan yang lantai jadi.

g. Jenis fasilitas toilet
   Balita yang jenis fasilitas toiletnya layak, bersama, dan tidak layak memiliki kecenderungan untuk mengalami diare masing-masing sebesar 0.428, 0.844, dan 0.445 kali lebih kecil dibandingkan dengan yang tidak memiliki fasilitas.
h. Sumber air minum utama
   Balita yang sumber air minum utamanya layak memiliki kecenderungan untuk mengalami diare sebesar 3.288 lebih besar dibandingkan dengan balita yang sumber air minumnya tidak layak. Terjadi *misleading* pada variabel sumber air minum utama. Seharusnya kecenderungan lebih besar untuk sumber air minum tidak layak, akan tetapi berdasarkan data yang ada ketika dimodelkan memberikan hasil yang sebaliknya. Sehingga dalam hal ini perlu peninjauan ulang terkait kriteria sumber air minum yang layak.

5. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, model analisis regresi logistik berganda dengan pendekatan SMOTE (*oversampling*) memiliki tingkat akurasi pengklasifikasian yang cukup baik untuk kelas minoritas pada kejadian diare pada balita di Indonesia tahun 2017 dan tidak terjadi *overfitting*. Selain itu, pada uji parsial (uji Wald) model dengan data *oversampling* menghasilkan lebih banyak variabel yang signifikan dibandingkan dengan model dengan data tanpa *oversampling*. Dengan demikian dapat disimpulkan bahwa penerapan metode SMOTE sangat tepat digunakan untuk meningkatkan keakurasian analisis regresi logistik berganda serta dapat menghindari terjadinya *overfitting* pada data diare balita di Indonesia tahun 2017 yang memiliki karakteristik *imbalanced* (rasio tidak berimbang).

6. DAFTAR PUSTAKA

[1] Untoro, Meida Cahyo dan Joko Lianto Buliali. (2018). *Penanganan imbalance class data laboratorium kesehatan dengan Majority Weighted Minority Oversampling Technique*. Jurnal Ilmiah Teknologi Sistem Informasi 4 (1), 2018, 23-29.

[2] Almeida, J., Barbosa, L., Pais, A., & Formosinho, S. (2007). *Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering.* Chemometrics and Intelligent Laboratory Systems, 2007(2007), 208-217

[3] Siringoringo, Rimbun. (2018). *Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE Dan K-Nearest Neighbor.* Jurnal ISD Vol.3 No.1 Januari - Juni 2018.

[4] Karsmakers, P., Pelckmans, K., & Suykens, J. a. K. (2007). Multiclass kernel logistic regression: a fixed-size implementation. 2007 International Joint Conference on Neural Networks.

[5] Harrington, P. (2012). Machine Learning in Action. Manning Publications Co.

[6] Barro, Rossi Azmatul, dkk. (2013). *Penerapan Synthetic Minority Oversampling Technique (SMOTE) terhadap Data Tidak Seimbang pada Pembuatan Model Komposisi Jamu.* Jurnal Xplore Vol.1, 2013, 1-6.

[7] Ganganwar, Vaishali. (2012). *An Overview of Classification Algorithms for Imbalanced Datasets.* International Journal of Emerging Technology and Advanced Engineering Vol.2 Issue 4, April 2012, 42-47.

[8] WHO. 2017. *Diarrhoeal Disease*. https://www.who.int/en/news-room/fact-sheets/detail/diarrhoeal-disease. Diakses pada 15 Februari 2020.

[9] Astuti, Anna Puji. 2019. *Variabel-variabel yang Mempengaruhi Kejadian Persalinan Lama di Indonesia (Analisis Data SDKI 2017).* Jakarta: Politeknik Statistika STIS.

[10] Chawla, Nitesh V., dkk. 2018. *SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary.* Journal of Artificial Intelligence Research 61, 2018, 863-905.