Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 20, No. 2, Mei 2021, pp. 417~426
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: 10.30812/matrik.v20i2.1174

Penerapan Metode *Machine Learning* dalam Klasifikasi Risiko Kejadian Berat Badan Lahir Rendah di Indonesia

Implementation of Machine Learning Method in Risk Classification on Low Birth weight in Indonesia

Istiqomatul Fajriyah Yuliati[1], Pardomuan Robinson Sihombing[2]

[1]Badan Kependudukan dan Keluarga Berencana Nasional (BKKBN) Jakarta
[2]Badan Pusat Statistik (BPS) Jakarta

Article Info

Article history:

Received, 22 April 2021
Revised, 18 Mei 2021
Accepted, 24 Mei 2021

Kata Kunci:

CART
machine learning
naïve bayes
SVM
random forest

ABSTRAK

Penurunan Angka Kematian Neonatal (AKN) menjadi salah satu target dalam mencapai tujuan Sustainable Development Goals (SDGs). Penyebab utama kejadian kematian neonatal adalah kondisi Berat Badan Lahir Rendah (BBLR). Oleh karena itu, diperlukan pemodelan klasifikasi untuk penentuan risiko kejadian bayi dengan BBLR yang diharapkan dapat menjadi solusi dalam menurunkan kelahiran bayi dengan BBLR di Indonesia. Data BBLR yang digunakan bersumber dari hasil Survei Demografi dan Kesehatan Indonesia (SDKI) 2017. Penelitian ini akan mengkaji penerapan beberapa metode machine learning dengan memperhatikan kasus *imbalanced* data dalam pemodelan klasifikasi. Adapun metode *meachine learning* yang digunakan adalah *Classification and Regression Tree* (CART), Naïve Bayes, *Random Forest* dan *Support Vector Machine* (SVM). Pemodelan klasifikasi dengan menggunakan teknik resample pada kasus imbalanced data dan set data besar terbukti mampu meningkatkan ketepatan klasifikasi khususnya terhadap kelas minoritas yang dapat diihat dari nilai sensitivity yang tinggi dibandingkan data asli (*tanpa treatment*). Selanjutnya, dari kelima model klasifikasi yang iuji menunjukkan bahwa model random forest memberikan kinerja terbaik berdasarkan nilai sensitivity, specificity, G-mean dan AUC tertinggi. Variabel terpenting/paling berpengaruh dalam klasifikasi resiko kejadian BBLR adalah jarak dan urutan kelahiran, pemeriksaan kehamilan, dan umur ibu.

Keywords:

CART
machine learning
naïve bayes
SVM
random forest

ABSTRACT

Reducing the Infant Mortality Rate (IMR) is one of the targets for achieving the Sustainable Development Goals (SDGs). The main cause of neonatal death is Low Birth Weight (LBW). Therefore, classification modeling is needed to determine the risk of LBW events which are expected to be a solution in reducing the birth rate of LBW babies in Indonesia. The LBW data used comes from the results of the 2017 Indonesian Demographic and Health Survey (IDHS). This study will examine the application of several machine learning methods by considering cases of data imbalance in classification modeling. This study will examine the application of several machine learning methods by considering cases of data imbalance in classification modeling to determine the risk of LBW babies which are expected to be a solution in reducing the birth of LBW babies in Indonesia. The meachine learning methods used are Classification and Regression Tree (CART), Naïve Bayes, Random Forest and Support Vector Machine (SVM). Classification modeling using resample technique in cases of unbalanced data and large data sets is proven to improve classification accuracy, especially for minority classes, seen from high sensitivity values compared to the original data (without treatment). Furthermore, from the five classification models tested, it was seen that the random forest model gave the best performance based on the highest sensitivity, specificity, G-mean, and AUC values. The most important / most influential variables in the risk classification of LBW incidence are the distance and order of births, antenatal care, and the mother's age.

Penulis Korespondensi:

Pardomuan Robinson Sihombing,
Badan Pusat Statistik Jakarta,
Email: robinson@bps.go.id

1. PENDAHULUAN

Indonesia bersama dengan negara-negara lain berkomitmen dalam mencapai tujuan Sustainable Development Goals (SDGs) dengan salah satu targetnya adalah pada tahun 2030 dapat menurunkan Angka Kematian Neonatal (AKN) setidaknya hinggal 12 per 1.000 kelahiran hidup. AKN adalah jumlah anak yang dilahirkan pada waktu tertentu dan meninggal dalam periode 28 hari pertama kehidupan dan dinyatakan sebagai angka per 1.000 kelahiran hidup [1]. Berdasarkan hasil Survei Demografi dan Kesehatan Indonesia (SDKI) 2017 menunjukkan bahwa dalam periode 5 tahun terakhir, AKN adalah 15 per 1.000 kelahiran hidup, artinya 1 dari 67 anak meninggal dalam bulan pertama kehidupannya[2]. Penyebab utama kejadian kematian neonatal adalah kondisi berat badan lahir rendah [3]. World Health Organization (WHO) mendefinisikan berat badan lahir rendah (BBLR) adalah bayi yang lahir dengan berat lahir kurang dari 2500 gram. SDKI 2017 menyatakan bahwa di antara kelahiran hidup dalam kurun waktu 5 tahun terakhir, 94 persen melaporkan berat lahir, 7 persen memiliki berat lahir rendah.

Klasifikasi adalah proses untuk menemukan model atau fungsi yang dapat menggambarkan dan membedakan kelas data atau konsep, dengan tujuan agar model tersebut dapat digunakan untuk memprediksi kelas yang belum diketahui dari suatu objek pengamatan [4]. Metode klasifikasi yang umum digunakan pada disiplin ilmu statistika adalah Analisis Diskriminan dan Regresi Logistik. Namun, semakin populernya era data yang menunjukkan bahwa terjadinya pertumbuhan pesat dari volume data yang luar biasa banyak sehingga menghasilkan set data besar (big data), maka sangat dibutuhkan alat analisis yang kuat dan sebrbaguna untuk mengungkap informasi berharga dari set data besar dan untuk mengubah set data besar tersebut menjadi pengetahuan yang terorganisir[4]. Disisi lain, pesatnya perkembangan teknologi kecerdasan buatan maka berkembanglah metode machine learning¸yaitu mesin yang dikembangkan untuk bisa belajar dengan sendirinya tanpa arahan dari penggunanya dimana dalam pengembangannya berdasarkan disiplin ilmu lain seperti statistika, matematika dan data mining. Metode klasifikasi pada machine learning yang sering digunakan pada metode machine learning adalah Classification and Regression Trees (CART), Random Forest, Naïve Bayes, Support Vector Machines (SVM), dan lain-lain.

Salah satu permasalahan dalam klasifikasi data adalah banyaknya data yang tidak seimbang antara kelas yang berbeda, dan ketika kondisi ketidakseimbangan ekstrim, masalah ini disebut rare event atau imbalanced data [5]. Meskipun banyak metode klasifikasi yang baik digunakan untuk klasifikasi data, namun sayangnya belum tentu metode tersebut tepat digunakan pada kondisi imbalanced data, karena metode tersebut didasarkan pada asumsi bahwa banyaknya data terdistribusi secara merata antara kelas yang berbeda. Referensi [6] menyatakan bahwa ketika metode klasifikasi digunakan pada kasus imbalanced data, maka pengklasifikasian cenderung menihilkan peluang dari kelas minoritas karena nilai prediksi akan cenderung pada kategori yang mayoritas, sehingga tingkat ketepatan klasifikasi yang dihasilkan menjadi kurang baik. Hal ini terjadi terutama untuk set data besar. Referensi [7] melakukan perbandingan beberapa metode resample untuk mengatasi kasus imbalanced data yang menunjukkan hasil bahwa metode both/combine sampling menghasilkan kinerja klasifikasi yang terbaik.

Sejumlah penelitian bidang kesehatan khususnya risiko bayi dengan BBLR telah banyak diteliti dengan diantaranya banyak menggunakan metode regresi logistik, sedangkan metode machine learning masih sangat jarang digunakan. Adapun penelitian terdahulu menggunakan metode machine learning ditampilkan pada Tabel 1.

Tabel 1. Penerapan Metode Machine Learning pada Penelitian Terdahulu

Tahun	Peneliti	Topik	Ketepatan Klasifikasi
2015	S.H. Sumartini dan S.W. Purnami [8]	Penggunaan Metode *Classification and Regression Trees* (CART) untuk Klasifikasi Rekurensi Pasien Kanker Serviks di RSUD Dr. Soetomo Surabaya	CART menghasilkan ketepatan klasifikasi sebesar 69,14 persen
2017	Mambang dan A. Byna [9]	Analisis Perbandingan Algoritma C.45, Random Forest dengan CHAID Decision Tree untuk Klasifikasi Tingkat Kecemasan Ibu Hamil	Random Forest menghasilkan ketepatan klasifikasi sebesar 64 persen
2017	C. Oganis, S. Musdalifah, dan D. Lusiyanti [10]	Klasifikasi Status Gizi Ibu Hamil untuk mengidentifikasi Bayi Berat Lahir Rendah menggunakan Metode Support Vector Machine (SVM) (Studi Kasus di Puskesmas Labuan)	Dengan sampel hanya 125 data, SVM menghasilkan ketepatan klasifikasi sebesar 92 persen
2020	P.L. Kumalasari [11]	Sistem Pengmabilan Keputusan untuk menentukan Proses Persalinan dengan Metode Naïve Bayes dan Forward Chaining	Naïve Bayes menghasilkan ketepatan klasifikasi sebesar 90,1 persen

Evaluasi risiko kejadian melahirkan bayi dengan BBLR menjadi salah satu persoalan yang menarik untuk dibahas. Oleh karena itu, dalam penelitian ini akan mengkaji penerapan beberapa metode machine learning dengan memperhatikan kasus imbalanced data dalam pemodelan klasifikasi untuk penentuan risiko kejadian bayi dengan BBLR yang diharapkan dapat menjadi solusi dalam menurunkan kelahiran bayi dengan BBLR di Indonesia.

2. METODE PENELITIAN

2.1 Sumber Data dan Variabel Penelitian

Data yang digunakan dalam penelitian ini adalah data sekunder yang bersumber dari hasil SDKI 2017 mengenai berat badan lahir pada bayi yang dilahirkan dengan kategori berat badan lahir normal dan rendah. SDKI 2017 dilaksanakan bersama Badan Pusat Statistik (BPS), Badan Kependudukan dan Keluarga Berencana Nasional (BKKBN) dan Kementerian Kesehatan. Tujuan utama SDKI 2017 adalah menyediakan estimasi terbaru indikator demografi dan kesehatan yang memberikan gambaran menyeluruh tentang kondisi terkini mengenai kependudukan, keluarga berencana (KB), kesehatan reproduksi, serta kesehatan ibu dan anak di Indonesia. SDKI 2017 ini dilakukan pada 34 provinsi di seluruh Indonesia, dengan targe respondennya adalah wanita umur 15-49 tahun, pria kawin/hidup bersama umur 15-54 tahum, dan remaja pria berstatus belum kawin umur 15-24 tahun. Namun, dalam penelitian ini akan digunakan objek pengamatannya adalah wanita umur 15-49 tahun yang memiliki bayi dilahirkan, dengan jumlah sampel sebanyak 16.344 bayi yang dilahirkan. Variabel respon dalam penelitian ini adalah status Berat Bayi Lahir (BBL), yaitu berat badan bayi yang ditimbang dalam waktu 1 jam pertama. Klasifikasi BBL terdiri dari: BBL normal, yaitu bayi dilahirkan dengan berat ≥ 2.500 gram; dan BBL rendah atau (BBLR), yaitu bayi dilahirkan dengan berat lahir <2.500gram tanpa memandang usia gestasi atau dahulu disebut sebagai prematur [12]. Sedangkan, variabel prediktor yang digunakan mengacu pada penelitian [13] yang dapat dilihat pada Tabel 2.

Tabel 2. Variabel Penelitian

Variabel Respon		Kategori
Y	Status Berat Bayi Lahir (BBL)	0: normal (≥ 2.500 gr) 1: rendah (< 2.500 gr)
Variabel Prediktor		**Kategori**
X_1	Kelahiran kembar	1: tidak kembar 2: kembar
X_2	Jenis kelamin bayi	1: perempuan 2: laki-laki
X_3	Jarak dan urutan lahir	1: anak pertama 2: < 2 tahun dan anak ke 2-3 3: ≥ 2 tahun dan anak ke 2-3 4: < 2 tahun dan anak ke 4+ 5: ≥ 2 tahun dan anak ke 4+
X_4	Suplemen zat besi	1: tidak konsumsi 2: konsumsi
X_5	Pemeriksaan kehamilan	1: tidak periksa 2: tradisional 3: medis
X_6	Riwayat keguguran	1: tidak ada riwayat 2: ada riwayat
X_7	Umur ibu	1: terlalu muda 2: tidak berisiko 3: terlalu tua
X_8	Pendidikan ibu	1: tidak sekolah/dasar 2: menengah/tinggi
X_9	Pekerjaan ibu	1: tidak bekerja 2: bekerja
X_{10}	Status ekonomi	1: bawah 2: menengah atas
X_{11}	Sumber air minum	1: tidak layak 2: layak
X_{12}	Kebiasaan merokok	1: tidak merokok 2: merokok
X_{13}	Daerah tempat tinggal	1: perdesaan 2: perkotaan

2.2 Pemodelan klasifikasi

Dalam pengklasifikasian data ada dua proses langkah, yang terdiri dari:

1. Langkah *learning* atau *training*, pada langkah ini dilakukan analisis dari serangkaian *training data* (yaitu objek pengamatan yang label kelasnya diketahui) untuk membangun model klasifikasi.
2. Langkah klasifikasi, pada langkah ini model yang telah dibangun digunakan untuk memprediksi label kelas untuk data yang diberikan (*testing data*). Langkah ini juga digunakan untuk mengetahui kinerja model klasifikasi.

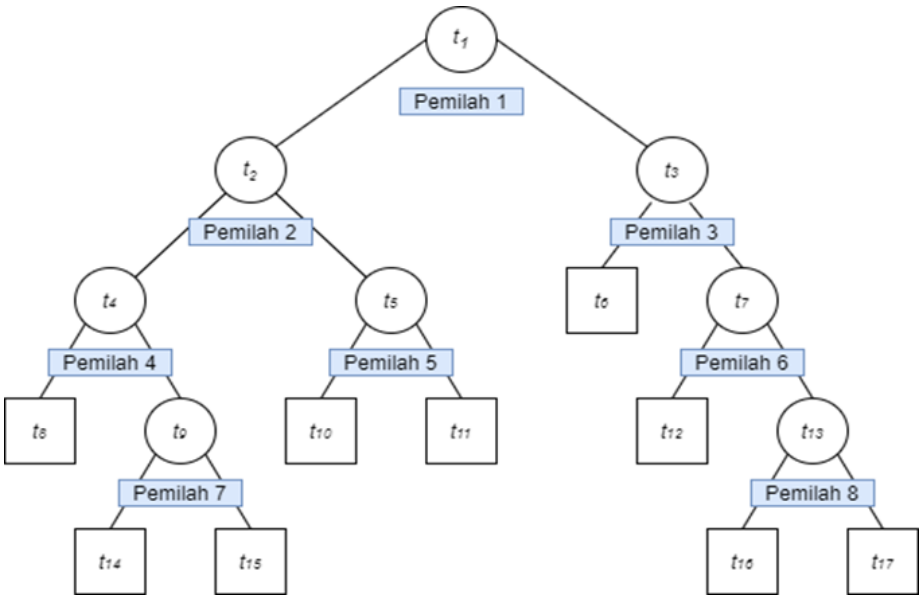
Oleh karena itu, perlu dibentuk 2 (dua) *dataset* yaitu *training data* dan *testing data*. Pada penelitian ini, pembagian *dataset* menggunakan metode deterministik/*holdout*, yaitu dengan menentukan rasio pembagian dari kedua dataset tersebut. Kavzogolu dkk (2012) dalam penelitian [14] menyatakan bahwa mereka melakukan studi tentang efek dari *training data* pada dua model klasifikasi yaitu SVM dan pohon keputusan. Hasil studi menunjukkan bahwa *error* terkecil didapatkan pada jumlah *training*

data 79,09 sampai 86,19 persen. Oeh karena itu, pada penelitian ini digunakan 80 persen dari keseluruhan data untuk *training data* dan sisanya 20 persen data digunakan untuk *testing data* yang dapat menghasilkan tingkat kinerja klasifikasi.

2.3 *Classification and Regression Tree* (CART)

Classification and Regression Tree (CART) adalah suatu metode nonparametrik untuk keperluan klasifikasi dengan cara membangun sebuah pohon klasifikasi yang diperoleh melalui penyekatan berulang terhadap suatu himpunan data menjadi sebuah simpul/*node* baru. CART ini menggunakan algoritma penyekatan rekursif secara biner (*binary recursive* partitioning) [15]. CART akan menghasilkan pohon klasifikasi jika variabel respon berskala kategorik dan akan menghasilkan pohon regresi jika variabel respon berupa data kontinu. CART memiliki tujuan untuk mendapatkan suatu kelompok data yang akurat sebagai penciri dari suatu pengklasifikasian.

Ilustrasi struktur pohon klasifikasi dapat dilihat pada Gambar 1. Simpul awal disebut *parent node* dinotasikan t_1, simpul dalam dinotasikan dengan t_2, t_3, t_4, t_7, t_9, dan t_{10}, serta simpul akhir disebut *terminal node* dinotasikan dengan t_5, t_6, t_8, t_{11}, t_{12}, t_{13}, t_{14} dan t_{15}.



Gambar 1. Struktur Pohon Klasifikasi [16]

2.4 *Random Forest*

Random Forest merupakan salah satu metode dalam *decision tree* atau pohon keputusan. Metode ini digunakan untuk membangun pohon keputusan yang terdiri dari *root node, internal node,* dan *leaf node* dengan mengambil atribut dan data secara acak sesuai ketentuan yang diberlakukan. *Root node* digunakan untuk mengumpulkan data, sebuah *inner node* yang berada pada *root node* berisi pertanyaan tentang data, dan sebuah *leaf node* digunakan untuk memecahkan masalah serta membuat keputusan. Pohon keputusan dimulai dengan cara menghitung nilai *entropy* sebagai penentu tingkat ketidakmurnian atribut dan nilai *information gain* seperti pada Persamaan (1) dan (2) [17].

$$Entropy(Y) = -\sum_i p\,(c|Y) log_2 p(c|Y) \quad (1)$$

dengan Y adalah himpunan kasus dan $p(c|Y)$ adalah proporsi nilai Y terhadap kelas c.

$$Information\ Gain\ (Y, a) = Entropy(Y) - \sum_{v \epsilon Values(a)} \frac{|Y_v|}{|Y_a|} Entropy(Y_v) \quad (2)$$

dengan $Values(a)$ adalah semua nilai yang mungkin dalam himpunan kasus a, Y_v adalah subkelas dari Y dengan kelas v yang berhubungan dengan kelas a, dan Y_a adalah semua nilai yang sesuai dengan a.

2.5 *Naïve Bayes*

Naïve bayes adalah metode klasifikasi statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu kelas. *Naïve bayes* didasarkan pada *Teorema Bayes* yang memiliki kemampuan klasifikasi. *Teorema Bayes* adalah memprediksi

probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya. *Teorema* tersebut dikombinasikan dengan *naïve* yang diasumsikan kondisi antar atribut saling bebas [4].

Klasifikasi *naïve bayes* yang mengacu pada *Teorema Bayes* dapat dilihat pada Persamaan (3).

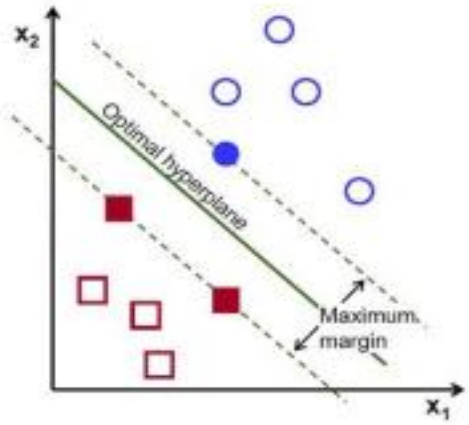
$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (3)$$

dengan:

X : data dengan kelas yang belum diketahui
C : suatu kelas spesifik dari data X
$P(C|X)$: probabilitas bersyarat dari kelas C berdasarkan kondisi X (probabilitas posterior)
$P(C)$: probabilitas kelas C (probabilitas prior)
$P(X|C)$: probabilitas bersyarat dari X berdasarkan kondisi kelas C, disebut sebagai *likelihood*
$P(X)$: probabilitas data X
i : kelas

2.6 *Support Vector Machine* (SVM)

Support Vector Machine (SVM) adalah suatu teknik yang dikembangkan oleh Vapnik pada tahun 1995 untuk melakukan prediksi, baik dalam kasus klasifikasi maupun regresi [18]. SVM merupakan metode *machine learning* yang dapat digunakan untuk klasifikasi baik data linear mapun non-linear. SVM masuk ke dalam kelas *supervised learning*. Ide dasar SVM adalah memaksimalkan batas *hyperplane* seperti pada Gambar 2.



Gambar 2. Optimal *Hyperplane* dengan *Maximum Margin*

Hyperplane (batas keputusan) pemisah terbaik antara kedua kelas diperoleh dengan mengukur *margin hyperplane* tersebut dan mencari titik maksimalnya. *Margin* merupakan jarak antara *hyperplane* tersebut dengan data terdekat dari masing-maisng kelas. Selanjutnya, data yang terdekat tersebut adalah *support vector*. Fungsi kernel yang akan digunakan pada penelitian ini adalah seperti pada Persamaan (4) dan Persamaan (5).

Kernel Linear

$$K(x, y) = (x^T y) \quad (4)$$

Kernal Radial

$$K(x, y) = exp - \left(\frac{|x-y|^2}{2y^2}\right) \quad (5)$$

2.7 Evaluasi Kinerja Klasifikasi

Evaluasi dilakukan untuk pemilihan metode pembagian *dataset* dan metode klasifikasi terbaik yang dilihat melalui ukuran kinerja klasifikasi. Ukuran kinerja klasifkasi yang digunakan dalam penelitian ini dengan memeperhatikan *confussion matrix. Confussion matrix* adalah alat yang berguna utnuk menganalisis seberapa baik atau seberapa akurat metode klasifikasi dapat mengenali objek pengamatan dari kelas yang berbeda [5]. Tabel 3 merupakan *confussion matrix* untuk klasifikasi biner. Bagian kolom menunjukkan label aktual pada setiap kelas, sedangkan bagian baris menunjukkan label kelas berdasarkan hasil prediksi.

Tabel 3. *Confussion Matrix*

Confussion Matrix		Kelas Aktual		Total
		Yes	***No***	
Kelas Prediksi	***Yes***	TP	FP	P'
	No	FN	TN	N'
Total		P	N	

Beberapa ukuran kinerja klasifikasi yang dapat diperoleh dari *confussion matrix* adalah seperti pada Persamaan (6), (7), (8), dan (9).

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

$$Sensitivity = \frac{TP}{TP+FN} \quad (7)$$

$$Specificity = \frac{TN}{TN+FP} \quad (8)$$

$$G-mean = \sqrt{Sensitivity\ x\ Specificity} \quad (9)$$

Akurasi adalah metode yang paling umum digunakan untuk menilai kinerja klasifikasi. Namun, untuk kasus *imbalanced data*, akurasi menempatkan lebih banyak bobot pada kelas mayoritas atau kelas yang tidak menjadi perhatian. Oleh karena itu, akurasi tidak boleh digunakan sebagai ukuran kinerja klasifikasi kasus *imbalanced data* [19], sehingga perlu juga memperhatikan *sensistivity* dan *specificity*. Jika tingkat akurasi tinggi, namun *sensitivity* atau *specificity* rendah, maka pengklasifikasian dapat dikatakan tidak baik. Selain itu, untuk kasus *imbalanced data* dapat diatasi dengan menggunakan *Geometric Mean* (*G-mean*) dan kurva *Receiver Operating Charateristic* (ROC) karena tidak tergantung pada distribusi objek pengamatan antar kelas [20]. Nilai tunggal yang dapat digunakan utnuk mengukur kinerja klasifikasi pada kurva ROC adalah *Area Under Curve the ROC* (AUC).

2.8 Tahapan Analisis Data

Tahapan analisis data pada penelitian ini adalah sebagai berikut:

1. Melakukan *resample* terhadap data dengan menggunakan teknik *combine/both sampling*
2. Melakukan pembagian data dengan teknik deterministik, yaitu 80 persen untuk *training data* dan 20 persen untuk *testing data*
3. Melakukan pemodelan klasifikasi dengan 5 (lima) metode *machine learning*, yaitu CART, random forest, naïve bayes, SVM linear, dan SVM radial.
4. Melakukan evaluasi kinerja klasifikasi terhadap 5 (lima) metode *machine learning* yang digunakan
5. Memilih metode *machine learning* terbaik untuk data BBL berdasarkan kriteria *specificity*, *sensitivity*, *G-mean*, dan AUC tertinggi
6. Pengolahan data menggunakan *software* R versi 4.0.5

3. HASIL DAN ANALISIS

Sebelum dilakukan pemodelan klasifikasi, maka disajikan terlebih dahulu statistik deskriptif mengenai variabel penelitian. Sampel penelitian menunjukkan bahwa sebanyak 7,3% bayi yang dilahirkan di Indonesia mengalami berat badan lahir rendah atau BBLR. Kondisi faktor dasar dan genetis seperti kelahiran kembar dan jenis kelamin bayi laki-laki mempengaruhi risiko kejaidan BBLR. Selain itu, faktor obstetrik seperti jarak yang terlalu dekat (< 2 tahun) dan urutan lahir (urutan ke-4 dan seterusnya) serta memiliki riwayat keguguran juga mempengaruhi risiko kejadian BBLR. Faktor nutrisi seperti suplemen zat besi dan sumber air minum diduga mempengaruhi risiko kejadian BBLR. Pemeriksaan kehamilan secara tradisional ataupun tidak diperiksa sama sekali lebih berisiko mengalami BBLR. Risiko kejadian BBLR juga dipengaruhi oleh faktor demografi dan psikososial, seperti umur ibu yang terlalu muda atau tua, pendidikan ibu, ibu bekerja, status ekonomi bawah, dan kebiasaan merokok. Sedangkan, menurut daerah tempat tinggal menunjukkan bahwa risiko kejadian BBLR lebih besar pada wilayah perdesaan dibandingkan perkotaan. Statistik deskriptif untuk variabel penelitian secara lebih lengkap dapat dilihat pada Tabel 4.

Pemodelan klasifikasi untuk kasus *imbalanced data* dengan metode *machine learning* tanpa dilakukan *treatment* diperoleh nilai akurasi dan *specificity* yang lebih tinggi dibandingkan dengan data yang sudah menggunakan *resample combine/both sampling*. Namun, kinerja klasifikasi lain pada pemodelan klasifikasi tanpa *treatment* seperti *sensitivity* dan *G-mean* justru memiliki nilai 0 dan nilai AUC lebih rendah dibandingkan dengan data yang disudah dilakukan *both sampling* pada kelima model seperti yang disajikan pada Tabel 5.

Pemodelan klasifikasi tanpa dilakukan *treatment* pada kelima model menunjukkan nilai kinerja klasifikasi yang sama, yaitu akurasi sebesar 0,924, *specificity* sebesar 0,924, *sensitivity* sebesar 0, *G-mean* sebesar 0 dan AUC sebesar 0,462. Hal ini sesuai bahwa nilai prediksi cenderung diklasifikasikan pada kelas yang mayoritas (kelas yang tidak diperhatikan) dibandingkan kelas

minoritas (kelas yang diperhatikan dalam hal ini BBLR), sehingga tingkat ketepatan pada pemodelan klasifikasi tanpa dilakukan *treatment* pada kelima model memberikan hasil yang tidak baik. Adanya kesalahan klasifikasi akan mengakibatkan kesalahan fatal dalam perencanaan ataupun pengambilan kebijakan pemerintah dalam penanganan kasus BBLR yang seharusnya akan berdampak pada penurunan AKN. Penerapan *both sampling* dalam penanganan kasus *imbalanced data* dapat meningkatkan ketepatan klasifikasi khususnya terhadap kelas minoritas. Hal ini dapat terlihat bahwa pada kelima model untuk nilai *sensitivity*, yang menunjukkan ukuran ketepatan klasifikasi pada kelas minoritas yang diprediksi secara benar oleh model, berkisar antara 0,521 sampai dengan 0,629. Selain itu, juga meningkatkan nilai *G-mean* dan AUC yang berkisar antara 0,523 sampai dengan 0,651. Pada sisi lain, terjadi penurunan nilai *specificity* menjadi berkisar antara 0,532 sampai dengan 0,674. Dengan kata lain, adanya penanganan kasus *imbalanced data* menghasilkan nilai *specificity* dan *sensitivity* menjadi lebih berimbang yang mengakibatkan nilai akurasi menjadi lebih rendah, yaitu berkisar antara 0,520 sampai dengan 0,648.

Pemodelan klasifikasi untuk memprediksi risiko kejadian BBLR dengan memperhatikan nilai *sensitivity, specificity, G-mean* dan AUC, maka model terbaiknya adalah model *random forest* dengan skema *combine/both sampling*. Hal ini dikarenakan model *random forest* tersebut memiliki nilai kinerja klasifikasi terbesar dibandingkan dengan model klasifikasi lainnya. Model tersebut memiliki nilai *sensitivity* sebesar 0,629, *specificity* sebesar 0,674, *G-mean* sebesar 0,651, dan AUC sebesar 0,651. Oleh karena ukuran kinerja klasifikasi pada model terbaik sudah diatas *cut off* (0,5), maka model tersebut dapat dikatakan baik. Hal ini menunjukkan model klasifikasi terbaik mampu mengklasifikasikan dengan tepat bayi yang baru dilahirkan ke dalam kelas berat bayi lahir normal atau rendah.

Tabel 4. Statistik Deskriptif

Variabel Prediktor		**Status BBL**	
		Normal	**Rendah**
Kelahiran kembar	Tidak kembar	92,7	7,3
	Kembar	88,1	11,9
Jenis kelamin bayi	Perempuan	92,7	7,3
	Laki-laki	92,6	7,4
Jarak dan urutan lahir	Anak pertama	92,8	7,2
	< 2 tahun dan anak ke 2-3	94,0	6,0
	≥ 2 tahun dan anak ke 2-3	92,3	7,7
	< 2 tahun dan anak ke 4+	91,5	8,5
	≥ 2 tahun dan anak ke 4+	93,2	6,8
Suplemen zat besi	Tidak konsumsi	93,2	6,8
	Konsumsi	92,6	7,4
Pemeriksaan kehamilan	Tidak periksa	92,0	8,0
	Tradisional	92,5	7,5
	Medis	93,5	6,5
Riwayat keguguran	Tidak ada Riwayat	92,6	7,4
	Ada Riwayat	93,1	6,9
Umur ibu	Terlalu muda	90,0	10,0
	Tidak berisiko	92,4	7,6
	Terlalu tua	93,4	6,6
Pendidikan ibu	Tidak sekolah/dasar	92,8	7,2
	Menengah/tinggi	92,6	7,4
Pekerjaan ibu	Tidak bekerja	92,8	7,2
	Bekerja	92,5	7,5
Status ekonomi	Bawah	92,5	7,5
	Menengah atas	92,8	7,2
Sumber air minum	Tidak layak	93,3	6,7
	Layak	92,6	7,4
Kebiasaan merokok	Tidak merokok	92,6	7,4
	Merokok	94,0	6,0
Daerah tempat tinggal	Perdesaan	92,6	7,4
	Perkotaan	92,7	7,3

Tabel 5. Perbandingan Kinerja Klasifikasi Pemodelan *Machine Learning*

Model	CART	Naïve Bayes	*Random Forest*	SVM Linear	SVM Radial
No Treatment					
Akurasi	0,924	0,924	0,924	0,924	0,924
Specificity	0,924	0,924	0,924	0,924	0,924
Sensitivity	0	0	0	0	0
G-mean	0	0	0	0	0
AUC	0,462	0,462	0,462	0,462	0,462
Both Sampling					
Akurasi	0,525	0,526	0,648	0,520	0,545
Specificity	0,532	0,530	0,674	0,532	0,556
Sensitivity	0,521	0,522	0,629	0,514	0,537
G-mean	0,526	0,526	0,651	0,523	0,547
AUC	0,526	0,526	0,651	0,523	0,547

Tabel 6 menyajikan nilai *mean decrease gini* dari model klasifikasi *random forest* terbaik. Hal ini menunjukkan bahwa 3 (tiga) variabel terpenting/paling berpengaruh dalam klasifikasi resiko kejadian BBLR adalah jarak dan urutan kelahiran, pemeriksaan kehamilan, umur ibu. Selanjutnya, 3 (tiga) variabel yang paling rendah pengaruhnya terhadap klasifikasi resiko kejadian BBLR adalah sumber air minum, kebiasaan merokok, dan kelahiran kembar.

Tabel 6. Variabel Prediktor yang Paling Berpengaruh menurut *Mean Decrease Gini*

Variabel Prediktor	*Mean Decrease Gini*
Jarak dan urutan kelahiran	158,698
Pemeriksaan kehamilan	107,589
Umur ibu	77,179
Jenis kelamin bayi	77,027
Pekerjaan ibu	74,714
Riwayat keguguran	73,286
Status ekonomi	72,374
Daerah tempat tinggal	70,809
Pendidikan ibu	63,939
Suplemen zat besi	63,439
Sumber air minum	55,756
Kebiasaan merokok	31,181
Kelahiran kembar	21,499

4. KESIMPULAN

Pemodelan klasifikasi dengan menggunakan teknik resample pada kasus *imbalanced data* dan set data besar terbukti mampu meningkatkan ketepatan klasifikasi khususnya terhadap kelas minoritas yang dapat diihat dari nilai sensitivity yang tinggi dibandingkan data asli (tanpa treatment). Selanjutnya, dari kelima model klasifikasi yang iuji menunjukkan bahwa model random forest memberikan kinerja terbaik berdasarkan nilai sensitivity, specificity, G-mean dan AUC tertinggi. Pada penelitian selanjutnya, pemodelan klasifikasi dengan skema both sampling dapat dibandingkan dengan teknik resample lain dalam penanganan kasus imbalanced data agar kinerja klasifikasi lebih optimal. Pada penelitian selanjutnya, pemodelan klasifikasi dengan skema *both sampling* dapat dibandingkan dengan teknik *resample* lain seperti *over sampling* dan *under sampling* dalam penanganan kasus *imbalanced data* agar kinerja klasifikasi lebih optimal. Selain itu dapat menambah variabel independen lainnya yang relevan dalam mempengaruhi BBLR

REFERENSI

[1] Bappenas, *Metadata Indikator Tujuan Pembangunan Berkelanjutan (TPB)/Sustainable Development Goals (SDGs) Indonesia Pilar Pembangunan Sosial*. Jakarta: Bappenas, 2020.

[2] Bkkbn, "Bkkbn Survei Demografi Dan Kesehatan Indonesia 2017," *10 Februari 2017*, 2017. https://www.bkkbn.go.id/detailpost/bkkbn-survei-demografi-dan-kesehatan-indonesia-2017 (accessed May 22, 2021).

[3] Kemenskes, *Profil Kesehatan Indonesia Tahun 2019*. Jakarta: Kementerian Kesehatan RI, 2019.

[4] J. Han, M. Kamber, and J. Pei, "Data Mining Concepts and Techniques," Third Edit., Elsevier, 2012.

[5] M. Maalouf and M. Siddiqi, "Weighted logistic regression for large-scale imbalanced and rare events data," *Knowledge-Based Syst.*, vol. 59, pp. 142–148, 2014, doi: https://doi.org/10.1016/j.knosys.2014.01.012.

[6] G. King and L. Zeng, "Logistic Regression in Rare Events Data," *Polit. Anal.*, vol. 9, no. 2, pp. 137–163, 2001, doi: 10.1093/oxfordjournals.pan.a004868.

[7] T. Purwa, "Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Data Imbalanced (Studi Kasus: Klasifikasi Rumah Tangga Miskin di Kabupaten Karangasem, Bali Tahun 2017)," *J. Mat. Stat. dan Komputasi*, vol. 16, no. 1, pp. 58–73, 2019.

[8] S. H. Sumartini and S. W. Purnami, "Penggunaan Metode Classification and Regression Trees (CART) untuk Klasifikasi Rekurensi

Pasien Kanker Serviks di RSUD Dr. Soetomo Surabaya," *J. Sains dan Seni ITS*, vol. 4, no. 2, pp. 211–216, 2015, [Online]. Available: https://www.neliti.com/publications/15687/penggunaan-metode-classification-and-regression-trees-cart-untuk-klasifikasi-rek.

[9] M. Mambang and A. Byna, "Analisis Perbandingan Algoritma C. 45, Random Forest Dengan Chaid Decision Tree Untuk Klasifikasi Tingkat Kecemasan Ibu Hamil," *Semin. Naisonal Teknol. Inf. dan Multimed. 2017*, vol. 5, no. 1, pp. 103–108, 2017.

[10] C. Oganis, S. Musdalifah, and D. Lusiyanti, "Klasifikasi Status Gizi Ibu Hamil Untuk Mengidentifikasi Bayi Berat Lahir Rendah (Bblr) Menggunakan Metode Support Vector Machine (Svm)(Studi Kasus Di Puskesmas Labuan)," *J. Ilm. Mat. Dan Terap.*, vol. 14, no. 2, pp. 144–151, 2017.

[11] P. L. Kumalasari, "Sistem Pengambilan Keputusan untuk Menentukan Proses Persalinan dengan Metode Naïve Bayes dan Forward Chaining," UNNES, 2020.

[12] M. S. Kosim, A. Yunanto, R. Dewi, G. I. Sarosa, and A. Usman, "Buku ajar neonatologi," *IDI, Jakarta*, 2008.

[13] A. B. Setyawan, K. A. Notodiputro, and Indahwati, "Pemodelan Regresi Logistik Pada Kasus Berat Badan Lahir Rendah (Bblr) Dan Pengaruh Agregasi Data Terhadap Hasil Pendugaan," IPB University, 2015.

[14] M. B. Johra, "Perbandingan Kernel Trick pada Non Linier Support Vector Machine (Studi Kasus: Pemilihan Penolong Persalinan di Provinsi Maluku Utara 2016).," Universitas Padjadjaran, 2018.

[15] R. J. Lewis, "An introduction to classification and regression tree (CART) analysis," in *Annual meeting of the society for academic emergency medicine in San Francisco, California*, 2000, vol. 14.

[16] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[17] K. Schouten, F. Frasincar, and R. Dekker, "An Information Gain-Driven Feature Study for Aspect-Based Sentiment Analysis," in *Natural Language Processing and Information Systems*, 2016, pp. 48–59.

[18] S. R. Gunn, "Support vector machines for classification and regression," *ISIS Tech. Rep.*, vol. 14, no. 1, pp. 5–16, 1998.

[19] M. Maalouf and T. B. Trafalis, "Rare events and imbalanced datasets: an overview," *Int. J. Data Mining, Model. Manag.*, vol. 3, no. 4, pp. 375–388, 2011.

[20] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," in *Icml*, 1997, vol. 97, pp. 179–186.