ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 10, Nomor 1, Tahun 2021, Halaman 11 - 20
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

PENANGANAN KLASIFIKASI KELAS DATA TIDAK SEIMBANG DENGAN RANDOM OVERSAMPLING PADA NAIVE BAYES (Studi Kasus: Status Peserta KB IUD di Kabupaten Kendal)

Reza Dwi Fitriani[1*], Hasbi Yasin[2], Tarno[3]
[1,2,3] Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
***e-mail** rezadwifitriani05@gmail.com

ABSTRACT

The Family Planning Program (KB) launched by the Government of Indonesia to address the problem of population control does not always produce the desired program results. In 2017, there were 7 users of the IUD contraceptive type of contraceptive who failed from 1,102 new IUD users in Kendal Regency so that the ratio of success and failure to the IUD KB program when compared to users of the new IUD KB is 0.64%: 99.36% . The ratio of success and failure of family planning programs which tend to be unbalanced makes it difficult to predict. One of the handling imbalanced data is oversampling, for example using Random Oversampling (ROS). Naive Bayes is used for classification because it's easy and efficient learning model. The data in this study used 14 independent variables and 1 dependent variable. The results of this study indicate that the G-mean of Naive Bayes is less than 60%. The G-mean of ROS-Naive Bayes is 96.6%. It can be concluded that in this research, the ROS-Naive Bayes method is better than the Naive Bayes method for detecting the success status of IUD family planning in Kendal Regency.

Keywords: Naive Bayes, Random Oversampling, G-mean

1. PENDAHULUAN

Klasifikasi merupakan sebuah proses untuk menemukan sebuah model yang menjelaskan dan membedakan konsep atau kelas data dengan tujuan memperkirakan kelas dari suatu objek yang kelasnya tidak diketahui (Tan *et al.*, 2006). Naive Bayes merupakan metode klasifikasi yang sering digunakan karena proses algoritmanya yang lebih cepat dan mudah serta *robust* terhadap data pencilan (Prasetyo, 2012). Klasifikasi dapat diterapkan dalam berbagai aspek sehingga seiring berjalannya waktu metode klasifikasi cukup banyak dikembangkan, namun terdapat permasalahan yang sering ditemui dalam klasifikasi yaitu masalah ketidakseimbangan data.

Ketidakseimbangan data terjadi ketika salah satu kelas memiliki jumlah yang jauh lebih besar dibanding kelas lainnya sehingga menyebabkan menurunnya kinerja klasifikasi pada kelas minoritas. Kinerja algoritma *machine learning* biasanya dievaluasi dengan akurasi hasil prediksi, namun hal ini tidak sesuai apabila terjadi ketidakseimbangan kelas (Chawla *et al.*, 2002). Metode *machine learning* cenderung memberi label berupa kelas mayoritas pada data yang diprediksi dan mengabaikan kelas minoritas sehingga hanya akan menghasilkan akurasi hasil prediksi yang baik bagi kelas mayoritas saja .

Salah satu kasus klasifikasi dengan rasio tidak seimbang adalah status keberhasilan pasien program Keluarga Berencana jenis alat kontrasepsi *Intra Uterine Device* (IUD). Kegagalan pemakaian alat kontrasepsi IUD setiap tahunnya cenderung di bawah 5% dari angka keberhasilannya, sebagai contoh pada tahun 2017 di Kabupaten Kendal tercatat sejumlah 7 pengguna KB jenis alat kontrasepsi *Intra Uterine Device* (IUD) mengalami

kegagalan dari 1.102 pengguna KB IUD baru secara sehingga jika dibandingkan antara jumlah berhasil dan tidak berhasil program KB IUD dengan berdasarkan pengguna KB IUD baru adalah sebesar 0,64 persen dibanding 99,36 persen (BKKBN, 2017). Adanya data tidak seimbang untuk proses klasifikasi menyebabkan hasil klasifikasi data minor menjadi tidak tepat atau tertutupi oleh prediksi data mayor sehingga dibutuhkan solusi untuk mengatasi masalah tersebut salah satunya yaitu melakukan *oversampling* misalnya dengan *Random Oversampling* (ROS) agar rasio ketimpangan kelas bisa dikurangi.

Penelitian klasifikasi dengan kelas tidak seimbang telah banyak dilakukan, salah satunya oleh Mutrofin *et al.* (2019) yang mengusulkan penerapan algoritma *k-Nearest Neighbor* pada kasus pemilihan calon siswa baru memberikan hasil untuk data yang tidak seimbang nilai k yang optimal yaitu $k \geq 100$. Ustyannie & Suprapto (2020) menerapkan metode *Random Oversampling* pada Regresi Logistik untuk data Thennar diperoleh akurasi terbaik dibanding kedua metode *sampling* lainnya yaitu sebesar 78,26%.

Penerapan algoritma yang mengabaikan kelas data yang tidak seimbang akan menghasilkan prediksi yang baik pada kelas mayor, sedangkan kelas minor diabaikan (Chen *et al.*, 2018). Algoritma klasifikasi akan mengalami penurunan performa jika menghadapi kelas data yang tidak seimbang (García *et al.*, 2012). Penelitian ini menggunakan metode *Random Oversampling* (ROS) yang dikombinasikan pada Naive Bayes untuk memprediksi keberhasilan pasien program KB jenis alat kontrasepsi IUD Kabupaten Kendal dengan kondisi kelas data tidak seimbang.

2. TINJAUAN PUSTAKA

2.1. *Imbalanced Class* Data

Imbalanced data merupakan kondisi data yang tidak berimbang antara kelas data satu dengan kelas data yang lain. Kondisi *imbalanced* data menjadi masalah dalam klasifikasi karena *classifier learning* akan condong memprediksi ke kelas data yang banyak (mayoritas) dibanding dengan kelas yang sedikit (minoritas). Akibatnya, dihasilkan akurasi prediksi yang baik terhadap kelas data *training* yang banyak (kelas mayoritas) sedangkan untuk kelas data *training* yang sedikit (kelas minoritas) akan dihasilkan akurasi prediksi yang buruk (Chawla, 2003).

2.2. *Feature Selection*

Langkah paling sederhana dalam *feature selection* adalah dengan mengamati setiap *feature* yang dibangkitkan secara independen dan menguji kemampuan diskriminasinya pada masalah yang harus diselesaikan (Prasetyo, 2014). Langkah ini membantu membuang *feature* dengan kemampuan diskriminasi yang buruk dan mempertahankan *feature* dengan kemampuan diskriminasi yang baik sehingga mampu mengurangi kompleksitas model dan waktu komputasi. *Feature selection* dapat dilakukan antara lain dengan menggunakan uji independensi *Chi-Square* dan uji independensi *Mann-Whitney*.

2.3. Naive Bayes

Algoritma Naive Bayes berakar pada teorema Bayes. Teorema Bayes merupakan teorema yang mengacu pada konsep probabilitas bersyarat (Tan *et al.*, 2006). Metode ini merupakan pendekatan statistik untuk melakukan inferensi induksi pada persoalan klasifikasi. Misalkan A dan B adalah kejadian dalam ruang sampel. Teorema Bayes secara matematis, teorema ini dapat diekspresikan sebagai berikut:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (1)$$

Menurut Prasetyo (2014), jika $\boldsymbol{X}$ merupakan vektor yang berisi fitur dan Y adalah label kelompok, Naive Bayes dituliskan dengan $P(Y|\boldsymbol{X})$. Nilai tersebut berarti probabilitas label

kelompok Y didapatkan setelah fitur-fitur $\boldsymbol{X}$ diamati. Notasi ini disebut juga probabilitas akhir untuk Y, sedangkan $P(Y)$ disebut probabilitas awal untuk Y. Formula Naive Bayes untuk klasifikasi adalah sebagai berikut:

$$P(Y|\,\boldsymbol{X}) = \frac{P(Y)\,\prod_{g=1}^{p} P(X_g|Y)}{P(X)} \qquad (2)$$

Dengan $Y = Y_i$, $i = 1, 2, 3 \ldots, k$
$P(Y|\,\boldsymbol{X})$ adalah probabilitas data dengan vektor $\boldsymbol{X}$ pada kelompok Y
$P(Y)$ adalah probabilitas awal kelompok Y
$\prod_{g=1}^{p} P(\boldsymbol{X_g}|Y)$ adalah probabilitas independen Y dari semua *feature* dalam vektor $\boldsymbol{X}$
$P(\boldsymbol{X})$ adalah probabilitas dari $\boldsymbol{X}$.

Probabilitas $P(\boldsymbol{X})$ selalu tetap sehingga dalam perhitungan prediksi dapat dihilangkan dan hanya menghitung bagian $P(Y)\prod_{g=1}^{p} P(\boldsymbol{X_g}|Y)$ dengan memilih nilai yang terbesar sebagai kelompok yang terpilih sebagai hasil prediksi. Sementara probabilitas independen $\prod_{g=1}^{p} P(\boldsymbol{X_g}|Y)$ merupakan pengaruh semua *feature* dari data terhadap setiap kelompok Y, yang dinotasikan dengan:

$$P(\boldsymbol{X}|\,Y = y) = \prod_{g=1}^{p} P(\boldsymbol{X_g}|Y = y) \text{ dan } \boldsymbol{X} = [X_1, X_2, X_3, \ldots, X_p] \qquad (3)$$

Umumnya, Bayes mudah dihitung untuk *feature* bertipe kategorik, namun untuk *feature* dengan tipe numerik (non kategorik) ada perlakuan khusus sebelum diproses menggunakan Naive Bayes (Prasetyo, 2012). Caranya adalah:

a. Melakukan diskritisasi pada setiap *feature* kontinu dan mengganti nilai *feature* kontinu tersebut dengan nilai interval diskrit. Pendekatan ini dilakukan dengan mentransformasi *feature* ke dalam *feature* ordinal.
b. Mengasumsikan bentuk tertentu dari distribusi probabilitas untuk *feature* kontinu dan memperkirakan parameter distribusi dengan data *training*. Distribusi Gaussian biasanya dipilih untuk mempresentasikan probabilitas bersyarat dari *feature* kontinu pada sebuah kelompok $P(X_g|Y)$, sedangkan distribusi Gaussian dikarakteristikkan dengan dua parameter yaitu *mean* (μ) dan *varian* (σ^2) untuk setiap kelompok Y_i, probabilitas bersyarat kelompok Y_i untuk *feature* X_g adalah $P(X = x_g|Y = y_1) = g(x_g, \mu_{gi}, \sigma_{gi})$, dengan

$$g(x_g, \mu_{gi}, \sigma_{gi}) = \frac{1}{\sqrt{2\pi}\sigma_{gt}} e^{\left(\frac{x_{gi}-\mu_{gi}}{2\sigma_{gt}}\right)^2},\ i = 1, 2 \text{ dan } g = 1, 2, 3, \ldots, p \qquad (4)$$

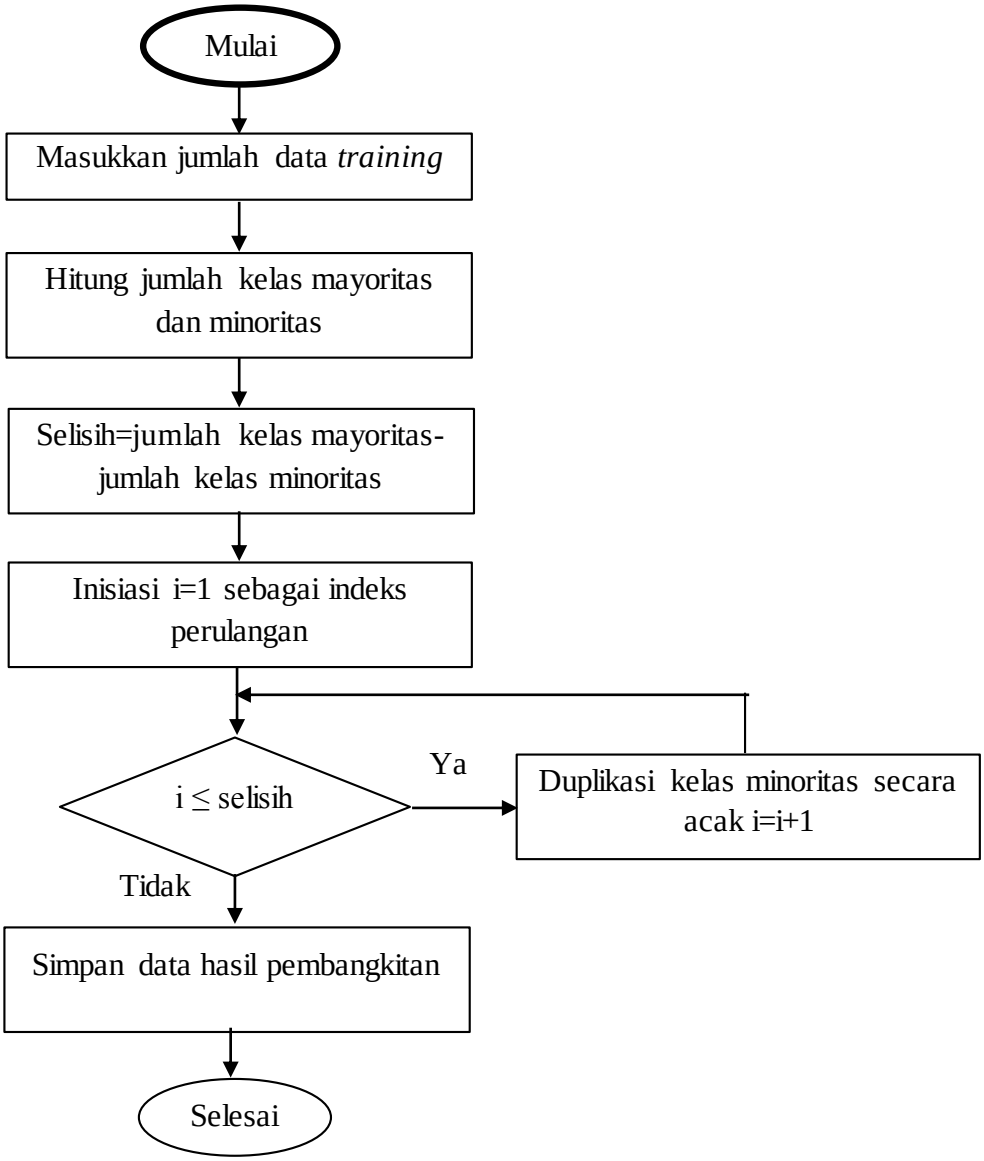
Parameter μ_{gi} diperoleh dari sampel $X_g\,(\bar{X})$ dari semua data *training* yang memiliki kelompok Y_i, sedangkan σ_{gi}^2 dapat diperkirakan dari *varian* sampel (s^2) dari data *training*.

Menurut Hastuti (2016) model klasifikasi yang tidak memerlukan asumsi non multikolinieritas dan normalitas antara lain Metode Naive Bayes dan *Decision Tree*. *Naïve Bayes* mengaplikasikan Teorema Bayes yang mengestimasi parameter mengikuti distribusi data sedangkan *Decision Tree* bersifat nonparametrik. Oleh karena itu, kedua metode tersebut dapat digunakan untuk mengklasifikasikan data yang tidak memenuhi asumsi normal.

2.4. *Random Oversampling* (ROS)

Random Oversampling (ROS) merupakan penambahan data dari kelas minoritas ke dalam data *training* secara acak. Proses penambahan ini diulang sampai jumlah data kelas minoritas sama dengan jumlah kelas mayoritas. Pertama dihitung selisih antara kelas mayoritas dengan kelas minoritas. Selanjutnya, dilakukan perulangan sebanyak hasil

penghitungan selisih sambil membaca data kelas minoritas secara acak dan ditambahkan ke dalam data *training* (Chawla *et al.*, 2002). Berikut adalah *flowchart* algoritma *Random Oversampling* (Saifudin & Wahono, 2015):



Gambar 1. *Flow Chart* Algoritma *Random Oversampling*

2.5. Ukuran Kinerja Klasifikasi

Kasus kelas tidak seimbang dengan kelas mayoritas 98-99% dari keseluruhan populasi akan menghasilkan hasil klasifikasi akan mencapai akurasi tinggi karena hanya melihat kelas mayoritas saja. Jelas bahwa untuk kasus kelas tak seimbang, akurasi klasifikasi tidak cukup sebagai ukuran kriteria standar (Pangastuti *et al.*, 2018). Evaluasi kinerja metode secara keseluruhan dapat dilakukan dengan menggunakan *geometric mean* (*G-mean*). *G-mean* merupakan rata-rata geometrik sensitivitas dan spesifisitas (Kubat & Matwin, 1997)

$$\text{Akurasi} = \frac{\text{TP+TN}}{(\text{TP+FP+FN+TN})} \times 100\% \quad (5)$$

$$Spesifisitas = \frac{TN}{(TN+FP)} \times 100\% \quad (6)$$

$$Sensitivitas = \frac{TP}{(TP+FN)} \times 100\% \quad (7)$$

$$G - Mean = \sqrt{Sensitivitas \times Spesifisitas} \quad (8)$$

2.6. *Holdout Cross Validation*

Salah satu cara untuk membagi data *training* dan data *testing* yang paling sederhana adalah *holdout cross validation*. *Holdout cross validation* akan membagi data menjadi 2 bagian dengan proporsi tertentu yang ditentukan oleh peneliti. Proporsi yang biasa digunakan oleh peneliti adalah 60/40, 70/30, atau 80/20 (Raschka, 2018). Proporsi data pelatihan lebih besar dibandingkan dengan data uji. Proporsi yang dipilih harus tepat yaitu tidak terlalu besar untuk data pelatihan karena akan menyebabkan validasi dengan data uji menjadi kurang mencerminkan keakuratan model sebenarnya.

3. METODE PENELITIAN

3.1. Data Penelitian

Data yang digunakan dalam penelitian ini adalah data status peserta Keluarga Berencana (KB) jenis alat kontrasepsi *Intra Uterine Device* (IUD) Kabupaten Kendal pada tahun 2018. Data tersebut merupakan data sekunder yang diperoleh dari formulir kartu status peserta KB K/IV Petugas Lapangan Keluarga Bencana (PLKB) Badan Kependudukan dan Keluarga Berencana Nasional (BKKBN) Kabupaten Kendal tahun 2018 dan dilakukan proses pengambilan data pada Januari 2019. Data diambil dari responden yang telah memasang KB jenis IUD sejak 2015, diperoleh sebanyak 250 pasien dan terdapat 15 variabel.

3.2. Variabel Penelitian

Variabel yang digunakan dalam penelitian ini terdiri dari variabel dependen (Y) dan beberapa variabel independen (X) dengan rincian sebagai berikut:

1. Variabel dependen
   Y = Status keberhasilan KB IUD, 1 = berhasil, 0 = tidak berhasil

2. Variabel independen
   X_1 = Usia, X_2 = Cara KB terakhir, X_3 = Status menyusui, X_4 = Pendarahan pervaginam, X_5 = Keputihan, X_6 = Riwayat tumor, X_7 = Berat badan, X_8 = Tekanan darah sistolik, X_9 = Tekanan darah diastolik, X_{10} = Posisi rahim, X_{11} = Radang, X_{12} = Tumor ganas *ginekologi*, X_{13} = Diabetes, dan X_{14} = Kelainan pembekuan darah.

3.3. Analisis Data

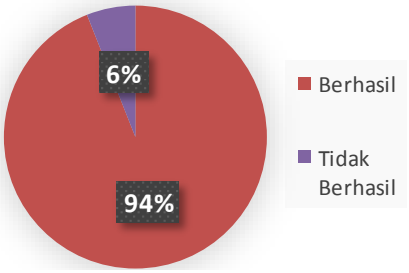
Software yang digunakan untuk pengolahan data ini adalah RStudio versi 1.1.463. Metode yang digunakan adalah *Synthetic Minority Oversampling Technique* (SMOTE), *Random Oversampling* (ROS), dan Naive Bayes. Langkah-langkah analisis data pada penelitian ini adalah sebagai berikut:

1. Input data
2. Melakukan analisis deskriptif
3. Melakukan *pre-processing*
4. Melakukan pembagian data *training* dan data *testing*
5. Melakukan *copy* data *training* menjadi tiga data *training*
6. Melakukan klasifikasi dengan Naive Bayes
7. Melakukan pengukuran kebaikan klasifikasi Naive Bayes
8. Melakukan klasifikasi dengan ROS-Naive Bayes
9. Melakukan pengukuran kebaikan klasifikasi ROS-Naive Bayes
10. Menentukan model klasifikasi terbaik diantara Naive Bayes dan ROS-Naive Bayes berdasarkan ukuran kebaikan klasifikasi.

4. HASIL DAN PEMBAHASAN

4.1. Rasio Variabel Dependen

Berdasarkan data status keberhasilan KB IUD 250 pasien di BKKBN Kabupaten Kendal tahun 2018 terlihat bahwa terdapat 235 pasien (94%) status program pasien KB berhasil dan 15 pasien lainnya tidak berhasil dalam menjalani program KB jenis IUD. Distribusi frekuensi status keberhasilan KB IUD dapat dilihat pada Gambar 3.



Gambar 2. Persentase Status Keberhasilan KB IUD

4.2. Feature Selection

a) Uji Independensi *Chi-Square*

Hipotesis yang digunakan pada uji ini yaitu:

H_0 : Tidak terdapat hubungan antara variabel independen dengan variabel dependen

H_1 : Terdapat hubungan antara variabel independen dengan variabel dependen

Taraf Signifikansi α=0,05

Statistik Uji

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right] \quad (9)$$

$$E_{ij} = \frac{(n_{i.})(n_{.j})}{N} \quad (10)$$

O_{ij}: Frekuensi obyek teramati baris ke-i dan kolom ke-j (n_{ij}).

E_{ij} : Frekuensi harapan obyek baris ke-i dan kolom ke-j.

derajat kebebasan dk= (r-1)(c-1)

r : banyaknya baris

c : banyaknya kolom

Kaidah Pengambilan Keputusan

H_0 ditolak apabila nilai $\chi^2 > \chi^2_{\alpha,(r-1)(c-1)}$ atau nilai signifikansi $< \alpha$.

Hasil perhitungan nilai *Chi-Square* dan probabilitas masing-masing variabel ditampilkan pada Tabel 1.

Tabel 1. Hasil Uji *Chi-Square*

Variabel	*Chi-Square*	Signifikansi
X_2	23,847	0,000
X_3	93,705	0,000
X_4	1,6285	0,202
X_5	18,861	0,000
X_6	3,931	0,047
X_{10}	39,623	0,000
X_{11}	18,573	0,000
X_{12}	43,608	0,000
X_{13}	22,182	0,000
X_{14}	3,653	0,056

Berdasarkan Tabel 2 dapat disimpulkan bahwa variabel X_2, X_3, X_5, X_6, X_{10}, X_{11}, X_{12}, X_{13}, dan X_{14} yang berhubungan dengan variabel dependen (Y).

b) Uji Independensi *Mann-Whitney*

Hipotesis yang digunakan pada uji ini yaitu:

$H_0 : M_x = M_y$

$H_1 : M_x \neq M_y$

Taraf Signifikansi α=0,05

Statistik Uji

$$T = S - \frac{n_1(n_1+1)}{2} \tag{11}$$

Pengujian dengan data berukuran besar (n_1, $n_2 > 20$) dapat didekati dengan distribusi normal sebagai berikut:

$$Z = \frac{T - \frac{(n_1 n_2)}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12} - \frac{n_1 n_2 (\sum t^3 - \sum t)}{12(n_1+n_2)(n_1+n_2-1)}}} \tag{12}$$

t : jumlah *ties*

n_1 : jumlah data populasi 1

n_2 : jumlah data populasi 2

Probabilitas hasil uji *Mann-Whitney* masing-masing variabel ditampilkan pada Tabel 2.

Tabel 2. Hasil Uji *Mann-Whitney*

Variabel	Signifikansi
X_1	0,003
X_7	0,002
X_8	0,085
X_9	0,000

Berdasarkan Tabel 3 dapat disimpulkan bahwa variabel X_1, X_7, dan X_9 yang berhubungan dengan variabel dependen (Y).

Pada analisis selanjutnya akan dibandingkan pula klasifikasi antara data *non feature selection* dan data dengan *feature selection* menggunakan model Naive Bayes dan ROS-Naive Bayes pada tiga bentuk *split* data yaitu 75/25, 80/20, dan 85/15.

4.3. Hasil Klasifikasi dengan Naive Bayes dan ROS-Naive Bayes

Model pembelajaran klasifikasi yang dibangun pada penelitian ini adalah Naive Bayes sedangkan metode *oversampling* yang digunakan yaitu ROS. Prediksi dilakukan pada data *training* dan data *testing* yang telah dibagi dengan metode *holdout cross validation*. Proporsi *split* data *training* dan data *testing* yang digunakan pada penelitian ini ada tiga macam yaitu 75/25, 80/20, dan 85/15 dengan membandingkan antara tanpa *feature selection* dan menggunakan *feature selection* pada data. Hasil klasifikasi data *test* dapat ditampilkan pada Tabel 3.

Tabel 3. Hasil Klasifikasi Data *Testing*

Feature Selection	Model	Akurasi	Sensitivitas	Spesifisitas	*G-mean*
Proporsi *Split* Data *Testing* 75/25					
Tidak	Naive Bayes	0,952	0,250	1,000	0,500
	ROS-Naive Bayes	0,921	1,000	0,915	0,957
Ya	Naive Bayes	0,952	0,250	1,000	0,500
	ROS-Naive Bayes	0,937	1,000	0,932	0,966
Proporsi *Split* Data *Testing* 80/20					
Tidak	Naive Bayes	0,960	0,333	1,000	0,577
	ROS-Naive Bayes	0,960	1,000	0,957	0,978
Ya	Naive Bayes	0,960	0,333	1,000	0,577
	ROS-Naive Bayes	1,000	1,000	1,000	1,000
Proporsi *Split* Data *Testing* 85/15					
Tidak	Naive Bayes	0,973	0,500	1,000	0,707
	ROS-Naive Bayes	0,973	1,000	0,971	0,986
Ya	Naive Bayes	0,973	0,500	1,000	0,707
	ROS-Naive Bayes	0,838	1,000	0,829	0,910

4.4. Perbandingan Hasil Klasifikasi

Hasil klasifikasi yang sudah diperoleh kemudian dibandingkan ukuran kebaikan klasifikasi berdasarkan model, penggunaan *feature selection*, dan *split data*.

a) Perbandingan Model

Rata-rata ukuran hasil klasifikasi data *testing* model Naive Bayes dan ROS-Naive Bayes. dapat ditampilkan pada Tabel 4.

Tabel 4. Perbandingan Model

Ukuran Kebaikan Model	Naive Bayes	ROS-Naive Bayes
Akurasi	0,962	0,938
Sensitivitas	0,361	1,000
Spesifisitas	1,000	0,934
G-mean	0,595	0,966

Berdasarkan Tabel 4 nilai akurasi kedua model berada di atas 93%. Setelah dilakukan pembangkitan data dengan ROS akurasinya lebih kecil dibandingkan sebelum dilakukan *balancing* data, namun nilai *G-mean*nya jauh lebih besar yaitu 0,966. Model sebelum dan sesudah dilakukan ROS mempunyai nilai sensitivitas yang sangat berbeda, apabila tidak

dilakukan ROS maka model Naive Bayes memiliki sensitivitas yang sangat kecil. Nilai sensitivitas yang kecil pada Naive Bayes mengindikasikan bahwa model tersebut tidak mampu melakukan klasifikasi secara benar untuk pasien tidak berhasil KB IUD. Model yang baik adalah model yang mampu melakukan klasifikasi secara tepat pada semua jenis kelas. Algoritma ROS memperbaiki hasil klasifikasi menjadi jauh lebih baik terlihat bahwa nilai *G-mean* lebih besar dibandingkan tanpa *balancing* sehingga model ROS-Naive Bayes dinilai yang terbaik untuk model klasifikasi data KB IUD Kabupaten Kendal.

b) Perbandingan Penggunaan Seleksi Fitur

Rata-rata ukuran hasil klasifikasi data *testing* penggunaan *feature selection* pada data dapat ditampilkan pada Tabel 5.

Tabel 5. Perbandingan Penggunaan *Feature Selection*

Ukuran Kebaikan Model	*Non Feature Selection*	*Feature Selection*
Akurasi	0,957	0,943
Sensitivitas	0,681	0,681
Spesifisitas	0,974	0,960
G-mean	0,784	0,777

Berdasarkan Tabel 5 hasil klasifikasi data *non feature selection* menghasilkan nilai *G-mean* yang lebih tinggi dibandingkan hasil klasifikasi data yang dilakukan *feature selection*. Hal ini dapat dikatakan model klasifikasi data *non feature selection* dinilai lebih baik dibandingkan klasifikasi data yang dilakukan *feature selection*.

5. KESIMPULAN

1. Terdapat 11 variabel independen yang memiliki asosiasi terhadap variabel respon, yaitu variabel usia (X_1), variabel cara KB terakhir (X_2), variabel status menyusui (X_3), variabel keputihan (X_5), variabel riwayat tumor (X_6), variabel berat badan (X_7), variabel darah diastolik (X_9), variabel posisi rahim (X_{10}), variabel radang (X_{11}), variabel tumor ganas *ginekologi* (X_{12}), dan variabel diabetes (X_{13}).
2. Berdasarkan hasil klasifikasi yang dilakukan dengan dua cara yaitu dengan *feature selection* dan *non feature selection*, terlihat bahwa rata-rata ukuran kebaikan klasifikasi model *non feature selection* lebih baik dibanding dengan *feature selection* karena rata-rata *G-mean* model *non feature selection* lebih besar dibandingkan dengan *feature selection*.
3. Hasil klasifikasi terbaik untuk diterapkan pada klasifikasi status keberhasilan program KB tahun 2018 adalah metode ROS-Naive Bayes. Hal ini dikarenakan nilai sensitivitas dan *G-mean* data *testing* yang lebih tinggi dibandingkan Naive Bayes sehingga apabila peneliti ingin prediksi secara lebih akurat untuk kelas tidak berhasil KB IUD Kabupaten Kendal maka lebih baik menggunakan metode ROS-Naive Bayes.

DAFTAR PUSTAKA

BKKBN. (2017). *Laporan Program KB Nasional, Dalap Tabel 8A Kumulatif*. Tersedia di: http://aplikasi.bkkbn.go.id/sr/Klinik/Laporan2013/Bulanan/Faskes2013Tabel8aKumulatif.aspx (Diakses pada: 26 February 2021).

Chawla, N. V (2003). C4.5 and Imbalanced Data Sets : Investigating the Effect of Sampling Method, Probabilistic Estimate, and Decision Tree Structure. *ICML Workshop Learning from Imbalanced Data Sets II*. Washington D.C.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* Vol. 16, No. 2, Hal: 321–357.

Chen, L., Fang, B., Shang, Z., & Tang, Y. (2018). Tackling class overlap and imbalance problems in software defect prediction. *Software Quality Journal* Vol. 26, No. 1, Hal: 97-125 doi: 10.1007/s11219-016-9342-6.

García, V., Sánchez, J. S. & Mollineda, R. A. (2012). On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, Vol: 25, No. 1, Hal: 13-21. doi: 10.1016/j.knosys.2011.06.013.

Hastuti, Y. (2016). Klasifikasi Karakteristik Mahasiswa Universitas Cokroaminoto Palopo Menggunakan Metode Naive Bayes dan Decision Tree. *Jurnal Dinamika* Vol. 07, No. 2, Hal: 34-41.

Kubat, M. & Matwin, S. (1997). Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. *Fourteenth International Conference on Machine Learning* Hal: 179-186.

Mutrofin, S., Mualif, A., Ginardi, R. V., & Fatichah, C. (2019). Solution of Class Imbalance of K-Nearest Neighbor for Data of New Student Admission Selection. *International Journal of Artificial Intelligence Research* Vol. 3, No. 2, Hal: 47-55. doi: 10.29099/ijair.v3i2.92.

Pangastuti, S. S., Fithriasari, K. & Irawan, N. (2018). *Perbandingan Metode Ensemble Random Forest dengan Smote-Boosting dan Smote-Bagging pada Klasifikasi Data Mining untuk Kelas Imbalance*. Surabaya: Institut Teknologi Sepuluh Nopember.

Prasetyo, E. (2012). *Data Mining - Konsep dan Aplikasi Menggunakan MATLAB*. Yogyakarta: ANDI.

Prasetyo, E. (2014). *Data Mining - Mengolah Data Menjadi Informasi Menggunakan Matlab*. Yogyakarta: ANDI.

Raschka, S. (2018). *Model evaluation, model selection, and algorithm selection in machine learning*.

Saifudin, A. & Wahono, R. P. (2015). Pendekatan Level Data untuk Menangani Ketidakseimbangan Kelas pada Prediksi Cacat Software. *Journal of Software Engineering* Vol. 3, No. 2, Hal: 47-55.

Tan, P., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. Boston: Pearson Education

Ustyannie, W., & Suprapto, S. (2020). Oversampling Method to Handling Imbalanced Dataset Problem in Binary Logistic Regression Algorithm. *Indonesian Journal of Computing and Cybernetics Systems* Vol. 14, No.1, Hal: 1-10. doi: 10.22146/ijccs.37415.