Leibniz : Jurnal Matematika
Volume 2, Nomor 2, Halaman 11–22
Juli 2022
e-ISSN: 2775-2356

PENERAPAN KLASIFIKASI *NAIVE BAYES* DENGAN ALGORITMA *RANDOM OVERSAMPLING* DAN *RANDOM UNDERSAMPLING* PADA DATA TIDAK SEIMBANG *CERVICAL CANCER RISK FACTORS*

Jus Prasetya

Program Studi Statistika, Fakultas Sains dan Teknologi,
Universitas YPPI Rembang
email: jusprasetya777@gmail.com

Abstrak: *Machine learning* adalah cabang ilmu komputer yang memanfaatkan pengalaman (peristiwa) pada masa lalu untuk dipelajari dan menggunakan pengetahuannya untuk membuat keputusan di masa depan. Pada *machine learning*, proses klasifikasi dilakukan untuk meminimalkan kesalahan klasifikasi maka dengan demikian akan memaksimalkan akurasi prediksi. Asumsi yang mendasari metode klasifikasi ini adalah bahwa data yang diteliti memiliki jumlah sampel yang seimbang setiap kelas yang tersedia. *Random Oversampling* adalah proses resamplingnya dengan cara memilih sampel dari kelas minoritas secara acak, sampel yang dipilih secara acak ini kemudian diduplikasi dan ditambahkan ke set pelatihan baru. *Random Undersampling* adalah proses resampling dengan sampel pada kelas mayoritas dalam set pelatihan dihilangkan secara acak sampai rasio antara kelas minoritas dan mayoritas berada pada tingkat yang diinginkan. Nilai AUC yang didapatkan pada klasifikasi *naive bayes* sebesar 0,5325 yang berarti klasifikasi gagal. Nilai AUC yang didapatkan pada klasifikasi *random oversampling-naive bayes* sebesar 0,62 yang berarti klasifikasi buruk. Nilai AUC yang didapatkan pada klasifikasi *random undersampling-naive bayes* sebesar 0,7013 yang berarti klasifikasi cukup baik.

Kata Kunci: *Machine learning, oversampling, undersampling, naive bayes, AUC*

Abstract: *Machine learning is a branch of computer science that utilizes past experiences (events) to learn and uses this knowledge to make future decisions. In machine learning, the classification process is carried out to minimize classification errors, thereby maximizing prediction accuracy. The assumption underlying this classification method is that the data studied have a balanced number of samples for each available class. Random Oversampling is a resampling process by selecting a sample from the minority class at random, this*

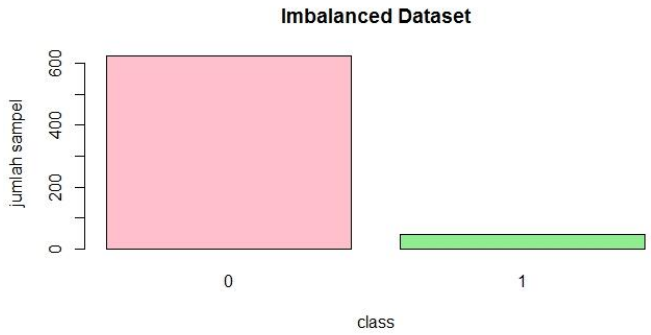
randomly selected sample is then duplicated and added to a new training set. Random Undersampling is a resampling process in which the sample in the majority class in the training set is removed randomly until the ratio between the minority and majority class is at the desired level. The AUC value obtained in the Naive Bayes classification is 0.5325, which means the classification failed. The AUC value obtained in the random oversampling-naive bayes classification is 0.62, which means the classification is bad. The AUC value obtained in the random undersampling-naive bayes classification is 0.7013, which means the classification is fair.

Keywords: *Machine learning, oversampling, undersampling, naive bayes, AUC*

PENDAHULUAN

Machine learning adalah cabang ilmu komputer yang memanfaatkan pengalaman (peristiwa) pada masa lalu untuk dipelajari dan menggunakan pengetahuannya untuk membuat keputusan di masa depan. *Machine learning* merupakan irisan ilmu komputer, teknik dan statistika. Tujuan *machine learning* adalah untuk menggeneralisasi pola atau membuat aturan yang tidak diketahui sebelumnya dari data yang telah diberikan (Dangeti, 2017).

Pada *machine learning*, proses klasifikasi dilakukan untuk meminimalkan kesalahan klasifikasi maka dengan demikian akan memaksimalkan akurasi prediksi. Asumsi yang mendasari metode klasifikasi ini adalah bahwa data yang diteliti memiliki jumlah sampel yang seimbang dari setiap kelas yang tersedia. Pada proses klasifikasi diasumsikan bahwa probabilitas dari kelas variabel target (respon) adalah seimbang. Namun, di banyak kasus dunia nyata seperti diagnostik medis, sebagian besar data klasifikasi cenderung condong ke nilai kelas negatif. Data dikatakan tidak seimbang jika setidaknya salah satu dari nilai variabel target memiliki jumlah sampel yang jauh lebih kecil jika dibandingkan dengan nilai variabel target pada kelas lainnya (Thabtah dkk, 2019).



Gambar 1. *Imbalanced Dataset Cervical Cancer Risk Factors*

Leibniz : Jurnal Matematika
Volume 2, Nomor 2, Halaman 11–22
Juli 2022
e-ISSN: 2775-2356

Ketidakseimbangan kelas adalah salah satu faktor paling berpengaruh dalam kinerja prediksi klasifikasi. Pengklasifikasi cenderung akan membuat model pembelajaran bias yang memiliki akurasi prediksi yang buruk atas kelas minoritas dibandingkan dengan kelas mayoritas. Ini karena sebagian besar pembelajaran pengklasifikasi, seperti *decision tree, backpropagation neural network, support vector machines*, dan lainnya dirancang berdasarkan dengan asumsi bahwa distribusi kelas relatif seimbang (Zheng, 2020).

Pada kasus data tidak seimbang terdapat tiga alternatif untuk mengatasi ketidakseimbangan. i) Pendekatan level algoritma, yaitu pendekatan yang mengatasi masalah distribusi kelas dengan memodifikasi tahapan proses *learning*. ii) Pendekatan *ensemble*, yaitu pendekatan dengan membagi kelas mayoritas menjadi beberapa himpunan bagian ukuran yang sama dengan kelas minoritas, kemudian sub himpunan dilatih secara terpisah sehingga masing-masing menghasilkan model pembelajaran dan menentukan model pembelajaran terbaik berdasarkan *majority voting*. iii) Pendekatan level data, yaitu pendekatan dengan menyesuaikan rasio kelas pada dataset sehingga mencapai distribusi kelas yang seimbang. Pendekatan level data terdiri dari dua metode yaitu *oversampling* dan *undersampling* (Osorio dkk, 2021).

Algoritma *oversampling* yaitu metode mereplikasi dan/atau menghasilkan sampel pada kelas minoritas sedangkan algoritma *undersampling* menghilangkan sampel pada kelas mayoritas. Menghilangkan sampel pada kelas mayoritas merupakan hal yang baik untuk menghindari kemungkinan terjadinya *noise* yang dapat memperburuk hasil klasifikasi (Rodríguez dkk, 2021).

Penyakit kanker serviks menjadi penyebab kematian yang signifikan di seluruh dunia, hal tersebut dapat dicegah dan disembuhkan dengan cara membuang jaringan yang terkena pada tahap awal. Dalam Fernandes dkk (2018) dijelaskan bahwa strategi komputasi untuk memprediksi hasil biopsi pasien berdasarkan pola resiko dari catatan medis pasien. Teknik klasifikasi yang diusulkan dengan algoritma *supervised deep learning* dan menghasilkan prediksi dengan akurasi area teratas di bawah kurva AUC sebesar 0,687.

Teknik klasifikasi *naive bayes* adalah metode klasifikasi probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari data yang diberikan yang akan digunakan untuk memprediksi probabilitas sampel kedalam suatu kelas (respon). Asumsi utama dari klasifikasi *naive bayes* adalah setiap variabel prediktor saling independen (Fallo, 2021).

Gorunescu (2011) menyatakan teorema bayes (aturan bayes atau rumus bayes) mendefinisikan probabilitas bersyarat antara dua peristiwa sebagai

$$P\{A_i|B\} = \frac{P\{B|A_i\}P\{A_i\}}{\sum_{i=1}^{n} P\{B|A_i\}P\{A_i\}}, P\{B\} > 0, P\{A_i\} > 0, i = 1,2, \dots, n$$

$P\{A_i|B\}$ adalah probabilitas *posterior* karena diturunkan dari, atau bergantung pada, nilai tertentu dari B, $P\{A_i\}$ adalah probabilitas prior karena tidak memperhitungkan informasi apa pun tentang B, $P\{B|A_i\}$ adalah *likelihood*, dan $P\{B\}$ *evidence*. Dalam konteks ini, rumus Bayes dapat ditulis sebagai :

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior probability}}{\text{evidence}}$$

The Bayes' formula dapat dituliskan dalam bentuk yang sederhana, dihubungankan dengan rumus probabilitas bersyarat. Berdasarkan rumus probabilitas yang menghubungkan dua kejadian A dan B, ($P\{B\} \neq 0$), maka dapat ditulis sebagai :

$$P\{A|B\} = \frac{P\{A \cap B\}}{P\{B\}},$$

dengan $P\{A|B\}$ merupakan probabilitas kejadian A dengan syarat B terjadi, atau rumus bayes dapat dituliskan dalam bentuk sederhana lainnya sebagai :

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

Teori keputusan bayesian adalah metode dasar statistik dalam bidang klasifikasi. Dalam teori keputusan, tujuan dasarnya adalah untuk meminimalkan probabilitas membuat keputusan yang salah, atau risiko yang diharapkan.

Dalam kasus variabel kontinu, untuk memperkirakan probabilitas bersyarat $P\{A_i|B_j\}$ perlu mengidentifikasi jenis distribusi dari variabel, dimana variabel sebagai variabel random kontinu. Asumsikan bahwa semua variabel kontinu terdistribusi secara normal, dengan estimasi parameternya yaitu *mean* dan *variance*. Maka fungsi kepadatan probabilitas dapat mengevaluasi probabilitas bersyarat $P\{A_i|B_j\}$ untuk setiap kelas secara terpisah, maka fungsi kepadatan probabilitas sebagai:

$$P\{A_i|B_j\} = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp\left(-\frac{(A_i - \mu_{ij})^2}{2\sigma_{ij}^2}\right) \quad (2)$$

Tujuan penelitian ini adalah mengatasi ketidakseimbangan kelas pada dataset *cervical cancer risk factors* dengan menggunakan pendekatan level data yakni *random oversampling* dan *random undersampling* dan melakukan teknik klasifikasi *naive bayes* dengan performa model dilihat berdasarkan nilai output akurasi, sensitivitas, spesifisitas, presisi dan *Area Under Curve (AUC) ROC.*

Leibniz : Jurnal Matematika
Volume 2, Nomor 2, Halaman 11–22
Juli 2022
e-ISSN: 2775-2356

METODE PENELITIAN

Jenis penelitian yang digunakan dalam penelitian ini adalah studi literatur dengan menggunakan data sekunder *cervical cancer risk factors* yang diperoleh dari *UCI Machine Learning Repository Data Set*s merupakan portal big data internasional. Data terdiri dari 23 variabel independen dan 1 variabel target (respon) yaitu *Biopsy* dengan nilai 0 (tidak kanker serviks) dan 1 (kanker serviks). Data *cervical cancer risk factors* dijelaskan pada Tabel 1 berikut.

Tabel 1. Variabel pada Data *cervical cancer risk factors*

No.	Variabel	Type
1.	Age	Numerik
2.	Number.of.sexual.partners	Numerik
3.	First.sexual.intercourse	Numerik
4.	Num.of.pregnancies	Numerik
5.	Smokes	Kategorik
6.	Smokes..packs.year.	Numerik
7.	Hormonal.Contraceptives	Kategorik
8.	Hormonal.Contraceptives..years.	Numerik
9.	IUD	Kategorik
10.	IUD..years.	Numerik
11.	STDs	Kategorik
12.	STDs..number.	Numerik
13.	STDs.condylomatosis	Kategorik
14.	STDs.vaginal.condylomatosis	Kategorik
15.	STDs.vulvo.perineal.condylomatosis	Kategorik
16.	STDs.syphilis	Kategorik
17.	STDs.pelvic.inflammatory.disease	Kategorik
18.	STDs.genital.herpes	Kategorik
19.	STDs.molluscum.contagiosum	Kategorik
20.	STDs.HIV	Kategorik
21.	STDs.Hepatitis.B	Kategorik
22.	STDs.HPV	Kategorik
23.	STDs..Number.of.diagnosis	Numerik
24.	Biopsy (Respon)	Kategorik

Analisis data diawali dengan melakukan penanganan pada data agar keseimbangan kelas terpenuhi dengan menggunakan metode *random oversampling* yaitu kelas minoritas akan ditingkatkan jumlah sampel kelasnya dan *random undersampling* yaitu kelas mayoritas akan diturunkan jumlah sampel kelasnya. Hasil keseimbangan kelas kemudian dilakukan teknik klasifikasi *naive bayes* dengan

menggunakan persamaan (1) jika variabel independen bersifat kategorik dan menggunakan persamaan (2) jika variabel independen bersifat numerik. Data akan dibagi menjadi dua bagian yaitu 80% data *training* yang berfungsi membuat model klasifikasi *naive bayes* dan 20% data *testing* yang berfungsi menguji kebaikan model yang dibuat. Performa model akan dilihat berdasarkan nilai output akurasi, sensitivitas, spesifisitas, presisi dan *Area Under Curve (AUC) ROC.*

Suyanto (2019) menyatakan terdapat beberapa ukuran yang digunakan untuk mengevaluasi performa model klasifikasi yaitu :

$$Akurasi = \frac{TP + TN}{P + N} \tag{3}$$

$$Sensitivitas = \frac{TP}{P} \tag{4}$$

$$Spesifisitas = \frac{TN}{N} \tag{5}$$

$$Presisi = \frac{TP}{TP + FP} \tag{6}$$

dengan *confusion matrix* seperti Tabel 2 berikut:

Tabel 2. *Confusion Matrix* untuk Evaluasi Model

	Prediksi : Yes	**Prediksi : No**	**Total**
Aktual :Yes	TP	FN	P
Aktual : No	FP	TN	N
Total	P’	N’	P+N

Lopez dkk (2013) menyatakan ukuran *Area Under Curve* dihitung sebagai daerah kurva ROC menggunakan persamaan berikut :

$$\mathrm{AUC} = \frac{1 + \mathrm{TP}_{\mathrm{rate}} - \mathrm{FP}_{\mathrm{rate}}}{2} \tag{7}$$

Gorunescu (2011) menjelaskan kriteria dasar yang digunakan untuk menyimpulkan hasil klasifikasi menggunakan nilai AUC adalah

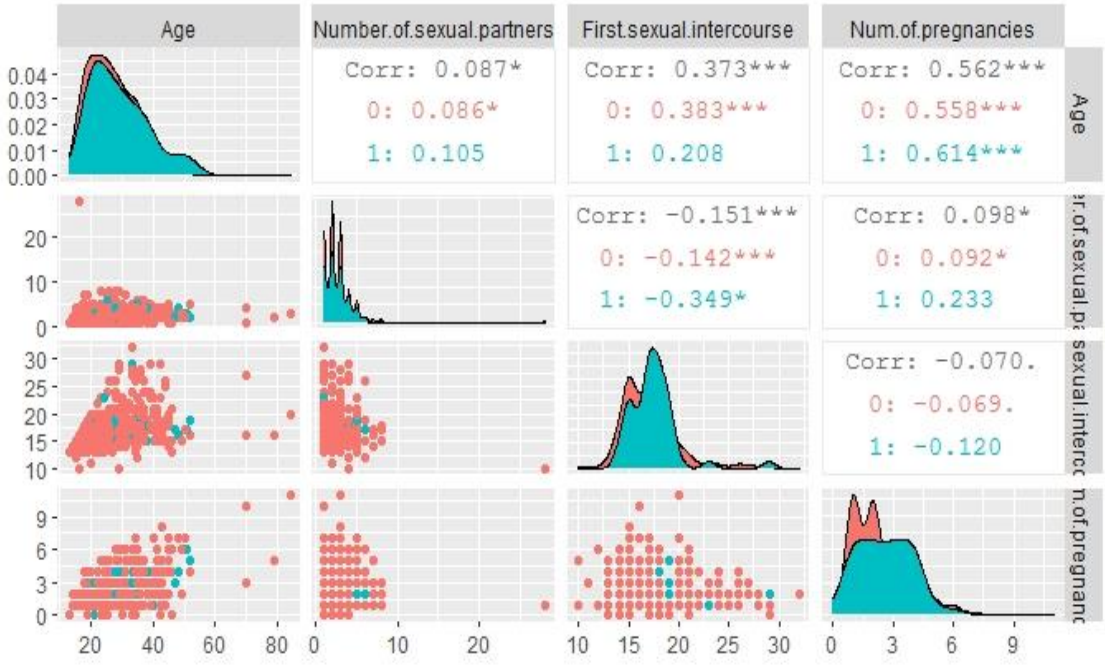
0.90 - 1.00 = Klasifikasi sempurna;
0.80 - 0.90 = Klasifikasi baik;
0.70 - 0.80 = Klasifikasi cukup baik;
0.60 - 0.70 = Klasifikasi buruk;
0.50 - 0.60 = Gagal.

Leibniz : Jurnal Matematika
Volume 2, Nomor 2, Halaman 11–22
Juli 2022
e-ISSN: 2775-2356

HASIL DAN PEMBAHASAN

Deskriptif Data

Data *Cervical Cancer Risk Factors* terdiri dari 23 variabel independen yaitu *Age, Number of sexual partners, First sexual intercourse, Num of pregnancies, Smokes, Smokes (packs/year),* dan lainnya sedangkan variabel target yaitu Biopsy dengan nilai 0 (tidak kanker serviks) dan 1 (kanker serviks). Kelas dengan status tidak kanker serviks sebanyak 623 sampel dan kelas dengan status kanker serviks sebanyak 45 sampel. Berikut deksriptif data secara visual dengan menghubungkan 4 variabel independen sebagai gambaran awal pada data.

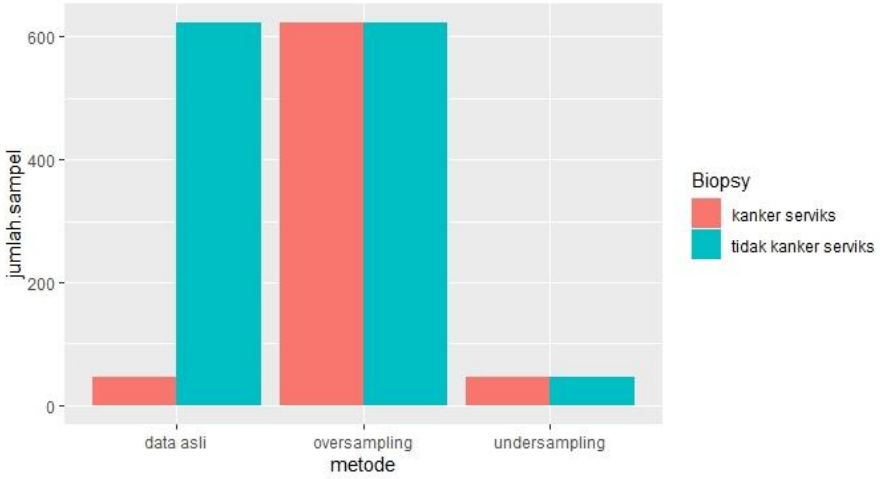


Gambar 2. Deskriptif Data

Berdasarkan hasil deskriptif secara visual di atas dengan menghubungkan beberapa variabel independen, pada grafik *scatter plot* yang menghubungkan variabel *age* dan *number of sexual partners*, *age* dan *first sexual intercourse*, *age* dan *num of pregnancies* dan lainnya secara visual kelas 0 (tidak kanker serviks) memiliki titik plot mayoritas dibandingkan dengan kelas 1 (kanker serviks). Nilai korelasi *age* dan *num of pregnancies* sebesar 0,562, nilai korelasi *age* dan *number of sexual partners* sebesar 0,087, nilai korelasi *num of pregnancies* dan *first sexual intercourse* memiliki nilai korelasi negatif sebesar -0,07.

Random Oversampling* dan *Random Undersampling

Ketidakseimbangan kelas diatasi menggunakan metode *random oversampling* dan *random undersampling*. *Random oversampling* akan meningkatkan kelas minoritas dengan cara menduplikasi sampel secara acak sehingga jumlah sampel pada kelas minoritas akan sama dengan kelas mayoritas. *Random undersampling* akan mengurangi kelas mayoritas dengan cara menghilangkan sampel secara acak sehingga jumlah sampel pada kelas mayoritas akan sama dengan kelas minoritas.



Gambar 3. Aktual Data, *Random Oversampling*, dan *Random Undersampling*

Berdasarkan Gambar 3 di atas, metode *random oversampling* meningkatkan kelas minoritas (kanker serviks) sehingga ukuran sampel pada kelas minoritas seimbang dengan kelas mayoritas (tidak kanker serviks) dan jumlah sampel menjadi 1246 sampel. Metode *random undersampling* mengurangi kelas mayoritas (tidak kanker serviks) sehingga ukuran sampel pada kelas mayoritas seimbang dengan kelas minoritas (kanker serviks) dan jumlah sampel menjadi 90 sampel.

Klasifikasi Naive Bayes

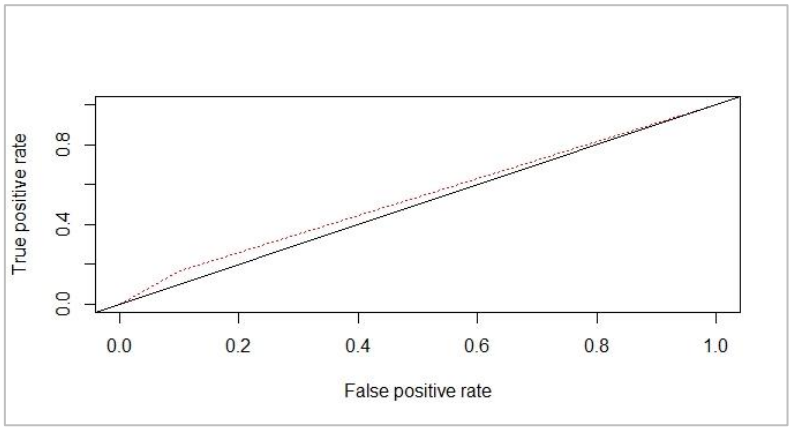
Berdasarkan *running* program *R-Studio* untuk data *cervical cancer risk factors* yang tidak seimbang (*imbalanced data*) maka model *naive bayes* yang dihasilkan data *training* berdasarkan *confusion matrix* pada data uji seperti pada Tabel 3 berikut ini.

Leibniz : Jurnal Matematika
Volume 2, Nomor 2, Halaman 11–22
Juli 2022
e-ISSN: 2775-2356

Tabel 3. *Confusion Matrix* untuk Evaluasi Model *Naive Bayes*

Biopsy		**Prediksi**		**Total**
		0	1	
Aktual	0	115	13	128
	1	5	1	6
Total		120	14	134

Tabel 3 menjelaskan bahwa kelas 0 (tidak kanker serviks) yang tepat diprediksi sebagai kelas 0 sebanyak 115 sampel dan terdapat misklasifikasi sebanyak 13 sampel. Kelas 1 (kanker serviks) yang tepat diprediksi sebagai kelas 1 sebanyak 1 sampel dan terdapat misklasifikasi 5 sampel lainnya. Berdasarkan persamaan (3), (4), (5) dan (6) maka nilai akurasi yang didapatkan sebesar 86,57%, sensitifitas 89,84%, spesifisitas 16,66% dan presisi 95,83%.



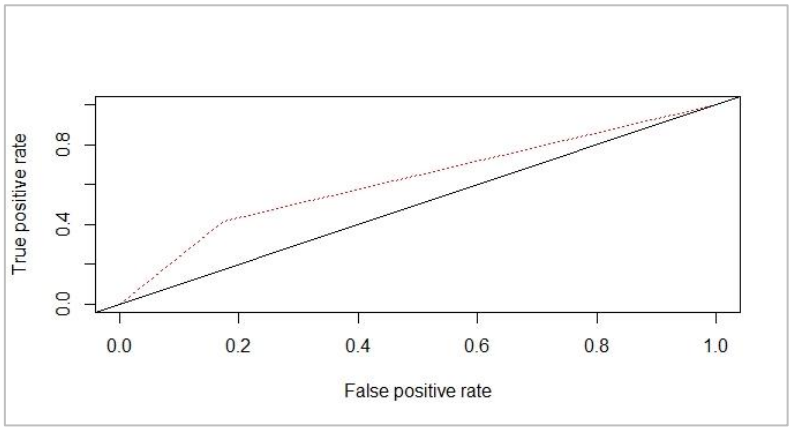
Gambar 4. *Area under Curve (AUC) ROC Naive Bayes*

Nilai *Area Under Curve* yang didapatkan berdasarkan persamaan (7) sebesar 0,5325 maka dapat disimpulkan bahwa hasil klasifikasi *naive bayes* pada data tidak seimbang (*imbalanced data*) adalah gagal. Kesimpulannya model *naive bayes* yang dihasilkan pada data tidak seimbang tidak dapat digunakan untuk melakukan prediksi klasifikasi pada data baru. Berdasarkan *running* program *R-Studio* untuk data *cervical cancer risk factors* yang seimbang (*balanced data*) dengan menggunakan algoritma *random oversampling* maka model *naive bayes* yang di hasilkan berdasarkan *confusion matrix* pada data uji seperti pada Tabel 4 berikut ini.

Tabel 4. *Confusion Matrix* untuk Evaluasi Model *ROS-Naive Bayes*

Biopsy		**Prediksi**		**Total**
		0	1	
Aktual	0	99	21	120
	1	76	54	130
Total		175	75	250

Tabel 4 menjelaskan bahwa kelas 0 (tidak kanker serviks) yang tepat diprediksi sebagai kelas 0 sebanyak 99 sampel dan terdapat misklasifikasi sebanyak 21 sampel. Kelas 1 (kanker serviks) yang tepat diprediksi sebagai kelas 1 sebanyak 54 sampel dan terdapat misklasifikasi 76 sampel lainnya. Berdasarkan persamaan (3), (4), (5) dan (6) maka nilai akurasi yang didapatkan sebesar 61,20%, sensitifitas 82,50%, spesifisitas 41,54% dan presisi 56,57%.



Gambar 5. *Area under Curve (AUC) ROC Random Oversampling Naive Bayes*

Nilai *Area Under Curve* yang didapatkan berdasarkan persamaan (7) sebesar 0,62 maka dapat disimpulkan bahwa hasil klasifikasi *naive bayes* dengan algoritma *random oversampling* adalah klasifikasi buruk. Kesimpulannya model *naive bayes* yang dihasilkan pada data seimbang dengan menggunakan algoritma *random oversampling* adalah buruk untuk digunakan melakukan prediksi klasifikasi pada data baru.

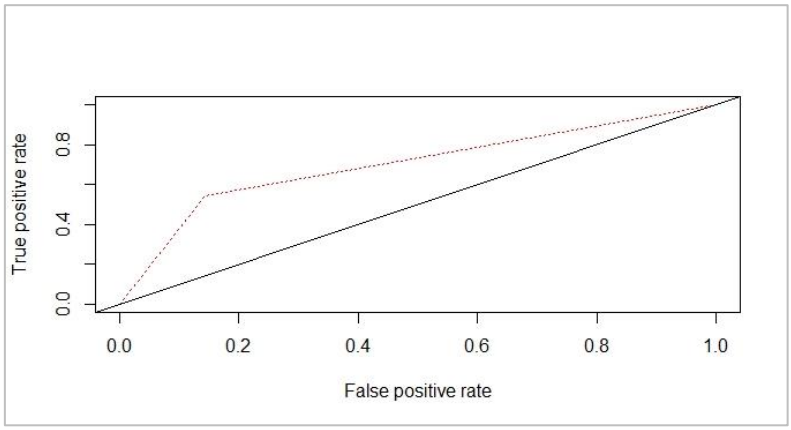
Berdasarkan *running* program *R-Studio* untuk data *cervical cancer risk factors* yang seimbang (*balanced data*) dengan menggunakan algoritma *random undersampling* maka model *naive bayes* yang di hasilkan berdasarkan *confusion matrix* pada data uji seperti pada Tabel 5 berikut ini.

Leibniz : Jurnal Matematika
Volume 2, Nomor 2, Halaman 11–22
Juli 2022
e-ISSN: 2775-2356

Tabel 5. *Confusion Matrix* untuk Evaluasi Model *RUS-Naive Bayes*

Biopsy		**Prediksi**		**Total**
		0	1	
Aktual	0	6	1	7
	1	5	6	11
Total		11	7	18

Tabel 5 menjelaskan bahwa kelas 0 (tidak kanker serviks) yang tepat diprediksi sebagai kelas 0 sebanyak 6 sampel dan terdapat misklasifikasi sebanyak 1 sampel. Kelas 1 (kanker serviks) yang tepat diprediksi sebagai kelas 1 sebanyak 6 sampel dan terdapat misklasifikasi 5 sampel lainnya. Berdasarkan persamaan (3), (4), (5) dan (6) maka nilai akurasi yang didapatkan sebesar 66,67%, sensitifitas 85,71%, spesifisitas 54,55% dan presisi 54,55%.



Gambar 6. *Area under Curve (AUC) ROC Random Undersampling Naive Bayes*

Nilai *Area Under Curve* yang didapatkan berdasarkan persamaan (7) sebesar 0,7013 maka dapat disimpulkan bahwa hasil klasifikasi *naive bayes* dengan algoritma *random undersampling* adalah klasifikasi cukup baik. Kesimpulannya model *naive bayes* yang dihasilkan pada data seimbang dengan menggunakan algoritma *random undersampling* adalah cukup baik untuk digunakan melakukan prediksi klasifikasi pada data baru.

SIMPULAN

Berdasarkan hasil pembahasan maka dapat disimpulkan algoritma *random oversampling* dapat meningkatkan kelas minoritas sehingga seimbang dengan kelas mayoritasnya dan jumlah sampel menjadi 1246 sampel sedangkan algoritma *random*

undersampling dapat mengurangi kelas mayoritasnya sehingga seimbang dengan kelas minoritasnya dan jumlah sampel menjadi 90 sampel. Metode klasifikasi dengan *naive bayes* pada data tidak seimbang (*imbalanced data*) menghasilkan nilai AUC sebesar 0,5325 yang berarti klasifikasi gagal. Metode klasifikasi *naive bayes* dengan menggunakan algoritma *random oversampling* menghasilkan nilai AUC sebesar 0,62 yang berarti klasifikasi buruk. Metode klasifikasi *naive bayes* dengan menggunakan algoritma *random undersampling* menghasilkan nilai AUC sebesar 0,7013 yang berarti klasifikasi cukup baik.

DAFTAR PUSTAKA

Dangeti, Pratap. (2017). *Statistics for Machine Learning*. Mumbai: Packt Publishing Ltd.

Fallo, S. I. (2021). *Support Vector Machine, Naive Bayes Classifier, dan Regresi Logistik Ordinal dalam Prediksi Cuaca* (Doctoral dissertation, Universitas Gadjah Mada).

Fernandes, Kelwin.dkk. (2018). *Supervised deep learning embeddings for the prediction of cervical cancer diagnosis*. Portugal. PeerJ Computer Science.

Gorunescu, Florin. (2011). *Data Mining: Concepts, Models, and Techniques*. Romania: Springer.

Lopez, Victoria.dkk. (2013). *An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics*. Spain. Elsevier Ltd.

Osorio, J. Hoyos.dkk. (2021). *Relevant information undersampling to support imbalanced data Classification*. Colombia. Elsevier Ltd.

Rodríguez, Néstor.dkk. (2021). *SOUL: Scala Oversampling and Undersampling Library for imbalance classification*. Spain. Elsevier Ltd.

Suyanto. (2019). *Data Mining untuk Klasifikasi dan Klasterisasi Data*. Bandung: Penerbit Informatika.

Thabtah, Fadi.dkk. (2019). *Data imbalance in classification : Experimental evaluation*. New Zealand. Elsevier Ltd.

Zheng, Wanwan & Jin, Mingzhe Jin. (2020). *The Effects of Class Imbalance and Training Data Size on Classifier Learning*: An Empirical Study Springer Nature Singapore Pte Ltd