JURNAL RESISTOR | ISSN 2598-7542 | E-ISSN 2598-9650
Vol. 3 No 1 – April 2020 | https://s.id/jurnalresistor
DOI: https://doi.org/10.31598
Publishing: Prahasta Publisher

Kombinasi *Oversampling* dan *Undersampling* dalam Menangani *Class Imbalanced* dan *Overlapping* pada Klasifikasi *Data Bank Marketing*

Anak Agung Gde Wahyu Sukma Erlangga[1], I Gede Aris Gunadi[2], I Made Gede Sunarya[3]

[1,2,3]Program Studi Ilmu Komputer, Program Pascasarjana, Universitas Pendidikan Ganesha
Jl. Udayana No.11, Banjar Tegal, Singaraja, Kabupaten Buleleng, Bali 81116, Indonesia

e-mail: anak.agung.gde@student.undiksha.ac.id[1], igedearisgunadi@undiksha.ac.id[2], sunarya@undiksha.ac.id[3]

Received : Januari, 2024	Accepted : April, 2024	Published : April, 2024

Abstract

Class imbalance can occur in various types of datasets, one of which is bank marketing datasets. The class imbalance can cause classification problems. To handle the problem, the SMOTE method can be used. However, the application of SMOTE can cause class overlapping which can also interfere with classification performance. Therefore, this research tries to handle it by combining the SMOTE method with undersampling methods consisting of ENN, NCL, and TomekLink. The classification algorithm used is Logistic Regression and the performance evaluation uses sensitivity, specificity and g-means of the model. The results of this study show that the resampling technique can improve the balance of model classification as shown by the g-means value in each dataset, namely, bank marketing (88.47%), credit card fraud (93%) and cerebral stroke (77.03%). The SMOTE-ENN combination provides the best performance in increasing the sensitivity of the model, which is shown by the bank marketing (94.05%) and cerebral stroke (80.1%) datasets. Meanwhile, the highest specificity is generated by the original datasets, respectively bank marketing (97.64%), credit card fraud (99.98%), cerebral stroke (100%).

Keywords: *bank marketing; class imbalance; class overlapping; oversampling; undersampling*

Abstrak

Ketidakseimbang kelas dapat terjadi pada berbagai jenis *datasets*, salah satunya pada *datasets bank marketing*. Ketidakseimbangan kelas itu dapat menyebabkan masalah klasifikasi. Untuk menangani masalah itu dapat digunakan metode SMOTE. Namun, penerapan SMOTE dapat menimbulkan *class overlapping* yang juga dapat mengganggu performa klasifikasi. Oleh sebab itu, penelitian ini mencoba menanganinya dengan mengkombinasikan metode SMOTE dengan metode *undersampling* yang terdiri dari ENN, NCL, dan TomekLink. Untuk algoritma klasifikasi yang digunakan adalah *Logistic Regression* dan evaluasi performa menggunakan *sensitivity*, *specificity* dan *g-means* dari model. Hasil penelitian ini menunjukan teknik *resampling* dapat meningkatkan keseimbangan klasifikasi model yang ditunjukan nilai *g-means* pada setiap *datasets* yaitu, *bank marketing* (88,47%), *credit card fraud* (93%) dan *cerebral stroke* (77,03%). Kombinasi SMOTE-ENN memberikan kinerja terbaik dalam meningkatkan *sensitivity* dari model, yang ditunjukan oleh *datasets bank marketing* (94,05%) dan *cerebral stroke* (80,1%). Sementara itu, *specificity* tertinggi dihasilkan *datasets original* berturut-turut *bank marketing* (97,64%), *credit card fraud* (99,98%), *cerebral stroke* (100%).

Kata Kunci: *bank marketing; class imbalance; class overlapping; oversampling; undersampling*

1. PENDAHULUAN

Berbagai layanan keuangan biasanya disediakan oleh bank, seperti pengiriman dan penerimaan uang, penyimpanan uang, deposito dan sebagainya. Untuk dapat mempromosikan layanan tersebut maka bank biasanya akan menerapkan suatu teknik pemasaran untuk mendapatkan nasabah. Salah satu teknik yang digunakan adalah *telemarketing* [1].

Telemarketing merujuk pada penggunaan telepon dan pusat panggilan untuk berinteraksi dengan calon pelanggan, melakukan penjualan kepada pelanggan, serta menyediakan layanan seperti pemrosesan pesanan dan menjawab pertanyaan [2]. Perusahaan menggunakan *Telemarketing* dengan tujuan untuk mencari peluang dan calon klien bagi produk atau layanan tertentu, dimana dengan menggunakan komunikasi jarak jauh, perusahaan atau organisasi akan memperoleh keuntungan lebih melalui pemasaran yang efektif. Untuk dapat mengetahui keberhasilan dari telemarketing dapat digunakan suatu teknik *machine learning*. *Machine Learning* (ML) adalah bagian dari kecerdasan buatan (AI), dimana tanpa diprogram secara eksplisit, sistem dapat secara otomatis belajar dan meningkatkan kinerja dari pengalaman mereka sendiri. [3]. *Datasets* yang dapat digunakan sebagai data latih adalah *datasets bank marketing* yang diambil dari *website* UCI Machine Learning yaitu, *dataset* kampanye pemasaran langsung (*telemarketing*) yang berasal dari lembaga perbankan Portugis. Namun, terdapat ketidakseimbangan kelas atau *class imbalance* pada dataset tersebut, dimana terdapat kelas yang jauh lebih dominan secara signifikan dari kelas lainnya.

Ketidakseimbangan kelas atau *class imbalanced* yang terjadi pada data tersebut adalah jumlah kelas yang tidak berlangganan deposito jauh lebih banyak dibandingkan dengan kelas yang berlanggan deposito. Kelas yang memiliki distribusi persentase lebih kecil akan disebut dengan kelas minoritas sebaliknya kelas yang memiliki persentase lebih besar akan disebut dengan kelas mayoritas [4]. Ketidakseimbangan data antara kelas akan mengakibatkan ketergantungan pada kelas mayoritas dalam proses pengklasifikasian [5]. Ketergantungan itu akan menyebabkan model *machine learning* baik dalam memprediksi kelas mayoritas tetapi sulit untuk memprediksi kelas minoritas [6], [7].

Berdasarkan permasalahan itu, maka ketidakseimbangan kelas memerlukan suatu penanganan sehingga model yang dihasilkan dapat bekerja dengan baik dan lebih andal dalam melakukan prediksi terhadap data yang diberikan. Teknik yang dapat digunakan untuk menangani masalah tersebut diantaranya adalah teknik *oversampling*, *undersampling* dan kombinasi dari kedua teknik tersebut. Teknik itu bekerja dengan meningkatkan jumlah sampel pada kelas minoritas (*oversampling*) dan menurunkan jumlah sampel pada kelas mayoritas (*undersampling*). Melalui cara ini, jumlah sampel pada kedua kelas menjadi lebih seimbang dan model dapat mempelajari pola dari kedua kelas secara merata dan akurat.

Penerapan metode *oversampling* yang terlalu banyak dapat menyebabkan terjadi *overfitting* [8] sedangkan penerapan *undersampling* yang berlebihan dapat menyebabkan hilangnya informasi penting dari *datasets* [9], [10]. *Synthetic Minority Oversampling Technique* (SMOTE) adalah salah satu teknik *oversampling* yang dapat mengurangi *overfitting* [11] jika diterapkan karena dalam melakukan *oversampling* SMOTE akan membangkitkan data sintetis. Hal itu dapat dilihat dari penelitian oleh [12] dimana semua matriks performa menunjukan *score* 1-2% lebih tinggi ketika menggunakan SMOTE saat memprediksi *lumpy skin disease*. Namun, SMOTE memiliki suatu kelemahan dimana data sintetis yang dihasilkan dapat menyebabkan terjadinya *class overlapping* atau tumpang tindih kelas berdasarkan penelitian [13], dimana *class overlapping* juga dapat memengaruhi performa dari model yang dihasilkan.

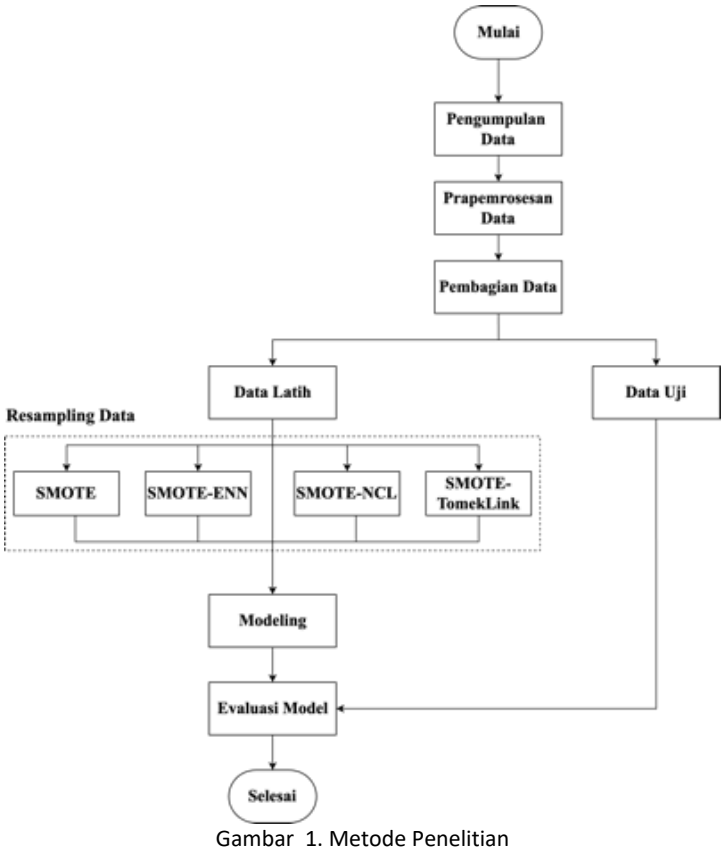
Oleh karena itu, demi menangani permasalahan itu maka peneliti mengusulkan untuk mengkombinasikan teknik SMOTE dengan teknik *undersampling*. SMOTE akan melakukan *oversampling* terhadap kelas minoritas dan kemudian teknik *undersampling* akan melakukan *undersampling* terhadap kelas mayoritas. Penerapan teknik *undersampling* ini diharapkan mampu membuat data yang dihasilkan lebih bersih dan terhindar dari data *noise* dan tumpang tindih kelas atau *class overlapping* yang diakibatkan oleh data sintetis yang dihasilkan dari metode SMOTE. Teknik *undersampling* yang akan coba dikombinasikan dengan teknik SMOTE pada penelitian ini diantaranya, *Edited Nearest Neighbor* (ENN),

Neighborhood Cleaning Rule (NCL) dan TomekLink. Sementara itu, untuk mengukur kinerja dari model yang dilatih menggunakan *confusion matrix* sehingga dapat diketahui *sensitivity*, *specificity*, dan *g-means* dari model yang dilatih. Selain itu, penelitian ini juga menggunakan 2 *datasets* berbeda untuk menguji kembali metode yang digunakan yaitu, *datasets credit card fraud* dan *datasets cerebral stroke* yang diambil dari *website* Kaggle, dimana kedua *datasets* tersebut merupakan *datasets* yang memiliki karakteristik yang mirip dengan *datasets bank marketin*g yaitu, ketidakseimbangan pada kelasnya.

Diharapkan dengan adanya penelitian ini dapat diketahui perbandingan performa dari model yang dilatih menggunakan data original, data yang di-*resampling* dengan SMOTE serta yang dilatih dengan data *resampling* kombinasi yaitu dengan SMOTE-ENN, SMOTE-NCL, dan SMOTE-TomekLink. Peneliti juga berharap penelitian ini akan memberikan informasi yang berguna untuk membangun model klasifikasi pada dataset yang mengalami ketidakseimbangan kelas. Selain itu, penelitian ini dapat digunakan sebagai referensi bagi peneliti lain yang tertarik dengan topik yang serupa.

2. METODE PENELITIAN

Gambar 1 merupakan skema dari metode penelitian yang dilakukan. Sementara itu, untuk tools yang digunakan adalah jupyter notebook, degan bantuan *library sckit-learn*, *pandas*, *numpy* dan *imblearn*.



Gambar 1. Metode Penelitian

2.1 Pengumpulan Data

Ini merupakan proses pengumpulan *datasets*, dimana *datasets* pada penelitian ini diambil dari *website* UCI Machine Learning, yaitu *datasets* kampanye pemasaran langsung (*telemarketing*) yang berasal dari lembaga perbankan Portugis. Data ini berjumlah 41188 buah data dengan 20 fitur dan sebuah variabel target. Tabel 1 menunjukan informasi dari setiap atribut yang dimiliki oleh *datasets*.

Table 1: Atribut *Datasets Bank Marketing*

Atribut	Informasi
age	Umur dari nasabah
job	Jenis pekerjaan dari nasabah
marital	Status perkawinan dari nasabah
education	Pendidikan terkahir dari nasabah
default	Memiliki kredit atau tidak
housing	Memiliki kredit perumahan atau tidak
loan	Mempunyai pinjaman pribadi atau tidak
contact	Jenis komunikasi yang digunakan
month	Bulan kontak terakhir tahun ini
day_of_week	Hari kontak terkahir minggu ini
duration	Durasi kontak
campaign	Jumlah kontak yang telah dilakukan selama kampanye ini terhadap nasabah terkait.
pdays	Jumlah hari yang berlalu setelah klien terakhir dihubungi dari kampanye sebelumnya
previous	Jumlah kontak yang dilakukan sebelum kampanye ini untuk klien terkait
poutcome	Hasil dari kampanye pemasaran sebelumnya
emp.var.rate	Tingkat variasi pekerjaan - indikator triwulanan
cons.price.idx	Indeks harga konsumen - indikator bulanan
cons.conf.idx	Indeks kepercayaan konsumen - indikator bulanan
euribor3m	Tingkat triwulan *Euribor* - indikator harian
nr.employed	Jumlah karyawan - indikator triwulanan

2.2 Prapemrosesan Data

Pada proses ini *datasets* yang telah diambil sebelumnya akan diolah sebelum dilakukan proses selanjutnya. Pada penelitian dilakukan beberapa tahap prapemrosesan diantarnya, penanganan nalai yang hilang, *encoding* fitur kategorikal, pemisahan variabel fitur dan target, serta standarisasi agar data memiliki rentang nilai yang sama.

2.3 Pembagian Data

Setelah tahap prapemrosesan selesai, data akan dibagi menjadi data uji dan data latih. Data latih digunakan untuk melatih model, sedangkan data uji digunakan untuk menguji model yang telah dilatih untuk mengetahui seberapa baik mereka berfungsi. Pada penelitian ini data latih dan data uji akan dibagi dengan proporsi 75% data latih dan 25% data uji menggunakan t*rain_test_split* dari *scikit-learn*.

2.4 *Resampling* Data

Pada tahap ini akan dilakukan *resampling* pada data fitur yang akan dijadikan data latih dengan tujuan untuk membuat distribusi kelas dari data menjadi seimbang. Metode yang akan digunakan adalah metode *oversampling* SMOTE, dimana metode ini menyeimbangkan kelas dengan meng-*generate* data sintetis dengan cara memilih data acak dari kelas minoritas dan di antara titik terpilih dan tetangganya akan dihasilkan data sintetik menggunakan metode interopolasi [8], [14]. Rumus (1) merupakan rumus yang digunakan untuk menghasilkan data sintetis tersebut [14].

$$\mathrm{x_{syn}} = \mathrm{x_i} + \mathrm{rand}(0,1) \times (\mathrm{x_{knn}} - \mathrm{x_i}) \qquad (1)$$

Keterangan:

x_{syn} : Data sintetik hasil replikasi.
x_i : Data yang akan direplikasi.
$rand(0,1)$: Bilangan random diantara 0 dan 1.
x_{knn} : Data terdekat dengan data yang akan direplikasi yang dihitung dengan *euclidian distance.*

Selain itu, metode SMOTE juga akan dikombinasikan dengan teknik *undersampling* diantaranya adalah ENN yang merupakan suatu metode *undersampling* pada kelas mayoritas yang dilakukan melalui penghapusan sampel yang memiliki label kelas yang berbeda dengan mayoritas dari tetangga terdekatnya sebanyak k [10], [15], [16]. Selanjutnya, teknik NCL yang

mengkombinasikan CNN dan ENN, dimana Teknik ENN untuk mengidentifikasi seluruh sampel yang termasuk dalam kelas mayoritas. Kemudian, CNN satu langkah digunakan untuk menghapus sampel pada kelas mayoritas yang jumlahnya lebih dari setengah jumlah kelas minoritas setelah proses sebelumnya dilakukan [17]. Terakhir, SMOTE akan dikombinasikan dengan teknik TomekLink yang menggunakan aturan tetangga terdekat untuk memilih *instance* dan menghilangkan *instance* kelas mayoritas yang lebih dekat dengan kelas minoritas [18].

2.5 *Modeling*

Pada tahap ini akan dilakukan *modeling* dimana model akan dilatih menggunakan data latih *original*, data hasil *resampling* dengan SMOTE serta data hasil resampling dengan teknik kombinasi yaitu SMOTE-ENN, SMOTE-NCL, dan SMOTE-TomekLink. Untuk metode atau algoritma *machine learning* yang digunakan untuk membuat model adalah *Logistic Regression*. Metode ini digunakan untuk menunjukkan hubungan antara variabel respons dan sekumpulan variabel prediktor, dalam hal ini, variabel respons adalah variabel biner atau dikotomus [19], [20]. Oleh sebab itu, biasanya algoritma ini digunakan untuk melakukan klasifikasi biner atau *binary classification*, dimana algoritma ini berguna untuk memperkirakan probabilitas posterior dari setiap kelas [21]. Rumus (2) merupakan persamaan untuk *Logistic Regression* multivariat dengan k variabel prediktor [19].

$$\pi(\mathrm{x}) = \frac{\exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}{1+\exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)} \quad (2)$$

Keterangan:

$\pi(x)$: Probabilitas bahwa variabel target atau biasa disebut Y sama dengan 1, yaitu probabilitas positif atau hasil regresi logistic yang ingin diprediksi.

exp : Fungsi exponen atau kebalikan dari logaritma natural.

β_0 : *Intercept* atau bias.

$\beta_1, \beta_2, \dots \beta_k$: Koefisien regresi regresi yang mengukur pengaruh masing-masing variabel prediktor $(x_1, x_2, \dots x_k)$ terhadap variabel target (Y).

$x_1, x_2, \dots x_k$: Variable prediktor atau fitur yang digunakan dalam regresi logistic.

2.6 Evaluasi Model

Setelah model selesai dilatih dengan menggunakan data latih maka akan dilakukan pengujian performansi dari model untuk mengetahui seberapa baik performa model dalam memprediksi data. Pada penelitian ini matriks evaluasi yang digunakan adalah *confusion matrix*. Berdasarkan nilai pada *confusion matrix* maka dapat dihitung matriks evaluasi yaitu *sensitivity* untuk menghitung kinerja model dalam memprediksi kelas positif; *specificity* untuk menghitung kinerja model dalam kelas memprediksi kelas negatif; *g-means* untuk menghitung seberapa seimbang kinerja model dalam memprediksi kedua kelas. Untuk rumus dari *sensitivity*, *specificity*, dan *g-means* dapat dilihat berturut-turut pada rumus (3), (4), dan (5).

$$\text{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \quad (3)$$

$$\text{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \quad (4)$$

$$\mathrm{G-Means} = \sqrt{\text{sensitivity} \times \text{specificity}} \quad (5)$$

Keterangan:

True Positive (TP) : Data bernilai positif yang diprediksi positif.

False Positive (FP) : Data bernilai negatif yang diprediksi positif.

True Negative (TN) : Data bernilai negatif yang diprediksi negatif.

False Negative (FN) : Data bernilai positif yang diprediksi negatif.

3. HASIL DAN PEMBAHASAN

Penelitian ini dimulai dari pengumpulan data, dimana untuk *datasets bank marketing* diambil dari *website* UCI Machine Learning, sedangkan dua *datasets* lainnya yang akan digunakan untuk menguji kembali metode yang digunakan diambil dari *website* Kaggle. Adapun *datasets* yang dikumpulkan adalah data yang memiliki ketidakseimbangan kelas yang dapat dilihat pada Tabel 2.

Tabel 1: Keterangan Setiap *Datasets*

Datasets	**Jumlah Data**	**Jumlah Fitur & Target**
Bank Marketing	41188	21
Credit Card Fraud	284807	31
Cerebral Stroke	43400	12

Semua *datasets* yang berhasil dikumpulkan akan melewati proses prapemrosesan sehingga nantinya data dapat digunakan untuk melakukan pengujian metode. Proses prapemrosesan yang dilakukan terdiri dari, penanganan data yang hilang, *encoding* pada fitur kategorikal, memisahkan variabel fitur dan target, serta standarisasi untuk menyamakan rentang nilai dari data. Setelah proses prapemrosesan selesai dilakukan akan dilanjutkan dengan pembagian data, dengan proporsi 75% untuk data latih dan 25% sebagai data uji yang merupakan pengaturan bawaan dari *train_test_split()* miliki *library scikit-learn*. Untuk hasil pembagian data dapat dilihat pada Tabel 3.

Tabel 2: Hasil Pembagian *Datasets*

Datasets	**Data Latih**	**Data Uji**
Bank Marketing	30891	10297
Credit Card Fraud	213605	71202
Cerebral Stroke	32541	10848

Setelah data dibagi menjadi data latih dan data uji maka akan dilakukan proses *resampling* untuk menyeimbangkan kelas dari setiap *datasets*, dimana metode yang akan digunakan diantaranya SMOTE, SMOTE-ENN, SMOTE-NCL, dan SMOTE-TomekLink. Untuk distribusi kelas sebelum dan sesudah dilakukan *resampling* pada data latih dapat dilihat pada Tabel 4 dan Tabel 5.

Tabel 3: Perbandingan Jumlah Kelas Sebelum *Resampling*

Datasets	**Positif**	**Negatif**
Bank Marketing	3480	27411
Credit Card Fraud	369	213236
Cerebral Stroke	578	31954

Datasets bank marketing memiliki 3480 kelas positif dan 27411 kelas negatif. Untuk *datasets credit card fraud* memiliki 369 kelas positif dan 213236 kelas negatif, dan terakhir *datasets cerebral stroke* memiliki 578 kelas positif dan 31954 kelas negatif. Hal itu menunjukan bahwa memang terjadi ketidakseimbangan kelas yang terjadi pada ketiga *datasets* tersebut dengan kelas negatif yang menjadi kelas mayoritas, sedangkan kelas positif adalah kelas minoritas.

Tabel 4: Perbandingan Jumlah Kelas Setelah *Resampling*

Datasets	**Teknik Resampling**	**Positif**	**Negatif**
Bank Marketing	SMOTE	27411	27411
	SMOTE-ENN	26390	23019
	SMOTE-NCL	27323	27411
	SMOTE-TomekLink	27340	27340
Credit Card Fraud	SMOTE	213236	213236

Datasets	Teknik Resampling	Positif	Negatif
	SMOTE-ENN	213236	212843
	SMOTE-NCL	213236	213236
	SMOTE-TomekLink	213236	213236
Cerebral Stroke	SMOTE	31954	31954
	SMOTE-ENN	30559	28102
	SMOTE-NCL	31732	31954
	SMOTE-TomekLink	31748	31748

Hasil *resampling* pada Tabel 5 menunjukan bahwa teknik-teknik *resampling* itu dapat menyeimbangkan distribusi kelas dari setiap *datasets*. Terlihat bahwa SMOTE dan SMOTE-TomekLink menghasilkan distribusi kelas yang sama. Sementara itu, untuk SMOTE-ENN dan SMOTE-NCL menghasilkan distribusi yang memiliki sedikit perbedaan terutama pada *datasets bank marketing* dan *cerebral stroke*. Hal itu terjadi karena teknik *undersampling* tersebut melakukan proses pembersihan data setelah pengaplikasian teknik SMOTE dengan tujuan agar menghapus *data overlapping* yang mungkin dihasilkan oleh SMOTE.

Setelah proses *resampling* dilakukan datasets siap untuk dilatih menggunakan algoritma Logistic Regression. Untuk mengevaluasi kinerja klasifikasi yang dihasilkan akan digunakan *confusion matrix*, dimana matriks yang dihitung adalah *sensitivity*, *specificity* dan *g-means*. Untuk perbandingan performansi klasifikasi berturut turut dapat dilihat pada Tabel 6, Tabel 7, dan Tabel 8.

Tabel 5: Hasil Evaluasi *Sensitivity*

Datasets	Teknik Resampling	Nilai
Bank Marketing	Original	42,59%
	SMOTE	90,09%
	SMOTE-ENN	94,05%
	SMOTE-NCL	89,83%
	SMOTE-TomekLink	90,17%
Credit Card Fraud	Original	61,79%
	SMOTE	88,62%
	SMOTE-ENN	88,62%
	SMOTE-NCL	88,62%
	SMOTE-TomekLink	88,62%
Cerebral Stroke	Original	0,00%
	SMOTE	78,57%
	SMOTE-ENN	80,10%
	SMOTE-NCL	78,06%
	SMOTE-TomekLink	78,06%

Berdasarkan Tabel 6 dapat dilihat bahwa *sensitivity* untuk data original sangat rendah, bahkan pada *datasets cerebral stroke* berada pada angka 0%, dimana itu menunjukan bahwa model sama sekali tidak bisa mengklasifikasi data dengan kelas positif. Untuk *sensitivity* tertinggi dimiliki oleh SMOTE-ENN, dengan nilai 94,05% pada *datasets bank marketing* dan 80,1% pada *datasets cerebral stroke*. Lalu, untuk *datasets credit card fraud* memiliki hasil yang sama yaitu sebesar 88,62% untuk setiap teknik yang diujikan. Hal itu juga ditunjukan dari penelitian sebelumnya, dimana SMOTE-ENN meningkatkan rata-rata kinerja model klasifikasi sebesar 24,01% [11].

Tabel 6. Hasil Evaluasi *Specificity*

Datasets	**Teknik Resampling**	**Nilai**
Bank Marketing	Original	97,64%
	SMOTE	86,53%
	SMOTE-ENN	83,22%
	SMOTE-NCL	86,61%
	SMOTE-TomekLink	86,53%
Credit Card Fraud	Original	99,98%
	SMOTE	97,59%
	SMOTE-ENN	97,47%
	SMOTE-NCL	97,59%
	SMOTE-TomekLink	97,58%
Cerebral Stroke	Original	100,00%
	SMOTE	75,53%
	SMOTE-ENN	73,17%
	SMOTE-NCL	75,66%
	SMOTE-TomekLink	75,62%

Specificity adalah suatu matriks evaluasi yang dapat menunjukan seberapa baik suatu model dapat memprediksi kelas negatif dari suatu data. Tabel 7 menunjukan bahwa setiap *datasets* original memiliki performa *specificity* yang begitu tinggi jika dibandingkan dengan teknik *resampling*, bahkan *datasets cerebral stroke* memiliki *specificity* sempurna yaitu 100%. Sebelumnya, itu disebabkan oleh suatu ketidakseimbangan kelas antara kelas positif (kelas minoritas) dan negatif (kelas mayoritas), sehingga model klasifikasi memprioritaskan klasifikasi terhadap kelas negatif (kelas mayoritas) bukan kelas positif (kelas minoritas). Hasil serupa juga ditemukan pada penelitian [7], dimana dari 5 *datasets* yang diujikan *specificity* tertinggi dihasilkan oleh *datasets original*.

Tabel 7. Hasil Evaluasi *G-Means*

Datasets	**Teknik Resampling**	**Nilai**
Bank Marketing	Original	64,48%
	SMOTE	88,29%
	SMOTE-ENN	88,47%
	SMOTE-NCL	88,21%
	SMOTE-TomekLink	88,33%
Credit Card Fraud	Original	78,60%
	SMOTE	93,00%
	SMOTE-ENN	92,94%
	SMOTE-NCL	93,00%
	SMOTE-TomekLink	92,99%
Cerebral Stroke	Original	0,00%
	SMOTE	77,03%
	SMOTE-ENN	76,56%
	SMOTE-NCL	76,85%
	SMOTE-TomekLink	76,83%

Tabel 8 adalah hasil evaluasi menggunakan *g-means*. Konsep dasar dari *g-means* adalah untuk memaksimalkan akurasi setiap kelas dengan menjaga keseimbangan di antara keduanya [14]. Oleh sebab itu, nilai *g-means* yang tinggi dapat mengindikasikan bahwa kedua kelas dalam data yang tidak seimbang memiliki performa yang baik [22]. Terlihat *datasets* original tidak memiliki *g-means* yang begitu baik dan performa klasifikasi yang tidak seimbang antar

kelas, dengan performa terendah ada pada datasets cerebral stroke yaitu 0%. Namun, setelah penerapan *resampling* terdapat peningkatkan yang signifikan terhadap nilai *g-means,* dimana pada *datasets bank marketing* SMOTE-ENN memiliki performa terbaik yaitu 88,47%. Lalu, pada *datasets cerebral stroke* SMOTE-NCL menjadi yang terbaik dengan 76,85% dan pada *datasets credit card fraud* performa terbaik dimiliki oleh SMOTE dan SMOTE-NCL walaupun sebenarnya tidak ada perbedaan nilai *g-means* yang begitu signifikan yang dihasilkan diantara semua teknik *resampling* yang digunakan. Penelitian [7] juga menunjukan hal serupa, dimana teknik *resampling* meningkatkan nilai *g-means* dari model, dengan nilai rata-rata tertinggi (89,36%) yang dihasilkan oleh teknik ADASYN yang dikombinasikan dengan algoritma klasifikasi *random forest*.

Berdasarkan evaluasi yang telah dilakukan dengan menggunakan *sensitivity*, *specificity*, dan *g-means* dapat dilihat bahwa penggunaan teknik *resampling* dapat meningkatkan kinerja dari model dalam menanganani ketidakseimbangan kelas yang terjadi pada *datasets*. Penggunaan kombinasi SMOTE-ENN menunjukan hasil terbaik dalam meningkatkan *sensitivity* pada *datasets bank marketing* dan *cerebral stroke*. Hal itu bisa menjadi indikasi bahwa teknik *undersampling* ENN mampu membersihkan data yang mungkin merupakan *data overlapping* dari teknik SMOTE pada kedua *datasets*. Hal itu juga ditunjukan melalui distribusi kelas positif dan negatif hasil SMOTE-ENN yang memiliki sedikit perbedaan atau tidak benar-benar 50%:50%, sehingga terjadi peningkatkan performa.

Namun, hal yang sedikit berbeda terjadi pada *datasets credit card fraud*, dimana setiap teknik *resampling* menghasilkan kinerja yang sama. Itu bisa mengindikasikan bahwa penerapan SMOTE pada *datasets* tersebut tidak menghasilkan terlalu banyak *data overlapping*. Berdasarkan Tabel 5 hanya teknik SMOTE-ENN yang melakukan penghapusan data dan jumlah yang dihapus pun tidak terlalu signifikan, sehingga performa yang dihasilkan sama. Terakhir, untuk *g-means* terlihat performansi yang dihasilkan bervariasi dimana tidak ada teknik yang begitu dominan, tetapi penerapan teknik *resampling* dapat dikatakan sudah mampu meningkatkan keseimbangan model dalam melakukan klasifikasi terhadap kedua kelas baik kelas positif dan kelas negatif.

4. KESIMPULAN

Hasil penelitian ini menunjukan bahwa teknik *resampling* pada *imbalance datasets* dapat memberikan dampak positif signifikan terhadap kinerja model klasifikasi dalam memprediksi baik kelas positif dan negatif. Itu ditunjukan dengan *g-means* pada setiap *datasets* yaitu, *bank marketing* (88,47%), *credit card fraud* (93%) dan cerebral stroke (77,03%). Kombinasi SMOTE-ENN memberikan kinerja terbaik dalam meningkatkan *sensitivity* dari model, yang ditunjukan oleh *datasets bank marketing* (94,05%) dan *cerebral stroke* (80,1%). Hal itu dapat menjadi indikasi bahwa teknik ENN dapat membersihkan *data overlapping* yang dapat dihasilkan teknik SMOTE.

Namun, pada *datasets credit card fraud* hanya terjadi sedikit pembersihan data yang mengindikasikan bahwa penerapan SMOTE pada *datasets* itu tidak menghasilkan banyak *data overlapping*. Oleh sebab itu, kinerja yang dihasilkan hampir sama antara teknik satu dengan lainnya dengan kisaran *sensitivity* (88,62%), *specificity* (97,59%) dan *g-means* (93%). Sementara itu, *specificity* tertinggi dihasilkan *datasets original* berturut-turut *bank marketing* (97,64%), *credit card fraud* (99,98%), *cerebral stroke* (100%) yang disebabkan oleh model yang masih memprioritaskan kelas negatif yang jauh lebih banyak. Dengan demikian, dapat dikatakan bahwa pemilihan teknik *resampling* harus disesuaikan dengan karakteristik *datasets* yang akan ditangani dan matriks evaluasi mana yang ingin diprioritaskan.

5. SARAN

Hasil dari penelitian ini dapat digunakan sebagai referensi pada penelitian selanjutnya, dimana dapat dilakukan seleksi fitur pada tahap prapemrosesan. Selain itu dapat dicoba kombinasi teknik *resampling* lainnya ataupun algoritma *machine learning* yang lebih canggih seperti *ensamble learning*.

DAFTAR PUSTAKA

[1] R. Sulaehani, “Prediksi Keputusan Klien Telemarketing Untuk Deposito Pada Bank Menggunakan Algoritma Naive Bayes Berbasis Backward Elimination,”

ILKOM Jurnal Ilmiah, vol. 8, no. 3, pp. 182–189, 2016.

[2] P. Kotler and K. L. Keller, *Manajemen Pemasaran*. Jakarta: PT. Indeks, 2016.

[3] I. Tahyudin, *Pengenalan Machine Learning Menggunakan Jupyter Notebook*. in Mechine Learning. Zahira Media Publisher, 2020. [Online]. Available: https://books.google.co.id/books?id=_uMREAAAQBAJ

[4] W. Ustyannie and S. Suprapto, "Oversampling Method to Handling Imbalanced Datasets Problem in Binary Logistic Regression Algorithm," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 14, no. 1, p. 1, Jan. 2020, doi: 10.22146/ijccs.37415.

[5] S. Mutmainah, "Penanganan Imabalance Data pada Klasifikasi Kemungkinan Penyakit Stroke," Yogyakarta, 2021. [Online]. Available: https://library.uii.ac.id/osr

[6] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Information (Switzerland)*, vol. 14, no. 1, Jan. 2023, doi: 10.3390/info14010054.

[7] C. Kaope and Y. Pristyanto, "The Effect of Class Imbalance Handling on Datasets Toward Classification Algorithm Performance," *Teknik Informatika dan Rekayasa Komputer*, vol. 22, no. 2, pp. 227–238, 2023, doi: 10.30812/matrik.v22i2.2515.

[8] P. Kaur and A. Gosain, "Comparing The Behavior of Oversampling and Undersampling Approach of Class Imbalance Learning by Combining Class Imbalance Problem with Noise," in *Advances in Intelligent Systems and Computing*, Springer Verlag, 2018, pp. 23–30. doi: 10.1007/978-981-10-6602-3_3.

[9] A. Guzmán-Ponce, R. M. Valdovinos, J. S. Sánchez, and J. R. Marcial-Romero, "A New Under-Sampling Method to Face Class Overlap and Imbalance," *Applied Sciences (Switzerland)*, vol. 10, no. 15, Aug. 2020, doi: 10.3390/app10155164.

[10] H. Guo, X. Diao, and H. Liu, "Embedding undersampling rotation forest for imbalanced problem," *Comput Intell Neurosci*, vol. 2018, 2018, doi: 10.1155/2018/6798042.

[11] H. Cai, S. Shen, Q. Lin, X. Li, and H. Xiao, "Predicting the Energy Consumption of Residential Buildings for Regional Electricity Supply-Side and Demand-Side Management," *IEEE Access*, vol. 7, pp. 30386–30397, 2019, doi: 10.1109/ACCESS.2019.2901257.

[12] S. Suparyati, Emma Utami, and Alva Hendi Muhammad, "Applying Different Resampling Strategies In Random Forest Algorithm To Predict Lumpy Skin Disease," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 4, pp. 555–562, Aug. 2022, doi: 10.29207/resti.v6i4.4147.

[13] N. P. Y. T. Wijayanti, E. N. Kencana, and I. W. Sumarjaya, "SMOTE: Potensi dan Kekurangannya Pada Survei," *E-Jurnal Matematika*, vol. 10, no. 4, p. 235, Nov. 2021, doi: 10.24843/mtk.2021.v10.i04.p348.

[14] N. Santoso, W. Wibowo, and H. Himawati, "Integration of Synthetic Minority Oversampling Technique for Imbalanced Class," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 13, no. 1, pp. 102–108, Jan. 2019, doi: 10.11591/ijeecs.v13.i1.pp102-108.

[15] M. Bach, A. Werner, and M. Palt, "The proposal of undersampling method for learning from imbalanced datasets," in *Procedia Computer Science*, Elsevier B.V., 2019, pp. 125–134. doi: 10.1016/j.procs.2019.09.167.

[16] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data," *J Biomed Inform*, vol. 107, Jul. 2020, doi: 10.1016/j.jbi.2020.103465.

[17] H. Wang and X. Liu, "Undersampling bankruptcy prediction: Taiwan bankruptcy data," *PLoS One*, vol. 16, no. 7 July, Jul. 2021, doi: 10.1371/journal.pone.0254030.

[18] S. Sawangarreerak and P. Thanathamathee, "Random forest with sampling techniques for handling imbalanced prediction of university student depression," *Information (Switzerland)*, vol. 11, no. 11, pp. 1–13, Nov. 2020, doi: 10.3390/info11110519.

[19] F. S. Pamungkas, B. D. Prasetya, and I. Kharisudin, "Perbandingan Metode Klasifikasi Supervised Learning pada Data Bank Customers Menggunakan Python," *PRISMA, Prosiding Seminar Nasional Matematika*, vol. 3, pp. 689–694, 2019, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/

[20] N. A. Firdausanti, R. A. Ningrum, and S. Qomariyah, "Comparisons of Logistic Regression and Support Vector Machines in Classification of Echocardiogram Dataset," *Inferensi*, vol. 5, no. 2, p. 85, Sep. 2022, doi: 10.12962/j27213862.v5i2.14121.

[21] Y. Li, N. Adams, and T. Bellotti, "A Relabeling Approach to Handling the Class Imbalance Problem for Logistic Regression," *Journal of Computational and Graphical Statistics*, vol. 31, no. 1, pp. 241–253, 2022, doi: 10.1080/10618600.2021.1978470.

[22] V. Sridhar, M. C. Padma, and K. A. R. Rao, *Emerging Research in Electronics, Computer Science and Technology: Proceedings of International Conference, ICERECT 2018*. in Lecture Notes in Electrical Engineering. Springer Nature Singapore, 2019. [Online]. Available: https://books.google.co.id/books?id=eXWUDwAAQBAJ