Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.1148198
Vol. 11, No. 4, Agustus 2024, hlm. 735-742 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

MODEL KLASIFIKASI DENGAN LOGISTIC REGRESSION DAN RECURSIVE FEATURE ELIMINATION PADA DATA TIDAK SEIMBANG

Sutarman[*1], Rimbun Siringoringo[2], Dedy Arisandi[3], Edi Kurniawan[4], Erna Budhiarti Nababan[5]

[1,2,3,4,5]Universitas Sumatera Utara, Medan
Email: [1]sutarman@usu.ac.id, [2]rimbun@student.usu.ac.id, [3]dedyarisandi@students.usu.ac.id, [4]edikurniawan@students.usu.ac.id, [5]ernabrn@usu.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 06 Desember 2023, diterima untuk diterbitkan: 8 Agustus 2024)

Abstrak

Logistic Regression merupakan metode pengklasifikasi yang sangat populer dan digunakan secara luas pada berbagai penelitian. Logistic Regression dapat memberikan hasil yang baik pada masalah klasifikasi maupun prediksi. Fitur dataset yang besar mengakibatkan beban komputasi, dan menurunkan kinerja klasifikasi. Terdapat tiga dataset yang digunakan pada penelitian ini yaitu Bank marketing, Glass, dan Musk II. Dataset tersebut bersumber dari UCI Repository dan memiliki karakteristik yang berbeda. Ada dua tantangan penggunaan dataset tersebut, yaitu ketidakseimbangan kelas, dan jumlah fitur yang besar. Ada dua tahapan utama pada penelitian ini, yaitu pemrosesan awal dan klasifikasi. Tahapan pemrosesan awal menerapkan seleksi fitur melalui recursive feature elimination, dan penyeimbangan data menggunakan teknik SMOTE. Tahapan klasifikasi menerapkan Logistic Regression. Teknik ridge regression (L2-regularization) diterapkan untuk menghindari overfitting pada tahap validasi model LR. Evaluasi kinerja model didasarkan pada matrik konfusi dan grafik ROC. Hasil penelitian menunjukkan bahwa seleksi fitur dan peyeimbangan kelas memiliki dampak yang baik. Melalui ROC, model LR+RFE+SMOTE memiliki luas sebesar 93%. Hasil ini lebih baik dibanding dengan empat model klasifikasi lainnya, yaitu Naïve Bayes, Decision Tree, K-NN, dan Random Forest.

Kata kunci: *logistic regression, recursive feature elimination, SMOTE*

CLASSIFICATION MODEL USING LOGISTIC REGRESSION AND RECURSIVE FEATURE ELIMINATION ON UNBALANCED DATA

Abstract

Logistic regression is a widely popular classification method extensively used in various studies. Logistic regression can yield good results in classification and prediction problems. The extensive features of the dataset can lead to computational burdens and reduced classification performance. Three datasets were utilized in this research: Bank Marketing, Glass, and Musk II. The dataset is sourced from the UCI Repository and contains various characteristics. There are two challenges associated with using this dataset: class imbalance and a large number of features. There are two main stages in this research: initial processing and classification. At the initial processing stage, feature selection is conducted through recursive feature elimination, and data balancing is achieved using the SMOTE technique. The classification stage applies logistic regression. The ridge regression technique (L2-regularization) is applied to prevent overfitting during the validation stage of the linear regression model. The model performance evaluation is based on confusion matrices and ROC graphs. The research results show that feature selection and class balancing have a positive impact. Through the Receiver Operating Characteristics (ROC) analysis, the LR+RFE+SMOTE model achieved an area under the curve of 93%. These results are better than those of four other classification models, namely Naïve Bayes, Decision Tree, K-NN, and Random Forest.

Keywords: *logistic regression, recursive feature elimination, SMOTE*

1. PENDAHULUAN

Logistic Regression (LR) merupakan metode pengklasifikasi yang sangat populer, berkinerja baik, dan telah diterapkan di berbagai bidang penelitian. Pada bidang medis, LR diterapkan pada klasifikasi kanker kolorektal atau colorectal cancer (CRC) (Feng et al., 2022). Penelitian tersebut menerapkan model multinomial LR pada dataset berdimensi besar yaitu TCGA RNASeq sebanyak 825 fitur. LR dapat menghasilkan performa accuracy, dan

precision yang sangat baik. LR dapat diterapkan pada prediksi pasca pancreatoduodenectomy (Ingwersen et al., 2023). Penelitian tersebut menggunakan data pancreatic cancer audit yang dikumpulkan antara Januari 2014 sampai Januari 2021. LR mampu memperoleh nilai area under curve (AUC) yang lebih baik dibandingkan dengan metode mesin pembelajaran lainnya.

Selain di bidang medis, penerapan LR juga memberikan kinerja yang baik pada bidang keuangan dan bisnis. LR dapat diterapkan untuk memprediksi kondisi keuangan perusahaan (Supsermpol et al., 2023). Penelitian tersebut menggunakan data keuangan pada 111 perusahaan yang terdaftar di bursa efek untuk rentang tahun 2023 sampai 2014. Model prediktif dengan LR mampu menemukan insight berharga bagi perusahaan yaitu faktor-faktor penentu pada aspek sumberdaya. Metode LR memberikan hasil yang sangat baik pada klasifikasi pendapatan rumah tangga (Strzelecka et al., 2020). Penelitian tersebut menggunakan 1000 data keuangan rumah tangga di Polandia. LR dapat mengidentifikasi dan mengevaluasi faktor sosial ekonomi yang menentukan utang rumah tangga.

Penelitian dan penerapan LR tidak terbatas pada data teks, tetapi juga pada citra. Penelitian yang dilakukan oleh (Tang et al., 2011) melakukan investigasi diagnostik pada masalah pra operasi berdasarkan data citra MRI. Penelitian tersebut menunjukkan bahwa LR dapat melakukan klasifikasi diagnostik dan mengungkap faktor-faktor penting pada data citra MRI yaitu faktor peningkatan kontras, penonjolan dinding lateral, dan tidak adanya garis normal. Penelitian yang dilakukan oleh (Jiao et al., 2021) menerapkan LR dengan model pelatihan embedding untuk klasifikasi 100 citra histopathological. Setelah fase pelatihan, LR memperoleh performa Area Under Curve (AUC) di atas 95%.

Selain kelebihan tersebut, LR memiliki tantangan jika menangani dataset yang besar. Dataset yang besar dapat mengakibatkan beban komputasi (Mouhajir et al., 2023). Hal tersebut dapat menjadi kendala dalam memperoleh kinerja algoritma yang baik. Untuk mengatasi hal tersebut, terdapat dua pendekatan yang dapat diterapkan, yaitu pendekatan level data dan level algoritma. Pada level data, upaya yang dilakukan adalah reduksi dimensi data (Kim & Shin, 2019), seleksi fitur (J. Wang et al., 2023), dan penyeimbangan kelas. Pada level algoritma, pendekatan yang dapat diterapkan adalah pemrosesan paralel (Mouhajir et al., 2023).

Penelitian ini menerapkan pendekatan level data melalui tahapan seleksi fitur dan penyeimbangan kelas pada masalah klasifikasi dataset tidak seimbang. Seleksi fitur merupakan metode populer dan andal dalam memilih variabel atau fitur yang relevan dalam rangka meningkatkan performa model klasifikasi. Seleksi fitur memiliki peranan yang besar untuk mereduksi dimensi data dan menungangi efek variabel yang tidak relevan pada kinerja model klasifikasi. Representasi variabel yang tidak relevan dapat mengakibatkan noise atau error klasifikasi. Terdapat banyak penelitian yang menyertakan proses pemilihan fitur sebagai solusi masalah tersebut. Penerapan LR lasso dan seleksi fitur acak dengan metode bostrap pada dataset android malware (Wichitaksorn et al., 2023). Seleksi fitur dapat mengurangi kompleksitas komputasi sehingga menghasilkan performa klasifikasi yang sangat baik. Penerapan seleksi fitur untuk mengurangi fitur yang redundansi pada lima dataset dari UCI. Seleksi fitur tersebut diterapkan pada kerangka mixed-integer exponential cone program (MIEXP) dan menghasikan performa AIC dan BIC yang sangat baik (Asgharieh Ahari & Kocuk, 2023). Penerapan LR dan seleksi fitur pada klasifikasi data medis melalui 825 fitur dataset TCGA RNASeq. Seleksi fitur dapat menghasilkan performa klasifikasi yang lebih baik (Feng et al., 2022). Penerapan LR pada prediksi churn (Kiguchi et al., 2022) menggunakan data digital game-based learning (DGBL). Penelitian tersebut menerapkan seleksi fitur, dan tuning hyper parameter untuk mencapai prediksi yang lebih baik. Penelitian tersebut membuktikan bahwa seleksi fitur dapat menjadi solusi menangani dimensi data yang besar.

Metode recursive feature elimination (RFE) merupakan salah satu algoritma seleksi fitur yang sangat baik. Dibandingkan dengan metode seleksi fitur lainnya, RFE memiliki kelebihan dalam mempertimbangkan relevansi fitur dataset (Han et al., 2023). Dengan menghapus fitur-fitur yang paling tidak penting secara rekursif, RFE dapat secara efektif mereduksi dimensi dataset dan tetap mempertahankan fitur-fitur yang paling informatif. Penelitian yang menerapkan metode RFE pada seleksi model fitur menunjukkan bahwa metode tersebut efektif dalam mengurangi generalisasi galat (error) klasifikasi. Diantaranya seleksi fitur dengan metode RFE serta klasifikasi menggunakan SVM pada dataset bioinformatics (Ding et al., 2022). Seleksi fitur dengan RFE serta klasifikasi dengan metode LR dapat meningkatkan prediktabilitas dan reliabilitas model klasifikasi. Hal tersebut didukung oleh penelitian menerapkan RFE pada seleksi fitur serta klasifikasi menggunakan LR pada diagnosis penyakit kanker (Mathew, 2019), screening penggunaan drugs (Ge et al., 2021), dan Internet of Things (Chalichalamala et al., 2023).

Masalah ketidakseimbangan data merupakan masalah pada machine learning. Ketidakseimbangan data dapat mengakibatkan bias estimasi dan ketidakakuratan prediksi (Charizanos et al., 2024). Penelitian terkait dengan LR pada umumnya terfokus pada membangun model yang probabilistik, pada saat yang sama, lebih sedikit perhatian pada masalah ketidakseimbangan data. Ketidakseimbangan data terjadi jika jumlah kelas

data (misalnya kelas "ya") lebih sedikit atau lebih banyak dibanding kelas data yang lain (misalnya "tidak"). Masalah ketidakseimbangan kelas merupakan masalah umum yang dapat ditemukan pada berbagai jenis aplikasi atau penelitian, seperti deteksi serangan siber (Merino et al., 2020), diagnosa penyakit (Li & Hsu, 2022), deteksi fraud (Gupta et al., 2023), dan bidang psikologi (Pulungan et al., 2023).

Masalah ketidakseimbangan kelas data secara signifikan mempengaruhi performa model klasifikasi. Dampak penyeimbangan kelas terhadap perbaikan performa klasifikasi telah terbukti melalui berbagai studi. Penerapan SMOTE pada masalah klasifikasi menggunakan metode two-layers k-nearest neighbour (kTLNN) (Sun et al., 2024). SMOTE berhasil meningkatkan performa support vector machine pada klasifikasi kelulusan mahasiswa (Hairani, 2021). Penerapan SMOTE dan klasifikasi menggunakan metode LSTM (Han et al., 2023) dapat mengatasi masalah ketidakseimbangan kelas pada data limbah kayu. Penerapan teknik ensemble dengan metode selective bagging memperbaiki akurasi algoritma pada segmentasi pelanggan (Zhu et al., 2023), penerapan teknik ensemble AdaBoost pada masalah ketidakseimbangan data (W. Wang & Sun, 2021) berhasil memperbaiki akurasi dan menurunkan tingkat error.

Sejalan dengan pendekatan di atas, penelitian ini menerapkan pendekatan level data, dengan dua skenario. Pertama menerapkan recursive feature elimination (RFE) pada seleksi fitur, dan skenario kedua adalah penyeimbangan kelas data menggunakan teknik SMOTE.

2. METODE PENELITIAN

2.1. Gambaran Dataset

Penelitian ini menggunakan tiga jenis dataset (Tabel 1) dengan karakteristik yang berbeda, yaitu (1) *Bank marketing*: dataset dengan rasio ketidakseimbangan sedang (0, 75), (2) *Glass*: dataset dengan rasio ketidakseimbangan tinggi (15, 46), dan (3) *Musk II*: dataset dengan jumlah fitur tinggi (166). Deskripsi dataset tersebut dapat dijabarkan dalam Tabel 1.

Tabel 1. Dataset Penelitian

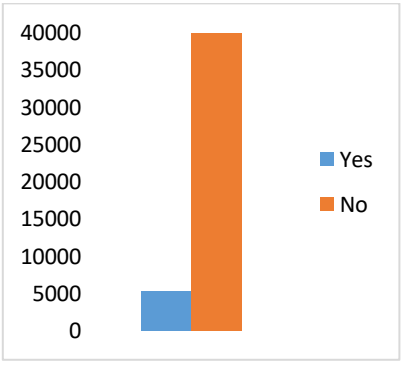
No	Dataset	Dimensi	Rasio
1	*Bank Marketing*	Fitur : 20 Data : 45.210 Kelas : 2	0,75
2	Glass	Fitur : 9 Data : 214	15,46
3	Musk II	Fitur : 166 Data : 460 Kelas : 2	1,44

Dataset *Bank Marketing* bersumber dari *UCI Machine learning repository*. Dataset tersebut terdiri dari 45.210 data dengan 20 fitur independen, dan satu fitur target. Fitur independen terdiri atas 10 fitur numerik, 3 fitur biner, dan 7 fitur kategorial. Daftar fitur dapat dijelaskan pada Tabel 2.

Tabel 2. Deskripsi Dataset *Bank Marketing*

No	Fitur	Fitur Subset	Tipe
1	age		numerik
2	job	admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed	kategorial
3	marital	divorced, married, single	kategorial
4	education	Basic_4y, basic_6y, basic_9y, high_school, illiterate, professional_course, university_degree	kategorial
5	default	yes, no	biner
6	balance		numerik
7	housing	yes, no	biner
8	loan	yes, no	biner
9	contact	cellular, telephone	kategorial
10	day_of_week	mon, tue, wed, thu, fri	kategorial
11	month	jan, feb, mar, …, nov, dec	kategorial
12	duration		numerik
13	campaign		numerik
14	pdays		numerik
15	previous		numerik
16	poutcome	failure, nonexistent, success	kategorial
17	emp_var_rate		numerik
18	cons_price_idx		numerik
19	cons_conf_idx		numerik
20	euribor3m		numerik
21	y	yes, no	biner

Rasio ketidakseimbangan Bank marketing tersebut tampilkan pada Gambar 1.



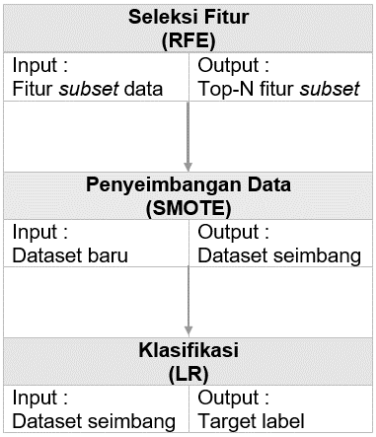
Gambar 1. Distribusi kelas dataset

2.2. Alur Kerja Penelitian

Alur kerja (workflow) penelitian ini dapat dilihat pada Gambar 2.

Berdasarkan Gambar 2, terdapat tiga tahapan utama yaitu seleksi fitur, penyeimbangan kelas, dan klasifikasi. RFE menerima input fitur subset pada dataset, dan memerikan *top-N fitur* paling relevan. Dataset baru dengan fitur paling relevan diberikan sebagai input pada SMOTE, dan memberikan

dataset dengan rasio ketidakseimbangan yang lebih kecil.



Gambar 2. *Workflow* penelitian

Recursive Feature Elimination (RFE) adalah metode pemilihan fitur yang secara berulang kali menghilangkan fitur yang paling tidak penting dari dataset berdasarkan kepentingannya (Wibawa & Novianti, 2017). RFE bekerja dengan melatih model pada data dan kemudian menghitung pentingnya setiap fitur. Fitur dengan kepentingan paling rendah kemudian dihapus, dan model dipasang ulang pada data yang tersisa. Proses ini diulang sampai jumlah fitur yang diinginkan tercapai. Detail alur kerja dapat dilihat pada Gambar 3.

Algoritma RFE dapat dijelaskan pada Algoritma 1.

Algoritma 1 : Recursive Feature Elimination

1. Latih model dengan menggunakan skema *k-fold CV*
2. Tentukan performa model
3. Tentukan fitur penting berdasarkan ranking setiap fitur
4. **For setiap subset-*i* (*i=1,2,3….n*) do :**
   Jadikan subset-*i* sebagai fitur terpenting
   Lakukan proses *Train/Test* model berdasarkan fitur-*i*
   Tentukan kembali performa model
   Tentukan kembali ranking fitur
5. **End For**
6. Tentukan *Top-N* fitur terpenting

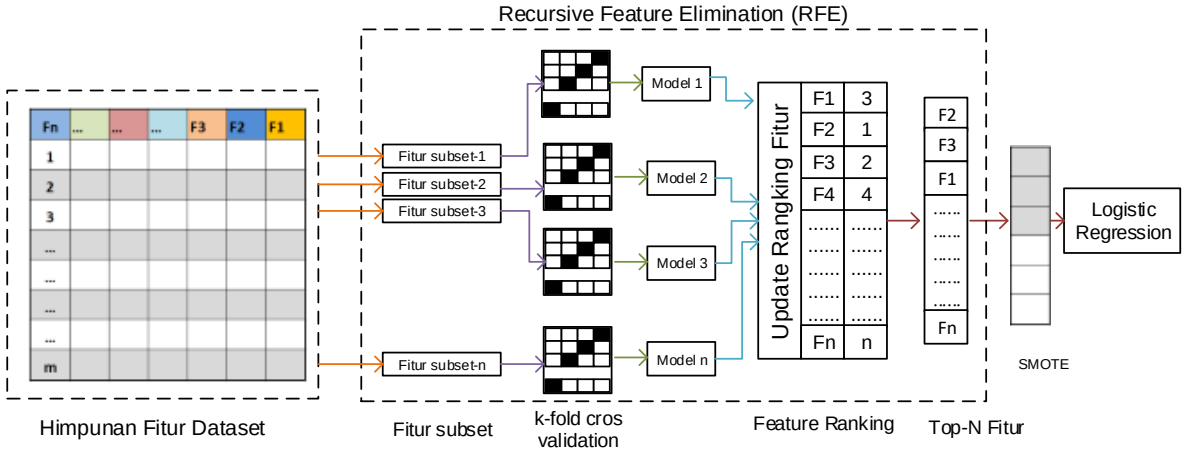
RFE menentukan fitur penting berdasarkan ranking setiap fitur. Pada penerapannya, RFE membutuhkan estimator sebagai *base model* untuk menentukan koefisien setiap fitur. RFE pada model penelitian ini menerapkan *support vector classifier* (SVC) dengan jenis kernel *linier*. Nilai parameter untuk jumlah fitur terseleksi adalah=6, jumlah fitur minimal=1, dan skema *k-fold cross validation* dengan nilai *k=5*.

Algoritma SMOTE dapat jelaskan pada Algoritma 2.

Algoritma 2 SMOTE

1. Input : Data minor $D^{(t)} = \{x_i \in X\}$, dengan i=1,2,3,…T
2. Parameter : T (jumlah data minor), N (persentase SMOTE), *k* (jumlah ketetanggaan)
3. Algoritma :
4. **for *i*=1 to T do :**
5. Temukan sebanyak *k* ketetanggaan terdekat untuk x_i
6. Tentukan $\hat{N} = [\frac{N}{100}]$
7. **while $\hat{N} \neq 0$ do :**
8. Pilih satu dari *k*-tetangga terdekat sebagai $\bar{x}$
9. Pilih bilangan acak $\alpha \in [0,1]$
10. Tentukan $\hat{x} = x_i + \alpha(\bar{x} - x_i)$
11. Tambahkan $\hat{x}$ ke *S*
12. $\hat{N} = \hat{N} - 1$
13. **End while**
14. **End for**
15. Output : Data sintetik *S*

Synthetic Minority Over-sampling Technique (SMOTE) merupakan metode yang populer diterapkan dalam rangka menangani ketidak seimbangan kelas. Teknik ini mensintesis sampel baru dari kelas minoritas untuk menyeimbangkan dataset dengan cara sampling ulang sampel kelas minoritas. Pendekatan ini bekerja dengan membuat replika dari data minoritas (*synthetic data*). SMOTE menambah jumlah data kelas minor agar setara dengan kelas mayor dengan cara membangkitkan data buatan atau sintetis.



Gambar 3. Model penelitian

Logistic Regression (LR) merupakan algoritma pembelajaran terawasi yang dapat bekerja pada data kategorial. Algoritma LR dapat diuraikan pada Algoritma 3.

Algoritma 3 Logistic Regression

1. **Input** : Fitur hasil seleksi
2. **Output** : Hasil klasifikasi
3. **Parameter:** C; jumlah iterasi; penalti , dan nilai toleransi, dan nilai *k* (jumlah *fold cross validation*)
4. **Algoritma** :
5. **For *j*=1 to *k* do :**
6. Set nilai target dengan regresi berdasarkan persamaan (1)
7. $Z = \frac{y_{j-P(1-d_j)}}{[P(1-d_j)(1-p(1-d_j)]}$ (1)
8. Inisialisasi bobot untuk setiap data d_j untuk $P(1|d_j)(1-P)(1|dj)$
9. Terapkan f(j) ke data dengan nilai class (z_j) dan bobot (w_j)
10. **End For**
11. Tentukan (label target:1) jika $P(1|d_j)$ >0,5; selainnya (label target:2)

Hasil klasifikasi dapat berbentuk variabel diskrit atau variabel biner. LR menerapkan fungsi *sigmoid* sebagai fungsi *cost*. Fungsi *sigmoid* logistik pada persamaan (2) dapat memetakan nilai riil hasil prediksi ke bentuk nilai probabilitas antara 0 dan 1.

$$P(x) = 1/1 + e\ \hat{}(-x) \quad (2)$$

P(x) adalah fungsi probabilitas estimasi yang bernilai antara 0 dan 1, *x* adalah fitur input data, konstanta *e* merupakan nilai *Euler* yang bernilai 2, 71828.

Teknik evaluasi dan estimasi performa pada penelitian ini menggunakan skema *k-fold cross-validation*, dengan *k=5*. Hal ini berarti, dataset dibagi menjadi 5 bagian atau *fold* yang sama, setiap *fold* berisi 20% dataset, kemudian dilakukan proses *learning* sebanyak 5 kali. Hasil partisi dataset dapat ditampilkan pada Tabel .3.

Tabel 3. Skema *k-fold cross-validation*

model	training	testing	total
Subset-1	36.168	9.042	45.210
Subset-2	36.168	9.042	45.210
Subset-3	36.168	9.042	45.210
Subset-4	36.168	9.042	45.210
Subset-5	36.168	9.042	45.210

Parameter LR yang diterapkan adalah nilai C=1, 0; jumlah iterasi=100; penalti =L2, dan nilai toleransi=0.0001. Bobot sangat mempengaruhi interaksi antar fitur dataset. Untuk mengontrol besarnya bobot terhadap interaksi fitur, maka parameter *L2-regularization* atau *ridge regression* diterapkan. Cara kerja *L2-regularization* adalah dengan menambahkan nilai norm penalti pada fungsi objektif pada persamaan (3). Pemberian *norm* ini dilakukan hanya terhadap bobot saja, karena bobot yang besar akan berpengaruh besar terhadap sedikit perubahan fitur masukan, Pemberian *norm* ini dilakukan hanya terhadap bobot saja. Semakin besar nilai bobot, semakin besar nilai penalti, begitu pula sebaliknya.

$$L(x,y) = \sum_{i=1}^{n} ((y_i - h_\theta(x_i))^2 + \lambda \sum_{i=1}^{n} \theta_t^2 \quad (3)$$

Matrik konfusi atau *confusion matrix* merupakan metode populer untuk mengukur kinerja *machine learning* khususnya masalah klasifikasi. Matrik konfusi untuk kalsifikasi biner dilihat pada Tabel 4.

Tabel 4. Matrik Konfusi Klasifikasi Biner

		Kelas prediksi		
		yes	no	Total
Kelas Aktual	yes	TP	FN	P
	no	FP	TN	N
	Total	P'	N'	P+N

Pada penelitian ini, daftar kriteria pengukuran kinerja model klasifikasi adalah accuracy, error rate, sensitivity, specificity dan precision. Perhitungan masing-masing kriteria tersebut didasarkan pada persamaan 4,5,6,7, dan 8 pada Tabel 5.

Tabel 5. Pengukuran Performa Prediksi dan Klasifikasi

Kriteria	Formula	
Accuracy, recognition rate	$\frac{(TP+TN)}{P+N}$	(4)
Error rate, misclassification rate	$\frac{FP+FN}{P+N}$	(5)
Sensitivity, true positive rate	$\frac{TP}{P}$	(6)
Specificity, true negative rate	$\frac{TN}{N}$	(7)
Precision	$\frac{TP}{TP+FP}$	(8)

True Positive (TP) : Total kelas 1 yang diklasifikasikan tepat sebagai kelas 1.

True Negative (TN) : Total kelas 0 yang tepat diklasifikasikan sebagai kelas 0

False Positive (FP) : Total kelas 0 yang diklasifikasikan sebagai kelas 1.

False Negative (FN) : Total kelas 1 yang diklasifikasikan sebagai kelas 0.

Kinerja model klasifikasi LR-RFE dievaluasi berdasarkan matrik konfusi untuk menghasilkan grafik *Receiver Over Characteristics* (ROC).

3. HASIL DAN PEMBAHASAN

Seleksi fitur dengan RFE menghasilkan enam fitur yaitu sebagai *top-N feature* yaitu {*euribor3m, job, marital, education, default, contact month, poutcome*}. RFE melakukan seleksi fitur terhadap fitur subset Secara rekursif. Jumlah fitur subset yang dipilih ada sejumlah 16 fitur yaitu {*euribor3m, blue-collar, housemaid, apr, aug, jul, jun, mar, may, nov, oct, failure, success, married lliterate, dec*}. Pada Besarnya pengaruh fitur (*feature importance*)

didasarkan pada nilai standar *error* dan probabilitas (P) fitur, dan dapat dijelaskan melalui Tabel 6.

Seleksi fitur menghasilkan dimensi dataset baru yang lebih kecil. Total data yang dihasilkan adalah 1690 data, dimana total data dengan label "Yes"=1150 (68, 04%), dan label "No"=540 (31, 95%).

Berdasarkan Tabel 6, fitur *euribor3m* memiliki nilai *feature importance* yang paling besar (nilai *error* terkecil). Sebagai informasi tambahan, fitur *euribor3m* atau *Euro interbank offered rate* (Euribor) adalah serangkaian suku bunga referensi yang diterbitkan setiap hari oleh *European Money Markets Institute* yang mencerminkan suku bunga rata-rata di delapan tingkat kematangan di mana bank-bank zona euro menawarkan untuk meminjamkan dana tanpa jaminan satu sama lain.

Tabel 6. Ranking Variabel

No	Fitur	Fitur Subset	Std.Err.	z	P>|z|
1	euribor3m	euribor3m	0.0091	-509.471	0
2	job	blue-collar	0.0283	-61.23	0
3	job	housemaid	0.0778	-41.912	0
4	month	apr	0.0913	-91.49	0
5	month	aug	0.0929	-74.053	0
6	month	jul	0.0935	-43.391	0
7	month	jun	0.0917	-52.55	0
8	month	mar	0.1229	53.989	0
9	month	may	0.0874	-168.815	0
10	month	nov	0.0942	-88.085	0
11	month	oct	0.1175	43.111	0
12	poutcome	failure	0.0363	-137.706	0
13	poutcome	success	0.0618	255.313	0
14	marital	married	0.2253	33.082	0.0009
15	education	illiterate	0.4373	30.084	0.0026
16	month	dec	0.1655	-25.579	0.0105

Metode Synthetic Minority Over Sampling (SMOTE) diterapkan untuk mengatasi ketidakseimbangan kelas data. Hasil penyeimbangan kelas menghasilkan "Yes"=1150 (54, 11%), dan label "No"=975 (45, 88%). Total data keseluruhan setelah proses SMOTE adalah 2.125. Hal lainnya yang dapat diungkap dari dataset adalah frekuensi pembelian berdasarkan variabel. Misalnya frekuensi pembelian berdasarkan variabel job. Job dengan kategori admin memiliki frekuensi pembelian tertinggi, diikuti oleh fitur blue-collar, dan technician.Penerapan RFE dan SMOTE memiliki pengaruh pada rasio ketidakseimbangan data. Perbandingan rasio ketidakseimbangan tersebut dijelaskan pada Tabel 7.

Tabel 7. Perbandingan Rasio Ketidakseimbangan

	Data Awal	RFE	SMOTE
Fitur	20	6	6
Kelas Yes	39. 922	1150	1150
Kelas No	5.288	540	970
Total	45.211	1.690	2.125
Rasio	7,5	2,12	1,18

Pada Tabel 7, efek penerapan RFE adalah reduksi dimensi data melalui seleksi fitur. Seleksi fitur juga memiliki pengaruh pada perbaikan rasio ketidakseimbangan data. RFE memperbaiki rasio ketidakseimbangan dari 7,5 menjadi 2,12. Penerapan SMOTE meningkatkan jumlah data sintesis pada kelas minor, sehingga mengurangi rasio ketidakseimbangan data dari 2,12 menjadi 1,18. Kinerja model diuraikan pada Tabel 8. Pada tabel tersebut, hasil klasifikasi model LR dibandingkan dengan model-model klasifikasi lainnya yaitu *naïve bayes*, *decision tree*, *k-nn* dan *random forest*.

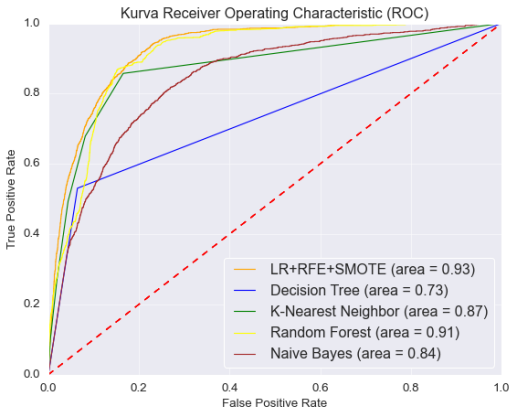
Dengan menggunakan dataset *Bank marketing*, data pada Tabel 8 memiliki interpretasi sebagai berikut:

- ***Accuracy***: Akurasi model dalam memprediksi nasabah berdasarkan dataset *bank marketing* untuk mengajukan deposito adalah 0,9290 (92, 90%). Hal ini memiliki pemahaman bahwa model RFE+SMOTE+LR berhasil memprediksi 92, 90% dari total dataset dengan benar, baik untuk kelas "Ya", maupun "No".
- ***Precision:*** Presisi model dalam memprediksi keputusan nasabah yang sebenarnya "Yes" adalah 0.9489 atau 94, 89%. Hal ini bermakna model dapat memprediksi 94, 89% dari total prediksi kelas "Yes" dengan benar. Namun demikian, terdapat kesalahan prediksi kelas "Yes" sebesar 5, 11%.
- ***Sensitivity***: Dari hasil pengujian, model penelitian ini tingkat *sensitivity* atau *recall* sebesar 0.9761 **(**97, 61%), yang berarti model berhasil mengenali 97, 61% data nasabah yang menjawab dengan "Yes". Namun, ini juga berarti model melewatkan 2, 39% data nasabah yang menjawab "Ya", yang sebenarnya (*False Negative*).

- ***F-Measure***: Data akhir hasil penyeimbangan dataset *Bank marketing* dengan SMOTE menghasilkan data dengan rasio 1, 18. Performa model dari sisi *accuracy* dan *sensitivity* tidak cukup untuk menggambarkan kinerja model secara keseluruhan. *F-Measure* dapat digunakan untuk mengukur kinerja model pada data tidak seimbang. Nilai *F1 Score* model ini adalah **0.9623** atau 96, 23%. Nilai ini menunjukkan bahwa model RFE+SMOTE+RL mampu membedakan kelas "Ya" dan "No" sebesar 96, 23%.
- ***G-Mean*****:** merupakan metrik statistik yang mengukur keberhasilan klasifikasi berimbang untuk kelas "Yes" dan "No". Nilai *G-mean* untuk dataset *bank marketing* menggunakan model RFE+SMOTE+LR adalah 0.7947 atau 79, 47%. Hal ini mengindikasikan bahwa keberhasilan klasifikasi berimbang untuk kedua kelas adalah 79, 47%.
- **Grafik ROC:** Kurva ROC menunjukkan *trade-off* antara *sensitivity* dan *specificity*. Pengklasifikasi yang memiliki kurva yang lebih dekat ke sudut kiri atas (luas kurva yang lebih besar) menunjukkan kinerja yang lebih baik. Pada Gambar 4. ROC model LR+RFE+SMOTE memiliki luas area 0, 93 atau 93%, lebih baik dibanding dengan model-model klasifikasi lainnya.

Tabel 8. Kinerja model secara global untuk tiga dataset

Dataset	Model	Accuracy	Error rate	Sensitivity	Specificity	Precision	F-measure	G-mean
Bank	LR+RFE+SMOTE	**0.9290**	**0.0710**	**0.9761**	0.3174	**0.9489**	**0.9623**	**0.7947**
Marketing	Naïve Bayes	0.8742	0.1258	0.9241	0.4767	0.9337	0.9289	0.6760
	Decision Tree	0.8891	0.1109	0.9357	**0.5229**	0.9390	0.9373	0.6451
	K-NN	0.9049	0.0951	0.9566	0.4848	0.9379	0.9471	0.6731
	Random Forest	0.8365	0.1635	0.9271	0.4401	0.9243	0.0001	0.6623
Glass	LR+RFE+SMOTE	**0.9215**	**0.0785**	**0.9808**	0.4326	0.9345	**0.9571**	0.7084
	Naïve Bayes	0.8807	0.1193	0.9241	**0.5349**	0.9407	0.9323	0.6370
	Decision Tree	0.8866	0.1134	0.9338	0.5192	0.9380	0.9359	0.6471
	K-NN	0.9049	0.0951	0.9566	0.4848	0.9379	0.9471	0.6731
	Random Forest	0.8420	0.1580	0.9500	0.0330	**0.9961**	0.0001	**0.9205**
Musk II	LR+RFE+SMOTE	**0.9215**	0.0785	**0.9808**	0.4326	0.9345	**0.9571**	**0.7084**
	Naïve Bayes	0.8807	0.1193	0.9241	0.5349	0.9407	0.9323	0.6370
	Decision Tree	0.8968	0.1032	0.9356	**0.5556**	0.9488	0.9422	0.6271
	K-NN	0.9049	0.0951	0.9566	0.4848	0.9379	0.9471	0.6731
	Random Forest	0.9134	0.0866	0.9717	0.4326	**0.9553**	0.0001	0.7080



Gambar 4. Kurva ROC model LR+RFE+SMOTE

4. KESIMPULAN

Logistic Regression telah terbukti sebagai algoritma klasifikasi yang efektif dalam membangun model prediktif. Kinerja model tersebut tidak hanya terletak pada disain algoritma, tetapi juga pada faktor lainnya. Pemrosesan awal, pemilihan fitur data merupakan hal esensial dalam meningkatkan kinerja model. Pada penelitian ini dibuktikan bahwa seleksi fitur mempunyai peran yang signifikan dalam meningkatkan akurasi.

DAFTAR PUSTAKA

ASGHARIEH AHARI, S., & KOCUK, B. 2023. A mixed-integer exponential cone programming formulation for feature subset selection in logistic regression. EURO Journal on Computational Optimization, 11, 100069. https://doi.org/https://doi.org/10.1016/j.ejco.2023.100069

CHALICHALAMALA, S., GOVINDAN, N., & KASARAPU, R. 2023. Logistic Regression Ensemble Classifier for Intrusion Detection System in Internet of Things. Sensors, 23(23). https://doi.org/10.3390/s23239583

CHARIZANOS, G., DEMIRHAN, H., & İÇEN, D. 2024. A Monte Carlo fuzzy logistic regression framework against imbalance and separation. Information Sciences, 655(August 2023), 119893. https://doi.org/10.1016/j.ins.2023.119893

DING, X., YANG, F., & MA, F. 2022. An efficient model selection for linear discriminant function-based recursive feature elimination. Journal of Biomedical Informatics, 129, 104070. https://doi.org/https://doi.org/10.1016/j.jbi.2022.104070

FENG, C. H., DISIS, M. L., CHENG, C., & ZHANG, L. 2022. Multimetric feature selection for analyzing multicategory outcomes of colorectal cancer: random forest and multinomial logistic regression models. Laboratory Investigation, 102(3), 236–244. https://doi.org/10.1038/s41374-021-00662-x

GE, C., LUO, L., ZHANG, J., MENG, X., & CHEN, Y. 2021. FRL: An Integrative Feature Selection Algorithm Based on the Fisher Score, Recursive Feature Elimination, and Logistic Regression to Identify Potential Genomic Biomarkers. BioMed Research International, 2021. https://doi.org/10.1155/2021/4312850

GUPTA, P., VARSHNEY, A., KHAN, M. R., AHMED, R., SHUAIB, M., & ALAM, S. 2023. Unbalanced Credit Card Fraud Detection Data: A Machine Learning-Oriented Comparative Study of Balancing Techniques. Procedia Computer Science, 218, 2575–2584. https://doi.org/https://doi.org/10.1016/j.procs.2023.01.231

HAIRANI, H. 2021. Peningkatan Konerja Metode SVM Menggunakan Metode KNN Imputasi dan K-Means-Smote untuk Klasifikasi Kelulusan Mahasiswa Universitas Bumigora. Jurnal Teknologi Informasi Dan Ilmu Komputer, 8(4), 713–718. https://doi.org/10.25126/jtiik.2021843428

HAN, Y., DU, Z., HU, X., LI, Y., CAI, D., FAN, J., & GENG, Z. 2023. Production prediction modeling of food waste anaerobic digestion for resources saving based on SMOTE-LSTM. Applied Energy, 352, 122024. https://doi.org/https://doi.org/10.1016/j.apenergy.2023.122024

INGWERSEN, E. W., STAM, W. T., MEIJS, B. J. V, ROOR, J., BESSELINK, M. G., GROOT KOERKAMP, B., DE HINGH, I. H. J. T.,

VAN SANTVOORT, H. C., STOMMEL, M. W. J., & DAAMS, F. 2023. Machine learning versus logistic regression for the prediction of complications after pancreatoduodenectomy. Surgery, 174(3), 435–440. https://doi.org/https://doi.org/10.1016/j.surg.2023.03.012

JIAO, Y., YUAN, J., QIANG, Y., & FEI, S. 2021. Deep embeddings and logistic regression for rapid active learning in histopathological images. Computer Methods and Programs in Biomedicine, 212, 106464. https://doi.org/https://doi.org/10.1016/j.cmpb.2021.106464

KIGUCHI, M., SAEED, W., & MEDI, I. 2022. Churn prediction in digital game-based learning using data mining techniques: Logistic regression, decision tree, and random forest. Applied Soft Computing, 118, 108491. https://doi.org/https://doi.org/10.1016/j.asoc.2022.108491

KIM, B., & SHIN, S. J. 2019. Principal weighted logistic regression for sufficient dimension reduction in binary classification. Journal of the Korean Statistical Society, 48(2), 194–206. https://doi.org/https://doi.org/10.1016/j.jkss.2018.11.001

LI, Y., & HSU, W. W. 2022. A classification for complex imbalanced data in disease screening and early diagnosis. Statistics in Medicine, 41(19), 3679–3695. https://doi.org/10.1002/sim.9442

MATHEW, T. E. 2019. A Logistic Regression with Recursive Feature Elimination Model for Breast Cancer Diagnosis. International Journal on Emerging Technologies, 10(3), 55–63

MERINO, T., STILLWELL, M., STEELE, M., COPLAN, M., PATTON, J., STOYANOV, A., & DENG, L. 2020. Expansion of Cyber Attack Data from Unbalanced Datasets Using Generative Adversarial Networks. In R. Lee (Ed.), Software Engineering Research, Management and Applications (pp. 131–145). Springer International Publishing. https://doi.org/10.1007/978-3-030-24344-9_8

MOUHAJIR, M., NECHBA, M., & SEDJARI, Y. 2023. High Performance Computing Applied to Logistic Regression: A CPU and GPU Implementation Comparison. 1–5. https://doi.org/10.1109/aibthings58340.2023.10291024

PULUNGAN, M. P., PURNOMO, A., & KURNIASIH, A. 2023. Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Kepribadian MBTI Menggunakan Naive Bayes Classifier. Jurnal Teknologi Informasi Dan Ilmu Komputer, 10(7), 1493–1502. https://doi.org/10.25126/jtiik.1077989

STRZELECKA, A., KURDYŚ-KUJAWSKA, A., & ZAWADZKA, D. 2020. Application of logistic regression models to assess household financial decisions regarding debt. Procedia Computer Science, 176, 3418–3427. https://doi.org/https://doi.org/10.1016/j.procs.2020.09.055

SUN, P., WANG, Z., JIA, L., & XU, Z. 2024. SMOTE-kTLNN: A hybrid re-sampling method based on SMOTE and a two-layer nearest neighbor classifier. Expert Systems with Applications, 238, 121848. https://doi.org/https://doi.org/10.1016/j.eswa.2023.121848

TANG, Z., FENG, X., QIAN, W., & SONG, J. 2011. Evaluation of magnetic resonance imaging criteria for Meckel's cave lesion: logistic regression analysis and correlation with surgical findings. Clinical Imaging, 35(5), 329–335. https://doi.org/https://doi.org/10.1016/j.clinimag.2010.08.013

WANG, J., WANG, H., NIE, F., & LI, X. 2023. Feature selection with multi-class logistic regression. Neurocomputing, 543, 126268. https://doi.org/https://doi.org/10.1016/j.neucom.2023.126268

WANG, W., & SUN, D. 2021. The improved AdaBoost algorithms for imbalanced data classification. Information Sciences, 563, 358–374. https://doi.org/https://doi.org/10.1016/j.ins.2021.03.042

WIBAWA, M. S., & NOVIANTI, K. D. P. 2017. Reduksi Fitur untuk Optimalisasi Klasifikasi Tumor Payudara Berdasarkan Data Citra FNA. Konferensi Nasional Sistem & Informatika, 73–78.

WICHITAKSORN, N., KANG, Y., & ZHANG, F. 2023. Random feature selection using random subspace logistic regression. Expert Systems with Applications, 217, 119535. https://doi.org/https://doi.org/10.1016/j.eswa.2023.119535

ZHU, B., QIAN, C., VANDEN BROUCKE, S., XIAO, J., & LI, Y. 2023. A bagging-based selective ensemble model for churn prediction on imbalanced data. Expert Systems with Applications, 227, 120223. https://doi.org/https://doi.org/10.1016/j.eswa.2023.120223