Xplore (2022), Vol. 11, No. 3, 226-237
DOI: 10.29244/xplore.v11i3.1015

Xplore: Journal of Statistics
Available at https://journal.stats.id/index.php/xplore
e-ISSN: 2655-2744

Analisis Regresi Logistik dan Cart untuk *Credit Scoring* dengan Penanganan Kelas Tak Seimbang*

Siwi Haryu Pramesti[1], Indahwati[2‡], and Utami Dyah Syafitri[3]

[123]Department of Statistics, IPB University, Indonesia
[‡]corresponding author: indahwati@apps.ipb.ac.id

Abstract

The absence of collateral for a type of credit will increase the bank's credit risk (failed to pay). Banks apply the precautionary principle by managing their credit portfolios so that potential hazards that occur can be measured and controlled in a model. Credit scoring describes how likely a debtor will fail with payments. This study aimed to compare logistic regression analysis and Classification and Regression Trees (CART) in classifying debtors to evaluate credit policies. One of the problems in classification is unbalanced data. Synthetic Minority Oversampling Technique (SMOTE) is a technique to handle the unbalanced problem in classification. The results show that the logistic regression model with SMOTE has higher sensitivity than the CART model, and there was no difference in Area Under Curve (AUC). The variables that have significant effects on the classification of debtors (good, bad) are level of education, homeownership status, and income.

Keywords: CART, credit risk, credit scoring, logistic regression, SMOTE

* Received: Aug 2022; Reviewed: Aug 2022; Published: Sep 2022

1. Pendahuluan

Menurut UU Nomor 10 (1998), kredit adalah penyediaan uang atau tagihan yang dapat dipersamakan dengan itu, berdasarkan persetujuan atau kesepakatan pinjam-meminjam antara bank dan pihak lain yang mewajibkan pihak peminjam umtuk melunasi hutangnya setelah jangka waktu tertentu dengan pemberian bunga. Kredit Tanpa Agunan (KTA) adalah salah satu produk pinjaman dari bank yang memberikan fasilitas kredit tanpa membebankan calon debitur untuk mempersiapkan suatu aset sebagai jaminan atas pinjaman tersebut.

Tingginya permintaan kredit tak membuat pihak bank dapat menerima semua permohonan yang ada. Untuk itu, perlu dilakukan suatu proses penyeleksian dalam rangka melihat calon debitur mana yang layak diberi pinjaman. Umumnya bank menggunakan bunga dan jaminan untuk meminimalkan risiko yang akan dihadapi, namun pada KTA yang tidak memiliki jaminan maka risiko yang dihadapi oleh pihak bank terhadap calon debitur semakin tinggi. Salah satu risikonya adalah risiko kredit, yaitu risiko yang disebabkan oleh ketidakmampuan (gagal bayar) dari debitur atas kewajibannya membayar pinjaman (Misdiati 2013). Penerapan prinsip kehati-hatian oleh bank diantaranya diterapkan melalui kemampuan bank untuk mengelola portofolio kredit yang dimiliki, sehingga potensi risiko yang terjadi dapat diukur dan dikontrol. Proses penilaian tersebut dinamakan dengan credit scoring.

Credit scoring menggambarkan seberapa besar kemungkinan debitur akan macet dengan pembayaran. Terdapat berbagai metode statistika untuk menghasilkan model pengklasifikasian seperti analisis diskriminan, regresi linier, regresi logistik, dan pohon keputusan (Hand and Hanley 1997).

Beberapa penelitian terkait klasifikasi debitur dalam ranah credit scoring diantaranya dilakukan oleh Yuli et al. (2012), Guangli et al. (2011), Waluyo et al. (2015), serta Tanjung et al. (2017). Yuli et al. (2012) membuat model tingkat kelancaran pembayaran kredit bank menggunakan model regresi logistik ordinal memperoleh nilai kesesuaian hasil prediksi sebesar 80,86%. Guangli et al. (2011) membandingkan model regresi logistik dan pohon keputusan memperoleh hasil bahwa regresi logistik menunjukkan model yang lebih baik. Waluyo et al. (2015) membandingkan model regresi logistik biner dan CART menunjukkan bahwa akurasi yang dihasilkan CART lebih baik. Tanjung et al. (2017) memperoleh ketepatan klasifikasi sebesar 84,5% dalam penerapan metode CART untuk menentukan faktor-faktor yang memengaruhi pembayaran kredit oleh nasabah.

Penelitian ini menggunakan data mengenai status kredit nasabah bank dimana terdapat ketidakseimbangan kelas pada data tersebut. Ketidakseimbangan kelas merupakan suatu kondisi terjadinya ketimpangan proporsi antara satu kelas dengan kelas lainnya. Umumnya, jumlah debitur macet jauh lebih sedikit dibandingkan debitur lancar. Permasalahan yang muncul pada kasus kelas tidak seimbang adalah prediksi model akan lebih mengarah pada kelas mayoritas, sedangkan kelas minoritas akan menghasilkan ketidaktepatan prediksi.

Ketidaktepatan prediksi akan berdampak pada kebijakan bank, sehingga hal tersebut perlu diatasi. Permasalahan klasifikasi pada ketidakseimbangan kelas dapat ditangani dengan berbagai pendekatan seperti pada data atau pada algoritma (Han et al. 2000). Salah satu metode penanganan ketidakseimbangan kelas yang popular adalah Synthetic Minority Oversampling Technique (SMOTE). Metode ini bekerja dengan cara membangkitkan amatan baru pada kelas minoritas dengan menggunakan konsep k-tetangga terdekat sehingga banyaknya amatan pada kelas minoritas bertambah dan menghasilkan proporsi yang setara dengan kelas mayoritas

(Chawla et al. 2002). Sari dan Irhamah (2019) dalam penelitiannya mengenai menunjukkan bahwa penanganan ketidakseimbangan kelas dengan metode SMOTE menghasilkan rata- rata nilai AUC yang lebih tinggi.

2. Metodologi

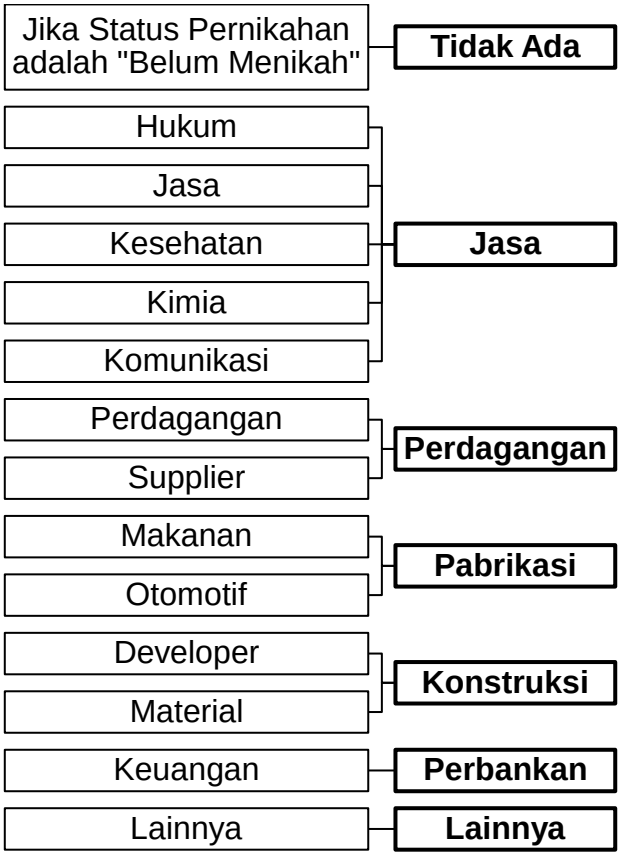
2.1 Data

Data yang digunakan dalam penelitian ini merupakan data debitur pada bank X pada tahun 2017. Terdiri atas 5.199 debitur, 12 peubah penjelas, dan 1 peubah respon. Peubah respon diperoleh dengan mengamati debitur selama satu tahun masa peminjaman, sehingga selanjutnya dapat diketahui apakah debitur tersebut tergolong debitur lancar atau debitur macet pada tahun 2018. Peubah-peubah yang digunakan tercantum dalam Tabel 1.

Tabel 1 Peubah-peubah yang terdapat pada data

No	Peubah	Skala Pengukuran
X1	Jenis Kelamin	Nominal
X2	Status Pernikahan	Nominal
X3	Tingkat Pendidikan	Ordinal
X4	Bidang Usaha	Nominal
X5	Bidang Usaha Pasangan	Nominal
X6	Status Kepemilikan Rumah	Nominal
X7	Jumlah Tanggungan	Rasio
X8	Persentase Pinjaman Disetujui	Rasio
X9	Usia	Rasio
X10	Pendapatan	Rasio
X11	Rasio Cicilan dan Pendapatan	Rasio
X12	Status Pekerjaan	Nominal
Y	Status Kredit	Nominal

Pada peubah bidang usaha dan bidang usaha pasangan dilakukan pengkategorian ulang terhadap kategori yang telah ada mengikuti Universitas Mercu Buana (2011). Hasil dari pengkategorian baru terdapat pada Gambar 1 untuk peubah bidang usaha dan peubah bidang usaha pasangan. Terdapat penambahan kategori baru “Tidak Ada” pada bidang usaha pasangan untuk nasabah dengan status pernikahan belum menikah.



Gambar 1 Pengkategorian peubah bidang usaha dan bidang usaha pasangan

2.2 Prosedur Analisi Data

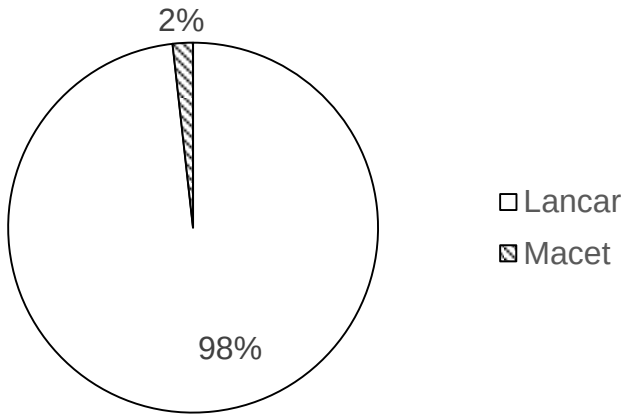
Adapun tahapan analisis data yang akan dilakukan dalam penelitian ini adalah sebagai berikut:

1. Melakukan eksplorasi data untuk mengetahui gambaran umum data
2. Melakukan praproses data dengan mengamati data hilang dan kategori peubah serta melakukan tindakan yang dibutuhkan
3. Mempersiapkan data yang digunakan dengan membagi dalam dua bagian. Data latih sebanyak 70% untuk membentuk model dan data uji sebanyak 30% untuk validasi yang diambil secara acak dengan sistem simple random sampling
4. Melakukan penanganan terhadap kondisi ketidakseimbangan data menggunakan metode SMOTE dengan banyaknya data bangkitan kelas minoritas setara dengan banyaknya data resampling kelas mayoritas. Terdapat 5 jenis SMOTE berdasarkan persentase jumlah data hasil SMOTE terhadap jumlah data latih. SMOTE1 sebesar 10%, SMOTE2 sebesar 20%, SMOTE3 sebesar 30%, SMOTE4 sebesar 40%, dan SMOTE5 sebesar 50%. Proses SMOTE dilakukan sebanyak 100 kali pengulangan untuk melihat kekonsistenan pengklasifikasian.
5. Melakukan analisis regresi logistik
6. Melakukan analisis CART
7. Mengevaluasi model dengan melihat tabel ketepatan klasifikasi dan area dibawah kurva ROC dengan menggunakan Youden Index
8. Memilih model terbaik dan interpretas

3. Hasil dan Pembahasan

3.1 Deskripsi Data

Data yang digunakan terdiri dari 5.199 debitur Bank X yang menggunakan jasa Kredit Tanpa Agunan (KTA) pada tahun 2017. Jumlah data yang termasuk kategori "Lancar" sebanyak 5107 (98%) dan kategori "Macet" sebanyak 92 (2%). Pada Gambar 2 ditunjukkan adanya ketidakseimbangan kelas antara kredit lancar dengan kredit macet.



Gambar 2 Diagram lingkaran status debitur bank X

3.2 Praproses Data

Terdapat 3 amatan hilang pada peubah pendapatan, 4 pencilan pada peubah tanggungan, 1 pencilan pada peubah persentase pinjaman disetujui, dan 1 pencilan pada peubah rasio cicilan dan pendapatan. Perlakuan yang diterima terhadap amatan hilang dan pencilan adalah dihilangkan, sehingga data yang digunakan dalam analisis berjumlah 5190.

3.3 Pembagian Data

Analisis regresi logistik dilakukan untuk mengklasifikasikan data KTA. Pengklasifikasian ini menggunakan data latih dan data uji dengan perbandingan 70:30. Tabel 2 menunjukkan komposisi pembagian data, terlihat bahwa data latih yang digunakan sebesar 3.633 dan data uji sebesar 1.557. Besarnya persentase kelas mayor dan kelas minor pada data latih maupun data uji memiliki perbandingan 98:2. Hal tersebut mengindikasikan bahwa data yang digunakan tidak seimbang.

Tabel 2 Pembagian data

Kelas	Data Latih	Data Uji
Lancar	3.567	1.532
Macet	66	25
Total	3.633	1.557

3.4 Synthetic Minority Oversampling Technique (SMOTE)

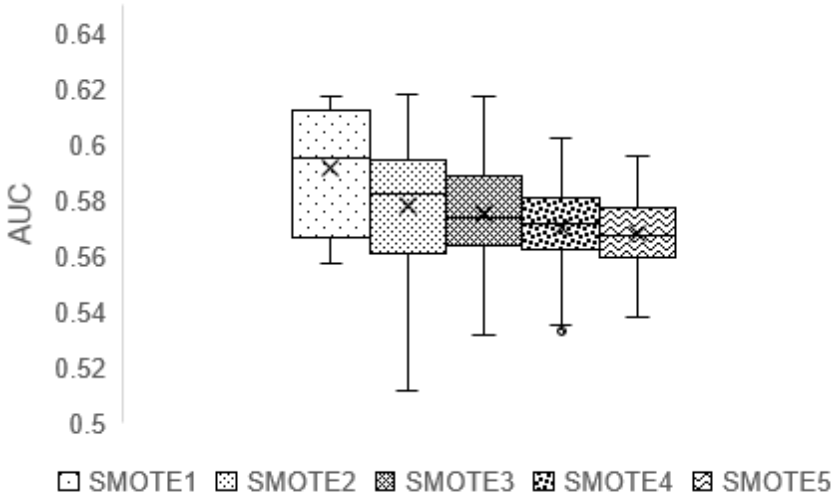
SMOTE yang dilakukan memiliki 5 jenis persentase data yang dihasilkan. Jumlah bangkitan kelas minoritas dan resampling kelas mayoritas dari jumlah data latih sebesar 10% untuk SMOTE1, 20% untuk SMOTE2, 30% untuk SMOTE3, 40% untuk SMOTE4, dan 50% untuk SMOTE5 sebagaimana dapat dilihat data hasil SMOTE pada Tabel 3. Penerapan SMOTE membangkitkan data sintesis pada kategori macet sebanyak data pada kategori lancar sehingga terjadi penyeimbangan kelas.

Tabel 3 Hasil kinerja SMOTE

	Lancar	Macet	Total
Data Latih	3.567	66	3.633
SMOTE1	184	168	336
SMOTE2	363	396	756
SMOTE3	554	528	1028
SMOTE4	726	792	1.518
SMOTE5	900	924	1.824

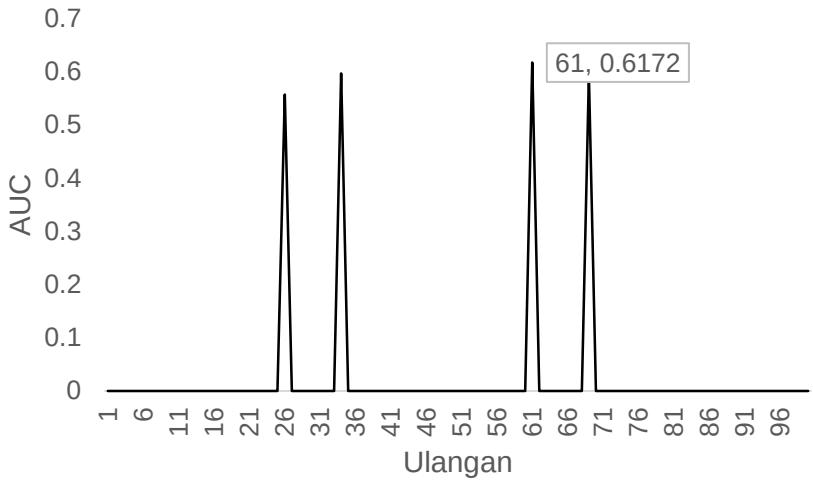
3.5 Regresi Logistik

Jumlah model yang terbentuk pada regresi logistik secara berurutan adalah 4 model pada SMOTE1, 35 model pada SMOTE2, 65 model pada SMOTE3, 75 model pada SMOTE4, dan 91 model pada SMOTE5. Nilai AUC hasil SMOTE ditunjukkan pada Gambar 3. Nilai rata-rata AUC SMOTE1 lebih tinggi daripada nilai rata-rata AUC SMOTE lainnya yaitu sebesar 0,5914.



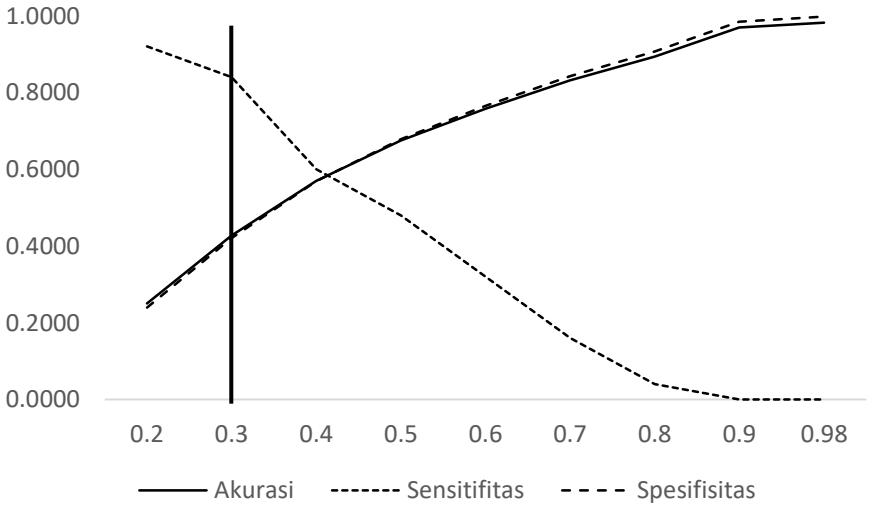
Gambar 3 Nilai AUC SMOTE pada regresi logistik

Pemilihan model ditentukan dengan melihat hasil kinerja SMOTE terbaik. Berdasarkan Gambar 4 SMOTE dengan nilai AUC tertinggi dimiliki oleh data hasil SMOTE1 pada ulangan ke 61 dengan nilai sebesar 0,6172.



Gambar 4 Sebaran nilai AUC pada SMOTE1 regresi logistik

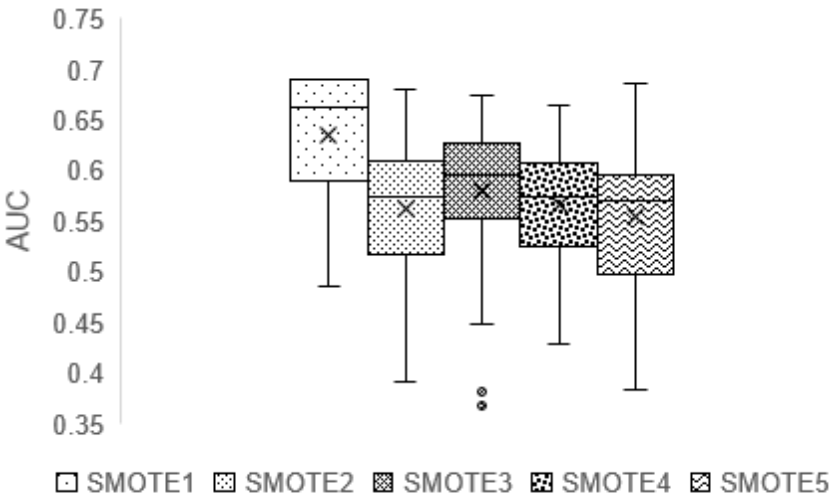
Hasil kinerja klasifikasi regresi logistik dari model ulangan ke 61 dapat dilihat pada Gambar 5. Nilai akurasi dan spesifisitas tertinggi pada titik potong 0,9800, sedangkan nilai sensitifitas tertinggi pada titik potong 0,2000. Nilai titik potong terbaik adalah titik 0,3000 dengan nilai akurasi sebesar 0,4271, nilai sensitifitas sebesar 0,8400, dan nilai spesifisitas sebesar 0,4204.



Gambar 5 Sebaran nilai akurasi, sensitifitas, dan spesifisitas regresi logisti

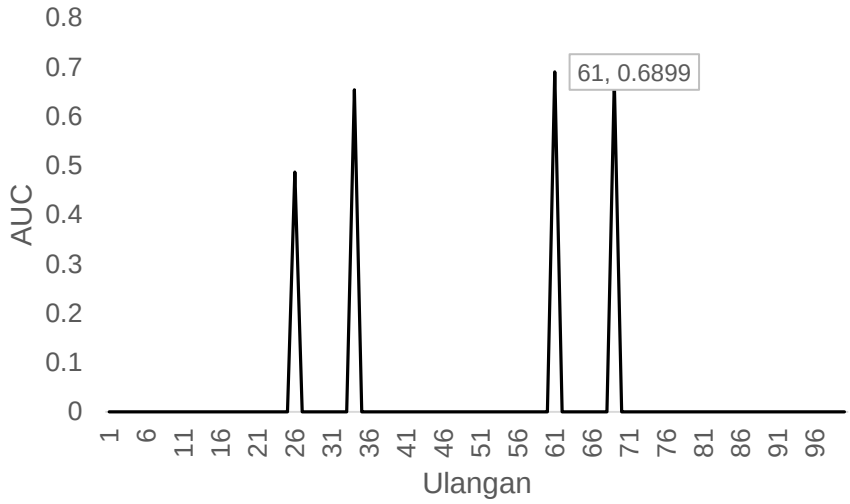
3.6 Classification and Regression Tree (CART)

Jumlah model yang terbentuk pada CART secara berurutan adalah 4 model pada SMOTE1, 35 model pada SMOTE2, 65 model pada SMOTE3, 75 model pada SMOTE4, dan 91 model pada SMOTE5. Nilai AUC SMOTE ditunjukkan pada Gambar 6. Nilai rata-rata AUC SMOTE1 lebih tinggi daripada nilai rata-rata AUC SMOTE lainnya yaitu sebesar 0,6251.



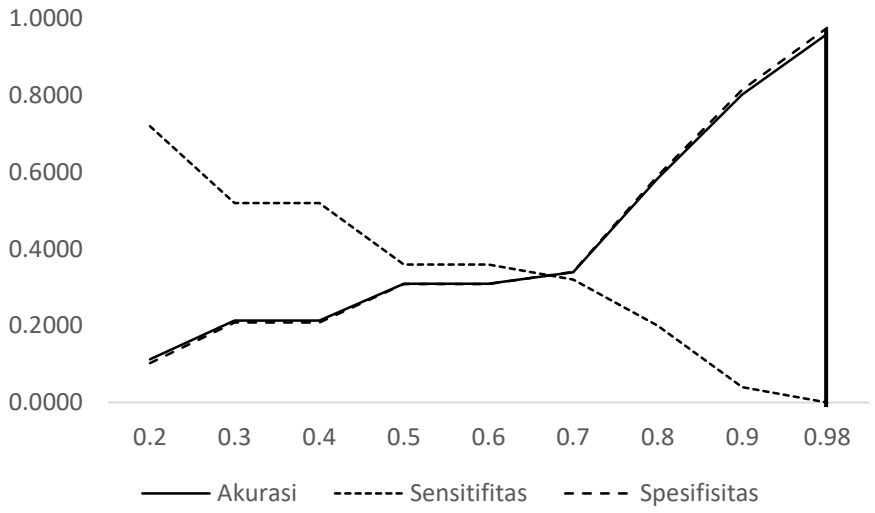
Gambar 6 Hasil kinerja SMOTE pada CART

Pemilihan model ditentukan dengan melihat hasil kinerja SMOTE terbaik. Berdasarkan Gambar 7 SMOTE dengan nilai AUC tertinggi dimiliki oleh data hasil SMOTE1 pada ulangan ke 61 dengan nilai sebesar 0,6899.



Gambar 7 Sebaran nilai AUC pada SMOTE3 CART

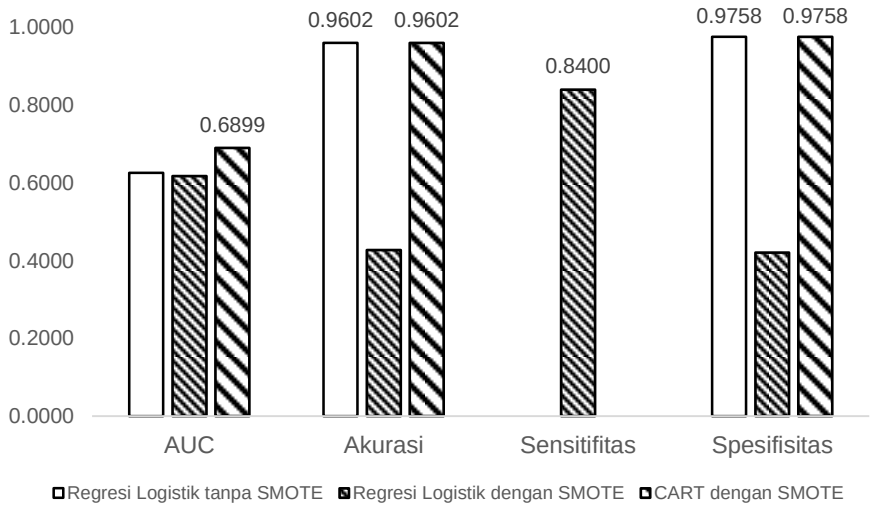
Hasil kinerja klasifikasi CART dari model ulangan ke 61 dapat dilihat pada Gambar 8. Nilai akurasi dan spesifisitas tertinggi pada titik potong 0,9800, sedangkan nilai sensitifitas tertinggi pada titik potong 0,2000. Nilai titik potong terbaik adalah titik 0,9800 dengan nilai akurasi sebesar 0,96018, nilai sensitifitas sebesar 0,0000, dan nilai spesifisitas sebesar 0,9758.



Gambar 8 Sebaran nilai akurasi, sensitifitas, dan spesifisitas CART

3.7 Perbandingan Kebaikan Model

Gambar 9 menunjukkan perbandingan nilai AUC, akurasi, sensitifitas, dan spesifisitas pada regresi logistik tanpa SMOTE, regresi logistik dengan SMOTE, dan CART dengan SMOTE. Terlihat bahwa penanganan SMOTE meningkatnya sensitifitas, yang semula bernilai 0,0000 menjadi 0,8400 pada regresi logistik. Nilai AUC tertinggi dimiliki oleh CART dengan SMOTE. Nilai akurasi dan spesifisitas pada regresi logistik tanpa SMOTE dan CART dengan SMOTE lebih tinggi daripada regresi logistik dengan SMOTE. Nilai sensitifitas pada CART dengan SMOTE bernilai 0,0000, sehingga disimpulkan bahwa Regresi Logistik dengan penanganan SMOTE adalah metode yang terbaik.



Gambar 9 Perbandingan nilai AUC, akurasi, sensitifitas, dan spesifisitas

Model klasifikasi status kredit dibangun dari 12 peubah penjelas yaitu jenis kelamin, status pernikahan, pendidikan, bidang usaha, bidang usaha pasangan, status kepemilikan rumah, jumlah tanggungan, persentase pinjaman disetujui, usia, pendapatan, rasio cicilan, dan status pekerjaan. Pendugaan model menghasilkan

nilai statistik G sebesar 139,7623 dan nilai-p < 0,0000 ini menunjukkan bahwa minimal terdapat satu peubah penjelas yang berpengaruh. Tabel 4 menunjukkan peubah pendidikan, peubah status kepemilikan rumah, dan peubah pendapatan berpengaruh dalam klasifikasi status kredit pada taraf nyata 5%.

Tabel 5 Dugaan parameter dan nilai *odds* model regresi logistik

Peubah	B	Rasio *Odds*	Nilai-p
Konstanta	-151,5877	0,0000	0,9917
Jenis Kelamin			
Wanita	0,7187	2,0518	0,0718
Status Pernikahan			
Menikah	0,1289	1,1375	0,7823
Janda/Duda	0,5092	1,6640	0,6553
Tingkat Pendidikan			
SD	1,0296	2,8001	0,2316
SMP	2,1022	8,1838	0,0018*
SMA	1,7911	5,9958	0,0001*
Diploma	1,8906	6,6233	0,0009*
Bidang Usaha			
Jasa	17,2994	$3,26 \times 10^{7}$	0,9897
Keuangan	35,2175	$1,97 \times 10^{5}$	0,9836
Pabrikasi	1,0745	2,9300	0,9995
Perdagangan	16,3645	$1,28 \times 10^{7}$	0,9903
Lainnya	17,2905	$3,23 \times 10^{7}$	0,9897
Bidang Usaha Pasangan			
Tidak Ada	18,0894	$7,18 \times 10^{7}$	0,9940
Jasa	19,5832	$3,20 \times 10^{8}$	0,9935
Keuangan	3,6744	$3,94 \times 10^{1}$	0,9991
Pabrikasi	2,0772	7,9800	0,9994
Perdagangan	17,2317	$3,05 \times 10^{7}$	0,9943
Lainnya	17,6419	$4,59 \times 10^{7}$	0,9941
Status Kepemilikan Rumah			
Keluarga	1,4704	4,3508	$4,17 \times 10^{-5}$*
Milik Sendiri	0,0469	1,0480	0,8963
Sewa Bulanan	0,9820	2,6698	0,2087
Sewa Tahunan	0,4619	1,5871	0,4418
Jumlah Tanggungan	0,0410	1,0418	0,7706
Usia	-0,0593	0,9424	0,9945
Pendapatan	$3,24 \times 10^{-11}$	1,0000	0,0232*
Rasio Cicilan dan Pendapatan	1,6444	5,1781	0,1362
Status Pekerjaan			
Tetap	18,9485	$1,70 \times 10^{8}$	0,9937

Interpretasi untuk parameter regresi akan lebih mudah dilihat dari nilai rasio *odds*. Pada peubah tingkat pendidikan, nasabah dengan tingkat Pendidikan SD memiliki peluang macet dibandingkan tidak macet *(odds*) 8,1838 kali dibandingkan nasabah dengan tingkat pendidikan sarjana, sedangkan nasabah dengan tingkat pendidikan SMP, SMA, dan diploma secara berurutan memiliki *odds* 8,1838, 5,9958, dan 6,6233 kali dibandingkan nasabah dengan tingkat pendidikan sarjana, Pada peubah status kepemilikan rumah, peluang untuk dikategorikan sebagai nasabah macet

dibandingkan tidak macet (*odds*) adalah sebesar 4,3508 kali untuk nasabah dengan status kepemilikan rumah milik keluarga dibandingkan dengan nasabah yang memiliki status kepemilikan rumah milik perusahaan, sedangkan nilai *odds* untuk nasabah dengan status kepemilikan rumah milik sendiri, sewa bulanan, dan sewa tahunan masing-masing sebesar 1,0480, 2,6698, dan 1,5871 kali dibandingkan dengan nasabah yang memiliki status kepemilikan rumah milik perusahaan. Pada peubah pendapatan, rasio *odds* menggambarkan setiap kenaikan pendapatan satu satuan maka kecenderungan untuk dikategorikan sebagai nasabah macet meningkat sebesar 1 kali.

4. Simpulan dan Saran

Data dengan kelas tidak seimbang mengakibatkan model hanya dapat mengklasifikasikan kelas mayor namun tidak dapat mengklasifikasikan kelas minor. Teknik sampling SMOTE pada data tidak seimbang mampu meningkatkan nilai sensitifitas. Regresi logistik dengan SMOTE dipilih sebagai model terbaik pada penelitian ini karena menghasilkan nilai sensitifitas yang lebih tinggi dibandingkan CART dengan SMOTE. Nilai AUC pada regresi logistik dengan SMOTE tidak jauh berbeda daripada CART dengan SMOTE. Regresi logistik dengan SMOTE menghasilkan tiga peubah berpengaruh dalam klasifikasi pada taraf nyata 5%, yaitu tingkat pendidikan, status kepemilikan rumah, dan pendapatan. Memperhatikan kembali saat memasukkan data, untuk menghindari munculnya amatan kosong atau amatan yang kurang sesuai. Kualitas data pada bank perlu ditingkatkan untuk mengetahui karakteristik debitur terutama pada peubah bidang usaha dan bidang usaha pasangan yang memiliki banyak kategori. Persentase terbesar terletak pada kategori lainnya yang sulit untuk diketahui secara rinci. Penelitian selanjutnya diharapkan untuk melakukan dengan analisis lain seperti *Random Forest* dapat dicoba pada penelitian selanjutnyaum.

Daftar Pustaka

Chawla VN, Bowyer KW, Hall LO, Kegelmeyer WP. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. 16: 321-357.

Guangli N, Wei R, Lingling Z, Yingjie T, Yong S. 2011. Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*. 38:15273-15285.

Han H, Wang WY, Mao BH. 2005. Borderline-SMOTE: a new over-sampling method in imbalance data sets learning. *Journal of Lecture Notes in Computer Science*. 3644:878-887.

Hand DJ, Henley WE. 1997. Statistical classification methods in consumer credit scoring. *J Royal Statist Society A*. 160: 523-541.

Misdiati L, Rahayu SP. 2013. Analisis klasifikasi kredit menggunakan metode newton truncated-kernel logistic regression (NTR-KLR) (studi kasus: data kredit bank “X”) [skripsi]. Surabaya: Institut Teknologi Sepuluh November.

Sari EDN, Irhamah. 2019. Analisis Sentimen Nasabah Pada Layanan Perbankan Menggunakan Metode Regresi Logistik Biner, Naïve Bayes Classifier (NBC), dan Support Vector Macine (SVM). *Jurnal Sains dan Seni ITS*. 8(2):2337-3520

Tanjung RH, Kartiko. 2017. Penerapan metode CART (classification and regression trees) untuk menentukan faktor-faktor yang mempengaruhi pembayaran kredit

oleh nasabah (studi kasus bank BRI unit aek tarum – sumatera utara. *Jurnal Statistika Industri dan Komputasi*. 2: 78-83.

Universitas Mercu Buana. 2011. *Membangun Usaha Sukses Sejak Usia Muda*. Jakarta. Salemba Empat.

[UU] Undang-undang Republik Indonesia Nomor 10 Tahun 1998 Tentang Perubahan Atas Undang-undang Nomor 7 Tahun 1992 Tentang Perbankan. 1998.

Yuli A, Uxti M, Herlina H. 2012. Model tingkat kelancaran pembayaran kredit bank menggunakan model regresi logistik ordinal (studi kasus: bank rakyat indonesia tbk unit pasar bintuhan). *Jurnal Gradien*. 8(2):809-814.

Waluyo A, Mukid MA, Wuryandari T. 2015. Perbandingan Analisis Klasifikasi Nasabah Kredit Menggunakan Regresi Logistik Biner dan CART (Classification and Regression Trees). *Jurnal Gaussian*. 4(2):215-225