E-ISSN 2621–086X

Indonesian Journal of Applied Statistics

Vol. 6, No. 2, pp. 152-161, November 2023
https://jurnal.uns.ac.id/ijas/article/view/79389
https://doi.org/10.13057/ijas.v6i2.79389

Klasifikasi Menggunakan Algoritma K-Nearest Neighbor pada Imbalance Class Data dengan SMOTE. (Studi Kasus: Nasabah Bank Perkreditan Rakyat "X")

Salsabilla Rizka Ardhana[1], Tatik Widiharih[2*], and Bagus Arya Saputra[3]
[123]Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro

***Corresponding author*:** widiharih@gmail.com

Abstract. *Rural Banks (Bank Perkreditan Rakyat/BPR) provide financial services to micro-businesses and low repayment communities, especially in rural areas. The main activity of the bank is lending. Customer credit classification is expected to assist BPR in anticipating potential bad loans. K-Nearest Neighbor classify current and potential bad credit status based on customer data from BPR "X" in Central Java in October 2022. K-Nearest Neighbor is effective against a large amount of training data and works based on the nearest neighbor. There is an imbalance class data which causes the classification process to focus more on the majority class. Imbalance class data is handled using Synthetic Minority Oversampling Technique (SMOTE) as an oversampling approach. Classification with the addition of SMOTE can improve the evaluation of classification accuracy, especially G-mean. G-mean is the most comprehensive measurement in term of accuracy, sensitivity and specificity in evaluating classification performance on imbalance class data. The results of this research were able to increase g-mean to 58.55% and sensitivity to 45.46% by implementing SMOTE. Based on the classification results, it is concluded that K-Nearest Neighbor with SMOTE at k = 19 and a proportion of training data to test data of 70:30 is a more appropriate classification model to use for customer credit status.*

Keywords: *Credit Status; K-Nearest Neighbor; Imbalance Class Data; SMOTE*

1. PENDAHULUAN

Bank Perkreditan Rakyat (BPR) adalah bank yang memiliki peran dalam menyediakan layanan keuangan untuk usaha mikro dan kecil (UMK) dan masyarakat berpenghasilan rendah terutama di daerah pedesaan. Kehadiran BPR ditunjukkan untuk membantu usaha kecil dalam berkembang serta memberikan layanan kebutuhan perbankan bagi ekonomi lemah yang tidak dapat dijangkau oleh bank umum. Kegiatan BPR mencakup kegiatan penyaluran dan penghimpunan dana tetapi BPR dilarang menerima simpanan giro [1].

Klasifikasi status kredit nasabah berdasarkan latar belakang nasabah seperti usia, jenis kelamin, pekerjaan, pendapatan, status perkawinan, dan pendidikan diharapkan dapat membantu BPR mengantisipasi kerugian yang ditimbulkan akibat kredit berpotensi macet. Klasifikasi merupakan proses menemukan dan membedakan kelas data untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui [2]. Metode klasifikasi yang digunakan dalam penelitian ini adalah *K-Nearest Neighbor*. *K-Nearest Neighbor* merupakan metode yang melakukan klasifikasi

Submitted: 08 Oktober 2023; ***Accepted:*** 28 Februari 2024

dengan menggunakan konsep kedekatan jarak. Prinsip kerja *K-Nearest Neighbor* dengan mencari jarak paling dekat antara data yang akan dievaluasi dengan *k* tetangga terdekat pada data latih.

Model klasifikasi yang baik dapat dilihat dari tingkat ketepatan dalam melakukan prediksi yang dapat dipengaruhi oleh ketidakseimbangan data. *Imbalance class data* terjadi ketika jumlah suatu label kelas lebih sedikit dibandingkan dengan jumlah label kelas data yang lainnya pada data latih. Menurut Singh dan Sharma [3] klasifikasi cenderung memiliki tingkat ketepatan yang baik pada kelas mayoritas tetapi sangat buruk pada kelas minoritas. Permasalahan *imbalance class data* dapat ditangani salah satunya dengan menggunakan *Synthetic Minority Oversampling Technique* (SMOTE). SMOTE merupakan pendekatan oversampling dengan membuat data sintetis baru pada kelas minoritas sehingga data minoritas menjadi seimbang dengan data mayoritas [4].

Penelitian terdahulu telah dilakukan oleh Permana dkk [5] dengan membandingkan *K-Nearest Neighbor* dengan C5.0 pada data kredit koperasi dengan tingkat akurasi C5.0 lebih tinggi dibandingkan *K-Nearest Neighbor*. Penelitian lain yang dilakukan oleh Sharma dan Kumar [6] pada klasifikasi *smart city* menghasilkan tingkat akurasi *K-Nearest Neighbor* yang lebih tinggi dibandingkan C5.0. Selain itu pada penelitian yang dilakukan Ramadhanti [7] yang membandingkan *K-Nearest Neighbor* dengan penyeimbang data *Adaptive Syntethic Sampling Approach* (ADASYN) dan SMOTE diperoleh bahwa SMOTE memiliki tingkat akurasi, sensitivitas, dan spesifisitas yang lebih baik dibandingkan ADASYN pada data klasifikasi rumah tangga miskin. Pada penelitian Umma [8] tentang klasifikasi status kemiskinan rumah tangga menyatakan bahwa pada metode C5.0, data yang diseimbangkan dengan SMOTE menghasilkan tingkat sensitivitas dan spesifisitas yang lebih baik dibandingkan dengan data yang tidak diseimbangkan.

Penelitian ini bertujuan untuk mengklasifikasikan status kredit nasabah berdasarkan latar belakang nasabah Bank Perkreditan Rakyat "X" di regional Jawa Tengah dengan metode *K-Nearest Neighbor* tanpa penanganan dan dengan penanganan *imbalance class data* menggunakan SMOTE. Variabel prediktor yang digunakan meliputi: jenis kelamin, usia, jangka waktu, jumlah pinjaman, pendapatan, status perkawinan, pekerjaan, jenjang pendidikan, dan jenis jaminan.

2. METODE PENELITIAN

Proses klasifikasi didasarkan pada empat komponen [9], yaitu :

a) Kelas merupakan variabel dependen dari model yang merupakan variabel kategori yang
   mewakili label pada objek setelah klasifikasinya, pada beberapa penelitian lain disebut juga variabel target.
b) Prediktor merupakan variabel independen dari model yang diwakili oleh karakteristik variabel dari data yang akan diklasifikasikan.
c) Data latih merupakan kumpulan data yang berisi nilai dari dua komponen sebelumnya, dan digunakan untuk melatih model dalam mengenali kelas yang sesuai, berdasarkan prediktor yang tersedia.
d) Data uji merupakan data baru yang akan diklasifikasikan oleh model yang telah dibuat sehingga menghasilkan ketepatan klasifikasi (kinerja model) dan dapat dievaluasi.

2.1 *K-Nearest Neighbor*

Menurut Han dan Kamber [2] klasifikasi adalah proses menemukan model yang bertujuan memperkirakan kelas dari objek yang tidak diketahui label kelasnya. Algoritma *K-Nearest Neighbor* (KNN) mengklasifikasikan data berdasarkan kedekatan jarak suatu data dengan data yang lain. KNN memiliki parameter *k* dengan nilai *k* paling sedikit 1 dan paling besar adalah hasil akar kuadrat dari jumlah data training. Pada klasifikasi dengan dua kelas sebaiknya menggunakan nilai *k* ganjil untuk menghindari kemungkinan kelas yang berbeda memiliki jumlah label yang sama [10].

Jumlah tetangga terdekat yang digunakan untuk memperkirakan label kelas pada data uji dinyatakan dengan nilai *k*. Setelah *k* tetangga terdekat terpilih, dilakukan pemilihan label kelas berdasarkan mayoritas label dari *k* tetangga terdekat tersebut. Label kelas hasil prediksi pada data uji ditetapkan dari kelas yang memiliki jumlah label tetangga paling banyak [11]. *K-Nearest Neighbor* juga didasarkan pada pengukuran jarak. Penelitian ini menggunakan jarak *euclidean* untuk pengukuran jarak seperti pada Persamaan 1:

$$d\left(x_i, y_i\right) = \sqrt{\sum_{l=1}^{p} \left(diff\left(x_{il}, y_{il}\right)\right)^2} \quad (1)$$

dengan $d\left(x_i, y_i\right)$ = Jarak *euclidean* data uji ke-i dan data latih ke-i, x_{il} = data uji ke-i pada variabel ke-l, y_{il} = data latih ke-i pada variabel ke-l , p = dimensi data variabel bebas, dan $diff\left(x_{il}, y_{il}\right)$ = selisih antara x_{il} dan y_{il}.

Menurut Prasetyo [13], perhitungan nilai ketidaksamaan yang digunakan dalam persamaan jarak *euclidean* bergantung pada tipe data yang digunakan seperti pada Tabel 1.

Tabel 1. Selisih Dua Data dengan Satu Atribut

Tipe Atribut	Formula Jarak
Nominal	$diff\left(x_{il}, y_{il}\right) = \{0,\ apabila\ x_{il} = y_{il}\ 1,\ apabila\ x_{il} \neq y_{il}$
Ordinal	$diff\left(x_{il}, y_{il}\right) = \left\|x_{il} - y_{il}\right\|/(c - 1)$ dengan c = banyaknya pengkategorian dalam x
Interval atau Rasio	$diff\left(x_{il}, y_{il}\right) = \left\|x_{il} - y_{il}\right\|$

Algoritma KNN menurut [13] adalah sebagai berikut :

1. Membagi data *training* dan data *testing*.
2. Menghitung jarak *euclidean* antara setiap data *training* dengan setiap data *testing*.
3. Mengurutkan jarak *euclidean* dari yang terkecil sampai terbesar.
4. Tentukan nilai $k = 2, 3, 4, \ldots, \sqrt{n}$, dengan *n* adalah jumlah data *training*.
5. Periksa kelas dari *k* tetangga terdekat.
6. Menetapkan kelas terbanyak dari *k* sebagai kelas data *testing*.
7. Evaluasi hasil klasifikasi dengan mengukur nilai akurasi.

2.2 *Synthetic Minority Oversampling Technique* (SMOTE)

Pada beberapa dataset seringkali ditemukan ketidakseimbangan data (*imbalance class data*). Ketidakseimbangan data terjadi saat jumlah objek dalam suatu kelas lebih banyak dibandingkan dengan kelas yang lain. Penerapan algoritma klasifikasi dapat gagal karena tidak mampu menggambarkan karakteristik data secara tepat. Hal tersebut dapat memberikan hasil performa yang buruk pada seluruh kelas data saat diberikan data yang tidak seimbang karena

sebagian besar algoritma klasifikasi mengasumsikan distribusi kelas yang seimbang [15]. Permasalahan *imbalance class data* dapat diatasi menggunakan SMOTE (*Synthetic Minority Oversampling Technique*) yaitu metode oversampling pada kelas minoritas dengan membuat sampel sintetis [4]. Pembangkitan data sintetis dilakukan dengan cara berikut:

a. Data Numerik
   Data yang memiliki jarak terdekat kemudian digunakan untuk membangkitan data sintetis baru menggunakan Persamaan (2).

$$X_{baru_{[a]}} = x + \left(x^{*} - x\right) \times rand[0,1]_a \quad (2)$$

   dengan $X_{baru_{[a]}}$ = data sintetis hasil dari replikasi ke *a,* x = data acak yang akan direplikasi, x^{*} = data acak yang memiliki jarak terdekat dari data yang akan direplikasi, $rand[0,1]_a$ = bilangan acak antara 0 sampai 1 untuk replikasi ke *a.*

b. Data Kategorik
   Pembangkitan data sintetis baru yang berskala kategorik dilakukan dengan memilih kelas mayoritas yang dievaluasi dengan *k* tetangga terdekat. Apabila terdapat kesamaan nilai kelas mayoritas yang muncul, maka akan dipilih secara acak. Suatu objek yang terpilih sebagai dasar pembangkitan data sintetis, pembangkitan data sintetis dilakukan pada masing-masing variabel sesuai dengan sifat data (numerik ataukah kategorik)

2.3 Evaluasi Ketepatan Hasil Klasifikasi

Pada penelitian ini, kinerja klasifikasi diukur melalui matriks konfusi (*confusion matrix*). Ukuran kinerja dari model klasifikasi dapat dievaluasi menggunakan perhitungan statistik diantaranya keakuratan, sensitivitas, spesifisitas, dan *g-mean* [16]. Perhitungan tersebut ditentukan melalui *true positive* (TP), *true negative* (TN), *false positive* (FP), *false negative* (FN). Tabel 2 merupakan *confusion matrix* untuk klasifikasi dua kelas yang menampilkan TP, TN, FP, FN.

Tabel 2**.** *Confusion* Matrix untuk Klasifikasi Dua Kelas

		Kelas Aktual	
		Positif	Negatif
Kelas Prediksi	Positif	TP	FP
	Negatif	FN	TN

$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

$$Error\ Rate = \frac{FP+FN}{TP+FP+TN+FN} \quad (4)$$

$$Sensitivitas = \frac{TP}{TP+FN} \quad (5)$$

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

$$G - mean = \sqrt{Sensitivitas \times Spesifisitas} \quad (7)$$

Keakuratan klasifikasi menunjukkan proporsi data yang diprediksi dengan benar. Secara ekuivalen, kinerja sebuah model dinyatakan dalam bentuk *error rate*. Sensitivitas menunjukkan proporsi data positif yang diprediksi dengan benar sebagai data positif. Spesifisitas menunjukkan proporsi data negatif yang diprediksi dengan benar sebagai data negatif. Evaluasi

ketepatan hasil klasifikasi keseluruhan pada metode klasifikasi dapat dihitung menggunakan *g-mean (geometric mean)* yaitu suatu pengukuran yang paling komprehensif dalam melakukan evaluasi kinerja klasifikasi khususnya pada persoalan *imbalance class data*. *G-mean* diartikan sebagai rata-rata *geometric* dari sensitivitas dan spesifisitas [16]. *G-mean* akan bernilai satu apabila seluruh amatan diklasifikasikan dengan tepat.

2.4 Metodologi Penelitian

Data yang digunakan dalam penelitian ini adalah data sekunder nasabah Bank Perkreditan Rakyat "X" di regional Provinsi Jawa Tengah pada bulan Oktober 2022. Unit observasi pada penelitian ini adalah pada regional Provinsi Jawa Tengah. Data yang digunakan terdiri dari 431 data nasabah.

Variabel Penelitian

Variabel yang digunakan dalam penelitian ini yang disajikan pada Tabel 3. Langkah-langkah analisis:

a. Membagi menjadi data *training* dan *testing* dengan proporsi 60:40 70:30 dan 80:20.
b. Melakukan klasifikasi dengan metode KNN pada data sebelum diseimbangkan.
c. Menghitung akurasi, *sensitifitas, specificitas* dan *G-mean.*
d. Mengatasi ketidakseimbangan data dengan SMOTE.
e. Melakukan klasifikasi dengan metode KNN pada data yang sudah diseimbangkan
f. Menghitung akurasi, *sensitifitas, specificitas* dan *G-mean.*
g. Interpretasi hasil klasifikasi.

Tabel 3. Variabel Penelitian

Nama Variabel	Keterangan	Tipe Variabel (Skala)
Status Kredit	0 : Lancar 1: Berpotensi Macet	Kategorik (Nominal)
Jenis Kelamin	1: Laki-laki 2 : Perempuan	Kategorik (Nominal)
Usia	1 : <35 tahun 2 : 35-50 tahun 3 : >50 tahun	Kategorik (Ordinal)
Jangka Waktu	1 : <12 bulan 2 : 12-36 bulan 3 : >36 bulan	Kategorik (Ordinal)
Jumlah Pinjaman	1 : <Rp.50.000.000 2 : Rp. 50.000.000-Rp. 100.000.000 3 : >Rp.100.000.000	Kategorik (Ordinal)
Pendapatan	1 : ≤5.000.000 2 : Rp. 5.000.001-Rp. 10.000.000 3 : >Rp.10.000.000	Kategorik (Ordinal)
Status Perkawinan	1 : Menikah 2 : Belum Menikah 3 : Duda/Janda	Kategorik (Nominal)
Pekerjaan	1 : Karyawan Swasta 2 : Pengajar (Guru, Dosen) 3 : Wiraswasta 4 : Peternak/Petani 5 : Pensiunan	Kategorik (Nominal)

Nama Variabel	Keterangan	Tipe Variabel (Skala)
	6 : Pegawai Pemerintahan Non Guru 7 : Lain-lain	
Jenjang Pendidikan	1 : SD 2 : SMP 3 : SMA/SMK 4 : DIPLOMA 5 : SARJANA	Kategorik (Ordinal)
Jenis Jaminan	1 : Tanpa Jaminan 2 : BPKB 3 : SHM 4 : Deposito	Kategorik (Nominal)

3. HASIL DAN PEMBAHASAN

Data yang digunakan sebanyak 431 data yang terdiri dari 394 status kredit lancar atau 91,42% dan 37 status kredit berpotensi macet atau 8,58%. Dataset dibagi menjadi dua bagian yaitu data latih dan uji dengan proporsi 60:40, 70:30, dan 80:20.

Klasifikasi KNN dari data nasabah. Nilai *k* tetangga terdekat pada *K-Nearest Neighbor* yang digunakan dalam penelitian ini yaitu 3,5,7,9,11,13,15,17 (menurut Hassanat *et al* [9]). Setiap *k* dihitung nilai *error rate* (menggunakan bantuan *Rstudio*), nilai *error rate* tertinggi dipilih sebagai *k* terbaik yang akan digunakan untuk klasifikasi dengan KNN. Nilai *error rate* dari data *training* disajikan pada Tabel 4.

Tabel 4. Tabel Nilai *Error Rate* Data *Training* pada Beberapa Nilai k

Nilai k	Proporsi Data Latih : Data Uji					
	60:40		70:30		80:20	
	Akurasi	*Error Rate*	Akurasi	*Error Rate*	Akurasi	*Error Rate*
3	0,9006	0,0994	0,8915	0,1085	0,9059	0,0941
5	**0,9181**	**0,0819**	0,9147	0,0853	0,9176	0,0824
7	0,9181	0,0819	0,9147	0,0853	0,9176	0,0824
9	0,9181	0,0819	0,9147	0,0853	0,9176	0,0824
11	0,9181	0,0819	0,9147	0,0853	0,9176	0,0824
13	0,9181	0,0819	0,9147	0,0853	0,9176	0,0824
15	0,9181	0,0819	0,9147	0,0853	0,9176	0,0824
17	0,9181	0,0819	0,9147	0,0853	0,9176	0,0824

Berdasarkan Tabel 4, dapat dilihat bahwa nilai *k* terbaik adalah 5 dengan proporsi data latih banding data uji 60:40 dengan nilai *error rate* 0,0819. Nilai *k=5* digunakan untuk klasifikasi status kredit nasabah dan diperoleh Confusion Matrix dari data *testing* pada Tabel 5.

Tabel 5. *Confusion Matrix* Data *Testing* Klasifikasi KNN

		Kelas Aktual	
		Lancar	Berpotensi Macet
Kelas Prediksi	Lancar	157	14
	Berpotensi Macet	0	0

Evaluasi ketepatan hasil klasifikasi dapat dihitung berdasarkan Tabel 5 dengan perhitungan sebagai berikut:

$$Akurasi = \frac{0+157}{0+0+157+14} \times 100 = 91,81\%$$

$$Sensitivitas = \frac{0}{0+14} \times 100\% = 0\%$$

$$Spesifisitas = \frac{157}{157+0} \times 100\% = 100\%$$

$$G - mean = \sqrt{0 \times 1} = 0$$

Nilai akurasi sebesar 91,81% sehingga ketepatan model dalam memperkirakan data adalah 91,81%, nilai spesifisitas sebesar 100% interpretasinya data status kredit lancar diprediksi seluruhnya kedalam status kredit lancar, nilai sensitivitas sebesar 0% interpretasinya bahwa status kredit berpotensi macet tidak dapat diprediksi dengan benar sebagai status kredit berpotensi macet. Pada nilai *g-mean* 0 interpretasinya model tidak berhasil mengklasifikasikan secara tepat. Ketidak berhasilan kinerja metode klasifikasi dalam mengklasifikasikan status kredit nasabah dapat disebabkan karena adanya *imbalance class data*.

Permasalahan *imbalance class data* menyebabkan tidak berhasilnya model dalam mengklasifikasikan kelas minoritas padahal kelas tersebutlah yang akan diteliti lebih lanjut. Pada penelitian ini data status kredit nasabah lancar (mayoritas) memiliki jumlah yang jauh lebih banyak dibandingkan dengan status kredit berpotensi macet (minoritas). Evaluasi ketepatan hasil klasifikasi yang dihasilkan diperoleh nilai *g-mean* sebesar 0 yang dapat disimpulkan bahwa model tidak berhasil mengklasifikasikan secara tepat. Model tidak dapat memprediksi kelas berpotensi macet pada data status kredit nasabah dikarenakan tidak berhasilnya model dalam mengklasifikasikan kelas minoritas berakibat pada tidak teridentifikasinya nilai *g-mean*. Permasalahan *imbalance class data* dapat ditangani menggunakan SMOTE. Teknik SMOTE yang digunakan dalam penelitian ini membangkitkan data sintetis untuk kelas minoritas sebanyak 10 kali dari data latih minoritas (dipilih supaya banyaknya data minoritas setelah dilakukan SMOTE mendekati banyaknya data mayoritas). SMOTE dilakukan pada kelas minoritas pada data *training*. Hasil sebelum dan sesudah SMOTE disajikan pada Tabel 6.

Tabel 6. Perbandingan Data Latih Sebelum dan Sesudah SMOTE

Kategori	Proporsi 60		Proporsi 70		Proporsi 80	
	Sebelum	Sesudah	Sebelum	Sesudah	Sebelum	Sesudah
0 (Lancar)	237 (91,15%)	237 (50,75%)	276 (91,39%)	276 (51.49%)	316 (91,33%)	316 (51,30%)
1 (Berpotensi Macet)	23 (8,85%)	230 (49,25%)	26 (8,61%)	260 (48,51%)	30 (8,67%)	300 (48,70%)
Jumlah	260 (100%)	467 (100%)	302 (100%)	536 (100%)	346 (100%)	616 (100%)

Proses klasifikasi *K-Nearest Neighbor* dengan penanganan *imbalance class data* sama seperti tanpa penanganan. Nilai *k* yang digunakan dalam analisis ini yaitu 3, 5, 7, 9, 11, 13, 15,17,19, 21 (menurut Hassanat *et al* [9]). Nilai *k* yang digunakan masing-masing dihitung kembali nilai *error rate* untuk menentukan nilai *k* terbaik yang akan digunakan menggunakan bantuan *RStudio*. Selengkapnya disajikan pada Tabel 7.

Tabel 7. Nilai *Error Rate* Data *Training* pada *Imbalance Class Data*

Nilai k	Proporsi Data Latih : Data Uji					
	60:40		70:30		80:20	
	Akurasi	*Error Rate*	Akurasi	*Error Rate*	Akurasi	*Error Rate*
3	0.6316	0.3684	0.6899	0.3101	0.6118	0.3882
5	0.6199	0.3801	0.6822	0.3178	0.6118	0.3882
7	0.6784	0.3216	0.6977	0.3023	0.6471	0.3529
9	0.6901	0.3099	0.6899	0.3101	0.6471	0.3529
11	0.7018	0.2982	0.6899	0.3101	0.6353	0.3647
13	0.7135	0.2865	0.7054	0.2946	0.6588	0.3412
15	0.7076	0.2924	0.7132	0.2868	0.6471	0.3529
17	0.7135	0.2865	0.7054	0.2946	0.6353	0.3647
19	0.7076	0.2924	0.7287	0.2713	0.6353	0.3647
21	0,7076	0.2924	0.7287	0.2713	0.6471	0.3529

Berdasarkan Tabel 7 nilai *k* terbaik adalah 19 pada proporsi data 70:30 dengan nilai *error rate* 0,2713. Nilai *k* tersebut digunakan dalam penentuan klasifikasi status kredit nasabah yang menghasilkan hasil klasifikasi (berupa *Confusion Matrix*) dari data *testing* disajikan pada Tabel 8.

Tabel 7. *Confusion Matrix* Data *Testing* pada Data Setelah SMOTE

		Kelas Aktual	
		Lancar	Berpotensi Macet
Kelas Prediksi	Lancar	89	6
	Berpotensi Macet	29	5

Evaluasi ketepatan hasil klasifikasi dapat dihitung berdasarkan Tabel 7 dengan perhitungan sebagai berikut:

$$Akurasi = \frac{5+89}{5+29+89+6} \times 100 = 72,87\%$$

$$Sensitivitas = \frac{5}{5+6} \times 100\% = 45,46\%$$

$$Spesifisitas = \frac{89}{89+29} \times 100\% = 100\%$$

$$G - mean = \sqrt{0,4546 \times 7542} = 58,55\%$$

Evaluasi ketepatan hasil klasifikasi pada metode KNN setelah dilakukan SMOTE dengan *k*=19 diperoleh nilai akurasi sebesar 72,87% sehingga ketepatan model dalam memperkirakan data adalah 72,87%. Nilai spesifisitas sebesar 75,42% diartikan bahwa 75,42% data status kredit lancar diprediksi secara benar kedalam status kredit lancar. Nilai sensitivitas sebesar 45,46% diartikan bahwa 45,46% data status kredit berpotensi macet diprediksi secara benar kedalam status kredit berpotensi macet. Pada nilai *g-mean* diperoleh nilai 58,55% sehingga dapat ditarik kesimpulan bahwa sebesar 58,55% model dapat mengklasifikasikan secara tepat.

Hasil klasifikasi dengan perbandingan data latih dan data uji 70:30 dari kedua metode klasifikasi yaitu *K-Nearest Neighbor* tanpa penanganan dan dengan penanganan *imbalance class data* menggunakan SMOTE berdasarkan evaluasi ketepatan hasil klasifikasi disajikan pada Tabel 9.

Tabel 9**.** Evaluasi Ketepatan Hasil Klasifikasi Metode Klasifikasi

Evaluasi Ketepatan Hasil Klasifikasi	*K-Nearest Neighbor*	
	Tanpa SMOTE	SMOTE
Akurasi	91,81%	72,87%
Spesifisitas	100%	75,42%
Sensitivitas	0%	45,46%
G-mean	0%	58,55%

Evaluasi ketepatan hasil klasifikasi setelah penanganan menggunakan SMOTE paling baik diperoleh nilai k = 19 pada *K-Nearest Neighbor* dengan perbandingan data latih banding data uji sebesar 70:30. Nilai sensitivitas pada *K-Nearest Neighbor* setelah penanganan menggunakan SMOTE mengalami peningkatan sebesar 45,46%. Nilai spesifisitas sebelum dilakukan penanganan *imbalance class data* menggunakan SMOTE sebesar 100% setelah dilakukan penanganan sebesar 75,42%. Penerapan SMOTE metode *K-Nearest Neighbor* mampu meningkatkan nilai *g-mean* sebesar 58,55%. Nilai *g-mean* diperoleh metode klasifikasi *K-Nearest Neighbor* setelah penanganan *imbalance class data* menggunakan SMOTE menghasilkan nilai yang paling tinggi. Kesimpulan berdasarkan analisis kedua metode klasifikasi yang digunakan yaitu pada data status kredit nasabah Bank Perkreditan Rakyat "X" di regional provinsi Jawa Tengah diperoleh bahwa metode klasifikasi *K-Nearest Neighbor* dengan penanganan *imbalance class data* menggunakan SMOTE pada proporsi data latih banding data uji sebesar 70:30 pada nilai k = 19 merupakan metode klasifikasi yang lebih tepat digunakan karena menghasilkan nilai *g-mean* yang lebih baik.

4. KESIMPULAN

Penerapan klasifikasi status kredit nasabah dengan *K-Nearest Neighbor* pada data tanpa penanganan ketidakseimbangan data tidak berhasil mengklasifikasikan secara tepat karena nilai sensitivitas dan *g-mean* nol. Masalah ini diduga disebabkan karena adanya ketidakseimbangan pada data (*imbalance class data*). Permasalahan *imbalance class data* dapat ditangani menggunakan SMOTE. Penerapan SMOTE pada metode *K-Nearest Neighbor* mampu meningkatkan nilai *g-mean* sebesar 58,55%. Berdasarkan hasil klasifikasi disimpulkan bahwa K-Nearest Neighbor dengan SMOTE pada k=19 dan proporsi data latih terhadap data uji sebesar 70:30 merupakan metode klasifikasi yang lebih tepat digunakan dalam mengklasifikasikan data status kredit nasabah karena nilai *g-mean* yang lebih baik.

DAFTAR PUSTAKA

[1] Hasan, N. I. *Pengantar Perbankan.* Jakarta: Referensi (Gaung Persada Press Group). 2014.
[2] Han, J., dan Kamber, M. 2006. *Data Mining Concepts and Techniques Second Edition*. San Fransisco: Morgan Kaufmann.
[3] Singh, P. dan Sharma, P. A. Analysis of Imbalanced Classification Algorithms: A Perspective View. *International Journal of Trend in Scientific Research and Development.* 3(2): 974-978. 2019.
[4] Chawla, N. V., Bowyer, K. W., Hall, L. O., dan Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research.* 16:321-357. 2002
[5] Permana, T., Siregar, A. M., Masruriyah, A. F. N., dan Juwita, A. R. 2020. Perbandingan Hasil Prediksi Kredit Macet pada Koperasi Menggunakan Algoritma KNN dan C5.0. *Conference on Innovation and Application of Science and Technology (CIASTECH 2020*). 734-746. 2020
[6] Sharma, P., dan Kumar, D. Comparative Analysis of KNN and C5.0 Algorithm for Smart

City Classification. *International Journal of Engineering and Technical Research (IJETR)*. 7(4): 54-56. 2017

[7] Ramadhanti, D. V. *Perbandingan SMOTE dan ADASYN pada Data Imbalance untuk Klasifikasi Rumah Tangga Miskin di Kabupaten Temanggung dengan Algoritma K-Nearest Neighbor*. Skripsi. Semarang: Universitas Diponegoro (tidak dipublikasikan). 2022.

[8] Umma, F. N. *Klasifikasi Status Kemiskinan Rumah Tangga dengan Algoritma C5.0 di Kabupaten Pemalang*. Skripsi. Semarang: Universitas Diponegoro (tidak dipublikasikan). 2021.

[9] Gorunescu, F. 2011. *Data Mining: Concepts, Models and Techniques*. Berlin: *Springer.*

[10] Hassanat, A. B., Abbadi, M. A., dan Altarawneh, G. A. Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach. *International Journal of Computer Science and Information Security (IJCSIS)*, 12 (8):33-39. 2014.

[11] Tan, P., Steinbach, M., dan Kumar, V. 2006. *Introduction to Data Mining*. Boston: Pearson Education.

[12] Sreemathy, J., dan Balamurugan, P. S. An Efficient Text Classification using KNN and Naïve Bayesian. *International Journal on Computer Science and Engineering*. 4(3): 392-396. 2012

[13] Prasetyo, E. *Data Mining Konsep dan Aplikasi Menggunakan MATLAB*. Yogyakarta: ANDI Yogyakarta. 2012.

[14] Primartha, R. *Belajar Machine Learning Teori dan Praktik*. Bandung: Penerbit Informatika. 2018.

[15] He, H., dan Gracia, E.A. Learning from Imbalanced Data. *IEEE Trans*. *Knowl. Discov*. 21(9) 1263-1284. 2009.

[16] Kubat, M., Holte, R., dan Matwin, S. Learning When Negative Examples Abound. *In European conference on machine learning* (pp. 146-153). Springer, Berlin, Heidelberg. 1997.