ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 15, Nomor 1, Tahun 2026, Halaman 280 - 288
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

KLASIFIKASI MENGGUNAKAN ALGORITMA *K-NEAREST NEIGHBOR* DAN C5.0 PADA *IMBALANCE CLASS DATA* DENGAN SMOTE (Studi Kasus : Nasabah Bank Perkreditan Rakyat "X")

Salsabilla Rizka Ardhana[1*], Tatik Widiharih[2], Bagus Arya Saputra[3]
[1,2,3] Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
***e-mail**: salsabillarizka1801@gmail.com

DOI: 10.14710/j.gauss.15.1.280-288

Article Info:
Received: 2024-08-24
Accepted: 2025-12-31
Available Online: 2026-08-24

Keywords:
Credit Status; K-Nearest Neighbor; C5.0; Imbalance Class Data; SMOTE

Abstract: Rural Banks (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 potentially bad loans. K-Nearest Neighbor and C5.0 classify current and potentially 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. C5.0 can improve classification accuracy and work by calculating entropy, gain, split info, and gain ratio to form a decision tree. There is an imbalance class data which causes the classification process to focus more on the majority class. Imbalance class data is handled using 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 compared to accuracy, sensitivity and specificity in evaluating classification performance on imbalance class data. Result show an increased G-Mean to 58.55% on KNN and 64.05% on C5.0. Based on the classification results, it is concluded that C5.0 with SMOTE is a more appropriate classification model for customer credit status.

1. PENDAHULUAN

Bank Perkreditan Rakyat (BPR) adalah lembaga penyalur keuangan yang membantu usaha kecil dalam berkembang dan memberikan layanan kebutuhan perbankan bagi ekonomi lemah yang tidak dapat dijangkau oleh bank umum. Kegiatan utama bank sebagai lembaga keuangan yaitu pemberian kredit. 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 memperkirakan kelas dari objek yang label kelasnya tidak diketahui (Han dan Kamber, 2006). Metode klasifikasi yang digunakan dalam penelitian ini adalah *K-Nearest Neighbor* dan pohon keputusan C5.0. *K-Nearest Neighbor* menggunakan konsep kedekatan jarak sedangkan C5.0 membentuk suatu pohon keputusan berdasarkan perhitungan *entropy, information gain, split info* dan *gain ratio.*

Model klasifikasi yang baik dapat dilihat dari tingkat ketepatan dalam melakukan prediksi yang dapat dipengaruhi oleh ketidakseimbangan data. 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 (2019) klasifikasi cenderung memiliki tingkat ketepatan yang baik pada kelas mayoritas tetapi sangat buruk pada kelas minoritas. Permasalahan *imbalance class data* dapat ditangani menggunakan SMOTE yaitu pendekatan oversampling yang mana kelas minoritas akan dibuat data sintesis baru.

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* dan C5.0 tanpa penanganan serta penanganan *imbalance class data* menggunakan SMOTE.

2. TINJAUAN PUSTAKA

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 (Hasan, 2014).

Han dan Kamber (2006) mendefinisikan klasifikasi sebagai 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. 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 suara dari *k* tetangga terdekat tersebut. Label kelas hasil prediksi pada data uji ditetapkan dari kelas yang memiliki jumlah suara tetangga paling banyak (Tan, et al., 2006). *K-Nearest Neighbor* juga didasarkan pada pengukuran jarak. Penelitian ini menggunakan jarak *euclidean* untuk pengukuran jarak seperti pada Persamaan (1) berikut (Sreemathy, 2012).

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

dengan $d(x_i, y_i)$ = 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 $diff(x_{il}, y_{il})$ = selisih antara x_{il} dan y_{il}.

Menurut Prasetyo (2012), 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	$diff(x_{il}, y_{il}) = \begin{cases} 0, apabila\ x_{il} = y_{il} \\ 1, apabila\ x_{il} \neq y_{il} \end{cases}$
Ordinal	$diff(x_{il}, y_{il}) = \lvert x_{il} - y_{il} \rvert / (n-1)$ dengan n = banyaknya pengkategorian dalam x
Interval atau Rasio	$diff(x_{il}, y_{il}) = \lvert x_{il} - y_{il} \rvert$

Algoritma C5.0 merupakan salah satu algoritma klasifikasi pada pohon keputusan. Perhitungan C5.0 terdiri dari perhitungan kemunculan kejadian, perhitungan *entropy, information gain*, *split info*, dan *gain ratio*. Proses perhitungan C5.0 diuraikan dalam langkah-langkah berikut (Kantardzic, 2011).

1. Menghitung nilai *entropy* dari kategorik setiap variabel independen (*X*) dan variabel dependen (*Y*) dengan persamaan sebagai berikut.

$$Entropy\left(X_{ij}\right) = -(p_{lj} \times log_2 p_{lj} + p_{mj} \times log_2 p_{mj}) \tag{2}$$

dengan X_{ij} = variabel prediktor ke-*j* kategori ke-*i,* p_{lj} = peluang banyaknya data status kredit lancar pada masing-masing kategori variabel ke-*j* dan p_{mj} = peluang banyaknya data status kredit berpotensi macet pada masing-masing kategori variabel ke-*j*.

Perhitungan *entropy* juga dilakukan untuk pada variabel dependen dengan rumus sebagai berikut.

$$Entropy\ (Y) = -(p_l \times log_2 p_l + p_m \times log_2 p_m) \quad (3)$$

dengan Y = variabel dependen, p_l = peluang banyaknya data status kredit lancar dan p_m = peluang banyaknya data status kredit berpotensi macet.

2. Menghitung nilai *gain* dari setiap variabel dengan persamaan berikut.

$$Gain\ (X_{(i)}) = Entropy\ (Y) - \sum_{j=1}^{k} \frac{n(S_i)}{n(S)} \times Entropy\ (X_{ij}) \quad (4)$$

dengan $X_{(i)}$ = variabel prediktor ke-*i,* k = banyaknya kategori pada variabel prediktor $X_{(i)}$, $n(S_i)$ = jumlah sampel pada kategori ke-*i,* dan $n(S)$ = jumlah semua sampel pada himpunan data.

3. Menghitung nilai *split info* dan *gain ratio* dari setiap variabel. Perhitungan *gain ratio* selanjutnya menggunakan persamaan (5) berikut.

$$Gain\ Ratio = \frac{Gain\ (X_{(i)})}{SplitInfo(X_{(i)})} \quad (5)$$

$Gain\ (X_{(i)})$ = *information gain* pada variabel prediktor X dan $SplitInfo(X_{(i)})$ =*split information* pada variabel prediktor X.

Atribut untuk *node* (simpul) dipilih berdasarkan nilai *gain ratio* tertinggi. Pendekatan ini menerapkan normalisasi pada *information gain* menggunakan *split information* yang menyatakan informasi potensial. *SplitInfo* dapat dihitung dengan persamaan (6).

$$SplitInfo\ (X_{(i)}) = -\sum_{i=1}^{k} \frac{n(S_i)}{n(S)} log_2 \frac{n(S_i)}{n(S)} \quad (6)$$

Perhitungan *gain ratio* akan berhenti saat tidak ada lagi atribut yang dapat dipecah menjadi lebih kecil. Setiap keputusan cabang akan dilihat pada nilai *entropy*. Jika *entropy* pada cabang bernilai 0, maka keputusan yang diambil pada cabang tersebut sesuai definisi kategori awal.

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 mernggambarkan 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 (He dan Gracia, 2009). Permasalahan *imbalance class data* dapat diatasi menggunakan SMOTE (*Synthetic Minority Oversampling Technique*) yaitu metode oversampling pada kelas minoritas dengan membuat sampel sintetis. Pembangkitan data sintetis dilakukan dengan cara berikut.

a. Data Numerik

Data yang memiliki jarak terdekat kemudian digunakan untuk membangkitan data sintetis baru menggunakan persamaan (7) berikut

$$X_{baru_{[a]}} = x + (x^* - x) \times rand[0,1]_a \quad (7)$$

dengan $X_{baru_{[a]}}$= data sintetis hasil dari replikasi ke *a,* *a*= indeks replikasi (1,2,3,…….,$(n_{mayor} - n_{minor})$), 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.

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. 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*. Perhitungan tersebut ditentukan melalui *true positive* (TP), *true negative* (TN), *false positive*(FP), *false negative* (FN). Tabel 3. berikut merupakan *confusion matrix* 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

$$Akurasi = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

$$Error\ Rate = \frac{FP + FN}{TP + FP + TN + FN} \tag{9}$$

$$Sensitivitas = \frac{TP}{TP + FN} \tag{10}$$

$$Spesifisitas = \frac{TN}{TN + FP} \tag{11}$$

$$G - mean = \sqrt{Sensitivitas \times Spesifisitas} \tag{12}$$

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*. Kubat dan Matwin (1997) mengemukakan bahwa *g-mean* diartikan sebagai rata-rata *geometric* dari sensitivitas dan spesifisitas. *G-mean* akan bernilai satu apabila seluruh amatan diklasifikasikan dengan tepat.

3. METODE PENELITIAN

Data yang digunakan dalam penelitian ini adalah jenis data sekunder yaitu data nasabah Bank Perkreditan Rakyat (BPR) "X" di regional Provinsi Jawa Tengah pada bulan Oktober 2022. Variabel yang digunakan terdiri dari status kredit sebagai variabel dependen serta jenis kelamin, usia, jangka waktu, jumlah pinjaman, pendapatan, status perkawinan, pekerjaan, jenjang pendidikan, dan jenis jaminan. *Software* yang digunakan untuk pengolahan data adalah *Microsoft Excel* dan *RStudio* versi 4.1.0, dengan format file yang digunakan adalah (.xlsx). Langkah-langkah analisis pada penelitian ini diuraikan sebagai berikut.

1. Menginput dataset.
2. Melakukan analisis deskriptif.
3. Membagi data menjadi data latih dan data uji.

4. Melakukan klasifikasi status kredit nasabah dengan menggunakan metode *K-Nearest Neighbor* dan C5.0 serta *K-Nearest Neighbor* dan C5.0 dengan penanganan data tidak seimbang menggunakan SMOTE.
   a. Klasifikasi menggunakan algoritma *K-Nearest Neighbor* dengan tahapan sebagai berikut.
      i. Menentukan nilai *k* yang akan digunakan. Dalam analisis ini digunakan nilai *k* yaitu 3, 5, 7,9,11,13,15,17,19,21,23,25, dan 27.
      ii. Menghitung jarak (kedekatan lokasi) menggunakan jarak *Euclidean* dengan perhitungan *differencing* berdasarkan tipe data yang digunakan.
      iii. Mengurutkan nilai jarak dari terkecil hingga terbesar.
      iv. Menentukan jarak terdekat sebanyak nilai *k* yang ditetapkan.
      v. Menentukan kelas yang paling banyak muncul dari *k* tetangga terdekat sebagaitkelas dari data uji.
      vi. Menghitung *error rate* klasifikasi.
      vii. Melakukan kembali langkah ii-vi sampai didapatkan nilai *error rate* yang paling rendah.
      viii. Menghitung ketepatan hasil klasifikasi.
   b. Klasifikasi menggunakan algoritma C5.0 dengan tahapan sebagai berikut.
      i. Menghitung nilai *entropy* total dari semua kasus dan tiap atribut.
      ii. Menghitung nilai *gain* dan *split info*.
      iii. Menentukan cabang dari masing-masing *node* dengan menghitung nilai *gain ratio* tertinggi dari variabel prediktor.
      iv. Membagi kelas dalam cabang yang telah ditentukan.
      v. Melakukan kembali langkah i-iii sampai semua kelas pada cabang mempunyai kelas masing-masing.
      vi. Menghitung ketepatan hasil klasifikasi.
   c. Pembangkitan data sintetis baru menggunakan SMOTE dengan tahapan sebagai berikut.
      i. Menentukan data minoritas yang akan dibangkitkan
      ii. Menentukan nilai *k* yang digunakan. Pada *RStudio 4.1.0* nilai *deafault k* untuk SMOTE adalah 5
      iii. Melakukan perhitungan jarak data minoritas menggunakan jarak *euclidean*
      iv. Mengurutkan nilai jarak dari terkecil hingga terbesar.
      v. Menentukan kelas yang paling banyak muncul dari *k* tetangga terdekat sebagai kelas data hasil pembangkitan.
   d. Klasifikasi menggunakan algoritma *K-Nearest Neighbor* dan C5.0 dengan menggunakan penanganan SMOTE dilakukan dengan langkah yang sama seperti tanpa penanganan *imbalance class data* hanya saja kelas minoritas pada data latih dibangkitkan data sintetis baru menggunakan SMOTE.
5. Melakukan interpretasi hasil klasifikasi menggunakan metode *K-Nearest Neighbor* dan C5.0 tanpa penanganan dan dengan penanganan pada *imbalance class data* dengan SMOTE.

4. 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. Nilai *k* tetangga terdekat pada *K-Nearest* Neighbor yang digunakan dalam penelitian ini yaitu 3,5,7,9,11,13,15,17,19,21, 23, 25, dan 27. Nilai *k* yang digunakan masing-masing dihitung

error rate Nilai *error rate* yang paling kecil digunakan untuk menentukan nilai *k* terbaik yang akan digunakan dalam klasifikasi. Berdasarkan nilai *error rate* diperoleh bahwa nilai *k* terbaik adalah 5 dengan proporsi data latih banding data uji yaitu 60:40. Nilai *k* tersebut digunakan dalam penentuan klasifikasi status kredit nasabah yang menghasilkan evaluasi ketepatan hasil klasifikasi yaitu nilai akurasi sebesar 91,81% sehingga ketepatan model dalam memperkirakan data adalah 91,81%, nilai spesifisitas sebesar 100% diartikan data status kredit lancar diprediksi seluruhnya kedalam status kredit lancar, nilai sensitivitas sebesar 0% diartikan bahwa status kredit berpotensi macet tidak dapat diprediksi dengan benar sebagai status kredit berpotensi macet. Pada nilai *g-mean* diperoleh nilai 0 sehingga dapat ditarik kesimpulan bahwa model tidak berhasil mengklasifikasikan secara tepat berdasarkan nilai *g-mean*. Tidak berhasilnya kinerja metode klasifikasi dalam mengklasifikasikan status kredit nasabah dapat disebabkan karena adanya *imbalance class data*.

Analisis klasifikasi C5.0 dilakukan dengan perhitungan *entropy, information gain, split info,* dan *gain ratio* yang akan membentuk suatu pohon keputusan. Hasil klasifikasi terbaik pada perbandingan data latih dan data uji 60:40 dengan evaluasi ketepatan hasil klasifikasi diperoleh nilai akurasi sebesar 91,81% sehingga ketepatan model dalam memperkirakan data adalah 91,81%, nilai spesifisitas sebesar 100% diartikan bahwa data status kredit lancar diprediksi seluruhnya kedalam status kredit lancar, nilai sensitivitas sebesar 0% diartikan bahwa status kredit berpotensi macet tidak dapat diprediksi dengan benar sebagai status kredit berpotensi macet. Pada nilai *g-mean* diperoleh nilai 0 sehingga dapat ditarik kesimpulan bahwa model tidak berhasil mengklasifikasikan secara tepat berdasarkan nilai *g-mean*. Pohon yang dihasilkan hanya terdiri dari 1 node karena seluruh data latih diklasifikasikan ke dalam kelas lancar sehingga seluruh aturan yang dihasilkan pada pohon keputusan akan menghasilkan hasil klasifikasi pada kategori status kredit lancar. Maka dari itu, perlu dilakukan penanganan pada *imbalance class data* dengan menggunakan SMOTE.

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 kedua metode 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, seperti yang disajikan pada Tabel 3. berikut.

Tabel 3. 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%)

Data latih yang telah dilakukan teknik SMOTE kemudian akan digunakan untuk mengklasifikasikan status kredit nasabah menggunakan metode *K-Nearest Neighbor* dan

C5.0. Proses klasifikasi *K-Nearest Neighbor* pada penanganan *imbalance class data* sama seperti tanpa penanganan hanya saja kelas minoritas pada latih terlebih dahulu dibangkitkan data sintetis dengan teknik SMOTE. Nilai *k* yang digunakan dalam analisis dengan SMOTE yaitu 3, 5, 7, 9, 11, 13, 15,17,19, 21,23,25, dan 27. Nilai *k* terbaik dipilih berdasarkan *error rate* yang paling kecil yaitu sebesar *k=19* pada perbandingan data latih dan data uji sebesar 70:30. Nilai *k* tersebut digunakan dalam penentuan klasifikasi status kredit nasabah yang menghasilkan evaluasi ketepatan hasil klasifikasi sebesar 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 lancer, 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.

Klasifikasi C5.0 pada imbalance class data dilakukan setelah kelas minoritas pada data latih dibentuk data sintetis baru dengan menggunakan SMOTE. Dari hasil klasifikasi diketahui bahwa proporsi 80:20 menghasilkan nilai *g-mean* paling tinggi sehingga dipilih sebagai proporsi terbaik. Pembentukan pohon keputusan dilakukan dengan menghitung nilai *entropy, information gain, split info,* dan *gain ratio*. Evaluasi ketepatan hasil klasifikasi pada metode C5.0 setelah dilakukan SMOTE diperoleh nilai akurasi sebesar 70,59% sehingga ketepatan model dalam memperkirakan data adalah 70,59%, nilai spesifisitas sebesar 71,79% diartikan bahwa 71,79% data status kredit lancar diprediksi secara benar kedalam status kredit lancar, nilai sensitivitas sebesar 57,14% diartikan bahwa 57,14% data status kredit berpotensi macet diprediksi secara benar kedalam status kredit berpotensi macet. Pada nilai *g-mean* diperoleh nilai 64,05% sehingga dapat ditarik kesimpulan bahwa sebesar 64,05% model dapat mengklasifikasikan secara tepat. Pohon keputusan yang terbentuk terdiri dari *node* akar hingga daun dimana berupa aturan-aturan dalam mengklasifikasikan data. Jadi, dapat diartikan bahwa pohon yang terbentuk berupa pola status kredit nasabah yang merupakan relasi-relasi dari atribut yang menemui keputusan pada *node* terakhir. Hasil klasifikasi metode C5.0 menghasilkan pola berupa lintasan *tree* dari *node* akar hingga daun, Pola status kredit nasabah yang dihasilkan merupakan relasi berupa aturan IF…THEN yang akan menemui keputusan pada node terakhir. Penggunaan algoritma C5.0 menghasilkan ukuran tingkat kepentingan pada setiap variabel independen yang digunakan. Tingkat kepentingan menunjukkan seberapa besar pengaruh variabel independen dalam membangun model klasifikasi tersaji pada Tabel 4.

Tabel 4. Tingkat Kepentingan Variabel Independen pada Klasifikasi C5.0

Variabel Independen	Penggunaan Variabel
Pendidikan	100%
Jenis Jaminan	94,32%
Usia	77,60%
Status Perkawinan	68,51%
Pekerjaan	61,04%
Jumlah Pinjaman	49,03%
Pendapatan	22,73%
Jangka Waktu	22,24%

Berdasarkan Tabel 4. variabel jenjang pendidikan merupakan variabel yang paling berpengaruh pada model klasifikasi yang diartikan bahwa faktor jenjang pendidikan nasabah sangat berpengaruh terhadap status kredit nasabah. Pada pohon keputusan yang terbentuk,

penggunaan variabel jenjang pendidikan sebesar angka 100% yang menunjukkan bahwa pohon keputusan yang dihasilkan sepenuhnya memerlukan variabel jenjang pendidikan. Variabel penting selanjutnya adalah jenis jaminan sebesar 94,32% yang menunjukkan besar pengaruh variabel jenis jaminan terhadap pembentukan pohon keputusan. Persentase variabel-variabel independen lain selanjutnya menunjukkan seberapa besar pengaruh variabel tersebut terhadap pembangunan pohon keputusan.

Evaluasi ketepatan hasil klasifikasi K-Nearest Neighbor dan C5.0 tanpa penanganan dan dengan penanganan *imbalance class data* menggunakan SMOTE disajikan pada Tabel 5.

Tabel 5. Evaluasi Ketepatan Hasil Klasifikasi Metode Klasifikasi

Perbandingan Data Latih : Data Uji	Evaluasi Ketepatan Hasil Klasifikasi	*K-Nearest Neighbor*		C5.0	
		Tanpa SMOTE	SMOTE	Tanpa SMOTE	SMOTE
60 : 40	Akurasi	91,81%	71,35%	91,81%	71,93%
	Spesifisitas	100%	73,89%	100%	74,52%
	Sensitivitas	0%	42,86%	0%	42,86%
	G-mean	0%	56,28%	0%	56,51%
70 : 30	Akurasi	91,47%	72,87%	91,47%	67,44%
	Spesifisitas	100%	75,42%	100%	70,34%
	Sensitivitas	0%	45,46%	0%	36,36%
	G-mean	0%	58,55%	0%	50,57%
80 : 20	Akurasi	91,76%	65,88%	91,76%	70,59%
	Spesifisitas	100%	66,67%	100%	71,79%
	Sensitivitas	0%	57,14%	0%	57,14%
	G-mean	0%	61,72%	0%	**64,05%**

Berdasarkan Tabel 23. diperoleh bahwa pada penelitian ini nilai akurasi dari keempat model klasifikasi berkisar antara 65% sampai 92%. 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 serta 80:20 pada C5.0. Nilai sensitivitas pada *K-Nearest Neighbor* dan C5.0 setelah penanganan menggunakan SMOTE mengalami peningkatan sebesar 45,46% pada *K-Nearest Neighbor* dan 57,14% pada C5.0. Nilai spesifisitas pada kedua metode sebelum dilakukan penanganan *imbalance class data* menggunakan SMOTE sebesar 100% setelah dilakukan penanganan sebesar 75,42% pada *K-Nearest Neighbor* dan 71,79% pada C5.0. Penerapan SMOTE pada kedua metode mampu meningkatkan nilai *g-mean* sebesar 58,55% pada *K-Nearest Neighbor* dan 64,05% pada C5.0. Nilai *g-mean* diperoleh metode klasifikasi C5.0 setelah penanganan *imbalance class data* menggunakan SMOTE menghasilkan nilai yang paling tinggi.

Kesimpulan berdasarkan analisis keempat metode klasifikasi yang digunakan yaitu pada data status kredit nasabah Bank Perkreditan Rakyat "X" di regional provinsi Jawa Tengah diperoleh bahwa metode klasifikasi C5.0 dengan penanganan *imbalance class data* menggunakan SMOTE pada proporsi data latih banding data uji sebesar 80:20 merupakan metode klasifikasi yang lebih tepat digunakan karena menghasilkan nilai *g-mean* yang lebih baik dibandingkan metode lain yang digunakan.

5. KESIMPULAN

Penerapan klasifikasi status kredit nasabah dengan *K-Nearest Neighbor* dan C5.0 menghasilkan nilai akurasi sebesar 91,81%, nilai spesifisitas sebesar 100%, tetapi nilai spesifisitas dan *g-mean* bernilai 0 yang diartikan bahwa model tidak berhasil mengklasifikasikan secara tepat. Masalah tersebut bisa disebabkan karena adanya ketidakseimbangan pada data (*imbalance class data*). Pada penelitian status kredit nasabah terjadi *imbalance class data* dengan proporsi kelas minoritas (status kredit berpotensi macet)

sebesar 8,85%. Permasalahan *imbalance class data* dapat ditangani menggunakan SMOTE. Penerapan SMOTE pada kedua metode mampu meningkatkan nilai *g-mean* sebesar 58,55% pada *K-Nearest Neighbor* dan 64,05% pada C5.0. Berdasarkan hasil klasifikasi diperoleh metode klasifikasi C5.0 dengan SMOTE merupakan metode klasifikasi yang lebih tepat digunakan dalam mengklasifikasikan data status kredit nasabah karena nilai *g-mean* yang lebih baik dibandingkan model lain.

DAFTAR PUSTAKA

Chawla, N. V., Bowyer, K. W., Hall, L. O., dan Kegelmeyer, W.P. 2002. *SMOTE: Synthetic Minority Over-Sampling Technique*. Journal of Artificial Intelligence Research, 16, 321-357.

Han, J., dan Kamber, M. 2006. *Data Mining Concepts and Techniques Second Edition*. San Fransisco: Morgan Kaufmann.

Hasan, N. I. (2014). *Pengantar Perbankan.* Jakarta: Referensi (Gaung Persada Press Group).

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

Kantardzic, M. 2011. *Data Mining: Concept, Models, Methods, and Algorithms.* 2nd edition. New Jersey: John W & Sons, Inc.

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

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

Singh, P. dan Sharma, P. A. 2019. *Analysis of Imbalanced Classification Algorithms: A Perspective View*. International Journal of Trend in Scientific Research and Development, 3(2), 974-978.

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

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