ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 15, Nomor 1, Tahun 2026, Halaman 154 - 165
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

ANALISIS KLASIFIKASI MENGGUNAKAN REGRESI LOGISTIK BINER DAN *K-NEAREST NEIGHBOR* PADA DATA *IMBALANCE*

Eva Fitriyani[1*], Tatik Widiharih[2], Bagus Arya Saputra[3]
[1,2,3] Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
***e-mail**: evafitriyani42@gmail.com

DOI: 10.14710/j.gauss.15.1.154-165

Article Info:
Received: 2024-08-21
Accepted: 2025-12-20
Available Online: 2026-05-29

Keywords:
KSP; SMOTE; ADASYN; Binary Logistic Regression; K-Nearest Neighbor; Accuracy.

Abstract: Savings and Loan Cooperative or (KSP) is a cooperative that conducts its business activities only saving and borrowing. KSP members come from various different backgrounds so that they can affect their behavior in carrying out their obligations. To find out the status of current or bad customer payments, a classification process is carried out. The division of KSP customer data is carried out in the classification process into two, namely training data and test data. In the classification process, there are often cases of data imbalance, so it is necessary to handle data imbalance in training data with SMOTE and ADASYN. SMOTE and ADASYN were chosen because these methods handle imbalance data by generating data from minor classes so as not to eliminate important parts of the data. Classification was performed with Binary Logistic Regression and K-Nearest Neighbor. Binary Logistic Regression is a regression where the dependent variable is binary. While K-Nearest Neighbor is a grouping method based on the closeness of the distance of a data with other data as many as k nearest neighbors. The results of this study indicate that the ADASYN Binary Logistic Regression method is the best method that can classify and predict the payment status of KSP customers because it produces the highest accuracy and G-mean, namely the accuracy value of 70.67% and G-Mean 67.63%.

1. PENDAHULUAN

Koperasi Simpan Pinjam atau (KSP) merupakan koperasi yang pelaksanaan kegiatannya hanya simpan pinjam. Anggota KSP berasal dari berbagai latar belakang yang berbeda sehingga dapat mempengaruhi perilakunya dalam menjalankan kewajiban. Latar belakang yang berbeda seperti jenis kelamin, usia, status pernikahan, pekerjaan, dan lain-lain dapat digunakan untuk mengklasifikasikan karakteristik nasabah dengan status pembayaran lancar atau tidak lancar.

Klasifikasi merupakan metode penelitian yang mengevaluasi sebuah objek data agar sesuai dengan kelas tertentu di antara kelas-kelas yang tersedia (Prasetyo, 2012). Metode klasifikasi yang digunakan yaitu Regresi Logistik Biner dan *K-Nearest Neighbor*. Metode yang dipilih berdasarkan pada jenis data yang sesuai dengan metode klasifikasi. Data yang digunakan pada penelitian tugas akhir ini bersifat biner pada variabel dependennya, sehingga menggunakan metode Regresi Logistik Biner. KNN dipilih karena metode ini merupakan salah satu bentuk model pendukung keputusan yang dapat mengklasifikasikan data berdasarkan jarak terdekat. KNN mempunyai beberapa keunggulan, yaitu: 1) pelatihan dilakukan dengan cepat; 2) mudah dipelajari dan sederhana; 3) tahan terhadap data pelatihan yang memiliki derau; dan 4) efektif dilakukan terhadap data latih yang besar (Bhatia dan Vandana, 2010).

Proses klasifikasi sering ditemukan kasus ketidakseimbangan data sehingga dilakukan penanganan data *imbalance* pada data latih. Data *imbalance* merupakan kondisi ketidakseimbangan data antara kelas yang satu dengan lainnya. SMOTE dan ADASYN merupakan metode *oversampling* yang dapat digunakan untuk mengatasi data *imbalance*.

Penelitian ini bertujuan untuk menentukan ketepatan klasifikasi dengan akurasi dan *G-mean* untuk mengevaluasi kinerja algoritma klasifikasi khususnya dalam permasalahan ketidakseimbangan kelas pada dataset.

2. TINJAUAN PUSTAKA

Koperasi Simpan Pinjam adalah koperasi yang melakukan usahanya hanya simpan pinjam sebagai satu-satunya usaha (Undang-Undang Nomor 17 Tahun 2012). Dana yang dimiliki oleh lembaga keuangan dapat diperoleh dari dua asal, yakni melalui pinjaman serta melalui dana internal yang dimiliki. Modal yang diperoleh melalui pinjaman berasal dari para anggota, koperasi lain, dan lembaga keuangan seperti bank. Modal sendiri merupakan modal yang disumbangkan oleh anggota koperasi dalam berbagai bentuk, termasuk simpanan wajib, simpanan pokok, simpanan sukarela, dan subsidi.

Klasifikasi adalah model data diskriminasi kelas yang bermaksud untuk mengestimasi kategori suatu objek dengan label yang belum teridentifikasi (Han & Kamber, (2011). Permasalahan pada proses klasifikasi yang dapat terjadi yaitu data *imbalance*. Data *imbalance* adalah kondisi ketidakseimbangan data yang terjadi antara satu kelas dengan kelas data lainnya. Data tidak seimbang mengakibatkan algoritma tersebut menghasilkan akurasi yang buruk pada seluruh kelas data dan gagal dalam mewakili karakteristik data distribusif secara tepat (He & Gracia, 2009).

Synthetic Minority Over-Sampling Technique (SMOTE) adalah teknik untuk mengatasi masalah data *imbalance* dengan cara *oversampling* di kelas minoritas yaitu dengan menciptakan sampel sintetis yang diperkenalkan oleh Nithes V. Chawla. Data sintetis dibangkitkan menggunakan Persamaan (1) (Choi, 2010).

$$x_{syn_k} = x_i + (x_{knn} - x_i) \times \gamma_k \quad (1)$$

x_{syn_k} merupakan data yang telah disintesis melalui metode SMOTE, x_i merupakan mewakili data pengamatan ke-i dari kelompok minoritas. x_{knn} adalah data dari kelompok minoritas yang memiliki jarak paling dekat dengan x_i, dan gamma (γ) adalah suatu bilangan acak yang berada dalam rentang antara 0 dan 1.

Adaptive Syntethic Sampling Approach (ADASYN) adalah teknik penanganan data *imbalance* dengan melakukan *oversampling* terhadap kelas minoritas menggunakan bobot distribusi untuk data kelas minoritas yang didasari oleh tingkat kesulitan belajar pada model. Data sintetis yang dihasilkan oleh data minoritas lebih sulit untuk dimengerti dibandingkan dengan data minoritas yang lebih mudah untuk dimengerti (He et al, 2008). ADASYN mempunyai parameter untuk menetapkan keseimbangan yang diharapkan (β) dan batasnya didefinisikan sebagai tingkat toleransi maksimum rasio ketidakseimbangan kelas (d_{th}). Langkah untuk membangkitkan data sintetis dengan ADASYN adalah sebagai berikut:

1. Menentukan nilai parameter ADASYN, yaitu nilai d_{th} (nilai toleransi maksimum data *imbalance*) dan β (nilai keseimbangan)
2. Menghitung derajat keseimbangan $d = \frac{m(minoritas)}{m(mayoritas)}$
3. Menghitung jumlah data sintetis yang diperlukan untuk membuat data kelas minoritas G = $(mmayoritas - mminoritas) \times \beta$
4. Menghitung rasio berdasarkan *K-Nearest Neighbor* menggunakan *Euclidean distance* dengan rumus berikut. $r_i = \frac{\Delta_i}{k}$
5. Melakukan normalisasi r_i sehingga $\hat{r}_i$ merupakan distribusi kerapatan (*density distribution*) dengan persamaan sebagai berikut: $\hat{r}_i = \frac{r_i}{\sum_{i=1}^{m(minoritas)} r_i}$
6. Menghitung jumlah data sintetis yang akan dihasilkan untuk setiap data minoritas, $gi = \hat{ri} \times G$, i = 1, 2, 3,…, $m(minoritas)$

7. Membangkitkan sampel data sintetis menggunakan Persamaan (2).

$$s_k = x_i + (x_{zi} - x_i) \times \lambda_k \quad (2)$$

s_k merupakan data yang telah disintesis melalui metode ADASYN, di mana x_i mewakili data pengamatan ke-i dari kelas minoritas, x_{zi} merupakan data dari kelas minoritas yang mempunyai jarak paling dekat dengan x_i yang dipilih secara acak, dan lambda merupakan (λ) suatu angka acak yang berada dalam rentang antara 0 dan 1.

Metode regresi logistik merupakan suatu metode analisis statistika yang menganalisis hubungan antara variabel independent dan variabel dependen bersifat kategorikal (Hosmer dan Lemeshow, 2000). Model regresi logistik yang variabel dependennya terdiri dari dua kategori disebut model Regresi Logistik Biner (dikotomus). Model statistik variabel respon biner, dengan hasil respon untuk setiap subjek adalah "kesuksesan" atau "kegagalan" (Agresti, 2007). Model regresi logistiknya seperti pada Persamaan (3).

$$\pi(\boldsymbol{x_i}) = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi})} \quad (3)$$

Sedangkan logit dari $\pi(\boldsymbol{x_i})$ adalah :

$$g(\boldsymbol{x_i}) = ln\left[\frac{\pi(\boldsymbol{x_i})}{1-\pi(\boldsymbol{x_i})}\right] = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} \quad (4)$$

Metode penduga maksimum likelihood dapat digunakan untuk menentukan pendugaan parameter pada metode Regresi Logistik Biner (Agresti, 2007). Hosmer Lemeshow (2000) menyatakan, dimisalkan ada sampel dari n pengamatan independen $(x_i, y_i), i = 1,2, \ldots, n$ dengan y_i dinyatakan dari variabel respon biner dan x_i merupakan nilai dari variabel prediktor untuk subjek ke-i, maka fungsi dari likelihoodnya merupakan perkalian dari fungsi densitasnya masing-masing.

1. Uji Rasio Likelihood

   Pengujian Rasio Likelihood dilakukan untuk mengetahui besar pengaruh dari koefisien β terhadap variabel dependen secara bersama (Hosmer dan Lemeshow,2000).

   Hipotesis:

   $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ (Seluruh variabel independent tidak ada yang mempengaruhi variabel dependen)

   H_1: paling sedikit ada satu $\beta_j \neq 0$ dengan $j = 1,2, \ldots, p$ (Sedikitnya terdapat satu variabel independent yang mempengaruhi variabel dependen)

   Statistik Uji:

   $$G = -2\, In\left[\frac{\text{likelihood tanpa variabel independen}}{\text{likelihood dengan variabel independen}}\right] \quad (5)$$

   Kriteria penolakan : H_0 ditolak jika $G > \chi^2_{(p,\alpha)}$ atau nilai $p - value < \alpha$

2. Uji Wald

   Uji Wald memiliki tujuan untuk memahami dampak dari setiap koefisien β_j secara individual dengan membandingkannya terhadap kesalahan standar, sehingga memungkinkan penentuan variabel independen dalam model memiliki pengaruh yang signifikan terhadap variabel dependen (Hosmer dan Lemeshow, 2000).

   Hipotesis:

   $H_0: \beta_j = 0$ dengan $j = 1,2, \ldots, p$ (Tidak ada yang berpengaruh antara variabel independen dengan variabel dependennya)

   $H_1: \beta_j \neq 0$ dengan $j = 1,2, \ldots, p$ (Ada yang memberikan pengaruh antara variabel independen dengan variabel dependennya)

   Statistik uji :

   $$W_j = \left\{\frac{\widehat{\beta_j}}{se\,(\widehat{\beta_j})}\right\} \quad (6)$$

   Kriteria penolakan: H_0 ditolak jika $|W_j| > Z_{\alpha/2}$ atau nilai $p - value < \alpha$

3. Uji Hosmer dan Lemeshow (Uji Kesesuaian Model)

Hosmer dan Lemeshow (2000), menjelaskan jika uji kesesuaian model adalah pengujian yang dilakukan untuk mengidentifikasi ada atau tidak perbedaan antara prediksi dan hasil pengamatan (model cocok atau tidak).

Hipotesis:

H_0 : Model cocok

H_1 : Model tidak cocok

Statistik Uji :

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n'_k \bar{\pi}_k)^2}{(n'_k \bar{\pi}_k)(1-\bar{\pi}_k)} \quad (7)$$

Kriteria penolakan: H_0 ditolak jika $\hat{C} > \chi^2_{(\alpha,g-2)}$ atau nilai $p - value < \alpha$

K-Nearest Neighbor (KNN) adalah metode pengelompokan data yang didasari oleh kedekatan sebuah data dengan data lainnya (Prasetyo, 2012). Prinsip kerja KNN yaitu mengelompokkan data uji berdasarkan banyaknya k tetangga terdekat dari data latih. Jumlah tetangga terdekat (*nearest neighbor*) dinyatakan dengan nilai k yang digunakan. Pemilihan nilai k yang tepat merupakan hal yang sangat penting dalam metode KNN. Cara menetapkan nilai k yang paling baik bisa dilakukan dengan cara *trial and error*, seperti yang dilakukan oleh Bagaskoroo dkk. (2018) yang melakukan percobaan dengan beberapa nilai k untuk mendapatkan model dengan nilai k yang menghasilkan akurasi paling tinggi. Klasifikasi *K-Nearest Neighbor* umumnya juga didasarkan pada jarak *Euclidean* antara data uji dan data latih. Jarak *Euclidean* merupakan perhitungan yang mengukur jarak dua titik dalam *Euclidean* space yang mempelajari hubungan antara sudut dan jarak. Menurut Sreemathy (2012), peringkat k tetangga terdekat berdasarkan kemiripan dihitung menggunakan jarak *Euclidean* dengan rumus seperti pada Persamaan (8).

$$\mathrm{d}(x_{i,} y_{i,}) = \sqrt{\sum_{l=1}^{p} (diff(x_{il,} y_{il}))^2} \quad (8)$$

Perhitungan nilai ketidaksamaan (diff) bergantung terhadap tipe data yang digunakan. Perhitungan nilai ketidaksamaan dilakukan berdasarkan tipe setiap variabel seperti pada Tabel 1 (Prasetyo, 2012).

Tabel 1. Ketidaksamaan Dua Data dengan Satu Atribut

Tipe Atribut	Formula Jarak
Nominal	$diff(x_{il,} y_{il}) = \begin{cases} 0, jika\ x_{il} = y_{il} \\ 1, jika\ x_{il} \neq y_{il} \end{cases}$
Ordinal	$diff(x_{il,} y_{il}) = \lvert x_{il} - y_{il} \rvert$/(n-1) n merupakan jumlah pengkategorian dalam x
Interval atau Rasio	$diff(x_{il,} y_{il}) = \lvert x_{il} - y_{il} \rvert$

Evaluasi hasil dibutuhkan untuk mengevaluasi dan melihat kinerja model yang didasari oleh data yang digunakan. Cara untuk mengevaluasi ukuran kinerja algoritma adalah dengan menggunakan *confusion matrix* yang ditunjukkan seperti Tabel 2.

Tabel 2. *Confusion Matrix*

		Kelas Prediksi	
		+	-
Kelas Sebenarnya	+	TP *(True Positive)*	FN *(False Negative)*
	-	FP *(False Positive)*	TN *(True Negative)*

Evaluasi kinerja model dapat dihitung berdasarkan akurasi, spesifisitas, sensitivitas, dan *G-mean* dengan rumus seperti berikut:

$$accuracy = \frac{TP+TN}{Total} \quad (9)$$

$$specificity = \frac{TN}{TN+FP} \quad (10)$$

$$sensitivity = \frac{TP}{TP+FN} \quad (11)$$

$$G - mean = \sqrt{Sensitivity \times Specificity} \quad (12)$$

Akurasi merupakan nilai ketepatan sebagai ukuran model untuk menentukan seberapa akurat prediksi. Spesifisitas adalah perhitungan untuk mengevaluasi kemampuan model untuk memprediksi kelas negatif yang diklasifikasikan. Sensivisitas adalah perhitungan yang digunakan untuk mengevaluasi kemampuan prediksi dari kelas positif suatu model. *G-mean* adalah rata-rata *geometric* dari sensitivitas dan spesifisitas, jika seluruh pengamatan diklasifikasikan dengan benar maka G-mean akan memilki nilai satu (Kubat Matwin (1997).

3. METODE PENELITIAN

Penelitan ini menggunakan jenis data sekunder, yaitu data nasabah salah satu koperasi simpan pinjam (KSP) di kota Semaran.Variabel pada penelitian ini terdiri dari dua variabel yaitu variabel dependen dan variabel independen. Variabel dependen merupakan kategori status pembayaran nasabah yang dibagi menjadi dua yaitu kredit lancar dan kredit macet. Variabel independen terdiri dari 8 variabel yaitu jenis kelamin, usia, status pernikahan, pekerjaan, tingkat pendidikan, jangka waktu, pendapatan, dan status kepemilikan rumah.

Pengolahan data dalam penelitian ini menggunakan Program R Studio versi 4.2.2 dan versi 4.1.2 serta *Microsoft Excel* 2019, dengan format file yang digunakan berupa (.txt). Langkah-langkah analisis pada penelitian ini adalah sebagai berikut:

1. Masukkan dataset.
2. Melakukan pembagian data menjadi data latih dan data uji.
3. Melakukan penanganan data *imbalance* dengan SMOTE dan ADASYN pada data latih.
4. Menentukan klasifikasi status pembayaran nasabah Koperasi Simpan Pinjam (KSP) dengan menggunakan metode Regresi Logistik Biner.
   a. Melakukan pembentukan model Regresi Logistik Biner
   b. Dilakukan pengujian secara serentak dengan uji Rasio Likelihood
   c. Dilakukan pengujian parsial dengan uji Wald
   d. Melakukan pengujian Hosmer and Lemeshow
   e. Menetapkan model akhir dari Regresi Logistik Biner
   f. Melakukan perhitungan nilai $\pi(\boldsymbol{x_i})$. Nilai $\pi(\boldsymbol{x_i})$ merupakan peluang dari kredit macet.
      i. Nilai $\pi(\boldsymbol{x_i}) < 0{,}5$ maka masuk ke dalam kelas 0 (kredit lancar).
      ii. Nilai $\pi(\boldsymbol{x_i}) \geq 0{,}5$ maka masuk ke dalam kelas 1 (kredit macet).
   g. Membentuk *confusion matrix* dan dilanjutkan dengan perhitungan ketepatan klasifikasi model memakai data uji.
5. Melakukan klasifikasi menggunakan metode *K-Nearest Neighbor*.
   a. Menentapkan nilai k tetangga
   b. Menghitung jarak data uji dengan setiap data latih.
   c. Memilih dari data ke-i berdasarkan jarak terdekat sebanyak k tetangga.
   d. Menentukan kelas dari data ke-i.
      i. Apabila kelas terbanyak yang muncul 0 pada k tetangga terdekat maka masuk ke dalam kelas 0 sehingga dikategorikan menjadi kredit lancar.
      ii. Apabila kelas terbanyak yang muncul 1 pada k tetangga terdekat maka masuk ke dalam kelas 1 sehingga dikategorikan menjadi kredit macet.

e. Membentuk *confusion matrix* dan melakukan perhitungan ketepatan klasifikasi model menggunakan data uji.

6. Memilih metode terbaik
7. Menginterpretasikan hasil klasifikasi metode Regresi Logistik Biner dan *K-Nearest Neighbor*.

4. HASIL DAN PEMBAHASAN

Kategori nasabah dibagi menjadi 2 yaitu nasabah dengan status pembayaran lancar dan macet. Nasabah dengan status pembayaran lancar sebanyak 324 orang atau sebesar 85% dan nasabah dengan status pembayaran macet sebanyak 55 orang atau sebesar 15%. Penelitian ini menggunakan perbandingan 80:20, didapatkan hasil untuk data latih sebanyak 304 dan data uji sebanyak 75. Data tersebut dikunci dengan nama train.orig2321 dan test2321.

Klasifikasi nasabah kredit di KSP dilakukan dengan menggunakan Regresi Logistik Biner dan *K-Nearest Neighbor*. Permasalahan data *imbalance* sering terjadi pada klasifikasi, ketidakseimbangan pada penelitian ini sebesar 14,47% sehingga termasuk dalam permasalahan *imbalance* tipe sedang karena berada pada kisaran 1%-20%. Penanganan data *imbalance* dilakukan menggunakan SMOTE dan ADASYN pada data latih. Proses klasifikasi menggunakan Regresi Logistik Biner dan *K-Nearest Neighbor* dilakukan setelah melakukan penanganan data *imbalance*.

Model awal SMOTE Regresi Logistik Biner dibuat dengan menggunakan estimasi parameter berdasarkan Tabel 3 yang merupakan hasil uji Wald.

Tabel 3. Hasil Uji Wald Model Awal

Variabel	Estimate	Std. Error	Nilai $\mathbf{W_j}$	p-value	Keterangan
$x_1(2)$	-0,8761	0,2405	-3,643	0,000269	Signifikan
$x_2(2)$	-0,6825	0,2899	-2,354	0,018574	Signifikan
$x_2(3)$	-3,0762	1,2165	-2,529	0,011444	Signifikan
$x_3(2)$	-3,9295	1,1550	-3,402	0,000669	Signifikan
$x_4(2)$	-16,8897	2742,7386	-0,006	0,995087	Tidak Signifikan
$x_4(3)$	-16,9601	2761,9214	-0,006	0,995100	Tidak Signifikan
$x_4(4)$	0,9660	0,2290	4,218	2,47e-05	Signifikan
$x_4(5)$	-2,3153	1,0748	-2,154	0,031221	Signifikan
$x_5(2)$	18,9662	923,4585	0,021	0,983614	Tidak Signifikan
$x_5(3)$	13,5790	923,4578	0,015	0,988268	Tidak Signifikan
$x_5(4)$	12,5527	4901,7091	0,003	0,997957	Tidak Signifikan
$x_5(5)$	13,6197	923,4581	0,015	0,988233	Tidak Signifikan
$x_6(2)$	0,6055	0,2719	2,227	0,025938	Signifikan
$x_6(3)$	-20,2889	1890,0677	-0,011	0,991435	Tidak Signifikan
$x_7(2)$	-20,0388	846,6071	-0,024	0,981116	Tidak Signifikan
$x_7(3)$	-18,7871	846,6071	-0,022	0,982296	Tidak Signifikan
$x_8(2)$	-0,5241	0,3536	-1.482	0,138262	Tidak Signifikan
$x_8(3)$	-0,6944	0,7219	-0,962	0,336134	Tidak Signifikan
constant	5,5580	1252,8048	0,004	0,996460	-

Model awal SMOTE Regresi Logistik Biner yang terbentuk yaitu $\pi(\boldsymbol{x_i}) = \frac{e^{g(x_i)}}{1+e^{g(x_i)}}$ sebagai peluang dari kredit macet dengan

$g(\boldsymbol{x})$ = 5,5580 – 0,8761$x_1(2)$ – 0,6825$x_2(2)$ – 3,0762$x_2(3)$ – 3,9295$x_3(2)$ -16,8897$x_4(2)$ – 16,9601$x_4(3)$ + 0,9660$x_4(4)$ – 2,3153$x_4(5)$ + 18,9662$x_5(2)$ + 13,5790x_5 (3) + 12,5527$x_5(4)$ + 13,6197x_5 (5) + 0,6055$x_6(2)$ –20,2889x_6 (3) - 20,0388x_7 (2) – 18,7871x_7 (3) – 0,5241x_8 (2) – 0,6944x_8 (3).

Hipotesis Uji Rasio Likelihood:

$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ (Semua variabel independen tidak ada yang mempengaruhi variabel dependen)

H_1: paling sedikit ada satu $\beta_j \neq 0$ dengan $j = 1,2, \ldots, p$ (Paling sedikit terdapat satu variabel independen yang mempengaruhi variabel dependen)

Statistik Uji:

$G = -2\,In\left[\frac{\text{likelihood tanpa variabel independen}}{\text{likelihood dengan variabel independen}}\right]$= 191,3515927

Pada taraf signifikansi α = 5% H_0 ditolak karena G = 191,3515927 > $X^2_{(18,0.05)}$ = 28,86932. Sehingga disimpulkan jika sedikitnya terdapat satu variabel independen yang mempengaruhi variabel status pembayaran.

Hipotesis Uji Wald:

$H_0: \beta_j = 0$ dengan $j = 1,2, \ldots, p$ (Tidak ada yang berpengaruh antara variabel independen dengan variabel dependennya)

$H_1: \beta_j \neq 0$ dengan $j = 1,2, \ldots, p$ (Ada yang memberikan pengaruh antara variabel independen dengan variabel dependennya)

Statistik uji :

$$W_j = \left\{\frac{\widehat{\beta_J}}{se\,(\widehat{\beta_J})}\right\}$$

Tabel 3 menunjukkan hasil uji Wald pada taraf signifikansi α = 5% H_0 ditolak apabila $|W_j| > Z_{\alpha/2}$ = 1,96 atau H_0 ditolak apabila nilai p-value < α, variabel yang mempengaruhi secara signifikan adalah variabel jenis kelamin (x_1), usia (x_2), status pernikahan (x_3), pekerjaan (x_4), dan jangka waktu (x_6). Karena ada beberapa variabel independen yang tidak signifikan maka dibuat model baru dengan mengeluarkan variabel independent yang tidak signifikan.

Perbedaan diantara prediksi dengan hasil observasi dapat diketahui dengan menggunakan Uji Hosmer dan Lemeshow atau uji kecocokan model.

Hipotesis:

H_0 : Model cocok

H_1 : Model tidak cocok

Statistik Uji :

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n'_k\bar{\pi}_k)^2}{(n'_k\bar{\pi}_k)(1-\bar{\pi}_k)}$$

p-value = 0,05573

Pada taraf signifikansi α = 5% H_0 ditolak jika $\hat{C} > X^2_{(0.05,8)}$ = 15,50731 atau H_0 ditolak jika nilai p-value < α, H_0 diterima karena $\hat{C}$ =15,18< $X^2_{(\alpha,g-2)}$ = 15,50731 atau p-value = 0,05573 > α= 0,05. Oleh karena itu disimpukan jika model cocok.

Model akhir SMOTE Regresi Logistik Biner yang terbentuk adalah $\pi(\boldsymbol{x_i}) = \frac{e^{g(x_i)}}{1+e^{g(x_i)}}$ sebagai peluang dari kredit macet dengan

g($\boldsymbol{x}$) = -0,2527 – 0,7415$x_1(2)$ – 0,2065$x_2(2)$ – 0,9291$x_2(3)$ – 1,5774$x_3(2)$ – 14,9528$x_4(2)$ – 15,7150$x_4(3)$ + 0,9699$x_4(4)$ –0,3282$x_4(5)$ + 0,5022$x_6(2)$ – 14,8752$x_6(3)$.

Ketepatan klasifikasi menggunakan SMOTE Regresi Logistik Biner didapatkan dari perhitungan *confusion matrix* sebagai berikut:

$$accuracy = \frac{TP+TN}{Total} \times 100\% = 57{,}33\%$$

$$specificity = \frac{TN}{TN+FP} \times 100\% = 54{,}59\%$$

$$sensitivity = \frac{TP}{TP+FN} \times 100\% = 72{,}73\%$$

$$G - mean = \sqrt{Sensitivity \times Specificity} = \sqrt{0{,}7273 \times 0{,}5459} = 63{,}01\%$$

Model awal dari ADASYN Regresi Logistik Biner dibuat dengan menggunakan estimasi parameter berdasarkan Tabel 4 yang merupakan hasil uji Wald.

Tabel 4. Hasil Uji Wald Model Awal

Variabel	Estimate	Std. Error	Nilai $\mathbf{W_j}$	p-value	Keterangan
$\boldsymbol{x_1}(2)$	-0,07378	0,2299	-0,321	0,74824	Tidak Signifikan
$\boldsymbol{x_2}(2)$	-0,5898	0,826	-2,087	0,03687	Signifikan
$\boldsymbol{x_2}(3)$	-19,27	2.191	-0,009	0,99298	Tidak Signifikan
$\boldsymbol{x_3}(2)$	-19,21	1.711	-0,011	0,99104	Tidak Signifikan
$\boldsymbol{x_4}(2)$	-18,20	7.602	-0,002	0,99809	Tidak Signifikan
$\boldsymbol{x_4}(3)$	-18,11	7.522	-0,002	0,99808	Tidak Signifikan
$\boldsymbol{x_4}(4)$	0,6769	0,2478	2,731	0,00631	Signifikan
$\boldsymbol{x_4}(5)$	-31,42	1.916	-0,016	0,98692	Tidak Signifikan
$\boldsymbol{x_5}(2)$	16,84	3.441	0,005	0,99610	Tidak Signifikan
$\boldsymbol{x_5}(3)$	-0,5469	3.176	0,000	0,99986	Tidak Signifikan
$\boldsymbol{x_5}(4)$	-2,775	13.550	0,000	0,99984	Tidak Signifikan
$\boldsymbol{x_5}(5)$	-2,149	3.176	-0,001	0,99946	Tidak Signifikan
$\boldsymbol{x_6}(2)$	0,03653	0,2524	0,145	0,88494	Tidak Signifikan
$\boldsymbol{x_6}(3)$	-35,74	5.467	-0,007	0,99478	Tidak Signifikan
$\boldsymbol{x_7}(2)$	71,19	4.908	-0,015	0,98843	Tidak Signifikan
$\boldsymbol{x_7}(3)$	-69,11	4.908	-0,014	0,98877	Tidak Signifikan
$\boldsymbol{x_8}(2)$	-0,8122	0,3811	-2,131	0,03306	Signifikan
$\boldsymbol{x_8}(3)$	-2,582	1,143	-2,260	0,02381	Signifikan
constant	70,52	6.616	0,011	0,99150	-

Model awal ADASYN Regresi Logistik Biner yang terbentuk yaitu $\pi(\boldsymbol{x_i}) = \frac{e^{g(x_i)}}{1+e^{g(x_i)}}$ sebagai peluang dari kredit macet dengan

$$g(\boldsymbol{x}) = 70{,}52 - 0{,}07378x_1(2) - 0{,}5898x_2(2) - 19{,}27x_2(3) - 19{,}21x_3(2) - 18{,}20x_4(2) - 18{,}11x_4(3) + 0{,}6769x_4(4) - 31{,}42x_4(5) + 16{,}84x_5(2) - 0{,}5469x_5(3) - 2{,}775x_5(4) - 2{,}149x_5(5) + 0{,}03653x_6(2) - 35{,}74x_6(3) - 71{,}19x_7(2) - 69{,}11x_7(3) - 0{,}8122x_8(2) - 2{,}582x_8(3)$$

Hipotesis Uji Rasio Likelihood:

$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ (Semua variabel independen tidak ada yang mempengaruhi variabel dependen)

H_1: paling sedikit ada satu $\beta_j \neq 0$ dengan $j = 1,2, \ldots, p$ (Paling sedikit terdapat satu variabel independen yang mempengaruhi variabel dependen)

Statistik Uji:

$$G = -2\,In\left[\frac{\text{likelihood tanpa variabel independen}}{\text{likelihood dengan variabel independen}}\right] = 201{,}1268546$$

Pada taraf signifikansi α = 5% H_0 ditolak karena G = 201,1268546 $> X^2_{(18,0.05)}$ = 28,86932. Sehingga disimpulkan sedikitnya terdapat satu variabel independen yang mempengaruhi variabel status pembayaran.

Hipotesis Uji Wald:

$H_0: \beta_j = 0$ dengan $j = 1,2, \ldots, p$ (Tidak ada yang berpengaruh antara variabel independen dengan variabel dependennya)

$H_1: \beta_j \neq 0$ dengan $j = 1,2, \ldots, p$ (Ada yang memberikan pengaruh antara variabel independen dengan variabel dependennya)

Statistik uji :

$$W_j = \left\{ \frac{\widehat{\beta}_J}{se\,(\widehat{\beta}_J)} \right\}$$

Tabel 4 menunjukkan hasil uji Wald pada taraf signifikansi α = 5% H_0 ditolak apabila $|W_j| > Z_{\alpha/2}$ = 1,96 atau H_0 ditolak apabila nilai p-value < α, variabel yang mempengaruhi secara signifikan adalah variabel usia (x_2), pekerjaan (x_4), dan status kepemilikan rumah (x_8). Karena ada beberapa variabel independen yang tidak signifikan maka dibuat model baru dengan mengeluarkan variabel independent yang tidak signifikan.

Perbedaan diantara prediksi dengan hasil observasi diketahui dengan menggunakan Uji Hosmer dan Lemeshow atau uji kecocokan model .

Hipotesis:

H_0 : Model cocok

H_1 : Model tidak cocok

Statistik Uji :

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n'_k \bar{\pi}_k)^2}{(n'_k \bar{\pi}_k)(1 - \bar{\pi}_k)}$$

p-value = 0,08706

Pada taraf signifikansi α = 5% H_0 ditolak jika $\hat{C} > X^2_{(0.05,8)}$ = 15,50731 atau H_0 ditolak jika nilai p-value < α, H_0 diterima karena $\hat{C} = 13{,}803 < X^2_{(\alpha,g-2)}$ = 15,50731 atau p-value = 0,08706 > α= 0,05. Oleh karena itu disimpukan jika model cocok.

Model akhir ADASYN Regresi Logistik Biner yang terbentuk adalah $\pi(\boldsymbol{x_i}) = \frac{e^{g(x_i)}}{1+e^{g(x_i)}}$ sebagai peluang dari kredit macet dengan

g($\boldsymbol{x}$) = 0,1552 – 0,3006x_2(2) - 2,0462x_2(3) – 15,5234x_4(2) - 15,5735x_4(3) + 0,8043x_4(4) – 1,6490x_4(5) – 0,8623x_8(2) - 1.6577x_8(3).

Ketepatan klasifikasi menggunakan ADASYN Regresi Logistik Biner didapatkan dari perhitungan *confusion matrix* sebagai berikut:

$$accuracy = \frac{TP+TN}{Total} \times 100\% = 70{,}67\%$$

$$specificity = \frac{TN}{TN+FP} \times 100\% = 71{,}88\%$$

$$sensitivity = \frac{TP}{TP+FN} \times 100\% = 63{,}64\%$$

$$G - mean = \sqrt{Sensitivity \times Specificity} = \sqrt{0{,}6364 \times 0{,}7188} = 67{,}63\%$$

Klasifikasi dengan metode SMOTE K-Nearest Neighbor diawali dengan melakukan perhitungan jarak antara data uji dengan setiap data latih. Perhitungan jarak *Euclidean* dilakukan secara manual pada data uji ke-1 dengan nilai k yang dipilih sebesar 3. Berikut ini merupakan perhitungan jarak *Euclidean* secara manual.

Data latih ke-1: $x_{11} = 1, x_{12} = 1, x_{13} = 1, x_{14} = 2, x_{15} = 4, x_{16} = 1, x_{17} = 3, x_{18} = 1$

Data uji ke-1: $y_{11} = 2, y_{12} = 2, y_{13} = 1, y_{14} = 1, y_{15} = 3, y_{16} = 1, y_{17} = 3, y_{18} = 1$

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

$$= \sqrt{1^2 + 0{,}5^2 + 0^2 + 1^2 + 0{,}25^2 + 0^2 + 0^2 + 0^2} = 1{,}520691$$

Perhitungan jarak *Euclidean* untuk data uji ke-1 dengan semua data latih telah dilakukan dan didapatkan jarak terdekat dengan 3 tetangga terdekat data latih sebesar 0 dan 0,5. Data latih yang memiliki jarak dengan data uji sebesar 0 sebanyak 2 dan 0,5 sebanyak 22. Selanjutnya dilakukan voting mayoritas dengan k = 3.

Tabel 5. Jarak *Euclidean* data uji ke-1 dengan 3 tetangga terdekat

Data ke-	Jarak *Euclidean*	Kelas Aktual
189	0	0
207	0	0
⋮	⋮	⋮
47	0,5	0
75	0,5	0
510	0,5	1

Tabel 5 menunjukkan bahwa data uji ke-1 diprediksi masuk dalam kelas 0 (kredit lancar) karena merupakan kelas mayoritas. Kelas aktual 0 sebanyak 22 dan kelas aktual 1 sebanyak 16. Ketepatan klasifikasi kategori nasabah kredit menggunakan metode SMOTE *K-Nearest Neighbor* dengan k=3 diperoleh dari perhitungan *confusion matrix*.

$$accuracy = \frac{TP+TN}{Total} \times 100\% = 58{,}67\%$$

$$specificity = \frac{TN}{TN+FP} \times 100\% = 60{,}94\%$$

$$sensitivity = \frac{TP}{TP+FN} \times 100\% = 45{,}45\%$$

$$G - mean = \sqrt{Sensitivity \times Specificity} = \sqrt{0{,}4545 \times 0{,}6094} = 52{,}63\%$$

Nilai k optimal ditentukan dengan melakukan *trial error*, dalam penelitian ini diujicobakan untuk k = 3, 5, 7, 9,… , 19. Nilai k yang memiliki persentase kesalahan prediksi terkecil dan akurasi teringgi merupakan k optimal. Persentase kesalahan prediksi terkecil pada SMOTE *K-Nearest Neighbor* yaitu 0,4133333 dan akurasi tertinggi sebesar 0,5866667 dengan nilai k= 3.

Klasifikasi dengan metode ADASYN *K-Nearest Neighbor* diawali dengan melakukan perhitungan jarak antara data uji dengan setiap data latih. Dilakukan perhitungan jarak *Euclidean* secara manual pada data uji ke-1 dengan nilai k yang dipilih sebesar 3. Berikut ini merupakan perhitungan jarak *Euclidean* secara manual.

Data latih ke-1: $x_{1\,1} = 2, x_{1\,2} = 2, x_{1\,3} = 1, x_{1\,4} = 4, x_{1\,5} = 2, x_{1\,6} = 2, x_{1\,7} = 2,\ x_{1\,8} = 1$

Data uji ke-1: $y_{1\,1} = 2, y_{1\,2} = 2, y_{1\,3} = 1, y_{1\,4} = 1, y_{1\,5} = 3, y_{1\,6} = 1, y_{1\,7} = 3,\ y_{1\,8} = 1$

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

$$= \sqrt{0^2 + 0^2 + 0^2 + 1^2 + 0{,}25^2 + 0{,}5^2 + 0{,}5^2 + 0^2} = 1{,}25$$

Setelah perhitungan jarak *Euclidean* untuk data uji ke-1 jarak terdekat dengan 3 tetangga terdekat data latih sebesar 0 dan 0,5. Data latih yang memiliki jarak dengan data uji sebesar 0 sebanyak 2 dan 0,5 sebanyak 82. Selanjutnya dilakukan voting mayoritas dengan k = 3.

Tabel 6. Jarak *Euclidean* data uji ke-1 dengan 3 tetangga terdekat.

Data ke-	Jarak *Euclidean*	Kelas Aktual
233	0	0
251	0	0
⋮	⋮	⋮
2	0,5	1
31	0,5	1
488	0,5	1

Tabel 6 menunjukkan bahwa data uji ke-1 diprediksi masuk kelas 1 (kredit macet) karena merupakan kelas mayoritas. Kelas aktual 0 sebanyak 23 dan kelas aktual 1 sebanyak 61. Ketepatan klasifikasi kategori nasabahkredit menggunakan metode SMOTE *K-Nearest Neighbor* dengan k=3 diperoleh dari perhitungan *confusion matrix*.

$$accuracy = \frac{TP+TN}{Total} \times 100\% = 69{,}33\%$$

$$specificity = \frac{TN}{TN+FP} \times 100\% = 71{,}88\%$$

$$sensitivity = \frac{TP}{TP+FN} \times 100\% = 54{,}55\%$$

$$G - mean = \sqrt{Sensitivity \times Specificity} = \sqrt{0{,}5455 \times 0{,}7188} = 62{,}62\ \%$$

Nilai k optimal ditentukan dengan melakukan *trial error*, dalam penelitian ini diujicobakan untuk k = 3, 5, 7, 9,… , 19. Nilai k yang memiliki persentase kesalahan prediksi terkecil dan akurasi tertinggi merupakan k optimal. Persentase kesalahan prediksi terkecil pada ADASYN *K-Nearest Neighbor* yaitu 0,3066667 dan akurasi tertinggi sebesar 0,6933333 dengan nilai k= 3.

5. KESIMPULAN

Klasifikasi menggunakan Regresi Logistik Biner dapat diterapkan dengan baik dan dapat menunjukkan variabel yang berpengaruh secara signifikan terhadap status pembayaran nasabah KSP. Variabel yang mempengaruhi secara signifikan pada klasifikasi dengan SMOTE Regresi Logistik Biner yaitu jenis kelamin, status perkawinan, pekerjaan, dan jangka waktu sedangkan pada metode ADASYN Regresi Logistik Biner yaitu usia, pekerjaan, dan status kepemilikan rumah. Hasil klasifikasi menggunakan *K-Nearest Neighbor* menunjukkan bahwa algoritma *K-Nearest Neighbor* telah berhasil diimplementasikan dengan baik dan didapatkan nilai k optimal sebesar 3 baik pada metode SMOTE *K-Nearest Neighbor* maupun metode ADASYN K-Nearest Neighbor. Hasil ketepatan klasifikasi menunjukkan jika metode ADASYN Regresi Logistik Biner merupakan metode terbaik yang dapat mengklasifikasi dan memprediksi status pembayaran nasabah KSP karena menghasilkan akurasi dan $G - mean$ tertinggi yaitu nilai akurasi sebesar 70,67% dan G-Mean 67,63%.

DAFTAR PUSTAKA

Agresti, A. 2007. *An Introduction to Categorical Data Analysis*. New York: John Wiley and Sons.

Bagaskoro, G., N., Fauzi, M.A., dan Adikara, P.P. 2018. Penerapan Klasifikasi Tweets pada Berita Twitter Menggunakan Metode K-Nearest Neighbor dan Query Expansion Berbasis Distributional Semantic. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Vol. 2, No. 10 Hal. 3849-3855.

Bhatia, M., Vandana., 2010. *Survey of Nearest Neighbor Techniques*. *International Journal of Computer Science and Information Security* 8, 1947-5500.

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.*

Choi, J. M. 2010. *A Selective Sampling Method for Imbalanced Data Learning on Support Vector Machines*. Graduate Theses and Dissertations, Paper 11529.

Han J, Kamber M, J. P. 2011. *Data Mining Concept and Techniques Third Edition*

He, H., Bai, Y., Garcia, E. A., dan Li, S. 2008. ADASYN: *Adaptive Syntethic Sampling Approach for Imbalanced Learning*. In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence) (pp. 1322-1328). IEEE.

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

Hosmer, D. W., dan Lemeshow, S. 2000. *Apllied Logistic Regression*. New York: John Wiley & Sons.

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.

Republik Indonesia. 2012. Undang-Undang Republik Indonesia Nomor 17 Tahun 2012. *Tentang Perkoperasian*. Pemerintah Pusat.

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