DOI: https://doi.org/10.26593/jrsi.v12i1.6380.1-10

Jurnal Rekayasa Sistem Industri
Volume 12 No 1 - April 2023
http://journal.unpar.ac.id/index.php/jrsi/index
ISSN 2339-1499 (*online*) – ISSN 0216-1036 (*print*)

Disampaikan : 08 Januari 2023
Direview : 19 Januari 2023
Diterima : 27 Maret 2023

Oversampling Sintetis Berbasis Kopula untuk Model Klasifikasi dengan Data yang Tidak Seimbang: Studi Kasus Prediksi Kredit Macet Nasabah Kartu Kredit

Fransiscus Rian Pratikto[1]
[1)] Fakultas Teknologi Industri, Jurusan Teknik Industri, Universitas Katolik Parahyangan
Jl. Ciumbuleuit 94, Bandung 40141
Email: frianp@unpar.ac.id

Abstract

A machine learning classification model for detecting abnormality is usually developed using imbalanced data where the number of abnormal instances is significantly smaller than the normal ones. Since the data is imbalanced, the learning process is dominated by normal instances, and the resulting model may be biased. The most common method for coping with this problem is synthetic oversampling. Most synthetic oversampling techniques are distance-based, usually based on the k-Nearest Neighbor method. Patterns in data can be based on distance or correlation. This research proposes a synthetic oversampling technique that is based on correlations in the form of the joint probability distribution of the data. The joint probability distribution is represented using a Gaussian copula, while the marginal distribution uses three alternatives distribution: the Pearson distribution system, empirical distribution, and the Metalog distribution system. This proposed technique is compared with several commonly used synthetic oversampling techniques in a case study of credit card default prediction. The classification model uses the k-Nearest Neighbor and is validated using the k-fold cross-validation. We found that the classification model using the proposed oversampling method with the Metalog marginal distribution has the greatest total accuracy.

Keywords: *synthetic oversampling, copula, classification model, Metalog distribution, k-Nearest Neighbor*

Abstrak

Model klasifikasi berbasis pembelajaran mesin untuk mendeteksi anomali biasanya didasarkan pada data dengan proporsi yang tidak seimbang. Proporsi data anomali biasanya jauh lebih kecil dibandingkan proporsi data non anomali. Ketidakseimbangan data menyebabkan model klasifikasi lebih banyak melakukan pembelajaran dengan data non anomali sehingga model bisa bias. Salah satu metode yang banyak digunakan untuk mengatasi masalah ini adalah oversampling sintetis. Oversampling sintetis umumnya didasarkan pada jarak dan didominasi metode berbasis *k-Nearest Neighbor*. Secara umum, pola data bisa berdasarkan jarak atau hubungan korelasional. Penelitian ini bertujuan menawarkan metode *oversampling* sintetis berdasarkan hubungan korelasional dalam bentuk distribusi probabilitas bersama dari data aslinya. Distribusi probabilitas bersama direpresentasikan dengan kopula Gaussian, sedangkan distribusi probabilitas marjinalnya direpresentasikan menggunakan tiga alternatf distribusi, yaitu sistem distribusi Pearson, distribusi empiris, dan sistem distribusi Metalog. Metode ini dibandingkan dengan beberapa metode oversampling lain yang umum digunakan untuk data yang tidak seimbang. Implementasi dilakukan dalam masalah kredit macet nasabah kartu kredit di suatu bank dengan metode klasifikasi *k-Nearest Neighbor* dengan ukuran kinerja akurasi total dengan metode validasi *k-fold cross validation*. Didapati bahwa model klasifikasi dengan metode *oversampling* usulan menggunakan distribusi marjinal Metalog memiliki akurasi total tertinggi.

Kata kunci: *oversampling* sintetis, kopula, model klasifikasi, distribusi Metalog. K-Nearest Neighbor

DOI: https://doi.org/10.26593/jrsi.v12i1.6380.1-10

Pendahuluan

Model klasifikasi berbasis pembelajaran mesin (*machine learning*) banyak digunakan untuk mendeteksi abnormalitas. Model seperti ini biasanya bertujuan mendeteksi secara dini abnormalitas yang mungkin terjadi sehingga bisa dilakukan langkah-langkah pencegahan atau antisipasi. Dalam bidang perbankan, model ini banyak digunakan untuk mendeteksi transaksi *fraud*, *credit scoring*, dan deteksi kredit macet. Model seperti ini biasanya dibangun menggunakan data dengan proporsi yang tidak seimbang, di mana banyaknya data yang tergolong abnormal/anomali (status transaksi fraud atau kredit macet) jauh lebih sedikit dibanding data kategori normal/non-anomali. Ketidakseimbangan ini menyebabkan model yang dihasilkan cenderung lebih akurat dalam memprediksi kejadian normal namun kurang akurat memprediksikan kejadian abnormal/anomali. Ini berdampak relatif tingginya false positive yang, dalam konteks kredit scoring misalnya, berpotensi merugikan bank karena nasabah yang diprediksi tidak macet ternyata kemudian macet.

Cara yang umum digunakan untuk mengatasi masalah ini adalah *resampling*, memodifikasi algoritma klasifikasi, pendekatan berbasis biaya, dan *ensemble* (Patel et al., 2020). *Resampling* pada dasarnya bertujuan menyeimbangkan data; ini bisa dilakukan dengan *undersampling* terhadap data yang proporsinya lebih besar dan/atau *oversampling* terhadap data yang proporsinya lebih kecil. Strategi modifikasi algoritma dilakukan dengan mengubah algoritma klasifikasi sehingga mampu menangani data yang tidak seimbang. Sementara itu, pendekatan berbasis biaya memperhitungkan konsekuensi biaya akibat kesalahan klasifikasi, baik yang berupa *false positive* maupun *false negative*, sedangkan strategi *ensemble* menggabungkan beberapa pendekatan sekaligus untuk menghasilkan model yang lebih akurat.

Resampling merupakan metode yang paling banyak digunakan dengan *oversampling* sintetis menjadi yang paling populer. Dalam *oversampling* sintetis, data baru (sintetis) dibuat berdasarkan data awal menggunakan sebuah algoritma dan kemudian digabungkan dengan data awal menjadi dataset dengan jumlah yang lebih besar. Metode *oversampling* sintetis sebagian besar didasarkan pada jarak yang umumnya didasarkan metode *k-Nearest Neighbor* (Cover & Hart, 1967). Salah satu penelitian tentang *resampling* melakukan eksperimen untuk meneliti pengaruh rasio ketidakseimbangan dan mendapati bahwa *undersampling* dan *oversampling* sama baiknya jika rasio ketidakseimbangan rendah (García et al., 2012), sementara jika rasionya tinggi *oversampling* lebih baik. Penelitian lain (Menardi & Torelli, 2014) menyarankan penggunaan metode *resampling* yang mirip dengan *boosting* dan *bagging* untuk meningkatkan akurasi jika ketidakseimbangan sangat tinggi.

Beberapa penelitian mengembangkan metode *resampling* seperti *synthetic minority oversampling technique* atau SMOTE (Chawla et al., 2002), *adaptive synthetic sampling approach for imbalanced learning* atau ADASYN (Haibo He et al., 2008), dan *random-walk oversampling approach* atau RWO (Zhang & Li, 2014). Sementara itu, penelitian lain melakukan modifikasi atau kombinasi dari teknik-teknik yang sudah ada sebelumnya, seperti Wang et al. (2014) yang mengkombinasikan SMOTE dan *particle swarm optimization* (PSO) ke dalam metode klasifikasi C5, dan Sáez et al. (2015) yang memodifikasi SMOTE menjadi SMOTE-IPF dengan mempertimbangkan data di perbatasan dan derau (*noise*) sehingga batas antar kelas menjadi lebih jelas. Selain itu terdapat penelitian dari Tahir et al. (2012) yang mengembangkan strategi *inverse random* untuk *undersampling*, dan penelitian Wong et al. (2014) yang menggunakan *fuzzy logic* dalam pengambilan sampel dari kelas mayoritas.

Penelitian ini mengusulkan metode *oversampling* sintetis yang didasarkan pada distribusi probabilitas. Karena data bersifat multivariabel, distribusi probabilitas dinyatakan dengan distribusi probabilitas bersama (*joint probability distribution*) berbentuk kopula Gaussian (Durante & Sempi, 2016). Kopula dipilih untuk mengakomodasi setiap data memiliki distribusi probabilitas marjinal yang berbeda-beda. Distribusi marjinalnya akan menggunakan dua alternatif, yaitu distribusi empiris dan sistem distribusi Metalog (Keelin, 2016). Sistem distribusi Metalog dikembangkan sebagai salah satu upaya untuk merepresentasikan kejadian probabilistik dengan cara yang lebih sederhana. Dalam sistem distribusi ini, setiap distribusi probabilitas

DOI: https://doi.org/10.26593/jrsi.v12i1.6380.1-10

teoritis direpresentasikan menggunakan deret terhingga fungsi logistik, sehingga setiap distribusi memiliki bentuk fungsi dasar yang sama. Dengan cara ini formulasi masalah dalam model-model probabilistik menjadi lebih sederhana.

Metode *oversampling* usulan ini akan dibandingkan dengan beberapa metode oversampling sintetis yang banyak digunakan mengatasi masalah data yang tidak seimbang, yaitu SMOTE, ANSMOTE, Borderline SMOTE, Density Based SMOTE, Safe Level SMOTE, Relocating Safe Level SMOTE, dan ADASYN. Perbandingan akan dilakukan dengan menerapkan metode usulan dan semua metode pembanding tersebut dalam contoh kasus prediksi kredit macet nasabah kartu kredit di suatu bank. Perbandingan akan dilakukan dengan menerapkan metode *oversampling* ke model klasifikasi berbasis *k-Nearest Neighbor*. Metode klasifikasi ini sengaja dipilih karena semua metode oversampling pembanding tersebut didasarkan pada metode klasifikasi ini sehingga jika metode *oversampling* usulan memiliki kinerja lebih baik maka itu disebabkan keunggulan metode *oversampling* tersebut, bukan karena pengaruh metode klasifikasinya.

Secara umum, pola data bisa didasarkan pada jarak atau hubungan korelasional. Saat ini belum ada metode oversampling sintetis yang didasarkan pada hubungan korelasional. Adanya metode oversampling sintetis berbasis hubungan korelasional akan memberi alternatif cara untuk mengatasi ketidakseimbangan data. Kebaruan penelitian ini adalah penggunaan metode *oversampling* berdasarkan hubungan korelasional yang direpresentasikan dengan distribusi probabilitas bersama, alih-alih jarak, sebagai dasar *resampling*. Penggunaan sistem distribusi Metalog untuk merepresentasikan distribusi marjinal juga merupakan hal baru dalam *data analytics*. Walaupun perbandingan dengan metode *oversampling* lain dilakukan menggunakan kasus dengan variabel kontinu, metode ini bisa digunakan untuk kasus dengan data diskrit.

Metodologi

Sistem Distribusi Pearson

Penentuan distribusi probabilitas dari suatu data dimulai dengan penentuan tipe distribusi dan dilanjutkan dengan uji kebaikan suai (*goodness of fit test*). Sistem distribusi Pearson merupakan sistem distribusi yang paling banyak digunakan saat ini. Dalam sistem distribusi ini, tipe distribusi dari suatu data dapat diestimasi berdasarkan nilai *skewness* kuadrat dan kurtosis seperti dalam Gambar 1.

Dalam sistem distribusi ini terdapat 12 keluarga distribusi yang fungsi matematisnya dapat diturunkan dari persamaan diferensial yang sama yang merupakan fungsi dari *skewness* kuadrat (β_1) dan kurtosis β_2 . Fungsi *density* dalam sistem distribusi Pearson $f(x)$ merupakan solusi dari persamaan diferensial berikut (Pearson, 1895):

$$\frac{f'(x)}{f(x)} + \frac{a+(x-\mu)}{b_0+b_1(x-\mu)+b_2(x-\mu)^2} = 0 \qquad \text{Pers. 1}$$

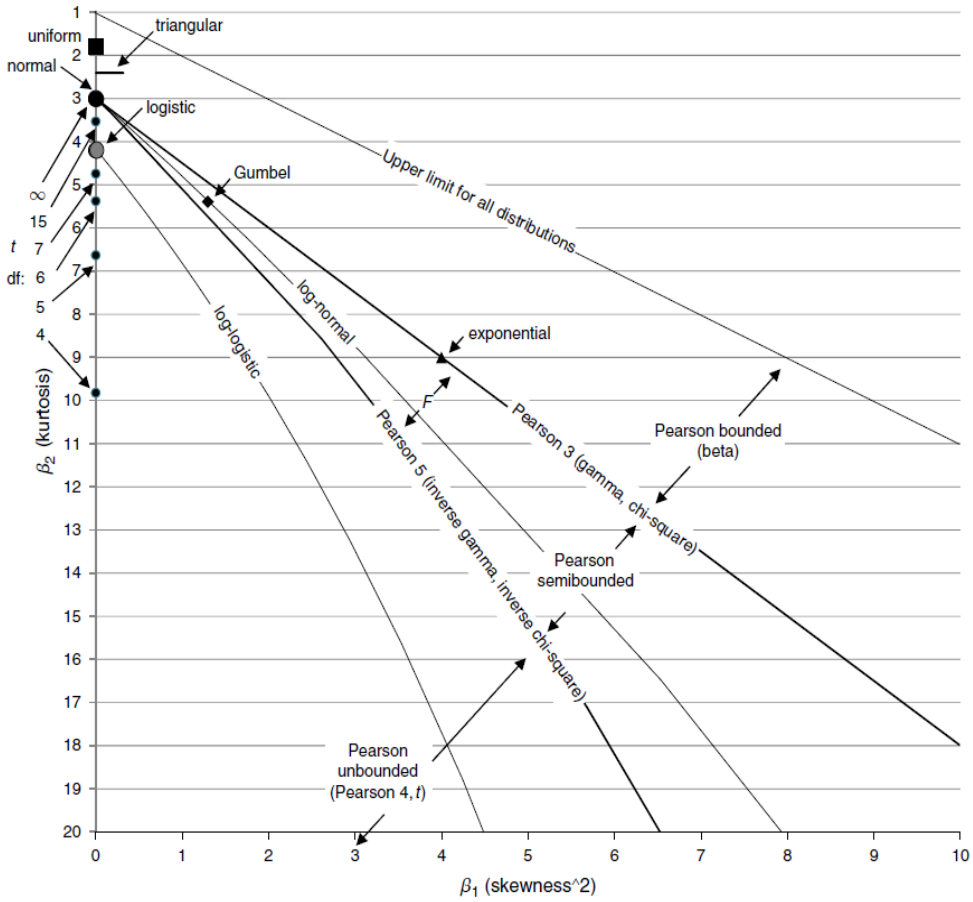
di mana

$$b_0 = \frac{4\beta_2-3\beta_1}{10\beta_2-12\beta_1-18}\mu_2$$
$$a = b_1 = \sqrt{\mu_2}\sqrt{\beta_1}\frac{\beta_2+3}{10\beta_2-12\beta_1-18}$$
$$b_2 = \frac{2\beta_2-3\beta_1-6}{10\beta_2-12\beta_1-18}$$

Sistem Distribusi Metalog

Sistem distribusi Metalog yang mencakup distribusi kontinu digunakan untuk merepresentasikan data acak ketika proses yang mendasarinya sulit dikarakterisasi sehingga fungsi distribusi teoritis yang bersesuaian sulit ditentukan (Keelin, 2016). Suatu distribusi probabilitas dinyatakan dengan dua fungsi, yaitu *probability density function* (PDF) dan fungsi *quantile* atau *inverse cumulative distribution function* (*inverse* CDF). Dalam sistem distribusi Metalog, fungsi *quantile* dan PDF-nya berbentuk deret terhingga dari fungsi logistik. Ini membuat sistem distribusi ini diklaim lebih sederhana secara matematis, lebih fleksibel dalam segi bentuk dan *boundedness*, lebih mudah dalam uji kebaikan suai, dan lebih mudah digunakan dalam simulasi karena memiliki fungsi *quantile* dalam bentuk tertutup (*closed-form*).

DOI: https://doi.org/10.26593/jrsi.v12i1.6380.1-10



Gambar 1. Grafik Cullen-Frey yang menggambarkan sistem distribusi Pearson (Keelin, 2016)

Misalkan x dan y merepresentasikan koordinat dari CDF di mana x adalah data dan $0 < y < 1$ adalah probabilitas kumulatif. Jika $\mathbf{x} = (x_1, \dots, x_m)$ dan $\mathbf{y} = (y_1, \dots, y_m)$ dengan $m \geq n$, fungsi umum *quantile* distribusi Metalog dengan n suku, dinotasikan dengan M_n adalah sebagai berikut (Keelin, 2016):

$$M_2(y: \mathbf{x}, \mathbf{y}) = a_1 + a_2 \ln\frac{y}{1-y} \qquad \text{Pers. 2}$$

$$M_3(y: \mathbf{x}, \mathbf{y}) = a_1 + a_2 \ln\frac{y}{1-y} + a_3(y-0{,}5)\ln\frac{y}{1-y} \qquad \text{Pers. 3}$$

$$M_4(y: \mathbf{x}, \mathbf{y}) = a_1 + a_2 \ln\frac{y}{1-y} + a_3(y-0{,}5)\ln\frac{y}{1-y} + a_4(y-0{,}5) \qquad \text{Pers. 4}$$

Sementara itu, untuk jumlah suku ganjil $n \geq 5$ fungsi *quantile*-nya adalah sebagai berikut:

$$M_n(y: \mathbf{x}, \mathbf{y}) = \mathrm{M}_{n-1} + a_n(y-0{,}5)^{(n-1)/2} \qquad \text{Pers. 5}$$

Sedangkan fungsi *quantile* untuk jumlah suku genap $n \geq 6$ adalah:

$$M_n(y: \mathbf{x}, \mathbf{y}) = \mathrm{M}_{n-1} + a_n(y-0{,}5)^{n/2-1}\ln\frac{y}{1-y} \qquad \text{Pers. 6}$$

Vektor $\mathbf{a} = (a_1, \dots, a_n)$ ditentukan dengan

$$\mathbf{a} = [\mathbf{Y}_n'\mathbf{Y}_n]^{-1}\mathbf{Y}_n'\mathbf{x} \qquad \text{Pers. 7}$$

dengan $\mathbf{Y}_n$ adalah matriks $m \times n$ berikut:

$$\mathbf{Y}_2 = \begin{bmatrix} 1 & \ln\frac{y_1}{1-y_1} \\ \vdots & \vdots \\ 1 & \ln\frac{y_m}{1-y_m} \end{bmatrix} \qquad \text{Pers. 8}$$

$$\mathbf{Y}_3 = \begin{bmatrix} 1 & \ln\frac{y_1}{1-y_1} & (y_1-0{,}5)\ln\frac{y_1}{1-y_1} \\ \vdots & \vdots & \vdots \\ 1 & \ln\frac{y_m}{1-y_m} & (y_m-0{,}5)\ln\frac{y_m}{1-y_m} \end{bmatrix} \qquad \text{Pers. 9}$$

$$\mathbf{Y}_4 = \begin{bmatrix} 1 & \ln\frac{y_1}{1-y_1} & (y_1-0{,}5)\ln\frac{y_1}{1-y_1} & y_1-0{,}5 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \ln\frac{y_m}{1-y_m} & (y_m-0{,}5)\ln\frac{y_m}{1-y_m} & y_m-0{,}5 \end{bmatrix} \qquad \text{Pers. 10}$$

DOI: https://doi.org/10.26593/jrsi.v12i1.6380.1-10

Untuk jumlah suku ganjil $n \geq 5$ matriks $\mathbf{Y}_n$ dinyatakan dengan

$$\mathbf{Y}_n = \begin{bmatrix} & (y_1 - 0{,}5)^{(n-1)/2} \\ \mathbf{Y}_{n-1} & \vdots \\ & (y_m - 0{,}5)^{(n-1)/2} \end{bmatrix} \quad \text{Pers. 11}$$

Sedangkan untuk jumlah suku genap $n \geq 6$

$$\mathbf{Y}_n = \begin{bmatrix} & (y_1 - 0{,}5)^{(n-1)/2} \ln\frac{y_1}{1-y_1} \\ \mathbf{Y}_{n-1} & \vdots \\ & (y_m - 0{,}5)^{(n-1)/2} \ln\frac{y_m}{1-y_m} \end{bmatrix} \quad \text{Pers. 12}$$

Sementara itu, fungsi umum PDF distribusi Metalog merupakan fungsi dari probabilitas kumulatif, dinotasikan dengan $m_n(y)$, dinyatakan dengan (Keelin, 2016):

$$m_2(y) = \frac{y(1-y)}{a_2} \quad \text{Pers. 13}$$

$$m_3(y) = \frac{1}{\frac{a_1}{y(1-y)} + a_3\left(\frac{y-0{,}5}{y(1-y)} + \ln\frac{y}{1-y}\right)} \quad \text{Pers. 14}$$

$$m_4(y) = \frac{1}{\frac{a_1}{y(1-y)} + a_3\left(\frac{y-0{,}5}{y(1-y)} + \ln\frac{y}{1-y}\right) + a_4} \quad \text{Pers. 15}$$

Untuk jumlah suku ganjil $n \geq 5$ fungsi PDF dinyatakan dengan

$$m_n(y) = \frac{1}{\frac{1}{m_{n-1}(y)} + a_n\frac{n-1}{2}(y-0{,}5)^{(n-3)/2}} \quad \text{Pers. 16}$$

Sedangkan untuk jumlah suku genap $n \geq 6$ fungsi PDF dinyatakan dengan

$$m_n(y) = \frac{1}{\frac{1}{m_{n-1}(y)} + a_n\left(\frac{(y-0{,}5)^{n/2-1}}{y(1-y)} + \left(\frac{n}{2}-1\right)(y-0{,}5)^{n/2-2}\ln\frac{y}{1-y}\right)} \quad \text{Pers. 17}$$

Sebagian besar distribusi probabilitas teoritis dapat dimodelkan dengan Metalog 4-suku sedangkan distribusi multimodal umumnya dapat direpresentasikan dengan Metalog 5-suku (Keelin, 2016).

Kopula Gaussian

Misalkan $F_i(X_i)$ adalah fungsi CDF dari variabel X_i untuk $i = 1, \ldots, \mathrm{p}$, dan misalkan $F(X_1, \ldots, X_\mathrm{p})$ adalah fungsi distribusi kumulatif bersama yang bersesuaian. Berdasarkan teorema Sklar, terdapat kopula C yang unik sedemikian sehingga (Durante & Sempi, 2016):

$$F(X_1, \ldots, X_p) = C\left(F_1(X_1), \ldots, F_p(X_p)\right) \quad \text{Pers. 18}$$

Dalam kopula Gaussian, kopula G didefinisikan sedemikian sehingga

$$F(X_1, \ldots, \mathrm{X}_p) = G\left(\Phi^{-1}\left(F_1(X_1)\right), \ldots, \Phi^{-1}\left(F_p(X_p)\right)\right) \quad \text{Pers. 19}$$

dengan $\Phi(\cdot)$ merepresentasikan fungsi distribusi kumulatif normal baku.

Membangkitkan Variat Acak Normal Multivariat

Misalkan distribusi probabilitas bersama yang direpresentasikan dengan kopula Gaussian G memiliki vektor rerata $\boldsymbol{\mu}$ dan matriks kovariansi $\boldsymbol{\Sigma}$. Vektor variat acak berukuran p dari distribusi probabilitas ini, dinotasikan dengan $\mathbf{Z}$, dapat dibangkitkan dengan (Rubinstein, 1981):

$$\mathbf{Z} = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{\Phi}^{-1}(\mathbf{U}) \quad \text{Pers. 20}$$

di mana $\mathbf{L}$ adalah matriks diagonal bawah sedemikian sehingga $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}'$ yang bisa diperoleh menggunakan dekomposisi Cholesky, dan $\mathbf{U}$ adalah vektor berukuran p dari bilangan random berdistribusi seragam kontinu dalam interval $(0,1)$.

Metode Oversampling Sintetis Usulan

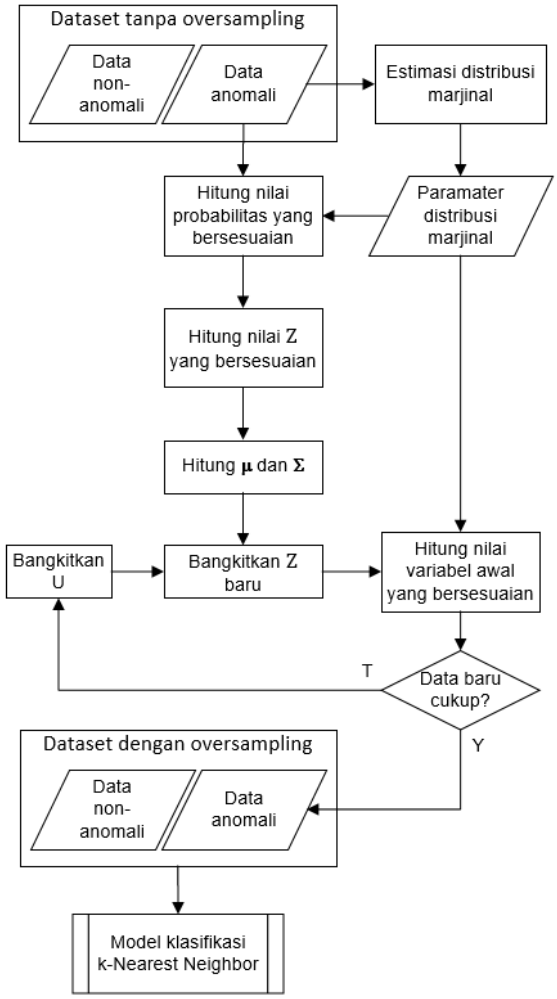
Metode *oversampling* sintetis yang diusulkan dalam penelitian ini mengikuti langkah-langkah berikut:

1. Penentuan distribusi probabilitas marjinal setiap variabel berdasarkan alternatif sistem distribusi berikut:
   a. Sistem distribusi Pearson
   b. Sistem distribusi Metalog
   c. Distribusi empiris
2. Penentuan distribusi probabilitas bersama menggunakan kopula Gaussian dengan parameter vektor rerata $\boldsymbol{\mu}$ dan matriks kovariansi $\boldsymbol{\Sigma}$
3. Membangkitkan variat acak baru berdasarkan vektor rerata $\boldsymbol{\mu}$ dan matriks kovariansi $\boldsymbol{\Sigma}$
4. Mentransformasikan variat acak ke nilai

DOI: https://doi.org/10.26593/jrsi.v12i1.6380.1-10

variabel yang bersesuaian berdasarkan distribusi probabilitas marjinalnya

Gambar 2 menampilkan metode *oversampling* usulan secara lebih detil. Untuk mendapatkan distribusi probabilitas bersama, dihitung nilai probabilitas yang bersesuaian dengan nilai setiap data berdasarkan distribusi probabilitas marjinalnya. Selanjutnya dihitung nilai variabel berdistribusi normal $N(0,1)$ yang bersesuaian dengan nilai probabilitas tersebut. Nilai terakhir inilah yang digunakan untuk menghitung parameter kopula Gaussian, $\boldsymbol{\mu}$ dan $\boldsymbol{\Sigma}$. Kedua parameter ini selanjutnya digunakan untuk menghasilkan data baru.



Gambar 2. Metode *oversampling* sintetis usulan

k-Nearest Neighbor

k-Nearest Neighbor adalah metode klasifikasi yang memberi label pada suatu data berdasarkan sebanyak *k* data yang paling dekat dengannya (Cover & Hart, 1967). Metode ini bersifat nonparametrik dan didasarkan pada jarak. Ukuran jarak yang biasa digunakan adalah Eucledian, Manhattan, Minkowski, dan Hamming (untuk data biner). Nilai k yang disarankan adalah akar kuadrat dari banyaknya data. Sebagai metode nonparametrik, model klasifikasi *k-Nearest Neighbor* akan memiliki probabilitas galat setidaknya sebesar probabilitas galat metode parametrik atau *Bayesian*. Cover & Hart (1967) menunjukkan bahwa dengan jumlah data yang relatif besar probabilitas galat tidak akan lebih besar dari dua kali probabilitas galat metode Bayesian.

Untuk mencegah *overfitting*, model klasifikasi akan divalidasi menggunakan *k-fold cross validation* dengan $k = 10$. Dalam metode validasi ini, dataset dibagi menjadi 10 sub dataset dengan proporsi yang seimbang. Selanjutnya dibuat model menggunakan 9 dari 10 sub dataset tersebut sementara 1 sub dataset lainnya digunakan untuk validasi. Proses ini diulang sebanyak 10 kali sedemikian sehingga setiap sub dataset satu kali digunakan untuk validasi (Theodoridis, 2015).

Studi Kasus

Metode *oversampling* sintetis usulan akan diimplementasikan dalam masalah prediksi kredit macet nasabah kartu kredit di suatu bank. Dataset yang digunakan dalam penelitian ini adalah data nasabah kartu kredit dengan jumlah *record* 17.707. Dataset ini terdiri dari 1 variabel target biner mengindikasikan macet atau tidak macet dan 20 variabel prediktor. Sebanyak 1.564 data berlabel macet dan 16.143 berlabel tidak macet. Dalam studi kasus ini dibuat model klasifikasi berbasis k-Nearest Neighbor. Variabel prediktornya meliputi:

1. **ncard**: jumlah kartu aktif yang dimiliki pelanggan
2. **outst**: total saldo pemakaian kartu kredit
3. **limit**: jumlah maksimum limit kartu kredit yang dapat digunakan
4. **balance**: jumlah tagihan pada bulan terakhir
5. **tusage**: total pemakaian kartu kredit (tunai dan retail) pada bulan terakhir
6. **tcash**: total pemakaian transaksi tunai pada bulan terakhir
7. **tretail**: total pemakaian transaksi retail pada bulan terakhir
8. **unpaid**: jumlah tagihan yang tidak terbayar pada bulan terakhir
9. **payrat**: rasio perbandingan jumlah yang dibayar dengan tagihan pada bulan terakhir
10. **percol**: persentasi overlimit

DOI: https://doi.org/10.26593/jrsi.v12i1.6380.1-10

11. **util3**: utilisasi kartu kredit selama 3 bulan terakhir
12. **usage3**: rata-rata pemakaian selama 3 bulan terakhir dibagi total limit
13. **payrat3**: rata-rata rasio pembayaran 3 bulan terakhir
14. **util6**: utilisasi kartu kredit selama 3 bulan sebelum 3 bulan terakhir
15. **usage6**: rata-rata pemakaian selama 3 bulan sebelum 3 bulan terakhir
16. **payrat6**: rata-rata rasio pembayaran 3 bulan sebelum 3 bulan terakhir
17. **balpcard**: jumlah tagihan yang tidak terbayar pada bulan terakhir dibagi jumlah kartu aktif
18. **unpaidplmt**: jumlah tagihan yang tidak terbayar pada bulan terakhir dibagi total limit
19. **tuseplmt**: total pemakaian kartu kredit (tunai dan retail) pada bulan terakhir dibagi total limit
20. **length**: jumlah tahun sejak pembukaan kartu kredit pertama kali

Oversampling dengan metode usulan dan tujuh metode pembanding (SMOTE, ANSMOTE, Borderline SMOTE, Density Based SMOTE, Safe Level SMOTE, Relocating Safe Level SMOTE, dan ADASYN) akan dilakukan untuk mencapai *balance level* sebesar 1 yang berarti proporsi jumlah data kredit macet dan tidak macet sama. Setelah dilakukan *oversampling* akan didapatkan tiga dataset dari metode usulan dan 7 dataset dari metode pembanding. Kesepuluh dataset ini selanjutnya akan menjadi input untuk model klasifikasi.

Implementasi metode *oversampling* sintetis usulan dan perbandingan dengan metode *oversampling* sintetis lain dilakukan dengan membangun model klasifikasi berbasis *k-Nearest Neighbor* dengan ukuran kinerja akurasi total. Model klasifikasi divalidasi dengan metode *k-fold cross validation*. Implementasi dilakukan dengan perangkat lunak RStudio v1.4.1717 dengan package `class` (Venables & Ripley, 2002) untuk klasifikasi *k-Nearest Neighbor*, `smotefamily` (Siriseriwan, 2019) untuk metode *oversampling* dalam keluarga SMOTE yang digunakan sebagai pembanding, `fitdistrplus` (Delignette-Muller & Dutang, 2015) untuk *distribution fitting* terhadap sistem distribusi Pearson, `rmetalog` (Faber & Jung, 2021) untuk *distribution fitting* terhadap sistem distribusi Metalog, dan `MASS` (Ripley et al., 2021) untuk membangkitkan variat acak berdasarkan kopula Gaussian.

Hasil dan Pembahasan

Proporsi data dalam dataset yang digunakan dalam studi kasus sangat tidak seimbang dengan *imbalanced ratio* $1.564/16.143 = 0{,}097$ yang masuk kategori ketidakseimbangan tinggi. *Oversampling* dilakukan terhadap data dengan label kredit macet di mana data sintetis yang dibangkitkan merupakan data 20 variabel prediktor seperti yang dijelaskan di subbab terdahulu. Dalam studi kasus ini tidak dilakukan *exploratoty data analysis* untuk menyeleksi variabel, sehingga semua variabel prediktor digunakan dalam model klasifikasi.

Distribusi probabilitas marjinal setiap variabel prediktor ditentukan berdasarkan tiga alternatif sistem distribusi, yaitu Pearson, Metalog, dan empiris. Hasil *distribution fitting* terhadap sistem distribusi Pearson tidak digunakan karena menghasilkan tingkat kesesuaian yang relatif buruk. Plot data dalam grafik Cullen-Frey menunjukkan tidak ada distribusi teoritis yang relatif dekat dengan data. Hasil *distribution fitting* terhadap sistem distribusi Metalog menghasilkan fungsi distribusi yang sesuai dengan rekapitulasi banyaknya suku untuk setiap variabel prediktor dapat dilihat di Tabel 1.

Sementara itu, distribusi empiris akan selalu sesuai dengan data. Walaupun demikian, penggunaan distribusi empiris memiliki kelemahan terkait batas rentang nilai dan kemungkinan terjadinya *overfitting*. Untuk selanjutnya, metode *oversampling* usulan akan didasarkan pada distribusi marjinal empiris dan Metalog.

Selanjutnya dilakukan pembuatan data sintesis menggunakan dua metode usulan dan tujuh metode pembanding, yaitu SMOTE, ANSMOTE, Borderline SMOTE, Density Based SMOTE, Safe Level SMOTE, Relocating Safe Level SMOTE, dan ADASYN. *Oversampling* sintetis dilakukan untuk mencapai *balance level* 1 sehingga untuk setiap variabel prediktor dibuat data sintetis sebanyak $16.143 - 1.564 = 14.579$ untuk setiap metode.

Di tahap selanjutnya, dibuat model klasifikasi dengan *k-Nearest Neighbor* menggunakan data yang sudah seimbang untuk setiap metode *oversampling*. Model untuk

DOI: https://doi.org/10.26593/jrsi.v12i1.6380.1-10

semua metode oversampling dibuat dengan spesifikasi yang sama dengan ukuran jarak Eucledian dan nilai $k = 179$.

Tabel 1. Banyaknya suku hasil *distribution fitting* terhadap distribusi Metalog

Variabel prediktor	Banyaknya suku
ncard	13
outst	13
limit	13
balance	13
tusage	13
tcash	3
tretail	13
unpaid	3
payrat	13
percol	3
util3	13
usage3	3
payrat3	13
util6	13
usage6	13
payrat6	13
balpcard	3
unpaidplmt	13
tuseplmt	3
length	13

Tabel 2 menampilkan nilai akurasi total dari model klasifikasi k-Nearest Neighbor yang dibuat menggunakan dataset nasabah kartu kredit yang sudah diseimbangkan jumlahnya menggunakan berbagai metode *oversampling*. Nilai akurasi total ini adalah rata-rata dari 10 pengulangan menggunakan *k-fold cross validation*.

Dapat dilihat dalam Tabel 2 bahwa metode usulan dengan distribusi marjinal Metalog memiliki akurasi total terbesar. Penggunaan model klasifikasi berbasis *k-Nearest Neighbor*, yang merupakan dasar dari hampir semua metode *oversampling* pembanding, menambah keyakinan terhadap hasil ini. Walaupun hasil ini menjanjikan, diperlukan eksperimen lebih ekstensif untuk menguji motode *oversampling* sintetis usulan ini dengan berbagai dataset dan teknik klasifikasi yang berbeda, sehingga bisa meningkatkan keyakinan terhadap metode ini.

Metode *ovesampling* sintetis usulan diimplementasikan dalam studi kasus yang seluruh variabelnya kontinu atau diasumsikan kontinu. Walaupun demikian, metode ini dapat diterapkan untuk dataset yang mengandung variabel diskrit. Selama variabel diskrit tersebut bisa ditentukan distribusi probabilitasnya, baik teoritis maupun empiris, distribusi probabilitas bersama dalam bentuk kopula Gaussian dapat didefinisikan sehingga data sintetis bisa dibuat.

Tabel 2. Akurasi total model klasifikasi *k-Nearest Neighbor* dengan berbagai metode *oversampling* sintetis

Model klasifikasi *k-Nearest Neighbor* dengan metode *oversampling* sintetis	Akurasi total
SMOTE	69,08%
ANSMOTE	69,33%
Borderline SMOTE	69,11%
Density Based SMOTE	80,77%
Safe Level SMOTE	72,83%
Relocating Safe Level SMOTE	75,55%
ADASYN	68,49%
Kopula Gaussian dengan distribusi marjinal empiris	79,98%
Kopula Gaussian dengan distribusi marjinal Metalog	82,13%

Menilai kinerja suatu metode *oversampling* tidaklah mudah. Akurasi total sejatinya adalah ukuran kinerja model klasifikasi dan tidak sepenuhnya tepat digunakan mengukur kinerja metode *oversampling* yang dipakai untuk menghasilkan data sintetis yang menjadi masukan model klasifikasi tersebut.

Secara intuitif, metode *oversampling* sintetis bisa dikatakan baik jika bisa menghasilkan data baru yang mirip dengan data aslinya. Namun, ukuran kemiripan menjadi rancu ketika setiap metode menggunakan cara yang berbeda dalam menghasilkan data baru. Metode-metode berbasis SMOTE mendasarkan kemiripan pada jarak, sementara metode usulan mendasarkannya pada distribusi probabilitas bersama. Dalam statistika, keduanya merupakan ukuran yang sahih digunakan untuk mengidentifikasi pola dalam data multivariat. Dalam konteks ini perbandingan "kemiripan" akan sulit dilakukan. Setidaknya, metode *oversampling* sintetis usulan ini memberi lebih banyak pilihan bagi para ilmuwan data untuk mengatasi masalah data yang tidak seimbang dalam model klasifikasi.

Kesimpulan

Penelitian ini mengusulkan metode *oversampling* sintetis berbasis distribusi probabilitas bersama dari data aslinya. Distribusi probabilitas bersama direpresentasikan dengan kopula Gaussian, dengan tiga alternatif distribusi marjinal, yaitu sistem distribusi Pearson, distribusi empiris, dan sistem distribusi Metalog. Metode usulan ini dibandingkan dengan beberapa metode *oversampling* yang umum digunakan untuk data yang tidak seimbang. Metode usulan diterapkan dalam masalah prediksi kredit macet nasabah kartu kredit dengan metode klasifikasi *k-Nearest Neighbor* dengan ukuran kinerja akurasi total dengan metode validasi *k-fold cross validation*. Didapati bahwa model klasifikasi dengan metode *oversampling* usulan dengan distribusi marjinal Metalog memiliki akurasi total tertinggi.

Diperlukan pengujian lebih ekstensif menggunakan berbagai dataset dengan berbagai metode klasifikasi untuk menambah keyakinan terhadap kinerja metode usulan ini. Metode usulan ini dapat dikembangkan lebih lanjut untuk mencakup distribusi probabilitas marjinal diskrit dan metode *ensemble*. Metode ini bisa menjadi alternatif cara untuk mengatasi masalah data yang tidak seimbang dalam penggunaan model klasifikasi di industri.

Data

Dataset dan program dalam bahasa pemrograman R yang digunakan dalam penelitian ini tersedia di Mendeley Data dengan tautan doi: 10.17632/jrss9jdjz9.1.

Daftar Pustaka

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus : An R Package for Fitting Distributions. Journal of Statistical Software, 64(4). https://doi.org/10.18637/jss.v064.i04

Durante, F., & Sempi, C. (2016). Principles of Copula Theory. Chapman and Hall/CRC. https://doi.org/10.1201/b18674

Faber, I., & Jung, J. (2021). rmetalog: The Metalog Distribution.

García, V., Sánchez, J. S., & Mollineda, R. A. (2012). On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. Knowledge-Based Systems, 25(1), 13–21. https://doi.org/10.1016/j.knosys.2011.06.013

Haibo He, Yang Bai, Garcia, E. A., & Shutao Li. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

Keelin, T. W. (2016). The Metalog Distributions. Decision Analysis, 13(4), 243–277. https://doi.org/10.1287/deca.2016.0338

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. Data Mining and Knowledge Discovery, 28(1), 92–122. https://doi.org/10.1007/s10618-012-0295-5

Patel, H., Singh Rajput, D., Thippa Reddy, G., Iwendi, C., Kashif Bashir, A., & Jo, O. (2020). A review on classification of imbalanced data for wireless sensor networks. International Journal of Distributed Sensor Networks, 16(4), 155014772091640. https://doi.org/10.1177/1550147720916404

Pearson, K. (1895). X. Contributions to the mathematical theory of evolution.—II. Skew variation in homogeneous material. Philosophical Transactions of the Royal Society of London. (A.), 186, 343–414. https://doi.org/10.1098/rsta.1895.0010

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., & Firth, D. (2021). MASS: Support Functions and Datasets for Venables and Ripley's MASS (7.3-54). https://cran.r-project.org/package=MASS

Rubinstein, R. Y. (Ed.). (1981). Simulation and the Monte Carlo Method. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316511

Sáez, J. A., Luengo, J., Stefanowski, J., & Herrera, F. (2015). SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with

filtering. Information Sciences, 291, 184–203. https://doi.org/10.1016/j.ins.2014.08.051

Siriseriwan, W. (2019). smotefamily: a collection of oversampling techniques for class imbalance problem based on SMOTE.

Tahir, M. A., Kittler, J., & Yan, F. (2012). Inverse random under sampling for class imbalance problem and its application to multi-label classification. Pattern Recognition, 45(10), 3738–3750. https://doi.org/10.1016/j.patcog.2012.03.014

Theodoridis, S. (2015). Machine Learning: A Bayesian and Optimization Perspective (1st ed.). Elsevier.

Venables, W. N., & Ripley, B. D. (2002). Modern Applied Statistics with S. Springer New York. https://doi.org/10.1007/978-0-387-21706-2

Wang, K.-J., Makond, B., Chen, K.-H., & Wang, K.-M. (2014). A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients. Applied Soft Computing, 20, 15–24. https://doi.org/10.1016/j.asoc.2013.09.014

Wong, G. Y., Leung, F. H. F., & Ling, S.-H. (2014). An under-sampling method based on fuzzy logic for large imbalanced dataset. 2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 1248–1252. https://doi.org/10.1109/FUZZ-IEEE.2014.6891771

Zhang, H., & Li, M. (2014). RWO-Sampling: A random walk over-sampling approach to imbalanced data classification. Information Fusion, 20, 99–116. https://doi.org/10.1016/j.inffus.2013.12.003