ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 8, Nomor 1, Tahun 2019, Halaman 149 - 160
Online di: http://ejournal-s1.undip.ac.id/index.php/gaussian

METODE *GENERALIZED MEAN DISTANCE-BASED K-NEAREST NEIGHBOR CLASSIFIER* (GMDKNN) UNTUK ANALISIS *CREDIT SCORING* CALON DEBITUR KREDIT TANPA AGUNAN (KTA)

Mei Sita Saraswati[1], Moch. Abdul Mukid [2], Abdul Hoyyi [3]
[1]Mahasiswa Departemen Statistika FSM Universitas Diponegoro
[2,3]Dosen Departemen Statistika FSM Universitas Diponegoro
meisitasaras@gmail.com

ABSTRACT

Unsecured Credit is one of the credit facilities provided by banks, where the prospective debtor can borrow some amount of fund from the bank without having to provide collateral. Credit scoring is a process that aims to assess the worthiness of credit applications and classify the credit applicants into prospective debtors whose the credit application is worthy to be accepted and prospective debtors whose the credit application should be rejected. One of the statistical methods that can be applied in examining the analysis of credit scoring is the Generalized Mean Distance-Based k-Nearest Neighbor (GMDKNN) classifier. Empirical study on this method uses 23,337 data of prospective debtor of unsecured credit in 2018, with the dependent variable being the credit scoring final decision and seven independent variables, i.e. age, child dependent, length of employment, age of the company, income, loan proposed, and duration of credit. Based on the feature selection test, all independent variables are significantly taking effect on the credit scoring final decision. The best classification model is obtained in the parameters k = 137 and p = -1 with the classification performance metrics represented by the values of APER = 0,2580, accuracy = 74,20%, sensitivity = 0,6083, specificity = 0,8393, AUC = 0,7238, and G-Mean = 0,7146.

Keywords: Unsecured Credit, credit scoring, classification, *Generalized Mean Distance-Based k-Nearest Neighbor* (GMDKNN).

1. PENDAHULUAN

Perkreditan bagi masyarakat perseorangan atau badan usaha adalah salah satu kegiatan usaha bank dalam rangka penyaluran dan penghimpunan dana. Menurut Rama dan Jones (2009), kredit merupakan sumber utama bank dalam menjalankan aktivitas usahanya termasuk pada perusahaan pembiayaan. Pemberian kredit, selain sebagai aktiva produktif terbesar suatu bank juga merupakan pembawa risiko tertinggi yang mampu mempengaruhi tingkat kesehatan bank (Firdaus & Ariyanti, 2009). Salah satu risiko dari aktivitas penyaluran kredit oleh bank yaitu munculnya kredit bermasalah atau yang sering disebut dengan kredit macet. Kredit macet merupakan suatu keadaan dimana debitur baik perorangan atau perusahaan tidak mampu membayar kredit bank tepat pada waktunya sehingga mengancam likuiditas bank tersebut.

Salah satu jenis kredit yang disalurkan oleh bank adalah Kredit Tanpa Agunan (KTA). KTA adalah sebuah produk bank, dimana calon debitur dapat meminjam sejumlah dana atau uang dari bank tanpa harus memberikan jaminan atau agunan, sehingga KTA memiliki risiko lebih terhadap penyelewengan kredit pinjaman oleh debitur dibandingkan jenis kredit lainnya.

Kebanyakan perbankan telah menerapkan prosedur *credit scoring* untuk menilai risiko berkaitan dengan pinjaman kepada calon debitur dan mengklasifikasi para pemohon kredit ke dalam dua kelompok yaitu calon debitur yang permohonan kreditnya layak untuk diterima dan calon debitur yang permohonan kreditnya sebaiknya ditolak. Untuk melaksanakan

proses *credit scoring*, telah diaplikasikan metode statistika yaitu bermacam teknik klasifikasi.

Metode klasifikasi yang baik adalah metode yang menghasilkan kesalahan yang kecil (Johnson & Wichern, 2007). Saat ini terdapat banyak metode statistika yang digunakan untuk mengklasifikasi, salah satunya adalah metode *Generalized Mean Distance-Based k-Nearest Neighbor* (GMDKNN). Menurut Gou *et al*. (2019), metode GMDKNN merupakan pengembangan dari metode *k-Nearest Neighbor* (KNN) dengan memperkenalkan *multi-generalized mean distances* dan *nested generalized mean distance* yang keduanya didasarkan pada karakteristik rata-rata tergeneralisasi. Berdasarkan hasil penelitian sebelumnya oleh Gou *et al*. (2019), diperoleh kesimpulan bahwa metode GMDKNN dapat meningkatkan performa klasifikasi berbasis KNN serta mampu mengatasi sensitivitas dari tetangga ukuran *k* yang seringkali secara signifikan menurunkan performa klasifikasi berbasis KNN. Maka dalam penulisan ini akan diterapkan metode GMDKNN untuk mengklasifikasikan keputusan akhir permohonan kredit calon debitur KTA pada proses *credit scoring* Bank "X".

2. TINJAUAN PUSTAKA

2.1. Kredit

Menurut Undang-Undang Perbankan nomor 10 tahun 1998 pengertian kredit adalah penyediaan uang atau tagihan yang dapat dipersamakan dengan itu, berdasarkan persetujuan atau kesepakatan pinjam meminjam antara bank dengan pihak lain yang mewajibkan pihak peminjam melunasi utangnya setelah jangka waktu tertentu dengan pemberian bunga. Menurut Sinungan (1991) tujuan utama pemberian kredit adalah untuk memperoleh keuntungan dalam bentuk bunga yang diterima oleh bank sebagai balas jasa dan biaya administrasi kredit yang dibebankan kepada nasabah.

2.2. *Credit scoring*

Credit scoring didefinisikan sebagai sebuah metode sistematis untuk mengevaluasi risiko kredit yang menyajikan suatu analisis konsisten dari faktor-faktor yang telah ditetapkan sebelumnya sebagai penyebab atau mempengaruhi level dari risiko (Fensterstock, 2005). *Credit scoring* membantu menentukan apakah kredit seharusnya diberikan kepada peminjam atau tidak (Morrison, 2004). Teknik yang paling umum digunakan dalam pembentukan model *credit scoring* adalah metode statistika yaitu klasifikasi.

2.3. *Data Mining*

Turban *et al*. (2005) mendefinisikan *data mining* sebagai proses menggunakan teknik statistika, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar. Tahapan *data mining* menurut Jiawei (2006) meliputi *data cleaning*, *data integration*, *data selection*, *data transformation*, *data mining*, *pattern evaluation*, dan *knowledge presentation*. Berdasarkan tugasnya, *data mining* dikelompokkan menjadi deskripsi, estimasi, prediksi, klasifikasi, *clustering*, dan asosiasi (Larose, 2005).

2.4. Seleksi Fitur

Langkah yang paling sederhana dalam memilih fitur adalah mengamati setiap fitur yang dibangkitkan secara independen dan menguji kemampuan diskriminasinya pada masalah yang harus diselesaikan (Prasetyo, 2014). Untuk tujuan tersebut, dapat diterapkan prosedur nonparametrik yaitu uji *Mann-Whitney*.

Penjelasan uji *Mann-Whitney* oleh Conover (1980) yaitu pada data yang terdiri dari dua sampel acak, dimana $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_{n1}$ menotasikan sampel acak sebanyak n_1 dari populasi

1 dan $\mathbf{y}_1$, $\mathbf{y}_2$, …, $\mathbf{y}_{n2}$ menotasikan sampel acak sebanyak n_2 dari populasi 2, kemudian ditentukan peringkat 1 hingga $n_1 + n_2$ untuk semua sampel data. Tahapan pengujian dijelaskan oleh Daniel (1989) yang diawali dengan penetapan hipotesis nol sebagai nilai rata-rata fitur dalam dua kelas sama dan hipotesis alternatifnya yaitu nilai rata-rata fitur dalam dua kelas berbeda. Selanjutnya dihitung statistik uji dengan rumus:

$$T = S - \frac{n_1(n_1 + 1)}{2}$$

dengan S adalah jumlah peringkat hasil-hasil pengamatan yang merupakan sampel dari populasi 1. Apabila entah n_1 atau n_2 lebih besar dari 20, maka teorema limit sentral dapat diterapkan, dengan demikian statistik uji:

$$Z = \frac{T - n_1 n_2/2}{\sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12} - \frac{n_1 n_2(\sum t^3 - \sum t)}{12(n_1 + n_2)(n_1 + n_2 - 1)}}}$$

dengan t adalah banyaknya angka sama untuk suatu peringkat. Kriteria uji yaitu H_0 ditolak jika $Z > Z_{\alpha/2}$ atau $Z < -Z_{\alpha/2}$ atau apabila p-value < α, dimana $Z_{\alpha/2}$ adalah nilai kritis yang diperoleh dari tabel distribusi normal standar.

2.5. Standarisasi Data

Untuk mencegah pengaruh perbedaan satuan antar atribut terhadap analisa data, dilakukan standarisasi terhadap nilai atribut (Larose, 2005). Salah satu metode standarisasi yang umum diterapkan adalah metode *z-score*. Formula untuk standarisasi atribut X dengan metode *z-score* adalah:

$$Z = \frac{X - \bar{X}}{s_X}$$

dengan Z adalah nilai setelah distandarisasi, X adalah nilai sebelum distandarisasi, $\bar{X}$ adalah rata-rata nilai sebelum distandarisasi, s_X adalah nilai standar deviasi atribut X.

2.6 Klasifikasi

Klasifikasi didefinisikan sebagai suatu pekerjaan yang melakukan pelatihan/pembelajaran terhadap fungsi target f yang memetakan setiap vektor (set fitur) x ke dalam satu dari sejumlah label kelas y yang tersedia. Pekerjaan pelatihan tersebut akan menghasilkan suatu model yang kemudian disimpan sebagai memori (Prasetyo, 2014). Dalam pembangunan model diperlukan adanya algoritma pelatihan (*learning algorithm*). Algoritma pelatihan mempunyai prinsip melakukan suatu pelatihan sehingga di akhir pelatihan model dapat memetakan (memprediksi) setiap vektor masukan ke label kelas keluaran dengan benar (Prasetyo, 2014).

2.7 Jarak *Euclid*

Ukuran ketidakmiripan yang paling umum digunakan dalam metode klasifikasi adalah jarak *Euclid* (Prasetyo, 2014). Jarak *Euclid* memberikan jarak lurus antara dua buah data $\mathbf{x_b} = (x_{1b}, x_{2b}, \ldots, x_{nb})$ dan $\mathbf{x_a} = (x_{1a}, x_{2a}, \ldots, x_{na})$ yang diformulasikan oleh persamaan berikut:

$$d(\mathbf{x_b}, \mathbf{x_a}) = \sqrt{\sum_{l=1}^{n}(x_{bl} - x_{al})^2}$$

dengan d adalah jarak antara data $\mathbf{x_b}$ dan $\mathbf{x_a}$, $\mathbf{x_b}$ adalah data uji, $\mathbf{x_a}$ adalah data latih, dan n adalah banyaknya fitur (dimensi) data.

2.8 *Generalized Mean Distance-Based k-Nearest Neighbor* (GMDKNN)

Gou *et al*. (2019) memperkenalkan metode *Generalized Mean Distance-based k-Nearest Neighbor Classifier* (GMDKNN) sebagai pengembangan dari metode *k-Nearest Neighbour* (KNN) dengan inovasi utama berupa *Multi-generalized mean distances* berbasis *multi-local mean vectors* serta *Nested generalized mean distance*. Gou *et al*. (2019) mendeskripsikan langkah-langkah untuk memberikan label kelas dari sampel uji **x** menggunakan metode GMDKNN, sebagai berikut:

a) Mencari k tetangga terdekat terhadap **x** dari himpunan $\mathbf{X}^{j}$ di setiap kelas $\boldsymbol{\omega}_j$, dinyatakan sebagai $\boldsymbol{X}_j^{NN} = \left\{\boldsymbol{x}_{ij}^{NN} \in R^d\right\}_{i=1}^{k}$. Penentuan *k nearest neighbor* diurutkan dari tetangga dengan jarak *Euclid* terdekat hingga tetangga dengan jarak *Euclid* terjauh terhadap **x**.

b) Menghitung *k local mean vectors* (vektor rata-rata lokal) menggunakan i ($1 \le i \le k$) tetangga terdekat terhadap **x** dari tiap kelas $\boldsymbol{\omega}_j$ berdasarkan rumus:

$$\boldsymbol{u}_{ij}^{NN} = \frac{1}{i}\sum_{l=1}^{i} \boldsymbol{x}_{ij}^{NN}, 1 \le i \le k$$

dinyatakan sebagai $\boldsymbol{U}_j^{NN} = \left\{\boldsymbol{u}_{ij}^{NN} \in R^d\right\}_{i=1}^{k}$. Jarak *Euclid* yang bersesuaian terhadap **x** diindikasikan sebagai himpunan: $D_j^{NN} = \left\{d\left(\boldsymbol{x}, \boldsymbol{u}_{1j}^{NN}\right), d\left(\boldsymbol{x}, \boldsymbol{u}_{2j}^{NN}\right), \ldots, d\left(\boldsymbol{x}, \boldsymbol{u}_{kj}^{NN}\right)\right\}$.

c) Menghitung *k generalized mean distances* menggunakan r ($1 \le r \le k$) jarak-jarak pada himpunan D_j^{NN} untuk kelas $\boldsymbol{\omega}_j$. *Generalized mean distance* (jarak rata-rata tergeneralisasi) r jarak pertama dari r pertama *local mean vector* terhadap **x** untuk setiap kelas dirumuskan sebagai:

$$g\left(\boldsymbol{x}, \boldsymbol{U}_{rj}^{NN}\right) = \left(\frac{1}{r}\sum_{i=1}^{r}\left(d\left(\boldsymbol{x}, \boldsymbol{u}_{ij}^{NN}\right)\right)^p\right)^{\frac{1}{p}}$$

dengan j = 1, 2, …, m dan $1 \le r \le k$

dimana $\boldsymbol{U}_{rj}^{NN} = \left\{\boldsymbol{u}_{ij}^{NN} \in R^d\right\}_{i=1}^{r}$ dan *k generalized mean distances* di kelas $\boldsymbol{\omega}_j$ dinotasikan sebagai $\boldsymbol{g}^j = \left\{g\left(\boldsymbol{x}, \boldsymbol{U}_{1j}^{NN}\right), g\left(\boldsymbol{x}, \boldsymbol{U}_{2j}^{NN}\right) \ldots g\left(\boldsymbol{x}, \boldsymbol{U}_{kj}^{NN}\right)\right\}$.

d) Menentukan sebuah *nested generalized mean distance* baru berdasarkan nilai *k generalized mean distances* yang telah diperoleh sebelumnya untuk setiap kelas dengan rumus:

$$G\left(\boldsymbol{x}, g^j\right) = \left(\frac{1}{k}\sum_{r=1}^{k}\left(g\left(\boldsymbol{x}, \boldsymbol{U}_{rj}^{NN}\right)\right)^p\right)^{\frac{1}{p}}, \text{ j = 1, 2, …, m.}$$

e) Mengklasifikasikan sampel uji **x** ke dalam kelas yang memiliki nilai *nested generalized mean distance* yang paling minimum sebagai:

$$\omega = \arg\min_{\omega_j} G\left(\boldsymbol{x}, g^j\right), \quad j = 1,2, \ldots, m$$

Gou *et al*. (2019) menyatakan tujuan utama dari metode GMDKNN adalah mengembangkan performa klasifikasi berbasis KNN dan mengatasi sensitivitas ukuran tetangga *k*. Dalam menentukan nilai maksimal parameter *k*, Beckmann *et al*. (2015) mengembangkan *rule of thumb* pada kasus klasifikasi kelas biner sebagai berikut:

$$\text{Max}(k) = odd\left(\sqrt{n}\right)$$

dengan nilai *n* adalah jumlah data latih.

2.9 Evaluasi Performansi Ketepatan Klasifikasi

Data aktual dan data hasil prediksi dari model klasifikasi disajikan dengan menggunakan matriks konfusi (*confusion matrix*) sebagai berikut:

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

keterangan:
TP: *True Positive*, data aktual positif dan diklasifikasikan positif
FP: *False Positive*, data aktual negatif namun diklasifikasikan positif
FN: *False Negative*, data aktual positif, namun diklasifikasikan negatif
TN: *True Negative*, data aktual negatif dan diklasifikasikan negatif

Berdasarkan nilai-nilai dalam matriks konfusi, dapat dilakukan perhitungan ukuran-ukuran performansi ketepatan klasifikasi yaitu *Apparent Error Rate* (APER), akurasi, sensitivitas, dan spesifisitas yang dirumuskan dalam Prasetyo (2014) sebagai berikut:

$$\text{APER} = \frac{(\text{FP} + \text{FN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \qquad \text{Sensitivitas} = \frac{\text{TP}}{(\text{TP+FN})}$$

$$\text{Akurasi} = 1 - \text{APER} \qquad \text{Spesifisitas} = \frac{\text{TN}}{(\text{FP+TN})}$$

Evaluasi performansi klasifikasi menggunakan *Geometric Mean* (*G* –Mean) dapat dirumuskan sebagai berikut (Yuchun *et al*., 2002):

$$G - Mean = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{FP+TN}}$$

Pada kasus klasifikasi kelas biner, nilai performansi klasifikasi *Area under the ROC Curve* (AUC) dapat didekati dengan nilai *Balanced Accuracy* (Bekkar *et al*., 2013) sebagai berikut:

$$AUC = \frac{1}{2}(sensitivity + specitifity)$$

3. METODE PENELITIAN

Data yang digunakan dalam penulisan penelitian ini adalah data sekunder, yaitu data calon debitur yang mengajukan permohonan Kredit Tanpa Agunan (KTA) ke Bank X yang terletak di Provinsi DKI Jakarta pada tahun 2018 dengan jumlah data sebanyak 23.337 data. Variabel dependen yang digunakan adalah keputusan akhir *credit scoring* dikelompokkan menjadi sebesar 9.832 calon debitur kredit ditolak yang dilabelkan dengan ‘1’ dan 13.505 calon debitur kredit diterima yang dilabelkan dengan ‘0’. Variabel independen yaitu tujuh variabel berskala numerik meliputi data usia calon debitur dalam satuan tahun, tanggungan anak dalam satuan jiwa, lama bekerja dan lama perusahaan dalam satuan tahun, pendapatan dan pinjaman yang diajukan dalam satuan rupiah, serta durasi kredit dalam satuan bulan. *Software* yang digunakan pada penelitian ini adalah Ms. Excel 2013, IBM SPSS Statistics 25, dan Matlab R2015a. Langkah-langkah analisis yang dilakukan pada penelitian ini sebagai berikut:

1. Melakukan seleksi fitur pada semua variabel independen (X) menggunakan uji *Mann-Whitney* pada *software* IBM SPSS Statistics 25.
2. Analisis data dengan metode *Generalized Mean Distance-based k-Nearest Neighbor* (GMDKNN) menggunakan *software* Matlab R2015a dengan tahapan sebagai berikut:
   a. Standarisasi pada semua variabel independen (X) menggunakan metode *z-score*.
   b. Mengelompokkan data terstandarisasi ke dalam masing-masing kelas.
   c. Membagi data yakni untuk data latih diambil 80% data pada masing-masing kelas dan untuk data uji diambil 20% data pada masing-masing kelas.

d. Menghitung jarak *Euclid* antara data latih dan data uji.
e. Menentukan nilai parameter *k* dan *p*.
f. Mengurutkan nilai jarak *Euclid* dari nilai terkecil hingga nilai terbesar.
g. Menentukan jarak terdekat sebanyak nilai *k* yang telah ditentukan pada masing-masing kelas.
h. Menghitung *k local mean vectors* $\boldsymbol{U}_j^{NN}$ dari masing-masing kelas menggunakan i ($1 \leq i \leq k$) tetangga terdekat.
i. Menghitung kembali jarak D_j^{NN} tiap data uji terhadap setiap nilai *local mean vector* $\boldsymbol{u}_{ij}^{NN}$ pada *k local mean vectors* $\boldsymbol{U}_j^{NN}$ untuk masing-masing kelas.
j. Menghitung *k generalized mean distances* menggunakan r ($1 \leq r \leq k$) jarak-jarak pada himpunan D_j^{NN} untuk masing-masing kelas.
k. Menentukan sebuah *nested generalized mean distance G* pada masing-masing kelas berdasarkan nilai *k generalized mean distances*.
l. Memilih kelas dengan nilai *nested generalized mean distance* terkecil dari setiap kelas yang ada sebagai kelas data uji.
m. Menghitung AUC hasil klasifikasi.
n. Melakukan kembali langkah e-m pada berbagai nilai parameter *k* dan *p* sehingga mendapatkan nilai AUC terbesar.

3. Menghitung ukuran performansi ketepatan klasifikasi berupa nilai akurasi, APER, sensitivitas, spesifisitas dan *G-Mean* pada model klasifikasi dengan nilai parameter *k* dan *p* yang menghasilkan AUC terbesar, menggunakan *software* Matlab R2015a.

4. HASIL DAN PEMBAHASAN

4.1 Seleksi Fitur

Dalam pengujian hipotesis *Mann-Whitney*, fitur yang tidak mempunyai informasi diskriminasi data terhadap kelas akan dibuang. H_0 pengujian yaitu nilai rata-rata fitur dalam dua kelas keputusan akhir *credit scoring* sama, sedangkan H_1 yaitu nilai rata-rata fitur dalam dua kelas keputusan akhir *credit scoring* berbeda. Dengan taraf signifikansi $\alpha = 5\%$ dan nilai $-Z_{\alpha/2} = -1{,}960$, berdasarkan hasil output *software* SPSS diperoleh hasil uji *Mann-Whitney* untuk setiap variabel independen yang dinyatakan dalam tabel sebagai berikut:

Tabel 1. Uji *Mann-Whitney*

Variabel	Z_{hitung}	Sig.	Keputusan
Usia	-21,744	0,000	H_0 ditolak
Tanggungan Anak	-12,849	0,000	H_0 ditolak
Lama Bekerja	-12,767	0,000	H_0 ditolak
Lama Perusahaan	-7,112	0,000	H_0 ditolak
Pendapatan	-38,867	0,000	H_0 ditolak
Pinjaman yang Diajukan	-27,500	0,000	H_0 ditolak
Durasi Kredit	-16,550	0,000	H_0 ditolak

Berdasarkan tabel di atas, dapat disimpulkan bahwa semua variabel independen memiliki nilai rata-rata fitur dalam dua kelas keputusan akhir *credit scoring* yang berbeda, sehingga semua variabel independen berpengaruh secara signifikan terhadap keputusan akhir *credit scoring* dan valid untuk dimasukkan dalam model klasifikasi.

4.2 Standarisasi Data

Sebelum dilakukan standarisasi, data terlebih dahulu dibagi menjadi data latih dengan proporsi sebesar 80% yaitu sebanyak 18.670 data dan data uji dengan proporsi sebesar 20%

yaitu sebanyak 4.667 data. Hasil standarisasi metode *z-score* tujuh variabel independen dari seluruhnya 18.670 data latih dinyatatakan dalam tabel sebagai berikut:

Tabel 2. Data Latih Setelah Distandarisasi

No	Usia	Tanggungan Anak	Lama Bekerja	Lama Perusahaan	Pendapatan	Pinjaman yang Diajukan	Durasi Kredit	Keputusan Akhir
1	-0,9293	-1,2791	-0,5582	-0,4734	0,2513	-0,2388	-1,2160	0
2	-0,7010	0,2264	-0,7936	0,1095	-0,0793	0,1626	0,3544	0
3	-0,0160	2,4848	0,5011	-0,1554	0,7644	0,5479	0,3544	0
4	1,0114	1,7320	1,3250	1,5403	-0,1403	-0,4796	-0,6925	0
5	0,6689	2,4848	-0,9113	1,9112	1,0653	2,1693	0,3544	0
…	…	…	…	…	…	…	…	…
18669	-1,3860	-1,2791	-0,7936	-0,8443	-0,6738	-0,2388	0,3544	1
18670	-1,5001	-0,5263	-1,0290	-0,8443	-0,5405	-0,2388	0,3544	1

Sedangkan hasil standarisasi metode *z-score* tujuh variabel independen dari seluruhnya 4.667 data uji dinyatatakan dalam tabel sebagai berikut:

Tabel 3. Data Uji Setelah Distandarisasi

No	Usia	Tanggungan Anak	Lama Bekerja	Lama Perusahaan	Pendapatan	Pinjaman yang Diajukan	Durasi Kredit	Keputusan Akhir
1	0,7831	1,7320	0,0303	-1,1622	-0,4540	-0,5598	-0,1691	0
2	-0,0160	0,2264	-0,3228	-0,6853	-0,6256	-0,6401	-1,2160	0
3	0,4406	-0,5263	-1,0290	-0,3674	-0,5998	-0,6401	-0,1691	0
4	-0,1302	0,2264	0,5011	-0,7913	-0,2621	0,0823	0,3544	0
5	-0,2444	0,2264	-1,0290	-1,3212	0,1881	-0,2388	-1,2160	0
…	…	…	…	…	…	…	…	…
4666	-1,1576	-1,2791	-1,1467	0,5864	-0,4676	-0,3993	0,3544	1
4667	2,0388	0,9792	2,3843	0,8514	-0,3243	-0,3993	-1,7394	1

4.3 Penentuan Klasifikasi Kelas untuk Observasi Terakhir Data Uji

Pada tahap ini akan digunakan metode *Generalized Mean Distance-Based k-Nearest Neighbor* (GMDKNN) untuk mengklasifikasikan kelas dari seluruh data uji, dan sebagai ilustrasi perhitungan digunakan data uji terakhir yaitu data uji ke-4.667. Sebelum melakukan metode klasifikasi, terlebih dahulu data latih dibagi berdasarkan label kelasnya yaitu data latih kelas diterima atau kelas 1 dengan jumlah 10.804 data dan data latih kelas ditolak atau kelas 2 dengan jumlah 7.866 data.

4.3.1 Perhitungan Jarak *Euclid*

Hasil perhitungan jarak Euclid untuk data uji ke-4667 terhadap seluruh data latih dapat ditampilkan dalam tabel berikut:

Tabel 4. Perhitungan Jarak *Euclid* untuk Data Uji Ke-4.667

4.3.2 Penentuan Parameter Model Klasifikasi

No	Usia	Tanggungan Anak	Lama Bekerja	Lama Perusahaan	Pendapatan	Pinjaman yang Diajukan	Durasi Kredit	Keputusan Akhir	Jarak *Euclid*
1	-0,9293	-1,2791	-0,5582	-0,4734	0,2575	-0,2388	-1,2160	0	4,9954
2	-0,7010	0,2264	-0,7936	0,1095	-0,0705	0,1626	0,3544	0	4,8459
3	-0,0160	2,4848	0,5011	-0,1554	0,7667	0,5479	0,3544	0	4,1852
…	…	…	…	…	…	…	…	…	…
18669	-1,3860	-1,2791	-0,7936	-0,8443	-0,6604	-0,2388	0,3544	1	5,8597
18670	-1,5001	-0,5263	-1,0290	-0,8443	-0,5281	-0,2388	0,3544	1	5,8115

Pada penelitian ini dengan jumlah data latih sebesar 18.670 diperoleh nilai *k* maksimal sebesar 137. Nilai-nilai parameter *k* yang diujicobakan yaitu: $\{1, 3, 5, 7, 9, 11, 13, 15, 17, 27, 37, 57, 77, 107, 137\}$. Berdasarkan penelitian Gou *et al*. (2019), maka nilai-nilai parameter *p* yang diujicobakan pada penelitian ini yaitu: $\{-13, -11, -9, -7, -5, -4, -3, -2, -1, 1, 2, 3\}$. Dengan bantuan program MatLab, diperoleh hasil nilai AUC untuk berbagai nilai parameter *k* dan *p* yang dapat dilihat pada Tabel sebagai berikut:

Tabel 5. Hasil Nilai AUC GMDKNN

Nilai *p* / Nilai *k*	-13	-11	-9	-7	-5	-4	-3	-2	**-1**	1	2	3
1	0,6585	0,6585	0,6585	0,6585	0,6585	0,6585	0,6585	0,6585	0,6585	0,6585	0,6585	0,6585
3	0,6736	0,6745	0,6750	0,6757	0,6753	**0,6766**	0,6764	0,6752	0,6751	0,6756	0,6729	0,6742
5	0,6825	0,6826	0,6830	**0,6845**	0,6816	0,6785	0,6799	0,6818	0,6821	0,6782	0,6775	0,6750
7	0,6851	0,6853	0,6866	0,6879	**0,6893**	0,6880	0,6878	0,6891	0,6868	0,6844	0,6852	0,6829
9	0,6875	0,6876	0,6884	0,6929	0,6929	0,6921	**0,6929**	0,6898	0,6891	0,6863	0,6894	0,6869
11	0,6874	0,6900	0,6922	**0,6933**	0,6921	0,6933	0,6931	0,6895	0,6899	0,6882	0,6895	0,6891
13	0,6894	0,6896	0,6907	0,6931	0,6959	**0,6963**	0,6936	0,6914	0,6913	0,6886	0,6891	0,6876
15	0,6903	0,6915	0,6916	0,6933	0,6958	**0,6974**	0,6935	0,6926	0,6923	0,6933	0,6895	0,6906
17	0,6915	0,6939	0,6925	0,6914	0,6949	0,6965	**0,6965**	0,6947	0,6947	0,6950	0,6922	0,6908
27	0,6955	0,6961	0,6971	0,6971	0,7001	0,6996	0,6989	0,7009	**0,7021**	0,7001	0,7011	0,7020
37	0,6961	0,6962	0,6987	0,6989	0,6999	0,7021	0,7020	0,7019	0,7030	0,7053	0,7049	**0,7054**
57	0,6978	0,6982	0,7001	0,6996	0,7005	0,7031	0,7053	0,7047	0,7080	0,7124	**0,7132**	0,7113
77	0,6993	0,6995	0,7026	0,7018	0,7049	0,7074	0,7109	0,7130	**0,7169**	0,7164	0,7148	0,7150
107	0,7016	0,7022	0,7038	0,7052	0,7100	0,7126	0,7169	0,7196	0,7197	**0,7214**	0,7180	0,7172
137	0,7015	0,7039	0,7051	0,7061	0,7135	0,7164	0,7195	0,7223	**0,7238**	0,7196	0,7178	0,7177

Dengan memperhatikan parameter yang menghasilkan nilai AUC terbesar, maka pada penelitian ini model klasifikasi GMDKNN dibangun dengan menggunakan nilai parameter *k* yaitu *k* = 137 dan nilai parameter *p* yaitu *p* = -1. Tingginya nilai parameter *k* ini menunjukkan bahwa model klasifikasi dapat lebih banyak memperhitungkan tetangga terdekat, serta mampu mengatasi sensitivitas terhadap tetangga terdekat ukuran *k*. Selain itu, berdasarkan nilai AUC juga ditunjukkan bahwa performa klasifikasi yang optimal cenderung dihasilkan oleh parameter *p* yang bernilai negatif.

4.3.3 *Multi Local Mean Vectors*

Setelah sebelumnya dihitung jarak *Euclid* untuk semua data latih, selanjutnya pada masing-masing kelas ditentukan *k* tetangga terdekat yaitu 137 tetangga terdekat berdasarkan kedekatan jarak *Euclid*. Dari 137 tetangga terdekat ini, pada masing-masing kelas kemudian dihitung nilai *multi local mean vectors* atau *k local mean vectors* menggunakan i tetangga terdekat dari tiap kelas, dimana i bernilai $1 \leq i \leq k$ atau i bernilai 1 hingga 137 sehingga diperoleh hasil perhitungan sejumlah 137 *local mean vectors* untuk masing-masing kelas 1 dan kelas 2. Setelah diperoleh nilai *multi local mean vectors*, kemudian dihitung kembali jarak Euclid antara *multi local mean vectors* terhadap data uji. Hasil perhitungan *multi local mean vectors* beserta jarak *Euclid* nya pada kelas 1 dapat ditampilkan pada Tabel sebagai berikut:

Tabel 6. *Multi Local Mean Vectors* Kelas 1 Beserta Jarak *Euclid* nya

No	Usia	Jumlah Tanggungan Anak	Lama Bekerja	Lama Perusahaan	Pendapatan	Pinjaman yang Diajukan	Durasi Kredit	Jarak *Euclid*
1	1,9247	0,9792	2,2666	0,8514	0,0079	-0,3993	-1,2160	0,6413
2	2,0388	0,9792	2,3255	0,8514	0,0381	-0,3190	-1,2160	0,6444
3	1,9627	0,9792	2,2666	0,8514	-0,1003	-0,4528	-1,2160	0,5888
.	…	…	…	…	…	…	…	…
136	1,7786	0,7910	2,1048	0,7228	0,0199	-0,4385	-1,0531	0,8882
137	1,7805	0,7924	2,1051	0,7122	0,0168	-0,4404	-1,0581	0,8839

Hasil perhitungan *multi local mean vectors* beserta jarak *Euclid* nya pada kelas 2 dapat ditampilkan pada Tabel sebagai berikut:

Tabel 7. *Multi Local Mean Vectors* Kelas 2 Beserta Jarak *Euclid* nya

No	Usia	Jumlah Tanggungan Anak	Lama Bekerja	Lama Perusahaan	Pendapatan	Pinjaman yang Diajukan	Durasi Kredit	Jarak *Euclid*
1	2,0388	0,9792	2,1489	0,9044	-0,1460	-0,8006	-1,7394	0,5011
2	2,0388	0,9792	2,2078	0,8779	-0,1836	-0,7605	-1,4777	0,5006
3	2,0388	0,9792	2,4236	1,0457	-0,2909	-0,6401	-1,5650	0,3589
.	…	…	…	…	…	…	…	…
136	1,7660	0,8353	2,1368	0,7131	-0,1923	-0,4687	-1,1159	0,7659
137	1,7672	0,8418	2,1352	0,7064	-0,1912	-0,4670	-1,1167	0,7655

4.3.4 *Multi Generalized Mean Distances*

Multi generalized mean distances pada tiap kelas, diperoleh dengan menghitung *generalized mean* atau *power mean* dari 1 jarak *Euclid* pertama, 2 jarak *Euclid* pertama, 3 jarak *Euclid* pertama, seterusnya hingga 137 jarak *Euclid* antara data uji ke-4.667 dengan *multi local mean vectors*. Dari hasil perhitungan ini kemudian pada masing-masing kelas diperoleh 137 *generalized mean distances* untuk parameter *p* = -1. Berdasarkan hasil perhitungan, *multi generalized mean distances* 137 tetangga terdekat di kelas 1 dapat dinyatakan sebagai berikut:

$$\boldsymbol{g}^1 = \{g(\boldsymbol{x}, \boldsymbol{U}_{1\,1}^{NN}), g(\boldsymbol{x}, \boldsymbol{U}_{2\,1}^{NN}), g(\boldsymbol{x}, \boldsymbol{U}_{3\,1}^{NN}) \ldots g(\boldsymbol{x}, \boldsymbol{U}_{137\,1}^{NN}) \}$$

$$\boldsymbol{g}^1 = \{0{,}6413;\ 0{,}6428;\ 0{,}6237; \ldots 0{,}7396 \}$$

adapun *multi generalized mean distances* 137 tetangga terdekat di kelas 2 dapat dinyatakan sebagai berikut:

$$\boldsymbol{g}^2 = \{g(\boldsymbol{x}, \boldsymbol{U}_{1\,2}^{NN}), g(\boldsymbol{x}, \boldsymbol{U}_{2\,2}^{NN}), g(\boldsymbol{x}, \boldsymbol{U}_{3\,2}^{NN}) \ldots g(\boldsymbol{x}, \boldsymbol{U}_{137\,2}^{NN})\}$$

$$\boldsymbol{g}^2 = \{0{,}5011;\ 0{,}5009;\ 0{,}4425; \ldots 0{,}5196\}$$

4.3.5 *Nested Generalized Mean Distance*

Nilai *nested generalized mean distance* pada tiap kelas diperoleh dengan menghitung *generalized mean* dari seluruh nilai *generalized mean distances* pada *multi generalized mean distances* untuk tiap kelas. Dengan bantuan aplikasi MatLab diperoleh nilai n*ested generalized mean distance* pada kelas 1 dan kelas 2, dimana pada kelas 1 hasil perhitungannya adalah sebagai berikut:

$$G(\boldsymbol{x}, g^1) = \left(\frac{1}{137}\sum_{r=1}^{137}\left(g(\boldsymbol{x}, \boldsymbol{U}_{r\,1}^{NN})\right)^{-1}\right)^{\frac{1}{-1}} = 0{,}6750$$

adapun pada kelas 2 diperoleh hasil perhitungan nilai *nested generalized mean distance* sebagai berikut:

$$G(\boldsymbol{x}, g^2) = \left(\frac{1}{137}\sum_{r=1}^{137}\left(g(\boldsymbol{x}, \boldsymbol{U}_{r\,2}^{NN})\right)^{-1}\right)^{\frac{1}{-1}} = 0{,}4383$$

Pada perhitungan ini nilai *nested generalized mean distance* pada kelas 2 adalah $G(\boldsymbol{x}, g^2) = 0{,}4383$ memiliki nilai yang lebih kecil dibandingkan nilai *nested generalized mean distance* pada kelas 1 adalah $G(\boldsymbol{x}, g^1) = 0{,}6750$, dengan demikian menggunakan metode klasifikasi GMDKNN data uji ke-4.667 yang merupakan data uji terakhir diklasifikasikan ke dalam kelas 2 atau kelas dengan label 1 yaitu kelas kategori keputusan akhir *credit scoring* ditolak.

4.4 Evaluasi Performansi Ketepatan Klasifikasi

Berdasarkan perhitungan yang telah dilakukan, hasil klasifikasi menggunakan metode GMDKNN dapat dilihat pada tabel Matriks Konfusi berikut:

Tabel 8. Matriks Konfusi Metode GMDKNN

Hasil Observasi (*Actual Class*)	Hasil Prediksi (*Predicted Class*)	
	Kelas 2 (Ditolak)	Kelas 1 (Diterima)
Kelas 2 (Ditolak)	1.196	770
Kelas 1 (Diterima)	434	2.267

Berdasarkan nilai matriks konfusi pada Tabel 8, klasifikasi memiliki ukuran kinerja sistem yang dinyatakan dengan nilai APER = 0,2580, akurasi = 74,20%, sensitivitas = 0,6083, spesifisitas =0,8393, AUC = 0,7238, dan *G-Mean* = 0,7146. Dengan nilai AUC = 0,7238 maka model klasifikasi berada dalam kategori *Good* dengan kata lain model klasifikasi metode GMDKNN merupakan model klasifikasi yang baik untuk diterapkan pada kasus klasifikasi keputusan akhir permohonan kredit calon debitur Kredit Tanpa Agunan (KTA) pada proses *credit scoring* Bank X di Provinsi DKI Jakarta.

5. KESIMPULAN

Metode GMDKNN dapat diterapkan dalam mengklasifikasikan keputusan akhir permohonan kredit calon debitur KTA pada proses *credit scoring* Bank X di Provinsi DKI Jakarta. Melalui seleksi fitur uji *Mann-Whitney*, variabel usia, tanggungan anak, lama bekerja, lama perusahaan, pendapatan, pinjaman yang diajukan, dan durasi kredit valid untuk dimasukkan dalam model klasifikasi. Parameter *k* dan *p* yang memberikan model klasifikasi dengan performansi ketepatan klasifikasi terbaik berdasarkan percobaan dengan cara *trial error* adalah k = 137 dan p = -1. Dalam penelitian ini diperoleh ukuran kinerja sistem klasifikasi yang dinyatakan dengan nilai APER = 0,2580, akurasi = 74,20%, sensitivitas =0,6083, spesifisitas = 0,8393, AUC = 0,7238, dan *G-Mean* = 0,7146. Dengan nilai AUC = 0,7238 maka model klasifikasi metode GMDKNN merupakan model klasifikasi yang baik untuk diterapkan pada kasus klasifikasi keputusan akhir permohonan kredit calon debitur KTA pada proses *credit scoring* Bank X di Provinsi DKI Jakarta.

DAFTAR PUSTAKA

Beckmann, M., Ebecken, N. F. F., Lima B. S. L. 2015. A KNN Undersampling Approach for Data Balancing. *Journal of Intelligent Learning Systems and Applications* Vol. 7: Hal. 104-116.

Bekkar, M., Djemma, H. K., & Alitouche, T. A. 2013. Evaluation Measures for Models Asessment Over Imbalanced Data Sets. *Journal of Information Engineering and Applications* Vol. 3, No. 10.

Conover, W. J. 1980. *Practical Nonparametric Statistics*. 2 ed. New York: John Wiley and Sons, Inc.

Daniel, W. W. 1989. *Statistika Nonparametrik Terapan*. Diterjemahkan oleh: Alex Tri Kantjono W. Jakarta: PT Gramedia. Terjemahan dari: Applied Nonparametric Statistics

Fensterstock, A. 2005. Credit Scoring and the Next Step. *Business Credit* Vol. 107, No. 3: Hal. 46-49.

Firdaus, R., & Ariyanti, M. 2009. *Manajemen Perkreditan Bank Umum*. Bandung : Alfabeta.

Gou, J., Ma, H., Ou, W., Zeng, S., Rao, Y., & Yang, H. 2019. A Generalized Mean Distance-based *k*-Nearest Neighbor Classifier. *Expert Systems With Applications* Vol. 115: Hal. 356-372.

Han, J., & Kamber, M. 2006. *Data Mining : Concept and Techniques*, Second Edition. London : Morgan Kaufmann Publishers.

Johnson, R. A., Winchern, D.W. 2007. *Applied Multivariate Statistical Analysis*. *Sixth Edition*. New Jersey: Prentice Hall International. Inc.

Larose, D. T. 2005. *Discovering Knowledge in Data: An Introduction to Data Mining*. New Jersey : John Wiley & Sons, Inc.

Morrison, J. 2004. Introduction to Survival Analysis in Business. *The Journal of Business Forecasting Methods & Systems* Vol. 23, No. 1: Hal. 18-22.

Prasetyo, E. 2014. *Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab*. Yogyakarta : ANDI.

Rama, D. V., & Jones, F. L. 2009. *Sisitem Informasi Akuntansi*. Jakarta : Salemba Empat.

Republik Indonesia. 1998. Undang-Undang No. 10 Tahun 1998 tentang Perbankan. Lembaran Negara Tahun 1998. Sekretariat Negara. Jakarta.

Sinungan, M. 1991. *Dasar-Dasar dan Teknik Managemen Kredit*. Jakarta : Bumi Aksara.

Turban, E., Aronson, J. E., Liang, T. 2005. *Decision Support Systems and Intelligent Systems*. New Jersey : Prentice-Hall, Inc.

Yuchun, T., Ya-Qing, Z., Chawla, N. V., & Sven, K. 2002. SVMs Modeling for Highly Imbalanced Classification. Journal of Latex Class Files Vol. 1, No. 11.