PENGARUH METODE PENGUKURAN JARAK DAN *SMOTE* PADA KLASIFIKASI PENILAIAN KREDIT

Rifal Fadilah[1], Yulison Herry Chrisnanto[2], Gunawan Abdillah[3]

[123,]Informatika, Fakultas Sains & Informatika, Universitas Jenderal Achmad Yani

Jln. Terusan Jend. Sudirman, Kota Cimahi, Jawa Barat 40525
[1] rifalfadilah20@if.unjani.ac.id, [2] yhc@if.unjani.ac.id*, [3]gunawanabdillah03@gmail.com

Abstract

Credit plays an important role in supporting the economic growth of society. However, credit card users are reminded to use credit cards wisely according to their needs, and always make payments on time according to the bills that have been set, in order to avoid bad credit problems.To overcome this problem, a machine learning that is able to detect as early as possible high-risk credit card accounts is crucial to reduce losses caused by bad credit payments.This research proposes the development of techniques in machine learning to help classify bad credit using the KNN algorithm by referring to previous research on distance matrix measurement. The German credit data set obtained from UCI Machine Learning has an imbalance class where the number of positive classes is greater than the negative class. From testing using various distance measurement methods and with the application of SMOTE. The test shows that handling class imbalance using SMOTE algorithm significantly improves the accuracy, precision, recall, f1-score, and g-mean of KNN. The implementation of 5 distance measurement methods shows that Manhattan distance provides the best performance. The Manhattan distance provides the highest accuracy result of 84% after applying the SMOTE technique in the K = 5 test, while the Canberra distance also shows strong performance. The Manhattan method is superior in providing accurate classification results compared to other methods.

Keywords : *bad credit, classification, KNN, distance matrix, SMOTE*

Abstrak

Kredit berperan penting untuk menunjang pertumbuhan ekonomi masyarakat. Meski demikian, pengguna kartu kredit diingatkan untuk menggunakan kartu kredit secara bijak sesuai kebutuhan, dan selalu melakukan pembayaran tepat waktu sesuai tagihan yang telah ditetapkan, guna menghindari masalah kredit macet Untuk mengatasi masalah tersebut dibutuhkan sebuah pembelajaran mesin yang mampu mendeteksi sedini mungkin terhadap akun kartu kredit yang beresiko tinggi menjadi krusial untuk mengurangi kerugian yang disebabkan oleh pembayaran kredit macet.penelitian kali ini mengusulkan pengembangan Teknik dalam mechine learning untuk membantu mengklasifikasi kredit macet dengan menggunakan algoritma KNN dengan mengacu kepada penelitian sebelumnya mengenai pengukuran matrix jarak. Dataset German credit data yang di dapatkan dari UCI Mechine Learning memiliki imbalance class Dimana jumlah class positif lebih besar dari class negative. Dari pengujian menggunakan berbagai metode pengukuran jarak serta dengan penerapan SMOTE. Pengujian menunjukkan penanganan ketidakseimbangan kelas menggunakan algoritma SMOTE meningkatkan akurasi, presisi, recall, f1-score, dan g-mean pada KNN secara signifikan. Implementasi 5 metode pengukuran jarak menunjukkan bahwa jarak Manhattan memberikan performa terbaik. Jarak Manhattan memberikan hasil akurasi tertinggi sebesar 84% setelah dilakukan penerapan Teknik SMOTE pada pengujian K = 5, sementara jarak Canberra juga menunjukkan performa yang kuat. Metode Manhattan lebih unggul dalam memberikan hasil klasifikasi yang akurat dibandingkan dengan metode lainnya.

Kata kunci : Kredit Macet, Klasifikasi*, KNN,* Pengukuran Jarak, *SMOTE*

1. PENDAHULUAN

Kredit memiliki peran penting dalam mendukung pertumbuhan ekonomi masyarakat. Berdasarkan Undang-Undang Republik Indonesia Nomor 10 Tahun 1998 tentang Perubahan atas Undang-Undang Nomor 7 Tahun 1992 tentang Perbankan, kredit didefinisikan sebagai penyediaan dana atau setara dengannya, yang dilakukan berdasarkan perjanjian dan kesepakatan antara calon nasabah dengan lembaga keuangan. Pemberian pinjaman atau kredit bukanlah hal yang asing di Indonesia, karena hampir di seluruh pelosok nusantara terdapat tempat atau lembaga yang khusus menyediakan pinjaman bagi individu yang membutuhkan[1]. Meskipun demikian, pengguna kartu kredit harus diingatkan untuk menggunakan kartu kredit secara bijak sesuai dengan kebutuhan mereka. Selain itu, mereka juga perlu selalu melakukan pembayaran tepat waktu sesuai dengan tagihan yang ditetapkan. Hal ini dilakukan untuk menghindari masalah kredit macet. Kredit macet terjadi ketika nasabah tidak mampu membayar atau melunasi pinjamannya kepada penyedia layanan kartu kredit. Situasi ini dapat berpotensi menimbulkan kerugian bagi kedua belah pihak, baik bagi nasabah maupun bagi penyedia layanan kartu kredit[2]. Untuk mengatasi masalah tersebut, diperlukan sistem pembelajaran mesin yang mampu mendeteksi sedini mungkin akun kartu kredit yang berpotensi tinggi untuk mengalami kredit macet. Hal ini menjadi krusial dalam upaya mengurangi kerugian yang timbul akibat pembayaran kredit yang tidak tepat waktu atau macet.

Klasifikasi merupakan salah satu teknik data mining yang paling populer. Tujuan dari teknik ini adalah untuk memperkirakan kelas objek yang labelnya tidak diketahui. Klasifikasi dapat digunakan dalam berbagai aspek, sehingga metodenya telah berkembang seiring berjalannya waktu. Namun, klasifikasi sering menghadapi masalah, salah satunya adalah ketidakseimbangan data. Pengklasifikasian cenderung membuat model pembelajaran menjadi bias, sehingga prediksi terhadap kelas minoritas memiliki akurasi yang lebih rendah dibandingkan dengan prediksi terhadap kelas mayoritas[3]. Klasifikasi data kelas tidak seimbang adalah masalah utama dalam data mining karena hampir semua algoritma klasifikasi akan lebih akurat untuk kelas mayoritas daripada kelas minoritas ketika bekerja dengan data tidak seimbang. Seringkali, data kelas minoritas diklasifikasikan sebagai kelas mayoritas karena dataset yang tidak seimbang menyebabkan hasil klasifikasi yang salah atau tidak akurat[4]. Dengan menggunakan algoritma klasifikasi tanpa mempertimbangkan keseimbangan kelas, prediksi yang akurat tentang kelas mayoritas dan minoritas diabaikan. Algoritma klasifikasi akan memiliki kinerja yang buruk jika digunakan langsung pada kumpulan data yang tidak seimbang[5].

Salah satu algoritma pembelajaran mesin yang sangat terkenal adalah *K-Nearest Neighbor* (KNN)[6]. Konsep dasar dari algoritma KNN adalah menggunakan fungsi jarak untuk mengukur kemiripan antara data dan mengelompokkannya berdasarkan tingkat kemiripan tersebut. Meskipun fungsi jarak *Euclidean* adalah yang paling umum digunakan dalam algoritma ini, namun bukan berarti itu merupakan pilihan terbaik untuk semua kasus klasifikasi. Oleh karena itu, penting untuk meneliti fungsi jarak yang optimal untuk setiap kasus klasifikasi. Sebuah studi sebelumnya telah mengeksplorasi penggunaan fungsi jarak terbaik dalam algoritma KNN untuk mengklasifikasikan penerima Kartu Indonesia Pintar. Hasil penelitian tersebut menunjukkan bahwa kombinasi fungsi jarak Mahalanobis dan KNN memberikan kinerja terbaik untuk data dengan dimensi yang besar, sedangkan kombinasi fungsi jarak Manhattan dan KNN lebih cocok untuk data dengan dimensi yang lebih kecil[6].

Pada penelitian selanjutnya yang membahas Fungsi jarak terbaik pada Algoritma KNN pada Klasifikasi Kebakaran Hutan dan Lahan menunjukkan bahwa fungsi jarak Manhattan menghasilkan performa terbaik dalam mengidentifikasi jenis kebakaran hutan dan lahan dengan akurasi 92,6%[7].

Penelitian terdahulu lainya menunjukkan bahwa menggunakan metode SMOTE untuk menangani ketidakseimbangan kelas data telah memberikan hasil yang signifikan. Algoritma ini berhasil meningkatkan nilai G-mean dan F-measure secara substansial dibandingkan dengan pendekatan yang hanya menggunakan KNN. Secara khusus, rata-rata nilai performa G-mean mencapai 81,0% untuk skema SMOTE + KNN, jauh lebih tinggi dibandingkan dengan skema KNN yang hanya mencapai 53,4%. Selain itu, rata-rata performa F-measure untuk skema SMOTE + KNN mencapai 81,8%, sedangkan untuk skema KNN hanya mencapai 38,7%. Hal ini menunjukkan bahwa integrasi SMOTE dengan KNN efektif dalam meningkatkan kemampuan algoritma untuk menangani ketidakseimbangan kelas dalam data klasifikasi[8].

Oleh karena itu, penelitian ini mengusulkan analisis dalam tiga aspek utama: pertama, pengklasifikasian kredit macet menggunakan algoritma k-NN; kedua, analisis pengaruh metode pengukuran jarak seperti Euclidean, Canberra, Chebyshev, Manhattan, dan Minkowski pada kinerja algoritma k-NN; ketiga, analisis pengaruh teknik SMOTE terhadap penyeimbangan kelas dalam dataset kredit macet serta kinerja model klasifikasi. Dataset yang digunakan diambil dari Kaggle, yang menyediakan data terstruktur yang relevan untuk tujuan penelitian ini.

2. TINJAUAN LITERATUR

2.1 Data Mining

Data mining adalah proses mengekstraksi dan mengidentifikasi informasi berguna dengan menggunakan matematika, statistik, kecerdasan buatan, dan pembelajaran mesin. Data mining adalah suatu proses menemukan hubungan yang berarti, pola, dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika[9]. Berdasarkan fungsinya, data mining dikelompokkan menjadi deskripsi, estimasi, prediksi, klasifikasi, clustering, dan asosiasi[10]. Pengolahan data terdiri dari tiga langkah utama. Pertama, data dipilih, dibersihkan, dan diproses sesuai pedoman dan pengetahuan ahli domain, yang menangkap dan mengintegrasikan data internal dan eksternal untuk memberikan tinjauan organisasi yang lengkap. Pada langkah ini, algoritma data mining digunakan untuk menggali data yang terintegrasi, memudahkan identifikasi informasi berharga. Namun, semakin besar data yang diolah, semakin besar pula waktu prosesnya.

2.2 Klasifikasi

Klasifikasi adalah metode untuk memprediksi kelas atau properti setiap instance data. Model prediksi memungkinkan untuk memprediksi nilai variabel yang tidak diketahui berdasarkan nilai variabel lainnya. Klasifikasi juga dikenal sebagai supervised learning karena kelas data telah ditentukan sebelumnya[11].

2.3 SMOTE

Synthetic Minority Over-Sampling Technique (SMOTE) adalah teknik untuk menyeimbangkan distribusi data dengan menambah jumlah sampel pada kelas minoritas sehingga sebanding dengan kelas mayoritas. Teknik ini menghasilkan data sintetis dari kelas minoritas. SMOTE bekerja dengan mencari k tetangga terdekat (nearest neighbors) untuk setiap sampel di kelas minoritas, lalu membuat data sintetis berdasarkan persentase duplikasi yang diinginkan dengan memilih secara acak antara sampel minoritas dan tetangga terdekatnya[12].

2.4 KNN

Algoritma *K-Nearest Neighbor* (KNN) mengklasifikasikan data menggunakan set data pembelajaran (*train data*), yang diambil dari K tetangga terdekatnya (*nearest neighbor),* di mana K adalah banyaknya tetangga terdekat[13]. Dengan kata lain, tujuan dari algoritma ini adalah untuk mengklasifikasikan objek baru berdasarkan atribut dan sampel data pelatihan.

Terdapat 5 fungsi jarak yang akan di uji pada penelitian ini, diantaranya adalah:

a. Eulidean

Euclidean Distance merupakan metode pengukuran jarak antara dua objek yang cukup terpopuler. Metode ini mengukur kedekatan antara dua buah objek yang memiliki nilai atribut bersifat numerik [13]. Selain itu, teknik ini juga berfungsi sebagai pengukur jarak geometris dalam algoritma dasar k-NN. Berikut merupkan rumus perhitungan Jarak Euclidean.

$$d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ij} - x_{jk})2} \qquad (1)$$

b. Canberra

2.5 Pada tahun 1966, Godfrey N. Lance dan William T. Williams pertama kali menggunakan metode pengukuran jarak Canberra. Ini biasanya digunakan untuk mengukur jarak antara dua titik dalam ruang vektor[14]. Jarak Canberra dapat dihitung menggunakan persamaan (2).

$$d_{ij} = \sum_{k=1}^{n} \frac{|x_{ij}-x_{jk}|}{|x_{ik}-x_{jk}|} \qquad (2)$$

c. Chebyshev

Metode Chebyshev menghitung jarak dengan menggunakan nilai absolut dari selisih pasangan koordinat titik[15].

Berikut merupakan perhitungan untuk jarak Chebyshev.

$$d(x,y) = max_{i=1}^{n}|x_i - y_i| \quad (3)$$

d. Manhattan
Jarak Manhattan, sering dikenal sebagai "city block," adalah metode pengukuran geometris antara dua objek. Perhitungan Manhattan digunakan untuk menghitung perbedaan absolut (mutlak) antara kordinat sepasang objek[16]. Berikut merupakan perhitungan untuk jarak Manhattan.

$$d(x,y) = \sum_{i=1}^{m}|x_i - y_i| \quad (4)$$

e. Minkowski
Jarak Minkowski, dinamai berdasarkan matematikawan Hermann Minkowski, adalah metrik yang menghitung jarak antara dua titik dalam ruang vektor bernorma. Berikut merupakan perhitungan Jarak Minkowski menggunakan persamaan (5).

$$d_p(x,y) = \sqrt[p]{\sum_{i=1}^{n}|x_i - y_i|^p} \quad (5)$$

2.6 Cross Validation

Salah satu metode validasi model adalah *cross-validation*, juga dikenal sebagai estimasi rotasi. Teknik ini terutama digunakan untuk memprediksi model dan menilai akurasi model prediktif dalam praktik. K-fold cross-validation adalah metode validasi silang yang membagi data menjadi K bagian yang sama besar [17]. Validasi silang dengan k-fold=10 menghilangkan bias pada data. Pelatihan dan tes dilakukan sebanyak K kali

2.7 Evaluation

Evaluasi klasifikasi dilakukan dengan menguji objek yang diklasifikasikan secara benar dan salah. Validasi ini berguna untuk menentukan jenis model terbaik berdasarkan hasil klasifikasi[8]. Evaluasi dalam penelitian ini menggunakan confusion matrix. Dimana Confusion matrix sendiri adalah teknik untuk mengevaluasi ketepatan model klasifikasi yang dikembangkan untuk memisahkan data ke dalam berbagai kelas. Kinerja model klasifikasi dapat ditentukan dengan mengukur tingkat akurasinya. Persentase kelompok data latih yang berhasil diklasifikasikan dapat digunakan untuk menentukan akurasi klasifikasi. Nilai accuracy, recall, precision, f1-score, dan g-mean akan ditentukan dengan menggunakan tabel confusion matrix untuk memeriksa hasil pengujian model klasifikasi[18].

TABEL I. CONFUSION MATRIX

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

Dari Tabel Confusion Matrix diatas akan ditentukan perhitungan evaluasi matrix seperti pada rumus berikut :

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

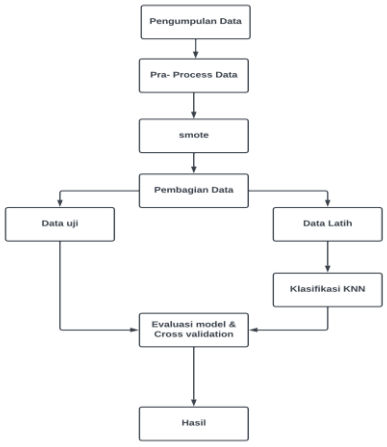
$$F1\text{-}Score = 2 \times \frac{Precision\ x\ Recall}{Precision+Recall} \quad (9)$$

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

$$G\text{-}mean = \sqrt{Recall\ \times Spesifisitas} \quad (11)$$

3. METODOLOGI PENELITIAN

3.1 Skema Alur Penelitian



Gambar 1 Alur Penelitian

3.2 Pengumpulan Data

Metode pengumpulan data adalah teknik atau cara yang dilakukan oleh peneliti untuk mengumpulkan informasi yang dibutuhkan dalam rangka mencapai tujuan penelitian. Pada penelitian ini, data yang digunakan berasal dari data starlog (german credit data) yang terdiri dari 1.000 data. Data ini diperoleh dari UCI Machine Learning Repository, dengan rincian 700 data yang memiliki label positif (0) dan 300 data yang memiliki label negatif (1).

3.3 Pra-Process Data

Preprocessing adalah tahapan kritis dalam analisis data yang bertujuan untuk mengubah data mentah dari proses pengambilan data menjadi format yang lebih efisien untuk digunakan dalam proses klasifikasi.[19]. Tujuan dari *Preprocessing* untuk meningkatkan kualitas dan kegunaan data sehingga model yang akan dibangun berkinerja dengan baik. Pada tahap ini, data yang telah dikumpulkan akan dilakukan *preprocessing* agar data dapat diolah. Data akan di bersihkan dari *missing value*, diubah menjadi nilai numerik, dan dilakukan standarisasi menggunakan *min-max scaller*.

3.4 SMOTE

Pada tahap ini dilakukan penerapan teknik SMOTE untuk mengatasi ketidakseimbangan kelas dalam dataset. SMOTE bekerja dengan menghasilkan contoh sintetik dari kelas minoritas untuk menyeimbangkan proporsi antara kelas mayoritas dan minoritas. Alasan penggunaan SMOTE adalah untuk meningkatkan kinerja model dengan memberikan model kesempatan belajar yang lebih baik dari data yang seimbang, mengurangi risiko overfitting yang sering terjadi dengan data kelas minoritas yang sedikit, memperbaiki metode evaluasi sehingga metrik seperti F1-score, recall, dan precision memberikan gambaran kinerja model yang lebih akurat, serta mempertahankan variabilitas data dengan menghasilkan contoh sintetik berdasarkan kombinasi dari contoh yang ada. Dengan demikian, diharapkan model dapat mengenali dan memprediksi data dari kelas minoritas dengan lebih efektif, sekaligus menjaga keseimbangan keseluruhan dalam dataset.

3.5 Pembagian Data

Pada tahap ini, data akan dibagi menjadi dua kelompok utama, kelompok data latih dan kelompok data uji. Data latih merupakan data yang telah dilengkapi dengan label dan digunakan untuk mengembangkan model klasifikasi. Di sisi lain, data uji merupakan data yang tidak memiliki label dan digunakan untuk menguji keefektifan model yang telah dibuat.

Pembagian ini penting karena memungkinkan evaluasi yang objektif terhadap kinerja model yang dibangun. Dengan membagi data menjadi kedua kelompok ini, tujuannya adalah untuk memastikan bahwa model yang dikembangkan dapat menggeneralisasi dan bekerja dengan baik ketika diterapkan pada data baru yang belum pernah dilihat sebelumnya[20]. Umumnya rasio pembagian data latih dan uji yang ideal adalah 70-90%, sehingga pada penelitian ini bobot pembagian data uji dan data latih adalah 80%/20%.

3.6 Klasifikasi

Setelah melewati langkah pembagian data selanjutkan akan dilakukan pembuatan model klasifikasi menggunakan teknik K-NN (K-Nearest Neighbor). Pembuatan model akan dilakukan sebanyak 2 kali dimana salah satunya akan menggunakan Teknik smote. Pembuatan model diawali dengan menentukan jumlah tetangga / neighbor, Dimana pada penelitian ini jumlah K yang digunakan bervariasi yaitu 1,3, dan 5. Fitur-fitur tersebut dihitung menggunakan rumus jarak Euclidean, Canberra, Chebishev, Manhattan dan minkowski.

3.7 Evaluasi

Langkah selanjutnya dilakukan evaluasi performa pada kedua model (KNN tanpa SMOTE dan KNN dengan SMOTE) menggunakan metrik evaluasi yang sesuai, matrix evaluasi yang akan digunakan adalah confusion matrix seperti akurasi, precision, recall, f1-score, dan G-mean, serta dilakukan pengecekan silang menggunakan Cross Validation. Jumlah fold yang dipilih akan menentukan seberapa akurat sebuah model, semakin tinggi jumlah fold maka semakin tinggi juga akurasinya sehingga dipilih fold = 10 untuk Cross Validation.

4. HASIL DAN PEMBAHASAN

4.1 Pengumpulan Data

Penelitian ini menggunakan data Starlog German Credit dari UCI Machine Learning Repository, yang terdiri dari 1.000 sampel data.

4.2 Preprocessing Data

Pada tahap ini, data yang telah dikumpulkan akan menjalani preprocessing untuk mempersiapkannya agar dapat diolah. Proses ini meliputi pembersihan data dari missing value, konversi data menjadi nilai numerik, serta standarisasi menggunakan min-max scaler.

```
...                          Missing Values (%)
    Checking Account                        0.0
    Duration                                0.0
    Credit History                          0.0
    Purpose                                 0.0
    Credit Amount                           0.0
    Savings Account                         0.0
    Employment                              0.0
    Installment Rate                        0.0
    Personal Status                         0.0
    Other Debtors                           0.0
    Residence                               0.0
    Property                                0.0
    Age                                     0.0
    Other Installment Plans                 0.0
    Housing                                 0.0
    Existing Credits                        0.0
    Job                                     0.0
    Maintenance                             0.0
    Telephone                               0.0
    Foreign Worker                          0.0
    Credit Risk                             0.0
```

Gambar 2. Menghilangkan *Missing Value*

	Checking Account	Duration	Credit History	Purpose	Credit Amount	Savings Account	Employment	Installment Rate	Personal Status	Other Debtors	...	Property	Age	Other Installment Plans	Housing	Existing Credits	Job	
0	0	6	4	4	1169	4	4	4	2	0	...	0	67	2	1	2	2	
1	1	48	2	4	5951	0	2	2	1	0	...	0	22	2	1	1	2	
2	3	12	4	7	2096	0	3	2	2	0	...	0	49	2	1	1	1	
3	0	42	2	3	7882	0	3	2	2	2	...	1	45	2	2	1	2	
4	0	24	3	0	4870	0	2	3	2	0	...	3	53	2	2	2	2	
..	..	..	..	..	..	..	..	..	..	..	...	..	..	..	..	..	..	
95	1	54	0	9	11945	0	1	3	2	0	...	3	58	2	0	1	2	
96	3	12	4	7	2012	4	3	4	1	0	...	2	61	2	1	1	2	
97	1	18	2	9	2622	1	2	4	2	0	...	2	34	2	1	1	2	
98	1	36	4	4	2337	0	4	4	2	0	...	0	36	2	1	1	2	
99	1	20	3	1	7057	4	3	3	2	0	...	1	36	0	0	2	3	

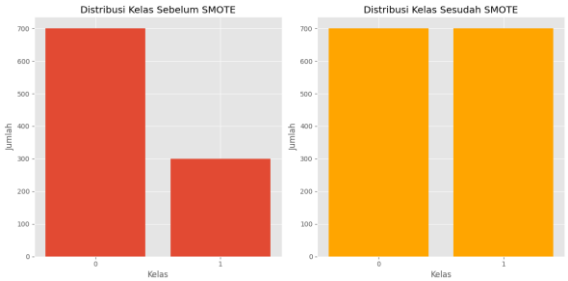
Gambar 3. Transformation Data

```
[[0.33333333 0.47058824 0.25       ... 0.         0.         0.        ]
 [1.         0.10294118 1.         ... 0.         0.         0.        ]
 [0.         0.11764706 0.5        ... 0.         0.         0.        ]
 ...
 [0.         0.11764706 0.5        ... 0.         0.         0.        ]
 [1.         0.29411765 1.         ... 0.         0.         0.        ]
 [0.         0.10294118 0.25       ... 0.         0.         0.        ]]
```

Gambar 4. Min-Max Scaller

4.3 SMOTE

Dataset German Credit Data sebelum dilakukan penyeimbangan data memiliki masing masing kelas kurang seimbang. Setelah dilakukan penyeimbangan data menggunakan Teknik SMOTE, frekuensi untuk masing masing kelas menjadi seimbang, seperti disajikan pada gambar 5.



Gambar 5. Penerapan Teknik *SMOTE*

Dari Gambar 5 dapat dilihat bahwa sebelum dilakukan penyeimbangan dataset jumlah masing masing kelas adalah 700 dan 300, sedangkan setelah dilakukan penyeimbangan dengan menggunakan Teknik SMOTE kedua kelas menjadi sama yaitu 700.

4.4 Klasifikasi

Setelah dilakukan penyeimbangan dataset menggunakan Teknik SMOTE maka akan dilakukan pelatihan model tanpa smote dan dengan smote menggunakan algoritma KNN dengan 5 pengukuran jarak yaitu Euclidean, Canberra, Chebishev, Manhattan dan minkowski serta menerapkan evaluasi cross validation. dengan perhitungan K=1,3 dan 5.

4.5 Euclidean

Pada Pengujian pertama, perhitungan dilakukan dengan menggunakan perhitungan jarak euclidean dengan K = 1,3 dan 5 dan cross validation = 10. Berikut adalah hasil performa model tanpa smote dan dengan smote.

TABEL II. EUCLIDEAN

Performa	KNN			KNN + SMOTE		
	1	3	5	1	3	5
Akurasi	0,74	0,74	0,74	0,78	0,80	0,79
Presisi	0,59	0,58	0,59	0,73	0,75	0,74
Recal	0,46	0,44	0,46	0,83	0,85	0,85
F-Measure	0,51	0,50	0,51	0,78	0,80	0,79
G-Mean	0,63	0,62	0,63	0,78	0,80	0,79
CV	0,67	0,73	0,73	0,78	0,77	0,78

Penggunaan teknik Synthetic Minority Over-sampling Technique (SMOTE) secara konsisten meningkatkan performa model KNN pada semua metrik yang diuji (Akurasi, Presisi, Recall, F-

measure, dan G-mean). Akurasi meningkat dari sekitar 0,74 menjadi hingga 0,80 saat menggunakan SMOTE. Presisi meningkat dari sekitar 0,58 menjadi hingga 0,75. Recall menunjukkan peningkatan signifikan dari sekitar 0,44 tanpa SMOTE menjadi hingga 0,85 dengan SMOTE. Untuk nilai K = 1, 3, dan 5, performa model tanpa SMOTE cenderung konstan, sementara dengan SMOTE, nilai K = 3 memberikan performa terbaik dalam hal Akurasi, Presisi, dan Recall. Dengan menggunakan SMOTE, model KNN tidak hanya lebih akurat tetapi juga lebih seimbang dalam mengklasifikasikan kelas minoritas, sebagaimana ditunjukkan oleh peningkatan nilai G-mean dari sekitar 0,62 tanpa SMOTE menjadi 0,80 dengan SMOTE.

4.6 Canberra

Pada Pengujian kedua, perhitungan dilakukan dengan menggunakan perhitungan jarak canberra dengan K = 1,3 dan 5. Berikut adalah hasil performa model tanpa smote dan dengan smote.

TABEL III. CANBERRA

Performa	KNN			KNN + SMOTE		
	1	3	5	1	3	5
Akurasi	0,72	0,73	0,75	0,80	0,81	0,81
Presisi	0,53	0,57	0,64	0,76	0,78	0,77
Recal	0,53	0,44	0,36	0,85	0,82	0,84
F-Measure	0,53	0,50	0,46	0,80	0,80	0,81
G-Mean	0,65	0,61	0,57	0,81	0,81	0,81
CV	0,69	0,71	0,73	0,78	0,80	0,77

Pada pengujian ke dua ini, akurasi meningkat dari sekitar 0,75 menjadi hingga 0,81 setelah diterapkan Teknik SMOTE. Presisi meningkat dari sekitar 0,64 menjadi hingga 0,77 dengan SMOTE. Recall menunjukkan peningkatan signifikan dari sekitar 0,36 menjadi hingga 0,84 dengan SMOTE. Untuk nilai K = 1, 3, dan 5, performa model tanpa SMOTE menunjukkan peningkatan kecil dengan meningkatnya nilai K, terutama pada metrik Akurasi dan Presisi, nilai K = 5 memberikan performa yang paling konsisten dan optimal. Dengan menggunakan SMOTE, model KNN tidak hanya lebih akurat tetapi juga lebih seimbang dalam mengklasifikasikan kelas minoritas, sebagaimana ditunjukkan oleh peningkatan nilai G-mean dari sekitar 0,57 menjadi 0,81 dengan SMOTE.

4.7 Cheyshev

Pada Pengujian ketiga, perhitungan dilakukan dengan menggunakan perhitungan jarak chebysev dengan K = 1,3 dan 5. Berikut adalah hasil performa model tanpa smote dan dengan smote.

TABEL IV. CHEYSHEV

Performa	KNN			KNN + SMOTE		
	1	3	5	1	3	5
Akurasi	0,70	0,71	0,70	0,73	0,72	0,71
Presisi	0,50	0,51	0,49	0,68	0,68	0,65
Recal	0,42	0,44	0,46	0,81	0,78	0,81
F-Measure	0,46	0,47	0,47	0,74	0,72	0,72
G-Mean	0,59	0,60	0,61	0,73	0,72	0,71
CV	0,67	0,66	0,71	0,75	0,75	0,74

Pada pengujian ke tiga ini, akurasi meningkat dari sekitar 0,70 menjadi hingga 0,73 dengan SMOTE. Presisi meningkat dari sekitar 0,50 menjadi hingga 0,68 dengan SMOTE. Recall menunjukkan peningkatan signifikan dari sekitar 0,42 menjadi hingga 0,81 dengan SMOTE. nilai K = 1 memberikan performa yang paling optimal. Dengan menggunakan SMOTE, model KNN tidak hanya lebih akurat tetapi juga lebih seimbang dalam mengklasifikasikan kelas minoritas, sebagaimana ditunjukkan oleh peningkatan nilai G-mean dari sekitar 0,59 menjadi 0,73 dengan SMOTE.

4.8 Manhattan

Pada Pengujian keempat, perhitungan dilakukan dengan menggunakan perhitungan jarak minkowski dengan K = 1,3 dan 5. Berikut adalah hasil performa model tanpa smote dan dengan smote.

TABEL V. MANHATTAN

Performa	KNN			KNN + SMOTE		
	1	3	5	1	3	5
Akurasi	0,74	0,73	0,74	0,82	0,81	0,84
Presisi	0,59	0,56	0,59	0,78	0,77	0,80
Recal	0,46	0,42	0,44	0,85	0,86	0,86
F-Measure	0,51	0,48	0,50	0,82	0,81	0,83

G-Mean	0,63	0,60	0,62	0,82	0,82	0,84
CV	0,68	0,75	0,74	0,79	0,82	0,80

Pada pengujian ke empat ini, akurasi meningkat dari sekitar 0,74 menjadi hingga 0,84 dengan SMOTE. Presisi meningkat dari sekitar 0,59 menjadi hingga 0,80 dengan SMOTE. Recall menunjukkan peningkatan signifikan dari sekitar 0,44 menjadi hingga 0,86 dengan SMOTE. nilai K = 5 memberikan performa yang paling optimal pada semua matrix evaluasi. Dengan menggunakan SMOTE, model KNN tidak hanya lebih akurat tetapi juga lebih seimbang dalam mengklasifikasikan kelas minoritas, sebagaimana ditunjukkan oleh peningkatan nilai G-mean dari sekitar 0,62 menjadi 0,84 dengan SMOTE.

4.9 Minkowski

Pada Pengujian kelima, perhitungan dilakukan dengan menggunakan perhitungan jarak minkowski dengan K = 1,3 dan 5. Berikut adalah hasil performa model tanpa smote dan dengan smote.

TABEL VI. MINKOWSKI

Performa	KNN			KNN + SMOTE		
	1	3	5	1	3	5
Akurasi	0,74	0,74	0,74	0,78	0,80	0,79
Presisi	0,59	0,58	0,59	0,73	0,75	0,74
Recal	0,46	0,44	0,46	0,83	0,85	0,85
F-Measure	0,51	0,50	0,51	0,78	0,80	0,79
G-Mean	0,63	0,62	0,63	0,78	0,80	0,79
CV	0,67	0,73	0,73	0,78	0,77	0,78

Pada pengujian ke lima ini, akurasi meningkat dari sekitar 0,74 menjadi hingga 0,80 dengan SMOTE. Presisi meningkat dari sekitar 0,58 menjadi hingga 0,75 dengan SMOTE. Recall menunjukkan peningkatan signifikan dari sekitar 0,44 menjadi hingga 0,85 dengan SMOTE. nilai K = 3 memberikan performa yang cukup optimal. Dengan menggunakan SMOTE, model KNN tidak hanya lebih akurat tetapi juga lebih seimbang dalam mengklasifikasikan kelas minoritas, sebagaimana ditunjukkan oleh peningkatan nilai G-mean dari sekitar 0,62 menjadi 0,80 dengan SMOTE.

4.10 Hasil Pengujian Pengukuran Jarak

Berdasarkan pengujian yang telah dilakukan sebelumnya, dilakukan perbandingan performa dari setiap fungsi jarak terbaik. Tabel 7 menyajikan perbandingan performa dari kelima metode pengukuran jarak dengan K terbaik setelah melalui tahap SMOTE.

TABEL VII. PERBANDINGAN PENGUKURAN JARAK

Performa	Euclidean	Canberra	Chebyshev	Manhattan	Minkowski
Akurasi	0,80	0,81	0,73	0,84	0,80
Presisi	0,75	0,77	0,68	0,80	0,75
Recall	0,85	0,84	0,81	0,86	0,85
F-Measure	0,80	0,81	0,74	0,83	0,80
G-Mean	0,80	0,81	0,73	0,84	0,80
CV	0,77	0,77	0,75	0,80	0,77

Jarak Manhattan secara konsisten menunjukkan performa terbaik dalam sebagian besar metrik evaluasi, menjadikannya pilihan paling efektif untuk klasifikasi dalam dataset ini. Manhattan menunjukkan akurasi tertinggi 0,84, ini menunjukkan penggunaan jarak Manhattan memberikan hasil yang lebih akurat dalam klasifikasi dibandingkan jarak Euclidean, Canberra, Chebyshev, dan Minkowski. Manhattan juga memberikan presisi tertinggi 0.80, diikuti oleh Canberra 0,77. Chebyshev memiliki presisi terendah 0,68. Manhattan memiliki recall tertinggi 0,86. Manhattan dan Canberra memiliki F-measure tertinggi (0,83 dan 0,81), menunjukkan keseimbangan yang baik antara presisi dan recall. Manhattan memiliki G-mean tertinggi 0,84, menunjukkan kemampuan terbaik dalam menyeimbangkan antara sensitivitas dan spesifisitas. Manhattan menunjukkan performa terbaik dalam validasi silang (CV) (0,80). Oleh karena itu, Model dengan jarak Manhattan lebih stabil dan konsisten dalam berbagai subset data.

5. Kesimpulan dan Saran

Penangan Distribusi kelas yang tidak seimbang pada dataset menggunakan algoritma SMOTE dapat meningkatkan nilai akurasi, presisi, recall, f1-score dan g-mean pada algoritma KNN. Hal tersebut menunjukan bahwa proses penanganan terhadap distribusi kelas yang tidak seimbang pada tahap pra-proses data memberikan pengaruh yang cukup signifikan.

Setelah melakukan implementasi 5 jenis metode pengukuran jarak pada algoritma KNN. Jarak Manhattan menunjukkan performa terbaik dibandingkan dengan jarak Euclidean, Canberra, Minkowski, dan Chebyshev. Jarak Manhattan konsisten memberikan akurasi tertinggi 0.84, presisi 0.80, Recal 0.86, F-measure 0.83, dan G-mean 0.84. Selain itu, Manhattan juga menunjukkan performa terbaik dalam validasi silang 0.80, menegaskan stabilitas dan konsistensinya dalam berbagai subset data. Jarak Canberra juga menunjukkan performa yang kuat dengan akurasi tinggi 0.81 dan presisi tertinggi 0.77.

Jarak Manhattan dan Canberra terbukti memberikan hasil klasifikasi yang lebih akurat dibandingkan jarak Euclidean, Minkowski, dan Chebyshev. Jarak Manhattan unggul dalam semua metrik evaluasi, termasuk akurasi, presisi, recall, F-measure, dan G-mean, menunjukkan keseimbangan yang baik antara sensitivitas dan spesifisitas. Sementara itu, jarak Canberra memberikan presisi yang cukup tinggi 0.77, yang penting untuk mengurangi false positives dalam klasifikasi kredit macet. Jarak Euclidean, Minkowski, dan Chebyshev menunjukkan performa yang kurang optimal, dengan Chebyshev memiliki presisi terendah 0.68. Dengan demikian, pilihan metode pengukuran jarak sangat mempengaruhi performa k-NN, dengan Manhattan sebagai pilihan yang lebih disarankan.

Daftar Pustaka:

[1] E. Ajeng, A. Putri, E. Nuraina, and E. E. Yusdita, "Upaya Pencegahan dan Penanganan Kredit Macet Ditinjau dari Persepsi Nasabah," *Jurnal Riset Akuntansi dan Perpajakan)*, vol. 7, no. 2, pp. 185–196, 2020.

[2] R. R. A. S. P. P. Putri Ayu Mardhiyah1, "Klasifikasi Untuk Memprediksi Pembayaran Kartu Kredit Macet Menggunakan Algoritma C4.5," *Jurnal Teknologia*, vol. 3, pp. 2654–5683, 2020, Accessed: Feb. 07, 2024. [Online]. Available: https://aperti.e-journal.id/teknologia/article/view/66/44

[3] J. Prasetya, "Penerapan Klasifikasi Naive Bayes Dengan Algoritma Random Oversampling Dan Random Undersampling Pada Data Tidak Seimbang Cervical Cancer Risk Factors," *Leibniz : Jurnal Matematika*, vol. 2, no. 2, pp. 11–22, 2022.

[4] M. Mustaqim, B. Warsito, and B. Surarso, "Combination of synthetic minority oversampling technique (Smote) and backpropagation neural network to handle imbalanced class in predicting the use of contraceptive implants," *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 5, no. 2, pp. 116–127, Jul. 2019, doi: 10.26594/register.v5i2.1705.

[5] M. Sulistiyono, Y. Pristyanto, S. Adi, and G. Gumelar, "Implementasi Algoritma Synthetic Minority Over-Sampling Technique untuk Menangani Ketidakseimbangan Kelas pada Dataset Klasifikasi," *SISTEMASI*, vol. 10, no. 2, p. 445, May 2021, doi: 10.32520/stmsi.v10i2.1303.

[6] I. M. K. Karo, A. Khosuri, and R. Setiawan, "Effects of Distance Measurement Methods in K-Nearest Neighbor Algorithm to Select Indonesia Smart Card Recipient," *2021 International Conference on Data Science and Its Applications (ICoDSA)*, pp. 209–214, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:244841914

[7] I. M. Karo Karo, A. Khosuri, J. S. I. Septory, and D. P. Supandi, "Pengaruh Metode Pengukuran Jarak pada Algoritma k-NN untuk Klasifikasi Kebakaran Hutan dan Lahan," *Jurnal Media Informatika Budidarma*, vol. 6, no. 2, p. 1174, Apr. 2022, doi: 10.30865/mib.v6i2.3967.

[8] R. Siringoringo, "Klasifikasi Data Tidak Seimbang Menggunakan Algoritma Smote Dan K-Nearest Neighbor," *Journal Information System Development (ISD)*, vol. 3, no. 1, pp. 44–49, 2018.

[9] D. Bencana and A. Akhlaqul, "Kinerja Metode C4.5 dalam Penyaluran Bantuan," *Prosiding Seminar Nasional Ilmu Komputer dan Teknologi Informasi*, vol. 3, no. 2, 2018.

[10] L. Setiyani, M. Wahidin, D. Awaludin, and S. Purwani, "Analisis Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Data Mining Naïve Bayes : Systematic Review," *Faktor Exacta*, vol. 13, no. 1, p. 35, Jun. 2020, doi: 10.30998/faktorexacta.v13i1.5548.

[11] P. Mayadewi and E. Rosely, "Prediksi Nilai Proyek Akhir Mahasiswa Menggunakan Algoritma Klasifikasi Data Mining," 2015. [Online]. Available: https://www.researchgate.net/publication/283570705

[12] E. Sutoyo, M. Asri Fadlurrahman, J. Telekomunikasi Jl Terusan Buah Batu, K. Dayeuhkolot, K. Bandung, and J. Barat, "Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network," *Jurnal Edukasi & Penelitian Informatika (JEPIN)*, vol. 6, no. 3, pp. 379–385, 2020.

[13] S. Mulyati, S. Maulana Husein, and K. Kunci, "Rancang Bangun Aplikasi Data Mining Prediksi Kelulusan Ujian Nasional Menggunakan Algoritma (KNN) K-Nearest Neighbor Dengan Metode Euclidean Distance Pada SMPN 2 PAGEDANGAN sistem dapat memprediksi dan mengklasifikasikan dengan baik dan cepat," pp. 65–73, 2020.

[14] S. Gultom, S. Sriadhi, M. Martiano, and J. Simarmata, "Comparison analysis of K-Means and K-Medoid with Ecluidience Distance Algorithm, Chanberra Distance, and Chebyshev Distance for Big Data Clustering," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, Oct. 2018. doi: 10.1088/1757-899X/420/1/012092.

[15] É. O. Rodrigues, "Combining Minkowski and Chebyshev: New distance proposal and survey of distance metrics using k-nearest neighbours classifier," 2018. [Online]. Available: www.elsevier.com/locate/patrec

[16] M. Nishom, "Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 4, no. 1, pp. 20–24, Jan. 2019, doi: 10.30591/jpit.v4i1.1253.

[17] H. Azis, P. Purnawansyah, F. Fattah, and I. P. Putri, "Performa Klasifikasi K-NN dan Cross Validation pada Data Pasien Pengidap Penyakit Jantung," *ILKOM Jurnal Ilmiah*, vol. 12, no. 2, pp. 81–86, 2020, doi: 10.33096/ilkom.v12i2.507.81-86.

[18] E. E. Ragil Dimas Himawan, "JEPIN (Jurnal Edukasi dan Penelitian Informatika) Perbandingan Akurasi Analisis Sentimen Tweet terhadap Pemerintah Provinsi DKI Jakarta di Masa Pandemi," *jurnal edukasi & penelitian informatika (JEPIN)*, vol. 7, no. 2, 2021, [Online]. Available: https://dx.doi.org/10.26418/jp.v7i1.41728

[19] R. Sumanti, P. Studi Sistem Informasi, U. Darwan Ali Jln Batu Berlian No, and S. - Kalimantan Tengah, "Klasifikasi Data Tingkat Kerawanan Kebakaran Menggunakan Algoritma CART," *Jurnal Informatika & Rekayasa Elektronika)*, vol. 7, no. 1, pp. 51–59, 2024, [Online]. Available: http://e-journal.stmiklombok.ac.id/index.php/jire

[20] A. Setiawan, R. Febrio Waleska, M. Adji Purnama, and L. Efrizoni, "KOMPARASI ALGORITMA K-NEAREST NEIGHBOR(K-NN), SUPPORT VECTOR MACHINE(SVM), DAN DECISION TREE DALAM KLASIFIKASI PENYAKIT STROKE," *Jurnal Informatika & Rekayasa Elektronika)*, vol. 7, no. 1, pp. 107–114, 2024, [Online]. Available: http://e-journal.stmiklombok.ac.id/index.php/jire ISSN.2620-6900