1213 PENGARUH METODE PENGUKURAN JARAK DAN SMOTE PADA KLASIFIKASI PENILAIAN KREDIT

By Rifal Fadilah

WORD COUNT 3564
TIME SUBMITTED 15-JUN-2024 09:28AM
PAPER ID 109654025

Template ini berlaku untuk terbitan Volume 6 Nomor 1 April 2023

PENGARUH METODE PENGUKURAN JARAK DAN SMOTE PADA KLASIFIKASI PENILAIAN KREDIT

Rifal Fadilah

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 calculation methods. 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 matrix 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 a strong performance. method Manhattan method is superior in providing accurate classification results compared to other methods. 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 metode perhitungan 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 berperan penting untuk menunjang pertumbuhan ekonomi masyarakat. Berdasarkan Undang-Undang Republik Indonesia Nomor 10

Tahun 1998 Tentang Perubahan Atas Undang-Undang Nomor 7 Tahun 1992 Tentang Perbankan (1998), menyatakan kredit adalah penyediaan dana yang dapat dipersamakan dengan itu, berdasarkan perjanjian serta kesepakatan bersama antara calon nasabah dan pihak lembaga keuangan. Pemberian pinjaman atau kredit sudah bukan sesuatu yang asing lagi di Indonesia bahkan hampir di seluruh pelosok nusantara selalu ada tempat atau lembaga yang khusus menyediakan pinjaman bagi setiap individu yang membutuhkan[1].

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. Kredit macet terjadi ketika nasabah tidak mampu membayar atau melunasi pinjamannya kepada penyedia layanan kartu kredit, yang pada akhirnya dapat menyebabkan kerugian bagi kedua belah pihak[2]. 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.

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. Jika distribusi kelas data tidak seimbang, jumlah kelas data *(instance)* yang satu lebih sedikit atau lebih banyak daripada jumlah kelas data lainnya[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 menyelidiki fungsi jarak yang optimal untuk setiap kasus klasifikasi. Sebuah penelitian sebelumnya telah menginvestigasi penggunaan fungsi jarak terbaik pada algoritma KNN untuk mengklasifikasikan penerima Kartu Indonesia Pintar. Hasil penelitian tersebut menunjukkan bahwa kombinasi fungsi jarak *Mahalanobis* dan KNN memberikan kinerja terbaik untuk dimensi data yang besar, sementara kombinasi fungsi jarak *Manhattan* dan KNN lebih baik untuk dimensi data yang lebih kecil. Pada penelitian selanjutnya dengan judul Pengaruh Metode Pengukuran jarak pada Algoritma KNN untuk klasifikasi kebakaran hutan dan Lahan menunjukkan hasil Metode pengukuran jarak *Manhattan* pada k-NN menghasilkan performansi terbaik dalam mengidentifikasi type kebakaran hutan dan lahan dengan akurasi 92,6 %[7].

Penelitian terdahulu selanjutnya yang membahas mengenai ketidak seimbangan data dengan menggunakan metode SMOTE. Algoritma ini terbukti mampu menangani ketidak seimbangan kelas data dengan menghasilkan nilai *G-mean* dan *F-measure* yang lebih tinggi dibandingkan dengan knn saja. Dimana rata-rata nilai performa *G-Mean* sebesar 81,0% untuk skema SMOTE + KNN dan 53,4% untuk skema KNN. Rata-rata performa *F-Measure* untuk skema SMOTE + KNN adalah sebesar 81,8% dan dengan skema KNN adalah 38,7%[8].

Oleh karena itu, penelitian ini menyajikan analisis tiga aspek, pengklasifikasian Kredit macet dengan algoritma k-NN, menganalisis pengaruh metode pengukuran jarak pada algoritma k-NN serta menganalisis pengaruh teknik smote terhadap *imbalance class* pada dataset kredit macet dan performa model klasifikasi. Adapun data set diperoleh dari kaggle. Sedangkan metode pengukuran jarak yang dianalisis adalah *Euclidean, Canberra, Chebyshev, Manhattan* dan *Minkowski.*

2. METODOLOGI PENELITIAN

2.1. Skema Alur Penelitian

Gambar 1 alur penelitian

Data mining adalah proses mengekstraksi dan mengidentifikasi informasi berguna dengan menggunakan matematika, statistik, kecerdasan buatan, dan pembelajaran mesin.Proses penemuan pola dalam data dikenal sebagai data mining. Menurut fungsinya, data mining dikelompokkan menjadi deskripsi, estimasi, prediksi, klasifikasi, clustering, dan asosiasi. Proses pengolahan data terdiri dari tiga langkah utama. Pertama, data dipilih, dibersihkan, dan diproses sebelum diproses. Proses ini dilakukan dengan mengikuti pedoman dan pengetahuan dari ahli domain, yang menangkap dan mengintegrasikan data internal dan eksternal ke dalam tinjauan organisasi yang lengkap.Pada langkah ini, algoritma data mining digunakan untuk menggali data yang terintegrasi, yang memudahkan identifikasi informasi berharga. Namun semakin besar data yang diolah maka semakin besar pula waktu prosesnya[9][10].

2.2. Pengumpulan Data

Metode pengumpulan data adalah Teknik atau cara yang dilakukan oleh peneliti untuk mengumpulkan data. Pengumpulan data dilakukan untuk memperoleh informasi yang dibutuhkan dalam rangka mencapai tujuan penelitian. Data yang digunakan pada penelitian ini merupakan data starlog (german credit data) dengan jumlah 1.000 data yang didapatkan dari UCI Machine Learning Repository, dimana 700 merupakan data positif (label 0) dan 300 merupakan data negative (label 1).

2.3. Pra-Process Data

Preprocessing adalah tahap penting dalam analisis data yang dilakukan untuk mengubah data mentah dari proses pengambilan data menjadi format yang lebih efisien untuk digunakan dalam proses klasifikasi[11]. 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*.

2.4. 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].

2.5. Pembagian Data

Pada tahap ini data akan dibagi menjadi dua kelompok yaitu kelompok data latih dan kelompok data uji. Data latih adalah data berlabel yang digunakan untuk membangun model klasifikasi. Data uji adalah data tidak berlabel yang digunakan untuk menguji model yang akan dibuat. *Sharing data* bertujuan untuk memastikan bahwa model yang dibangun dapat diterapkan pada data baru[13]. 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%.

2.6. 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[14]. 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[15]. Dengan kata lain, tujuan dari algoritma ini adalah untuk mengklasifikasikan objek baru berdasarkan atribut dan sampel data pelatihan.

Terdapat 5 metode pengukuran jarak yang akan di uji pada penelitian ini, yaitu Euclidean, Canberra, Chebyshev, Manhattan dan Minkowski.

a. Eulidean

Euclidean Distance merupakan metode pengukuran jarak antara dua objek terpopuler. Metode ini cocok digunakan pada objek yang memiliki nilai atribut Euclidean[15]. Selain itu, teknik ini juga merupakan fungsi dari jarak geometris dari algoritma dasar k-NN. Jarak geometris dapat dihitung dengan menggunakan persamaan (1). Dij adalah jarak antara objek i dan j, dan variabel x menunjukkan objek.

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

b. Canberra

Godfrey N. Lance dan William T. Williams pertama kali menggunakan metode pengukuran jarak Canberra pada tahun 1966. Metode ini biasanya digunakan untuk mengukur jarak antara dua titik dalam ruang vektor[16]. 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[17]. Jarak dua buah objek berdasarkan Chebyshev dapat dihitung menggunakan persamaan (3).

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

d. Manhattan

Jarak Manhattan, juga dikenal sebagai "*city block*," adalah pengukuran geometris antara dua objek. Dalam banyak kasus, hasil perhitungan menggunakan Manhattan lebih baik dibandingkan dengan Euclidean. Jarak Manhattan dapat dihitung menggunakan persamaan (4).

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

Jika hasil dari rumus tersebut besar, tingkat kemiripan antara kedua objek akan semakin rendah. Sebaliknya, jika hasilnya kecil, tingkat kemiripan antara kedua objek akan semakin tinggi. Objek yang dimaksud adalah data latih dan data uji yang kemiripannya akan dihitung[18].

e. Minkowski

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

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

2.7. Cross Validation

Salah satu metode validasi model adalah *cross-validation*, juga dikenal sebagai estimasi rotasi. Teknik ini utamanya digunakan untuk melakukan prediksi model dan memperkirakan seberapa akurat sebuah model prediktif ketika diterapkan dalam praktiknya. K-fold cross-validation adalah teknik validasi silang yang memecah data menjadi K bagian dari set data dengan ukuran yang sama[19]. Validasi silang dengan k-fold=10 menghilangkan bias pada data. Pelatihan dan tes dilakukan sebanyak K kali.

2.8. Evaluasi

Evaluasi klasifikasi didasarkan pada pengujian terhadap objek yang benar dan objek yang salah. Validasi ini digunakan untuk menentukan jenis model terbaik berdasarkan hasil klasifikasi[8]. Evaluasi dalam penelitian ini menggunakan confusion matrix. Confusion matrix 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[20].

TABEL I. Confusion matrix

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

Dari Hasil Confusion Matrix akan ditentukan nilai Akurasi, Presisi, Recall, F-measure, dan G-mean 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. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Data yang digunakan pada penelitian ini merupakan data starlog (german credit data) dengan jumlah 1.000 data yang didapatkan dari UCI Machine Learning Repository.

3.2 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

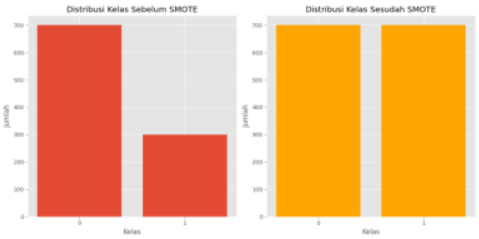
Gambar 3 transformation data

```
[[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

3.3 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.

3.4 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.

3.5 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.

3.6 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 cukup konsisten dan optimal dalam hal Akurasi, Presisi, Recall, dan F-measure. 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.

3.7 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 cukup optimal dalam hal Akurasi, Presisi, Recall, dan F-measure. 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.

3.8 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 cukup optimal dalam hal Akurasi, Presisi, Recall, dan F-measure. 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.

3.9 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 dalam hal Akurasi, Presisi, Recall, dan F-measure. 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.

3.10 Perbandingan Pengukuran Jarak

Berdasarkan Pengujian yang telah dilakukan sebelumnya, dilakukan perbandingan performansi setiap metode pengukuran jarak terbaik. Tabel 8 menyajikan perbandingan performansi 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.

4. 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.

1213 PENGARUH METODE PENGUKURAN JARAK DAN SMOTE PADA KLASIFIKASI PENILAIAN KREDIT

ORIGINALITY REPORT

28%
SIMILARITY INDEX

PRIMARY SOURCES

1	ejurnal.stmik-budidarma.ac.id Internet	262 words — 6%
2	jurnal.fikom.umi.ac.id Internet	124 words — 3%
3	123dok.com Internet	108 words — 3%
4	ejournal.medan.uph.edu Internet	94 words — 2%
5	journal.univpancasila.ac.id Internet	84 words — 2%
6	ojs.unsulbar.ac.id Internet	78 words — 2%
7	sistemasi.ftik.unisi.ac.id Internet	49 words — 1%
8	Hendriyo Mokodompit, Nurnaningsih Nico Abdul, Elvie Fatmah Mokodongan. "PONDOK PESANTREN MODERN DARUL MADINAH WONOSARI KABUPATEN BOALEMO DENGAN PENDEKATAN ARSITEKTUR TROPIS", JAMBURA Journal of Architecture, 2024 Crossref	48 words — 1%

9	e-journal.stmik-bnj.ac.id Internet	31 words — 1%
10	docplayer.info Internet	30 words — 1%
11	e-journal.stmiklombok.ac.id Internet	27 words — 1%
12	repository.uin-suska.ac.id Internet	23 words — 1%
13	www.researchgate.net Internet	20 words — < 1%
14	usir.salford.ac.uk Internet	19 words — < 1%
15	jurnal.yoctobrain.org Internet	18 words — < 1%
16	www.msn.com Internet	17 words — < 1%
17	adoc.pub Internet	16 words — < 1%
18	media.neliti.com Internet	16 words — < 1%
19	repository.unhas.ac.id Internet	15 words — < 1%
20	nanopdf.com Internet	11 words — < 1%

21	core.ac.uk Internet	10 words — < 1%
22	Lecture Notes in Computer Science, 2015. Crossref	9 words — < 1%
23	Waris, Muhammad, Khurshid Ahmad, Muhammad Kabir, and Maqsood Hayat. "Identification of DNA binding proteins using evolutionary profiles position specific scoring matrix", Neurocomputing, 2016. Crossref	8 words — < 1%
24	ejournal.akprind.ac.id Internet	8 words — < 1%
25	jurnal.unimed.ac.id Internet	8 words — < 1%
26	repository.universitasbumigora.ac.id Internet	8 words — < 1%
27	Ratih Purwati, Gunawan Ariyanto. "Pengenalan Wajah Manusia berbasis Algoritma Local Binary Pattern", Emitor: Jurnal Teknik Elektro, 2017 Crossref	6 words — < 1%

EXCLUDE QUOTES	OFF	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	OFF	EXCLUDE MATCHES	OFF