Journal of Information System Research (JOSH)
Volume 3, No. 2, January 2022, pp 118−124
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v3i2.1262

Analisa Distance Metric Algoritma K-Nearest Neighbor Pada Klasifikasi Kredit Macet

Khairul Fadli Margolang[1,*], Muhammad Mizan Siregar[1], Sugeng Riyadi[1], Zakarias Situmorang[2]

[1] Ilmu Komputer, Universitas Potensi Utama, Medan
JL. KL. Yos Sudarso Km. 6,5 No. 3-A, Tanjung Mulia, Medan, Sumatera Utara, Indonesia
[2] Ilmu Komputer, Universitas Katolik Santo Thomas, Medan
Jl. Setia Budi, Kp. Tengah, Kec. Medan Tuntungan, Kota Medan, Sumatera Utara, Indonesia
Email: [1,*]khairulfadhlim@gmail.com, [2]mizan.siregar1@gmail.com, [3]adhie.ogenk@gmail.com, [4]zakarias65@yahoo.com
Email Penulis Korespondensi: khairulfadhlim@gmail.com
Submitted: **20/01/2022**; Accepted: **31/01/2022**; Published: **31/01/2022**

Abstrak−*Data mining* merupakan salah satu metode yang dapat mengklasifikasikan data ke dalam kelas-kelas yang berbeda berdasarkan fitur-fitur di dalam data tersebut. Menggunakan *data mining*, dapat di klasifikasikan kategori kredit macet berdasarkan data pemberian kredit dari koperasi kepada anggotanya. *K-Nearest Neighbor* sebagai salah satu algoritma *data mining* digunakan dalam penelitian ini untuk mengklasifikasikan kategori kredit macet dengan menggunakan variasi *distance metric* seperti *chebyshev*, *euclidean*, *mahalanobis* dan *manhattan*. Hasil evaluasi menggunakan *10-fold cross validation* menunjukkan bahwa *euclidean distance* memilki nilai akurasi, presisi, F1 dan sensitivitas tertinggi dibandingkan *distance metric* lainnya. *Chebyshev distance* memiliki akurasi, presisi dan sensitivitas terendah dan *mahalanobis distance* memiliki nilai F1 terendah. *euclidean* dan *manhattan distance* memiliki nilai reabilitas tertinggi untuk klasifikasi kelas *true positive* dan *true negative*. *Mahalanobis distance* memiliki nilai reabilitas terendah untuk klasifikasi kelas *false positive* sedangkan *chebyshev distance* memiliki reabilitas terendah untuk klasifikasi kelas *false negative*.

Kata Kunci: Data Mining; Distance Metric; Klasifikasi; K-Nearest Neighbor; Kredit Macet

Abstract−Data mining is a method that can classify data into different classes based on the features in the data. With data mining, non-performance loan categories can be classified based on data on lending from cooperatives to their members. This study uses K-Nearest Neighbor to classify non-performance loan categories with various distance metric variations such as Chebyshev, Euclidean, Mahalanobis, and Manhattan. The evaluation results using 10-fold cross-validation show that the Euclidean distance has the highest accuracy, precision, F1, and sensitivity values compared to other distance metrics. Chebyshev distance has the lowest accuracy, precision, sensitivity, while Mahalanobis distance has the lowest F1 value. Euclidean and Manhattan distances have the highest reliability values for true-positive and true-negative class classifications. Mahalanobis distance has the lowest reliability value for false-positive class classification, while Chebyshev distance has the lowest value for false-negative class classification.

Keyword: Data Mining; Classification; Distance Metric; K-Nearest Neighbor; Non-Performing Loan

1. PENDAHULUAN

Pesatnya pertumbuhan koperasi di Indonesia, pemberian kredit bagi anggota koperasi semakin diperketat dengan cara menerapkan prinsip yang diterapkan bank dunia, yaitu melalui *credit analysis*. Pihak koperasi akan melakukan survei terhadap calon penerima kredit melalui analisis 5C, yaitu *character* (karakter), *capacity* (kapasitas), *capital* (kapital), *condition of economic* (kondisi ekomoni) dan *collateral* (agunan). Analisis ini dilakukan agar resiko kredit macet dapat diminimalisir [1].

Data mining dapat membantu pihak koperasi dalam menganalisis kredit macet calon penerima kredit dengan cara membandingkan data lama (data pemberian kredit sebelumnya) dengan data baru (data hasil survei calon penerima kredit) dan mengelompokkannya dalam bentuk klasifikasi kredit macet atau kredit tidak macet [2]. Dengan proses seleksi, eksplorasi dan pemodelan terhadap data sebelumnya, *data mining* mampu menemukan pengetahuan berupa hubungan antara fitur satu dengan fitur lainnya yang sebelumnya tidak diketahui atau sering disebut sebagai *knowledge discovery in database* [3].

Sebagai salah satu penerapan *data mining*, *clustering* membagi data dalam beberapa kluster dengan melihat tingkat kesamaan data dengan kluster yang dibentuk. Dengan prinsip klusterisasi ini diperoleh kelompok-kelompok data yang memiliki kemiripan tertentu sehingga lebih mudah untuk dilakukan identifikasi terhadap data tersebut [4].

K-Nearest Neighbor (K-NN) merupakan salah satu algoritma *data mining* yang banyak digunakan di dalam penelitian, khususnya yang berhubungan dengan klasifikasi objek di dalam data. Dengan prinsip mencari kelompok berjumlah k dari sekumpulan data set, K-NN menghitung jarak terdekat antara data *training* dengan data *test* untuk menghasilkan klasifikasi data [5].

Inti dari algoritma K-NN adalah penghitungan jarak (*distance metric*) karena pemilihan *distance metric* yang tepat akan mempengaruhi performa algoritma ini dalam mengklasifikasikan data. Beberapa teknik pengukuran data yang dapat digunakan dalam algoritma K-NN diantaranya adalah *chebychev distance*, *euclidean distance*, *mahalonobis distance* dan *manhattan distance* [6].

Penelitian ini membandingkan *chebychev distance*, *euclidean distance*, *mahalanobis distance* dan *manhattan* dalam klasifikasi kredit macet menggunakan algoritma K-NN. Keempat *distrance metric* ini dievaluasi

Journal of Information System Research (JOSH)
Volume 3, No. 2, January 2022, pp 118−124
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v3i2.1262

akurasi, presisi, F-1 *score* dan sensitivitasnya dalam mengklasifikasikan data menggunakan teknik *10-Fold Cross Validation* dan reabilitasnya berdasarkan nilai *False Positive Rate* (FPR) dan *False Negative Rate* (FNR) menggunakan teknik *confusion matrix*. Berdasarkan perbandingan akurasi, presisi, sensitivitas, FPR dan FNR ini kemudian dianalisa mana *distance metric* terbaik dan terburuk untuk kasus klasifikasi data set kredit macet yang digunakan.

2. METODOLOGI PENELITIAN

Data yang digunakan dalam penelitian ini adalah data primer yang diperoleh langsung dari koperasi Mutiara Sejahtera berupa data 61 anggota koperasi yang diolah menjadi fitur-fitur yang dibutuhkan di dalam penelitian seperti Karyawan Tetap (apakah anggota koperasi merupakan karyawan tetap), Lama Keanggotaan (berapa lama anggota tersebut menjadi anggota koperasi), Jumlah Pinjaman (besarnya pinjaman yang diajukan anggota), Lama Pinjaman (waktu pinjaman hingga dilunasi anggota), Pinjaman Tempat Lain (apakah anggota memiliki pinjaman di tempat lain) dan Kredit Macet (apakah pinjaman anggota tersebut termasuk kredit macet atau tidak). Dari 61 data ini, diambil 10 data sebagai sampel yang ditampilkan pada Tabel 1 dan Tabel 2.

Tabel 1. Sampel Data Koperasi (1-4)

No	KT	LK	JP	LP	PTL	KM
1	Y	AL	PS	WL	T	T
2	Y	AL	PS	WL	T	T
3	Y	AL	PB	WL	T	T
4	Y	AS	PS	WL	T	T

Tabel 2. Sampel Data Koperasi (5-10)

No	KT	LK	JP	LP	PTL	KM
5	Y	AL	PK	WL	T	T
6	Y	AB	PK	WL	T	T
7	Y	AB	PK	WS	T	T
8	T	AS	PK	WS	T	T
9	Y	AL	PB	WP	Y	Y
10	Y	AL	PK	WL	T	T

Keterangan: KT = Karyawan Tetap, LK = Lama Keanggotaan, JP = Jumlah Pinjaman, LP = Lama Pinjaman, PTL = Pinjaman Tempat Lain, KM = Kredit Macet, Y = Ya, T = Tidak, AL = Anggota Lama, AS = Anggota Sedang, AB = Anggota Baru, PB = Pinjaman Besar, PS = Pinjaman Sedang, PK = Pinjaman Kecil, WL = Waktu Lama, WS = Waktu Sedang, WP = Waktu Pendek

Klasifikasi kredit macet atau tidak macet dari data set di atas dilakukan menggunakan algoritma K-NN dengan langkah sebagai berikut:

a. Menentukan jumlah *neighbor*
   Dalam penelitian ini, digunakan variasi jumlah *neighbor* sebanyak 1 sampai 10 dalam proses klasifikasi.
b. Menghitung jarak data
   Masing-masing data diukur jaraknya dengan kelas Kredit Macet dan Kredit Tidak Macet menggunakan empat variasi *distance metric*, yaitu *chebychev*, *euclidean*, *mahalanobis* dan *manhattan*.
c. Mengklasifikasikan data
   Berdasarkan jarak yang diperoleh, masing-masing data diklasifikasikan apakah termasuk kelas Kredit Macet atau kelas Kredit Tidak Macet.
d. Evaluasi *Cross Validation*
   Hasil klasifikasi diukur nilai akurasi, presisi, F-1 *score* dan sensitivitasnya menggunakan teknik *10-fold Cross Validation*.
e. Evaluasi *Confusion Matrix*
   Hasil prediksi kelas data diukur reabilitasnya menggunakan *confusion matrix* dengan melihat *false positive rate* dan *false negative rate* yang dihasilkan masing-masing variasi *distance metric*.

Journal of Information System Research (JOSH)
Volume 3, No. 2, January 2022, pp 118−124
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v3i2.1262

Penghitungan jarak variasi *distance metric* dilakukan dengan menggunakan persamaan (1) sampai (4) sebagai berikut:

a. *Chebychev Distance* [7]

$$CD\ (x,y) = \log_{\lambda-\infty} \sqrt{\sum_{j=1}^{N} |x-y|^{\lambda}} \quad (1)$$

b. *Euclidean Distance* [7].

$$ED\ (x,y) = \sqrt{\sum_{j=1}^{N} |x-y|^{2}} \quad (2)$$

c. *Mahalanobis Distance* [8]

$$MD\ (x,y) = \sqrt{\sum_{j=1}^{N} |x-y|^{T}\ |x-y|} \quad (3)$$

d. *Manhattan Distance* [7]

$$ManD\ (x,y) = \sum_{j=1}^{N} |x-y| \quad (4)$$

Evaluasi *10-fold cross validation* untuk menghitung nilai *accuracy*, *precision*, *F-1 score* dan *recall* masing-masing variasi *distance metric* dilakukan menggunakan persamaan (5) sampai (8) sebagai berikut [9]:

1. Nilai *Accuracy*

$$NA = \frac{TP+TN}{TP+FN+FP+TN} \quad (5)$$

2. Nilai *Precision*

$$NP = \frac{TP}{TP+FP} \quad (6)$$

3. Nilai *Recall*

$$NR = \frac{TP}{TP+FN} \quad (7)$$

4. Nilai *F-1 Score*

$$NF1 = 2\ \frac{NR\ .\ NP}{NR+NP} \quad (8)$$

Evaluasi *confusion* untuk menghitung nilai *false positive rate* (FPR) dan *false negative rate* (FNR) masing-masing variasi *distance metric* dilakukan menggunakan persamaan (9) dan (10) sebagai berikut [10]:

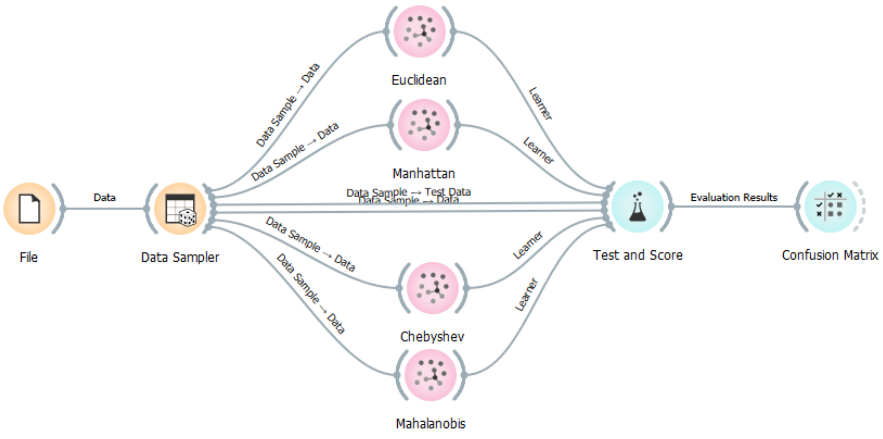
1. *False Positive Rate*

$$FPR = \frac{FP}{FP+TN} \quad (9)$$

2. *False Negative Rate*

$$FNR = \frac{FN}{FN+TP} \quad (10)$$

Widget aplikasi *Orange 3.30* seperti File, Data Sampler, Learner k-NN Euclidean, Manhattan, Chebyshev serta Mahalanobis, Test and Score dan Confusion Matrix digunakan untuk membentuk model klasifikasi, seperti terlihat pada Gambar 1.

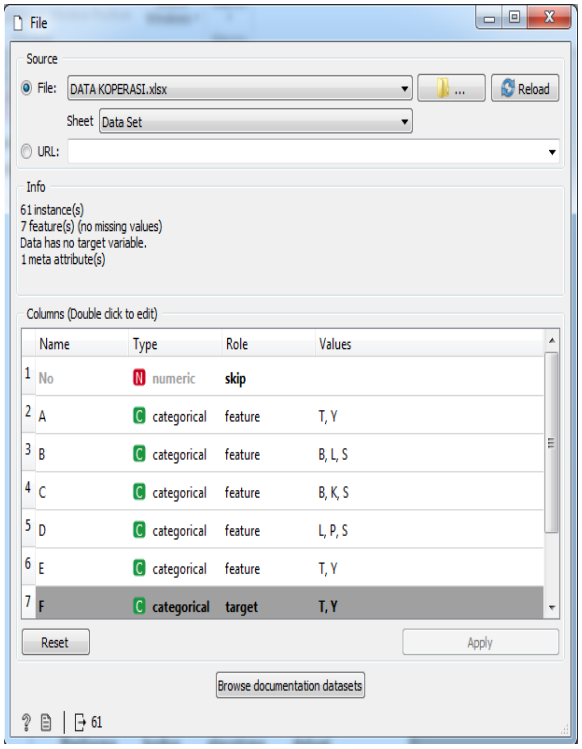


Gambar 1. Model Klasifikasi Menggunakan *Orange 3.30*

Journal of Information System Research (JOSH)
Volume 3, No. 2, January 2022, pp 118−124
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v3i2.1262

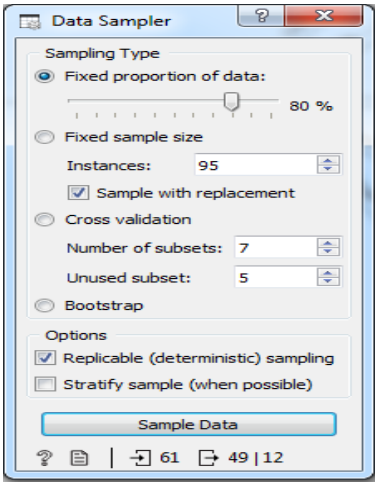
3. HASIL DAN PEMBAHASAN

Widget File digunakan untuk membuka file data set dan memilih fitur F sebagai target sehingga diperoleh informasi fitur dan target di dalam data set seperti terlihat pada Gambar 2.



Gambar 2. Informasi Fitur dan Target

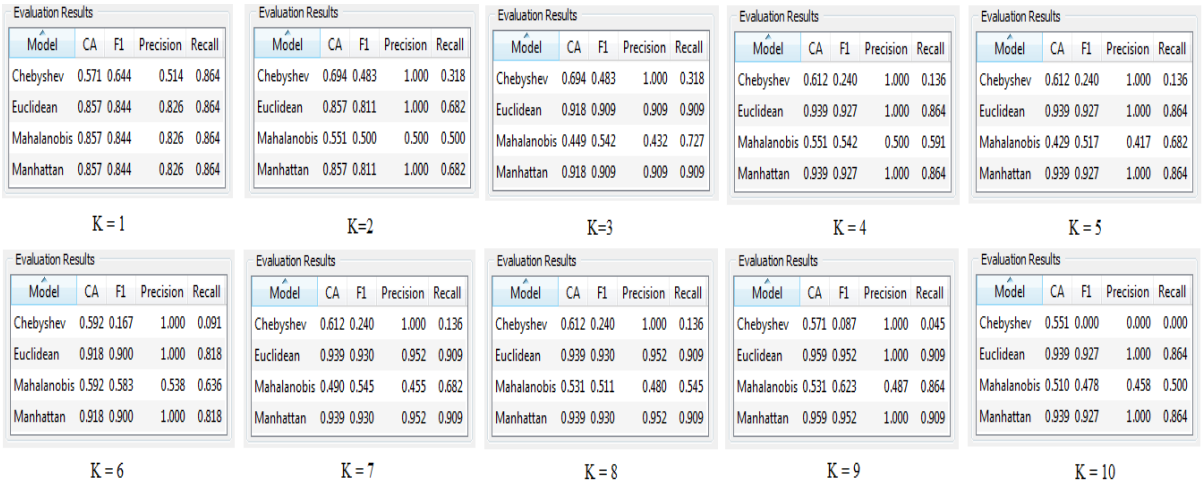
Widget Data Sampler digunakan untuk memilih data *training* dan data *test* secara acak dari data set dengan komposisi perbandingan 8:2, seperti terlihat pada Gambar 3.



Gambar 3. Data *Training* dan Data *Test*

Klasifikasi data dilakukan dengan 10 tahapan menggunakan variasi jumlah *neighbor* sebanyak 1 sampai 10 pada masing-masing *widget* Learner. Hasil evaluasi klasifikasi berupa nilai *accuracy*, *precision* dan *recall* diperoleh dari *widget* Test and Score seperti terlihat pada Gambar 4.

Journal of Information System Research (JOSH)
Volume 3, No. 2, January 2022, pp 118−124
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v3i2.1262



Gambar 4. *10 Fold Cross Validation* Jumlah K = 1 Sampai K = 10

Dari Gambar 4, dihitung rata-rata nilai *accuracy*, *precision*, *F1* dan *recall* masing-masing variasi *distance metrix*, sehingga menhasilkan nilai-nilai seperti ditunjukkan pada Tabel 3.

Tabel 3. Rata-Rata Nilai *Cross Validation*

DM	RA	RP	RF	RR
C	0.6121	0.2824	0.8514	0.2098
E	0.9204	0.9057	0.9639	0.8592
Ma	0.5491	0.5685	0.5093	0.6591
Mb	0.8796	0.8638	0.9167	0.8228

Keterangan: DM = *Distance metrix*, RA = Rata-rata *accuracy*, RP = Rata-tara *precision*, RF = Rata-rata *F1-score*, RR = Rata-rata *recall*, C = *Chebyshev*, E = *Euclidean*, Ma = *Mahalanobis*, Mb = *Manhattan*

Dari Tabel 3, terlihat *distance metric* dengan *accuracy* tertinggi adalah *euclidean* dengan nilai 0.9204 sedangkan yang terendah adalah *chebyshev* dengan nilai 0.6121.

Distance metric dengan *precision* tertinggi adalah *euclidean* dengan nilai 0.9057 sedangkan yang terendah adalah *chebyshev* dengan nilai 0.2824.

Distance metric dengan *F1 score* tertinggi adalah *euclidean* dengan nilai 0.9639 sedangkan yang terendah adalah *mahalanobis* dengan nilai 0.5093.

Distance metric dengan *recall* tertinggi adalah *euclidean* dengan nilai 0.8592 sedangkan yang terendah adalah *chebyshev* dengan nilai 0.2098.

Evaluasi reabilitas masing-masing *distance metric* dalam mengklasifikasikan data set diperoleh dari *widget* Confusion Metric berupa nilai *true negative*, *false positive*, *false negative* dan *true positive* seperti terlihat pada Tabel 4 dan Tabel 5.

Tabel 4. Hasil *Confusion Matrix* (*Chebyshev*, *Euclidean* dan *Mahalanobis*)

DM	K	TN	FP	FN	TP
C	1	9	18	3	19
	2	27	0	15	7
	3	27	0	15	7
	4	27	0	19	3
	5	27	0	19	3
	6	27	0	20	2
	7	27	0	19	3
	8	27	0	19	3
	9	27	0	21	1
	10	27	0	22	0
E	1	23	4	3	19

Journal of Information System Research (JOSH)
Volume 3, No. 2, January 2022, pp 118−124
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v3i2.1262

	2	27	0	7	15
	3	25	2	2	20
	4	27	0	3	19
	5	27	0	3	19
	6	27	0	4	18
	7	26	1	2	20
	8	26	1	2	20
	9	27	0	2	20
	10	27	0	3	19
Ma	1	13	14	6	16
	2	16	11	11	11
	3	6	21	6	16
	4	14	13	9	13
	5	6	21	7	15
	6	15	12	8	14
	7	9	18	7	15
	8	14	13	10	12
	9	7	20	3	19
	10	14	13	11	11

Tabel 5. Hasil *Confusion Matrix* (*Manhattan*)

K	TN	FP	FN	TP
1	23	4	3	19
2	27	0	7	15
3	25	2	2	20
4	27	0	3	19
5	27	0	3	19
6	27	0	4	18
7	26	1	2	20
8	26	1	2	20
9	27	0	2	20
10	27	0	3	19

Keterangan: K = Jumlah *neighbor*, TN = *True negative*, FP = *False positive*, FN = *False negative*, TP = *True positive*

Dari Tabel 4 dan Tabel 5, dihitung rata-rata nilai *true negative, false positive, false negative* dan *true positive* masing-masing variasi *distance metrix*, sehingga menghasilkan nilai-nilai seperti ditunjukkan pada Tabel 6.

Tabel 6. Rata-Rata Nilai *Confusion Matrix*

DM	TN	FP	FN	TP
C	25.2	1.8	17.2	4.8
E	26.2	0.8	3.1	18.9
Ma	11.4	15.6	7.8	14.2
Mb	26.2	0.8	3.1	18.9

Dari Tabel 6, dihitung nilai *false positive rate* (FPR) dan *false negative rate* (FNR) menggunakan persamaan (9) dan (10) sebagai berikut:

$$FPR_C = \frac{1.8}{1.8 + 25.2} = 0.0667$$

$$FPR_E = \frac{0.8}{0.8 + 26.2} = 0.0296$$

$$FPR_{Ma} = \frac{15.6}{15.6 + 11.4} = 0.5777$$

Journal of Information System Research (JOSH)
Volume 3, No. 2, January 2022, pp 118−124
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v3i2.1262

$$FPR_{Mb} = \frac{0.8}{0.8 + 26.2} = 0.0296$$
$$FNR_{C} = \frac{17.2}{17.2 + 4.8} = 0.7818$$
$$FNR_{E} = \frac{3.1}{3.1 + 18.9} = 0.1409$$
$$FNR_{Ma} = \frac{7.8}{7.8 + 14.2} = 0.3545$$
$$FNR_{Mb} = \frac{3.1}{3.1 + 18.9} = 0.1409$$

Dari hasil perhitungan *false positive rate* dan *false negative rate* di atas, terlihat bahwa *distance metric euclidean* dan *manhattan* memiliki nilai *false positive rate* terendah dengan nilai 0,0296 dan *distance metric mahalanobis* memiliki nilai tertinggi dengan nilai 0,5777.

Distance metric euclidean dan *manhattan* memiliki nilai *false negative rate* terendah dengan nilai 0,1409 sedangkan *distance metric chebyshev* memiliki nilai tertinggi dengan nilai 0,7818.

4. KESIMPULAN

Dari hasil klasifikasi 61 data anggota koperasi Mutiara Sejahtera menggunakan algoritma *logistic regression* dengan variasi *distamce metric* seperti *chebyshev, euclidean, manhattan* dan *mahalanobis,* diperoleh bahwa algoritma dengan akurasi, presisi, *F1 score* dan sensitifitas tertinggi adalah *distance metric euclidean*. *Distance metric chebyshev* memiliki nilai akurasi, presisi dan sensitivitas terendah sedangkan *distance metric mahalanobis* memiliki nilai *F1 score* terendah. Dari segi reabilitas, *distance metric euclidean* dan *manhattan* memiliki reabilitas terbaik dalam mengklasifikasikan data pada kelas Kredit Macet dan Kredit Tidak Macet dilihat dari nilai *false positive rate* dan *false negative rate* yang dihasilkan. *Distance metric mahalanobis* memiliki reabilitas terburuk dalam mengklasifikasikan data pada kelas Kredit Macet dilihat dari nilai *false positive rate* yang dihasilkan. *Distance metric chebyshev* memiliki reabilitas terburuk dalam mengklasifikasikan data pada kelas Kredit Tidak Macet, dilihat dari nilai *false negative rate* yang dihasilkan.

REFERENCES

[1] Christnatalis, Saragih, R. R., Tambunan, B. C. (2021) Data Mining Algorithm C4.5 Classification Determination Credit Eligibility For Jaya Bersama Cooperatives (KORJABE). *Jurnal Teknologi dan Sistem Informasi*. 8(1): 59-68.
[2] Rizky, S. A., Yesputra, R., Santoso. (2021) Prediksi Kelancaran Pembayaran Cicilan Calon Debitur Dengan Metode K-Nearest Neighbor. *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*. 7(2): 195-202.
[3] Nofitri, R., Irawati, N. (2019) Integrasi Metode Neive Bayes Dan Software Rapidminer Dalam Analisis Hasil Usaha Perusahaan Dagang. *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*. 6(1): 35-42.
[4] Iqbal, M. (2019) Klasterisasi Data Jamaah Umroh Pada Auliya Tour & Travel Menggunakan Metode K-Means Clustering. *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*. 5(2): 97-104.
[5] Arnomo, R. A., Saptomo, W. L. Y., Harsadi, P. (2018) Implementasi Algoritma K-Nearest Neighbor Untuk Identifikasi Kualitas Air (Studi Kasus: PDAM Kota Surakarta). *Jurnal TIKomSiN*. 6(1): 1-5.
[6] Alfeilat, H. A. A., Hassanat, A. B. A., Lasassmeh, O., Tarawneh, A. S, Alhasanat, M. B., Salman, H. S. E., Prasath, V. B. S. (2019) Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review. *Big Data*. 7(4): 1-28.
[7] Iswanto, Tulus, Sihombing, P. (2021) Comparison of Distance Models on K-Nearest Neighbor Algorithm in Stroke Disease Detection. *Applied Technology and Computing Science Journal*. 4(1): 63-68.
[8] Fan, H., Chen, Y., Huang, S., Zhang, X., Guan, H. (2018) Post-fault Transient Stability Assessment Based on k-Nearest Neighbor Algorithm with Mahalanobis Distance. *International Conference on Power System Technology (POWERCON)*. 4417-4423.
[9] Farokhah, L. (2020) Implementasi K-Nearest Neighbor Untuk Klasifikasi Bunga Dengan Ekstraksi Fitur Warna RGB. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*. 7(6): 1129-1135.
[10] Markoulidakis, I., Rallis, I., Georgoulas, I., Kopsiaftis, G., Doulamis, A., Doulamis, N. (2021). Multiclass Confusion Matrix Reduction Method and Its Application on Net Promoter Score Classification Problem. *Technologies*. 9(81): 58-73.