Statistika, Vol. 22 No. 1, 95 – 104
Mei 2022

Penerapan Metode *Modified K-Nearest Neighbor* Pada Pengklasifikasian Status Pembayaran Kredit Barang Elektronik dan *Furniture*

SELSA AMELIA[1], MEMI NOR HAYATI[2], SURYA PRANGGA[3]

[1,2,3]Program Studi Statistika, Jurusan Matematika, Fakultas MIPA Universitas Mulawarman, Indonesia
e-mail: selsapanjaitan@gmail.com

ABSTRAK

Klasifikasi merupakan serangkaian proses pembentukan model dari suatu objek ke dalam kelompok untuk memprediksi kelas dari suatu objek yang belum diketahui sebelumnya. *Modified K-Nearest Neighbor* (MK-NN) merupakan salah satu metode klasifikasi pengembangan dari algoritma *K-Nearest Neighbor* (K-NN) yang menambahkan proses validitas serta *weight voting* (pembobotan) untuk mengatasi tingkat akurasi rendah dari algoritma K-NN. Penelitian ini bertujuan untuk mengetahui hasil pengklasifikasian status pembayaran kredit barang elektronik dan *furniture* serta tingkat akurasi klasifikasi pada metode MK-NN. Data yang digunakan adalah data debitur PT. KB Finansia Multi Finance Tahun 2020 dengan status pembayaran kredit lancar dan tidak lancar serta menggunakan 7 variabel bebas yaitu usia, jumlah tanggungan, lama tinggal, pendapatan, masa kerja, besar pembayaran kredit, dan lama peminjaman kredit. Berdasarkan penelitian yang telah dilakukan, diperoleh nilai akurasi sebesar 84,61% dengan K optimal yaitu K = 5 pada proporsi 90% : 10%.

Kata Kunci: Klasifikasi, *K-Nearest Neighbor*, *Modified K-Nearest Neighbor*, Kredit.

ABSTRACT

Classification is a series of process of forming a model of an object into groups to predict the class of an object that has not been known before. Modified K-Nearest Neighbor (MK-NN) is one of the classification methods developed from the K-Nearest Neighbor (K-NN) algorithm which adds a process of validity and weight voting to overcome the low level of accuracy of the K-NN algorithm. This study aims to determine the results of classifying credit payment status for electronic goods and furniture as well as the accuracy of the classification using the MK-NN method. The data used is debtor data for the 2020 KB Finansia Multi Finance Company with current and non-current credit payment status and uses 7 independent variables, namely age, number of dependents, length of stay, income, years of service, amount of credit payments, and length of loan. Based on the research that has been done, an accuracy value of 84.61% is obtained with optimal K, namely K = 5 at a proportion of 90%: 10%.

Keywords: Classification, K-Nearest Neighbor, Modified K-Nearest Neighbor, Credit.

1. PENDAHULUAN

Klasifikasi adalah salah satu teknik dari *data mining* yang melihat atribut dari kelompok data yang sudah didefinisikan sebelumnya. Atribut-atribut tersebut dapat digunakan sebagai variabel dalam penentuan kelas suatu objek baru. Tujuan dari klasifikasi yaitu untuk menentukan kelas dari suatu objek yang belum diketahui kelasnya dengan akurat (Imanda, Hidayat, & Furqon, 2018). Algoritma K-NN merupakan salah satu metode yang paling banyak digunakan untuk mencari solusi dari permasalahan klasifikasi data terhadap objek berdasarkan nilai K atau tetangga terdekatnya. Metode K-NN memerlukan ukuran jarak untuk menentukan kedekatan suatu objek yang terdapat pada data *testing* berdasarkan dengan tetangga terdekat yang dimiliki (Saxena, Khan, dan Singh, 2014). Akan tetapi, metode K-NN memiliki beberapa kelemahan dalam penentuan nilai K yang menyebabkan rendahnya kinerja akurasi yang dihasilkan oleh metode K-NN. Oleh karena itu, beberapa penelitian telah dilakukan dalam hal untuk perbaikan

metode K-NN, baik dari segi akurasi maupun dari segi optimasi nilai K, sehingga terbentuklah algoritma *Modified K-Nearest Neighbor* (MK-NN).

Modified K-Nearest Neighbor (MK-NN) merupakan pengembangan dari metode K-NN yang bertujuan untuk mengatasi masalah tingkat akurasi yang rendah pada algoritma K-NN. Algoritma MK-NN menambahkan dua prosedur dalam melakukan klasifikasi yaitu proses validasi dari data *training* serta proses pembobotan terbesar (*weight voting*) dari tetangga terdekatnya (*nearest neighbor*). Validasi dan bobot terbesar yang dimiliki oleh MK-NN mampu mengatasi kelemahan dari klasifikasi berdasarkan jarak terdekat sehingga metode MK-NN jauh lebih unggul dibandingkan metode algoritma K-NN (Parvin, Alizadeh, & Bidgoli, 2008).

Seiring berjalannya waktu, meningkatnya kebutuhan manusia akan barang dan jasa menimbulkan pembiayaan akan kebutuhan juga semakin meningkat. Kebutuhan manusia yang semakin meningkat mengakibatkan semakin banyak pula lembaga keuangan dari sektor perbankan terbentuk khususnya perusahaan pembiayaan konsumen. Penyaluran kredit merupakan salah satu permasalahan atau kendala yang sering muncul pada perusahaan pembiayaan konsumen karena terdapat banyak persyaratan yang harus dipertimbangkan sebelum pendanaan tersebut disalurkan kepada debitur. Permasalahan dalam penyaluran kredit menyebabkan kredit tidak lancar dari debitur sehingga menimbulkan kerugian dan resiko yang berdampak pada lembaga pembiayaan bukan bank itu sendiri. Banyaknya resiko kredit yang tidak lancar memungkinkan suatu perusahaan pembiayaan konsumen harus lebih bijak dalam memilih debitur yang akan dibantu pembiayaannya sehingga tidak membuat perusahaan mengalami kerugian yang besar. Oleh karena itu, tahap analisis menggunakan *Modified K-Nearest Neighbor* dapat memprediksi karakteristik calon debitur yang mengajukan permohonan kredit.

2. METODE PENELITIAN

Data pada penelitian ini menggunakan data sekunder dari data debitur dari PT. KB Finansia Multi Finance Bontang Tahun 2020 yang terdiri dari 133 data debitur. Variabel penelitian yang digunakan dalam penelitian ini terdiri dari variabel terikat (Y) yaitu status pembayaran kredit yang memiliki dua kategori status pembayaran yakni kredit lancar dan kredit tidak lancar serta terdapat 7 variabel bebasnya (X) yakni usia debitur, jumlah tanggungan, lama tinggal, masa kerja, pendapatan, besar pembayaran, dan lama peminjaman.

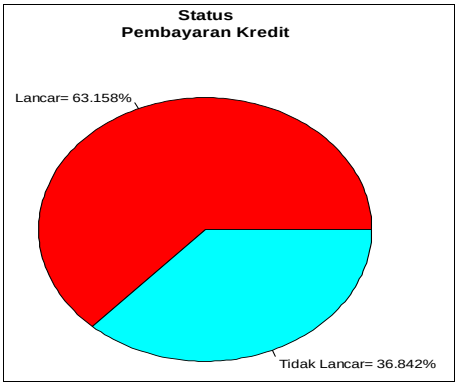
Tahapan analisis yang digunakan pada penelitian ini yakni sebagai berikut:

1. Melakukan analisis statistika deskriptif.
2. Melakukan standarisasi data pada semua variabel.
3. Melakukan pengacakan data.
4. Melakukan pembagian data *training* dan data *testing* berdasarkan proporsi.
5. Penentuan nilai *k-fold cross validation.*
6. Penentuan jumlah data dalam *subset*.
7. Menghitung jarak Euclid antar data *training* dan data *testing*.
8. Melakukan klasifikasi berdasarkan tertangga terdekat.
9. Memilih nilai K optimal berdasarkan proporsi.
10. Menghitung jarak Euclid antar data *training*.
11. Menghitung validitas pada data *training*.
12. Menghitung *weight voting*.
13. Menghitung akurasi ketepatan prediksi.

3. HASIL DAN PEMBAHASAN

Gambaran Umum mengenai Status Pembayaran Kredit

Gambaran data debitur berdasarkan status pembayaran kredit menggunakan analisis statistika deskriptif. Tahapan statistika deskriptif dalam penelitian ini menggunakan *software R* yang dapat dijelaskan pada Gambar 1 sebagai berikut.



Gambar 1 Persentase Status Pembayaran Kredit

Gambar 1 dapat diketahui bahwa terdapat 84 debitur yang memiliki status pembayaran kredit lancar dengan persentase 63,16% dan sisanya terdapat 49 debitur yang memiliki status pembayaran kredit tidak lancar dengan persentase 36,84%. Dari data tersebut dapat dilihat bahwa secara keseluruhan data debitur dengan status pembayaran kredit lancar lebih banyak dibandingkan debitur dengan status pembayaran kredit tidak lancar.

Standarisasi Data

Standarisasi data dilakukan agar semua variabel berada dalam jangkauan yang sama sehingga proporsi pengaruh pada fungsi klasifikator dapat seimbang. Proses standarisasi data hanya dilakukan pada seluruh variabel bebas. Hasil standarisasi data dapat dilihat pada Tabel 1 sebagai berikut.

Tabel 1 Data Hasil Standarisasi

Sampel	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
DB-1	L	1,41	-0,04	-0,89	-0,61	0,39	-0,29	0,37
DB-2	TL	0,10	-0,63	-0,26	-0,46	-0,54	-0,25	-0,41
DB-3	L	-0,12	1,16	0,20	-0,90	-0,84	-0,27	-1,18
DB-4	L	-1,54	-1,23	0,65	-0,90	-0,84	-0,43	0,37
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
DB-133	L	1,41	1,16	0,38	2,03	-0,23	-0,28	0,11

Pengacakan Data

Pengacakan data pada bertujuan agar semua data memiliki kesempatan yang sama untuk menjadi data *training* dan data *testing*. Selain itu, pengacakan data juga dilakukan agar hasil klasifikasi terpercaya serta ketika dilakukan pengacakan lagi hasil klasifikasi yang didapatkan tetap. Adapun hasil pengacakan data dapat dilihat pada Tabel 2 sebagai berikut.

Tabel 2 Hasil Pengacakan Data

Sampel	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
DB-96	TL	1,74	-0,04	-0,26	3,94	0,39	0,25	2,16
DB-25	L	1,74	-0,04	1,57	-0,46	0,39	-0,43	-1,95
DB-133	L	1,41	1,16	0,38	2,03	-0,23	-0,28	0,11
DB-98	L	-0,45	-1,23	-0,80	-0,46	0,39	-0,26	-1,18
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
DB-7	L	1,41	1,16	0,65	1,74	0,39	0,25	-0,41

Pembagian Data *Training* dan Data *Testing*

Penelitian ini menggunakan 3 proporsi dalam pembagian data *training* dan data *testing* yaitu 50% : 50%, 70% : 30%, dan 90% : 10%. Data yang berada pada urutan pertama akan digunakan menjadi data *training* dan sisanya akan menjadi data *testing*. Peneliti menggunakan proporsi 90% : 10% sebagai contoh dalam menguraikan perhitungan analisis. Penggunaan proporsi 90% : 10% menandakan bahwa 90% dari hasil pengacakan data pada urutan pertama akan menjadi

data *training* dan 10% sisanya akan menjadi data *testing*. Sedemikian sehingga, diperoleh 120 data *training* dan 13 data *testing*.

Penentuan Nilai *k-fold Cross Validation*

k-fold cross validation merupakan metode yang digunakan untuk mengetahui rata-rata keberhasilan dengan cara melakukan perulangan tiap *subset* dari pengacakan data yang didapat. Penggunaan *k* dilakukan dalam pembagian sebuah himpunan secara acak yang akan menjadi *subset*. Tahapan *k-fold cross validation* diawali dengan membagi data sejumlah *k-fold* pada ukuran yang sama, kemudian proses *training* dan *testing* dilakukan sebanyak *k* kali. Peneliti menggunakan 10-*fold cross validation* sehingga jumlah *fold* yang diperlukan yaitu sebanyak 10 *fold*. Penggunaan 10 *fold* tersebut akan membentuk sebuah *subset* sedemikian sehingga jumlah *subset* yang digunakan sebanyak 10 *subset*. Sebagai contoh, pada *fold* ke-1 terdapat 9 kombinasi *subset* yang berbeda sebagai data *training* dan sisanya terdapat 1 *subset* sebagai data *testing*, selanjutnya proses *training* dan *testing* dilakukan sampai *fold* ke-10.

Penentuan Jumlah Data dalam *Subset*

Setelah menentukan *k* pada *k-fold cross validation* yakni 10-*fold cross validation*, maka dilakukan perhitungan jumlah data yang akan diamati dari masing-masing *subset*. Perhitungan jumlah data dalam *subset* hanya menggunakan banyaknya data *training* dari proporsi yang ditentukan yaitu misalnya pada proporsi 90% : 10% dengan jumlah data *training* sebanyak 120 serta nilai *k-fold cross validation* yang ditentukan yaitu 10-*fold cross validation*. Jumlah data dalam *subset* yang diperoleh yaitu sebanyak 12 data pengamatan dan dapat dilihat pada Tabel 3 sebagai berikut.

Tabel 3 Hasil Penentuan Jumlah Data dalam *Subset*

Subset	Sampel	Y	Jumlah Sampel
Subset 1	DB-96	TL	12
	DB-25	L	
	DB-133	L	
	⋮	⋮	
	DB-5	L	
Subset 2	DB-1	L	12
	DB-15	L	
	DB-104	L	
	⋮	⋮	
	DB-13	TL	
⋮	⋮	⋮	⋮
Subset 10	DB-42	L	12
	DB-60	L	
	DB-52	L	
	⋮	⋮	
	DB-72	TL	

Perhitungan Jarak Euclid antar Data *Training* dan Data *Testing*

Penelitian ini menggunakan *k-fold cross validation* yaitu 10-*fold cross validation* sedemikian sehingga 1 *subset* akan berlaku sebagai data *testing* dan *subset* lainnya sebagai data *training*. Setiap *subset* memiliki giliran untuk dijadikan sebagai data *testing* contohnya, ketika *subset* 1 dijadikan sebagai data *testing* maka *subset* 2, *subset* 3 sampai dengan *subset* 10 akan menjadi data *training*. Peneliti menggunakan data *testing* ke-1 dari *subset* 1 (DB-96) sebagai contoh dalam melakukan proses perhitungan jarak Euclid yaitu sebagai berikut.

$$d_{(1,1)} = \sqrt{(1{,}41-1{,}74)^2+\cdots+(0{,}37-2{,}16)^2}$$

$$= 4{,}97$$

$$d_{(2,1)} = \sqrt{((-0{,}45)-1{,}74)^2+\cdots+(0{,}37-2{,}16)^2}$$

$= 5{,}58$

⋮

$$d_{(108,1)} = \sqrt{(-0{,}01-1{,}74)^2+\cdots+(-2{,}20-2{,}16)^2}$$

$= 6{,}36$

Perhitungan jarak Euclid dilakukan hingga dilakukan hingga *subset* 10 sebagai data *testing* dan *subset* 1, *subset* 2, sampai dengan *subset* 9 sebagai data *training* yakni data *testing* ke-12 (DB-72) berasal dari *subset* 10 dengan data *training* ke-108 (DB-68) berasal dari *subset* 9

Klasifikasi Berdasarkan Nilai K (Tetangga Terdekat)

Jarak Euclid untuk setiap data *testing* diurutkan dari jarak yang paling dekat sampai dengan jarak yang paling jauh. Kemudian, tiap data *testing* akan diklasifikasikan berdasarkan nilai K yakni K=1, K=3, K=5, K=7, dan K=9. Batas K-NN menyatakan batas K yang menggambarkan jumlah tertangga terdekat yang menjadi acuan untuk menentukan kategori pada variabel terikat. Sebagai contoh, jika K awal yang digunakan yakni 1 atau 1-NN maka hanya maka hanya menggunakan 1 tetangga terdekat atau *rank* pertama dalam menentukan kelas dari data *testing*, jika K awal yang digunakan adalah 3 atau 3-NN maka hanya menggunakan 3 tetangga terdekat atau *rank* pertama sampai ketiga dalam menentukan kelas dari data *testing*, dan seterusnya sampai batas K awal adalah 9 atau 9-NN. Peneliti menggunakan data *testing* ke-1 dari *subset* 1 (DB-96) sebagai contoh dalam melakukan penentuan hasil klasifikasi data *testing*. Jumlah data *testing* pada *subset* 1 pada perhitungan ini yakni terdapat 12 data *testing* dengan 108 data *training* yang dapat dilihat pada Tabel 4 sebagai berikut.

Tabel 4 Hasil Pengurutan Jarak Euclid 108 Data *Training* Terhadap Data *Testing* ke-1 dari *Subset* 1 (DB-96)

Rank	Data Training		Data Testing ke-1 dari subset 1 (DB-96)	Batas K-NN	Hasil Klasifikasi Data Testing
	Sampel	Klasifikasi	$d_{(a,b)}$		
1	DB-75	L	2,67	1-NN	L
2	DB-62	L	3,01		
3	DB-52	L	3,29	3-NN	L
⋮	⋮	⋮	⋮	⋮	⋮
108	DB-74	L	6,60		

Hal tersebut dilakukan hal yang sama pada data *testing* ke-2 dari *subset* 1 (DB-25), data *testing* ke-3 dari *subset* 1 (DB-133), sampai dengan data *testing* ke-12 dari *subset* 10 (DB-72) untuk mendapatkan hasil klasifikasi data *testing* tiap *subset*.

Pemilihan Nilai K Optimal Berdasarkan Proporsi Terbaik

Pemilihan nilai K optimal dan proporsi terbaik dapat dilakukan dengan melihat nilai akurasi yang dihasilkan dari masing-masing *subset*. Perhitungan akurasi diawali dengan menghitung akurasi tiap *subset* berdasarkan nilai K dengan menggunakan kemudian dilakukan perhitungan rata-rata akurasi masing-masing *subset*. Adapun contoh hasil perhitungan akurasi menggunakan 1-NN dapat dilihat pada Tabel 5 sebagai berikut.

Tabel 5 Akurasi Perbandingan Hasil Klasifikasi pada *Subset* 1

Data Testing	Klasifikasi berdasarkan Nilai K					Klasifikasi Pada Data Asli (Y)
	1	3	5	7	9	
DB-96	L*	L*	L*	L*	L*	TL
DB-25	TL*	TL*	L	L	L	L
⋮	⋮	⋮	⋮	⋮	⋮	⋮
DB-5	L	L	L	L	L	L
Prediksi Benar	5	7	9	6	7	
$A_{(1,j)}$	0,42	0,58	0,75	0,50	0,58	

Berdasarkan Tabel 5 dapat dilihat bahwa sel dengan angka yang diberi tanda (*) memiliki perbedaan klasifikasi dengan klasifikasi pada data aslinya. Penggunaan batas 1 tetangga terdekat (1-NN) terdapat 5 data debitur yang tepat diklasifikasikan. Penggunaan batas 3 tetangga terdekat (3-NN) terdapat 7 data debitur yang tepat diklasifikasikan, dan seterusnya hingga penggunaan batas 9 tetangga terdekat (9-NN) untuk sehingga mendapat hasil klasifikasinya. Hasil klasifikasi yang benar (prediksi tepat) masing-masing nilai K akan dihitung jumlahnya kemudian dilanjutkan dengan menghitung akurasi hasil klasifikasi dari *subset* 1. Semakin banyak jumlah prediksi klasifikasi yang tepat maka semakin baik juga nilai K optimal yang didapat untuk mengklasifikasikan kategori status pembayaran kredit. Langkah-langkah tersebut akan dilakukan hal yang sama pada *subset* 2, *subset* 3, sampai dengan *subset* 10 untuk mendapatkan nilai akurasi tiap *subset.*

Setelah diperoleh akurasi tiap *subset*, kemudian dilakukan perhitungan rata-rata keseluruhan nilai akurasi berdasarkan nilai K. Semua tahapan tersebut diterapkan juga pada 2 proporsi yang berbeda yakni pada proporsi 50% : 50% dan 70% : 30%. Kemudian, dilakukan pemilihan proporsi terbaik dari 3 proporsi berdasarkan nilai akurasi tertinggi. Berikut ini merupakan nilai akurasi dengan 10-*fold cross validation* dari beberapa proporsi data *training* dan data *testing* yang dapat dilihat pada Tabel 6.

Tabel 6 Percobaan Pembagian Data *Training* dan Data *Testing* dengan 10-*fold cross validation*

Nilai K	Proporsi		
	50%:50%	70%:30%	90%:10%
1	57,68%	51,5%	60,83%
3	52,2%	57,74%	58,32%
5	51,17%	55,48%	64,18%
7	53,05%	61,29%	61,67%
9	50,04%	51,74%	57,49%

Berdasarkan Tabel 6 dapat dilihat bahwa nilai akurasi tertinggi terdapat pada proporsi 90% : 10% yakni pada batas 5 tetangga terdekat (5-NN) dengan nilai akurasi sebesar 64,18%. Sedemikian sehingga, nilai K optimal yang digunakan pada analisis selanjutnya yaitu K = 5 dengan 120 data *training* dan data *testing*.

Perhitungan Jarak Euclid antar Data *Training*

Tahapan ini bertujuan untuk mendapatkan tetangga terdekat yang akan digunakan pada proses pencarian validasi data *training*. Pada penelitian kali ini, jarak Euclid dihitung berdasarkan 120 data *training* terhadap 120 data *training*. Peneliti menggunakan data *training* ke-1 yaitu sampel pertama (DB-96) yang akan menjadi contoh dalam perhitungan jarak Euclid yang dapat dilihat sebagai berikut.

$$d_{(1,1)} = \sqrt{(1{,}74-1{,}74)^2+\cdots+(2{,}16)-2{,}16)^2}$$
$$= 0$$
$$d_{(2,1)} = \sqrt{(1{,}74-1{,}74)^2+\cdots+((-1{,}95)-2{,}16)^2}$$
$$= 6{,}33$$
$$\vdots$$
$$d_{(120,120)} = \sqrt{((-0{,}01)-(-0{,}01)^2+\cdots+((-2{,}21)-(-2{,}21))^2}$$
$$= 0$$

Perhitungan jarak Euclid dilakukan hingga data *training* ke-120 (DB-72) dengan data *training* ke-120 (DB-72). Setelah dilakukan perhitungan jarak Euclid pada data *training*, langkah selanjutnya yaitu mengurutkan jarak Euclid dari jarak yang paling dekat sampai dengan jarak yang paling jauh. Berikut adalah data *rank* jarak Euclid untuk data *training* ke-1 yaitu sampel pertama (DB-96) yang dapat dilihat pada Tabel 7.

Tabel 7 Hasil Pengurutan Jarak Euclid 120 Data *Training* Terhadap Data *Training* ke-1 (DB-96)

Rank	Data Training	Data Training ke-1 (DB-96)	Nilai K	Hasil $d_{(a,b)}$ Berdasarkan Nilai K
	Sampel	$d_{(a,b)}$		
1	DB-75	$d_{(65,1)}$= 2,67	1	$d_{(65,1)}$= 2,67
2	DB-62	$d_{(61,1)}$= 3,01		$d_{(61,1)}$= 3,01
3	DB-133	$d_{(3,1)}$= 3,24	3	$d_{(3,1)}$= 3,24
4	DB-52	$d_{(111,1)}$= 3,29		$d_{(111,1)}$= 3,29
5	DB-5	$d_{(12,1)}$= 3,32	5	$d_{(12,1)}$= 3,32
⋮	⋮	⋮		
120	DB-118	$d_{(69,1)}$= 7,61		

Berdasarkan Tabel 7 dapat diketahui bahwa terdapat 120 jarak dalam perhitungan jarak Euclid untuk data *training*. Dalam penelitian ini, karena nilai K optimal yang didapat dengan menggunakan 10-*fold cross validation* yaitu K = 5 maka jumlah tetangga terdekat akan diambil dari tetangga pertama sampai dengan tetangga kelima.

Validitas pada Data *Training*

Setelah dilakukan pengurutan jarak Euclid, selanjutnya dilakukan tahapan validitas. Nilai fungsi validitas diperoleh dari nilai kesamaan (*similiarity*) antara label kelas data *training* dengan label kelas jarak terdekat pada data *training*. Berdasarkan K optimal yang didapat sebelumnya yakni K = 5 maka jumlah tetangga yang terpilih dalam proses validitas (*H*) sebanyak 5 tetangga (5-NN). Perhitungan validitas pada data *training* dapat dilakukan dengan yakni sebagai berikut.

$$\text{Validitas}(1) = \frac{1}{5} \times S\big(lbl(65,1), lbl(61,1), \ldots, lbl(12,1)\big)$$
$$= \frac{1}{5} \times (0 + 0 + \cdots + 0)$$
$$= 0$$
$$\text{Validitas}(2) = \frac{1}{5} \times S(lbl(103,2), lbl(119,2), \ldots, lbl(90,2))$$
$$= \frac{1}{5} \times (0 + 0 + \cdots + 1)$$
$$= 0{,}4$$
$$\vdots$$
$$\text{Validitas}(120) = \frac{1}{5} \times S(lbl(107,120), lbl(35,120), \ldots, lbl(28,120))$$
$$= \frac{1}{5} \times (1 + 0 + \cdots + 0)$$
$$= 0{,}2$$

Perhitungan validitas data *training* dilakukan hingga data *training* ke-120 (DB-72). Peneliti menggunakan data *training* ke-1 yaitu sampel pertama (DB-96) sebagai contoh perhitungan validitas data *training*. Berikut hasil perhitungan validitas dengan menggunakan data *training* ke-1 (DB-96) pada Tabel 8.

Tabel 8 Hasil Perhitungan Validitas Data *Training*

Nilai K	Sampel	Data *Training*	Klasifikasi pada Data Asli		$S(a,b)$	Validitas(a)
		$d_{(a,b)}$	$d_{(a)}$	$d_{(b)}$		
1	DB-75	$d_{(65,1)}$= 2,67	L	TL	0	0
	DB-62	$d_{(61,1)}$= 3,01	L	TL	0	
3	DB-133	$d_{(3,1)}$= 3,24	L	TL	0	
	DB-52	$d_{(111,1)}$= 3,29	L	TL	0	
5	DB-5	$d_{(12,1)}$= 3,32	L	TL	0	
Total $S(a,b)$					0	

Berdasarkan Tabel 8 dapat diketahui bahwa secara keseluruhan dari kelima jarak data *training*, hasil klasifikasi jarak pada data *training* dengan klasifikasi jarak terdekat pada data *training* berbeda (tidak tepat) maka diperoleh nilai validitas pada *training* ke-1 (DB-96) yaitu 0. Hal ini dikarenakan secara keseluruhan klasifikasi data *training* dengan klasifikasi jarak terdekat pada data *training* tidak sama yakni untuk data *training* ke-65 adalah label kelas TL (kredit tidak lancar) sedangkan data *training* pertama adalah label kelas L (kredit tidak lancar). Selanjutnya dilakukan hal yang sama proses perhitungan validitas pada data *training* ke-2, data *training* ke-3, sampai dengan data *training* ke-120.

Perhitungan *Weight Voting*

Weight voting merupakan proses memasukkan nilai validasi data *training* serta perhitungan jarak Euclid yang telah diperoleh dengan melihat nilai bobot terbesar berdasarkan nilai K optimal. Perhitungan *weight voting* dilakukan sebagai contoh pada data *testing* ke-1 pada sampel ke-121 (DB-93) yakni sebagai berikut.

$$W_{(1,1)} = \text{Validitas}(1) \times \frac{1}{d(1,121)+0,5}$$
$$= 0 \times \frac{1}{6,11+0,5}$$
$$= 0$$
$$W_{(2,1)} = \text{Validitas}(2) \times \frac{1}{d(2,121)+0,5}$$
$$= 0,4 \times \frac{1}{4,69+0,5}$$
$$= 0,07$$
$$\vdots$$
$$W_{(120,13)} = \text{Validitas}(120) \times \frac{1}{d(120,13)+0,5}$$
$$= 0,2 \times \frac{1}{3,15+0,5}$$
$$= 0,06$$

Perhitungan *weight voting* dilakukan hingga data *training* ke-120 (DB-72) dan data *testing* ke- 13 (DB-7). Setelah dilakukan perhitungan *weight voting*, maka dilakukan proses pengurutan nilai *weight voting* dari bobot terbesar sampai dengan bobot terkecil (*descending*) tiap data *testing* berdasarkan nilai K optimal yaitu K = 5. Kemudian dilakukan *voting* kelas berdasarkan nilai *weight voting* yang bertujuan untuk memilih klasifikasi pada data *testing* yang paling banyak muncul. Berikut nilai *weight voting* untuk data *testing* ke-1 pada sampel ke-121 (DB-93) yang dapat dilihat pada Tabel 9 sebagai berikut.

Tabel 9 Hasil Pengurutan Nilai *Weight Voting* pada Data *Testing* ke-1 (DB-93)

Data Training	Data Testing ke-1 pada Sampel ke-121 (DB-93)	Nilai K	Hasil Klasifikasi Data Testing	Klasifikasi Awal Data Testing ke-1 pada Sampel ke-121 (DB-93)
Sampel	$W_{(a,b)}$			
DB-74	$W_{(91,121)}$= 0,43	1	L	L
DB-11	$W_{(26,121)}$= 0,42		L	
DB-98	$W_{(4,121)}$= 0,38	3	L	
DB-4	$W_{(56,121)}$= 0,37		L	
DB-18	$W_{(17,121)}$= 0,34	5	L	
⋮	⋮	⋮	⋮	⋮
DB-8	$W_{(119,121)}$= 0			

Berdasarkan Tabel 9 diketahui bahwa DB-74 merupakan data *training* ke-91 yang terdekat dari data *testing* pertama ditandai dengan nilai bobot yang paling besar dibandingkan bobot lainnya. Secara keseluruhan dari tetangga pertama sampai dengan tetangga kelima, hasil prediksi klasifikasi pada data *testing* ke-1 (DB-93) adalah label kelas L (status pembayaran kredit lancar). Selanjutnya dilakukan hal yang sama tahapan *weight voting* (bobot) pada data *testing* ke-2, data *testing* ke-3, sampai dengan data *testing* ke-13 (DB-7).

Akurasi Prediksi MK-NN

Setelah diperoleh hasil prediksi klasifikasi tiap data *testing*, selanjutnya melakukan perbandingan hasil klasifikasi dari metode MK-NN dengan klasifikasi data aslinya untuk semua percobaan 13 data *testing*. Berikut tabel hasil perbandingan klasifikasi yang dapat dilihat pada Tabel 10.

Tabel 10 Hasil Perbandingan Klasifikasi Metode MK-NN

Data Testing	Prediksi MK-NN	Klasifikasi Data Asli
DB-93	L	L
DB-94	L	L
DB-99	TL	TL
DB-116	L	L
DB-29	TL	TL
DB-69	TL	TL
DB-31	L	L
DB-108	L	L
DB-77	TL	TL
DB-67	L*	TL
DB-59	L*	TL
DB-105	L	L
DB-7	L	L

Berdasarkan Tabel 10 dapat dilihat bahwa sel yang diberi tanda (*) memiliki perbedaan klasifikasi dengan klasifikasi pada data aslinya. Terdapat 2 debitur tidak tepat diklasifikasikan dan sisanya terdapat 11 debitur yang tepat diklasifikasian pada status pembayaran kredit barang elektronik dan *furniture*. Berikut perhitungan akurasi prediksi tepat dari metode MK-NN yaitu sebagai berikut.

$$\text{Akurasi} = \frac{11}{13} \times 100\%$$
$$= 84{,}61\%$$

Berdasarkan perhitungan akurasi dari metode MK-NN dengan menggunakan K optimal yakni K = 5 diperoleh nilai akurasinya yaitu sebesar 84,61% yang menggambarkan bahwa ketepatan klasifikasi dengan menggunakan 5 tetangga terdekat (5-NN) pada proporsi 90% : 10% sebagian besar data *testing* memiliki kesamaan hasil kasifikasi dengan klasfikasi dari data aslinya. Metode MK-NN dapat mengklasifikasikan status pembayaran kredit yakni dari 9 debitur yang memiliki status pembayaran kredit lancar, terdapat 7 debitur yang tepat diklasifikasikan dan sisanya terdapat 2 debitur yang tidak tepat diklasifikasikan. Dari 4 debitur yang memiliki status pembayaran kredit tidak lancar, terdapat 4 debitur yang tepat diklasifikasikan.

4. SIMPULAN DAN SARAN

Kesimpulan

Metode klasifikasi dengan menggunakan *Modified K-Nearest Neighbor* (MK-NN) dalam memprediksi status pembayaran kredit barang elektronik dan *furniture* di PT. KB Finansia Multi Finance tahun 2020 berdasarkan proporsi 90% : 10% dengan K optimal yakni K = 5 diperoleh 11 debitur yang tepat diklasifikasikan dan sisanya terdapat 2 debitur yang tidak tepat diklasifikasikan. Adapun nilai akurasi yang diperoleh dengan menggunakan MK-NN yakni sebesar 84,61%.

Saran

Dalam penelitian selanjutnya dapat menggunakan nilai *k-fold* lainnya pada *k-fold cross validation*. Selain itu dapat menggunakan metode klasifikasi seperti halnya *Genetic Modified K-Nearest Neighbor* (GMK-NN), *Support Vector Machine* (SVM), *Artificial Neural Network* (ANN), atau metode pengklasifikasian lainnya. Selain itu, dapat dikembangkan dengan menggunakan data yang memiliki lebih dari dua kategori pada variabel terikatnya.

DAFTAR PUSTAKA

Fernanda, S.I., Ratnawati, D.E., & Adikara P.P. (2017). Identifikasi Penyakit Diabetes Melitus Menggunakan Metode Modified K-Nearest Neighbor (MKNN). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*. Vol. 1 No. 6, 507-513.

Gorunescu, F. (2011). *Data Mining: Concepts, Models, and Techniques*. Verlag Berlin Heidelberg: Springer.

Han, J.W, Kamber, M. and Pei, J. (2012). *Data Mining Concepts and Techniques 3rd Edition*. San Fransisco: Morgan Kaufmann Publishers.

Imanda, A.C, Hidayat, N., & Furqon, M.T. (2018). Klasifikasi Kelompok Varietas Unggul Padi Menggunakan Modified K-Nearest Neighbor. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*. 2(8), 2392-2399.

Mustafa. (2014). Perancangan Aplikasi Prediksi Kelulusan Tepat Waktu Bagi Mahasiswa Baru Dengan Teknik Data Mining (Studi Kasus : Data Akademik Mahasiswa STMIL Dipanedgara). *Citec Jurnal*. 1(3).

Pandie, E. S. Y. (2012). Implementasi Algoritma Data Mining K-Nearest Neighbor (KNN) Dalam Pengambilan Keputusan Pengajuan Kredit. *Jurnal Ilmu Komputer Universitas Nusa Cendana.*

Parvin, H., Alizadeh, H., & Minati, B. (2010). A Modification on K-Nearest Neighbor Classifier. *Global Journal of Computer Science and Technology*, 10 (14), 37- 41.

Rodiyansyah, S.F dan Winarko, E. (2013). Klasifikasi Posting Twitter Kemacetan Lalu Lintas Kota Bandung Menggunakan Naive Bayesian Classification. *Jurnal Universitas.*

Saxena, K., Khan, Z., and Singh, S. (2014). Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm. *International Journal of Computer Science Trends and Technology* (IJCST), 2(4), 36-43.