PENERAPAN ALGORITMA *K-NEAREST NEIGHBOR* UNTUK PENENTUAN RESIKO KREDIT KEPEMILIKAN KENDARAAN BEMOTOR

Henny Leidiyana
Program Pasca Sarjana Magister Ilmu Komputer STMIK Nusa Mandiri
Email : tamma_nuruka@yahoo.com

ABSTRAK

Sejalan dengan pertumbuhan bisnis, kredit merupakah masalah yang menarik untuk diteliti. Beberapa riset bidang komputer untuk mengurangi resiko kredit telah banyak dilakukan dalam rangka menghindarai kehancuran suatu perusahaan pembiayaan. Paper ini membahas algoritma k-Nearest Neighbor (kNN) yang diterapkan pada data konsumen yang menggunakan jasa keuangan kredit kendaraan bermotor. Hasil testing untuk mengukur performa algoritma ini menggunakan metode Cross Validation, Confusion Matrix dan kurva ROC dan menghasilkan akurasi dan nilai AUC berturut-turut 81,46 % dan 0,984. Karena nilai AUC berada dalam rentang 0,9 sampai 1,0 maka metode tersebut masuk dalam kategori sangat baik (*excellent*).

Kata kunci : K-Nearest Neighbor, Cross Validation, Confusion matrix, ROC

ABSTRACT

In line with the growth and business development, credit issues remain to be studied and revealed interesting. Some of the research field of computers has done much to reduce the credit risk of causing harm to the company. In this study, k-Nearest Neighbor (kNN) algorithm is applied to the data of consumers who have good credit financing motorcycle that consumers are troubled or not. From the test results to measure the performance of the algorithms using the test method Cross Validation, Confusion Matrix and ROC curves, it is known that the accuracy value of 81.46% and AUC values of 0.984. This methodes is include excellent classification because the AUC value between 0.90-1.00.

Keywords: K-Nearest Neighbor, Cross Validation, Confusion matrix, ROC

1. Pendahuluan

Dari penelitian-penelitian yang pernah dilakukan, evaluasi resiko kredit merupakan masalah yang menarik dalam analisa keuangan. Penelitian mengenai analisis kelayakan pemberian kredit untuk konsumen khususnya kredit kepemilikan barang dengan metode klasifikasi *data mining* telah banyak dilakukan. Dalam penulisan ini akan dibahas mengenai penerapan algoritma *k-nearest neighbor* untuk penentuan resiko kredit kepemilikan kendaraan bermotor.

2. Landasan Teori

2.1. Leasing

Menurut Surat Keputusan Bersama Menteri Keuangan, Perindustrian dan Perdagangan No.1169/KMK.01/1991 tanggal 21 Nopember 1991 tentang kegiatan Sewa Guna Usaha, *Leasing* (Noerlina, 2007) adalah setiap kegiatan pembiayaan perusahaan dalam bentuk penyediaan barang-barang modal untuk digunakan oleh suatu perusahaan untuk jangka waktu tertentu, berdasarkan pembayaran-pembayaran berkala disertai dengan hak pilih (opsi) bagi perusahaan tersebut untuk membeli barang-barang modal yang bersangkutan atau memperpanjang jangka waktu *leasing* berdasarkan nilai sisa yang telah disepakati.

2.2. Kredit

Kredit adalah penyerahan barang, jasa, atau uang dari satu pihak (kreditor/pemberi pinjaman) atas dasar kepercayaan kepada pihak lain (nasabah atau pengutang/*borrower*) dengan janji membayar dari penerima kredit kepada pemberi kredit pada tanggal yang telah disepakati kedua belah pihak (Rivai, 2006). Agar kredit yang diberikan mencapai sasaran, yaitu aman, maka analisis kredit perlu dilakukan. Analisis kredit (Rivai, 2006) adalah kajian yang dilakukan untuk mengetahui kelayakan dari suatu permasalahan kredit. Melalui hasil analisis kreditnya, dapat diketahui apakah usaha nasabah layak (*feasible*), *marketable* (hasil usaha dapat dipasarkan), *profitable* (menguntungkan), serta dapat dilunasi tepat waktu. Untuk mewujudkan hal tersebut, perlu dilakukan persiapan kredit, yaitu dengan mengumpulkan informasi dan data untuk bahan analisis. Kualitas hasil analisis tergantung pada kualitas SDM, data yang diperoleh, dan teknik analisis.

2.3. Data Mining

Data Mining (Witten, 2011) didefinisikan sebagai proses penemuan pola dalam data. Berdasarkan tugasnya, *data mining* dikelompokkan menjadi deskripsi, estimasi, prediksi, klasifikasi, *clustering* dan asosiasi (Larose, 2005). Proses dalam tahap *data mining* (Gambar 1.) terdiri dari tiga langkah Utama (Sumathi, 2006), yaitu :

a. *Data Preparation*

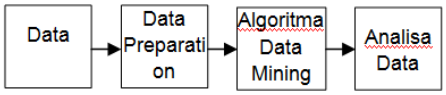
Pada langkah ini, data dipilih, dibersihkan, dan dilakukan *preprocessed* mengikuti pedoman dan *knowledge* dari ahli domain yang menangkap dan mengintegrasikan data internal dan eksternal ke dalam tinjauan organisasi secara menyeluruh.

b. Algoritma *data mining*

Penggunaan algoritma *data mining* dilakukan pada langkah ini untuk menggali data yang terintegrasi untuk memudahkan identifikasi informasi bernilai.

c. Fase analisa data

Keluaran dari data mining dievaluasi untuk melihat apakah *knowledge* domain ditemukan dalam bentuk *rule* yang telah diekstrak dari jaringan.



Gambar 1 Langkah-langkah dalam Proses *Data Mining* (Maimon & Rokach, 2010)

2.4. Klasifikasi

Klasifikasi adalah proses penemuan model (atau fungsi) yang menggambarkan dan membedakan kelas data atau konsep yang bertujuan agar bisa digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui (Han, 2006). Algoritma klasifikasi yang banyak digunakan secara luas, yaitu *Decision/classification* trees, *Bayesian classifiers/ Naïve Bayes classifiers*, *Neural networks*, Analisa Statistik, Algoritma Genetika, *Rough sets*, *k-nearest neighbor*, Metode *Rule Based*, *Memory based reasoning*, dan *Support vector machines* (SVM).

Klasifikasi data terdiri dari 2 langkah proses. Pertama adalah *learning* (fase *training*), dimana algoritma klasifikasi dibuat untuk menganalisa data *training* lalu direpresentasikan dalam bentuk *rule* klasifikasi. Proses kedua adalah klasifikasi, dimana data tes digunakan untuk memperkirakan akurasi dari *rule* klasifikasi (Han, 2006). Proses klasifikasi didasarkan pada empat komponen (Gorunescu, 2011) :

a. Kelas

Variabel dependen yang berupa kategorikal yang merepresentasikan 'label' yang terdapat pada objek. Contohnya: resiko penyakit jantung, resiko kredit, *customer loyalty*, jenis gempa.

b. Predictor

Variabel independen yang direpresentasikan oleh karakteristik (atribut) data. Contohnya: merokok, minum alkohol, tekanan darah, tabungan, aset, gaji.

c. Training dataset

Satu set data yang berisi nilai dari kedua komponen di atas yang digunakan untuk menentukan kelas yang cocok berdasarkan *predictor.*

d. Testing dataset

Berisi data baru yang akan diklasifikasikan oleh model yang telah dibuat dan akurasi klasifikasi dievaluasi

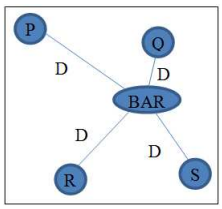
2.5. *K-nearest neighbor*

k-nearest neighbor (kNN) termasuk kelompok *instance-based learning*. Algoritma ini juga merupakan salah satu teknik *lazy learning*. kNN dilakukan dengan mencari kelompok k objek dalam data *training* yang paling dekat (mirip) dengan objek pada data baru atau data *testing* (Wu, 2009).

Contoh kasus, misal diinginkan untuk mencari solusi terhadap masalah seorang pasien baru dengan menggunakan solusi dari pasien lama. Untuk mencari solusi dari pasien baru tersebut digunakan kedekatan dengan kasus pasien lama, solusi dari kasus lama yang memiliki kedekatan dengan kasus baru digunakan sebagai solusinya.

Terdapat pasien baru dan 4 pasien lama, yaitu P, Q, R, dan S (Gambar 2). Ketika ada pasien baru maka yang diambil solusi adalah

solusi dari kasus pasien lama yang memikili kedekatan terbesar.



Gambar 2. ilustrasi kasus algoritma kNN

Misal D1 adalah jarak antara pasien baru dengan pasien P, D2 adalah jarak antara pasien baru dengan pasien Q, D3 adalah jarak antara pasien baru dengan pasien R, D4 adalah jarak antara pasien baru dengan pasien S. Dari ilustrasi gambar terlihat bahwa D2 yang paling terdekat dengan kasus baru. Dengan demikian maka solusi dari kasus pasien Q yang akan digunakan sebagai solusi dari pasien baru tersebut.

Ada banyak cara untuk mengukur jarak kedekatan antara data baru dengan data lama (data *training*), diantaranya e*uclidean distance* dan *manhattan distance* (*city block distance*), yang paling sering digunakan adalah *euclidean distance* (Bramer,2007), yaitu:

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2}$$

Dimana a = a_1,a_2, …, a_n, dan b = b_1, b_2, …, b_n mewakili n nilai atribut dari dua *record*.

Untuk atribut dengan nilai kategori, pengukuran dengan *euclidean distance* tidak cocok. Sebagai penggantinya, digunakan fungsi sebagai berikut (Larose, 2006):

different(a_i,b_i) { 0 jika $a_i = b_i$

= 1 selainnya

Dimana a_i dan b_i adalah nilai kategori. Jika nilai atribut antara dua *record* yang dibandingkan sama maka nilai jaraknya 0, artinya mirip, sebaliknya, jika berbeda maka nilai kedekatannya 1, artinya tidak mirip sama sekali. Misalkan atribut warna dengan nilai merah dan merah, maka nilai kedekatannya 0, jika merah dan biru maka nilai kedekatannya 1.

Untuk mengukur jarak dari atribut yang mempunyai nilai besar, seperti atribut pendapatan, maka dilakukan normalisasi. Normalisasi bisa dilakukan dengan *min-max normalization* atau *Z-score standardization* (Larose, 2006). Jika data *training* terdiri dari atribut campuran antara numerik dan kategori, lebih baik gunakan *min-max normalization* (Larose, 2006).

Untuk menghitung kemiripan kasus, digunakan rumus (Kusrini, 2009):

$$\text{Similarity (p, q)} = \frac{\sum_{i=1}^{n} f(pi, qi)\, X\, wi}{wi}$$

Keterangan :

P = Kasus baru

q = Kasus yang ada dalam penyimpanan

n = Jumlah atribut dalam tiap kasus

i = Atribut individu antara 1 sampai dengan n

f = Fungsi *similarity* atribut i antara kasus p dan kasus q

w = Bobot yang diberikan pada atribut ke-i

2.6. Metode Evaluasi dan Validasi Algoritma Klasifikasi Data mining

Untuk mengukur akurasi algoritma klasifikasi, metode yang dapat digunakan antara lain *cross validation, confusion matrix*, dan kurva ROC (*Receiver Operating Characteristic*). Untuk mengembangkan aplikasi (*development*) berdasarkan model yang dibuat, digunakan Rapid Miner.

a. *Cross Validation*

Cross validation adalah pengujian standar yang dilakukan untuk memprediksi *error rate*. Data *training* dibagi secara random ke dalam beberapa bagian dengan perbandingan yang sama kemudian *error rate* dihitung bagian demi bagian, selanjutnya hitung rata-rata seluruh *error rate* untuk mendapatkan *error rate* secara keseluruhan.

b. *Confusion matrix*

Metode ini menggunakan tabel matriks seperti padaTabel 1 jika data set hanya terdiri dari dua kelas, kelas yang satu dianggap sebagai positif dan yang lainnya negatif (Bramer, 2007).

Tabel 1 Model *Confusion Matrix* (Bramer, 2007)

Klasifikasi yang benar	Diklasifikasikan sebagai	
	+	-
+	true positives	false negatives
-	false positives	true negatives

True positives adalah jumlah *record* positif yang diklasifikasikan sebagai positif, *false positives* adalah jumlah *record* negatif yang diklasifikasikan sebagai positif, *false negatives* adalah jumlah *record* positif yang diklasifikasikan sebagai negatif, *true negatives* adalah jumlah *record* negatif yang diklasifikasikan sebagai negative, kemudian masukkan data uji. Setelah data uji dimasukkan ke dalam *confusion matrix*, hitung nilai-nilai yang telah dimasukkan tersebut untuk dihitung jumlah *sensitivity* (*recall*), *specificity*, *precision* dan *accuracy*.

Sensitivity digunakan untuk membandingkan jumlah TP terhadap jumlah *record* yang positif sedangkan *specificity* adalah perbandingan jumlah TN terhadap jumlah *record* yang negatif. Untuk menghitung digunakan persamaan di bawah ini (Han, 2006) :

$$sensitivity = \frac{TP}{P}$$

$$specificity = \frac{TN}{N}$$

$$precision = \frac{TP}{TP + FP}$$

$$accuracy = sensitivity\frac{P}{(P + N)} + specificity\frac{N}{(P + N)}$$

Keterangan:

TP = jumlah *true positives*

TN = jumlah *true negatives*

P = jumlah *record* positif

N = jumlah *tupel* negatif

c. FP = jumlah *false positives*

Cross validation adalah pengujian standar yang dilakukan untuk memprediksi *error rate*. Data *training* dibagi secara random ke dalam beberapa bagian dengan perbandingan yang sama kemudian *error rate* dihitung bagian demi bagian, selanjutnya hitung rata-rata seluruh *error rate* untuk mendapatkan *error rate* secara keseluruhan.

d. Kurva ROC

Kurva ROC menunjukkan akurasi dan membandingkan klasifikasi secara visual. ROC mengekspresikan *confusion matrix*. ROC adalah grafik dua dimensi dengan *false positives* sebagai garis horisontal dan *true positives* sebagai garis vertikal (Vecellis, 2009). *The area under curve* (AUC) dihitung untuk mengukur perbedaan performansi

metode yang digunakan. AUC dihitung menggunakan rumus (Liao, 2007) :

$$\theta^r = \frac{1}{mn}\sum_{j=1}^{n}\sum_{i=1}^{m}\psi\left(xi^r, xj^r\right)$$

Dimana

$$(X,Y) = \begin{cases} 1 & Y < X \\ \frac{1}{2} & Y = X \\ 0 & Y > X \end{cases}$$

K = jumlah algoritma klasifikasi yang dikomparasi

X = output *positif*

Y = output *negatif*

3. Hasil Penelitian

Penelitian ini menggunakan 558 *record* transaksi kredit baik yang bermasalah maupun yang tidak bermasalah, yang diperoleh dari sebuah *leasing* yang berlokasi di Cikarang. Semua atribut pada data *training* bernilai kategori, seperti terlihat pada Tabel 2. data *training* terdiri dari 14 atribut, dimana 13 atribut merupakan prediktor dan 1 atribut label. Untuk mendapatkan data yang berkualitas, dilakukan *preprocessing*. Setelah dilakukan *preprocessing*, 558 *record* direduksi dengan menghilangkan duplikasi menjadi 481 *record* untuk data *training*.

Tabel 2 Daftar atribut dan nilainya

no	Atribut	Nilai atribut
1	status perkawinan	Menikah
		belum menikah
		janda/duda
2	jumlah tanggungan	tidak ada
		1 orang
		2-3 orang
		> 3 orang
3	pendidikan terakhir	>S1
		S1
		Diploma
		SLTA
		SLTP
		SD
		tidak sekolah
4	Usia	< 21 tahun atau > 60 tahun
		21 - 55 tahun
		55 - 60 tahun
5	kepemilikan rumah	milik sendiri (PBB/srtfkt/AJB/rek listrik)
		milik sendiri (PBB a/n orang lain)
		KPR
		orang tua
		Keluarga
		dinas
		sewa/kontrak >= tenor
		sewa/kontrak < tenor
		Kost
6	lama tinggal	> 5 tahun
		3 - 5 tahun
		1 - 3 tahun
		< 1 tahun
7	kondisi rumah	Permanen
		non permanen
8	jenis pekerjaan	PNS
		TNI/POLRI
		Jaksa
		Karyawan
		wiraswasta kecil
		wiraswasta menengah
		wiraswasta besar
		Konsultan
		Dokter
		Dosen
		Guru
		Pengacara
		Pensiunan
9	status perusahaan	BUMN/D
		swasta besar
		swasta menengah
		swasta kecil
		perorangan

		lembaga pendidikan
		lembaga negara
10	status kepegawaian	tetap
		kontrak
		harian buruh pabrik
		harian buruh non pabrik
		pemilik
11	Masa kerja/usaha	> 5 tahun
		2 - 5 tahun
		< 2 tahun
12	penghasilan perbulan	> 3x angsuran dengan slip
		> 3x angsuran tanpa slip
		> 2x angsuran dengan slip
		> 2x angsuran tanpa slip
		> 1x angsuran dengan slip
		> 1x angsuran tanpa slip
		< 1x angsuran
13	pembayaran pertama	> 30 %
		20 - 30%
		10 - 20 %
		< 10 %
14	Remark	bad cust
		Good cust

Untuk mengukur jarak antar atribut, akan diberikan bobot pada atribut. Bobot jarak ini diberikan nilai antara 0 sampai dengan 1. Nilai 0 artinya jika atribut tidak berpengaruh dan sebaliknya nilai 1 jika atribut sangat berpengaruh.

Tabel 3 pembobotan atribut

no	Atribut	Bobot
1	status perkawinan	0.5
2	jumlah tanggungan	1
3	pendidikan terakhir	0.5
4	usia	0.5
5	kepemilikan rumah	0.8
6	lama tinggal	1
7	kondisi rumah	0.6
8	jenis pekerjaan	0.8
9	status perusahaan	0.5
10	status kepegawaian	0.8
11	Masa kerja/usaha	1
12	penghasilan perbulan	1
13	pembayaran pertama	1

Contoh penentuan kedekatan antar nilai atribut terdapat pada table 4, misalkan untuk atribut status perkawinan terdiri dari tiga nilai kategori, yaitu menikah, belum menikah, dan duda/janda.

Tabel 4 kedekatan nilai atribut status perkawinan

Atribut	Nilai atribut 1	Nilai atribut 2	Bobot
status perkawinan	Menikah	Menikah	0
	Menikah	Belum menikah	1
	Menikah	Duda/janda	0.5
	Belum menikah	Belum menikah	0
	Belum menikah	Duda/janda	0.5
	Duda/janda	Duda/janda	0

Pembobotan nilai atribut dilakukan untuk 13 atribut prediktor. Setelah itu hitung kemiripannya. Misal sebuah data konsumen baru akan diklasifikasi apakah bermasalah atau tidak dalam pembayaran angsuran motor maka dilakukan perhitungan kedekatan antara kasus baru dibandingkan dengan data kasus lama (data *training*).

Tabel 5 berisi sampel data *training* yang merupakan kasus lama dan akan diukur kedekatannya dengan kasus yang baru.

Tabel 5 sampel data *training*

status perkawinan	jumlah tanggungan	pendidikan terakhir	usia	kepemilikan rumah	lama tinggal	kondisi rumah
menikah	tidak ada	SLTA	< 21/> 60 thn	Ortu	3-5	permanen
menikah	>3	SLTA	21-55	KPR	3-5	permanen

Tabel 5 sampel data *training* (lanjutan)

jenis pekerjaan	status perusahaan	status kepegawaian	masa kerja	penghasilan perbulan	total pembayaran pertama	remark
karyawan	swasta menengah	kontrak	<2	> 2x ang slip	10-20%	bad
karyawan	swasta menengah	tetap	>5	> 3x ang slip	10-20%	good

Misalkan ada kasus baru pada data *testing* dengan nilai atribut seperti pada tabel 6. Kasus baru tersebut akan dihitung kedekatannya dengan kasus lama yang terdapat pada data training table 5.

Tabel 6 sampel data *testing*

status perkawinan	jumlah tanggungan	pendidikan terakhir	Usia	kepemilikan rumah	lama tinggal	kondisi rumah
Belum menikah	Tidak ada	SLTP	< 21/> 60 thn	Ortu	>5	permanen

Tabel 6 sampel data *testing* (lanjutan)

jenis pekerjaan	status perusahaan	statudegs kepegawaian	masa kerja	penghasilan perbulan	total pembayaran pertama	remark
karyawan	swasta kecil	Kontrak	<2	> 2x ang slip	10-20%	bad

Perhitungan kedekatan kasus baru pada data *testing* (Tabel 6) dengan 2 kasus lama pada data *training* (Tabel 5), yaitu:

Kedekatan kasus baru dengan kasus nomor 1

A: Kedekatan bobot atribut status perkawinan (menikah dengan belum menikah) = 1

B: Bobot atribut status perkawinan = 0.5

C: Kedekatan bobot jumlah tanggungan (tidak ada dengan tidak ada) = 0

D: Bobot atribut jumlah tanggungan = 1

E: Kedekatan bobot pendidikan terakhir (SLTA dengan SLTA) = 0

F: Bobot atribut pendidikan terakhir = 0.5

G: Kedekatan bobot usia (< 21 tahun/ > 60 tahun dengan < 21 tahun/ > 60 tahun) = 0

H: Bobot atribut usia = 0.5

I: Kedekatan bobot kepemilikan rumah (orang tua dengan orang tua) = 1

J: Bobot atribut kepemilikan rumah = 0.8

K: Kedekatan bobot lama tinggal (3-5 tahun dengan >5tahun) = 0.5

L: Bobot atribut lama tinggal = 1

M: Kedekatan bobot kondisi rumah (permanen dengan permanen) = 0

N: Bobot atribut kondisi rumah = 0.6

O: Kedekatan bobot jenis pekerjaan (karyawan dengan karyawan) = 0

P: Bobot atribut jenis pekerjaan = 0.8

Q: Kedekatan bobot status perusahaan (swasta menengah dengan swasta kecil) = 0.5

R: Bobot atribut status perusahaan = 0.5

S: Kedekatan bobot status kepegawaian (kontrak dengan kontrak) = 0

T: Bobot atribut status kepegawaian = 0.8

U: Kedekatan bobot masa kerja/usaha (<2 tahun dengan <2 tahun) = 0

V: Bobot atribut masa kerja/usaha = 1

W: Kedekatan bobot penghasilan perbulan (>2x ang slip dengan >2x ang slip) = 0

X: Bobot atribut penghasilan perbulan = 1

Y: Kedekatan bobot pembayaran pertama (10-20% dengan 10-20%) = 0

Z: Bobot atribut pembayaran pertama = 1

Similarity = [(A*B) + (C*D) + (E*F) + (G*H) + (I*J) + (K*L) + (M*N) + (O*P) + (Q*R) + (S*T) + (U*V) + (W*X) + (Y*Z)] / (B+D+F+H+J+ L+N+P+R+T+V+X+Z)

= [(1*0.5) + (0*1) + (0*0.5) + (0*0.5) + (1*0.8) + (0.5*1) + (0*0.6) + (0*0.8) +(0.5*0.5) + (0* 0.8) + (0* 1) + (0* 1) + (0*1)] / (0.5+1+0.5+0.5+0.8+1+0.6+ 0.8 +0.5+0.8+1+1+1)

= (0.5+0+0+0+0.8+0.5+0+0.25+ 0+0+0+0+0)/10

= 2.05/10

= 0.205

Kedekatan kasus baru dengan kasus nomor 2

A : Kedekatan bobot atribut status perkawinan (menikah dengan belum menikah) = 1

B : Bobot atribut status perkawinan = 0.5

C : Kedekatan bobot jumlah tanggungan (>3 dengan tidak ada) = 1

D : Bobot atribut jumlah tanggungan = 1

E : Kedekatan bobot pendidikan terakhir (SLTA dengan SLTA) = 0

F : Bobot atribut pendidikan terakhir = 0.5

G : Kedekatan bobot usia (21-55 tahun dengan < 21 tahun/ > 60 tahun) = 0.5

H : Bobot atribut usia = 0.5

I : Kedekatan bobot kepemilikan rumah (KPR dengan orang tua) = 1

J : Bobot atribut kepemilikan rumah = 0.8

K : Kedekatan bobot lama tinggal (3-5 tahun dengan >5tahun) = 0.5

L : Bobot atribut lama tinggal = 1

M : Kedekatan bobot kondisi rumah (permanen dengan permanen) = 0

N : Bobot atribut kondisi rumah = 0.6

O : Kedekatan bobot jenis pekerjaan (karyawan dengan karyawan) = 0

P Bobot atribut jenis pekerjaan = 0.8

Q : Kedekatan bobot status perusahaan (swasta menengah dengan swasta kecil) = 0.5

R : Bobot atribut status perusahaan = 0.5

S : Kedekatan bobot status kepegawaian (tetap dengan kontrak) = 1

T : Bobot atribut status kepegawaian = 0.8

U : Kedekatan bobot masa kerja/usaha (>5 tahun dengan <2 tahun) = 1

V : Bobot atribut masa kerja/usaha = 1

W : Kedekatan bobot penghasilan perbulan (>3x ang slip dengan >2x ang slip) = 0.5

X : Bobot atribut penghasilan perbulan = 1

Y : Kedekatan bobot pembayaran pertama (10-20% dengan 10-20%) = 0

Z : Bobot atribut pembayaran pertama = 1

Similarity = [(A*B) + (C*D) + (E*F) + (G*H) + (I*J) + (K*L) + (M*N) + (O*P) + (Q*R) + (S*T) + (U*V) + (W*X) + (Y*Z)] / (B+D+F+H+J+ L+N+P+R+T+V+X+Z)

= [(1*0.5) + (1*1) + (0*0.5) + (0.5*0.5) + (1*0.8) + (0.5*1) + (0*0.6) + (0*0.8) + (0.5*0.5) + (1* 0.8) + (1* 1) + (0.5* 1) + (0*1)]/ (0.5+1+0.5+0.5+0.8+1 + 0.6+ 0.8+0.5+0.8+1+1+1)

= (0.5+1+0+0.25+0.8+0.5+0+ 0.25+0.8+1+0.5+0)/10

= 5.6/10

= 0.56

Setelah dihitung nilai kedekatannya yang terendah adalah kasus nomor 1. Dengan demikian kasus yang terdekat dengan kasus baru adalah kasus nomor 1. Jadi kemungkinan konsumen baru tersebut akan bermasalah dalam pembayaran angsurannya.

4. Pengujian Algoritma

1. *Cross Validation*

Dalam penelitian ini digunakan *10 fold-cross validation* dimana 481 record pada data *training* dibagi secara random ke dalam 10 bagian dengan perbandingan yang sama kemudian *error rate* dihitung bagian demi bagian, selanjutnya hitung rata-rata seluruh *error rate* untuk mendapatkan *error rate* secara keseluruhan.

2. *Confusion Matrix*

Tabel 7 adalah table *confusion matrix* yang dihasilkan dengan menggunakan algoritma kNN. Perhitungan kedekatan kasus lama pada data *training* dengan kasus baru pada data *testing*, diketahui dari 481 data, 162 diklasifikasikan *bad* , 15 data diprediksi *bad* tetapi ternyata *good*, 233 data *class good* diprediksi sesuai, dan 75 data diprediksi *good* ternyata *bad*. Tingkat akurasi penerapan algoritma kNN ini sebesar 81.46%.

Tabel 7 Model *Confusion Matrix* untuk Metode kNN

accuracy: 81.46% +/- 4.16% (mikro: 81.44%)			
	true bad	true good	class precision
pred. bad	162	75	68.35%
pred. good	15	233	93.95%
class recall	91.53%	75.65%	

3. Kurva ROC

Hasil perhitungan divisualisasikan dengan kurva ROC. Kurva ROC pada gambar

mengekspresikan *confusion matrix* dari Tabel 7. Garis horizontal adalah *false positives* dan garis vertikal *true positives*. Terlihat pada table, nilai AUC sebesar 0.984.

Gambar 3 Kurva ROC dengan Metode kNN

Untuk klasifikasi *data mining*, nilai AUC dapat dibagi menjadi beberapa kelompok (Gorunescu, 2011).

a. 0.90-1.00 = klasifikasi sangat baik
b. 0.80-0.90 = klasifikasi baik
c. 0.70-0.80 = klasifikasi cukup
d. 0.60-0.70 = klasifikasi buruk
e. 0.50-0.60 = klasifikasi salah

Berdasarkan pengelompokkan di atas maka dapat disimpukan bahwa metode kNN termasuk klasifikasi sangat baik karena memiliki nilai AUC antara 0.90-1.00.

5. Kesimpulan

Dalam penelitian ini dilakukan penerapan algoritma kNN pada data konsumen yang mendapat pembiayaan kredit motor. Agar didapat data yang berkualitas, dilakukan *preprosesing* sebelum diterapkan ke dalam algoritma. Kedekatan antara kasus baru dengan kasus lama dilakukan untuk menentukan termasuk kelas mana kasus baru tersebut. Untuk mengukur kinerja algoritma tersebut digunakan metode *Cross Validation, Confusion Matrix* dan Kurva ROC, diketahui nilai *accuracy* 81.46% dan termasuk klasifikasi sangat baik karena memiliki nilai AUC antara 0.90-1.00, yaitu sebesar 0.984.

Daftar Pustaka

Bramer, Max. 2007. *Principles of Data Mining*. London : Springer

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

Han, J.,&Kamber, M. 2006.*Data Mining Concept and Tehniques*.San Fransisco : Morgan Kauffman.

Kusrini&Luthfi,E.T. 2009. Algoritma Data Mining. Yogyakarta : Andi Publishing.

Larose, D. T. 2005. *Discovering Knowledge in Data*. New Jersey : John Willey & Sons, Inc.

Liao. 2007. *Recent Advances in Data Mining of Enterprise Data* : *Algorithms and Application*. Singapore : *World Scientific Publishing*

Maimon, Oded&Rokach, Lior. 2005. *Data Mining and Knowledge Discovey Handbook*. New York : Springer

Noerlina. 2007. Perancangan Sistem Informasi Berbasis *Object Oriented*. Jakarta : MitraWacana Media.

Rivai, Veithzal.,&Veithzal, Andria Permata. 2006. *Credit Management Handbook*. Jakarta : Raja GrafindoPersada.

Sumathi, & S., Sivanandam, S.N. 2006. *Introduction to Data Mining and its Applications*. Berlin Heidelberg New York: Springer

Vercellis, Carlo. 2009. *Business Intelligent: Data Mining and Optimization for Decision Making*. Southern Gate, Chichester, West Sussex : John Willey & Sons, Ltd.

Witten, I. H., Frank, E., & Hall, M. A. 2011. *Data Mining: Practical Machine Learning and Tools*. Burlington : Morgan Kaufmann Publisher