ISSN 2541-1004

PERBANDINGAN ALGORITMA K-NEAREST NEIGHBOR DAN DECISION TREE UNTUK PENENTUAN RISIKO KREDIT KEPEMILIKAN MOBIL

Devi Yunita
Prodi Teknik Informatika Universitas Pamulang
Jl. Raya Puspitek No.46, Serpong – Tangerang Selatan
email: devienitha@gmail.com

ABSTRAK

Perbandingan Algoritma K-Nearest Neighbor Dan Decision Tree untuk Risiko Kredit Kepemilikan Mobil Kredit adalah sarana agar orang atau perusahaan dapat meminjam modal atau uang dan membayarnya dalam tempo yang sudah ditentukan. Agar kredit yang diberikan sesuai tujuan atau sasaran, yaitu aman, maka perlu diakukan analisis kredit. Analisis kredit adalah kajian yang dilakukan untuk mengetahui kelayakan dari suatu permasalahan kredit. Dalam penelitian analisa kredit ini menggunakan perbandingan Algoitma K-nearest neighbor (K-NN) yang merupakan penelitian menggunakan metode dengan mencari kedekatan antara kriteria kasus baru dengan kriteria kasus lama berdasarkan kriteria kasus yang paling mendekati, dan menggunakan Metode Decision tree yang merupakan metode yang ada pada teknik klasifikasidalam data mining. Hasil penelitian dengan menggunakan aplikasi Rapid Miner menunjukan bahwa Algoritma K-Nearest Neighbor (K-NN) memiliki nilai akurasi yang lebih baik

Kata Kunci : Analisa Kredit, Algoritma K-NN, Metode Desicion Tree, Rapid Miner, Data Mining

1. PENDAHULUAN

Kredit merupakan suatu fasilitas yang memungkinkan seseorang atau badan usaha meminjam uang untuk membeli produk dan membayarnya kembali dalam jangka waktu yang ditentukan. Berdasarkan UU No.10 tahun 1998 menjelaskan bahwa kredit adalah penyediaan uang atau tagihan yang dapat dipersamakan dengan itu, sesuai persetujuan atau kesepakatan pinjam meminjam antara bank dengan pihak lain yang mewajibkan pihak peminjam untuk melunasi utangnya setelah jangka waktu tertentu dengan pemberian bunga.

Pemberian Kredit mempunyai unsur-unsur yang harus disepakati oleh pihak yang terlibat dalam kredit tersebut yang meliputi kepercayaan, waktu, risiko, dan prestasi. Tujuan pemberian kredit pada umumnya adalah mencari kentungan berbentuk imbalan atau bagi hasil.

Algoitma *K-nearest neighbor* (K-NN) merupakan penelitian menggunakan metode dengan mencari kedekatan antara kriteria kasus baru dengan beberapa kriteria kasus lama berdasarkan criteria kasus yang paling mendekati. Algoitma *K-nearest neighbor* (K-NN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data latih yang jaraknya paling dekat dengan objek tersebut. Ketepatan algoritma K-NN ini sangat di pengaruhi oleh ada atau tidaknya kriteria-kriteria yang tidak relevan, atau jika bobot kriteria tersebut tidak setara dengan relevansinya terhadap klasifikasi.

Metode *Decision tree* merupakan metode yang ada pada teknik klasifikasi dalam data mining. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon pohon keputusan yang mempresentasikan aturan. Pohon keputusan juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara jumlah calon variable input dengan sebuah variabel target.

2. METODOLOGI PENELITIAN

Metode penelitian yang digunakan dalam penelitian ini menggunakan model *Cross-Standard Industry for Data Mining* (CRISP-DM) yang terdiri dari 6 fase, yaitu: *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, *Deployment*

Data yang digunakan dalam penelitian ini berupa data primer. Dimana data tersebut diperoleh dari hasil transaksi kredit baik yang bermasalah maupun yang tidak bermasalah di BPR Prima Sukses Sejahtera.

Variabel-variabel penelitian yang memiliki pengaruh besar dalam penentuan keputusan adalah variabel tak bebas yakni kelancaran, sedangkan variabel bebas meliputi penghasilan, pinjaman, jangka waktu dan jaminan.

Data yang digunakan dalam penelitian ini berupa data primer. Dimana data tersebut diperoleh dari hasil kuisioner yang disebarkan kepada pelanggan PT Trigatra Komunikatama. Obyek dari penelitian ini adalah hal-hal yang berkaitan dengan pelayanan kepada pelanggan.

Setelah data yang diperlukan diperoleh, kemudian ditentukan variabel-variabel yang akan digunakan dalam penelitian. Variabel respon yang digunakan dalam penelitian ini adalah kepuasan, yaitu pelanggan yang puas dan tidak puas terhadap pelayanan PT. Trigatra Komunikatama. Sedangkan variabel bebas yang digunakan antara lain Realibility, Responsiveness, Assurance, Empathy, Tangibles.

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.

K-Nearest Neighbor (KNN)

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..

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

$$D(A,B) = \sqrt{\sum_{k=1}^{d}(\mathrm{Ak} - \mathrm{Bk})^2} \qquad (1)$$

Algoritma C4.5

Algoritma C4.5 dan pohon keputusan merupakan dua model yang tak terpisahkan, karena untuk membangun sebuah pohon keputusan, dibutuhan algoritma C4.5. Di akhir tahun 1970 hingga di awal tahun 1980-an, J. Ross Quinlan seorang peneliti di bidang mesin pembelajaran mengembangkan sebuah model pohon keputusan yang dinamakan ID3 (Iterative Dichotomiser), walaupun sebenarnya proyek ini telah dibuat sebelumnya oleh E.B. Hunt, J. Marin, dan P.T. Stone. lalu Quinlan membuat suatu algoritma dari pengembangan ID3 yang diberi nama C4.5 yang dasar *supervised learning*.

Decision tree adalah metode pada teknik klasifikasi dalam data mining. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon pohon keputusan yang mempresentasikan aturan. Pohon keputusan dapat digunakan untuk mengeksplorasi data yaitu dengan menemukan hubungan tersembunyi antara jumlah calon variable input dengan sebuah variabel target

Gambar 16. Model Pohon Keputusan

K-Nearest Neighbor

Rule based atau algoritma berbasis aturan merupakan cara terbaik untuk merepresentasikan sejumlah bit data atau pengetahuan (Han & Kamber, 2006). Rule based biasanya dituliskan dalam bentuk logika *IF-THEN* atau jika dibuat persamaannya yaitu :

IF condition THEN conlusion
contoh sebuah rule yaitu :
IF age=youth AND student=yes THEN buys_computer=yes

Pernyataan IF dari persamaan di atas dikenal sebagai rule antecedent atau precondition sedangkan pernyataan THEN disebut sebagai rule consequent. Dalam rule antecedent biasanya menyertakan satu atau lebih atribut (misalnya atribut age dan student) dan menggunakan logika AND jika menggunakan lebih dari satu atribut. Rule consequent merupakan prediksi kelas, dalam contoh di atas prediksinya yaitu membeli komputer atau buys_computer=yes

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.

Confusion matrix

Confusion matrix adalah suatu metode yang biasanya digunakan untuk melakukan perhitungan akurasi pada konsep data mining. Rumus ini melakukan perhitungan dengan 4 keluaran, yaitu: *recall, precision, acuraccy* dan *error rate*. *Recall* adalah *proporsi* kasus positif yang diidentifikasi dengan benar. *Precision* adalah *proporsi* kasus dengan hasil positif yang benar. *Accuracy* adalah perbandingan kasus yang diidentifikasi benar dengan jumlah semua kasus. *Error Rate* adalah kasus yang diidentifikasi salah dengan sejumlah semua kasus.

Keakuratan hasil *klasifikasi* dapat diukur dengan menggunakan *confusion matrix*. *Confusion matrix* adalah media yang berguna untuk menganalisis seberapa baik *Class*ifier dapat mengenali tupel dari kelas yang berbeda. Misalkan terdapat dua kelas, maka akan diistilahkan sebagai *tupel positif* dan *tupel negative*. True positif mengacu pada *tupel positif* yang diberi label dengan tepat oleh *Class*ifier, sementara true negative adalah *tupel positif* yang diberi label dengan tepat oleh *Class*ifier. *False positif* adalah *tupel negative* yang diberi label tidak tepat, false negatif adalah *tupel positif* yang diberi label dengan tidak tepat. Istilah ini berguna untuk menganalisis kemampuan *Class*ifier dan diringkas seperti table berikut.

Tabel 1. Model *Confusion matrix*

	C1	C2
C	*true positives*	*false negatives*
C	*false positives*	*true negatives*

Misalkan terdapat *Confusion matrix* 2x2 seperti pada table, maka rumus yang akan digunakan untuk menghitung akurasi adalah sebagai berikut:

$$Akurasi = \frac{True\ positif + true\ negatif}{True\ positif + true\ negatif + false\ positif + false\ negatif} \text{ x } 100\% \quad \ldots(2)$$

3. HASIL DAN PEMBAHASAN

Business Understanding

Total konsumen yang mendapatkan pembiayaan kredit mobil sebanyak 110 orang, 34 diantaranya bermasalah dalam pembayaran angsuran. Ini merupakan permasalahan yang terjadi pada *leasing* diakibatkan oleh analisa yang kurang akurat. Sampai saat ini belum diketahui pula algoritma yang paling akurat dalam melakukan penentuan kelayakan pemberian kredit bagi konsumen. Untuk itu maka dalam penelitian ini akan dilakukan komparasi algoritma K-Nearest Neighbor dan *Decision tree.*

Data Preprocessing

Data yang diperoleh untuk penelitian ini sebanyak 110 *record* transaksi kredit baik yang bermasalah maupun yang tidak bermasalah. Untuk mendapatkan data yang berkualitas, beberapa teknik *preprocessing* digunakan, yaitu:

a. *data validation*, untuk mengidentifikasi dan menghapus data yang ganjil (*outlier/noise*), data yang tidak konsisten, dan data yang tidak lengkap (*missing value*)
b. *data integration and transformation, untuk meningkatkan akurasi dan efisiensi algoritma. Data yang digunakan dalam penulisan ini bernilai kategorikal. untuk model neural network, data ditransformasi ke dalam angka menggunakan software RapidMiner.*
c. *data size reduction and dicretization*, untuk memperoleh data set dengan jumlah atribut dan *record* yang lebih sedikit tetapi bersifat informatif. Di dalam data *training* yang digunakan dalam penelitian ini, dilakukan seleksi atribut dan penghapusan data duplikasi menggunakan *software* RapidMiner.

no	namaz	alamat	status perkawinan	jumlah tanggungan	pendidikan	usia	kepemilikan rumah	lama tinggal	kondisi rumah	jenis pekerjaan	status perusahaan	status ke
1	NOVI ANGGRAINI	KECAMATAN KEMBANGAN	menikah	2	S1	27	keluarga	27	permanen	wirawasta	PERORANGAN	PEMILIK
2	WAHYUDI	KECAMATAN CILODONG	belum menikah	0	S1	23	orang tua	23	non permanen	Karyawan	SWASTA KECIL	KONTRAK
3	ELPA YUNAEDI	KECAMATAN CINERE	menikah	2	SMA	27	orang tua	27	non permanen	Karyawan	SWASTA KECIL	KONTRAK
4	MUHAMMAD FERRY	KECAMATAN BOJONG GEDE	menikah	3	S1	30	keluarga	30	permanen	Karyawan	SWASTA MENENGAH	TETAP
5	WAHYUDI MARUUK	KECAMATAN KEMANG	menikah	3	S2	49	milik sendiri	19	permanen	Guru	LEMBAGA PENDIDIKAN	TETAP
6	HARIANTO	KECAMATAN CIPAYUNG	belum menikah	0	SMA	26	keluarga	27	permanen	wirawasta	SWASTA MENENGAH	PEMILIK
7	ALI PANGESTU	KECAMATAN SUKMAJAYA	belum menikah	0	SMA	26	orang tua	26	non permanen	Karyawan	SWASTA KECIL	KONTRAK
8	AHMAD BASRI	KECAMATAN CINERE	belum menikah	0	S1	25	orang tua	25	permanen	Karyawan	SWASTA KECIL	KONTRAK
9	EDI MULYONO	KECAMATAN CIPAYUNG	belum menikah	0	S1	29	orang tua	29	permanen	Karyawan	SWASTA KECIL	KONTRAK

Gambar 2 Data Sampel

Teknik Analisis

a. Analisa Perhitungan Algoritma C4.5
   Menentukan nilai keputusan ya, keputusan tidak dan *Entropy* dari suatu kasus berdasarkan atribut *Windy, humidity, outlook* dan *temperature*. Kemudian, lakukan penyeleksian atribut dengan menghitung Gain tertinggi.
   1) Hitung nilai Gain masing-masing atribut
   2) menghitung nilai Gain pada baris Status Perkawinan

3) menghitung nilai Gain pada baris Pendidikan
4) menghitung nilai Gain pada baris Pendidikan
5) menghitung nilai Gain pada baris usia
6) menghitung nilai Gain pada baris Kepemilikan Rumah
7) menghitung nilai Gain pada baris Lama Tinggal
8) menghitung nilai Gain pada baris kondisi rumah
9) menghitung nilai Gain pada baris jenis pekerjaan
10) menghitung nilai Gain pada baris Status perusahaan
11) menghitung nilai Gain pada baris Status kepegawaian
12) menghitung nilai Gain pada baris masa kerja
13) menghitung nilai Gain pada baris penghasilan perbulan
14) menghitung nilai Gain pada baris pembayaran pertama

b. Analisa Perhitungan Algoritma *K-Nearest Neighbor*

Untuk mengukur jarak antar atribut maka akan diberikan bobot pada masing-masing 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 bepengaruh.

Pebobotan nilai atribut dilakukan untuk 13 atribut rediktor. Setelah itu dihitung kemiripannya. Misalkan sebuah data konsumen baru akan diklasifikasi apakah bermasalah atau tidak dalam pembayaran angsuran mobil, maka akan dilakukan perhitungan dengan kedekatan antara kasuss baru dengan data kasus lama (data training).

Setelah dihitung nilai kedekatannya dari nilai tersebut diketahui bahwa nilai tertinggi adalah kasus nomor 1. Dengan demikian kasus yang terdekat dengan kasus baru adalah kasus nomor 1, jadi kemungkinan calon kreditur baru tersebut tidak akan bermasalah dalam pembayaran angsurannya.

Pengujian Model

Model yang telah dibentuk kemudian diuji tingkat akurasinya dengan memasukan data uji yang berasal dari data training. Karena data yang didapat dalam penelitian ini setelah proses preprocessing hanya 110 data maka digunakan metode cross validation untuk menguji tingkt akurasi.

Dengan metode *K-NN*, menghasilkan kondisi seperti pada Tabel 4.17 Diketahui dari 110 data, 75 data diklasifikasikan *lancar* sesuai dengan prediksi yang dilakukan dengan metode K-NN, lalu 1 data diprediksi lancar tetapi ternyata *tidak lancar*, 33 data tidak lancar diprediksi sesuai, dan 1 data diprediksi tidak *lancar* ternyata *lancar*.

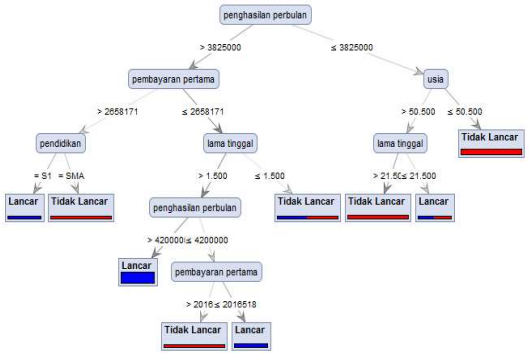
Table 2. Confusion Matrix Algoritma K-NN

Akurasi: 98.18 % +/- 3,64 (mikro 98,18%)			
	True lancer	True tidak lancar	Class precision
Prediksi lancer	75	1	98,68%
Prediksi tidak lancer	1	33	97,06%
Class recall	98,68%	97,06%	

Dengan metode *Decision tree*, menghasilkan kondisi seperti pada Tabel 4.18 Diketahui dari 110 data, 73 data diklasifikasikan *lancar* sesuai dengan prediksi yang dilakukan dengan metode K-NN, lalu 7 data diprediksi lancar tetapi ternyata *tidak lancar*, 27 data tidak lancar diprediksi sesuai, dan 3 data diprediksi tidak *lancar* ternyata *lancar*.

Table 3. Confusion Matrix Decision tree

Akurasi: 90,91 % +/- 5,75 (mikro 90,91%)			
	True lancer	True tidak lancar	Class precision
Prediksi lancer	73	7	91,25%
Prediksi tidak lancer	3	27	90,00%
Class recall	96,06%	79,41%	



Gambar 3. *Decision Tree*

4. KESIMPULAN

a. Analisa kredit menggunakan Perbandingan Algoritma K-Nearest Neighbor dan Metode Decison Tree dapat meningkatkan tingkat ketelitian bagi Kredit Pemilikan Mobil (KPM) dalam menyeleksi konsumen baru yang akan melakukan kredit mobil.

b. Pada hasil penelitian menunjukan penggunaan Algoritma K-Nearest Neighbor lebih akurat dalam penentuan kelayakan konsumen dengan nilai akurasi 98,18%. Nilai keakuratan ini bertujuan untuk menghindari terjadinya kredit macet.

DAFTAR PUSTAKA

Siamat, Dahlan. (2005) Manajemen Lembaga Keuangan: Kebijakan Moneter dan Perbankan. Jakarta: Fakultas Ekonomi UI

Rapid-I GmbH. (2010).*Rapid Miner User Manual*. Dortmund: Rapid-I GmbH

Sogala, Satchidananda S. *Comparing the Efficacy of the Decision Trees with LogisticRegression for Credit Risk Analysis*. India

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

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

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

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

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

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