p-ISSN 2460-9986
e-ISSN 2476-9436

KLASIFIKASI *DATA MINING* DALAM MENENTUKAN PEMBERIAN KREDIT BAGI NASABAH KOPERASI

Ika Menarianti

Fakultas Matematika Ilmu Pengetahuan Alam dan Teknologi Informasi, Jurusan Pendidikan Teknologi Informasi, Universitas PGRI Semarang, Jl. Dr. Cipto – Lontar No. 1 Semarang; Telp.024-8451279.
Email: kmnrt201086@gmail.com

Abstrak

Kredit adalah penyediaan dana untuk transaksi pinjam meminjam atas persetujuan dan kesepakatan antara pihak bank atau instansi keuangan dengan nasabahnya, serta mewajibkan peminjam untuk membayar utang dalam jangka waktu tertentu dan pemberian jasa. Pemberian kredit dilakukan dengan mengidentifikasi dan menilai faktor yang mempengaruhi resiko kredit. Hilangnya pendapatan dan ancaman profitabilitas merupakan hal yang perlu diwaspadai dalam pemberian kredit. Klasifikasi *data mining* dapat digunakan untuk membantu para analis kredit dalam menentukan pemberian kredit pada nasabah. Proses klasifikasi dilakukan untuk mendapatkan atribut penentu. Hasil proses klasifikasi dievaluasi menggunakan *cross validation, confusion matrix, ROC Curve* dan *T-test* untuk mengetahui klasifikasi yang paling akurat dalam menentukan pemberian kredit bagi nasabah koperasi.

Kata kunci: klasifikasi *data mining, cross validation, confussion matrix*, *ROC Curve, T-test*

Abstract

Credit is the provision of funds for lending and borrowing transactions with the consent and agreement between the bank or financial institution and its customers, as well as requiring the borrower to repay the debt within a certain period and the provision of services. Lending is done by identifying and assessing the factors affecting the credit risk. Loss of income and profitability threat is a things that have to be awared in the provision of credit. Classification of data mining can be used to assist in determining a credit analyst lending to customers. The classification process is done to get a decisive attribute. The results of the classification process was evaluated using cross validation, confusion matrix, ROC Curve and T-test to determine the classification of the most accurate in determining the provision for customer credit cooperatives

Keywords: *Classification of data mining, cross validation, confusion matrix, ROC curve, T-test*

1. PENDAHULUAN

Perbankan dan instansi keuangan memiliki peranan yang strategis dalam pembangunan nasional. Badan usaha yang menghimpun dana dari masyarakat dalam bentuk simpanan dan menyalurkannya kepada masyarakat dalam bentuk kredit atau bentuk-bentuk lainnya dalam rangka meningkatkan taraf hidup masyarakat banyak di bahas pada Undang-Undang Perbankan No.10 tahun 1998.

Menurut Pasal 1 angka 11 Undang-Undang Nomor 10 Tahun 1998, kredit adalah penyediaan uang atau tagihan yang dapat dipersamakan dengan itu, berdasarkan persetujuan atau kesepakatan pinjam meminjam antara bank atau instansi keuangan dengan pihak lain yang mewajibkan pihak peminjam untuk melunasi utangnya setelah jangka waktu tertentu dengan jumlah bunga.

Peraturan dan regulasi perbankan berubah dengan diterapkannya *internet banking*. Semua data nasabah yang terkait pinjaman baik lancar maupun bermasalah dapat dilihat serta diverifikasi melalaui *internet banking*. Sehingga resiko kredit dapat ditekan. Sebaliknya instansi keuangan seperti koperasi di luar perbankan belum memiliki pusat data, hal ini dapat meningkatkan resiko kredit yang mengancam profitabilitas. Koperasi adalah merupakan usaha kekeluargaan dengan tujuan untuk mensejahterakan anggotanya (UUD 1945 pasal 33 ayat 1).

Koperasi memiliki kebijakan yang berbeda-beda dalam pemberian kredit. Tetapi pada umumnya pemberian kredit dipengaruhi beberapa faktor seperti kepercayaan, kesepakatan, jangka waktu, risiko dan balas jasa (Kasmir, 2010). Analis kredit perlu mengidentifikasi dan menilai faktor-faktor yang dapat mempengaruhi nasabah dalam pengembalian kredit (Costa et al., 2007).

Pengukuran yang akurat dan kemampuan manajemen yang baik dalam menghadapi risiko kredit merupakan upaya penyelamatan unit operasi ekonomi dan bermanfaat untuk sistem keuangan yang stabil dan sehat secara keseluruhan dan pembangunan ekonomi yang berkesinambungan (Ma & Guo, 2010). Kegagalan mengidentifikasi risiko kredit mengarah pada hilangnya pendapatan dan memperluas kredit untuk risiko kredit yang bertipe buruk adalah ancaman bagi profitabilitas (Zurada & Kunene, 2011)

p-ISSN 2460-9986
e-ISSN 2476-9436

Kesalahan analisa kredit dapat menyebabkan risiko kredit, seperti menghilangnya nasabah, ketidakpastian pembayaran dana pinjaman bahkan ketidakmampuan nasabah dalam mengembalikan pinjaman dana kredit. Untuk melindungi dana kredit, digunakan jaminan yang harus disediakan oleh pihak nasabah sebagai beban nasabah. Pemberian kredit dengan jaminan dapat berupa: jaminan benda berwujud (tanah, bangunan, kendaraan bermotor, kebun, perhiasan dan lain-lain), jaminan tidak berwujud (sertifikat tanah, sertifikat saham, sertifikat obligasi, SK pengangkatan kerja dan lain-lain) dan jaminan orang (jaminan yang diberikan oleh seseorang yang menyatakan kesanggupan untuk menanggung segala resiko apabila kredit tersebut macet).

Kriteria penilaian kredit seperti sifat atau watak seseorang, kemampuan membayar, penggunaan dana, kondisi sosial, ekonomi dan politik serta jaminan yang diajukan diperlukan untuk memberikan informasi mengenai itikad baik dan kemampuan membayar seorang nasabah (Kasmir, 2010). Komponen yang mempengaruhi risiko kredit, adalah kemungkinan debitur akan gagal membayar dalam memenuhi kontrak pembayaran, klaim yang akan ditanggung debitur jika tidak memenuhi kewajiban membayar dan nominal yang hilang akibat risiko default atau gagal bayar.

Teknik klasifikasi *data mining* dapat digunakan untuk menentukan risiko kredit. *Data Mining* adalah kegiatan yang meliputi pengumpulan dan pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data yang berukuran besar (Santoso, 2007). Keluaran yang dihasilkan oleh klasifikasi *data mining* dapat digunakan untuk memperbaiki pengambilan keputusan bagi analis kredit dalam pemberian kredit.

Pemilihan algoritma klasifikasi *data mining* untuk menentukan resiko kredit yang terjadi pada transaksi peminjaman berdasarkan beberapa penelitian sebelumnya. Teknik klasifikasi *data mining* dalam menentukan peningkatan kualitas kredit dan penurunan resiko kredit dengan menggunakan *Logistic Regression, Discriminant Analysis, K-Nearest Neighboor, TAN Technique, Naive Bayes, Decission Tree (C45), Associative Classification*, *Neural Network* dan *Support Vector Machine* (Yu et al., 2007).

Klasifikasi *data mining* dalam memeriksa dan mengkomparasi 4 teknik *data mining* pada dua set data kredit untuk menghasilkan dua keluaran yaitu "*good customer*" dan "*bad customer*". Klasifikasi yang di gunakan adalah *Logistic Regression, Decission Tree, Support Vector Machine* dan *Neural Network* (Yu et al., 2010).

Komparasi dalam menentukan metode yang paling baik performanya dalam mendeteksi resiko kredit. Klasifikasi yang digunakan adalah *Bayesian Network, Naive Bayes, Support Vector Machine, Linier Logistic Regression, K-Nearest Neighboor, C45, RIPPER* dan *RBF* (Peng & Kou, 2008).

Tingkat akurasi klasifikasi untuk menentukan pemberian kredit dengan membandingkan klasifikasi *Logistic Regression, Neural Network, RBFNN, Support Vector Machine, K-Nearest Neighbor* dan *Decision Tree* (Zurada & Kunene, 2011).

2. METODE

Penelitian ini menggunakan penelitian eksperimen dengan tahapan penelitian: pengumpulan data, pengolahan awal data, metode yang digunakan, eksperimen dan pengujian model serta evaluasi dan validasi hasil klasifikasi.

2.1. Pengumpulan Data

Penentuan jenis dan sumber data untuk memperoleh data yang benar-benar akurat merupakan hal yang sangat penting. Sumber data pada penelitian ini adalah data kredit yang diambil dari Koperasi Borobudur Agung pada tahun-tahun sebelumnya sebagai acuan untuk menemukan pola-pola tertentu yang bisa dijadikan atribut penentu. Data yang dapat digunakan dalam penelitian ini adalah data agunan, data pinjaman dan data piutang lancar.

Tabel 1. Data Agunan

No. Agunan	Tgl. Masuk	Nama Nasabah	Barang Jaminan
285	02-Jan-08	Mei Wulandari	BPKB Honda H5521VW
286	08-Jan-08	Chaprista RH	BPKB Honda Civic H7619DC
287	09-Jan-08	Sukardi	BPKB Suzuki RC 110 H5115YY
291	14-Jan-08	Sri Wahyuni	BPKB Starlet AD8310DG
293	18-Jan-08	Prilia Sukawati	BPKB Honda H6348EY
294	24-Jan-08	Sarmin	HM No.262 Luas 107m2 Gayamsari
....			

Tabel 1 berisi nama nasabah dan barang yang dijadikan sebagai jaminan. Tabel 2 berisi nama nasabah, tanggal mulai peminjaman, jumlah yang dipinjam, nilai tunggakan dan keterangan sudah jatuh tempo atau lebih dari jatuh tempo.

p-ISSN 2460-9986
e-ISSN 2476-9436

Tabel 2. Data piutang lancar

Nama	Tgl. pinjam	Jmlh pnjm	Tunggak	Ket
Suyanto	26-Apr-05	5.000.000	800.000	JT
Usman Rais	22-Jul-06	5.000.000	3.390.000	JT
Sri Murhin	19-Jan-07	41.500.000	31.000.000	JT
Sunariah	31-Mei-07	9.000.000	4.250.000	JT
Purwanto	18-Ags-07	4.000.000	667.500	3bln
Fitria R	21-Ags-07	5.000.000	1.042.500	5bln
Nuriyah	22-Sep-07	5.000.000	1.100.000	4bln
.....				

Tabel 3 berisi nama nasabah, penanggung jawab, jumlah pinjaman, iuran pokok yang harus dibayarkan, jumlah angsuran, jasa (bunga) yang harus dibayarkan dan jatuh tempo pembayaran

Tabel 3. Data pinjaman nasabah

Nama	Pen. Jwb	Jmlh pjm*	Pkk*	x	Jasa*	JT
Suharti	**Pengelola**	**2.500**	**208**	**12**	**55**	**Feb**
Sukadiyo	**Hartono**	**4.000**	**166**	**24**	**88**	**Feb**
Sugiyarto	**Pengelola**	**5.000**	**208**	**24**	**100**	**Feb**
Tatang S	**Pengelola**	**3.500**	**350**	**10**	**77**	**Des**
Agus B	**Agus B**	**50.000**	**4.166**	**12**	**1.100**	**Feb**
Mujiono	**Pengelola**	**2.500**	**1.240**	**10**	**55**	**Des**
Fitria R	**Pengelola**	**5.000**	**208**	**24**	**100**	**Jan**
.....						

*) *dalam ribuan*

Pengambilan data dilakukan dengan melihat sistem yang berjalan pada Koperasi Borobudur Agung seperti Gambar 1.

2.2. Pengolahan Awal Data

Proses pengolahan awal data diperlukan untuk menyiapkan data yang benar-benar valid sebelum diproses. Pengolahan dilakukan dengan membersihkan data yang ganda, menyamakan batasan data, pengelompokan data, melakukan seleksi fitur dan *pre-procesing* data (Gorunescu, 2011).

2.2.1. Integrasi Data

Data yang dapat digunakan dalam proses penentuan kredit adalah data piutang lancar, data agunan dan data pinjaman. Integrasi data adalah cara menggabungkan beberapa data dari tabel yang berbeda dengan melihat kesamaan data berdasarkan atribut kunci (*primary key*), atribut tamu (*foreign k*ey) hingga melihat ketergantungan fungsionalnya (*functional dependency*). Integrasi data diperlukan karena perlu dilakukan seleksi fitur untuk mendapatkan pola yang merujuk pada hasil pemberian kredit.

2.2.2. Seleksi fitur (atribut)

Seleksi fitur dilakukan dengan mengambil sebagian variabel pada seluruh atribut yang ada untuk dijadikan atribut penentu dalam melakukan pemberian keputusan. Fitur yang diambil adalah atribut yang memiliki sifat ketergantungan fungsional dan merupakan bagian dari *super key*. Berikut merupakan hasil seleksi fitur:

Tabel 4. Seleksi fitur

Atribut	Nilai	Kategori
Jenis Kelamin	1	Laki-laki
	2	Perempuan
Agunan	1	Motor
	2	Mobil
	3	Bangunan
	4	Jamsostek
	5	Tidak Ada
Penanggung Jawab	1	Pengelola
	2	Anggota
Jumlah Pinjam	1	<=5.000.000
	2	<=15.000.000
	3	>15.000.000
Jangka Waktu	1	Pendek (<=6 bulan)
	2	Menengah(<=12 bln)
	3	Panjang (>12 bulan)
Status Kredit	1	Lancar
	2	Bermasalah

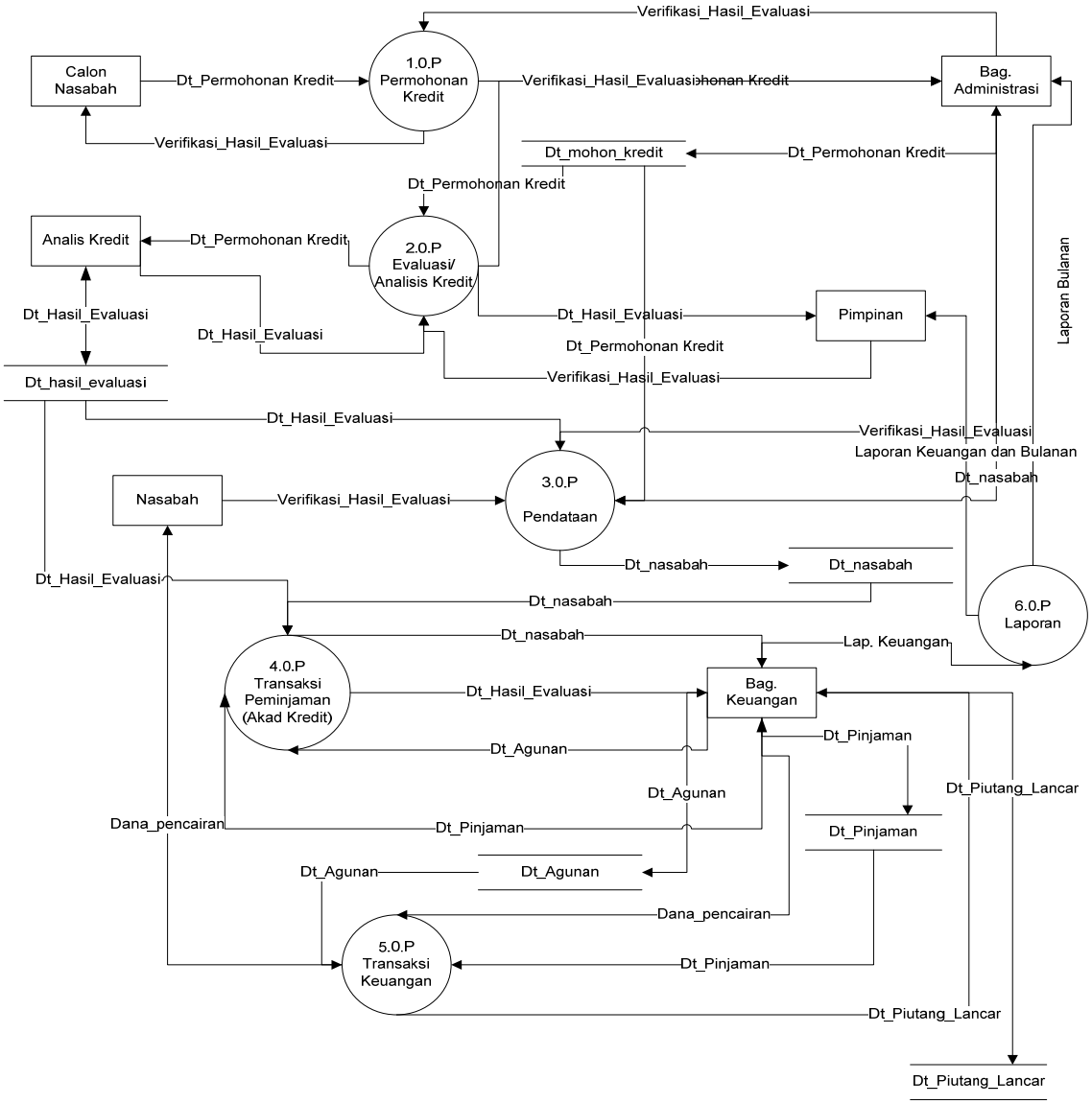
2.2.3. Data cleansing

Proses *cleansing* merupakan tahapan yang penting, dimana data dibersihkan dari data yang tidak diperlukan (seperti: no.anggota, nama, alamat) dan menghapus data yang sama (*redudancy*). Hal ini dimaksudkan untuk menjaga nilai ketergantungan fungsionalnya.

2.2.4. Transformasi data

Pada proses transformasi data, data dikelompokkan berdasarkan kriteria yang sama untuk mempermudah pengolahan data selanjutnya, yang dapat dilihat pada Tabel 5.

p-ISSN 2460-9986
e-ISSN 2476-9436



Gambar 1. *Data Flow Diagram* sistem yang berjalan pada Koperasi Borobudur Agung

Tabel 5. Hasil proses pre-processing data

Jns_Klmn	Agunan	Pen. Jawab	Jml.Pinjam	Jangka Waktu	Status kredit
Perempuan	Motor	Pengelola	2.500.000	Menengah	Lancar
Laki-laki	Mobil	Pengelola	5.000.000	Menengah	Lancar
Perempuan	Bangunan	Pengelola	6.000.000	Menengah	Lancar
Laki-laki	Motor	Pengelola	1.500.000	pendek	Lancar
Perempuan	Motor	Pengelola	5.000.000	Menengah	Bermasalah
Laki-laki	Motor	Anggota	15.000.000	pendek	Lancar
Laki-laki	Bangunan	Anggota	5.000.000	Menengah	Lancar
Perempuan	Motor	Anggota	2.500.000	Menengah	Bermasalah
Laki-laki	Motor	Pengelola	3.000.000	Menengah	Lancar

p-ISSN 2460-9986
e-ISSN 2476-9436

Laki-laki	Motor	Pengelola	3.000.000	Menengah	Lancar
Perempuan	Motor	Pengelola	1.500.000	Menengah	Bermasalah
Laki-laki	Bangunan	Pengelola	10.000.000	Panjang	Bermasalah
.....					

2.3. Metode yang digunakan

Metode yang digunakan dalam penelitian ini adalah *cross validation, confussion matrix, ROC curve* dan *T-Test*. Hal ini dilakukan untuk melihat sejauh mana perbedaan data setelah dan sebelum dilakukan *pre-processing* data. Untuk menentukan klasifikasi yang digunakan pada suatu masalah diperlukan cara sistematis untuk mengevaluasi bagaimana metode yang bekerja dan membandingkannya dengan yang lain. Klasifikasi *data mining* yang digunakan adalah *Logistic Regression, Discriminant Analys, K-Nearest Neighbour, Naive Bayes, Decision Tree, Neural Network* dan *Support Vector Machine*.

Evaluasi klasifikasi didasarkan pada pengujian pada obyek benar dan salah (Gorunescu, 2011). Validasi data digunakan untuk menentukan jenis terbaik dari skema belajar yang digunakan, berdasarkan data pelatihan untuk melatih skema pembelajaran untuk memaksimalkan penggunaan data (Witten et al., 2011).

2.3.1. *Cross* Validation

Setiap kelas pada kelompok data harus diwakili dalam proporsi yang tepat antara data *training* dan data *testing*. Data dibagi secara acak pada masing-masing kelas dengan perbandingan yang sama. Untuk mengurangi bias yang disebabkan oleh sampel tertentu, seluruh proses *training* dan *testing* diulangi beberapa kali dengan sampel yang berbeda. Tingkat kesalahan pada iterasi yang berbeda akan dihitung rata-ratanya untuk menghasilkan *error rate* secara keseluruhan. Model yang memberikan rata-rata kesalahan terkecil adalah model yang terbaik.

2.3.2. *Confusion Matrix*

Confussion matrix melakukan pengujian untuk memperkirakan obyek yang benar dan salah (Gorunescu, 2011). Urutan pengujian ditabulasikan dalam *confusion matrix* dimana kelas yang diprediksi ditampilkan di bagian atas matriks dan kelas yang diamati di bagian kiri. Setiap sel berisi angka yang menunjukkan berapa banyak kasus yang sebenarnya dari kelas yang diamati untuk diprediksi.

Tabel 6. Model *confussion matrix*

	Nilai Aktual	
Nilai Prediksi	TP	FN
	FP	TN

Keterangan :

TP = tupel postif yang diklasifikasikan positif.

TN = tupel negatif yang diklasifikasikan negatif.

FP = tupel positif yang diklasifikasikan negatif.

FN = tupel negatif yang diklasifikasikan positif.

Untuk menghitung tingkat akurasi pada matriks digunakan:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Sensitivitas dan spesifisitas tidak memberikan informasi untuk nilai diagnosa yang benar. Maka perlu adanya PPV (nilai prediksi positif) dimana proporsi kasus dengan hasil tes "positif" adalah:

$$PPV = \frac{TP}{TP + FP} \quad (2)$$

dan membutuhkan NPV (nilai prediksi negatif) dengan proporsi kasus dengan hasil tes "negatif" yang dituliskan pada persamaan 3.

$$NPV = \frac{TN}{TN + FN} \quad (3)$$

Tingkat kesalahan diperoleh dari persamaan 4.

$$Tingkat\ kesalahan = \frac{FN}{positif + negatif} \quad (4)$$

Keterangan:

positif = TP + FN

negatif = FP + TN.

2.3.3. *ROC Curve*

ROC curve banyak digunakan dalam penelitian *data mining* dalam menilai hasil prediksi (Gorunescu, 2011). Secara teknis *ROC curve* dibagi dalam dua dimensi, dimana tingkat TP di letakkan pada sumbu Y dan tingkat FP di letakkan pada sumbu X. Tetapi untuk merepresentasikan grafis yang menentukan klasifikasi mana yang lebih baik, digunakan metode yang menghitung luas daerah dibawah ROC yang disebut AUC (Area Under the ROC Curve) yang diartikan sebagai probabilitas (Witten et al., 2011).

AUC mengukur kinerja diskriminatif dengan memperkirakan probabilitas *output* dari sampel yang dipilih secara acak dari populasi positif atau negatif. Semakin besar AUC, semakin kuat klasifikasi yang digunakan (Yu et al., 2007).

Panduan tingkat keakuratan klasifikasi dengan menggunakan AUC:

- 0,90 – 1,00 = klasifikasi yang baik
- 0,80 – 0,90 = klasifikasi yang baik
- 0,70 – 0,80 = klasifikasi yang adil atau sama
- 0,60 – 0,70 = klasifikasi rendah
- 0,50 – 0,60 = kegagalan

2.3.4. *T-Test*

T-test termasuk kedalam metode statistik yang digunakan untuk mempelajari pengambilan keputusan parameter populasi dari sampel yang ada. Dalam sebuah kegiatannya ada dua hal dasar yaitu adanya data yang berasal dari sampel dan adanya perlakuan dengan tujuan tertentu terhadap sampel. Dalam hal ini adalah melakukan pengujian atas perbedaan antara sebelum dan sesudah dilakukan suatu aksi. *T-Test* adalah suatu metode pengujian hipotesis dengan menggunakan satu individu (objek penelitian) dikenai dua perlakuan yang berbeda.

2.3.5. Logistic *Regression*

Logistic Regression adalah variasi regresi yang digunakan ketika variabel dependen bersifat biner (Yu et al., 2010). Model ini dapat memprediksi hasil diskrit dari satu kelompok variabel yang mungkin akan berlangsung terus menerus, kategorikal atau keduanya (Keramati & Yousefi, 2011). Tujuan dari model ini adalah untuk mendapatkan persamaan regresi yang dapat memprediksi dua atau lebih kelompok objek yang dapat ditempatkan yaitu apakah pinjaman harus diklasifikasikan sebagai pinjaman yang baik atau pinjaman yang buruk (Santoso, 2007).

Diberikan kelompok sampel dengan jumlah dimensi dan label kelas $y_i \in \{1, 2, \ldots, K\}$. *Logistic Regression* dapat diterapkan ke dalam klasifikasi biner dengan $y \in \{0, 1\}$. Maka probabilitas *posterior* sampel x dapat dihitung:

$$\log it(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \quad (5)$$

$\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ merupakan parameter yang dicari. Perbandingan antara $\left(\frac{p}{1-p}\right)$ disebut dengan *odds ratio.*

2.3.6. *Discriminant Analys*

Discriminant Analysis mencari fungsi diskriminan yang merupakan kombinasi linier dari variabel-variabel, yang memisahkan obyek ke dalam dua kelompok atau kelas. Data dibagi ke dalam variabel x_1= *yield point* dan x_2=*ultimate strength*. Kemudian fungsi diskriminan ditentukan dengan menentukan kombinasi linier:

$$z = \omega' x = \omega_1 x_1 + \omega_2 x_2 \quad (6)$$

dan menentukan vektor observasi $(n_1 + n_2)$ dimana $x_{11}, x_{12}, .., x_{1n}$ dan $x_{21}, x_{22}, .., x_{2n}$. Transformasikan ke besaran skalar $z_{11}, z_{12}, .., z_{1n}$ dan $z_{21}, z_{22}, .., z_{2n}$. Kemudian mencari rata-rata: $\bar{z_1} = \omega' \bar{x}_1$ dan $\bar{z_2} = \omega' \bar{x}_2$. Selanjutnya tentukan jarak kuadrat untuk nilai maksimum:

$$\omega = S_{pl}\left(\overline{x_1} - \overline{x_2}\right) \quad (7)$$

$x_{11}, x_{12}, .., x_{1n}$ dari kelompok populasi satu dan $x_{21}, x_{22}, .., x_{2n}$ dari kelompok populasi dua. Setiap vektor x_{ij} terdiri dari p variabel. Kombinasi linier $z = \omega' x$ mentransformasikan setiap vektor observasi menjadi besaran skalar.

2.3.7. *K-Nearest Neighbour*

Algoritma ini menghasilkan batas klasifikasi non-linier. *K-nearest neighbor* merupakan salah satu metode pengklasifikasian data berdasarkan kesamaan dengan label data (Larose, 2006). Untuk menghitung kesamaan dapat digunakan matriks jarak dimana satuan jaraknya menggunakan satuan Euclidean.

$$d(x, y) = \|x - y\|^2 = \sqrt{\sum_{i=1}^{n}\left(x_i - y_i\right)^2} \quad (8)$$

Matriks d(x,y) adalah jarak skalar dari kedua vektor x dan y dari matriks dengan ukuran d dimensi. Adapun rumus untuk menghitung kedekatan antara dua kasus ditunjukkan pada persamaan 9.

$$similarity(T, S) = \frac{\sum_{i=1}^{n} f\left(t_i, s_i\right) * w_i}{w_i} \quad (9)$$

S merupakan kasus yang ada dalam penyimpanan, T adalah kasus yang baru, n adalah jumlah atribut, i merupakan atribut individu, f adalah fungsi similarity atribut i antara kasus T dan S sedangkan w adalah bobot yang diberikan kepada atribut ke-i. Kedekatan kasus biasanya berada pada nilai antara 0 dan 1. Nilai 0 artinya kedua kasus mutlak tidak memiliki kesamaan dan jika nilai 1 maka kedua kasus tersebut mutlak memiliki kesamaan.

2.3.8. *Naive Bayes*

Klasifikasi *naive bayes* mengasumsikan keberadaan (atau ketidakberadaan) spesifikasi fitur tertentu pada kelas yang tidak terkait dengan keberadaan lain (Keramati & Yousefi, 2011). Untuk mendapatkan nilai probabilitas pada sebuah sampel diberikan sebuah teorema Bayes:

$$P(h|x) = \frac{P(x|h)P(h)}{P(x)} \quad (10)$$

P(h) adalah nilai probabilitas prior dari hipotesa pada sebuah sampel disebut priori. *P*(x) merupakan *evidence* dari probabilitas data pelatihan. P(h|x) adalah nilai probabilitas h yang mempengaruhi x (*posterior density*), sedangkan P(x|h) merupakan probabilitas x kepada h yang disebut *likelihood*.

$$posterior = \frac{likelihood \times prior}{evidence} \quad (11)$$

Kemudian gunakan probabilitas *m-estimasi*

$$\frac{n_c + mp}{n + m} \quad (12)$$

n_c adalah total nilai dari contoh sampel pada atribut yang dimiliki kelas C, n adalah total nilai keseluruhan sampel, m adalah nilai ekivalen yang konstan dari ukuran sampel, p adalah probabilitas prior.

2.3.9. *Decision Tree*

Decision Tree mengklasifikasikan sampel secara *top-down*, mulai dari simpul akar dengan menjaga jarak dengan hasil dari tes *node* internal, sampai simpul daun yang dicapai oleh kelas label yang ditugaskan (Yu et al., 2007). Keuntungan paling signifikan dari pohon keputusan adalah kenyataan bahwa pengetahuan dapat diekstraksi dan direpresentasikan dalam bentuk aturan klasifikasi *if-then* (Yu et al., 2010).

Teori entropi diadopsi untuk memilih pemecahan atribut yang tepat untuk algoritma C4.5, dengan menyatakan jumlah rata-rata informasi yang dibutuhkan untuk mengklasifikasikan sampel. Untuk menghitung nilai entropy digunakan persamaan 13.

$$entropy(S) = \sum_{i=1}^{n} -p_i . log_2 p_i \quad (13)$$

S merupakan himpunan kasus, n adalah jumlah partisi S, dan p_i adalah proporsi S_i terhadap S. Ketika *output* data atau variabel *dependent* S dikelompokkan berdasarkan atribut A, dinotasikan dengan $gain(S, A)$. Hasil dari atribut mendapatkan i*nformation g*ain yang didefinisikan pada persamaan 14.

$$gain(S, A) = entropy(S) - \sum_{i=}^{n} \frac{S_i}{S} . entropy(S_i) \quad (14)$$

S merupakan himpunan kasus, A adalah atribut, n adalah jumlah partisi atribut A, S_i adalah proporsi S_i terhadap S dan S adalah jumlah kasus dalam himpunan. Sebuah prosedur tambahan dilakukan untuk menghindari pohon yang menghasilkan *overfits* data yang kompleks.

2.3.10. *Neural Network*

Model *neural network* digunakan dalam berbagai aplikasi seperti pemetaan non-linier, pengenalan pola, pendekatan fungsi, klasifikasi dan optimasi (Santoso, 2007). Setiap proses dimulai dari *neuron input* yang dikirimkan melalui *neuran* pada lapisan berikutnya untuk membawa *output* ke lapisan *neuron output*. *Neuron* adalah unit pemroses yang sangat vital dalam suatu operasi *neural network*.

Menghitung jumlah n signal input x_{ij}=1,2,...,n yang diberi bobot dan menghasilkan nilai 1 bila jumlah diatas batas tertentu dan 0 bila dibawah batas dapat ditulis pada persamaan 15.

$$y = \varphi\left(\sum_{j}^{n} w_j x_j - u\right) \quad (15)$$

φ(.) adalah fungsi aktivasi dan w adalah bobot sesuai dengan input ke-j. Bias dinyatakan sebagai b yang mempunyai fungsi menaikkan atau menurunkan net input untuk fungsi aktivasi dan neuron dinyatakan dengan k.

$$u_k = \sum_{j=1}^{m} w_k x_j \quad (16)$$

dan

$$y_k = \varphi(u_k + b_k) \quad (17)$$

$x_1,x_2,...,x_m$ adalah signal input dan $w_1,w_2,...,w_m$ adalah bobot dari synapsis k. u_k adalah kombinasi linier dari output yang dihasilkan signal, b_k adalah bias, φ adalah fungsiaktivasi dan y_k adalah signal output dari neuron yang bersangkutan.

Pemakaian bias mempengaruhi *output neuron*.

$$v_k = \sum_{j=0}^{m} w_{kj} x_j \quad (18)$$

dan $y_k = \varphi(v_k)$. Dimana f(*) adalah fungsi aktivasi dan bk adalah bias. Sehingga fungsi aktivasi *sigmoid*:

p-ISSN 2460-9986
e-ISSN 2476-9436

$$f(x) = \frac{1}{1 + e^{-ax}} \quad (19)$$

Tujuan dari proses learning adalah menemukan bobot w dan bias b, sehingga *network* secara tepat menghasilkan *output* {-1,+1} untuk setiap data *training* yang dimasukkan. *Error* adalah selisih antara target yang sebenarnya dan keluaran dari *network* pada unit *output*, dimana E adalah *error training*.

$$E(w) = \frac{1}{2}\sum_{k=1}^{m}(d_k - y_k)^2 \quad (20)$$

2.3.11. *Support Vector Machine*

Support Vector Machine merupakan perpaduan pemodelan linier untuk menangani tugas klasifikasi dalam memecahkan masalah non-linier. Teknik ini berusaha untuk menemukan fungsi pemisah yang optimal yang bisa memisahkan dua kelompok data dari dua kelas yang berbeda. Formulasi masalah optimasi SVM untuk klasifikasi linier di dalam *primal space* adalah:

$$min\frac{1}{2}\|w\|^2 \quad (21)$$

dengan subjek

$$y_i(wx_i + b) \geq 1, i = 1, \ldots, l \quad (22)$$

x_i adalah data input, y_i adalah keluaran data. x_i, w, b adalah parameter-parameter yang kita cari nilainya.

2.4. Eksperimen dan pengujian

Metode pengujian ini mengikuti cara pengukuran dengan mengukur tingkat akurasi dari masing-masing algoritma berdasarkan data set kredit yang dibagi kedalam variabel-variabel penentu keputusan(Yu et al., 2007). Dari hasil *pre-processing* data, terdapat 588 data kredit dengan total data nasabah yang tidak bermasalah sebanyak 514 data dan 74 data nasabah bermasalah dalam keharusan membayar kredit. Berikut merupakan salah satu perhitungan klasifikasi menggunakan algoritma logistic regression.

Tabel 7. Data Pengujian

No	x_1	x_2	x_3	x_4	x_5	Status_kredit
1	1	1	0	1	1	Lancar
2	0	0	0	0	0	Lancar
3	1	0	0	0	1	Bermasalah
4	0	1	1	1	0	Lancar
5	0	1	1	0	1	Lancar
6	1	0	1	0	0	Bermasalah
7	0	0	0	0	0	Lancar
8	0	0	0	0	1	Lancar
9	1	0	0	0	0	Bermasalah
10	0	1	0	1	1	Bermasalah

Dari model diatas akan didapatkan hasil:

$$exp(\beta_0) = \frac{P(y-1|x_1 = x_2 = x_3 = x_4 = x_5 = 0)}{P(y-0|x_1 = x_2 = x_3 = x_4 = x_5 = 0)} = \frac{6}{4} = 1.5$$

$$exp(\beta_1) = \frac{1}{1.5} = 0.67 \qquad exp(\beta_2) = \frac{3}{1.5} = 2$$

$$exp(\beta_3) = \frac{2}{1.5} = 1.33 \qquad exp(\beta_4) = \frac{2}{1.5} = 1.33$$

$$exp(\beta_5) = \frac{3}{1.5} = 2$$

Sehingga persamaan fungsi *logistic regression* bisa dituliskan sebagai:

$$P(y-1|x_1, x_2, \ldots, x_k) = \frac{e^{-1.5+0.67x_1+2x_2+1.33x_3+1.33x_4+2x_5}}{1+e^{-1.5+0.67x_1+2x_2+1.33x_3+1.33x_4+2x_5}} \quad (23)$$

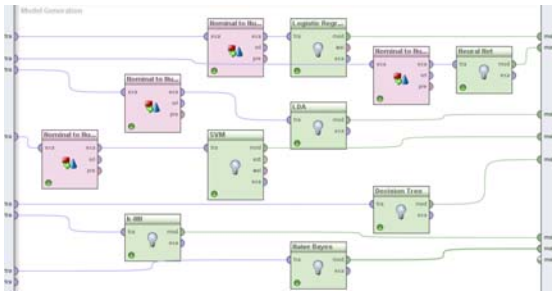
Hasil *knowledge representation* untuk *logistic regression*:

$$\begin{aligned}\text{Status\_Kredit} = &-1.5 + (0.67 * \text{Jenis\_kelamin}) \\ &+ (-2.00 * \text{Agunan}) \\ &+ (-1.33 * \text{Pen. jawab}) \\ &+ (1.33 * \text{jumlah\_pinjam}) \\ &+ (2.00 * \text{jangka waktu})\end{aligned}$$

Jika hasil perhitungan bernilai positif maka status kredit bernilai Lancar dan jika bernilai negatif maka status kredit bernilai Bermasalah.

2.5. Evaluasi dan validasi hasil

Evaluasi dan validasi hasil klasifikasi dilakukan dengan menggunakan bantuan *software Rapid Miner*, dimana semua data di pisah secara acak menjadi dua yaitu data *testing* dan data *training*.



Gambar 2. Proses evaluasi menggunakan *Rapid Miner*

p-ISSN 2460-9986
e-ISSN 2476-9436

3. HASIL DAN PEMBAHASAN

Proses yang pertama dilakukan adalah membandingkan proses pengujian dengan menggunakan 10-*fold cross validation*.

Gambar 3. Proses *10-fold cross validation.*

Berdasarkan proses *cross validation* pada Gambar 3, diperoleh hasil dari tingkat akurasi pada Tabel 8.

Tabel 8. Hasil Eksperimen

Algoritma	Accuracy (%)	Precision (%)		Type error (%)
		Positif	Negatif	
LR	87,41	87,41	0,00	-
DA	87,41	87,41	0,00	-
KNN	76,71	86,89	9.09	11,90
NB	83,56	87,41	12,50	2,38
DT	87,41	87,41	0,00	-
NN	86,73	87,33	0,00	6,802
SVM	86,39	86,39	0,00	-

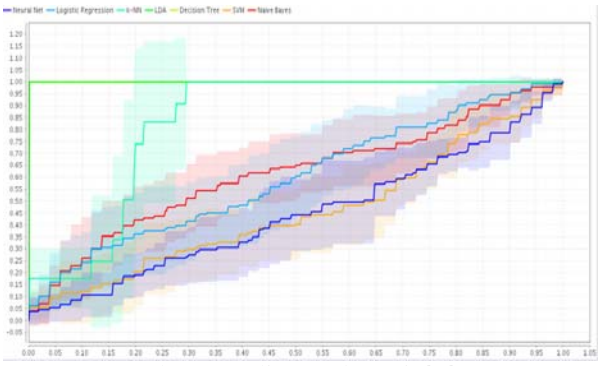
Berdasarkan dari analisis pengujian *confussion matrix*, hasil diperoleh sebagai berikut:

Tabel 9. Hasil pengujian *confussion matrix*

	LR	DA	KNN	NB	DT	NN	SVM
Accu-racy*)	87,4	87,4	76,7	83,6	87,4	86,7	86,4
AUC	1,000	1,000	0,879	0,469	1,000	0,565	0,482

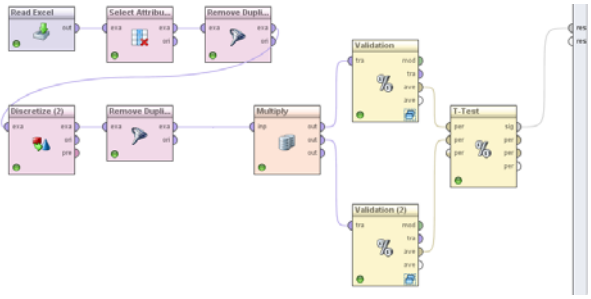
*) dalam %

Sedangkan hasil pengujian menggunakan *ROC Curve* adalah



Gambar 4. Hasil pengujian *ROC Curve*

Penentua lebih lanjut adalah dengan menggunakan pengujian dengan memanfaatkan uji statistik yaitu dengan menggunakan uji *T-Test.*



Gambar 5. Model pengujian *T-Test*

Pada proses pengujian akan dibandingkan antara dua algoritma secara bergantian sehingga akan didapatkan hasil perbandingan seperti pada Tabel 10.

Tabel 10. Hasil uji *T-Test*

	LR	DA	KNN	NB	DT	NN	SVM
LR		1.000	**0.363**	**0.326**	**0.441**	**0.062**	0.661
DA	1.000		**0.147**	**0.225**	0.662	0.460	1.000
KNN	**0.363**	**0.147**		0.755	**0.352**	**0.035**	0.225
NB	**0.326**	**0.225**	0.755		0.495	**0.290**	0.661
DT	**0.441**	0.662	**0.352**	0.495		**0.104**	0.460
NN	**0.062**	0.460	**0.035**	**0.290**	**0.104**		1.000
SVM	0.661	1.000	0.225	0.460	0.460	1.000	

4. SIMPULAN

Dari hasil analisis komparasi dengan menggunakan *cross validation, confusion matrix, ROC curve* dan *T-Test* pada beberapa algoritma klasifikasi *data mining* dapat disimpulkan bahwa algoritma yang paling akurat adalah algoritma *Logistic Regression.* karena memiliki nilai akurasi tertinggi yaitu 87,41% dengan uji *T-test* paling dominan terhadap algortima lainnya, dengan nilai AUC paling tinggi yaitu 1.000.

Algoritma *Neural Network* walaupun nilai AUC-nya kecil yaitu 0.565 tetapi setelah dilakukan uji *T-Test,*

p-ISSN 2460-9986
e-ISSN 2476-9436

algoritma ini memiliki sifat yang dominan dengan nilai akurasi cukup tinggi yaitu 86,73% sehingga dapat disimpulkan bahwa algoritma ini cukup akurat.

Algoritma *Discriminant Analysis* dan *Decision T*ree meskipun memiliki akurasi dan nilai AUC yang tinggi yaitu 87,41 % dan 1.000, tetapi berdasarkan uji *T-Test* bukan merupakan algoritma yang dominan namun masih cukup baik untuk kasus penentuan pemberian kredit.

Algoritma yang memiliki kinerja kurang memuaskan adalah *Support Vector Machine*, meskipun nilai akurasinya sebesar 86,39% dan *Naive Bayes* dengan tingkat akurasi sebesar 83,56%. Sedangkan *K-Nearest Neighbor* merupakan algoritma dengan nilai paling rendah yaitu dengan tingkat akurasi sebesar 76,71%.

Dengan kata lain seleksi fitur yang telah dilakukan mempengaruhi hasil akurasi. Tingkat akurasi yang dicapai dapat membantu para analis kredit dalam pengambilan keputusan mengenai pemberian kredit bagi nasabah koperasi.

5. REKOMENDASI

Selanjutnya perlu dilakukan pengujian kualitas data pada *dataset* yang akan digunakan untuk klasifikasi *data mining* dengan metode pengukuran pada analisis komparasi seperti metode *Delong-Pearson*.

Hasil penelitian dapat dikembangkan dengan mengoptimasi hasil eksperiman mengguanakan *Adaboost*.

6. UCAPAN TERIMAKASIH

Terima kasih peneliti sampaikan kepada Koperasi Borobudur Agung yang telah memberikan data sebagai data eksperimen dan pengujian.

7. DAFTAR PUSTAKA

Costa, G. et al., 2007. Data Mining for Effective Risk Analysis in a Bank Intelligence Scenario.

Feng, W., Zhao, Y. & Deng, J., 2009. Application of SVM Based on Principal Component Analysis to Credit Risk Assessment in Comercial Bank. In *Global Congress on Intelligent Systems*. China, 2009.

Gorunescu, F., 2011. *Data Mining: Concepts, Model and Techniques*. Berlin, Jerman: Springer.

Kasmir, 2010. *Dasar-Dasar Perbankan*. 1st ed. Jakarta, Indonesia: PT. Raja Grafindo Persada.

Keramati, A. & Yousefi, N., 2011. A Proposed Classification of Data Mining Techniques in Credit Scoring. In *Proceedings of the 2011 International Conference on Industrial Engineering and Operations Management*. Kuala Lumpur, Malaysia, 2011.

Larose, D.T., 2006. *Data Mining Methods and Models*. New Jersey, United States of America: John Wiley & Sons, Inc.

Ma, H. & Guo, Y., 2010. Credit Risk Evaluation Based on Artificial Intelligence Technology. In *2010 Iternational Conference on Artificial Intelligence and Computational Intelligence*. China, 2010.

Peng , Y. & Kou, G., 2008. A Comparative Study of Classification methods in Financial Risk Detection. In *Fourth International Conference on Networked Computing and Advanced Information management*. China, 2008.

Santoso, B., 2007. *Data Mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis*. 1st ed. Yogyakarta, Indonesia: Graha Ilmu.

Witten, I.H., Frank, E. & Hall, M.A., 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd ed. Burlington, United States of America: Morgan Kaufmann.

Yu, L. et al., 2007. Application and Comparison of Classification Techniques in Controlling Credit Risk. In P.M. Pardalos, ed. *Recent Advances in Data Mining of Enterprise Data: Algorithms and Applications*. Singapore: World Scientific. Ch. 2.

Yu, H., Huang, X., Hu, X. & Cai, H., 2010. A Comparative Study on Data Mining Algorithms for Individual Credit Risk Evaluation. In *2010 International Conference on Management of e-Commerce and e-Government*. China, 2010.

Zurada, J., 2010. In *Could Decission Trees Improve the Classification Accuracy and Interpretability of Loan Granting Decission*. Hawai, 2010. Proceedings of the 43rd Hawaii International Conference on System Sciences-2010.

Zurada, J. & Kunene, K.N., 2011. In *Comparisons of The Performance of Computational Intelligence Methods for Loan Granting Decisions*. Hawaii, 2011. Proceedings of 44th Hawaii International Conference on System Sciences-2011.