Pencarian Nasabah dengan Menggunakan Data Mining dan Algoritma C 4.5 Koperasi Maduma Subang

Timbo Faritcan Parlaungan Siallagan
STMIK Subang
timbo.siallagan@yahoo.co.id

Abstract - **Credit is the provision of money or bills that can be equated with it, based on consent or agreement between the bank and the borrowing other party require that borrowers pay off its debt after a certain period of time by the giving of flowers. Although the Lender has approved a credit proposed by the debtor, but the credit analysis to be done of debtors who have been approved so that the cause of non-performing loans can be examined and get a good classification for the determination of the appropriateness of granting credit. In granting credit need to analyse the needs of creditors, then that must be known in advance is the principles that need to be ditegakan in the framework of granting credit. Things that need to be considered in granting credit to customers is the principle 6 C's Analysis. With these problems, then there is need for troubleshooting existing solutions, by making a decision support System. Thus this decision support System will be able to meet the expectations to be achieved. The algorithm C 4.5 is algorithms used to create the decision tree. Decision tree classification method and prediction is a very powerful and famous. Getting rich in information or knowledge that is conceived by training data, the accuracy of the decision tree will be increased.**

Keywords - **credit Analysis, principles 6 C's Analysis, Data Mining and algorithms C4.5**

I. PENDAHULUAN

Dalam pemberian Kredit perlu menganalisa kebutuhan kreditur, maka yang harus diketahui terlebih dahulu adalah prinsip-prinsip yang perlu ditegakan dalam rangka pemberian Kredit. Hal-hal yang perlu diperhatikan dalam pemberian kredit bagi nasabah adalah *Prinsip 6 C's Analysis* yaitu sebagai berikut:

Character adalah keadaan watak dari nasabah, baik dalam kehidupan pribadi maupun dalam lingkungan usaha. Kegunaan dari penilaian terhadap karakter ini adalah untuk mengetahui sampai sejauh mana kemauan nasabah untuk memenuhi kewajibannya (*willingness to pay*) sesuai dengan perjanjian yang telah ditetapkan. Sebagai alat untuk memperoleh gambaran tentang karakter dari calon nasabah tersebut, dapat ditempuh melalui upaya antara lain:

a. Meneliti riwayat hidup calon nasabah;
b. Meneliti reputasi calon nasabah tersebut di lingkungan usahanya;
c. Meminta *bank to bank information* (Sistem Informasi Debitur);
d. Mencari informasi kepada asosiasi-asosiasi usaha dimana calon nasabah berada.
e. Mencari informasi apakah calon nasabah suka berjudi;
f. Mencari informasi apakah calon nasabah memiliki hobi berfoya-foya.

Capital adalah jumlah dana/modal sendiri yang dimiliki oleh calon nasabah. Semakin besar modal sendiri dalam perusahaan, tentu semakin tinggi kesungguhan calon nasabah dalam menjalankan usahanya dan bank akan merasa lebih yakin dalam memberikan kredit. Modal sendiri juga diperlukan bank sebagai alat kesungguhan dan tangung jawab nasabah dalam menjalankan usahanya karena ikut menanngung resiko terhadap gagalnya usaha. Dalam praktik, kemampuan capital ini dimanifestasikan dalam bentuk kewajiban untuk menyediakan *self-financing*, yang sebaiknya jumlahnya lebih besar daripada kredit yang dimintakan kepada bank.

Capacity adalah kemampuan yang dimiliki calon nasabah dalam menjalankan usahanya guna memperoleh laba yang diharapkan. Kegunaan dari penilaian ini adalah untuk mengetahui sampai sejauh mana calon nasabah mampu untuk mengembalikan atau melunasi utang-utangnya secara tepat waktu dari usaha yang diperolehnya.

Pengukuran capacity tersebut dapat dilakukan melalui berbagai pendekatan berikut ini:

a. *Pendekatan historis*, yaitu menilai *past performance*, apakah menunjukkan perkembangan dari waktu ke waktu.
b. *Pendekatan finansial*, yaitu menilai latar belakang pendidikan para pengurus
c. *Pendekatan yuridis*, yaitu secara yuridis apakah calon nasabah mempunyai kapasitas untuk mewakili badan usaha yang diwakilinya untuk mengadakan perjanjian kredit dengan bank.
d. *Pendekatan manajerial*, yaitu menilai sejauh mana kemampuan dan keterampilan nasabah melaksanakan fungsi-fungsi manajemen dalam memimpin perusahaan.
e. *Pendekatan teknis*, yaitu untuk menilai sejauh mana kemampuan calon nasabah mengelola faktor-faktor produksi seperti tenaga kerja, sumber bahan baku,

peralatan-peralatan , administrasi dan keuangan, industrial relation sampai pada kemampuan merebut pasar.

Collateral adalah barang-barang yang diserahkan nasabah sebagai agunan terhadap kredit yang diterimanya. Collateral tersebut harus dinilai oleh bank untuk mengetahui sejauh mana resiko kewajiban finansial nasabah kepada bank. Pada hakikatnya bentuk collateral tidak hanya berbentuk kebendaan tetapi juga collateral yang tidak berwujud seperti jaminan pribadi *(borgtocht), letter of guarantee, letter of comfort*, rekomendasi dan avalis.

Condition of Economy, yaitu situasi dan kondisi politik , sosial, ekonomi , budaya yeng mempengaruhi keadaan perekonomian pada suatu saat yang kemungkinannya memengaruhi kelancaran perusahaan calon debitur. Untuk mendapat gambaran mengenai hal tersebut, perlu diadakan penelitian mengenai hal-hal antara lain:

a. Keadaan konjungtur
b. Peraturan-peraturan pemerintah
c. Situasi, politik dan perekonomian dunia
d. Keadaan lain yang memengaruhi pemasaran

Constraint adalah batasan dan hambatan yang tidak memungkinkan suatu bisnis untuk dilaksanakan pada tempat tertentu, misalnya pendirian suatu usaha pompa bensin yang disekitarnya banyak bengkel las atau pembakaran batu bata.

Dari keenam prinsip diatas, yang paling perlu mendapatkan perhatian account officer adalah character, dan apabila prinsip ini tidak terpenuhi, prinsip lainnya tidak berarti. Dengan perkataan lain, permohonannya harus ditolak. Proses penilaian masing-masing kriteria pada kreditur di salah satu BPR dalam hal ini masih kurang memadai dalam membuat keputusan yang spesifik untuk memecahkan permasalahan kredit macet pada bank tersebut. Dibawah ini adalah gambar grafik laporan status kredit macet yang diambil dari laporan tahun 2012 untuk bulan oktober dan bulan November, gambar diambil dari tempat penelitian yaitu salah Koperasi Maduma Subang.

TABEL I
LAPORAN STATUS KREDIT MACET

Uraian	%	2013 Nominal (Rp)	%	2012 Nominal (Rp)
Cadangan Umum	10 %	17,255,723	10 %	17,612,825
Pendidikan	5%	8,627,861	5%	8,806,412
Anggota	49%	84,553,041	49%	86,301,840
Pengelola	16%	27,609,156	16%	28,180,519
Pengawas & Pengurus	20%	34,511,445	20%	35,225,649

Sumber: Laporan Koperasi Maduma Subang Oktober dan Nopember Tahun 2012

Oleh karena itu Sistem Pendukung Keputusan salah satu komponen yang cukup penting dalam sistem informasi. Dengan permasalahan tersebut, maka perlu adanya solusi pemecahan masalah yang ada, dengan membuat suatu Sistem Pendukung Keputusan. Dengan demikian Sistem Pendukung Keputusan ini nantinya dapat memenuhi harapan yang ingin dicapai. Namun, perlu diperhatikan juga bahwa nasabah yang telah disetujui juga tidak semuanya pembayar kredit yang baik, artinya ada beberapa nasabah yang telah disetujui tapi beberapa bulan kemudian pembayarannya lebih dari batas jatuh tempo atau bahkan menunggak. Pembayaran yang tidak tepat waktu jika tidak diwaspadai sejak dini maka akan menjadi suatu faktor kerugian bagi perusahaan tersebut. Oleh karena itu diperlukan suatu penggalian data terhadap nasabah atau Debitur. Ada beberapa atribut yang menyertai data debitur yaitu Nama_Nasabah, Jenis_Kelamin,Umur,Jumlah_Pinjaman, Jangka_Waktu,Jumlah_Angsuran_Perbulan,Type_Pinjaman ,Jenis_Pinjaman,Bi_Sektor_Ekonomi,Col,Bi_Golongan_Debitur,Bi_Golongan_Penjamin,Saldo_Nominatif, Tunggakan_Pokok, Tunggakan_Bunga, Status_Kredit.Banyak penelitian membahas mengenai penentuan kelayakan pemberian kredit dengan berbagai algoritma data mining. Seperti penelitian yang dilakukan Abbas Heiat (2011) menyatakan bahwa Risiko bagi lembaga keuangan untuk memberikan kredit yang diminta tergantung pada seberapa baik mereka membedakan pemohon kredit yang baik dari para pemohon kredit macet. Di bawah ini adalah beberapa penelitian yang berkaitan dengan masalah kredit yaitu:

a. Jiang (2009) membuat model untuk memprediksi nasabah yang bermasalah dan tidak bermasalah dalam pembayaran kredit dengan menggunakan model algoritma C4.5. Data yang digunakan diambil dari perusahaan German credit yang merupakan perusahaan pembiayaan. Jiang mengambil beberapa atribut dan kemudian dimasukkan ke dalam model untuk memprediksi persentase nasabah yang bermasalah. Pada penelitian ini, peneliti menyatakan hasil penelitiannya yaitu Statistik menunjukkan bahwa biaya *misclassifying* kredit lancar dan kredit macet adalah 5 ~ 20 kali dari *misclassifying* kredit lancar dan kredit macet.
b. S. Satchidananda and J. B. Simha (2006) Penelitian ini membandingkan dua model algoritma untuk analisa resiko kredit, yaitu Pohon Keputusan dan Regresi Logistik. Data diambil dari dua bank yang berbeda, kemudian untuk mengelompokkan kasus positif dan negatif maka dilakukan klustering data dengan menggunakan k-means. Hasil analisa dari masing-masing model dikomparasi dan kemudian diukur, kemudian didapatkan bahwa algoritma pohon keputusan mempunyai tingkat akurasi yang tinggi dibandingkan algoritma regresi logistik. Penelitian ini

masih dalam proses untuk menyelidiki kinerja yang diusulkan, Pendekatan dibandingkan dengan teknik klasifikasi lainnya untuk credit scoring sehingga tingkat akurasinya belum dapat diketahui.

c. C. Firmansyah (2011), juga melakukan penelitian dengan judul "Penerapan Algoritma Klasifikasi C4.5 untuk Penentuan Kelayakan Pemberian Kredit Koperasi" Nilai *accuracy*, *precision*, dan *recall* nya dari data training dapat dihitung dengan menggunakan Rapid Miner. Setelah diuji coba dengan metode *crossvalidation*, didapatkan hasil pengukuran terhadap data trainingnya yaitu hanya mencapai *accuracy* = 79.50%, *precision* = 86.50% dan *recall* = 91.00% [5]. Hasil pengujian tersebut berdasarkan 5 parameter, merupakan masalah penelitian Teknik Informatika yang masih bisa di tingkatkan akurasinya. mengapa Hasil penelitian sebelumnya "Penerapan Algoritma Klasifikasi C4.5 untuk Penentuan Kelayakan Pemberian Kredit Koperasi" Nilai *accuracy nya hanya 79.50 % ?* oleh karena itu penulis berkesempatan untuk meningkatkan hasil penelitian yang dilakukan oleh Firmansyah dengan cara menambahkan 1 parameter sehingga jumlah parameter nya menjadi 6 parameter dalam mengklasifikasikan Nasabah atau *Debitur*"

II. RUMUSAN MASALAH

Dari hasil identifikasi masalah yang terdapat di salah satu Koperasi adalah meningkatnya jumlah kredit macet. Berdasarkan laporan data kredit nasabah pada bulan Oktober dan November tahun 2012 diketahui bahwa jumlah kredit macet semakin meningkat, maka perlu didukung dengan system pendukung keputusan kelayakan pemberian kredit bagi nasabah sehingga masalah tersebut dapat terpecahkan.

Tujuan Penelitian

Berdasarkan latar belakang dan rumusan masalah diatas, maka penelitian ini bertujuan untuk mengantisipasi jumlah nasabah yang melakukan pembayaran melewati jatuh tempo yang sudah ditetapkan agar tidak terjadi kenaikan jumlah kredit macet yang berpotensi terjadinya kerugian pada pihak Koperasi

Manfaat Penelitian

a. Manfaat bagi masyarakat
   Manfaat hasil penelitian ini adalah agar petugas Analys Kredit dapat mengetahui dan memiliki standar untuk menentukan pemberian kredit kepada calon Nasabah sehingga dapat meminimalisir terjadinya kredit macet yang berdampak terjadinya kerugian bagi Koperasi atau pihak nasabah.
b. Manfaat bagi IPTEK
   Hasil penelitian ini diharapkan dapat memberikan sumbangan penerapan model SPK Kelayakan Pemberian Kredit Nasabah dengan Metode C4.5 berdasarkan prisip *6 C's Analysis*.

III. TINJAUAN PUSTAKA

Kredit

Kredit adalah penyediaan uang atau tagihan yang dapat dipersamakan dengan itu, berdasarkan persetujuan atau kesepakatan pinjam meminjam antara bank dengan pihak lain yang mewajibkan pihak peminjam untuk melunasi utangnya setelah jangka waktu tertentu dengan pemberian bunga sedangkan nasabah adalah pihak yang menggunakan jasa bank. Dalam penelitian ini kita akan membahas masalah kredit sehingga kita akan membahas pula masalah nasabah atau debitur, Nasabah debitur adalah Nasabah yang memperoleh fasilitas kredit atau pembiayaan

Dalam proses kredit ada beberapa atribut yang dijadikan bahan analisis pemberian kredit yaitu:

a. Nama nasabah
b. Jenis kelamin
c. Umur
d. Jumlah pinjaman
e. Jangka waktu
f. Jumlah angsuran per bulan
g. Type pinjaman
h. Jenis pinjaman
i. Bidang sektor ekonomi
j. Col
k. Bidang golongan debitur
l. Bidang golongan penjamin
m. Saldo nominatif
n. Tunggakan pokok
o. Tunggakan bunga
p. Status kredit

Data Mining

Definisi Data Mining

a. Mengekstrak atau "mining" pengetahuan dari kumpulan data yg sangat besar
b. Ekstraksi informasi yg berguna dari data, dimana sebelumnya tidak diharapkan, tidak dikenal & implisit
c. Eksplorasi & analisis, secara otomatis atau semi-otomatis dari sekumpulan data yg sangat besar untuk memperoleh pola2 data yg berarti
d. Proses analisis database yg besar secara semi-otomatis untk menemukan pola yang valid, baru, berguna dan dapat dipahami manusia

Data mining merupakan bagian dari proses Knowledge Discovery in Databases (KDD) – Proses transformasi data mentah menjadi informasi berguna. Dibawah ini adalah gambar proses kerja data mining.

Pada dasarnya *data mining* terdiri dari :

a. *Predictive*, metode yang menggunakan beberapa variabel yang ada untuk memprediksi nilai masa depan

(belum diketahui) dari variabel lain. Contoh : *classification, regression, biases/anomalies detection*.

b. Descriptive, metode yang mengungkapkan pola dalam data, agar mudah diinterpretasikan oleh pengguna. Contoh : *clustering, association rules, sequential patterns.*

Pengelompokan Data Mining

Data mining dibagi menjadi beberapa kelompok berdasaarkan tugas yang dapat dilakukan, yaitu [12]:

Deskripsi

Terkadang peneliti dan analis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data. Sebagai contoh, petugas pengumpul suara mungkin tidak dapat menemukan keterangan atau fakta bahwa siapa yang tidak cukup profesional akan sedikit di dukung dalam pemilihan presiden. Deskripsi dari pola dan kecenderungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan.

Estimasi

Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik daripada ke arah kategori. Model dibangun menggunakan *record* lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi. Sebagai contoh, akan dilakukan estimasi tekanan darah sistolik pada pasien rumah sakit berdasarkan umur pasien, jenis kelamin, indeks berat badan, dan level sodium darah. Hubungan antara tekanan darah sistolik dan nilai variabel prediksi dalam proses pembelajaran akan menghasilkan model estimasi. Model estimasi yang dihasilkan dapat digunakan untuk kasus baru lainnya. Contoh lain yaitu estimasi nilai indeks prestasi kumulatif mahasiswa program pasca sarjana dengan melihat nilai indeks prestasi mahasiswa tersebut pada saat mengikuti program sarjana.

Prediksi

Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada dimasa mendatang.

Contoh prediksi dalam bisnis dan penelitian adalah:

a. Prediksi harga beras dalam tiga bulan yang akan datang.
b. Prediksi presentase kenaikan kecelakaan lalu lintas tahun depan jika batas bawah kecepatan dinaikan .

Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan (untuk keadaan yang tepat) untuk prediksi.

Klasifikasi

Dalam klasifikasi, terdapat target variabel kategori. Sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam 3 kategori, yaitu pendapatan tinggi, pendapatan sedang, pendapatan rendah.

Contoh lain klasifikasi dalam bisnis dan penelitian adalah:

a. Menentukan apakah suatu transaksi kartu kredit merupakan transaksi yang curang atau bukan.
b. Memperkirakan apakah suatu pengakuan hipotek oleh nasabah merupakan suatu kredit yang baik atau buruk.
c. Mendiagnosis seorang penyakit pasien untuk mendapatkan termasuk kategori penyakit apa.

Pengklusteran

Pengklusteran merupakan pengelompok *record*, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. Kluster adalah kumpulan record yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan *record-record* dengan kluster lain.

Pengklusteran berbeda dengan klasifikasi yaitu tidak adanya variabel target dalam pengklusteran. Pengklusteran tidak mencoba untuk melakukan klasifikasi, mengestimasi, atau memprediksi nilai dari variabel target. Akan tetapi, algoritma pengklusteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan (homogen), yang mana kemiripan *record* dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan *record* dalam kelompok lain akan bernilai minimal.

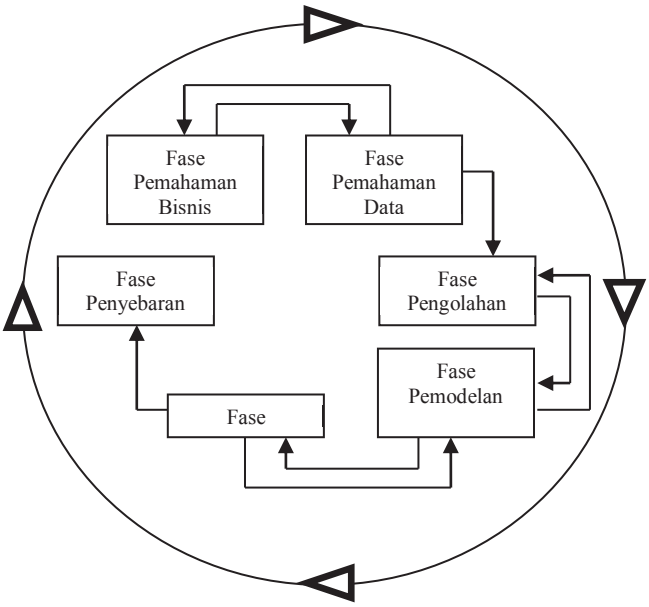
Contoh pengklusteran dalam bisnis dan penelitian adalah:

a. Mendapatkan kelompok-kelompok konsumen untuk target pemasaran dari suatu produk bagi perusahaan yang tidak memiliki dana pemasaran yang besar.
b. Untuk tujuan audit akuntansi, yaitu malakukan pemisahan terhadap perilaku finansial dalam baik dan mencurigakan.
c. Melakukan pengklusteran terhadap ekspresi dari gen, untuk mendapatkan kemiripan perilaku dari gen dalam jumlah besar.

CRISP-DM (Cross Industry Standard Process For Data Mining) EEE 2006

CRISP-DM (*Cross Industry Standard Process For Data Mining*) yang dikembangkan tahun 1996 oleh analisis dari beberapa industri seperti DaimlerChrysler, SPSS, dan NCR. CRISP DM menyediakan standar proses *data mining* sebagai strategi pemecahan masalah secara umum dari bisnis atau unit penelitian. Dalam CRISP-DM, Sebuah proyek *data mining* memiliki siklus hidup yang terbagi dalam enam fase . Keseluruhan fase berurutan yang ada tersebut bersifat adaptif. Fase berikutnya dalam urutan

bergantung kepada keluaran dari fase sebelumnya. Hubungan penting antar fase digambarkan dengan panah. Sebagai contoh, jika proses berada pada fase *modeling*. Berdasar pada perilaku dan karakteristik model, proses mungkin harus kembali pada fase *data preparation* untuk perbaikan lebih lanjut terhadap data atau berpindah maju kepada fase *evaluation.*



Gambar 1 Proses Data Mining Menurut CRISP-DM

Dibawah ini adalah enam fase CRISP-DM:

1. Fase Pemahaman Bisnis (*Business Understanding Phase*)
   a. Penentuan tujuan proyek dan kebutuhan secara detail dalam lingkup bisnis atau unit penelitian secara keseluruhan.
   b. Menerjemahkan tujuan dan batasan menjadi formula dari permasalahan *data mining* .
   c. Menyiapkan strategi awal untuk mencapai tujuan
2. Fase Pemahaman Data (*Data Understanding Phase*)
   a. Mengumpulkan data.
   b. Menggunakan analisis penyelidikan data untuk mengenali lebih lanjut data dan pencarian pengetahuan awal.
   c. Mengevaluasi kualitas data.
   d. Jika diinginkan, pilih sebagian kecil grup data yang mungkin mengandung pola dari permasalahan.

3. Fase Pengolahan Data (*Data Preparation Phase*)
   a. Siapkan dari data awal, kumpulan data yang akan digunakan untuk keseluruhan fase berikutnya. Fase ini merupakan pekerjaan berat yang perlu dilaksanakan secara intensif.
   b. Pilih kasus dan variabel yang ingin dianalisis dan yang sesuai analisis yang akan dilaksanakan.
   c. Lakukan perubahan pada beberapa variabel jika dibutuhkan.
   d. Siapkan data awal sehingga siap untuk perangkat pemodelan.

4. Fase Pemodelan (*Modeling Phase*)
   a. Pilih dan aplikasikan teknik pemodelan yang sesuai.
   b. Kalibrasi aturan model untuk mengoftimalkan hasil.
   c. Perlu diperhatikan bahwa beberapa teknik mungkin untuk digunakan pada permasalahan *data mining* yang sama.
   d. Jika diperlukan, proses dapat kembali ke fase pengolahan data untuk menjadikan data kedalam bentuk yang sesuai dengan spesifikasi kebutuhan teknik *data mining* tertentu.
5. Fase Evaluasi (*Evaluation Phase*)
   a. Mengevaluasi satu atau lebih model yang digunakan dalam fase pemodelan untuk mendapatkan kualitas dan efektivitas sebelum disebarkan untuk digunakan.
   b. Menetapkan apakah terdapat model yang memenuhi tujuan pada fase awal.
   c. Menentukan apakah terdapat permasalahan penting dari bisnis atau penelitian yang tidak tertangani dengan baik.
   d. Mengambil keputusan berkaitan dengan penggunaan hasil dari *data mining*.

6. Fase Penyebaran (*Deployment Phase*)
   a. Menggunakan model yang dihasilkan. Terbentuknya model tidak menandakan telah terselesainya proyek.
   b. Contoh sederhana penyebaran: Pembuatan laporan.
   c. Contoh kompleks penyebaran: penerapan proses data mining secara paralel pada departemen lain.

Algoritma C4.5

Algoritma C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan. Pohon keputusan merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami. Dan mereka juga dapat diekspresikan dalam bentuk bahasa basis data seperti *Structured Query Language* untuk mencari *record* pada kategori tertentu.

Berikut adalah algoritma C4.5, yaitu :
Input: an attribute-valued dataset D
1: Tree = {}
2: if D is “pure” OR other stopping criteria met then
3: terminate
4: end if
5: for all attribute a ∈ D do

6: Compute information-theoretic criteria if we split on a
7: end for
8: abest = Best attribute according to above computed criteria
9: Tree = Create a decision node that tests abest in the root
10: Dv = Induced sub-datasets from D based on abest
11: for all Dv do
12: Treev = C4.5(Dv)
13: Attach Treev to the corresponding branch of Tree
14: end for
15: return Tree

Ada beberapa tahap dalam membuat sebuah pohon keputusan dengan algoritma C4.5, yaitu :

1. Menyiapkan data training. Data training biasanya diambil dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokkan ke dalam kelas-kelas tertentu.
2. Menentukan akar dari pohon. Akar akan diambil dari atribut yang terpilih,dengan cara menghitung nilai Gain dari masing-masing atribut, nilai Gain yang paling tinggi yang akan menjadi akar pertama. Sebelum menghitung nilai Gain dari atribut, hitung dahulu nilai entropy yaitu :

   Entropy (S) = $\sum_{i=1}^{n} - pi * log_2 pi$ (1)

   Keterangan :
   S : himpunan kasus
   n : jumlah partisi S
   pi : proporsi dari Si terhadap S
   Entropi menyatakan *impurity* suatu kumpulan objek dan digunakan untuk memilih nilai optimal untuk memecahkan node berdasarkan maksimalisasi informasi. Jika semua objek memiliki label kelas yang sama maka entropinya adalah 0 dan akan meningkat nilai entropi hingga maksimum ketika semua kelas sama-sama didistribusikan.
3. Kemudian hitung nilai Gain dengan metode *information gain* :

   Gain(S,A) = Entropy(S) – $\sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i)$

   Keterangan :
   S : himpunan kasus
   A : atribut
   n : jumlah partisi atribut A
   |Si| : jumlah kasus pada partisi ke-i
   |S| : jumlah kasus dalam S
4. Ulangi langkah ke-2 hingga semua tupel terpartisi.
5. Proses partisi pohon keputusan akan berhenti saat :
   a. Semua tupel dalam node N mendapat kelas yang sama.
   b. Tidak ada atribut di dalam tupel yang dipartisi lagi.
   c. Tidak ada tupel di dalam cabang yang kosong.

Confusion Matrix dan Kurva ROC

Mengingat bahwa evaluasi kinerja model klasifikasi didasarkan pada tuntutan (pengujian) yang memperkirakan obyek tersebut benar dan salah. Hitungannya ini bisa ditabulasikan dalam bentuk yang disebut *confusion matrix*. Secara singkat, *confusion matrix* memberikan perincian mendetail mengenai *misclassifications*. Kelas yang diprediksi akan ditampilkan di bagian atas matriks, dan kelas diamati di sisi kiri. Setiap sel berisi sejumlah menunjukkan berapa banyak kasus yang sebenarnya dari kelas yang diamati diberikan ditugaskan oleh model ke kelas diprediksi diberikan. Untuk lebih jelasnya berikut ini adalah gambar 5 contoh model *confusion matrix.*

TABEL II
CONTOH MODEL CONFUSION MATRIX

CLASSIFICATION		PREDICTED CLASS	
		Class = YES	Class = NO
OBSERVED CLASS	Class = YES	*a* (*true positive*-TP)	*b* (*false negative* -FN)
	Class = NO	*c* (*false positive*-FP)	*d* (*true negative*-TN)

Setelah data uji dimasukkan ke dalam confusion matrix, hitung nilai-nilai yang telah dimasukkan tersebut untuk dihitung jumlah *precision, recall* dan *accuracy*. *Sensitivity* digunakan untuk membandingkan jumlah true positives terhadap jumlah tupel yang positives sedangkan *specificity* adalah perbandingan jumlah true negatives terhadap jumlah tupel yang negatives. Untuk menghitung digunakan persamaan di bawah ini :

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

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

Sekarang beberapa gagasan dasar tentang kurva ROC (***R**eceiver **O**perating **C**haracteristic*) digunakan secara luas dalam menilai hasil prediksi. Kurva ROC juga biasanya digunakan dalam pembelajaran mesin dan penelitian data mining. Salah satu yang mengadopsi kurva ROC dalam pembelajaran mesin adalah Spackman, yang memperlihatkan diperlukannya ROC kurva dalam mengevaluasi dan membandingkan algoritma. Dalam masalah klasifikasi, kurva ROC adalah teknik untuk memvisualisasikan, mengatur dan memilih pengklasifikasi, berdasarkan kinerja mereka.
Secara teknis, ROC kurva, juga dikenal sebagai grafik ROC, adalah dua-dimensi grafik di mana tingkat TP diplot pada sumbu Y dan tingkat FP diplot pada sumbu-X. Dengan cara ini, grafik ROC menggambarkan perbandingan antara keuntungan ("*true positives*") dengan biaya ("*false positives*").

III. METODE PENELITIAN

Penelitian yang dilaksanakan adalah jenis penelitian eksperimen, yaitu melakukan pengujian tingkat akurasi algoritma C 4.5 dalam pengklasifikasian nasabah kredit

lancar dan kredit macet. Data eksperimen diambil dari tempat penelitian yaitu di salah satu koperasi.

Ada beberapa tahap yang dilakukan dalam melakukan eksperimen ini, penulis menggunakan model *Cross-Standard Industry for Data Mining* (CRISP-DM) yang terdiri dari 6 tahap, yaitu :

a. Tahap *business understanding*
Penelitian pendahuluan dilakukan dengan melakukan observasi ke tempat penelitian untuk melihat dan mengetahui secara langsung kondisi dan permasalahan yang terjadi. Terdapat peningkatan jumlah kredit macet pada laporan kredit tahun 2012, ini dikarenakan masih sulitnya menentukan klasifikasi kredit lancar dan kredit macet dengan akurasi yang baik sehingga perlu dikembangkan model klasifikasi yang baru.

b. Tahap *data understanding*.
Data diperoleh dari koperasi pada tahun 2012. Data tersebut sebanyak 700 record memiliki atribut Nama_Nasabah, Jenis_Kelamin, Umur, Jumlah_Pinjaman,Jangka_Waktu,Jumlah_Angsuran_Perbulan,Type_Pinjaman, Jenis_Pinjaman, Bi_Sektor_Ekonomi,Col,Bi_Golongan_Debitur, Bi_Golongan_Penjamin,Saldo_Nominatif,Tunggakan_Pokok,Tunggakan_Bunga,Status_Kredit. Nilai dari semua atribut yang ada di tabel, merupakan nilai kategorikal dan bukan nilai angka, misalnya seperti atribut umur, yaitu debitur yang berusia 17 th sampai 40 tahun termasuk dalam kategori muda, sedangkan debitur yang berusia 41 tahun sampai 55 tahun termasuk kategori paruh baya, dan kategori ketiga adalah debitur yang berusia diatas 55 tahun termasuk kategori tua. Tabel 6 di bawah ini ditampilkan nama atribut, kategori, dan nilai angka (rangenya)

c. Tahap *data preparation*
Tabel di bawah ini menunjukkan data transaksi kredit baik yang bermasalah maupun yang tidak bermasalah.

digunakan dan record yang duplikasi. Untuk itu maka diperlukan tehnik dalam preprocessing yaitu :

a. Data cleaning bekerja untuk membersihkan nilai yang kosong ,tidak konsisten atau mungkin tupel yang kosong (missing values dan noisy).
b. Data integration berfungsi menyatukan tempat penyimpanan (arsip) yang berbeda ke dalam satu data. Dalam hal ini, ada dua arsip yang diambil sebagai data warehouse yaitu data anggota dan data kredit.
c. Data *reduction*. Jumlah atribut dan tupel yang digunakan untuk data training mungkin terlalu besar, hanya beberapa atribut yang diperlukan sehingga atribut yang tidak diperlukan akan dihapus. Tupel dalam data set mungkin terjadi duplikasi atau terdapat tupel yang sama, sehingga untuk memperkecil jumlah tupel, tupel yang sama akan dijadikan dalam satu tupel untuk mewakili tupel tersebut akan terlihat pada tabel 8 di bawah :

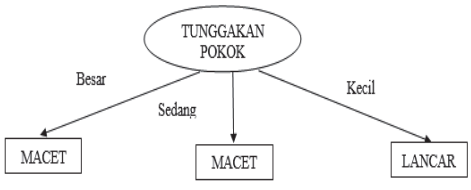
IV. ALGORITMA C 4.5

Tahap ini juga dapat disebut tahap *learning* karena pada tahap ini data training diklasifikasikan oleh model dan kemudian menghasilkan sejumlah aturan. Model yang digunakan dalam tahap ini menggunakan algoritma C4.5. Seperti yang telah dijelaskan sebelumnya, ada beberapa tahap yang harus dilalui dalam membentuk pohon keputusan, tentunya algoritma C4.5 digunakan untuk membuat pohon keputusan.

TABEL III
HASIL PERHITUNGAN INFORMATION GAIN

Node			Jumlah Kasus	MACET Si	LANCAR Si	Entropy	Gain Information	Keterangan
1	TOTAL		103	53	50	0,999388		
	jenis_kelamin						0,005801	
		Laki-laki	48	27	21	0,988699		
		Perempuan	55	26	29	0,997853		
	umur						0,028436	
		muda	64	28	36	0,988699		
		paruh baya	39	25	14	0,941829		
		tua	0	0	0	0,000000		
	jml_pinjaman						0,002287	
		kecil	89	45	44	0,999909		
		sedang	10	6	4	0,970951		
		besar	4	2	2	1,000000		
	jangka waktu						0,008102	
		cepat	42	22	20	0,998364		
		sedang	11	4	7	0,945660		
		lambat	50	27	23	0,995378		
	jml_angsuran_per_bulan						0,002667	
		kecil	38	19	19	1,000000		
		sedang	60	32	28	0,996792		
		besar	5	2	3	0,970951		
	bi_gol_penjamin						0,094982	
		Perorangan	21	18	3	0,591673		
		Tanpa Penjamin	82	35	47	0,984496		
	saldo_nominatif						0,006527	
		kecil	92	46	46	1,000000		
		sedang	9	6	3	0,918296		
		besar	2	1	1	1,000000		
	tunggakan_pokok						0,303630	Gain information tertinggi
		kecil	73	24	49	0,913662		
		sedang	12	11	1	0,413817		
		besar	18	18	0	0,000000		
	tunggakan_bunga						0,146177	
		kecil	84	35	49	0,979869		
		sedang	18	17	1	0,309543		
		besar	1	1	0	0,000000		

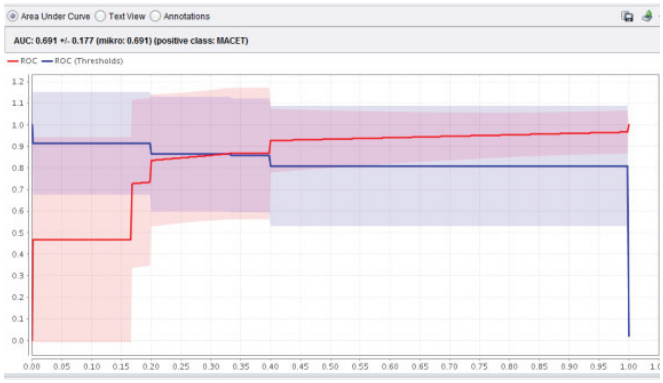
Dari hasil perhitungan diatas, maka didapatkan model pohon keputusan seperti berikut:



Gambar 3 Pohon Keputusan Menggunakan Algoritma C4.5

Kurva ROC

Kurva ROC menunjukkan *trade-off* antara *true positive rate* (proporsi tuple positif yang teridentifikasi dengan benar) dan *false positive rate* (proporsi tuple negatif yang teridentifikasi salahsebagai positif) dalam suatu model. Untuk mengukur ketelitian dari suatu model, kita dapat mengukur area di bawah kurva ROC.



Gambar 4 Kurva Akurasi C4.5

Gambar 22 menunjukkan grafik ROC dengan nilai AUC (Area Under Curve) dengan C 4.5sebesar 0.691. Akurasi AUC dikatakan sempurna apabila nilai AUC mencapai 1.000 dan akurasinya buruk jika nilai AUC dibawah 0.500.

Dengan kurva ROC, kita dapat melihat trade off antara tingkat dimana suatu model dapat mengenali tuple positif secara akurat dan tingkat dimana model tersebut salah mengenali tuple negatif sebagai tuple positif. Kurva ROC terdiri atas sumbu vertikal yang menyatakan true positive rate, dan sumbu horizontal yang menyatakan false positive rate.

Jika memiliki true positif (sebuah tupel positif yang benar diklasifikasikan) maka pada kurva ROC akan bergerak ke atas dan plot titik. Sebaliknya, jika tupel milik kelas "tidak" ketika memiliki false positif, maka kurva ROC bergerak ke kanan dan plot titik. Proses ini diulang untuk setiap tupel tes (setiap kali bergerak ke atas kurva untuk true positif atau terhadap hak untuk false positif).

Hasil pengujian Dataset Kredit menggunakan Metode C.4.5

accuracy: 87.36% +/- 17.79% (mikro: 87.38%)

	true LANCAR	true MACET	class precision
pred. LANCAR	46	9	83.64%
pred. MACET	4	44	91.67%
class recall	92.00%	83.02%	

Gambar 5 Nilai Akurasi

Dari gambar di atas diperoleh jumlah *True Negative* (TN) sebanyak 46 sebagai *false* dan sesuai dengan klasifikasi, *False Positive* (FP) sebanyak 9 diprediksi *false* ternyata hasil prediksi *True Positive* (TP) sebanyak 44 diklasifikasi sebagai *True* dan sesuai dengan prediksi yang dilakukan menggunakan *cross validation* dan *False Negative* (FN) sebanyak 4 dan klasifikasinya *true* ternyata hasil klasifikasinya *false*. Tingkat akurasi yang diperoleh.

V. KESIMPULAN

Dengan dihasilkannya klasifikasi kelayakan pemberian kredit nasabah dengan jumlah atribut 8 menghasilkan akurasi 87.36 % merupakan tingkat akurasi yang baik, sehingga kelancaran proses penilaian kelayakan kredit dapat tercipta dan pembayaran terlambat (menunggak) sudah terprediksi dari awal untuk dapat diwaspadai agar dapat meminimalisir meningkatnya kredit macet.

Tingkat akurasi algoritma C 4.5 eksperiment ini sudah mencapai tingkat baik, sehingga dapat meningkatkan ketelitian dalam proses klasifikasi dan prediksi dengan cara menambahkan beberapa atribut dari histori pembayaran kredit nasabah yang ada sehingga dihasilkan pola klasifikasi yang lebih akurat.

DAFTAR PUSTAKA

[1] Y. Jiang, "Credit Scoring Model Based on the Decision Tree and the Simulated Annealing Algorithm," Learning, no. 2007, pp. 18–22, 2009.
[2] J. B. Simha, "Comparing decision trees with logistic regression for credit risk analysis," 2006.
[3] K. K. Lai, L. Yu, L. Zhou, and S. Wang, "Credit Risk Evaluation with Least Square Support Vector Machine," Evaluation, pp. 490–495, 2006.
[4] J. Zurada, "Could Decision Trees Improve the Classification Accuracy and Interpretability of Loan Granting Decisions ?," Information Systems, pp. 1–9, 2010.
[5] *S. Sogala and P. D, "Comparing the Efficacy of the Decision Trees with Logistic Regression for Credit Risk Analysis."*
[6] F. C. Li, "The Hybrid Credit Scoring Model based on KNN Classifier," Sixth International Conference on Fuzzy Systems and Knowledge Discovery. IEEE Computer Society, 2009.
[7] (Central Connecticut State LAROSE, DANIEL T. University), DISCOVERING KNOWLADGE IN DATA. Canada: John Wiley & Sons, Inc., Hoboken, New Jersey., 2005.
[8] Prof . F l or i n G or u n e s c u, Data Mining Concepts, Models and Tecniques. Berlin: Springer Verlag Berlin Heidelberg, 2011, p. 16.
[9] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. Mclachlan, A. Ng, B. Liu, P. S. Yu, Z. Z. Michael, S. David, and J. H. Dan, "Top 10 algorithms in data mining," Knowledge and Information Systems, pp. 1–37, 2008.
[10] J. A. Bastos and R. Archive, "Credit scoring with boosted decision trees," no. 8156, 2008.
[11] K. Xindong Wu, "The Top Ten Algorithms in Data Mining," 2009.
[12] M. K. Jiawei, Data Mining Concepts and Techniques. 2006.