ISSN : 2355-9365 e-Proceeding of Engineering : Vol.2, No.1 April 2015 | Page 1768

IMPLEMENTASI KLASIFIKASI DECISION TREE DENGAN ALGORITMA C4.5 DALAM PENGAMBILAN KEPUTUSAN PERMOHONAN KREDIT OLEH DEBITUR (STUDI KASUS: BANK PASAR DAERAH ISTIMEWA YOGYAKARTA)

Rafik Khairul Amin, Dra.Indwiarti ,M.Si,Yuliant Sibaroni,S.Si., M,T
Prodi lmu Komputasi Fakultas Informatika
Universitas Telkom
rafikkhairulamin@gmail.com, indwindwi@gmail.com, ysibaroni@gmail.com

Abstrak

Meminjam dangan cara kredit sudah merupakan hal biasa di masyarakat. Sebelum mendapatkan kredit, seseorang harus melalui *survey* yang akan dilakukan oleh seorang analisis kredit untuk mengetahui apakah pemohon kredit layak atau tidak layak untuk mendapat kredit.Seorang analisis kredit harus benar-benar teliti dalam memprediksi pemohon kredit tersebut dalam pemberian kredit agar tidak terjadi kredit macet. Perlu adanya suatu penunjang keputusan untuk membantu seorang analisis kreditdalammemprediksi pemohon kredit.

Pohon keputusan adalah model prediksi menggunakan struktur pohon atau struktur berhirarki. Pohon keputusan merupakan salah satu metode klasifikasi yang paling populer karena mudah untuk dipahami. C4.5 merupakan algoritma pohon keputusan yang sering digunakan untuk membuat suatu pohon keputusan karena memiliki tingkat akurasi yang tinggi dalam menentukan keputusan.

Algoritma C4.5 adalah suksesor dari ID3 dimana pemilihan *root* dan *parent* bukan hanya berdasar *information gain* saja tetapi juga *split information* untuk mendapatkan *Gain Ratio*.

Dataset yang digunakan dalam penelitian ini yaitu sebanyak 1000 data dengan proporsi 70% disetujui dan 30% data debitur yang ditolak.

Dalam laporan ini dibahas kinerja algoritma pohon keputusan C4.5 pada identifikasi kelayakan kredit oleh debitur. Dari penelitian yang dilakukan, diketahui nilai *precision* terbesar dicapai oleh algoritma C4.5 dengan partisi data 90%:10% dengan nilai sebesar 78,08 %. Nilai *recall* terbesar partisi data 80%:20% dengan nilai sebesar 96,4 %.

Dari hasil data latih yang sama,ID3 menghasilkan *precision* sebesar 71,51% dan *recall* sebesar 92,09%

Hasil akhir dari penelitian ini membuktikan bahwa pada kasus ini algoritma C4.5 memiliki tingkat akurasi yang tinggi dan lebih baik dari ID3.

Kata kunci :Pohon Keputusan, C4.5, Kelayakan Kredit Debitur, *Gain Ratio*.

I. PENDAHULUAN

Perkembangan teknologi informasi yang semakin pesat pada saat ini selalu berusaha untuk memenuhi kebutuhan dan kemudahan dalam pencarian,penyajian dan penanganan data. Hampir semua bidang membutuhkan kemudahan untuk penanganan informasi yang mereka miliki, terutama untuk bidang bisnis dalam dunia keuangan seperti perbankan.

Kredit merupakan sumber utama penghasilan bagi bank,sekaligus sumber resiko operasi bisnis terbesar.sebagian dana operasional bank diputarkan dalam kredit. Bila kegiatan bisnis yang satu ini berhasil, akan berhasil pula operasi bisnis mereka[1]. Memberikan kredit adalah pekerjaan mudah, tetapi untuk menarik kembali kredit macet atau bermasalah dari debitur dibutuhkan keahlian, pengalaman, serta waktu dan biaya yang cukup besar. Kredit macet atau bermasalah merupakan duri dalam daging bank, karena kredit macet dalam jumlah besar dapat mengganggu sendi kehidupan ekonomi dan akan menggerogoti dana operasional bank itu sendiri, serta membahayakan likuiditas keuangan bank dan menurunkan kepercayaan masyarakat dan luar negeri terhadap profesionalisme pengelolaan bisnis perbankan nasional. Resiko kredit macet atau bermasalah tadi dapat diperkecil dengan melakukan evaluasi kredit secara profesional, yang dilakukan sebelum pengambilan keputusan untuk memberikan kredit.

Dalam melakukan evaluasi permintaan kredit, seorang analisis kredit akan meneliti kondisi calon debitur yang diperkirakan dapat mempengaruhi kemampuan mereka dalam memenuhi kewajiban kepada bank. Untuk meneliti kondisi tersebut, analisis akan perlu mengumpulkan data-data tentang calon debitur ini baik yang kuantitatif seperti data keuangan, maupun kualitatif seperti penilaian terhadap pengelolaan perusahaan dan sebagainya. Kemudian data-data ini akan diolah dan diproses sesuai prosedur pada bank tersebut sebelum akhirnya diambil keputusan apakah layak untuk memperoleh pinjaman kredit dari bank.

Terdapat beberapa algoritma klasifikasi data salah satunya yaitu pohon keputusan atau *decision tree*.Algoritma C4.5 merupakan pengembangan dari algoritma konvensional induksi pohon keputusan yaitu ID3. Algoritma yang merupakan pengembangan dari ID3 ini dapat mengklasifikasikan data dengan metode pohon keputusan yang memiliki kelebihan dapat mengolah data numerik (kontinyu) dan diskret, dapat menangani nilai atribut yang hilang, menghasilkan aturan-aturan yang mudah di interpretasikan, dan tercepat diantara algoritma-algoritma yang menggunakan memori utama di komputer. Pada penerapan beberapa kasus teknik klasifikasi, algoritma ini mampu menghasilkan akurasi dan performansi yang baik.

Selain itu Penelitian tentang algoritma tersebut sebelumnya juga pernah dilakukan oleh Lilis Setyowati dengan menggunakan algoritma decision tree C4.5,ada penelitian tersebut dijelaskan tentang cara memilih pegawai berdasarkan karakteristik calon pegawai yang menghasilkan akurasi yang besar yaitu 81,25%.selain itu penelitian tentang C4.5 juga pernah dilakukan oleh Gelar Nurcahya tentang klasifikasi data konsumen telemarketing untuk deposito pada bank,karena metode yang digunakan tepat untuk diimplementasikan pada kasus tersebut berdasarkan klasifikasi dari data *record*, maka dari itu penulis tertarik menggunakan algoritma tersebut dalam kasus dengan data yang berisi atribut dan record yang berbeda yaitu persetujuan penerimaan kredit di Bank dengan atribut berupa karakteristik debitur yang mempunyai record sesuai dengan historis yang telah tersimpan pada data bank.

II. LANDASAN TEORI

II.1 Kredit

Ikatan Akuntan Indonesia (2004:31.4) mendefinisikan kredit sebagai berikut: Kredit adalah pinjaman uang atau tagihan yang dapat dipersamakan dengan itu berdasarkan persetujuan atau kesepakatan pinjam-meminjam antara bank dan pihak lain yang mewajibkan pihak peminjam untuk melunasi

imbalan, atau pembagian hasil keuntungan. Hal yang termasuk dalam pengertian kredit yang diberikan adalah kredit dalam

pembelian surat berharga nasabah yang dilengkapi dengan Note Purchase Agreement [1].

Atribut dapat dibedakan dalam tipe-tipe yang berbeda bergantung pada tipe nilai yang diterima.berikut ini beberapa jenis atribut beserta contohnya [10].

II.2.1 Atribut Kategorikal

Salah satu tipe yang domainnya merupakan sebuah himpunan

symbol berhingga[10].Contoh : jenis kelamin (L,P), status(Menikah,Belum Menikah).

Atribut kategorikal dibedakan menjadi dua tipe,yaitu :

1. Nominal : sebuah atribut dikatakan nominal jika nilai – nilainya tidak dapat diurutkan. Contoh : Jenis Kelamin, Warna Mata.
2. Ordinal : disebut ordinal jka nilai – nilainya dapat diurutkan dengan beberapa cara, contoh : ranking (rasa dari keripik kentang pada skala 1-10).sifat dari ordinal adalah pembeda dan urutan.

II.2.2 Atribut Numerik

Domain dari tipe numerik adalah berupa bilangan riil atau integer[10].Contoh : Umur dan Gaji = bilangan riil positif.atribut numeric dibedakan menjadi :

1. Interval : mempunyai sifat bahwa perbedaan antara nilai – nilainya sangat berarti. Contoh : tanggal.
2. Rasio : dalam atribut jenis ini, baik beda maupun rasio sangat berarti. Contoh : panjang, waktu, jumlah.

Kita dapat menyatakan umur 20 tahun adalah dua kali lebih tua dari umur 10 tahun.

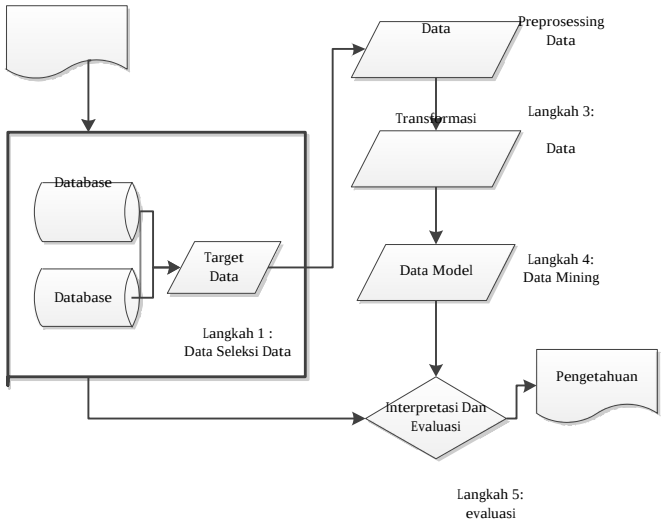
II.2.3 Atribut berdasarkan jumlah nilainya

Berdasarkan nilai,data dapat dikategorikan menjadi 2 yaitu [10]:

1. Atribut diskrit : atribut yang hanya menggunakan sebuah nilai berhingga atau himpunan nilai tak berhingga yang dapat dihitung. Contoh: himpunan kata dalam kumpulan dokumen.
2. Atribut kontinyu : atribut yang menggunakan bilangan riil sebagai nilai atribut. Contoh : suhu, ketinggian atau berat.

II.3 *Knowledge Discovery in Database (KDD)*

Knowledge discoveryin databases (KDD) adalah keseluruhan proses non-trivial untuk mencari dan mengidentifikasi pola (pattern) dalam data, dimana pola yang ditemukan bersifat sah, baru, dapat bermanfaat dan dapat dimengerti.KDD berhubungan dengan teknik integrasi dan penemuan ilmiah, interprestasi dan visualisasi dari pola-pola sejumlah kumpulan data.



Gambar II. 1 Alur Tahapan KDD

II.3.1 Seleksi Data

Tujuan dari seleksi data adalah menciptakan himpunan data target , pemilihan himpunan data, atau memfokuskan pada subset variabel atau sampel data, dimana penemuan (discovery) akan dilakukan.

II.3.2*Pre-processing/* Pembersihan data

Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi). Dilakukan pula proses *enrichment*, yaitu proses memperkaya data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

II.3.3 Transformasi

Data transformation adalah prosses Mentransformasi atau menggabungkan data ke dalam bentuk yang sesuai untuk penggalian lewat operasi *summary* atau *aggregation*.

II.3.4 Data mining

Pemilihan tugas data mining merupakan sesuatu yang penting yaitu berupa pemilihan goal dari proses KDD yang pada kasus ini yaitu klasifikasi apakah seorang debitur layak mendapatkan kredit atau tidak.

II.3.5 Interpretasi/ Evaluasi

Tahap ini merupakan bagian dari proses KDD yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya. Jika pola atau informasi yang dihasilkan masih bertentangan dengan fakta, maka perlu dilakukan pengkajian ulang pada data dan proses yang dilakukan.

II.4 *Decision Tree* (Pohon keputusan)

II.4.1 Perhitungan Data Menjadi Model Tree

Sebelum kita menuju ke arah ekstraksi data ke dalam bentuk model tree, tentumya ada beberapa proses yang harus diperhatikan dalam pembentukan struktur pohon ini, yaitu:

a. Pilih *root* berdasarkan *gain ratio* terbesar
b. Pilih internal *root* /cabang *root* berdasar gain ratio terbesar setelah menghapus atribut yang telah terpilih sebagai *root*
c. Ulangi sampai semua atribut terhitung nilai *gain rationya*.

Parameter yang tepat digunakan untuk mengukur efektifitas suatu atribut dalam melakukan teknik pengklasifikasian sampel data, salah satunya adalah dengan menggunakan *information gain*.Sebelum mencari nilai gain, terlebih dahulu mencari peluang kemunculan suatu *record* dalam atribut (*entropy*)

1. Penghitungan Nilai *Entropy*
   Untuk mendapatkan nilai *information gain*, terlebih dahulu kita harus mengetahui parameter lain yang mempengaruhi nilai gain, dimana parameter ini sangat diperlukan untuk mendapatkan nilai gain. Parameter tersebut adalah *entropy*.Parameter ini sering digunakan untuk mengukur heterogenitas suatu kumpulan sampel data. Secara matematis nilai *entropy* dapat dihitung dengan menggunakan formula sebagai berikut :

$$Entropy(S) = \sum_{i}^{c} - p_i \log 2 p_i \qquad (2.1)$$

   C = jumlah nilai yang ada pada atribut target (jumlah kelas)
   P_i = jumlah sampel pada kelas i

   Dari formula diatas dapat kita cermati bahwa apabila hanya terdapat 2 kelas dan dari kedua kelas tersebut memiliki komposisi jumlah sampel yang sama, maka *entropy*nya = 0.
2. Perhitungan *information gain*
   Ketika kita sudah mendapatkan nilai *entropy*, maka langkah selanjutnya adalah melakukan perhitungan terhadap *information gain*.Berdasarkan perhitungan matematis *information gain* dari suatu atribut A dapat diformulasikan sebagai berikut :

$$Gain(S, A) = Entropy(S) - \sum_{Vevslustion(A)} \frac{|s_v|}{s} entropy(s_v) \qquad (2.2)$$

   A : atribut

   V : menyatakan suatu nilai yang mungkin untuk atribut A
   Values (A) : himpunan nilai-nilai yang mungkin untuk atribut A
   $|S_v|$: jumlah smpel untuk nilai v
   |S| : jumlah seluruh sampel data
   Entropy (S) : entropy untuk asmpel-sampel yang memiliki nilai v

3. Gain ratio
   Untuk menghitung *gain ratio* kita perlu ketahui suatu term baru yang disebut *split information*. *Split information* dihitung dengan formula sebagai berikut:

$$split\ information = -\sum_{t=1}^{c} \frac{si}{s} \log 2 \frac{si}{s} \qquad (2.3)$$

   Dimana :

   S_1 sampai S_c adalah c subset yang dihasilkan dari pemecahan S dengan menggunakan atribut A yang mempunyai banyak C nilai.

   Selanjutnya *gain ratio* dihitung dengan cara

$$gain\ ratio = \frac{Gain\ (S,A)}{split\ information\ (S,A)} \qquad (2.4)$$

II.5 Algoritma C4.5

Berikut ini adalah algoritma C4.5 :
Input : sample training,label training, atribut

- Membuat simpul akar untuk pohon yang dibuat
- Jika semua sample positif, berhenti dengan suatu pohon dengan satu simpul akar, beri label (+)
- Jika semua sample negatif, berhenti dengan suatu pohon degan satu simpul akar, beri label (-)
- Jika atribut kosong, berhenti dengan suatu pohon dengan satu simpul akar, dengan label sesuai nilai yang terbanyak yang ada pada label training
- Untuk yang lain, mulai
  - A ←atribut yang mengklasifikasikan sampel dengan hasil terbantik (berdasarkan gain ratio)

ISSN : 2355-9365 e-Proceeding of Engineering : Vol.2, No.1 April 2015 | Page 1771

- Atribut keputusan untuk simpul akar ← A
- Untuk setiap nilai V_i yang mungkin untuk A

ISSN : 2355-9365 e-Proceeding of Engineering : Vol.2, No.1 April 2015 | Page 1772

- Tambahan cabang dibawah akar yang berhubungan dengan A=V_i
- Tentukan sampel Sv_i sebagai subset dari sampel yang mempunyai nilai untuk V_i untuk atribut A
- Jika sampel SV_i kosong
  - Dibawah cabang tambahkan simpul daun dengan label = niilai yang terbanyak yang ada pada label training
  - Yang lain tambahkan cabang baru dibawah cabang yang sekarang

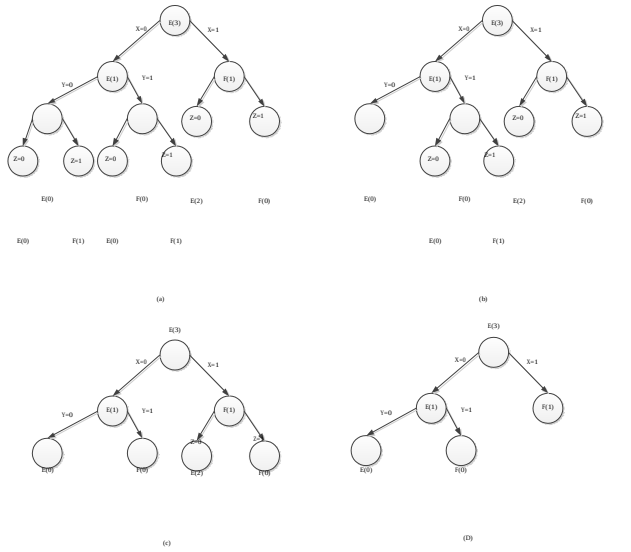
Berhenti

II.6 Pruning

Teknik pruning merupakan teknik yang digunakan untuk menyederhanakan struktur pohon yang telah dibangun oleh algoritma C4.5 diatas. Teknik pruning ini digunakan untuk mengantisipasi banyaknya level dan besarnya (lebatnya) hutan dalam struktur tree[3]. Sehingga struktur tree menjadi lebih sederhana dan distribusi kelas dapat terjaga. Alasan lain kenapa harus dilakukan pruning adalah karena dalam teknik klasifikasi yang akan dijalankan nantinya akan mengeluarkan rule (pola) yang dibentuk berdasarkan struktur tree, jadi jika struktur tree tidak teratur atau kurang sederhana, maka rule yang dihasilkan pun akan rumit untuk diimplementasikan.
Ada beberapa hal yang perlu dilakukan ketika menyederhanakan struktur pohon :

1. Membuat tabel distribusi terpadu dengan menyertakan semua nilai kejadian pada setiap rule
2. Menghitung tingkat independensi antara kriteria pada suatu rule, yaitu antara atribut dan target atribut.
3. Mengeliminasi kriteria yang tidak perlu, yaitu yang tingkat independensinya tinggi.

Berikut ini adalah contoh pemotongan tree :



Gambar II. 2*prune*

1. *Prepruning* yaitu menghentikan pembangunan suatu *subtree* lebih awal, yaitu dengan memutuskan untuk tidak lebih jauh mempartisi *data training*. Pada pendekatan prepruning, sebuah pohon dipangkas dengan cara menghentikan pembangunannya jika partisi yang akan dibuat dianggap tidak signifikan.
2. *Postpruning* yaitu menyederhanakan pohon dengan cara membuang beberapa cabang *subtree* setelah pohon selesai dibangun. Metode *postpruning* ini merupakan metode *standard* untuk algoritma C4.5.

Pemangkasan pohon juga dapat digunakan untuk mengatasi *overfitting*.*Overfitting* terjadi karena ada *noise data training*, yaitu data yang tidak relevan sehingga mengakibatkan pohon memiliki subtree yang panjang dan tidak seimbang. Misal internal node memiliki kelas YA = 5 dan TIDAK = 1. Data yang berada pada kelas TIDAK merupakan *noise*, sehingga apabila data tersebut diolah akan menghasilkan pohon dengan subtree yang panjang. *Overfitting* juga dapat terjadi karena *data training* yang sedikit.
Langkah pemangkasan pohon :

1. hitung *Pessimistic error rate* parent
2. hitung *Pessimistic error rate* child
3. jika *Pessimistic error rate* pada child > parent, maka lakukan *pruning*
4. jika *Pessimistic error rate* pada child<parent, maka lakukan lanjutkan *spilt*

Rumus Pessimistic error rate untuk prepruning:

$$e = \frac{r + \frac{z^2}{2n} + z\sqrt{\frac{r}{n} - \frac{r^2}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \qquad (2.5)$$

Dimana:
r = nilai perbandingan *error rate*
n = total *sample*

$$z = \Phi^{-1}(c)$$

c = *confidence level*

II.7 Performansi

Untuk mengevaluasi informasi dari sistem dapat dilakukan dengan cara menghitung akurasi sistem berdasar inputan data training dan data uji. Berikut ini adalah tabel *confusion matrix* dalam dua kelas diterima dan ditolak :

Untuk permasalahan dalam klasifikasi, pengukuran yang biasa digunakan adalah *precision, recall* dan *accuracy*. Karena kelayakan kredit merupakan *binary classification*, maka *precision*, *recall* dan *accuracy* dapat dihitung dengan cara seperti pada Tabel II.1.

ISSN : 2355-9365 e-Proceeding of Engineering : Vol.2, No.1 April 2015 | Page 1773

Ada dua metode dalam melakukan pemangkasan dalam pohon keputusan, yaitu :

Tabel II.1 Tabel Penilaian

	Diidentifikasisebagai tidak layak	Diidentifikasi sebagai layak
Tidak Layak	a	b
Layak	c	d

1. *Precision*

Precision adalah bagian data yang di ambil sesuai dengan informasi yang dibutuhkan. Rumus *precision* adalah

$$precision = \frac{(d)}{(b+d)} \times 100\% \qquad (2.6)$$

Dalam klasifikasi binari, *precision* dapat disamakan dengan *positive predictive value* atau nilai prediktif yang positif .

2. *Recall*

Recall adalah pengambilan data yang berhasil dilakukan terhadap bagian data yang relevan dengan *query*. Rumus *Recall*adalah :

$$recall = \frac{(d)}{(c+d)} \times 100\% \qquad (27)$$

Dalam klasifikasi binari, *recall* disebut juga dengan *sensitivity*. Peluang munculnya data relevan yang diambil sesuai dengan query dapat dilihat dengan *recall*.proporsi kasus positif yang diidentifikasi dengan benar.

3. *Accuracy*

Accuracy adalah persentase dari total data ujicoba yang benar diidentifikasi. Rumus *Accuracy*adalah :

$$accuracy = \frac{(a+d)}{(totalsample)} \times 100\% \qquad (2.8)$$

III. ANALISIS DAN PERANCANGAN

III.1 Analisis Data

Berikut ini adalah penjelasan tentang penggunaan data pada tugas akhir ini, dimana data yang digunakan mempunyai beberapa tipe data yang berbeda-beda pada atributnya.

Tipe data yang dapat digunakan dalam teknik klasifikasi adalah kategorikal dan diskrit, maka dari itu selain dari jenis tipe data tersebut harus diubah menjadi kategorikal dan diskrit. Berikut ini spesifikasi atribut yang di gunakan dalam data analisis kelayakan kredit oleh debitur:

Tabel III. 1 Tabel atribut

Atribut	Keterangan	kode atribut
Atribut 1	Status rekening giro	A1
Atribut 2	Jangka waktu (dalam sebulan)	A2
Atribut 3	Sejarah kredit	A3
Atribut 4	Jumlah kredit	A5
Atribut 5	Rekening tabungan / obligasi	A6
Atribut 6	Lama kerja	A7
Atribut 7	Status pribadi dan jenis kelamin	A9
Atribut 8	Lama domisili	A11
Atribut 9	Kepemilikan	A12
Atribut 10	Rencana angsuran lain	A14
Atribut 11	Jumlah kredit yang sudah ada di bank	A16
Atribut 12	Jumlah orang yang bertanggung jawab	A18
Atribut 13	Telepon	A19
Atribut 14	class	Yes/No

III.1.1Kelompok Atibut Kategorikal

Salah satu tipe yang domainnya merupakan sebuah himpunan simbol berhingga.

Atribut ke 3 :Sejarah kredit
Atribut ke 7 :Status pribadi dan jenis kelamin
Atribut ke 9 :Kepemilikan
Atribut ke 10: Rencana angsuran lain
Atribut ke 11: Jumlah kredit yang sudah ada di bank
Atribut ke 13: Telepon

Untuk atribut atribut yang memiliki karakteristik kategorikal tidak dilakukan perubahan data karena sudah dapat dikenali dalam proses klasifikasi.

III.1.2Kelompok Atribut Numerik

Domain dari tipe numeric adalah berupa bilangan riil atau integer.

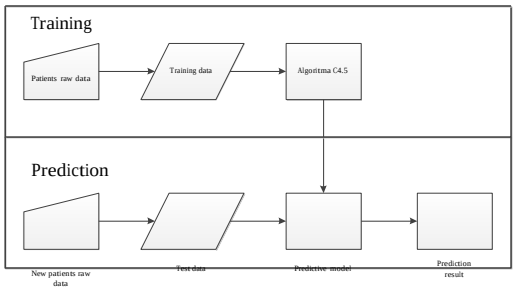
Pada aribut dengan jenis numerik, maka perlu dirubah agar terkategorisasi. Cara pengkategorian data pada Tugas Akhir kali ini yaitu oleh *expert*/ orang yang mempunyai ilmu dibidang perkreditan yang ada di bank tersebut. Berikut ini adalah tetapan diskretisasi oleh *expert*.

Atribut ke 1 :Status rekening giro yang ada
Atribut ke 2 :Jangka waktu (dalam sebulan)
Atribut ke 4 :Jumlah kredit
Atribut ke 5 :Rekening tabungan / obligasi
Atribut ke 6 :Lama kerja
Atribut ke 8 :Lama domisili
Atribut ke 12: Jumlah orang yang bertanggung jawab

ISSN : 2355-9365 e-Proceeding of Engineering : Vol.2, No.1 April 2015 | Page 1775

III.2 Model Perancangan

Model perancangan merupakan bagan yang menunjukkan alur kerja atau apa yang sedang dikerjakan di dalam sistem secara keseluruhan dan menjelaskan urutan dari prosedur-prosedur yang ada di dalam sistem.

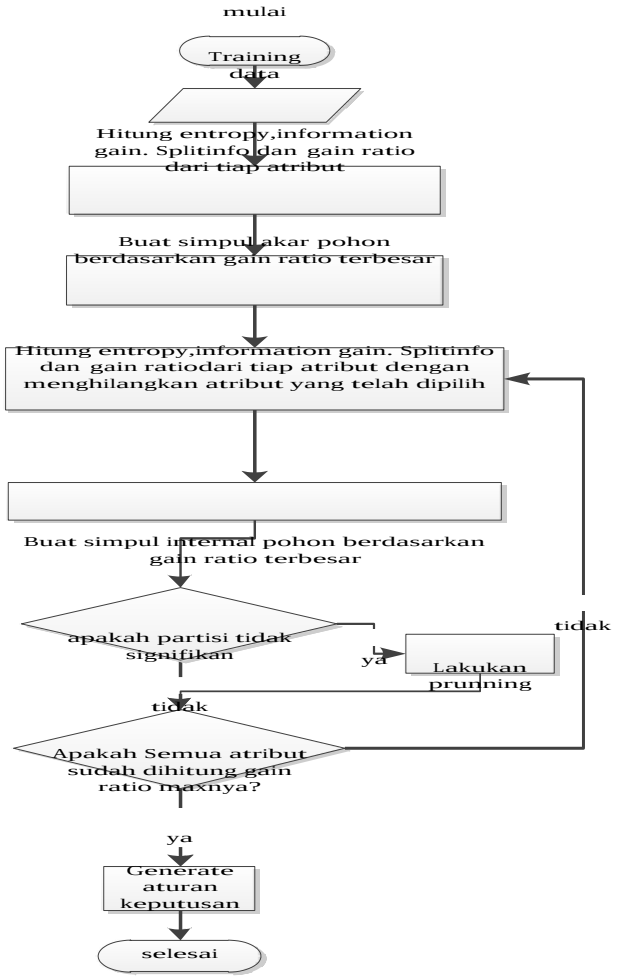


Keterangan proses :

1. Training : Barisan data mentah yang telah melalui proses KDD kemudian menjadi sebuah data latih yang kemudian diproses mengguanakan algoritma C4.5
2. Prediction : Barisan data baru yang menjadi data test untuk memprediksi dan menguji rule yang terbentuk dari keluaran algoritma C4.5 yang akan menghasilkan sebuah prediksi.

III.3 Pohon keputusan C4.5

Berikut ini alur algoritma C4.5 yang akan menghasilkan rule.



Gambar III. 2*Flowchart* Pohon Keputusan C4.5

Berikut keterangan dari tiap proses:

1. Data training dimasukkan.
2. Hitung *Gain Ratio*, *Split Info* dan *entropy* dari masing-masing atribut data training yang ada.
3. Buat simpul akar dari pemilihan atribut yang memiliki *Gain Ratio* terbesar.
4. Hitung *Gain Ratio*, *Split Info* dan *entropy* dari masing-masing atribut dengan menghilangkan atribut yang telah dipilih sebelumnya.
5. Buat simpul *internal* dari pemilihan atribut yang memiliki *Gain Ratio* terbesar.
6. Lakukan pemotongan pohon (*prunning*)
7. Cek apakah semua atribut sudah dibentuk pada pohon. Jika belum, maka ulangi proses d dan e, jika sudah maka lanjut pada proses berikutnya.
8. Lakukan pemangkasan pohon untuk menghilangkan cabang-cabang yang tidak perlu.
9. Kemudian aturan keputusan di*generate* mengikuti pohon yang telah dibentuk sebelumnya.

IV. IMPLEMENTASI SISTEM

IV.1 Data yang digunakan

Data yang digunakan dalam tugas akhir ini merupakan data *credit approval* data atribut yang diperoleh dari Bank Pasar Daerah Istimewa Yogyakarta. Data merupakan data

kategorikal dan tidak ada *mising value* pada data. Jumlah data yang digunakan sebanyak 1000 data *record* yang diambil dalam rentang watu Januari sampai dengan Juli 2014.

Data juga dipecah menjadi 4 dipartisi, yaitu:partisi 90% data *training* :10% data ujicoba, partisi 80% data *training* :20% data ujicoba, partisi 70% data *training* :30% data ujicoba dan partisi 60% data *training* :40% data ujicoba. Masing-masing pasrtisi dicoba menggunakan nilai confidence yang berbeda yaitu menggunakan nilai sebesar 85%,90%,95%, dan 98%.

IV.2 Proses mining C4.5

Dalam proses mining C4.5, proses yang dilakukan adalah sebagai berikut:

1. Hitung frekuensi kemunculan masing-masing nilai atribut pada data survey.
2. Hitung nilai *Entropy* dari masing-masing nilai atribut.
3. Hitung nilai *Information Gain* dengan menggunakan nilai *Entropy* yang telah dihitung sebelumnya.
4. Hitung nilai *Split Info* dari tiap atribut.
5. Hitung nilai *Gain Ratio* menggunakan nilai *Information Gain* dan *Split Info*.
6. Ambil nilai *Gain Ratio* terbesar dan jadikan simpul akar.
7. Hilangkan atribut yang dipilih sebelumnya dan ulangi perhitungan nilai *Entropy*, *Information Gain*, *Split Info* dan *Gain Ratio* dengan memilih *Gain Ratio* terbesar dan dijadikan simpul *internal* pohon.
8. Ulangi perhitungan tersebut hingga semua atribut pohon memiliki kelas.
9. Jika semua pohon sudah memilik kelas, maka tampilkan pohon keputusan awal dan generate aturan keputusan awal.

Berikut contoh Perhitungan C4.5 yang digenerate oleh aplikasi:

C4.5 » Perhitungan C4.5

Menu: Perhitungan C4.5 | Lakukan Mining C4.5 | Pohon Keputusan C4.5

Opsi: Hapus Semua Data

NO	ATRIBUT GAIN RATIO MAX	ATRIBUT	NILAI ATRIBUT	JUMLAH KASUS TOTAL	JUMLAH KASUS YES	JUMLAH KASUS NO	ENTROPY	INFORMATION GAIN	SPLIT INFO	GAIN RATIO
1	status_rekening_giro	Total	Total	800	561	239	0.8798			0
2	status_rekening_giro	status_rekening_giro	A11	209	110	99	0.998	0.0939	1.8166	0.0517
3	status_rekening_giro	status_rekening_giro	A12	226	133	93	0.9773	0.0939	1.8166	0.0517
4	status_rekening_giro	status_rekening_giro	A13	55	43	12	0.7668	0.0939	1.8166	0.0517
5	status_rekening_giro	status_rekening_giro	A14	310	275	35	0.5086	0.0939	1.8166	0.0517
6	status_rekening_giro	jangka_waktu	A21	53	25	28	0.9977	0.0118	0.3518	0.0335
7	status_rekening_giro	jangka_waktu	A22	747	536	211	0.8588	0.0118	0.3518	0.0335
8	status_rekening_giro	sejarah_kredit	A30	33	13	20	0.9673	0.0379	1.7095	0.0222

Gambar IV. 1Perhitungan C4.5

Berikut contoh pohon keputusan C4.5 dan aturan partisi yang digenerate oleh aplikasi, lebih lengkapnya berada di lampiran:

C4.5 » Pohon Keputusan

Menu: Perhitungan C4.5 | Lakukan Mining C4.5 | Pohon Keputusan C4.5

Opsi: Hapus Semua Data

Pohon Keputusan:

```
status_rekening_giro = A11 (Yes = 110, No = 99) : ?
|  jangka_waktu = A21 (Yes = 2, No = 10) : No
|  jangka_waktu = A22 (Yes = 108, No = 89) : ?
|  |  sejarah_kredit = A33 (Yes = 2, No = 7) : ?
|  |  |  rekening_tabungan_obligasi = A61 (Yes = 1, No = 7) : No
|  |  |  rekening_tabungan_obligasi = A62 (Yes = 0, No = 0) : No
|  |  |  rekening_tabungan_obligasi = A63 (Yes = 0, No = 0) : No
|  |  |  rekening_tabungan_obligasi = A64 (Yes = 0, No = 0) : No
|  |  |  rekening_tabungan_obligasi = A65 (Yes = 1, No = 0) : Yes
|  |  sejarah_kredit = A32 (Yes = 60, No = 55) : ?
|  |  |  jumlah_kredit_yang_ada_di_bank = A162 (Yes = 1, No = 4) : No
|  |  |  jumlah_kredit_yang_ada_di_bank = A161 (Yes = 59, No = 51) : Yes
|  |  |  jumlah_kredit_yang_ada_di_bank = A163 (Yes = 0, No = 0) : Yes
|  |  |  jumlah_kredit_yang_ada_di_bank = A164 (Yes = 0, No = 0) : Yes
|  |  sejarah_kredit = A31 (Yes = 5, No = 9) : No
|  |  sejarah_kredit = A30 (Yes = 3, No = 6) : No
|  |  sejarah_kredit = A34 (Yes = 38, No = 12) : Yes
```

Gambar IV. 2 Pohon Keputusan

IV.3 Penentu keputusan

Penentu keputusan adalah proses pencocokan aturan keputusan C4.5 dengan data survey yang belum memiliki keputusan atau kelas, tujuannya agar data survey tersebut memiliki keputusan.

Data yang akan diberi keputusan diinput melalui form input data survey, lalu dari data tersebut dibaca nilai atribut dari masing-masing atribut kemudian dilakukan pencocokan aturan atau rule keputusan C4.5.

Pertama adalah dengan mengisikan form penentu keputusan

Penentu Keputusan

Input Data Survey	
Nama Pemohon	:
Status Rekening Giro	: A11
Jangka Waktu	: A21
Sejarah Kredit	: A30
Jumlah Kredit Plavon	: A51
Rekening Tabungan Obligasi	: A61
Lama Kerja	: A71
Status Pribadi dan Jenis Kelamin	: A91
Lama Domisili	: A111
Kepemilikan	: A121
Rencana Angsuran Lain	: A141
Jumlah Kredit yang ada di Bank	: A161
Jumlah Orang yang bertanggung Jawab	: A181
Telepon	: A191
Input	

Gambar IV. 3 Form Penentu Keputusan

Setelah input, program akan mengeluarkan hasil dari data latih yang telah dilakukan mining sebelumnya

NO	NAMA PEMOHON	STATUS REKENING GIRO	JANGKA WAKTU	SEJARAH KREDIT	JUMLAH KREDIT PLAVON	REKENING TABUNGAN OBLIGASI	LAMA KERJA	STATUS PRIBADI DAN JENIS KELAMIN	LAMA DOMISILI	KEPEMILIKAN	RENCANA ANGSURAN LAIN	JUMLAH KREDIT YANG SUDAH ADA DI BANK	JUMLAH ORANG YANG BERTANGGUNG JAWAB	TELEPON
1	jalaludin	A11	A22	A31	A52	A63	A74	A91	A113	A123	A143	A163	A182	A191

Gambar IV. 4 Inputan

Mengeluarkan hasil keputusan

Gambar IV. 5 Hasil Keputusan

IV.4 Analisa hasil pohon keputusan

Setelah pohon dibentuk, selanjutnya dilakukan perbandingan data yang merupakan data ujicoba dimana data tersebut dilakukan pengklasifikasian menggunakan C4.5 yang telah dibentuk. Kemudian kelas yang terbentuk dibandingkan dan dihitung nilai *error rate*nya.

Hasil klasifikasi algoritma C4.5 dapat dilihat di tabel penilaian berikut:

Tabel IV. 1 Tabel Penilaian C4.5

	cofidence level	Partisi training			
		60	70	80	90
Akurasi	85%	71,25	73,33	74,5	73
	90%	71,25	73,33	74,5	73
	95%	72	73,33	72,5	73
	98%	72,25	70,67	72,5	71

Tabel IV. 2 Tabel Penilaian C4.5 100% data *training*

	Diidentifikasi No oleh C4.5	Diidentifikasi Yes oleh C4.5
Keputusan Asli: No = 300	125	175
Keputusan Asli: Yes = 700	67	633

1. Precision = 633 / (175 + 633) * 100% = 78.34 %
2. Recall = 633 / (67 + 633) * 100% = 90.43 %
3. Accuracy = (125 + 633) / (700 + 300) * 100% = 75.8 %

Berikut adalah tabel klasifikasi dengan kelompok akurasi terbesar yaitu menggunakan nilai confidence 90%.

Tabel IV. 3 Tabel Penilaian C4.5 Partisi Data 90%:10%

	Diidentifikasi No oleh C4.5	Diidentifikasi Yes oleh C4.5
Keputusan Asli: No = 32	16	16
Keputusan Asli: Yes = 68	11	57

1. Precision = 57 / (16 + 57) * 100% = 78.08 %
2. Recall = 57 / (11 + 57) * 100% = 83.82 %
3. Accuracy = (16 + 57) / (68 + 32) * 100% = 73 %

Tabel IV. 4 Tabel Penilaian C4.5 Partisi Data 80%:20%

	Diidentifikasi No oleh C4.5	Diidentifikasi Yes oleh C4.5
Keputusan Asli: No = 61	15	46
Keputusan Asli: Yes = 139	5	134

1. Precision = 134 / (46 + 134) * 100% = 74.44 %
2. Recall = 134 / (5 + 134) * 100% = 96.4 %
3. Accuracy = (15 + 134) / (139 + 61) * 100% = 74.5 %

Tabel IV. 5 Tabel Penilaian C4.5 Partisi Data 70%:30%

	Diidentifikasi No oleh C4.5	Diidentifikasi Yes oleh C4.5
Keputusan Asli: No = 93	21	72
Keputusan Asli: Yes = 207	8	199

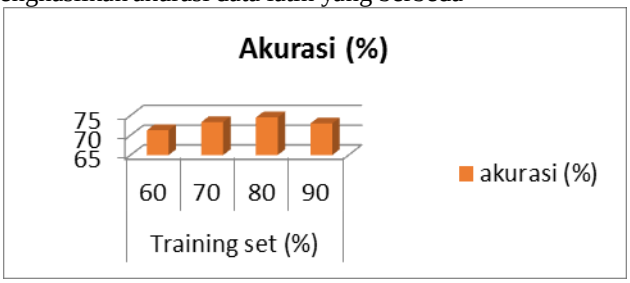
1. Precision = 199 / (72 + 199) * 100% = 73.43 %
2. Recall = 199 / (8 + 199) * 100% = 96.14 %
3. Accuracy = (21 + 199) / (207 + 93) * 100% = 73.33 %

Tabel IV. 6 Tabel Penilaian C4.5 Partisi Data 60%:40%

	Diidentifikasi No oleh C4.5	Diidentifikasi Yes oleh C4.5
Keputusan Asli: No = 123	24	99
Keputusan Asli: Yes = 277	16	261

1. Precision = 261 / (99 + 261) * 100% = 72.5 %
2. Recall = 261 / (16 + 261) * 100% = 94.22 %
3. Accuracy = (24 + 261) / (277 + 123) * 100% = 71.25 %

Dapat kita lihat dari beberapa percobaan partisi data diatas, menghasilkan akurasi data latih yang berbeda



Gambar IV. 6 Akurasi Data

V. KESIMPULAN DAN SARAN

V.1 Kesimpulan

Dari pengukuran kinerja algoritma yang telah dilakukan, dapat disimpulkan algoritma C4.5 memiliki kinerja (*precision*, *recall*, dan *accuracy*) sesuai jumlah data latih dan data test . Dari percobaan yang telah dilakukan,yang memiliki nilai *accuracy* dan *recall* tertinggi pada partisi data 80%:20% dengan nilai confidence 90%.

Partisi data 80%:20% merupakan partisi terbaik karena memiliki nilai *accuracy* dan *recall* yang paling tinggi daripada partisi lainnya yaitu dengan *accuracy* 74,5% dan *recall* 96,4% , akan tetapi perlu diperhatikan dari nilai *precision* terbesar pada partisi 90%:10% sebesar 78,08%, hal ini dikarenakan beberapa faktor, faktor yang paling mempengaruhi adalah data latih yang banyak menyebabkan kondisi

dan rule yang terbentuk dapat lebih menangani variasi dari data test.

Akan tetapi perlu diperhatikan bahwa tidak selalu data latih yang besar membuat *accuracy,precision,* dan *recall* semakin tinggi, tergantung dengan kualitas data yang dijadikan sebagai data latih.

V.2 *Saran*

Saran-saran yang bisa disampaikan adalah sebagai berikut:

1. Aplikasi ini masih bisa dikembangkan untuk algoritma pohon keputusan lainnya dan untuk metode *pruning* yang digunakan juga masih bisa dikembangkan lagi.
2. Algoritma C4.5 pada aplikasi ini tidak bisa mengklasifikasi data yang mengandung *missing value* atau data *numeric*, sehingga dapat lebih disempurnakan lagi.
3. Aplikasi ini masih bisa dioptimasi lagi pada algoritmanya sehingga dapat mempercepat proses *mining* dan proses penentu keputusan.

DAFTAR PUSTAKA

[1] Hidayati, Ery. 2003. Sistem Pendukung Keputusan Berbasis Logika Fuzzy Untuk Analisis Kelayakan Kredit. Fakultas Matematika dan Ilmu Pengetahuan Alam : Institut teknologi sepuluh nopember

[2] Moertini, Veronica S. 2003. *Towards The Use Of C4.5 Algorithm For Classifying Banking Dataset*. Universitas katolik parahyangan : Bandung.

[3] Santosa, Budi. Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis.2007. Graha Ilmu : Yogyakarta.

[4] Feri, Sulianta, & Dominikus,juju. 2010. "Data Mining : Meramalkan Bisnis Perusahaan", PT.Elex Media Komputindo Jakarta.

[5] Kusrini & luthfi, E.T. 2009. *Algoritma Data Mining*. Yogyakarta:Andi publishing.

[6] Basuki Achmad dan Iwan Syarif. 2003. Decision Tree. Politeknik Elektronika Negeri Surabaya.

[7] Dwiantara, L., & Hadi, R. 2004. *Manajemen Logistik: Pedoman Praktis Bagi Sekretaris dan Staf Administrasi*. Jakarta: PT.Grasindo

[8] Lambert M., Stock R. 2001. *Strategic Logistics Management*. New York: McGraw-Hill.

[9] Wysocki R.K. 2006. *Effective Software Project Management*. Willey: Indianapolis.

[10] Hermawati, Fajar Astuti. Data Mining.2013.ANDI: Yogyakarta.