PERBANDINGAN PERFORMANSI ALGORITMA DECISION TREE C5.0, CART, DAN CHAID: KASUS PREDIKSI STATUS RESIKO KREDIT DI BANK X

Yogi Yusuf W.
Jurusan Teknik Industri, Universtias Katolik Parahyangan
Jalan Ciumbuleuit 94 Bandung
e-mail: yogi@home.unpar.ac.id

ABSTRAKSI

Perkembangan teknologi informasi yang pesat telah mempengaruhi cara penilaian resiko kredit yang semula dengan cara human judgment bergeser ke arah cara yang formal dan objektif yaitu melalui credit scoring. Banyak teknik yang dapat membantu dalam pembangunan model credit scoring. Pada perkembangan terbaru, teknik-teknik yang terdapat di dalam data mining mulai banyak digunakan khususnya teknik decision tree telah menjadi teknik yang populer. Ada beberapa algoritma decision tree yaitu C5.0, CART, dan CHAID yang dapat digunakan untuk membangun model tree. Ketiga algoritma tersebut menghasilkan model tree yang berbeda untuk set data yang sama. Model yang berbeda dapat memberikan keakuratan yang berbeda pula. Algoritma C5.0 memberikan rata-rata tingkat keakuratan sebesar 87,72%, CART 87,27%, dan CHAID 87,15%. Dari analisis statistik diperoleh bahwa tidak ada perbedaan performansi yang signifikan di antara ketiga algoritma tersebut.

***Kata kunci**: Credit Scoring, Data Mining, Decision Tree, Algoritma C5.0, CART, CHAID*

1. PENDAHULUAN

Bank mempunyai peranan yang penting dalam pembangunan. Melalui perbankan dana masyarakat dapat dihimpun melalui tabungan, deposito, giro dan selanjutnya disalurkan kembali ke pihak-pihak yang membutuhkan dana dalam bentuk pinjaman. Di Indonesia bank dibagi ke dalam dua jenis (Jusuf, 1992), yaitu:

a. Bank Umum yaitu bank yang dapat memberikan lalu lintas pembayaran. Bank umum dapat menkhususkan diri untuk melaksanakan kegiatan tertentu atau memberikan perhatian yang lebih besar kepada kegiatan tertentu atau memberikan perhatian yang lebih besar kepada kegiatan tertetnu, seperti pembiayaan jangka panjang dan pembiayaan untuk mengembangkan koperasi dan golongan ekonomi lemah.
b. Bank Perkreditan Rakyat yaitu bank yang menerima simpanan hanya dalam bentuk deposito berjangka, tabungan, dan/atau bentuk lainnya yang dipersamakan dengan itu, seperti memberikan kredit, menyediakan biaya bagi nasabah berdasarkan prinsip bagi hasil. Bank Perkreditan Rakyat menempatkan dananya dalam bentuk Sertifikat Bank Indonesia, deposito berjangka, sertifikat deposito, dan/atau tabungan pada bank lain.

Bank mempunyai peranan yang esensial dalam penyaluran kredit kepada pihak-pihak yang membutuhkan. Fungsi pokok kredit yaitu memenuhi pelayanan terhadap kebutuhan masyarakat dalam rangka memperlancar perdagangan, produksi, dan jasa-jasa bahkan konsumsi yang kesemuanya itu ditujukan untuk meningkatkan kesejahteraan manusia. Salah satu unsur dalam kredit adalah adanya janji dan kesanggupan membayar dari debitur kepada kreditur (Firdaus, 2004).

Pembayaran hutang dari debitur kepada kreditur tidak selamanya berjalan sesuai dengan perjanjian. Ketidaklancaran pembayaran hutang oleh debitur dapat memunculkan kredit macet. Kredit macet ini dapat disebabkan oleh faktor eksternal seperti kondisi ekonomi yang tidak kondusif dan debitur yang nakal, atau faktor internal yaitu kekurangmampuan pihak bank dalam menilai resiko calon debitur. Faktor eksternal sulit dikontrol oleh pihak bank, sementara faktor internal dapat dikontrol oleh pihak bank.

Perkembangan teknologi informasi yang pesat telah mempengaruhi cara penilaian resiko yang semula dengan cara *human judgment* bergeser ke arah cara yang formal dan objektif yaitu melalui *credit scoring*. Tujuan dari *credit scoring* ini adalah membantu pihak penyedia kredit mengkuantifikasi resiko finansial sehingga keputusan dapat diambil dengan cepat dan lebih akurat (Chye, 2004).

Banyak teknik yang dapat membantu dalam pembangunan model *credit scoring*. Pada perkembangan terbaru, teknik-teknik yang terdapat di dalam *data mining* mulai banyak digunakan. Khususnya teknik *decision tree* telah menjadi teknik yang populer karena *decision tree* yang dihasilkan mudah diinterpretasikan dan divisualisasikan (Chye, 2004).

Ada beberapa algoritma *decision tree* yaitu C5.0, CART, dan CHAID yang dapat digunakan untuk membangun model *tree*. Ketiga algoritma tersebut menghasilkan model *tree* yang berbeda untuk set data yang sama. Model yang berbeda dapat memberikan keakuratan yang berbeda pula. Makalah ini membahas perbandingan performansi model *tree*

yang dihasilkan oleh algoritma C5.0, CART, dan CHAID pada masalah *credit scoring*.

2. CREDIT SCORING

Banyaknya permohonan kredit menuntut kreditor harus mampu mengevaluasi pemohon kredit dengan objektif, akurat, dan konsisten. Evaluasi tersebut dapat dibantu dengan *credit scoring*. Isac mendefinisikan *credit scoring* sebagai *tool* yang melibatkan penggunaan model statistik untuk mengevaluasi seluruh informasi yang tersedia dengan objektif dalam pengambilan keputusan kredit (Noe, 1997).

Manfaat yang dapat diperoleh dari penerapan *credit scoring* adalah peningkatan kecepatan dan konsistensi proses aplikasi pinjaman dan memungkinkan otomatisasi proses peminjaman (Chye, 2004); adanya kemampuan belajar sepanjang waktu karena model *credit scoring* didasarkan pada perhitungan statistik dari data masa lalu (Glassman, 1997).

Penerapan model statistik prediktif membutuhkan dua faktor (Glassman, 1997):

1. Teknologi yang memungkinkan model bekerja dengan cepat sehingga kecepatan proses memberikan waktu respon yang dapat diterima, dan
2. Basis data yang menyediakan input bagi model prediktif.

Seiring dengan kemajuan teknologi informasi yang sangat pesat disertai dengan harga yang semakin terjangkau, sangat dimungkinkan bagi perusahaan kecilpun menggunakan model statistik dalam mengevaluasi kredit.

Model *credit scoring* dibangun dengan menggunakan sampel kredit masa lalu dalam jumlah yang besar. Sampel tersebut dibagi kedalam dua kelas yaitu kredit yang baik (pembayaran dilakukan tepat waktu) dan kredit yang bermasalah (pembayaran dilakukan tidak tepat waktu atau tidak dapat melakukan pembayaran). Berdasarkan pola masa lalu, kombinasi karakteristik peminjam yang membedakan peminjam yang baik dan yang buruk menghasilkan skor sebagai estimasi resiko dari tiap peminjam baru.

3. DECISION TREE

Decision tree merupakan salah satu teknik yang dapat digunakan untuk melakukan klasifikasi terhadap sekumpulan objek atau *record*. Teknik ini terdiri dari kumpulan *decision node*, dihubungkan oleh cabang, bergerak ke bawah dari *root node* sampai berakhir di *leaf node*. Pengembangan *decision tree* dimulai dari *root node*, berdasarkan konvensi ditempatkan di bagian atas diagram *decision tree*, semua atribut dievaluasi pada *decision node*, dengan tiap *outcome* yang mungkin menghasilkan cabang. Tiap cabang dapat masuk baik ke *decision node* yang lain ataupun ke *leaf node*.

Persyaratan yang harus dipenuhi dalam penerapan algoritma *decision tree*:

1. Algoritma *decision tree* merepresentasikan *supervised learning*, dan oleh karena itu membutuhkan variabel target *preclassified*.
2. *Training data set* harus kaya dan bervariasi.
3. Kelas atribut target harus diskrit.

4. CART (Classification and Regression Trees)

Pada permulaan proses, *training set* yang terdiri dari record yang sudah diklasifikasikan harus tersedia. *Training set* digunakan untuk membangun *tree* yang memungkinkan penempatan suatu kelas ke dalam variabel target dari record baru yang didasarkan pada nilai-nilai variabel yang lain atau variabel indepeden.

CART membangun *binary tree* dengan memecah record pada tiap node berdasarkan fungsi variabel input tunggal. Tugas pertama yang dijalankan adalah menentukan variabel independen yang menjadi *splitte*r terbaik. *Splitter* terbaik adalah *splitter* yang menurunkan keanekaragaman set record dengan jumlah penurunan terbesar.

Split awal menghasikan dua node. Pada tiap node, *splitter* terbaik berikutnya akan dicari. Proses pemecahan ini terus dilakukan sampai tidak ada pemecahan yang secara signifikan menurunkan keanekargaman node. Node yang tidak dipecah lagi disebut *leaf node*.

Pemecahan record pada tiap node menyebabkan jumlah record yang semakin kecil dari *root node* ke *child node* sampai ke *leaf node*. Semakin sedikit jumlah record semakin kurang representatif node tersebut. Akibatnya adalah model *tree* hanya dapat memprediksi secara akurat untuk record yang berada pada *training set*, tetapi tidak dapat memprediksi record baru yang berasal dari luar *training set* secara akurat atau *overtraining*. Untuk mengurangi *overtraining*, pemangkasan pohon atau *pruning* dapat dilakukan. *Pruning* menghasilkan beberapa kandidat *subtree*.

Beberapa kandidat *subtree* dipilih berdasarkan kemampuannya dalam memprediksi record baru. Pemilihan tersebut membutuhkan set data baru yaitu *test set* yang berisi record baru yang berbeda dengan record yang ada pada *training set*. Tiap kandidat *subtree* digunakan untuk memprediksi record yang ada dalam *test set*. *Subtree* yang memberikan eror terkecil terpilih sebagai model *tree*.

Langkah terakhir adalah mengevaluasi *subtree* terpilih dengan menerapkannya pada set data baru yaitu *validation set*. Nilai eror yang diperoleh dari *validation set* digunakan untuk memprediksi *expected performance* model prediksi.

5. C5.0

C5.0 menghasilkan *tree* dengan jumlah cabang per node bervariasi. C5.0 memperlakukan variabel kontinyu sama dengan yang dilakukan oleh CART, tetapi untuk variabel kategorikal C5.0 memperlakukan nilai variabel kategorikal sebagai *splitter*.

Strategi pengembangan *decision tree* dengan menggunakan alogritma C5.0 adalah sebagai berikut:

1. Pada tahap awal, *tree* digambarkan sebagai *node* tunggal yang merepresentasikan *training set*.
2. Jika sampel seluruhnya berisi kelas yang sama, maka *node* tersebut menjadi *leaf* dan dilabeli dengan kelas tersebut.
3. Jika tidak, algoritma dengan menggunakan ukuran berbasis entropi (*information gain*) akan memilih variabel prediktor yang akan memisahkan record ke dalam kelas-kelas individual. Variabel tersebut menjadi variabel tes atau keputusan pada node tersebut.
4. Cabang dikembangkan untuk tiap nilai yang diketahui dari variabel tes, dan sampel dipartisi berdasarkan cabang tersebut.
5. Algoritma menggunakan proses yang sama secara rekursif membentuk *decision tree*.
6. Partisi rekursif berakhir hanya ketika satu dari kondisi-kondisi berikut terpenuhi:
   a. Seluruh record pada *node* tertentu memiliki kelas yang sama.
   b. Tidak ada atribut yang tersisa pada record yang dapat dipartisi lebih lanjut. Dalam kasus ini suara mayoritas digunakan. *Node* tersebut menjadi *leaf node* dan dilabeli dengan kelas yang menjadi mayoritas dalam record yang ada.
   c. Tidak ada record untuk cabang variabel tes. Dalam kasus ini, *leaf* terbentuk dengan mayoritas kelas sebagai label record tersebut.

6. CHAID (Chi-Squared Automatic Interaction Detection)

Pembangunan *tree* dengan CHAID berbeda dengan CART dan C5.0. CART dan C5.0 membangun *tree* dengan *overfitting* data, kemudian melakukan *pruning*, CHAID akan menghentikan pembangunan *tree* sebelum *overfitting* terjadi.

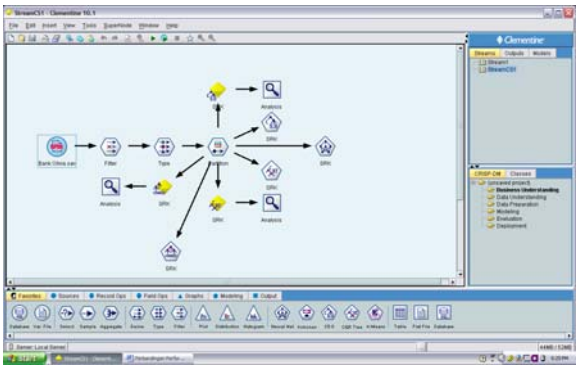
Tiap variabel prediktor dipertimbangkan sebagai *splitter*. Tahap pertama dalam investigasi ini adalah menggabungkan kategori-kategori yang berkorespondensi dengan nilai variabel target yang sama. Seluruh variabel prediktor yang tidak menghasilkan perbedaan yang signifikan dalam nilai variabel target digabung.

Dalam tahap kedua, tiap grup dari tiga atau lebih prediktor dipecah kembali dengan seluruh pembagian biner yang mungkin. Jika pemecahan ini menghasilkan perbedaan yang signifikan, maka pemecahan tersebut dipertahankan.

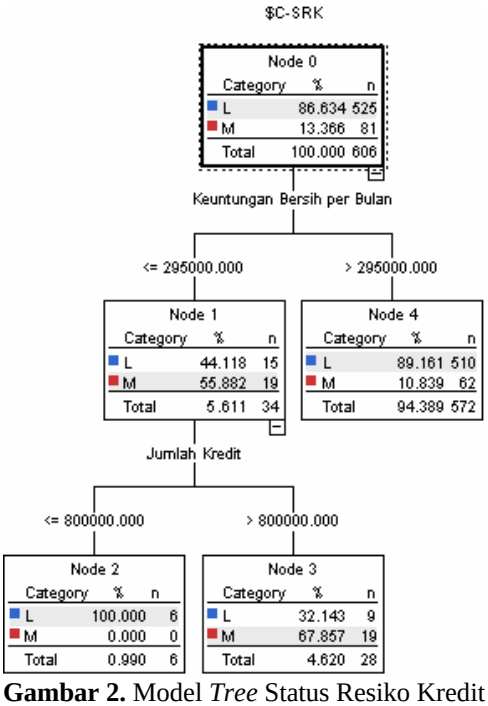
Setelah tiap variabel prediktor dikelompokan untuk menghasilkan keanekargaman kelas yang maksimum dalam variabel target, tes x^2 diterapkan pada kelompok tersebut. Prediktor yang menghasilkan kelompok-kelompok yang paling berbeda dipilih sebagai *splitter* pada node tersebut.

7. PENGEMBANGAN MODEL TREE

Model *tree* yang dikembangkan adalah model *tree* untuk prediksi status resiko kredit. Set data terdiri dari 11 variabel prediktor yaitu jenis kelamin, umur, jumlah kredit, lama pinjaman, angsuran per bulan, total angsuran per bulan, jenis pengajuan, sektor ekonomi, omzet per bulan, keuntungan bersih per bulan, jaminan, dan nilai jaminan; dan 1 variabel target yaitu status resiko kredit. Pengembangan model *tree* ini dibantu dengan Clementine. Gambar 1 menunjukan tahapan pengembangan model dalam Clementine.



Gambar 1. Tahapan Pengembangan Model *Decision Tree*



Gambar 2. Model *Tree* Status Resiko Kredit

Model *tree* yang dihasilkan sangat tergantung pada komposisi record pada tiap set data training, test, dan validation. Untuk menghindari bias, pengembangan model dilakukan sebanyak 20 kali dengan komposisi record pada tiap set data berbeda dari satu pengembangan ke pengembangan lainnya. Tiap pengembangan model menghasilkan model yang berbeda. Gambar 2 menunjukan contoh model *decision tree* prediksi status resiko kredit dan tabel 1 menunjukan performansi (tingkat keakuratan prediksi dalam % tepat) untuk tiap model *tree* yang dihasilkan dari algoritma *decision tree*.

Tabel 1. Performansi Algoritma Decision Tree

Model	*C5.0*	*CART*	*CHAID*
1	85,26	86,86	84,62
2	88,04	88,04	89,7
3	84,64	86,27	83,33
4	89,51	88,52	88,85
5	87,96	90,15	88,32
6	86,69	87,66	86,69
7	86,96	87,29	86,29
8	87,89	83,85	84,16
9	88,27	86,97	87,3
10	85,71	87,66	86,69
11	88	87,14	88
12	86,27	87,58	87,58
13	88,12	84,82	86,47
14	90,64	86,14	92,13
15	88,85	90,24	90,24
16	88,7	87,38	85,05
17	88,82	86,26	87,54
18	89,9	88,6	88,93
19	87,94	85,46	86,17
20	86,29	87,63	84,95

8. ANALISIS PERBANDINGAN PERFORMANSI

Algoritma C5.0 menghasilkan rata-rata tingkat keakuratan sebesar 87,72% dengan standar deviasi 1,56; CART rata-rata sebesar 87,28% dan standar deviasi 1,51; CHAID rata-rata sebesar 87,15% dan standar deviasi 2,19.

Analisis variansi dilakukan untuk menguji apakah ada perbedaan variansi secara signifikan diantara ketiga algoritma tersebut. Hasil analisis menunjukan bahwa tidak ada perbedaan variansi secara signifikan. Tabel 2 menunjukan ringkasan hasil analisis variansi.

Tabel 2. Analisis Variansi

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Antar Algoritma	3.622	2	1.811	.571	.568
Dalam Algoritma	180.710	57	3.170		
Total	184.332	59			

Selain uji variansi, dilakukan juga perbandingan rata-rata yang memberikan hasil bahwa tidak ada perbedaan rata-rata secara signifikan di antara ketiga algoritma tersebut. Tabel 3 menunjukan ringkasan perbandingan rata-rata.

Tabel 3. Perbandingan Rata-rata Performansi

Algoritma (I)	Algoritma (J)	Mean Difference (I-J)	Sig.
C5.0	CART	.44700	.431
	CHAID	.57250	.314
CART	C5.0	-.44700	.431
	CHAID	.12550	.824
CHAID	C5.0	-.57250	.314
	CART	-.12550	.824

9. PENUTUP

Algoritma C5.0, CART, dan CHAID mempunyai performansi, yaitu tingkat keakuratan dalam memprediksi status resiko kredit, yang tidak berbeda secara signifikan.

Model *tree* yang dihasilkan sangat tergantung pada komposisi record dalam *training set* dan *testing set*, untuk itu diperlukan pengujian perbandingan performansi diantara model-model *tree* yang dihasilkan.

PUSTAKA

Berry, M.J.A., dan Linoff, G., 1997. *Data Mining Techniques: For Marketing, Sales, and Customer Support*. New York: John Wiley & Sons.

Cyhe, K.H., Chin, T.W., dan Peng, G.C., 2004. Credit Scoring Using Data Mining Techniques. *Singapore Management Review* 26 (2): 25-47.

Firdaus, Rachmat H., dan Maya Arianti, 2004. *Manajemen Perkreditan Bank Umum*. Bandung : Alfabeta.

Glassman, C.A., dan Wilkins, H.M., 1997. Credit Scoring: Probabilities and Pitfalls. *Journal of Retail Banking Services* 19 (2): 53-56.

Han, J., dan Kamber, M., 2001. *Data Mining: Concepts and Techniques.* San Diego: Academic Press.

Isaac, F., 2006. Small Business Credit Scoring. *Business Credit* 108 (3): 20-210

Jusuf, Jopie, 1992. *Panduan Dasar Untuk Account Officer*. Jakarta : Intermedia.

Larose, D.T., 2005. *Discovering Knowledge in Data: An Introduction to Data Mining*. New Jersey: Wiley-Interscience.

Noe, J., 1997. Credit Scoring. America's Community Banker 6 (8): 29-33.

Thomas, L.C., Oliver, R.W., dan Hand D.J., 2005. A Survey of the Issue in Consumer Credit Modelling Research. *Journal of the Operational Research Society* 56: 1006-1015.