

Credit Scoring Menggunakan Algoritma Classification And Regression Tree (CART)

Hermawan¹, Santun Irawan²

Magister Informatika, Universitas Sriwijaya
Indonesia, Palembang

hermawanmdp@gmail.com, santunirawan03@gmail.com

Reza Firsandaya Malik

Fakultas Ilmu Komputer, Universitas Sriwijaya
Jl. Raya Palembang – Prabumulih Km 32 Indralaya, OI
Sumatera Selatan 30662
rezafm@unsri.ac.id

Abstrak— *Credit Scoring*, adalah proses penilaian permohonan kredit yang dilakukan oleh lembaga kreditur/ pihak yang memberikan kredit kepada debitur selaku penerima kredit tersebut. Keluaran dari *credit scoring* adalah layak atau tidaknya calon debitur tersebut untuk menerima kredit. Tahapan ini adalah tahapan yang paling penting didalam proses kredit. Kesalahan di dalam tahapan ini akan berdampak besar pada keseluruhan tahapan pemberian kredit, dan secara global berpengaruh terhadap lembaga itu sendiri. Bidang ilmu dari teknologi informasi, yang bisa membantu *credit scoring* adalah *data mining*. Salah satu algoritma yang bisa digunakan di dalam *data mining* adalah Classification And Regresion Tree (CART). Penggunaan algoritma ini untuk *credit scoring* akan menghemat waktu, usaha dan biaya serta dengan cepat, tepat dan efektif menganalisis kelayakan calon debitur. Model yang dibentuk dari algoritma CART di *domain credit scoring* memberikan rata-rata tingkat akurasi sebesar 75,20 % dan dikategorikan sebagai *fair classification*.

Keywords—*data mining, credit scoring, CART.*

I. PENDAHULUAN

Credit scoring, sebuah tahapan penting dalam proses pemberian kredit kepada calon debitur. Penentuan kelayakan pemberian kredit kepada calon debitur yang dilakukan oleh lembaga keuangan. Kesalahan sering terjadi pada tahapan penting ini. Akan tetapi, seiring waktu akan terlihat apakah calon debitur bisa dikategorikan sebagai calon debitur yang layak atau tidak layak untuk diberikan kredit. Data historis tersebut biasanya akan tercatat di dalam *database* lembaga keuangan yang bersangkutan. Data historis tersebut kemudian akan menjadi sumber pengetahuan yang berharga. Data historis dengan ukuran skala besar tersebut, tentunya tidak mungkin dianalisa secara manual [1]. Penggunaan data mining akan sangat bermanfaat jika digunakan pada data historis tersebut. Data mining akan memprediksi calon debitur baru,

mengklasifikasikan apakah calon debitur layak atau tidak untuk diberikan kredit. Prediksi tersebut dibuat berdasarkan karakteristik data calon debitur baru, dengan dasar pertimbangan data historis yang sudah ada di lembaga keuangan. Pengetahuan dan hasil yang didapatkan dari *data mining* tersebut yang akan menjadi bahan pertimbangan lembaga keuangan untuk membuat keputusan.

Berbagai model *credit scoring* dan algoritma klasifikasi di dalam bidang ilmu data mining bisa digunakan untuk *credit scoring*. Algoritma-algoritma tersebut dikelompokkan ke dalam kategori model statistik, kategori model *artificial intelligent*, kategori model hybrid dan kategori model metode ensemble [2]. Beberapa algoritma klasifikasi yang sudah pernah digunakan di dalam penelitian lainya adalah Support Vector Machine (SVM), Artificial Neural Network (ANN), Classification and Regression Tree (CART). Penelitian terhadap *credit scoring* ini akan menggunakan algoritma CART, dengan pertimbangan bahwa CART adalah algoritma sederhana, tetapi *powerfull*, yang menggunakan pohon keputusan untuk memprediksi layak dan tidaknya calon debitur untuk mendapatkan kredit. Algoritma ini juga diakui sebagai salah satu dari beberapa algoritma terbaik di dalam data mining. [3].

Penelitian sebelumnya telah melakukan komparasi beberapa metode seperti algoritma ID3, C4.5 dan CART pada area penelitian medical data mining. Hasilnya menunjukkan bahawa algoritma CART memberikan hasil akurasi yang lebih baik dibandingkan metode lainnya. [4]. Komparasi lainnya dilakukan dengan algoritma Chi Squared Automatic Interaction Detection (CHAID), dan hasilnya CART masih menunjukkan hasil akurasi yang lebih baik walaupun tidak signifikan.[5]. Beberapa penelitian yang terkait algoritma CART telah banyak dilakukan. Algoritma CART pernah diaplikasikan di dalam bidang medical data mining, khususnya untuk mendiagnosa penyakit hepatitis. Setelah membandingkan hasil akurasi

Prosiding
ANNUAL RESEARCH SEMINAR 2016

6 Desember 2016, Vol 2 No. 1

ISBN : 979-587-626-0 | UNSRI

<http://ars.ilkom.unsri.ac.id>

metode ini dengan beberapa metode lainnya seperti ID3 dan C4.5, algoritma CART menunjukkan hasil akurasi terbaik dengan tingkat akurasi sebesar 83,2 % [4]. Masih di bidang medical data mining, kombinasi metode CART dengan Random Forest untuk deteksi penyakit diabetes type II menunjukkan hasil akurasi sebesar 77% [6]. CART juga digunakan untuk menganalisis faktor-faktor yang mempengaruhi urbanisasi di area Jawa Timur, yang memberikan hasil berupa model pohon dan tingkat persentasi sebesar 75,92 % [7]. Penelitian lain mengenai CART juga menunjukkan hasil baik. Metode ensemble yang diterapkan pada algoritma CART juga menunjukkan bahwa algoritma ini mampu juga beradaptasi dengan metode ensemble, dan mampu memberikan peningkatan akurasi pada klasifikasi tingkat kemiskinan [8].

II. CLASSIFICATION AND REGRESSION TREE (CART)

CART adalah algoritma yang termasuk sebagai kelompok metode statistik non parametric, yang dikembangkan oleh oleh Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. CART dapat menggambarkan hubungan antara variabel respon (variabel dependen/variabel target) dengan satu atau lebih variabel prediktor (variabel independent) [7]. Algoritma ini akan menghasilkan keluaran berupa pohon klasifikasi, jika variabel targetnya adalah tipe data kategorik. Sedangkan, apabila variabel targetnya adalah tipe data numeric atau kontinu, algoritma CART akan menghasilkan pohon regresi. [9]

Metodologi CART terdiri dari 4 langkah yaitu :

1. Tree Building

Pembuatan pohon keputusan, yang dimulai dari root node. CART akan melakukan perhitungan variabel dan nilai yang terbaik untuk dibagi ke dalam tepat dua child node. Proses ini akan dilakukan secara rekursif berulang-ulang. Kriteria pemilahan terbaik umumnya diukur menggunakan impurity measure. Beberapa fungsi impurity measure yang bisa digunakan adalah gini impurity, information gain, twoing dan entropy. [10]

2. Stopping Tree Building process

Algoritma CART awalnya dibuat secara rekursif tanpa ada batasan, sampai pada tahap tidak bisa dibuat lagi simpul atau node. Perkembangan berikutnya dari algoritma ini, kemudian dibuat batasan tertentu untuk penghentian pembentukan pohon keputusan. Batasan tersebut, biasanya adalah jika tidak ada penurunan keheterogenan yang berarti. Batasan lainnya yang bisa digunakan adalah

batasan minimum, batasan level pohon, dan kedalaman pohon. [10]

3. Tree Pruning

Algoritma CART dengan pembentukan pohon keputusan secara rekursif, otomatis biasanya juga akan menghasilkan pohon keputusan yang kompleks juga. Tahapan ini akan memangkas pohon keputusan dengan kompleksitas yang tinggi, sehingga akan menghasilkan pohon keputusan yang mudah dimengerti, dan tetap dengan pertimbangan bahwa akurasi yang dihasilkan juga harus baik. Pengujian akan dilakukan dengan menggunakan pengukuran cost complexity minimum.

4. Optimal Tree Selection

Tahapan ini adalah tahapan lanjutan dari tahapan *tree pruning*. Dimana tahapan ini akan memilih pohon keputusan yang paling optimum dan terbaik. Pengukuran dilakukan dengan melakukan dengan *test sample estimate* dan *V cross validation estimate*.

III. METODE PENELITIAN

Penelitian ini akan menggunakan lingkungan sistem operasi Windows 7 64 bit, dengan spesifikasi perangkat keras menggunakan Intel Core I-5 dan memory dengan kapasitas 10 GB, serta harddisk 500 GB. Pengujian algoritma CART akan dilakukan dengan menggunakan tools data mining Rapidminer, dengan library Simple CART yang berasal dari *Weka*. Kemudian validasi akurasi model yang dibangun akan menggunakan metode *Cross Validation*, yaitu K-Fold Cross Validation. Nilai K yang digunakan adalah 10, sehingga metode validasi yang digunakan adalah *10-Fold Cross Validation*. Analisa hasil penelitian, menggunakan *Confusion Matrix* dan kurva *Receiver Operating Characteristic (ROC)*. Analisa penelitian akan menunjukkan tingkat akurasi, presisi, recall, dan hasil kurva ROC akan menunjukkan hasil penelitian dalam bentuk visual.

IV. DATASET

Dataset public yang digunakan adalah dataset penilaian kredit yang berasal dari University of California (UCI), yaitu yaitu German.data-numeric. Dataset "german.data-numeric" berisikan data nasabah yang dikategorikan sebagai yang tergolong baik atau tidak. Dataset "german.data-numeric" adalah dataset hasil modifikasi "german data" oleh universitas Strathclyde, dimana data-data polynominal diubah menjadi data-data numerik. Dataset penelitian dapat di unduh di alamat

Prosiding ANNUAL RESEARCH SEMINAR 2016

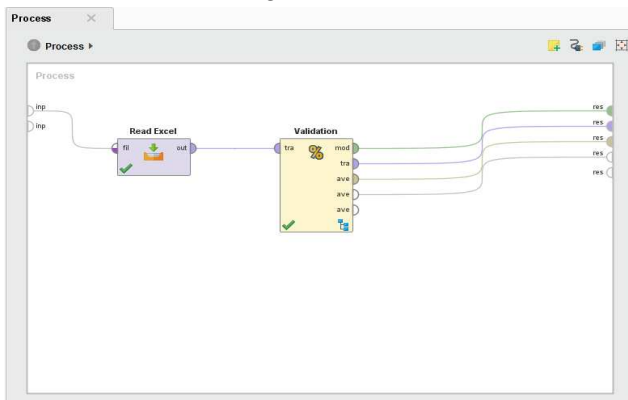
6 Desember 2016, Vol 2 No. 1

[http://archive.ics.uci.edu/ml/datasets/Statlog+\(German+Credit+Data\)](http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)).

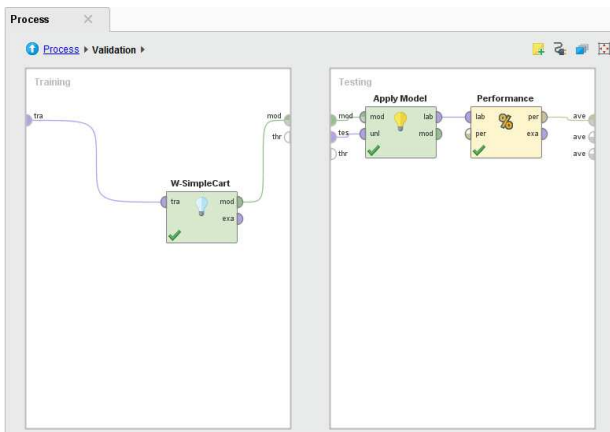
V. HASIL DAN PEMBAHASAN

Penelitian ini menghasilkan output berupa persentase tingkat akurasi dari prediksi dari model *credit scoring* yang dibentuk oleh algoritma CART. Kemudian juga mengklasifikasi hasil output model melalui kurva ROC, sesuai dengan aturan klasifikasi internasional yang telah disepakati.

5.1 Model Credit Scoring



Gambar 5.1 Model Proses utama Data Mining



Gambar 5.2 Model Algoritma CART

Gambar 5.1 dan Gambar 5.2 menunjukkan model *credit scoring* yang diusulkan. Gambar 5.1 menunjukkan bahwa data yang diambil adalah data yang sudah ditransformasikan ke dalam bentuk Microsoft Excel, dan model akan menggunakan metode validasi K-Fold Cross Validation dengan nilai K yang

digunakan adalah 10. Sedangkan gambar 5.2 menunjukkan tahapan inti dari model, yaitu pemrosesan data menggunakan algoritma CART melalui pembagian dua tahapan utama yaitu tahapan training dan tahapan testing.

5.2 Confusion Matrix

Accuracy : 75,20 % +/- 4,09 % (micro: 75,20 %)

Parameter	True 1	True 2	class precision
pred 1	616	164	78,97 %
pred 2	84	136	61,82 %
Class Recall	88,00 %	45,33 %	

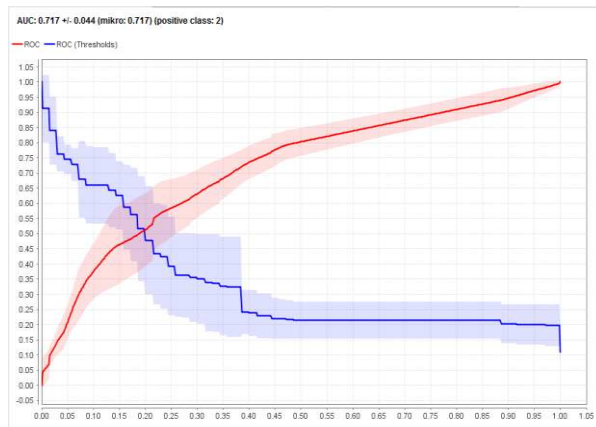
Tabel 5.1 Confusion Matrix

Gambar 5.2 menunjukkan bahwa dari 1000 data yang diuji, sejumlah 616 data hasil prediksi “1/Benar/Layak” dan hasilnya juga benar (prediksi akurat), sedangkan hasil prediksi “1/Benar/Layak” tetapi pada kenyataannya “2/Salah/Tidak Layak” sejumlah 164 data. Sedangkan kebalikkannya, sejumlah 84 data di prediksi sebagai “2/Salah/Tidak Layak” dan juga hasilnya juga “2/Salah/Tidak Layak” (prediksi akurat), dan sejumlah 136 data menunjukkan bahwa prediksi “2/Salah/Tidak Layak” akan tetapi pada kenyataannya bahwa data tersebut harusnya dikelompokkan/diklasifikasikan sebagai data yang “1/Benar/Layak”. Tingkat akurasi menunjukkan bahwa persentase tingkat akurasi yang dihasilkan adalah sebesar 75,20 %. Berikut adalah tabel hasil uji coba hasil penelitian :

Parameter	Persentase
Accuracy	75,20 %
Precision	62,29 %
Recall	45,33 %

Tabel 5.2 Tabel Hasil Uji Coba

5.3 Kurva ROC



Hasil kurva *Receiver Operating Characteristic* (ROC) menunjukkan hasil nilai *Area Under Curve* (AUC) sebesar 0,717. Sesuai aturan klasifikasi, nilai tersebut dikategorikan sebagai cukup (*fair classification*).

VI. KESIMPULAN

Data Mining adalah disiplin ilmu yang salah satu tugasnya adalah tugas prediktif, yaitu memprediksikan nilai suatu atribut berdasarkan atribut data lainnya. Oleh karena itu, bisa dimanfaatkan di dalam *credit scoring*/ penilaian kredit. Dimana hasil dari klasifikasi tersebut menentukan apakah calon debitur bisa dikategorikan sebagai debitur yang layak atau tidak untuk diberikan kredit. Hasil prediksi tersebut yang bisa dijadikan pertimbangan dan dasar keputusan bagi lembaga keuangan untuk memberikan kredit. Hasil penelitian

menunjukkan bahwa algoritma CART bisa digunakan untuk *credit scoring*, penilaian kredit dengan akurasi sebesar 75,20 %. Akan tetapi, tantangan yang umum selalu terjadi di dalam data mining adalah bagaimana meningkatkan tingkat akurasi dari model /algoritma yang diusulkan. Sehingga topik penelitian (*future research*) selanjutnya harus berfokus bagaimana meningkatkan tingkat akurasi model *credit scoring* yang diusulkan.

DAFTAR PUSTAKA

- [1] H.-P. Kriegel, K. M. Borgwardt, P. Kröger, A. Pryakhin, M. Schubert, and A. Zimek, "Future trends in data mining," *Data Min Knowl Disc*, vol. 15, pp. 87–97, 2007.
- [2] X.-L. Li, "An Overview of Personal Credit Scoring: Techniques and Future Work," *Int. J. Intell. Sci.*, vol. 02, no. 24, pp. 182–190, 2012.
- [3] X. Wu, V. Kumar, Q. J. Ross, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, *Top 10 algorithms in data mining*, vol. 14, no. 1. 2008.
- [4] G. Sathyadevi, "Application of CART algorithm in hepatitis disease diagnosis," *Int. Conf. Recent Trends Inf. Technol. ICRTIT 2011*, pp. 1283–1287, 2011.
- [5] M. A. Kadir and S. Teknik, "Perbandingan Performansi Algoritma Decision Tree CART dan CHAID."
- [6] M. K. Sabariah, "Early Detection of Type II Diabetes Mellitus with Random Forest and Classification and Regression Tree (CART)," pp. 238–242, 2014.
- [7] A. Hartati, I. Zain, and S. Suprih, "Analisis CART (Classification And Regression Trees) pada Faktor-Faktor yang Mempengaruhi Kepala Rumah Tangga di Jawa Timur Melakukan Urbanisasi," vol. 1, no. 1, 2012.
- [8] K. Jombang, M. J. Muttaqin, B. W. Otok, and P. Rahayu, "Metode Ensemble pada CART untuk Perbaikan Klasifikasi Kemiskinan," pp. 1–6.
- [9] J. Morgan, "Classification and Regression Tree Analysis," *PM931 Dir. Study Heal. Policy Manag.*, no. 1, p. 16, 2014.
- [10] F. Eka and I. Zain, "Klasifikasi Pengangguran Terbuka Menggunakan CART (Classification and Regression Tree) di Provinsi Sulawesi Utara," *J. SAINS DAN SENI POMITS*, vol. 3, no. 1, pp. 2337–3520, 2014.