PENINGKATAN AKURASI PADA ALGORITMA C4.5 MENGGUNAKAN ADABOOST UNTUK MEMINIMALKAN RESIKO KREDIT

Aldi Nurzahputra[1*], Much Aziz Muslim [2]
[1,2] Program Studi Teknik Informatika, Fakultas MIPA, Universitas Negeri Semarang
[*]Email: aldinurzah96@gmail.com

Abstrak

Tingkat akurasi dalam penilaian risiko pemohon kredit sangat penting bagi organisasi pemberi pinjaman. Data pemohon kredit yang besar dapat diolah menjadi informasi yang dapat digunakan sebagai pendukung keputusan dalam menentukan permohoanan kredit. Pengolahan data tersebut termasuk dalam bidang data mining. Salah satu metode yang dapat diterapkan dalam permohonan kredit, yaitu klasifikasi. Terdapat beberapa algoritma klasifikasi salah satunya yaitu pohon keputusan atau decision tree. Algoritma decision tree yang terkenal ialah C4.5. Algoritma C4.5 dapat diterapkan dalam mengklasifikasi permohonan kredit. Penelitian ini menggunakan German Credit Card dataset. Adapun tujuan penelitian ini yaitu meningkatkan akurasi dari algoritma C4.5 dengan menerapkan adaboost dalam mengklasifikasi permohonan kredit dengan membandingkan hasil sebelum dan sesudah diterapkan adaboost. Validasi dalam penelitian ini menggunakan 10 fold cross validation. Sedangkan pengukuran akurasi diukur dengan confussion matrix. Hasil percobaan menunjukan terdapat peningkatan akurasi 3.7%. Akurasi penerapan algoritma C4.5 saja mencapai 70.5%. Sedangkan akurasi pnerapan algoritma C4.5 dengan adaboot mencapai 74.2%.

Kata Kunci: C4.5, *Adaboost, Data Mining*, *German Credit Card.*

1. PENDAHULUAN

Tingkat akurasi dalam penilaian risiko pemohon kredit sangat penting bagi organisasi pemberi pinjaman. Penilaian yang tepat atas kelayakan pemohon kredit memungkinkan lembaga keuangan meningkatkan volume kredit namun juga meminimalkan kemungkinan kerugian. Data pemohon kredit yang besar dapat diolah menjadi informasi yang dapat digunakan sebagai pendukung keputusan dalam menentukan permohoanan kredit. Pengolahan data tersebut termasuk dalam bidang *data mining. Data mining* menurut Han dkk. (2012) adalah proses yang mempekerjakan satu atau lebih teknik pembelajaran komputer *(machine learning)* untuk menganalisis dan mengekstraksi pengetahuan *(knowledge)* secara otomatis. Salah satu metode yang dapat diterapkan dalam permohona kredit, yaitu klasifikasi.

Menurut Han J, et al. (2012) Klasifikasi adalah suatu proses yang digunakan untuk menemukan model (atau fungsi) dengan menggambarkan dan membedakan kelas data atau konsep. 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 beberapa kelebihan. Adapun kelebihannya 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 (Quinlan, 2014). Pada penerapan beberapa kasus teknik klasifikasi, algoritma ini mampu menghasilkan akurasi dan performansi yang baik.

Algoritma C4.5 dapat diterapkan dalam bidang perbankan dalam mengklasifikasi permohonan kredit. Proses tersebut ada pada tahap analisis kredit. Analisis kredit merupakan suatu proses analisis dengan menggunakan pendekatan-pendekatan dan rasio-rasio keuangan untuk menentukan kebutuhan kredit yang wajar. Tujuan analisis kredit ialah untuk melihat/menilai suatu usaha atas dasar kelayakan usaha, menilai risiko usaha dan bagaimana mengelolanya, dan memberikan kredit. (Wahyono & Cahyono, 2015). Adapun tahapan yang dilakukan oleh bank dalam analisis kredit antara lain: (1) Permohonan Kredit, (2) Pengumpulan data dan pengamatan jaminan, serta (3) Analisis kredit. Permohonan kredit adalah Tahap pertama dalam pemberian kredit dengan pengajuan permohonan kredit oleh calon debitur. Permohonan ini bisa diajukan

secara tertulis tetapi dalam prakteknya lebih banyak dilakukan secara lisan. Tahap kedua ialah pengumpulan data dan pengamatan jaminan yang merupakan tahap tindakan. Apabila permohonan kredit dinilai layak, maka pihak bank akan melakukan pengumpulan data lapangan baik menyangkut data pribadi maupun reputasi dan hal-hal lain yang berkaitan dengan bisnis calon debitur. Tahap ketiga adalah analisis kredit dimana tahapan yang paling menentukan dalam analisis dan pengambilan keputusan pemberian kredit terhadapa layak atau tidak permohonan kredit calon debitur. Pihak bank dituntut obyektif dan konsisten atas hasil analisis dengan berpegang pada prinsip-prinsip kelayakan kredit.

Penelitian ini menggunakan *German Credit Card dataset*. *Dataset* yang digunakan dalam penelitian ini diperoleh dari UCI *Repository of Machine Learning Datasets*. *Dataset* merupakan kumpulan dari objek dan sifat atau karakteristik dari suatu objek itu sendiri (atribut). Penelitian ini menerapkan *ensemble learning adaboost* pada proses *meta-learning* untuk klasifikasi permohonan kredit.

Tujuan dari penelitian ini yaitu meningkatkan akurasi dari algoritma C4.5 dengan menerapkan *adaboost* dalam mengklasifikasi permohonan kredit dengan membandingkan hasil sebelum dan sesudah diterapkan *adaboost*. Validasi dalam penelitian ini menggunakan 10 *fold cross validation*. Sedangkan pengukuran akurasi diukur dengan *confussion matrix.*

2. METODOLOGI

Decision tree merupakan metode yang ada pada teknik klasifikasi dalam data mining. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang mempresentasikan aturan. Pohon keputusan juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara jumlah calon variable input dengan sebuah variabel target (Ahmad, 2012).

Metodologi yang digunakan dalam penelitian ini terdiri dari tiga tahap. Tahap pertama ialah studi literatur dengan mengumpulkan teori dan fakta terkait algoritma C4.5 dengan *meta-learning.* Tahap kedua ialah pengumpulan data. Data yang digunakan dalam penelitian ini adalah German Credit Card dataset. Tahap ketiga adalah pengolahan data dan uji coba dengan menerapkan algoritma C4.5 dengan penambahan *adaboost.* Kemudian melakukan perbandingan tingkat akurasi algoritma C4.5 dengan algoritma C4.5 dan *adaboost*.

Algoritma C4.5

Banyak algoritma yang dapat dipakai dalam pembentukan pohon keputusan, salah satunya adalah algoritma C4.5. Algoritma C4.5 membuat pohon keputusan dari atas ke bawah, di mana atribut paling atas merupakan akar, dan yang paling bawah dinamakan daun. Keuntungan dalam metode ini adalah efektif dalam menganalisis sejumlah besar atribut dari data yang ada dan mudah dipahami oleh pengguna akhir (Da & Ji, 2014).

Secara umum Algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut:

1. Pilih atribut sebagai akar
2. Buat cabang untuk masing-masing nilai
3. Bagi kasus dalam cabang
4. Ulangi proses untuk masing-masing cabang sampai semua kasus pada cabang memiliki kelas yang sama. (Kamagi & Hansun, 2014)

Algoritma C4.5 merupakan salah satu algoritma yang telah secara luas digunakan, khususnya di area machine learning yang memiliki beberapa perbaikan dari algoritma sebelumnya, ID3, yaitu dalam hal metode pemangkasannya (*prunning*). Adapun perbaikannya adalah sebagai berikut:

1. Algoritma C4.5 menghitung gain ratio untuk masing-masing atribut, dan atribut yang memiliki nilai yang tertinggi akan dipilih sebagai simpul. Penggunaan gain ratio ini memperbaiki kelemahan dari ID3 yang menggunakan information gain.
2. Pemangkasan dapat dilakukan pada saat pembangunan pohon (*tree*) ataupun pada saat proses pembangunan pohon selesai.
3. Mampu menangani *continues attribute*.
4. Mampu menangani *missing* data.

5. Mampu membangkitkan rule dari sebuah pohon. (Muzakir & Wulandari, 2016)

Pemilihan atribut sebagai akar didasarkan pada nilai gain tertinggi dari atribut-atribut yang ada. Rumus yang digunakan untuk menghitung gain ditunjukan pada persamaan 1.

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

Dimana S merupakan himpunan kasus dan A merupakan atribut data. Nilai n merupakan jumlah partisi atribut A dan |Si| merupkan jumlah kasus pada partisi ke-i. Jumlah kasus ditunjukan dengan |S|.

Sebelum mendapatkan nilai Gain adalah dengan mencari nilai Entropi. Entropi digunakan untuk menentukan seberapa informatif sebuah masukan atribut untuk menghasilkan sebuah atribut. Rumus dasar dari Entropi ditunjukan pada persamaan 2.

$$Entropy(S) = \sum_{i=1}^{n} - pi * \log_2 pi \quad (2)$$

Dimana n merupakan jumlah partisi S dan pi merupakan proporsi dari Si terhadap S.

***Adaptive Boosting* (*Adaboost*)**

Boosting adalah pendekatan pada *machine learning* untuk meningkatkan peraturan prediksi yang akurat dengan menggabungkan banyak peraturan yang relatif lemah dan tidak akurat. *Adaptive* boosting (*adaboost*) merupakan salah satu dari beberapa varian pada algoritma *boosting* (Liu, 2015). *Adaboost* merupakan *ensemble learning* yang sering digunakan pada algoritma *boosting*.

Algoritma *AdaBoost* dari Freund dan Schapire (1995) merupakan algoritma penguat praktis pertama, dan tetap menjadi salah satu yang paling banyak digunakan dan dipelajari, dengan aplikasi di berbagai bidang. *Boosting* bisa dikombinasikan dengan *classifier* algoritma yang lain untuk meningkatkan performa klasifikasi. Tentunya secara intuitif, penggabungan beberapa model akan membantu jika model tersebut berbeda satu sama lain.

Adaboost dan variannya telah sukses diterapkan pada beberapa bidang (*domain*) karena dasar teorinya yang kuat, prediksi yang akurat, dan kesederhanaan yang besar. Langkah-langkah pada algoritma *adaboost* adalah sebagai berikut.

a. *Input*: Suatu kumpulan *sample* penelitian dengan label $\{(x_i, y_i), \dots, (x_N, y_N)\}$, suatu *component learn* algoritma, jumlah perputaran T.
b. *Initialize*: Bobot suatu sampel pelatihan $w_i^1 = 1/N$, untuk semua i=1,...,N
c. Do for t= 1, ...,T
   1. Gunakan c*omponent learn* algoritma untuk melatih suatu komponen klasifikasi, h_t, pada *sample* bobot pelatihan.
   2. Hitung kesalahan pelatihannya pada h_t: $\varepsilon_t = \sum_{i=1}^{N} w_i^t , y_i \neq h_t(x_i)$
   3. Tetapkan bobot untuk *component classifier* $h_t = \alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$
   4. *Update* bobot *sample* pelatihan $w_i^{t+1} = \frac{w_i^t \exp\{-\alpha_t y_i h_t(x_i)\}}{C_t}, i = 1, \dots, N$ C_t adalah suatu konstanta normalisasi.
d. *Output*: $f(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$.

3. HASIL DAN PEMBAHASAN

Penelitian ini meggunakan tool data mining, Weka 3.6. Weka merupakan toolyang dilengkapi algoritma machine learning untuk keperluan data mining (Thornton, 2013). Pengujian model dilakukan menggunakan *German Credit Card dataset* yang diambil dari UCI *repository of machine learning*. *German Credit Card dataset* terdiri dari 20 artibrut untuk klasifikasi permohona kredit dianggap berisiko buruk atau baik dengan jumlah 1000 pemohon kredit. Adapun kelas yang digunakan dalam data tersebut ialah good dan bad. Berikut atribut yang digunakan pada dataset ditunjukan pada Tabel 1.

Tabel 1. Atribut German Credti Card Dataset

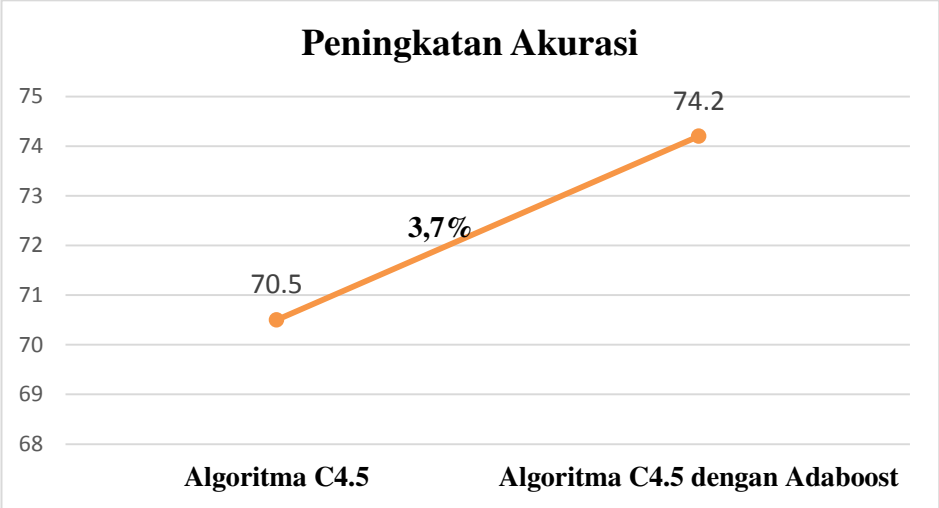
Atribut	Nama	Tipe
1.	*Status of existing checking account*	*kualitatif*
2.	*Duration in month*	*numerik*
3.	*Credit history*	*kualitatif*
4.	*Purpose*	*kualitatif*
5.	*Credit amount*	*numerik*
6.	*Savings account/bonds*	*kualitatif*
7.	*Present employment since*	*kualitatif*
8.	*Installment rate in percentage of disposable income*	*numerik*
9.	*Personal status and sex*	*kualitatif*
10.	*Other debtors / guarantors*	*kualitatif*
11.	*Present residence since*	*numerik*
12.	*Property*	*kualitatif*
13.	*Age in years*	*numerik*
14.	*Other installment plans*	*kualitatif*
15.	*Housing*	*kualitatif*
16.	*Number of existing credits at this bank*	*numerik*
17.	*Job*	*kualitatif*
18.	*Number of people being liable to provide maintenance for*	*numerik*
19.	*Telephone*	*kualitatif*
20.	*foreign worker*	*kualitatif*

Penerapan adaboost dengan algoritma C4.5 pada penelitian ini menggunakan *Number of performed Iterations* 100. Adapun hasil perbandingan penerapan algoritma C4.5 pada *German Credit Card* ditunjukan pada Tabel 2.

Tabel 2. Hasil Pebandingan Akurasi Model

Perbandingan Model	**Algoritma C4.5**	**Algoritma C4.5 dengan *Adaboost***
Akurasi (%)	70.5	74.2

Hasil perbandingan tersebut menunjukan bahwa penerapan *adaboost* pada *German Credit Card* dapat meningkatkan akurasi algoritma C4.5 sebesar 3.7%. Peningkatkannya ditunjukan pada Gambar 1 berikut.



Gambar 1. **Peningkatan Akurasi Algoritma C4.5**

Dari pengolahan data yang sudah dilakukan dengan metode *boosting* yaitu *adaboost*, terbukti dapat meningkatkan akurasi algoritma C4.5 pada *German Credit Card*. Data yang digunakan dapat diklasifikasikan dengan baik ke dalam klasifikasi *good* dan *bad*.

4. KESIMPULAN

Pengujian model dilakukan menggunakan *German Credit Card dataset* yang diambil dari *UCI repository of machine learning* untuk mengklasifikasi permohonan kredit. Pengolahan data dilakukan dengan penerapan algoritma C4.5 saja dengan hasil akurasi 70.5 %. Sedangkan penerapan algoritma C4.5 dengan penambahan *adaboost* diperoleh hasil akurasi mencapai 74.2%. Hasil penelitian tersebut menunjukan bahwa algoritma *boosting*, *adaboost* dapat meningkatkan algoritma C4.5 pada *German Credit Card* mencapai peningkatan akurasi 3.7%.

DAFTAR PUSTAKA

Dai, W. and Ji, W., 2014. A mapreduce implementation of C4. 5 decision tree algorithm. *International Journal of Database Theory and Application*, *7*(1), pp.49-60.

Freund, Y. and Schapire, R.E., 1995, March. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory* (pp. 23-37). Springer Berlin Heidelberg.

Han, J., Micheline, K. & Jian, P. 2012. *Data mining: Concepts and Techniques* (3th ed.). Waltham, MA: Elsevier/Morgan Kaufmann.

Ian H Witten, Eibe Frank, and Mark A Hall, *Data Mining Practical Machine Learning Tools and Techniques*, 3rd ed. USA: Morgan Kaufmann Publishers, 2011.

Kamagi, D. H., & Hansun, S., 2014. Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa. ULTIMATICS. 6(1), 15-18.

Liu, H., Tian, H.Q., Li, Y.F. and Zhang, L., 2015. Comparison of four Adaboost algorithm based artificial neural networks in wind speed predictions. Energy Conversion and Management, 92, pp.67-81.

Muzakir, A. and Wulandari, R.A., 2016. Model Data Mining sebagai Prediksi Penyakit Hipertensi Kehamilan dengan Teknik Decision Tree. Scientific Journal of Informatics, 3(1), pp.19-26.

Quinlan, J.R., 2014. C4. 5: programs for machine learning. Elsevier.

Thornton, C., Hutter, F., Hoos, H.H. and Leyton-Brown, K., 2013, August. Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 847-855). ACM.

Wahyono, T. and Cahyono, A.D., 2015. Mitigasi Risiko Kredit: Studi Model-Model Sistem Pendukung Keputusan Permohonan Kredit Pada Koperasi Simpan Pinjam.