

Algoritma *Decision Tree* Dengan Menggunakan *Pruning* dan *Missing Value* Untuk Prediksi Kredit Macet

Nur Hatima Ina Arifin¹, Wali Ja'far Shudiq²

^{1,2} Universitas Nurul Jadid, Probolinggo

Info Artikel

Riwayat Artikel

Diterima: 27-05-2022

Disetujui: 30-06-2022

Kata Kunci

Decision Tree ;

Pruning;

Missing Value;

Kredit Macet;

ABSTRAK

Sebagai industri yang mempunyai perkembangan sangat pesat yang di ikuti dengan resiko kredit pada setiap kegiatan usaha perbankan. Kredit merupakan penghasilan utama dari sebuah perbankan dan juga merupakan sumber resiko terbesar karena sebagian besar dana operasional bank berputar dalam bentuk kredit. Proses data nasabah, data jaminan hingga pada proses pengambilan keputusan menjadi peran penting untuk menekan dan mengurangi resiko bagi pihak pengajuan kredit. Syarat perbankan menjadi indikator pemberian pinjaman kepada nasabah agar layanan yang diberikan tepat dan tidak memberikan efek buruk pada manajemen bank , misalnya kredit macet [2]. Tujuan dari penelitian ini adalah untuk memprediksi tingkat resiko kredit yang terjadi pada sebuah perbankan. Berdasarkan beberapa metode yang telah diusulkan dan telah dilakukan prediksi kredit macet, menghasilkan nilai akurasi terendah. Dalam penelitian ini menggunakan metode *pruning* pada *missing value* dengan menerapkan Algoritma *decision tree* untuk prediksi kredit macet dengan jumlah dataset sebanyak 732 record dengan 11 variabel dan terdapat data yang tidak lengkap (*missing value*) pada salah satu variable dataset kredit macet. Pada pengujian algoritma *decision tree* dengan metode *pruning* berhasil meningkatkan nilai akurasi menjadi 75.80%, sedangkan algoritma *decision tree* tanpa metode *pruning* hasil akurasi adalah 72.61%.

*chatiem.indah@gmail.com

1. PENDAHULUAN

Situasi pada lingkungan internal dan eksternal sebuah perbankan yang mempunyai perkembangan sangat pesat yang di ikuti dengan resiko pada setiap kegiatan usaha perbankan sehingga meningkatkan kebutuhan praktek tata kelola bank yang sehat (*good corporate governance*) dan penerapan manajemen resiko sesuai dengan ketentuan yang berlaku [1]. Kredit adalah sumber utama dari penghasilan bagi sebuah perbankan dan merupakan sumber resiko terbesar, karena sebagian besar dari dana operasional bank berputar dalam bentuk kredit. Proses data nasabah, data jaminan hingga pada proses pengambilan keputusan menjadi peran penting untuk menekan dan mengurangi resiko bagi pihak pengajuan kredit. Syarat perbankan menjadi indikator pemberian pinjaman kepada nasabah agar layanan yang diberikan tepat dan tidak memberikan efek buruk pada manajemen bank , misalnya kredit macet [2]. Karena pada Nurja Muamalah Pondok Pesantren Nurul jadid Paiton Probolinggo Jawa Timur mengalami banyak kemacetan sehingga perlu dilakukan prediksi kredit macet

untuk menghasilkan pengambilan keputusan yang tepat. Minimnya manfaat yang diperoleh dari database kredit macet seolah – olah membuat pekerjaan mengambil data ini menjadi *non added value* atau kurang memberikan nilai tambah, salah satu cara memanfaatkan data tersebut adalah dengan mengolahnya sehingga pola atau kecenderungan pada data tersebut dapat ditemukan. Dengan mengetahui pola yang terdapat pada data kredit macet, maka hal tersebut akan dapat digunakan dalam membantu pengambilan keputusan bagi pihak bank dalam memberikan sebuah keputusan pengambilan kredit.

Klasifikasi adalah metode analisis data yang bisa memprediksi sebuah label kelas dari sampel untuk diklasifikasi. Berbagai teknik klasifikasi diusulkan diberbagai bidang seperti *machine learning*, sistem pakar dan statistik. Umumnya, model klasifikasi dilatih terlebih dahulu pada kumpulan data (yaitu, rangkaian pelatihan) dengan label kelas yang telah diketahui, pengklasifikasi terlatih ini dilakukan prediksi label kelas pada sampel baru [5].

Decision Tree Adalah pohon keputusan mengklasifikasikan sampel dengan cara top-down, dimulai dari simpul akar dan terus bergerak sesuai dengan hasil tes pada simpul internal, sampai simpul daun tercapai dan label kelasnya ditugaskan [5].

Algoritma C4.5 adalah algoritma yang dikembangkan dari algoritma ID3. Oleh sebab itu algoritma C4.5 memiliki prinsip dasar kerja yang hampir sama dengan algoritma ID3 [9]. Dalam memilih atribut algoritma C4.5 menggunakan Gain Ratio pada persamaan dibawah ini:

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

Di mana :

S : Himpunan Kasus

A : Atribut

N : Jumlah partisi atribut A

$|S_i|$: jumlah kasus pada partisi ke i

$|S|$: jumlah kasus dalam S

Sedangkan nilai perhitungan entropy dapat dilihat dari persamaan berikut:

$$entropy(A) = - \sum_{i=1}^N \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

Keterangan:

S : Himpunan Kasus

A : Atribut

N : jumlah partisi atribut A

$|S_i|$: jumlah kasus pada partisi ke i

$|S|$: jumlah kasus dalam S

Information Gain merupakan *attribute selection measure* yang diterapkan untuk pemilihan test atribut pada setiap *node* yang terdapat pada *tree*. Atribut yang memiliki information gain tertinggi yang dipilih sebagai *test attribute* dari suatu *node*. Dalam proses perhitungannya, *gain* dapat terjadi atau tidak pada sebuah missing value [10].

Metode Pruning adalah Penyederhanaan rule dalam C4.5 dengan cara menghitung angka kesalahan untuk setiap cabang (*leaf node*) dan menerima bahwa setiap angka kesalahan akan menjadi kesalahan yang buruk, penyederhanaan rule dalam C4.5 merupakan sebuah metode pemangkasan pohon keputusan yang membandingkan entropy yang ada pada pohon keputusan, mengubah pohon keputusan yang dihasilkan dalam beberapa *rule* [9].

Pada penelitian ini metode *post pruning* yang digunakan adalah *error based pruning* (EBP). Algoritma EBP biasa digunakan pada algoritma *decision tree* C4.5, dalam algoritma ini mengizinkan pergantian *subtree* dengan salah satu *leaf node* nya untuk membuat *decision tree* yang lebih simple.

EBP mulai melakukan *pruning* pada *internal node* dari bagian bawah *decision tree*. Pemeriksaan setiap *internal node* dilakukan dengan menggunakan persamaan berikut :

$$e'(t) < e'(T_t) + \text{std}(e'(T_t))$$

dengan :

$e'(t)$: *error rate internal node*

$e'(T_t)$: *error rate subtree* dengan *internal node* t^1 .

dengan :

$n(t)$: jumlah *node* yang diperiksa *error rate* nya.

Apabila *error rate* nya menjadi lebih kecil maka *pruning subtree* dilakukan [13].

Missing value adalah sebuah data atau informasi yang tidak tersedia pada atribut tertentu. Dataset *missing value* bisa terjadi karena nilainya tidak relevan atau tidak bisa tercatat pada saat pengumpulan data karena adanya privasi. *Missing value* dapat diatasi dengan melakukan beberapa hal seperti mengurangi objek data, memperkirakan nilai *missing value*, tidak menggunakan *missing value* untuk menganalisis data, dan mencari nilai rata-rata pada sebuah atribut yang mempunyai *missing value* [9].

Evaluasi diterapkan untuk menguji kapasitas metode klasifikasi, dalam penelitian ini untuk menguji keakuratan metode klasifikasi yang di ukur dengan akurasi adalah menggunakan *precision* dan *recall* [11].

Precision dipresentasikan dari kelas data buruk yang dikelaskan dengan benar dan kelas yang seharusnya baik tetapi dikelaskan sebagai kelas data buruk. Berikut persamaan untuk menghitung *Precision* :

$$\text{Precision} = \frac{TP}{TP+FP}$$

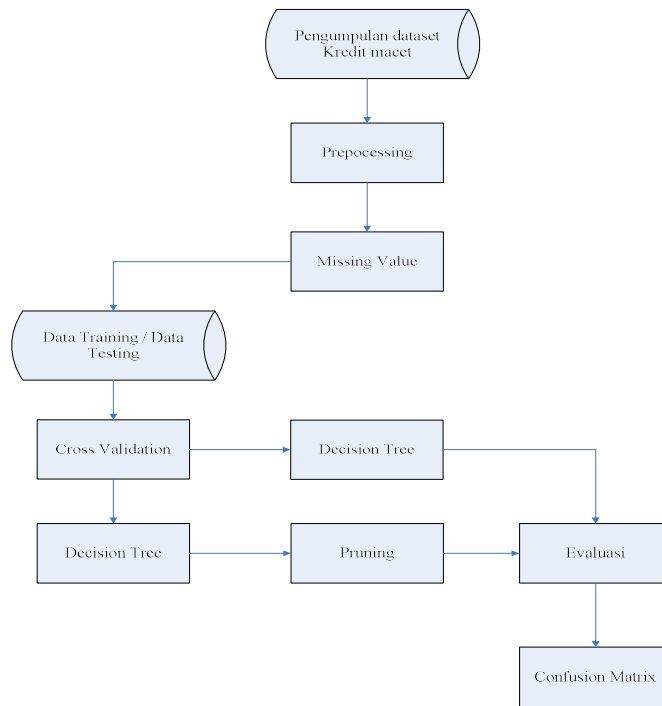
Recall mempunyai definisi sebagai presentase antara kelas data buruk yg dikelaskn dengan benar dan kelas data buruk yang salah diprediksikan ke kelas yang baik. Adapun persamaannya adalah sebagai berikut:

$$\text{Recall} = \frac{TP}{TP + FP}$$

Validasi adalah suatu tindakan yang membuktikan bahwa suatu proses/ metode dapat memberikan hasil yang konsisten sesuai dengan spesifikasi yang telah ditetapkan dan terdokumentasi dengan baik [6].

2. METODE

Data yang digunakan dalam penelitian ini adalah data kredit macet di Nurja Muamalah Pondok Pesantren Nurul jadid Paiton Probolinggo Jawa Timur. Pengumpulan data dilakukan dengan menggunakan metode wawancara, setelah wawancara selesai kemudian pengumpulan data dilakukan oleh Bpk. Fuadi, S.Pd.i selaku manager di Nurja Muamalah Pondok Pesantren Nurul jadid Paiton Probolinggo Jawa Timur. Pada dataset kredit macet terdapat beberapa data yang atributnya tidak memiliki nilai / *Missing Value* dikarenakan adanya data nasabah yang tidak lengkap. Sebelum penelitian ini dilakukan terlebih dahulu dilakukan proses pengolahan data awal (*Pre-prosessing*) agar data siap untuk diolah. (*Pre-prosessing*) dilakukan untuk memilih beberapa atribut yang memiliki nilai terdekat untuk menentukan prediksi kredit macet, pada penelitian ini pemilihan variabel dan atribut dipilih secara acak dan dilakukan secara manual.

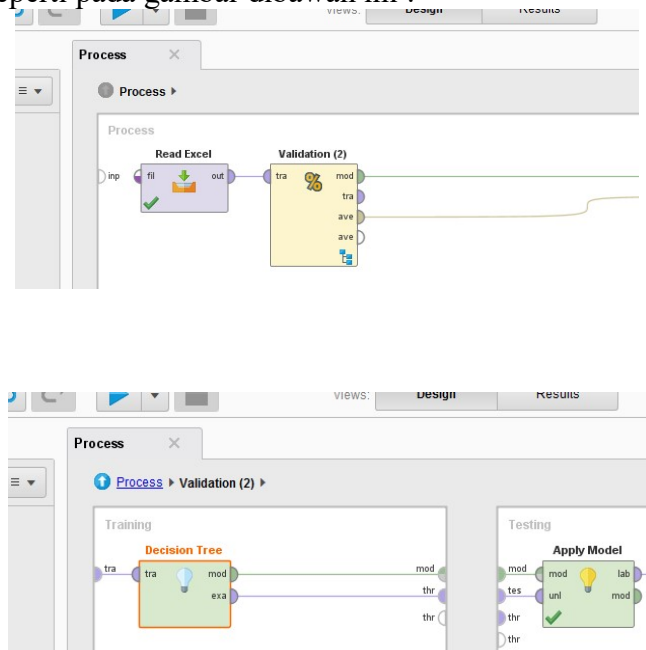


Gambar 1. Metode yang di usulkan

3. HASIL DAN PEMBAHASAN

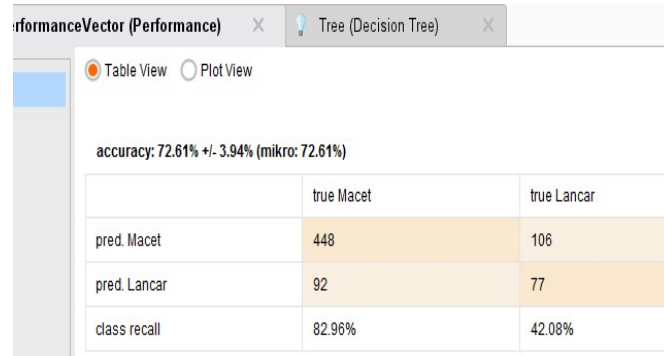
Algoritma Decision Tree C4.5

Pengujian yang dilakukan pada metode *Decision Tree* menggunakan *cross validation* dengan *information gain* sebagai *attribute selection measure* untuk sebuah *missing value*. Untuk menggunakan akurasi pada setiap algoritma, penelitian ini menggunakan *cross-validation*. Berikut hasil pengujian dataset menggunakan metode *Decision Tree* menggunakan *tools rapidminer* seperti pada gambar dibawah ini :



Gambar 2. Desain Model Validasi C4.5

Dari eksperimen yang dilakukan maka hasil akurasi yang didapat pada Algoritma *decision tree* menggunakan *information gain* ditunjukkan pada gambar berikut :



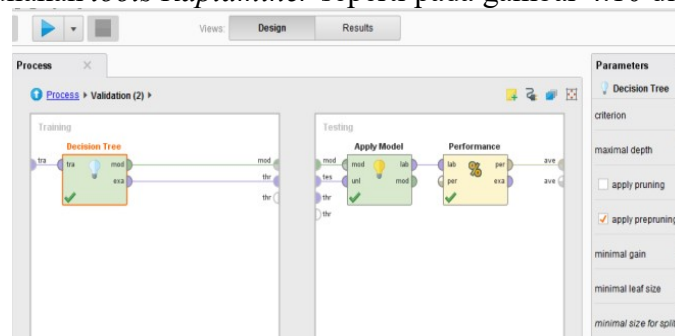
	true Macet	true Lancar
pred. Macet	448	106
pred. Lancar	92	77
class recall	82.96%	42.08%

Gambar 3. Hasil Akurasi Algoritma C4.5

Dari hasil eksperimen di atas dapat disimpulkan bahwa nilai akurasi yang didapat adalah 72.61% sehingga perlu dilakukan metode pruning untuk meningkatkan nilai akurasinya.

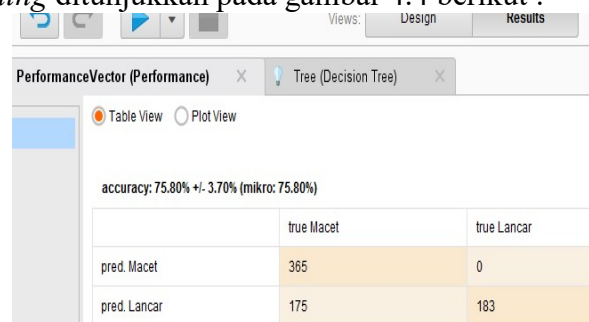
Algoritma Decision Tree C4.5 dengan Pruning

Pada tahapan ini adalah melakukan pemangkasan pada setiap cabang dengan membandingkan nilai entropi pada pohon keputusan sehingga bisa menambah nilai akurasi pada algoritma *decision tree*. Berikut hasil akurasi yang didapat menggunakan metode C4.5 dengan pruning menggunakan *tools Rapidminer* seperti pada gambar 4.10 dibawah ini :



Gambar 4. Algoritma C4.5 dengan Pruning

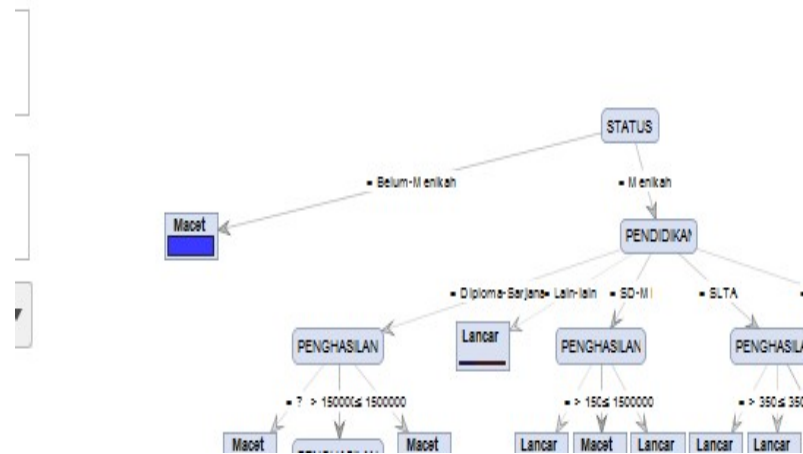
Dari eksperimen yang dilakukan maka hasil akurasi yang didapat pada Algoritma *decision tree* menggunakan *pruning* ditunjukkan pada gambar 4.4 berikut :



	true Macet	true Lancar
pred. Macet	385	0
pred. Lancar	175	183

Gambar 5. Hasil Akurasi Algoritma C4.5 dengan Pruning

Dari hasil eksperimen di atas dapat disimpulkan bahwa nilai akurasi yang dihasilkan adalah 75.80% sehingga nilai akurasi yang didapat menggunakan Algoritman C4.5 dengan metode *pruning* lebih tinggi nilai akurasinya dan menghasilkan pohon keputusan seperti pada gambar 4.6 dibawah ini :



Gambar 5. Pohon Keputusan Algoritma *Decision Tree* dengan metode *Pruning*

Evaluasi Hasil Penelitian

Pada tabel 4.1 adalah hasil dari perhitungan *Confusion matrix* algoritma C4.5 menggunakan pruning yang memiliki nilai akurasi 75.80%.

Tabel 4.1 Hasil *Confusion Matrix* Algoritma C4.5

	True Macet	Tue Lancar	Class Precision
Pred. Macet	448	106	80,87%
Pred.Lancar	92	77	45,56%
Class Recall	82,96%	42,08%	

Penjelasan dari perhitungan *confusion matrix* dia atas adalah :

- Nilai *true positive* dari label macet adalah 448 data, pada label lancar adalah 92 data.
- *Class recall* adalah kolom yang berisi hasil penelitian yang di klasifikasikan secara tepat dan benar. Berikut adalah contoh perhitungan untuk true macet :

$$TP = \frac{TP}{TP + FN} = \frac{448}{448 + 92} = \frac{448}{540} = 0,8296 = 82,96 \%$$

- *Class Precision* adalah baris yang berisi hasil penelitian yang di klasifikasikan secara tepat dan benar. Berikut ini adalah contoh perhitungan untuk true lancar:

$$TP = \frac{TP}{TP + FN} = \frac{448}{92 + 77} = \frac{448}{169} = 2,65088 = 42,08 \%$$

Perbandingan Algoritma Decision C4.5 dan Decision Tree C4.5 dengan Pruning

Hasil eksperimen Algoritma C4.5, dan C4.5 dengan pruning dapat dilihat pada tabel dibawah ini:

Tabel 4.2 Perbandingan Algoritma *Decision Tree* dan *Decision Tree* dengan Pruning

<i>Decision Tree</i>	<i>Decision Tree + Pruning</i>
72.61%	75.80%

Dari tabel 4.2 yang di hasilkan dari eksperimen menggunakan algoritma *decision tree* dan *decision tree* tanpa metode *pruning* menghasilkan akurasi 72.61% , sedangkan algoritma *decision tree* dengan *pruning* 75.80% , sehingga akurasinya bertambah 3%. Dengan demikian algoritma *decision tree* dengan *pruning* lebih menghasilkan nilai akurasi terbaik dibandingkan dengan algoritma *decision tree* tanpa *pruning*.

4. KESIMPULAN DAN SARAN

Pada penelitian ini metode yang digunakan adalah Algoritma *Decision Tree C4.5* dengan metode *pruning* untuk prediksi kredit macet dengan jumlah data 723 record dan terdapat data yang tidak lengkap (*missing value*) pada salah satu variable dataset kredit macet. Pada pengujian model dilakukan dengan menggunakan *cross validation* dengan *information gain* sebagai *attribute selection measure*, kemudian menghasilkan akurasi untuk algoritma *decision tree* tanpa metode *pruning* hasil akurasinya adalah 72.61% sedangkan *decision tree* menggunakan metode *pruning* menghasilkan akurasi 75.80%, sehingga nilai akurasi menjadi lebih tinggi dibandingkan dengan tanpa menggunakan metode *pruning*.

Untuk mengembangkan penelitian ini, berikut beberapa saran yang diusulkan:

Penelitian dapat dikembangkan lebih lanjut dengan menggunakan metode data mining lainnya serta menambahkan variable lain yang bisa meningkatkan nilai akurasinya.

Harapan pada penelitian selanjutnya dapat optimasi menggunakan metode Algoritma Genetika atau algoritma lainnya.

5. DAFTAR PUSTAKA

- [1] P. T. Bank and S. Go, "ISSN 2303-1174 E.Y.Watopa., S.Murni. I.S.Saerang. Analisis Penerapan" vol. 5, no. 2, pp. 323–333, 2017.
- [2] S. Adi and E. Winarko, "Klasifikasi Data NAP (Nota Analisis Pembiayaan) untuk Prediksi Tingkat Keamanan Pemberian Kredit (Studi Kasus : Bank Syariah Mandiri Cabang Luwuk Sulawesi Tengah)," IJCCS (Indonesian J. Comput. Cybern. Syst., vol. 9, no. 1, pp. 1–12, 2015.
- [3] B. Sudiyatno and A. Fatmawati, "Pengaruh Resiko Kredit dan Efisiensi Operasional Terhadap Kinerja Bank (Studi Empirik pada Bank yang Terdaftar di Bursa Efek Indonesia)," J. Organ. dan Manajemen, Vol. 9, Nomor 1, Maret 2013, 73-86; Fak. Ekon. Univ. Stikubank Semarang, vol. 9, no. 1, pp. 73–86, 2013.
- [4] A. F. Atiya, "Bankruptcy prediction for credit risk using neural networks: A survey and new results," IEEE Trans. Neural Networks, vol. 12, no. 4, pp. 929–935, 2001.
- [5] L. Yu, G. Chen, A. Koronios, S. Zhu, and X. Guo, "Application and Comparison of Classification Techniques in Controlling Credit Risk," Recent Adv. Data Min. Enterp. Data Algorithms Appl., pp. 2007–2007, 2007.
- [6] E. Prasetyo, mengolah data mining menjadi informasi menggunakan matlab. Yogyakarta: Andi, 2014.
- [7] J. Fürnkranz, "A Comparison of Pruning Methods for Relational Concept Learning," Knowl. Discov. Databases Pap. from 1994 AAAI Work., pp. 371–382, 1994.
- [8] Fajar astuti herawati, Data Mining. Yogyakarta, 2013.
- [9] Mochammad faid, Data mining dengan weka dan netbeans. istiqlal publishing, 2016.
- [10] S. Lorena, W. Zarman, and I. Hamidah, "Analisis Dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Memprediksi Masa Studi Mahasiswa Berdasarkan Data Nilai Akademik," Pros. Semin. Nas. Apl. Sains dan Teknol., no. November, 2014.

- [11] Sumarlin, "Implementasi Algoritma K-NNNeighbor Sebagai Pendukung Keputusan Klasifikasi Penerima Beasiswa PPA dan BBM," J. Sist. Inf. Bisnis.
- [12] Ripley BD. 1996. Pattern Recognition and Neural Networks. Cambridge : Cambridge University Press
- [13] Quinlan JR. 1992. C4.5 : Programs For Machine Learning. San mateo, CA: Morgan Kaufmann.