JTIF | Jurnal INOVATIF WIRA WACANA Vol. 2 No. 2 Agustus 2023, 153-159 p**ISSN (**2961-7316**)**
eISSN (2962-5998**)**

Penggunaan Data Mining Algoritma C4.5 dalam Menentukan Klasifikasi Nasabah Potensialdi PT. ADIRA Finance Soe

(*Use of Data Mining Algorithm C4.5 in Determining Potential Customer Classification at PT. ADIRA Finance Soe*)

Wanto I. Missa[1], Emerensye S.Y. Pandie[2], Tiwuk Widiastuti[3]

[1,2,3]Program Studi Ilmu Komputer, Universitas Nusa Cendana

E-mail :[1]missawanto88@gmail.com, [2]emerensyepandie@staf.undana.ac.id, [3]tritiwuk@gmail.com

KEYWORDS:
Customer, C4.5 Algorithm, Decision Tree, K-Fold Cross Validation

ABSTRACT

Classification of Potential Customers is an activity carried out to get customers who really have potential by paying attention to several aspects that support so that credit recipients are eligible to receive loans. PT. Adira Finance SoE is also an institution that provides loan credit services to customers and often the problem faced is the acceptance of customers who are not in accordance with procedures, causing credit payments arrears by customers. By using the C4.5 algorithm, it is hoped that it can help solve existing problems where the C4.5 algorithm is used to study characteristics based on previous customer data in order to avoid customers who are in arrears on credit. Tests carried out on 1000 data records using several attributes to support the C4.5 algorithm where the attributes used are income, number of dependents, age, marital status, occupation, place of residence. The results of this test get a decision tree that determines whether the customer is eligible or not to get a loan based on the existing attributes of the previous customer data. System testing is done by applying the K-Fold Cross Validation algorithm when it has been calculated using the C4.5 algorithm with the results of 1000 records where 956 records are in the right classification class and 44 records are in the wrong classification class.

KATA KUNCI:
Nasabah, Algoritma C4.5, Decision Tree, K-Fold Cross Validation

ABSTRAK

Klasifikasi Nasabah Potensial merupakan suatu kegiatan yang dilakukan untuk mendapatkan nasabah yang benar-benar berpotensi dengan memperhatikan beberapa aspek yang menunjang agar penerima kredit layak menerima pinjaman. PT. Adira Finance SoE juga merupakan sebuah lembaga yang memberikan layanan kredit pinjaman terhadap nasabah dan sering kali masalah yang dihadapi ialah adanya penerimaan nasabah yang tidak sesuai prosedur sehingga menyebabkan adanya tunggakan pembayaran kredit oleh para nasabah. Dengan menggunakan algoritma C4.5 diharapkan dapat membantu menyelesaikan masalah yang ada dimana algoritma C4.5 digunakan untuk mempelajari karakteristik berdasarkan data nasabah sebelumnya agar menghindari ada nasabah yang melakukan tunggakan kredit. Pengujian yang dilakukan terhadap 1000 record data dengan menggunakan beberapa atribut sebagai penunjang algoritma C4.5 dimana atribut yang digunakan ialah, penghasilan, jumlah tanggungan, umur, status perkawinan, pekerjaan, tempat tinggal. Hasil pengujian ini mendapatkan pohon keputusan (decision tree) yang menentukan nasabah tersebut layak atau tidak mendapatkan pinjaman berdasarkan atribut-atribut yang ada dari data nasabah sebelumnya. Pengujian system dilakukan dengan menerapkan algoritma K-Fold Cross Validation pada saat telah dihitung dengan menggunakaan algoritma C4.5 dengan hasil dari 1000 record dimana 956 record berada pada kelas klasifikasi yang tepat dan 44 record berada pada kelas klasifikasi yang tidak tepat.

PENDAHULUAN

Pada era globalisasi saat ini dimana kebutuhan ekonomi sangatlah penting sehingga banyak sekali orang yang berusaha untuk mendapatkan uang dengan berbagai cara. Salah satu cara secara cepat yang sering dilakukan untuk mendapatkan uang adalah dengan mengajukan pinjaman/kredit kepada lembaga-lembaga perbankan atau kantor yang menjalankan pinjaman kredit kepada masyarakat salah satunya adalah PT. Adira Finance.

Adira dinamika multi finance merupakan sebuah perusahaan yang bergerak pada bidang peminjaman dana secara kredit tentunya dengan klasifikasi kriteria yang sesuai. Masalah yang sering dihadapi ialah adanya penerimaan nasabah yang tidak sesuai prosedur sehingga menyebabkan adanya tunggakan pembayaran kredit oleh para nasabah. Atributyang akan digunakan pada penelitian ini yaitu menggunakan data dari para nasabah yang sudah ada (data dari tahun 2015). Data yang digunakan yaitu data nasabah peminjam yang diambil dari Kantor Adira Finance SoE, dan data yang didapat berjumlah 1.000 *record*. Variabel yang dipakai dalam penelitian ini terdiri dari variabel umur, status perkawinan, pekerjaan, penghasilan, jumlah tanggungan, dan tempat tinggal. Dan variabel target yang ingin dicapai pada penelitian ini adalah layak atau tidak layak nya para calon peminjam mendapat kredit atau tidak.

Beberapa penelitian sebelumnya menggunakan algoritma C4.5 dilakukan oleh [1] Implementasi Algoritma C4.5 Untuk Klasifikasi Tingkat Kepuasan Pembeli Online Shop, [5] Klasifikasi Nasabah Menggunakan Algoritma C4.5 Sebagai Dasar Pemberian Kredit, [6] Implementasi Data Mining Klasifikasi Nasabah Potensial Menggunakan Algoritma C4.5. Hasil penelitian-penelitian tersebut mampu mendapatkan hasil berupa nasabah yang potensial dengan tingkat akurasi yang baik.

METODE PENELITIAN

Data miming memiliki beberapa tahapan untuk menghasilkan output yang berupa pengetahuan dimana output tersebut digunakan sebagai kontribusi atau pengetahuan baru, tahapan tersebut yaitu seleksi data, tahap *preprocessing* data, transformasi data, data mining atau penambangan datadan diakhiri tahap interpretasi dan evaluasi. Berikut merupakan penjalasan dari tahapan tersebut.

1. *Dataselection*

Proses awal penambangan data KDDdiawali dengan prosesseleksi datayang berasal dari data operasionalyang sudah terkumpul. Kemudiandatahasilseleksiyang digunakan dalam proses data mining, selanjutnya disimpan pada suatu berkas yangdipisahkan dari basis dataoperasional.Data sekunder berupa data nasabah dari Kantor Adira Finance SoE yang sudah ada sebelumnya. Untuk mendapat pohon keputusan digunakan atribut umur, status perkawinan, pekerjaan, penghasilan, jumlah tanggungan, dan tempat tinggal berdasarkan dari data nasabah yang sudah adasebelumnya.Berikut contoh data hasil seleksi seperti terlihat pada tabel 1.

Tabel 1. Data Selection PT ADIRA Finance Soe

No	Alamat (Kel)	Umur	Status	Pekerjaan	Penghasilan	Jumlah Tanggungan	Tempat Tinggal	Klasifikasi
1	Tuafanu	Dewasa	Menikah	Wiraswasta	Besar	Banyak	Kontrak	Tidak layak
2	Tubuhue	Dewasa	Belum Menikah	PNS	Kecil	Sedikit	Rumah	Layak
3	Oeekam	Paruh Baya	Menikah	Wiraswasta	Besar	Sedikit	Rumah	Layak
4	Oenasi	Dewasa	Belum Menikah	Wiraswasta	Besar	Banyak	Kontrak	Layak
5	Tuapukas	Paruh Baya	Menikah	PNS	Besar	Banyak	Rumah	Layak
6	Loli	Paruh Baya	Menikah	Wiraswasta	Kecil	Sedikit	Rumah	Layak
7	Oeekam	Dewasa	Menikah	Wiraswasta	Kecil	Banyak	Kontrak	TidakLayak
8	Kobekamusa	Dewasa	Menikah	PNS	Kecil	Sedikit	Rumah	Layak

9	Kampung Maleset	Dewasa	Menikah	Wiraswasta	Besar	Sedikit	Rumah	Layak

2. *Pre-processing/cleaning*

Preprocessing merupakan proses *cleaning*atau pembersihan yangmencakup diantaranyamembersihkan duplikasi data, pemeriksaan data yang berubah-ubah, dan memperbaiki jika ada kesalahan dalam data.prosescleaning data ini perlu dilakukan sebelum proses dataminingmenurut tahapan KDD.

3. *Transformation*

*Transformasi*merupakan suatu proses perubahan pada data yang telah ada, sehingga data yang sebelumnya tidak sesuai tersebut menjadi sesuai untuk proses penambangan data. Proses *transformasi*pada KDD adalah proses yang sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.Pada tahap pengkodean, data yang sudah diterima dengan benar diolah diberi kode sehingga tidak menimbulkan kebingungan saat melakukan pengisian ulang di aplikasinya sendiri. Dimana data alamat yang sudah didapat diberi kode Kel, dan juga memberikan *range* data pada atribut umur, penghasilan dan jumlah tanggungan.Hasil dari tahap *transformasi* datasepertiterlihat pada tabel 1, 2, dan 3.

Tabel1.*Range* Data Umur

No	Umur	*Range*
1	Dewasa	(20-40 Tahun)
2	Paruh Baya	(41-60 Tahun)

Tabel 2.*Range* Data Penghasilan

No	Umur	*Range*
1	Besar	(Diatas dari Rp. 2.500.000)
2	Kecil	(Dibawah dari Rp. 2.000.000)

Tabel 3. *Range* Jumlah Tanggungan

No	Umur	*Range*
1	Banyak	(Lebih dari 3 orang)
2	Sedikit	(Kurang dari 3 orang)

4. DataMining

Data mining merupakan proses pencarian pola atau informasi yang menarik pada data yang sudah di seleksi dan di transformasi sebelumnya dan kemudian data tersebut diolahdenganteknikataumetodedata mining tertentu.Dataminingmemilik teknik atau metodeatau algoritmayang sangat beragam.Pemilihanmetode manakah yang tepat sangat bergantung pada tujuan dan proses KDD yang ingin dicapai.

Algoritma C4.5 adalah peningkatan dari algoritma ID3, pada proses pohon keputusan data yang ada diubah bentuk menjadi model pohon kemudian diubah lagi menjadi *rule* setelah itu *rule* tersebut disederhanakan.

Langkah awal yang dilakukan dalam membentuk pohon keputusan adalah memilih atau menentukan atributyang menjadi akar dari pohon keputusan. Langkah dalam menentukan variabel yang menjadi akar adalah dengan menggunakan nilai *entropy*, *gain*, *split info*, dan *gain ratio*.

1. Entropy

*Entropy*merupakan suatu parameter atau kriteria dalam mengukur tingkat keberagaman

(heterogenitas) dari kumpulan data. Tingkat keberagaman suatu kumpulan data yang semakin besarakan berakibat pada nilai dari *entropy*yang jugasemakin besar. Rumus dalam menghitung *entropy* sebagai berikut:

$$Entropy\ (S) = \sum_{j=1}^{|k} - pj\ log2\ pj$$

Keterangan:
k = jumlah kelas
pj = jumlah proporsi (peluang) untuk kelas j

2. Gain

*Gain*merupakan ukuran efektifitas suatu variable dalam klasifikasi data. *Gain* suatu variabel merupakan selisih dari nilai *entropy* total dengan nilai *entropy* dari variabel tersebut. *Gain* dapat dirumuskan sebagai berikut:

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

Keterangan:
S : Himpunan Kasus
A : Atribut
N : Jumlah Partisi dalam A
|Si| : Jumlah kasus pada partisi ke-i
|S| : Jumlah Kasus dalam S

Nilai *gain* yang dihasilkan, akan digunakan dalam penentuan variabel yang menjadi nodedari suatu pohon keputusan.

3. Split Info

Split Info digunakan sebagai pembagi dari *Gain* (A) yang akan menghasilkan*Gain Ratio.*

$$SplitInfoA(D) = -\sum_{j=1}^{v} \frac{Dj}{D}\ log2\ (\frac{Dj}{D})$$

Keterangan:
D : Jumlah Kasus
A : Atribut
v : nilai yang mungkin dari variabel A
|Dj| : jumlah sampel nilai v
|D| : jumlah sampel seluruh sampel data
Entropy (Sv) : *Entropy* dari sampel yang memiliki nilai v

4. Gain Ratio

Dalam mengatasi masalah berupa nilai pada atribut yang sangat bervarisai dapat digunakan *Gain Ratio*dimana *gain ratio* ini merupakan salah satu ukuran lain yang dapat mengatasi masalah tersebut. Nilai

Gain Ratio tertinggi akan dipilih sebagai atribut *test* untuk simpul.

$$Gain\ Ratio\ (S,A) = \frac{Gain\ (S,A)}{Split\ Information(S,A)}$$

Keterangan:
S : Himpunan Kasus
A : Atribut

K-Fold CrossValidation

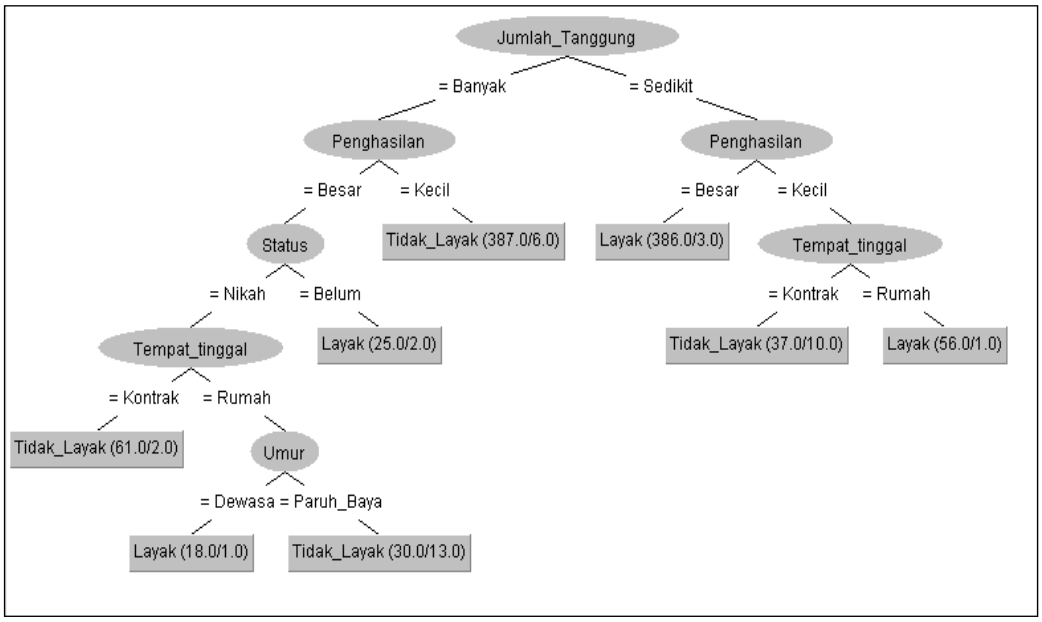
K-Fold cross validation adalahsalah satu metode untuk menilai model pengujian dari data pembelajaran dan data uji untuk mengolah data dengan seimbang [9]. Dalam *cross validation* data akan dibagi menjadi k buah partisi dengan ukuran yang sama variabel data ke-i (1,2,3,) danseterusnya [4]. selanjutnya proses pengujian dan data latih dilakukan sebanyak k kali.

5. *Interpretation* / *evalution*

Tahap *interpretation*adalahmerupakan bagian dari proses KDD dimana tahapan ini mencakuppemeriksaan data dan penenentuanuntuk menjawab pertanyaan apakahpolaatauinformasiyangditemukanbertentangan dengan fakta atau hipotesis yang ada.Pola data yang dihasilkan dari proses data mining inidiharapkan akan mudah dimengerti oleh pihak yang berkepentingan.

HASIL DAN PEMBAHASAN

Berdasarkan hasil perhitungan terhadap 1000 *record* data selanjutnya melewati tahapan penerapan algoritma menggunakan algoritma C4.5 di*software* WEKA 3.8, hasil yang didapat bisa dilihat pada gambar 1.



Gambar 1. Hasil *Decision Tree*

Dari gambar 3.1 diatas maka didapatlah *rule* seperti dibawah ini:

1. *IF* jumlah tanggung banyak, penghasilan besar, status nikah, tempat tinggal rumah, umur dewasa, *THEN* layak
2. *IF* jumlah tanggung banyak, penghasilan besar, status belum menikah *THEN*layak

3. *IF* jumlah tanggung sedikit penghasilan besar *THEN*layak
4. *IF* jumlah tanggung sedikit, penghasilan kecil, tempat tinggal rumah *THEN*layak
5. *IF* jumlah tanggung banyak, penghasilan besar, status nikah, tempat tinggal rumah, umur paruh baya *THEN* tidaklayak
6. *IF*jumlahtanggungbanyak,penghasilanbesar,statusnikah,tempattinggal,kontrak*THEN* tidak layak
7. *IF* jumlah tanggung banyak, penghasilan kecil *THEN* tidaklayak
8. *IF* jumlah tanggung sedikit, penghasilan kecil, tempat tinggal kontrak *THEN* tidak layak

Hasil PengujianMenggunakanAlgoritma*K-Fold CrossValidation*

Dari 10-fold cross validation yang dipakai untuk menguji kevalidan data dengan menggunakan algoritma C4.5 maka didapatlah hasil yang dapat dilihat pada gambar2.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         956               95.6    %
Incorrectly Classified Instances        44                4.4    %
Kappa statistic                          0.912
Mean absolute error                      0.0636
Root mean squared error                  0.1846
Relative absolute error                 12.7178 %
Root relative squared error             36.9276 %
Total Number of Instances             1000
```

Gambar 2. Hasil Uji *K-Fold Cross Validation*

KESIMPULAN DAN SARAN

Hasil penelitian "Penggunaan Data Mining Algoritma C4.5 dalam Menentukan Klasifikasi data Nasabah Potensial Di PT. Adira Finance SoE", dapat disimpulkan sebagai berikut:

1. Hasil Analisa data dengan metode C4.5 dapat berjalan dengan baik. Sistem mampu membagi 1000 *record* data kedalam 2 kelompok klasifikasi kelas layak dan tidaklayak.
2. Pengujian sistem menggunakan metode *K-Fold Cross Validation* dengan k = 10, dilakukan terhadap 1000 record. hasil dari pengujian tersebut mendapat nilai kebenaran klasifikasi melalui algoritma C4.5 dengan tingkat persentase yang cukup besar yaitu lebih dari95%

Saran dari penelitian ini adalah sebagai berikut:

1. Pada penelitian lanjutan tentang kasus ini dapat menggunakan algoritma yang berbeda untuk membandingkan kebenaran tingkat klasifikasipola baik itu jenis klasifikasi maupun clustering.
2. Penambahan jumlah data yang lebih banyak lagi pada penelitianselanjutnya untuk lebih dapat memperoleh pola data yang rinci dengan tingkat uji validitas yang lebih baik.

DAFTAR PUSTAKA

[1] Febriyanto, D.B., Handoko, L., Wahyuli, W., Aisyah, H. &Rumini, R. 2018. Implementasi Algoritma C4. 5 Untuk Klasifikasi Tingkat Kepuasan Pembeli Online Shop. *JURIKOM (JurnalRisetKomputer)*, 5(6):569–575.

[2] PITALOKA, G.F. t.t. SYNTHETIC MINORITY OVERSAMPLING TECHNIQUEUNTUK MENGATASI IMBALANCECLASS.

[3] Rani, L.N. 2016. Klasifikasi Nasabah Menggunakan Algoritma C4. 5 SebagaiDasar Pemberian Kredit. *INOVTEK Polbeng-Seri Informatika*, 1(2):126–132.

[4] Sholihah, H., Satria, F. & Muslihudin, M. 2018. Implementasi Algoritma C4. 5 Klasifikasi Nasabah Potensial ADIRA Dinamika Multi Finance Pringsewu. *Konferensi Nasional Sistem Informasi (KNSI)2018*.

[5] Suntoro, J. 2019. *DATA MINING: Algoritma dan Implementasi dengan Pemrograman php*. Elex Media Komputindo.

[6] Triayudi, A. &Susilawati, R. 2015. Klasifikasi Calon Nasabah Pembiayaan Pada Pt Sinar Mitra Sepadan Finance MenggunakanAlgoritma C4. 5. *ProTekInfo (Pengembangan Riset dan Observasi Teknik Informatika)*, 2:59–62.

[7] Utama, I.G.B.R. & SE, M. 2018. *Statistik Penelitian Bisnis dan Pariwisata (Dilengkapi Studi Kasus Penelitian)*. PenerbitAndi.