

PENERAPAN DATA MINING UNTUK MEMPREDIKSI POLA NASABAH MENGGUNAKAN ALGORITMA C4.5

Donny Maulana¹⁾, Raden Rangga Bramantya Putra²⁾

Program Studi Teknik Informatika Fakultas Teknik
Universitas Pelita Bangsa
donny.maulana@pelitabangsa.ac.id

Disetujui, 30 Desember 2019

Abstraksi

Data Mining merupakan proses analisa data dari sudut yang berbeda dan mengolahnya menjadi informasi-informasi penting yang bisa digunakan untuk meningkatkan keuntungan. Secara teknis, data mining dapat disebut juga sebagai proses untuk menemukan korelasi atau pola dari ratusan atau ribuan field. Pada data mining dapat digunakan metode decision tree untuk melakukan klasifikasi. Kredit tidak lagi menjadi hal yang asing bagi masyarakat luas, banyaknya kebutuhan membuat masyarakat mengambil kredit untuk memenuhi kebutuhan konsumtif mereka. Sering kali pembayaran kredit yang macet membuat bank kerepotan, karena semakin banyaknya nasabah yang membayar secara macet dapat berdampak buruk untuk kesehatan bank. Oleh karena itu, data nasabah dari Bank XYZ menjadi bahan acuan untuk menganalisis pola nasabah pemohon kredit. Pemohon kredit termasuk dalam kategori lancar, atau macet. Untuk mengatasi masalah tersebut, dapat digunakan metode decision tree. Sehingga penelitian ini bisa dijadikan acuan pihak Bank untuk menilai nasabah dengan record data yang ada untuk pengambilan kredit selanjutnya. Informasi yang ditampilkan berupa tingkat akurasi data nasabah lancar dan macet. Kemudian hasil akurasi dari aplikasi yang diimplementasikan akan dibandingkan dengan hasil menggunakan software rapidminer. Sehingga diperoleh akurasi dengan decision tree sebesar 95%. Dapat disimpulkan bahwa algoritma C4.5 dapat digunakan untuk melakukan klasifikasi dengan cukup baik dan dapat mengukur klasifikasi nasabah.

Kata kunci: algoritma C4.5, data mining, klasifikasi, decision tree.

Abstract

Data mining is the process of analyzing data from different angles and process them into important information that can be used to increase profits. Technically, data mining can also be called as a process of finding correlation or patterns of hundreds or thousand of field. In data mining decision tree method can be used for the classification. Credit is no longer a strange thing for the general public, many need to make people take credit to meet their consumer needs. Often jammed credit payments made bank hassles, because more and more customers who pay jams can be bad for the health of banks. Thereof, customer data from Bank XYZ become are ference for analyzing pattern of customer credit applicants. Credit applicants included in the current category, or jammed. To overcome these problems, can use decision tree method. So this research could be used as a reference for Bank to assess clients with existing data record for further credit decision. The information is displayed in the form of the accuracy of customer data smoothly and jammed. Then the accuracy of the results of application that is implemented will be compared with results using software Rapidminer. This obtained with less accuracy of decision tree by 95%. It can be concluded that the C4.5 algorithm can be used to classify fairly well and can measure customer classifications.

Keywords: algorithm C4.5, data mining, classification, decision tree.

1. Pendahuluan

Kredit bukan merupakan sesuatu yang asing lagi bagi masyarakat luas, masyarakat desa saja mengenal kredit. Banyak masyarakat mengambil kredit untuk memenuhi kebutuhan mereka, seperti, untuk membangun rumah, kebutuhan sehari-hari, modal, biaya anak sekolah, dan lain-lain. Kredit adalah kepercayaan.

PT. Bank XYZ adalah salah satu bank yang menyediakan layanan Kredit Pegawai kepada para nasabahnya. Pengajuan kredit Pegawai bisa dilakukan oleh pemohon yang telah memenuhi syarat dari pihak bank. Pemberian kredit ini masih menjadi masalah yang krusial, karena pihak bank masih sangat sulit untuk menentukan pola debitur atau pemohon yang menyebabkan kredit lancar atau macet. Banyak faktor yang menyebabkan kredit macet, diantaranya jumlah penghasilan tiap bulan dan banyaknya kebutuhan ekonomi. Pembayaran kredit yang macet membuat performa bank menjadi buruk.

Pada Bank XYZ Cabang kredit yang banyak diambil adalah kredit pegawai, di mana hanya yang berpenghasilan tetapi yang diperbolehkan mengambil kredit tersebut. Syarat untuk mengajukan kredit pada Bank XYZ juga sangatlah mudah, tetapi walaupun demikian, masih ada juga nasabah yang membayar kredit secara macet. Apalagi jika nasabah tersebut tiba-tiba menghilang, tidak terlihat lagi dalam instansinya. Sedangkan untuk nasabah yang meninggal sebelum kreditnya lunas, biaya itu ditanggung oleh asuransi.

Pembayaran kredit nasabah yang macet ini dapat berdampak negatif untuk kesehatan Bank, karena apabila nilai NPL Non Performing Loan pada Bank ini mencapai 2% dapat dikatakan Bank tidak memiliki performa yang bagus atau bahkan bangkrut, karena seharusnya nilai NPL pada Bank adalah di bawah 2%. Untuk menekan nilai NPL, pihak Bank harus selektif terhadap nasabah kredit yang membayar secara macet pada kredit. Sehingga penelitian ini bisa dijadikan sebagai acuan pihak Bank untuk menilai nasabah dengan record data yang ada untuk pengambilan kredit selanjutnya.

Oleh karena itu, data nasabah dari Bank XYZ menjadi bahan acuan untuk menganalisa pola nasabah pemohon kredit. Pemohon kredit termasuk dalam kategori lancar, diperlukan perhatian khusus, kurang lancar, atau macet. Dalam kasus ini digunakan teknik data mining decision tree untuk mengklasifikasi pola nasabah dengan menggunakan pohon keputusan algoritma C4.5.

Algoritma C4.5 diperkenalkan oleh Quinlan 1996 sebagai versi perbaikan dari ID3. Dalam ID3, pada decision tree hanya bisa digunakan untuk fitur bertipe kategorikal saja nominal atau ordinal, sedangkan untuk tipe numerik interval atau rasio tidak dapat digunakan. Perbaikan yang membedakan algoritma C4.5 dari ID3 adalah dapat menangani fitur dengan tipe numerik, melakukan pemotongan decision tree, dan penurunan rule set. Algoritma C4.5 juga menggunakan kriteria gain dalam menentukan fitur yang menjadi pemecah node pada pohon yang diinduksi.

Berdasarkan pertimbangan di atas dan dorongan untuk menerapkan pengetahuan yang diperoleh, maka peneliti melakukan suatu penelitian ilmiah yang membahas tentang “penerapan data mining untuk memprediksi pola nasabah menggunakan algoritma c4.5”.

2. Tinjauan Studi

2.1. Data Mining

Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar. Data mining merupakan serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual.

2.2. Algoritma C4.5

Algoritma C4.5 adalah algoritma yang sangat populer digunakan oleh banyak peneliti di dunia, hal ini dijelaskan oleh Xindong Wu dan Vipin Kumar dalam bukunya yang berjudul *The Top Ten Algorithms in Data Mining*. Algoritma C4.5 merupakan pengembangan dari algoritma ID3 yang diciptakan oleh J. Rose Quinlan.

3. Desain Penelitian/Metodologi

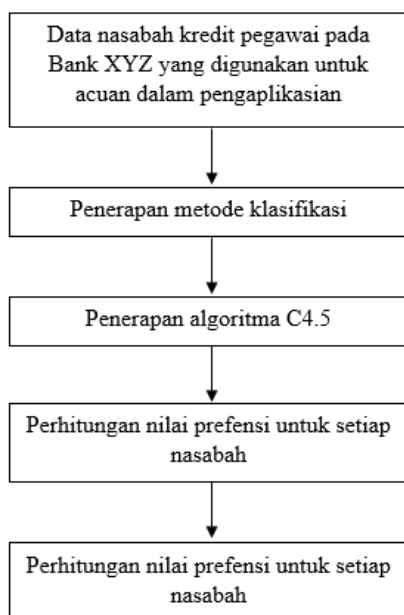
3.1. Metode Pengumpulan Data

Metode pengumpulan data yang dilakukan peneliti untuk mendapatkan data yang valid adalah sebagai berikut:

1. Metode Interview atau Wawancara
Metode wawancara dilakukan dengan melakukan wawancara terhadap pegawai Bank bagian kredit XYZ untuk mendapatkan informasi seputar syarat kredit dan kriteria kredit. Selain itu juga untuk dijadikan atribut pendukung pada penelitian ini.
2. Metode Studi Literatur
Pada metode studi literatur, peneliti mengumpulkan, membaca, mempelajari, dan mencatat literatur dari jurnal maupun buku yang berkaitan dengan algoritma C4.5. Dari metode studi literatur diharapkan dapat memberi gambaran yang dapat diimplementasikan pada aplikasi yang digunakan peneliti dalam menyelesaikan penelitian ini.

3.2. Teknik Analisa Data

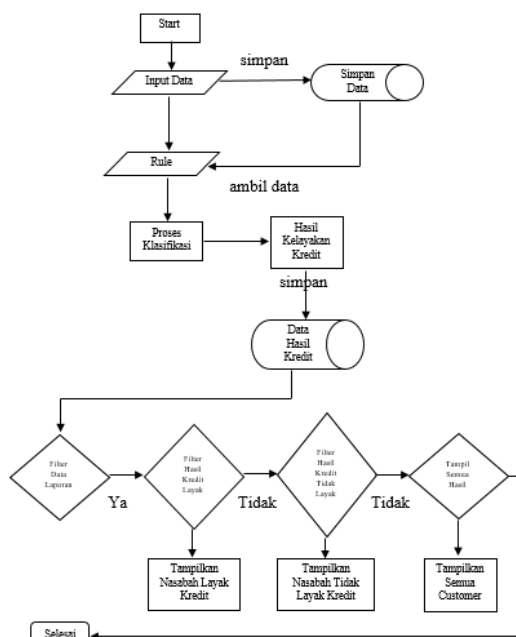
Metode yang diusulkan untuk proses seperti ini menggunakan metode klasifikasi dengan algoritma C4.5. Metode yang diusulkan untuk proses penentuan nasabah kredit adalah seperti gambar di bawah ini:



Gambar 1. Alur yang diusulkan

3.3. Metode Yang Diusulkan

Dalam penelitian ini akan dilakukan analisa menggunakan algoritma C4.5. Data dihitung dengan algoritma sesuai dengan metodenya, kemudian dicari hasil akurasi. Dalam tahapan ini akan dilakukan beberapa langkah pengujian data yaitu seperti berikut:



Gambar 2. Flowchart Program

4. Hasil Penelitian Dan Pengujian

4.1. Pengujian Data

Pada penelitian ini digunakan 1000 data nasabah untuk menentukan akurasi klasifikasi. Terdapat 300 data nasabah yang digunakan sebagai data sample untuk perhitungan entropy dan gain secara manual. Berikut data testing yang digunakan :

Tabel 1. Tabel Data Testing

No	Nama	No Rekening	Status Pernikahan	Status Rumah	Umur	Penghasilan	Angsuran	Jangka Waktu	Status Label
1	Yuli Andika	72210111	MPLKANI	rumah sendiri	45	3100000	<=300000	12	Macet
2	Arifa Pita Vikiati	72210112	MPLKANI	rumah sendiri	42	3000000	>300000	12	Lancar
3	Akshita Cahilla	72210113	RELIAN MPLKANI	rumah sendiri	24	4000000	>300000-1000000	24	Macet
4	Alhadi Naji	72210114	MPLKANI	kontrak	22	4800000	>300000-1000000	24	Lancar
5	Analysia Wijaya	72210115	MPLKANI	rumah sendiri	24	3700000	>300000	24	Lancar
6	Agil Mubandah Rahmat	72210116	RELIAN MPLKANI	kontrak	22	4800000	>300000-1000000	24	Lancar
7	Arifa Pita	72210117	RELIAN MPLKANI	kontrak	24	3700000	>300000	24	Lancar
8	Andeananda	72210118	MPLKANI	kontrak	24	4800000	<=300000	24	Macet
9	Fauzan Agustin Pratomo	72210119	MPLKANI	rumah sendiri	43	3000000	>300000-1000000	36	Lancar
10	Hana Azzila	72210120	MPLKANI	rumah sendiri	35	4000000	<=300000	24	Lancar
11	Hana Hanika	72210121	RELIAN MPLKANI	rumah sendiri	21	3800000	<=300000	24	Macet
12	Rian Fadhil	72210122	MPLKANI	rumah sendiri	48	3000000	>300000	12	Lancar
13	Mardiana Mubana Almal	72210123	MPLKANI	rumah sendiri	24	4800000	>300000-1000000	24	Macet
14	Mubandah Adh Husna	72210124	RELIAN MPLKANI	rumah sendiri	24	3800000	<=300000	24	Macet
15	Mubandah Fiqri Rahmat	72210125	MPLKANI	rumah sendiri	24	3800000	<=300000	36	Macet
16	Nida Fadhil Dini	72210126	MPLKANI	kontrak	21	4700000	>300000-1000000	36	Lancar
17	Rafi Mubandah	72210127	MPLKANI	kontrak	28	3100000	>300000	24	Lancar
18	Suliyah Kurniawan	72210128	RELIAN MPLKANI	kontrak	24	3100000	<=300000	24	Macet
19	Suzana Devi Vack	72210129	MPLKANI	kontrak	38	3800000	>300000-1000000	12	Lancar
20	Sulhan Zulfa Aul	72210130	MPLKANI	rumah sendiri	38	4800000	<=300000	36	Lancar
21	Siti Ananda Azzila	72210131	MPLKANI	rumah sendiri	48	3800000	<=300000	24	Macet
22	Mubandah Nabil Alimurrahman	72210132	MPLKANI	kontrak	21	3800000	>300000	24	Lancar
23	Nida Khatiyah	72210133	RELIAN MPLKANI	kontrak	22	4800000	>300000-1000000	12	Macet
24	Zuliyah Azzila	72210134	MPLKANI	rumah sendiri	38	3800000	<=300000	12	Macet
25	Naura Hanu Hanafi	72210135	MPLKANI	kontrak	24	3000000	<=300000	24	Macet

4.2. Pemodelan Data Mining Menggunakan Algoritma C4.5

Berikut ini merupakan langkah-langkah perhitungan nilai Entrophy dan Gain pada setiap atribut yang memiliki label Lancar dan Macet :

1. Perhitungan Nilai Entrophy

Langkah awal algoritma C4.5 adalah mencari nilai entrophy. Pertama, tentukan terlebih dahulu nilai entrophy total dalam kasus. Rumusnya sebagai berikut :

$$Entropy = -\sum pi * \log_2 pini = 1 \quad (4.1)$$

	Simpul	Jumlah Kasus	Kolektibilitas	
			Lancar	Macet
Total		300	156	144
Status Pernikahan	Belum Menikah	96	51	45
	Menikah	204	105	99
Status Rumah	Rumah Sendiri	198	99	99
	Kontrak	102	57	45
Umur	<=25	103	54	49
	>25	197	102	95
Umur	<=35	157	80	77
	>35	143	76	67
Umur	<=45	199	101	98
	>45	101	55	46
Penghasilan	<=3900000	111	0	111
	>3900000	189	156	33
Penghasilan	<=4650000	196	52	144
	>4650000	104	104	0
Angsuran	<=700000	144	29	115
	>=800000-1500000	92	63	29
Jangka Waktu	12	59	26	33
	24	130	74	56
	36	111	56	55

Maka :
Entropy Total (S)

$$\begin{aligned}
 & \text{Entropy total (101,99)} \\
 & = (-156/300) * \log_2(156/300) + (-144/300) * \log_2(144/300) \\
 & = (-0,52) * (-0,94341647163) + (-0,48) * (-1,05889368905) \\
 & = 0,49057656524 + 0,50826897074 \\
 & = 0,999
 \end{aligned}$$

Selanjutnya lakukan perhitungan entropy pada tiap atribut berdasarkan pada jumlah kasus per subset atribut, yaitu :

a. Perhitungan nilai subset pada atribut status pernikahan

$$\begin{aligned}
 & \text{Entropy Belum Menikah (51, 45)} \\
 & = (-51/96) * \log_2(51/96) + (-45/96) * \log_2(45/96) \\
 & = (-0,531) * (-0,91253715875) + (-0,469) * (-1,09310940439) \\
 & = 0,48455723129 + 0,51266831066 \\
 & = 0,997
 \end{aligned}$$

$$\begin{aligned}
 & \text{Entropy Menikah (105, 99)} \\
 & = (-105/204) * \log_2(105/204) + (-99/204) * \log_2(99/204) \\
 & = (-0,515) * (-0,9581798243) + (-0,485) * (-1,04306872189) \\
 & = 0,49346260951 + 0,50588833011 \\
 & = 0,999
 \end{aligned}$$

b. Perhitungan nilai subset pada atribut status rumah

$$\begin{aligned}
 & \text{Entropy Rumah Sendiri (99, 99)} \\
 & = (-99/198) * \log_2(99/198) + (-99/198) * \log_2(99/198) \\
 & = (-0,5) * (-1) + (-0,5) * (-1) \\
 & = 0,5 + 0,5 \\
 & = 1
 \end{aligned}$$

$$\begin{aligned}
 & \text{Entropy Kontrak (57, 45)} \\
 & = (-57/102) * \log_2(57/102) + (-45/102) * \log_2(45/102) \\
 & = (-0,559) * (-0,8395353278) + (-0,441) * (-0,896164189) \\
 & = 0,46930024824 + 0,52063236032 \\
 & = 0,99
 \end{aligned}$$

c. Perhitungan nilai subset pada atribut umur

$$\begin{aligned}
 & \text{Entropy } \leq 25 \text{ (54, 49)} \\
 & = (-54/103) * \log_2(54/103) + (-49/103) * \log_2(49/103) \\
 & = (-0,524) * (-0,93161302502) + (-0,476) * (-1,07179068307) \\
 & = 0,48816522511 + 0,51017236514 \\
 & = 0,998
 \end{aligned}$$

$$\begin{aligned}
 & \text{Entropy } > 25 \text{ (102, 95)} \\
 & = (-102/197) * \log_2(102/197) + (-18/60) * \log_2(18/60) \\
 & = (-0,518) * (-0,94962647748) + (-0,482) * (-1,05219621113) \\
 & = 0,49190651533 + 0,50715857376 \\
 & = 0,999
 \end{aligned}$$

d. Perhitungan nilai subset pada atribut umur

$$\begin{aligned}
 & \text{Entropy } \leq 35 \text{ (80, 77)} \\
 & = (-80/157) * \log_2(80/157) + (-77/157) * \log_2(77/157) \\
 & = (-0,510) * (-0,972692654) + (-0,490) * (-1,0278342082) \\
 & = 0,49607325354 + 0,50363876201 \\
 & = 1
 \end{aligned}$$

$$\begin{aligned}
 & \text{Entropy } > 35 \text{ (76, 67)} \\
 & = (-76/143) * \log_2(76/143) + (-67/143) * \log_2(67/143) \\
 & = (-0,531) * (-0,91194382333) + (-0,469) * (-1,09378214632) \\
 & = 0,48424217019 + 0,51298382662 \\
 & = 0,997
 \end{aligned}$$

e. Perhitungan nilai subset pada atribut umur

$$\text{Entropy } \leq 45 \text{ (101, 98)}$$

$$\begin{aligned}
 &= (- (101/199) * \log_2 (101/199) + (- (98/199) * \log_2 (98/199)) \\
 &= (- (0,508) * (-0,97841313779) + (- (0,492) * (-1,02191477643)) \\
 &= 0,49703387399 + 0,50278207 \\
 &= 1
 \end{aligned}$$

Entropy >45 (55, 46)

$$\begin{aligned}
 &= (- (55/101) * \log_2 (55/101) + (- (46/101) * \log_2 (46/101)) \\
 &= (- (0,544) * (-0,87685176922) + (- (0,456) * (-1,13464952669)) \\
 &= 0,47700736246 + 0,51740018417 \\
 &= 0,994
 \end{aligned}$$

f. Perhitungan nilai subset pada atribut penghasilan

Entropy <=3900000 (0, 111)

$$\begin{aligned}
 &= (- (0/111) * \log_2 (0/111) + (- (111/111) * \log_2 (111/111)) \\
 &= (- (0) * (0) + (- (1) * (0)) \\
 &= 0 + 0 \\
 &= 0
 \end{aligned}$$

Entropy >3900000 (156, 33)

$$\begin{aligned}
 &= (- (156/189) * \log_2 (156/189) + (- (33/189) * \log_2 (33/189)) \\
 &= (- (0,825) * (-0,27684020535) + (- (0,175) * (-2,51784830486)) \\
 &= 0,22839316942 + 0,44062345335 \\
 &= 0,668
 \end{aligned}$$

g. Perhitungan nilai subset pada atribut penghasilan

Entropy <=4650000 (52, 144)

$$\begin{aligned}
 &= (- (52/196) * \log_2 (52/196) + (- (144/196) * \log_2 (144/196)) \\
 &= (- (0,265) * (-1,91427012597) + (- (0,735) * (-0,44478484267)) \\
 &= 0,50728158338 + 0,32691685936 \\
 &= 0,835
 \end{aligned}$$

Entropy >4650000 (104, 0) 6

$$\begin{aligned}
 &= (- (104/104) * \log_2 (104/104) + (- (0/104) * \log_2 (0/104)) \\
 &= (- (1) * (0) + (- (0) * (0)) \\
 &= 0 + 0 \\
 &= 0
 \end{aligned}$$

h. Perhitungan nilai subset pada atribut angsuran

Entropy <=700000 (29, 115)

$$\begin{aligned}
 &= (- (29/144) * \log_2 (29/144) + (- (115/144) * \log_2 (115/144)) \\
 &= (- (0,201) * (-2,31194400631) + (- (0,798) * (-0,32443495049)) \\
 &= 0,46470074526 + 0,25889909049 \\
 &= 0,725
 \end{aligned}$$

Entropy >=800000-1500000 (63, 29)

$$\begin{aligned}
 &= (- (63/92) * \log_2 (63/92) + (- (29/92) * \log_2 (29/92)) \\
 &= (- (0,685) * (-0,54628203255) + (- (0,315) * (-1,66558096093)) \\
 &= 0,3742031923 + 0,52465800269 \\
 &= 0,899
 \end{aligned}$$

Entropy >1500000 (64, 0)

$$\begin{aligned}
 &= (- (64/64) * \log_2 (64/64) + (- (0/64) * \log_2 (0/64)) \\
 &= (- (1) * (0) + (- (0) * (0)) \\
 &= 0 + 0 \\
 &= 0
 \end{aligned}$$

i. Perhitungan nilai subset pada atribut jangka waktu

Entropy 12 (26, 33)

$$\begin{aligned}
 &= (- (26/59) * \log_2 (26/59) + (- (33/59) * \log_2 (33/59)) \\
 &= (- (0,441) * (-1,18220333122) + (- (0,559) * (-0,83824893)) \\
 &= 0,52135166906 + 0,46858115187 \\
 &= 0,99
 \end{aligned}$$

Entropy 24 (74, 56)

$$\begin{aligned}
 &= (- (74/130) * \log_2 (74/130) + (- (56/130) * \log_2 (56/130)) \\
 &= (- (0,570) * (-0,8129144474) + (- (0,430) * (-1,21501289097)) \\
 &= 0,46336123501 + 0,52245554311 \\
 &= 0,986
 \end{aligned}$$

Entropy 36 (56, 55)

$$\begin{aligned}
 &= (- (56/111) * \log_2 (56/111) + (- (55/111) * \log_2 (55/111)) \\
 &= (- (0,505) * (-0,98706094429) + (- (0,495) * (-1,01305615283)) \\
 &= 0,49846577686 + 0,50146279564 \\
 &= 1
 \end{aligned}$$

2. Perhitungan Nilai Gain

Setelah semua perhitungan entropy pada masing-masing subset atribut selesai, kemudian lakukan perhitungan nilai gain, dengan rumus :

$$\text{Gain (S,A)} = \text{Entropy(S)} - \sum |S_i|/|S| * \text{Entropy (S)}_{n_i=i} \quad (4.2)$$

a. Gain status pernikahan

$$\begin{aligned}
 &= 0,999 - (96/300 * 0,997) + (204/300 * 0,999) \\
 &= 0,999 - (0,31904) + (0,67932) \\
 &= 0,999 - 0,998 \\
 &= 0,001
 \end{aligned}$$

b. Gain status rumah

$$\begin{aligned}
 &= 0,999 - (198/300 * 1) + (102/300 * 0,99) \\
 &= 0,999 - (0,66) + (0,3366) \\
 &= 0,999 - 0,997 \\
 &= 0,002
 \end{aligned}$$

c. Gain umur

$$\begin{aligned}
 &= 0,999 - (103/300 * 0,998) + (197/300 * 0,999) \\
 &= 0,999 - (0,343) + (0,656) \\
 &= 0,999 - 0,999 \\
 &= 0
 \end{aligned}$$

d. Gain umur

$$\begin{aligned}
 &= 0,999 - (157/300 * 1) + (143/300 * 0,997) \\
 &= 0,999 - (0,523) + (0,476) \\
 &= 0,999 - 0,999 \\
 &= 0
 \end{aligned}$$

e. Gain umur

$$\begin{aligned}
 &= 0,999 - (199/300 * 1) + (101/300 * 0,994) \\
 &= 0,999 - (0,663) + (0,335) \\
 &= 0,999 - 0,998 \\
 &= 0,001
 \end{aligned}$$

f. Gain penghasilan

$$\begin{aligned}
 &= 0,999 - (111/300 * 0) + (189/300 * 0,668) \\
 &= 0,999 - (0) + (0,421) \\
 &= 0,999 - 0,4217
 \end{aligned}$$

$$= 0,578$$

g. Gain Penghasilan

$$\begin{aligned}
 &= 0,999 - (196/300 * 0,835) + (104/300 * 0) \\
 &= 0,999 - (0,546) + (0) \\
 &= 0,999 - 0,546 \\
 &= 0,453
 \end{aligned}$$

h. Gain angsuran

$$\begin{aligned}
 &= 0,999 - (144/300 * 0,725) + (64/300 * 0) + (92/300 * 0,899) \\
 &= 0,999 - (0,348) + (0) + (0,276) \\
 &= 0,999 - 0,624 \\
 &= 0,375
 \end{aligned}$$

i. Gain jangka waktu

$$\begin{aligned}
 &= 0,999 - (59/300 * 0,99) + (130/300 * 0,986) + (111/300 * 1) \\
 &= 0,999 - (0,1947) + (0,4273) + (0,37) \\
 &= 0,999 - 0,9983 \\
 &= 0,0007
 \end{aligned}$$

Berikut adalah tampilan hasil hitungan entropy dan gain yang selengkapnya dalam bentuk tabel :

Tabel 4.3 Tabel Hasil Perhitungan Nilai Entropy dan Gain

Nilai gain tertinggi dijadikan node akar pertama. Untuk node akar selanjutnya, dilakukan perhitungan lagi sampai semua tupel terpenuhi.

4.3 Perancangan Proses

1. Dari hasil pengujian sistem menggunakan bahasa pemrograman php dengan metode decision tree C4.5 dengan jumlah data sebanyak 300 data record, maka menghasilkan 279 data teridentifikasi benar dan 21 data teridentifikasi salah. Untuk menentukan akurasi kinerja dari sistem ini maka nilai hasil pengujian akan dimasukkan ke dalam persamaan berikut:

a. Akurasi

Perhitungan akurasi dilakukan dengan cara membagi jumlah data yang diklasifikasi secara benar dengan total sample data testing yang diuji.

$$\begin{aligned}
 &= \text{hasil benar} / \text{jumlah data} \times 100\% \\
 &= 279 / 300 \times 100\% \\
 &= 93\%
 \end{aligned}$$

b. Precision

Nilai precision dihitung dengan cara membagi jumlah data benar yang bernilai positif (True Positive) dibagi dengan jumlah data yang bernilai positif (True Positive) dan data salah yang bernilai (False Positive).

$$\begin{aligned}
 &= \text{tptp} / (\text{tptp} + \text{fp}) \times 100\% \\
 &= 0 / (0 + 21) \times 100\% \\
 &= 0\%
 \end{aligned}$$

c. Recall

Sedangkan recall dihitung dengan cara membagi data benar yang bernilai positive (tp) dengan hasil penjumlahan dari data benar yang bernilai positif (tp) dan data salah yang bernilai negative (fn).

$$\begin{aligned}
 &= \text{tptp} / (\text{tptp} + \text{fn}) \times 100\% \\
 &= 0 / (0 + 0) \times 100\% \\
 &= 0\%
 \end{aligned}$$

2. Dari hasil pengujian menggunakan software rapidminer dengan metode decision tree C4.5 dengan jumlah data

8

sebanyak 300 data record, maka menghasilkan 291 data teridentifikasi benar dan 9 data teridentifikasi salah. Berikut adalah hasil akurasi, precision, dan recall:

a. Akurasi

Perhitungan akurasi dilakukan dengan cara membagi jumlah data yang diklasifikasi secara benar dengan total sample data testing yang diuji.

$$\begin{aligned}
 &= \text{hasil benar} / \text{jumlah data} \times 100\% \\
 &= 291 / 300 \times 100\% \\
 &= 97\%
 \end{aligned}$$

b. Precision

Nilai precision dihitung dengan cara membagi jumlah data benar yang bernilai positif (True Positive) dibagi dengan jumlah data yang bernilai positif (True Positive) dan data salah yang bernilai (False Positive).

$$\begin{aligned} &= \frac{tp}{tp+fp} \times 100\% \\ &= \frac{140}{140+4} \times 100\% \\ &= 97,22\% \end{aligned}$$

c. Recall

Sedangkan recall dihitung dengan cara membagi data benar yang bernilai positive (tp) dengan hasil penjumlahan dari data benar yang bernilai positif (tp) dan data salah yang bernilai negative (fn).

$$\begin{aligned} &= \frac{tp}{tp+fn} \times 100\% \\ &= \frac{140}{140+151} \times 100\% \\ &= 48,10\% \end{aligned}$$

Dari kedua hasil percobaan diatas, menggunakan software Rapidminer dan pemrograman php dapat diambil rata-rata sebagai berikut:

$$\begin{aligned} &= \frac{\text{jumlah akurasi software} + \text{jumlah akurasi Rapidminer2}}{2} \\ &= \frac{93\% + 97\%}{2} \\ &= 95\% \end{aligned}$$

5. Kesimpulan

Berdasarkan hasil penelitian ini maka di dapat beberapa kesimpulan sebagai berikut:

1. Bahwa penerapan data mining klasifikasi pola nasabah menggunakan algoritma C4.5 pada Bank XYZ menghasilkan akurasi yang tinggi, dengan menggunakan tools Rapidminer, yaitu 95%.

Daftar Pustaka