Teorema: *Teori dan Riset Matematika*, *7*(1), 101–116, Maret 2022
p-ISSN 2541-0660, e-ISSN 2597-7237 © 2022
https://jurnal.unigal.ac.id/index.php/teorema/article/view/6692

IMPLEMENTASI ALGORITMA C5.0 DALAM MENGANALISA KELAYAKAN PENERIMA KERINGANAN UKT MAHASISWA ITK

Rizkya Nur Amalda[1], Nashrul Millah[2], Irma Fitria[3]
[1,3] Institut Teknologi Kalimantan, Jl. Soekarno-Hatta KM. 15, Balikpapan, Kalimantan Timur, Indonesia
[2] Universitas Airlangga, Jl. Mulyorejo, Surabaya, Jawa Timur, Indonesia
Email: rizkiaalda9@gmail.com

ABSTRACT

The COVID-19 is not only detrimental in terms of health, but also weakens various sectors in Indonesia, especially the economic sector. Government policies that limit public activities have an impact on weakening economic activity. This can be in line with the number of UKT (Single Tuition Fee) appeals from Kalimantan Institute of Technology (ITK) students. The bureaucrats have facilitated students who wish to apply for UKT relief. However, the submission process is still done manually and is prone to human errors. In analyzing the application, we need a method that can help in making a decision whether the application is acceptable or not, one of which is the C5.0 Algorithm method. The C5.0 algorithm is an algorithm with a decision tree model that can process data into a rules that can later be used as suggestions in decision making. It is to be hoped that this system will be able to assist ITK in making decisions and determining the amount of UKT more precisely and objectively. Based on the results of the evaluation of 108 data from ITK students who requested an appeal for UKT, there was an accuracy of 91% in the decision making system, while the accuracy of the UKT determination system were 80%. The C5.0 algorithm produces a decision tree with 4 classification rules in the decision making system and 21 classification rules in the UKT determining system.

Keywords: Algorithm C5.0, decision making system, decision tree, decrease of ukt, determination of ukt.

ABSTRAK

COVID-19 tidak hanya merugikan dari sisi kesehatan saja, namun juga melemahkan berbagai sektor di Indonesia, khususnya sektor ekonomi. Kebijakan pemerintah yang membatasi aktivitas masyarakat berimbas pada melemahnya kegiatan ekonomi. Hal ini selaras dengan banyaknya pengajuan banding UKT (Uang Kuliah Tunggal) dari mahasiswa Institut Teknologi Kalimantan (ITK). Pihak birokrat telah memfasilitasi mahasiswa yang ingin mengajukan keringanan UKT. Namun, proses pengajuan tersebut masih dilakukan secara manual dan rentan terjadi *human error*. Dalam menganalisa pengajuan tersebut, diperlukan suatu metode yang dapat membantu dalam memberikan keputusan apakah pengajuan tersebut layak diterima atau tidak, salah satunya dengan metode Algoritma C5.0. Algoritma C5.0 merupakan algoritma *decision tree* yang dapat menganalisa data menjadi sekumpulan aturan yang nantinya dapat dijadikan masukan dalam pengambilan keputusan. Diharapkan dengan adanya sistem ini mampu membantu pihak ITK dalam mengambil keputusan dan menentukan besaran UKT secara lebih tepat dan objektif. Berdasarkan hasil evaluasi terhadap 108 data mahasiswa ITK yang mengajukan banding UKT diperoleh akurasi sebesar 91% pada sistem pengambilan keputusan, sedangkan pada sistem penentu UKT diperoleh akurasi sebesar 80%. Algoritma C5.0 menghasilkan pohon keputusan dengan 4 aturan klasifikasi pada sistem pengambilan keputusan dan 21 aturan klasifikasi pada sistem penentu UKT.

Kata kunci: Algoritma C5.0, keringanan ukt, penentuan ukt, pohon keputusan, sistem pengambilan keputusan.

Dikirim: 06 Desember 2021; Diterima: 14 Februari 2022; Dipublikasikan: 30 Maret 2022
Cara sitasi: Amalda, R. N., Millah, N., & Fitria, I. (2022). Implementasi Algoritma C5.0 dalam menganalisa kelayakan penerima keringanan ukt mahasiswa itk. *Teorema*: *Teori dan Riset Matematika,* *7*(1), 101-116. DOI: http://dx.doi.org/10.25157/teorema.v7i1.6692

PENDAHULUAN

Kehadiran COVID-19 tidak hanya merugikan dari sisi kesehatan saja, namun juga melemahkan berbagai sektor di Indonesia. Hampir seluruh sektor terdampak, salah satu yang mengalami dampak serius akibat pandemi ini adalah sektor ekonomi. Kebijakan pemerintah yang membatasi aktivitas masyarakat berimbas pada melemahnya kegiatan bisnis yang kemudian berpengaruh pada perekonomian Indonesia. Kinerja ekonomi yang melemah ini ditandai dengan meluasnya PHK, hingga meningkatnya pengangguran. Berdasarkan data Badan Pusat Statistik (BPS), meningkatnya pengangguran terjadi di semua kelompok usia, tingkat pengangguran tertinggi terjadi pada penduduk usia 20-24 tahun sebesar 17,66% pada Februari 2021, meningkat 3,36% dibandingkan periode yang sama tahun lalu sebesar 14,3%. Hal ini selaras dengan banyaknya pengajuan keringanan UKT (Uang Kuliah Tunggal) dari mahasiswa Institut Teknologi Kalimantan (ITK) yang merasa keberatan dengan besaran UKT yang diterima dikarenakan kondisi ekonomi keluarga yang sedang menurun (Jalil *et al*., 2020; Rizaty, 2021).

Institut Teknologi Kalimantan (ITK) sebagai perguruan tinggi negeri telah menyelenggarakan pelayanan secara adil terkait kesejahteraan mahasiswanya. Salah satu bentuk pelayanan tersebut adalah dengan memfasilitasi pengajuan keringanan UKT dari mahasiswa ITK. Uang Kuliah Tunggal (UKT) merupakan sistem pembayaran perkuliahan dimana biaya ini ditanggung oleh setiap mahasiswa sesuai kemampuan ekonomi dari masing-masing mahasiswa. Saat ini proses pendataan mahasiswa yang mengajukan keringanan UKT masih dilakukan secara manual. Proses ini tidak akan menjadi masalah besar apabila data yang diolah dalam jumlah yang sedikit, namun bila data yang diolah dalam jumlah yang banyak, maka akan diperlukan waktu dan tenaga yang cukup lama dalam melakukan seleksi berkas. Selain itu, sistem konvensional ini juga rentan akan terjadinya *human error* dari panitia penyeleksi. Karenanya diperlukan suatu metode yang dapat mengklasifikasi data, dengan memanfaatkan teknologi *data mining* (Larytasari & Susanti, 2019).

Data mining merupakan proses pembelajaran komputer (*machine learning*) dengan menggunakan satu atau lebih teknik pembelajaran dengan mengeksplorasi pola dari suatu data untuk didapatkan informasi yang berguna. Penelitian ini menerapkan bidang *data mining* khususnya teknik klasifikasi dengan menggunakan metode *decision tree*. *Decision tree* adalah teknik klasifikasi terhadap sekumpulan objek atau data dengan representasi pohon, salah satu algortima *decision tree* adalah algoritma C5.0. Algoritma C5.0 merupakan algoritma *decision tree* yang dapat menganalisa data menjadi sekumpulan aturan yang diharapkan nantinya dapat dijadikan masukan dalam pengambilan keputusan. Alasan penelitian ini menggunakan algoritma C5.0, karena algoritma ini memiliki beberapa kelebihan diantaranya dapat menangani *missing value* dan data dalam jumlah yang besar. Selain itu, algoritma ini juga dapat melakukan *training* data dalam waktu yang relatif cepat untuk digunakan dalam *testing* data. Penelitian ini menggunakan bahasa pemrograman R dikarenakan R sangat mendukung dalam pengolahan data serta menyediakan beragam *package*, salah satunya *package* untuk algoritma C5.0. (Benediktus & Oetama, 2020; Kastawan *et al*., 2018; Pardede *et al*., 2019; Riadi *et al*., 2020; W., 2007).

Beberapa penelitian mengenai algoritma C5.0 mengatakan bahwa algoritma ini lebih baik dalam melakukan klasifikasi, seperti pada penelitian Kastawan yang menunjukkan akurasi sebesar 96,08%. Sedangkan pada penelitian Larytasari menggunakan algoritma C4.5 diperoleh akurasi sebesar 77.95% dengan variabel yang digunakan terdiri dari UKT, pendapatan, voltase listrik, pajak, PBB, jumlah tanggungan dan kepemilikan rumah. Dilihat dari penelitian sebelumnya, variabel yang akan digunakan pada penelitian ini terdiri dari UKT, penghasilan, jumlah tanggungan dan selisih UKT dengan penghasilan. Algoritma C5.0 pernah diterapkan sebelumnya pada studi kasus dan variabel yang berbeda dan akan diaplikasikan pada data pengajuan keringanan UKT dari mahasiswa ITK. Diharapkan dengan adanya sistem ini dapat membantu pihak ITK dalam mengambil keputusan secara lebih tepat dan objektif. Adapun tujuan dari penelitian ini untuk mengetahui hasil penerapan algoritma C5.0 dalam menentukan kelayakan penerima keringanan UKT serta

menentukan besaran UKT terbaru dari pengajuan keringanan UKT (Kastawan *et al*., 2018; Larytasari & Susanti, 2019).

METODE PENELITIAN

Sub bab ini berisi kajian pustaka dari beberapa referensi sebagai penunjang pada penelitian ini.

1. Pohon Keputusan

Decision tree atau pohon keputusan merupakan salah satu metode klasifikasi yang bersifat prediktif. *Decision tree* dapat mengubah *database* menjadi sekumpulan aturan dengan representasi pohon keputusan. Aturan dapat dengan mudah dipahami, karenanya *decision tree* merupakan teknik klasifikasi yang mudah untuk dipelajari dan sangat populer digunakan. *Decision tree* merupakan teknik klasifikasi terhadap sekumpulan objek atau data dengan representasi pohon keputusan. Pohon keputusan umumnya digunakan untuk eksplorasi pola dan melihat korelasi antara sejumlah variabel *input* dengan variabel *output*. Adapun bagan *decision tree* terdiri dari tiga bagian yaitu: (Hadi, 2017; Itiqomah *et al*., 2019; Manurung, 2020; Marcania, 2019).

1. *Root Node* : Merupakan node akar yang terletak paling atas dari struktur pohon keputusan.
2. *Internal Node* : Merupakan *node* percabangan,
3. *Leaf Node* : Merupakan node keputusan, hanya terdapat satu *input* dan tidak memiliki *output* hanya memiliki satu *input* dan memiliki minimal dua *output*.

Adapun syarat yang perlu dipenuhi dalam penerapan algoritma *decision tree* adalah diperlukan *training dataset* yang menyediakan variabel target sebagai *supervisor* atau guru dalam pembelajaran mesin, *training dataset* harus banyak dan beragam, serta variabel target bertipe kategorik (Hadi, 2017; Hutabarat, 2018; W, 2007).

2. Algoritma C5.0

Algoritma C5.0 merupakan metode *data mining* dengan algoritma klasifikasi yang berbasis pada teknik *decision tree*. Algoritma ini disempurnakan dari algoritma ID3 dan C4.5 oleh Ross Quinlan pada tahun 1987. Algoritma ini dinilai lebih baik dibanding algoritma sebelumnya dalam hal akurasi dan memori. Kecepatan dalam membuat model dinilai sangat cepat dibanding algoritma lainnya. Algoritma ini juga dapat menangani atribut yang bernilai diskrit maupun kontinu. Kelebihan inilah yang membuat algoritma C5.0 dinilai unggul dibanding algoritma lainnya. Algoritma ini bermula dari semua atribut yang dijadikan akar dari pohon keputusan. Kemudian dipilih atribut yang memiliki nilai *gain* tertinggi untuk dijadikan *root node.* Selanjutnya atribut lainnya akan dievaluasi dengan cara yang serupa untuk mendapatkan akar node selanjutnya. Adapun persamaan untuk menghitung *information gain* dari keseluruhan kasus adalah sebagai berikut (Hutabarat, 2018; Itiqomah *et al*., 2019; Manik *et al*., 2018; Pardede *et al*., 2019).

$$I(S_1, S_2, \dots S_m) = -\sum_{i=1}^{m} P_i \log_2(P_i) \quad (1)$$

Dengan $I(S_1, S_2, \dots S_m)$ merupakan informasi dari keseluruhan kasus pada kelas i, m merupakan banyaknya kelas, P_i merupakan proporsi dari S_i terhadap S, S_i merupakan jumlah kasus pada kelas i, S merupakan jumlah kasus. Langkah selanjutnya yaitu menghitung *information gain* dari keseluruhan kasus pada kelas i dan kategori j sebagai berikut.

$$I(S_{1j}, S_{2j}, \dots, S_{mj}) = -\sum_{i,j=1}^{my} P_{ij} \log_2(P_{ij}) \quad (2)$$

$I(S_{1j}, S_{2j}, \dots, S_{mj})$ merupakan informasi dari keseluruhan kasus pada kelas i dan kategori j. P_{ij} merupakan proporsi kelas i pada kategori j dari S_{ij} terhadap S. S_{ij} menunjukkan jumlah kasus pada kelas i dan kategori j. y menunjukkan banyaknya kategori. $\log_2(x)$ umum digunakan dalam konteks teknik komputer, kebanyakan bahasa komputer mengandung logaritma natural berbasis dua. Sedangkan untuk menghitung nilai *entropy* ditunjukkan pada persamaan berikut.

$$E(A_j) = \sum_{i,j=1}^{my} \frac{S_{1j}+\cdots+S_{mj}}{S} \times I(S_{1j}, S_{2j}, \ldots S_{mj}) \quad (3)$$

$E(A_j)$ merupakan *entropy* atribut A pada kategori j, S_{ij} merupakan jumlah kasus dari kelas i dan kategori j dari atribut A, $\frac{S_{1j}+\cdots+S_{mj}}{S}$ merupakan informasi dari jumlah kasus kelas i dan kategori j terhadap S. Untuk mendapatkan nilai *gain* selanjutnya dihitung dengan formula sebagai berikut.

$$Gain(A) = I(S_1, S_2, \ldots S_m) - E(A) \quad (4)$$

Gain dapat diartikan sebagai ukuran efektifitas atau seberapa informatif suatu atribut dalam melakukan klasifikasi data. $E(A)$ merupakan total *entropy* dari keseluruhan kategori pada atribut A. Proses dilakukan hingga kategori sampel tidak dapat dilakukan *split* atau pemisahan.

Pseudo Code algoritma C5.0:

1. Periksa kasus dasar.
2. Untuk setiap atribut, hitung perolehan informasi yang dinormalisasi untuk pemisahan atribut.
3. Pilih atribut terbaik yang memiliki perolehan informasi tertinggi.
4. Temukan simpul keputusan terbaik, sebagai simpul akar.
5. Ulangi pada atribut yang diperoleh dengan memisahkan yang terbaik dan menambahkan simpul-simpul tersebut sebagai anak simpul (Benediktus & Oetama, 2020).

3. Confusion Matrix

Confusion matrix merupakan metode evaluasi yang menggunakan tabel matriks. Matriks ini terdiri dari beberapa sel, setiap sel berisi angka yang menyatakan jumlah data uji yang diklasifikasi dengan benar dan jumlah data uji yang salah diklasifikasi. Adapun tabel *confusion matrix* sebagai berikut.

Tabel 1. Confusion matrix

Actual	Prediction	
	Class +	Class -
Class +	TP (*True Positive*)	FN (*False Negative*)
Class -	FP (*False Positive*)	TN (*True Negative*)

Dimana TP merupakan jumlah data positif yang terklasifikasi dengan benar oleh sistem. TN merupakan jumlah data negatif yang terklasifikasi dengan benar oleh sistem. FN merupakan jumlah data positif yang terklasifikasi negatif oleh sistem. FP merupakan jumlah data negatif yang terklasifikasi positif oleh sistem. Berikut formula untuk mengukur akurasi, *error*, presisi, *recall* dan *specifity* berdasarkan *confusion matrix* (Kastawan *et al*., 2018).

$$Accuracy\ (\%) = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (5)$$

$$Error\ (\%) = \frac{FN+FP}{TP+TN+FP+FN} \times 100\% \quad (6)$$

$$Precision\ (\%) = \frac{TP}{TP+FP} \times 100\% \quad (7)$$

$$Recall\ (\%) = \frac{TP}{TP+FN} \times 100\% \quad (8)$$

$$Specifity\ (\%) = \frac{TN}{TN+FP} \times 100\% \quad (9)$$

Sub bab ini menjelaskan uraian penelitian secara garis besar.

1. Studi Literatur

Studi literatur adalah metode pengumpulan informasi dengan mencari referensi atau sumber yang berkaitan dengan permasalahan yang dibahas dalam penelitian ini, baik dari sumber elekronik yang didapat dari internet maupun dari buku teks. Referensi atau sumber yang diperoleh, kemudian dikaji untuk memperkuat landasan teori dalam penelitian yang dilakukan. Studi literatur dilakukan dengan mencari jurnal-jurnal, artikel, buku atau skripsi penelitian terkait dengan *data mining*, klasifikasi, *decision tree*, algoritma C5.0, dan *confusion matrix* (Makih, 2019).

2. Sumber Data

Data yang digunakan pada penelitian ini adalah data mahasiswa ITK yang mengajukan keringanan UKT periode tahun 2020.

3. *Preprocessing Data*

Preprocessing Data merupakan langkah awal sebelum dilakukan pengolahan data. Tahap ini dilakukan pemilihan variabel, karena tidak semua variabel pada data dibutuhkan dalam mencapai tujuan penelitian, sehingga perlu diambil beberapa variabel yang akan digunakan sebagai variabel *input* dan variabel target. Selain itu, dilakukan *data cleaning* dengan membuang data yang mengandung duplikasi dan *missing value.* Proses ini dilakukan secara manual menggunakan *excel* untuk memastikan bahwa data yang dipilih layak untuk diproses pada model (Marcania, 2019).

4. Pembagian Data

Tahap ini *dataset* yang telah melewati *preprocessing data* dibagi menjadi dua bagian, yaitu sebagian dialokasikan pada *data training* dan sebagian pada *data testing*. *Data training* digunakan untuk membentuk model *decision tree*, sedangkan *data testing* akan digunakan untuk klasifikasi berdasarkan model yang telah terbentuk. Pembagiannya dilakukan secara acak dengan teknik *random*. Pada sistem pengambilan keputusan proporsi yang dipakai 80% *data training* dan 20% *data testing*. Sedangkan pada sistem penentu UKT proporsi yang digunakan 95% *data training* dan 5% *data testing* (Marcania, 2019).

5. Transformasi Data

Transformasi data merupakan proses mengubah data ke dalam bentuk tertentu sebelum diproses ke *data mining*. Transformasi dilakukan dengan mengubah tipe data numerik menjadi data kategorik (Putri, 2016).

6. Implementasi Algoritma C5.0

Tahap ini dibentuk model *decision tree* untuk mengetahui pola penting dalam basis data dengan mengimplementasi Algoritma C5.0. Data *training* yang telah dibentuk sebelumnya, digunakan untuk membentuk model pohon keputusan. Model ini yang nantinya akan digunakan untuk memprediksi data *testing*. Adapun proses pembentukan pohon keputusan dijelaskan sebagai berikut (Rahmayanti *et al*., 2020).

1. Menghitung *Information Gain* dari keseluruhan kasus.
2. Menghitung *Information Gain* dan *Entropy* dari setiap kemungkinan posisi pemecahan atribut.
3. Menghitung *Entropy* total dan nilai *Gain* dari setiap kemungkinan posisi pemecahan atribut. Gain tertinggi akan dipilih menjadi *root node*.
4. Membuat cabang dari setiap kategori *root node*. Jika *Entropy* atribut bernilai nol maka percabangan berhenti dan menjadi *leaf node. Leaf node* diputuskan berdasar pada mayoritas kelas.

5. Jika hasil percabangan menjadi *internal node*, artinya perlu dilakukan *step* yang sama seperti cara sebelumnya.
6. Proses dilakukan hingga kategori atribut tidak dapat dilakukan *split* (pemisahan) lagi.

Berikut *code* algoritma C5.0 menggunakan software R.

```
library("C50")

#Menghasilkan dan menampilkan summary model
Status_model <- C5.0(input_training_set, class_training_set, control =
C5.0Control(label="Status"))
summary(Status_model)

#Plot decission tree
plot(Status_model)

#Menggunakan model untuk prediksi testing set
predict(Status_model, input_testing_set)

#Menyimpan hasil prediksi testing set ke dalam kolom hasil_prediksi
input_testing_set$Status <- data[-indeks_training_set,]$Status
input_testing_set$hasil_prediksi <- predict(Status_model, input_testing_set)
print(input_testing_set)
```

Gambar 1. *Code R*

7. Pengujian Model Klasifikasi

Setelah didapatkan model *decision tree*, tahap terakhir adalah mengimplementasi *data testing* pada aturan model yang sudah terbentuk, yang kemudian hasilnya disimpan pada kolom hasil prediksi. Hasil prediksi ini dibandingkan dengan data sebenarnya untuk melihat keberhasilan model dalam melakukan prediksi. Untuk menguji keberhasilan model dapat menggunakan metode *confusion matrix* yang terdiri dari metrik *accuracy*, *error*, *precision*, *recall* dan *specifity*. Akurasi untuk mengukur seberapa baik model dalam melakukan prediksi dan *error* untuk mengukur tingkat kesalahan model dalam melakukan prediksi. Presisi untuk melihat tingkat ketepatan antara informasi pada data sebenarnya dengan jawaban yang diberikan oleh sistem. *Recall* digunakan untuk melihat seberapa sensitif model dalam mendeteksi data berlabel positif. S*pecifity* digunakan untuk melihat seberapa sensitif model dalam mendeteksi data berlabel negatif (Nugroho, 2019).

HASIL DAN PEMBAHASAN

1. Sistem Pengambilan Keputusan

a. *Preprocessing Data*

Data yang telah tersedia, selanjutnya dilakukan *data selection* dengan memilih variabel *input* dan variabel target yang akan dianalisis. Variabel *input* yang digunakan pada sistem ini terdiri dari variabel selisih dan variabel tanggungan, sedangkan variabel target yang digunakan yaitu variabel status. Setelah ditentukan variabel, langkah selanjutnya adalah proses *cleaning data,* yaitu menghilangkan duplikasi dan *missing value* pada data.

Proses ini diperoleh sebanyak 108 data yang layak digunakan dari 258 data keseluruhan. Berikut potongan data beserta variabel yang digunakan pada penelitian ini.

Tabel 2. Preprocessing data

D	UKT	Penghasilan	Selisih	Tanggungan	Status
1	2000000	2209500	209500	3	Disetujui
2	9000000	2900000	-6100000	3	Disetujui
3	6000000	4971200	-1028800	4	Disetujui
4	9500000	9673000	173000	2	Disetujui
5	6000000	2900000	-3100000	2	Disetujui
6	4000000	2518966	-1481034	3	Disetujui
7	4000000	2500000	-1500000	1	Disetujui
8	9000000	8500000	-500000	3	Disetujui
9	9000000	3189315	-5810685	1	Disetujui
10	6000000	3500000	-2500000	1	Disetujui
11	6000000	6000000	0	5	Disetujui
12	6000000	4000000	-2000000	2	Disetujui
13	12000000	12562000	562000	4	Disetujui
14	10000000	8000000	-2000000	2	Disetujui
15	1000000	1000000	0	1	Disetujui
16	10000000	8000000	-2000000	3	Disetujui
17	8000000	12000000	4000000	2	Tidak Disetujui
18	4000000	1500000	-2500000	2	Disetujui
19	9000000	8000000	-1000000	3	Disetujui
20	2000000	1500000	-500000	2	Disetujui

Variabel selisih merupakan nilai pengurangan dari penghasilan terbaru dengan nilai UKT sebelum diajukan keringanan, variabel tanggungan menunjukkan jumlah tanggungan keluarga, dan variabel status menunjukkan keputusan disetujui atau tidak pengajuan tersebut.

b. *Pembagian* Data

Tahap ini data yang telah melalui proses *preprocessing data* akan dibagi ke dalam dua bagian. Data akan dibagi secara acak dengan proporsi 80% *data training* dan 20% *data testing*. Sehingga diperoleh *data training* sebanyak 86 data dan *data testing* sebanyak 22 data.

c. *Implementasi* Algoritma C5.0

Tahap selanjutnya yaitu proses pembentukan pohon keputusan dengan algoritma C5.0. Pembentukan ini diawali dengan menentukan node awal sebagai akar dari pohon keputusan. Algoritma C5.0 akan memeriksa semua kemungkinan pemecahan nilai dari masing-masing atribut dan memilih posisi pemecahan terbaik yaitu yang memiliki kriteria *Gain* paling besar. *Root node* dipilih berdasarkan atribut yang memiliki nilai *Gain* terbesar. Berdasarkan hasil perhitungan *Gain* diperoleh pemecahan terbaik dari masing-masing atribut yang telah ditransformasi nilainya ke dalam skala ordinal adalah sebagai berikut.

Tabel 3. Transformasi data

Atribut	Kategori	Keterangan
Selisih	$> 1.400.000$	Banyak
	$\leq 1.400.000$	Sedikit
Tanggungan	> 2	Banyak
	≤ 2	Sedikit

Berdasarkan hasil pada tabel 3, diperoleh masing-masing atribut dibagi ke dalam dua kategori. Berikut tabel pembagian kasus dalam menentukan *root node* beserta perhitungan manual untuk *mendapatkan* nilai *Gain* dari masing-masing atribut.

Tabel 4. Pembagian kasus

Atribut	Nilai	Jumlah	Disetujui	Tidak Disetujui
		86	**82**	**4**
Selisih	Banyak	12	8	4
	Sedikit	74	74	0
Tanggungan	Banyak	28	28	0
	Sedikit	58	54	4

Tabel 4 berisikan informasi jumlah data di masing-masing nilai atribut. Berdasarkan data ini, kemudian dilakukan perhitungan sebagai berikut.

1) Perhitungan nilai *Information Gain* dari total keseluruhan atribut:

$$
\begin{aligned}
I(82{,}4) &= -\sum_{i=1}^{2} P_i \ \log_2(P_i) \\
&= \left(-\frac{82}{86}\log_2\left(\frac{82}{86}\right)\right) + \left(-\frac{4}{86}\log_2\left(\frac{4}{86}\right)\right) \\
&= \left(-\frac{82}{86} \times \frac{\ln\left(\frac{82}{86}\right)}{\ln 2}\right) + \left(-\frac{4}{86} \times \frac{\ln\left(\frac{4}{86}\right)}{\ln 2}\right) \\
&= 0.066 + 0.206 \\
&= 0.272
\end{aligned}
$$

2) Perhitungan *Information Gain* dan *Entropy* pada atribut Selisih dengan kategori banyak:

$$
\begin{aligned}
I(8,4) &= \left(-\frac{8}{12}\log_2\left(\frac{8}{12}\right)\right) + \left(-\frac{4}{12}\log_2\left(\frac{4}{12}\right)\right) \\
&= \left(-\frac{8}{12} \times \frac{\ln\left(\frac{8}{12}\right)}{\ln 2}\right) + \left(-\frac{4}{12} \times \frac{\ln\left(\frac{4}{12}\right)}{\ln 2}\right) \\
&= 0.390 + 0.528 \\
&= 0.918
\end{aligned}
$$

$$
E(A_j) = \sum_{i,j=1}^{my} \frac{S_{1j} + \cdots + S_{mj}}{S} \times I(S_{1j}, S_{2j}, \ldots S_{mj})
$$

$$
E(A_j) = \frac{12}{86} \times 0.918 = 0.128
$$

3) Perhitungan Information *Gain* dan *Entropy* pada atribut Selisih dengan kategori sedikit :

$$I(74,0) = \left(-\frac{74}{74}\log_2\left(\frac{74}{74}\right)\right) + 0$$
$$= \left(-\frac{74}{74} \times \frac{\ln\left(\frac{74}{74}\right)}{\ln 2}\right)$$
$$= 0$$

$$E(A_j) = \frac{74}{86} \times 0 = 0$$

4) Perhitungan *Entropy* total dan nilai *Gain* pada atribut Selisih :

$$E(A) = 0.128 + 0 = 0.128$$

$$Gain(A) = I(S_1, S_2) - E(A)$$
$$= 0.272 - 0.128$$
$$= 0.144$$

Dengan melakukan perhitungan yang sama pada atribut tanggungan, sehingga hasil perhitungan dapat dirangkum pada tabel perhitungan berikut.

Tabel 5. Perhitungan node akar

Atribut	Kategori	Jumlah	Disetujui	Tidak Disetujui	Information Gain	Entropy	Total Entropy	Gain
		86	**82**	**4**	**0,272**			
Selisih	Banyak	12	8	4	0,918	0,128	0,128	0,144
	Sedikit	74	74	0	0	0		
Tanggungan	Banyak	28	28	0	0	0	0,244	0,028
	Sedikit	58	54	4	0,362	0,244		

Terlihat bahwa nilai *Gain* paling tinggi terdapat pada atribut Selisih, maka atribut Selisih dipilih menjadi *root node* pada model *decision tree*. *Entropy* atribut yang bernilai nol ada pada kategori sedikit, sehingga kategori ini menuju *leaf node* dengan status disetujui karena nilai mayoritas ada pada kelas disetujui. Setelah diperoleh node akar, dengan cara yang sama, dilakukan evaluasi pada atribut lainnya untuk memperoleh node *parent* selanjutnya. Sehingga hasil perhitungan dapat dirangkum pada tabel perhitungan berikut.

Tabel 6. Perhitungan node internal

Atribut	Kategori	Jumlah	Disetujui	Tidak Disetujui	Information Gain	Entropy	Total Entropy	Gain
		12	8	4	0,918			
Tanggungan	Banyak	5	5	0	0	0	0,575	0,343
	Sedikit	7	3	4	0,985	0,575		
Selisih	Banyak	4	4	0	0	0	0,667	0,251
	Sedikit	8	4	4	1	0,667		

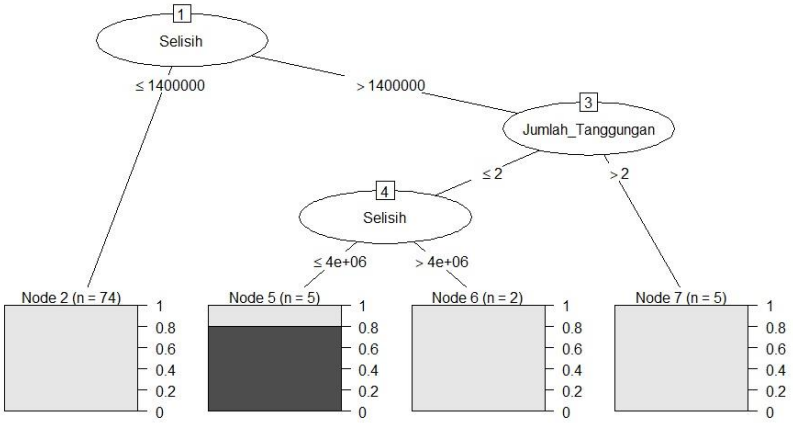
Berdasarkan tabel 6 diketahui nilai *Gain* tertinggi ada pada atribut tanggungan. Sehingga, atribut tanggungan dipilih sebagai node *parent* selanjutnya. Selain itu terdapat *Entropy* atribut yang bernilai nol yaitu pada kategori banyak, yang artinya kategori tersebut

menuju *leaf node* dengan status disetujui karena nilai mayoritas ada pada kelas disetujui. Dengan cara yang sama, berikut hasil perhitungan pada atribut selisih.

Tabel 7. Perhitungan node internal

Atribut	Kategori	Jumlah	Disetujui	Tidak Disetujui	Information Gain	Entropy
		7	**3**	**4**	**0,985**	
Selisih	Banyak	2	2	0	0	0
	Sedikit	5	1	4	0,722	0,516

Berdasarkan tabel 7 diketahui nilai *Gain* tertinggi ada pada kategori sedikit. Selain itu terdapat *Entropy* atribut yang bernilai nol yaitu pada kategori banyak, yang artinya kategori tersebut menuju *leaf node* dengan status disetujui karena nilai mayoritas ada pada kelas disetujui. Sedangkan pada kategori sedikit mayoritas pada kelas tidak disetujui, namun ada satu yang disetujui. Berikut adalah *plot decision tree* berdasarkan perhitungan algoritma C5.0 menggunakan *software* R.



Gambar 2. Model pohon keputusan

d. Pengujian Model Klasifikasi

Dengan menerapkan *data* testing pada *rules* model yang sudah terbentuk, diperoleh hasil prediksi yang dinyatakan dalam *confusion matrix* sebagai berikut.

Tabel 8. Confusion matrix

Actual	Prediction	
	Disetujui	**Tidak Disetujui**
Disetujui	19	1
Tidak Disetujui	1	1

Berdasarkan hasil pada tabel 8, diketahui yang diprediksi dengan benar ada sebanyak 20 data dan yang salah diprediksi ada sebanyak 2 data. Berdasarkan hasil ini, diperoleh hasil pengujian dengan metrik pengukur sebagai berikut.

1) Akurasi

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$
$$= \frac{20}{22} \times 100\% = 90.9 \approx 91\%$$

2) *Error*

$$Error = \frac{FN + FP}{TP + TN + FP + FN} \times 100\%$$
$$= \frac{2}{22} \times 100\% = 9.09 \approx 9\%$$

3) Presisi

$$Precision = \frac{TP}{TP + FP} \times 100\%$$
$$= \frac{19}{20} \times 100\% = 95\%$$

4) *Recall*

$$Recall = \frac{TP}{TP + FN} \times 100\%$$
$$= \frac{19}{20} \times 100\% = 95\%$$

5) *Specifity*

$$Specifity = \frac{TN}{TN + FP} \times 100\%$$
$$= \frac{1}{2} \times 100\% = 50\%$$

Berdasarkan tabel 8, diperoleh akurasi, presisi dan *recall* yang sangat baik yaitu diatas 90%, namun diperoleh *specifity* hanya sebesar 50%. Hal ini *menunjukan* bahwa model lebih condong memprediksi disetujui dibanding tidak disetujui. Ini disebakan karena data latih yang tidak seimbang antara disetujui dengan tidak disetujui.

2. Sistem Penentu UKT

a. *Preprocessing Data*

Tahap ini dilakukan *data selection* dengan memilih variabel *input* dan variabel target yang akan dianalisis. Variabel *input* yang digunakan pada sistem ini terdiri dari variabel UKT lama, penghasilan terbaru dan jumlah tanggungan, sedangkan variabel target yang digunakan yaitu variabel UKT terbaru. Setelah ditentukan variabel, langkah selanjutnya adalah proses *cleaning data,* yaitu menghilangkan duplikasi dan *missing value* pada data. Proses ini diperoleh sebanyak 108 data yang layak digunakan dari 258 data keseluruhan.

Variabel UKT lama menunjukkan nilai UKT sebelum diajukan keringanan, variabel penghasilan menunjukkan nilai penghasilan terbaru dari orang tua/wali, variabel tanggungan menunjukkan jumlah tanggungan keluarga, dan variabel UKT terbaru menunjukkan nilai UKT terbaru baik yang disetujui maupun tidak disetujui pengajuannya.

b. Pembagian Data

Tahap ini data yang telah melalui proses *preprocessing data* akan dibagi ke dalam dua bagian. Data akan dibagi secara acak dengan proporsi 95% *data training* dan 5% *data testing*. Sehingga diperoleh *data training* sebanyak 103 data dan *data testing* sebanyak 5 data.

c. Implementasi Algoritma C5.0

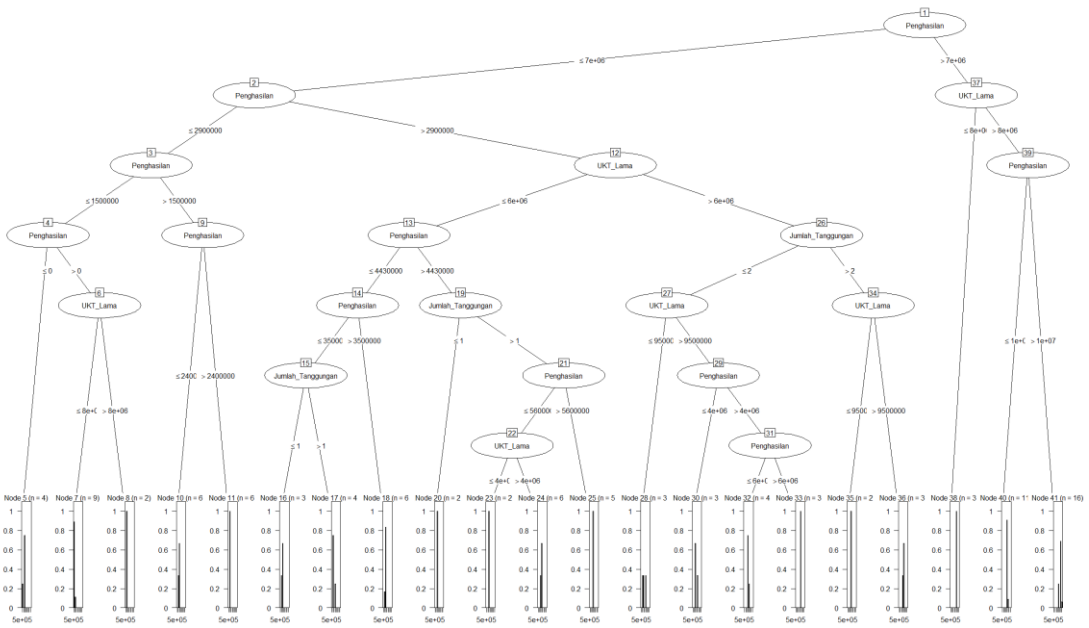
Pada tahap ini akan dibentuk model pohon keputusan (*decision tree*) dengan mengimplementasi algoritma C5.0. Algoritma C5.0 akan memeriksa semua kemungkinan

pemecahan nilai dari masing-masing atribut dan memilih posisi pemecahan terbaik yaitu yang memiliki kriteria *Gain* paling besar. Berikut atribut-atribut hasil pencarian semua kemungkinan pemecahan terbaik yang telah ditransformasi nilainya ke dalam skala ordinal.

Tabel 9. Transformasi data

Atribut	Kategori	Keterangan
Penghasilan	> 7000000	Banyak
	≤ 7000000	Sedikit
UKT Lama	> 8000000	Banyak
	≤ 8000000	Sedikit
Penghasilan	> 10000000	Banyak
	≤ 10000000	Sedikit
Penghasilan	> 2900000	Banyak
	≤ 2900000	Sedikit
Penghasilan	> 1500000	Banyak
	≤ 1500000	Sedikit
Penghasilan	> 2400000	Banyak
	≤ 2400000	Sedikit
Penghasilan	> 0	Banyak
	≤ 0	Sedikit
UKT Lama	> 6000000	Banyak
	≤ 6000000	Sedikit
Penghasilan	> 4430000	Banyak
	≤ 4430000	Sedikit
Penghasilan	> 3500000	Banyak
	≤ 3500000	Sedikit
Jumlah Tanggungan	> 1	Banyak
	≤ 1	Sedikit
Penghasilan	> 5600000	Banyak
	≤ 5600000	Sedikit
UKT Lama	> 4000000	Banyak
	≤ 4000000	Sedikit
Jumlah Tanggungan	> 2	Banyak
	≤ 2	Sedikit
UKT Lama	> 9500000	Banyak
	≤ 9500000	Sedikit
Penghasilan	> 4000000	Banyak
	≤ 4000000	Sedikit
Penghasilan	> 6000000	Banyak
	≤ 6000000	Sedikit

Berdasarkan hasil pada tabel 9, diketahui masing-masing atribut dibagi ke dalam dua kategori. Berikut adalah *plot decision tree* berdasarkan perhitungan algoritma C5.0 dan simulasi menggunakan *software* R.



Gambar 3. Model pohon keputusan

d. Pengujian Model Klasifikasi

Dengan menerapkan *data testing* pada *rules model* yang sudah terbentuk, diperoleh hasil prediksi yang dinyatakan dalam *confusion matrix* sebagai berikut.

Tabel 10. *Confusion matrix*

Actual	Classification			
	500,000	**2,000,000**	**4,000,000**	**6,000,000**
500,000	1	1		
2,000,000		1		
4,000,000			1	
6,000,000				1

Berdasarkan hasil pada table 10, diketahui yang diprediksi dengan benar ada sebanyak 5 data dan yang salah diprediksi ada sebanyak 1 data. Berdasarkan hasil ini, diperoleh hasil pengujian dengan metrik pengukur sebagai berikut.

1) Akurasi

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$
$$= \frac{4}{5} \times 100\% = 80\%$$

2) *Error*

$$Error = \frac{FN + FP}{TP + TN + FP + FN} \times 100\%$$
$$= \frac{1}{5} \times 100\% = 20\%$$

Tabel 11. Hasil perhitungan metrik

Metrik	UKT			
	500.000	2.000.000	4.000.000	6.000.000
Presisi	100%	50%	100%	100%
Recall	50%	100%	100%	100%
Specifity	50%	100%	100%	100%

Berdasarkan tabel 11 diperoleh akurasi model yang cukup bagus yaitu sebesar 80% serta presisi, *recall* dan *specifity* yang sangat bagus untuk UKT 4.000.000 dan 6.000.000. Namun model tidak cukup baik dalam memprediksi UKT 500.000 dan 2.000.000 dilihat dari presisi, *recall* dan *specifity* yang hanya sebesar 50%. Jika dibandingkan dengan penelitian sebelumnya yang menggunakan algoritma C4.5 dalam menentukan besaran UKT mahasiswa, algoritma C5.0 memperoleh akurasi yang lebih unggul dari algoritma C4.5.

KESIMPULAN

Berdasarkan hasil simulasi, pada sistem pengambilan keputusan diperoleh pohon keputusan dengan 4 aturan klasifikasi. Sedangkan pada sistem penentu UKT diperoleh pohon keputusan dengan 21 aturan klasifikasi. Pada sistem pengambilan keputusan diperoleh akurasi sebesar 91%. Selain itu, ketepatan model dalam melakukan prediksi mencapai 95%, sama halnya dengan keberhasilan model dalam mendeteksi data yang berlabel disetujui yaitu sebesar 95%, namun sebaliknya, sensitivitas model kurang dalam mendeteksi data yang berlabel tidak disetujui yaitu sebesar 50%. Pada sistem penentu UKT diperoleh akurasi sebesar 80%. Selain itu, presisi atau ketepatan model dalam melakukan prediksi pada kategori 500.000, 4.000.000 dan 6.000.000 mencapai 100%, sedangkan presisi pada kategori 2.000.000 hanya sebesar 50%. Pada metrik *recall* dan *specifity* diperoleh hasil 100% pada kategori UKT 2.000.000, 4.000.000 dan 6.000.000, sedangkan pada kategori 500.000 diperoleh nilai *recall* dan *specifity* sebesar 50%.

REKOMENDASI

Memaksimalkan atau menambah kriteria lainnya yang lebih relevan agar tercapai model keputusan yang lebih rasional, mengatasi *imbalance* pada label target, serta diharapkan pada pengembangan selanjutnya dapat dibuat versi aplikasi agar lebih mudah dalam penggunaan dan aksesnya.

UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih kepada Irma Fitria, S.Si., M.Si. dan Nashrul Millah, S.Si., M.Si. selaku Dosen Pembimbing Utama dan Dosen Pembimbing Pendamping yang telah membimbing dan memotivasi penulis hingga akhir. Penulis juga mengucapkan terima kasih kepada Primadina Hasanah, S.Si., M.Sc., Abrari Noor Hasmi, S.Si., M.Si., dan Indira Anggriani, S.Si., M.Si. selaku dosen penguji yang telah memberi banyak saran dan masukan untuk perbaikan dari penelitian ini.

DAFTAR PUSTAKA

Benediktus, N., & Oetama, R. S. (2020). Algoritma klasifikasi decision tree c5.0 untuk memprediksi performa akademik siswa. *14 ULTIMATICS*, *12*(1). https://www.kaggle.com/aljarah/xAPI-Edu-Data

Hadi, F. (2017). Penerapan data mining dalam menganalisa pemberian pinjaman dengan menggunakan metode algoritma c5.0 (studi kasus : koperasi jasa keuangan syariah kelurahan lambung bukik). *Jurnal Komputer Teknologi Informasi*, *4*(2), 214–223.

Hutabarat, C. (2018). Penerapan data mining untuk memprediksi permintaan produk kartu perdana internet menggunakan algoritma c5.0 (studi kasus: vidha ponsel). *Jurnal Pelita Informatika*, *6*(4), 419–424.

Itiqomah, F., Susanti, Y., & Zukhronah, E. (2019). Klasifikasi status kredit nasabah bmt menggunakan algoritma c5.0. *Seminar & Conference Procedings*, 73–78. http://jurnal.umt.ac.id/index.php/cpu/article/view/1684

Jalil, A., Kasnelly, S., & Studi Ekonomi Syariah Sekolah Tinggi Agama Islam An-Nadwah Kuala Tungkal, P. (2020). Meningkatnya Angka Pengangguran Ditengah Pandemi (COVID-19). *Al-Mizan : Jurnal Ekonomi Syariah*, *3*(1), 45–60. http://www.ejournal.an-nadwah.ac.id/index.php/almizan/article/view/142

Kastawan, P. W., Wiharta, D. M., & Sudarma, M. (2018). Implementasi algoritma c5.0 pada penilaian kinerja pegawai negeri sipil. *Majalah Ilmiah Teknologi Elektro*, *17*(3), 371. https://doi.org/10.24843/mite.2018.v17i03.p11

Larytasari, L. A., Susanti, Y., & Respatiwulan. (2019). Penentuan ukt mahasiswa uns dengan algoritma iterative dichotomiser three dan classification version 4.5. *Seminar & Conference Proceedings,* 94-100. https://jurnal.umt.ac.id/index.php/cpu/article/view/1687

Makih, A. (2019). *Algoritma predikat pegawai dengan tiga variabel menggunakan fuzzy inference system tsukamoto*. *100*. https://doi.org/10.31227/osf.io/gkje5

Manik, R., Pristiwanto., & Tampubolon, K. (2018). Prediksi kolektibilitas kredit anggota dengan algoritma c5.0 (studi kasus : cu damai sejahtera medan). *Jurnal Riset Komputer (JURIKOM)*, *5*(2), 151–160.

Manurung, M. A. (2020). Implementasi data mining algoritma c5.0 dalam sertifikasi produk pengguna tanda sni pada air minum dalam kemasan (studi kasus : balai riset dan standardisasi industri medan). *Journal of Computer System and Informatics (JoSYC)*, *1*(3), 199–206.

Marcania, M. (2019). Prediksi pengangkatan karyawan dengan metode klasifikasi algoritma c5.0 (studi kasus pt. kiyokuni indonesia factory-2). *Thesist*, *0*, 71.

Nugroho, K. S. (2019). Confusion matrix untuk evaluasi model pada supervised learning. Diakses dari: https://ksnugroho.medium.com/confusion-matrix-untuk-evaluasi-model-pada-unsupervised-machine-learning-bc4b1ae9ae3f

Pardede, M., Buulolo, E., & Ndruru, E. (2019). Implementasi algoritma c5.0 pada kelulusan peserta ujian kemahiran berbahasa indonesia (ukbi) pada balai bahasa sumatera utara. *KOMIK (Konferensi* Nasional *Teknologi Informasi Dan Komputer)*, *3*(1), 64–72. https://doi.org/10.30865/komik.v3i1.1569

Putri, Y. R., Mukhlash, I., & Hidayat, N. (2016). Prediksi pola kecelakaan kerja pada perusahaan non ekstraktif menggunakan algoritma decision tree: c4.5 dan c5.0. *Jurnal Sains Dan Seni Pomits, 2*(1), 1-6.

Rahmayanti, V., Azhar, Y., & Pramudita, A. E. (2020). Penerapan algoritma c5.0 pada analisis faktor-faktor pengaruh kelulusan tepat waktu mahasiswa teknik informatika UMM. *Jurnal Repositor*, *1*(2), 131. https://doi.org/10.22219/repositor.v1i2.545

Riadi, M., Azhar, Y., & Wicaksono, G. W. (2020). Implementasi algoritma c5.0 dan k-medoids untuk klasterisasi ibu hamil beresiko tinggi. *Jurnal Repositor*, *2*(4), 511. https://doi.org/10.22219/repositor.v2i4.696

Rizaty, M. A. (2021). Bps: tingkat pengangguran anak muda semakin tinggi saat pandemi. Diakses dari: https://databoks.katadata.co.id/datapublish/2021/08/31/bps-tingkat-pengangguran-anak-muda-semakin-tinggi-saat-pandemi

W., Y. Y. (2007). Perbandingan performansi algoritma decision tree c5.0, cart, dan chaid: kasus prediksi status resiko kredit di bank x. *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, *0*(0), 1907–5022. https://journal.uii.ac.id/Snati/article/view/1628