

Model Prediksi Kelulusan Tepat Waktu Mahasiswa Berdasarkan Algoritma Decision Tree C4.5

Bustomi Alwi
Universitas Buana Perjuangan
Karawang, Indonesia
ifl17.bustomialwi@mhs.ubpkarawang.ac.id

Sutan Faisal
Universitas Buana Perjuangan
Karawang, Indonesia
sutan.faisal@ubpkarawang.ac.id

Hilda Yulia Novita
Universitas Buana Perjuangan
Karawang, Indonesia
hilda.yulia@ubpkarawang.ac.id

Abstract— Kelulusan mahasiswa tepat waktu merupakan suatu cerminan keberhasilan perguruan tinggi dalam menjalankan sistem pendidikan. Penelitian ini bertujuan untuk mengetahui prediksi kelulusan mahasiswa. Penelitian ini menggunakan 4 atribut dan 1 label. Penelitian ini dilakukan untuk menganalisis atribut dalam memprediksi lulus tepat waktu menggunakan algoritma Decision Tree C4.5. Data yang digunakan dalam penelitian ini berasal dari web Kaggle sebanyak 524 data dan memiliki 14 atribut serta 1 label. Dalam penelitian ini, dataset dibagi menjadi dua, yaitu data training yang berjumlah 60% dan data testing yang berjumlah 40%. Pengujian ini dilakukan secara manual menggunakan Excel dengan data training. Pengujian menggunakan Python dilakukan dengan data testing, dan pengujian menggunakan Weka juga menggunakan data testing. Hasil yang diperoleh dari Weka menunjukkan nilai accuracy 84%, precision 88%, dan recall 84%. Sedangkan, accuracy Python menghasilkan 93%. Untuk hasil perhitungan manual, accuracy mendapatkan hasil 84%, precision 75%, dan recall 100%.

Kata kunci — Prediksi kelulusan mahasiswa, perguruan tinggi, Decision Tree C4.5

I. PENDAHULUAN

Umumnya mahasiswa menginginkan lulus tepat pada waktunya. Namun kenyataannya, banyak mahasiswa yang tidak bisa lulus tepat waktu. Bukan hanya mahasiswa yang menginginkan lulus tepat pada waktunya, tetapi pihak perguruan tinggi juga menginginkan mahasiswa untuk lulus tepat waktu. Berdasarkan Badan Akreditasi Nasional, kualitas perguruan tinggi yang berkualitas diukur berdasarkan standar, salah satunya yaitu persentase kelulusan tepat waktu untuk setiap program (BAN-PT 2019).

Faktor yang menyebabkan mahasiswa lulus tidak tepat waktu adalah mahasiswa yang gagal dalam belajar, apalagi dengan adanya pandemi COVID-19. Menurut (Ulfa & Mikdar, 2020), pandemi COVID-19 berdampak kepada mahasiswa untuk perkuliahan, yaitu menyesuaikan kuliah daring, daerah terpencil yang terhambat internet, dan lainnya yang mempengaruhi perilaku belajar, sosial, dan kesehatan yang dapat memberikan dampak negatif, yaitu mahasiswa kesulitan belajar online.

Penelitian yang telah membahas perbandingan algoritma Decision Tree oleh Anam & Santoso (2018), dalam penelitiannya mengatakan bahwa algoritma C4.5 dan Naïve Bayes dalam model mengklasifikasi penerimaan beasiswa menunjukkan algoritma C4.5 mendapatkan hasil yang lebih baik dibandingkan Naïve Bayes. Berdasarkan perbandingan kinerja dari akurasi, algoritma C4.5 mendapatkan 96,4% dengan proses waktu 0 detik, namun hasil algoritma Naïve Bayes yaitu 95,11% dengan proses waktu 0 detik.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk memprediksi mahasiswa lulus tepat waktu menggunakan algoritma Decision Tree C4.5.

II. DATA DAN METODE

A. Kelulusan Mahasiswa

Kelulusan mahasiswa sangat berpengaruh kepada pihak perguruan tinggi yang disebutkan dalam BAN-PT Nomor 3 (Tahun 2019) dalam instrumen tentang akreditasi perguruan tinggi. Dalam Luaran dan Capaian Tridharma, fokus penilaian kriteria yang menyanggung tentang kelulusan. Produktivitas dalam rancangan pendidikan dapat dinilai dari efisiensi pendidikan dan waktu pendidikan pada mahasiswa. Penelusuran hasil lulusan, umpan balik dari perguruan tinggi, dan persepsi publik terhadap kualitas ukuran lulusan telah sesuai dalam capaian pembelajaran lulusan.

B. Prediksi

Prediksi merupakan perkiraan yang dapat membantu suatu permasalahan sejak dini yang diharapkan bisa membantu dalam mengambil perbaikan atau keputusan sehingga dapat membantu mengoptimalkan dan memantau suatu permasalahan dengan mengolah data [1].

C. Algoritma Decision Tree C4.5

Algoritma C4.5 adalah salah satu pengembangan dari algoritma ID3. Algoritma C4.5 adalah teknik klasifikasi yang sering digunakan oleh para peneliti. Dari hasil perhitungan algoritma C4.5, akan terbentuk pohon keputusan atau decision tree [2].

Untuk menghitung algoritma decision tree C4.5, terbagi menjadi dua langkah. Yang pertama adalah menghitung nilai entropy. Setelah menghitung nilai entropy, dilanjutkan dengan menghitung nilai gain, yang akan menjadi akar. Rumus penghitungan decision tree yaitu:

Menghitung entropy

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

Keterangan:

S: Himpunan kasus.

N: Jumlah partisi dalam atribut.

Pi: Proporsi dari Si, pada S

Menghitung gain

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

Keterangan:

S: Himpunan kasus

A: Atribut

N: Jumlah partisi dalam atribut

|Si|: Jumlah kasus pada partisi ke-i

|S|: Jumlah kasus S

D. Confusion Matrix

Confusion matrix adalah metode untuk melakukan penghitungan tingkat akurasi. Menurut [3] confusion matrix merupakan tabel yang menyatakan prediksi dari data jumlah uji yang benar dan salah. Rumus confusion matrix dalam menghitung akurasi, presisi, dan recall adalah:

$$akurasi = \frac{TP + TN}{total}$$

$$presisi = \frac{TP + TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

Keterangan:

TP (True Positive) = Jumlah dokumen dari kelas satu yang benar yang mengklasifikasikan menjadi kelas satu.

TN (True Negative) = Jumlah dokumen dari kelas nol yang benar mengklasifikasikan menjadi kelas nol.

FP (False Positive) = Jumlah dokumen dari kelas nol yang salah mengklasifikasikan menjadi kelas satu.

FN (False Negative) = Jumlah dokumen dari kelas satu yang salah mengklasifikasikan menjadi kelas nol.

Tabel 1 Confusion Matrix		
	1	0
1	TP	FN
0	FP	TN

E. Google Colaboratory

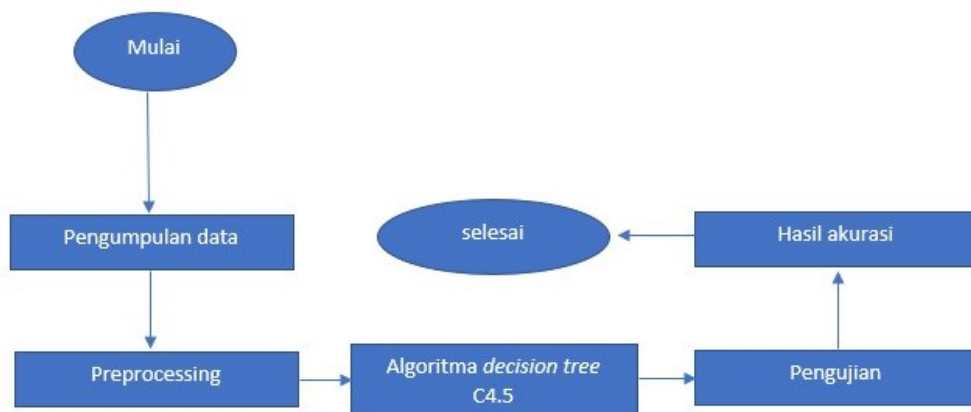
Google Colaboratory sekarang lebih dikenal dengan sebutan Google Colab atau hanya Colab. Google Colaboratory memiliki fitur menarik yang disajikan kepada penggunanya. Google juga mendukung dengan GPU yang bersifat open source. Tujuan utama Google membuat akses gratis yang disediakan untuk umum adalah agar bisa menjadikannya acuan sebagai standar utama dalam machine learning [4].

F. Weka

Weka merupakan aplikasi yang dikembangkan pada tahun 1994 dan pertama diperkenalkan di Universitas Waikato, Selandia Baru. Kelebihan dari aplikasi Weka adalah terdapat algoritma disertai dengan machine learning, dan algoritma yang terdapat dalam aplikasi selalu terupdate. Aplikasi Weka merupakan aplikasi yang terdiri dari beberapa algoritma machine learning dan bisa digunakan untuk generasi dari kumpulan data [5].

G. Presedur Penelitian

Gambar 3.2 Alur Prosedur Penelitian merupakan langkah dalam melakukan penelitian yang dimulai dari pengumpulan data, preprocessing, pengaplikasian algoritma Decision Tree C4.5, pengujian, lalu memperoleh hasil akurasi dan mengevaluasi.



Gambar 1 Prosedur Penelitian

Pada Gambar 4, Tahapan Algoritma Decision Tree menjelaskan tentang langkah-langkah pengaplikasian algoritma Decision Tree C4.5.

- a) Membuat pengujian atribut entropy dan gain.

Menghitung *entropy*

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

Keterangan

S : Kumpulan kasus.

N : Jumlah partisi dalam atribut.

Pi : Proporsi dari Si, pada S

Menghitung *gain*

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

Keterangan

S : Kumpulan kasus

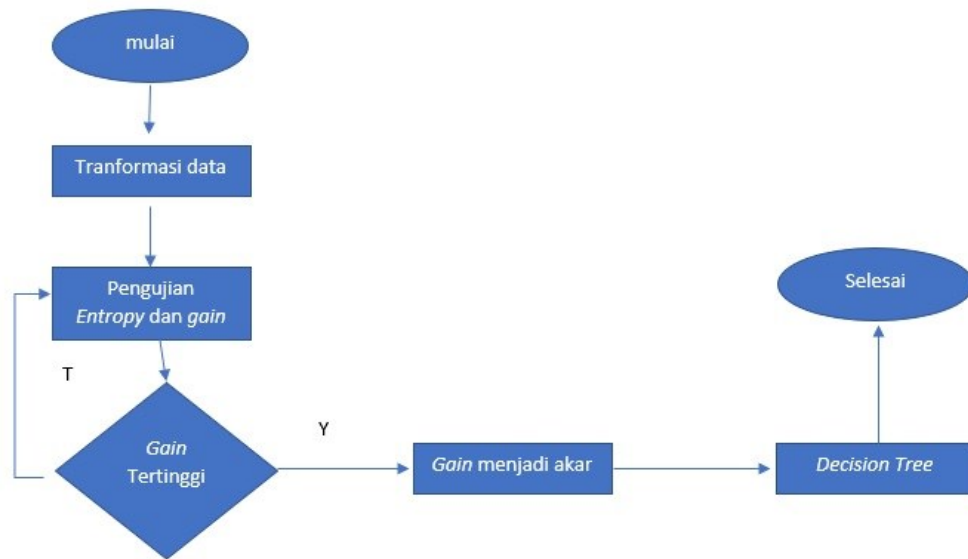
A : Atribut

N : Jumlah partisi dalam atribut

SI : Jumlah kasus dari partisi ke-i

S : jumlah kasus S

- b) Mencari nilai gain tertinggi.
 c) Jika gain tertinggi tidak ditemukan maka kembali melakukan pengujian atribut ulang.
 d) Jika gain tertinggi ditemukan maka dapat diketahui bahwa partisi data tersebut sesuai dengan atribut.
 e) Gain tertinggi menjadi *root* atau *node* akar.
 f) pembuatan *decision tree*.

Gambar 2 Tahapan pengujian algoritma *decision tree* c4.5

III. HASIL DAN PEMBAHASAN

A. Proses Pengolahan Data

Dataset ini merupakan dataset yang belum dilakukan preprocessing, yang diambil dari website Kaggle.com dengan data yang diupload pada bulan September tahun 2020, yang memiliki 524 data dan 14 atribut, serta keterangan lulus (tepat).

NAMA	JENIS KELAMIN	STATUS MAHASISWA	UMUR	STATUS NIKAH	IPS 1	IPS 2	IPS 3	IPS 4	IPS 5	IPS 6	IPS 7	IPS 8	IPK	STATUS KELULUSAN
STEVEN JULIAN	LAKI - LAKI	MAHASISWA	24	BELUM MENIKAH	3.17	2.7	3.23	2.41	3	2.47	1.75	0	2.75	TEPAT
LEYLA TRIYANA PRATIWI	PEREMPUAN	MAHASISWA	26	BELUM MENIKAH	3.6	3.5	3.42	2.85	3.31	2.95	2.18		3.39	TEPAT
VERIS SOFIYAN PRAYOGA	LAKI - LAKI	MAHASISWA	29	BELUM MENIKAH	2.67	2.66	2.93	3.14	2.92	2.64	2.88	0.5	2.81	TEPAT
ADITYA AKBAR NUGRAHA	LAKI - LAKI	MAHASISWA	27	BELUM MENIKAH	2.48	2.86	2.09	2.55	2.55	2.43	2.55	2.17	2.82	TEPAT
ERNA EKA RIYANTI	PEREMPUAN	MAHASISWA	25	BELUM MENIKAH	3.19	3.08	3.31	2.83	3.36	2.73	3.06	0	3.09	TEPAT
FARID DWI NORYANTO	LAKI - LAKI	MAHASISWA	24	BELUM MENIKAH	3.1	2.98	3.17	3.25	3.41	3.08	3.43	3	3.23	TEPAT
DAFIK HADI WINOTO	LAKI - LAKI	MAHASISWA	24	BELUM MENIKAH	2.98	2.68	2.23	2.86	2.25	2.64	1.52	2.1	2.54	TEPAT
WAHYU FITRIYANTO	LAKI - LAKI	MAHASISWA	24	BELUM MENIKAH	3.45	3.15	3.54	3.78	3.42	3.88	2.5	4	3.56	TEPAT
IMAM SURYO SUSANTO	LAKI - LAKI	MAHASISWA	26	BELUM MENIKAH	3.31	3.02	3.48	3.7	3.19	3.21	2.58	4	3.4	TEPAT
AHMAD SUTOPO	LAKI - LAKI	MAHASISWA	24	BELUM MENIKAH	2.62	2.73	2.11	3.36	2.79	3.17	2.45	0	2.97	TEPAT
HERI SUFA'AT	LAKI - LAKI	MAHASISWA	24	BELUM MENIKAH	3.24	3.06	2.79	2.73	3.02	2.59	3.08	0	3.08	TEPAT
MURYA AMIEN NUR PRABOWO	LAKI - LAKI	MAHASISWA	26	BELUM MENIKAH	2.83	3.02	3.02	3.63	3.21	3.09	2.57	3	3.33	TEPAT
DWI HESTYNA PRIHASTANTY	PEREMPUAN	MAHASISWA	32	BELUM MENIKAH	3	3.3	3.14	3.14	2.84	3.13	3.25	0	3.17	TERLAMBAT
MURYA ARIEF BASUKI	PEREMPUAN	BEKERJA	29	BELUM MENIKAH	3.5	3.3	3.7	3.29	3.53	3.72	3.73	0	3.54	TERLAMBAT
MARTINA FLAVERIA	LAKI - LAKI	BEKERJA	25	MENIKAH	3.05	2.92	2.27	2.65	3	3.45	3.24	2	3.09	TERLAMBAT
MASLUH JAMIL	LAKI - LAKI	BEKERJA	30	BELUM MENIKAH	2.64	3	2.9	2.77	3.25	3.5	3.22	2	3.08	TERLAMBAT
MURDIANTORO	LAKI - LAKI	MAHASISWA	32	BELUM MENIKAH	2.45	2.45	2.14	1.95	1.82	2.66	1.61	0	2.32	TERLAMBAT
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
NIMATUL JANNAH	PEREMPUAN	MAHASISWA	23	BELUM MENIKAH	3.02	2.94	3.25	2.87	3	2.94	3.09	3	3.16	TEPAT
DINDU SETYO WICAKSONO	LAKI - LAKI	MAHASISWA	23	BELUM MENIKAH	3.1	3.06	3	3.23	2.79	3	2.41	3	2.16	TEPAT

Gambar 3 *Dataset* sebelum dilakukan *Preprocessing*

Data yang telah terkumpul, maka tahap selanjutnya adalah melakukan pembersihan data dan menseleksi atribut agar memperoleh atribut-atribut yang sesuai. Pada Gambar 4.2, yaitu tabel yang sudah melalui tahapan preprocessing. Untuk pembagian jumlah data, dari 524 data, jumlah data training adalah 314 data, sedangkan jumlah data testing adalah 213 data.

$$\begin{aligned}\text{Baik} &= \left(-\frac{160}{219}\right) * \text{Imlog2} \left(\frac{160}{219}\right) + \left(-\frac{59}{219}\right) * \text{Imlog2} \left(\frac{59}{219}\right) \\ &= 0.840611557\end{aligned}$$

b) Menghitung *gain*

Gain Status Mahasiswa

$$\begin{aligned}0.999818247 - \left(\frac{187}{315} * 0.999484234\right) + \left(\frac{1}{315} * 1\right) \\ = 0.81282284\end{aligned}$$

Gain Umur

$$\begin{aligned}0.999818247 - \left(\frac{160}{315} * 0.896038234\right) + \left(\frac{155}{315} * 0.907165768\right) \\ = 0.99107087\end{aligned}$$

Gain Status Nikah

$$\begin{aligned}0.999818247 - \left(\frac{160}{315} * 1\right) + \left(\frac{155}{315} * 0.999249248\right) \\ = 0.98357581\end{aligned}$$

Gain IP 4

$$\begin{aligned}0.999818247 - \left(\frac{96}{315} * 0\right) + \left(\frac{219}{315} * 0.840611557\right) \\ = 1.58424342\end{aligned}$$

C. Pengujian

Pengujian dari data yang telah diolah dilakukan dengan menggunakan dua metode yang berbeda: pengujian perhitungan diuji menggunakan confusion matrix dan menggunakan software Weka, serta pengujian pemrograman Python yang menggunakan library Scikit-learn.

1. Pengujian *confusion matrix*

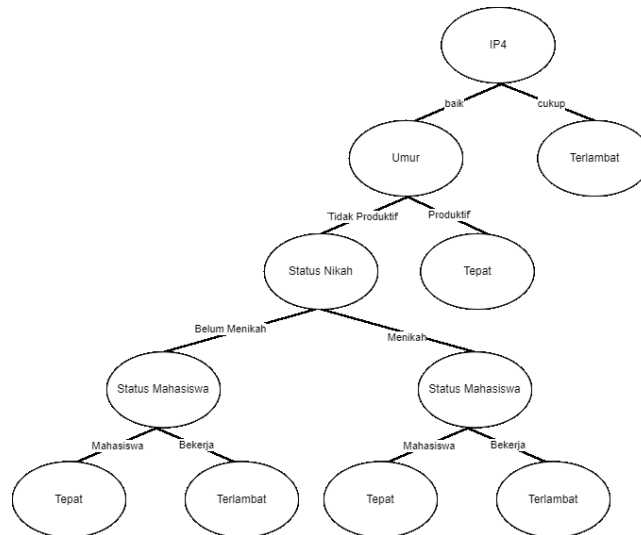
Tabel 2 Perhitungan <i>Confusion Matrix</i>		
	Tepat	Terlambat
Tepat	110	50
Terlambat	0	155
Total	315	

$$\begin{aligned}\text{Accuracy} &= \frac{TP+TN}{\text{jumlah total prediksi}} = \frac{110+155}{315} * 100\% \\ &= 71\%\end{aligned}$$

$$\begin{aligned}\text{Precision} &= \frac{TP}{TP+FP} = \frac{110}{110+0} * 100\% \\ &= 100\%\end{aligned}$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{110}{110+50} * 100\%$$

Berdasarkan hasil Tabel 2, yang menunjukkan hasil pengujian menggunakan confusion matrix, dalam accuracy menghasilkan nilai 71%, precision mendapatkan 100%, recall mendapatkan 84%, sedangkan f-measure mendapatkan hasil 69%. Perhitungan confusion matrix dengan total data 315, di mana TP berjumlah 110, FN berjumlah 50, FP berjumlah 0, dan TN berjumlah 155.



Gambar 6 Hasil pohon keputusan menggunakan excel

2. Pengujian python

Nilai akurasi dihitung dari prediksi yang benar dari variabel y dan y_{pred} dengan menggunakan perintah `accuracy_score`. Sedangkan nilai presisi adalah rasio jumlah positif benar dan positif palsu dalam mengklasifikasi label positif di label negatif. Nilai recall menunjukkan kemampuan klasifikasi untuk menemukan seluruh sampel positif.

```

✓ [15] #nilai akurasi
0s      akurasi = accuracy_score(y, y_pred)
        print('nilai akurasi : %d persen'%(akurasi*100))

⇨ nilai akurasi : 84 persen

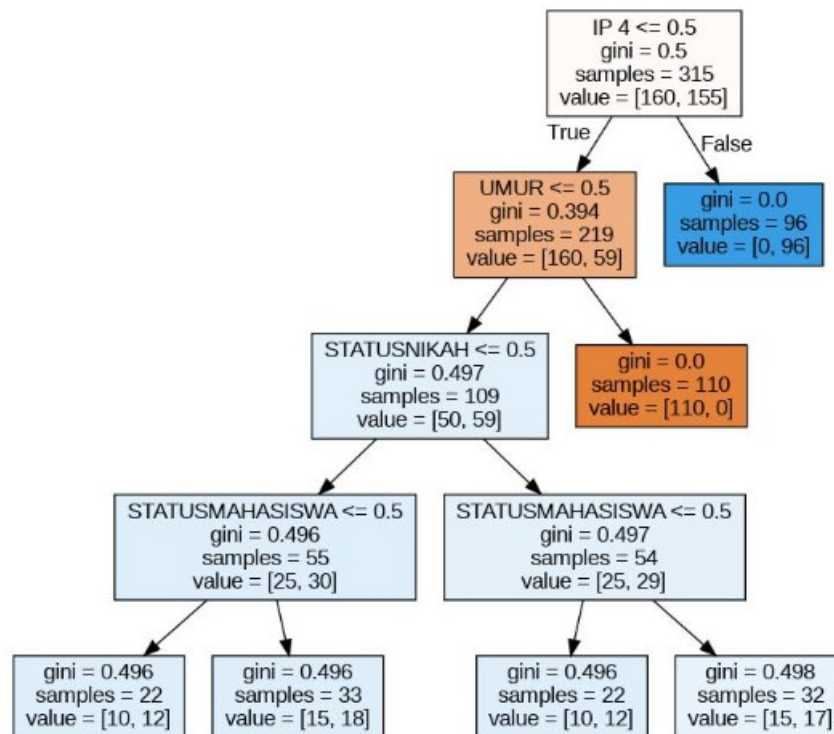
✓ [16] #nilai presisi
0s      presisi = precision_score(y, y_pred)
        print('nilai presisi : %d persen'%(presisi*100))

⇨ nilai presisi : 75 persen

✓ [17] #nilai recall
0s      recall = recall_score(y, y_pred)
        print('nilai recall : %d persen'%(recall*100))

⇨ nilai recall : 100 persen
  
```

Gambar 7 hasil akurasi presisi dan recall



Gambar 8 Hasil pohon keputusan dari python

IV. KESIMPULAN DAN SARAN

Penerapan algoritma Decision Tree C4.5 melalui tahapan-tahapan, di antaranya yaitu melakukan perhitungan nilai entropy dan mencari nilai gain. Jika telah mendapatkan nilai gain tertinggi, maka dapat diketahui bahwa partisi data telah sesuai dengan atribut. Kemudian, nilai gain tertinggi menjadi node akar sampai perhitungan nilai atribut dari nilai gain tertinggi tersebut membentuk semua cabang terpartisi. Setelah proses tersebut selesai, tahapan selanjutnya yaitu pembuatan pohon keputusan atau decision tree. Algoritma Decision Tree C4.5 dalam permasalahan kelulusan mahasiswa tepat waktu diukur berdasarkan nilai accuracy, precision, dan recall dengan pembagian data 40% menjadi data testing, sedangkan 60% menjadi data training. Hasil pengujian yang menggunakan Weka memiliki nilai accuracy 71%, precision 100%, recall 69%. Sedangkan, accuracy Python mendapatkan hasil 84%, precision 75%, dan recall 100%.

Saran untuk penelitian selanjutnya yaitu melakukan perbandingan algoritma Decision Tree C4.5 dengan algoritma prediksi lain agar mengetahui perbedaan dan bisa dijadikan acuan bagi penelitian selanjutnya dengan permasalahan yang sama.

PENGAKUAN

Naskah ilmiah ini merupakan sebagian penelitian pada Tugas Akhir milik Bustomi Alwi dengan judul "Prediksi Algoritma Decision Tree C4.5 dalam Ketepatan Waktu Kelulusan Mahasiswa," yang dibimbing langsung oleh Bapak Sutan Faisal, M.Kom, dan Ibu Hilda Yulia Novita, M.Kom.

DAFTAR PUSTAKA

- [1] D. Y. Hakim Tanjung, "Optimalisasi Algoritma C4.5 untuk Prediksi Kerusakan Mesin ATM," *INFOSYS (INFORMATION Syst. J.*, vol. 6, no. 1, p. 12, 2021, doi: 10.22303/infosys.6.1.2021.12-21.
- [2] F. M. Hana, "Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5," *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 4, no. 1, pp. 32–39, 2020, doi: 10.47970/siskom-kb.v4i1.173.
- [3] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," *J. Sains Komput. Inform.*, vol. 5, no. 2, pp. 697–711, 2021.
- [4] A. A. Rizal *et al.*, "Peningkatan Efektifitas Programming Dengan Pelatihan Python for Data Science Bagi Komunitas Programming Pondok Pesantren," vol. 1, no. 1, pp. 13–19, 2021.
- [5] S. Turnip and P. Siltionga, "Analisis Pola Penyebaran Penyakit dengan Menggunakan Algoritma C4.5," *J. Tek. Inform. Unika St. Thomas*, vol. 03, no. 479, pp. 3–7, 2018.