PREDIKSI TINGKAT KELULUSAN MAHASISWA MENGGUNAKAN ALGORITMA C4.5
(Studi Kasus: Informatika Universitas AMIKOM Yogyakarta)

Arif Budiman[1], Arief Setyanto[2], Ferry Wahyu Wibowo[3]

Jurusan Magister Teknik Informatika
Universitas AMIKOM Yogyakarta

arif.6122@students.amikom.ac.id[1], arief_s@amikom.ac.id[2],
ferry.w@amikom.ac.id[3]

Abstrak

Mahasiswa merupakan aset penting yang dimiliki oleh institusi penyelenggara pendidikan. Lama masa studi mahasiswa dalam menyelesaikan kuliahnya merupakan salah satu bagian penilaian akreditasi bagi perguruan tinggi, namun masih banyak mahasiswa yang tidak dapat meyelesaikan kuliahnya selama 8 semester karena berbagai macam faktor akademik seperti nilai IPK semester 1, IPK semester 2, IPK semester 3, IPK semester 4 dan Jumlah SKS. Sehingga hal itu mempengaruhi ketepatan kelulusan mahasiswa dalam proses penyelenggaraan pendidikan di perguruan tinggi. Hasil prediksi ini dapat digunakan oleh pengelola program studi dalam membina mahasiswanya dalam membuat keputusan yang tepat untuk meningkatkan ketepatan masa studi. Penelitian ini berfokus untuk menguji kelayakan prediksi tingkat kelulusan mahasiswa Universitas AMIKOM Yogyakarta menggunakan algoritma C4.5 dan dibandingkan dengan algoritma ID3 dan CART. Pengujian dengan 10-fold cross validation sekaligus evaluasi kinerja model menggunakan tool RapidMiner. Hasil penelitian menggunkan model algoritma C4.5 menggunakan pembagian data yang paling optimal yaitu 70:30 dengan tingkat recall sebesar 95.59% dan akurasi sebesar 76.10%

Kata Kunci: Prediksi, Kelulusan Mahasiswa, Algoritma C4.5

1. Pendahuluan

Mahasiswa merupakan aset penting yang dimiliki oleh institusi penyelenggara pendidikan. Lama masa studi mahasiswa dalam menyelesaikan kuliahnya merupakan salah satu bagian penilaian akreditasi bagi perguruan tinggi, namun masih banyak mahasiswa yang tidak dapat meyelesaikan kuliahnya selama 8 semester karena berbagai macam faktor. Sehingga hal itu mempengaruhi ketepatan kelulusan mahasiswa dalam proses penyelenggaraan pendidikan di perguruan tinggi. *Data Mining* adalah teknik yang memanfaatkan data dalam jumlah yang besar yang tersedia didalam database untuk menguraikan penemuan pengetahuan (Turban, dkk., 2005). Terdapat banyak metode dalam data mining diantaranya seperti *Decision Trees*, *Bayesian*, *Artificial Neural Networks*, *Nearest Neighbor*, *Support Vector Machines* dan lainnya (Suyanto, 2017). Penelitian ini berfokus untuk menguji kelayakan prediksi tingkat kelulusan mahasiswa Universitas AMIKOM Yogyakarta menggunakan

algoritma C4.5 dan dibandingkan dengan algoritma ID3 dan CART untuk menemukan algoritma yang memiliki tingkat akurasi tertinggi.

Penelitian terkait mengenai prediksi tingkat kelulusan mahasiswa antara lain seperti pada penelitian (Pertiwi, dkk., 2017) bertujuan untuk mengklasifikasikan tingkat *drop out* berdasarkan provinsi di indonesia menjadi dua bagian yaitu Tinggi dan Rendah .Dalam penelitian dataset yang dikumpulkan berjumlah 128 baris data dengan pembagian data 60% sebagai data latih dan 40% sebagai data uji berdasarkan 7 variabel pendukungnya antara lain seperti jumlah penduduk miskin, jumlah pembiayaan daerah, jumlah pendapatan daerah, jumlah belanja daerah, tpak, rasio gini, dan tingkat putus sd. Pengujian yang digunakan dalam penelitian ini adalah *confussion matrix* dan menghasilkan tingkat akurasi sebesar 71.2%.

Penelitian (Amin dkk., 2015) yang menggunakan data berjumlah 1000 data dengan 70% data diterima dan sisanya 30% ditolak. Dari hasil pengujian pembagian data menggunakan 5 cara antara lain 100%, 90%;10%, 80%:20%, 70%:30%, 60%:40% menunjukan bahwa pembagian 80%:20% yang terbaik karena memiliki tingkat akurasi dan presisi tertinggi dan sekaligus menunjukan tingkat akurasi algoritma C4.5 sebesar 74,5%.

Penelitian (Anam dan Santoso, 2018) mencoba membandingkan tingkat akurasi antara algoritma C4.5 dan dan *Naive Bayes* dengan hasil pengujian menunjukan algoritma C4.5 memiliki tingkat akurasi sebesar 96.40% lebih baik dari tingkat akurasi algoritma *Naive Bayes* sebesar 95.11%. Pengujian divalidasi dengan 10-fold cross validation dan evalusi menggunakan *confusion matrix*.

Berdasarkan latar belakang masalah yang telah dijelaskan sebelumnya maka dalam penelitian penulis akan menganalisis prediksi tingkat kelulusan mahasiswa menggunakan algoritma C4.5 dengan menggunakan 5 variabel antara lain IPK Semester 1, IPK Semester 2, IPK Semester 3, IPK Semester 4, dan Jumlah SKS. dan akan dievaluasi menggunakan *confusion matrix*.

2. Kajian Literatur

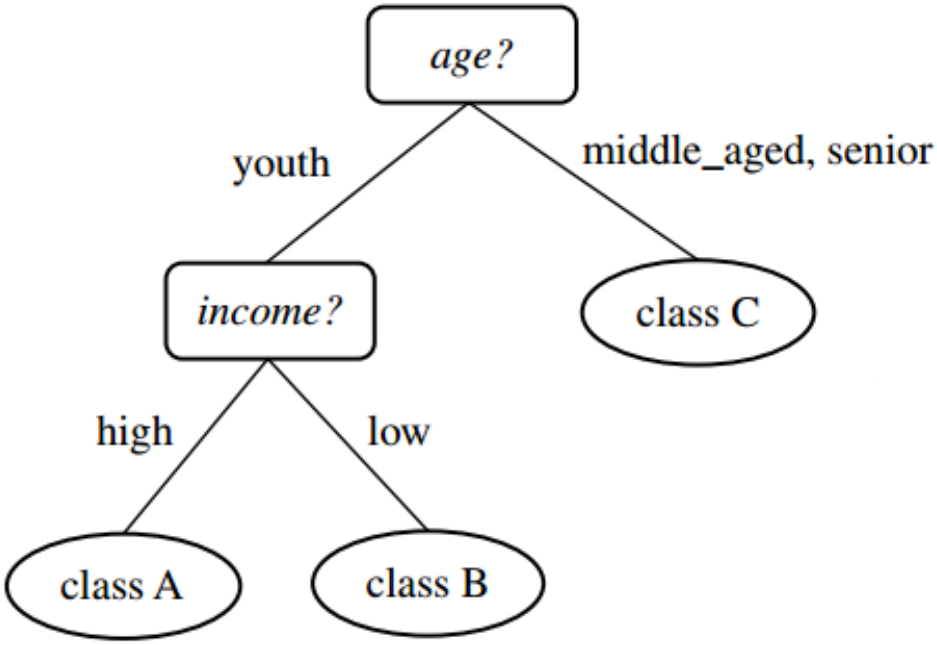
2.1 Data Mining

Data mining merupakan proses untuk menemukan sebuah pengetahuan dengan mencari pola tersembunyi dalam database yang menggunakan menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat (Turban, dkk., 2005). Data mining merupakan serangkaian proses untuk

menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basis data yang ada yang biasanya berupa data yang sangat besar dengan tujuan untuk memanipulasi data menjadi informasi yang lebih berharga (Kusrini dan Luthfi, 2009).

2.2 Decision Tree

Pohon keputusan merupakan metode klasifikasi dan prediksi yang kuat dan terkenal, metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang dapat merepresentasi aturan yang akan dihasilkan. Pohon keputusan beguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel dan variabel target. Terdapat beberapa algoritma yang dapat dipakai dalam pembentukan pohon keputusan antara lain ID3, CART dan C4.5 (Kusrini dan Luthfi, 2009). contoh pohon keputusan seperti yang ditampilkan pada Gambar 1



Gambar 1 Pohon keputusan (Han, dkk., 2012)

Pada pohon keputusan klasifikasi keputusannya adalah Class A, Class B atau Class C. Apabila Age = Youth & Income = high (Class A), Age = Youth & Income = Low (Class B), Age = Middle_aged, Senior (Class C)

2.3 Algoritma C4.5

Algoritma C4.5 digunakan untuk membentuk pohon keputusan yang dapat digunakan untuk membentuk pohon keputusan yang ditemukan oleh John

Ross Quinlan. Algoritma C4.5 merupakan pengembangan dari algoritma ID3. Berbeda dengan ID3 yang menggunakan information gain dalam algoritma C4.5 pemilihan atribut dilakukan dengan menggunakan *Gain Ratio*. Gain Ratio digunakan untuk Untuk mengatasi atribut yang memiliki nilai yang sangat bervariasi dan dihitung berdasarkan *Split Information* (Suyanto, 2017).

2.4 Evaluasi Kinerja

Evaluasi digunakan untuk menguji model klasifikasi data mining untuk mengetahui kinerja sistem. Metode yang digunakan antara lain :

1. *k-fold Cross Validation*

Merupakan salah satu teknik memvalidasi akurasi sebuah model yang dibangun berdasarkan data set tertentu. Metode ini membagi data set menjadi dua bagian, yaitu data training dan data testing. Data set dibagi menjadi sejumlah k-buah partisi secara acak. Kemudian dilakukan sejumlah k-kali proses klasifikasi, dimana masing-masing proses pengujian menggunakan data partisi ke-k sebagai data testing danmemanfaatkan sisa partisi lainnya sebagai *data training* (Anam dan Santoso, 2018).

2. *Confusion Matrix*.

Confusion matrix adalah tool yang digunakan sebagai evaluasi model klasifikasi untuk memperkirakan objek yang benar atau salah. Sebuah matrix dari prediksi yang akan dibandingkan dengan kelas sebenarnya atau dengan kata lain berisi informasi nilai sebenarnya dan prediksi pada klasifikasi (Gonureschu, 2011). seperti yang ditampilkan pada Tabel 1

Tabel 1 *Confusion Matrix* Dua Kelas

Classification	*Predicted Class*	
	Class = Yes	Class = No
Class = Yes	TP (*True Positive*)	FN (*False Negative*)
Class = No	FP (*False Positive*)	TN (*True Negative*)

Berikut persamaan untuk menghitung *akurasi, presisi,* dan *recall* pada *confusion matrix* adalah ditunjukkan pada persamaan 1,2, dan 3. (Gonureschu, 2011).

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (1)$$

$$\Pr esisi = \frac{TP}{TP + FP} \times 100\% \quad (2)$$

$$\text{Re}\, call = \frac{TP}{TP + FN} \times 100\% \quad (3)$$

2.5 *RapidMiner*

RapidMiner merupakan perangakat lunak untuk melakukan analisis terhadap *data mining*, text mining dan analisis prediksi. *RapidMiner* menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik (Fathushabib, 2016).

3. Metode Penelitian

Berikut adalah gambaran proses *Data Mining* dengan menggunakan C4.5 yaitu:

1. *Menghitung* nilai *entropy* total dan *entropy* dari masing-masing atribut dilakukan dengan persamaan 4 untuk mengukur tingkat homogenitas dan *purity* dari data.

$$Entropy(S) = \sum_{i}^{c} - P_i Log_2 P_i \quad (4)$$

Keterangan:

S : himpuanan kasus

c : jumlah nilai yang terdapat pada atribut target (jumlah kelas)

pi : rasio antar jumlah sampel dikelas i dengan jumlah semua sampel pada himpunan data

2. *Menghitung* nilai *information gain* dilakukan dengan persamaan 5 dengan menggunakan nilai *entropy* yang telah diperoleh untuk mengukur efektifitas suatu atribut dalam mengklasifikan data.

$$Gain(S, A) = Entropy(S) - \frac{|S_v|}{|S|} Entropy(S_v) \quad (5)$$

Keterangan :

A : atribut

V : menyatakan suatu nilai yang mungkin untuk atribut A

Value(A) : himpunan nilai-nilai yang mungkin untuk atribut A

$|S_v|$: jumlah sampel untuk nilai v

$|S|$: jumlah seluruh sampel data

Entropy (S_v) : *entropy* untuk sampel-sampel yang memiliki nilai v

1. Menghitung nilai *split information* dilakukan dengan persamaan 6 dari setiap atribut berisi normalisasi dari *information gain* yang memperhitungkan *entropy* dari distribusi probabilitas subset setelah dilakukan proses partisi

$$SplitInformation(S,A) = \sum_{i=1}^{c} -\frac{|S_i|}{|S|} Log_2 \frac{|S_i|}{|S|} \quad (6)$$

Keterangan:

S : himpunan sampel data

S_1 – Sc : sub himpunan sampel data yang terbagi berdasarkan jumlah variasi nilai pada atribut A

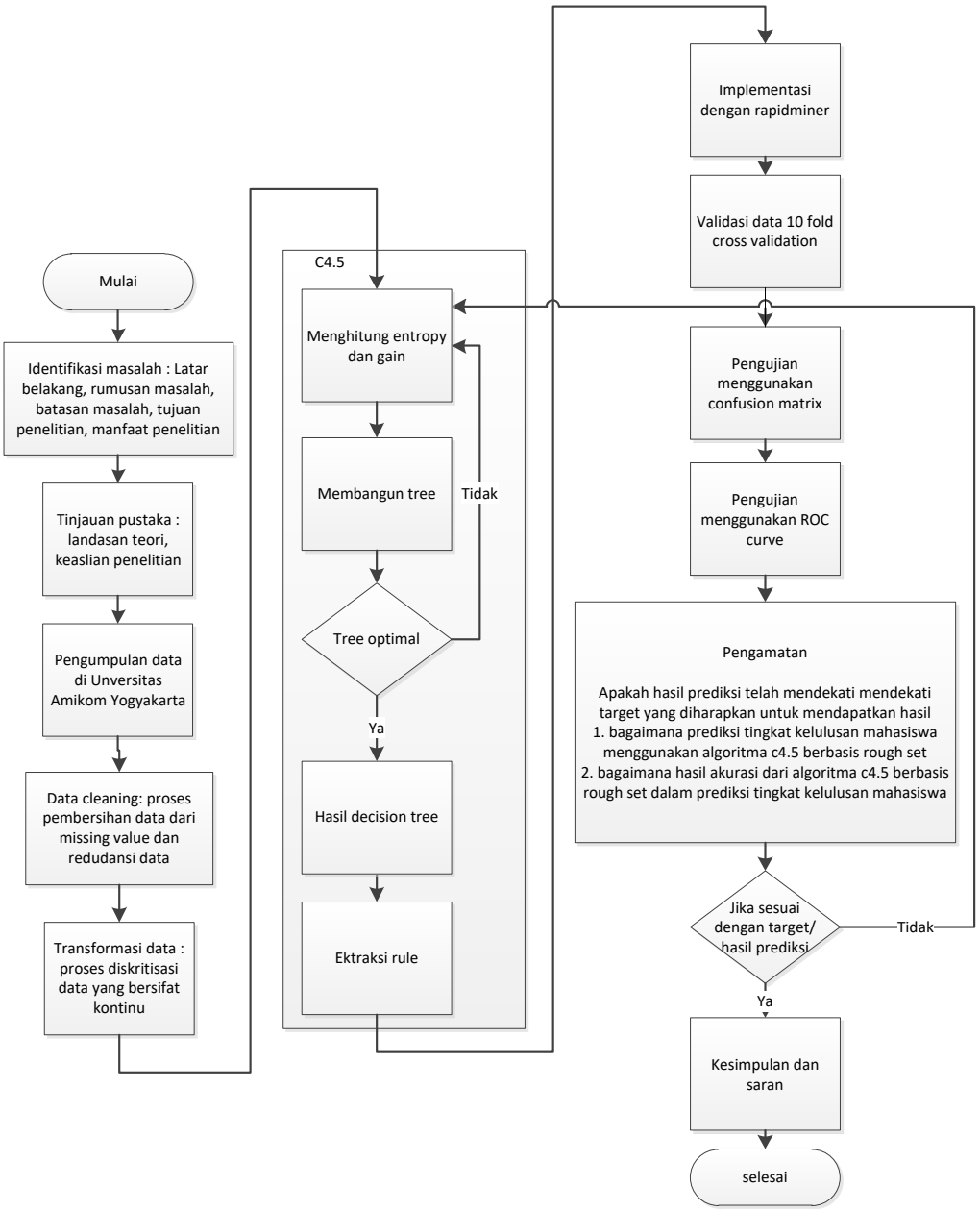
2. Menghitung gain ratio dilakukan dengan persamaan 7 menggunakan nilai *information gain* dan *split information*

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInformation(S,A)} \quad (7)$$

3. Mengambil nilai *gain ratio* terbesar sebagai simpul akar.
4. Menghilangkan atribut yang sudah dipilih sebelumnya dan ulangi perhitungan nilai *entropy*, *information gain*, *split info*, dan *gain ratio* dengan memilih *gain ratio* terbesar sebagai simpul internal pohon.
5. Mengulangi perhitungan tersebut hingga semua atribut memiliki kelas

3.1. Alur Penelitian

Secara umum alur penelitian yang dilakukan mengacu pada kerangka penelitian seperti pada Gambar 2.



Gambar 2 Alur Penelitian

4. HASIL DAN PEMBAHASAN

1. Pengumpulan Data

Proses pengambilan data terkait yang dibutuhkan untuk memprediksi kelulusan mahasiswa pada Universitas AMIKOM Yogyakarta. Data yang didapat merupakan data mahasiswa jurusan Informatika tahun 2011 sehingga diperoleh 705 baris data

2. *Data Selection*

   Proses pemilihan data untuk menghasilkan data yang sesuai dengan kebutuhan dalam prediksi kelulusan mahasiswa. Dalam prediksi kelulusan mahasiswa dibutuhkan data mahasiswa yang telah lulus. Data yang dapat digunakan adalah data mahasiswa tahun kelulusan 2015 dan 2016

3. Data Cleaning

   Proses pembersihan terhadap data yang dilakukan untuk memastikan data yang diperoleh sebelumnya dapat digunakan serta bebas dari duplikasi, kesalahan dan *validation rules* sudah sesuai.

4. Transformasi Data

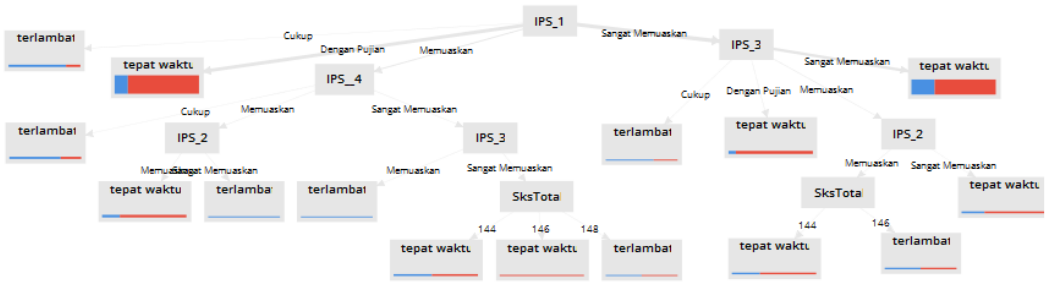
   Transformasi data dilakukan untuk mengubah data menjadi nilai dengan format tertentu. Hal ini dilakukukan untuk membagi data yang memiliki cakupan yang luas kedalam beberapa kelompok kecil. Seperti yang ditampilkan pada Tabel 2

Tabel 2 Transformasi Data

Atribut	Nilai atribut	keterangan
IPK semester 1	2,00 < IPK < 2,50	Cukup
	2,50 < IPK < 3,00	Memuaskan
	3,00 < IPK < 3,50	Sangat memuaskan
	IPK > 3,50	Dengan pujian
IPK semester 2	2,00 < IPK < 2,50	Cukup
	2,50 < IPK < 3,00	Memuaskan
	3,00 < IPK < 3,50	Sangat memuaskan
	IPK > 3,50	Dengan pujian
IPK semester 3	2,00 < IPK < 2,50	Cukup
	2,50 < IPK < 3,00	Memuaskan
	3,00 < IPK < 3,50	Sangat memuaskan
	IPK > 3,50	Dengan pujian
IPK semester 4	2,00 < IPK < 2,50	Cukup
	2,50 < IPK < 3,00	Memuaskan
	3,00 < IPK < 3,50	Sangat memuaskan
	IPK > 3,50	Dengan pujian
Jumlah SKS	144	144
	146 148 150	146 148 150

5. Permodelan

   Proses penggunaan algoritma C4.5 untuk melakukan klasifikasi terhadap dataset yang telah dihasilkan sebelumnya. Hasil klasifikasi dapat digunakan untuk memprediksi tingkat kelulusan mahasiswa. Berikut adalah hasil pohon keputusan dari algoritma C4.5 seperti yang ditampilkan pada Gambar 3



Gambar 3 Pohon Keputusan

Dari pohon keputusan tersebut dapat diambil *rules* sebagai berikut:

I IPS_1 = Cukup: terlambat {terlambat=12, tepat waktu=3}

IPS_1 = Dengan Pujian: tepat waktu {terlambat=29, tepat waktu=155}

IPS_1 = Memuaskan

| IPS__4 = Cukup: terlambat {terlambat=10, tepat waktu=4}

| IPS__4 = Memuaskan

| | IPS_2 = Memuaskan: tepat waktu {terlambat=4, tepat waktu=15}

| | IPS_2 = Sangat Memuaskan: terlambat {terlambat=3, tepat waktu=0}

| IPS__4 = Sangat Memuaskan

| | IPS_3 = Memuaskan: terlambat {terlambat=2, tepat waktu=0}

| | IPS_3 = Sangat Memuaskan

| | | SksTotal = 144: tepat waktu {terlambat=5, tepat waktu=6}

| | | SksTotal = 146: tepat waktu {terlambat=0, tepat waktu=3}

| | | SksTotal = 148: terlambat {terlambat=1, tepat waktu=1}

IPS_1 = Sangat Memuaskan

| IPS_3 = Cukup: terlambat {terlambat=2, tepat waktu=1}

| IPS_3 = Dengan Pujian: tepat waktu {terlambat=2, tepat waktu=21}

| IPS_3 = Memuaskan

| | IPS_2 = Memuaskan

| | | SksTotal = 144: tepat waktu {terlambat=3, tepat waktu=6}

| | | SksTotal = 146: terlambat {terlambat=4, tepat waktu=4}

| | IPS_2 = Sangat Memuaskan: tepat waktu {terlambat=3, tepat waktu=8}

| IPS_3 = Sangat Memuaskan: tepat waktu {terlambat=41, tepat waktu=108}

6. Pengujian

Proses pengujian terhadap *rules* yang terbentuk dari pemodelan *data training* menggunakan algoritma C4.5 dilakukan dengan menggunakan *Ten-fold cross validation* dan *confussion matrix*. Sehingga ditetahui kemampuan model dalam memprediksi tingkat kelulusan mahasiswa yang ditampilkan

pada tabel 3 dan pengukuran kinerja permodelan dibagi menjadi beberapa bagian sepeti ditunjukan pada tabel 4 dan 5

Tabel 3 *Confussion Matrix*

Aktual	Prediksi	
	Tepat Waktu	Terlambat
Tepat Waktu	319	16
Terlambat	93	28

Tabel 4 Detail Pengukuran Kinerja Algoritma ID3

No	Data Latih	Data Uji	Presisi	Recall	Akurasi
1	90%	10%	77.07%	95.13%	75.64%
2	80%	20%	77.04%	93.73%	74.85%
3	70%	30%	77.24%.	95.22%,	75.88%
4	60%	40%	76.64%	93.73%	74.45%
5	50%	50%	75.96%	90.83%	72.12%

Tabel 5 Detail Pengukuran Kinerja Algoritma C4.5

No	Data Latih	Data Uji	Presisi	Recall	Akurasi
1	90%	10%	76.87%	95.59%	75.65%
2	80%	20%	76.28%	93.21%	73.70%
3	70%	30%	77.43%.	95.22%,	76.10%
4	60%	40%	77.14%	94.08%	75.21%
5	50%	50%	76.55%	92.50%	73.65%

Tabel 5 Detail Pengukuran Kinerja Algoritma CART

No	Data Latih	Data Uji	Presisi	Recall	Akurasi
1	90%	10%	77.02%	94.90%	75.47%
2	80%	20%	77.04%	93.73%	74.85%
3	70%	30%	77.24%.	95.22%,	75.88%
4	60%	40%	76.64%	93.73%	74.45%
5	50%	50%	75.87%	90.42%	71.82%

4. Kesimpulan

Berdasarkan hasil implementasi algoritma C4.5 pada prediksi tingkat kelulusan mahasiswa informatika di Universitas AMIKOM Yogyakarta dapat diambil beberapa kesimpulan bahwa nilai akurasi tertinggi sebesar 76.10% pada

algoritma C4.5 dengan pembagian data yang paling optimal yaitu 70:30. Faktor yang paling mempengaruhi kelulusan mahasiswa adalah IPK Semester 1. Pada penelitian selanjutnya dapat digunakan algoritma lain seperti JST, *Naïve Bayes* dan *SVM* untuk dianalisis dan dibandingkan. Selain itu dapat juga menambahkan variabel yang mempengaruhi kelulusan dengan menggunakan variasi data yang lebih banyak

Daftar Pustaka

Anam, C., dan Santoso, H. B. (2018). Perbandingan Kinerja Algoritma C4 . 5 dan Naive Bayes untuk Klasifikasi Penerima Beasiswa. *Jurnal Ilmiah Ilmu-Ilmu Teknik Vol.8 No.1 Edisi Mei 201*, *8*(1), 13–19.

Fathushabib. (2016). *Data Mining Untuk Prediksi Mahasiswa Non Aktif*. STMIK AMIKOM Yogyakarta.

Gonureschu, F. (2011). *Data Mining : Concepts, Models and Tecniques*. New York: Springer - Verlag Berlin Heidelberg.

Han, J., Kamber, M., dan Pei, J. (2012). *Data mining Concepts and Techniques 3nd Edition*. Waltham: The Morgan Kaufmann Publisher.

Kusrini, dan Luthfi, E. T. (2009). *Agoritma Data Mining*. Yogyakarta: Penerbit Andi.

Suyanto. (2017). *Data Mining Untuk Klasifikasi dan Klasterisasi Data*. Bandung: Penerbit Informatika.

Turban, E., Aronson, J. E., dan Liang, T.-P. (2005). *Decision Support Systems and Intelligent Systems* (7th ed.). USA: Pearson Education.