
Penerapan Fungsi Data Mining Klasifikasi untuk Prediksi Masa Studi Mahasiswa Tepat Waktu pada Sistem Informasi Akademik Perguruan Tinggi

Irwan Budiman*¹, Muliadi², Retma Ramadina³

^{1,2,3} Program Studi Ilmu Komputer FMIPA Universitas Lambung Mangkurat Banjarbaru (UNLAM),
telp/fax of institution/affiliation

e-mail: *¹irwan.budiman@unlam.ac.id, ²muliadizz@gmail.com ³retmaradina@gmail.com

Abstrak

Waktu kelulusan merupakan permasalahan umum bagi pihak program studi dan mahasiswa, karena kedua pihak tersebut sama-sama tidak dapat memprediksi masa studi tepat waktu mahasiswa. Teknik data mining klasifikasi dapat memecahkan masalah ini, yakni dengan menggunakan algoritma C4.5. Hasil penelitian menunjukkan persentasi sebesar 73% terhadap data yang telah diuji. Dengan demikian, pengolahan data ini dapat digunakan oleh program studi untuk memprediksi masa studi tepat waktu mahasiswa.

Kata kunci: C4.5, Klasifikasi, Data Mining, SIA

1. PENDAHULUAN

Data mining sendiri merupakan sebuah proses ekstraksi informasi untuk menemukan pola (*pattern recognition*) yang penting pada tumpukan data dalam database sehingga menjadi pengetahuan (*knowledge discovery*) [2]. Fungsi-fungsi dalam data mining antara lain fungsi deskripsi, fungsi estimasi, fungsi prediksi, fungsi klasifikasi, fungsi clustering dan fungsi asosiasi. Dalam penelitian ini adanya pemanfaatan database yang tidak maksimal, sehingga tidak terdapat penemuan dan pembentukan pola yang bisa digunakan sebagai klasifikasi untuk mengetahui sejauh mana mahasiswa menjalani masa studi[1].

Klasifikasi merupakan sebuah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep dan kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang kelasnya tidak diketahui. *Decision tree* adalah salah satu metode klasifikasi yang paling populer karena mudah untuk diinterpretasi oleh manusia. Model klasifikasi ini terdiri dari dua yaitu deskripsi dan prediksi. Algoritma *decision tree* yang paling terkenal adalah C4.5. Algoritma ini termasuk algoritma terbaik dalam data mining dan berada pada posisi pertama[4].

Pada Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia No.49 tahun 2014 Tentang Standar Nasional Pendidikan Tinggi pasal 17 ayat 2 “144 sks untuk diploma empat dan program sarjana”, dan ayat 3 “4 (empat) sampai 5 (lima) tahun untuk program diploma empat dan program sarjana”. Hal ini sesuai dengan peraturan akademik Universitas Lambung Mangkurat pasal 28 ayat 5 “Seorang dinyatakan lulus Program Sarjana (S1) apabila telah menempuh jumlah kredit antara 144-160 SKS”.

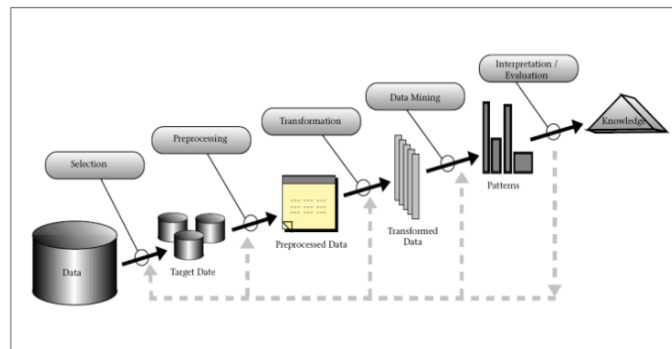
Jika masa studi mahasiswa dapat diketahui lebih dini maka pihak jurusan Ilmu Komputer dapat melakukan tindakan-tindakan yang diperlukan supaya mahasiswa dapat lulus dalam waktu kurang dari lima tahun atau tepat lima tahun berdasarkan Peraturan Menteri Pendidikan.

2. METODE PENELITIAN

Prosedur Penelitian

1. Analisis Data Mining

Untuk analisis data mining ini meliputi tahapan proses *Knowledge Discovery in Database* (KDD). Keunggulan dari proses KDD adalah proses terorganisir untuk mengidentifikasi pola yang valid, baru, berguna dan dapat dimengerti dari sebuah data *set* yang besar dan kompleks. Dalam proses KDD melibatkan algoritma untuk mengeksplorasi data, mengembangkan dan menemukan model yang pola sebelumnya tidak diketahui. KDD juga digunakan untuk memahami keadaan data, analisis dan prediksi. Adapun tahapan dari proses *Knowledge Discovery in Database* (KDD), yaitu:



Gambar 1. Proses Metode KDD

a. *Data Selection*

Pada tahap ini dilakukan pemilihan data yang akan digunakan dalam proses *data mining*. Data yang digunakan dalam penelitian ini berasal dari database Sistem Informasi Akademik di Program Studi Ilmu Komputer FMIPA Universitas Lambung Mangkurat yang sudah ada dan terbentuk tetapi belum sesuai dengan format *dataset* penelitian.

b. *Pre-Processing* dan *Cleaning Data*

Dalam tahap ini kehandalan data ditingkatkan, termasuk pembersihan data seperti menangani data yang tidak lengkap, menghilangkan gangguan atau outlier. Tahapan *Pre-Processing* dan *Cleaning Data* perlu dilakukan agar data terhindar dari duplikasi data, data yang tidak konsisten, memperbaiki kesalahan data, atau dilakukan penambahan data untuk menunjang sistem yang dibuat.

c. *Transformation*

Setelah melakukan tahapan *Pre-Processing* dan *Cleaning Data* kemudian dilanjutkan dengan tahapan *Transformation*, yaitu tahap untuk mengubah data dari bentuk asalnya menjadi data yang siap untuk ditambang. Karena hal tersebut dapat memudahkan dalam proses penggalian data untuk menemukan suatu pengetahuan baru. Berikut terdapat konsep dari data *set* setelah melalui tahap transformasi dan akan di *mining*.

Tabel 1. Konsep Data set

Nama	IP1	IP2	IP3	IP4	Lama Studi	Keterangan
Nadya Rosana	2.2	1.53	2.17	1.88	5	Tepat Waktu
M.Fazeriansyah	3	3.56	2.18	2.32	6	Tidak Tepat Waktu
Mufidah	2.63	3	4	4	5	Tepat Waktu
...
...
Riski Yulianto	2	1.53	1.84	2	6	Tidak Tepat Waktu

d. *Data Mining*

Pada tahap ini memilih tugas data *mining* yang cocok dan menentukan tipe data *mining* yang akan digunakan. Pada tahap ini juga dilakukan pemilihan algoritma dan implementasi dari algoritma yang dipilih. Proses *data mining* ini melakukan pencarian pola atau informasi menarik dalam data terpilih dengan menerapkan algoritma C4.5. Algoritma C4.5 termasuk ke dalam 10 (sepuluh) algoritma terbaik dalam data *mining* dan menempati posisi pertama sebagai algoritma yang paling populer untuk penggunaannya [5]. Adapun langkah-langkah dari algoritma C4.5, yaitu sebagai berikut:

- 1) Pilih atribut sebagai akar
- 2) Buat cabang untuk masing-masing nilai
- 3) Bagi kasus dalam cabang
- 4) Ulangi proses untuk masing-masing cabang sampai semua kasus pada cabang memiliki kelas yang sama

e. *Interpretation / Evaluasi*

Pada tahap ini dilakukan *interpretation/evaluasi* dari pola yang dihasilkan dari hasil klasifikasi data dengan menggunakan algoritma C4.5. Pada tahap ini juga penggunaan dan umpan balik secara keseluruhan pada pola dan hasil penemuan diperoleh dengan data *mining*.

Algoritma C4.5

Algoritma C4.5 merupakan perkembangan dari algoritma ID3, dikembangkan oleh Quinlan Ross pada tahun 1993. Algoritma C4.5 merupakan penerapan dari algoritma Hunt yang juga diterapkan pada ID3 dan CART. Algoritma ini menggunakan proses rekursif dalam membentuk sebuah pohon keputusan. Ketika sebuah atribut dipilih menjadi *node* pemecahan atau cabang, maka atribut tersebut tidak diikuti lagi dalam perhitungan nilai *information gain*. Proses rekursif berhenti ketika atribut memiliki cabang berupa kelas atau tidak ada lagi atribut yang bisa dipecah untuk menjadi cabang pohon keputusan. Pohon keputusan adalah model prediksi menggunakan struktur pohon atau struktur berhirarki. Contoh dari pohon keputusan dapat dilihat berikut ini [3].



Gambar 2. Model Pohon Keputusan

3. HASIL DAN PEMBAHASAN

Berikut hasil penerapan metode KDD dan algoritma C4.5.

1.1. Data Selection

Data yang digunakan adalah database SIA Program Studi Ilmu Komputer FMIPA Universitas Lambung Mangkurat Banjarbaru. Data yang digunakan adalah data mahasiswa dari angkatan 2006-2010. Ada beberapa data yang akan digunakan untuk proses selanjutnya yaitu data mahasiswa berupa NIM, Nama, IP 1, IP 2, IP 3, IP 4, Lama Studi dan Keterangan.

1.2. Pre-Processing dan Cleaning Data

Pada database SIA Ilmu Komputer ini, semua data sudah relevan, tidak *missing value* dan tidak *redundant* dan sesuai dengan atribut yang ditentukan. Hal ini merupakan syarat awal dalam melakukan data *mining*. Dalam database ini, terdapat dataset yang berjumlah 148 data.

1.3. Transformation

Pada tahap ini perubahan pada data dilakukan agar lebih mudah untuk proses *mining*. Perubahan awal yang dilakukan adalah mengambil data nilai-nilai matakuliah setiap mahasiswa, dari nilai-nilai matakuliah mahasiswa didapatkan nilai IP Semester setiap tahun ajaran semesternya. Setelah mendapatkan data berupa nilai IP Semester setiap mahasiswa, kemudian perubahan selanjutnya yaitu menampilkan nilai matakuliah Tugas Akhir setiap mahasiswa dengan syarat terdapat nilai Tugas Akhirnya, untuk mengetahui status lulus dari mahasiswa.

Setelah terjadi beberapa perubahan data maka didapatkan hasil akhir query yang nantinya akan dilanjutkan pada proses *mining*. Sehingga terbentuk kumpulan variabel-variabel seperti NIM, Nama, IP1, IP2, IP3, IP4, Lama Studi dan Keterangan. Berikut tampilan dataset mahasiswa.

NIM	Nama	IP1	IP2	IP3	IP4	Lama	Keterangan
J1F106201	Nadya Rosana	2.39	2.61	2.74	3	5	Tepat Waktu
J1F106202	M. Fazeriansyah	2.77	2.42	0	2.86	6	Tidak Tepat Waktu
J1F106206	Mufidah Nur Phasya	2.98	3.63	3.39	3.59	5	Tepat Waktu
J1F106208	Setyo Wahyu Saputro	2.75	2.24	3	3.35	5	Tepat Waktu
J1F106210	Willianto Caniago	3.34	3.39	3.61	3.87	5	Tepat Waktu
J1F106211	Lailatul Qadariah	3.25	3.55	3.63	3.63	5	Tepat Waktu
J1F106215	Rizki Yulianto	2.91	1.84	2.66	2.62	6	Tidak Tepat Waktu
J1F106218	M.Noor Arifin	2.43	1.39	2.39	1.85	6.5	Tidak Tepat Waktu
J1F106218	Sapto Maryanto	2.66	2.34	2.92	3.17	6	Tidak Tepat Waktu
J1F106219	Christanti Sulandari	2.86	3.05	2.89	3.48	5	Tepat Waktu
J1F106223	Marina Rulianti Pardede	2.66	2.71	3.18	3.39	5	Tepat Waktu
J1F106224	Arief Arrachman	2	1.54	2.66	1.85	6	Tidak Tepat Waktu
J1F106225	Erikson Simanungkalit	2.34	1.61	1.89	2	6.5	Tidak Tepat Waktu
J1F106226	Ahmad Fikri Azmi	2.57	2.32	2.94	2.75	5	Tepat Waktu
J1F106228	Intan Sari Indah Sundari	2.89	2.63	3.08	3.43	5	Tepat Waktu
J1F106231	Gusti Puja Rizki Hidayat	2.32	1.55	2.33	2.59	7	Tidak Tepat Waktu
J1F106232	Friska Abadi	2.91	2.47	3.03	3.41	5	Tepat Waktu
J1F106236	Rizki Amalia	3.64	3.42	3.18	4	5	Tepat Waktu
J1F106239	Noor Mina Awalil	2.91	2.16	2.97	3.4	5.5	Tidak Tepat Waktu
J1F106241	Hafizh Ramadhanani	2.7	2.84	3.11	3.37	5	Tepat Waktu
J1F106242	Azis Syafi'i Z	2.64	1.97	2.59	2.42	3.5	Tidak Tepat Waktu
J1F106243	Sherly Melati Ramadhan	2.61	2.16	3.38	2.96	5	Tepat Waktu
J1F107001	Aria Marwan Putra	3	2.71	3.19	3.5	5	Tepat Waktu
J1F107003	Rahmat Hidayatullah	2.21	2.22	2.82	3.19	6.5	Tidak Tepat Waktu
J1F107006	Yuniar Rezekiani	2.36	2.03	2.63	2.64	6	Tidak Tepat Waktu
J1F107007	Muhammad Hasmi	2.86	2.26	3.26	3.48	5.5	Tidak Tepat Waktu
J1F107008	Itmi Yunita	2.93	2.55	3.19	3.62	4.5	Tepat Waktu
J1F107011	Dewi Nurwanti	3.07	3.06	3.45	3.12	5	Tepat Waktu
J1F107015	Fikky Hitaparisuda	2.21	2.22	2.71	3.21	4.5	Tepat Waktu
J1F107018	Baron Hidayat	2.71	2.31	2.89	3.12	7.5	Tidak Tepat Waktu

Gambar 3. Dataset Mahasiswa

1.4. Data Mining

Data *mining* merupakan tahapan untuk menemukan pola atau informasi dalam sekumpulan data dengan menggunakan teknik dan algoritma tertentu. Pada penerapan data *mining* ini menggunakan algoritma C4.5 untuk memprediksi masa studi mahasiswa. Dari proses transformasi data sebelumnya diperoleh data dengan beberapa atribut yang nantinya akan *dimining* dan dibagi menjadi data *training* dan data *testing*. Dalam proses *mining* ini setiap nilai IP Semester dari setiap mahasiswa akan dibuat kategori berdasarkan IP dari mahasiswa agar lebih memudahkan dalam memproses atau membentuk pohon keputusan berdasarkan data yang telah ada. Berikut kategori IP.

Tabel 2. Kategori IP

Kategori	Keterangan	Range IP
A	Sangat Baik	3 – 4
B	Baik	2 – 2,99
C	Cukup	1 – 1,99
D	Kurang	0 – 0,99

Dari proses sebelumnya format dataset berbentuk seperti ini.

Tabel 3. Dataset Mahasiswa

Nama	IP1	IP2	IP3	IP4	Lama a Studi	Keterangan
Nadya Rosana	2.2	1.53	2.17	1.88	5	Tepat Waktu
M.Fazeriansyah	3	3.56	2.18	2.32	6	Tidak Tepat Waktu
Mufidah	2.63	3	4	4	5	Tepat Waktu

...
...
Riski Yulianto	2	1.53	1.84	2	6	Tidak Waktu	Tepat

Setelah diubah menjadi bentuk kategori berdasarkan IP Semester kemudian format data menjadi seperti ini.

Tabel 4. Dataset Mahasiswa berkategori

Nama	IP1	IP2	IP3	IP4	Lama Studi	Keterangan
Nadya Rosana	B	C	C	D	5	Tepat Waktu
M.Fazeriansyah	A	A	C	C	6	Tidak Waktu Tepat
Mufidah	B	A	A	A	5	Tepat Waktu
...
...
Riski Yulianto	B	C	C	C	6	Tidak Waktu Tepat

Dataset berjumlah 148 akan dibagi menjadi dua yaitu data *training* berjumlah 110 data sedangkan data *testing* berjumlah 30 data. Tidak ada penentuan untuk membuat jumlah dari data *training* dan data *testing*, apabila jumlah data *training* semakin banyak maka akan semakin bagus dalam menghasilkan pengetahuan. Tahap pertama yang dilakukan adalah memproses data *training* dengan menentukan nilai *entropy* dari total data. Kemudian menentukan dan memilih nilai *gain* yang tertinggi dari semua atribut. Berikut rumus untuk perhitungan *gain* dan *entropy*.

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

Dengan :

S : Himpunan Kasus

A : Atribut

n : Jumlah partisi atribut A

|S_i| : Jumlah kasus pada partisi ke-i

|S| : Jumlah kasus dalam S

$$Entropy(S) = \sum_{i=1}^n - p_i * \log_2 p_i \quad (2)$$

Dengan :

S : Himpunan Kasus

A : Atribut

n : Jumlah partisi atribut S

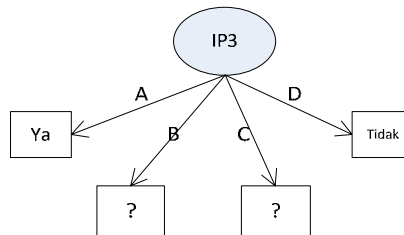
p₁ : Proporsi dari S₁ terhadap S

Node 1. Dari 110 data terdapat 87 data yang menyatakan mahasiswa tepat waktu dan 23 data yang menyatakan mahasiswa tidak tepat waktu.

Tabel 5. Perhitungan Node 1

	Jumlah	Ya	Tidak	Entropy	Gain
	110	87	23	0.739741449	
IP1-A	41	40	1	0.16542734	1.234329
IP1-B	67	45	22	0.913240161	
IP1-C	2	2	0	0	
IP1-D	0	0	0	0	
IP2-A	71	66	5	0.367498495	0.844076
IP2-B	35	19	16	0.994693795	
IP2-C	3	2	1	0.918295834	
IP2-D	1	0	1	0	
IP3-A	25	25	0	0	1.305584
IP3-B	26	12	14	0.995727452	
IP3-C	58	50	8	0.616166193	
IP3-D	1	0	1	0	
IP4-A	77	68	9	0.520334616	0.64687
IP4-B	31	19	12	0.962900415	
IP4-C	2	0	2	0	
IP4-D	0	0	0	0	

Pada perhitungan node 1 didapatkan hasil nilai *entropy* dan *gain* dari masing-masing IP. Nilai *gain* tertinggi dihasilkan oleh IP3, dan dari perhitungan IP3 didapatkan hasil A mendapatkan jumlah Ya sebesar 25, dan D mendapatkan jumlah Tidak bernilai 1, jadi keluaran dari A = Ya dan D = Tidak. Sedangkan pada IP3 B dan C masih sama-sama berisi maka dilakukan perhitungan kembali sampai perhitungan akhir. Berikut akar yang dihasilkan dari nilai *gain* tertinggi.



Gambar 4. Akar pertama yang dihasilkan dari nilai gain tertinggi pada Node 1

Node 2. Perhitungan IP3 B dan C

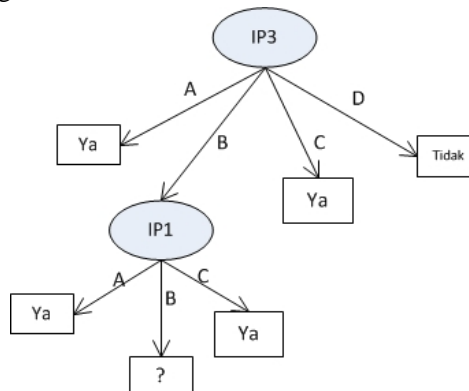
Table 6. Perhitungan Node 2 IP3-B

IP3-B		885	62	A23	0.995727452	
	IP1-A	22	21	1	0.266764988	1.603586
	IP1-B	61	39	22	0.943226155	
	IP1-C	2	2	0	0	
	IP2-A	47	42	5	0.488908591	1.156801
	IP2-B	34	18	16	0.997502546	
	IP2-C	3	2	1	0.918295834	
	IP2-D	1	0	1	0	
	IP4-A	52	43	9	0.664695305	0.940266
	IP4-B	31	19	12	0.962900415	
	IP4-C	2	0	2	0	

Tabel 7. Perhitungan Node 2 IP3-C

IP3-C		85	62	23	0.616166	
	IP1-A	22	21	1	0.266765	1.224025
	IP1-B	61	39	22	0.943226	
	IP1-C	2	2	0	0	
	IP2-A	47	42	5	0.488909	0.77724
	IP2-B	34	18	16	0.997503	
	IP2-C	3	2	1	0.918296	
	IP2-D	1	0	1	0	
	IP4-A	52	43	9	0.664695	0.560705
	IP4-B	31	19	12	0.9629	
	IP4-C	2	0	2	0	

Pada perhitungan diatas IP3-B dan IP3-C sama-sama menghasilkan IP1 sebagai nilai *gain* tertinggi. Untuk mendapatkan hasil dari IP3 yang belum ditentukan, maka antara IP3-B dan IP3-C dipilih nilai entropy yang paling besar. Nilai *gain* tertinggi dihasilkan dari IP1, yang diantaranya A = Ya, C = Ya dan B akan dihitung kembali. Berikut hasil perhitungan dari node2.



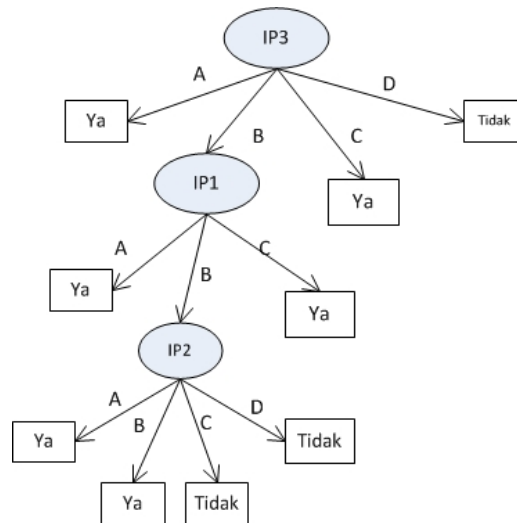
Gambar 5. Akar yang dihasilkan node 2

Node 3. Perhitungan IP B

Tabel 8. Perhitungan Node 3 IP-B

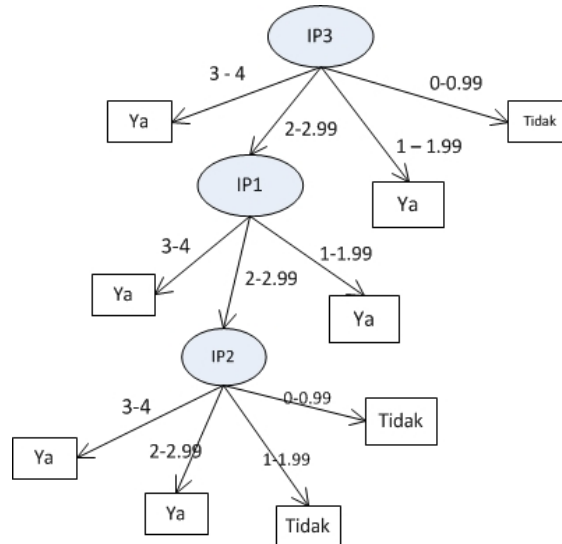
IP1-B		61	39	22	0.943226155	
	IP2-A	30	26	4	0.609511907	1.120208
	IP2-B	27	18	9	0.975119065	
	IP2-C	3	0	3	0.918295834	
	IP2-D	1	0	1	0	
	IP4-A	35	27	8	0.775512658	0.891702
	IP4-B	4	12	12	1	
	IP4-C	2	0	2	0	

Nilai *gain* tertinggi dihasilkan IP2 dengan keluaran A = Ya, B = Ya, C = Tidak dan D = Tidak. Karena tidak ada akar yang dihasilkan lagi, maka perhitungan sudah selesai. Berikut pohon keputusan (*tree*) yang sudah terbentuk.



Gambar 6. Pohon Keputusan yang telah terbentuk

Berikut tampilan pohon keputusan beserta aturannya dalam bentuk range IP yang sudah ditentukan sebelumnya.



Gambar 7. Pohon Keputusan yang dihasilkan oleh proses *mining* menggunakan data *training*

Aturan dari pohon keputusan (*tree*) yang sudah terbentuk.

- Jika IP3 = 3-4 maka Ya (Tepat Waktu)
- Jika IP3 = 2-2.99 dan IP1 = 3-4 maka Ya (Tepat Waktu)
- Jika IP3 = 2-2.99 dan IP1 = 2-2.99 dan IP2 = 3-4 maka Ya (Tepat Waktu)
- Jika IP3 = 2-2.99 dan IP1 = 2-2.99 dan IP2 = 2-2.99 maka Ya (Tepat Waktu)
- Jika IP3 = 2-2.99 dan IP1 = 2-2.99 dan IP2 = 1-1.99 maka Tidak (Tidak Tepat Waktu)
- Jika IP3 = 2-2.99 dan IP1 = 2-2.99 dan IP2 = 0-0.99 maka Tidak (Tidak Tepat Waktu)
- Jika IP3 = 2-2.99 dan IP1 = 0-0.99 maka Ya (Tepat Waktu)
- Jika IP3 = 1-0.99 maka Ya (Tepat Waktu)
- Jika IP3 = 0-0.99 maka Tidak (Tidak Tepat Waktu)

1.5. Evaluation

Pohon keputusan yang telah terbentuk kemudian diuji dengan menggunakan data *training* dan data *testing*. Pada pengujian data *training* yang berjumlah 110 data menggunakan aturan dari pohon keputusan yang telah terbentuk, data yang diprediksi yang sesuai dengan data real berjumlah 92 data atau sebesar 83,64 %. Sedangkan data yang tidak sesuai berjumlah 18 data atau sebesar 16,36 %. Hal ini sesuai dengan nilai akurasi yang dihasilkan oleh program. Pada program nilai statistik *kappa* yang dihasilkan sebesar 0,4608 yang berarti penilaian kuantitatifnya sedang.

Dalam hal pengujian data terhadap data *training* dikatakan cukup berhasil karena persentasi yang dihasilkan mencapai 80%. Hal ini dikarenakan data yang digunakan untuk diuji adalah data yang diproses untuk pembentukan pohon keputusan yaitu data *training* itu sendiri.

Pengujian selanjutnya dilakukan pada data *testing* yang berjumlah 30 data. Jumlah data yang sesuai adalah 22 data atau sebesar 73,33 % dan data yang tidak sesuai adalah 8 data atau sebesar 26,67 %. Nilai statistik *kappa* yang dihasilkan sebesar 0.0400 yang berarti pengukuran secara kuantitatifnya sedikit berdasarkan kriteria pada nilai *kappa*. Pada pengujian data *testing* persentase yang dihasilkan lebih rendah daripada data *training*, hal ini dikarenakan data *testing*

adalah data real yang diprediksi menggunakan pohon keputusan yang telah terbentuk dan diproses dari data training.

Berdasarkan akurasi yang telah dihasilkan oleh data training dan data testing tidak mencapai persentase 100 %, hal ini dikarenakan ada beberapa faktor yang sangat mempengaruhi dalam proses *mining*. Beberapa faktornya diantara lain :

- a. Jumlah data yang digunakan untuk proses *mining*
- b. Varian data yang kurang banyak, semakin data itu bervariasi semakin banyak menghasilkan pengetahuan

4. KESIMPULAN

Kesimpulan yang diperoleh dari penelitian ini adalah:

1. Terdapat beberapa langkah dalam memprediksi masa studi mahasiswa tepat waktu.
 - a. Selection Data
Data yang dipilih adalah database SIA Ilmu Komputer FMIPA Universitas Lambung Mangkurat.
 - b. Cleaning Data
Pada tahap ini semua data sudah benar dan sesuai dengan atribut yang ditentukan untuk proses selanjutnya.
 - c. Transformation Data
Ada dua belas (12) perubahan pada data untuk terbentuk menjadi format data *training* dan data *testing* untuk dilakukan proses *mining*.
 - d. Data Mining
Pada tahapan ini terdapat tiga (3) kali iterasi yang menghasilkan pohon keputusan (*tree*) untuk dijadikan dasar aturan untuk memprediksi masa studi mahasiswa.
 - e. Evaluation
Pengujian dilakukan pada data training dan data testing. Berdasarkan pengujian yang telah dilakukan persentasi yang dihasilkan oleh data training sebesar 80% dengan jumlah 110 data dan data testing sebesar 73% dengan jumlah 30 data.
2. Penerapan data mining dengan teknik klasifikasi menggunakan algoritma C4.5 yang dilakukan menghasilkan sebuah informasi dalam memprediksi masa studi tepat waktu mahasiswa di program studi Ilmu Komputer FMIPA UNLAM, dengan tingkat akurasi yang dihasilkan sebesar 73,33% dengan jumlah 30 data *testing* dari 140 data *set*.

5. SARAN

Penerapan data mining pada penelitian ini yang menggunakan algoritma C4.5 untuk fungsi klasifikasi dapat dilakukan perbandingan dengan algoritma lainnya untuk kasus yang sama. Sehingga dapat diketahui algoritma yang mana yang lebih baik untuk diterapkan pada bidang akademik.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Ikhtison Mekongga yang telah memberi dukungan terhadap penelitian ini.

DAFTAR PUSTAKA

- [1] Larose, D.T, 2005. *Discovering Knowledge in Data: An Introduction to Data mining*. John Willey & Sons, Inc.
- [2] Larose, D.T, 2005. *Data Mining Method and Models*”, Inc Publication.
- [3] Retma Ramadina, 2015. *Rancang Bangun Aplikasi Data Mining untuk Prediksi Masa Studi Mahasiswa Tepat Waktu dengan Algoritma C4.5*, Program Studi Ilmu Komputer, Universitas Lambung Mangkurat, Banjarbaru.
- [4] Wu, X. 2008. “*Top 10 Algorithms in Data Mining*”, Department of Computer Science, University of Vermont Burlington, VT, USA.