KLASIFIKASI LAMA STUDI MAHASISWA MENGGUNAKAN METODE ALGORITMA C5.0 PADA STUDI KASUS DATA KELULUSAN MAHASISWA FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM UNIVERSITAS MULAWARMAN TAHUN 2017

Daniel Dalbergio[1*], Memi Nor Hayati[1], Yuki Novia Nasution[1]

[1]Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Mulawarman, Indonesia

Corresponding author: daniel.dalbergio1996@gmail.com

Abstrak. Klasifikasi adalah suatu teknik pembentukan model dari data yang belum terklasifikasi untuk digunakan mengklasifikasikan data baru. Klasifikasi merupakan pengelompokan sampel berdasarkan ciri-ciri persamaan dengan menggunakan variabel target sebagai kategori. Decision tree algoritma C5.0 merupakan implementasi dari Algoritma C4.5 yang memproduksi pohon keputusan. Tujuan dalam penelitian ini adalah mengetahui hasil klasifikasi dan ketepatan klasifikasi pada metode algoritma C5.0 dan pohon keputusan yang dibentuk menggunakan Algoritma C5.0 untuk mengklasifikasi Masa Studi mahasiswa yang lulus pada tahun 2017. Data yang digunakan adalah data masa studi mahasiswa Fakultas Matematika dan Ilmu Pengetahuan Alam Samarinda Tahun 2017 dan digunakan 4 Variabel yaitu jenis kelamin, program studi, indeks prestasi kumulatif (IPK) dan asal daerah untuk memprediksi masa studi. Laju error yang dihitung menggunakan nilai Apparent Error Rate (APER) dipergunakan sebagai evaluator pada metode Algoritma C5.0 pada klasifikasi Masa Studi. Kesalahan klasifikasi yang dihasilkan adalah 15,79 %. Hal ini menunjukkan bahwa dari 19 orang, terdapat 16 orang yang tepat diklasifikasikan.

Kata Kunci: Algoritma C5.0, APER, *Decision Tree,* Klasifikasi, Masa Studi.

1 PENDAHULUAN

Data adalah catatan atas kumpulan fakta. Data dapat diperoleh, disimpan, diolah, dipakai dan sebagainya. Salah satu bentuk pengolahan suatu data yaitu *data mining*. *Data mining* suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam *database*. Data mining juga merupakan proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk menguraikan, mengidentifikasi informasi yang bermanfaat, dan pengetahuan yang terkait dari berbagai database besar [1].

Untuk memanfaatkan data yang tersebar sangat banyak di dunia digital ini, diperlukan suatu alat untuk mengolah dokumen-dokumen sehingga informasinya dapat terserap dan tersajikan dengan baik. Salah satu bentuk dari pengolahan suatu data yaitu *data mining*. *Data mining* adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam *database*. Data mining juga merupakan proses yang menggunakan teknik statistik, matematika, kecerdasan buatan dan *machine learning* untuk menguraikan dan mengidentifikasi informasi yang bermanfaat [2].

Lulus tepat waktu dengan masa studi di bawah 5 tahun tentunya menjadi harapan setiap mahasiswa. Lama studi merupakan bagian dari penilaian evaluasi sebuah program studi, yang
dalam proses evaluasinya dapat dilakukan lebih awal dengan mengklasifikasi data historis mahasiswa lulusan untuk memprediksi lama studi mahasiswa aktif yang informasinya belum diketahui sebelumnya, sehingga program studi dapat mengidentifikasi secara dini karakteristik mahasiswa yang berpotensi lulus dengan kategori tepat waktu.

Masa studi adalah salah satu tolak ukur keberhasilan dalam studi seorang mahasiswa. Semakin cepat lama studi mahasiswa dalam menempuh perkuliahan, dapat mengindikasikan bahwa mahasiswa tersebut rajin dan cerdas. Salah satu kewajiban universitas adalah mengontrol lama studi yang ditempuh mahasiswa yaitu maksimal 7 tahun akademik untuk program sarjana dan program diploma 4 atau sarjana terapan berdasarkan Standar Nasional Pendidikan Tinggi Nomor 44 Tahun 2015, sedangkan dalam peraturan Rektor Universitas Mulawarman Nomor 12 Tahun 2017 menyatakan bahwa masa studi mahasiswa program Strata satu (S1) yaitu antara 8 sampai dengan 14 semester.

2 TUJUAN PENELITIAN

Tujuan penelitian ini yaitu:

a. Untuk mengetahui hasil klasifikasi pada masa studi kelulusan seluruh mahasiswa Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Mulawarman tahun ajaran 2017.
b. Untuk mengetahui ketepatan hasil klasifikasi metode algoritma C5.0 menggunakan APER.

3 METODOLOGI PENELITIAN

Pada penelitian ini, analisis data dilakukan dengan langkah-langkah berikut:

1. Analisis Statistika Deskriptif

   Penyajian data dilakukan dengan membuat tabulasi penyajian data dalam bentuk diagram batang.

2. Membagi data training (data latih) dan data testing (data uji)

   Sebelum melakukan proses klasifikasi dengan algoritma C5.0, langkah pertama yang dibuat adalah membagi data training dan data testing, kemudian dilakukan randomisasi terlebih dahulu agar setiap data memiliki kesempatan yang sama untuk menjadi data training dan data testing. Pada metode ini digunakan proporsi data training dan data testing 90:10.

3. Analisis Algoritma C5.0

   C5.0 menghasilkan tree dengan jumlah cabang per node bervariasi. Algoritma C5.0 merupakan pengembangan dari algoritma C4.5. Strategi pengembangan decision tree dengan menggunakan Algoritma C5.0 adalah sebagai berikut [3]:

   1. Penentuan Atribut Akar

      Langkah pertama untuk membangun pohon keputusan yaitu pemilihan atribut akar, dengan menghitung nilai *gain* dengan meggunakan:

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

      dengan:

      S : Himpunan kasus

      S_i : Himpunan kasus pada partisi ke-I

      A : Atribut

      m : Jumlah partisi

      $|S_i|$: Jumlah kasus pada partisi ke-I

      $|S|$: Jumlah kasus dalam S

      Sebelum mendapatkan nilai *gain* dicari terlebih dahulu nilai *entropy* menggunakan:

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

      dengan :

      S : Himpunan Kasus

      p_i : Proporsi dari S_i terhadap S

   2. Penentuan Cabang

      Menentukan cabang masing-masing atribut dengan cara yang sama seperti mencari atribut akar. Untuk menentukan cabang pada metode C5.0 secara manual peneliti menggunakan *software Microsoft Excel* 2010.

   3. Penentuan pohon klasifikasi maksimal

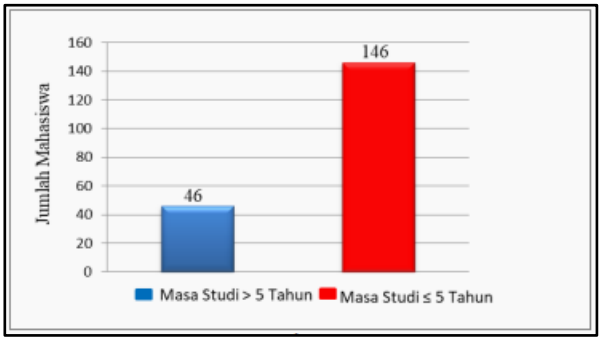
      Kelas dibagi dalam cabang dan apabila cabang mempunyai dua kelas maka yang dipilih kelas yang terbanyak dan proses diulang untuk masing-

masing cabang sampai semua kelas pada cabang memiliki kelasnya masing-masing.

4 HASIL DAN PEMBAHASAN

4.1 Statistika Deskriptif

Tahap awal yang dilakukan dalam penelitian ini adalah analisis deskriptif, bertujuan untuk menggambarkan karakteristik data kelulusan seluruh mahasiswa Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Mulawarman (FMIPA UNMUL) Tahun 2017. Karakteristik yang digambarkan pada analisis deskriptif adalah masa studi (Y), jenis kelamin (X_1), program studi (X_2), IPK (X_3) dan asal daerah (X_4). Berdasarkan gambar 1, mahasiswa yang lulus dengan masa studi kurang dari sama dengan 5 tahun ada 146 mahasiswa dan mahasiswa yang lulus dengan masa studi lebih dari 5 tahun ada 44 mahasiswa dari total sebanyak 190 mahasiswa lulusan tahun 2017 FMIPA UNMUL.



Gambar 1: Diagram Batang Variabel Masa Studi

4.2 Klasifikasi Decision Tree Algoritma C5.0

Pemilihan node akar adalah tahap pertama dalam pembentukan pohon klasifikasi dengan menggunakan Persamaan (1) dan Persamaan (2).

a. Perhitunga *entropy* total menggunakan Persamaan (1), diperoleh nilai *entropy* total adalah sebesar 0.7048.
b. Perhitungan entropy variabel jenis kelamin menggunakan persamaan (2), diperoleh perhitungan *gain* diatas dapat diketahui bahwa *gain* jenis kelamin adalah sebesar 0,0105.

Berdasarkan Tabel 1 menunjukkan bahwa *gain* tertinggi ada di variabel IPK sehingga IPK dijadikan sebagai *node* akar (*node* 1). Kemudian diperoleh cabang untuk *node* akar ada tiga, yaitu IPK < 3, 3,00 ≤ IPK < 3,50, IPK ≥ 3,50.

Berdasarkan Tabel 2 menunjukkan bahwa *gain* tertinggi ada di variabel program studi sehingga program studi dijadikan sebagai *node* 2. Maka cabang untuk *node* 4 ada empat, yaitu statistika, fisika, kimia, biologi.

Tabel 1: Hasil Perhitungan *Gain* dan *Entropy* untuk *Node* Akar Masa Studi

Node	Variabel		Jumlah Kasus (S)	≤ 5 Tahun	> 5 Tahun	*Entropy*	*Gain*
1	Total		171	131	40	0,7848	
	Jenis Kelamin	Perempuan	112	90	22	0,7147	0,0105
		Laki-laki	59	41	18	0,8874	
	Program Studi	Statistika	36	28	8	0,7642	0,0121
		Fisika	24	18	6	0,8113	
		Kimia	76	61	15	0,6944	
		Biologi	35	24	12	0,9183	
	IPK	**IPK < 3**	**17**	**6**	**11**	**0,9367**	**0,0798**
		3,00 ≤ IPK < 3,50	**111**	**86**	**25**	**0,7696**	
		IPK ≥ 3,50	**43**	**39**	**4**	**0,4465**	
	Asal Daerah	Samarinda	48	40	8	0,6500	0,0075
		Luar Samarinda	123	91	32	0,8270	

Tabel 2: Hasil Perhitungan *Gain* dan *Entropy* untuk *Node* Akar Masa Studi

Node	Variabel		Jumlah Kasus (S)	≤ 5 Tahun	> 5 Tahun	*Entropy*	*Gain*
2	Total		17	6	11	0,9367	
	Jenis Kelamin	Perempuan	12	3	9	0,8113	0,3356
		Laki-Laki	5	3	2	0,9710	
	Program Studi	**Statistika**	**4**	**2**	**2**	**1**	**0,5105**
		Fisika	**1**	**1**	**0**	**0**	
		Kimia	**4**	**3**	**1**	**0,8113**	
		Biologi	**8**	**0**	**8**	**0**	
	Asal Daerah	Samarinda	3	2	1	0,9183	0,0638
		Luar Samarinda	14	4	10	0,8631	

Tabel 3: Hasil Perhitungan *Gain* dan *Entropy* untuk *Node* Akar Masa Studi

Node	Variabel		Jumlah Kasus (S)	≤ 5 Tahun	> 5 Tahun	*Entropy*	*Gain*
3	Total		111	87	25	0,7660	
	Jenis Kelamin	**Perempuan**	**68**	**58**	**10**	**0,6024**	**0,1808**
		Laki-laki	**43**	**29**	**14**	**0,9257**	
	Program Studi	Statistika	22	16	6	0,8454	0,0215
		Fisika	22	17	5	0,7732	
		Kimia	49	39	10	0,7300	
		Biologi	18	15	3	0,7425	
	Asal Daerah	Samarinda	32	26	6	0,6962	0,0022
		Luar Samarinda	79	61	18	0,7909	

Berdasarkan Tabel 3 menunjukkan bahwa *gain* tertinggi ada di variabel Jenis Kelamin sehingga Jenis Kelamin dijadikan sebagai *node* 3. Kemudian diperoleh cabang untuk *node* 3 ada dua, yaitu laki-laki dan perempuan.

4.3 Ketepatan Klasifikasi Algoritma C5.0

Diperoleh tabel matriks konfusi hasil klasifikasi C5.0 menggunakan *data testing* sebagai berikut:

Tabel 4: Matriks Konfusi Klasifikasi C5.0

Kelas	Kelas Prediksi		Total
	> 5 tahun	≤ 5 tahun	
> 5 Tahun	1	2*	3
≤ 5 tahun	1*	15	16
Total	2	17	19

Pada Tabel 4, diberikan objek yang tepat diklasifikasikan dan yang salah diklasifikasikan untuk masing-masing kelompok. Tanda (*) pada angka-angka di Tabel 5 menyatakan jumlah obyek kelompok tertentu salah diklasifikasikan oleh metode ini. Berdasarkan Tabel 5 diketahui dengan menggunakan teknik metode Algoritma C5.0, bahwa:

a. Dari 3 orang yan diklasifikasikan ke dalam masa studi dengan kategori masa studi > 5 tahun diperoleh 1 orang tepat diklasifikasikan ke dalam kategori masa studi > 5 tahun dan 2 orang tidak tepat diklasifikasikan ke dalam kategori masa studi > 5 tahun.
b. Dari 16 orang yang diklasifikasikan ke dalam masa studi dengan kategori masa studi ≤ 5 tahun diperoleh 1 orang tidak tepat diklasifikasikan ke dalam kategori masa studi $\leq$ 5 tahun dan 15 orang tepat diklasifikasikan ke dalam kategori masa studi $\leq$ 5 tahun.

Setelah diketahui klasifikasi untuk tiap-tiap kelompok maka laju *error* hasil klasifikasi secara total dapat diketahui dengan menghitung nilai APER, maka diperoleh nilai APER sebagai berikut:

$$APER = \frac{1+2}{1+2+1+15} = \frac{3}{19} = 15{,}79\%$$

Tabel 5: Hasil Perhitungan Klasifikasi C5.0

	Persentase
APER	15,79%
Ketepatan Klasifikasi	84,21%

Laju *error* dihitung menggunakan nilai APER dipergunakan sebagai evaluaor pada metode Algoritma C5.0 pada klasifikasi Masa Studi. Kesalahan klasifikasi yang dihasilkan adalah 15,79%. Hal ini menunjukkan bahwa dari 190 orang, terdapat 19 orang yang diklasifikasikan. Sedangkan nilai ketepatan klasifikasi yang dihasilkan adalah sebesar 84,21% hal ini menunjukkan bahwa dari 19 orang terdapat 16 orang yag tepat diklasifikasikan.

5 KESIMPULAN

Berdasarkan hasil analisis dan pembahasan yang dilakukan, diperoleh kesimpulan sebagai berikut:

1. Metode algoritma C5.0 dengan menggunakan proporsi *data training* dan *testing sebesar* 90:10, diperoleh bahwa dari 19 objek pengamatan terdapat bahwa 3 pengamatan yang tidak tepat diklasifikasikan dan 16 pengamatan yang tepat diklasifikasikan.
2. Hasil ketepatan klasifikasi pada metode algoritma C5.0, diperoleh nilai persentase ketepatan klasifikasi metode algoritma C5.0 sebesar 84,21%. Kemudian kesalahan klasifikasi (APER) yang dihasilkan adalah sebesar 15,79%. Dengan demikian metode algoritma C5.0 merupakan metode yang baik dalam pengklasifikasian data masa studi kelulusan seluruh mahasiswa FMIPA UNMUL tahun 2017.

DAFTAR PUSTAKA

[1] Efraim, T., Aronson, J.E., & Liang, T.P. (2005). *Decision Support System and Intelligent Systems. Yogyakarta*: ANDI.

[2] Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining.* New Jersey: John Wiley & Sons.

[3] Yusuf W, Y. (2007). Perbandingan Performasi Algoritma Decision Tree C5.0, CART dan CHAID: Kasus Prediksi Status Resiko Kredit di Bank X. *Seminar Nasional Aplikasi Teknologi Informasi.*