

PREDIKSI KELULUSAN TEPAT WAKTU MAHASISWA MENGGUNAKAN ALGORITMA C4.5 PADA STMIK DHARMA WACANA

¹Muhammad Wisnu Mukti, ²Tri Aristi Saputri

¹Teknik Informatika STMIK Dharma Wacana,

²STMIK Dharma Wacana

¹wisnumukti.wm3@gmail.com, ²aristy@dharmawacana.ac.id

ABSTRAK

Penilaian akreditasi perguruan tinggi salah satu faktornya yaitu kelulusan mahasiswanya. Namun yang menjadi masalah adalah bagaimana cara mengetahui kelulusan mahasiswa tepat waktu dan tidak tepat waktu. Pada penelitian ini diulas bagaimana prediksi mahasiswa lulus tepat waktu dengan metode data mining algoritma C4.5. Data mahasiswa lulus pada penelitian ini digunakan sebagai data training. Tujuan dari penelitian ini yaitu mengetahui ketepatan waktu kelulusan mahasiswa, sehingga dapat memberikan informasi bagi perguruan tinggi untuk membuat kebijakan guna perbaikan dimasa yang akan datang.

Keyword: Data mining, Penilaian.

1 PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi saat ini mempengaruhi dibutuhkannya tingkat akurasi suatu data dalam kehidupan sehari-hari. Namun tingginya kebutuhan akan informasi ini terkadang belum diimbangi dengan memadainya penyajian informasi, sehingga informasi tersebut masih harus digali dan terkadang berasal dari data yang jumlahnya sangat besar. Data dalam suatu sistem informasi dapat dimanfaatkan demi menunjang kegiatan pengambilan keputusan, namun tidak hanya mengandalkan data operasional, diperlukan pula analisis agar potensi informasi dapat digali dari bongkahan data yang ada (Nuqson, 2010) tidak terkecuali data mahasiswa pada sebuah institusi perguruan tinggi.

STMIK Dharma Wacana Metro adalah organisasi pendidikan pada bidang teknologi informasi dan multimedia. STMIK Dharma Wacana Metro terdiri dari jurusan Sistem Informasi dan Teknik Informatika. Ketepatan kelulusan mahasiswa STMIK Dharma Wacana Metro adalah informasi yang sangat penting untuk meningkatkan pelayanan agar mahasiswa merasa nyaman dan dapat lulus tepat waktu sesuai target. Perguruan tinggi perlu mengetahui faktor yang dapat mempengaruhi tingkat ketepatan kelulusan mahasiswa sesuai dengan masa studi yang telah ditetapkan, diantaranya adalah indeks prestasi, usia mahasiswa, dan berbagai faktor lainnya.

Berdasarkan data tersebut, diketahui bahwa setiap tahun ajaran tidak semua mahasiswa dapat menyelesaikan pendidikannya dengan tepat waktu. Hal ini dapat mengakibatkan jumlah mahasiswa masuk tidak sesuai dengan angka kelulusan pada tahun angkatannya. Oleh sebab itu, dibutuhkan suatu sistem prediksi kelulusan mahasiswa yang dapat digunakan untuk memprediksi kemungkinan mahasiswa dapat lulus tepat waktu atau tidak.

Teknik *Data mining* digunakan untuk membantu pencarian informasi berharga pada data mahasiswa untuk mengetahui faktor pendukung dan penghambat kelulusan mahasiswa. *Data mining* memiliki beberapa padanan, yaitu *knowledge discovery* maupun *pattern recognition*. Istilah *knowledge discovery* yaitu penemuan pengetahuan secara tepat yang digunakan sebagai tujuan utama dari *data mining* demi mendapatkan pengetahuan yang masih tersembunyi dalam bongkahan data (Susanto, Suryadi, 2010).

Salah satu metode pengolahan data pada *data mining* yaitu klasifikasi, yang memecah data menjadi beberapa bagian untuk membuat sebuah keputusan. Dengan algoritma C4.5, pola kelulusan mahasiswa dapat dikenali dan kemudian bisa digunakan dalam upaya strategi kegiatan perkuliahan. (Mardi, 2016). Algoritma ini dikenal dan diminati karena memiliki kelebihan, yaitu dapat mengolah data diskret serta numerik, dapat pula menangani kehilangan nilai atribut, serta dapat menghasilkan aturan yang mudah diinterpretasikan. Performa algoritma ini dibandingkan algoritma lain termasuk yang tercepat (Tjahyono, 2010).

Berdasarkan permasalahan tersebut, pemanfaatan algoritma C.45 pada *data mining* dapat dilakukan untuk membantu menemukan informasi prediksi kelulusan mahasiswa STMIK Dharma Wacana Metro. Dengan demikian, proses tersebut akan membantu program studi dalam menentukan prediksi status kelulusan mahasiswa serta menentukan kebijakan dalam mencapai target ketepatan kelulusan mahasiswa.

Masalah pada penelitian ini dirumuskan sebagai berikut : Klasifikasi prediksi tingkat kelulusan mahasiswa STMIK Dharma Wacana Metro menggunakan algoritma C4.5.

Penelitian ini memiliki tujuan sebagai berikut : Mendapatkan model yang terbentuk dari klasifikasi kelulusan mahasiswa STMIK Dharma Wacana Metro menggunakan algoritma C4.5.

2 LITERATUR REVIEW

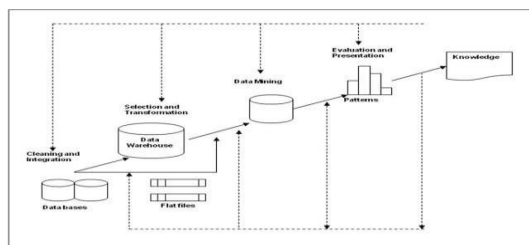
2.1 Pengertian Data Mining

Disiplin ilmu yang muncul dan mengacu pada berbagai macam metode yang secara otomatis, menjelajahi, menganalisa, dan memodelkan repository data besar dalam upaya mengidentifikasi pola baru, valid, berguna serta dapat dipahami merupakan definisi dari *data mining*. Penambahan data melibatkan menyimpulkan algoritma yang mengeksplorasi data untuk membuat dan mengembangkan model yang menyediakan kerangka kerja untuk menemukan dalam data pola yang sebelumnya tidak diketahui untuk analisis dan prediksi (Haim Dahan, 2013). *Data mining* termasuk salah satu proses inti dalam *Knowledge Data Discovery* (KDD). Hal ini disebabkan oleh sebagian besar pekerjaan pada KDD yang berfokus pada *data mining* (Selvia Lorena Br Ginting, 2014). Dari beberapa disiplin ilmu yang digabungkan ini, tujuan *data mining* untuk memperbarui teknik tradisional supaya dapat mengatasi :

- Data yang jumlahnya sangat besar.
- Dimensi pada data yang tinggi.
- Perbedaan sifat setiap data.

2.2 Tahapan Data Mining

Dalam satu rangkaian kegiatan, *data mining* dibagi pada beberapa *fase* yang diilustrasikan pada Gambar 1. *Fase* tersebut bersifat interaktif, dimana pengguna secara langsung terlibat atau melalui perantara *knowledge base*.



Gambar 1. Tahapan data mining

Tahapan *data mining* yaitu :

1. Data cleaning (Pembersihan data)

Pembersihan data yaitu proses menghapus noise, data yang tidak relevan maupun konsisten. Pada umumnya, perolehan data yang berasal dari *database* instansi ataupun hasil pengalaman, mempunyai *input* kurang lengkap, yaitu hilangnya data, kesalahan *entri*, dan data *error*.

2. Data integration (Integrasi data)

Integrasi data adalah gabungan data dari *database* yang berbeda pada database baru. Kegiatan integrasi data dengan atribut yang mengidentifikasi entitas unik seperti nama, nomor pelanggan, jenis produk dan lain-lain.

3. data selection (Seleksi data)

Tidak semua data dalam *database* digunakan, analisa hanya dilakukan pada data yang sesuai. Oleh karena itu perlu dilakukan seleksi data dalam *database*.

4. data transformation (Transformasi data)

Beberapa metode pada *data mining* membutuhkan format data secara khusus untuk dapat diaplikasikan. Data disesuaikan serta digabungkan pada format untuk dapat diproses pada *data mining*.

5. Proses mining

Merupakan tahapan utama untuk menemukan pola pengetahuan berharga yang tersembunyi pada suatu *dataset*.

6. Evaluasi pola (pattern evaluation)

Untuk mengidentifikasi pola yang telah terbentuk pada *knowledge base* yang ditemukan. Kemudian pola tersebut dievaluasi dan dinilai apakah hipotesa telah tercapai.

7. Presentasi pengetahuan (knowledge presentation)

Merupakan visualisasi serta penyajian pengetahuan terhadap metode yang dipakai dalam memperoleh pengetahuan. Tahap terakhir yaitu bagaimana membuat formulasi keputusan yang diperoleh dari hasil analisa.

2.3 Algoritma C4.5

Secara umum, sebuah keputusan yang berasal dari algoritma C4.5 adalah sebagai berikut :

- Menghitung jumlah data dengan syarat tertentu berdasarkan anggota atribut hasil.
- Memilih atribut untuk dijadikan *node*.
- Bentuk cabang pada tiap anggota *node*.
- Pastikan apakah ada nilai *entropy* anggota *node* yang memiliki nilai nol. Apabila ada, kemudian menentukan bentuk daun. Apabila semua *entropy* dari anggota *node* merupakan nol, proses kemudian dihentikan.
- Apabila anggota dari *node* terdapat nilai *entropy* yang lebih besar dari pada nol, maka lakukan proses dari pertama dengan *node* yang menjadi syarat hingga seluruh anggota dari *node* memiliki nilai nol.

Node merupakan atribut yang memiliki nilai *gain* tertinggi dari semua atribut. Rumus menghitung nilai *gain* dinyatakan dalam persamaan berikut :

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

Keterangan :
S : Himpunan kasus
A : Atribut
n : Jumlah partisi pada atribut
 $|S_i|$: Jumlah kasus yang terdapat di partisi ke-*i*
 $|S|$: Jumlah kasus pada *S*

Sementara untuk menghitung nilai *Entropy* adalah berikut ini :

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

Keterangan : S : Himpunan kasus

N : Jumlah partisi pada S

P_i : Jumlah kasus dalam partisi ke- i

3 METODOLOGI

3.1 Metode Pengumpulan Data

Metode pada proses pengumpulan data adalah sebagai berikut :

1. Metode Pengamatan secara Langsung
Melakukan pengamatan secara langsung pada bagian akademik STMIK Dharma Wacana Metro guna memperoleh data yang dibutuhkan.
2. Metode Wawancara
Melakukan wawancara terhadap pihak terkait pada bagian akademik supaya mendapatkan gambaran dan penjelasan secara terperinci.
3. Metode Studi Pustaka
Sumber yang menjadi rujukan untuk sumber data maupun literature adalah buku, jurnal dan lain-lain.

3.2 Penentuan Atribut

Prediksi kelulusan mahasiswa menggunakan atribut sebagai berikut :

- a. Jenis Kelamin

Atribut ini dibagi dua jenis yakni laki-laki (L) dan perempuan (P).

- b. Usia

Atribut usia dibagi dua bagian, kurang dari atau sama dengan 22 tahun dan lebih dari atau sama dengan 23 tahun.

- c. Indeks Prestasi Semester (1-4)

Atribut indeks prestasi semester adalah nilai yang didapat dari sistem kredit semester (sks) dari mata kuliah yang telah ditempuh pada satu semester .

- d. Keterangan

Atribut keterangan yaitu data yang digunakan dalam menentukan hasil keputusan. Pada atribut keterangan hanya terdapat dua nilai yaitu tidak tepat waktu serta tepat waktu.

4 HASIL DAN PEMBAHASAN

Data mahasiswa diambil dari mahasiswa alumni yang telah dinyatakan lulus untuk memprediksi kelulusan tepat waktu. Dari data mentah yang awalnya 148 mahasiswa menjadi 57 mahasiswa karena terdapat langkah seleksi data. Terdapat 6 atribut dengan keterangan yang menyatakan bahwa mahasiswa tersebut lulus tepat pada waktunya dan lulus tidak tepat dengan waktunya.

Data Preprocessing

Keseluruhan data didapat dari *file database* mahasiswa, berikut ini data awal sebelum diproses dinyatakan pada tabel 1

Tabel 1. Dataset Awal Mahasiswa

NO	JK	NPM	TAHUN MASUK	USIA	IPS 1	IPS	IPS 3	IPS 4
1	L	14010001	2014	24	3.09	3.0	3.27	2.90
2	L	15010025	2015	23	3.16	3.33	3.16	3.0
3	L	15010031	2015	22	1.63	2	2.90	3.36
4	L	15010010	2015	22	3.36	3.54	3.0	2.81
5	L	15010005	2015	21	3.0	2.81	2.63	2.41
6	L	15010001	2015	23	3.16	3.33	3.16	3.54
7	L	15010006	2015	22	3.72	3.54	2.90	3.63
8	P	15010012	2015	22	3.45	3.36	2.90	3.09
9	P	15010015	2015	21	3.90	3.81	3.54	3.81
10	P	15010014	2015	23	3.54	3.63	3.18	2.27
.	.	.	.	.	.	.	.	.
56	P	15020004	2015	21	3.72	3.45	3.50	3.18
57	P	11010009	2011	21	3.54	3.81	3.72	3.09

Pemilihan model disesuaikan dengan kebutuhan tujuan penelitian. Dalam penelitian ini tidak semua atribut dihitung, pilih atribut yang dinilai berpengaruh saja yang kemudian digunakan dalam perhitungan penelitian.

Tabel 2. Dataset dengan Keterangan

JK	USIA	IPS 1	IPS	IPS 3	IPS 4	KETERANGAN
L	24	3.09	3.0	3.27	2.90	TEPAT WAKTU
L	23	3.16	3.33	3.16	3.0	TEPAT WAKTU
L	22	1.63	2	2.90	3.36	TIDAK TEPAT WAKTU
L	22	3.36	3.54	3.0	2.81	TEPAT WAKTU
L	21	3.0	2.81	2.63	2.41	TEPAT WAKTU
L	23	3.16	3.33	3.16	3.54	TEPAT WAKTU
L	22	3.72	3.54	2.90	3.63	TEPAT WAKTU
P	22	3.45	3.36	2.90	3.09	TEPAT WAKTU
P	21	3.90	3.81	3.54	3.81	TEPAT WAKTU
P	23	3.54	3.63	3.18	2.27	TEPAT WAKTU
.	.	.	.	.	.	.
P	21	3.72	3.45	3.50	3.18	TEPAT WAKTU
P	21	3.54	3.81	3.72	3.09	TEPAT WAKTU

Pada tabel 4 ini beberapa atribut pada dataset awal sudah tidak disertakan lagi karena dianggap tidak berpengaruh terhadap perhitungan metode yang digunakan penelitian.

4.1 Hasil Eksperimen

Jumlah data yang digunakan sebanyak 57 *record*, 21 dinyatakan lulus tidak tepat waktu dan 36 dinyatakan lulus tepat waktu, sehingga perhitungan node nya dinyatakan sebagai berikut:

NODE	Jumlah Data	Tepat Waktu	Tidak Tepat Waktu	Entropy	Gain
TOTAL	57	36	21	0,949452015	
JK	L	27	15	0,99107606	0,016157619
	P	30	21	0,881290899	
USIA	<=22	35	22	0,951762676	0,0002
	>=23	22	14	0,945660305	
IPS 1	1,0-2,0	1	0	0	0,017273779
	2,01-3,0	6	3	1	
	3,01-4,0	50	32	0,942683189	
IPS 2	1,0-2,0	2	1	1	0,02278377
	2,01-3,0	9	4	0,99107606	
	3,01-4,0	46	31	0,910878379	
IPS 3	1,0-2,0	1	0	0	0,026306786
	2,01-3,0	15	10	0,918295834	
	3,01-4,0	41	26	0,947435136	
IPS 4	1,0-2,0	1	0	0	0,025863481
	2,01-3,0	19	12	0,949452015	
	3,01-4,0	37	24	0,93526914	

Pada atribut jenis kelamin diperoleh data 27 laki-laki dan 30 perempuan dengan data laki-laki lulus tepat waktu 15 mahasiswa dan tidak tepat waktu 12 mahasiswa maka diperoleh entropi 0,99107606, sedangkan data perempuan tepat waktu 21 mahasiswa dan tidak tepat waktu 9 mahasiswa maka diperoleh entropi 0,881290899. Sehingga didapat nilai gain jenis kelamin 0,016157619.

Pada atribut usia diperoleh data usia <= 22 tahun sebanyak 35 mahasiswa dan usia >=23 sebanyak 22 mahasiswa dengan data usia <= 22 tahun 22 mahasiswa tepat waktu dan 13 mahasiswa tidak tepat waktu maka diperoleh entropi 0,951762676, sedangkan data usia >= 23 14 mahasiswa tepat waktu dan 8 mahasiswa tidak tepat waktu maka diperoleh entropi 0,945660305. Sehingga didapat nilai gain 0,0002.

Pada atribut IPS 1 diperoleh data ip 1,0-2,0 sebanyak 1 mahasiswa tidak tepat waktu dengan entropi 0, ip 2,01-3,0 sebanyak 6 mahasiswa dengan 3 mahasiswa tepat waktu dan 3 mahasiswa tidak tepat waktu dengan entropi 1,000000, ip 3,01-4,0 sebanyak 50 mahasiswa dengan 32 mahasiswa tepat waktu dan 18 mahasiswa tidak tepat waktu dengan entropi 0,942683189. Sehingga didapat nilai gain 0,017273779.

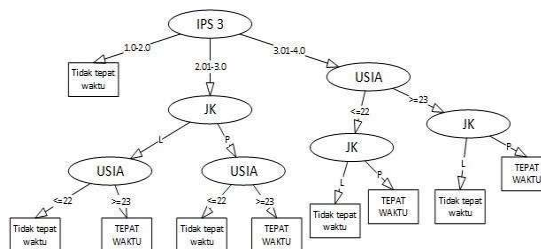
Pada atribut IPS 2 diperoleh data ip 1,0-2,0 sebanyak 2 mahasiswa dengan 1 mahasiswa tepat waktu dan 1 mahasiswa tidak tepat waktu dengan entropi 1,000000, ip 2,01-3,0 sebanyak 9 mahasiswa dengan 4 mahasiswa tepat waktu dan 5 mahasiswa tidak tepat waktu dengan entropi 0,99107606, ip 3,01-4,0 sebanyak 46 mahasiswa dengan 31 mahasiswa tepat waktu dan 15 mahasiswa tidak tepat waktu dengan entropi 0,910878379. Sehingga didapat nilai gain 0,02278377.

Pada atribut IPS 3 diperoleh data ip 1,0-2,0 sebanyak 1 mahasiswa tidak tepat waktu dengan entropi 0, ip 2,01-3,0 sebanyak 15 mahasiswa dengan 10 mahasiswa tepat waktu dan 5 mahasiswa tidak tepat waktu dengan entropi 0,918295834, ip 3,01-4,0 sebanyak 41 mahasiswa dengan 26 mahasiswa tepat waktu dan 15 mahasiswa tidak tepat waktu dengan entropi 0,947435136. Sehingga didapat nilai gain 0,026306786.

Pada atribut IPS 4 diperoleh data ip 1,0-2,0 sebanyak 1 mahasiswa tidak tepat waktu dengan entropi 0, ip 2,01-3,0 sebanyak 19 mahasiswa dengan 12 mahasiswa tepat waktu dan 7 mahasiswa tidak tepat waktu dengan entropi 0,949452015, ip

3,01-4,0 sebanyak 37 mahasiswa dengan 24 mahasiswa tepat waktu dan 13 mahasiswa tidak tepat waktu dengan entropi 0,93526914. Sehingga didapat nilai gain 0,025863481.

Dari perhitungan nilai *information gain* pada semua atribut, maka dipilih atribut ips 3 sebagai akar atau *root* karena memiliki nilai *gain* yang paling tinggi diantara atribut-atribut yang ada.



Gambar 3. Pohon Keputusan

Pohon keputusan terbentuk dari perhitungan gain tertinggi yaitu IPS 3 yang kemudian dijadikan sebagai akar. Pada IPS 3 dengan *range* 1,0-2,0 memiliki keterangan tidak tepat waktu, sedangkan dengan *range* 2,01-3,0 dengan perhitungan ulang entropi dan gain menghasilkan cabang jenis kelamin dan usia laki-laki <=22 tahun tidak tepat waktu >=23 tepat waktu dan perempuan <=22 tahun tidak tepat waktu >=23 tepat waktu, dengan *range* 3,01-4,0 usia <=22 tahun jenis kelamin laki-laki tidak tepat waktu dan perempuan tepat waktu, sedangkan usia >=23 jenis kelamin laki-laki tidak tepat waktu dan perempuan tepat waktu.

5 CONCLUSION

Berdasarkan perhitungan entropi dan gain pada setiap atribut diatas didapat root IPS 3 sebagai akar karena memiliki nilai gain paling tinggi diantara atribut yang ada yaitu 0,026306786. Klasifikasi pada algoritma C4.5 memiliki kompleksitas yang tinggi karena pada setiap atribut entropi harus ditelusuri untuk mendapatkan nilai gain. Proses yang sama juga dilakukan untuk membentuk pohon keputusan yang kemudian dijadikan sebuah model.

REFERENCES

- Astuti, I. P. (2017). Prediksi Ketepatan Waktu Kelulusan Dengan Algoritma Data Mining C4.5. *Fountain of Informatics Journal*, 42-45.
- Dahan, H., Cohen, S., Rokach, L., & Maimon, O. (2014). *Proactive Data Mining with Decision Trees*. Israel: Springer.
- Hirmawan, D. (n.d.). Aplikasi Data Mining Menggunakan Algoritma ID3 untuk Mengklasifikasi Kelulusan Mahasiswa Pada Universitas Dian Nuswantoro Semarang.

- Huda, N. M. (2010). APLIKASI DATA MINING UNTUK MENAMPILKAN INFORMASI TINGKAT KELULUSAN MAHASISWA.
- Kartini, D. (2017). Penerapan Data Mining Dengan Algoritma Neural Network (Backpropagation) Untuk Prediksi Lama Studi Mahasiswa. *Prosiding Seminar nasional Sisfotek*, 235-241.
- Larose, D. T., & Larose, C. D. (2015). *Data Mining and Predictive Analytics*. Canada: John Wiley & Sons.
- Mardi, Y. (2016). Data Mining: Klasifikasi Menggunakan Algoritma C4.5. *Edik Informatika*, 213-219.
- Mayadewi, P., & Rosely, E. (2015). PREDIKSI NILAI PROYEK AKHIR MAHASISWA MENGGUNAKAN ALGORITMA KLASIFIKASI DATA MINING. *Seminar Nasional Sistem Informasi Indonesia*, 1-6.
- Prabowo, F. E., & Kodar, A. (2019). Analisis Prediksi Masa Studi Mahasiswa Menggunakan Algoritma Naive Bayes. *Jurnal Ilmu Teknik dan Komputer*, 147-151.
- Pusfitasari, D., Saputri, T. A., & Perdana, A. (2021). SISTEM PAKAR PENENTUAN MENU PEMENUHAN GIZI IBU HAMIL. *IRobot*.
- Susanto, S., & Suryadi, D. (2010). *PENGANTAR DATA MINING Menggali Pengetahuan dari Bongkahan Data*. Yogyakarta: ANDI Yogyakarta.