MPLEMENTASI ALGORITMA C4.5 UNTUK MENENTUKAN ATURAN REKOMENDASI CALON PENERIMA BEASISWA

Ratna Rahmawati Rahayu
Sistem Informasi, STMIK Bani Saleh
ratnaridw4n@gmail.com

ABSTRAK

Database sebagai media untuk penyimpanan dan pengolahan data sudah menjadi kebutuhan di setiap instansi, perusahaan dan individu. Setiap transaksi atau kejadian disimpan dalam database, sehingga ukuran database pun semakin besar. Salah satunya database akademik yang menyimpan data identitas dan prestasi mahasiswa. Dengan memanfaatkan database yang sudah ada diharapkan dapat menentukan rekomendasi mahasiswa yang layak sebagai calon penerima beasiswa.
Salah satu teknik untuk menemukan informasi yang berguna dari database yang berukuran besar adalah algoritma C4.5 yang merupakan metode classification dari data mining. Dari database akademik diambil beberapa field yang mendukung dalam persyaratan beasiswa, kemudian diolah berdasarkan tahapan data mining dengan teknik algoritma C4.5, dan yang hasilnya berupa decision tree yang dapat digunakan untuk menentukan rekomendasi kelayakan mahasiswa sebagai calon penerima mahasiswa.

Kata Kunci : Algoritma C4.5, Data Mining, Pohon Keputusan, Beasiswa

ABSTRACT

The database as a medium for storing and processing data has become a necessity in every agency, company and individual. Each transaction or event is stored in the database, so the database size is even greater. One of them is an academic database that stores student identity and achievement data. By utilizing the existing database, it is expected to determine the recommendations of eligible students as prospective scholarship recipients.
One technique for finding useful information from large databases is the C4.5 algorithm which is a classification method of data mining. From academic databases, some fields are supported in the scholarship requirements, then processed based on the stages of data mining with C4.5 algorithm technique, and the results are in the form of a decision tree that can be used to determine student feasibility recommendations as prospective student recipients.

Keywords*: C4.5 algorithm, data mining, decision tree, scholarship*

PENDAHULUAN

Beasiswa memegang peranan yang sangat penting dalam dunia pendidikan. Karena beasiswa merupakan pemberian berupa bantuan keuangan yang diberikan kepada mahasiswa untuk dapat menyelesaikan pendidikan yang ditempuh. Beasiswa dapat diberikan oleh pemerintah, yayasan, institusi atau lembaga lainnya. Setiap jenis beasiswa memiliki persyaratan-persyaratan tertentu. Secara umum mekanisme pelaksanaannya, setiap mahasiswa mendaftar dan menyerahkan dokumen pengurusan beasiswa ke unit yang ditunjuk. Setelah pemberkasan ditingkat institusi selanjutnya dokumen diserahkan ke instansi beasiswa terkait untuk diseleksi lebih lanjut. Hal tersebut dapat menyebabkan penumpukan dokumen pendaftaran, karena pada saat pengumuman pendaftaran beasiswa belum dilampirkan daftar mahasiswa yang direkomendasikan untuk menerima beasiswa tersebut.

Rekomendasi mahasiswa yang layak untuk menerima beasiswa berdasarkan data yang memenuhi persyaratan. Dengan data mining, database akademik yang diantaranya berisikan sekumpulan data mahasiswa yang telah merima beasiswa dapat dimanfaatkan untuk menemukan pola-pola atau aturan-aturan yang menarik untuk menentukan kriteria-kriteria mahasiswa yang telah menerima beasiswa. Dan selanjutnya dari kriteria-kriteria tersebut dapat dibuatkan aplikasi untuk menentukan rekomendasi calon penerima beasiswa. Sehingga mahasiswa yang dapat mendaftar beasiswa

hanya yang tercantum dalam daftar rekomendasi calon penerima beasiswa. Dan tentunya ini akan mengurangi jumlah pendaftar dan penumpukan dokumen.

Dalam penelitian ini bertujuan menemukan kriteria-kriteria untuk menentukan rekomendasi calon penerima beasiswa dengan algoritma C4.5. Algoritma C4.5 merupakan salah satu metode klasifikasi dari data mining yang hasil dalam bentuk *decision tree* (pohon keputusan).

Hasil dari penelitian ini diharapkan dapat memberikan beberapa manfaat. Diantaranya dapat diimplementasikan untuk menentukan rekomendasi calon penerima beasiswa untuk berbagai jenis beasiswa, karena setiap beasiswa memiliki beberapa perbedaan persyaratan yang berkemungkinan *decision tree* yang dihasilkan akan berbeda. Selain itu dapat digunakan sebagai seleksi awal dalam penerimaan pendaftaran beasiswa. Dengan adanya tulisan ini diharapkan dapat dijadikan gambaran untuk tulisan selanjutnya.

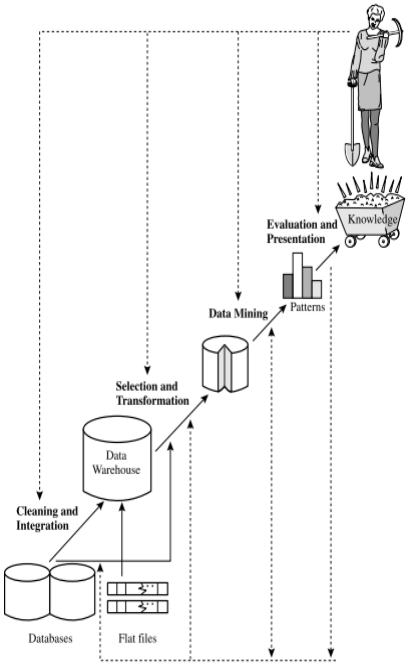
METODE

1. Data Mining

Data mining dapat diartikan sebagai ilmu penggalian informasi yang berguna dari *dataset* atau *database* yang berukuran besar (Gorunescu, 2011:4).

Menurut Jiawei Han dan Micheline Kamber (2006:7) tahapan proses data mining terdiri dari :

1. *Data Cleaning*, menghilangkan noise dan data yang tidak konsisten.
2. *Data integration*, menggabungkan beberapa sumber data.
3. *Data selection*, memilih data yang relevan untuk dianalisis yang diambil dari database.
4. *Data transformation*, mengubah data kedalam bentuk yang sesuai dengan model yang digunakan.
5. *Data mining*, proses aplikasi dengan suatu metode untuk mengekstrak pola data.
6. *Pattern evaluation*, mengidentifikasi pola yang benar-benar menarik yang mewakili pengetahuan yang didasarkan pada beberapa pengukuran.
7. *Knowledge presentation*, teknik representasi *knowledge* yang digunakan untuk menyajikan *knowledge* kepada pengguna.



Gambar 1. Tahapan proses data mining

2. Pohon Keputusan

Menurut Kusrini dan Emha Taufiq Luthfi (2009:13) bahwa pohon keputusan merupakan salah satu bentuk keluaran dari metode klasifikasi yang sangat kuat dan terkenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami. Dan pohon keputusan tersebut juga dapat diekspresikan dalam bentuk bahasa basis data seperti *Structured Query Language* untuk mencari *record* pada kategori tertentu.

Pohon keputusan terdiri dari node-node internal yang mewakili keputusan-keputusan terkait titik perpecahan, dan cabang-cabang yang mewakili partisi data yang diberi label dengan kelas mayoritas (Zaki, Mohammed J. Meira, Wagner, 2014:482)

Banyak algoritma yang dapat dipakai dalam pembentukan pohon keputusan, antara lain ID3, CART, dan C4.5. Algoritma C4.5 merupakan pengembangan dari algoritma ID3 (Larose, 2005:116)

Untuk penelitian ini algoritma yang digunakan untuk membuat pohon keputusan adalah algoritma C.45

3. Algoritma C4.5

Algoritma C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan, yang secara umum dilakukan dengan cara :

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

Untuk memilih atribut sebagai akar, dengan menghitung nilai gain setiap atribut. Untuk atribut yang memiliki nilai gain tertinggi dijadikan sebagai akar.

Rumus Gain :

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

S : himpunan kasus
A : atribut
n : jumlah partisi atribut A
$|S_i|$: jumlah kasus pada partisi ke-i
|S| : jumlah kasus dalam S

Rumus Entropy :

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

S : himpunan kasus
N : jumlah partisi S
Pi : perbandingan Si terhadap S (Si/S)

Dalam penelitian ini, data sampel yang digunakan terbatas hanya pada program studi Sistem Informasi semester 6 (enam) tahun akademik 2017/2018 dengan jumlah 37. Untuk pengumpulan data identitas dan prestasi mahasiswa dengan mengambil dari database akademik yang sudah ada. Data yang digunakan dalam penelitian ini adalah : Tanggal Lahir, Tahun Lulus SLTA, Penghasilan Orang Tua, Prestasi, IPK dan Status Beasiswa.

Untuk memudahkan dalam penelitian ini, data yang sudah ada ditransformasi terlebih dahulu, disesuaikan dengan kebutuhan penelitian ini. Bentuk transformasi yang digunakan sebagai berikut :

1. Usia ≤ 21, isian data "Ya" atau "Tidak" diambil dari pengolahan data Tanggal Lahir.
2. Baru Lulus, isian data "Ya" atau "Tidak" diperoleh apabila Tahun Lulus SMU/SMK sama dengan tahun berjalan.
3. Penghasilan Orang Tua (Juta), dikelompokkan menjadi "< 2", "2 … 3" dan "> 3".
4. Prestasi, isian data "Ya" atau "Tidak". Prestasi berupa kegiatan kejuaraan yang pernah diikuti.
5. IPK, dikelompokkan sebagai berikut :

ISIAN DATA	IPK
Sangat Baik	IPK ≥ 3
Baik	2.75 < IPK < 3
Cukup	2 ≤ IPK ≤ 2.75

6. Beasiswa, isian data "Ya" atau "Tidak".

Sehingga setelah dilakukan pengambilan dan transformasi ke bentuk yang telah ditentukan maka data yang digunakan dalam penelitian ini dapat ditunjukkan pada tabel 1.

Tabel 1. Data Sampel

NO.	Usia ≤ 21	Baru Lulus	Penghasilan Orang Tua (Juta)	Prestasi	IPK	Beasiswa
1	Ya	Ya	>3	Tidak	Baik	Tidak
2	Tidak	Tidak	2 - 3	Ya	Sangat Baik	Tidak
3	Tidak	Tidak	2 - 3	Ya	Sangat Baik	Tidak
4	Ya	Ya	>3	Tidak	Sangat Baik	Tidak
5	Ya	Tidak	>3	Tidak	Baik	Tidak
6	Ya	Tidak	2 - 3	Ya	Cukup	Tidak
7	Ya	Ya	2 - 3	Tidak	Baik	Tidak
8	Ya	Tidak	>3	Ya	Sangat Baik	Ya
9	Tidak	Tidak	<2	Ya	Cukup	Tidak
10	Ya	Tidak	>3	Ya	Sangat Baik	Ya
11	Ya	Tidak	2 - 3	Tidak	Sangat Baik	Tidak
12	Ya	Tidak	>3	Tidak	Cukup	Tidak
13	Ya	Tidak	2 - 3	Tidak	Cukup	Tidak
14	Tidak	Tidak	>3	Ya	Sangat Baik	Tidak
15	Ya	Ya	<2	Tidak	Sangat Baik	Tidak
16	Tidak	Tidak	2 - 3	Tidak	Baik	Tidak
17	Ya	Ya	<2	Tidak	Sangat Baik	Tidak
18	Ya	Ya	<2	Ya	Cukup	Tidak
19	Ya	Tidak	2 - 3	Tidak	Baik	Tidak
20	Ya	Tidak	>3	Ya	Sangat Baik	ya
21	Ya	Ya	>3	Ya	Sangat Baik	Ya
22	Ya	Ya	2 - 3	Tidak	Cukup	Tidak
23	Ya	Ya	>3	Tidak	Sangat Baik	Ya
24	Ya	Ya	2 - 3	Tidak	Baik	Tidak
25	Ya	Ya	2 - 3	Tidak	Cukup	Tidak
26	Ya	Tidak	<2	Tidak	Sangat Baik	Tidak
27	Ya	Tidak	2 - 3	Tidak	Sangat Baik	Tidak
28	Ya	Ya	2 - 3	Ya	Cukup	Tidak
29	Ya	Ya	2 - 3	Ya	Sangat Baik	Ya
30	Tidak	Tidak	>3	Tidak	Sangat Baik	Tidak
31	Ya	Ya	2 - 3	Tidak	Baik	Tidak
32	Tidak	Tidak	2 - 3	Tidak	Baik	Tidak
33	Ya	Ya	2 - 3	Tidak	Cukup	Tidak
34	Ya	Ya	>3	Ya	Sangat Baik	Ya
35	Ya	Ya	>3	Tidak	Baik	Tidak
36	Tidak	Tidak	<2	Tidak	Sangat Baik	Tidak
37	Ya	Ya	<2	Ya	Sangat Baik	Ya

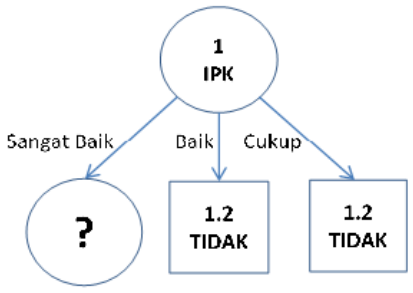
HASIL DAN PEMBAHASAN

Hasil pengolahan dengan Algoritma C4.5 berupa pohon keputusan dengan menghitung nilai entropy dan gain. Untuk atribut yang memiliki nilai gain tertinggi akan dijadikan node. Untuk hasil perhitungan nilai entropy dan gain data sampel dapat dilihat pada tabel 2.

Tabel 2. Hasil Perhitungan Gain untuk Node 1

NODE	ATRIBUT	PARTISI	JML. KASUS (S)	JML. YA (S_1)	JML. TIDAK (S_2)	ENTROPY	GAI
1	BEASISWA		37	8	29	0.753	
	Usia ≤ 21						0.0
		Ya	29	8	21	0.850	
		Tidak	8	0	8	-	
	Baru Lulus						0.0
		Ya	18	5	13	0.852	
		Tidak	19	3	16	0.629	
	Penghasilan Orang Tua (Juta)						0.1
		<2	7	1	6	0.592	
		2 - 3	17	1	16	0.323	
		>3	13	6	7	0.996	
	Prestasi						0.2
		Ya	14	7	7	1.000	
		Tidak	23	1	22	0.258	
	IPK						**0.24**
		Sangat Baik	19	8	11	0.982	
		Baik	9	0	9	-	
		Cukup	9	0	9	-	

Nilai gain terbesar dimiliki atribut IPK, maka atribut IPK dijadikan node 1. Pada atribut IPK untuk kategori "Sangat Baik" masih memiliki label "Ya" dan "Tidak", sedangkan kategori "Baik" dan "Cukup" hanya memiliki satu label yaitu "Tidak". Dari tabel 2 dapat digambarkan pohon keputusan seperti yang terdapat pada gambar 2.



Gambar 2. Pohon keputusan untuk node 1

Karena kategori "Sangat Baik" masih belum memiliki keputusan karena masih memiliki label "Ya" dan "Tidak", maka dilakukan perhitungan entropy dan gain berdasarkan IPK dengan kategori "Sangat

Baik" dan hasilnya dapat dilihat pada tabel 3.

Tabel 3. Hasil Perhitungan Gain untuk Node 1.1

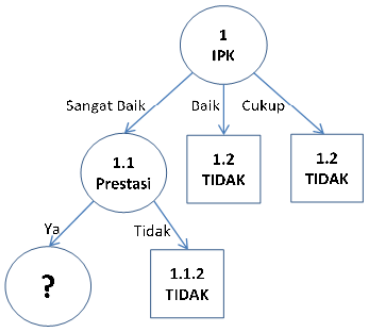
NODE	ATRIBUT	PARTISI	JML. KASUS (S)	JML. YA (S_1)	JML. TIDAK (S_2)	ENTROPY	GAIN
1.1	IPK	Sangat Baik	19	8	11	0.982	
	Usia ≤ 21						0.25
		Ya	14	8	6	0.985	
		Tidak	5	0	5	-	
	Baru Lulus						0.09
		Ya	8	5	3	0.954	
		Tidak	11	3	8	0.845	
	Penghasilan Orang Tua (Juta)						0.16
		< 2	5	1	4	0.722	
		2 - 3	5	1	4	0.722	
		> 3	9	6	3	0.918	
	Prestasi						**0.28**
		Ya	10	7	3	0.881	
		Tidak	9	1	8	0.503	

Berdasarkan data di tabel 3, nilai gain terbesar dimiliki atribut Prestasi maka atribut Prestasi dijadikan node 1.1. Atribut Prestasi memiliki dua kategori yang masing-masing masih memiliki lebih dari 1 (satu) label. Agar atribut Prestasi menghasilkan keputusan maka dilakukan perhitungan nilai support.

Atribut Prestasi memiliki kategori :

- ✓ Kategori = "Ya"
  - Label "Ya" = 7, nilai support = 7/10 = 0.7
  - Label "Tidak" =3, nilai support 3/10 = 0.3
- ✓ Kategori = "Tidak"
  - Label "Ya" = 1, nilai support = 1/9 = 0.11
  - Label "Tidak" = 8, nilai support = 8/9 = 0.89

Dari hasil tersebut dapat disimpulkan untuk atribut Prestasi akan menghasilkan keputusan kategori "Tidak" dengan label "Ya", karena yang memiliki nilai support terkecil. Sehingga dapat digambarkan pohon keputusan seperti yang terdapat pada gambar 3.



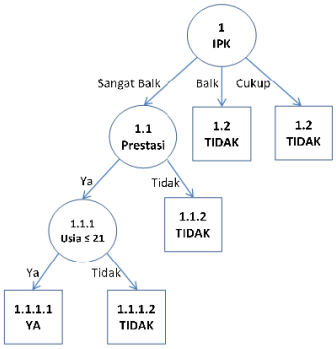
Gambar 3. Pohon keputusan untuk node 1.1

Karena kategori "Ya" masih belum memiliki keputusan karena masih memiliki label "Ya" dan "Tidak", maka dilakukan perhitungan entropy dan gain berdasarkan Prestasi dengan kategori "Ya" dan hasilnya dapat dilihat pada tabel 4.

Tabel 4. Hasil Perhitungan Gain untuk Node 1.1.1

NODE	ATRIBUT	PARTISI	JML. KASUS (S)	JML. YA (S_1)	JML. TIDAK (S_2)	ENTROPY	GAIN
1.1.1	Prestasi	Ya	10	7	3	0.881	
	Usia ≤ 21						**0.881**
		Ya	7	7	0	-	
		Tidak	3	0	3	-	
	Baru Lulus						0.281
		Ya	4	4	0	-	
		Tidak	6	3	3	1.000	
	Penghasilan Orang Tua (Juta)						0.216
		< 2	1	1	0	-	
		2 - 3	3	1	2	0.918	
		> 3	6	5	1	0.650	

Pada tabel 4, nilai gain tertinggi terdapat pada atribut Usia ≤ 21 dengan kategori "Ya" hanya memiliki label "Ya" dan kategori "Tidak" hanya memiliki label "Tidak". Karena semua kategori sudah memiliki hanya 1 (satu) label maka perhitungan nilai entropy dan gain berakhir, karena setiap cabang pohon keputusan sudah memiliki label. Sehingga dapat digambarkan pohon keputusan seperti yang terdapat pada gambar 4.



Gambar 4. Pohon keputusan hasil algoritma C4.5

Berdasarkan pohon keputusan pada gambar 4, dapat dibuat aturan sebagai berikut:

If IPK = “Sangat Baik” Then

If Prestasi = “Ya” Then

If Usia≤21 Then
Beasiswa = ”YA”
Else
Beasiswa = ”TIDAK”

Else If Prestasi = “Tidak”
Beasiswa = ”TIDAK”

If IPK = “Baik” Then
Beasiswa = ”TIDAK”

If IPK = “Cukup” Then
Beasiswa = ”TIDAK”

Dari pohon keputusan dan aturan yang dihasilkan maka dapat dibuatkan keputusan bahwa mahasiswa yang dapat direkomendasikan untuk calon penerima beasiswa adalah mahasiswa yang memiliki IPK ≥ 3, pernah mengikuti kegiatan kejuaraan dan usia ≤ 21 pada saat direkomendasikan.

PENUTUP

1. Kesimpulan

Dalam penelitian mengenai implementasi algoritma C4.5 untuk menentukan aturan rekomendasi calon penerima beasiswa dengan studi kasus mahasiswa program studi Sistem Informasi semester 6 (enam) tahun akademik 2017/2018, dapat dibuat kesimpulan sebagai berikut:

1. Penelitian bertujuan untuk menentukan aturan atau pola yang dimiliki oleh mahasiswa penerima beasiswa, dengan mengetahui aturan/pola tersebut dapat diaplikasikan untuk menentukan rekomendasi mahasiswa yang layak untuk menerima beasiswa.
2. Atribut data mahasiswa yang digunakan adalah Tanggal Lahir, Tahun Lulus SMU/SMK, Penghasilan Orang Tua, IPK, Prestasi dan Status Beasiswa. Untuk memudahkan, data-data tersebut ditransformasi terlebih dahulu disesuaikan dengan kebutuhan penelitian.
3. Berdasarkan label Beasiswa, dilakukan perhitungan menentukan nilai entropy dan nilai gain untuk setiap atribut. Atribut yang memiliki nilai gain tertinggi dijadikan node akar. Lalu berdasarkan node tersebut dilakukan perhitungan kembali untuk menentukan nilai entropy dan nilai gain. Nilai gain tertinggi dijadikan node. Hal tersebut dilakukan sampai semua cabang sudah memiliki label.
4. Setiap tabel gain dan entropi dibuatkan pohon keputusan sehingga memudahkan dalam menentukan keputusan untuk setiap atribut yang ada.
5. Sedangkan untuk memudahkan dalam pembuatan program maka pohon keputusan dapat dibuatkan dalam bentuk aturan.
6. Berdasarkan data sampel sebanyak 37 (tiga puluh tujuh) mahasiswa semester 6 (enam) program studi Sistem Informasi TA 2017/2018 diperoleh aturan bahwa mahasiswa yang memperoleh beasiswa adalah yang memiliki IPK sangat baik, berprestasi dan berusia kurang dari sama dengan 21 (dua puluh satu) tahun.
7. Perolehan beasiswa tidak dipengaruhi oleh penghasilan orang tua dan tahun kelulusan SMU/SMK.

2. Saran

Untuk menambah pengayaan dan diperoleh aturan yang lebih spesifik untuk setiap jenis beasiswa maka dapat disarankan sebagai berikut :

1. Dalam penelitian selanjutnya dengan menambah atribut-atribut lain.

2. Untuk atribut beasiswa lebih spesifik dengan mencantumkan jenis atau program beasiswa, agar terlihat lebih jelas aturan-aturan yang terbentuk untuk setiap program beasiswa. Karena setiap program beasiswa umumnya memiliki persyaratan atau ketentuan yang berbeda.
3. Hasil penelitian dapat diimplementasikan dalam bentuk aplikasi sehingga untuk menentukan mahasiswa yang layak untuk direkomendasikan sebagai calon penerima beasiswa sudah tersaring oleh sistem.

DAFTAR PUSTAKA

Gorunescu F. 2011. *Data Mining : Concept, Model and Techniques*. Intelligent System Reference Library Volume 12. Springer.

Han, Jiawei. Kamber, Micheline. 2006. *Data Mining : Concept and Techniques*. Morgan Kaufmann Publishers.

Kusrini. Luthfi, Emha Taufiq. 2009. Algoritma Data Mining. Penerbit Andi.

Larose, Daniel T. 2005. *Discovering Knowledge in Data : An Introduction to Data Mining*. John Willey & Sons, Inc.

Zaki, Mohammed J. Meira, Wagner. 2014. *Data Mining and Analysis : Fundamental Concepts and Algorithms*. Cambridge University Press.