PERANCANGAN DAN PEMBUATAN APLIKASI *DATAMINING* BERBASIS WEB MENGGUNAKAN ALGORITMA C4.5 UNTUK MEMPREDIKSI KELULUSAN MAHASISWA S-1 PADA UNIVERSITAS DARMA PERSADA

Disusun oleh :

Junarto

2008230026

Berdasarkan data kelulusan Universitas Darma Persada, banyak mahasiswa yang dinyatakan lulus tidak tepat waktu. Sementara itu sistem informasi yang ada saat ini belum mampu menggali lebih jauh informasi dari banyaknya data yang tersimpan untuk dijadikan sebuah pengetahuan. Oleh karena itu penulis melakukan penelitian dengan membuat aplikasi penerapan data mining dengan metode klasifikasi menggunakan algoritma C4.5. Algoritma C4.5 ini merupakan algoritma *decision tree* terbaik menurut IEEE. Algoritma C4.5 bekerja secara iteratif dengan menghitung nilai Gain tertinggi yang diperoleh berdasarkan penghitungan Entropy masing-masing atribut. Atribut yang digunakan dalam penelitian ini diantaranya: Jenis Kelamin, Jurusan, Nilai Indeks Prestasi Semester 1 hingga Indeks Prestasi Semester 4. Aplikasi ini menghasilkan informasi berupa pohon keputusan dan aturan-atturan (*rules*) yang berguna saat melakukan prediksi kelulusan. Setiap keputusan yang muncul merupakan hasil pencocokan nilai masukan dengan aturan-aturan hasil pembelajaran data kasus.

Kata kunci : *Data Mining*, Algoritma C4.5, Aplikasi Web.

PENDAHULUAN

Dengan kemajuan teknologi informasi dewasa ini, kebutuhan akan informasi yang akurat sangat dibutuhkan dalam kehidupan sehari-hari, sehingga informasi akan menjadi suatu elemen penting dalam perkembangan masyarakat saat ini dan waktu mendatang. Namun kebutuhan informasi yang tinggi kadang tidak diimbangi dengan penyajian informasi yang memadai, sering kali informasi tersebut masih harus digali ulang dari data yang jumlahnya sangat besar. Kemampuan teknologi informasi untuk mengumpulkan dan menyimpan berbagai tipe data jauh meninggalkan kemampuan untuk menganalisis, meringkas dan mengekstrak pengetahuan dari data. Metode tradisional untuk menganalisis data yang ada, tidak dapat menangani data dalam jumlah besar.

Pemanfaatan data yang ada di dalam sistem informasi untuk menunjang kegiatan pengambilan keputusan, tidak cukup hanya mengandalkan data operasional saja, diperlukan suatu analisis data untuk menggali potensi-potensi informasi yang ada. Para pengambil keputusan berusaha untuk memanfaatkan gudang data yang sudah dimiliki untuk menggali informasi yang berguna membantu mengambil keputusan, hal ini mendorong munculnya cabang ilmu baru untuk mengatasi masalah penggalian informasi atau pola yang penting atau menarikdari data dalam jumlah besar, yang disebut dengan *data mining*. Penggunaan teknik *data mining* diharapkan dapat memberikan pengetahuan-pengetahuan yang sebelumnya tersembunyi di dalam *database* sehingga menjadi informasi yang berharga.

Perguruan tinggi saat ini dituntut untuk memiliki keunggulan bersaing dengan memanfaatkan semua sumber daya yang dimiliki. Selain sumber daya sarana, prasarana, dan manusia, sistem informasi adalah salah satu sumber daya yang dapat digunakan untuk meningkatkan

keunggulan bersaing. Dilihat dari data kelulusan mahasiswa Universitas Darma Persada pada angkatan tahun 2008, sebagianlulusan Program Sarjana (S1) reguler di Universitas Darma Persada menempuh masa studi lebih dari 8 semester dari yang dijadwalkan yaitu 8 semester. Oleh karena itu, dengan memanfaatkan data induk mahasiswa dan data akademik mahasiswa, dapat diketahui pola kelulusan berdasarkan nilai mata kuliah pada empat semester pertama.
Untuk dapat mengetahui pola tersebut, dibutuhkan algoritma metode *data mining* yang sesuai dengan kasus tersebut. Berdasarkan *Paper* yang dipresentasikan pada *International Conference on Data Mining* (ICDM), IEEE mengidentifikasi 10 algoritma *data mining* terbaik. Dengan memperhatikan kelebihan suatu algoritma dalam mempelajari data, terpilihlah Algoritma C4.5 sebagai algoritma terbaik diantara 10 nominasi algoritma *data mining.*

LANDASAN TEORI

1. Data Mining

Data Mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam basis data. Data Mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai basis data besar (Kusrini & Emha Taufiq Luthfi, 2009:25).

Data Mining adalah salah satu bidang yang berkembang pesat karenabesarnya kebutuhan akan nilai tambah dari database skala besar yang makin banyak terakumulasi sejalan dengan pertumbuhan teknologi informasi. (Iko Pramudiono, 2003:1).Data mining, sering juga disebut Knowledge Discovery in Database (KDD),adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Keluaran dari data mining ini bisa dipakai untuk memperbaiki pengambilan keputusan di masa depan. Sehingga istilah pattern recognition jarang digunakan karena termasuk bagian dari data mining (Azevedo, A. Santos & Manuel F,2008:30).

Menurut Gartner Group, data mining adalah suatu proses menemukan hubungan yang berarti, pola, dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika (Larose, 2005).

2. Teknik Klasifikasi

Teknik Klasifikasi dalam data mining dikelompokkan ke dalam Teknik Pohon Keputusan, Bayesian (Naïve Bayesian dan Bayesian Belief Networks), Jaringan Saraf Tiruan (Backpropagation), Teknik yang berbasis konsep dari penambangan aturan-aturan asosiasi, dan teknik lain (K-Nearest Neighboor, algoritma genetik, teknik dengan pendekatan himpunan rough dan fuzzy)

Setiap teknik memiliki kelebihan dan kekurangannya sendiri. Data dengan profil tertentu mungkin paling optimal jika diklasifikasi dengan teknik tertentu, atau dengan kata lain, profil data tertentu dapat mendukung termanfaatkannya kelebihan dari teknik ini.

Secara umum, Proses Klasifikasi dapat dilakukan dalam dua tahap, yaitu proses belajar dari data training dan klasifikasi kasus. Pada proses belajar, Algoritma Klasifikasi mengolah data training untuk menghasilkan sebuah model. Setelah model diuji dan dapat diterima, pada tahap klasifikasi, model tersebut digunakan untuk memprediksi kelas dari kasus baru untuk membantu proses pengambilan keputusan (Han dkk.,2012).

3. Pohon Keputusan

Pohon Keputusan atau Decision Tree merupakan representasi sederhana dari teknik klasifikasi untuk sejumlah kelas berhingga, dimana simpul internal maupun simpul akar ditandai dengan nama atribut, rusuk-rusuknnya diberi label nilai atribut yang mungkin dan simpul daun ditandai dengan kelas-kelas yang berbeda.(Fajar Astuti, 2013).

4. Algoritma C4.5

Secara umum algorima C4.5 untuk membangun pohon keputusan adalah sebagai berikut :

a. Pilih atribut sebagai node akar.

b. Buat cabang untuk tiap-tiap nilai.

c. Bagi kasus dalam cabang.

d. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama (Kusrini & Luthfi, 2011)

Untuk memilih atribut sebagai node akar, didasarkan pada nilai Gain tertinggi dari atribut-atribut yang ada. Untuk menghitung Gain digunakan rumus seperti tertera dalam persamaan berikut :

$$Gain(S,A) = Entrotopy(S) - \sum_{i=1}^{n} * Entropy(Si)$$

Keterangan :

S : himpunan kasus

A: Atribut

n : jumlah partisi atribut

|Si| : jumlah kasus pada partisi ke –i

|S| : jumlah kasus dalam S

Setelah mendapatkan nilai Gain, ada satu hal lagi yang perlu kita lakukan perhitungan, yaitu mencari nilai Entropy. Entropy digunakan untuk menentukan seberapa informatif sebuah masukan atribut untuk menghasilkan keluaran atribut. Rumus dasar dari Entropy tersebut adalah sebagai berikut :

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

Keterangan :

S : Himpunan kasus

A : Fitur

n : Jumlah partisi S

pi : Proporsi dari Si terhadap S

PEMBAHASAN

Berdasarkan prosedur algoritma C4.5 dalam melakukan proses mining untuk membangun pohon keputusan, adalah sebagai berikut :

1. Pilih atribut sebagai node 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 node akar, didasarkan pada nilai Gain tertinggi dari atribut-atribut yang ada. Berikut penerapannya dengan mengambil data dari tabel mahasiswa :

Untuk memudahkan penjelasan mengenai algoritma C4.5, berikut penjabarannya penghitungan detilnya:

Node	Atribut	Nilai Atribut	Jumlah Kasus Total	Jumlah Kasus Tidak Tepat Waktu	Jumlah Kasus TepatWaktu	Entropy	Gain
1	Total	Total	80	22	58	0.8485	
	Jenis kelamin						0.0042
		L	21	7	14	0.9183	
		P	59	15	44	0.8179	
	jurusan						0.0237
		Sastra Cina	10	5	5	1	
		Sastra Inggris	16	4	12	0.8113	
		Sastra Jepang	54	13	41	0.7963	
	Nilai ips1kategori						0.1428
		Cumlaude	20	1	19	0.2864	
		Kurang Memuaskan	3	2	1	0.9183	
		Memuaskan	26	13	13	1	
		Sangat Memuaskan	31	6	25	0.7088	
	Nilaiips2 kategori						0.2392
		Cumlaude	21	1	20	0.2762	
		Kurang Memuaskan	9	7	2	0.7642	
		Memuaskan	31	13	18	0.9812	
		Sangat Memuaskan	19	1	18	0.2975	
	Nilai ips3 kategori						0.2405
		Cumlaude	11	1	10	0.4395	
		Kurang Memuaskan	7	6	1	0.5917	
		Memuaskan	29	13	16	0.9923	
		Sangat Memuaskan	33	2	31	0.3298	
Node	Atribut	Nilai Atribut	Jumlah Kasus Total	Jumlah Kasus Tidak Tepat Waktu	Jumlah Kasus TepatWaktu	Entropy	Gain
	Nilai ips4kategori						0.3154
		Cumlaude	20	1	19	0.2864	
		Kurang Memuaskan	12	11	1	0.4138	
		Memuaskan	19	7	12	0.9495	
		Sangat Memuaskan	29	3	26	0.4798	

Pada Tabel 4.1 diatas telah dipaparkan rincian hitungan Entropi tiap nilai atribut dan Gain untuk suatu atribut. Berikut ini contoh penerapan hitungan rumus Entropi *node* awal yang dilakukan pada atribut Jenis Kelamin:

Nilai entropi total node 1:

$$Entropi\ (S) = \left(-\left(\frac{22}{80}\right) \times \log_2 \frac{22}{80}\right) + \left(-\left(\frac{58}{80}\right) \times \log_2 \frac{58}{80}\right) = 0.8485$$

Nilai entropi atribut Laki-laki :

$$Entropi\ (S) = \left(-\left(\frac{7}{21}\right) \times \log_2 \frac{7}{21}\right) + \left(-\left(\frac{14}{21}\right) \times \log_2 \frac{14}{21}\right) = 0.9183$$

Nilai entropi atribut Perempuan :

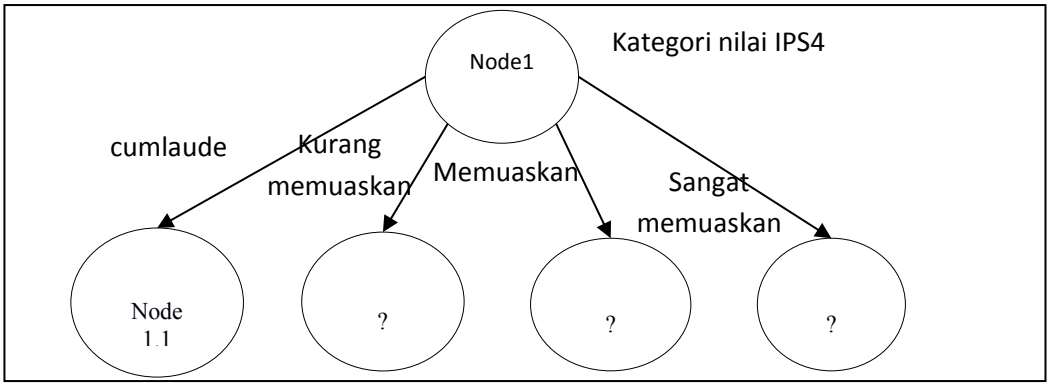
$$Entropi\ (S) = \left(-\left(\frac{15}{59}\right) \times \log_2 \frac{15}{59}\right) + \left(-\left(\frac{44}{59}\right) \times \log_2 \frac{44}{59}\right) = 0.8179$$

Nilai Gain untuk jenis kelamin:

$$Gain(Jenis\ Kelamin) = 0.8485 - \left(\left(\frac{21}{80}\right) \times 0.9183 + \left(\frac{59}{80}\right) \times 0.8179\right) = 0.0042$$

Perhitungan tersebut akan terus berjalan pada setiap node hingga seluruh atribut diketahui nilai gain nya. Apabila sudah diketahui, langkah selanjutnya adalah menentukan atribut yang akan menjadi node akar. Pada node 1 ini atribut dengan nilai gain terbesar adalah kategori nilai ips4. Karena nilai atribut tersebut belum mengklasifikasikan nilai (Ya atau Tidak), maka nilai atribut tersebut akan menjadi node pada level selanjutnya.

Dengan demikian berdasarkan perhitungan Node 1 sudah dapat menggambarkan pohon keputusan seperti gambar di bawah ini :



KESIMPULAN

Aplikasi Data Mining dengan metode klasifikasi algoritma c4.5 ini dapat memproses data mahasiswa menjadi sebuah aturan-aturan yang berguna sebagai masukan penentu keputusan walaupun jumlah data dapat mempengaruhi persentase keakuratannya.

Berdasarkan hasil evaluasi dengan membandingkan keputusan asli dengan keputusan hasil mining c4.5, maka dapat dihitung persentase keakuratan sehingga memperoleh nilai sebesar 85%.

DAFTAR PUSTAKA

A.S Rosa dan Salahuddin M. 2011. Modul Pembelajaran Rekayasa Perangkat Lunak (Terstruktur dan Berorientasi Objek). Bandung : Modula.

Astuti H., Fajar. 2013. Data Mining. Yogyakarta : Penerbit Andi.

Dharwiyanti, Sri dan Wahono, Romi, Satria. 2003. Pengantar Unified Modelling Language (UML):http://ikc.dinus.ac.id/umum/yanti-uml.php

Han, Jiawei. 2012. Data Mining : Concepts and Techniques, USA : Morgan Kaufmann Publishers.

Hariyanto, BAmbang. 2007. Esensi-esensi Bahasa Pemrograman Java, Bandung: Penerbit Informatika.

Irwansyah, Edy. 2014. Pengantar Teknologi Informasi. Yogyakarta : Deepublish.

Kadir, Abdul. 2008. Pengantar Teknologi Informasi. Yogyakarta : Penerbit Andi.

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

Larose, Daniel. 2005. Discovering Knowledge in Data : An Introduction to Data Mining. Canada : Wiley Publishing.

Munawar. 2005. Pemodelan Visual dengan UML. Yogyakarta : Graha Ilmu.

Raharjo, Budi. 2011. Panduan Mudah Untuk Pelajar :Mahasiswa dan Praktis. Bandung : Modula.