Penerapan Algoritma C4.5 Untuk Penentuan Jurusan Mahasiswa

Liliana Swastina
Program Studi Sistem Informasi
Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Indonesia
Banjarmasin, Indonesia
lilisera@gmail.com

Abstrak—**Banyak kasus dijumpai bahwa pemilihan jurusan yang tidak sesuai dengan kemampuan, kepribadian, minat dan bakat dapat mempengaruhi mahasiswa dalam mengikuti perkuliahan.**

Penggunaan pendekatan algoritma klasifikasi data mining akan diterapkan untuk menentukan jurusan dalam bidang studi yang akan diambil oleh mahasiswa, sehingga mahasiswa tidak salah dalam memilih jurusan yang akan di tempuh selama belajar pada perguruan tinggi. Algoritma C4.5 digunakan untuk menentukan jurusan yang akan diambil oleh mahasiswa sesuai dengan latar belakang, minat dan kemampuannya sendiri. Parameter pemilihan jurusan adalah Indeks Prestasi Kumulatif Semester 1 dan 2 .

Hasil eksperimen dan evaluasi menunjukan bahwa Algoritma Decision Tree C4.5 akurat diterapkan untuk penentuan kesesuaian jurusan mahasiswa dengan tingkat akurasi 93,31 % dan akurasi rekomendasi jurusan sebesar 82,64%.

Kata kunci – Klasifikasi, Penentuan Jurusan Mahasiswa, Algoritma C4.5.

I. PENDAHULUAN

Kecenderungan yang terjadi saat ini, banyak siswa kelas XII yang tidak tahu minatnya dan bakatnya serta akan memilih jurusan apa selepas SMU nanti [*1*]. Akibat yang buruk terjadi setelah itu, yaitu keengganan belajar dan menurunnya kualitas serta prestasi akademik karena siswa merasa salah dalam memilih jurusan [*2*]. Banyak kasus dijumpai bahwa pemilihan jurusan yang tidak sesuai dengan kemampuan, kepribadian, minat dan bakat dapat mempengaruhi mahasiswa dalam mengikuti perkuliahan. Dalam beberapa penelitian psikologi pendidikan, minat dan bakat siswa diketahui cukup terkait dengan prestasi akademiknya [*3*].

Sementara itu, data mining adalah proses yang menggunakan statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar [*4*].

Data mining disisi lain adalah kegiatan 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.

Untuk penentuan semacam ini, Zhiwu Liu, dkk [*5*] telah menggunakan datamining untuk melakukannya. Mereka memanfaatkan sifat prediksi yang dimiliki pohon keputusan. Dalam datamining banyak algoritma yang dapat dipakai dalam pembentukan pohon keputusan, antara lain: Algoritma ID3, CART, dan C4.5. Algoritma C4.5 merupakan pengembangan dari algoritma ID3 [*6*].

Rong Cao dan Lizhen Xu [*7*] menggunakan Algoritma C4.5 untuk menganalisa penjualan. Sementara itu, dalam bidang pendidikan, Ossy N [*8*] melakukan penelitian mengenai bagaimana sebuah model fuzzy dapat digunakan juga untuk membuat klasifikasi siswa yang mengikuti suatu kelas dengan kemungkinan berhasil atau gagal. Lebih jauh, Wen-Chih Chang, dkk [*9*], telah melakukan penelitian untuk mengukur kemampuan belajar siswa. Mereka menggunakan algoritma K-Means untuk membentuk klaster-klaster kemampuan.

Bahar melakukan penelitian tentang kurang akuratnya proses pemilihan jurusan dengan sistem manual pada SMA, sehingga perlu suatu penggunaan metode untuk mengelompokan siswa dalam proses pemilihan jurusan. Bahar menggunakan algoritma Fuzzy C-Means untuk mengelompokkan data siswa SMA berdasarkan Nilai mata pelajaran inti untuk proses penjurusan [*10*]. Sumanto, melakukan penelitian tentang kurang akuratnya mahasiswa dalam pemilihan peminatan Tugas Akhir [*11*] yang sesuai dengan

ilmu yang dikuasai oleh mahasiswa sangat berpengaruh dengan nilai tugas akhir dengan menerapkan Fuzzy C-Means untuk memudahkan mahasiswa dalam pemilihan peminatan tugas akhir dengan baik, sesuai dengan kemampuan mahasiswa dengan tingkat akurasi sebesar 82 %.

Berdasarkan pertimbangan diatas, pendekatan datamining dengan penerapan algoritma Decision Tree C4.5 akan dilakukan untuk menentukan jurusan yang akan diambil oleh mahasiswa sesuai dengan latar belakang, minat dan kemampuannya sendiri. Dengan demikian peluang untuk sukses dalam studi di perguruaan tinggi semakin besar.

Dengan demikian diharapkan algoritma Decision Tree C4.5 mampu menjadi alat pendukung keputusan yang digunakan oleh pihak Perguruan Tinggi dalam proses penentuan jurusan mahasiswa.

II. METODOLOGI PENELITIAN

Metode yang digunakan dalam paper ini adalah metode penelitian eksperimen, yang terdiri dari: (1) Pengumpulan data, (2) Pengolahan data awal, (3) Model yang diusulkan, (4) Pengujian model dan (5) Evaluasi dan validasi model.

A. Pengumpulan Data

Data sekunder adalah data yang diperoleh secara tidak langsung bersumber dari dokumentasi, literatur, buku, jurnal dan informasi lainnya yang ada hubungannya dengan masalah yang diteliti. Data sekunder pada penelitian ini adalah : buku-buku, jurnal tentang algoritma Decision Tree C4.5 dan data mining serta data mahasiswa baru STMIK Indonesia Banjarmasin tahun 2008 s.d 2009. Sedangkan Data primer adalah data yang diperoleh dari hasil penelitian. Data primer dalam penelitian ini adalah data hasil uji dengan menggunakan algoritma Decision Tree C4.5

B. Pengolahan Awal Data

Data yang didapatkan dari BAA STMIK Indonesia Banjarmasin yaitu data mahasiswa dengan atribut NIM, Nama, Tanggal Lahir, Asal Sekolah, Nilai UN, Jurusan yang dipilih. Data lain yang akan diolah adalah Indeks Prestasi Kumulatif Semester 1 dan Indeks Prestasi Kumulatif Semester 2.

C. Model Yang Diusulkan

Gartner Group dalam [*6*] menyebutkan bahwa data mining adalah:

1. Proses menelusuri pengetahuan baru
2. Pola dan tren yang dipilah dari jumlah data yang besar yang disimpan dalam repositori atau tempat penyimpanan dengan menggunakan teknik pengenalan pola serta statistik dan tehnik matematika

Di sisi lain, Data mining adalah proses yang menggunakan statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar [*4*]. Sehingga Data mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual [*12*].

Model yang diusulkan untuk menentukan jurusan mahasiswa adalah algoritma Decision Tree C4.5. Model ini akan dibandingkan dengan Model Naïve Bayes.

Tahapan Algoritma Decision Tree C4.5:

1) Menyiapkan data training
2) Menentukan akar dari pohon.
3) Hitung nilai Gain:

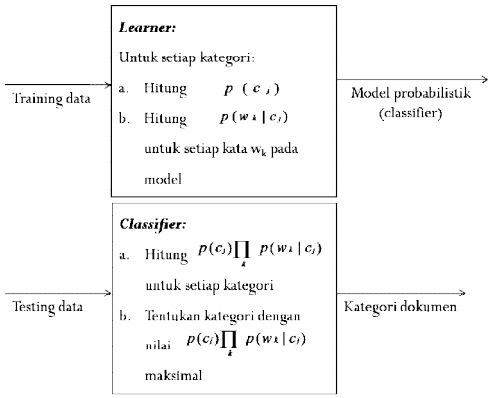
$$Entropy(S) = \sum_{i=1}^{n} - pi * \log_2 pi \qquad (1)$$

4) Ulangi langkah ke-2 hingga semua tupel terpartisi

$$Gain(S, A) = S - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * S_i \qquad (2)$$

5) Proses partisi pohon keputusan akan berhenti saat semua tupel dalam node N mendapat kelas yang sama dan atau tidak ada atribut di dalam tupel yang dipartisi lagi dan atau tidak ada tupel di dalam cabang yang kosong

Sedangkan Naïve Bayes merupakan salah satu metode machine learning yang menggunakan perhitungan probabilitas. Gambaran proses klasifikasi dengan algoritma Naïve Bayes:



GAMBAR 1. Tahapan Proses Klasifikasi dengan

Algoritma Naïve Bayes

D. Pengujian Model

Model yang telah dikembangkan dalam penelitian ini akan diterapkan pada data mahasiswa baru STMIK Indonesia Banjarmasin tahun 2008 s.d 2009 melalui suatu simulasi menggunakan RapidMiner. Data Sampel terdiri dari atribut Nama, Jenis Kelamin, Umur, Asal Sekolah, Jurusan Asal Sekolah, Nilai UAN, IPK Semester 1, IPK Semester 2.

Sebanyak 90 % data akan digunakan untuk membangun struktur pohon keputusan melalui metode C4.5. Sedangkan 10 % lainnya digunakan sebagai data uji.

E. Evaluasi dan Validasi Hasil

Evaluasi dilakukan dengan menganalisa hasil klasifikasi. Pengukuran data dilakukan dengan confusion matrix [*13*] dan ROC Curve (AUC) [*14*] untuk mengevaluasi hasil dari algoritma Decision Tree C4.5.

Confusion matrix merupakan sebuah table yang terdiri dari banyaknya baris data uji yang diprediksi benar dan tidak benar oleh model klasifikasi. Tabel ini diperlukan untuk mengukur kinerja suatu model klasifikasi [*15*]

TABEL 1. Confusion Matrix

		PREDICTED CLASS	
		CLASS = 1	***CLASS = 0***
ACTUAL CLASS	***CLASS = 1***	*F11*	*F10*
	CLASS = 0	*F01*	*F00*

Bentuk tabel confusion matrix dapat dilihat pada Tabel 1. Perhitungan akurasi dengan tabel confusion matrix adalah sebagai berikut:

$$Akurasi = \frac{F\,11+F\,00}{F\,11+F\,10+F\,01+F00} \qquad (3)$$

Penjelasan tentang pengukuran *Precision* dan *recall* dapat di lihat pada Tabel 2 dan perhitungan di berikut:

TABEL 2 Perhitungan Precision dan Recall

	Relevant	**Not Relevant**
Retrieved	A	B
Not Retrieved	C	D

Precision didefinisikan sebagai rasio item relevan yang dipilih terhadap semua item yang terpilih. Precision merupakan probabilitas bahwa sebuah item yang dipilih adalah relevan. Dapat diartikan sebagai kecocokan antara permintaan informasi dengan jawaban terhadap permintaan itu. *Precision* dihitung dengan rumus:

$$A/(A+B) \qquad (4)$$

Sedangkan *Recall* didefinisikan sebagai rasio dari item relevan yang dipilih terhadap total jumlah item relevan yang tersedia. *Recall* merupakan probabilitas bahwa suatu item yang relevan akan dipilih. *Recall* dapat dihitung dengan jumlah rekomendasi yang relevan yang dipilih oleh user dibagi dengan jumlah semua rekomendasi yang relevan baik dipilih maupun rekomendasi yang tidak terpilih. R*ecall* dapat dihitung dengan rumus:

$$A/(A+C) \qquad (5)$$

Precision and Recall dapat diberi nilai dalam bentuk angka dengan menggunakan perhitungan presentase (1-100%) atau dengan menggunakan bilangan antara 0-1. Sistem rekomendasi akan dianggap baik jika nilai *precision* and *recall*nya tinggi.

Sedangkan *F1* digunakan untuk representasi dari penggabungan antara *Precision* dan *Recal. F1* dapat dihitung menggunakan rumus dibawah ini:

$$F1 = 2PR / (P + R) \qquad (6)$$

Nilai *F1* merupakan tingkat akurasi terhadap sistem dalam memberikan rekomendasi yang diinginkan. Sistem akan diaanggap baik jika memiliki tingkat akurasi (*F1*) yang tinggi.

ROC *(Receivier Operating Characteristic) Curve* adalah grafik antara sensitifitas (*true positive rate*) pada sumbu Y dengan 1-spesifisitas pada sumbu X (*false positive rate*), seakan-akan menggambarkan tawar menawar antara sensitivitas dan spesifisitas, yang tujuannya adalah untuk menentukan *cut off point* pada uji diagnostic yang bersifat kontinyu [*14*]

Evaluasi pengukuran RapidMiner yaitu membandingkan nilai akurasi, nilai *precision*, dan nilai *recall* antara algoritma Decision Tree C4.5 dengan Algoritma Naive Bayes.

Validasi hasil penelitian dilakukan dengan mengambil sampel secara acak sebanyak 100 data mahasiswa. Data diuji dengan algoritma Decision Tree C4.5 dan Algoritma Naive Bayes, hasilnya akan dibandingkan.

III. HASIL

Uji pertama melalui data sample yaitu data angkatan 2008, Field data NRP, NAMA, Tempat Lahir dan Tanggal Lahir dihilangkan untuk mendapatkan akurasi yang lebih tinggi.

Selain itu, Untuk membentuk pohon keputusan maka atribut IPK Semester 1 dan IPK Semester 2 perlu di klasifikasi menjadi:

TABEL 3. Klasifikasi Nilai

NO	IPK SEMESTER	KLASIFIKASI
1	IPK>= 3,00	A
2	IPK>= 2,75	B
3	IPK< 2,75	C

Menghasilkan performance vector seperti terlihat dalam tabel 4. Menghasilkan akurasi yang di dapat pada uji pertama 93,31%. Nilai precision dan recall berturut-turut: 94,23% dan 92, 45%

TABEL 4. Perfomance Vector C4.5

	True Tidak sesuai	True Sesuai	Class Precision
Pred. Tidak sesuai	97	8	92,38%
Pred. Sesuai	6	98	94,23%
Class Recall	94,17%	92,45%	

Pada perhitungan nilai AUC didapatkan sebesar 0,961 (Gambar 2) dengan demikian maka klasifikasi keakuratan tes diagnostiknya termasuk dalam kategori sangat baik. Sedangkan data yang menghasilkan output "tidak sesuai", akan direkomendasikan jurusan baru.

Hasil perhitungan rekomendasi tanpa melibatkan field Umur, menunjukan hasil akurasi **76,64%**. Sedangkan dengan melibatkan field Umur, didapatkan bentuk pohon (Gambar 3) dengan akurasi **82,64%.** (Gambar 4)

Hasil Algoritma Decision Tree C4.5 dievaluasi dan dibandingkan dengan Algoritma Naive Bayes.

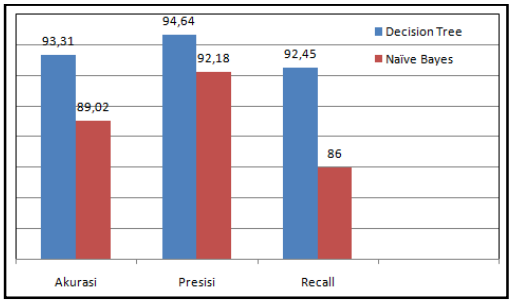
TABEL 5. Perfomance Vector Naïve Bayes

	True Tidak sesuai	True Sesuai	Class Precision
Pred. Tidak sesuai	95	15	96,36%
Pred. Sesuai	8	91	91,92
Class Recall	92,23%	85,85%	

Dari hasil eksperimen didapatkan hasil sebagaimana tabel 6:

TABEL 6 Perbandingan Validasi Hasil Prediksi Kesesuai Jurusan

	C4.5 (%)	Naive Bayes (%)
Akurasi	93,31	89,02
Presisi	94,23	91,92
Recall	92,45	85,85



GAMBAR 5. Grafik Perbandingan Hasil Prediksi Kesesuaian Jurusan

Kemudian pada model dilewatkan data uji untuk mendapatkan hasil Akurasi Rekomendasi. Algoritma Decision Tree C4.5 dievaluasi dan dibandingkan dengan Algoritma Naive Bayes. Hasilnya ditampilkan pada tabel 7:

TABEL 7 Perbandingan Hasil Validasi Prediksi Rekomendasi

	C4.5 (%)	Naive Bayes (%)
Akurasi	82,64	66,36

IV. KESIMPULAN

Dari hail uji, Algoritma Decision Tree C4.5 memprediksi lebih akurat dari pada Algoritma Naive Bayes dalam penentuan kesesuaian jurusan dan rekomendasi jurusan mahasiswa.

Dengan demikian dapat disiimpulkan bahwa Algoritma Decision Tree C4.5 akurat diterapkan untuk penentuan kesesuaian jurusan mahasiswa

dengan tingkat keakuratan 93,31 % dan akurasi rekomendasi jurusan sebesar 82,64%.

Dengan adanya penerapan Decision Tree C4.5 diharapkan mampu memberikan solusi bagi mahasiswa dan dapat membantu STMIK Indonesia dalam menentukan jurusan yang sesuai yang akan ditempuh oleh mahasiswa selama studi sehingga peluang untuk sukses dalam studi di perguruaan tinggi semakin besar

Namun terdapat beberapa hal yang perlu penulis sarankan bagi pengembangan penelitian ini antara lain, seperti dilakukan pengujian data dengan menambahkan beberapa kolom yang data yang baru. Dapat juga dilakukan perbandingan dengan metode algoritma lain yang mendukung pengujian data yang ada, sehingga bisa didapat tingkat akurasi yang lebih baik lagi.

REFERENSI

[1] Indri Savitri, M.Psi, Sekolahkan Anak Tanpa Memaksa, - Lembaga Psikologi Terapan Universitas Indonesia, 2006.

[2] Drs. H. Mulyadi, *Diagnosis Kesulitan Belajar*. Yogyakarta: Nuha Litera, 2010.

[3] Musrofi M, *Melesatkan Prestasi Akademik Siswa*. Yogyakarta: Pedagogia, 2010.

[4] Turban Efraim, Aronson Jay E, and Liang, *Decision Support Systems and Intelligent Systems*, 7th ed.: Prentice Hall, Upper Saddle River, NJ, 2005.

[5] Zhiwu Liu and Xiuzhi Zhang, "Prediction and Analysis for Students' Marks Based on Decision Tree Algorithm," in *2010 Third International Conference on Intelligent Networks and Intelligent Systems*, 2010.

[6] Larose Daniel T, *Discovering knowledge in data: An Introduction to Data Mining.*: Wiley Interscience, 2005.

[7] Rong Cao and Lizhen Xu, "Improved C4.5 Algorithm for the Analysis of Sales," in *2009 Sixth Web Information Systems and Applications Conference*, 2009.

[8] Nykänen Ossi, "Inducing Fuzzy Models for Student Classification," *Educational Technology & Society*, vol. 9(2), pp. 223-234, 2006.

[9] Chang Wen-Chih, "Integrating IRT to Clustering Student's Ability with K-Means," in *2009 Fourth International Conference on Innovative Computing, Information and Control*, 2009.

[10] Bahar, *Penentuan Jurusan Sekolah Menengah Atas Dengan Algoritma Fuzzy C-Means*. Semarang, Indonesia, 2011.

[11] Sumanto, *Penerapan Fuzzy C-Means dalam Pemilihan Peminatan Tugas Akhir*. Jakarta, Indonesia, 2010.

[12] Kusrini and Emha Taufiq Lutfhfi, *Algoritma Data Mining.*: ANDI Yogyakarta, 2009.

[13] Tan Pang-Ning, Michael Stienbach, and Vivin Kumar, *Introduction To Data Mining*, J. Taylor, Ed. Stanford, 2005.

[14] Florin Gorunescu, *Data Mining Concept Model Technique.*, 2011.

[15] Iwan Ariawan. (2011, Juni) Catatan materi kuliah dr Iwan Ariawan, MS. [Online]. http://www.scribd.com/doc/15123416/Kurva-Receiver-Operating-Characteristic

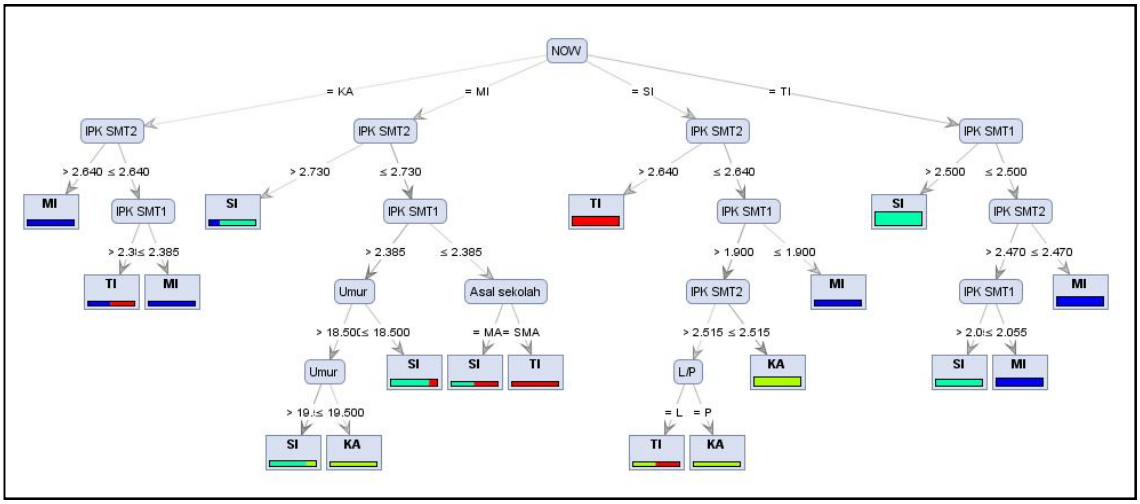
[16] Meller Thomas and all et, "New Classification Algorithms for Developing Online Program Recommendation Systems," in *2009 International Conference on Mobile, Hybrid, and On-line Learning*, 2009.

[17] Witten Ian H. and Eibe Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco: Morgan Kaufmann Publishers is an imprint of Elsevier, 2005.

[18] Jiawei Han, *Data Mining : Concepts and Techniques.*, 2006.

[19] Tom M. Mitchell, *Machine Learning*, McGraw Hill, Ed., 2005.

GAMBAR 2 Curva ROC



GAMBAR 3 Pohon keputusan rekomendasi jurusan dengan melibatkan 'Umur'

Table View Plot View

accuracy: 82.64% +/- 10.44% (mikro: 82.52%)

	true MI	true SI	true KA	true TI	class precision
pred. MI	26	1	1	1	89.66%
pred. SI	3	36	3	2	81.82%
pred. KA	0	1	12	2	80.00%
pred. TI	2	1	1	11	73.33%
class recall	83.87%	92.31%	70.59%	68.75%	

GAMBAR 4 Hasil perhitungan akurasi pohon keputusan rekomendasi jurusan dengan melibatkan 'Umur'