PERBANDINGAN PERFORMANSI ALGORITMA C4.5 DAN CART DALAM KLASIFIKSI DATA NILAI MAHASISWA PRODI TEKNIK KOMPUTER POLITEKNIK NEGERI PADANG

Indri Rahmayuni*
*Dosen Jurusan Teknologi Informasi
Politeknik Negeri Padang
rahmayuni@gmail.com

Abstrak

Dalam beberapa tahun terakhir, penggunaan *data mining* di dunia pendidikan yang dikenal sebagai *educational data mining* (EDM) semakin berkembang. Namun sebagian besar penggunaan itu dilakukan pada data yang berasal dari pendidikan berbasis web, komputer, dan *e-learning*. Padahal sebagian besar institusi pendidikan, terutama di negara-negara berkembang masih menggunakan sistem kelas tradisional. Data yang didapat dari kelas tradisional ini belum dieksploitasi dengan baik untuk memberikan dukungan dan bimbingan bagi siswa demi meningkatkan kualitas pendidikan.

Program studi Teknik Komputer merupakan salah satu program studi favorit di Politeknik Negeri Padang. Tahun pertama perkuliahan terutama semester pertama merupakan masa yang krusial bagi mahasiswa baru prodi Teknik Komputer.. Proses pendidikan di program studi Teknik Komputer didukung data hasil studi (nilai) mahasiswa tiap semesternya. Penggunaan *data mining* terhadap data hasil studi mahasiswa pada semester pertama diharapkan dapat memberikan pengetahuan mata kuliah apa saja yang paling krusial dalam menentukan kelulusan mahasiswa pada semester pertama..

Pada penelitian ini, *data mining* diterapkan menggunakan model proses CRISP-DM yang menyediakan proses standar penggunaan *data mining* pada berbagai bidang. Metode pohon keputusan (algoritma C4.5 dan CART) digunakan dalam klasifikasi karena hasil metode ini mudah dipahami dan diinterpretasikan. Kalkulus, Fisika, Algoritma dan Pemrogramaan, Pengantar TI, dan Praktek Dasar Pemorgraman merupakan mata kuliah paling krusial pada semester pertama.

Kata kunci : *educational data mining, klasifikasi, CART, C4.5*

1. PENDAHULUAN

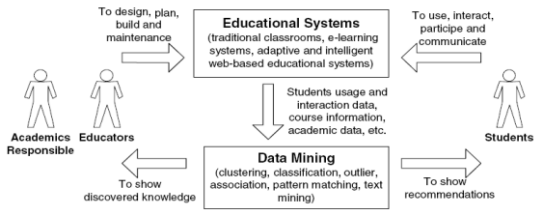
Pada era teknologi saat ini, data dan informasi menjadi bagian penting di berbagai bidang. Semua pihak berlomba mengumpulkan data dan informasi yang digunakan untuk mencapai kesuksesan. Awalnya, dengan munculnya komputer dan sarana penyimpanan data masal, data dikumpulkan dan disimpan dengan cepat. Sayangnya, koleksi-koleksi data tersebut dengan cepat menjadi sangat besar dan berlimpah. Dari data yang berlimpah ini, muncul pertanyaan mengenai hal-hal apa saja yang dapat dipelajari dari keseluruhan data dan informasi tersebut. Untuk menjawabnya dibutuhkan penyimpulan data secara otomatis, ekstraksi dari esensi informasi yang disimpan, serta penemuan pola yang ada dalam data. Proses ini dikenal sebagai *data mining*.

Program studi Teknik Komputer merupakan salah satu program studi baru yang berada dibawah Jurusan Teknologi Informasi Politeknik Negeri Padang. Program studi Teknik Komputer mulai dibuka pada tahun 2005. Walaupun baru berumur 8 tahun, program studi Teknik Komputer merupakan salah satu program studi favorit di Politeknik Negeri Padang.

Sebagai salah satu program studi favorit, Teknik Komputer harus terus melakukan perbaikan-perbaikan dalam sistem pendidikannya untuk mencapai kualitas yang lebih baik. Program studi Teknik Komputer perlu mengerahkan seluruh sumber daya yang dimiliki untuk membantu mahasiswa menyelesaikan pendidikan mereka dengan prestasi akademik yang baik dan meminimalisir tingkat ketidaklulusan mahasiswa. Salah satu caranya adalah dengan melakukan klasifikasi data nilai mahasiswa untuk mengetahui mata kuliah apa saja yang paling krusial pada semester pertama..

2. EDUCATIONAL DATA MINING

Komunitas *Educational Data mining* (EDM) pada www.educationaldatamining.org men-definisikan EDM sebagai sebuah disiplin ilmu yang sedang berkembang, dengan fokus pada pengembangan metode-metode untuk mengeksploitasi keunikan data yang berasal dari proses pendidikan dan menggunakan metode-metode tersebut untuk lebih memahami siswa serta sistem pembelajarannya.



Gambar 1. Alur Data mining pada Pendidikan [ROM07]

Romero dan Venture [ROM07] menggambarkan *data mining* pada sistem pendidikan (Gambar 2-1) sebagai suatu alur yang melibatkan tiga aktor yaitu pendidik dan penanggung jawab akademik sebagai pihak penyelanggara pendidikan serta siswa sebagai pengguna pendidikan. Melalui proses *data mining* terhadap sistem pendidikan, pendidik dan penanggung jawab pendidikan dapat mengetahui temuan/pengetahuan yang dihasilkan, sedangkan siswa mendapatkan rekomendasi terkait hasil tersebut.

Dari survei yang dilakukan Romero dan Ventura [ROM07], sebagian besar *data mining* pada dunia pendidikan dilakukan pada kelas berbasis web, pendidikan jarak jauh atau *e-learning*. Beberapa penelitian yang menerapkan *data mining* pada data pendidikan dari kelas tradisional memperlihatkan bahwa metode pohon keputusan merupakan metode yang paling banyak digunakan dan menghasilkan kualitas hasil yang lebih baik daripada metode lainnya [DEK09]. Hasil klasifikasi metode pohon keputusan juga lebih mudah dipahami dan diinterpretasikan.

2.1 Metode EDM

Dalam pengerjaannya, terdapat berbagai metode pada EDM. Metode-metode tersebut diambil dari berbagai literatur antara lain: *data mining* dan *machine learning*, *psychometrics* dan area statistik lainnya, *information visualization*, serta *computational modeling*. Ryan J. Baker [BAK*inpress*] membagi metode-metode EDM tersebut ke dalam lima kelompok:

(1) Prediksi
Terdapat dua tipe penggunaan utama prediksi dalam EDM. Pada beberapa kasus, metode prediksi dapat digunakan untuk mempelajari atribut apa yang paling penting dalam sebuah model prediksi. Pada tipe penggunaan kedua, metode prediksi digunakan untuk memprediksi keluaran data baru berdasarkan model prediksi yang dihasilkan.

(2) Pengklusteran
Pada EDM, pengklusteran banyak digunakan untuk kasus-kasus pengelompokan data seperti untuk menyelidiki persamaan dan perbedaan antara sekolah, untuk menyelidiki persamaan dan perbedaan antara siswa, atau untuk menyelidiki pola perilaku siswa.

(3) *Relationship Mining*
Tujuan *relationship mining* adalah untuk mengetahui hubungan antara variabel, dalam suatu kumpulan data yang memiliki sejumlah besar variabel. Hal ini untuk mengetahui variabel mana yang paling berhubungan dengan variabel tunggal acuan, atau mencoba untuk menemukan mana hubungan antara dua variabel yang paling kuat.

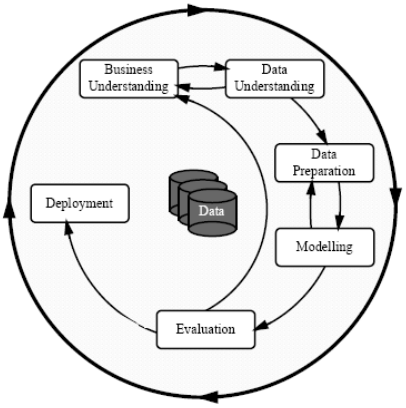
(4) Penyaringan Data Untuk Penilaian Manusia
Terdapat dua tujuan utama penyaringan data untuk penilaian manusia: identifikasi dan klasifikasi. Ketika data disaring untuk identifikasi, data ditampilkan dengan cara yang memungkinkan manusia untuk dapar mengidentifikasi pola yang ada dengan mudah.

(5) Penemuan Dengan Model
Dalam penemuan dengan model, model dari fenomena dikembangkan melalui prediksi, pengklusteran, atau dalam beberapa kasus menggunakan rekayasa pengetahuan. Model ini kemudian digunakan sebagai komponen dalam analisis lain, seperti prediksi atau *relationship mining*.

2.2 Cross-Industry Standard Process for Data Mining (CRISP-DM)

Data mining telah diterapkan di hampir seluruh bidang industri dan pengetahuan. Dengan semakin luasnya penerapan *data maning* tersebut, terdapat keinginan dari sekelompok analis *data mining* yang mewakili DaimlerChrysler, SPSS, dan NCR untuk membuat sebuah model proses *data mining* yang netral terhadap jenis industri, *tools*, dan aplikasi.

Para analis tersebut bergabung dan membangun *Cross-Industry Standard Process for Data mining* (CRISP-DM) pada tahun 1996. CRISP-DM menyediakan proses standar pelaksanaan *data mining* untuk menyelesaikan masalah dalam sebuah bisnis atau penelitian yang dapat digunakan semua orang [CHA00].



Gambar 2. Model Proses CRISP-DM [CHA00]

Berdasarkan CRISP-DM, sebuah proyek *data mining* merupakan sebuah siklus hidup yang terdiri atas enam fase (Gambar 2-2):

(1) Pemahaman Bisnis
Fase awal ini berfokus pada pemahaman tujuan dan kebutuhan proyek dari perspektif bisnis, kemudian mengubah pengetahuan ini ke dalam definisi masalah dan desain rencana awal *data mining* untuk mencapai tujuan proyek.

(2) Pemahaman Data
Fase pemahaman data dimulai dengan pengumpulan data awal dan dilanjutkan dengan aktifitas-aktifitas lain untuk mengenal data, mengidentifikasi permasalahan kualitas data, atau untuk mendeteksi subset data yang menarik untuk membentuk hipotesis bagi informasi yang tersembunyi.

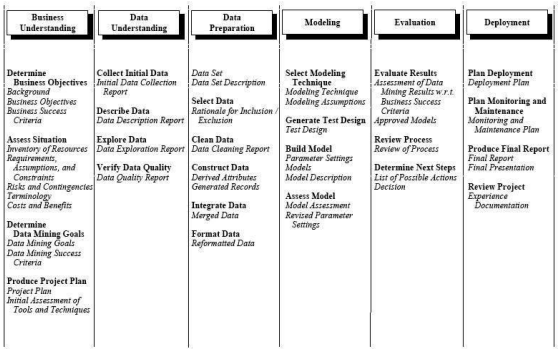
(3) Persiapan Data
Fase persiapan data meliputi seluruh aktifitas yang dilakukan untuk membangun dataset akhir (data yang akan digunakan sebagai masukan bagi aplikasi pemodelan) dari data mentah awal. Proses persiapan data biasanya dilakukan berulang kali untuk memastikan kualitas data telah dicapai. Aktifitas persiapan data antara lain pemilihan tabel, *record*, dan atribut, serta pembersihan dan transformasi data.

(4) Pemodelan
Pada fase ini, berbagai jenis teknik pemodelan dipilih dan diaplikasikan serta parameter-parameternya dikalibrasi untuk mendapatkan hasil yang optimal. Biasanya terdapat beberapa teknik untuk jenis permasalahan *data mining* yang sama. Beberapa teknik juga memiliki kebutuhan akan bentuk data yang spesifik. Oleh karena itu, seringkali proses persiapan data dibutuhkan kembali.

(5) Evaluasi
Pada tahap ini, model yang tampak memiliki kualitas tinggi dari perspektif analisis data telah dihasilkan. Sebelum dilanjutkan ke tahap penerapan, model yang dihasilkan dievaluasi dan di-review tiap langkah pembuatannya untuk memastikan model tesebut telah mencapai tujuan bisnis dengan tepat. Tujuan utumanya adalah untuk menentukan apakah terdapat beberapa permasalahan bisnis yang tidak dicakup dengan baik. Pada akhir fase ini, keputusan mengenai pengaplikasian hasil *data mining* harus dapat dicapai.

(6) Penerapan
Pekerjaan yang dilakukan pada fase penerapan ini tergantung pada kebutuhan dan tujuan proyek *data mining*, dari yang paling sederhana seperti pembuatan laporan, hingga yang paling kompleks seperti pengimplementasian proses *data mining* ke dalam sistem organisasi.

Panduan CRISP-DM yang dikeluarkan oleh Konsorsium CRIPS-DM [CHA00] memberikan contoh-contoh pekerjaan yang dapat dilakukan untuk tiap fasenya beserta hasil keluaran yang didapatkan (Gambar 2-3). Contoh pekerjaan dan keluaran ini merupakan contoh secara umum dimana penerapannya tergantung pada jenis industri dan kebutuhan serta tujuan *data mining*.



Gambar 3. Contoh Pekerjaan dan Keluaran Fase CRISP-DM [CHA00]

2.3 Algoritma C4.5

Algoritma C4.5 merupakan algoritma yang digunakan untuk membangun sebuah pohon keputusan (*decision tree*) dari data. Algoritma C4.5 merupakan pengembangan dari algoritma ID3 yang juga merupakan algoritma untuk membangun sebuah pohon keputusan. Algoritma C4.5 secara rekursif mengunjungi tiap simpul keputusan, memilih percabangan optimal, sampai tidak ada cabang lagi yang mungkin dihasilkan [LAR05].

Algoritma C4.5 menggunakan konsep *information gain* atau *entropy reduction* untuk memilih percabangan yang optimal. Misalkan terdapat sebuah variabel X dimana memiliki sejumlah k nilai yang mungkin dengan probabilitas p*1*, p*2*, …, p*k*. *Entropy* menggambarkan keseragaman data dalam variabel X. *Entropy* variabel X (H(X)) dihitung dengan menggunakan persamaan 2.1.

$$H(X) = -\sum_{j} p_j \log_2(p_j) \quad (2.1)$$

Misalkan terdapat sebuah kandidat simpul yang akan dikembangkan (S), yang membagi data T ke dalam sejumlah subset T_1, T_2, …, T_k. Dengan menggunakan persamaan *entropy* diatas, nilai *entropy* tiap subset dihitung ($H_S(T_i)$). Kemudian total bobot subset simpul S dihitung dengan menggunakan persamaan 2.2.

$$H_s(T) = \sum_{i=1}^{k} p_i \, H_s(T_i) \qquad (2.2)$$

dimana P_i merupakan proporsi *record* pada subset i. Semakin seragam sebuah subset terhadap kelas-kelas pembaginya, maka semakin kecil nilai *entropy*. Nilai *entropy* paling kecil adalah 0, yang dicapai ketika *record* subset berada pada satu kelas yang sama. Sedangkan nilai *entropy* paling tinggi adalah 1, yang dicapai ketika *record* subset terbagi sama rata pada untuk tiap kelas. Semakin kecil nilai *entropy*, semakin baik subset tersebut.

Dari nilai-nilai *entropy* yang didapat, nilai *information gain* untuk simpul S dihitung melaui persamaan 2.3.

$$\text{gain}(S) = H(T) - H_S(T) \qquad (2.3)$$

Pada algoritma C4.5, nilai *information gain* dihitung untuk seluruh simpul yang mungkin dikembangkan. Simpul yang dikembangkan adalah simpul yang memiliki nilai *information gain* yang paling besar.

2.4 Algoritma CART

Metode CART ini pertama kali diajukan oleh Leo Breiman et al. pada tahun 1984. Pohon keputusan yang dihasilkan CART merupakan pohon biner dimana tiap simpul wajib memiliki dua cabang. CART secara rekursif membagi *record*s pada data latihan ke dalam subset-subset yang memiliki nilai atribut target (kelas) yang sama.

Algoritma CART mengembangkan pohon keputusan dengan memilih percabangan yang paling optimal bagi tiap simpul. Pemilihan dilakukan dengan menghitung segala kemungkinan pada tiap variabel.

Misalkan $\Phi(s|t)$ merupakan nilai "kebaikan" kandidat cabang s pada simpul t, maka nilai $\Phi(s|t)$ dapat dihitung sebagai (persamaan 2.4) [LAR05]:

$$\emptyset(s|t) = 2P_L P_R \sum_{j=1}^{\#kelas} |P(j|t_L) - P(j|t_R)| \qquad (2.4)$$

dimana

t_L = *simpul anak kiri dari simpul t*

t_R = *simpul anak kanan dari simpul t*

$$P_L = \frac{\textit{jumlah record pada } t_L}{\textit{jumlah seluruh record pada data latihan}}$$

$$P_R = \frac{\textit{jumlah record pada } t_R}{\textit{jumlah seluruh record pada data latihan}}$$

$$P(j|t_L) = \frac{\textit{jumlah record kelas j pada } t_L}{\textit{jumlah record pada simpul t}}$$

$$P(j|t_R) = \frac{\textit{jumlah record kelas j pada } t_R}{\textit{jumlah record pada simpul t}}$$

Nilai $\sum_{j=1}^{\#kelas} |P(j|t_L) - P(j|t_R)|$ maksimal ketika *record* yang berada pada cabang kiri atau kanan simpul memiliki kelas yang sama (seragam). Nilai maksimal yang dicapai sama dengan jumlah kelas pada data. Misalkan jika data terdiri atas dua kelas, maka nilai maksimal $\sum_{j=1}^{\#kelas} |P(j|t_L) - P(j|t_R)|$ adalah 2.

Semakin seragam *record* pada cabang kiri atau kanan, maka semakin tinggi nilai $\sum_{j=1}^{\#kelas} |P(j|t_L) - P(j|t_R)|$. Nilai maksimal $2P_L P_R$ sebesar 0.5 dicapai ketika cabang kiri dan kanan memiliki jumlah *record* yang sama. Kandidat percabangan yang dipilih adalah kandidat yang memiliki nilai $\Phi(s|t)$ paling besar.

2.5 Penelitian Terkait

Saat ini penelitian terhadap *data mining* dan sistem pendidikan semakin banyak dilakukan. Penelitian mengenai *data mining* di dunia pendidikan telah lama ada (sejak tahun 1990an) dan baru dikelompokkan menjadi sebuah bidang penelitian Educational *Data mining* pada tahun 2005 ketika sekelompok peneliti *data mining* membuat organisasi penelitian *data mining* di dunia pendidikan yang dapat diakses di www.educationaldatamining.org.

Mulai tahun 2008, organisasi ini mengadakan konferensi tahunan EDM yang membahas penelitian-penelitian *data mining* di dunia pendidikan di seluruh dunia.

Sebagaimana dijelaskan sebelumnya, penelitian terkait prediksi pada EDM semakin banyak dilakukan, salah satunya mengenai prestasi akademik siswa. Beberapa diantaranya adalah:

(1) Jing Luan [LUA02] melakukan penelitian di beberapa univeristas di Amerika Serikat untuk memprediksi siswa *community college* yang memenuhi syarat untuk pindah ke universitas. Model yang dihasilkan ditujukan untuk menyediakan pola profil siswa berdasarkan data demografi, finansial, pelajaran yang diambil dan nilai siswa. Penelitian ini menggunakan algoritma *neural networks*, C4.5, dan CART.

(2) Erdogan dan Timor [ERD05] melakukan penelitian terhadap mahasiswa di Univeristas Maltepe Turki untuk mengetahui hubungan antara hasil ujian masuk universitas dengan kesuksesan mereka dalam proses perkuliahan. Penelitian ini menggunakan algoritma pengklusteran K-means.

(3) Gérard Lassibille dan Lucía Navarro Gómez [LAS07] melakukan penelitian terhadap 7000

mahasiswa universitas-universitas di Spanyol untuk mengetahui faktor utama yang mempengaruhi ketidaklulusan (*drop out*) mereka. Penelitian ini menunjukkan bahwa jenis kelamin (hanya di universitas teknik), umur ketika masuk, nilai ujian masuk, jenis SMU, sumber biaya kuliah, pendidikan orang tua, serta status tempat tinggal berpengaruh terhadap ketidaklulusan mahasiswa di Spanyol.

(4) Gerben W. Dekker [DEK09] melakukan penelitian untuk memprediksi ketidaklulusan mahasiswa (*drop out*) tahun pertama di Departemen Teknik Elektro Universitas Teknologi Eindhoven karena tingkat ketidaklulusan yang mencapai 40%. Data nilai akademik mahasiswa digunakan dalam penelitian ini. Model proses CRISP-DM dipakai sebagai acuan pelaksanaan penelitian dengan menggunakan algoritma C4.5 dan CART untuk melakukan prediksi. Dari penelitian ini diketahui bahwa nilai Aljabar Linier, Kalkulus, Jaringan, serta nilai rata-rata mata pelajaran IPA di SMU menjadi faktor penentu utama ketidaklulusan mahasiswa.

3. Data Yang Digunakan

Penelitian ini menggunakan data nilai semester pertama mahasiwa Program Studi Teknik Komputer Politeknik Negeri Padang dari angkatan 2006 sampai 2010. Tiap angkatan terdiri atas ±90 orang mahasiswa yang terdiri atas mahasiswa undangan (PMDK) dan mahasiswa jalur ujian masuk dengan latar belakang pendidikan SMU IPA dan SMK Teknik.

Data nilai semester pertama mahasiswa angkatan 2006-2010 didapat dari program studi Teknik Komputer PNP dengan format Ms.Excel. Pada semester pertama, mahasiswa harus mengambil sebelas mata kuliah yang terdiri atas mata kuliah teori dan praktikum (Tabel III-4) dengan nilai SKS yang berbeda. Selain nilai tiap mata kuliah, data nilai mahasiswa juga memiliki atribut indeks prestasi (IP) serta atribut status kelulusan pada semester terkait, apakah lulus, tidak lulus, atau lulus percobaan.

Tabel 1. Daftar Mata Kuliah Semester Pertama

No	Mata Kuliah	SKS
1	Kalkulus	2
2	Matematika Diskrit	2
3	Fisika	2
4	Bahasa Inggris 1	2
5	Pendidikan Agama	2
6	Pengantar Teknologi Informasi	3
7	Algoritma & Pemrograman	3
8	Prak. Dasar Pemrograman	1
9	Dasar Elektronika	3
10	Prak. Dasar Elektronika	1
11	Prak. Aplikasi Komputer	1

Nilai tiap mata kuliah terdiri atas tiga bentuk yaitu nilai angka, nilai huruf, nilai bobot. Selain itu juga diambil data nilai indeks prestasi (IP) mahasiswa pada semester pertama.

3.1 Persiapan Data

Data nilai yang digunakan adalah data nilai semester pertama mahasiswa program studi Teknik Komputer Politeknik Negeri Padang dari angkatan 2006-2010. Data ini digunakan sebagai acuan dalam penentuan status kelulusan mahasiswa dan telah diverifikasi baik oleh pihak program studi Teknik Komputer maupun bagian akademik Politeknik Negeri Padang, maka dapat dipastikan bahwa data ini memiliki kualitas yang sangat baik.

Setelah atribut nilai mata kuliah dari angkatan 2006 sampai 2010 memiliki format yang sama, dilakukan pula pengelompokan nilai mata kuliah untuk memperkecil variansi nilai atribut. Pengelompokan alternatif yang dilakukan adalah dengan mambagi nilai mata kuliah dikelompokkan ke dalam 4 kelompok *Best* (A, A-), *Good* (B+, B, B-), *Pass* (C+, C, C-), dan *Fail* (D, E).

3.2 Pendefinisian Atribut Kelas

Atribut kelas yang digunakan adalah atribut kelas yang didefinisikan secara manual. Atribut kelas dibuat dengan mengelompokkan nilai indeks prestasi semester satu mahasiswa yang diambil dari data nilai mahasiswa. Pada penelitian ini digunakan atribut kelas manual dengan mengelompokkan IP atas dua kelompok yaitu Atas dan Bawah yang mewakili posisi IP mahasiswa terhadap nilai IP 2.84. Nilai 2.84 ini didapat dengan menggunakan metode distribusi normal.

4. Klasifikasi dan Hasil

Klasifikasi dilakukan menggunakan aplikasi WEKA Explorer. Proses klasifikasi dilakukan terhadap data nilai mahasiswa dengan pengelompokkan IP sebagai atribut kelasnya. Klasifikasi dilakukan menggunakan algoritma CART dan C4.5. Proses klasifikasi dilakukan menggunakan metode *cross-validation*. Hasil klasifikasi ditampilkan pada Tabel 2.

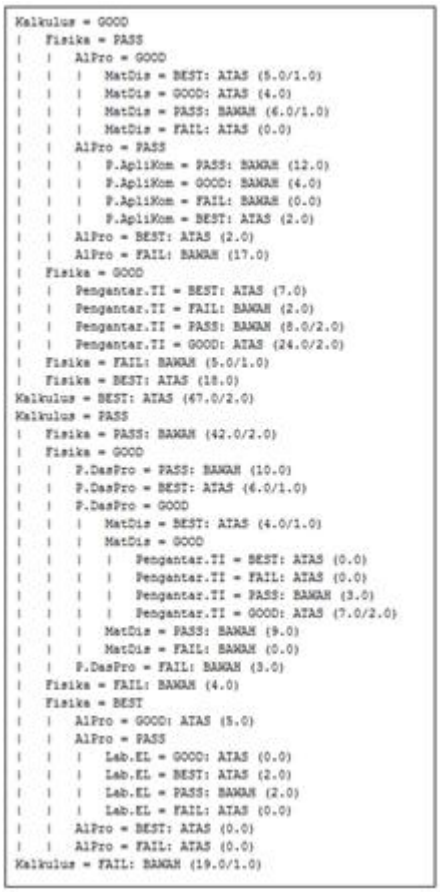
Tabel 2. Hasil Klasifikasi Data

Data	Training	
Algoritma	C4.5	CART
Akurasi	85.61%	84.95%
Recall	0.848	0.848
Precision	0.865	0.853
F-measure	0.856	0.850
Atribut	Kalkulus	Alpro
	Fisika	Fisika
	Alpro	Das.EL
	PengantarTI	Kalkulus
	P.DasPro	MatDis
		Pengantar.TI

Pada tabel 4.2 tersebut dapat diketahui bahwa algoritma C4.5 memberikan akurasi paling baik (85.61%), sedangkan algoritma CART memberikan hasil sedikit dibawahnya (84.95%). Hal ini terjadi karena algoritma C4.5 membangun pohon dengan jumlah cabang tiap simpul sesuai dengan jumlah nilai simpul tersebut. Selain itu algoritma C4.5 lebih cocok digunakan untuk data yang bersifat non-numerik seperti data nilai mahasiswa yang dikelompokkan kedalam empat kelompok (Best, *Good*, *Pass*, dan *Fail*). Berbeda dengan algoritma CART dengan konsep pohon biner lebih cocok digunakan untuk data yang bersifat numerik.

Gambar 4.1 memperlihatkan pohon keputusan yang dihasilkan oleh algoritma C4.5. Dari gambar ini dapat dilihat atribut-atribut yang digunakan dalam klasifikasi dan posisinya.

Dari hasil klasifikasi tersebut didapatkan bahwa mata kuliah Kalkulus, Fisika, Algoritma dan Pemrograman, Pengantar Teknologi Informasi, dan Praktek Dasar Pemrograman merupakan mata kuliah yang paling krusial pada semester pertama perkuliahan di Program Studi Teknik Komputer Politeknik Negeri Padang.



Gambar 4. Pohon Keputusan Klasifikasi Data Nilai Mahasiswa Menggunakan Algoritma C4.5

5. Kesimpulan

Dari penelitian yang dilakukan, dapat disimpulkan beberapa hal :

(1) Algoritma C4.5 memberikan akurasi yang lebih baik dari pada algoritma CART dalam klasifikasi data nilai mahasiswa.
(2) Algoritma C4.5 memberikan hasil lebih baik karena data nilai mahasiswa berupa data kelompok yang cocok dengan sifat klasifikasi algoritma C4.5
(3) Algoritma CART memberikan hasil dibawah C4.5 karena CART lebih cocok digunakan untuk data berjenis numerik.
(4) Kalkulus, Fisika, Algoritma dan Pemrograman, Pengantar Teknologi Informasi, dan Praktek Dasar Pemrograman merupakan mata kuliah yang paling krusial pada semester pertama perkuliahan di Program Studi Teknik Komputer Politeknik Negeri Padang.

6. Daftar Pustaka

[BAK*inpress*] Baker, R.S.J.d. (in press). *Data mining for Education*. To appear in McGraw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition). Oxford, UK:Elsevier.

[CHA00] Chapman, P., et.al. (2000). *CRISP-DM 1.0: Step-by-Step Data mining Guide*. CRISP-DM Consortium.

[DEK09] Dekker, W. Gerben., et.al. (2009). *Predicting Students Drop Out: A Case Study*. Proceedings of the 2nd International Conference on Educational *Data mining*. 41-50.

[ERD05] Erdogan, S.Z, Timor, M. (2005). *A Data mining Applications in Student Database*. Journal of Aeronautics and Space Technologies. Vol 2(2). 53-57.

[LAR05] Larose, D.T. (2005). *Discovering Knowledge in Data: An Introduction to Data mining*. Wiley Interscience. Ney Jersey.

[LAS07] Lassibille, G., Gomez, L. N. (2007). *Why Do Higher Education Students Drop Out? Evidence from Spain*. Education Economics. Vol 16(1). 89-105.

[LUA02] Luan, J. (2002). *Data mining and Its Applications in Higher Education.* New Directions for Institutional Research. Vol 133. 17-36.

[ROM07] Romero, C., Ventura, S. (2007). *Educational data mining: A survey from 1995 to 2005*. Expert System with Application. Vol 33. 135-146.