PENERAPAN KLASIFIKASI DENGAN ALGORITMA CART UNTUK PREDIKSI KULIAH BAGI MAHASISWA BARU

Mardiani
Jurusan Sistem Informasi, STMIK MDP Palembang
Jln. Rajawali No.14 Palembang 30113
Telp. (0711) 376400, Faks. (0711) 376360
E-mail: mardiani@stmik-mdp.net

ABSTRAK
Perguruan tinggi swasta, umumnya beresiko memiliki mahasiswa baru yang masih mempunyai kemungkinan tidak melanjutkan kuliah untuk seterusnya. Hal ini disebabkan oleh berbagai alasan, diantaranya adalah diterimanya mahasiswa tersebut di perguruan tinggi lain, kondisi dan lingkungan perguruan tinggi di perguruan tinggi tersebut, nilai yang didapat pada semester 1 dan masih banyak sebab lain. Penelitian ini menggunakan fungsionalitas data mining klasifikasi untuk memprediksi mahasiswa mana yang akan bertahan terus kuliah dan yang mana yang tidak meneruskan kuliah pada semester berikutnya. Atribut yang terukur, baik yang numeric maupun nominal, diambil dari nilai IPK pada semester 1, grade kelompok penilaian saat ujian masuk dan dari gelombang ujian saringan masuk yang mana, yang diikuti mahasiswa baru. Algoritma yang dipakai adalah algoritma klasifikasi CART, dari data training akan ditentukan hasilnya didata testing, dan dari hasil tersebut, bisa didapat pola bagi mahasiswa baru yang kemungkinan akan melanjutkan kuliah di semester 2 atau tidak. Hasilnya didapatkan bahwa penyebab terbesar, mahasiswa tidak melanjutkan kuliah pada semester berikutnya adalah nilai IPK yang dibawah 1.

Kata Kunci: klasifikasi, CART, mahasiswa

1. PENDAHULUAN

Berbeda dengan perguruan tinggi negeri, perguruan tinggi swasta mendapatkan mahasiswa, dengan harus berusaha mencari mahasiswanya sendiri. Kemudian setelah mendapatkan mahasiswa, perguruan tinggi swasta masih memiliki kemungkinan yang tinggi, kehilangan mahasiswanya tersebut. Ini misalnya, terjadi pada mahasiswa yang baru saja masuk kuliah, kemudian berhenti atau pindah tempat kuliah dengan berbagai alasan. Alasan-alasan tersebut, ada yang tidak bisa diukur, misalnya karena diterima di perguruan tinggi lain, dan ada yang bisa diukur, misalnya dari besarnya uang pembayaran masing-masing mahasiswa.

Atribut pertama dalam memprediksi kelanjutan kuliah dari mahasiswa baru, bisa diambil dari gelombang USM mana mahasiswa tersebut ikut. Hal ini dipertimbangkan karena biasanya mahasiswa yang ikut USM gelombang pertama lebih serius untuk kuliah, dibanding yang ikut gelombang terakhir yang biasanya banyak mempunyai pilihan, juga tes di tempat lain. Atribut kedua adalah dari besarnya pembayaran setiap mahasiswa yang ditentukan dari *grade* hasil tes. Semakin baik hasil tes, maka akan semakin kecil biaya yang dikeluarkan mahasiswa baru. Kemudian atribut ketiga diambil dari nilai IPK yang didapat pada semester satu. Hal ini dikarenakan, jika hanya menggunakan dua atribut diatas, maka mahasiswa tersebut belum memiliki NPM dan belum merasakan lingkungan dan suasana kuliah yang sesungguhnya.

Dari alasan tersebut diatas, maka penelitian ini memprediksi kelanjutan aktif atau tidaknya mahasiswa dari semester satu ke semester dua, dengan data yang diambil adalah dari salah satu perguruan tinggi swasta, untuk seluruh mahasiswa Jurusan Sistem Informasi, dengan data *training* menggunakan data mahasiswa angkatan 2010 dan data *testing* menggunakan data mahasiswa angkatan 2011.

Penelitian ini bertujuan untuk membantu manajemen untuk mengetahui pola, penyebab berkurangnya jumlah mahasiswa pada semester awal, guna mengantisipasi, mencegah dan berusaha mengurangi jumlah mahasiswa yang berhenti tersebut.

2. TINJAUAN PUSTAKA

2.1 *Data mining*

Menurut Nafisah (2008), kebutuhan manusia akan data dan informasi tidak dapat dipungkiri. Bahkan, sekarang melalui dunia teknologi, arus informasi dapat beredar dengan cepat dan mudah. Data perlu diorganisasikan dan dikontrol menjadi informasi agar lebih mudah difahami. Pengolahan data menjadi informasi tersebut perlu dilakukan secara hati-hati agar informasi yang dihasilkan memiliki kualitas yang baik. Seiring dengan hal ini maka algoritma *data mining* juga turut berkembang pada basis data yang besar. *Data mining* adalah suatu teknologi untuk mengekstrak pengetahuan atau yang dikenal sebagai informasi dari kumpulan data, sehingga hasilnya bisa dipergunakan untuk pengambilan keputusan.

Pengelompokan *Data mining* dibagi menjadi beberapa kelompok, menurut Kusrini dan Luthfi (2009):

a. Deskripsi, yang merupakan cara untuk menggambarkan pola dan kecendrungan yang terdapat dalam data yang dimiliki.
b. Estimasi, yang hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik daripada ke arah kategori. Model yang dibangun menggunakan *record* lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi.
c. Prediksi, prediksi menerka sebuah nilai yang belum diketahui dan juga memperkirakan nilai untuk masa mendatang.
d. Klasifikasi, terdapat target variabel kategori, misal penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu tinggi, sedang dan rendah.
e. Pengklusteran, yang merupakan pengelompokan *record*, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan.
f. Asosiasi, yang bertugas menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja.

2.2 Klasifikasi

Seiring dengan perkembangan pengetahuan *data mining* dan komponen-komponennya, *data mining* tidak lagi dimonopoli oleh bidang teknologi informasi. Pemakaiannya telah semakin meluas ke bidang lain misalnya pada bidang kesehatan, pertanian, asuransi dan lain-lain.

Khusus untuk fungsionalitas *data mining* menggunakan klasifikasi, beberapa penelitian telah di lakukan, misalnya oleh Widagdo (2010) yang menggunakan pembentukan pohon klasifikasi biner untuk studi kasus penyakit diabetes, penelitian Wibowo dan Purwarianti (2011) yang menerapkan *bagging* untuk memperbaiki hasil prediksi nasabah perusahaan asuransi dan penggunaan algoritma klasifikasi lainnya yang misalnya dilakukan oleh Sinambela (2008) pada data status daerah kabupaten di Indonesia.

. Penelitian-penelitian yang telah dilakukan berkaitan dengan klasifikasi *data mining*, ada yang menggunakan aplikasi yang sudah tersedia dan ada juga yang membuat sendiri aplikasi baru yang sesuai dengan algoritma yang dibuat. Kemudian juga ada yang membandingkan antara dua algoritma klasifikasi yaitu CART dan CHAID oleh Kadir (2007) untuk menentukan hasil kredit, serta algoritma CART dan MARS oleh Otok (2005) untuk klasifikasi kasus perbankan. Penelitian perbandingan seperti ini akan menghasilkan logika dari algoritma mana yang nantinya akan menghasilkan kesimpulan terbaik terhadap kasus masing-masing penelitian.

2.3 Algoritma CART

Menurut Susanto dan Suryadi (2010), pada klasifikasi algoritma CART (*Classification and Regression Trees*), sebuah *record* akan diklasifikasikan ke dalam salah satu dari sekian klasifikasi yang tersedia pada variabel tujuan berdasarkan nilai-nilai variabel prediktornya.

Langkah-langkah Algoritma CART:

a. Susunlah calon cabang (*candidate split*) yang dilakukan terhadap seluruh variabel prediktor. Daftar yang berisi calon cabang disebut calon cabang mutakhir.
b. Berikan penilaian keseluruhan calon cabang mutakhir dengan menghitung besaran $\Phi(s|t)$
c. Tentukan cabang yang memiliki kesesuaian $\Phi(s|t)$. Setelah noktah keputusan tidak ada lagi, algoritma CART dihentikan.

Kesesuaian (*goodness*) $\Phi(s|t)$ dari calon cabang s pada noktah keputusan t, didefiniskan sebagai persamaan-persamaan berikut:

$$\Phi(s|t) = 2P_L P_R Q(s|t) \quad (1)$$

$$Q(s|t) = \sum_{j=1}^{\text{jumlahkategori}} \left| P(j|t_L) - P(j|t_R) \right| \quad (2)$$

t_L = cabang kiri dari noktah keputusan t
t_R = cabang kanan dari noktah keputusan t

$$P_L = \frac{\text{calon cabang kiri } t_L}{\text{data latihan}} \quad (3)$$

$$P(j|t_L) = \frac{\text{j calon cabang kiri } t_L}{\text{noktah keputusan } t} \quad (4)$$

$$P_R = \frac{\text{calon cabang kanan } t_R}{\text{data latihan}} \quad (5)$$

$$P(j|t_R) = \frac{\text{j calon cabang kanan } t_R}{\text{noktah keputusan } t} \quad (6)$$

3. METODOLOGI PENELITIAN

Langkah-langkah penelitian yang dilakukan adalah perumusan masalah, penentuan teknik yang akan dipergunakan, preproses data, transformasi data, analisis hasil, dan penarikan kesimpulan.

Preposes data dilakukan, karena data yang didapatkan merupakan data yang masih berantakan dan harus diolah lagi terlebih dahulu, sebelum memasuki proses data selanjutnya. Setelah data di transformasi atau diolah, kemudian masuk kepada tahapan analisis data. Dari data *training* yang telah dianalisis, kemudian dibuat prediksi klasifikasi untuk kemungkinan-keungkinan yang akan datang, bagi data *testing* jenis yang sama.

4. HASIL DAN PEMBAHASAN

4.1 Pembersihan Data

Data *training* yang diambil berjumlah 365 yaitu dari seluruh data mahasiswa Jurusan Sistem Informasi angkatan 2010 yang telah melewati semester 2, sementara data *testing* berjumlah 188 yaitu dari seluruh data mahasiswa Jurusan Sistem Informasi angkatan 2011 yang baru saja masuk semester 2. Data-data tersebut didapat dari bagian akademik dalam bentuk *database* yang masih harus melalui proses *cleansing*. Proses tersebut kemudian mendapatkan data mahasiswa dengan 5 atribut yang akan dianalisis yaitu NPM, IPK semester 1 (0 sampai 4), *grade* masuk (A,B, dan C), gelombang USM (I,II,III,IV, dan V) dan keaktifan selanjutnya (ya dan tidak). Untuk data *training*, seluruh data tersebut telah terisi, namun untuk data *testing*, atribut keaktifan selanjutnya masih kosong dan diisi tanda tanya.

4.2 Kesesuaian Calon Cabang dan Noktah Keputusan

Pada data *training*, untuk menentukan puncak pohon keputusan, diambil tiga atribut utama yaitu IPK, *Grade* dan Gelombang. Kemudian data-data tersebut disusun menjadi calon cabang (*candidate split*) terhadap seluruh variabel prediktor secara lengkap, sehingga terbentuk daftar calon cabang mutakhir seperti pada tabel berikut ini:

Tabel 1. Tabel Daftar Calon Mutakhir

Nama Calon Cabang	*Calon Cabang Kiri*	*Calon Cabang Kanan*
1	IPK ≤1	IPK >1
2	IPK ≤2	IPK >2
3	IPK ≤3	IPK >3
4	Grade=A	Grade=(B,C)
5	Grade=B	Grade=(A,C)
6	Grade=C	Grade=(A,B)
7	Gel=I	Gel=(II,III,IV,V)
8	Gel=II	Gel=(I,III,IV,V)
9	Gel=III	Gel=(I,II,IV,V)
10	Gel=IV	Gel=(I,II,III,V)
11	Gel=V	Gel=(I,II,III,IV)

Kemudian dihitung nilai *candidate split purity left* P_L dan *purity right* P_R menggunakan persamaan (3) dan (5) sebagai berikut:

Tabel 2. Tabel Perhitungan P_L dan P_R

No	P_L	P_R
1	0,096	0,904
2	0,260	0,740
3	0,534	0,466
4	0,685	0,315
5	0,249	0,751
6	0,066	0,934
7	0,132	0,868
8	0,205	0,795
9	0,137	0,863
10	0,216	0,784
11	0,310	0,690

Selanjutnya dengan persamaan (4) dan (6) dihitung $P(j|t_L)$ dan $P(j|t_R)$ untuk kemungkinan aktif dan yang tidak aktif sebagai berikut:

Tabel 3. Tabel Perhitungan $P(j|t_L)$ dan $P(j|t_R)$

No	*Aktif*	$P(j\|t_L)$	$P(j\|t_R)$
1	ya	0,171	0,930
	tidak	0,829	0,070
2	ya	0,811	0,874
	tidak	0,189	0,126
3	ya	0,974	0,724
	tidak	0,026	0,276
4	ya	0,860	0,852
	tidak	0,140	0,148
5	ya	0,857	0,858
	tidak	0,143	0,142
6	ya	0,833	0,859
	tidak	0,167	0,141
7	ya	0,938	0,845
	tidak	0,063	0,155
8	ya	0,867	0,855
	tidak	0,133	0,145
9	ya	0,900	0,851
	tidak	0,100	0,149
10	ya	0,823	0,867
	tidak	0,177	0,133
11	ya	0,823	0,873
	tidak	0,177	0,127

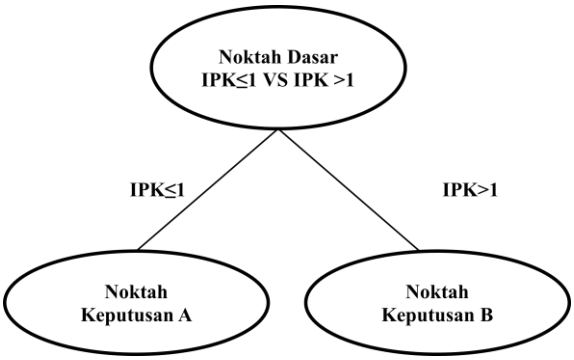
Dari daftar diatas, kemudian barulah dihitung nilai kesesuaian (*goodness*) untuk calon cabang $\Phi(s|t)$, dengan hasil perhitungan cabang pertama seperti ditunjukkan pada tabel berikut:

Tabel 4. Tabel Kesesuaian untuk Calon Cabang

No	$2P_LP_R$	$Q(s\|t)$	$\Phi(s\|t)$
1	0,173	1,518	0,263
2	0,385	0,127	0,049
3	0,498	0,502	0,250
4	0,432	0,016	0,007
5	0,374	0,001	0,000
6	0,123	0,052	0,006
7	0,228	0,184	0,042
8	0,327	0,023	0,008
9	0,236	0,098	0,023
10	0,339	0,089	0,030
11	0,427	0,100	0,043

 ISSN: 1907-5022

Hasil perhitungan kesesuaian (*goodness*) Φ(s|t) untuk calon cabang, menunjukkan bahwa calon cabang yang tertinggi nilai besarannya adalah nomor calon cabang 1 sebesar 0,263, yaitu cabang kiri IPK≤1 dan cabang kanan IPK>1, maka berarti calon cabang inilah yang dipilih sebagai *root node* pada tahap ini. Karena cabang selanjutnya IPK≤1 dan juga IPK>1 belum memberikan satu noktah keputusan, yaitu dua-duanya masih memiliki anggota dengan dua pilihan ya dan tidak, maka kedua cabang ini nantinya akan bercabang lagi. Dari hasil kesesuaian diatas juga terlihat bahwa, atribut pertama yaitu IPK memiliki nilai besaran yang lebih tinggi dibanding kedua atribut lainnya, yaitu *grade* dan gelombang. Cabang pertama dari hasil perhitungan diatas, seperti ditunjukkan pada gambar berikut:



Gambar 1. Pohon Keputusan Cabang Pertama

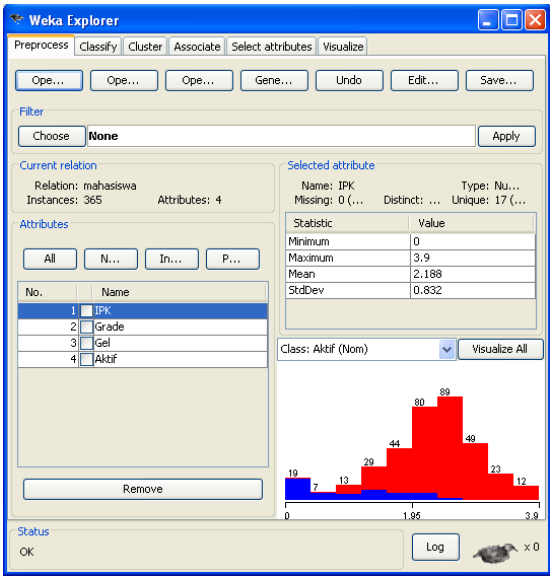
Kemudian cabang lainnya akan terus dihitung dengan cara yang sama menggunakan iterasi selanjutnya, setelah terlebih dahulu menghilangkan cabang nomor 1 yang terpilih tadi.

Hasil ini akan relevan jika menggunakan data dari tahun sebelumnya lagi sebagai perbandingan. Pada data mahasiswa angkatan 2008, terhitung 2,9% mahasiswa yang tidak melanjutkan kuliah dari total seluruhnya. Pada saat semester 2, sebanyak 71,4% memiliki nilai IPK dibawah 1, artinya data histori ini juga memiliki *root node* IPK ≤1.

4.3 Aplikasi Algoritma CART pada WEKA

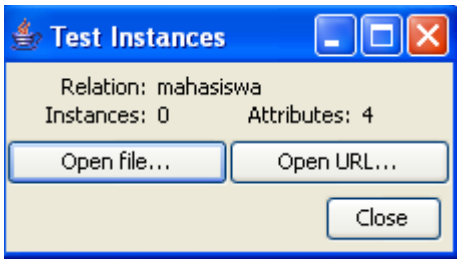
Penggunaan Aplikasi WEKA 3.6.1 untuk menunjukkan keseluruhan dari hasil data *testing* aktif atau tidak aktif berdasarkan data *training*. *File* yang dibuka pertama adalah *file* berisi 365 jumlah mahasiswa angkatan 2010 secara lengkap beserta semua atribut aktif dan tidak aktif yang telah terisi.

File data yang digunakan adalah berbentuk *excel* yang kemudian dikonversi menjadi *ekstension* arff, agar dapat dibaca oleh WEKA. *Attribute* IPK bertipe *numeric*, sementara *attribute Grade*, Gel dan Aktif bertipe *nominal spesification* yang menggunakan kurung kurawal dan koma untuk mengenumerasi nila-nilai yang mungkin. Berikut tampilan WEKA setelah membuka data *training*.



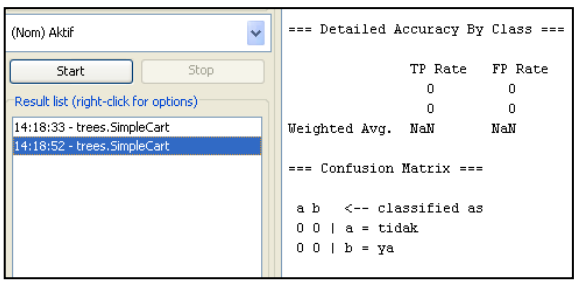
Gambar 2. Tampilan WEKA Data *Training*

Setelah data *training* dibuka, *Use training Set* untuk data *training*, sementara untuk data *testing* digunakan *supllied test set* dengan algoritma yang dipilih adalah *Classify SimpleCart.* Data *testing* dipilih, dan berikut ini gambar yang menunjukkan nama relasi data dan banyak atribut data *testing*.



Gambar 3. Tampilan *Test Instances*

Setelah membuka data *testing*, WEKA tidak akan menunjukkan hasil klasifikasi data *testing* tersebut secara langsung pada *run information*. *Run Information* akan menunjukkan *Confusion Matrix* yang semuanya bernilai 0. Berikut tampilan *run information confusion matrix* pada data *testing*.



Gambar 4. Tampilan *Confusion Matrix*

Hasil baru data *testing* dapat dilihat pada *visualize classifier errors* yang disimpan dalam bentuk *file* WEKA dan dibuka dengan aplikasi lain,

misal *notepad* atau w*ordpad* untuk membaca hasil klasifikasi data *testing*. *File* hasil tersebut akan sedikit berbeda dengan data *testing* sebelumnya, dengan berubahnya nama relasi mahasiswa, menjadi mahasiswa *predicted*, dan bertambahnya lagi dua atribut baru yaitu *attribute Instance number* bertipe *numeric* dan *attribute predicted* aktif bertipe *nominal spesification* dengan pilihan ya atau tidak.

5. KESIMPULAN

Dari hasil perhitungan kesesuaian calon cabang dan noktah keputusan serta penggunaan aplikasi WEKA secara keseluruhan, didapatkan beberapa kesimpulan, yaitu :

a. Dari 11 kemungkinan kombinasi menggunakan 3 atribut utama mahasiswa yang aktif atau tidak aktif pada semester 2, didapatkan *root node* yaitu IPK≤1.
b. Hasil klasifikasi pada data *testing* sebanyak 188 mahasiswa angkatan 2011 diperkirakan 168 yang akan terus melanjutkan kuliah pada semester 2 dan sisanya sebanyak 20 mahasiswa yang tidak akan melanjutkan kuliah.

Hal ini kemungkinan akan memberikan hasil yang agak berbeda, jika menggunakan data yang lebih banyak lagi, algoritma-algoritma yang lain dan juga beberapa *tools* yang lain sebagai perbandingan hasil klasifikasi.

PUSTAKA

Kadir, M. A., (2007), *Perbandingan Performansi Algoritma Decision Tree CART dan CHAID*, Seminar Nasional Aplikasi Teknologi Informasi (SNATI).

Kusrini, dan Luthfi, E.T., (2009), *Algoritma Data mining*, Penerbit Andi.

Nafisah, R., K., (2008), *Perkembangan Algoritma untuk Menghitung Pola yang Sering Muncul pada Basis Data yang Besar*, Institut Teknologi Bandung.

Otok, B. W., (2005), *Klasifikasi Perbankan dengan Pendekatan CART dan MARS*, Jurnal Widya Manajemen dan Akuntansi.

Sinambela, Y. E. S., (2008), *Penerapan Metode Pohon Klasifikasi dengan Algoritma CART pada Data Status Daerah Kabupaten di Indonesia*, Institut Pertanian Bogor.

Susanto, S., dan Suryadi, D., (2010), *Pengantar Data mining*, Penerbit Andi.

Wibowo,A., dan Purwarianti, A., (2011), *Penerapan Bagging untuk Memperbaiki Hasil Prediksi Nasabah Perusahaan Asuransi X*, Konferensi Nasional ICT-M.

Widagdo, K. A., (2010), *Pembentukan Pohon Klasifikasi Biner dengan Algoritma CART (Studi Kasus Penyakit Diabetes Suku Pima Indian)*, Universitas Diponegoro.