Penggunaan Algoritma *SLIQ* untuk Pengklasifikasian Kinerja Akademik Mahasiswa

(Studi kasus : Data Akademik Mahasiswa Fakultas Teknologi Informasi UNISBANK)

Arief Jananto
Program Studi Sistem Informasi
Fakultas Teknologi Informasi, Universitas Stikubank
email : arief@unisbank.ac.id

Abstract

Academic data increases every year in line with the increase of students. Abundant data store is also an abundance of information. Data mining technology is a tool for extracting information on large databases and has been widely used in many domains. Predicting student performance (study evaluation) is an activity to determine a future state based on existing data. Data in the field of academic research has been done with various methods and algorithms, but the use of algorithm SLIQ (Supervised Learning In Quest) has not been done.

SLIQ is an algorithm developed by the IBM's Quest project team in 1996 for mining large datasets. SLIQ algorithm classify and predict the students performance, beginning with the data cleaning, conducted election training and testing data. By calculating gini index of each attribute and then selecting the smallest gini index data table is split according to the criteria until find the same class. From the results of the calculation process can produce a set of rules that can be used to predict student performance.

From the experiment it can be concluded that the algorithm SLIQ with decision tree technique can be used as an alternative in designing a system datamining applications. Tests conducted system showed that the constructed model can be used to predict the performance of new students. The resulting accuracy of the model system in fact has a lower score than the accuracy of other applications that are used as a comparison of Tanagra. Advantages of the proposed system is in its design does not need complex calculations in obtaining the gini index attributes.

Keywords*: SLIQ algorithm, predicting performance, gini index*

PENGANTAR

Teknologi data mining merupakan salah satu alat bantu untuk penambangan data pada basis data berukuran besar dan dengan spesifikasi tingkat kerumitan yang telah banyak digunakan pada banyak domain. Penambangan data pada bidang akademik telah banyak dilakukan. Beberapa penelitian juga telah banyak dilakukan dengan menggunakan teknik data mining untuk menambang berbagai informasi pada domain akademik seperti Candra dan Nandhini (2005) melakukan penelitian untuk memprediksi kinerja siswa dengan teknik klasifikasi dengan algoritma induksi pohon keputusan dan *native bayes*. Menurut penelitian ini, kinerja siswa dapat dinilai melalui serangkaian tes terhadap perilaku seperti *aptitude*, *analytical*, *logical*, *Communication ability* dan *technical*. Untuk membangun model digunakan metode *decision tree (*pohon keputusan). Dengan menggunakan metode pembelajaran pohon maka dataset dapat di*split* ke dalam sejumlah subset yang didasarkan pada pengujian terhadap nilai atribut. Untuk membangun pohon keputusan digunakan perhitungan *gini impurity* dan *information theory.*

Kalles dan Pierrakeas (2006) melakukan penelitian untuk menganalisa kinerja siswa pada sistem pembelajaran jauh dengan algoritma genetik dan pohon keputusan. Pada penelitiannya, Kalles dan Pierrakeas (2006) melakukan analisa kinerja suatu kelompok siswa pada sebuah perguruan tinggi di Hellenic Ope University (HOU) melalui kemampuan

pengerjaan pekerjaan rumah yang pada akhirnya akan diperoleh hubungannya dengan tingkat keberhasilan di ujian akhir (*final exam)*.

Al-Radaideh, dkk.,(2006) menganalisa dan mengevaluasi data akademik untuk mendapatkan kinerja dari siswa yang selanjutnya dapat digunakan mengetahui kualitas perguruan tinggi. Pohon keputusan yang dibangun menggunakan sejumlah atribut prediktor berupa *HSGrade, Fund, TDept, TDegree, HKind, Study-Type, T-Gender, St-Depart, St-Gender.*

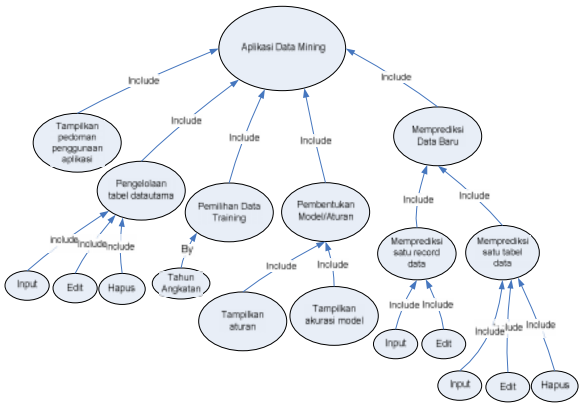
Pramudyo (2008) melakukan penelitian tentang klasifikasi mahasiswa baru berdasarkan prediksi indeks prestasi pada semester I(studi kasus program studi Teknik Informatika Universitas Bina Darma Palembang) dengan menggunakan metoda case base reasoning**.** Atribut yang digunakan sebagai prediktor adalah sejumlah nilai raport yang berjumlah 28 buah matapelajaran di Sekolah Menengah Atas atau jenjang pendidikan sebelumnya. Data tersebut dijadikan sebagai data training yang kemudian dengan metode *Case Base Reasoning*, dapat diperoleh suatu model yang selanjutnya dapat digunakan untuk mengklasifikasi data mahasiswa baru untuk mendapatkan indek prestasi (ip) semester 1 dengan rentang 1 sampai dengan 4.

Pada penelitian yang dilakukan difokuskan pada penggunaan algoritma SLIQ (*Supervised Learning In Quest*) untuk mengklasifikasikan data akademik dari mahasiswa lama (berupa data nilai hasil tes penerimaan mahasiswa baru (PMB) saat mahasiswa lama mendaftar, beberapa atribut identitas diri dan data indeks prestasi kumulatifnya (ipk)) untuk membangun sebuah model yang kemudian model tersebut digunakan untuk memprediksi kinerja mahasiswa baru. Data hasil tes PMB berupa nilai tes potensi akademik (tpa), nilai tes bahasa inggris. Data atribut identitas diri seperti jenis kelamin, asal sekolah, kategori sekolah dan usia, sedangkan indeks prestasi kumulatif sebagai indikator kinerja digunakan untuk mendapatkan label kelas dari record data yang ada. Berdasarkan ipk kinerja mahasiswa dikelompokan dalam dua kelas yaitu untuk ipk kurang dari 2.00 adalah kelas 'Tidak Mampu' dan untuk ipk lebih besar atau sama dengan 2.00 adalah kelas 'Mampu'.

METODOLOGI PENELITIAN

Fungsi-Fungsi Produk

Produk Aplikasi dibangun dalam bentuk form-form, sehingga dalam sebuah form bisa berisi satu atau lebih fungsi yang digunakan untuk mengelola suatu tabel data menjadi suatu aturan. Adapun fungsi-fungsi yang ada dapat dimanfaatkan oleh pengguna dengan mudah. Gambar 1. merupakan diagram fungsi dari produk aplikasi data mining yang akan dibangun.



Gambar 1. Diagram hirarki fungsi produk

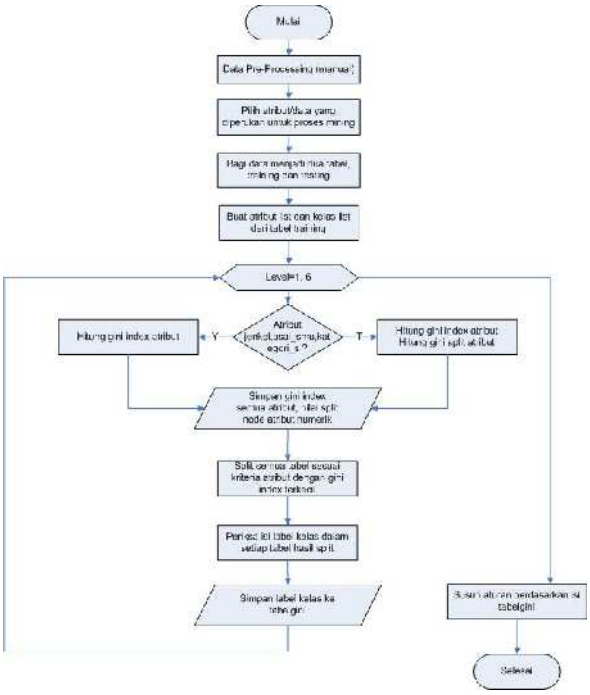
Berikut ini penjelasan dari masing-masing fungsi yang tersedia pada aplikasi ini :

1. Fungsi Tampilkan pedoman penggunaan aplikasi, digunakan untuk menampilkan pedoman dan aturan bagi pengguna yang akan menggunakan aplikasi ini.
2. Fungsi Pengelolaan tabel data utama, digunakan untuk mengelola tabel data utama, yang meliputi input, edit dan hapus.
3. Fungsi pemilihan data training, digunakan untuk menampilkan sebuah formulir yang berisi pilihan mengenai tahun angkatan dari data yang dijadikan data training maupun testing.
4. Fungsi Pembentukan model/aturan, digunakan untuk menampilkan pilihan proses aturan untuk menghasilkan daftar aturan, akurasi model, menyimpan ke excel dan pilihan tutup.
5. Fungsi Memprediksi data baru, digunakan untuk menampilkan beberapa pilihan yang digunakan untuk memprediksi data tunggal (satu record data) maupun memprediksi data

tabel (lebih dari satu record data).

Diagram Alir Proses Data Mining

Aliran proses data mining diperlihatkan pada gambar 2.



Gambar 2. Diagram Aliran Proses Data Mining

Data Pre_Processing

Pada gambar 2. Data Pra-Processing yang berisi kegiatan data *integration, data cleaning, data selection* dan *data transformation* dilakukan secara manual oleh admin hingga data siap untuk proses data mining.

Pembentukan Data Training dan Testing

Tahap berikutnya, memilih dan menyalin atribut-atribut berikut datanya ke dalam tabel baru yang nantinya akan digunakan selama proses datamining. Atribut-atribut ini dibagi dua kelompok yaitu sebagai atribut predictor dan atribut target. Adapun yang menjadi atribut *predictor* adalah jenis kelamin, usia, asal smu, kategori sekolah, nilai tes potensi akademik, nilai tes bahasa inggris, sedangkan sebagai atribut target adalah kelas. Tabel baru diberi nama data awal. Data yang ada dalam tabel dataawal kemudian dipecah menjadi tabel training dan tabel testing dengan pembagian 75% dan 25%. Pembagian dilakukan dengan memeriksa dari nomor record, dimana tabel training di ambil dari nomor record 1 hingga nomor record lebih kecil dari ke 75/100 dari jumlah record . Sedangkan untuk data testing diambil dari nomor record lebih besar dari 75/100 x jumlah record.

Selanjutnya, data training kemudian akan dibuat daftar pasangan atribut yang berisi dua buah atribut untuk atribut-atribut yang bukan kelas. Sedangkan untuk atribut kelas(target) akan berisi tiga buah atribut. Untuk atribut non kelas berisi pasangan identitas nomor record (IdRec). Kemudian untuk atribut bertipe numerik maka data record dapat langsung diurutkan (sort) secara ascending mengikuti urutan data pada atribut yang bukan idrecord. Sedangkan untuk atribut yang bukan numerik atau kategori tidak perlu diurutkan.

Tahap Pembelajaran

Tahap pembelajaran (*training*) menerima masukkan berupa record data training. Selanjutnya dengan mempersiapkan jumlah perulangan sebanyak jumlah atribut prediktor, maka mulailah dilakukan pemeriksaan tiap atribut dengan perhitungan nilai gini *split* index untuk menentukan atribut mana yang akan menjadi atribut *split*. Perhitungan nilai gini *split* index dibagi dua yaitu untuk atribut kategori hanya dihitung nilai gini *split* indexnya saja dengan cara mencari distribusi frekuensi dari masing-masing nilai kategori terhadap label kelas dan menyimpannya ke dalam histogram yang selanjutnya akan dihitung nilai gini indexnya, kemudian nilai gini index tersebut dijumlahkan untuk mendapatkan nilai *gini split* nya. Kemudian untuk atribut yang memiliki nilai gini *split* terkecil yang akan dijadikan sebagai atribut pemecah. Untuk atribut numeric selain dihitung *gini split index*-nya juga dihitung nilai *split point* untuk menentukan pada nilai berapa atribut numeric akan *split*. Pada penelitian ini semua atribut kategori hanya berisi dua buah nilai (*binary).* Contoh, untuk atribut jenis kelamin (data tahun 2005 yang berisi 433 record data training) diperoleh nilai dalam tabel histogram pada tabel 1.

Tabel 1. Histogram data jenis kelamin

	Mampu	Tidak Mampu	

Perempuan	135	25	160
Laki-laki	215	58	273
	350	83	433

Karena dataset dibagi menjadi 2 partisi maka dihitung gini index untuk masing-masing partisi sebagai berikut :

$$gini(t) = 1 - \sum_{j=1}^{m} P_J^2$$

$$Gini(Perempuan) = 1 - (135/160)^2 - (25/160)^2$$
$$= 0{,}26367188$$
$$\text{Gini(Laki - laki)} = 1 - (215/273)^2 - (58/273)^2$$
$$= 0{,}33463484$$

Sedangkan gini atribut(*split*) dapat dihitung :

$$gini_{split} = \frac{n_1}{n} gini(Perempuan) + \frac{n_2}{n} gini(Laki-laki)$$
$$= \frac{160}{433} x 0{,}26367188 + \frac{273}{433} x 0{,}33463484$$
$$= 0{,}30841296$$

Semua atribut prediktor dihitung dan dibandingkan dengan aturan yang telah disepkati. Hasil perhitungan disimpan dalam sebuah tabel gini dengan menyertakan nomor putaran perhitungan gini index. Setelah diperoleh atribut yang bernilai index gini terkecil maka selanjutnya tabel dengan atribut berindex gini terkecil dibagi(*split*) menjadi dua cabang (binary). Untuk atribut kategori, maka tabel yang berindex gini terkecil langsung dipecah menjadi dua kelompok tabel baru. Selanjutnya tabel yang lain juga dipecah dengan mengikuti idrecord yang ada pada tabel root. Hasil pecahan tersebut kemudian diuji apakah memiliki kelas yang sama atau belum. Kalau sudah maka proses *split*ing selesai, sedangkan jika belum maka aka dilakukan putaran berikutnya untuk mencari pemecahan selanjutnya. Jika sudah selesai semua maka kemudian dapat disusun suatu aturan. Proses penyusunan aturan dilakukan dengan cara mengkonversi isi *field* kriteria yang ada di tabelgini, sehingga dihasilkan daftar aturan seperti tampak pada gambar 3.

Gambar 3. Daftar aturan dalam bentuk *TreeView*

HASIL DAN PEMBAHASAN

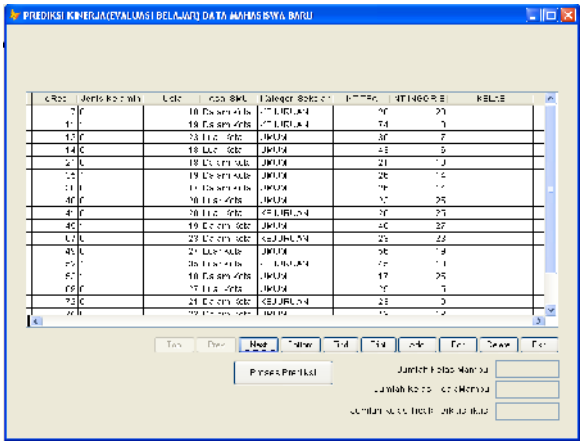
Sebagai data pelatihan digunakan data mahasiswa fakultas teknologi informasi tahun angkatan 2005 dan 2006. Untuk tahun 2005 terdapat jumlah record sebanyak 577 record setelah melalui proses praprosesing secara manual. Dan untuk tahun 2006 terdapat data sebanyak 490 record data.

Proses Training

Pada aplikasi ini setiap atribut prediktor hanya dibatasi satu kali muncul dalam tree/pohon, sehingga jumlah level dari pohon dibuat menjadi enam level. Berdasarkan pada algoritma yang digunakan dan pemecahan atau *split*/partisi data bersifat binary (dibagi dua bagian) maka diperoleh aturan maksimal 2 pangkat n (2^n), dimana n=6 sehingga 2 pangkat 6 yaitu 64 aturan. Proses training dilakukan dua kali, dimana dengan menggunakan data mahasiswa tahun angkatan 2005 dan 2006. Dimana kemudian setiap tahun angkatan dibagi menjadi dua bagian yaitu 75 persen sebagai data pelatihan untuk menghasilkan aturan dan 25 persen sisanya digunakan untuk menguji akurasi dari aturan yang dihasilkan.

User Interface

User interface yang digunakan untuk menjadi perantara antara sistem komputasi dengan penguna dengan tampilan seperti pada gambar 4. Konsep perancangan sistem aplikasi ini selengkapnya ada pada lampiran 1 dan 2 yang akan menguraikan DFD dari level 0 dan level 1.



Gambar 4. Tampilan antar muka prediksi data mahasiswa baru

Hasil Pengujian

Pada data tahun angkatan 2005 diperoleh urutan atribut berdasarkan tingkat pengaruh terhadap pencapaian kinerja adalah jenis kelamin, usia, asal_smu, Nilai TPA, Nilai Tes Inggris dan kategoris. Dari hasil pengujian dengan data testing menunjukan akurasi model untuk tahun angkata 2005 adalah sebesar 41,67 % seperti tampak pada gambar 5.(a).

Hasil perhitungan tanpa memperhitungkan kelas 'Tidak Terklasifikasi'
pada saat pengklasifikasian/pencocokan terhadap data testing

CONFUSION MATRIK

	Kelas Sebenarnya	
Prediksi	Mampu	Tidak Mampu
Mampu	34	5
Tidak Mampu	79	26

Jumlah record = 144

Akurasi model (prediksi benar) = 41.67 %

(a)

Hasil perhitungan tanpa memperhitungkan kelas 'Tidak Terklasifikasi'
pada saat pengklasifikasian/pencocokan terhadap data testing

CONFUSION MATRIK

	Kelas Sebenarnya	
Prediksi	Mampu	Tidak Mampu
Mampu	68	8
Tidak Mampu	37	9

Jumlah record = 122

Akurasi model (prediksi benar) = 63.11 %

(b)

Gambar 5. Matrik hasil prediksi terhadap data tahun angkatan 2005

Berdasarkan gambar 5.(a), dari 113 record berlabel kelas 'Mampu' dapat terprediksi dengan tepat sebagai kelas 'Mampu' sebanyak 34 record sedangkan yang dapat terprediksi sebagai kelas 'Tidak Mampu' sebanyak 79 record. Untuk record berlabel kelas 'Tidak Mampu' dari jumlah 31 record dapat terprediksi dengan tepat sebanyak 26 record sebagai record berlabel kelas 'Tidak Mampu' dan sebanyak 5 record terprediksi sebagai record berlabel kelas 'Mampu'.

Selanjutnya dilakukan uji coba untuk data angkatan tahun 2006, dengan jumlah record data sebanyak 368. Pada data tahun angkatan 2006 urutan atribut berdasarkan pengaruhnya terhadap kinerja (evaluasi belajar) mahasiswa adalah jenis kelamin, usia kemudian berikunya atribut nt_tpa, nt_inggris, asal_smu dan kategori_s. Untuk dua atribut terakhir dapat bervariasi. Diperoleh hasil perhitungan seperti tampak pada gambar 5.(b). Akurasi model aturan berdasarkan data tahun 2006 memiliki tingkat akurasi model 63,11%.

Hasil Pengujian Dengan Perangkat Lunak Lain

Perangkat lunak lain yang digunakan untuk melakukan proses data mining terhadap data akademik tersebut adalah Tanagra versi 4.1., merupakan perangkat lunak yang bebas untuk didapat (*free download)*. Dengan menggunakan teknik pembelajaran (supervised learning) dan algoritma C4.5 yang merupakan algoritma pendahulunya. Pada TANAGRA, tidak terdapat pembatasan jumlah kemunculan sebuah atribut, sehingga ada beberapa atribut yang muncul lebih dari satu kali. Dengan menggunakan TANAGRA untuk memproses data tahun angkatan 2005 diperoleh hasil dalam confusion matrix, dengan akurasi model dari data tahun angkatan 2005 adalah 86,11%, sedangakan untuk data tahun angkatan 2006 diperoleh akurasi model sebesar 86,89%.

Hasil Prediksi Pada Data Mahasiswa Baru

Uji coba untuk memprediksi data mahasiswa baru (dalam hal ini data yang diperoleh adalah tahun 2007), dengan menggunakan aturan yang diperoleh dari data training tahun angkatan 2005 dan 2006 diperoleh hasil seperti tampak pada tabel 2.

Tabel 2. Tingkat akurasi prediksi data mahasiswa baru

	Sebenarnya(2005)		Sebenarnya(2006)	
Prediksi	Mampu	Tidak Mampu	Mampu	Tidak Mampu
Mampu	129	18	130	18
Tidak Mampu	283	83	282	83
Akurasi Prediksi	41,33 %		41,52 %	

KESIMPULAN

Dari hasil penelitian yang telah dilakukan dapat disimpulkan hal-hal sebagai berikut:

1. Aplikasi datamining yang dihasilkan dengan menggunakan algoritma *Supervised Learning In Quest* (SLIQ) yang dapat digunakan untuk memprediksi kinerja (evaluasi belajar) mahasiswa baru.
2. Hasil pengujian dengan aplikasi sejenis lain menunjukkan skor yang lebih rendah pada aplikasi yang dibuat. Hal ini dapat dikarenakan adanya pembatasan jumlah kemunculan dari atribut pada pohon dalam aplikasi yang dibuat dan jumlah data pelatihan yang sedikit, serta distribusi nilai data pada atribut numerik.
3. Keunggulan dari aplikasi ini adalah data yang diolah dapat bertipe kategori dan numerik secara bersama-sama dengan tingkat akurasi yang masih cukup baik tanpa harus menggunakan perhitungan yang terlalu rumit. Kelemahan-kelemahan dari aplikasi yang dibuat ini adalah aplikasi tidak dapat melakukan preprosesing data sehingga sistem tidak dapat mengendalikan kebenaran nilai suatu data.

DAFTAR PUSTAKA

Agathe, dan Kalina, 2005, *Educational Data Mining: a Case Study*,Pôle Universitaire Léonard de Vinci, France

Al-Radaideh, Q.A., Al-Shawakfa, E.M., dan Al-Najjar, M.I., 2006, *Mining Student Data Using Decision Trees,* The 2006 International Arab Conference on Information Technology (ACIT'2006).

Chandra, B., dan Varghese, P.P., 2008, *Fuzzy SLIQ Decision Tree Algorithm*

Chandra, E., dan Nandhini, K., 2005, *Predicting Student Performance using Classification Techniques,* Proceedings of SPIT – IEEE Colloquium and International Conference, Mumbai, India Volumen 5, 83.

Han, J., dan Kamber, M., 2001, *Data Mining: Concepts and Techniques,* SunFransisco : Morgan Kaufmann Publishers

Kalles, D., dan Pierrakeas, C., 2006, *Analyzing Student Performance in Distance Learning with Genetic Algorithms and Decsion Trees,* Hellenic Open University.

Mehta, M., Agrawal, R., dan Rissanen, J., 1996, *SLIQ: A Fast Scalable Classifier for Data Mining*

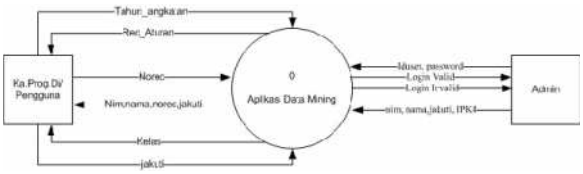
Pramudyo, A., S., 2008, *Case Based Reasoning untuk Klasifikasi Mahasiswa Baru Berdasarkan Prediksi Indeks Prestasi pada Semester I (Studi Kasus Program Studi Teknik Informatika Universitas Bina Darma Palembang),* Tesis, Jurusan Ilmu-Ilmu Komputer FMIPA UGM, Yogyakarta.

Shafer J., Mehta M., dan Agrawal R., 1996, *SPRINT: A Scalable Classifier for Data Mining*

Tan, P., Steinbach, M., dan Kumar, V., 2006, *Introduction to Data Mining*, Pearson Education.

Yan, H., Ma, R., dan Tong, X., 2005, *SLIQ in data mining and application in the generation unit's bidding decision system of electricity market,* International Power Engineering Conference (IPEC2005)

Lampiran 1. DFD Level 0 Aplikasi Datamining Prediksi Kinerja Mahasiswa Baru



Lampiran 2. DFD Level 1 Aplikasi Datamining Prediksi Kinerja Mahasiswa Baru