RANCANG BANGUN SISTEM REKOMENDASI BEASISWA MENGGUNAKAN ALGORITMA KLASIFIKASI C4.5 PADA UNIVERSITAS DIAN NUSWANTORO

Yosoa Putra Raharja

In the last 4 years (2010-2013), the number of students in Udinus much as 9289 students. Based on parental income data, there are 1405 income of the parents of students who are in the range of 3 million rupiah. Among students parents income was in the range below the 3 million dollars, 222 students of which have the status of "defaulters". Status defaulter is the status of these students who do not pay the tuition fee. This was due to the economic conditions that are not capable of. Of course, this number can be reduced with the scholarship program. Scholarship is given to a student allowance or as an aid student learning costs. A large number of scholarship applicants led to the difficulty of determining scholarship recipients. Data mining has been proven and widely used to solve various problems that exist either by the application of classification method. This study uses the C4.5 classification algorithm which is an extension of a classification algorithm ID3. From the research proves that C4.5 algorithm can be applied to perform classification based datasets grantee applicants and recipients. The results of the classification and the level of accuracy that is formed depends on the type, the type and number of datasets.

***Index Terms*—Scholarships, Data Mining, Classification, C4.5**

I. PENDAHULUAN

Universitas Dian Nuswantoro adalah salah satu lembaga pendidikan perguruan tinggi yang berdiri pada tahun 2001 berdasarkan SK Mendiknas RI No. 169/D/O/2001. Dalam 4 tahun terakhir (2010 – 2013), jumlah mahasiswa Udinus sebanyak 9289 mahasiswa. Berdasarkan data penghasilan orang tua, terdapat 1405 mahasiswa yang penghasilan orang tuanya berada di kisaran 3 juta rupiah. Diantara mahasiswa yang penghasilan orang tua berada di kisaran di bawah 3 juta rupiah, 222 mahasiswa di antaranya memiliki status "mangkir". Status mangkir adalah status dimana mahasiswa tersebut tidak melakukan pembayaran uang kuliah. Hal tersebut disebabkan karena kondisi ekonomi yang tidak mampu. Tentunya jumlah ini dapat tekan dengan adanya program beasiswa. Beasiswa adalah tunjangan yang diberikan kepada pelajar atau mahasiswa sebagai bantuan biaya belajar [1].

Udinus menawarkan fasilitas berupa bantuan beasiswa kepada seluruh mahasiswanya melalui Biro Kemahasiswaan (Bima). Bima terdapat 18 jenis beasiswa yang di tawarkan baik dari pemerintah daerah maupun swasta. Supaya beasiswa ini dapat tepat sasaran tentunya dibutuhkan proses seleksi. Banyaknya pemohon beasiswa di Udinus menuntut untuk proses seleksi rekomendasi dapat dilakukan dengan mudah, cepat dan tentunya tepat sasaran.

Data mining merupakan proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar [2]. Klasifikasi merupakan salah satu metode dalam data mining. Kalasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukannya kedalam kelas tertentu dari sejumlah kelas yang tersedia [3].

Oleh karena itu, dalam penelitian ini akan menerapkan algoritma C4.5. Algoritma C4.5 merupakan salah satu algoritma Decision Tree yang merupakan perkembangan dari algoritma Iterative Dichotomiser 3 (ID3) yang mudah dimengerti serta dapat membangun pohon keputusan dengan cepat.

II. ALGORITMA KLASIFIKASI C4.5

Pada penelitian ini, metode yang digunakan adalah Algoritma Klassifikasi C4.5 untuk klasifikasi penerima beasiswa.

A. Algortima Klasifikasi

Klasifikasi merupakan salah satu teknik atau metode dalam data mining. Dan metode klasifikasi termasuk dalam jenis *supervised learning*. *Supervised learning* adalah *machine learning* yang membutuhkan label sebagai tujuan dari pelatihan data (*data training*) [4].

Klasifikasi sendiri adalah teknik data mining yang digunakan untuk memprediksi kelompok keanggotaan untuk setiap contoh data. Klasifikasi merupakan proses penemuan model (atau fungsi) yang menggambarkan dan membedakan kelas data atau konsep yang bertujuan agar bias digunakan untuk memprediksi kelas dari obyek yang label kelasnya tidak diketahui. Klasifikasi terdiri dari dua tahap atau langkah proses. Tahap pertama adalah learning (fase training), yaitu dimana algoritma klasifikasi dibuat untuk

menganalisa data training atau data latih lalu direpresentasikan kedalam bentuk rule atau aturan klasifikasi. Sedangkan proses yang kedua adalah proses klasifikasi, dimana data test digunakan untuk memperkirakan akurasi dari aturan klasifikasi yang telah terbentuk [5]. Pada proses klasifikasi, dipengaruhi oleh empat komponen :

a. Class Label
   Variable Dependent yang bertipe kategorikal yang mempresentasikan label yang terdapat pada objek.
b. Predictor
   Variable Independent yang direpresentasikan oleh karakteristik atau atribut – atribut data.
c. Data Training
   Suatu set data yang berisi nilai dari kedua komponen di atas yang digunakan untuk menentukan kelas yang cocok berdasarkan predictor.
d. Data Testing
   Berisi data baru yang akan diklasifikasikan oleh model yang telah terbuat dan akurasi klasifikasi dievaluasi.

B. Algoritma C4.5

Algoritma C4.5 dan pohon keputusan merupakan dua model yang tak terpisahkan. Algoritma C4.5 merupakan salah satu algoritma klasifikasi yang kuat dan cukup banyak digunakan atau di implementasikan untuk pengklasifikasian dalam berbagai hal. Algoritma C4.5 juga biasa disebut J48 yang merupakan implementasi dari algoritma C4.5 pada WEKA.

Algoritma C4.5 merupakan pengembangan dari algoritma ID3 (Iterative Dichotomiser 3). Serangkaian perbaikan yang dilakukan pada algoritma ID3 mencapai puncaknya dengan menghasilkan sebuah system praktis dan berpengaruh untuk pembentukan pohon keputusan. Perbaikan tersebut meliputi metode untuk menangani numeric attributes, missing values, noisy data, dan aturan yang menghasilkan aturan dari tree [6].

Saat menyusun sebuah pohon keputusan pertama yang harus dilakukan adalah menentukan atribut mana yang akan menjadi simpul akar dan atribut mana yang akan menjadi simpul selanjutnya. Pemilihan atribut yang baik adalah atribut yang memungkinkan untuk mendapatkan pohon keputusan yang paling kecil ukurannya. Atau atribut yang bisa memisahkan obyek menurut kelasnya. Secara heuristik atribut yang dipilih adalah atribut yang menghasilkan simpul yang paling ”purest” (paling bersih). Ukuran purity dinyatakan dengan tingkat impurity, dan untuk menghitungnya, dapat dilakukan dengan menggunakan konsep Entropy, Entropy menyatakan impurity suatu kumpulan objek. Jika diberikan sekumpulan objek dengan label/output s yang terdiri dari objek berlabel 1, 2 sampai n. Entropy dari objek dengan n kelas ini dapat dihitung dengan rumus berikut :

$$Entropy(S) = \sum_{j=1}^{n} - Pj \log 2 \; Pj$$

Keterangan :
S adalah himpunan (dataset) kasus
n adalah banyaknya partisi S
pj adalah probabilitas yang di dapat dari kelas dibagi total kasus

Kemudian setelah menghitung Entropy, hitung Information Gain. Information gain adalah kriteria yang paling populer untuk pemilihan atribut. Information gain dapat dihitung dari output data atau variabel dependent y yang dikelompokkan berdasarkan atribut A, dinotasikan dengan gain (y,A). Information gain, gain (S,A), dari atribut A relatif terhadap output data S adalah :

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

Dimana nilai(A) adalah semua nilai yang mungkin dari atribut A, dan S_i adalah subset dari y dimana A mempunyai nilai i.

Berbeda dengan algoritma ID3, padam algoritma C4.5 menggunakan gain ratio untuk memperbaiki information gain dengan menggunakan rumus berikut :

$$Gain\ Ratio\ (S,A) = \frac{Gain\ (S,A)}{SplitInfo\ (S,A)}$$

Dimana S merupakan ruang (sample) data yang digunakan untuk training dan A adalah atribut. Gain (S,A) merupakan information gain pada atribut A, SplitInfo adalah nilai split information pada atribut A yang didapatkan dengan rumus berikut.

$$SplitInfo\ (S,A) = -\sum_{i=1}^{n} \frac{Si}{S} \log 2 \frac{Si}{S}$$

. Secara umum langkah algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut [7]:

a. Hitung Gain Ratio, Split Info dan entropy dari masing-masing atribut data training yang ada.
b. Buat simpul akar dari pemilihan atribut yang memiliki Gain Ratio terbesar.
c. Hitung Gain Ratio, Split Info dan entropy dari masing-masing atribut dengan menghilangkan atribut yang telah dipilih sebelumnya.
d. Buat simpul internal dari pemilihan atribut yang memiliki Gain Ratio terbesar.
e. Cek apakah semua atribut sudah dibentuk pada pohon. Jika belum, maka ulangi proses d dan e, jika sudah maka lanjut pada proses berikutnya.
f. Lakukan pemangkasan pohon untuk menghilangkan cabang-cabang yang tidak perlu.

C. Prunning

Pohon keputusan(decision tree) biasanya terlalu luas dan sering mengandung struktur atau banyak cabang yang tidak

perlu, dan umumnya disarankan untuk menyederhanakan sebelum digunakan agar mendapatkan hasil yang lebih baik. Pemangkasan ini salain mampu meningkatkan akurasi pohon keputusan juga berguna untuk menyederhanakan struktur pohon keputusan yang dihasilkan sehingga memudahkan dalam pembacaan. Untuk memangkas/pendekatan pruning ada dua cara :

a. Prepruning menghentikan proses pembuatan cabang pada titik tertentu. Semakin besar perulangan pembuatan cabang yang diperbolehkan, semakin besar pula kompleksitas dari pohon keputusan yang didapat jika data beragam, namun jika jumlah perulangan terlalu kecil, diagram pohon yang dihasilkan menjadi kurang akurat.
b. Postpruning memotong cabang pohon yang kurang mereprensentasikan data setelah sebuah pohon keputusan terbentuk. Kelas yang diberikan akan diukur dari jumlah persebaran label yang ada pada cabang tersebut.

D. Evaluasi dan Validasi

Untuk melakukan evaluasi dan validasi pada metode klasifikasi data mining dapat dilakukan dengan melakukan pengujian confusion matrix atau curva ROC (receiver operating characteristic).

a. Confusion Matrix

Confusion matrix memberikan keputusan yang diperoleh dalam traning dan testing, confusion matrix memberikan penilaian performance klasifikasi berdasarkan objek dengan benar atau salah (gurunescu). Confusion matrix berisi informasi aktual (actual) dan prediksi (predicted) pada sistem klasifikasi.

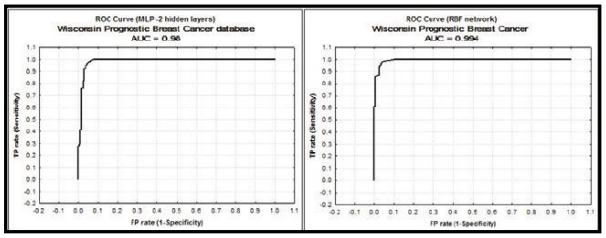
Classification	Predicted Class		
Observed Class		Class = Yes	Class = No
	Class = Yes	A (true positif – tp)	B (false negative – fn)
	Class = No	C (false positif – fp)	D (true negative – tn)

Keterangan:

- True Positive (tp) = proporsi positif dalam data set yang diklasifikasikan positif
- True Negative (tn) = proporsi negative dalam data set yang diklasifikasikan negative
- False Positive (fp) = proporsi negatif dalam data set yang diklasifikasikan potitif
- FalseNegative(fn) = proporsi negative dalam data set yang diklasifikasikan negative

b. Curva ROC

Curve ROC (Receiver Operating Characteristic) adalah cara lain untuk mengevaluasi akurasi dari klasifikasi secara visual [8]. Sebuah grafik ROC adalah plot dua dimensi dengan proporsi positif salah (fp) pada sumbu X dan proporsi positif benar (tp) pada sumbu Y. Titik (0,1) merupakan klasifikasi yang sempurna terhadap semua kasus positif dan kasus negatif. Nilai positif salah adalah tidak ada (fp = 0) dan nilai positif benar adalah tinggi (tp = 1). Titik (0,0) adalah klasifikasi yang memprediksi setiap kasus menjadi negatif {-1}, dan titik (1,1) adalah klasifikasi yang memprediksi setiap kasus menjadi positif {1}. Grafik ROC menggambarkan trade-off antara manfaat (‘true positives’) dan biaya (‘false positives’). Berikut tampilan dua jenis kurva ROC (discrete dan continous).



Dapat disimpulkan bahwa, satu point pada kurva ROC adalah lebih baik dari pada yang lainnya jika arah garis melintang dari kiri bawah ke kanan atas didalam grafik. Tingkat akurasi dapat di diagnosa sebagai berikut [7]:

Akurasi 0.90 – 1.00 = Excellent classification
Akurasi 0.80 – 0.90 = Good classification
Akurasi 0.70 – 0.80 = Fair classification
Akurasi 0.60 – 0.70 = Poor classification
Akurasi 0.50 – 0.60 = Failure

E. Kerangka Pemikiran

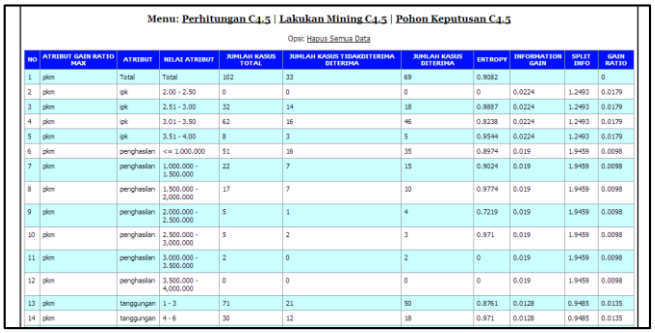
Dalam melakukan penelitian tugas akhir ini, dibuat sebuah kerangka pimikiran yang berguna sebagai pedoman dalam melakukan penelitian sehingga penelitian yang dilakukan dapat berjalan konsisten.

Masalah dalam penelitian ini adalah proses penentuan penerima beasiswa yang sulit dilakukan sehingga digunakan algoritma decision tree C4.5 untuk mengklasifikasi pendaftar beasiswa. Desain dalam penelitian ini menggunakan model CRISP-DM dan pemrograman PHP yang digunakan sebagai aplikasi model pengembangannya. Kemudian dilakukan pengujian terhadap kinerja algoritma Decision tree menggunakan confusion matrix dalam evaluasi. Setelah dilakukan evaluasi dan pengujian maka dapat diketahui rule – rule dari pohon keputusan, sehingga dapat ditarik diketahui klasifikasi pendaftar beasiswa.

III. HASIL PENELITIAN

A. *Implementasi*

Pada penelitian ini, menerapkan algoritma klasifikasi C4.5 yang diimplementasi menggunakan bahasa pemrograman PHP. Berikut merupakan hasil implementasinya.

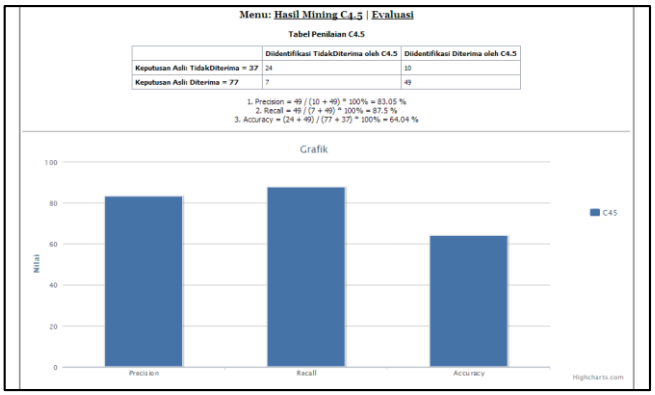


Menu: Perhitungan C4.5 | Lakukan Mining C4.5 | Pohon Keputusan C4.5

Opsi: Hapus Semua Data

NO	ATRIBUT GAIN RATIO MAX	ATRIBUT	NILAI ATRIBUT	JUMLAH KASUS TOTAL	JUMLAH KASUS TIDAKDITERIMA DITERIMA	JUMLAH KASUS DITERIMA	ENTROPY	INFORMATION GAIN	SPLIT INFO	GAIN RATIO
1	pkm	Total	Total	102	33	69	0.9082			0
2	pkm	ipk	2.00 - 2.50	0	0	0	0	0.0224	1.2493	0.0179
3	pkm	ipk	2.51 - 3.00	32	14	18	0.9887	0.0224	1.2493	0.0179
4	pkm	ipk	3.01 - 3.50	62	16	46	0.8238	0.0224	1.2493	0.0179
5	pkm	ipk	3.51 - 4.00	8	3	5	0.9544	0.0224	1.2493	0.0179
6	pkm	penghasilan	<= 1.000.000	51	16	35	0.8974	0.019	1.9459	0.0098
7	pkm	penghasilan	1.000.000 - 1.500.000	22	7	15	0.9024	0.019	1.9459	0.0098
8	pkm	penghasilan	1.500.000 - 2,000.000	17	7	10	0.9774	0.019	1.9459	0.0098
9	pkm	penghasilan	2.000.000 - 2.500.000	5	1	4	0.7219	0.019	1.9459	0.0098
10	pkm	penghasilan	2.500.000 - 3.000.000	5	2	3	0.971	0.019	1.9459	0.0098
11	pkm	penghasilan	3.000.000 - 3.500.000	2	0	2	0	0.019	1.9459	0.0098
12	pkm	penghasilan	3.500.000 - 4,000.000	0	0	0	0	0.019	1.9459	0.0098
13	pkm	tanggungan	1 - 3	71	21	50	0.8761	0.0128	0.9485	0.0135
14	pkm	tanggungan	4 - 6	30	12	18	0.971	0.0128	0.9485	0.0135

Tampilan Hasil Perhitungan Mining



Tampilan Hasil Evaluasi Mining

IV. PENUTUP

A. *Kesimpulan*

Dari penelitian yang telah dilakukan, dapat ditarik kesimpulan sebagai berikut :

a. Tingkat akurasi dari implementasi algoritma klasifikasi C4.5 dipengaruhi oleh beberapa hal seperti jenis, jumlah, isi dataset dan jumlah partisi data set.
b. Dari hasil uji coba partisi data set ditemukan tingkat akurasi tertinggi pada jumlah partisi data sebanyak 90% dan menghasilkan akurasi sebanyak 92,31%. Dan semakin besar jumlah partisi data maka akan menghasilkan jumlah akurasi yang semakin tinggi pula.
c. Hasil penelitian ini dapat dijadikan sebagai pertimbangan metode baru dalam proses penerimaan beasiswa pada Universitas Dian Nuswantoro

B. *Penelitian Selanjutnya*

Beberapa ide yang yang dapat digunakan untuk penelitian selanjutnya adalah sebagai berikut :

d. Dapat dilakukan pengolahan data lebih lanjut seperti pemilihan jenis attribute yang lain agar dapat menghasilkan tingkat akurasi yang lebih baik.
e. Agar mendapatkan hasil klasifikasi yang lebih maksimal disarankan melakukan penelitian lebih lanjut untuk memodifikasi algoritma C4.5, menggabungkan dengan algoritma lain dan atau menggunakan algoritma – algoritma optimasi data mining seperti PSO, Algoritma Genetika, pembobotan atribut dan lain lain.
f. Penelitian selanjutnya dapat dikembangkan dengan menggunakan jenis data serupa namun dengan metode yang lain, seperti clustering yang dapat digunakan sebagai perbandingan.

REFERENCES

V. BIBLIOGRAPHY

[1] E. Setiawan, "Kamus Besar Bahasa Indonesia (KBBI)," Kemdikbud (Pusat Bahasa), 2012 - 2014. [Online]. Available: http://kbbi.web.id/. [Accessed 25 Maret 2014].

[2] P. Tan, Introduction to Data Mining, Boston: Pearson Education, 2006.

[3] E. Prasetyo, Data Mining - Konsep dan Aplikasi Menggunakan Matlab, Yogyakarta: Penerbit Andi, 2012.

[4] M. Mohri, A. Rostamizadeh and A. Talwalkar, Fondations of Machine Learning, MIT Press, 2012.

[5] J. Han and M. Kamber, Data Mining : Concepts and Techniques 2nd Edition, San Francisco: Morgan Kaufmann Publishers, 2006.

[6] I. H. Witten, E. Frank and M. A. Hall, DATA MINING - Practical Machine Learning Tools and Techniques (3rd ed), Elsevier Inc., 2011.

[7] F. Gorunescu, Data Mining Concept Model Technique, India: Springer, 2011.

[8] C. Vercellis, Bussiness Intelligence : Data Mining and Optimization for Decision Making, John Wiley & Sons, Ltd, 2009.

[9] "Direktorat Jendral Pendidikan Tinggi," 2012. [Online]. Available: http://www.dikti.go.id/?page_id=397&lang=id. [Accessed 25 Maret 2014].