PENERAPAN DATA MINING UNTUK REKOMENDASI BEASISWA PADA SMA MUHAMMADIYAH GUBUG MENGGUNAKAN ALGORITMA C4.5

Dina Maurina, Ahmad Zainul Fanani S.Si, M.Kom

Jurusan Teknik Informatika FIK UDINUS, Jl. Nakula No. 5-11 Semarang-50131

dinamaurinaa@gmail.com

Abstrak - Klasifikasi rekomendasi beasiswa dilakukan untuk mengklasifikasi apakah siswa akan mendapatkan rekomendasi beasiswa sesuai dengan bobot yang akan di nilai. Klasifikasi dilakukan menggunakan data mining algoritma C4.5. Data yang digunakan yaitu data jurusan, kelas, jumlah nilai, penghasilan orangtua, dan jumlah saudara kandung. Proses data mining pada data training akan menghasilkan pohon keputusan atau rule. Metode evaluasi yang dilakukan dalam penelitian ini yaitu menggunakan confusion matrix dan nilai akurasi, untuk sekali pengujian tingkat akurasi yang dihasilkan yaitu 77%. hal ini membuktikan bahwa algoritma C4.5 cukup akurat dalam menentukan rekomendasi beasiswa pada SMA Muhammadiyah Gubug.

Kata Kunci :Beasiswa, Data Mining, Klasifikasi, C4.5

I. PENDAHULUAN

Pendidikan merupakan faktor utama dalam pembentukkan pribadi manusia. Pendidikan sangat berperan dalam membentuk baik atau buruknya pribadi manusia. Dengan hal tersebut, pemerintah sangat serius menangani bidang pendidikan, sebab dengan sistem pendidikan yang baik diharapkan muncul generasi penerus bangsa yang berkualitas dan mampu menyesuaikan diri untuk hidup bermasyarakat, berbangsa dan bernegara.

Beasiswa merupakan tunjangan yang diberikan kepada pelajar atau mahasiswa sebagai bantuan biaya belajar **[1]**. Beasiswa yang di dapat dari SMA Muhammadiyah Gubug merupakan beasiswa yang diberikan oleh 3 instansi yang bergerak di bidang perbankan yang selama 6 bulan memberikan beasiswa kepada siswa yang terpilih yang diseleksi langsung oleh pihak bank. Tetapi pihak sekolah harus menyeleksi siswa terlebih dahulu dan selanjutnya pihak bank yang akan menentukan.

Dari masalah yang ada, maka SMA Muhammadiyah Gubug harus mempunyai suatu aplikasi pendukung yang dapat mengambil keputusan dan mengklasifikasikan nilai siswa berdasarkan siswa yang berprestasi dan tidak mampu agar memudahkan dalam menentukan siswa yang akan direkomendasikan mendapatkan beasiswa dari instansi tersebut adalah tepat sasaran dan memberikan solusi dan membantu siswa kurang mampu yang berprestasi.

II. TEORI PENUNJANG

2.1 Klasifikasi

Salah satu proses dalam data mining adalah klasifikasi, pada klasifikasi diberikan sejumlah *record* yang dinamakan *training set*, yang terdiri dari kelas untuk *record*. Tujuan dari klasifikasi adalah untuk menemukan model dari *training set* yang membedakan record

kedalam kategori atau kelas yang sesuai, model tersebut kemudian digunakan untuk mengklasifikasikan *record* yang kelasnya belum diketahui sebelumnya.

Komponen-komponen utama dari proses klasifikasi antara lain [**10**]:

1. Kelas, merupakan variabel tidak bebas yang merupakan label dari hasil klasifikasi. Sebagai contoh adalah kelas loyalitas pelanggan, kelas badai atau gempa bumi, dan lain-lain.
2. Prediktor, merupakan variable bebas suatu model berdasarkan dari karakteristik atribut data yang diklasifikasi, misalnya merokok, minum-minuman beralkohol, tekanan darah, status perkawinan, dan sebagainya.
3. Set data pelatihan, merupakan sekumpulan data lengkap yang berisi kelas dan predictor untuk dilatih agar model dapat mengelompokan ke dalam kelas yang tepat. Contohnya adalah grup pasien yang telah di-test terhadap serangan jantung, grup pelanggan di suatu supermarket, dan sebagainya.
4. Set data uji, berisi data-data baru yang akan dikelompokan oleh model guna mengetahui akurasi dari model yang telah dibuat.

2.2 Algoritma C4.5

Algoritma C4.5 merupakan algortima yang digunakan untuk membentuk pohon keputusan. Pohon keputusan merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami. Dan mereka juga dapat diekspresikan dalam bentuk bahasa basis data seperti *Structured Query Language* untuk mencari *record* pada kategori tertentu [**8**].

Ada beberapa tahap dalam membuat sebuah pohon keputusan dengan algoritma C4.5 [**10**].

1. Menyiapkan data training. *Data training* biasanya diambil dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokan ke dalam kelas-kelas tertentu.
2. Menentukan akar dari pohon. Akar akan diambil dari atribut yang terpilih, dengan cara menghitung nilai gain dari masing-masing atribut, nilai gain yang paling tinggi yang akan menjadi akar pertama. Sebelum menghitung nilai gain dari atribut, hitung dahulu nilai entropy. Untuk menghitung nilai entropy digunakan rumus :

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

Keterangan :
S = himpunan kasus
n = jumlah partisi S
pi = proporsi S_i terhadap S

3. Kemudian hitung nilai *gain* menggunakan rumus :

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

Keterangan :
S = Himpunan kasus
A = fitur
n = jumlah partisi atribut A
|Si| = proporsi Si terhadap S
|S| = jumlah kasus dalam S

4. Ulangi langkah ke-2 hingga semua record terpartisi.
5. Proses partisi pohon keputusan akan berhenti saat :
   a. Semua *record* dalam simpul N mendapat kelas yang sama.
   b. Tidak ada atribut di dalam record yang dipartisi lagi.

c. Tidak ada *record* di dalam cabang yang kosong.

2.3 *Confusion Matrix*

Confusion Matrix adalah tools yang digunakan untuk evaluasi model klasifikasi untuk memperkirakan objek yang benar atau salah. Sebuah matrix dari prediksi yang akan dibandingkan dengan kelas yang asli dari inputan atau dengan kata lain berisi informasi nilai actual dan prediksi pada klasifikasi **[11]**.

Classification	Predicted class	
	Class = Yes	Class = No
Class = Yes	a (*true positive*-TP)	b (*false negative*-FN)
Class = No	c (*false positive*-FP)	d(*true negative*-TN)

Evaluasi dan validasi hasil dihitung menggunakan rumus akurasi, *precision recall dan f-measure* berikut ini**[13]** :

1. Akurasi

Perhitungan akurasi dilakukan dengan cara membagi jumlah data yang diklasifikasi secara benar dengan total sample *data testing* yang diuji.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Precision

Menghitung nilai *precision* dengan cara membagi jumlah data benar yang bernilai positif *(True Positive)* dibagi dengan jumlah data benar yang bernilai positif *(True Positive)* dan data salah yang bernilai positif *(False Negative)*.

$$Precision = \frac{TP}{TP+FP}$$

3. Recall

Sedangkan *recall* dihitung dengan cara membagi data benar yang bernilai positive *(True Positive)* dengan hasil penjumlahan dari data benar yang bernilai positif *(True Positive)* dan data salah yang bernilai negatif *(False Negative)*.

$$Recall = \frac{TP}{TP+FN}$$

4. F-Measure

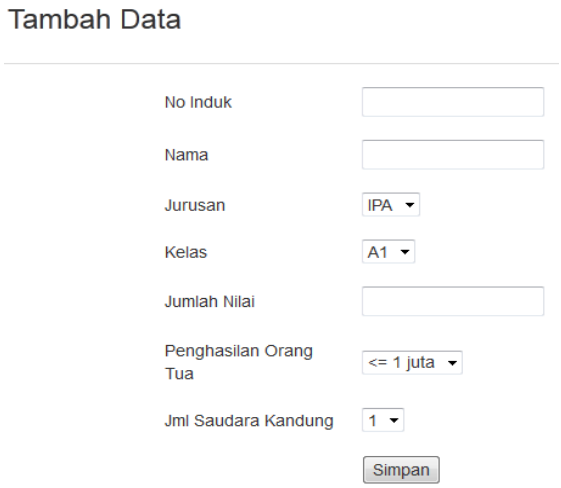
Nilai *F-Measure* didapat dari perhitungan pembagian hasil dari perkalian *precision* dan *recall* dengan hasil penjumlahan *precision* dan *recall*, kemudian dikalikan dua.

$$F - Measure = 2 * \frac{precision * recall}{precision + recall}$$

III. HASIL &IMPLEMENTASI

Pada penelitian ini, penerapan algoritma klasifikasi C4.5 telah diimplementasikan menggunakan bahasa pemrograman PHP. Dibawah ini merupakan hasil implementasinya.

1. Form untuk menambahkan data siswa



2. Form data yang telah ditambahkan

No	NIS	Nama	Jurusan	Kelas	Jml Nilai	Penghasilan Ortu	Sdr Kandung	#
1	12.7022	ERIC FIRDYAN	IPS	S2	722	1-2 juta	3-4 Orang	edit | hapus
2	12.7027	FUJI RIFAYANTI	IPS	S2	987	1-2 juta	1-2 Orang	edit | hapus
3	12.7032	HIDAYAH F	IPS	S2	954	<= 1 juta	1-2 Orang	edit | hapus
4	12.7045	KHOIRUL ANAM	IPS	S2	959	<= 1 juta	3-4 Orang	edit | hapus
5	12.7054	M. TRI	IPS	S2	939	<= 1 juta	3-4 Orang	edit | hapus
6	12.7050	MIA DIYAN	IPS	S2	974	<= 1 juta	3-4 Orang	edit | hapus
7	12.7076	NUR ROHMAN	IPS	S2	1009	<= 1 juta	1-2 Orang	edit | hapus
8	12.7087	REYNITA AYU	IPS	S2	944	1-2 juta	1-2 Orang	edit | hapus
9	12.7086	SINGGIH P	IPS	S2	950	1-2 juta	> 4 Orang	edit | hapus
10	13.0000	SINTIA KIKI	IPS	S2	937	1-2 juta	> 4 Orang	edit | hapus
11	12.7089	SITI MUAROPAH	IPS	S2	903	1-2 juta	> 4 Orang	edit | hapus
12	12.3456	Dina Maurina	IPA	A1	900	<= 1 juta	1-2 Orang	edit | hapus
13	12346	AYU	IPA	A1	980	<= 1 juta	1-2 Orang	edit | hapus
14	16677	LALA	IPS	S2	800	<= 1 juta	3-4 Orang	edit | hapus
15	1222	NINA	IPS	S2	999	<= 1 juta	3-4 Orang	edit | hapus
16	11111	DDD	IPA	A1	994	<= 1 juta	1-2 Orang	edit | hapus
17	67777	dinnna	IPA	A1	970	<= 1 juta	1-2 Orang	edit | hapus

3. Form uji kelayakan

No	NIS	Nama	Jurusan	Kelas	Jml Nilai	Penghasilan Ortu	Sdr Kandung	Lulus/Tidak
1	12.7022	ERIC FIRDYAN	IPS	S2	722	1-2 juta	3-4 Orang	TIDAK
2	12.7027	FUJI RIFAYANTI	IPS	S2	987	1-2 juta	1-2 Orang	TIDAK
3	12.7032	HIDAYAH F	IPS	S2	954	<= 1 juta	1-2 Orang	TIDAK
4	12.7045	KHOIRUL ANAM	IPS	S2	959	<= 1 juta	3-4 Orang	YA
5	12.7054	M. TRI	IPS	S2	939	<= 1 juta	3-4 Orang	TIDAK
6	12.7050	MIA DIYAN	IPS	S2	974	<= 1 juta	3-4 Orang	YA
7	12.7076	NUR ROHMAN	IPS	S2	1009	<= 1 juta	1-2 Orang	TIDAK
8	12.7087	REYNITA AYU	IPS	S2	944	1-2 juta	1-2 Orang	TIDAK
9	12.7086	SINGGIH P	IPS	S2	950	1-2 juta	> 4 Orang	YA
10	13.0000	SINTIA KIKI	IPS	S2	937	1-2 juta	> 4 Orang	TIDAK
11	12.7089	SITI MUAROPAH	IPS	S2	903	1-2 juta	> 4 Orang	TIDAK
12	12.3456	Dina Maurina	IPA	A1	900	<= 1 juta	1-2 Orang	TIDAK
13	12346	AYU	IPA	A1	980	<= 1 juta	1-2 Orang	TIDAK
14	16677	LALA	IPS	S2	800	<= 1 juta	3-4 Orang	TIDAK
15	1222	NNA	IPS	S2	999	<= 1 juta	3-4 Orang	YA
16	11111	DDD	IPA	A1	994	<= 1 juta	1-2 Orang	TIDAK
17	67777	dinnna	IPA	A1	970	<= 1 juta	1-2 Orang	TIDAK

4. Dari hasil yang di dapatkan, untuk data training dan data testing 80%:20% yaitu data training sebanyak 88 dan data testing sebanyak 22, mendapatkan akurasi yang cukup baik yaitu 77%, precision 83%, recall 55%, dan F-Measure 66%.

Presentase Data	*Data Training*	*Data Testing*	Akurasi	Precision	Recall	F-Measure
80%:20%	88	22	77%	83%	55%	66%

IV. KESIMPULAN DAN SARAN

Kesimpulan

Penerapan metode pohon keputusan terhadap data siswa SMA Muhammadiyah Gubug memiliki tingkat akurasi yang cukup baik dalam menyelesaikan klasifikasi rekomendasi beasiswa, dengan demikian metode pohon keputusan merupakan metode yang cukup sesuai untuk penyelesaian studi kasus dalam pemilihan siswa yang mendapatkan rekomendasi beasiswa. Tingkat akurasi yang dihasilkan oleh metode tersebut adalah 77%

Saran

1. Penelitian selanjutnya sebaiknya menggunakan data yang lebih banyak agar menghasilkan *rules* yang lebih akurat.
2. Penelitian selanjutnya sebaiknya menggunakan atribut yang lebih banyak agar menghasilkan data yang lebih akurat.
3. Pengujian metode ini belum sampai pada implementasi yang menghitung iterasi c4.5 kemudian menghasilkan rule dan dapat menentukan keputusan, maka perlu dibuat sistem aplikasi yang diperuntukkan untuk pihak sekolah.
4. Untuk penelitian selanjutnya sebaiknya uji data sampai 3 kali atau lebih dengan presentase yang berbeda untuk mengetahui perbandingannya.

DAFTAR PUSTAKA

[1] "Kamus Besar Bahasa Indonesia," [online].Available: http://kbbi.web.id/ [Accessed 21 Desember 2014].

[2] Hermawati, F.A, *Data Mining*. Yogyakarta : ANDI, 2013.

[3] Nofriansyah, Dicky.*Konsep Data Mining Vs Sistem Pendukung Keputusan*.Yogyakarta:Deepublish,2014.

[4] Yosoa Putra Raharja, "*Rancang Bangun Sistem Rekomendasi Beasiswa Menggunakan Algoritma Klasifikasi C4.5 Pada Universitas Dian Nuswantoro*," Universitas Dian Nuswantoro, Semarang, skripsi 2014.

[5] Swastina, L. Penerapan Algoritma C4.5 Untuk Penentuan Jurusan Mahasiswa. *GEMA AKTUALITA,Vol.2 No.1*. 93-98, Juni 2013.

[6] Sunjana, "Klasifikasi Data Nasabah Sebuah Asuransi Menggunakan Algoritma C4.5," *Seminar Nasional Aplikasi Teknologi Informasi*, pp. D31-D34, Juni 2010.

[7] K. Hastuti, "ANALISIS KOMPARASI ALGORITMA KLASIFIKASI DATA MINING UNTUK PREDIKSI MAHASISWA NON AKTIF," *Semantik*, pp. pp. 241-249, 2012.

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

[9] Jiawei Han and Micheline Kamber, *Data Mining Concepts and Techniques*, 2nd ed. San Francisco: Morgan Kaufmann Publishers, 2006.

[10] Rahmadya T. H dan Herlawati Prabowo P. W, *Penerapan Data Mining dengan Matlab*. Bandung: Rekayasa Sains, 2013.

[11] Dwi Untari, "DATA MINING UNTUK MENGANALISA PREDIKSI MAHASISWA BERPOTENSI NON-AKTIF MENGGUNAKAN METODE DECISION TREE C4.5," Universitas Dian Nuswantoro, Semarang, pdf skripsi 2014.

[12] V Wiratna Sujarweni, *Metodologi Penelitian*. Yogyakarta: PUSTAKABARUPERSS, 2014.

[13] Eko Prasetyo, *Mengolah Data Menjadi Informasi Menggunakan Matlab*, Aldo Sahala, Ed. Yogyakarta: ANDI, 2014.