ALGORITMA C4.5 UNTUK KLASIFIKASI CALON PESERTA LOMBA CERDAS CERMAT SISWA SMP N 1 WINONG TINGKAT KABUPATEN

Candraningsih[1], Bowo Nurhadiyono[2]
Universitas Dian Nuswantoro, Ilmu Komputer, Teknik Informatika
Jl.Imam Bonjol 205, Semarang, Jawa Tengah, 50131, (024) 3517261
E-mail : chandraningsih1604@gmail.com[1], bowo.nurhadiono@dsn.dinus.ac.id[2]

Abstrak *- Sekolah merupakan lembaga penyelenggara pendidikan akademik bagi siswa. Dalam proses pembelajaran di sekolah dalam jangka waktu tertentu maka akan terkumpul sejumlah data yang besar yang nantinya akan menyulitkan pihak sekolah untuk mengolah data tersebut sehingga berpengaruh dalam peningkatan mutu siswa yang dihasilkan, dan dalam skala besar akan menurunkan prestasi sekolah dilihat dari sedikitnya prestasi dari siswa yang mendapatkan gelar juara dalam sebuah perlombaan. Salah satu faktor penyebab menurunnya prestasi akademik SMP N 1 Winong adalah banyaknya data dan kriteria yang digunakan dalam proses seleksi calon peserta lomba cerdas cermat sehingga pihak sekolah kurang tepat dalam mengirimkan perwakilan lomba. Data mining dapat menggali informasi dari data yang jumlahnya sangat besar dengan metode- metode tertentu untuk mendapat informasi atau ilmu pengetahuan yang baru. Dengan metode klasifikasi yang digunakan dapat diketahui apakah siswa layak menjadi calon peserta lomba atau tidak. Oleh karena itu data mining bisa digunakan untuk mengklasifikasikan data calon peserta lomba sebagai sarana untuk menerapkan algoritma C4.5 dalam proses seleksi calon peserta lomba cerdas cermat siswa SMP N 1 Winong tingkat kabupaten. Hasil klasifikasi dari algoritma C4.5 untuk mengetahui tingkat akurasi dalam membuat klasifikasi calon peserta lomba cerdas cermat. Hasil evaluasi diperoleh bahwa algoritma C4.5 memiliki akurasi 95,45%. Rule yang diperoleh dari klasifikasi dengan algoritma C4.5 jika diterapkan dalam data baru diperoleh hasil validasi dengan tingkat akurasi 90,63%.*

Kata Kunci*: data mining, klasifikasi, algoritma C4.5, confussion matrik, akurasi, lomba cerdas cermat*

Abstract *- School was an institution of academic education administrator for students. The learning process in schools within a certain time period to produce a large amount of data would be difficult for the school for processing such data so that the effect of improving the quality of students produced. On a large scale would lower the school achievement seen from at least the achievements of students who obtain title in a race. One of the factors causing the decline in academic achievement SMP N 1 Winong was the number of data and criteria used in the selection process of candidates quiz competition so that the schools are less precise in the race to send a representative. Data mining could digged up information from very large amounts of data with specific methods to obtain information or a new science. Through the classification method used could be known whether the student deserves to be prospective competitor or not. Therefore, data mining could be used to classify the data prospective participants as a means to implement the algorithm C4.5 in the selection process of candidates quiz competition students of SMP N 1 Winong district level. Results of algorithm C4.5 classification used to determine the level of accuracy in classifying candidates quiz competition. Results of the evaluation showed that the algorithm C4.5 had an accuracy of 95.45%. Rule derived from C4.5 classification algorithm if implemented in the new data obtained validation results with an accuracy rate of 90.63%.*

.

Keyword : *data mining, classification, C4.5 algorithm, confussion matrix, accuracy, cerdas cermat comprtition*

I. PENDAHULUAN

Sekolah merupakan lembaga penyelenggara pendidikan akademik bagi siswa. Dalam proses pembelajaran di sekolah dalam jangka waktu tertentu, maka akan terkumpul sejumlah data yang sangat besar. Berangkat dari banyaknya data yang akan diolah mengakibatkan pihak sekolah kesulitan untuk menentukan kriteria yang akan diolah. Kumpulan data tersebut akan diproses lebih lanjut dengan data mining untuk memperoleh pola baru yang dapat digunakan untuk meningkatkan efektifitas dalam proses pembelajaran.

Data mining merupakan proses yang menggunakan teknik statistic, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar [1]. Istilah data mining memiliki hakikat sebagai disipin ilmu yang tujuan utamanya adalah untuk menemukan, menggali, atau menambang pengetahuan dari data atau informasi yang kita miliki [*2*]. Metode, teknik, maupun algoritma yang digunakan dalam data mining sangatlah bervariasi. Pemilihan metode dan algoritma yang tepat, semuanya bergantung dengan tujuan dan proses secara keseluruhan. Data mining mampu menganalisa data yang sangat besar sehingga mampu memberikan informasi maupun arti bagi pendukung keputusan yang akan diambil nantinya.

Data sampel yang digunakan dalam penelitian kali ini data nilai siswa calon peserta lomba cerdas cermat tingkat kabupaten yang setiap tahunnya pihak sekolah mengirimkan perwakilannya. Banyaknya kriteria yang digunakan membuat pihak sekolah kesulitan dalam mengirimkan peserta lomba. Melalui penelitian ini diharapkan pihak sekolah akan tepat dalam mengirimkan perwakilan lomba cerdas cermat tersebut.

II. METODE YANG DIUSULKAN

2.1 Tinjauan Studi

Tabel 1. Penelitian terkait

No	Nama Peneliti	Judul
1	Anik Andriani (2012)	Penerapan Algoritma C4.5 pada Program Klasifikasi Mahasiswa Dropout
2	Angga Ginanjar Mabrur, Riani Lubis (2012)	Penerapan Data Mining untuk Memprediksi Kriteria Nasabah Kredit
3	Budanis Dwi Meilani dan Achmad Fauzi Slamet (2010)	Klasifikasi Data Karyawan Untuk Menentukan Jadwal Kerja Menggunakan Metode Decision Tree

Dari penelitian diatas penulis berusaha mengembangkan dari penelitian yang sudah ada. Oleh karenanya penulis menggunakan metode klasifikasi dan algoritma C4.5 untuk mencari tingkat akurasi dari penelitian yang telah dilakukan.

2.1 Pengelompokan Data Mining

Ada beberapa teknik yang dimiliki data mining berdasarkan tugas yang bisa dilakukan, antara lain [3]:

a. Deskripsi
   Para peneliti biasanya mencob menemukan cara mendeskripsikan pola dan trend yang tersembunyi dalam data.
b. Estimasi
   Estimasi hamper sama dengan klasifikasi, kecuali variable tujuan yang lebih kearah numeric daripada kategori.
c. Prediksi
   Prediksi memiliki beberapa kemiripan dengan estimasi dan klasifikasi. Hanya saja jika prediksi hasilnya menunjukkan sesuatu yang belum terjadi (mungkin terjadi di masa depan).

d. Klasifikasi
   Dalam klasifikasi variable, tujuan bersifat kategorik. Misalnya, kita akan mengklasifikasikan pendapatan dalam tiga kelas, yaitu pendapatan tinggi pendapatan sedang, dan pendapatan rendah
e. Clustering
   Clustering lebih condong kea rah pengelompokan rcord, pengamatan, atau kasus dalam kelas yang memiliki kemiripan.
f. Asosiasi
   Mengidentifikasi hubungan antara berbagai peristiwa yang terjadi pada suatu waktu.

2.2 Klasifikasi

Klasifikasi merupakan salah satu teknik dalam data mining. Klasifikasi (taksonomi) merupakan proses penempatan objek atau konsep tertentu ke dalam satu set kategori berdasarkan objek yang digunakan. Salah satu teknik klasifikasi yang popular digunakan adalah decision tree [4]. Klasifikasi sendiri terbagi menjadi dua tahap, yaitu pengklasifikasian dan pembelajaran. Pada tahap pembelajaran, sebuah algoritma klasifikasi akan membangun sebuah model klasifikasi dengan cara menganalisis training data. Tahap pembelajaran dapat juga dipandang sebagai tahap pembentukan fungsi atau pemetaan y=f(x) di maana y adalah kelas hasil prediksi dan X adalah truple yang ingin diprediksi kelasnya.

2.3 Algoritma C4.5

Algoritma C4.5 merupakan salah satu algoritma yang telah secara luas digunakan, khususnya di area machine learning yang memiliki beberapa perbaikan dari algoritma sebelumnya yaitu ID3. Algoritma C4.5 dan ID3 model yang tak terpisahkan, karena membangun sebuah pohon keputusan, dibutuhkan algoritma C4.5 Diakhir tahun 1980- an, J. Ross Quinlan seorang peneliti di bidang mesin pembelajaran mengembangkan sebuah model pohon keputusan yang dinamakan ID3. Ada beberapa tahapan dalam membuat sebuah pohon keputusan dalam algoritma C4.5 [5] yaitu:

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

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

Keterangan :
S = Himpunan Kasus
n = Jumlah Partisi S
pi = proporsi S_i terhadap S

Kemudian setelah nilai entropy pada masing-masing atribut sudah diperoleh maka hitung nilai *gain* dengan menggunakan rumus :

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

Keterangan :
S= Himpunan kasus
A = Fitur
n = jumlah partisi atribut A
$|S_i|$ = Proporsi S_i terhadap S
|S| = Jumlah Kasus dalam S

3 METODE PENELITIAN

3.1 Objek Penelitian

Penulis melakukan penelitian di SMP Negeri 1 Winong yang beralamtkan di JL. Raya Wnong Gabus KM 0.5. Penelitian ini dilakukan pada bulan September 2014. Adapun penelitian ini dilakukan untuk mengetahui calon siswa yang diprediksi untuk mengikuti lomba cerdas cermat yang tingkat kabupaten yang setiap tahun selalu rutin diadakan.

3.2 Teknik Analisis Data

Data yang digunakan dalam penelitian ini adalah data berdasarkan kriteria yang digunakan dalam perhitungan, yaitu pada siswa kelas VIII semester ganjil tahun ajaran 2013/2014 SMP Negeri 1 Winong yang digunakan untuk perhitungan alternatif tertinggi penentuan siswa yang akan mengikuti lomba cerdas cermat tingkat Kabupaten. Metode yang diusulkan untuk proses seperti yang telah dijelaskan di atas yaitu metode klasifikasi dengan algoritma yang digunakan adalah algoritma C4.5 dengan kriteria yang digunakan sebagai berikut :

1. Nilai Bahasa Indonesia
2. Nilai Bahasa Inggris
3. Nilai Biologi
4. Nilai Fisika
5. Nilai Matematika
6. Nilai IPS
7. Nilai keaktifan (meliputi keaktifan mengerjakan soal dan menjawab pertanyaan ketika bimbingan belajar berlangsung
8. Perolehan skor IQ
9. Nilai Bimbingan belajar

Tabel 3.1 Data yang digunakan

B.Ind	B.Ing	BIO	FIS	MAT	IPS	Aktif	IQ	Bim	Hasil
93	84	86	82	83	84	A	107	83	L
80	84	80	83	82	78	K	94	85	TL
80	76	84	84	85	82	K	102	80	TL
80	83	85	83	81	85	A	128	86	TL
88	83	85	90	90	80	K	90	83	L

Tabel 3.2 Konversi Nilai

Nilai	Klasifikasi
86-100	A
71-85	B
56-70	C
41-55	D
≤ 40	E

Tabel 3.3 Konversi IQ

Range	Kategori	Klasifikasi
>= 140	Genius	5
120 – 139	Superior	4
110 – 119	Diatas rata- rata	3
90 – 109	Rata- rata	2
<= 89	Dibawah rata-rata	1

Tabel 3.4 Hasil Konversi

B.Ind	B.Ing	BIO	FIS	MAT	IPS	Aktif	IQ	Bim	Hasil
A	B	A	B	B	B	A	2	B	L
B	B	B	B	B	B	K	2	B	TL
B	B	B	B	B	B	K	2	B	TL
B	B	B	B	B	B	A	4	A	TL
A	B	B	A	A	B	K	2	B	L

Tabel 3.5 Perhitungan gain dan entropi

Node	Atribut	Kategori	Jml_kasus	L	TL	Entropi	Gain
1	Total		132	26	106	0.71	
	B.Ind	A	15	11	4	0.82	0.13 Gain tertinggi
		B	117	15	102	0.55	
	B.Ing	A	16	8	8	1	0.05
		B	116	18	98	0.62	
	Bio	A	14	8	6	0.98	0.06
		B	118	18	100	0.61	
	Fis	A	26	4	2	0.91	0.04
		B	106	22	104	0.66	
	Mat	A	13	9	4	0.61	0.1
		B	119	17	102	0.09	
	IPS	A	12	6	6	1	0.03
		B	120	20	120	0.65	
	Bim	A	11	5	6	0.99	0.02
		B	121	21	100	0.66	
	N.Aktif	A	26	20	6	0.71	0
		K	106	82	24	0.72	
	IQ	1	1	0	1	0	0.01
		2	111	23	88	0.73	
		3	13	1	12	0.39	
		4	7	2	5	0.86	
		5	0	0	0	0	

3.3 CONFUSSION MATRIX

Correct Classfication	*Classified as*	
	+	-
+	*True positives*	*False Negatives*
–	*False positives*	*True Negatives*

Tabel 3.6 Confussion Matrix

Kolom menyatakan prediksi klasifikasi, sedangkan baris menyatakan klasifikasi sebenarnya. Evaluasi dengan *confusion matrix* menghasilkan nilai *accuracy*, dimana a*ccuracy* dalam klasifikasi adalah persentase ketepatan *record* data yang diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi. False positive (FP) adalah hasil yang diprediksi positif (yes) namun pada klasifikasi sebenarnya hasilnya negative (no). False negative adalah hasil yang diprediksi negative (no) namun pada klasifikasi sebenarnya hasilnya positif (yes).

Tingkat akurasi dari seluruh klasifikasi ditentukan dengan jumlah klasifikasi yang benar dibagi dengan total jumlah record klasifikasi.

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

$$= \frac{20+106}{20+6+106}$$

$$= 0.9545$$

Untuk menghitung prosentasi akurasi, maka tingkat sukses dikalikan 100%. Ini berarti prosentase error dapat dicari dengan cara 100% dikurangi dengan prosentase sukses.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} * 100\%$$

$$= \frac{20+106}{20+6+106} * 100\%$$

$$= 95.45\%$$

4 PENUTUP

4.1 Kesimpulan

Berdasarkan hasil penelitian pada klasifikasi penjurusan siswa dapat diambil beberapa kesimpulan sebagai berikut:

1. Klasifikasi proses seleksi calon peserta lomba siswa SMP N 1 Winong dapat mengklasifikasikan siswa dalam tahapan lolos atau tidaknya dalam seleksi.
2. Dari 132 data siswa yang digunakan menunjukkan tingkat akurasi dengan algoritma C4.5 sebesar 95,45% dan pengujian data uji baru sebanyak 32 data pada tahun sebelumnya diperoleh tingkat akurasi sebesar 90,63%.
3. Penerapan *rules* dari algoritma C4.5 selanjutnya diterapkan pada bahasa pemrogramman PHP yang digunakan dalam klasifikasi hail proses seleksi yang berupa lolos atau tidaknya siswa sebagai calon peseta lomba.

REFERENCES

[1] Adi Suwondo, Dian Asmarajati, and Heri Surahman, "Algoritma C4.5 Berbasis Adaboost untuk Prediksi Penyakit Jantung Koroner," Juni 2013.

[2] Fatayat and Joko Risanto, "Proses Data Mining dalam Meningkatkan Sistem Pembelajaran pada Pendidikan Sekolah Menengah Pertama," 2013.

[3] Anik Andriani, "Penerapan algoritma C4.5 Pada program klasifikasi mahasiswa dropout," 2012.

[4] Bain.K, Holisatul Munawaroh, and Yeni Kustiyahningsih, "Perbandingan algoritma ID3 dan C5.0 dalam identifikasi penjurusan siswa SMA," Juni 2013.

[5] Swastina Liliana, "Penerapan Algoritma C4.5 Untuk Penentuan Jurusan Mahasiswa," *Gema Aktualita*, Juni 2013.