Jurnal Inkofar * Volume 1 No. 2, Desember 2018 * ISSN: 2615-3645 (Print) / 2581-2920 (Online)
Tersedia secara online di: http://www.politeknikmeta.ac.id/meta/ojs/

ALGORITMA C4.5 UNTUK KLASIFIKASI CALON PESERTA LOMBA CERDAS CERMAT SISWA SMP DENGAN MENGGUNAKAN APLIKASI RAPID MINER

Dian Ardiansyah [1], Walim Walim[2]

[1] Pascasarjana Magister Ilmu Komputer / Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Nusa Mandiri / dian.did@bsi.ac.id
[2] Pascasarjana Magister Ilmu Komputer / Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Nusa Mandiri / walim.wam@bsi.ac.id

ABSTRACT

The school is a provider of education for students. The student is an input component in the educational system which further processed in the education process so that the quality of being human according to the purpose of national education. In the learning process in schools within a certain period then it will have accumulated large amounts of data that would later complicate the school to process such data so influential in increasing the quality of students is produced, and in large scale will lower school achievement seen from at least the achievements of students who get the title in a race. One of the factors the causes of declining student academic performance is the number of data and criteria used in the selection process for potential participants to the activities of the intelligent races carefully so that the school is less precise in sending representatives of the destination. Data mining can dig out information from a very large amount of data with specific methods to obtain information or new science. A method of classification is used to determine whether students qualify potential participants of the race or not. Therefore, data mining can be used to classify the data potential participants of the race as a means to implement the algorithm C 4.5 in the selection process of candidates participants race savvy JUNIOR students carefully. The results of the classification of the algorithm C 4.5 to know the level of accuracy in making the classification of potential participants of the race closely intelligent students of JUNIOR HIGH SCHOOL. Evaluation of the results obtained that the algorithm C 4.5 81.81% accuracy.

Keywords : Algorithm C4.5, Careful, Intelligent Race Classification, Accuracy, Confusion Matrix, Data Mining

ABSTRAK

Sekolah adalah suatu lembaga penyelenggara pendidikan bagi siswa. Siswa merupakan sebuah komponen masukan dalam sistem pendidikan yang selanjutnya diproses dalam proses pendidikan sehingga menjadi manusia yang berkualitas sesuai dengan tujuan pendidikan nasional. 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 siswa adalah banyaknya data dan kriteria yang digunakan dalam proses seleksi calon

peserta untuk kegiatan lomba cerdas cermat sehingga pihak sekolah kurang tepat dalam mengirimkan perwakilan lombanya. Data mining dapat menggali informasi dari data yang jumlahnya sangat besar dengan metode-metode tertentu untuk mendapat informasi atau ilmu pengetahuan yang baru. Metode klasifikasi digunakan untuk mengetahui 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. Hasil klasifikasi dari algoritma C4.5 untuk mengetahui tingkat akurasi dalam membuat klasifikasi calon peserta lomba cerdas cermat siswa SMP. Hasil evaluasi diperoleh bahwa algoritma C4.5 memiliki akurasi 81,81%.
Kata Kunci : Algoritma C4.5, lomba cerdas cermat, klasifikasi, akurasi, Confusion matrix, data mining

1. PENDAHULUAN

Sekolah merupakan lembaga penyelengara pendidikan akademik bagi siswa. Siswa adalah komponen masukan dalam sistem pendidikan yang selanjutnya diproses dalam proses pendidikan sehingga menjadi manusia yang berkualitas sesuai dengan tujuan pendidikan nasional. Dalam proses pemebelajaran di sekolah dalam jangka waktu tertentu maka akan terkumpul sejumlah data yang besar yang nantinya akan menyulitkan pihak sekolah untuk mengolah data. 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 mengekstrasi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar. Data sampel yang digunakan yaitu data nilai siswa calon peserta lomba cerdas cermat yang setiap tahunnya pihak sekolah mengirimkan perwakilannya.

2. LANDASAN TEORI

2.1. *Data mining*

Data mining telah banyak menarik perhatian di dunia sistem informasi dan di masyarakat secara keseluruhan dalam beberapa tahun ini, karena ketersediaan luas dalam jumlah besar data dan kebutuhan segera untuk mengubah data tersebut menjadi informasi yang berguna dan pengetahuan. Informasi dan pengetahuan yang diperoleh dapat digunakan untuk aplikasi mulai dari pasar analis, deteksi penipuan, dan retensi pelanggan, untuk pengendalian produksi dan ilmu pengetahuan eksplorasi (Han & Kember, 2012).

Data mining adalah proses menemukan korelasi baru yang bermakna, pola dan tren dengan memilah-milah sejumlah besar data yang tersimpan dalam repositori, menggunakan teknologi penalaran pola serta teknik-teknik statistik dan matematika (Larose & Larose, 2014).

1. Pengelompokan *Data mining*

Ada beberapa teknik yang dimiliki *data mining* berdasarkan tugas yang bisa dilakukan, antara lain:

a. Deskripsi
   Para peneliti biasanya mencoba menemukan cara mendeskripsikan pola dan trend yang tersembunyi dalam data.

b. *Estimasi*
   Estimasi hampir sama dengan klasifikasi, kecuali *variable* tujuan yang lebih kearah *numeric* dari pada 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 kategori. Misalnya, kita akan mengklasifikasikan pendapatan dalam tiga kelas, yaitu pendapatan tinggi pendapatan sedang, dan pendapatan rendah.
e. *Clustering*
   Clustering lebih condong kearah pengelompokan record, 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*.

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) dimana 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 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:

$$Entropy(S) = \sum_{i=1}^{n} - pi * Log2\ pi \qquad \text{persamaan(1)}$$

Keterangan:

S : Himpunan Kasus
n : Jumlah Partisi S
pi : Proporsi dari Si terhadap S

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

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|si|}{|s|} * Entropy(Si)$$ persamaan (2)

Keterangan :

S : *Himpunan Kasus*
A : *Fitur*
n : *Jumlah partisi atribut A*
|Si| : *Jumlah kasus pada partisi ke-i*
|S| : *Jumlah kasus dalam S*

2.4. Confusion matrix

Confusion matrix adalah *tool* 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 aktual dan prediksi pada klasifikasi.

Tabel 1. Tabel *Confusion matrix* 2 Kelas

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)*

Rumus untuk menghitung tingkat akurasi pada matrik adalah:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} = \frac{A+D}{A+B+C+D}$$

2.5. Rapid Miner

Rapid Miner merupakan perangakat lunak yang bersifat terbuka (*open source*). *Rapid Miner* adalah sebuah solusi untuk melakukan analisis terhadap *data mining*, *text mining* dan analisis prediksi. *Rapid Miner* menggunakan berbagai teknik *deskriptif* dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik. *Rapid Miner* memiliki kurang lebih 500 operator *data mining*, termasuk operator untuk input, output, data *preprocessing* dan *visualisasi*. *Rapid Miner* merupakan *software* yang berdiri sendiri untuk analisis data dan sebagai mesin *data mining* yang dapat diintegrasikan pada produknya sendiri. *Rapid Miner* ditulis dengan munggunakan bahasa java sehingga dapat bekerja di semua sistem operasi.

Beberapa fitur dari *Rapid Miner*, antara lain:

1. Banyaknya *algoritma data mining*, seperti *decision tree* dan *self-organization map*.
2. Bentuk grafis yang canggih, seperti tumpang tindih diagram *histogram*, *tree chart* dan *3D Scatter plots*.
3. Banyaknya variasi *plugin*, seperti *text plugin* untuk melakukan analisis teks.
4. Menyediakan prosedur *data mining* dan *machine learning* termasuk: ETL (*extraction, transformation, loading*), data *preprocessing*, *visualisasi*, *modelling* dan evaluasi.
5. Proses *data mining* tersusun atas operator-operator yang *nestable*, dideskripsikan dengan XML, dan dibuat dengan GUI. Mengintegrasikan proyek *data mining* Weka dan statistika R.

3. HASIL DAN PEMBAHASAN

Data yang digunakan dalam penelitian ini adalah data berdasarkan kriteria yang digunakan dalam perhitungan, yaitu pada siswa SMP kelas VIII yang digunakan untuk perhitungan alternatif tertinggi penentuan siswa yang akan mengikuti lomba cerdas cermat. Metode yang diusulkan untuk proses seperti yang telah dijelaskan diatas yaitu metode klasifikasi dengan *algoritma* yang digunakan adalah *algoritma C4.5* dengan kriteria yang digunakan sebagai berikut:

1. Nama Siswa
2. Nilai Bahasa Indonesia
3. Nilai Bahasa Inggris
4. Nilai Ilmu Pengetahuan Alam (IPA)
5. Nilai Ilmu Pengetahuan Sosial (IPS)
6. Nilai Matematika
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	IPA	IPS	MTK	AKTIF	IQ	BIM	HASIL
90	90	88	98	88	A	454	85	Lolos
90	90	97	80	75	K	432	83	Lolos
86	88	86	90	83	K	433	84	Lolos
85	88	90	90	84	A	437	84	Lolos
84	86	90	90	83	A	433	85	Lolos

Setelah itu data yang sudah ada nilai-nilai mata pelajaran dan bimbingan di konversi menggunakan ini:

Tabel 3.2. Konversi Nilai

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

Selain nilai mata pelajaran dan bimbingan yang di konversi, nilai IQ juga di konversi menggunakan ini:

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	IPA	IPS	MTK	AKTIF	IQ	BIM	HASIL
A	A	A	A	A	A	2	A	Lolos
A	A	A	B	B	K	1	B	Lolos
A	A	A	A	B	K	1	B	Lolos
A	A	A	A	B	A	1	B	Lolos
B	A	A	A	B	A	1	A	Lolos

Setelah selesai semua nilai di konversi, langkah selanjutnya menentukan gain dengan cara perhitungan gain dan entropy.
Langkah-langkah membuat algoritma C4.5 dengan memakai data *training* yang berjumlah 33, yaitu:

1. Siapkan data *training* yaitu tabel 3.1 yang berjumlah 33 data.
2. Hitung jumlah calon peserta lomba cerdas cermat yang lolos dan tidak lolos berdasarkan nilai setiap atribut.
3. Hitung nilai *entropy* total dimana diketahui calon peserta cerdas cermat yang lolos sebanyak 21 siswa dan yang tidak lolos sebanyak 12 siswa.

$$Entropy(S) = \sum_{i=1}^{n} - pi * Log2\ pi$$

=(-(21/33)*(LOG10(21/33)/LOG10(2)))-((12/33)*(LOG10(12/33)/LOG10(2)))
= 0,945660305

4. Hitung nilai *gain* untuk masing-masing atribut. Kemudian tentukan nilai gain tertinggi. Atribut dengan nilai tertinggi maka atribut tersebut dijadikan sebagai akar. Sebagai contoh hitung nilai gain untuk atribut mata pelajaran IPS yaitu:

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

=0,945660305-((13/33*0)+(20/33*0,970950594))
=0,357205399

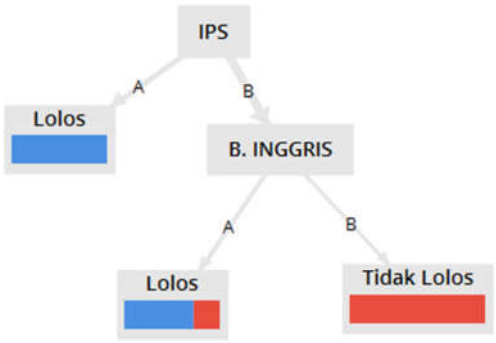
Perhitungan nilai *entropy* dan *gain* untuk semua atribut dilakukan untuk mendapatkan nilai *gain* tertinggi yang akan dijadikan sebagai akar. Hasil perhitungannya terlihat ditabel dibawah ini:

Tabel 3.5. Perhitungan gain dan entropy

Node	Atribut	Kategori	Jumlah Kasus	L	TL	Entropy	Gain
1	Total		33	21	12	0,945660305	
	B.Ind	A	20	12	8	0,970950594	-0,021463567
		B	13	8	5	0,961236605	
	B.Ing	A	21	18	3	0,591672779	0,274131037
		B	12	3	9	0,811278124	
	IPA	A	21	15	6	0,863120569	0,032765397
		B	12	6	6	1	
	IPS	A	13	13	0	0	0,357205399
		B	20	8	12	0,970950594	
	MTK	A	13	8	5	0,961236605	0,000889487
		B	20	13	7	0,934068055	
	Aktif	A	17	10	7	0,977417818	0,007699256
		K	16	11	5	0,896038233	
	IQ	1	32	20	12	0,954434003	0,020148544
		2	1	1	0	0	
	BIM	A	8	5	3	0,954434003	0,066213569
		B	25	18	7	0,855450811	

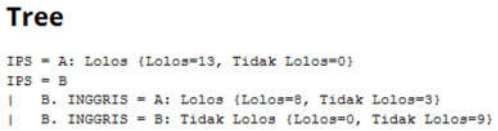
Sumber: Pengolahan Data (2017)

Berdasarkan hasil perhitungan *gain* pada tabel 1 terlihat atribut mata pelajaran IPS mempunyai nilai *gain* tertinggi yaitu 0,357205399 sehingga atribut IPS bisa dijadikan sebagai simpul akar dari pohon keputusan.



Gambar 3.1. Pohon Keputusan

Adapun hasil berupa penjelasan teks atau logika programnya seperti dibawah ini:



Gambar 3.2. Hasil Penjelasan berupa teks

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

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$
$$= \frac{18+9}{18+9+3+3}$$
$$= 0{,}8181 = 81{,}81\%$$

4. KESIMPULAN

Berdasarkan hasil pengujian pada klasifikasi Calon peserta cerdas cermat diambil beberapa kesimpulan sebagai berikut:

1. Klasifikasi proses seleksi calon peserta lomba siswa SMP dapat mengklasifikasikan siswa dalam tahapan lolos atau tidaknya dalam seleksi.
2. Dari 33 data siswa yang digunakan menunjukkan tingkat akurasi dengan algoritma C4.5 sebesar 81,81%.

5. REFERENSI

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

Han, J., & Kember, M. (2012). Data Mining Concepts & Techniques, Simon Fraser University Academic Press, US.

Fatayat and Joko Risanto, "Proses *Data mining* dalam Meningkatkan Sistem Pembelajaran pada Pendidikan Sekolah Menengah Pertama," 2013.

Kusrini, & Luthfi, E., T. (2009) Algoritma *data mining*. Yogyakarta: Andi Publishihng.

Larose, D. T., & Larose, C. D. (2014). *Discovering Knowledge in Data*. https://doi.org/10.1002/9781118874059

