ALGORITMA C4.5 DALAM DATA MINING UNTUK MENENTUKAN KLASIFIKASI KELULUSAN CALON MAHASISWA BARU (Sudi Kasus : AMIK-DP)

Fitriany

AMIK Depati Parbo Kerinci
Email : fitamik1@yahoo.com

ABSTRAK

Algoritma C4.5 adalah salah satu algoritma yang terdapat pada data mining yang merupakan salah satu algoritma yang digunakan untuk melakukan klasifikasi data dengan membentuk pohon keputusan. Pohon keputusan algoritma C4.5 dibangun dengan beberapa tahap yang meliputi pemilihan atribut sebagai akar, membuat cabang untuk tiap-tiap nilai dan membagi kasus dalam cabang. Tahap-tahap ini akan diulangi untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Dari penyelesaian pohon keputusan maka akan didapatkan beberapa rule suatu kasus. Dalam hal ini penulis mengklasifikasikan kelulusan dari calon mahasiswa baru pada suatu Perguruan Tinggi (AMIK Depati Parbo Kerinci) tidak hanya berdasarkan kriteria Nilai Ujian Tertulis saja tetapi juga keahlian lainnya yang berhubungan dengan komputer, kehadiran waktu ujian wawancara dan nilai rata-rata UAN waktu tamat SMA yang juga menjadi tolak ukur seseorang untuk dapat diterima di sebuah perguruan tinggi. Dengan penerapan algoritma C4.5 ini akan dapat membantu pihak akademik dalam menentukan calon mahasiswa baru yang benar-benar ingin mengikuti perkuliahan.

Kata kunci : data mining, klasifikasi, algoritma C4.5, pohon keputusan

PENDAHULUAN

Mahasiswa merupakan bagian yang penting bagi suatu Universitas. Banyaknya mahasiswa yang berminat masuk pada perguruan tinggi tersebut sangat berpengaruh pada kualitas dari perguruan tinggi. Berdasarkan hal tersebut, salah satu cara untuk untuk menentukan kelulusan calon mahasiswa baru menggunakan bahasa sehari-hari yang nantinya dalam pengambilan keputusan akan lebih mudah dipahami guna menjadikan sebagai pembelajaran untuk kedepannya dalam menghasilkan para calon mahasiswa baru yang bermutu.

Peneliti mencoba untuk memberikan informasi dalam menentukan kelulusan calon mahasiswa baru di AMIK Depati Parbo Kerinci. Selama ini informasi yang diberikan dalam menentukan kelulusan calon mahasiswa baru hanya bisa dilihat berdasarkan Nilai Tes Seleksi Penerimaan Mahasiswa Baru/Nilai Tertulis, yang menunjukkan kualitas calon mahasiswa baru secara umum tanpa mengetahui keahlian yang benar-benar dimiliki oleh calon mahasiswa tersebut.

Dalam pengambilan keputusan penerimaan mahasiswa baru ini dibatasi dengan memperhatikan beberapa atribut yaitu Nilai Ujian Tertulis, wawancara, nilai ujian praktek dan nilai UAN. Metode yang digunakan adalah metode Algoritma C4.5 yang meruakan salah satu konsep algoritma yang ada pada data mining. Algoritma C4.5 adalah salah satu algoritma yang digunakan untuk melakukan klasifikasi data dengan membentuk pohon keputusan sehingga nantinya memberikan data yang lengap dan akan lebih mudah bagi pihak universitas menentukan seorang calon mahasiswa baru diterima atau tidak di universitas (AMIK Depati Parbo).

Permasalahan

Permasalahan yang ada adalah bagaimana memperoleh pengetahuan dan mengklasifikasikan calon mahasiswa baru yang akan diterima pada sebauh universitas (AMIK Depati Parbo) untuk menentukan siapa saja yang berhak masuk atau diterima berdasarkan Nilai Ujian Tertulis, wawancara, nilai ujian praktek dan nilai UAN (Ujian Akhir Nasional). Pengambilan keputusan untuk kelulusan calon mahasiswa baru AMIK Depati Parbo Kerinci masih dalam bentuk data-data yang pasti, yang pada prinsipnya masih sulit dalam proses pengambilan keputusan. Metode yang digunakan untuk menyajikan informasi kelulusan mahasiswa adalah dengan membentuk pohon keputusan dengan Algoritma C4.5 dan menguji kebenaran hasil klasifikasi kelulusan calon mahasiswa baru dengan menggunakan aplikasi *WEKA GUI Chooser*.

Data Mining

Data mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual. Kata *mining* berarti usaha untuk mendapatkan sedikit barang berharga dari sejumlah besar material dasara (Pramudiono, 2003). *Data mining* merupakan proses pencarian pola relasi-relasi yang tersembunyi dalam sejumlah data yang besar dengan tujuan untuk melakukan klasifikasi, estimasi, prediksi, *association rule*, *clustering*, deskripsi dan visualisasi (Han dan Kamber, 2001).

Klasifikasi

Klasifikasi dan prediksi adalah dua bentuk analisis data yang bisa digunakan untuk mengekstrak model dari data yang berisi kelas-kelas atau untuk memprediksi *trend* data yang akan datang. Klasifikasi memprediksi data dalam bentuk kategori, sedangkan prediksi memodelkan fungsi-fungsi dari nilai yang kontinyu. Misalnya model klasifikasi bisa dibuat untuk mengelompokkan aplikasi peminjaman pada bank apakah berisiko atau aman, sedangkan model prediksi bisa dibuat untuk diprediksi pengeluaran untuk membeli peralatan komputer dari pelanggan potensial berdasarkan pendapatan dan lokasi tinggalnya.

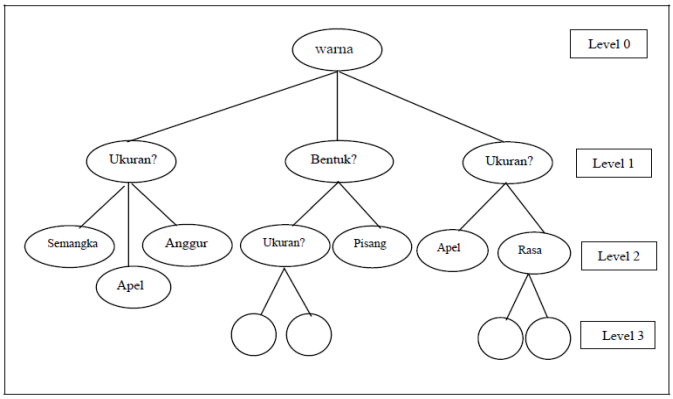
Secara umum, proses klasifikasi dapat dilakukan dalam dua tahap, yaitu proses belajar dari data pelatihan dan klasifikasi kasus baru. Pada proses belajar, algoritma klasifikasi mengolah data pelatihan untuk menghasilkan sebuah model. Setelah model diuji dan dapat diterima, pada tahap klasifikasi, model tersebut digunakan untuk memprediksi kelas dari kasus baru untuk membantu proses pengambilan keputusan (Han, 2001; Quinlan, 1993). Kelas yang dapat diprediksi adalah kelas-kelas yang sudah terdefinisi pada data pelatihan. Karena proses klasifikasi kasus baru cukup sederhana, penelitian lebih banyak ditujukan untuk memperbaiki teknik-teknik pada proses belajar, seperti gambar berikut.

***Decision Tree* (Pohon Keputusan)**

Decision Tree (Pohon Keputusan) adalah sebuah diagram alir yang mirip dengan struktur pohon, di mana setiap *internal node* menotasikan atribut yang diuji, setiap cabangnya merepsentasikan hasil dari atribut tes tersebut, dan *leaf node* merepresentasikan kelas-kelas tertentu atau distribusi dari kelas-kelas (Han & Kamber, 2001).

Seringkali untuk mengklasifikasikan obyek, kita ajukan urutan pertanyaan sebelum bisa kita tentukan kelompoknya. Jawaban pertanyaan pertama akan mempengaruhi pertanyaan berikutnya dan seterusnya. Dalam *decision tree*, pertanyaan pertama akan kita tanyakan pada simpul akar pada level 0. Jawaban dari pertanyaan ini dikemukakan dalam cabang-cabang. Jawaban dalam cabang akan disusul dengan pertanyaan kedua lewat simpul yang berikutnya pada *level* 1. Dalam setiap *level* ditanyakan nilai atribut melalui sebuah simpul. Jawaban dari

pertanyaan ini dikemukakan lewat cabang-cabang. Langkah ini akan berakhir di suatu sumpul jika pada simpul tersebut sudah ditemukan kelas atau jenis obyeknya. Kalau dalam satu tingkat suatu obyek sudah diketahui termasuk dalam kelas tertentu, maka kita berhenti di *level* tersebut. Jika tidak, maka dilanjutkan dengan pertanyaan di *level* berikutnya hingga jelas ciri-cirinya dan jenis obyek dapat ditentukan (Santosa, 2007), seperti pada gambar 4.4 dibawah ini contoh penggunaan metode *decision tree* untuk menentukan jenis buah.



Gambar 1. Contoh penggunaan metode Decision Tree untuk menentukan jenis buah

Algoritma C4.5

Algoritma C4.5 adalah algoritma klasifikasi data dengan teknik pohon keputusan yang terkenal dan disukai karena memiliki kelebihan-kelebihan. Kelebihan ini misalnya : dapat mengolah data numerik (kontinyu) dan diskret, dapat menangani nilai atribut yang hilang, menghasilkan aturan-aturan yang mudah diinterpretasikan dan tercepat di antara algoritma-algoritma yang menggunakan memori utama di komputer (Quinlan, 1993; Han *et al.*, 2001; Berry *et al.*, 1997; Ruggieri, 2001).

Pada tahap belajar dari data, algoritma C4.5 mengkonstruksi pohon keputusan dari data pelatihan, yang berupa kasus-kasus atau rekord-rekord (tupel) dalam basisdata. Setiap kasus berisikan nilai dari atribut-atribut untuk sebuah kelas. Setiap atribut dapat berisi data diskret atau kontinyu (numerik). C4.5 juga menangani kasus yang tidak memiliki nilai untuk sebuah atau lebih atribut. Akan tetapi, atribut kelas hanya bertipe diskret dan tidak boleh kosong.

Tiga prinsip kerja algoritma C4.5 pada tahap belajar dari data adalah :

1. Pembuatan pohon keputusan.

   Obyektif dari algoritma pohon keputusan adalah mengkonstruksi struktur data pohon (dinamakan pohon keputusan) yang dapat digunakan untuk memprediksi kelas dari sebuah kasus atau rekord baru yang belum memiliki kelas. C4.5 mengkonstruksi pohon keputusan dengan strategi *divide* dan *conquer*. Pada awalnya, hanya dibuat node akar dengan menerapkan algoritma *divide* dan *conquer*. Algoritma ini memilih pemecahan kasus-kasus yang terbaik dengan menghitung dan membandingkan *gain ratio*, kemudian pada node-node yang terbentuk di level berikutnya, algoritma *divide* dan *conquer* akan diterapkan lagi. Demikian seterusnya sampai terbentuk daun-daun. Algoritma C4.5 dapat menghasilkan pohon keputusan, dengan simbol kotak menyatakan simpul dan elips menyatakan daun.
2. Pemangkasan pohon keputusan dan evaluasi (opsional).

   Karena pohon yang dikonstruksi dapat berukuran besar dan tidak mudah "dibaca", C4.5 dapat menyederhanakan pohon dengan melakukan pemangkasan berdasarkan nilai tingkat kepercayaan (*confidence level*). Selain untuk pengurangan ukuran pohon, pemangkasan juga bertujuan untuk mengurangi tingkat kesalahan prediksi pada kasus (*record*) baru.
3. Pembuatan aturan-aturan dari pohon keputusan (opsional).

   Aturan-aturan dalam bentuk *if-then* diturunkan dari pohon keputusan dengan melakukan penelusuran dari akar sampai ke daun. Setiap *node* dan syarat pencabangannya akan diberikan di *if*, sedangkan nilai pada daun akan menjadi ditulis di *then*. Setelah semua aturan dibuat, maka aturan akan disederhanakan (digabung atau diperumum).

Jika aturan-aturan dari pohon tidak dibuat, maka klasifikasi kasus baru dapat dilakukan dengan menggunakan pohon keputusan.

Komputasi *Gain Ratio* pada Konstruksi Pohon C4.5

Pada konstruksi pohon C4.5, di setiap simpul pohon, atribut dengan nilai *gain ratio* yang tertinggi dipilih sebagai atribut test atau split untuk simpul. Rumus dari *gain ratio* adalah : *gain ratio(a) = gain(a) / split info(a)*

dimana *gain(a)* adalah *information gain* dari atribut *a* untuk himpunan sampel *X* dan *split info(a)* menyatakan entropi atau informasi potensial yang didapat pada pembagian *X* menjadi *n* sub himpunan berdasarkan telaahan pada atribut *a*. Sedangkan *gain(a)* didefinisikan sebagai :

$$gain\ (a) = info(X) - info_a(X)$$

Sedangkan rumus *split info(a)* adalah

$$split\ info(a) = -\sum_{i=1}^{n} \frac{|X_i|}{|X|} \times \log_2\left(\frac{|X_i|}{|X|}\right)$$

dimana *Xi* menyatakan sub himpunan ke-*i* pada sampel *X*.

Alasan penggunaan *gain ratio(a)* pada C4.5 (bukan *gain(a)*) sebagai kriteria pada atribut yang memiliki banyak nilai unik. Pembagian *gain(a)* dengan *split info(a)* dimaksudkan untuk sampel *X1…Xn*, dimana *n* adalah jumlah nilai unik pada atribut dan *Xi* adalah sub sampel yang memiliki nilai atribut *a = i*.

Untuk menghitung nilai *info a(X)*, jika *a* adalah atribut diskret, maka sampel *X* dibagi menjadi sub sampel *X1…Xn*, dimana *n* adalah jumlah nilai unik pada atribut *a* dan *Xi* adalah sub sampel yang memiliki nilai atribut *a = i*.

Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut :

Pilih atribut sebagai akar

1. Buat cabang untuk tiap-tiap nilai
2. Bagi kasus dalam cabang
3. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama

Untuk memilih atribut sebagai akar, didasarkan pada nilai gain tertinggi dari atribut-atribut yang ada. Untuk menghitung gain digunakan rumus seperti tertera dalam persamaan berikut :

$$Gain(S, A) = Enropy(s) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.1)$$

Di mana :

S : himpunan kasus

A : atribut

n : jumlah partisi atribut A

|Si| : jumlah kasus pada partisi ke-i

|S| : jumlah kasus dalam S

Sementara itu, perhitungan nilai entropi dapat dilihat pada persamaan berikut :

$$Entropy\ (S) = \sum_{i=1}^{n} \frac{|Si|}{|S|} - pi * log_2\ pi \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.2)$$

Di mana :

S : himpunan kasus

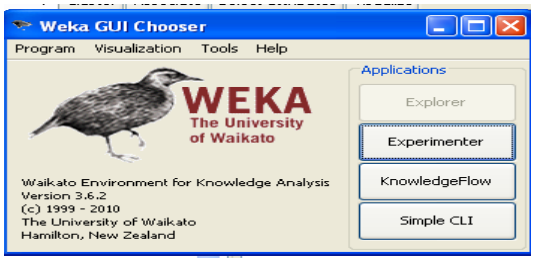
A : fitur

n : jumlah partisi S

*p*i : proporsi dari Si terhadap S

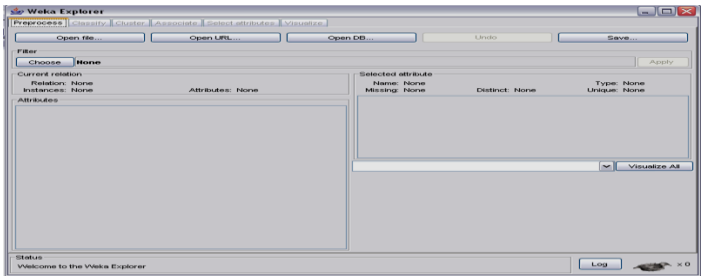
WEKA GUI Chooser

WEKA GUI Chooser adalah tampilan utama yang akan dilihat *user* pada saat pertama kali membuka perangkat lunak *WEKA*. Tampilan utama tersebut memberikan 4 pilihan *GUI WEKA*, yaitu *Simple CLI, Experimenter, Explorer,* dan *Knowledge Flow*, seperti pada gambar 2 berikut.



Gambar 2. WEKA GUI Chooser

GUI Explorer adalah *GUI WEKA* yang paling mudah digunakan dan menyediakan semua fitur *WEKA* dalam bentuk tombol dan tampilan visualisasi yang menarik dan lengkap. *Preprocess*, klasifikasi, asosiasi, *clustering*, pemilihan atribut, dan visualisasi dapat dilakukan dengan mudah dan menyenangkan di sini, seperti pada gambar 3 berikut ini.



Gambar 3. WEKA GUI Explorer

Format Data dalam WEKA

Misalnya diketahui sekumpulan data dan ingin dibangun sebuah *decision tree* dari data tersebut, maka data tersebut harus disimpan dalam *format 'flat', ARFF* karena *WEKA* perlu mengetahui beberapa informasi tentang tiap atribut yang tidak dapat disimpulkan secara otomatis dari nilai-nilainya.

File ARFF (*Attribute-Relation File Format*) adalah sebuah file teks *ASCII* yang berisi daftar *instances* dalam sekumpulan atribut. *File ARFF* dikembangkan oleh *Machine Learning Project* di *Department of Computer Science of The University of Waikato* untuk digunakan dalam perangkat lunak *WEKA*.

Pengubahan format data ini dapat dilakukan dengan mudah. Misalkan data dalam format .xls (lihat tabel 5), buka data tersebut dari Microsoft Excel dan simpan sebagai *.csv*. Selanjutnya, buka file tersebut dari *Microsoft Word*, notepad, atau editor teks lainnya dan data sudah berubah dalam format *comma-separated*. Hasilnya, data tersebut sudah dapat digunakan sebagai inputan dalam WEKA.

Pastikan bahwa data dalam format .arff tersebut sudah memenuhi:

- Data dipisahkan dengan koma, dengan kelas sebagai atribut terakhir.
- Bagian *header* diawali dengan @RELATION.
- Tiap atribut ditandai dengan @ATTRIBUTE. Tipe-tipe data dalam WEKA: numeric (REAL atau INTEGER), nominal, String, dan *Date*.
- Bagian data diawali dengan @DATA

PEMBAHASAN

Data kelulusan calon mahasiswa baru pada Perguruan Tinggi Swasta AMIK Depati Parbo Kerinci tahun ajaran 2009/2010 memiliki *form*at seperti

berikut : No. Ujian, Nama, Tempat/ Tanggal Lahir, Jenis Kelamin, Alamat, Nilai Ujian Tertulis, Keterangan Kelulusan.

Dari data-data tersebut, yang diambil sebagai variabel keputusannya adalah nilai kelulusan ya, dan tidak. Sedangkan yang diambil sebagai variabel penentu dalam pembentukan pohon keputusan adalah Nilai Ujian Tertulis, Nilai Wawancara, Nilai Ujian Praktek dan Rata-rata NEM.

Pemilihan variabel-variabel tersebut dengan pertimbangan bahwa jumlah nilai variabelnya tidak banyak sehingga diharapkan kelulusan mahasiswa yang masuk dalam satu klasifikasi nilai variabel tersebut cukup banyak. Terdapat sampel 40 orang mahasiswa yang mengikuti seleksi calon mahasiswa baru (SPMB) dengan memperhatikan parameter/atribut penilaian. Berarti dengan jumlah data 40, maka akan mendapatkan 6 kelas, data tersebut akan dikelompokkan berdasarkan atribut sebagai berikut.

1. Mengelompokkan Nilai Ujian Tertulis, pengelompokkan nilai ujian tertulis ini berdasarkan dari hasil ujian yang didapat oleh calon mahasiswa baru tersebut sehingga nilai tersebut dikelompokkan seperti terlihat pada tabel berikut ini :

Tabel 1. Klasifikasi Nilai Ujian Tertulis

Nilai Ujian Tertulis	Klasifikasi
0 – 15	1
16 – 32	2
33 – 47	3
48 – 63	4
64 – 79	5
>80	6

2. Mengelompokkan Nilai Wawancara berdasarkan keikutsertaan ataupun kehadiran pada pelaksanaan tes wawancara, seperti terlihat pada tabel berikut ini ;

Tabel 2. Klasifikasi Wawancara

Wawancara	Klasifikasi
Hadir	Y
Tidak Hadir	T

3. Mengelompokkan Nilai Praktek, berdasarkan hasil yang didapat pada pelaksanaan ujian praktek, seperti terlihat pada tabel berikut ini :

Tabel 3. Nilai Praktek

Nilai Ujian Praktek	Klasifikasi
0 – 15	1
16 – 32	2
33 – 47	3
48 – 63	4
64 – 79	5
>80	6

4. Mengelompokkan Nilai Rata-rata NEM, pengelompokan ini diambil dari nilai NEM yang diraih oleh calon mahasiswa baru tersebut pada waktu tamat SMA atau sederajat, pengelompokan nilai Rata-rata NEM tersebut dapat dilihat pada tabel berikut :

Tabel 4. Nilai Rata-rata NEM

Nilai Rata-rata NEM	Klasifikasi
0,00 – 1,59	1
1,60 – 3,29	2
3,30 – 4,79	3
4,80 – 6,39	4
6,40 – 7,90	5
>8,00	6

*Form*at data akhir setelah dilakukan pra-proses tampak seperti tabel berikut :

Tabel 5. *Form*at Data Akhir

No.	No. Ujian	Nilai Tertulis	wawancara	Nilai Praktek	Rata-rata NEM	Kelulusan
1	70001	6	Y	6	5	YA
2	70002	6	Y	5	6	YA
3	70003	6	Y	6	5	YA
4	70004	5	T	5	6	TIDAK
5	70005	6	T	5	5	TIDAK
6	70006	6	Y	5	5	YA
7	70007	5	Y	5	4	TIDAK
8	70008	5	Y	6	4	YA
9	70019	6	Y	6	5	YA
10	70010	6	Y	5	5	YA
11	70011	5	Y	5	6	YA
12	70012	6	Y	5	5	YA
13	70013	5	T	5	4	TIDAK
14	70014	6	Y	5	5	YA
15	70015	5	Y	6	4	YA
16	70016	5	Y	5	5	TIDAK

17	70017	6	Y	6	6	YA
18	70018	5	Y	5	5	TIDAK
19	70019	6	Y	5	5	YA
20	70020	6	Y	6	6	YA
21	70021	5	T	5	5	TIDAK
22	70022	6	Y	6	5	YA
23	70023	6	Y	6	6	YA
24	70024	5	T	6	5	YA
25	70025	6	Y	5	6	YA
26	70026	5	Y	5	4	YA
27	70027	6	Y	6	5	YA
28	70028	5	Y	5	4	TIDAK
29	70029	5	T	5	5	TIDAK
30	70030	5	Y	6	4	YA
31	70031	6	Y	6	6	YA
32	70032	6	Y	5	4	TIDAK
33	70033	6	Y	5	6	YA
34	70034	6	Y	5	5	YA
35	70035	6	Y	5	4	TIDAK
36	70036	6	Y	6	4	YA
37	70037	5	Y	5	6	TIDAK
38	70038	5	Y	5	6	TIDAK
39	70039	6	Y	5	4	TIDAK
40	70040	6	Y	6	4	YA

Format data akhir pada tabel di atas didapat berdasarkan dari atribut yang sudah dikelompokan atau diklasifikasi, misalkan data pada tabel 4.1 Nilai Ujian Tertulis adalah “81”, setelah diklasifikasi menjadi “6”, Wawancara adalah “hadir”, maka setelah klasifikasi menjadi “Y”, Nilai Praktek adalah “80” berubah menjadi “6”, dan Rata-rata NEM adalah “6.32” berubah menjadi “5” dan seterusnya.

Pohon Keputusan

Dari *form*at data akhir kelulusan calon mahasiswa baru maka akan dilakukan klasifikasi data algoritma C4.5 dengan membuat pohon keputusan. Seperti yang telah dijelaskan sebelumnya, algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut :

1. Pilih atribut sebagai akar.
2. Buat cabang untuk tiap-tiap nilai.
3. Bagi kasus dalam cabang.
4. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Dalam kasus yang tertera pada tabel 4.6 di atas, akan dibuat pohon keputusan untuk menentukan klasifikasi kelulusan calon mahasiswa baru (ya dan tidak) dengan melihat Nilai Ujian Tertulis, Nilai Wawancara, Nilai Ujian Praktek dan Nilai Rata-rata NEM.

Untuk memilih atribut sebagai akar, didasarkan pada nilai *gain* tertinggi dari atribut-atribut yang ada. Untuk menghitung *gain* digunakan rumus (2.1), sedangkan untuk menghitung nilai *entrop*y dapat dilihat pada rumus (2.2).

Dengan menggunakan dua persamaan di atas maka akan didapatkan *entropy* dan *gain* yang digunakan sebagai akar dalam membuat pohon keputusan.

- Menghitung jumlah kasus, jumlah kasus untuk keputusan "Ya", jumlah kasus untuk keputusan "Tidak", dan kasus yang dibagi berdasarkan atribut Nilai Ujian Tertulis, Nilai Wawancara, Nilai Ujian Praktek, Nilai Rata-rata NEM. Setelah itu, lakukan perhitungan *gain* untuk setiap atribut.

Menghitung Nilai Entropy tiap-tiap atribut :

***Entropy* (Total)**

$$Entrupy\ TotalA = \left(-\frac{26}{40} * log_2\left(\frac{26}{40}\right)\right) + \left(-\frac{14}{40} * log_2\left(\frac{14}{40}\right)\right) = 0{,}934068$$

Entropy(total) adalah menghitung nilai total keputusan ya (26) dan tidak (14), sedangkan 40 adalah jumlah keseluruhan kasus.

Atribut Nilai Rata-rata NEM

$$Entropy(1) = \left(-\frac{0}{0} * log_2\left(\frac{0}{0}\right)\right) + \left(-\frac{0}{0} * log_2\left(\frac{0}{0}\right)\right) = 0$$

$$Entropy(2) = \left(-\frac{0}{0} * log_2\left(\frac{0}{0}\right)\right) + \left(-\frac{0}{0} * log_2\left(\frac{0}{0}\right)\right) = 0$$

$$Entropy(3) = \left(-\frac{0}{0} * log_2\left(\frac{0}{0}\right)\right) + \left(-\frac{0}{0} * log_2\left(\frac{0}{0}\right)\right) = 0 Entrupy\ (4)$$

$$= \left(-\frac{6}{12} * log_2\left(\frac{6}{12}\right)\right) + \left(-\frac{6}{12} * log_2\left(\frac{6}{12}\right)\right) = 1$$

$$Entropy(5) = \left(-\frac{12}{17} * log_2\left(\frac{12}{17}\right)\right) + \left(-\frac{5}{17} * log_2\left(\frac{5}{17}\right)\right) = 0{,}873981$$

$$Entropy(6) = \left(-\frac{8}{11} * log_2\left(\frac{8}{11}\right)\right) + \left(-\frac{3}{11} * log_2\left(\frac{3}{11}\right)\right) = 0{,}845351$$

Menghitung nilai *entrophy* atribut Nilai Rata-rata NEM berdasarkan dari tiap-tiap kelas (1, 2, 3, 4, 5, 6) pada Atribut Nilai Rata-rata NEM.

a. Atribut Nilai Ujian Tertulis

$$Entropy(1) = \left(-\frac{0}{0} * log_2\left(\frac{0}{0}\right)\right) + \left(-\frac{0}{0} * log_2\left(\frac{0}{0}\right)\right) = 0$$

$$Entropy(2) = \left(-\frac{0}{0} * log_2\left(\frac{0}{0}\right)\right) + \left(-\frac{0}{0} * log_2\left(\frac{0}{0}\right)\right) = 0$$

$$Entropy(3) = \left(-\frac{0}{0} * log_2\left(\frac{0}{0}\right)\right) + \left(-\frac{0}{0} * log_2\left(\frac{0}{0}\right)\right) = 0$$

$$Entropy(4) = \left(-\frac{0}{0} * log_2\left(\frac{0}{0}\right)\right) + \left(-\frac{0}{0} * log_2\left(\frac{0}{0}\right)\right) = 0$$

$$Entropy(5) = \left(-\frac{6}{16} * log_2\left(\frac{6}{16}\right)\right) + \left(-\frac{10}{16} * log_2\left(\frac{10}{16}\right)\right) = 0{,}954434$$

$$Entropy(6) = \left(-\frac{20}{24} * log_2\left(\frac{20}{24}\right)\right) + \left(-\frac{4}{24} * log_2\left(\frac{4}{24}\right)\right) = 0{,}650022$$

Menghitung nilai *entrophy* Atribut Nilai Ujian Tertulis berdasarkan dari tiap-tiap kelas (1, 2, 3, 4, 5, 6) pada Atribut Nilai Ujian Tertulis.

b. Atribut Nilai Ujian Praktek

$$Entropy(1) = \left(-\frac{0}{0} * log_2\left(\frac{0}{0}\right)\right) + \left(-\frac{0}{0} * log_2\left(\frac{0}{0}\right)\right) = 0$$

$$Entropy(2) = \left(-\frac{0}{0} * log_2\left(\frac{0}{0}\right)\right) + \left(-\frac{0}{0} * log_2\left(\frac{0}{0}\right)\right) = 0$$

$$Entropy(3) = \left(-\frac{0}{0} * log_2\left(\frac{0}{0}\right)\right) + \left(-\frac{0}{0} * log_2\left(\frac{0}{0}\right)\right) = 0$$

$$Entropy(4) = \left(-\frac{0}{0} * log_2\left(\frac{0}{0}\right)\right) + \left(-\frac{0}{0} * log_2\left(\frac{0}{0}\right)\right) = 0$$

$$Entropy(5) = \left(-\frac{11}{25} * log_2\left(\frac{11}{25}\right)\right) + \left(-\frac{14}{25} * log_2\left(\frac{14}{25}\right)\right) = 0{,}989588$$

$$Entropy(6) = \left(-\frac{15}{15} * log_2\left(\frac{15}{15}\right)\right) + \left(-\frac{0}{15} * log_2\left(\frac{0}{15}\right)\right) = 0$$

Menghitung nilai *entrophy* Nilai Ujian Praktek berdasarkan dari tiap-tiap kelas (1, 2, 3, 4, 5, 6) pada Atribut Nilai Ujian Praktek.

c. Atribut Wawancara

$$Entropy(Y) = \left(-\frac{25}{34} * \ log_2\left(\frac{25}{34}\right)\right) + \left(-\frac{9}{34} * \ log_2\left(\frac{9}{34}\right)\right)$$
$$= 0{,}833765$$
$$Entropy(T) = \left(-\frac{1}{6} * \ log_2\left(\frac{1}{6}\right)\right) + \left(-\frac{5}{6} * \ log_2\left(\frac{5}{6}\right)\right) = 0{,}650022$$

Menghitung nilai *entrophy* Atribut Wawancara berdasarkan dari tiap-tiap kelas (Y,T) pada Atribut Wawancara.

Menghitung Nilai Gain tiap-tiap atribut :

Gain (Total, Nilai Rata-rata NEM)

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|Rata\ NEM_i|}{|Total|} * Entropy(Rata\ NEM_i)$$
$$= 0.934068 - \left(\left(\frac{12}{40} * 1\right) + \left(\frac{17}{40} * 0.873981\right) + \left(\frac{11}{40} * 0.845351\right)\right)$$
$$= 0.030155$$

Gain (Total, Nilai Tertulis)

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|Nilai\ Tertulis_i|}{|Total|} * Entropy(Nilai\ Tertulis_i)$$
$$= 0.934068 - \left(\left(\frac{16}{40} * 0.954434\right) + \left(\frac{24}{40} * 0.650022\right)\right) = 0{,}162281$$

Gain(Total, Nilai Ujian Praktek)

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|Nilai\ Praktek_i|}{|Total|} * Entropy(Nilai\ Praktek_i)$$
$$= 0.934068 - \left(\left(\frac{25}{40} * 0.989588\right) + \left(\frac{55}{40} * 0\right)\right) = 0.315576$$

Gain(Total, Nilai Wawancara)

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|Nilai\ Wawancara_i|}{|Total|} * Entropy(Nilai\ Wawancara_i)$$
$$= 0.934068 - \left(\left(\frac{34}{40} * 0.833765\right) + \left(\frac{6}{40} * 0.650022\right)\right) = 0{,}1227864$$

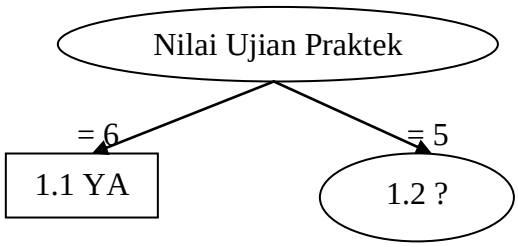
Setelah nilai *entropy* dan *gain* dihitung, kemudian hasil dari perhitungan tersebut dimasukkan ke dalam tabel Perhitungan Node 1 berikut ini.

Tabel 6. Perhitungan Node 1

Node			Jumlah Kasus (S)	Ya (S1)	Tidak (S2)	Entropy	Gain
1	TOTAL		40	26	14	0.934068	
	Nilai Rata-rata NEM						0.030155
		1	0	0	0		
		2	0	0	0		
		3	0	0	0		
		4	12	6	6	1	
		5	17	12	5	0.873981	
		6	11	8	3	0.845351	
	Nilai Ujian Tertulis						0.162281
		1	0	0	0		
		2	0	0	0		
		3	0	0	0		
		4	0	0	0		
		5	16	6	10	0.954434	
		6	24	20	4	0.650022	
	Nilai Ujian Praktek						0.315576
		1	0	0	0		
		2	0	0	0		
		3	0	0	0		
		4	0	0	0		
		5	25	11	14	0.989588	
		6	15	15	0	0	
	Nilai Wawancara						0.127864
		YT	34 6	25 1	9 5	0.833765 0.650022	

Dari perhitungan pada tabel diatas dapat diketahui bahwa atribut dengan *gain* tertinggi adalah Nilai Ujian Praktek sebesar 0.315576. Berarti Nilai Ujian Praktek dapat menjadi node akar. Ada dua nilai atribut dari Nilai Ujian Praktek yaitu “5” dan “6”. Dari nilai atribut tersebut, nilai “6” sudah Lulus, sehingga tidak

perlu dilakukan perhitungan, tetapi nilai atribut "5" masih perlu dilakukan perhitungan lagi, seperti pada gambar dibawah ini.



Gambar 4. Pohon Keputusan Hasil Perhitungan Node 1

- Selanjutnya adalah menyelesaikan untuk menghitung Node 1.2 sebagai akar, sama dengan cara yang diatas dengan menghitung nilai *entropy* dari atribut yang tersisa yaitu Nilai Rata-rata NEM, Nilai Ujian Tertulis dan Nilai Wawancara, setelah di hitung *entropy*, kemudian menghitung *gain* untuk tiap-tiap atribut.

Setelah nilai *entropy* dan *gain* dihitung, kemudian hasil dari perhitungan tersebut dimasukkan ke dalam tabel Perhitungan Node 1.2 dibawah ini.

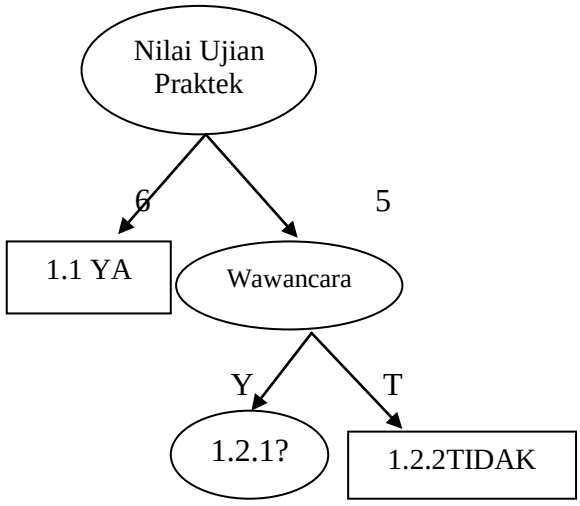
Tabel 7. Perhitungan Node 1.2

Node			Jumlah Kasus (S)	Ya (S1)	Tidak (S2)	Entropy	Gain
1.2	Nilai Ujian Praktek-5		25	11	14	0.989588	
	Nilai Rata-rata NEM						0.110683
		1	0	0	0		
		2	0	0	0		
		3	0	0	0		
		4	7	1	6	0.591673	
		5	11	6	5	0.99403	
		6	7	4	3	0.985228	
	Nilai Ujian Tertulis						0.143544
		1	0	0	0		
		2	0	0	0		
		3	0	0	0		
		4	0	0	0		

	5	13	3	10	0.77935	
	6	12	8	4	0.91829	
					6	
Wawancara						0.1953
						69
	Y	20	11	9	0.99277	
					4	
	T	5	0	5	0	

Dari hasil tabel diatas dapat diketahui bahwa atribut *gain* tertinggi adalah Nilai Wawancara, yaitu sebesar 0.195369, berarti Wawancara dapat menjadi node akar. Ada dua nilai atribut dari Wawancara yaitu "Y" dan "T". Dari nilai atribut tersebut, nilai "T" sudah Tidak Lulus, sehingga tidak perlu dilakukan perhitungan, tetapi nilai atribut "Y" masih perlu dilakukan perhitungan lagi.

Pohon keputusan yang terbentuk dari perhitungan Node 1.2 adalah seperti pada gambar 5 dibawah ini.



Gambar 5. Pohon Keputusan Hasil Perhitungan Node 1.2

- Selanjutnya adalah menyelesaikan untuk menghitung Node 1.2.1 sebagai akar, sama dengan cara yang diatas dengan menghitung nilai *entropy* dari atribut yang tersisa yaitu Nilai Rata-rata NEM dan Nilai Ujian Tertulis, setelah di hitung *entropy*, kemudian menghitung *gain* untuk tiap-tiap atribut.

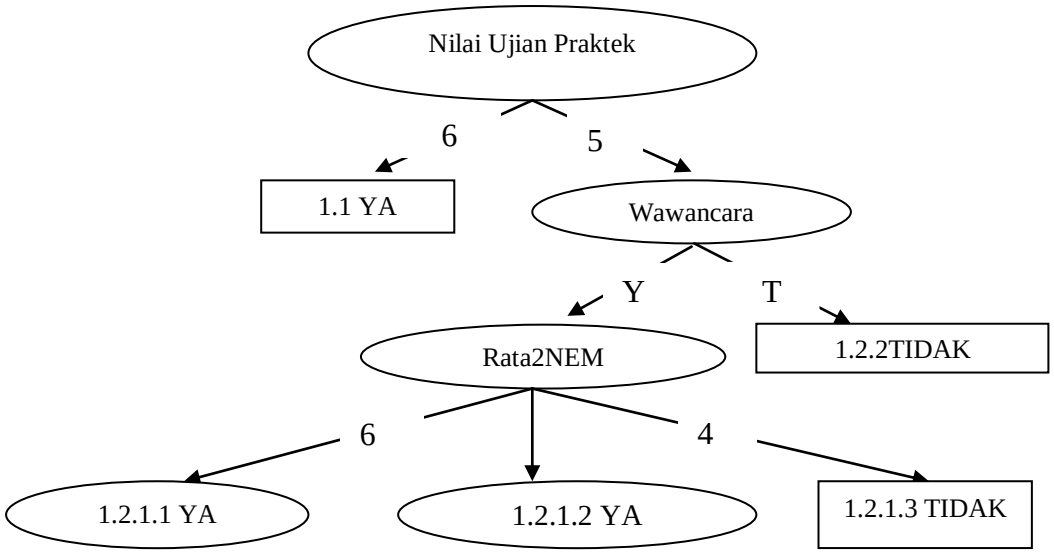
Setelah nilai *entropy* dan *gain* dihitung, kemudian hasil dari perhitungan tersebut dimasukkan ke dalam tabel Perhitungan Node 1.2.1 dibawah ini.

Tabel 8. Perhitungan Node 1.2.1

Node			Jumlah Kasus (S)	Ya (S1)	Tidak (S2)	Entropy	Gain
1.2.1	Wawancara – Y		20	11	9	0.992774	
	Nilai Rata-rata NEM						0.197737
		1	0	0	0		
		2	0	0	0		
		3	0	0	0		
		4	6	1	5		
		5	8	6	2		
		6	6	4	3		
	Nilai Ujian Tertulis						0.117255
		1	0	0	0		
		2	0	0	0		
		3	0	0	0		
		4	0	0	0		
		5	8	6	2		
		6	12	8	4		

Dari hasil tabel Perhitungan Node 1.2.1 diatas dapat diketahui bahwa atribut *gain* tertinggi adalah Rata-rata NEM, yaitu sebesar 0.197737, berarti Rata-rata NEM dapat menjadi node akar selanjutnya. Ada tiga nilai atribut dari Nilai Rata-rata NEM yaitu “4”, “5” dan “6”. Dari nilai atribut tersebut, nilai “4” sudah Tidak Lulus, sehingga tidak perlu dilakukan perhitungan, tetapi nilai atribut “5” dan “6” masih perlu dilakukan perhitungan lagi.

Pohon keputusan yang terbentuk pada saat ini adalah seperti terlihat pada gambar dibawah ini.



Gambar 6. Pohon Keputusan Hasil Perhitungan Node 1.2.1

Berdasarkan pohon keputusan terakhir yang terbentuk pada gambar di atas, maka aturan atau *rule* yang terbentuk adalah sebagai berikut :

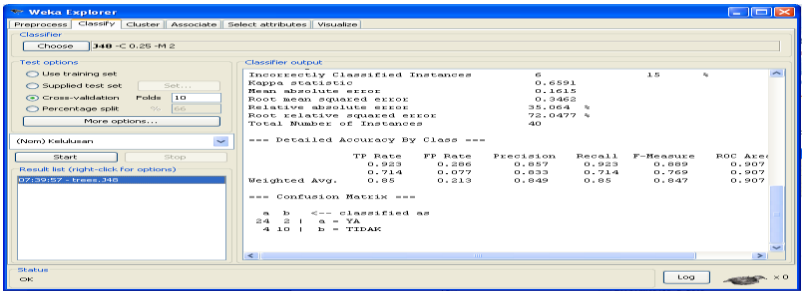
1. Jika Nilai Ujian Praktek = 6, maka Kelulusan = YA
2. Jika Nilai Ujian Praktek = 5 dan Wawancara = T, Maka Kelulusan = TIDAK
3. Jika Nilai Ujian Praktek = 5 dan Wawancara = Y dan Rata-rata NEM = 6 dan Nilai Tertulis = 6, maka kelulusan = YA
4. Jika Nilai Ujian Praktek = 5 dan Wawancara = Y dan Rata-rata NEM = 6, dan Nilai Tertulis = 5, maka kelulusan = TIDAK
5. Jika Nilai Ujian Praktek = 5 dan Wawancara = Y dan Rata-rata NEM = 5, dan Nilai Tertulis = 6, maka kelulusan = YA
6. Jika Nilai Ujian Praktek = 5 dan Wawancara = Y dan Rata-rata NEM = 5 dan Nilai Tertulis = 5, maka kelulusan = TIDAK
7. Jika Nilai Ujian Praktek = 5 dan Wawancara = Y dan Rata-rata NEM = 4, maka kelulusan = TIDAK

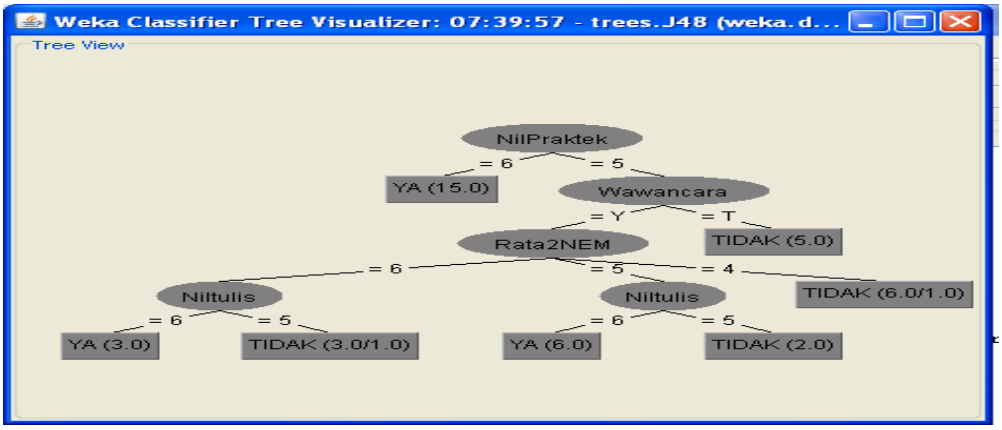
Berdasarkan dari *rule/knowledge* yang dihasilkan terdapat beberapa *rule* yang cukup sesuai dengan kejadian yang terjadi didalam menentukan kelulusan calon mahasiswa baru, dimana mahasiswa yang memiliki nilai yang tinggi akan lulus dalam seleksi calon mahasiswa baru.

Implementasi

Untuk menguji kebenaran dari hasil pengolahan data yang dikerjakan secara manual pada Pembahasan diatas, kita dapat menggunakan salah satu *software* aplikasi *WEKA Knowledge Explorer* dengan langkah-langkah sebagai berikut :

1. Seluruh variabel-variabel (terdiri dari atribut kondisi dan atribut keputusan) yang digunakan untuk menentukan kelulusan calon mahasiswa baru disimpan pada *microsoft excel* dengan nama file datatesis.xls (yang berisi kasus atau kriteria dalam mengasilkan *rule*) seperti yang terlihat pada Gambar 5.1.
2. file datatesis.xls kemudian simpan sebagai .csv. Selanjutnya, buka file tersebut dari *Microsoft Word*, *notepad*, atau editor teks lainnya dan data sudah berubah dalam format *comma-separated*. Lalu sesuaikan data tersebut dengan menambahkan informasi awal hasilnya, data tersebut sudah dapat digunakan sebagai inputan dalam *WEKA*.
3. Agar data tersebut dapat digunakan pada aplikasi *WEKA decision tree*, maka data tersebut harus disimpan dalam format 'ARFF' (Gambar 5.2). File ARFF (*Attribute-Relation File Format*) adalah sebuah file teks ASCII yang berisi daftar *instances* dalam sekumpulan atribut. File ARFF dikembangkan oleh *Machine Learning Project* di *Department of Computer Science of The University of Waikato* untuk digunakan dalam perangkat lunak *WEKA*. Data dalam format .arff tersebut dapat dipenuhi dengan cara:
    - Data dipisahkan dengan koma, dengan kelas sebagai atribut terakhir.
    - Bagian *header* diawali dengan @RELATION.
    - Tiap atribut ditandai dengan @ATTRIBUTE. Tipe-tipe data dalam *WEKA*: numerik(REAL atau INTEGER), nominal, String, dan *Date*.
    - Bagian data diawali dengan @DATA





Gambar 8. Hasil *Visualisasi Tree*

Kesimpulan

Berdasarkan pembahasan diatas dapat diambil kesimpulan sebagai berikut :

1. Sistem pengklasifikasian kelulusan calon mahasiswa baru menggunakan algoritma C4.5 dapat digunakan dalam pengambilan keputusan untuk mencari alternatif yang baik.
2. Seorang peserta yang ikut seleksi calon mahasiswa baru dinyatakan diterima atau tidak tergantung pada pihak universitas berdasarkan pertimbangan beberapa kriteria yang ditetapkan.

DAFTAR PUSTAKA

Iko Pramudiono. (2003). "*Jurnal Pengantar Data Mining: Menambang Permata Pengetahuan di Gunung Data*"

Kusrini dan Luthfi Taufiq Emha. (2009). " *Algoritma Data Mining*". Yogyakarta : Andi

Kusrini dan Sri Hartati. (2007). "*Seminar Nasional Teknologi 2007 (SNT 2007}, Penggunaan Pohon Keputusan untuk Menanalisis Kemungkinan Pengunduran Diri Calon Mahasiswa di STMIK Amikom Yogyakart"*

Riduwan. (2003). *Dasar-dasar Statistika".Bandung .Alfabeta Mahasiswa dengan Metode Klasifikasi Decision Tree"*

Sani Susanto dan Dedy Suryadi. (2010). *"Pengantar Data Mining Menggali Pengetahuan dari Bongkahan Data".* Yogyakarta. Andi

Veronica Sri Moertini. (2007). "*Disertasi Pengembangan Skalabilitas Algoritma Klasifikasi C4.5 dengan Pendekatan Konsep Operator Relasi (Studi Kasus:Pra-Pengolahan dan Klasifikasi Citra Batik)*"

Veronica S. Moertini. "*Jurnal Penanganan Atribut Citra dengan Wavelet untuk Pengembangan Aloritma C4.5*"

Yudho Giri Sucahyo. (2003). "*Jurnal Data Mining Menggali Informasi yang Terpendam"*