Jurnal SimanteC

**Vol. 5, No. 2 Juni 2016** **ISSN 2088-2130;e-ISSN 2502-4884**

# APLIKASI SISTEM PENDUKUNG KEPUTUSAN MENGGUNKAAN ALGORITMA C4.5. UNTUK PENJURUSAN SMA

**Yeni Kustiyahningsih[1)], Eza Rahmanita[2)]**
[1]Program Studi Manajemen Informatika, Universitas Trunojoyo
[2]Program Studi Teknik Informatika, Universitas Trunojoyo
Jl. Raya Telang PO.BOX 2 Kamal Bangkalan
E-mail: ykustiyahninsih@yahoo.com , rahmanita@gmail.com

## ABSTRAK

Algoritma C4.5 merupakan pengembangan dari algoritma ID3, kelebihan algoritma C4.5 dapat menangani atribut kontinyu dan diskrit, kemudian dapat menangani *training* data dengan *missing value*, serta menggunakan *gain ratio* untuk memperbaiki *information gain*. Selama ini untuk menentukan penjurusan sekolah SMA masih dilakukan dengan cara manual. Semakin tahun terjadi peningkatan jumlah siswa dan syarat untuk penjurusan juga semakin komplek, sehingga diperlukan sistem aplikasi penjurusan untuk membantu pihak sekolah dalam mempercepat dan efisiensi penjurusan sekolah. Apabila penjurusan sesuai dengan kemampuan dan minat siswa, maka mereka dapat belajar dengan nyaman dan lulusan yang dihasilkan juga mendapat nilai yang maksimal, sehingga rata-rata nilai meningkat. Tujuan penelitian ini adalah membuat klasifikasi penjurusan siswa menggunakan algoritma C4.5 (metode decision tree) untuk mempermudah dan mempercepat penentuan penyeleksian pemilihan jurusan sehingga proses yang dihasilkan dari seleksi ini lebih akurat dan objektif. Adapun Kriteria penjurusan adalah nilai Matematika, Fisika, Biologi, Kimia, Nilai Psikotest (IQ), Saran Psikotest, Angket/Minat Siswa, Saran Bimbingan Konseling. Hasil dari klasifikasi algoritma akan di analisa untuk menentukan *recall, pressicion, accuracy* terbesar dan Nilai *error rate* terkecil yang dicapai. Dari skenario uji coba yang dilakukan nilai akurasi yang dihasilkan setelah dilakukan *pruning* lebih baik dari pada tanpa *pruning*.

**Kata Kunci** : Algoritma C4.5, Penentuan Jurusan, *decission tree*, akurasi, kriteria.

## ABSTRACT

*The C4.5 algorithm is the development of ID3 algorithm that the excess of C4.5 can handle continuous and discrete attributes and training data with the missing value, and uses the gain ratio to improve information gain. During this time, to determine the high school majors is still done manually. For more years the increasing number of students and the requirements for majors are increasingly complex, it is necessary to have an application system of majors to help school in accelerating the efficiency of the school majors. If the majors are in accordance with the student's abilities and interests, the result will be that the students can learn comfortably and can get maximum score, and thus the average value increases. The purpose of this research is to make a classification placement of students using the C4.5 algorithm (decision tree method) to facilitate and accelerate the determination of the screening department election so that the resulting process of this selection is more accurate and more objective. The majors' criteria for consideration are the score of Mathematics, Physics, Biology, Chemistry, Psychological Value (IQ), Psychological advice, Questionnaire / Students' Interests, and Counseling. The results of the algorithm classification will be analyzed to determine any recall, precision, greatest accuracy and value of the smallest error rate achieved. From the performed test-scenario, the accuracy values produced after pruning is better than without pruning.*

***Keywords**: Algorithm C 4.5, Determination Department, decision tree, accuracy, criteria.*

## PENDAHULUAN

Pendidikan merupakan pilar utama dalam kemajuan suatu bangsa. Suatu negara dikatakan maju apabila pendidikan negara tersebut berkembang pesat dan memadai. Pendidikan menjadi salah satu elemen penting meningkatkan kecerdasan bangsa. Sekolah adalah bagian dari pendidikan dimana terdapat proses belajar mengajar siswa dalam meningkatkan kecerdasan, pengembangan ilmu pengetahuan dan pengabdian kepada masyarakat. SMA merupakan jenjang pendidikan menengah pada pendidikan formal di Indonesia. Berdasarkan kebijakan Dirjen Mandikdasmen No. 12/C/Kep/TU/2008, 12 Februari 2008 tentang panduan penyusunan Laporan Hasil Belajar (LHB) Kurikulum Tingkat Satuan Pendidikan (KTSP) Bagian E. Butir 1 s.d. 5 yang substansinya memberikan rambu-rambu operasional tentang dasar-dasar penjurusan di SMA diantaranya menetapkan perlunya penjurusan pada siswa kelas X SMA yang akan naik ke kelas XI. Penjurusan akan disesuaikan dengan minat dan kemampuan siswa, dengan tujuan agar pembelajaran siswa menjadi lebih terarah. Penjurusan yang tersedia di SMA meliputi Ilmu Pengetahuan Alam (IPA), Ilmu Pengetahuan Sosial (IPS) dan bahasa.

Proses penjurusan di SMA Muhammadiyah Bangkalan dilakukan pada saat siswa berada di kelas X dan akan naik ke kelas XI. Setelah wali kelas menerima seluruh nilai semester maka wali kelas akan memutuskan apakah siswa tersebut naik atau tidak. Jika siswa tersebut dinyatakan naik maka selanjutnya akan dilakukan proses penjurusan oleh tim yang terdiri dari Wakil Kepala Sekolah Bidang Kurikulum, Guru Bimbingan Konseling, Wali Kelas X dan Guru Mata Pelajaran yang berkaitan dengan penjurusan. Masalah yang sering terjadi dalam proses penjurusan adalah keterlambatan nilai siswa dari para wali kelas, kriteria penjurusan juga semakin komplek, akibatnya proses penjurusan tidak efisien dan kurang tepat, ditambah lagi dengan banyaknya jumlah siswa kelas X. Tahun ajaran 2013/2014 tercatat siswa kelas X sejumlah ± 220 orang. Dari sini sangat dikhawatirkan terjadi kekeliruan dalam penyeleksian, sehingga menyebabkan kesalahan dalam pemilihan jurusan dan akhirnya tidak sesuai dengan minat dan kemmapuan siswa.

Tujuan dalam penelitian ini untuk mengatasi masalah tersebut, yaitu dengan melakukan proses *data mining*. Metode *data mining* yang digunakan adalah klasifikasi. Klasifikasi merupakan proses untuk menemukan model yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Pada penelitian ini menggunakan algoritma C4.5. Untuk aplikasi klasifikasi penjurusan siswa SMA. Algoritma C4.5. Lebih efektif, hasil ketepatan prediksi dan nilai kesalahan (*error rate*) lebih baik dari ID3[1]. Pengukuran kinerja yang dilakukan menggunakan sekelompok data uji untuk mengetahui *prosentase precision*, *recall* dan *accuracy*, menunjukkan bahwa algoritma C4.5 memiliki tingkat akurasi yang lebih tinggi dari pada algoritma ID3[2]. Oleh karena itu pada penelitian ini menggunakan algoritma C4.5 dalam melakukan klasifikasi penentuan jurusan SMA dengan tujuan membantu pihak sekolah dalam menentukan penjurusan SMA dengan dengan cepat dan akurat serta *rule* yang dihasilkan akan digunakan sebagai penentu keputusan, sehingga hasilnya dapat mengklasifikasikan penjurusan sekolah sesuai kemampuan minat dan bakat siswa.

## TINJAUAN PUSTAKA

### *Data Mining*

*Data mining* merupakan proses pencarian pola dan relasi-relasi yang

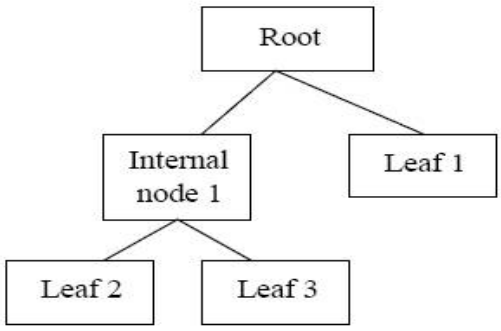
tersembunyi dalam sejumlah data yang besar dengan tujuan untuk melakukan klasifikasi, estimasi, prediksi, *association rule, clustering,* deskripsi dan visualisasi. *Data mining* adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual[3].

## Pohon Keputusan (*Decision Tree*)

Salah satu metode *data mining* yang umum digunakan adalah pohon keputusan. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan *rule*. Pohon keputusan adalah salah satu metode klasifikasi yang paling popular karena mudah untuk diinterpretasi oleh manusia. Konsep dari pohon keputusan adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan [3].


Gambar 1. Konsep pohon keputusan

Pohon keputusan merupakan himpunan aturan IF...THEN. Setiap *path* dalam *tree* dihubungkan dengan sebuah aturan, di mana premis terdiri atas sekumpulan *node-node* yang ditemui, dan kesimpulan dari aturam terdiri atas kelas yang terhubung dengan *leaf* dari *path* [4].




Gambar 2. Konsep Dasar Pohon Keputusan

## Pohon Keputusan C4.5

Algoritma *Classification version* 4.5 atau biasa disebut C4.5 adalah pengembangan dari algoritma ID3. Oleh karena pengembangan tersebut, algoritma C4.5 mempunyai prinsip dasar kerja yang sama dengan algoritma ID3. Perbedaan utama C4.5 dari ID3 adalah C4.5 dapat menangani atribut kontinyu dan diskrit, C4.5 dapat menangani *training data* dengan *missing value*, Hasil pohon keputusan C4.5 akan dipangkas setelah dibentuk, pemilihan atribut yang dilakukan dengan menggunakan *Gain Ratio. Information gain* pada ID3 lebih mengutamakan pengujian yang menghasilkan banyak keluaran. Dengan kata lain, atribut yang memiliki banyak nilailah yang dipilih sebagai *splitting* atribut. Sebagai contoh, pembagian terhadap atribut yang berfungsi sebagai *unique identifier*, seperti product_ID, akan menghasilkan keluaran dalam jumlah yang banyak, di mana setiap keluaran hanya terdiri dari satu *tuple*. Partisi semacam ini tentu saja bersifat *pure*, sehingga informasi yang dibutuhkan untuk mengklasifikasi D berdasarkan partisi seperti ini adalah sebesar Infoproduct_ID(D) = 0. Sebagai akibatnya, *information gain* yang dimiliki atribut product_ID menjadi maksimal. Padahal, jelas sekali terlihat bahwa partisi semacam ini tidaklah berguna. Karena itu algoritma C4.5 yang merupakan suksesor dari ID3 menggunakan *gain ratio* seperti pada Persamaan 1 untuk memperbaiki *information gain*, dengan rumus *gain ratio* [3] :

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInfo(S,A)} \quad (1)$$

Dimana:
S = Ruang (data) sample yang digunakan untuk training.
A = atribut. *Gain*(S,A) = *information gain* atribut A *SplitInfo*(S,A) = *split information* atribut A.

Atribut dengan nilai *Gain Ratio* tertinggi dipilih sebagai atribut test untuk simpul. Dengan *gain* adalah *information gain.* Pendekatan ini menerapkan normalisasi pada *information gain* dengan menggunakan apa yang disebut sebagai *split information*. *SplitInfo* menyatakan *entropy* atau informasi potensial dengan rumus untuk *SplitInfo* menggunakan Persamaan 2.

$$SplitInfo(S,A) = -\sum_{i=1}^{n} \frac{S_i}{S} \log_2 \frac{S_i}{S} \qquad (2)$$

Dimana:
S = Ruang (data) sample yang digunakan untuk training.
A = Attribut.
$S_i$ = Jumlah sample untuk atribut i

***Reduced Error Prunning (REP)***

*Reduced Error Pruning* merupakan salah satu algoritma *postpruning.* Algoritma ini membagi data menjadi dua, yaitu *training data* dan *test data. Training data* adalah data yang digunakan untuk membentuk pohon keputusan, sedangkan *test data* digunakan untuk menghitung nilai *error rate* pada pohon setelah dipangkas. Cara kerja REP adalah dengan memangkas *internal node* yang dimulai dari *internal node* paling bawah ke atas. Pemangkasan dilakukan dengan cara mengganti atribut dengan *leaf node* yang memiliki kelas yang dominan muncul. Setelah itu *test data* diproses menggunakan *rule* hasil pemangkasan, kemudian dihitung nilai *error rate*nya. *Test data* juga diproses dengan rule awal, yaitu rule yang terbentuk sebelum pohon dipangkas, kemudian dihitung nilai *error rate*nya. Apabila nilai *error rate* yang dihasilkan dari pemangkasan pohon lebih kecil, maka pemangkasan dilakukan.

***Pre Prunning***

*Prepruning* yaitu menghentikan pembangunan suatu *subtree* lebih awal, yaitu dengan memutuskan untuk tidak lebih jauh mempartisi *data training.* Rumus *prepruning* menggunakan persamaan 3 [3].

$$e = \frac{r + \frac{z^2}{2n} + z\sqrt{\frac{r}{n} - \frac{r^2}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \qquad (3)$$

Dimana:
r = nilai perbandingan *error rate*
n = total *sample*
$z = \Phi^{-1}(c)$
c = *confidence level*

## METODE

### Pendefinisian Masalah

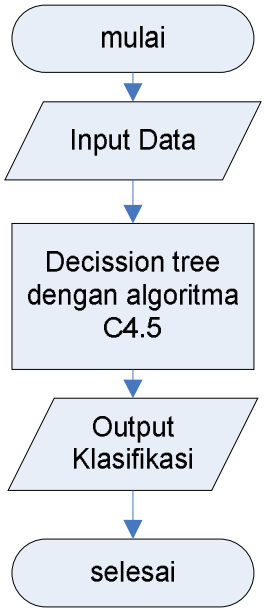
Tahap ini merupakan tahapan bagaimana suatu permasalahan dirumuskan berdasarkan latar belakang atau tahap pendahuluan dari penelitian.

1. Studi Literatur dan review jurnal.
   Dukungan teori dan bahan – bahan bacaan, jurnal atau paper mengenai rekayasa perangkat lunak, metode klasifikasi C 4.5., *decission support system* (DSS).
2. Survey, pengumpulan data dan informasi.
   Tahap ini bertujuan untuk mengetahui dan melihat secara langsung dan lebih mendetail permasalahan yang akan diteliti, sehingga diperoleh data–data atau informasi yang diperlukan. Data yang digunakan dalam algoritma C4.5. akan dipecah menjadi beberapa bagian , yaitu:
   - Data *trainning*: digunakan untuk membentuk pohon keputusan.
   - Data *testing*: digunakan untuk ujicoba pada pohon yang telah

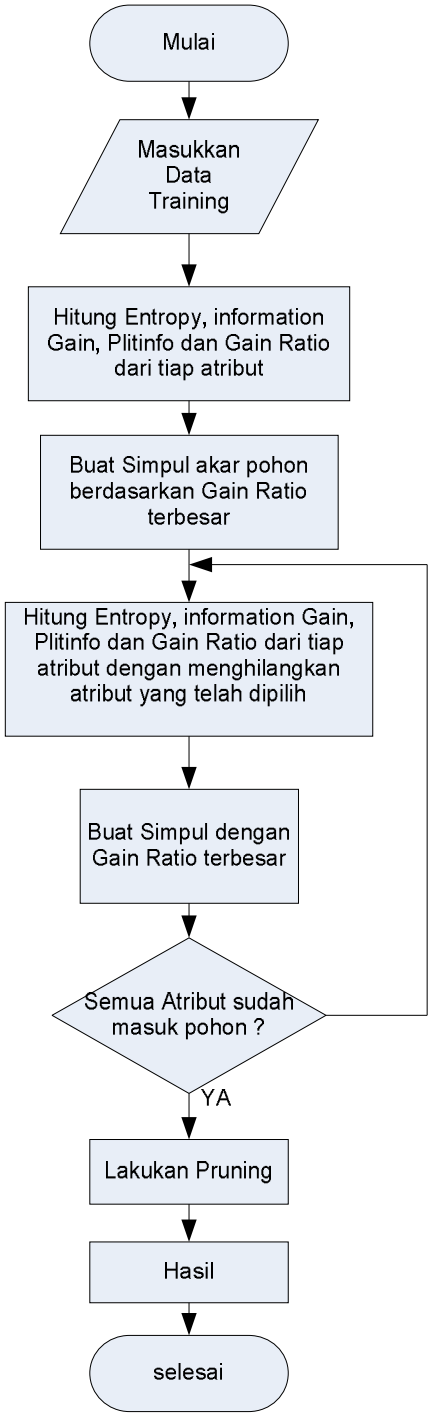
dibentuk guna menghitung nilai *error rate.*

- Data *test pruning*: digunakan untuk menguji akurasi pada pohon yang telah dibentuk guna proses pemangkasan pohon.

3. Analisis kebutuhan sistem
   Analisis Kebutuhan sistem adalah gambaran kebutuhan sistem yang akan dibangun secara keseluruhan. Untuk mengidentifikasi gambaran sistem secara keseluruhan adalah dengan melakukan pengamatan (observasi), kemudian mengidentifikasi kebutuhan perangkat lunak yg dibutuhkan untuk Aplikasi penjurusan
4. Perancangan sistem
   Tahap ini adalah tahap dimana dibuat suatu rancangan untuk membangun aplikasi, mulai dari fitur-fitur atau konten, rancangan *user interface*, *Flowchart* sistem dan *flowcha*rt metode (Gambar 3 dan Gambar 4), *desain usecase (*Gambar 5*), Activity diagram,* desain *data base* (conceptual data model / CDM dan physical data model / PDM).




Gambar 3. Flowchart sistem




Gambar 4 Flowchart algoritma *C 4.5.*

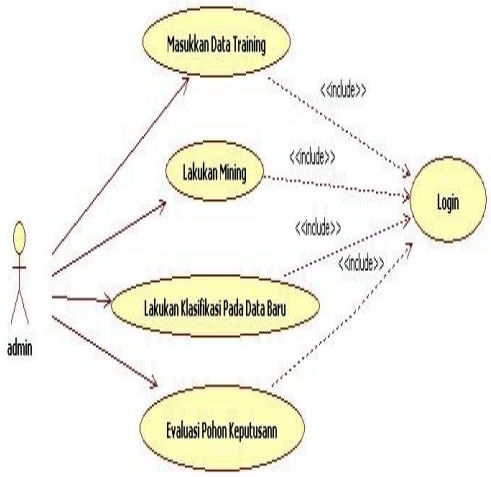
Berikut keterangan dari tiap proses yang terdapat pada Gambar 4.:

a) Data training di inputkan.
b) Hitung *gain ratio, split info* dan *entropy* dari masing-masing atribut data training yang ada.

c) Buat simpul akar dari pemilihan atribut yang memiliki *gain ratio* terbesar.
d) Hitung *gain ratio, split info* dan *entropy* dari masing-masing atribut dengan menghilangkan atribut yang telah dipilih sebelumnya.
e) Buat simpul internal dari pemilihan atribut yang memiliki *gain ratio* terbesar. Cek apakah semua atribut sudah dibentuk pada pohon.
f) Jika belum, maka ulangi proses d dan e, jika sudah maka lanjut pada proses berikutnya.
g) Lakukan pemangkasan pohon untuk menghilangkan cabang-cabang yang tidak perlu.Kemudian aturan keputusan di*generate* mengikuti pohon yang telah dibentuk sebelumnya.

Dalam sistem informasi ini, hanya terdapat satu pengguna yaitu, admin. Admin tersebut dapat melakukan tugas sebagai berikut (sepert pada Gambar 5) :

a. Input data training, yaitu melakukan proses menambah, data yang akan dijadikan data training.
b. Mining, yaitu proses pengolahan data dari data training yang kemudian akan membentuk sebuah aturan *(rule)*. Dari mining akan menghasilkan pembentukan pohon, pada pohon tersebut akan terbentuk sebuah rule yang akan menjadi aturan dari data yang akan diklasifikasikan.
c. Lakukan klasifikasi pada data baru, yaitu input data siswa yang belum penjurusan.
d. Evaluasi pohon keputusan yaitu melakukan perbandingan antar partisi, kemudian dicari nilai akurasi yang terbaik. Dari nilai akurasi yang terbaik itu lah yang akan dijadikan acuan pada proses klasifikasi selanjutnya.




Gambar 5. *Use Case Diagram*

## HASIL DAN PEMBAHASAN

### Data yang Digunakan

Data merupakan data kategorikal dan tidak ada *mising value* pada data. Jumlah data yang digunakan sebanyak 200 data. Dalam implementasinya, data dipecah menjadi beberapa bagian, yaitu:

**Proses Mining C4.5**

Dalam proses mining C4.5, proses yang dilakukan adalah sebagai berikut:

1. Hitung frekuensi kemunculan masing-masing nilai atribut pada data survey.
2. Hitung nilai *Entropy* dari masing-masing nilai atribut.
3. Hitung nilai *Information Gain* dengan menggunakan nilai *Entropy* yang telah dihitung sebelumnya.
4. Hitung nilai *Split Info* dari tiap atribut.
5. Hitung nilai *Gain Ratio* menggunakan nilai *Information Gain* dan *Split Info.*
6. Ambil nilai *Gain Ratio* terbesar dan jadikan simpul akar.
7. Hilangkan atribut yang dipilih sebelumnya dan ulangi perhitungan nilai *Entropy*, *Information Gain*, *Split Info* dan

*Gain Ratio* dengan memilih *Gain Ratio* terbesar dan dijadikan simpul *internal* pohon.

8. Ulangi perhitungan tersebut hingga semua atribut pohon memiliki kelas.
9. Jika semua pohon sudah memilik kelas, maka tampilkan pohon keputusan awal dan generate aturan keputusan awal.

**Skenario 1**

Pada skenario 1 ini data yang digunakan yaitu 125 data training dan 75 data testing. Data skenario 1 seperti ditunjukkan pada Tabel 1.

Tabel 1. Tabel Data Skenario 1

| Jumlah Data | Algoritma C 4.5 | |
|---|---|---|
| | Data Training | Data Testing |
| IPA (140) | 121 | 19 |
| IPS (60) | 54 | 6 |
| Jumlah | 175 | 25 |

**Analisa Algoritma**

Setelah pohon dibentuk, selanjutnya dilakukan perbandingan dengan data yang merupakan data *testing*, data yang digunakan ada 25 data dimana data tersebut dilakukan pengklasifikasian menggunakan *rule* C4.5. yang telah dibentuk. Kemudian kelas yang terbentuk dibandingkan dan dihitung nilai *error rate*nya. Setelah proses klasifikasi, kemudian dihitung kinerja dari masing-masing algoritma yang meliputi akurasi, *error rate*, *precision* dan *recall*. Berikut hasil kinerja C 4.5 seperti pada Tabel 2.

Pada Skenario 1 terdapat penilaian kinerja algoritma C4.5. Penilaian kinerja diperoleh dari hasil klasifikasi rule algoritma dengan data testing. Dari Hasil penilaian kinerja diketahui algoritma C 4.5 nilai akurasi C4.5 sebesar 96% sedangkan tanpa *pruning* sebesar 87%.

Tabel 2. Kinerja Algoritma C 4.5.

| **Kinerja** | **Skenario 1** | |
|---|---|---|
| | **C4.5** | **C4.5 Post Pruning** |
| | **25:75** | **125:75** |
| Akurasi | 94% | 96% |
| Error Rate | 7% | 5% |
| Precision | 93,51% | 94,74% |
| Recall | 98,63% | 98,63% |

**KESIMPULAN**

Adapun kesimpulan dari penelitian ini adalah :

1. Dari pengukuran kinerja algoritma yang telah dilakukan, dapat disimpulkan algoritma C4.5 post pruning memiliki kinerja (*precision, recall,* dan *accuracy*) yang lebih baik dibandingkan tanpa pruning
2. Metode post pruning merupakan metode pruning yang lebih baik dari pada pre pruning, hal ini dapat dilihat pada nilai akurasi sebesar 96% sedangkan tanpa *pruning* sebesar 87%.

**DAFTAR PUSTAKA**

[1] Hardikar S, Shrivastava A, Choudhary V.,"Comparison between ID3 and C4.5 in Contrast to IDS", *VSRD-IJCSIT*, Vol. 2 (7), pp. 659-667, 2012.

[2] Kumar, S. Anupama, dan M.N, Vijayalakshmi, "Efficiency of Decision Trees in Predicting Student's Academic Performance", *Computer Science & Information Technology (CS & IT) DOI: 10.5121/csit.2011.1230*. 1: 5-8, 2011.

[3] Larose, Daniel T., *Discovering Knowledge in Data: An Introduction to Data Mining.* New Jersey : John Willey & Sons, Inc, 2005.

[4] Romansyah, F., Sitanggang I. S., dan Nurdiati, S, "Fuzzy Decision Tree dengan Algoritma ID3 pada Data Diabetes", *Internetworking Indonesia Journal Vol. 1/No. 2*, 2009.