PREDIKSI NILAI PROYEK AKHIR MAHASISWA MENGGUNAKAN ALGORITMA KLASIFIKASI DATA MINING

Paramita Mayadewi[1)], Ely Rosely[2)]
[1,2]D3 Manajemen Informatika, Fakultas Ilmu Terapan, Universitas Telkom
Jl. Telekomunikasi, Ters, Buah Batu, Bandung, 40257
Telp : (022) 7564108, Fax : (022) 7565200
E-mail : paramita@tass.telkomuniversity.ac.id[1)], ely.rosely@tass.telkomuniversity.ac.id[2)]

Abstrak

Jumlah data yang tersimpan dalam database perguruan tinggi bertambah dengan cepat. Sebagian data-data tersebut berisi informasi tersembunyi mengenai kinerja mahasiswa dan belum banyak dimanfaatkan untuk memperbaiki kualitas kinerja mahasiswa. Data mining pendidikan digunakan untuk mempelajari data yang tersedia di bidang pendidikan dan membawa keluar pengetahuan tersembunyi yang ada pada data tersebut. Penelitian ini bertujuan untuk membuat aturan yang dapat memprediksi nilai proyek akhir mahasiswa program diploma manajemen informasi berdasarkan nilai-nilai matakuliah yang mendukung penyusunan proyek akhir dengan menggunakan model klasifikasi data mining. Penelitian yang dilakukan juga akan menganalisis prestasi mahasiswa pada matakuliah yang mendukung penyusunan proyek akhir dengan pencapaian nilai proyek akhir mereka. Prediksi ini diharapkan dapat membantu dalam mengidentifikasi nilai berdasarkan matakuliah yang mendukung proyek akhir mereka. Berdasarkan penelitian yang telah dilakukan, analisis prediksi menggunakan ID3 memiliki akurasi sebesar 62,66%, CHAID 63,66% dan Naïve Bayes 65,67%.

***Kata kunci**: data mining, klasifikasi, proyek akhir*

Abstract

The amount of data stored in the database of colleges is growing rapidly. Most of these data contain hidden information about student performance and have not been widely used to improve the quality of student performance. Education data mining is used to study the data available in the field of education and bring out the hidden knowledge that exist in the data. This study aims to create a rule that can predict student performance of diploma program of information management based on the values of the subjects that support the preparation of their final projects using classification data mining. Research will also analyze the achievements of students in subjects that support the final project with the achievement of their final project. This prediction is expected to assist in identifying student achievement based courses that support their final projects. Based on the research that has been done, predictive analysis using ID3 have an accuracy of 62,66%, CHAID 63,66% and Naïve Bayes 65,67%.

***Keywords**: data mining, classification, final project*

1. PENDAHULUAN

Data mining adalah proses pencarian pola data yang tidak diketahui atau tidak diperkirakan sebelumnya. Konsep data mining dapat diterapkan dalam berbagai bidang seperti pemasaran, pendidikan, kesehatan, pasar saham, customer relationship management (CRM), teknik, dan lain sebagainya. Educational Data Mining (EDM) adalah proses mengubah data mentah dari sistem akademik menjadi informasi yang berguna untuk mengambil keputusan dan menjawab pertanyaan penelitian. EDM fokus pada metode pengembangan yang menemukan pengetahuan pada data yang berasal dari lingkungan pendidikan. Berbagai teknik data mining seperti klasifikasi, clustering, dan rule mining dapat diterapkan untuk membawa keluar berbagai pengetahuan tersembunyi dari data pendidikan.

Penelitian menggunakan data mining dalam dunia pendidikan telah dilakukan oleh Abeer dan Ibrahim (2014) [1] untuk melakukan studi prediksi kinerja siswa menggunakan model klasifikasi dengan algoritma decision tree ID3. Keluaran yang dihasilkan dari studi tersebut adalah sebuah model aturan yang digunakan dalam memprediksi nilai akhir siswa. Jai dan David (2014) [2] melakukan studi prediksi kinerja siswa menggunakan model klasifikasi dengan menggunakan algoritma klasifikasi dan membandingkan kinerja algoritma tersebut berdasarkan studi yang dilakukan. Algoritma yang digunakan adalah ID3, J48, REP Tree, Simple Cart, NB Tree, MLP (Multilayer Perceptron) serta Decision Table. Studi yang dilakukan menggunakan software WEKA. Hasil studi menunjukkan algoritma MLP menghasilkan kinerja yang lebih baik dibandingkan dengan algoritma yang

lainnya. Kalpesh, Aditya, Amiraj, Rohit dan Vipul (2013) [3] melakukan studi dalam mempredikasi kinerja siswa menggunakan algoritma klasifikasi ID3 dan C4.5. Studi yang dilakukan menggunakan software RapidMiner. Surjeet dan Saurabh (2012) [4] menerapkan algoritma klasifikasi C4.5, ID3 dan CHART dalam studi yang mereka lakukan untuk memprediksi kinerja mahasiswa teknik. Studi yang dilakukan juga membandingkan kinerja algoritma tersebut. Brijesh dan Saurabh (2011) [5] melakukan studi prediksi performansi siswa menggunakan model klasifikasi. Algoritma yang digunakan adalah Bayesian. Bahar (2011) [7] melakukan penelitian tentang kurang akuratnya proses pemilihan jurusan dengan sistem manual pada SMA, sehingga perlu suatu penggunaan metode untuk mengelompokkan siswa dalam proses pemilihan jurusan. Bahar menggunakan algoritma Fuzzy C-Means untuk mengelompokkan data siswa SMA berdasarkan nilai mata pelajaran inti untuk proses penjurusan.

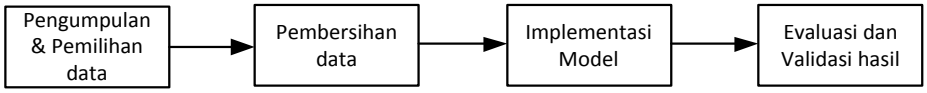
Salah satu indikator keberhasilan mahasiswa D3 Manajemen Informatika dalam menyelesaikan proyek akhir mereka adalah pemahaman mereka terhadap matakuliah yang mendukung proyek akhir mereka. Matakuliah tersebut adalah algoritma dan pemrograman, perancangan basis data, analisis dan perancangan sistem serta pemrograman web. Kebanyakan dari mahasiswa mengulang mengambil matakuliah algoritma dan pemrograman serta matakuliah perancangan basis data. Tujuan utama dari studi ini adalah menggunakan data mining untuk memprediksi nilai proyek akhir mahasiswa berdasarkan nilai-nilai matakuliah yang mendukung proyek akhir mereka. Kemungkinan seorang mahasiswa mengulang sebuah matakuliah dipertimbangan pula dalam studi yang dilakukan. Model data mining yang digunakan dalam studi ini menggunakan model klasifikasi data mining. Metode klasifikasi yang digunakan adalah decision tree (pohon keputusan). Metode Naïve Bayes digunakan sebagai pembanding metode Decision Tree. Untuk metode klasifikasi decision tree digunakan algoritma ID3 dan CHAID.

Klasifikasi adalah teknik yang dilakukan untuk memprediksi class atau properti dari setiap instance data. Model prediksi memungkinkan untuk memprediksi nilai-nilai variabel yang tidak diketahui berdasarkan nilai variabel lainnya. Klasifikasi memetakan data ke dalam kelompok-kelompok kelas yang telah ditetapkan sebelumnya. Klasifikasi disebut juga dengan supervised learning karena kelas data telah ditentukan sebelumnya. Decision tree merupakan model prediksi menggunakan struktur pohon atau struktur hirarki. Konsep dari decision tree adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan. Keunggulan dari penggunaan decision tree adalah kemampuannya untuk mem-break down proses pengambilan keputusan yang kompleks menjadi lebih sederhana sehingga pengambil keputusan akan mudah untuk menginterpresikan solusi dari permasalahan. Decision tree juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi. Metode klasifikasi Naive Bayes merupakan sebuah metoda klasifikasi yang berakar pada teorema Bayes. Konsep dasar teori Bayes itu pada dasarnya adalah peluang bersyarat. Metode ini memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya.

Berdasarkan pada penjelasan diatas, pemilihan metode klasifikasi dengan menggunaan decision tree didasarkan pada kemudahan dalam melakukan identifikasi dan melihat hubungan antara faktor-faktor yang mempengaruhi suatu masalah sehingga dapat dicari penyelesaian terbaik dengan memperhitungkan faktor-faktor tersebut. Adapun pemilihan metode Naïve Bayes adalah hanya memerlukan sejumlah kecil data pelatihan untuk mengestimasi parameter yang dibutuhkan untuk klasifikasi. Selain itu, Naïve Bayes mampu menangani nilai yang hilang dengan mengabaikan instansi selama perhitungan estimasi peluang. Hasil dari model klasifikasi tersebut akan dibandingkan.

2. METODOLOGI PENELITIAN

Tahapan studi yang dilakukan terdiri dari: (1) Pengumpulan data, (2) Pembersihan data, (3) Implementasi model, dan (4) Evaluasi dan validasi hasil.



Gambar 1. Tahapan Studi

2.1 Pengumpulan dan Pemilihan Data

Data yang digunakan dalam studi ini merupakan data mahasiswa program studi D3 Manajemen Informatika angkatan 2009 – 2011. Jumlah data awal adalah 750 *record*. Dalam hal ini, data yang berasal dari berbagai tabel telah disimpan dalam satu tabel. Dalam tahap ini, ditentukan pula atribut-atribut yang akan digunakan dalam studi. Atribut-atribut yang tidak diperlukan dalam studi seperti nama mahasiswa, alamat, jenis kelamin, dan lain sebagainya tidak digunakan dalam studi yang dilakukan. Pemilihan atribut yang digunakan dalam studi adalah atribut nilai matakuliah yang mendukung pengerjaan proyek akhir mahasiswa sebagaimana yang telah dijelaskan sebelumnya. Kemungkinan mahasiswa mengulang matakuliah-matakuliah tersebut juga dipertimbangkan dalam studi yang dilakukan. Tabel 1 menjelaskan atribut dalam studi yang dilakukan.

Tabel 1. Atribut Mahasiswa

Atribut	Keterangan	Nilai Atribut
PA	Proyek Akhir	{A,B}
Alpro	Nilai Algoritma	{A,B,C,D}
Peranc_DB	Nilai Perancangan Basisdata	{A,B,C,D}
APSI	Nilai Analisis & Perancangan Sistem	{A,B,C,D}
Pemrog_Web	Nilai Pemrograman Web	{A,B,C,D}
U_Alpro	Informasi mengulang matakuliah Algoritma	{ya, tidak}
U_Peranc_DB	Informasi mengulang matakuliah Perancangan Basisdata	{ya, tidak}
U_APSI	Informasi mengulang matakuliah APSI	{ya, tidak}
U_Pemrog_web	Informasi mengulang matakuliah Pemrograman Web	{ya, tidak}

2.2 Pembersihan Data

Pembersihan data merupakan langkah yang dilakukan sebelum masuk pada proses mining pada data. Pembersihan data berisi beberapa kegiatan yang tujuan utamanya adalah melakukan pengenalan dan perbaikan pada data yang akan diteliti. Perlunya perbaikan pada data yang akan diteliti disebabkan karena data mentah cenderung tidak siap untuk di-*mining*. Kasus yang sering terjadi adalah adanya *missing values* pada data.

Missing value dalam *dataset* mahasiswa berasal dari data-data yang atributnya tidak memiliki nilai informasi. Informasi ini tidak diperoleh dimungkinkan karena proses yang terjadi saat penggabungan data. Penanganan *missing value* pada studi ini dilakukan dengan pengurangan obyek data (*under sampling*). Hasil dari pembersihan data yang dilakukan, terdapat 699 *record* dari jumlah awal sebanyak 750 *record*.

2.3 Implementasi Model Data Mining

Dalam tahap implementasi ini dilakukan penerapan algoritma model klasifikasi yang akan digunakan, yaitu ID3, CHAID dan Naïve Bayes. Penerapan algoritma menggunakan perangkat lunak RapidMiner.

RapidMiner merupakan perangkat lunak yang bersifat *open source*. RapidMiner merupakan solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. RapidMiner merupakan perangkat lunak yang berdiri sendiri untuk analisis data dan sebagai mesin data mining yang diintegrasikan pada produknya sendiri [6].

2.4 Evaluasi dan Validasi Hasil

Hasil dari implementasi model data mining akan terdapat sebuah *rule*. *Rule* yang dihasilkan akan digunakan sebagai dasar prediksi nilai yang akan dilakukan. Sebelumnya, *rule* tersebut harus dievaluasi dan divalidasi sehingga diketahui seberapa akurat hasil prediksi yang akan dilakukan. Evaluasi dan validasi hasil *rule* klasifikasi dilakukan dengan *confusion matrix* [8] dan ROC (*Receiver Operating Characteristic*) Curve [9].

Tabel 2. Contoh Penggalan Rule yang dihasilkan

```
Jika (alpro=A) AND (peranc_db=A) AND (webpro=A) AND (apsi=A) AND (u_alpro=T)AND
(u_peranc_db=T) AND (u_apsi=T) AND (u_webpro=T) THEN PA = A.
Jika (alpro=A) AND (peranc_db=A) AND (webpro=A) AND (apsi=A) AND (u_alpro=Y)AND THEN PA = A.
……..
```

Confusion matrix adalah suatu metode yang digunakan untuk melakukan perhitungan akurasi pada konsep data mining. Evaluasi dengan *confusion matrix* menghasilkan nilai akurasi, presisi dan *recall*. Akurasi dalam klasifikasi adalah persentase ketepatan record data yang diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi [8]. Presisi atau *confidence* adalah proporsi kasus yang diprediksi positif yang juga positif benar pada data yang sebenarnya. *Recall* atau *sensitivity* adalah proporsi kasus positif yang sebenarnya yang diprediksi positif secara benar [10].

Tabel 3. Model *Confusion Matrix* [8]

Correct Classification	**Classified as**	
	+	-
+	True positives (A)	False negatives (B)
-	False positives (C)	True negatives (D)

Perhitungan akurasi dengan tabel *confusion matrix* adalah sebagai berikut:

Akurasi = (A+D)/(A+B+C+D) (1)

Presisi didefinisikan sebagai rasio item relevan yang dipilih terhadap semua item yang terpilih. Presisi dapat diartikan sebagai kecocokan antara permintaan informasi dengan jawaban terhadap permintaan tersebut. Rumus presisi adalah:

Presisi = A/(C+A) (2)

Recall didefinisikan sebagai rasio dari item relevan yang dipilih terhadap total jumlah item relevan yang tersedia. *Recall* dihitung dengan rumus:

Recall = A/(A+D) (3)

Presisi dan *Recall* dapat diberi nilai dalam bentuk angka dengan menggunakan perhitungan persentase (1-100%) atau dengan menggunakan bilangan antara 0-1. Sistem rekomendasi akan dianggap baik jika nilai presisi dan *recall*nya tinggi.

Kurva ROC menunjukkan akurasi dan membandingkan klasifikasi secara visual. ROC mengekspresikan *confusion matrix*. ROC adalah grafik dua dimensi dengan *false positive* sebagai garis horizontal dan *true positive* sebagai garis vertikal. AUC (*the area under curve*) dihitung untuk mengukur perbedaan performansi metode yang digunakan. ROC memiliki tingkat nilai diagnosa yaitu [9]:

a. Akurasi bernilai 0,90 – 1,00 = *excellent classification*
b. Akurasi bernilai 0,80 – 0,90 = *good classification*
c. Akurasi bernilai 0,70 – 0,80 = *fair classification*
d. Akurasi bernilai 0.60 – 0,70 = *poor classification*
e. Akurasi bernilai 0.50 – 0.60 = *failure*

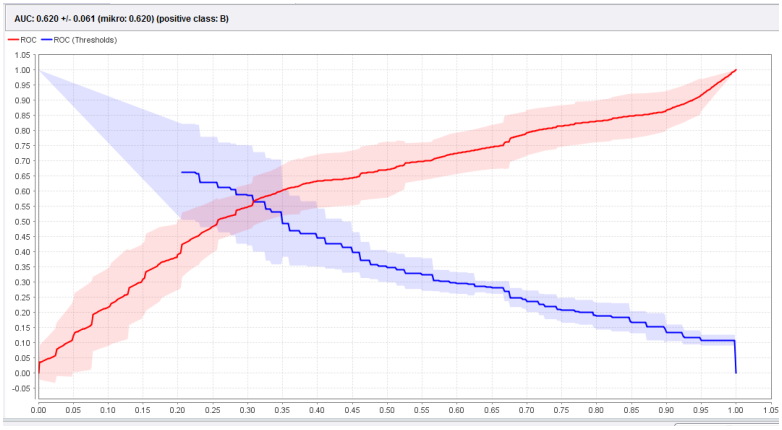
3. HASIL DAN PEMBAHASAN

Klasifikasi dilakukan dengan RapidMiner menggunakan algoritma ID3, CHAID serta Naïve Bayes. Pengujian hasil klasifikasi ditunjukkan dengan *x-validation* menggunakan 10 *fold*. Hasil *x-validation* dengan algoritma ID3, ditunjukkan dalam tabel 4. Nilai akurasi adalah 62,66%. Nilai presisi dan *recall* berturut-turut 61,11% dan 39,67%.

Tabel 4. *x-Validation* dengan *Decision Tree* ID3

***Accuracy* = 62,66%**	*True* A	*True* B	*Class Precision*
Pred. A	317	184	63,27%
Pred. B	77	121	61,11%
Class Recall	80,46%	39,67%	

Hasil pengolahan ROC untuk algoritma ID3 adalah 0.620, yang dapat dilihat pada gambar 2 dengan tingkat diagnosa *poor classification*.



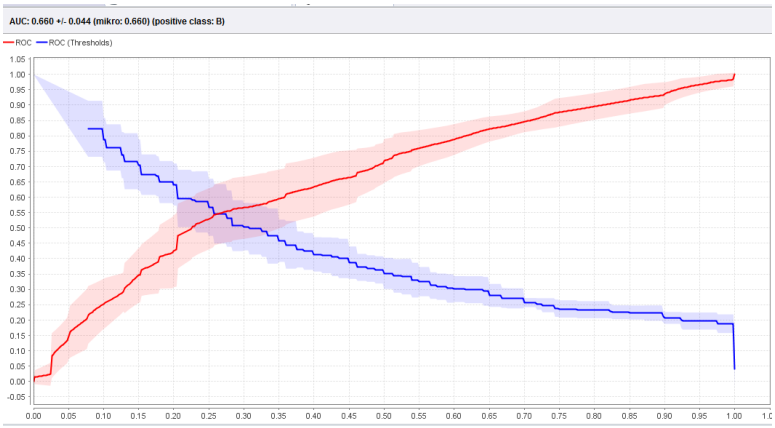
Gambar 2. Hasil ROC dari ID3 yang dihasilkan oleh RapidMiner

Hasil *x-validation* dengan algoritma CHAID, ditunjukkan dalam tabel 5. Nilai akurasi adalah 63,66%. Nilai presisi dan *recall* berturut-turut 61,97% dan 43,28%.

Tabel 5. *x-Validation* dengan *Decision Tree* CHAID

***Accuracy* = 63,66%**	*True* A	*True* B	*Class Precision*
Pred. A	313	173	64,40%
Pred. B	81	132	61,97%
Class Recall	79,44%	43,28%	

Hasil pengolahan ROC untuk algoritma CHAID adalah 0.660, yang dapat dilihat pada gambar 3 dengan tingkat diagnosa *poor classification*.



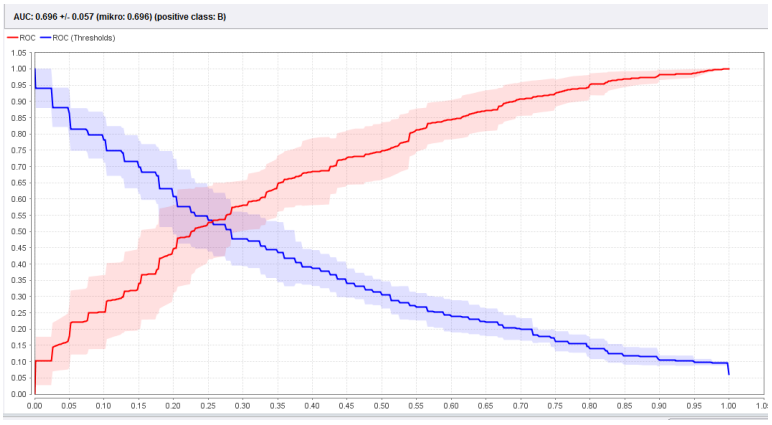
Gambar 3. Hasil ROC dari CHAID yang dihasilkan oleh RapidMiner

Hasil *x-validation* dengan algoritma Naïve Bayes, ditunjukkan dalam tabel 6. Nilai akurasi adalah 65,67%. Nilai presisi dan *recall* berturut-turut 62,08% dan 54,75%.

Tabel 6. *x-Validation* dengan Naïve Bayes

***Accuracy* = 65,67%**	*True* A	*True* B	*Class Precision*
Pred. A	292	138	67,91%
Pred. B	102	167	62,08%
Class Recall	74,11%	54,75%	

Hasil pengolahan ROC untuk algoritma Naïve Bayes adalah 0.696, yang dapat dilihat pada gambar 4 dengan tingkat diagnosa *poor classification*.



Gambar 4. Hasil ROC dari Naïve Bayes yang dihasilkan oleh RapidMiner

Kinerja (tingkat akurasi) model yang dihasilkan oleh tiga algoritma yang digunakan yaitu algoritma ID3, algoritma CHAID dan algoritma Naïve Bayes dalam memprediksi nilai proyek akhir mahasiswa, dijelaskan dalam tabel 2 berikut ini.

Tabel 7. Perbandingan Nilai Akurasi dan AUC Algoritma ID3, CHAID dan Naïve Bayes

Algoritma	Akurasi	AUC
ID3	62,66%	0,620
CHAID	63,66%	0,660
Naïve Bayes	65,67%	0,696

4. SIMPULAN DAN SARAN

4.1 Simpulan

Berdasarkan analisis data kinerja mahasiswa pada mata kuliah pendukung proyek akhir mereka menggunakan algoritma ID3, CHAID dan Naïve Bayes berdasarkan literatur yang digunakan, dapat disimpulkan bahwa ketiga algoritma tersebut tidak dapat diterapkan untuk menentukan nilai proyek akhir berdasarkan mata kuliah pendukung proyek akhir. Hasil evaluasi dan validasi menggunakan *confusion matrix* serta kurva ROC menunjukkan tingkat diagnosa *poor classification*. Berdasarkan hasil ini pula dapat disimpulkan bahwa tidak terdapat hubungan dan pengaruh yang kuat terhadap nilai proyek akhir mahasiswa berdasarkan pencapaian mereka pada mata kuliah pendukung proyek akhir, termasuk apakah mahasiswa pernah mengulang mata kuliah tersebut atau tidak.

4.2 Saran

Perlu dilakukan studi lebih lanjut dalam kasus prediksi nilai proyek akhir mahasiswa, misalnya dengan menggunakan metode validasi yang berbeda serta melakukan perbandingan dengan beberapa algoritma lainnya sehingga diperoleh algoritma dengan tingkat akurasi yang baik. Selain itu, dapat pula melibatkan beberapa atribut lainnya seperti jumlah kehadiran siswa, motivasi belajar siswa dan lain sebagainya.

5. DAFTAR RUJUKAN

[1] Abeer Badr, Ibrahim Sayed, 2014. Data Mining:A predictionfor student's Performance Using Clasification Method. *World Journal of Computer Application and Technology 2(2): 43-47*.

[2] Jai Ruby, K.David, 2014. *Predicting the Performance of Students in Higher Education Using Data Mining Classification Algorithms – A Case Study*, *International Journal for Research in Applied Science & engineering Technology (IJRASET).*Volume 2 Issue XI, November 2014.

[3] Kalpesh A, Aditya G, Amiraj D, Rohit J, Vipul H., 2013. Predicting Students's Performance Using ID3 and C4.5 Classification Algorithms, *International Journal of Data Mining & Knowledge Management Process* Vol. 3, No.5, September 2013.

[4] Surjeet K, Saurabh Pal, 2012. Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification, *World of Computer Science and Information Technology Journal* Vol. 2, No. 251-56.

[5] Brijesh K, Saurabh Pal, 2011. Data Mining: A Prediction for Performance Improvement using Classification, *International Journal of Computer Science and Information Security* Vol 9, No. 4, April 2011.

[6] Bala Deshpande. *Decision Tree Digest – An eBook.* SimaFore.

[7] Bahar, 2011. Penentuan Jurusan Sekolah Menengah Atas Dengan Algoritma Fuzzy C-Means. Jurnal Teknologi Informasi.

[8] Han, J & Kamber, M, 2006. Data Mining Concepts & Techniques 2nd Edition. San Fransisco: Elsevier.

[9] Gorunescu, F, 2011. Data Mining Concepts, Model and Techniques. Berlin: Springer.

[10] Powers D, 2011. Evaluation: From Precision, Recall, and F-Measure to ROC, Infomedness, Markedness & Correlation, *Journal of Machine Learning Technologies*, 37-63.