PENCARIAM JURUSAN SUBANG DENGAN ALGORITMA C 4.5 DAN DATA MINING STMIK SUBANG

Timbo Faritcan Parlaungan Siallagan

Program Studi Teknik Informatika, STMIK Subang
Jl. Marsinu No. 5 - Subang, Tlp. 0206-417853 Fax. 0206-411873
E-mail: timbo.siallagan@yahoo.com

ABSTRAKSI

Penentuan Program Studi di perguruan tinggi tidak bisa dianggap mudah Banyak kasus dijumpai bahwa penentuan program studi yang tidak sesuai, kepribadian, minat dan bakat dapat mempengaruhi mahasiswa dalam mata kuliah. Berdasarkan pertimbangan diatas, penelitian akan mengunakan pendekatan algoritma C 4.5 untuk menentukan program studi yang akan diambil oleh mahasiswa sesuai dengan latar belakang, minat dan kemampuannya sendiri. Adapun parameter dalam pemilihan Program Studi adalah Indeks Tes Nilai Ujian Saringan Masuk. Berdasarkan hasil eksperimen dan evaluasi yang dilakukan maka dapat disimpulkan bahwa Algoritma C 4.5 akurat diterapkan untuk penentuan kesesuaian program studi mahasiswa dengan tingkat akurasi rekomendasi program sebesar 73,46 %.

Kata Kunci: ***Klasifikasi, Penentuan Program Studi, Algoritma C4.5***

1. Pendahuluan

1.1. Latar Belakang

Setiap tahun, siswa kelas XII SMU/SMK dan Sederajat yang berencana melanjutkan jenjang pendidikannya ke perguruan tinggi harus memutuskan pilihan, ke bidang atau jurusan apa akan melanjutkan pendidikannya kelak. Dan ini adalah sesuatu yang cukup sulit untuk diputuskan oleh kebanyakan siswa SMU/SMK dan Sederajat, terutama yang tidak banyak memiliki referensi dan mencari informasi terkait dengan pendidikan tinggi. Kecenderungan yang terjadi saat ini, banyak siswa kelas XII yang tidak tahu minatnya dan bakatnya serta akan memilih prodi apa selepas SMU nanti (Indri Savitri, M.Psi, 2006). Keputusan para siswa, terkadang dipengaruhi oleh pendapat orang tua, guru, teman atau figur-figur yang diidolakan (M, 2010).

Dengan hanya mendasarkan pendapat orang lain dan tanpa menelaah kemampuannya, seorang siswa bisa membuat keputusan yang sangat bertolak belakang dengan minat dan bakatnya. Akibat yang buruk terjadi setelah itu, yaitu keengganan belajar dan menurunnya kualitas serta prestasi akademik karena siswa merasa salah dalam memilih jurusan (Mulyadi, 2010).Memilih program studi di perguruan tinggi tidak bisa dianggap mudah dan sepele. Banyak kasus dijumpai bahwa pemilihan jurusan yang tidak sesuai dengan kemampuan, kepribadian, minat dan bakat dapat mempengaruhi mahasiswa dalam mengikuti perkuliahan. Dalam beberapa penelitian psikologi pendidikan, minat dan bakat siswa diketahui cukup terkait dengan prestasi akademiknya (M, 2010). Sehingga siswa calon mahasiswa dapat menentukan studi sesuai dengan kemampuannya.Untuk penentuan semacam ini, Zhiwu Liu, dkk (Liu & Zhang, 2010) telah menggunakan pendekatan pohon keputusan (decision tree) untuk melakukannya. Mereka memanfaatkan sifat prediksi yang dimiliki pohon keputusan. Banyak algoritma yang dapat dipakai dalam pembentukan pohon keputusan, antara lain: Algoritma ID3, CART, dan C4.5. Algoritma C4.5 merupakan pengembangan dari algoritma ID3 (Daniel T, 2005). Rong Cao dan Lizhen Xu (Cao & Xu, 2009) menggunakan Algoritma C4.5 untuk menganalisa penjualan. Sementara itu, dalam bidang pendidikan, Ossy N (Ossi, 2006) melakukan penelitian mengenai bagaimana sebuah model fuzzy dapat

digunakan juga untuk membuat klasifikasi siswa yang mengikuti suatu kelas dengan kemungkinan berhasil atau gagal. Lebih jauh, Wen-Chih Chang, dkk (Wen-Chih, 2009), telah melakukan penelitian untuk mengukur kemampuan belajar siswa. Mereka menggunakan algoritma K-Means untuk membentuk klaster-klaster kemampuan.

Karena itu adalah mungkin untuk menggunakan pendekatan algoritma klasifikasi data mining untuk menentukan jurusan dalam bidang studi yang akan diambil oleh mahasiswa. Identifikasi ini penting diketahui di awal studi sehingga calon mahasiswa tidak salah dalam memilih jurusan yang akan di tempuh selama belajar pada perguruan tinggi. Upaya rekomendasi untuk pemilihan studi semacam ini juga mendapat perhatian dari Thomas Meller, dkk (Thomas & et, 2009). Mereka membangun program online untuk sistem rekomendasi dengan mengunakan pendekatan algoritma naive bayes dan algoritma J48.

Bahar melakukan penelitian tentang kurang akuratnya proses pemilihan jurusan dengan sistem manual pada Sekolah Menengah Atas menyebabkan perlunya suatu penggunaan metode komputasi untuk mengelompokan siswa dalam proses pemilihan jurusan menggunakan algoritma Fuzzy C-Means untuk mengelompokkan data siswa Sekolah Menengah Atas berdasarkan Nilai mata pelajaran inti untuk proses penjurusan (Bahar, 2011). Penelitian ini juga menguji tingkat akurasi algoritma Fuzzy C-Means dalam penentuan jurusan pada Sekolah Menengah Atas.

Demikian juga dengan Sumanto, melakukan penelitian tentang kurang akuratnya mahasiswa dalam pemilihan peminatan Tugas Akhir (Sumanto, 2010)yang sesuai dengan ilmu yang dikuasai oleh mahasiswa sangat berpengaruh dengan nilai tugas akhir dengan menerapkan Fuzzy C-Means untuk memudahkan mahasiswa dalam pemilihan peminatan tugas akhir dengan baik, sesuai dengan kemampuan mahasiswa dengan tingkat akurasi sebesar 82 %. Berdasarkan pertimbangan diatas, penelitian akan mengunakan pendekatan algoritma Decision Tree C 4.5 untuk menentukan jurusan yang akan diambil oleh mahasiswa sesuai dengan latar belakang, minat dan kemampuannya sendiri. Dengan demikian peluang untuk sukses dalam studi di perguruaan tinggi semakin besar.

1.2. Identifikasi Masalah

Penerapan algoritma Decision Tree C4.5 dalam menentukan Program Studi".

1.3. Tujuan

Menentukan Program Studi mahasiswa yang lebih akurat dengan menggunakan algoritma C 4.5

1.4. Manfaat

Manfaat teoritis penelitian ini yaitu diharapkan dapat menjadi referensi untuk penerapan model algoritma Decision Tree C 4.5 bagi praktisi atau peneliti lain untuk diterapkan pada kasus penelitian yang lain untuk penentuan prodi mahasiswa berdasarkan Nilai Ujian Saringan Masuk.

Manfaat praktis dari penelitian ini adalah diharapkan dapat membantu pihak akademik khususnya manajemen Sekolah Tinggi Keguruan dan Ilmu Pendidikan (STMIK Subang) untuk meningkatkan akurasi dalam proses penentuan Program Studi berdasarkan Nilai Ujian Saringan Masuk

Manfaat kebijakan yaitu diharapkan agar algoritma Decision Tree C4.5 mampu menjadi alat pendukung keputusan yang digunakan oleh pihak Perguruan Tinggi dalam proses penentuan jurusan mahasiswa.

1.5. Metodologi Penelitian

Metode penelitian yang akan digunakan dalam pembuatan sistem penentu kualitas agar-agar tepung ini adalah metode prancangan perangkat lunak *Waterfall*. Pengembangan metode *Waterfall* sendiri melalui beberapa tahapan yaitu:

- Penelitian Lapangan (*Field Research*), Penelitian dilakukan langsung turun kelapangan untuk mendapatkan data dan informasi yang dibutuhkan.
- Penelitian Kepustakaan (*Library Research*), Penelitian ini bertujuan untuk mendapatkan data yang bersifat teori seperti mengumpulkan buku-buku atau bahan lainnya.
- Observasi, Observasi yang dilakukan penulis adalah mengamati secara langsung data yang diperoleh.
- Analisis Perangkat Lunak, Kegiatan analisis perangkat lunak meliputi analisis spesifikasi perangkat lunak yang akan digunakan sebagai alat bantu penelitian.
- Perancangan Perangkat Lunak, Perancangan perangkat lunak meliputi perancangan keras dan perancangann antarmuka dari hasil analisis.
- Implementasi Perangkat Lunak, Implementasi dari hasil analisis dan perancangan perangkat lunak.
- Pengujian Perangkat Lunak, Pengujian terhadap perangkat lunak yang telah diimplementasikan.

2. Tinjauan Pustaka

2.1. Data Mining

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan yaitu Deskripsi. Terkadang peneliti dan analisis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecenerungan yang terdapat dalam data. Sebagai contoh, petugas pengumpulan suara mungkin tidak dapat menemukan keterangan atau fakta bahwa siapa yang tidak cukup profesional akan sedikit didukung dalam pemilihan presiden. Deskripsi dari pola dan kecenderungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan.

Estimasi. Estimasi hampir sama dengan klasifikasi,kecuali variabel target estimasi lebih ke arah numerik daripada ke arah kategori. Model dibangun menggunakan record lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya , pada peninjauan berikutnya estimasi nilai dari variabeltarget dibuat berdasarkan umur pasien, jenis kelamin, indeks berat badan, dan level sodium daarah. Hubungan antara tekanan arah sistolik dan nilai variabel prediksi dalam proses pembelajaran akan menghasilkan model estimasi. Model estimasi yang dihasilkan dapat digunakan untuk kasus baru lainnya.

Contoh lain yaitu estimasi nilai indeks prestasi kumulatif mahasiswa program pascasarjana dengan melihat nilai indeks prestasi mahasiswa tersebut pada saat mengikuti program sarjana

Pertama memeriksa atribut yang menyediakan Informasi Gain tertinggi dalam rangka untuk membagi data training berdasarkan pada atribut itu. Hitung informasi yang diinginkan untuk mengklasifikasikan set dan entropi setiap atribut.

Informasi Gain adalah informasi minus entropi. Informasi dari 2 kelas I
(SYes, SNo) = I(9,5) = -9/14 log2(9/14) – 5/14 log2(5/14) = 0.94
Untuk Age terdapat 3 nilai:

- age<=30 (2 yes and 3 no),
- age31..40 (4 yes and 0 no)
- dan age>40 (3 yes 2 no)

Entropy(age) = 5/14 (-3/5log(3/5)-2/5log(2/5)(09) + 0 + 5/14(0.9709)
= 5/14 (-2/5 log(2/5)-3/5log(3/5)) + 4/14 (0)
= 5/14(0.97
= 0.6935
Gain(age) = 0.94 – 0.6935 = 0.2465

Untuk Income terdapat 3 nilai:

- income high (2 yes and 2 no),

- income medium (4 yes and 2 no)
- dan income low (3 yes 1 no)

Entropy(income) = 4/14(-2/4log(2/4)-2/4log(2/4)) + 6/14 (-4/6log(4/6) – 2/6log(2/6)) + 4/14 (-3/4log(3/4)-1/4log(1/4))
= 4/14 (1) + 6/14 (0.918) + 4/14 (0.811)
= 0.285714 + 0.393428 + 0.231714
= 0.9108
Gain(income) = 0.94 – 0.9108 = 0.0292

Untuk Student terdapat 2 nilai:

- student yes (6 yes and 1 no)
- dan student no (3 yes 4 no)

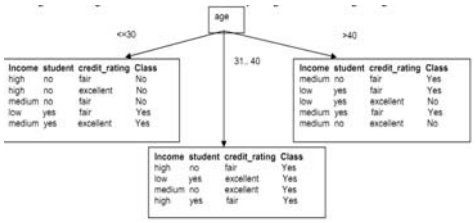
Entropy(student) = 7/14(-6/7log(6/7)) + 7/14(-3/7 log(3/7)-4/7log(4/7)
= 7/14(0.5916) + 7/14(0.9852)
= 0.2958 + 0.4926
= 0.7884
Gain (student) = 0.94 – 0.7884 = 0.1516

Untuk Credit_Rating terdapat 2 nilai:

- credit_ratingfair (6 yes and 2 no)
- dan credit_ratingexcellent (3 yes 3 no)

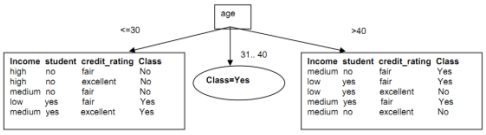
Entropy(credit_rating) = 8/14(-6/8log(6/8)-2/8log(2/8)) + 6/14(-3/6log(3/6) - 3/6log(3/6))
= 8/14(0.8112) + 6/14(1)
= 0.4635 + 0.4285
= 0.8920
gain(credit_rating) = 0.94 – 0.8920 = 0.479

Sejak Usia memiliki Informasi Gain tertinggi kita mulai membelah dataset menggunakan atribut usia



Gambar 1 Pembagian node keputusan berdasar usia

setelah semua record di bawah cabang age31 .. 40 semua class Yes, kita dapat mengganti daun dengan class = Yes



Gambar 2 Pergantian node keputusan

Proses yang sama dari pemecahan harus terjadi untuk dua cabang yang tersisa. Untuk cabang $age_{<=30}$ terdapat atribut income, student and credit_rating.
Informasi yang saling bergantung adalah:

$$I(S_{Yes}, S_{No}) = I(2,3) = -2/5 \log2(2/5) – 3/5\log2(3/5) = 0.97$$

Untuk Income terdapat 3 nilai:

- $income_{high}$ (0 yes and 2 no),
- $income_{medium}$ (1 yes and 1 no)
- dan $income_{low}$ (1 yes and 0 no)

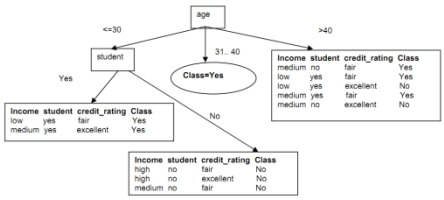
Entropy(income) = 2/5(0) + 2/5 (-1/2log(1/2)-1/2log(1/2)) + 1/5 (0)
= 2/5 (1)
= 0.4
Gain(income) = 0.97 – 0.4 = 0.57

Untuk Student terdapat 2 nilai:

- $student_{yes}$ (2 yes and 0 no)
- dan $student_{no}$ (0 yes 3 no)

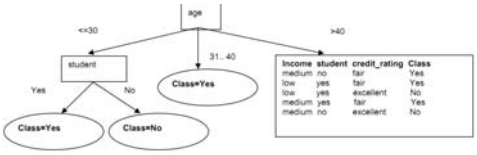
Entropy(student) = 2/5(0) + 3/5(0) = 0
Gain (student) = 0.97 – 0 = 0.97

Kemudian dapat membagi pada atribut student tanpa memeriksa atribut lainnya sejak Informasi Gain dimaksimalkan



Gambar 2 Pembagian node keputusan berdasar student

Karena kedua cabang baru dari kelas yang berbeda, kita membuat mereka menjadi node daun dengan kelas masing-masing sebagai label:



Gambar **Error! No text of specified style in document.**3 Pergantian Node Keputusan

Sekali lagi proses yang sama diperlukan untuk cabang lain dari age.
Informasinya adalah $I(S_{Yes}, S_{No})= I(3,2)= -3/5 \log2(3/5) – 2/5\log2(2/5)=0.97$

Untuk Income terdapat 2 nilai:

- $income_{medium}$ (2 yes and 1 no)
- and $income_{low}$ (1 yes and 1 no)

Entropy(income) = 3/5(-2/3log(2/3)-1/3log(1/3)) + 2/5 (-1/2log(1/2)-1/2log(1/2))
= 3/5(0.9182)+2/5 (1)
= 0.55+0. 4
= 0.95
Gain(income) = 0.97 – 0.95 = 0.02

Untuk Student terdapat 2 nilai:

- $student_{yes}$ (2 yes and 1 no)
- dan $student_{no}$ (1 yes and 1 no)

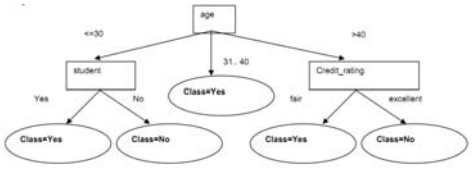
Entropy(student) = 3/5(-2/3log(2/3)-1/3log(1/3)) + 2/5(-1/2log(1/2)-1/2log(1/2))
= 0.95
Gain (student) = 0.97 – 0.95 = 0.02

Untuk Credit_Rating terdapat 2 nilai:

- $credit\_rating_{fair}$ (3 yes and 0 no)
- dan $credit\_rating_{excellent}$ (0 yes and 2 no)

Entropy(credit_rating) = 0
Gain(credit_rating) = 0.97 – 0 = 0.97

Kemudian kita bagi berdasarkan pada credit_rating. Pembagian ini memberi masing-masing partisi dengan record dari kelas yang sama. Kita hanya perlu membuat ke node daun dengan label kelas terlampir



Gambar 4 Pergantian Node Keputusan

3. Analisa

3.1 Algoritma Decision Tree C4.5

Sering disebut dengan pohon keputusan (*decision tree*). Mirip sebuah struktur pohon dimana terdapat node internal yang mendeskripsikan atribut-atribut, setiap cabang menggambarkan hasil dari atribut yang diuji, dan setiap daun menggambarkan kelas (Kusrini & Lutfhfi, 2009).

3.2 Gambaran Umum

Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut (Kusrini & Lutfhfi, 2009):

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

3.3 Tahapan Algoritma Decision

Tree C4.5

1. Menyiapkan data training
2. Menentukan akar dari pohon:
3. Hitung nilai Gain:]Ulangi langkah ke-2 hingga semua tupel terpartisi]Proses partisi pohon keputusan akan berhenti saat:
4. Semua tupel dalam node N mendapat kelas yang sama
5. $Gain\ (S, A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy\ (S_i)$

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

4. Hasil dan Pembahasan

4.1 Implementasi

Model/Metode

Data hasil modifikasi yang akan dipergunakan untuk pengujian sebagaimana Tabel

Data Sampel terdiri dari atribut Nama, TBS,TP,TK,TPA,TKA dan UN

Untuk pengujiannya menggunakan Software data mining yaitu RapidMiner, untuk uji pertama melalui data sample yaitu data angkatan 2013, pada bagian, NAMA akan dihilangkan untuk mendapatkan akurasi yang lebih tinggi, pada bagian keterangan untuk penentuan program studi ada 2 kategori yaitu sesuai dan tidak sesuai dijadikan sebagai label dalam RapidMiner sehingga untuk hasilnya menggunakan software RapidMiner bisa dilihat pada gambar dibawah:

Tabel 2 Klasifikasi Nilai

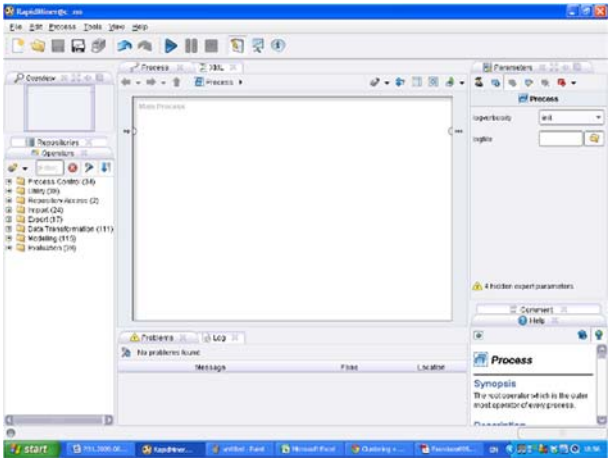
No	NILAI	Klasifikasi
1	0-55	Rendah
2	56-75	Sedang
3	76-100	Tinggi

Tabel 3 Data hasil Modifikasi

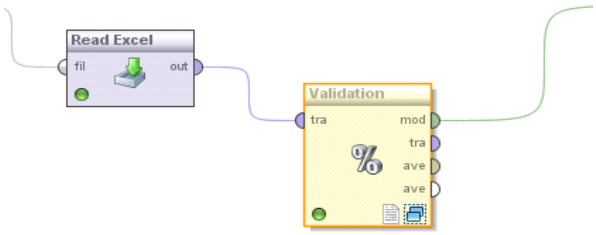
ExampleSet (125 examples, 2 special attributes, 7 regular attributes)

Role	Name	Type	Statistics	Range
id	NO	integer	avg = 63 +/- 36.228	[1.000 ; 125.000]
label	HASIL	binominal	mode = lulus (84), least = tida	lulus (84), tidak lulus (41)
regular	T B S	integer	avg = 67.080 +/- 11.277	[45.000 ; 88.000]
regular	T P	integer	avg = 66.280 +/- 9.928	[45.000 ; 88.000]
regular	T K	integer	avg = 67.160 +/- 11.064	[45.000 ; 89.000]
regular	T PA	integer	avg = 66. avg = 67.160 +/- 11.064	.000 ; 88.000]
regular	T K A	integer	avg = 69.472 +/- 10.671	[45.000 ; 88.000]
regular	UN	integer	avg = 67.264 +/- 10.870	[45.000 ; 89.000]
regular	PROGRAM STUDI	binominal	mode = pbi (125), least = pbi (	pbi (125)

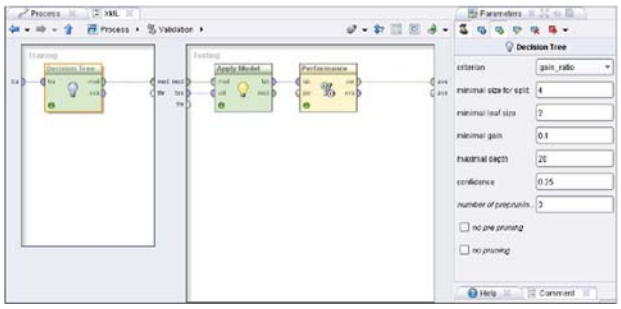
Hasil dari pengolahan data excel menggunakan aplikasi RapidMiner versi 5 pada gambar di bawah ini disajikan editor pengolahan data, langkah awal dipilih menu File-> new, atau dengan menekan icon di pojok kiri atas di bawah menu pilihan File, adapun hasilnya tampak sebagai berikut :



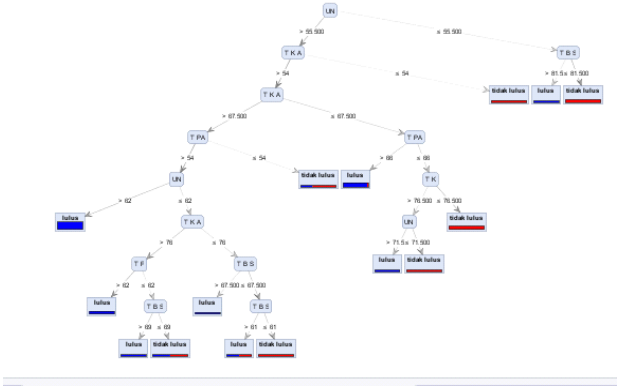
Gambar 6 Tampilan RapidMiner



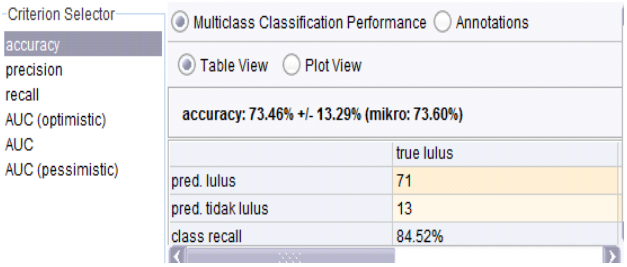
Gambar 7 Numerical X-Validation



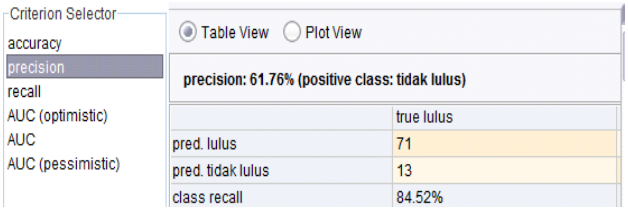
Gambar 8 Decision Tree Clasification Model



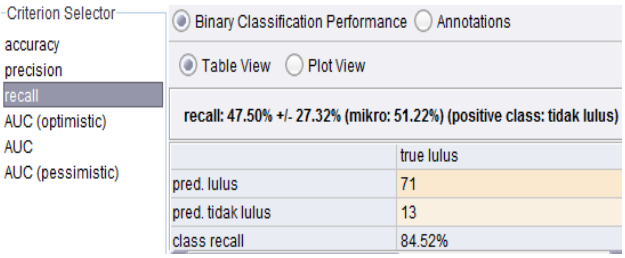
Gambar 9 Decision Tree



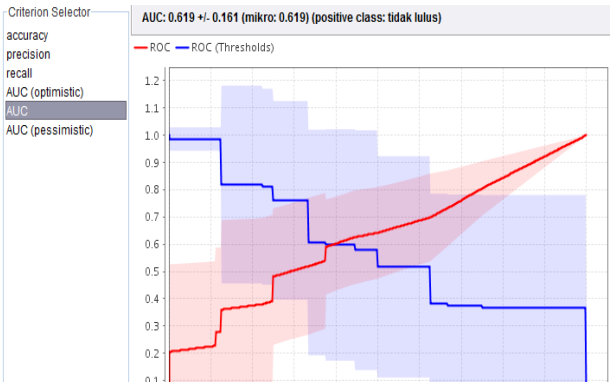
Gambar 10 Performance Vector (Accuracy)



Gambar 11 Performance Vector (Precision)



Gambar 12 Performance Vector (Recall)



Gambar 13Curva ROC

Berdasarkan dari hasil akurasi yang didapat maka nilai AUC nya sebesar 0,619 maka klasifikasi keakuratan tes diagnostiknya termasuk dalam kategori tidak baik.

5. Kimpulan

Dengan adanya penerapan Decision Tree C 4.5 diharapkam mampu memberikan solusi bagi mahasiswa dan dapat membantu STMIK Subang dalam menentukan program studi yang sesuai yang akan ditempuh oleh mahasiswa selama studi sehingga peluang untuk sukses dalam studi di perguruaan tinggi semakin besar

Berdasarkan hasil eksperimen dan evaluasi yang dilakukan maka dapat disiimpulkan bahwa Algoritma Decision Tree C 4.5 akurat diterapkan untuk akurasi rekomendasi program studi sebesar 73,46%

Pustaka

M. North, Data Mining for the Masses,Washington, USA: Agami Press, 2012.

Florin Gorunescu, Data Mining Concept Model Technique, 2011.

Bahar. (2011). Penentuan Jurusan Sekolah Menengah Atas Dengan Algoritma Fuzzy C-Means . Semarang, Indonesia.

A. A. Aziz, N. H. Ismail and F. Ahmad, "Mining Students' Academic Performance, Journal of Theoretical and Applied Information Technology, vol. 53 No.3, 2013.

Liu, Z., & Zhang, X. (2010). Prediction and Analysis for Students' Marks Based on Decision Tree