PREDIKSI HASIL PENJURUSAN SISWA SEKOLAH MENENGAH ATAS DENGAN MENGGUNAKAN ALGORITMA *DECISION TREE* C.45

Imam Sujai[1], Purwanto[2], H.Himawan[3]
[123]Pasca Sarjana Teknik Informatika Universitas Dian Nuswantoro

ABSTRACT

The Majors is one of the placement or distribution process in the selection of high school students teaching program. In these majors, students are given the opportunity to choose majors that best matches the characteristics themselves. The accuracy in choosing majors can determine the success of student learning. In contrast, an excellent opportunity for students will be lost due to lack of inaccuracy in determining the majors. In the 2013 curriculum, majors in high school started in class X after being accepted as a student, so the school should really be able to classify students on the correct corresponding majors talents and interests of students. In studies using the C4.5 algorithm to create a predictive model results placement of students because this method has been used a lot in previous studies to predict the various cases problems with good results. This is evident from the results of the C4.5 algorithm generates a classification accuracy of 96.04% value with a precision of 95.96% class, class recall of 95.92% while the value of AUC (Area Under the Curve) of 0948 + / - 0.028 with very good category. It can be concluded that in order to predict the value of the majors C4.5 algorithm produces accuracy that is very good value.

Keywords: C4.5, accuracy, precision class, recall class, AUC

1. PENDAHULUAN

1.1. Latar Belakang

Kurikulum SMA pada tahun 2013 berubah dari Kurikulum Tingkat Satuan Pendidikan (KTSP) menjadi Kurikulum 2013 yang mengubah struktur kurikulum dan proses penjurusan. Pada saat menggunakan KTSP penjurusan dilakukan pada saat siswa akan naik ke kelas XI yaitu jurusan Ilmu Pengetahuan Alam (IPA), Ilmu Pengetahuan Sosial (IPS) dan Bahasa, sedangkan pada Kurikulum 2013 penjurusan dilakukan pada saat setelah proses Penerimaan Peserta didik Baru (PPDB) yaitu Jurusan Matematika dan Ilmu Alam (MIA), Ilmu-ilmu Sosial (IIS) dan Bahasa. Karena perbedaan waktu pelaksanaan penjurusan maka berubah pula instrumen yang digunakan untuk penjurusan yaitu menggunakan nilai Ujian Nasional (UN) Matematika dan IPA SMP, Nilai Tes Penjurusan dan hasil Tes *Intelligence Quotient* (IQ).

Perbedaan mendasar pada model penjurusan pada KTSP sudah pasti tergantung pada prestasi nilai mata pelajaran ilmu pasti sedangkan pada model penjurusan Kurikulum 2013 pada nilai UN Matematika dan IPA, Nilai Tes Penjurusan dan IQ sehingga dari pihak sekolah harus bisa mengklasifikasikan siswa pada jurusan yang sesuai dengan kemampuan dan minat. Diperlukan sebuah model klasifikasi untuk memprediksi hasil penjurusan pada sistem penjurusan Kurikulum 2013. Banyak algoritma untuk mengklasifikasi data diantaranya adalah algoritma *decision tree, naïve bayes, k-nearest neighbor, support vector machine* dan lain-lain.

Pada penelitian lain yang dihasilkan bahwa klasifikasi dengan menggunakan metode *nearest neighbor* tidak lebih akurat dari algoritma C4.5 tetapi proses klasifikasi membutuhkan waktu lama yang lebih banyak dan memerlukan proses yang lebih panjang [1]. Survey yang dilakukan Clifton Phua menyatakan

bahwa Algoritma C4.5 dapat membantu tidak hanya untuk membuat prediksi yang akurat, tetapi juga menjelaskan pola-pola di dalamnya. Ini berkaitan dengan atribut yang hilang, seleksi, memperkirakan tingkat kesalahan, kompleksitas induksi pohon keputusan. Dalam prediksi akurasi algoritma C4.5 lebih daik dari CART dan ID3 [2]. Pada pengujian hasil prediksi kelulusan dengan algoritma Naiva Bayes menghasilkan nilai akurasi sebesar 80,85% sedangkan pengujian dengan menggunakan algoritna C.45 menghasilkan nilai akurasi sebesar 85,70% yang merupakan nilai akurasi yang lebih baik dari algoritma *Naïve Bayes* [3]

Pada penelitian ini akan digunakan algoritma C4.5 untuk mengklasifikasi nilai penjurusan. Teknik klasifikasi Algoritma *decision tree C4.5* menggunakan pohon keputusan yang memiliki kelebihan seperti dapat mengolah data numerik (kontinyu) dan diskret serta dapat menangani nilai atribut yang hilang. Diperlukan sebuah model yang mempunyai kecepatan atau efesiensi waktu komputasi.

1.2. Rumusan Masalah

a. Perbedaan waktu proses penjurusan pada KTSP dan Kurikulum 2013 Sekolah Menengah Atas (SMA) membutuhkan klasifikasi hasil penjurusan yang lebih baik.
b. Model klasifikasi penjurusan siswa dari penelitian sebelumnya memiliki akurasi yang kurang optimal.

1.3. Tujuan

a. Secara umum tujuan penelitian ini adalah mendapatkan hasil klasifikasi hasil penjurusan yang lebih baik.
b. Tujuan Secara Spesifik adalah untuk meningkatkan akurasi model klasifikasi penjurusan siswa dengan tiga model kriteria *gain_ratio*, *information_gain* dan *gini_index*.

1.4. Manfaat

a. Manfaat Penelitian bagi mayarakat adalah membantu sekolah untuk memproses penjurusan siswa baru dengan tepat sesuai minat bakat siswa.
b. Manfaat Peneltian bagi IPTEKS adalah memberikan pengetahuan tentang kemampuan algoritma C4.5 dalam melakukan klasifikasi penjurusan siswa sekolah.

2. TINJAUAN PUSTAKA

2.1. Penelitian yang Relevan

Beberapa penelitian yang sudah pernah ada yang berhubungan dengan permasalahan penelitian yang menjadi rujukan penulis adalah sebagai berikut.

Tabel 1. Penelitian Terkait

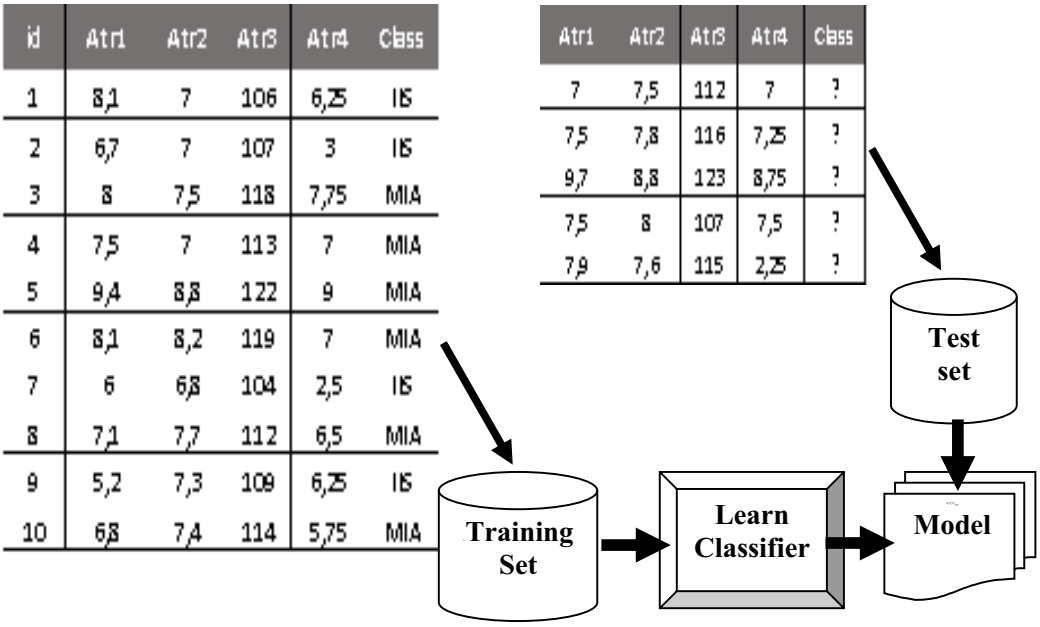
Peneliti	Judul	Masalah	Metode	Hasil
Kusrini, 2009 [4]	Perbandingan Metode Nearest Neighbor dan Algoritma C4.5 untuk Menganalisis Kemungkinan Pengunduran Diri Calon Mahasiswa di STMIK Amikom	Dari jumlah 1954 calon mahasisw yang diterima pada tahun 2006/2007 sebagai calon mahasiswa baru di STMIK AMIKOM Yogyakarta, 498 calon mahasiswa mengundurkan diri	*Nearest Neighbor* dan *C.45*	Metode *k-nearest neighbor*,proses *testing* memerlukan waktu yang lebih lama dibanding dengan menggunakan algoritma *C4.5*

	Yogyakarta	dengan cara tidak melakukan registrasi 25,5% calon mahasiswa yang mungkin potensial		karena setiap kasus baru akan dicocokan dengan semua kasus yang lama
Clifton Phua, Damminda Alahakoon dan Vincent Lee, 2005 [2]	*Minority report in Fraud Detection: Classification of Skewed Data*	Deteksi penipuan penelitian yang ada	*Backpropa-gatuon, naïve bayes dan C4.5*	Algoritma C4.5 menjadi algoritma yang terbaik dari algoritma *Backpropagation* (BP), dan algoritma naïve bayes
Marselina & Ernastuti, 2010 [3]	*Graduation prediction of gunawarma university students using algorithm and naïve bayes, C4.5 algortihm*	Bagaimana memprediksi tingkat kelulusan mahasiswa	*naïve bayes dan C4.5*	Nilai akurasi algoritma C4.5 sebesar 85,7% lebih tinggi dari nilai akurasi algoritma *nauve bayes* sebesar 80,85%
Budanis Dwi Meilani Achmad dan Fauzi Slamat,2012 [5]	Klasifikasi Data Karyawan untuk Menentukan Jadwal Kerja Menggunakan Metode *Decision Tree*	Penentuan Jadwal Kerja Karyawan	*ID3, C4.5* dan *CART*	Tingkat akurasi sebesar 87% pada algoritma C4.5
Sunjana, 2010 [6]	Klasifikasi Data Nasabah sebuah Asuransi Menggunakan Algoritma C4.5	Mengetahui lancar atau tidak lancarnya nasabah asuransi	*C4.5*	pola dapat digunakan untuk memperkirakan nasabah yang bergabung, sehingga perusahaan bisa mengambil keputusan menerima atau menolak calon nasabah
Ahmad Saikhu, Joko Lianto dan Umi Hanik, 2011 [7]	*Fuzzy Decision Tree* dengan algoritma C4.5 pada data diabetes Indian Pima	Deteksi diabetes pada masyakarat keturunan Indian Pima	*C4.5*	Nilai Akurasi sebesar 78,91%
Angga Raditya [8]	Implementasi *Data Mining Clasification* untuk mencari Pola Prediksi Hujan dengan menggunakan Algoritma C4.5	Mencari pola prediksi hujan dengan menggunakan data cuaca tahun 2007 sebagai *training* dan data cuaca tahun 2008 dan 2009 sebagai *testing*	C4.5	Menghasilkan tingkat akurasi di atas 70% dan hampir mendekati 80%

2.2. Landasan Teori

2.1.1 Klasifikasi

Model klasifikasi digunakan untuk pemodelan deskriptif sebagai perangkat penggambaran untuk membedakan objek-objek dari *class* yang berbeda dan pemodelan prediktif yang digunakan untuk memprediksi label *class* untuk *record* yang tidak diketahui atau tidak dikenal. Pada teknik klasifikasi merupakan pendekatan sistematis untuk membangun model klasifikasi dari sekumpulan data input. Metode yang digunakan dalam klasifikasi adalah *Decision Tree* (Pohon Keputusan), *Rule-Based* (Berbasis Aturan), *Neural Network* (Jaringan Syaraf), *Support Vector Machine* (SVM), *Naïve Bayes*, *K-Nearest Neighbor*.



Gambar 1. Contoh Model Klasifikasi

Untuk mengevaluasi performansi sebuah model yang dibangun oleh algoritma klasifikasi dapat dilakukan dengan menghitung jumlah dari *test record* yang diprediksi secara benar (akurasi) oleh model tersebut. Rumus akurasi adalah sebagai berikut:

$$Akurasi = \frac{\text{Jumlah prediksi benar}}{\text{Jumlah total prediksi}} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \qquad 2.1$$

2.1.2 *Data mining*

Data mining merupakan suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam database. *Data mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan dan *machine learning* untuk mengekstrasi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar [4]. *Data mining* adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual dari sekumpulan data [5]. *Data mining* merupakan analisis dari peninjauan kumpulan data untuk menemukan hubungan yang tidak diduga dan meringkas data dengan cara yang berbeda dengan

sebelumnya, yang dapat diphami dan bermanfaat bagi pemilik data [6]. Beberapa faktor yang mendorong kemajuan *data mining*antara lain [6]:

a. Pertumbuhan yang cepat dalam kumpulan data.
b. Penyimpanan data dalam *data warehouse*, sehingga seluruh perusahaan memiliki akses ke dalam database yang andal.
c. Adanya peningkatan akses data melalui navigasi web dan internet.
d. Tekanan kompetisi bisnis untuk meningkatkan penguasaan pasar dalam globalisasi ekonomi.
e. Perkembangan teknologi perangkat lunak untuk *data mining* (ketersediaan teknologi).

Perkembangan yang hebat dalam kemampuan komputasi dan perkembangan kapasitas media penyimpanan.

2.1.3 Pohon Keputusan (*decision tree*)

Pohon keputusan merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami. Pohon keputusan dapat mengeksplorasi data, menemukan hubungan tersembunyi antara calon variabel input dengan sebuah variabel target. Pohon keputusan juga memadukan antara eksplorasi data dan pemodelan dan sebuah pohon keputusan adalah sebuah struktur yang dapat digunakan untuk membagi kumpulan data yang besar menjadi himpunan-himpunan *record* yang lebih kecil dengan menerapkan serangkaian aturan keputusan. Dengan masing-masing rangkaian pembagian, anggota himpunan hasil menjadi mirip satu dengan yang lain [7].

2.1.4 Algoritma C.45

Algoritma C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan untuk memprediksi atau mengklasifikasi yang sangat kuat. Pohon keputusan dapat membagi kumpulan data yang besar menjadi himpunan-himpunan *record* yang lebih kecil dengan menerapkan serangkaian aturan keputusan. Dalam algoritma C4.5 untuk membangun pohon keputusan hal pertama yang dilakukan yaitu memilih atribut sebagai akar. Kemudian dibuat cabang untuk tiap-tiap nilai di dalam akar tersebut. Langkah berikutnya yaitu membagi kasus dalam cabang, kemudian 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 di bawah ini:

$$Gain\ (S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * \text{Entropy (Si)} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad 2.2$$

Keterangan :

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 ini:

$$Entropy(S) = \sum_{i=1}^{n} - \text{pi} * \log_2 \text{pi} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad 2.3$$

Keterangan :

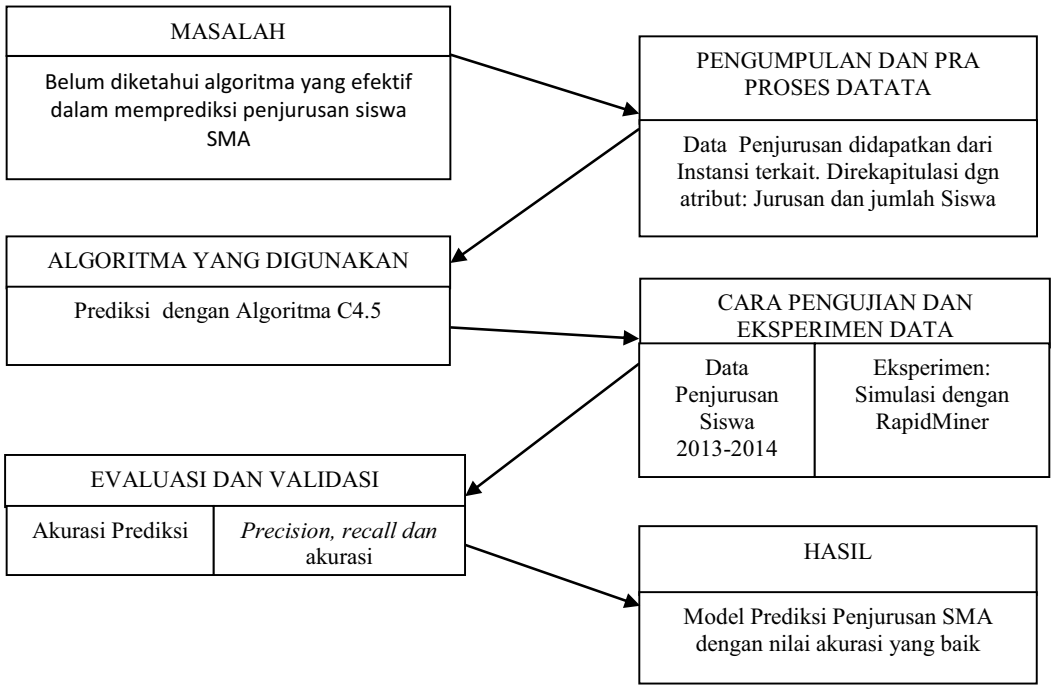
S : himpunan kasus
A : fitur
n : jumlah partisi S
Pi : proporsi dari S_i terhadap S

2.1.5 *Cross Validation*

Cross Validation menurut Payam Refaeizadeh, Lei Tang dan Huan Lu merupakan metode statistik untuk mengevaluasi dan membandingkan algoritma pembelajaran dengan membagi data menjadi dua segmen, satu digunakan untuk belajar atau melatih model dan yang lainnya digunakan untuk memvalidasi model. Dalam *cross validation* pelatihan dan validasi set harus menyeberang dalam putaran berturut-turut sehinga setiap titik data memiliki kesempatan yang divalidasi. Bentuk dasar *cross validation* adalah *k-fold cross validation*. Bentuk lain dari *cross validation* adalah kasus khusus dari *k-fold cross validation* atau melibatkan putaran berulang *k-fold cross validation*. *cross validation* digunakan untuk mengevaluasi atau membandingkan algoritma pembelajaran sebagai berikut: dalam setiap iterasi, satu atau lebih algoritma pembelajaran menggunakan k 1 lipatan data untuk mempelajari satu atau lebih model dan selanjutnya model belajar diminta untuk membuat prediksi tentang data di lipatan validasi. Kinerja setiap algoritma pembelajaran pada setiap flip dapat dilacak menggunakan beerapa metrik kinerja yang telah ditentukan seperti akurasi. Setelah selesai, k sampel dari metrik kinerja yang telah ditentukan untuk masing-masing algoritma. Metodologi yang berbeda seperti rata-rata dapat digunakan untuk mendapatkan ukuran agregat dari sampel tersebut atau sampel ini dapat digunakan dalam uji hipotesis statistik menunjukkan bahwa satu algoritma lebih unggul dari yag lain.

3. METODE PENELITIAN

Berangkat dari masalah, penelitian dilakukan dengan tahapan- tahapan seperti pada gambar berikut ini.



Gambar 2. Diagram Alir Metode Penelitian

3.1. Pengumpulan Data

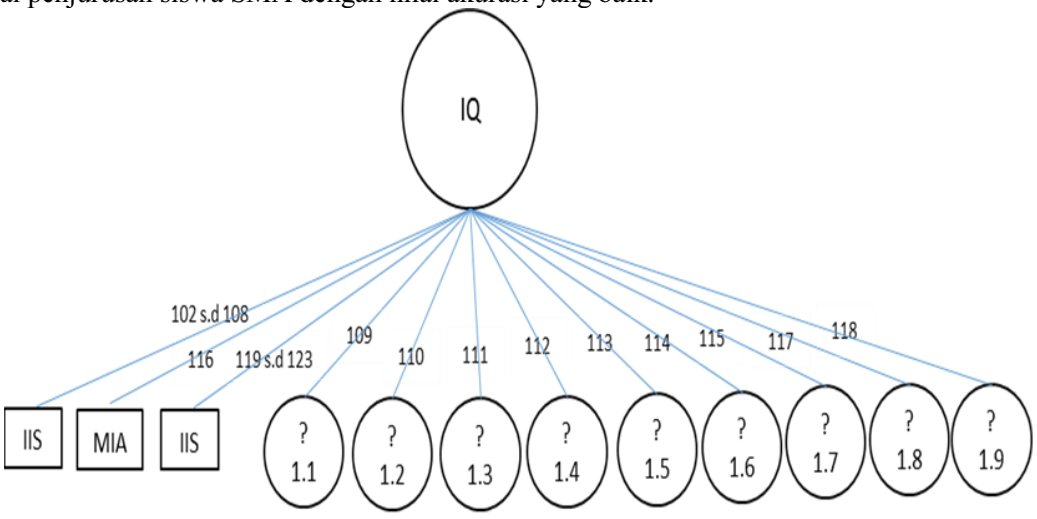
Data uji yang digunakan pada penelitian ini bersumber dari data privat hasil penjurusan siswa SMA Negeri 4 Tegal Tahun Pelajaran 2013/2014. Data penjurusan ini terdiri dari 227 siswa yang akan dibagi jurusan menjadi dua yaitu Jurusan Matematika dan Ilmu Pengetahuan Alam (MIA) serta Jurusan Ilmu-ilmu Sosial (IIS).

3.2. Praproses Data

Pada tahapan ini data diolah dengan cara pertama membuang duplikasi data dan memeriksa data yang memiliki fitur tidak lengkap. Pada tahapan ini memahami data sangat penting untuk mendapatkan data uji yang baik. Data yang ada pada umumnya banyak *noise*, berukuran besar dan dapat merupakan campuran dari berbagai sumber. Mengapa perlu praproses data karena data mentah yang ada sebagian besar tidak komplet dengan berisi data yang hilang/ kosong, kekurangan atribut yang sesuai dan hanya berisi data *aggregate*. Data yang banyak *noise* dengan berisi data yang *outlier*, berisi *error*. Data yang tidak konsisten dengan berisi nilai yang berbeda dalam suatu kode atau nama. Pada penelitian ini pada tahapan praproses data mentah yang akan diujikan hanya dihilangkan pada atribut nama yang tidak diikutkan dalam proses pengujian. data yang diolah akan langsung diklasifikasi menggunakan algoritma C4.5 untuk dihitung nilai akurasinya.

Hasil dari tahapan ini yaitu atribut-atribut data yang sudah bersih akan digunakan untuk tahapan klasifikasi. Atribut-atribut yang terpilih dan memenuhi syarat yang dapat digunakan dalam proses klasifikasi nantinya.

Algoritma yang bisa digunakan untuk klasifikasi diantaranya algoritma *Support Vector Machine (SVM)*, *Naïve Bayes, K-Nearest Neighbor, C4.*5 dan lain-lain. Pada penelitian ini digunakan algoritma C4.5 untuk mengklasifikasi data penjurusan siswa SMA untuk menghasilkan model untuk memprediksi hasil nilai penjurusan siswa SMA dengan nilai akurasi yang baik.



Gambar 3. Pembentukan Pohon Keputusan Prediksi Nilai Penjurusan Siswa SMA

Konsep kerja algoritma C4.5 dalam memprediksi hasil nilai penjurusan siswa SMA adalah dengan pembentukan pohon keputusan. Proses pohon keputusan adalah pertama dengan mencari jumlah data jurusan MIA dan IIS per *record* data dari masing-masing atribut yang ada pada data uji kemudian mencari entropi dan gain dari masing-masing *record* pada setiap atribut, kemudian cari gain tertinggi untuk

membentuk node utama dalam pembentukan pohon keputusan. Ulangi langkah-langkah tersebut sampai terbentuk pohon keputusan.

Dari data yang sudah dilakukan pengolahan awal akan dihasilkan data uji. Pada penelitian ini digunakan data uji berupa nilai penjurusan siswa SMA yang terdiri dari atribut nilai Ujian Nasional mata pelajaran Matematika dan IPA, Nilai Hasil Tes Penjurusan dan Nilai Hasil Tes IQ.

3.3. Eksperimen dan Pengujian Model

Algoritma C4.5 yang digunakan dalam penelitian ini akan diterapkan pada data nilai penjurusan siswa SMA melalui suatu model simulasi. Dari data tersebut akan diklasifikasikan menjadi data *training* dan data *testing*. Pada penelitian ini menggunakan metode *k-fold crooss validation*, data akan menjadi data *training* dan *testing* dengan memberikan nilai validasi, sebagai contoh nilai validasi 10 artinya dari semua data yang diujikan 90% sebagai data *training* dan 10% sebagai data *testing*. Pada penelitian ini akan diujikan dengan nilai validasi 3, 7, 10,20 30 dan 40 yang kemudian akan dicari nilai akurasi yang tertinggi.

Evaluasi yang akan dilakukan menggunakan parameter *f-measure* yang terdiri dari perhitungan precision dan *recall* yang merupakan metode pengukuran efektifitas yang biasa dilakukan pada proses klasifikasi. Pada penelitian ini nilai *precision* dan *recall* didasarkan pada hasil keluaran menggunakan *confusion matrix*.

Tabel 2. *Confusion Matrix*

		Predicated Class	
		True	**False**
Actual Class	Positif	TP	FP
	Negatif	FN	TN

Keterangan :
TP : True Positif
TN : True Negatif
FP : False Positif
FN : False Negatif

Tabel 2 merupakan tabel *confusion matrix* yang terdiri dari dua baris dan dua kolom yang berisi jumlah *True* Positif, *True* Negatif, *False* Positif dan *False* Negatif yang berfungsi untuk analisis secara rinci untuk menghasilkan prediksi dengan nilai akurasi yang baik. Dan apabila klasifikasi data tidak seimbang akan menghasilkan nilai akurasi yang buruk.

Recall dan *Precision* merupaka dua kriteria yang berfungsi untuk mengevaluasi efektifitas kinerja sistem.

a. *Precision* merupakan bagian data yang diambil secara acak sesuai informaasi yang dibutuhkan. Rumus *precision* :

$$Precison\ (p) = \frac{TP}{TP+FP} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots . \quad 3.1$$

b. *Recall* merupakan pengambilan data yang berhasil dilakukan terhadap bagian data yang relevan dengan *query*. Rumus *recall* :

$$Recall\ (r) = \frac{TP}{TP+FN} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.\ 3.2$$

c. *Akurasi* merupakan prosentase dari total jurusan IPA dan IPS yang bernilai benar diidentifikasi. Rumus *akurasi*:

$$Akurasi\ = \frac{TP+TN}{TP+FP+TN+FN} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..\ 3.3$$

Kurva ROC digunakan apabila respon diagnosis lebih dari sua jenis respon atau bilangan respon kontinu. Kurva ini menghubungkan nilai *sensitivitas* dengan 1-*specifisitas*. Area di bawah kurva ROC dapat digunakan untuk menilai keakuratan suatu diagnosis. Kurva ROC adalah alat visual yang berguna untuk membandingkan dua model klasifikasi. Kurva ROC menunjukkan *trade-off* antara *true positive rate* (proporsi *tuple* positif yang teridentifikasi dengan benar) dan *false positive rate* (proporsi tuple negatif yang teridentifikasi salah sebagai positif) dalam suatu model.

Dengan kurva ROC, kita dapat melihat *trade off* antara tingkat suatu model dapat mengenali *tuple* positif secara akurat dan tingkat model tersebut salah mengenali *tuple* negatif sebagai *tuple* positif. Kurva ROC terdiri atas sumbu vertikal yang menyatakan *true positive rate*, dan sumbu horizontal yang menyatakan *false positive rate*. Jika memiliki *true* positif (sebuah *tuple* positif yang benar diklasifikasikan) maka pada kurva ROC akan bergerak ke atas dan plot titik. Sebaliknya, jika *tuple* milik kelas “tidak” ketika memiliki false positif, maka kurva ROC bergerak ke kanan dan plot titik. Proses ini diulang untuk setiap *tuple* tes (setiap kali bergerak ke atas kurva untuk true positif atau terhadap hak untuk false positif). Untuk mengukur ketelitian dari suatu model, kita dapat mengukur area di bawah kurva ROC.

4. HASIL PENELITIAN DAN PEMBAHASAN

Dari uji coba yang dilakukan menghasilkan masing-masing kriteria menjadi tabel-tabel sebagai berikut.

Tabel 3. Hasil Eksperimen dengan Menggunakan *3 Fold Cross Validation*

Kriteria	***Precision Class***	***Recall Class***	***Akurasi***	***Nilai AUC***
Gain_ratio	92,93	91,09	92,95	0.924 +/- 0.056
Information_gain	91,92	86,67	90,31	0.913 +/- 0.013
Gini_index	92,07	89,32	92,07	0.948 +/- 0.028

Tabel 4. Hasil Eksperimen dengan Menggunakan *7 Fold Cross Validation*

Kriteria	***Precision Class***	***Recall Class***	***Akurasi***	***Nilai AUC***
Gain_ratio	95,96	94,06	95,59	0.936 +/- 0.053
Information_gain	95,96	90,48	93,83	0.928 +/- 0.060
Gini_index	94,95	94,00	95,15	0.939 +/- 0.054

Tabel 5. Hasil Eksperimen dengan Menggunakan *10 Fold Cross Validation*

Kriteria	***Precision Class***	***Recall Class***	***Akurasi***	***Nilai AUC***
Gain_ratio	94,95	92,16	94,27	0.892 +/- 0.142
Information_gain	93,94	92,08	93,83	0.921 +/- 0.057
Gini_index	92,93	92,93	93,83	0.891 +/- 0.142

Hasil Eksperimen dengan Menggunakan *20 Fold Cross Validation*

Kriteria	*Precision Class*	*Recall Class*	*Akurasi*	*Nilai AUC*
Gain_ratio	94,95	94,95	95,59	0.825 +/- 0.221
Information_gain	93,94	93,00	94,27	0.938 +/- 0.092
Gini_index	92,93	93,88	94,27	0.785 +/- 0.223

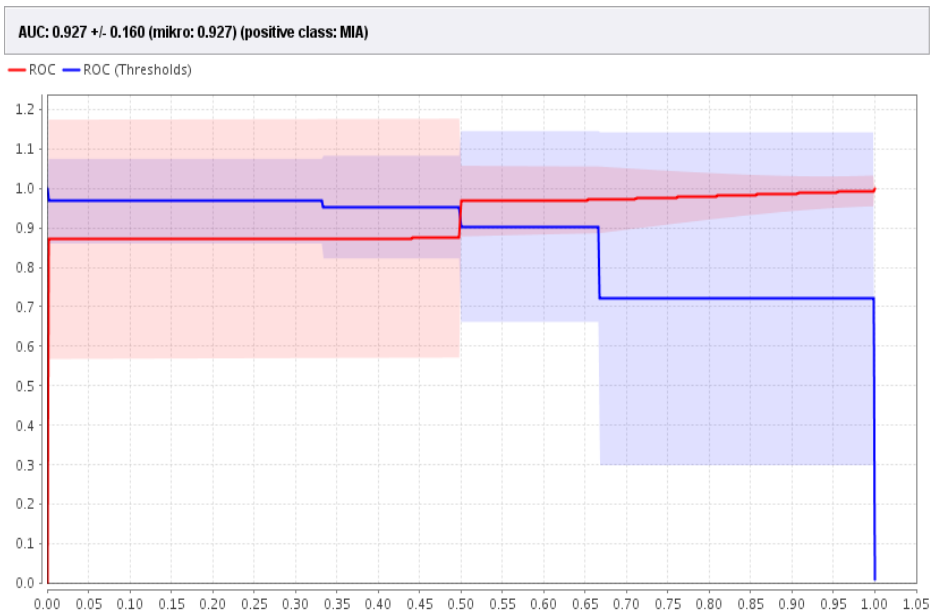
Tabel 4.5 Hasil Eksperimen dengan Menggunakan *30 Fold Cross Validation*

Kriteria	*Precision Class*	*Recall Class*	*Akurasi*	*Nilai AUC*
Gain_ratio	94,95	93,07	94,71	0.826 +/- 0.224
Information_gain	93,94	91,18	93,39	0.921 +/- 0.142
Gini_index	92,93	91,09	92,95	0.802 +/- 0.230

Tabel 4.6 Hasil Eksperimen dengan Menggunakan *40 Fold Cross Validation*

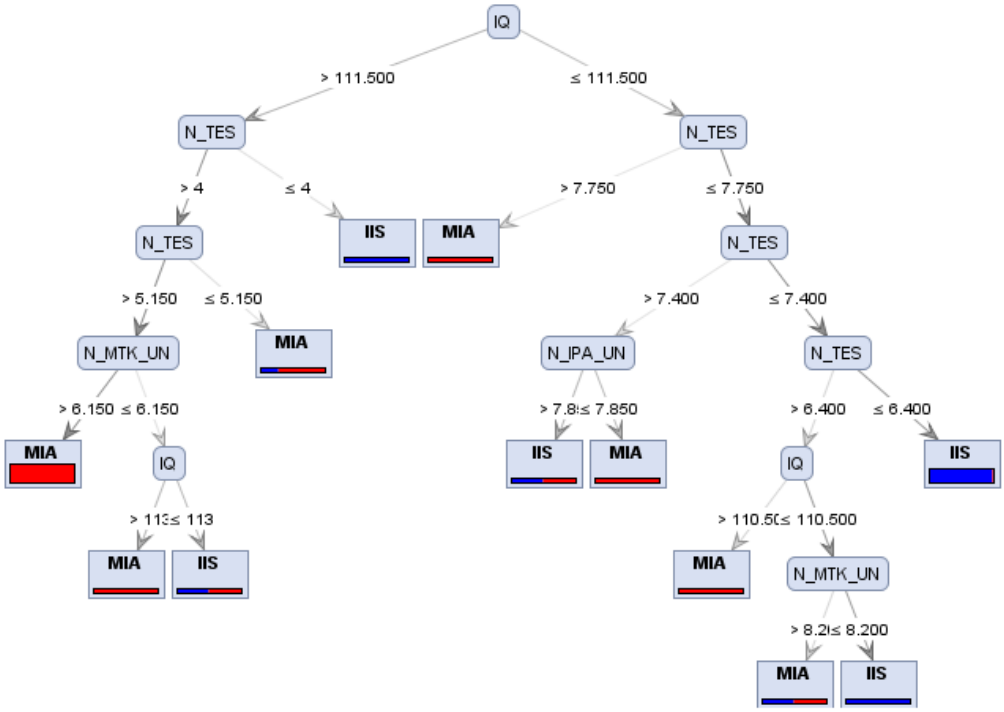
Kriteria	*Precision Class*	*Recall Class*	*Akurasi*	*Nilai AUC*
Gain_ratio	94,95	95,92	96,04	0.731 +/- 0.246
Information_gain	93,94	93,00	94,27	0.927 +/- 0.160
Gini_index	92,93	93,88	94,27	0.731 +/- 0.246

Dari rangkuman tabel didapatkan nilai tertinggi dari *precision class* sebesar 95,96% menggunakan *7 fold cross validation* dengan kriteria *gain_ratio* dan *information_gain*. Nilai tertinggi dari *recall class* sebesar 95,92% menggunakan *40 fold cross validation* dengan kriteria *gain_ratio*. Nilai tertinggi *akurasi* sebesar 96,04% menggunakan *40 fold cross validation* dengan kriteria *gain_ratio*. Nilai tertinggi dari *AUC* sebesar 0.948 +/- 0.028 menggunakan *3 fold cross validation* dengan kriteria *gini_index*.



Gambar 4. Kurva ROC dengan Kriteria *Information Gain* pada *40 Fold Cross Validation*

Berdasarkan kurva ROC di atas didapatkan informasi bahwa nilai *AUC* kriteria *information gain* sebesar 0.927 +/- 0.160 dengan tingkat klasifikasi sangat baik, sedangkan nilai *AUC* kriteria *gain_ratio* sebesar 0.731 +/- 0.246 dan kriteria *gini_index* sebesar 0.731 +/- 0.246 dengan tingkat klasifikasi cukup baik.



Gambar 5. Pohon Keputusan (*Decision* Tree) dengan Kriteria *Information Gain* pada *40 Fold Cross Validation*

5. KESIMPULAN

Dari hasil pengujian sampai dengan tahap evaluasi dihasilkan kesimpulan bahwa algoritma C4.5 mendapatkan nilai akurasi tertinggi dari penggunaan tiga model kriteria *gain_ratio*, *information_gain* dan *gini_index*, menghasilkan akurasi tertinggi pada model kriteria *gain_ratio* sebesar 96,04%. Rata-rata Nilai *AUC* di antara 0.80 – 0.90 dengan klasifikasi Baik.

UCAPAN TERIMAKASIH

Pada kesempatan ini perkenankan penulis menyampaikan ucapan terima kasih yang setulus-tulusnya kepada : Bapak Dr. Ir. Edi Noersasongko, M.Kom, Bapak Dr. Abdul Syukur, MM, Bapak Purwanto, SSi., M.Kom, Ph.D dan H. Himawan, M.kom atas bantuannya dalam penyusunan tesis ini.

PERNYATAAN ORIGINALITAS

"Saya menyatakan dan bertanggung jawab dengan sebenarnya bahwa artikel ini adalah hasil karya saya sendiri kecuali cuplikan dan ringkasan yang masing-masing telah saya jelaskan sumbernya. **[Imam Sujai – P31.2011.01059]**

DAFTAR PUSTAKA

[1] Pasal 18 Undang-Undang Republik Indonesia No. 20. *Sistem Pendidikan Nasional*. Indonesia:Presiden Republik Indonesia.

[2] Clifton Phua, Damminda Alahakoon, and Vincent Lee. *Minority Report in Fraud Detection: Classification 0f Skewed Data*. School of Business Systems, Faculty of Information Techlology Monash University, Clayton 3800, Australia.

[3] Marselina Silvia Suhartinah, Ernastuti. 2010, *Graduation Prediction of Gunadarma University Students Using Algorith and Naïve Bayes C4.5 Algorithm*. Undergraduate Program, Faculty of Industrial Engineering, Gunadarma University, Jakarta.

[4] Turban, E. 2005. *Decision Support System and Intelligent Systems*. Yogyakarta: Andi Offset.

[5] Iko Pramudiono, I. 2003. *Pengantar Data Mining:Menambang Permata Pengetahuan di Gunung Data*. Copyright@2003 Ilmu Komputer.com

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

[7] Berry, Michael J.A. dan Gordon S. Linoff. 2004. *Data Mining Techniques for Marketing, Sales, Customer Relationship Management.* Second Edition. Wiley Publishing, Inc.