PENERAPAN ALGORITMA C4.5 DALAM PEMILIHAN PROGRAM STUDI FAKULTAS ILMU KOMPUTER (Studi Kasus Sekolah Menengah Atas Negeri 1 Tambun Utara)

Ultach Enri
Fakultas Ilmu Komputer Program Studi Teknik Informatika
Universitas Singaperbangsa Karawang (UNSIKA)
Jl. HS. Ronggo Waluyo, Puseurjaya, Teluk Jambe
Karawang, Jawa Barat 41361
http://www.unsika.ac.id
ultach@staff.unsika.ac.id

Naskah diterima 15 Maret 2018

ABSTRACT

The right decision for course selection after completing high school is one of the most important thing, because it will affect their future. This research aim to find and determine the course in computer science faculty, based on their grade in each variable. The variable will be used to get the model in decision tree form using C4.5 algorithm. The C4.5 algorithm is one of the classification method in data mining.

Key Word : *Course Selection, C4.5, Classification*

ABSTRAK

Pemilihan Program Studi yang tepat bagi siswa-siswi setelah menempuh pendidikan sekolah menengah atas adalah salah satu hal yang penting dikarenakan akan mempengaruhi masa depan siswa. Penelitian ini dilakukan untuk mengetahui dan menentukan pemilihan Program Studi pada fakultas ilmu komputer yang dapat dipilih oleh siswa setelah lulus, sesuai dengan nilai-nilai masing-masing variabel yang digunakan untuk mendapatkan model berupa decision tree dengan menggunakan algoritma C4.5. Dimana algoritma C4.5 adalah salah satu algoritma dalam metode klasifikasi dalam data mining.

Kata Kunci : Pemilihan Program Studi, C4.5, Klasifikasi

I. PENDAHULUAN

Pentingnya pengambilan Program Studi dalam melanjutkan pendidikan perguruan tinggi setelah lulus, apabila salah dalam upaya pengambilan Program Studi, maka akan membawa dampak negatif terhadap siswa seperti prestasi yang tidak optimal dan merasa tidak mampu menguasai materi perkuliahan sehingga nilai hasil studi akan tidak memuaskan dan siswa tersebut tidak akan mendapatkan ilmu yang bermanfaat.

SMAN 1 Tambun Utara yang berdiri sejak tahun 1996 merupakan salah satu Sekolah Menengah Atas di Kabupaten Bekasi yang terletak di Jalan Raya Sriamur, Tambun Utara, Kabupaten Bekasi. Dimana Setiap tahunnya meluluskan lebih dari 200 siswa yang terdiri dari jurusan IPA dan IPS. Sebanyak 50 % dari lulusan tersebut melanjutkan pendidikan ker perguruan tinggi. Masih banyak siswa yang belum bisa menentukan Program Studi yang akan dipilih setelah lulus terutama di Program Studi yang bernaung di bawah fakultas ilmu komputer.

Algoritma C4.5 diperkenalkan oleh Quinlan pada tahun 1996 sebagai versi perbaikan dari ID3 (Iterative Dichotomiser), dimana di dalam ID3, induksi *decision tree* hanya bisa dilakukan pada fitur bertipe kategorikal (nominal atau ordinal), sedangkan tipe numerik tidak dapat digunakan.Perbaikan yang membedakan antara algoritma C4.5 dan ID3 adalah selain dapat menangani tipe data numerik, dapat juga melakukan pemotongan(*pruning*) *decision tree*, dan penurunan rule set. (Prasetyo, 2014). Didalam algoritma C4.5 ini, pohon-pohon keputusan yang dibentuk berdasarkan kriteria-kriteria pembentuk keputusan (Nofriansyah, 2014).

Algoritma C4.5 di harapkan mampu untuk memberikan solusi dalam penentuan Program Studi bagi siswa tersebut

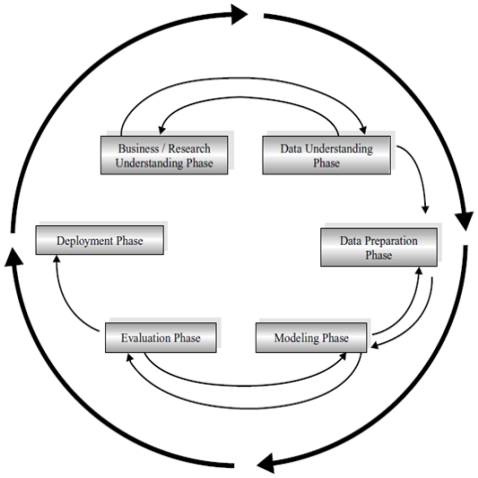
II. TINJAUAN PUSTAKA

2.1 Data Mining

Data mining sering juga di sebut *knowledge discovery in database*(KDD), adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar (Santosa, 2007), yang disimpan di dalam repositori, menggunakan teknologi pengenalan pola dan juga teknik statistik dan matematika (Larose, 2005), Dengan kata lain data mining mempunyai kegiatan utama yaitu mengumpulkan, menemukan, dan menggali atau menambang pengetahuan dari data.

Fungsi-fungsi yang terdapat dalam data mining adalah (Larose, 2005) : Fungsi deskripsi (*Description*), Fungsi estimasi (*Estimation*), Fungsi prediksi (*Prediction*), Fungsi klasifikasi (*Classification*), Fungsi pengelompokan (*Clustering*), Fungsi asosiasi (*Associaton*).

Cross-Industry Standard Process for Data Mining (CRISP-DM) dkembangkan tahun 1996 oleh analis yang mewakili DaimlerChrysler, SPSS dan NCR. CRISP-DM menyediakan kepemilikan dan tersedia proses standar yang bebas untuk data mining yang sesuai sebagai strategi pemecahan masalah secara umum dari bisnis atau unit penelitian. Dalam CRISP-DM, sebuah proyek *data mining* memiliki siklus hidup yang terbagi dalam enam fase (Gambar 2.1).



Gambar 2.1 Proses Data Mining menurut CRISP-DM (Larose, 2005)

2.2 Algoritma C4.5

Pohon keputusan merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal, salah satu algoritma yang paling terkenal adalah C4.5 yang dikembangkan oleh Quinlan pada tahun 1996 sebagai perbaikan dari ID3, dimana algoritma C4.5 ini menggunakan konsep *information gain* atau *entropy reduction* untuk memilih pembagian yang optimal(Han & Kamber, 2006).

Ada tiga kelompok penting yang menjadi syarat pengujian pada node (Prasetyo, 2014), yaitu:

1. Fitur biner, yaitu fitur yang hanya mempunyai dua nilai berbeda.
2. Fitur bertipe kategorikal, yaitu fitur yang nilainya bertipe kategorikal (nominal atau ordinal) bisa mempunyai beberapa nilai berbeda.
3. Fitur bertipe numerik, yaitu fitur bertipe numerik dengan syarat pengujian dalam node (akar maupun internal) dinyatakan dengan pengujian perbandingan.

Secara umum tahapan algoritma C4.5 untuk membangun pohon keputusan (Kusrini & Luthfi, 2009) adalah:

1. Pilih atribut sebagai akar, dengan cara menghitung nilai gain dari masing-masing atribut, nilai *gain* yang tertinggi akan menjadi akar. Sebelum menghitung nilai *gain* kita harus menghitung terlebih dahulu nilai *entropy*.

Entropy(S) $\sum_{i=1}^{n} - pi.log_2 Pi$ =
Dimana:
S = Himpunan kasus
n = Jumlah partisi S
pi = proporsi S_i terhadap S

Setelah itu hitung nilai gain, dengan rumus sebagai berikut:
Gain(S,A)
= *Entropy* (S) $- \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy\ (Si)$

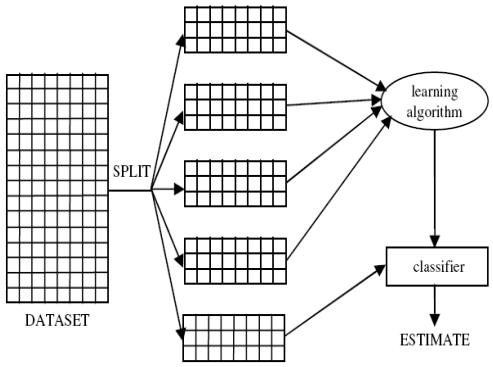
Dimana :
S = Himpunan kasus
A = fitur
n = jumlah partisi atribut A
$|S_i|$ = proporsi S_i terhadap S
$|S|$ = jumlah kasus dalam S

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.

2.3 Pengujian K-Fold Cross Validation

Salah satu pendekatan alternatif untuk "*train* dan *test*" yang sering di adopsi dalam beberapa kasus (dan beberapa lainnya terlepas dari ukurannya) yang di sebut dengan *k-fold cross validation* (Bramer, 2007), dengan cara menguji besarnya error pada data test(Santosa, 2007). Kita gunakan k-1 sampel untuk training dan 1 sampel sisanya untuk testing. Misalnya ada 10 subset data, kita menggunakan 9 subset untuk training dan 1 subset sisanya untuk testing. Ada 10 kali training dimana pada masing-masing training ada 9 subset data untuk training dan 1 subset digunakan untuk testing. Dari situ lalu di hitung rata-rata error dan standar deviasi error (Santosa, 2007). Setiap bagian k pada gilirannya digunakan sebagai ujian menetapkan dan k lainnya - 1 bagian

digunakan sebagai training set (Bramer, 2007), atau disebut juga dengan *leave-one-out*. Pendekatan ini mempunyai keuntungan penggunaan sebanyak mungkin data sebagai set data latih sehingga data latih yang digunakan hampir seluruh data dalam set data, sedangkan kelemahan dari pendekatan ini yaitu komputasi yang mahal untuk mengulang prosedur sebanyak N kali. (Prasetyo, 2014).



Gambar 2.2. ***K-fold Cross-validation*** **(Bramer, 2007)**

2.5 Evaluasi

Penerapan teknik klasifikasi diharapkan dapat memberikan hasil yang benar, akan tetapi tidak dapat dipungkiri bahwa kinerja sistem tidak bisa bekerja 100% benar. Oleh karena itu sebuah sistem klasifikasi harus diukur kinerjanya, dimana umumnya yang digunakan adalah matriks confusion, yang merupakan tabel yang mencatat hasil kerja klasifikasi. (Prasetyo, 2014)

1. Akurasi

Akurasi adalah presentase ketepatan record data yang diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi(Han & Kamber, 2006). Untuk mengukur akurasi digunakan formula sebagai berikut:

$$Akurasi = \frac{TP + TN}{P}$$

Dimana TP merupakan jumlah record positif yang dilabelkan secara benar oleh algoritma klasifikasi, sedangkan TN adalah jumlah record negatif yang dilabelkan secara salah oleh model algoritma klasifikasi dan P adalah total semua record yang dievaluasi Semua algoritma klasifikasi berusaha untuk membentuk model yang mempunyai akurasi yang tinggi.

2. Kappa

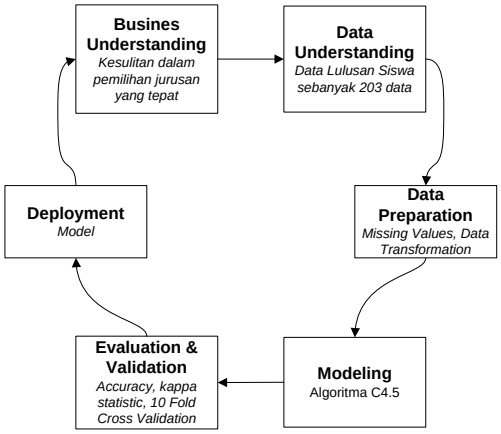
Kappa menunjukan analisis diantara kelas-kelas yang berbeda, semakin tinggi nilai kappa, maka akan dipertimbangkan sebagai performa yang memiliki kinerja yang bagus(Mittal & Gill, 2014).

Klasifikasi nilai kappa (Altman, 1991) diperlihatkan pada tabel 2.1

Tabel 2.1 Klasifikasi nilai Kappa

Nilai *K*	Strength of Agreement
< 0.20	Rendah (Poor)
0.21 – 0.40	Lumayan (Fair)
0.41 – 0.60	Cukup (Moderate)
0.61 – 0.80	Kuat (Good)
0.81 – 1.00	Sangat Kuat (Very Good)

III. METODOLOGI PENELITIAN



Gambar 3.1 Kerangka Pemikiran

Pada gambar 3.1, penelitian ini bertujuan untuk meneliti dengan

menggunakan Algoritma C4.5 yang bertujuan untuk memberikan klasifikasi dalam memprediksi Program Studi mana yang cocok untuk siswa yang telah lulus dari sekolah menengah atas untuk kasus ini berdasarkan data lulusan dari Sekolah Menengah Atas Negeri (SMAN) 1 Tambun Utara.

Data yang diperoleh sebanyak 203 data lulusan. Dan dalam tahapan preprocessing akan dilakukan prosedur missing values dalam menangani data yang kosong dan juga data transformation untuk merubah data-data yang berupa numeric menjadi nominal. Setelah itu akan diterapkan algoritma C4.5 untuk mendapatkan model, dan akan di validasi dengan 10 *fold cross validation* dan juga diukur tingkat *accuracy* dan nilai kappa nya.

3.1. Analisa Kebutuhan

Jenis penelitian yang digunakan dalam penelitian ini adalah model eksperimen. Penelitian eksperimen ini menggunakan algoritma C4.5 untuk memprediksi Program Studi Program

Studi terutama untuk Program Studi yang berada di Fakultas Ilmu Komputer. Jenis data yang digunakan dalam penelitian ini adalah data primer yang diperoleh penulis dari data alumni SMAN 1 Tambun Utara, sedangkan *software* yang digunakan sebagai alat bantu dalam penelitian ini adalah *Rapidminer* 5.3 dan *Microsoft Excel* 2007.

3.2. Desain Penelitian

Pada penelitian ini akan digunakan algoritma *C4.5*, berikut adalah langkah-langkah yang digunakan : Pengumpulan Data, Pengolahan Data (Cleaning Data(Missing values, Data Transformation), Metode yang digunakan, Eksperimen dan Pengujian Model, Evaluasi. Data yang digunakan dalam penelitian ini adalah data lulusan siswa SMAN 1 Tambun Utara selama 1 tahun, yaitu dari 2015 sampai dengan 2016, terdapat sebanyak 203 *records*. Pembagian data untuk penelitian ini dibagi menjadi dua bagian, yaitu data *training* dan data *testing*. Data yang akan digunakan untuk *training* adalah 90% dari data yang sudah dikumpulkan dan 10% sebagai *testing*. Pada penelitian ini metode yang akan diusulkan adalah dengan menggunakan C4.5.

Setelah dilakukan proses *cleaning data* dan *Transformation*, dan juga pemilihan Program Studi yang hanya berada di fakultas ilmu komputer, seperti terlihat pada tabel 3.1.

Tabel 3.1 Data Setelah Preprocesing

Nama Siswa	B. Indonesia	B. Inggris	Matematika	Fisika	Kimia	Biologi	Jurusan	Hasil
X1	BAIK	CUKUP	KURANG	KURANG	CUKUP	KURANG	IPA	Teknik Komputer
X2	CUKUP	CUKUP	KURANG	BAIK	KURANG	KURANG	IPA	Teknik Informatika
X3	CUKUP	CUKUP	KURANG	CUKUP	CUKUP	BAIK	IPA	Teknik Komputer
X4	BAIK	BAIK	BAIK	BAIK SEKALI	BAIK	BAIK	IPA	Teknik Informatika
X5	BAIK	CUKUP	BAIK	BAIK SEKALI	BAIK	BAIK	IPA	Teknik Komputer
X6	BAIK	BAIK	BAIK	BAIK	BAIK	CUKUP	IPA	Teknik Informatika
X7	BAIK	CUKUP	CUKUP	BAIK SEKALI	KURANG	BAIK SEKALI	IPA	Ilmu Teknologi
X8	BAIK	BAIK	BAIK	BAIK SEKALI	BAIK	BAIK	IPA	Sistem Informasi
X9	BAIK	BAIK	CUKUP	BAIK	KURANG	BAIK	IPA	Informatika
X10	BAIK	BAIK SEKALI	CUKUP	CUKUP	CUKUP	KURANG	IPA	Sistem Komputer
X11	BAIK	BAIK SEKALI	BAIK SEKALI	BAIK	BAIK SEKALI	BAIK SEKALI	IPA	Teknik Informatika
X12	BAIK SEKALI	BAIK SEKALI	BAIK SEKALI	BAIK	BAIK SEKALI	BAIK	IPA	Manajemen Teknik Informatika
X13	BAIK SEKALI	BAIK SEKALI	BAIK SEKALI	BAIK	BAIK SEKALI	BAIK	IPA	Teknik Informatika
X14	BAIK SEKALI	BAIK SEKALI	BAIK SEKALI	BAIK SEKALI	BAIK SEKALI	BAIK SEKALI	IPA	Komputer Akuntansi
X15	BAIK SEKALI	BAIK SEKALI	BAIK SEKALI	BAIK SEKALI	BAIK SEKALI	BAIK SEKALI	IPA	Teknik Informatika
X16	BAIK SEKALI	BAIK SEKALI	BAIK SEKALI	BAIK SEKALI	BAIK SEKALI	BAIK	IPA	Teknik Komputer
X17	CUKUP	BAIK	CUKUP	CUKUP	CUKUP	CUKUP	IPS	Teknik Informatika
X18	BAIK	BAIK	BAIK	BAIK	BAIK	BAIK	IPS	Ilmu Teknologi

Selanjutnya hasil dari proses preprocessing tersebut akan di aplikasikan dengan menggunakan algoritma C4.5 selanjutnya hasil yang didapat akan divalidasi dengan *10 folds x-validation*, untuk mendapatkan *Nilai Accuracy* dan nilai *Kappa*.

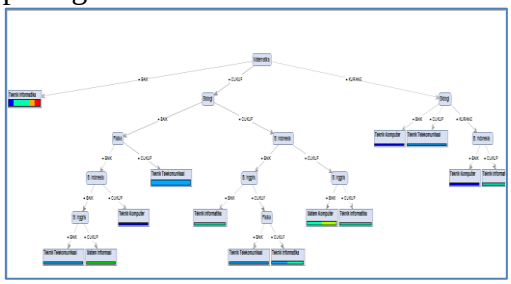
Atribut yang digunakan untuk modeling dengan menggunakan algoritma C4.5 adalah nilai Bahasa Indonesia, nilai Bahasa Inggris,nilai Matematika, nilai Fisika, nilai Kimia, nilai Biologi, jurusan serta hasil program studi sebagai label. Akar diambil dari atribut yang terpilih pada proses data *preparation*, dengan cara menghitung nilai *entropy* dan *gain* dari masing-masing atribut, nilai *gain* yang paling tinggi yang akan menjadi akar pertama.

Setelah diperoleh model dari hasil *training*, maka akan dilakukan evaluasi terhadap model tersebut. Proses evaluasi akan dilakukan dengan menggunakan *cross validation*, dengan menguji model

yang terbentuk dengan data secara acak yang dipisahkan dengan mengunakan 10 *folds cross validation*. Model yang sudah diperoleh akan diukur nilai accuracy dan juga nilai Kappa untuk menguji model yang terbentuk dengan algoritma C4.5.

IV. HASIL PENELITIAN DAN PEMBAHASAN

Setelah dikakukan dengan menggunakan algoritma C4.5, maka akan di dapat model berupa pohon keputusan(Decision Tree), yang tambak pada gambar 4.1.



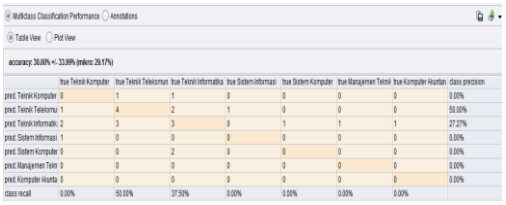
Gambar 4.1 Pohon Keputusan Prediksi Program Studi

Data Set Testing, hasil dari testing dengan menggunakan modeling yang diperoleh dari data training tampak pada tabel 4.1

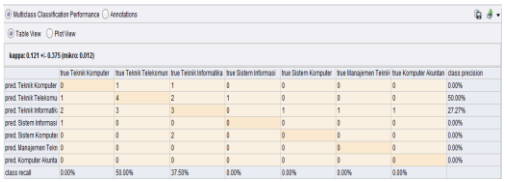
Tabel 4.1 Hasil Data Testing

Hasil	Prediction	B indonesia	B Inggris	Matematika	Fisika	Kimia	Biologi	Jurusan
Teknik Komputer	Teknik Komputer	BAIK	CUKUP	KURANG	KURANG	CUKUP	KURANG	IPA
Teknik Telekomunikasi	Teknik Telekomunikasi	CUKUP	CUKUP	KURANG	CUKUP	CUKUP	CUKUP	IPA
Teknik Informatika	Teknik Informatika	CUKUP	CUKUP	KURANG	CUKUP	CUKUP	KURANG	IPA
Teknik Komputer	Teknik Komputer	CUKUP	CUKUP	KURANG	CUKUP	CUKUP	BAIK	IPA
Teknik Telekomunikasi	Teknik Telekomunikasi	BAIK	BAIK	CUKUP	BAIK	CUKUP	BAIK	IPA
Teknik Informatika	Teknik Informatika	BAIK	BAIK	CUKUP	BAIK	BAIK	CUKUP	IPA
Teknik Telekomunikasi	Teknik Telekomunikasi	BAIK	CUKUP	CUKUP	BAIK	BAIK	CUKUP	IPA
Teknik Telekomunikasi	Teknik Telekomunikasi	CUKUP	CUKUP	CUKUP	CUKUP	BAIK	BAIK	IPA
Teknik Komputer	Teknik Komputer	CUKUP	CUKUP	CUKUP	BAIK	BAIK	BAIK	IPA
Teknik Informatika	Teknik Informatika	CUKUP	CUKUP	CUKUP	BAIK	CUKUP	CUKUP	IPA
Teknik Telekomunikasi	Teknik Telekomunikasi	BAIK	CUKUP	CUKUP	CUKUP	CUKUP	BAIK	IPA
Sistem Informasi	Sistem Informasi	BAIK	CUKUP	CUKUP	BAIK	BAIK	BAIK	IPA
Teknik Telekomunikasi	Teknik Telekomunikasi	CUKUP	KURANG	CUKUP	CUKUP	BAIK	BAIK	IPA
Teknik Informatika	Teknik Informatika	BAIK	CUKUP	CUKUP	CUKUP	CUKUP	CUKUP	IPA
Teknik Telekomunikasi	Teknik Informatika	BAIK	CUKUP	CUKUP	CUKUP	CUKUP	CUKUP	IPA
Sistem Komputer	Sistem Komputer	CUKUP	BAIK	CUKUP	CUKUP	CUKUP	CUKUP	IPA
Teknik Telekomunikasi	Teknik Telekomunikasi	BAIK	BAIK	CUKUP	CUKUP	BAIK	BAIK	IPA
Teknik Informatika	Teknik Informatika	BAIK	BAIK	BAIK	BAIK	BAIK	BAIK	IPA
Manajemen Teknik Informatika	Teknik Informatika	BAIK	BAIK	BAIK	BAIK	BAIK	BAIK	IPA
Teknik Informatika	Teknik Informatika	BAIK	BAIK	BAIK	BAIK	BAIK	BAIK	IPA
Komputer Akuntansi	Teknik Informatika	BAIK	BAIK	BAIK	BAIK	BAIK	BAIK	IPA
Teknik Informatika	Teknik Informatika	BAIK	BAIK	BAIK	BAIK	BAIK	BAIK	IPA
Teknik Komputer	Teknik Informatika	BAIK	BAIK	BAIK	BAIK	BAIK	BAIK	IPA
Teknik Informatika	Sistem Komputer	CUKUP	BAIK	CUKUP	CUKUP	CUKUP	CUKUP	IPS

Dari hasil testing modeling yang diperoleh dari data training yang di olah dengan menggunakan algoritma C4.5 dapat terlihat hasil prediksi memberikan hasil yang hampir akurat, akan tetapi kita tetap harus melakukan evaluasi terhadap model yang di hasilkan.



Gambar 4.2 Nilai Accuracy



Gambar 4.3 Nilai Kappa Statistic

Dari nilai accuracy yang di dapat sebesar 30% dan nilai kappa 0.121 yang termasuk ke dalam kategori rendah. Hal ini dimungkinkan karena data yang digunakan lebih sedikit tidak sebanding dengan jumlah program studi yang di gunakan.

V. KESIMPULAN

Dari hasil penelitian yang dilakukan dari tahap awal hingga pengujian penerapan C4.5 untuk memprediksi program studi, telah didapatkan jawaban dari pertanyaan-pertanyaan penelitian yang telah diidentifikasi sebelumnya, yaitu pada penelitian ini secara umum penerapan algoritma C4.5 dapat memberikan model yang dapat dipergunakan untuk memprediksi program studi pada perguruan tinggi yang akan ditempuh oleh siswa, akan tetapi karena keterbatasan penelitian ini perlu disarankan untuk melakukan penelitian selanjutnya untuk mendapatkan akurasi yang lebih baik. Adapun saran-saran yang dapat diberikan, yaitu :

a. Penggunaan data yang lebih besar dan pembatasan pengklasifikasian program studi.
b. Penggunaan attribut selection.
c. Penggunaan algoritma yang berbeda untuk penelitian sejenis.
d. Penelitian ini dapat dikembangkan dengan menggunakan metode

optimasi, seperti *Particle Swarm Optimization* (PSO), *Genetic Algorithm* (GA), *Simulated Annealing* (SA), *Ant ColonyOptimization* (ACO) serta *Artificial Bee Colony Algorithm* (ABC).

VI. DAFTAR PUSTAKA

[1] Altman, D. G. (1991). Practical Statistics for Medical Research. London : Chapman & Hall/CRC.

[2] Bramer, M., (2007). Principles of Data Mining. London : Springer.

[3] Han, J. & Kamber, M. (2006). Data Mining Concept and Techniques. San Fransisco : Elsevier.

[4] Kecman, V., (2001). Learning and Soft Computing. The MIT Press Cambridge, Massachusetts London, England.

[5] Kusrini., & Luthfi, E. T. (2009). Algoritma Data Mining. Yogyakarta : Andi Offset.

[6] Larose, D. T., (2005). " Discovering Knowledge in Data," . Canada : Wiley Interscience .

[7] Mittal, P., & Gill, N. S. (2014) A Comparative Analysis of Classification Techniques on Medical Data Sets. IJRET : International Journal of Research in Engineering and Technology, Volume : 03 Number : 06, pp.454-460.

[8] Nofriansyah, D. (2014). Konsep Data Mining vs Sistem Pendukung Keputusan. Yogyakarta : Deepublish.

[9] Prasetyo, E., (2014). Data Mining Mengolah Data Menjadi Informasi dengan menggunakan Matlab. Yogyakarta : Andi Offset.

[10] Santosa, B., (2007). Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis. Yogyakarta : Graha Ilmu.

[11] Santosa, B., (2007). Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis. Yogyakarta : Graha Ilmu.