KAJIAN PERBANDINGAN TEKNIK KLASIFIKASI ALGORITMA C4.5, *NAÏVE BAYES* DAN *CART* UNTUK PREDIKSI KELULUSAN MAHASISWA (STUDI KASUS : STMIK ROSMA KARAWANG)

Priati

Sekolah Tinggi Manajemen Informatika dan Komputer LIKMI

Jl. Ir. H. Juanda 96 Bandung 40132

tie.assiroj@gmail.com

Abstrak

Beberapa penelitian tentang kelulusan mahasiswa telah banyak dilakukan. Dalam penelitian ini dilakukan komparasi algoritma C4.5, *naïve bayes*, dan *CART* yang diaplikasikan terhadap data mahasiswa STMIK Rosma tahun 2005 sampai 2008 untuk jenjang Diploma III dan Strata 1.

Hasil pengujian dengan mengukur kinerja dari ketiga algoritma tersebut menggunakan metode pengujian *Confusion Matrix* dan kurva *ROC*, diketahui bahwa algoritma C4.5 dan algoritma *CART* memiliki nilai *accuracy* yang sama tinggi yaitu 95,6012% serta paling rendah adalah *accuracy* algoritma *naïve bayes* sebesar 89,5894%. Nilai AUC untuk algoritma *naïve bayes* menunjukkan nilai tertinggi yaitu 0,97 disusul algoritma C4.5 dengan nilai AUC 0,923 dan algoritma CART dengan nilai AUC 0,922. Melihat nilai AUC dari ketiga algoritma tersebut, maka semua algoritma termasuk kelompok klasifikasi yang sangat baik, karena nilai AUC-nya antara 0,90-1,00.

Kata kunci: kelulusan, algoritma C4.5, *Naïve Bayes, CART, Confusion Matrix,* Kurva *ROC*

1. PENDAHULUAN

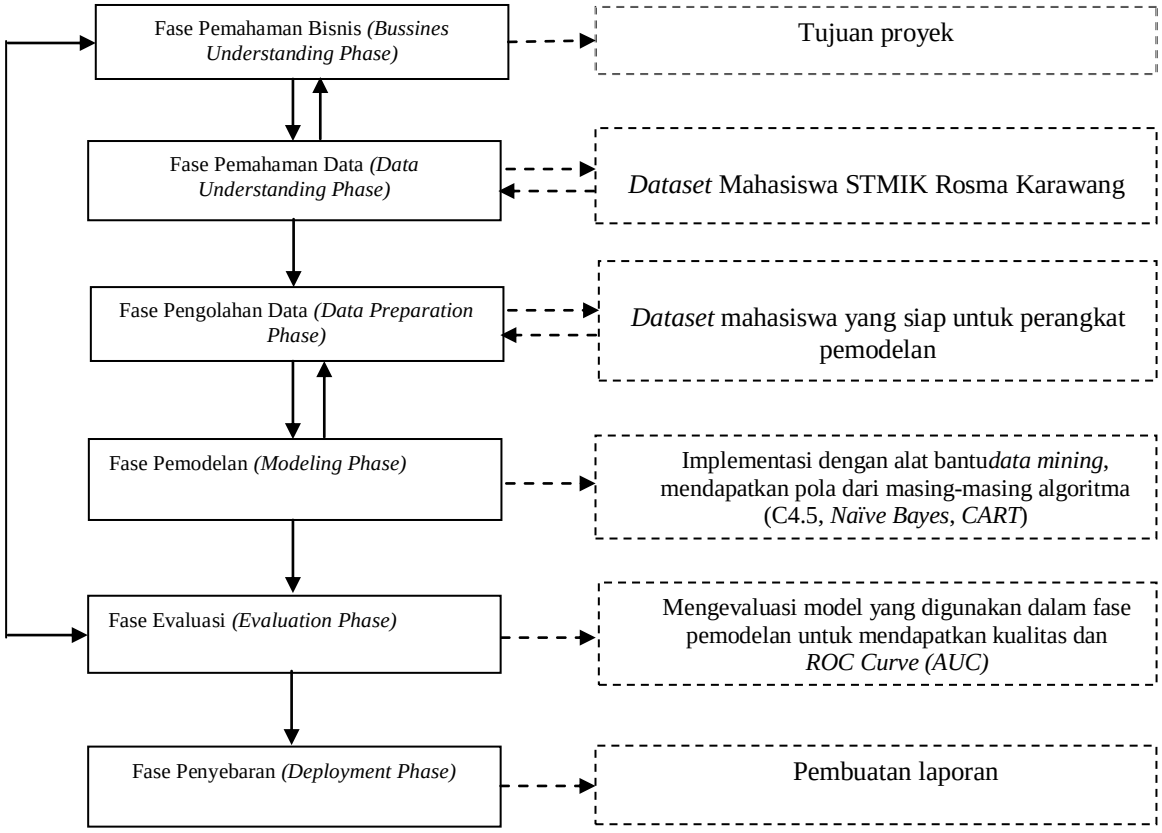
Data mahasiswa yang semakin banyak jumlahnya dari waktu ke waktu sangat disayangkan apabila tidak dianalisa.Klasifikasi data mahasiswa pada bidang pendidikan merupakan tugas penting dalam memprediksi kelulusan, bahkan dapat membantu pihak perguruan tinggi dalam mengambil keputusan atau kebijakan. Dengan demikian sangat

penting dalam memprediksi kelulusan secara dini untuk meningkatkan jumlah kelulusan mahasiswa.

Banyak penelitian sudah dilakukan dalam prediksi kelulusan namun belum ada yang membandingkan hasil dari tiga algoritma *(*C4.5, *naïve bayes* dan *CART)*.Ketiga algoritma tersebut digunakan dalam memprediksi kelulusan mahasiswa dengan tujuan agar algoritma terpilih merupakan algoritma yang paling akurat sehingga dapat melakukan prediksi kelulusan secara dini.Ketiga algoritma tersebut termasuk dalam sepuluh klasifikasi *data mining* yang paling popular.

2. METODOLOGI PENELITIAN

Penelitian ini didesain dengan merujuk pada model CRISP-DM *(Cross Industry Standard Process for Data Mining)*. Gambar 1 adalah tahapan yang dilakukan dalam penelitian.



Gambar 1. Tahapan Penelitian

3. EVALUASI DAN ALAT UKUR

Hasil evaluasi algoritma dapat ditampilkan dengan menggunakan *Confusion Matrix* (Tan, 2005).*Confusion Matrix* adalah salah satu alat ukur berbentuk *matrix* 2x2 yang digunakan untuk mendapatkan jumlah ketepatan klasifikasi dataset terhadap kelas lulus dan tidak lulus pada algoritma yang dipakai tiap kelas yang diprediksi memiliki empat kemungkinan keluaran yang berbeda, yaitu *true positive* (TP) dan *true negatives* (TN) yang menunjukkanketepatan klasifikasi. Jika prediksi keluaran bernilai positif sedangkan nilai aslinya adalah negatif maka disebutdengan *false positive* (FP) dan jika prediksi keluaran bernilai negatif sedangkan nilai aslinya adalah positif makadisebut dengan *false negative* (FN). Tabel 1 menyajikan bentuk *confusion matrix* seperti yangtelah dijelaskan sebelumnya.

Tabel 1. *Confusion Matrix* untuk Klasifikasi Kelas (Tan S, 2005)

		Predicated Class	
Actual Class		*Yes*	*No*
	Yes	*True Positive (TP)*	*False Negative (FN)*
	No	*False Positive (FP)*	*True Negative (TN)*

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

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+TN}$$

Pengukuran data dilakukan dengan *confusion matrix* (Tan S, 2005) dan ROC *Curve* (AUC) (Gorunescu, 2011) untuk mengevaluasi hasil dari algoritma *Decision Tree* C4.5. *Confusion matrix* merupakan sebuah table yang terdiri dari banyaknya baris data uji yang diprediksi benar dan tidak benar oleh model klasifikasi.Tabel ini diperlukan untuk mengukur kinerja suatu model klasifikasi (Ariawan, 2009).

Tabel 2. *Confusion Matrix* (Ariawan, 2009)

		Predicated Class	
Actual Class		*Class = 1*	*Class = 0*
	Class = 1	F 11	F 10
	Class = 0	F 01	F 00

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

$$\text{Akurasi} = \frac{F\ 11 + F\ 00}{F\ 11 + F\ 10 + F\ 01 + F\ 00}$$

ROC *(Receivier Operating Characteristic) Curve* adalah grafik antara sensitifitas *(true positive rate)* pada sumbu Y dengan 1-spesifisitas pada sumbu X *(false positive rate)*, seakan-akan menggambarkan tawar menawar antara sensitivitas dan spesifisitas, yang tujuannya adalah untuk menentukan *cut off point* pada uji *diagnostic* yang bersifat kontinyu. Untuk klasifikasi *data mining*, nilai AUC dapat dibagi menjadi beberapa kelompok (Gorunescu, 2011).

a. 0.90-1.00 = Klasifikasi sangat baik
b. 0.80-0.90 = Klasifikasi baik
c. 0.70-0.80 = Klasifikasi cukup
d. 0.60-0.70 = Klasifikasi buruk
e. 0.50-0.60 = Klasifikasi salah

4. TEKNIK KLASIFIKASI ALGORITMA C4.5, *NAÏVE BAYES* DAN *CART*

4.1. *CLASSIFICATION*

Classification adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Model itu sendiri bisa berupa aturan "jika-maka", berupa *decision tree*, formula matematis atau *neural network.Decision tree* adalah salah satu metode *classification* yang paling populer karena mudah untuk diinterpretasi oleh manusia.

4.2. ALGORITMA C4.5

Ada beberapatahapan dalam membangun sebuah pohon keputusan dengan Algoritma C4.5 yaitu (Kusrini, 2009) :

a. Menyiapkan *data training. Data training* biasanya diambil dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokkan ke dalam kelas-kelas tertentu.
b. Menentukan akar dari pohon. Akar akan diambil dari atribut atau variabel yang terpilih, dengan cara menghitung nilai gain dari masing-masing atribut atau variabel, nilai gain yang paling tinggi yang akan menjadi akar pertama. Sebelum menghitung nilai gain dari atribut atau variabel, hitung dahulu nilai *entropy*. Untuk menghitung nilai *entropy* digunakan rumus:

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

Keterangan:

S : himpunan kasus

A : atribut atau variabel

n : jumlah partisi S

pi : proporsi dari Si terhadap S

c. Kemudian hitung nilai *Gain* yang menggunakan rumus:

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i)$$

Keterangan :

S = himpunan kasus

A = fitur

n = jumlah partisi atribut atau variabel A

│Si│ = proporsi Si terhadap S

│S│ = jumlah kasus dalam S

d. Ulangi langkah ke-2 hingga semua *record* terpartisi.

e. Proses partisipohonkeputusan akan terhenti saat :

(1) Semua *record* dalam simpul N mendapat kelas yang sama.

(2) Tidak ada atribut atau variabel didalam *record* yang dipartisi lagi.

(3) Tidak ada *record* didalam cabang yang kosong.

4.3 ALGORITMA *NAÏVE BAYES*

Algoritma *Naïve Bayes*adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class.*NaïveBayes* didasarkan pada teorema Bayes yang memiliki kemampuan klasifikasi serupa dengan *decision tree* dan *neural network*. *NaïveBayes* terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar (Prasetyo, 2012). Prediksi Bayes didasarkan pada formula teorema Bayes dengan formula umum sebagai berikut :

$$P(H|X) = \frac{P(H|X) * P(H)}{P(X)}$$

Dimana:

X : Data dengan class yang belum diketahui

H : Hipotesis data X merupakan suatu class spesifik.

P(H|X) : Probabilitas hipotesis H berdasar kondisi X (posteriori probability)

P(H) : Probabilitas hipotesis H (prior probability)

P(X|H) : Probabilitas X berdasar kondisi pada hipotesis H

P(X) : Probabilitas dari X

*Naïve Bayes*adalah penyederhanaan metode *Bayes*. Teorema *Bayes* disederhanakan menjadi:

$$P(H|X) = P(X|H) * P(X)$$

Bayes rule diterapkan untuk menghitung posterior dan probabilitas dari data sebelumnya. Dalam analisis Bayesian, klasifikasi akhir dihasilkan dengan menggabungkan kedua sumber informasi (*prior* dan *posterior)* untuk menghasilkan *probabilitas* menggunakan aturan *Bayes.*

4.4. ALGORITMA *CART*

Ciri khas algoritma *CART* ini adalah noktah keputusan yang selalu bercabang 2 atau bercabang biner.Algoritma *CART* ini pertama kali digagas oleh Leo Breiman, Jerome Friedman, Richard Olshen, dan Charles Stone(Larose, 2005).Algoritma ini juga masuk dalam *The Top 10 Algorithm in Data Mining* (Wu dan Kumar, 2009).

Langkah – langkah pada algoritma *CART* adalah sebagai berikut (Susanto, dkk., 2010):

a. Langkah pertama, susunlah calon cabang *(Candidate Split)*. Penyusunan ini dilakukan terhadap seluruh variabel *predictor* secara lengkap *(Exhaustive).* Daftar yang berisi calon cabang disebut daftar calon cabang mutakhir.
b. Langkahkedua adalah menilai kinerja keseluruhan calon cabang yang ada pada daftar calon cabang mutakhir dengan jalan menghitung nilai besaran kesesuaian, Φ (s|t).
c. Langkah ketiga adalah menentukan calon cabang manakah yang akan benar-benar dijadikan cabang dengan memilih calon cabang yang memiliki nilai kesesuaian Φ (s|t) terbesar. Setelah itu gambarkanlah percabangan. Jika tidak ada lagi noktah keputusan, pelaksanaan algoritma *CART*akan dihentikan.
d. Namun, jika masih terdapat noktah keputusan, pelaksanaan algoritma dilanjutkan dengan kembali ke langkah kedua, dengan terlebih dahulu membuang calon cabang yang telah berhasil menjadi cabang sehingga mendapatkan daftar calon cabang yang baru.
e. Pohon keputusan yang dihasilkan *CART* merupakan pohon biner dimana tiap simpul wajib memiliki dua cabang. *CART* secara rekursif membagi *records* pada data latihan

ke dalam *subset-subset* yang memiliki nilai atribut atau variabel target (kelas) yang sama.

AlgoritmaCARTmengembangkan pohon keputusan dengan memilih percabangan yang paling optimal bagi tiap simpul.Pemilihan dilakukan dengan menghitung segala kemungkinan pada tiap variabel.

Misalkan $\Phi(s|t)$ merupakan nilai "kebaikan" kandidat cabang s pada simpul t, maka nilai $\Phi(s|t)$ dapat dihitung sebagai persamaan berikut (Larose, 2005):

$$\Phi(s|t) = 2P_L P_R \sum_{j=1}^{\#kelas} |P(j|t_L) - P(j|t_R)|$$

Dimana :

t_L=simpul anak kiri dari simpul t

t_R=simpul anak kanan dari simpul t

P_L=jumlah *record* pada t_L jumlah seluruh *record* pada data latihan

P_R=jumlah *record* pada t_R jumlah seluruh *record* pada data latihan

P j|t_L =jumlah *record* kelas j pada $t_L$jumlah *record* pada simpul t

P j|t_R =jumlah *record* kelas j pada$t_R$jumlah record *pada* simpult

Maksimal ketika *record*yang berada pada cabang kiri atau kanan simpul memiliki kelas yang sama (seragam). Nilai maksimal yang dicapai sama dengan jumlah kelas pada data. Misalkan jika data terdiri atas dua kelas, maka nilai maksimal adalah 2.

$$\sum_{j=1}^{\#kelas} |P(j|t_L) - P(j|t_R)|$$

Semakin seragam*record* pada cabang kiri atau kanan, maka semakin tinggi nilai. Nilai maksimal $2P_LP_R$sebesar 0.5 dicapai ketika cabang kiri dan kanan memiliki jumlah *record* yang sama. Kandidat percabangan yang dipilih adalah kandidat yang memiliki nilai $\Phi(s|t)$ paling besar.

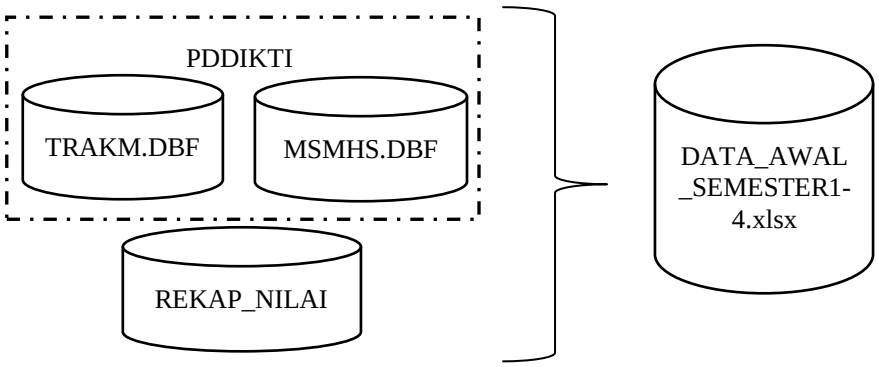
5. HASIL DAN PEMBAHASAN

5.1. FASE PEMAHAMAN BISNIS *(BUSINESS UNDERSTANDING PHASE)*

Tujuan proyek dalam penelitian ini adalah mengkaji dan membuat model hasil komparasi algoritma C4.5, *Naïve Bayes, CART*, serta menentukan algoritma mana yang paling akurat dan menghasilkan *rule* prediksi kelulusan mahasiswa STMIK Rosma Karawang sehingga dapat dijadikan acuan untuk meningkatkan jumlah kelulusan mahasiswa di tahun-tahun kelulusan berikutnya.

5.2. FASE PEMAHAMAN DATA *(DATA UNDERSTANDING PHASE)*

Sumber data utama yang digunakan dalam penelitian ini adalah data mahasiswa STMIK Rosma Karawang jenjang DIII dan S1 pada tahun 2005 sampai dengan tahun 2008 dengan format xlsx dan DBF.Data dalam tabel 1.1 adalah gabungan dari data yang ada di folder REKAP_NILAI, data PDDIKTI yaitu master mahasiswa (MSMHS.DBF) dan transkrip nilai (TRAKM.DBF).Hasil penggabungan tersebut adalah DATA_AWAL_SEMESTER1-4.xlsx seperti pada gambar 2.



Gambar 2. *Dataset* Mahasiswa 2005-2008

Hasil evaluasi terhadap kualitas data adalah masih terdapat data yang rangkap atau double dan ditemukan banyak nilai kosong atau null yang disebut dengan *missing value.* Sehingga dataset yang didapatkan sejumlah 682 *record* seperti terlihat pada tabel 3.

Tabel 3. *Dataset* Jumlah Mahasiswa STMIK Rosma 2005-2008

No	Prodi	Jenjang	Angkatan	Jumlah
1	Manajemen Informatika (57401)	DIII (E)	2005-2008	197
2	Komputerisari Akuntansi (57402)	DIII (E)	2005-2008	142
3	Teknik Informatika (55201)	S1 (C)	2005-2008	227
4	Sistem Komputer (56201)	S1 (C)	2005-2008	25
5	Sistem Informasi (57201)	S1 (C)	2005-2008	91
Total				682

5.3. FASE PENGOLAHAN DATA *(DATA PREPARATION PHASE)*

Data mahasiswa terdiri dari beberapa variabel antara lain IP Semester1 (IPS1), IP Semester2 (IPS2), IP Semester 3 (IPS3), IP Semester 4 (IPS4), Jumlah SKS yang sudah

ditempuh (JMLSKS), PRODI, Jenjang (JNJNG), Jenis Kelamin (JNSKLMN), Tanggal Kelulusan (KELULUSAN). Pengkategorian datanya sebagai berikut:

(1) Variabel Indeks Prestasi Semester (IPS)

Jenis data IPS diambil dari semester 1, semester 2, semester 3 dan semester 4 serta dikategorikan menjadi 3 seperti ditampilkan pada tabel 4.

Tabel 4. Kategori IPK

BESAR	IPS≥3,50
SEDANG	2,75<IPS<3,50
KECIL	IPS≤2,75

(2) Variabel Jumlah SKS yang telah ditempuh (JMLSKS)

Jenis data JMLSKS merupakan data real yang dikategorikan berdasarkan rata-rata jumlah SKS yang telah ditempuh oleh mahasiswa selama empat semester dari semester satu sampai semester empat seperti terlihat pada tabel 5.

Tabel 5. Kategori Jumlah SKS

Kategori	Keterangan
KECIL	SKS<67
BESAR	SKS≥67

(3) Variabel Program Studi (PRODI)

Kategori Prodi dapat dilihat pada tabel 6.

Tabel 6. Kode Program Studi

Kode	Prodi	Kategori
55201	Teknik Informatika (S1)	TIC
56201	Sistem Komputer (S1)	SKC
57201	Sistem Informasi (S1)	SIC
57401	Manajemen Informatika (D3)	MIE
57402	Komputerisasi Akuntansi (D3)	KAE

(4) Variabel Jenjang (JNJNG)

Jenis data Jenjang dikategorikan seperti pada tabel 7.

Tabel 7. Kode Jenjang

Kode	Jenjang
C	Strata Satu (S1)
E	Diploma Tiga (DIII)

(5) Variabel Jenis Kelamin (JNSKLMN)

Jenis data dari variabel Jenis Kelamin dikategorikan menjadi Laki-laki dengan inisial L dan Perempuan dengan inisial P.

Data yang berjenis numerikal seperti tahun lulus harus dilakukan proses inisiasi data terlebih dahulu kedalam bentuk nominal. Untuk melakukan inisiasi tahun lulus dapat dilakukan dengan:

(a) Mahasiswa dari setiap angkatan yang sudah terdapat tahun kelulusan dinyatakan "lulus".

(b) Mahasiswa dari setiap angkatan yang belum terdapat tahun kelulusan dinyatakan "tidak lulus".

Tabel 8 adalah *dataset* mahasiswa tahun 2005-2008 yang belum diinisiasi dan tabel 9 adalah *dataset* mahasiswa tahun 2005-2008 yang siap untuk perangkat pemodelan.

Tabel 8. *Dataset* Mahasiswa tahun 2005-2008 yang belum diinisiasi

NIM	IPS1	IPS2	IPS3	IPS4	JUMLAHSKS	PRODI	JENJANG	JENISKELAMIN	KELULUSAN
2005 01 15 1 001	3,05	2,35	2,67	2,27	83	SIC	C	L	26-Agust-10
2005 01 15 1 002	3,15	2,30	3,09	2,88	90	SIC	C	L	17-Jul-09
2005 01 15 1 003	2,59	2,70	2,83	3,00	87	SIC	C	L	13-Jul-09
2005 01 15 1 005	3,45	2,43	2,87	3,13	90	SIC	C	P	14-Jul-09
2005 01 15 1 006	2,90	1,65	2,52	2,18	83	SIC	C	L	20-Agust-11
2005 01 15 1 008	2,95	2,00	2,24	2,00	79	SIC	C	L	20-Agust-11
2005 01 15 1 010	2,90	2,09	3,13	2,96	90	SIC	C	L	17-Jul-09
2005 01 15 1 012	3,00	2,48	3,17	2,88	90	SIC	C	L	14-Jul-09
2005 01 15 1 014	2,80	2,48	3,04	3,58	90	SIC	C	P	11-Jul-09
2005 01 15 1 015	3,10	3,09	3,22	2,92	90	SIC	C	L	11-Jul-09
2005 01 15 1 016	2,70	1,78	2,35	2,33	90	SIC	C	P	11-Jul-09
2005 01 15 3 017	2,85	0,02	2,00	0,07	46	SIC	C	L	
2005 01 15 3 018	2,90	1,47	2,68	0,00	64	SIC	C	P	
2005 01 15 3 019	3,20	3,14	3,04	1,00	70	SIC	C	P	
2005 01 15 7 020	3,70	3,48	3,26	3,50	90	SIC	C	P	15-Jul-09
2005 01 25 1 001	2,64	3,05	3,08	3,20	88	TIC	C	L	11-Jul-09
2005 01 25 1 002	3,09	2,50	3,17	3,17	92	TIC	C	L	11-Jul-09
2005 01 25 1 003	2,23	2,32	2,67	0,00	68	TIC	C	L	
2005 01 25 1 004	3,23	3,32	3,58	3,25	92	TIC	C	L	11-Jul-09
2005 01 25 1 005	0,17	0,00	0,00	0,00	18	TIC	C	L	
2005 01 25 1 006	2,18	2,27	0,50	0,00	52	TIC	C	L	
2005 01 25 1 007	3,64	3,09	3,21	2,50	92	TIC	C	L	15-Jul-09
2005 01 25 1 008	2,73	2,36	2,29	0,81	89	TIC	C	L	11-Jul-09
2005 01 25 1 009	3,00	2,91	2,83	3,38	92	TIC	C	L	11-Jul-09

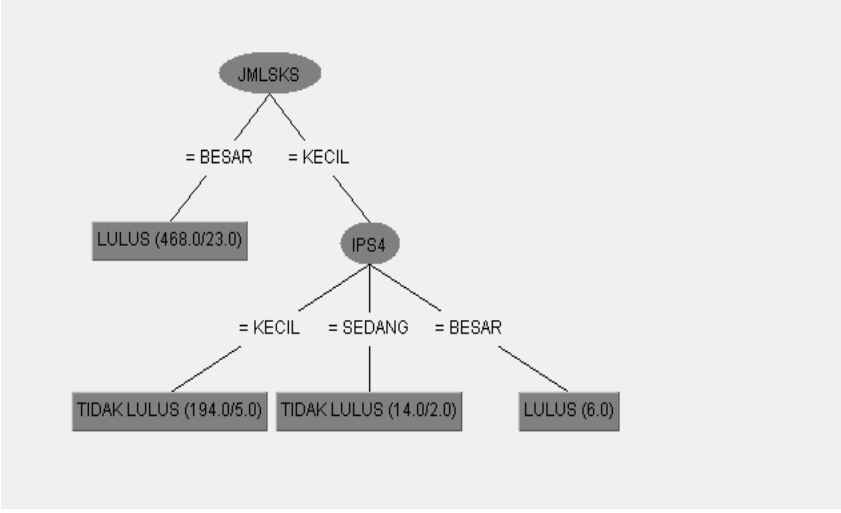
Tabel 9. *Dataset* Mahasiswa tahun 2005-2008 yang Siap Perangkat Pemodelan

IPS1	IPS2	IPS3	IPS4	JMLSKS	PRODI	JNJNG	JNSKLMN	KELULUSAN
SEDANG	KECIL	KECIL	KECIL	BESAR	SIC	C	L	LULUS
SEDANG	KECIL	SEDANG	SEDANG	BESAR	SIC	C	L	LULUS
KECIL	KECIL	SEDANG	SEDANG	BESAR	SIC	C	L	LULUS
SEDANG	KECIL	SEDANG	SEDANG	BESAR	SIC	C	P	LULUS
SEDANG	KECIL	KECIL	KECIL	BESAR	SIC	C	L	LULUS
SEDANG	KECIL	KECIL	KECIL	BESAR	SIC	C	L	LULUS
SEDANG	KECIL	SEDANG	SEDANG	BESAR	SIC	C	L	LULUS
SEDANG	KECIL	SEDANG	SEDANG	BESAR	SIC	C	L	LULUS
SEDANG	KECIL	SEDANG	BESAR	BESAR	SIC	C	P	LULUS
SEDANG	SEDANG	SEDANG	SEDANG	BESAR	SIC	C	L	LULUS
KECIL	KECIL	KECIL	KECIL	BESAR	SIC	C	P	LULUS
SEDANG	KECIL	KECIL	KECIL	KECIL	SIC	C	P	TIDAK LULUS
BESAR	SEDANG	SEDANG	BESAR	BESAR	SIC	C	P	LULUS
KECIL	SEDANG	SEDANG	SEDANG	BESAR	TIC	C	L	LULUS
SEDANG	KECIL	SEDANG	SEDANG	BESAR	TIC	C	L	LULUS
KECIL	KECIL	KECIL	KECIL	BESAR	TIC	C	L	TIDAK LULUS
SEDANG	SEDANG	BESAR	SEDANG	BESAR	TIC	C	L	LULUS
KECIL	KECIL	KECIL	KECIL	KECIL	TIC	C	L	TIDAK LULUS
KECIL	KECIL	KECIL	KECIL	KECIL	TIC	C	L	TIDAK LULUS
BESAR	SEDANG	SEDANG	KECIL	BESAR	TIC	C	L	LULUS
KECIL	KECIL	KECIL	KECIL	BESAR	TIC	C	L	LULUS
SEDANG	SEDANG	SEDANG	SEDANG	BESAR	TIC	C	L	LULUS
SEDANG	SEDANG	SEDANG	SEDANG	BESAR	TIC	C	L	LULUS
KECIL	KECIL	KECIL	KECIL	KECIL	TIC	C	L	TIDAK LULUS

5.4. FASE PEMODELAN *(MODELING PHASE)*

(1) Algoritma C4.5

Berikut ini adalah hasil penelitian dengan menggunakan *dataset* mahasiswa tahun 2005-2008 seperti pada tabel 9 yang selanjutnya akan diolah dengan alat bantu *data mining* WEKA 3.6.1 dengan algoritma C4.5. Gambar 3 adalah pohon keputusan yang terbentuk.



Gambar 3. Pohon Keputusan Algoritma C4.5

Dari sembilan variabel yang digunakan terlihat hanya dua variabel yang membentuk pohon, yaitu variabel JMLSKS (Jumlah SKS) dan IPS4 (IP Semester 4).Sedangkan variabel IPS1, IPS2, IPS3, Jenjang, Jenis Kelamin dan PRODI tidak terlihat dari pohon keputusan. Yang menjadi simpul akar adalah JMLSKS (Jumlah SKS) karena memiliki *gain* tertinggi. Jika JMLSKS BESAR maka "lulus" sedangkan jika JMLSKS KECIL maka lihat IPS4 (IP Semester 4).Jika IPS4 KECIL dan SEDANG maka "tidak lulus" sedangkan jika IPS4 BESAR maka "lulus".

Confusion Matrix untuk algoritma C4.5 dapat dilihat pada tabel 10. *Confusion Matrix* menunjukkan ketepatan klasifikasi atau kesesuaian dengan prediksi yang dilakukan dengan metode C4.5.

Tabel 10. *Confusion Matrix* Algoritma C4.5

```
=== Confusion Matrix ===

   a   b   <-- classified as
 451   7 |   a = LULUS
  23 201 |   b = TIDAK LULUS
```

Akurasi = $\frac{451+201}{451+7+23+201}$ x 100% = 95.6012 %

Nilai kurva *ROC* yang dihasilkan adalah 0,923.

Keterangan tabel 1.8 adalah:

(a) Jumlah data *real* yang LULUS dan diprediksi LULUS adalah 451.

(b) Jumlah data *real* yang TIDAK LULUS dan diprediksi TIDAK LULUS adalah 201.

(c) Jumlah data *real* yang TIDAK LULUS dan diprediksi LULUS adalah 23

(d) Jumlah data *real* yang LULUS dan diprediksi TIDAK LULUS adalah 7.

(2) *Naïve Bayes*

Berikut ini adalah hasil penelitian dengan menggunakan *dataset* mahasiswa tahun 2005-2008 pada tabel 1.7 yang selanjutnya akan diolah dengan alat bantu *data mining* WEKA 3.6.1 dengan algoritma *Naïve Bayes*. Hasil yang didapat sudah memberikan informasi atau pengetahuan yang sesuai dengan tujuan proyek.Tabel 11 adalah *Confusion Matrix* yang terbentuk dari pemodelan dengan algoritma *Naïve Bayes*.

Tabel 11. *Confusion Matrix* Algoritma *Naïve Bayes*

```
=== Confusion Matrix ===

   a   b   <-- classified as
 407  51 |   a = LULUS
  20 204 |   b = TIDAK LULUS
```

Akurasi = $\frac{407 + 204}{407 + 204 + 51 + 20}$ x 100% = 89,5894 %

Nilai kurva *ROC* yang dihasilkan adalah 0,97.

(3) *CART*

Penelitian selanjutnya adalah dengan menggunakan *dataset* mahasiswa tahun 2005-2008 seperti pada tabel 1.7 dengan algoritma *CART* dan diolah dengan alat bantu*data mining* WEKA 3.6.1. Hasil yang didapat sudah memberikan informasi atau pengetahuan yang sesuai dengan tujuan proyek.Tabel 12 adalah *Confusion Matrix* yang terbentuk dari pemodelan dengan algoritma *CART*.

Tabel 12. *Confusion Matrix* Algoritma *CART*

```
=== Confusion Matrix ===

   a   b   <-- classified as
 451   7 |   a = LULUS
  23 201 |   b = TIDAK LULUS
```

Akurasi = $\frac{451 + 201}{451 + 201 + 7 + 23} x\ 100\%$ = 95.6012%

Nilai kurva *ROC* yang dihasilkan adalah 0,922.

5.5 FASE EVALUASI *(EVALUATION PHASE)*

Pebandingan hasil perhitungan nilai kurca *ROC* (*AUC)* untuk algoritma C4.5, *Naïve Bayes*, dan *CART* dapat dilihat pada tabel 13.

Tabel 13. Komparasi Nilai AUC

Algoritma	Nilai AUC
C4.5	0.923
Naïve Bayes	0.97
CART	0.922

6. ANALISIS HASIL KOMPARASI

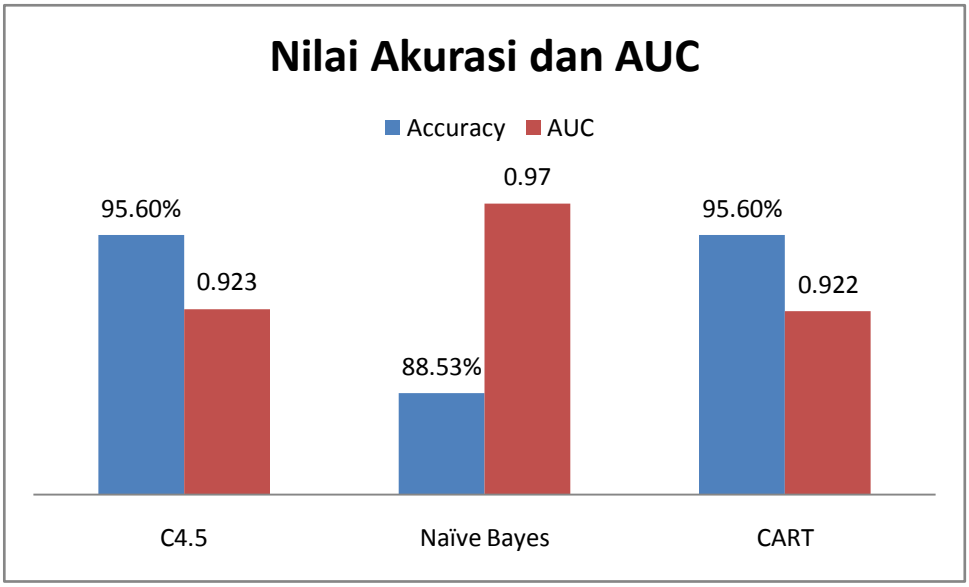
Model yang dihasilkan algoritma C4.5, *Naïve Bayes*, dan *CART* diuji menggunakan metode *Cross Validation*, terlihat algoritma C4.5 dan algoritma *CART*

memiliki nilai *accuracy* yang sama dan paling tinggi, sedangkan yang terendah adalah *Naïve Bayes.*

Tabel 14. Komparasi Nilai *Accuracy* dan AUC

	C4.5	*Naïve Bayes*	*CART*
Accuracy	95,6012%	88,5341%	95,6012%
AUC	0,923	0,97	0,922

Tabel 1.12 membandingkan A*ccuracy* dan AUC dari tiap algoritma. Terlihat bahwa nilai *Accuracy* algoritma C4.5 dan algoritma *CART* memiliki nilai yang sama yaitu 95,6012%. Nilai AUC paling tinggi adalah *Naïve Bayes*. Berdasarkan pengelompokkan tabel 1.12 maka dapat disimpukan bahwa algoritma C4.5, *Naïve Bayes*, dan *CART* termasuk klasifikasi sangat baik karena memiliki nilai AUC antara 0.90-1.00. Grafik nilai akurasi dan nilai AUC masing-masing algoritma terlihat pada gambar 4.



Gambar 4. Nilai Akurasi dan AUC masing-masing Algoritma

7. KESIMPULAN

Hasil penelitian yang diperoleh sudah sesuai dengan tujuan penelitian yaitu sebagai berikut:

a. Dapat membandingkan tingkat akurasi yang dihasilkan masing-masing algoritma. Dengan alat bantu WEKA prediksi tingkat kelulusan mahasiswa STMIK Rosma

Karawang, Algoritma C4.5 menghasilkan akurasi 95,6012%, Naïve Bayes 89,5894% dan CART 95,6012%.

b. Algoritma C4.5, *Naïve Bayes* dan *CART* dapat digunakan untuk memprediksi kelulusan mahasiswa. Untuk mengukur ketiga fungsi algoritma tersebut digunakan *confusion matrix* dan kurva ROC dengan hasil bahwa algoritma yang memiliki tingkat akurasi yang paling tinggi adalah algoritma C4.5 dan algoritma *CART*. Sedangkan algoritma yang menghasilkan kurva ROC paling tinggi adalah algoritma *Naïve Bayes.* Akurasi algoritma C4.5 dan algoritma *CART* memberikan hasil yang sama yaitu 95,6012%. Hal ini terjadi karena algoritma C4.5 membangun pohon dengan jumlah pohon tiap simpul sesuai dengan jumlah nilai simpul tersebut, seperti kasus data yang peneliti lakukan terhadap data nilai mahasiswa yang dikelompokkan kedalam dua kelompok (lulus dan tidak lulus) sehingga akan sama dengan algoritma *CART* dengan konsep cabang pohon biner.
c. Algoritma C4.5 dan algoritma *CART* memberikan akurasi yang lebih baik daripada *Naïve Bayes* dalam klasifikasi data mahasiswa STMIK Rosma Karawang.
d. Algoritma C4.5 dan algoritma *CART* memberikan hasil lebih baik karena data mahasiswa STMIK Rosma merupakan data kelompok yang cocok dengan sifat klasifikasi algoritma C4.5 dan algoritma *CART*.
e. Kelulusan mahasiswa dapat diprediksi lebih dini yaitu pada semester 4.
f. *Data mining* dengan algoritma C4.5 dan *CART* dapat diimplementasikan untuk memprediksi kelulusan mahasiswa STMIK Rosma dengan dua kategori yaitu lulus dan tidak lulus. Variabel yang berpengaruh dalam hasil prediksi adalah Jumlah SKS yang telah ditempuh (JMLSKS) dan Indek Prestasi Semester 4 (IPS4).
g. Dapat menjabarkan masing-masing algoritma kedalam *rule*.
h. Dapat menerapkan masing-masing algoritma dalam melakukan prediksi terhadap kelulusan mahasiswa STMIK Rosma Karawang.

Untuk penelitian selanjutnya dapat menambah variabel lain, selain dari variabel yang dilakukan peneliti serta menggunakan perpaduan algoritma lainnya seperti *k-nearest neighbor* dan *neural network.*

8. DAFTAR PUSTAKA

[1] Andi Wahju Rahardjo, Emanuel danHartono Arie. 2008.*Pengembangan Aplikasi Pengenalan Karaketer Alfanumerik Dengan Menggunakan AlgoritmaNeural Network Three-Layer Backpropagation.*Jurnal Informatika. Vol.4, No.1. 49 – 58.

[2] Ariawan, Iwan. 2009. *Catatan materi kuliah dr Iwan Ariawan, MS.* [Online]. http://www.scribd.com/doc/15123416/Kurva-Receiver-Operating-Characteristic. Diakses tanggal 19 Oktober 2015, 20:15 WIB

[3] Al-Radaideh, Q.A. 2006. *Mining Student Data Using Decission Tree. International Arab Conference on Informational Technology* (ACIT).

[4] Basuki, Ahmad dan Syarif, Iwan. 2003. *Decision Tree*. Surabaya : PoliteknikElectronika Negeri Surabaya ITS.

[5] Bramer, M. 2007. *Principles of Data Mining.* London: Springer

[6] Gorunescu, Florin. 2011. *Data Mining Concept Model Technique.*

[7] Han, J, dan Kamber, M. 2001. *Data Mining Concepts and Techniques.* Morgan Kaufman Pub.USA

[8] Huda, Nuqson Masykur. 2010. *Aplikasi Data Mining Untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa*.Semarang

[9] Jananto A. 2010. *Penggunaan Algoritma SLIQ untuk Pengklasifikasian Kinerja Akademik Mahasiswa.* Jurnal Teknologi Informasi DINAMIK Vol XV, No.1 : 66-72

[10] Jefri. 2013. *Implementasi Algoritma C4.5 Dalam Aplikasi Untuk Memprediksi Jumlah Mahasiswa Yang Mengulang Mata Kuliah Di STMIK AMIKOM Yogyakarta,* Yogyakarta

[11] Kamagi, David Hartanto dan Hansun Seng. 2014.*Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa*. ULTIMATICS Vol. VI, No. 1

[12] Karypis, George, dkk. 2007.*CHAMELEON:A Hierarchical Clustering Algorithm UsingDynamicModeling*. Diambil darihttp://www.users.cs.umn.edu/~kumar/papers/chameleon.ps. Diakses tanggal 19 Oktober 2015, 20:20 WIB.

[13] Kursini, Luthfi. E. T. 2009. *Algoritma Data Mining.*PT Andi Offset.

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

[15] M, Mohammed, dkk. 2012.*Mining Educational Data to Improve Students' Performance: A Case Study.*International Journal of Information and Communication Technology Research Volume 2 No. 2

[16] Maimon, Rockah. 2005. *Data Mining and Knowledge Discovery Handbook*. Springer Heidelberg. Berlin

[17] Moertini, Veronica Sri. 2007.*Pengembangan Skalabilitas Algoritma Klasifikasi C4.5 Dengan Pendekatan Konsep Operator Relasi, studi kasus: pra-pengolahan dan klasifikasi citra batik.* Bandung

[18] Nilakant, K. 2004. *Application of Data Mining in Constraint Based Intelligent Tutoring System.*www.cosc.canterbury.ac.nz/research/reports/HonsReps/2004/hons_0408.pdf. diakses tanggal 12 Oktober 2015

[19] Nugroho, Yuda Septian. 2014. *Data Mining Menggunakan Algoritma Naïve Bayes Untuk Klasifikasi Kelulusan Mahasiswa Universitas Dian Nuswantoro*. Fasilkom UDINUS Semarang

[20] Nugroho, Yusuf Sulistyo.2014. *Penerapan Algoritma C4.5 Untuk Klasifikasi Predikat Kelulusan Mahasiswa Fakultas Komunikasi Dan Informatika Universitas Muhammadiyah Surakarta*. Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Yogyakarta

[21] Oscar Ong, Johan.2013. *Implementasi Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing President University*.Jurnal Ilmiah Teknik Industri. vol. 12.No. 1.pp. 10-13.

[22] Prasetyo, Eko. 2012. *Data Mining : Konsep dan Aplikasi menggunakan MATLAB, 1st ed.*PT Andi Offset

[23] Rahmayuni,Indri. 2014. *Klasifikasi Data Karakteristik Mahasiswa Menggunakan Algoritma C4.5 Dan Cart (Studi Kasus Educational Data Mining)*.Jurnal Teknologi Informasi & Pendidikan ISSN : 2086 – 4981. Vol. 7 No. 1

[24] Ridwan, Mujib, dkk.2013. *Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naïve Bayes Classifier*. Malang

[25] STMIK ROSMA Karawang. 2010. *Buku Panduan Akademik Mahasiswa Tahun Ajaran 2010-2011*. Karawang. Jawa Barat

[26] Susanto, Sani, dkk. 2010. *Pengantar Data Mining Menggali Pengetahuan Dari Bongkahan Data*.ANDI Yogyakarta.

[27] Suhartinah, Marselina Silvia dan Ernastuti. 2010. *Graduation Prediction OfGunadarma University StudentsUsing Algorithm And Naive Bayes C4.5 Algorithm.*Undergraduate Program, Faculty of Industrial Engineering,Gunadarma University

[28] Sunjana. 2010. *Aplikasi MiningData Mahasiswa Dengan Metode Klasifikasi Decision Tree.*Seminar Nasional Aplikasi Teknologi Informasi.

[29] Swastina, Liliana. 2013. *Penerapan Algoritma C4.5 Untuk Penentuan Jurusan Mahasiswa.* Jurnal GEMA AKTUALITA, Vol. 2 No. 1

[30] Tan S, Kumar P, Steinbach M. 2005. *Introduction to Data Mining.* Addison Wesley. J. Taylor, Ed. Stanford

[31] Turban, E, dkk.2005. *Decicion Support Systems and Intelligent Systems*. PT Andi Offset

[32] Witten, Ian H, dkk. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc.San Francisco, CA, USA

[33] Wu & Kumar. 2009. *The Top Ten Algorithms in Data Mining.* USA: CRC Press

[34] Yadav, Surjeet Kumar, dkk. 2012. *Data Mining Applications: A comparative Study for Predicting Student's performance.* International Journal Of Innovative Technology & Creative Engineering (ISSN:2045-711). Vol.1 No.12

[35] http://www.cs.waikato.ac.nz/ml/weka/ diakses tanggal 19 Oktober 2015, 19:59 WIB