KAJIAN PENERAPAN ALGORITMA C4.5 DAN *NAÏVE BAYES* UNTUK KLASIFIKASI PENERIMA BEASISWA KOPERTIS

STUDI KASUS AMIK BSI YOGYAKARTA

TESIS

MUHAMAD TABRANI

(14000641)

PROGRAM PASCASARJANA MAGISTER ILMU KOMPUTER
SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
NUSA MANDIRI
JAKARTA
2014

SURAT PERNYATAAN ORISINALITAS

Yang bertanda tangan di bawah ini,saya:

Nama	: Muhamad Tabrani
Nim	: 14000641
Program studi	: Magister Ilmu Komputer
Jenjang	: Strata Dua (S2)
Konsentrasi	: *e-business*

Dengan ini menyatakan bahwa tesis yang telah saya buat dengan judul ”Kajian Penerapan Algoritma C4.5 dan *Naïve Bayes* untuk Klasifikasi Penerima Beasiswa Kopertis : Studi Kasus AMIK BSI Yogyakarta” adalah hasil karya sendiri, dan semua sumber baik yang kutip maupun yang dirujuk telah saya nyatakan dengan benar dan tesis belum pernah diterbitkan atau dipublikasikan dimanapun dan dalam bentuk apapun.

Demikianlah surat pernyataan ini saya buat dengan sebenar-benarnya. Apabila dikemudian hari ternyata saya memberikan keterangan palsu atau ada pihak lain yang mengklaim bahwa tesis yang telah saya buat adalah hasil karya milik seseorang atau badan tertentu, saya bersedia diproses baik secara pidana maupun perdata dan kelulusan saya dari Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri dicabut/dibatalkan.

Jakarta, 04 Maret 2014
Yang menyatakan

Materai 6000

Muhamad Tabrani

HALAMAN PENGESAHAN

Tesis ini diajukan oleh :

Nama	:	Muhamad Tabrani
NIM	:	14000641
Program Studi	:	Magister Ilmu Komputer
Jenjang	:	Strata Dua (S2)
Konsentrasi	:	*e-business*
Judul Tesis	:	**"Kajian Penerapan Algoritma C4.5 dan *Naïve Bayes* untuk Klasifikasi Penerima Beasiswa Kopertis : Studi Kasus AMIK BSI Yogyakarta"**

Telah berhasil dipertahankan dihadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Magister Ilmu Komputer (M.Kom) pada Program Pascasarjana Ilmu Komputer Sekolah Manajemen Informatika dan Komputer Nusa Mandiri (STMIK Nusa Mandiri).

Jakarta, 04 Maret 2014
Pascasarjana Ilmu Komputer
STMIK Nusa Mandiri
Direktur

Prof. Dr. Ir. Kaman Nainggolan, MS

DEWAN PENGUJI

Penguji I	:	Dr. Ir Probowo Pudjo Widodo MS	 .....
Penguji II	:	Windu Gata M.Kom	 .....
Penguji III / Pembimbing	:	Mochamad Wahyudi,MM, M.Kom, M.Pd	 .....

ABSTRAK

Setiap tahunnya masing–masing perguruan tinggi mendapatkan beasiswa dari Koordinator Koordinator Perguruan Tinggi Swasta (Kopertis) beasiswa yang ditujukan untuk mahasiswa di perguruan tinggi swasta yaitu Beasiswa Peningkatan Prestasi Akademik (Beasiswa PPA) dan Beasiswa Bantuan Belajar Mahasiswa (Beasiswa BBM). Proses pengajuan beasiswa PPA dan BBM melalui dua tahap seleksi yaitu seleksi pertama merupakan tahap seleksi di perguruan tinggi untuk menentukan calon penerima beasiswa yang akan diusulkan ke kopertis. Tahap seleksi yang kedua yaitu tahap seleksi di kopertis. Banyaknya mahasiswa yang mengajukan beasiswa tersebut serta melebihi kouta yang diberikan mengakibatkan proses penyeleksian penerima memakan waktu yang lama karena penyeleksian harus sesuai dengan kriteria agar penerima beasiswa tepat sasaran. Berdasarkan permasalahan tersebut perlu suatu tindakan untuk menentukan penerima beasiswa yang tepat. Tujuan Penelitian ini adalah membuat klasifikasi mahasiswa penerima beasiswa dengan algoritma C4.5, dan *Naïve Bayes* serta membandingkan hasil klasifikasi kedua algoritma tersebut agar diperoleh algoritma terbaik. Hasil klasifikasi dari tiga algoritma dievaluasi dan divalidasi dengan *confusion matrix* dan kurva ROC, hasilnya diperoleh algoritma terbaik untuk klasifikasi mahasiswa *dropout* yaitu algoritma C4.5 dengan tingkat akurasi 86.88%. Sehingga dapat diterapkan untuk permasalahan penentuan penerima beasiswa.

Kata Kunci:
Algoritma C4.5, *Naïve Bayes.*

DAFTAR ISI

1. PENDAHULUAN

Setiap tahunnya masing–masing perguruan tinggi mendapatkan beasiswa dari Koordinator Koordinator Perguruan Tinggi Swasta (Kopertis) beasiswa yang ditujukan untuk mahasiswa di perguruan tinggi swasta yaitu Beasiswa Peningkatan Prestasi Akademik (Beasiswa PPA) dan Beasiswa Bantuan Belajar Mahasiswa (Beasiswa BBM). Beasiswa PPA adalah beasiswa yang diberikan kepada para mahasiswa jenjang Strata Satu dan Diploma Tiga yang mempunyai prestasi akademik baik yaitu minimal Indeks Prestasi Kumulatif (IPK) minimal 3,00. Sedangkan beasiswa BBM adalah beasiswa yang diberikan kepada para mahasiswa jenjang Strata Satu dan Diploma Tiga yang mempunyai prestasi akademik minimal IPK 2,50 dan aktif dalam kegiatan kemahasiswaan serta kondisi orangtuanya kurang mampu. Tujuan dari pemberian beasiswa PPA dan BBM ini untuk membantu mahasiswa yang kurang mampu dalam membayar biaya studi di perguruan tingginya.

Hasil seleksi penerima beasiswa yang dilakukan secara manual dapat berbeda-beda tergantung dari pengambil keputusan yang terlibat dalam proses penentuan penerima beasiswa (Karismariyanti, 2011). Hal ini sering menimbulkan kesalahan dalam penentuan calon penerima beasiswa seperti terpilihnya penerima beasiswa yang kurang tepat. Selain itu proses pengambilan keputusan untuk menentukan calon penerima beasiswa yang dilakukan secara manual sulit dan membutuhkan waktu lama. Untuk membantu dalam pengambilan keputusan diperlukan suatu klasifikasi mahasiswa penerima beasiswa.

Klasifikasi merupakan salah satu dari teknik data mining yang merupakan proses penempatan objek atau konsep tertentu ke dalam satu set kategori

berdasarkan objek yang digunakan. Dari berbagai teknik klasifikasi yang paling populer digunakan adalah *decision tree* (Han & Kamber, 2006). Sedangkan algoritma yang dapat dipakai dalam *decision tree* salah satunya adalah Algoritma C4.5. Selain menggunakan konsep *decision tree* dengan algoritma C4.5, dalam membuat klasifikasi dapat juga diterapkan algoritma *Naive Bayes*.

Komparasi algoritma C4.5 dan *Naive Bayes* dilakukan untuk mencari hasil klasifikasi dengan tingkat akurasi yang tinggi dalam prediksi mahasiswa penerima beasiswa. Rule algoritma hasil komparasi dengan tingkat akurasi tertinggi selanjutnya diaplikasikan dalam membuat program seleksi penerima beasiswa PPA dan BBM. Program seleksi penerima beasiswa PPA dan BBM yang digunakan untuk mempermudah dan mempercepat proses seleksi di perguruan tinggi dapat dibangun dengan menggunakan *software* Microsoft Visual Basic 6.0 dalam menerapkan *rule* hasil klasifikasi mahasiswa penerima beasiswa.

2. LANDASAN/KERANGKA PEMIKIRAN

A. Data Mining

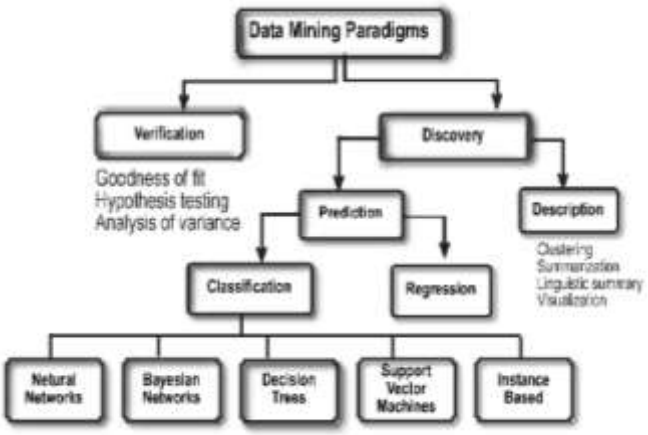
Data mining merupakan sebuah proses terpadu dari analisis data yang terdiri dari serangkaian kegiatan yang berjalan berdasarkan pada pendefinisian tujuan dari apa yang akan dianalisis sampai pada interpretasi dan evaluasi hasil (Giudici & Figini, 2009). *Data mining* juga dapat diartikan sebagai proses dalam menemukan pola dari sebuah set data dimana proses tersebut harus otomatis atau biasanya semi-otomatis dan pola yang dihasilkan harus berarti bahwa pola tersebut memberikan beberapa keuntungan (Witten, Frank, & Hall, 2011).

Menurut Giudici & Figini (2009), berbagai tahapan dari proses *data mining* adalah:

1. Definisi tujuan untuk analisa
2. Seleksi, organisasi, dan pra-perawatan data
3. Eksplorasi analisis data dan mentransformasinya
4. Spesifikasi dari metode statistik
5. Analisis data berdasarkan metode yang dipilih
6. Evaluasi dan perbandingan dari metode yang digunakan dan pilihan model akhir untuk analisis
7. Interpretasi dari model yang dipilih dan penggunaannya dalam *decision process*

B. Klasifikasi

Klasifikasi atau taksonomi adalah proses menempatkan suatu objek atau konsep kedalam satu set kategori berdasarkan objek atau konsep yang bersangkutan (Gorunescu, 2011). Ada banyak metode data mining yang digunakan untuk tujuan yang berbeda-beda. Metode klasifikasi digunakan untuk membantu dalam memahami pengelompokkan data. Klasifikasi sendiri merupakan cabang dari *discovery data mining* seperti yang ditunjukkan pada gambar 2.1 (Maimon & Rokach, 2010) .



Sumber: (Maimon & Rokach, 2010)

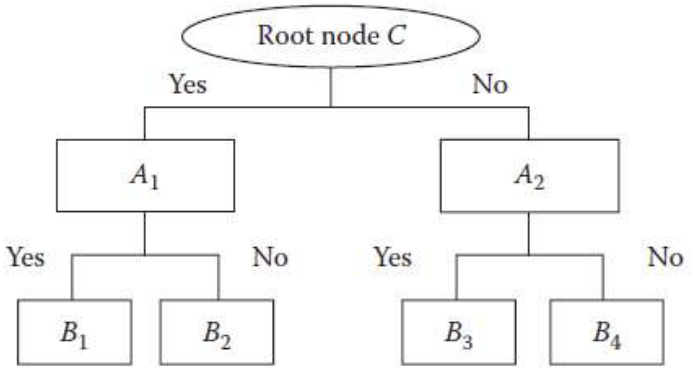
Gambar 2.1. Data mining taksonomi

C. 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.

Banyak algoritma yang dapat dipakai dalam pembentukan pohon keputusan, antara lain ID3, CART, dan C4.5 (Larose, 2006). Data dalam pohon keputusan biasanya dinyatakan dalam bentuk tabel dengan atribut dan *record*. Atribut menyatakan suatu parameter yang dibuat sebagai kriteria dalam pembentukan pohon. Misalkan untuk menentukan main tenis, kriteria yang diperhatikan adalah cuaca, angin, dan temperatur.

Seperti ditunjukkan dalam Gambar 2.2, *decision tree* tergantung pada aturan *if-then*, tetapi tidak membutuhkan parameter dan metrik. Struktur sederhana dan dapat ditafsirkan memungkinkan *decision tree* untuk memecahkan masalah atribut *multi-type*. *Decision tree* juga dapat mengelola nilai-nilai yang hilang atau data *noise* (Dua & Xian, 2011).



Sumber: (Dua & Xian, 2011

Gambar 2.2 Contoh Struktur *Decision Tree*

D. Algoritma C4.5

Algoritma C4.5 dan pohon keputusan merupakan dua model yang tak terpisahkan, karena untuk membangun sebuah pohon keputusan, dibutuhan algoritma C4.5. Di akhir tahun 1970 hingga di awal tahun 1980-an, J. Ross Quinlan seorang peneliti di bidang mesin pembelajaran mengembangkan sebuah model pohon keputusan yang dinamakan ID3 (*Iterative Dichotomiser*), walaupun sebenarnya proyek ini telah dibuat sebelumnya oleh E.B. Hunt, J. Marin, dan P.T. Stone. Kemudian Quinlan membuat algoritma dari pengembangan ID3 yang dinamakan C4.5 yang berbasis *supervised learning*.

Menurut (Witten, Frank, & Hall, 2011) serangkaian perbaikan yang dilakukan pada ID3 mencapai puncaknya dengan menghasilkan sebuah sistem praktis dan berpengaruh untuk *decision tree* yaitu C4.5. Perbaikan ini meliputi metode untuk menangani *numeric attributes*, *missing values*, *noisy* data, dan aturan yang menghasilkan *rules* dari *trees*.

Ada beberapa tahapan dalam membuat sebuah pohon keputusan dalam algoritma C4.5 (Larose, 2005) yaitu :

1. Mempersiapkan data *training*. Data *training* biasanya diambil dari data histori yang pernah terjadi sebelumnya atau disebut data masa lalu dan sudah dikelompokkan dalam kelas-kelas tertentu.
2. Menghitung akar dari pohon. Akar akan diambil dari atribut yang akan terpilih, dengan cara menghitung nilai gain dari masing-masing atribut, nilai gain yang paling tinggi yang akan menjadi akar pertama. Sebelum menghitung nilai gain

dari atribut, hitung dahulu nilai *entropy*. Untuk menghitung nilai *entropy* digunakan rumus :

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

Keterangan :
S= Himpunan kasus
n = jumlah partisi S
P_i = proporsi S_i terhadap S

Kemudian hitung nilai gain menggunakan rumus :

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

Keterangan :
S = Himpunan Kasus
A = Fitur
n = jumlah partisi atribut A
|Si| = Proporsi Si terhadap S
|S| = jumlah kasus dalam S

3. Ulangi langkah ke 2 dan langkah ke 3 hingga semua *record* terpatisi

4. Proses partisi pohon keputusan akan berhenti saat :

   a. semua *record* dalam simpul N mendapat kelas yang sama.

   b. Tidak ada atribut didalam *record* yang dipartisi lagi

   c. Tidak ada *record* didalam cabang yang kosong

E. Naïve Bayes

Salah satu metode yang sangat penting dalam klasifikasi adalah metode *Naïve Bayes*. Metode ini juga disebut *idiot's Bayes, simple Bayes, independence Bayes*. Yang menjadikan metode ini sangat penting karena metode ini sangat mudah

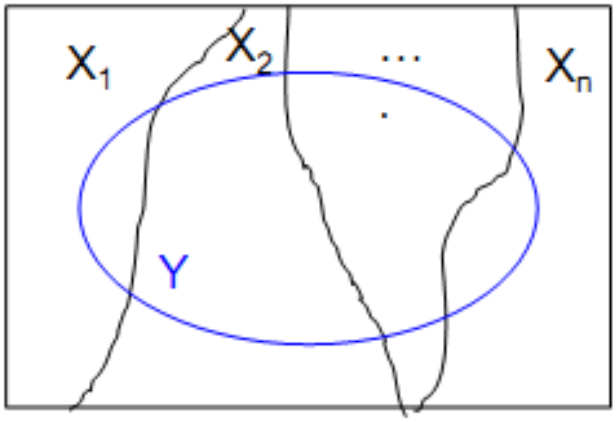
dibangun, dan tidak memerlukan skema estimasi parameter berulang yang rumit. Hal ini menunjukkan bahwa metode *Naïve Bayes* dapat diterapkan dalam data set yang besar. Selain itu metode *Naïve Bayes* sangat mudah digunakan sehingga pengguna yang tidak terampil dalam teknik klasifikasi (Wu & Kumar, 2009).

Klasifikasi Bayes didasarkan pada teorema Bayes, diambil dari nama seorang ahli matematika yang juga menteri *Prebysterian* Inggris, Thomas Bayes (1702-1761), yaitu (Bramer, 2007) dimana teorema Bayes ditulis dengan rumus sebagai berikut:

$$P(x|y) = \frac{P(y|x)\,P(x)}{P(y)}$$

Keterangan :

y = data dengan kelas yang belum diketahui
x = hipotesis data y merupakan suatu kelas spesifik
P(x | y) = probabilitas hipotesis x berdasar kondisi y (*posteriori probability*)
P(x) = probabilitas hipotesis x (*prior probability*)
P(y | x) = probabilitas y berdasarkan kondisi pada hipotesis x
P(y) = probabilitas dari y



Sumber: (Bramer, 2007)

Gambar 2.4 *Posterior Probability* (Probabilitas X_i di dalam Y) dan *Prior Probability* (Probabilitas Y di dalam X_i)

F. Evaluasi dan Validasi Metode Klasifikasi Data Mining

Evaluasi dan validasi hasil klasifikasi dengan data mining pada penelitian ini digunakan metode *Confusion Matrix* dan kurva ROC (*Receiver Operating Characteristic*).

1. Confusion Matrix

Metode ini hanya menggunakan tabel matriks seperti pada Tabel 2.1, jika dataset hanya terdiri dari dua kelas, kelas yang satu dianggap sebagai positif dan yang lainnya negatif (Bramer, 2007).

Evaluasi dengan *confusion matrix* menghasilkan nilai *accuracy*, *precison*, dan *recall*. *Accuracy* dalam klasifikasi adalah persentase ketepatan *record* data yang diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi (Han & Kamber, 2006). Sedangkan *precision* atau *confidence* adalah proporsi kasus yang diprediksi positif yang juga positif benar pada data yang sebenarnya. *Recall* atau *sensitivity* adalah proporsi kasus positif yang sebenarnya yang diprediksi positif secara benar (Powers, 2011).

Tabel 2.3 Model *Confusion Matrix*

Correct Classification	*Classified as*	
	+	-
+	*True positives*	*False negatives*
-	*False positives*	*True negatives*

Sumber: (Han & Kamber, 2006)

True Positive adalah jumlah *record* positif yang diklasifikasikan sebagai positif, *false positive* adalah jumlah *record negative* yang diklasifikasikan sebagai positif, *false negative* adalah jumlah *record* positif yang diklasifikasikan sebagai negative, *true negative* adalah jumlah *record negative* yang diklasifikasikan sebagai negative, kemudian masukkan data uji. Setelah data uji dimasukkan ke dalam *confusion matrix*, hitung nilai-nilai yang telah dimasukkan tersebut untuk dihitung jumlah *sensitivity (recall), Specifity, precision,* dan *accuracy. Sensitivity* digunakan untuk membandingkan jumlah *t_pos* terhadap jumlah *record* yang positif sedangkan *Specifity, precision* adalah perbandingan jumlah *t_neg* terhadap jumlah *record* yang negative. Untuk menghitung digunakan persamaan dibawah ini (Han & Kamber, 2006).

$$\text{Sensitifity} = \frac{t\_pos}{pos} \quad (2.5)$$

$$\text{Specifity} = \frac{t\_neg}{neg} \quad (2.6)$$

$$\text{Precision} = \frac{t\_pos}{t\_pos+f\_pos} \quad (2.7)$$

$$\text{accuracy} = \text{Sensitivity}\frac{pos}{(pos+neg)} + \text{Specifity}\frac{neg}{(pos+neg)} \quad (2.8)$$

Keterangan :

t_pos = Jumlah *true positives*
t_neg = Jumlah *true negative*
p = Jumlah *record positives*
n = Jumlah *tupel negatives*
f_pos = Jumlah *false positives*

2. Kurva ROC

Kurva ROC menunjukkan akurasi dan membandingkan klasifikasi secara visual. ROC mengekspresikan *confusion matrix*. ROC adalah grafik dua dimensi dengan *false positives* sebagai garis horizontal dan *true positive* sebagai garis

vertical (Vercellis, 2009). *The area under curve* (AUC) dihitung untuk mengukur perbedaan performansi metode yang digunakan. AUC digunakan dengan menggunakan rumus (Liao, 2007):

$$\theta^r = \frac{1}{mn} \sum_{j=1}^{n} \sum_{i=1}^{m} \psi\,(xt^r, xj^r) \tag{2.9}$$

Dimana :

$$\boldsymbol{\psi}(\mathbf{X,Y}) = \begin{cases} 1 & Y < X \\ \frac{1}{2} & Y = X \\ 0 & Y > X \end{cases} \tag{2.10}$$

Keterangan :

X = Output positif

Y = Output negatif

G. Tinjauan Studi

Beberapa peneliti telah melakukan proses penelitian terdahulu yang terkait dengan tema dan data mining, diantaranya:

1. Simulasi Pendukung Keputusan Penerima Beasiswa Menggunakan Metode *Composite Performance Index* (Karismariyanti, 2011)
   a. Latar Belakang Masalah: Keputusan yang diambil dari penentuan penerima beasiswa hanya bergantung pada pengambil keputusan saja.
   b. Tujuan Penelitian: Membuat sistem yang mampu mengolah data alternatif-alternatif keputusan yaitu Sistem Pendukung Keputusan (*Decision Support System*)

c. Metodologi: Penelitian ini menggunakan metode *Composite Performance Index*.

d. Hasil Penelitian: Sistem pendukung keputusan yang dibangun berupa simulasi dari data yang telah dimasukkan berdasarkan kriteria yang telah dipilih oleh pengambil keputusan. Nilai alternatif berubah-ubah menyesuaikan bobot data

2. *A Decision Support Prototype Tool for Predicting Student Performance in an ODL Environment* (Kotsiantis & Pintelas, 2004)

   a. Latar Belakang: Pada lembaga-lembaga pembelajaran jarak jauh *dropout* siswa sering sekali terjadi. Alat yang dapat secara otomatis mengenali siswa dengan kemungkinan *dropout* yang tinggi cukup berguna untuk guru/pengajar mengambil tindakan pencegahan dan konsultasi untuk mengurangi jumlah siswa *dropout*.

   b. Tujuan Penelitian: Membangun perangkat lunak untuk memprediksi kinerja siswa untuk mengetahui mahasiswa dengan kinerja belajar rendah dan berpotensi tinggi *dropout*.

   c. Metodologi: Pembangunan perangkat lunak untuk memprediksi siswa menggunakan algoritma *Decision Tree, Neural Networks, Naïve Bayes, Instance-Based Learning, Rule-Based Learning, dan Support Vector Machine.*

   d. Hasil Penelitian: Hasil penelitian berupa Machine Learning berbasis *decision support* untuk prediksi performance mahasiswa dengan hasil prediksi *continue* atau *dropout*. Data diambil dari Universitas Terbuka Hellenic. Set

data yang diuji antara lain: *sex, age, marital status, number of children, occupation, computer knowledge, job associated with computers*.

3. Sistem Pendukung Keputusan Cerdas Dalam Penentuan Penerima Beasiswa (Santiary, 2012)
    a. Latar Belakang : Penentuan penerima beasiswa secara manual menyebabkan pengelolaan data beasiswa yang tidak efisien terutama dari segi waktu dan banyaknya perulangan proses.
    b. Tujuan Penelitian: Membuat sistem pendukung keputusan cerdas dalam penentuan penerima beasiswa.
    c. Metodologi: Penelitian ini menggunakan *Fuzzy Multiple Attribute Decision Making* (FMADM) dengan metode *Simple Additive Weighting* (SAW).
    d. Hasil Penelitian: Sistem pendukung keputusan untuk membantu menentukan penerima beasiswa dengan menggunakan logika fuzzy FMADM dengan menggunakan metode SAW dapat mempercepat proses penentuan penerima beasiswa dengan perhitungan yang akurat dalam memberikan rekomendasi.
4. *Predicting Students Drop Out: A Case Study* (Dekker, *et all*, 2009)
    a. Latar Belakang Masalah: Tingkat mahasiswa *dropout* pada departemen Teknik Elektro di Universitas Teknik Eindhoven mencapai 40%. Mendeteksi kelompok mahasiswa beresiko pada tahap awal sangat penting untuk menjaga mahasiswa dari kemungkinan *dropout*. Hal ini memungkinkan pihak institusi untuk memberikan pengarahan kepada mahasiswa yang membutuhkan.
    b. Tujuan Penelitian: Membuat klasifikasi mahasiswa beresiko *dropout* pada tahun pertama kuliah.

c. Metodologi: Penelitian ini menggunakan metode klasifikasi popular dari Weka, dengan membandingkan dua algoritma *Decision Tree* yaitu *Simple Cart* dan C4.5, *Bayesian Classifier (BayesNet), Logistic Model (Simple Logistic), Rule-based learner (JRip) dan Random Forest*.

d. Hasil Penelitian: Penelitian ini mengklasifikasi mahasiswa yang beresiko *dropout* pada semester awal di Departemen Teknik Elektro Universitas Teknologi Eindhoven periode 2000-2009 sejumlah 648 mahasiswa baru dengan menggunakan algoritma Simple Cart, C.45(J48), *Bayesian Network, Simple Logistic, JRip, dan Random Forest*. Hasil dari penelitian ini menunjukkan tingkat akurasi dari semua algoritma yang digunakan antara 75% sampai dengan 80%.

H. Tinjauan Obyek Penelitian

AMIK "BSI Yogyakarta" merupakan salah satu perguruan tinggi di Yogyakarta yang beralamatkan di Jl. Ringroad Barat, Ambarketawang, Gamping, Sleman, Yogyakarta. Perguruan tinggi ini mengelola program pendidikan Diploma Tiga (D3) jurusan Manajemen Informatika. Awal didirikan pada tahun 1993 di Jalan Wates km 3 Kalibayem, Kasihan, Bantul, Yogyakarta perguruan tinggi ini baru membuka Program Pendidikan Satu Tahun. Pada tahun 2004 AMIK "BSI Yogyakarta" mengambil alih ijin penyelenggaraan Akademi Manajemen Informatika dan Komputer dari AMIK "Proactive", sejak saat itulah AMIK "BSI Yogyakarta" membuka Program Pendidikan Diploma Tiga. Kemudian pada tahun 2010 AMIK "BSI Yogyakarta" menempati gedung baru yang berada di Jalan Ringroad Barat, Ambarketawang, Gamping, Sleman, Yogyakarta.

Sistem perkuliahan di AMIK "BSI Yogyakarta" sudah didukung oleh teknologi informasi yang bagus dari sistem pendaftaran mahasiswa baru, pembayaran, informasi akademik, sampai ujian sudah dilakukan secara *online* yaitu dengan menggunakan laman www.bsi.ac.id

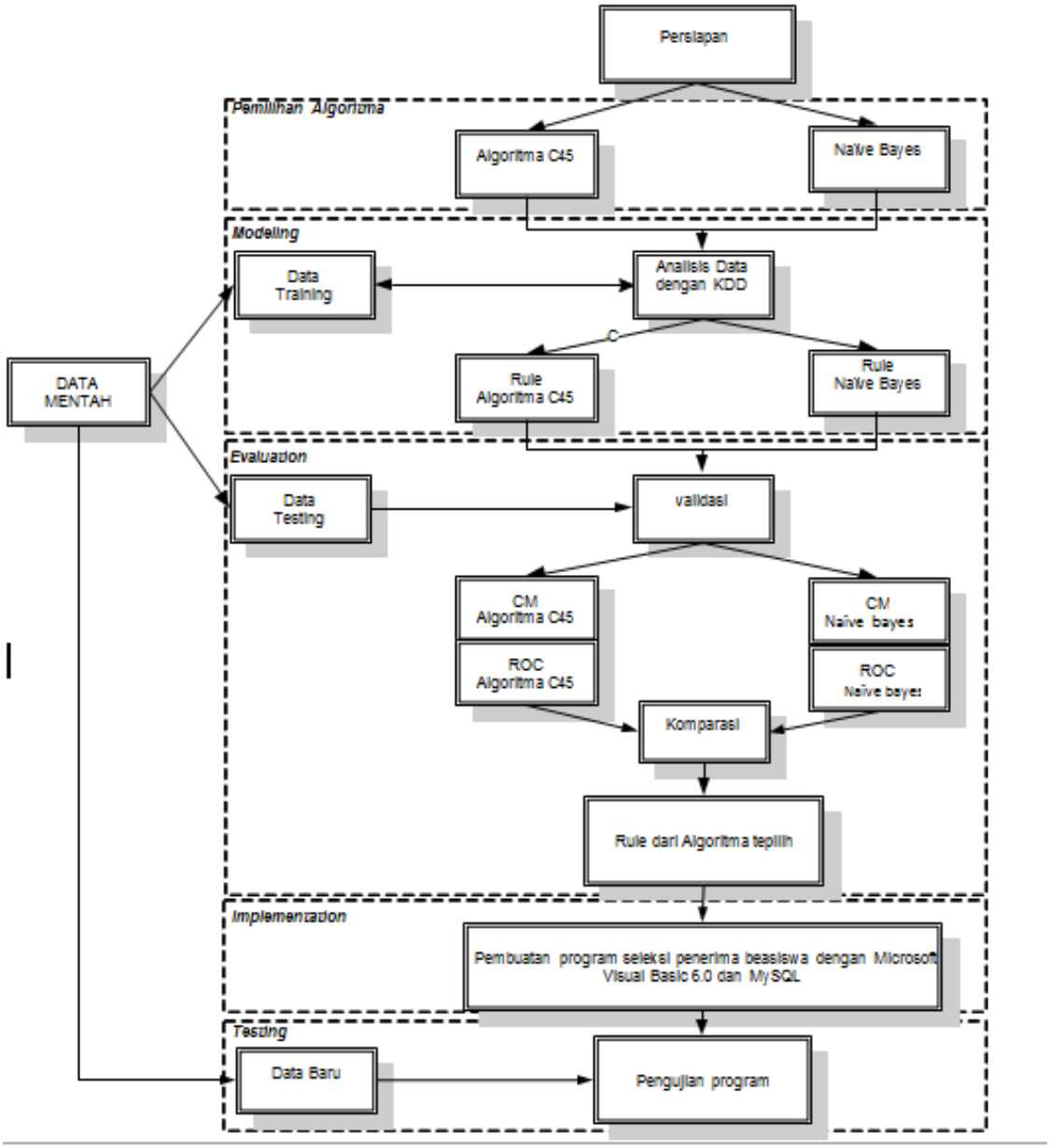
AMIK "BSI Yogyakarta" memiliki fasilitas dan sarana akademik berupa tiga ruang kelas teori yang aktif digunakan, ruang praktikum sebanyak dua laboratorium komputer yang dilengkapi dengan fasilitas internet. Kampus ini juga dilengkapi oleh *hotspot area* yang bisa diakses oleh dosen, karyawan, dan mahasiswa.

I. Kerangka Pemikiran

Permasalahan dalam penelitian ini pengolahan data masih secara manual yang mengakibatkan seleksi penerima beasiswa tidak tepat sasaran serta meningkatnya jumlah mahasiswa yang berminat mengajukan beasiswa. Tujuan penelitian ini untuk membuat klasifikasi mahasiswa yang berpotensi menerima beasiswa. Untuk mencapai tujuan tersebut maka dilakukan klasifikasi dengan dua algoritma yaitu C4.5 dan *Naïve Bayes*. Data mahasiswa yang diklasifikasi adalah data mahasiswa AMIK "BSI Yogyakarta". Data yang diambil untuk klasifikasi sejumlah 133 data mahasiswa. Kemudian data mentah tersebut dibagi menjadi dua yaitu data training dan data testing.

Data *training* diolah dengan Sembilan langkah di *Knowledge Discovery in Databases* (KDD). Hasilnya berupa *rule* klasifikasi mahasiswa dari algoritma C4.5, dan *Naïve Bayes*. Setelah itu *rule* klasifikasi mahasiswa dari algoritma tersebut di validasi dengan menggunakan data *testing* yang hasilnya berupa

Confusion Matrix dan kurva ROC dari validasi *rule* algoritma C4.5, dan *Naïve Bayes*. Dari hasil tersebut dipilih algoritma dengan tingkat akurasi tertinggi. Untuk penerapan digunakan data baru dengan menggunakan algoritma terpilih, kemudian dievaluasi. Proses pada kerangka pemikiran ini seperti pada gambar 2.6



Gambar 2.6 Kerangka Pemikiran

J. Hipotesis

Hipotesis merupakan jawaban sementara dari pertanyaan penelitian. Biasanya hipotesis dirumuskuan dalam bentuk hubungan antara dua variabel yaitu variabel bebas dan variabel terikat. Hipotesis berfungsi untuk menentukan ke arah pembuktian, artinya hipotesis merupakan pernyataan yang harus dibuktikan. Dalam penelitian ini ada beberapa hipotesis yang dapat dijabarkan dalam hipotesis mayor dann hipotesis minor. Hipotesis mayor lebih bersifat umum sedangkan hipotesis minor lebih bersifat khusus.

1. Hipotesis Mayor (Umum)

   Tingginya tingkat mahasiswa penerima beasiswa ditentukan berdasarkan prestasi akademik.

2. Hipotesis Minor (Khusus)

   a. Semakin tinggi IPK mahasiwa semakin besar peluang mendapatkan beasiswa

   b. Makin rendah penghasilan orangtua/wali semakin tinggi potesi mendapatkan beasiswa.

3. METODE PENELITIAN

A. Kerangka Pendekatan Penelitian

Penelitian untuk mengklasifikasi mahasiswa penerima beasiswa PPA dan BBM di AMIK “BSI Yogyakarta” ini juga termasuk penelitian kuantitatif karena pada penelitian ini meneliti secara ilmiah dengan cara sistematis terhadap fenomena penerimaan beasiswa kopertis (PPA dan BBM) serta mencari tingkat akurasi hubungan antara data nilai IPK mahasiswa yang mengajukan beasiswa, penghasilan orang tua dengan tingkat potensi penerimaan beasiswa kopertis (PPA dan BBM).

B. Model dan Variabel

Data mahasiswa di AMIK "BSI Yogyakarta" tahun angkatan 2008 sampai dengan 2013 terdapat 133 mahasiswa dengan mahasiswa yang mengajukan beasiswa, untuk siswa yang mendapatkan beasiswa kopertis sebanyak 60 mahasiswa baik dalam bentuk beasiswa BBM ataupun beasiswa PPA, Sampai saat ini belum diketahui algoritma yang paling akurat dalam melakukan klasifikasi mahasiswa yang berpotensi diterima dalam beasiswa kopertis (PPA dan BBM). Oleh karena itu dalam penelitian ini akan dilakukan komparasi algoritma C4.5, dan *Naïve Bayes*, untuk mengetahui algoritma yang paling akurat dalam klasifikasi mahasiswa yang berpotensi di terima beasiswa.

Atribut dan nilai atribut diperoleh dari tabel data mahasiswa dan data nilai. Adapun atribut yang digunakan dalam penelitian ini antara lain:

Tabel 3.1 Atribut nilai dan katageri

No	Atribut	Nilai	Nilai Baru
1	Semester	2	2
		4	4
		6	6
2	IPK	1.00 - 1.99	1.00 - 1.99
		2.00 - 2.75	2.00 - 2.75
		2.76 - 3.50	2.76 - 3.50
		3.51 - 4.00	3.51 - 4.00
3	Orang Tua	Ada	Ada
		Wafat	Yatim
4	Penghasilan	< 1500000	Rendah
		1500000 - 2500000	Sedang
		2500000 - 3500000	Tinggi
		> 3500000	Sangat Tinggi
5	Beasiswa	Tidak	Tidak
		Penerima	Penerima
6	Berkas	Lengkap	Lengkap
		Tidak Lengkap	Tidak Lengkap
7	remark	DAPAT	DAPAT
		TIDAK	TIDAK

C. Populasi dan sempel penelitian

Tahap pertama dalam penelitian ini adalah pemilihan sampel dengan mengidentifikasi populasi target, yaitu populasi spesifik yang relevan dengan tujuan penelitian atau masalah dalam penelitian. Populasi dalam penelitian ini adalah mahasiswa AMIK "BSI Yogyakarta".

Sedangkan sampel merupakan bagian dari elemen-elemen populasi yang hendak diteliti. Dalam penelitian ini sampel yang diambil adalah mahasiswa AMIK "BSI Yogyakarta" yang mengajukan permohonan beasiswa kopertis (BBM dan PPA) dari tahun 2008 sampai dengan tahun 2013. Dalam pengambilan sampel dilakukan data *validation* yang bertujuan untuk mengidentifikasi dan menghapus data yang ganjil (*outlier/noise*), data yang tidak konsisten dan data yang tidak lengkap (*missing value*) hasilnya diambil sebanyak 133 data mahasiswa yang mengajukan beasiswa Kopertis baik yang diterima maupun yang ditolak.

4. HASIL DAN PEMBAHASAN

A. Hasil

Penelitian ini bertujuan untuk menentukan akurasi kelayakan pemberian beasiswa yang dibandingkan dengan menggunakan metode algoritma C4.5,dan *naïve bayes*, Setelah itu membandingkan nilai akurasi dari kedua metode tersebut, dalam menentukan hasil penelitian ini menggunakan data *training* berjumlah 106 data dan data *testing* berjumlah 26.

1. Algoritma C4.5

Langkah-langkah untuk membuat algoritma C.45 dengan memakai data *training* yang berjumlah 106, yaitu :

a. Siapkan data *training* yaitu tabel 3.3 yang berjumlah 106 data.

b. Hitung jumlah mahasiswa penerima beasiswa kopertis (PPA dan BBM) yang gagal berdasarkan nilai tiap atribut.

c. Hitung nilai *entropy* total dimana diketahui penerima beasiswa kopertis sebanyak 46 mahasiswa dan yang gagal berjumlah 60.

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

=((-46/106) * $\log_2$ (46/106))+((-60/106) * $\log_2$ (60/106))
= 0.98738

d. Hitung nilai *gain* untuk masing-masing atribut. Kemudian tentukan nilai gain tertinggi. Atribut dengan nilai *gain* tertinggi maka atribut tersebut dijadikan sebagai akar. Sebagai contoh hitung nilai *gain* untuk atribut Semester yaitu:

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

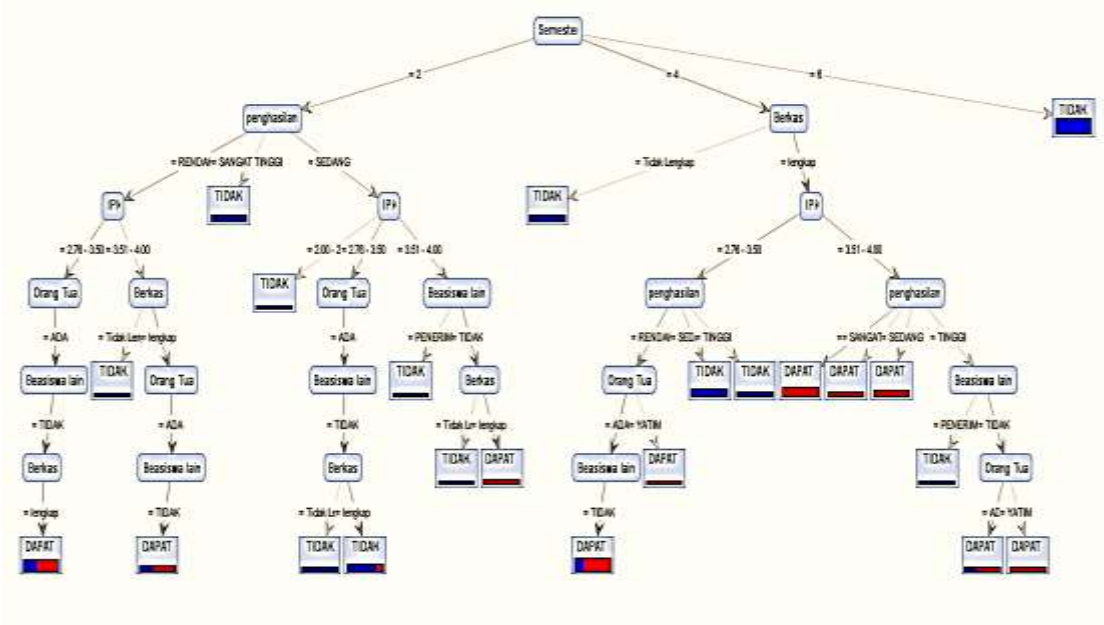
=0.98738- ((36/106*0.96408)+(52/106*0.96124) +(18/106*0))
= **0.188407**

Perhitungan nilai *entropy* dan *gain* untuk semua atribut dilakukan untuk mendapatkan nilai *gain* tertinggi yang akan dijadikan sebagai akar. Hasil perhitungannya terlihat di tabel di bawah ini:

Tabel 4.1 Hasil nilai *entropy* dan *gain* untuk menentukan simpul akar dengan data *training*

Node			Jumlah Kasus (S)	DAPAT	TIDAK	Entropy	Gain
1	Total		106	46	60	0.98738	
	Semester						0.188407
		2	36	14	22	0.96408	
		4	52	32	20	0.96124	
		6	18	0	18	0	
	IPK						0.0956666
		2.00 - 2.75	4	0	4	0	
		2.76 - 3.50	61	20	41	0.91273	
		3.51 - 4.00	41	26	15	0.94744	
	Orang Tua						0.0839632
		ADA	99	39	60	0.96729	
		YATIM	7	7	0	0	
	Penghasilan						0.1159426
		RENDAH	50	32	18	0.94268	
		SEDANG	36	9	27	0.81128	
		TINGGI	10	3	7	0.88129	
		SANGAT TINGGI	10	2	8	0.72193	
	Beasiswa						0.0317983
		TIDAK	102	46	56	0.99306	
		PENERIMA	4	0	4	0	
	Berkas						0.0740404
		LENGKAP	97	46	51	0.99808	
		TIDAK LENGKAP	9	0	9	0	

Pembentukan simpul-simpul dengan perhitungan *gain* diperoleh *decision tree* untuk klasifikasi mahasiswa penerima beasiswa seperti pada gambar di bawah ini:



2. Algoritma Naïve Bayes

Data *training* yang digunakan untuk metode *naïve bayes* menggunakan data pada table 3.3. Dengan mencari *prior probability* untuk nilai yang diterima dan tidak diterima untuk semua jumlah data. Jika diketahui dalam data *training*, jumlah data 106, siswa yang diterima beasiswa dalam kelas DAPAT 46 *record* dan yang tidak diterima dalam kelas TIDAK 60 record. Berikut hasil perhitungan *prior probability* dengan menggunakan rumus (3.3) dan (3.4) :

P(DAPAT,n) = 46/106 = 0.434

P(TIDAK,n) = 60/106= 0.566

Setelah itu mencari masing-masing setiap *class* atribut. Berikut hasil perhitungan *prior probability* untuk semester 2 dalam katagori DAPAT:

P(2,DAPAT) = 14/36 = 0.3889

Berikut hasil perhitungan *priori probability* untuk masing-masing atribut, terdapat pada tabel dibawah ini.

Tabel 4.2 Hasil nilai *prior probability* dengan data *training*

		Kasus	DAPAT	TIDAK	p(x|C1)	
					DAPAT	TIDAK
TOTAL		106	46	60	0.434	0.566
Semester						
	2	36	14	22	0.3889	0.6111
	4	52	32	20	0.6154	0.3846
	6	18	0	18	0	1
IPK						
	2.00 - 2.75	4	0	4	0	1
	2.76 - 3.50	61	20	41	0.3279	0.6721
	3.51 - 4.00	41	26	15	0.6341	0.3659
Orang Tua						
	Ada	99	39	60	0.3939	0.6061
	Yatim	7	7	0	1	0
Penghasilan						
	RENDAH	50	32	18	0.64	0.36
	SEDANG	36	9	27	0.25	0.75
	TINGGI	10	3	7	0.3	0.7
	SANGAT TINGGI	10	2	8	0.2	0.8
Beasiswa						
	Tidak	102	46	56	0.451	0.549
	Penerima	4	0	4	0	1
Berkas						
	Lengkap	97	46	51	0.4742	0.5258
	Tidak lengkap	9	0	9	0	1

B. Pembahasan

Hasil dari pengujian model yang telah dilakukan, dilakukan pengujian tingkat akurasi dengan menggunakan confussion matrix dan kurva ROC/AUC (*Area Under Cover*).

1. *Confussion Matrix*

Tabel 4.3 adalah perhitungan akurasi data training menggunakan algoritma C4.5. Diketahui dari 106 data training, dengan menggunakan metode algoritma C4.5 didapat 48 data prediksi tidak sesuai dengan tidak, 1 prediksi tidak ternyata Dapat, 12 data prediksi Dapat ternyata tidak , dan 45 data prediksi dapat sesuai dengan Dapat.

Tabel 4.3 *Confussion Matrix* data training Untuk Algoritma C4.5

accuracy: 87.74%			
	true TIDAK	true DAPAT	class precision
pred. TIDAK	48	1	97.96%
pred. DAPAT	12	45	78.95%
class recall	80.00%	97.83%	

Tabel 4.4 adalah perhitungan akurasi data training menggunakan *naïve bayes*. Diketahui dari 106 data training, dengan menggunakan metode *naïve bayes* didapat 48 data prediksi tidak sesuai dengan Tidak, 4 prediksi Tidak ternyata Dapat, 12 data prediksi Dapat ternyata tidak, dan 42 data predisi Dapat sesuai dengan Dapat.

Tabel 4.4 *Confussion Matrix* Data Training Untuk *Naïve Bayes*

accuracy: 84.91%			
	true TIDAK	true DAPAT	class precision
pred. TIDAK	48	4	92.31%
pred. DAPAT	12	42	77.78%
class recall	80.00%	91.30%	

2. Evaluasi dengan Kurva ROC

ROC memiliki tingkat nilai diagnosa yaitu(Gorunescu, 2011):

Akurasi bernilai 0.90 – 1.00 = *excellent classification*
Akurasi bernilai 0.80 – 0.90 = *good classification*
Akurasi bernilai 0.70 – 0.80 = *fair classification*
Akurasi bernilai 0.60 – 0.70 = *poor classification*
Akurasi bernilai 0.50 – 0.60 = *failure*

Hasil yang didapat dari pengolahan ROC untuk algoritma C4.5 dengan menggunakan data training sebesar 0.970 dapat dilihat pada gambar 4.2 dengan tingkat diagnosa *excellent classification*

Gambar 4.2 Grafik ROC dalam algoritma C4.5 dengan data *training*

C. Analisis Evaluasi Komparasi Model

Dari hasil pengujian diatas, dengan dilakukan evaluasi. *Confussion Matrix* dengan memperhatikan kolom training pada akurasi, algoritma C.45 memiliki tingkat akurasi yang paling tinggi dengan tingkat akurasi 87.74%, Sehingga dapat dikatakan algoritma tersebut adalah algoritma terbaik yang dapat digunakan untuk penentuan kelayakan penerimaan beasiswa kopertis

Berdasarkan kolom ROC *Confussion Matrix* pada training algoritma C4.5 memiliki tingkat ROC paling tinggi, yaitu 0.970 termasuk dalam kategori *excellent classification* dan pada testing algoritma C.45 memiliki tingkat ROC yaitu 0.883, termasuk dalam katagori *good classification.*

Dengan menggunakan perbandingan data training dengan data testing, yaitu 80 berbanding 20, maka untuk akurasi dapat dilihat dalam tabel 4.15:

Tabel 4.15 Perbandingan Akurasi

Metode	Akurasi		Perbandingan Akurasi
	Training	Testing	
Algoritma C4.5	87.74%	81.48%	86.88%
Naïve Bayes	84.91%	85.19%	84.96%

Oleh karena itu, berdasarkan tabel 4.13 Algoritma C.45 yang memiliki tingkat akurasi yang paling tinggi, sehingga baik digunakan untuk penentuan kelayakan pemberian beasiswa dengan persentase 86.88%.

4. KESIMPULAN

Beberapa kesimpulan yang dapat diambil dari hasil penelitian yang dilakukan adalah:

1. Dalam penelitian ini dilakukan pembuatan model menggunakan Algoritma C4.5 dan *Naïve Bayes* menggunakan data mahasiswa yang mengajukan beasiswa. Model yang dihasilkan, dikomparasi untuk mengetahui algoritma yang paling baik dalam penentuan penerima beasiswa kopertis. Untuk mengukur kinerja ketiga algoritma tersebut digunakan metode pengujian *Confusion Matrix* dan Kurva ROC. Diketahui bahwa algoritma C4.5 memiliki nilai *accuracy* paling tinggi, diikuti oleh metode N*aïve bayes*

2. Diketahui bahwa algoritma C4.5 memiliki akurasi tinggi sehingga dapat dikatakan bahwa metode tersebut merupakan algoritma yang paling baik dalam pengklasifikasian data.
3. Diketahui bahwa penerapan data baru menggunakan Algoritma C4.5 dan pemrograman visual basic menghasilkan data yang sesuai dengan prediksi lebih besar di banding dengan yang tidak sesuai dengan prediksi, sehingga dapat dikatakan bahwa program tersebut tersebut dapat digunakan untuk penentuan mahasiswa penerima beasiswa kopertis

Berdasarkan kesimpulan yang diambil dari hasil penelitian, maka saran untuk penelitian ini adalah sebagai berikut:

1. Penelitian ini dapat dikembangkan dengan algoritma yang lain seperti *Neural Network, Statistical Analysis, Genetic Algorithms, Rough Sets, Rule-based methods, Memory based reasoning, Support vector machine* dan lain sebagainya dan menambahkan atribut-atribut baru.
2. Penelitian semacam ini dapat dikembangkan pada unit lain seperti prediksi calon penerima beasiswa prestasi yang ada di BSI
3. Untuk penelitian selanjutnya dengan permasalahan yang sama dan dengan metode yang sama dapat ditingkatkan salah satunya dengan melakukan *pruning* terhadap algoritma C4.5 sehingga pohon yang terbentuk tidak terlalu besar bahkan mungkin untuk jumlah data yang besar sekalipun. Ini dilakukan untuk mengefisienkan kinerja dari algoritma C4.5 tanpa mengurangi keakuratannnya.

5. DAFTAR PUSTAKA

[1] Bramer, M. (2007). *Principles of Data Mining.* United Kingdom: Springer.
[2] Deker, G. W., Pechenizkiy, M., Vleeshouwers, J. M,. (2009). Predicting Students Drop Out: A Case Study.
[3] Giudici, P., & Figini, S. (2009). *Applied Data Mining for Business and Industry Second Edition.* United Kingdom: John Willey and Sons.
[4] Gorunescu, F. (2011). *Data Mining Concepts, Models and Techniques.* Berlin: Springer.
[5] Han, J., & Kamber, M. (2006). *Data Mining Concepts And Techniques 2nd Edition.* San Fransisco: Elsevier.
[6] Jones, P. (2001). *Visual Basic: A Complete Course.* London: Continuum.
[7] Karismariyanti, M. (2011). Simulasi Pendukung Keputusan Penerima Beasiswa Menggunakan Metode Composite Performance Index. *Jurnal Teknologi Informasi Vol.1* , 54-59.
[8] Kothari, C. R. (2004). *Research Methodology Methods and Techniques, Second Revised Edition.* New Delhi: New Age International Publishers.
[9] Kotsiantis, S., & Pintelas. P. (2004). A Decision Support Prototype Tool for Predicting Student Performance in an ODL Environment. Greece: University of Patras.
[10] Larose, D. T. (2005). *Discovering Knowledge in Data An Introduction to Data Mining.* New Jersey: John Willey and Sons.
[11] Liao, T. W. (2007). Enterprise Data Mining: A Review and Research Directions. *Recent Advances in Data Mining of Enterprise Data: Algorithms and Applications* , 1-109.
[12] Liberty, J. (2005). *Programming Visual Basic 2005.* USA: O'Reilly Media.
[13] Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook Second Edition.* London: Springer.
[14] Powers, D. (2011). Evaluation: From Precision, Recall and F-Measure To ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies* , 37-63.
[15] Santiary, P. A. (2012). Sistem Pendukung Keputusan Cerdas dalam Penentuan Beasiswa. *Jurnal Logic Vol.12 No.2* , 87-91.
[16] Suryana, T. (2009). *Visual Basic.* Yogyakarta: Graha Ilmu.
[17] Vercellis, C. (2009). *Business Intelligence.* United Kingdom: John Wiley and Sons.
[18] Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining Practical Machine Learning Tools and Techniques Third Edition.* Burlington: Elsevier.
[19] Wu, X., & Kumar, V. (2009). *The Top Ten Algorithms in Data Mining.* USA: CRC Press.
[20] www.dikti.go.id/.januari-13,2014. http://www.dikti.go.id/files/Lemkerma/kepmen232-2000.txt