PEMBENTUKAN MODEL KLASIFIKASI DATA LAMA STUDI MAHASISWA STMIK INDONESIA MENGGUNAKAN *DECISION TREE* DENGAN ALGORITMA NBTREE

Syam Gunawan[1], Pritasari Palupiningsih[2]
[1,2] Program Studi Sistem Informasi, STMIK Indonesia
e-mail: [1]syam@stmik-indonesia.ac.id, [2]prita@stmik-indonesia.ac.id

ABSTRAK

Salah satu kriteria penilaian pada akreditasi program studi adalah penilaian terhadap lama studi mahasiswa yang lulus tepat waktu.tidak sedikit mahasiswa yang menempuh masa studi melebihi standar kelulusan yang telah ditetapkan. Sehingga penting bagi program studi untuk mengetahui mahasiswa mana saja yang memiliki kemungkinan lulus tidak tepat waktu. Untuk itu diperlukan adanya prediksi lama studi mahasiswa. Salah satu cara untuk dapat memprediksi lama studi mahasiswa adalah dengan membangun model klasifikasi. Penelitian ini bertujuan untuk membangun model prediksi lama studi mahasiswa menggunakan Decision Tree dengan algoritme NBTree. Data yang digunakan adalah data nilai akademik serta data cuti akademik mahasiswa. Hasil yang diperoleh adalah model klasifikasi berupa Naïve Bayes Decision Tree dengan akurasi 73,45%.

Kata Kunci*:Lama Studi Mahasiswa, Decision Tree, Algoritma NBTree*

1. PENDAHULUAN

Program Studi merupakan garda terdepan dalam penyelenggaraan pendidikan dari sebuah Perguruan Tinggi, sehingga program studi senantiasa melakukan evaluasi guna meningkatkan mutu dan efisiensi perguruan tinggi termasuk peningkatan kualitas lulusan. Selain itu, salah satu kriteria penilaian pada akreditasi program studi adalah penilaian terhadap lama studi mahasiswa yang lulus tepat waktu. Masa studi mahasiswa telah diatur dalam ketetapan Kementerian Pendidikan dan Kebudayaan Direktorat Jenderal Pendidikan Tinggi tentang Sistem Pendidikan Tinggi yang menyebutkan bahwa untuk memenuhi standar kompetensi lulusan bagi mahasiswa program sarjana (S1) beban wajib yang harus ditempuh adalah paling sedikit 144 - 160 satuan kredit semester (sks) dengan masa studi selama 8 - 10 semester atau 4 – 5 tahun.

Pada institusi pendidikan perguruan tinggi, data mahasiswa dan data jumlah kelulusan mahasiswa dapat menghasilkan informasi yang berlimpah berupa jumlah kelulusan setiap tahunnya, profil, dan hasil akademik mahasiswa selama menempuh proses kegiatan belajar mengajar di perguruan tinggi. Adanya informasi mengenai lama studi mahasiswa tentu akan menjadi pendukung suatu pengambilan keputusan yang tepat bagi manajemen Perguruan Tinggi dalam mengambil langkah berikutnya.

Permasalahan yang sering terjadi adalah masih banyaknya jumlah mahasiswa yang lulus dengan masa studi melampaui waktu yang telah ditetapkan atau tidak tepat waktu. Hal ini dapat mempengaruhi mutu lulusan Perguruan Tinggi. Sehingga penting bagi program studi untuk mengetahui mahasiswa mana saja yang memiliki kemungkinan lulus tidak tepat waktu. Untuk itu diperlukan adanya prediksi lama studi mahasiswa. Salah satu cara untuk dapat memprediksi lama studi mahasiswa adalah dengan membangun model klasifikasi. Selama ini STMIK Indonesia belum memiliki model klasifikasi lama studi mahasiswa yang dapat digunakan sebagai prediksi jumlah lulus tepat waktu. Padahal data mahasiswa sangat berlimpah, hanya saja data-data tersebut belum dimanfaatkan untuk dianalisis lebih jauh.

Data lama studi mahasiswa berukuran besar dapat dianalisis menggunakan teknik klasifikasi. Salah satu metode dalam klasifikasi adalah *decision tree* yang akan menghasilkan model klasifikasi. Salah satu algoritma yang dapat diterapkan dalam metode *decision tree* adalah NBTree. Model klasifikasi yang terbentuk akan dapat digunakan dalam prediksi. Penelitian ini bertujuan untuk menerapkan metode *decision tree* dengan algoritma NBTree untuk membentuk model klasifikasi. Hasil dari penelitian ini adalah terbentuknya model klasifikasi data lama studi mahasiswa STMIK Indonesia yang nantinya dapat digunakan untuk prediksi jumlah mahasiswa lulus tepat waktu.

2. TINJAUAN PUSTAKA

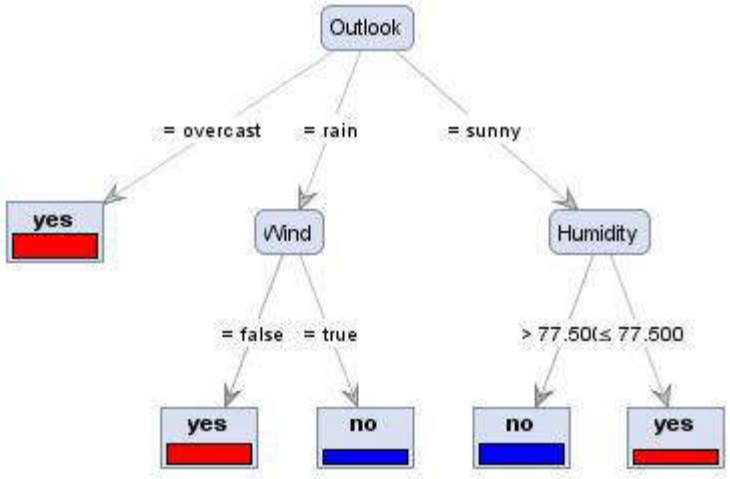
2.1 Decision Tree

Decision Tree adalah sebuah struktur pohon, dimana setiap *node* pohon merepresentasikan atribut yang telah diuji, setiap cabang merupakan suatu pembagian hasil uji, dan *node* daun merepresentasikan kelompok kelas tertentu. Level *node* teratas dari sebuah *decision Tree* adalah node akar (*root*) yang biasanya berupa atribut yang paling memiliki pengaruh terbesar pada suatu kelas tertentu. Pada umumnya *decision Tree* melakukan strategi pencarian secara *top-down* untuk solusinya. Pada proses mengklasifikasi data yang tidak diketahui, nilai atribut akan diuji dengan cara melacak jalur dari *node* akar (*root*) sampai *node* akhir (daun) dan kemudian akan diprediksi kelas yang dimiliki oleh suatu data baru tertentu. [1]

Terdapat 3 jenis node yang terdapat pada decision tree, yaitu:

a. *Root node*, merupakan *node* paling atas, pada*node*ini tidak ada *input* dan bisa tidak mempunyai *output* atau mempunyai *output* lebih dari satu.
b. *Internal Node*, merupakan *node* percabangan. Pada *node* ini terdapat percabangan. Pada *node* ini terdapat satu *input* dan memiliki *output* mininmal dua.
c. *Leaf node* atau terminal *node*, merupakan *node* akhir. Pada *node* ini terdapat satu *input* dan tidak mempunyai *output*.

Contoh decision tree dapat dilihat dari Gambar 1berikut :



Gambar 1. Contoh *Decision Tree*

2.2 Algoritma NBTree

Salah satu algoritme pembentukan *decision tree* adalah algoritme NBTree. Algoritme NBTree merupakan algoritme hasil penggabungan teknik *decision treeclassifier* dengan *naïve-bayes classifier*. Algoritme ini akan membangun *decision tree* dengan *node* yang mengandung *univariatesplit* seperti *decision tree* biasa, tetapi pada node *leaf* terkandung *naïve-bayes classifier.* [2]

Algoritma NBTree (Kohavi 1996)

Input : himpunan T yang terdiri dari *instance* dengan label
Output : sebuah pohon keputusan dengan pengkategori Naïve-bayes pada daun

a. Hitung *utility* untuk setiap atribut $\mathbf{X_i}$, $\mu(X_i)$. Untukatribut kontinyu, dibuat sebuah *threshold.*
b. Misalkan $\mathbf{j} = \arg\max_i(\mu_i)$, adalah atribut dengan nilai *utility* tertinggi.
c. Jika μ_j tidak lebih tinggi dibanding nilai *utility* yang dimiliki *node* yang sekarang, buat model Naïve-Bayes untuk *node* yang sekarang dan kembali ke langkah 1.
d. Bagi *T* menurut pengujian di $\mathbf{X_j}$. Jika $\mathbf{X_j}$ adalah kontinyu, sebuah pembagian menggunakan *threshold* dibuat untuk semua nilai yang mungkin.
e. Untuk setiap *child*, panggil algoritme secara rekursif untuk membagi *T* yang sesuai dengan pengujian dari *child.*

Dengan memberikan sekumpulan *instance* ke suatu *node*, algoritme NBTree akan melakukan evaluasi *utility of split* untuk setiap atribut. Jika *utility* terbesar dari semua atribut lebih tinggi dibanding *utility* yang dimiliki node yang sekarang, maka akan dilakukan pembagian *instance-instance* yang adaberdasarkan atribut tersebut. [2]

Utilityof node dihitung dengan melakukan diskretisasi pada data yang ada dan menghitung estimasi akurasi 5-*fold cross validation* dari penggunaan naïve-bayes di *node* tersebut. Sedangkan *utility of split* adalah jumlah bobot dari *utility of node*, dimana bobot yang diberikan ke sebuah node sebanding dengan jumlah *instance* yang diturunkan *node* tersebut. Pembagian ditetapkan signifikan jika reduksi relatif terhadap kesalahan lebih bagus dari 5% dan setidaknya terdapat 30 *instance* di *node* tersebut. Hal ini untuk menghindari terjadinya pembagian dengan nilai yang kecil.[2]

2.3 Naïve Bayes Classifier

Klasifikasi Naive Bayes dapat diuraikan sebagai berikut :

Asumsi bahwa setiap *instance* direpresentasikan dengan sebuah vektor $\boldsymbol{X}=(x_1,x_2,\ldots,x_n)$, dimana $x_1,x_2,\ldots,xn$ adalah ukuran dari atribut $A_1,A_2,\ldots,A_n$. Andaikan terdapat kelas sejumlah *m* yaitu $C_1,C_2,\ldots,C_m$. Diberikan suatu *instance*$\boldsymbol{X}$ yang belum diketahui kelasnya, dengan menggunakan teorema Bayesian, *posterior probability* dari $\boldsymbol{X}$ terhadap C_1 ditunjukkan pada persamaan (1).

$$P(C_l|\boldsymbol{X}) = [P(\boldsymbol{X}|C_l)P(C_l)]/P(\boldsymbol{X}) = [P(\boldsymbol{X}|C_l)P(C_l)]/[\sum_{k=1}^{m} P(Ck)P(\boldsymbol{X}|Ck)] \quad (1)$$

Class prior probability dapat diduga dengan $P(C_l)=S_l/S$, dimana s_l adalah jumlah dari data pelatihan dengan kelas C_l dan s adalah jumlah total data pelatihan.

Naive Bayes menduga conditionally independent antara satu atribut dan atribut lainnya dengan menggunakan persamaan (2).

$$P(X|Cl) = \prod_{k=1}^{n} P(X_k|C_l) \quad (2)$$

$P(x_k|C_l)$dapat diduga dari data. Sehingga dengan menggunakan persamaan (3) dapat diperoleh nilai peluang $P(Cl|\boldsymbol{X})$.

$$P(Cl|\boldsymbol{X}) = P(C_l)\prod_{k=1}^{n} P(X_k|C_l)/[\sum_{h=1}^{m} P(Ch)P(X_k|C_h)] \quad (3)$$

Untuk menggolongkan sebuah data $\boldsymbol{X}$ yang belum diketahui kelasnya, $P(C_l|\boldsymbol{X})$ dievaluasi untuk setiap kelas C_l Data $\boldsymbol{X}$ akan dimasukkan dalam kelas C_l jika dan hanya jika $P(C_l|X)>P(C_j|X)$, $1\le j\le m, j\ne l$.[4]

2.4 *Confusion Matrix*

Confusion matrix merupakan sebuah tabel yang berisi jumlah banyaknya *test record* yang diprediksi secara benar dan tidak benar oleh model klasifikasi. Bentuk dari *confusion matrix* terlihat pada Tabel 1. Setiap entri pada f_{ij} pada tabel ini menyatakan banyaknya *record* dari kelas i yang diprediksi ke dalam kelas j.

Tabel 1.*Confussion Matrix*

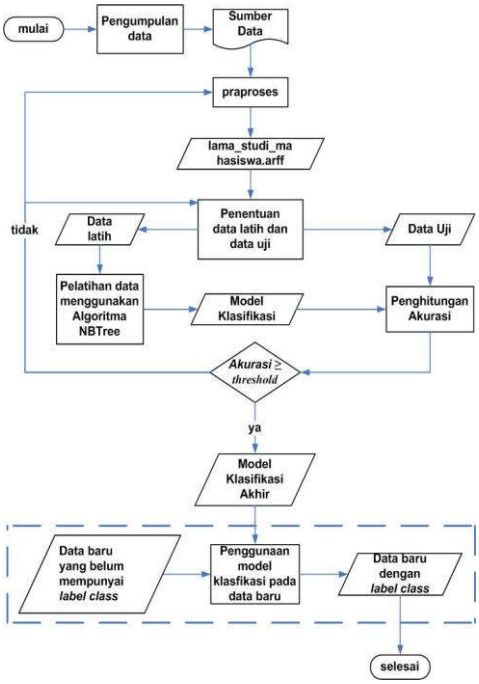
		Kelas yang diprediksi	
		Kelas = 1	Kelas = 0
Kelas aktual	Kelas = 1	f_{11}	f_{11}
	Kelas = 2	f_{01}	f_{00}

Informasi dari *confusion matrix* diperlukan untuk menentukan kinerja suatu model klasifikasi. Informasi ini dapat diringkas ke dalam suatu nilai seperti akurasi .[5]

$$\text{akurasi} = \frac{\text{banyaknya prediksi yang benar}}{\text{total banyaknya prediksi}} = \frac{f_{11}+f_{00}}{f_{11}+f_{10}+f_{01}+f_{00}}$$

3. METODE PENELITIAN

Data yang digunakan dalam penelitian ini merupakan data mahasiswa 2013-2015 di STMIK Indonesia. Penelitian ini dilakukan secara bertahap sesuai tahapan yang telah disusun pada Gambar 2. Nilai *threshold* yang digunakan pada penelitian ini adalah 70%.



Gambar 2. Metode Penelitian

3.1. Praproses Data

Data dari sumber dikumpulkan dan dilakukan tahapan praproses datasebagai berikut :

a. Seleksi data

Pada tahap ini, dilakukan pengelompokkan mahasiswa berdasarkan lama studinya dan memilih atribut-atribut yang sesuai dengan kategori permasalahan.

b. Pembersihan data

Pada data dilakukan pembersihan data untuk memperbaiki data yang hilang atau kosong, data yang mengandung *noise*, dan data yang tidak konsisten.

c. Integrasi data

Pada tahap ini dilakukan penggabungan data dari berbagai sumber ke suatu basis data. Kemudian dilakukan proses reduksi data, dimana data yang tidak relevan dan data yang redudansi dibuang.

d. Transformasi data

Proses perubahan bentuk ke dalam bentuk data yang tepat agar dapat digunakan untuk proses selanjutnya. Proses ini meliputi penyeragaman nama atribut.

3.2. Pembagian Data Latih dan Data Uji

Proses pembagian data menjadi data latih dan data uji dilakukan dengan menggunakan 10 *foldcross validation*. Data latih akan digunakan untuk membentuk model klasifikasi. Sedangkan data uji akan digunakan untuk menghitung akurasi yang diperoleh dari model klasifikasi.

3.3. Klasifikasi

Pada proses klasifikasi dilakukan pembentukan model klasifikasi menggunakan metode *decision tree*. Kemudian dilakukan penghitungan akurasi dari model klasifikasiyang terbentuk. Dari proses klasifikasi ini akan diperoleh model klasifikasi yang dapat digunakan untuk mengisi label kelas dari data baru yang belum diketahui label kelasnya.

3.4. Model Klasifikasi

Proses pembentukan model klasifikasi dari data mahasiswa menggunakan data tabel gabungan dan kelas targetnya. Pembentukan model klasifikasiini menggunakan salah satu algoritme pembentukan *decision tree* yaitu algoritme NBTree.

3.5. Penghitungan Akurasi

Tahap ini adalah tahap untuk menghitung akurasi dari model klasifikasi yang diperoleh dari proses klasifikasi. Metode yang digunakan dalam proses penghitungan akurasi ini adalah dengan menggunakan *confussion matrix*.

Jika hasil akurasi yang diperoleh sudah memenuhi nilai *threshold*, maka model klasifikasi itu akan digunakan untuk menentukan label kelas dari data baru. Akan tetapi jika akurasi yang diperoleh belum memenuhi nilai *threshold*, maka proses klasifikasi akan diulang dengan menggunakan proporsi data latih dan data uji yang berbeda atau mengulang tahap praproses dengan objek yang berbeda.

4. HASIL DAN PEMBAHASAN

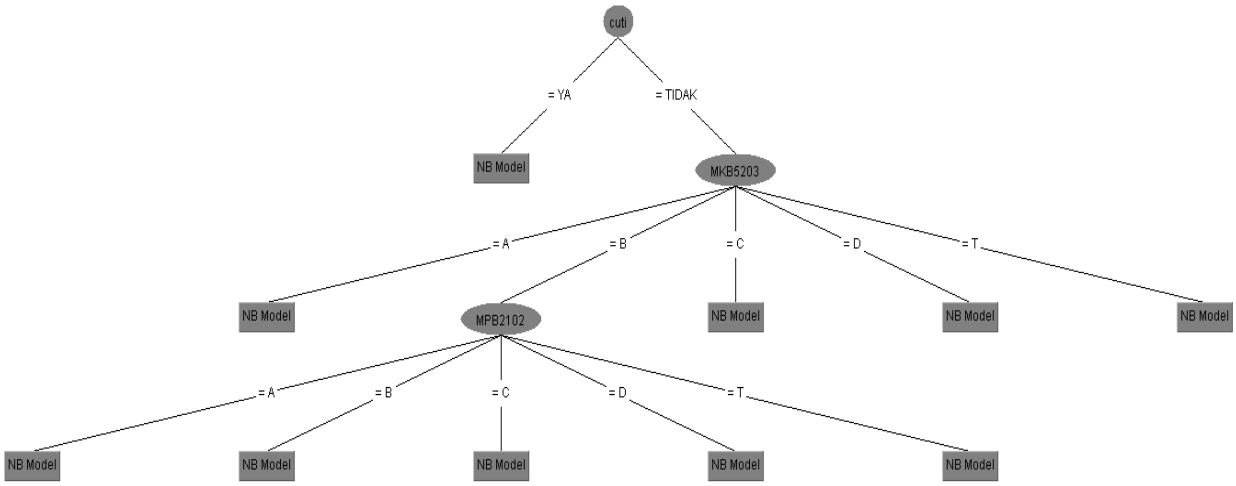
4.1. Praproses Data

Data yang digunakan pada penelitian ini adalah data nilai akademik mahasiswa, data cuti akademik, dan data ketepatan waktu lulus mahasiswaprogram studi sistem informasi tahun 2013 – 2015.Tidak semua data yang terdapat pada data nilai akademik mahasiswa digunakan pada penelitian ini. Setelah melalui tahap pemilihan atribut, terdapat 36 atribut yang digunakan pada penelitian ini. Atribut itu terdiri dari 34 mata kuliah semester 1 sampai semester 4, cuti kuliah dan ketepatan lulus studi. Atribut ketepatan lulus studi menjadi kelas dari data yang digunakan pada penelitian ini.

Proses selanjutnya adalah proses pembersihan data. Salah satu tujuan proses pembersihan data adalah untuk mengganti data yang kosong. Jika terdapat nilai atribut yang kosong untuk suatu record, akan diganti dengan nilai T. Dimana nilai T ini berarti mahasiswa tersebut tidak mengambil mata kuliah tersebut. Hal ini terjadi karena terdapat perbedaan kurikulum antara mahasiswa yang lulus tahun 2013 dengan mahasiswa yang lulus tahun 2014 – 2015. Selain itu nilai masing-masing atribut mata kuliah terdiri dari A, B, C, D, dan T. Tidak terdapat nilai E, dikarenakan salah satu persyaratan sidang skripsi adalah tidak terdapat nilai E. Sehingga mahasiswa yang sudah lulus, otomatis tidak memiliki nilai E. Kemudian masing-masing data tersebut digabungkan menjadi satu.

4.2. Klasifikasi

Proses klasifikasi dilakukan dalam dua tahap, yaitu pembentukan model klasifikasi berupa *decision tree* dan penghitungan akurasi dari *decision tree* yang terbentuk.Pembentukan *decision tree* dilakukan dengan menggunakan algoritme NBTree. Proses pembentukannya dilakukan dengan perangkat lunakWeka. Gambar 3 merupakan*decision tree* yang dibentuk. Pada *decisiontree* yang dihasilkan, diketahui bahwa tidak semua atribut yang digunakan muncul sebagai *node* pada *decision tree*.



Gambar 3. Model Klasifikasi Data Lama Studi Mahasiswa Program Studi Sistem Informasi STMIK Indonesia

Dari gambar 3 di atas, dapat dilihat bahwa dari 35 atribut yang digunakan, hanya 3 atribut yang muncul pada *decision tree* tersebut, yaitu cuti kuliah, MKB5203 (Mata kuliah Sistem Operasi) dan MPB2102 (Mata kuliah Komputer dan Masyarakat). *Leaf node* yang dihasilkan dari *decision tree* yang dibentuk dari algoritme NBTree merupakan sebuah model naïve bayes, dimana model ini berisi peluang untuk masing-masing kelas, dan peluang setiap atribut terhadap masing-masing kelas.

Berdasarkan *decisiontree* yang terbentuk, dapat dibuat 10 aturan. Sebagai contoh, aturanyang terbentuk dari *decision tree* pada Gambar 3, adalah sebagai berikut :

a. Aturan 1:
   Jika nilai atribut cuti adalah YA, maka penentuan kelas dapat dihitung menggunakan model naïve bayes 1.
b. Aturan 6:
   Jika nilai atribut cuti adalah TIDAK, niali atribut MKB5203 adalah B, dan nilai atribut MPB2012 adalah A, maka penentuan kelas dapat dihitung menggunakan model naïve bayes 6.

Nilai yang dihitung dengan menggunakan model naïvebayes adalah *conditionally independent* antara satu atribut dan atribut lainnya dengan menggunakan persamaan (2). Kemudian dilakukan penghitungan peluang suatu *record* termasuk ke dalam setiap kelas yang dihitung dengan Persamaan (3). Pengitungan ini dilakukan untuk semua kelas yang ada. Untuk penentuan kelas, diambil berdasarkan nilai peluang suatu record termasuk ke dalam suatu kelas yang paling tinggi.

4.3. Penghitungan Akurasi

Untuk menghitung akurasi dari model klasifikasiyang terbentuk digunakan *confusion matrix*. *Confusion matrix* yang diperoleh dari model klasifikasitersebut terdapat pada Tabel 2.

Tabel 2.*Confusion matrix* dari model klasifikasi

		Kelas yang diprediksi	
		Kelas = Tepat	Kelas = Tidak Tepat
Kelas Aktual	Kelas = Tepat	270	35
	Kelas = Tidak Tepat	115	145

Penghitungan akurasi dengan menggunakan *confusion matrix* adalah sebagai berikut:

$$\text{akurasi} = \frac{\text{banyaknya prediksi yang benar}}{\text{total banyaknya prediksi}}$$

Dengan menggunakan data pada tabel *confusion matrix*, dapat dihitung akurasi dari model klasifikasi.

$$\text{akurasi} = \frac{270 + 145}{565}$$

$$= \frac{415}{565}$$

$$= 0{,}7345$$

Hasil akurasi yang diperoleh adalah 73,45%. Sedangkan nilai *threshold* yang digunakan adalah 70%. Sehingga nilai akurasi yang diperoleh telah memenuhi *threshold* yang diberikan.

5. KESIMPULAN

Berdasarkan penelitian yang dilakukan dalam membentuk model klasifikasi untuk data lama studi mahasiswa, dapat diambil kesimpulan sebagai berikut :

a. Telah terbentuk model klasifikasi untuk yang memiliki 10 aturan klasifikasi dengan akurasi 73,45%.
b. Lama studi mahasiswa dapat dideskripsikan oleh nilai akademik dan data cuti akademik mahasiswa.

6. SARAN

a. Data yang digunakan bisa ditambah lebih dari 3 tahun,agar penelitiannya agar model data klasifikasinya lebih variatif.
b. Gunakan metode penelitian yang lain, agar bisa membandingkan hasil penelitian NBTree

DAFTAR PUSTAKA

[1] Hastuti K., 2012,*Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Mahasiswa Non Aktif*, Seminar Nasional Teknologi Informasi & Komunikasi Terapan, Semarang, 23 Juni 2012

[2] Kohavi, R., 1996,*Scaling Up the Accuracy of Naïve-Bayes Classifiers : a Decision-Tree Hybrid*, AAAI.

[3] [3] Williams N, Zander S, & Armitage G., 2006,*Evaluating Machine Learning Algorithms for Automated Network Application Identification*, CAIA.

[4] Tan P, Michael S, dan Vipin K., 2005,*Introduction to Data mining*, Pearson Education, Inc, Boston.

[5] Han J, Kamber M., 2006,*Data Mining : Concepts and Techniques*,Morgan Kaufman Publisher, San Francisco.