KAJIAN PENERAPAN *ALGORITMAC4.5, NEURAL NETWORK* DAN *NAÏVE BAYES* UNTUK KLASIFIKASI MAHASISWA YANG BERMASALAH DALAM REGISTRASI

HERU SULISTIONO
mildlaser3@gmail.com
081282400050
Program Studi Teknik Informatika, Fakultas Teknik, Matematika dan IPA
Universitas Indraprasta PGRI
Jl. Nangka 58 Tanjung Barat Jagakarsa Jakarta Selatan

Abstract. Registration is a registration activities conducted in each semester. In an educational institution, for the administrationis veryimportant. If an educational institution having problems in the administration will be useless and can notlast long. Registration done old studentin each semester is the best way for educational institutions including the University Indraprasta PGRI, such as to be able to determine the number of classes that will be prepared at the beginning of the semestertuition. From the data known to thedecrease in the number of students who will be attending for the nextterm. Therefore,any factorthat causes many problems in the registration of students. Purpose of this researchis to create a classification problem or astudent whois not in the registration,in the study conducted comparison algorithm C4.5, naïvebayes and neural network which is applied to the data in the registration troubled students. This study aimed to measure the accuracy of the comparativestudy of classification algorithms in 3 pieces that students have trouble registering. From the test resultsto measure the performance of the three algorithms using Cross Validation testing methods, Confusion Matrix and the ROC curve, it is known that Naive Bayes algorithm has he high estaccuracy value, ie91.57%, followed by C4.5 method with the accuracy of 91.43% and the lowest Neural Network is a method with a value of89.02% accuracy.

Keyword: Data Mining, Algoritma C4.5, Naïve Bayes, Neural Network.

PENDAHULUAN

Perguruan tinggi merupakan penyelenggara pendidikan akademik bagi mahasiswa, Perguruan tinggi diharapkan menyelenggarakan pendidikan yang berkualitas bagi mahasiswa sehingga menghasilkan sumber daya manusia yang berilmu, cakap dan kreatif. Semakin bertambah jumlah perguruan tinggi maka semakin meningkat pula jumlah sumber daya manusia berkualitas yang dihasilkan perguruan tinggi. Salah satu faktor yang menentukan kualitas perguruan tinggi adalah persentasi kemampuan mahasiswa untuk menyelesaikan studi tepat waktu.

Kualitas yang ada pada diri mahasiswa dapat dilihat dari prestasi akademik yang diraihnya. Prestasi akademik merupakan perubahan yang mencakup kecakapan tingkah laku maupun kemampuan yang dapat bertambah selama beberapa waktu yang tidak disebabkan proses pertumbuhan tetapi karena adanya situasi belajar, sehingga dipandang sebagai bukti usaha yang diperoleh mahasiswa (Sobur, 2006).

Database perguruan tinggi menyimpan data akademik, administrasi dan biodata mahasiswa. Data tersebut apabila digali dengan tepat maka dapat diketahui pola atau pengetahuan untuk mengambil keputusan. Serangkaian proses mendapatkan pengetahuan atau pola dari kumpulan data disebut dengan data mining. Data mining memecahkan masalah dengan menganalisis data yang telah ada dalam database. Perguruan tinggi perlu melakukan prediksi perilaku mahasiswa untuk mencegah secara dini kegagalan akademik

mahasiswa. Penelitian yang dilakukan oleh Gerben W. Dekker menyebutkan bahwa monitoring dan dukungan terhadap mahasiswa di tahun pertama sangat penting dilakukan. Mahasiswa jurusan teknik elektro Universitas Eindhoven yang berhenti studi pada tahun pertama mencapai hingga 40%. Kurikulum yang sulit dianggap sebagai salah satu penyebab tingginya jumlah mahasiswa drop out. Selain itu, nilai, prestasi, kepribadian, latar belakang sosial mempunyai peran dalam kesuksesan akademik mahasiswa. Dekker menggunakan algoritma *Decision tree, Bayesian classifiers, logistic models, rule-based learner* dan random *forest*.

Dalam penelitian ini, dilakukan analisis komparasi tiga algoritma klasifikasi data mining yaitu algoritma *C4.5, Neural Network* dan *Naïve Bayes* sehingga dapat diketahui algoritma yang paling akurat untuk memprediksi mahasiswa yang bermasalah dalam registrasi.

TINJAUAN PUSTAKA

Data Mining

Data mining merupakan proses ektraksi pengetahuan dari data yang besar. Sesuai fungsinya, data mining adalah proses pengambilan pengetahuan dari volume data yang besar yang disimpan dalam basis data, data *warehouse*, atau informasi yang disimpan dalam repository (Han dan Kamber, 2006).

Gartner Group dalam (Larose, 2005) menyebutkan bahwa data mining adalah proses menelusuri pengetahuan baru, pola dan tren yang dipilih dari jumlah data yang besar yang disimpan dalam repositori atau tempat penyimpanan dengan menggunakan teknik pengenalan pola serta statistik dan teknik matematika.

Algoritma Klasisfikasi Data *Mining*

Klasifikasi (Han, 2006) adalah proses penemuan model atau fungsi yang menggambarkan dan membedakan kelas data atau konsep dengan tujuan agar dapat digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui.

Klasifikasi data terdiri dari 2 langkah proses. Pertama adalah fase training, yaitu dimana algoritma klasifikasinya dibuat untuk menganalisa data training lalu direfresentasikan dalam bentuk rule klasifikasi. Proses kedua adalah klasifikasi (Han, 2006) dimana data tes digunakan untuk memperkirakan dari rule klasifikasi.

Salah satu teknik klasifikasi yang paling popular digunakan dalam proses data mining adalah *decision tree* (Gorunescu, 2011). *Decision tree* merupakan salah satu metode klasifikasi yang menggunakan representasi struktur pohon (*tree*) dimana setiap *node* merepresentasikan atribut, cabangnya merepresentasikan nilai dari atribut, dan daun merepresentasikan kelas. *Node* yang paling atas dari *decision tree* disebut sebagai *root*.

Pada *decision tree* terdapat 3 jenis *node*, 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 hanya terdapat satu *input* dan mempunyai *output* minimal dua.
c. *Leaf node atau terminal node*, merupakan *node* akhir, pada *node* ini hanya terdapat satu *input* dan tidak mempunyai *output*.

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. Dan mereka juga dapat diekspresikan dalam bentuk bahasa basis data seperti *Structured Query Language* untuk mencari *record* pada kategori tertentu. Pohon keputusan juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel *input* dengan sebuah variabel target.

Sebuah model pohon keputusan terdiri dari sekumpulan aturan untuk membagi sejumlah populasi yang heterogen menjadi lebih kecil, lebih *homogeny* dengan memperhatikan pada variabel tujuannya. Sebuah pohon keputusan mungkin dibangun dengan seksama secara manual atau dapat tumbuh secara otomatis dengan menerapkan salah satu atau beberapa algoritma pohon keputusan untuk memodelkan himpunan data yang belum terklasifikasi.

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. Dan mereka juga dapat diekspresikan dalam bentuk bahasa basis data seperti *Structured Query Language* untuk mencari *record* pada kategori tertentu. Pohon keputusan juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel *input* dengan sebuah variabel target.

Sebuah model pohon keputusan terdiri dari sekumpulan aturan untuk membagi sejumlah populasi yang heterogen menjadi lebih kecil, lebih *homogeny* dengan memperhatikan pada variabel tujuannya. Sebuah pohon keputusan mungkin dibangun dengan seksama secara manual atau dapat tumbuh secara otomatis dengan menerapkan salah satu atau beberapa algoritma pohon keputusan untuk memodelkan himpunan data yang belum terklasifikasi.

Variabel tujuan biasanya dikelompokkan dengan pasti dan model pohon keputusan lebih mengarah pada perhitungan *probability* dari tiap-tiap *record* terhadap kategori-kategori tersebut atau untuk mengklasifikasi *record* dengan mengelompokkannya dalam satu kelas. Pohon keputusan juga dapat digunakan untuk mengestimasi nilai dari variabel *continue* meskipun ada beberapa teknik yang lebih sesuai untuk kasus ini.

Algoritma C4.5

Algoritma C4.5 yaitu pohon keputusan yang merupakan metode klasifikasi dan prediksi yang sangat kuat dan kental. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Pohon keputusan juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel *input* dengan sebuah variabel target. Karena pohon keputusan memadukan antara eksplorasi data dan pemodelan, algoritma ini sangat bagus sebagai langkah awal dalam proses pemodelan.

Algoritma *Neural Network*

Algoritma *neuralnetwork* yang paling populer adalah *Backpropagation*, algoritma *backpropagation* melakukan pembelajaran pada jaringan saraf *multi layer feed forward* yang terdiri dari tiga lapisan/*layer*, yaitu: lapisan *input*, lapisan tersembunyi, dan lapisan keluaran (Han & Kamber, 2006). Pada beberapa diagram *neural network* dimungkinkan terdapat lebih dari satu lapisan tersembunyi, meskipun kebanyakan hanya mengandung satu lapisan tersembunyi yang dirasa cukup untuk berbagai tujuan (Larose, 2006).

Algoritma *neuralnetwork* yang paling populer adalah *Backpropagation*, algoritma *backpropagation* melakukan pembelajaran pada jaringan saraf *multi layer feed forward* yang terdiri dari tiga lapisan/*layer*, yaitu: lapisan *input*, lapisan tersembunyi, dan lapisan keluaran (Han & Kamber, 2006). Pada beberapa diagram *neural network* dimungkinkan

terdapat lebih dari satu lapisan tersembunyi, meskipun kebanyakan hanya mengandung satu lapisan tersembunyi yang dirasa cukup untuk berbagai tujuan (Larose, 2006).

Backpropagation merupakan suatu algoritma yang menggunakan metode pembelajaran terbimbing (*supervised learning*) yang dikenalkan oleh Rumelhart dkk. *Backpropagation* merupakan algoritma *neural network* untuk klasifikasi yang menggunakan *gradient descent*, *backpropagation* mencari satu set bobot yang dapat memodelkan data sehingga dapat meminimalkan jarak kuadrat rata-rata antara prediksi kelas jaringan dan label kelas yang sebenarnya dari *tuple* data (Han & Kamber, 2006).

Algoritma Naïve Bayes

Salah satu metode yang sangat penting dalam klasifikasi adalah metode *Naïve Bayes*. Metode ini juga disebut *idiot's Bayes, simple Bayes, independenceBayes*. 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 dan Kumar, 2009).

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 dan Kumar, 2009).

METODE

Jenis Penelitian

1. Penelitian Experimental
   Penelitian experimental merupakan penelitian yang bersifat uji coba, memanipulasi dan mempengaruhi hal-hal yang terkait dengan seluruh variable atau atribut.
2. Penelitian Perbandingan atau studi komparasi yakni dengan membandingkan algoritma *C4.5, Neural Network* dan *Naïve Bayes*.

Kegiatan penelitian ini melalui beberapa tahap dalam pengembangannya yaitu:

1. Studi pendahuluan
   Kegiatan yang dilakukan pada saat studi pendahuluan yaitu mengumpulkan materi-materi kepustakaan yang berhubungan dengan pengambilan judul. Kemudian langkah selanjutnya yaitu survey langsung ke tempat penelitian Tata Usaha Fakultas Teknik Matematika dan Ilmu Pengetahuan Alam serta BAAK Universitas Indraprasta PGRI. Dari penelitian tersebut akan di dapat beberapa kriteria yang digunakan untuk penelitian lebih lanjut.
2. Data.
   Setelah dilakukannya survey maka didapatkan data yang akan digunakan dalam penelitian ini, yang terdiri dari beberapa atribut yaitu data mahasiswa dan lain-lain.
3. Pengolah Data.
   Setelah mendapatkan data, maka data diolah menggunakan metode data mining, yaitu pertama kali menggunakan aplikasi *microsoft office excel*.

Populasi

Populasi dalam penelitian ini adalah pemilihan sampel dengan mengidentifikasi populasi target yaitu populasi yang relevan dengan tujuan masalah dalam penelitian.

Sampel

Penentuan sampel dilakukan dengan teknik *proposive* (teknik penarikan sampel) yaitu penarikan sampel dengan memiliki target dan tujuan tertentu dalam hal ini menerapkan model pengambilan keputusan dalam pemilihan 1 dari 3 buah algoritma berdasarkan metode data mining klasifikasi mahasiswa teknik informatika yang bermasalah dalam registrasi. Dalam pengambilan sampel dilakukan data *validation* yang bertujuan untuk mengidentifikasi, menghapus data yang ganjil (*outlier/noise*), data yang tidak konsisten dan data yang tidak lengkap (*missing value)*.

Variabel Penelitian

Dalam melakukan klasifikasi mahasiswa teknik informatika yang bermasalah dalam registrasi pada penelitian ini akan dilakukan komparasi algoritma *C4.5, Neural Network* dan *Naïve Bayes* untuk mengetahui algoritma yang paling akurat dalam klasifikasi mahasiswa teknik informatika yang bermasalah dalam registrasi.

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

1. Jenis Kelamin (JK)
   Merupakan atribut yang berisi data jenis kelamin mahasiswa.
2. Kelompok Belajar (KB)
   Merupakan atribut yang berisi data waktu kuliah yang dipilih mahasiswa.
3. IPK semester satu (ipk smt1)
   Merupakan atribut yang berisi data nilai Indeks Prestasi Kumulatif (IPK) pada semester awal kuliah yaitu di semester satu.
4. Kota
   Merupakan atribut yang berisi data kota asal mahasiswa.
5. Jur SLTA
   Merupakan atribut yang berisi data jurusan pada saat mahasiswa di SLTA.
6. Bekerja
   Merupakan atribut yang berisi status mahasiswa bekerja atau tidak selama kuliah.
7. Penghasilan Ortu
   Merupakan atribut yang berisi penghasilan orangtua/wali perbulannya.
8. Biaya Studi
   Merupakan atribut yang berisi sumber biaya yang digunakan mahasiswa untuk membayar biaya kuliah.
9. Beasiswa
   Merupakan atribut yang berisi status penerimaan beasiswa oleh mahasiswa.
10. Keterangan
    Merupakan atribut yang berisi data keterangan aktif atau cuti mahasiswa.

HASIL DAN PEMBAHASAN

Algoritma C.4.5

Berikut ini adalah langkah-langkah klasifikasi mahasiswa dengan algoritma C4.5.

1. Siapkan data *training* yaitu tabel 3.3 yang berjumlah 747 data.
2. Hitung jumlah kelas registrasi dan bermasalah berdasarkan nilai tiap atribut.
3. Hitung nilai *entropy* total dimana diketahui jumlah kelas yang registrasi berjumlah 656 dan kelas bermasalah berjumlah 91.

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

= (-656/747) * $\log_2$ (656/747) + (-91/747) * $\log_2$ (91/747)
= 0,535

4. 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 jenis kelamin yaitu:

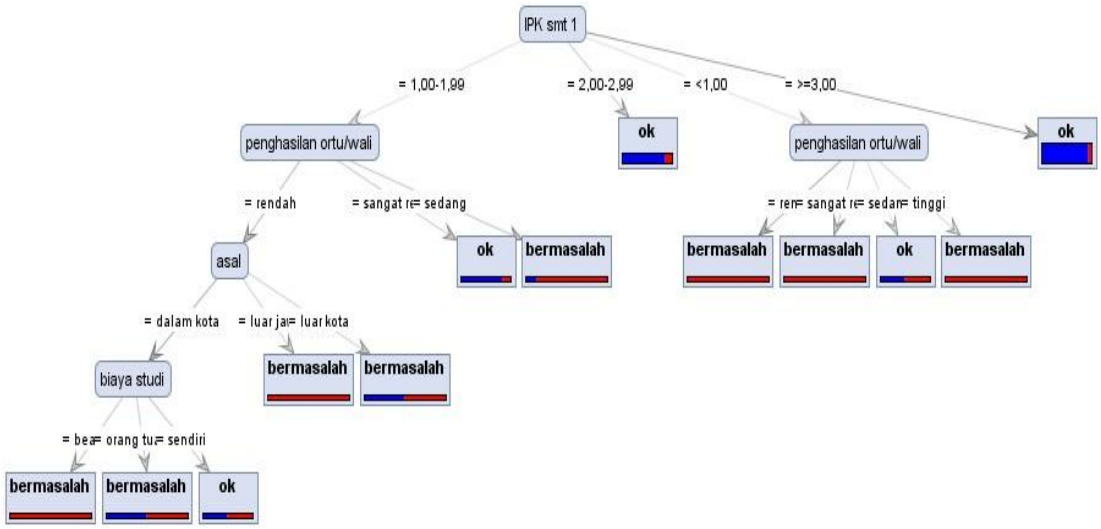
$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy\ (Si)$$

= 0,535 – ((482/747*0, 875)+(265/747*0,426))
= 0,0047662

Perhitungan nilai *entropy* dan *gain* untuk semua atribut dilakukan untuk mendapatkan nilai *gain* tertinggi yang akan dijadikan sebagai akar.

Tentukan simpul-simpul lainnya dengan mencari *gain* tertinggi dari masing-masing perhitungan *node* berdasarkan *node* yang sudah ditentukan hingga ditemukan hasil klasifikasinya.

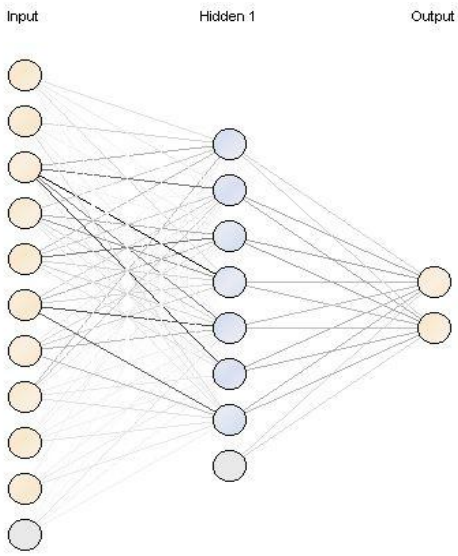
Pembentukan simpul-simpul dengan perhitungan *gain* diperoleh *decision tree* untuk klasifikasi mahasiswa bermasalah dalam registrasi terlihat seperti pada gambar 4.1



Gambar 1
Decision Tree klasifikasi mahasiswa dengan C4.5

Algoritma *Neural Network*

Klasifikasi mahasiswa dengan *Neural Network* menggunakan data *training* yang sama dengan data *training* yang digunakan untuk klasifikasi mahasiswa dengan algoritma C4.5 sejumlah 747 *record* data. Hasil pengolahan data *training* tersebut diperoleh *neural net* seperti pada gambar 4.2 dengan menggunakan metode *neural network* menghasilkan tiga *layer*, yaitu *input layer* yang terdiri dari sepuluh simpul yang terdiri dari sepuluh simpul yang sama dengan jumlah atribut *predictor* dan satu buah simpul bias. *Hidden layer* terdiri dari sembilan simpul yang terdiri dari delapan simpul ditambah satu simpul bias. *Output layer* yang merupakan hasil klasifikasi terdiri dari dua simpul yaitu OK dan Bermasalah.



Gambar 2
Neural Net yang dihasilkan dengan algoritma *Neural Network*

Bobot awal untuk *input layer*, *hidden layer*, dan bias diinisialisasi secara acak (biasanya antara -0.1 sampai dengan 1.0). Simpul bias terdiri dari dua, yaitu pada *input layer* yang terhubung dengan simpul-simpul pada *hidden layer*, dan pada *hidden layer* yang terhubung pada *output layer*. Setelah semua nilai awal diinisialisasi, kemudian dihitung masukan, keluaran, dan *error*. Selanjutnya membangkitkan *output* untuk simpul menggunakan fungsi aktifasi sigmoid. Setelah didapat nilai dari fungsi aktifasi, hitung nilai e*rror* antara nilai yang diprediksi dengan nilai yang sesungguhnya. Setelah nilai e*rror* dihitung, selanjutnya dibalik ke *layer* sebelumnya (*backpropagated*). Nilai *Error* yang dihasilkan dari langkah sebelumnya digunakan untuk memperbarui bobot relasi. Hasil perhitungan akhir *backpropagation* fungsi aktifasi untuk simpul pada *hidden layer* terdapat pada Tabel 4.3. Kolom pertama pada Tabel 4.3 merupakan atribut yang dinyatakan berupa simpul pada *input layer* seperti pada Gambar 4.2. Sedangkan Kolom satu sampai delapan mewakili jumlah simpul pada *hidden layer*.

Tabel 1 Nilai bobot akhir *Hidden Layer*

No	Simpul	Hidden Layer (Sigmoid)						
		1	2	3	4	5	6	7
1	JK	7.232	-1.929	-2.606	0.683	5.497	-0.555	-0.262
2	KB	-0.933	4.248	7.649	-1.083	-5.49	-0.278	0.462
3	IPK smt 1	-2.539	-12.27	-11.182	-21.512	-13.026	-20.056	-5.994
4	Asal	1.584	-0.734	2.642	-6.089	-5.275	-1.336	-5.607
5	Jurusan	9.301	1.525	-4.469	-1.422	1.684	-5.191	6.347
6	Penghasilan ortu	14.155	-0.27	7.364	1.684	5.561	-2.524	5.279
7	Biaya Studi	-1.986	3.299	0.557	-3.063	-1.515	0.726	1.675
8	Bekerja/Tidak	-6.756	1.415	3.669	3.026	-5.355	-1.295	-3.343
9	Beasiswa/Tidak	-0.305	-0.132	0.135	1.445	4.069	2.169	1.975
10	KET	2.368	0.545	-0.557	1.827	4.671	2.568	2.184
11	Threshold	0.640	0.227	-0.444	-1.234	-5.477	-1.992	-1.364

Pada tabel 4.3 kolom simpul menerangkan atribut yang dinyatakan kolom simpul *input layer*. Sedangkan kolom satu sampai delapan menerangkan jumlah simpul pada *hidden layer*. Untuk nilai akhir fungsi aktifasi *output layer* dapat dilihat pada tabel 4.4 dibawah ini:

Tabel 2 Nilai bobot akhir *Output Layer*

Class	**output (sigmoid)**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**threshold**
OK	5.266	2.949	-8.125	6.225	-10.227	7.953	6.912	-2.988
Bermasalah	-5.266	-2.949	8.125	-6.225	10.227	-7.953	-6.912	2.988

Pada tabel 4.4 Class terdiri dari OK dan Bermasalah yang merupakan nilai hasil klasifikasi. Nilai yang terdapat pada kolom satu sampai delapan merupakan nilai bobot akhir pada *output layer*.

Algoritma Naïve Bayes

Penggunaan algoritma *Naïve Bayes* menggunakan data *training* pada Tabel 3.3 dimulai dengan melakukan perhitungan *probabilitas prior* untuk mengetahui nilai yang diterima dan tidak diterima untuk semua jumlah data. Pada data training jumlah data sebanyak 747 data, dimana kelas OK sebanyak 656 record dan yang Bermasalah sebanyak 91 record.

Evaluasi dan Validasi Model

Hasil dari pengujian model yang telah dilakukan yaitu dengam algoritma C4.5, *Neural Network*, dan *Naïve Bayes*, dilakukan pengujian tingkat akurasi dengan menggunakan *confussion matrix* dan kurva ROC/AUC (*Area Under Cover*).

1. *Confusion Matrix*
   a) Perhitungan akurasi data *training* menggunakan algoritma C4.5. Diketahui data *training* terdiri dari 747 *record* data, 649 data diklasifikasikan OK dan 57 data diprediksi OK tetapi ternyata Bermasalah, 7 data dinyatakan OK tetapi Bermasalah, dan, 34 data secara benar diklasifikasikan Bermasalah .
   b) Perhitungan akurasi data *testing* menggunakan algoritma C4.5. Diketahui data *testing* terdiri dari 187 *record* data, 160 data diklasifikasikan OK dan 14 data diprediksi OK tetapi ternyata Bermasalah, 7 data secara benar diklasifikasikan Bermasalah dan 6 data diprediksi Bermasalah ternyata OK.
   c) Perhitungan akurasi data *training* menggunakan algoritma *Neural Network*. Diketahui data *training* terdiri dari 747 *record* data, 627 data diklasifikasikan OK dan 53 data diprediksi OK tetapi ternyata Bermasalah, 38 data secara benar diklasifikasikan Bermasalah dan 29 data diklasifikasikan Bermasalah tetapi ternyata OK.

Confusion matrix untuk data *testing* dengan metode *neural network.* Diketahui dari 187 data *testing*, 159 diklasifikasikan OK, 10 data diprediksi OK tetapi ternyata Bermasalah, 11 data diprediksi dengan benar untuk klasifikasi Bermasalah, dan 7 data diprediksi Bermasalah ternyata OK.

Perhitungan akurasi data *training* menggunakan algoritma *Naïve Bayes*. Diketahui data *training* terdiri dari 747 *record* data, 646 data diklasifikasikan OK dan 53 data

diprediksi OK tetapi ternyata Bermasalah, 38 data secara benar diklasifikasikan Bermasalah dan 10 data diprediksi OK tetapi ternyata Bermasalah.

Perhitungan akurasi data *testing* menggunakan algoritma *Naïve Bayes*. Diketahui data *testing* terdiri dari 187 *record* data, 163 data diklasifikasikan OK dan 12 data diprediksi OK tetapi ternyata Bermasalah, 9 data secara benar diklasifikasikan Bermasalah dan 3 data diprediksi Bermasalah ternyata OK.

Dari hasil *confusion matrix*diatas, selanjutnya dilakukan perhitungan nilai *accuracy*, *precision*, dan *recall*. Perbandingan nilai *accuracy*, *precision*, dan *recall* yang telah dihitung untuk metode C4.5, *naïve bayes*, dan *neural network*.

Tabel 3Perbandingan Nilai A*ccuracy, Precision*, dan *Recall*

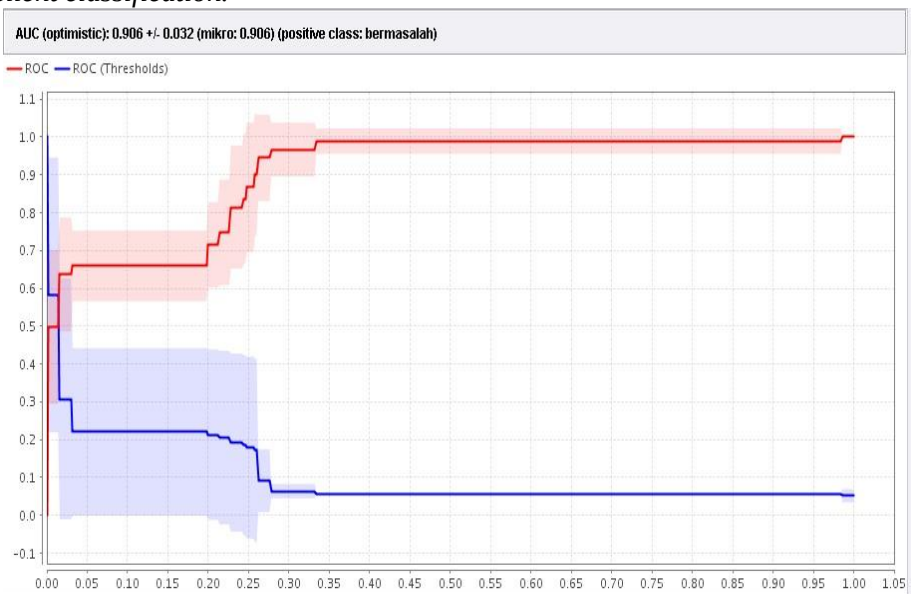
	C4.5		***Neural Network***		***Naïve Bayes***	
	training	*testing*	*training*	*testing*	*training*	*testing*
Accuracy	91.43	89.3	89.02	90.91	91.57	91.99
Precision	85.71	53.85	56.29	69.17	81.67	75.00
Recall	37.67	35.00	41.78	50.00	41.89	45.00

2. Kurva ROC

Hasil perhitungan divisualisasikan dengan kurva ROC (*Receiver Operating Characteristic*) atau AUC (*Area Under Curve*). ROC memiliki tingkat nilai diagnosa yaitu (Gorunescu, 2011):

a. Akurasi bernilai 0.90 – 1.00 = *excellent classification*
b. Akurasi bernilai 0.80 – 0.90 = *good classification*
c. Akurasi bernilai 0.70 – 0.80 = *fair classification*
d. Akurasi bernilai 0.60 – 0.70 = *poor classification*
e. Akurasi bernilai 0.50 – 0.60 = *failure*

Hasil yang didapat dari pengolahan ROC untuk algoritma C4.5 dengan menggunakan data *training* sebesar 0.906 dapat dilihat pada gambar 4.3 dengan tingkat diagnosa *excellent classification.*



Gambar 3 Kurva ROC data *training* untuk metode C4.5

Hasil yang didapat dari pengolahan ROC untuk algoritma C4.5 dengan menggunakan data testing sebesar 0.836 dapat dilihat pada gambar 4.4 dengan tingkat diagnosa *excellent classification.*

Perbandingan hasil perhitungan nilai AUC untuk metode C4.5, *Neural Network*, dan *Naïve Bayes*.

Tabel 4Perbandingan Nilai AUC

	C4.5		***Neural Network***		***Naïve Bayes***	
	training	*testing*	*training*	*testing*	*training*	*testing*
AUC	0.906	0.836	0.730	0.867	0.791	0.857

Analisis Evaluasi Komparasi Model

Berdasarkan pengujian dan evaluasi hasil klasifikasi dengan algoritma C4.5, *Neural Network*, dan *Naïve Bayes* pada tabel 4.14, dapat kita lihat hasilnya yaitu tingkat akurasi menggunakan data *training* tertinggi adalah dengan algoritma *Naïve Bayes* dengan tingkat akurasi 91,57%, sedangkan pada tingkat akurasi tertinggi dengan menggunakan data *testing* yaitu dengan algoritma *Naïve Bayes* juga memiliki nilai akurasi paling tinggi yaitu 91,99%.

Berdasarkan kolom ROC pada tabel 4.14, pada data *training* algoritma C4.5 memiliki tingkat ROC paling tinggi, yaitu 0.906 dan pada *testing* algoritma *Neural Network* memiliki tingkat ROC yang paling tinggi, yaitu 0.867 termasuk dalam katagori *excellent classification*.

Dengan menggunakan perbandingan data *training* dengan data *testing*, yaitu 80 berbanding 20.

Tabel 5Perbandingan Akurasi

Metode	*Confussion Matrix*		Perbandingan Komparasi
	Training	*Testing*	
Algoritma C4.5	91.43%	89.30%	93.05%
Neural Network	89.02%	90.91%	89.56%
Naïve Bayes	91.57%	91.99%	93.58%

Berdasarkan hasil perbandingan akurasi pada tabel 4.16, Algoritma *Naïve Bayes* memiliki tingkat akurasi yang paling tinggi, sehingga baik digunakan untuk klasifikasi mahasiswa yang bermasalah dalam registrasi dengan persentase 93,58%.

PENUTUP

Simpulan

Dalam penelitian ini dilakukan pembuatan model menggunakan algoritma C4.5, *Naïve Bayes*dan *Neural Network*menggunakan data mahasiswa. Model yang dihasilkan, dikomparasi untuk mengetahui algoritma yang paling baik dalam mengatasi masalah pada mahasiswa yang bermasalah pada registrasi. Untuk mengukur kinerja ketiga algoritma tersebut digunakan metode pengujian *Confusion Matrix* dan Kurva ROC, diketahui bahwa algoritma Naïve Bayes memiliki nilai *accuracy* dan AUC paling tinggi

Dengan demikian algoritma Naïve Bayes merupakan algoritma terbaik dan dapat memberikan pemecahan dalam permasalahan mahasiswa yang bermasalah dalam registrasi.

Saran

Agar penelitian ini bisa ditingkatkan, berikut adalah saran-saran yang diusulkan:

1. Hasil penelitian ini diharapkan bisa digunakan pada pihak lembaga, untuk lebih meningkatkan akurasi analisa mahasiswa yang bermasalah dalam registrasi.
2. Model klasifikasi mahasiswa mahasiswa yang bermasalah dalam registrasi diharapkan dapat diterapkan pada sistem sehingga dapat dijadikan pendukung pengambilan kebijakan pihak manajemen dalam mengurangi jumlah mahasiswa yang bermasalah dalam registrasi
3. Untuk mendukung pengambilan keputusan dan pengembangan system informasi manajemen strategik, model ini dapat diterapkan pada sekolah dengan menerapkan sistem yang menggunakan perangkat keras dan perangkat lunak, disertai dengan pembuatan *Standard Operational Procedure* dan pelatihan bagi *end-user*.

DAFTAR PUSTAKA

Alpayadin, Ethem. 2010. **"Introduction to Machine Learning"**, The MIT Press, London, 2010

Bramer, Max. 2007. **Principles of Data Mining**. London: Springer. ISBN-10: 1-84628-765-0, ISBN-13: 978-1-84628-765-7.

Deker, G. W., Pechenizkiy, M., Vleeshouwers, J. M,. 2009. **Predicting Students Drop Out: A Case Study**.

Dua, S. & Xian Du. 2011. **Data Mining and Machine Learning in Cybersecurity**. USA: Taylor & Francis Group. ISBN-13: 978-1-4398-3943-0

Gorunescu, F. 2011. **Data Mining Concept Model and Techniques**. Berlin: Springer. ISBN 978-3-642-19720-8

Guidici, P. & Figini, S. 2009. **Applied Data Mining for Business and Industry (2nd ed)**. Italy. John Wiley & Sons, Ltd. ISBN: 978-0-470-05886-2

Han, J., & Kamber, M. 2006. **Data Mining Concept and Tehniques**. San Fransisco: Morgan Kauffman. ISBN 13: 978-1-55860-901-3

Irnawati. 2011. **Penerapan Data Mining Untuk Pengambilan Keputusan Pada Sistem Informasi Akademik: Studi Kasus Universitas Indraprasta PGRI**

Kusrini, & Luthfi, E. T. 2009. **Algoritma Data Mining**. Yogyakarta: Andi Publishing.

Kusumadewi, S. 2004. **Membangun Jaringan Syaraf Tiruan Menggunakan Matlab & Excel Link**. Yogyakarta: Graha Ilmu.

Larose, D. T. 2005. **Discovering Knowledge in Data**. New Jersey: John Willey & Sons, Inc. ISBN 0-471-66657-2.

Liao, T. W., & Triantaphyllou, E. 2007. **Recent Advances in Data Mining of Enterprise Data: Algorithms and Applications**. Series on Computers and Operations Research, Vol 6. USA: World Scientific. ISBN-13 978-981-277-985-4, ISBN-10 981-277-985-X

Maimon, Oded., & Rokach, Lior. 2010. **Data Mining and Knowledge Discovery Handbook**, 2nd Edition. New York: Springer. ISBN 978-0-387-09822-7

Pauziah, Ulpa. 2012. **Kajian Komparasi Algoritma C45, Naive Bayes, Dan Neural Network Dalam Pemilihan Penerima Beasiswa: Studi Kasus SMA Muhammadiyah 4 Jakarta**.