 ISBN 979 - 26 - 0255 - 0

ANALISIS KOMPARASI ALGORITMA KLASIFIKASI DATA MINING UNTUK PREDIKSI MAHASISWA NON AKTIF

Khafiizh Hastuti
Universitas Dian Nuswantoro, Semarang
E-mail : afis@staff.dinus.ac.id

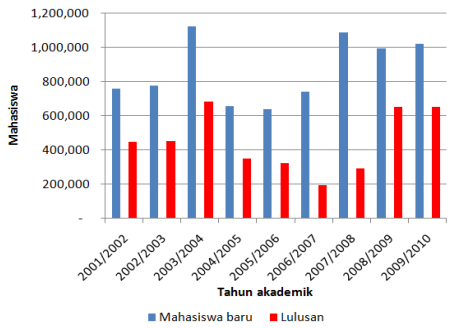
ABSTRAK

Mahasiswa non aktif adalah mahasiswa yang berhenti studi dan tidak melakukan registrasi administratif. Mahasiswa yang memiliki status non aktif memiliki kecenderungan untuk drop out. Tingginya persentase mahasiswa dengan status non aktif mempengaruhi nilai akreditasi universitas. Perlu diketahui faktor-faktor penyebab mahasiswa memiliki status non aktif. Teknik klasifikasi data mining dapat digunakan untuk prediksi mahasiswa non aktif. Banyak algoritma klasifikasi data mining yang dapat digunakan, sehingga perlu dilakukan komparasi untuk mengetahui tingkat akurasi dari masing-masing algoritma. Algoritma yang digunakan adalah logistic regression, decision tree, naïve bayes dan neural network. Data yang digunakan sebanyak 3861 mahasiswa program studi Teknik Informatika, Sistem Informasi dan Desain Komunikasi Visual Universitas Dian Nuswantoro. Hasil dari proses klasifikasi dievaluasi dengan menggunakan cross validation, confusion matrix, ROC Curve dan T-Test untuk mengetahui algoritma klasifikasi data mining yang paling akurat untuk prediksi mahasiswa non aktif.

Kata kunci : mahasiswa non aktif, klasifikasi data mining

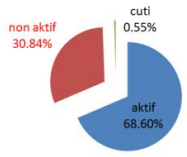
1. Pendahuluan

Perguruan tinggi merupakan penyelenggara pendidikan akademik bagi mahasiswa [1]. Lima lembaga perguruan tinggi diantaranya adalah universitas, institut, sekolah tinggi, akademi dan politeknik. Data yang diperoleh dari Pusat Statistik Pendidikan Badan Penelitian dan Pengembangan Departemen Pendidikan Nasional Republik Indonesia [2] menyebutkan bahwa jumlah lembaga penyelenggara perguruan tinggi mengalami peningkatan setiap tahunnya. Sampai dengan tahun 2010 tercatat 3011 perguruan tinggi diselenggarakan di Indonesia. Perguruan tinggi diharapkan menyelenggarakan pendidikan yang berkualitas bagi mahasiswa sehingga menghasilkan sumber daya manusia yang berilmu, cakap dan kreatif [3]. 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. Berdasarkan matriks penilaian instrument akreditasi program studi Badan Akreditasi Nasional Perguruan Tinggi [4] bahwa persentase mahasiswa yang lulus tepat waktu merupakan salah satu elemen penilaian akreditasi universitas.
Data dari Pusat Statistik Pendidikan Badan Penelitian dan Pengembangan Departemen Pendidikan Nasional Republik Indonesia [2] pada tahun akademik 2001/2002 sampai dengan 2009/2010 menunjukkan bahwa perguruan tinggi menerima rata-rata sebanyak 868.050 mahasiswa baru dan meluluskan rata-rata 451.168 mahasiswa setiap tahunnya. Jumlah lulusan perguruan tinggi ternyata hanya mencapai 51,97% dari jumlah mahasiswa baru setiap tahun. Artinya, terdapat 48,03% mahasiswa yang tidak diketahui statusnya. Ketidakjelasan status tersebut bisa jadi karena mahasiswa menempuh studi tidak tepat waktu, memiliki status non-aktif (mangkir) atau bahkan *drop out*.



Gambar 1: Perbandingan jumlah mahasiswa baru dan lulusan perguruan tinggi di Indonesia tahun akademik 2001/2002 sampai dengan 2009/2010

Saat ini, masalah kegagalan studi siswa dan faktor-faktor penyebabnya menjadi topik yang menarik untuk diteliti [5]. Perguruan tinggi perlu mendeteksi perilaku mahasiswa yang memiliki status "tidak diinginkan" tersebut sehingga dapat diketahui faktor-faktor penyebab kegagalannya. Beberapa penyebab kegagalan mahasiswa diantaranya rendahnya kemampuan akademik, faktor pembiayaan, domisili saat menempuh studi dan faktor lainnya. Universitas Dian Nuswantoro merupakan salah satu perguruan tinggi swasta terbesar di Jawa Tengah yang memiliki 13.416 mahasiswa. Gambar 2 menunjukkan bahwa terdapat 30,84% atau sekitar 4.138 mahasiswa dengan status non aktif. Tingginya persentasi mahasiswa yang memiliki status non aktif menyebabkan tingginya persentasi mahasiswa lulus tidak tepat waktu. Hal tersebut menjadi sangat penting bagi manajemen universitas mengingat persentasi mahasiswa lulus tepat waktu adalah salah satu elemen penilaian akreditasi yang ditetapkan oleh Badan Akreditasi Nasional. Manajemen memerlukan tindakan untuk mengetahui faktor-faktor penyebab mahasiswa memiliki status non aktif.



Gambar 2: Mahasiswa Universitas Dian Nuswantoro tahun 2011

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 [6]. Serangkaian proses mendapatkan pengetahuan atau pola dari kumpulan data disebut dengan *data mining* [7]. *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 Kotsiantis, Pierrakeas dan Pintelas [12] menyebutkan bahwa sangat penting bagi dosen untuk mendeteksi mahasiswa yang cenderung *drop out* sebelum mereka memasuki pertengahan masa studi. Beberapa algoritma klasifikasi *data mining* telah digunakan untuk memprediksi perilaku mahasiswa yang berpotensi *drop out* diantaranya *decision tree, neural network, naïve bayes, instance-based learning, logistic regression* dan *support vector machine*. Hasilnya, *naïve bayes* menunjukkan hasil yang paling akurat. Penelitian yang dilakukan oleh Gerben W. Dekker [11] 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 empat algoritma klasifikasi data mining yaitu *logistic regression, decision tree, naïve bayes* dan *neural network* dengan menggunakan 3681 data set mahasiswa yang terdiri atas data demografi dan akademik mahasiswa sehingga dapat diketahui algoritma yang paling akurat untuk memprediksi mahasiswa non-aktif.

2. Data Mining

Menurut Witten [7], 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. Penelitian tentang klasifikasi algoritma data mining untuk prediksi mahasiswa yang memiliki potensi *drop-out* dilakukan oleh Sotiris Kotsiantis [13] dengan menggunakan 354 mahasiswa Hellenic Open University sebagai data set. Kotsiantis mengelompokkan 2 kelompok atribut yaitu: berbasis kurikulum dan kinerja mahasiswa. Atribut kelompok berbasis kurikulum terdiri atas: jenis kelamin, usia, status marital, jumlah anak, pekerjaan, kemampuan komputer, hubungan pekerjaan dengan komputer. Adapun atribut dalam kelompok kinerja mahasiswa terdiri atas: tatap muka ke-1, tugas ke-1, tatap muka ke-2, tugas ke-2. Kotsiantis menggunakan 6 (enam) algoritma yaitu: *decision tree, neural network, naïve bayes, instance-based learning, logistic regression* dan *support vector machine.*
Dekker [11] melakukan komparasi algoritma *Decision tree, Bayesian classifiers, logistic models, rule-based learner* dan *random forest* dengan menggunakan 648 data set mahasiswa. Dalam penelitian tersebut *decision tree* menunjukkan tingkat akurasi yang paling tinggi.
C. Marquez-Vera, C. Romero dan S. Ventura [5] melakukan komparasi sepuluh algoritma untuk memprediksi kegagalan studi mahasiswa sebagai berikut: JRip, NNge, OneR, Prism dan Ridor, J48, SimpleCart, ADTree, RandomTree dan REPTree. Romero menggunakan 670 data set mahasiswa UAPUAZ yang diketahui: 610 lulus dan 60 gagal menempuh studi. Dari penelitian tersebut ADTree menunjukkan tingkat akurasi yang paling tinggi yaitu mencapai 97%.
Zlatko J. Kovacic [14] meng-eksplore latar belakang *socio-demographic* mahasiswa dengan menggunakan atribut (usia, jenis kelamin, etnis, pendidikan, status pekerjaan dan kekurangan) dan *study environment* (program studi dan *course block*). Penelitian tersebut dilakukan untuk mengetahui faktor-faktor yang mempengaruhi kelancaran atau drop out

mahasiswa di Open Polytechnic of New Zealand dengan menggunakan 450 data set mahasiswa dengan menggunakan algoritma CART.

2.1 Teknik Klasifikasi Data Mining

2.2 Logistic Regression

Menurut Budi Santoso [8], tujuan dari model ini adalah untuk mendapatkan persamaan regresi yang dapat memprediksi dua atau lebih kelompok objek yang dapat ditempatkan yaitu apakah mahasiswa diklasifikasikan sebagai mahasiswa aktif atau non aktif.

Diberikan set sampel dengan jumlah dimensi dan label kelas $y_i \in \{1, 2, \ldots, K\}$. Kemudian LR mencoba untuk memperkirakan probabilitas posterior dari sampel x yang baru. LR dapat diterapkan ke dalam klasifikasi biner dengan $y \in \{0, 1\}$. Maka probabilitas posterior sampel *x* dapat dihitung:

$$(= 0|) = \frac{-(+)}{1 + -(+)} \quad (1)$$

dan

$$(= 1|) = \frac{1}{1 + -(+)}, \quad (2)$$

Model matematika yang digunakan adalah:

$$( ) = \frac{}{1-} = + + + \cdots + \quad (3)$$

Dimana , , ,… , merupakan parameter yang dicari. Perbandingan antara — disebut dengan *odds ratio*.

2.3 Decision Tree

Decision tree merupakan salah satu teknik klasifikasi data mining yang paling populer. Decision tree sesuai digunakan untuk kasus yang memiliki ciri-ciri sebagai berikut [8]:

1. Data atau contoh dinyatakan dengan pasangan atribut dan nilainya.
2. Label atau output data biasanya bernilai diskrit.
3. Data mempunyai *missing value*

Teori entropi diadopsi untuk memilih pemecahan atribut yang tepat untuk algoritma C4.5, dengan menyatakan jumlah rata-rata informasi yang dibutuhkan untuk mengklasifikasikan sampel.

Untuk menghitung nilai entropy digunakan rumus:

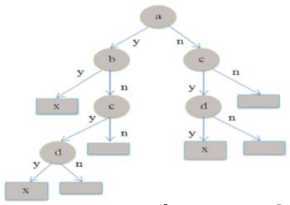
$$( ) = - . \quad (4)$$

Dimana *S* merupakan himpunan kasus, *n* adalah jumlah partisi *S*, dan adalah proporsi terhadap S.Ketika output data atau variabel dependent *S* dikelompokkan berdasarkan atribut *A*, dinotasikan dengan (,). hasil dari atribut mendapatkan *information gain* yang didefinisikan sebagai:

$$(,) = ( ) - - . ( ) \quad (5)$$

Dimana *S* merupakan himpunan kasus, *A* adalah atribut, *n* adalah jumlah partisi atribut A, adalah proporsi terhadap *S* dan S adalah jumlah kasus dalam himpunan. Sebuah prosedur tambahan dilakukan untuk menghindari pohon yang menghasilkan *overfits* data yang kompleks.

Untuk memudahkan pembacaan pohon keputusan dibuatlah set aturan if-then, dimana satu aturan digeneralisasikan ke seluruh simpul daun. Sebuah ilustrasi dijelaskan pada gambar 3, dimana sebuah aturan (rule) dengan struktur yang sama tetapi berbeda atribut, seperti:



Gambar 3: Decision Tree untuk Disjungsi Sederhana

2.4 Naïve Bayes

Naïve Bayes merupakan salah satu penerapan teorema Bayes. Naïve Bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai output [8].

Untuk mendapatkan nilai probabilitas pada sebuah sampel diberikan sebuah teorema Bayes:

$$(h| \ ) = \frac{(\ |h) \ (h)}{(\ )} \quad (6)$$

Dimana*P(h)* adalah nilai probabilitas prior dari hipotesa pada sebuah sampel, biasa disebut dengan priori. *P(x)* merupakan evidence dari probabilitas data pelatihan. *P(h|x)* adalah nilai probabilitas *h* yang mempengaruhi *x* (*posterior density*), sedangkan *P(x|h)* merupakan probabilitas *x* kepada *h* yang disebut dengan likelihood.

$$= \frac{h \quad \times}{} \quad (7)$$

Kemudian gunakan probabilitas *m-estimasi* :

$$\frac{+}{+} \quad (8)$$

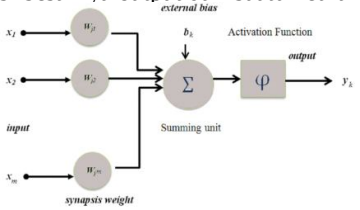
Keterangan:

n_c : total nilai dari contoh sampel pada atribut yang dimiliki kelas *C*
n : n total nilai keseluruhan sampel
m: nilai ekivalen yang konstan dari ukuran sampel
p: probabilitas prior

2.5 Neural Network

Artificial Neural Network (ANN) terinspirasi dari jaringan saraf makhluk hidup. Menurut Budi Santoso [8], kelebihan utama Neural network adalah kemampuan memprediksi, kecepatan dan *robust* terhadap missing data.
Neuron adalah unit pemroses informasi dalam neural network yang terdiri atas:

1. Set synapsis atau link penghubung yang ditandai dengan adanya bobot.
2. Penambah, untuk menjumlahkan signal input yang diberi bobot yang disebut kombinasi linier.
3. Fungsi aktivasi, untuk membatasi besarnya output dari suatu neuran.



Gambar 4: Model Neuron

Menghitung jumlah *n* signal input $x_{ij}=1,2,...,n$ yang diberi bobot dan menghasilkan nilai 1 bila jumlah di atas batas tertentu dan 0 bila di bawah batas. Secara Matematis dapat ditulis:

$$= \quad - \quad (9)$$

dimana φ(.) adalah fungsi aktivasi dan *w* adalah bobot sesuai dengan input ke-j. Dalam model neuron, bias dinyatakan sebagai *b* yang mempunyai fungsi menaikkan atau menurunkan net input untuk fungsi aktivasi. Neuron dinyatakan dengan *k*, didiskripsikan secara matematis:

$$= \quad (10)$$

dan

$$= (\quad + \quad) \quad (11)$$

Dimana $x_1,x_2,...,x_m$ adalah signal input dan $w_1,w_2,...,w_m$adalah bobot dari synapsis k. adalah kombinasi linier dari output yang dihasilkan signal. adalah bias, adalah fungsiaktivasi dan adalah signal output dari neuron yang bersangkutan. Pemakaian bias mempengaruhi output neuron

= + atau

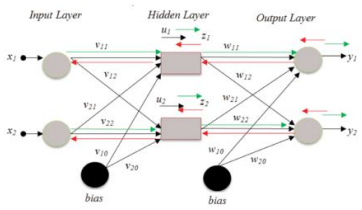
$$= \quad (12)$$

dan = ( ). Dimana *f(*)* adalah fungsi aktivasi dan b_k adalah bias. Sehingga fungsi aktivasi sigmoid didefinisikan sebagai berikut:

$$( ) = \frac{1}{1 \ +} \quad (13)$$

Untuk mekanisme back propagasi dengan menggunakan multi layer perceptron, ditunjukkan pada gambar 5 di bawah ini:

 ISBN 979 - 26 - 0255 - 0



Gambar 5: Mekanisme Back Propagasi pada multi layer perceptron

2.6 Mahasiswa non-aktif

Dalam sistem pendidikan perguruan tinggi di Indonesia terdapat lima status mahasiswa yaitu: mahasiswa aktif, mahasiswa non aktif, mahasiswa cuti, mahasiswa keluar dan mahasiswa drop out (berhenti studi). Yang dimaksud mahasiswa non aktif adalah mahasiswa yang tidak melakukan registrasi administratif setiap awal semester gasal [10]. Menurut peraturan akademik, mahasiswa yang memiliki status non aktif selama empat semester berturut-turut dikategorikan sebagai mahasiswa drop out.

Mahasiswa dengan status non aktif mengakibatkan mahasiswa tersebut tidak dapat menyelesaikan studi tepat waktu. Data yang diperoleh dari Universitas Dian Nuswantoro pada tahun 2011 menunjukkan bahwa dari 13.416 mahasiswa terdapat 30,84% atau 4.138 mahasiswa memiliki status non aktif. Keadaan ini sangat memprihatinkan mengingat pada tahun 2009, Universitas Dian Nuswantoro telah men-*drop out*-kan sebanyak 3.432 mahasiswa. Artinya, bahwa dalam jangka waktu dua tahun rata-rata terdapat 2.069 mahasiswa yang memiliki status non aktif tiap tahunnya.Manajemen universitas selama ini menganggap bahwa faktor kemampuan financial mahasiswa sebagai faktor utama munculnya kasus mahasiswa memiliki status non aktif.

3. Eksperimen dan Pengujian Model

Pada tahap ini dilakukan eksperimen dan teknik pengujian yang digunakan untuk mengukur tingkat akurasi masing-masing algoritma berdasarkan data set mahasiswa yang digunakan yaitu sebanyak 3.861 mahasiswa dari program studi Teknik Informatika, Sistem Informasi dan Desain Komunikasi Visual jenjang strata satu tahun akademik 2005 sampai dengan 2009. Tercatat 1.018 mahasiswa memiliki status non aktif dan 2.843 mahasiswa dengan status aktif. Atribut yang digunakan ditunjukkan dalam tabel 1 berikut ini:

Tabel 1: Atribut data mahasiswa

Program Studi	Teknik Informatika, Sistem Informasi, Desain Komunikasi Visual
Jenis Kelamin	Laki-laki, Perempuan
Usia saat mendaftar	<=20, <=25, <=30, >30
Kota asal	Semarang, Non Semarang, Luar Jawa
Status Domisili	Rumah sendiri, Kos atau sewa, Tidak diketahui
Agama	Islam, Katolik, Kristen, Hindu, Budha
Marital	Single, Menikah
Asal Sekolah	Negeri, Swasta
Status Kerja	Tidak bekerja, Sambil bekerja, Tidak diketahui
Asal biaya	Orang tua, Sendiri, Beasiswa, Tidak diketahui
Pekerjaan orang tua	PNS, Polisi/TNI, Swasta, Wiraswasta, Petani/nelayan, Tidak diketahui
Penghasilan orang tua	< 1jt, >= 1jt - < 3jt, >= 3jt - < 5jt, >= 5jt, Tidak diketahui
IPS 1	Indek Prestasi semester 1
IPS 2	Indek Prestasi semester 2
IPS 3	Indek Prestasi semester 3
IPS 4	Indek Prestasi semester 4
SKS 1	Jumlah Satuan Kredit Semester yang diambil pada semester 1
SKS 2	Jumlah Satuan Kredit Semester yang diambil pada semester 2
SKS 3	Jumlah Satuan Kredit Semester yang diambil pada semester 3
SKS 4	Jumlah Satuan Kredit Semester yang diambil pada semester 4
Status skripsi	Belum skripsi, Skripsi
Status akademik	Aktif, Non aktif

Diperlukan cara yang sistematis untuk mengevaluasi kinerja suatu metoda. Evaluasi klasifikasi didasarkan pada pengujian pada obyek benar dan salah [9]. Validasi data digunakan untuk menentukan jenis terbaik dari skema belajar yang digunakan, berdasarkan data pelatihan untuk melatih skema pembelajaran untuk memaksimalkan penggunaan data [7].

1. *Cross Validation*

Merupakan pengujian standar yang dilakukan untuk memprediksi *error rate*. Setiap kelas pada data set harus diwakili dalam proporsi yang tepat antara data *training* dan data *testing*. Data dibagi secara acak pada masing-masing kelas dengan perbandingan yang sama. Untuk mengurangi bias yang disebabkan oleh sampel tertentu, seluruh proses *training* dan *pengujian* diulangi beberapa kali dengan sampel yang berbeda. Tingkat kesalahan pada iterasi yang berbeda akan dihitung rata-ratanya untuk menghasilkan *error rate* secara keseluruhan.

2. *Confusion Matrix*

Evaluasi model klasifikasi didasarkan pada pengujian untuk memperkirakan obyek yang benar dan salah [9], urutan pengujian ditabulasikan dalam *confusion matrix* dimana kelas yang diprediksi ditampilkan dibagian atas matriks dan kelas yang diamati disisi kiri. Setiap sel berisi angka yang menunjukkan berapa banyak kasus yang sebenarnya dari kelas yang diamati untuk diprediksi.

Tabel 2: Confusion Matrix untuk 2 kelas

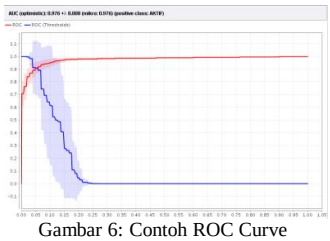
CLASSIFICATION	PREDICTED CLASS	
	Class = YES	Class = No
Class = YES	a (true positive-TP)	b (false negative-FN)
Class = No	c (true positive-TP)	d (false negative-FN)

Hasil klasifikasi dapat dihitung tingkat akurasinya berdasarkan kinerja matriks. Untuk menghitung tingkat akurasi pada matriks digunakan:

$$= \frac{+}{+ \quad + \quad +}$$

3. ROC *Curve*

Kurva ROC dibagi dalam dua dimensi, dimana tingkat TP diplot pada sumbu Y dan tingkat FP diplot pada sumbu X. Tetapi untuk merepresentasikan grafis yang menentukan klasifikasi mana yang lebih baik, digunakan metode yang menghitung luas daerah dibawah kurva ROC yang disebut AUC (*Area Under the ROC Curve*) yang diartikan sebagai probabilitas [9]. AUC mengukur kinerja diskriminatif dengan memperkirakan probabilitas output dari sampel yang dipilih secara acak dari populasi positif atau negatif, semakin besar AUC, semakin kuat klasifikasi yang digunakan. Karena AUC adalah bagian dari daerah unit persegi, nilainya akan selalu antara 0,0 dan 1,0.



Gambar 6: Contoh ROC Curve

Tabel 3 merupakan panduan keakuratan klasifikasi dengan menggunakan AUC:

Tabel 3: Klasifikasi AUC

Performance	Klasifikasi
0,90 – 1,00	Paling baik
0,80 – 0,90	Baik
0,70 – 0,80	Adil atau sama
0,60 – 0,70	Rendah
0,50 – 0,60	Gagal

3. T-Test

T-Test adalah metode pengujian hipotesis dengan menggunakan satu individu (objek penelitian) dengan menggunakan dua perlakuan yang berbeda. Walaupun dengan menggunakan objek yang sama tetapi sampel tetap terbagi menjadi dua yaitu data dengan perlakuan pertama dan data dengan perlakuan kedua. *Performance* dapat diketahui dengan cara membandingkan kondisi objek penelitian pertama dan kondisi objek pada penelitian kedua.

 ISBN 979 - 26 - 0255 - 0

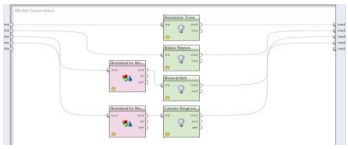
4. Performance Komparasi

Perbandingan *performance* masing-masing algoritma diitunjukkan dalam tabel 4 di bawah ini:

Tabel 4: Perbandingan *performance* algoritma

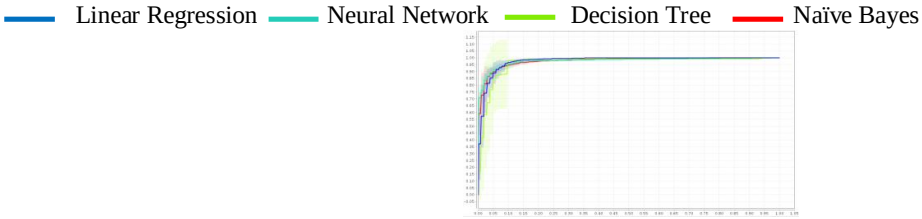
	LR	DT	NB	NN
Accuracy	81,64	95,29	93,47	94,56
AUC	0,933	0,963	0,976	0,976

Berdasarkan tabel 4 dapat diketahui bahwa algoritma *decision tree* memiliki nilai *accuracy* tertinggi yaitu 95,29%, *neural network* 94,56%, *naïve bayes* 93,47% *dan logistic regression* 81,64%.
Sedangkan pada uji ROC *curve* menunjukkan bahwa *neural network* dan *naïve bayes* mencapai nilai AUC yang terbaik yaitu 0,976, kemudian *decision tree* 0,963 dan *logistic regression* 0,933.
Berikut ini adalah uji komparasi empat algoritma dengan menggunakan ROC *curve* melalui framework rapid miner yang ditunjukkan dalam gambar 7:



Gambar 7: Desain model komparasi menggunakan Compare ROCs

Hasil komparasi dengan menggunakan ROC *curve* ditunjukkan dalam gambar 8 di bawah ini:



Gambar 8: Grafik ROC *curve* untuk model algoritma *logistic regression, decision tree, naïve bayes* dan *neural network*

Hasil komparasi algoritma *logistic regression, decision tree, naïve bayes* dan *neural network* menunjukkan bahwa *neural network* dan *naïve bayes* mencapai nilai AUC yang paling baik yaitu 0,976.

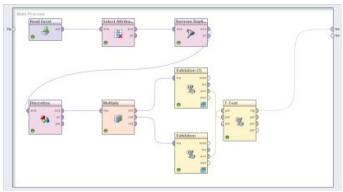
Analisis hasil komparasi

Lima klasifikasi performance akurasi AUC ditunjukkan dalam tabel 5 sebagai berikut:

Tabel 5: klasifikasi AUC

Performance	Klasifikasi
0,90 – 1,00	Paling baik
0,80 – 0,90	Baik
0,70 – 0,80	Adil atau sama
0,60 – 0,70	Rendah
0,50 – 0,60	Gagal

Berdasarkan hasil perhitungan AUC dari masing-masing algoritma, dapat diketahui bahwa dalam penelitian ini algoritma *logistic regression* (0,933), *decision tree* (0,963) , *naïve bayes* (0,976) dan *neural network* (0,976) sehingga keempat algoritma tersebut termasuk dalam klasifikasi paling baik.
Untuk penentuan lebih lanjut akan digunakan pengujian dengan memanfaatkan uji statistik yaitu dengan menggunakan uji T-Test.



Gambar 9: Model T-Test pada *naïve bayes* dan *neural network*

Dalam pengujian ini, akan dibandingkan dua algoritma secara bergantian sehingga akan didapatkan hasil perbandingan seperti tabel 6:

Tabel 6: Uji statistik T-Test

	LR	DT	NB	NN
LR		**0,000**	**0,000**	**0,000**
DT	**0,000**		0,966	0,054
NB	**0,000**	0,966		**0,047**
NN	**0,000**	0,054	**0,047**	

Berdasarkan uji statistik T-Test pada tabel 6 di atas dapat dianalisis bahwa algoritma *logistic regression* memiliki nilai yang sangat dominan terhadap algoritma yang lainnya.
Dengan demikian, berdasarkan uji T-Test yang dilakukan terhadap seluruh algoritma dapat diperoleh hasil perbandingan seperti dalam tabel 7 di bawah ini:

Tabel 7: Hasil perbandingan seluruh pengujian

	LR	DT	NB	NN
Accuracy	81,64	95,29	93,47	94,56
AUC	0,933	0,963	0,976	0,976
T-TEST	DOMINAN	TIDAK DOMINAN	DOMINAN	DOMINAN

Berdasarkan tabel 7 dapat disimpulkan bahwa dalam uji T-Test, algoritma *logistic regression* menunjukkan hasil bahwa algoritma tersebut paling dominan terhadap algoritma yang lain meskipun nilai *accuracy* tidak menunjukkan hasil yang tertinggi yaitu 81,64% dan nilai AUC 0,933.
Algoritma *decision tree* meskipun memiliki *accuracy* tertinggi yaitu 95,29% dan nilai AUC mencapai 0,963 ternyata dalam uji T-Test tidak dominan terhadap algoritma yang lain.
Algoritma *naïve bayes* mencapai *accuracy* 93,47%, nilai AUC 0,976 dan dalam pengujian T-Test bersifat dominan terhadap algoritma yang lain.
Algoritma *neural network* memiliki nilai *accuracy* yang cukup tinggi yaitu 94,56%, dalam uji AUC mencapai hasil maksimal yaitu 0,976 dan bersifat dominan terhadap algoritma yang lain.

5. Kesimpulan dan saran

Berdasarkan komparasi algoritma klasifikasi data mining yaitu *logistic regression, decision tree, naïve bayes* dan *neural network* untuk prediksi mahasiswa non aktif dengan menggunakan 3861 data set mahasiswa Universitas Dian Nuswantoro, maka dapat dianalisa hasil bahwa *decision tree* merupakan algoritma yang paling akurat, namun demikian *decision tree* tidak dominan terhadap algoritma yang lain. *Logistic regression* merupakan algoritma yang paling dominan di antara algoritma yang lain meskipun akurasinya paling rendah. Berdasarkan nilai AUC, *logistic regression, decision tree, naïve bayes* dan *neural network* masuk dalam kategori *excellent classification.*
Saran untuk penelitian selanjutnya adalah dengan menambahkan beberapa algoritma klasifikasi *data mining* untuk dikomparasi seperti *K-Nearest Neighbours, Support Vector Machine, Linear Discriminant Analysis* (LDA), dan lain-lain. Menambah jumlah atribut seperti nilai ujian nasional calon mahasiswa, prestasi siswa di sekolah asal, jalur masuk pendaftaran siswa dan lain-lain. adapun untuk melakukan analisis komparasi dapat menggunakan metode pengukuran *delong pearson.*

DAFTAR PUSTAKA

[1] Peraturan Pemerintah Republik Indonesia Nomor 66 tahun 2010 tentang Perubahan atas Peraturan Pemerintah Nomor 17 tahun 2010 tentang Pengelolaan dan Penyelenggaraan Pendidikan.

[2] Pusat Statistik Pendidikan Badan Penelitian dan Pengembangan Departemen Pendidikan Nasional Republik Indonesia.

[3] Undang Undang Republik Indonesia Nomor 20 tahun 2003 tentang Sistem Pendidikan Nasional.

[4] Buku VI Matriks Penilaian Instrumen Akreditasi Program Studi Badan Akreditasi Nasional Perguruan Tinggi, 2008.

[5] C. Marquez-Vera, C. Romero, and S. Ventura, "Predicting School Failure Using Data Mining," *Journal of Educational Data Mining*, 2011.

[6] Alaa El-Halees, "Department of Computer Science," *Mining Students Data to Analyze Learning Behaviour: A Case Study*, *Journal of Educational Data Mining,*2009.

[7] Ian H. Witten, frank Eibe, and Mark A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed., Asma Stephan and Burlington, Eds. United States of America: Morgan Kaufmann, 2011.

[8] Budi Santoso, *Data Mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis*, 1st ed. Yogyakarta, Indonesia: Graha Ilmu, 2007.

[9] Florin Gorunescu, *Data Mining: Concepts, Model and Techniques*, Prof. Janusz Kacprzyk and Prof. Lakhmi C. Jain, Eds. Berlin,

Jerman: Springer, 2011, vol. 12.

[10] Keputusan Rektor Universitas Dian Nuswantoro nomor: 075/KEP/UDN-01/IV/2009 tentang Peraturan Akademik Universitas Dian Nuswantoro tahun akademik 2009/2010.

[11] Gerben W. Dekker, "Predicting Students Drop Out: A Case Study," *In International Conference on Educational Data Mining, Cordoba, Spain*, 41-50, 2009.

[12] S.B. Kotsiantis, C.J. Pierrakeas, and P.E. Pintelas, "Preventing Student Dropout in Distance Learning Using Machine Learning Techniques," *In International Conference on Knowledge-Based Intelligent Information & Engineering Systems, Oxford*, 3-5, 2003.

[13] Sotiris Kotsiantis, "Educational Data Mining: A Case Study for Predicting Dropout-Prone Students," *Int. J. of Knowledge Engineering and Soft Data Paradigms*, vol. X, 2010.

[14] Zlatco J. Kovacic, "Early Prediction of Student Success: Mining Students Enrolment Data," in *Proceedings of Informing Science & IT Education Conference (InSITE)*, 2010.