KAJIAN KOMPARASI ALGORITMA C4.5, *NAÏVE BAYES* DAN *NEURAL NETWORK* DALAM PEMILIHAN PENERIMA BEASISWA (Studi Kasus pada SMA Muhammadiyah 4 Jakarta)

Ulfa Pauziah
Program Studi Teknik Informatika, Universitas Indraprasta PGRI
Email: pelangi_ulfa@yahoo.com

ABSTRAK

Salah satu masalah pendidikan saat ini yang dihadapi oleh bangsa Indonesia adalah bagaimana meningkatkan mutu pendidikan disetiap jenjang. Adakalanya memang benar sang pelajar tidak mampu secara finansial, tapi tak jarang dari kalangan mampu pun memanfaatkan surat ini. Yang penting bisa sekolah gratis. Jikalau targetnya adalah sekolah gratis tanpa pertanggungjawabkan " beasiswa" yang diterima, seharusnya lembaga pendidikan mematok beberapa peraturan yang memacu prestasi penerima beasiswa. Oleh karena itu dalam penelitian ini dilakukan komparasi algoritma C4.5, *naïve bayes* dan *neural network* yang diaplikasikan terhadap data siswa yang menerima beasiswa. Penelitian ini bertujuan untuk mengukur tingkat akurasi dari kajian komparasi 3 buah algoritma dalam pemilihan calon penerima beasiswa di SMA Muhammadiyah 4 Jakarta. Dari hasil pengujian dengan mengukur kinerja ketiga algoritma tersebut menggunakan metode pengujian *Cross Validation*, *Confusion Matrix* dan Kurva ROC, diketahui bahwa algoritma *Naïve Bayes* memiliki nilai *accuracy*paling tinggi, yaitu 86.5672%, diikuti oleh metode *C4.5* dengan *accuracy* sebesar 67.1642% dan yang terendah adalah metode *Neural Network* dengan nilai accuracy 82.0896%.

Kata Kunci*: C4.5, Naïve Bayes, Neural Network.*

Pendahuluan

Pendidikan di Indonesia selalu berhadapan dengan persoalan kemiskinan. Kemiskinan ini kemudian menjadi alasan seseorang mengajukan permohonan beasiswa kepada lembaga pendidikan. Berbagai langkahpun ditempuh, antara lain melalui ketersediaan dana pembebasan biaya sekolah melalui program Surat Keterangan Tidak Mampu (SKTM) yang dikeluarkan dari kantor kecamatan setempat. Agar penerima beasiswa ini terus terpacu meningkatkan prestasi akademisnya, tidak keliru apabila lembaga pendidikan tidak serta merta mengabulkan permohonan beasiswa yang diajukan.

Untuk mengatasi masalah tersebut, maka diuji menggunakan 3 buah metode algoritma yaitu algoritma C4.5, algoritma naïve bayes dan neural network. Dari ketiga buah metode tersebut akan dikaji metode mana yang paling akurat digunakan untuk mengukur tingkat kelayakan para siswa/siswi dalam menerima bantuan beasiswa.

Tinjauan Pustaka

Beasiswa adalah penghasilan bagi yang menerimanya (Anneahira, 2012). Dalam ketentuan pasal 4 ayat (1) UU PPh/2000 pengertian penghasilan adalah tambahan kemampuan ekonomis dengan nama dan dalam bentuk apapun yang diterima atau diperoleh dari sumber Indonesia atau luar Indonesia yang dapat digunakan untuk konsumsi atau menambah kekayaan wajib pajak (WP).

Data Mining adalah sebuah proses, yang mana dalam melakukan prosesnya harus sesuai dengan prosedur dari proses tersebut, yaitu CRISP-DM (*Cross-Industry Standard Process for Data Mining),* yang terdiri dari keseluruhan proses, *preprosesing* data, pembentukan model, model evaluasi, dan tahap akhir penyebaran model (Larose, 2005).

Quinlan (1993) mengemukakan bahwa C4.5 adalah algoritma yang digunakan untuk klasifikasi data yang dapat mengolah data/atribut numerik, algoritma ini dapat mengatasi nilai atribut yang hilang, dan dapat mengatasi data kontinyu dan *pruning* / penyederhanaan.

Algoritma Naive bayes merupakan salah satu metode pengklasifikasi berpeluang sederhana yang berdasarkan pada penerapan Teorema Bayes dengan asumsi antar variabel penjelas saling bebas (independen).

Neural network adalah satu set unit input/output yang terhubung dimana tiap relasinya memiliki bobot. Neural network dimaksudkan untuk mensimulasikan perilaku system biologi susunan syaraf manusia, yang terdiri dari sejumlah besar unit pemroses yang disebut neuron, yang beroperasi secara parallel (Alpayadin, 2010).

Weka (*Waikato Environment for Knowledge Analysis*) adalah sebuah alat (*tool*) yang merupakan aplikasi *data mining* berbasis *open source (GPL)* yang ditulis dengan Java.

Metodologi Penelitian

Jenis penelitian ini adalah sebagai berikut:

1. Penelitian Eksperimental
   Penelitian eksperimental merupakan penelitian yang bersifat uji coba, memanipulasi dan mempengaruhi hal-hal yang terkait dengan seluruh variabel atau atribut.
2. Penelitian Perbandingan atau studi komparasi yakni dengan membandingkan antara tiga macam algoritma yaitu algoritma C4.5, Naïve Bayes dan Neural Network.

Hasil dan Pembahasan

Penelitian ini bertujuan untuk menentukan akurasi kelayakan pemberian beasiswa yang dibandingkan dengan menggunakan metode algoritma C4.5, *Naïve bayes*, dan *Neural Network*. Setelah itu membandingkan nilai akurasi ketiga metode tersebut, dalam menentukan hasil penelitian ini menggunakan data *training* berjumlah 234 data dan data *testing* berjumlah 197 data.
Langkah-langkah untuk membuat algoritma C.45 dengan memakai data *training* yang berjumlah 197, yaitu:

1. Siapkan data training. Data testing yang digunakan penelitian ini sebanyak 197 data.
2. Hitung nilai entropy
   Setelah dilakukan perhitungan *entropy* dengan menggunakan rumus dibawah ini didapat *entropy* sebagai berikut:

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

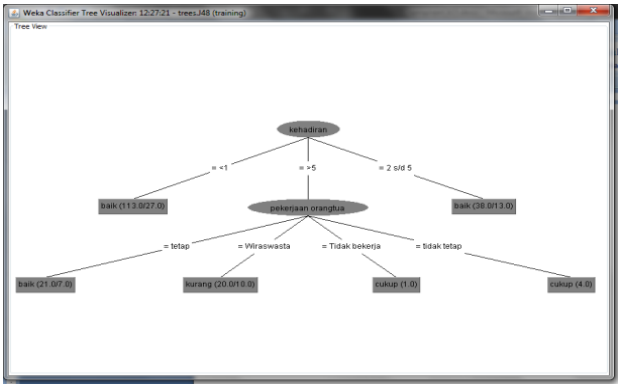
$$= (80/197 . \log_2 (80/197)) + (-117/173 . \log_2 (117/197))$$
$$= - 2.0518$$

3. Setelah itu, hitung nilai *gain* untuk setiap atribut, lalu pilih nilai *gain* yang tertinggi. Nilai gain tertinggi itulah yang akan menjadi akar dari pohon. Perhitungan Gain menggunakan rumus dibawah ini. Misalkan untuk atribut kehadiran akan didapat gain sebagai berikut:

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

Gain (S, A) = -2.0518 – (113/197*-2.0010) + (38/197*-2.0364) + (46/197*-3.1402)

= - 2.03008431



Gambar 1. Gambar Pohon Keputusan dari C4.5

Dari gambar pohon keputusan diatas didapat aturan-aturan sebagai berikut:

1. IF kehadiran = <1: baik AND kehadiran = >5 THEN diterima
2. IF pekerjaan orangtua tetap: baik THEN diterima
3. IF pekerjaan orangtua = Wiraswasta: kurang THEN terima
4. IF pekerjaan orangtua = Tidak bekerja: cukup THEN terima
5. IF pekerjaan orangtua = tidak tetap: cukup THEN terima
6. IF kehadiran = 2 s/d 5: baik THEN terima

Tabel 2. Hasil dan Nilai Entropy dan Gain untuk menentukan simpul akar dengan data training

Simpul		Kasus	Terima	Tidak	Entropy	Gain
Kelas		131	57	74	-2.0245	-3.1023
	X-1	19	6	13	-2.2105	
	X-2	7	3	4	-2.0297	
	X-3	24	8	16	-2.1699	
	XI IPA-1	22	10	12	-2.0120	
	XI IPA-2	36	16	20	-2.0179	
	XI IPS	23	8	15	-2.1402	
Pekerjaan orangtua		131	57	74	-2.0245	-1.3984
	wiraswasta	39	18	21	-2.0086	
	tetap	78	24	54	-2.2310	
	tidak tetap	11	8	3	-2.3339	
	tidak bekerja	3	1	2	-2.1699	
Penghasilan orangtua		131	57	74	-2.0245	-2.6255
	<1000000	53	26	27	-2.0005	
	1050000-2000000	37	17	20	-2.0095	
	>2000000	38	7	31	-2.7343	
	tidak ada	3	1	2	-2.1699	
Pengeluaran		131	57	74	-2.0245	-2.4436
	<1000000	56	27	29	-2.0018	
	1050000-2000000	47	17	30	-2.1148	
	>2000000	28	7	21	-2.4150	

Simpul		Kasus	Terima	Tidak	Entropi	Gain
Jumlah tanggungan		131	57	74	-2.0245	-3.7971
	1	10	4	6	-2.0589	
	2-4 orang	101	37	64	-2.1070	
	>4	20	10	10	-2.0000	
Kehadiran		131	57	74	-2.0245	-1.9300
	<1	76	37	39	-2.0010	
	2 s/d 5	24	9	15	-2.0931	
	>5	31	5	26	-2.8860	
Kegiatan pengembangan diri		131	57	74	-2.0245	-3.0005
	baik	89	38	51	-2.0311	
	cukup	29	11	18	-2.0866	
	kurang	13	2	11	-2.9414	

Perhitungan *entropy* dan *gain* untuk semua atribut dilakukan, untuk mendapatkan nilai gain tertinggi. Hasil perhitungan seluruh atribut terlihat pada Tabel 1.

Perhitungan Naïve Bayes.

Dengan mencari *prior probability* untuk nilai yang diterima dan tidak diterima untuk semua jumlah data. Jika diketahui dalam data *training*, jumlah data 197, siswa yang diterima beasiswa dalam kelas terima 80 *record* dan yang tidak diterima dalam kelas tidak 119 record.

Tabel 3. Perhitungan probabilitas prior

P		Terima	Tidak	P(X|Ci)	
				Terima	Tidak
total		80	117	0.1663	0.2432
Kelas				0.0000	0.0000
	X-1	12	24	0.0249	0.0499
	X-2	13	22	0.0270	0.0457
	X-3	13	19	0.0270	0.0395
	XI IPA-1	10	12	0.0208	0.0249
	XI IPA-2	16	20	0.0333	0.0416
	XI IPS	16	20	0.0333	0.0416
Pekerjaan orangtua				0.0000	0.0000
	wiraswasta	34	37	0.0707	0.0769
	tetap	35	73	0.0728	0.1518
	tidak tetap	8	5	0.0166	0.0104
	tidak bekerja	3	2	0.0062	0.0042
Penghasilan orangtua				0.0000	0.0000
	<1000000	44	44	0.0915	0.0915
	1050000-2000000	25	24	0.0520	0.0499
	>2000000	10	47	0.0208	0.0977
	tidak ada	1	2	0.0021	0.0042

Pengeluaran				0.0000	0.0000
	<1000000	45	46	0.0936	0.0956
	1050000-2000000	26	44	0.0541	0.0915
	>2000000	9	27	0.0187	0.0561
Jumlah tanggungan				0.0000	0.0000
orangtua	1	4	7	0.0083	0.0146
	2-4 orang	60	93	0.1247	0.1933
	>4	16	17	0.0333	0.0353
Ranking				0.0000	0.0000
	1	8	0	0.0166	0.0000
	2 s/d 4	17	0	0.0353	0.0000
	5 s/d 10	41	4	0.0852	0.0083
	tidak ada	14	113	0.0291	0.2349
Nilai raport				0.0000	0.0000
	<60	0	0	0.0000	0.0000
	61-70	0	8	0.0000	0.0166
	71-80	38	102	0.0790	0.2121
	>80	42	7	0.0873	0.0146
Kehadiran				0.0000	0.0000
	<1	58	55	0.1206	0.1143
	2 s/d 5	16	22	0.0333	0.0457
	>5	6	40	0.0125	0.0832
Kepribadian				0.0000	0.0000
	baik	80	113	0.1663	0.2349
	cukup	0	2	0.0000	0.0042
	kurang	0	2	0.0000	0.0042
Kegiatan pengembangan diri				0.0000	0.0000
	baik	57	74	0.1185	0.1538
	cukup	19	24	0.0395	0.0499
	kurang	4	19	0.0083	0.0395

Untuk menentukan kasus baru termasuk kelas mana, dilakukan perhitungan probabilitas *posterior* berdasarkan probabilitas *prior* probabilitas *posterior* untuk menentukan data testing termasuk klasifikasi mana terdapat pada Tabel 4. Misalkan diambil sebuah data testing X dengan nilai seperti pada Tabel 3 kolom dua, untuk menentukan kelas mana, dilakukan perhitungan probabilitas *posterior* yang hasilnya terdapat pada Tabel 3 kolom tiga dan empat.

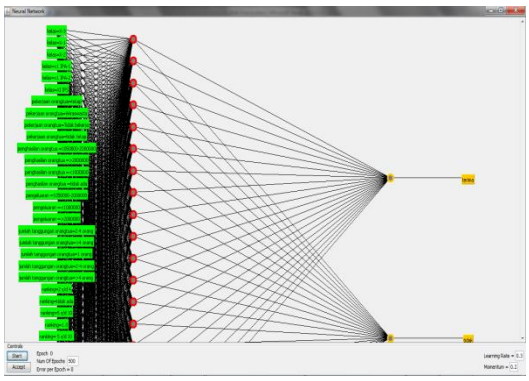
Tabel 4. Perhitungan Probabilitas *Posterior*

Data X		P(X|Ci)	
Atribut	**Nilai**	**Terima**	**Tidak**
Kelas	X-1	0.0249	0.0499
Pekerjaan orangtua	tetap	0.0728	0.1518
Penghasilan orangtua	1050000-2000000	0.0520	0.0499
Pengeluaran	>2000000	0.0187	0.0561
Jumlah tanggungan orangtua	2-4 orang	0.1247	0.1933
Ranking	5 s/d 10	0.0852	0.0083
Nilai raport	71-80	0.0790	0.2121
Kehadiran	2 s/d 5	0.0333	0.0457
Kepribadian	baik	0.1663	0.2349
Kegiatan pengembangan diri	cukup	0.0395	0.0499

Dari hasil perhitungan di atas, didapat nilai $P(X|C_i)$ dan $P(X|C_i)$ $P(C_i)$ lebih besar untuk *remark = terima* sehingga dapat disimpulkan bahwa data testing tersebut termasuk klasifikasi *good*.

Neural Network

Gambar 2 adalah *neural net* yang dihasilkan dari pengolahan data *training* dengan metode *neural network* adalah *multilayer perceptron* yang dihasilkan dari data *training* pada Tabel . Terdiri dari tiga *layer*, yaitu *Input layer* terdiri dari 21 simpul, sama dengan jumlah atribut prediktor ditambah satu simpul bias. Pada pembahasan ini digunakan satu*hidden layer* yang terdiri dari 20 simpul diambah satu simpul bias. Di bagian *output layer* terdapat dua simpul yang mewakili atribut kelas yaitu *terima* dan *tidak.*



Gambar 2. *Neural net*

Tabel 5 adalah nilai akhir fungsi aktifasi pada *output layer*. Kolom pertama pada Tabel 5 menyatakan *class*, yaitu atribut kelas yang dinyatakan dengan simpul pada *output layer* seperti pada gambar 2.

Tabel 5. Nilai Bobot Akhir untuk Output Layer

	output (Sigmoid)																					
class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	threshold
terima	1.6758	3.1446	-0.3007	2.1844	-2.916	1.9137	-2.659	0.3096	0.6891	3.3385	2.7988	-1.0518	-0.2924	-3.0044	1.0532	-2.3046	1.0117	-1.3144	0.166	-2.5816	1.272	-1.7857
tidak	-1.673	-3.141	0.2989	-2.1962	2.901	-1.912	2.6751	-0.3282	-0.7028	-3.3253	-2.8029	1.0377	0.2852	3.009	-1.0573	2.3226	-1.0048	1.3137	-0.1248	2.573	-1.2865	1.7842

Evaluasi dan Validasi

1. Pengujian model

Model yang telah dibentuk diuji tingkat akurasinya dengan memasukan data uji yang berasal dari data *training*. Karena data yang didapat dalam penelitian ini setelah proses *preprocessing* hanya 197 data maka digunakan metode *cross validation* untuk menguji tingkat akurasi. Untuk nilai akurasi model untuk metode C4.5 sebesar 67.1642% metode *naïve bayes* sebesar 86.5672 %, dan metode *neural network* sebesar 82.0896 %.

Confision Matrix

Tabel 6. Model *Confusion Matrix* untuk Metode C4.5

```
=== Confusion Matrix ===

  a  b   <-- classified as
 22  6 |  a = terima
  1 38 |  b = tidak
```

Tabel 7. Model *Confusion Matrix* untuk Metode Naïve bayes

```
=== Confusion Matrix ===

  a  b   <-- classified as
 21  7 |  a = terima
  2 37 |  b = tidak
```

Tabel 8. Model *Confusion Matrik* Metode *Neural Network*

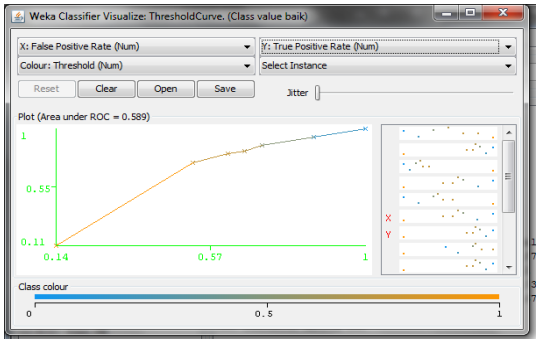
```
=== Confusion Matrix ===

  a  b   <-- classified as
 22  6 |  a = terima
  6 33 |  b = tidak
```

Dari tiga table *confusion matrix*, 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* dapat dilihat pada Tabel 9.

Tabel 9. Komparasi Nilai *Accuracy, Precision*, dan *Recall*

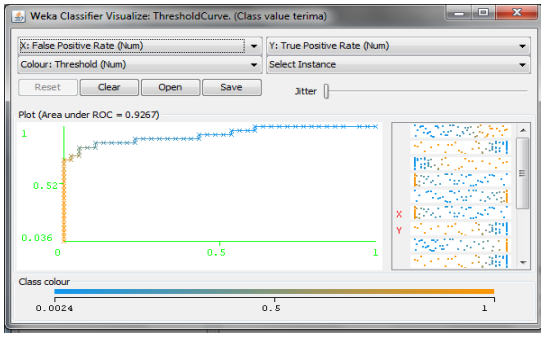
	C4.5	*Naïve Bayes*	*Neural network*
Accuracy	67.1642%	86.5672 %	82.0896 %
Precision	0.727%	0.913%	0.786%
Recall	0.87%	0.75%	0.786%

Kurva ROC



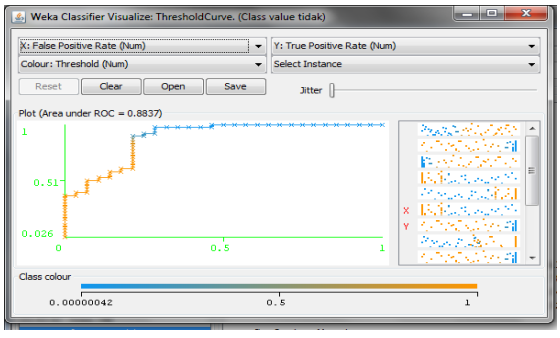
Gambar 3. Kurva ROC Algoritma C4.5

Kurva ROC pada gambar 3 diatas mengekspresikan *confusion matrix* Garis X adalah *false positives* dan garis Y *true positives*, sebesar 0.589.



Gambar 4. Kurva ROC Naïve Bayes

Hasil yang didapat dari pengolahan ROC untuk *Naïve Bayes* sebesar 0.9267 dapat dilihat pada gambar 4.



Gambar 5. Kurva ROC Neural Network

Hasil yang didapat dari pengolahan ROC untuk *Neural Network* sebesar 0.8837 dapat dilihat pada gambar 5.

Setelah dilakukan perhitungan pada kurva ROC maka didapatlah hasil perbandingan tersebut. Perbandingan hasil perhitungan nilai AUC untuk metode C4.5, *naïve bayes*, dan *neural network* dapat dilihat pada Tabel 10.

Tabel 10. Komparasi Nilai AUC

	C4.5	*Naïve Bayes*	*Neural network*
AUC	0.589	0.9267	0.8837

Analisis hasil komparasi

Model yang dihasilkan dengan metode C4.5, *naïve bayes*, dan *neural network* diuji menggunakan metode *Cross Validation*, terlihat perbandingan nilai *accuracy*, *precision*, dan *recall* pada Tabel 11.

Tabel 11. Komparasi Nilai A*ccuracy* dan AUC

	C4.5	*Naïve Bayes*	*Neural network*
Accuracy	67.1642%	86.5672 %	82.0896 %
AUC	0.589	0.9267	0.8837

Tabel 11 membandingkan *accuracy* dan AUC dari tiap metode. Terlihat bahwa nilai *accuracy* Naïve Bayes paling tinggi begitu pula dengan nilai AUC-nya. Untuk metode *C4.5* dan *Neural Network* juga menunjukan nilai yang sesuai.

Penerapan Algoritma Terpilih

Tabel 12. Data Baru untuk Penerapan Algoritma Terpilih

kelas	pekerjaan orangtua	penghasilan orangtua	pengeluaran	jumlah tanggungan	ranking	nilai raport	kehadiran	kepribadian	kegiatan peng- diri	Hasil
X-1	Wiraswasta	<1000000	<1000000	2-4 orang	1	>80	<1	baik	cukup	terima
X-1	tetap	>2000000	1050000-2000	2-4 orang	tidak ada	71-80	2 s/d 5	baik	kurang	tidak
X-2	Wiraswasta	<1000000	<1000000	>4 orang	tidak ada	71-80	2 s/d 5	baik	baik	tidak
X-2	Wiraswasta	<1000000	<1000000	2-4 orang	5 s/d 10	>80	<1	baik	cukup	terima
X-3	tetap	>2000000	>2000000	2-4 orang	5 s/d 10	>80	<1	baik	baik	terima
X-3	tetap	>2000000	>2000000	2-4 orang	tidak ada	71-80	>5	baik	baik	tidak
x1 IPA-1	Wiraswasta	<1000000	<1000000	2-4 orang	5 s/d 10	>80	<1	baik	baik	terima
x1 IPA-2	Wiraswasta	<1000000	<1000000	2-4 orang	5 s/d 10	>80	<1	baik	baik	terima
XI IPS	tidak tetap	<1000000	<1000000	2-4 orang	5 s/d 10	71-80	2 s/d 5	baik	baik	terima
XI IPS	tetap	>2000000	1050000-2000	2-4 orang	tidak ada	61-70	<1	kurang	baik	tidak

Hasil penerapan *rule* algoritma Naïve Bayes terhadap data baru sejumlah 10 record data dimana 6 data diprediksi terima dan 4 data diprediksi tidak. Dengan tingkat akurasi sebesar 70%. Dan dapat dilihat juga tabel dari *Confusion Matrix* pada tabel 13.

Tabel 13. *Confussion Matrix* Data Baru dengan Algoritma *Naïve Bayes*

```
=== Confusion Matrix ===

 a b   <-- classified as
 4 2 | a = terima
 1 3 | b = tidak
```

Kesimpulan

Dalam penelitian ini dilakukan pembuatan model menggunakan algoritma C4.5, *naïve bayes* dan *neural network* menggunakan data siswa yang menerima beasiswa di sekolah. Model yang dihasilkan, dikomparasi untuk mengetahui algoritma yang paling baik dalam pemilihan penerima beasiswa. 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 pemilihan penerima beasiswa di sekolah.

Saran

1. Hasil penelitian ini diharapkan bisa digunakan pada sekolah, untuk lebih meningkatkan akurasi analisa penerima beasiswa bagi siswa.
2. Untuk mendukung pengambilan keputusan dan pengembangan system informasi manajemen strategik, model ini dapat diterapkan pada sekolah dengan menerapkan system yang menggunakan perangkat keras dan perangkat lunak, disertai dengan pembuatan *Standard Operational Procedure* dan pelatihan bagi *end-user*.

Daftar Pustaka

Alpayadin, E. (2010). *Introduction to Machine Learning*. London: The MIT Press.

Anneahira. (2012). Tujuan Beasiswa Dalam http://www.anneahira.com/beasiswa.htm(diakses pada tanggal 3 Juni 2012).

Kusrini, dan Luthfi, Emha Taufik. (2009). *Algoritma Data Mining.* Edisi I. Yogyakarta:Andi Publishing.

Larose, Daniel. T. (2005). *Discovering Knowledge in Data*. New Jersey: John Willey & Sons, Inc.