ANALISIS KOMPARASI ALGORITMA KLASIFIKASI DATA MINING UNTUK STATUS KEMISKINAN RUMAH TANGGA

Erfan Karyadiputra, Galih Mahalisa, dan Agus Alim Muin
Fakultas Teknologi Informasi, Universitas Islam Kalimantan
E-mail : erfantsy@gmail.com

ABSTRAK

Kemiskinan merupakan salah satu permasalahan yang sering dihadapi dalam upaya peningkatan kesejahteraan suatu negara. Di Indonesia, sebagai negara berkembang angka kemiskinan masih cukup tinggi sehingga berdampak pada kesenjangan sosial di masyarakat. Oleh sebab itu, tersedianya data kemiskinan yang akurat, komprehensif dan berkesinambungan merupakan salah satu instrumen penting bagi pengambil kebijakan dalam memfokuskan perhatian pada pendistribusian bantuan sesuai rumah tangga sasaran (RTS). Data kemiskinan yang baik dapat menjadi informasi terpercaya untuk mengevaluasi kebijakan pemerintah dalam mengentaskan kemiskinan. Pemerintah telah banyak memberikan solusi kebijakan untuk mengurangi dampak kemiskinan salah satunya berupa program bantuan sosial yang pendistribusian setiap jenis bantuan tersebut berbeda-beda sesuai dengan kategori rumah tangga sasaran (RTS) berdasarkan kriteria kemiskinan non-monetary. Dalam penelitian ini akan menganalisis algoritma klasifikasi data mining dengan membandingkan hasil klasifikasi dari algoritma naive bayes dengan algoritma decision tree menggunakan confusion matrix. Dari hasil komparasi tersebut dapat dibandingkan nilai akurasi terbaik sehingga algoritma yang didapatkan merupakan algoritma yang paling akurat dalam melakukan klasifikasi status kemiskinan rumah tangga.

Kata Kunci : Data mining, Decision Tree, Klasifikasi, Naïve Bayes

ABSTRACT

Poverty is one of the problems frequently encountered in efforts to improve the well-being of a country. In Indonesia, as a developing country poverty rates are still high enough to have an impact on social inequalities in society. Therefore, the availability of poverty data is accurate, comprehensive and sustainable is an important instrument for policy makers to focus on the distribution of aid according targeted households (RTS). Poverty data that could either be reliable information to evaluate government policies to alleviate poverty. The government has made a policy solution to reduce the impact of poverty one form of social assistance programs that the distribution of each type of such assistance varies according to the category of targeted households (RTS) based on the criteria of non-monetary poverty. In this study will analyze the classification of data mining algorithms by comparing the results of the algorithm Naive Bayes classification with decision tree algorithm using the confusion matrix. The comparison of the results can be compared to the best accuracy values obtained so that the algorithm is an algorithm is most accurate in the classification status of household poverty.

Keywords: Data mining, Decision Tree, classification, Naive Bayes

PENDAHULUAN

Kemiskinan masih menjadi salah satu permasalahan yang harus ditanggulangi secara tuntas terutama di negara-negara berkembang. Di Indonesia sebagai salah satu negara berkembang angka kemiskinan masih cukup tinggi sehingga berdampak pada kesenjangan sosial di masyarakat. Negara berkembang seperti Indonesia dicirikan dengan kemiskinan karena pada umumnya di negara berkembang, permasalahan pendapatan yang rendah dan kemiskinan masih merupakan masalah utama dalam pembangunan ekonomi. Usaha Pemerintah dalam hal ini Kementrian Sosial memiliki beberapa program penanggulangan kemiskinan diantaranya dengan membentuk beberapa program bantuan sosial berbasiskan keluarga yang pendistribusian setiap jenis bantuan tersebut berbeda-beda sesuai dengan kategori rumah tangga sasaran (RTS). Tersedianya data kemiskinan yang akurat, komprehensif dan berkesinambungan merupakan salah satu instrumen penting bagi pengambil kebijakan dalam memfokuskan perhatian pada pendistribusian bantuan sesuai rumah tangga sasaran (RTS). Data kemiskinan yang baik dapat menjadi informasi terpercaya untuk mengevaluasi kebijakan pemerintah dalam mengentaskan kemiskinan. Penelitian yang akan dilakukan adalah menganalisis komparasi algoritma klasifikasi *data mining* dengan membandingkan hasil klasifikasi dari beberapa algoritma seperti *naive bayes* dan algoritma *decision tree* menggunakan *confusion matrix.* Dari hasil komparasi tersebut dapat dibandingkan nilai akurasi terbaik sehingga algoritma yang didapatkan merupakan algoritma yang tepat dalam melakukan klasifikasi status kemiskinan rumah tangga.

METODE PENELITIAN

Adapun metode penelitian yang dilakukan pada penelitian ini adalah sebagai berikut: Metode pengumpulan data yang digunakan dalam penelitian ini adalah berasal dari hasil pendataan PPLS 2011. Variabel yang akan digunakan pada penelitian ini adalah **:**

Tabel 1. Penjelasan Variabel dan Kategori

Variabel	Keterangan	Skala	Kategori
Y	Status Rumah Tangga Sasaran (RTS)	Nominal	1 : Sangat Miskin RTSM
			2 : Miskin RTM
X1	Jenis Kelamin KRT	Nominal	1 : Laki-Laki
			2 : Perempuan
X2	Umur KRT	Numerik	-
X3	Pendidikan KRT	Nominal	0 : Tidak Punya Ijazah
			1 : SD/Sederajat
			2 : SMP/Sederajat
			3 : SMA/Sederajat
X4	Lapangan Usaha KRT	Nominal	1 : Pertanian (Padi & Palawija)
			2 : Hortikultura
			3 : Perkebunan
			4 : Perikanan Tangkap

			5 : Perikanan Budidaya
			6 : Peternakan
			7 : Kehutanan & Pertanian Lain
			8 : Pertambangan / Penggalian
			9 : Bangunan / Konstruksi
			10 : Pedagang
X5	Status Kependudukan dalam Pekerjaan Kepala Rumah Tangga	Nominal	1 : Berusaha Sendiri
			2 : Berusaha dibantu Buruh tidak tetap / tidak dibayar
			3 : Berusaha dibantu Buruh tetap / dibayar
			4 : Buruh/ Karyawan/ Pegawai Swasta
			5 : Pekerja Bebas
			6 : Pekerja Keluarga/Tidak dibayar
X6	Status Penguasaan Bangunan Tempat Tinggal	Nominal	1 : Milik Sendiri
			2 : Kontrak/ Sewa
X7	Jenis Atap Terluas	Nominal	1 : Beton
			2: Genteng
			3 : Sirap
			4 : Seng
			5 : Asbes
			6 : Ijuk/ Rumbia
X8	Kualitas Atap	Nominal	1 : Biasa/Kualitas Sedang
			2 : Jelek/Kualitas Rendah
X9	Jenis Dinding Terluas	Nominal	1 : Tembok
			2 : Kayu
			3 : Bambu
X10	Kualitas Dinding	Nominal	1 : Biasa/Kualitas Sedang
			2 : Jelek/Kualitas Rendah
X11	Jenis Lantai	Nominal	1 : Kayu / bambu
			2 : Semen Tanpa Plester
X12	Sumber Air Minum	Nominal	1 : Air Kemasan
			2 : Air Ledeng
			3 : Air Terlindung
			4 : Air Tidak Terlindung
X13	Bahan Bakar Utama Memasak	Nominal	1 : Listrik/ Gas/ Elpiji
			2 : Minyak Tanah
			3 : Kayu
X14	Sumber Penerangan	Nominal	1 : Listrik PLN
			2 : Listrik Non-PLN
			3 : Tidak Ada Listrik
X15	Jumlah Keluarga	Numerik	-
X16	Jumlah Individu	Numerik	-

Dataset diatas memiliki 1 variabel sebagai kelas yaitu status rumah tangga sangat miskin (RTSM) dan status rumah tangga miskin (RTM) dan 16 variabel sebagai atribut. Sebagian besar variabel atribut bertipe data *nominal* kecuali atribut umur, atribut jumlah keluarga dan atribut jumlah individu.

Metode Pengolahan Data Awal

Tahapan selanjutnya setelah pengumpulan data maka data tersebut kemudian diolah agar dapat diproses dalam *data mining*.

Eksperimen dan Pengujian Metode

Terdapat beberapa tahapan dalam melakukan eksperimen, yaitu :

Tahap 1 : Tahap pertama melakukan pengujian klasifikasi pertama menggunakan algoritma *naive bayes* menggunakan data original yang masih terdapat data kosong. Kemudian dilakukan validasi model klasifikasi dilakukan terhadap data *testing*.

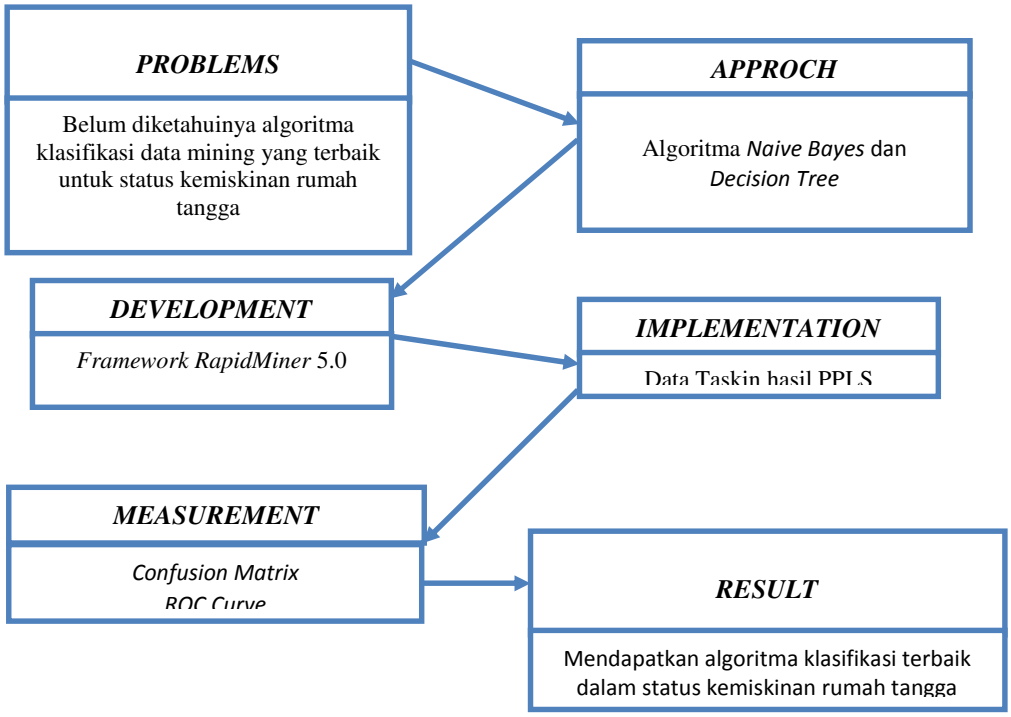
Tahap 2 : Setelah melakukan pengujian dengan data original maka selanjutnya dilakukan pengisian data kosong menggunakan *replace missing value.* Pengisian data kosong dilakukan untuk membuat model keputusan yang baik, sehingga harus menggunakan data yang baik pula (lengkap, benar, konsisten, terintegrasi).

Tahap 3 : Tahap selanjutnya memisahkan dataset menjadi 2 jenis data, yakni data *training* dan data *testing* dengan menggunakan pembagian persentase dari jumlah dataset. Apabila data *training* yang digunakan sebanyak 10% dari dataset maka persentase data *testing* adalah 90%. Dan apabila data *training* yang digunakan sebanyak 20% dari dataset maka persentase data *testing* adalah 80%. Jumlah data *training* apabila dijumlah dengan data *testing* akan menjadi jumlah dataset tersebut. Jumlah persentase pada data *trainning* nantinya akan ditentukan berdasarkan perulangan pada proses *cross-validation*. Contohnya, proses *cross-validation* pada iterasi ke 4 dari total 10 iterasi, maka persentase jumlah data *training* adalah 40% dan persentase jumlah data *testing* adalah 60%. Proses pemisahan data *training* dan data *testing* mengikuti proses *X-validation* dengan jumlah iterasi mulai dari 2,3,4,5,6,7,8,9 dan 10 dan agar data yang digunakan pada proses itu tetap maka akan menggunakan dataset yang sama.

Tahap 4 : Mengevaluasi ketepatan beberapa algoritma klasifikasi seperti *naïve bayes* dan *decision tree* dan validasi hasil dilakukan pengujian 10-*fold cross-validation* yang akan mengulang pengujian sebanyak 10 kali dan hasil pengukuran adalah nilai rata-rata dari 10 kali pengujian.

Tahap 5 : Melakukan pengujian *performance* dengan menggunakan *confusion matrix* sehingga dapat diketahui hasil akurasi dan nilai AUC. Nilai AUC digunakan untuk menentukan hasil klasifikasi kedalam klasifikasi sangat baik, klasifikasi baik, klasifikasi cukup, klasifikasi buruk dan klasifikasi salah.

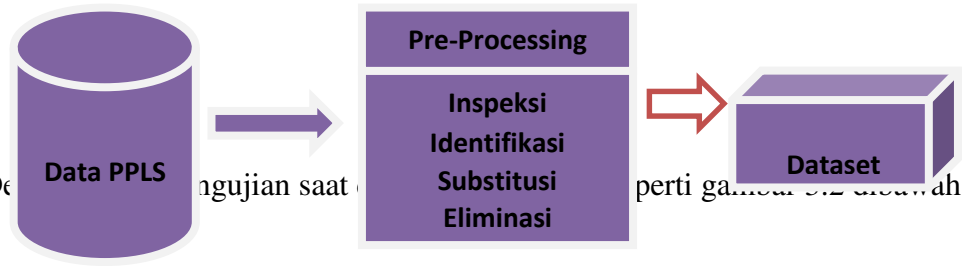
Kerangka Pemikiran



HASIL DAN PEMBAHASAN

Hasil Eksperimen dan Pengujian Model/Metode

Sebelum melakukan eksperimen dan pengujian model, data-data yang dikumpulkan terlebih dahulu diolah agar dapat diproses dalam *data mining*. Pada tahap pertama pengujian dilakukan dengan data original yang sebagian masih memiliki data kosong dan belum diolah/dimodifikasi. Kemudian pengujian dilakukan menggunakan algoritma *naive bayes* dengan validasi model klasifikasi dilakukan terhadap data *testing* dengan teknik 10-*folds cross validation.*



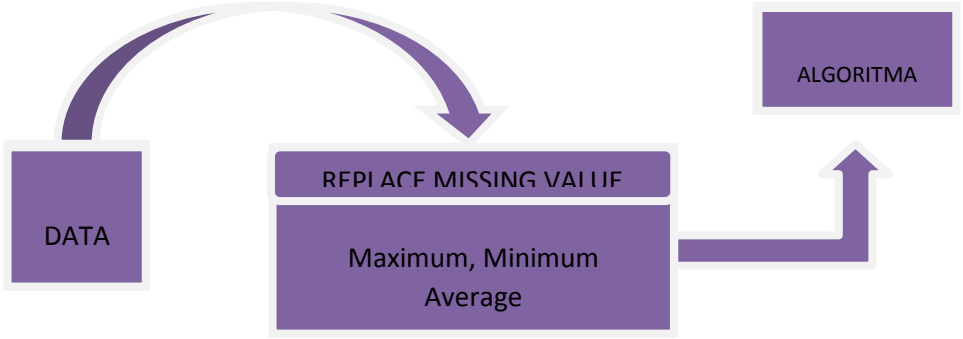
De ngujian saat perti ga ah ini :

Gambar 1 Pre-Processing

Tabel 3. 1 Akurasi Naive Bayes Menggunakan Data Original

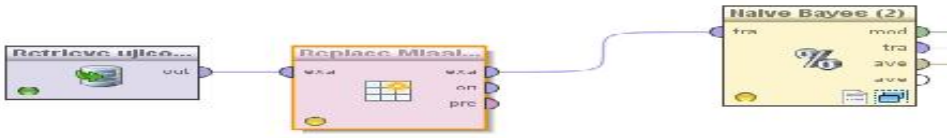
accuracy = 89.79% +/- 2.34% (mikro: 89.79%)			
	True Miskin	True Sangat Miskin	Class precision
Pred Miskin	478	42	91.92%
Pred Sangat Miskin	63	445	87.60%
Class recall	88.35%	91.38%	

Dari hasil pengujian pada tabel 3. menggunakan metode *naive bayes* dengan data original, maka didapatkan akurasi sebesar 89,79% dengan nilai AUC sebesar 0.959. Selanjutnya dilakukan *pre-processing* untuk memastikan data yang akan diolah dalam *data mining* adalah data yang baik dan lengkap sehingga menghasilkan model keputusan yang baik pula. Salah satu teknik yang dapat digunakan untuk pengisian data kosong yang terdapat dalam *data mining* seperti metode *replace missing values.*



Gambar 3. Model Pengisian Data Menggunakan Replace Missing value

Beberapa metode *replace missing values* dalam *data mining* dalam pengisian data kosong diantaranya berdasarkan nilai *minimum, maximum, average dan zero*. Oleh sebab itu perlu dilakukan pengujian dari setiap teknik tersebut sehingga didapatkan teknik terbaik yang sesuai dengan karakter data yang diuji coba. Dari beberapa percobaan tersebut didapatkan metode *replace missing values* bertipe *average* yang cocok. oleh sebab itu selanjutnya metode *replace missing values* bertipe *average* akan digunakan pada percobaan selanjutnya. Hasil pengujian yang dilakukan dengan data yang telah dilengkapi dan dimodifikasi menggunakan metode *replace missing values* bertipe *average* ketika diimplementasi menghasilkan data sebagai berikut :



Gambar.4 Implementasi Pengisian Data Menggunakan Replace Missing value

Hasil pengujian yang dilakukan dengan data yang telah dilengkapi dan dimodifikasi menggunakan metode *replace missing values* ketika diimplementasi menghasilkan data sebagai berikut :

Tabel 3. Akurasi Naive Bayes Menggunakan Data Modifikasi

accuracy = 85.80% +/- 3.44% (mikro: 85.80%)			
	Pred Miskin	Pred Sangat Miskin	Class precision
Pred Miskin	449	54	89.26%
Pred Sangat Miskin	92	433	82.48%
Class recall	82.99%	88.91%	

Dari hasil pengujian membuktikan bahwa algoritma *naive bayes* sudah dapat diterapkan untuk mengidentifikasi status kemiskinan rumah tangga miskin dengan akurasi 85.80% .

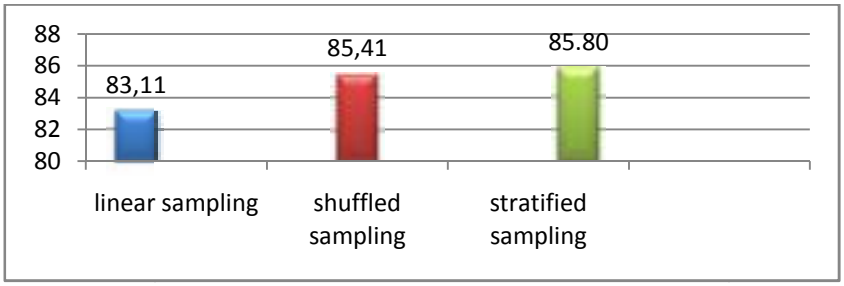
Evaluasi Dan Validasi Hasil

Sebelum melakukan evaluasi dan validasi, terlebih dahulu akan dilakukan percobaan dengan mengganti jenis *parameter* pada *X-validation* yaitu *sampling type* yang terdiri atas tiga jenis, diantaranya *linear sampling*, *shuffled sampling*, dan *stratified sampling* sehingga didapatkan *sampling type* yang terbaik dan sesuai untuk digunakan pada data yang diuji.

Tabel 3. Perbandingan Sampling Type

Sampling Type	***Linear Sampling***	***Shuffled Sampling***	***Stratified Sampling***
Accuracy	83.18	85.41	85.80
AUC	0.931	0.937	0.930

Tabel 3. terlihat penggunaan *X-Validation* dengan *sampling type stratified* memiliki tingkat akurasi yang sedikit lebih baik daripada *sampling type* yang lain meskipun dalam hal kehandalan klasifikasi masih sedikit lebih unggul penggunaan *sampling type stratified*. Perbandingan akurasi dari beberapa tipe validasi dalam bentuk grafik seperti gambar dibawah ini



Gambar 5 Grafik Perbandingan Tipe Validasi

Percobaan selanjutnya akan dilakukan pengujian algoritma *naive bayes* dengan teknik *folds cross validation* dengan pengujian data mulai 2,3,4,5,6,7,8,9 dan 10 kemudian dievaluasi dan dibandingkan dengan algoritma *decision tree*.

Tabel 3. 4 Hasil Pengujian Naive Bayes

Validation	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Accuracy	85.70	85.80	**86.28**	85.50	85.70	85.80	86.19	85.60	85.80
Precision	82.96	82.44	83.31	82.12	82.47	82.96	83.21	82.63	82.75
Recall	87.89	89.33	88.92	89.12	88.91	88.31	89.13	88.71	88.90
AUC	0.937	0.935	**0.940**	0.927	0.938	0.935	0.935	0.938	0.930

Berdasarkan penelitian dari berbagai percobaan yang ekstensif dan pembuktian teoritis, menunjukkan bahwa penggunaan 10-*fold cross-validation* adalah pilihan terbaik untuk mendapatkan hasil validasi yang akurat. Untuk mendapatkan metode terbaik maka hasil *naive bayes* akan dibandingkan dengan metode *decision tree* menggunakan *confusion matrix.*. Ketika diimplementasi menghasilkan data sebagai berikut :

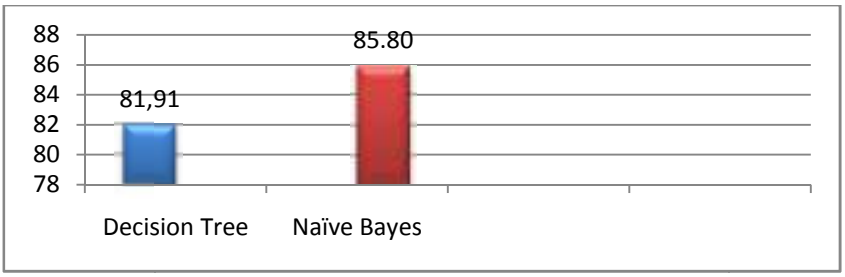
Tabel 5 Akurasi Algoritma Decision Tree

accuracy: 81.91% +/- 3.37% (mikro: 81.91%)			
	Pred Miskin	Pred Sangat Miskin	class precision
Pred Miskin	364	9	97.59%
Pred Sangat Miskin	177	478	72.98%
class recall	67.28%	98.15%	

Tabel 6 Hasil Perbandingan Performance Algoritma

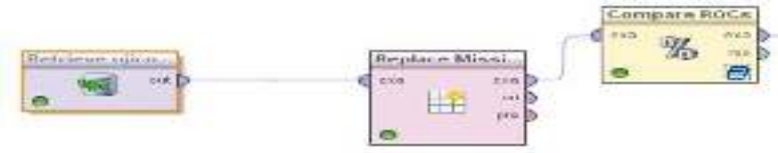
Metode	***Naive Bayes***	***Decision Tree***
Accuracy	**85.80**	81.91
AUC	**0.930**	0.829

Pada tabel 3.6 secara umum *naive bayes* lebih baik berdasarkan keunggulan tingkat akurasi sebagai tingkat kedekatan antara nilai prediksi dengan nilai aktual jika dibandingkan dengan metode *decision tree*. Perbandingan akurasi algoritma *decision tree* dengan *naive bayes* dalam bentuk grafik seperti gambar dibawah ini :



Gambar 7. Grafik Pebandingan Akurasi Decision Tree Dan Naive Bayes

Selain pengujian *performance* dengan menggunakan *confusion matrix* diatas, pengujian juga dilakukan dengan menggunakan *ROC Curve* sehingga dapat diketahui perbedaan tingkat kehandalan klasifikasi masing-masing algoritma.



Gambar 8 Pengujian Performance Menggunakan Campare ROC

Hasil perbandingan tingkat kehandalan klasifikasi seperti tabel dibawah ini :

Tabel 7. Hasil Perbandingan Performance AUC

Tingkat Klasifikasi	***Naive Bayes***	***Decision Tree***
Excellent	0.930	-
Good	-	0.829
Fair	-	-
Poor	-	-
Failure	-	-

Berdasarkan hasil tersebut diatas dan dengan menerapkan klasifikasi *performance* keakurasian AUC diatas maka hasil penelitian ini dapat dibagi menjadi 2 (dua) pengklasifikasian yaitu *excellent classification* yaitu algoritma *Naive Bayes* (0.930) kemudian *good classification* untuk algoritma *decision tree* (0.829).

KESIMPULAN

Dari hasil penelitian yang dilakukan dari tahap awal hingga pengujian, dan dari hasil perbandingan dapat disimpulkan bahwa secara umum *naive bayes* lebih baik jika dibandingkan dengan algoritma *decision tree* berdasarkan keunggulan tingkat akurasi sebagai tingkat kedekatan antara nilai klasifikasi dengan nilai aktual dan berdasarkan keunggulan dalam hal kehandalan dalam klasifikasi karena memiliki nilai AUC yang lebih baik.

DAFTAR PUSTAKA

Aprilla, D., Baskoro, D. A., Ambarwati, L., & Wicaksana, I. W. (2013). Belajar Data Mining dengan RapidMiner. Jakarta: academia.edu.

Bellazzi, R., & Zupanb, B. (2008). *Predictive Data Mining In Clinical Medicine: Current Issues And And Guidelines.* International Journal Of Medical Informatics.

Dunham, M. (2003). *Data Mining Introuctory and Advanced Topics.* New Jersey: Prentice Hall.

Gorunescu, F. (2010). *Data Mining: Concept, Models and Techniques.* Romania: Springer.

Han, J., & Kamber, M. (2007). *Data Mining : Concepts and Techniques* (Second ed.). (M. R. Jim Gray, Ed.) San Francisco, United States of America: Morgan Kaufmann Publishers.

Jhingan, M. (2004). *ekonomi pembangunan dan perencanaan.* jakarta: Raja grafindo persada..

Larose. (2006). *Data Mining Methods And Models.* Canada: John Wiley & Sons, Inc.

Lewis, & J, R. (2000). *An Introduction to Classification And Regression Trees (CART) Analysis.* Presented at the 2000.

Maimon, O. (2010). *Data Mining And Knowledge Discovery Handbook.* London: Springer.

Martin, J. (1990). *Information Enginering Book II Planning and Analysis 2nd Edition.* New Jersey: Prentice-Hall.

Mirawanti, Y. (2012). *Pebandingan Metode Regresi Logistik Ordinal Dengan Jaringan Syaraf Tiruan Fungsi Radial Basis [thesis].* surabaya: Institut Teknologi Sepuluh November Surabaya.

Prakosa, H. M. (2011). *Klasifikasi Kesejahteraan Rumah Tangga Di Provinsi Jawa Timur Dengan Pendekatan Bootstrap Aggregatting Classification And Regression Trees (CART Bagging).* Surabaya: Institute Teknologi Sepuluh September.

Pratama, D. A. (2011). *Klasifikasi Kesejahteraan Rumah Tangga di Jawa Timur dengan Pendekatan Multivariate Adaptive Regression Spline Bootstrap Aggregating (MARS Bagging).* Surabaya: Institute Teknologi Sepuluh September.

Widyandono, I. (2010). *Klasifikasi Kesejahteraan Rumah Tangga Di Provinsi Jawa Timur Dengan Pendekatan Cart Arcing [Thesis].* Surabaya: Institute Teknologi Sepuluh September.

Wijaya, A. (2011). *Analisis Kemiskinan Di Provinsi Lampung Dengan Pendekatan Analisis Kemiskinan Di Provinsi Lampung Dengan Pendekatan Spatial Autoregressive Model (Linear Contiguity Method) [Thesis].* Surabaya: Institute Teknologi Sepuluh September.