

Komparasi Algoritma Data Mining Untuk Akurasi Penentuan Beasiswa Kurang Mampu IAIN Syekh Nurjati Cirebon

Arif Maulana

Fakultas Ilmu Komputer, Dian Nuswantoro University

Email: id.arifmaulana@gmail.com

Abstract: The scholarships reserved for students who are economically disadvantaged to pay the cost of study in college. The number of criteria used for the determination of scholarships made it difficult to make a decision on granting a scholarship. Besides the time it takes too long. The purpose of this study is the classification of students are not able to Decision Tree, k-NN, NN, which uses the algorithm C4.5. Classification results are evaluated and validated using the area under the curve (AUC) of the ROC curve. Given this comparison can facilitate and accelerate decision-making to provide scholarships for students can not afford.

Keywords: classification, scholarship, k-NN, neural network, decision tree, random forest, random tree, data mining.

1 PENDAHULUAN

Pemberian beasiswa terhadap mahasiswa yang tidak mamapu merupakan wujud dari pemerintah akan kepedulian terhadap masyarakat khususnya bagi para siswa/mahasiswa.

Keefektifan pemberian beasiswa terhadap mahasiswa haruslah tepat guna. Komparasi merupakan salah satu metode untuk mengukur dari pendukung keputusan terhadap calon mahasiswa yang ingin mendapatkan beasiswa.

2 TINJAUAN STUDI

Analisis mengenai komparasi klasifikasi data mining sebelumnya sudah dilakukan salah satunya oleh Hastuti, Khafiizh Analisis Komparasi Algoritma Klasifikasi Untuk Prediksi Mahasiswa Non Aktif Semantik 2012. pp. 241-249. ISSN 979 - 26 - 0255 – 0.

3 LANDASAN TEORI

3.1 Data Mining

Data Mining adalah suatu proses yang menggunakan data statistic, matematika, kecerdasan buatan serta machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat untuk orang lain dan dalam ilmu pengetahuan yang terkait dari berbagai database yang ada.(Turban Efraim, & liang, 2005)

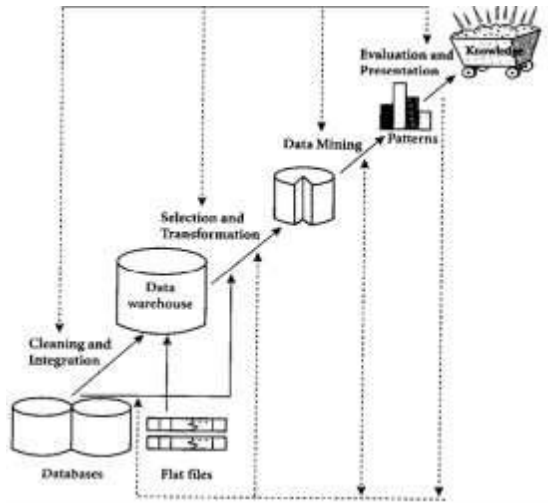
Sehingga Data mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual.(Kusrini & Taufik, 2009)

Menurut (B Neel, 2011) teknik data mining didukung oleh tiga teknologi yaitu pengumpulan data secara masif (besar), multiprocessor pada computer dan algoritma data mining. Tugas dari data mining yaitu Deskriptif – menemukan gambaran pola yang menarik dari data dan Prediktif - memprediksi perilaku dari model berdasarkan data yang ada.

Menurut Larose, data mining dapat dibagi menjadi beberapa kelompok yaitu:

- 1.Deskripsi, 2.Estimasi, 3.Prediksi, 4.Klasifikasi, 5.Klustering, 6. Asosiasi

Tahapan dalam data mining dapat digambarkan pada gambar berikut.



Gambar 1. Tahapan Data Mining

3.2 Klasifikasi

Klasifikasi adalah proses menentukan suatu obyek kedalam suatu kelas atau kategori yang telah ditentukan. Penentuan obyek dapat menggunakan suatu model tertentu beberapa model yang bisa digunakan antara lain: *classification (IF-THEN) rules*, *decision trees*, formula matematika atau *neural networks* (Han, J., Kamber, M., dan Pei, J., 2006).

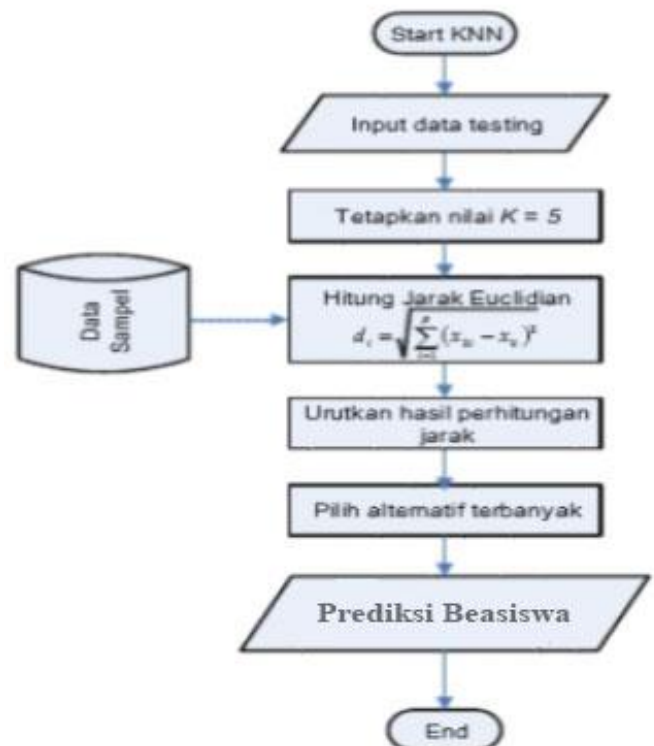
3.3 Algoritma K-nearest neighbor (k-NN)

K-nearest neighbor adalah salah satu contoh instance-based learning, dengan set data pelatihan (training set) disimpan, sehingga klasifikasi untuk record baru yang belum terklasifikasi dapat ditemukan hanya dengan membandingkannya

dengan record paling mirip dalam training set (Larose 2005).

Jika A adalah sekumpulan data pelatihan maka ketika data uji d disajikan, algoritma akan menghitung jarak antara setiap data dalam D dengan data uji d . perhitungan jarak dilakukan menggunakan *Euclidian distance*.

Kemudian, k buah data dalam D yang memiliki jarak terdekat dengan d diambil. Himpunan k merupakan *k-nearest neighbor*.

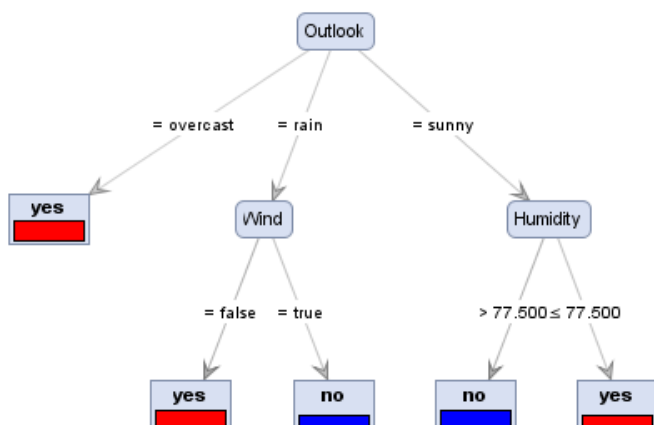


Gambar 2. flowchart k-NN

3.4 Algoritma decision tree (C4.5)

Algoritma C4.5 banyak digunakan peneliti untuk melakukan tugas klasifikasi. *Output* dari algoritma C4.5 adalah sebuah pohon keputusan atau sering dikenal dengan *decission tree*. Dalam beberapa penelitian algoritma C4.5 ini menjadi pilihan terbaik dibandingkan dengan beberapa algoritma klasifikasi lain.

Decision tree sendiri merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal . Dalam *decission tree* ini data yang berupa fakta dirubah menjadi sebuah pohon keputusan yang berisi aturan dan



tentunya dapat lebih mudah dipahami dengan bahasa alami. Model pohon keputusan banyak digunakan pada kasus data dengan *output* yang bernilai diskrit. Walaupun tidak menutup kemungkinan dapat juga digunakan untuk kasus data dengan atribut numeric.

Gambar 3. Pohon keputusan pada dataset golf

3.5 Algoritma random tree (RT)

Dalam matematika dan ilmu komputer, sebuah pohon acak adalah pohon atau arborescence yang terbentuk oleh proses stokastik. Jenis pohon acak meliputi :

1. Uniform spanning tree
2. Random minimal spanning tree
3. Random binary tree
4. Random recursive tree
5. Treap
6. Rapidly exploring random tree
7. Brownian tree
8. Random forest
9. Branching process

3.6 Algoritma neural network (NN)

Menurut (Budi Santoso, 2007), kelebihan utama Neural network adalah kemampuan memprediksi, kecepatan dan *robust* terhadap missing data.

Neuron adalah suatu unit pemroses informasi dalam neural network yang terdiri atas:

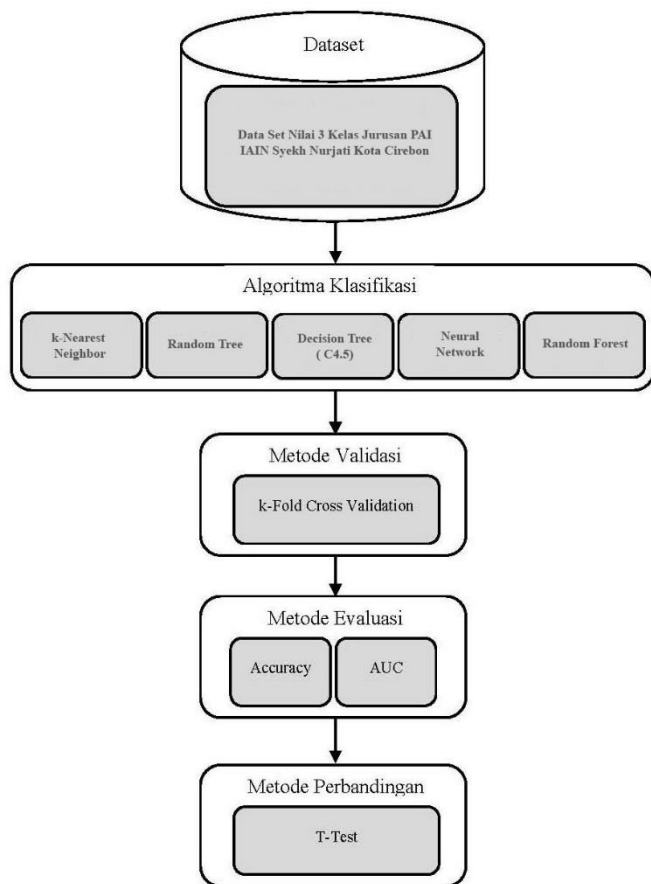
1. Set synapsis atau link penghubung yang ditandai dengan adanya bobot.
2. Penambah, untuk menjumlahkan signal input yang diberi bobot yang disebut kombinasi linier.
3. Fungsi aktivasi, untuk membatasi besarnya output dari suatu neuron.

3.7 Algoritma random forest (RF)

Iterative Dichotomiser 3 (ID3) merupakan sebuah metode yang digunakan untuk membangkitkan pohon keputusan. Algoritma pada metode ini berbasis pada Occam's razor: lebih memilih pohon keputusan yang lebih kecil (teori sederhana) dibanding yang lebih besar. Tetapi tidak dapat selalu menghasilkan pohon keputusan yang paling kecil dan karena itu Occam's razor bersifat heuristik. Occam's razor diformalisasi menggunakan konsep dari entropi informasi.

4 METODE PENELITIAN

Metode penelitian yang digunakan penulis dalam penelitian ini yaitu dengan mengkomparasikan 5 algoritma klasifikasi dengan menggunakan 1 dataset dan untuk teknik evaluasi menggunakan perbandingan AUC, accuracy dari setiap masing-masing algoritma, untuk struktur penelitian tergambar dalam gambar dibawah ini :



Gambar 4.Struktur Metodologi

Dalam mengetahui prediksi beasiswa terhadap mahasiswa yang kurang mampu dalam penelitian ini menggunakan perbandingan dengan 5 metode algoritma data mining yaitu : 1. k-Nearest Neighbor, 2.Decison Tree, 3.Random Tree, 4.Neural Network, 5.Random Forest.

Struktur dalam penelitian ini antara lain yaitu: 1.Dataser, 2.Algoritma klasifikasi, 3.Metode validasi 4.Metode evaluasi, 5.Metode perbandingan menggunakan T-test.

4.1 Data Set

Pada penelitian ini penulis mengambil dataset berupa dataset privat yang diperoleh dari Institut Agama Islam Negri Syekh Nurjati Kota Cirebon, dataset yang diambil berupa daftar nilai,penelusuran ekonomi,serta keaktifan dalam keagamaan.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Nama	Kelas	IPK(1-2)	Penghasilan	Prestasi	Pesantren	Rumah	Listrik	Tanah	Luas Bangun	Air	WC	Jarak	Orang	Kendaraan	KelBeasiswa
IMAM ISMAIL A		3.800	cukup	TIDAK	YA	sangat mem	sangat mem	memadai	memadai	cukup	sangat mem	kurang	memadai	memadai	TIDAK
FUJIRIZQI K A		3.420	memadai	TIDAK	YA	sangat mem	sangat mem	tidak ada	kurang	memadai	sangat mem	kurang	memadai	memadai	YA
RIEKA ROHI A		3.290	sangat mem	TIDAK	YA	sangat mem	sangat mem	cukup	cukup	cukup	memadai	kurang	memadai	memadai	TIDAK
DAWAR RAJI A		3.850	sangat mem	TIDAK	YA	sangat mem	sangat mem	sangat mem	cukup	cukup	sangat mem	kurang	memadai	memadai	TIDAK
ALIEFYA ANI A		3.320	cukup	TIDAK	YA	sangat mem	sangat mem	sangat mem	sangat mem	memadai	sangat mem	kurang	memadai	memadai	TIDAK
IRFA ANNIAI A		3.120	memadai	TIDAK	YA	sangat mem	sangat mem	cukup	kurang	memadai	sangat mem	kurang	sangat mem	memadai	TIDAK
TIAS SAFITRI A		3.200	sangat mem	YA	YA	sangat mem	sangat mem	memadai	memadai	cukup	sangat mem	kurang	memadai	memadai	TIDAK
NAUFAL MU A		3.750	memadai	YA	TIDAK	cukup	sangat mem	memadai	cukup	memadai	sangat mem	cukup	sangat mem	memadai	TIDAK
NURHABIBA A		3.700	memadai	TIDAK	YA	sangat mem	sangat mem	tidak ada	tidak ada	cukup	sangat mem	memadai	sangat mem	memadai	TIDAK
NAFLATURI A		3.550	cukup	TIDAK	YA	sangat mem	sangat mem	memadai	memadai	cukup	sangat mem	kurang	cukup	kurang	YA
MALIKIBRAH A		3.160	memadai	TIDAK	YA	sangat mem	sangat mem	kurang	kurang	cukup	sangat mem	memadai	memadai	memadai	TIDAK
TTI SUGAR A		3.430	memadai	TIDAK	YA	sangat mem	sangat mem	memadai	memadai	cukup	sangat mem	kurang	memadai	memadai	TIDAK
ERMAN A		3.650	sangat mem	TIDAK	YA	sangat mem	sangat mem	tidak ada	tidak ada	memadai	sangat mem	kurang	sangat mem	memadai	TIDAK
DUDUNG AI A		3.580	sangat mem	TIDAK	YA	sangat mem	sangat mem	memadai	memadai	cukup	memadai	kurang	memadai	memadai	TIDAK
FITRIYANI PI A		3.480	cukup	TIDAK	YA	sangat mem	sangat mem	memadai	cukup	cukup	sangat mem	kurang	memadai	memadai	TIDAK
AAR ABDUL A		3.290	sangat mem	YA	YA	sangat mem	sangat mem	memadai	cukup	kurang	sangat mem	kurang	memadai	memadai	TIDAK
DITA PARANI A		3.280	memadai	TIDAK	YA	sangat mem	sangat mem	cukup	cukup	cukup	sangat mem	Lux	memadai	memadai	TIDAK
DEWI SETIA A		3.390	sangat mem	TIDAK	YA	sangat mem	sangat mem	kurang	kurang	cukup	memadai	kurang	memadai	kurang	YA
NUR KOMAR A		3.110	sangat mem	TIDAK	TIDAK	kurang	sangat mem	kurang	kurang	cukup	memadai	kurang	memadai	kurang	YA
JUJU KHAFI A		3.330	sangat mem	TIDAK	YA	sangat mem	sangat mem	memadai	cukup	memadai	sangat mem	kurang	sangat mem	memadai	TIDAK
ANDINA PUTI A		3.100	sangat mem	YA	YA	sangat mem	sangat mem	cukup	cukup	memadai	sangat mem	cukup	sangat mem	memadai	TIDAK
NIDA NADINI A		3.640	cukup	TIDAK	YA	sangat mem	sangat mem	memadai	cukup	cukup	sangat mem	memadai	memadai	memadai	TIDAK
IMAM MAULI A		3.260	cukup	TIDAK	YA	sangat mem	sangat mem	sangat mem	sangat mem	memadai	sangat mem	memadai	sangat mem	memadai	TIDAK
NATASYA SL A		3.950	memadai	TIDAK	YA	sangat mem	sangat mem	cukup	cukup	cukup	sangat mem	kurang	memadai	memadai	TIDAK
MURUL INAY A		3.220	cukup	TIDAK	YA	sangat mem	sangat mem	memadai	memadai	memadai	sangat mem	cukup	memadai	memadai	TIDAK
ROSSI FEBRI A		3.620	memadai	TIDAK	YA	sangat mem	sangat mem	memadai	cukup	cukup	sangat mem	memadai	memadai	memadai	TIDAK
ALFIRRAIDH A		2.800	memadai	TIDAK	YA	sangat mem	sangat mem	cukup	cukup	cukup	sangat mem	cukup	cukup	cukup	TIDAK

Gambar 5.Dataset

4.2 Evaluasi

Pada tahap ini peneliti melihat nilai hasil dari *accuracy* dari *confusion matrix* dan nilai *area under curve* (AUC) dari *ROC curve* yang digunakan sebagai indikator pengukuran tingkat akurasi performansi dari masing – masing algoritma klasifikasi.Nilai *accuracy* dalam penelitian ini diperoleh dari tabel *confusion matrix* RapidMiner. AUC mengukur performansi metode klasifikasi berdasarkan *ROC curve*. Nilai AUC ditunjukkan dalam skala 0 sampai 1 dimana angka 0 menunjukkan tingkat negatif dan angka 1 menunjukkan tingkat positif.

4.3 Perbandingan

Peneliti menggunakan metode perbandingan uji beda parametrik *t-test* untuk membandingkan akurasi algoritma klasifikasi. Nilai akurasi yang diperoleh dibandingkan menggunakan *t-test* untuk memastikan apakah ada perbedaan signifikan pada akurasi algoritma.

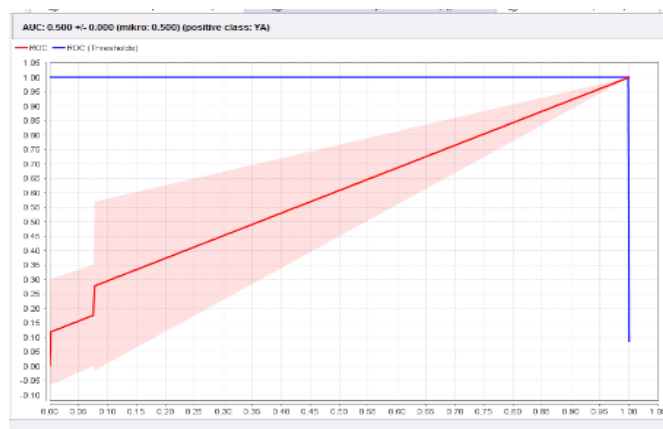
5 HASIL PENELITIAN

Pada eksperimen yang dilakukan penulis menggunakan aplikasi bantu yaitu Rapidminer 6.2 dengan spesifikasi computer Intel Core i3, dengan RAM 6 Gb serta menggunakan OS Windows 8.

5.1 DT

accuracy: 89.04% +/- 3.91% (mikro: 88.96%)			
	true TIDAK	true YA	class precision
pred. TIDAK	132	16	89.19%
pred. YA	1	5	83.33%
class recall	99.25%	23.81%	

Tabel.1 Performa DT



Gambar 6.AUC DT

5.2 K-NN

accuracy: 82.50% +/- 7.15% (mikro: 82.47%)			
	true TIDAK	true YA	class precision
pred. TIDAK	121	15	88.97%
pred. YA	12	6	33.33%
class recall	90.98%	28.57%	

Tabel.2 Performa k-NN



Gambar 7.AUC k-NN

5.3 RT

accuracy: 86.46% +/- 4.21% (mikro: 86.36%)			
	true TIDAK	true YA	class precision
pred. TIDAK	133	21	86.36%
pred. YA	0	0	0.00%
class recall	100.00%	0.00%	

Tabel.3 Performa RT

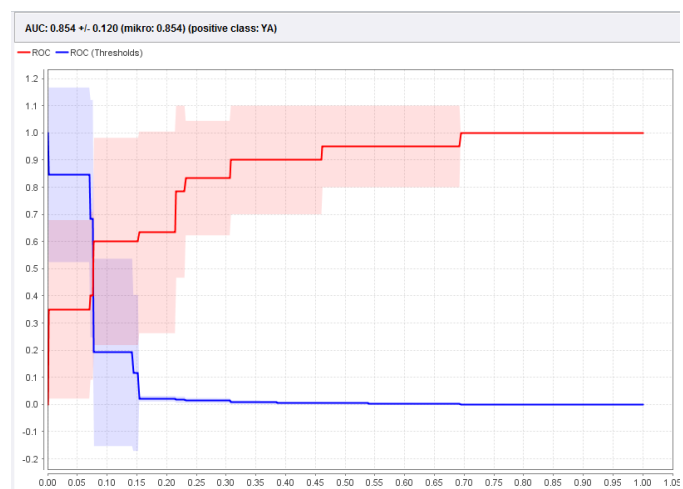


Gambar. 8 AUC RT

5.4 NN

accuracy: 89.67% +/- 4.11% (mikro: 89.61%)			
	true TIDAK	true YA	class precision
pred. TIDAK	131	14	90.34%
pred. YA	2	7	77.78%
class recall	98.50%	33.33%	

Tabel.4 Performa NN

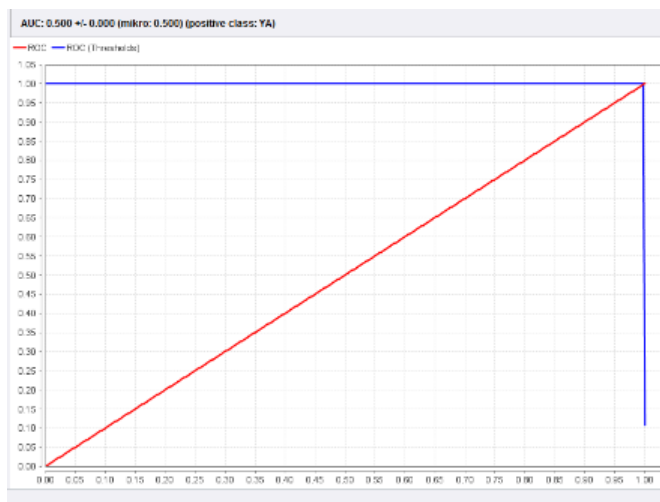


Gambar.9 AUC NN

5.5 ID3

accuracy: 86.46% +/- 4.21% (mikro: 86.36%)			
	True TIDAK	true YA	class precision
pred. TIDAK	133	21	86.36%
pred. YA	0	0	0.00%
class recall	100.00%	0.00%	

Tabel.5 Performa ID3



Gambar.10 AUC ID3

Hasil pengukuran dari *confusion matrix* yang diterapkan pada dataset tersebut tambak bahwa algoritma C4.5 khususnya *descision tree* memiliki tingkat akurasi yang tinggi yaitu 88,96 %, yang diikuti oleh *neural network* 87.02%, *k-NN* 87.01%, serta RF 86.56%

	k-NN	DT	RT	NN	ID3
Accuracy	82.47	88.96	86.36	89.61	86.36
	%	%	%	%	%

Tabel 6. *confution matrix*

Hasil yang didapat dalam klasifikasi 5 algoritma klasifikasi pada data mining dengan menggunakan dataset tersebut adalah sebagai berikut :

Dari gambar T-Test terlihat algoritma C4.5 decision tree menunjukkan yang paling baik

dibandingkan dengan algoritma k-NN, RT, NN, RF dengan tingkat selisih yang tidak terlalu signifikan.

T-Test Significance

	0.890 +/- 0.039	0.825 +/- 0.071	0.865 +/- 0.042	0.897 +/- 0.041	0.865 +/- 0.042
0.890 +/- 0.039		0.023	0.221	0.735	0.221
0.825 +/- 0.071			0.187	0.014	0.187
0.865 +/- 0.042				0.123	1.000
0.897 +/- 0.041					0.123
0.865 +/- 0.042					

Probabilities for random values with the same result

Bold values are smaller than alpha=0.050 which indicates a probably significant difference between the actual mean values!

Gambar.11 Hasil T-test

6 Kesimpulan

Dari hasil komparasi 5 algoritma yang digunakan untuk prediksi pemberian beasiswa mahasiswa kurang mampu pada IAIN Syekh Nurjati, dengan dataset pengabungan dari report nilai,serta penelusuran ekonomi bahwa algoritma C4.5 dengan algoritma decision tree merupakan algoritma yang paling akurat dengan nilai presentase 88.96% dibandingkan dengan algoritma C4.5 lainnya.saran dalam penelitian ini yaitu, memperbanyak atribut yang dihitung agar tingkat akurasi makin membaik.

REFERENSI

Turban Efraim, Aronson Jay E, and Liang, *Decision Support Systems and Intelligent Systems*, 7th ed.: Prentice Hall, Upper Saddle River, NJ, 2005

Kusrini and Emha Taufiq Lutfhfi, *Algoritma Data Mining.*: ANDI Yogyakarta, 2009.

Larose DT. 2005. *Discovering Knowledge in Data An Introduction to Data Mining*. New Jersey (US): J Wiley.

Lindawati, *Data Mining dengan Teknik Clustering Dalam Pengklasifikasian Data Mahasiswa Studi Kasus Prediksi Lama Studi Mahasiswa Universitas Bina Nusantara*, 2008.

Han, J., Kamber, M., dan Pei, J., *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2006.

X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. Mclachlan, A. Ng, B. Liu, P. S. Yu, Z. Z. Michael, S. David, and J. H. Dan, *Top 10 algorithms in data mining*. 2007.

J. Han and M. Kamber, *Data Mining: Concepts and Techniques Second Edition*. Elsevier, 2006.

D. Widiastuti, “Analisa Perbandingan Algoritma SVM, Naïve Bayes, dan Decission Tree dalam Mengklasifikasikan Serangan (Attack) pada Sistem Pendeteksi Intrusi,” *Jur. Sist. Inf. Univ. Gunadarma*, 2007.

Kusrini and L. E. Taufiq, *Algoritma Data Mining*. Yogyakarta: Andi Offset, 2009.

B. Santosa, *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*, Edisi Pert. Yogyakarta: Graha Ilmu, 2007.