KOMPARASI ALGORITMA KLASIFIKASI DATA MINING MODEL C4.5 DAN NAIVE BAYES UNTUK PREDIKSI PENYAKIT DIABETES

RINGKASAN TESIS

FATMAWATI

14000945

PROGRAM PASCASARJANA MAGISTER ILMU KOMPUTER

SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER

NUSA MANDIRI

JAKARTA

2015

KOMPARASI ALGORITMA KLASIFIKASI DATA MINING MODEL C4.5 DAN NAIVE BAYES UNTUK PREDIKSI PENYAKIT DIABETES

TESIS

FATMAWATI

14000945

PROGRAM PASCASARJANA MAGISTER ILMU KOMPUTER

SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER

NUSA MANDIRI

JAKARTA

2015

HALAMAN PENGESAHAN

Tesis ini diajukan oleh :

Nama	:	Fatmawati
NIM	:	14000945
Program Studi	:	Magsiter Ilmu Komputer
Jenjang	:	Strata Dua (S2)
Konsentrasi	:	*Managemen Information System*
Judul Tesis	:	"Komparasi Algoritma Klasifikasi Data Mining Model C4.5 dan *Naive bayes* Untuk Prediksi Penyakit Diabetes"

Telah berhasil dipertahankan dihadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Magister Ilmu Komputer (M.Kom) pada Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri (STMIK Nusa Mandiri).

Jakarta, 05 Desember 2015
Pascasarjana Magister Ilmu Komputer
STMIK Nusa Mandiri
Direktur

Prof. Dr. Ir. R. Eko Indrajit, MSC, MBA

D E W A N P E N G U J I

Penguji I	:	Dr. Sularso Budilaksono, M.Kom	.....
Penguji II	:	Dr. Sfenrianto, M.Kom	.....
Pembimbing	:	Dr. Windu Gata, M.Kom	.....

DAFTAR ISI

Halaman

COVER1 i
COVER2 ii
HALAMAN PENGESAHAN..... **Error! Bookmark not defined.**
DAFTAR ISI iv
ABSTRAK 1
I. PENDAHULUAN 1
II. KAJIAN LITERATUR 2
1.1. Diabetes 2
1.2. Data Mining 4
1.3. Klasifikasi 5
1.4. Algoritma C4.5 5
1.5. Naive Bayes 7
1.6. Rapid Miner 7
1.7. Evaluasi dan Validasi 8
III. METODE PENELITIAN 10
IV. PEMBAHASAN 12
V. PENUTUP 17
1. KESIMPULAN 17
2. SARAN 17

ABSTRAK

Penyakit diabetes merupakah salah satu penyakit yang mematikan, faktor resiko tinggi dalam keluarga yang menyebabkan diabetes dikarenakan orang gemuk yang tidak melakukan latihan fisik, dan orang-orang yang tidak memiliki gaya hidup sehat dan makanan yang berlebihan dari apa yang dibutuhkan oleh tubuh. Berdasarkan data *history* penderita diabetes dapat dibuat rekomendasi prediksi penyakit diabetes yang dapat membantu tenaga kesehatan. Klasifikasi merupakan salah satu teknik dari *data mining* yang dapat digunakan untuk membantu prediksi. Klasifikasi dapat dilakukan dengan *Decision Tree* yaitu dengan algoritma C4.5 dan *Naive Bayes*. Penelitian ini bertujuan membuat klasifikasi dan menerapkan klasifikasi *data mining*. Hasil klasifikasi data di evaluasi dengan menggunakan *Confusion Matrix* dan kurva *ROC* untuk mengetahui tingkat hasil akurasi menggunakan algoritma *Decision Tree* yaitu sebesar 73.30% dan nilai AUC dari kurva *ROC* adalah 0.733 sedangkan algoritma *Naive Bayes* sebesar 75.13% nilai AUC dari kurva *ROC* 0.810 sehingga dapat dikatakan bahwa algoritma *Naive Bayes* memiliki hasil prediksi yang baik dalam memprediksi penyakit diabetes seorang pasien.

Kata kunci: prediksi penyakit diabetes, algoritma C4.5, model *Decision Tree, Naive Bayes*

I. PENDAHULUAN

Diabetes merupakan penyakit gangguan metabolik menahun akibat pankreas tidak memproduksi cukup insulin atau tubuh tidak dapat menggunakan insulin yang diproduksi secara efektif. Insulin adalah hormon yang mengatur keseimbangan kadar gula darah. Akibatnya terjadi peningkatan konsentrasi glukosa didalam darah (hiperglikemia).

Penyakit diabetes disebabkan oleh peningkatan kadar glukosa dalam darah, apabila kadar glukosa darah meningkat dalam jangka waktu yang lama maka akan menyebabkan komplikasi seperti gagal ginjal, kebutaan dan serangan jantung (Jayalskshmi & Santhakumaran, 2010).

Estimasi terakhir IDF (*International Diabetes Federation*), terdapat 382 juta orang yang hidup dengan diabetes di dunia pada tahun 2013. Dari berbagai penelitian epidemiologis di indonesia yang dilakukan oleh pusat-pusat diabetes, seekitar tahun 1980-an prevalensi diabetes melitus pada penduduk usia 15 tahun ke atas sebesar 1,5-

2,3% dengan prevalensi di daerah rural/perdesaan lebih rendah dibandingkan perkotaan.

Teknik analisa konvensional secara manual yang selama ini digunakan tidak lagi efektif digunakan untuk mendiagnosa. Seiring dengan perkembangan sistem berbasis pengetahuan medis tuntutan akan adanya penggunaan sistem pengetahuan berbasis komputer sebagai teknik analisa dalam mendiagnosa penyakit menjadi seemakin penting. Oleh karenanya, saat inilah waktu yang tepat untuk mengembangkan sistem pengetahuan berbasis komputer yang modern, efektif dan efisien dalam mendiagnosa penyakit (Neshat, Mehdi & Yaghobi, 2009).

Oleh karena itu, penelitian ini dilakukan untuk membantu menyelesaikan permasalahan tersebut dengan *data mining* yang berfungsi untuk memprediksi penyakit diabetes, diperlukan suatu metode atau teknik yang dapat mengolah data-data yang sudah ada. Salah satu metodenya menggunakan teknik *data mining*.Penggunaan *data mining* dengan algoritma C4.5 dan *Naive Bayes* sebagai pilihan untuk diagnosa penyakit diabetes dapat menjadi alternatif pilihan yang tepat, tetapi sampai saat ini belum diketahui algoritma yang paling akurat dalam memprediksi penyakit diabetes.

Pada penelitian ini akan dilakukan komparasi *data mining* algoritma C4.5 dan *Naive Bayes* untuk mengetahui algoritma yang memiliki akurasiyang lebih tinggi dalam mendeteksi penyakit diabetes.

II. KAJIAN LITERATUR

1.1. Diabetes

Diabetes adalah epidemi yang paling cepat berkembang di Barat dunia. Satu dari tiga anak-anak Amerika akan tumbuh dan terkena diabetes, 24% orang dewasa Amerika resisten insulin dan 45% orang dewasa di atas usia 60 resisten insulin (Mason, 2005). Diabetes adalah salah satu penyebab utama kematian di banyak negara dan

penyebab utama kebutaan, gagal ginjal, dan *nontraumatic* amputasi (Robert, Zgonis, & Driver, 2006)**.**

Faktor penyebab diabetes adalah (Nurrahmani, 2012):

1. Gen diabetes dalam keluarga

   Gen merupakan sel pembawa sifat yang dapat diwariskan orang tua kepada keturunannya, dan diabetes merupakan penyakit yang bisa diwariskan.

2. Insulin dan gula darah

   Insulin adalah karena ketidakmampuan beta sel-sel di pankreas untuk memproduksi insulin (Mason, 2005). Produksi ini disebabkan oleh tingginya kadar gula dalam darah, sehingga menyebakan insulin diproduksi semakin tinggi.

3. Kegemukan (Obesitas)

   Pada kegemukan atau obesitas sel-sel lemak yang menggemuk yang jumlahnya lebih banyak dari pada keadaan tidak gemuk, sehingga menyebabkan reistensi terhadap insulin dimana gula darah sulit masuk kedalam sel, sehingga gula darah tetap tinggi (hiperglikemi) sehingga terjadilah diabetes, khususnya terjadi pada diabetes tipe2.

4. Asma

   Penderita yang mengalami penyakit asma diharuskan untuk mengkonsumsi obat asma, sehingga memicu terjadinya diabetes, dikarenakan hormon yang digunakan pada obat asma tersebut adalah steroid yang berkerja berlawanan dengan insulin yang menaikkan gula darah.

5. KB

   Pil kontrasepsi merupakan salah satu obat yang mengandung hormon streoid dengan anti insulin rendah.

1.2. Data Mining

Data mining merupakan teknologi baru yang sangat berguna untuk membantu perusahaan-perusahaan menemukan informasi yang sangat penting dari gudang data mereka. Beberapa aplikasi data mining fokus pada prediksi, mereka meramalkan apa yang akan terjadi dalam situasi baru dari data yang menggambarkan apa yang terjadi di masa lalu (Witten, Frank, & Hall, 2011).

Data mining juga merupakan bagian dari *Knowledge Discovery in Database* (KDD) yang merupakan proses ekstraksi informasi yang berguna, tidak diketahui sebelumnya dan tersembunyi dari data (Bramer, 2007).

Secara garis besar *Knowledge Discovery in Database* (KDD) dapat dijelaskan sebagai berikut (Kusrini & Luthfi, 2009):

1. *Data Selection*

   Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD di mulai. Data hasil selksi yang akan digunakan untuk proses rpisah *data mining*, disimpan dalam suatu berkas terpisah dari basis data operasional.

2. *Pre-Processing/Cleaning*

   Proses *cleaning* antara lain membuang duplikasi data, memeriksa data yang inkonsisten dan memperbaiki kesalahan pada data. Pada proses ini dilakukan juga proses *enrichment*, yaitu proses memperkaya data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD.

3. *Transformation*

   Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*.

4. *Data Mining*

   Data Mining adalah proses mencari pola atau informasi menarikdalam data terpilih dengan menggunakan teknik atau metode tertentu.

5. *Interpretation/Evaluation*

6. Pola informasi yang dihasilkan dari proses *data mining* diterjemahkan menjadi bentuk yang lebih mudah dimengerti oleh pihak yang berkepentingan.

1.3. Klasifikasi

Klasifikasi merupakan bagian dari prediksi, dimana nilai yang diprediksi berupa label. Klasifikasi menentukan *class* atau grup untuk tiap contoh data, *input* dari model klasifikasi adalah atribut dari contoh data (data *samples*) dan *output*nya adalah *class* dari data *samples* itu sendiri, dalam *machine learning* untuk membangun model klasifikasi digunakan metode *supervised learning* (HuiHuang, 2006). Metode *supervised learning* yaitu metode yang mencoba untuk menemukan hubungan antara atribut masukan dan atribut target, hubungan yang ditemukan diwakili dalam struktur yang disebut model.

Dalam klasifikasi kita dapat menentukan orang atau objek kedalam suatu kategori tertentu, contoh untuk masalah klasifikasi adalah menentukan apakah seseorang pasien "mengidap" atau "tidak mengidap" penyakit tertentu. Informasi tentang pasien sebelumnya digunakan sebagai bahan untuk melatih algoritma untuk mendapatkan *rule* atau aturan.

Salah satu tujuan klasifikasi adalah untuk meningkatkan kehandalan hasil yang diperoleh dari data (Kahramanli & Allahverdi, 2008).

1.4. Algoritma C4.5

Algoritma C4.5 diperkenalkan oleh J. Ross Quinlan yang merupakan perkembangan dari algoritma ID3, algoritma tersebut digunkan untuk pohon keputusan. Pohon keputusan dianggap sebagai salah satu pendekatan yang paling populer, dalam klasifikasi pohon keputusan terdiri dari sebuah *node* yang membentuk akar, *node* akar tidak memiliki inputan. *Node* lain yang bukan sebagai akar tetapi memiliki tepat satu inputan disebut *node internal* atau *test node*, sedangkan *node* lainnya dinamakan daun. Daun mewakili nilai target yang paling tepat dari salah satu *class* (Maimon & Rokack, 2010).

Langkah-langkah membangun pohon keputusan menggunakan algoritma C4.5 adalah sebagai berikut (Kusrini & Luthfi, 2009):

1. Pilih atribut sebagai akar.

   Pemilihan atribut sebagai akar berdasarkan pada nilai *gain* tertinggi dari atribut-atribut yang ada. Untuk menghitung nilai *gain* tertinggi digunkan persamaan berikut:

$$Gain\ (S, A) = Entropy(S) - \sum_{i=0}^{n} \frac{|S_i|}{|S|} * Entropy(Si)$$

Keterangan:

S : himpunan kasus

A : atribut

n : jumlah partisi atribut A

|Si| : jumlah kasus pada partisi ke-i

|S| : jumlah kasus dalam S

Nilai entropi dapat dihitung dengan cara berikut:

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

Dimana:

S : himpunan kasus

n : jumlah partisi S

Pi : proporsi dari S_i tehadap S

2. Buat cabang untuk tiap-tiap nilai.
3. Bagi kasus dalam cabang.

   Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

1.5. Naive Bayes

Naive Bayes merupakan metode yang tidak memiliki aturan, *naive bayes* menggunakan cabang matematika yang dikenal dengan teori probabilitas untuk mencari peluang terbesar dari kemungkinan klasifikasi, dengan cara melihat frekuensi tiap klasifikasi pada data training. *Naive Bayes* merupakan metode klasifikasi populer dan masuk dalam sepuluh algoritma terbaik dalam data mining, algoritma ini juga dikenal dengan nama *Idiot's Bayes, Simple Bayes* dan *Independence Bayes* (Bramer, 2007).Klasifikasi *Bayes* di dasarkan pada teorema bayes, diambil dari nama seoranga ahli matematika yang juga menteri Prebysterian Inggris, Thomas Bayes(1702-1761). Yaitu:

$$P(x|y) = \frac{P(y|x)\,P(x)}{P(y)}$$

Keterangan:

Y : data dengan kelas yang belum diketahui

X : hipotesis data y merupakan suatu kelas spesifik

P(x|y) : probabilitas hipotesis x berdasarkan kondisi y (posteriori probability)

P(x) : probabilitas hipotesis x (prior probability)

P(y|x) : probabilitas y berdasarkan kondisi pada hipotesis x

p(y) : probabilitas dari y

1.6. Rapid Miner

Rapid Miner merupakan perangkat lunak yang dibuat oleh Dr. Markus Hofmann dari Institute of Technology Blanchardstown dan Raif Klinkenberg dari rapid-i.com dengan tampilan GUI (*Graphical User Interface*) sehingga memudahkan pengguna dalam menggunakan perangkat lunak ini. Perangkat lunak ini bersifat *open source* dan dibuat dengan menggunakan bahasa java dibawah lisensi GNU *Public License* dan Rapid Miner dapat dijalankan disistem operasi manapun. Dengan menggunakan Rapid

Miner, tidak dibutuhkan kemampuan koding khusus, karena semua fasilitas sudah disediakan. Rapid Miner dikhususkan untuk penggunaan data mining.

1.7. Evaluasi dan Validasi

Validasi adalah proses mengevaluasi akurasi prediksi dari sebuah model, validasi mengacu untuk mendapatkan prediksi dengan menggunakan model yang ada kemudian membandingkan hasil yang diperoleh dengan hasil yang diketahui (Gorunescu, 2011).

Untuk mengevaluasi model digunakan metode *confusion matrix*, dan kurva ROC (*Receiver Operating Characteristic*).

1. Confusion Matrix

Confusion matrix memberikan rincian klasifikasi, kelas yang diprediksi akan ditampilkan di bagian atas matrix dan kelas yang diobservasi ditampilkan di bagian kiri (Gorunescu, 2011). Evaluasi model *confussion matrix* menggunakan tabel seperti matrix dibawah ini:

Tabel 1. Matrik Klasifikasi untuk Model 2 Class

Classification	Predicted Class		
		Class=Yes	Class=No
Observed Class	Class=Yes	(True Positive – TP)	(False Negative – FN)
	Class=No	(False Positive – FP)	(True Negative – TN)

Sumber: Gorunescu(2011)

Akurasi dapat dihitung dengan menggunakan rumus berikut:

$$Accurasy = \frac{TP+TN}{TP+TN+FP+FN}$$

TP : Jumlah kasus positif yang diklasifikasikan sebagai positif

FP : Jumlah kasus negatif yang diklasifikasikan sebagai positif

TN : Jumlah kasus negatif yang diklasifikasikan sebagai negatif

FN : Jumlah kasus positif yang diklasidikasikan sebagai negatif

1. Kurva ROC

Kurva ROC banyak digunakan untuk menilai hasil prediksi, kurva *ROC* adalah teknik untuk memvisualisasikan, mengatur, dan memilih pengklasifikasian berdasarkan kinerja mereka (Gorunescu, 2011).

Kurva *ROC* adalah *tool* dua dimensi yang digunakan untuk menilai kinerja klasifikasi yang menggunakan dua *class* keputusan, masing-masing objek dipetakan ke salah satu elemen dari himpunan pasangan, positif atau negatif. Pada kurva ROC, TP *rate* diplot pada sumbu Y dan FP *rate* diplot pada sumbu X.

Untuk klasifikasi *data mining*, nilai AUC dapat dibagi menjadi beberapa kelompok (Gorunescu, 2011).

a. 0.90-1.00 = *Excellent Classification*
b. 0.80-0.90 = *Good Classification*
c. 0.70-0.80 = *Fair Classification*
d. 0.60-0.70 = *Poor Classification*
e. 0.50-0.60=*Failur*

The Area Under Curve (AUC) dihitung untuk mengukur perbedaan performasi metode yang digunakan. AUC dihitung menggunakan rumus (Liao & Triantaphyllou, 2007):

$$\theta^r = \frac{1}{mn}\sum_j^n = 1 \sum_i^m = 1 \psi(xi^r, xj^r$$

Dimana

$$\psi(X,Y) = \begin{cases} 1 & Y < X \\ \frac{1}{2} & Y = X \\ 1 & Y > X \end{cases}$$

X= Output Positif

Y = Output Negatif

III. METODE PENELITIAN

Dalam penelitian ini dilakukan beberapa langkah yang dilakukan dalam proses penelitian.

1. Pengumpulan data

Pada tahap ini ditentukan data yang akan diproses. Mencari data yang tersedia, memperoleh data tambahan yang dibutuhkan, mengintegrasikan semua data kedalam data set, termasuk variabel yang diperlukan dalam proses.

Data yang diperoleh adalah data sekunder karena diperoleh dari Pima Indian diabetes database dalam UCI (singkatan dari Pima Diabetes). Masalah yang harus dipecahkan di sini adalah prediksi terjadinya diabetes melitus dalam waktu 5 tahun dengan menggunakan Pima yang berisi 786 orang yang diperiksa dan sebanyak 500 pasien tidak terdeteksi terkena penyakit diabetes, sehingga 268 pasien terdeteksi penyakit diabetes. Dengan atribut dari penyakit diabetes adalah berapa kali hamil, konsentrasi glukosa, tekanan darah, ketebalan lipatan kulit, serum insulin, indeks massa tubuh, diabetes silsilah fungsi dan umur dan kelas sebagai label yang terdiri atas ya dan tidak

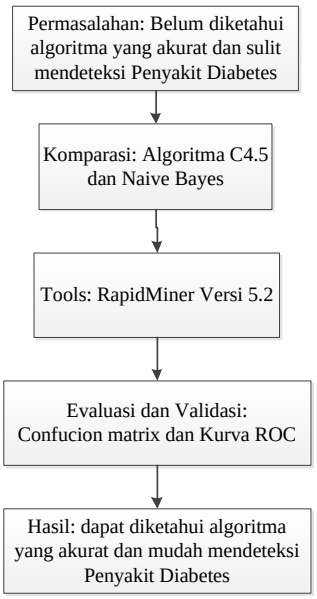
2. Pengolahan data awal

Ditahap ini dilakukan penyeleksian data, data dibersihkan dan ditransformasikan kebentuk yang diinginkan sehingga dapat dilakukan persiapan dalam pembuatan model. Jumlah data awal yang diperoleh dari pengumpulan data yaitu sebanyak 768 data, namun tidak semua data dapat digunakan dan tidak semua atribut digunakan karena harus melalui beberapa tahap pengolahan awal data (*preparation data*)**.**

3. Metode yang diusulkan

Pada tahap ini data dianalisis, dikelompokan variabel mana yang berhubungan dengan satu sama lainnya. Setelah data dianalisis lalu diterapkan model-model yang sesuai dengan jenis data. Pembagian data kedalam data latihan (*training data*) dan data uji (*testing data*) juga diperlukan untuk pembuatan model.

Pada tahap modeling ini dilakukan pemprosesan data traning sehingga akan membahas metode algoritma yang diuji dengan memasukan data penyakit diabetes kemudian di analisa dan dikomparasi.



Gambar 1. Metode yang diusulkan

Metode yang diusulkan dari penelitian ini, dimulai dari problem (permasalahan) analisa penyakit diabetes kemudian dibuat *approach* (model) dalam bentuk algoritma C4.5 dan *Naive Bayes* untuk memecahkan permasalahan, sedangkan *software* yang digunakan dalam penelitian ini adalah *Rapidminer* 5.2, *Microsoft Excell* 2013, pengujian evaluasi dan validasi untuk mengukur akurasi menggunakan *confusion matrix* dan *kurva ROC*, serta hasil dari penelitian didapat di antara ke dua algoritma tersebut, didapat algoritma yang terbaik dalam prediksi penyakit diabetes.

4. Eksperimen dan pengujian metode

Pada tahap ini model yang diusulkan akan diuji untuk melihat hasil berupa *rule* yang akan dimanfaatkan dalam pengambilan keputusan. Tahap modeling untuk menyelesaikan prediksi penyakit diabetes dengan menggunakan dua metode yaitu algoritma algoritma C4.5 dan *Naive Bayes.* Penelitian ini termasuk penelitian eksperimen, dimana penelitian ini dimulai dengan menentukan model yang

digunakan, memasukan data training dan testing kedalam model dan mengujinya dengan tools rapidminer.

5. Evaluasi dan validasi

Pada tahap ini dilakukan evaluasi terhadap model yang ditetapkan untuk mengetahui tingkat keakurasian model. Pada penelitian ini digunakan penerapan algoritma C4.5 dengan menentukan nilai *weight* terlebih dahulu. Setelah didapatkan nilai akurasi dan AUC terbesar, nilai *weight* tersebut akan dijadikan nilai yang akan digunakan untuk mencari nilai akurasi dan AUC tertinggi. Sedangkan penerapan algoritma *Naive Bayes* beracuan pada nilai *weight* pada algoritma tersebut. Setelah ditemukan nilai akurasi yang paling ideal dari parameter tersebut langkah selanjutnya adalah menentukan nilai *weight*. Setelah ditemukan nilai akurasi yang paling ideal dari parameter tersebut langkah selanjutnya adalah menentukan *weight* sehingga terbentuk struktur algoritma yang ideal untuk pemecahan masalah tersebut.

IV. PEMBAHASAN

1. Tahap Pengumpulan Data

Data yang digunakan dalam penelitian ini bersumber dari alamat web: http://archive.ics.uci.edu/ml/. Data merupakan hasil pemeriksaan terhadap 768 orang, 500 orang tidak terdeteksi penyakit diabetes dan 268 orang terdeteksi menderita penyakit diabetes. Pada data diabetes ini terdiri dari 9 atribut, 8 atribut predictor dan 1 atribut tujuan. Seperti terlihat pada Tabel 2:

Tabel 2. Data Pasien Diabetes

No.	Jumlah Hamil	Konsentrasi Glukosa	Tekanan Darah	Lipatan Kulit	Serum Insulin	IMB	Riwayat Diabetes	Umur	Hasil
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1

4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0
7	3	78	50	32	88	31.0	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
10	8	125	96	0	0	0.0	0.232	54	1
11	4	110	92	0	0	37.6	0.191	30	0
12	10	168	74	0	0	38.0	0.537	34	1
13	10	139	80	0	0	27.1	1.441	57	0
14	1	189	60	23	846	30.1	0.398	59	1
15	5	166	72	19	175	25.8	0.587	51	1
16	7	100	0	0	0	30.0	0.484	32	1
17	0	118	84	47	230	45.8	0.551	31	1
18	7	107	74	0	0	29.6	0.254	31	1
19	1	103	30	38	83	43.3	0.183	33	0
20	1	115	70	30	96	34.6	0.529	32	1
21	3	126	88	41	235	39.3	0.704	27	0
22	8	99	84	0	0	35.4	0.388	50	0
23	7	196	90	0	0	39.8	0.451	41	1
24	9	119	80	35	0	29.0	0.263	29	1
25	11	143	94	33	146	36.6	0.254	51	1

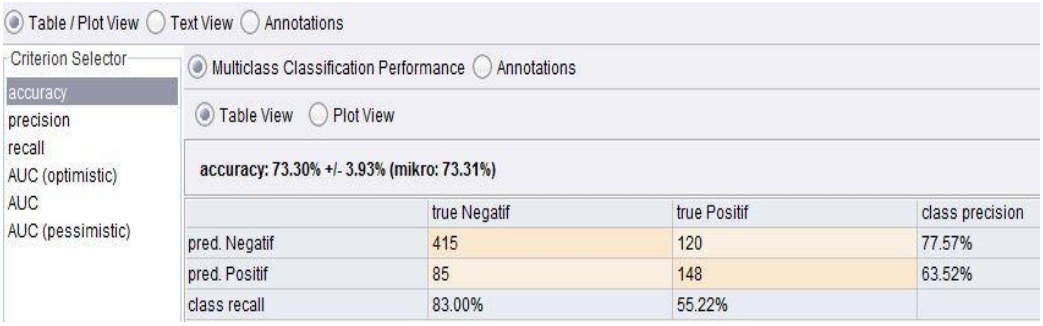
Sumber: http://archive.ics.uci.edu/ml/.

Dari tabel 2 diatas merupakan sample dari data penyakit diabetes dan data yang didapat tidak disertai keterangan yang menjelaskan maksud secara rinci mengenai maksud data, sehingga peneliti harus menganalisa dengan langkah awal melakukan pencarian

informasi mengenai diabetes. Setelah melakukan pencarian tersebut, maka didapat beberapa informasi dan keterangan yang dapat membuat peneliti lebih memahami mengenai data pasien diabetes.

2. Tahap Pengolahan Data

Dari data ekperimen akan diujikan dengan menggunakan metode 10-*fold cross-validation*, dimana data secara acak (*random*)akan dibagi menjadi 10 bagian. Pembagian menjadi 10 bagian merupakan metode yang paling tepat untuk mendapatkan estimasi terbaik menentukan kesalahan. Setiap bagian akan dihitung tingkat kesalahan setelah itu secara keseluruhan akan dihitung rata-ratanya. Setelah dilakukan klasifikasi model data, maka tahap selanjutnya melakukan pengujian akurasi data uji, metode yang digunakan untuk menganalisa model klasifikasi yaitu:



	true Negatif	true Positif	class precision
pred. Negatif	415	120	77.57%
pred. Positif	85	148	63.52%
class recall	83.00%	55.22%	

Gambar 2. Hasil Akurasi Prediksi Penyakit Diabetes Menggunakan Algoritma C4.5

Berdasarkan gambar 4 menunjukan bahwa, diketahui dari 768 data pasien penyakit diabetes, ada 500 orang tidak terdeteksi diabetes tetapi pada hasil tabel *confusion matrix* diatas ada 415 pasien diprediksi negatif maka hasilnya sesuai dengan prediksi yaitu negatif, 120 pasien diprediksi negatif tetapi hasilnya adalah positif, sedangkan 268 orang diprediksi positif tetapi pada gambar diatas menunjukan bahwa ada 85 pasien yang diprediksi positif tetapi hasilnya adalah negatif dan 148 diprediksi positif maka hasilnya sesuai dengan prediksi yaitu positif dan tingkat akurasi dengan menggunakan algoritma C4.5 adalah 73.30%, dan dapat dihitung untuk mencari nilai *accuracy*, yaitu:

Keterangan:

TP = 148

FP = 120

TN= 415

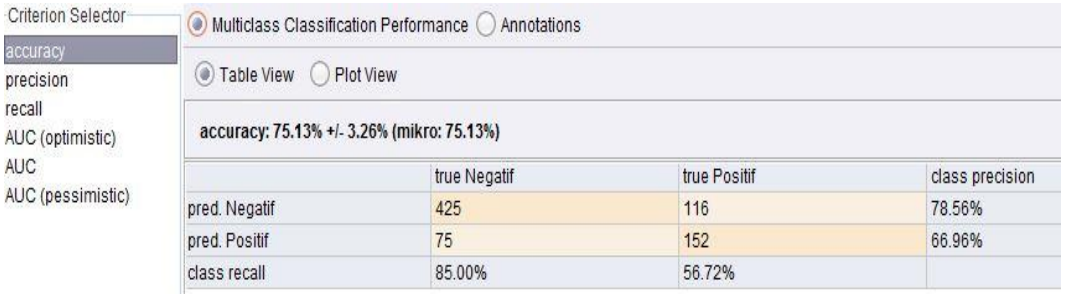
FN = 85

Akurasi = (TP+TN) / (TP+TN+FP+FN)

= (148+415) / (148+415+120+85)

= 0.7330 (73.30%)

Sedangkan untuk algoritma Naive Bayes akan menghasilkan nilai seperti di bawah ini:



	true Negatif	true Positif	class precision
pred. Negatif	425	116	78.56%
pred. Positif	75	152	66.96%
class recall	85.00%	56.72%	

Gambar 3. Hasil Akurasi Prediksi Penyakit Diabetes Menggunakan Algoritma Naive Bayes

Berdasarkan gambar 5 menunjukan bahwa, diketahui dari 768 data pasien penyakit diabetes, ada 500 orang tidak terdeteksi diabetes tetapi pada hasil tabel *confusion matrix* diatas ada 425 pasien diprediksi negatif maka hasilnya sesuai dengan prediksi yaitu negatif, 116 pasien diprediksi negatif tetapi hasilnya adalah positif, sedangkan 268 orang diprediksi positif tetapi pada gambar diatas menunjukan bahwa ada 75 pasien yang diprediksi positif tetapi hasilnya adalah negatif dan 152 diprediksi positif maka hasilnya sesuai dengan prediksi yaitu positif tingkat akurasi dengan menggunakan algoritma *Naive Bayes* adalah 75,13%, dan dapat dihitung untuk mencari nilai *accuracy*, yaitu:

Keterangan:

TP = 152

FP =116

TN= 425

FN = 75

Akurasi = (TP+TN) / (TP+TN+FP+FN)

= (152+425) / (152+425+116+75)

= 0.7513 (75.13%)

Hasil pengujian *confucion matrix* diatas diketahui bahwa model algoritma C4.5 mempunyai akurasi 73.30% sedangkan model Naive Bayes memiliki akurasi 75.13%, tingkat akurasi Naive Bayes lebih tinggi dibandingkan dengan algoritma C4.5 sebesar 1.83%.

3. Analisa Hasil Komparasi

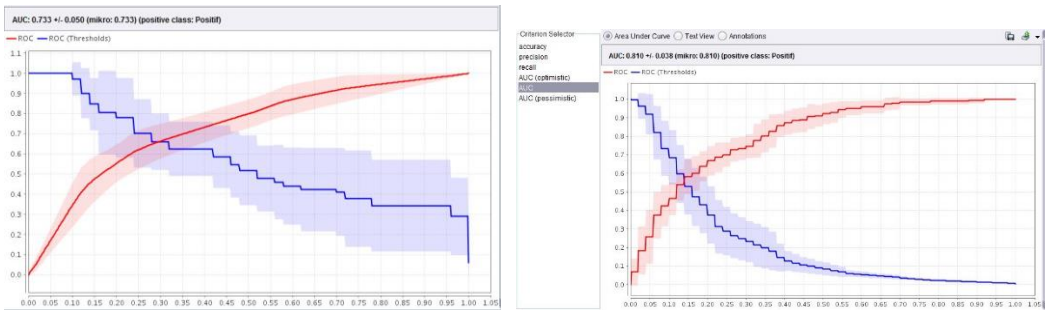
Dari hasil pengujian diatas baik evaluasi menggunakan *confusion matrix* maupun kurva *ROC* untuk model klasifikasi algoritma C4.5 dan *Naive Bayes* sebagai berikut:

Tabel 3. Hasil Komparasi Algoritma C4.5 dan Naive Bayes

	Accuracy	AUC
Decision Tree	73.30%	0.733
Naive Bayes	75.13%	0.810

4. Pengujian dan Evaluasi

Untuk evaluasi menggunakan kurva *ROC* sehingga menghasilkan nilai AUC (*Area Under Curve*) untuk model algoritma *Decision Tree* menghasilkan nilai 0.733 dengan nilai diagnosa Fair *Classification* sedangkan untuk algoritma *Naive Bayes* menghasilkan nilai 0.810 dengan nilai diagnosa *Good Classification* dan selisih nilai keduanya sebesar 1.83%. dapat dilihat pada gambar dibawah ini:



Gambar 4. Kurva ROC dengan Algoritma C4.5 dan Naive Bayes

Dengan demikian algoritma *Naive Bayes* dapat memberikan solusi untuk permasalahan dalam prediksi penyakit diabetes.

V. PENUTUP

1. KESIMPULAN

Berdasarkan hasil pengujian dan analisis bahwa pengujian ini bertujuan untuk mengetahui diantara model algoritma C4.5 dan *Naive Bayes* yang memiliki akurasi paling tinggi untuk memprediksi penyakit diabetes. Hasil perbandingan antara C4.5 dan *Naive Bayes* diukur tingkat akurasinya menggunakan pengujian *Confusion Matrix* dan Kurva ROC. Berdasarkan hasil pengukuraan tingkat akurasi kedua algoritma tersebut, diketahui bahwa nilai akurasi C4.5 adalah 73.30% dan nilai AUC adalah 0.733, sedangkan nilai akurasi *Naive Bayes* 75.13% dan nilai AUC adalah 0.810 dapat disimpulkan bahwa dengan menggunakan model *Naive Bayes* lebih tinggi tingkat akurasinya, dengan peningkatan akurasi sebesar 1.83% dan peningkatan nilai AUC sebesar 0.077 sedangkan hasil pengujian dari prediksi diabetes hasilnya termasuk *Good Clasification.*

2. SARAN

Beberapa saran yang dapat diberikan untuk pengembangan dan perbaikan diwaktu yang akan datang adalah:

1. Dapat melakukan komparasi lebih dari tiga metode algoritma dan dengan data yang lebih banyak lagi untuk prediksi penyakit diabetes sehingga diperoleh algoritma dengan tingkat akurasi yang lebih tinggi.
2. Dapat menggunakan metode optimasi seperti AdaBoost, dan lain-lain.

DAFTAR PUSTAKA

Bramer, M.(2007). *Principles of Data Mining* London: Springer Clark. L.A., Kochanska, G., & Ready, R. (2000). Mothers' personality and its interaction with child temperament as predictors of parenting behavior. Journal of Personality and Social Psychology, 79, 274-285.

Gorunescu, F. (2011). *Data Mining Concepts, Models and Technique*. Berlin: Springer

Hui-Huang, H. (2006). *Advanced Data mining Technologies in Bioinformatics*. United States of America: Idea Group Publishing.

Jayalakshmi, T., Santhakumaran, A. (2010). *Improved Gradient Descent Back Propagation Neural Network for Diagnoses of Type II Diabetes Militus. Global Journal of Computer Science and Technology*. Vol.9 Issue 5.

Kahramanli, H., & Allahverdi, N. (2008). *Design of A Hybrid System for the Diabetes and Heart Diseases. Expert System with Application,* 82-89

Kusrini, & Luthfi, T. E.(2009). *Algoritma Data Mining*. Yogyakarta: Penerbit Andi

Maimon, o., & Rokach, L. (2010). *Data Mining and Knowladge Discovery Handbook Second Edition.* New York:Springer.

Mason, R. (2005). The Natural Diabetes Cure. Usa: 4th Printing Spring 2012.

Mehdi Neshat, and Mehdi Yaghobi. (2009, October, 20-22). "*Designing a Fuzzy Expert System of Diagnosing the Hepatitis B Intensity Rate and Comparing it with Adaptive Neural Network Fuzzy System". Proceeding of the world congress on engineering and computer science* 2009,Vol II, WCECS 2009, ISBN:978-988-18210-2-7. pp 1-6, October 20-22.

Nurrahmani, U. (2012). Stop!Diabetes Mellitus. Yogyakarta: Familia

Robert, F. G., Zgonis, T., & Driver, V. R. (2006). Diabetic Foot Disorders: A Clinical Practice Guideline (2006 Revision). The Journal Of Foot & Ankle Surgery , 3.

Witten, I. H., Frank, E., & Hall, M.A.(2011) *Data Mining Practical Machine Learning Tools And Technique*. Burlington, Usa: Morgan kaufmann Publishers.