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

Fatmawati
Sistem Informasi
STMIK Nusa Mandiri Jakarta
Jl. Damai No.8, Warung Jati Barat (Margasatwa), Jakarta Selatan, Indonesia
fatmawati.fmw@bsi.ac.id

Abstract—*Diabetes is one of the deadly disease, high risk factors in families that cause diabetes because fat people who do not do physical exercise, and those who do not have a healthy lifestyle and diet excess of what is needed by the body. Based on the history data diabetics can be made on the prediction of diabetes that can help health professionals. Classification is one of data mining techniques that can be used to help predict .Classification can be done with that Decision Tree algorithm C4.5 and Naive Bayes. This study aims to classify and apply data mining classification. Results of data classification in the evaluation using the Confusion Matrix and ROC curve to determine the level of accuracy results using algorithms Decision Tree that is equal to 73.30% and the AUC of the ROC curve was 0733 while the algorithm Naive Bayes amounted to 75.13% AUC values of the ROC curve of 0.810, so it can be said that the algorithm Naive Bayes have the result of a good predictor in predicting diabetes patient.*

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

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 akurasi yang lebih tinggi dalam mendeteksi penyakit diabetes.

BAHAN DAN METODE

I. Kajian Literatur

Studi kasus mengenai prediksi penyakit Diabetes sudah cukup banyak. Berikut ini beberapa penelitian terkait mengenai prediksi penyakit diabetes.

Menurut (Purnama dan Supriyanto, 2013) dalam penelitiannya yang berrjudul Deteksi Penyakit Diabetes Type II Dengan *Naive Bayes* Berbasis *Particle Swarm Optimization*, metode yang digunakan dalam penelitiannya adalah *naive bayes* berbasis *particle swarm optimization* (PSO) untuk meningkatkan akurasi dalam deteksi penyakit diabetes. *Data set* yang digunakan sejumlah 598 pasien dengan parameter sebagai berikut: usia, jenis kelamin, kolesterol total, HDL, LDL, trigliserid, hemoglobin, lekosit, trombosit, tekanan darah, riwayat diabetes, olahraga, merokok, hamil. Dalam penelitiannya algoritma *naive bayes* berbasis *particle swarm optimization* terbukti akurat dengan akurasi 98.16% dan memiliki nilai AUC 0.99 dikategorikan ke dalam *excellent classification*. Nilai ini membuktikan bahwa algoritma *naive bayes* berbasis *particle swarm optimization* dapat meningkatkan akurasi pada deteksi penyakit diabetes *type* II.

Menurut (Andriani, 2013) dalam penelitiannya yang berjudul Sistem Prediksi Penyakit Diabetes Berbasis *Decision Tree*. Penelitian ini bertujuan membuat klasifikasi data diabetes dan menerapkannya dalam pengembangan sistem prediksi penyakit diabetes. Hasil klasifikasi data diabetes di evaluasi dengan *confusion matrix* dan kurva ROC (*Receiver Operating Characteristic*) untuk mengetahui tingkat akurasi hasil klasifikasi. Evaluasi yang dilakukan menunjukan hasil yang termasuk *Excellent Classification*. *Rule* hasil klasifikasi diimplementasikan untuk pembuatan sistem prediksi penyakit diabetes.

Menurut Romansyah, Sitanggang dan Nurdiati (2009) dalam penelitiannya yang berrjudul *Fuzzy Decision Tree* Dengan *Algoritme* ID3 Pada Data Diabetes. Penelitian ini bertujuan untuk menerapkan salah satu teknik klasifikasi yaitu Fuzzy ID3 (*Iterative Dichotomiser*) *Decision Tree* pada data hasil pemeriksaan lab pasien dan menemukan aturan klasifikasi pada data diabetes yang menjelaskan dan membedakan kelas-kelas atau konsep sehingga dapat digunakan untuk memprediksi penyakit diabetes berdasarkan nilai dari atribut lain yang diketahui. Dari hasil penelitian bahwa algoritme ID3 memiliki kinerrja yang baik dalam membentuk *fuzzy decision tree* untuk data diabetes yang ada. Nilai akurasi terbaik yang didapat dari model yaitu 94,15% diperoleh pada *fuzziness control threshold* ($Ø_r$)=75% dan *leaf decision threshold* ($Ø_n$)=8% atau 10%.nilai $Ø_r$ dan $Ø_n$ sangat berpengaruh terhadap jumlah aturan yang dihasilkan, nilai $Ø_r$ yang terlalu tinggi akan menyebabkan turunnya nilai akurasi. Dilain pihak, nilai $Ø_n$ yang terlalu rendah juga dapat menyebabkan akurasis menurun.

a. *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 *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 menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu.

5) *Interpretation/Evaluation*
Pola informasi yang dihasilkan dari proses *data mining* diterjemahkan menjadi bentuk yang lebih mudah dimengerti oleh pihak yang berkepentingan.

b. 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).

c. Algoritma C4.5

Algoritma C4.5 diperkenalkan oleh J. Ross Quinlan yang merupakan perkembangan dari algoritma ID3, algoritma tersebut digunakan 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.

d. *Decision Tree*

Pohon keputusan merupakan salah satu metode klasifikasi dan prediksi yang sangat kuat dan terkenal dalam penerapan *data mining.* Pada dasarnya *Decision Tree* mengubah data menjadi pohon keputusan (*decision tree*) dan aturan-aturan keputusan(*rule*).

Pohon keputusan juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target.

Sebuah pohon keputusan mungkin dibangun dengan saksama secara manual atau dapat tumbuh secara otomatis dengan menerapkan salah satu atau beberapa algoritma pohon keputusan untuk memodelkan himpunan data yang belum terklasifikasi. Banyak algoritma yang dipakai dalam pembentukan pohon keputusan antara lain ID3, CART dan C4.5 (Larose, 2005).

Variabel tujuan biasanya dikelompokan dengan pasti dan model pohon keputusan lebih mengarah pada perhitungan probabilitas dari tiap-tiap *record* kategori-kategori tersebut atau

untuk mengklasifikasikan *record* dengan mengkelompokannya dalam satu kelas. Decision tree membuat set rule yang paling efisien dan kemungkinan terkecil yang membuatnya menjadi *predictive* model yang baik. Jika terdapat *overlap* diantara dua prediktor maka yang terbaik dari keduanya yanga akan diambil. Pada sistem *rule induction*, keduanya akan diambil pada sistem ruleinductin salah satunya akan menjadi lemah atau kurang akurat.

Kelebihan-kelebihan *decision tree* adalah sebagai berikut:

1. Menyediakan *visual result*
2. Dibangun berdasarkan *rule-rule* yang dapat dimengerti dan dipahami
3. Bersifat *predictive*
4. Memungkinkan untuk melakukan prediksi
5. Menampilkan apa yang penting.

Algoritma akan mengidentifikasi atribut yang paling relevan dan akan mengidentifikasi suatu *set rule* yang akan memberikan *presentase* kemungkinan akan terjadi hal demikian dikemudian hari.

Decision tree dibuat dengan menggunakan sebuah teknik yang disebut *recursive partitioning.* Algoritma akan mendifinisikan atribut yang paling relevan dan akan men-split data yang ada berdasarkan atribut tersebut. Setiap *partision* disebut sebagai *rule*. Proses akan di ulang terus untuk setiap subgroup sampai ditemukan sebuah *good stopping point*. *Information gain measure* digunakan untuk memilih atribut mana yang akan dites pada *node tree* (Han & Kamber, 2006). Atribut dengan *information gain* yang paling tinggi akan dipilih sebagai test *atribute* untuk *current node*.

e. *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 seorang 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

f. *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.

g. 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).

Mengevaluasi akurasi dari model klasifikasi sangat penting, akurasi dari sebuah model mengindikasikan kemampuan model tersebut untuk memprediksi *class* target (Vercellis, 2009).

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*
Observe	*Class=*	*(True Positive*	*(False Negative –*

d Class	*Yes*	*– TP)*	*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

2) 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

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

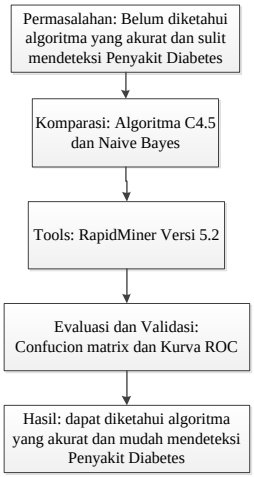
X= *Output Positif*

Y = *Output Negatif*

2. Metode Penelitian

a. Kerangka Pemikiran

Jenis penelitian yang digunakan dalam penelitian ini adalah model penelitian eksperimen. Penelitian ini bertujuan untuk melakukan perbandingan dan evaluasi pada algoritma klasifikasi data mining. Penelitian eksperimen ini menekankan pada teori-teori yang sudah ada. Pada penelitian ini, jenis penelitian yang diambil adalah eksperimen komparatif ini dilandasi oleh kerangka pemikiran pemecahan masalah seperti pada gambar 1.



Sumber: Hasil Penelitian (2015)

Gambar 1. Kerangka Pemikiran

Kerangka pemikiran 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. *Tools* yang digunakan *Rapid Miner* versi 5.2, data penelitian diambil dari, hhtp://archive.ics.uci.edu/ml/, 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,

b. Langkah- Langkah Penelitian

Dalam penelitian ini, digunakan sistem komputer dengan konfigurasi *hardware* yang terdiri dari Processor AMD E1-2100 APU, RAM 2 GB, hardisk 160 GB, dan *software* yang terdiri dari Sistem operasi Windows 8 Pro dan *tools data mining* rapid Miner.

Penelitian ini dilakukan dengan menjalankan beberapa langkah proses penelitian yaitu:

1. Pengumpulan data

2. Pengolahan awal data
3. Pengukuran penelitian
4. Analisa komparasi hasil

HASIL DAN PEMBAHASAN

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. Dari tabel 2 di atas dapat dirincikan sebagai berikut:

Tabel 3. Atribut, Tipe, Ukuran dan Nilai Atribut

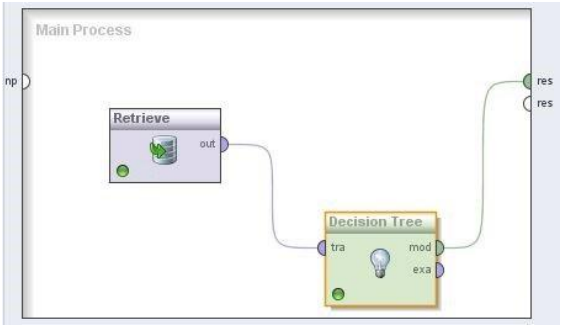
No.	Atribut	Tipe	Ukuran	Nilai Atribut
1.	Jumlah Hamil	*Integer*	*Scale*	Angka
2.	Konsentrasi Glukosa	*Integer*	*Scale*	Angka
3.	Tekanan Darah	*Integer*	*Scale*	Angka
4.	Lipatan Kulit	*Integer*	*Scale*	Angka

5.	Serum Insulin	*Integer*	*Scale*	Angka
6.	IMB	*Real*	*Scale*	Angka
7.	Riwayat Diabetes	*Real*	*Scale*	Angka
8.	Umur	*Integer*	*Scale*	Angka
9.	Hasil	String	Binominal	Positif, Negatif

Sumber: Hasil Penelitian (2015)

1. Pengolahan Awal Data

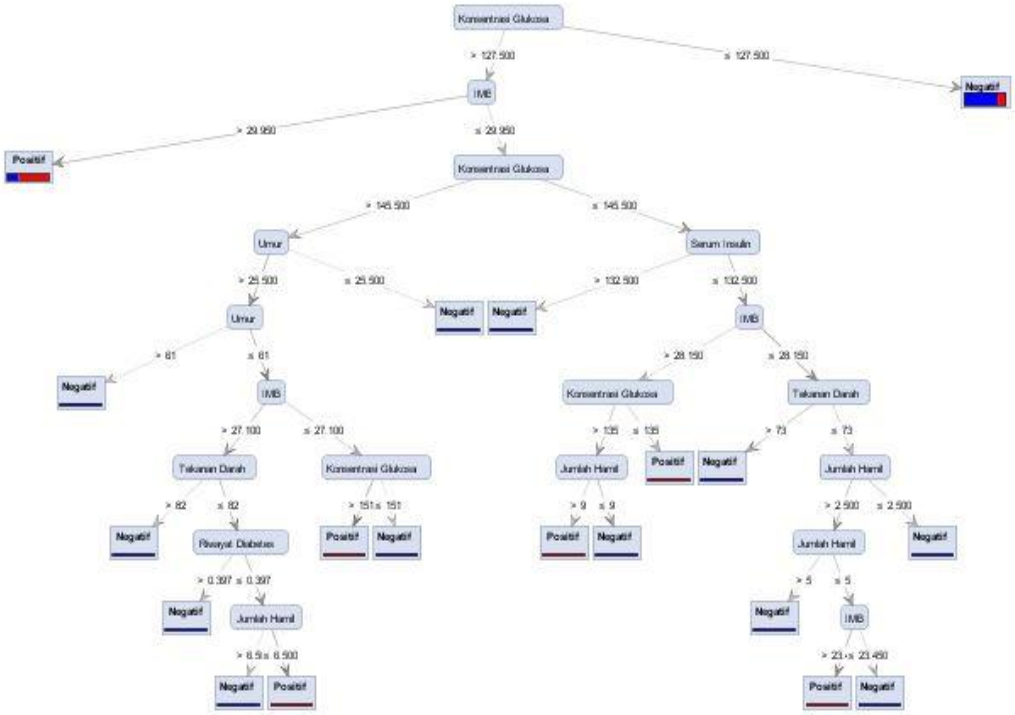
Pada tahap ini menentukan data yang akan diproses. pada tabel 2 dapat dibuat suatu model pohon keputusan dengan menggunakan *Rapid Miner* 5.2 dengan desain model sebagai berikut berikut:



Sumber: Hasil Penelitian (2015)

Gambar 2. Desain Model Algoritma C4.5

Pada gambar 2 diatas merupakan bentuk desain model dari dataset pasien diabetes kemudian di relasikan ke algoritma C4.5 menggunakan *tools Rapid Miner* versi 5.2. Dengan desain diatas menghasilkan sebuah pohon keputusan sebagai berikut:



Sumber: Hasil Penelitian (2015)

Gambar 3. Hasil Klasifikasi Menggunakan Algoritma C4.5

Pada Gambar 3 diatas terdapat 15 *rule* merupakan hasil dari klasifikasi dengan menggunakan model algoritma C4.5 dapat dijelaskan sebagai berikut:

R1: *IF* Konsentrasi Glukosa ≤127.500 *THEN* Hasil=Negatif

R2: *IF* Konsentrasi Glukosa >127.500 ^ IMB >29.950 *THEN* Hasil=Positif

R3: *IF* Konsentrasi Glukosa >127.500 ^ IMB ≤29.950 ^ Konsentrasi Glukosa ≤145.500 ^ Serum Insulin >132.500 *THEN* Hasil=Negatif

R4: *IF* Konsentrasi Glukosa >127.500 ^ IMB ≤29.950 ^ Konsentrasi Glukosa ≤145.500 ^ Serum Insulin ≤132.500 ^ IMB >28.150 ^ Konsentrasi Glukosa ≤135 *THEN* Hasil=Positif

R5: *IF* Konsentrasi Glukosa >127.500 ^ IMB ≤29.950 ^ Konsentrasi Glukosa ≤145.500 ^ Serum Insulin ≤132.500 ^ IMB >28.150 ^ Konsentrasi Glukosa >135 *THEN* Hasil=Negatif

R6: *IF* Konsentrasi Glukosa >127.500 ^ IMB ≤29.950 ^ Konsentrasi Glukosa ≤145.500 ^ Serum Insulin ≤132.500 ^ IMB ≤28.150 ^ Tekanan Darah >73 *THEN* Hasil=Negatif

R7: *IF* Konsentrasi Glukosa >127.500 ^ IMB ≤29.950 ^ Konsentrasi Glukosa ≤145.500 ^ Serum Insulin ≤132.500 ^ IMB ≤28.150 ^ Tekanan Darah ≤73 ^ Jumlah Hamil ≤2.500 *THEN* Hasil=Negatif

R8: *IF* Konsentrasi Glukosa >127.500 ^ IMB ≤29.950 ^ Konsentrasi Glukosa ≤145.500 ^ Serum Insulin ≤132.500 ^ IMB ≤28.150 ^ Tekanan Darah ≤73 ^ Jumlah Hamil >2.500 ^ Jumlah Hamil ≤5 *THEN* Hasil=Positif

R9: *IF* Konsentrasi Glukosa >127.500 ^ IMB ≤29.950 ^ Konsentrasi Glukosa ≤145.500 ^ Serum Insulin ≤132.500 ^ IMB ≤28.150 ^ Tekanan Darah ≤73 ^ Jumlah Hamil >2.500 ^ Jumlah Hamil >5 *THEN* Hasil=Negatif

R10: *IF* Konsentrasi Glukosa >127.500 ^ IMB ≤29.950 ^ Konsentrasi Glukosa >145.500 ^ Umur ≤25.500 *THEN* Hasil=Negatif

R11: *IF* Konsentrasi Glukosa >127.500 ^ IMB ≤29.950 ^ Konsentrasi Glukosa >145.500 ^ Umur >25.500 ^ Umur >61 *THEN* Hasil=Negatif

R12: *IF* Konsentrasi Glukosa >127.500 ^ IMB ≤29.950 ^ Konsentrasi Glukosa >145.500 ^ Umur >25.500 ^ Umur ≤61 ^ IMB ≤27.100 *THEN* Hasil=Positif

R13: *IF* Konsentrasi Glukosa >127.500 ^ IMB ≤29.950 ^ Konsentrasi Glukosa >145.500 ^ Umur >25.500 ^ Umur ≤61 ^ IMB >27.100 ^ Tekanan Darah >82 *THEN* Hasil=Negatif

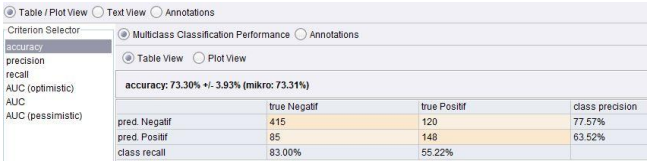
R14: *IF* Konsentrasi Glukosa >127.500 ^ IMB ≤29.950 ^ Konsentrasi Glukosa >145.500 ^ Umur >25.500 ^ Umur ≤61 ^ IMB >27.100 ^ Tekanan Darah ≤82 ^ Riwayat Diabetes >0.397 *THEN* Hasil=Negatif

R15: *IF* Konsentrasi Glukosa >127.500 ^ IMB ≤29.950 ^ Konsentrasi Glukosa >145.500 ^ Umur >25.500 ^ Umur ≤61 ^ IMB >27.100 ^ Tekanan Darah ≤82 ^ Riwayat Diabetes ≤0.397 *THEN* Hasil=Positif

2. Pengukuran Penelitian

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:

a. *Confucion Matrix*:



Sumber: Hasil Penelitian (2015)

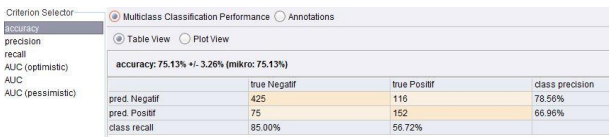
Gambar 4. Hasil Akurasi Prediksi Penyakit Diabetes Menggunakan Algoritma C4.5

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:



Sumber: Hasil Penelitian (2015)

Gambar 5. 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:

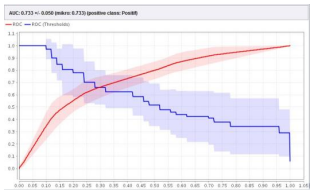
Keterangan:
TP = 152
FP =116
TN= 425
FN = 75

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%.

b. Kurva *ROC (Receiver Operating Characteristic)*

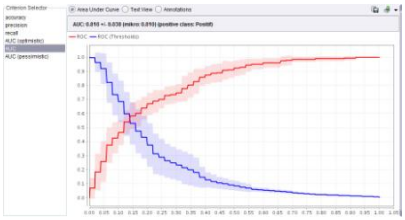
Data uji diatas akan dinilai hasil prediksi dengan menggunakan grafik ROC untuk algoritma C4.5, visualisasi dari grafik ROC yaitu:



Sumber: Hasil Penelitian (2015)

Gambar 6. Grafik ROC dari Model Algoritma C4.5

Dari gambar 6 terdapat grafik ROC dengan nilai AUC (*Area Under Curve*) sebesar 0.733 dimana diagnosa hasilnya *Fair Classification*. Sedangkan visualisasi grafik ROC dengan model algoritma *Naive Bayes* sebagai berikut:



Sumber: Hasil Penelitian (2015)

Gambar 7. Grafik ROC dari Model *Naive Bayes*

Dari gambar 7 terdapat grafik ROC dengan nilai AUC (*Area Under Curve*) sebesar 0.810 dimana diagnosa hasilnya *Good Classification.*

3. Analisa Hasil Komparasi

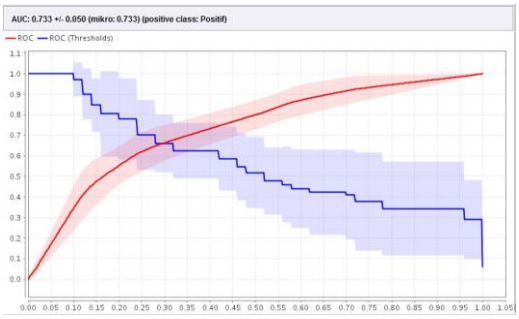
Dari hasil pengujian diatas baik evaluasi menggunakan *confusion matrix* maupun kurva *ROC* untuk model klasifikasi algoritma C4.5 dan *Naive Bayes* sebagai berikut:

Tabel 4. Hasil Komparasi Algoritma C4.5 dan *Naive Bayes*

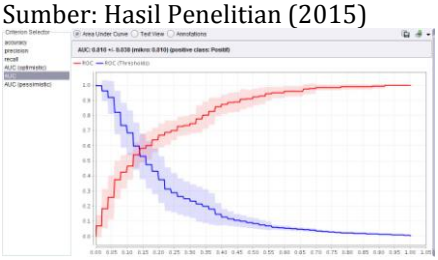
	Accuracy	AUC
Decision Tree	73.30%	0.733
Naive Bayes	75.13%	0.810

Sumber: Hasil Penelitian (2015)

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:



Sumber: Hasil Penelitian (2015)



Sumber: Hasil Penelitian (2015)

Gambar 8. Kurva ROC dengan Algoritma C4.5 dan *Naive Bayes*

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

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.*

REFERENSI

Andriani, Anik (2013). Sistem Prediksi penyakit Diabetes Berbasis *Decision Tree*. Jurnal Bianglala Informatika Vol. I No. 1 September 2013.

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

Han, J., & Kamber, M. (2006). *Data Mining Concept and Tehniques*. San Fransisco: Morgan Kauffman

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

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

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

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.

Purnama, Parida dan Catur Supriyanto (2013). Deteksi Penyakit Diabetes Type II Dengan *Naive Bayes* Berbasis *Particle Swarm Optimization.* Jurnal Teknologi Informasi, Volume 9 Nomor 2, Oktober 2013, ISSN 1414-9999.

Romansyah, F, I.S. Sitanggang, S, & Nurdianti (2009). *Fuzzy Decision Tree* Dengan *Algoritme* ID3 Pada Data Diabetes. *Internet Working* Indonesia *Journal* Vol. I/No. 2, 2009. *University of California Irvine Machine learning Repository*. Dikutip 02 November 2015, dari http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/heart/hear.dat

Vercellis, C. (2009). *Business Intelligence: Data Mining Optimization for Decision making*. United Kingdom: John Wiley & Sons.

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

BIODATA PENULIS

Fatmawati, M.Kom. Tangerang, 28 Agustus 1990. Tahun 2013 lulus dari Program Strata Satu (S1) Program Studi Sistem Informasi STMIK Nusa Mandiri Jakarta. Tahun 2015 lulus dari Program Strata Dua (S2) Program Studi Ilmu Komputer STMIK Nusa Mandiri Jakarta. Mengajar di kampus STMIK Nusa Mandiri Jakarta.