KONVERSI DATA TRAINING TENTANG PENYAKIT HIPERTENSI MENJADI BENTUK POHON KEPUTUSAN DENGAN TEKNIK KLASIFIKASI MENGGUNAKAN TOOLS RAPID MINER 4.1

Muhammad Syahril

Program Studi Sistem Informasi, STMIK Triguna Dharma

mhd.syahril@gmail.com

ABSTRAK: Penyajian data untuk menghasilkan nilai informasi sering kali ditampilkan dalam bentuk tabulasi. Kalau data yang ditampilkan memiliki kapasitas kecil, mungkin tidak terlalu sulit untuk mencerna kandungan informasi tersebut. Tetapi bila data yang disajikan memiliki kapasitas yang sangat besar, dikhawatirkan adanya kendala untuk menyerap informasi secara tepat dan cepat. Hal ini dikarenakan bahwa dibutuhkan waktu yang cukup lama untuk membaca data yang ditampilkan secara rinci hingga akhir data. Data yang akan dibahas pada tulisan ini adalah berkaitan dengan data historis yang berisi tentang kriteria seorang pasien yang berpotensi penyakit hipertensi atau tidak. Data historis yang ditampilkan akan dikonversi menjadi bentuk sebuah pohon keputusan. Dengan demikian penyerapan informasi akan menjadi lebih mudah dan lebih cepat untuk dilakukan. Tulisan ini mengimplementasikan disiplin ilmu Data Mining menggunakan teknik Klasifikasi Pohon Keputusan serta diaplikasikan dengan tools Rapid Miner 4.1.

Kata Kunci: Klasifikasi Pohon Keputusan, Rapid Miner 4.1

A. PENDAHULUAN

Konsep Data Mining merupakan upaya menggali informasi yang terpendam dalam timbunan data yang berjumlah besar. Data mining bukanlah suatu bidang yang sama sekali baru. Salah satu kesulitan untuk mendefinisikan data mining adalah kenyataan bahwa data mining mewarisi banyak aspek dan teknik dari bidang-bidang ilmu yang sudah mapan terlebih dulu. Beberapa solusi yang dapat diselesaikan oleh Data Mining adalah dalam bidang pasar dan manajemen, keuangan, telekomunikasi, keuangan, astronomi dan bidang-bidang lainnya.

Salah satu teknik Data Mining yang umum digunakan adalah Teknik Klasifikasi. Didalam Klasifikasi, ada beberapa metode klasifikasi yang digunakan, anatara lain: pohon keputusan, *rule based*, *neural network, support vector machine, naive bayes, dan nearest neighbour.* Dan pada kajian ini penulis akan menggunakan teknik pohon keputusan, karena beberapa alasan:

1. Dibandingkan dengan *classifier* JST atau *bayesian*, sebuah pohon keputusan mudah diinterpretasi/ ditangani oleh manusia.
2. Sementara training JST dapat menghabiskan banyak waktu dan ribuan iterasi, pohon keputusan efisien dan sesuai untuk himpunan data besar.
3. Algoritma dengan pohon keputusan tidak memerlukan informasi tambahan selain yang terkandung dalam data training (yaitu, pengetahuan domain dari distribusidistribusi pada data atau kelas-kelas).
4. Pohon keputusan menunjukkan akurasi klasifikasi yang baik dibandingkan dengan teknik-teknik yang lainnya.

B. POHON KEPUTUSAN

Pohon keputusan merupakan metode klasifikasi dan prediksi yang sangat kuat. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang mempresentasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami. Dan mereka juga dapat diekspresikan dalam bentuk bahasa basis data seperti *Structured Query Language* untuk mencari *record* pada katogori tertentu.

Pohon keputusan juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target. Karena pohon keputusan memadukan antara eksplorasi data dan pemodelan, dia sangat bagus sebagai langkah awal dalam proses pemodelan bahkan ketika dijadikan sebagai model akhir dari beberapa teknik lain.

Proses pada pohon keputusan adalah mengubah bentuk data (tabel) menjadi model pohon, mengubah model pohon menjadi *rule,* dan menyederhanakan *rule* (Basuki & Syarif, 2003)

Pohon Keputusan adalah struktur flowchart yang menyerupai tree (pohon), dimana setiap simpul internal menandakan suatu tes pada atribut, setiap cabang merepresentasikan hasil tes, dan simpul daun merepresentasikan kelas atau distribusi kelas. Alur pada Pohon Keputusan di telusuri dari simpul akar ke simpul daun yang memegang prediksi kelas untuk contoh tersebut. Pohon Keputusan mudah untuk dikonversi ke aturan klasifikasi (classification rules)(Zalilia, 2007).

Tinjauan pada kasus ini adalah bagaimana cara mengubah data historis menjadi sebuah pohon keputusan. Yang dapat memprediksi keputusan dalam hal penentuan apakah seseorang pasien baru berpotensi menderita hipertensi hanya dengan menentuan kriteria-kriteria yang dibutuhkan saja. Tanpa harus melalui proses pemeriksaan yang memakan waktu. Tentunya dengan memanfaatkan *Tools Rapid Miner* untuk menetapkan sebuah pohon keputusan yang bersumber dari data historis, untuk keperluan pemeriksaan dimasa mendatang.

C. KLASIFIKASI POHON KEPUTUSAN

Klasifikasi adalah sebuah proses untuk menemukan model yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang kelasnya tidak diketahui (Tan *et all*, 2004).

Di dalam klasifikasi diberikan sejumlah record yang dinamakan *training set*, yang terdiri dari beberapa atribut, atribut dapat berupa kontinyu ataupun kategoris, salah satu atribut menunjukkan kelas untuk *record*.

D. ALGORITMA POHON KEPUTUSAN

Banyak algoritma yang dapat dipakai dalam pembentukan pohon keputusan, antara lain ID3, CART, dan C4.5 (Larose,2005). Algoritma C4.5 merupakan pengembangan dari dari algoritma ID3 (Larose, 2005).

Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut :

a. Pilih atribut sebagai akar
b. Buat cabang untuk tiap-tiap nilai
c. Bagi kasus dalam cabang\
d. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Untuk memilih atribut sebagai akar, didasarkan pada nilai *gain* tertinggi dari atribut-atribut yang ada. Untuk menghitung *gain* digunakan rumus seperti tertera dalam persamaan berikut.

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

Keterangan :

S : himpunan kasus
A : atribut
n : jumlah partisi atribut A
|Si| : jumlah kasus pada partisi ke-i
|S| : jumlah kasus dalam S

Sementara itu, penghitungan nilai entropi dapat dilihat pada persamaan 2 berikut.

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

Keterangan :
S : himpunan kasus
A : fitur
n : jumlah partisi S
pi : proporsi dari S_i terhadap S

E. MODEL KLASIFIKASI

Model Klasifikasi terdiri dari (Tan et all, 2006):

1. Pemodelan Deskriptif yaitu dapat bertindak sebagai suatu alat yang bersifat menjelaskan untuk membedakan antara objek dengan kelas yang berbeda.
2. Pemodelan Prediktif dimana model klasifikasi juga dapat menggunakan prediksi label kelas yang belum diketahui recordnya.

F. TUJUAN KLASIFIKASI

Tujuan dari klasifikasi adalah untuk dapat menjelaskan sebagai berikut:

1. Menemukan model dari training set yang membedakan record kedalam kategori atau kelas yang sesuai, model tersebut kemudian digunakan untuk mengklasifikasikan record yang kelasnya belum diketahui sebelumnya pada test set.
2. Mengambil keputusan dengan memprediksikan suatu kasus, berdasarkan hasil klasifikasi yang diperoleh .

G. KONSEP DATA DALAM POHON KEPUTUSAN

Konsep data dalam pohon keputusan diberikan seperti pada Gambar 1, yakni sebagai berikut:

1. Data dinyatakan dalam bentuk tabel dengan atribut dan record.
2. Atribut menyatakan suatu parameter yang dibuat sebagai kriteria dalam pembentukan tree. Misalkan untuk menentukan seorang pasien berpotensi menderita hipertensi, kriteria yang diperhatikan adalah berat bedan, usia dan jenis kelamin. Salah satu atribut merupakan atribut yang menyatakan data solusi per-item data yang disebut dengan target atribut.
3. Atribut memiliki nilai-nilai yang dinamakan dengan instance. Misalkan atribut berat badan mempunyai instance berupa *overweight, average dan underweight*

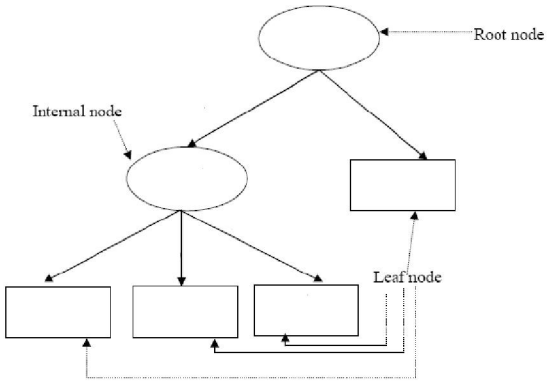
Gambar 1. Konsep Pohon Keputusan

H. KOMPOSISI POHON KEPUTUSAN

Sebagaimana sebuah pohon, komposisi Pohon Keputusan terdiri dari beberapa bagian yang disebut dengan simpul. Adapun pohon keputusan tersebut mempunyai 3 tipe simpul yaitu (Zalilia, 2007):

1. *Root Node* atau simpul akar dimana tidak ada masukan edge dan nol atau lebih keluaran edge (tepi),
2. *Internal Node* atau simpul internal, masing-masing memiliki satu masukan edge dan 2 atau lebih edge keluaran,
3. *Leaf Node* atau simpul daun disebut juga sebagai simpul akhir, masing-masing memiliki satu masukan edge dan tidak ada edge keluaran.

Pada Pohon Keputusan (Gambar 2) setiap simpul daun menandai label kelas. Simpul yang bukan simpul akhir terdiri dari akar dan simpul internal yang terdiri dari kondisi tes atribut pada sebagian record yang mempunyai karakteristik yang berbeda. Simpul akar dan simpul internal ditandai dengan bentuk oval dan simpul daun ditandai dengan bentuk segi empat (Han, 2001).



Gambar 2. Komposisi Pohon Keputusan

I. PENYIAPAN DAN PENGOLAHAN DATA TRAINING

Hal yang terpenting dalam penyelesaian kasus Pohon Keputusan adalah ketersediaan data training atau data histori. Untuk sampel kasus, penulis mencoba untuk memilih kasus hipertensi. Adapun data training yang dipakai adalah dalam format Microsoft Excel diberian pada Tabel 1 berikut.

Tabel 1. Data Training: Hipertensi.xls

Nama Pasien	Berat Badan	Usia	Jenis Kelamin	Hipertensi?
Oki lukman	Overweight	Tua	Perempuan	Ya
Pasha ungu	Overweight	Tua	Laki-laki	Ya
Budi anduk	Overweight	Tua	Laki-laki	Ya
Indra bekti	Overweight	Tua	Laki-laki	Ya
Luna maya	Overweight	Muda	Perempuan	Ya
Tukul	Overweight	Muda	Laki-laki	Ya
Afgan	Average	Tua	Laki-laki	Ya
Desta	Average	Tua	Laki-laki	Ya
Ringgo	Average	Muda	Laki-laki	Tidak
Ruben	Average	Muda	Laki-laki	Tidak
Titi kamal	Average	Muda	Perempuan	Tidak
Aurakasih	Average	Tua	Perempuan	Tidak
Jengkelin	Average	Tua	Perempuan	Tidak
Ari untung	Average	Muda	Laki-laki	Tidak
Gita gutawa	Underweight	Muda	Perempuan	Tidak
Fedi nuril	Underweight	Muda	Laki-laki	Tidak
Dian sastro	Underweight	Tua	Perempuan	Tidak
Nicholas	Underweight	Tua	Laki-laki	Tidak

J. TRANSFORMASI DATA

Tools *Rapid Miner* tidak meminta format data asli (*.xls) sebagaimana yang disiapkan di atas, melainkan membutuhkan format data yang lebih mudah diproses kedalam sistem aplikasinya. Format yang dimaksud adalah *comma separated value* atau *.csv. Oleh karena itu kita perlu mengubah format **hipertensi.xls** di atas menjadi **hipertensi.csv**

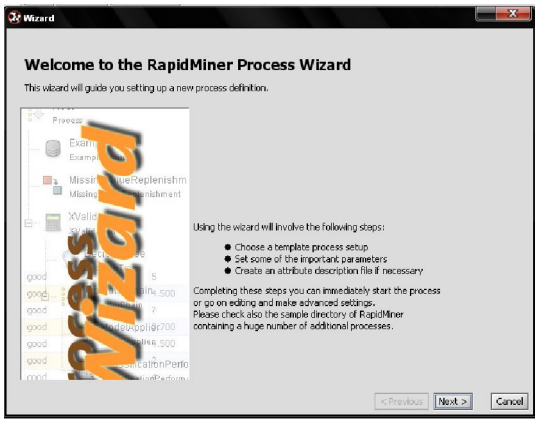
K. INSTALASI TOOLS DATA MINING *RAPID MINER*

Setelah ketersediaan data training dengan format yang diminta sudah dapat disiapkan, maka pastikan pula Tools *Rapid Miner* sudah terinstalasi secara baik.

L. BEKERJA DENGAN *RAPID MINER* 4.0

Bukalah Tools Aplikasi *Rapid Miner* 4.0 yang telah diinstalasi kemudian ikuti panduan berikut :

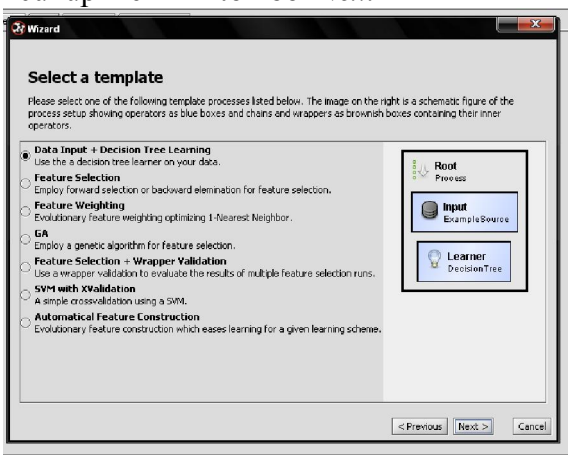
- Tampilan interface dari Rapid Miner wizard adalah sebagai berikut (Gambar 3). Setelah Aplikasi *Rapid Miner* 4.1 sudah aktif, Klik Menu *File* pilih *Wizard*



Gambar 3. Interface Wizard Rapid Miner

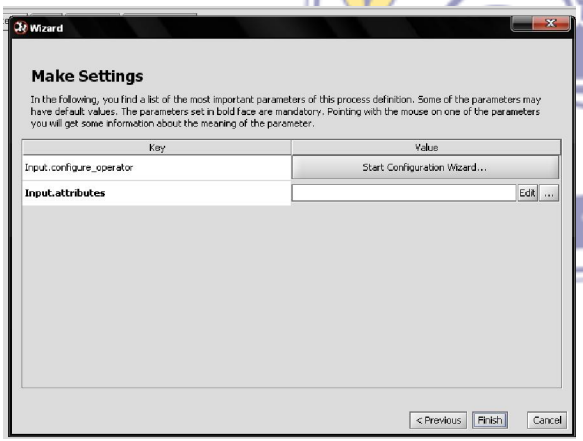
- Pada tampilan jendela wizard di atas, kita diberitahu tentang tahapan atau step yang akan dilakukan untuk memproses data

berdasarkan prosedur standar yang disiapkan oleh Tools *Rapid Miner*, kita cukup memilih tombol *Next*



Gambar 4. Wizard Memilih Template

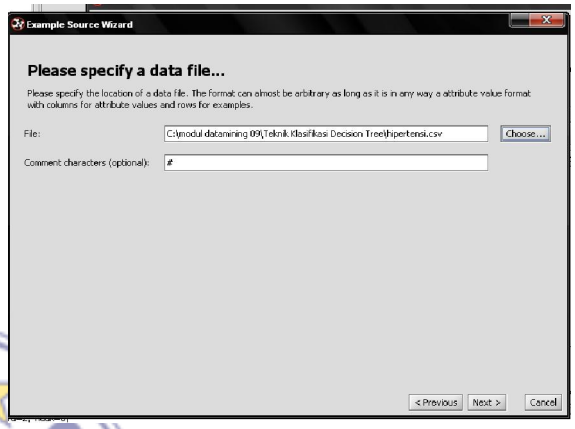
- Jendela yang muncul berikutnya adalah jendela *wizard Select Template* (Gambar 4). Pada kotak dialog ini Pilihlah opsi Data Input + Pohon *Decision Learning*. Kemudian kembali pilih tombol *Next* untuk melanjutkan.



Gambar 5. Pengaturan Setting Data Training

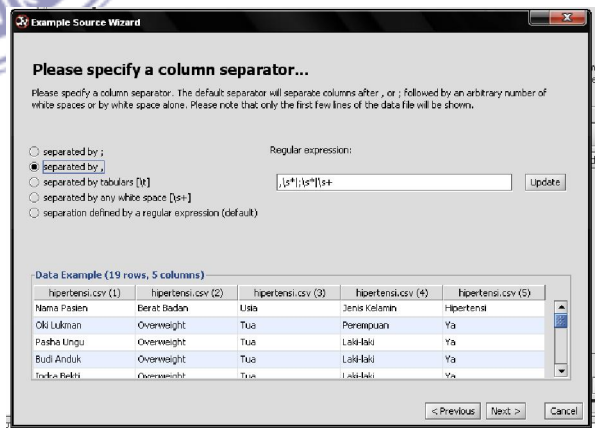
- Selanjutnya akan muncul kotak dialog *wizard Make Setting* (Gambar 5). Tahap ini akan membawa kita pada proses pengaturan konfigurasi data training sehingga data yang kita siapkan tersebut dapat diproses oleh sistem *Rapid Miner*. Untuk memulainya Klik Tombol *Start Configuration Wizard...,* maka akan muncul kotak dialog *Welcome To Example Source Wizard*, klik tombol *Next* untuk melanjutkan.

- Tahapan yang muncul berikutnya adalah *Please Specify a Data File* pada kotak dialog *Example Source Wizard* (Gambar 6). Disini kita diminta untuk memanggil data *training* yang sudah kita siapkan di awal yaitu *file* hipertensi.csv. Pilihlah tombol *Choose* untuk menetapkan nama dan alamat file hipertensi.csv dimaksud.



Gambar 6. Pemanggilan Data Training

- Setelah nama dan alamat file data *training* (hipertensi.csv) ditampilkan maka langkah berikutnya adalah mengklik *Next* untuk memulai melakukan pengaturan data *training*. Selanjutnya akan muncul kotak dialog berikut (Gambar 7).

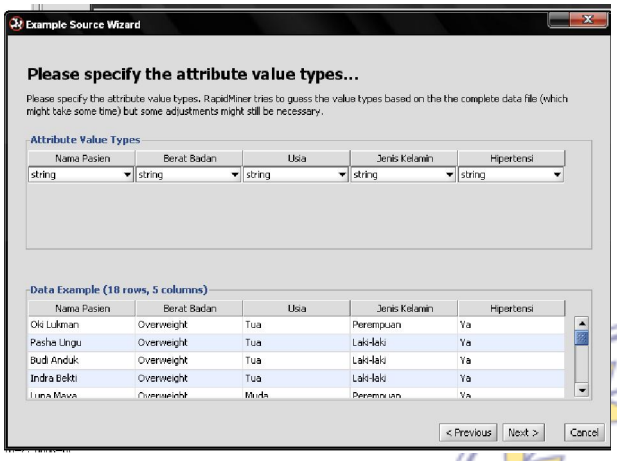


Gambar 7. Penentuan Jenis Pemisah Data Field

- Pada kotak dialog diatas kita diminta untuk menentukan tanda pemisah terhadap *field* data training (hipertensi.csv). Pada kasus ini kita cukup memilih "separated by, " yaitu opsi yang kedua yang menandakan bahwa

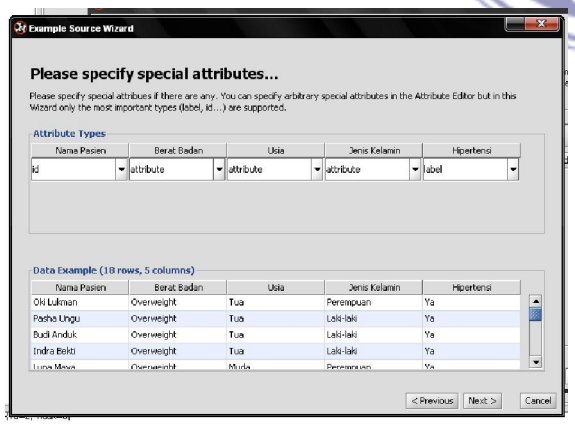
file data training yang kita siapkan adalah berformat comma separated value. Untuk melanjutkan klik tombol *Next*.

- Centang atau tandai pada pilihan *Use First Row For Column Names,* agar baris pertama dari data training dijadikan sebagai judul kolom. Klik *Next* untuk melanjutkan.



Gambar 8. Penentuan Type Atribut Data

- Pada kotak dialog *Please specify the attribute value types*…. (Gambar 8), aturlah *type* data untuk setiap atribut dan kolom dengan pilihan string (teks kalau bertipe data huruf) kemudian Klik tombol *Next* untuk melanjutkan.



Gambar 9. Penentuan Tipe Data Atribut

- Pada kotak dialog diatas (Gambar 9), kita diminta untuk mengatur tipe data untuk semua kolom yang ada. Untuk kolom Nama Pasien pilihlah tipe data Id, untuk kolom Berat Badan, Usia dan Jenis Kelamin Pilih tipe attribut. Sedangkan untuk kolom Hipertensi pilih tipe label seperti yang ditunjukkan oleh gambar di atas. Kembali Klik *Next* untuk melanjutkan.

- Setelah konfigurasi data dilakukan maka tahap selanjutnya adalah menyimpan data training berikut konfigurasi yang telah dilakukan tersebut ke dalam sebuah nama file baru yang berformat *.aml. Misalnya hipertensi.aml

- Klik *Finish* pada kotak dialog yang muncul untuk menegaskan nama file yang telah kita tentukan.

- Pada kotak dialog yang muncul berikutnya, Pilihlah tombol .... Pada input Attribut untuk membuka file *.aml (hipertensi.aml) yang telah kita buat sebelumnya. Secara umum *Rapid Miner* menyimpan data yang telah kita konfigurasi diatas ke dalam alamat berikut : " C:\Program Files\Rapid-I\RapidMiner-4.x \ hipertensi.aml"

- Maka bukalah *file* yang diminta (hipertensi.aml) dengan memiilih tombol Open pada kotak dialog yang aktif. Selanjutnya klik tombol *Finish*

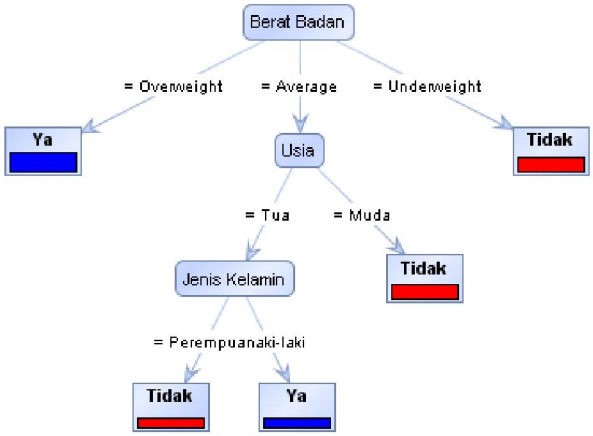
- Berikutnya kita akan dibawa ke halaman depan *Rapid Miner*. Klik Tombol *checklist* warna hijau pada panel atas, untuk mem validasi operator dan pengaturan yang telah di buat. Pastikan muncul tulisan *Process Ok* pada panel bawah status.

- Klik Tombol F5 pada *keyboard* untuk melihat hasil aplikasi

- Sebelum hasil ditampilkan biasanya selaku muncul pesan untuk menyimpan hasilnya yang akan ditampilkan dalam bentuk file baru yang berformat *.xml. Pilih tombol YES untuk menyimpan hasil output dalam bentuk *.xml (yaitu : hipertensi.xml)

- Perhatikan lokasi penyimpanan yaitu lokasi default *Rapid Miner*, klik save untuk mengakhiri langkah penyimpanan, dan melihat hasil pohon keputusan dari data training yang telah diteliti.

- Berikut hasil akhir data training studi kasus klasifikasi penyakit hipertensi yang telah diubah menjadi berupa pohon keputusan yang mudah difahami, dapat dilihat pada Gambar 10.



Gambar 10. Hasil Konversi Data Training Menjadi Pohon Keputusan

M. SIMPULAN

Data yang ditampilkan dalam bentuk pohon keputusan lebih mudah dan cepat untuk difahami dibandingkan bila data disajikan dalam bentuk tabulasi. *Rapid Miner* 4.1 berperan cukup baik dalam mengkonversikan data training yang ada. Sehingga walaupun data yang dikonversikan dalam bentuk sebuah pohon keputusan berasal dari data yang sangat besar, namun pada dasarnya nilai kandungan data yang ditampilkan punya kecenderungan memiliki pola yang dapat menghasilkan data yang cukup informatif.

N. DAFTAR PUSTAKA

Jogiyanto H. M. 1999. *Analisa dan Desain.* Yogyakarta: Andi Offset.

Kusrini, Emha Taufiq Luthfi. 2009. *Algoritma Data Mining.* Yogyakarta : Penerbit Andi

Kusrini. 2008. *Aplikasi Sistem Pakar*, Yogyakarta: Penerbit Andi.

----. 2006. *Sistem Pakar Teori dan Aplikasi*. Yogyakarta: Penerbit Andi.

Larose, Daniel T. 2005. *Discovering Knowledge in Data: An Introduction to Data Mining.* John Willey & Sons, Inc.

Laboratorim Data Mining. 2011. Klasifikasi Decision Tree dalam "http://datamining-lab.com". Yogyakarta : Fakultas Teknologi Industri Universitas Islam Indonesia

Rapid Miner4.1 manual

Santosa, Budi. 2007. *Data Mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis: Teori dan Aplikasi.* Yogyakarta : Graha Ilmu

Susanto, Sani., dan Suryadi, Dedi. 2010. *Pengantar Data Mining: Menggali Pengetahuan Dari Bongkahan Data.* Yogyakarta : Andi