PENERAPAN ALGORITMA C4.5 UNTUK PREDIKSI PENYAKIT HEPATITIS

Wisti Dwi Septiani
Manajemen Informatika
Akademi Manajemen Informatika dan Komputer Bina Sarana Informatika Jakarta
Jl. RS. Fatmawati N0. 24 Pondok Labu, Jakarta Selatan
wisti.wst@bsi.ac.id

ABSTRACT

Hepatitis is a inflammation disease of the liver because infection that attacks and causes damage to cells and liver function. Hepatitis is a disease precursor of liver cancer. Hepatitis can damage liver function as neutralizing poisons and digestive system in the body that break down nutrients and then spread to all organs of the body that very important for humans. Research of predicting disease hepatitis have been carried out by previous researchers. This research using the method of classification data mining algorithm C4.5. The measurement ofmethod using cross validation, confusion matrix and ROC curve. The result of this research is a decision tree rule with 77.29% accurate.

***Keywords** : Hepatitis, Data Mining, Algorithm C4.5*

PENDAHULUAN

Dewasa ini dalam dunia kesehatan, diagnosis penyakit menjadi hal yang sangat sulit dilakukan.Namun demikian catatan rekam medis telah menyimpan gejala-gejala penyakit pasien dan diagnosis penyakitnya.Hal seperti ini tentu sangat berguna bagi para ahli kesehatan.Mereka dapat menggunakan catatan rekam medis yang sudah ada sebagai bantuan untuk mengambil keputusan tentang diagnosis penyakit pasien.(Prasetyo, 2012).

Penyakit hepatitis merupakan penyakit peradangan hati karena infeksi virus yang menyerang dan menyebabkan kerusakan pada sel-sel dan fungsi organ hati.Penyakit hepatitis merupakan penyakit cikal bakal dari kanker hati.Penyakit hepatitis dapat merusak fungsi organ hati sebagai penetral racun dan sistem pencernaan makanan dalam tubuh yang mengurai sari-sari makanan untuk kemudian disebarkan ke seluruh organ tubuh yang sangat penting bagi manusia.Hepatitis atau peradangan hati merupakan salah satu dari banyaknya jenis penyakit hati, yang lainnya seperti pembengkakan hati (*fatty liver*) dan kanker hati (*cirrhosis)*. Di Indonesia, pada tahun 2007 penyakit hati merupakan salah satu dari sepuluh besar penyakit penyebab kematian terbesar di Indonesia (Departemen Kesehatan RI, 2009).

Seiring dengan perkembangan ilmu pengetahuan dan teknologi informasi, kehadiran cabang ilmu baru di bidang komputer *data mining* telah menarik banyak perhatian dalam dunia sistem informasi. Literatur mengenai pembahasan prediksi hepatitis telah dilakukan dengan beberapa metode. Berikut metode-metode yang pernah digunakan untuk menyelesaikan prediksi penyakit hepatitis:

Tabel 1. Tinjauan Studi Terdahulu

Peneliti	Masalah	Metode	Hasil
- Lale Ozyilmaz - Tulay Yildirim (2003)	Prediksi penyakit hepatitis dengan tiga algoritma: - *Multilayer Perceptron(MLP)* - *Radial Basis Function (RBF)* - *Conic Section Function Neural Network(CSFNN)*	Framework: Matlab	Akurasi: - MLP : 81,375 % - RBF : 85% - CSFNN : 90%
- Bekir Karlik (2011)	Prediksi penyakit hepatitis dengan dua algoritma : - *Backpropagation* - *Naïve Bayes*	- 10Fold Cross Valdiation - Confusion Matrix - ROC Area - Framework RapidMiner	Akurasi : - *86% Naïve Bayes* - *98% Backpropagation*
- Varun Kumar - Vijay Sharathi - Gayatri Devi (2012)	Prediksi penyakit hepatitis dengan algoritma *Support Vector Machine (SVM)* dengan fitur seleksi.	- Chi-Square - Fitur Seleksi - Framework RapidMiner	Akurasi : - 79,33% SVM - 83,12% fitur seleksi
- Ahmed Mohamed Samir Ali Gamal Eldin (2011)	Prediksi penyakit hepatitis menggunakan CART dengan 939 sampel (199 virus melakukan pembelahan dan 740 tidak melakukan pembelahan)	- 10Fold Cross Valdiation - Confusion Matrix - Sensitivity - Specificity - Framework Matlab	Data Training : Accuracy 99% Sensitivity 98% Spesificity 99% Data Testing: Accuracy 96% Sensitivity 95,5% Spesificity 98,6%

Decision tree telah banyak digunakan untuk melakukan klasifikasi dan prediksi di berbagai bidang.Algoritma C4.5 merupakan salah satu metode dalam *decision tree.*

Decision tree mengubah fakta yang sangatbesar menjadi pohon keputusan yang merepresentasikan aturan (Suhartinah, 2010).Untuk itu dalam penelitian ini akan dilakukan perhitungan menggunakan metode algoritma C4.5 untuk prediksi penyakit hepatitis.

BAHAN DAN METODE

Data Miningtelah menarik banyak perhatian dalam dunia sistem informasi dan dalam masyarakat secara keseluruhan dalam beberapa tahun terakhir, karena ketersediaan luas dalam jumlah besar data dan kebutuhan segera untuk mengubah data tersebut menjadi informasi yang berguna dan pengetahuan.*Data mining* adalah untuk mengekstrasikan atau "menambang" pengetahuan dari kumpulan banyak data (Han dan Kamber, 2007).

Data mining, sering juga disebut *knowledge discovery in database (KDD)*, adalah kegiatan yang meliputi pengumpulan, pemakain data historis untuk menentukan pola keteraturan, pola hubungan dalam set data berukuran besar (Santosa, 2007).

Berdasarkan tugasnya, *data mining* dikelompokkan menjadi 6 yaitu deskripsi, estimasi, prediksi, klasifikasi, clustering, dan asosiasi (Larose, 2005). Klasifikasi (taksonomi) adalah proses menempatkan objek tertentu (konsep) dalam satu set kategori, berdasarkan masing-masing objek (konsep) *property* (Gorunescu, 2011). Proses klasifikasi didasarkan pada empat komponen mendasar yaitu kelas, prediktor, *training set*, dan pengujian *dataset*.

Diantara model klasifikasi yang paling popular adalah *Decision/Classification Trees, Bayesian Classifiers/Naïve Bayes Classifiers, Neural Networks, Statistical Analysis, Genetic Algorithms, Rough Sets, K-Nearest Neighbor Classifier, Rule-based Methods, Memory Based Reasoning, Support Vector Machines* (Gorunescu, 2011).

Algoritma C4.5. *Decision Tree* menyerupai struktur *flowchart*, yang

masing-masng internal *node*-nya dinyatakan sebagai atribut pengujian, setiap cabang mewakili *output* dari pengujian, dan setiap *node* daun (*terminal node*) menentukan label *class*.*Node* paling atas dari sebuah pohon adalah *node* akar (Han & Kamber, 2007).

Algoritma C4.5 menggunakan konsep *information gain* atau *entropy reduction* untuk memilih pembagian yang optimal (Larose, 2005). Tahapan dalam membuat pohon keputusan dengan algoritma C4.5 (Gorunescu, 2011) yaitu:

1. Mempersiapkan data *training*, dapat diambil dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokkan dalam kelas-kelas tertentu.
2. Menentukan akar dari pohon dengan menghitung nilai *gain* yang tertinggi dari masing-masng atribut atau berdasarkan nilai *index entropy* terendah. Sebelumnya dihitung terlebih dahulu nilai *index entropy*, dengan rumus:

$$Entropy\ (i) = -\sum_{j=1}^{m} f(i,j).\log 2\ f[(i,j)]$$

3. Hitung nilai *gain* dengan rumus:

$$Entropy\ split = \sum_{i=1}^{p} \left(\frac{n1}{n}\right).\ IE(i)$$

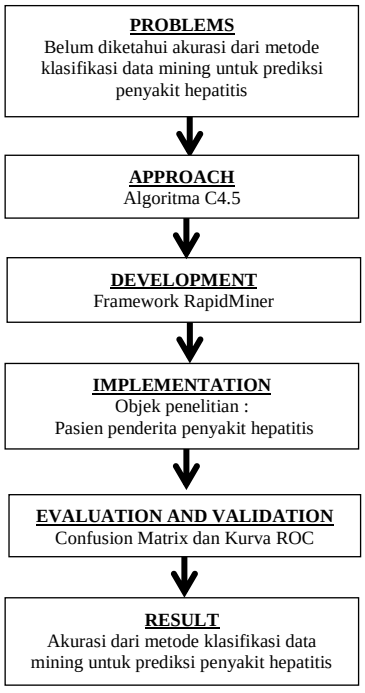
4. Ulangi langkah ke-2 hingga semua *record* terpartisi. Proses partisi pohon keputusan akan berhenti disaat:

   a. Semua tupel dalam *record* dalam simpul N mendapat kelas yang sama.

   b. Tidak ada atribut dalam *record* yang dipartisi lagi.

   c. Tidak ada *record* di dalam cabang yang kosong.

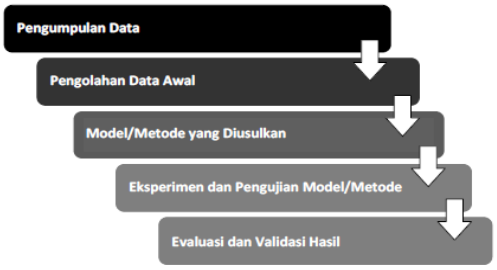
Dalam menyelesaikan penelitian perlu dibuat sebuah kerangka pemikiran yang berguna sebagai pedoman atau acuan penelitian ini sehingga penelitian dapat dilakukan secara konsisten.Penelitian ini terdiri dari beberapa tahap seperti terlihat pada gambar 1 di bawah ini.Permasalahan pada penelitian ini adalah belum diketahui akurasi dari metode klasifikasi data mining untuk prediksi penyakit hepatitis. Oleh sebab itu metode yang digunakan untuk memecahkan masalah adalah Algoritma C4.5 dengan melakukan pengujian terhadap kinerja metode tersebut.Pengujian metode dilakukan dengan cara*confusion matrix*dan kurva ROC serta menggunakan *toolsRapidMiner*Berikut ini adalah kerangka permikiran dari penelitian ini:



Sumber: Hasil Penelitian (2013)

Gambar 1. Kerangka Pemikiran

Pada penelitian ini data yang digunakan adalah data yang didapat dari *Machine Learning Repository*UCI (Universitas California Invene) dengan alamat web: http://archive.ics.uci.edu/ml/. Dalam penelitian ini akan dilakukan beberapa langkah-langkah atau tahapan penelitian seperti gambar di bawah ini:



Sumber: Hasil Penelitian (2013)

Gambar 2.Tahapan Penelitian

1. Pengumpulan Data

Teknik pengumpulan data ialah teknik atau cara-cara yang dapat digunakan untuk menggunakan data (Riduwan, 2008).Dalam pengumpulan data terdapat sumber data, sumber data yang dihimpun langsung oleh peneliti disebut dengan sumber primer, sedangkan apabila melalui tangan kedua disebut sumber sekunder (Riduwan, 2008). Data pada penelitian ini merupakan data sekunder yang diperolah dari *Machine Learning Repository* UCI (Universitas California, Invene) dengan alamat web http://archive.ics.uci.edu/ml/machine-learning-databases/hepatitis/. Data yang dikumpulkan adalah data pemeriksaan pasien penyakit hepatitis oleh G. Gong (Carnegie – Mellon University) di Yugoslavia pada November 1988. Data terkumpul sebanyak 155 data dengan 123 pasien penyakit hepatitis yang hidup dan 32 pasien penyakit hepatitis yang mati dengan atribut *age, sex, steroid, antivirals, fatigue,malaise, anorexia, liver_big, liver_firm,spleen_palpable, spiders, ascites, varices,bilirubin, alk_phosphate, sgot, albumin, protime, histology,* dan *class* (atribut hasil prediksi).

2. Pengolahan Data Awal

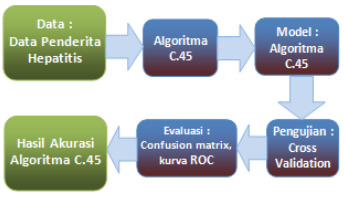
Untuk mendapatkan data yang berkualitas, beberapa teknik yang dilakukan adalah sebagai berikut (Vecellis, 2009):

a. *Data validation*, untuk mengidentifikasi dan menghapus data yang ganjil (*outlier/noise*), data yang tidak konsisten, dan data yang tidak lengkap (*missing value*).
b. *Data integration and transformation*, untuk meningkatkan akurasi dan efisiensi algoritma. Data yang digunakan dalam penelitian ini bernilai kategorikal.
c. *Data size reduction and dicritization*, untuk memperoleh *dataset* dengan jumlah atribut dan *record* yang lebih sedikit tetapi bersifat informatif.

Dari proses pengolahan awal data di atas diperoleh sebanyak 155 data dengan 123 data dengan kelas "HIDUP" dan 32 data dengan kelas "MATI".

3. Metode yang Diusulkan

Dalam penelitian ini metode yang diusulkan adalah metode klasifikasi *data mining* algoritma C4.5. Dalam tahapan ini akan dilakukan beberapa langkah-langkah sebagai berikut:



Sumber: Hasil Penelitian (2013)

Gambar 3.Model yang diusulkan

HASIL DAN PEMBAHASAN

Eksperimen dan Pengujian Metode.Pada tahap ini dilakukan eksperimen dan pengujian metode yang digunakan yaitu menghitung dan mendapatkan *rule-rule* yang ada pada algoritma yang diusulkan yaitu Algoritma C.45. Langkah-langkah yang dilakukan sebagai berikut:

1. Menghitung jumlah kasus "LIFE" dan "DIE" serta nilai *Entropy* dari semua kasus. Dari data *training* yang ada diketahui jumlah kasus yang "LIFE" sebanyak123 *record*, dan jumlah kasus yang "DIE" adalah sebanyak 32 *record* total kasuskeseluruhan adalah 155 kasus. Sehingga didapat *entropy* keseluruhan:

$$Entropy = -\sum_{j=1}^{m} f(i,j).\log 2\, f[(i,j)]$$

= (-123/155 * log2 (123/155)) + (-32/155 * log2 (32/155))

= 0,7346

2. Hitung nilai *entropy* dan nilai *gain* masing-masing atribut. Nilai *gain* tertinggi adalah atribut yang menjadi *root* dari pohon keputusan yang akan dibuat. *Entropy* atribut dihitung dengan rumus sebagai berikut:

$$Entropy\ split = \sum_{i=1}^{p} \left(\frac{n1}{n}\right).\ IE(i)$$

Terdapat 10 atribut yaitu *age, steroid, malaise, liver_big, spiders, varices, bilirubin, sgot, albumin,* dan *protime.*

Menghitung *entropy* dan *gain* bagi atribut *age*.

<= 32,5	= 40/155
> 32,5	= 155/155
<= 49	= 110/155
> 49	= 45/155
<= 61,5	= 144/155
> 61,5	= 11/155

Atribut *age* <= 32,5 terdiri dari 38 class "LIFE" dan 2 class "DIE", untuk atribut *age* > 32,5 terdiri dari 85 class "LIFE" dan 30 untuk class "DIE", untuk atribut *age* <= 49 terdiri dari 89 class "LIFE" dan 21 class "DIE", untuk atribut *age* > 49 terdiri dari 34 class "LIFE" dan 11 class "DIE", untuk atribut *age* <= 61,5 terdiri dari 114 class "LIFE" dan 30 class "DIE", untuk atribut *age* > 61,5 terdiri dari 9 class "LIFE" dan 2 class "DIE".

Maka *entropy* untuk atribut *age* adalah sebagai berikut :

$E_{<=32,5}$[38,2] = (-38/40 * $\log_2$(38/40)) + (-2/40 * $\log_2$ (2/40) = 0,2863

$E_{>32,5}$[85,30] = (-85/115 * $\log_2$(85/115)) +(-30/115 * $\log_2$ (30/115)) = 0,8280

$E_{<=49}$[89,21] = (-89/110 * $\log_2$ (89/110)) + (-21/110 * $\log_2$ (21/110))= 0,7033

$E_{>49}$[34,11] = (-34/45 * $\log_2$ (34/45)) + (-11/45 * $\log_2$ (11/45)) = 0,8023

$E_{<=61,5}$[114,30] =(-114/144 * $\log_2$(114/144))+(-30/144 * $\log_2$ (30/144)) = 0,7382

$E_{>61,5}$[9,2] = (-9/11 * $\log_2$ (9/11)) + (-2/11 * $\log_2$ (2/11)) = 0,6840

E split age = (40/155 * (0,2863)) + (115/155 * (0,8280))= (110/155 * (0,7033)) + (45/155 * (0,8023)) = (144/155 * (0,7382)) + (11/155 * (0,6840)) = 0,6882 + 0,7320 + 0,7343 = 2,1545

Gain age = 0,7346 – 2,1545
= – 1,42

Dengan cara yang sama, dilakukan perhitungan *entropy* dan *gain* bagi atribut lainnya yaitu *steroid, malaise, liver_big, spiders, varices, bilirubin, sgot, albumin,* dan *protime.*

E splitsteroid= (79/155 * (0,6145)) + (76/155 * (0,8314))= 0,7208

Gainsteroid = 0,7346 – 0,7208
= 0,0137

E splitmalaise= (94/155 * (0,4553)) + (61/155 * (0,9559))= 0,6523
Gainmalaise = 0,7346 – 0,6523
= 0,0822

E splitliver_big = (130/155 * (0,7657)) + (25/155 * (0,5293))= 0,7275
Gainliver_big = 0,7346 – 0,7275
= 0,0070

E splitspiders= (104/155 * (0,4566)) + (51/155 * (0,9863))= 0,6308
Gainspiders = 0,7346 – 0,6308
= 0,1037

E splitvarices= (137/155 * (0,6180)) + (18/155 * (0,9640))=0,6581
Gainvarices = 0,7346 – 0,6581
= 0,0764

E splitbilirubin= (105/155 * (0,4220)) + (50/155 * (0,9953)) = 0,6069 + 0,7333 = 1,3402
Gainbilirubin = 0,7346 – 1,3402
= – 0,6056

E splitsgot=(102/155 * (0,6722)) + (53/155 * (0,8329))=0,7271
Gainsgot = 0,7346 – 0,7271
= 0,0074

E splitalbumin=(7/155 * (0)) + (148/155 * (0,6522))=0,6227
Gainalbumin = 0,7346 – 0,6227
= 0,1119

E split protime= (20/155 * (0,9340)) + (135/155 * (0,5861))=0,1205 + 0,5104=0,6309
Gain protime = 0,7346 – 0,6309
= 0,1037

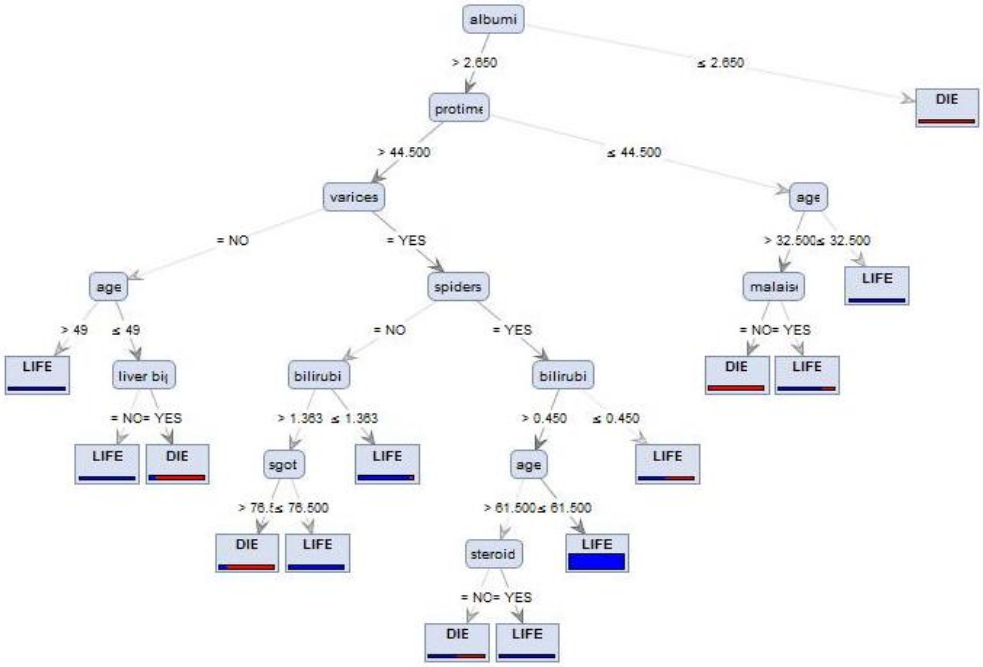
Tabel 2. Nilai *entropy* dan *gain* untukpenentuan *root*

Simpul	Entropy	Gain
	0,7346	
Age		-1,42
<= 32,5 thn	0,2863	
> 32,5 thn	0,8280	
<= 49 thn	0,7033	
> 49 thn	0,8023	
<= 61,5 thn	0,7382	
> 61,5 thn	0,6840	
Steroid		0,0137
Yes	0,6145	
No	0,8314	
Malaise		0,0822
Yes	0,4553	
No	0,9559	
Liver_big		0,0070
Yes	0,7657	
No	0,5293	
Spiders		0,1037
Yes	0,4566	
No	0,9863	
Varices		0,0764
Yes	0,6180	
No	0,9640	
Bilirubin		-0,6056
<= 1,363	0,4220	
> 1,363	0,9953	
<= 0,450	0,9182	
> 0,450	0,7297	
Sgot		0,0074
<= 76,500	0,6722	
> 76,500	0,8329	
Albumin		**0,1119**
<= 2,650	0	
> 2,650	0,6552	
Protime		0,1037
<= 44,500	0,9340	
> 44,500	0,5861	

Sumber: Hasil Penelitian (2013)

Dari tabel 2 dapat dilihat nilai *gain* tertinggi ada pada atribut *albumin* yakni 0,1119 sehingga didapat bahwa atribut *albumin* adalah akar (*root*) dari pohon keputusan. Kemudian dilakukan kembali perhitungan nilai *entropy* dan *gain* untuk menentukan simpul 1.1, nilai yang dihitung berdasarkan atribut *albumin*<= 2,650 dan atribut *albumin*> 2,650.

Dari tabel perhitungan menentukan simpul 1.1 untuk atribut *albumin*> 2,650 diperoleh gain tertinggi yaitu protime dengan nilai 0,2092sehingga atribut tersebut dijadikan simpul 1.1. Untukmenetukan simpul selanjutnya, dilakukan perhitungan nilai *entropy* dan *gain* dengan cara yang sama, sehingga diperoleh pohon keputusan seperti gambar di bawah.



Sumber: Hasil Penelitian (2013)

Gambar 4. Pohon keputusan hasil Algoritma C4.5

Dari pohon keputusan pada gambar 4 didapatkan *rule* untuk memprediksi penyakit hepatitis. *Rule* yang didapat sebagai berikut :

R1: Jika albumin <= 2,650 maka pasien "DIE".

R2: Jika albumin > 2,650 dan protime > 44,500 dan varices = NO dan age > 49 tahun maka pasien "LIFE".

R3: Jika albumin > 2,650 dan protime > 44,500 dan varices = NO dan age <= 49 tahun dan liver_big = NO maka pasien "LIFE"

R4: Jika albumin > 2,650 dan protime > 44,500 dan varices = NO dan age <= 49 tahun dan liver_big = YES maka pasien "DIE"

R5: Jika albumin > 2,650 dan protime > 44,500 dan varices = YES dan spiders = NO dan bilirubin > 1,363 dan sgot > 76,500 maka pasien "DIE".

R6: Jika albumin > 2,650 dan protime > 44,500 dan varices = YES dan spiders = NO dan bilirubin > 1,363 dan sgot <= 76,500 maka pasien "LIFE".

R7: Jika albumin > 2,650 dan protime > 44,500 dan varices = YES dan spiders = NO dan bilirubin <= 1,363 maka pasien "LIFE".

R8: Jika albumin > 2,650 dan protime > 44,500 dan varices = YES dan spiders = YES dan bilirubin > 0,450 dan age > 61,5 tahun dan steroid = NO maka pasien "DIE".

R9: Jika albumin > 2,650 dan protime > 44,500 dan varices = YES dan spiders = YES dan bilirubin > 0,450 dan age > 61,5 tahun dan steroid = YES maka pasien "LIFE".

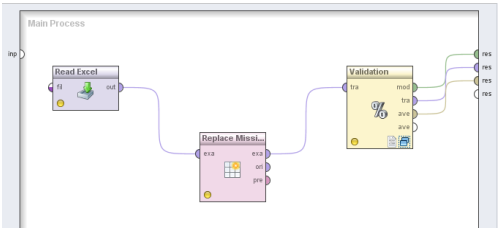
R10: Jika albumin > 2,650 dan protime > 44,500 dan varices = YES dan spiders = YES dan bilirubin > 0,450 dan age <= 61,5 tahun maka pasien "LIFE".

R11: Jika albumin > 2,650 dan protime > 44,500 dan varices = YES dan spiders = YES dan bilirubin <= 0,450 maka pasien "LIFE".

R12: Jika albumin > 2,650 dan protime <= 44,500 dan age > 32,5

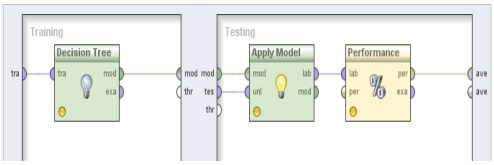
tahun dan malaise = NO maka pasien "DIE".
R13: Jika albumin > 2,650 dan protime <= 44,500 dan age > 32,5 tahun dan malaise = YES maka pasien "LIFE".
R14: Jika albumin > 2,650 dan protime <= 44,500 dan age <= 32,5 tahun maka pasien "LIFE".

Pengujian dengan 10-*Fold Cross Validation*untuk model Algoritma C4.5ini menggunakan aplikasi RapidMiner seperti berikut:



Sumber: Hasil Penelitian (2013)

Gambar 5.Pengujian 10-*Fold Cross Validation*



Sumber: Hasil Penelitian (2013)

Gambar 6.*Validation* Model Algoritma C4.5

Evaluasi dan Validasi Hasil. Setelah data diolah maka dapat diuji tingkat akurasinya untuk melihat kinerja darimetode Algoritma C4.5.Penelitian ini bertujuan untuk melihat akurasi analis data pasien penderita penyakit hepatitis, menilai kemungkinan kelangsungan hidup penderita apakah hidup atau mati. Pengujian tingkat akurasi dilakukan dengan menggunakan *confussion matrix* dan kurva ROC/AUC (*Area Under Cover*).

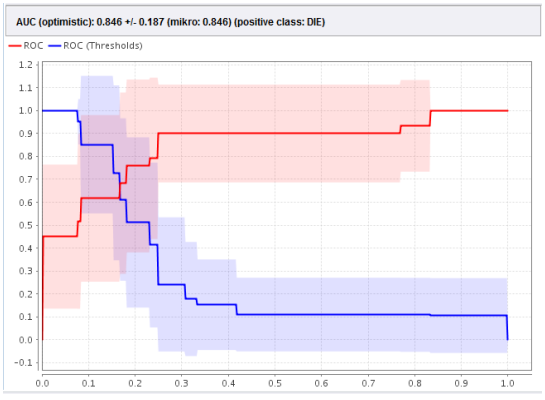
Tabel 3 merupakan hasil perhitungan akurasi data *training* menggunakan Algoritma C4.5. Diketahui tingkat akurasinya 77,29%. Dari 155 data sebanyak 103 data diprediksikan sesuai yaitu 103 data "LIFE" dan 15 data yang diprediksikan "LIFE" tetapi ternyata "DIE". Dan sebanyak 20 data diprediksi "DIE" ternyata termasuk klasifikasi "LIFE" dan sebanyak 17 data diprediksi sesuai yaitu "DIE". Tabel *confusion matrix* disajikan pada tabel 4.7 dan gambar6 adalah grafik AUC (*Area Under Cover*) dari model Algoritma C4.5 yaitu 0,846. Garis horizontal adalah*false* positif dan garis vertikal *false* negatif.

Tabel 3. Tabel *Confusion Matrix* Algoritma C4.5

accuracy: 77.29% +/- 11.33% (mikro: 77.42%)

	true LIFE	true DIE	class precision
pred. LIFE	103	15	87.29%
pred. DIE	20	17	45.95%
class recall	83.74%	53.12%	

Sumber: Hasil Penelitian (2013)



Sumber: Hasil Penelitian (2013)

Gambar 7.Grafik AUC (*Area Under Curve*)

Performance keakurasian AUC (Gorunescu, 2010) dapat diklasifikasikan menjadi lima kelompok yaitu:

1. 0,90 – 1,00 = *Exellent Classification*
2. 0,80 – 0,90 = *Good Classification*

3. 0,70 – 0,80 = *Fair Classification*
4. 0,60 – 0,70 = *Poor Classification*
5. 0,50 – 0,60 = *Failure Classification*

Berdasarkan klasifikasi tersebut maka dapat disimpulkan bawah Algoritma C4.5termasuk dalam *Good Classification* untuk prediksi penyakit hepatitis.

KESIMPULAN

Dari hasil penelitian yang telah dilakukan pada data pasien penderita penyakit hepatitis maka dapat disimpulkan bahwa metode klasifikasi data mining Algoritma C4.5 menghasilkan akurasi 77,29% dan nilai AUC 0,846 yang termasuk dalam *Good Classification*. Dengan demikian dapat disimpulkan bahwa metode ini akurat dalam melakukan prediksi untuk penyakit hepatitis.

Agar penelitian ini bisa ditingkatkan berikut ini adalah saran-saran untuk mendapatkan hasil yang lebih baik:

1. Penelitian ini dapat dikembangkan dengan metode optimasi seperti PSO (*Particle Swarm Optimization*), GA (*Genetic Algorithm*), dan lainnya untuk meningkatkan akuras dari metode.
2. Penelitian ini dapat dikembangkan lagi menggunakan metode klasifikasi lainnya seperti *Naïve Bayes*, *Neural Network*, *KNN*, dan lain-lain.
3. Tidak semua kasus atau permasalahan harus diselesaikan dengan satu algoritma pada *data mining.* Karena belum tentu algoritma yang digunakan merupakan algoritma yang paling akurat.Oleh karena itu untuk menentukkan algoritma yang paling akurat ini perlu dilakukan komparasi beberapa algortima.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada semua pihak terkait, yang telah membantu dalam penelitan dan penulisan artikel ilmiah ini.

DAFTAR PUSTAKA

Bramer, M. (2007). *Principles of Data Mining.* London: Springer.

Eldin, Ahmed. (2011). *A Data Mining Approach for the Prediction of Hepatitis C Virus protease Cleavage Sites.*Cairo : *International Journal of Advanced Computer Science and Applications Vol 2 No.12.*

Gorunescu, Florin. (2011). *Data Mining: Concepts and Techniques*. Verlag berlin Heidelberg: Springer.

Han, J., & Kamber, M. (2007). *Data Mining Concepts and Techniques.* San Fransisco: Mofgan Kaufan Publisher.

Karlik. (2011). *Hepatitis Disease Diagnosis Using Backpropagation and the Naive Bayes Classifiers*. Turkey : *Journal of Science and Technology Vol. 1 No. 1.*

Kumar, Varun & Sharathi, Vijay & Devi, Gayathri (2012). *Hepatitis Prediction Model based on Data Mining Algorithm and Optimal Feature Selection to Improve Predictive Accuracy*. Vellore : *International Journal of*

Computer Applications (0975-8887) Volume 51 - No. 19.

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

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

Liao. (2007). *Recent Advances in Data Mining of Enterprise Data: Algorithms and Application* . Singapore: World Scientific Publishing.

Myatt, Glenn J. (2007). *Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining*. New Jersey: John Wiley & Sons, Inc.

Ozyilmaz, Lale & Yildirim, Tulay. (2003). *Artificial Neural Network for Diagnosis of Hepatitis Disease*.

Riduwan. (2008). *Metode dan Teknik Menyusun Tesis.* Bandung: Alfabeta.

Santosa, B. (2007). *Data Mining Teknik Pemanfaat Data Untuk Keperluan Bisnis.* Yogyakarta: Graha Ilmu.

Shukla, A., Tiwari, R., & Kala, R. (2010). *Real Life Application of Soft Computing.* Taylor and Francis Groups, LLC.

UCI (Universitas California, Invene) *Machine Learning Repository*dengan alamat website http://archive.ics.uci.edu/ml/machine-learning-databases/hepatitis/ Akses : 5 Januari 2013 pukul 10:00

Vercellis, C. (2009). *Business Intelligent: Data Mining and Optimization for Decision Making.* Southern Gate: John Willey & Sons Inc.

Witten, H. I., Eibe, F., & Hall, A. M. (2011). *Data Mining Machine Learning Tools and Techiques.* Burlington: Morgan Kaufmann Publisher.

Wu, X., & Kumar, V. (2009). *The Top Ten Algorithms in Data Mining.* Boca Raton: CRC Press.