KOMPARASI METODE KLASIFIKASI *DATA MINING ALGORITMA C4.5* DAN *NAIVE BAYES* UNTUK PREDIKSI PENYAKIT HEPATITIS

Wisti Dwi Septiani
Program Studi Manajemen Informatika
AMIK BSI Jakarta
Jl. Kramat Raya No. 18 Jakarta Pusat
Wisti.wst@bsi.ac.id

Abstract *— Hepatitis is an 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 and Naïve Bayes is then performed comparative to both methods., The measurement of two method using cross validation, confusion matrix and ROC curve. The result of this research is the best algorithm that can be used to predict disease hepatitis.*

Intisari — 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. Penelitian dalam hal memprediksi penyakit hepatitis telah banyak dilakukan oleh para peneliti terdahulu. Penelitian ini menggunakan metode klasifikasi data mining Algoritma C4.5 dan Naïve Bayes kemudian dilakukan perbandingan kedua metode. Pengukuran dua metode tersebut menggunakan confusion matrix dan kurva ROC. Hasil penelitian ini adalah algoritma terbaik yang dapat digunakan untuk memprediksi penyakit hepatitis.

Kata Kunci*: Hepatitis, Data Mining, Algorithm C4.5, Naïve Bayes*

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).

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	Tahun	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 mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan (Suhartinah, 2010). Klasifikasi Bayes juga dikenal dengan *Naïve Bayes*, memiliki kemampuan sebanding dengan pohon keputusan dan *Neural Network* (Han & Kamber, 2007). Untuk itu dalam penelitian ini akan dilakukan perbandingan metode klasifikasi data mining yaitu Algoritma C4.5 dan Naïve Bayes. Kemudian akan dilakukan komparasi terhadap kedua metode tersebut sehingga didapatkan algoritma terbaik untuk prediksi penyakit hepatitis.

BAHAN DAN METODE

Data Mining

Data mining telah 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:

$$Entropy\ split\ = \sum_{\mathrm{i=1}}^{\mathrm{p}} \binom{\mathrm{n1}}{\mathrm{n}}.\ IE(i)$$

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.

Naïve Bayes

Kata *Naïve*, yang terkesan merendahkan berasal dari asumsi independensi pengaruh nilai suatu atribut dari probalilitas pada kelas yang diberikan terhadap nilai atribut lainnya (Bramer, 2007). Penggunaan teorema Bayes pada algoritma Naïve Bayes yaitu dengan mengkombinasikan prior probability dan probabilitas bersyarat dalam sebuah rumus yang bisa digunakan untuk menghitung probabilitas tiap klasifikasi yang mungkin (Bramer, 2007). Model independence ini menghasilkan pemecahan yang terbaik. Efektifitas metode Naïve Bayes juga terlihat pada contoh dalam Hand dan Yu (2001) dan perbandingan empiris lebih jauh, dengan hasil yang sama, terdapat pada Domingos dan Pazzani (1997) (Wu, 2009). Klasifikasi Bayes didasarkan pada teorema Bayes, diambil dari nama seorang ahli matematika yang juga menteri Prebysterian Inggris, Thomas Bayes (1702-1761), yaitu (Bramer, 2007):

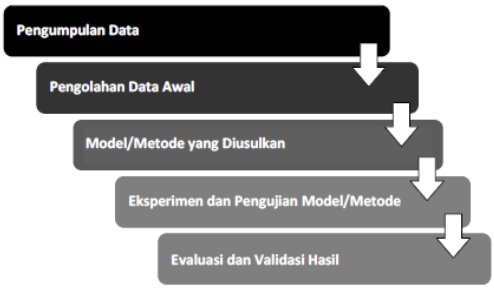
$$P(x|y) = \frac{P(y|x)\,P(x)}{P(y)}$$

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
P(x) = probabilitas hipotesis x
P(y|x) = probabilitas y berdasarkan kondisi pada hipotesis x
P(y) = probabilitas dari y

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 dan *Naïve Bayes* dengan melakukan pengujian terhadap kinerja metode tersebut. Pengujian metode dilakukan dengan cara *confusion matrix* dan kurva ROC serta menggunakan *tools RapidMiner*. Berikut ini adalah kerangka permikiran dari penelitian ini:

Metode Penelitian

Pada penelitian ini data yang digunakan adalah data penyakit hepatitis 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:



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 dan Naive Bayes. Pengujian model menggunakan *Cross Validation*, evaluasi dengan *Confusion Matrix* dan kurva *ROC* sehingga dihasilkan akurasi dari kedua metode tersebut. Lalu akan dilakukan komparasi terhadap dua metode tersebut sehingga didapatkan algoritma yang akurat untuk memprediksi penyakit hepatitis.

HASIL DAN PEMBAHASAN

Eksperimen dan Pengujian Metode Algotirma C4.5

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" sebanyak 123 *record*, dan jumlah kasus yang "DIE" adalah sebanyak 32 *record* total kasus keseluruhan 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:

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

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 split steroid = (79/155 * (0,6145)) +
(76/155 * (0,8314))
= 0,7208

Gain steroid = 0,7346 – 0,7208
= 0,0137

E split malaise = (94/155 * (0,4553)) +
(61/155 * (0,9559))
= 0,6523

Gain malaise = 0,7346 – 0,6523
= 0,0822

E split liver_big = (130/155 * (0,7657)) +
(25/155 * (0,5293))
= 0,7275

Gain liver_big = 0,7346 – 0,7275
= 0,0070

E split spiders = (104/155 * (0,4566)) +
(51/155 * (0,9863))
= 0,6308

Gain spiders = 0,7346 – 0,6308
= 0,1037

E split varices = (137/155 * (0,6180)) +
(18/155 * (0,9640))
= 0,6581

Gain varices = 0,7346 – 0,6581
= 0,0764

E split bilirubin = (105/155 * (0,4220)) +
(50/155 * (0,9953))
= 0,6069 + 0,7333
= 1,3402

Gain bilirubin = 0,7346 – 1,3402
= – 0,6056

E split sgot = (102/155 * (0,6722)) +
(53/155 * (0,8329))
= 0,7271

Gain sgot = 0,7346 – 0,7271
= 0,0074

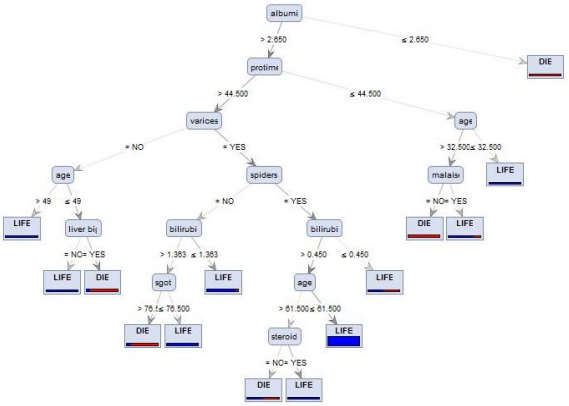
E split albumin = (7/155 * (0)) + (148/155 * (0,6522))
= 0,6227
Gain albumin = 0,7346 – 0,6227
= 0,1119
E split protime = (20/155 * (0,9340)) + (135/155 * (0,5861))
= 0,1205 + 0,5104
= 0,6309
Gain protime = 0,7346 – 0,6309
= 0,1037

Tabel 2. Nilai *entropy* dan *gain* untuk penentuan *root*

Simpul	Jml Kasus	Life	Die	Entropy	Gain
Jumlah kasus	155	123	32	0,7346	
Age					
<= 32,5 thn	40	38	2	0,2863	
> 32,5 thn	115	85	30	0,8280	
<= 49 thn	110	89	21	0,7033	-1,42
> 49 thn	45	34	11	0,8023	
<= 61,5 thn	144	114	30	0,7382	
> 61,5 thn	11	9	2	0,6840	
Steroid					
Yes	79	67	12	0,6145	0,0137
No	76	56	20	0,8314	
Malaise					
Yes	94	85	9	0,4553	0,0822
No	61	38	23	0,9559	
Liver_big					
Yes	130	101	29	0,7657	0,0070
No	25	22	3	0,5293	
Spiders					
Yes	104	94	10	0,4566	0,1037
No	51	29	22	0,9863	
Varices					
Yes	137	116	21	0,6180	0,0764
No	18	7	11	0,9640	
Bilirubin					
<= 1,363	105	96	9	0,4220	
> 1,363	50	27	23	0,9953	-
<= 0,450	3	2	1	0,9182	0,6056
> 0,450	152	121	31	0,7297	
Sgot					
<= 76,500	102	84	18	0,6722	0,0074
> 76,500	53	39	14	0,8329	
Albumin					
<= 2,650	7	0	7	0	**0,1119**
> 2,650	148	123	25	0,6552	
Protime					
<= 44,500	20	13	7	0,9340	0,1037
> 44,500	135	116	19	0,5861	

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 0,2092 sehingga atribut tersebut dijadikan simpul 1.1. Untuk menetukan simpul selanjutnya, dilakukan perhitungan nilai *entropy* dan *gain* dengan cara yang sama, sehingga diperoleh pohon keputusan seperti gambar di bawah ini:

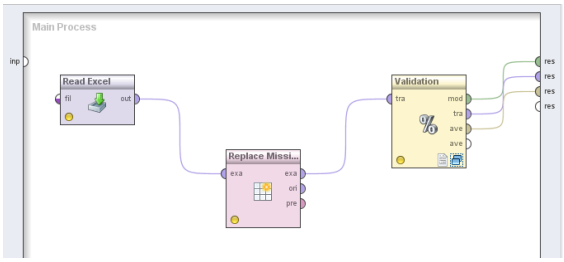


Gambar 4. Pohon keputusan hasil Algoritma C4.5
Sumber: Hasil Olahan Data, 2015

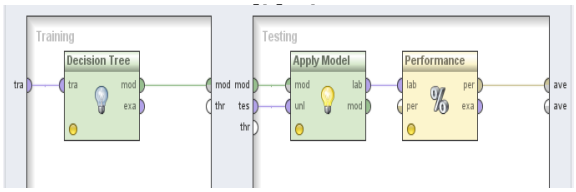
Dari pohon keputusan pada gambar 4 didapatkan *rule* untuk memprediksi penyakit hepatitis. *Rule* yang didapat sebagai berikut :
R1: Jika albumin <= 2,650 maka pasien "DIE".
R2: Jika albumin > 2,650 dan protime > 44,500 dan varices = NO dan age > 49 tahun maka pasien "LIFE".
R3: Jika albumin > 2,650 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 C4.5 ini menggunakan aplikasi RapidMiner seperti berikut:



Gambar 5a. Pengujian 10-*Fold Cross*



Gambar 5b. *Validation* Model Algoritma C4.5

Eksperimen dan Pengujian Metode Naïve Bayes

Naïve Bayes adalah model kedua yang akan dihitung. Langkah-langkah yang akan dilakukan adalah menghitung nilai probabilitas prior, yaitu probabilitas nilai "LIFE" dan "DIE" masing-masing atribut terhadap total kasus "LIFE" dan "DIE" dari seluruh data.

Tabel 3. Perhitungan nilai probabilitas prior

Atribut	Kasus	Life	Die	P(X|Ci)	
				Life	Die
Total kasus	155	123	32	0,793548387	0,206451613
Age <= 32,5 tahun	40	38	2	0,95	0,05
Age > 32,5 tahun	115	85	30	0,739130435	0,260869565
Age <= 49 tahun	110	89	21	0,809090909	0,190909091
Age > 49 tahun	45	34	11	0,755555556	0,244444444
Age <= 61,5 tahun	144	114	30	0,791666667	0,208333333
Age > 61,5 tahun	11	9	2	0,818181818	0,181818182
Steroid = YES	79	67	12	0,848101266	0,151898734
Steroid = NO	76	56	20	0,736842105	0,263157895
Malaise = YES	94	85	9	0,904255319	0,095744681
Malaise = NO	61	38	23	0,62295082	0,37704918
Liver_big = YES	130	101	29	0,776923077	0,223076923
Liver_big = NO	25	22	3	0,88	0,12
Spiders = YES	104	94	10	0,903846154	0,096153846
Spiders = NO	51	29	22	0,568627451	0,431372549
Varices = YES	137	116	21	0,846715328	0,153284672
Varices = NO	18	7	11	0,388888889	0,611111111
Bilirubin <= 1,363	105	96	9	0,914285714	0,085714286
Bilirubin > 1,363	50	27	23	0,54	0,46
Bilirubin <= 0,450	3	2	1	0,666666667	0,333333333
Bilirubin > 0,450	152	121	31	0,796052632	0,203947368
Sgot <= 76,500	102	84	18	0,823529412	0,176470588
Sgot > 76,500	53	39	14	0,735849057	0,264150943
Albumin <= 2,650	7	0	7	0	1
Albumin > 2,650	148	123	25	0,831081081	0,168918919
Protime <= 44,500	20	13	7	0,65	0,35
Protime > 44,500	135	116	19	0,859259259	0,140740741

Untuk menentukan kelas dari kasus baru maka dilakukan perhitungan probabilitas posterior berdasarkan probabilitas prior yang telah dihitung sebelumnya dan telah disajikan pada tabel 3. Perhitungan probabilitas posterior untuk menentukan data testing termasuk klasifikasi yang mana, sebagai contoh diambil kasus seperti tabel 4 berikut, dimana X tersebut adalah data yang akan diprediksi hasilnya.

Tabel 4. Perhitungan nilai probabilitas prior

Data X untuk kasus terbaru		P(X|Ci)	
Atribut	Nilai	Life	Die
Age	<= 49 tahun	0,809090909	0,190909091
Steroid	NO	0,736842105	0,263157895
Malaise	YES	0,904255319	0,095744681
Liver_big	YES	0,776923077	0,223076923
Spiders	NO	0,568627451	0,431372549
Varices	YES	0,846715328	0,153284672
Bilirubin	<= 44,500	0,666666667	0,333333333
Sgot	<= 76,500	0,823529412	0,176470588
Albumin	> 2,650	0,831081081	0,168918919
Protime	<= 44,500	0,65	0,35

Berdasarkan nilai probabilitas prior masing-masing atribut yang telah dihitung pada tabel 4 maka dapat dilihat rule yang diperoleh untuk atribut di atas seperti berikut ini :

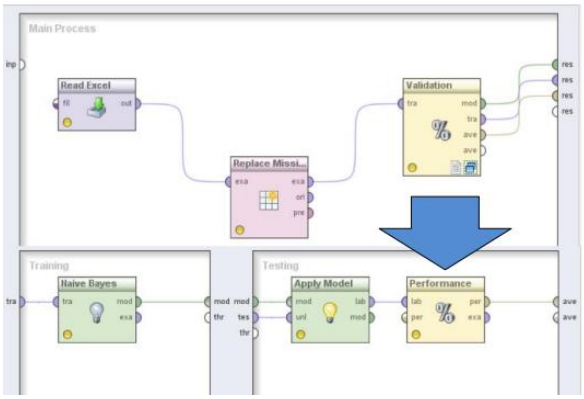
1. Hitung probabilitas "LIFE" untuk setiap atribut
P(LIFE)P(Age<=49|LIFE)P(Steroid=NO|LIFE)P(Malaise=YES|LIFE)P(Liver_big=YES|LIFE)P(Spiders=NO|LIFE)P(Varices=YES|LIFE)P(Bilirubin<=44,500|LIFE)P(Sgot<=76,500|LIFE)P(Albumin>2,650|LIFE)P(Protime<=44,500|LIFE)=0,793548387*0,809090909*0,736842105*0,904255319*0,776923077*0,568627451*0,846715328*0,666666667*0,823529412*0,831081081*0,65 =0,047459605

2. Hitung probabilitas "DIE" untuk setiap atribut
P(DIE)P(Age<=49|DIE)P(Steroid=NO|DIE)P(Malaise=YES|DIE)P(Liver_big=YES|DIE)P(Spiders=NO|DIE)P(Varices=YES|DIE)P(Bilirubin<=44,500|DIE)P(Sgot<=76,500|DIE)P(Albumin>2,650|DIE)P(Protime<=44,500|DIE)=0,206451613*0,190909091*0,263157895*0,095744681*0,223076923*0,431372549*0,153284672*0,333333333*0,176470588*0,168918919*0,35=5,09424E-08

3. Bandingkan hasil dari probabilitas "LIFE" dan "DIE"
Probabilitas "LIFE" = 0,047459605
Probabilitas "DIE" = 5,09424E-08

Dikarenakan 0,047459605 > 5,09424E-08, maka dapat disimpulkan bahwa data testing tersebut termasuk klasifikasi "LIFE".
Rule1: Jika probabilitas "LIFE" lebih besar dari probabilitas "DIE" maka hasil adalah "LIFE"
Rule2: Jika probabilitas "DIE" lebih besar dari probabilitas "LIFE" maka hasil adalah "DIE".

Pengujian dengan 10-*Fold Cross Validation* untuk model *Naïve Bayes* ini menggunakan aplikasi RapidMiner seperti berikut:



Gambar 6. Pengujian 10-*Fold Cross Validation* Model *Naïve Bayes*

Evaluasi dan Validasi Hasil

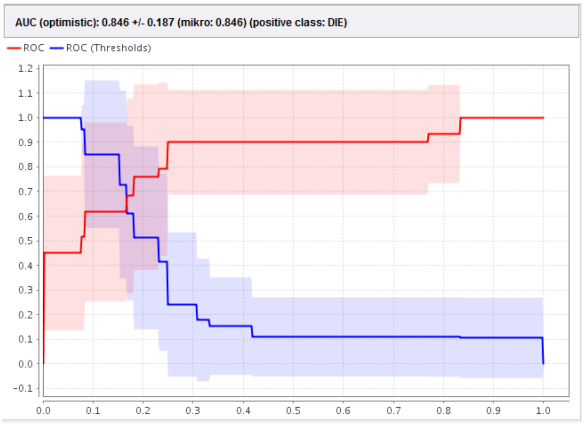
Setelah data diolah maka dapat diuji tingkat akurasinya untuk melihat kinerja dari metode Algoritma C4.5. Penelitian ini bertujuan untuk melihat akurasi analis data pasien penderita penyakit hepatitis, menilai kemungkinan kelangsungan hidup penderita apakah hidup atau mati. Pengujian tingkat akurasi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

Tabel 5 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 5 dan gambar 7 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 5. 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%	



Gambar 7. Grafik AUC (*Area Under Curve*) Algoritma C4.5

Tabel 6 merupakan hasil perhitungan akurasi data training menggunakan Naïve Bayes. Diketahui tingkat akurasinya 83,71%. Dari 155 data sebanyak 106 data diprediksikan sesuai yaitu 106 data "LIFE" dan 8 data yang diprediksikan "LIFE" tetapi ternyata "DIE". Dan sebanyak 17 data diprediksi "DIE" ternyata termasuk klasifikasi "LIFE" dan sebanyak 24 data diprediksi sesuai yaitu "DIE". Tabel confusion matrix disajikan pada tabel 6 dan gambar 8 adalah grafik AUC (*Area Under Cover*) dari model *Naïve Bayes*, garis horizontal adalah false positif dan garis vertikal false negatif.

Tabel 6. Tabel *Confusion Matrix Naïve Bayes*

accuracy: 83.71% +/- 8.89% (mikro: 83.87%)

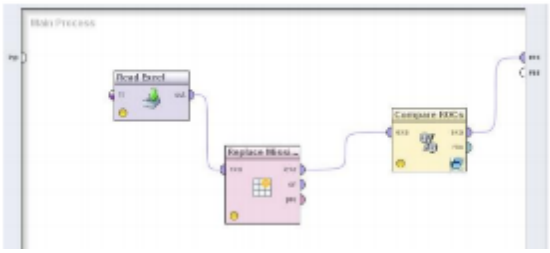
	true LIFE	true DIE	class precision
pred. LIFE	106	8	92.98%
pred. DIE	17	24	58.54%
class recall	86.18%	75.00%	

Gambar 8. Grafik AUC (*Area Under Curve*) *Naïve Bayes*

Analisis dan Evaluasi Komparasi Model

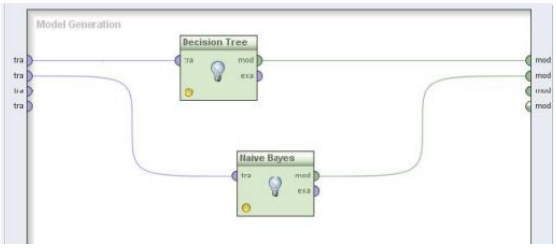
Berikut ini adalah pengujian performance dengan menggunakan *Confusion Matrix* dan *ROC Curve*. Model evaluasi komparasi dengan

menggunakan *ROC Curve* secara visual pada *framework* RapidMiner seperti berikut ini:



Gambar 9. Desain Model Komparasi dengan *ROC Curve*

Pada gambar 9, dalam modul *Compare ROC's* berisi beberapa model seperti berikut:



Gambar 10. Algoritma-algoritma dalam Modul ROC's

Berdasarkan dari analisa pengujian masing-masing algoritma di atas maka dapat dirangkumkan hasilnya sebagai berikut:

Tabel 7. Perbandingan Performance Metode

	C4.5	Naive Bayes
Accuracy	77,29%	83,71%
AUC	0,846	0,812

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 bahwa Algoritma C4.5 dan *Naive Bayes* termasuk algoritma yang akurat untuk memprediksi penyakit hepatitis karena nilai AUC termasuk dalam predikat *Good Classification* (0,80–0,90).

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*. Naive Bayes menghasilkan akurasi 83,71% dan nilai AUC 0,812. Dengan demikian dapat disimpulkan bahwa kedua metode ini akurat dalam melakukan prediksi untuk penyakit hepatitis.

Melihat dari hasil perbandingan kedua algoritma tersebut memang dapat dinyatakan bahwa Algoritma C4.5 lebih unggul dari Naive Bayes karena memiliki nilai AUC 0,846 dengan kategori *Good Clasification*.

Akan tetapi jika ditelusuri lebih lanjut ternyata masih belum bisa dinyatakan sebagai algoritma yang lebih unggul. Menurut pengujian berdasarkan *Accuracy*, algoritma terbaik adalah *Naive Bayes*. Sedangkan menurut pengujian berdasarkan ROC Curve (AUC) algoritma yang terbaik adalah Algoritma C4.5. Agar penelitian ini bisa ditingkatkan berikut ini adalah saran-saran untuk mendapatkan hasil yang lebih baik:

1. Penelitian ini dapat dikembangkan lebih lanjut dengan melakukan uji statistik dengan menggunakan uji T-Test dengan membandingkan kedua algoritma untuk melihat algoritma mana yang lebih dominan atau signifikan berdasarkan nilai probabilitas.
2. Penelitian ini dapat dikembangkan dengan metode optimasi seperti PSO (*Particle Swarm Optimization*), GA (*Genetic Algorithm*), dan lainnya untuk meningkatkan akuras dari metode.
3. Penelitian ini dapat dikembangkan lagi dengan membandingkan dengan metode lainnya seperti *Neural Network, SVM, KNN*, dan lain-lain.
4. 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.

REFERENSI

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.

BIODATA PENULIS

Wisti Dwi Septiani, M.Kom. Lahir di Jakarta, 17 September 1986. Penulis adalah Staff Pengajar di AMIK BSI Jakarta sejak tahun 2008-sekarang. Penulis menyelesaikan Studi Strata I (S1) di Kampus STMIK PGRI Tangerang dengan Jurusan Sistem Informasi dengan gelar S.Kom dan menyelesaikan Studi Strata 2 di Pascasarjana STMIK Nusa Mandiri Jakarta jurusan Ilmu Komputer dengan gelar M.Kom. Selain mengajar, penulis juga sudah pernah membuat artikel ilmiah sebelumnya dan diterbitkan di Jurnal Techno Vol. XI No. 1 Maret 2014 dengan judul Penerapan Algoritma C4.5 Untuk prediksi Penyakit Hepatitis.