

ANALISA KOMPARASI ALGORITMA NAIVE BAYES DAN C4.5 UNTUK PREDIKSI PENYAKIT LIVER

Eva Rahmawati

Program Studi Teknik Informatika, STMIK Nusa Mandiri

Jl. Kramat Raya No. 18, Jakarta Pusat

Eva.rijal@gmail.com

ABSTRACT

Liver disease is one of the deadliest diseases in the world. Several studies have been conducted to diagnose patients properly but still unknown what method was accurate in predicting liver disease. Data mining is the science that uses past data as a reference to get a new knowledge. One of the data mining algorithm is a classification algorithm. Data are obtained from the UCI which consists of 583 records with 11 fields. In this research, comparative Naïve Bayes and C4.5 algorithms using software algorithms KNAME to know which are the most accurate in predicting liver disease. The results of the second test is known that the algorithm C4.5 algorithm has the highest accuracy value is 72.845% while the Naïve Bayes algorithm has a value of 63.362% accuracy. Thus C4.5 algorithm can more accurately predict liver disease.

Kata kunci:C4.5 , Naive Bayes, Liver

PENDAHULUAN

Berdasarkan data *World Health Organization* (WHO), virus hepatitis B kronis diperkirakan menyerang 350 juta orang didunia, terutama Asia Tenggara dan Afrika dan menyebabkan kematian 1,2 Juta orang pertahun. Dari jumlah itu 15-25% yang terinfeksi kronis meninggal dunia karena komplikasi dari sirosis dan kanker hati. Hati sebagai organ yang memiliki tugas utama sebagai penetral racun ditubuh menjadikan racun-racun yang selama ini masuk melalui tubuh kita dari makanan atau lingkungan mampu dinetralisir oleh hati. Salah satu penyakit yang meyerang hati adalah hepatitis atau Liver. Penyakit Liver merupakan peradangan hati yang disebabkan oleh infeksi virus, bakteri atau bahan-bahan beracun sehingga hati tidak dapat melakukan fungsinya dengan baik. Dalam bidang kesehatan, kesalahan dalam mendiagnosa penyakit yang dialami pasien adalah tanggung jawab yang paling berat

untuk diemban oleh ahli kesehatan. Kesalahan dalam mendiagnosa penyakit dapat menyebabkan hal yang membahayakan bagi kesehatan pasien bahkan dapat menyebabkan kematian (Neshat dkk, 2012).

Penggunaan *data mining* dengan model *Naive Bayes* dan *Algoritma C4.5* dalam mendiagnosa penyakit hati dapat menjadi alternatif pilihan yang tepat. Namun sampai saat ini belum diketahui algoritma yang paling akurat dalam penentuan diagnosa untuk prediksi penyakit ini. Untuk itu maka dalam penelitian ini akan dilakukan komparasi metode algoritma Naive bayes dan Algoritma C4.5 untuk mengetahui algoritma yang memiliki akurasi lebih tinggi dalam mendeteksi penyakit hati.

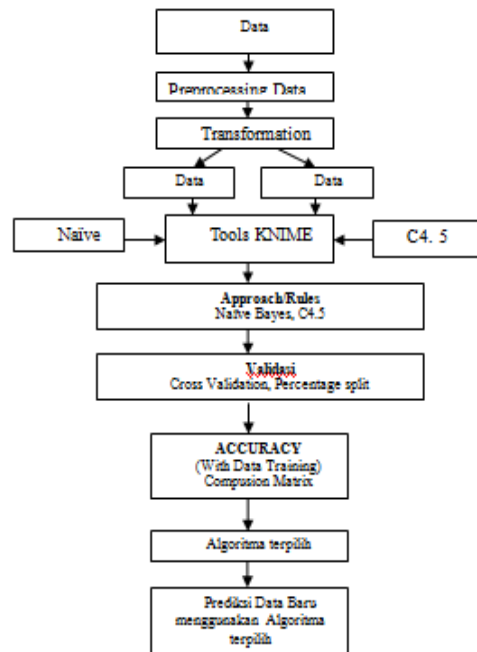
Naive Bayes merupakan salah satu metode pengklasifikasian berpeluang sederhana yang berdasarkan pada penerapan Teorama Bayes dengan asumsi antar variabel penjelas saling bebas (independen) (Han, Kember, 2006).

Konsep dari algoritma C4.5 adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan (*rule*). Algoritma C4.5 memetakan nilai atribut menjadi class yang dapat diterapkan untuk klasifikasi baru (Wu, Kumar, 2009). Untuk menerapkan algoritma Naive Bayes dan C4.5 ini digunakan perangkat lunak Kname. Data yang digunakan dalam penelitian ini bersumber dari alamat web: <http://archive.ics.uci.edu/ml/>. Data yang diteliti merupakan hasil pemeriksaan terhadap 583 orang dari wilayah Andhra Pradesh, India dengan 11 *field*. Hasil dari penerapan model *Naive Bayes dan C4.5* ini kemudian akan dikomparasi tingkat akurasi menggunakan metode Confusion Matrix untuk mengetahui algoritma yang memiliki tingkat akurasi yang paling tinggi, sehingga tujuan Pengidentifikasi penyakit Liver dengan Metode algoritma Naive Bayes dan C4.5 dapat tercapai.

BAHAN DAN METODE

Jenis penelitian yang dilakukan dalam penelitian ini adalah jenis penelitian eksperimen. Jenis penelitian eksperimen dibagi dua, yaitu eksperimen absolut dan eksperimen komparatif. Eksperimen absolut mengarah kepada dampak yang dihasilkan dari eksperimen, sedangkan eksperimen komparatif yaitu membandingkan dua objek yang berbeda, misalnya membandingkan dua algoritma yang berbeda dengan melihat hasil statistik masing-masing mana yang lebih baik (Depkes, 2009). Pada penelitian ini, jenis penelitian yang diambil adalah Eksperimen komparatif. Penelitian eksperimen komparatif ini dilandasi oleh kerangka pemikiran pemecahan

masalah seperti terlihat pada gambar 1:



Sumber: Data penelitian(2015)

Gambar 1
kerangka pemikiran pemecahan masalah

1. Pengumpulan Data

Teknik pengumpulan data ialah teknik atau cara-cara yang dapat digunakan untuk menggunakan data. Dalam pengumpulan data terdapat sumber data, sumber data yang terhimpun langsung oleh peneliti disebut dengan sumber primer, sedangkan apabila melalui tangan kedua disebut sumber sekunder (Riduan,2008). Data yang diperoleh adalah data sekunder karena diperoleh dari Data yang digunakan dalam penelitian ini bersumber dari alamat web: <http://archive.ics.uci.edu/ml/>. Data ini merupakan hasil pemeriksaan terhadap 583 orang dari wilayah Andhra Pradesh, India yang diperiksa dengan hasil 416 orang terdeteksi menderita penyakit hati dan 167 orang tidak terdeteksi menderita penyakit hati. Sumber data terdiri dari 441 orang berjenis

kelamin laki-laki dan 142 orang berjenis kelamin perempuan.

Variabel yang terdapat pada data pasien liver tersebut sebagai berikut:

- a. Age (usia)
- b. Gender (Jenis kelamin)
- c. Total Bilirubin (Bilirubin Total)
- d. Direct Bilirubin (Bilirubin Langsung)
- e. Alkaline Phosphatase (ALP)
- f. Serum Glutamic Pyruvic Transaminase (SGPT) / Alanin Aminotransferase (ALT)
- g. Serum Glutamic Oxaloacetic Transaminase (SGOT)/ Aspartate Aminotransferase (AST)
- h. Total Protein (Protein Total)
- i. Albumin
- j. Albumin-Globulin Ratio (A/G Rasio)
- k. Liver Patient (yes/no)

Berdasarkan data yang diperoleh, tidak perlu lagi dilakukan *data integration*, hal ini disebabkan karena tempat penyimpanan yang digunakan hanya bersumber dari satu tempat penyimpanan saja, sehingga tidak diperlukan adanya proses penyatuan tempat penyimpanan. Sedangkan proses *data cleaning* dan *data reduction* perlu dilakukan. Hal ini disebabkan karena dalam data yang diperoleh masih terdapat kemungkinan adanya data yang bernilai kosong, tidak konsisten atau mungkin tupel yang kosong (*missing values* dan *noisy*) serta adanya kemungkinan terjadi duplikasi atau terdapat tupel yang sama. Data pasien penyakit liver bisa di lihat pada tabel 1 berikut:

Tabel 1. Data pasien Penyakit liver

Age	Gender	TB	DB	AP	SGPT	SGOT	TP	Albumin	A/G Ratio	Liver Patient
65	Female	0.7	0.1	187	16	13	6.8	3.3	0.9	1
62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
62	Male	7.3	4.1	490	60	63	7	3.3	0.89	1
58	Male	1	0.4	182	14	20	6.8	3.4	1	1
72	Male	3.9	2	195	27	59	7.3	2.4	0.4	1
46	Male	1.8	0.7	208	19	14	7.6	4.4	1.3	1
26	Female	0.9	0.2	154	16	12	7	3.5	1	1
29	Female	0.9	0.3	202	14	11	6.7	3.6	1.1	1

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

1. Data Mining

Data Mining merupakan teknologi baru yang sangat berguna untuk membantu perusahaan-perusahaan menemukan informasi yang sangat penting dari gudang data mereka. *Data mining* adalah perpaduan dari ilmu statistik, kecerdasan buatan, dan penelitian bidang *database* (Han, 2006). Data mining didefinisikan sebagai proses tentang memecahkan masalah dengan menganalisis data yang berada dalam database (Witten, 2011).

Nama *data mining* berasal dari kemiripan antara pencarian informasi yang bernilai dari *database* yang besar dengan menambang sebuah gunung untuk sesuatu yang bernilai (Sumathi, Sivanandam, 2006).

Data mining didefinisikan sebagai proses pengenalan pola dalam data. Data mining, sering disebut *knowledge discovery in database* (KDD), adalah suatu kegiatan yang meliputi pengumpulan, pemakaian data

historis untuk menentukan keteraturan, pola atau hubungan dalam set data berukuran besar (Witten, 2011). Keluaran dari data mining ini bias dipakai untuk memperbaiki

gambilan keputusan di masa depan (Santoso, 2007).

2. Algoritma 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, Max, 2007).

Klasifikasi Naive Bayes adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class. Klasifikasi bayesian didasarkan pada teorema Bayes, diambil dari nama seorang ahli matematika yang juga menteri Prebysterian Inggris, Thomas Bayes (1702-1761) (Larose, 2005). Klasifikasi bayesian memiliki kemampuan klasifikasi serupa dengan decision tree dan neural network (Maimon, Oded and Rokach, Lior, 2010).

$$P(x|y) = \frac{P(y|x) P(x)}{P(y)} \quad (\text{Persamaan 1})$$

keterangan :

y = data dengan kelas yang belum diketahui

x = hipotesis data y merupakan suatu kelas spesifik

$P(x|y)$ = probabilitas hipotesis x berdasar 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

3. Decision Tree (C4.5)

Algoritma C4.5 merupakan bagian dari kelompok algoritma decision trees dan merupakan kategori 10 algoritma yang paling populer (Han, 2006). Algoritma C4.5 diperkenalkan oleh J. Ross Quinlan seorang peneliti dibidang mesin pembelajaran yang merupakan perkembangan dari algoritma ID3 (*Iterative Dichotomiser*), algoritma tersebut digunakan untuk membentuk 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 (Maimon, Oded and Rokach, Lior, 2010).

Salah satu metode data mining yang umum digunakan adalah pohon keputusan. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan rule. Pohon keputusan adalah salah satu metode klasifikasi yang paling populer karena mudah untuk diinterpretasi oleh manusia. Konsep dari pohon keputusan adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan. Untuk mengklasifikasikan obyek diperlukan urutan pertanyaan sebelum dapat dibuat kelompoknya. Jawaban dari setiap pertanyaan akan mempengaruhi pertanyaan berikutnya dan selanjutnya. Dalam

pohon keputusan (*decision tree*) pertanyaan-pertanyaan pertama akan ditanyakan pada simpul akar. Jawaban dari pertanyaan ini dikemukakan dalam cabang-cabang. Jawaban dalam cabang akan disusul dengan pertanyaan kedua lewat simpul berikutnya. Langkah ini akan berakhir disuatu simpul jika sudah jelas kelas atau obyek yang kita cari. Pada dasarnya konsep dari algoritma C4.5 adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan (*rule*). C4.5 adalah algoritma yang cocok untuk masalah klasifikasi dan data mining, C4.5 memetakan nilai atribut menjadi class yang dapat diterapkan untuk klasifikasi baru (Wu, 2009). Seperti persamaan berikut:

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

(persamaan 2)

Dimana:

S : Himpunan kasus

A : Atribut

n : Jumlah partisi atribut A

|S_i| : Jumlah kasus pada partisi ke-i

|S| : Jumlah kasus dalam S

Nilai entropi dapat dihitung dengan cara berikut:

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

(persamaan 3)

Dimana:

S : Himpunan kasus

n : Jumlah partisi S

P_i : Proporsi dari S_i terhadap S

Buat cabang untuk tiap-tiap nilai, Bagi kasus dalam cabang

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

HASIL DAN PEMBAHASAN

Data Training yang digunakan adalah sebanyak 583 dari 416 (empat ratus enam belas) pasien positif penyakit liver dan 167 (seratus enam

puluh tujuh) Tetapi dalam data tersebut masih mengandung duplikasi dan anomali atau inkonsisten data maka dengan ini dilakukan *replace missing*. Sehingga Data *training* yang digunakan pada model *Naive Bayes* adalah data yang sama seperti yang digunakan pada model pohon keputusan C4.5. Dari data yang diperoleh sebanyak 579 record, 414 record menunjukkan data positif terkena penyakit liver sedangkan sebanyak 165 record menunjukkan negative terkena liver.

1. Dengan algoritma Naive Bayes

Dalam membuat model algoritma *Naive Bayes* terlebih dahulu kita mencari *probabilitas* hipotesis untuk masing-masing class P(H). Hipotesis yang ada yaitu pasien terkena penyakit LIVER (positif) dan pasien tidak terkena penyakit Liver (negative) dengan menggunakan persamaan 1 sebagai berikut:

$$P(\text{Terkena Liver}) = 414 : 579 \\ = 0.715025$$

$$P(\text{Tidak Terkena liver}) = 165 : 579 \\ = 0.28497$$

Hasil perhitungan *probabilitas prior* dengan menggunakan *Naive Bayes* dapat dilihat pada tabel dibawah ini.

Tabel 2. Probabilitas Prior

ATRIBUT	Kasus (S)	Jml	Positif	Negatif	P (X C)	
		(S _i)	(S _i)	(S _i)	Positif	Negatif
TOTAL	579	414	165	0.715026	0.284974	
USIA	<=14	16	8	8	0.019324	0.048485
	15-49	334	232	102	0.560386	0.618182
	>=50	229	174	55	0.420293	0.333333
Jenis	Laki-Laki	439	323	116	0.780193	0.70303
		439	323	116	0.780193	0.70303
Kelamin	Perempuan	140	91	49	0.219807	0.29697
		140	91	49	0.219807	0.29697
TB	<=1	303	183	120	0.442029	0.727273
		303	183	120	0.442029	0.727273

	1 sd 2	105	73	32	0.17632 9	0.19393 9
	2 sd 3	51	43	8	0.10386 5	0.04848 5
	3 sd 9	68	63	5	0.15217 4	0.03030 3
	>=9	52	52	0	0.12560 4	0
	<=1	414	259	155	0.62560 4	0.93939 4
DB	1 sd 2	63	58	5	0.14009 7	0.03030 3
	2 sd 5	54	49	5	0.11835 7	0.03030 3
	>5	48	48	0	0.11594 2	0
	<=140	35	23	12	0.05555 6	0.07272 7
ALP	141 sd 210	261	151	110	0.36473 4	0.66666 7
	211 sd 280	106	87	19	0.21014 5	0.11515 2
	281 sd 420	99	83	16	0.20048 3	0.09697
	>=420	78	70	8	0.16908 2	0.04848 5
	<=40	332	204	128	0.49275 4	0.77575 8
SGPT	41 sd 80	141	113	28	0.27294 7	0.16969 7
	81 sd 120	38	32	6	0.07729 5	0.03636 4
	121 sd 200	33	30	3	0.07246 4	0.01818 2
	>200	35	35	0	0.08454 1	0
	<=41	288	172	116	0.41545 9	0.70303
SGOT	42 sd 82	137	104	33	0.25120 8	0.2
	83 sd 164	82	69	13	0.16666 7	0.07878 8
	>164	72	69	3	0.16666 7	0.01818 2
	>=6	408	288	120	0.69565 2	0.72727 3
TP	5 sd 6	128	95	33	0.22946 9	0.2
	<5	43	31	12	0.07487 9	0.07272 7
	4 sd 6	209	128	81	0.30917 9	0.49090 9
Albumin	3 sd 4	255	195	60	0.47101 4	0.36363 6
	2 sd 3	109	87	22	0.21014 5	0.13333 3
	0 sd 2	6	4	2	0.00966 2	0.01212 1
	>=3	3	3	0	0.00724 6	0
AG Ratio	2 sd 3	21	14	7	0.03381 6	0.04242 4
	<2	555	397	158	0.95893 7	0.95757 6

Pada probabilitas *prior* terdapat dua *class* yang dibentuk, yaitu:

class diagnosa = Liver
class diagnosa = Sehat atau tidak terkena penyakit Liver
Probailitas prior digunakan untuk menentukan *class* pada kasus baru yang terlebih dahulu dihitung *probabilitas posteriornya*. Jika ada kasus baru seperti yang terlihat pada tabel berikut:

Tabel 3. Probabilitas Posterior

Data X	P (X Ci)			
	Atribut	Nilai	Positif	Negatif
USIA	34	0.560386	0.618182	
Jenis Kelamin	Male	0.219807	0.29697	
total Bilirubbin	4	0.152174	0.030303	
Direct Bilirubbin	2	0.140097	0.030303	
ALP	4	0.200483	0.09697	
SGPT	5	0.084541	0	
SGOT	4	0.166667	0.018182	
Total Protein	2	0.229469	0.2	
Albumin	2	0.471014	0.363636	
AG Ratio	3	0.210145	0.133333	

Setelah diketahui *probabilitas* setiap atribut terhadap *probabilitas* tiap *class* atau $P(X|Ci)$, maka langkah selanjutnya adalah menghitung total keseluruhan probabilitas tiap *class*
 $P(X|diagnosa = Liver)$

$$= 0.560386 \times 0.219807 \times 0.152174 \times 0.140097 \times 0.200483 \times 0.084541 \times 0.166667 \times 0.229469 \times 0.471014 \times 0.210145$$

$$= 1.68488E-07$$

$P(X|diagnosa =Negatif)$

$$= 0.618182 \times 0.29697 \times 0.030303 \times 0.030303 \times 0.09697 \times 0 \times 0.018182 \times 0.2 \times 0.363636 \times 0.133333$$

$$= 0$$

$$P (X|diagnosa = Liver) P (Liver) = 1.68488E-07 \times 0.715026 = 1.20473E-07$$

$$P(X|\text{diagnose} = \text{Negatif}) P(\text{Negatif}) = 0 \times 0.284974 = 0$$

Dari hasil perhitungan tersebut diketahui nilai $P(X|\text{Liver})$ lebih besar dari pada nilai $P(X|\text{Negatif})$, sehingga dapat disimpulkan bahwa untuk kasus tersebut masuk kedalam klasifikasi Positif Liver.

Model yang telah dibentuk diuji tingkat akurasi dengan memasukkan data uji yang berasal dari data training dengan menggunakan metode cross validation dan split percentage untuk menguji tingkat akurasi. Dengan Split percentage dengan menggunakan 60:40 maka diperoleh data training sebanyak 347 dan data testing 232 maka didapat hasil akurasi dengan menggunakan perhitungan akurasi maka akan diperoleh nilai akurasi sebanyak 63.362% dan error 36.638%. Seperti tertera pada tabel 4.

Tabel 4 Akurasi Naïve Bayes

Class\Prediksi	Positif	Negative
Positif	93	76
Negative	9	54

Untuk pengujian menggunakan metode *cross validation* sebanyak 7,8,10 kali pengujian maka akan didapat seperti tertera pada tabel berikut:

Tabel 5. Pengujian menggunakan cross 7

Naïve Bayes			
row ID	Error in %	Size of Test Set	Error count
fold 0	43.373	83	3
fold 1	38.554	83	6
fold 2	43.373	83	3
fold 3	35.366	82	6
fold 4	38.554	83	9

fold 5	32.53	83	2
Fold 6	35.366	82	7

Didalam pengujian menggunakan Cross 7 data yang digunakan untuk testing 82 data secara acak maka akan didapatkan hasil terbaik pada fold ke- 5 dengan error sebanyak 32.53% untuk lebih jelasnya dapat dilihat pada tabel 5.

Tabel 6. Pengujian menggunakan cross 8

Naïve Bayes			
row ID	Error in %	Size of Test Set	Error count
fold 0	46.575	73	34
fold 1	37.5	72	27
fold 2	30.137	73	22
fold 3	45.833	72	33
fold 4	31.944	72	23
fold 5	32.877	73	24
Fold 6	37.5	72	27
Fold 7	37.5	72	27

Didalam pengujian menggunakan Cross 8 data yang digunakan untuk testing 72 data secara acak maka akan didapatkan hasil terbaik pada fold ke- 2 dengan error sebanyak 30.137% untuk lebih jelasnya dapat dilihat pada tabel 6.

Tabel 7. Pengujian menggunakan cross 10

Naïve Bayes			
row ID	Error in %	Size of Test Set	Error Count
fold 0	39.655	58	2
fold 1	43.103	58	3
fold 2	37.931	58	2
fold 3	37.931	58	5
fold 4	39.655	58	2
fold 5	37.931	58	3
Fold 6	32.759	58	2
Fold 7	39.655	58	9

			3
			2
Fold8	41.379	58	4
			1
Fold9	29.825	57	7

Didalam pengujian menggunakan Cross 10 data yang digunakan untuk testing 57 data secara acak maka akan didapatkan hasil terbaik pada fold ke- 6 dengan error sebanyak 32.759 % untuk lebih jelasnya dapat dilihat pada tabel 7.

2. Dengan Algoritma C4.5

Dalam membuat pohon keputusan lebih dahulu kita hitung jumlah class yang terkena penyakit Liver dan tidak serta nilai entropy dengan menggunakan persamaan 3 dari masing-masing class berdasarkan atribut yang telah ditentukan dengan menggunakan data training yang dihitung setelah mendapatkan nilai entropy dari masing-masing class maka didapat nilai gain dengan menggunakan persamaan 2 tiap atribut Hasil perhitungan dapat dilihat pada tabel 8 berikut:

Tabel 8. perhitungan Gain Node 1

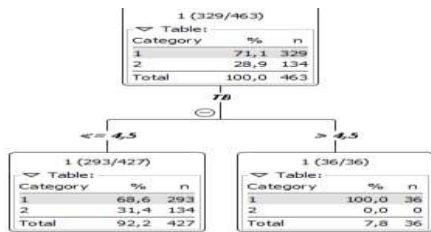
ATRIBUT	Jml Kasus (S)	Positif (Si)	Negatif (Si)	Entropy	Gain
TOTAL	579	414	165	0.25953	0.002355
USIA	<=14	16	8	0.30103	
	15-49	334	232	0.267248	
	>=50	229	174	0.239418	
Jenis Kelamin	Laki-Laki	439	323	0.25078	
	Perempuan	140	91	0.281183	
total Bilirubin	<=1	303	183	0.291574	
	1 sd 2	105	73	0.267024	
	2 sd 3	51	43	0.188671	
	3 sd 9	68	63	0.114078	
	>=9	52	52	0	
Direct Bilirubin					0.028595

bin	<=1	414	259	155	0.287179
	1 sd 2	63	58	5	0.120393
	2 sd 5	54	49	5	0.133978
	>5	48	48	0	0
ALP	<=140	35	23	12	0.279213
	141 sd 210	261	151	110	0.295649
	211 sd 280	106	87	19	0.204226
	281 sd 420	99	83	16	0.192106
	>=420	78	70	8	0.143613
					0.020815
SGPT	<=40	332	204	128	0.289549
	41 sd 80	141	113	28	0.216465
	81 sd 120	38	32	6	0.189423
	121 sd 200	33	30	3	0.132302
	>200	35	35	0	0
					0.020925
SGOT	<=41	288	172	116	0.292767
	42 sd 82	137	104	33	0.239769
	83 sd 164	82	69	13	0.189889
	>164	72	69	3	0.075222
Total Protein	>=6	408	288	120	0.263095
	5 sd 6	128	95	33	0.247876
	<5	43	31	12	0.257134
					0.006527
Albumin	4 sd 6	209	128	81	0.289954
	3 sd 4	255	195	60	0.236949
	2 sd 3	109	87	22	0.218422
	0 sd 2	6	4	2	0.276435
AG Ratio	>=3	3	3	0	0
	2 sd 3	21	14	7	0.276435
	<2	555	397	158	0.259414

Dari hasil perhitungan entropy dan gain yang terdapat pada tabel 8 terlihat bahwa atribut Direct Bilirubin (yang diberi tanda merah) mempunyai nilai gain paling tinggi yaitu 0,02859. Oleh karena itu Direct Bilirubin akan menjadi akar (node pertama) dari pohon keputusan yang terbentuk.

Setelah didapatkan nilai gain yang tertinggi maka selanjutnya dihitung dengan cara yang sama dengan menggunakan persamaan entropy untuk mendapatkan nilai entropy dan persamaan gain untuk mendapatkan nilai gain.

Setelah dilakukan hasil perhitungan dan gain, maka akan terbentuk pohon keputusan seperti gambar 2



Gambar 2
pohon Keputusan Liver Node 1

Model yang telah dibentuk diujitingkat akurasi dengan memasukkan atau uji yang berasal dari data training dengan menggunakan metode cross validation dan split persentase untuk mengujitingkat akurasi. Dengan menggunakan split persentase 60:40 maka diperoleh data training sebanyak 343 dan data testing 232 maka didapat hasil akurasi dengan menggunakan persamaan 2.5 maka akan diperoleh nilai akurasi sebanyak 72,845% dan error 27,155%. Seperti tertera pada tabel 9

Tabel 9. Nilai Akurasi Algoritma C4.5

Class\Prediksi (class)	Positif	Negative
Positif	169	0
Negative	63	0

Untuk pengujian menggunakan metode cross validation sebanyak 7,8,10 kali pengujian maka akan didapat seperti tertera pada tabel berikut:

Tabel 10
Pengujian menggunakan cross 7

ALgoritma C4.5			
row ID	Error in %	Size of Test Set	Error Count
fold 0	38.554	83	32
fold 1	22.892	83	19
fold 2	25.301	83	21
fold 3	24.39	82	20
fold 4	33.753	83	28
fold 5	32.53	83	27
Fold 6	21.951	82	18

Didalam pengujian menggunakan Cross 7 data yang digunakan untuk testing 82 data secara acak maka akan didapatkan hasil terbaik pada fold ke- 6 dengan error sebanyak 21.951% untuk lebih jelasnya dapat dilihat pada tabel 10.

Tabel 11. Pengujian menggunakan cross 8

Algoritma C4.5			
row ID	Error in %	Size of Test Set	Error Count
fold 0	26.027	73	9
fold 1	27.778	72	0
fold 2	27.397	73	0
fold 3	30.556	72	2
fold 4	22.222	72	6
fold 5	28.767	73	1
Fold 6	30.556	72	2
Fold 7	34.722	72	5

Dalam pengujian menggunakan Cross 8 data yang digunakan untuk testing 72 data secara acak maka akan didapatkan hasil terbaik pada fold ke- 4 dengan error

sebanyak 22.222% untuk lebih jelasnya dapat dilihat pada tabel 11

Tabel 12
pengujian menggunakan cross 10

ALgoritma C4.5			
row ID	Error in %	Size of Test Set	Error Count
fold 0	25.862	58	15
fold 1	41.379	58	24
fold 2	25.862	58	15
fold 3	29.31	58	17
fold 4	32.759	58	19
fold 5	20.69	58	12
Fold 6	24.138	58	14
Fold7	31.034	58	18
Fold8	22.414	58	13
Fold9	31.579	57	18

Didalam pengujian menggunakan Cross 10 data yang digunakan untuk testing 57 data secara acak maka akan didapatkan hasil terbaik pada fold ke- 5 dengan error sebanyak 20.69 % untuk lebih jelasnya dapat dilihat pada tabel 12.

3. Komparasi Acuraccy
Sebelum diterapkan pada data baru, terlebih dahulu dilakukan pengujian akurasi terhadap model yang telah terbentuk dengan menggunakan data testing. Hasil pengujian dengan menggunakan data training dan data testing 60:40 dapat dilihat pada tabel 13:

Tabel 13Nilai Accuracy

Metode	Accuracy	Error
Naïve Bayes	63362%	36.638 %.
Algoritma C4.5	72845%	27155%

Tabel 13 menunjukkan nilai akurasi terhadap metode yang digunakan peneliti, maka diperoleh hasil akurasi tertinggi dicapai oleh Algoritma C4.5 dengan akurasi sebesar 72.845 %

sedangkan algoritma Naïve Bayes hanya mencapai 63.8282%.

KESIMPULAN

Dalam penelitian ini dilakukan pembuatan model menggunakan algoritma Naïve bayes dan C4.5 menggunakan data Pasien Penderita Liver. Model yang dihasilkan, dikomparasi untuk mengetahui algoritma yang paling baik dalam penentuan Identifikasifikasi penyakit Liver. Untuk mengukur kinerja kedua algoritma tersebut digunakan metode pengujian Cross Validation, dan Split Percentace, dan pengukuranya dengan menggunakan *confusion matrix*, Algoritma C4.5 memiliki Akurasi yang lebih tinggi dengan nilai 69.828% dibandingkan Naïve Bayes dengan nilai 63.362%. Dengan demikian algoritma C4.5 dapat memberikan pemecahan untuk permasalahandalam mengidentifikasi penyakit Liver.

UCAPAN TERIMA KASIH

Dalam penyelesaian penelitian ini penulis tidak lupa mengucapkan terimakasih kepada semua pihak yang telah membantu untuk terselesaikannya penelitian ini kepada:

- 1) Ketua STMIK Nusa Mandiri
- 2) Rekan-rekan di STMIK Nusa Mandiri yeng telah memberikan motivasi dalam pengerjaan penelitian ini.

DAFTAR PUSTAKA

[1] Neshat, Mehdi, Mehdi Sargolzaei, Adel Nadjaran Toosi, dan Azra Masoumi. (2012). *Hepatitis Disease*

- Diagnosis using hybrid case based reasoning and particle swarm optimization.* International Scholarly Research Network: ISRN Artificial Intelligence Volume 2012
- [2] Han, J., and Kamber, M. (2006). *Data Mining Concepts and Techniques*. San Fransisco: Morgan Kauffman.
- [3] Wu, X., and Kumar, V. (2009). *The Top Ten Algorithms in Data Mining*. Boca Raton, London, New York: Taylor & Francis Group, LLC.
- [4] Departemen Kesehatan Republik Indonesia. (2009). *Profil Kesehatan Indonesia 2009*. Jakarta
- [5] Riduwan. (2008). *Metode dan Teknik Menyusun Tesis*. Alfabeta. (Bandung.
- [6] Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning and Tools*. Burlington: Morgan Kaufmann Publisher.
- [7] Sumathi, S., and Sivanandam, S. (2006). *Introduction to Data Mining and its Applications*. Verlag Berlin Heidelberg: Springer.
- [8] Santoso, B. (2007). *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta: Graha Ilmu.
- [9] Bramer, Max. (2007). *Principles of Data Mining*. London: Springer
- [10] Larose, D.T. (2005). *Discovering Knowledge in Data*. New Jersey: John Willey & Sons, Inc.
- [11] Maimon, Oded and Rokach, Lior. (2010). *Data Mining and Knowledge Discovery Handbook*. New York: Springer.