PENERAPAN ALGORITMA DECISION TREE C4.5 UNTUK DIAGNOSA PENYAKIT STROKE DENGAN KLASIFIKASI DATA MINING PADA RUMAH SAKIT SANTA MARIA PEMALANG

Sigit Abdillah A11.2011.06469
Program Studi Teknik Informatika – S1
Fakultas Ilmu Komputer
Universitas Dian Nuswantoro, Jl. Nakula 1 No. 5-11 Semarang
Sigit.abdillah99@gmail.com

ABSTRAK

Data Mining adalah proses ekstraksi sebelumnya tidak dikenal dan dipahami dari database berukuran besar dan digunakan untuk membuat keputusan bisnis yang penting.Studi kasus yang digunakan dan diterapkan dalam tugas akhir ini adalah data pasien penyakit saraf khususnya penyakit *Stroke* untuk dikelola menggunakan algoritma C4.5 dengan metode klasifikasi *data mining*.
Stroke termasuk penyakit pembuluh darah otak ditandai dengan kematian jaringan otak yang terjadi karena berkurangnya aliran darah dan oksigen ke otak. Salah satu cara untuk mempelajari Stroke yaitu dengan ilmu data mining tepatnya menggunakan algoritma C4.5, hasil laporan ini menentukan pasien penyakit stroke dengan variabel yang diketahui kemudian diolah menggunakan data mining algoritma C4.5.

Kata kunci :*Data mining*, algoritma C4.5, klasifikasi, *stroke*

I. PENDAHULUAN

Pada dunia kesehatan secara teknis sudah mengenal data mining dalam cakupan luas menjadi potensial informasi. Contohnya dalam bidang rekam medis sudah menggunakan beberapa teknik data mining modern pada beberapa kasus yang ada seperti klasifikasi dan data prediktif.Macam-macam kasus tersebut diantaranya yaitu terdapat teknik *Naïve Bayes classification* (NBC) yang diterapkan pada bidang kesehatan contohnya seleksi embrio, dan teknik data mining *Decision Tree*untuk mendeteksi dan memvalidasi hipertensi pada

kehamilan di rumah sakit ataupun instansi kesehatan lainnya[2].

Untuk menganalisa data dalam jumlah besar yang tersimpan pada *database*, biasanya digunakan teknik *data mining.*Meski telah umum digunakan pada industri keuangan dan telekomunikasi, teknik data mining mulai diterapkan secara intensif dibidang kesehatan. Sebagai contoh, Mayo Clinic bekerjasama dengan IBM menerapkan teknik data mining pada pasien dengan kesamaan jenis kelamin, usia dan riwayat kesehatan untuk mengetahui respon terhadap pengobatan tertentu.

Data Mining adalah proses ekstraksi sebelumnya tidak dikenal dan dipahami dari database berukuran besar dan digunakan untuk membuat keputusan bisnis yang penting[1].

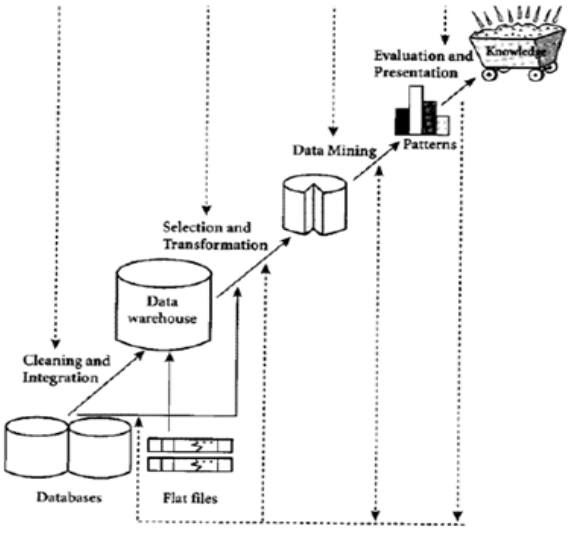
Pada dunia kesehatan secara teknis sudah mengenal data mining dalam cakupan luas menjadi potensial informasi. Contohnya dalam bidang rekam medis sudah menggunakan beberapa teknik data mining modern pada beberapa kasus yang ada seperti klasifikasi dan data prediktif. Macam-macam kasus tersebut diantaranya yaitu terdapat teknik *Naïve Bayes classification* (NBC) yang diterapkan pada bidang kesehatan contohnya seleksi embrio, dan teknik data mining *Decision Tree*untuk mendeteksi dan memvalidasi hipertensi pada kehamilan di rumah sakit ataupun instansi kesehatan lainnya.

Dalam hal ini studi kasus yang dibahas adalah mengenai salah satu penyakit berbahaya bagi manusia yang dapat menyebabkan kematian yaitu penyakit stroke, penyakit ini terbagi menjadi dua yaitu stroke mayor dan stroke minor yang dapat mengancam jiwa seseorang, dan dapat terjadi karena ada gangguan suplai darah pada sebagian atau seluruh organ otak.

Dari hal yang telah dijabarkan diatas, akan dilakukan pengujian mengenai mengangkat permasalahan tersebut sebagai Penerapan Algoritma Decision Tree C4.5 Untuk Diagnosa Penyakit Stroke Dengan Klasifikasi Data Mining Pada Rumah Sakit.

II. METODOLOGI

A. Tahap-tahap Data Mining



Gambar 1.Tahapan Data Mining

1. Pembersihan data (*Cleaning data*)

   Untuk menghilangkan data yang tidak diperlukan, data yang diperoleh dari tahap pengambilan dataset akan disaring untuk menghasilkan data yang benar-benar dibutuhkan. umumnya data tersebut memiliki nilai yang tidak sempurna seperti data yang hilang.Selain itu, ada juga atribut-atribut data yang tidak rsesuai dengan pemrosesan data mining yang akan digunakan.Data-data yang tidak relevan itu juga lebih baik dibuang karena keberadaannya bisa mengurangi mutu atau akurasi dari hasil data mining nantinya. Pembersihan data juga akan mempengaruhi performasi dari sistem data mining karena data yang ditangani akan berkurang jumlah dan kompleksitasnya.

2. Integrasi data

   Datayang akan digunakan untuk data mining tidak hanya berasal dari satu database tetapi juga berasal dari beberapa database atau file teks. Integrasi data dilakukan pada atribut-aribut yang mengidentifikasikan entitas-entitas yang unik seperti atribut nama, jenis produk, nomor pelanggan dan lain-lain. Pada tahap ini hal yang perlu dilakukan untuk lebih detail dan cermat karena kesalahan pada integrasi data bisa menghasilkan hasil yang menyimpang dan bahkanmenyesatkan pengambilan keputusan pada akhirnya. Dalam integrasi data ini juga perlu dilakukan transformasi dan pembersihan data karena seringkali data dari dua database berbeda tidak sama cara penulisannya atau bahkan data yang ada di satu database ternyata tidak ada di database lainnya.

3. Seleksi Data

   Data diseleksi untuk menentukan variabel apa saja yang akan diambil agar tidak terjadi kesamaan dan perulangan yang tidak diperlukan dalam pengolahan teknik data mining.Sebagai contoh, sebuah kasus yang meneliti faktor kecenderungan orangmembeli dalam kasus *market basket analysis*, tidak perlu mengambil namapelanggan, cukup dengan id pelanggan saja.

4. Transformasi data

   Pengubahan data menjadi format ekstensi yang sesuai untuk pengolahan dalam*data mining.*Beberapa metode *data mining* membutuhkan format data yangkhusus sebelum bisa diproses dalam teknik data mining. Misalnya sebagian metode standarseperti analisis asosiasi dan klastering hanya bisa

menerima input datakategorikal. Karenanya data berupa angka numerik yang berlanjut perlu dibagimenjadi beberapa interval.

5. Proses *mining*,

Untuk memproses teknik utama saat metode diterapkan agar menemukanpengetahuan berharga, data yang terkumpulkan sesuai prosedur harus di terapkan pada proses mining setelah data melalui tahap transformasi.

6. Evaluasi pola

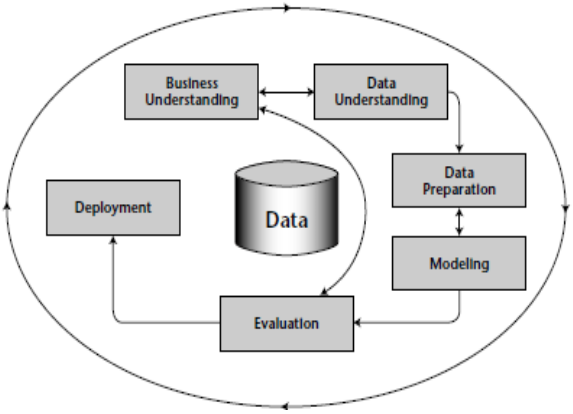
Tahap ini yaitu mengidentifikasi pola-pola menarik kedalam *knowledge based* yangdiidentifikasikan. Dalam tahap ini hasil dari teknik *data mining* berupa pola-pola yangkhas maupun model prediksi dievaluasi untuk menilai apakah kajian yang ada sudah memenuhi target yang diinginkan. Jika ternyata hasil yang diperoleh tidak sesuai kajian adabeberapa alternatif dengan mencoba metode *data mining* lain agarlebih sesuai, atau menerima hasil ini sebagai suatu hasil yang di luar dugaanyang mungkin bermanfaat.

7. Presentasi pengetahuan

Merupakan visualisasi dan penyajian pengetahuan mengenai metode yangdigunakan untuk memperoleh pengetahuan yang diperoleh pengguna. Tahapterakhir dari proses data miningadalah bagaimana memformulasikan keputusanatau aksi dari hasil analisis yang didapat. Ada kalanya hal ini harus melibatkanorang-orang yang tidak memahami data mining.Karenanya presentasi hasil *datamining.*

B. CRISP-DM (Cross Industry Standart Process for Data Mining)

CRISP-DM (CRoss-Industry Standard Process for Data Mining) merupakan proses standar yang biasa digunakan dalam penerapan ilmu data mining.



Gambar 2.CRISP-DM

1. Business Understanding

Memahami tujuan dan kebutuhan dari sudut pandang bisnis, kemudian menterjemakan pengetahuan ini ke pendefinisian masalah dalam data mining. Selanjutnya akan ditentukan

rencana dan strategi untuk mencapai tujuan tersebut.

Menerjemahkan tujuan dan batasan dari data yang diambil dari rumah sakit menjadi formula dari permasalahan data mining mulai dari menyiapkan strategi awal hingga metode yang dibutuhkan untuk mencapai tujuan.

2. Data Understanding

Pengumpulan data yang akan dilanjutkan mendeteksi adanya bagian yang menarik dari data yang dapat digunakan untuk hipotesa untuk informasi yang tersembunyi.

3. Data Preparation

Tahap ini meliputi semua kegiatan untuk membangun dataset akhir (data yang akan diproses pada tahap pemodelan) dari data mentah. Tahap ini dapat diulang beberapa kali.

Tahap ini adalah pemilihan atribut data yang digunakan, serta pembagian data menjadi dua kelompok yaitu data testing dan data training yang akan diimplementasikan pada analisa dan pembahasan

Tabel 1.atribut pengolahan data

Atribut	Detail penggunaan	
Id	✓	Nilai unique
Umur	✓	Nilai Model
Jenis kelamin	✓	No
Suhu tubuh	X	No
Denyut nadi	X	Nilai Model
Pernafasan	X	Nilai Model
Kesadaran	X	No
Sulit bicara	X	No
Gerak terbatas	X	Nilai Model
Badan lemas	X	No
Mual muntah	X	No
Factor keturunan	X	No
Diabetes mellitus	✓	Nilai Model
Hipertensi	✓	Nilai Model
Kolesterol	X	Nilai Model
Hemoglobin	X	Nilai Model
Kadar gula acak	X	Nilai Model
Kolesterol total	X	Nilai Model
Ket. Stroke	✓	Label Target

Setelah dilakukan pemilihan atribut pada proses pengolahan data dengan berdiskusi dengan pakar dibidangnya yaitu kepala rekam medis rumah sakit tempat melaksanakannya penelitian sebagai perwakilan dari pakar bidang kesehatan pada rumah sakit umum Santa Maria tersebut, apa sajakah variabel pendukung yang lebih

mempengaruhi tingkat keakuratannya dalam pengaruh penyakit stroke, penkonversian atribut ini berbguna untuk memudahkan dalam melakukan perhitungan dan analisa dalam tahap

No.	Umur	Hiper-tensi	Diabetes	Ket. Stroke
001	35	Ya	Positif	Tidak
002	49	Tidak	Negatif	Ya
003	55	Ya	Positif	Ya
004	57	Tidak	Negatif	Tidak
005	50	Ya	Negatif	Ya
006	57	Tidak	Positif	Tidak
007	61	Ya	Positif	Ya
008	49	Ya	Negatif	Tidak
009	30	Ya	Positif	Tidak
010	44	Tidak	Negatif	Ya
011	72	Ya	Positif	Tidak
012	81	Tidak	Positif	Ya
013	27	Tidak	Positif	Ya
014	68	Tidak	Positif	Ya
015	67	Ya	Negatif	Tidak
016	45	Tidak	Negatif	Ya
:	:	:	:	:
156	37	Ya	Negatif	Tidak

data mining

Table 2.ilustrasi atribut yang akan digunakan dalam pemodelan

Kemudian lakukan pengkonversian data agar mudah dilakukan pengolahan teknik data mining.

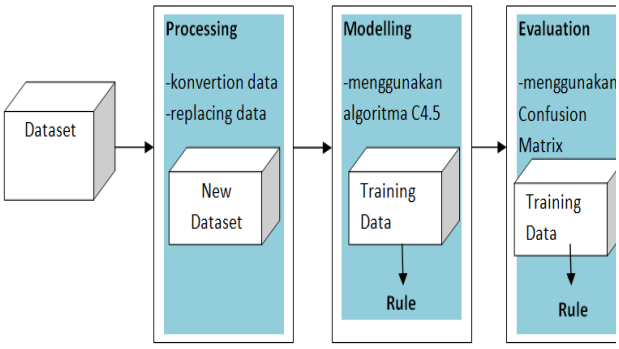
Table 3.

Id	Umur	Hipertensi	Diabetes	Ket. Stroke
001	Dewasa	Ya	Positif	Tidak
002	Tua	Tidak	Negatif	Ya
003	Tua	Ya	Positif	Ya
004	Tua	Tidak	Negatif	Tidak
005	Tua	Ya	Negatif	Ya
006	Tua	Tidak	Positif	Tidak
007	Tua	Ya	Positif	Ya
008	Tua	Ya	Negatif	Tidak
009	Muda	Ya	Positif	Tidak
010	Dewasa	Tidak	Negatif	Ya
011	Tua	Ya	Positif	Tidak
012	Tua	Tidak	Positif	Ya
013	Tua	Tidak	Positif	Ya
014	Tua	Tidak	Positif	Ya
015	Tua	Ya	Negatif	Tidak
016	Dewasa	Tidak	Negatif	Ya
:	:	:	:	:
156	Dewasa	Ya	Negatif	Tidak

Table 4.data setelah dikonversi

4. Modeling

Dalam tahap ini akan dilakukan pemilihan dan penerapan berbagai teknik pemodelan dan beberapa parameternya akan disesuaikan untuk mendapatkan nilai yang optimal.



Gambar 3 Model penelitian yang diusulkan

5. Evaluation

Melakukan evaluasi terhadap keefektifan dan kualitas model tujuan yang ditetapkan pada fase awal (Business Understanding).Kunci dari tahap ini adalah menentukan apakah ada masalah bisnis yang belum dipertimbangkan. Di akhir dari tahap ini harus ditentukan penggunaan hasil proses data mining.

6. Deployment

Pengetahuan atau informasi yang telah diperoleh akan diatur dan dipresentasikan dalam bentuk khusus sehingga dapat digunakan oleh pengguna. Tahap deployment dapat berupa pembuatan laporan Dalam banyak kasus, tahap deployment melibatkan konsumen, di samping analis data, karena sangat penting bagi konsumen untuk memahami tindakan apa yang harus dilakukan untuk menggunakan model yang telah dibuat.

Data yang digunakan dalam penelitian ini adalah sumber data primer. Data yang dikumpulkan yaitu data pasien berpenyakit stroke

C. DECISION TREE ALGORITMA C4.5

*Decision Tree.*Pohon (*Tree*) adalah sebuah struktur data yang yan terdiri dari simpul (*node*) dan rusuk (*edge*). Simpul pada sebuah pohon keputusan dibedakan menjadi tiga, akar simpul, simpul percabangan, dan simpul akhir[12].

Pada pohon keputusan ini bisa memberikan keuntungan berwujud visualisasi dari pemecahan masalah yang diolah menggunakan teknik data mining yang membuat protocol dari prediksinya dapat diamati, maka dari itu konsep ini termasuk fleksibel dan atraktif. Pohon keputusan ini sendiri juga sudah banyak digunakan pada berbagai bidang ilmu pengetahuan, salah satunya yaitu bidang kesehatan untuk diagnosa penyakit pasien, ilmu computer pada struktur data, psikologi untuk teori pengambilan keputusan, dan lain-lain.

Dalam pohon keputusan sangat berhubungan dengan algoritma C4.5, karena dasar algoritma C4.5 adalah pohon keputusan.Algoritma data mining C4.5 merupakan salah satu algoritma yang

digunakan untuk melakukan klasifikasi atau segmentasi atau pengelompokan yang bersifat prediktif.Cabang-cabang pohon keputusan merupakan pertanyaan klasifikasi dan daun-daunnya merupakan kelas-kelas atau segmen-segmennya.

Rumus menghitung entropy pada algoritma C4.5

Entropi (S) = $\sum_{i=1}^{k} -pi\ log_2\ pi$ ……..(2.1)

Keterangan :

S : Himpunan (dataset) kasus

k : Banyaknya partisi S

Pi : Probabilitaas yang didapat dari Sum(Ya) atau Sum(Tidak) dibagi total kasus

Setelah mendapatkan entropi dari keseluruhan kasus, lakukan analisis pada setiap atribut dan nilai-nilainya dan hitung entropinya. Langkah berikutnya yaitu dengan menghitung Gain, rumus daripada Gain adalah sebagai berikut:

Gain(A) = Entropi (S) - $\sum_{i=1}^{k} \frac{|Si|}{|S|}$ *x Entropi (Si.)*

..……..……..(2.2)

Keterangan :

S : himpunan kasus

A : atribut

n : jumlah partisi atribut A

|*Si*| : jumlah kasus pada partisi ke-i

|*S*| *:* jumlah kasus dalam S

D. MATRIKS KONFUSI

Confusion Matrix adalah tool yang digunakan untuk evaluasi model klasifikasi untuk memperkirakan objek yang benar atau salah. Sebuah matrix dari prediksi yang akan dibandingkan dengan kelas yang asli dari inputan atau dengan kata lain berisi informasi nilai aktual dan prediksi pada klasifikasi.

Table 5.contoh*confusion matrix*

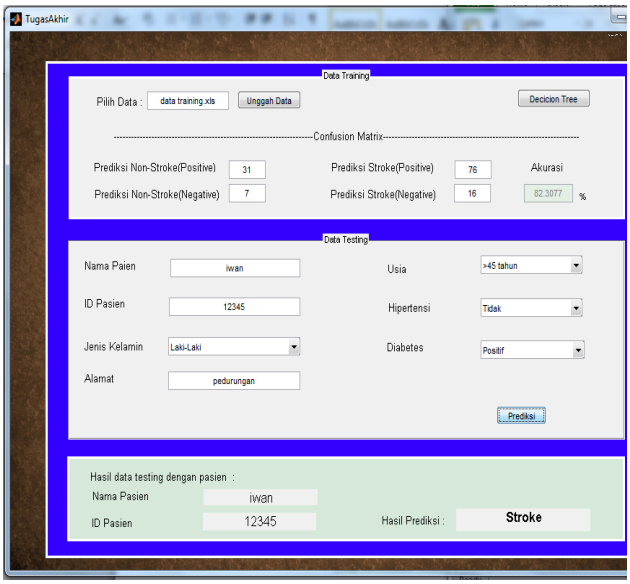
Classification	Predicted class	
	Class = Yes	Class = No
Class=Yes	a (true positive-TP)	b (false negative-FN)
Class=No	c (false positive-FP)	d (true negative-TN)

Rumus untuk menghitung tingkat akurasi pada matriks adalah:

Akurasi = $\frac{TP\ +TN}{TP+FP+TN+FN}$ = $\frac{a+d}{a+b+c+d}$....(2.3)

III. ANALISA DAN PEMBAHASAN

a. Pengoperasian Sistem



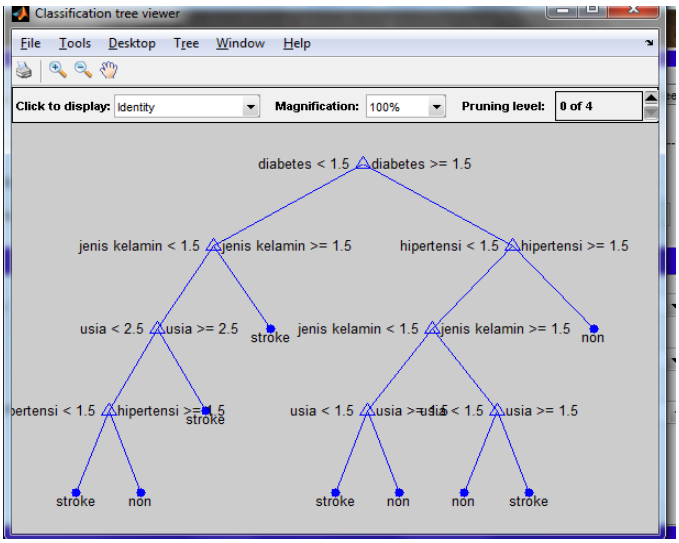
Gambar 4 Input Data Training pada Sistem

Data diinputkan pada sistem yang dibuat menggunakan matlab versi R2010a, isikan semua data yang dibutuhkan sesuai form yang tersedia. Pada gambar 4.3 langkah yang pertama dijalankan yaitu mengunggah file yang akan diolah dalam sistem, file yang digunakan adalah ‘data training TA’ yang berekstensi ‘.xls’ yaitu data pasien penyakit stroke setelah dikonversi, data tersebut mengandung empat variabel pendukung yaitu jenis kelamin, usia, jipertensi, dan diabetes. Serta mempunyai satu variabel target sebagai klasifikasi keputusan Stroke atau Non-stroke.

Dalam file ‘data training.xls’ tersebut terdapat tingkat akurasi sebesar 82,3077%, dengan cara menghitungnya yaitu menjumlahkan nilai Prediksi Stroke(positive) denganPrediksi Non-stroke(positive) kemudian membaginya dengan seluruh elemen variabel yang ada yaitu Prediksi Stroke(positive), Prediksi Stroke(negative), Prediksi Non-stroke(positive), dan Prediksi Non-stroke(negative). Nilainya yaitu (76+31)/(76+16+7+31) x 100% = 82,3077%.

b. Pohon Keputusan

Untuk mendukung aturan aturan yang terbentuk dari data pasien stroke yang diperoleh kedalam sistem maka dibentuklah pohon keputusan, selain berfungsi sebagai penentuan *rules*atau aturan klasifikasi penyakit stroke, sistem pohon keputusan ini juga mempresentasikan bagaimana seorang pasien bisa terserang stroke dari beberapa variabel yang tersedia dari data pasien penyakit stroke.



Gambar 5. Pohon Keputusan pada Sistem

Berikut penjelasan data training dan data testing yang akan digunakan dalam proses uji coba tingkat akurat data pasien, dari 156 data akan dibagi menjadi dua bagian yaitu data training yang berjumlah 130 data pasien dan sisanya pada data testing yaitu berjumlah 26 data pasien.

Table 6. Pembagian Data Testing dan Data Training

	jumlah	persentase
Data Training	130	83,33%
Data Testing	26	16,67%

c. Confusion Matrix

Table 7.confusion matrix dari data testing

	True Stroke	True Nonstroke
Prediksi Stroke	15	6
Prediksi Nonstroke	0	5

Pada tabel 7 tersebut menjelaskan bahwa jumlah tabel (Prediksi Stroke - True Stroke) atau (a) yaitu 15 merupakan jumlah pasien diklasifikasikan Stroke, jumlah (Prediksi Stroke – True Nonstroke) atau (b) adalah 6 merupakan jumlah pasien yang diklasifikasikan Stroke tetapi masuk kedalam NonStroke, jumlah (Prediksi NonStroke – True Stroke) atau (c) adalah 0 merupakan jumlah pasien yang diklasifikasikan NonStroke tetapi masuk kedalam Stroke, sedangkan jumlah (Prediksi NonStroke – TrueNonStroke) adalah 5 merupakan jumlah pasien yang diklasifikasikan NonStroke.

Accuracy = (a+d)/(a+b+c+d)

= (15+5)/(15+0+6+5) x 100%

= 76,92%

IV. KESIMPULAN

Berdasarkan hasil penelitian dari permasalahan yang dikembangkan dapat disimpulkan bahwa untuk studi kasus penyakit stroke dapat memanfaatkan teknik klasifikasi data mining dengan algoritma C4.5 sebagai klasifikasi stroke atau non-stroke. Data yang digunakan sebagai penelitian disini adalah data pasien penyakit stroke rumah sakit yang sifatnya rahasia

Dari metode klasifikasi data mining dengan algoritma C4.5 dan pengaplikasian pohon keputusanyang membentuk aturan tersebut terdapat akurasi pada data training yang berjumlah 130 dari 156 data pasien sebesar 82,31% sedangkan akurasi pada data testing yang berjumlah 26 dari 156 data pasien sebesar 76,92%. Perhitungan keduanya menggunakan*confusion matrix.*

V. DAFTAR PUSTAKA

[1] R. A. Prasetyo, "Aplikasi Data Mining Asociation Rule Untuk

Menampilkan Informasi Pola Penyebaran Penyakit ISPA Menggunakan Algoritma Apriori (Studi Kasus di Poliklinik Universitas Dian Nuswantoro)", Departemen Universitas Dian Nuswantoro, vol. 1, pp. 2, 2013.

[2] Murtanto, A, "klasifikasi biaya pasien rawat inap penyakit jantung menggunakan teknik data mining attribute important (ai) dan algoritma naive bayes," Skripsi Teknik Kendali Universitas Halu Oleo, 2014.

[3] I. K. Gama, I. K. W. Yasa dan I. Hartini "Kepatuhan Kontrol Penderita Hipertensi Dengan Kejadian Stroke", Keperawatan Politeknik Kesehatan Denpasar,vol. 1, pp. 4-5. 2011

[4] M. K. Mukhlis. 2011. "*Diagnosa Kemungkinan Pasien Terkena Stroke Dengan Menggunakan Metode Naive Bayes Dan Metode Jaringan Syaraf Tiruan Berbasis Web"*. Surabaya: Institut Negeri Sepuluh Nopember

[5] S.A. Aji, M. Sarosa dan S. Onny. 2014. "*Klasifikasi Stroke Berdasarkan Kelainan Patologis Dengan Leraning Vector Quantization"*. Surabaya.

[6] A. Linda. 2012. "*Sistem Pakar Pendeteksi Kemungkinan Penyakit Stroke"*.Palembang: Universitas Bina Dharma

[7] dr. pinzon, R. dr Asanti L. "Awas Stroke! pengertian, gejala, tindakan, perawatan, & pencagahan". Andi

[8] E. Prasetyo, DATA MINING – "Konsep dan Aplikasi Menggunakan Matlab",Yogyakarta: CV. ANDI, 2012

[9] E. Prasetyo, "DATA MINING - Mengolah Data Menjadi Informasi Menggunakan Matlab", Yogyakarta: CV. ANDI, 2014.

[10] P.P. Widodo, R.T. Handayanto dan Herlawati, "Penerapan Data Mining Dengan Matlab",Bandung: Rekayasa Sains, 2013.

[11] D. Retnosari, "Sistem Aplikasi Data Mining Untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa,", Departemen Teknik Informatika Universitas Islam Kalimantan, vol. 1, pp. 16-17, 2013.

[12] F.A. Himawati, Data Mining, Yogyakarta: ANDI, 2013