MODEL ALGORITMA C4.5 UNTUK PREDIKSI PENYAKIT JANTUNG

Abdul Rohman[1)] dan M.Rochcham[2)]
Email[1]: abdulrohman15@gmail.com
Email[2]: muhrochan@gmail.com

ABSTRACT

Penyakit jantung merupakan salah satu penyakit yang mematikan di dunia, dan perlu dikaji dan dianalisis. Dalam penelitian ini dilakukan prediksi penyakit jantung menggunakan algoritma *C4.5* dan terbentuk model algoritma. Dari hasil pengujian dengan mengukur metode *C4.5* menggunakan, *confusion matrix*, dan *curve ROC*, diketahui bahwa algoritma *C4.5* menghasilkan nilai akurasi 86,59 %, nilai AUC yang diperoleh 0.957 dan masuk kategori kelompok klasifikasi yang sangat baik.

Kata Kunci: Algoritma, C4.5, Jantung

PENDAHULUAN

Banyak penelitian tentang prediksi penyakit jantung dengan teknik klasifikasi *Data Mining*, diantaranya penelitian yang dilakukan oleh Palaniappan dan Awang yaitu dengan melakukan komporasi 3 metode yaitu *Naives Bayes*, *Decision Tree*, dan *Artificial Neural Network (ANN)* dan hasilnya *Decision Tree* menghasilkan nilai terbaik (Palaniappan dan Awang, 2008)

Anbarasi dkk (2010) dalam penelitiannya yaitu memprediksi kelangsungan hidup penyakit jantung dengan dengan menggunakan metode *Naïve Bayes, Decision Tree* dan *Clasification Via Clustering*. Dan hasilnya metode *Decision Tree* mendapat nilai terbaik.

Kotsiantis (2007) dalam review papernya, *Decision Tree* mempunyai kelebihan dalam mengolah dataset penyakit jantung yaitu kecepatan dalam klasifikasi, bersifat diskrit dalam tiap atribut, binari dan *continue*, serta adanya transparansi pengetahuan atau klasifikasi.

Berdasarkan atas penelitian diatas, peneliti akan memilih metode *Decision Tree* atau C4.5 dalam memprediksi penyakit jantung sehingga terbentuk modelnya, dengan mengoptimal atribut-atribut yang berasal dari dataset yang terpercaya untuk memprediksi penyakit jantung dengan tujuan agar akurasi menjadi meningkat.

KAJIAN PUSTAKA

Penyakit Jantung

Berdasarkan dataset penyakit jantung di UCI (*Univercity of California Irvine*) terdapat 14 atribut yaitu umur, jenis kelamin, jenis sakit dada, tekanan darah, kolestrol, kadar gula, elekrokardiografi, tekanan darah, angina induksi, oldpeak, segmen_st, flaurosopy, denyut jantung dan hasil sebagai label yang terdiri atas *healthy* (sehat) dan *sick* (sakit). Semua atribut tersebut selain hasil merupakan hal-hal yang mempengaruhi terjadinya penyakit jantung.

Algoritma C4.5

Algoritma Decision Tree banyak digunakan untuk membangun sebuah pohon keputusan, dimana pada awalnya berupa data set menjadi model pohon keputusan (Gorunescu, 2011).

Ada 6 tahap dalam membuat sebuah pohon keputusan dengan algoritma *C4.5*yaitu:

1. Mempersiapkan data *training*.
2. Menentukan *entrophy*

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

3. Menghitung nilai *gain*

$$gain = -\sum_{i=1}^{p} {}^{ni}_{n} IE(i)$$

4. Menghitung Split Information

$$Split\ Information = -\sum_{t=1}^{c} \frac{s1}{s} log_2 \frac{s1}{s}$$

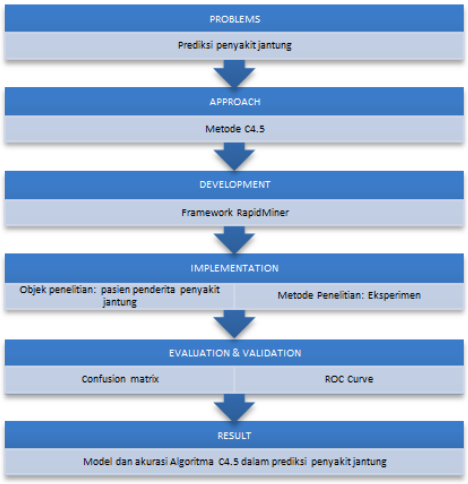
5. Menghitung gain ratio

$$Gainratio(S.A) = \frac{Gain(S.A)}{SplitInformation(S.A)}$$

6. Ulangi langkah ke-2 hingga semua *record* terpartisi

Kerangka Pemikiran

Kerangka pemikiran dalam proposal ini dimulai dari kurang sadarnya masyarakat atas gejala penyakit jantung serta kurang akuratnya penerapan algoritma C4.5 dalam memprediksi penyakit jantung.



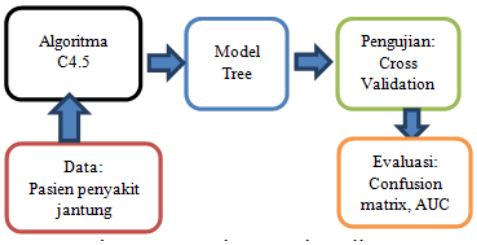
Gambar 1. Kerangka Pemikiran

METODOLOGI PENELITIAN

Dalam penelitian ini menggunakan data pasien yang melakukan pemeriksaaan penyakit jantung yang didapat dari UCI (*Universitas California, Invene*) *Machine Learning Repository* (Jasoni dan Steinbrunn, 2011). Hasil yang didapat sebanyak 867 orang yang diperiksa dan sebanyak 364 pasien terdeteksi sakit, sehingga 503 pasien terdeteksi sehat. Dataset tersebut adalah penggabungan antara dataset dari Cleveland yang terdiri dari 303 pasien, data dari statlog yang terdiri dari 270 pasien, dan data dari hungaria terdiri dari 294 pasien.

Penelitian ini adalah penelitian *experiment* yang melibatkan penyelidikan tentang perlakuan pada parameter dan variabel yang semuanya tergantung pada peneliti itu sendiri. *softwared*an *hardware* sebagai alat bantu dalam penelitian ini adalah sebagai berikut:

Model yang diusulkan pada penelitian ini adalah menggunakan algoritma *C4.5*yaitu:



Gambar 2. Metode yang diusulkan

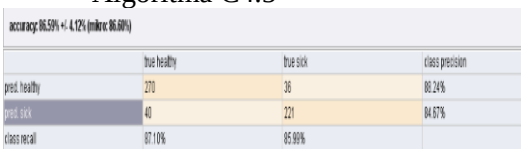
HASIL PEMBAHASAN

Dalam pengujian K-*Fold Cross Validation* Algoritma C4.5, peneliti menggunakan 10 kali percobaan dengan sampling type Stratified (bertingkat-tingkat) dengan menggunakan use local random seed karena hasil akurasinya juga lebih tinggi.

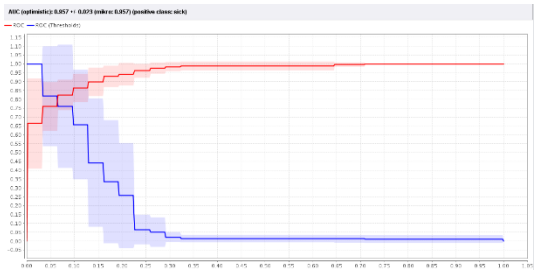
Metode klasifikasi bisa dievaluasi berdasarkan beberapa kriteria seperti tingkat akurasi, kecepatan, kehandalan, skalabilitas, dan interpretabilitas (Vercellis, 2009). Hasil pengujian model yang dilakukan adalah untuk mengukur tingkat akurasi dan AUC (*Area Under Curve*) dari prediksi penyakit jantung dengan metode *cross validation*

Hasil dari pengujian model yang telah dilakukan adalah untuk mengukur tingkat akurasi dan AUC (*Area Under Curve*).

Tabel 2. Model *Confusion Matrix* untuk Algoritma C4.5



Grafik ROC dengan nilai AUC (*Area Under Curve*) sebesar 0.957



Gambar 3. Nilai AUC dalam grafik ROC algoritma C4.5

Dari hasil pengujian diatas, baik evaluasi menggunakan *counfusion matrix* maupun ROC *curve* bahwa hasil pengujian algoritma C4.5

memiliki nilai akurasi sebesar 86,59% dengan nilai AUC 0,957

Tabel 3. Pengujian algoritma C4.5

	Accuracy	AUC
C4.5	86,59	0,957

Berdasarkan pengelompokan akurasi data mining maka akurasi 0.90-1.00 termasuk *Excellent classification* (Gorunescu, 2011):
Akurasi 0.90-1.00 = *Excellent classification*

SIMPULAN

Dalam penelitian ini dilakukan pengujian model dengan menggunakan algoritma *C4.5* dengan menggunakan data pasienyang menderita penyakit jantung atau tidak. Model yang dihasilkan diuji untukmendapatkan nilai *accuracy*, dan AUC dari setiap algoritmasehingga didapat pengujian dengan menggunakan *C4.5* didapat nilai *accuracy* adalah 86,59 % dengan nilai AUC adalah0.957, dan masuk kategori kelompok klasifikasi yang sangat baik, karena nilai AUC antara 0.90 sampai1.00

DAFTAR PUSTAKA

Anbarasi, M., Anupriya, E., and Iyengar, N.CH.S.N. 2010. Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm. *International Journal of Engineering Science and Technology*. Vol 2(10): 5370-5376.

Gorunescu, F. 2011. *Data mining: Concepts Models And Technique*. Springer. Berlin 2011.

Jasoni, A., and Steinbrunn, W. *UCI MAchine Learning Repository*: Retrieved from UCI MAchine Learning Repository: http:// archive.ics.uci.edu/ml/datasets/Heart+Disease, 2011.

Kotsiantis, S.B. 2007. Supervised Machine Learning: A Review of Classification Techniques. *Informatic.* Vol 31: 249-268

Palaniappan, S., and Awang, R. 2008. Intelligent Heart Disease Prediction System Using Data Mining Techniques. *International Journal of Computer Science and Network Security*. Vol 8(8):343-350.

Vercellis, C., 2009. *Business Intelligence: Data Mining and Optimization for Decision Making Decision Making*. Southern Gate, Chichester, West Sussex. United Kingdom: John Wiley & Sons Ltd.