# MODEL ALGORITMA C4.5 UNTUK PREDIKSI PENYAKIT JANTUNG

**Abdul Rohman[a1)] dan M.Rochcham[a2)]**
[a]Program Studi DIII Teknik Elektronika Fakultas Teknik Universitas Pandanaran
Jl. Banjarsari Barat No.1 Pedalangan, Banyumanik, Semarang 50268
Email[1]: abdulrohman@unpand.ac.id
Email[2]: muhrochan@gmail.com

## ABSTRACT

Penyakit jantung terjadi akibat adanya penyumbatan sebagian atau total dari suatu pembuluh darah. Akibatnya adanya peyumbatan, maka dengan sendirinya suplai energi kimiawi ke otot jantung akan berkurang, sehingga terjadi gangguan keseimbangan antara suplai dan kebutuhan. 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

Industri kesehatan memiliki sejumlah besar data kesehatan, namun sebagian besar data tersebut tidak diolah untuk mengetahui informasi tersembunyi untuk dijadikan pengambilan keputusan yang efektif oleh para praktisi kesehatan. Pengambilan keputusan atas dasar data dan informasi yang akurat akan menghasilkan keputusan dan prediksi penyakit menjadi tepat sasaran.

Penyakit jantung di Indonesia merupakan penyakit nomor satu yang mendorong angka kematian yang cukup tinggi, sehingga sampai sekarang penyakit tersebut ditakuti oleh manusia. Oleh karena itu penyakit jantung perlu diprediksi yaitu dengan menggunakan klasifikasi data mining sehingga praktisi kesehatan dalam pengambilan keputusan bisa lebih tepat dan akurat.

Banyak penelitian prediksi penyakit jantung dengan teknik klasifikasi *Data Mining*, diantaranya dilakukan oleh Palaniappan dan Awang dengan melakukan komporasi 3 metode yaitu *Naives Bayes, Decision Tree*, dan *Artificial Neural Network (ANN)* dengan total kasus 909 dan 15 atribut. Hasil dari penelitian tersebut metode *Decision Tree* menghasilkan nilai terbaik (Palaniappan dan Awang, 2008)

Penelitian yang dilakukan oleh Anbarasi dkk (2010) dalam memprediksi kelangsungan hidup penyakit jantung dengan berdasarkan 909 kasus dan 6 Atribut dengan menggunakan metode *Naïve Bayes, Decision Tree* dan *Clasification Via Clustering*. Hasil penelitian tersebut metode *Decision Tree* menghasilkan nilai terbaik.

Selain itu juga Kotsiantis (2007) dalam review papernya menjelaskan, bahwa metode *Decision Tree* mempunyai kelebihan-kelebihan dalam mengolah dataset penyakit jantung yaitu dari segi; kecepatan dalam klasifikasi, tiap atribut bersifat diskrit, binari dan kontinue, serta 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

Jantung adalah organ berupa otot, berbentuk kerucut, berongga dengan basisnya diatas dan puncaknya dibawah. Yang fungsinya untuk memompa bersih ke seluruh tubuh dan darah kotor ke paru-paru. Jika terjadi gangguan pada jantung maka fungsi pemompaan darah akan terganggu bahkan bisa mengakibatkan kematian.

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* digunakan untuk membangun sebuah pohon keputusan yang mudah dimengerti, fleksibel, dan menarik karena dapat divisualisasikan dalam bentuk gambar Pohon keputusan adalah salah satu metode klasifikasi yang paling populer karena mudah untuk diinterpretasi oleh manusia. Pohon keputusan adalah model prediksi menggunakan struktur pohon atau struktur berhirarki. Konsep dari pohon keputusan adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan (Gorunescu, 2011).

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

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

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

3. Hitung nilai *gain* dengan rumus:

$$\text{Gain} = -\sum_{i=1}^{p} \frac{ni}{n}.IE(i) \quad (2)$$

4. Untuk menghitung gain ratio perlu diketahui suatu term baru yang disebut Split Information dengan rumus:

$$SplitInformation = -\sum_{t=1}^{c} \frac{S_1}{S} log_2 \frac{S_1}{S} \quad (3)$$

5. Selanjutnya menghitung gain ratio

$$Gainratio(S,A) = \frac{Gain(S,A)}{SplitInformation(S,A)} \quad (4)$$

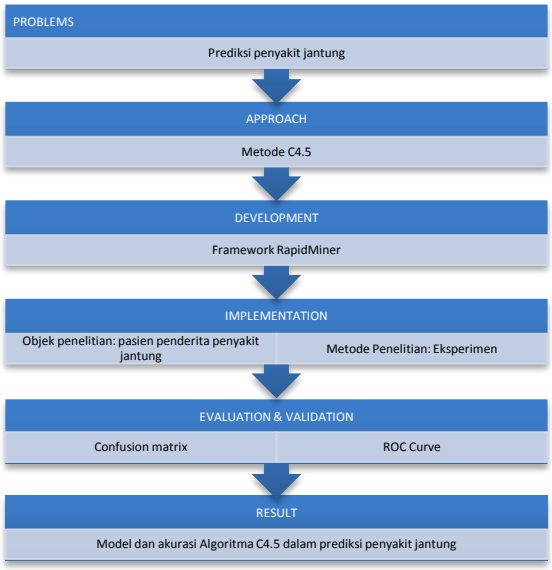
6. Ulangi langkah ke-2 hingga semua *record* terpartisi

Proses partisi pohon keputusan akan berhenti disaat:

a. Semua tupel dalam *record* dalam simpul m mendapat kelas yang sama
b. Tidak ada atribut dalam *record* yang dipartisi lagi
c. Tidak ada *record* didalam cabang yang kosong.

**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.

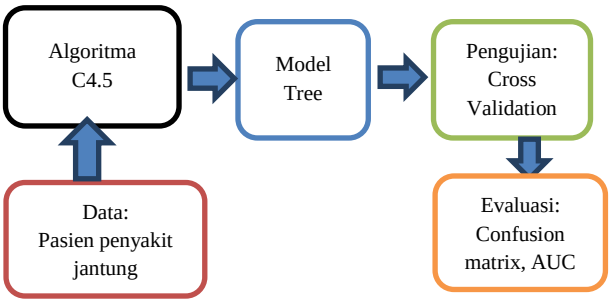
Penelitian ini adalah penelitian *experiment* yang melibatkan penyelidikan tentang perlakuan pada parameter dan variabel yang semuanya tergantung pada peneliti itu sendiri. *software*dan *hardware* sebagai alat bantu dalam penelitian ini adalah sebagai berikut:

Tabel 1.Spesifikasi *hardware* dan *software*

| Software | Hardware |
|---|---|
| Sistem operasi: Windows XP SP III 32 bit<br><br>Data Mining: RapidMiner Versi 5 | CPU: Dual Core 1,7 Ghz Ram 2 GB, Hdd 160Gb |

Data yang diperoleh dari UCI akan di *preprocessing* terlebih dahulu supaya data berkualitas dengan cara manual.Jadi data yang diolah dan diteliti sebanyak 567 dengan keadaan sakit sejumlah 257 orang dan keadaan sehat sejumlah 310 orang.

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




Gambar 2. Metode yang diusulkan

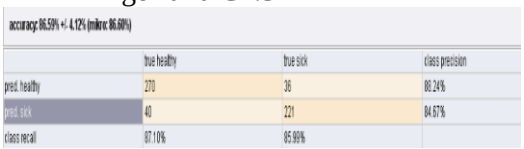
## HASIL PEMBAHASAN

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.

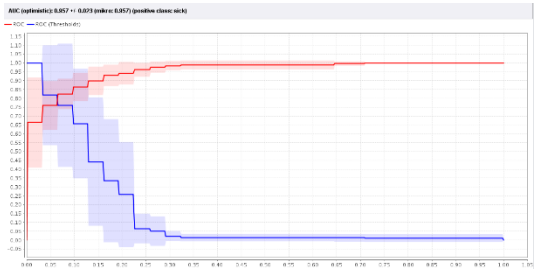
Metode klasifikasi bisa dievaluasi berdasarkan beberapa kriteria seperti tingkat akurasi, kecepatan, kehandalan, skalabilitas, dan interpretabilitas (Vercellis, 2009). Hasil pengujian model yang dilakukan adalah untuk mengukur tingkat akurasi dan AUC (*Area Under Curve*) dari prediksi penyakit jantung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

Tingkat akurasi dapat di diagnosa pada dasarnya sebagai berikut (Gorunescu, 2011):

1. Akurasi 0.90-1.00 = *Excellent classification*
2. Akurasi 0.80-0.90 = *Good classification*
3. Akurasi 0.70-0.80 = *Fair classification*
4. Akurasi 0.60-0.70 = *Poor classification*
5. Akurasi 0.50-0.60 = *Failure*

Karena nilai AUC dalam penelitian ini adalah 0.957, maka masuk kategori kelompok klasifikasi yang sangat baik, karena nilai AUC antara 0.90 sampai1.00

## 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.