

PENERAPAN ALGORITMA C4.5 BERBASIS ADABOOST UNTUK PREDIKSI PENYAKIT JANTUNG

Abdul Rohman *)

Abstrak

Penyakit jantung adalah terjadinya penyumbatan sebagian atau total dari suatu lebih pembuluh darah, akibatnya adanya penyumbatan maka dengan sendirinya suplai energi kimiawi ke otot jantung berkurang, sehingga terjadi gangguan keseimbangan antara suplai dan kebutuhan. Penelitian dalam memprediksi penyakit jantung telah banyak dilakukan oleh beberapa peneliti terdahulu. Dalam penelitian ini akan dilakukan prediksi penyakit jantung menggunakan metode algoritma C4.5 dan ditingkatkan kinerja dari algoritma C4.5 menggunakan metode adaboost yang di implementasikan pada data pasien penyakit jantung. Dari hasil pengujian dengan mengukur metode C4.5 berbasis adaboost menggunakan, confusion matrix, dan curve ROC, diketahui bahwa algoritma C4.5 menghasilkan nilai akurasi 86,59 %, nilai AUC yang diperoleh 0.957 dan setelah dioptimasi dengan menggunakan metode adaboost menjadi 92,24 %, nilai AUC menjadi 0.982. dengan melihat nilai accuracy dan AUC setelah dioptimasi, maka algoritma C4.5 berbasis adaboost masuk katagori kelompok klasifikasi yang sangat baik, karena nilai AUC antara 0.90 – 1.00

Kata Kunci:*C4.5, Adaboost, 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.

*) Dosen Jurusan Teknik Elektronika Fakultas Teknik Universitas Pandanaran

Banyak penelitian prediksi penyakit jantung dengan teknik klasifikasi *Data mining*, diantaranya dilakukan oleh Palaniappan dan Awang [1] 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.

Penelitian yang dilakukan oleh Anbarasi dkk [2] 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 S.B. Kotsiantis 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 [3]

Berdasarkan atas penelitian diatas, peneliti akan memilih metode *Decision tree* dalam memprediksi penyakit jantung. Dalam penelitian ini akan dilakukan penerapan algoritma *Decision tree (C4.5)* menggunakan metode *adaboost* dengan mengoptimal atribut-atribut yang berasal dari dataset yang terpercaya untuk memprediksi penyakit jantung dengan tujuan agar akurasi menjadi meningkat.

PERMASALAHAN

Berdasarkan hasil penelitian sebelumnya algoritma C4.5 dalam prediksi penyakit jantung akurasiya masih belum mencapai level excellence dan seberapa akurat model C4.5 yang ditingkatkan dengan metode *adaboost* dalam memecahkan masalah prediksi jantung.

PEMBAHASAN

a. 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 (University of California Irvine) terdapat 14 atribut yaitu umur, jenis kelamin, jenis sakit dada, tekanan darah, kolestrol, kadar gula, elektrokardiografi, 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.

b. Algoritma C4.5

Algoritma *decision tree* digunakan untuk membangun sebuah pohon keputusan yang mudah dimengerti, fleksibel, dan menarik karena dapat divisualisasikan dalam bentuk gambar [4] 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.

Ada beberapa tahap dalam membuat sebuah pohon keputusan dengan algoritma C4.5[4] 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-masing atribut atau berdasarkan nilai *index entropy* terendah. Sebelumnya dihitung terlebih dahulu nilai *index entropy*, dengan rumus:

$$Entropy(i) = \sum_{j=1}^m f(i, j) \cdot \log_2 f(i, j)$$

3. Hitung nilai *gain* dengan rumus:

$$gain = - \sum_{i=1}^p \frac{n_i}{n} \cdot IE(i) \quad (2.2)$$

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

$$SplitInformation = - \sum_{t=1}^c \frac{S_1}{S} \log_2 \frac{S_1}{S} \quad (2.3)$$

5. Selanjutnya menghitung gain ratio

$$Gainratio(S, A) = \frac{Gain(S, A)}{SplitInformation(S, A)} \quad (2.4)$$

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.

c. Metode Adaboost berbasis Adaboost

Setelah melakukan tahapan dalam membuat sebuah pohon keputusan dengan algoritma C4.5, dilakukan pemberian bobot pada pohon tunggal sehingga menghasilkan hipotesa baru dan sebuah pohon keputusan baru dengan langkah-langkah sebagai berikut [5]:

Metode adaboost adalah sebagai berikut:

1. Inisialisasi bobot data $\{W_n\}$ dengan $W_n^{(m)}$ untuk $n = 1, 2, \dots, N$.
2. For $m = 1 \dots, M$.
 - a. Training $y_m(x)$ dengan meminimalkan fungsi kesalahan (*error function*) sebagai berikut:

$$J_m = \sum_{n=1}^N W_n^{(m)} I(y_m(x_n) \neq t_n)$$

- b. Evaluasi kesalahan

$$\varepsilon_m = \frac{\sum_{n=1}^N W_n^{(m)} (y_m(x_n) \neq t_n)}{\sum_{n=1}^N W_n^{(m)}}$$

Dan kemudian digunakan evaluasi

$$a_m = \ln \left\{ \frac{1 - \varepsilon_m}{\varepsilon_m} \right\}$$

- c. Memperbaiki (update) bobot data

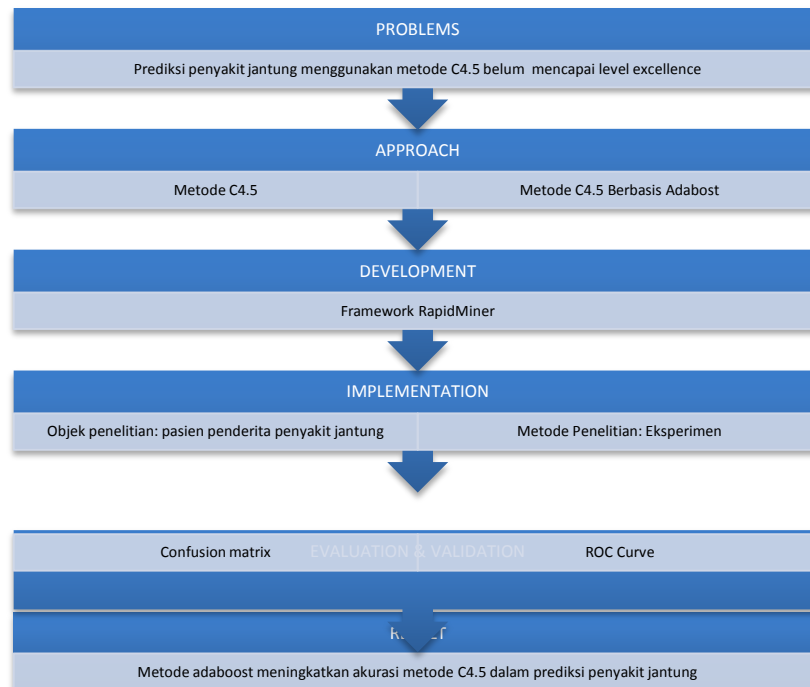
$$w_n^{(m+1)} = w_n^{(m)} \exp(a_m I(y_m(x_n) \neq t_n))$$

3. Membuat prediksi menggunakan model terakhir sebagai berikut

$$Y_m(x) = \text{sign}(\sum_{m=1}^M a_m y_m(x))$$

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 6.1 Kerangka Pemikiran

METODELOGI

Dalam penelitian ini menggunakan data pasien yang melakukan pemeriksaan penyakit jantung yang didapat dari UCI (*Universitas California, Invene*) *Machine Learning Repository* [6]. 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.

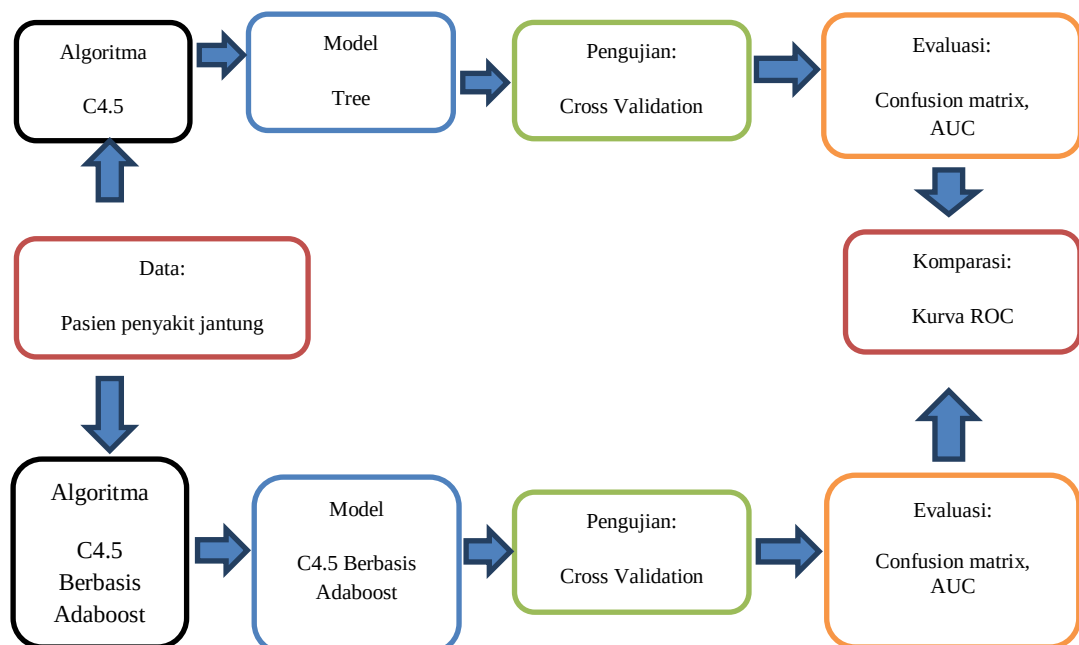
Penelitian ini adalah penelitian *experiment* yang melibatkan penyelidikan tentang perlakuan pada parameter dan variabel yang semuanya tergantung pada peneliti itu sendiri. *software* dan *hardware* sebagai alat bantu dalam penelitian ini adalah sebagai berikut:

Tabel 7.1 Spesifikasi hardware dan software

Software	Hardware
Sistem operasi: Windows XP SP III 32 bit	CPU: Dual Core 1,7 Ghz
Data Mining: RapidMiner Versi 5	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 dan algoritma C4.5 berbasis *adaboost* yaitu:



Gambar 7.1 Metode yang diusulkan

HASIL PEMBAHASAN

Dalam pengujian *K-Fold Cross Validation* Algoritma C4.5 dan Algoritma C4.5 berbasis Adaboost, peneliti juga 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 [7]. Hasil pengujian model yang dilakukan dalam bab tiga adalah untuk mengukur tingkat akurasi dan AUC (*Area Under Curve*) dari prediksi penyakit jantung dengan metode *cross validation*

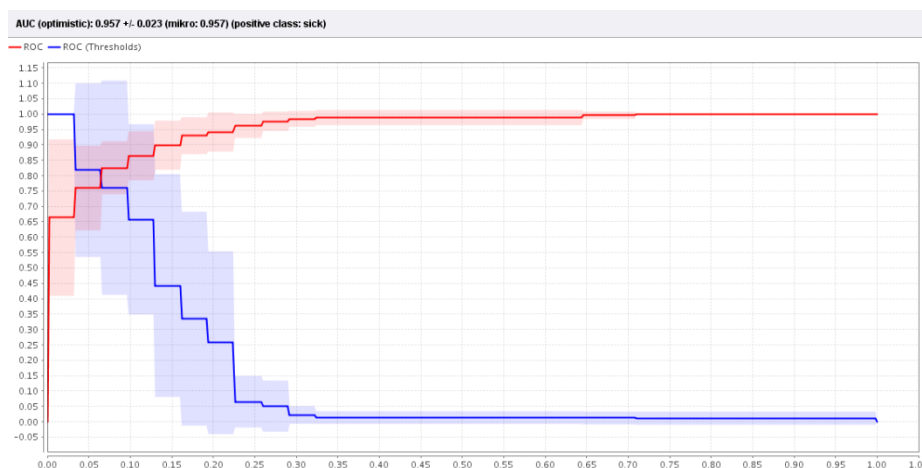
a. Hasil Pengujian Model Algoritma C4.5

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

Tabel 8.1 Model *Confusion Matrix* untuk Algoritma C4.5

accuracy: 86.59% +/- 4.12% (mikro: 86.60%)			
	true healthy	true sick	class precision
pred. healthy	270	36	88.24%
pred. sick	40	221	84.67%
class recall	87.10%	85.99%	

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



Gambar 8.1 Nilai AUC dalam grafik ROC algoritma C4.5

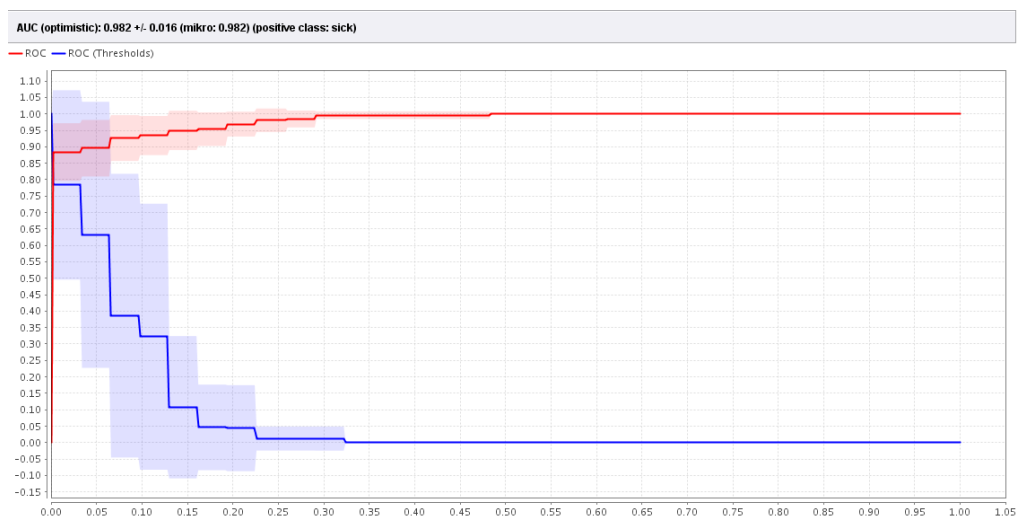
b. Hasil Pengujian Model Algoritma C4.5 berbasis Adaboost

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

Tabel 8.2 Model *counfusion matrix* untuk Algoritma C4.5 berbasis Adaboost

accuracy: 92.24% +/- 4.17% (mikro: 92.24%)			
	true healthy	true sick	class precision
pred. healthy	201	26	99.09%
pred. sick	19	232	92.43%
class recall	93.87%	90.27%	

Gambar 8.2 menunjukkan grafik ROC dengan nilai AUC (*Area Under Curve*) sebesar 0.982



Gambar 8.2Nilai AUC dalam grafik ROC algoritma C4.5 berbasis Adaboost

c. Analisis Evaluasi dan Validasi Model

Dari hasil pengujian diatas, baik evaluasi menggunakan *counfusion matrix* maupun *ROC curve* terbukti bahwa hasil pengujian algoritma C4.5 berbasis Adaboost memiliki nilai akurasi yang lebih tinggi dibandingkan dengan algoritma C4.5. Nilai akurasi untuk model algoritma C4.5 sebesar 86,59% dan nilai akurasi untuk model algoritma C4.5 berbasis Adaboost sebesar 92.24% dengan selisih akurasi 5,65%, dapat dilihat pada Tabel 8.2 dibawah ini:

Tabel 8.3 Pengujian algoritma C4.5 dan C4.5 berbasis Adaboost

	Accuracy	AUC
C4.5	86,59	0,957
C4.5 Berbasis Adaboost	92,24	0,982

Untuk evaluasi menggunakan ROC *curve* sehingga menghasilkan nilai AUC (*Area Under Curve*) untuk model algoritma C4.5 menghasilkan nilai 0.957 dan algoritma C4.5 berbasis Adaboost menghasilkan nilai 0.982 dengan nilai diagnosa *Excellent Classification*, dan selisih nilai keduanya sebesar 0.025. Jadi akurasi dari Algoritma C4.5 ke Algoritma C4.5 berbasis Adaboost meningkat 6,42% dan nilai AUC dari Algoritma C4.5 ke Algoritma C4.5 berbasis Adaboost meningkat 0,26%.

KESIMPULAN DAN SARAN

Dalam penelitian ini dilakukan pengujian model dengan menggunakan algoritma C4.5 dan algoritma C4.5 berbasis *adaboost* dengan menggunakan data pasien yang menderita penyakit jantung atau tidak. Model yang dihasilkan diuji untuk mendapatkan nilai *accuracy*, dan AUC dari setiap algoritma sehingga didapat pengujian dengan menggunakan C4.5 didapat nilai *accuracy* adalah 86,59 % dengan nilai AUC adalah 0.957. sedangkan pengujian dengan menggunakan C4.5 berbasis *adaboost* didapatkan nilai *accuracy* 92.24 % dengan nilai AUC adalah 0.982. selain itu juga peneliti membandingkan dengan algoritma C4.5 berbasis Bagging didapat *accuracy* 91,89% dan nilai AUC 0,963. Maka dapat disimpulkan pengujian model penyakit jantung dengan menggunakan algoritma C4.5 berbasis *adaboost* lebih baik dari pada C4.5 sendiri, dengan peningkatan akurasi sebesar 6,42% dan peningkatan nilai AUC sebesar 0,26%.

Dengan demikian dari hasil pengujian model diatas dapat disimpulkan bahwa C4.5 berbasis *adaboost* memberikan pemecahan untuk permasalahan penyakit jantung lebih akurat.

Dari hasil pengujian yang telah dilakukan dan hasil kesimpulan yang diberikan maka ada saran atau usul yang di berikan antara lain:

1. Dalam Penelitian ini dilakukan menggunakan metode algoritma C4.5 dan algoritma C4.5 berbasis metode *Adaboost*. Mencoba mengurangi beberapa atribut dan mencoba kembali dengan algoritma lain dengan mengoptimalisasi selain *adaboost* yang menghasilkan tingkat akurasi tinggi.
2. Hasil penelitian ini diharapkan bisa digunakan untuk rumah sakit untuk meningkatkan akurasi dalam prediksi penyakit jantung.

DAFTAR PUSTAKA

- [1] Sellappan Palaniappan & Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques," *IJCSNS International Journal of Computer Science and Network Security*, vol. 8, Agustus 2008.
- [2] E. Anupriya & Inyegar M. Anbarasi, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm," *International Journal of Engineering Science and Technology*, vol. 02, 2010..
- [3] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," *Department of Computer Science and Technology University of Peloponnese, Greece*, pp. 249-268, juli 2007.
- [4] Florin Gorunescu, *Data mining concepts models and technique*. berlin: Springer, 2011.
- [5] Wu, Xingdong, *The Top Ten Algorithm in Data Mining*. Minnesota: Taylor & Francis Group, 2009
- [6] A Jasoni and W Steinbrunn, *UCI MAchine Learning Repository*.: Retrieved from UCI MAchine Learning Repository:<http://archive.ics.uci.edu/ml/datasets/Heart+Disease>, 2011.
- [7] C. Vercellis, *Business Intelligence: Data Mining and Optimization for Decision Making Decision Making*. Southern Gate, Chichester, West Sussex, United Kingdom: John Wiley & Sons Ltd, 2009.