SEMINAR TESIS

Penerapan Algoritma C4.5 Dengan Seleksi Atribut Berbasis Algoritma Genetika Dalam Prediksi Penyakit Jantung

Edy
371 210 1184
Rekayasa Perangkat Lunak

Latar Belakang Masalah

- WHO telah memperkirakan angka kematian 12 juta setiap tahunnya terjadi disebabkan oleh penyakit jantung.
- Diagnosis medis dipandang sebagai tugas penting namun rumit yang perlu dijalankan secara tepat dan efisien.
- Terdapat beberapa teknik data mining yang populer seperti Algoritma C4.5 dan Naïve Bayes yang dipergunakan untuk memprediksi penyakit jantung (Srinivas, Raghavendra, & Govardhan, 2010, p. 1344)
- Algoritma C4.5 merupakan pengklasifikasian yang paling sederhana, namun sulit dalam menangani data dimensi tinggi

Latar Belakang Masalah (cont)

- Algoritma genetika dapat digunakan untuk mengurangi data aktual agar dapat mendapatkan subset yang optimal dalam memprediksi penyakit jantung.
- Seleksi atribut adalah sebuah proses pemilihan subset fitur dari fitur asli. Tujuan dari seleksi atribut adalah mengidentifikasi fitur dalam kumpulan data yang sama pentingnya, kemudian membuang fitur lainnya seperti informasi yang tidak relevan dan berlebihan (Maimon & Rokach, 2010)
- Pada penelitian ini akan dilakukan prediksi penyakit jantung menggunakan metode Algoritma C4.5 dengan seleksi atribut berbasiskan algoritma genetika.

Identifikasi Masalah

- Algoritma C4.5 dapat memecahkan masalah dalam hal prediksi, namun memiliki kekurangan yaitu pada data yang berukuran besar dapat mengurangi akurasi dan model yang terbentuk sulit dibaca karena terlalu kompleks.

Ruang Lingkup

- Penelitian ini dilakukan dengan proses eksperimen pada sebuah dataset dengan menggunakan algoritma C4.5. Algoritma genetika berbasis seleksi atribut akan digunakan untuk penentuan atribut yang akan dipergunakan untuk meningkatkan hasil prediksi dari algoritma C4.5 pada kasus prediksi penyakit jantung.

Rumusan Masalah

- Seberapa akurat algoritma C4.5 dibandingkan dengan seleksi atribut yang berbasiskan algoritma genetika dalam proses memprediksi penyakit jantung?

Tujuan Penelitian

- Penelitian ini bertujuan untuk menerapkan algoritma genetika untuk seleksi atribut sehingga dapat mengurangi dimensi dari data, serta mengidentifikasi fitur dalam kumpulan data dengan metode algoritma C4.5 dalam menganalisa prediksi penyakit jantung.

Manfaat Penelitian

- Dengan adanya data mining mempunyai potensi untuk menghasilkan lingkungan yang kaya pengetahuan yang dapat secara signifikan dalam meningkatkan kualitas keputusan klinis.
- Penelitian ini dapat bermafaat bagi para praktisi kesehatan dalam mendeteksi penyakit jantung sejak dini, agar dapat melakukan penanganan lebih awal.
- Penanganan lebih awal dapat memberikan dampak peningkatan kualitas kehidupan manusia seperti dalam hal budaya kerja yang baik, makanan yang pantas untuk dimakan, hinggal model hidup sehat.

Peneliti	Tahun	Metode	Hasil Akurasi
Anbarasi, Anupriya, Iyengar	2010	a. Naives Bayes b. Decision Tree c. Classification via Clustering	a. Naives Bayes 96.5% b. Decision Tree 99.2% c. Classification via Clustering 88.3%
Srinivas, Kavihta, Govrdhan	2010	Naives Bayes	84.14%
Ansari, Soni, Sharma	2011	Naives Bayes Decision Tree Clasification via Clustering	Akurasi: Naives Bayes 96.5% Decision Tree 99.2% CC 88.3%

Tinjauan Pustaka

- Pada penelitian ini melakukan tinjauan pustaka antara lain:
  1. Penyakit Jantung
  2. Data Mining
  3. Algoritma C4.5
  4. Seleksi Atribut
  5. Algoritma Genetika

Penyakit Jantung

- Jantung berfungsi untuk memompa darah melalui pembuluh darah dengan kontraksi yang ritmis dan berulang-ulang.
- Ada banyak penyebab penyakit jantung seperti pola hidup, kelainan bawaan sejak lahir, dan pola makan yang tidak sehat (Morrow, 2007, p. 34).

Faktor Terkena Resiko Penyakit Jantung

Faktor-Faktor yang Menambah Resiko Terkena Penyakit Jantung	
Dapat diubah	Tidak Dapat Diubah
• Merokok • Koresterol tinggi • Tekanan darah tinggi • Diabetes • Kegemukan • Stress • Kurang berolahraga	• Faktor genetika, misalnya kolesterol tinggi karena keturunan • Masalah gender: lebih banyak pria terkena penyakit jantung daripada wanita • Usia

Data Mining

- Data mining merupakan kegiatan melakukan ekstraksi untuk mendapatkan informasi penting yang sifatnya implisit dan sebelumnya tidak diketahui, dari suatu data. Data mining didefinisikan sebagai proses menemukan pola dalam data. Proses ini otomatis atau (biasanya) semi-otomatis (Witten, Frank, & Hall, 2011, p. 3)
- Data-data tersebut kemungkinan memiliki nilai
- Melakukan ekstraksi untuk mendapatkan informasi penting yang sifatnya implisit dan sebelumnya tidak diketahui, dari suatu data (Witten, 2011)

Data Mining

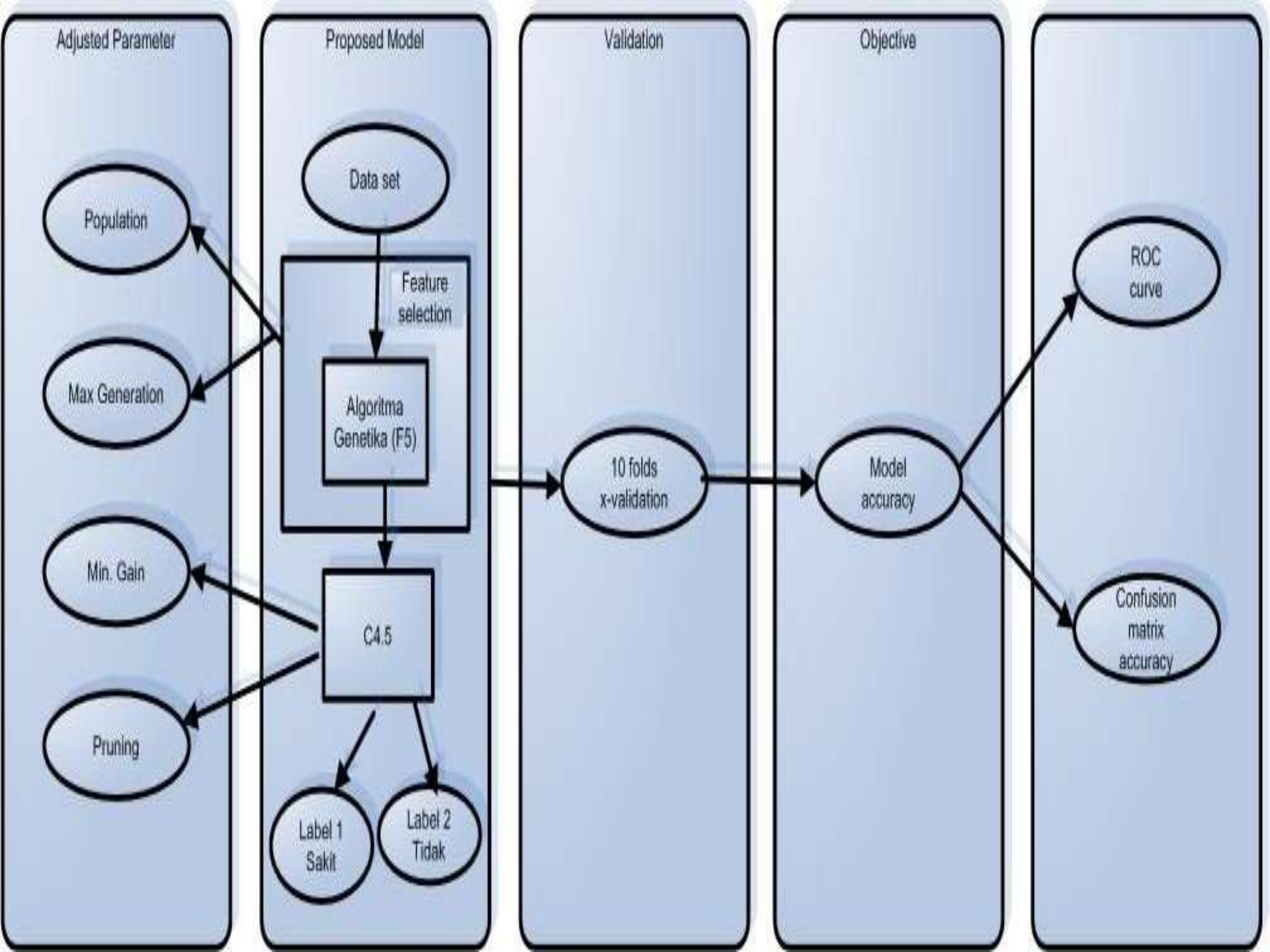
- Ada beberapa peran utama dalam data mining antara lain :
  - *Estimation* → Linear Regression, Neural Network, Support Vector Machine
  - *Prediction* →: Linear Regression, Neural Network, Support Vector Machine
  - *Classification* → Naive Bayes, K-Nearest Neighbor, C4.5, ID3, CART, Linear Discriminant Analysis
  - *Clustering* → K-Means, K-Medoids, Self-Organizing Map (SOM), Fuzzy C-Means
  - *Association* → A priori algorithm, FP-Growth algorithm, GRI algorithm

Algoritma C4.5

- C4.5 adalah bagian dari algoritma untuk klasifikasi dalam pembelajaran *machine learning* dan data mining. C4.5 merupakan algoritma yang cocok digunakan untuk masalah klasifikasi pada machine learning dan data mining (Wu & Kumar, 2009, p. 3).
- Ukuran pemilihan atribut ini juga disebut sebagai ***splitting rules*** karena menentukan bagaiman data akan dipisahkan kesetiap cabang.
- C4.5 yang merupakan pengembangan dari ID3 menggunakan ***Information gain*** untuk ukuran pemilihan atribut.

Seleksi Atribut

- Dapat mengurangi dimensi data, hal ini memungkinkan lebih efektif dalam operasi yang lebih cepat dari beberapa algoritma data mining.
- Dengan adanya seleksi atribut membuat algoritma data mining lebih cepat dan lebih efektif.


Adjusted Parameter
Proposed Model
Validation
Objective
Population
Max Generation
Min. Gain
Pruning
Data set
Feature selection
Algoritma Genetika (F5)
C4.5
Label 1 Sakit
Label 2 Tidak
10 folds x-validation
Model accuracy
ROC curve
Confusion matrix accuracy

Metodologi Penelitian

- Metode penelitian pada penelitian ini adalah penelitian eksperimen dengan tahapan penelitian seperti berikut:
  1. Pengumpulan data
  2. Pengolahan awal data
  3. Model/metode yang diusulkan
  4. Pengujian model
  5. Evaluasi dan validasi hasil

Metode Pengumpulan Data

- Data yang digunakan dalam penelitian ini merupakan dataset yang didapat dari UCI Machine Learning Repository pada bagian Heart Disease Dataset dengan menggunakan 14 atribut seperti: usia dalam tahun, jenis kelamin, tipe nyeri pada dada, dll.

Pengolahan Awal Data

- Dari seluruh data yang terkumpul, ada beberapa tahap pengolahan data (Han & Kamber, 2006, p. 7) yang dilakukan:

1. *Data cleaning*

   Data cleaning bekerja untuk membersihkan nilai yang kosong ,tidak konsisten atau mungkin tupel yang kosong (*missing values* dan *noisy*).

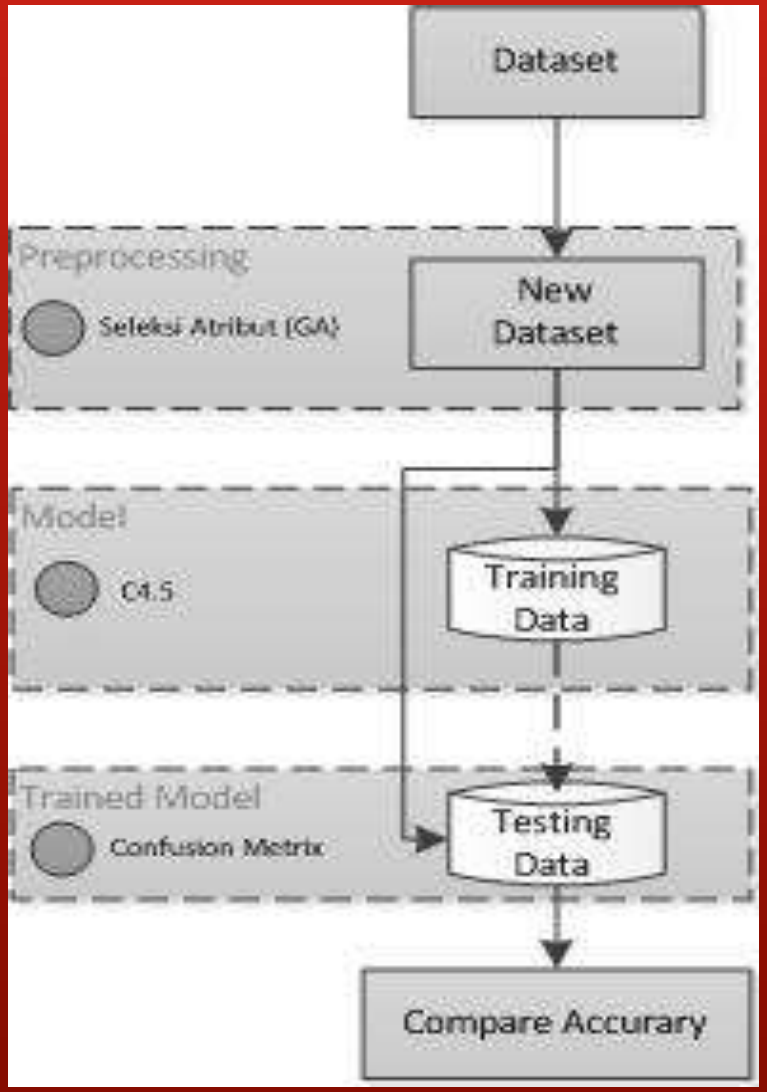
2. *Data integration*

   Data integration berfungsi menyatukan tempat penyimpanan (arsip) yang berbeda ke dalam satu data.

3. *Data reduction*

   Jumlah atribut dan tupel yang digunakan untuk data training mungkin terlalu besar, hanya beberapa atribut yang diperlukan sehingga atribut yang tidak diperlukan akan dihapus

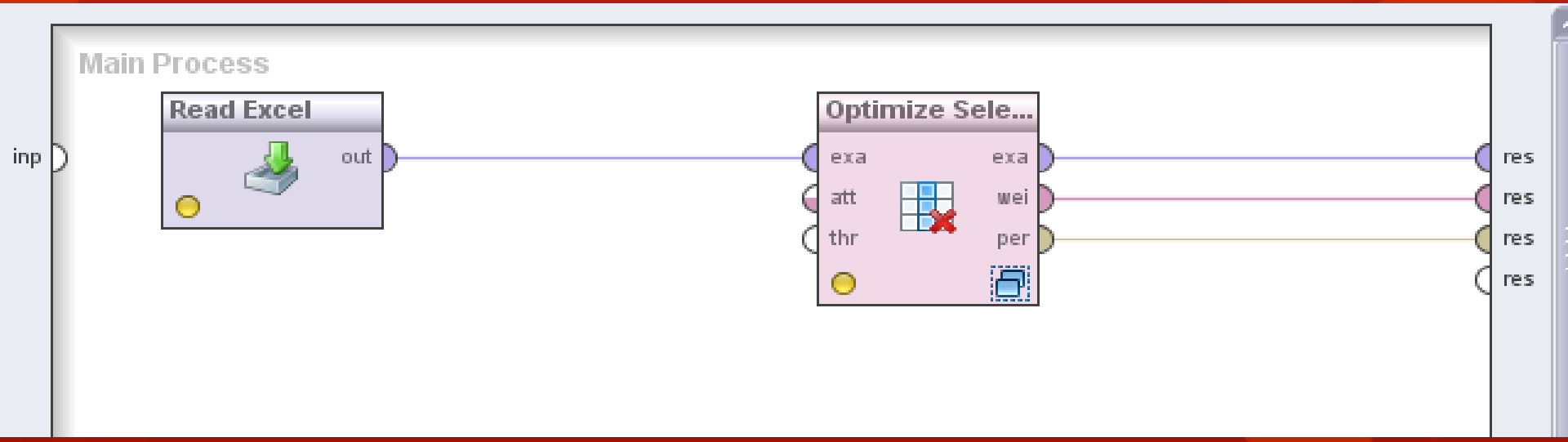
Model Yang Diusulkan



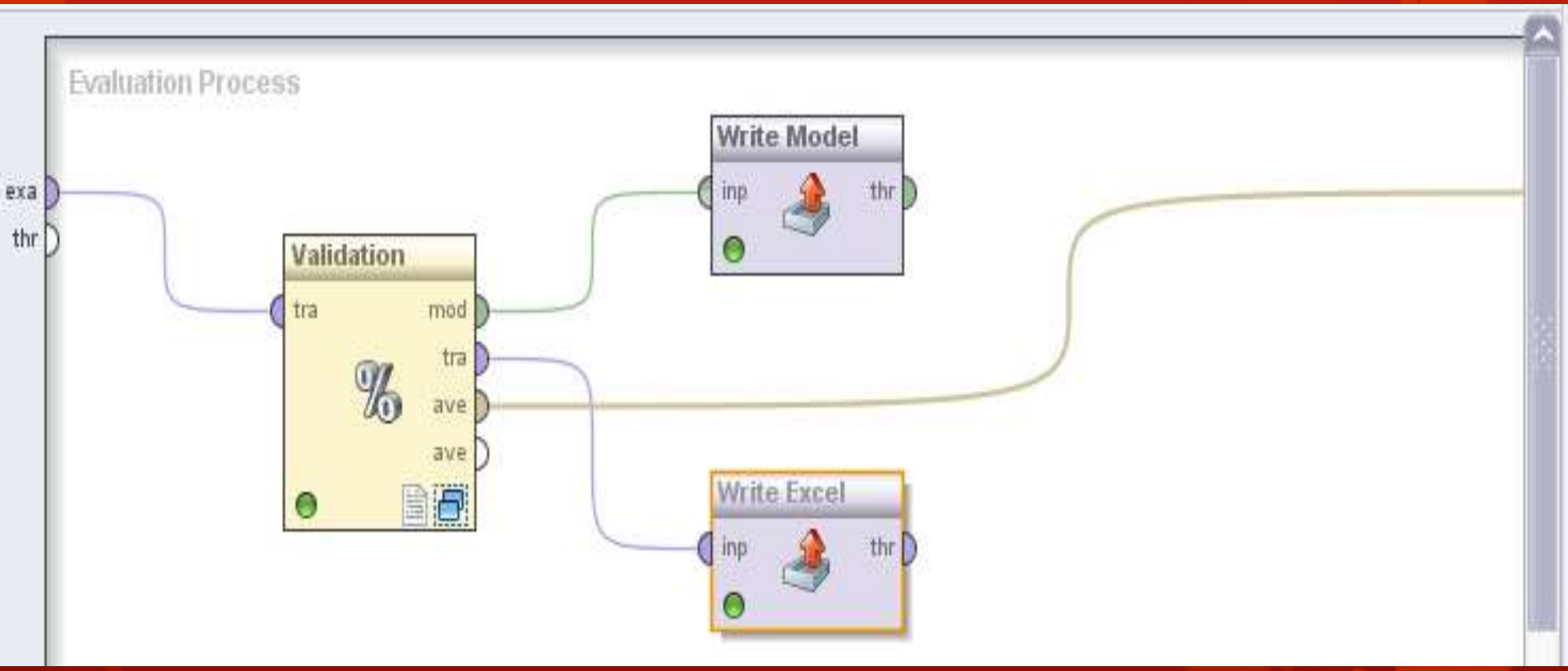
Pengujian Model

- Dalam melakukan penelitian ini diperlukan proses pengujian model yang diusulkan.
- Proses pengujian model menggunakan bagian dari dataset yang ada.
- Semua dataset kemudian diuji dengan metode yang diusulkan pada aplikasi Rapid Miner 5

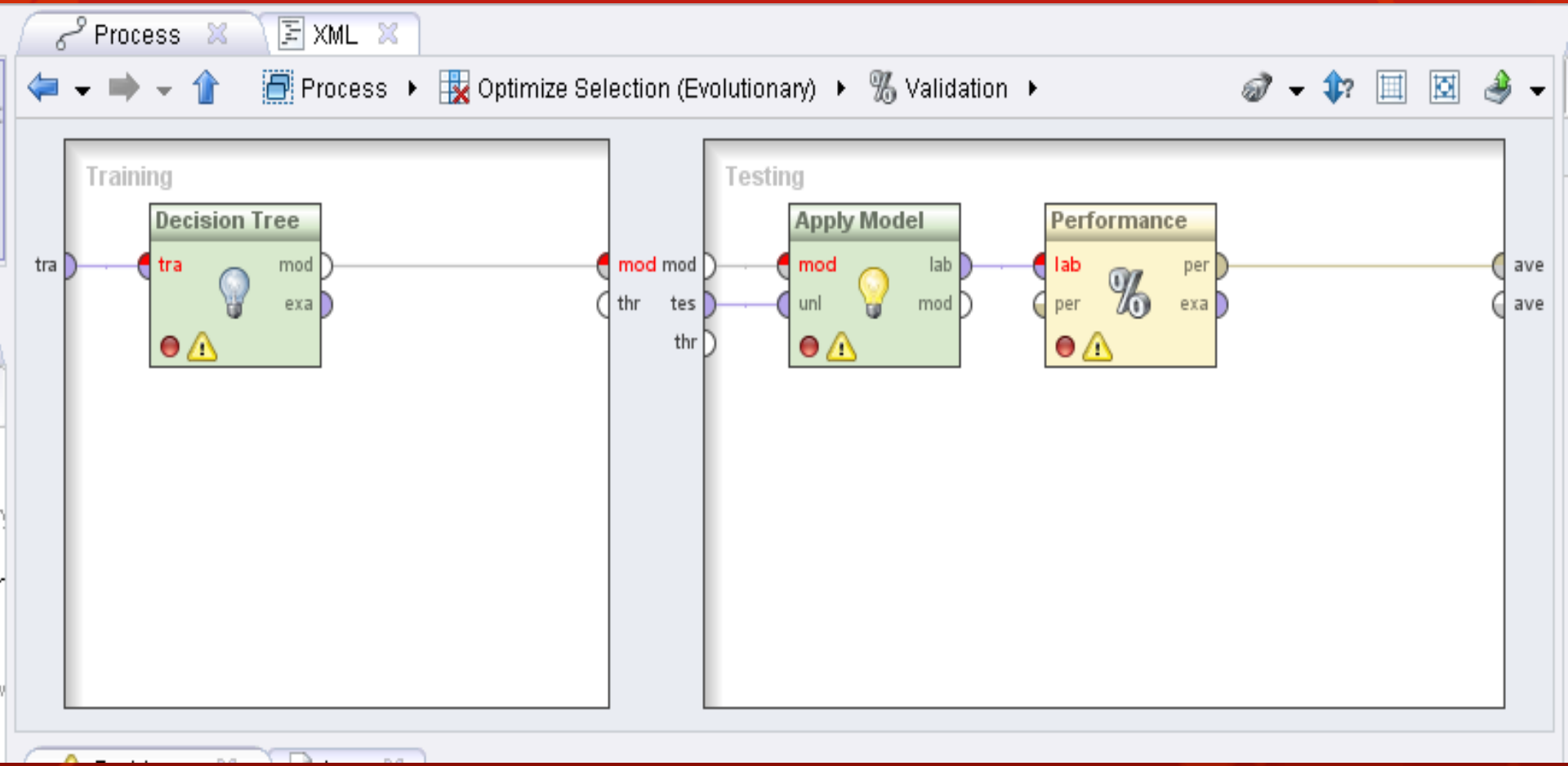
Pengujian Model



Pengujian Model



Pengujian Model



Hasil Penelitian

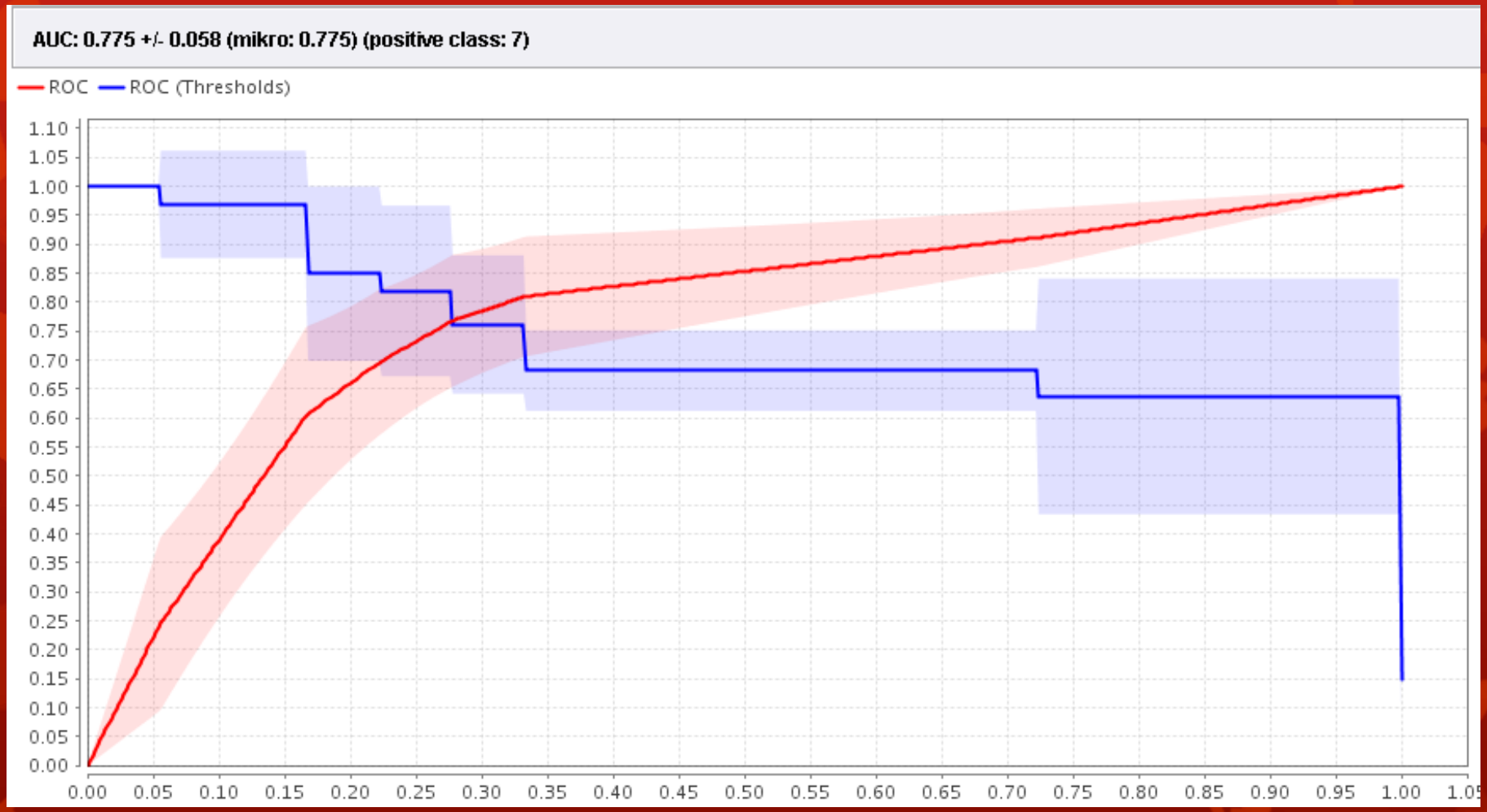
	Algoritma C4.5	Algoritma C4.5 + GA
Akurasi Model	77.77%	77.81%
AUC	0.775	0.777

Confusion Matrix

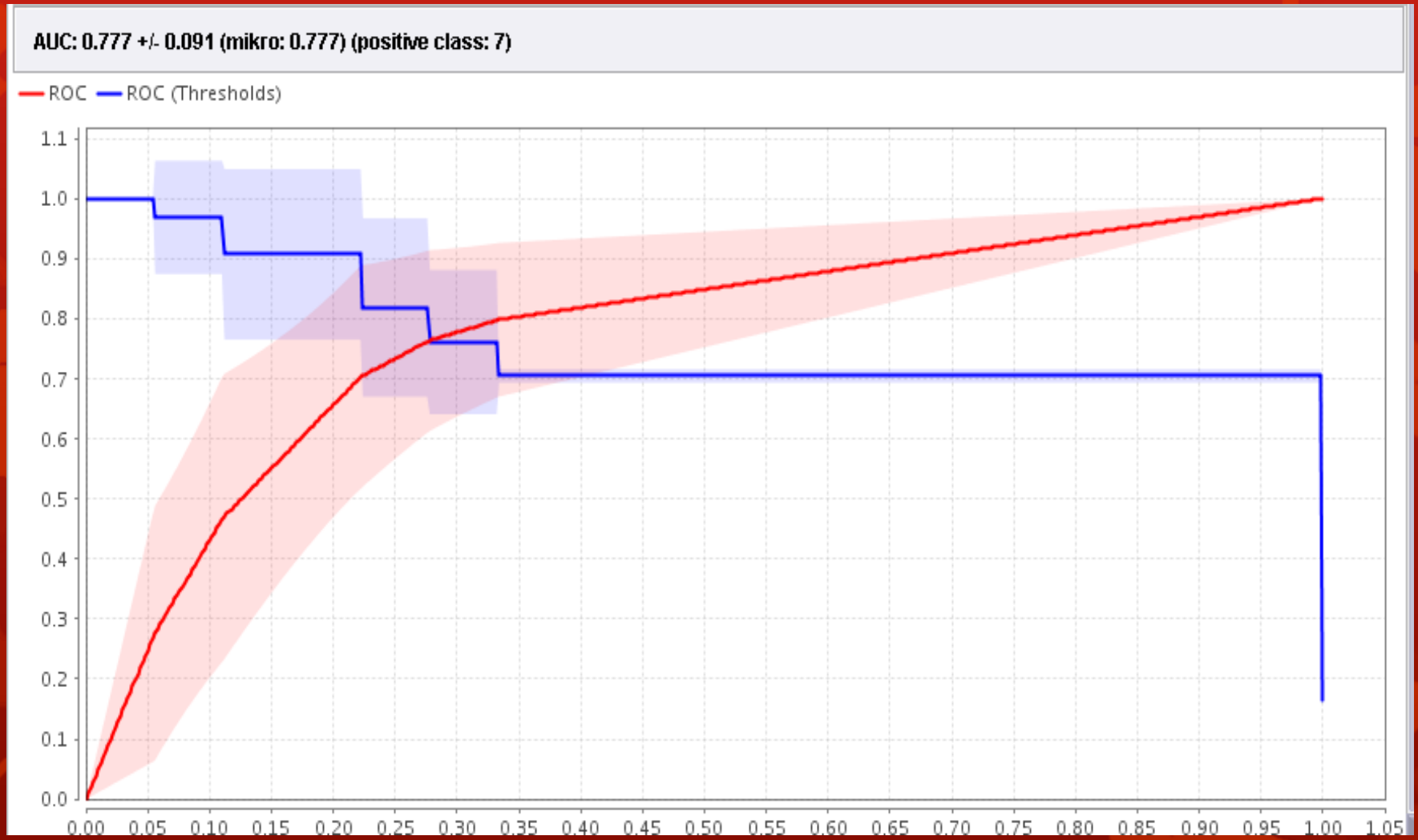
Akurasi Algoritma C4.5 **77.77% +/-5.38% (mikro:77.74%)**			
	true 3	**true 7**	***class precision***
pred. 3	141	28	83.43%
pred. 7	39	93	70.45%
class recall	78.33%	76.86%	

Akurasi Algoritma C4.5 (GA) **77.81% +/- 8.70% (mikro:77.74%)**			
	true 3	**true 7**	***class precision***
pred. 3	141	28	83.43%
pred. 7	39	93	70.45%
class recall	78.33%	76.86%	

AUC untuk model algoritma C4.5



AUC untuk model algoritma C4.5 yang dioptimasi dengan GA



AUC untuk model algoritma C4.5 yang dioptimasi dengan GA

- Hasil yang diperoleh dari metode C4.5 + GA, yakni dengan nilai **AUC: 0,777**. Dari 4 nilai untuk *confusion matrix* menghasilkan tabel sebagai berikut:

	Kenyataan F	**Kenyataan T**
Prediksi F	141	28
Prediksi T	39	93

AUC untuk model algoritma C4.5 yang dioptimasi dengan GA

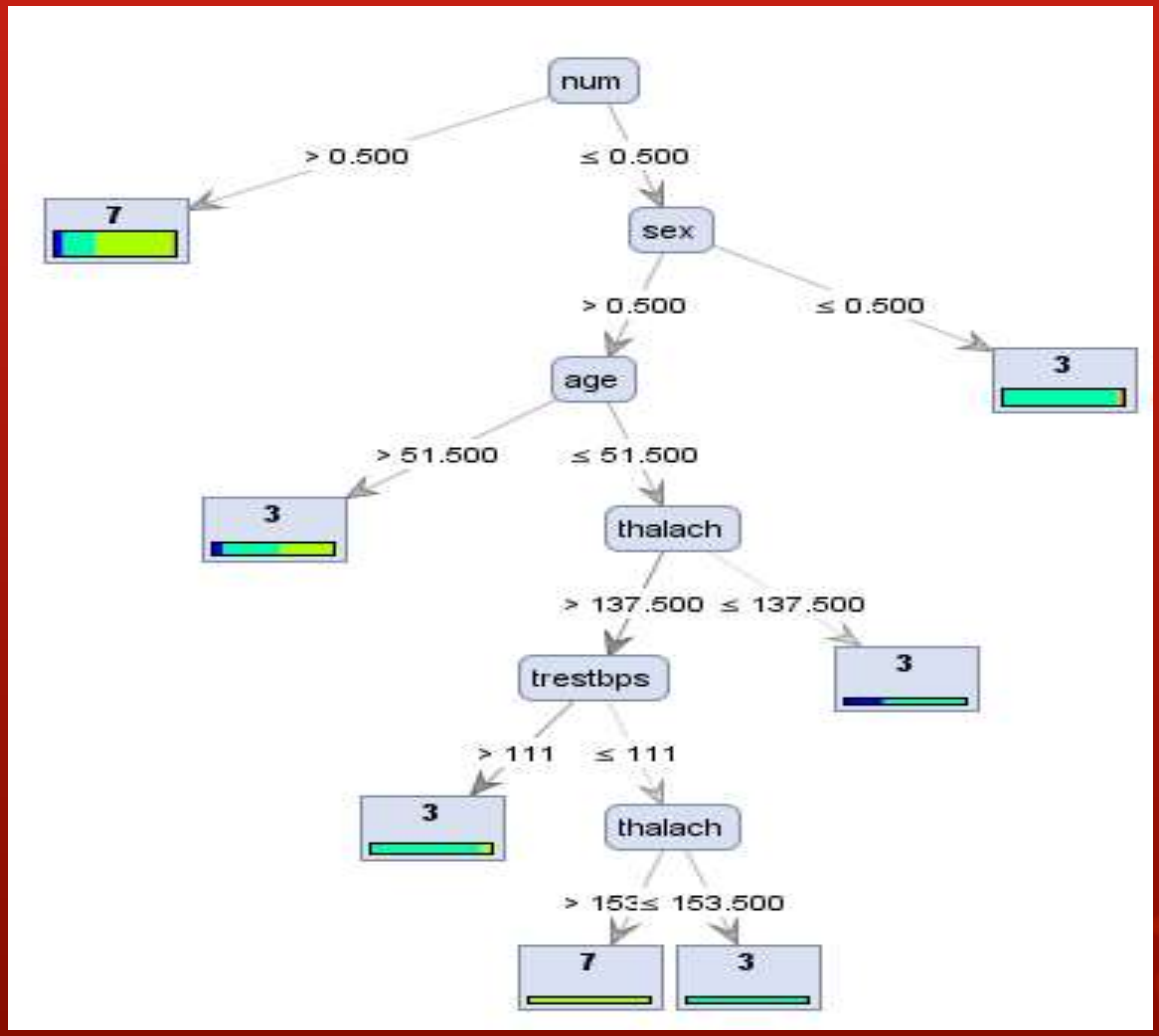
- Nilai akurasi dari *confusion matrix* tersebut adalah sebagai berikut:

$$akurasi = \frac{(TN + TP)}{(TN + FN + TP + FP)}$$

$$= \frac{(141 + 93)}{(141 + 28 + 93 + 39)}$$

$$= 0{,}7774 = 77{,}74\%$$

Model Setelah Seleksi Atribut



Interpretasi Model Setelah Optimasi

- R1: IF num > 0.5 THEN class = 7
- R2: IF num ≤ 0.5 AND sex ≤ 0.5 THEN class = 3
- R3: IF num ≤ 0.5 AND sex > 0.5 AND age > 51.5 THEN class = 3
- R4: IF num ≤ 0.5 AND sex > 0.5 AND age ≤ 51.5 AND thalach ≤ 137.5 THEN class = 3
- R5: IF num ≤ 0.5 AND sex > 0.5 AND age ≤ 51.5 AND thalach > 137.5 AND trestbps > 111 THEN class = 3
- R5: IF num ≤ 0.5 AND sex > 0.5 AND age ≤ 51.5 AND thalach > 137.5 AND trestbps ≤ 111 AND thalach < 153.5 THEN class = 7
- R5: IF num ≤ 0.5 AND sex > 0.5 AND age ≤ 51.5 AND thalach > 137.5 AND trestbps ≤ 111 AND thalach ≤ 153.5 THEN class = 3

Kesimpulan

- Dari penelitian yang dilakukan model yang terbentuk dengan algoritma C4.5 sendiri sudah memiliki akurasi yang cukup baik yaitu sebesar 77.77%
- Dengan proses seleksi atribut oleh algoritma genetika, model yang terbentuk dapat ditingkatkan lagi menjadi 77.81% dalam pengklasifikasian penyakit jantung.

Perbandingan C4.5 dg C4.5+GA

	Algoritma C4.5	Algoritma C4.5 + GA
Sukses prediksi pasien sakit	93	93
Sukses prediksi pasien sehat	141	141
Akurasi Model	77.77%	77.81%
AUC	0.775	0.777

Saran

- Untuk menghasilkan prediksi yang lebih akurat dari penelitian ini, diperlukan pembersihan *(data cleansing)* dari masukan data yang tidak konsisten dan data yang rusak atau yang disebut dengan data sampah, pada tahap pengolahan awal.
- Menggunakan teknik atau metode optimisasi lain, seperti *Particle Swarm Optmization, Backward Elimination, Forwad Selection*, atau yang lainnya. Dengan optimisasi lain, mungkin dapat menghasilkan nilai yang lebih akurat dan lebih baik
- Menambahkan jumlah data yang lebih besar dan atribut yang lebih banyak, sehingga hasil pengukuran yang akan didapatkan lebih baik lagi.