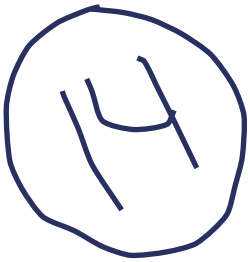
Konferensi Nasional Ilmu Sosial & Teknologi (KNiST)
Maret 2017, pp. 7~12

DIAGNOSA PREDIKSI PENYAKIT JANTUNG DENGAN MODEL ALGORITMA *NAÏVE BAYES* DAN ALGORITMA C4.5



Tri Retnasari[1], Eva Rahmawati[2]

[1]STMIK Nusa Mandiri Sukabumi
e-mail: retna3sari@gmail.com

[2]STMIK Nusa Mandiri Sukabumi
e-mail: eva.rijal@gmail.com

Abstrak

WHO menyebutkan penyakit jantung merupakan penyakit pembunuh orang didunia nomor 1, yang tentu saja telah merenggut banyak nyawa di berbagai belahan dunia. Dari data yang diperoleh di tahun 2015 saja ada 15 juta orang yang ada didunia meninggal akibat penyakit jantung, jika dipersentasikan pada tahun tersebut dengan total keseluruhan kematian di tahun tersebut jumlanya sekitar 54%. Tentu saja dengan data tersebut penyakit jantung merupakan penyakit yang mengkhawatirkan dan perlu diwaspadai oleh setiap orang. Sekitar 35 persen kematian di Indonesia disebabkan oleh penyakit jantung. Menurut Federasi Jantung Dunia, angka kematian akibat penyakit jantung koroner di Asia Tenggara mencapai 1,8 juta kasus pada tahun 2014. Banyak penelitian yang dilakukan untuk mendiagnosa pasien dengan benar namun belum diketahui metode apa yang akurat dalam memprediksi penyakit jantung. Dalam penelitian ini dilakukan komparasi algoritma Naïve Bayes dan C4.5 untuk mengetahui algoritma mana yang paling akurat dalam memprediksi penyakit jantung. Hasil pengujian kedua algoritma tersebut diketahui bahwa Algoritma Naïve Bayes memiliki nilai akurasi yang paling tinggi yaitu 86.67% sedangkan algoritma C4.5 memiliki nilai akurasi 83.70%. Dengan demikian algoritma Naïve Bayes dapat memprediksi penyakit jantung lebih baik. Manfaat bagi peneliti mampu membandingkan nilai akurasi metode data mining dengan metode yang lainnya

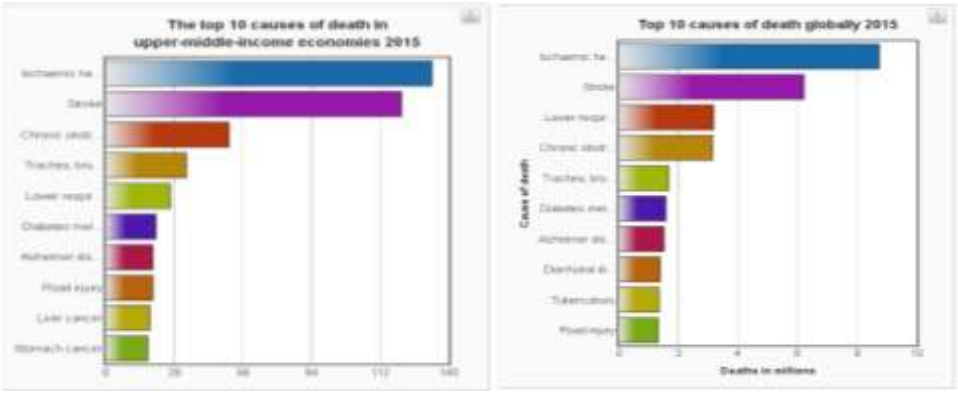
***Keywords**:* Penyakit Jantung, algoritma Naïve Bayes, Algoritma C4.5

1. Pendahuluan

Jantung merupakan salah satu organ terpenting dalam tubuh. Organ berukuran sebesar kepalan tangan ini berfungsi memompa dan menyebarkan darah yang mengandung oksigen ke seluruh tubuh. Berbagai penelitian telah banyak dilakukan untuk mengenali ciri-ciri awal seseorang terkena penyakit jantung, karena penyakit jantung bisa dialami dari mulai bayi, remaja, dewasa dan juga orangtua.

Ada banyak jenis gangguan pada jantung atau macam-macam penyakit jantung yang perlu diketahui dan diwaspadai. Banyak masyarakat awam yang menganggap jika penyakit jantung itu hanya penyakit jantung koroner, padahal penyakit jantung koroner adalah salah satu dari jenis penyakit jantung. (Wajhillah, 2014), berdasarkan dari data yang didapatkan dari (Organization, 2015), kemungkinan jumlah penderita jantung di Indonesia akan meningkat.

Diterima 22 Februari 2017; Revisi 06 Maret 2017; Disetujui 15 Maret 2017



Sumber: (Organization, 2015)

Gambar 1. Grafik 10 Penyakit Penyebab kematian di Negara Maju dengan Pendapatan Tinggi dan Rendah

Penelitian yang telah dilakukan untuk mendiagnosa pasien diantaranya:

a. (Wajhillah, 2014) menggunakan algoritma C4.5 untuk mengoptimasi prediksi penyakit jantung. Hasil penelitian bahwa nilai akurasi algoritma klasifikasi C4.5 senilai 81,25%, sedangkan untuk nilai akurasi algoritma klasifikasi C4.5 berbasis PSO sebesar 93,75%.
b. (Soni, Ansari, Sharma, & Soni, 2011) menggunakan *Decision Tree and Bayesian Classification* untuk meningkatkan dan mengurangi ukuran data aktual agar mendapatkan subset optimal atribut yang cukup saat memprediksi penyakit jantung. Hasil penelitian adalah bahwa keakuratan *Decision Tree* dan *Naïve Bayes* lebih meningkatkan setelah menerapkan algoritma genetika untuk mengurangi ukuran data aktual untuk mendapatkan bagian yang optimal dari atribut yang cukup untuk prediksi penyakit jantung.
c. (Pramunendar, Dewi, & Asari, 2013) menggunakan Algoritma *Back Propagation* Neural Network dengan Metode Adaboost. Hasil penelitian bahwa nilai akurasi hasil prediksi menggunakan algoritma BPNN adalah 96,65 % dan algoritma BPNN dengan metode Adaboost menjadi 99,29 %.

Neural network adalah satu set unit *input/output* yang terhubung dimana tiap relasinya memiliki bobot (Ramdhani, 2016; Han, 2006). *Neural Network* dimaksudkan untuk mensimulasikan perilaku sistem biologi susunan syaraf manusia (Alpaydin, 2010). Naive Bayes merupakan metode klasifikasi populer dan masuk dalam sepuluh algoritma terbaik dalam data mining, algoritma ini juga dikenal dengan nama Idiot's Bayes, Simple Bayes, dan Independence Bayes (Bramer, 2007). Menurut (Han, 2006) Algoritma C4.5 merupakan bagian dari kelompok algoritma *decision trees* dan merupakan kategori 10 algoritma yang paling populer.

Berdasarkan beberapa penelitian tersebut diatas, untuk menangani kelemahan-kelemahan yang masih ada maka akan diterapkan komparasi algoritma Naïve Bayes dan C4.5 untuk mengetahui algoritma mana yang paling akurat dalam memprediksi penyakit jantung. Penerapan algoritma Naïve Bayes dan C4.5 dalam prediksi awal penyakit jantung dengan menggunakan data set yang digunakan berasal dari UCI *Machine Learning Repository*. Hasil prediksi awal yang didapatkan dapat digunakan oleh para petugas medis sebagai alat bantu dalam penentuan penyakit jantung dan langkah awal penanganannya.

2. Metode Penelitian

Tahapan dalam kerangka penelitian dibagi menjadi 4 bingkai dasar, yaitu:

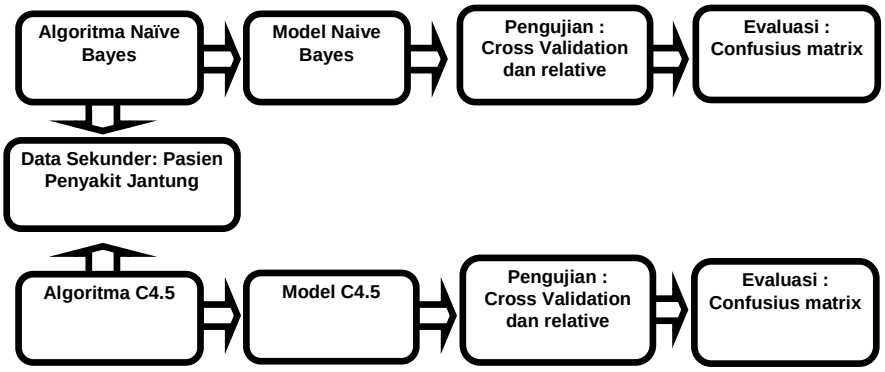
a. Pengumpulan Data.
   Pada penelitian ini, merupakan data sekunder. Dalam pengumpulan data sumber dapat di peroleh dari *University of California Irvine (UCI) Machine Learning Repository* untuk di jadikan objek penelitian dan mencari data tambahan melalui buku-buku, jurnal, publikasi dan lain-lain untuk di jadikan rujukan penulisan dan penelitian.
b. Pengolahan Awal Data.
   Dilakukan penyeleksian data, data dibersihkan dan ditransformasikan bentuk yang diinginkan sebelum dilakukan pembuatan model

c. Model atau Metode yang Diusulkan atau Dikembangkan.
Data yang diteliti dianalisa kemudian di kelompokan variabel mana yang berhubungan dengan satu sama lainnya, lalu dibuatkan model yang sesuai dengan jenis data. Pembagian data kedalam data latihan (*training* data) dan data uji (*testing* data) juga diperlukan untuk pembuatan model.

d. Eksperimen dan pengujian model atau metode pengujian model diusulkan pada model yang akan diuji untuk melihat hasil berupa *rule* yang akan dimanfaatkan dalam mengambil keputusan hasil penelitian.

e. Evaluasi dan Validasi Hasil.
Pada penelitian ini dilakukan evaluasi terhadap model yang ditetapkan untuk mengetahui tingkat keakurasian model.

Berikut ini adalah metode yang diusulkan dalam penelitian ini:



Gambar 2. Metode yang Diusulkan dalam Penelitian

3. Pembahasan

Pada penelitian ini, data kasus digunakan 270 data set penyakit jantung yang diperoleh dari *University of California Irvine (UCI) Machine Learning Repository* yang bersifat publik yang akan dibagi menjadi data *training* dan *testing*. Data set terdiri dari 13 atribut yaitu umur, jenis kelamin, jenis sakit dada, tekanan darah, kolestrol, kadar gula, elekrokardiografi, tekanan darah, angina induksi, *oldpeak*, segmen_st, *flaurosopy*, dan denyut jantung.

Data pasien penyakit jantung yang diambil dari UCI Machine Learning Repository berupa format *comma separated values* (CSV) seperti yang digambarkan pada tabel 1 berikut:

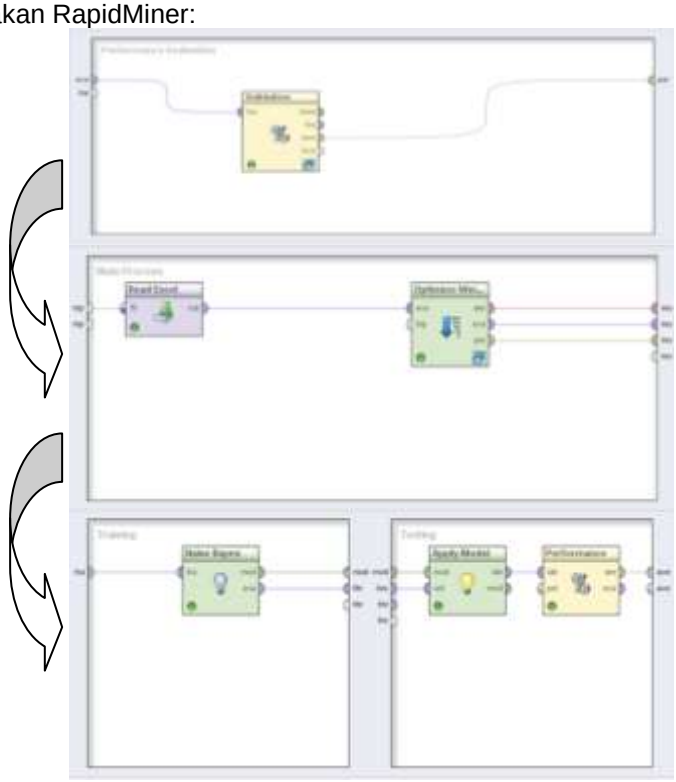
Tabel 1. Data pasien penyakit jantung

Sex	Chest pain type	Resting blood pressure	Serum cholestoral in mg/dl	Fasting blood sugar > 120 mg/dl	Resting electrocardiographic	Maximum heart rate achieved	Exercise induced angina	Oldpeak	The slope of the peak exercise ST segment	Number of major vessels (0-3) colored by flourosopy	Thal: 3 = normal; 6 = fixed defect; 7 = reversable defect	Variabel yang akan diprediksi
1.0	4.0	130.0	322.0	0.0	2.0	109.0	0.0	2.4	2.0	3.0	3.0	2
0.0	3.0	115.0	564.0	0.0	2.0	160.0	0.0	1.6	2.0	0.0	7.0	1
1.0	2.0	124.0	261.0	0.0	0.0	141.0	0.0	0.3	1.0	0.0	7.0	2
1.0	4.0	128.0	263.0	0.0	0.0	105.0	1.0	0.2	2.0	1.0	7.0	1
0.0	2.0	120.0	269.0	0.0	2.0	121.0	1.0	0.2	1.0	1.0	3.0	1
1.0	4.0	120.0	177.0	0.0	0.0	140.0	0.0	0.4	1.0	0.0	7.0	1
1.0	3.0	130.0	256.0	1.0	2.0	142.0	1.0	0.6	2.0	1.0	6.0	2
1.0	4.0	110.0	239.0	0.0	2.0	142.0	1.0	1.2	2.0	1.0	7.0	2
1.0	4.0	140.0	293.0	0.0	2.0	170.0	0.0	1.2	2.0	2.0	7.0	2
0.0	4.0	150.0	407.0	0.0	2.0	154.0	0.0	4.0	2.0	3.0	7.0	2
1.0	4.0	135.0	234.0	0.0	0.0	161.0	0.0	0.5	2.0	0.0	7.0	1
1.0	4.0	142.0	226.0	0.0	2.0	111.0	1.0	0.0	1.0	0.0	7.0	1
1.0	3.0	140.0	235.0	0.0	2.0	180.0	0.0	0.0	1.0	0.0	3.0	1
1.0	1.0	134.0	234.0	0.0	0.0	145.0	0.0	2.6	2.0	2.0	3.0	2
0.0	4.0	128.0	303.0	0.0	2.0	159.0	0.0	0.0	1.0	1.0	3.0	1
0.0	4.0	112.0	149.0	0.0	0.0	125.0	0.0	1.6	2.0	0.0	3.0	1
1.0	4.0	140.0	311.0	0.0	0.0	120.0	1.0	1.8	2.0	2.0	7.0	2
1.0	4.0	140.0	203.0	1.0	2.0	155.0	1.0	3.1	3.0	0.0	7.0	2
1.0	1.0	110.0	211.0	0.0	2.0	144.0	1.0	1.8	2.0	0.0	3.0	1
1.0	1.0	140.0	199.0	0.0	0.0	178.0	1.0	1.4	1.0	0.0	7.0	1
1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	2.0	2.0	7.0	2
1.0	2.0	130.0	245.0	0.0	2.0	180.0	0.0	0.2	2.0	0.0	3.0	1
1.0	4.0	115.0	303.0	0.0	0.0	181.0	0.0	1.2	2.0	0.0	3.0	1

Sumber: (Repository, 2017)

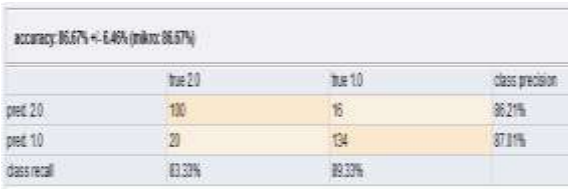
3.1. Model Naïve Bayes

Berikut adalah gambar pengujian data penyakit Jantung dengan algoritma Naïve Bayes menggunakan RapidMiner:



Gambar 3. Pengujian Penyakit jantung menggunakan algoritma Naïve Bayes

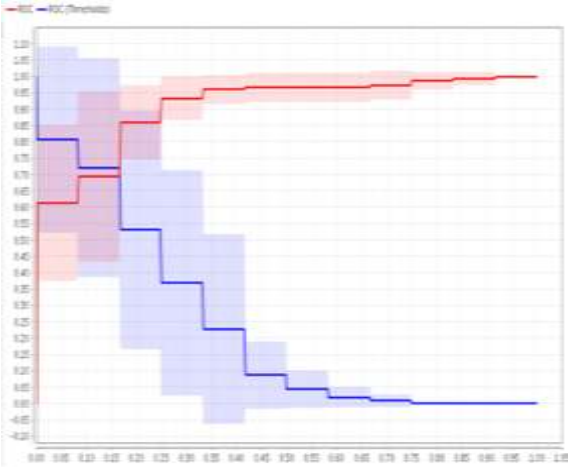
a. *Confusion Matrix*
Gambar 3. menunjukkan hasil dari *confusion matrix* model Naive Bayes. Berdasarkan gambar 3.2 dapat diketahui bahwa dari 270 data, 100 data diperediksi 2 (positif), kemudian 16 data diprediksi 2 (positif) tetapi ternyata hasilnya prediksi 1 (negatif). Kemudian 134 data 1 (negatif) diprediksi sesuai dengan prediksi yang dilakukan dengan model Naive Bayes, dan 20 data prediksi 1 (negatif) tetapi ternyata hasil prediksi nya 2 (positif).



accuracy: 86.67% +/- 6.46% (mikro: 86.67%)			
	true 2.0	true 1.0	class precision
pred. 2.0	100	16	86.21%
pred. 1.0	20	134	87.01%
class recall	83.33%	89.33%	

Gambar 4. Hasil pengujian *Confusion Matrix* untuk Model Naïve Bayes

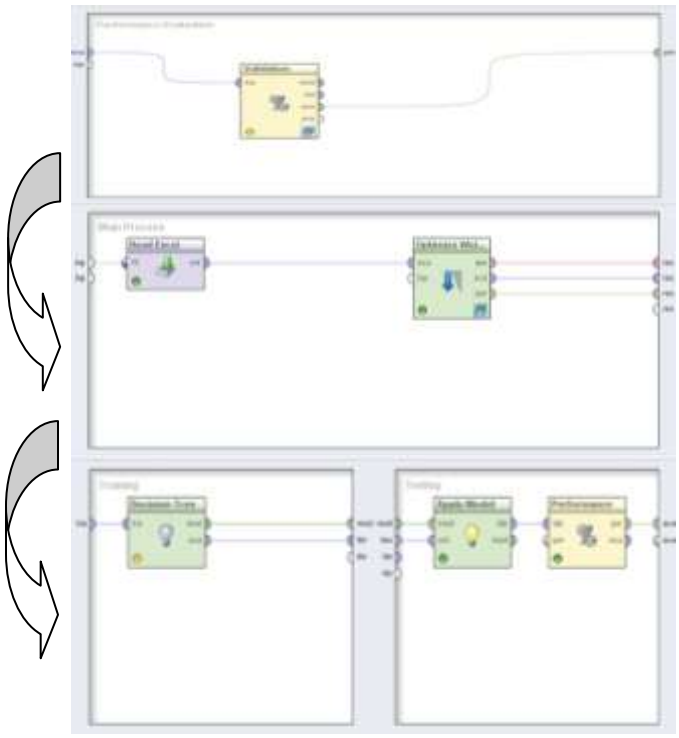
b. Kurva ROC
Hasil perhitungan divisualisasikan dengan kurva ROC. Perbandingan kedua class bisa dilihat pada gambar 5. yang merupakan kurva ROC untuk algoritma Naive Bayes. Menghasilkan nilai AUC (*Area Under Curve)* sebesar 0.909 dengan nilai akurasi Klasifikasi (*Excellent Classification).*



Gambar 5. Kurva ROC dengan model Naive Bayes

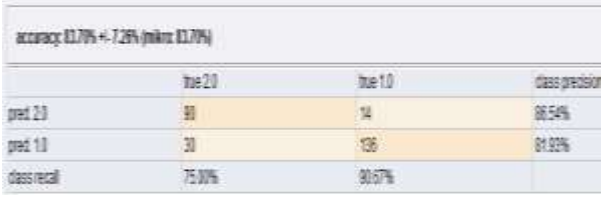
3.2. **Model C4.5**

3.3. Berikut adalah gambar pengujian data penyakit Jantung dengan algoritma C4.5 menggunakan RapidMiner:



Gambar 6. Pengujian Penyakit jantung menggunakan algoritma C 4.5

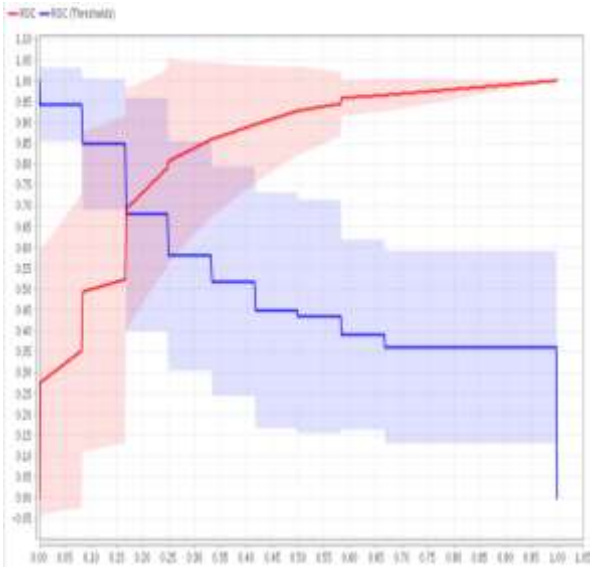
a. *Confusion Matrix*
Gambar 7. menunjukkan hasil dari *confusion matrix* model C 4.5. Berdasarkan gambar 3.5 dapat diketahui bahwa dari 270 data, 90 data diperediksi 2 (positif), kemudian 14 data diprediksi 2 (positif) tetapi ternyata hasilnya prediksi 1 (negatif). Kemudian 136 data 1(negatif) diprediksi sesuai dengan prediksi yang dilakukan dengan model C 4.5, dan 30 data prediksi 1 (negatif) tetapi ternyata hasil prediksi nya 2 (positif).



	true 2.0	true 1.0	class precision
pred. 2.0	90	14	86.54%
pred. 1.0	30	136	81.93%
class recall	75.00%	90.67%	

Gambar 7. Hasil pengujian *Confusion Matrix* untuk Model C4.5

b. Kurva ROC
Hasil perhitungan divisualisasikan dengan kurva ROC. Perbandingan kedua class bisa dilihat pada gambar 8. yang merupakan kurva ROC untuk algoritma C4.5. Menghasilkan nilai AUC (*Area Under Curve)* sebesar 0.834 dengan nilai akurasi Klasifikasi (*good classification*).



Gambar 8. Kurva ROC dengan model C4.5

Hasil analisis dari model algoritma Naïve Bayes dan C4.5 dirangkumkan dalam tabel dibawah ini.

Tabel 2. Komparasi Nilai *Accuracy* dan AUC

	Naïve Bayes	***C 4.5***
Accuracy	86.67%	83.70%
AUC	0.909	0.834

Tabel 4.1 membandingkan *Accuracy* dan AUC dari tiap model. Terlihat bahwa nilai *accuracy* dan AUC *Naïve Bayes* lebih tinggi dibandingkan *C 4.5.* Penerapan *Naïve Bayes* untuk prediksi penyakit jantung menghasilkan selisih nilai akurasi sebesar 2.97%. Untuk evaluasi menggunakan ROC *curve* sehingga menghasilkan nilai AUC (*Area Under Curve*) untuk model algoritma Naïve Bayes mengasilkan nilai 0.909 dengan nilai akurasi *Excellent Classification*, sedangkan untuk algoritma C 4.5 menghasilkan nilai 0.834 dengan nilai akurasi *Good Classification*, dan selisih nilai keduanya sebesar 0.075.

4. Simpulan

Meskipun diketahui bahwa algoritma *Naïve Bayes* memiliki akurasi yang paling tinggi namun untuk penelitian selanjutnya dapat ditambahkan untuk meningkatkan akurasi dan mengurangi prosedur pemeriksaan medis sehingga biaya untuk tes jantung bisa lebih sedikit dan agar penelitian ini bisa ditingkatkan. Untuk penelitian berikutnya dapat dilakukan optimasi peningkatan nilai akurasi dan membandingkan metode yang lain dapat juga pembuatan aplikasi *Decission Support System* (DSS) dari metode Algoritma Naïve Bayes.

Referensi

Alpaydin, E. (2010). *Introduction to Machine Learning.* London: The MIT Press.

Bramer, M. (2007). *Principles of Data Mining.* London: Springer.

Han, J. a. (2006). *Data Mining Concepts adn Techniques.* San Fransisco: Morgan Kauffman.

Organization, W. H. (2015). *http://www.who.int/mediacentre/factsheets/fs310/en/index1.html*. Retrieved Maret 06, 2017, from http://www.who.int/mediacentre/factsheets/fs310/en/index.html : http://www.who.int

Pramunendar, R., Dewi, I., & Asari, H. (2013). Penentuan Prediksi Awal Penyakit Jantung Menggunakan Algoritma Back Propagation Neural Network dengan Metode Adaboost. *SEMANTIK, ISBN:979-26-0266-6*, 298-304.

Ramdhani, Y. (2016). KOMPARASI ALGORITMA LDA DAN NAÏVE BAYES DENGAN OPTIMASI FITUR UNTUK KLASIFIKASI CITRA TUNGGAL PAP SMEAR. *INFORMATIKA*, *2*(2).

Repository, U. M. (2017). *https://archive.ics.uci.edu/ml/datasets/Heart+Disease*. Retrieved Febuari 20, 2017, from https://archive.ics.uci.edu/ml/datasets.html: https://archive.ics.uci.edu.

Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *International Journal of Computer Applications (0975– 8887)*, 43-48.

Wajhillah, R. (2014). Optimasi Algoritma Klasifikasi c4.5 Berbasis Particle Swarm Optimization Untuk Prediksi Penyakit Jantung. *Swabumi vol I No. 1,* , 26-36.