PENERAPAN PARTICLE SWARM OPTIMIZATION (PSO) UNTUK SELEKSI ATRIBUT DALAM MENINGKATKAN AKURASI PREDIKSI DIAGNOSIS PENYAKIT HEPATITIS DENGAN METODE ALGORITMA C4.5

Lis Saumi Ramdhani
Program Studi Manajemen Informatika
Akademi Manajemen Informatika dan Komputer BSI Sukabumi
Jl. Cemerlang No. 8 Sukakarya, Sukabumi
E-mail: lis.lud@bsi.ac.id

ABSTRACT

Hepatitis is a chronic disease that is chronic, at which time the person has been infected, the condition is still healthy and not showing signs and symptoms Typical but transmission continues to run. So from that process are still many people who do not recognize the symptoms of hepatitis. There have been many researchers who conducted the study to predict hepatitis, one of which applies the method C4.5. In this research, C4.5 algorithm optimization using Particle Swarm Optimization to improve prediction accuracy. After testing the two models namely the algorithm C4.5 and C4.5 Optimization using Particle Swarm Optimization, the results obtained are algorithms. Thus obtained test using values obtained C4.5 where accuracy is 79,33% and the AUC value is 0,655, while Optimization testing using C4.5Particle Swarm Optimization with accuracy values obtained 85,00% and AUC values were 0,718 at the level of diagnosis fair classification. So that the two methods have different levels of accuracy that is equal to 5,67% and the difference in AUC value of 0,063.

Keyword*s: Hepatitis, Algoritma C4.5, Selection Attributes, Particle Swarm Optimization*

I. PENDAHULUAN

Hati adalah salah satu organ yang paling penting dari tubuh manusia yang memiliki pengaruh yang tinggi terhadap kinerja bagian tubuh lainnya. Penyakit hati yang meliputi berbagai jenis penyakit hepatitis yang serius berbahaya dan fatal (Shariati and Haghighi, 2010).

Penyakit hepatitis merupakan masalah kesehatan masyarakat di dunia termasuk di Indonesia, yang terdiri dari Hepatitis A, B, C, D dan E. Hepatitis biasanya berhubungan dengan perilaku hidup bersih dan sehat. Penyakit hepatitis disebabkan oleh infeksi (virus, bakteri, parasit), infeksi tersebut bisa berasal dari lingkungan yang kurang bersih dan makanan serta minuman yang kurang bersih pula. Selain disebabkan oleh infeksi bisa juga disebabkan oleh obat-obatan, konsumsi alkohol maupun lemak yang berlebih (Kementerian Kesehatan RI, 2014)

Penelitian untuk memprediksi penyakit hepatitis, seperti penelitian yang dilakukan oleh Rouhani dan Haghighi (2009) tentang diagnosis penyakit hepatitis oleh Support Vector Mechine (SVM) dan Artificial Neural Network (ANN), penelitian selanjutnya yang dilakukan oleh Shariati dan Haghighi (2010) tentang Komparasi dari ANFIS Neural Network dan Support Vector Machine dimana yang lebih akurat untuk memprediksi penyakit hepatitis yaitu SVM. Penelitian yang dilakukan Sathyadevi (2011) tentang aplikasi dari algoritma cart dalam mendiagnosis penyakit hepatitis dan didalamnya tidak hanya menggunakan algoritma cart saja tetapi sekaligus membandingkan algoritma *decision tree* yang lain seperti algoritma ID3 dan algoritma C4.5. Penelitian yang dilakukan oleh Septiami (2014) tentang penggunaan algoritma C4.5 untuk diagnosis penyakit hepatitis.

Berdasarkan penelitian tersebut diatas, untuk menangani kelemahan-kelemahan yang masih ada maka akan diterapkan algoritma *decision tree* C4.5 dipadukan dengan *Particle Swarm Optimization (PSO)* yang akan digunakan untuk optimasi seleksi atribut untuk meningkatkan akurasi *decision tree* C4.5 dalam memprediksi penyakit hepatitis.

II. TINJAUAN PUSTAKA

a. Hepatitis

Hati adalah salah satu organ yang paling penting dari tubuh manusia yang memiliki pengaruh yang tinggi terhadap kinerja bagian tubuh lainnya. Penyakit hati yang meliputi berbagai jenis penyakit hepatitis yang serius berbahaya dan fatal (Shariati and Haghighi, 2010).

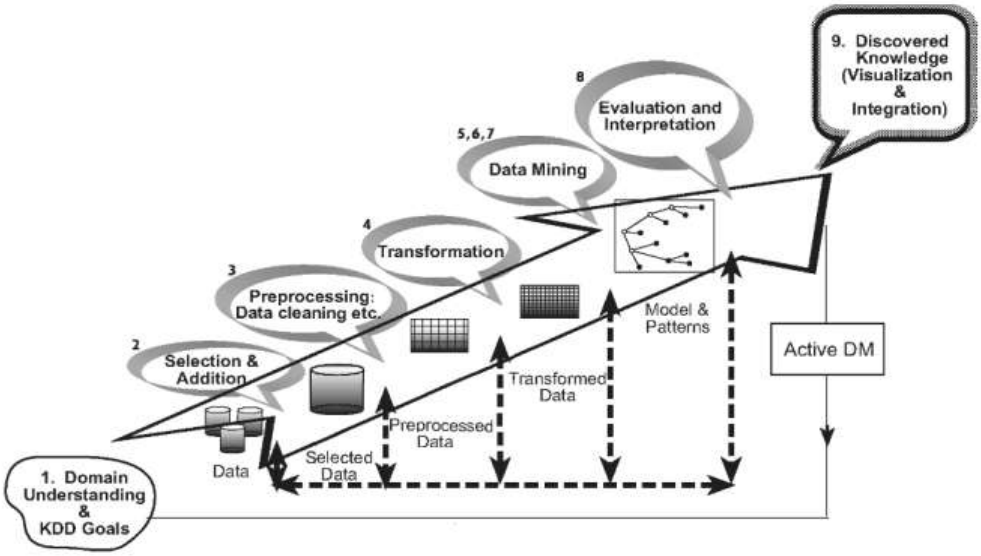
Hati adalah salah satu jenis organ yang sangat vital di tubuh kita. Sama halnya dengan

jenis-jenis organ vital lainnya seperti jantung, paru-paru, pankreas, ginjal, dsb, hati mempunyai peran yang sangat penting untuk tubuh kita dan jika terdapat gangguan fungsi hati sedikit saja maka tubuh akan merasakan dampaknya. Hepatitis biasanya dipakai untuk semua jenis peradangan pada sel-sel hati, yang bisa disebabkan oleh infeksi (virus, bakteri, parasit), obat-obatan (termasuk obat tradisional), konsumsi alkohol, lemak yang berlebih dan penyakit autoimmune (Kemenkes RI, 2014).

b. Knowladge Discovery in Database (KDD)

Data mining adalah proses menemukan pengetahuan yang menarik dari sejumlah data besar yang disimpan dalam database, data warehouse, atau repositori informasi lainnya (Han & Kamber, 2007). Data mining bisa dikatakan sebagai pencarian otomatis pola dalam basis data besar, menggunakan teknik komputasional campuran dari statistik, pembelajaran mesin dan pengenalan pola (Prasetyo, 2014 p.4).

Ada istilah lain yang mempunyai makna yang sama dengan data mining yaitu *knowledge discovery in database* (KDD). KDD bertujuan untuk memanfaatkan data dalam basis data dengan mengolahya sehingga menghasilkan informasi baru yang berguna.



Sumber: (Maimon & Rokach, 2010 p.3)

Gambar 1. Tahapan Proses *Knowladge Discovery in Database*

c. Algoritma C4.5

Algoritma C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan. Pohon keputusan merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang mempresentasikan aturan (Kusrini & Lutfhi, 2009).

Ada beberapa tahapan dalam membuat sebuah *decision tree* dalam algoritma C4.5 (Larose, 2005) yaitu :

1. Mempersiapkan data *training*. Data *training* biasanya diambil dari data histori yang pernah terjadi sebelumnya atau disebut data masa lalu dan sudah dikelompokkan dalam kelas-kelas tertentu.
2. Menghitung akar dari pohon. Akar akan diambil dari atribut yang akan terpilih, dengan cara menghitung nilai gain dari masing-masing atribut, nilai gain yang paling tinggi yang akan menjadi akar pertama. Sebelum menghitung nilai gain dari atribut, hitung dahulu nilai *entropy*.

Untuk menghitung nilai *entropy* digunakan rumus :

$$Entropy(S) = \sum_{i=0}^{n} - pi * \log_2 pi \quad (2.1)$$

Keterangan :
S= Himpunan kasus
n = jumlah partisi S
pi = proporsi Si terhadap S

Kemudian hitung nilai gain menggunakan rumus :

$$Gain(S, A) = Entropy(S) - \sum_{i=0}^{n} \frac{|Si|}{|S|} * Entropy(Si) \quad (2.2)$$

Keterangan :
S = Himpunan Kasus
A = Fitur
n = jumlah partisi atribut A
|Si| = Proporsi Si terhadap S
|S| = jumlah kasus dalam S

3. Ulangi langkah ke 2 dan langkah ke 3 hingga semua *record* terpatisi
4. Proses partisi *decision tree* akan berhenti saat :

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

d. Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) adalah teknik perhitungan evolusi yang dikembangkan oleh Kennedy dan Eberhart tahun 1995 (Cho, Lee & Jun, 2011). PSO dikembangkan dari kecerdasan berkelompok dan didasarkan pada penelitian perilaku gerakan kawanan burung dan ikan. Sementara mencari makanan, burung-burung yang baik tersebar atau pergi bersama-sama sebelum mereka menemukan tempat di mana mereka dapat menemukan makanan. Sementara burung mencari makanan dari satu tempat ke tempat lain, selalu ada burung yang bisa mencium makanan yang sangat baik, yaitu burung yang jelas dari tempat di mana makanan dapat ditemukan, memiliki informasi sumber daya makanan yang lebih baik. Karena mereka adalah transmisi informasi, khususnya informasi yang baik setiap saat ketika mencari makanan dari satu tempat ke tempat lain, yang dilakukan oleh informasi yang baik, burung-burung akhirnya akan berbondong-bondong ke tempat di mana makanan dapat ditemukan (Bai, 2010).

Pada algoritma PSO ini, pencarian solusi dilakukan oleh suatu populasi yang terdiri dari beberapa partikel. Populasi dibangkitkan secara random dengan batasan permasalahan yang dihadapi. Setiap partikel merepresentasikan partikel

atau solusi dari permasalahan yang dihadapi. Partikel tersebut mencari solusi yang optimal dengan melintasi ruang pencarian dengan cara partikel terkait melakukan penyesuaian terhadap posisi terbaik dari setiap partikel tersebut (*local best*) dan posisi partikel terbaik dari seluruh kawanan (*global best*) selama melintasi ruang pencarian. Jadi penyebaran informasi terjadi dalam partikel itu sendiri dan antara suatu partikel dengan partikel terbaik dari seluruh kawanan selama proses pencarian solusi. Setelah itu, dilakukan proses pencarian untuk mencari posisi terbaik setiap partikel dalam jumlah iterasi tertentu sampai didapatkan posisi relatif yang tetap (*steady*) atau mencapai batas iterasi yang telah ditetapkan. Pada setiap iterasi (t), setiap solusi yang direpresentasikan oleh posisi partikel i, dievaluasi performanya dengan cara memasukkan solusi tersebut ke dalam *fitness function*.

Setiap partikel diperlakukan seperti titik pada suatu dimensi ruang tertentu kemudian terdapat dua faktor yang memberikan karakter terhadap status partikel pada ruang pencarian yaitu posisi partikel (X) dan kecepatan partikel (Y). Formulasi matematika yang menggambarkan posisi dan kecepatan partikel suatu ruang dimensi tertentu sebagai berikut:

$$X_i(t) = x_{i1}(t), x_{i2}(t), \ldots, x_{in}(t) \quad (2.3)$$

$$Vi(t) = v_{i1}(t), v_{i2}(t), \ldots, v_{in}(t) \quad (2.4)$$

Persamaan di bawah (2.5) digunakan untuk menggambarkan kecepatan partikel baru berdasarkan kecepatan sebelumnya, jarak antara posisi saat ini dengan posisi partikel terbaik (*local best*), dan jarak antara posisi saat ini dengan posisi terbaik dalam kawanan (*global best*). Kemudian partikel terbang menuju posisi yang baru berdasarkan persamaan (2.6)

$$V_i(t) = v_{i1}(t-1) + c_1 r_1 (X_i^L - X_i(t-1) + c_2 r_2 (X_i^G - \ldots_i(t-1)) \quad (2.5)$$

$$X_{i\,(t)} = v_i(t) + X_i(t-1) \quad (2.6)$$

Dimana :
$V_i(t)$ = Kecepatan partikel ke-i pada iterasi ke-i
$X_{i\,(t)}$ = Posisi partikel saat ini pada partikel ke –i pada iterasi ke-i
t = *Iterasi*
X_i^L = *local best* dari particle ke-i
X^G = *global best* dari seluruh kawanan
c_1 = *learning factor*
c_2 = *learning factor*
r_1 = bilangan random yang bernilai anatar 0 sampai 1
r_2 = bilangan random yang bernilai anatar 0 sampai 1

e. Validasi (*K-Fold Cross Validation)*

Cross-Validasi adalah metode statistik mengevaluasi dan membandingkan algoritma pembelajaran dengan membagi data menjadi dua segmen: satu digunakan untuk belajar atau melatih model dan lainnya yang digunakan untuk memvalidasi model (Witten et al., 2011). Di dalam *cross validation*, set pelatihan dan set validasi harus *cross-over* secara berturut-turut sehingga setiap data memiliki kesempatan untuk tervalidasi. *K-Fold cross validation* merupakan bentuk dasar lintas validasi dimana kasus *k-fold cross validation* akan melibatkan putaran berulang sebanyak K *validation.* Misalnya pada kasus 10 *fold cross validation* maka data akan dibagi menjadi 10 set bagian, kemudian akan dilakukan 10 kali putaran (iterasi) untuk pengujian dan validasi.

f. Confusion Matrix (Accuracy)

Confusion Matrix (Gorunescu, 2011) adalah metode evaluasi model klasifikasi berdasarkan perhitungan objek testing, dimana data hasil prediksi ada diantara dua kelas (*mislabeling*) yaitu menghasilkan kelas positif dan kelas negatif. Metode ini menggunakan tabel matriks seperti pada Tabel 2.1 (Bramer, 2007):

Tabel 1 Model *Confusion matrix* (Bramer, 2007)

Klasifikasi yang benar	Diklasifikasikan sebagai	
	+	-
+	true positives	false negatif
-	false positives	true negatives

Pada tahap evaluasi menggunakan *confusion matrix* yang dilakukan menggunakan *tool rapid miner* akan diperoleh nilai *accuracy, sensitivity, specificity, PPV* dan *NPV.*
Akurasi dapat dihitung menggunakan rumus:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.7)$$

Dimana:
TP : Jumlah kasus positif yang diklasifikasikan sebagai positif
FP : Jumlah kasus negatif yang diklasifikasikan sebagai positif
TN : Jumlah kasus negatif yang diklasifikasikan sebagai negatif
FN : Jumlah kasus positif yang diklasifikasikan sebagai negatif

Sensitivitas dan spesifitas dapat digunakan sebagai ukuran statistik dari kinerja klasifikasi biner, sensitivitas dan spesifitas digunakan untuk mengukur model yang paling baik dan untuk memilih model yang paling efisien. Sensitivitas mengukur proporsi *true positive* yang diidentifikasikan dengan benar, spesifitas mengukur proporsi *true negative* yang diidentifikasikan dengan benar, dapat dihitung menggunakan rumus:

$$Sensitivity = \frac{Number\ of\ TruePositives}{Number\ of\ TruePositives + Number\ of\ False\ Negatives} \quad (2.8)$$

$$Specificity = \frac{Number\ of\ TrueNegatives}{Number\ of\ TrueNegatives + Number\ of\ False\ Positives} \quad (2.9)$$

Sensitivity juga dapat dikatakan *true positive rate* (TP *rate*) atau *recall*. Nilai *sensitivity* 100% berarti menunjukan bahwa pengklasifikasian mengakui sebuah kasus yang diamati positif. Misalnya semua orang yang memiliki penyakit jantung dinyatakan sakit.

Sedangkan untuk *PPV* (*Prediktive Positif Value*) adalah proporsi kasus dengan hasil diagnosa positif, *NPV* (*Prediktif Negative Value*) adalah proporsi kasus dengan hasil diagnosa negatif, dapat dihitung menggunakan rumus:

$$PPV = \frac{Number\ of\ TruePositives}{Number\ of\ TruePositives + Number\ of\ False\ Positives} \quad (2.10)$$

$$NPV = \frac{Number\ of\ TrueNegatives}{Number\ of\ TrueNegatives + Number\ of\ False\ Negatives} \quad (2.11)$$

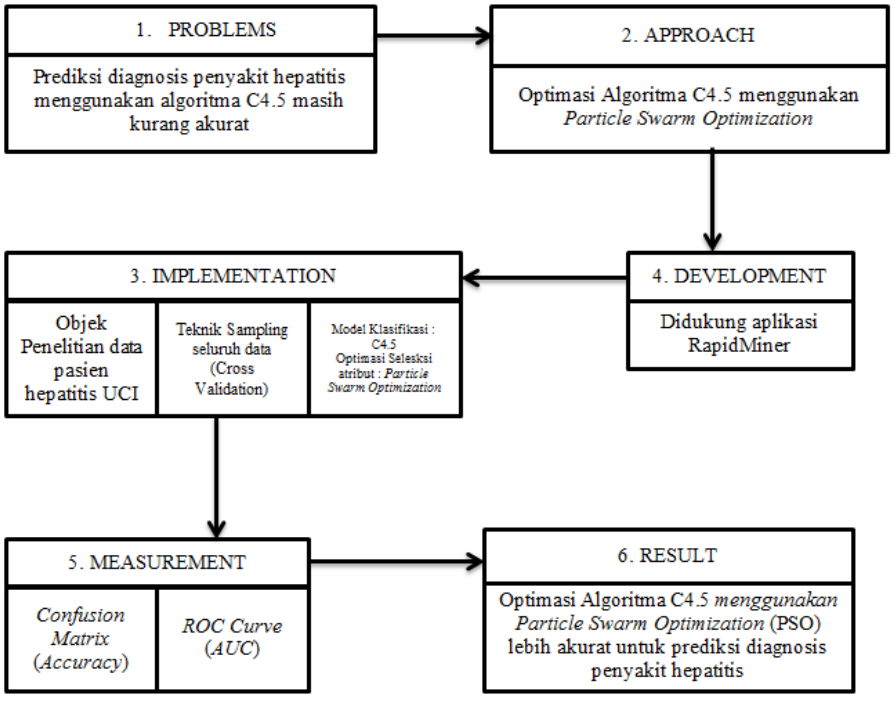
g. ROC Curve

Kurva ROC (*Receiver Operating Characteristic*) adalah ilustrasi grafis dari kemampuan diskriminan dan biasanya diterapkan untuk masalah klasifikasi biner (Fine, 2012), Secara teknik, kurva ROC juga disebut grafik ROC, grafik ROC terdiri dari dua dimensi grafik yaitu TP rate diletakan pada sumbu Y, sedangkan FP rate diletakan pada sumbu X. Untuk mengukur nilai grafik ROC, itu menggunakan teknik AUC (*Area Under Curve*), teknik ini dapat dibagi menjadi beberapa kelompok (Gorunescu, 2011)yaitu:

a. 0.90-1.00 = *Excellent Classification*
b. 0.80-0.90 = *Good Classification*
c. 0.70-0.80 = *Fair Classification*
d. 0.60-0.70 = *Poor Classification*
e. 0.50-0.60 = *Failure*

Kerangka Pemikiran

Model kerangka yang digunakan pada penelitian ini dapat digambarkan sebagai berikut:



Gambar 2. Kerangka Pemikiran

Tinjauan Studi

Penelitian yang dilakukan oleh Septiani (2014) yang berjudul Penerapan Algoritma C4.5 Untuk Prediksi Penyakit Hepatitis. Pada Penelitiannya membahas tentang penggunan algoritma C4.5 untuk memprediksi penyakit hepatitis. Hasil yang didapat dati penelitian ini sebesar 77, 92%.

Penelitian yang diakukan oleh Sathyadevi (2011) yang berjudul Application Of Cart Algorithm In Hepatitis Disease Diagnosis. Pada penelitian ini membahas penggunaan algoritma C4.5, algoritma ID3 dan algortima Cart untuk mengklasifikasi penyakit hepatitis dan membandingkan efektivitas, koreksi rate diantara ketiganya. Hasil yang didapat dari penelitian ini algoritma Cart sebesar 83,184% lebih akurat dibandingkan dengan algoritma ID3 dan algoritma C4,5.

Comparison of anfis neural network with several other anns and support vector machine for diagnosing hepatitis and thyoid, penelitian yang dilakukan oleh Shariati dan Haghighi (2010) dengan cara meningkatkan akurasi keseluruhan sistem diagnosis penyakit hepatitis dibandingkan penelitian sebelumnya. Akurasi terbaik keseluruhan sistem diagnosis untuk penyakit hepatitis sebelumnya adalah 97,6% menjadi 98,77%, maka ada peningkatan sebesar 1,2%.

Penelitian yang dilakukan oleh Rouhani dan Haghighi (2009) yang berjudul The Diagnosis of Hepatitis Disease by Support Vector Machines and Artificial Neural Networks. Penelitian ini menghasilkan akurasi keseluruhan lebih dari 96,4% untuk SVM dan Artificial Neural Network.

Tinjauan Objek Studi

Pada penelitian ini objek yang dijadikan bahan penelitian adalah *statlog database* yang berasal dari http://archive.ics.uci.edu/ml/datasets/Hepatitis sebagai subset dari dataset publik yang digunakan dalam proyek statlog eropa. Proyek tersebut melakukan perbandingan kinerja mesin pembelajaran, statistik dan algoritma jaringan syaraf tiruan pada dataset dari dunia nyata pada daerah industri termasuk dalam bidang kedokteran. Dataset yang digunakan terdiri dari 20 atribut.

III. METODE PENELITIAN

Jenis Penelitian

Jenis penelitian yang digunakan dalam penelitian ini menggunakan metode *experiment*, yaitu penelitian yang melibatkan penyelidikan kepada beberapa variable menggunakan tes tertentu yang dikendalikan sendiri oleh peneliti. Metode yang digunakan untuk memprediksi adalah algoritma C4.5 dengan *particle swarm optimization* yang akan digunakan untuk melakukan optimasi seleksi atribut.

Pengumpulan Data

Metode pengumpulan data dibagi menjadi dua sumber data yaitu data primer dan data sekunder. Data primer yaitu data yang dikumpulkan dari sumbernya langsung, sedangkan data sekunder yaitu data yang dikumpulkan dari peneliti sebelumnya. Data yang digunakan pada penelitian ini menggunakan data sekunder. Data penelitian ini diambil dari data pasien hepatitis yang di dapat dari *University of California Irvine (UCI) Machine Learning Data Repository*., dengan jumlah 155 *record* yang terdiri dari 123 *record* (79,35%) pasien hidup dan 32 *record* (20,65%) pasien meninggal. Dengan jumlah 20 atribut yang terdiri dari:

Pengolahan Awal Data

Validasi Data

Data dalam penlitian ini diambil dari *UCI Machine Learning* yang terdiri dari 19 atribut *predictor* dan 1 atribut hasil. Pada dataset tersebut akan dilakukan validasi data dengan menghilangkan data *missing value*. Data yang didapat dari *UCI Machine Learning* sebanyak 155 *record*, dari data tersebut terdapat data *missing value* sebanyak 5 *record*. Kemudian data missing value tersebut akan dihilangkan agar data menjadi valid. Pengolahan data juga dapat berupa konversi nilai agar mempermudah pembentukan model.

Pemisahan Dataset *Training* dan Data *Testing*

Pemisahan dataset *training* dan *testing* pada penelitian ini data training an data testing akan dipisah dengan menggunakan 10 *fold cross validation.* Dataset tersebut akan dibagi menjadi 10 bagian dan akan dilakukan pengulangan sebanyak 10 pengulangan. Contoh pada iterasi ke-3, jika bagian ketiga dijadikan sebagai data testing maka sisa bagian lainnya akan digunakan sebagai data training. Pengambilan data tersebut dilakukan secara acak agar semua

data dapat menjadi data training juga menjadi data testing.

Metode Yang Diusulkan

Metode yang diusulkan pada penelitian ini adalah melakukan pengolahan dataset sehingga mendapatkan variabel-variabel yang telah terseleksi dengan menggunakan *Particle Swarm Optimization.* Kemudian menggunakan metode 10 cross validation yaitu data testing dan data training, Data training dan data testing yang atributnya telah dioptimasi digunakan pada tahapan selanjutnya. Atribut yang sama digunakan untuk menghasilkan akurasi data yang sesuai. Data training diuji dengan menggunakan metode algoritma C4.5 sehingga menghasilkan sebuah metode baru dalam proses prediksi Sedangkan data testing menghasilkan model evaluation yang diukur dengan nilai *Accuracy* dan *AUC*

Eksperimen dan Pengujian Model

Penelitian yang akan dilakukan dalam eksperimen ini adalah dengan menggunakan komputer untuk melakukan proses perhitungan terhadap model yang diusulkan. Tahapan eksperimen pada penelitian ini adalah:

1. Menyiapkan dua dataset untuk eksperimen
2. Melakukan *training* dan *testing* terhadap model C4.5 dan mencatat hasil Accuracy dan *AUC*
3. Melakukan *training* dan *testing* terhadap model C4.5 dengan menggunakan *PSO* dan mencatat hasil *Accuracy* dan *AUC.*

Evaluasi dan Validasi Hasil

Pada tahap ini akan dilakukan proses pengujian metode yang diusulkan dengan mengevaluasi perbandingan hasil *Accuracy* dan *AUC* seluruh eksperimen antara menggunakan algoritma C4.5 dengan algoritma C4.5 dan *PSO*, dan memvalidasi model prediksi yang dianggap paling optimal semakin tinggi nilai *Accuracy* semakin baik pula metode yang digunakan.

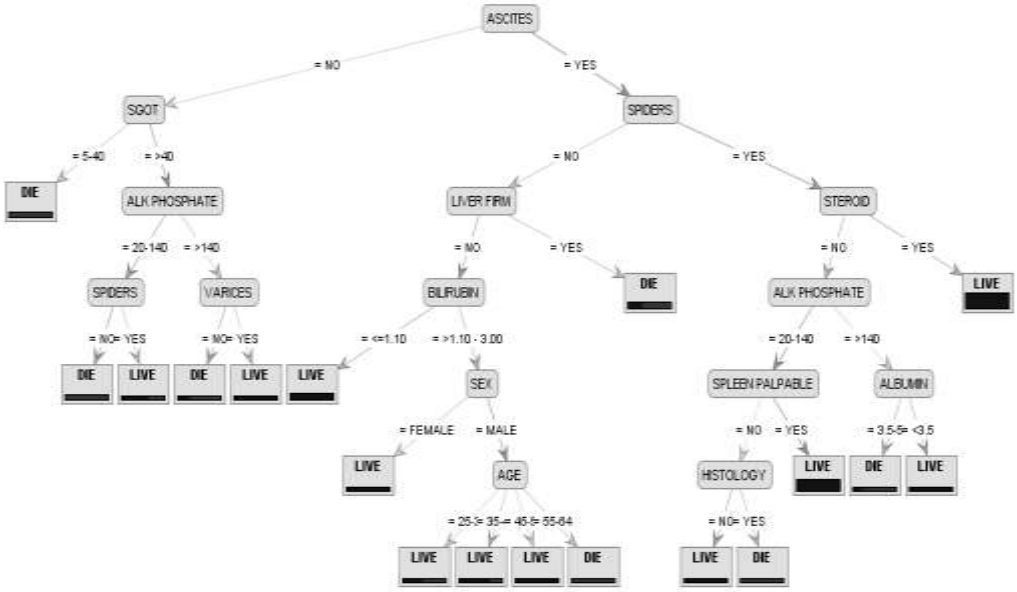
IV. HASIL PENELITIAN DAN PEMBAHASAN

Hasil

Tujuan dari penelitian ini adalah menerapkan *Particle Swarm Optimization* untuk mengeliminasi atribut input pada metode Algoritma C4.5, untuk meningkatkan akurasi prediksi penyakit hepatitis. Hasil dari penelitian ini berupa hasil proses pengolahan kualitatif dan kuantitatif yang telah dikumpulkan dengan perhitungan berdasarkan model yang diusulkan. Penelitian akan dilakukan terhadap semua dataset yang tersedia. Eksperimen dan pengujian dalam penelitian dilakukan melakukan prediksi terhadap dataset dengan C4.5 tanpa PSO, dan metode seleksi atribut dengan menggunakan PSO. Eksperimen akan dilakukan terhadap dataset yang telah divalidasi.

Eksperimen dan Pengujian Model C4.5

Pembuatan model C4.5 dilakukan pada dataset yang terdiri dari 19 atribut yang merupakan atribut dari diagnosis penyakit hepatitis dan class yang merupakan hasil akhir prediksi. Data kemudian di validasi agar proses pelatihan dapat berjalan dengan cepat dan mampu digunakan untuk melakukan pelatihan. Model dari algoritma C4.5 yaitu berupa pohon keputusan, untuk dapat membuat pohon keputusan, langkah pertama adalah menghitung jumlah *class* yang terkena penyakit hepatitis yang hidup dan yang meninggal dari masing-masing *class* berdasarkan atribut yang telah ditentukan dengan menggunakan data *training*. Kemudian menghitung Entropy (Total). Setelah didapatkan hasil perhitungan *entropy* dan *gain*, maka pohon keputusan yang terbentuk dapat dilihat seperti gambar di bawah ini:



Gambar 3 Pohon Keputusan Klasifikasi Penyakit Hepatitis menggunakan algoritma C4.5

Dari pohon keputusan tersebut maka diperoleh aturan-aturan atau *rule* sebagai berikut:

1) R1 : if ASCITES = NO and SGOT = 5-40 then class = DIE

2) R2 : if ASCITES = NO and SGOT = >40 and ALK PHOSPHATE = 20-140 and SPIDERS = NO then Class = DIE

3) R3 : if ASCITES = NO and SGOT = >40 and ALK PHOSPHATE = 20-140 and SPIDERS = YES then Class = LIVE

4) R4 : if ASCITES = NO and SGOT = >40 and ALK PHOSPHATE = >140 and VARICES = NO then Class = DIE

5) R5 : if ASCITES = NO and SGOT = >40 and ALK PHOSPHATE = >140 and VARICES = YES then Class = LIVE

6) R6 : if ASCITES = YES and SPIDERS = NO and LIVER FIRM = NO and BILIRUBIN = <=1.10 then Class = LIVE

7) R7 : if ASCITES = YES and SPIDERS = NO and LIVER FIRM = NO and BILIRUBIN = >1.10 - 3.00 and SEX = FEMALE then Class = LIVE

8) R8 : if ASCITES = YES and SPIDERS = NO and LIVER FIRM = NO and BILIRUBIN = >1.10 - 3.00 and SEX = MALE and AGE = 25-34 then Class = LIVE

9) R9 : if ASCITES = YES and SPIDERS = NO and LIVER FIRM = NO and BILIRUBIN = >1.10 - 3.00 and SEX = MALE and AGE = 35-44 then Class = LIVE

10) R10 : if ASCITES = YES and SPIDERS = NO and LIVER FIRM = NO and BILIRUBIN = >1.10 - 3.00 and SEX = MALE and AGE = 45-54 then Class = LIVE

11) R11 : if ASCITES = YES and SPIDERS = NO and LIVER FIRM = NO and BILIRUBIN = >1.10 - 3.00 and SEX = MALE and AGE = 55-64 then Class = DIE

12) R12 : if ASCITES = YES and SPIDERS = NO and LIVER FIRM = YES then Class = DIE

13) R13 : if ASCITES = YES and SPIDERS = YES and STEROID = NO and ALK PHOSPHATE = 20-140 and SPLEEN PALPABLE = NO and HISTOLOGY = NO then Class = LIVE

14) R14 : if ASCITES = YES and SPIDERS = YES and STEROID = NO and ALK PHOSPHATE = 20-140 and SPLEEN PALPABLE = NO and HISTOLOGY = YES then Class = DIE

15) R15 : if ASCITES = YES and SPIDERS = YES and STEROID = NO and ALK PHOSPHATE = 20-140 and SPLEEN PALPABLE = YES then Class = LIVE

16) R16 : if ASCITES = YES and SPIDERS = YES and STEROID = NO and ALK PHOSPHATE = >140 and ALBUMIN = 3.5-5.0 then Class DIE

17) R17 : if ASCITES = YES and SPIDERS = YES and STEROID = NO and ALK PHOSPHATE = >140 and ALBUMIN = <3.5 then Class = LIVE

18) R18 : if ASCITES = YES and SPIDERS = YES and STEROID = YES then Class = LIVE

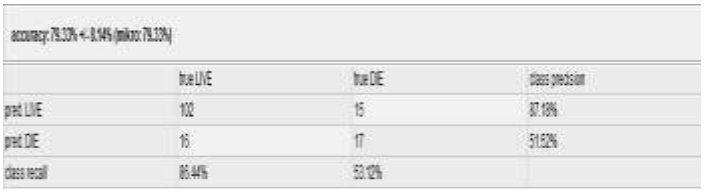
Hasil Pengujian Dengan Algoritma C4.5

Hasil dari uji coba yang dilakukan yaitu untuk menghasilkan nilai *accuracy* dan nilai *AUC* (*Area Under Curve*).

a. **Evaluasi Model Dengan *Confusion Matrix***

Model *confusion matrix* akan membentuk matrix yang terdiri dari *true positif* atau tupel positif dan *true negatif* atau tupel negatif, kemudian masukan data *testing* yang sudah disiapkan ke dalam *confusion matrix* sehingga didapatkan hasil pada tabel di bawah ini:

Tabel 3 *Confusion Matrix* Algoritma Klasifikasi C4.5 pada data *testing*



accuracy: 78.33% +/- 8.44% (mikro: 78.33%)			
	true LIVE	true DIE	class precision
pred. LIVE	102	15	87.18%
pred. DIE	16	17	51.52%
class recall	86.44%	53.13%	

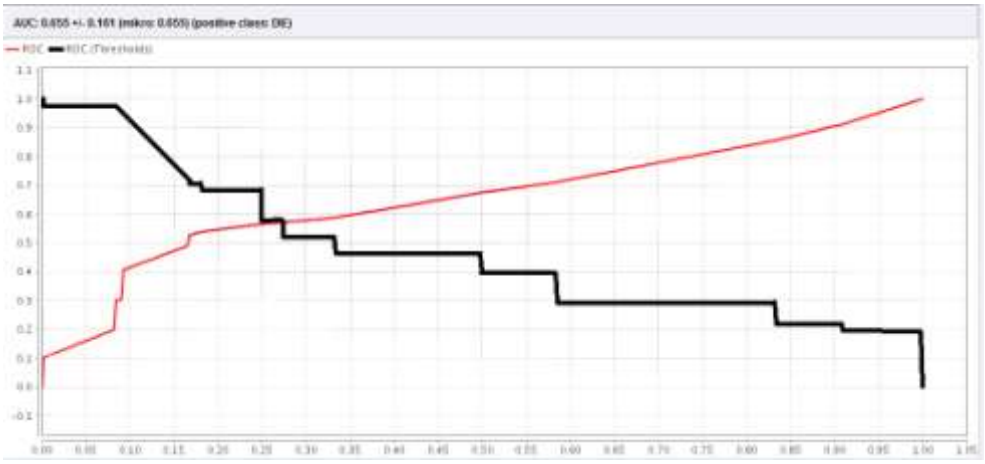
Berdasar tabel diatas dari data *testing* terdapat rincian jumlah *True Positive (TP)* 102, *False Negative (FN)* 15, *False Positive (FP)* adalah 16 dan *True Negative (TN)* 17. Dari data tersebut maka dapat dihitung nilai *accuracy, sensitvity, specifity, PPV dan NPV*. Data hasil olahan dapat dilihat pada tabel di bawah:

Tabel 4 Nilai *Accuracy*, *Sensitivity*, *Specificity*, *PPV* dan *NPV* C4.5

	Nilai
Accuracy	0,793
Sensitivity	0,872
Specificity	0,515
PPV	0,864
NVP	0, 531

b. **Evaluasi dengan *ROC Curve***

Hasil pengujian terhadap data *testing* untuk algoritma C4.5 terhadap nilai ROC diketahui pada gambar di bawah ini:

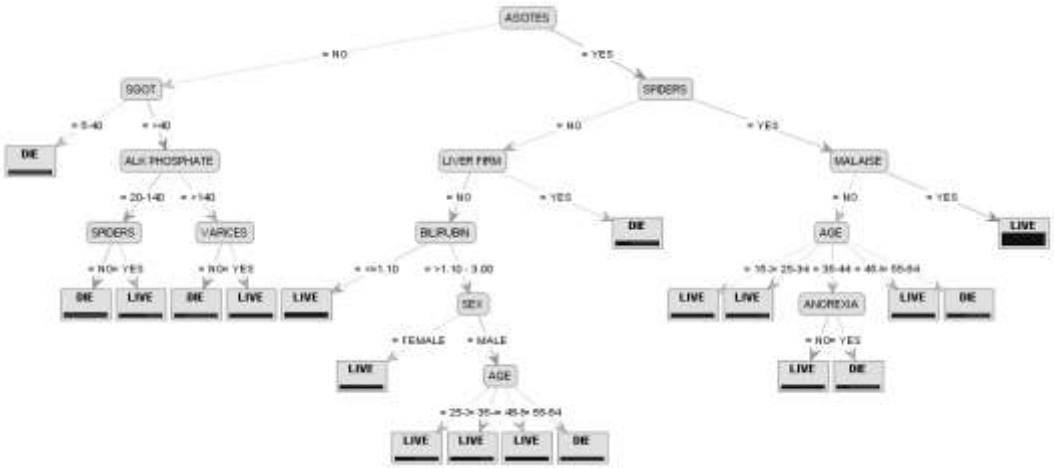


Gambar 4. Nilai AUC Algoritma C4.5 dalam grafik ROC

Berdasarkan nilai AUC sebesar 0,655 yang ditunjukan gambar di atas maka akurasi memiliki tingkat *Poor Classification*.

Eksperimen dan pengujian model PSO dan C4.5

Dengan PSO data yang akan diolah akan diberikan bobot untuk membantu meningkatkan hasil perhitungan, pemberian bobot ini ini diberikan secara acak dengan menentukan nilai minimum dan maksimum bobot. Setelah itu setiap partikel akan memiliki bobot sendiri dalam dataset, dan algoritma C4.5 akan diterapkan dan dihitung tingkat akurasinya. Setelah semua partikel dihitung, akan dicari partikel dengan nilai akurasi terbaik. Perulangan selanjutnya, partikel lainnya akan secara acak bergerak kearah partikel terbaik agar dapat menemukan bobot yang lebih baik lagi. Proses ini terus berulang sampai pada batas perulangan yang diijinkan. Maka pohon keputusan yang terbentuk dari model PSO dan C4.5 sebagai berikut:



Gambar 5. Pohon Keputusan Klasifikasi Penyakit Hepatitis PSO dan C4.5

Dari pohon keputusan tersebut maka diperoleh aturan-aturan atau *rule* sebagai berikut:

1) if ASCITES = NO and SGOT = 5-40 then class = DIE

2) if ASCITES = NO and SGOT = >40 and ALK PHOSPHATE = 20-140 and SPIDERS = NO then class = DIE

3) if ASCITES = NO and SGOT = >40 and ALK PHOSPHATE = 20-140 and SPIDERS = YES then class = LIVE

4) if ASCITES = NO and SGOT = >40 and ALK PHOSPHATE = >140 and VARICES = NO then class = DIE

5) if ASCITES = NO and SGOT = >40 and ALK PHOSPHATE = >140 and VARICES = YES then class = LIVE

6) if ASCITES = YES and SPIDERS = NO and LIVER FIRM = NO and BILIRUBIN = <=1.10 then class = LIVE

7) if ASCITES = YES and SPIDERS = NO and LIVER FIRM = NO and BILIRUBIN = >1.10 - 3.00 and SEX = FEMALE then class = LIVE

8) if ASCITES = YES and SPIDERS = NO and LIVER FIRM = NO and BILIRUBIN = >1.10 - 3.00 and SEX = MALE and AGE = 25-34 then class = LIVE

9) if ASCITES = YES and SPIDERS = NO and LIVER FIRM = NO and BILIRUBIN = >1.10 - 3.00 and SEX = MALE and AGE = 35-44 then class = LIVE

10) if ASCITES = YES and SPIDERS = NO and LIVER FIRM = NO and BILIRUBIN = >1.10 - 3.00 and SEX = MALE and AGE = 45-54 then class = LIVE

11) if ASCITES = YES and SPIDERS = NO and LIVER FIRM = NO and BILIRUBIN = >1.10 - 3.00 and SEX = MALE and AGE = 55-64 then class = DIE

12) if ASCITES = YES and SPIDERS = NO and LIVER FIRM = YES then class = DIE

13) if ASCITES = YES and SPIDERS = YES and MALAISE = NO and AGE = 15-24 then class = LIVE

14) if ASCITES = YES and SPIDERS = YES and MALAISE = NO and AGE = 25-34 then class = LIVE

15) if ASCITES = YES and SPIDERS = YES and MALAISE = NO and AGE = 35-44 and ANOREXIA = NO then class = LIVE

16) if ASCITES = YES and SPIDERS = YES and MALAISE = NO and AGE = 35-44 and ANOREXIA = YES then class = DIE

17) if ASCITES = YES and SPIDERS = YES and MALAISE = NO and AGE = 45-54 then class = LIVE

18) if ASCITES = YES and SPIDERS = YES and MALAISE = NO and AGE = 55-64 then class = DIE

19) if ASCITES = YES and SPIDERS = YES and MALAISE = YES then class = LIVE

Dari pohon keputusan tersebut maka diperoleh aturan-aturan atau *rule* sebagai berikut:

20) if ASCITES = NO and SGOT = 5-40 then class = DIE

21) if ASCITES = NO and SGOT = >40 and ALK PHOSPHATE = 20-140 and SPIDERS = NO then class = DIE

22) if ASCITES = NO and SGOT = >40 and ALK PHOSPHATE = 20-140 and SPIDERS = YES then class = LIVE

23) if ASCITES = NO and SGOT = >40 and ALK PHOSPHATE = >140 and VARICES = NO then class = DIE

24) if ASCITES = NO and SGOT = >40 and ALK PHOSPHATE = >140 and VARICES = YES then class = LIVE

25) if ASCITES = YES and SPIDERS = NO and LIVER FIRM = NO and BILIRUBIN = <=1.10 then class = LIVE

26) if ASCITES = YES and SPIDERS = NO and LIVER FIRM = NO and BILIRUBIN = >1.10 - 3.00 and SEX = FEMALE then class = LIVE

27) if ASCITES = YES and SPIDERS = NO and LIVER FIRM = NO and BILIRUBIN = >1.10 - 3.00 and SEX = MALE and AGE = 25-34 then class = LIVE

28) if ASCITES = YES and SPIDERS = NO and LIVER FIRM = NO and BILIRUBIN = >1.10 - 3.00 and SEX =

MALE and AGE = 35-44 then class = LIVE

29) if ASCITES = YES and SPIDERS = NO and LIVER FIRM = NO and BILIRUBIN = >1.10 - 3.00 and SEX = MALE and AGE = 45-54 then class = LIVE

30) if ASCITES = YES and SPIDERS = NO and LIVER FIRM = NO and BILIRUBIN = >1.10 - 3.00 and SEX = MALE and AGE = 55-64 then class = DIE

31) if ASCITES = YES and SPIDERS = NO and LIVER FIRM = YES then class = DIE

32) if ASCITES = YES and SPIDERS = YES and MALAISE = NO and AGE = 15-24 then class = LIVE

33) if ASCITES = YES and SPIDERS = YES and MALAISE = NO and AGE = 25-34 then class = LIVE

34) if ASCITES = YES and SPIDERS = YES and MALAISE = NO and AGE = 35-44 and ANOREXIA = NO then class = LIVE

35) if ASCITES = YES and SPIDERS = YES and MALAISE = NO and AGE = 35-44 and ANOREXIA = YES then class = DIE

36) if ASCITES = YES and SPIDERS = YES and MALAISE = NO and AGE = 45-54 then class = LIVE

37) if ASCITES = YES and SPIDERS = YES and MALAISE = NO and AGE = 55-64 then class = DIE

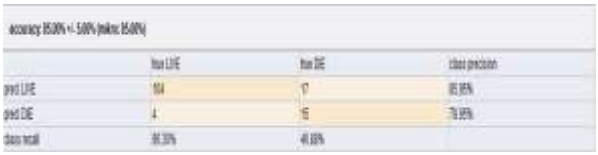
38) if ASCITES = YES and SPIDERS = YES and MALAISE = YES then class = LIVE

Hasil Pengujian Dengan PSO dan C4.5

Evaluasi Model Dengan *Confusion Matrix*

Model *confusion matrix* akan membentuk matrix yang terdiri dari *true positif* atau tupel positif dan *true negatif* atau tupel negatif, kemudian masukan data *testing* yang sudah disiapkan ke dalam *confusion matrix* sehingga didapatkan hasil pada tabel di bawah ini:

Tabel 5 *Confusion Matrix* PSO - Algoritma C4.5 pada data *testing*



Berdasar tabel diatas dari data *testing* terdapat rincian jumlah *True Positive (TP)* 104, *False Negative (FN)* 15, *False Positive (FP)* adalah 17 dan *True Negative (TN)* 4. Dari data tersebut maka dapat dihitung nilai *accuracy*, *sensitvity*, *specifity dan NPV*. Data hasil olahan dapat dilihat pada tabel di bawah:

Tabel 6. Nilai *Accuracy*, *Sensitivity*, *Specificity*, *PPV* dan *NPV* PSO-C4.5

	Nilai
Accuracy	0,85
Sensitivity	0,859
Specificity	0,789
PPV	0,962
NPV	0,468

Evaluasi dengan *ROC Curve*

Hasil pengujian terhadap data *testing* untuk PSO-algoritma C4.5 terhadap nilai ROC diketahui pada gambar di bawah ini:

Gambar 6. Nilai AUC PSO-Algoritma C4.5 dalam grafik ROC

Berdasarkan nilai AUC sebesar 0,718 yang ditunjukan gambar di atas maka akurasi memiliki tingkat *Fair Classification*.

V. KESIMPULAN

Hasil penelitian untuk nilai akurasi algoritma C4.5 senilai 79,33%, sedangkan untuk nilai akurasi Optimasi algoritma C4.5 menggunakan PSO sebesar 85,00% sehingga tampak selisih nilai akurasi yaitu sebesar 5,67%. Sedangkan evaluasi menggunakan *ROC curve* diperoleh hasil untuk algoritma C4.5 bernilai 0,655 dengan tingkat diagnosa *Poor Classification* dan Optimasi algoritma C4.5 menggunakan PSO bernilai 0,718 dengan tingkat diagnosa *Fair Classification*, didapatkan selisih nilai AUC sebesar 0,063.
Dari 19 atribut yang terdapat pada dataset *UCI Machine Learning Data Repository*, kemudian selanjutnya diseleksi menjadi hanya 11 atribut yang digunakan dalam menentukan prediksi penyakit hepatitis, atribut-atribut tersebut yaitu: ASCITES, SGOT, ALK PHOSPHATE, VARICES, SPIDERS, LIVER FIRM, BILIRUBIN, SEX, MALAISE, AGE, ANOREXIA.
Sehingga dapat disimpulkan bahwa penerapan teknik optimasi *particle swarm optimization* mampu menyeleksi atribut pada C4.5, sehingga menghasilkan tingkat akurasi diagnosis penyakit hepatitis yang lebih baik dibanding dengan menggunakan metode individual algoritma C4.5.

VI. DAFTAR PUSTAKA

[1] Bai, Q. (2010). Anaysis of Particle Swarm Optimization Algorithm. *Computer and Information Science-CCSE*, 180-184.

[2] Bramer, M. (2007). *Principles od Data Mining*. London: Spinger.

[3] Cho, Y. J., Lee, H., & Jun, C. H. (2011). Optimization of Decision Tree for Classification Using a Particle Swarm. *IEMS*, 272-278.

[4] Dawson, C. W. (2009). *Projects in Computing and Information Systems a student's guide.* Harlow, UK: Addison-Wesley.

[5] Dugdale, D. C. (2013, Februari 2). *Prothombin Time*. Dipetik Maret 17, 2015, dari Medline Plus: http://www.nlm.nih.gov/medlineplus/ency/article/003652

[6] Fine, J. (2012). *An Overview Of Statistical Methods in Diagnostic Medicine.* Chapel Hill.

[7] Gorunescu, F. (2011). *Data Mining Concepts, Models and Techniques.* Berlin: Springer.

[8] Han, J., & Kamber, M. (2006). *Data Mining Concepts and Techniques.* San Fransisco: Morgan Kauffman.

[9] Han, J., & Kamber, M. (2007). *Data Mining Concepts and Techniques Second Edition.* San Francisco: Morgan Kauffman.

[10] Kementerian Kesehatan RI. (2014). *Situasi dan Analisis Hepatitis.* Jakarta: Pusat Data dan Informasi.

[11] Kothari, C. R. (2004). *Research Methology Methods and Techniques.* India: New Age International Limited.

[12] Kusrini, & Luthfi, E. T. (2009). *Algoritma Data Mining.* Yogyakarta: ANDI.

[13] Larose, D. T. (2005). *Discovering Knowledge in Data An Introduction to Data Mining.* New Jersey: John Wiley & Sons, Inc.

[14] Larose, D. T. (2007). *Data Mining Methods and Models.* New Jersey: John Wiley & Sons, Inc.

[15] Liao, T. W., & Triantaphyllou, E. (2007). *Recent Advances in Data Mining of Enterprise Data Algorithms and Applications.* USA: World Scientific Publishing Co, Pte.Ltd.

[16] Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook.* New York: Springer .

[17] Prasetyo, E. (2014). *Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab.* Yogyakarta: Andi Offset.

[18] Rouhani, M., & Haghighi, M. M. (2009). The Diagnosis of Hepatitis Disease by Support Vector Machines and Artificial Neural Networks. *IEEE-IACSIT*, 456-458.

[19]Sathyadevi, G. (2011). Application of CART Algorithm In Hepatitis Disease Diagnosis. *IEEE-ICRTIT*, 1283-1287.

[20]Setiyorini, T. (2014). *Penerapan Metode Bagging Untuk Mengurangi Data Noise Pada Neural Netwok Estimasi Kuat Tekan Beton.* Jakarta.

[21]Shariati, S., & Haghighi, M. M. (2010). Comparison of anfis neural network with several other anns and support vector machine for diagnosing hepatitis and thyroid. *IEEE*, 596-599.

[22]Sugiyono. (2009). *Metode Penelitian Kuantitatif Kualitatif dan R&D.* Bandung: Alfabeta.

[23]Sumanto. (2014). *Statistika Deskriptif*. Yogyakarta: Center of Academic Publishing Service.

[24]Tan, F., Fu, X., Zhang, Y., & Bourgeois, A. G. (2007). A Genetic Algorithm-Based Method for Feature Subset Selection. *Soft Computing*, 111-120.

[25]Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning and Tools.* Burlington: Morgan Kaufmann Publisher.

[26]Wu, X., & Kumar, V. (2009). *The Top Ten Algorithms in Data Mining.* London: Taylor & Francis Group, LLC.

[27]Wulandari, N. (2014). *Seleksi Variabel Particle Swarm Optimization Untuk Prediksi Produksi Kelapa Sawit Dengan Menggunakan Artificial Neural Network*. Jakar

[28]Xu, & Donald. (2009). *Clustering.* Canada: A JOHN WILEY & SONS, INC.