ANALISA DATA MINING UNTUK PREDIKSI PENYAKIT HEPATITIS DENGAN MENGGUNAKAN METODE *NAIVE BAYES* DAN *SUPPORT VECTOR MACHINE*

Eka Wulansari Fridayanthie
Program studi Manajemen Informatika AMIK "BSI Jakarta"
Jl. RS Fatmawati No. 24 Pondok Labu, Jakarta Selatan. Indonesia
Email : eka.ewf@bsi.ac.id

ABSTRACT

In the case of hepatitis disease prediction has been solved by a method using Support Vector Machine (SVM) .Penyakit hepatitis is an inflammatory disease of the liver due to viral infection that attacks and cause damage to cells and organs function hati.Penyakit forerunner hepatitis is a disease of the liver cancer. Attributes or variables that have as many as 20 attributes which consists of 19 attributes preditor and 1 as the output destination attribute used to differentiate the results of the examination. Invene dataset from the University of California (UCI) Machine Learning Repository 583 as the data used and replace missing after the data is used only to evaluate the data 153 SVMyang approach proposed in the study ini.Hasil simulations showed that by developing this model achieved a reduction in dimensions and identification hati.Salah cancer of the optimization algorithm is quite popular is Naïve Bayes. In this study, will be used also classification algorithm Support Vector Machine (SVM) will be used to establish a predictive classification model of hepatitis.

Keywords : *Hepatitis,Naïve Bayes , Support Vector Machine*

I. PENDAHULUAN

Diagnosis medis dipandang sebagai tugas penting namun rumit yang perlu dijalankan secara tepat dan efisien. Otomatisasi sistem ini akan sangat mengutungkan. Namun, sayangnya semua dokter tidak memiliki keahlian khusus dalam setiap bagian keahlian dan terlebih lagi ada kekurangan dari nara sumber di tempat tertentu (Ansari, dkk, 2011: 43). Oleh karena itu, sistem diagnosis otomatis secara medis mungkin akan sangat bermanfaat dengan membawa semua hal itu. Sesuai informasi berbasis komputer dan/atau sistem pendukung keputusan dapat membantu dalam mencapai pengujian klinis dengan biaya yang terjangkau. Tujuan penelitian ini adalah melakukan analisia dan komparasi metode klasifikasi *data mining* sehingga diperoleh metode yang paling akurat di negara pada umumnya untuk prediksi penyakit hepatitis.

Hasil penelitian ini dapat digunakan sebagai rekomendasi dan masukan bagi ahli kesehatan dalam membuat prediksi penyakit hepatitis,Membantu administrasi perguruan tinggi untuk memberikan peringatan dini dan pembimbingan awal bagi mahasiswa yang kemungkinan tidak lulus tepat waktu.Ruang lingkup penelitian ini terbatas pada penggunaan metode Support Vector Machine dan *Naïve Bayes*, dalam memprediksi penyakit hepatitis dan melakukan perbandingan akurasi kedua metode tersebut. Parameter yang diuji pada data adalah *age, sex, steroid, antivirals, fatigue, malaise, anorexia, liver_big, liver_firm, spleen_palpable, spdiders, ascites, varices, bilirubin, alk_phosphate, sgot, albumin, protime, histology,* dan *class* (atribut hasil prediksi).

II. LANDASAN TEORI

2.1. Pengertian Penyakit Hepatitis

Hepatitis merupakan penyakit yang menimbulkan peradangan pada hati (*liver*), kadang-kadang menyebabkan kerusakan permanen.Penyakit ini sering disebabkan oleh virus dan zat-zat kimia tertentu yang masuk ke hati, termasuk obat-obatan dan alkohol. Virus hepatitis juga ada beberapa jenis yang menyerang hati, tepatnya pada sel-sel hati.

Peradangan ini, paling sering disebabkan oleh virus, walaupun dapat juga oleh sebab-sebab lain. Berkaitan dengan virus yang menyerang dan kondisi penyakit, hepatitis digolongkan sebagai berikut :

1. Hepatitis A (Hepatitis Infeksi)
2. Hepatitis B (Hepatitis Serum)
3. Hepatitis C (Hepatitis Non-A/Non-B)
4. Hepatitis D (Hepatitis Delta)
5. Hepatitis E (Hepatitis Enterik)
6. Hepatitis F
7. Hepatitis G
8. Hepatitis Kronis

2.2. Data Mining

Menurut Witten data mining adalah pemecahan masalah dengan menganalisa data yang sudah ada sebelumnya, dan didefinisikan sebagai proses dari penemuan pola pada suatu data (Witten,dkk,2011, :39) Menurut Gartner Group data mining adalah suatu proses menemukan hubungan yang berarti, pola dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik statistik dan matematika (Larose, 2005:11).

2.3. Algoritma Klasifikasi

Klasifikasi merupakan salah satu tujuan yang banyak dihasilkan dalam *data mining*. Klasifikasi merupakan proses pengelompokkan sebuah variabel ke 13 dalam kelas yang sudah ditentukan (Larose, 2005:95). Data mining mampu mengolah data dalam jumlah besar, setiap data terdiri dari kelas tertentu bersama dengan variable dan faktor faktor penentu kelas variabel tersebut.Dengan data mining, peneliti dapat menentukan suatu kelas dari variabel data yang dimiliki.

2.4. Pengujian K-Fold Cross Validation

Cross Validation adalah teknik validasi dengan membagi data secara acakkedalam k bagian dan masing-masing bagian akan dilakukan proses klasifikasi(Han & Kamber, 2007). Dengan menggunakan *cross validation* akan dilakukanpercobaan sebanyak k. Data yang digunakan dalam percobaan ini adalah data*training* untuk mencari nilai *error rate* secara keseluruhan. Secara umum pengujian nilai k dilakukan sebanyak 10 kali untuk memperkirakan kurasiestimasi. Dalam penelitian ini nilai k yang digunakan berjumlah 10 atau 10-*foldCross Validation.*

2.5. Algoritma *Support Vector Machine*

Support Vector Machine (SVM) diperkenalkan oleh Vapnik, Boser dan Guyon pada tahun 1992.SVM merupakan salah satu teknik yang relatif baru dibandingkan dengan teknik lain, tetapi memiliki performansi yang lebih baik di berbagai bidang aplikasi seperti bioinformatika, pengenalan tulisan tangan, klasifikasi teks, klasifikasi diagnosis penyakit dan lain sebagainya (Feng-Chia, 2009). Dalam kata lain, hanya sejumlah titik penting untuk klasifikasi tujuan dalam kerangka svm dan dengan demikian harus diambil (Huang, Yang, King, & Lyu, 2008).*Support Vector Machine (SVM)* adalah *metode learning machine* yang bekerja atas prinsip *Structural Risk Minimization (SRM)* dengan tujuan menemukan *hyperplane* terbaik yang memisahkan dua buah *class* pada *input*

space (Bellotti & Crook, 2007).*Hyperplane* terbaik adalah *hyperplane* yang terletak ditengah-tengah antara dua set obyek dari dua *class*. *Hyperplane* pemisah terbaik antara kedua *class* dapat ditemukan dengan mengukur *margin hyperplane* tersebut dan mencari titik maksimalnya.*Margin* adalah jarak antara *hyperplane* tersebut dengan *pattern* terdekat dari masing-masing *class.Pattern* yang paling dekat ini disebut sebagai *support vector* (Aydin, Karakose & Akin, 2011).

2.6. *Naive Bayes*

Klasifikasi Bayes juga dikenal dengan *Naïve Bayes*, memiliki kemampuan sebanding dengan dengan pohon keputusan dan *Neural Network* (Han & Kamber, 2007). Klasifikasi Bayes adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu kelas (Kusrini, 2009). *Naïve Bayes* dapat menggunakan penduga kernel kepadatan, yang meningkatkan kinerja jika asumsi normalitas sangat tidak benar, tetapi juga dapat menangani atribut numeric menggunakan diskritisasi diawasi (Witten & Frank, 2011). Teknik *Naïve Bayes* (NB) adalah salah satu bentuk sederhana dari *Bayesian* yang jaringan untuk klasifikasi. Sebuah jaringan *Bayes* dapat dilihat sebagai diarahkan sebagai tabel dengan distribusi probabilitas gabungan lebih dari satu set diskrit dan variabel stokastik (Pearl 1988) (Liao, 2007).

Metode ini penting karena beberapa alasan, termasuk berikut. Hal ini sangat mudah untuk membangun, tidak perlu ada yang rumit Parameter estimasi skema berulang. Ini berarti dapat segera diterapkan untuk besar Data set. Sangat mudah untuk menafsirkan, sehingga pengguna tidak terampil dalam teknologi classifier dapat memahami mengapa itu adalah membuat klasifikasi itu membuat. Dan, sangat penting, hal itu sering sangat baik: Ini mungkin bukan classifier terbaik dalam setiap diberikan aplikasi, tetapi biasanya dapat diandalkan untuk menjadi kuat dan melakukan dengan sangat baik (Wu, 2009).

2.7. *Confusion matrix*

Confusion matrix memberikan keputusan yang diperoleh dalam *traning* dan *testing, confusion matrix* memberikan penilaian *performance* klasifikasi berdasarkan objek dengan benar atau salah (Gorunescu, 2011).*Confusion matrix* berisi informasi aktual *(actual)* dan prediksi *(predicted)* pada sistem klasifikasi.

Tabel 1. *Confusion Matrix*

Classification	Predicted Class		
Observed Class		Class = Yes	Class = No
	Class = Yes	*a true positive* - TP)	*b (false negative* - FN)
	Class = No	*c (false positive* - FP)	d *(true negative* - TN)

Keterangan:

True Positive (TP) = proporsi positif dalam data set yang diklasifikasikan positif.

True Negative (TN) = proporsi negative dalam data set yang diklasifikasikan negative.

False Positive (FP) = proporsi negatif dalam data set yang diklasifikasikan potitif.

FalseNegative (FN) = proporsi negative dalam data set yang diklasifikasikan negatif.

Berikut adalah persamaan model *confusion matrix (*Han & Kamber, 2006):

a. Nilai *Accuracy* adalah proporsi jumlah prediksi yang benar. Dapat dihitung
dengan menggunakan persamaan:
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

b. *Sensitivity* digunakan untuk membandingkan proporsi TP terhadap tupel yang positif, yang dihitung dengan menggunakan persamaan:
$$Sensitivity = \frac{TP}{TP + FN}$$

c. *Specificity* digunakan untuk membandingan proporsi TN terhadap tupel yang negatif, yang dihitung dengan menggunakan persamaan:
$$Specificity = \frac{TN}{TN + FP}$$

d. PPV *(positive predictive value)* adalah proporsi kasus dengan hasil diagnosa
positif, yang dihitung dengan menggunakan persamaan:
$$PPV = \frac{TP}{TP + FP}$$

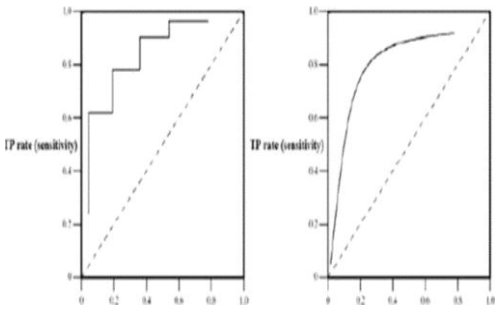
e. NPV *(negative predictive value)* adalah proporsi kasus dengan hasil diagnosa
negatif, yang dihitung dengan menggunakan persamaan:
$$PPV = \frac{TN}{TN + FN}$$

2.8. Kurva ROC

Kurva ROC *(Receiver Operating Characteristic)* adalah alat visual yang berguna untuk membandingkan dua model klasifikasi. ROC mengekspresikan *confusion matrix*. ROC adalah grafik dua dimensi dengan *false positives* sebagai garis horisontal dan *true positives* sebagai garis vertikal (Vecellis, 2009). Dengan kurva ROC, kita dapat melihat *trade off* antara tingkat dimana suatu model dapat mengenali tuple positif secara akurat dan tingkat dimana model tersebut salah mengenali tuple negatif sebagai tuple positif.

Sebuah grafik ROC adalah plot dua dimensi dengan proporsi positif salah (fp) pada sumbu *X* dan proporsi positif benar (tp) pada sumbu *Y*. Titik (0,1) merupakan klasifikasi yang sempurna terhadap semua kasus positif dan kasus negatif. Nilai positif salah adalah tidak ada (fp = 0) dan nilai positif benar adalah tinggi (tp = 1). Titik (0,0) adalah klasifikasi yang memprediksi setiap kasus menjadi negatif {-1}, dan titik (1,1) adalah klasifikasi yang memprediksi setiap kasus menjadi positif {1}.

Grafik ROC menggambarkan *trade-off* antara manfaat *(true positive)* dan biaya *(false positives).* Berikut tampilan dua jenis kurva ROC *(discrete* dan *continous).*



Gambar 1. Grafik ROC *(discrete* dan *continous)* (Gorunescu, 2011)

Poin diatas garis diagonal merupakan hasil klasifikasi yang baik, sedangkan point dibawah garis diagonal merupakan hasil klasifikasi yang buruk. Dapat disimpulkan bahwa, satu point pada kurva ROC adalah lebih baik dari pada yang lainnya jika arah garis melintang dari kiri bawah ke kanan atas didalam grafik.

Untuk tingkat akuransi nilai AUC dalam klasifikasi *data mining* dibagi menjadi lima kelompok (Gorunescu, 2011), yaitu:

1. 0.90 - 1.00 = klasifikasi sangat baik *(excellent classification)*
2. 0.80 - 0.90 = klasifikasi baik *(good classification)*
3. 0.70 - 0.80 = klasifikasi cukup *(fair classification)*
4. 0.60 - 0.70 = klasifikasi buruk *(poor classification)*
5. 0.50 - 0.60 = klasifikasi salah *(failure)*

III. METODE PENELITIAN

Dalam menyelesaikan penelitian, penulis membuat sebuah kerangka pemikiran yang berguna sebagai pedoman atau acuan penelitian ini sehingga penelitian dapat dilakukan secara konsisten. Penelitian ini terdiri dari beberapa tahap seperti terlihat pada gambar 1. Permasalahan pada penelitian ini adalah belum diketahuinya metode yang tepat dengan akurasi terbaik untuk prediksi penyakit hepatitis.

Untuk itu metode yang digunakan yaitu *Naïve Bayes,* dan *Support Vector Machine* untuk memecahkan masalah dilakukan pengujuan terhadap kinerja ketiga metode tersebut. Pengujian metode dilakukan dengan cara *confusion matrix* dan kurva ROC. Untuk mengembangkan aplikasi berdasarkan metode yang dibuat, digunakan *tools* RapidMiner. Berikut Tahapan-tahapan yang dilakukan pada penelitian ini :

3.1. Pengumpulan Data

Teknik pengumpulan data ialah teknik atau cara-cara yang dapat digunakan untuk menggunakan data (Riduwan, 2008). Dalam pengumpulan data terdapat sumber data, sumber data yang dihimpun langsung oleh peneliti disebut dengan sumber primer, sedangkan apabila melalui tangan kedua disebut sumber sekunder (Riduwan, 2008). Data pertama yang diperolah adalah data sekunder karena diperoleh dari UCI (Universitas California, Invene) *Machine Learning Repository* dengan alamat web http://archive.ics.uci.edu/ml/machine-learning-databases/hepatitis/.

Data yang dikumpulkan adalah data pemeriksaan pasien penyakit hepatitis oleh G. Gong (Carnegie – Mellon University) di Yugoslavia pada November 1988. Data terkumpul sebanyak 155 data dengan 123 pasien penyakit hepatitis yang hidup dan 32 pasien penyakit hepatitis yang mati dengan atribut *age, sex, steroid, antivirals, fatigue, malaise, anorexia, liver_big, liver_firm, spleen_palpable, spiders, ascites, varices, bilirubin, alk_phosphate, sgot, albumin, protime, histology,* dan *class* (atribut hasil prediksi)

3.2. Pengolahan Data Awal

Data yang diperoleh untuk penelitian ini sebanyak 155 *record* pasien pemeriksaan penyakit hepatitis baik yang hidup atau mati dan data kedua yaitu 538 *record* pasien pemeriksaan penyakit hati baik yang terdeteksi sakit atau tidak . Tetapi dalam data tersebut masih mengandung duplikasi dan anomali atau inkonsisten data maka dengan ini dilakukan *replace missing*.

3.3. Model atau Metode yang Diusulkan

Dalam penelitian ini akan dilakukan analisis komparasi menggunakan tiga metode klasifikasi data mining. Metode yang diusulkan untuk pengolahan data mahasiswa adalah pengunaan Algoritma C4.5, *Naïve Bayes* dan *Neural Network*. Data diolah sesuai dengan algoritmanya masing-masing, yakni data penyakit hepatitis diolah menggunakan metode Algoritma C4.5, *Naïve Bayes* dan *Neural Network*, setelah diolah dan menghasilkan model, maka

terhadap model yang dihasilkan tersebut dilakukan pengujian menggunakan *K-Fold Cross Validation*, kemudian dilakukan evaluasi dan validasi hasil dengan *confusion matrix* dan kurva *ROC*. Tahap selanjutnya adalah membandingkan hasil akurasi dan AUC dari setiap model, sehingga diperoleh model dari metode klasifikasi yang mana yang memperoleh nilai akurasi dan AUC tertinggi.

Hasil pengujian dengan akurasi yang paling tinggi adalah metode yang akan digunakan untuk prediksi penyakit hepatitis. Berikut gambaran kateristik dari masing-masing metode:

a. *Naïve Bayes* yaitu metode yang menghitung probabilitas antara kemunculan data yang satu dengan data yang lainnya.
b. *Support Vector Machine* yaitu metode *metode learning machine* yang bekerja atas prinsip *Structural Risk Minimization (SRM)* dengan tujuan menemukan *hyperplane* terbaik yang memisahkan dua buah *class* pada *input space*

IV. PEMBAHASAN

4.1. Pengolahan Data Awal

Data yang diperoleh untuk penelitian ini sebanyak 155 *record* pasien pemeriksaan penyakit hepatitis baik yang hidup atau mati dan data kedua yaitu 538 *record* pasien pemeriksaan penyakit hati baik yang terdeteksi sakit atau tidak . Tetapi dalam data tersebut masih mengandung duplikasi dan anomali atau inkonsisten data maka dengan ini dilakukan *replace missing*.

Tabel 2. *Missing Data* pada *Data Training*

ExampleSet (155 examples, 1 special attribute, 19 regular attributes) View Filter (155 / 155): all

Row No.	class	age	sex	steroid	antivirals	fatigue	malaise	anorexia	liver big	liver firm	spleen palp...	spiders	ascites	varice
1	LIFE	30	FEMALE	NO	YES	YES	YES	YES	NO	YES	YES	YES	YES	YES
2	LIFE	50	MALE	NO	YES	NO	YES	YES	NO	YES	YES	YES	YES	YES
3	LIFE	78	MALE	YES	YES	NO	YES	YES	YES	YES	YES	YES	YES	YES
4	LIFE	31	MALE	YES	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES
5	LIFE	34	MALE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
6	LIFE	34	MALE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
7	DIE	51	MALE	NO	YES	NO	YES	NO	YES	YES	NO	NO	YES	YES
8	LIFE	23	MALE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
9	LIFE	39	MALE	YES	YES	NO	YES	YES	YES	NO	YES	YES	YES	YES
10	LIFE	30	MALE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
11	LIFE	39	MALE	NO	NO	YES	YES	YES	NO	NO	YES	YES	YES	YES
12	LIFE	32	MALE	YES	NO	NO	YES	YES	YES	NO	YES	NO	YES	YES
13	LIFE	41	MALE	YES	NO	NO	YES	YES	YES	NO	YES	YES	YES	YES
14	LIFE	30	MALE	YES	YES	NO	YES	YES	YES	NO	YES	YES	YES	YES
15	LIFE	47	MALE	NO	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES
16	LIFE	38	MALE	NO	YES	NO	NO	NO	YES	YES	YES	YES	NO	YES
17	LIFE	66	MALE	YES	YES	NO	YES	YES	YES	YES	YES	YES	YES	YES
18	LIFE	40	MALE	NO	YES	NO	YES	YES	YES	NO	YES	YES	YES	YES
19	LIFE	38	MALE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
20	LIFE	38	MALE	NO	NO	YES	YES	YES	NO	NO	YES	YES	YES	YES
21	LIFE	22	FEMALE	YES	NO	NO	YES	YES	YES	YES	YES	YES	YES	YES
22	LIFE	27	MALE	YES	YES	NO	NO	NO	NO	NO	NO	NO	YES	YES

Parameter-parameter di atas akan dapat mudah diketahui dengan menggunakan *tools* dari *software framework* RapidMiner versi 5.3.005. Pada penelitian ini *Support Vector Machine (SVM)* digunakan karena diketahui dari hasil penelitian sebelumnya bahwa *Support Vector Machine (SVM)* memiliki kemampuan generalisasi yang sangat baik untuk memecahkan masalah walaupun dengan sampel yang terbatas. eksperimen menggunakan metode *support vector machine* menghasilkan tingkat akurasi sebesar 75.30 % dan mempunyai nilai AUC sebesar 0.780. Dari hasil tersebut diketahui bahwa keberhasilan dari *Support Vector Machine (SVM)* sangat dipengaruhi oleh pemilihan atribut yang tepat. Semakin

banyak atribut dan informasi yang digunakan akan mengakibatkan banyaknya waktu dan biaya yang dikorbankan bahkan akan mengurangi tingkat akurasi dan kompleksitas yang lebih tinggi.

Mengingat pentingnya seleksi atribut dalam *Support Vector Machine (SVM)* maka diterapkan *Particle swarm optimization (PSO)* untuk melakukan tugas tersebut. *Particle swarm optimization (PSO)* diketahui dapat digunakan sebagai teknik optimasi untuk mengoptimalkan subset fitur. Algoritma PSO sederhana dan memiliki kompleksitas yang lebih rendah.sehingga dapat memastikan solusi optimal dengan menyesuaikan pencarian global dan lokal, sehingga kinerja klasifikasi *Support Vector Machine (SVM)* dapat ditingkatkan.

Eksperiment dilakukan kembali dengan menerapkan *Particle swarm optimization (PSO)* untuk seleksi atribut dalam *Support Vector Machine (SVM)* dan dilakukan penyesuaian pada parameter *C, ε* dan *population.* Dari 20 variabel prediktor dilakukan seleksi atribut sehingga menghasikan terpilihnya 15 atribut yang dihasilkan.

4.2. Evaluasi dan Validasi Hasil

Model yang diusulkan pada penelitian tentang prediksi penyakit hepatitis adalah dengan menerapkan *support vector machine* dan *support vector machine* berbasis *Particle swarm optimization*. Penerapan algoritma *support vector machine* dengan menentukan nilai *weight* terlebih dahulu. Setelah didapatkan nilai akurasi dan AUC terbesar, nilai *weight* tersebut akan dijadikan nilai yang akan digunakan untuk mencari nilai akurasi dan AUC tertinggi.

Sedangkan penerapan algoritma *support vector machine* berbasis *Particle swarm optimization* beracuan pada nilai *weight* pada algoritma tersebut. Setelah ditemukan nilai akurasi yang paling ideal dari parameter tersebut langkah selanjutnya adalah menentukan nilai *weight*,sehingga terbentuk struktur algoritma yang ideal untuk pemecahan masalah tersebut.

Berdasarkan Tabel tersebut menunjukan bahwa, tingkat akurasi dengan menggunakan algoritma SVM adalah sebesar 68,42%, dan dapat dihitung untuk mencari nilai *accuracy, sensitivity, specificity, ppv,* dan *npv* hasilnya dan dapat dihitung untuk mencari nilai *accuracy, sensitivity, specificity,* ppv, dan npv pada persamaan dibawah ini:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{77 + 29}{77 + 29 + 8 + 41} = 0.6838$$

$$Sensitivity = \frac{TP}{TP + FN} = \frac{77}{77 + 41} = 0.6525$$

$$Specificity = \frac{TN}{TN + FP} = \frac{29}{29 + 8} = 0.7838$$

$$PPV = \frac{TP}{TP + FP} = \frac{77}{77 + 8} = 0.9058$$

$$NPV = \frac{TN}{TN + FN} = \frac{29}{29 + 41} = 0.4142$$

Tabel 3. Nilai *Accuracy, Sensitivity, Specificity, ppv* dan *npv* Metode *svm*

	Nilai (%)
Accuracy	68.38
Sensitivity	65.25
Specificity	78.38
PPV	90.58
NPV	41.42

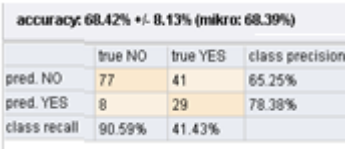
4.3. Hasil Pengujian Metode *Support Vector Machine*

1. *Confusion Matrix*

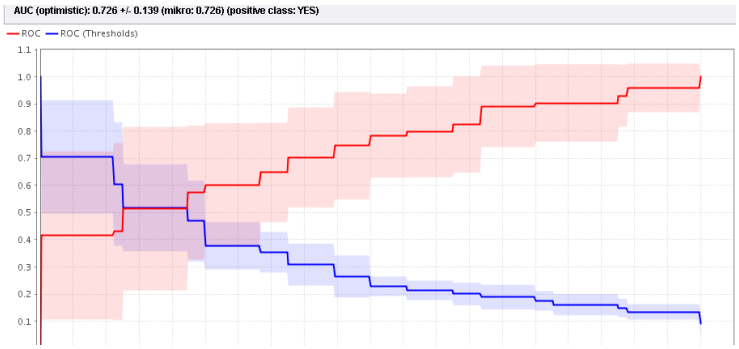
Tabel 4. menunjukkan hasil dari *confusion matrix* metode *support vector machine*

Tabel 4. Hasil *Confusion Matrix* untuk Metode *Support Vector Machine*

Accuracy :68.42 %			
	True: NO	True:YES	Precision
Pred. NO	77	41	71.64%
Pred. YES	8	29	89.81%
Class recall	90.58%	41.43%	



2. Kurva ROC
Hasil perhitungan divisualisasikan dengan kurva ROC. Perbandingan kedua *class* bisa dilihat pada Gambar yang merupakan kurva ROC untuk algoritma *Support Vector Machines.*



Gambar 2. Kurva ROC dengan Metode *Support Vector Machines*

Kurva ROC pada gambar 2 mengekspresikan *confusion matrix* dari Gambar Garis horizontal adalah *false positives* dan garis vertikal *true positives.* Menghasilkan nilai AUC *(Area Under Curve)* sebesar 0.726 dengan nilai akurasi klasifikasi cukup *(fair classification).*

4.4. Hasil Pengujian Model *Support Vector Machine* berbasis Algoritma *Particle Swarm Optimization* (PSO)

Hasil pengujian dengan menggunakan model *Support Vector Machine* didapatkan hasil pada table.

1. *Confusion Matrix*

Tabel diketahui dari 153 data, 33 diklasifikasikan *ya* sesuai dengan prediksi yang dilakukan dengan metode SVM berbasis *Particle Swarm Optimization* (PSO), lalu 7 data diprediksi *ya* tetapi ternyata hasilnya prediksi tidak, 33 data diprediksi tidak ternyata hasil prediksinya.

Tabel 5. Model *Confusion Matrix* untuk Metode *Support Vector Machine* Berbasis *Naïve Bayes*

Accuracy :83.71 %			
	True: NO	*True: YES*	*Precision*
Pred. YES	106	8	92.98%
Pred. NO	17	24	56.54%
Class recall	86.18%	75.00%	

Berdasarkan Tabel 5 tersebut menunjukan bahwa, tingkat akurasi dengan menggunakan algoritma SVM berbasis *Naïve Bayes* adalah sebesar 8.,71%, dan dapat dihitung untuk mencari nilai *accuracy, sensitivity, specificity, ppv,* dan *npv* hasilnya dan dapat dihitung untuk mencari nilai *accuracy, sensitivity, specificity,* ppv, dan npv pada persamaan dibawah ini:

Accuracy = TP + TN
= 78 + 33
= 0.7161
TP + TN + FP + FN = 78 + 33 + 7 + 37

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{78}{78+37} = 0.6782$$

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{33}{33+7}$$

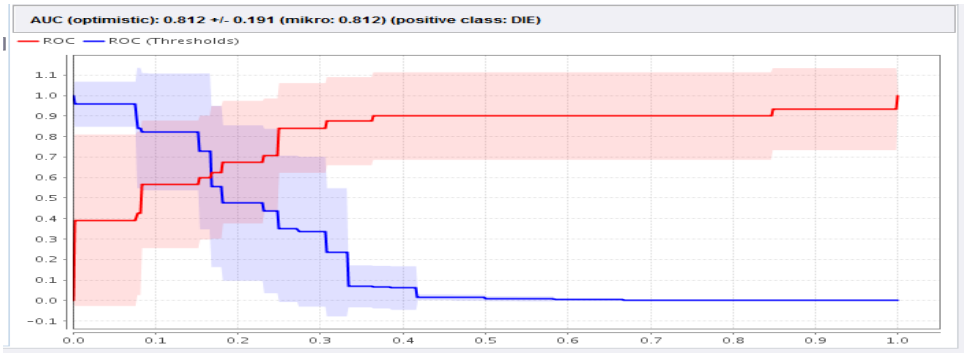
$$\text{PPV} = \frac{TP}{TP + FP} = \frac{78}{78+7} = 0.9176$$

$$\text{NPV} = \frac{TN}{TN + FN} = \frac{33}{33+37} = 0.4714$$

Tabel 6. Nilai *accuracy, sensitivity, specificity, ppv,* dan *npv* Metode *Support Vector Machine* berbasis *Particle Swarm Optimization*

	Nilai (%)
Accuracy	83.71
Sensitivity	67.82
Specificity	82.50
PPV	91.76
NPV	47.14

2. Kurva ROC
Hasil perhitungan divisualisasikan dengan kurva ROC.Perbandingan kedua metode komparasi bisa dilihat pada Gambar 3 yang merupakan kurva ROC untuk algoritma *Support Vector Machines* berbasis *Naïve Bayes* .Kurva ROC pada gambar 3 mengekspresikan *confusion matrix* dari Tabel 4. Garis horizontal adalah *false positives* dan garis vertikal *true positives*



Gambar 3. Kurva ROC dengan Metode *Support Vector Machines berbasis Naïve Bayes*

Dari Gambar 3 terdapat grafik ROC dengan nilai AUC (*Area Under Curve*) sebesar 0.812 dimana diagnosa hasilnya *Fair classification*

3. *Atribute weight*
Hasil *Atribute weight* yang didapat dari penelitian ini adalah tidak ada atribut yang bernilai 0 (nol) atau yang tidak berpengaruh,jadi semua atribut berpengaruh pada pada penelitian ini.

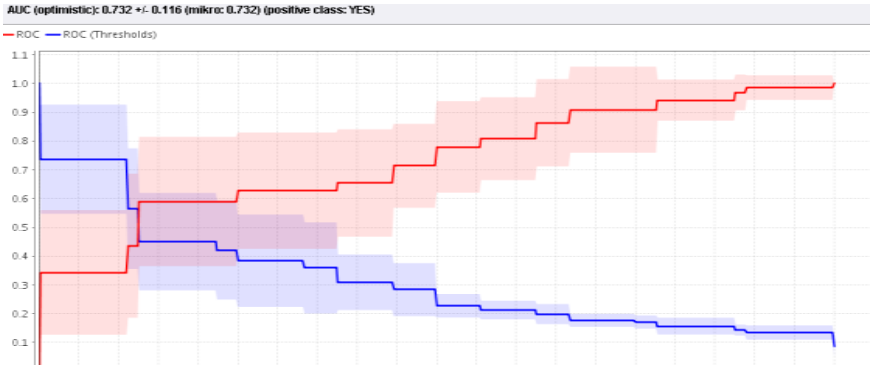
Tabel 7. Perbandingan *Performance* Metode *Dataset*

	SVM	***Naïve Bayes***
Accuracy	68.42%	83.71%
AUC	0,732	0.812

4.5. Analisis Evaluasi dan Validasi Model

Dari hasil pengujian diatas, baik evaluasi menggunakan *counfusion matrix* maupun ROC *curve* terbukti bahwa hasil pengujian algoritma SVM berbasis PSO memiliki nilai akurasi yang lebih tinggi dibandingkan dengan algoritma SVM Nilai akurasi untuk model algoritma SVM sebesar 68.38% dan nilai akurasi untuk model algoritma SVM berbasis Naïve Bayes sebesar 71.62 % dengan selisih akurasi 3.24%,

Untuk evaluasi menggunakan ROC *curve* sehingga menghasilkan nilai AUC (*Area Under Curve*) untuk model algoritma SVM mengasilkan nilai 0.726 dengan nilai diagnosa *Fair Classification*, sedangkan untuk algoritma SVMberbasis PSO (Particle Swarm Optimization) menghasilkan nilai 0.732 dengan nilai diagnose *Fair Classification*, dan selisih nilai keduanya sebesar 0.006. Dapat dilihat pada Gambar dibawah ini.



Gambar 4. Kurva ROC *Support vector machine* berbasis *Particle Swarm Optimization*

Dengan demikian algoritma SVM berbasis PSO dapat memberikan solusi untuk permasalahan dalam prediksi hasil prediksi penyakit hepatitis. Untuk rinciannya dapat dilihat pada Tabel .dan Gambar .

V. PENUTUP

5.1. Kesimpulan

Dalam penelitian ini dilakukan pengujian model dengan menggunakan *Support Vector Machines* dan *Support Vector Machines* berbasis *Particle Swarm Optimization* dengan menggunakan data penyakit hepatitis yang terkena penyakit atau tidak.

Model yang dihasilkan diuji untuk mendapatkan nilai *accuracy, precision, recall* dan AUC dari setiap algoritma sehingga didapat pengujian dengan menggunakan *support vector machines* didapat nilai *accuracy* adalah 68.38 % dan nilai AUC adalah 0.726. Sedangakan pengujian dengan mengunakan *support vector machines* berbasis *Naïve Bayes* didapatkan nilai *accuracy* 83.71 % dengan nilai dan nilai AUC adalah 0.812.

5.2. Saran

Agar penelitian ini bisa ditingkatkan, berikut adalah saran-saran yang diusulkan:

1. Penelitian ini diharapkan dapat digunakan pihak medis sebagai bahan pertimbangan memprediksi penyakit hepatitis, sehingga dapat

meningkatkan akurasi dalam prediksi prediksi penyakit hepatitis.
2. Penelitian ini dapat dikembangkan dengan metode optimasi lainnya seperti *Ant Colony Optimization* (ACO), *Genetic Algorithm* (GA), dan lainnya.

DAFTAR PUSTAKA

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

Aydin, I., Karakose, M., & Akin, E. (2011). A multi-objective artificial immune algorithm for parameter optimization in support vector machine.*Computer Engineering Department* , 120-129.

Badrul, Mohammad (2012). Prediksi Hasil Pemilu Legislatif Dki Jakarta Dengan Metode *Neural Network* Berbasis *Particle Swarm Optimization* Tesis, Magister Ilmu Komputer,STMIK Nusa Mandiri, Jakarta

Dong, Y., Xia, Z., Tu, M., & Xing, G. (2007). An Optimization Method For Selecting Parameters In Support Vector Machines. *Sixth International Conference On Machine Learning And Applications* , 1.

Handayanna,Frisma (2012). Penerapan *Particle Swarm Optimization* Untuk Seleksi Atribut Pada Metode *Support Vector Machine* Untuk Prediksi Penyakit DiabetesTesis, Magister Ilmu Komputer,STMIK Nusa Mandiri,Jakarta

Huang, K., Yang, H., King, I., & Lyu, M. (2008).*Machine Learning Modeling Data Locally And Globally.* Berlin Heidelberg: Zhejiang University Press, Hangzhou And Springer-Verlag Gmbh.

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

Lasut, Desiyanna (2012). Prediksi Loyalitas Pelanggan Pada Perusahaan Penyedia Layanan Multimedia Dengan Algoritma C4.5 Berbasis *Particle Swarm Optimization* Tesis,Program Studi Teknik Informatika Program Pasca Sarjana Magister Komputer,STMIK Eresha,Jakarta

Maimon, O. (2010). *Data Mining And Knowledge Discovery Handbook.* New York Dordrecht Heidelberg London: Springer.

Masripah, Siti (2011). Algoritma klasifikasi c4.5 berbasis *particle swarm optimization* untuk evaluasi penentuan kelayakan pemberian kredit Koperasi syariah Tesis, Magister Ilmu Komputer,STMIK Nusa Mandiri,Jakarta

Septiani, Dwi Wisti (2013). Analisa Dan Komparasi Metode Klasifikasi Data Mining Algoritma C4.5, *Naïve Bayes*,Dan *Neural Network* Untuk Prediksi Penyakit Hepatitis Tesis, Magister Ilmu Komputer,STMIK Nusa Mandiri,Jakarta

Salappa, A., Doumpos, M., & Zopounidis, C. (2007). Feature SelectionAlgorithms in Classification Problems: An Experimental Evaluation. *SystemsAnalysis, Optimization and Data Mining in Biomedicine* , 199-212.

Park, T. S., Lee, J. H., & Choi, B. (2009).Optimization for Artificial NeuralNetwork with Adaptive

inertial weight of particle swarm optimization.*CognitiveInformatics, IEEE International Conference* , 481-485.

Rinawati (2012).Penerapan *Particle Swarm Optimization* Untuk Seleksi Atribut Pada Metode *Support Vector Machine* Untuk Penentuan Penilaian Kredit Tesis, Magister Ilmu Komputer,STMIK Nusa Mandiri,Jakarta

Sousa, T., Silva, A., & Neves, A. (2004). Particle Swarm Based Data Mining Algorithms for Classification Tasks. *Parallel Computing , 30*, 767-783.

Witten, I. H., Eibe, F., & Hall, M. A. (2011).*Data Mining: Practical Machine Learning Tools and Techniques 3D Edition*. United State.

X. Hu, R. Eberhart, and Y. Shi. *Recent advances in particle swarm*, , IEEE Congress on Evolutionary Computation 2004, Portland, Oregon, USA