Prediksi Penyakit Hepatitis Menggunakan Algoritma Naïve Bayes Dengan Seleksi Fitur Algoritma Genetika

Duwi Cahya Putri Buani
Program Studi Teknik Informatika
dputribuani8@gmail.com

Abstract – *Hepatitis is one of the world's health-threatening diseases, so there is a need for specialized treatment for hepatitis. The field of Information Technology is one that can be applied to predict hepatitis disease. With the branch of data mining science of health problems can be overcome. Naïve bayes is one of the existing algorithms in data mining that can be used to make predictions, so that the accuracy of naive bayes algorithm predictions is increased so that feature selection using a genetic algorithm can be used. From the application of the method it can be predicted that hepatitis disease has accuracy of 96, 77%, this prediction result increased from previous research by using the same algorithm that naïve bayes algorithm without done selection of the result result of accuracy is 83, 71%, difference of result of previous research with this research is 13.06%.*

Keywords: Genetic Algorithm, Naïve Bayes, Feature Selection

Abstraksi - Hepatitis adalah satu dari penyakit yang merupakan ancaman kesehatan didunia, sehingga perlu adanya penangannan khusus untuk penyakit hepatitis. Bidang Teknologi Informasi adalah salah satu yang dapat diaplikasikan untuk memprediksi penyakit hepatitis. Dengan cabang ilmu data mining permasalahan kesehatan tersebut dapat ditanggulangi. Naïve bayes adalah salah satu dari algoritma yang ada didalam data mining yang dapat digunakan untuk melakukan prediksi, agar akurasi dari predikisi algoritma naïve bayes meningkat maka dapat digunakan seleksi fitur dengan menggunakan algoritma genetika, dari penerapan metode tersebut dapat dihasilkan predikisi untuk penyakit hepatitis memiliki akurasi sebesar 96, 77%, hasil prediksi ini meningkat dari penelitian sebelumnya dengan menggunakan algoritma yang sama yaitu algoritma naïve bayes tanpa dilakukan seleksi fitur hasil akurasinya adalah 83, 71%, selisih hasil penelitian sebelumnya dengan penelitian ini adalah 13.06%.

Kata Kunci: ***Algoritma Genetika, Naïve Bayes, Seleksi Fitur***

A. PENDAHULUAN

Hepatitis adalah satu dari penyakit yang merupakan ancaman kesehatan didunia. Pasien yang terkena virus hepatitis tidak yakin bagaimana dapat terserang penyakit ini.Virus hepatitis dapat menimbulkan problema pasca akut bahkan dapat terjadi *cirroshis hepatitis* dan *karsinoma hepatoseluler primer*. Sepuluh persen dari infeksi virus hepatitis akan menjadi kronik dan 20 % penderita hepatitis kronik ini dalam waktu 25 tahun sejak tertular akan mengalami cirroshis hepatis dan karsinoma hepatoselluler (hepatoma).

Indonesia merupakan salah satu negara yang memiliki edemisitas tinggi Hepatitis B, terbesar kedua di negara sout east asen regional (SEAR) setelah negara Myanmar. Berdasarkan hasil riset kesehatan dasar (Riskesdas), studi dan uji saring darah PMI maka diperkirakan 100 orang di Indoneisa, 10 orang diantaranya terindentifikasi virus hepatitis B dan C. Sehingga saat ini diperkirakan 28 juta penduduk Indonesia terserang penyakit Hepatitis dan 14 Juta diantaranya berpotensi untuk menjadi kronis dan dari yang kronis tersebut 1,4 juta berpotensi menjadi kangker hati (RI, 2014). Dari uraian diatas maka diperlukan penanganan yang khusus untuk penyakit hepatitis, salah satunya adalah dengan melakukan diagnose dini terhadap penyakit hepatitis, salah satu cara untuk melakukan detiksi dini terhadap penyakit hepatitis adalah dengan memanfaatkan Teknologi Informasi, Seiring dengan perkembangan ilmu pengetahuan dan teknologi informasi, kehadiran cabang ilmu baru dibidang komputer, data mining telah menarik banyak perhatian dalam dunia sistem informasi (Septiani, 2014). Data mining adalah salah satu cabang ilmu dari teknologi Informasi yang algoritma didalam metode data mining dapat digunakan untuk melakukan prediksi.

Penelitian sebelumnya yang dilakukan oleh Wisti Dwi Septiani dalam Jurnal Tecno Vol. XI No. 1, Maret 2014 dengan menggunakan salah satu metode data mining yaitu dengan menggunakan metode C45 untuk memprediksi penyakit hepatitis memperoleh akurasi sebesar 77,29% (Septiani, 2014) dan pada tahun 2017 Wisti Dwi Septiani juga melakukan riset yang sama dengan

membandingkan atau melakukan komparasi dua metode data mining, metode tersebut adalah C45 dan Naïve Bayes dan hasilnya adalah metode Naïve Bayes memiliki akurasi prediksi yang lebih besar yaitu sebesar 83,71% (Septiani, 2017). Dari penelitian diatas maka penulis mencoba melakukan penelitian dengan menggunakain metode data mining Naïve Bayes dengan melakukan seleksi fitur menggunakan Algoritma Genetika.

B. TINJAUAN PUSTAKA

1. Data Mining

Data mining merupakan perpaduan dari ilmu statistik, kecerdasan buatan (sitem pakar) dan penelitian dalam bidang database, untuk itu diperlukan penyaringan melalui sejumlah besar material data atau melakukan penyelidikan dengan cerdas tentang keberadaan suatu data yang memiliki nilai (Witten, Frank, & Hall, 2011).

Knowledge discovery from data (KDD) juga merupakan bagian dari proses data mining, dimana dalam proses penjelajahan pengetahuan dimulai dari beberapa database dengan melakukan proses cleaning dan integration sehingga menghasilkan data warehouse. Selanjutnya melakukan proses selection dan transformation kemudian sebut sebagai data mining untuk menemukan pola dan mendapatkan pengetahuan dari data(Han, Kamber, & Pei, 2012).

2. Naive Bayes

Naïve Bayes merupakan suatu bentuk klasifikasi data dengan menggunakan metode probabilitas dan statistik. Metode ini pertama kali dikenalkan oleh ilmuwan Inggris Thomas Bayes, yaitu digunakan untuk memprediksi peluang yang terjadi di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai teorema Bayes. Metode Teorema bayes kemudian dikombinasikan dengan naive yang diasumsikan dengan kondisi antar atribut yang saling bebas (Bramer, 2007).

Naïve Bayes adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class(Kusrini & Luthfi, 2009). Naïve Bayes didasarkan pada teorema Bayes yang memiliki kemampuan klasifikasi serupa dengan decision tree dan neural network. Naïve Bayes terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar.

Bayes rule digunakan untuk menghitung probabilitas suatu class. *Algoritma Naive Bayes* memberikan suatu cara mengkombinasikan peluang terdahulu dengan syarat kemungkinan menjadi sebuah formula yang dapat digunakan untuk menghitung peluang dari tiap kemungkinan yang terjadi.

Berikut adalah bentuk umum dari *teorema bayes*:

$$P(H|X) = \frac{P(H|X|P(H)}{P(X)}$$

Keterangan:

X = Data dengan *class* yang belum diketahui.

H = *Hipotesis* data X merupakan suatu *class spesifik*.

P(H|X) = *Probabilitas hipotesis* H berdasarkan kondisi X (*posteriori probability*).

P(H) = *Probabilitas Hipotesis* H (*prior probability*).

P(X|H) = *Probabilitas* X berdasar kondisi pada *Hipotesis* H

P(X) = Probabilitas dari X.

Metode algoritna Naïve bayes merupakan penyederhanaan *metode bayes*. Untuk mempermudah pemahaman, maka *Teorema Bayes* disederhanakan menjadi:

P(H|X) = P(X|H) P(X) *Metode Bayes rule* digunakan dan diterapkan untuk melakukan penghitungan terhadap *posterior* dan *probabilitas* dari data sebelumnya. Dalam analisis *bayesian*, fungsi klasifikasi akhir dihasilkan dengan menggabungkan kedua sumber informasi (*prior dan posterior*) untuk menghasilkan probabilitas menggunakan aturan *bayes*.

Naïve Bayes memiliki kelemahan yaitu attribut atau fitur independen sering salah dan hasil estimasi probabilitas tidak dapat berjalan optimal. Untuk mengatasi kelemahan tersebut salah satu caranya dengan metode pembobotan attribut untuk meningkatkan akurasi dari Naïve Bayes tersebut (Zaidi & Cerquides, 2013).

3. Algoritma Genetika

Algoritma genetika (AG) diperkenalkan pertama kali oleh John Holland (1975) dari Universitas Michigan, John Holland mengatakan bahwa setiap masalah yang berbentuk adaptasi(alami maupun buatan) dapat diformulasikan ke dalam terminologi genetika (Suryanto, 2007). Algoritma genetika merupakan suatu algoritma pencarian berdasarkan pada mekanisme seleksi alam dan genetika alam. Algoritma genetika dimulai dengan sekumpulan solusi awal(individu) yang disebut populasi. Satu hal yang sangat penting adalah bahwa satu individu menyatakan satu solusi. Populasi awal akan berevolusi menjadi populasi baru melalui

serangkaian iterasi (generasi). Pada akhir iterasi, algoritma genetika mengembalikan satu anggota populasi yang terbaik sebagai solusi untuk masalah yang dihadapi(Desiani & Muhammad, 2006). Pada setiap iterasi, proses evolusi yang terjadi adalah sebahai berikut:

a) Dua individu dipilih sebagai orang tua (*parent*) berdasarkan mekanisme tertentu. Kedua parent ini kemudian dikawinkan melalui operator crossover (kawin silang) untuk menghasilkan dua individu anak atau offspring.
b) Dengan probabilitas tertentu, dua individu anak ini mungkin mengalami perubahan gen melalui operator mutation.
c) Suatu skema penggantian *(replacement scheme)* tertentu diterapkan sehingga menghasilkan populasi baru.

Proses ini terus berulang sampai kondisi berhenti *(stopping condition)* tertentu. Kondisi berhenti bisa berupa jumlah iterasi tertentu, waktu tertentu, atau ketika variansi individu-individu dalam populasi tersebut sudah lebih kecil dari suatu nilai tertentu yang diinginkan.

C. METODE PENLITIAN

Penelitian dilakukan dengan cara melakukan eksperimen dalam bentuk sistem penunjang keputusan untuk memprediksi pasien pengidap penyakit hepatitis, dengan mengguanakan Metode data mining yaitu algoritma Naïve Bayes dengan Fitur Seleksi Algoritma Genetika. Data yang digunakan untuk melakukan penelitian adalah data primer dan data sekunder. Untuk mengukur tingkat akurasi dari prediksi menggunakan tools Rapid Miner 5.

Langkah-langkah yang dilakukan dalam penelitian ini adalah sebagai berikut:

1) Pengumpulan data

Pada tahap ini ditentukan data yang akan diproses. Mencari data yang tersedia, memperoleh data tambahan yang dibutuhkan, mengintegrasikan semua data kedalam data set, termasuk variabel yang diperlukan dalam proses.

2) Pengolahan data awal

Ditahap ini dilakukan penyeleksian data, data dibersihkan dan ditransformasikan kebentuk yang diinginkan sehingga dapat dilakukan persiapan dalam pembuatan model.

3) Metode yang diusulkan

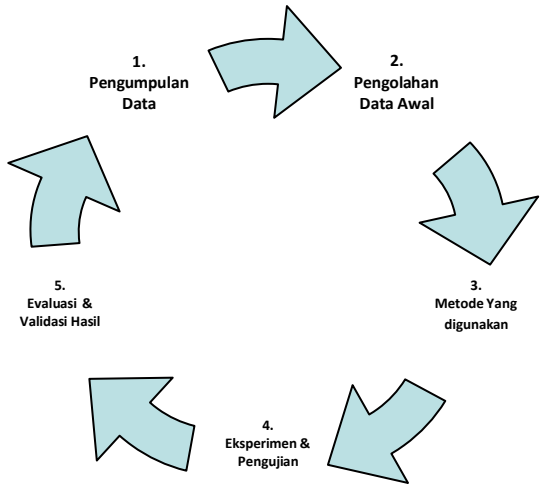
Pada tahap ini data dianalisis, dikelompokan variabel mana yang berhubungan dengan satu sama lainnya. Setelah data dianalisis lalu diterapkan model-model yang sesuai dengan jenis data.

4) Eksperimen dan pengujian metode

Pada tahap ini model yang diusulkan akan diuji untuk melihat hasil berupa rule yang akan dimanfaatkan dalam pengambilan keputusan.

5) Evaluasi dan validasi

Pada tahap ini dilakukan evaluasi terhadap model yang ditetapkan untuk mengetahui tingkat keakurasian model. Berikut gambar langkah-langkah dalam tahapan penelitian:



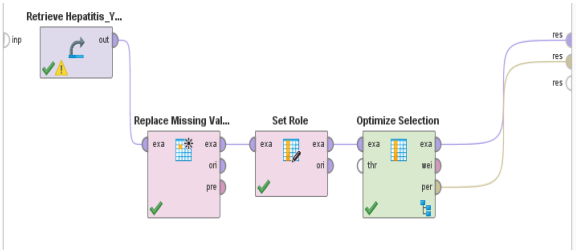
Sumber: (Buani, 2017)

Gambar 1. Langkah-langkah dalam melakukan penelitian

D. HASIL DAN PEMBAHASAN

Penelitian menggunakan tools Rapid Miner untuk melakukan olah data, berikut adalah tahapan dalam penggunaan rapid miner:

1. Pengujian Menggunakan Naïve Bayes dengan fitur seleski Algoritma Genetika.

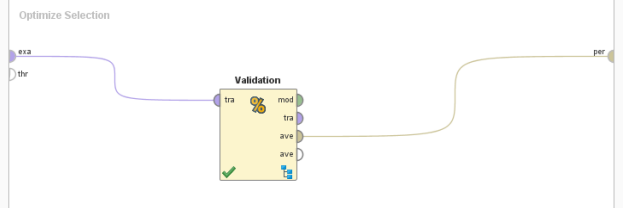


Sumber: (Buani, 2017)

Gambar 2. Koneksi Data dengan Optimaze Selection

Database Diagnosa Hepatitis dihubungkan dengan *Replace Missing Values* untuk memperbaiki data yang tidak lengkap, selanjutnya dihubungkan dengan *Set Role Attribute* untuk menjadikan salah satu dari atribut yang ada didalam data diagnosa hepatitis menjadi Id atau label dari semua atribut pada dataset, yang dijadikan label dari dataset diagnosa penyakit hepatitis adalah atribut Class, selanjutnya Set Role dihubungkan dengan operator *optimize selection (evolutionary)* untuk dilakukan pemilihan attribut-attribut yang relevan dengan proses prediksi hasil diagnose hepatitis. Didalam *optimaize selection (evalutionary)* terdapat

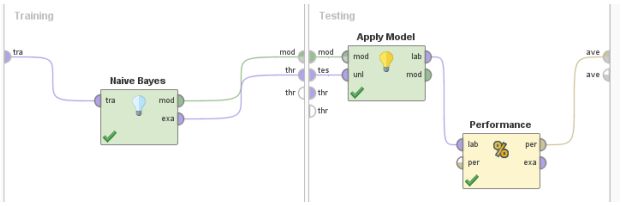
proses *cross validation* seperti yang terlihat pada Gambar 3.



Sumber: (Buani, 2017)

Gambar 3. Penggunaan Operator Cross Validation

Cross Validation yang digunakan dalam penelitian ini adalah 10-fold validation. Dataset yang berjumlah 155 data dengan 20 atribut akan dibagi menjadi 10 bagian. Dimana setiap bagian akan dibentuk secara acak. Prinsip adari 10-fold validation adalah 1:9, dimana 1 bagian menjadi data testing dan 9 bagian menjadi data traning, sehingga 10 bagian tersebut dapat berkesempatan menjadi data testing. Setelah dilakukan traning dan testing maka dapat diukut tingkat akurasinya. Di dalam cross validation terdapat proses penerapan algoritma naïve bayes seperti yang terdapat pada gambar 4.



Sumber: (Buani, 2017)

Gambar 4. Penggunaan Model Naïve Bayes

2. Hasil akurasi dari pengujian menggunakan Naïve Bayes dengan fitur seleksi algoritma genetika.

Setelah dilakukan penerapan model algoritma naïve bayes dengan optimasi menggunakan algoritma genetika maka hasil dapat dilihat pada Table 1.

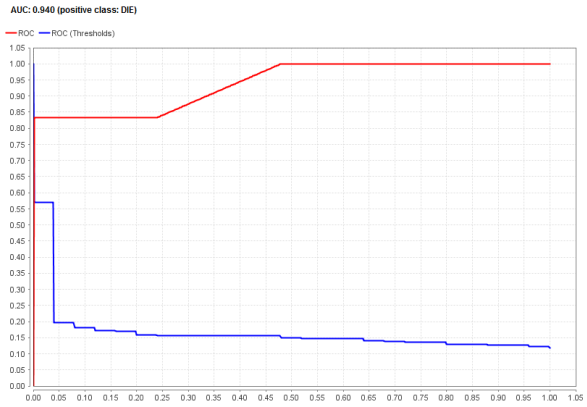
Table 1. Confussion Matrix

	True Life	Ture Die	Class Precision
Life	25	1	96,15%
Die	0	5	100%
Class Recall	100%	83,33%	
Accuracy	96,77%		

Sumber: (Buani, 2017)

Dari table diatas dapat diambil kesimpulan bahwa hasil prediksi menggunakan Naïve Bayes dengan seleksi fitur algoritma genetika tingkat akurasinya adalah 96,77%, hasil akurasi ini meningkat dari penelitian sebelumnya dengan menggunakan metode yang sama tetapi tidak menggunakan seleksi fitur sebesar 83,71% dan dengan menggunakan algoritma C45 hasil akurasi prediksinya adalah 77,29%.

3. Evaluasi dengan kurva ROC.



Sumber:(Buani, 2017)

Gambar 5. Kurva AUC *(Area Under Curve)*

Pada Gambar 5 menunjukkan grafik ROC dengan nilai AUC *(Area Under Curve)* sebesar 0.94 dengan tingkat diagnosa akurasi adalah *Excellent Classification* untuk model naive bayes dengan seleksi fitur algoritma genetika.

E. KESIMPULAN

Penelitian ini dilakukan untuk menguji hasil prediksi dari algoritma Naïve Bayes dengan seleksi fitur algoritma genetika, dah hasil prediksi yang didapatkan dalam pengujian ini adalah 96,77% hasil ini meningkat dari penelitian yang sebelumnya menggunakan data yang sama dan algoritma yang sama yaitu algoritma naïve bayes hasil prediksinya adalah 83, 71%, selisih dari penelitian sebelumnya dengan penelitian ini adalah 13.06%, selisih ini membuktikan bahwa tingkat akurasi dari algoritma naïve bayes setelah dilakukan seleksi fitur menggunakan algoritma genetika tingkat akurasinya lebih baik.

DAFTAR PUSTAKA

Bramer, M. (2007). Principles of Data Mining. Spinger. London: Springer. https://doi.org/10.1007/978-1-84628-766-4

Buani, D. C. P. (2017). Laporan Akhir Penelitian Mandiri. STMIK Nusa Mandiri Jakarta.

Desiani, A., & Muhammad, A. (2006). Konsep Kecerdasan Buatan. Yogyakarta: Cv. Andi Offset.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. (D. Cerra, Ed.), San Francisco, CA, itd: Morgan Kaufmann (Third Edit). San Francisco: The Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00001-0

Kusrini, & Luthfi, E. T. (2009). Algoritma Data Mining, 2009.

RI, K. K. (2014). InfoDATIN: Situasi dan Analisi Hepatitis. Pusat Data Dan Informasi. https://doi.org/24427659

Septiani, W. D. (2014). Penerapan Algoritma C4.5 Untuk Prediksi Penyakit Hepatitis Wisti. Jurnal Techno, Xi(1), 69–78.

Septiani, W. D. (2017). Komparasi Metode Klasifikasi Data Mining Algoritma C4.5 Dan Naive Bayes Untuk Prediksi Penyakit Hepatiti. Pilar Nusa Mandiri, 13(1), 76–84.

Suryanto. (2007). Artificial Intelligent, Searching, Reasoning Planning dan Learning. Bandung: Informatika Bandung.

Witten, I. H., Frank, E., & Hall, M. a. (2011). Data Mining: Practical Machine Learning Tools and Techniques (Google eBook). (J. Gray, Ed.), Complementary literature None (Second Edi). United States of America: Morgan Kaufmann. https://doi.org/0120884070, 9780120884070

Zaidi, N., & Cerquides, J. (2013). Alleviating Naive Bayes attribute independence assumption by attribute weighting. The Journal of Machine …, 14, 1947–1988. Retrieved from http://dl.acm.org/citation.cfm?id=2567725