Implementasi Data Mining Untuk Klasifikasi Penyakit Liver Dengan C4.5 Adaboost

Wiwid Wahyudi
Progdi Desain Grafis STEKOM Semarang, *wi2d.wahyudi@gmail.com*
Jl. Majapahit605, Semarang, telp/fax : 024-6717201-02

ABSTRACT

With the development of the times, it is undeniable that science and technology as well as the ease of development of the internet make it easier for people to identify liver disease, and become strong supporters of diseases with special needs. The initiative of computer experts with artificial intelligence or artificial intelligence, bioinformatics is defined as the application of computational and analytical tools to capture and view biological data. The amount of data on liver disease that continues to increase requires several methods to process and draw conclusions and information from the data, which is expected to be able to improve the quality of data or information as well as the efficiency and effectiveness of data

processing. ultimately facilitate or assist in policy making, especially in tackling the problem of liver disease. To support this, data mining techniques can be used to find valuable information from a collection of information or historical data of the liver. Research on liver using data mining classifications has been done, both comparisons of several classifications of data mining models or improvements to data mining classifications. Research on liver has been carried out and research has been carried out. To conduct this research, it is necessary to have a study of previous related research. In order to know what methods were used, what kind of data was studied, and what kind of model was produced. The dataset used does not contain missing values, so there is no need to preprocess the data. In this study, the C4.5, Support Vector Machine (SVM) and C4.5 Adaboost classification methods were used. The results of this study indicate that the C4.5 Adaboost method is 77.12% and the sensitivity value is 76.40%, where the value is greater than the two other methods of analysis used

Keywords: Liver,Classification, SVM, C4.5, Adaboost.

I. PENDAHULUAN

Dengan berkembangnya zaman, tidak dapat dipungkiri bahwa perkembangan ilmu pengetahuan dan teknologi serta kemudahan internet telah mempermudah masyarakat dalam mengidentifikasi penyakit liver, dan menjadi pendukung kuat penyakit berkebutuhan khusus. Inisiatif para ahli komputer dengan kecerdasan buatan atau kecerdasan buatan, bioinformatika didefinisikan sebagai aplikasi alat komputasi dan analisis untuk menangkap dan menafsirkan data biologis.

Jumlah data tentang penyakit liver yang terus meningkat memerlukan beberapa metode untuk mengolah dan mengambil kesimpulan dan informasi dari data tersebut, yang diharapkan akan mampu meningkatkan kualitas data atau informasi serta efisiensi dan efektivitas pengolahan data. Sehingga pada akhirnya memudahkan atau membantu dalam pembuatan kebijakan terutama dalam menanggulangi masalah penyakit liver.

Untuk mendukung hal ini dapat digunakan teknik data mining untuk menggali informasi yang berharga dari kumpulan informasi atau histori data liver.Penelitian tentang liver dengan menggunakan klasifikasi data mining sudah pernah dilakukan, baik komparasi beberapa klasifikasi data mining models ataupun improvement terhadap klasifikasi data mining.

Penelitian tentang liver telah banyak dilakukan dan telah dipublikasikan. Untuk melakukan penelitian ini, perlu ada kajian terhadap penelitian yang terkait sebelumnya. Agar dapat mengetahui metode apa saja yang digunakan, data seperti apa yang diproses, dan model seperti apa yang dihasilkan. Dataset yang digunakan tidak mengandung missing values sehingga tidak perlu dilakukan preprocessing data. Pada

penelitian ini digunakan metode klasifikasi C4.5, Support Vector Machine (SVM) dan C4.5 Adaboost. Hasil dari penelitian ini menunjukkan bahwa metode C4.5 Adaboost dengan sebesar 77.12% dan nilai sensitivitas sebesar 76.40%, dimana nilai tersebut lebih besar jika dibandingkan dengan ke dua metode analisis lainnya.

II. RUMUSAN MASALAH

Dari latar belakang di atas dapat diketahui rumusan masalahnya, yaitu belum diketahuinya teknik terbaik untuk penanganan imbalance data dan error klasifikasi, untuk mengoptimalkan akurasi Klasifikasi liver dan .bagaimana perbandingan dari hasil klasifikasi C4.5 Adaboost dan Suport Vector Machine (SVM) dalam diagnosis penyakit liver

III. TUJUAN PENELITIAN

Tujuan penelitian ini adalah Untuk meningkatkan akurasi pada algoritma C4,5 dalam Klasifikasi Status liver dan mengetahui perbandingan dari hasil klasifikasi C4.5 Adaboost dan Suport Vector Machine (SVM) dalam penentuan penyakit liver

III. KAJIAN PUSTAKA

A. Data Mining

Data Mining (DM) adalah inti dari proses Knowledge Discovery in Database (KDD), melibatkan algoritma yang mengeksploarasi data, mengembangkan model dan menemukan pola yang tidak diketahui sebelumnya. Model ini digunakan untuk memahami fenomena dari data, analisis dan prediksi. Aksesibilitas dan banyaknya data membuat Knowledge Discovery dan Data Mining menjadi masalah yang cukup penting dan dibutuhkan.

Menurut Larose (2007) berdasarkan tugasnya, data mining dibagi menjadi 6 kelompok [7], yaitu:

1. Deskripsi
   Terkadang peneliti dan analisis secara sederhana mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data..
2. Estimasi
   Estimasi hampir sama dengan klasifikasi, tetapi variable target estimasi lebih ke arah numerik daripada kategori.
3. Prediksi
   Prediksi hampir sama dengan estimasi dan klasifikasi, tetapi dalam prediksi akan menghasilkan nilai.
4. Klasifikasi
   Dalam klasifikasi terdapat target variabel kategori.
5. Cluster
   Mengelompokan *record*, pengamatan dan membentuk kelas obyek-obyek yang memiliki kemiripan. Tujuan dari algoritma *cluster* adalah dengan memecahkan setiap data dalam *dataset* menjadi kelompok kelompok yang homogen. Kelompok data ini biasanya disebut sebagi *cluster*. Setiap *cluster* yang terbentuk akan terdiri dari data yang sejenis dan berbeda dengan data pada *cluster* lainnya. Pengelompokkan ini sama dengan cara kerja otak manusia, dimana ilmu pengetahuan dikelompokkan dalam setiap bidangnya. Dengan adanya pengelompokkan, data yang dapat diolah dengan lebih spesifik sesuai dengan tujuan penelitian. Pemecahan data kedalam *cluster* data juga diterapkan pada tahappengolahan awal data dalam proses data mining, sehingga dapat diterapkan metode data mining data mining kedalam setiap *cluster* data. Proses *clustering* juga dapat mengurangi jumlah ataupun dimensi data yang diolah.

B. Liver Disorder

Hati merupakan organ intestinal paling besar dalam tubuh manusia. Beratnya rata-rata 1.2-1.8 kg atau kira-kira 2.5% berat badan orang dewasa. Di dalamnya terjadi pengaturan metabolisme tubuh dengan fungsi yang sangat kompleks dan juga proses-proses penting lainnya bagi kehidupan, seperti penyimpanan

energi, pembentukan protein dan asam empedu, pengaturan metabolisme kolesterol, dan detoksifikasi racun atau obat yang masuk dalam tubuh. Gangguan fungsi hati seringkali dihubungkan dengan beberapa penyakit hati tertentu. Beberapa pendapat membedakan penyakit hati menjadi penyakit hati akut atau kronis. Dikatakan akut apabila kelainan-kelainan yang terjadi berlangsung seampai 6 bulan, sedangkan penyakit hati kronis berarti gangguan yang terjadi sudah berlangsung lebih dari 6 bulan. Ada satu bentuk penyakit hati akut yang fatal, yakni kegagalan hati fulminan, yang berarti perkembangan mulai dari timbulnya penyakit hati hingga kegagalan hati yang berakibat kematian (fata) terjadi dalam kurang dari 4 minggu. Beberapa penyebab penyakit hati antara lain: 1. Infeksi virus hepatitis, dapat ditularkan memalui selaput mukosa, hubungan seksual atau darah (parenteral). 2. Zat-zat toksik, seperti alkohol atau obat-obatan tertentu. 3. Genetik atau keturunan, seperti hemochromtosis. 4. Gangguan imunologis, seperti hepatitis autoimun, yang ditimbulkan karena adanya perlawanan sistem pertahanan tubuh terhadap jaringan tubugnya sendiri. Pada hepatitis autoimun, terjadi perlawanan terhadap sel-sel hati yang berakibat timbulnya peradangan kronis. 5. Kanker, seperti Hepatocellular Carcinoma dapat disebabkan oleh senyawa karsinogenik antara lain aflatoksin, polivinil klorida (bahan pembuat plastik), virus, dan lain-lain. Hepatitis B dan C maupun sirosis hati juga dapat berkembang menjadi kanker hati (Depkes RI, 2007)

C. Suport Vector Machine (SVM)

Support Vector Machine (SVM) salah satu metode learning machine yang bekerja atas prinsip Structural Risk Minimization (SRM) dengan tujuan menemukan hyperplane terbaik yang memisahkan dua buah class pada input space. Selain itu SVM juga bertujuan untuk meminimalkan batas atas dari general eror. Keuntungan lain menggunakan SVM adalah metode ini dapat dianalisis secara teoritis menggunakan konsep teori pembelajaran komputasi. Prinsip dasar SVM adalah linier classifier, kemudian dikembangkan untuk dapat bekerja pada kasus non linier dengan memasukkan konsep kernel pada ruang kerja berdimensi tinggi . Problem klasifikasi dapat diterjemahkan dengan usaha menemukan garis (hyperplane) yang memisahkan antara kedua kelompok tersebut.

D. Algoritma C4.5

Algoritma C4.5 adalah hasil dari pengembangan algoritma ID3 (Iterative Dichotomiser) yang dikembangkan oleh [18]. Algoritma ini digunakan untuk membangun sebuah pohon keputusan yang mudah dimengerti, fleksibel, dan menarik karena dapat divisualisasikan dalam bentuk gambar. Sebelumnya diakhir tahun 1970 sampai awal tahun 1980 J. Ross Quinlan, seorang peneliti dibidang machine learning, membuat sebuah algoritma decision tree yang dikenal dengan ID3 (Iterative Dichotomiser). Kalau ID3, pengukuran seleksi atribut ditentukan oleh Information Gain, sedangkan C4.5 pengukuran seleksi atribut ditentukan oleh GainRatio.

E. Algoritma Esemble Adaboost

Adaboost adalah algoritma yang ide dasarnya adalah untuk memilih dan menggabungkan sekelompok pengklasifikasi lemah untuk membentuk klasifikasi yang kuat. Parameter yang telah dipelajari tersebut akan digunakan sebagai penelitian untuk dapat meningkatkan akurasi pengklasifikasi dasar C4.5 melalui iterasi yang tepat. [14] menjelaskan teknik pembobotan pada algoritma Adaboost sebagai berikut :

Inisialisasi bobot data $\{ W_n \}$ dengan $W_n^{(m)}$ untuk n = 1, 2,.....N For m = 1 ,M

a. Training $Y_m(x)$ dengan meminimalkan fungsi kesalahan (*error function*) sebagai berikut :

$$J_m = \sum_{n=1}^{N} W_n^{(m)} I(y_m(x_n) \neq t_n)$$

b. Evaluasi Kesalahan

$$\varepsilon_m = \frac{\sum_{n=1}^{N} W_n^{(m)} (y_m(x_n) \neq t_n)}{\sum_{n=1}^{N} W_n^{(m)}}$$

c. Dan kemudian digunakan evaluasi

$$a_m = ln\left\{\frac{1-\varepsilon_m}{\varepsilon_m}\right\}$$

d. Memperbaiki (Update) bobot data

$$W_n^{(m+1)} = W_n^{(m)} \exp(a_m I(y_m(x_n) \neq t_n))$$

e. Membuat prediksi menggunakan model terakhir sebagai berikut:

$$Y_m(x) = sign(\sum_{m=1}^{M} a_m y_m(x))$$

IV. PEMBAHASAN PENELITIAN

Dalam eksperimen ini, Awal Optimasi AdaBoost pada Algoritma C4.5 untuk Klasifikasi liver yaitu penerapan Algoritma C4.5 dengan SVM dan C4.5+AdaBoost pada Tool Rapid Miner kemudian didapatkan hasil yang bisa dilihat pada tabel dibawah ini :

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari UCI Machine Learning berupa data Liver Disorder dari semua hasil tes dengan responden laki-laki. Data Liver Disorder memuat tujuh atribut, dengan lima variabel pertama adalah semua tes darah yang dianggap sensitif terhadap gangguan hati yang mungkin timbul akibat konsumsi alkohol yang berlebihan. Adapun kategori diagnosis dibedakan menjadi tidak beresiko dan beresiko.

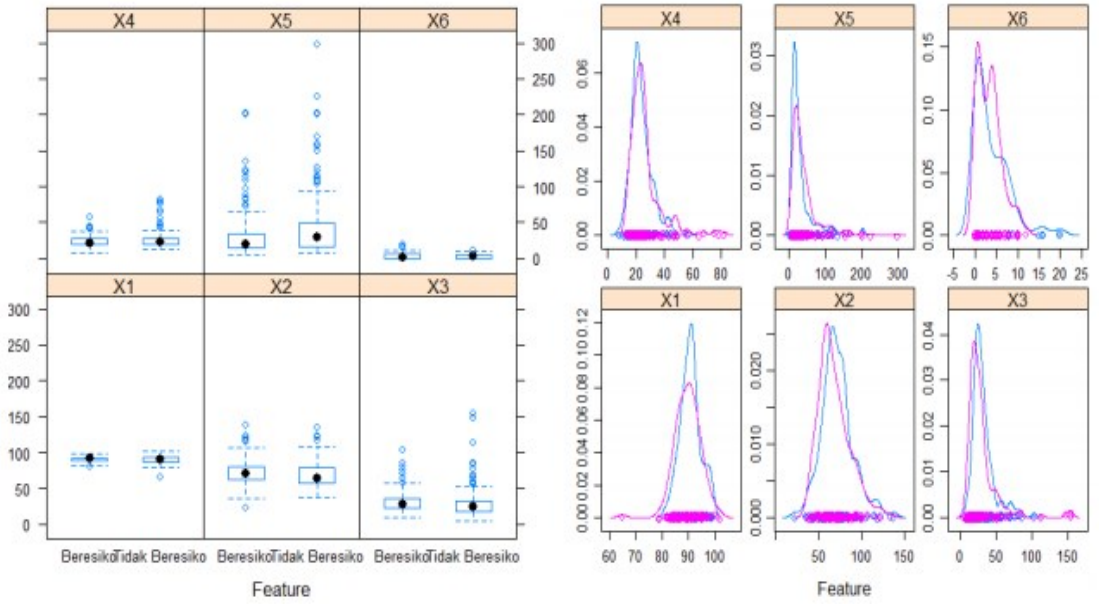
Tabel 4.1 Deskripsi Data

VARIABEL	KODE	DESKRIPSI
Selector	Y	Jenis resiko liver dibedakan menjadi (0) Tidak dan (1) beresiko
MCV	X_1	Rata-rata volume *corpuscular*
Alkpos	X_2	*Alkaline Phosphotase*
SGPT	X_3	*Alanine Aminotrensfrease*
SGOT	X_4	*Aspartate Aminotransferase*
Gammagt	X_5	*Gamma Glutamyl Transpeptidase*
Drinks	X_6	Jumlah konsumsi minuman yang mengandung alkohol per hari

Tabel 4.2 Sample Data

Selector (Y)	MCV (X1)	Alkphos (X2)	SGPT (X3)	SGOT (X4)	Gammagt (X5)	Drinks (X6)
Beresiko	85	92	45	27	31	0
Tidak Beresiko	85	64	59	32	23	0
Tidak Beresiko	86	54	33	16	54	0
Tidak Beresiko	91	78	34	24	36	0

Dari data penelitian diagnosis liver yang digunakan, variabel data penelitian dapat direpresentasikan pada visualiasi data seagai berikut:



Gambar 5.1 Visualisasi Data Penelitian

Dari visualisai data di atas, terlihat bahwa semua atribut data memiliki data outlier bahkan data ekstrim. Hal ini terlihat dari banyaknya titik data yang berada diluar boxplot. Selain itu jika dilihat dari density plot (kanan) terlihat pola grafik ynag condong ke arah kiri dengan data banyak terkonsentrasi pada sumbu X. Dari hasil visualisasi data penulis melakukan analisis regresi logistik untuk melihat hasil prediksi data liver disorder dengan metode statistika yang umum digunakan, sebelum dilakukan analisis dengan metode Machine Learning.

Tabel 5.2 Hasil Perbandingan Confusion Matrix

Metode	**Diagnosis**	**Tidak Beresiko**	**Beresiko**	**Total Akurasi**	**Total Sensitifitas**
C 4.5	Tidak Beresiko	82	23	0.6518	0.5540
	Beresiko	6	49		
SVM	Tidak Beresiko	185	28	0.7536	0.6552
	Beresiko	15	117		
C4.5Adaboost	Tidak Beresiko	185	28	**0.7712**	**0.7640**
	Beresiko	15	117		

Berdasarkan Tabel 5.2, menjelaskan hasil prediksi, total akurasi dan tingkat sensitifitas dari keenam metode machine learning. Bannyaknya data diagnosis Tidak Beresiko' yang dapat di prediksi dengan tepat oleh Machine Learning SVM adalah 185 data dan banyaknya diagnosis 'Tidak Beresiko' yang di prediksi salah Observed ada 28 data. Sehingga total akurasinya sebesar 75.36% dengan total sensitivitas sebesar 65.52%. Sedangkan untuk data diagnosis Beresiko yang dapat di prediksi dengan benar sebanyak 15 data dan 117 data tidak di prediksi dengan tepat. Untuk metode lainnya dapat di interpretasikan dengan cara yang sama. Nilai akurasi ini digunakan untuk melihat jenis diagnosis yang paling sesuai.

Pada hasil keseluruhan terlihat nilai akurasi untuk model algoritma C4.5 sebesar 65.18% dan nilai akurasi untuk model C4.5 +AdaBoost sebesar 77.12%. Akurasi Metode C4.5+Adaboost menghasilkan tingkat akurasi yang lebih baik dibandingkan dengan menggunakan C4.5 versi standar. Hal tersebut seperti dikatakan oleh [9] bahwa adaboost dapat memberikan keuntungan, lebih efektif dan akurat dalam

pengklasifikasian. Terbukti bahwa hasil pengujian algoritma C4.5 + AdaBoost memiliki nilai akurasi yang lebih baik dibandingkan dengan algoritma C4.5 tunggal dan algoritma SVM.

V. KESIMPULAN DAN SARAN

A. Kesimpulan

Dari hasil analisis dan pembahasan dalam penelitian ini, maka dapat diambil kesimpulan sebagai berikut :

1. Dari hasil klasifikasi yang diperoleh dengan menggunakan metode C4.5, Suport Vector Machine (SVM) dan C4.5 Adaboost memiliki hasil klasifikasi yang berbeda-beda setiap metodenya, seperti yang terlihat pada Tabel 5.2 diatas.
2. Metode yang paling baik untuk diagnosis liver adalah metode C4.5 Adaboost dengan tingkat akurasi sebesar 77.12% dan nilai sensitivitas sebesar 76.40%, dimana nilai tersebut lebih besar jika dibandingkan dengan kedua metode analisis lainnya.

B. Saran

Pada penelitian selanjutnya peneliti perlu memperhatikan jumlah atribut dan record dataset karena akan mempengaruhi nilai klasifikasi yang berbeda-beda.