IMPLEMENTASI DECISION TREE UNTUK MENDIAGNOSIS PENYAKIT LIVER

Intan Setiawati [1] , Adityo Permana Wibowo[2] , Arief Hermawan [3]

[1),3)] Magister Teknologi Informasi Universitas Teknologi Yogyakarta
[2)] Program Studi Teknik Informatika Universitas Teknologi Yogyakarta
email : intansetiawati@student.uty.ac.id[1)], adityopw@staff.uty.ac.id[2)], ariefdb@staff.utya.c.id[3)]

Abstraksi

Hati merupakan salah satu organ manusia yang paling penting. UCI Machine Learning Repository mempunyai banyak dataset, salah satunya adalah dataset ILPD (Indian Liver Patient Dataset). Penelitian ini membahas tentang klasifikasi penyakit liver pada dataset ILPD menggunakan Algoritma Decision Tree C4.5. Berdasarkan hasil pengolahan yang dilakukan, didapatkan bahwa Algoritma Decision Tree C4.5 menghaasilkan nilai akurasi sebesar 72.67% dan juga membuktikan bahwa dari 11 variabel penyakit liver yang ada pada dataset ILPD, hanya 2 variabel (Almine Alminotransferase) yang menjadi pokok dalam penentuan penyakit liver.

Kata Kunci*: *Decision Tree, Diagnosa, Klasifikasi, Penyakit Liver

Pendahuluan

Hati merupakan salah satu organ tubuh manusia yang paling penting. Fungsi hati adalah sebagai detoksifikasi racun yang ada dalam tubuh manusia serta mengendalikan kolesterol dan lemak yang ada dalam tubuh manusia. Jika organ hati mengalami kerusakan maka kesehatan pun akan terganggu, bahkan bisa mengalami kematian [1].

Beberapa penyakit yang menyerang hati salah satunya adalah Hepatitis. Menurut data WHO virus Hepatitis B menyerang 350 juta orang di dunia terutama Asia Tenggara dan Afrika yang menyebabkan kematian sebanyak 1,2 juta pertahun [2].

Pasien dengan penyakit hati semakin lama semakin meningkat, diantara penyebabnya adalah seringnya mengkonsumsi alcohol dan obat-obatan secara berlebihan. Selain itu menghirup gas berbahaya dan asupan makanan yang terkontaminasi juga menjadi penyebab penyakit hati [3].

Penelitian ini mencoba untuk mengolah dataset yang diambil dari database UCI *Machine Learning Repository,* yaitu *Indian Liver Patient Dataset* (ILPD Dataset) [4], dari dataset tersebut akan diolah menggunakan Algoritma C4.5 untuk mengetahui variabel mana yang paling berpengaruh dalam mendeteksi penyakit liver.

Berdasarkan pembahasan latar belakang diatas, penelitian ini akan membahas klasifikasi penyakit liver menggunakan Algoritma Decision Tree C4.5 dengan menggunakan ILPD Dataset. Penelitian ini juga akan membuktikan bahwa dari 10 variabel penentu penyakit Liver adakah variabel yang paling berpengaruh.

Tinjauan Pustaka

Penelitian Sebelumnya

Penelitian yang menggunakan ILPD Dataset sudah banyak dilakukan, pada tahun 2011 penelitian tentang membandingkan algoritma klasifikasi dengan menggunakan ILPD Dataset, dari penelitian tersebut menunjukkan bahwa algoritma KNN, Backpropagation, dan SVM memberikan hasil yang terbaik dari pada algoritma Naïve Bayes dan Decision Tree C4.5 [3]. Kemudian, pada tahun 2012 dilakukan penelitian dengan membandingkan dataset pasien penyakit liver dari ILPD dan UCLA. Dari dua dataset tersebut terdapat atribut yang sama yaitu Alkphos, SGPT, SGOT. Dari 3 atribut tersebut diolah menggunakan ANOVA dan MANOVA, menghasilkan bahwa tidak ada perbedaan yang signifikan antara pasien non-hati dari dataset UCI dan India dengan pasien non-hati dari AS dan India [5]. Selanjutnya, masih ditahun yang sama, penelitian tentang klasifikasi menggunakan dataset penyakit liver dilakukan dengan menggunakan Metode Rotation Forest yang dimodifikasi. Metode tersebut dilakukan modifikasi untuk mengetahui kedekatan dengan metode klasifikasi lainnya. Hasilnya, modifikasi Metode Rotation Forest diusulkan dengan metode Multilayer Perception dalam pemilihan fitur subset secara acak untuk dataset UCI. Modifikasi Metode Rotation Forest diusulkan dengan metode Nearest Neighbor untuk pemilihan fitur berdasarkan korelasi untuk dataset hari India [6].

Pada tahun 2015, dilakukan penelitian dengan membandingkan metode klasifikasi Support Vector Machine (SVM) dengan Naïve Bayes Classifier (NBC) untuk mengolah ILPD Dataset. Pembanding 2 metode tersebut adalah akurasi klasifikasi dan

waktu eksekusi dalam melakukan klasifikasi penyakit liver. Hasil penelitian tersebut menunjukakn bahwa dalam hal akurasi klasifikasi Metode SVM menghasilkan nilai akurasi lebih baik dari pada NBC, sedangkan dalam hal waktu pemrosesan, NBC lebih cepat dari pada SVM [7]. Masih di tahun yang sama, penelitian menggunakan ILPD Dataset juga pernah dilakukan. Penelitian tersebut membandingkan kinerja metode Naïve Bayes Calssifier (NBC) dengan Algoritma C4.5 dalam proses klasifikasi penyakit liver. Penelitian tersebut membandingkan nilai akurasi dalam proses klasifikasi. Hasilnya Algoritma C4.5 menghasilkan nilai akurasi sebesar 69,828%, sedangkan Algoritma NBC menghasilkan nilai akurasi sebesar 63,362. Dengan demikian Algoritma C4.5 memberikan pemecahan yang lebih baik dari pada Algoritma NBC dalam hal klasifikasi penyakit liver [2].

Kemudian pada tahun 2018, juga pernah dilakukan penelitian menggunakan ILPD Dataset. Penelitian tersebut membahas tentang perbandingan nilai akurasi algoritma klasifikasi dalam hal pemrosesan klasifikasi penyakit liver. Algoritma klasifikasi yang digunakan antara lain Naïve Bayes Classifier (NBC), Ada Boost, J48, Bagging, dan Random Forest. Dalam proses perbandingan metodenya, digunakan aplikasi bantuan yaitu WEKA. Dari penelitian tersebut menghasilkan bahwa Metode Random Forest memberikan kinerja yang lebih baik dari pada Metode NBC, Ada Boost, J48, dan Bagging [8]. Pada tahun yang sama, dilakukan penelitian menggunakan ILPD Dataset juga, hanya saja metode yang digunakan adalah kombinasi rule Algoritma Cart dan Algoritma Ripper, atau biasa disingkat dengan C-Ripper. Penelitian tersebut mencoba mengolah ILPD Dataset dengan seleksi fitur dan tanpa seleksi fitur. Hasilnya menunjukkan bahwa kombinasi Algoritma Cart dan Ripper dengan seleksi fitur menghasilkan nilai akurasi sebesar 70%, sedangkan tanpa seleksi fitur menghasilkan nilai akurasi sebesar 81% sehingga dengan demikian Algoritma Cart dan Ripper lebih bagus nilai akurasinya jika tidak melakukan seleksi fitur [1].

Selain penelitian yang menggunakan ILPD Dataset, penelitian lain yang tidak menggunakan ILPD Dataset tetapi menggunakan Algoritma C4.5 juga sudah banyak dilakukan. Pada tahun 2017, pernah dilakukan penelitian mengenai klasifikasi nilai kelayakan calon debitur baru untuk pembiayaan sepeda motor menggunakan Decision Tree C4.5. luaran pada penelitian tersebut yaitu nilai kelayakan yang terdiri dari Lunas dan Tarikan. Jika nilai kelayakan Lunas, maka calon debitur tersebut diprediksi lancar dalam hal pembayaran kredit, sedangkan jika nilai kelayakannya Tarikan, maka diprediksi calon debitur tersebut akan berpotensi kredit macet. Hasil penelitian tersebut menunjukan bahwa Algoritma C4.5 menghasilkan nilai akurasi lebih dari 70% dengan waktu proses kurang dari 15 menit [9].

Penelitian lain tentang algoritma C4.5 pernah dilakukan untuk memprediksi tingkat kesuburan pria. Factor yang mempengaruhi tingkat kesuburan pria antara lain, lama duduk, konsumsi alkohol, dan kebiasaan merokok. Dari penelitian tersebut menunjukkan bahwa Algoritma C4.5 menghasilkan nilai akurasi sebesar 92% dalam memprediksi tingkat kesuburan pria [10].

Klasifikasi

Klasifikasi merupakan suatu proses menemukan model yang dapat membedakan dan menggambarkan kelas pada suatu data. Model tersebut terbentuk berdasarkan analisis data pelatihan. Model turunan dapat direpresentasikan dalam beberapa bentuk, antara lain bentuk aturan IF-THEN, aturan pohon keputusan, aturan rumus matematika, atau jaringan saraf [11].

Banyak metode untuk membangun model klasifikasi, yaitu Naïve Bayes Classifier, Support Vector Machine, dan K-NN [11].

Dengan kata sederhananya, klasifikasi adalah proses pengelompokkan yang sudah diketahui dengan jelas jumlah dan nama kelompoknya.

Decision Tree

Decision *Tree Learning* (DTL) merupakan salah satu teknik pembelajaran mesin (*Machine Learning*) yang menggunakan aturan klasifikasi berstruktur sekuensial hirarki dengan cara mempartisi himpunan data latih secara rekursif [12]. Ada beberapa metode yang termasuk DTL, salah duanya adalah ID3 dan C4.5. Metode ID3 digunakan untuk data kategorial, sedangkan C4.5 digunakan untuk kategorial dan numerik. Metode ID3 menggunakan Information Gain, sedangkan C4.5 menggunakan Gain Ratio [12].

Entropy

Pada dasarnya entropy adalah perhitungan probabilitas yagn diadopsi pada Algoritma Decision Tree C4.5 untuk mengukur tingkat distribusi kelas apda sebuah dataset. Perhitungan nilai entropy menggunakan rumus yang terlihat pada persamaan dibawah ini [13].

$$\text{Entropy(X)} = \sum_{i=0}^{m} -pi * log2\ pi$$

Keterangan:
X : himpunan kasus
m : jumlah partisi variabel dari himpunan s
p_i : probabilitas kasus dalam partisi ke-i

Information Gain

Information Gain merupakan perubahan dari *entropy* yang sudah dibagi atributnya pada sebuah dataset, menjadi subset terkecil. Perhitungan

information gain merupakan selisih antara *entropy* dataset sebelum dan sesudah pembagian. Pembagian terbaik akan menghasilkan *entropy* yang paling kecil sehingga berdampak pada information gain yang terbesar. Pemilihan atribut sebagai *root* berdasarkan nilai gain tertinggi dari atribut yang ada. Perhitungan information gain menggunakan persamaan dibawah ini [13].

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \left(\frac{|Si|}{|S|} * Entropy(Si) \right)$$

Keterangan:
S : Himpunan kasus
A : Variabel penentu
n : Jumlah partisi A
|Si| : jumlah kasus pada partisi ke-i
|S| : Jumlah kasus dalam S

Gain Ratio

Sebelum perhitungan *gain ratio*, perlu dihitung terlebih dahulu *Split Information* seperti pada persamaan dibawah ini [12].

$$SplitInformation(S,A) \equiv \sum_{i=1}^{c} -\frac{|S_i|}{|S|} \log \frac{|S_i|}{|S|}$$

Keterangan:
S : himpunan sampel data
S_1 : subhimpunan sampel data yang terbagi berdasarkan jumlah variasi nilai pada atribut A.

Selanjutnya *gain ratio* dirumuskan dengan Information Gain dibagi Split Information, seperti dibawah ini [12].

$$GainRatio(S,A) \equiv \frac{Gain(S,A)}{SplitInformation(S,A)}$$

Rumus bisa mereduksi bias dalam penentuan atribut pemilah terbaik.

Metode Penelitian

Penelitian yang dilakukan meliputi pengolahan dataset ILPD menggunakan bantuan aplikasi. Dalam hal ini aplikasi bantuan yang dimaksud adalah Rapidminer versi 7.4.

Dataset

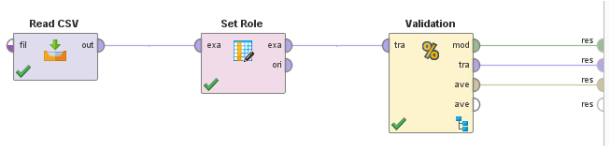
Dataset yang digunakan pada penelitian ini diambil dari database UCI *Machine Learning Repository* [4]. ILPD dataset berisi 583 data klinis dengan 10 atribut dengan target output positif *liver* sebanyak 416 dan *negative liver* sebanyak 167. Dari 583 data yang diproses, 433 data digunakan sebagai data latih, 150 data digunakan sebagai data uji. Atribut yang ada dalam *dataset* ILPD seperti ditampilkan pada Tabel 1.

Tabel 1. Atribut Dataset ILPD

No	Atribut	Tipe data	Keterangan
1.	Age	*Numeric*	Umur pasien
2.	Gender	*Text*	Jenis Kelamin pasien
3.	TB (Total Bilirubin)	*Numeric*	Pigmen berwarna kuning kecoklatan yang ada didalam empedu, darah dan tinja.
4.	DB *(Direct Bilibruin)*	*Numeric*	Pigmen berwarna jingga kuning sisa dari perombakan sel darah merah (langusng)
5.	Alkphos (Alkaline Phosphotase)	*Numeric*	Enzim hydrolase yang terutama ditemukan pada sebagian besar organ tubuh, terutama di tulang, tulang dan plasenta.
6.	Sgpt_AA *(Almine Aminotransferase)*	*Numeric*	Enzim yang sering dijumpai di serum darah dan berbagai jaringan tubuh, tetapi sering dikaitkan dengan kerusakan hati
7.	Sgot_AA *(Aspartate Aminotransferase)*	*Numeric*	Enzim yang berkaitan dengan kinerja organ hati
8.	TP (Total Proteins)	*Numeric*	Berisi Albumin dan Globulin
9.	ALB *(Albumin)*	*Numeric*	Protein Utama pada darah yang diproduksi hati
10.	A/G (Ratio Albumin Globulin Ratio)	*Numeric*	Perbandingan albumin dan globulin yang merupakan konstituen utama protein yang ditemukan dalam darah
11.	Class Variabel	*Numeric*	Yes: Positif liver No: Negatif liver

Pemodelan Klasifikasi

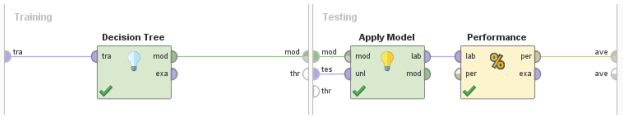
Pada penelitan ini, pemodelan klasifikasi menggunakan aplikasi bantuan yaitu Rapidminer. Pada aplikasi Rapidminer menggunakan 3 (tiga) komponen utama dan 3 (tiga) komponen bantuan. Komponen utama yang digunakan yaitu Read CSV yang berfungsi untuk mengambil dataset yang berbasis .CSV. Kemudian disambungkan dengan Set Role yang berfungsi untuk menentukan tipe data variable, yaitu tipe data Label. Kemudian untuk menentukan jumlah data latih dan data uji menggunakan komponen Split Validation. Susunan penggunaan komponen tersebut seperti pada Gambar 1.



Gambar 1. Komponen Perhitungan Rapidminer

Kemudian untuk proses pemodelan klasifikasi, nilai akurasi, dan pembentukan rule Decision Tree C4.5 digunakan bebrapa komponen, antara lain Decision Tree untuk pemodelan pohon keputusannya. Komponen Apply Model dan Performance digunakan untuk perhitungan nilai akurasi dan pembentukan rule pohon keputusan.

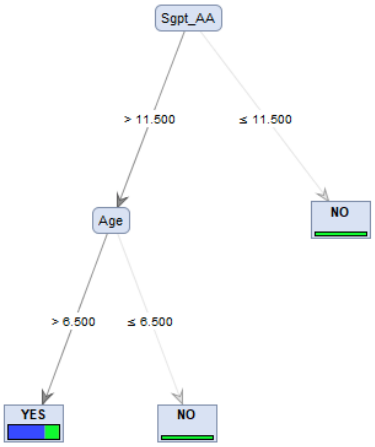
Susunan penggunaan komponen tersebut seperti pada Gambar 2.



Gambar 2. Komponen perhitungan Decision Tree

Hasil dan Pembahasan

Pemrosesan dataset ILPD dengan Decision Tree menggunakan Rapidminer menghasilkan pohon keputusan dan nilai akurasi. Pohon keputusan yang dihasilkan nantinya akan digunakan untuk rule atau aturan dalam klasifikasi penyakit liver. Pohon keputusan yang dihasilkan terdiri dari 2 node. Seperti terlihat pada Gambar 3.



Gambar 3. Pohon keputusan Dataset ILPD

Berdasarkan Gambar 3, terlihat bahwa hasil pemrosesan dataset ILPD hanya 2 node yang terdiri dari Almino Aminotransferase (Spgt_AA) dan umur (Age), dimana SPGT_AA menjadi *root* pada pohon tersebut. Dengan kata lain bahwa variabel penentu penyakit liver yang ada pada dataset ILPD jika diolah menggunakan Decision Tree hanya 2 variabel (SPGT_AA dan Age) yang paling berpengaruh untuk menentukan atau mengklasifikasikan seseorang menderita penyakit liver.

Setelah terbentuk aturan klasifikasi penyakit liver, selanjutnya adalah perhitungan nilai akurasi berdasarkan Dataset ILPD yang digunakan. Untuk menghitung nilai akurasi, dari 583 data yang digunakan, sebanyak 433 digunakan sebagai data latih dan sebanyak 150 digunakan sebagai data uji. Nilai akurasi yang dihasilkan adalah 72.67%, seperti terlihat pada Gambar 4.

accuracy: 72.67%

	true YES	true NO	class precision
pred. YES	107	41	72.30%
pred. NO	0	2	100.00%
class recall	100.00%	4.65%	

Gambar 4. Nilai Akurasi

Kesimpulan dan Saran

Penelitian ini menunjukkan bahwa hanya 2 variabel (SPGT_AA dan Age) diantara 10 variabel pada dataset ILPD yang paling berpengaruh dalam penentuan klasifikasi penyakit liver. Penelitian ini juga menunjukkan hasil akurasi sebesar 72.67% dalam penentuan klasifikasi penyakit liver menggunakan Dataset ILPD.

Daftar Pustaka

[1] D. Restiani, "KOMBINASI ALGORITMA C-RIPPER UNTUK MENDIAGNOSIS PENYAKIT LIVER," *J. Tek. Inform.*, vol. 11, no. 1, pp. 31–36, 2018.

[2] E. Rahmawati, "ANALISA KOMPARASI ALGORITMA NAIVE BAYES DAN C4.5 UNTUK PREDIKSI PENYAKIT LIVER," *J. Techno Nusa Mandiri*, vol. XII, no. 2, pp. 27–37, 2015.

[3] B. V. Ramana, P. M. Surendra, P. Babu, and P. N. B. Venkateswarlu, "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis," *Int. J. Database Manag. Syst.*, vol. 3, no. 2, pp. 101–114, 2011.

[4] S. P. V. Ramana, Bendi Venkata; Babu, "UCI Machine Learning Repository," *https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)*, 2012. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset).

[5] B. V. Ramana, P. M. Surendra, P. Babu, and P. N. B. Venkateswarlu, "A Critical Comparative Study of Liver Patients from USA and INDIA : An Exploratory Analysis," *Int. J. Comput. Sci.*, vol. 9, no. 3, pp. 506–516, 2012.

[6] B. V. Ramana, P. M. Surendra, and P. Babu, "Liver Classification Using Modified Rotation Forest," *Int. J. Eng. Res. Dev.*, vol. 1, no. 6, pp. 17–24, 2012.

[7] S. Vijayarani and S. Dhayanand, "Liver Disease Prediction using SVM and Naïve Bayes Algorithms," *Int. J. Sci. Eng. Technol. Res.*, vol. 4, no. 4, pp. 816–820, 2015.

[8] A. Pathan, D. Mhaske, S. Jadhav, R. Bhondave, and K. Rajeswari, "Comparative Study of Different Classification Algorithms on ILPD Dataset to Predict Liver," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 6, no. II, pp. 388–394, 2018.

[9] B. A. Hermanto, "Klasifikasi Nilai Kelayakan Calon Debitur Baru Menggunakan Decision Tree C4.5," *Indones. J. Comput. Cybern. Syst.*, vol. 11, no. 1, pp. 43–54, 2017.

[10] A. Amrulloh and A. P. Wibowo, "IMPLEMENTASI ALGORITMA DECISSION TREE UNTUK MENGKLASIFIKASI KONDISI KESUBURAN PRIA," *J. Apl. Sains, Informasi, Elektron. dan Komput.*, vol. 1, no. 1, pp. 7–11, 2019.

[11] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Tecniques*, 3rd ed. Wyman Street, Waltham, MA, USA: Elsevier Inc., 2012.

[12] Suyanto, *Machine Learning Tingkat Dasar dan Lanjut*, 1st ed. Bandung: Informatika Bandung, 2018.

[13] B. Hermanto, "Prediksi Potensi Pelunasan Sebagai Salah Satu Kriteria Kelayakan Calon Debitur Menggunakan Decision Tree C4.5," Universitas Gadjah Mada, 2016.