Jurnal Cahaya Mandalika, Vol. 5, No. 1, 2024, e-ISSN: 2721-4796, p-ISSN: 2828-495X
Available online at: http://ojs.cahayamandalika.com/index.php/JCM
Akreditasi Sinta 5 SK. Nomor: 1429/E5.3/HM.01.01/2022

KLASIFIKASI PENYAKIT SIROSIS HATI DENGAN CART

Yasmin Roni Mz[1§], I Komang Gde Sukarsa[2], I Gusti Ayu Made Srinadi[3]
[1,2,3]Program Studi Matematika, Fakultas MIPA – Universitas Udayana
[§]*Corresponding Author:* yasminroni2907@gmail.com[§] , gedesukarsa@unud.ac.id[2] , srinadi@unud.ac.id[3]

Abstract: *The nonparametric exploratory is a method that can be used to see the relationship between the dependent variable and the independent variable. One of the types of nonparametric exploratory methods is the CART. CART is a method that presents large amounts of data to be processed in the form of a decision tree so that it becomes valuable and easy to understand information. This research aims to build a decision tree model based on medical records from patients with liver cirrhosis using the CART. This research also used 16 independent variables of 276 data that will be used as research objects. The results of this study obtained a decision tree model with an independent variable. The first used as the root node is hepatomegaly because the hepatomegaly variable has a more homogeneous value compared to the other independent variables and that there were eight groups in this research. However, due to the nature of the CART method which is unstable and very sensitive to new data, and highly dependent on the number of samples, the accuracy rate in this study is less than 70%, this is because the data in one group is unbalanced if compared to data in the other group.*
Keywords*: CART, Liver Cirrhosis, Decision Tree, Nonparametric Regression, Entropy*

PENDAHULUAN

Jaringan – jaringan yang mempunyai satu atau lebih fungsi di dalam tubuh makhluk hidup disebut dengan organ (KBBI, 2022). Salah satu organ di dalam tubuh manusia adalah organ hati yang terletak pada bagian kanan atas perut tepat di bawah tulang rusuk yang normalnya berwarna merah kecoklatan. Hati memiliki fungsi vital dalam proses metabolisme tubuh sehingga, apabila hati tidak berfungsi dengan baik dapat menyebabkan peradangan atau inflamasi.

Penyakit inflamasi pada hati salah satunya adalah sirosis hati yaitu keadaan patologis karena terdapat luka pada hati sehingga hati membentuk jaringan parut untuk mengantikan jaringan normal pada hati. Jaringan parut yang terbentuk secara terus menerus dapat menghalangi aliran darah ke organ.

Penyakit sirosis hati menurut WHO menyebutkan pertambahan 3 – 4 juta orang/tahun dengan 3% populasi manusia menderita penyakit ini. Penyebab penyakit sirosis hati pada negara barat dan Indonesia memiliki perbedaan apabila, pada negara barat penyebab sirosis hati adalah karena kebiasaan meminum alkohol sedangkan pada negara Indonesia penyebab sirosis hati karena penyakit hepatitis B dan hepatitis C.

Gejala awal penyakit sirosis hati adalah peradangan pada hati karena melawan infeksi oleh bakteri sehingga hati tertutup oleh lemak yang disebut dengan *fatty liver* atau disebut juga dengan hati berlemak.

Sirosis hati terdiri dari empat stadium sehingga untuk dapat mendiagnosis seorang pasien mengalami stadium sirosis hati level satu maupun level lainnya dapat ditinjau dengan melihat rekam medis pasien dari pasien yang bersangkutan. Salah satu metode yang dapat digunakan untuk mendiagnosis seorang pasien tersebut adalah metode CART.

Metode CART pertama kali digagas oleh Leo Breiman, Jerome Friedman, Richard Olshen, dan Charles Stone pada tahun 1984. Metode ini menghasilkan *decision tree* yang memiliki ciri memcah simpul hanya menjadi dua cabang. CART bekerja dengan membagi data menjadi dua kelompok yang semakin homogen berdasarkan atribut tertentu. Pohon keputusan yang dihasilkan mirip dengan hierarki keputusan yang mengarah dari akar (node pertama) ke daun (node terakhir) dengan setiap node memiliki nilai atribut tertentu. Hasil dari setiap percabangan di pohon keputusan adalah klasifikasi

This is an open-access article under the CC-BY-SA License.

untuk data kategori atau prediksi untuk data numerik.

Pada penelitian ini akan diklasifikasikan stadium seorang pasien penderita sirosis hati pada masa mendatang berdasarkan rekam medis mereka. Rekam medis tersebut yang akan dijadikan dasar untuk menentukan stadium pasien sirosis hati disebut variabel prediktor. Adapun stadium sirosis hati yang akan ditentukan berdasarkan variabel prediktor disebut variabel respon.

Sehingga dari uraian latar belakang diatas akan dipaparkan hasil klasifikasi stadium pasien yang menderita sirosis hati dengan menggunakan metode CART.

METODE PENELITIAN

Penelitian ini menggunakan data yang bersumber dari kaggle dengan nama *liver cirrhosis prediction,* berikut akan disajikan gambaran umum dari data :

Tabel 2. 1 Gambaran Umum Data

PEUBAH	**JENIS**	**KETERANGAN**
Stage	Ordinal	1 : Stadium I 2 : Stadium II 3 : Stadium III 4 : Stadium IV
Drug	Nominal	*Placebo* *D-penicillamine*
Age	Rasio	-
Sex	Nominal	F : *Female* M : *Male*
Ascites	Nominal	Y : *Yes* (Ya) N : *No* (Tidak)
Hepatomegaly	Nominal	Y : *Yes* (Ya) N : *No* (Tidak)
Spiders	Nominal	Y : *Yes* (Ya) N : *No* (Tidak)
Edema	Nominal	Y : edema & tanpa terapi diuretik S : edema tanpa diuretik/ edema teratasi dengan diuretik N : edema meskipun dengan terapi diuretik
Bilirubin	Rasio	-
Cholesterol	Rasio	-
Albumin	Rasio	-
Copper	Rasio	-
Alk_Phos	Rasio	-
SGOT	Rasio	-
Triglycerides	Rasio	-
Platelets	Rasio	-
Prothrombin	Rasio	-

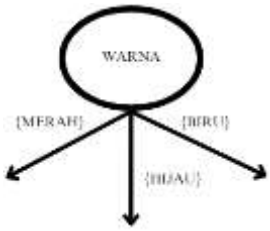
METODE PEMBELAJARAN

Metode pembelajaran adalah salah satu pendekatan dalam pembelajaran mesin (*machine learning*). Metode ini terbagi menjadi dua yaitu *supervised learning* atau metode pembelajaran terawasi yaitu model mempelajari dari data yang telah berlabel yang bertujuan untuk mengajarkan model menghubungkan *input* dan *output* yang diinginkan sehingga model dapat memprediksi secara akurat data baru meskipun belum pernah dilihat sebelumnya. Contoh dari supervised learning adalah prediksi seseorang akan bermain bola basket outdoor atau tidak berdasarkan *behavior* sebelumnya.

Metode pembelajaran kedua adalah *unsupervised learning*. Kebalikan dari metode sebelumnya, pada *unsupervised learning* model diharapkan dapat menemukan kelompok atau asosiasi yang alami dalam data mengidentifikasi pola atau struktur yang tersembunyi dalam data tanpa adanya informasi target yang jelas. Contoh dari metode *unsupervised learning* adalah seorang guru yang ingin mengelompokan siswa – siswa berdasarkan pada kriteria kesamaan IQ atau mengelompokan berdasarkan umur dan tinggi badan maupun ketiganya.

1.1 Decision Tree

Decision tree adalah sebuah struktur berhirarki yang menggambarkan model klasifikasi maupun regresi yang bertugas menguraikan data yang kemudian disajikan menjadi sebuah pohon keputusan. Berdasarkan algoritmanya *decision tree* dapat menghasilkan pohon berbentuk biner maupun non – biner.



Gambar 2. 1 Contoh DT Non – Biner

Decision tree non – biner nilai atribut akan terbagi menjadi lebih dari dua himpunan bagian tidak kosong yang berbeda. Misalnya nilai atribut warna = {merah, hijau, biru} maka, akan terdapat himpunan bagian yaitu : {(merah), (hijau), (biru)}.

Gambar 2. 2 Contoh DT Biner

Selanjutnya pada *decision tree* biner nilai – nilai atribut terbagi menjadi dua himpunan bagian tidak kosong yang berbeda dengan lebih dari satu kemungkinan bentuk percabangan apabila jumlah nilai atributnya lebih dari dua. Misalnya nilai atribut warna = {merah, hijau, biru} maka, akan terdapat tiga kemungkinan himpunan bagian yaitu : {{(merah), (hijau, biru)}, {(hijau), (merah, biru)}, {(biru), (merah, hijau)}}.

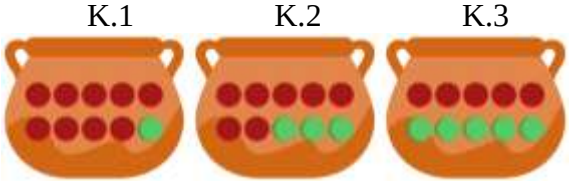
1.2 Entropi

Entropi adalah metode yang digunakan untuk mencari informasi ketidakteraturan atau ketidakpastian dari suatu data (Gray, 2011). Jika kumpulan sampel semakin heterogen, maka nilai entropinya akan semakin besar. Suatu kumpulan data yang terbagi menjadi c kelas memiliki nilai entropi yang berada pada interval 0 hingga $\log_2 c$ dengan c adalah banyak kelompok pada varibel terikat. Sehingga apabila himpunan data terkelompok dalam empat kelas nilai entropi maksimum adalah $\log_2 c = \log_2 4 = 2$. Sehingga apabila nilai entropi maksimum proporsi jumlah data antar kelas adalah sama namun, apabila nilai entropi minimum atau 0 maka, kelas tersebut mempunyai tidak memiliki keberagaman atau homogen.

Nilai entropi dapat dihitung dengan hasil penjumlahan setiap probabilitas amatan sebanyak $\log_2 p_i$ yang secara matematis dapat ditulis sebagai berikut:

$$Entropi(S) = -\sum_{i=1}^{n} p_i(\log_2 p_i)$$

Formulasi dari perhitungan nilai entropi dapat iilustrasikan seperti berikut ini :



Gambar 2. 3 Ilutrasi Perhitungan Entropi

Terdapat tiga buah kendi yang berisi 10 bola. Bola – bola tersebut berwarna merah dan hijau dengan proporsi bola merah dan hijau berbeda pada masing – masing kendi. Selanjutnya akan dihitung nilai entropi pada ketiga kendi dengan keterangan p_1 adalah probabilitas bola merah dan p_2 probabilitas bola hijau seperti pada Tabel 2.1 berikut ini:

Tabel 2. 2 Perhitungan Nilai Entropi

	p_1	p_2	$Entropi(S)$
K.1	0.9	0.1	0.47
K.2	0.7	0.3	0.89
K.3	0.5	0.5	1

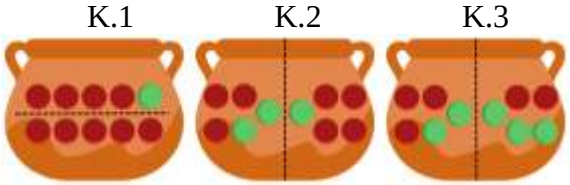
Terlihat pada kendi tiga entropi bernilai maksimum karena terdapat dua kelompok pada kendi sehingga interval pada kejadian tersebut adalah 0 hingga $\log_2 c = \log_2 2 = 1$. Nilai entropi maksimum berarti proporsi antara bola merah dan bola hijau pada kendi tiga adalah seimbang. Sedangkan pada kendi satu dan kendi dua proporsi bola merah dan hijau tidak seimbang sehingga nilai entropi hanya mendekati nilai maksimum.

1.3 Information Gain

Information gain ini adalah selisih antara entropi awal sebelum data tersebut dipartisi atau $Entropi\ (S)$ dengan rata – rata terboboti dari entropi masing – masing bagian $Entropi(S_1)$, $Entropi(S_2)$ hingga $Entropi(S_n)$ yang dituliskan secara matematis sebagai berikut :

$$IG = Entropi(S) - \sum_{i=1}^{n} \frac{|S_n|}{|S|} Entropi(S_n)$$

Formulasi dari perhitungan nilai information gain dapat dijelaskan sebagai berikut :



Gambar 2. 4 Ilutrasi Perhitungan IG

Pada subab 2.3 telah dijelaskan perhitungan nilai entropi sehingga pada subab selanjutnya akan diilustrasikan perhitungan dari information gain dengan keterangan $Entropi\ (S_1)$ adalah nilai entropi pada bagian atas kendi satu atau bagian kiri pada kendi dua dan kendi tiga, sedangkan $Entropi\ (S_2)$ adalah nilai entropi pada bagian bawah kendi satu atau bagian kanan kendi dua dan kendi tiga. Untuk mempermudah melihat perbedaanya nilai telah disajikan dalam bentuk Tabel 2.2 di bawah ini :

Tabel 2. 3 Perhitungan Nilai IG

	p_1	p_2	$E(S_1)$	p_1	p_2	$E(S_2)$
K.1	0.8	0.2	0.72	1	0	0
K.2	0.6	0.4	0.97	0.8	0.2	0.72
K.3	0.6	0.4	0.97	0.4	0.6	0.97

	$E(S)$	$\frlvert S_1 \frrvert$ / $\frlvert S \frrvert$	$E(S_1)$	$\frlvert S_2 \frrvert$ / $\frlvert S \frrvert$	$E(S_2)$	IG
K.1	0.47	½	0.72	½	0	0.11
K.2	0.89	½	0.97	½	0.72	0.03
K.3	1	½	0.97	½	0.97	0.03

Sehingga kesimpulannya adalah *information gain* nilainya akan semakin besar jika partisi yang dilakukan menghasilkan partisi baru yang bersifat lebih homogen atau kelas pada partisi tersebut cenderung didominasi pada salah satu kelas saja.

1.4 Algoritma Cart

Salah satu ciri dari metode ini adalah memecah simpul hanya menjadi dua cabang (biner). Adapun langkah – langkah dari algoritma CART sendiri adalah :

1. Untuk setiap variabel bebas akan ditentukan kemungkinan penyekatan yang terjadi berdasarkan tipe data. Bagi data yang bertipe numerik akan didapatkan $n - 1$, untuk data yang bertipe nominal akan didapatkan sebanyak $2^{L-1} - 1$ calon cabang dan bagi data yang bertipe ordinal akan didapatkan $L - 1$ calon cabang.
2. Menyusun calon cabang yang dilakukan terhadap semua variabel prediktor. Adapun daftar yang berisikan calon cabang tersebut dinamakan calon cabang mutakhir.
3. Selanjutnya yaitu menilai kinerja seluruh calon cabang yang berada pada daftar calon cabang mutakhir dengan cara menghitung nilai besaran kesesuaian.
4. Setelah mendapatkan nilai kesesuaian untuk semua calon cabang, maka dipilih nilai terbesar untuk dipecah terlebih dahulu.

HASIL DAN PEMBAHASAN

Setelah dilakukan pembersihan data yang meliputi penghapusan data yang hilang dan memperbaiki kesalahan penulisan didapatkan data baru yang berjumlah 276 data dengan rincian stadium 1 sebanyak 12 data, stadium 2 sebanyak 59 data, stadium 3 sebanyak 111 data dan stadium 4 sebanyak 94 data. Selanjutnya akan digunakan perbandingan sebesar 90 : 10 untuk perbandingan data latih dan data uji. Sehingga dari 248 data yang digunakan sebagai data uji terdapat satu variabel terikat dengan skala pengukuran ordinal yang terdiri dari empat kelas dan 16 variabel bebas dengan skala pengukuran nominal dan rasio.

Pada tahap pertama akan ditentukan terlebih dahulu untuk cabang kanan dan cabang kiri pada penelitian ini, yang mana variabel yang bertipe nominal pada penelitian ini akan memiliki masing – masing satu calon cabang hal ini karena variabel nominal pada penelitian ini hanya memiliki dua nilai $(2^{2-1}-1=1)$ kecuali, untuk variabel edema, karena variabel ini akan memiliki tiga calon cabang $(2^{3-1}-1=3)$. Selanjutnya akan didapatkan 18 simpul untuk calon cabang kanan dan kiri.

Kemudian dilakukan perhitungan nilai indeks gini yang disajikan dalam Tabel 3.1 di bawah ini:

Tabel 3. 1 Nilai Indeks Gini

Simpul	$\boldsymbol{P_L}$	$\boldsymbol{P_R}$	$\boldsymbol{i(t)}$
1	0.49355	0.50806	0.49828
2	0.50081	0.50081	0.49839
3	0.87823	0.12339	0.21350
4	0.93145	0.06855	0.12770
5	0.48387	0.51613	0.49948
6	0.70968	0.29032	0.41207
7	0.84677	0.15323	0.25950
8	0.09274	0.90726	0.16828
9	0.06048	0.93952	0.11365
10	0.48387	0.51613	0.49948
11	0.49718	0.50444	0.49836
12	0.50081	0.50081	0.49839
13	0.49355	0.50806	0.49828
14	0.50081	0.50081	0.49839
15	0.50081	0.50081	0.49839
16	0.48629	0.51532	0.49796
17	0.50081	0.50081	0.49839
18	0.46089	0.54073	0.49520

Setelah didapatkan nilai dari indeks gini pada setiap cabang kemudian akan dipilih cabang pertama untuk dipecah terlebih dahulu yang diurutkan berdasarkan pada nilai goodness of split yang dapat dilihat pada Tabel 3.2 berikut ini:

Tabel 3. 2 Rangking Goodness of Split

Simpul	$\Phi(s\|t)$	Rangking
1	-0.176	12
2	-0.1656	6
3	-0.461	15
4	-0.4979	16
5	-0.1139	1
6	-0.2357	13
7	-0.3948	14
8	-0.5025	17
9	-0.5963	18
10	-0.1525	2
11	-0.1697	9
12	-0.1535	3
13	-0.1615	5
14	-0.1755	11
15	-0.1725	10
16	-0.1693	8
17	-0.169	7
18	-0.1541	4

Terlihat bahwa nilai goodness of split yang terbesar berada pada simpul kelima yaitu sebesar -0.11 dengan variabel hepatomegaly pada cabang kiri N dan cabang kanan adalah Y.

Selanjutnya digunakan variabel hepatomegaly sebagai root node dan proses pembentuka pohon keputusan berulang kembali secara rekursif.

Berdasarkan pohon keputusan yang dihasilkan didapatkan delapan kelompok penduga yang disajikan dalam Tabel 3.3

Tabel 3. 3 Kelompok Penduga Yang Dihasilkan

		Observasi	
		N	%
1	Prothrombin ≥11	15	6%
2	Cholesterol <352	37	15%
3	Tryglicerides <122	15	6%
	Tryglicerides ≥122	7	3%
4	Albumin <3.5	7	3%
	Albumin ≥3.5	37	15%
5	Prothrombin <10	7	3%
	Prothrombin ≥10	17	7%
6	Age <18e+3	10	4%
	Age ≥18e+3	10	4%
7	Age <17e+3	15	6%
	Age ≥17e+3	12	5%
8	SGOT <64	10	4%
	SGOT ≥64	50	20%

KESIMPULAN DAN SARAN

Dari pembahasan pada penelitian ini dapat diketahui CART memiliki beberapa keunggulan jika dibandingkan dengan metode klasifikasi lainnya, yaitu hasilnya lebih mudah diinterpretasikan karena tersaji dalam bentuk visual dan lebih cepat penghitungannya, selain itu CART dapat diterapkan untuk himpunan data yang berjumlah besar dengan variabel yang banyak namun terlepas dari itu semua,

Metode ini juga memiliki kelemahan seperti tidak stabil dalam decision tree hal ini karena CART sangat sensitif dengan data baru dan CART sangat bergantung dengan jumlah sampel sehingga apabila sampel data learning dan testing berubah maka pohon keputusan yang dihasilkan juga ikut berubah.

Saran yang ingin disampaikan oleh peneliti untuk hasil dari penelitian ini yaitu bagi peneliti selanjutnya diharapkan untuk meningkatkan tingkat akurasinya baik dengan cara menerapkan metode untuk menangani unbalanced data maupun mencari data yang seimbang apabila ingin menggunakan metode CART.

DAFTAR PUSTAKA

Gray, R. M. (2011). Entropy and Information Theory (2nd ed.). Springer Science Business Media.

Hermawati, Fajar Agus. (2013). Data Mining. Yogyakarta : Penerbit ANDI.

Hermawati, Fajar Agus. 2013. Data Mining. Yogyakarta : Penerbit ANDI.

Kurniasih, T. (2018). Sistem Organ Manusia. 1(1). Yogyakarta: Penerbit Deepublish.

Larose, D. T., Larose, C. D. (2014). Discovering Knowledge in Data: An Introduction to Data Mining (2nd ed.). John Wiley & Sons.

Santosa, Budi dan Ardian Umam. 2018. Data Mining dan Big Data Analytics. Yogyakarta : Penebar Media Pustaka.

Sulistyoningrum, E., & Murtisiwi, L. (2020). Gambaran Peresepan Pasien Sirosis Hati di Instalasi Rawat Jalan Rumah Sakit Panti Waluyo Surakarta. Jurnal Farmasi Stikes Nasional, 9(1), 1–7.

Suniantara, I.K.P., Sukarsa, I.K.G., Srinadi, I.G.A.M. (2008). Penerapan Metode Regresi Berstruktur Pohon untuk Memprediksi Berat Badan Bayi Lahir Studi Kasus: RSUD Wangaya.

Suntoro, Joko. 2019. Data Mining : Algoritma dan Implementasi dengan Pemrograman PHP. Jakarta : PT Elex Media Komputindo.

Susanto, Sani dan Dedy Suryadi. 2010. Pengantar Data Mining : Menggali Pengetahuan dari Bongkahan Data. Yogyakarta : Penerbit ANDI.

Suyanto. 2017. Data Mining untuk Klasifikasi dan Klasterisasi Data. Bandung : Penerbit Informatika.

Suyanto. 2018. Machine Learning Tingkat Dasar dan Lanjut. Bandung : Penerbit Informatika.

Suyanto. 2021. Artificial Intelligence : Searching, Reasoning, Planning dan Learning. Bandung : Penerbit Informatika.

Universitas Udayana, P. S. M. (2019). Pedoman Pelaksanaan Seminar dan Tugas Akhir.