JURNAL DISPROTEK
ISSN: 2088-6500 (p); 2548-4168 (e)
Vol 13, No. 2, Juli 2022, hlm. 131-136
DOI: 10.34001/jdpt.v12i2

KLASIFIKASI HEPATITIS C VIRUS MENGGUNAKAN ALGORITMA C4.5

CLASSIFICATION OF HEPATITIS C VIRUS USING ALGORITHM C4.5

Susanto[1*], Nuri[2]

[1,2]Sekolah Tinggi Teknik Pati
Email : [1*]susanto0033@gmail.com, [2]nuri.indramayu@gmail.com
*Penulis Korespondensi

Abstrak - Hepatitis C Virus (HCV) merupakan Penyakit gangguan pada hati manusia yang menyebabkan peradangan pada sel-sel dan kinerja organ hati manusia. Penyakit HCV sangat berbahaya bagi tubuh manusia karena penyakit ini merupakan penyakit yang berasal dari penyakit kanker hati. Penyebab penyakit HCV adalah penyalahgunaan alkohol, penggunaan alat suntik bekas, dan tranfusi darah oleh orang yang telah terkena penyakit HCV. Adanya permasalahan tersebut, maka tujuan peneliti adalah melakukan penerapan sebuah Algoritma yang digunakan untuk mengklasifikasikan penyakit HCV. Algoritma yang pilih peneliti adalah Algoritma C4.5 yang merupakan salah satu jenis Algoritma klasifikasi. Penerapan Algoritma C4.5 digunakan untuk memperoleh sebuah keputusan, jika seseorang terkena HCV ataupun tidak. Algoritma C4.5 menggunakan atribut-atribut data yang ada untuk menghasilkan suatu keputusan terhadap penyakit HCV. Algoritma C4.5 diimplementasikan dengan metode Adaboost untuk memperoleh hasil yang maksimal. Metode Adaboost merupakan metode penunjang yang digunakan untuk meningkatkan hasil akurasi keputusan dari Algoritma C4.5, sehingga menghasilkan nilai akurasi yang tinggi. Nilai akurasi yang dihasilkan sebesar 95,60% (semula 94,43% menjadi 95,60%).

Kata kunci: Algoritma C.4.5, Adaboost, Hepatitis C Virus

Abstract *- Hepatitis C Virus (HCV) is a disease of the human liver that causes inflammation of the cells and the performance of the human liver. HCV disease is very dangerous for the human body because this disease is a disease that comes from liver cancer. The causes of HCV disease are alcohol abuse, use of used syringes, and blood transfusions by people who have been exposed to HCV disease. Given these problems, the aim of the researcher is to implement an algorithm that is used to classify HCV disease. The algorithm that the researcher chose is the C4.5 algorithm, which is one type of classification algorithm. The application of the C4.5 Algorithm is used to obtain a decision, if a person is exposed to HCV or not. The C4.5 algorithm uses the existing data attributes to produce a decision on HCV disease. The C4.5 algorithm is implemented using the Adaboost method to obtain maximum results. The Adaboost method is a supporting method used to improve decision accuracy results from the C4.5 Algorithm, resulting in a high accuracy value. The resulting accuracy value is 95.60% (originally 94.43% to 95.60%)*

Keywords*: C4.5 Algorithm, Adaboost , Hepatitis C Virus*

1. PENDAHULUAN

Pada tahun 2019, terdapat 325 juta orang di dunia hidup dengan mengidap virus hepatitis B dan C. Hepatitis C adalah penyakit hati yang disebabkan oleh virus hepatitis C (HCV). Virus hepatitis C adalah virus yang ditularkan melalui darah, misalnya dengan digunakannya peralatan kesehatan yang tidak aman, narkoba suntikan, tranfusi darah, serta praktik seksual yang mengarah pada paparan darah. Sebagian besar dari penderita yang terinfeksi hepatitis C kronis akan berkembang hingga menjadi sirosis atau kanker hati [1].

Hepatitis adalah penyakit pada hati yang mengalami peradangan, yang diakibatkan oleh virus, jamur,bakteri, dan parasit dengan cara menginfeksi hati atau bisa saja tidak dengan cara infeksi, yaitu seperti mengonsumsi alkohol,obat-obatan, autoimun dan metabolik. Tingkat peradangan pada hati dapat berupa akut maupun kronik. Peradangan pada tingkatan kronik dapat menyebabkan kerusakan hati yang bervariasi dalam kurun waktu minimal 6 bulan[2].

Jenis virus RNA yaitu Virus hepatitis C merupakan virus yang bertugas membawa isyarat untuk membentuk DNA/ptotein baru. Apabila virus menginfeksi RNA, maka akan mengubah bentuk dari DNA. Sel memiliki sifat alami dalam mempertahankan diri agar terus hidup atau mati setelah mengubah pola kode DNA yang baru melalui perubahan pola RNA. Pola kode DNA akan berbeda dengan aslinya jika HCV berhasil menginfeksi RNA[3].

Berdasarkan penelitian sebelumnya yang telah dilakukan menggunakan algoritma Neural Network (NN) menghasilkan tingkat akurasi yang tinggi yaitu sebesar 94,12% dalam memprediksi penyakit HCV dibandingkan dengan hasil akurasi algoritma C4.5 yang hanya mendapatkan nilai sebesar 75,3%.

Dari permasalahan tersebut, peneliti akan melakukan implementasi Metode baru untuk menunjang hasil akurasi dari Algoritma C4.5, sehingga memperoleh hasil yang maskimal dalam pengklasifikasian penyakit HCV. Metode Adaboost merupakan metode yang akan digunakan peneliti dalam melakukan penelitian ini.

1.1. Tujuan penelitian tersebut adalah untuk membantu para dokter dalam pengklasifikasian penyakit HCV, sehingga berkurangnya kasus kematian akibat penyakit HCV yang terjadi di beberapa negara, terutama di Indnesia.

Hepatitis C

Hepatitis C merupakan penyakit yang dapat menyerang organ hati. Tahap hepatitis C diantaranya adalah inflamasi, fibrosis, sirosis, dan gagal hati atau kanker hati. Seseorang yang mengalami hepatitis C pada usia tua cenderung lebih susah untuk disembuhkan. Wanita secara biologis lebih mudah sembuh dengan sendirinya karena memiliki sistem imun lebih baik. Gejala yang mungkin dirasakan oleh seseorang yang mengidap penyakit hepatitis C adalah rasa lelah, demam, diare, mual, dan bola mata atau kulitnya akan berwarna kekuningan [4]. Deteksi penyakit hepatitis C bisa dilakukan dengan mengambil sampel darah yaitu untuk melihat jumlah sel darah merah, sel darah putih, hemoglobin, dan trombosit. Selain itu perlu juga untuk melihat jumlah enzim *aspartate aminotransferase* (AST) dan *alanaine aminotransferase* (ALT) dalam darah. Meningkatnya enzim AST dan ALT bisa mengindikasikan adanya gangguan organ hati[5].

1.2. Algoritma *C4.5*

Algoritma C4.5 digunakan untuk membangun sebuah pohon keputusan yang mudah dimengerti, fleksibel, dan menarik karena dapat divisualisasikan dalam bentuk gambar [6] Pohon keputusan adalah salah satu metode klasifikasi yang paling populer karena mudah untuk diinterpretasi oleh manusia. Pohon keputusan adalah model prediksi menggunakan struktur pohon atau struktur berhirarki. Konsep dari pohon keputusan adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan.

Ada beberapa tahap dalam membuat sebuah pohon keputusan dengan algoritma *C4.5* [6] yaitu:

a. Mempersiapkan data *training*, dapat diambil dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokan dalam kelas-kelas tertentu.
b. Menentukan akar dari pohon dengan menghitung nilai *gain* yang tertinggi dari masing-masing atribut atau berdasarkan nilai *index entropy* terendah. Sebelumnya dihitung terlebih dahulu nilai *index entropy*, dengan rumus:

$$Entropy(i) = \sum_{j=1}^{m} \mathrm{f}(i,j).\,2\ \mathrm{f}[(i,j)]$$

c. nilai *gain* dengan rumus:

$$\mathrm{gain} = -\sum_{i=1}^{p} \frac{ni}{n}.IE(i)$$

d. Untuk menghitung gain ratio perlu diketahui suatu term baru yang disebut Split Information dengan rumus:

$$SplitInformation = -\sum_{t=1}^{c} \frac{S_1}{S} log_2 \frac{S_1}{S}$$

e. Selanjutnya menghitung gain ratio

$$Gainratio(S,A) = \frac{Gain(S,A)}{SplitInformation(S,A)}$$

f. Ulangi langkah ke-2 hingga semua *record* terpartisi Proses partisi pohon keputusan akan berhenti disaat:
   1) Semua tupel dalam *record* dalam simpul m mendapat kelas yang sama
   2) Tidak ada atribut dalam *record* yang dipartisi lagi
   3) Tidak ada *record* didalam cabang yang kosong.

1.3. Metode C4.5 brbasis *Adaboost*

Setelah melakukan tahapan dalam membuat sebuah pohon keputusan dengan algoritma *C4.5*, dilakukan pemberian bobot pada pohon tunggal sehingga menghasilkan hipotesa baru dan sebuah pohon keputusan baru [7].

2. METODE PENELITIAN

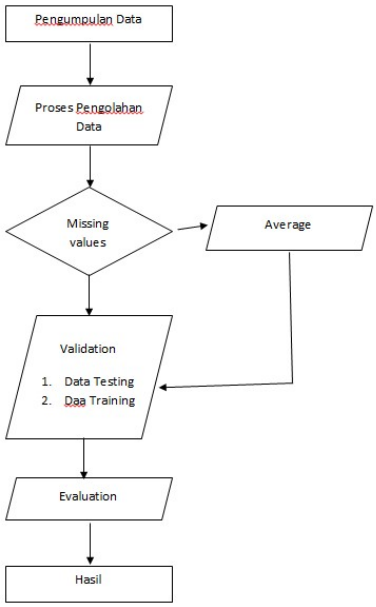
Data yang digunakan berasal dari UCI dataset. Terdiri dari 11 atribut dan 1 label dengan record sebanyak 615.

Category	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
0	0	38.500	52.500	7.700	22.100	7.500	6.930	3.230	106	12.100	69
0	0	38.500	70.300	18	24.700	3.900	11.170	4.800	74	15.600	76.500
0	0	46.900	74.700	36.200	52.600	6.100	8.840	5.200	86	33.200	79.300
0	0	43.200	52	30.600	22.600	18.900	7.330	4.740	80	33.800	75.700
0	0	39.200	74.100	32.600	24.800	9.600	9.150	4.320	76	29.900	68.700
0	0	41.600	43.300	18.500	19.700	12.300	9.920	6.050	111	91	74
0	0	46.300	41.300	17.500	17.800	8.500	7.010	4.790	70	16.900	74.500
0	0	42.200	41.900	35.800	31.100	16.100	5.820	4.600	109	21.500	67.100
0	0	50.900	65.500	23.200	21.200	6.900	8.690	4.100	83	13.700	71.300
0	0	42.400	86.300	20.300	20	35.200	5.460	4.450	81	15.900	69.900
0	0	44.300	52.300	21.700	22.400	17.200	4.150	3.570	78	24.100	75.400
0	0	46.400	68.200	10.300	20	5.700	7.360	4.300	79	18.700	68.600
0	0	36.300	78.600	23.600	22	7	8.560	5.380	78	19.400	68.700
0	0	39	51.700	15.900	24	6.800	6.460	3.380	65	7	70.400
0	0	38.700	39.800	22.500	23	4.100	4.630	4.970	63	15.200	71.900

ExampleSet (615 examples, 1 special attribute, 11 regular attributes)

Gambar 1.1 Sample data yang digunakan

Pada Gambar 1.1 menunjukkan data sample yang masih memiliki missing value. Missing value berpengaruh langsung terhadap hasil akurasi yang didapatkan menggunakan Machine Learning. Peningkatan akurasi akan dilakukan dengan mengisi missing value dengan menggunakan metode tambahan. Pada penelitian ini akan menggunakan metode yang berbeda dengan metode yang telah digunakan pada penelitian yang terkait. Algoritma C4.5 dipilih untuk mengklasifikasi Hepatitis C Virus dengan melakukan model range pada atribut category.



Gambar 1.2. Alur Penelitian

Seperti yang ditujukkan pada Gambar 1.2 Data awal diolah dengan melakukan validation. Validation dilakukan untuk menghilangkan data yang missing dan menghapus data yang tidak diperlukan. Missing value dalam dataset diubah dengan menggunakan model average. Model Average diterapkan di setiap atribut data yang memiliki missing values, sehingga menghasilkan data baru tanpa missing values dengan jumlah record tetap.

Data baru tanpa missing values akan diolah menggunakan Algoritma C4.5 untuk mengklasifikasi Hepatitis C Virus. Algortima C4.5 akan menghasilkan suatu pohon keputusan (Decission Tree) yang akan digunakan untuk menentukan keputusan terhadap sebuah kondisi. Setelah penerapan Algortima C4.5 akan dilakukan proses evaluasi. Evaluasi bertujuan untuk mengerahui tingkat akurasi, precision, dan recall dari data yang telah diproses menggunakan algoritma C4.5. Hasil Akurasi yang akan digunakan sebagai hasil akhir dari penelitian.

3. HASIL DAN PEMBAHASAN

Algoritma C4.5 merupakan salah satu algoritma klasifikasi data mining. Menurut Gorunescu dalam (Sunge, 2018) Algortima dalam kasifikasi yang banyak digunakan ialah Decision Tree. Dikarenakan sangat mudah dimengerti dan dijabarkan oleh banyak pengguna juga mudah dipahami dimana cabang pohon disimpulkan dalam bentuk klasifikasi. Algoritma C4.5 menggunakan konsep information gain atau entropy reduction untuk memilih pembagian yang optimal (Larose, 2005). Tahapan dalam membuat pohon keputusan dengan algoritma C4.5 (Gorunescu, 2011) yaitu:

1. Mempersiapkan data training, dapat diambil dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokkan dalam kelas-kelas tertentu..
2. Menentukan akar dari pohon dengan menghitung nilai gain yang tertinggi dari masing-masng atribut atau berdasarkan nilai index entropy terendah. Sebelumnya dihitung terlebih dahulu nilai index entropy, dengan rumus:

$$Entropy\ (i) = -\sum_{j=1}^{m} f(i,j).log\,2\,f[(i,j)]$$

3. Hitung nilai gain dengan rumus:

$$Entropy\ split\ = \sum_{i=1}^{p} \binom{n1}{n}.\ IE(i)$$

4. Ulangi langkah ke-2 hingga semua record terpartisi. Proses partisi pohon keputusan akan berhenti disaat:
   a. Semua tupel dalam record dalam simpul N mendapat kelas yang sama.
   b. Tidak ada atribut dalam record yang dipartisi lagi.
   c. Tidak ada record di dalam cabang yang kosong

Peneliti menggunakan metode Adaboost untuk meningkatkan hasil kurasi dari penerapan algoritma C4.5. Metode Adaboost berfungsi untuk mengisi cabang kosong dari pohon keputusan Algoritma C4.5 sehingga hasil akurasi akan meningkat. Dengan penerapan metode Adasoost diperoleh hasil akurasi yang lebih baik dibandingkann dengan hanya menggunakan Algoritma C4.5.

Berikut adalah hasil yang diperoleh sebelum dan sesudah penerapan Metode Adaboost dalam Algoritma C4.5 :

accuracy: 94.43% +/- 1.08% (micro average: 94.43%)

	true 0	true 1	class precision
pred. 0	1920	86	95.71%
pred. 1	38	182	82.73%
class recall	98.06%	67.91%	

Gambar 1.3 Sebelum Penerapan Metode Adaboost

accuracy: 95.61% +/- 2.16% (micro average: 95.60%)

	true 0	true 1	class precision
pred. 0	1929	69	96.55%
pred. 1	29	199	87.28%
class recall	98.52%	74.25%	

Gambar 1.4 Setelah Penerapan Metode Adaboost

Pada Gambar 1.3 menunjukkan nilai dari pengolahan data menggunakan Machine Learning yang belum ditambahkan Metode Adaboost. Sedangkan, Pada Gambar 1.4 menunjukkan hasil dari penerapan Metode Adaboost pada Algoritma C4.5 dan memiliki hasil yang lebih tinggi dibandingkan sebelum penerapan metode Adaboost. Hasil yang diperoleh memiliki peningkatan sebesar 1,22 % setelah penerapan metode Adaboost.

4. KESIMPULAN

Penelitian ini dilakukan untuk mengklasifikasi penyakit Hepattis C Virus dengan menggunakan Algoritma C4.5. Penerapan Algoritma C4.5 memiliki kelebihan dalam melakukan komputasi data dengan waktu yang singkat. Untuk meningkatkan hasil dari Algoritms C4.5 dilakukan penerapan meted lain yaitu Metode Adaboost, sehingga hasil nilai akurasi akan meningkat, Nilai akurasi yang dihasilkan dari Algritma C4.5 dengan Adaboost sebesar 95,60%. Peneliti berharap hasil yang telah diperoleh dapat ditingkatkan menjadi lebih baik dengan memanfaatkan penggabungan metode-metode yang lain.

UCAPAN TERIMA KASIH

Penelitian ini melibatkan beberapa pihak yang ada di sekitar peneliti. Peneliti mengucapkan terima kasih kepada pihak yang telah terlibat dalam proses pembuatan artikel ini, terutama pada pihak yang telah menyediakan perangkat komputer sebagai media pendukung dalam proses penelitian.

DAFTAR PUSTAKA

[1] World Health Organization (WHO), 2019. Hepatitis C. Retrieved from https://www.who.int/news=room/fact-sheets/detail/hepatitis-c

[2] Raharja, K. Y. (2021). *PERBANDINGAN KINERJA ALGORITMA GAUSSIAN NAIVE BAYES DAN K-NEAREST NEIGHBOR (KNN) UNTUK MENGKLASIFIKASI PENYAKIT HEPATITIS C VIRUS (HCV)* (Doctoral dissertation, Universitas Muhammadiyah Jember).

[3] Al Kindhi, B., Sardjono, T. A., & Purnomo, M. H. (2018). Optimasi Support Vector Machine untuk Memprediksi Adanya Mutasi pada DNA Hepatitis C Virus. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI)*, *7*(3), 317-323.

[4] John, T. M. S., 2008. Signs and Symptomps that May be Associated with Hepatitis C. *Hepatitis C Choices.* Caring Ambassadors Program, Inc., pp

[5] Sandt, L., 2008. Understanding Hepatitis C disease. *Hepatitis C Choices.* Caring Ambassadors Program, Inc., pp. 23-42.

[6] Florin Gorunescu, *Data mining concepts models and technique*. berlin: Springer, 2011

[7] Wu, Xingdong, *The Top Ten Algorithm in Data mining*. Minnesota: Taylor & Francis Group, 2009.

[8] Saputri, N. D. (2021). *Komparasi penerapan metode Bagging dan Adaboost pada Algoritma c4. 5 untuk prediksi Penyakit Stroke* (Doctoral dissertation, UIN Sunan Ampel Surabaya).

[9] Rabbani, M. N., Yusuf, A., & Rolliawati, D. (2021). Komparasi Model Prediksi Daftar Ulang Calon Mahasiswa Baru Menggunakan Metode Decision Tree Dan Adaboost. *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, *10*(1), 18-24.

[10] Destyar, M. R. (2021). *IMPLEMENTASI METODE MODIFIED K-NEAREST NEIGHBOR UNTUK PREDIKSI HASIL TREATMENT PENYAKIT HEPATITIS C* (Doctoral dissertation, Universitas Muhammadiyah Jember).

[11] Jaumi, J. (2020). *Kajian sistematis pengaruh pemberian sofosbuvir/velpatasvir terhadap viral load penderita hepatitis C kronik* (Doctoral dissertation, Universitas Hasanuddin).

[12] Destyar, M. R. (2021). *IMPLEMENTASI METODE MODIFIED K-NEAREST NEIGHBOR UNTUK PREDIKSI HASIL TREATMENT PENYAKIT HEPATITIS C* (Doctoral dissertation, Universitas Muhammadiyah Jember).