

KOMBINASI ALGORITMA C-RIPPER UNTUK MENDIAGNOSIS PENYAKIT *LIVER*

Dian Restiani

Teknik Informatika, STMIK AMIKOM Purwokerto
Jl. Let. Jend. Pol. Sumarto Watumas Purwokerto
dianrestianii@gmail.com

ABSTRAK

Hati merupakan organ yang paling besar dan penting bagi tubuh kita. Kita tidak bisa hidup tanpa hati. Penyakit hati tidak mudah ditemukan pada stadium awal. Penanganan pasien penyakit hati pada stadium awal akan memperpanjang usia pasien. Untuk mendeteksi penyakit hati, pasien harus melakukan tes darah. Kesalahan dalam mendiagnosis penyakit dan penentuan konsumsi obat dapat menyebabkan hal yang membahayakan bagi kesehatan pasien bahkan dapat menyebabkan kematian. Ini menjadikan para peneliti dan praktisi memusatkan perhatiannya untuk mendeteksi/mendiagnosis penyakit *liver* dan mencegahnya karena penyakit ini bisa menimbulkan kematian jika sudah akut. Metode yang digunakan dalam penelitian ini yaitu CRISP-DM. Algoritma yang digunakan dalam penelitian ini adalah kombinasi CART dan RIPPER dengan menggunakan *dataset* diambil dari *repository database* UCI *Indian Liver Patient Dataset* yang terdiri dari data klinis pasien yang terdeteksi positif dan negatif penyakit *liver*. Hasil penelitian ini dengan menggunakan kombinasi *rule* hasil algoritma CART dan algoritma RIPPER dari *confusion matrix* didapatkan hasil akurasi sebesar 81%, sensitifitas/*recall* 0,63%, spesifisitas 0,91%, *precision* 0,80% dan *f-measure* sebesar 0,70% pada *dataset* ILPD tanpa melakukan seleksi fitur, sedangkan pada hasil akurasi *dataset* ILPD dengan melakukan seleksi fitur yaitu akurasi 70% dengan sensitifitas/*recall* 0,84%, spesifisitas 0,59%, *precision* 0,58% dan *f-measure* 0,70%. Dari hasil tersebut dapat disimpulkan bahwa untuk mendiagnosis penyakit *liver* dengan menggunakan kombinasi algoritma CART dan RIPPER tanpa melakukan seleksi fitur.

Kata kunci: *Diagnosis, Penyakit Liver, Algoritma, CRISP-DM, CART, RIPPER, Seleksi Fitur*

ABSTRACT

The liver is the largest organ and important for our body. We can not live without the heart. Liver disease is not easy to find in the early stages. For detection of liver disease patients in the early stages will prolong the patient's age. To detect liver disease, the patient must perform a blood test. Mistakes in diagnosing illness and determining drug consumption can cause harm to the health of the patient and even cause death. This allows researchers and practitioners to focus on detecting/diagnosing liver disease and preventing it because it can lead to death if it is acute. The method used in this research is CRISP-DM. The algorithm used in this study is a combination of CART and RIPPER using dataset taken from the UCI Indian Liver patient data repository Dataset consisting of clinical data of patients who detected positive and negative liver disease. The result of this research by using combination of rule result from CART algorithm and RIPPER algorithm of confusion matrix obtained accuracy 81%, sensitivity/*recall* 0.63%, specificity 0.91%, precision 0.80% and f-measure 0.70% on the ILPD dataset without performing feature selection, whereas on the accuracy results of the ILPD dataset by performing feature selection is accuracy 70%, with recall/sensitivity 0.84%, specificity 0.59%, precision 0.58% and f-measure 0.70 %. From these results it can be concluded that to diagnose liver disease by using a combination of CART and RIPPER algorithms without performing feature selection.

Keywords: *Diagnosis, Liver Disease, Algorithm, CRISP-DM, CART, RIPPER, Feature Selection*

DOI : 10.15408/jti.v11i1.6660

I. PENDAHULUAN

Hati merupakan organ yang paling besar penting bagi kita. Hati berfungsi sebagai pembentukan dan sekresi empedu, selain itu sebagai penyimpanan glikogen, sintesa urea, metabolisme kolestrol dan lemak serta detoksifikasi racun. Manusia tidak bisa hidup apabila hati mengalami kegagalan [1].

Salah satu penyakit yang meyerang hati adalah hepatitis atau *liver*. Penyakit *liver* merupakan peradangan hati yang disebabkan oleh infeksi virus, bakteri atau bahan-bahan beracun sehingga hati tidak dapat melakukan fungsinya dengan baik. Dalam bidang kedokteran, kesalahan dalam mendiagnosis penyakit yang dialami oleh seorang pasien adalah tanggung jawab yang paling berat para ahli kesehatan [2].

Angka pasti prevalensi dan kejadian penyakit *liver* di Indonesia pada tahun 2007 belum diketahui tingkat kenaikannya, tetapi berdasarkan data WHO (*World Health Organization*) pada tahun 2007 menunjukkan bahwa Indonesia termasuk dalam peringkat endemik yang tinggi [3].

Dalam mendiagnosis penyakit *liver* dokter akan mengajukan serangkaian pertanyaan dan mengevaluasi tanda-tanda dari gejala umum. Selanjutnya, dokter akan memeriksa tanda-tanda fisik untuk penyakit *liver*. Ada tiga jenis utama dari tes yang dilakukan dalam diagnosis penyakit *liver* yaitu tes darah, tes pencitraan dan analisis jaringan [4].

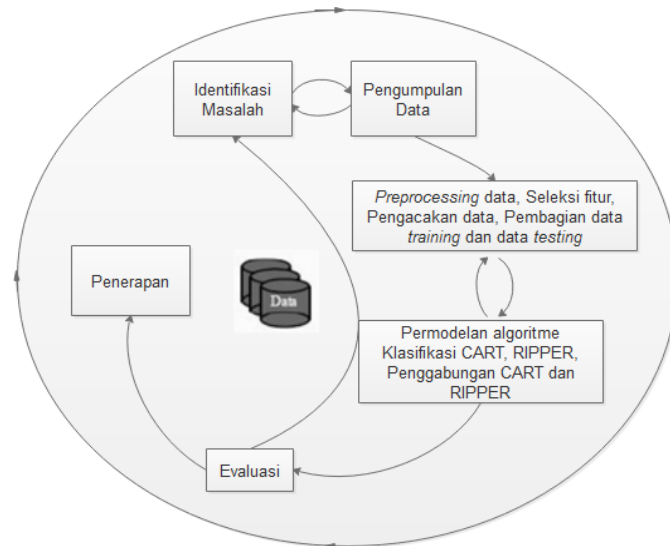
Dalam proses klasifikasi terdapat banyak algoritma yang dapat digunakan. Pada penelitian sebelumnya, Rahmawati pada tahun 2016 melakukan penelitian dengan klasifikasi untuk mendeteksi penyakit *liver*, ukuran perbandingan berdasarkan *Cross Validation* dan *Percentage Split* dengan tingkat akurasi

algoritma C4.5 69,828% dan algoritma *Naive Bayes* 63,362%. Pada tahun 2014, Widodo melakukan penelitian dengan tujuan untuk menghasilkan rule atau aturan yang sederhana tetapi memiliki akurasi yang cukup tinggi untuk mengklasifikasi pasien penyakit *liver*. Hasil Penelitian ini menunjukkan bahwa dari keempat algoritma yang digunakan, algoritma OneR menghasilkan performa nilai *accuracy* dan *precision* yang tertinggi, sedangkan untuk algoritma C4.5 menghasilkan nilai *sensitivity* dan nilai *spesificity* paling besar di antara algoritma lainnya. Algoritma RIPPER (*Repeated Incremental Prunning to Error Reduction*), algoritma ini akurat dalam menghasilkan *rule* terpendek (Astuti dan Yuliyanti, 2015). Selain itu salah satu algoritma dalam *data mining* yaitu CART (*Classification And Regression Trees*). Algoritma ini cocok diterapkan untuk data dengan variabel yang banyak dan kompleks. CART merupakan salah satu metode atau algoritma dari salah satu teknik eksplorasi data yaitu teknik pohon keputusan [5].

Pada Penelitian ini mencoba menkombinasikan algoritma CART dan RIIPER dalam mendeteksi penyakit *liver* pada *dataset* ILPD (*Indian Liver Patient Dataset*) dengan judul “Kombinasi Algoritma CART dan RIPPER untuk Mendiagnosis Penyakit *Liver* Berbasis *Correlation Based Feature Selection*”.

II. METODE PENELITIAN

Dalam penelitian ini dilakukan beberapa kegiatan *preprocessing* data, seleksi fitur, pengacakan data, penggunaan metode klasifikasi, evaluasi, penerapan yaitu identifikasi masalah, pengumpulan data, dan penulisan laporan tersusun dalam metode CRISP-DM.



Gambar 1. Kerangka penelitian

Dari gambar kerangka penelitian di atas dapat dijelaskan setiap tahapannya sebagai berikut:

1. Identifikasi Masalah
Upaya mengetahui permasalahan serta menggunakan metode yang baru untuk menganalisis performa algoritma pada penyakit *liver*.
2. Pengumpulan Data
Dalam penelitian ini data sekunder yang digunakan merupakan *dataset Indian Liver Patient Dataset (ILPD)*, diambil dari database *UCI Machine Learning Repository* yang terdiri dari 583 *record* dengan 414 *record* memiliki status sakit *liver* dan 169 *record* memiliki status sehat/tidak sakit *liver*.
3. *Preprocessing* data dan Seleksi fitur, Pengacakan data, Pembagian data *training* dan data *testing*
Untuk tahap *preprocessing* data yang pertama adalah persiapan data yaitu mengkonversi data ILPD dari bentuk CSV menjadi file dengan ekstensi *arff*, selanjutnya adalah penanganan *missing value* pada 4 data dengan cara mensubstitusi (mencari nilai rata-rata) nilai dari atribut. Selanjutnya adalah proses *smote* untuk meningkatkan atau membuat data menjadi lebih banyak dengan *tools weka*, dilanjutkan dengan seleksi fitur yaitu dengan algoritma CFS (*Correlation based Feature Selection*). Pengacakan data untuk mendapatkan hasil performa *dataset* ILPD yang paling optimal. Untuk pembagian data *training* dan data *testing* diambil persentase 80% dan 20% [5].

4. Permodelan Algoritma Klasifikasi CART, RIPPER, Penggabungan CART dan RIPPER

Untuk permodelan dilakukan *test option* dengan *supplied test set*, di tab ini lah data *testing* dimasukan sebagai data yang akan kita lihat performanya dari pola-pola yang telah dihasilkan pada data *training* baik dari algoritma CART ataupun RIPPER dan pada data yang dilakukan proses seleksi fitur ataupun tanpa seleksi fitur.

5. Evaluasi
Nilai confusion matrix yang dihasilkan menjadi parameter untuk mengevaluasi kinerja sistem yang dirancang dalam proses klasifikasi, meliputi nilai *accuracy*, *sensitivity*, *spesificity/recall*, *precision* dan *f-measure*.
6. Penerapan
Disimpulkan hasil performa penggabungan algoritma CART dan RIPPER dalam mendiagnosis penyakit *liver*.

III. HASIL DAN PEMBAHASAN

1. Identifikasi masalah
Melakukan beberapa studi pendahuluan dengan mempelajari literatur-literatur yang berkaitan dengan penelitian penulis yaitu untuk mendiagnosis penyakit *liver* serta pemilihan algoritma yang sesuai.
2. Pengumpulan data
Dataset pada penelitian ini diambil dari *repository database UCI Machine Learning Repository*, ILPD (*Indian Liver Patient Dataset*) terdiri dari 583 data klinis dan 11 atribut dengan target output positif *liver*

sebanyak 414 dan *negative liver* sebanyak 169. Atribut yang ada dalam *dataset* ILPD ditampilkan pada tabel berikut:

Tabel 1. Atribut pada dataset ILPD

No	Atribut	Tipe data
1	Age	Numeric
2	Gender	Kategorial
3	TB (Total Bilirubin)	Numeric
4	DB (Direct Bilirubin)	Numeric
5	Alkphos (Alkaline Phosphotase)	Numeric
6	Sgpt_AA (Alamine Aminotransferase)	Numeric
7	Sgot_AA (Aspartate Aminotransferase)	Numeric
8	TP (Total Proteins)	Numeric
9	ALB (Albumin)	Numeric
10	A/G (Ratio Albumin and Globulin Ratio)	Numeric
11	Class Variabel (positive liver (1) and negative liver (2))	Numeric

(Sumber: UCI Machine Learning Repository)

3. *Preprocessing* data dan Seleksi fitur, Pengacakan data, Pembagian data *training* dan data *testing*

```
@relation livertdataset

@attribute Age numeric
@attribute Gender {Male,Female}
@attribute TB numeric
Dst...

@data
1 65,Female,0.7,0.1,187,16,18,6.8,3.3,0.9,
,1 62,Male,10.9,5.5,699,64,100,7.5,3.2,0.74
Dst...
```

Gambar 2. Konversi data ILPD ke format arff

Selanjutnya penanganan *missing value*, Hasil perhitungan untuk nilai substitusi pada *record* yang *missing* sebagai berikut:

Persamaan 4.1.

$$Mean = \frac{\sum nilai_atribut \times 10}{\sum data}$$

$$= \frac{0,9+0,74+0,89+1+0,4+1,3+.....+0,37+1,1+1+1+1,5}{579}$$

$$= 1$$

Sehingga nilai yang dihasilkan digunakan untuk mengisi nilai atribut yang kosong.

Tabel 2. Data perbandingan setelah dilakukan proses *smote*

Dataset	Jumlah data
ILPD asli	583
ILPD hasil <i>smote</i>	2.168

(Sumber: Data ILPD setelah diolah)

Hasil seleksi atribut yang dilakukan *software weka* adalah.

```
Selected attributes: 1,3,4,5,6,7,10 : 7
Age
TB
DB
Alkphos
Sgp_AA
Sgot_AA
A/G
```

Gambar 3. Hasil *runing weka* hasil proses seleksi fitur dengan algoritma CFS

Maka dari proses ini dihasilkan 2 *dataset*, *dataset* 1 hasil *smote* dan *dataset* 2 hasil seleksi fitur.

Pengacakan data dilakukan agar data tidak bias, *dataset* ILPD yang didapatkan dari UCI Machine Learning Repository, persebaran datanya belum merata, maksudnya persebaran data dengan keputusan *yes* dan *no* terhadap penyakit *liver* belum merata, dilakukan pengacakan data ini agar data keputusannya merata.

Persentase pembagian antara data *training* dan data *testing* yaitu 80% dan 20% dari total record pada *dataset*, jadi untuk data *training* adalah 80% dari 2.168 dihasilkan sebanyak 1734 data dan untuk data *testing* adalah 20% dari 2.168 dihasilkan sebanyak 434 data.

4. Permodelan Algoritma Klasifikasi CART, RIPPER, Penggabungan CART dan RIPPER

Tabel 3. Jumlah rule hasil penggabungan pada *dataset*

Jenis <i>dataset</i>	algoritma CART	algoritma RIPPER	Total <i>rule</i>
Dataset 1	104 rule	36 rule	140 rule
Dataset 2	57 rule	26 rule	83 rule

(Sumber: Dataset ILP setelah diolah)

5. Evaluasi

Tabel 4. Hasil confusion matrix berdasarkan pengujian dari dataset 1

99 (True Positive)	57 (False Negative)
26 (False Positive)	252 (True Negative)

(Sumber: Dataset 1 setelah diolah)

Selanjutnya menghitung nilai akurasi, sensitifitas/recall, spesifisitas, precision dan f-measure berdasarkan Tabel 4.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{99+252}{99+252+26+57} = 0,81$$

$$Sensitifitas / Recall = \frac{TP}{TP+FN} = \frac{99}{99+57} = 0,63$$

$$Spesifisitas = \frac{TN}{TN+FP} = \frac{252}{252+26} = 0,91$$

$$Precision = \frac{TP}{TP+FP} = \frac{99}{99+26} = 0,80$$

$$F-Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2 \times 0,80 \times 0,63}{0,80 + 0,63} = 0,70$$

Tabel 5. Hasil confusion matrix berdasarkan pengujian dari dataset 2

148 (True Positive)	28 (False Negative)
105 (False Positive)	153 (True Negative)

(Sumber: Dataset 2 setelah diolah)

Selanjutnya menghitung nilai akurasi, sensitifitas/recall, spesifisitas, precision dan f-measure berdasarkan tabel 5.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{148+153}{148+153+105+28} = 0,70$$

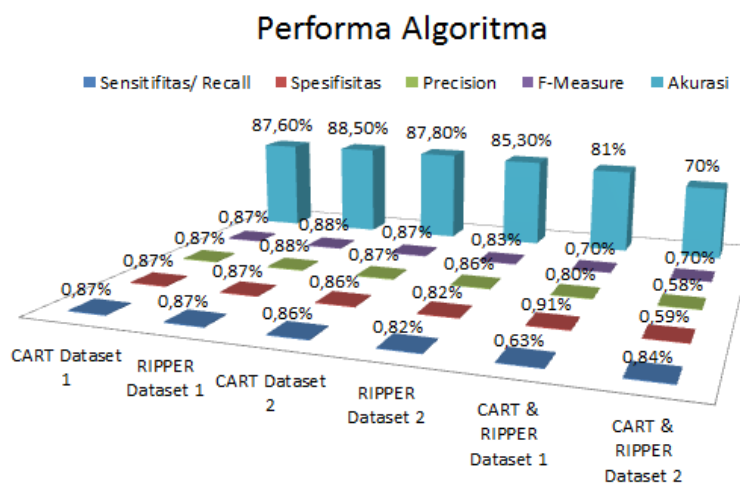
$$Sensitifitas / Recall = \frac{TP}{TP+FN} = \frac{148}{148+28} = 0,84$$

$$Spesifisitas = \frac{TN}{TN+FP} = \frac{153}{153+105} = 0,59$$

$$Precision = \frac{TP}{TP+FP} = \frac{148}{148+105} = 0,58$$

$$F-Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2 \times 0,58 \times 0,84}{0,58 + 0,84} = 0,70$$

Sehingga berdasarkan hasil performa dari masing-masing algoritma dan hasil penggabungan rule algoritma pada dua dataset adalah sebagai berikut:



Gambar 4. Perbandingan hasil pengujian data testing pada masing-masing performa algoritma

Dengan menggunakan rule penggabungan algoritma CART dan RIPPER memiliki nilai tinggi pada dataset ILPD tanpa melakukan seleksi fitur dengan algoritma CFS hasilnya adalah 81%, sedangkan akurasi yang dicapai

untuk dataset ILPD yang dilakukan seleksi fitur dengan algoritma CFS adalah 70%.

6. Penerapan

Pada tahap ini dilakukan penarikan kesimpulan untuk penelitian ini, bahwa untuk menentukan diagnosis penyakit *liver* dengan metode penggabungan *rule* dari algoritma CART dan RIPPER lebih baik ketika tidak dilakukan seleksi fitur karena tingkat akurasi lebih tinggi dibandingkan dengan dilakukan seleksi fitur.

IV KESIMPULAN

Dengan menggunakan *rule* hasil algoritma CART dan algoritma RIPPER yang telah digabungkan dari *confusion matrix* didapatkan hasil akurasi sebesar 81% dengan sensitifitas/*recall* sebesar 0,63%, spesifisitas sebesar 0,91%, *precision* sebesar 0,80% dan *f-measure* sebesar 0,70% pada dataset ILPD tanpa melakukan seleksi fitur, sedangkan pada hasil akurasi *dataset* ILPD dengan melakukan seleksi fitur yaitu akurasi sebesar 70% dengan sensitifitas/*recall* sebesar 0,84%, spesifisitas sebesar 0,59%, *precision* sebesar 0,58% dan *f-measure* sebesar 0,70%. Dari hasil tersebut dapat disimpulkan bahwa untuk mendiagnosis penyakit *liver* dengan menggunakan penggabungan algoritma CART dan RIPPER tanpa melakukan seleksi fitur.

DAFTAR PUSTAKA

- [1] Widodo, Pudji. 2014. Rule-Based Classifier Untuk Mendeteksi Penyakit Liver. *Bianglala Informasi* Vol. II No. 1 Maret 2014
- [2] Rahmawati, Eva. 2015. Analisa Komparasi Algoritma Naive Bayes Dan C4.5 Untuk Prediksi Penyakit Liver. *Jurnal Techno Nusa Mandiri* Vol. XII No. 2, September 2015.
- [3] Departemen Kesehatan Republik Indonesia. (2007). *Pharmaceutical Care Untuk Penyakit Hati*. Direktorat Bina Farmasi Komunitas Dan Klinik Ditjen Bina Kefarmasian Dan Alat Kesehatan Departemen Kesehatan RI. Diakses tanggal 2 September 2017.
- [4] Samiadi, Lika Aprilia. 2017. Berbagai Tes untuk Mendiagnosis Penyakit Hati. <https://hellosehat.com/pusat-kesehatan/penyakit-hati-penyakit-liver/berbagai-tes-untuk-mendiagnosis-penyakit-hati>. Diakses tanggal 1 September 2017.
- [5] Kumara, Rendragraha., Supriyanto, Catur. *Klasifikasi Data Mining Untuk Penerimaan Seleksi Calon Pegawai Negeri Sipil 2014 Menggunakan Algoritma Decision Tree*. *Jurnal Skripsi, Program Sarjana, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro*.