Komparasi Tingkat Akurasi *Support Vector Machine* (SVM) dan *C4.5* dalam Mengklasifikasikan Keberlangsungan Hidup Pasien Hepatitis

Putu Susi Oktaviani[1], Rima Dias Ramadhani[2], Tri Ginanjar Laksana[3], Andhika Elok Amalia[4]
[1,2]Putu (S1 Informatika, Teknologi Industri dan Informatika, Institut Teknologi Telkom Purwokerto)
[1,2]Jl.D.I. Panjaitan No.128, 53147, Purwokerto Selatan, Indonesia
e-mail: 14102034@ittelkom-pwt.ac.id[1)], rima@ittelkom-pwt.ac.id[2)], anjarlaksana@ittelkom-pwt.ac.id [3)], andhika.amalia@ittelkom-pwt.ac.id

ABSTRAK

Proses mencari kesesuain algoritma dalam pengelompokan *Dataset* hepatitis tidak mudah, berdasarkan pada atribut-atribut yang terdapat pada *dataset*. Algoritma yang digunakan pada penelitian ini yaitu SVM dan C4.5, pada kedua algoritma ini belum diketahui akurasi yang sesuai dalam mengklasifikasi hepatitis. Hepatitis sudah masuk ke dalam kategori 5 penyakit mematikan di dunia. Pada penelitian ini, membandingkan algoritma SVM dan C4.5 untuk didapatkan kesesuaian hasil akurasi. Dimana algoritma SVM dipilih, karena memiliki kemampuan berupa fungsi linier berdimensi tinggi. Sedangkan C4.5 dipilih karena memiliki kelebihan dalam mempresentasikan hasil data membentuk sebuah pohon keputusan yang mudah dipahami. Hasil yang di dapatkan dalam penelitian ini menunjukan bahwa nilai akurasi C4.5 lebih tinggi dari SVM. Dengan perolehan nilai 80,6452% pada C4.5 dan 80,3279% untuk hasil SVM. Penelitian ini diharapkan dapat memberikan solusi dalam pengelompokan *dataset* hepatitis.

Kata Kunci: Akurasi, C4.5, Hepatitis, Klasifikasi, SVM.

ABSTRACT

The process of finding algorithm compatibility in hepatitis Dataset grouping is not easy, based on the attributes contained in the dataset. The algorithms used in this study are SVM and C4.5, in which the two algorithms have not been known to be accurate in classifying hepatitis. Hepatitis has entered the category 5 deadly disease in the world. In this study, comparing SVM and C4.5 algorithms to obtain the accuracy of the results. Where the SVM algorithm is chosen, because it has the ability in the form of a high dimensional linear function. Whereas C4.5 is chosen because it has advantages in presenting the results of the data to form an easy-to-understand decision tree. The results obtained in this study indicate that the C4.5 accuracy value is higher than SVM. With the acquisition of 80.6452% in C4.5 and 80.3279% for SVM results. This research is expected to provide a solution in grouping hepatitis datasets.

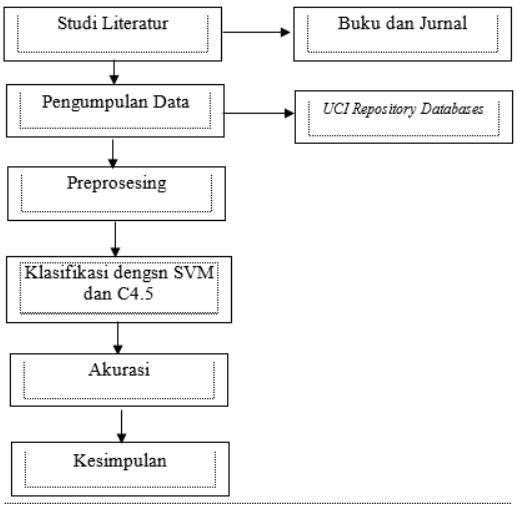
Keyword: Accuration, C4.5, Classification, Hepatitis, SVM.

I. PENDAHULUAN

Penyakit hepatitis merupakan peradangan dan kerusakan hepatosit di hati [1]. Hepatitis dianggap sebagai penyakit yang mematikan, bahkan hepatitis sudah dianggap sebagai penyebab kematian kelima di seluruh dunia. Penyebab peradangan hati disebabkan oleh infeksi (virus, bakteri, parasit), genetika, alkohol, penggunaan obat-obatan dan lemak yang berlebih dan penyakit autoimmune penyakit ini biasanya berkaitan dengan bagaimana pola hidup bersih yang dilakukan sehari - hari [2]. Berdasarkan data dari *World Health Organization* (WHO) tahun 2013, angka kematian hepatitis terus terjadi dan peningkatan terjadi sebanyak 22 persen sejak tahun 2000. Laporan ini memperkirakan infeksi hepatitis dan komplikasinya merenggut 1,45 juta jiwa di tahun 2013. Hepatitis telah membunuh 1,34 juta jiwa di tahun 2015, sebuah angka yang sebanding dengan tuberkulosis dan HIV / AIDS. Seiring adanya kemajuan di bidang teknologi kini sudah dapat dimanfaatkan dalam bidang informatika. Salah satu teknologi yang dapat dimanfaatkan dalam kasus ini adalah data mining. Data *mining* merupakan proses ekstraksi untuk mendapatkan informasi penting dengan cara menemukan sebuah pola dan pengetahuan yang menarik dari sejumlah data yang besar [3]. Terdapat beberapa metode dalam data *mining* salah satunya adalah klasifikasi. Klasifikasi merupakan cara untuk mendefinisikan suatu kelas maupun suatu objek, dengan menggunakan sejumlah data yang telah diketahui data induknya [4] . Terdapat beberapa

algoritma dalam klasifikasi diantaranya adalah SVM dan C4.5

II. METODE PENELITIAN



Gambar 14 Alur Metode Penelitian

Pada Gambar 1 merupakan gambaran dari alur penelitian. Tahapan ke 3 yaitu *pre-processing* dilakukan tahapan mengganti nilai yang hilang (*missing value)* dan normalisasi. Kemudian masuk Pada tahapan membentuk algoritma menggunakan algoritma *Support Vector Machine* dan C4.5. dilanjutkan adalah mencari nilai akurasi pada ke dua algoritma tersebut.

Pada penelitian ini data yang digunakan diadopis dari https://archive.ics.uci.edu/ml/datasets/hepatitis dengan dataset *Hepatitis Domain*. Dataset yang dipakai berasal dari *UCI Repository of Machine Learning Databases* dengan jumlah data sebanyak 3100 data yang terdiri dari 155 *record* dan 20 atribut.

Atribut	Data									
Id	1	2	3	4	5	6	7	...	...	155
Clas	2	2	2	2	2	2	1	...	...	1
X1	30	50	78	31	34	34	51	...	...	43
X2	2	1	1	1	1	1	1	...	...	1
X3	1	1	2	?	2	2	1	...	...	2
X4	2	2	2	1	2	2	2	...	...	2
X5	2	1	1	2	2	2	1	...	...	1
X6	2	2	2	2	2	2	2	...	...	2
X7	2	2	2	2	2	2	1	...	...	2
X8	1	1	2	2	2	2	2	...	...	2
X9	2	2	2	2	2	2	2	...	...	2
X10	2	2	2	2	2	2	1	...	...	1
X11	2	2	2	2	2	2	1	...	...	1
X12	2	2	2	2	2	2	2	...	...	1
X13	2	2	2	2	2	2	2	...	...	2
X14	1.00	0.90	0.70	0.70	1.00	0.90	?	...	...	1.20
X15	85	135	96	46	?	95	?	...	...	100
X16	18	42	32	52	200	28	?	...	...	19
X17	4.0	3.5	4.0	4.0	4.0	4.0	?	...	...	3.1
X18	?	?	?	80	?	75	?	...	...	42
X19	1	1	1	1	1	1	1	...	...	2

III. HASIL PENELITIAN

Dalam mencari hasil penelitian, tahapan per tahapan dilakukan dalam memperoses data.

1. Tahap Pre-processing

Tahap pre-processing merupakan tahapan di mana dataset akan diteliti kelengkapan dan juga kebersihan data. Database sangat rentan terhadap data yang noise, hilang, dan tidak konsisten. Data berkualitas rendah akan menghasilkan hasil mining berkualitas rendah. Secara umum, metode untuk menangani nilai yang hilang pada fitur adalah pada bagian dari sequential methods biasa disebut preprocessing methods [5].

$$\overline{x} = \sum_{i=0}^{n} \frac{xi}{N}$$

2. Normalisasi Data

Tahapan selanjut nya setelah data sudah lengkap, kemudian data melewati tahapan normalisasi. Dimana pada tahapan ini, data akan di ubah range nilai nya. *Range* nilai yang di terapkan yaitu [1,-1]. Tujuan dari proses ini adalah bertujuan untuk mendapatkan data dengan ukuran yang lebih kecil yang mewakili data yang asli tanpa kehilangan karakteristik sendirinya [6]. Perhitungan menormalisasikan data, sebagai berikut:

$$\gamma i = \frac{Vi - \min A}{\max A - \min A} (new \quad maxA - new\ minA) + new\_minA$$

3. *Cross Validation*

Setelah data bersih dari missing value dan telah melewati tahapan normalisasi data, kemudian membagi data dengan menggunakan metode *Cross Validation*. Pada tahapan ini, data akan di bagi menjadi 2 yaitu menjadi latih dan uji.

4. Algoritma Support Vector Machine

Data yang akan digunakan pada klasifikasi yaitu data yang sudah melewati proses pre-processing dan juga sudah di cross validasi.

Tahapan pertama yang dilakukan pada algoritma SVM yaitu mencari nilai bobot dan bias. Proses pencarian bobot dan bias dilakukan dengan cara eliminasi. Data yang digunakan adalah data latih yang di dapat dari akurasi terbesar pada Matlab.

Yi (b+W_1 X_1+ W_2 X_2)≥1

Di mana jika :

f(x)=1 jika<w,x>+b ≥1
f(x)=-1 jika<w,x>+b <-1

Data kemudian di terapkan pada persamaan diatas dimana Y merupakan kelas dan nilai dari masing-masing atribut juga dimasukkan ke dalam persamaan.

$$1\ (w_1\ -0.3521 + w_2\ 1 + w_3\ -1 + w_4\ 1 + w_5\ 1 + w_6\ 1 + w_7\ 1 + w_8\ -1 + w_9\ 1 + w_{10}\ 1 + w_{11}\ 1 + w_{12}\ 1 + w_{13}\ 1 + w_{14}\ -0.8181 + w_{15}\ -0.5613 + w_{16}\ -0.9873 + w_{17}\ -0.1162 + w_{18}\ -0.3 + w_{19}\ -1 + b) \geq 1$$

Proses eliminasi dilakukan sebanyak 94 kali sesuai dengan jumlah data latih yang digunakan. Hasil dari eliminasi untuk mendapatkan hasil bobot dan bias dapat dilihat pada Tabel di bawah ini :

ATRIBUT	BOBOT
AGE	0,092
SEX	-0,272
STEROID	-0,044
ANTIVIRALS	-0,017
FATIGUE	-0,135
MALAISE	-0,18
ANOREXIA	0,182
LIVER BIG	0,251
LIVER FIRM	0,046
SPLEEN PALPABLE	-0,135
SPIDERS	-0,258
ASCITES	-0,251
VARICES	-0,073
BILIRUBIN	0,254
ALK PHOSPAHATE	-0,098
SGOT	-0,073
ALBUMIN	-0,316

Langkah berikut nya setelah mendapatkan nilai w untuk setiap atribut nya dan mendapatkan nilai b nya, selanjutnya akan di lakukan tahapan menentukan label dari *support vektor* atau masuk ke fungsi pemisahan optimal. Dalam menentukan label untuk setiap titik data (xi) dengan menggunakan fungsi :

$$f(x) = sign\ (W_1X_1 + W_2X_2 + \ldots\ldots\ldots + W_nX_n + b\ )$$

5. Algoritma C4.5

Data latih yang sudah siap kemudian akan langsung di eksekusi. Data yang digunakan berjumlah 124 data, kemudian dari data total di cari banyak jumlah kemunculan data yang masuk ke dalam kelas hidup dan mati. Pada saat eksekusi, terdapat 124 jumlah data, terdiri dari 99 kasus mati dan 25 kasus hidup. Perhitungan C4.5 hanya berdasarkan berapa nilai entropy dan Gain Ratio terbesar. Ada pun persamaan untuk mencari nilai entropy sampai dengan nilai gain adalah sebagai berikut :

$$info\ (D) = -\sum_{i=1}^{m} Pi\ log_2\ Pi$$

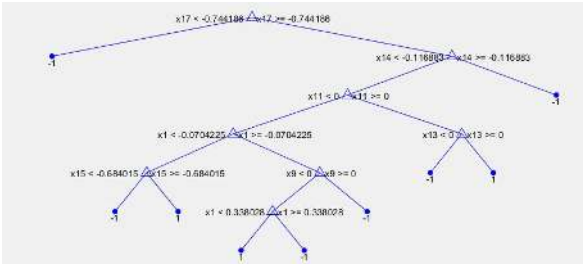
Persamaan Info Entropy :

$$info\ A(D) = \sum_{j=1}^{v} \frac{||D_{j||}}{||D||} \times info\ D_j$$

Persamaan Gain :

$$Gain\ (A) = info(D) - info_a(D)$$

Proses perhitungan menggunakan ke 3 persamaan diatas dilakukan sebanyak 19 kali. Sampai semua atribut masuk ke dalam pohon. Berikut merupakan gambaran hasil dari pohon keputusan dari data set Hepatitis :



6. Akurasi

Pada klasifikasi algoritma *Support Vector Machine* dengan percobaan dilakukan dari rentang 0,1 – 0,9 dalam melakukan pembagian data. Hasil dari *cross validasi* ini akan sangat berpengaruh terhadap hasil akurasi. Berikut Tabel 1, tabel hasil akurasi SVM:

Tabel 2 Hasil Akurasi SVM

Pembagian	Akurasi
0,1	80,0000%
0,2	80,0000%
0,3	80,0000%
0,4	**80,3279%**
0,5	79,2208%
0,6	79,3478%
0,7	79,6296%
0,8	79,6748%
0,9	79,7101%

Pada Tabel 1 di atas ditujukan hasil akurasi pada klasifikasi SVM yang di lakukan iterasi sebanyak 10 kali pada saat *cross-validation*. Dengan rentang pembagian data 0,1-0,9 yang diacak. Perhitungan nilai akurasi, berdasarkan pada hasil *confusion matrix*. Hasil *confusion matrix* dapat dilihat pada Tabel 2 di bawah ini :

Tabel 3 Hasil Confusion Matrix

	Hidup	**Mati**
Aktual hidup	10	2
Aktual mati	9	40

Nilai pada tabel 2 diatas, diambil dari hasil *cross validation* dengan pembagian 0,4 untuk data uji yang digunakan. Mencari nilai akurasi menggunakan persamaan seperti di bawah ini :

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$
$$= \frac{10 + 40}{10 + 40 + 2 + 9}$$
$$= \frac{50}{61} = 0{,}8196$$

Selanjut nya klasifikasi algoritma *C4.5* dengan percobaan dilakukan dari rentang 0,1 – 0,9 dalam melakukan pembagian data. Hasil dari *cross validasi* ini akan sangat berpengaruh terhadap hasil akurasi. Berikut Tabel 3, tabel hasil akurasi C4.5:

Tabel 3 Hasil Akurasi C4.5

Pembagian	**Akurasi**
0,1	80,0000%
0,2	**80,6452%**
0,3	79,0323%
0,4	79,0323%
0,5	79,2208%
0,6	78,4946%
0,7	79,6296%
0,8	79,8387%
0,9	79,8516%

IV. PEMBAHASAN

Proses klasifikasi pada algoritma *Support Vector Machine* dengan C4.5 diawali pada tahapan penghilangan atribut-atribut yang nilai nya hilang. Kemudian dilanjutkan dengan menormalisasikan *dataset* tersebut. Tahapan ini sangat berguna untuk algoritma klasifikasi yang melibatkan pengukuran terhadap jarak, normalisasi ini juga digunakan untuk mencegah atribut yang memiliki rentang yang besar. Berikut nya tahapan membagi data, data dibagi dengan menggunakan *cross validation*, dari proses ini menghasilkan data latih dan data uji.

Data latih digunakan untuk mengklasifikasi dan membentuk sebuah model *classifer*. Kemudian data uji digunakan untuk melakukan pengujian data pada Model *Classifier*. Model *classifer* yang dihasilkan untuk melakukan perhitungan dimana nantinya perhitungan tersebut diambil sebagai suatu tetapan. Ketetapan yang digunakan dalam algoritma svm adalah mencari bobot, *bias, support vektor* dan juga *hyperplane* dan juga penentuan kelas yang optimal. Sedangkan untuk C4.5 tetapan yang digunakan adalah *entropy* dan juga gain untuk membangun sebuah pohon keputusan sebagai implementasinya.

Proses klasifikasi yang dimiliki ke dua algoritma ini berbeda beda dalam proses mengklasifikasikan. Terlihat dari hasil akurasi yang ada dan terlihat pada saat melakukan proses perhitungan manual. Pada SVM saat mengklasifikasikan terlebih dahulu mencari bobot dari setiap atribut nya. Di dalam pembobotan ini di pengaruhi dari jumlah data pada satu *record* data nya. Proses pengklasifikasian svm mendapatkan nilai akurasi yang rendah terlihat pada saat memproses SVM secara manual. Pada saat proses SVM tergolong klasifikasi yang generalisasi, karena pada proses nya mengambil perhitungan berdasarkan *hyperplane*.

Sedangkan pada algoritma C4.5 hasil akurasi menunjukan cukup lebih baik dari SVM. Hal tersebut dapat terjadi saat pengujian. Dapat dikatakan lebih baik dari svm karena dalam tahapan klasifikasi tidak dilakukan secara general. Yang dimaksudkan tidak dengan *general*, di setiap akan memulai eksekusi di setiap atributnya C4.5 mengelompokkan setiap atributnya dengan sangat baik dan memberi perlabelan atau *range* nilai pada setiap proses eksekusi per atribut nya. Sehingga klasifikasi nya dapat dikatakan lebih khusus. Karena dilakukan dengan memperlakukan proses per atribut nya. Di buktikan juga pada saat satu kali proses mencari nilai *entropy* per partisi data, hanya untuk mendapatkan salah satu nilai *gain* terbesar dari setiap atribut. Pada eksekusi di penelitian ini, terjadi hal yang sulit dikontrol pada saat menampilkan data. Karena menggunakan *cross validation* dengan variasi pengacakan, pada C4.5 selain data akan selalu diacak, hal ini juga berpengaruh terhadap proses pembentukan pohon setiap kali akan menjalankan pengacakan pada tools namun nilai akurasi tetap stabil. Hasil akurasi menggunakan data test menunjukan hasil yang sama dengan pembagian jumlah data.

V. PENUTUP

A. Kesimpulan

Pada penelitian ini dapat disimpulkan berdasarkan pada hasil dan analisa, bahwa hasil klasifikasi dengan SVM dan C4.5 itu menghasilkan rentang nilai akurasi yang tidak jauh berbeda. Hal ini dapat terjadi karena pada tahapan awal, sebelum membagi data dilakukan proses normalisasi. Dimana data diubah range nilai nya menjadi 1 dan -1. Pada C4.5 mendapatkan hasil akurasi lebih tinggi dari SVM karena dalam tahapan klasifikasian nya, C4.5 memproses persatu data atribut. Beda hal nya dengan SVM yang cenderung mengklasifikasikan nya secara general, cakupan nya lebih luas. Hasil yang diperoleh C4.5 sebesar 80,6452% tidak jauh berbeda dengan SVM yang mendapatkan nilai akurasi 80,3279%*Saran*

B. Saran

Banyak nya keterbatasan dalam penelitian ini, maka perlu adanya beberapa hal yang harus dilakukan dalam penelitian yang akan datang, diantaranya:

1. Untuk mengetahui tingkat yang lebih baik dari ke dua algoritma ini adalah bisa diuji atau digunakan data hepatitis yang memiliki *record* data yang lebih banyak .
2. Membandingkan antara tools yang satu dengan yang lain
3. Membandingkan hasil akurasi menggunakan semua variasi cross validasi.
4. Mencoba melakukan tahapan *pruning* dan optimasi pada C4.5.

DAFTAR PUSTAKA

[1] M. Rouhani and M. M. Haghighi, "The Diagnosis Of Hepatitis Diseases By Support Vector Machines and Artificial Neural Networks," *2009 Int. Assoc. Comput. Sci. Inf. Technol. - Spring Conf. IACSIT-SC 2009*, pp. 456–458, 2009.

[2] L. Saumi Ramdhani, "Penerapan Particle Swarm Optimization (PSO) Untuk Seleksi Atribut Dalam Meningkatkan Akurasi Prediksi Diagnosis Penyakit Hepatitis Dengan Metode Algoritma C4.5," *Swabumi*, vol. IV, no. 1, pp. 1–15, 2016.

[3] H. Jiawei, M. Kamber, J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.

[4] G. Suwardika, "Pengelompokan Dan Klasifikasi Pada Data Hepatitis Dengan Menggunakan Support Vector Machine (SVM), Classification And Regression Tree (Cart) Dan Regresi Logistik Biner," *J. Educ. Res. Eval.*, vol. 1, pp. 183–191, 2017.

[5] G. Tsoumakas, I. Katakis, and I. Vlahavas, *Data Mining and Knowledge Discovery Handbook*. 2010.

[6] H. Penelitian and F. Teknik, "Prediksi Curah Hujan dengan Jaringan Saraf Tiruan," *Pros. Fak. Tek.*, vol. 6, no. 1, pp. 978–979, 2012.