INTI NUSA MANDIRI
DOI: https://doi.org/10.33480/inti.v15i1.1421
VOL. 15. NO. 1 AGUSTUS 2020
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: PPPM Nusa Mandiri

OPTIMASI ALGORITMA C4.5 MENGGUNAKAN ALGORITMA GENETIKA UNTUK PREDIKSI PENYAKIT HEPATITIS

Wisti Dwi Septiani

Program Studi Sistem Informasi
Universitas Bina Sarana Informatika
www.bsi.ac.id
wisti.wst@bsi.ac.id

Abstract— Hepatitis is an infectious disease which is a public health problem that affects morbidity, mortality, public health status, life expectancy, and other socio-economic impacts. Hepatitis is inflammation of the liver that can increase into liver cancer. The most common causes of hepatitis are those caused by hepatitis B and C viruses. Previous research used the data mining classification method C.45 algorithm and showed an accuracy rate of 77.29%. The purpose of this study is to improve the accuracy of the C.45 algorithm by optimization using genetic algorithm. The results of this further research are the decision tree and an increase in the accuracy rate of 12.42% from 77.29% to 89.71%.

Keywords*: algoritma C4.5, data mining, hepatitis, genetic algortima*

Abstrak—Hepatitis merupakan salah satu penyakit menular yang menjadi masalah kesehatan masyarakat yang berpengaruh terhadap angka kesakitan, angka kematian, status kesehatan masyarakat, angka harapan hidup, dan dampak sosial ekonomi lainnya. Hepatitis adalah peradangan hati yang bisa berkembang menjadi kanker hati. Penyebab paling umum Hepatitis adalah yang disebabkan oleh Virus Hepatitis B dan C. Penelitian sebelumnya menggunakan metode klasifikasi data mining Algoritma C.45 dan menunjukkan tingkat akurasi 77.29%. tujuan dari penelitian ini adalah untuk meningkatkan akurasi algoritma C4.5 dengan cara dilakukan optimasi menggunakan algortima genetika. Hasil dari penelitian lanjutan ini adalah pohon keputusan dan terjadi kenaikan tingkat akurasi sebesar 12.42 % dari 77.29 % menjadi 89.71 %.

Kata kunci: algoritma C4.5, algoritma genetika, data mining, hepatitis

PENDAHULUAN

Penyakit hepatitis merupakan penyakit peradangan hati karena infeksi virus yang menyerang dan menyebabkan kerusakan pada sel-sel dan fungsi organ hati. Penyakit hepatitis merupakan penyakit cikal bakal penyakit kanker hati . Indonesia merupakan salah satu negara yang memiliki edemisitas tinggi Hepatits B, terbesar kedua di negara South East Asean Regional (SEAR) setelah negara Myanmar (Buani, 2018). Prevalensi Hepatitis di Indonesia pada tahun 2013 sebesar 1,2% meningkat dua kali dibanding tahun 2007 dan semakin meningkat pada penduduk di atas 15 tahun. Menurut Prasetyo dalam (Septiani, 2014) saat ini dalam dunia kesehatan, rekam medis telah menyimpan gejala-gejala penyakit pasien dan diagnosis penyakitnya. Hal seperti ini tentu sangat berguna bagi para ahli kesehatan untuk dapat dijadikan sebagai bantuan dalam mengambil keputusan terhadap diagnosis penyakit pasien.

Seiring dengan perkembangan ilmu pengetahuan dan teknologi informasi, kehadiran cabang ilmu data mining telah menarik banyak perhatian dalam dunia sistem informasi. Data mining yang sering juga disebut *knowledge discovery in database* (KDD) adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menentukan pola keteraturan, pola hubungan dalam set data berukuran besar (Santosa, 2007).

Tinjauan studi data mining dalam hal prediksi penyakit hepatitis telah dipublikasikan dengan beberapa metode yaitu prediksi hepatitis menggunakan *Backpropagation* dan *Naïve Bayes* (Karlık, 2011), *Support Vector Machine (SVM)* dengan fitur seleksi (Kumar et al., 2012), Komparasi metode *Naïve Bayes* dan *Support Vector Machine* (Fridayanthie, 2015), serta Komparasi Algoritma Berbasis *Neural Network* (Saputra, 2017).

Pada penelitian ini dilakukan optimasi terhadap Algoritma C4.5 menggunakan algoritma Genetika. Optimasi adalah proses menyelesaikan suatu masalah tertentu supaya berada pada kondisi yang paling menguntungkan dari suatu

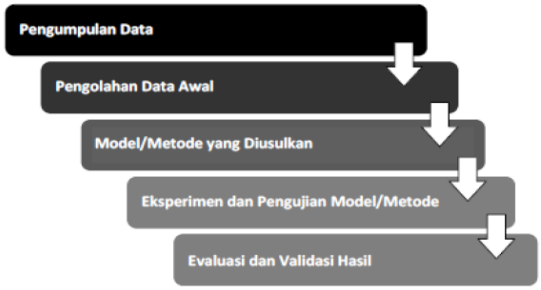
INTI NUSA MANDIRI
DOI: https://doi.org/10.33480/inti.v15i1.1421

sudut pandang, yaitu berhubungan dengan pencarian nilai minimum atau nilai maksimum (Buani, 2016). Pada penerapan optimasi menggunakan algoritma genetika terdapat peningkatan akurasi dari 97,66% menjadi 99,33% pada agoritma Naïve Bayes untuk prediksi *fertility* (Buani, 2016)dan peningkatan akurasi dari 83,81% menjadi 86,47% pada algoritma C4.5 untuk prediksi *phising website* (Sunge, 2018).

Decision Tree 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. Tujuan dari penelitian ini adalah melakukan optimasi algoritma C4.5 menggunakan algoritma genetika untuk meningkatkan nilai akurasinya sehingga prediksi yang dihasilkan lebih baik dan akurat.

BAHAN DAN METODE

Dalam menyelesaikan penelitian perlu dibuat rancangan penelitian yang berguna sebagai pedoman atau acuan penelitian. Rancangan penelitian sebagai berikut:

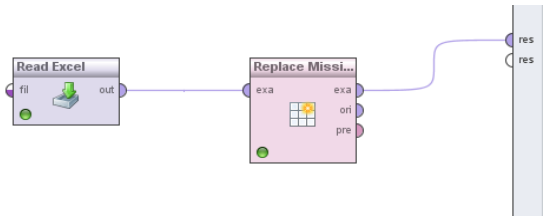


Sumber: (Septiani, 2020)
Gambar 1. Rancangan Penelitian

Penelitian dilakukan dengan cara melakukan eksperimen dalam bentuk sistem penunjang keputusan untuk prediksi pasien penyakit hepatitis menggunakan metode klasifikasi data mining yaitu Algoritma C4.5 dengan fitur seleksi Algoritma Genetika. Tools yang digunakan untuk mengukur akurasi dari penelitian ini adalah Rapid Miner.

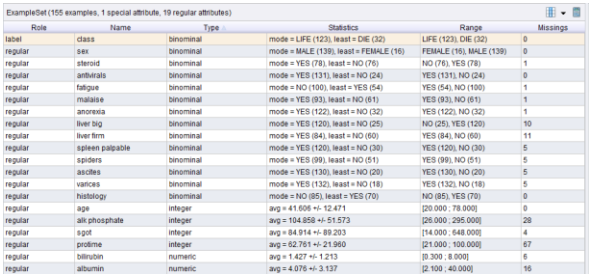
Sumber data yang digunakan pada penelitian ini adalah data sekunder berupa data penyakit hepatitis yang didapat dari *Machine Learning Repository* UCI (Universitas California Invene) dengan alamat web: http://archives.ics.uci.edu/ml/.

Pengolahan data awal dilakukan: (1) *Data validation*, untuk mengidentifikasi dan menghapus data yang ganjil, data yang tidak konsisten dan data yang tidak lengkap. (2) Data integration and transformation, untuk meningkatkan akurasi dan efisiensi algoritma. Data yang digunakan dalam penelitian ini bersifat kategorikal. (3) *Data size reduction and dicritization* untuk memperoleh dataset dengan jumlah atribut dan *record* lebih sedikit tetapi bersifat informatif.



Sumber: (Septiani, 2020)
Gambar 2. Model *Replace Missing*

Pada gambar 2 dilakukan pengolah data awal berupa model *replace missing* yang bertujuan untuk menghilangkan duplikasi dan anomali atau inkonsistensi data.


ExampleSet (155 examples, 1 special attribute, 19 regular attributes)

Role	Name	Type	Statistics	Range	Missings
label	class	binominal	mode = LIFE (123), least = DIE (32)	LIFE (123), DIE (32)	0
regular	sex	binominal	mode = MALE (139), least = FEMALE (16)	FEMALE (16), MALE (139)	0
regular	steroid	binominal	mode = YES (78), least = NO (76)	NO (76), YES (78)	1
regular	antivirals	binominal	mode = YES (131), least = NO (24)	YES (131), NO (24)	0
regular	fatigue	binominal	mode = NO (100), least = YES (54)	YES (54), NO (100)	1
regular	malaise	binominal	mode = YES (93), least = NO (61)	YES (93), NO (61)	1
regular	anorexia	binominal	mode = YES (122), least = NO (32)	YES (122), NO (32)	1
regular	liver big	binominal	mode = YES (120), least = NO (25)	NO (25), YES (120)	10
regular	liver firm	binominal	mode = YES (84), least = NO (60)	YES (84), NO (60)	11
regular	spleen palpable	binominal	mode = YES (120), least = NO (30)	YES (120), NO (30)	5
regular	spiders	binominal	mode = YES (99), least = NO (51)	YES (99), NO (51)	5
regular	ascites	binominal	mode = YES (130), least = NO (20)	YES (130), NO (20)	5
regular	varices	binominal	mode = YES (132), least = NO (18)	YES (132), NO (18)	5
regular	histology	binominal	mode = NO (85), least = YES (70)	NO (85), YES (70)	0
regular	age	integer	avg = 41.606 +/- 12.471	[20.000 ; 78.000]	0
regular	alk phosphate	integer	avg = 104.858 +/- 51.573	[26.000 ; 295.000]	28
regular	sgot	integer	avg = 84.914 +/- 89.203	[14.000 ; 648.000]	4
regular	protime	integer	avg = 62.761 +/- 21.960	[21.000 ; 100.000]	67
regular	bilirubin	numeric	avg = 1.427 +/- 1.213	[0.300 ; 8.000]	6
regular	albumin	numeric	avg = 4.076 +/- 3.137	[2.100 ; 40.000]	16

Sumber: (Septiani, 2020)
Gambar 3. *Missing atributes*

Gambar 3 merupakan hasil dari missing atributes yang dilakukan dan data yang inkonsistensi akan dihilangkan. Dari data yang diperoleh sebanyak 155 *record* pasien proses pengolahan data awal didapatkan hasil atribut yang digunakan adalah *age, sex, steroid, antivirals, fatigue, malaise, anorexia, liver_big, liver_firm, spleen_palpable, spiders, ascites, varices, bilirubin, alk_phosphate, sgot, albumin, protime, histology,* dan *class (*atribut hasil prediksi*)*. Sebanyak 15 5 data dengan 123 data dengan *class* "HIDUP" dan 32 data dengan *class* "MATI".

Berdasarkan tugasnya, *data mining* dikelompokkan menjadi 6 yaitu deskripsi, estimasi, prediksi, klasifikasi, clustering, dan asosiasi (Larose, 2005). Klasifikasi (taksonomi) adalah proses menempatkan objek tertentu (konsep) dalam satu set kategori, berdasarkan masing-masing objek (konsep) *property* (Gorunescu, 2011). Proses klasifikasi didasarkan pada empat komponen

mendasar yaitu kelas, prediktor, *training set*, dan pengujian *dataset*.
Diantara model klasifikasi yang paling popular adalah *Decision/Classification Trees, Bayesian Classifiers/Naïve Bayes Classifiers, Neural Networks, Statistical Analysis, Genetic Algorithms, Rough Sets, K-Nearest Neighbor Classifier, Rule-based Methods, Memory Based Reasoning, Support Vector Machines* (Gorunescu, 2011).

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:

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

3. Hitung nilai gain dengan rumus:

$$Entropy\ split = \sum_{i=1}^{p} \binom{n1}{n}.\ IE(i) \quad .....(2)$$

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

Pada tahun 1970 Algoritma Genetica (GA) diperkenalkan oleh John Holland di Universitas Michigan, bahwa dari bagian masalah merupakan bentuk dari adaptasi dari alam maupun buatan yang dapat diformulakan menjadi bagian genetika (Sunge, 2018). Menurut Desiani dan Muhammad dalam (Buani, 2018) Algoritma genetika merupakan suatu algoritma pencarian berdasarkan pada mekanisme seleksi alam dan genetika alam. Algoritma genetika dimulai dengan sekumpulan solusi awal(individu) yang disebut populasi. Satu hal yang sangat penting adalah bahwa satu individu menyatakan satu solusi. Populasi awal akan berevolusi menjadi populasi baru melalui serangkaian iterasi (generasi). Pada akhir iterasi, algoritma genetika mengembalikan satu anggota populasi yang terbaik sebagai solusi untuk masalah yang dihadapi. Pada setiap iterasi, proses evolusi yang terjadi adalah sebahai berikut:

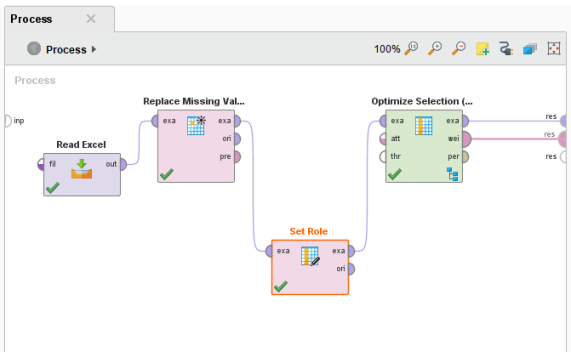
a) Dua individu dipilih sebagai orang tua (*parent*) berdasarkan mekanisme tertentu. Kedua parent ini kemudian dikawinkan melalui operator crossover (kawin silang) untuk menghasilkan dua individu anak atau offspring.
b) Dengan probabilitas tertentu, dua individu anak ini mungkin mengalami perubahan gen melalui operator mutation.
c) Suatu skema penggantian *(replacement scheme)* tertentu diterapkan sehingga menghasilkan populasi baru.

Proses ini terus berulang sampai kondisi berhenti *(stopping condition)* tertentu. Kondisi berhenti bisa berupa jumlah iterasi tertentu, waktu tertentu, atau ketika variansi individu-individu dalam populasi tersebut sudah lebih kecil dari suatu nilai tertentu yang diinginkan.

HASIL DAN PEMBAHASAN

Pengujian dengan tools RapidMiner untuk mengolah data dengan tahapan sebagai berikut:

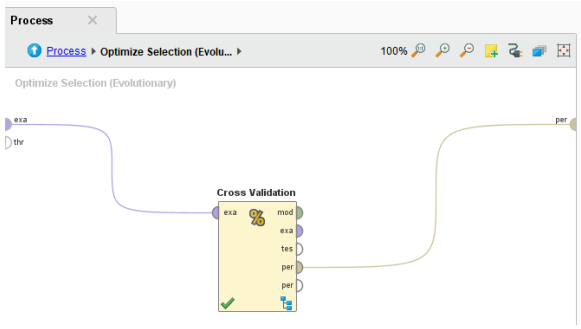
1. Pengujian Menggunakan Algoritma C4.5 dengan fitur seleksi Algoritma Genetika.



Sumber: (Septiani, 2020)
Gambar 4. *Optimize Selection* (*Evolutionary*)

Berdasarkan Gambar 4, database hepatitis dihubungkan dengan *Replace Missing Value* untuk mencari apakah masih ada *record* yang kosong. Setelah itu dihubungkan dengan *Set Role* untuk menentukan atribut yang menjadi label dalam *dataset*. Selanjutnya *Set Role* dihubungkan dengan *Feature Optimize Selection* (*Evolutionary*) dan didalamnya diberikan *Cross Validation*.

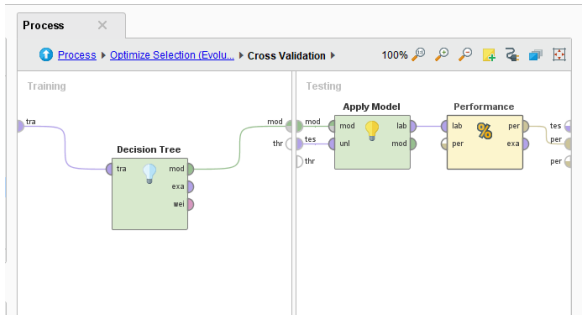
DOI: https://doi.org/10.33480/inti.v15i1.1421



Sumber: (Septiani, 2020)

Gambar 5. *Cross Validation*

Berdasarkan gambar 5, penggunaan *Cross Validation* dalam prediksi hepatitis terdiri dari *10-fold validation* berjumlah 155 data yang terbagi 20 atribut dipecah menjadi 10 atribut dan setiap bagian dibagi secara acak. Dengan perbandingan 1:9 dimana 1 bagian merupakan *data testing* dan 9 bagian dijadikan *data training*. Pada *Cross Validation* terdapat tahap dalam penggunaan algoritma *Decision Tree*.



Sumber: (Septiani, 2020)

Gambar 6. *Model Decision Tree*

Pada gambar 6 setelah model *Decision Tree* maka tahap terakhir dilakukan proses terhadap model tersebut untuk menampilkan hasil berupa tingkat akurasi,

2. Hasil akurasi dari pengujian Algoritma C4.5 dengan fitur seleksi Algoritma Genetika.

T

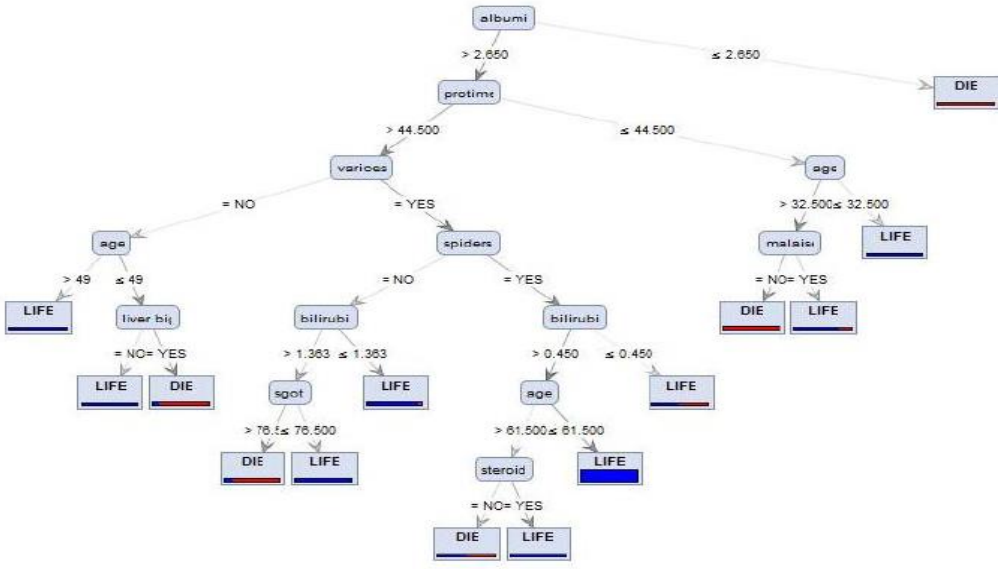
Tabel 1. Akurasi Algoritma *Decision Tree*

	True LIFE	True DIE	Class Precision
Pred LIFE	120	13	90.23%
Pred DIE	3	19	86.36%
Class Recall	97.56%	59.38%	
Accuracy		**89.71%**	

Sumber: (Septiani, 2020)

Berdasarkan data di Tabel 1 dapat diambil kesimpulan bahwa hasil prediksi menggunakan Decision Tree dengan fitur seleksi Algoritma Generika tingkat akurasinya adalah 89.71%. Terjadi peningkatan nilai akurasi dari penelitian sebelumnya yaitu 77.29% (Septiani, 2014) dengan menggunakan metode yang sama tetapi tidak menggunakan fitur seleksi.

Hasil dari penggunaan Algoritma C4.5 ini adalah pohon keputusan seperti pada gambar 7 yang menghasilkan 14 *rule*.



Sumber: (Septiani, 2020)

Gambar 7. Pohon Keputusan

VOL. 15. NO. 1 AGUSTUS 2020
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: PPPM Nusa Mandiri

Berdsasarkan gambar 7 dari pohon keputusan didapatkan 14 *rule* untuk memprediksi penyakit hepatitis. *Rule* yang didapat sebagai berikut :
R1: Jika albumin <= 2,650 maka pasien "DIE".
R2: Jika albumin > 2,650 dan protime > 44,500 dan varices = NO dan age > 49 tahun maka pasien "LIFE".
R3: Jika albumin > 2,650 dan protime > 44,500 dan varices = NO dan age <= 49 tahun dan liver_big = NO maka pasien "LIFE".
R4: Jika albumin > 2,650 dan protime > 44,500 dan varices = NO dan age <= 49 tahun dan liver_big = YES maka pasien "DIE"
R5: Jika albumin > 2,650 dan protime > 44,500 dan varices = YES dan spiders = NO dan bilirubin > 1,363 dan sgot > 76,500 maka pasien "DIE".
R6: Jika albumin > 2,650 dan protime > 44,500 dan varices = YES dan spiders = NO dan bilirubin > 1,363
dan sgot <= 76,500 maka pasien "LIFE".
R7: Jika albumin > 2,650 dan protime > 44,500 dan varices = YES dan spiders = NO dan bilirubin <= 1,363 maka pasien "LIFE".
R8: Jika albumin > 2,650 dan protime > 44,500 dan varices = YES dan spiders = YES dan bilirubin > 0,450 dan age > 61,5 tahun dan steroid = NO maka pasien "DIE".
R9: Jika albumin > 2,650 dan protime > 44,500 dan varices = YES dan spiders = YES dan bilirubin > 0,450 dan age > 61,5 tahun dan steroid = YES maka pasien "LIFE".
R10: Jika albumin > 2,650 dan protime > 44,500 dan varices = YES dan spiders = YES dan bilirubin > 0,450 dan age <= 61,5 tahun maka pasien "LIFE".
R11: Jika albumin > 2,650 dan protime > 44,500 dan varices = YES dan spiders = YES dan bilirubin <= 0,450 maka pasien "LIFE".
R12: Jika albumin > 2,650 dan protime <= 44,500 dan age > 32,5 tahun dan malaise = NO maka pasien "DIE".
R13: Jika albumin > 2,650 dan protime <= 44,500 dan age > 32,5 tahun dan malaise = YES maka pasien "LIFE".
R14: Jika albumin > 2,650 dan protime <= 44,500
dan age <= 32,5 tahun maka pasien "LIFE".

KESIMPULAN

Rules yang dihasilkan dari pengujian penggunaan fitur seleksi Algoritma Genetika yang diterapkan pada Algoritma C4.5 dapat dijadikan kontribusi dalam pengambilan keputusan terhadap penyakit hepatitis. Evaluasi dalam pengujian menggunakan Algoritma C4.5 dengan seleksi fitur Algoritma Genetika ini didapatkan nilai akurasi 89.71%. Hasil penelitian ini menunjukkan adanya peningkatan nilai akurasi sebesar 12.42 % dari penelitian sebelumnya 77.29 % tanpa fitur seleksi. Sehingga dapat disimpulkan bahwa penggunaan fitur seleksi mampu meningkatkan nilai akurasi. Penelitian ini dapat dijadikan masukkan untuk dilanjutkan kembali dengan metode optimasi lain seperti Adabbost dan PSO. Berdasarkan penelitian ini dapat diberikan saran untuk diadakannya penelitian lebih lanjut dengan melakukan pengujian dengan metode lain seperti SVM, Nural Network, ataupun komparasi dari beberapa metode klasifikasi data mining.

REFERENSI

Buani, D. C. P. (2016). Optimasi Algoritma Naive Bayes Dengan Menggunakan Algoritma Genetika Untuk Prediksi Kesuburan (Fertility). *Jurnal Evolusi*, *4*(1), 55–64. https://doi.org/10.31294/evolusi.v4i1.3397

Buani, D. C. P. (2018). Prediksi Penyakit Hepatitis Menggunakan Algoritma Naive Bayes Dengan Seleksi Fitur Algoritma Genetika. *Jurnal Evolusi*, *6*(2), 1–5. https://doi.org/10.31294/evolusi.v6i2.4381

Fridayanthie, E. W. (2015). Analisa Data Mining Untuk Prediksi Penyakit Hepatitis Dengan Menggunakan Metode Naive Bayes dan Support Vector Machine. *Jurnal Khatulistiwa Informatika*, *3*(1), 24–36.

Gorunescu, F. (2011). *Data Mining: Concepts and Techniques*. Springer.

Karlık, B. (2011). Hepatitis Disease Diagnosis Using Backpropagation and the Naive Bayes Classifiers. *Journal of Science and Technology*, *1*(1), 49–62. https://www.academia.edu/20478023/Hepatitis_Disease_Diagnosis_Using_Backpropagation_and_the_Naive_Bayes_Classifiers

Kumar, V., Sharathi, V., & Devi, G. (2012). Hepatitis Prediction Model based on Data Mining Algorithm and Optimal Feature Selection to Improve Predictive Accuracy. : *: International Journal of Computer Applications (0975-8887)*, *51*(19).

Larose, D. T. (2005). *Discovering Knowledge in Database*. John Willey & Sons Inc.

Santosa, B. (2007). *Data Mining Teknik Pemanfaat Data Untuk Keperluan Bisnis*. Graha Ilmu.

Saputra, S. (2017). Komparasi Algoritma Berbasis Neural Network dalam Mendeteksi Penyakit

Hepatitis. *Faktor Exacta Exacta*, *10*(1), 40–49. https://doi.org/10.30998/faktorexacta.v10i1.1304

Septiani, W. D. (2014). Penerapan Algoritma C4.5 Untuk Prediksi Penyakit Hepatitis. *Techno Nusa Mandiri*, *11*(1), 69–78. https://doi.org/10.33480/techno.v11i1.172

Septiani, W. D. (2020). *Laporan Akhir Penelitian Mandiri: Optimasi Algoritma C4.5 Menggunakan Algoritma Genetika Untuk Prediksi Penyakit Hepatitis*. Universitas Bina Sarana Informatika.

Sunge, A. S. (2018). Optimasi Algoritma C4.5 Menggunakan Genetic Algoritma Dalam Memprediksi Website Phishing. *Seminar Nasional Inovasi Dan Tren (SNIT)*, 92–96. http://seminar.bsi.ac.id/snit/index.php/snit-2018/article/view/47