Prediksi Harapan Hidup Penderita Hepatitis Kronik Menggunakan Metode-Metode Klasifikasi

Siti Khomsah
Program Studi Teknik Informatika
Fakultas Ilmu Komputer Universitas Alma Ata
Yogyakarta
sitikhomsah5@gmail.com

***Abstrak*—Data Riskesdas 2013 menunjukkan 28 juta penduduk Indonesia terinfeksi hepatitis B atau C. Potensi penderita hepatitis kronik sebesar empat belas juta dan satu koma empat juta diantaranya berpotensi menjadi penderita kanker hati. Perawatan bagi pasien hepatitis B kronik bertujuan memperpanjang harapan hidup pasien. Hepatitis C merupakan penyebab utama kanker hati dan sirosis. Vaksin yang tepat bagi penderita hepatitis kronik belum ditemukan sehingga pengobatannya hanya bertujuan memperpanjang harapan hidup pasien. Masa depan kesehatan pasien hepatitis kronik atau akut dapat diukur dari gejala-gejala hasil pemeriksaan baik fisik maupun laboratorium. Berdasarkan hasil pemeriksaan, dokter dapat memprediksi apakah pasien berisiko meninggal dunia karena penyakit tersebut sehingga dapat memberikan perlakuan yang tepat pada pasien. Data mining adalah salah satu teknik untuk menemukan pola informasi dari *dataset* pasien hepatitis. Pola informasi tersebut digunakan untuk membangun model yang dapat memprediksi resiko kematian pasien hepatitis. Klasifikasi adalah salah satu teknik dalam data mining untuk analisis prediksi. Penelitian bertujuan menerapkan metode data mining klasifikasi untuk memprediksi harapan hidup penderita hepatitis kronik. Fokus penelitian adalah membandingkan beberapa metode klasifikasi dan akurasinya dalam memprediksi harapan hidup pasien hepatitis. Metode yang diajukan adalah K-NN, Naive Bayes, D-Tree, dan Random forest. Model yang dirancang akan diuji menggunakan 155 data penderita hepatitis kronik atau akut. Performance model diukur berdasarkan nilai akurasi dan AUC. Model yang dirancang akan diuji menggunakan 155 data penderita hepatitis kronik atau akut. Kinerja model diukur berdasarkan nilai akurasi dan AUC. Metode validasi menggunakan *k-fold cross validation* dengan k = 10. Hasil pengujian model menunjukkan Random forest merupakan metode yang paling akurat yaitu mencapai 79.35%. Nilai AUC Naive Bayes, D-Tree, dan Random forest lebih dari 0.8, artinya ketiga model tersebut bagus sebagai *classifier*. Sedangkan nilai AUC K-NN adalah 0.7 artinya K-NN hanya pada level *fair* atau cukup.**

Kata kunci— hapatitis; resiko;data mining; klasifikasi;prediksi

I. PENDAHULUAN

Hepatitis adalah salah satu jenis penyakit endemis di beberapa negara berkembang, termasuk Indonesia. Penyakit ini disebabkan oleh infeksi jamur, bakteri, virus, obat-obatan, konsumsi alkohol, lemak berlebihan, atau penyakit *autoimmune*. Ada 5 jenis hepatitis mulai dari ringan sampai dengan kronik, yaitu A, B, C, D, E. Data riset kesehatan dasar (Riskesdas) 2013 menunjukkan bahwa setiap 100 orang di Indonesia terdapat 10 penduduk yang terinfeksi virus hepatitis C atau B. Sehingga diperkirakan terdapat 28 juta penduduk yang terinfeksi, 14 juta orang diantaranya berpotensi menjadi hepatitis kronik, dan 1,4 juta dari yang kronik tersebut berpotensi terkena kanker hati [1]. Pada tahun 2013, Indonesia termasuk negara endemis hepatitis B pada urutan kedua tertinggi di Asia Tenggara[1]. Serangakaian tes untuk diagnosis hepatitis biasanya dilakukan setelah ada indikasi atau gejala yang dirasakan pasien atau ditemukan tidak sengaja pada pemeriksaan lainnya. Hepatitis kronik seperti hepatitis B, C, atau D dapat berubah menjadi akut dan menimbulkan sirosis bahkan kanker hati. Saat pasien sudah dinyatakan mengidap hepatitis kronik maka berpotensi menjadi hepatitis akut bahkan berisiko kematian. Dokter tidak dapat menentukan harapan hidup penderita pasien hepatitis kronik atau akut.

Data mining klinik adalah penerapan metode data mining untuk tujuan menggali informasi data medis dan data klinis [2] [3]. Dengan metode ini, kondisi pasien dimasa masa depan dapat diprediksi berdasarkan observasi data pasien lainnya atau di masa lalu [4] [5]. Salah satu metode prediksi adalah klasifikasi. Berbagai metode klasifikasi diuji coba untuk melihat akurasi hasil prediksi pada data pasien hepatitis [5] [6][7].

Penelitian ini bermaksud menerapkan metode data mining klasifikasi untuk memprediksi harapan hidup penderita hepatitis kronik. Fokus penelitian adalah membandingkan beberapa metode klasifikasi dan akurasinya dalam memprediksi harapan hidup pasien hepatitis.

II. TINJAUAN PUSTAKA

Data mining adalah disiplin bidang ilmu komputer yang bermaksud menggali informasi dan pola pengetahuan dari data kumpulan data besar [8]. Metode data mining klasifikasi dapat digunakan untuk analisis prediksi data klinis. Beberapa diantaranya adalah K-NN, Decesion Tree, Random forest, Naive Bayes.

Penelitian [5] menyebutkan metode Decesion Tree (D-Tree) adalah teknik yang paling sering digunakan untuk klasifikasi dan prediksi diantaranya adalah ID3 dan C.45[5]. Penelitian terkait *dataset* pasien hepatitis adalah uji coba beberapa algoritma klasifikasi seperti Naive Bayes, BayesNet, Random forest, Naive Bayes Updatable, J48, dan Multi Layer Perceptron menggunakan *dataset* dari UCI *learning repository*[6]. Hasilnya adalah akurasi model dan kecepatan proses menunjukkan Naive Bayes merupakan metode terbaik untuk *dataset* tersebut[6]. Penelitian yang lain adalah penerapan metode klasifikasi Logistic Regression, Decision Tree (D-Tree), Linear Support Vector, dan Naive Bayes pada *dataset* hepatitis. Tujuannya untuk mengklasifikasikan apakah seseorang akan tetap hidup atau mati [9]. Penelitian yang lain adalah pengembangan model untuk mengidentifikasi pasien beresiko tinggi kanker hati, menggunakan teknik analisis prediksi data mining [4]. Penelitian yang lain adalah tentang penalaran berbasis kasus untuk penyakit hepatitis. Penelitian ini mengkombinasikan dua metode yaitu PSO dan CBR (*Case-Based Reasoning*) untuk menegakkan diagnosis penyakit hepatitis. Data diperoleh dari *dataset* UCI *machine learning repository* dan digunakan untuk membandingkan lima metode klasifikasi yang lain dan metode CBR-PSO mendapatkan akurasi tertinggi yaitu of 93.25%[10]. Hasil perbandingan akurasi dua metode yaitu algoritma C4.5 dengan Naive Bayes untuk prediksi harapan hidup pasien hepatitis menyimpulkan akurasi C.45 hanya 77,29% sedangkan akurasi Naive Bayes mencapai 83,71% [11].

III. METODOLOGI PENELITIAN

A. Akuisisi Data

Dataset yang digunakan adalah data pasien hepatitis yang diunduh dari *repository* UCI *learning*. *Dataset* berisi sejumlah atribut gejala medis beserta identifikasi apakah penderita hepatitis hidup (*live*) atau mati (*die*) jika memiliki gejala medis tersebut. Total data sebanyak 155 *record*. Atribut yang menunjukkan gejala sejumlah 19 dan 1 atribut kelas keputusan. Atribut kelas keputusan berisi nilai 1 untuk "die" dan 2 untuk "live". Keterangan atribut terdapat pada Tabel 1.

B. Analisis Data

Analisis data berguna untuk menentukan kebutuhan proses selanjutnya. Analisis ini bertujuan mengidentifikasi distribusi data, nilai atribut yang hilang (*missing value),* atribut yang digunakan dan yang tidak digunakan.

1) Atribut Dataset dan Domain Nilai

Atribut *dataset* hepatitis penelitian ini terdiri dari enam atribut numerik dan empat belas atribut binomial.

TABEL 3 ATRIBUT DATA

No. Atribut	Atribut	Domain Nilai	Keterangan
1.	Kelas/label keputusan	DIE, LIVE	Label yang menunjukkan pasien hidup/ mati karena gejala yang ditemukan
2.	Umur	Angka numerik	Umur pasien
3.	Jenis Kelamin	Laki-laki, perempuan	Jenis kelamin pasien
4.	STEROID	No, Yes	Apakah mendapatkan terapi steroid?
5.	ANTIVIRAL	No, Yes	Apakah mendapatkan terapi antiviral?
6.	FATIGUE	No, Yes	Apakah mengalami symstoms/gejala kelelahan akut?
7.	MALAISE	No, Yes	Apakah mengalami symstoms/ gejala malaise (rasa tidak nyaman)?
8.	ANOREXIA	No, Yes	Apakah mengalami symstoms/ gejala anorexia (muntah setiap maka)?
9.	LIVER BIG	No, Yes	Apakah kondisi hati/liver membesar?
10.	LIVER FIRM	No, Yes	Apakah kondisi hati/liver mengeras?
11.	SPLEEN PALPABLE	No, Yes	Apakah ada gejala spleen palpable/ limfa lebih jelas/besar dari normal?

No. Atribut	Atribut	Domain Nilai	Keterangan
12.	SPIDERS	No, Yes	Apakah ada gejala Spider/ pembuluh darah upnormal pada kulit (pembuluh darah mengumpul dan menonjol pada permukaan kulit)?
13.	ASCITES	No, Yes	Terjadi penumpukan cairan pada rongga perut?
14.	VARICES	No, Yes	Terjadi pembekakan vena esophagus (varises)?
15.	BILIRUBIN:	Angka numerik	Nilai kadar bilirubin dalam darah
16.	ALK PHOSPHATE	Angka numerik	Kadar Alkalin Phospate dalam liver
17.	SGOT	Angka numerik	Nilai SGOT
18.	ALBUMIN	Angka numerik	Kadar Albumin
19.	PROTIME	Angka numerik	Uji Masa protrombhine
20.	HISTOLOGY	No, Yes	Apakah dilakukan pemeriksaan dengan histology (biopsy hati)?

2) Distribusi Atribut Kelas Keputusan

TABEL 4 DISTRIBUSI LABEL KEPUTUSAN

No	Kelas	Jumlah
1.	Die	32
2.	Live	123

3) Atribut dengan Missing Value

Atribut dengan *missing value* diindikasikan oleh nilai "?". Distribusi frekuensi digunakan untuk mengidentifikasi jumlah *missing value* setiap atribut ditunjukkan oleh Tabel 3.

TABEL 5 MISSING VALUE

No. Atribut	Atribut	Jumlah Missing Value
1.	KELAS/LABEL KEPUTUSAN	0
2.	UMUR	0
3.	JENIS KELAMIN	0
4.	STEROID	1
5.	ANTIVIRAL	0
6.	FATIGUE	1
7.	MALAISE	1
8.	ANOREXIA	1
9.	LIVER BIG	10
10.	LIVER FIRM	11
11.	SPLEEN PALPABLE	5
12.	SPIDERS	5
13.	ASCITES	5
14.	VARICES	5
15.	BILIRUBIN:	6
16.	ALK PHOSPHATE	29
17.	SGOT	4
18.	ALBUMIN	16
19.	PROTIME	67
20.	HISTOLOGY	0

4) Atribut yang Digunakan

Sembilan belas atribut gejala digunakan untuk proses klasifikasi meskipun ada atribut yang nilai *missing value*-nya tinggi. Misalnya atribut PROTIME.

C. Pre-Prosessing

1) Data Cleanning

Data *missing value* adalah data atribut yang nilainya "?". Untuk mengatasi data *missing value* tersebut, setiap data atribut yang bernilai "?" diubah menjadi 0.

2) Transformasi Data Bertipe Binomial ke Tipe Numerik.

TABEL 6 KONVERSI DATA

No. Atrbut	Atribut	Domain Nilai	Transformasi Nilai
1.	Kelas/label keputisan	DIE, LIVE	Die=1, Live=2
2.	Jenis Kelamin	Laki-laki, perempuan	Laki-Laki =1, Perempuan=2
3.	STEROID	No, Yes	No=1, Yes =2
4.	ANTIVIRAL	No, Yes	No=1, Yes =2
5.	FATIGUE	No, Yes	No=1, Yes =2
6.	MALAISE	No, Yes	No=1, Yes =2
7.	ANOREXIA	No, Yes	No=1, Yes =2
8.	LIVER BIG	No, Yes	No=1, Yes =2
9.	LIVER FIRM	No, Yes	No=1, Yes =2
10.	SPLEEN PALPABLE	No, Yes	No=1, Yes =2
11.	SPIDERS	No, Yes	No=1, Yes =2
12.	ASCITES	No, Yes	No=1, Yes =2
13.	VARICES	No, Yes	No=1, Yes =2
14.	HISTOLOGY	No, Yes	No=1, Yes =2

D. Metode Klasifikasi

1) K-Nearest Neighbor

Metode ini mencari kesaman kasus dengan menghitung kedekatan antara kasus baru dengan kasus lama, yaitu berdasarkan pada pencocokan bobot dari sejumlah fitur yang ada. Nilai similiaritas dihitung mengunakan persamaan (1).

$$similarity(T,S) = \frac{\sum_{i=1}^{n} f(T_i, Si) \times w_t}{w_t} \quad (1)$$

Keterangan:
T : Kasus baru
S : Kasus yang ada dalam penyimpanan
n : Jumlah atribut dalam setiap kasus
i : Atribut individu antara 1 sampai dengan n
f : fungsi similarity atribut i antara kasus T dan kasus S
w : bobot yang diberikan pada atribut ke-i

Kedekatan biasanya berada pada nilai 0 sampai dengan 1. Nilai 0 artinya kedua kasus mutlak tidak mirip, dan nilai 1 kasus mutlak mirip.

2) Naive Bayes Classifier

Klasifikasi Naive Bayes berdasarkan pada persamaan (2).

$$P(x|y) = P(y|x) \text{ x } P(x) / P(y) \quad (2)$$

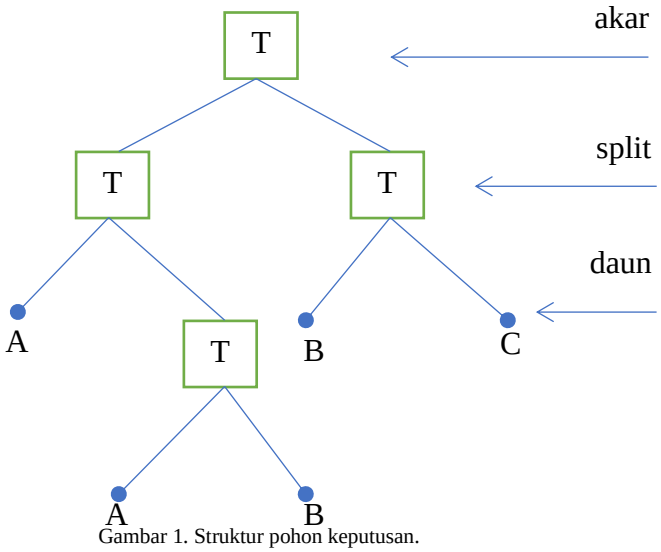
Keterangan:
Y = data dengan kelas yang belum diketahui
X = hipotesis data y merupakan suatu kelas spesifik
P(x|y) = probabilitas hipotesis x berdasarkan kondisi y
P(x) = probabilitas hipotesis x

P(y|x) = probabilitas y berdasarkan kondisi pada hipotesis x
P(y) = probabilitas dari y

3) Decision Trees (D-Tree)

D-Tree (pohon keputusan) adalah salah satu jenis algoritma data mining yang paling popular untuk klasifikasi dan prediksi. D-Tree menstrukturkan himpunan data menjadi struktur pohon yang terdiri dari simpul akar, cabang dan simpul daun. Simpul akar berada di bagian puncak struktur pohon. Simpul merepresentasikan atribut, cabang merepresentasikan hasil, dan daun merepresentasikan keputusan.

D-Tree didefinisikan sebagai langkah-langkah klasifikasi yang bekerja secara rekursif mempartisi suatu kumpulan data ke himpunan-himpunan data yang lebih kecil berdasarkan pengujian-pengujian di setiap cabang (atau simpul) pohon. Pohon tersebut tersusun dari satu simpul akar (yang dibentuk dari semua data), kemudian membentuk satu himpunan simpul internal (percabangan), dan kemudian membentuk satu himpunan simpul terminal (daun). Ilustrasi struktur pohon ada pada Gambar 1.



Gambar 1. Struktur pohon keputusan.

Pada Gambar 1, setiap kotak disebut sebagai simpul yang didalamnya terdapat proses T yang secara rekursif membagi data menjadi kelompok-kelompok data yang lebih kecil. Label A, B, dan C yang ada di setiap daun adalah label kelas yang ditetapkan untuk setiap satu observasi. Setiap simpul T dalam pohon keputusan hanya memiliki satu buah simpul induk dan dua atau lebih simpul anak [12].

4) Random Forest

(RF) adalah *classifier* dalam tipe pohon keputusan. RF muncul karena pohon yang dihasilkan D-Tree tidak fleksibel ketika digunakan mengklasifikasi data baru. Prinsip kerja RF adalah membuat banyak pohon klasifikasi dari *dataset*. Algoritma RF menerapkan *bootsrap aggregation (Bagging)* yang diperkenalkan oleh Breimans [13]. *Bagging* merupakan pembelajaran *ensemble* atau penggabungan beberapa algoritma *classifier* yang bertujuan untuk menghindari masalah varians yang tinggi, membuat pohon keputusan lebih stabil dan meningkatkan akurasi [13]. Langkah-langkah RF yaitu[14] :

a. Proses dimulai dari membuat *dataset bootstrap* dengan ukuran sama dengan *dataset* asli yang anggota *dataset*-nya diambil secara acak dari *dataset* asli. Satu data dapat dipilih acak lebih dari satu kali.
b. Pohon dibentuk dari *dataset bootstrap* namun hanya menggunakan subset variabel pada setiap langkahnya. Pembentukan pohon ini tidak menggunakan langkah *pruning* (pemangkasan).
c. Ulangi langkah a dan b sehingga terbentuk banyak pohon dari *bootstrap* atau $n_{tree.}$
d. Memprediksi data baru menggunakan pohon-pohon n_{tree} yang terbentuk. Hasil keputusan setiap pohon akan disimpan dan diakumulasi sesuai jenis labelnya. Keputusan akhir prediksi adalah jenis label dengan jumlah mayoritas.

Keluaran dari *classifier* diperoleh dari gabungan prediksi semua pohon untuk kombinasi keputusan.

E. Pengujian dan Validasi

1) Pengujian

Pengujian menggunakan validasi model *k-fold cross validation* dengan nilai k=10. Artinya pada saat uji model, dataset dibagi menjadi sepuluh bagian (partisi) sama besar. Nilai k juga menunjukkan jumlah pengulangan pengujian. Setiap pengulangan, satu

partisi data berperan sebagai data uji sedangkan 9 partisi lainnya sebagai data latih. Setiap iterasi, partisi yang menjadi data latih dan data testing berbeda-beda.

Kinerja model dilihat hasil akurasi dan kurva ROC-AUC (*Receiver Operating Characteristic-Area Under Curve*). Akurasi menggunakan *confusion matrix* pada persamaan (3). Komponen *confusion matrix* terdiri dari empat kondisi hasil prediksi yaitu:

- True Positives (TP) adalah hasil prediksi maupun data aktual menyatakan pasien hidup.
- False Positives (FP) adalah hasil prediksi menyatakan mati namun data aktual menyatakan pasien hidup.
- True Negatives (TN) adalah hasil prediksi dan data aktual menyatakan pasien mati.
- False Negatives (FN) adalah hasil prediksi menyatakan pasien hidup namun data aktual menyatakan pasien mati.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

2) Presisi

Presisi adalah nilai rata-rata TP (*TP rate*) yang terprediksi benar, dihitung mengunakan persamaan (4). Nilai presisi tersebut menunjukkan sensivitas model yang dibangun.

$$Presisi = \frac{jumlah\ TP}{jumlah\ \ TP+jumlah\ FN} \quad (4)$$

3) Recall

Recall menunjukkan *specivicity* model. *Recall* adalah perbandingan antara jumlah *record* yang relevan dengan jumlah total *record* dalam basisdata. Perhitungan *recall* mengunakan persamaan (5).

$$Recall = \frac{jumlah\ TN}{jumlah\ \ TN+jumlah\ FP} \quad (5)$$

4) ROC- AUC

Kurva ROC adalah kura yang memetakan nilai TP pada sumbu *y* dan FP pada sumbu *x*. Hasil AUC diklafikasikan berdasarkan kelas berikut [15]:

0.90 - 1.00 = *excellent classification*
0.80 - 0.90 = *good classification*
0.70 - 0.80 = *fair classification*
0.60 - 0.70 = *poor classification*
0.50 - 0.60 = *failure*

IV. PENGUJIAN DAN PEMBAHASAN

A. Pengujian

Pengujian menggunakan *k-fold cross validation* dengan k = 10. *Dataset* berjumlah 155 *record* dibagi menjadi sepuluh partisi secara acak. Sepuluh partisi tersebut dibagi menjadi sembilan partisi sebagai data latih dan satu partisi sebagai data uji. Data latih digunakan untuk membangun model sedangkan data uji untuk menguji model yang telah dibangun. Hasil pengujian dengan empat model yaitu K-NN, Naive Bayes, D-Tree, dan Random forest sebagai berikut:

1) K-Nearest Neigbor (K –NN)

Tabel 5 adalah hasil pengujian klasifikasi dengan K-NN .

TABEL 7 PENGUJIAN PERTAMA DENGAN K-NN

	True Positif "Live"	**True Negatif "Die"**	**Presisi**	**AUC**	**Akurasi(3)**
Prediksi "Live"	105	28	78.95%	0.7	70,31%
Prediksi "Die"	18	4	18.18%		
Kelas Recall	85.37%	12.50%			

2) Naive Bayes

Tabel 6 adalah hasil pengujian menggunakan metode Naive Bayes.

TABEL 8 PENGUJIAN DENGAN NAIVE BAYES

	True Positif "Live"	**True Negatif "Die"**	**Kelas Presisi**	**AUC**	**Akurasi(3)**
Prediksi "Live"	86	5	94.51%	0.84	72,90%
Prediksi "Die"	37	27	42.19%		
Kelas Recall	69.92%	84.38%			

3) Decission Tree (D-Tree)

Tabel 7 adalah hasil pengujian mengunakan D-Tree.

TABEL 9 PENGUJIAN DENGAN D-TREE

	True Positif "Live"	True Negatif "Die"	Kelas Presisi	AUC	Akurasi(3)
Prediksi "Live"	99	15	86.84%	0.81	74,84%
Prediksi "Die"	24	17	41.46%		
Kelas Recall	80.49%	53.12%			

4) Random Forest

Tabel 8 adalah hasil pengujian mengunakan *Random forest.*

TABEL 10 PENGUJIAN DENGAN RANDOM FOREST

	True Positif "Live"	True Negatif "Die"	Kelas Presisi	AUC	Akurasi(3)
Prediksi "Live"	117	26	86.82 %	0.81	79,35%
Prediksi "Die"	6	6	50 %		
Kelas Recall	95.12%	18.75%			

B. Analisis Hasil

Perbandingan akurasi dan AUC hasil pengujian dirangkum pada Tabel 9. Hasil pengujian menunjukkan urutan akurasi tertinggi adalah algoritma Random forest dengan akurasi 79.35%, disusul akurasi D-Tree sebesar 74.84%, Naive Bayes sebesar 72.90%, dan KNN sebesar 70.31%. Empat model yang diuji memiliki akurasi yang hampir sama. Selain akurasi, kinerja model ditunjukkan dengan nilai AUC yang semuanya mempunyai nilai lebih besar sama dengan 7. Nilai AUC K-NN sebesar 0.7 artinya model K-NN adalah *classifier* pada level *fair*. Sedangkan Naive Bayes, D-Tree dan Random forest nilai AUC-nya diatas 0.8 artinya termasuk *classifier* dengan level *good*. Akurasi metode Naive Bayes dengan penelitian sebelumnya [11] jauh berbeda hasilnya. Hal ini perlu diselidiki lebih dalam karena data yang digunakan adalah sama. Karena *pre-prosessing* pada penelitian [11] tidak dijelaskan maka bisa diduga perbedaan *pre-processing* akan menyebabkan akurasi model. Meskipun *missing value* pada atribut PROTIME cukup besar namun tetap dianggap sebagai atribut penentu dalam prediksi.

TABEL 11 MATRIK AKURASI KLASIFIKASI

Algoritma	Akurasi	AUC
K-NN	70.31	0.7
Naive Bayes	72.90	0.84
Decision Tree (D-Tree)	74.84	0.81
Random forest	79.35	0.81

V. KESIMPULAN DAN SARAN

Metode K-NN, Naive Bayes, D-Tree, dan Random forest dapat dipakai untuk membangun model prediksi harapan hidup penderita hepatitis kronik. Akurasi empat metode tersebut antara 70.31% sampai dengan 79.35 %. Random forest adalah metode yang paling tinggi akurasinya yaitu 79.35%. Akurasi model juga sangat bergantung pada kondisi data yang digunakan dan juga tahap *pre-processing.*

Penelitian berikutnya adalah bagaimana mencapai akurasi lebih dari 80% dengan memodifikasi metode- metode tersebut atau menggunakan metode lainnya. Pengujian lebih lanjut sebaiknya menggunakan data sampel yang lebih banyak, dan berupa data primer yang diperoleh dari klinik atau RS di Indonesia.

REFERENSI

[1] Pusdatin Kemenkes RI, "Infodatin." Pusdatin Kemenkes RI, Jakarta, 2014.

[2] S. GraciaJacob and R. Geetha Ramani, "Data Mining in Clinical Data Sets: A Review," *Int. J. Appl. Inf. Syst.*, vol. 4, no. 6, pp. 15–26, 2012.

[3] E. M. F. El Houby, "A Survey On Applying Machine Learning Techniques For Management Of Diseases," *J. Appl. Biomed.*, vol. 16, no. 3, pp. 165–174, 2018.

[4] M. Kurosaki *et al.*, "Data Mining Model Using Simple And Readily Available Factors Could Identify Patients At High Risk For Hepatocellular Carcinoma In Chronic Hepatitis C," *J. Hepatol.*, vol. 56, no. 3, pp. 602–608, 2012.

[5] S. O. Hussien, S. S. Elkhatem, N. Osman, and A. O. Ibrahim, "A Review of Data Mining Techniques for Diagnosing Hepatitis," in *Sudan Conference on Computer Science and Information Technology (SCCSIT) 2017*, 2017, vol. 101, no. 1, pp. 41–46.

[6] T. Karthikeyan and P. Thangaraju, "Analysis of Classification Algorithms Applied to Hepatitis Patients," *Int. J. Comput. Appl.*, vol. 62, no. january, pp. 25–

30, 2013.

[7] F. M. Ba-alwi and H. M. Hintaya, “Comparative Study for Analysis the Prognostic in Hepatitis Data: Data Mining Approach,” *Int. J. Sci. Eng. Res.*, vol. 4, no. 8, pp. 680–685, 2013.

[8] J. Weihan, Michelin Kamber, and J. Pei, “Data Mining:Concepts and Technicques.” Morgan Kauffman, 2011.

[9] K. S. Bhargav, T. D. Kumari, D. S. S. B. Toha, and V. B, “Application of Machine Learning Classification Algorithms on Hepatitis Dataset,” *Int. J. Appl. Eng. Res.*, vol. 13, no. 16, pp. 12732–12737, 2018.

[10]M. Neshat, M. Sargolzaei, A. N. Toosi, and A. Masoumi, “Hepatitis Disease Diagnosis Using Hybrid Case Based Reasoning and Particle Swarm Optimization,” *ISRN Artif. Intell.*, vol. 2012, 2012.

[11] W. D. Septiani, P. Studi, and M. Informatika, “Komparasi Metode Klasifikasi Data Mining Algoritma C4.5 Dan Naive Bayes Untuk Prediksi Penyakit Hepatitis,” *J. Pilar Nusa Mandiri*, vol. 13, no. 1, pp. 76–84, 2017.

[12] C. E. Brodley and M. A. Friedl, “Decision Tree Classification Of Land Cover From Remotely Sensed Data,” *Remote Sens. Environ.*, vol. 61, no. 3, pp. 399–409, 1997.

[13] A. T. Azar, H. I. Elshazly, A. E. Hassanien, and A. M. Elkorany, “A Random Forest Classifier for Lymph Diseases,” *Comput. Methods Programs Biomed.*, vol. 113, no. 2, pp. 465–473, 2013.

[14] A. Liaw and M. Wiener, “Classification and Regression by Random Forest,” *R News*, vol. 2, no. December, pp. 18–22, 2002.

[15] F. Gorunescu, “Data Mining: Concepts, Models and Techniques,” *Springer*. 2011.