PREDIKSI HARAPAN HIDUP PASIEN PENDERITA HEPATITIS B MENGGUNAKAN *CLASSIFICATION AND REGRESSION TREES*

Ahmad Saparudin, Ulfa Khaira, Tiya Maulidina, Muqsith Naufal Fikri, Doli Sumarlin, dan Irvan Wahyudi
Fakultas Sains dan *Teknologi, Universitas Jambi*
email: ahmadsafarudinsman2@gmail.com

ABSTRACT

Data mining is the two words that are familiar are heard by some people. In the world of business data mining is often utilized to support the company's business strategies or formulate. Not only in business, data mining is also utilized in various other areas of education, agriculture, animal husbandry, and medicine. This research was conducted in the classification of the patient sufferer hepatisis B use classification and refression trees (CART) to generate a decision tree that can be implemented to predict the life expectancy of the patient sufferer hepatitis B. After doing this research, produced a decision tree with 93.6% prediction accuracy.

Kata Kunci: *Prediksi, Harapan Hidup Pasien Hepatitis B, Klasifikasi, Klasifikasi, CART.*

1. PENDAHULUAN

Hepatitis B merupakan salah satu penyakit hati yang disebakan oleh VHB (Virus Hepatitis B). VHB merupakan virus yang menyebar tidaklah melalui makanan ataupun kontak biasa, melainkan menyebar melalui darah ataupun cairan tubuh dari penderita yang terinfeksi. Sebagai contoh adalah soerang bayi yang dapat terinfeksi dari ibunya selama proses kelahirannya. Bagi seorang tenaga medis umum memprediksi penyakit spesialis tidaklah mudah apalagi hingga harus memprediksi harapan hidup orang yang pengidap penyakit khusus tersebut. Oleh sebab itu diperlukan teknologi yang dapat memudahkan tenaga medis dalam memprediksi suatu penyakit spesialis (dalam hal ini adalah harapan hidup). Salah satu bidang kelimuan teknologi yang dapat digunakan untuk memprediksi penyakit spesialis adalah *data mining*.

Data mining merupakan suatu konsep yang digunakan untuk menemukan pengetahuan yang tersembunyi di dalam database. Data mining merupakan proses semi otomatik yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi pengetahuan potensial dan berguna yang tersimpan di dalam database besar (Turban, 2005). Data mining adalah bagian dari proses KDD (*Knowledge Discovery in Databases*) yang terdiri dari beberapa tahapan seperti pemilihan data, pra pengolahan, transformasi, data mining, dan evaluasi hasil. Ditinjau dalam biding bisnis, saat ini data mining telah dimanfaatkan oleh banyak perusahaan sebagai keunggulan kompetitif. Penggunaan data mining khususnya *task*nya sebagai prediksi dalam sebuah organisasi (atau perusahaan) umumnya digunakan untuk strategi differensiensinya dengan perusahaan lain. Tidak hanya dalam bidang bisnis, data mining saat ini juga telah banyak diimplementasikan di dibang kesehatan untuk memprediksi berbagai macam penyakit.

Berdasarkan uraian diataslah dilandasinya penelitian ini guna memprediksi harapan hidup pasien penderita hepatitis B menggunakan classification *and regression trees* (metode *data mining*). Dimana pohon keputusan yang dihasilkan nantinya akan diimplementasikan kedalam aplikasi berbasis *desktop* guna memudahkan pengguna (khalayak umum) untuk mengetahui hasil dari pemprediksian pasien hepatitis B.

2. METODE PENELITIAN

Dalam penelitian ini dataset yang digunakan adalah data pasien pengidap penyakit hepatitis B yang diperoleh melalui organisasi pengedia dataset global (*https://datahub.io/machine-learning)*. Dalam dataset tersebut terdapat 19 atribut yang dijadikan dasar dalam memprediksi harapan hidup pasien penderita hepatitis B. Atribut yang dimaksudkan tersebut meliputi*: age, sex, steroid, antivirals, fatigue, malaise, anorexia, liver big, liver firm, spleen palpable, spiders, ascites, varices, bilirubin, alk phosphate, sgot, albumin, protime*, dan *histology*.

Pengolahan data dilakukan dengan cara sebagai berikut:

1) Seleksi Data

Pada proses ini dilakukan kegiatan meminimalkan jumlah data yang digunakan untuk proses mining dengan tetap

merepresentasikan data aslinya. Dalam proses ini data yang dieliminasi adalah *record-record* yang memiliki banyak *missing value* disebagian besar atributnya. Sebagai contoh adalah *record* yang memiliki *null value* (*NaN*) di 11 dari 19 atribut yang ada.

2) Pengisian *Missing Value*

Pengisian *missing value* dilakukan terhadap *record-record* yang lulus dalam penyaringan tahap pertama (seleksi data). Dalam pengisian *missing value* ini terdapat perlakuan khusus terhadap beberapa atribut/variabel. Perlakuan khusus yang dimaksudkan yakni pengimplementasian sistem pengisian *missing value* dengan prinsip *modus* (nilai yang sering muncul) terhadap atribut *steroid, fatigue, malaise, anorexia, liver_big, liver_firm, spleen_palpable, spiders, ascites, varices,* dan *histology*. Serta pengimplementasian prinsip *mean* (nilai rata-rata) terhadap atribut *bilirubin, alk_phosphate, sgot, albumin,* dan *protime*.

3) Inisialisasi *Target* dan Transformasi

Inisialisasi target yang dimaksudkan disini merupakan penginisialisasian target klasifikasi yang akan dilakukan. Karena tujuan akhir dalam penelitian ini adalah untuk memprediksi harapan hidup pasien penderita hepatitis B maka terdapat dua jenis target, yakni: *live* dan *die.* Penelitian ini dilakukan dengan mentransformasi *live* menjadi 1 dan *die* menjadi 0, serta *gender/sex male* menjadi 1 dan *female* menjadi 0. Dalam hal ini transformasi bertujuan untuk mempermudahkan pembentukan *decision tree* nantinya.

4) Proses *Mining*

Dalam proses peminingan ini dilakukan pembentukan pohon keputusan menggunakan algoritma atau metode *classification and regression trees (CART)* dengan skema pembentukan 70% dari data digunakan sebagai *training set* dan 30% dari data digunakan sebagai *testing set.* Pembentukan pohon keputusan dilakukan dengan cara:

a. Pembentukan pohon klasifikasi

Tahap pertama membentuk pohon klasifikasi digunakan *sampel data Learning* (L) yang masih bersifat heterogen. Setiap pemilahan hanya bergantung pada nilai yang berasal dari suatu variabel independen. Rumus kemungkinan pemilah yaitu jika variabel prediktor kontinu = n – 1 pemilahan, jika variabel prediktor kategori nominal = $2^{L-1} - 1$ pemilahan, dan jika variabel prediktor kategori ordinal = L – 1 pemilahan. Sampel tersebut akan dipilah berdasarkan aturan pemilahan dan kriteria *goodness-of-split*. Untuk mengukur tingkat keheterogenan suatu kelas dari suatu node tertentu dalam pohon klasifikasi dikenal dengan istilah impurity measure i (t). Ukuran ini akan membantu dalam menemukan fungsi pemilah yang optimal (Siahaan, 2016). Fungsi keheterogenan yang digunakan adalah indeks Gini yakni:

$$i(t) = \sum_{j \neq i} p\,(j|t)p(i|t) \qquad (1)$$

$p(j|t)$ adalah peluang j pada *node* t. *Goodness of split* merupakan suatu evaluasi pemilahan oleh pemilah s pada *node* t. *Goodness of split* $\Phi(s,t)$ didefinisikan sebagai penurunan keheterogenan. Kualitas ukuran dari seberapa baik pemilah s dalam menyaring data menurut kelas merupakan ukuran penurunan keheterogenan dari suatu kelas dan didefinisikan sebagai

$$\Phi(s,t) = \Delta i(s,t) \qquad (2)$$

$$\Phi(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \qquad (3)$$

b. Penentuan *terminal node*

Tahap kedua adalah penentuan terminal node. Suatu node t akan menjadi *terminal node* atau tidak, akan dipilah kembali bila pada node t tidak terdapat penurunan keheterogenan secara berarti atau adanya batasan minimum n seperti hanya terdapat satu pengamataan pada tiap node anak. Umumnya jumlah kasus minimum dalam suatu terminal akhir adalah 5, dan apabila hal itu terpenuhi maka pengembangan pohon dihentikan (Lewis, 2000).

c. Penandaan Label Kelas

Tahap ketiga yaitu penandaan label kelas. Penandaan label kelas pada *terminal node* dilakukan berdasarkan aturan jumlah terbanyak, yaitu:

$$p(j_0|t) = max_j p(j|t) = max_j \frac{N_j(t)}{N(t)} \qquad (4)$$

dimana $p(j|t)$ adalah proporsi kelas j pada *node* t, $N_j(t)$ adalah jumlah pengamatan kelas j pada *node* t dan $N(t)$ adalah jumlah pengamatan pada *node* t. Label kelas terminal *node* t adalah j_0 yang memberi nilai dugaan kesalahan pengklasifikasian *node* t terbesar. Proses pembentukan pohon klasifikasi berhenti saat terdapat hanya satu pengamatan dalam tiap-tiap node anak atau adanya batasan minimum n, semua pengamatan dalam tiap node anak identik, dan adanya batasan jumlah level/kedalaman pohon maksimal (Siahaan, 2016)

d. Pemangkasan Pohon Klasifikasi

Setelah terbentuk pohon maksimal, tahap selanjutnya adalah *pemangkasan* pohon untuk mencegah terbentuknya pohon klasifikasi yang berukuran besar dan kompleks. Pemangkasan (*pruning*) yaitu suatu penilaian ukuran pohon

tanpa mengorbankan ketepatan melalui pengurangan node pohon sehingga mencapai ukuran pohon yang layak. Ukuran pemangkasan yang digunakan untuk memperoleh ukuran pohon yang layak adalah *Cost Complexity Minimum* (Breiman, 1993). Ukuran *complexity* adalah sebagai berikut:

$$R_\alpha(t) = R(T) + \alpha|\tilde{T} \quad (5)$$

dimana, $R(T)$ adalah *resubtution estimate* (proporsi kesalahan pada sub pohon), α adalah kompleksitas parameter (*complexity parameter*) dan $|\tilde{T}|$ adalah ukuran banyaknya node terminal pohon T.

Menurut Siahaan (2016), *cost complexity prunning* menentukan suatu pohon bagian $T(\alpha)$ yang meminimumkan $R_\alpha(t)$ pada seluruh pohon bagian, atau untuk setiap nilai α, dicari pohon bagian $T(\alpha) < T\ max$ yang meminimumkan $R_\alpha(t)$ yaitu:

$$R_\alpha(T(\alpha)) = min_{T<T_{max}} R\alpha(T) \quad (6)$$

e. Penentuan Pohon Klasifikasi Optimal

Setelah dilakukan pemangkasan diperoleh pohon klasifikasi optimal yang berukuran sederhana namun memberikan nilai pengganti yang cukup kecil. Penduga pengganti yang sering digunakan adalah *validate* silang lipat V (*Cross Validation V-Fold Estimates*). Penduga validasi silang lipat V sering digunakan apabila amatan yang tidak cukup besar. Amatan dalam L dibagi secara aca menjadi bagian V bagian yang saling lepas dengan ukuran kurang lebih sama besar untuk setiap kelasnya. Pohon $T^{(V)}$ dibentuk dari $L - L_v$ dengan v = 1, 2, …, V. Misalkan $d^{(V)}(x)$ adalah hasil pengklasifikasian, penduga sampel uji untuk $R(T_1^{(v)})$ yaitu:

$$R^{ts}(T_t^{(v)}) = \frac{1}{Nv} \sum_{(x_n j_n)\epsilon L_v} X(d^{(V)}(x_n) \neq j_n) \quad (7)$$

dengan $Nv = {}^{N}/_{V}$ adalah jumlah amatan dalam L_v. Kemudian dilakukan prosedur yang sama menggunakan seluruh L, maka penduga validasi silang lipat V untuk $(T_t^{(v)})$ adalah

$$R^{Cv}(T_t) = \frac{1}{V} \sum_{v=1}^{v} R^{ts}\left(T^{(v)}\right) \quad (8)$$

5) Evaluasi Klasifikasi/Pohon Keputusan

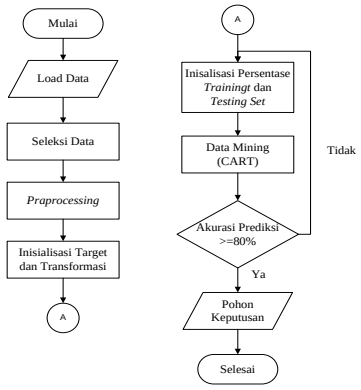
Dalam tahapan ini dilakukan evaluasi klasifikasi yang dihasilkan dari tahapan sebelumnya. Evaluasi ini dilakukan dengan cara melihat nilai keakuratan prediksi. Apabila tingkat keakuratan prediksi mencapai ≥ 80% maka akan dilanjutkan ke tahapan berikutnya. Namun apabila tingkat keakuratan prediksi adalah < 80% maka dilakukan pemrosesan *mining* ulang (kembali ke tahapan sebelumnya) dengan perlakuakn mengubah persentase *training set* dan *testing set* hingga tingkat akurasi klasifikasi yang dihasilkan sesuai dengan harapan (≥ 80%).

6) *Presentasi* Pengetahuan

Dalam beberapa literatur presentasi pengetahuan merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna. Tahap terakhir dari proses data mining adalah bagaimana memformulasikan keputusan atau aksi dari hasil analisis yang didapat. Ada kalanya hal ini harus melibatkan orang-orang yang tidak memahami data mining. Karenanya presentasi hasil data mining dalam bentuk pengetahuan yang bisa dipahami semua orang adalah satu tahapan yang diperlukan dalam proses data mining.

Dalam tahapan presentasi pengetahuan ini dilakukan pemvisualisasian hasil akhir (klasifikasi) kedalam bentuk aplikasi berbasis desktop (GUI) guna memudahkan pengguna dalam memahami hasil *mining* ataupun untuk memprediksi harapan hidup pasien pengidap penyakit hepatitis B.

Tahapan diatas dapat divisualisaikan kedalam bentuk *flowchart* berikut:



Gambar 1 Flowchart Tahapan Mining

3. HASIL DAN PEMBAHASAN

Setelah dilakukannya penelitian dan/atau percobaan didapatkan hasil sebagai berikut:

1) Data Sampel

No	Age	Sex	Steroid	…	Class
1	30	Male	False	…	Live
2	50	Female	False	…	Live
3	78	Female	True	…	Live
…	…	…	…	…	…

Tabel 1 Data Sampel

Attribute	*Sum of Missing Value*
age	0
sex	0

steroid	1
antivirals	0
fatigue	1
malaise	1
anorexia	1
liver big	10
liver firm	11
spleen palpable	5
spiders	5
ascites	5
varices	5
bilirubin	6
alk phosphate	29
sgot	4
albumin	16
protime	67
histology	0
class	0

Tabel 2 Jumlah Missing Value tiap Atribut

2) Data Hasil Pengisian *Missing Value* dan Transformasi

No	Age	Sex	Steroid	…	Class
3	31	0	True	…	1
97	44	0	False	…	1
130	54	0	True	…	1
…	…	…	…	…	…

Tabel 3 Data Sampel setelah Praprocessing

Attribute	*Sum of Missing Value*
age	0
sex	0
steroid	0
antivirals	0
fatigue	0
malaise	0
anorexia	0
liver big	0
liver firm	0
spleen palpable	0
spiders	0
ascites	0
varices	0
bilirubin	0
alk phosphate	0
sgot	0
albumin	0
protime	0
histology	0
class	0

Tabel 4 Jumlah Missing Value setelah Praprocessing

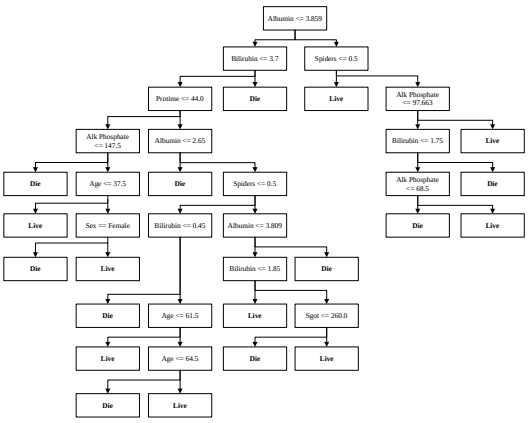
	Age	Bilirubin	Alk Phosphate
Count	155.000	155.000	155.000
Mean	41.200	1.428	105.325
STD	12.566	1.188	46.406
Min	7.000	0.300	26.000
25%	32.000	0.800	78.000
50%	39.000	1.000	102.000
75%	50.000	1.500	119.500
Max	78.000	8.000	295.000

Tabel 5 Ringkasan Variabel Data Sampel

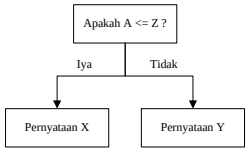
	Sgot	Albumin	Protime
Count	155.000	155.000	155.000
Mean	85.894	3.817	61.852
STD	88.479	0.617	17.193
Min	14.000	2.100	0.000
25%	32.500	3.500	57.000
50%	59.000	3.900	61.852
75%	99.000	4.200	65.000
Max	648.000	6.400	100.000

Tabel 6 Ringkasan Variabel Data Sampel

3) Pohon Keputusan (*Decision Tree*)



Gambar 2 Pohon Keputusan Klasifikasi Harapan Hidup Pasien Hepatitis B



Gambar 3 Format Pohon Keputusan

4) Keakurasian Prediksi

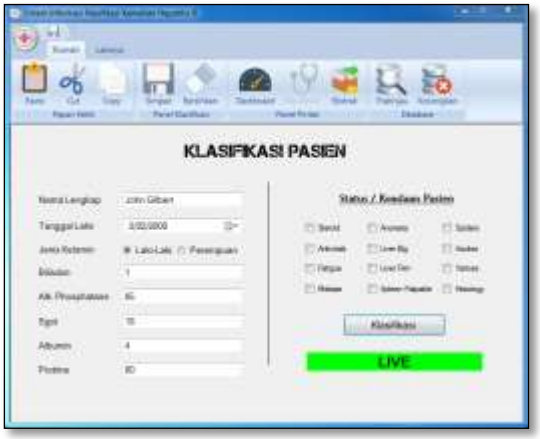
Pengujian keakurasian prediksi yang dihasilkan dalam penelitian ini dilakukan dengan menggunakan pengkodean berikut:

```
import sklearn.metrics
prediksi = dtc.predict(X_test)
akurasi =
sklearn.metrics.accuracy_score(y_test,prediksi)
```

Setelah dilakukannya pengujian didapatkan hasil keakuratan prediksi: 0.93617 (skala 0-1) atau sama halnya dengan 93.6 %.

5) Aplikasi Prediksi Harapan Hidup Pasien Penderita Hepatitis B

Karena keakurasian prediksi harapan hidup yang diperoleh diatas ambang batas (≥ 80%), maka *rule-rule* dari decision tree yang dihasilkan diimplementasikan kedalam aplikasi *desktop*. Adapun antar muka dari aplikasi desktop yang dihasilkan yakni sebagai berikut:



Gambar 4 Antar Muka Aplikasi Prediksi Harapan Hidup Pasien Hepatitis B

Berdasarkan penelitian/percobaan yang telah dilakukan dapat diutarakan bahwa penggunaan algoritma *classification and regression trees* mampu untuk memprediksi harapan hidup pasien penderita hepatitis B dengan tingkat keakurasian yang tinggi (≥ 80%). Dalam penelitian ini apabila ditelaah lebih lanjut pemprediksian harapan hidup pasien penderita hepatitis B hanya memerlukan variabel *age*, *gender/sex*, *bilirubin*, *alk phosphate*, *sgot*, *albumin*, *protime*, dan *spiders* (*Rule* yang terbentuk terlampir). Sedangkan *steroid*, *antivirals*, *fatigue*, *malaise*, *anorexia*, *liver big*, *liver firm*, dan *spleen palpable* tidak ikut berpartisipasi dalam memprediksi (berdasarkan pohon keputusan yang terbentuk).

Secara perhitungan (algoritma CART) variabel-variabel seperti *steroid*, *antivirals*, dan sebagainya tidak masuk kedalam pohon keputusan karena tidak adanya *record* yang dipengaruhi secara signifikan oleh variabel-variabel tersebut. Sedangkan secara realitanya variabel-variabel seperti *liver big* dan *liver firm* sangat berperan dalam memprediksi harapan hidup pasien penderita hepatitis B. Walaupun demikian, pohon keputusan yang terbentuk dalam penelitian ini juga tidak melanggar aturan medis tersebut. Sebagai contoh nyata adalah *liver* yang berperan dalam memproduksi *bilirubin.* Sehingga apabila *liver* mengalami masalah, maka *bilirubin* yang dihasilkan tidaklah sama ketika *liver* tersebut dalam kedaan normal. Dengan kata lain dengan diketahuinya *bilirubin* seseorang, maka hal tersebut mampu mencerminkan keadaan *liver* dari orang tersebut.

Dalam visualisasi akhir (aplikasi) yang dikembangkan, inputan variabel *age* (umur) dimanipulasi kedalam bentuk tanggal lahir. Hal ini dilakukan karena variabel *age* berperan dalam memprediksi harapan hidup pasien penderita hepatitis B seiring berjalannya waktu. Sebagai antisipasi perubahan *logic* dari pohon keputusan yang terbentuk, maka mengubah teknik inputan variabel *age* menjadi tanggal lahir merupakan solusi yang terbaik. Dalam kasus ini, sebagai contoh adalah apabila seorang perempuan yang lahir pada tanggal 16 November 1981 dengan *albumin*: 3.859, *bilirubin*: 3.7, *protime*: 44.0, dan *alk phosphate*: 147.5. Apabila orang tersebut diprediksi pada tanggal 16 November 2018 maka orang tersebut terklasifikasi kedalam *class live.* Sedangkan apabila orang tersebut diprediksi pada tanggal 16 November 2019 maka orang tersebut terklasifikasi kedalam *class die.* Dengan kata lain apabila sistem mengira perempuan tersebut berumur ≤ 37.5 tahun maka orang tersebut terklasifikasi kedalam *class live.* Sedangkan apabila umurnya > 37.5 tahun, maka orang tersebut terklasifikasi kedalam *class live.*

Dalam kasus diatas apabila jenis penginputan variabel *age* berupa *single textbox* maka prediksi yang dihasilkan tidaklah dinamis terhadap waktu. Oleh sebab itu dalam aplikasi/sistem yang dikembangkan *value* dari variabel *age* diperoleh melalui tanggal lahir dari orang tersebut.

4. KESIMPULAN

Dari penelitian yang telah dilakukan, dapat disimpulkan bahwa algoritma *classification and regression trees* (CART) mampu menghasilkan pohon keputusan dengan tingkat akurasi yang tinggi (dalam penelitian ini yakni 93.6 %). Pohon keputusan yang dihasilkan dalam

penelitian ini dapat digunakan atau diimplementasikan sebagai algoritma dasar dalam pengembangan perangkat lunak prediksi harapan hidup pasien penderita hepatitis B. Dalam metode *classification and regression trees, records training set* berbanding lurus dengan tingkat akurasi prediksi yang dihasilkan.

5. REFERENSI

Breiman, L., Friedman, J.H., Olsen, R.A., dan Stone, C.J. 1993. *Classification and Regression Trees*. New York: Chapman & Hall.

Siahaan, David, dkk. 2016. *The Application of Classification and Regression Tree (CART) and Ordinal Logistic Regression in Education (Case Study: Predicate of Bachelor Degree's Graduation at Faculty of Mathematics and Natural Sciences)*. *Jurnal Eksponensial*. Vol. 7. No. 1.

Turban, E., dkk. 2005. *Decision Support System and Intelligent System - 7th ed*. London: Pearson Education, Inc.