ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 3, Nomor 4, Tahun 2014, Halaman 731 - 739
Online di: http://ejournal-s1.undip.ac.id/index.php/gaussian

PEMBENTUKAN POHON KLASIFIKASI BINER DENGAN ALGORITMA QUEST (*QUICK, UNBIASED, AND EFFICIENT STATISTICAL TREE*) PADA DATA PASIEN LIVER

Muhammad Rosyid Abdurrahman[1], Dwi Ispriyanti[2], Alan Prahutama[3]
[1]Mahasiswa Jurusan Statistika FSM Universitas Diponegoro
[2,3]Staf Pengajar Jurusan Statistika FSM Universitas Diponegoro

ABSTRACT

In this modern era of fast food commonly found that sometimes have chemical substances and the increasing number of motor vehicles that cause the uncontrolled circulation of air pollution that can affect the health of the human liver. To assist in analyzing the presence of liver disorders in humans can be used QUEST (Quick, Unbiased, and Efficient Statistical Tree) algorithm to classify the characteristics of the patient's liver by liver function tests performed in clinical laboratories. QUEST construct rules to predict the class of an object from the values of predictor variables. The tree is constructed by partitioning the data by recuresively, where class and the values of the predictor variables of each observation in the data sample is known. Each partition is represented by a node in the tree. QUEST is one of the binary classification tree method. The results of the classification tree is formed, an important variable in classifying a person affected by liver disease or not, that is the variable Direct Bilirubin, Alkaline Phosphatase, Serum Glutamic Oxaloacetic Transaminase (SGOT), and age of the patient. Accuracy of the QUEST algorithm classifying liver patient data by 73,4 %.

Keywords: binary classification trees, QUEST algorithm, liver patient data.

1. PENDAHULUAN

Hati merupakan salah satu organ terbesar dalam tubuh manusia yang mempunyai banyak fungsi bagi tubuh. Organ ini mempunyai fungsi yang kompleks, sehingga mudah terpengaruh gangguan penyakit. Oleh karena memiliki fungsi yang kompleks, tidak mudah dalam mendiagnosis penyakit liver (gangguan fungsi hati).

Dalam mendiagnosis ada atau tidak penyakit liver dapat digunakan acuan dari hasil tes fungsi hati yang dilaksanakan di laboratorium. Tes tersebut antara lain yaitu *transaminase serum*, fosfatase alkali, total bilirubin, bilirubin terkonjugasi, total protein, albumin, serta rasio albumin dan globulin. Dari hasil tes tersebut dapat dilihat hasil tes yang signifikan sebagai ciri adanya gangguan fungsi hati dengan menggunakan algoritma pohon klasifikasi karena dapat memperoleh informasi mengenai data klasifikasi pasien penyakit liver.

Menurut Rokach dan Maimon (2008) dan Maroco *et al* (2011), algoritma pohon klasifikasi merupakan pendekatan nonparametrik. Algoritma QUEST merupakan algoritma pohon klasifikasi yang menghasilkan variabel tak bias dan memiliki dua simpul setiap penyekatan. QUEST dapat diterapkan pada data dengan variabel respon dua kategori berupa data nominal dan variabel prediktor dengan varibel berbentuk kategorik maupun kontinu. Pembentukan pohon klasifikasi dengan QUEST melalui dua langkah, yakni pemilihan variabel penyekat dan menentukan titik sekat. Pemilihan variabel penyekat digunakan uji ANOVA F, Levene F, dan chi-kuadrat Pearson dan dipilih variabel yang signifikan dengan *p-value* terkecil. Pemilihan variabel penyekat digunakan untuk menentukan titik sekat, yakni suatu nilai yang dapat mempartisi atau membagi data ke dalam dua simpul yang berbeda.

2. TINJAUAN PUSTAKA

2.1 Tinjauan Umum Hati

Hati yang merupakan alat tubuh yang paling besar adalah pusat dari metabolisme tubuh. Dalam hati terjadi proses-proses sintesa, modifikasi, penyimpanan, pemecahan serta ekskresi dari berbagai macam zat yang dibutuhkan untuk hidup. Fungsi dari hati adalah banyak dan beranekaragam serta sangat rumit (Soemohardjo *et al*, 1983).

Penyakit liver (gangguan pada hati) dapat berasal dari pola konsumsi makanan yang salah atau zat-zat kimia yang terkandung dalam obat, seperti antibiotik, parasetamol, dan makanan yang dikonsumsi manusia dalam bentuk hidangan cepat saji.

Tes fungsi hati merupakan suatu kumpulan analisis laboratorium yang berkaitan dengan hati. Tes-tes yang diperiksa untuk tujuan ini meliputi tes-tes yang secara rutin diperiksa untuk mengadakan penyaringan, apakah ada penyakit liver atau tidak.

2.2 Data Mining

Menurut Berry dan Linoff (2004), data mining adalah eksplorasi dan analisis data dalam jumlah besar untuk menemukan aturan yang berarti. Menurut Larose (2005), data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yakni deskripsi, prediksi, klasifikasi, *clustering*, dan asosiasi.

2.2.1 Pohon Keputusan

Pohon keputusan dalam data mining adalah model prediksi yang digunakan untuk menggambarkan bentuk pengklasifikasian dan model regresi (Rokach dan Maimon, 2008). Pohon keputusan dibagi menjadi dua, yaitu pohon regresi dan pohon klasifikasi. Pada pohon klasifikasi variabel respon berupa data kategorik, sedangkan pada pohon regresi variabel respon berbentuk data kontinu.

Pohon klasifikasi adalah aturan untuk memprediksi kelas dari sebuah objek dari nilai-nilai variabel prediktor. Pohon dibentuk melalui penyekatan data secara berulang, di mana kelas dan nilai-nilai variabel prediktor setiap amatan pada data contoh sudah diketahui. Masing-masing sekatan (*split*) data dinyatakan sebagai simpul (*node*) dalam pohon yang terbentuk (Loh dan Shih, 1997).

Menurut Maroco *et al* (2011), pohon klasifikasi adalah pengklasifikasi nonparametrik yang membangun pohon keputusan hirarki dengan membelah data antar kelas kriteria pada langkah yang diberikan seperti sebuah aturan "*if-then*" yang diterapkan pada variabel prediktor, menjadi dua *child node* (anak simpul) atau lebih secara berulang, dari simpul akar yang berisi seluruh sampel.

Pohon klasifikasi terdiri dari beberapa simpul. Dalam penelitian ini, t menyatakan *node* atau simpul. Simpul akar (t_0) merupakan tempat informasi keseluruhan data yang akan di pecah untuk pertama kali. Simpul dalam yaitu simpul dari hasil pemecahan simpul sebelumnya, namun masih dapat di pecah kembali ke beberapa simpul. Simpul akhir yaitu simpul yang tidak dapat dipecah lagi.

2.2.1.1 Algoritma QUEST

Algoritma QUEST (*Quick, Unbiased, and Efficient Statistical Tree*) pertama kali dikenalkan tahun 1997 oleh Loh dan Shih. QUEST termasuk ke dalam pohon klasifikasi biner, yaitu pohon yang menghasilkan dua simpul setiap sekatnya. Pohon klasifikasi QUEST dibentuk dari pemilihan variabel penyekat, pemilihan titik sekat untuk variabel yang dipilih, dan proses pemberhentian.

Pemilihan variabel penyekat digunakan untuk menentukan simpul yang akan dibentuk. Variabel penyekat dapat dipilih dari variabel prediktor kontinu maupun

kategorik. Variabel prediktor kontinu digunakan uji ANOVA F untuk menguji perbedaan rata-rata antar kelas pada variabel prediktor kontinu X atau digunakan uji Levene F untuk untuk menguji kesamaan varian antar kelas. Sedangkan untuk variabel kategorik digunakan uji chi-kuadrat Pearson (χ^2) untuk menguji kebebasan antar variabel respon Y dan variabel prediktor kategorik X (Loh dan Shih, 1997).

Menurut Loh dan Shih (1997), pemilihan variabel penyekat pada QUEST adalah memilih nilai *p-value* terkecil dari hasil uji setiap variabel prediktor terhadap variabel respon dan dibandingkan dengan koreksi Bonferroni. Menurut Grabczewski (2014), koreksi Bonferroni diusulkan untuk menggunakan tingkat kepercayaan lebih kecil ketika beberapa uji dilakukan. Ketika n uji dilakukan maka masing-masing uji menggunakan tingkat kepercayaan $\frac{\alpha}{n}$, dengan α adalah tingkat signifikansi.

Berikut adalah langkah-langkah pemilihan variabel penyekat:

1. Setiap variabel prediktor X yang terdapat dalam masing-masing simpul dilakukan
   a. Uji ANOVA F, untuk menguji perbedaan rata-rata antar kelas pada setiap variabel prediktor kontinu X.
   b. Uji χ^2, untuk menguji kebebasan antar variabel respon Y dan variabel prediktor kategorik X.
      Hipotesis:
   c. Dari uji ANOVA F dan χ^2 akan didapatkan *p-value* dari analisis menggunakan SPSS.
2. Variabel dengan *p-value* terkecil dipilih dan dilambangkan dengan $X*$.
3. Jika *p-value* terkecil kurang dari $\frac{\alpha}{M}$, maka variabel prediktor $X*$ dipilih sebagai variabel penyekat untuk simpul. Jika tidak, dilanjutkan ke tahap 4.
4. Jika *p-value* terkecil yang didapat $\geq \frac{\alpha}{M}$, maka:
   a. Untuk setiap variabel prediktor kontinu X, dilakukan uji Levene F untuk menguji kesamaan varian.
   b. Didapatkan *p-value* dengan bantuan SPSS dari setiap variabel yang diuji menggunakan uji Levene F.
   c. Variabel prediktor dengan *p-value* terkecil dipilih dan dinotasikan dengan $X**$.
   d. Jika *p-value* terkecil kurang dari nilai $\frac{\alpha}{M+M_1}$, maka $X**$ dipilih menjadi variabel penyekat untuk simpul. Jika tidak, maka penyekatan simpul berhenti.

2.2.1.1.2 Algoritma Menentukan Titik Sekat

Proses penentuan titik sekat pada QUEST dibedakan menjadi dua, yaitu proses penentuan titik sekat dengan variabel prediktor kontinu dan proses penentuan titik sekat dengan variabel prediktor kategorik.

a. Menentukan Titik Sekat dengan Variabel Prediktor Kontinu

Misalkan variabel respon adalah Y memiliki dua kelas, yaitu $j = A, B$. Misalkan pula variabel X dipilih untuk menyekat simpul t, maka langkah berikutnya adalah menentukan titik sekat.

1. Misalkan $\bar{x}_A$ dan s_A^2 menyatakan rata-rata dan varian sampel kelas A, sedangkan $\bar{x}_B$ dan s_B^2 adalah rata-rata dan varian sampel kelas B.
2. Untuk menentukan titik sekat yang merupakan suatu angka yang dapat membagi atau mempartisi suatu data ke dalam dua data baru yang direpresentasikan oleh dua simpul yang berbeda maka digunakan bantuan persamaan kuadrat$ax^2 + bx + c =$

0 yang diperoleh dari penggunaan fungsi ln pada ke dua sisi dari persamaan yang diberikan Loh dan Shih (1997) berikut ini

$$\frac{p(A|t)}{s_A}.\phi\{(x-\bar{x}_A)/s_A\} = \frac{p(B|t)}{s_B}.\phi\{(x-\bar{x}_B)/s_B\},$$

dengan $\phi(x) = \frac{1}{\sqrt{2\pi}}.\exp(-\frac{x^2}{2})$ menunjukkan fungsi densitas normal standar, maka

$a = s^2{}_A - s^2{}_B$

$b = 2(\bar{x}_A s^2{}_B - \bar{x}_B s^2{}_A)$

$$c = (\bar{x}_B s_A)^2 - (\bar{x}_A s_B)^2 + 2s^2{}_A s^2{}_B \,.\ln\left(\frac{p(A|t).s_B}{p(B|t).s_A}\right)$$

3. Sebuah simpul disekat pada $x = d$ di mana d didefinisikan sebagai berikut:
   a. Jika $a = 0$, maka

$$d = \begin{cases} \frac{\bar{x}_A+\bar{x}_B}{2} - \frac{s^2{}_A \ln\{\frac{p(A|t)}{p(B|t)}\}}{(\bar{x}_A-\bar{x}_B)} & ,\bar{x}_A \neq \bar{x}_B \\ \bar{x}_A & ,\bar{x}_A = \bar{x}_B \end{cases}$$

   b. Jika $a \neq 0$ maka :
      (i) Jika $b^2 - 4ac < 0$, maka $d = \frac{(\bar{x}_A+\bar{x}_B)}{2}$
      (ii) Jika $b^2 - 4ac \geq 0$, maka :
         A. Ditentukan bahwa $d = \frac{-b\pm\sqrt{b^2-4ac}}{2a}$ yang lebih dekat dengan rata-rata kelas, sehingga menghasilkan dua simpul tak kosong.
         B. Jika tidak, $d = \frac{(\bar{x}_A+\bar{x}_B)}{2}$.

Untuk mendapatkan penyekatan biner, QUEST hanya menggunakan satu dari dua akar sebagai titik sekat, yaitu akar yang nilainya mendekati rata-rata sampel dari setiap kelas (Loh dan Shih, 1997).

b. Menentukan Titik Sekat dengan Variabel Prediktor Kategorik

Misalkan *X* merupakan variabel prediktor kategorik dengan nilai $\{b_1, \dots, b_I\}$. Transformasikan *X* ke dalam variabel kontinu ξ untuk setiap kelas. Lebih jelasnya, maka dapat dilakukan dengan langkah-langkah berikut.

1. Setiap nilai kategori dari variabel prediktor *X* pada simpul terpilih ditransformasikan ke dalam vektor *dummy I* dimensi, $v = (v_{1,\dots,}v_I)'$,
   dimana $v_i = \begin{cases} 1 & b_i; i = 1,2,\dots,I \\ 0 & \text{lainnya} \end{cases}$
2. Menghitung rata-rata kesuluruhan kelas *j* dari v.
   Diketahui :
   v_i : vektor kategori ke- i
   $\bar{v}$: vektor rata-rata untuk semua pengamatan pada simpul t
   $\bar{v}^{(j)}$: rata-rata semua pengamatan pada simpul t untuk kelas respon j
   f_i : banyaknya pengamatan untuk v_i
   n_i : banyaknya pengamatan pada simpul t kelompok respon j untuk v_i
   N_t : total banyaknya pengamatan pada simpul t
   $N_{j,t}$: jumlah pengamatan pada simpul t kelompok respon j.

$$\bar{v} = \frac{\sum_{i=1}^{I} f_i v_i}{N_t}$$

$$\bar{v}^{(j)} = \frac{\sum_{i=1}^{I} n_i v_i}{N_{j,t}},$$

Ditentukan matriks I x I antar kelas (B) dan total kelas (T)

$B = \sum_{j=1}^{J} N_{j,t}\left(\bar{v}^{(j)} - \bar{v}\right)(\bar{v}^{(j)} - \bar{v})'$ (8)

$T = \sum_{i=1}^{I} f_i(v_i - \bar{v})(v_i - \bar{v})'$ (9)

3. Dilakukan *singular value decomposition* (SVD) pada T untuk memperoleh $T = QDQ'$, dimana Q adalah matriks ortogonal I x I, $D = diag(d_1, \dots, d_I)$. Misalkan $D^{-\frac{1}{2}} = diag(d_1^*, \dots, d_I^*)$ dimana $d_i^* = d_i^{-1/2}$ jika $d_i > 0$, dan 0 untuk lainnya.
4. Sederhanakan nilai tunggal pada $D^{-\frac{1}{2}}Q'BQD^{-\frac{1}{2}}$ untuk menentukan vektor eigen a yang berhubungan dengan nilai eigen yang terbesar.
5. Koordinat diskriminan terbesar dari v dapat diproyeksikan dengan $\xi = a'D^{-\frac{1}{2}}Q'v$. QUEST mengaplikasikan algoritma pemilihan titik sekat untuk variabel kontinu yang di kasus ini diwakili ξ untuk menentukan titik sekat.

2.2.1.1.3 Proses Pemberhentian Penyekatan Simpul

Proses penyekatan dilakukan secara berulang sampai tidak mungkin untuk dilanjutkan. Menurut Rokach dan Maimon (2008), penyekatan dihentikan karena:

1. Pada simpul hanya terdapat kasus yang berasal dari salah satu kelas variabel respon.
2. Kedalaman pohon maksimal telah tercapai.
3. Jumlah kasus di simpul terminal kurang dari jumlah minimal kasus untuk menjadi orang tua simpul (*parent node*).
4. Jika simpul disekat, jumlah kasus dalam satu atau lebih anak simpul akan kurang dari jumlah minimal kasus untuk anak simpul, maka proses penyekatan simpul berhenti.
5. Jika semua variabel prediktor mempunyai nilai signifikansi lebih besar dari nilai alpha (α) yang ditentukan, maka simpul tidak dapat disekat.

2.3 Ketepatan Pohon Klasifikasi

Sebuah sistem yang melakukan klasifikasi diharapkan dapat melakukan klasifikasi semua set data dengan benar, namun kinerja suatu sistem klasifikasi ini tidaklah sepenuhnya terhindar dari kesalahan. Bentuk kesalahannya adalah dalam mengklasifikasikan objek baru ke dalam suatu kelas (*missclasification*).

Menurut Prasetyo (2012), matriks konfusi merupakan tabel pencatat hasil kerja klasifikasi. Tabel 1 merupakan matriks konfusi yang melakukan klasifikasi masalah biner . Setiap sel f_{ij} dalam matriks menyatakan jumlah rekord (data) dari kelas i yang hasil prediksinya masuk ke kelas j. Misalnya, sel f_{11} adalah jumlah data dalam kelas A yang secara benar dipetakan ke kelas A, dan f_{10} adalah data dalam kelas A yang dipetakan secara salah ke kelas B.

Tabel 1 Matriks Konfusi

f_{ij}		Kelas hasil prediksi (j)	
		Kelas = A	Kelas = B
Kelas asli (i)	Kelas = A	f_{11}	f_{10}
	Kelas = B	f_{01}	f_{00}

Maka dapat dihitung tingkat akurasi dan tingkat kesalahan prediksi :

1. Akurasi $= \frac{f_{11}+f_{00}}{f_{11}+f_{10}+f_{01}+f_{00}}$

2. Tingkat salah prediksi = $\frac{f_{10}+f_{01}}{f_{11}+f_{10}+f_{01}+f_{00}}$

3. METODOLOGI PENELITIAN

Data yang digunakan dalam pembuatan pohon klasifikasi dengan algoritma *QUEST* ini, adalah data pasien penyakit hati (*liver*) yang dikumpulkan dari timur laut Andrha Pradesh di India. Data yang digunakan terdiri dari 414 pasien liver dan 165 bukan pasien liver dengan 140 pasien berjenis kelamin perempuan dan 439 pasien berjenis kelamin laki-laki. Data diperoleh dari situs internet, http://archive.ics.uci.edu/ml yang disediakan oleh UCI (*University of California at Irvine*).

Variabel dalam penelitian ini terdiri dari variabel prediktor dan variabel respon. Variabel respon pada penelitian adalah status pasien, dimana pada penelitian ini dikelompokkan kedalam dua kategori yaitu:

1. Status 0 = A, yaitu kelompok pasien liver.
2. Status 1 = B, yaitu kelompok pasien non liver.

Variabel prediktor pada penelitian ini adalah hasil tes fungsi hati

1. Jenis Kelamin (JK) berupa data kategorik dengan 0 menyatakan jenis kelamin perempuan dan 1 menyatakan jenis kelamin laki-laki
2. Umur pasien (Umur) berupa data kontinu
3. Total Bilirubin (TB) berupa data kontinu
4. Bilirubin Terkonjugasi (DB) berupa data kontinu
5. Fosfatase Alkali (FA) berupa data kontinu
6. *Serum Glutamic Pyruvic Transaminase* (SGPT) berupa data kontinu
7. *Serum Glutamic Oxaloacetic Transaminase* (SGOT) berupa data kontinu
8. Total Protein (TP) berupa data kontinu
9. Albumin (ALB) berupa data kontinu
10. Rasio Albumin dan Globulin (AG) berupa data kontinu

Tahapan analisis data yang dilakukan menggunakan metode algoritma QUEST adalah sebagai berikut:

1. Memasukkan data Status sebagai variabel respon dan JK, Umur, TB, DB, FA, SGPT, SGOT, TP, ALB serta AG sebagai variabel prediktor.
2. Pemilihan variabel penyekat untuk menentukan simpul yang akan dibentuk.
3. Menentukan titik sekat untuk menyekat simpul yang terbentuk sekaligus memisahkan informasi data ke dalam dua simpul.
4. Proses pemilihan variabel split dan titik split dilakukan berulang-ulang sampai proses pembentukan pohon klasifikasi biner berhenti dengan peraturan pemberhentian yang diberlakukan.
5. Interpretasi pohon klasifikasi yang terbentuk.

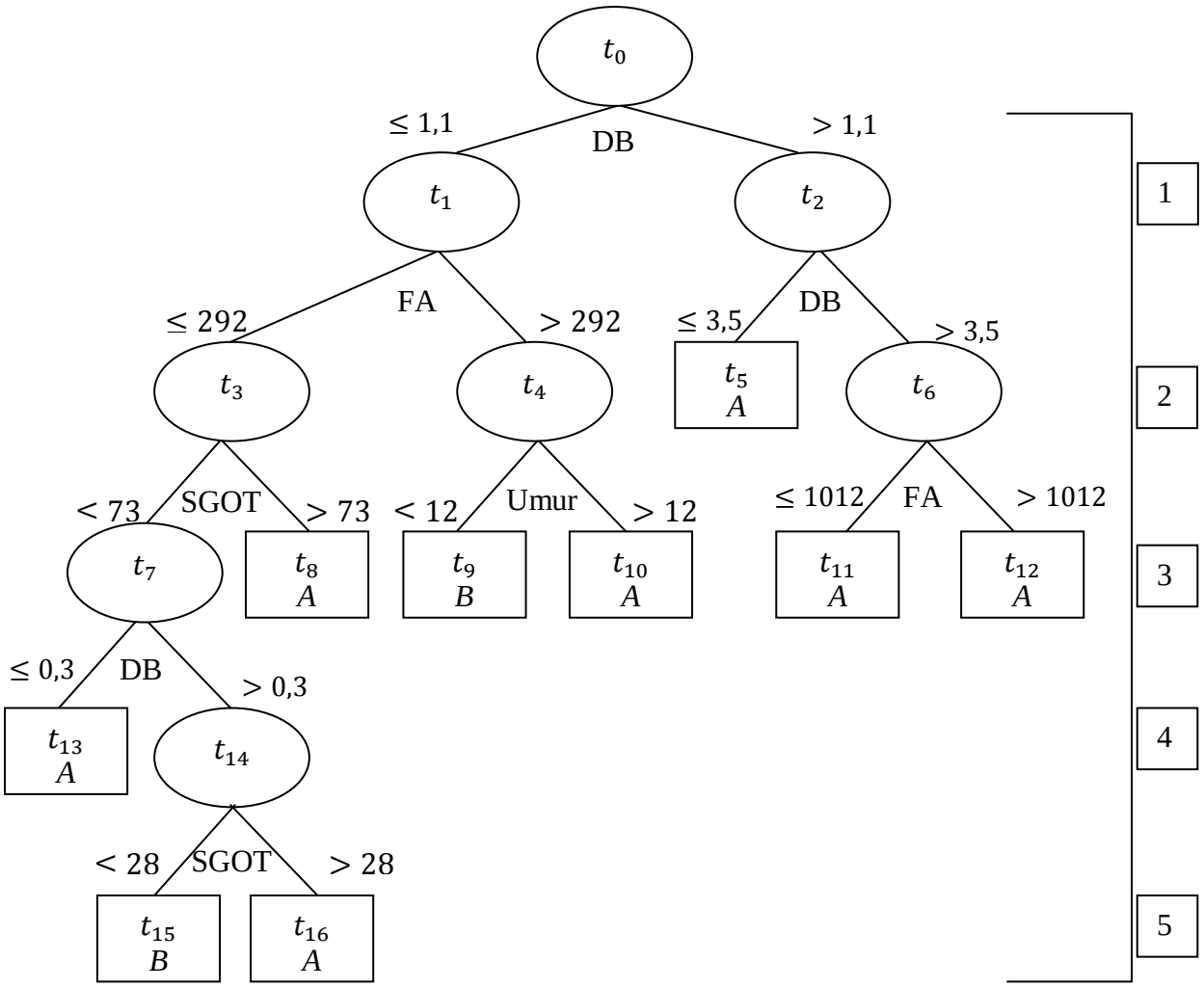
4. Hasil dan Pembahasan

Variabel respon Y terdapat dua kelas, yaitu $j = A, B$. Kelas A menyatakan data pasien liver dan B menyatakan data bukan pasien liver. Jika banyaknya kasus dari masing-masing kelas dinyatakan dalam notasi maka:

$N_A = 414 \qquad N_B = 165 \qquad N = 579$

N_A menyatakan banyaknya kasus pasien liver, N_B menyatakan banyaknya kasus bukan pasien liver, dan N menyatakan banyaknya kasus keseluruhan.

Pada data ini, proses pembentukan pohon klasifikasi biner dibentuk sampai data pada orang tua simpul terdapat satu kasus dan pada anak simpul terdapat satu kasus, nilai batas pengambilan keputusan signifikansi $\alpha = 0{,}05$. Pada t_0 yang merupakan simpul akar, akan di sekat (dipecah) menjadi dua simpul. Untuk mendapatkan variabel penyekat yang memecah simpul t_0, dilakukan uji ANOVA F untuk semua variabel prediktor kontinu sedangkan variabel prediktor kategorik di uji menggunakan Pearson χ^2. Untuk mendapatkan variabel penyekat digunakan variabel prediktor yang signifikan dengan *p-value* terkecil. Pemecahan simpul dilakukan sampai aturan pemberhentian yang berlaku terpenuhi.



Gambar 1 Pohon Klasifikasi Biner yang Terbentuk

Notasi t menunjukkan simpul. Kedalaman pohon yang terbentuk adalah 5 level, dengan t_0 adalah simpul akar menunjukkan kedalaman 0 level. Notasi A menyatakan kelas pasien liver serta B menyatakan kelas bukan pasien liver. Dari Gambar 1 ditunjukkan bahwa simpul t_5, t_8, t_{10}, t_{11}, t_{12}, t_{13}, dan t_{16} diprediksi sebagai kelas pasien liver. Sedangkan t_9 dan t_{15} diprediksi sebagai kelas bukan pasien liver.

4.1 Interpretasi Pohon Klasifikasi yang Terbentuk

4.1.1. Aturan Klasifikasi yang Terbentuk

Hasil dari penerapan algoritma QUEST ke data pasien liver berupa pohon klasifikasi dapat dilihat pada Gambar 1. Dengan menggunakan algoritma QUEST, dihasilkan suatu pohon dengan aturan klasifikasi yang digunakan untuk memprediksi seseorang terkena penyakit liver atau tidak. Dari pohon yang terbentuk, variabel yang penting dalam mengklasifikasikan seseorang terkena penyakit liver atau tidak, yaitu variabel X_4 (DB), X_5 (FA), X_7 (SGOT), dan X_2 (Umur) yang selanjutnya variabel (hasil tes fungsi hati) tersebut dapat disebut sebagai ciri-ciri pasien liver.

Dari pohon klasifikasi yang terbentuk terdapat sembilan simpul akhir, sehingga didapat aturan klasifikasi yang dapat dilihat pada Tabel 2 berikut

Tabel 2 Hasil Klasifikasi Data Pasien Liver

Klasifikasi	Simpul	Status Prediksi Kelas	Aturan yang Terbentuk
1	13	Pasien liver	FA $\leq$ 292, SGOT $\leq$ 73, dan DB $\leq$ 0,3
2	15	Bukan pasien liver	FA $\leq$ 292, SGOT $\leq$ 73, DB $>$ 0,3 dan SGOT $\leq$ 28
3	16	Pasien liver	0,3 $<$ DB $\leq$ 1,1, 28 $<$ SGOT $\leq$ 73, dan FA $\leq$ 292
4	8	Pasien liver	DB $\leq$ 1,1, FA $\leq$ 292, dan SGOT $>$ 73
5	9	Bukan pasien liver	DB $\leq$ 1,1, FA $>$ 292, Umur $\leq$ 12
6	10	Pasien liver	DB $\leq$ 1,1, FA $>$ 292, dan Umur $>$ 12
7	5	Bukan pasien liver	DB $>$ 1,1 dan DB $\leq$ 3,5
8	11	Pasien liver	DB $>$ 1,1, DB $>$ 3,5, dan FA $\leq$ 1012
9	12	Pasien liver	DB $>$ 1,1, DB $>$ 3,5, dan FA $>$ 1012

4.1.2 Ukuran Ketepatan Prediksi

Uji ketepatan pohon klasifikasi dalam mengklasifikasikan data dapat dilakukan menggunakan matriks konfusi pada Tabel 3 berikut

Tabel 3 Matriks Konfusi Hasil Klasifikasi

f_{ij}		Kelas Hasil Prediksi (j)	
		Kelas = A	Kelas = B
Kelas asli (i)	Kelas = A	396	18
	Kelas = B	136	29

Akurasi$= \frac{396+29}{396+18+136+29} = 0{,}734$

Tingkat kesalahan prediksi $= \frac{18+136}{396+18+136+29} = 0{,}266$

Maka tingkat akurasi pohon klasifikasi dalam mengklasifikasikan data sebesar 0,734 atau 73,4 %, dan dengan tingkat kesalahan memprediksi sebesar 0,266 atau 26,6 %.

5. KESIMPULAN

Berdasarkan hasil dan pembahasan yang telah dipaparkan sebelumya, maka dapat diambil kesimpulan bahwa dari pohon yang terbentuk, variabel yang penting dalam mengklasifikasikan seseorang terkena penyakit liver atau tidak, yaitu variabel X_4 (DB), X_5 (FA), X_7 (SGOT), dan X_2 (Umur) yang selanjutnya variabel (hasil tes fungsi hati) tersebut dapat disebut sebagai ciri-ciri pasien liver. Tingkat akurasi pohon klasifikasi dalam mengklasifikasikan data sebesar 73,4 %.

6. DAFTAR PUSTAKA

Bache, K. dan Lichman, M. 2013. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Berry, M. J. A. dan Linoff, G. S. 2004. *Data Mining Techniques*. Wiley Publishing, Inc: Indiana.

Grabczewski, K. 2014. *Meta-Learning in Decision Tree Induction*. Springer International Publishing, Switzerland.

Kim, H. dan Loh, W.-Y. 2001. Classification Trees with Unbiased Multiway Splits. *Am. Statist. Assoc*. 96; 590-604.

Larose, D. T. 2005. *Discovering Knowledge in Data "An Introduction to Data Mining".* John Wiley dan Sons, Inc: New Jersey.

Loh, W.-Y. dan Shih, Y.-S. 1997. Split Selection Methods for Classification Trees, *Statistica Sinica 7*; 815-840.

Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I. dan de Mendonca, A. 2011. Data mining methods in the prediction of Dementia. *BMC Research Notes*; 4:299.

Prasetyo, E. 2012. *Data Mining: Konsep dan Aplikasi Menggunakan MATLAB*. C.V Andi Offset: Yogyakarta.

Rokach, L. dan Maimon, O. 2008. *Data Mining with Decision Trees "Theory and Apllication"*. World Scientific Publishing Co. Pte. Ltd. : USA.

Soemohardjo, S., Soeleiman, B. H., Widjaya, A. dan Muljanto. 1983. *Tes Faal Hati "Dasar-dasar Teoritik dan Pemakaian dalam Klinik"*. Penerbit Alumni: Bandung.