Metode Pohon Klasifikasi pada Data Respon Biner

Achmad Zainul Wafah
Dosen STMIK AsiA Malang

ABSTRAK

Pelanggaran terhadap asumsi dalam metode parametrik menyebabkan suatu data tidak dapat dianalisis dengan metode yang dilandasi asumsi tersebut, namun dalam kenyataannya data tersebut perlu dianalisis karena ingin diperoleh informasi yang terdapat dalam data. Oleh karena itu digunakan metode nonparametrik sehingga data tersebut dapat dianalisis dan diperoleh informasi yang bermanfaat. Metode pohon klasifikasi merupakan metode nonparametrik yang dapat digunakan untuk menduga dan mengelompokkan suatu obyek ke dalam salah satu kategori peubah respon. Metode ini dapat digunakan untuk mencari peubah yang berpengaruh terhadap kematian dini pada neonatus (bayi yang baru lahir) dan membentuk model pohon klasifikasi. Model pohon optimal digunakan sebagai model untuk pendugaan terhadap status neonatus karena model ini memiliki struktur pohon yang sederhana dan memiliki tingkat kesalahan relatif validasi silang lipat-10 yang lebih kecil dari pada model pohon maksimal, yaitu masing-masing sebesar 0,383 dan 0,586. Dari 3 peubah prediktor yang terdapat dalam data, hanya berat lahir (X_1) dan skor apgar (X_2) yang mempengaruhi pembetukan model. Neonatus yang memiliki berat lahir lebih dari 2027,5 gram dan memiliki skor apgar 2 atau 6 atau 7 dikategorikan neonatus akan tetap hidup dengan peluang sebesar 0,958.

***Kata kunci**: Nonparametrik, Metode Pohon Klasifikasi , Neonates, Data Biner.*

ABSTRACT

The violation to the assumptions in parametric methods cause a data can not be analyzed with the methods based on that assumptions, but actually the data must be analyzed so that we get such information. Because of that, nonparametric method is used. The method of classification tree is a nonparametric method which is used to predict and classify an object into one of categories of the independent variable. This method can be used to look for a variable that influencing the early death of neonatus and build a classification tree model. The optimal tree model is used in predicting the neonatus status because this model have a simple tree structure and 10-vold cross validated relative cost less than maximal tree, they are 0.383 and 0.586 respectively. From 3 predictors, only the birth weight (X_1) and apgar score (X_2) are influencing the model. Neonatus with birth weight more than 2027.5 grams and apgar score 2 or 6 or 7 are classified to survive with life probability of 0.958.

***Keywords:** Nonparametrik, Metode Pohon Klasifikasi , Neonates, Data Biner.*

PENDAHULUAN

Analisis regresi logistik memiliki beberapa asumsi yang harus dipenuhi agar diperoleh hasil yang sahih dan akurat. Asumsi tersebut antara lain sebaran galat berdistribusi normal dan tidak terdapat korelasi antara peubah prediktor yang membentuk model. Asumsi-asumsi tersebut menyebabkan suatu data tidak dapat dianalisis dengan metode regresi logistik karena ada asumsi yang tidak terpenuhi, namun dalam kenyataannya data tersebut harus dianalisis karena ingin diperoleh informasi. Oleh karena itu dalam jurnal ini akan dibahas suatu metode yang dapat digunakan untuk memprediksi atau mengklasifikasikan suatu obyek ke dalam salah satu kategori dari peubah respon, yaitu metode pohon klasifikasi.

Metode pohon klasifikasi merupakan metode nonparametrik sehingga tidak diperlukan pemenuhan asumsi kenormalan data. Struktur data dapat dilihat secara visual sehingga memudahkan eksplorasi data dan pengambilan keputusan berdasarkan model yang diperoleh. Namun perhitungan metode pohon klasifikasi sulit dilakukan secara manual karena banyaknya kombinasi untuk mencari peubah yang paling dominan sehingga digunakan software CART (*Classification and Regression Trees*), tidak ada penetapan validasi silang lipat-v (v-*vold cross validation*) dan jenis fungsi keheterogenan simpul.

KAJIAN TEORI

1. Pohon Klasifikasi

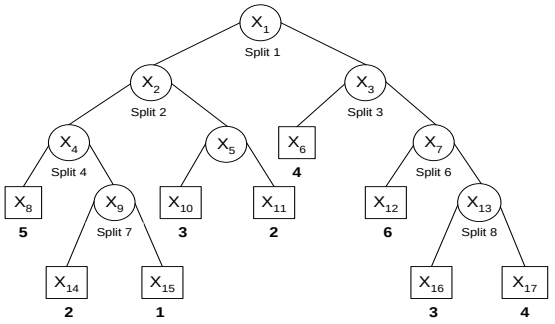
Menurut Breiman *et al.* (1984), metode pohon regresi dan pohon klasifikasi adalah metode yang digunakan untuk menggambarkan hubungan antara peubah respon dengan satu set peubah prediktor. Pohon klasifikasi bertujuan untuk menghasilkan pengklasifikasian yang akurat dan menjelaskan prediksi data baru dalam tiap kategori yang terdapat dalam respon.

Pembentukan pohon klasifikasi memerlukan 4 komponen, yaitu:

a. Gugus pertanyaan dikotomus dengan bentuk " apakah $x_i' \in A$ (Apakah x_i' anggota dari A)?" dan $A \subset X$ (A himpunan bagian dari X), di mana x_i' merupakan hasil amatan dan X adalah ruang peubah prediktor (dengan i = 1,2,...,N, di mana N merupakan banyaknya amatan). Jawaban dari pertanyaan tersebut menentukan pemilahan bagi ruang peubah prediktor. Amatan dengan jawaban "ya" masuk ke ruang A, sedangkan jawaban "tidak" masuk ke ruang komplemen A. Ruang yang terbentuk dari jawaban tersebut disebut simpul.
b. *Kriteria Goodness-of-Split* merupakan alat evaluasi bagi pemilahan yang dilakukan pemilah s pada simpul t berupa persamaan
$\Phi(s,t) = i(t) - P_L i(t_L) - P_R(t_R)$, (2.1).
di mana:
- $i(t)$ = fungsi keheterogenan pada simpul t.
- P_L = proporsi data yang masuk ke simpul kiri.
- P_R = proporsi data yang masuk ke simpul kanan.
c. Ukuran yang digunakan untuk menentukan ukuran pohon yang layak yaitu dengan menggunakan sampel uji (*test sample*) atau sampel validasi silang lipat-v.
d. Aturan penandaan label kelas pada setiap simpul terminal.

2. Struktur Pohon Klasifikasi

Menurut Breiman et al. (1984), pohon klasifikasi dibentuk dengan perulangan pemilah pada setiap simpul menjadi dua bagian himpunan turunan. Proses pemilahan dimulai dari simpul utama (root node) yang berisi data yang akan dipilah. Pemilahan dilakukan pada tiap simpul sampai didapatkan suatu simpul akhir. Peubah yang memilah pada simpul utama adalah peubah terpenting dalam menentukan kelas dari amatan. Sebagai ilustrasi pandang struktur pohon kalsifikasi berikut:



Gambar 2.1 Pohon Klasifikasi dengan 6 Kelas

Gambar 1: Pohon Klasifikasi dengan 6 Kelas

Gambar dan Tabel diletakkan di dalam kelompok teks dan diberi keterangan. Gambar diikuti dengan judul gambar yang diletakkan di bawah gambar yang bersangkutan. Tabel diberi judul tabel yang diletakkan di atas tabel yang bersangkutan. Judul gambar dan judul tabel diberi nomor urut. Tabel yang ditampilkan tanpa garis vertikal, sedangkan garis horisontal hanya ditampilkan 3 garis horisontal utama yaitu 2 garis horisontal untuk item judul kolom dan garis penutup dari baris paling bawah.

3. Pembentukan Pohon Klasifikasi

Breiman et al. (1984), menjelaskan bahwa untuk membentuk pohon klasifikasi diperlukan Learning Sample L yang terdiri dari N amatan dengan data (xn,jn), di mana $x_n \in X$ dan $j_n \in \{1,2,...,J\}$, n= 1,2,...,N. Proses pembentukan pohon klasifikasi terdiri dari 3 tahapan, yaitu pemilahan pemilah, penentuan simpul terminal dan penandaan label kelas.

4. Pemangkasan Pohon Klasifikasi

Menurut Breiman, et al. (1984), untuk mendapatkan ukuran pohon yang layak dilakukan pemangkasan dengan ukuran cost complexity minimal. Untuk sebarang pohon T yang merupakan sub pohon dari pohon terbesar Tmax (T<Tmax) ukuran cost complexity adalah:

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}| \qquad (2.10)$$

di mana:

R(T)= Resubstitution Estimate (Penduga Pengganti)

α = Parameter cost complexity

$|\tilde{T}|$ = Ukuran banyaknya simpul terminal pohon T.

5. Penentuan Pohon Klasifikasi Optimal

Pohon klasifikasi yang terbentuk bisa berukuran sangat besar dan kompleks dalam menggambarkan struktur data, sehingga

dilakukan pemangkasan. Pemangkasan dilakukan dengan penilaian ukuran sebuah pohon tanpa mengabaikan kebaikan struktur pohon dan ketepatan klasifikasi melalui pengurangan simpul pohon, sehingga dicapai penghematan gambaran. Pemangkasan dilakukan dengan memangkas bagian pohon yang kurang penting (simpul anak yang memiliki tingkat keheterogenan sama atau lebih dari simpul induk), sehingga diperoleh pohon optimal.

PEMBAHASAN

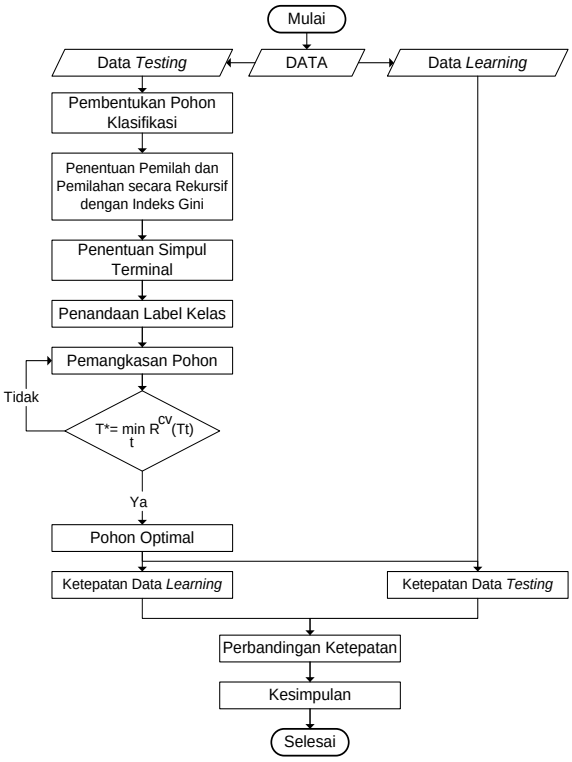
Dalam penelitian ini digunakan data rumah sakit dr. Saiful Anwar Malang. Peubah respon yaitu status neonatus (bayi yang baru lahir dengan usia 0-7 hari) dan peubah penjelas (faktor-faktor kematian dini pada neonatus) adalah sebagai berikut:

-X1 = Berat lahir (gram)
-X2 = Skor apgar
-X3 = Masa gestasi (minggu)
-Y =0 :Neonatus mengalami kematian
= 1 : Neonatus tetap hidup

Metode pohon klasifikasi dilakukan dengan bantuan software CART. Alur dari penelitian adalah sebagai berikut:

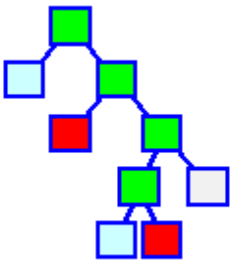
1. Membagi data secara acak menjadi dua kelompok, sebanyak 2/3 dari data sebagai data learning yang akan digunakan dalam pembentukan model sedangkan 1/3 bagian data digunakan sebagai data testing yang akan dugunakan untuk validasi (estimasi tingkat kesalahan pengklasifikasian sehingga diperoleh ketepatan model).
2. Membentuk pohon klasifikasi dengan menggunakan data learning.
   a) Menentukan pemilah terbaik dan pemilahan secara rekursif berdasarkan kriteria pemilahan. Pemilah terbaik adalah pemilah yang memberikan penurunan keheterogenan tertinggi. Proses pemilahan dilakukan dengan terlebih dahulu menentukan fungsi pemilah, dalam tugas akhir ini digunakan fungsi pemilahan Indeks Gini.
   b) Menentukan simpul terminal. Simpul terminal diperoleh bila simpul t didapatkan sehingga tidak terdapat penurunan keheterogenan secara berarti.
   c) Penandaan label kelas dengan aturan jumlah terbanyak.
3. Memangkas pohon yang terbentuk berdasarkan cost-complexity minimum dan contoh validasi silang lipat-10.
4. Memilih pohon optimal dengan nilai kesalahan relatif validasi silang lipat-10 minimum.
5. Menguji keakuratan model klasifikasi yang telah diperoleh dari metode pohon klasifikasi dengan menggunakan data testing.

Gambar dibawah merupakan skema analisis klasifikasi pohon:



Gambar 2: Tahap analisis klasifikasi pohon

Model pohon optimal yang terbentuk dari analisis pohon klasifikasi dapat dilihat pada gambar berikut:



Gambar 3. Skema analisis klasifikasi pohon

Penjabaran dari gambar pohon maksimal adalah:

a) Node berwarna hijau merupakan simpul dalam, sedangkan yang lain merupakan simpul terminal.
b) Warna yang berbeda pada simpul terminal menunjukkan tingkat kehomogenan dari tiap simpul. Warna merah menujukkan tingkat kehomogenan simpul yang rendah dan wana warna biru tua menunjukkan tingkat

keheterogenan yang tinggi, sedangkan warna lain merupakan peralihan dari warna merah ke warna biru tua yang menunjukkan tingkat keheterogenan sesuai dengan peralihan warna tersebut.

c) Node pertama merupakan simpul dalam pertama, node kedua merupaka simpul dalam kedua, dan seterusnya. Terminal node pertama merupakan simpul terminal pertama, terminal node kedua merupakan simpul terminal kedua dan seterusnya.
d) Simpul dalam pertama merupakan himpunan dari peubah bayi yang akan dipilah menjadi dua bagian yang lebih homogen.
e) Model pohon optimal yang terbentuk memiliki 4 buah simpul dalam dan 5 buah simpul terminal.
f) Dari Gambar 3 dapat diketahui bahwa simpul terminal yang memiliki tingkat kehomogenan tinggi adalah simpul terminal 2 dan 4 sedangkan simpul terminal yang lain memiliki tingkat kehomogenan rendah.

PENUTUP

Hasil yang diperoleh dari metode pohon klasifikasi terhadap data neonatus pada bayi berat lahir rendah adalah sebagai berikut:

1. Model pohon maksimal dan model pohon optimal menghasilkan tingkat kesalahan relatif validasi silang lipat-10 berturut-turut sebesar 0,586 dan 0,383, dengan struktur model pohon optimal yang lebih sederhana dari pada struktur model pohon maksimal.
2. Perubah prediktor yang mempengaruhi kematian neonatus dini berdasarkan model optimal adalah berat lahir (X1) dan skor apgar (X2),sedangkan masa gestasi tidak mempengaruhi kematian neonatus dini.

Neonatus yang lahir dikategorikan akan tetap hidup bila memiliki berat lahir lebih dari 2027,5 gram dan memiliki skor apgar 2 atau 6 atau 7 dengan peluang sebesar 0,958.

Karena penelitian ini masih terbatas pada peubah respon dengan kategori biner, perlu kajian untuk peubah respon berskala ordinal atau memiliki lebih dari dua kategori dengan metode klasifikasi yang lain.

DAFTAR PUSTAKA

1. Anonymous. 2004. Scientific Software. www.hearne.com. Australia. Tanggal akses:11 Oktober 2005.
2. Breiman, L. 1992. Some Properties of Splitting Criteria. Statistics Department, University of California, Berkeley.
3. Breiman, L., J.H. Friedman, R.A. Olsen and C.J. Stone. 1984. Classification and Regression Trees. Champman and Hall. New York.
4. Lewis, R.J. 2000. An Introduction to Classification and Regression Tree (CART) Analysis.www.saem.org. Tanggal akses: 02 Agustus 2005.
5. Steinberg, D. and P. Colla. 1997. *CART-Classification and Regression Trees*. Salford Systems. San Diego.
6. Venables, W.N. and B.D. Ripley. 1994. Modern Applied Statistics with S-Plus, 2nd ed. Sringer-Verlag. New York.