BAB III
REGRESI LOGISTIK BINER DAN *CLASSIFICATION AND REGRESSION TREES* (CART)

3.1 Regresi Logistik Biner

Regresi logistik berguna untuk meramalkan ada atau tidaknya karakteristik berdasarkan prediksi seperangkat variabel prediktor. Regresi logistik menghasilkan rasio peluang (*odds ratio/OR*) terkait dengan nilai setiap variabel prediktor. *Odds ratio* dari suatu kejadian diartikan sebagai peluang peristiwa yang terjadi dibagi dengan peluang suatu peristiwa yang tidak terjadi.

$$Odds\ Ratio = \frac{p}{(1-p)} \tag{3.1}$$

dengan:

p = peluang dari peristiwa yang terjadi

$p-1$ = peluang dari peristiwa yang tidak terjadi

Regresi logistik biasanya digunakan untuk memprediksi variabel yang bersifat kategorik (biasanya dikotomi) oleh seperangkat variabel prediksi. Dengan adanya sifat variabel yang kategorikal, analisis fungsi diskriminan biasanya digunakan jika semua variabel prediktor berbentuk data kontinu dan terdistribusi dengan baik. Analisis logit digunakan jika semua variabel prediktor bersifat kategorik dan regresi logistik dipilih jika variabel prediktor memuat campuran variabel kontinu dan kategorik.

Analisis regresi logistik biner digunakan untuk melihat pengaruh sejumlah variabel prediktor $x_1, x_2, x_3, \dots, x_k$ terhadap variabel respon y yang berupa variabel respon biner dan hanya mempunyai dua nilai. Model regresi logistik biner berdistribusi Bernoulli. Distribusi Bernoulli adalah distribusi dari peubah acak yang hanya mempunyai dua kategori, misalnya sukses atau gagal serta untung atau rugi.

Jika data hasil pengamatan memiliki p buah variabel prediktor yaitu $x_1, x_2, x_3, \dots, x_p$ dan satu variabel respon Y, dengan Y mempunyai dua kemungkinan nilai yaitu 0 dan 1, maka:

$Y = 1$ menyatakan bahwa respon memiliki kriteria yang ditentukan

$Y = 0$ menyatakan bahwa respon tidak memiliki kriteria yang ditentukan

Jika variabel Y berdistribusi Bernoulli dengan parameter $\pi(x_i)$, maka fungsi distribusi peluang menjadi:

$$f(y_i) = [\pi(x_i)]^{y_i}[1 - \pi(x_i)]^{1-y_i} \qquad y_i = 0,1 \tag{3.2}$$

sehingga diperoleh:

untuk $y_i = 0$ $f(0) = [\pi(x_i)]^0[1 - \pi(x_i)]^{1-0} = 1 - \pi(x_i)$

untuk $y_i = 1$ $f(1) = [\pi(x_i)]^1[1 - \pi(x_i)]^{1-1} = \pi(x_i)$

Hosmer dan Lemeshow (2000: 31), model umum regresi logistik dengan p buah variabel prediktor dibentuk dengan nilai $\pi(x) = E(Y = 1|x)$, $\pi(x)$ dinotasikan sebagai berikut:

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \tag{3.3}$$

dengan $g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

Fungsi $\pi(x)$ merupakan fungsi non linear sehingga untuk membuatnya menjadi fungsi linear harus dilakukan transformasi logit agar dapat dilihat hubungan antara variabel respon (y) dengan variabel prediktornya (x). Bentuk logit dari $\pi(x)$ adalah $g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right]$ sehingga diperoleh:

$$logit\,[\pi(x)] = g(x) = \ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{3.4}$$

$g(x)$ merupakan fungsi hubungan dari model regresi logistik yang disebut fungsi hubungan logit.

Bukti:

$$
\begin{aligned}
g(x) &= \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] \\
&= \ln\left[\frac{\dfrac{e^{(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}}{1+e^{(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}}}{1-\dfrac{e^{(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}}{1+e^{(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}}}\right] \\
&= \ln\left[\frac{\dfrac{e^{(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}}{1+e^{(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}}}{\dfrac{1+e^{(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}}{1+e^{(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}}-\dfrac{e^{(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}}{1+e^{(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}}}\right] \\
&= \ln\left[\frac{\dfrac{e^{(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}}{1+e^{(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}}}{\dfrac{1}{1+e^{(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}}}\right] \\
&= \ln\left[\frac{e^{(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}}{1+e^{(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}}\left(1+e^{(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}\right)\right] \\
&= \ln\left[e^{(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}\right] \\
&= \beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p
\end{aligned}
$$

3.2 Penaksiran Parameter

Pada regresi linear umumnya digunakan metode kuadrat terkecil untuk menaksir parameter β. Berdasarkan asumsi yang biasa digunakan untuk regresi linear (misalnya asumsi kenormalan ataupun kehomogenan varians), metode kuadrat terkecil akan menghasilkan penaksir parameter dengan sifat-sifat statistik yang diinginkan (tak bias dan memiliki varians minimum). Namun apabila metode kuadrat terkecil ini diterapkan untuk model dengan variabel respon biner, maka penaksir parameter yang dihasilkan tidak lagi memiliki sifat-sifat statistik yang diinginkan tersebut, yaitu ada asumsi homoskedastisitas yang tidak mungkin dipenuhi oleh distribusi Bernoulli. Hal ini disebabkan karena varians distribusi Bernoulli berubah-ubah bergantung pada nilai peluang suksesnya. Oleh karena itu, pendekatan yang digunakan untuk mengatasi hal tersebut adalah dengan metode kemungkinan maksimum atau *Maximum Likelihood Estimation* (MLE).

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \quad , \ \beta = \pi(x_i) \tag{3.5}$$

dengan:

i = 1,2, … , n

y_i = pengamatan pada variabel respon ke-*i*

(x_i) = peluang untuk variabel prediktor ke-*i*

Untuk mempermudah perhitungan, maka dilakukan penaksiran parameter β dengan cara memaksimumkan fungsi logaritma kemungkinannya (*log-likelihood*), yaitu:

$$L(\beta) = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1-y_i)\ln[1-\pi(x_i)]\} \tag{3.6}$$

Bukti:

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i}$$

$$\begin{aligned} \ln l(\beta) &= \ln\left(\prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i}\right) \\ &= \sum_{i=1}^{n} \left(\ln\left(\left[\pi(x_i)^{y_i}\right][1-\pi(x_i)]^{1-y_i}\right)\right) \\ &= \sum_{i=1}^{n} \left(\ln\left[\pi(x_i)^{y_i}\right] + ln\,[1-\pi(x_i)]^{1-y_i}\right) \\ &= \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1-y_i)\ln[1-\pi(x_i)]\} \end{aligned}$$

Untuk mendapatkan nilai penaksiran koefisien regresi logistik ($\widehat{\beta}$) dilakukan dengan membuat turunan pertama $L(\beta)$ terhadap β dan disamakan dengan nol (Herrhyanto, 2003:97).

$$L(\beta) = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1-y_i)\ln[1-\pi(x_i)]\}$$

$$= \sum_{i=1}^{n}\{y_i \ln[\pi(x_i)]\} + \sum_{i=1}^{n}\{(1-y_i)\ln[1-\pi(x_i)]\}$$

$$= \sum_{i=1}^{n}\{y_i \ln[\pi(x_i)]\} + \left(\sum_{i=1}^{n} 1 - \sum_{i=1}^{n} y_i\right)\ln[1-\pi(x_i)]$$

$$= \sum_{i=1}^{n}\{y_i \ln[\pi(x_i)]\} + \left(n - \sum_{i=1}^{n} y_i\right)\ln[1-\pi(x_i)]$$

turunkan ln L(β) terhadap $\pi(x_i)$, yaitu:

$$\frac{d \ln L(\beta)}{d\beta} = \frac{\sum_{i=1}^{n} y_i}{\pi(x_i)} + \frac{n-\sum_{i=1}^{n} y_i}{1-\pi(x_i)}(-1)$$

$$\frac{\sum_{i=1}^{n} y_i}{\pi(x_i)} - \frac{n-\sum_{i=1}^{n} y_i}{1-\pi(x_i)}$$

$$\frac{d \ln L(\beta)}{d\beta} = 0$$

$$\frac{\sum_{i=1}^{n} y_i}{\pi(x_i)} - \frac{n-\sum_{i=1}^{n} y_i}{1-\pi(x_i)} = 0$$

$$\frac{(1-\pi(x_i))(\sum_{i=1}^{n} y_i) - \pi(x_i)\,(n-\sum_{i=1}^{n} y_i)}{\pi(x_i)(1-\pi(x_i))} = 0$$

$$(1-\widehat{\pi(x_i)})(\textstyle\sum_{i=1}^{n} y_i) - \widehat{\pi(x_i)}\,(n-\sum_{i=1}^{n} y_i) = 0$$

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i\,\widehat{\pi(x_i)} - n\,\widehat{\pi(x_i)} + \sum_{i=1}^{n} y_i\,\widehat{\pi(x_i)} = 0$$

$$\sum_{i=1}^{n} y_i - n\,\widehat{\pi(x_i)} = 0$$

$$\sum_{i=1}^{n} y_i = n\,\widehat{\pi(x_i)}$$

$$\widehat{\pi(x_i)} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}$$

Karena $\beta = \pi(x_i)$, maka didapatkan $\hat{\beta}$ yang merupakan penduga kemungkinan maksimum.

3.3 Uji Signifikansi Parameter

Pengujian terhadap parameter model dilakukan untuk memeriksa peranan variabel-variabel prediktor yang ada dalam model terhadap variabel responnya.

Pengujian terhadap parameter ini dilakukan melalui statistik G. Maharani *et al*. (2007: 39), statistik uji G yaitu uji rasio kemungkinan maksimum *(maximum likelihood ratio test)* yang digunakan untuk menguji peranan variabel prediktor di dalam model secara bersama-sama dengan rumusannya sebagai berikut:

$$G = -2 \ln \left[\frac{L_0}{L_p} \right] \tag{3.7}$$

dengan:

L_0 = *likelihood* tanpa variabel prediktor

L_p = *likelihood* dengan *p* variabel prediktor

Langkah-lagkah pengujiannya sebagai berikut:

1). Rumusan Hipotesis

$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$

H_1: paling sedikit ada satu $\beta_i \neq 0$, $i = 1,2, \ldots, p$

2). Besaran yang diperlukan

Hitung L_0, L_p

3). Statistik Uji

$$G = -2 \ln \left[\frac{L_0}{L_p} \right]$$

4). Kriteria Pengujian

Dengan mengambil taraf nyata α, maka tolak H_0 jika $G > \chi^2_{(\alpha,v)}$.

5). Kesimpulan

Penafsiran H_0 diterima atau ditolak

Selanjutnya dengan menggunakan uji *Wald*, akan dilakukan pengujian secara individu terhadap signifikansi parameter model. Menurut Hosmer dan Lemeshow (2000: 16), statistik Uji *Wald* didefinisikan sebagai:

$$W = \left(\frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \right)^2 \tag{3.8}$$

dengan:

$\hat{\beta}_i$ = penaksir dari β_i

$SE(\hat{\beta}_i)$ = penaksir galat baku dari β_i

Uji *Wald* ini akan menunjukkan apakah suatu variabel prediktor signifikan atau layak untuk masuk dalam model atau tidak. Uji *Wald* ini diperoleh dengan membandingkan penaksir kemungkinan maksimum dari parameter, yaitu β_i dengan penaksir galat bakunya. Adapun langkah-langkah pengujiannya adalah sebagai berikut:

1). Rumusan hipotesis

$H_0: \beta_i = 0, i = 1,2,\dots,p$

$H_1: \beta_i \neq 0, i = 1,2,\dots,p$

2). Besaran yang diperlukan

$\hat{\beta}_i$ dan $SE(\hat{\beta}_i) = \sqrt{\left(\sigma^2(\hat{\beta}_i)\right)}$

3). Statistik Uji

$$W = \left(\frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}\right)^2$$

4). Kriteria Pengujian

Tolak H_0 jika $|W| > \chi^2_{(\alpha,1)}$

5). Kesimpulan

Penafsiran H_0 diterima atau ditolak

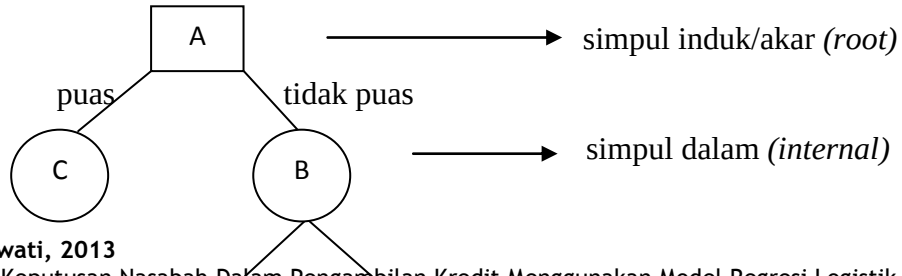
3.4 *Classification and Regression Trees* (CART)

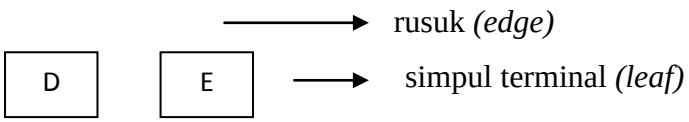
CART adalah salah satu metode atau algoritma dari salah satu teknik eksplorasi data, yaitu teknik pohon keputusan. Metode ini dikembangkan oleh Leo Breiman, Jerome H. Friedman, Richard A. Olshen dan Charles J. Stone sekitar tahun 1980-an.

CART merupakan metode statistika nonparametrik yang dapat menggambarkan hubungan antara variabel respon dengan satu atau lebih variabel prediktor. CART dikembangkan untuk topik analisis klasifikasi, baik untuk variabel respon kategorik maupun kontinu. CART menghasilkan sebuah pohon klasifikasi (*classification trees*), jika variabel responnya kategorik dan menghasilkan pohon regresi (*regression trees*), jika variabel responnya kontinu. Variabel respon dalam penelitian ini berskala kategorik, sehingga metode yang akan digunakan adalah metode pohon klasifikasi.

CART dapat menyeleksi variabel-variabel dan interaksi-interaksi variabel yang paling penting dalam penentuan hasil. Tujuan utama CART adalah untuk mendapatkan suatu kelompok data yang akurat sebagai penciri dari suatu pengklasifikasian. CART mempunyai beberapa kelebihan dibandingkan dengan metode pengelompokan yang klasik, seperti hasilnya lebih mudah diinterpretasikan, lebih akurat, dan lebih cepat penghitungannya. Menurut Yohannes dan Webb (Otok, 2009: XVI-2), tingkat kepercayaan yang dapat digunakan dalam pengklasifikasian data baru pada CART adalah akurasi yang dihasilkan oleh pohon klasifikasi yang murni dibentuk dari data yang mempunyai kesamaan kondisi.

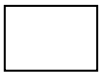
CART merupakan metode yang bisa diterapkan untuk himpunan data yang memiliki jumlah besar, variabel prediktornya banyak dengan skala variabel campuran dilakukan melalui prosedur pemilahan biner, sejauh terlihat dalam Gambar 3.1 berikut.

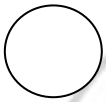




Gambar 3.1
Diagram CART

Keterangan:


= simbol keputusan


= simbol kejadian tidak pasti

Pada Gambar 3.1 di atas A, B, C, D dan E merupakan variabel prediktor yang terpilih untuk menjadi simpul. A merupakan simpul induk atau simpul akar, B merupakan simpul dalam, sementara C, D dan E merupakan simpul akhir atau simpul terminal yang tidak bercabang lagi. Setiap simpul terminal merupakan titik akhir dari suatu pemilahan berstruktur pohon, simpul ini tidak bisa dipilah kembali menjadi simpul lain atau dengan kata lain simpul terminal merupakan simpul yang mengandung amatan-amatan yang homogen dan akhirnya akan dimasukkan sebagai suatu kelas tertentu. Variabel prediktor yang dianggap berpengaruh terhadap variabel respon adalah variabel prediktor yang muncul sebagai pemisah.

Tahapan dalam pembuatan pohon klasifikasi adalah membuat pohon yang besar yaitu dengan simpul yang banyak. Pohon yang terbentuk kemudian disederhanakan dengan cara memangkas beberapa cabang untuk mendapatkan struktur pohon yang layak dengan aturan-aturan tertentu sehingga terbentuk sebuah pohon optimal.

3.5 Langkah-langkah Algoritma Pohon Klasifikasi CART

Algoritma penyusunan pohon klasifikasi dan pohon regresi telah banyak digunakan dalam berbagai macam penelitian. Beberapa algoritma tersebut,

diantaranya C4.5 dan C5, CHAID, CART, dan QUEST. Pada prinsipnya algoritma-algoritma tersebut sebagai berikut:

1. Identifikasi variabel penjelas dan nilainya (atau levelnya kalau itu adalah variabel kategorik) yang dapat digunakan sebagai pemisah keseluruhan data menjadi dua atau lebih subset data.
2. Lakukan iterasi terhadap proses nomor 1 terhadap subset-subset yang ada sampai ditemukan salah satu dari dua hal berikut:
   a. semua subset sudah homogen nilainya
   b. tidak ada lagi variabel prediktor yang bisa digunakan
   c. jumlah amatan di dalam subset sudah terlalu sedikit untuk menghasilkan pemisahan yang memuaskan
3. Lakukan pemangkasan (*pruning*), jika pohon yang dihasilkan dinilai terlalu besar.

Proses identifikasi variabel prediktor dan nilai yang menjadi batas pemisah dapat dilakukan dengan berbagai cara dan berbagai kriteria. Namun tujuan dari pemisahan ini pada berbagai metode adalah sama, yaitu mendapatkan subset-subset yang memiliki nilai variabel respon yang lebih homogen daripada sebelum dilakukan pemisahan.

Algoritma pembentukan pohon klasifikasi CART terdiri dari empat tahapan, yaitu:

1). Pemilihan pemilah (*Classifier)*
2). Penentuan simpul terminal
3). Penandaan label kelas
4). Penentuan pohon klasifikasi optimal

1. Pemilihan Pemilah

Pada tahap ini dicari pemilah dari setiap simpul yang menghasilkan penurunan tingkat keheterogenan paling tinggi. Untuk mengukur tingkat keheterogenan suatu kelas dari suatu simpul tertentu dalam pohon klasifikasi dikenal dengan istilah *impurity measure*. Fungsi impuritas yang dapat digunakan didalam pembentukan pohon klasifikasi CART adalah Indeks Gini. Derajat

impurity yang tinggi menunjukkan simpul tersebut belum homogen, sedangkan sebuah simpul dengan derajat *impurity* yang rendah menunjukkan simpul tersebut sudah homogen. Jika kelas obyek dinyatakan dengan *k*, *k* = 1,2,..,*m*, dimana *m* adalah jumlah kelas untuk variabel/output respon *y*, maka nilai impuritas dari simpul menggunakan Indeks Gini dapat dituliskan persamaannya sebagai berikut:

$$Gini(t) = 1 - \sum_{i=1}^{m} [P(k|t)]^2 \quad (3.9)$$

dengan

$[P(k|t)]$ = frekuensi relatif dari kelas *j* pada simpul *t*

m = jumlah kelas

Jika nilai Indeks Gini, $Gini(t) = 0$, maka semua data dari simpul tersebut sudah berada pada kelas yang sama (homogen). Misalkan dilakukan pemisahan (*spliting*) sebuah simpul menggunakan Indeks Gini. Jika simpul *t* di *split* kedalam *k* partisi (anak), maka kualitas *split* dihitung sebagai berikut:

$$Gini_{split} = \sum_{i=1}^{k} \frac{n_i}{n} Gini(t) \quad (3.10)$$

dengan

n_i = Jumlah record pada anak ke-i

n = Jumlah record pada simpul

2. Penentuan Simpul Terminal

Suatu simpul *t* akan menjadi simpul terminal atau tidak akan dipilah kembali, apabila pada simpul *t* tidak terdapat penurunan keheterogenan secara berarti (sudah homogen) atau adanya batasan minimum *n* seperti hanya terdapat satu pengamatan pada tiap simpul anak. Menurut Breiman (Otok, 2009: XVI-3), pada umumnya jumlah kasus minimum dalam suatu terminal akhir adalah 5, dan

apabila hal itu terpenuhi maka pengembangan pohon dihentikan. Sementara itu, menurut Steinberg dan Colla (Otok, 2009: XVI-3), jumlah kasus yang terdapat dalam simpul terminal yang homogen adalah kurang dari 10 kasus.

3. Penandaan Label Kelas

Penandaan label kelas pada simpul terminal dilakukan berdasarkan aturan jumlah terbanyak. Misalkan pada kasus klasifikasi keputusan pembelian komputer (ya, tidak), dalam salah satu simpul terminal yang dihasilkan terdapat jumlah keputusan ya dan keputusan tidak. Jumlah terbanyak dari keputusan tersebut dijadikan label kelas simpul terminal.

4. Penentuan Pohon Klasifikasi Optimal

Pohon klasifikasi yang berukuran besar akan memberikan nilai penaksir pengganti paling kecil, sehingga pohon ini cenderung dipilih untuk menaksir nilai dari variabel respon. Tetapi ukuran pohon yang besar akan menyebabkan nilai kompleksitas yang tinggi, karena struktur data yang digambarkan cenderung kompleks, sehingga perlu dipilih pohon optimal yang berukuran sederhana tetapi memberikan nilai penaksir pengganti cukup kecil. Ada dua jenis penaksir pengganti, yaitu penaksir sampel uji (*test sample estimate*) dan penaksir validasi silang lipat (*cross validation K-fold estimate*).

Validasi silang merupakan salah satu teknik untuk menduga *error rate.* Beberapa teknik yang lain diantaranya adalah: *holdout, leave one* dan *bootstrapping. K-fold cross validation* membagi data menjadi *k* bagian terpisah, satu data menjadi data *testing* dan *k-1* bagian menjadi data *training* sehingga terdapat *k* pasang data *training-testing. K-fold cross validation* dapat digunakan untuk data berukuran kecil ataupun besar. Aspek terpenting dalam validasi silang adalah kestabilan dari penaksiran yang diperoleh. Kestabilan pohon dapat bernilai rendah, jika mengandung terlalu banyak variabel prediktor.

Salah satu cara untuk mendapatkan pohon optimum yaitu dengan pemangkasan (*pruning*). Pemangkasan dilakukan dengan jalan memangkas bagian pohon yang kurang penting sehingga didapatkan pohon optimal. Ukuran

pemangkasan yang digunakan untuk memperoleh ukuran pohon yang layak adalah *cost complexity minimum.*

Sebagai ilustrasi, untuk sembarang pohon T yang merupakan sub pohon dari pohon terbesar $T_{max}(T < T_{max})$ ukuran *cost complexity* yaitu:

$$R_\alpha(T_k) = R(T_k) + \alpha|\tilde{T}_k| \quad (3.11)$$

dengan:

$R(T_k)$ = tingkat kesalahan klasifikasi dari pohon bagian $T_k$untuk k =1

$\tilde{T}_k$ = himpunan simpul terminal pada T_k

$|\tilde{T}_k|$ = banyak simpul terminal pada $\tilde{T}_k$

α = parameter *cost-complexity*

Untuk *binary tree*, parameter *cost-complexity* bernilai 0,5 yang berarti sebuah simpul selalu dikembangkan menjadi dua simpul anak. Tingkat kesalahan klasifikasi (*misclassification error*) pada simpul *t* dinyatakan dengan:

$$error\ (t) = 1 - {}^{max}_{i} P(i|t) \quad (3.12)$$

Contoh menghitung *misclassification error* jika sebuah simpul sudah diketahui:

$C1: 2\ \ C2: 6$

$$P(C1) = \frac{2}{8} = 0{,}25 \qquad P(C2) = \frac{6}{8} = 0{,}75$$

$$error = 1 - \max(0{,}25\ ;\ 0{,}75)$$
$$= 1 - 0{,}75$$
$$= 0{,}25$$

3.6 Contoh Kasus Pembentukan Pohon Keputusan dengan Algoritma CART

Data keputusan pembelian komputer

Age	Income	Student	Credit_rating	Class: buys_computer
Youth	High	No	Fair	No
Senior	High	No	Fair	Yes

Senior	Low	No	Fair	Yes
Senior	High	Yes	Fair	Yes
Youth	Low	Yes	Excellent	Yes

Klasifikasi dibagi menjadi dua kelas, yaitu:

C0 : *No* dan C1 : *Yes*

Atribut *age* mempunyai dua kemungkinan nilai yaitu {*youth, senior*}, dimana masing-masing nilai dapat diuraikan sebagai berikut:

1) *Record* yang mempunyai atribut *age=youth* ada 2; 1 *record* dikelas *No* (*record* ke-1) dan 1 *record* dikelas *Yes* (*record* ke-5), berarti C0 : 1 dan C1 : 1. Besarnya Indeks Gini dari simpul ini (A), adalah:

$$P(C0) = \frac{1}{2} \qquad P(C1) = \frac{1}{2}$$

$$Gini(A) = 1 - \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right)$$

$$= 0{,}5$$

2) *Record* yang mempunyai atribut *age=senior* ada 3; ketiganya ada dikelas *Yes* (*record* ke-2, ke-3, dan ke-4), berarti C0 : 0 dan C1 : 3. Besarnya Indeks Gini dari simpul ini (B), adalah:

$$P(C0) = \frac{0}{3} \qquad P(C1) = \frac{3}{3}$$

$$Gini(A) = 1 - \left(\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2\right)$$

$$= 0$$

Selanjutnya, hitung $Gini_{split}$ untuk atribut *age*:

$$Gini\ (age) = \frac{2}{5} \times 0{,}5 + \frac{3}{5} \times 0$$

$$= 0{,}20$$

Atribut *income* mempunyai dua kemungkinan nilai yaitu {*high*, *low*}, dimana masing-masing nilai dapat diuraikan sebagai berikut:

1) *Record* yang mempunyai atribut *income=high* ada 3; 1 *record* dikelas *No* (*record* ke-1) dan 2 *record* dikelas *Yes* (*record* ke-2 dan ke-4), berarti C0 : 1 dan C1 : 2. Besarnya Indeks Gini dari simpul ini (A), adalah:

$$P(C0) = \frac{1}{3} \qquad P(C1) = \frac{2}{3}$$

$$Gini(A) = 1 - \left(\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right)$$

$$= 0{,}44$$

2) *Record* yang mempunyai atribut *income=low* ada 2; keduanya berada dikelas *Yes* (*record* ke-3 dan ke-5), berarti C0 : 0 dan C1 : 2. Besarnya Indeks Gini dari simpul ini (B), adalah:

$$P(C0) = \frac{0}{2} \qquad P(C1) = \frac{2}{2}$$

$$Gini(A) = 1 - \left(\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right)$$

$$= 0$$

Selanjutnya, hitung $Gini_{split}$ untuk atribut *income*:

$$Gini(income) = \frac{3}{5} \times 0{,}44 + \frac{2}{5} \times 0$$

$$= 0{,}27$$

Atribut *student* mempunyai dua kemungkinan nilai yaitu {*no*, *yes*}, dimana masing-masing nilai dapat diuraikan sebagai berikut:

1) *Record* yang mempunyai atribut *student=no* ada 3; 1 *record* dikelas *No* (*record* ke-1) dan 2 *record* dikelas *Yes* (*record* ke-2 dan ke-3), berarti C0 : 1 dan C1 : 2. Besarnya Indeks Gini dari simpul ini (A), adalah:

$$P(C0) = \frac{1}{3} \qquad P(C1) = \frac{2}{3}$$

$$Gini(A) = 1 - \left(\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right)$$
$$= 0{,}44$$

2) *Record* yang mempunyai atribut *student=yes* ada 2; keduanya berada dikelas *Yes* (*record* ke-4 dan ke-5), berarti C0 : 0 dan C1 : 2. Besarnya Indeks Gini dari simpul ini (B), adalah:

$$P(C0) = \frac{0}{2} \qquad P(C1) = \frac{2}{2}$$
$$Gini(A) = 1 - \left(\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right)$$
$$= 0$$

Selanjutnya, hitung $Gini_{split}$ untuk atribut *student*:

$$Gini(student) = \frac{3}{5} \times 0{,}44 + \frac{2}{5} \times 0$$
$$= 0{,}27$$

Atribut *credit_rating* mempunyai dua kemungkinan nilai yaitu {*fair*, *excellent*}, dimana masing-masing nilai dapat diuraikan sebagai berikut:

1) *Record* yang mempunyai atribut *credit_rating=fair* ada 4; 1 *record* dikelas *No* (*record* ke-1) dan 3 *record* dikelas *Yes* (*record* ke-2, ke-3, dan ke-4), berarti C0 : 1 dan C1 : 3. Besarnya Indeks Gini dari simpul ini (A), adalah:

$$P(C0) = \frac{1}{4} \qquad P(C1) = \frac{3}{4}$$
$$Gini(A) = 1 - \left(\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right)$$
$$= 0{,}38$$

2) *Record* yang mempunyai atribut *credit_rating=excellent* ada 1; *record* berada dikelas *Yes* (*record* ke-5), berarti C0 : 0 dan C1 : 1. Besarnya Indeks Gini dari simpul ini (B), adalah:

$$P(C0) = \frac{0}{1} \qquad P(C1) = \frac{1}{1}$$

$$Gini(A) = 1 - \left(\left(\frac{0}{1}\right)^2 + \left(\frac{1}{1}\right)^2\right)$$
$$= 0$$

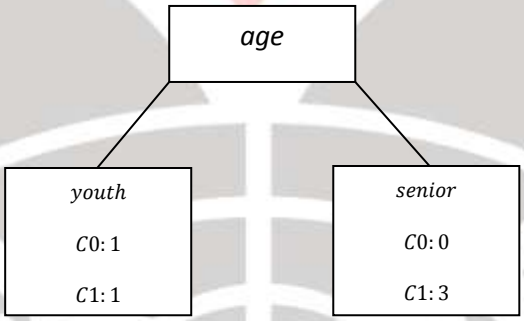
Selanjutnya, hitung $Gini_{split}$ untuk atribut *credit_rating*:

$$Gini(credit\_rating) = \frac{4}{5} \times 0{,}38 + \frac{1}{5} \times 0$$
$$= 0{,}30$$

Sehingga didapat matriks perhitungan sebagai berikut:

Kelas	X_1 (*age*)		X_2 (*income*)		X_3 (*student*)		X_4 (*credit_rating*)	
	youth	*senior*	*high*	*low*	*no*	*yes*	*fair*	*excellent*
No	$C0:1$	$C0:0$	$C0:1$	$C0:0$	$C0:1$	$C0:0$	$C0:1$	$C0:0$
Yes	$C1:1$	$C1:3$	$C1:2$	$C1:2$	$C1:2$	$C1:2$	$C1:3$	$C1:1$
Gini	0,20		0,27		0,27		0,30	

Dari keempat atribut tersebut, nilai Gini atribut *age* paling kecil sehingga dipilih sebagai pemilah pertama. Pohon keputusan sementara menjadi:



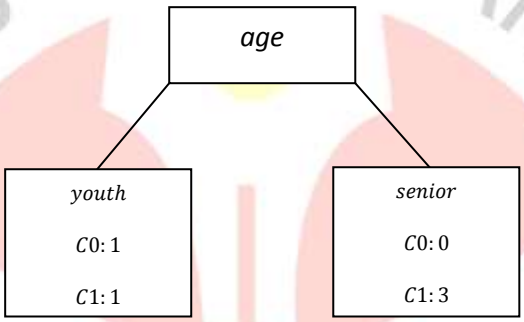
Gambar 3.2
Pohon Keputusan Sementara

Dari kedua simpul atribut *age*, simpul *youth* belum homogen sehingga perlu memilih calon pemilah selanjutnya (*income*, *student*, *credit_rating*) untuk data di record 1 dan 5. Dengan langkah yang sama seperti di atas, maka diperoleh matriks perhitungan sebagai berikut:

Kelas	X_2 (*income*)	X_3 (*student*)	X_4 (*credit_rating*)

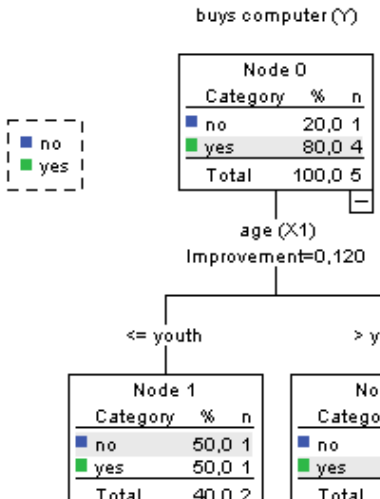
	high	*low*	*no*	*yes*	*fair*	*excellent*
No	$C0:1$	$C0:0$	$C0:1$	$C0:0$	$C0:1$	$C0:0$
Yes	$C1:0$	$C1:1$	$C1:0$	$C1:1$	$C1:0$	$C1:1$
Gini	0		0		0	

Karena nilai Indeks Gini semua simpul sudah nol, artinya setiap *record* dalam simpul berada dalam kelas yang sama (homogen) maka proses pembuatan pohon dihentikan sehingga didapatkan pohon optimum sebagai berikut:



Gambar 3.3
Pohon Keputusan Optimum

Pengolahan data menggunakan SPSS diperoleh pohon optimum sebagai berikut:



Gambar 3.4
Pohon Klasifikasi Optimum

Berdasarkan pohon optimum yang diperoleh dari kedua proses pengolahan data diatas, terlihat bahwa variabel yang berpengaruh secara signifikan dalam klasifikasi pembelian komputer adalah variabel X_1 (*age*) serta menghasilkan dua simpul terminal. Pada usia ≥ *youth*, terdapat satu orang yang memutuskan tidak membeli komputer dan satu orang yang memutuskan membeli komputer. Sedangkan untuk usia > *youth*, sebanyak tiga orang yang memutuskan untuk membeli komputer. Jadi dapat disimpulkan bahwa usia > *youth* lebih cenderung untuk membeli komputer.