

# Classification and Regression Tree (Cart) Analysis Pada Penderita Skizofrenia di RSJKO Soeprapto Daerah Bengkulu

Nurul Komariah<sup>1</sup>, Sigit Nugroho<sup>2</sup>, Jose Rizal<sup>2</sup>

<sup>1</sup>Alumni Jurusan Matematika Fakultas MIPA Universitas Bengkulu

<sup>2</sup>Staf Pengajar Jurusan Matematika Fakultas MIPA Universitas Bengkulu

## ABSTRAK

Penelitian ini bertujuan untuk mengetahui deskripsi, model pohon klasifikasi dan model pohon regresi pada penderita skizofrenia di Rumah Sakit Jiwa dan Ketergantungan Obat (RSJKO) Soeprapto Daerah Bengkulu. Metode yang digunakan adalah Analisis CART (*Classification and Regresssion Tree*). Analisis statistik nonparametrik ini dirancang untuk menyajikan kaidah keputusan berbentuk pohon biner dengan menggunakan prosedur penyekatan berulang. Analisis ini memiliki keunggulan dalam menangani data dengan struktur yang kompleks, efisien dalam terminologi perhitungan, sangat tangguh dalam menangani pencilan, dan dapat menangani sembarang data kontinu/ numerik dan kategorik. Pohon klasifikasi dibangun dengan menggunakan dua kaidah, yaitu kaidah pemisahan *Gini* dan kaidah pemisahan *Twoing*, sedangkan pohon regresi dibangun dengan menggunakan metode *validasi silang*. Hasil penelitian menunjukkan bahwa 81.7% pasien merupakan pasien dengan diagnosa skizofrenia paranoid. Sisanya sebanyak 18.3% berturut-turut adalah skizofrenia residual, hebefrenik, katatonik dan tidak terinci. Nilai ketepatan klasifikasi dengan menggunakan kaidah *Gini* adalah 86.9%, sedangkan dengan kaidah *Twoing* adalah 85.6%. Pada pohon regresi diperoleh proporsi varian yang dijelaskan oleh model sebesar 89.7%. Ini menunjukkan bahwa model pohon klasifikasi dan pohon regresi yang dibangun adalah cukup baik.

Kata kunci : *Skizofrenia, CART, Gini, Twoing, Validasi silang*

## PENDAHULUAN

Skizofrenia merupakan penyakit otak yang timbul akibat ketidakseimbangan pada *dopamine*, yaitu salah satu sel kimia dalam otak. Penyakit ini ditandai dengan adanya disorganisasi kepribadian yang cukup parah, distorsi realita dan ketidakmampuan berinteraksi dalam kehidupan sehari-hari<sup>[1]</sup>.

Penderita skizofrenia memiliki frasa-frasa kata yang hanya dapat dimengerti oleh dirinya sendiri dan seringkali merasa “aneh” dengan bagian tubuh mereka sendiri. Pada umumnya penderita tidak merasakan emosi apa-apa serta tidak mampu merespon stimulus emosi dengan benar. Mereka seringkali menunjukkan aktivitas motorik dan ekspresi wajah yang aneh. Selain itu penderita juga melakukan gerakan yang tak lazim tanpa berhenti atau mempertahankannya dalam periode waktu yang lama<sup>[1]</sup>.

Penderita skizofrenia khususnya di Kota Bengkulu seringkali ditelantarkan oleh keluarganya sehingga hidup sendiri atau berkeliaran di jalan tanpa ada perhatian dan penanganan khusus. Bahkan tidak sedikit yang menitipkan anggota keluarga mereka yang mengidap skizofrenia tersebut ke rumah sakit jiwa. Satu-satunya rumah sakit yang

menangani para penderita skizofrenia di Provinsi Bengkulu adalah *Rumah Sakit Jiwa dan Ketergantungan Obat (RSJKO) Soeprapto Daerah Bengkulu*.

Penderita skizofrenia yang dirawat di RSJKO Soeprapto Daerah Bengkulu memiliki latar belakang dan karakteristik yang berbeda-beda. Perbedaan ini meliputi jenis kelamin, umur, tingkat pendidikan, status perkawinan, dan lain-lain. Demikian pula halnya dengan jenis skizofrenia yang mereka derita. Untuk mengetahui deskripsi penderita skizofrenia yang dirawat di RSJKO Soeprapto Daerah Bengkulu, lamanya mereka menjalani perawatan, serta jenis-jenis skizofrenia yang paling banyak ditemukan pada penderita tersebut, maka dalam tulisan ini akan dibuat pohon klasifikasi dan pohon regresi menggunakan Analisis CART (*Classification and Regression Tree*) guna mendeskripsikan hal tersebut di atas.

Tujuan penelitian ini adalah untuk mengetahui deskripsi pasien skizofrenia, mengetahui model pohon klasifikasi (*Classification Tree*) diagnosa skizofrenia yang diderita, serta model pohon regresi (*Regression Tree*) lamanya penderita skizofrenia dirawat inap di RSJKO Soeprapto Daerah Bengkulu.

## TINJAUAN PUSTAKA

Classification and Regression Trees adalah metode klasifikasi menggunakan data historis untuk membangun suatu pohon keputusan. Metodologi CART mulai dikembangkan pada tahun 80-an oleh Breiman, Friedman, Olshen, dan Stone dalam makalah mereka yang berjudul "*Classification and Regression Trees*" (1984).

CART adalah suatu analisis diskriminan non-parametrik yang dirancang untuk menyajikan kaidah keputusan berbentuk pohon biner yang membagi data pada learning sampel dalam batasan linier univariat. Analisis ini menghasilkan kelompok data hirarkis yang dimulai dari *node root* untuk keseluruhan learning sampel dan berakhir pada kelompok kecil pengamatan yang homogen. Pada setiap *terminal node* diberikan label kelas atau nilai yang diramalkan, sehingga menghasilkan struktur pohon yang dapat ditafsirkan sebagai pohon keputusan<sup>[11]</sup>.

Keuntungan dari penggunaan analisis CART<sup>[11]</sup> adalah sebagai berikut :

1. Merupakan bentuk statistika non-parametrik, sehingga tidak memerlukan asumsi sebaran dan uji hipotesis.
2. Tidak memerlukan variabel untuk dipilih sebelumnya.
3. Sangat efisien dalam terminologi perhitungan.
4. Dapat menangani dataset dengan struktur yang kompleks.
5. Sangat tangguh dalam menangani outlier, umumnya algoritma pemisahan akan mengisolasi outlier pada individu node atau beberapa node.
6. Dapat menggunakan sembarang kombinasi data kontinu/numerik dan kategorik.
7. Hasilnya invarian dengan transformasi monoton dari variabel respon, artinya penggantian sembarang variabel dengan algoritmanya atau nilai akar kuadrat, tidak akan menyebabkan struktur pohon berubah.

## Pohon Klasifikasi

*Classifier*<sup>[12]</sup> atau kaidah klasifikasi adalah suatu cara sistematis dalam memprediksi suatu kasus masuk dalam kelas tertentu. Untuk memberikan formulasi yang lebih tepat, maka disusun suatu himpunan pengukuran  $(x_1, x_2, \dots, x_n)$  sebagai faktor pengukuran (*measurement*

vector). Semua vektor pengukuran yang mungkin berada di dalamnya didefinisikan sebagai ruang pengukuran  $X$ .

Andaikan suatu kasus atau objek mempunyai  $J$  kelas yaitu  $1, 2, \dots, j$  dan misalkan  $C$  adalah himpunan kelas tersebut dengan  $C = \{1, 2, \dots, j\}$ . Suatu cara sistematis dalam memprediksi anggota kelas tersebut adalah dengan menggunakan suatu aturan yang menempatkan anggota kelas dalam  $C$  tersebut pada setiap vektor pengukuran  $x$  dalam  $X$ .

**Definisi 1.**

“Suatu *classifier* atau aturan klasifikasi adalah suatu fungsi  $d(x)$  pada  $X$  sehingga untuk setiap  $x$ ,  $d(x)$  adalah sama dengan salah satu dari  $\{1, 2, \dots, j\}$ .”

Cara lain untuk melihat *classifier* adalah dengan mendefinisikan  $A_j$  sebagai subset dari  $X$  dimana  $d(x)$  sama dengan  $j$  sehingga  $A_j = \{x ; d(x) = j\}$ .

Himpunan  $A_1, \dots, A_j$  adalah disjoint dan

$$X = \bigcup_j A_j$$

sehingga  $A_j$  adalah bentuk partisi dari  $X$ .

**Definisi 2.**

“*Classifier* adalah suatu partisi pada  $X$  dalam  $J$  yang memisahkan himpunan bagian  $A_1, \dots, A_j \ni X = \bigcup A_j$ . Sehingga untuk setiap  $x \in A_j$  kelas prediksinya adalah  $j$ .”

Dalam konstruksi klasifikasi sistematis, semua data historis dirangkum dalam suatu *learning sample*. Learning sampel merupakan sampel data yang digunakan untuk membangun pohon klasifikasi.

Sehingga diperoleh definisi rumus-rumus sebagai berikut (gambar 2) :

1. Proporsi pengamatan pada kelas ke- $j$  terhadap jumlah keseluruhan pengamatan.

$$p(j) = \frac{N_j}{N} \tag{1}$$

2. Jumlah pengamatan pada kelas ke- $j$ .

$$N_j = \sum_{s=1}^k N_j(s) \tag{2}$$

3. Jumlah pengamatan pada node  $s$ .

$$N(s) = \sum_{j=1}^k N_j(s) \tag{3}$$

4. Peluang pengamatan pada node  $s$ .

$$p(s) = \frac{N(s)}{N} \tag{4}$$

5. Peluang bersama pengamatan pada node  $s$  kelas ke- $j$ .

$$p(j,s) = \frac{N_j(s)}{N} \tag{5}$$

6. Peluang bersyarat pengamatan pada node  $s$  kelas ke- $j$ .

$$p(j|s) = \frac{p(j, s)}{p(s)} = \frac{\frac{N_j(s)}{N}}{\frac{N(s)}{N}} = \frac{N_j(s)}{N(s)} \quad (6)$$

Dari persamaan (3) dan (6) diperoleh rumus berikut :

$$\begin{aligned} p(1|s) + p(2|s) + p(3|s) + \dots + p(k|s) &= \sum_{j=1}^k p(j|s) = \sum_{j=1}^k \frac{N_j(s)}{N(s)} \\ &= \frac{1}{N(s)} \cdot N(s) = 1 \end{aligned}$$

Suatu ukuran impurity pada node  $t$  disimbolkan dengan  $i(t)$ , dimana  $i(t)$  merupakan suatu fungsi peluang kelas  $p(1|t), p(2|t), \dots, p(k|t)$ . Sehingga secara matematis dapat dituliskan dengan :

$$i(t) = f [p(1|t), p(2|t), \dots, p(k|t)] \quad (7)$$

### Definisi 3.

“*Impurity function* adalah suatu fungsi  $f$  yang didefinisikan pada himpunan  $(p_1, p_2, \dots, p_k)$  yang memenuhi  $p_j \geq 0, j = 1, \dots, k, \sum_j p_j = 1$  dengan kriteria sebagai berikut :

1.  $f$  akan *maksimum unik* pada titik  $(\frac{1}{k}, \frac{1}{k}, \dots, \frac{1}{k})$ . Dengan kata lain, masing-masing kelas dalam populasi memiliki peluang yang sama.
2.  $f$  akan *minimum unik* pada titik  $(1, 0, 0, \dots, 0), (0, 1, 0, \dots, 0), \dots, (0, 0, 0, \dots, 1)$ .
3.  $f$  adalah fungsi simetrik dari  $p_1, p_2, \dots, p_k$ .”

Misalkan jumlah terminal node pada pohon klasifikasi adalah  $\tilde{T}$  dan diketahui himpunan  $I(t) = i(t) \cdot p(t)$ . Rumus impurity pohon (*tree impurity*) didefinisikan dengan :

$$I(\tilde{T}) = \sum_{t \in \tilde{T}} I(t) = \sum_{t \in \tilde{T}} i(t) \cdot p(t) \quad (8)$$

Suatu pohon klasifikasi dibangun berdasarkan aturan pemisahan (*splitting rule*), yaitu aturan yang memisahkan learning sampel ke dalam bagian yang lebih kecil. Setiap kali data yang ada harus dibagi menjadi dua bagian dengan homogenitas maksimum.

Anggap  $t$  adalah node ayah yang dipisahkan oleh pembagian  $x$  menjadi dua node anak, yaitu node anak kiri  $t_L$  dan node anak kanan  $t_R$ . Masing-masing node anak tersebut mempunyai peluang,  $p_L$  dan  $p_R$  dengan  $p_R = 1 - p_L$  (gambar 3). Pada sembarang terminal node, akan dipilih pembagian yang paling mengurangi nilai  $I(t)$  dengan kata lain akan akuivalen dengan memaksimalkan perubahan fungsi impurity node  $t$  pada  $x$  sebagai berikut :

$$\Delta I(x, t) = I(t) - I(t_L) - I(t_R)$$

atau

$$\Delta I(x, t) = I(t) - p_L \cdot I(t_L) - p_R \cdot I(t_R) \quad (9)$$

Nilai pemisahan terbaik  $\Delta i(x, t)$  menunjukkan perubahan impurity node  $t$  pada  $x$  dengan  $t_L \cup t_R = 1$ . Oleh karena itu, pemisahan terbaik dari  $t$  adalah :

$$\arg \max_x \Delta i(x, t) = \arg \max_x (i(t) - p_L i(t_L) - p_R i(t_R)) \quad (10)$$

Nilai optimal  $x^*$  dapat ditentukan dengan cara memaksimalkan  $\Delta i(x, t)$  dengan  $x$  yang berbeda pada masing-masing node  $t$ . Prosedur semacam ini memungkinkan untuk membangun pohon keputusan dari sembarang bentuk pohon maksimum.

Karena nilai  $i(t)$  pada kenyataannya adalah konstan, sehingga hasilnya akan ekuivalen dengan :

$$\begin{aligned} x^* = \arg \max_x \Delta i(x, t) &= \arg \max_x (-p_L i(t_L) - p_R i(t_R)) \\ &= \arg \min_x (p_L i(t_L) + p_R i(t_R)) \end{aligned} \quad (11)$$

dimana  $t_L$  dan  $t_R$  adalah fungsi eksplisit dari  $x$ .

Di dalam teori ada beberapa fungsi impurity, tetapi yang secara luas digunakan dalam prakteknya, yaitu Kaidah Pemisahan Gini (*Gini Splitting Rule*) dan Kaidah Pemisahan Twoing (*Twoing Splitting Rule*).

### Gini Splitting Rule

*Gini Splitting Rule* atau disebut juga indeks Gini (*Gini Index*) adalah kaidah yang paling umum digunakan dalam memecahkan permasalahan pohon klasifikasi. Data impurity didefinisikan dengan menggunakan ukuran varian (*variance measure*). Misalkan 1 adalah semua pengamatan pada node  $t$  kelas ke- $j$  dan 0 untuk yang lainnya. Kemudian estimasi varian contoh untuk node  $t$  pengamatan sebagai berikut :

$$p(j|t)(1 - p(j|t))$$

Indeks Gini pada node  $t$  kelas ke- $j$  didefinisikan dengan rumus sebagai berikut :

$$\begin{aligned} i(t) &= \sum_{j=1}^k p(j|t) (1 - p(j|t)) = \sum_{j=1}^k p(j|t) - p^2(j|t) \\ &= \sum_{j=1}^k p(j|t) - \sum_{j=1}^k p^2(j|t) \\ &= 1 - \sum_{j=1}^k p^2(j|t) \end{aligned} \quad (12)$$

Sehingga diperoleh perubahan fungsi impurity node  $t$  oleh pemisahan  $x$  sebagai berikut :

$$\Delta i(x, t) = - \sum_{j=1}^k p^2(j|t) + p_L \cdot \sum_{j=1}^k p^2(j|t_L) + p_R \cdot \sum_{j=1}^k p^2(j|t_R) \quad (13)$$

Pemisahan terbaik dari  $t$  dirumuskan dengan :

$$\arg \max_x \Delta i(x, t) = \arg \max_x \left\{ - \sum_{j=1}^k p^2(j|t) + p_L \cdot \sum_{j=1}^k p^2(j|t_L) + p_R \cdot \sum_{j=1}^k p^2(j|t_R) \right\} \quad (14)$$

Indeks Gini akan mencari learning sampel untuk kelas paling besar dan mengisolasinya dari sisa data tersebut. Kaidah ini bekerja dengan baik pada data berukuran besar.

### Twoing Splitting Rule

Tidak seperti kaidah *Gini*, kaidah *Twoing* tidak mencari nilai maksimal dari ukuran impurity. Sebagai gantinya kaidah ini mencoba untuk menyeimbangkan konstruksi pohon dengan cara seolah-olah membagi learning sampel menjadi dua kelas. Sehingga pengamatan dapat dibedakan antara faktor umum yang berada pada tingkat teratas dan karakteristik khusus yang berada pada tingkat yang lebih rendah.

Misalkan terdapat himpunan kelas learning sampel  $C = \{1, 2, \dots, k\}$ . Himpunan tersebut dibagi menjadi dua bagian yaitu:  $C_1 = \{c_1, c_2, \dots, c_n\}$  dan  $C_2 = C \setminus C_1$  sedemikian sehingga semua pengamatan yang berada pada  $C_1$  mempunyai kelas dummy 1, sedangkan sisanya mempunyai kelas dummy 2<sup>[11]</sup>.

Kemudian akan dihitung nilai  $\Delta I(x, t)$  untuk  $x$  yang berbeda “jika hanya ada dua kelas dummy”. Karena nilai  $\Delta I(x, t)$  bergantung pada  $C_1$ , maka nilai  $\Delta I(x, t, C_1)$  adalah maksimal. Dengan kata lain, kaidah *Twoing* adalah suatu aturan yang digunakan untuk menemukan kombinasi superkelas pada setiap node seolah-olah kenaikan impurity telah dimaksimalkan hanya oleh dua kelas  $C = \{1, 2\}$ .

Walaupun kaidah *Twoing* dapat diterapkan terutama untuk data dengan jumlah kelas yang besar, kelemahannya terdapat pada kecepatan perhitungan. Asumsikan bahwa learning sampel mempunyai  $J$  kelas, kemudian himpunan  $C$  dipisahkan menjadi  $C_1$  dan  $C_2$  dengan  $2^{J-1}$  cara. Pada kasus dimana terdapat 11 data kelas pada learning sampel, maka akan terbentuk 1024 kombinasi.

Pada kaidah *twoing*, tidak ada ukuran impurity yang spesifik<sup>[25]</sup>. Sehingga untuk sembarang node, pemisahan yang terbaik ditentukan dengan cara memaksimalkan perubahan impurity pada node anak kanan  $t_R$  dan node anak kiri  $t_L$ . Ini mengakibatkan timbulnya perbedaan definisi kaidah *twoing* oleh para peneliti, antara lain :

1. Chee Jen Chang (2002)

$$I(t) = \frac{n_L \cdot n_R}{4} \left( \sum_j |p(j|t_L) - p(j|t_R)| \right)^2 \quad (15)$$

2. David Feldman (2003)

$$d_r(t) = \frac{n_L \cdot n_R}{4} \sum_{j=1}^J |p(j|t_L) - p(j|t_R)| \quad (16)$$

3. Roman Timofeev (2004)

$$\Delta I(t) = \frac{n_L \cdot n_R}{4} \left[ \sum_{j=1}^J |p(j|t_L) - p(j|t_R)| \right]^2 \quad (17)$$

4. Anton Andriyashin (2005)

$$S_L(t) = \{j \mid p(j|t_L) \geq p(j|t_R)\}$$

$$\max_{S_1} \Delta(x, t, S_1) = \frac{N_1 \cdot N_2}{4} \left[ \sum_{j=1}^J |p(j|t_1) - p(j|t_2)| \right]^2 \quad (18)$$

5. Ozge Sezgin (2006)

$$\Delta(x) = \frac{N_1 \cdot N_2}{4} \left[ \sum_{j=1}^J p(j|t_1) p(j|t_2) \right]^2 \quad (19)$$

6. Zambon, et al (2006)

$$\text{Twining}(\Delta) = \frac{N_1 \cdot N_2}{4} \left( \sum_{j=1}^J (p(j|t_1) - p(j|t_2)) \right)^2 \quad (20)$$

### Pohon Regresi

Konstruksi pohon pada pohon klasifikasi dan pohon regresi tidak jauh berbeda, yang membedakannya adalah jenis variabel responnya. Pohon regresi adalah jenis pohon keputusan dengan variabel respon kontinu. Tujuan utama pohon regresi adalah untuk menghasilkan struktur pohon prediktor atau kaidah prediksi. Prediktor ini menjalankan dua tujuan utama, yaitu : (1) untuk memprediksi ketelitian variabel respon atau nilai baru dari variabel prediktor, (2) untuk menjelaskan hubungan antara variabel respon dan variabel prediktor<sup>[27]</sup>.

Prediktor pada pohon regresi dibangun dengan mendeteksi heterogenitas (dalam kaitan dengan varian pada variabel prediktor) yang ada di dalam data tersebut. Pohon regresi melakukannya dengan penyekatan berulang data ke dalam kelompok atau terminal node yang secara internal lebih homogen dibandingkan node di atasnya. Pada masing-masing terminal node, nilai rata-rata dari variabel respon dianggap sebagai nilai prediksi.

Terdapat dua kaidah pemisahan atau fungsi impurity pada pohon regresi, yaitu *Least Square (LS) function* dan *the Least Absolute Deviation (LAD) function*. Karena mekanisme kedua aturan tersebut sama, maka disini hanya akan dibahas ukuran impurity LS. Menggunakan kriteria ini, impurity node diukur dengan jumlah kuadrat node, **SS(t)** yang didefinisikan sebagai berikut :

$$SS(t) = \sum_{i=1}^{n(t)} (y_{(i)} - \bar{y}_{(t)})^2 \quad (21)$$

Misalkan diberikan fungsi impurity, **SS(t)** yang dipisahkan oleh  $x$  pada node kiri  $t_L$  dan node kanan  $t_R$ . Sehingga pemisahan terbaik (*goodness of split*) dapat ditunjukkan dengan rumus sebagai berikut :

$$\arg\max_x f(x, t) = \arg\max_x (SS(t) - SS(t_L) - SS(t_R)) \quad (22)$$

Alternatif lain dari fungsi impurity pada pohon regresi adalah dengan menggunakan nilai varian dari node anak kiri dan node anak kanan. Nilai varian pada suatu node  $t$  didefinisikan dengan rumus berikut :

$$s^2(t) = \frac{1}{n(t)} \sum_{i=1}^{n(t)} [y_{(i)} - \bar{y}_{(t)}]^2 \quad (23)$$

Sehingga diperoleh perubahan fungsi impurity node  $t$  oleh pemisahan  $x$  berikut :

$$\Delta x^2(t) = x^2(t) - p_L \cdot x^2(t_L) - p_R \cdot x^2(t_R) \quad (24)$$

Nilai perubahan impurity  $\Delta x^2(t)$  menunjukkan nilai pemisahan terbaik node  $t$  oleh pemisahan  $x$ , yang dirumuskan dengan :

$$\underset{x}{\operatorname{argmax}} (\Delta x^2(t)) = \underset{x}{\operatorname{argmax}} (x^2(t) - p_L \cdot x^2(t_L) - p_R \cdot x^2(t_R)) \quad (25)$$

Sama halnya dengan pohon klasifikasi, nilai optimal  $x^*$  dapat ditentukan dengan cara memaksimalkan  $\Delta x^2(t)$  dengan  $x$  yang berbeda pada masing-masing node  $t$ . Karena nilai  $x^2(t)$  adalah konstan, sehingga hasilnya akan ekuivalen dengan :

$$\begin{aligned} x^* &= \underset{x}{\operatorname{argmax}} (\Delta x^2(t)) = \underset{x}{\operatorname{argmax}} (-p_L \cdot x^2(t_L) - p_R \cdot x^2(t_R)) \\ &= \underset{x}{\operatorname{argmin}} (p_L \cdot x^2(t_L) + p_R \cdot x^2(t_R)) \end{aligned} \quad (26)$$

Sehingga diperoleh kesimpulan bahwa pemisahan yang terbaik adalah nilai  $\Delta x^2(t)$  yang paling tinggi dengan jumlah varian  $(p_L \cdot x^2(t_L) + p_R \cdot x^2(t_R))$  yang paling rendah. Prosedur ini menghasilkan node anak kiri dan node anak kanan yang lebih homogen dibandingkan node ayahnya. Masing-masing bagian membagi pengamatan pada node anak kiri dan node anak kanan sehingga nilai rata-rata dari variabel prediktor pada suatu terminal node adalah lebih rendah dibandingkan nilai rata-rata node ayah.

## METODE PENELITIAN

Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dengan metode dokumentasi, yaitu pengambilan data yang bersumber arsip registrasi pasien skizofrenia yang dirawat di masing-masing ruangan/bangsas di RSJKO Soeprapto Daerah Bengkulu periode 01 Januari 2007 sampai dengan 31 Oktober 2007. Adapun yang menjadi populasi dalam penelitian ini adalah seluruh pasien skizofrenia yang menjalani rawat inap di RSJKO Soeprapto Daerah Bengkulu yang dibagi dalam tiga ruangan, yaitu : ruang Anggrek, Murai, dan Rajawali.

Untuk membangun pohon klasifikasi dan pohon regresi, maka dibuat langkah-langkah pengerjaan sebagai berikut :

1. Data yang diperoleh dibuat dalam bentuk tabel pada program *Microsoft Excel*.
2. Untuk menentukan proporsi pada setiap variabel, digunakan software program *SPSS* versi 16.0. Dalam hal ini variabel berbentuk kategorik harus dibuat menjadi variabel *dummy* untuk memudahkan perhitungan.
3. Langkah selanjutnya adalah membuat model pohon klasifikasi dan pohon regresi.

Pohon Klasifikasi	Pohon Regresi
Respon ( $Y$ ) = diagnosa skizofrenia	Respon ( $Y$ ) = lamanya pasien dirawat
Prediktor ( $X_i$ ), $i = 1, 2, \dots, 8$	Prediktor ( $X_i$ ), $i = 1, 2, \dots, 8$
$X_1$ = jenis kelamin	$X_1$ = jenis kelamin
$X_2$ = finansial/ cara pembayaran	$X_2$ = finansial/ cara pembayaran
$X_3$ = umur	$X_3$ = umur
$X_4$ = pendidikan	$X_4$ = pendidikan
$X_5$ = pekerjaan	$X_5$ = pekerjaan
$X_6$ = status pernikahan	$X_6$ = status pernikahan



$X_7$ = daerah asal $X_8$ = lama dirawat	$X_7$ = daerah asal $X_8$ = diagnosa skizofrenia
---------------------------------------------	-----------------------------------------------------

## HASIL DAN PEMBAHASAN

Deskripsi pasien skizofrenia yang dirawat inap di RSJKO Soeprpto Bengkulu meliputi antara lain : jenis kelamin, finansial, diagnosa, tingkat pendidikan, status pekerjaan, status pernikahan, dan daerah asal. Hasil penelitian menunjukkan bahwa mayoritas pasien skizofrenia berjenis kelamin laki-laki sebanyak 80.4% dengan latar belakang kalangan tidak mampu. Ini dibuktikan dengan finansial menggunakan Surat Keterangan Tidak Mampu dengan persentase sebanyak 68.6%. Kemudian disusul dengan Askin, Askes dan Umum. Diagnosa jenis skizofrenia yang diderita adalah 81.7% skizofrenia paranoid, 9.2% skizofrenia residual, 7.8% skizofrenia hebefrenik, dan sisanya masing-masing 0.7% skizofrenia katatonik dan tidak terinci.

Sebaran tingkat pendidikan pasien skizofrenia cukup merata, antara lain : lulusan SD sebanyak 34.6%, lulusan SLTA sebanyak 26.1%, lulusan SLTP sebanyak 25.5%, dan lulusan S.1 sebanyak 2.6%. Sisanya sebanyak 10.5% tidak menempuh pendidikan formal. Ini berarti skizofrenia dapat diderita seseorang tanpa memandang latar belakang pendidikannya. Status pekerjaan dapat dilihat dari mayoritas pasien skizofrenia Tidak Bekerja dengan persentase sejumlah 60.1%. Status Tidak Bekerja disini dapat berupa ibu rumah tangga, mahasiswa, pelajar, atau pengangguran. Status pernikahan cukup beragam, yaitu Belum Kawin sebanyak 56%, Kawin sebanyak 35.9%, sisanya masing-masing 1.3% dan 2.6% memiliki status Janda dan Duda.

Daerah asal pasien skizofrenia tersebar merata di seluruh wilayah Provinsi Bengkulu. Jumlah terbanyak berasal dari Kota Bengkulu sebanyak 20.9%, Rejang Lebong 17%, Bengkulu Selatan 14.4% dan Bengkulu Utara 13.7%. Kabupaten-kabupaten pemekaran, daerah di luar propinsi Bengkulu, serta gepeng (gelandangan dan pengemis) memiliki jumlah yang tidak terlalu banyak.

Jumlah minimum *parent node* dan *child node* serta jumlah maksimum kedalaman pada pohon klasifikasi dan pohon regresi yang dibangun dapat ditentukan sebelumnya dengan menggunakan software SPSS versi 16.0. Pohon klasifikasi yang dibangun dengan menggunakan kaidah Gini memiliki 5 kedalaman maksimum, jumlah minimum *parent node* sebanyak 5 dan *child node* sebanyak 1. Hasil yang diperoleh adalah 23 node, yaitu 1 node root, 10 internal node dan 12 terminal node. Sedangkan pohon klasifikasi yang dibangun dengan menggunakan Kaidah Twoing memiliki 5 kedalaman maksimum, jumlah minimum *parent node* sebanyak 5 dan *child node* sebanyak 2. Hasil yang diperoleh adalah 19 node, yaitu 1 node root, 8 internal node dan 10 terminal node. Variabel yang paling berpengaruh pada pohon klasifikasi baik menggunakan kaidah Gini maupun Twoing adalah variabel "Daerah Asal".

Untuk menangani pohon dengan kompleksitas yang tinggi, maka dilakukan prosedur pembabatan pohon (*pruning tree*). Pada pohon klasifikasi yang dibangun dengan menggunakan kaidah Gini, diperoleh nilai ketepatan klasifikasi untuk keseluruhan data penelitian adalah 86.9 %. Setelah dilakukan *pruning tree*, nilainya adalah tetap. Pada pohon klasifikasi yang dibangun dengan menggunakan kaidah Twoing, diperoleh nilai ketepatan

klasifikasi untuk keseluruhan data penelitian adalah 86.9 %. Setelah dilakukan *pruning tree*, nilainya turun menjadi 85.6 % .

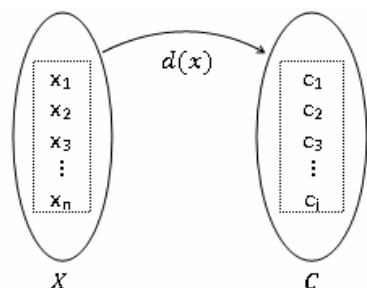
Pohon regresi yang dibangun dengan menggunakan metode *cross-validation* memiliki 5 kedalaman maksimum dengan jumlah minimum *parent node* sebanyak 2 dan *child node* sebanyak 1. Hasil yang diperoleh adalah 23 node, yaitu 1 node root, 10 internal node dan 12 terminal node. Variabel yang paling berpengaruh pada pohon regresi adalah variabel “Umur”. Pada pohon regresi diperoleh proporsi varian yang dijelaskan oleh model sebesar 89.7%. Setelah dilakukan *pruning tree*, nilainya turun menjadi 86.9%.

Penurunan nilai ketepatan klasifikasi dan proporsi varian yang dijelaskan oleh model setelah dilakukan *pruning tree* disebabkan karena ketelitian dan kekuratan pohon lebih tinggi sebelum dilakukan *pruning tree*. Nilai sebelum dilakukan *pruning tree* menunjukkan bahwa model pohon klasifikasi dan pohon regresi yang dibangun adalah cukup baik.

**KESIMPULAN**

Dari analisis hasil penelitian ini dapat diberikan kesimpulan sebagai berikut :

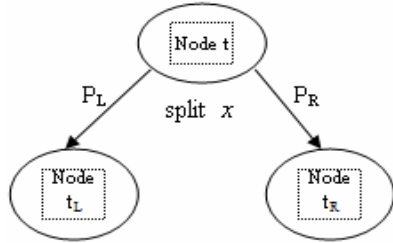
1. Mayoritas penderita skizofrenia di RSJKO adalah sebagai berikut :
  - a. Jenis kelamin laki-laki
  - b. Diagnosa Skizofrenia Paranoid
  - c. Menggunakan Surat Keterangan Tidak Mampu
  - d. Tingkat Pendidikan lulusan SD
  - e. Tidak memiliki Pekerjaan
  - f. Status Belum Kawin
  - g. Daerah Asal Kota Bengkulu
2. Variabel yang paling berpengaruh pada pohon klasifikasi adalah “Daerah Asal”.
3. Variabel yang paling berpengaruh pada pohon regresi adalah “Umur”.
4. Pohon klasifikasi dengan menggunakan kaidah Gini setelah dilakukan *pruning tree* memiliki ketepatan klasifikasi 86.9%, sedangkan kaidah Twoing 85.6%.
5. Proporsi varian yang dijelaskan oleh model pohon regresi adalah 89.7%, setelah *pruning tree* menjadi 86.9%.
6. *Pruning tree* dilakukan untuk menangani pohon dengan kompleksitas yang tinggi.



Gambar 1. Definisi *Classifier* pada suatu fungsi  $d(x)$

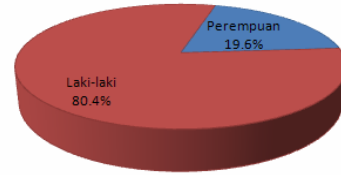
		Node						
		1	2	...	s	...	t	
Kelas	1				$N_1(s)$			
	2				$N_2(s)$			
	⋮				⋮			
	$j$	$N_j(1)$	$N_j(2)$	...	$N_j(s)$	...	$N_j(t)$	$N_j$
	⋮				⋮			
$k$				$N_k(s)$				
				$N(s)$			$N$	

Gambar 2. Ilustrasi Perhitungan Rumus Pohon Klasifikasi



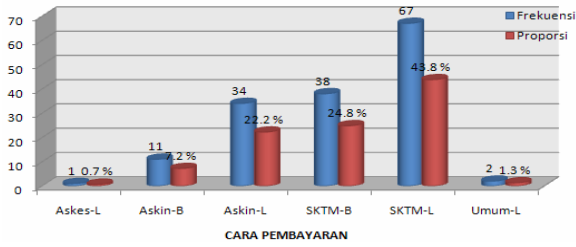
Gambar 3. Algoritma pemisahan pada CART

Proporsi Jenis Kelamin Pasien Skizofrenia



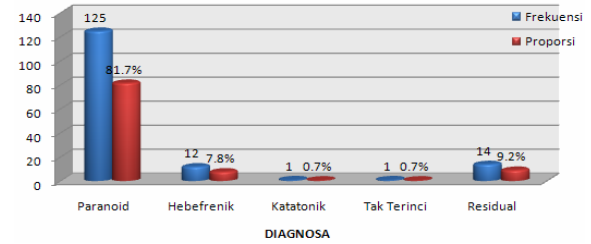
Gambar 4. Grafik Deskripsi Jenis Kelamin Pasien Skizofrenia

Deskripsi Finansial Pasien Skizofrenia



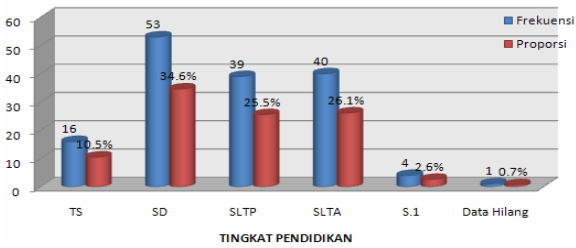
Gambar 5. Grafik Deskripsi Finansial Pasien Skizofrenia

Deskripsi Diagnosa Pasien Skizofrenia



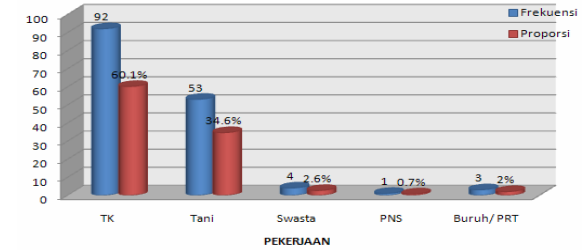
Gambar 6. Grafik Deskripsi Diagnosa Pasien Skizofrenia

Deskripsi Tingkat Pendidikan Pasien Skizofrenia



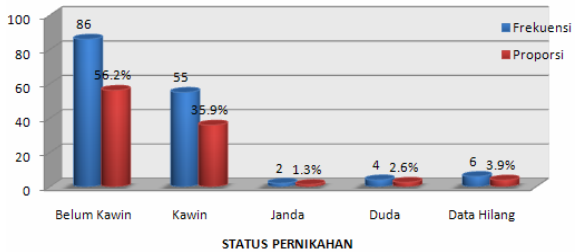
Gambar 7. Grafik Deskripsi Tingkat Pendidikan Pasien Skizofrenia

Deskripsi Status Pekerjaan Pasien Skizofrenia



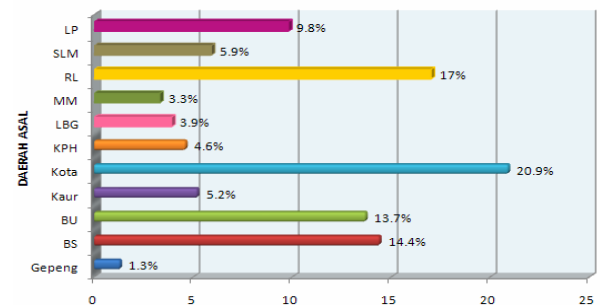
Gambar 8. Grafik Deskripsi Status Pekerjaan Pasien Skizofrenia

Deskripsi Status Pernikahan Pasien Skizofrenia



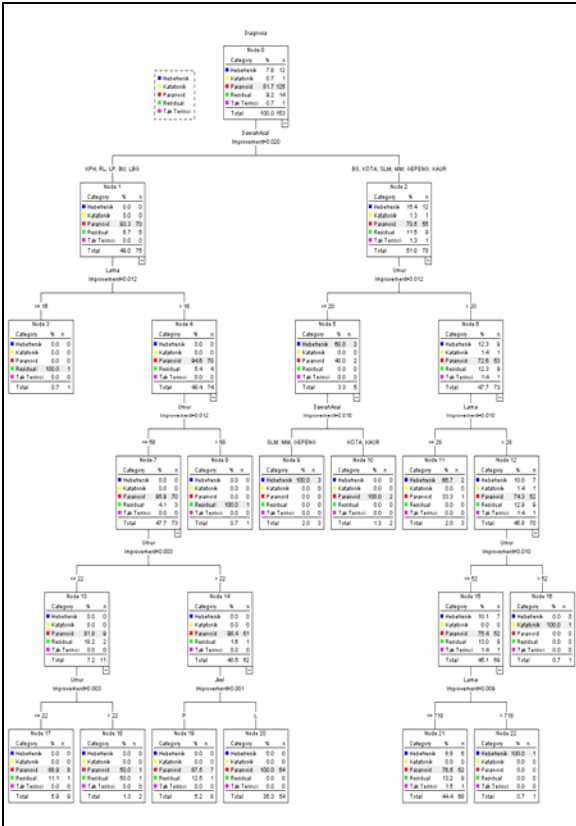
Gambar 9. Grafik Deskripsi Status Pernikahan Pasien Skizofrenia

Proporsi Daerah Asal Pasien Skizofrenia

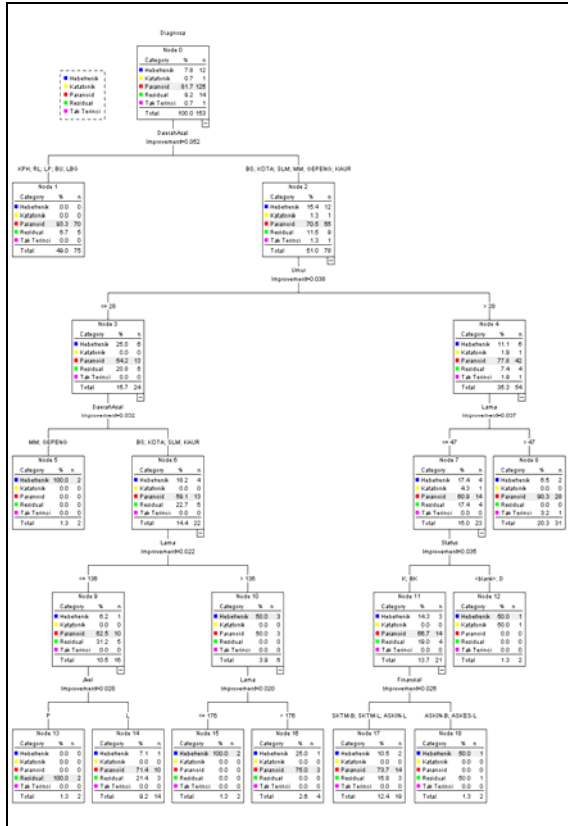


Gambar 10. Grafik Deskripsi Daerah Asal Pasien Skizofrenia

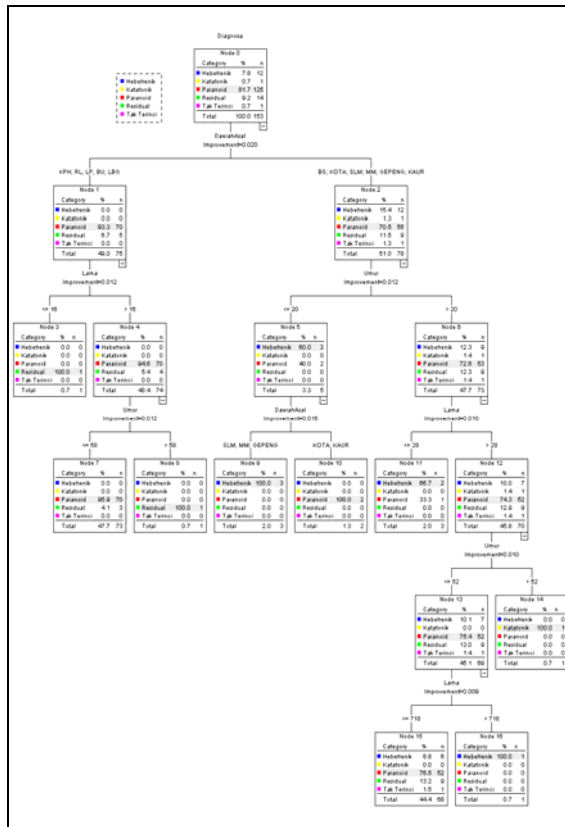
# Classification and Regression Tree ...



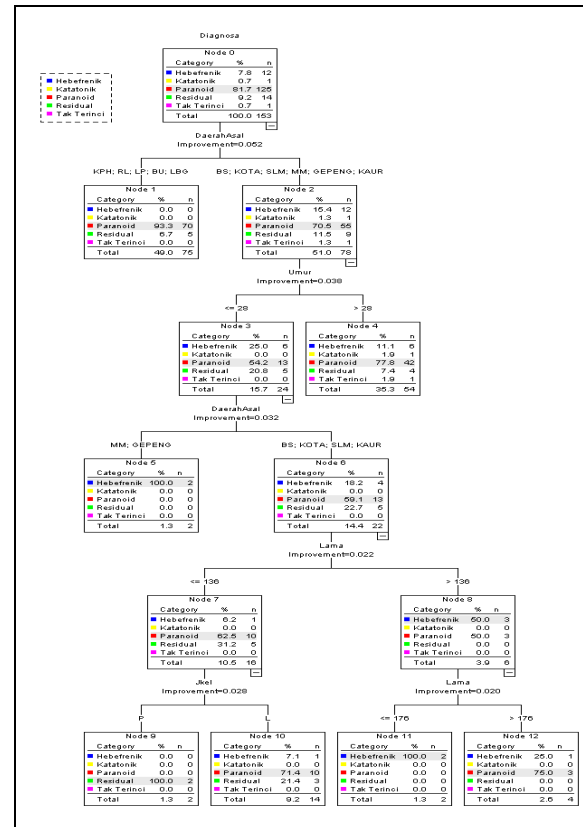
Gambar 11. Pohon Klasifikasi dengan menggunakan kaidah Gini



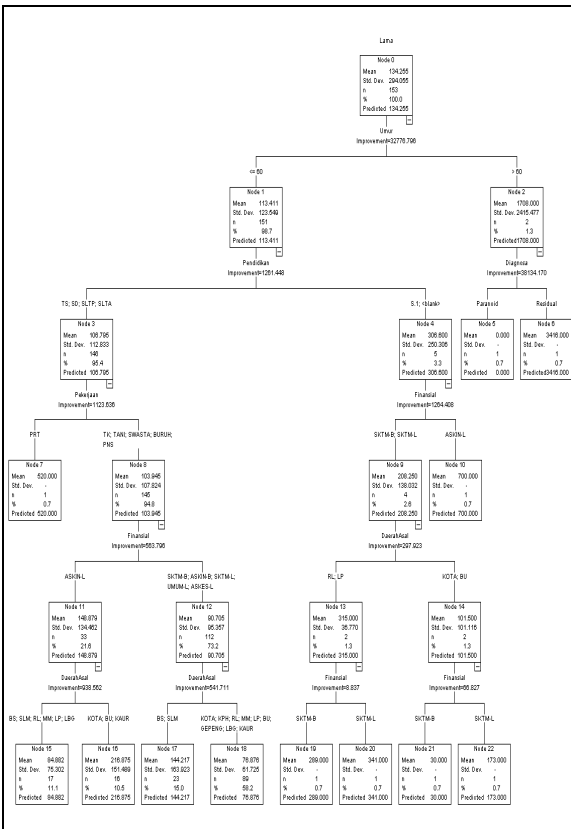
Gambar 12. Pohon Klasifikasi dengan menggunakan kaidah Twoing



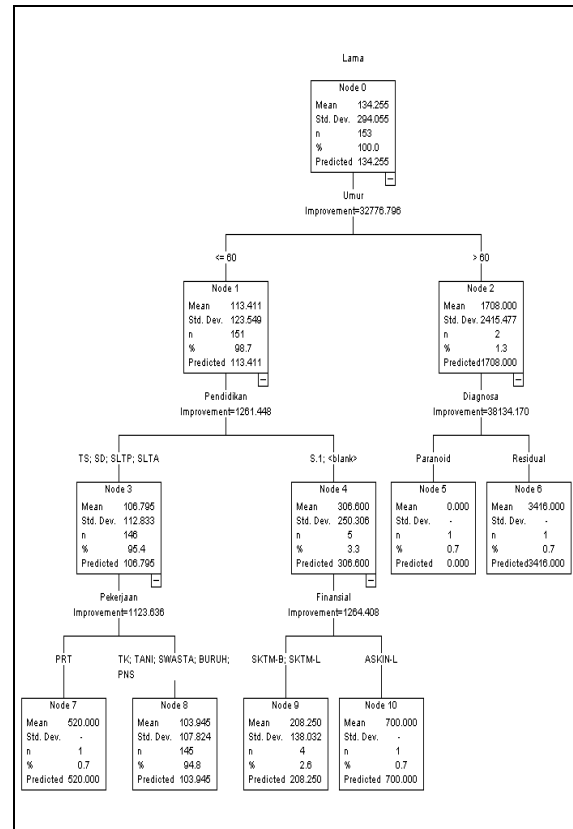
Gambar 13. Pruning tree pada pohon klasifikasi dengan menggunakan kaidah Gini



Gambar 14. Pruning tree pada pohon klasifikasi dengan menggunakan kaidah Twing



Gambar 15. Pohon Regresi dengan menggunakan cross-validation



Gambar 16. Pruning Tree pada Pohon Regresi

## DAFTAR PUSTAKA

- [1] Admin. 2004. *Skizofrenia*. Fakultas Psikologi Universitas Muhammadiyah Surakarta. [www.ums.ac.id/fakultas/psikologi/modules.php?name=News&file=article&sid=14-24](http://www.ums.ac.id/fakultas/psikologi/modules.php?name=News&file=article&sid=14-24)
- [2] Anonim. 2004a. *Preventing Schizophrenia*. [www.schizophrenia.com/prev1.htm](http://www.schizophrenia.com/prev1.htm)
- [3] Anonim. 2004b. *Symptoms and Diagnosis of Schizophrenia*. [www.schizophrenia.com/ami/diagnosis/organic.html](http://www.schizophrenia.com/ami/diagnosis/organic.html)
- [4] Anonim. 2005. *Faktor-faktor Penyebab Gangguan Jiwa*. [www.balipost.co.id/balipostcetaK/2005/8/3/k4.htm](http://www.balipost.co.id/balipostcetaK/2005/8/3/k4.htm)
- [5] Anonim. 2007a. *Classification and Regression Trees*. [www.cems.uwe.ac.uk/~rblawton/classification%20and%20regression%20trees.ppt](http://www.cems.uwe.ac.uk/~rblawton/classification%20and%20regression%20trees.ppt)
- [6] Anonim. 2007b. *Decision Tree Learning*. [www.wikipedia.com](http://www.wikipedia.com)
- [7] Anonim. 2007c. *Kesehatan Jiwa*. [faperta.ugm.ac.id/articles/kesehatan\\_jiwa.pdf](http://faperta.ugm.ac.id/articles/kesehatan_jiwa.pdf)
- [8] Anonim. 2007d. *Skizofrenia*. [id.wikipedia.org/wiki/Skizofrenia](http://id.wikipedia.org/wiki/Skizofrenia)
- [9] Anonim. 2007e. *Tree Structured Classifier*. [www.stat.psu.edu/~jiali/course/stat597e/notes2/trees.pdf](http://www.stat.psu.edu/~jiali/course/stat597e/notes2/trees.pdf)
- [10] Anonim. 2007f. *SPSS Classification Trees<sup>TM</sup> 13.0*. <https://www.washington.edu/uware/spss/docs/ClassificationTrees13.0.pdf>
- [11] Andriyashin, A. 2005. *Financial Applications of Classification and Regression Trees*. Center of Applied Statistics and Economics Humboldt University. Berlin. [edoc.hu-berlin.de/master/andriyashin-anton-2005-03-24/PDF/andriyashin.pdf](http://edoc.hu-berlin.de/master/andriyashin-anton-2005-03-24/PDF/andriyashin.pdf)
- [12] Breiman, L., J. H. Friedman, R. A. Olshen & C. J. Stone. 1984. *Classification and Regression*

- Trees*. Monterey, California, U.S.A.: Wadsworth, Inc.
- [13] Chee, J. C. 2002. *Partitioning Groups using Classification and Regression Tree in Biomedical Research*.  
[binfo.ym.edu.tw/edu/seminars/pdf/CART\\_YMUBC.pdf](http://binfo.ym.edu.tw/edu/seminars/pdf/CART_YMUBC.pdf)
- [14] Devore, J. L. 2004. *Probability and Statistics for Engineering and The Sciences*. Sixth Edition. Thomson Brooks/Cole : Canada.
- [15] Feldman, D. & S. Gross. 2003. *Mortgage Default : Classification Trees Analysis*. The Pinhas Sapir Center for Development. Tel-Aviv University.  
[sapir.tau.ac.il/papers/sapir-wp/3-03.pdf](http://sapir.tau.ac.il/papers/sapir-wp/3-03.pdf)
- [16] Hadi, S. 1977. *Metodologi Research*. Jilid II. Yayasan Penerbitan Fakultas Psikologi Universitas Gajah Mada. Yogyakarta.
- [17] Hens, N., L. Bruckers, M. Arbyn, M. Aerts & G. Molenberghs. 2002. *Classification Tree Analysis of Cervix Cancer Screening in the Belgian Health Interview Survey 1997*. Arch Public Health. 60 : 275-294.  
[www.iph.fgov.be/aph/pdf/aphfull60\\_275\\_294.pdf](http://www.iph.fgov.be/aph/pdf/aphfull60_275_294.pdf)
- [18] Johnson, R. A. 2000. *Probability and Statistics for Engineering*. Sixth Edition. Prentice Hall International, Inc. New Jersey : USA.
- [19] Kandouw, A., JES Kandouw, S. D. Elvira & I. Ariawan. 2007. *Proporsi Gangguan Depresi pada Penyalahguna Zat yang Menjalani Rehabilitasi di RS Marzoeki Mahdi*. Cermin Dunia Kedokteran No. 156. Departemen Psikiatri Fakultas Kedokteran Universitas Indonesia. Jakarta.  
[www.kalbe.co.id/.../156\\_08ProporsiGangguanDepresipenyalahguna.html](http://www.kalbe.co.id/.../156_08ProporsiGangguanDepresipenyalahguna.html)
- [20] Kucukkoçoglu, G. & O. Sezgin. 2007. *IPO Mechanism Selection by Using Classification and Regression Trees (CART)*. Başkent University, Faculty of Economics and Administrative Sciences, Management Department, Bağlıca, Ankara. Turkey.
- [21] Kuntjoro, Z. S. 2002. *Mengenal Gangguan Jiwa Pada Lansia*.  
[www.e-psikologi.com/usia/140502.htm](http://www.e-psikologi.com/usia/140502.htm)
- [22] Kutner, M.H., C.J. Nachtsheim, J. Neter & W. Li. 2005. *Applied Linear Statistical Models*. Fifth Edition. Mc Graw-Hill International Edition.
- [23] Lazarusli, C. A. 2005. *Dia Sanggup Melakukannya Berjam-jam*. Katarsis Edisi I Maret 2005.  
[www.unika.ac.id/fakultas/psikologi/katarsis/katarsis.pdf](http://www.unika.ac.id/fakultas/psikologi/katarsis/katarsis.pdf)
- [24] Rachmat, A. 2007. *Manipulasi Tree*. Handout Struktur Data Prodi Teknik Informatika UKDW.
- [25] Sezgin, O. 2006. *Statistical Methods In Credit Rating*. Department of Financial Mathematics. The Middle East Technical University. Turkey  
[www3.iam.metu.edu.tr/iam/images/2/21/ÖzgeSezginthesis.pdf](http://www3.iam.metu.edu.tr/iam/images/2/21/ÖzgeSezginthesis.pdf)
- [26] Timofeev, R. 2004. *Classification and Regression Trees (CART) Theory and Applications*. Center of Applied Statistics and Economics Humboldt University. Berlin. [edoc.hu-berlin.de/master/timofeev-roman-2004-12-20/PDF/timofeev.pdf](http://edoc.hu-berlin.de/master/timofeev-roman-2004-12-20/PDF/timofeev.pdf)
- [27] Yohannes, Y. & P. Webb. 1999. *Classification and Regression Trees, CART<sup>TM</sup>: A User Manual For Identifying Indicators of Vulnerability to famine and Chronic Food Insecurity*. International Food Policy Research Institute. Washington, U.S.A. [www.ifpri.org/pubs/microcom/micro3.pdf](http://www.ifpri.org/pubs/microcom/micro3.pdf)
- [28] Zambon, M., R. Lawrence, A. Bunn & S. Powell. 2006. *Effect of Alternative Splitting Rules on Image Processing Using Classification Tree Analysis*. Photogrammetric Engineering & Remote Sensing Vol. 72, No. 1 : 25–30.
- [29] Zhou, Z. H. 2007. *Data Mining Chapter 5 : Classification and Regression*. Department of Computer Science and Technology Nanjing University. China