# KLASIFIKASI DIAGNOSIS PENYAKIT KANKER PAYUDARA DENGAN PENDEKATAN REGRESI LOGISTIK BINER DAN METODE *CLASSIFICATION AND REGRESSION TREES* (CART)

Rifqy Marwah Akhsanti[1], Widyanti Rahayu[2], Vera Maya Santi[2]
Program Studi Matematika Fakultas, Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta
Jl. Pemuda 10, Rawamangun, Jakarta Timur 13220, Indonesia
Rifqi_almarwah@yahoo.com

**Abstrak**

Kanker merupakan masalah penyakit yang dapat menyebabkan kematian bagi penderitanya. Salah satu kanker yang sering dialami oleh kalangan wanita adalah kanker payudara. Regresi logistik biner dan Metode CART (Classification And Regression Trees) diterapkan pada data pasien kanker payudara RS. Dharmais tahun 2015 untuk mengetahui faktor pengaruh timbulnya kanker payudara yang diklasifikasikan menurut dua kategori yaitu jinak dan ganas. Faktor yang digunakan adalah usia, usia menarche, usia menopause, obesitas, riwayat keluarga penderita kanker (genetik), tidak menyusui anaknya, penggunaan KB. Analisis regresi logistik biner yang terbentuk menghasilkan faktor yang berpengaruh signifikan terhadap hasil diagnosis kanker adalah usia dan riwayat keluarga penderita kanker (genetik), model ini mampu mengklasifikasikan sebesar 90.5%. Metode CART menghasilkan pohon klasifikasi optimum dengan empat simpul terminal dan memperoleh nilai ketepatan klasifikasi sebesar 93%.

**Kata kunci** : diagnosis kanker, Regresi Logistik Biner, Metode CART

## 1. PENDAHULUAN

Salah satu permasalahan penyakit tidak menular yang sering muncul di masyarakat adalah kanker. Kanker merupakan suatu penyakit yang disebabkan oleh pertumbuhan sel-sel jaringan tubuh yang abnormal cenderung menyerang jaringan sekitarnya dan menyebar melalui jaringan ikat, darah dan menyerang organ-organ penting lainnya dalam tubuh. Kanker dapat terjadi di berbagai organ di setiap tubuh manusia, salah satunya adalah kanker payudara. Kanker payudara adalah suatu pertumbuhan jaringan payudara abnormal yang pertumbuhannya berlebihan dan bertambah banyak secara tidak terkendali. Menurut data WHO tahun 2000, terdapat 22% dari seluruh kasus kanker adalah kanker payudara.

Pengklasifikasian merupakan salah satu metode statistika dengan cara mengelompokkan atau mengklasifikasikan suatu data yang disusun secara matematis. Masalah klasifikasi sering dijumpai pada kehidupan nyata, salah satunya pada masalah diagnosis pasien kanker payudara. Ada penyelesaian masalah klasifikasi yang perlu diperhatikan dalam memilih metode klasifikasi yang tepat.

Analisis yang dapat digunakan dalam metode klasifikasi adalah regresi logistik biner dan metode CART (Classification and Regression Trees). Analisis regresi logistik biner merupakan suatu metode analisis data yang digunakan untuk mencari hubungan antara variabel respon yang memiliki dua kategorik dengan satu atau lebih variabel bebas yang berskala kategorik maupun kontinu.

Metode CART (Classification and Regression Trees) adalah metode statistic yang digunakan untuk menggambarkan hubungan antara variabel respon dengan satu atau lebih variabel bebas yang dikembangkan untuk metode klasifikasi, baik untuk variabel respon yang kategorik maupun kontinu. Jika variabel respon yang dimiliki bertipe kategorik maka CART menghasilkan pohon klasifikasi, sedangkan apabila variabel respon yang dimiliki bertipe kontinu maka CART menghasilkan pohon regresi[2].

Pada tulisan ini akan dilakukan penelitian terhadap pasien kanker payudara dengan mengaplikasikan dengan model Regresi Logistik Biner dan metode Classification and Regression Trees (CART), sehingga dapat mengetahui hasil diagnosis kanker payudara dengan pola klasifikasi dari faktor-faktor risiko yang dapat menimbulkan penyakit kanker payudara.

## 2. LANDASAN TEORI

### 2.1 Faktor Risiko Kanker Payudara

Penunjang terjadinya risiko kanker payudara disebabkan dari serangkaian faktor genetik, hormonal, dan lingkungan. Adapun berikut disajikan beberapa faktor-faktor risiko yang berhubungan dengan timbulnya penyakit kanker payudara diantaranya adalah :

a. Usia
b. Usia *Menarche*
c. Usia *Menopause*
d. Obesitas
e. Riwayat Keluarga Penderita Kanker (genetik)
f. Tidak Menyusui Anak
g. Pengguna KB

### 2.2 Regresi Logistik Biner

Menurut Hosmer dan Lemeshow (2000), metode regresi logistik adalah suatu metode analisis statistika yang mendeskripsikan hubungan antara variabel respon (Y) dengan satu atau lebih variabel bebas (X) yang berskala kategori maupun kontinu. Model regresi logistik dibentuk dengan nilai $P(Y_i = 1|X_i)$ sebagai $\pi_i$ , yang dinotasikan sebagai berikut

$$\pi_i = \frac{\exp(X^T\beta)}{1 + \exp(X^T\beta)}$$

Suatu fungsi dari $\pi_i$ dicari dengan menggunakan transformasi logit, yaitu

$$Logit\ (\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = X_i^T\beta$$

Model regresi logistik dengan $k$ variabel prediktor ke -$j$ berupa data kategori adalah

$$Logit\ \ (\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \sum_{l=1}^{k_j-1} \beta_{jl}\, D_{jl} + \cdots + \beta_k\, X_{ik}$$

Metode yang dapat menduga parameter dalam regresi logistik adalah metode maksimum likelihood. Fungsi distribusi peluang untuk $Y_i$ adalah

$$f(y_i) = \pi_i^{y_i}[1-\pi_i]^{1-y_i} \quad , \quad \text{i} = 1, 2, \ldots, \text{n.}$$

dengan $Y_i$ bernilai 0 dan 1 untuk masing-masing amatan. Karena $Y_i$ diasumsikan saling bebas, maka fungsi likelihoodnya adalah

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

Uji signifikansi terhadap parameter-parameter model dilakukan baik secara serentak maupun secara parsial dan uji kelayakan model. Pengujian parameter model secara serentak menggunakan uji Rasio *Likelihood*, dengan hipotesis adalah

$H_0 :\ \beta_0 = \beta_1 = \cdots = \beta_k = 0$

$H_1$ : minimal ada satu $\beta_i \neq 0$ dimana $i = 1, 2, \ldots, k$

Statistik uji yaitu Uji G

$$G = -2\,ln\left[\frac{L_1}{L_0}\right]$$

Kriteria penolakan (Tolak $H_0$) adalah jika nilai $G > \chi^2_{db,\alpha}$

Pengujian parameter model secara pasial menggunakan uji Wald, dengan hipotesis adalah

$H_0$ : $\beta_i = 0$

$H_1$ : $\beta_i \neq 0$ dimana i $= 1, 2, \ldots, k$

Statistik uji yaitu Uji W

$$W = \left[\frac{\beta_i}{SE(\beta_i)}\right]^2$$

Kriteria penolakan (Tolak $H_0$) adalah jika nilai W $> \chi^2_{\alpha,1}$

Uji Kelayakan Model dilakukan untuk menilai apakah model sesuai dengan data atau tidak menggunakan uji *Goodness of fit*, dengan hipotesis adalah

$H_0$ : Tidak trdapat perbedaan signifikan antara klasifikasi yang diprediksi dengan klasifikasi yang diamati.

$H_1$ : Terdapat perbedaan signifikan antara klasifikasi yang diprediksi dengan klasifikasi yang diamati.

Kriteria penolakan (Tolak $H_0$) adalah p-*value* $> 0.05$

Interpretasi koefisien untuk model regresi logistik biner dapat dilakukan dengan menggunakan nilai rasio oddsnya. Odds dari suatu kejadian digambarkan sebagai peluang dari peristiwa yang terjadi dibagi oleh peluang dari peristiwa yang tidak terjadi. Nilai rasio odds didefinisikan sebagai berikut

$$\hat{\psi} = \exp(\hat{\beta}_\iota) = \exp(g(1) - g(0))$$

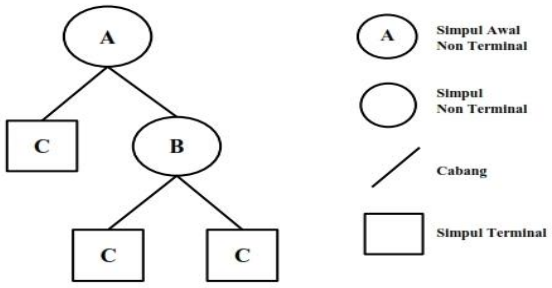
Interpretasi dari rasio *odds* adalah kecenderungan untuk nilai $Y = 1$ pada $X = 1$ sebesar $\psi$ kali dibandingkan $X = 0$.

### 2.3 Metode *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. Freidman, Richard A. Olshen dan Charles J. Stone sekitar tahun 1980-an.

Teknik dalam membuat pohon klasifikasi CART dikenal dengan istilah *Binary Recursive Partitioning*. Proses disebut *binary* karena setiap *parent node* akan selalu mengalami pemecahan ke dalam tepat dua *child node*. Sedangkan *recursive* berarti bahwa proses pemecahan tersebut akan diulang kembali pada setiap *child node* dari hasil pemecahan terdahulu. Proses ini akan terus dilakukan sampai tidak ada kesempatan lagi untuk melakukan pemecahan berikutnya. Dan istilah *partitioning* mengartikan bahwa learning sample yang dimiliki dipecah ke dalam partisi-partisi yang lebih kecil.

Misalkan A,B, dan C merupakan variabel-variabel yang terpilih untuk menjadi simpul dalam pembentukan pohon klasifikasi sederhana




Algoritma pembentukan pohon klasifikasi terdiri dari empat tahapan, yaitu kriteria pemecahan simpul nontermnal; penandaan label kelas; menentukan simpul terminal; dan menentukan pohon optimum.

### 2.3.1 Kriteria Pemecahan Simpul Nonterminal

Kriteria pemecahan terbaik dibentuk berdasarkan *impurity* (fungsi keragaman). Fungsi impuritas yang dapat digunakan adalah Indeks Gini. Nilai impuritas menggunakan Indeks Gini pada simpul , maka $i(t)$ :

$$i(t) = 1 - \sum_j P^2(j|t)$$

Jika sebuah pemecahan $s_k$ dari simpul $t$ dipilah menjadi dua maka memberikan proporsi kanan yaitu $p_R$ dari data pada $t$ ke dalam $t_R$ dan proporsi kiri yaitu $p_L$ dari data pada $t$ ke dalam $t_L$, maka penurunan impuritas sebagai berikut :

$$\Delta i(s_k, t) = i(t) - p_R\, i(t_R) - p_L\, i(t_L)$$

Pemecahan $s^*$ sebagai pemecahan teraik dari suatu simpul $t$ yaitu pemecahan yang memberikan penurunan impuritas terbesar yaitu

$$i(s^*, t) = max_{s_k \in S}\, \Delta i(s_k, t)$$

### 2.3.2 Penandaan Label Kelas

Penandaan label pada kelas dibentuk dari aturan $j(t)$ yitu kelas yang memiliki $P(j|t)$ terbesar ditetapkan sebagai kelasnya. Aturan penetapan kelas $j^*(t)$ yaitu Jika $P(j|t) = max_i\ p(j|t)$ maka $j^*(t) = j$. Jika nilai maksimum dicapai oleh dua atau lebh kelas, maka penetapan $j^*(t)$ dari kelas yang mana saja.

### 2.3.3 Menentukan Simpul Terminal

Suatu simpul *t* akan menjadi simpul terminal atau tidak akan dipilah kembali, jika jumlah pengamatannya kurang dari jumlah minimum. Umumnya jumlah pengamatan minimum pada simpul sebesar 5 dan terkadang berjumlah 1 (Breiman *et al.* 1993*).*

### 2.3.4 Menentukan Pohon Optimum

Menurut Breiman *et al.* (1993), salah satu cara mendapatkan pohon optimum yaitu dengan pemangkasan (*pruning).* Pemangkas berturut-turut memangkas pohon bagian yang kurang penting. Tingkat kepentingan sebuah pohon bagian diukur berdasarkan ukuran biaya kompleksitas (*cost complexity).* Persamaannya adalah:

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

Pemangkasan pohon klasifikasi optimum dimulai dengan mengambil simpul anak kanan $t_R$ dan simpul anak kiri $t_L$. Jika $(t) = R(t_R) + R(t_L)$ , maka simpul anak $t_R$ dan $t_L$ dipangkas. Proses tersebut diulang sampai tidak ada lagi pemangkasan yang mungkin.

## 3. PEMBAHASAN

### 3.1 Analisis Data Hasil Diagnosis Kanker Kanker Payudara dengan Regresi Logistik Biner

Pendugaan model regresi logistik biner dengan menggunakan tujuh variabel bebas dilakukan pengujian secara serentak dilakukan uji hipotesis :

$H_0$ : $\beta_1 = \beta_2 = \cdots = \beta_7 = 0$

$H_1$ : minimal ada satu $\beta_i \neq 0$ ; $i = 1, 2, \ldots, 7$

$$G = -2\,[L_0 - L_1] = 246.865 - 101.081$$
$$= 145.865$$

Kriteria pengambilan keputusan adalah karena nilai $> \chi^2_{(7,0.05)} = 14.067$ , maka Tolak $H_0$ artinya minimal ada satu $\beta_i$ yang tidak sama dengan nol.

Selanjutnya akan diuji parameter secara parsial yaitu melihat pengaruh setiap parameter pada model secara individual. Hasil pengujian parameter diperoleh menggunakan statistik uji-Wald, akan ditunjukkan dalam tabel dibawah ini dengan hipotesis sebagai berikut

$H_0$ : $\beta_i = 0$

$H_1$ : $\beta_i \neq 0$ i $= 1, 2, \ldots, 7$

| Variabel Bebas | Wald | p-*value* |
|---|---|---|
| **Constant** | 36.217 | **0.000*** |
| **Usia** | 40.512 | **0.000*** |
| Usia *Menarche* | 0.835 | 0.361 |
| Usia *Menopause* | 3.325 | 0.068 |
| Obesitas | 0.144 | 0.704 |
| **RKPK** | 7.898 | **0.005*** |
| TMA | 1.994 | 0.158 |
| KB | 0.092 | 0.761 |

Uji Wald menguji koefisien dengan menghasilkan dua variabel pada $= 0.05$ , yaitu nilai p-value $< \alpha$ adalah variabel usia dan riwayat keluarga penderita kanker.

Maka Model Regresi Logistik Biner terbaik yang diperoleh adalah

$$Logit\ (\pi_i) = -16.620 + 0.403\ Usia + 1.704\ RKPK_{(1)}$$

Uji kelayakan model di atas diperoleh dengan mencari nilai goodness of fit yaitu jika model layak maka nilai p-*value* $> 0.05$ sehingga hasil pengujian model di atas menunjukkan nilai p-value sebesar 0.289. Maka kriteria keputusan adalah Terima $H_0$, dapat disimpulkan bahwa model tersebut layak untuk digunakan untuk analisis selanjutnya untuk melihat faktor risiko timbulnya kanker terhadap hasil diagnosis penyakit kanker payudara.

Interpretasi koefisien untuk model regresi logistik biner dapat dilakukan dengan melihat nilai rasio oddsnya. Nilai duga rasio odds untuk kedua variabel adalah

| Variabel Bebas | Penduga Rasio *Odds* |
|---|---|
| Usia $X_1$ | 1.497 |
| RKPK $X_{5(1)}$ | 5.496 |

Dari tabel di atas dapat diinterpretasikan bahwa nilai rasio *odds* pada variabel usia adalah 1.497 bernilai lebih dari 1 maka variabel tersebut memberikan pengaruh positif terhadap faktor risiko menentukan hasil diagnosis kanker payudara. Dan variabel RKPK berpeluang 243 kali lebih besar berisiko kanker dibandingkan dengan pasien yang tidak memiliki RKPK.

Hasil ketepatan klasifikasi data diagnosis kanker payudara menggunakan *Apparent Error Rate* (APER)

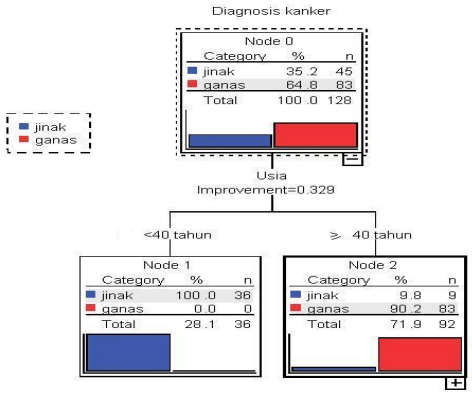
| Observasi | Hasil Prediksi | | Persentase Total Prediksi |
|---|---|---|---|
| | Jinak | Ganas | |
| Diagnosis kanker payudara jinak | 55 | 13 | 80.9% |
| Diagnosis kanker payudara jinak | 5 | 116 | 95.9% |
| Persentase keseluruhan | | | 90.5% |

Persentase hasil missklasifikasi dari 189 pasien adalah 9.5%. Dengan perhitungan nilai ketepatan klasifikasi adalah

$$1 - APER = \frac{(55+116)}{(55+13+5+116)} = \frac{171}{189} = 90.5\ \%$$

### 3.2 Analisis Data Hasil Diagnosis Kanker Kanker Payudara dengan Metode Classification And Regression Trees (CART)

Pemecahan simpul nonterminal berdasarkan dari nilai fungsi keragaman $\Delta I(s_k, t)$ kriteria variabel pemecahan $s_k$ yang terpilih adalah usia sebagai pemilah terbaik karena memiliki nilai keragaman yang paling maksimal dimana $\Delta I(s_k, t)$ sebesar 0.329. Akan ditunjukkan pada gambar di bawah ini




Setelah terpilih pemilihan terbaik, maka pada simpul utama yaitu simpul 0 berisi 128 objek yang selanjutnya akan dipilah untuk memecah simpul $t$ menjadi dua buah simpul yaitu simpul $t_R$ terbentuk akibat kriteria variabel usia pasien $\geq$ 40 tahun dan untuk simpul $t_L$ terbentuk akibat kriteria variabel pasien yang memiliki usia < 40 tahun.

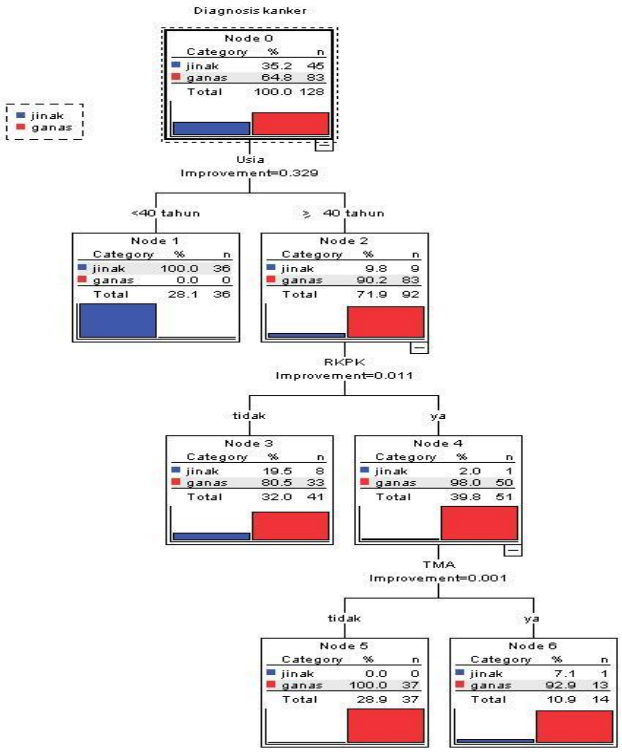
Selanjutnya pemberian label kelas pada simpul-simpul yang telah terbentuk. Prosedur pemberian label kelas berdasarkan aturan penetapan kelas adalah jika $P(j|t) = max_i\, P(j|t)$ maka $j^*(t) = j$ , dimana $j^*(t)$ adalah kelas yang diidentifikasi pada simpul $t$ .

$$P(jinak|t) = \frac{45}{128} = 0.352$$

$$P(ganas|t) = \frac{83}{128} = 0.648$$

Sehingga simpul diberi label kelas “Diagnosis Kanker Ganas” , karena memiliki peluang kelas ganas lebih besar dibandingkan kelas jinak. Proses pelabelan kelas ini berlaku pada semua simpul terutama pada simpul akhir (terminal), karena simpul terminal adalah simpul yang sangat penting dalam memprediksi suatu objek pada kelas tertentu jika objek berada pada simpul terminal tersebut.

Menentukan simpul terminal diperoleh dari hasil pohon klasifikasi optimum melalui pemangkasan pohon, sehingga pada pohon klasifikasi optimum diagnosis kanker payudara diperoleh 7 simpul terdiri 3 simpul nonterminal dan 4 simpul terminal. Variabel yang masuk dalam pohon klasifikasi adalah Usia, RKPK, dan Tidak menyusui anak. Variabel usia adalah variabel pertama sebagai penyekat. Hal ini menyatakan bahwa variabel tersebut merupakan variabel yang paling dominan dalam pembentukan pohon klasifikasi. Berikut ini merupakan bentuk pohon klasifikasi optimum setelah pohon dipangkas




Pohon klasifikasi optimum ini memiliki 4 kelas yang menentukan hasil diagnosis kanker payudara. Klasifikasi yang terbentuk adalah

a) Diagnosis kanker payudara pada pasien yang memiliki riwayat tidak menyusui bayinya memiliki potensi 92.9% akan mengidap kanker ganas dan untuk pasien memiliki riwayat tidak menyusui bayinya berpotensi mendapatkan hasil diagnosis kanker payudara adalah jinak sebesar 7.1%.
b) Pasien yang tidak memiliki riwayat tidak menyusui bayinya juga memiliki potensi tinggi sebesar 100% dapat terdiagnosis kanker payudara adalah ganas.
c) Adanya riwayat keluarga penderita kanker payudara pada pasien terdiagnosis kanker payudara adalah jinak berpotensi 19.5% dan pada pasien yang mengidap kanker ganas akan berpotensi memiliki riwayat keluarga penderita kanker payudara adalah sebesar 80.5%.
d) Untuk usia pasien yang memiliki kriteria < 40 tahun berpotensi besar akan mengidap kanker jinak yaitu 100%

Pohon klasifikasi optimum yang telah dihasilkan kemudian diuji tingkat ketepatan atau akurasi dalam mengelompokkan data testing. Uji keakuratan klasifikasi pohon dengan menggunakan

$$R(d) = \frac{1}{N} \sum_{(x_n\,,\,j_n)\in\mathcal{L}}^{N} x(d(\widetilde{x_n}) \neq j_n)$$

$$= \frac{9+0}{128} = 0.070$$

Nilai $R(d) = 0.070$ maka nilai ketepatan klasifikasi adalah $1 - R(d) = 0.93 = 93\%$. Hasil dari klasifikasi optimum dapat dilihat pada tabel sebagai berikut

| Observasi | Hasil Prediksi | | Persentase Total Prediksi |
|---|---|---|---|
| | Jinak | Ganas | |
| Diagnosis kanker payudara jinak | 36 | 9 | 80% |
| Diagnosis kanker payudara jinak | 0 | 83 | 100% |
| Persentase keseluruhan | | | 93% |

## 4. KESIMPULAN

Dari hasil analisis diperoleh kesimpulan bahwa model regresi logistik biner mampu mengklasifikasi dengan nilai ketepatan klasifikasi sebesar 90.5%. Dan untuk variabel faktor pengaruh timbulnya penyakit kanker yang signifikan terhadap diagnosis kanker payudara yaitu Usia dan Riwayat keluarga penderita kanker. Model logit terbaik yang diperoleh adalah

$$Logit\ (\pi_i) = -16.620 + 0.403\ X_1 + 1.704 X_{5(1)}$$

Pohon klasifikasi CART yang terbentuk menghasilkan pohon optimum dengan empat simpul terminal. Variabel yang masuk ke dalam pohon klasifikasi adalah Usia, RKPK, dan Tidak menyusui anak. Nilai ketepatan klasifikasi pohon optimum pada diagnosis kanker payudara sebesar 93%.

**Daftar Pustaka**


[1] Agresti, A. 2007. "*An Introduction to Categorical Data Analysis*". Second Edition. John Wiley and Sons, Inc. New Jersey.

[2] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. 1993. "*Classification and Regression Trees*". Chapman and Hall, New York.

[3] Johnson, R.A., and Wichern, D.W. 2007. "*Applied Multivariate Statistical Analysis*". Sixth Edition. Printice Hall . New Jersey.

[4] Hosmer, D.H., and Lemeshow, S. 2000. "*Applied Logistic Regression*". Second Edition. John Wiley and Sons, Inc. New York.

[5] Montgomery, D.C., and Peck, E.A. 1992. "*Introduction to Linear Regression Analysis 2sd Edition*". John Wiley and Sons, Inc. New York.

[6] Webb, P., and I., Yohannes. 1999. "*Classification and Regression Trees, CART*". International Food Policy Research Institute, Washington D.C.