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

Skripsi
Disusun untuk melengkapi syarat-syarat
guna memperoleh gelar Sarjana Sains

RIFQY MARWAH AKHSANTI
3125110233

PROGRAM STUDI MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS NEGERI JAKARTA
2017

ABSTRACT

RIFQY MARWAH AKHSANTI, 3125110233. Classification diagnosis of breast cancer disease with Binary Logistic Regression and method of CART (*Classification And Regression Trees*). Thesis. Faculty of Mathematics and Natural Science Jakarta State University. 2017.

Cancer is a disease problem that can use death to the sufferer. One of cancer that is often experienced among women is breast cancer. The methods of Binary Logistic Regression and CART (Classification and Regression Trees) are applied to data of breast cancer patients Dharmais hospital 2015 to determine factors influence the incidence of breast cancer is classified according two categories of benign and malignant. The variables used are age, the age of menarche, menopause age, obesity, family history of cancer, not breastfeeding her child, and the use of KB. Binary Logistic Regression that formed the influential factors of significantly to the results of the diagnosis of cancer is an age and a family history of cancer, this model is able to classify of 90.5%. *CART method is produces to optimum classification of tree with four terminal nodes and obtain a value of* 93% *classification accuracy.*

Keywords *: diagnosis of cancer, binary logistic regression, CART of method.*

ABSTRAK

RIFQY MARWAH AKHSANTI, 3125110233. Klasifikasi Diagnosis Penyakit Kanker Payudara Dengan Pendekatan Regresi Logistik Biner dan Metode *Classification And Regression Trees* (CART). Skripsi. Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta. 2017.

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.

PERSEMBAHANKU...

" Hidup itu ibarakan "Matahari dan Bulan", Dilihat atau tidak ia tetap bersinar. Dihargai atau tidak ia tetap menerangi. Diterimaksihi atau tidak ia tetap berbagi"

" Teruskanlah berbuat baik, berkata baik, memberi nasihat yang baik, walaupun tidak ramai orang mengenalimu, cukuplah Allah mengenalimu lebih daripada yang lain"

Skripsi ini kupersembahkan untuk Mamah, Papah, Ijal, dan Caca.

"Terima kasih atas dukungan, do'a, serta kasih sayang kalian".

KATA PENGANTAR

Puji syukur kepada Allah SWT atas pengetahuan dan kemampuan sehingga penulis dapat menyelesaikan skripsi yang berjudul "Klasifikasi Diagnosis Penyakit Kanker Payudara Dengan Pendekatan Regresi Logistik Biner dan Metode *Classification And Regression Trees* (CART)" yang merupakan salah satu syarat dalam memperoleh gelar Sarjana Jurusan Matematika Universitas Negeri Jakarta.

Skripsi ini berhasil diselesaikan tidak terlepas dari adanya bantuan dari berbagai pihak. Oleh karena itu, dalam kesempatan ini penulis ingin menyampaikan terima kasih terutama kepada:

1. Ibu Dra. Widyanti Rahayu, M.Si., selaku Dosen Pembimbing I dan Ibu Vera Maya Santi, M.Si., selaku Dosen Pembimbing II, yang telah meluangkan waktunya dalam memberikan bimbingan, saran, nasehat serta arahan sehingga skripsi ini dapat menjadi lebih baik dan terarah.

2. Ibu Dr. Lukita Ambarwati, S.Pd, M.Si., selaku Koordinator Prodi Matematika FMIPA UNJ yang telah banyak membantu dan memberikan saran untuk penulis.

3. Bapak Drs. Mulyono, M.Kom., selaku Pembimbing Akademik atas segala bimbingan dan kerja sama Bapak selama perkuliahan, dan seluruh Bapak/Ibu dosen atas pengajarannya yang telah diberikan selama perkuliahan, serta karyawan/karyawati FMIPA UNJ yang telah memberikan informasi yang penulis butuhkan dalam menyelesaikan skripsi.

4. Mamah dan Papap tercinta yang selalu mendo'akan, serta mendukung, memberi motivasi, dan setia membantu penulis dengan penuh cinta dan kasih sayang yang tulus.

5. Adik Laki-laki penulis Rizal dan adik perempuan penulis Tasya yang terus memberi semangat, mendoakan penulis, dan selalu membantu ketika penulis mengalami kesulitan dalam penulisan skripsi ini.

6. Keluarga besar tercinta Maende, Pa Aki (Alm), Mbah, Aki Sanusi (Alm), Uwa, Ate, Abah, Yeye, Om dan semua Sepupu yang tak pernah henti mendo'akan dan memberi semangat kepada penulis.

7. Teman alumni Ma'had Al-Zaytun 2005 terkhusus Raisa, Lina, Puji, Luthfi, Dian, Neni, Cut, Odie, Arum, Titin, dan Ama sebagai teman separuh hidup penulis hingga saat ini dan semoga akan seterusnya.

8. Teman-teman Matematika 2011 terimakasih sudah menjadi keluarga baru mengisi canda tawa haru dalam kehidupan penulis. Teman BONS-ku tersayang Nurul, Nita, Tyan, Puti, Dytta, Dinna dan Muti yang selalu setia memberikan perhatian dan semangat serta Indah Hoi selaku teman seperjuangan penulis

9. Asih sebagai teman kamar dan teman separuh hidup penulis yang selalu ada dikala senang maupun susah dan teman sehobi nonton drama penulis yaitu Tiyan.

10. Kak Denis telah menjadi sosok sebagai kakak, teman, sekaligus orang tua yang selalu senantiasa ada bagi penulis.

11. Teteh Vera yang telah memberikan banyak bantuan berupa wawasan dan memotivasi penulis.

12. Adik tingkat matematika 2012 terutama Meila, Jen, Yohana, Bety , Sharah, Yuli yang telah membantu penulis dan juga sebagai teman seperjuangan penulis, serta adik tingkat matematika 2015 terutama Istu, Kikin, Dara, Dinda, Gabela, Juli yang selalu menyemangati dan menghibur penulis.

Penulis menyadari bahwa skripsi ini masih jauh dari sempurna. Masukan dan kritikan akan sangat berarti. Semoga skripsi ini dapat bermanfaat bagi pembaca sekalian.

Jakarta, 10 Januari 2017

Rifqy Marwah Akhsanti

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

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 (Corrwin, 2009). Kanker dapat terjadi di berbagai organ di setiap tubuh manusia, salah satunya adalah kanker payudara. Ristarolas (2009) menyatakan 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.

Kanker payudara merupakan penyebab utama dalam insiden dan kematian oleh kanker pada wanita. Insidensi berdasar *Age Standardized Ratio* (ASR) tahun 2000, kanker payudara sebesar 20,6 (20,6/100.000 penduduk) dan *mortality* (ASR) tahun 2000 akibat kanker payudara di Indonesia sebesar 10,1 (10,1/100.000 penduduk) dengan jumlah kematian akibat kanker payudara sebesar 10.753. Dan pada tahun 2005 diperkirakan *mortality* (ASR) sebesar 10,9/100.000 penduduk dengan jumlah kematian akibat kanker payudara sebanyak 12.352 (Ramli, 2003). Kanker payudara tidak hanya menyerang wanita saja, laki-laki juga berisiko terke-

na kanker payudara walaupun tidak sebesar pada wanita.

Terdapat dua klasifikasi tipe yang diambil pada diagnosis kanker berdasarkan tingkat keganasannya yaitu jinak dan ganas. Kanker payudara yang jinak umumnya ditemukan benjolan berbentuk kelereng yang berukuran kurang dari 2 cm, sedangkan pada kanker payudara yang ganas bentuk benjolan membesar pada payudara sehingga menyebabkan nyeri, terdapat pula kerutan dan mengeluarkan cairan, serta mengalami perubahan pada kulit dan ukuran payudara. Oleh karena itu, untuk menyelesaikan masalah diagnosis kanker pada pasien kanker payudara diselesaikan dengan metode klasifikasi.

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. Analisis ini tidak memerlukan asumsi multivariat normal atau kesamaan matriks varian kovarian, sehingga regresi logistik dapat diterapkan dalam berbagai skala data.

Metode CART (*Classification and Regression Trees*) adalah metode statistik 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.

Penelitian sebelumnya tahun 2012, I Nyoman Putrayasa melakukan penelitian dengan menerapkan analisis regresi logistik biner dan metode CART untuk melihat hubungan antara status desa dan potensi desa yang mempengaruhi status tingkat kemajuan desa di provinsi Bali. 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.

1.2 Perumusan Masalah

Permasalahan yang akan dibahas pada penulisan ini yaitu bagaimana pengklasifikasian faktor yang mempengaruhi diagnosis kanker pada pasien kanker payudara dengan menggunakan model Regresi Logistik Biner dan metode CART (*Classification and Regression Trees*)?

1.3 Pembatasan Masalah

Pembatasan masalah dalam penelitian ini adalah :

1. Metode CART (*Classification and Regression Trees*) terdiri dari dua metode yang berbeda yaitu pohon klasifikasi dan pohon regresi, dalam pembahasan ini hanya dilakukan pada pembentukan pohon klasifikasi.

2. Penduga parameter pada model Regresi Logistik Biner yang digunakan adalah Metode Maksimum *Likelihood* dan Metode Newton-Raphson

3. Pembentukan pohon klasifikasi dilakukan dengan klasifikasi *Binary Recrusive Partitioning*.

4. Pada penelitian ini data yang digunakan adalah data pasien terdiagnosis kanker payudara di RS.Dharmais pada tahun 2015.

1.4 Tujuan Penulisan

Adapun tujuan yang ingin dicapai dalam penulisan skripsi ini adalah :

1. Menjelaskan bagaimana model Regresi Logistik Biner dan metode CART (*Classification and Regression Trees*) adalah beberapa metode dalam teknik klasifikasi.

2. Mengetahui hasil diagnosis kanker pada pasien kanker payudara berdasarkan faktor yang mempengaruhi dengan model Regresi Logistik Biner dan metode CART (*Classification and Regression Trees*).

3. Mendapatkan suatu kelompok data sebagai penciri dari suatu pengklasifikasian.

4. Mengaplikasikan model Regresi Logistik Biner dan metode CART (*Classification and Regression Trees*) pada data RS.Dharmais untuk pasien kanker payudara.

1.5 Manfaat Penulisan

Manfaat yang diharapkan dari penulisan skripsi ini adalah :

1. Bagi penulis, menambah wawasan tentang metode klasifikasi dan mampu mengklasifikasikan dengan model Regresi Logistik Biner dan metode CART (*Classification and Regression Trees*).

2. Bagi pihak lain, dapat dijadikan sebagai referensi dan informasi bahan tambahan untuk pengkajian yang lebih lanjut.

1.6 Metode Penelitian

Skripsi ini merupakan studi literatur, yaitu melakukan penelitian kajian teori dengan cara menganalisis pengetahuan yang ada dalam pustaka. Sumber kajian pustaka dapat berupa buku, jurnal, skripsi dan laporan penelitian yang membahas tentang pengklasifikasian dengan menggunakan model Regresi Logistik Biner dan metode CART (*Classification and Regression Trees*).

BAB II

LANDASAN TEORI

Pada bab ini akan dibahas tentang penyakit kanker payudara dan faktor risiko kanker payudara, bagaimana bentuk model regresi logistik. Sebagai awal, ada beberapa teori yang mendasari penelitian ini akan dijelaskan mengenai analisis regresi, metode maksimum *likelihood*, metode newton-raphson, pengujian signifikansi parameter dan rasio *odds*.

2.1 Kanker Payudara

Menurut Luwia (2003), kanker payudara merupakan kanker yang berasal dari kelenjar, saluran kelenjar dan jaringan penunjang payudara. Ketika sejumlah sel di dalam payudara tumbuh dan berkembang dengan tidak terkendali inilah dikatakan kanker payudara. Penyebab dari kanker payudara tidak diketahui dengan pasti, namun terdapat serangkaian faktor genetik, hormonal, dan lingkungan. Penyebab tersebut dapat menunjang terjadinya kanker payudara.

Genetik merupakan faktor penting terjadinya kanker payudara akibat kelainan genetik sebesar $5-10\%$. Riwayat keluarga yang perlu dicatat diantaranya adalah kanker payudara pada ibu atau saudara perempuan yang terkena kanker payudara pada umur di bawah 50 tahun atau keponakan dengan jumlah lebih dari dua (Luwia, 2003).

Hormon estrogen adalah hormon yang berperan dalam proses tumbuh kembang organ seksual wanita. Hormon estrogen justru sebagai penyebab awal

kanker pada sebagian wanita. Hal ini disebabkan adanya reseptor estrogen pada sel-sel epitel saluran kelenjar susu. Hormon estrogen yang menempel pada saluran ini, lambat laun akan mengubah sel-sel epitel tersebut menjadi kanker (Luwia, 2003). Pengunaan KB hormonal seperti pil, suntik KB, dan susuk yang mengandung banyak dosis estrogen meningkatkan risiko timbulnya kanker payudara (John Cleese, 2010). Adapun faktor ingkungan dan gaya hidup juga dapat menjadi pemicu kanker payudara. Lingkungan tersebut berupa zat makanan, zat kimia, infeksi dan faktor fisik seperti paparan radiasi bahan-bahan radioaktif, dan trauma (Luwia, 2003).

2.1.1 Faktor Risiko Kanker Payudara

Banyak faktor yang diprediksi mempunyai hubungan timbulnya kanker payudara (John Cleese, 2010). 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 :

1. Usia

   Wanita yang berusia lebih dari 40 tahun mempunyai risiko kanker payudara lebih besar dibandingkan usia kurang dari 40 tahun. Banyak kasus kanker payudara yang ditemukan terjadi pada wanita berusia antara 40-64 tahun (Wilensky dan Lincoln, 2008).

2. Usia *Menarche* (pertama menstruasi)

   Pada wanita yang riwayat *menarche*nya lambat insedensinya lebih rendah akan tetapi *menarche* awal (dibawah 12 tahun) termasuk dalam faktor risiko terjadinya kanker payudara (Luwia, 2003).

3. Usia *Menopause*

   Wanita yang usia *menopause*nya terlambat atau lebih dari 50 tahun mempunyai risiko terkena kanker payudara lebih besar dibandingkan wanita yang usia *menopause*nya normal yaitu usia kurang dari 50 tahun (Luwia, 2003).

4. Obesitas

   Wanita yang mengalami obesitas cenderung akan terkena kanker payudara. Risiko pada obesitas akan meningkat karena sintesis estrogen pada timbunan lemak (Rasjidi, 2009).

5. Riwayat keluarga penderita kanker payudara (genetik)

   Risiko terkena kanker payudara meningkat pada wanita yang mempunyai ibu atau saudara perempuan yang terkena kanker payudara. Semua saudara dari penderita kanker payudara memiliki peningkatan risiko mengalami kanker payudara (Wilensky dan Lincoln, 2008).

6. Tidak menyusui anak

   Menyusui merupakan salah satu faktor penting yang memberikan proteksi terhadap risiko kanker payudara. Wanita yang tidak menyusui bayinya, mempunyai risiko yang tinggi terkena kanker payudara dibandingkan dengan wanita yang menyusui bayinya (Bustan, 2007).

7. Penggunaan KB

   Penggunaan KB hormonal seperti pil atau suntik tidak dianjurkan lebih dari lima tahun dan untuk wanita yang telah berusia di atas 35 tahun. Hal ini dapat meningkatkan risiko terkena kanker payudara (Luwia, 2003).

2.2 Analisis Regresi

Analisis regresi merupakan suatu metode statistika untuk menganalisis data yang menggambarkan hubungan antara satu atau lebih variabel bebas dengan variabel respon. Misal diberikan himpunan data (X_i, Y_i) dimana $i = 1, 2, \ldots, n$. Secara umum hubungan antar Y dan X dapat didefinisikan sebagai berikut :

$$Y_i = f(x_i) + \varepsilon_i \tag{2.1}$$

dengan $f(x_i)$ adalah suatu fungsi regresi yang belum diketahui dan ingin ditaksir, dan ε_i adalah suatu variabel acak yang menggambarkan variasi Y di sekitar $f(x)$ (Eubank, 1999).

Regresi parametrik terdapat beberapa asumsi menentukan model, sehingga diperlukan pengujian untuk memenuhi asumsi tersebut. Model regresi berganda adalah

$$Y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ \ldots \ + \beta_k X_{ik} + \varepsilon \tag{2.2}$$

Pada model regresi tersebut biasanya parameter ditaksir menggunakan metode kuadrat terkecil. Metode kuadrat terkecil merupakan salah satu metode yang paling sering digunakan untuk menduga parameter regresi, penduga kuadrat terkecil ini diperoleh dengan meminimumkan jumlah kuadrat galat (JKG) yang didefinisikan sebagai berikut

$$JKG = \varepsilon^{'}\varepsilon = (Y - Xb)^{'}(Y - Xb)$$

Sehingga diperoleh penduga kuadrat terkecil untuk parameter β adalah

$$\widehat{\beta} = b = (X'X)^{-1}X'Y$$

Masalah multikolinieritas merupakan masalah yang sering timbul pada regresi linier yaitu suatu kondisi dimana terdapat suatu hubungan atau ketergantungan linier dari setiap variabel bebas dalam model regresi. Jika variabel tersebut saling berkolerasi maka akan sulit mendapatkan penaksir yang baik bagi koefisien regresi, sehingga untuk mendeteksi adanya multikolinieritas dapat dicari dengan menggunakan nilai VIF (*Variance Inflation Factor*) didefinisikan sebagai berikut

$$VIF = \frac{1}{1 - R_j^2}$$

dengan R_j^2 adalah koefisien determinasi.

Besarnya nilai VIF ini bergantung pada nilai koefisien determinasi (R_j^2) yang dihasilkan. Menurut Montgomery dan Peck (1992) jika nilai VIF lebih besar dari 10 menunjukkan adanya multikolinieritas antara variabel-variabel bebas yang diamati.

2.3 Model Regresi Logistik

Penerapan regresi klasik pada variabel bebas kategorik merupakan awal mula lahirnya regresi logistik. 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. Apabila variabel respon Y_i memiliki kategori biner atau memiliki dua kemungkinan maka regresi logistik yang digunakan adalah

regresi logistik biner. Kategori-kategori tersebut biasanya diberi kode 1 (kejadian sukses) dan 0 (kejadian gagal).

Peluang Y_i sukses untuk suatu variabel X_i adalah π_i atau didefinisikan sebagai $P(Y_i = 1|X_i) = \pi_i$ sedangkan peluang Y_i gagal untuk suatu variabel X_i adalah $1 - \pi_i$ atau didefinisikan sebagai $P(Y_i = 0|X_i) = 1 - \pi_i$.

Misal diberikan model regresi untuk respon biner adalah sebagai berikut

$$Y_i = X_i^T \beta + \varepsilon_i \tag{2.3}$$

dimana

$X_i^T = \begin{pmatrix} 1 & X_{i1} & X_{i2} & \cdots & X_{ik} \end{pmatrix}$

$\beta^T = \begin{pmatrix} \beta_0 & \beta_1 & \beta_2 & \cdots & \beta_k \end{pmatrix}$

Karena variabel respon Y_i bernilai 0 dan 1. Sehingga kondisi tersebut mengakibatkan bahwa Y_i adalah berdistribusi Bernoulli.

Akan dicari nilai harapan untuk variabel respon yaitu

$$\begin{aligned} E(Y_i) &= \sum_{y_i}^{1} y_i \; P(Y_i = y_i|X_i) \\ &= 1 \times P(Y_i = 1|X_i) \; + \; 0 \times P(Y_i = 0|X_i) \\ &= 1 \times \pi_i \; + \; 0 \times (1 - \pi_i) \\ &= \pi_i \end{aligned}$$

Dari Persamaan (2.3) dapat dibentuk $E(Y_i) = E(X_i^T \beta) + E(\varepsilon_i)$. Karena $E(\varepsilon_i) = 0$ dan $X_i^T \beta$ adalah suatu konstanta, maka $E(Y_i) = X_i^T \beta$. Sehingga

$$E(Y_i) = \pi_i = X_i^T \beta \tag{2.4}$$

Ada beberapa masalah pada model Persamaan (2.3) yang harus diperhatikan. Pertama, jika variabel respon yang dimiliki adalah biner, maka *error* ε_i hanya memiliki dua nilai yaitu

1. $\varepsilon_i = 1 - X_i^T\beta$, untuk $Y_i = 1$
2. $\varepsilon_i = -X_i^T\beta$, untuk $Y_i = 0$

Sehingga *error* tidak menyebar normal. Hal ini melanggar asumsi dalam regresi linier yang menyatakan bahwa *error* harus menyebar normal. Untuk permasalahan yang kedua, perhatikan bahwa :

$$\begin{aligned}
\sigma^2_{Y_i} &= E[Y_i - E(Y_i)]^2 \\
&= \sum_{y_i}^{i}[Y_i - E(Y_i)]^2 \; P(Y_i = y_i|X_i) \\
&= [1 - E(Y_i)]^2 P(Y_i = 1|X_i) \; + \; [0 - E(Y_i)]^2 P(Y_i = 0|X_i) \\
&= (1 - \pi_i)^2\pi_i \; + \; (0 - \pi_i)^2(1 - \pi_i) \\
&= (1 - 2\pi_i + \pi_i^2)\pi_i \; + \; \pi_i^2(1 - \pi_i) \\
&= \pi_i - 2\pi_i^2 + \pi_i^3 + \pi_i^2 - \pi_i^3 \\
&= \pi_i - \pi_i^2 \\
&= \pi_i(1 - \pi_i)
\end{aligned}$$

Dapat dilihat dari Persamaan (2.3) dan Persamaan (2.4), bahwa $X_i^T\beta = \pi_i$ (suatu konstanta) sehingga dapat dibentuk

$$\begin{aligned}
\sigma^2_{\varepsilon_i} &= \sigma^2_{Y_i} \\
&= \pi_i(1 - \pi_i) \\
&= X_i^T\beta(1 - X_i^T\beta)
\end{aligned}$$

Oleh karena itu, dapat disimpulkan bahwa variansi dari *error* bergantung pada $X_i^T\beta$, karena $X_i^T\beta$ tidak konstan (heteroskedastis). Hal ini melanggar asumsi bahwa variansi dari *error* harus konstan untuk semua variabel bebas (homoskedastis). Masalah selanjutnya yang ketiga adalah karena variabel respon Y_i yang dimiliki biner, maka model Persamaan (2.3) melanggar asumsi bahwa Y_i harus distribusi normal. Terakhir, karena π_i adalah peluang, maka nilai π_i harus terletak antara 0 dan 1 atau $0 \leq \pi_i \leq 1$. Sedangkan dalam Persamaan (2.4) π_i merupakan fungsi linier sehingga nilai π mengambil dari sepanjang garis bilangan riil.

Dari masalah diatas maka regresi dengan variabel respon yang memiliki kategori biner tidak dapat dimodelkan dengan fungsi linier. Oleh karena itu, perlu dicari suatu fungsi penghubung untuk mentransformasi model dalam persamaan (2.3) dan (2.4) sehingga π_i dapat dihubungkan dengan prediktor linier. Fungsi penghubung yang tepat untuk π_i berupa peluang dan memiliki jangkauan nilai dari 0 sampai 1 adalah penghubung logit yaitu $\ln(\frac{\pi_i}{1-\pi_i})$, dilambangkan dengan *Logit* (π_i). Fungsi tersebut memodelkan logaritma natural dari *odds*, $\frac{\pi_i}{1-\pi_i}$. *Odds* adalah rasio dari peluang kejadian sukses terhadap peluang kejadian gagal. Dengan demikian diperoleh model regresi logistik sebagai berikut

$$Logit\ (\pi_i) = \ln(\frac{\pi_i}{1-\pi_i}) = X_i^T\beta \tag{2.5}$$

Model regresi logistik sering disebut juga model logit. Mengingat bahwa nilai π_i dibatasi pada jangkauan $0-1$, karena logit dapat berupa sebarang bilangan riil, sehingga model ini tidak mempunyai masalah struktural yang dimiliki model peluang linier.

Selanjutnya model regresi logistik di atas dapat dikonversi ke dalam ben-

tuk *odds* yaitu

$$\begin{aligned} \ln[\frac{\pi_i}{1-\pi_i}] &= X_i^T\beta \\ \exp\left(\ln[\frac{\pi_i}{1-\pi_i}]\right) &= \exp(X_i^T\beta) \\ \frac{\pi_i}{1-\pi_i} &= \exp(X_i^T\beta) \end{aligned} \tag{2.6}$$

Model tersebut juga dapat dikonversi ke dalam bentuk peluang π_i, yaitu

$$\begin{aligned} \ln[\frac{\pi_i}{1-\pi_i}] &= X_i^T\beta \\ \exp\left(\ln[\frac{\pi_i}{1-\pi_i}]\right) &= \exp(X_i^T\beta) \\ \frac{\pi_i}{1-\pi_i} &= \exp(X_i^T\beta) \\ \pi_i &= (1-\pi_i)\exp(X_i^T\beta) \\ \pi_i &= \exp(X_i^T\beta) - \pi_i\exp(X_i^T\beta) \\ \pi_i + \pi_i\exp(X_i^T\beta) &= \exp(X_i^T\beta) \\ \pi_i(1+\exp(X_i^T\beta)) &= \exp(X_i^T\beta) \\ \pi_i &= \frac{\exp(X_i^T\beta)}{1+\exp(X_i^T\beta)} \end{aligned}$$

2.4 Metode Maksimum *Likelihood*

Dalam analisis regresi linier klasik metode yang paling banyak digunakan untuk menduga parameter adalah metode kuadrat terkecil (*least square*) yaitu parameter dicari dengan cara meminimumkan jumlah kuadrat dari galat.

Tetapi metode tersebut tidak dapat digunakan untuk menduga parameter pada regresi logistik, karena melanggar asumsi variansi *error* yaitu tidak konstan (homoskedastis).

Metode kuadrat terkecil diasumsikan bahwa variansi *error* bersifat konstan (homogen), sedangkan dalam regresi logistik variansi dari *error* tidak konstan (heterogen). Metode yang dapat digunakan untuk menduga parameter regresi logistik adalah metode Maksimum *Likelihood*. Metode maksimum *likelihood* digunakan untuk menduga parameter dalam regresi logistik dengan cara memaksimumkan fungsi *likelihood*nya (Hosmer dan Lemeshow, 2000). Karena Y_i menyebar Bernoulli maka fungsi distribusi peluang untuk Y_i adalah

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

dengan Y_i bernilai 0 atau 1 untuk masing-masing amatan. Karena Y_i setiap amatan diasumsikan saling bebas, maka fungsi *likelihood*nya adalah

$$\mathrm{l}(\beta) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} (\pi_i)^{y_i}[1-\pi_i]^{1-y_i} \tag{2.7}$$

Secara matematis akan mempermudah dalam perhitungan dengan menggunakan fungsi *log likelihood*

$$\begin{aligned} \mathrm{L}(\beta) &= \ln[\mathrm{l}(\beta)] \\ &= \sum_{i=1}^{n} [y_i \ln \pi_i + (1-y_i) \ln[1-\pi_i]] \\ &= \sum_{i=1}^{n} [y_i \ln \pi_i + \ln[1-\pi_i] - y_i \ln[1-\pi_i]] \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} [y_i(\ln \pi_i - \ln[1-\pi_i]) + \ln[1-\pi_i]] \\
&= \sum_{i=1}^{n} \left[y_i \ln \frac{\pi_i}{1-\pi_i} + \ln \frac{1}{1+e^{X_i^T \beta}} \right] \\
&= \sum_{i=1}^{n} \left[y_i (X_i^T \beta) - \ln[1+e^{X_i^T \beta}] \right]
\end{aligned}
\tag{2.8}
$$

Untuk memperoleh nilai β maka memaksimumkan nilai $\mathrm{L}(\beta)$ dengan cara mendiferensialkan $\mathrm{L}(\beta)$ terhadap β_0 dan β_j, untuk $j = 1, 2, \ldots, k$, kemudian samakan hasilnya dengan 0. Persamaan ini dapat ditulis sebagai berikut

$$
\begin{aligned}
\frac{\partial(\mathrm{L}(\beta))}{\partial(\beta_0)} &= \sum_{i=1}^{n} \left(y_i - \frac{e^{X_i^T \beta}}{1+e^{X_i^T \beta}} \right) \\
0 &= \sum_{i=1}^{n} (y_i - \pi_i)
\end{aligned}
\tag{2.9}
$$

dan

$$
\begin{aligned}
\frac{\partial(\mathrm{L}(\beta))}{\partial(\beta_j)} &= \sum_{i=1}^{n} \left(y_i X_{ij} - \frac{X_{ij} e^{X_i^T \beta}}{1+e^{X_i^T \beta}} \right) \\
0 &= \sum_{i=1}^{n} X_{ij} (y_i - \pi_i)
\end{aligned}
\tag{2.10}
$$

untuk $j = 1, 2, \ldots, k$.

Apabila dibentuk dalam matriks, untuk turunan pertama dari fungsi *likelihood* adalah

$$\mathrm{L}^{'}(\beta) = \begin{pmatrix} \frac{\partial(\mathrm{L}(\beta))}{\partial(\beta_0)} \\ \frac{\partial(\mathrm{L}(\beta))}{\partial(\beta_i)} \\ \vdots \\ \frac{\partial(\mathrm{L}(\beta))}{\partial(\beta_k)} \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_{11} & X_{21} & \cdots & X_{n1} \\ X_{12} & X_{22} & \cdots & X_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1k} & X_{2k} & \cdots & X_{nk} \end{pmatrix} \left(\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_n \end{pmatrix} \right)$$

atau

$$\begin{aligned} \mathrm{L}^{'}(\beta) &= X^T Y_i - X^T \pi_i \\ &= X^T (Y_i - \pi_i) \end{aligned} \tag{2.11}$$

Sebelum mencari penduga parameter, sebaiknya dicari terlebih dahulu penduga variansi dan kovariansi dari parameter β. Penduga variansi dan kovariansi diperoleh dari matriks turunan kedua untuk fungsi *likelihood*, sehingga bentuk turunan kedua dari fungsi *likelihood* adalah

$$\begin{aligned} \frac{\partial^2(\mathrm{L}(\beta))}{\partial(\beta_0^2)} &= -\sum_{i=1}^{n} \frac{e^{X_i^T \beta}}{(1 + e^{X_i^T \beta})^2} \\ &= -\sum_{i=1}^{n} \pi_i \left(1 - \pi_i\right) \end{aligned} \tag{2.12}$$

$$\begin{aligned} \frac{\partial^2(\mathrm{L}(\beta))}{\partial(\beta_j)\partial(\beta_0)} &= -\sum_{i=1}^{n} \frac{X_{ij} e^{X_i^T \beta}}{(1 + e^{X_i^T \beta})^2} \\ &= -\sum_{i=1}^{n} X_{ij} \pi_i \left(1 - \pi_i\right) \end{aligned} \tag{2.13}$$

$$\begin{aligned}\frac{\partial^2(\mathrm{L}(\beta))}{\partial(\beta_j^2)} &= -\sum_{i=1}^{n}\frac{X_{ij}^2 e^{X_i^T\beta}}{(1+e^{X_i^T\beta})^2} \\ &= -\sum_{i=1}^{n} X_{ij}^2\pi_i\left(1-\pi_i\right) \qquad (2.14)\end{aligned}$$

$$\begin{aligned}\frac{\partial^2(\mathrm{L}(\beta))}{\partial(\beta_p)\partial(\beta_j)} &= -\sum_{i=1}^{n}\frac{X_{ij}X_{ip} e^{X_i^T\beta}}{(1+e^{X_i^T\beta})^2} \\ &= -\sum_{i=1}^{n} X_{ij}X_{ip}\pi_i\left(1-\pi_i\right) \qquad (2.15)\end{aligned}$$

untuk $j, p = 1, 2, \ \ldots \ , k$

Sebuah matriks berukuran $(k+1)\times(k+1)$ yang semua elemen anggotanya adalah negatif turunan kedua dari fungsi *likelihood*, notasikan matriks tersebut dengan $\mathbf{I}(\beta)$. Matriks ini dinamakan matriks informasi, sehingga bentuk matriks dapat didefinisikan

$$\mathbf{I}(\beta) = \begin{pmatrix} -\frac{\partial^2(\mathrm{L}(\beta))}{\partial(\beta_0^2)} & -\frac{\partial^2(\mathrm{L}(\beta))}{\partial(\beta_1)\partial(\beta_0)} & \cdots & -\frac{\partial^2(\mathrm{L}(\beta))}{\partial(\beta_k)\partial(\beta_0)} \\ -\frac{\partial^2(\mathrm{L}(\beta))}{\partial(\beta_0)\partial(\beta_1)} & -\frac{\partial^2(\mathrm{L}(\beta))}{\partial(\beta_1^2)} & \cdots & -\frac{\partial^2(\mathrm{L}(\beta))}{\partial(\beta_k)\partial(\beta_1)} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{\partial^2(\mathrm{L}(\beta))}{\partial(\beta_0)\partial(\beta_k)} & -\frac{\partial^2(\mathrm{L}(\beta))}{\partial(\beta_1)\partial(\beta_k)} & \cdots & -\frac{\partial^2(\mathrm{L}(\beta))}{\partial(\beta_k^2)} \end{pmatrix}$$

Dengan sifat perkalian matriks, maka matriks informasi $\mathbf{I}(\beta)$ dapat ditulis dengan

$$\mathbf{I}(\beta) = X^T V X$$

dimana X adalah matriks dari variabel bebas berukuran $n \times (k+1)$, dan V merupakan matriks diagonal yang anggota elemen diagonalnya adalah $\pi_i(1-\pi_i)$, sehingga bentuk matriks X adalah

$$X = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \cdots & \vdots & \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix}$$

dan bentuk matriks V adalah

$$V = \begin{pmatrix} \pi_1(1-\pi_1) & 0 & \cdots & 0 \\ 0 & \pi_2(1-\pi_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_n(1-\pi_n) \end{pmatrix}$$

Adapun matriks variansi dan kovariansi dari parameter β diperoleh dengan cara menginverskan matriks informasi, yang dinotasikan dengan $Var(\beta) = \mathbf{I}^{-1}(\beta)$. Dengan demikian, penduga untuk variansi parameter β adalah elemen-elemen diagonal dari matriks $Var(\beta)$ dan untuk penduga kovariansi parameter β adalah elemen-elemen non diagonal dari matriks $Var(\beta)$ sehingga dapat dihitung pada nilai penduga $\beta(\widehat{\beta})$.

Karena persamaan *likelihood* tidak linier dalam parameter β, untuk menyelesaikan solusi persamaan nonlinier di atas maka dibutuhkan proses iterasi unuk

menduga parameter secara numerik. Salah satu metode yang dapat digunakan adalah Newton Raphson.

2.5 Metode Newton-Raphson

Menduga parameter menggunakan metode maksimum *likelihood* akan menghasilkan persamaan *likelihood* yang nonlinier. Untuk mendapatkan penduga parameter yang dapat menyelesaikan persamaan nonlinier tersebut digunakan metode Newton-Raphson. Berikut akan diberikan prosedur untuk mencari parameter β dengan menggunakan metode Newton Raphson, yaitu :

1. Pilih taksiran awal untuk $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$, misalkan $\widehat{\beta}^0 = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_k \end{pmatrix}^0$

2. Hitung nilai taksiran untuk β pada iterasi ke-$(q+1)$ dimana $q = 0, 1, 2, \ldots$, maka $\widehat{\beta}^{(q+1)}$, sehingga

$$\begin{aligned} \widehat{\beta}^{(q+1)} &= \widehat{\beta}^{(q)} - [-\mathbf{I}(\widehat{\beta}^{(q)})]^{-1}\mathrm{L}'(\widehat{\beta}^{(q)}) \\ &= \widehat{\beta}^{(q)} + [\mathbf{I}(\widehat{\beta}^{(q)})]^{-1}\mathrm{L}'(\widehat{\beta}^{(q)}) \\ &= \widehat{\beta}^{(q)} + [X^T V X]^{-1} X^T (Y_i - \pi_i) \end{aligned}$$

dimana

$\mathrm{L}'(\widehat{\beta}^{(q)})$ merupakan matriks turunan pertama yang dihitung nilai parameter $\beta = \widehat{\beta}^{(q)}$ pada fungsi *log likelihood*

$\mathbf{I}(\widehat{\beta}^{(q)})$ adalah matriks informasi yang dihitung nilainya pada $\beta = \widehat{\beta}^{(q)}$

3. Hentikan proses iterasi jika $\widehat{\beta}^{(q+1)} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_k \end{pmatrix}^{(q+1)} \approx \widehat{\beta}^{(q)} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_k \end{pmatrix}^{(q)}$

   kemudian ambil $\widehat{\beta}^{(q+1)} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_k \end{pmatrix}^{(q+1)}$ sebagai taksiran $\widehat{\beta} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_k \end{pmatrix}^{0}$

2.6 Pengujian Signifikansi Parameter

Setelah menaksir parameter maka langkah selanjutnya yang dilakukan adalah menguji signifikansi parameter tersebut. Untuk itu pengujian terhadap parameter model dilakukan untuk memeriksa peranan variabel-variabel bebas yang berpengaruh signifikan terhadap variabel respon. Pengujian signifikansi parameter dilakukan sebagai berikut :

1. Uji Serentak

   Uji Serentak disebut juga uji kesesuaian model, pengujian hipotesis ini bertujuan untuk melihat peranan variabel bebas terhadap variabel respon dalam model secara bersama-sama.

   Perumusan hipotesisnya adalah :

   $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$

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

Statistik uji yang digunakan untuk menguji hipotesis tersebut adalah statistik uji G atau uji rasio *likelihood* :

$$\begin{aligned} G &= -2\ln\left[\frac{L_0}{L_1}\right] \\ &= -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)}{\prod_{i=1}^{n}\widehat{\pi}_i^{\,y_i}(1-\widehat{\pi}_i)^{(1-y_i)}}\right] \end{aligned} \tag{2.16}$$

atau

$$G = 2\left(\sum_{i=1}^{n}\left[y_i\ln(\widehat{\pi}_i) + (1-y_i)\ln(1-\widehat{\pi}_i)\right] - \left[n_1\ln(n_1) + n_0\ln(n_0) - n\ln(n)\right]\right)$$

dimana

n_1	:	banyaknya amatan yang berkategori 1
n_0	:	banyaknya amatan yang berkategori 0
n	:	banyaknya amatan $(n_1 + n_0)$
L_0	:	*Likelihood* tanpa variabel bebas tertentu
L_1	:	*Likelihood* dengan variabel bebas tertentu

Apabila terima H_0 maka statistik uji G mengikuti distribusi *chi-square* (χ^2), sehingga untuk memperoleh keputusan dilakukan perbandingan nilai χ^2 tabel dengan derajat bebas k. Kriteria penolakan (tolak H_0) adalah jika nilai $G > \chi^2_{db,\alpha}$, sehingga model regresi tersebut cocok untuk menjelaskan hubungan antara variabel bebas terhadap variabel respon.

2. Uji Parsial

   Pengujian hipotesis ini digunakan untuk menguji pengaruh setiap β_i secara individual. Hasil pengujian secara parsial akan menunjukkan apakah suatu variabel bebas layak untuk masuk dalam model atau tidak.

   Perumusan hipotesisnya adalah :

   $H_0 : \beta_i = 0$

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

   Statistik uji yang digunakan untuk menguji hipotesis tersebut adalah statistik uji *Wald*

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

   dengan $\widehat{\beta}_i$ = nilai dugaan β_i dan $SE(\widehat{\beta}_i) = \sqrt{var(\widehat{\beta}_i)}$

   Apabila terima H_0 maka statistik uji Wald dibandingkan dengan sebaran *chi-square* (χ^2) pada taraf nyata α dan derajat bebas sama dengan 1, sehingga kriteria penolakan (tolak H_0) jika nilai $W_j > \chi^2_{\alpha,1}$ maka variabel bebas (X_j) dengan $j = 1, 2, \ \ldots \ , k$ memiliki pengaruh terhadap variabel respon (Y) pada taraf nyata α.

3. Uji Kelayakan Model

   Uji kelayakan model digunakan untuk menilai apakah model sesuai dengan data atau tidak. Untuk mengetahui apakah model sesuai atau tidak berdasarkan hasil uji Hosmer dan Lemeshow atau sering disebut juga *Goodness of fit test*. Pengujian ini dilakukan untuk menguji hipotesis sebagai berikut

H_0 : Tidak terdapat perbedaan signifikan antara klasifikasi yang diprediksi dengan klasifikasi yang diamati

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

Pengambilan keputusan dilakukan dengan melihat nilai *goodness of fit test* yang diukur dengan nilai *chi-square*. Keputusan uji Hosmer dan Lemeshow adalah apabila p-*value* > 0.05 maka tolak H_0, maka tidak ada perbedaan perbedaan yang signifikan antara klasifikasi yang diprediksi dengan klasifikasi yang diamati. Oleh karena itu dapat disimpulkan bahwa model mampu memprediksi nilai data. Sedangkan apabila p-*value* < 0.05 maka terdapat perbedaan yang signifikan antara klasifikasi yang diprediksi dengan klasifikasi yang diamati, dengan demikian model tidak mampu memprediksi nilai data.

2.7 Rasio *Odds*

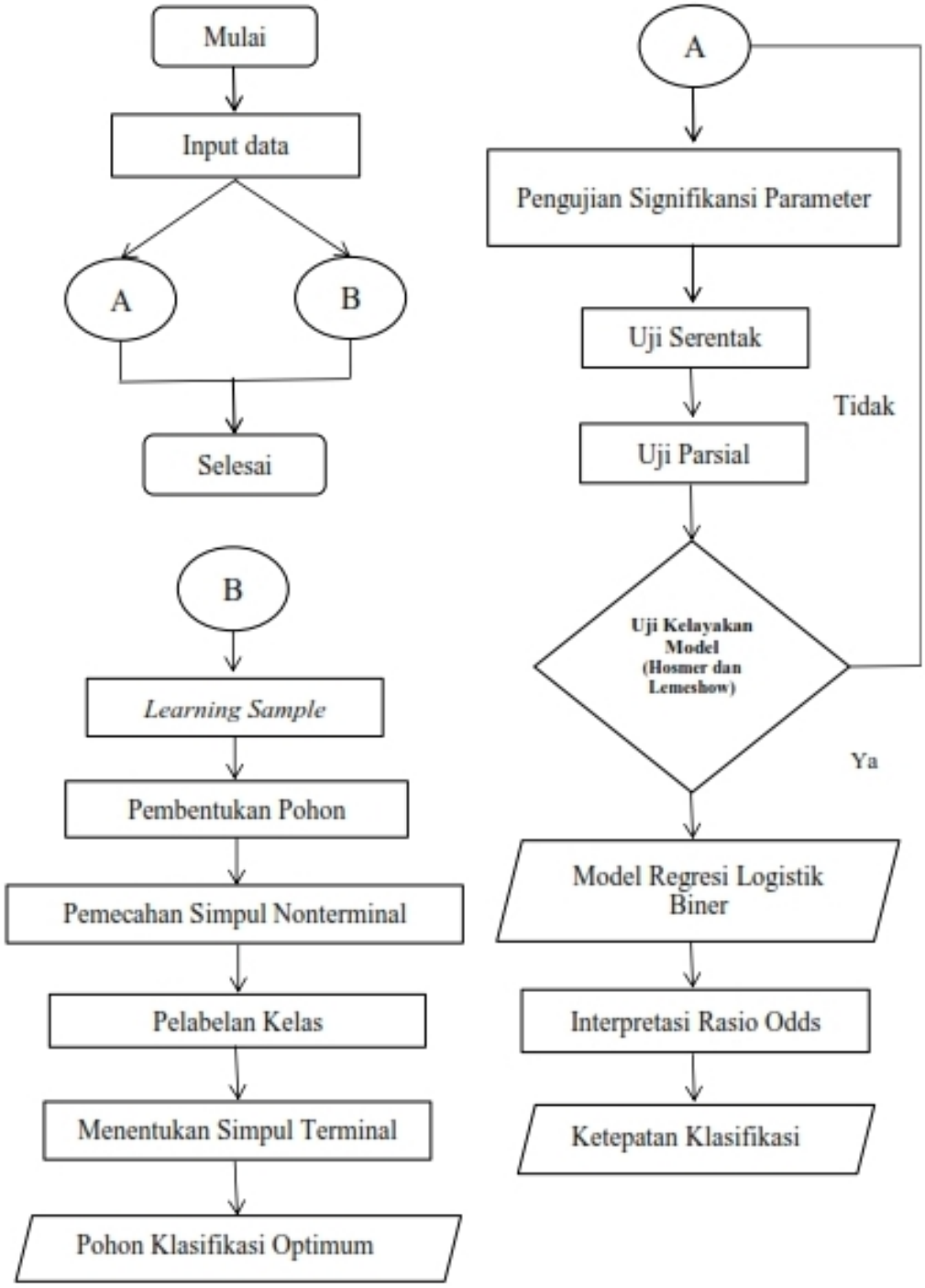
Setelah model dinyatakan layak dalam menggambarkan hubungan variabel bebas dan variabel respon maka langkah selanjutnya adalah menginterpretasikan model tersebut dalam penarikan kesimpulan. Secara bahasa *odds* mempunyai arti yang sama dengan peluang, akan tetapi di dalam statistik peluang dan *odds* mempunyai konsep berbeda. *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

$$\begin{aligned}\psi &= \frac{\pi(1) \setminus [1-\pi(1)]}{\pi(0) \setminus [1-\pi(0)]} \\ &= \frac{\pi(1)\ [1-\pi(0)]}{\pi(0)\ [1-\pi(1)]} \\ &= \frac{\exp^{\beta_0+\beta_1}}{\exp^{\beta_0}} = \exp^{\beta_1}\end{aligned}$$

Sedangkan nilai log rasio *odds* adalah

$$\begin{aligned}\ln\psi &= \ln(\exp^{\beta_1}) = \beta_1 \\ &= g(1) - g(0)\end{aligned}$$

Interpretasi dari rasio *odds* ini adalah kecenderungan untuk $Y = 1$ pada $X = 1$ sebesar ψ kali dibandingkan pada $X = 0$. Bila nilai $\psi = 1$, maka antara kedua variabel tersebut tidak terdapat hubungan. Bila nilai $\psi < 1$, maka antara kedua variabel terdapat hubungan negatif terhadap perubahan nilai X yang bernilai benar dan sebaliknya bila nilai $\psi > 1$.



Gambar 2.1: Diagram Alir Model Regresi Logistik Biner dan Metode CART

BAB III

PEMBAHASAN

3.1 Klasifikasi Regresi Logistik Biner

Analisis regresi logistik biner merupakan salah satu metode klasifikasi dalam pendekatan parametrik yaitu untuk menganalisis data dengan variabel respon yang memiliki dua kategori dengan satu atau lebih variabel prediktor berskala kategorik maupun kontinu. Hosmer dan Lemeshow (2000), menjelaskan bahwa model regresi logistik biner dibentuk dengan nilai $P(Y_i = 1|X_i)$ sebagai π_i, yang dinotasikan sebagai berikut

$$\pi_i = \frac{\exp(X_i^T\beta)}{1 + \exp(X_i^T\beta)} = \frac{e^{(\beta_0+\beta_1 x_i)}}{1 + e^{(\beta_0+\beta_1 x_i)}} \tag{3.1}$$

Suatu fungsi dari π_i dicari dengan mentransformasi model dalam Persamaan (2.3) dan (2.4) sehingga diperoleh fungsi penghubung yang tepat untuk π_i yaitu penghubung *Logit* (π_i) yang didefinisikan sebagai berikut

$$\textit{Logit}\ (\pi_i) = \ln(\frac{\pi_i}{1 - \pi_i}) = X_i^T\beta \tag{3.2}$$

Secara umum jika terdapat variabel bebas berupa data kategori dan variabel tersebut memiliki banyak p kategori yang mungkin, maka diperlukan sebanyak $p - 1$ variabel *dummy* pada model. Dengan demikian, model regresi

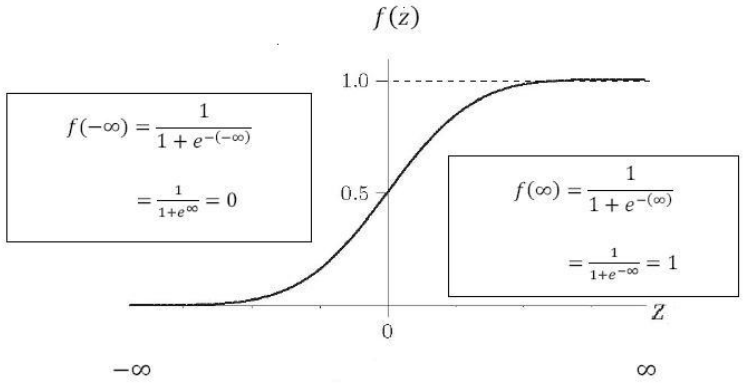
logistik dengan k variabel bebas ke-j berupa data kategori menjadi

$$Logit(\pi_i) = \ln(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 X_{i1} + \ \ldots \ + \sum_{l=1}^{k_j-1} B_{jl} D_{jl} + \ \ldots \ + \beta_k X_{ik} \quad (3.3)$$

dimana

x_j	:	variabel bebas ke-j berupa data dengan p_j kategori
$p_j - 1$	:	jumlah variabel *dummy*
D_{jl}	:	$p_j - 1$ variabel *dummy*
B_{jl}	:	koefisien dari variabel *dummy*
l	:	$1, 2, \ \ldots \ , p_j - 1$

Kurva regresi logistik digambarkan dalam bentuk S signoid dengan nilai z berkisar antara $-\infty$ sampai dengan $+\infty$ dan nilai $f(z)$ bergerak dari 0 sampai 1. Ketika nilai z mendekati $-\infty$ maka nilai $f(z)$ bergerak mendekati nilai 0, dan jika nilai z mendekati $+\infty$ maka nilai $f(z)$ bergerak mendekati nilai 1. Kurva regresi logistik dapat dilihat pada Gambar 3.1 sebagai berikut :



Gambar 3.1: Kurva Regresi Logistik

3.1.1 Ketepatan Klasifikasi

Menurut Johnson dan Wichern (2007) Ketepatan klasifikasi adalah suatu evaluasi yang melihat peluang kesalahan yang dilakukan oleh suatu fungsi klasifikasi. Jika analisis regresi logistik digunakan untuk mengklasifikasikan data amatan, maka perlu diuji keakuratan pada amatan dari sampel lain. Ketepatan hasil klasifikasi dapat dihitung dengan nilai *Apparent Error Rate* (APER) yaitu persentase dari amatan yang salah dalam pengklasifikasian terhadap jumlah total amatan.

Menentukan nilai APER dapat dihitung dengan mudah dalam bentuk *crosstab*, jika menghitung ketepatan klasifikasi adalah penjumlahan dari n_{11} dan n_{22} dibagi dengan total amatan N , maupun sebaliknya jika menentukan salah dalam klasifikasi yaitu penjumlahan dari n_{12} dan n_{21} dibagi dengan total amatan N. N_1 adalah total amatan yang dibentuk dari grup π_1 dan N_2 adalah total amatan yang dibentuk dari grup π_2. Berikut disajikan dalam tabel untuk menghitung ketepatan klasifikasi sebagai berikut :

Tabel 3.1: Tabel Ketepatan Klasifikasi

Observasi	Prediksi		Total
	$\widehat{\pi_1}$	$\widehat{\pi_2}$	Prediksi
π_1	n_{11}	n_{12}	N_1
π_2	n_{21}	n_{22}	N_2
Total	N_1	N_2	N

dengan

n_{11} : Total amatan dalam grup π_1 yang tepat dan diklasifikasikan sebagai grup π_1

n_{12} : Total amatan dalam grup π_1 yang salah dan diklasifikasikan sebagai grup π_2

n_{21} : Total amatan dalam grup π_2 yang salah dan diklasifikasikan sebagai grup π_1

n_{22} : Total amatan dalam grup π_2 yang tepat dan diklasifikasikan sebagai grup π_2

sehingga nilai APER dapat didefinisikan sebagai

$$APER \text{ (dalam \%)} = \frac{n_{12} + n_{21}}{N_1 + N_2}$$

Ketepatan prediksi klasifikasi secara tepat dapat dihitung menggunakan nilai *Total Accuracy Rate* $(1 - APER)$ yaitu persentase dari amatan yang tepat dalam pengklasifikasian terhadap jumlah total amatan, didefinisikan sebagai

$$1 - APER \text{ (dalam \%)} = \frac{n_{11} + n_{22}}{N_1 + N_2}$$

3.2 Metode *Classification And Regression Trees* (CART)

CART (*Classification And Regression Trees*) 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. Menurut Breiman *et, al.* (1993), CART merupakan metodologi statistik nonparametrik yang dikembangkan untuk analisis klasifikasi, baik untuk variabel respon kategorik maupun kontinu.

CART menghasilkan suatu pohon klasifikasi (*classification trees*) jika variabel responnya kategorik, dan menghasilkan pohon regresi (*regression trees*) jika variabel responnya kontinu. Klasifikasi CART ini memiliki banyak keunggulan, diantaranya pada variabel-variabel dalam CART baik variabel respon maupun prediktor tidak menggunakan asumsi distribusi tertentu, variabel responnya dapat bertipe kategorik (nominal ataupun ordinal) maupun kontinu, tidak berlaku adanya transformasi data dan interpretasinya mudah dipahami. Tujuan metode ini adalah untuk mendapatkan suatu kelompok data yang akurat sebagai penciri dari suatu pengklasifikasian.

3.2.1 Pohon Klasifikasi

Sebuah data dapat dinyatakan sebagai vektor komponen berdasarkan pengukuran karakteristik tertentu, dinyatakan sebagai vektor $\widetilde{x}$. Misalkan banyak karakteristik adalah M maka $\widetilde{x}$ dituliskan sebagai vektor komponen M, yaitu $\widetilde{x} = x_1, x_2, \ \ldots \ , x_M$. Himpunan dari vektor tersebut dinotasikan dengan X disebut ruang amatan, $X = \{\widetilde{x_n} = (x_{n1}, x_{n2}, \ \ldots \ , x_{nM}) \ ; \ n = 1, 2, \ \ldots \ , N\}$. Berdasarkan suatu kriteria tertentu data ini dapat dikelompokkan menjadi j kelas dimana $j = 1, 2, \ \ldots \ , J$. Sebut C sebagai himpunan dari semua kelas, dituliskan $C = \{1, 2, \ \ldots \ , J\}$. Akan dicari suatu cara klasifikasi yang sistematis untuk memprediksi keanggotaan suatu kelas $j = 1, 2, \ \ldots \ , J$ untuk setiap $\widetilde{x} \ \epsilon \ X$. Penjabaran cara klasifikasi ini menurut Breiman et al (1993) menggunakan definisi-definisi berikut

Definisi 3.2.1. Suatu pengklasifikasian adalah suatu fungsi $d(\widetilde{x})$ yang didefinisikan pada X, sedemikian sehingga untuk setiap $\widetilde{x} \ \epsilon \ X$, $d(\widetilde{x})$ sama dengan salah satu dari $1, 2, \ \ldots \ , J$. Maka dapat ditulis $d(\widetilde{x}) = j$ untuk suatu $j \ \epsilon \ C$

Pembentukan suatu klasifikasi berdasarkan sekumpulan kelas yang sudah ditentukan dari data disebut *Learning sample* atau *training sample*. Data *Learning sample* terdiri dari N data amatan dan M karakteristik yang diukur.

Definisi 3.2.2. Diberikan *Learning sample* $\mathcal{L} = \{(\widetilde{x}_1, j_1), (\widetilde{x}_2, j_2), \ \ldots \ , (\widetilde{x}_N, j_N)\}$, dimana $\widetilde{x} \ \epsilon \ X$ dan $j_n \ \epsilon \ 1, 2, \ \ldots \ , J = C$, $n = 1, 2, \ \ldots \ , N$. N menyatakan besar sampel.

Pada suatu prediktor diberikan suatu fungsi $d(\widetilde{x})$ yang didefinisikan pada X dan juga diberi nilai dari C. Notasikan $R^*(d)$ sebagai tingkat kesalahan klasifikasi. Keakuratan dari suatu prediktor diukur dengan menaksir $R^*(d)$ yaitu dengan cara menguji prediktor pada pengamatan sejumlah data yang telah diketahui klasifikasinya. Didefinisikan ruang $X \times C$ sebagai himpunan semua pasangan $(\widetilde{x}, j)$ dimana $\widetilde{x} \ \epsilon \ X$ dan $j \ \epsilon \ C$. Misalkan $P(A, j)$ dimana $A \ \subset \ X$ adalah probabilitas pada $X \times C$. Suatu data yang diambil secara random dari suatu populasi dan mempunyai $P(A, j)$ adalah vektor pengukuran $\widetilde{x}$ yang ada di A dan kelasnya adalah j. Asumsikan bahwa *learning sample* $\mathcal{L}$ terdiri dari N data $\{(\widetilde{x}_1, j_1), \ \ldots \ , (\widetilde{x}_N, j_N)\}$ yang saling bebas dan random dengan probabilitas $P(A, j)$. Bentuk $d(\widetilde{x})$ dengan menggunakan $\mathcal{L}$, kemudian ambil sampel baru yang berukuran besar dari populasi yang sama dengan $\mathcal{L}$ dan amati klasifikasinya. Probabilitas bahwa d akan salah mengklasifikasikan sampel baru tersebut didefinisikan dengan $R^*(d)$.

Definisi 3.2.3. Ambil $(\widetilde{x}, Y)$, $\widetilde{x} \ \epsilon \ X$, $Y \ \epsilon \ C$ sebagai sampel baru dari populasi dengan probabilitas $P(A, j)$ berarti $P(\widetilde{x} \ \epsilon \ A, Y = j) = P(A, j)$ dan $(\widetilde{x}, Y)$ independen atau bebas pada $\mathcal{L}$, maka dapat didefinisikan $R^*(d) = P(d(\widetilde{x}) \neq Y)$.

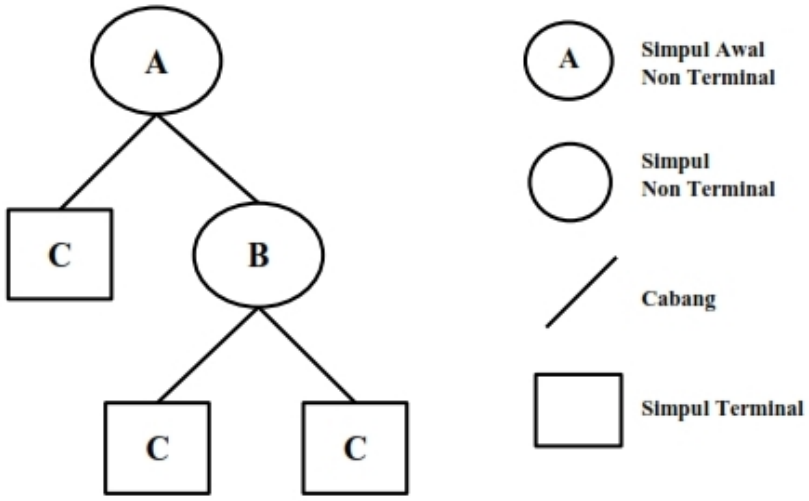
Perhitungan probabilitas $P(d(\widetilde{x}) \neq Y|\mathcal{L})$ mengacu pada himpunan $\mathcal{L}$. Notasi lain adalah $P(d(\widetilde{x}) \neq Y|\mathcal{L})$ yaitu probabilitas kesalahan mengklasifikasikan pada sampel baru bila diberikan *learning sample* $\mathcal{L}$. Dalam praktek, hanya data $\mathcal{L}$ yang digunakan untuk membentuk $d(\widetilde{x})$ maupun menaksir $R^*(d)$. Menurut (Breiman et al, 1993) salah satu cara untuk mencari taksiran dari $R^*(d)$ adalah dengan menggunakan taksiran resubtitusi, yang didefinisikan sebagai berikut

$$R(d) = \frac{1}{N}\sum_{n=1}^{N} X(d(\widetilde{x}_n) \neq j_n) \tag{3.4}$$

3.2.2 Struktur Pohon Klasifikasi CART

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.

Pada Gambar (3.2) Misalkan A, B dan C merupakan variabel-variabel yang terpilih untuk menjadi simpul dalam pembentukan pohon klasifikasi sederhana. Akan ditunjukkan pohon CART sederhana pada gambar sebagai berikut :



Gambar 3.2: Struktur Pohon Klasifikasi CART

Berikut dijelaskan keterangan dari Gambar (3.2) di atas adalah

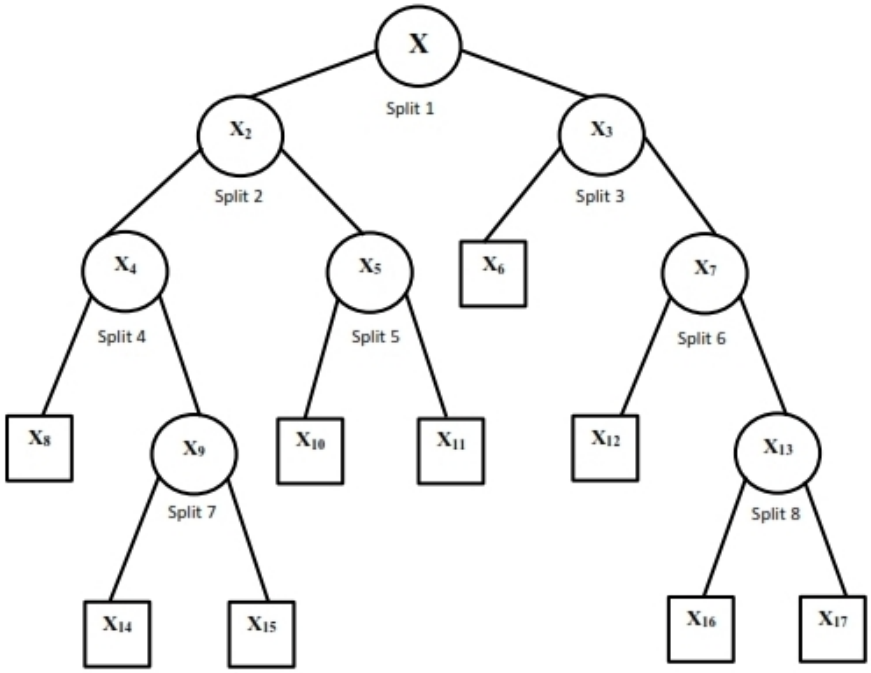
1. A merupakan simpul nonterminal yang digambarkan dengan lingkaran yaitu simpul nonterminal yang paling awal terbentuknya suatu pohon.

2. B adalah Simpul nonterminal digambarkan dengan lingkaran yang merupakan subset dari simpul nonterminal diatasnya yang memenuhi kriteria pemecahan tertentu.

3. Cabang digambarkan dengan dua garis lurus yaitu cabang dari simpul nonterminal. Cabang merupakan tempat kriteria pemecahan dari masing-masing simpul nonterminal.

4. C adalah simpul terminal digambarkan dengan persegi yang merupakan tempat diprediksikan sebuah objek pada kelas tertentu.

Struktur pohon klasifikasi atau bentuk pohon klasifikasi biner dibentuk dengan cara pemecahan berulang kali subset-subset dimulai dari X menjadi dua subset yang saling asing. Misal bentuk pohon klasifikasi maksimum dengan variabel yang terpilih menjadi simpul dalam jumlah besar yaitu terdapat 8 variabel, akan ditunjukkan pada gambar sebagai berikut :

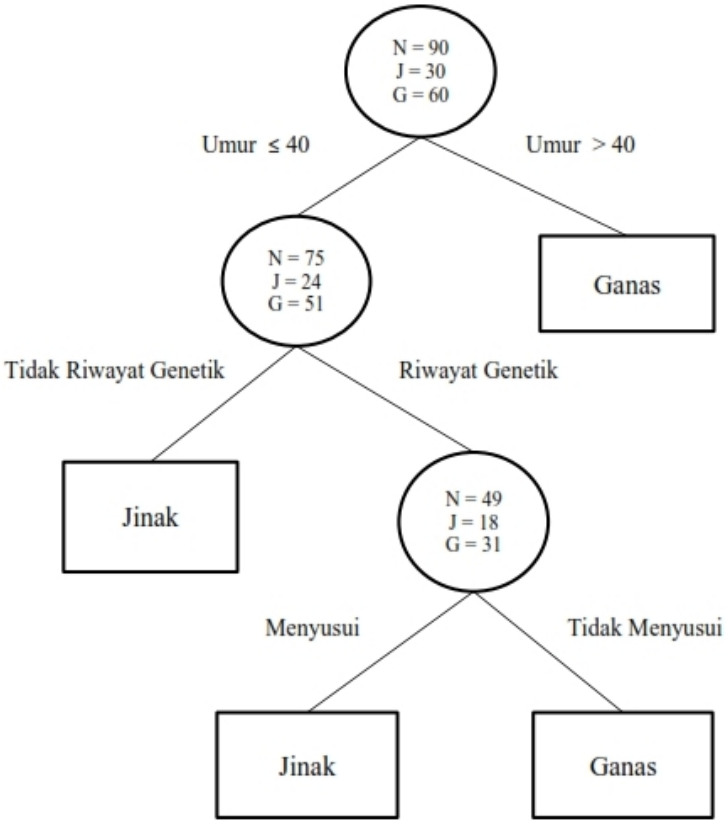


Gambar 3.3: Pohon Klasifikasi Maksimum

Pada Gambar (3.3) definisikan bahwa simpul t adalah subset dari X dan dimulai dengan simpul awal $t_1 = X$ maka $X = X_2 \cup X_3$ dimana X_2 dan X_3 saling asing. Demikian juga dengan $X_4 \cup X_5 = X_2$ dan seterusnya. Subset yang tidak dapat dipecah lagi disebut subset terminal seperti $X_8, X_{10}, X_{11}, X_{12}$ dan lainnya. Subset-subset terminal tersebut membentuk suatu partisi dari X. Setiap subset terminal mewakili suatu kelas tertentu. Suatu kelas mungkin memiliki dua atau lebih subset terminal. Subset terminal didefinisikan sebagai simpul terminal dan

subset nonterminal didefinisikan sebagai simpul nonterminal.

Interpretasi dari pohon klasifikasi yang telah terbentuk disajikan dalam suatu aturan hubungan yaitu aturan pengklasifikasian yang terbentuk **if... then...** (**jika... maka...**). Adapun diberikan contoh gambar dalam pembentukan pohon klasifikasi dari suatu data pasien kanker payudara dengan aturan-aturan klasifikasi yang telah terbentuk bahwa seorang pasien akan terdiagnosis kanker ganas atau jinak dilihat dari faktor-faktor yang mempengaruhi timbulnya penyakit kanker akan dijelaskan pada Gambar (3.4) adalah sebagai berikut



Gambar 3.4: Contoh Pohon Klasifikasi CART

Pada contoh Gambar (3.4) diatas pada simpul pertama diberikan total data sebesar 90 pasien dipilah oleh faktor usia dengan masing-masing pasien yang mengidap kanker jinak dan ganas adalah sebesar 30 dan 60. Pada simpul kedua terdapat 75 pasien yang dipilah oleh faktor riwayat genetik dengan kanker jinak sebesar 24 dan kanker ganas sebesar 51. Selanjutnya pada simpul terakhir dipilah oleh faktor tidak menyusui anak diperoleh 49 pasien dengan total 18 pasien mengidap kanker jinak dan 31 pasien mengidap kanker ganas. Berikut hasil diagnosis dari pohon klasifikasi yang telah terbentuk adalah

1. **Jika** umur > 40 **maka** kelas hasil diagnosis penyakit kanker menunjukkan "ganas"

2. **Jika** umur ≤ 40 dan tidak memiliki riwayat genetik **maka** kelas hasil diagnosis penyakit kanker menunjukkan "jinak"

3. **Jika** umur ≤ 40 **dan** memiliki riwayat genetik **dan** tidak menyusui anak **maka** kelas hasil diagnosis penyakit kanker menunjukkan "ganas"

4. **Jika** umur ≤ 40 **dan** memiliki riwayat genetik **dan** menyusui anak **maka** kelas hasil diagnosis penyakit kanker menunjukkan "jinak"

3.3 Pembentukan Pohon Klasifikasi CART

Menurut Webb dan Yohannes (1999) pada pembentukan pohon klasifikasi CART dilakukan metode awal yang digunakan untuk melihat proporsi banyaknya objek pada setiap kelas dalam $\mathcal{L}$. Pada *learning sample* $\mathcal{L}$ dengan banyaknya kelas adalah j. Misalkan N_j adalah banyaknya objek dalam kelas j, maka nilai probabilitas $\{\pi(j)\}$ ditaksir dari objek sebagai proporsi $\{\frac{N_j}{N}\}$.

Dalam sebuah simpul t, diberikan

$N(t)$: Banyaknya onjek pada $\mathcal{L}$ dimana $\widetilde{x}_n \, \epsilon \, t$.

$N_j(t)$: Banyaknya objek pada kelas j berada di simpul t.

$\frac{N_j(t)}{N(t)}$: Proporsi objek dalam kelas j pada $\mathcal{L}$ berada di simpul t.

Sehingga probabilitas dari sebuah objek merupakan anggota kelas j berada dalam simpul t adalah

$$\begin{aligned} p(j,t) &= \pi_j \frac{N_j(t)}{N_j} \\ &= \frac{N_j}{N} . \frac{N_j(t)}{N_j} \\ &= \frac{N_j(t)}{N} \end{aligned} \tag{3.5}$$

Jika $P(t)$ adalah probabilitas beberapa objek akan berada dalam simpul t, maka berdasarkan Persamaan (3.5) diperoleh

$$\begin{aligned} P(t) &= \sum_{j}^{J} P(j,t) \\ &= p(1,t) + p(2,t) + \ \ldots \ + p(J,t) \\ &= \frac{N_1}{N} + \frac{N_2}{N} + \ \ldots \ + \frac{N_j}{N} \\ &= \frac{N(t)}{N} \end{aligned} \tag{3.6}$$

Jika $P(j|t)$ adalah probabilitas sebuah objek adalah anggota kelas j pada $\mathcal{L}$ yang berada dalam simpul t, maka berdasarkan Persamaan (3.6) diperoleh

$$P(j|t) = \frac{p(j,t)}{P(t)} = \frac{\frac{N_j(t)}{N_j}}{\frac{N(t)}{N}} = \frac{N_j(t)}{N(t)} \tag{3.7}$$

Persamaan (3.7) tersebut memenuhi

$$\sum_j P(j|t) = 1$$

3.3.1 Kriteria Pemecahan Simpul Nonterminal

Pada tahap ini dicari pemilah dari setiap simpul yang menghasilkan penurunan tingkat heterogen paling tinggi. Misal himpunan S dari pemecahan-pemecahan s_k untuk setiap simpul t dihasilkan oleh aturan pemecahan. Kriteria pemecahan terbaik ini dibentuk berdasarkan *impurity* (fungsi keragaman).

Definisi 3.3.1. Fungsi *impurity* adalah sebuah fungsi ϕ yang didefinisikan pada himpunan semua J komponen oleh $(p_1, p_2, \ \ldots \ p_j)$; yang memenuhi $p_j \geq 0$ dan $\sum_j p_j = 1$ dimana $j = 1, 2, \ \ldots \ , J$.

Fungsi *impurity* ϕ memenuhi kriteria dengan ketentuan sebagai berikut :

1. ϕ maksimum apabila memenuhi nilai-nilai

$$(p_1, p_2, \ \ldots \ p_j) = \left(\frac{1}{1}, \frac{1}{2}, \ \ldots \ \frac{1}{j}\right)$$

2. ϕ minimum apabila memenuhi nilai-nilai

$$(p_1, p_2, \ \ldots \ p_j) = (1, 0, \ \ldots \ 0), \ (0, 1, \ \ldots \ 0), \ \ldots \ \ (0, 0, \ \ldots \ 1)$$

3. ϕ adalah suatu fungsi simetri dari $p_1, p_2, \ \ldots \ p_J$.

Definisi 3.3.2. Diberikan suatu fungsi *impurity* ϕ, definisikan ukuran *impurity* $i(t)$ pada sembarang simpul t sebagai

$$i(t) = \phi(P(1|t),\ P(2|t),\ \ldots\ , P(J|t))$$

Fungsi impuritas yang dapat digunakan adalah Indeks Gini. Bila impuritas suatu simpul semakin tinggi maka semakin heterogen simpul tersebut (Breiman et al, 1993). Nilai impuritas menggunakan Indeks Gini pada simpul t, maka $i(t)$ dapat ditulis sebagai berikut

$$i(t) = 1 - \sum_{j} P^2(j|t) \tag{3.8}$$

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, atau dinyatakan dengan $p_R = \frac{p(t_R)}{p(t)}$ dan $p_L = \frac{p(t_L)}{p(t)}$ maka penurunan *impurity* sebagai berikut :

$$\Delta i(s_k, t) = i(t) - p_R\ .\ i(t_R) - p_L\ .\ i(t_L) \tag{3.9}$$

Definisikan pemecahan s^* sebagai pemecahan terbaik dari suatu simpul t yaitu pemecahan yang memberikan penurunan *impurity* terbesar atau dapat dinyatakan

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

Berdasarkan Persamaan (3.9) $\Delta i(s_k, t)$ akan maksimum apabila diperoleh $p_R i(t_R)$ dan $p_L i(t_L)$ minimum. Hal ini menunjukkan bahwa pemecahan dilakukan untuk membuat dua simpul baru yang nilai keragamannya lebih kecil (homogen) jika

dibandingkan dengan simpul awalnya.

3.3.2 Penandaan Label Kelas Pada Simpul Terminal

Menentukan label kelas pada simpul terminal dilakukan berdasarkan aturan jumlah terbanyak. Misalkan sebuah pohon T telah dibentuk dan mempunyai simpul terminal.

Definisi 3.3.3. Aturan penetapan sebuah kelas $j \in \{1, 2, \ldots, J\}$ untuk setiap simpul terminal $t \in T$ dinyatakan dengan $j(t)$

Penandaan label pada kelas dibentuk dengan aturan $j(t)$ yaitu kelas yang memiliki $P(j|t)$ terbesar ditetapkan sebagai kelasnya.

Definisi 3.3.4. Aturan penetapan kelas $j^*(t)$ dinyatakan sebagai berikut :
jika $P(j|t) = \max_i p(i|t)$ maka $j^*(t) = j$. Jika nilai maksimum dicapai oleh dua atau lebih kelas, maka penetapan $j^*(t)$ dapat diambil dari kelas yang mana saja.

Sehingga bila dari semua kelas yang ada, kelas j memiliki $P(j|t)$ yang maksimum maka kelas ke $j = j^*(t)$ yang ditetapkan sebagai kelas.

3.3.3 Menentukan Simpul Terminal

Ukuran dari pohon akan terus meningkat bila pemecahan tetap berjalan. Penghentian suatu pemecahan diperlukan untuk mendapatkan solusi yang baik. Peraturan penghentian pemecahan (*stop-splitting rule*) adalah suatu kriteria yang dibuat untuk mengakhiri pemecahan, sehingga dapat menentukan kapan suatu simpul menjadi simpul terminal, yang bertujuan membatasi ukuran dari pohon klasifikasi.

Misalkan telah dilakukan beberapa pemecahan sampai pada sebuah himpunan dari simpul terminal. Notasikan himpunan dari simpul-simpul terminal dengan T dan definisikan simpul *impurity* adalah $I(t) = i(t)\, p(t)$. Sehingga *impurity* pohon $I(T)$ dinyatakan sebagai

$$I(T) = \sum_{t\ \epsilon\ T} I(T) = \sum_{t\ \epsilon\ T} i(t)p(t)$$

Proposisi 3.3.1. Pemilihan s yang memaksimalkan $\Delta I(s_k, t)$ ekuivalen dengan pemilihan s yang meminimalkan pohon *impurity* $I(t)$

Ambil sembarang simpul $t\ \epsilon\ T$, dengan menggunakan sebuah pemecahan s_k, pemecahan simpul tersebut ke dalam $t_{\mathcal{L}}$ dan t_R. Pohon baru $T^{'}$ mempunyai *impurity*

$$I(T^{'}) = \sum_{T-\{t\}} I(t) + I(t_{\mathcal{L}}) + I(t_R)$$

Penurunan dalam *impurity* pohon sebesar :

$$I(T) - I(T^{'}) = I(t) - I(t_{\mathcal{L}}) - I(t_R)$$

Penurunan ini hanya bergantung pada simpul t dan pemecahan s_k. Karena itu, memaksimumkan penurunan dalam *impurity* dengan pemecahan pada t akan ekuivalen dengan :

$$\Delta I(s_k, t) = I(t) - I(t_{\mathcal{L}}) - I(t_R) \tag{3.10}$$

Dengan proporsi $p_{\mathcal{L}}$ dan p_R yang masuk ke dalam $t_{\mathcal{L}}$ dan t_R dimana $p_R = \frac{p(t_R)}{p(t)}$, $p_{\mathcal{L}} = \frac{p(t_{\mathcal{L}})}{p(t)}$ dan $p_{\mathcal{L}} + p_R = 1$ maka Persamaan (3.10) dapat ditulis sebagai berikut

:

$$\begin{aligned}\Delta I(s_k, t) &= [i(t) - p_{\mathcal{L}}\ i(t_{\mathcal{L}}) - p_R\ i(t_R)]\, p(t) \\ &= \Delta i(s_k, t) p(t)\end{aligned}$$

Karena $\triangle I(s_k, t)$ berbeda dari $\Delta i(s_k, t)$ pada faktor $p(t)$, maka pemecahan s^* yang sama akan memaksimumkan keduanya. Kondisi berakhirnya suatu pemecahan atau penentuan suatu simpul menjadi simpul terminal adalah sebagai berikut :

Definisi 3.3.5. Taksiran resubsitusi $r(t)$ dari probabilitas salah klasifikasi yang diberikan oleh data apabila jatuh pada simpul t adalah $r(t) = 1 - \max_j P(j|t)$

Definisi 3.3.6. Probabilitas salah dalam klasifikasi yang diberikan data pada simpul t terhadap keseluruhan adalah $R(t) = r(t)\ p(t)$, maka taksiran resubtitusi untuk tingkat kesalahan klasifikasi secara keseluruhan pada simpul t adalah

$$R(t) = \sum_{t \epsilon T} R(t)$$

Kondisi ini merupakan kriteria untuk mengakhiri pemecahan ketika pemecahan selanjutnya tidak menurunkan semua tingkat kesalahan klasifikasi secara nyata.

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). Maka selanjutnya t tidak dipilah lagi tetapi dijadikan simpul terminal dan hentikan pembuatan pohon.

3.3.4 Menentukan Pohon Optimum

Ukuran pohon yang besar akan menyebabkan nilai kompleksitas yang tinggi karena struktur data yang digambarkan cenderung kompleks, sehingga perlu dipilih pohon optimum yang berukuran sederhana tetapi memiliki kesalahan pengklasifikasian yang cukup kecil. 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*).

Definisi 3.3.7. Misalkan sembarang pohon T yang merupakan sub pohon dari pohon terbesar $T_{\max}$ ($T < T_{\max}$), sehingga ukuran biaya kompleksitas didefinisikan sebagai berikut

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

dimana $R_\alpha(T)$ adalah tingkat kesalahan klasifikasi pada pohon bagian T, $R(T)$ adalah proporsi kesalahan pada sub pohon, $|\widetilde{T}|$ adalah ukuran banyaknya simpul terminal pohon T, dan α adalah parameter biaya kompleksitas.

Ukuran kompleksitas menentukan pohon bagian $T(\alpha)$ yang meminimumkan $R_\alpha(T)$ pada seluruh pohon bagian untuk setiap nilai α, sehingga dicari pohon bagian $T(\alpha) < T_{\max}$ yang meminimumkan $R_\alpha(T)$ yaitu

$$R_\alpha(T(\alpha)) = \min_{T<T_{\max}} R_\alpha(T)$$

Pemangkasan pohon klasifikasi dimulai dengan mengambil t_R yang merupakan simpul anak kanan dan t_L yang merupakan simpul anak kiri dari simpul t. Jika $R(t) = R(t_R) + R(t_L)$, maka simpul anak t_R dan t_L dipangkas. Proses

tersebut diulang sampai tidak ada lagi pemangkasan yang mungkin.

3.4 Data

Data yang digunakan dalam penelitan ini merupakan data sekunder yang diperoleh dari RS. Dharmais yaitu terdaftar sebagai pasien yang mengidap penyakit kanker payudara pada tahun 2015. Jumlah data set yang diambil adalah sebanyak 189 pasien. Dalam penelitian ini variabel yang diteliti terdiri dari variabel respon dan variabel bebas.

Untuk variabel respon yang digunakan adalah tingkat keganasan diagnosis penyakit kanker payudara, sedangkan untuk variabel bebas yang digunakan adalah faktor-faktor yang mempengaruhi timbulnya penyakit kanker payudara dengan 7 variabel, yaitu usia, usia *menarche* (pertama menstruasi), usia *menopause*, obesitas, riwayat keluarga penderita kanker payudara (genetik), tidak menyusui anak, dan penggunaan KB (Kelsey dkk , 1991). Ketujuh variabel tersebut sudah dikodekan sesuai dengan kategorinya masing-masing.

Berikut ini merupakan definisi operasional variabel tersebut :

1. **Tingkat keganasan kanker payudara (Y)**

   Diagnosis kanker pada pasien kanker payudara ditentukan berdasarkan tingkat keganasannya yaitu jinak dan ganas. Dengan melihat gejala-gejala yang timbul, yaitu pada kanker payudara yang jinak umumnya ditemukan benjolan berbentuk kelereng yang berukuran kurang dari 2 cm, sedangkan pada kanker payudara yang ganas bentuk benjolan membesar pada payudara sehingga menyebabkan nyeri, terdapat pula kerutan dan mengeluarkan cairan, serta mengalami perubahan pada kulit dan ukuran payudara. Skala data

untuk variabel tingkat keganasan kanker payudara termasuk dalam skala nominal yang dibedakan menjadi dua kategori, yaitu

Tabel 3.2: Pengkategorian dan Pemberian Kode Berdasarkan Tingkat Keganasan Kanker

Tingkat keganasan kanker payudara	Kode
Pasien penyakit kanker payudara terdiagnosis kanker jinak	0
Pasien penyakit kanker payudara terdiagnosis kanker ganas	1

2. **Usia** (X_1)

   Terdapat banyak kasus wanita kanker yang ditemukan yaitu pada usia antara $40 - 60$ tahun memiliki risiko mengidap kanker payudara. Skala data untuk variabel usia yang mengidap kanker payudara dikategorikan dalam skala ordinal. Berikut merupakan pengkategorian pasien kanker payudara berdasarkan kelompok usia, yaitu

   Tabel 3.3: Pengkategorian dan Pemberian Kode Berdasarkan Usia

Kelompok Usia	Kode
Usia < 40 tahun	0
Usia ≥ 40 tahun	1

3. **Usia *Menarche*** (X_2)

   Pada diagnosis tingkat keganasan kanker dilihat apakah pasien mengidap kanker payudara memiliki riwayat *menarche* lambat atau cepat. Adapun skala data variabel usia *menarche* dikategorikan dalam skala ordinal. Berikut merupakan pengkategorian pasien kanker payudara berdasarkan kelompok usia *menarche*, yaitu

Tabel 3.4: Pengkategorian dan Pemberian Kode Berdasarkan Usia *Menarche*

Kelompok Usia *Menarche*	Kode
Usia *menarche* lambat ≥ 12 tahun	0
Usia *menarche* cepat < 12 tahun	1

4. **Usia *Menopause*** (X_3)

   Diagnosis tingkat keganasan payudara juga dapat dilihat dari usia *menopause* apakah terlambat atau melebihi usia *menopause* semestinya (50 tahun), untuk skala data variabel usia *menopause* dikategorikan dalam skala ordinal. Berikut merupakan pengkategorian pasien kanker payudara berdasarkan kelompok usia *menopause*, yaitu

Tabel 3.5: Pengkategorian dan Pemberian Kode Berdasarkan Usia *Menopause*

Kelompok Usia *Menopause*	Kode
Usia *menopause* cepat ≤ 50 tahun	0
Usia *menopause* lambat > 50 tahun	1

5. **Obesitas** (X_4)

   Apabila pola hidup yang tidak baik salah satunya akan berakibat mengalami kenaikan berat badan yang berlebih (obesitas), hal ini cenderung juga berisiko terkena kanker payudara. Skala data untuk variabel obesitas berisiko mengidap kanker payudara dikategorikan dalam skala nominal. Berikut merupakan pengkategorian pasien kanker payudara berdasarkan obesitas, yaitu

Tabel 3.6: Pengkategorian dan Pemberian Kode Berdasarkan Obesitas

Status Obesitas	Kode
Pasien kanker payudara yang berstatus tidak mengalami obesitas	0
Pasien kanker payudara yang berstatus mengalami obesitas	1

6. **Riwayat Keluarga Penderita Kanker Payudara (Genetik)** (X_5)

   Risiko terkena kanker payudara meningkat apabila mempunyai ibu atau saudara perempuan yang mengidap kanker payudara, dan semua saudara dari penderita kanker payudara pun memiliki peningkatan berisiko akan mengalami kanker payudara. Skala data untuk variabel riwayat keluarga penderita kanker payudara dikategorikan dalam skala nominal. Berikut merupakan pengkategorian pasien kanker payudara berdasarkan genetik, yaitu

Tabel 3.7: Pengkategorian dan Pemberian Kode Berdasarkan Genetik

Riwayat Keluarga Penderita Kanker	Kode
Pasien kanker payudara yang tidak memiliki riwayat keluaga penderita kanker (genetik)	0
Pasien kanker payudara yang memiliki riwayat keluaga penderita kanker (genetik)	1

7. **Tidak Menyusui Anak** (X_6)

   Apabila wanita yang tidak menyusui bayinya, mempunyai risiko tinggi terkena kanker payudara dibandingkan dengan wanita yang menyusui bayinya. Hal ini merupakan faktor penting mengakibatkan timbulnya kanker payudara pada wanita. Skala data untuk variabel riwayat keluarga penderita kanker payudara dikategorikan dalam skala nominal. Berikut merupakan pengkategorian pasien kanker payudara berdasarkan tidak menyusui anak, yaitu

Tabel 3.8: Pengkategorian dan Pemberian Kode Berdasarkan Tidak Menyusui Anak

Riwayat Tidak Menyusui Anak	Kode
Pasien kanker payudara yang memiliki riwayat menyusui anak	0
Pasien kanker payudara yang memiliki riwayat tidak menyusui anak	1

8. **Penggunaan KB** (X_7)

   Penggunaan KB tidak dianjurkan dalam jangka panjang, hal ini menunjukkan jika pemakaian yang berlebih dapat meningkatkan risiko timbulnya kanker payudara. Skala data untuk variabel penggunaan KB dikategorikan dalam skala nominal. Berikut merupakan pengkategorian pasien kanker payudara berdasarkan penggunaan KB, yaitu

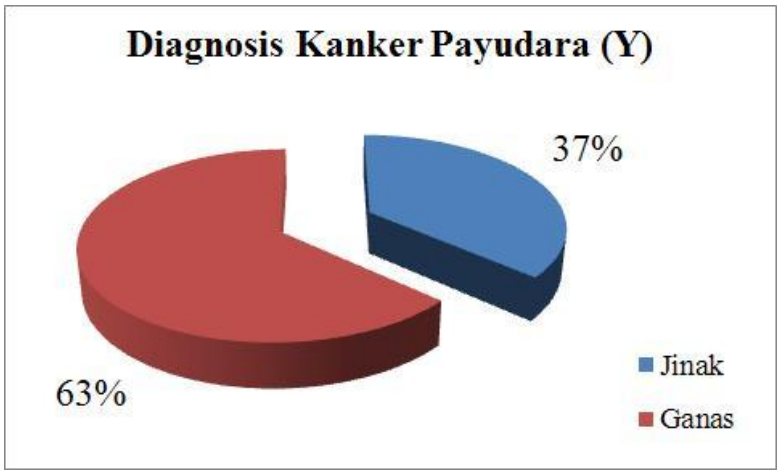
Tabel 3.9: Pengkategorian dan Pemberian Kode Berdasarkan Penggunaan KB

Penggunaan KB	Kode
Pasien kanker payudara melakukan penggunaan KB tidak jangka panjang	0
Pasien kanker payudara melakukan penggunaan KB jangka panjang	1

3.5 Deskripsi Statistik

Pada bagian ini akan disajikan ringkasan data yang berhubungan dengan diagnosis kanker payudara dan faktor-faktor yang mempengaruhinya, yaitu usia, usia *menarche* (pertama menstruasi), usia *menopause*, obesitas, riwayat keluarga penderita kanker payudara, tidak menyusui, dan pengguna KB.

Data banyaknya pasien terkumpul akan digunakan dalam penelitian ini adalah 189 pasien. Jumlah pasien kanker payudara dalam kategori terdiagnosis kanker jinak sebanyak 69 pasien dan untuk kategori terdiagnosis kanker ganas sebanyak 120 pasien. Hal ini menunjukkan bahwa sebagian besar pasien termasuk dalam kategori terdiagnosis kanker ganas. Berikut gambar yang menunjukkan persentase diagnosis kanker payudara sebagai berikut

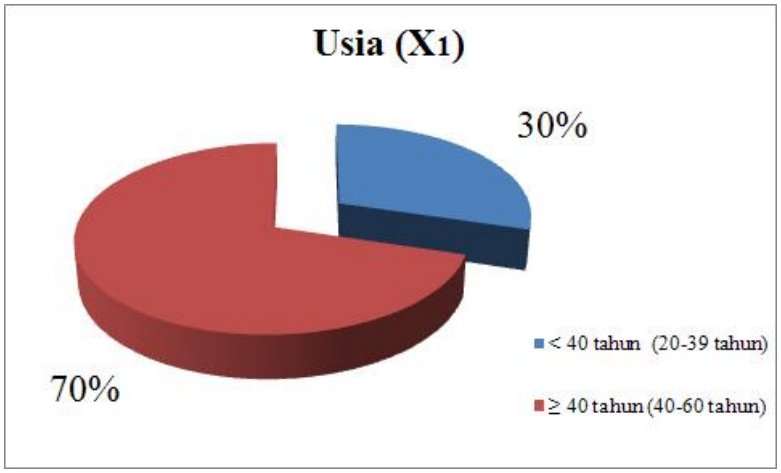


Gambar 3.5: Persentase Diagnosis Kanker Payudara

Dari gambar di atas terlihat bahwa pasien payudara terdiagnosis kanker ganas memiliki persentase sebesar 63% dan pasien payudara terdiagnosis kanker jinak memiliki persentase lebih rendah yaitu sebesar 37%. Hal ini menunjukkan bahwa jumlah pasien terbesar adalah dimiliki oleh pasien yang telah mengidap penyakit

kanker ganas, sehingga persentase seseorang akan terdiagnosis kanker ganas pada kanker payudara cukuplah tinggi.

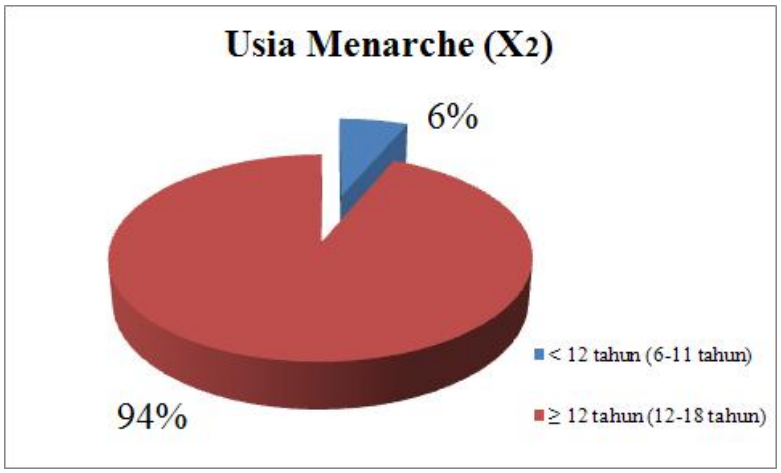
Berdasarkan dari faktor yang mempengaruhi hasil diagnosis kanker payudara adalah usia yang akan dikategorikan dalam dua kategori yaitu dengan usia selang $20 - 39$ tahun termasuk dalam kategori usia < 40 tahun yang berjumlah 56 pasien dan untuk usia selang $40 - 60$ tahun termasuk dalam kategori usia ≥ 40 tahun berjumlah 133 pasien. Akan diberikan persentase diagnosis kanker payudara dalam bentuk gambar sebagai berikut



Gambar 3.6: Persentase Usia

Berdasarkan Gambar (3.6) terlihat bahwa faktor yang mempengaruhi diagnosis kanker payudara berdasarkan usia pasien yang termasuk < 40 tahun memiliki persentase sebesar 30% dan usia pasien yang termasuk ≥ 40 tahun memiliki persentase terbesar yaitu sebesar 70%. Terlihat jelas dari hasil persentase diatas bahwa mayoritas sebagian besar pasien mengalami timbulnya kanker payudara pada usia diatas 40 tahun. Oleh karena itu kontribusi agar dapat memperkecil risiko timbulnya penyakit kanker payudara dengan melakukan pemeriksaan secara klinis sedini mungkin.

Faktor yang mempengaruhi diagnosis kanker payudara selanjutnya adalah usia *menarche* yang dikategorikan dalam dua kategori yaitu apabila usia *menarche* masuk dalam kategori usia < 12 tahun yang memiliki usia selang $6 - 11$ tahun berjumlah 12 dan untuk usia selang $12 - 18$ tahun masuk dalam kategori usia ≥ 12 tahun berjumlah 177 pasien. Akan diberikan persentase usia *menarche* dalam bentuk gambar sebagai berikut

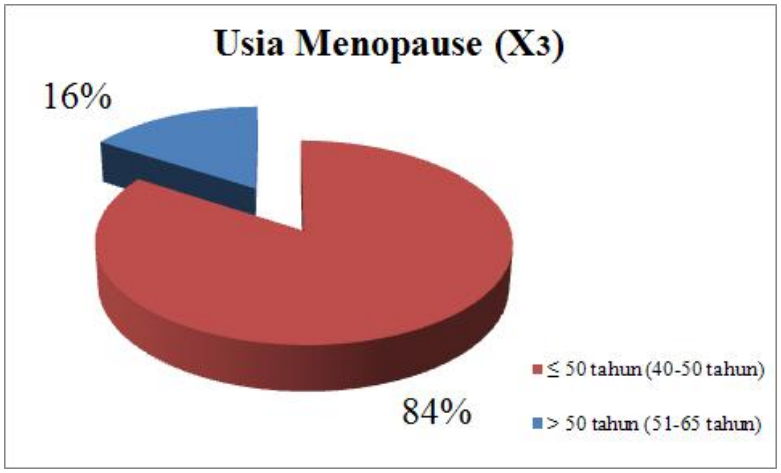


Gambar 3.7: Persentase Usia *Menarche*

Gambar di atas menunjukkan bahwa faktor yang mempengaruhi diagnosis kanker payudara berdasarkan usia *menarche* yang termasuk < 12 tahun memiliki persentase sangat rendah yaitu sebesar 6% dan begitupun sebaliknya untuk usia *menarche* yang termasuk ≥ 12 tahun memiliki persentase sangat besar hingga 94%. Pada persentase ini jumlah pasien mengalami usia *menarche* pada < 12 tahun sangatlah kecil.

Selanjutnya berdasarkan dari faktor yang mempengaruhi hasil diagnosis kanker payudara adalah usia *menopause* yang dikategorikan dalam dua kategori yaitu dengan usia *menopause* selang $40 - 50$ tahun termasuk dalam kategori usia *menopause* ≤ 50 tahun berjumlah 159 pasien dan untuk usia selang $51 - 65$ tahun

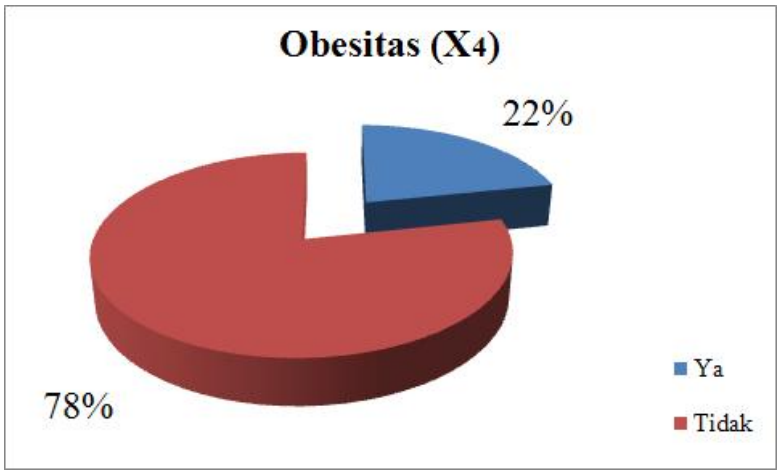
termasuk dalam kategori usia > 50 tahun berjumlah 30 pasien. Akan diberikan persentase usia *menopause* dalam bentuk gambar sebagai berikut



Gambar 3.8: Persentase Usia *Menopause*

Pada Gambar (3.8) memperlihatkan bahwa faktor yang mempengaruhi diagnosis kanker payudara berdasarkan usia *menopause* yang termasuk ≤ 50 tahun memiliki persentase sebesar 16% dan usia pasien yang termasuk > 50 tahun memiliki persentase terbesar yaitu sebesar 84%. Ini menunjukkan sebagian besar jumlah pasien memiliki riwayat telah melewati masa menstruasi pada usia > 50 tahun.

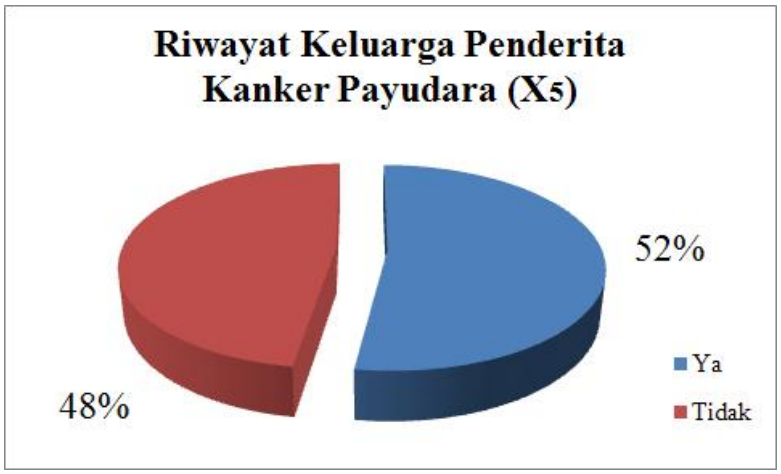
Diagnosis kanker payudara memiliki faktor yang mempengaruhi berikutnya adalah obesitas yang dikategorikan dalam dua kategori yaitu apabila pasien mengalami berat badan obesitas maka termasuk kategori "ya" dengan jumlah pasien sebanyak 41. Dan apabila pasien tidak mengalami berat badan obesitas maka masuk dalam kategori "tidak" dengan jumlah sebanyak 148 pasien. Akan ditunjukkan persentase obesitas dalam bentuk gambar sebagai berikut



Gambar 3.9: Persentase Obesitas

Gambar diatas menunjukkan bahwa nilai persentase bahwa pasien yang mengalami berat badan obesitas sebesar 22% dan persentase untuk pasien yang tidak mengalami berat badan obesitas memiliki nilai sebesar 78%. Pada penelitian ini pada nilai kasus pasien yang memiliki riwayat obesitas cukuplah kecil dibandingkan sebagian besar pasien tidak mengalami riwayat obesitas.

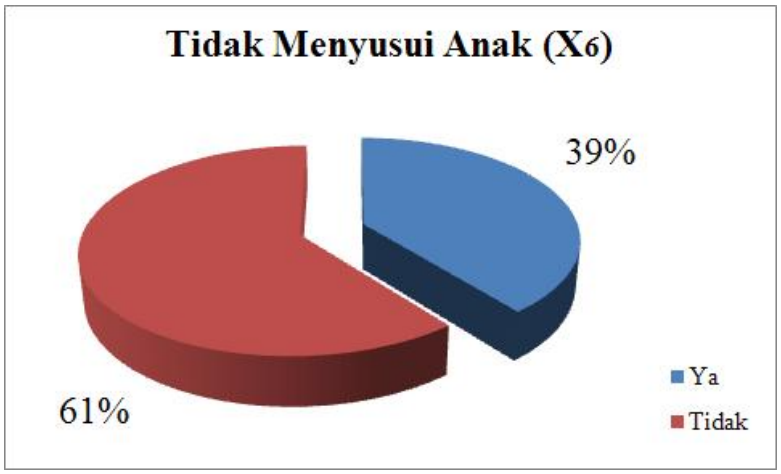
Salah satu faktor yang mempengaruhi hasil diagnosa kanker payudara berikutnya adalah memiliki riwayat keluarga penderita kanker payudara (genetik) yang dikategorikan menjadi dua kategori yaitu jika pasien memiliki riwayat keluarga penderita kanker payudara maka masuk dalam kategori "ya" dengan jumlah sebanyak 99 pasien dan masuk dalam kategori "tidak" dengan jumlah pasien sebanyak 90 merupakan pasien tidak memiliki riwayat keluarga penderita kanker payudara. Dalam faktor ini menunjukkan jumlah yang hampir seimbang antara keduanya. Akan diperlihatkan dalam bentuk gambar untuk persentase riwayat keluarga penderita kanker payudara sebagai berikut



Gambar 3.10: Persentase Riwayat Penderita Kanker Payudara

Pada gambar (3.10) dijelaskan bahwa persentase dimiliki oleh pasien yang terdapat riwayat keluarga penderita kanker payudara yaitu sebesar 52% dan nilai persentase sebesar 48% dimiliki oleh pasien yang tidak terdapat riwayat keluarga penderita kanker payudara. Perbandingan nilai persentase dari keduanya sangatlah tipis, sehingga hampir dominan memiliki riwayat keluarga penderita kanker dan tidak memiliki riwayat penderita kanker.

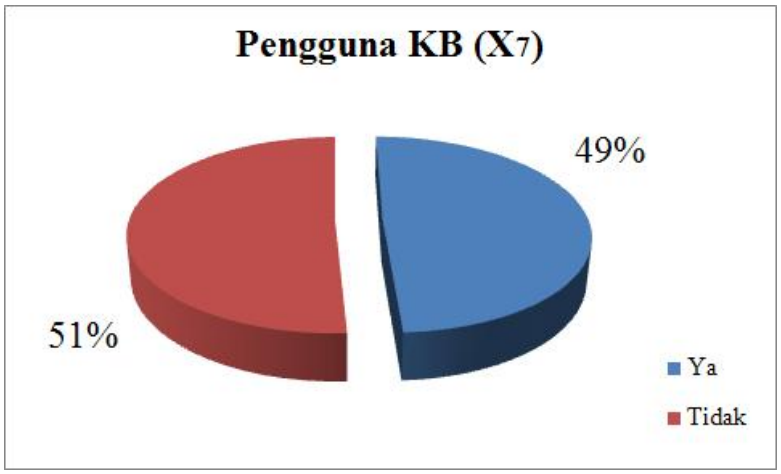
Tidak menyusui anak adalah salah satu faktor yang mempengaruhi diagnosis kanker payudara yang dikategorikan dalam dua kategori yaitu apabila pasien tidak pernah menyusui anaknya atau tidak pernah memiliki anak maka dikategorikan "ya" dengan jumlah pasien sebanyak 74 dan apabila faktor tidak menyusui anak dikategorikan "tidak" maka pasti pasien menyusui anaknya pada fase pasca melahirkan dengan jumlah sebanyak 115. Akan ditunjukkan penyajian dalam gambar terhadap nilai persentase tidak menyusui anak sebagai berikut



Gambar 3.11: Persentase Tidak Menyusui Anak

Terlihat pada gambar diatas bahwa persentase yang diperoleh faktor pasien yang tidak menyusui anaknya pada fase setelah melahirkan sebesar 39% dan faktor pasien yang menyusui anaknya ditunjukkan persentase sebesar 61%. Dari seluruh pasien yang terdiagnosis kanker ada diantara pasien yang tidak mengalami fase menyusui anaknya setelah melahirkan, ini menunjukkan bahwa adanya faktor-faktor lain yang menyebabkan seorang ibu tidak dapat menyusui anaknya.

Dan faktor terakhir yang dapat mempengaruhi diagnosis kanker payudara adalah menggunakan program keluarga berencana (KB) yang akan dikategorikan dalam dua kategori yang sama dengan sebelumnya yaitu jika pasien adalah pengguna KB maka masuk dalam kategori ”ya” dan jika pasien adalah tidak pengguna KB maka masuk dalam kategori ”tidak”, dengan masing-masing kategori memiliki jumlah pasien sebanyak 93 dan 96 pasien. Hal ini menunjukkan bahwa pasien yang menggunakan program KB hampir sama besarnya dengan yang tidak menggunakan program KB. Di bawah ini akan disajikan persentase pengguna KB dalam bentuk gambar sebagai berikut



Gambar 3.12: Persentase Pengguna KB

Sesuai dalam gambar ditunjukkan bahwa pasien pengguna KB dan pasien tidak pengguna KB memiliki nilai persentase yang tipis yaitu masing-masing sebesar 49% dan 51%.

Ringkasan data persentase diagnosis penyakit kanker payudara dengan faktor-faktor yang mempengaruhinya dapat dilihat dalam tabel seagai berikut

Tabel 3.10: Tabel Nilai Rata-Rata, Standar Deviasi, dan Varians Variabel Respon dan Variabel Bebas

Variabel	Rata-Rata	Standar Deviasi	Varians
Diagnosis Kanker (Y)	0.64	0.481	0.232
Usia (X_1)	40.70	8.632	74.510
Usia *Menarche* (X_2)	0.06	0.244	0.060
Usia *Menopause*(X_3)	0.16	0.366	0.134
Obesitas (X_4)	0.22	0.413	0.171
RPKP (X_5)	0.51	0.501	0.251
TMA (X_6)	0.39	0.489	0.240
Penggunaan KB (X_7)	0.49	0.501	0.251

Rata-rata persentase pasien terdiagnosis penyakit kanker payudara adalah sebesar 0.64% dengan diagnosis kanker payudara terbanyak adalah kanker ganas.

Persentase usia pasien (X_1) yang mengidap kanker payudara tertinggi pada usia 59 tahun sedangkan untuk usia terendah adalah 24 tahun. Rata-rata persentase usia *menarche* (X_2) dan usia *menopause* (X_3) adalah sebesar 0.06% dan 0.16%. Pada pasien yang memiliki obesitas (X_4) memiliki rata-rata persentase sebesar 0.22%. Persentase pasien yang memiliki riwayat keluarga penderita kanker payudara (X_5) tertinggi adalah pada pasien yang terdiagnosis kanker ganas. Rata-rata persentase pasien yang tidak menyusui anaknya (X_6) adalah sebesar 0.39%, dan persentase pasien yang menggunakan KB berkepanjangan memiliki nilai rata-rata sebesar 0.49%.

3.6 Analisis Data Hasil Diagnosis Kanker Payudara dengan Regresi Logistik Biner

Untuk mengetahui hasil diagnosis kanker dengan faktor-faktor penyebab timbulnya kanker payudara dapat menggunakan analisis regresi logistik. Selain itu dapat diketahui besarnya pengaruh setiap faktor dalam menentukan peluang seseorang akan terdiagnosis kanker payudara jinak atau ganas.

3.6.1 Model Regresi Logistik Biner

Model regresi logistik mengansumsikan tidak diperbolehkan adanya multikolinieritas diantara variabel bebas, karena jika terdapat multikolinieritas *standard error* dari koefisien regresinya akan membesar sehingga dimungkinkan dari masing-masing variabel bebas akan tidak signifikan. Untuk mendeteksi adanya multikolinieritas dapat dilihat dari nilai VIF (*Variance Inflation Factor*). Adanya multikolinieritas jika nilai VIF pada setiap variabel bebas memiliki nilai melebihi

5. Nilai VIF dari variabel faktor pengaruh diagnosis penyakit kanker payudara ditunjukkan pada Tabel 3.11 sebagai berikut

Tabel 3.11: Tabel Nilai VIF untuk Setiap Variabel Bebas

Variabel Bebas	Nilai VIF
X_1	1.658
X_2	1.019
X_3	1.721
X_4	1.056
X_5	1.073
X_6	1.201
X_7	1.182

Tabel di atas menunjukkan bahwa semua variabel bebas memiliki nilai VIF kurang dari 5. Hal ini menunjukan bahwa tidak terjadi multikolinieritas pada setiap variabel yang akan diamati.

Setelah melihat multikolinieritas dari setiap variabel bebas maka selanjutnya akan menduga koefisien pada model. Akan ditunjukkan hasil dugaan koefisien model regresi logistik biner pada tabel sebagai berikut

Tabel 3.12: Tabel Dugaan Koefisien Model Regresi Logistik Biner

Variabel	Koefisien
Constant	-16.620
X_1	0.403
$X_{2(1)}$	1.216
$X_{3(1)}$	-2.420
$X_{4(1)}$	0.244
$X_{5(1)}$	1.704
$X_{6(1)}$	0.813
$X_{7(1)}$	0.165

Berdasarkan dari Tabel 3.12 diperoleh model regresi logistik biner (*full model*) adalah

$$Logit(\pi_i) = -16.620 + 0.403X_1 + 1.216X_{2(1)} - 2.420X_{3(1)}$$
$$+ 0.244X_{4(1)} + 1.704X_{5(1)} + 0.813X_{6(1)} + 0.165X_{7(1)}$$

3.6.2 Pengujian Parameter Secara Serentak

Pendugaan model regresi logistik biner dengan menggunakan tujuh variabel bebas dilakukan uji hipotesis yaitu

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

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

Diketahui nilai statistik-G -2*Log likelihood* (tanpa variabel bebas yang masuk dalam model) sebesar 246.865 dan hasil nilai statistik-G -2*Log likelihood* (variabel bebas masuk dalam model) sebesar 101.081. Berdasarkan hasil data tersebut maka diperoleh nilai statistik-G sebesar 145.865 dengan p-*value* adalah 0.000. Untuk lebih jelasnya dapat dilihat pada Lampiran 3

Kriteria pengambilan keputusan dilakukan dengan membandingkan nilai statistik-G dengan nilai *chi-square* dari tabel, $\chi^2_{(7,0.05)} = 14.067$. Karena nilai statistik-G $> \chi^2_{(7,0.05)}$, sehingga keputusannya adalah Tolak H_0 artinya sedikitnya ada satu β_i yang tidak sama dengan nol pada taraf nyata 5%. Dapat diambil kesimpulan bahwa adanya pengaruh signifikan secara simultan dari faktor-faktor risiko timbulnya kanker terhadap hasil diagnosis kanker payudara.

3.6.3 Pengujian Parameter Secara Parsial

Selanjutnya akan diuji parameter secara parsial yaitu melihat pengaruh setiap parameter pada model secara individual. Hasil pengujian parameter diper-

oleh menggunakan statistik uji-Wald, akan ditunjukkan dalam tabel dibawah ini dengan hipotesis sebagai berikut

$H_0 : \beta_i = 0$

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

Tabel 3.13: Tabel Pengujian Secara Parsial Pemodelan Awal Dengan Uji-Wald

Variabel Bebas	Wald	p-value
Constant	**36.217**	**0.000***
Usia	**40.512**	**0.000***
Usia *Menarche* (1)	0.835	0.361
Usia *Menopause* (1)	3.325	0.068
Obesitas (1)	0.144	0.704
RKPK (1)	**7.898**	**0.005***
TMA (1)	1.994	0.158
KB (1)	0.092	0.761

Berdasarkan tabel diatas bahwa uji Wald menguji koefisien regresi logistik dengan menghasilkan dua variabel yang nyata pada $\alpha = 5\% = 0.05$, maka dapat diinterpretasikan yaitu :

1. Untuk variabel usia diperoleh nilai Wald sebesar 40.512 dengan p-*value* = $0.000 < \alpha$, maka koefisien regresi pada variabel usia adalah signifikan

2. Dan nilai Wald sebesar 7.898 diberikan pada variabel riwayat keluarga penderita kanker (RKPK) dengan p-*value* = $0.005 < \alpha$ sehingga koefisien regresi pada variabel RKPK adalah signifikan

Dari pernyataan diatas pengujian hipotesis yang dihasilkan adalah menolak H_0 atau berarti variabel usia dan riwayat keluarga penderita kanker memberikan pengaruh parsial yang signifikan terhadap hasil diagnosis kanker payudara.

Pada tahap selanjutnya adalah pemilihan model terbaik dengan pengambilan keputusan pada taraf nyata 5%, sehingga terdapat dua variabel yang berpen-

garuh secara signifikan. Variabel-variabel tersebut adalah usia (X_1) dan riwayat keluarga penderita kanker ($X_{5(1)}$), hasil selengkapnya dapat dilihat pada Lampiran 3. Model Regresi Logisik Biner terbaik yang terbentuk adalah

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

3.6.4 Uji Kelayakan Model

Uji *Hosmer dan Lemeshow* adalah uji *Goodness of fit test*, yaitu uji untuk menentukan apakah model yang terbentuk layak atau tidak. Pengujian ini digunakan untuk menguji hipotesis sebagai berikut

H_0 : Tidak terdapat perbedaan signifikan antara klasifikasi yang diprediksi dengan klasifikasi yang diamati

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

Hasil pengujian dapat dilihat pada Lampiran 3, menunjukkan bahwa nilai *chi-square* mendekati 9.672 dengan nilai signifikasi sebesar 0.289. Karena nilai p-*value*$>$ 0.05 maka dapat disimpulkan bahwa uji hipotesis H_0 diterima. Sehingga tidak terdapat perbedaan signifikan antara klasifikasi yang diprediksi dengan klasifikasi yang diamati, maka model regresi ini layak digunakan analisis selanjutnya untuk melihat faktor risiko timbulnya kanker terhadap hasil diagnosis kanker payudara.

3.6.5 Interpretasi Rasio *Odds*

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

bebas dapat dilihat pada tabel sebagai berikut

Tabel 3.14: Tabel Rasio *Odds* Model Regresi Logistik Biner

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

Kesimpulan dari tabel dapat diinterpretasikan bahwa nilai rasio *odds* pada variabel usia adalah sebesar 1.497, karena rasio *odds* variabel X_1 tersebut bernilai > 1 maka variabel tersebut memberikan pengaruh positif terhadap faktor risiko menentukan hasil diagnosis kanker payudara. Dan untuk variabel riwayat keluarga penderita kanker memiliki nilai rasio *odds* sebesar 5.496, ini menunjukkan bahwa pada pasien yang memiliki riwayat keluarga penderita kanker berpeluang 243 kali lebih besar risiko terdiagnosis kanker payudara dibandingkan dengan pasien yang tidak memiliki riwayat keluarga penderita kanker payudara.

3.6.6 Ketepatan Klasifikasi

Selanjutnya akan dilakukan perhitungan klasifikasi untuk melihat peluang ketepatan klasifikasi model. Berikut hasil ketepatan klasifikasi data diagnosis kanker payudara

Tabel 3.15: Tabel Ketepatan Pengklasifikasian Diagnosis Kanker Payudara

Observasi	Hasil prediksi		Presentase
	jinak	ganas	Total Prediksi
Diagnosis kanker payudara jinak	55	13	80.9%
Diagnosis kanker payudara ganas	5	116	95.9%
Presentase keseluruhan			**90.5%**

Berdasarkan tabel dapat dilihat bahwa dari 68 pasien yang terdiagnosis kanker payudara jinak sebanyak 55 pasien diklasifikasikan dengan benar, sedangkan dari 121 pasien yang terdiagnosis kanker payudara ganas sebanyak 116 pasien yang diklasifikasikan dengan benar. Persentase masing-masing hasil klasifikasi sebesar 80.9% dan 95.9% dengan hasil missklasifikasi dari 189 pasien adalah 9.5%. Dengan perhitungan nilai ketepatan klasifikasi sebagai berikut

$$\frac{55+116}{55+13+5+116} = \frac{171}{189} = 0.905 = 90.5\%$$

Secara keseluruhan, model regresi logistik yang telah terbentuk mampu mengklasifikasikan dengan baik dengan memperoleh nilai kesalahan klasifikasi cukup kecil sehingga model ini dapat dikatakan sudah baik diterapkan untuk menentukan hasil diagnosis kanker payudara.

3.7 Analisis Data Hasil Diagnosis Kanker Payudara dengan Metode (CART)

Pembagian data *Learning sample* dilakukan secara acak menjadi 2 bagian (*training testing*) dengan proporsi 70% : 30%. Dengan demikian data *training* berjumlah 132 buah, sedangkan data *testing* berjumlah 57 buah.

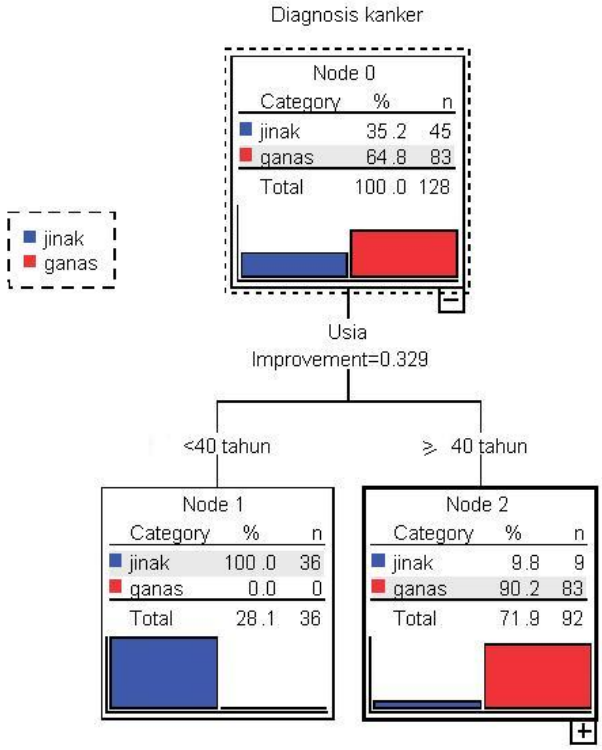
3.7.1 Proses Pemecahan Simpul Nonterminal

Berdasarkan nilai fungsi keragaman $\Delta I(s_k, t)$, kriteria variabel pemecahan s_k yang terpilih adalah variabel usia sebagai pemilah terbaik karena memiliki nilai keragaman yang paling maksimal dimana $\Delta I(s_k, t)$ sebesar 0.329. Akan ditunjukkan pada tabel dibawah ini

Tabel 3.16: Tabel Hasil Kriteria Pemilah Terbaik

No	Variabel	Nilai keragaman (*Improvement*)	*Split*	Simpul Kiri	Simpul Kanan
1	Usia	0.329	$<$ 40 tahun	36	
			$\geq$ 40 tahun		92
2	Riwayat Keluarga Penderita Kanker	0.011	Tidak	41	
			Ya		51
3	Tidak Menyusui Anak	0.001	Tidak	37	
			Ya		14

Setelah terpilih pemilihan terbaik, maka pada simpul utama yaitu *node* 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. Untuk lebih jelasnya proses pemilahan dapat dilihat pada Gambar 3.12 sebagai berikut :



Gambar 3.13: Variabel Pemilah Simpul Terbaik

Proses serupa terus berjalan pada simpul-simpul lainnya. Proses pemilahan yang berulang-ulang akan berhenti apabila sudah tidak dimungkinkan lagi dilakukan proses pemilahan karena pada pohon klasifikasi maksimum tersisa simpul akhir yang memiliki anggota kelas yang sama (homogen).

3.7.2 Pelabelan Kelas

Pada bagian ini dibahas mengenai pemberian label kelas pada simpul-simpul yang telah terbentuk. Prosedur pemberian label kelas berdasarkan Definisi

3.3.4 yaitu aturan penetapan kelas adalah jika $P(j|t) = \max_i p(i|t)$ maka $j^*(t) = j$, dimana $j^*(t)$ adalah kelas yang diidentifikasi pada simpul t. Sebagai contoh pada Gambar 3.12 didefinisikan sebagai berikut

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

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

Sehingga simpul diberi label kelas "Diagnosis Kanker Ganas", dikarenakan peluang kelas ganas lebih besar dari peluang 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.

3.7.3 Proses Menentukan Simpul Terminal

Pada proses pemangkasan pohon dilakukan untuk mengurangi kompleksitas pohon agar menjadi sederhana. Dengan pemangkasan jumlah simpul akan berkurang sehingga jumlah simpul terminal juga akan berkurang. Hasil pohon klasifikasi dapat diperoleh melalui proses pemangkasan pohon, sehingga didapat pohon yang memiliki 7 simpul diantaranya adalah 3 simpul nonterminal dan 4 simpul terminal.

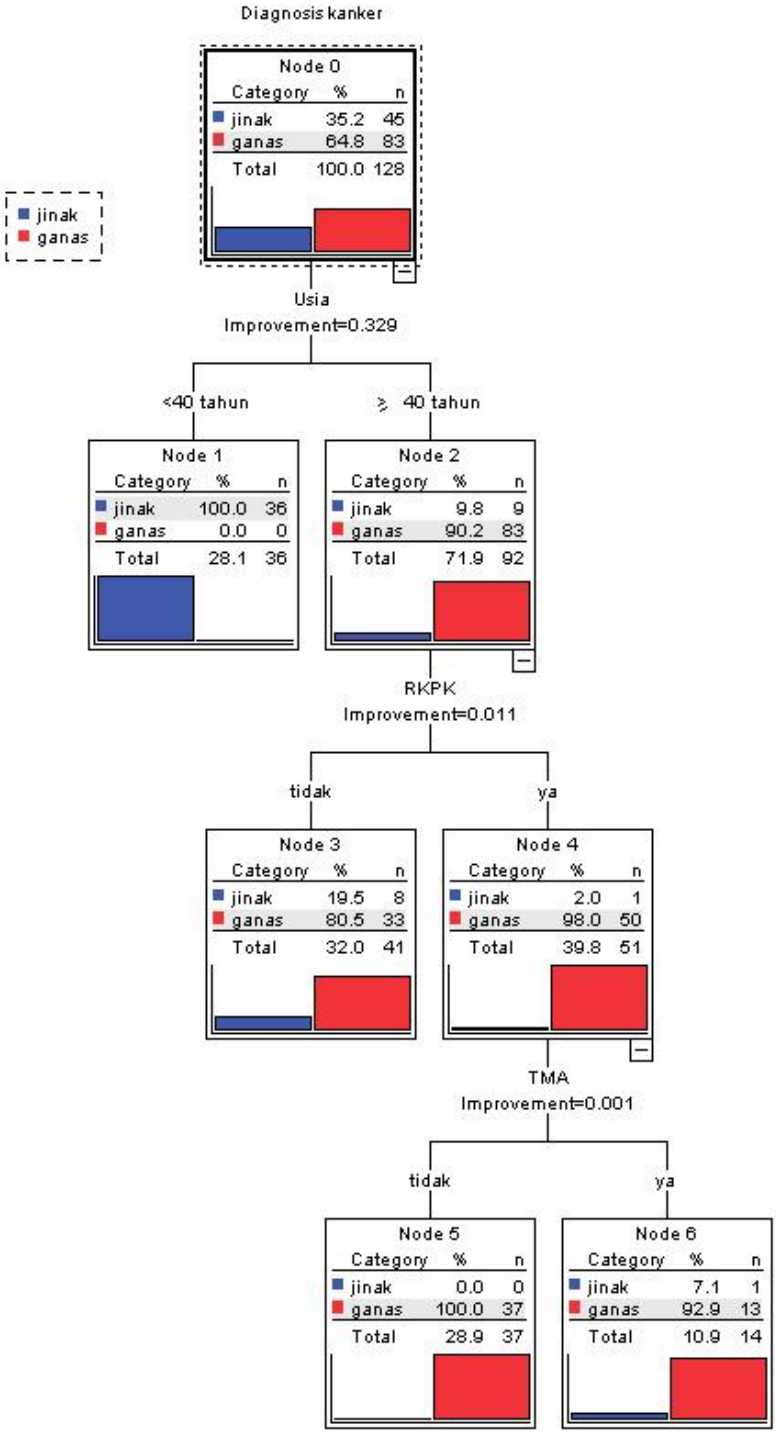
Variabel bebas yang masuk ke dalam pohon klasifikasi hanya ada tiga dari seluruh variabel yang dapat mempengaruhi hasil diagnosis kanker payudara yaitu usia, riwayat keluarga penderita kanker payudara, dan tidak menyusui. Variabel usia adalah variabel pertama sebagai penyekat. Hal ini menyatakan bahwa variabel tersebut merupakan variabel yang paling dominan dalam pembentukan

pohon klasifikasi.

Diagnosis kanker payudara berjumlah 128 pada simpul pertama (*node* 0) dipilah menjadi kelompok kiri dan kanan oleh variabel usia yaitu usia pasien yang memiliki < 40 tahun sebanyak 36 pasien mengelompok pada simpul 1 (kiri) sedangkan untuk usia pasien ≥ 40 tahun sebanyak 92 mengelompok pada simpul 2 (kanan). Pada simpul 1 merupakan simpul terminal. Penurunan nilai keragaman pada simpul pertama sebesar 0.329 ditunjukkan oleh *improvement*.

Terdapat 92 terdiagnosa kanker payudara pada simpul 2 yang akan dipilah lagi menjadi dua kelompok oleh variabel riwayat keluarga penderita kanker payudara. Diagnosis penyakit kanker payudara yang sebagian besar pasiennya adalah tidak memiliki riwayat keluarga penderita kanker payudara sebanyak 41 pasien mengelompok pada simpul 3 (kiri) sedangkan untuk diagnosis penyakit kanker payudara yang sebagian besar pasiennya memiliki riwayat penderita kanker payudara sebanyak 51 pasien mengelompok pada simpul 4 (kanan). Dan simpul terminal lainnya adalah simpul 3, dengan penurunan nilai keragaman pada simpul 2 adalah 0.011.

Selanjutnya pada simpul 4 sebanyak 51 pasien akan dipilah kembali menjadi dua kelompok oleh variabel tidak menyusui untuk masing-masing kelompok adalah pada simpul 5 (kiri) dan simpul 6 (kanan). Diagnosa kanker payudara tidak memiliki riwayat tidak menyusui bayinya dan memiliki riwayat tidak menyusui bayinya masing-masing sebanyak 37 pasien dan 14 pasien. Untuk penurunan nilai keragaman pada simpul 4 adalah sebesar 0.001 dan pada pemilah terakhir ini simpul terminal diperoleh adalah 5 dan 6. Dari pernyataan diatas dapat dilihat lebih jelas ditunjukkan pada gambar 3.13 dibawah ini sebagai berikut



Gambar 3.14: Pohon Klasifikasi Optimum

3.7.4 Interpretasi Pohon Klasifikasi Optimum

Hasil pohon klasifikasi setelah proses pemangkasan diperoleh variabel usia merupakan variabel prediktor yang paling berpengaruh, sehingga menjadi pemilah terbaik dari simpul nonterminal. Pohon klasifikasi optimum ini memiliki 4 kelas yang menentukan hasil diagnosis kanker payudara. Klasifikasi yang terbentuk adalah

1. 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%.

2. Pasien yang tidak memiliki riwayat tidak menyusui bayinya juga memiliki potensi tinggi sebesar 100% dapat terdiagnosis kanker payudara adalah ganas.

3. 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%.

4. 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

$$\begin{aligned} R(d) &= \frac{1}{N}\sum_{(x_n,j_n)\epsilon\mathcal{L}}^{N} X(d(\widetilde{x_n}) \neq j_n) \\ &= \frac{9+0}{128} = \frac{9}{128} = 0.070 \end{aligned}$$

Dengan nilai $R(d) = 0.070$, maka ketepatan klasifikasinya adalah $1 - 0.070 = 0.93$ atau 93%. Hasil dari klasifikasi optimum dengan menggunakan data *testing* dapat dilihat pada tabel sebagai berikut

Tabel 3.17: Tabel Tingkat Ketepatan Klasifikasi Pohon Optimum

Observasi	Hasil prediksi		Presentase
	jinak	ganas	Total Prediksi
Diagnosis kanker payudara jinak	36	9	80%
Diagnosis kanker payudara ganas	0	83	100%
Presentase keseluruhan	28.1%	71.9%	**93**%

Tingkat ketepatan pengklasifikasian pada metode ini adalah 93%. Dari seluruh total 45 pasien yang terdiagnosis kanker jinak ada sebanyak 36 pasien diklasifikasikan dengan benar, sedangkan dari 83 pasien yang terdiagnosis kanker ganas diklasifikasikan dengan benar seluruhnya yaitu 83 pasien. Kesalahan untuk hasil diagnosis kanker jinak yang diprediksi akan mengidap kanker ganas adalah sebanyak 9 pasien. Sehingga dapat dikatakan model ini sudah baik dalam mendiagnosis kanker payudara.

BAB IV

PENUTUP

4.1 Kesimpulan

Dari hasil analisis dan pembahasan mengenai klasifikasi diagnosis penyakit kanker payudara dengan pendekatan regresi logistik biner dan metode CART diperoleh kesimpulan sebagai berikut :

1. Hasil klasifikasi regresi logistik biner menunjukkan bahwa terdapat 55 pasien yaitu sebesar 80.5% yang diklasifikasikan secara benar akan mengidap kanker jinak sedangkan pada pasien berisiko tinggi mengidap kanker ganas yang diklasifikasikan secara benar sebesar 95.9% yaitu sebanyak 116 pasien.

   Variabel faktor pengaruh timbulnya penyakit kanker yang signifikan terhadap diagnosis kanker payudara yaitu usia dan riwayat keluarga penderita kanker. Model logit yang diperoleh adalah

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

   Dalam analisis ini mampu mengklasifikasikan diagnosis kanker payudara secara keseluruhan dengan nilai ketepatan klasifikasi sebesar 90.5%.

2. Proses pembentukan pohon CART menghasilkan variabel yang masuk ke dalam pohon klasifikasi optimum yaitu usia, riwayat keluarga penderita kanker, dan tidak menyusui anak.

Pohon optimum ini menghasilkan empat simpul terminal yang diklasifikasikan sebagai berikut

- Pasien yang memiliki riwayat tidak menyusui anak memiliki potensi sebesar 92.9% akan mengidap kanker ganas.
- Adapun memiliki potensi tinggi sebesar 100% dapat juga terjadi pada pasien yang menyusui anaknya.
- Tidak adanya riwayat keluarga penderita kanker payudara juga memiliki potensi mengidap kanker ganas sebesar 80.5%.
- Diagnosis penyakit kanker adalah jinak berpotensi sebesar 100% apabila usia pasien memiliki kriteria < 40 tahun.

Nilai ketepatan klasifikasi secara keseluruhan pada pohon optimum untuk pasien kanker payudara sebesar 93%, maka klasifikasi ini baik menggambarkan hasil diagnosis kanker payudara.

4.2 Saran

Faktor usia dan riwayat keluarga penderita kanker sangat berhubungan erat terhadap hasil diagnosis kanker payudara, oleh karena itu disarankan bagi seseorang pada usia tertentu dan memiliki riwayat keluarga penderita kanker payudara sebaiknya melakukan pemeriksaan lebih mendalam karena bila dibiarkan berdampak buruk berisiko dapat terdiagnosis memiliki penyakit kanker payudara.

Pada skripsi ini metode klasifikasi yang digunakan adalah analisis regresi logistik biner dan CART, sehingga pada penelitian selanjutnya dapat memilih metode klasifikasi berbeda yaitu analisis regresi logistik multinomial dengan mengambil penelitian lebih dari dua variabel respon dan dapat menggunakan

variabel respon berkategori kontinu pada metode CART yang akan menghasilkan pohon regresi.

DAFTAR PUSTAKA

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

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

Bustan, M.N. 2007. "*Epidemiologi Penyakit Tidak Menular*". Rineka Cipta, Jakarta.

Corwin, EJ. 2009. "*Buku Saku Patofisiologi*". Buku Kedokteran EGC, Jakarta.

Eubank, R. 1999. "*Nonparametric Regression and Spline Smoothing*". Marcel Dekker, Inc. New York.

Selviani Handayani, S.H, Purnami, S.W. 2014. "Pendekatan Metode *Classification and Regression Trees* untuk Diagnosis Tingkat Keganasan pada Pasien Kanker Tiroid". *Jurnal: Sains dan Seni Pomits, Vol.3, No.1.*

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

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

KemKes. R.I. 2015. "*Stop Kanker*". InfoDATIN, Jakarta.

Luwia, M.S. 2003. "*Problematik dan Perawatan Payudara*". Kawan Pustaka, Jakarta.

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

Putrayasa, I Nyoman. 2012. "*Model Regresi Logistik Biner dan Metode CART dalam Klasifikasi Status Desa di Bali*". SKRIPSI: Institut Pertanian Bogor.

Ramli, Muchlis. 1997. "*Epidemological Analysis of Risk Factor for Breast Cancer in Indonesia*". FKUI.

Rasjidi, Imam. 2009. "*Deteksi Dini dan Pencegahan Kanker Pada Wanita*". Sagung Seto, Jakarta.

Waluyo, A., Abdul Mukid, M., dan Wuryandari, T., . 2014. "Perbandingan Klasifikasi Nasabah Kredit Menggunakan Regresi Logistik Biner Dan CART (*Classification And Regression Trees*)". *Jurnal: Media Statistika, Vol.7, No.2.*

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

LAMPIRAN-LAMPIRAN

LAMPIRAN 1

Data Pasien Penyakit Kanker Payudara Tahun 2015

No	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	0	0	0	0	0	0	0	1
2	0	0	0	0	1	1	0	1
3	1	1	0	0	0	1	1	0
4	1	1	0	0	0	1	1	0
5	0	0	0	0	1	0	1	0
6	1	1	0	0	0	0	0	1
7	1	1	0	0	0	1	0	0
8	1	1	0	1	0	1	0	1
9	1	1	0	1	0	1	1	0
10	1	1	0	0	0	0	1	0
11	0	0	0	0	0	0	0	1
12	0	0	0	0	0	1	0	1
13	1	1	1	0	0	0	0	1
14	0	0	0	0	1	1	0	1
15	1	1	0	1	0	0	1	0
16	0	0	0	0	0	0	1	0
17	0	0	0	0	0	0	0	1
18	1	1	0	0	0	1	1	0
19	1	1	0	0	0	0	0	1
20	1	1	0	0	0	1	1	0
21	0	0	0	0	1	1	0	1
22	1	1	0	1	0	1	0	0
23	0	0	0	0	0	1	0	0
24	1	1	0	0	0	0	0	1
25	1	1	0	0	0	1	0	1
26	0	1	0	0	0	0	0	0
27	1	1	0	1	0	0	1	0
28	1	1	0	1	0	1	0	0
29	1	1	0	0	0	1	1	1
30	0	0	0	0	0	1	0	1

No	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
31	0	0	0	0	0	1	0	1
32	1	1	0	0	1	1	1	0
33	0	0	0	0	1	1	1	1
34	1	1	0	0	0	1	0	0
35	0	1	0	0	0	0	0	1
36	1	1	0	1	0	1	1	0
37	0	0	0	0	1	1	1	0
38	1	1	0	0	0	0	1	0
39	0	0	0	0	0	1	1	0
40	0	0	0	0	0	1	1	0
41	1	1	0	0	0	1	1	1
42	0	0	0	0	0	1	0	0
43	1	1	0	0	0	1	0	1
44	1	1	0	0	0	0	0	0
45	0	0	0	0	0	0	1	0
46	0	0	0	0	0	0	1	0
47	0	0	0	0	1	1	0	0
48	0	0	0	0	0	0	0	1
49	1	1	0	0	0	1	0	0
50	1	1	0	0	0	0	1	1
51	1	1	1	0	0	1	0	0
52	0	1	0	0	0	0	0	1
53	1	1	0	0	0	0	1	0
54	1	1	0	1	0	1	1	0
55	0	0	0	0	0	0	0	1
56	0	0	0	0	0	0	0	1
57	1	1	0	0	0	0	1	1
58	0	0	0	0	1	1	0	1
59	1	1	0	0	0	1	0	1
60	0	0	0	0	0	0	0	1
61	0	0	0	0	0	0	0	1
62	1	1	0	0	0	1	0	1
63	1	1	0	0	0	0	0	1
64	0	1	0	0	0	0	0	0
65	1	1	0	0	1	0	1	1
66	1	1	0	0	1	1	0	1

No	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
67	1	1	0	0	0	1	0	1
68	1	1	0	1	0	1	1	1
69	1	1	0	0	1	1	0	0
70	1	1	0	0	1	1	0	1
71	1	1	0	0	1	0	1	0
72	1	1	0	1	0	1	1	1
73	0	1	0	0	0	1	1	1
74	0	0	0	0	0	0	0	1
75	0	0	0	0	0	0	1	0
76	0	0	0	0	0	0	0	1
77	1	1	0	0	0	1	0	0
78	1	1	0	1	1	1	1	0
79	1	1	0	0	0	0	1	1
80	1	1	0	0	0	0	0	0
81	1	1	0	0	0	1	0	0
82	0	0	0	0	0	0	0	0
83	0	0	0	0	0	0	0	1
84	1	1	0	0	1	0	0	1
85	1	1	0	0	0	1	0	1
86	0	0	0	0	0	1	1	1
87	0	0	0	0	0	0	0	1
88	1	1	0	0	0	0	0	0
89	1	1	0	1	0	0	0	0
90	1	1	0	0	0	1	0	0
91	1	1	0	0	0	0	0	0
92	1	1	0	0	0	1	0	1
93	0	0	1	0	0	0	1	0
94	1	1	0	0	1	0	1	0
95	1	1	0	1	0	0	1	0
96	1	1	0	1	0	1	0	0
97	1	1	0	1	1	1	0	0
98	0	0	0	0	0	0	1	0
99	0	0	0	0	0	0	1	0
100	0	0	0	0	0	1	0	0
101	1	1	0	0	0	1	0	0
102	1	1	1	1	0	0	1	0

No	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
103	1	1	0	0	1	0	0	1
104	1	1	0	0	0	1	0	0
105	0	1	0	0	0	0	0	1
106	1	1	1	0	0	0	0	1
107	0	0	0	0	0	1	0	0
108	1	1	0	0	0	0	0	0
109	0	0	0	0	0	1	1	0
110	0	0	0	0	0	1	0	0
111	0	0	0	0	1	1	0	1
112	0	0	0	0	0	1	0	1
113	1	1	0	1	0	1	1	0
114	1	1	0	1	0	1	0	0
115	0	1	0	0	1	0	1	0
116	1	1	0	0	0	1	0	1
117	1	1	0	1	0	0	0	0
118	1	1	0	0	0	0	0	1
119	1	1	0	0	0	1	0	0
120	1	1	0	1	0	1	0	1
121	1	1	0	0	0	1	0	0
122	0	1	0	0	0	0	0	1
123	1	1	0	0	0	1	0	0
124	1	1	0	0	0	0	0	0
125	1	1	0	0	0	1	0	0
126	1	1	0	1	0	1	0	0
127	0	0	0	0	1	1	0	0
128	1	1	0	1	0	0	1	0
129	1	1	0	0	0	1	0	0
130	1	1	0	1	0	1	0	0
131	1	1	0	0	1	0	1	1
132	1	1	0	1	0	1	0	0
133	0	0	1	0	0	0	1	1
134	1	1	1	0	1	1	0	1
135	1	1	0	0	0	0	0	1
136	0	1	0	0	0	0	0	1
137	1	1	0	0	1	1	0	1
138	1	1	0	0	0	1	0	1

No	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
139	1	1	1	0	0	0	0	1
140	1	1	1	0	0	1	1	0
141	0	0	0	0	0	0	1	0
142	0	0	0	0	1	0	1	0
143	1	1	0	0	0	0	0	1
144	0	0	0	0	0	1	0	1
145	0	1	0	0	0	0	1	0
146	1	1	0	0	0	1	1	0
147	1	1	0	0	0	1	0	1
148	1	1	0	0	1	0	1	0
149	1	1	0	0	0	0	0	1
150	1	1	0	0	0	0	0	1
151	1	1	0	0	1	1	0	1
152	0	0	0	0	0	1	0	0
153	1	1	0	1	0	0	1	0
154	1	1	0	0	1	0	1	0
155	1	1	0	0	0	0	0	0
156	0	0	0	0	1	0	1	1
157	1	1	0	0	0	1	1	1
158	1	1	0	0	1	1	1	0
159	0	1	0	0	0	0	0	1
160	1	1	0	0	0	0	1	1
161	0	0	0	0	0	0	1	0
162	1	1	0	0	1	1	0	1
163	1	1	0	0	0	0	0	1
164	1	1	0	1	0	1	0	1
165	1	1	0	0	1	0	1	0
166	0	0	0	0	1	0	1	1
167	1	1	0	0	1	1	0	1
168	1	1	0	0	0	0	0	1
169	1	1	0	1	1	1	1	0
170	0	1	0	1	0	0	0	0
171	1	1	0	0	0	1	1	1
172	1	1	0	0	0	0	1	1
173	1	1	0	0	0	0	1	1
174	1	1	0	0	0	1	0	1

No	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
175	0	1	0	0	0	0	0	1
176	0	0	0	0	1	0	1	0
177	1	1	0	0	0	1	1	0
178	0	0	0	0	0	1	1	0
179	1	1	0	0	1	0	1	0
180	0	0	0	0	0	1	0	1
181	1	1	1	0	1	1	0	1
182	1	1	0	0	0	1	0	1
183	1	1	0	0	0	0	1	0
184	1	1	0	1	1	0	1	1
185	1	1	0	0	0	1	1	0
186	0	0	0	0	0	1	0	1
187	1	1	0	0	0	0	1	1
188	1	1	1	1	0	0	0	1
189	1	1	1	0	1	1	1	0

LAMPIRAN 2

Output Data Deskriptif dan Multikolinieritas

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
Diagnosis kanker	189	0	1	.64	.481	.232
Usia	189	24	59	40.70	8.632	74.510
Usia menarche	189	0	1	.06	.244	.060
Usia menopause	189	0	1	.16	.366	.134
Obesitas	189	0	1	.22	.413	.171
RKPK	189	0	1	.51	.501	.251
TMA	189	0	1	.39	.489	.240
KB	189	0	1	.49	.501	.251
Valid N (listwise)	189					

Regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.788[a]	.621	.606	.302

a. Predictors: (Constant), KB, Obesitas, Usia menarche, RKPK, Usia, TMA, Usia menopause

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	27.024	7	3.861	42.321	.000[a]
	Residual	16.511	181	.091		
	Total	43.534	188			

a. Predictors: (Constant), KB, Obesitas, Usia menarche, RKPK, Usia, TMA, Usia menopause

b. Dependent Variable: Diagnosis kanker

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-1.506	.139		-10.864	.000		
	Usia	.051	.003	.916	15.537	.000	.603	1.658
	Usia menarche	.056	.091	.028	.613	.540	.981	1.019
	Usia menopause	-.385	.079	-.293	-4.881	.000	.581	1.721
	Obesitas	.006	.055	.005	.106	.916	.947	1.056
	RKPK	.164	.046	.171	3.608	.000	.932	1.073
	TMA	.076	.049	.078	1.549	.123	.832	1.201
	KB	.021	.048	.022	.438	.662	.846	1.182

a. Dependent Variable: Diagnosis kanker

LAMPIRAN 3

Output Data dengan Analisis Regresi Logistik Biner

Logistic Regression

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	189	100.0
	Missing Cases	0	.0
	Total	189	100.0
Unselected Cases		0	.0
Total		189	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
jinak	0
ganas	1

Block 0: Beginning Block

Classification Table[a,b]

			Predicted		
			Diagnosis kanker		Percentage Correct
	Observed		jinak	ganas	
Step 0	Diagnosis kanker	jinak	0	68	.0
		ganas	0	121	100.0
	Overall Percentage				64.0

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.576	.152	14.458	1	.000	1.779

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	X1	103.360	1	.000
		X2	2.075	1	.150
		X3	16.499	1	.000
		X4	.008	1	.927
		X5	4.377	1	.036
		X6	.254	1	.614
		X7	1.152	1	.283
	Overall Statistics		117.320	7	.000

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	145.865	7	.000
	Block	145.865	7	.000
	Model	145.865	7	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	101.081[a]	.538	.737

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.672	8	.289

Classification Table[a]

Observed			Predicted		
			Diagnosis kanker		Percentage Correct
			jinak	ganas	
Step 1	Diagnosis kanker	Jinak	55	13	80.9
		Ganas	5	116	95.9
	Overall Percentage				90.5

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	X1	.403	.063	40.512	1	.000	1.497
	X2(1)	1.216	1.330	.835	1	.361	3.373
	X3(1)	-2.420	1.327	3.325	1	.068	.089
	X4(1)	.244	.642	.144	1	.704	1.276
	X5(1)	1.704	.606	7.898	1	.005	5.496
	X6(1)	.813	.576	1.994	1	.158	2.255
	X7(1)	.165	.544	.092	1	.761	1.180
	Constant	-16.620	2.762	36.217	1	.000	.000

a. Variable(s) entered on step 1: X1, X2, X3, X4, X5, X6, X7.

LAMPIRAN 4

Output Data dengan Metode CART

Classification Tree

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	Diagnosis kanker
	Independent Variables	Usia, Usia menarche, Usia menopause, Obesitas, RKPK, TMA, KB
	Validation	Split Sample
	Maximum Tree Depth	5
	Minimum Cases in Parent Node	50
	Minimum Cases in Child Node	10
Results	Independent Variables Included	Usia, RKPK, TMA
	Number of Nodes	7
	Number of Terminal Nodes	4
	Depth	3

Risk

Sample	Estimate	Std. Error
Training	.070	.023
Test	.066	.032

Growing Method: CRT
Dependent Variable: Diagnosis kanker

Classification

		Predicted		
Sample	Observed	jinak	ganas	Percent Correct
Training	jinak	36	9	80.0%
	ganas	0	83	100.0%
	Overall Percentage	28.1%	71.9%	93.0%
Test	jinak	19	4	82.6%
	ganas	0	38	100.0%
	Overall Percentage	31.1%	68.9%	93.4%

Growing Method: CRT
Dependent Variable: Diagnosis kanker

Tree Table

Sample	Node	jinak		ganas		Total		Predicted Category	Parent Node	Primary Independent Variable		
		N	Percent	N	Percent	N	Percent			Variable	Improvement	Split Values
Training	0	45	35.2%	83	64.8%	128	100.0%	ganas				
	1	36	100.0%	0	.0%	36	28.1%	jinak	0	Usia	.329	<= <40 tahun
	2	9	9.8%	83	90.2%	92	71.9%	ganas	0	Usia	.329	> <40 tahun
	3	8	19.5%	33	80.5%	41	32.0%	ganas	2	RKPK	.011	tidak
	4	1	2.0%	50	98.0%	51	39.8%	ganas	2	RKPK	.011	ya
	5	0	.0%	37	100.0%	37	28.9%	ganas	4	TMA	.001	tidak
	6	1	7.1%	13	92.9%	14	10.9%	ganas	4	TMA	.001	ya
Test	0	23	37.7%	38	62.3%	61	100.0%	ganas				
	1	19	100.0%	0	.0%	19	31.1%	jinak	0	Usia	.329	<= <40 tahun
	2	4	9.5%	38	90.5%	42	68.9%	ganas	0	Usia	.329	> <40 tahun
	3	4	17.4%	19	82.6%	23	37.7%	ganas	2	RKPK	.011	tidak
	4	0	.0%	19	100.0%	19	31.1%	ganas	2	RKPK	.011	ya
	5	0	.0%	9	100.0%	9	14.8%	ganas	4	TMA	.001	tidak
	6	0	.0%	10	100.0%	10	16.4%	ganas	4	TMA	.001	ya

Growing Method: CRT
Dependent Variable: Diagnosis kanker

SURAT PERNYATAAN KEASLIAN SKRIPSI

Dengan ini saya yang bertanda tangan di bawah ini, mahasiswa Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta:

Nama	:	Rifqy Marwah Akhsanti
No. Registrasi	:	3125110233
Program Studi	:	Matematika

Menyatakan bahwa skripsi ini yang saya buat dengan judul "**Klasifikasi Diagnosis Penyakit Kanker Payudara Dengan Pendekatan Regresi Logistik Biner dan Metode *Classification And Regression Trees* (CART)**" adalah :

1. Dibuat dan diselesaikan oleh saya sendiri.
2. Bukan merupakan duplikat skripsi yang pernah dibuat oleh orang lain atau jiplakan karya tulis orang lain.

Pernyataan ini dibuat dengan sesungguhnya dan saya bersedia menanggung segala akibat yang timbul jika pernyataan saya tidak benar.

Jakarta, 10 Februari 2017

Yang membuat pernyataan

Rifqy Marwah Akhsanti

DAFTAR RIWAYAT HIDUP

RIFQY MARWAH AKHSANTI Lahir di Subang, 30 September 1993. Anak pertama dari pasangan Bapak Jumadi dan Ibu Nurhayati. Bertempat tinggal di Griya Mutiara Blok B No 17 RT 003/RW 002, Kec. Purwadadi, Kab. Subang, Jawa Barat 41261.

No. Ponsel : 082213594450

Email : rifqi.almarwah@gmail.com

Riwayat Pendidikan : Penulis mengawali pendidikan di SDN Cikampek Utara IV pada tahun 1998 - 2004. Setelah itu, penulis melanjutkan ke MTs Ma'had Al-Zaytun Indramayu hingga tahun 2008. Kemudian tetap kembali melanjutkan ke MA Ma'had Al-Zaytun Indramayu dan lulus tahun 2011. Di Tahun yang sama penulis melanjutkan ke Universitas Negeri Jakarta (UNJ), jurusan Matematika, melalui jalur SNMPTN Undangan.

Riwayat Organisasi : Selama di bangku perkuliahan, penulis aktif di berbagai organisasi kemahasiswaaan yaitu Tank FMIPA, Desabinaan FMIPA. Dalam dua tahun pertama, penulis mendapat kepercayaan sebagai staff Profesi dan Keilmuan BEMJ Matematika, dan sebagai ketua acara kegiatan "Career Sharing".

Riwayat Pekerjaan : Penulis mulai menjadi pengajar privat matematika untuk SD, SMP, dan SMA sejak tahun 2011 hingga saat ini. Pada tahun 2014, penulis juga menjadi pegawai magang di PT. Bank Muamalat.Tbk. selama kurang lebih 3 bulan serta di Kementrian Perdagangan R.I. selama kurang lebih 1 bulan. Dan di tahun 2015 penulis menjadi pengajar matematika di Ganesha Operation (GO) Jakarta selama kurang lebih 1 tahun.