PERBANDINGAN METODE REGRESI LOGISTIK BINER DAN *CLASSIFICATION AND REGRESSION TREES* (CART) UNTUK KLASIFIKASI DIAGNOSA PENYAKIT DIABETES MELLITUS (DM)

SKRIPSI

Disusun Oleh
PUJI WIDIARTI
H72216062

PROGRAM STUDI MATEMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI SUNAN AMPEL
SURABAYA

2020

PERNYATAAN KEASLIAN

Saya yang bertanda tangan di bawah ini,

Nama	:	PUJI WIDIARTI
NIM	:	H72216062
Program Studi	:	Matematika
Angkatan	:	2016

Menyatakan bahwa saya tidak melakukan plagiat dalam penulisan skripsi saya yang berjudul " PERBANDINGAN METODE REGRESI LOGISTIK BINER DAN *CLASSIFICATION AND REGRESSION TREES* (CART) UNTUK KLASIFIKASI DIAGNOSA PENYAKIT DIABETES MELLITUS (DM) ". Apabila suatu saat nanti terbukti saya melakukan tindakan plagiat, maka saya bersedia menerima sanksi yang telah ditetapkan.

Demikian pernyataan keaslian ini saya buat dengan sebenar-benarnya.

Surabaya, 23 Juli 2020

Yang menyatakan,

PUJI WIDIARTI

NIM. H72216062

LEMBAR PERSETUJUAN PEMBIMBING

Skripsi oleh

Nama : PUJI WIDIARTI

NIM : H72216062

Judul Skripsi : PERBANDINGAN METODE REGRESI LOGISTIK BINER DAN *CLASSIFICATION AND REGRESSION TREES* (CART) UNTUK KLASIFIKASI DIAGNOSA PENYAKIT DIABETES MELLITUS (DM)

telah diperiksa dan disetujui untuk diujikan.

Surabaya, 10 Juli 2020

Pembimbing

Wika Dianita Utami, M.Sc

NIP. 19920610208012003

PENGESAHAN TIM PENGUJI SKRIPSI

Skripsi oleh

Nama	:	PUJI WIDIARTI
NIM	:	H72216062
Judul Skripsi	:	PERBANDINGAN METODE REGRESI LOGISTIK BINER DAN *CLASSIFICATION AND REGRESSION TREES* (CART) UNTUK KLASIFIKASI DIAGNOSA PENYAKIT DIABETES MELLITUS (DM)

Telah dipertahankan di depan Tim Penguji
pada tanggal 23 Juli 2020

Mengesahkan,
Tim Penguji

Penguji I

Wika Dianita Utami, M.Sc
NIP. 199206102018012003

Penguji II

Dr. Moh. Hafiyusholeh, M.Si
NIP. 198002042014031001

Penguji III

Putroue Keumala Intan, M.Si.
NIP. 198805282018012001

Penguji IV

Nurissaidah Ulinnuha, M.Kom
NIP. 199901102201403200 4

Mengetahui,
Plt. Dekan Fakultas Sains dan Teknologi
UIN Sunan Ampel Surabaya

Dr. Hj. Evi Fatimatul Rusydiyah, M.Ag
NIP. 197312272005012003

KEMENTERIAN AGAMA
UNIVERSITAS ISLAM NEGERI SUNAN AMPEL SURABAYA
PERPUSTAKAAN
Jl. Jend. A. Yani 117 Surabaya 60237 Telp. 031-8431972 Fax.031-8413300
E-Mail: perpus@uinsby.ac.id

LEMBAR PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademika UIN Sunan Ampel Surabaya, yang bertanda tangan di bawah ini, saya:

Nama	:	PUJI WIDIARTI
NIM	:	H72216062
Fakultas/Jurusan	:	SAINTEK/MATEMATIKA
E-mail address	:	pujiwidiarti2@gmail.com

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Perpustakaan UIN Sunan Ampel Surabaya, Hak Bebas Royalti Non-Eksklusif atas karya ilmiah :
■ Skripsi ☐ Tesis ☐ Desertasi ☐ Lain-lain (.....)
yang berjudul :

PERBANDINGAN METODE REGRESI LOGISTIK BINER DAN *CLASSIFICATION AND REGRESSION TREES* (CART) UNTUK KLASIFIKASI DIAGNOSA PENYAKIT DIABETES MELLITUS (DM)

beserta perangkat yang diperlukan (bila ada). Dengan Hak Bebas Royalti Non-Ekslusif ini Perpustakaan UIN Sunan Ampel Surabaya berhak menyimpan, mengalih-media/format-kan, mengelolanya dalam bentuk pangkalan data (database), mendistribusikannya, dan menampilkan/mempublikasikannya di Internet atau media lain secara ***fulltext*** untuk kepentingan akademis tanpa perlu meminta ijin dari saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan atau penerbit yang bersangkutan.

Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak Perpustakaan UIN Sunan Ampel Surabaya, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak *Cipta dalam karya ilmiah saya ini.*

Demikian pernyataan ini yang saya buat dengan sebenarnya.

Surabaya, 23 Juli 2020

Penulis

(PUJI WIDIARTI)

ABSTRAK

PERBANDINGAN METODE REGRESI LOGISTIK BINER DAN *CLASSIFICATION AND REGRESSION TREES* (CART) UNTUK KLASIFIKASI DIAGNOSA PENYAKIT DIABETES MELLITUS (DM)

Diabetes Mellitus merupakan suatu penyakit yang mempunyai karakteristik hiperglikemia (kadar gula darah tinggi $\geq$ 187 mg/dL) sedangkan kadar gula darah normal pada tubuh manusia $\leq$ 140 mg/dL. Penyakit ini mempunyai risiko tinggi sehingga diperlukan tes laboratorium untuk menentukan apakah pasien mengidap penyakit Diabetes Mellitus. Oleh karena itu, perlu metode klasifikasi secara tepat untuk mengklasifikasikan data sesuai dengan kriteria *class*. Variabel yang digunakan dalam penelitian ini yaitu usia, faktor keturunan, kadar gula darah, obesitas dan pola makan sedangkan variabel respon yaitu diagnosa pasien (negatif atau positif Diabetes Mellitus). Metode klasifikasi yang digunakan adalah Regresi Logistik Biner dan CART (*Classfication And Regression Trees*). Berdasarkan data *testing* sebesar 30% pengklasifikasian menggunakan metode Regresi Logistik Biner diperoleh ketepatan hasil klasifikasi sebesar 91,67% dan terdapat empat faktor yang berpengaruh, yaitu usia, faktor keturunan, kadar gula darah dan obesitas. Sementara itu, metode CART diperoleh ketepatan hasil klasifikasi sebesar 90,27% dan terdapat 13 *rule*, yaitu 5 *rule* dengan kategorik positif Diabetes Mellitus dan 8 *rule* dengan kategorik negatif Diabetes Mellitus. Dalam penelitian ini, metode klasifikasi menggunakan Regresi Logistik Biner lebih baik daripada CART.

Kata kunci: Diabetes Mellitus, Klasifikasi, Regresi Logistik Biner, CART

ABSTRACT

COMPARISON OF BINARY LOGISTIC REGRESSION AND CLASSIFICATION AND REGRESSION TRESS (CART) METHODS FOR THE CLASSIFICATION OF DIAGNOSES OF DIABETES MELLITUS (DM)

Diabetes Mellitus is a disease that has the characteristics of hyperglycemia (blood sugar level $\geq$ 187 mg/dL) while the normal blood sugar level in the human body $\leq$ 140 mg/dL. This disease has a high risk laboratory testing is needed to determine whether the patient has Diabetes Mellitus. Therefore, an appropriate classification method is needed to classify data according to class criteria. The variables used in this study are age, heredity, blood sugar levels, obesity and eating patterns while the response variable is the patients diagnosis (negative or positive Diabetes Mellitus). The classification methods used are Binary Logistic Regression and CART (Classification And Regression Trees). Based on testing data for 30% of classification using the Binary Logistic Regression method obtained the accuracy of the classification results of 91,67% and obtained four influential factors, that is age, heredity, blood sugar levels and obesity. Meanwhile, the CART method obtained the accuracy of the classification results of 90,27 % and obtained 13 rules, that is 5 rules with positive categories of Diabetes Mellitus and 8 rules with negative categories of Diabetes Mellitus. In this study, the classification method using Binary Logistic Regression is better than CART.

Keywords: Diabetes Mellitus, *Classification*, Binary Logistic Regression, CART

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Kerusakan yang terjadi pada organ kerja tubuh manusia mempunyai dampak buruk serta menjadi salah satu sumber permasalahan di dunia ini. Terdapat berbagai macam jenis penyakit di dunia ini yang berujung kematian, salah satu penyakit tersebut adalah Diabetes Mellitus (DM). Diabetes Mellitus atau dikenal kencing manis merupakan penyakit yang disebabkan oleh insulin yang kurang baik dan gangguan fungsi insulin dalam tubuh penderita. Diabetes Mellitus yaitu penyakit yang mempunyai ciri hiperglikemia (gula darah tinggi), hal ini dikarenakan terjadi gangguan pada kerja insulin, sekresi insulin atau kedua-duanya (IDF, 2017).

Berdasarkan data dari *Internasional of Diabetic Ferderation* (IDF), kasus Diabetes Mellitus dengan karakteristik kelainan sekresi insulin dan/atau kelainan sensitivitas insulin mencapai angka sebesar 90%. IDF menyatakan tingkat prevalensi global pengidap Diabetes Mellitus (DM) pada tahun 2016 mencapai 8,3% dari totalitas penduduk di dunia serta mengklaim kenaikan pada tahun 2016 menjadi 387 juta kasus. Kasus tersebut menjadi permasalahan kesehatan di dunia dikarenakan tingkat insiden penyakit ini terus mengalami peningkatan, baik di negara industri maupun negara berkembang, termasuk di Indonesia. Pada tahun 2017, dengan angka penderita Diabetes Mellitus terus meningkat Indonesia menjadi urutan keenam di dunia (IDF, 2017).

Terdapat beberapa faktor penyebab seseorang mengidap penyakit Diabetes Mellitus, antara lain faktor genetik, faktor lingkungan, faktor obesitas, faktor demografi dan lain sebagainya. Beberapa faktor risiko tersebut harus diperhatikan agar berdampak signifikan terhadap penurunan angka penderita Diabetes Mellitus serta penurunan kematian. Dengan demikian, perlu adanya program pengendalian Diabetes Mellitus, dengan cara mengendalikan faktor risiko tersebut (Misnadiarly, 2006).

Faktor risiko dapat dibedakan menjadi dua, yaitu faktor risiko dapat diubah dan tidak dapat diubah. Faktor risiko yang dapat diubah seperti pola hidup yang kurang baik misalnya kebiasaan merokok, kurang berolahraga dan mengkonsumsi makanan tidak sehat. Sementara itu, faktor risiko yang tidak dapat diubah misalnya faktor genetik, jenis kelamin dan usia.

Walaupun telah diketahui beberapa faktor yang berpengaruh terhadap pengidap Diabetes Mellitus, akan tetapi dalam menentukan seseorang terserang Diabetes Mellitus sangat sulit untuk ditentukan (WHO, 2006). Seiring dengan perkembangan teknologi dokter spesialis menggunakan rekam medis dan uji laboratorium untuk memperoleh data yang valid. Data yang didapatkan dari hasil rekam medis tersebut selanjutnya akan didiagnosa oleh dokter apakah pasien mengidap Diabetes Mellitus. Berdasarkan data tersebut dapat dibentuk model matematis, yaitu dengan metode klasifikasi (Widagdo, 2010).

Metode klasifikasi salah satunya dapat diselesaikan menggunakan metode statistik. Pada penyelesaian masalah klasifikasi (pengelompokkan) perlu diperhatikan dalam menentukan metode klasifikasi yang tepat, misalkan ingin mengklasifikasikan pasien yang mengidap penyakit Diabetes Mellitus dan tidak mengidap Diabetes Mellitus. Apabila mengklasifikasikan pasien pengidap penyakit Diabetes Mellitus ke dalam kelompok pasien yang tidak mengidap

penyakit Diabetes Mellitus maka akan menimbulkan suatu kesalahan yang fatal. Pengklasifikasian suatu objek dapat menggunakan pendekatan metode parametrik dan nonparametrik (Widagdo, 2010).

Salah satu metode statistik dengan pendekatan parametrik adalah Regresi Logistik Biner. Analisis Regresi merupakan salah satu metode statistik yang digunakan untuk mempelajari hubungan antara variabel independen dengan variabel dependen. Regresi Logistik merupakan salah satu metode yang digunakan untuk menganalisa data yang mempunyai variabel dependen berupa data kategorik (Margasari, 2014).

Regresi Logistik tidak mengasumsikan hubungan linier antar variabel independen dan dependen, dikarenakan bentuk varibel dependen berupa kategorik. Variabel dependen dengan dua kategorik (biner) yang bertolak belakang dapat dilakukan pengujian statistika menggunakan metode Regresi Logistik Biner. Variabel dependen biasanya diprediksi dengan nilai 0 dan 1, dimana 0 menyatakan bahwa pasien negatif Diabetes Mellitus dan 1 menyatakan pasien positif Diabetes Mellitus (Rumaendra, 2016).

Sementara itu, klasifikasi dapat menggunakan metode pendekatan nonparametrik. Pendekatan nonparametrik tidak bergantung terhadap asumsi tertentu, sehingga dapat memperoleh fleksibilitas yang lebih besar untuk menganalisa suatu data. Akan tetapi, tetap memiliki tingkat akurasi yang tinggi serta mudah dalam penerapannya. Salah satu metode statistik untuk klasifikasi dengan pendekatan nonparametrik adalah metode *Classification And Regression Trees* (CART). Pada tahun 1984, metode tersebut diperkenalkan oleh Leo Breiman, Richard A. Olshen, Jerome H. Friedman serta Charles J. Stone. Keempat ilmuwan tersebut memperkenalan metode *Classification And Regression Trees* (CART) sebagai metode pohon klasifikasi dan regresi. CART dikembangkan untuk

menganalisa klasifikasi, baik variabel dependen kontinu maupun kategorik. CART akan menghasilkan *output* pohon klasifikasi (*classification trees*) apabila variabel Y berupa kategorik dan menghasilkan *output* pohon regresi (*regression trees*) apabila variabel Y berupa kontinu atau numerik (Breiman dkk., 1984).

Metode CART mempunyai kelebihan struktur data yang digunakan dapat dilihat secara visual, dapat mengeksplorasi struktur data yang kompleks dan bersifat nonparametrik sehingga tidak membutuhkan asumsi tertentu yang sering tidak terpenuhi oleh suatu data, serta proses pengklasifikasian lebih mudah dilakukan dengan cara menelusuri pohon klasifikasi yang dihasilkan (Breiman dkk., 1993).

Berdasarkan penelitian yang telah dilakukan oleh Wella Rumaenda (2016) dengan judul Perbandingan Klasifikasi Penyakit Hipertensi Menggunakan Regresi Logistik Biner dan Algoritma C4.5. Dalam penelitian tersebut bertujuan untuk membandingkan ketepatan klasifikasi antara metode Regresi Logistik Biner dengan Algoritma C4.5. Berdasarkan metode yang digunakan hasil klasifikasi menggunakan metode Regresi Logistik Biner diperoleh ketepatan klasifikasi sebesar 72,5352% sedangkan menggunakan Algoritma C4.5 didapatkan ketepatan klasifikasi sebesar 64,0845% (Rumaendra, 2016).

Penelitian lain yang telah dilakukan oleh R. Lestawati, Rais dan I. T. Utami (2016) dengan judul Perbandingan antara Metode CART (*Classification And Regression Tress*) dan Regresi Logistik (*Logistic Regression*) dalam Mengklasifikasikan Pasien Penderita DBD (Demam Berdarah *Dengue*). Penelitian tersebut bertujuan untuk memperoleh ketepatan hasil klasifikasi antara metode CART dan Regresi Logistik serta variabel yang signifikan terhadap penyakit DBD. Berdasarkan metode yang digunakan hasil klasifikasi menggunakan metode CART diperoleh ketepatan klasifikasi sebesar 76,3% dengan variabel yang signifikan

yaitu hepatomegali, epitaksis, melena dan diare sedangkan hasil klasifikasi menggunakan metode Regresi Logistik Biner diperoleh ketepatan klasifikasi sebesar 76,7% dan variabel yang signifikan yaitu hepatomegali (Lestawati dkk., 2018).

Penelitian lain yang telah dilakukan oleh Rifqy Marwah Akhsanti, Widyanti Rahayu dan Vera Maya Santi (2018) dengan judul Klasifikasi Diagnosis Penyakit Kanker Payudara dengan Pendekatan Regresi Logistik Biner dan Metode *Classification And Regression Tress* (CART). Penelitian tersebut bertujuan untuk mengetahui faktor pengaruh timbulnya kanker payudara yang diklasifikasikan menurut dua kategorik, yaitu jinak dan ganas serta mengetahui ketepatan hasil klasifikasi dari dua metode tersebut. Berdasarkan penelitian, hasil yang diperoleh dengan analisis Regresi Logistik Biner faktor yang berpengaruh signifikan terhadap hasil diagnosis kanker payudara adalah usia dan riwayat keluarga penderita kanker (RKPK) serta ketepatan hasil klasifikasi yang diperoleh sebesar 90,5%. Sementara itu, metode CART menghasilkan pohon klasifikasi optimum dengan empat simpul terminal dan ketepatan hasil klasifikasi yang diperoleh sebesar 93% (Akhsanti dkk., 2018).

Berdasarkan latar belakang tersebut, tidak selamanya akurasi dari metode CART lebih baik daripada Regresi Logistik Biner untuk klasifikasi suatu kasus, begitupula sebaliknya. Oleh karena itu, dalam penelitian ini akan dilakukan klasifikasi penyakit Diabetes Mellitus menggunakan perbandingan metode Regresi Logistik Biner dan CART. Dengan demikian, dalam penelitian ini penulis mengambil judul "PERBANDINGAN METODE REGRESI LOGISTIK BINER DAN *CLASSIFICATION AND REGRESSION TREES* (CART) UNTUK KLASIFIKASI DIAGNOSA PENYAKIT *DIABETES MELLITUS* (DM)".

1.2. Rumusan Masalah

Berdasarkan latar belakang yang diuraikan di atas, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana model Regresi Logistik Biner dalam klasifikasi penyakit Diabetes Mellitus?
2. Bagaimana struktur pohon klasifikasi dalam penyakit Diabetes Mellitus menggunakan metode *Classification And Regression Trees* (CART)?
3. Bagaimana hasil perbandingan tingkat akurasi klasifikasi penyakit Diabetes Mellitus menggunakan metode Regresi Logistik Biner dan *Classification And Regression Trees* (CART)?

1.3. Tujuan Penelitian

Adapun tujuan dalam penelitian ini adalah sebagai berikut:

1. Untuk mengetahui model Regresi Logistik Biner dalam klasifikasi penyakit Diabetes Mellitus.
2. Untuk mengetahui struktur pohon klasifikasi dalam penyakit Diabetes Mellitus menggunakan metode *Classification And Regression Trees* (CART).
3. Untuk mengetahui perbandingan tingkat akurasi klasifikasi penyakit Diabetes Mellitus menggunakan metode Regresi Logistik Biner dan *Classification And Regression Trees* (CART).

1.4. Manfaat Penelitian

Adapun manfaat yang dapat diambil dari penelitian ini adalah sebagai berikut:

1. Bagi Penulis:

   Dapat memperdalam wawasan keilmuan mengenai metode klasifikasi Regresi Logistik Biner dan metode *Classification And Regression Trees* (CART) dalam tingkat kesehatan.

2. Bagi Pembaca:

   Dapat menambah ilmu pengetahuan atau menjadi rujukan bagi pembaca dalam melakukan penelitian selanjutnya dengan metode terbaik untuk pengklasifikasian pada tingkat kesehatan.

3. Bagi Kementerian Kesehatan Indonesia:

   Penelitian ini dapat membantu untuk memprediksi seseorang apakah mengidap penyakit Diabetes Mellitus.

1.5. Batasan Masalah

Adapun batasan masalah dalam penelitian ini terdiri dari variabel dependen dan independen. Varibel dependen berskala kategorik (0 = negatif Diabetes Mellitus, 1 = positif Diabetes Mellitus). Sementara itu, variabel independen yang digunakan dalam penelitian ini meliputi usia, faktor keturunan, kadar gula darah, obesitas serta pola makan.

1.6. Sistematika Penulisan

Adapun sistematika penulisan dalam penelitian ini adalah sebagai berikut:

BAB I PENDAHULUAN

Bagian ini menjelaskan mengenai permasalahan apa saja yang menjadi latar belakang penelitian, yaitu dalam mengklasifikasikan pasien yang mengidap penyakit Diabetes Mellitus ke dalam kelompok pasien yang tidak mengidap penyakit Diabetes Mellitus akan terjadi kesalahan yang cukup fatal, sehingga diperlukan metode klasifikasi yang tepat dengan membandingkan pendekatan parametrik dan nonparametrik yaitu menerapkan metode Regresi Logistik Biner sebagai pendekatan parametrik dan metode *Classification And Regression Trees* (CART) sebagai pendekatan nonparametrik. Selain itu, pada bagian ini penulis juga menjelaskan mengenai rumusan masalah, tujuan penelitian, batasan masalah penelitian, manfaat penelitian serta sistematika penulisan.

BAB II KAJIAN PUSTAKA

Bagian ini menjelaskan mengenai teori-teori yang mendukung untuk menyelasikan permasalahan yang berhubungan dengan penelitian. Sebagai tinjauan terhadap beberapa literatur yang terkait dengan topik penelitian yang meliputi Diabetes Mellitus, faktor-faktor, tahapan dalam pengklasifikasian menggunakan metode Regresi Logistik Biner dan metode *Classification And Regression Trees* (CART) serta metode untuk mengukur tingkat keakurasian dari kedua metode tersebut.

BAB III METODOLOGI PENELITIAN

Bagian ini menjelaskan mengenai jenis penelitian, sumber data, operasional variabel dan tahapan-tahapan yang dilakukan dalam penelitian disertai dengan diagram alir (*flowchart*).

BAB IV HASIL DAN PEMBAHASAN

Bagian ini menjelaskan mengenai perhitungan untuk mengklasifikasikan penyakit DM berdasarkan faktor-faktor yang mempengaruhi sehingga membentuk model pada masing-masing metode. Selain itu, diperoleh tingkat keakurasian dari model yang didapat.

BAB V PENUTUP

Bagian ini menjelaskan mengenai kesimpulan dari analisis hasil dan saran.

BAB II

TINJAUAN PUSTAKA

Topik yang akan dibahas dalam penelitian ini adalah klasifikasi diagnosis penyakit Diabetes Mellitus menggunakan metode Regresi Logistik Biner dan *Classification And Regression Trees* (CART). Dalam bab ini akan dibahas terlebih dahulu mengenai Diabetes Mellitus, faktor-faktor, Regresi Logistik Biner dan CART.

2.1. Diabetes Mellitus

World Health Organization (WHO) menyatakan bahwa Diabetes Mellitus (DM) tidak dapat didefinisikan secara ringkas dan jelas. Namun, DM dapat dikatakan sebagai kumpulan permasalahan yang kompleks mengenai anatomi dan kimiawi akibat beberapa faktor dimana terdapat defisiensi insulin (Sudoyo dkk, 2009).

Diabetes Mellitus atau yang sering disebut dengan penyakit kencing manis disebabkan oleh insulin yang kurang baik dan gangguan fungsi insulin dalam tubuh penderita. Diabetes Mellitus merupakan suatu penyakit yang mempunyai karakteristik hiperglikemia, hal ini terjadi karena gangguan pada sekresi insulin, kerja insulin atau kedua-duanya. Hiperglikemia yang krogis pada Diabetes Mellitus akan menyebabkan beberapa kerusakan organ pada tubuh manusia, antara lain ginjal, saraf, mata, jantung dan pembuluh darah (ADA, 2009).

Menurut Dorland Diabetes Mellitus merupakan suatu kondisi dimana terjadi peningkatan kadar gula darah yang melebihi nilai normal (kronis) pada

tubuh manusia sebagai akibat dari gangguan pada metabolisme karbohidrat, protein dan lemak karena kekurangan hormon insulin (Dorland, 2005). Kadar gula darah yang melebihi nilai normal yaitu apabila kadar gula darah sewaktu sama atau lebih dari 200 mg/dl. Secara umum, diagnosis Diabetes Mellitus ditandai dengan gejala, antara lain *polyuria* (banyak kencing), *polifagia* (banyak makan), *polydipsia* (banyak minum) serta penurunan berat badan yang sebabnya tidak jelas (Misnadiarly, 2006).

Seseorang mengidap penyakit Diabetes Mellitus disebabkan oleh beberapa faktor, antara lain faktor genetik, faktor lingkungan, faktor obesitas dan faktor demografi. Beberapa faktor risiko tersebut harus diperhatikan agar berdampak signifikan terhadap penurunan angka penderita Diabetes Mellitus serta penurunan kematian. Berdasarkan *World Health Organization* (WHO), beberapa faktor risiko penderita Diabetes Mellitus, antara lain faktor genetik, usia, kadar gula darah, pola makan dan indeks massa tubuh (IMT) (WHO, 2006).

2.2. Faktor-faktor Risiko Diabetes Mellitus

1. Usia

   Apabila dilihat dari faktor risiko, seseorang yang berumur lebih dari atau mencapai 45 tahun rentan berpotensi mengidap Diabetes Mellitus dibandingkan dengan seseorang yang berusia kurang dari 45 tahun (ADA, 2009). Kategori usia terhadap Diabetes Mellitus disajikan pada Tabel 2.1.

Tabel 2.1 Kategori Usia

Kategori	Usia (Tahun)
Normal	< 45
Tidak Normal	≥ 45

Berdasarkan Tabel 2.1, bahwa seseorang dengan usia kurang dari 45 mempunyai potensi kecil atau tidak terjadi Diabetes Mellitus. Sementara itu, seseorang yang berusia lebih dari atau mencapai 45 tahun mempunyai potensi mengidap Diabetes Mellitus (ADA, 2009).

2. Fungsi Silsilah Diabetes Mellitus (genetik)

Timbulnya penyakit *Diabetes Mellitus* sangat dipengaruhi oleh faktor keturunan (genetik). Apabila terjadi mutasi gen mengakibatkan kerusakan metabolisme yang berujung timbulnya Diabetes Mellitus. Risiko anak mempunyai Diabetes Mellitus adalah 15% apabila salah satu orang tuanya mengidap Diabetes Mellitus. Sementara itu, apabila kedua orang tua mempunyai Diabetes Mellitus, maka risiko anak untuk mengidap Diabetes Mellitus adalah 75%. Orang yang ibu kandungnya mempunyai Diabetes Mellitus risikonya 10% – 30% lebih besar daripada orang yang mempunyai ayah dengan Diabetes Mellitus. Hal ini dikarenakan penurunan gen sewaktu dalam kandungan lebih besar dari ibu. Apabila saudara kadung menderita Diabetes Mellitus, maka risiko menderita Diabetes Mellitus 10% dan 90% apabila yang menderita merupakan saudara kembar identik. Pengukuran fungsi silsilah Diabetes Mellitus adalah sebagai berikut (Diabetes UK, 2010):

$$DPF = \frac{\sum_i K_i(88 - ADM_i) + 20}{\sum_i K_j(ALC_j - 14) + 50} \tag{2.1}$$

Dihitung ketika tes hanya menggunakan data yang tersedia di hari tersebut, dimana:

DPF = fungsi silsilah Diabetes Mellitus

i = jarak dari seluruh keluarga yang memiliki diabetes berdasarkan hari

dimana subjek menjalani tes.

j = jarak dari seluruh keluarga yang tidak memiliki diabetes berdasarkan hari dimana subjek menjalani tes.

K = jumlah persen plasma pembawa sifat yang dibagikan oleh keluarga (50%, jika keluarga tersebut orang tua/saudara kandung), (25%, jika dari kakek-nenek, paman-bibi), dan (12,5%, jika saudara sepupu pertama).

ADM = umur keluarga saat terdiagnosa diabetes.

ACL = umur keluarga pada waktu terakhir kali melakukan tes non-diabetes.

Kategori fungsi silsilah Diabetes Mellitus dapat disajikan pada Tabel 2.2.

Tabel 2.2 Kategori Fungsi Silsilah Diabetes Mellitus

Kategori	**Fungsi Silsilah Diabetes Mellitus (Faktor Genetik)**
Normal	$\leq 0,49$
Pra-Diabetes Mellitus	0,50 - 0,53
Tidak Normal	$\geq 0,54$

Berdasarkan Tabel 2.2, apabila seseorang memiliki keluarga dengan riwayat Diabetes Mellitus, kemungkinan besar orang tersebut mengidap Diabetes Mellitus. Seseorang mengalami pra-Diabetes Mellitus apabila nilai faktor genetiknya antara 0,50 - 0,53. Sementara itu, seseorang dengan nilai faktor genetik lebih dari 0,53 dikatakan mengidap Diabetes Mellitus (Diabetes UK, 2010).

3. Kadar Gula Darah

Seseorang terdiagnosis mengidap Diabetes Mellitus dapat dilihat melalui hasil pengukuran kadar gula darah sewatu (2 jam sesudah makan). Kategori kadar gula darah sewaktu disajikan pada Tabel 2.3 (Depkes, 2005).

Tabel 2.3 Kategori Kadar Gula Darah Sewaktu

Kategori	**Kadar Gula Darah Sewaktu (2 Jam Sesudah Makan)**
Normal	≤ 140 mg/dL
Pra-Diabetes Mellitus	141 – 186 mg/dL
Diabetes Mellitus	≥ 187 mg/dL

Berdasarkan Tabel 2.3, apabila hasil pengukuran kadar gula darah sewaktu $\geq$ 187 mg/dL menunjukkan potensi seseorang mengidap Diabetes Mellitus, hasil pengukuran menunjukkan kondisi Pra-Diabetes Mellitus yaitu kadar gula darah sewaktu antara 141 - 186 mg/dL. Sementara itu, kadar gula darah sewaktu dikatakan normal yaitu $\leq$ 140 mg/dL (Depkes, 2005).

4. Indeks Massa Tubuh (IMT)

Nilai dari Indeks Massa Tubuh (IMT) diperoleh dari pengukuran berat badan (BB) dalam bentuk satuan kilogram dan tinggi badan (TB) dalam satuan meter. Kemudian, dilakukan pengukuran menggunakan rumus hitung IMT:

$$IMT = \frac{BB(kg)}{(TB)^2(m)} \tag{2.2}$$

IMT dapat digunakan untuk mengetahui apakah seseorang mempunyai berat badan yang ideal atau belum. Untuk mengetahuinya, dapat dijelaskan pada Tabel

2.4.

Tabel 2.4 Kriteria Indeks Massa Tubuh (IMT)

Kategori	Indeks Massa Tubuh (IMT)
BB Normal	≤ 33
BB Tidak Normal	≥ 34

Berdasarkan Tabel 2.4, hasil pengukuran IMT yang masuk kategori tidak normal (obesitas) perlu untuk diwaspadai. Obesitas merupakan salah satu faktor risiko yang berperan penting terhadap penyakit Diabetes Mellitus. Seseorang dengan obesitas mempunyai masukan kalori yang berlebih. Sel beta pada kelenjar pankreas akan mengalami kelelahan sehingga tidak mampu untuk memproduksi insulin yang cukup untuk mengimbangi masukan kalori yang berlebih. Dengan demikian, kadar gula darah akan tinggi sehingga akan menjadi Diabetes Mellitus (Kaban & Sempakata, 2007).

5. Pola Makan

Seseorang dengan gaya hidup tidak baik, seperti pola makan tidak normal. Pola makan yang kurang baik misalnya mengkonsumsi alkohol, merokok dan konsumsi lemak berlebihan dapat memicu seseorang mengidap Diabetes Mellitus (Kemenkes, 2010).

Dengan pemanfaatan diagnosa serta data laboratorium, teknologi informasi dapat memberikan solusi untuk menyelesaikan permasalahan klasifikasi Diabetes Mellitus. Cabang ilmu data mining yang secara khusus berurusan mengenai informasi, penyimpanan, penarikan serta penggunaan data dapat memecahkan permasalahan dan juga pengambilan keputusan yang berasal dari dunia media (ADA, 2009).

2.3. Klasifikasi

Pada klasifikasi, terdapat target pada variabel dependen berupa kategorik. Sebagai contoh, pengklasifikasian diagnosa Diabetes Mellitus dapat dibagi dalam dua kategorik, yaitu positif Diabetes Mellitus dan negatif Diabetes Mellitus (Kusrini & Luthfi, 2009). Proses klasifikasi akan menggunakan data berupa kategorik yang bernilai diskrit/kontinu. Menurut Goronescu, dalam proses klasifikasi didasarkan pada empat komponen dasar yang sangat penting, komponen tersebut adalah sebagai berikut (Goronescu, 2011):

1. *Class*

   Variabel dependen dari model, merupakan variabel berupa kategorik yang merepresentasikan label terhadap objek setelah klasifikasi. Contohnya terdapat *class* pada klasifikasi penyakit Diabetes Mellitus, yaitu *class* positif Diabetes Mellitus dan *class* negatif Diabetes Mellitus (Goronescu, 2011).

2. Prediktor

   Variabel independen dari model, merupakan faktor-faktor yang mempengaruhi variabel dependen. Contohnya adalah data gejala dan hasil uji laboratorium mengenai faktor-faktor yang mempengaruhi Diabetes Mellitus, misalnya faktor genetik (Goronescu, 2011).

3. *Training* dataset

   Kumpulan data yang terdiri dari dua komponen di atas yang digunakan untuk melatih model dalam mengenali *class* yang sesuai berdasarkan variabel independen yang tersedia (Goronescu, 2011).

4. *Testing* dataset

   Kumpulan data baru yang akan diklasifikasikan menggunakan model yang

telah dibentuk sebelumnya. Dengan demikian, akan memperoleh hasil akurasi klasifikasi dan dapat dievaluasi (Goronescu, 2011).

2.4. Regresi Logistik Biner

Pada dasarnya analisis regresi merupakan suatu ilmu mengenai hubungan antara variabel dependen dengan satu atau lebih variabel independen, dengan maksud untuk memprediksi dan mengestimasi nilai-nilai variabel dependen berdasarkan nilai variabel independen yang telah diketahui (Ghozali & Imam, 2005).

Regresi Logistik Biner merupakan salah satu metode analisis data yang digunakan untuk mengetahui keterkaitan antara beberapa variabel independen (x) terhadap variabel dependen (y) yang mempunyai dua kategori (biner) dimana variabel dependen (y) bernilai salah dan benar (Hosmer & Lemeshow, 2000).

Metode yang cocok untuk mengklasifikasikan kejadian dengan variabel dependen (y) yang mempunyai dua kategori yaitu menggunakan metode regresi logistik, misalnya dalam mengklasifikasikan pasien apakah menderita penyakit Diabetes Mellitus atau tidak dengan memperhatikan beberapa faktor yang mempengaruhinya. *Output* dari variabel dependen (y) yang akan terbentuk dari pasien tersebut yaitu negatif Diabetes Mellitus atau positif Diabetes Mellitus (Hosmer & Lemeshow, 2000).

Apabila terdapat variabel independen sebanyak *k*, maka probabilitas untuk memperoleh ”NEGATIF” (y = 0) dapat dinyatakan dalam bentuk $P(Y = 0|x) = \pi(x)$. Sementara itu, probabilitas untuk memperoleh ”POSITIF” (y = 1) dapat dinyatakan dalam bentuk $P(Y = 1|x) = \pi(x)$. Variabel (x) merupakan variabel independen yang dapat bersifat kualitatif, misalkan $x = 0$ atau $x = 1$ dan seterusnya. Bentuk umum fungsi regresi logistik dengan variabel independen

sebanyak *i* dirumuskan dalam bentuk persamaan sebagai berikut (Rizki, dkk, 2015):

$$\pi(x_i) = \frac{e^{(\beta_0+\beta_1 x_{1j}+\beta_2 x_{2j}+\ldots+\beta_i x_{ij})}}{1+e^{(\beta_0+\beta_1 x_{1j}+\beta_2 x_{2j}+\ldots+\beta_i x_{ij})}} \tag{2.3}$$

dimana,

$\pi(x_i)$ = probabilitas suatu kejadian dengan nilai $0 \leq \pi(x_i) \leq 1$

β_0 = konstanta

x_{ij} = variabel independen ke-*i* (*i* = 1,2,3,...) pada data ke-*j*

β_i = koefisien dari variabel independen ke-*i* (*i* = 1,2,3,...)

Pada regresi logistik biner, variabel dependen (y) merupakan variabel dengan dua kategorik (biner) yaitu mempunyai nilai 0 atau 1. Regresi logistik biner mempunyai tujuan untuk menduga pola keterkaitan antara variabel *x* dengan $\pi(x_i)$. Dimana nilai $\pi(x_i)$ adalah nilai probabilitas suatu kejadian yang disebabkan oleh variabel *x* sehingga kemungkinan *output* yang diperoleh dari fungsi logistik bernilai 0 atau 1. Apabila $\pi(x_i)$ nilai harapan dan kategori satu terjadi maka $0 \leq \pi(x_i) \leq 1$. Dengan demikian, nilai transformasi logit dari nilai $\pi(x_i)$ diperoleh persamaan sebagai berikut (Zaen, 2019)

$$\pi(x_i) = \frac{e^{(\beta_0+\beta_1 x_{1j}+\beta_2 x_{2j}+\ldots+\beta_i x_{ij})}}{1+e^{(\beta_0+\beta_1 x_{1j}+\beta_2 x_{2j}+\ldots+\beta_i x_{ij})}}$$

$$e^{(\beta_0+\beta_1 x_{1j}+\beta_2 x_{2j}+\ldots+\beta_i x_{ij})} = \pi(x_i)(1+e^{(\beta_0+\beta_1 x_{1j}+\beta_2 x_{2j}+\ldots+\beta_i x_{ij})})$$

$$e^{(\beta_0+\beta_1 x_{1j}+\beta_2 x_{2j}+\ldots+\beta_i x_{ij})} = \pi(x_i)+\pi(x_i)e^{(\beta_0+\beta_1 x_{1j}+\beta_2 x_{2j}+\ldots+\beta_i x_{ij})}$$

$$\pi(x_i) = e^{(\beta_0+\beta_1 x_{1j}+\beta_2 x_{2j}+\ldots+\beta_i x_{ij})} - \pi(x_i)e^{(\beta_0+\beta_1 x_{1j}+\beta_2 x_{2j}+\ldots+\beta_i x_{ij})}$$

$$\pi(x_i) = (1-\pi(x_i))e^{(\beta_0+\beta_1 x_{1j}+\beta_2 x_{2j}+\ldots+\beta_i x_{ij})}$$

$$\frac{\pi(x_i)}{1-\pi(x_i)} = e^{(\beta_0+\beta_1 x_{1j}+\beta_2 x_{2j}+\ldots+\beta_i x_{ij})}$$

$$ln(\frac{\pi(x_i)}{1-\pi(x_i)}) = \beta_0+\beta_1 x_{1j}+\beta_2 x_{2j}+\ldots+\beta_i x_{ij}$$

$$g(x_i) = ln(\frac{\pi(x_i)}{1-\pi(x_i)}) = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + ... + \beta_i x_{ij} \quad (2.4)$$

$$= \beta_0 + \sum_{j=0}^{n} \beta_i x_{ij}$$
$$= X\beta$$

dengan

j = data ke-*n*

β_i = koefisien variabel *x*

x_i = variabel independen ke-*i* pada data ke-*j*

Apabila dinyatakan dalam bentuk matriks adalah sebagai berikut:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1j} \\ 1 & x_{21} & x_{22} & \cdots & x_{2j} \\ 1 & x_{31} & x_{32} & \cdots & x_{3j} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i1} & x_{i2} & \cdots & x_{ij} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_i \end{bmatrix}$$

Pengklasifikasian suatu objek dilakukan berdasarkan nilai probabilitas logistiknya. Pengklasifikasian tersebut berdasarkan kriteria sebagai berikut (Rumaendra, 2016)

a. Klasifikasikan objek pada *class* negatif Diabetes Mellitus (Y = 0), apabila nilai $\pi(x_i) \leq 0,5$.

b. Klasifikasikan objek pada *class* positif Diabetes Mellitus (Y = 1), apabila nilai $\pi(x_i) > 0,5$.

1. Estimasi Parameter

Estimasi parameter dilakukan menggunakan metode *Maximum Likelihood*

yang tujuannya untuk memaksimumkan fungsi *likelihood*. Pengujian parameter terhadap koefisien β secara univariabel terhadap variabel dependen (y) yaitu dengan membandingkan parameter hasil *Maximum Likelihood*, dugaan β dengan standar *error* dari parameter tersebut (Rizki, dkk, 2015).

Misalkan dalam percobaan dilakukan sebanyak N kali untuk setiap variabel independen sehingga diperoleh pengamatan sebanyak y_i dengan peluang "sukses" $\pi(x_i)$ dan peluang "gagal" $(1-\pi(x_i))$ dimana *i* = 1,2,3,...,n maka y_i berdistribusi binomial dengan fungsi kepadatan sebagai berikut (Hosmer & Lemeshow, 2000)

$$\pi(Y_i = y_i) = \pi(x_i)^{y_i}[1-\pi(x_i)]^{x_i - y_i}, \quad y_i = 0, 1 \tag{2.5}$$

Apabila x_i dan y_i merupakan pasangan variabel independen dan dependen pada pengamatan ke-*i* yang diasumsikan bahwa setiap pasangan dari suatu pengamatan saling bebas dengan pasangan pengamatan lainnya, yang artinya suatu kejadian pada pasangan data tersebut tidak mempengaruhi terjadinya kejadian pada pasangan data yang lain. Oleh karenanya, berlaku aturan perkalian tersebut maka dari Persamaan (2.5) fungsi *likelihood* merupakan gabungan dari fungsi distribusi masing-masing pasangan yaitu sebagai berikut (Zaen, 2019):

$$\ell(\beta; y_i) = \prod_{i=1}^{N} \pi(Y_i = y_i)$$

$$\ell(\beta; y_i) = \prod_{i=1}^{N} (\pi(x_i)^{y_i}(1-\pi(x_i))^{n_i - y_i} \tag{2.6}$$

$y_i = 0, 1$ *i* = 1,2,...,*n*

Dari Persamaan (2.6), maka diperoleh fungsi log *likelihood* sebagai berikut:

$$\begin{aligned}
\log \ell(\beta; y_i) &= \log \prod_{i=1}^{N} (\pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i} \\
&= \sum_{i=1}^{N} y_i \log \pi(x_i) + (1-y_i)\log(1-\pi(x_i)) \\
&= \sum_{i=1}^{N} y_i \log\left(\tfrac{e^{(g(x_i))}}{1+e^{(g(x_i))}}\right) + (1-y_i)\log\left(1-\tfrac{e^{(g(x_i))}}{1+e^{(g(x_i))}}\right) \\
&= \sum_{i=1}^{N} y_i[\log(e^{(g(x_i))}) - \log(1+e^{(g(x_i))})] + (1-y_i) \\
&\quad \log\left(\tfrac{1}{1+e^{(g(x_i))}}\right) \\
&= \sum_{i=1}^{N} y_i[\log(e^{(g(x_i))} - \log(1+e^{(g(x_i))})] + (1-y_i) \\
&\quad [\log 1 - \log(1+e^{(g(x_i))})] \\
&= \sum_{i=1}^{N} y_i[\log(e^{(g(x_i))} - \log(1+e^{(g(x_i))})] - (1-y_i) \\
&\quad \log(1+e^{(g(x_i))}) \\
&= \sum_{i=1}^{N} [y_i \log(e^{(g(x_i))} - y_i \log(1+e^{(g(x_i))}) - \\
&\quad \log(1+e^{(g(x_i))}) + y_i \log(1+e^{(g(x_i))})] \\
&= \sum_{i=1}^{N} [y_i \log(e^{(g(x_i))}) - \log(1+e^{(g(x_i))})]
\end{aligned}$$

$$\log \ell(\beta; y_i) = \sum_{i=1}^{N} (y_i g(x_i)) - \log(1+e^{(g(x_i))}) \tag{2.7}$$

Karena $g(x_i) = ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = x_i\beta$, maka:

$$\begin{aligned}
\ell(\beta) &= \sum_{i=1}^{N} (y_i x_i \beta - \log(1+e^{(x_i\beta)}) \\
&= \sum_{i=1}^{N} y_i x_i \beta - \sum_{i=1}^{N} \log(1+e^{(x_i\beta)} \\
&= \sum_{i=1}^{N} y_i \left(\sum_{j=0}^{n} \beta_i x_{ij}\right) - \sum_{i=1}^{N} \log(1+e^{\left(\sum_{j=0}^{n} \beta_i x_{ij}\right)})
\end{aligned}$$

Kemudian Persamaan (2.7) diturunkan terhadap $\beta_0, \beta_1, \beta_2, ..., \beta_n$ dan hasil dari penurunan tersebut disamadengankan nol. Dengan demikian diperoleh:

a) Turunan fungsi log *likelihood* terhadap β_0

$$\frac{\partial \ell(\beta)}{\partial \beta_0} = 0$$

$$\frac{\partial(\sum_{i=1} y_i\left(\sum_{j=0}^{n} \beta_i x_{ij}\right) - \sum_{i=1}^{N} \log(1+e^{\left(\sum_{j=0}^{n} \beta_i x_{ij}\right)})}{\partial \beta_0} = 0$$

$$\frac{\partial(\sum_{i=1}^{N} y_i \beta_0 x_{i0} - \sum_{i=1}^{N} \log(1+e^{\left(\sum_{j=0}^{n} \beta_i x_{ij}\right)})}{\partial \beta_0} = 0$$

$$\sum_{i=1}^{N} y_i x_{i0} - \sum_{i=1}^{N} x_{i0} \frac{e^{\left(\sum_{j=0}^{n} \beta_i x_{ij}\right)}}{1+e^{\left(\sum_{j=0}^{n} \beta_i x_{ij}\right)}} = 0$$

$$\sum_{i=1}^{N} y_i x_{i0} - \sum_{i=1}^{N} \pi(x_i) = 0$$

b) Turunan fungsi log *likelihood* terhadap β_1

$$\frac{\partial \ell(\beta)}{\partial \beta_1} = 0$$

$$\frac{\partial(\sum_{i=1}^{N} y_i\left(\sum_{j=0}^{n} \beta_i x_{ij}\right) - \sum_{i=1}^{N} \log(1+e^{\left(\sum_{j=0}^{n} \beta_i x_{ij}\right)})}{\partial \beta_1} = 0$$

$$\frac{\partial(\sum_{i=1}^{N} y_i \beta_1 x_{i1} - \sum_{i=1}^{N} \log(1+e^{\left(\sum_{j=0}^{n} \beta_i x_{ij}\right)})}{\partial \beta_1} = 0$$

$$\sum_{i=1}^{N} y_i x_{i1} - \sum_{i=1}^{N} x_{i1} \frac{e^{\left(\sum_{j=0}^{n} \beta_i x_{ij}\right)}}{1+e^{\left(\sum_{j=0}^{n} \beta_i x_{ij}\right)}} = 0$$

$$\sum_{i=1}^{N} y_i x_{i1} - \sum_{i=1}^{N} x_{i1} \pi(x_i) = 0$$

c) Turunan fungsi log *likelihood* terhadap β_2

$$\frac{\partial \ell(\beta)}{\partial \beta_2} = 0$$

$$\frac{\partial(\sum_{i=1}^{N} y_i\left(\sum_{j=0}^{n} \beta_i x_{ij}\right) - \sum_{i=1}^{N} \log(1+e^{\left(\sum_{j=0}^{n} \beta_i x_{ij}\right)})}{\partial \beta_2} = 0$$

$$\frac{\partial(\sum_{i=1}^{N} y_i \beta_2 x_{i2} - \sum_{i=1}^{N} \log(1+e^{\left(\sum_{j=0}^{n} \beta_i x_{ij}\right)})}{\partial \beta_2} = 0$$

$$\sum_{i=1}^{N} y_i x_{i2} - \sum_{i=1}^{N} x_{i2} \frac{e^{\left(\sum_{j=0}^{n} \beta_i x_{ij}\right)}}{1+e^{\left(\sum_{j=0}^{n} \beta_i x_{ij}\right)}} = 0$$

$$\sum_{i=1}^{N} y_i x_{i2} - \sum_{i=1}^{N} x_{i2} \pi(x_i) = 0$$

d) Turunan fungsi log *likelihood* terhadap β_n

$$\frac{\partial \ell(\beta)}{\partial \beta_n} = 0$$

$$\frac{\partial(\sum_{i=1}^{N} y_i\left(\sum_{j=0}^{n}\beta_i x_{ij}\right)-\sum_{i=1}^{N}\log(1+e^{\left(\sum_{j=0}^{n}\beta_i x_{ij}\right)})}{\partial\beta_n}=0$$

$$\frac{\partial(\sum_{i=1}^{N} y_i\beta_n x_{in}-\sum_{i=1}^{N}\log(1+e^{\left(\sum_{j=0}^{n}\beta_i x_{ij}\right)})}{\partial\beta_n}=0$$

$$\sum_{i=1}^{N} y_i x_{in}-\sum_{i=1}^{N} x_{in}\frac{e^{\left(\sum_{j=0}^{n}\beta_i x_{ij}\right)}}{1+e^{\left(\sum_{j=0}^{n}\beta_i x_{ij}\right)}}=0$$

$$\sum_{i=1}^{N} y_i x_{in}-\sum_{i=1}^{N} x_{in}\pi(x_i)=0$$

Dengan demikian, diperoleh persamaan umumnya sebagai berikut:

$$\begin{aligned}\frac{\partial\ell(\beta)}{\partial\beta_n} &= \textstyle\sum_{i=1} y_i x_{ia}-\sum_{i=1} x_{ia}\pi(x_i)\\ &= \sum_{i=1}^{N} x_{ia}\left[y_i-\pi(x_i)\right]\end{aligned} \tag{2.8}$$

dengan $\pi(x_i)=\frac{e^{\sum_j \beta_i x_{ij}}}{1+e^{\sum_j \beta_i x_{ij}}}$, $\quad j=0,1,2,...,n, \quad a=0,1,2,...,k.$

Dalam bentuk matriks diperoleh sebagai berikut:

$$X^t(Y-\pi(x_i)) = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{a1} & x_{a2} & \cdots & x_{aN} \end{bmatrix} \cdot \begin{bmatrix} y_1-\pi(x_1) \\ y_2-\pi(x_2) \\ y_3-\pi(x_3) \\ \vdots \\ y_N-\pi(x_N) \end{bmatrix}$$

dimana,

t = Indeks iterasi

Tahap selanjutnya adalah menentukan turunan kedua dari fungsi log *likelihood* terhadap $\beta_0, \beta_1, \beta_2, ..., \beta_n$, sehingga diperoleh persamaan sebagai berikut:

a) Turunan kedua fungsi log *likelihood* terhadap β_0

Turunan kedua dari $\ell(\beta)$ terhadap $\beta_0\beta_0$, adalah sebagai berikut:

$$\begin{aligned}
\frac{\partial^2 \ell(\beta)}{(\partial\beta_0)^2} &= \frac{\partial^2}{(\partial\beta_0)^2}\left(\sum_{i=1}^{N} y_i\left(\sum_{j=0}^{n}\beta_i x_{ij}\right) - \sum_{i=1}^{N}\log(1+e^{\left(\sum_{j=0}^{n}\beta_i x_{ij}\right)})\right)\\
&= \frac{\partial}{\partial\beta_0}\sum_{i=1}^{N} x_{0i}\left[y_i - \pi(x_i)\right]\\
&= \frac{\partial}{\partial\beta_0}\sum_{i=1}^{N} x_{0i}\left[y_i - \frac{e^{\sum_j \beta_i x_{ij}}}{1+e^{\sum_j \beta_i x_{ij}}}\right]\\
&= -\sum_{i=1}^{N} x_{0i}x_{0i}\left[\frac{e^{\sum_j \beta_i x_{ij}}(1+e^{\sum_j \beta_i x_{ij}})-(e^{\sum_j \beta_i x_{ij}})^2}{(1+e^{\sum_j \beta_i x_{ij}})^2}\right]\\
&= -\sum_{i=1}^{N} x_{0i}x_{0i}\left[\frac{e^{\sum_j \beta_i x_{ij}}}{(1+e^{\sum_j \beta_i x_{ij}})^2} - \left(\frac{e^{\sum_j \beta_i x_{ij}}}{(1+e^{\sum_j \beta_i x_{ij}})}\right)^2\right]\\
&= -\sum_{i=1}^{N} x_{0i}x_{0i}[\pi(x_i) - (\pi(x_i))^2]\\
\frac{\partial^2 \ell(\beta)}{(\partial\beta_0)^2} &= -\sum_{i=1}^{N} x_{0i}x_{0i}[\pi(x_i)(1-\pi(x_i))]
\end{aligned}$$

Kemudian turunan kedua dari $\ell(\beta)$ terhadap $\beta_1\beta_0$, adalah sebagai berikut:

$$\begin{aligned}
\frac{\partial^2 \ell(\beta)}{\partial\beta_1\partial\beta_0} &= \frac{\partial^2}{(\partial\beta_1\beta_0}\left(\sum_{i=1}^{N} y_i\left(\sum_{j=0}^{n}\beta_i x_{ij}\right) - \sum_{i=1}^{N}\log(1+e^{\left(\sum_{j=0}^{n}\beta_i x_{ij}\right)})\right)\\
&= \frac{\partial}{\partial\beta_0}\sum_{i=1}^{N} x_{0i}\left[y_i - \pi(x_i)\right]\\
&= \frac{\partial}{\partial\beta_1}\sum_{i=1}^{N} x_{0i}\left[y_i - \frac{e^{\sum_j \beta_i x_{ij}}}{1+e^{\sum_j \beta_i x_{ij}}}\right]\\
&= -\sum_{i=1}^{N} x_0 x_{1i}\left[\frac{e^{\sum_j \beta_i x_{ij}}(1+e^{\sum_j \beta_i x_{ij}})-(e^{\sum_j \beta_i x_{ij}})^2}{(1+e^{\sum_j \beta_i x_{ij}})^2}\right]\\
&= -\sum_{i=1}^{N} x_0 x_{1i}\left[\frac{e^{\sum_j \beta_i x_{ij}}}{(1+e^{\sum_j \beta_i x_{ij}})^2} - \left(\frac{e^{\sum_j \beta_i x_{ij}}}{(1+e^{\sum_j \beta_i x_{ij}})}\right)^2\right]\\
&= -\sum_{i=1}^{N} x_0 x_{1i}[\pi(x_i) - (\pi(x_i))^2]\\
\frac{\partial^2 \ell(\beta)}{\partial\beta_1\partial\beta_0} &= -\sum_{i=1}^{N} x_0 x_{1i}[\pi(x_i)(1-\pi(x_i))]\\
&\vdots
\end{aligned}$$

Kemudian turunan kedua dari $\ell(\beta)$ terhadap $\beta_n\beta_0$, adalah sebagai berikut:

$$\begin{aligned}
\frac{\partial^2 \ell(\beta)}{\partial\beta_n\partial\beta_0} &= \frac{\partial^2}{\partial\beta_n\beta_0}\left(\sum_{i=1}^{N} y_i\left(\sum_{j=0}^{n}\beta_i x_{ij}\right) - \sum_{i=1}^{N}\log(1+e^{\left(\sum_{j=0}^{n}\beta_i x_{ij}\right)})\right)\\
&= \frac{\partial}{\partial\beta_0}\sum_{i=1}^{N} x_{0i}\left[y_i - \pi(x_i)\right]\\
&= \frac{\partial}{\partial\beta_n}\sum_{i=1}^{N} x_{0i}\left[y_i - \frac{e^{\sum_j \beta_i x_{ij}}}{1+e^{\sum_j \beta_i x_{ij}}}\right]
\end{aligned}$$

$$
\begin{aligned}
&= -\sum_{i=1}^{N} x_{0i}x_{ni}\left[\frac{e^{\sum_j \beta_i x_{ij}}(1+e^{\sum_j \beta_i x_{ij}})-(e^{\sum_j \beta_i x_{ij}})^2}{(1+e^{\sum_j \beta_i x_{ij}})^2}\right] \\
&= -\sum_{i=1}^{N} x_{0i}x_{ni}\left[\frac{e^{\sum_j \beta_i x_{ij}}}{(1+e^{\sum_j \beta_i x_{ij}})^2} - \left(\frac{e^{\sum_j \beta_i x_{ij}}}{(1+e^{\sum_j \beta_i x_{ij}})}\right)^2\right] \\
&= -\sum_{i=1}^{N} x_{0i}x_{ni}[\pi(x_i) - (\pi(x_i))^2] \\
\frac{\partial^2 \ell(\beta)}{\partial \beta_n \partial \beta_0} &= -\sum_{i=1}^{N} x_{0i}x_{ni}[\pi(x_i)(1-\pi(x_i))]
\end{aligned}
$$

b) Turunan kedua fungsi log *likelihood* terhadap β_1

Turunan kedua dari $\ell(\beta)$ terhadap $\beta_0\beta_1$, adalah sebagai berikut:

$$
\begin{aligned}
\frac{\partial^2 \ell(\beta)}{\partial \beta_0 \partial \beta_1} &= \frac{\partial^2}{\partial \beta_0 \partial \beta_1}\left(\sum_{i=1}^{N} y_i\left(\sum_{j=0}^{n}\beta_i x_{ij}\right) - \sum_{i=1}^{N}\log(1+e^{\left(\sum_{j=0}^{n}\beta_i x_{ij}\right)})\right) \\
&= \frac{\partial}{\partial \beta_0}\sum_{i=1}^{N} x_{1i}\left[y_i - \pi(x_i)\right] \\
&= \frac{\partial}{\partial \beta_0}\sum_{i=1}^{N} x_{1i}\left[y_i - \sum_{i=1}^{N}\pi(x_i)\right] \\
&= -\sum_{i=1}^{N} x_{1i}x_{0i}\left[\frac{e^{\sum_j \beta_i x_{ij}}(1+e^{\sum_j \beta_i x_{ij}})-(e^{\sum_j \beta_i x_{ij}})^2}{(1+e^{\sum_j \beta_i x_{ij}})^2}\right] \\
&= -\sum_{i=1}^{N} x_{1i}x_{0i}\left[\frac{e^{\sum_j \beta_i x_{ij}}}{(1+e^{\sum_j \beta_i x_{ij}})^2} - \left(\frac{e^{\sum_j \beta_i x_{ij}}}{(1+e^{\sum_j \beta_i x_{ij}})}\right)^2\right] \\
&= -\sum_{i=1}^{N} x_{1i}x_{0i}[\pi(x_i) - (\pi(x_i))^2] \\
\frac{\partial^2 \ell(\beta)}{\partial \beta_0 \partial \beta_1} &= -\sum_{i=1}^{N} x_{1i}x_{0i}[\pi(x_i)(1-\pi(x_i))]
\end{aligned}
$$

Kemudian turunan kedua dari $\ell(\beta)$ terhadap $\beta_1\beta_1$, adalah sebagai berikut:

$$
\begin{aligned}
\frac{\partial^2 \ell(\beta)}{(\partial \beta_1)^2} &= \frac{\partial^2}{(\partial \beta_1)^2}\left(\sum_{i=1}^{N} y_i\left(\sum_{j=0}^{n}\beta_i x_{ij}\right) - \sum_{i=1}^{N}\log(1+e^{\left(\sum_{j=0}^{n}\beta_i x_{ij}\right)})\right) \\
&= \frac{\partial}{\partial \beta_1}\sum_{i=1}^{N} x_{1i}\left[y_i - \pi(x_i)\right] \\
&= \frac{\partial}{\partial \beta_1}\sum_{i=1}^{N} x_{1i}\left[y_i - \frac{e^{\sum_j \beta_i x_{ij}}}{1+e^{\sum_j \beta_i x_{ij}}}\right] \\
&= -\sum_{i=1}^{N} x_{1i}x_{1i}\left[\frac{e^{\sum_j \beta_i x_{ij}}(1+e^{\sum_j \beta_i x_{ij}})-(e^{\sum_j \beta_i x_{ij}})^2}{(1+e^{\sum_j \beta_i x_{ij}})^2}\right] \\
&= -\sum_{i=1}^{N} x_{1i}x_{1i}\left[\frac{e^{\sum_j \beta_i x_{ij}}}{(1+e^{\sum_j \beta_i x_{ij}})^2} - \left(\frac{e^{\sum_j \beta_i x_{ij}}}{(1+e^{\sum_j \beta_i x_{ij}})}\right)^2\right] \\
&= -\sum_{i=1}^{N} x_{1i}x_{1i}[\pi(x_i) - (\pi(x_i))^2] \\
\frac{\partial^2 \ell(\beta)}{(\partial \beta_1)^2} &= -\sum_{i=1}^{N} x_{1i}x_{1i}[\pi(x_i)(1-\pi(x_i))]
\end{aligned}
$$

$$\vdots$$

Kemudian turunan kedua dari $\ell(\beta)$ terhadap $\beta_n\beta_1$, adalah sebagai berikut:

$$\begin{aligned}
\frac{\partial^2 \ell(\beta)}{\partial \beta_n \partial \beta_1} &= \frac{\partial^2}{\partial \beta_n \partial \beta_1} \left(\sum_{i=1}^{N} y_i \left(\sum_{j=0}^{n} \beta_i x_{ij} \right) - \sum_{i=1}^{N} \log(1 + e^{\left(\sum_{j=0}^{n} \beta_i x_{ij} \right)}) \right) \\
&= \frac{\partial}{\partial \beta_1} \sum_{i=1}^{N} x_{1i} \left[y_i - \pi(x_i) \right] \\
&= \frac{\partial}{\partial \beta_n} \sum_{i=1}^{N} x_{1i} \left[y_i - \frac{e^{\sum_j \beta_i x_{ij}}}{1+e^{\sum_j \beta_i x_{ij}}} \right] \\
&= -\sum_{i=1}^{N} x_{1i} x_{ni} \left[\frac{e^{\sum_j \beta_i x_{ij}} (1+e^{\sum_j \beta_i x_{ij}}) - (e^{\sum_j \beta_i x_{ij}})^2}{(1+e^{\sum_j \beta_i x_{ij}})^2} \right] \\
&= -\sum_{i=1}^{N} x_{1i} x_{ni} \left[\frac{e^{\sum_j \beta_i x_{ij}}}{(1+e^{\sum_j \beta_i x_{ij}})^2} - \left(\frac{e^{\sum_j \beta_i x_{ij}}}{(1+e^{\sum_j \beta_i x_{ij}})} \right)^2 \right] \\
&= -\sum_{i=1}^{N} x_{1i} x_{ni} [\pi(x_i) - (\pi(x_i))^2] \\
\frac{\partial^2 \ell(\beta)}{\partial \beta_n \partial \beta_1} &= -\sum_{i=1}^{N} x_{1i} x_{ni} [\pi(x_i)(1 - \pi(x_i))]
\end{aligned}$$

c) Turunan kedua fungsi log *likelihood* terhadap β_2

Turunan kedua dari $\ell(\beta)$ terhadap $\beta_0\beta_2$, adalah sebagai berikut:

$$\begin{aligned}
\frac{\partial^2 \ell(\beta)}{\partial \beta_0 \partial \beta_2} &= \frac{\partial^2}{\partial \beta_0 \partial \beta_2} \left(\sum_{i=1}^{N} y_i \left(\sum_{j=0}^{n} \beta_i x_{ij} \right) - \sum_{i=1}^{N} \log(1 + e^{\left(\sum_{j=0}^{n} \beta_i x_{ij} \right)}) \right) \\
&= \frac{\partial}{\partial \beta_0} \sum_{i=1}^{N} x_{2i} \left[y_i - \pi(x_i) \right] \\
&= \frac{\partial}{\partial \beta_0} \sum_{i=1}^{N} x_{2i} \left[y_i - \frac{e^{\sum_j \beta_i x_{ij}}}{1+e^{\sum_j \beta_i x_{ij}}} \right] \\
&= -\sum_{i=1}^{N} x_{2i} x_{0i} \left[\frac{e^{\sum_j \beta_i x_{ij}} (1+e^{\sum_j \beta_i x_{ij}}) - (e^{\sum_j \beta_i x_{ij}})^2}{(1+e^{\sum_j \beta_i x_{ij}})^2} \right] \\
&= -\sum_{i=1}^{N} x_{2i} x_{0i} \left[\frac{e^{\sum_j \beta_i x_{ij}}}{(1+e^{\sum_j \beta_i x_{ij}})^2} - \left(\frac{e^{\sum_j \beta_i x_{ij}}}{(1+e^{\sum_j \beta_i x_{ij}})} \right)^2 \right] \\
&= -\sum_{i=1}^{N} x_{2i} x_{0i} [\pi(x_i) - (\pi(x_i))^2] \\
\frac{\partial^2 \ell(\beta)}{\partial \beta_0 \partial \beta_2} &= -\sum_{i=1}^{N} x_{2i} x_{0i} [\pi(x_i)(1 - \pi(x_i))]
\end{aligned}$$

Kemudian turunan kedua dari $\ell(\beta)$ terhadap $\beta_1\beta_2$, adalah sebagai berikut:

$$\begin{aligned}
\frac{\partial^2 \ell(\beta)}{\partial \beta_1 \beta_2} &= \frac{\partial^2}{\partial \beta_1 \beta_2} \left(\sum_{i=1}^{N} y_i \left(\sum_{j=0}^{n} \beta_i x_{ij} \right) - \sum_{i=1}^{N} \log(1 + e^{\left(\sum_{j=0}^{n} \beta_i x_{ij} \right)}) \right) \\
&= \frac{\partial}{\partial \beta_1} \sum_{i=1}^{N} x_{2i} \left[y_i - \pi(x_i) \right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\partial}{\partial \beta_1} \sum_{i=1}^{N} x_{2i} \left[y_i - \frac{e^{\sum_j \beta_i x_{ij}}}{1+e^{\sum_j \beta_i x_{ij}}} \right] \\
&= -\sum_{i=1}^{N} x_{2i} x_{1i} \left[\frac{e^{\sum_j \beta_i x_{ij}} (1+e^{\sum_j \beta_i x_{ij}}) - (e^{\sum_j \beta_i x_{ij}})^2}{(1+e^{\sum_j \beta_i x_{ij}})^2} \right] \\
&= -\sum_{i=1}^{N} x_{2i} x_{1i} \left[\frac{e^{\sum_j \beta_i x_{ij}}}{(1+e^{\sum_j \beta_i x_{ij}})^2} - \left(\frac{e^{\sum_j \beta_i x_{ij}}}{(1+e^{\sum_j \beta_i x_{ij}})} \right)^2 \right] \\
&= -\sum_{i=1}^{N} x_{2i} x_{1i} [\pi(x_i) - (\pi(x_i))^2] \\
\frac{\partial^2 \ell(\beta)}{\partial \beta_1 \beta_2} &= -\sum_{i=1}^{N} x_{2i} x_{1i} [\pi(x_i)(1 - \pi(x_i))] \\
&\vdots
\end{aligned}
$$

Kemudian turunan kedua dari $\ell(\beta)$ terhadap $\beta_n \beta_2$, adalah sebagai berikut:

$$
\begin{aligned}
\frac{\partial^2 \ell(\beta)}{\partial \beta_n \partial \beta_2} &= \frac{\partial^2}{\partial \beta_n \partial \beta_2} \left(\sum_{i=1}^{N} y_i \left(\sum_{j=0}^{n} \beta_i x_{ij} \right) - \sum_{i=1}^{N} \log(1 + e^{\left(\sum_{j=0}^{n} \beta_i x_{ij} \right)}) \right) \\
&= \frac{\partial}{\partial \beta_1} \sum_{i=1}^{N} x_{2i} \left[y_i - \pi(x_i) \right] \\
&= \frac{\partial}{\partial \beta_n} \sum_{i=1}^{N} x_{2i} \left[y_i - \frac{e^{\sum_j \beta_i x_{ij}}}{1+e^{\sum_j \beta_i x_{ij}}} \right] \\
&= -\sum_{i=1}^{N} x_{2i} x_{ni} \left[\frac{e^{\sum_j \beta_i x_{ij}} (1+e^{\sum_j \beta_i x_{ij}}) - (e^{\sum_j \beta_i x_{ij}})^2}{(1+e^{\sum_j \beta_i x_{ij}})^2} \right] \\
&= -\sum_{i=1}^{N} x_{2i} x_{ni} \left[\frac{e^{\sum_j \beta_i x_{ij}}}{(1+e^{\sum_j \beta_i x_{ij}})^2} - \left(\frac{e^{\sum_j \beta_i x_{ij}}}{(1+e^{\sum_j \beta_i x_{ij}})} \right)^2 \right] \\
&= -\sum_{i=1}^{N} x_{2i} x_{ni} [\pi(x_i) - (\pi(x_i))^2] \\
\frac{\partial^2 \ell(\beta)}{\partial \beta_n \partial \beta_2} &= -\sum_{i=1}^{N} x_{2i} x_{ni} [\pi(x_i)(1 - \pi(x_i))]
\end{aligned}
$$

Misalkan turunan pertama dari $\ell(\beta)$ terhadap β_a, $a \leq n$ adalah sebagai berikut:

$$
\begin{aligned}
\frac{\partial \ell(\beta)}{\partial \beta_a} &= \frac{\partial}{\partial \beta_a} \left(\sum_{i=1}^{N} y_i \left(\sum_{a=0}^{n} \beta_i x_{ai} \right) - \sum_{i=1}^{N} \log(1 + e^{\left(\sum_{a=0}^{n} \beta_i x_{ai} \right)}) \right) \\
&= \sum_{i=1}^{N} \left[y_i x_{ai} - \frac{e^{\sum_a \beta_i x_{ai}}}{1+e^{\sum_a \beta_i x_{ai}}} \cdot x_{ai} \right] \\
\frac{\partial \ell(\beta)}{\partial \beta_a} &= \sum_{i=1}^{N} x_{ai} [y_i - \pi(x_i)]
\end{aligned}
$$

Maka, turunan kedua dari $\ell(\beta)$ terhadap β_b, $b \leq n$ adalah sebagai berikut:

$$
\frac{\partial^2 \ell(\beta)}{\partial \beta_b \partial \beta_a} = \frac{\partial}{\partial \beta_b} \sum_{i=1}^{N} x_{ai} [y_i - \pi(x_i)]
$$

$$= -\sum_{i=1}^{N}\left[\frac{x_{ai}x_{bi}e^{\sum_j \beta_i x_{ij}}(1+e^{\sum_j \beta_i x_{ij}})-(x_{ai}x_{bi}e^{\sum_j \beta_i x_{ij}})^2}{(1+e^{\sum_j \beta_i x_{ij}})^2}\right]$$

Dengan demikian, diperoleh persamaan umumnya sebagai berikut:

$$\frac{\partial^2 \ell(\beta)}{\partial \beta_b \partial \beta_a} = -\sum_{i=1}^{N} x_{ai}x_{bi}[\pi(x_i)(1-\pi(x_i))], \quad a,b = 0,1,2,...,n \tag{2.9}$$

Apabila dinyatakan dalam bentuk matriks adalah sebagai berikut:

$$X^t V X = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & x_{13} & \cdots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & x_{k3} & \cdots & x_{kn} \end{bmatrix} \cdot$$

$$\begin{bmatrix} \pi(x_1)[1-\pi(x_i)] & 0 & 0 & \cdots & 0 \\ 0 & \pi(x_2)[1-\pi(x_i)] & 0 & \cdots & 0 \\ 0 & 0 & \pi(x_3)[1-\pi(x_i)] & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \pi(x_n)[1-\pi(x_i)] \end{bmatrix} \cdot$$

$$\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & \cdots & x_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

Untuk estimasi parameter $\hat{\beta}$ dalam penelitian ini, metode numerik yang digunakan adalah metode *Newton Raphson*. *Newton Raphson* merupakan suatu metode iterasi numerik yang dapat digunakan untuk mengestimasi hampiran akar-akar sistem persamaan linier dan non linier (Zaen, 2019).

Algoritma iterasi menggunakan metode *Newton Raphson* untuk memperoleh nilai $\hat{\beta}$ dapat dilakukan melalui tahap-tahap sebagai berikut (Rizki, dkk, 2015):

1) Memasukkan nilai dugaan awal untuk β misal $\hat{\beta} = 0$.
2) Menghitung nilai matriks dari $X^t(Y - \pi(x_i)$ dan X^tVX, kemudian hitung nilai invers dari matriks X^tVX.
3) Untuk setiap $(i + 1)$ dihitung dugaan baru, yaitu $\hat{\beta}_{i+1} = \hat{\beta}_i +$ $[X^tVX]^{-1}$ $[X^t(Y - \pi(x_i)]$.
4) Iterasi berakhir, apabila diperoleh $\hat{\beta}_{i+1} \cong \hat{\beta}_i$.

2. Uji G

Statistik uji G merupakan uji rasio kemungkinan maksimum (*maximum likelihood ratio test*) yang digunakan untuk menguji pengaruh variabel independen di dalam model secara serentak (Hosmer & Lemeshow, 2000). Persamaannya adalah sebagai berikut:

$$G = -2\, ln \frac{\left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{\pi} \pi_i^{y_i}(1-\pi_i)^{1-y_1}} \tag{2.10}$$

dimana:

n_0 = banyaknya pengamatan yang mempunyai nilai $Y = 0$

n_1 = banyaknya pengamatan yang mempunyai nilai $Y = 1$

n = banyaknya total pengamatan

dengan nilai $n = n_1 + n_0$, $n_1 = \sum_{i=1}^{\pi} y_i$ dan $n_0 = \sum_{i=1}^{\pi}(1 - y_i)$

Tahap-tahap pengujian menggunakan uji G adalah sebagai berikut:

a) Rumusan Hipotesis

$H_0 : \beta_1 = \beta_2 = ... = \beta_p$

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

b) Besaran yang diperlukan

Menghitung nilai n, n_0 dan n_1

c) Statistik Uji

Menghitung nilai G

d) Kriteria Pengujian

Untuk memperoleh suatu keputusan yang optimal maka perlu membandingkan nilai *Chi-Square* dengan nilai X^2 tabel dimana $db = k - 1$ dan k merupakan banyaknya variabel independen. Apabila nilai uji G lebih besar dari nilai *Chi-Square* dengan banyaknya variabel dalam model atau dapat ditulis dengan $G > X^2{}_{(db,\alpha)}$ atau $(Sig.) < \alpha$ sebesar 0,05, maka H_0 ditolak (Hosmer & Lemeshow, 2000).

e) Kesimpulan

Menarik kesimpulan dari hipotesis H_0 diterima atau ditolak. Apabila H_0 ditolak, maka kesimpulannya variabel independen secara serentak mempengaruhi variabel dependen. Sementara itu, apabila H_0 diterima, maka kesimpulannya variabel independen secara serentak tidak mempengaruhi variabel dependen. Setelah melakukan Uji G, langkah selanjutnya adalah melakukan Uji Wald (Hosmer & Lemeshow, 2000).

3. Uji Wald

Uji Wald merupakan teknik pengujian yang digunakan dalam uji parsial. Uji tersebut bertujuan untuk mengetahui apakah setiap variabel independen berpengaruh terhadap model atau tidak. Uji Wald dapat diperhitungkan dengan cara membandingkan parameter yang ditaksir dengan galat baku dari parameter

tersebut (Hosmer & Lemeshow, 2000), hal ini dapat dirumuskan sebagai berikut:

$$W = \left(\frac{\beta_j}{SE(\beta_j)} \right) \tag{2.11}$$

dimana:

W = Uji Wald

β_j = estimasi parameter

SE = standar *error*

Untuk menghitung nilai standar *error* dapat menggunakan persamaan sebagai berikut:

$$SE = \frac{s}{\sqrt{n}} \tag{2.12}$$

dimana:

s = standar deviasi populasi

n = jumlah sampel

Tahap-tahap pengujian menggunakan Uji Wald adalah sebagai berikut:

a) Rumusan Hipotesis

$H_0 : \beta_j = 0, \quad j = 1, 2, ..., p$

$H_1 : \beta_j \neq 0, \quad j = 1, 2, ..., p$

b) Besaran yang diperlukan

Menghitung nilai β_j dan SE

c) Statistik Uji

Menghitung nilai W

d) Kriteria Pengujian

Apabila nilai $|W| > Z_{\alpha/2}$ atau $(Sig.) < \alpha$, maka H_0 ditolak

e) Kesimpulan

Menarik kesimpulan dari hipotesis H_0 diterima atau ditolak. Apabila H_0 ditolak, maka kesimpulannya variabel independen secara individu mempengaruhi variabel dependen. Sementara itu, apabila H_0 diterima, maka kesimpulannya variabel independen secara individu tidak mempengaruhi variabel dependen (Hosmer & Lemeshow, 2000).

4. Nilai P (P-*Value*) atau (Sig.)

Nilai P-*Value* atau yang sering disebut (Sig.) merupakan nilai kesalahan yang diperoleh peneliti dari hasil perhitungan statistik (Hasil Uji Statistik). Sementara itu, nilai α (*alpha*) merupakan kesalahan maksimal yang telah ditetapkan oleh peneliti. Nilai P-*Value* digunakan untuk mengambil keputusan (tolak/terima H_0) dari suatu hipotesis yaitu dengan cara membandingkan antara nilai P-*Value* atau (Sig.) dengan taraf α (*alpha*) tertentu. Dalam penelitian ini, nilai taraf α (*alpha*) yang digunakan sebesar 5% atau 0,05 hal ini mempunyai arti bahwa paling besar 5 kesalahan yang dapat ditolerir oleh peneliti. Untuk kriteria pengambilan keputusan yang digunakan adalah sebagai berikut (Zaen, 2019):

a) Apabila nilai P-*Value* atau (Sig.) $>$ α (0,05), maka H_0 diterima.

b) Apabila nilai P-*Value* atau (Sig.) $<$ α (0,05), maka H_0 ditolak.

2.5. *Classification And Regression Trees* (CART)

CART merupakan salah satu algoritma atau metode dari salah satu teknik eksplorasi data, yaitu teknik pohon keputusan. Metode CART dikembangkan sejak tahun 1980-an oleh ilmuwan yang bernama Leo Breiman, Jerome H. Friedman, Richard A. Olshen dan Charles J. Stone (Breiman dkk., 1984).

CART merupakan salah satu metode statistika nonparametrik untuk

menggambarkan keterkaitan antara variabel dependen dengan satu atau lebih variabel independen. Metode CART dikembangkan untuk analisis klasifikasi, dimana variabel dependen berupa kategorik maupun kontinu. Apabila varaibel dependen berskala kategorik, maka CART akan menghasilkan pohon klasifikasi (*classification trees*) dan apabila variabel dependen berskala kontinu, maka CART akan menghasilkan pohon regresi (*regression trees*). Variabel dependen dalam penelitian ini berskala kategorik, sehingga metode yang akan digunakan adalah metode pohon klasifikasi (Breiman dkk., 1993).

Tujuan utama metode CART adalah untuk memperoleh suatu kelompok data yang akurat sebagai identitas dari suatu pengklasifikasian. Metode CART mempunyai beberapa kelebihan dibandingkan dengan metode pengelompokan yang klasik, antara lain hasil yang diperoleh lebih mudah untuk diinterpretasikan, lebih cepat perhitungannya dan lebih akurat (Breiman dkk., 1984).

CART merupakan metode yang dapat diterapkan untuk himpunan data yang berjumlah besar dan variabel independennya banyak. Pada *learning sample* dengan banyaknya *class* adalah *u*, maka probabilitas banyaknya objek pada tiap-tiap *class* tersebut adalah sebagai berikut (Breiman dkk., 1984)

$$(\pi_u) = \frac{N_u}{N} \tag{2.13}$$

Dimana:

(π_u) = probabilitas prior

N_u = jumlah objek pada *class u*

N = jumlah objek pada data *training*

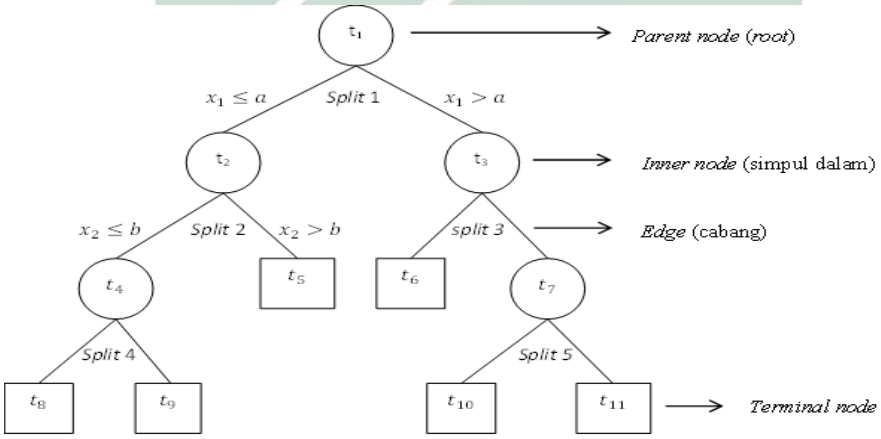
Dalam sebuah *node* t, probabilitas sebuah objek merupakan anggota *class* u

dan berada dalam *node* t adalah sebagai berikut (Breiman dkk., 1984)

$$P(u,t) = \pi_u \frac{N_u(t)}{N_u}$$
$$= \frac{N_u}{N} . \frac{N_u(t)}{N_u}$$

$$P(u,t) = \frac{N_u(t)}{N} \tag{2.14}$$

Diagram pohon klasifikasi CART dapat ditunjukkan pada Gambar 2.1.



Gambar 2.1 Diagram CART

Berdasarkan Gambar 2.1, t_1, t_2, t_3, ..., t_{11} merupakan variabel independen yang terpilih menjadi *node*. t_1 merupakan *parent node* atau simpul induk, t_2, t_3, t_4 dan t_7 merupakan *inner node* atau simpul dalam, sementara t_5, t_6, t_8, t_9, t_{10} dan t_{11} merupakan *terminal node* atau simpul akhir yang tidak dapat bercabang lagi. Setiap *terminal node* adalah titik terakhir dari suatu pemilahan berstruktur pohon, *node* tersebut nantinya akan dimasukkan sebagai suatu kelas tertentu. Variabel independen yang dianggap berpengaruh terhadap variabel dependen adalah variabel independen yang muncul sebagai pemisah (Breiman dkk., 1984).

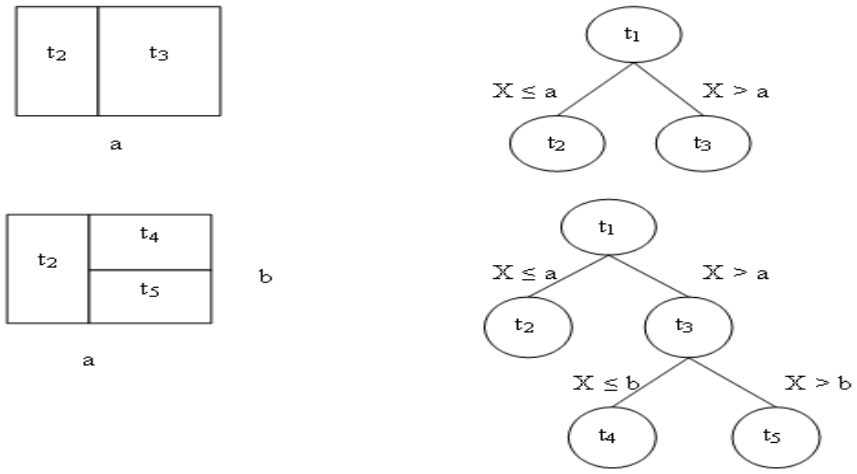
Proses kerja pada CART untuk membentuk sebuah pohon klasifikasi

disebut dengan istilah *binary recursive partitioning*. Proses *binary* yaitu setiap *parent node* akan selalu mengalami pemecahan tepat dua *child node*. Proses *recursive* berarti bahwa pada proses *binary* tersebut akan diulang kembali disetiap *child nodes* sebagai hasil pemecahan terdahulu, sehingga *child nodes* tersebut sekarang menjadi *parent nodes*. Proses pemecahan ini akan dilakukan perulangan sampai tidak dapat melakukan pemecahan berikutnya. Sementara itu, proses *partitioning* berarti bahwa *learning sample* yang dimiliki akan dipecah ke dalam partisi-partisi yang lebih kecil (Breiman dkk., 1984).

Splitting criterion (kriteria pemecahan) didasarkan pada nilai-nilai dari variabel independen yang digunakan. Misalkan variabel dependen *Y* dengan berskala kategorik dan variabel-variabel independen X_1, X_2, …, X_k. Proses *binary recursive partitioning* dapat diilustrasikan sebagai proses penyekatan atau pembagian dari ruang berdimensi *k* dari variabel-variabel *X* kedalam sekatan-sekatan yang berbentuk persegi panjang dan tidak saling bertumpang tindih. Pertama, pilih salah satu variabel independen X_k dan nilai X_k misalkan S_1 terpilih untuk memecah ruang berdimensi k kedalam dua bagian. Bagian pertama adalah sekatan berisi objek-objek yang mana $X_k \leq S_1$ (Breiman dkk., 1984). Sementara itu, bagian lainnya adalah sekatan yang berisi objek-objek dengan nilai $X_k > S_1$. Selanjutnya, masing-masing dari sekatan tersebut dipecah kembali dengan cara yang sama dengan variabel independen (variabel independen yang terpilih dapat X_k kembali atau yang lainnya) dengan nilai tertentu. Proses ini akan terus berlanjut hingga diperoleh sekatan-sekatan yang lebih kecil dengan berisikan objek-objek yang homogen atau seragam. Homogen yang dimaksud adalah objek-objek yang terdapat dalam sekatan tersebut merupakan anggota satu *class* yang sama. Akan tetapi, pada kenyataannya keadaan seperti tersebut tidaklah mutlak selalu diperoleh. Proses *splitting* akan terus berlanjut hingga diperoleh

pohon klasifikasi yang optimal atau maksimal (Breiman dkk., 1984).

Pembentukan partisi dapat ditunjukkan pada Gambar 2.2.



Gambar 2.2 Pembentukan Partisi

Tahapan dalam pembentukan pohon klasifikasi dengan cara membuat pohon besar dengan *node* yang banyak. Pohon yang terbentuk selanjutnya disederhanakan dengan cara memangkas beberapa cabang untuk memperoleh struktur pohon yang layak dengan memperhatikan aturan-aturan tertentu sehingga terbentuk suatu pohon optimal (Breiman dkk., 1984).

Tahap-tahap Pembentukan Pohon Klasifikasi CART adalah sebagai berikut:

1. Pemilihan Pemilah

   Pada tahap ini akan dicari pemilah dari setiap *node* yang menghasilkan penurunan tingkat keheterogenan paling tinggi. Untuk mengukur tingkat keheterogenan suatu *class* dari suatu *node* tertentu dalam pohon klasifikasi disebut dengan *impurity measure* (Breiman dkk., 1993). Pemilihan pemilah masing-masing root *nodes* didasarkan pada aturan *goodness of split criterion*. *Goodness of split criterion* tersebut berdasarkan pada fungsi *impurity*, fungsi

impurity yang dapat diterapkan dalam pembentukan pohon klasifikasi CART adalah Indeks Gini. Derajat *impurity* yang tinggi menunjukkan bahwa *node* tersebut belum homogen. Sementara itu, suatu *node* dengan derajat *impurity* yang rendah menunjukkan bahwa *node* tersebut sudah homogen (Breiman dkk., 1993). Apabila *class* suatu objek dinyatakan dengan u, $u = 1, 2, \ldots, m$, dimana m adalah jumlah *class* untuk varibel dependen (y). Misalkan Indeks Gini untuk mengukur *impurity* dari suatu partisi D, dirumuskan sebagai berikut(Breiman dkk., 1993):

$$Gini\ (D) = 1 - \sum_{u} P^2(u|t)] \tag{2.15}$$

Dimana:

$Gini\ (D)$ = ukuran *impurity* dari partisi D

u = jumlah *class* pada variabel *y*

$P(u|t)$ = peluang *class* u pada *node* t

Pembagian biner diperiksa dengan cara menjumlahkan ukuran *impurity* dari setiap partisi yang diperoleh dari pembagian tersebut. Misalkan pembagian yang dilakukan pada parameter A mempartisi D menjadi D_1 dan D_2. Sehingga Indeks Gini dari pembagian pada parameter A tersebut adalah sebagai berikut (Breiman dkk., 1993)

$$Gini_A\ (D) = \frac{|D_1|}{|D|} \cdot Gini(D_1) + \frac{|D_2|}{|D|} \cdot Gini(D_2) \tag{2.16}$$

dimana:

$Gini_A\ (D)$ = *impurity* dari partisi D pada parameter A

$Gini\ (D_1)$ = *impurity* dari partisi pertama D

$Gini\ (D_2)$ = *impurity* dari partisi kedua D

D = partisi D

D_1 = partisi pertama D

D_2 = partisi kedua D

Penurunan tingkat *impurity* yang dihasilkan dari suatu pembagian biner terhadap parameter A adalah sebagai berikut (Breiman dkk., 1993):

$$\triangle\, Gini\,(A) = Gini\,(D) - Gini_A\,(D) \tag{2.17}$$

dimana:

$\triangle\, Gini(A)$ = tingkat *impurity*

$Gini_A\,(D)$ = *impurity* dari partisi D pada parameter A

$Gini\,(D)$ = *impurity* dari partisi pertama D

Parameter yang memaksimalkan penurunan dari tingkat *impurity* tersebut yang akan dipilih sebagai *splitting* parameter.

2. Penentuan *Terminal Node*

Suatu *node* ke-*t* akan menjadi *terminal node* atau tidak akan dipilih kembali, jika pada *node* tersebut tidak terdapat penurunan keheterogenan (homogen) atau terdapat batasan minimum *n* seperti hanya ada satu pengamatan pada tiap *partition node* (Lewis and Roger, 2000). Menurut Breiman, pada umumnya jumlah kasus maksimum dalam suatu terminal akhir adalah 5, apabila hal tersebut terpenuhi maka pengembangan pohon dihentikan (Breiman dkk., 1984).

3. Pemberian Label *Class*

Pemberian label *class* merupakan proses pengidentifikasian tiap-tiap *nodes* pada suatu *class* tertentu. Pelabelan *class* tidak hanya dilakukan pada *terminal nodes* saja, *non-terminal nodes* dan *root node* juga dilakukan proses ini. Hal ini dikarenakan setiap *non-terminal nodes* mempunyai kesempatan untuk menjadi *terminal nodes* (Breiman dkk., 1993). Walaupun setiap *nodes* mengalami proses ini, akan tetapi *terminal nodes* yang paling membutuhkan proses ini. Hal ini dikarenakan *terminal nodes* merupakan *nodes* yang paling penting dalam memprediksi suatu objek pada *class* tertentu. Terdapat dua aturan dalam pelabelan *class*, masing-masing aturan berdasarkan pada dua macam *misclassification cost* yang berbeda. Misalkan sebuah pohon klasifikasi sudah terbentuk dan mempunyai *terminal nodes* ($\bar{T}$), pelabelan *class* untuk mengidentifikasi sebuah *class* $u \in \{1, 2, ..., U\}$ pada setiap *terminal nodes* $t \in \bar{T}$. *Class* yang akan diidentifikasi pada *node* $t \in \bar{T}$ dinotasikan dengan $u(t)$ (Breiman dkk., 1993).

a) Aturan pertama, pada aturan ini berdasarkan nilai *misclassification cost* yang sama pada setiap *class*. Aturan ini disebut dengan *the plurality rule* (aturan keragaman).

   Didefinisikan aturan pelabelan *class* adalah $u*(t)$ (Breiman dkk., 1993),

$$P(u|t) = max_i P(i|t) \Rightarrow u*(t) = u \tag{2.18}$$

b) Aturan kedua, pada aturan ini berdasarkan nilai *misclassification cost* yang berbeda untuk setiap *class*.

   Didefisinikan aturan pelabelan *class* $u*(t), u*(t) = i_0$ jika i_0 meminimalkan

(Breiman dkk., 1993).

$$\sum_{u} C(i|u)P(u|t) \tag{2.19}$$

dimana $C(i|u) > 0$ jika $i \neq u$ dan $C(i|u) = 0$ jika $i = u$.

Resubtitution estimate $R(t)$ merupakan probabilitas terjadinya *misclassification* yang dialami oleh beberapa objek pada *node* t tertentu. Proses *splitting* pada *node* t mengakibatkan $R(t)$ yang kecil pada kedua *node* baru yang terbentuk. Setiap pemecahan *node* t menjadi *node* $t_L dant_R$ berlaku (Breiman dkk., 1993):

$$R(t) \geq R(t_L) + R(t_R) \tag{2.20}$$

4. Pemangkasan Pohon Klasifikasi

Pada sebuah *maximal tree* (T_{max}) yang dihasilkan dari proses pembentukan pohon dapat terjadi dua permasalahan, yaitu:

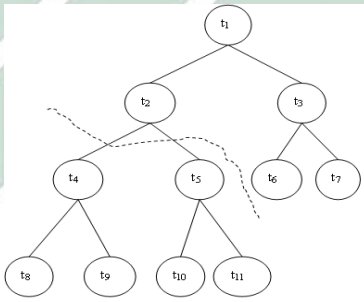
a) Walaupun T_{max} mempunyai tingkat akurasi yang tinggi dengan nilai *misclassification cost* yang kecil, akan tetapi T_{max} dapat mengakibatkan *over-fitting* atau pencocokan nilai yang terlalu kompleks pada data baru. Hal ini dikarenakan terlalu banyak variabel independen yang digunakan untuk melakukan prediksi terhadap data baru tersebut (Breiman dkk., 1993).

b) Sebagai akibat dari *point* a), T_{max} akan sulit untuk diinterpretasikan atau dipahami. Oleh karena itu, T_{max} akan lebih tepat disebut sebagai *complex tree* yang kompleksitasnya ditentukan dari jumlah *terminal nodes* dimiliki (Breiman dkk., 1993).

Proses *tree pruning* (pemangkasan pohon) dilakukan untuk menghilangkan dua permasalahan tersebut dengan cara memangkas atau memotong T_{max}

menjadi beberapa pohon klasifikasi (*T*) yang ukurannya lebih kecil (Breiman dkk., 1984).

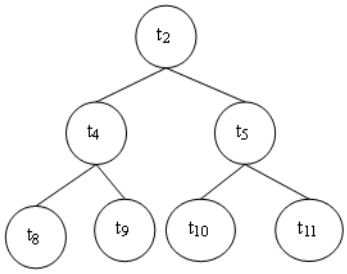
Sebuah *node* t' disebut dengan *descendant* (anak) dari *node* t dan *node* t disebut dengan *ancestor* dari *node* t' apabila kedua *node* tersebut dapat terhubungkan oleh jalur yang bergerak dari atas hingga bawah (*node* t berada di atas *node* t') (Breiman dkk., 1984). Diberikan contoh pohon klasifikasi T pada Gambar 2.3.



Gambar 2.3 Pohon Klasifikasi T

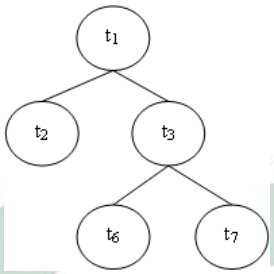
Berdasarkan pohon klasifikasi T pada Gambar 2.3, t_4, t_5, t_8, t_9, t_{10} dan t_{11} merupakan *descendant* dari t_2 namun tidak dengan t_6 dan t_7. Begitu pula dengan t_4, t_2 dan t_1 merupakan *ancestor* dari t_9 tapi t_3 bukan *ancestor* dari t_9.

Suatu *branch* T_t dari T dengan *root node* $t \in T$ terdiri dari *node* t itu sendiri dengan semua *descendant* dari t dalam T (Breiman dkk., 1984). Hal ini ditunjukkan pada Gambar 2.4.



Gambar 2.4 *Branch* T_{t_2}

Pruning (pemangkasan) sebuah *branch* T_t dari sebuah pohon T adalah mengeliminasi atau menghasilkan semua *descendant* t dari *T* kecuali *root node*nya (t itu sendiri). Pohon hasil pemangkasan dinotasikan $T–T_t$ (Breiman dkk., 1984). Hal ini ditunjukkan pada Gambar 2.5.



Gambar 2.5 Pohon Klasifikasi $T - T_{t_2}$ (*Pruned Subtree*)

Apabila T' diperoleh dari T sebagai hasil dari pemangkasan suatu *branch*, maka T' disebut sebagai *pruned subtree* dari T dan dinotasikan dengan $T' < T$. Sebagai catatan dimana T' dan T mempunyai *root node* yang sama (Breiman dkk., 1984). Gambar 2.5 menunjukkan *pruned subtree*.

Metode yang digunakan dalam proses pemangkasan pohon berdasarkan pada *minimal cost complexity pruning* adalah sebagai berikut (Breiman dkk., 1984):

$$R(T) = \sum_{t \in \bar{T}} r(t)P(t) = \sum_{t \in \bar{T}} R(t) \tag{2.21}$$

dimana:

$R(T)$ = *tree misclassification cost/tree resubtitution cost*

$R(t)$ = *node misclassification cost*

Diketahui *subtree* $T < T_{max}$ didefinisikan *complexity* dari *subtree* ini adalah $|\bar{T}|$, yaitu jumlah *terminal nodes* yang dimiliki T, $\alpha \geq 0$ mrupakan *node complexity* dan *cost complexity measure* ($R_\alpha(T)$) dirumuskan sebagai berikut

(Breiman dkk., 1993)

$$R_\alpha(T) = R(T) + \alpha|\bar{T}| \tag{2.22}$$

Nilai dari α dapat diartikan sebagai *penalty* dari setiap *terminal node* berdasarkan kompleksitasnya. Semakin besar nilai α, maka kompleksitas *node*nya juga besar dan ukuran pohon yang dihasilkan akan semakin kecil, begitu sebaliknya apabila nilai α semakin kecil, maka kompleksitas *node*nya juga kecil dan ukuran pohon yang dihasilkan semakin besar. Dengan demikian, apabila sebuah T_{max} mempunyai *terminal node* yang hanya terdapat satu objek (homogen), sehingga nilai α pada T_{max} adalah 0 dan *cost complexity measure* mencapai nilai minimum karena $R(T_{max}) = 0$. Nilai α akan terus meningkat pada saat proses *pruning* dan akan mencapai nilai terbesar ketika *terminal node* = *root node* (Breiman dkk., 1993).

Setiap nilai α dicari pohon bagian $T(\alpha) < T_{max}$ yang meminimumkan $R_\alpha(T)$, hal ini dirumuskan sebagai berikut (Breiman dkk., 1993):

$$R_\alpha((T)) = min_{T<T_{max}} R_\alpha(T) \tag{2.23}$$

Proses pemangkasan pohon klasifikasi dilakukan dengan cara mengambil t_L yang merupakan *node* anak kiri dan t_R yang merupakan *node* anak kanan dari T_{max} yang dihasilkan dari *root node* t. Apabila diperoleh *root node* dan dua *node* anak yang memenuhi persamaan $R(t) = R(t_L) + R(t_R)$, maka *node* anak t_L dan t_R dipangkas. Sehingga diperoleh pohon T_1 yang memenuhi kriteria $R(T_1) = R(T_{max})$. Proses tersebut akan diulang hingga tidak terdapat pemangkasan yang mungkin (Breiman dkk., 1993).

5. Penentuan Pohon Klasifikasi Optimal

Setelah dilakukan pemangkasan bagian pohon yang kurang penting, maka akan diperoleh pohon klasifikasi optimal. Metode pemangkasan pohon pada sebelumnya memperoleh urutan *subtree* $T_1 > T_2 > ... > \{t\}$, karena pohon klasifikasi yang diperoleh berjumlah cukup banyak, maka permasalahan yang timbul adalah bagaimana cara untuk menentukan pohon klasifikasi yang optimal (Breiman dkk., 1993). Apabila menggunakan Persamaan (2.21)

$$R(T) = \sum_{t \in \bar{T}} r(t)P(t) = \sum_{t \in \bar{T}} R(t)$$

maka T_1 akan dipilih sebagai pohon klasifikasi optimal, karena nilai *resubtitution estimate* dari T_1 paling kecil. Oleh karena itu, metode *resubtitution estimate* merupakan metode yang bias untuk mengestimasi nilai *true misclassification cost*. Ada dua metode tak bias untuk mengestimasi nilai *true misclassification cost* yaitu penduga sampel uji *test sample estimate* dan penduga validasi silang lipat V (*cross validation V-fold estimate*) (Breiman dkk., 1993).

Test sample estimate digunakan apabila ukuran data dengan jumlah besar (L). Kemudian data (L) tersebut dibagi menjadi dua bagian L_1 dan L_2. Misalkan L_1 sebanyak N_1 objek dan L_2 sebanyak N_2 objek. Selanjutnya, dibuat T_{max} menggunakan L_2 dan dilakukan pemangkasan hingga diperoleh $\{T_k\} = T_1 > T_2 > ... > \{t\}$. L_2 digunakan untuk memprediksi pada masing-masing *tree* dan dihitung banyak objek yang terjadi kesalahan klasifikasi (*misclassification*) (Breiman dkk., 1993).

Sementara itu, *cross validation* digunakan jika ukuran data berjumlah kecil. Dalam *V-fold cross validation learning sample L* dibagi secara acak dalam V bagian terpisah, sebanyak *V-1* bagian sebagai *learning sample* dan sisa (*V-(V-1)*) sebagai *test sample*. Setiap bagian terdapat objek dengan jumlah yang

sama (hampir sama) (Breiman dkk., 1993).

Untuk *binary tree*, parameter *cost complexity* bernilai 0,5 yang berarti suatu *node* selalu dikembangkan menjadi dua *child node* (Breiman dkk., 1984). Tingkat kesalahan klasifikasi (*misclassification error*) pada *node* ke-*t* didefinisikan dalam persamaan sebagai berikut (Breiman dkk., 1993):

$$error(t) = 1 - max_i P(i|t) \tag{2.24}$$

2.6. Ketepatan Klasifikasi

Hasil dari proses klasifikasi akan diperoleh kelas yang benar dan salah. Untuk menghitung ketepatan ketepatan klasifikasi (akurasi) dapat disajikan dalam bentuk tabel *confusion matrix*. *Confusion matrix* merupakan suatu tabel yang berisi jumlah data uji yang salah diklasifikasikan dan jumlah data uji yang benar diklasifikasikan (Johnson dan Wichern, 2007). *Confusion matrix* untuk klasifikasi dengan kategorik biner ditunjukkan pada Tabel 2.5.

Tabel 2.5 *Confusion Matrix* untuk Klasifikasi Biner

Klasifikasi		**Kelas Prediksi**	
Kelas		1	0
Kelas	1	TP	FN
Aktual	0	FP	TN

Ketepatan klasifikasi dapat dinyatakan dalam persamaan sebagai berikut:

$$Akurasi = \frac{TP + TN}{TP + FN + FP + TN} \times 100\% \tag{2.25}$$

dimana,

TP (*True Positive*) adalah jumlah data dari kelas 1 yang benar dan diklasifikasikan sebagai kelas 1.

TN (*True Negative*) adalah jumlah data dari kelas 0 yang benar dan diklasifikasikan sebagai kelas 0.

FP (*False Positive*) adalah jumlah data dari kelas 0 yang salah dan diklasifikasikan sebagai kelas 1.

FN (*False Negative*) adalah jumlah data dari kelas 1 yang salah dan diklasifikasikan sebagai kelas 0.

2.7. Integrasi Keilmuan

Pemerintah Indonesia saat ini sedang mengupayakan pencegahan penyakit Diabetes Mellitus guna mengurangi jumlah penderita Diabetes Mellitus. Hal tersebut dilakukan dengan cara mengidentifikasi apakah seseorang mengidap Diabetes Mellitus agar dapat melakukan perawatan yang tepat untuk pasien. Sebagaimana firman Allah Swt. yang tertuang dalam Al-Qur'an surah Al-Isra ayat 82, yakni sebagai berikut:

وَنُنَزِّلُ مِنَ ٱلْقُرْءَانِ مَا هُوَ شِفَآءٌ وَرَحْمَةٌ لِّلْمُؤْمِنِينَ ۙ ●●●● ﴿٨٢﴾

Artinya:

"Dan Kami turunkan dari Al Quran suatu yang menjadi penawar dan rahmat bagi orang-orang yang beriman .... (Al-Isra : 82)".

Berdasarkan potongan surah Al Isra ayat 82 menunjukkan bahwa Allah tidak akan menurunkan penyakit kecuali menurunkan pula penawar (obatnya). Pada ayat tersebut menjelaskan bahwa obat yang paling mujarab salah satunya yang telah disediakan oleh Allah untuk manusia yaitu Al-Qur'an. Sebagai umat manusia, tentu harus menyakini bahwa Allah akan menyembuhkan suatu penyakit

walau separah apapun penyakit tersebut, asalkan tetap berusaha dan berdoa kepada-Nya. Hal ini dijelaskan juga pada H.R. Muslim yang artinya *"Setiap penyakit ada obatnya. Apabila ditemukan obat yang tepat untuk suatu penyakit, akan sembuhlah penyakit itu dengan izin Allah 'azza wajalla"*. Bentuk usaha yang dapat dilakukan manusia yaitu dengan melakukan perawatan medis yang tepat berdasarkan pada diagnosa. Dengan begitu, pendeteksian dini suatu penyakit dengan tepat dapat menunda dan/atau mencegah kematian akibat dari suatu penyakit.

BAB III

METODE PENELITIAN

Pada bab ini akan dijelaskan metodologi penelitian sedemikian sehingga penelitian ini dapat terarah dengan baik dalam hal materi dan waktu pengerjaannya.

Metodologi penelitian untuk penelitian suatu studi kasus memuat uraian mengenai jenis penelitian, sumber data, pengumpulan data dan tahap penelitian. Penjelasan dalam metodologi penelitian dilengkapi dengan diagram alur (*flowchart*) yang berfungsi untuk menjelaskan uraian tahap-tahap penelitian.

3.1. Jenis Penelitian

Pada penelitian ini bersifat kuantitatif, yaitu penelitian yang lebih menekankan pada analisis data yang diolah menggunakan metode statistik.

3.2. Sumber Data

Penelitian ini menggunakan data pasien yang positif Diabetes Mellitus dan negatif Diabetes Mellitus yang diperoleh dari Rumah Sakit X pada tahun 2019 (Arvianti, 2019).

3.3. Pengumpulan Data

Dataset *Diabetes Mellitus* yang digunakan dalam penelitian ini sejumlah 240 sampel pasien yang terdiri dari 72 negatif Diabetes Mellitus dan 168 positif Diabetes Mellitus. Berdasarkan sampel tersebut diambil sebanyak 70% untuk

training dan 30% untuk *testing*. Kemudian, dari data *training* diperoleh model matematis yang akan digunakan untuk mengklasifikasikan data *testing*.

1. Variabel Penelitian

   Variabel yang digunakan dalam penelitian ini adalah sebagai berikut:

   a) Variabel dependen, yaitu diabetes mellitus yang terdiri dari dua kategorik negatif Diabetes Mellitus (0) dan positif Diabetes Mellitus (1).

   b) Variabel independen, yaitu usia, faktor keturunan Diabetes Mellitus, kadar gula darah, obesitas serta pola makan.

2. Definisi Operasional Variabel

   Definisi operasional variabel-variabel dalam penelitian ini adalah sebagai berikut:

   a) Diabetes Mellitus (Y)

      Variabel Diabetes Mellitus dalam penelitian ini merupakan variabel dependen ”Y” yang berskala kategorik. Variabel ini merupakan *output* yang menyatakan bahwa pasien termasuk dalam *class* negatif Diabetes Mellitus atau *class* positif Diabetes Mellitus.

   b) Usia (X_1)

      Usia dalam penelitian ini merupakan variabel independen sebagai uji pengendalian apakah pasien negatif Diabetes Melitus atau negatif Diabetes Mellitus, data berupa skala kategorik (0 = ”Muda” dan 1 = ”Tua”).

   c) Faktor Keturunan (X_2)

      Faktor keturunan dalam penelitian ini merupakan variabel independen sebagai uji untuk pengendalian apakah pasien negatif Diabetes Mellitus atau

positif Diabetes Mellitus, data berupa skala kategorik (0 = ”Tidak Ada” dan 1 = ”Ada”).

d) Kadar Gula Darah (X_3)

Kadar gula darah dalam penelitian ini merupakan variabel independen sebagai uji pengendalian apakah pasien negatif Diabetes Mellitus atau negatif Diabetes Mellitus, data berupa skala kategorik (0 = ”Normal” dan 1 = ”Tidak Normal”) dihitung sejak dilakukan pemeriksaan kadar gula darah sewaktu (2 jam setelah makan) dengan satuan mg/dL.

e) Obesitas (X_4)

Obesitas dalam penelitian ini merupakan variabel independen sebagai uji pengendalian apakah pasien negatif Diabetes Mellitus atau positif Diabetes Mellitus, data berupa skala kategorik (0 = ”Normal” dan 1 = ”Tidak Normal”) dihitung sejak dilakukan pemeriksaan Indeks Massa Tubuh (IMT) pasien.

f) Pola Makan (X_5)

Pola makan dalam penelitian ini merupakan variabel independen sebagai uji pengendalian apakah pasien negatif Diabetes Mellitus atau negatif Diabetes Mellitus, data berupa skala kategorik (0 = ”Baik” dan 1 = ”Buruk”).

3.4. Tahap-tahap Penelitian

Dalam penelitian ini dilakukan melalui beberapa tahap. Tahapan tersebut akan dijelaskan sebagai berikut:

1. Mencari dan membaca referensi terkait dengan topik penelitian.
2. Mengumpulkan dataset pasien yang positif *Diabetes Mellitus* dan negatif *Diabetes Mellitus*.

3. Mendeskripsikan dan memberi label pada variabel-variabel yang digunakan.

4. Membagi data secara acak menjadi dua bagian, yaitu data *training* dan data *testing* ukuran menggunakan standar *training-testing* yaitu 70% untuk data *training* dan 30% untuk data *testing*.

5. Melakukan klasifikasi menggunakan metode Regresi Logistik Biner melalui beberapa tahap, tahap klasifikasi menggunakan metode Regresi Logistik Biner adalah sebagai berikut:

   a) *Input* Data *Training*

      Melakukan *input* data *training* penyakit Diabetes Mellitus sebagai variabel dependen (Y) dan variabel independen (X).

   b) Penaksir Parameter

      Untuk memperoleh model regresi logistik biner, langkah awal yang dilakukan adalah mencari nilai taksiran parameter yaitu $\beta_1, \beta_2, \beta_3, \beta_4$ dan β_5. Pada tahap ini dilakukan proses eliminasi variabel independen apabila variabel tersebut mempunyai nilai signifikansi lebih dari 0,05. Proses tersebut akan berhenti apabila tiap-tiap variabel independen telah signifikan.

   c) Uji Signifikansi (Sig.)

      Pada uji signifikansi variabel independen (X) yang mempunyai pengaruh terhadap variabel dependen (Y) dapat dilakukan menggunakan dua uji, yaitu uji G dan uji Wald.

      1) Uji G

         $H_0 : \beta_1 = \beta_2 = ... = \beta_p$

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

         Pada pengujian ini menggunakan Persamaan (2.11). Dengan kriteria

penolakan (tolak H_0) jika $G > X^2_{(db,\alpha)}$ atau $(Sig.) < \alpha$. Dengan demikian, dapat ditarik kesimpulan bahwa variabel independen secara serentak mempengaruhi variabel dependen.

2) Uji Wald

$H_0 : \beta_j = 0, \quad j = 1, 2, ..., p$

$H_1 : \beta_j \neq 0, \quad j = 1, 2, ..., p.$

Pengujian hipotesis pada Uji Wald menggunakan Persamaan (2.12). Dengan kriteria penolakan (tolak H_0) apabila $|W| > Z_{\alpha/2}$ atau $(Sig.) < \alpha$. Dengan demikian, dapat disimpulkan bahwa variabel independen secara individu berpengaruh terhadap variabel dependen.

d) Model Terbaik

Menentukan model terbaik regresi logistik menggunakan Persamaan (2.3).

6. Langkah selanjutnya, melakukan klasifikasi menggunakan metode CART melalui beberapa tahap, tahap klasifikasi menggunakan metode CART adalah sebagai berikut:

a) Membentuk pohon klasifikasi maksimal menggunakan data *training* dengan langkah-langkah sebagai berikut:

1) Melakukan pemilihan pemilah masing-masing *nodes* berdasarkan aturan dari pemilihan Indeks Gini menggunakan Persamaan (2.15) dan (2.16). Setelah diperoleh nilai Indeks Gini, langkah selanjutnya melakukan evaluasi hasil dari pemilahan menggunakan uji *goodness of split criterion* menggunakan Persamaan (2.17).

2) Menentukan *terminal nodes*. Suatu *node* akan menjadi *terminal node* ketika dalam *node* tersebut terdapat kasus yang homogen atau terdapat jumlah kasus minimum 5.

3) Menentukan label *class*, pemberian label *class* dilakukan berdasarkan dua aturan. Aturan pertama yaitu *the plurality rule* menggunakan Persamaan (2.18). Aturan kedua berdasarkan nilai *misclassification cost* yang berbeda untuk setiap *class* menggunakan Persamaan (2.19).

b) Memangkas pohon *pruning* berdasarkan kriteria *cost complexity minimum* menggunakan Persamaan (2.21).

c) Memilih pohon klasifikasi yang optimal menggunakan *test sample estimates*.

d) Membuat kesimpulan dari hasil analisis.

7. *Input* Data *Testing*

Melakukan *input* data *testing* pada masing-masing metode yaitu regresi logistik biner dan CART.

8. Prediksi Data Pasien Diabetes Mellitus

Data *testing* selanjutnya diprediksi menggunakan model dari masing-masing metode yang telah terbentuk.

9. Perhitungan Tingkat Akurasi

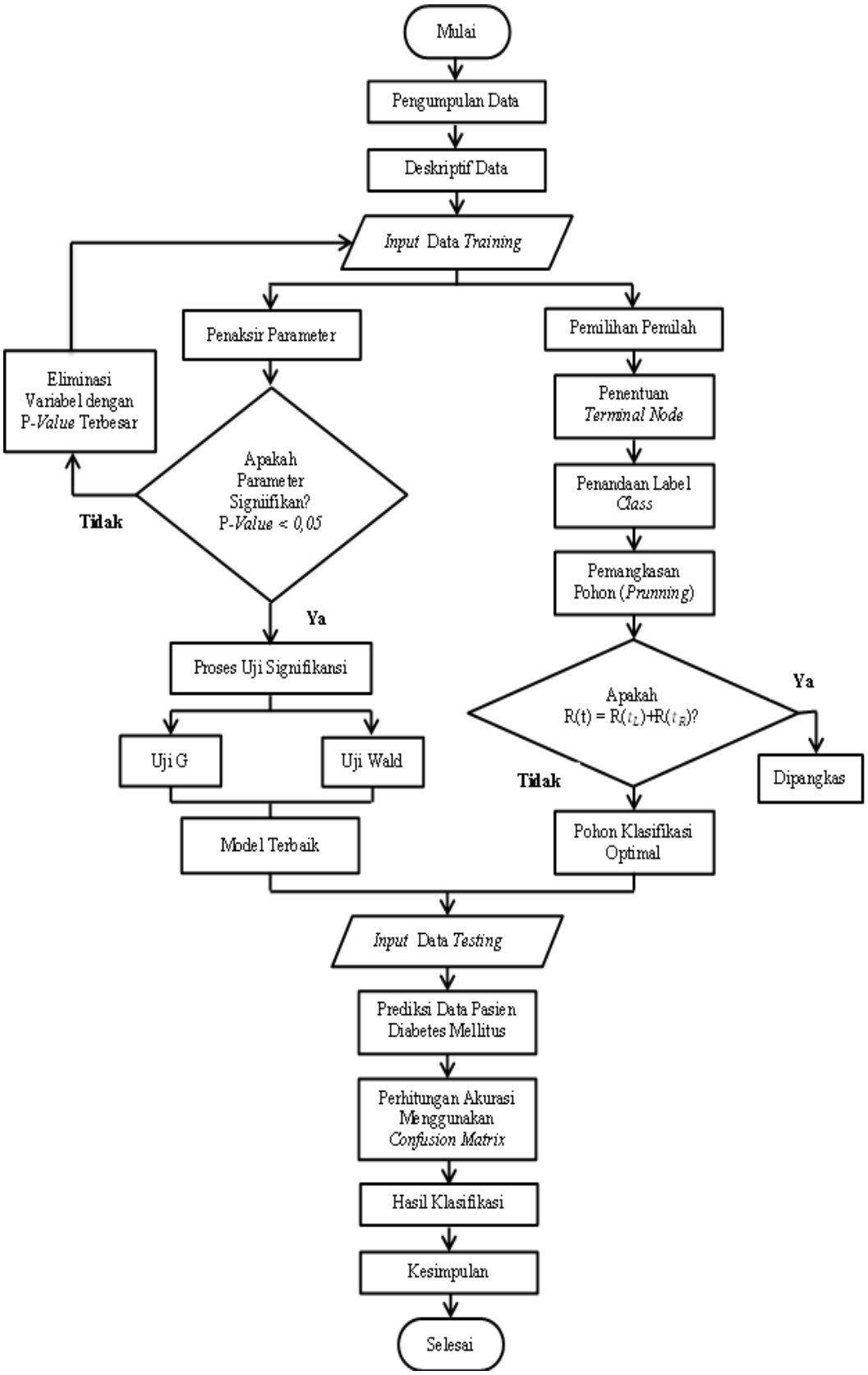
Menghitung ketepatan klasifikasi dari metode Regresi Logistik Biner dan CART menggunakan Persamaan (2.25).

10. Kesimpulan

Menyimpulkan dari hasil penelitian.

3.5. *Flowchart*

Tahap-tahap penelitian ini dapat diilustrasikan menggunakan diagram alir *(flowchart)* yang disajikan pada Gambar 3.1.



Gambar 3.1 Diagram Alir (*Flowchart*)

BAB IV

HASIL DAN PEMBAHASAN

Pada penelitian ini akan dibahas analisis klasifikasi untuk penentuan pasien positif Diabetes Mellitus (DM) atau negatif Diabetes Mellitus (DM) menggunakan Metode Regresi Logistik Biner dan *Classification And Regression Trees* (CART). Hasil analisis klasifikasi dari kedua metode tersebut akan dibandingkan untuk mengetahui metode yang lebih baik untuk mendiagnosa pasien Diabetes Mellitus.

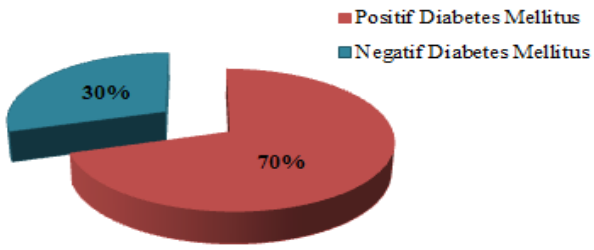
4.1. Deskriptif Data Diabetes Mellitus

Proses pertama yang dilakukan untuk klasifikasi adalah melakukan analisis data. Jumlah sampel yang digunakan dalam penelitian sebanyak 240 pasien yang terdiri dari variabel Y yaitu, positif DM dan negatif DM serta variabel X, yaitu usia (X_1), keturunan (X_2), kadar gula darah (X_3), obesitas (X_4) dan pola makan (X_5). Sampel dari data tersebut disajikan pada Tabel 4.1.

Tabel 4.1 Data Pasien Diabetes Mellitus pada Rumah Sakit X

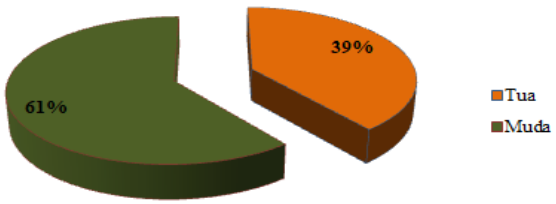
Data	X_1	X_2	X_3	X_4	X_5	Y
1	1	0	0	0	1	Negatif
2	0	0	0	1	1	Negatif
3	0	1	1	1	0	Positif
4	0	1	0	1	1	Positif
⋮	⋮	⋮	⋮	⋮	⋮	⋮
240	0	1	0	1	1	Positif

Berdasarkan Tabel 4.1, sampel yang digunakan dalam penelitian masing-masing variabel X maupun Y berupa kategorik. Pada variabel Y, ”0” menyatakan negatif Diabetes Mellitus sedangkan ”1” menyatakan positif Diabetes Mellitus. Diagnosis penyakit Diabetes Mellitus ditunjukkan pada Gambar 4.1.



Gambar 4.1 *Pie Chart* Diagnosis Penyakit Diabetes Mellitus

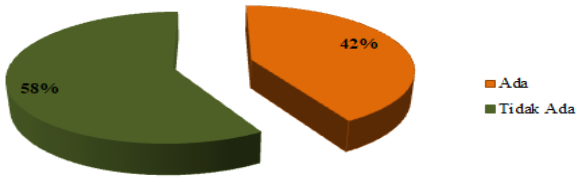
Berdasarkan *pie chart* pada Gambar 4.1, menunjukkan diagnosis jumlah pasien positif Diabetes Mellitus sebesar 70% atau sebanyak 118 pasien. Sementara itu, jumlah pasien terdiagnosis negatif Diabetes Mellitus sebesar 30% atau sebanyak 50 pasien. Selanjutnya, berdasarkan variabel usia disajikan pada Gambar 4.2.



Gambar 4.2 *Pie Chart* Usia

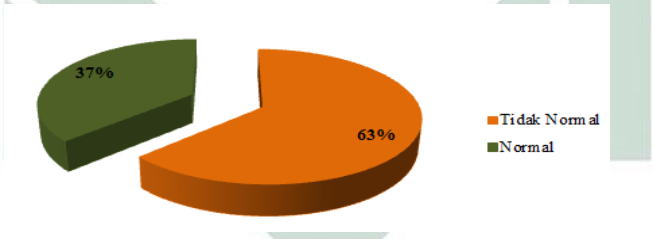
Pada Gambar 4.2, menunjukkan jumlah faktor usia pasien dengan kategorik ”muda” sebesar 61% atau sebanyak 102 pasien, dikatakan muda apabila usia pasien tersebut$<$ 45 tahun. Sementara itu, jumlah pasien dengan kategorik ”tua”

sebesar 39% atau sebanyak 66 pasien, dikatakan tua apabila usia pasien ≥ 45 tahun. Selanjutnya, berdasarkan variabel keturunan disajikan pada Gambar 4.3.



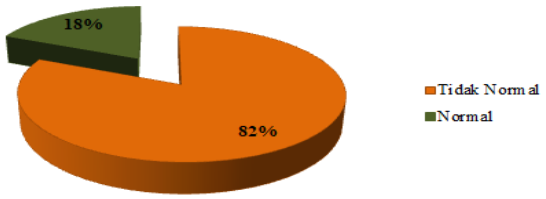
Gambar 4.3 *Pie Chart* Faktor Keturunan

Pada Gambar 4.3, menunjukkan jumlah faktor keturunan pasien dengan kategorik ”tidak ada” (tidak mempunyai riwayat keturunan DM) sebesar 58% atau sebanyak 98 pasien. Sementara itu, jumlah faktor keturunan pasien dengan kategorik ”ada” (mempunyai riwayat keturunan DM) sebesar 42% atau sebanyak 70 pasien. Selanjutnya, berdasarkan variabel gula darah disajikan pada Gambar 4.4.



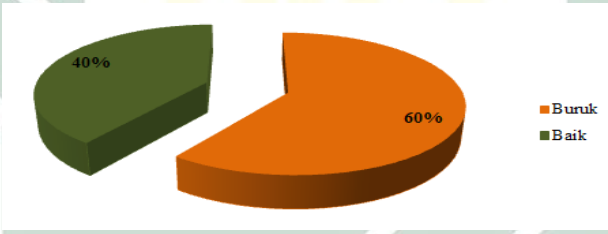
Gambar 4.4 *Pie Chart* Kadar Gula Darah

Pada Gambar 4.4, menunjukkan jumlah kadar gula darah pasien dengan kategorik ”normal” sebesar 37% atau sebanyak 62 pasien, dikatakan normal apabila pasien mempunyai kadar gula darah ≤ 140 mg/dL. Sementara itu, jumlah faktor usia pasien dengan kategorik ”tidak normal” sebesar 63% atau sebanyak 106 pasien, dikatakan tidak normal apabila pasien mempunyai kadar gula darah > 140 mg/dL. Selanjutnya, berdarakan variiabel obesitas disajikan pada Gambar 4.5.



Gambar 4.5 *Pie Chart* Obesitas

Pada Gambar 4.5, menunjukkan jumlah obesitas pasien dengan kategorik ”normal” sebesar 18% atau sebanyak 31 pasien, dikatakan normal apabila pasien mempunyai tingkat obesitas ≤ 33. Sementara itu, jumlah faktor usia pasien dengan kategorik ”tidak normal” sebesar 82% atau sebanyak 137 pasien, dikatakan tidak normal apabila pasien mempunyai tingkat obesitas ≥ 34. Selanjutnya, berdasarkan variabel pola makan disajikan pada Gambar 4.6.



Gambar 4.6 *Pie Chart* Pola Makan

Pada Gambar 4.6, menunjukkan jumlah pola makan pasien dengan kategorik ”baik” sebesar 40% atau sebanyak 67 pasien, dikatakan kategorik baik apabila pasien mengkonsumsi makan-makanan yang bergizi dan sehat. Sementara itu, jumlah faktor usia pasien dengan kategorik ”buruk” sebesar 60% atau sebanyak 101 pasien, dikatakan kategorik buruk apabila pasien mengkonsumsi makanan yang tidak baik bagi tubuh.

Langkah selanjutnya, dilakukan pembagian data menjadi dua bagian yaitu data *training* dan data *testing* dengan proporsi 70% : 30%. Dengan demikian, data

training sebanyak 168 data dan data *testing* sebanyak 72 data. Selanjutnya data tersebut akan dilakukan analisis menggunakan Metode Regresi Logistik Biner dan *Classification And Regression Trees* (CART).

4.2. Analisis Klasifikasi Regresi Logistik Biner

Tahap awal yang dilakukan untuk pembentukan model Regresi Logistik Biner adalah dengan melakukan penaksiran parameter model.

1. Penaksiran Parameter

   Pada tahap ini, seluruh variabel dari data *training* diuji terlebih dahulu untuk memperoleh variabel independen yang berpengaruh terhadap variabel dependen dengan cara mengeliminasi varibel independen jika nilai signifikansi variabel tersebut lebih besar dari nilai signifikansi yang ditetapkan yaitu nilai *alpha* sebesar 0,05. Dengan demikian, jika nilai signifikansi variabel independen kurang dari 0,05, maka variabel-variabel tersebut yang akan dimasukkan ke dalam model Regresi Logistik Biner. Penaksiran parameter β untuk seluruh variabel pada iterasi pertama disajikan pada Tabel 4.2.

Tabel 4.2 Penaksiran Parameter pada Iterasi Pertama

Koefisien	Taksiran Parameter (β)	Std. *Error*	z *value*	Sig.	Nilai AIC
(*Intercept*) (β_0)	-6,7136	1,3853	-4,846	1,26e-06 ***	81,93312
Usia (β_1)	1,4696	0,8155	1,802	0,0715	
Keturunan (β_2)	5,5796	1,2722	4,386	1,16e-05 ***	
Gula Darah (β_3)	3,9630	0,7072	5,604	2,10e-08 ***	
Obesitas (β_4)	4,7321	1,1352	4,169	3,06e-05 ***	
Pola Makan (β_5)	-0,3562	0,7257	-0,491	0,6235	

Berdasarkan Tabel 4.2, menunjukkan nilai signifikansi masing-masing variabel. Dari variabel usia, keturunan, gula darah dan obesitas mempunyai nilai signifikansi kurang dari 0,05. Sementara itu, untuk variabel pola makan mempunyai nilai signifikansi lebih besar dari 0,05 artinya variabel pola makan tidak berpengaruh terhadap penyakit Diabetes Mellitus. Dengan demikian, variabel pola makan dieliminasi dan selanjutnya variabel yang lain akan di uji seperti sebelumnya. Pada iterasi kedua ini penaksiran parameter disajikan pada Tabel 4.3.

Tabel 4.3 Penaksiran Parameter pada Iterasi Kedua

Koefisien	**Taksiran Parameter** (β)	**Std.** ***Error***	**z *value***	**Sig.**	**Nilai AIC**
(*Intercept*) (β_0)	-6,9683	1,3238	-5,264	1,41e-07 ***	
Usia (β_1)	1,6115	0,7540	2,137	0,0326 *	
Keturunan (β_2)	5,6213	1,2993	4,327	1,51e-05 ***	80,17622
Gula Darah (β_3)	3,9764	0,7037	5,651	1,59e-08 ***	
Obesitas (β_4)	4,6913	1,1394	4,117	3,83e-05 ***	

Berdasarkan Tabel 4.3, menunjukkan nilai signifikansi variabel usia, keturunan, gula darah dan obesitas masing-masing sebesar 0,0326, 1,51e-05, 1,59e-08 dan 3,83e-05. Nilai signifikansi masing-masing variabel tersebut kurang dari 0,05 maka variabel tersebut mempunyai pengaruh secara signifikan terhadap penyakit Diabetes Mellitus. Dengan demikian, variabel-variabel tersebut dapat dimasukkan dalam model Regresi Logistik Biner sebagai berikut:

$$\pi(x_i) = \frac{e^{(-6{,}9683+1{,}6115X_1+5{,}6213X_2+3{,}9764X_3+4{,}6913X_4)}}{1+e^{(-6{,}9683+1{,}6115X_1+5{,}6213X_2+3{,}9764X_3+4{,}6913X_4)}} \qquad (4.1)$$

Langkah selanjutnya, akan dilakukan beberapa uji terhadap model yang diperoleh.

2. Uji Signifikansi Model

   a. Uji G

   Uji G dilakukan untuk mengetahui pengaruh variabel independen terhadap variabel dependen secara serentak. Hipotesis dalam pengujian ini H_0 ditolak jika nilai G lebih besar dari nilai X^2 tabel. Diperoleh hasil Uji G pada Tabel 4.4.

Tabel 4.4 Uji G

G^2	G	$X^2_(db, \alpha)$
134,634490	11, 60323	9,48773

   Berdasarkan hasil pengujian secara serentak, diperoleh nilai G sebesar 11,60323 sedangkan nilai X^2_{Tabel} sebesar 9,487729. Hal ini menunjukkan bahwa nilai $G > X^2$, maka tolak H_0 yang artinya variabel usia (X_1), keturunan (X_2), gula darah (X_3) dan obesitas (X_4) secara serentak mempunyai pengaruh terhadap variabel DM (Y).

   b. Uji Wald

   Uji Wald dilakukan untuk mengetahui pengaruh variabel independen terhadap variabel dependen secara individu. Hipotesis dalam pengujian ini H_0 ditolak jika nilai p-*value* $< (\alpha = 0,05)$. Bedasarkan Tabel 4.5 :

Tabel 4.5 Nilai Z_{value} dan P_{value}

Variabel	Z_{value}	P_{value} **/ Sig.**
(*Intercept*)	-5,264	1,41e-07 ***
Usia (X_1)	2,137	0,0326 *
Keturunan (X_2)	4,327	1,51e-05 ***
Gula Darah (X_3)	5,651	1,59e-08 ***
Obesitas (X_4)	4,117	3,83e-05 ***

Berdasarkan hasil pengujian secara individu, diperoleh nilai P_{value} masing-masing variabel independen $X_1 = 0,0326$, $X_2 = 1,51e-05$, $X_3 = 1,59e-08$ dan $X_4 = 3,83e-05$. Masing-masing nilai P_{value} tersebut $< (\alpha = 0,05)$. Oleh karena itu, secara individu variabel independen X_1, X_2, X_3 dan X_4 mempunyai pengaruh terhadap variabel diabetes (Y).

3. Diagnosa Diabetes Mellitus

Setelah diperoleh suatu model Regresi Logistik Biner dan masing-masing uji terpenuhi, selanjutnya akan dilakukan diagnosa penyakit Diabetes Mellitus dari data *testing* yang telah disiapkan. Data tersebut diuji menggunakan model yang diperoleh sebelumnya.

Misalkan untuk mengetahui seseorang menderita penyakit DM dengan ciri-ciri muda, tidak mempunyai riwayat keturunan DM, kadar gula darah normal dan memiliki tingkat kegemukan (obesitas) normal. Maka dapat dihitung menggunakan model Regresi Logistik Biner yang telah dibentuk sebagai berikut:

$$\pi(x_i) = \frac{e^{(-6,9683+1,6115X_1+5,6213X_2+3,9764X_3+4,6913X_4)}}{1+e^{(-6,9683+1,6115X_1+5,6213X_2+3,9764X_3+4,6913X_4)}}$$

$$= \frac{e^{(-6,9683+1,6115(0)+5,6213(0)+3,9764(0)+4,6913(0))}}{1+e^{(-6,9683+1,6115(0)+5,6213(0)+3,9764(0)+4,6913(0))}}$$

$$= \frac{e^{-6,9683}}{1+e^{-6,9683}}$$

$$= 0{,}000940367$$

Berdasarkan model tersebut seseorang dengan ciri-ciri muda, tidak mempunyai riwayat keturunan DM, kadar gula darah normal dan tingkat kegemukan (obesitas) normal sebesar $0{,}000940367 \leq 0{,}5$. Sehingga seseorang dengan ciri-ciri tersebut dinyatakan negatif Diabetes Mellitus.

Dengan demikian, akan diperoleh diagnosa pasien positif dan negatif Diabetes Mellitus pada Tabel 4.6 sebagai berikut:

No.	**Aktual**	Probabilitas	**Prediksi**	**No.**	**Aktual**	Probabilitas	**Prediksi**	**No.**	**Aktual**	Probabilitas	**Prediksi**
1	Negatif	0.09305	Negatif	25	Positif	0.84546	Positif	49	Positif	0.84546	Positif
2	Negatif	0.20636	Negatif	26	Positif	0.99934	Positif	50	Positif	0.09305	Negatif
3	Negatif	0.20094	Negatif	27	Positif	0.84546	Positif	51	Positif	0.96480	Positif
4	Negatif	0.33951	Negatif	28	Positif	0.99934	Positif	52	Positif	0.84546	Positif
5	Negatif	0.00094	Negatif	29	Positif	0.84546	Positif	53	Negatif	0.09305	Negatif
6	Negatif	0.33951	Negatif	30	Positif	0.99987	Positif	54	Negatif	0.33951	Negatif
7	Negatif	0.00094	Negatif	31	Positif	0.99301	Positif	55	Negatif	0.00469	Negatif
8	Negatif	0.09305	Negatif	32	Positif	0.99301	Positif	56	Negatif	0.20636	Negatif
9	Negatif	0.96592	Positif	33	Positif	0.96480	Positif	57	Negatif	0.20636	Negatif
10	Negatif	0.04779	Negatif	34	Positif	0.84546	Positif	58	Negatif	0.84546	Positif
11	Negatif	0.04779	Negatif	35	Positif	0.99934	Positif	59	Negatif	0.20636	Negatif
12	Negatif	0.04779	Negatif	36	Positif	0.99934	Positif	60	Positif	0.84546	Positif
13	Negatif	0.84546	Positif	37	Positif	0.96480	Positif	61	Positif	0.98581	Positif
14	Negatif	0.20094	Negatif	38	Positif	0.99934	Positif	62	Positif	0.96592	Positif
15	Negatif	0.84546	Positif	39	Positif	0.96480	Positif	63	Positif	0.96480	Positif
16	Positif	0.99934	Positif	40	Positif	0.96592	Positif	64	Positif	0.99934	Positif
17	Positif	0.96592	Positif	41	Positif	0.84546	Positif	65	Positif	0.99987	Positif
18	Positif	0.99934	Positif	42	Positif	0.99301	Positif	66	Positif	0.84546	Positif
19	Positif	0.99934	Positif	43	Positif	0.96592	Positif	67	Positif	0.96480	Positif
20	Positif	0.96480	Positif	44	Positif	0.96480	Positif	68	Positif	0.99934	Positif
21	Positif	0.96480	Positif	45	Positif	0.93273	Positif	69	Positif	0.84546	Positif
22	Positif	0.96480	Positif	46	Positif	0.96480	Positif	70	Positif	0.99934	Positif
23	Positif	0.09305	Negatif	47	Positif	0.98581	Positif	71	Positif	0.93273	Positif
24	Positif	0.96480	Positif	48	Positif	0.96480	Positif	72	Positif	0.96480	Positif

Berdasarkan Tabel 4.6, menunjukkan bahwa terdapat sampel yang tidak sesuai antara data prediksi dengan data aktual, yaitu pada sampel ke-9, 13, 15, 23, 50 dan 58. Pada sampel data ke-9, 13, 15 dan 58 pasien diprediksi positif Diabetes Mellitus, akan tetapi pada data sebenarnya pasien negatif Diabetes Mellitus. Sementara itu, pada sampel data ke-23 dan ke-50 pasien diprediksi negatif Diabetes Mellitus, akan tetapi pada data sebenarnya pasien positif Diabetes Mellitus. Oleh karena itu, dalam prediksi ini terjadi 6 (enam) kesalahan klasifikasi.

4. Ketepatan Klasifikasi

Untuk mengetahui tingkat keakurasian dari metode ini dapat menggunakan tabel *confusion matrix* pada Tabel 4.7 sebagai berikut:

Tabel 4.7 Ketepatan Klasifikasi Model Terbaik Regresi Logistik Biner Berdasarkan Data *Training* & *Testing*

Klasifikasi		*Class* Prediksi		Akurasi	*Error*
		1	0		
Training					
Class Aktual	1	112	6	92,86%	7,14%
	0	6	44		
Testing					
Class Aktual	1	48	2	91,67%	8,33%
	0	4	18		

Berdasarkan Tabel 4.7 dan Persamaan 2.28, maka nilai akurasi klasifikasi Diabetes Mellitus menggunakan Regresi Logistik Biner adalah sebagai berikut:

a. Ketepatan klasifikasi berdasarkan data *training*

$$\text{Akurasi} = \frac{TP+TN}{TP+FN+FP+TN} \text{ x } 100\%$$
$$= \frac{112+44}{112+6+6+44} \text{ x } 100\%$$
$$= \frac{156}{72} \text{ x } 100\%$$
$$= 92,86\%$$

Dari perhitungan tersebut dapat disimpulkan bahwa ketepatan klasifikasi berdasarkan data *training* penyakit Diabetes Mellitus menggunakan model terbaik Regresi Logistik Biner sebesar 92,86%.

b. Ketepatan klasifikasi berdasarkan data *testing*

$$\text{Akurasi} = \frac{TP+TN}{TP+FN+FP+TN} \text{ x } 100\%$$
$$= \frac{48+18}{48+4+2+18} \text{ x } 100\%$$
$$= \frac{66}{72} \text{ x } 100\%$$
$$= 91,67\%$$

Dari perhitungan tersebut dapat disimpulkan bahwa ketepatan klasifikasi berdasarkan data *testing* penyakit Diabetes Mellitus menggunakan model terbaik Regresi Logistik Biner sebesar 91,67%.

Selanjutnya, akan dilakukan analisis klasifikasi menggunakan *Classification And Regression Trees* (CART).

4.3. Analisis Klasifikasi Menggunakan *Classification And Regression Trees* (CART)

Tahap awal yang dilakukan untuk pembentukan model CART adalah dengan menentukan pemilihan pemilah.

1. Pemilihan Pemilah

Pemilihan variabel pemilah yang terbaik dari kelima variabel independen

berdasarkan kriteria *goodness of split*. Suatu *split* s akan digunakan sebagai pemecah *node* t menjadi dua buah *node* yaitu *node* t_L dan t_R. Nilai *goodness of split* yang paling besar akan menjadi *root node*.

Dalam penelitian ini data *training* yang digunakan sebanyak 168 data, terdapat 50 data pasien negatif DM dan 118 data pasien positif DM.

Total data = 168 *record*

Negatif = 50 *record*

Positif = 118 *record*

Berdasarkan data tersebut akan dilakukan perhitungan *Gini Indeks* menggunakan Persamaan (2.18):

$$\begin{aligned}\text{Gini (D)} &= 1 - (\tfrac{50}{168})^2 - (\tfrac{118}{168})^2 \\ &= 1 - 0,08858 - 0,49334 \\ &= 0,41808\end{aligned}$$

Selanjutnya akan dilakukan perhitungan *Gini Indeks* juga pada masing-masing variabel X.

a. Pada variabel usia (X_1) memiliki dua kemungkinan nilai kejadian, yaitu ”Muda” (0) dan ”Tua” (1), masing-masing nilai kejadian dapat diuraikan sebagai berikut:

   1) Nilai kejadian ”Muda” (0) pada variabel X_1 = 102, sebanyak 41 kejadian terdapat pada *class* Negatif dan 61 kejadian pada *class* Positif. Besarnya *Gini Indeks* dari *node* ini (D_0) adalah sebagai berikut:

$$\begin{aligned}Gini(D_0) &= 1 - (\tfrac{41}{102})^2 - (\tfrac{61}{102})^2 \\ &= 1 - 0,16157 - 0,35765 \\ &= 0,48078\end{aligned}$$

2) Nilai kejadian ”Tua” (1) pada variabel X_1 = 66, sebanyak 9 kejadian terdapat pada *class* Negatif dan 57 kejadian pada *class* Positif. Besarnya *Gini Indeks* dari *node* ini (D_1) adalah sebagai berikut:

$$\begin{aligned} Gini(D_1) &= 1 - (\tfrac{9}{66})^2 - (\tfrac{57}{66})^2 \\ &= 1 - 0,01859 - 0,74586 \\ &= 0,23555 \end{aligned}$$

Kemudian, dilakukan perhitungan *Gini Split* dari variabel X_1 menggunakan Persamaan (2.19) sebagai berikut:

$$\begin{aligned} Gini_A(D) &= (\tfrac{102}{168} \text{ x } 0,48078) + (\tfrac{66}{168} \text{ x } 0,23553) \\ &= 0,38444 \end{aligned}$$

Selanjutnya, dilakukan perhitungan *decrease impurity* menggunakan Persamaan (2.20) sebagai berikut:

$$\begin{aligned} \Delta Gini(D) &= Gini(D) - Gini_A(D) \\ &= 0,41808 - 0,38444 \\ &= 0,03364 \end{aligned}$$

b. Pada variabel keturunan (X_2) memiliki dua kemungkinan nilai kejadian, yaitu ”Tidak Ada” (0) dan ”Ada”(1), masing-masing nilai kejadian dapat diuraikan sebagai berikut:

1) Nilai kejadian ”Tidak Ada” (0) pada variabel X_2 = 98, sebanyak 43 kejadian terdapat pada *class* Negatif dan 55 kejadian pada *class* Positif. Besarnya *Gini Indeks* dari *node* ini (D_0) adalah sebagai berikut:

$$\begin{aligned} Gini(D_0) &= 1 - (\tfrac{43}{98})^2 - (\tfrac{55}{98})^2 \\ &= 1 - 0,19252 - 0,31497 \\ &= 0,49251 \end{aligned}$$

2) Nilai kejadian ”Ada” (1) pada variabel X_2 = 70, sebanyak 7 kejadian

terdapat pada *class* Negatif dan 63 kejadian pada *class* Positif. Besarnya Indeks Gini dari *node* ini (D_1) adalah sebagai berikut:

$$\begin{aligned} Gini(D_1) &= 1 - (\tfrac{7}{70})^2 - (\tfrac{63}{70})^2 \\ &= 1 - 0,01 - 0,81 \\ &= 0,18 \end{aligned}$$

Kemudian, dilakukan perhitungan *Gini Split* dari variabel X_2 menggunakan Persamaan (2.19) sebagai berikut:

$$\begin{aligned} Gini_A(D) &= (\tfrac{98}{168} \text{ x } 0,49250) + (\tfrac{70}{168} \text{ x } 0,18) \\ &= 0,36229 \end{aligned}$$

Selanjutnya, dilakukan perhitungan *decrease impurity* menggunakan Persamaan (2.20) sebagai berikut:

$$\begin{aligned} \Delta Gini(D) &= Gini(D) - Gini_A(D) \\ &= 0,41808 - 0,36229 \\ &= 0,05579 \end{aligned}$$

c. Pada variabel gula darah (X_3) memiliki dua kemungkinan nilai kejadian, yaitu ”Normal” (0) dan ”Tidak Normal” (1), masing-masing nilai kejadian dapat diuraikan sebagai berikut:

1) Nilai kejadian ”Normal” (0) pada variabel X_3 = 62, sebanyak 38 kejadian terdapat pada *class* Negatif dan 24 kejadian pada *class* Positif. Besarnya *Gini Indeks* dari *node* ini (D_0) adalah sebagai berikut:

$$\begin{aligned} Gini(D_0) &= 1 - (\tfrac{38}{62})^2 - (\tfrac{24}{62})^2 \\ &= 0,37565 - 0,14984 \\ &= 0,47451 \end{aligned}$$

2) Nilai kejadian ”Tidak Normal” (1) pada variabel X_3 = 106, sebanyak 12 kejadian terdapat pada *class* Negatif dan 94 kejadian pada *class* Positif.

Besarnya *Gini Indeks* dari *node* ini (D_1) adalah sebagai berikut:

$$\begin{aligned} Gini(D_1) &= 1 - (\tfrac{12}{106})^2 - (\tfrac{94}{106})^2 \\ &= 1 - 0,01282 - 0,78640 \\ &= 0,20078 \end{aligned}$$

Kemudian, dilakukan perhitungan *Gini Split* dari variabel X_3 menggunakan Persamaan (2.19) sebagai berikut:

$$\begin{aligned} Gini_A(D) &= (\tfrac{62}{168} \text{ x } 0,47451) + (\tfrac{106}{168} \text{ x } 0,20078) \\ &= 0,3018 \end{aligned}$$

Selanjutnya, dilakukan perhitungan *decrease impurity* menggunakan Persamaan (2.20) sebagai berikut:

$$\begin{aligned} \Delta Gini(D) &= Gini(D) - Gini_A(D) \\ &= 0,41808 - 0,3018 \\ &= 0,11628 \end{aligned}$$

d. Pada variabel obesitas (X_4) memiliki dua kemungkinan nilai kejadian, yaitu ”Normal” (0) dan ”Tidak Normal” (1), masing-masing nilai kejadian dapat diuraikan sebagai berikut:

   1) Nilai kejadian ”Normal” (0) pada variabel X_4 = 31, sebanyak 22 kejadian terdapat pada *class* Negatif dan 9 kejadian pada *class* Positif. Besarnya *Gini Indeks* dari *node* ini (D_0) adalah sebagai berikut:

$$\begin{aligned} Gini(D_0) &= 1 - (\tfrac{22}{31})^2 - (\tfrac{9}{31})^2 \\ &= 1 - 0,50364 - 0,08429 \\ &= 0,41207 \end{aligned}$$

   2) Nilai kejadian ”Tidak Normal” (1) pada variabel X_4 = 137, sebanyak 28 kejadian terdapat pada *class* Negatif dan 109 kejadian pada *class* Positif. Besarnya *Gini Indeks* dari *node* ini (D_1) adalah sebagai berikut:

$$Gini(D_1) = 1 - (\frac{28}{137})^2 - (\frac{109}{137})^2$$
$$= 1 - 0,04177 - 0,63301$$
$$= 0,32522$$

Kemudian, dilakukan perhitungan *Gini Split* dari variabel X_4 menggunakan Persamaan (2.19) sebagai berikut:

$$Gini_A(D) = (\frac{31}{168} \text{ x } 0,41207) + (\frac{137}{168} \text{ x } 0,32522)$$
$$= 0,34124$$

Selanjutnya, dilakukan perhitungan *decrease impurity* menggunakan Persamaan (2.20) sebagai berikut:

$$\Delta Gini(D) = Gini(D) - Gini_A(D)$$
$$= 0,41808 - 0,34124$$
$$= 0,07684$$

e. Pada variabel pola makan (X_5) memiliki dua kemungkinan nilai kejadian, yaitu ”Baik” (0) dan ”Buruk” (1), masing-masing nilai kejadian dapat diuraikan sebagai berikut:

1) Nilai kejadian ”Baik” (0) pada variabel X_5 = 67, sebanyak 14 kejadian terdapat pada *class* Negatif dan 53 kejadian pada *class* Positif. Besarnya *Gini Indeks* dari *node* ini (D_0) adalah sebagai berikut:

$$Gini(D_0) = 1 - (\frac{14}{67})^2 - (\frac{53}{67})^2$$
$$= 1 - 0,04366 - 0,62575$$
$$= 0,33059$$

2) Nilai kejadian ”Tidak Normal” (1) pada variabel X_5 = 101, sebanyak 36 kejadian terdapat pada *class* Negatif dan 65 kejadian pada *class* Positif. Besarnya *Gini Indeks* dari *node* ini (D_1) adalah sebagai berikut:

$$Gini(D_1) = 1 - (\frac{36}{101})^2 - (\frac{65}{101})^2$$

$$= 1 - 0,12705 - 0,41417$$

$$= 0,45878$$

Kemudian, dilakukan perhitungan *Gini Split* dari variabel X_5 menggunakan Persamaan (2.19) sebagai berikut:

$$Gini_A(D) = (\tfrac{67}{168} \text{ x } 0,33059) + (\tfrac{101}{168} \text{ x } 0,45878)$$

$$= 0,40765$$

Selanjutnya, dilakukan perhitungan *decrease impurity* menggunakan Persamaan (2.20) sebagai berikut:

$$\Delta Gini(D) = Gini(D) - Gini_A(D)$$

$$= 0,41808 - 0,40765$$

$$= 0,01043$$

Dari perhitungan tersebut, kriteria pemilihan pemilah iterasi pertama disajikan pada Tabel 4.8.

Tabel 4.8 Kriteria Pemilihan Pemilah Iterasi Pertama

X	L				R				Gini A	Δ Gini
	Record	Negatif	Positif	Gini	Record	Negatif	Positif	Gini		
Usia	102	41	61	0.48078	66	9	57	0.23554	0.38443	0.03365
Keturunan	98	43	55	0.4925	70	7	63	0.18	0.36229	0.05579
Gula Darah	62	38	24	0.47451	106	12	94	0.20078	0.3018	0.11628
Obesitas	31	22	9	0.41207	137	28	109	0.32522	0.34124	0.07684
Pola Makan	67	14	53	0.33059	101	36	65	0.45878	0.40765	0.01043

Berdasarkan Tabel 4.8, ”L” merupakan calon cabang *node* kiri dan ”R” merupakan calon cabang *node* kanan. Dari kelima variabel tersebut nilai $\Delta Gini$

X_3 paling besar sehingga variabel Gula Darah dipilih menjadi pemilah pertama. Pasien dengan gula darah normal (0) ditetapkan menjadi calon cabang kiri sedangkan pasien dengan gula darah tidak normal (1) ditetapkan menjadi calon cabang kanan.

a. Gula Darah normal (0) dan Gula Darah tidak normal (1)

1) Gula Darah Normal (0):

Total data = 62 *record*

Negatif = 38 *record*

Positif = 24 *record*

$$\begin{aligned} Gini(D_0) &= 1 - (\tfrac{38}{62})^2 - (\tfrac{24}{62})^2 \\ &= 1 - 0,37565 - 0,14984 \\ &= 0,47451 \end{aligned}$$

2) Gula Darah Tidak Normal (1):

Total data = 106 *record*

Negatif = 12 *record*

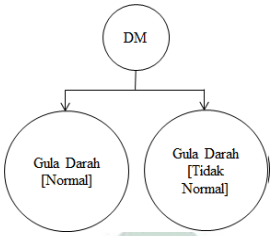
Positif = 94 *record*

$$\begin{aligned} Gini(D_1) &= 1 - (\tfrac{12}{106})^2 - (\tfrac{94}{106})^2 \\ &= 1 - 0,01282 - 0,78640 \\ &= 0,20078 \end{aligned}$$

$$\begin{aligned} Gini_A(D) &= (\tfrac{24}{106}) \cdot 0,47451 + (\tfrac{94}{106}) \cdot 0,20078 \\ &= 0,28549 \end{aligned}$$

$$\begin{aligned} \Delta Gini(D) &= 0,41808 - 0,28549 \\ &= 0,13259 \end{aligned}$$

Dengan demikian, pohon yang terbentuk pada iterasi pertama seperti Gambar 4.7.



Gambar 4.7 Pohon Hasil Iterasi Pertama

Selanjutnya dilakukan tahap rekursif pada cabang kiri. Data yang digunakan pada iterasi kedua ini berdasarkan data cabang kiri dari iterasi pertama yang diperoleh nilai *Gini Indeks* = 0,47451.

Kemudian akan dilakukan perhitungan *Gini Indeks* juga pada masing-masing variabel X seperti pada langkah iterasi pertama. Dari perhitungan tersebut, kriteria pemilihan pemilah iterasi kedua disajikan pada Tabel 4.9.

Tabel 4.9 Kriteria Pemilihan Pemilah Iterasi Kedua

X	L				R				Gini A	Δ Gini
	Record	Negatif	Positif	Gini	Record	Negatif	Positif	Gini		
Usia	42	31	11	0,38662	20	7	13	0,455	0,40868	0,06583
Keturunan	35	31	4	0,20245	27	7	20	0,38409	0,28155	0,19296
Gula Darah	62	38	24	0,47451	0	0	0	0	0,47451	0
Obesitas	17	15	2	0,20761	45	23	22	0,49975	0,41965	0,05486
Pola Makan	16	8	8	0,5	46	30	16	0,45369	0,46564	0,00887

Berdasarkan Tabel 4.9, dari kelima variabel tersebut nilai $\Delta Gini\ X_2$ paling besar sehingga variabel Keturunan dipilih menjadi pemilah kedua. Pasien yang

tidak mempunyai riwayat keturunan DM (0) ditetapkan menjadi calon cabang kiri sedangkan pasien yang mempunyai riwayat keturunan DM (1) ditetapkan menjadi calon cabang kanan.

a. Tidak Ada Keturunan (0) dan Ada Keturunan (1)

1) Tidak Ada Keturunan (0):

Total data = 35 *record*

Negatif = 31 *record*

Positif = 4 *record*

$$Gini(D_0) = 1 - (\tfrac{31}{35})^2 - (\tfrac{4}{35})^2$$
$$= 1 - 0,78449 - 0,01306$$
$$= 0,20245$$

2) Ada Keturunan (1):

Total data = 27 *record*

Negatif = 7 *record*

Positif = 27 *record*

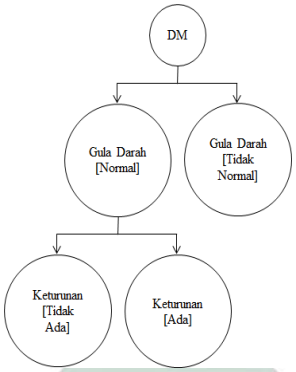
$$Gini(D_1) = 1 - (\tfrac{7}{27})^2 - (\tfrac{20}{27})^2$$
$$= 1 - 0,06722 - 0,54870$$
$$= 0,38408$$

$$Gini_A(D) = (\tfrac{35}{62}) \cdot 0,20245 + (\tfrac{27}{62}) \cdot 0,38408$$
$$= 0,28155$$

$$\Delta Gini(D) = 0,47451 - 0,28155$$
$$= 0,19296$$

Dengan demikian, pohon yang terbentuk pada iterasi kedua seperti Gambar

4.8.



Gambar 4.8 Pohon Hasil Iterasi Kedua

Selanjutnya dilakukan tahap rekursif pada cabang kiri. Data yang digunakan pada iterasi ketiga ini berdasarkan data cabang kiri dari iterasi kedua yang diperoleh *Gini Indeks* = 0,20245.

Kemudian, akan dilakukan perhitungan *Gini Indeks* juga pada masing-masing variabel X seperti pada langkah sebelumnya. Dari perhitungan tersebut, kriteria pemilihan pemilah iterasi ketiga disajikan pada Tabel 4.10.

Tabel 4.10 Kriteria Pemilihan Pemilah Iterasi Ketiga

X	L				R				Gini A	Δ Gini
	Record	Negatif	Positif	Gini	Record	Negatif	Positif	Gini		
Usia	27	24	3	0,19753	8	7	1	0,21875	0,20238	6,8E-05
Keturunan	35	31	4	0,20245	0	0	0	0	0,20245	0
Gula Darah	35	31	4	0,20245	0	0	0	0	0,20245	0
Obesitas	9	9	0	0	26	22	4	0,26036	0,19341	0,00904
Pola Makan	7	4	3	0,4898	28	27	1	0,06888	0,15306	0,04939

Berdasarkan Tabel 4.10, dari kelima variabel tersebut nilai $\Delta Gini\ X_5$ paling besar sehingga variabel Pola Makan dipilih menjadi pemilah ketiga. Pasien

dengan Pola Makan yang Baik (0) ditetapkan menjadi calon cabang kiri sedangkan pasien dengan Pola Makan yang Buruk (1) ditetapkan menjadi calon cabang kanan.

a. Pola Makan Baik (0) dan Pola Makan Buruk (1)

1) Pola Makan Baik (0):

Total data = 7 *record*

Negatif = 4 *record*

Positif = 3 *record*

$$Gini(D_0) = 1 - (\tfrac{4}{7})^2 - (\tfrac{3}{7})^2$$
$$= 1 - 0,32653 - 0,18367$$
$$= 0,4898$$

2) Pola Makan Buruk (1):

Total data = 28 *record*

Negatif = 27 *record*

Positif = 1 *record*

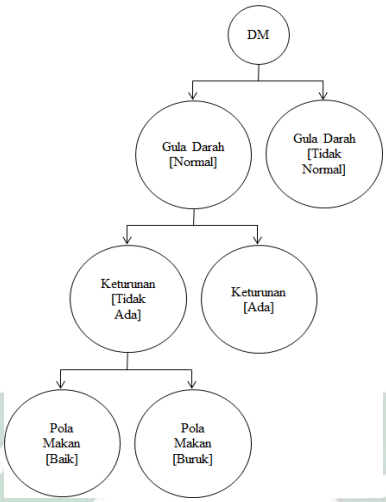
$$Gini(D_1) = 1 - (\tfrac{27}{28})^2 - (\tfrac{1}{28})^2$$
$$= 1 - 0,92985 - 0,00128$$
$$= 0,06887$$

$$Gini_A(D) = (\tfrac{7}{35}) \cdot 0,4898 + (\tfrac{28}{35}) \cdot 0,06887$$
$$= 0,15306$$

$$\Delta Gini(D) = 0,20245 - 0,15306$$
$$= 0,04939$$

Dengan demikian, pohon yang terbentuk pada iterasi ketiga seperti Gambar

4.9.



Gambar 4.9 Pohon Hasil Iterasi Ketiga

Selanjutnya dilakukan tahap rekursif pada cabang kiri. Data yang digunakan pada iterasi keempat ini berdasarkan data cabang kiri dari iterasi ketiga yang diperoleh *Gini Indeks* = 0,4898.

Kemudian, akan dilakukan perhitungan *Gini Indeks* juga pada masing-masing variabel X seperti langkah sebelumnya. Dari perhitungan tersebut, kriteria pemilihan pemilah iterasi keempat disajikan pada Tabel 4.11.

Tabel 4.11 Kriteria Pemilihan Pemilah Iterasi Keempat

X	L				R				Gini A	Δ Gini
	Record	Negatif	Positif	Gini	Record	Negatif	Positif	Gini		
Usia	4	2	2	0,5	3	2	1	0,44444	0,47619	0,01361
Keturunan	7	4	3	0,4898	0	0	0	0	0,4898	0
Gula Darah	7	4	3	0,4898	0	0	0	0	0,4898	0
Obesitas	2	2	0	0	5	2	3	0,48	0,34286	0,14694
Pola Makan	7	4	3	0,4898	0	0	0	0	0,4898	0

Berdasarkan Tabel 4.11, dari kelima variabel tersebut nilai $\Delta Gini\ X_4$ paling besar sehingga variabel Obesitas dipilih menjadi pemilah keempat. Pasien dengan Obesitas Normal (0) ditetapkan menjadi calon cabang kiri sedangkan pasien dengan Obesitas Tidak Normal (1) ditetapkan menjadi calon cabang kanan.

a. Obesitas Normal (0) dan Obesitas Tidak Normal (1)

1) Pola Makan Baik (0):

Total data = 2 *record*

Negatif = 2 *record*

Positif = 0 *record*

$$Gini(D_0) = 1 - (\tfrac{2}{2})^2 - (\tfrac{0}{2})^2$$
$$= 1 - 1 - 0$$
$$= 0$$

2) Obesitas Tidak Normal (1):

Total data = 5 *record*

Negatif = 2 *record*

Positif = 3 *record*

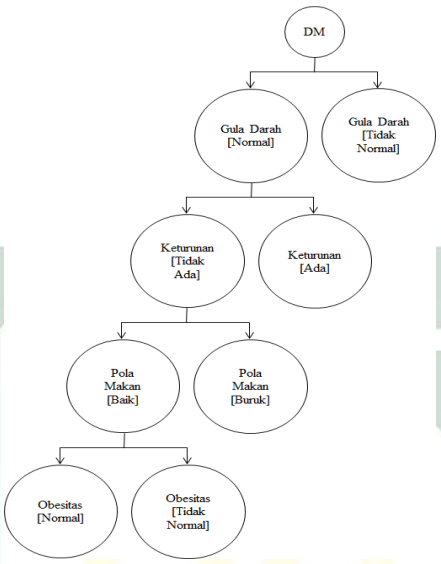
$$Gini(D_1) = 1 - (\tfrac{2}{5})^2 - (\tfrac{3}{5})^2$$
$$= 1 - 0,16 - 0,36$$
$$= 0,48$$

$$Gini_A(D) = (\tfrac{2}{7}) \cdot 0 + (\tfrac{5}{7}) \cdot 0,48$$
$$= 0,34286$$

$$\Delta Gini(D) = 0,4898 - 0,34286$$

$$= 0,14694$$

Dengan demikian, pohon yang terbentuk pada iterasi keempat seperti Gambar 4.10.



Gambar 4.10 Pohon Hasil Iterasi Keempat

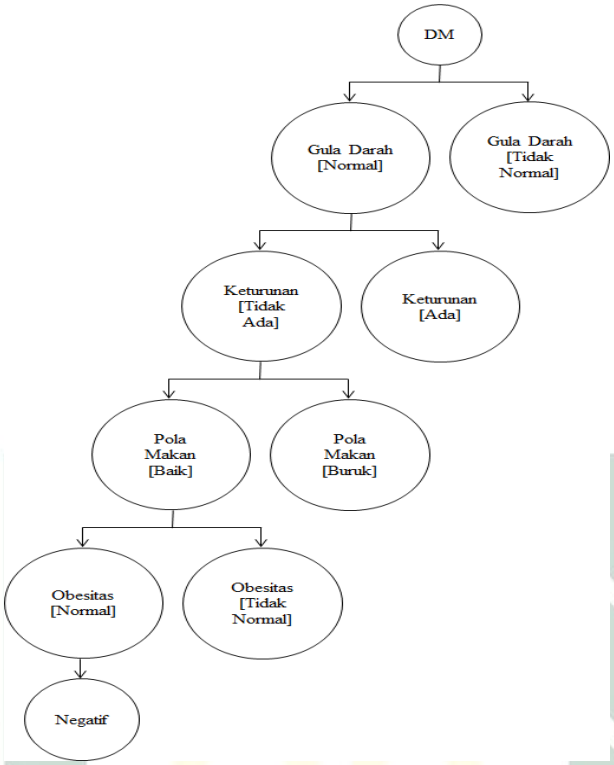
Selanjutnya dilakukan tahap rekursif pada cabang kiri dengan nilai *Gini Indeks* dari Obesitas Normal= 0.

Kemudian akan dilakukan perhitungan *Gini Indeks* juga pada masing-masing variabel X seperti pada langkah sebelumnya. Dari perhitungan tersebut, kriteria pemilihan pemilah iterasi kelima disajikan pada Tabel 4.12.

Tabel 4.12 Kriteria Pemilihan Pemilah Iterasi Kelima

X	L				R				Gini A	Δ Gini
	Record	Negatif	Positif	Gini	Record	Negatif	Positif	Gini		
Usia	2	2	0	0	0	0	0	0	0	0
Keturunan	2	2	0	0	0	0	0	0	0	0
Gula Darah	2	2	0	0	0	0	0	0	0	0
Obesitas	2	2	0	0	0	0	0	0	0	0
Pola Makan	2	2	0	0	0	0	0	0	0	0

Berdasarkan Tabel 4.12, dari kelima variabel tersebut nilai $\Delta Gini$ = 0 sehingga proses rekursif pada cabang kiri berhenti dan ditetapkan sebagai *class* Negatif, dikarenakan pada iterasi kelima data hanya terdiri dari *class* Negatif. Dengan demikian, pohon yang terbentuk dari iterasi kelima seperti Gambar 4.11.



Gambar 4.11 Pohon Hasil Iterasi Kelima

Selanjutnya akan dilakukan proses rekursif pada variabel Obesitas cabang kanan dengan nilai *Gini Indeks* = 0,48.

Kemudian akan dilakukan perhitungan *Gini Indeks* juga pada masing-masing variabel X seperti pada langkah sebelumnya. Dari perhitungan tersebut, kriteria pemilihan pemilah iterasi keenam disajikan pada Tabel 4.13.

Tabel 4.13 Kriteria Pemilihan Pemilah Iterasi Keenam

X	L				R				Gini A	Δ Gini
	Record	Negatif	Positif	Gini	Record	Negatif	Positif	Gini		
Usia	2	0	2	0	3	2	1	0,44444	0,26667	0,21333
Keturunan	5	2	3	0,48	0	0	0	0	0,48	0
Gula Darah	5	2	3	0,48	0	0	0	0	0,48	0
Obesitas	0	0	0	0	5	2	3	0,48	0,48	0
Pola Makan	5	2	3	0,48	0	0	0	0	0,48	0

Berdasarkan Tabel 4.13, dari kelima variabel tersebut nilai $\Delta Gini\ X_1$ paling besar sehingga variabel Usia dipilih menjadi pemilah kelima. Pasien dengan Usia Muda (0) ditetapkan menjadi calon cabang kiri sedangkan pasien dengan Usia Tua (1) ditetapkan menjadi calon cabang kanan.

a. Usia Muda (0) dan Usia Tua (1)

1) Usia Muda (0):

Total data = 2 *record*

Negatif = 0 *record*

Positif = 2 *record*

$$Gini(D_0) = 1 - (\tfrac{0}{2})^2 - (\tfrac{2}{2})^2$$
$$= 1 - 0 - 1$$
$$= 0$$

2) Usia Tua (1):

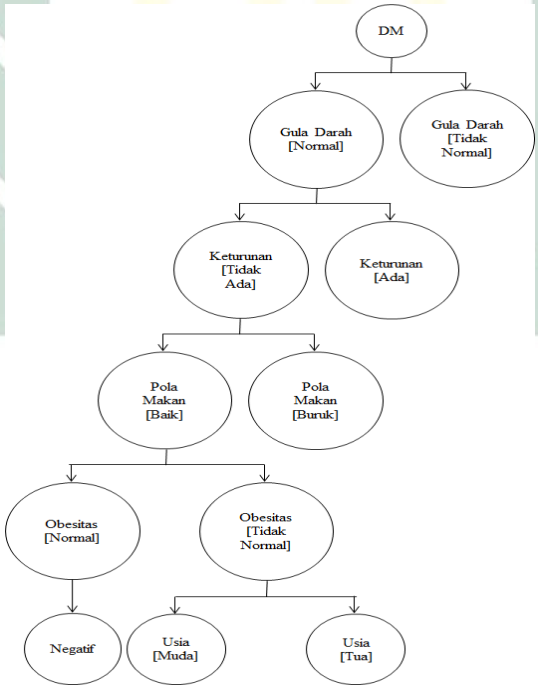
Total data = 3 *record*

Negatif = 2 *record*

Positif = 1 *record*

$$Gini(D_1) = 1 - (\frac{2}{3})^2 - (\frac{1}{3})^2$$
$$= 1 - 0,44444 - 0,11111$$
$$= 0,44445$$
$$Gini_A(D) = (\frac{2}{5}) \cdot 0 + (\frac{3}{5}) \cdot 0,44445$$
$$= 0,26667$$
$$\Delta Gini(D) = 0,48 - 0,26667$$
$$= 0,21333$$

Dengan demikian, pohon yang terbentuk pada iterasi keenam seperti Gambar 4.12.



Gambar 4.12 Pohon Hasil Iterasi Keenam

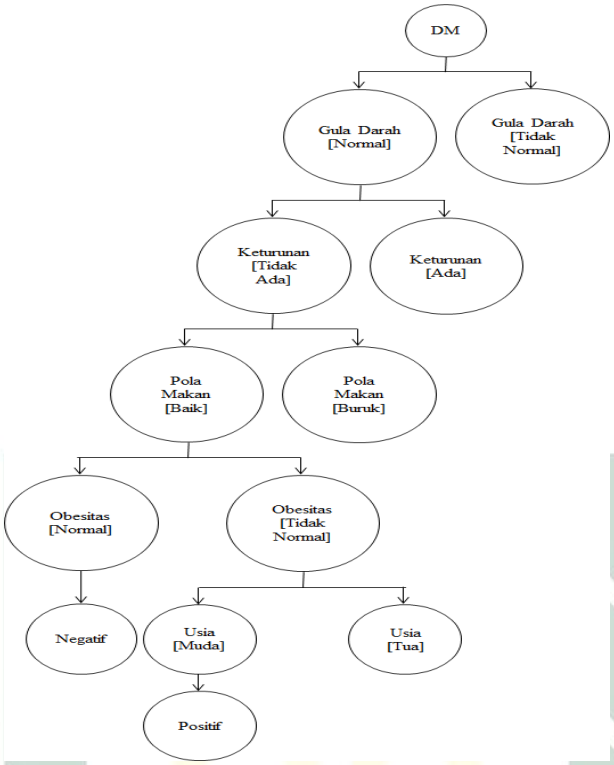
Selanjutnya dilakukan tahap rekursif pada cabang kiri dengan nilai *Gini Indeks* dari Usia Muda = 0.

Kemudian akan dilakukan perhitungan *Gini Indeks* juga pada masing-masing variabel X seperti pada langkah sebelumnya. Dari perhitungan tersebut, kriteria pemilihan pemilah iterasi ketujuh disajikan pada Tabel 4.14.

Tabel 4.14 Kriteria Pemilihan Pemilah Iterasi Ketujuh

X	L				R				Gini A	Δ Gini
	Record	Negatif	Positif	Gini	Record	Negatif	Positif	Gini		
Usia	2	0	2	0	0	0	0	0	0	0
Keturunan	2	0	2	0	0	0	0	0	0	0
Gula Darah	2	0	2	0	0	0	0	0	0	0
Obesitas	0	0	0	0	2	0	2	0	0	0
Pola Makan	2	0	2	0	0	0	0	0	0	0

Berdasarkan Tabel 4.14, dari kelima variabel tersebut nilai $\Delta Gini$ = 0 sehingga proses rekursif pada cabang kiri berhenti dan ditetapkan sebagai *class* Positif, dikarenakan pada iterasi ketujuh data hanya terdiri dari *class* Positif. Dengan demikian, pohon yang terbentuk dari iterasi ketujuh seperti Gambar 4.13.



Gambar 4.13 Pohon Hasil Iterasi Ketujuh

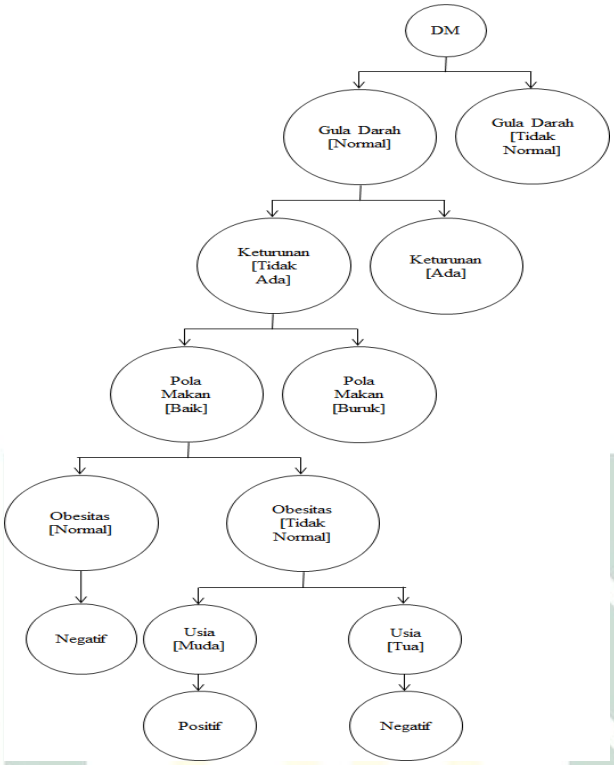
Selanjutnya akan dilakukan proses rekursif pada cabang kanan dengan nilai *Gini Indeks* dari Usia Tua = 0,44445.

Kemudian akan dilakukan perhitungan *Gini Indeks* juga pada masing-masing variabel X seperti pada langkah sebelumnya. Dari perhitungan tersebut, kriteria pemilihan pemilah iterasi ketujuh disajikan pada Tabel 4.15.

Tabel 4.15 Kriteria Pemilihan Pemilah Iterasi Kedelapan

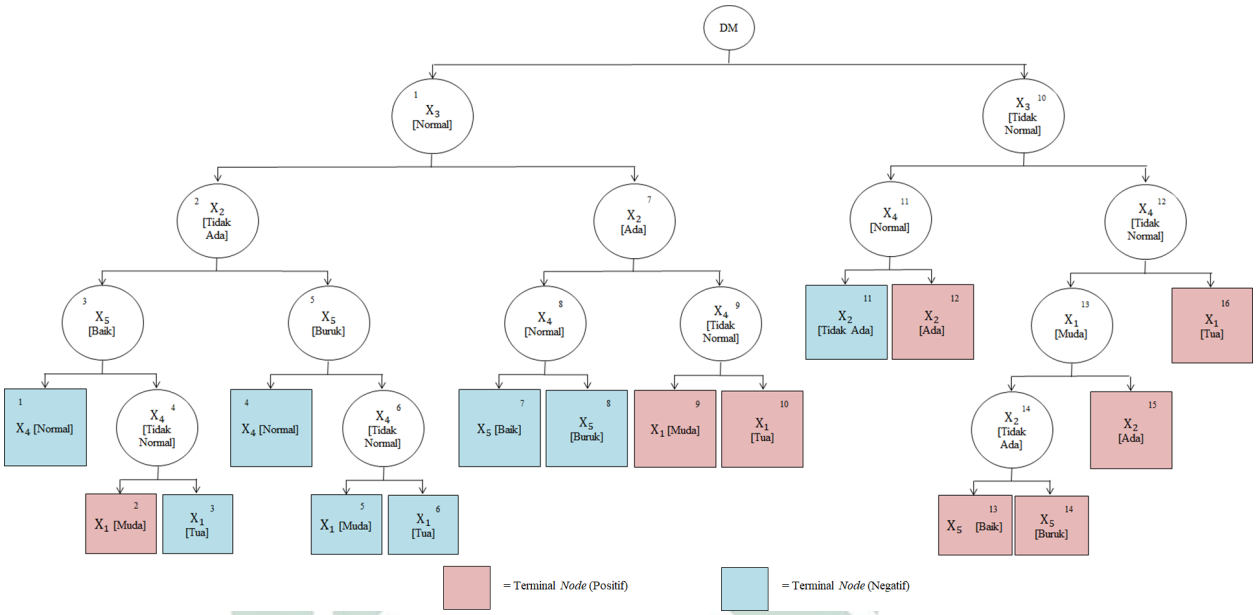
X	L				R				Gini A	Δ Gini
	Record	Negatif	Positif	Gini	Record	Negatif	Positif	Gini		
Usia	0	0	0	0	3	2	1	0,44444	0,44444	0
Keturunan	3	2	1	0,44444	0	0	0	0	0,44444	0
Gula Darah	3	2	1	0,44444	0	0	0	0	0,44444	0
Obesitas	0	0	0	0	3	2	1	0,44444	0,44444	0
Pola Makan	3	2	1	0,44444	0	0	0	0	0,44444	0

Berdasarkan Tabel 4.15, dari kelima variabel tersebut nilai $\Delta Gini = 0$ sehingga proses rekursif pada cabang kanan berhenti dan ditetapkan sebagai *class* Negatif, dikarenakan pada iterasi kedelapan data *record* < 5 serta peluang *class* Negatif lebih besar dari pada *class* Positif yaitu 0,67 : 0,33. Dengan demikian, pohon yang terbentuk dari iterasi kedelapan seperti Gambar 4.14.



Gambar 4.14 Pohon Hasil Iterasi Kedelapan

Proses rekursif ini berlanjut hingga semua variabel X baik cabang kiri maupun kanan telah masuk dalam pohon klasifikasi maksimal. Setelah seluruh variabel X dilakukan tahap rekursif, diperoleh pohon klasifikasi maksimal pada iterasi ke-30 seperti Gambar 4.15.

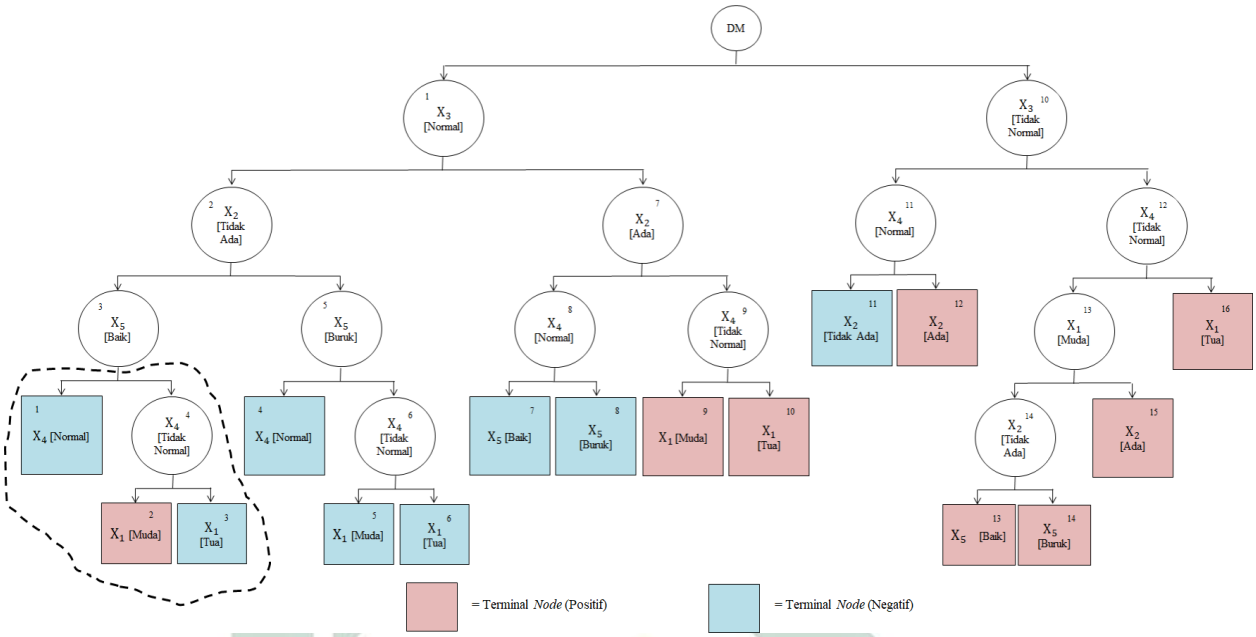


Gambar 4.15 Pohon Hasil Iterasi Ke-30

Berdasarkan Gambar 4.15, pohon klasifikasi maksimal yang diperoleh dari tahap rekursif terdiri dari 16 terminal *node*, 14 *inner node* dan tingkat kedalaman sebanyak 5. Pohon klasifikasi yang dihasilkan tersebut sangat besar sehingga akan dilakukan tahap *tree pruning*.

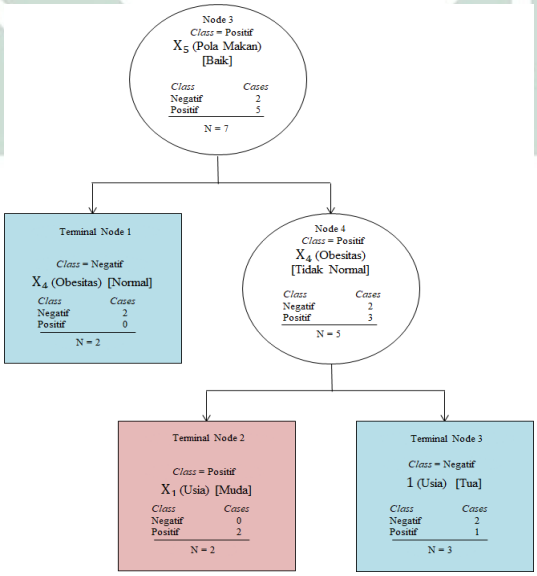
2. Pemangkasan Pohon Klasifikasi Maksimal (*Tree Prunning*)

Tahap *Tree Pruning* dilakukan dengan pengambilan t_L (*node* kiri) dan t_R (*node* kanan) dari T_max yang dihasilkan oleh *root node t*. Apabila dua *chile node* dan *root node* yang memenuhi persamaan $R(t) = R(t_L) + R(t_R)$, maka *chile node* t_L dan t_R dipangkas.



Gambar 4.16 Pohon Klasifikasi Maksimal yang akan Dipangkas

Pada Gambar 4.16, *node* pertama yang akan dipangkas yaitu *node* 3. *Node* yang akan dipangkas ditunjukkan pada Gambar 4.17.



Gambar 4.17 *Node* Pertama yang akan Dipangkas

Tahap perhitungan *prunning* sebagai berikut:

a. *Root Node* 3:

Dalam *node* 3 terdiri dari 2 *class* negatif dan 5 *class* positif. Nilai $R(t)$ diperoleh dari $r(t) \cdot P(t)$, dimana $r(t) = 1 - max_j P(j|t)$. Probabilitas tiap *class* dalam *node* 3 (*t*) sebagai berikut:

$$\text{P(Negatif} \mid \text{t)} = \frac{2}{7} = 0,28571$$

$$\text{P(Positif} \mid \text{t)} = \frac{5}{7} = 0,71429$$

Sehingga diperoleh

$$\text{r(t)} = 1 - 0,71429 = 0,28571$$

Nilai probabilitas objek dalam *node* 3 adalah sebagai berikut:

$$\text{P(t)} = \frac{N(node\ 3)}{N} = \frac{7}{168} = 0,04167$$

Dengan demikian,

$$\text{R(t)} = \text{r (t)} \times \text{P(t)} = 0{,}28571 \times 0{,}04167 = 0{,}01190$$

b. Terminal *Node* 1 dan *Node* 4

∗ Terminal *Node* 1

Probabilitas tiap *class* dalam terminal *node* 1 (t_L) sebagai berikut:

$$\text{P(Negatif} \mid t_L) = \frac{2}{2} = 1$$

$$\text{P(Positif} \mid t_L) = \frac{2}{2} = 0$$

Sehingga diperoleh

$$\text{r}(t_L) = 1 - 1 = 0$$

Nilai probabilitas objek dalam terminal *node* 1 adalah sebagai berikut:

$$\text{P}(t_L) = \frac{N(terminal\ node\ 1)}{N} = \frac{2}{168} = 0,01190$$

Dengan demikian,

$$\text{R}(t_L) = \text{r}\ (t_L) \times \text{P}(t_L) = 0 \times 0{,}01190 = 0$$

* *Node* 4 Probabilitas tiap *class* dalam *node* 4 (t_R) sebagai berikut:

$$\text{P(Negatif } |t_R) = \frac{2}{5} = 0,4$$

$$\text{P(Positif } |t_R) = \frac{3}{5} = 0,6$$

Sehingga diperoleh

$$\text{r}(t_R) = 1 - 0,6 = 0,4$$

Nilai probabilitas objek dalam *node* 4 adalah sebagai berikut:

$$\text{P}(t_R) = \frac{N(node\ 4)}{N} = \frac{5}{168} = 0,02976$$

Dengan demikian,

$$\text{R}(t_R) = \text{r}\ (t_R) \times \text{P}(t_R) = 0,4 \times 0,02976 = 0,01190$$

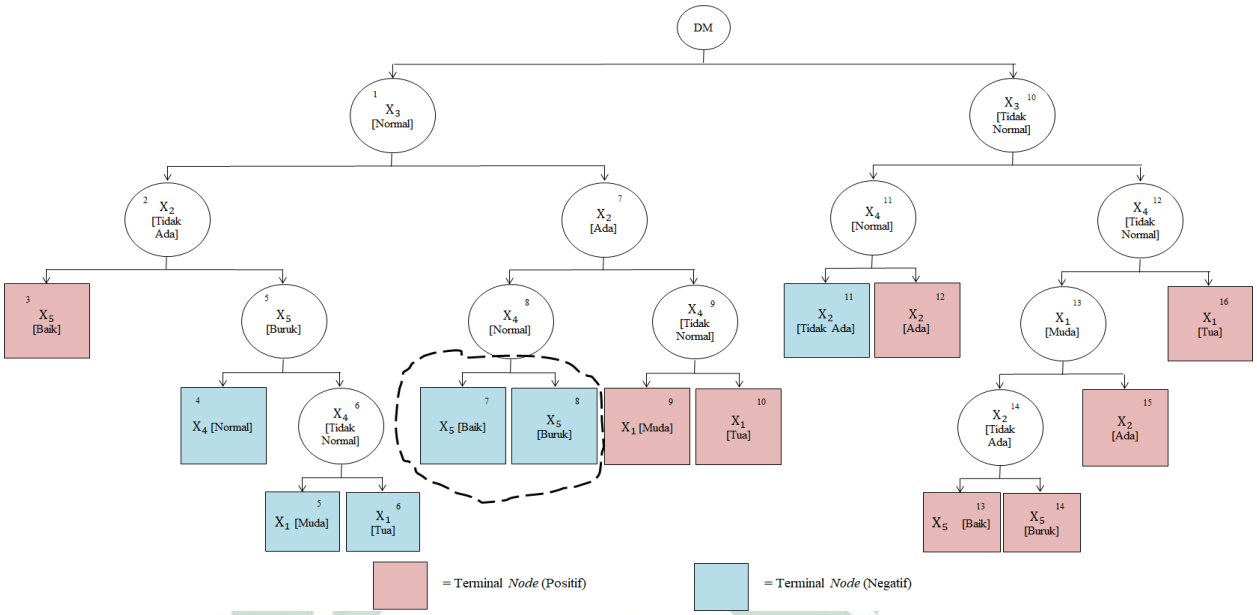
Oleh karena itu, persamaan

$$\text{R(t)} = \text{R}(t_L) + \text{R}(t_R)$$

$$0,01190 = 0 + 0,01190$$

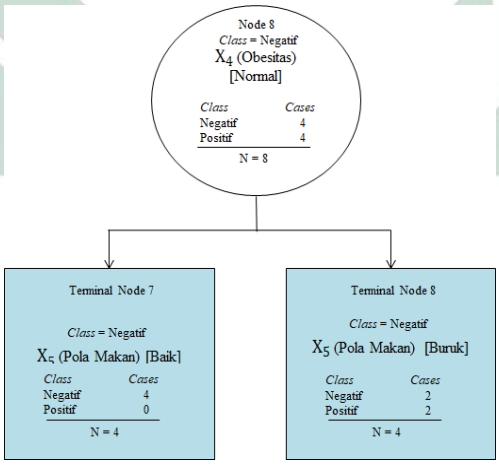
$$0,01190 = 0,01190$$

terpenuhi, sehingga dilakukan pemangkasan pada *node* 3 dan diperoleh pohon klasifikasi seperti Gambar 4.18:



Gambar 4.18 Pohon Klasifikasi Setelah Pemangkasan *Node* 3

Selanjutnya, dilakukan *prunning* pada *node* 8. *Node* yang akan dipangkas ditunjukkan pada Gambar 4.19.



Gambar 4.19 *Node* Kedua yang akan Dipangkas

Tahap perhitungan *prunning* sebagai berikut:

a. *Root Node* 8:

Dalam *node* 8 terdiri dari 4 *class* negatif dan 4 *class* positif. Probabilitas tiap *class* dalam *node* 8 (*t*) sebagai berikut:

$$\text{P(Negatif} \mid \text{t)} = \frac{4}{8} = 0,5$$

$$\text{P(Positif} \mid \text{t)} = \frac{4}{8} = 0,5$$

Sehingga diperoleh

$$\text{r(t)} = 1 - 0,5 = 0,5$$

Nilai probabilitas objek dalam *node* 8 adalah sebagai berikut:

$$\text{P(t)} = \frac{N(node\ 8)}{N} = \frac{8}{168} = 0,04762$$

Dengan demikian,

$$\text{R(t)} = \text{r (t)} \times \text{P(t)} = 0,5 \times 0,04762 = 0,02381$$

b. Terminal *Node* 7 dan Terminal *Node* 8

∗ Terminal *Node* 7

Probabilitas tiap *class* dalam terminal *node* 7 (t_L) sebagai berikut:

$$\text{P(Negatif} \mid t_L) = \frac{4}{4} = 1$$

$$\text{P(Positif} \mid t_L) = \frac{0}{4} = 0$$

Sehingga diperoleh

$$\text{r}(t_L) = 1 - 1 = 0$$

Nilai probabilitas objek dalam terminal *node* 7 adalah sebagai berikut:

$$\text{P}(t_L) = \frac{N(terminal\ node\ 7)}{N} = \frac{4}{168} = 0,02381$$

Dengan demikian,

$$\text{R}(t_L) = \text{r}\ (t_L) \times \text{P}(t_L) = 0 \times 0,02381 = 0$$

∗ Terminal *Node* 8 Probabilitas tiap *class* dalam terminal *node* 8 (t_R) sebagai berikut:

$$\text{P(Negatif } |t_R) = \frac{2}{4} = 0,5$$

$$\text{P(Positif } |t_R) = \frac{2}{4} = 0,5$$

Sehingga diperoleh

$$\text{r}(t_R) = 1 - 0,5 = 0,5$$

Nilai probabilitas objek dalam terminal *node* 8 adalah sebagai berikut:

$$\text{P}(t_R) = \frac{N(terminal\ node\ 8)}{N} = \frac{4}{168} = 0,4762$$

Dengan demikian,

$$\text{R}(t_R) = \text{r}\ (t_R) \times \text{P}(t_R) = 0,5 \times 0,4762 = 0,02381$$

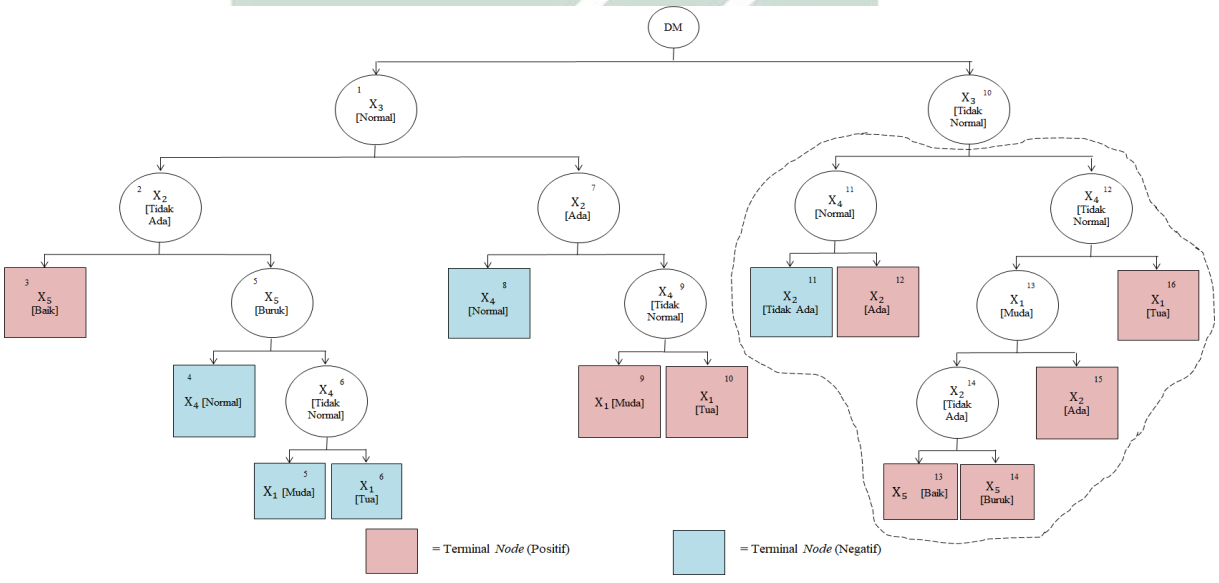
Oleh karena itu, persamaan

$$\text{R(t)} = \text{R}(t_L) + \text{R}(t_R)$$

$$0,02381 = 0 + 0,02381$$

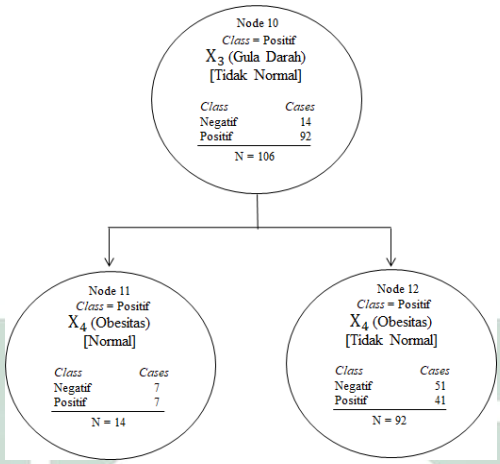
$$0,02381 = 0,02381$$

terpenuhi, sehingga dilakukan pemangkasan pada *node* 8 dan diperoleh pohon klasifikasi seperti Gambar 4.20:



Gambar 4.20 Pohon Klasifikasi Setelah Pemangkasan *Node* 8

Selanjutnya, dilakukan *prunning* pada *node* 10. *Node* yang akan dipangkas ditunjukkan pada Gambar 4.21.



Gambar 4.21 *Node* Ketiga yang akan Dipangkas

Tahap perhitungan *prunning* sebagai berikut:

a. *Root Node* 10:

Dalam *node* 10 terdiri dari 14 *class* negatif dan 92 *class* positif. Probabilitas tiap *class* dalam *node* 10 (*t*) sebagai berikut:

$$\text{P(Negatif} \mid \text{t)} = \frac{14}{106} = 0,13208$$

$$\text{P(Positif} \mid \text{t)} = \frac{92}{106} = 0,86792$$

Sehingga diperoleh

$$\text{r(t)} = 1 - 0,86792 = 0,13208$$

Nilai probabilitas objek dalam *node* 10 adalah sebagai berikut:

$$\text{P(t)} = \frac{N(node\ 10)}{N} = \frac{106}{168} = 0,63095$$

Dengan demikian,

$$\text{R(t)} = \text{r (t)} \times \text{P(t)} = 0{,}13208 \times 0{,}63095 = 0{,}08333$$

b. *Node* 11 dan *Node* 12

* *Node* 11

Probabilitas tiap *class* dalam *node* 11 (t_L) sebagai berikut:

$$\text{P(Negatif } |t_L) = \frac{7}{14} = 0,5$$

$$\text{P(Positif } |t_L) = \frac{7}{14} = 0,5$$

Sehingga diperoleh

$$\text{r}(t_L) = 1 - 0,5 = 0,5$$

Nilai probabilitas objek dalam *node* 11 adalah sebagai berikut:

$$\text{P}(t_L) = \frac{N(terminal\ node\ 11)}{N} = \frac{14}{168} = 0,08333$$

Dengan demikian,

$$\text{R}(t_L) = \text{r}\ (t_L) \times \text{P}(t_L) = 0,5 \times 0,08333 = 0,04167$$

* *Node* 12

Probabilitas tiap *class* dalam *node* 12 (t_R) sebagai berikut:

$$\text{P(Negatif } |t_R) = \frac{51}{92} = 0,55435$$

$$\text{P(Positif } |t_R) = \frac{41}{92} = 0,44565$$

Sehingga diperoleh

$$\text{r}(t_R) = 1 - 0,55435 = 0,44565$$

Nilai probabilitas objek dalam *node* 12 adalah sebagai berikut:

$$\text{P}(t_R) = \frac{N(node\ 12)}{N} = \frac{92}{168} = 0,54762$$

Dengan demikian,

$$\text{R}(t_R) = \text{r}\ (t_R) \times \text{P}(t_R) = 0,44565 \times 0,54762 = 0,22405$$

Oleh karena itu, persamaan

$$\text{R(t)} = \text{R}(t_L) + \text{R}(t_R)$$

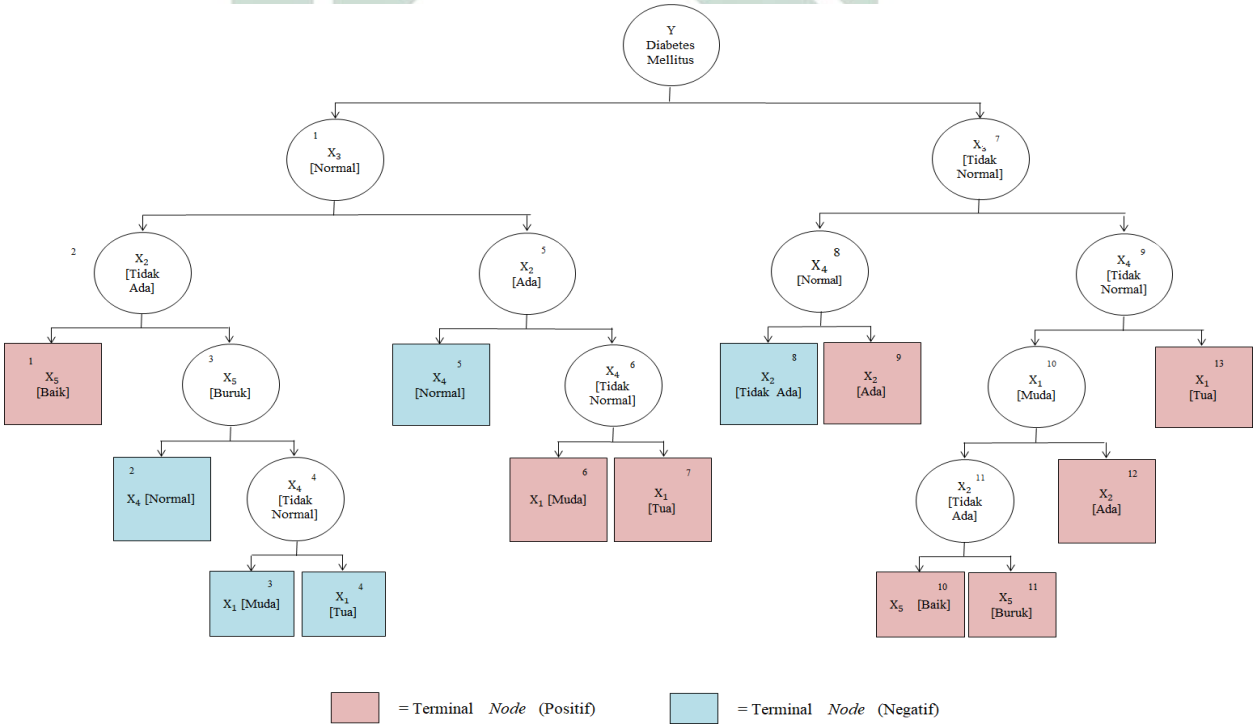
$$0{,}13208 = 0{,}04167 + 0{,}22405$$

$$0{,}13208 \neq 0{,}26572$$

tidak terpenuhi, sehingga tidak dilakukan pemangkasan pada *node* 10. Langkah ini dilakukan pada setiap *root node* dan dua *child node* hingga memperoleh pohon klasifikasi optimal.

3. Penentuan Pohon Klasifikasi Optimal

Setelah dilakukan tahap *tree prunning* pada setiap *root node* dan dua *child node* dihasilkan pohon klasifikasi optimal seperti Gambar 4.22.



Gambar 4.22 Pohon Klasifikasi Optimal

Berdasarkan Gambar 4.22, pohon klasifikasi tersebut terdapat 13 terminal *node*, 11 *child node*, 1 *root node* dengan tingkat kedalaman sebanyak 5.

4. Ketepatan Klasifikasi

Pohon klasifikasi optimal yang telah diperoleh, selanjutnya diuji keakuratan

dalam mengklasifikasikan data *training* dan data *testing*. Perhitungan ketepatan klasifikasi menggunakan tabel *confusion matrix*.

a. Ketepatan Klasifikasi Data *Training*

Data *Training* yang digunakan dalam penelitian ini sebanyak 168 sampel.Ketepatan klasifikasi data *training* disajikan pada Tabel 4.16.

Tabel 4.16 Ketepatan Klasifikasi Pohon Optimal Berdasarkan Data *Training*

Klasifikasi		*Class* Prediksi		Akurasi	*Error*
		1	0		
Class Aktual	1	112	6	90,48%	9,52%
	0	10	44		

Berdasarkan Tabel 4.16, ketepatan klasifikasi (akurasi) diperoleh dari Persamaan (2.28).

$$\text{Akurasi} = \left(\frac{TP+TN}{TP+FN+FP+TN}\right) \times 100\%$$
$$= \left(\frac{112+40}{112+6+10+40}\right) \times 100\%$$
$$= \left(\frac{152}{168}\right) \times 100\%$$
$$= 90,48\%$$

Dari perhitungan tersebut dapat disimpulkan bahwa ketepatan klasifikasi berdasarkan data *training* penyakit Diabetes Mellitus menggunakan CART sebesar 90,48%.

b. Ketepatan Klasifikasi Menggunakan Data *Testing*

Data *Testing* yang digunakan dalam penelitian ini sebanyak 74 sampel. Ketepatan klasifikasi data *testing* disajikan pada Tabel 4.17.

Tabel 4.17 Ketepatan Klasifikasi Pohon Optimal Berdasarkan Data *Testing*

Klasifikasi		*Class* Prediksi		Akurasi	*Error*
		1	0		
Class Aktual	1	49	1	90,27%	9,73%
	0	6	16		

Berdasarkan Tabel 4.17, ketepatan klasifikasi (akurasi) diperoleh dari Persamaan (2.28).

$$\begin{aligned}\text{Akurasi} &= \left(\frac{TP+TN}{TP+FN+FP+TN}\right) \times 100\% \\ &= \left(\frac{49+16}{49+1+7+16}\right) \times 100\% \\ &= \left(\frac{65}{72}\right) \times 100\% \\ &= 90,27\%\end{aligned}$$

Dari perhitungan tersebut dapat disimpulkan bahwa ketepatan klasifikasi berdasarkan data *testing* penyakit Diabetes Mellitus menggunakan CART sebesar 90,27%.

4.4. Analisis Hasil

Berdasarkan hasil model klasifikasi penyakit Diabetes Mellitus yang diperoleh dari metode Regresi Logistik Biner dan CART akan diinterpretasikan sebagai berikut:

1. Regresi Logistik Biner

   Dari beberapa uji yang telah terpenuhi, dapat diinterpretasikan model Regresi Logistik Biner yang diperoleh sebagai berikut:

$$\pi(x_i) = \frac{e^{(-6,9683+1,6115X_1+5,6213X_2+3,9764X_3+4,6913X_4)}}{1+e^{(-6,9683+1,6115X_1+5,6213X_2+3,9764X_3+4,6913X_4)}}$$

Variabel-variabel yang mempunyai pengaruh terhadap terjadinya penyakit Diabetes Mellitus adalah:

a. Usia (X_1), semakin tua usia seseorang maka peluang seseorang mengidap penyakit Diabetes Mellitus sebesar $1,6115$ kali daripada seseorang dengan usia muda, dengan nilai variabel yang lain tetap.

b. Faktor Keturunan (X_2), seseorang yang mempunyai riwayat keturunan Diabetes Mellitus berpeluang mengidap penyakit Diabetes Mellitus sebesar $5,6213$ kali daripada seseorang yang tidak mempunyai riwayat keturunan Diabetes Mellitus , dengan nilai variabel yang lain tetap.

c. Kadar Gula Darah (X_3), semakin tinggi kadar gula darah seseorang maka peluang seseorang mengidap penyakit Diabetes Mellitus sebesar $3,9764$ kali daripada seseorang dengan kadar gula darah normal , dengan nilai variabel yang lain tetap.

d. Obesitas (X_4), semakin bertambah berat badan seseorang maka peluang seseorang mengidap penyakit Diabetes Mellitus sebesar $4,6913$ kali daripada seseorang yang berat badannya normal/ideal, dengan nilai variabel yang lain tetap.

Perbandingan data *training* dan data *testing* yang digunakan dalam penelitian ini yaitu 70% : 30%. Berdasarkan data *training* tersebut terbentuk model Regresi Logistik Biner terbaik dengan ketepatan klasifikasi (akurasi) untuk data *training* sebesar 92,86% dan data *testing* sebesar 91,67%.

2. *Classification And Regression Trees (CART)*

Pada model pohon klasifikasi optimal yang terbentuk diperoleh beberapa aturan (*rule*) klasifikasi sebagai berikut:

a. X_3 (Gula Darah) = normal $\cup$ X_2 (Keturunan) = tidak ada $\cup$ X_5 (Pola Makan) = baik $\Rightarrow$ Positif DM.

Apabila pasien mempunyai kadar gula darah yang normal, tidak terdapat riwayat keturunan DM dan pola makannya baik, maka pasien tersebut dikatakan positif Diabetes Mellitus.

b. X_3 (Gula Darah) = normal $\cup$ X_2 (Keturunan) = tidak ada $\cup$ X_5 (Pola Makan) = buruk $\cup$ X_4 (Obesitas) = normal $\Rightarrow$ Negatif DM.

Apabila pasien mempunyai kadar gula darah yang normal, tidak ada riwayat keturunan DM, pola makannya buruk dan mempunyai tingkat kegemukan normla, maka pasien tersebut dikatakan negatif Diabetes Mellitus.

c. X_3 (Gula Darah) = normal $\cup$ X_2 (Keturunan) = tidak ada $\cup$ X_5 (Pola Makan) = buruk $\cup$ X_4 (Obesitas) = tidak normal $\cup$ X_1 (Usia) = muda $\Rightarrow$ Negatif DM.

Apabila pasien mempunyai kadar gula darah yang normal, tidak terdapat riwayat keturunan DM, pola makannya buruk, mempunyai tingkat kegemukan (obesitas) tidak normal dan berusia muda, maka pasien tersebut dikatakan negatif Diabetes Mellitus.

d. X_3 (Gula Darah) = normal $\cup$ X_2 (Keturunan) = tidak ada $\cup$ X_5 (Pola Makan) = buruk $\cup$ X_4 (Obesitas) = tidak normal $\cup$ X_1 (Usia) = tua $\Rightarrow$ Negatif DM.

Apabila pasien mempunyai kadar gula darah yang normal, tidak terdapat riwayat keturunan DM, pola makannya buruk, mempunyai tingkat kegemukan (obesitas) tidak normal dan berusia tua, maka pasien tersebut dikatakan negatif Diabetes Mellitus.

e. X_3 (Gula Darah) = normal $\cup$ X_2 (Keturunan) = ada $\cup$ X_4 (Obesitas) = normal $\Rightarrow$ Negatif DM.

Apabila pasien mempunyai kadar gula darah yang normal, terdapat riwayat keturunan DM dan mempunyai tingkat kegemukan (obesitas) normal, maka pasien tersebut dikatakan negatif Diabetes Mellitus.

f. X_3 (Gula Darah) = normal $\cup$ X_2 (Keturunan) = ada $\cup$ X_4 (Obesitas) = tidak normal $\cup$ X_1 (Usia) = muda $\Rightarrow$ Positif DM.

Apabila pasien mempunyai kadar gula darah yang normal, terdapat riwayat keturunan DM, mempunyai tingkat kegemukan (obesitas) tidak normal dan berusia muda, maka pasien tersebut dikatakan positif Diabetes Mellitus.

g. X_3 (Gula Darah) = normal $\cup$ X_2 (Keturunan) = ada $\cup$ X_4 (Obesitas) = tidak normal $\cup$ X_1 (Usia) = tua $\Rightarrow$ Positif DM.

Apabila pasien mempunyai kadar gula darah yang normal, terdapat riwayat keturunan DM, mempunyai tingkat kegemukan (obesitas) tidak normal dan berusia tua, maka pasien tersebut dikatakan positif Diabetes Mellitus.

h. X_3 (Gula Darah) = tidak normal $\cup$ X_4 (Obesitas) = normal $\cup$ X_2 (Keturunan) = tidak ada $\Rightarrow$ Negatif DM.

Apabila pasien mempunyai kadar gula darah yang tidak normal, mempunyai tingkat kegemukan (obesitas) normal dan tidak terdapat riwayat keturunan DM, maka pasien tersebut dikatakan negatif Diabetes Mellitus.

i. X_3 (Gula Darah) = tidak normal $\cup$ X_4 (Obesitas) = normal $\cup$ X_2 (Keturunan) = ada $\Rightarrow$ Positif DM.

Apabila pasien mempunyai kadar gula darah yang tidak normal, mempunyai tingkat kegemukan (obesitas) normal dan terdapat riwayat keturunan DM, maka pasien tersebut dikatakan positif Diabetes Mellitus.

j. X_3 (Gula Darah) = tidak normal $\cup$ X_4 (Obesitas) = tidak normal $\cup$ X_1 (Usia) = muda $\cup$ X_2 (Keturunan) = tidak ada $\cup$ X_5 (Pola Makan) = baik $\Rightarrow$ Positif DM.

Apabila pasien mempunyai kadar gula darah yang tidak normal, mempunyai tingkat kegemukan (obesitas) tidak normal, berusia muda, tidak terdapat riwayat keturunan DM dan pola makannya baik, maka pasien tersebut dikatakan positif Diabetes Mellitus.

k. X_3 (Gula Darah) = tidak normal $\cup$ X_4 (Obesitas) = tidak normal $\cup$ X_1 (Usia) = muda $\cup$ X_2 (Keturunan) = tidak ada $\cup$ X_5 (Pola Makan) = buruk $\Rightarrow$ Positif DM.

Apabila pasien mempunyai kadar gula darah yang tidak normal, mempunyai tingkat kegemukan (obesitas) tidak normal, berusia muda, tidak terdapat riwayat keturunan DM dan pola makannya buruk, maka pasien tersebut dikatakan positif Diabetes Mellitus.

l. X_3 (Gula Darah) = tidak normal $\cup$ X_4 (Obesitas) = tidak normal $\cup$ X_1 (Usia) = muda $\cup$ X_2 (Keturunan) = ada $\Rightarrow$ Positif DM.

Apabila pasien mempunyai kadar gula darah yang tidak normal, mempunyai tingkat kegemukan (obesitas) tidak normal, berusia muda dan terdapat riwayat keturunan DM, maka pasien tersebut dikatakan positif Diabetes Mellitus.

m. X_3 (Gula Darah) = tidak normal $\cup$ X_4 (Obesitas) = tidak normal $\cup$ X_1 (Usia) = tua $\Rightarrow$ Positif DM.

Apabila pasien mempunyai kadar gula darah yang tidak normal, mempunyai tingkat kegemukan (obesitas) tidak normal dan berusia tua, maka pasien tersebut dikatakan positif Diabetes Mellitus.

Dalam penelitian ini, perbandingan data *training* dan data *testing* yang digunakan yaitu 70% : 30%. Berdasarkan data *training* tersebut terbentuk pohon klasifikasi optimal dengan ketepatan klasifikasi (akurasi) untuk data *training* sebesar 90,48% dan data *testing* sebesar 90,27%.

Dengan demikian, berdasarkan model yang terbentuk dari masing-masing metode dapat disimpulkan bahwa untuk data *training* metode Regresi Logistik Biner mempunyai tingkat akurasi lebih tinggi 2,38% daripada CART. Sementara itu, untuk data *testing* metode Regresi Logistik Biner mempunyai tingkat akurasi lebih tinggi 1,4% daripada CART. Oleh karena itu, pada penelitian ini metode Regresi Logistik Biner lebih baik dibandingkan metode CART.

BAB V

PENUTUP

Pada bab ini akan diberikan simpulan dan saran-saran yang dapat diambil berdasarkan materi-materi yang telah dibahas pada bab-bab sebelumnya.

5.1. Simpulan

Berdasarkan hasil dan pembahasan yang telah dilakukan pada bab sebelumnya, sehingga simpulan yang dapat diambil dalam penelitian ini adalah sebagai berikut :

1. Dari pengujian data pasien Diabetes Mellitus menggunakan Regresi Logistik Biner diperoleh model sebagai berikut:

$$\pi(x_i) = \frac{e^{(-6{,}9683+1{,}6115X_1+5{,}6213X_2+3{,}9764X_3+4{,}6913X_4)}}{1+e^{(-6{,}9683+1{,}6115X_1+5{,}6213X_2+3{,}9764X_3+4{,}6913X_4)}}$$

   Berdasarkan model tersebut, variabel yang mempunyai pengaruh terhadap penyakit Diabetes Mellitus, yaitu usia, faktor keturunan, kadar gula darah dan obesitas.

2. Dari pengujian data pasien Diabetes Mellitus menggunakan *Classification And Regression Trees* (CART) diperoleh pohon klasifikasi optimal yang terdiri dari 11 *child node*, 13 terminal *node* yang berarti terdapat 13 aturan (*rule*). Aturan tersebut terdiri dari 5 *rule* dengan kategorik negatif Diabetes Mellitus dan 8 *rule* dengan kategorik positif Diabetes Mellitus.

3. Ketepatan klasifikasi dari model Regresi Logistik Biner sebesar 92,86% untuk data *training* dan data *testing* sebesar 91,67%. Sementara itu, akurasi dari pohon klasifikasi optimal CART sebesar 90,48% untuk data *training* dan data *testing* sebesar 90,27%. Dari kedua model diperoleh, untuk kasus klasifikasi penyakit Diabetes Mellitus dengan 5 variabel independen yaitu usia, faktor keturunan, kadar gula darah, obesitas dan pola makan metode Regresi Logistik Biner lebih baik apabila dibandingkan dengan metode CART. Hal ini dikarenakan metode Regresi Logistik Biner mempunyai tingkat akurasi lebih tinggi dibandingkan metode CART baik pada *training* maupun data *testing*.

5.2. Saran

Berdasarkan analisis hasil yang telah dilakukan, adapun saran yang dapat diberikan adalah sebagai berikut:

1. Dalam kasus klasifikasi penyakit Diabetes Mellitus menggunakan Regresi Logistik Biner Maupun CART belum 100% akurat, sehingga untuk penelitian selanjutnya dapat menggunakan metode lain, misalnya *Random Forest*, SVM atau metode lainnya agar diperoleh nilai akurasi yang lebih tinggi.

2. Untuk penelitian berikutnya, dapat dilakukan penambahan faktor-faktor yang mempengaruhi penyakit Diabetes Mellitus, misalnya jenis kelamin, kadar insulin, tekanan darah dan faktor yang lainnya.

3. Jumlah data yang digunakan dapat diperbanyak, semakin banyak data yang digunakan semakin kecil nilai *error*nya.

4. Untuk penelitian selanjutnya, dapat dikembangkan *software* untuk deteksi penyakit Diabetes Mellitus.

DAFTAR PUSTAKA

Adnyana, L., Hense, Budhiarta, A. G. 2006. *Penatalaksanaan Pasien Diabetes Mellitus di Poliklinik Rumah Sakit Sanglah Denpasar*. Jurnal Penyakit Dalam. Volume 7 No. 3, pp. 186-192. Denpasar.

Akhsanti, R. M., Rahayu, W. and Santi, V. M. 2018. *Klasifikasi Diagnosis Penyakit Kanker Payudara dengan Pendekatan Regresi Logistik Biner dan Metode Classification And Regression Trees (CART)*. Jurnal Matematika dan Terapan, Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta. Jakarta.

American Diabetes Association (ADA). 2012. *Medical Advice for People with Diabetes in Emergency Situations*. American Diabetes Association Journal. http://care.diabetesjournal.org/content (Diakses tanggal 7 Oktober 2019).

Arvianti, Risti V. 2019. *Klasifikasi Penyakit Diabetes Mellitus Tipe 2 Menggunakan Learning Vector Quantization (LVQ))*. Skripsi. Universitas Islam Negeri Sunan Ampel Surabaya: Surabaya.

Breiman, L., Friedman, J., Olsen, R. A., dan Stone, C. 1984. *Classification And Regression Trees*. Wadsworth Belmonth: California.

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

Departemen Kesehatan. 2005. *Pharmaceutical Care untuk Penyakit Diabetes Mellitus*. Direktorat Bina Farmasi Komunitas Dan Klinik : Jakarta.

Diabetes UK. 2010. *Diabetes in the UK 2010. Key Statistics on Diabetes*: United Kingdom.

Dorland, W. 2005. *Kamus Kedokteran Dorland*. EGC: Jakarta.

Ghozali, Imam. 2005. *Aplikasi Analisis Multivariate Dengan Program SPSS*. Penerbit Universitas Diponegoro: Semarang.

Goronescu, F. 2011. *Data Mining: Concept, Models and Techniques*. Springer: Verlag Berlin Heidelberg.

Hosmer, D. W., dan Lemeshow. S. 2000. *Applied Logistic Regression*. John Wiley & Sons, Inc: New York.

International of Diabetic Ferderation. 2017. *IDF Diabetes Atlas*. Eighth Edition. www.idf.org/diabetesatlas (Diakses tanggal 7 Oktober 2019)

Johnson, R.A. dan Wichern, D. W. 2007. *Applied Multivariate Statistical Analysis*. Sixth Edition. Pearson Education Inc: USA.

Kaban, Sempakata. 2007. *Diabetes Tipe 2 di Kota Sibolga Tahun 2005*. Majalah Kedokteran Nusantara. Volume 40 No. 2. Medan.

Kementerian Kesehatan. 2010. *Petunjuk Teknis Pengukuran Faktor Risiko Diabetes Mellitus*.

Kusrini, Luthfi, E. T. 2009. *Algoritma Data Mining*. Penerbit Andi: Yogyakarta.

Lestawati, R., Rais dan Utami, I. T. 2018. *Perbandingan antara Metode CART (Classification and Regression Tress) dan Regresi Logistik (Logistic Regression) dalam Mengklasifikasikan Pasien Penderita DBD (Demam Berdarah Dengue)*. Jurnal Ilmiah Matematika dan Terapan, Volume 15 No. 1, pp 98-107,

Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Tadulako. Palu.

Lewis dan Roger, J. 2000. *An Introduction to Classification and Regression Trees (CART) Analysis*. Annual Meeting of Society for Academic Emergency Medicine of Sanfransisco, UCLA Medical Center: California.

Margasari, A. 2014. *Penerapan Metode CART (Classification And Regression Trees) dan Analisis Regresi Logistik Biner pada Klasifikasi Profil Mahasiswa FMIPA Universitas Brawijaya*. Jurnal Mahasiswa Statistik, Vol. 2, No. 4, FMIPA Universitas Brawijaya. Malang.

Misnadiarly. 2006. *Diabetes Mellitus Gangren, Ulcer, Infeksi, Mengenali Gejala, Menanggulangi dan Mencegah Komplikasi*. Pustaka Obor Populer: Jakarta.

Mustamir. 2008. *5 Metode Penyembuhan dari Langit*. Lingkaran: Yogyakarta.

Rizki, F., Widodo, D. A., Wulandari, S. P. 2015. *Faktor Risiko Penyakit Anemia Gizi Besi pada Ibu Hamil di Jawa Timur Menggunakan Analisis Regresi Logistik*. *Jurnal Sains dan Seni ITS*. Vol. 4, No. 2;2337 - 3520.

Rumaendra, Wella. 2016. *Perbandingan Klasifikasi Penyakit Hipertensi Menggunakan Regresi Logistik Biner dan Algoritma C4.5*. Universitas Diponegoro: Semarang.

Sudoyo, A., Setiyohadi, B., Simadibrata, M., dan Setiati, S. 2009. *Buku Ajar Ilmu Penyakit dalam Jilid 3*. Edisi ., Interna Publishing: Jakarta.

WHO. 2006. *Definition, Diagnosis and Classification of Diabetes Mellitus and its Complication*. WHO.

Widagdo, K. A. 2010. *Pembentukan Pohon Klasifikasi Biner dengan Algoritma CART (Classification And Regression Trees) (Studi Kasus Penyakit Diabetes Suku Pima Indian)*. Skripsi. Universitas Diponegoro: Semarang.

Widarjono, Agus. 2010. *Analisis Statistika Multivariat Terapan*. UPP STIM YKPN: Yogyakarta.

Zaen, Nanida J. 2019. *Diagnosis Penyakit Stroke dengan Metode Regresi Logistik Biner*. Skripsi. Universitas Islam Negeri Sunan Ampel Surabaya: Surabaya.