UJIAN TENGAH SEMESTER

DATA MINING

Diajukan Untuk Melengkapi UTS Mata Kuliah Data Mining

Disusun Oleh :

Thalita Safa Azzahra (140610190053)

Dosen : Sinta Septi P, S.Si., M.Stat

PROGRAM STUDI S-1 STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS PADJADJARAN

JATINANGOR

2021

SOAL NOMOR 1

1. Perhatikan pernyataan berikut ini, apakah termasuk ke dalam persoalan data mining atau tidak.
   a. Membagi pelanggan sebuah perusahaan berdasarkan jenis kelaminnya.
      → Tidak.
   b. Membagi pelanggan sebuah perusahaan berdasarkan pendapatannya.
      → Tidak.
   c. Mengurutkan database siswa berdasarkan NPM.
      → Tidak.
   d. Memprediksi hasil pelemparan sepasang dadu.
      → Tidak, karena perhitungan tersebut merupakan perhitungan probabilitas.
   e. Memprediksi harga saham perusahaan di masa depan menggunakan metode historis catatan.
      → Ya, termasuk data mining karena dapat dibuat model yang bisa memprediksi nilai harga saham yang akan datang sehingga termasuk data mining pemodelan prediktif.
   f. Pemantauan detak jantung pasien untuk kelainan.
      → Ya, termasuk deteksi anomali pada data mining.
   g. Pemantauan gelombang seismik untuk aktivitas gempa.
      → Ya, termasuk klasifikasi pada data mining.

SOAL NOMOR 2

2. Klasifikasikan atribut berikut sebagai biner, diskrit, atau kontinu. Juga mengklasifikasikan mereka sebagai kualitatif (nominal atau ordinal) atau kuantitatif (interval atau rasio). Beberapa kasus mungkin memiliki lebih dari satu interpretasi, jadi tunjukkan secara singkat alasan Anda jika menurut Anda mungkin ada beberapa ambiguitas.
   Contoh: Umur dalam tahun. Jawaban: Diskrit, kuantitatif, rasio
   a. Waktu dalam WIB, WIT atau WITA.
      → Biner, Kualitatif, Ordinal
   b. Tingkat kecerahan yang diukur dengan pengukur cahaya.
      → Kontinu, Kuantitatif, Rasio
   c. Sudut dalam derajat antara 0^0 dan 360^0.
      → Kontinu, Kuantitatif, Rasio
   d. Ketinggian diatas permukaan laut.
      → Kontinu, Kuantitatif, Interval atau Rasio
   e. Jumlah pasien di rumah sakit.
      → Diskrit, Kuantitatif, Rasio

SOAL NOMOR 3

3. Apakah perbedaan antara reduksi dimensi berdasarkan agregasi dan reduksi dimensi berdasarkan teknik seperti PCA dan SVD? Jelaskan!

J AW A B :

Reduksi dimensi berdasarkan agregasi seperti Data cube aggregation yang memungkinkan analisis data pada berbagai tingkat abstraksi sehingga menyediakan akses cepat ke data yang dirangkum dan diringkas. Sedangkan reduksi dimensi berdasarkan Teknik seperti pca dan svd (hierarki), untuk cabang dapat memungkinkan cabang untuk dikelompokkan ke dalam wilayah, berdasarkan alamatnya

SOAL NOMOR 4

4. John mengukur tekanan semua ban yang masuk ke garasi untuk mengganti oli dan mencatat nilainya. Tanpa sepengetahuannya, pengukur bannya salah kalibrasi dan menambahkan 3 psi untuk setiap pembacaan. Menurut definisi noise yang digunakan dalam data mining, apakah kesalahan yang ditimbulkan oleh pengukur ban ini dianggap noise? Berikan alasan Anda.

J AW A B :

Noise merupakan random error atau varians dalam variabel yang diukur. Dalam data mining *noise* merujuk pada modifikasi dari nilai original atau nilai aslinya, atau juga bisa disebut data yang berisi nilai – nilai yang salah atau anomali. Berdasarkan permasalahan diatas, artinya John membaca nilai – nilai yang salah atau nilai anomali yang disebabkan oleh pengukuran tekanan ban yang salah kalibrasi sehingga harus menambahkan 3psi kedalam setiap nilai pengamatan. Sehingga kesalahan ini dapat dianggap sebagai *noise.*

SOAL NOMOR 5

5. Diberikan masalah klasifikasi dengan 2 *class*, dimana P adalah kelas positif dan N adalah kelas negatif, kita dapat mendeskripsikan kinerja (*performance*) algoritma dengan menggunakan istilah sbb: TP, FP, TN, dan FN.
   a. Merujuk pada apa masing-masing istilah tersebut?

J AW A B :

- TP merupakan *true positive*, yaitu tupel kelas positif yang diberi label dengan benar oleh pengklasifikasian. Terjadi ketika memprediksi positif dan prediksi itu benar dengan aktualnya.

- FP merupakan *false positive*, yaitu tupel kelas negatif yang diberi label dengan benar oleh pengklasifikasian. Terjadi ketika memprediksi positif tetapi aktualnya tidak benar.
- TN merupakan *true negative*, yaitu tupel kelas negatif yang salah diberi label sebagai positif. Terjadi ketika memprediksi negatif dan benar aktualnya.
- FN merupakan *false negative*, yaitu tupel kelas yang salah diberi label sebagai positif. Terjadi ketika memprediksi negatif tetapi aktualnya tidak benar.

b. Tempatkan 4 istilah yang tercantum di atas pada bagian a ke dalam slot yang sesuai pada tabel di bawah.

J AW A B :

		Prediksi	
		Positif	Negatif
Aktual	Positif	TP (*True Positive)*	FN (*False Negative)*
	Negatif	FP (*False Positive)*	TN (True *Negative)*

c. Berikan rumus akurasi dalam TP, TN, FP, dan FN.

J AW A B :

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} x\ 100\%$$

SOAL NOMOR 6

Gunakan Heart Disease Dataset pada Tugas Pre-processing beberapa waktu lalu.

A. RUMUSAN PERMASALAHAN DATA

Permasalahan yang diangkat dalam kasus ini adalah mengklasifikan setiap pasien apakah terdiagnosis penyakit jantung atau tidak berdasarkan faktor atau variabel yang ada. Tentunya dalam pengklasifikasian ini diperlukan model yang paling baik sehingga hasil yang didapat akurat.

B. *PRE-PROCESSING DATA*

Preprocessing data merupakan teknik awal data mining untuk mengubah data mentah atau biasa dikenal dengan *raw* data yang dikumpulkan dari berbagai sumber menjadi informasi yang lebih bersih dan bisa digunakan untuk pengolahan selanjutnya. Proses ini bisa juga disebut dengan langkah awal

untuk mengambil semua informasi yang tersedia dengan cara membersihkan, memfilter, dan menggabungkan data-data tersebut

1. Data Cleaning

Data Cleaning berfungsi untuk membersihkan data dengan mengisi nilai – nilai yang hilang, *smoothing noise* data, mengidentifikasi dan menghapus *outlier*, serta menyelesaikan inkonsistensi

a. *Incomplete (Missing Value)*

Data yang tidak akurat karena informasi yang hilang menyebabkan informasi yang ada di dalamnya tidak relevan. Missing value sering terjadi ketika ada masalah dalam proses pengumpulan, seperti kesalahan dalam entry data atau masalah dalam penggunaan biometrik.

```
> ##DATA CLEANING
> #Cek Missing Value
> colSums(is.na(as.data.frame(data)))
     Age      Sex       CP trestbps     chol      fbs  restecg
       0        0        0        0        0        0        0
  thalac    exang  oldpeak    slope       ca     thal      num
       0        0        0        0        4        2        0
```

Berdasarkan *output* di atas dapat dilihat bahwa terdapat *missing value* yaitu pada variabel **ca** dan variabel **thal**. Untuk menangani missing value, kita bisa mengabaikan bagian kumpulan data yang hilang yang disebut dengan tupel. Namun, cara ini hanya dapat dilakukan jika kita memiliki kumpulan big data yang memiliki beberapa missing value dalam tupel yang sama. Namun, jika data yang kita miliki tidak terlalu besar, pendekatan lain yang bisa digunakan adalah dengan mengisi missing value tersebut dengan memasukan suatu nilai secara manual maupun menggunakan proses komputasi. Biasanya, missing value akan diisi dengan mean atau modus, tergantung dari jenis datanya.

Missing value yang terdapat pada data ini akan kita isi dengan imputasi nilai modus dari setiap variabel terkait.

```
> #Mengatasi Missing Value
> #karena nilai variabel ca dan thal bersifat kategori maka
> #penanganannya menggunakan modus
> mode <- function(data){
+   uqx <- unique(data)
+   tab <- table(data)
+   sort(uqx)[tab == max(tab)]
+ }
> data$ca[is.na(data$ca)] <- mode(data$ca)
> data$thal[is.na(data$thal)] <- mode(data$thal)
> colSums(is.na(as.data.frame(data))) #cek ulang
     Age      Sex       CP trestbps     chol      fbs  restecg
       0        0        0        0        0        0        0
```

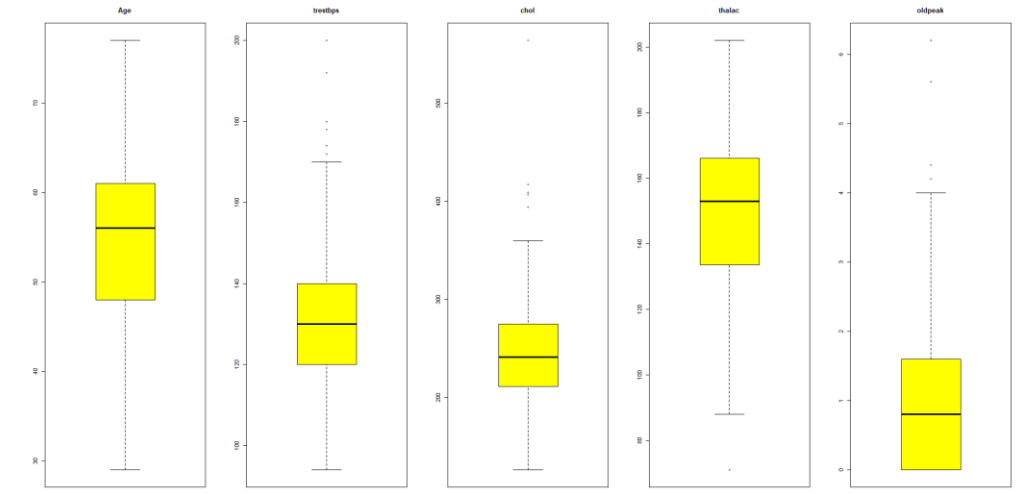
```
  thalac     exang   oldpeak     slope        ca      thal       num
       0         0         0         0         0         0         0
```

Berdasarkan *output* di atas, dapat dilihat bahwa pada data sudah tidak terdapat *missing value.*

b. *Noisy (Outlier)*

Berisi kesalahan atau nilai-nilai outlier yang menyimpang yang tidak sesuai dengan data yang lainnya.

```
> #Noisy Data
> #outlier
> #variabel bukan numerik: sex, cp, fbs, restecg, exang, slope, ca,
thal, num
> #variabel tersebut tidak perlu dicek outlier
> boxplot(data)
> par(mfrow=c(2,3))
> bp1 = boxplot(data$Age,col="yellow",main="Age") ; bp1$out
numeric(0)
> bp2 = boxplot(data$trestbps, col = "yellow", main = "trestbps")
; bp2$out
[1] 172 180 200 174 178 192 180 178 180
> bp3 = boxplot(data$chol, col="yellow",main="chol") ; bp3$out
[1] 417 407 564 394 409
> bp4 = boxplot(data$thalac, col="yellow",main="thalac") ; bp4$out
[1] 71
>  bp5  =  boxplot(data$oldpeak,  col="yellow",main="oldpeak")  ;
bp5$out
[1] 6.2 5.6 4.2 4.2 4.4
```



Berdasarkan *output* di atas, dapat dilihat bahwa dari lima (5) variabel numerik yang di *check outlier* nya, hanya empat (4) variabel yang memiliki *outlier* yaitu variabel trestbps, chol, thalac, dan oldpeak. *Trestbps* merupakan variabel *Resting Blood Pressure* yaitu tekanan darah normal yang biasanya bernilai 120/80 mmHg. Nilai *serum cholesterol* yang sehat adalah sebesar <200 mg/dL. Karena nilai

outlier yang ditemuka benar – benar jauh dari angka normal, maka untuk penganannya, kita akan me - *remove outlier* pada data.

```
>#terdapat outlier pada variabel trestbps, chol,thalac, oldpeak,
> #removeoutlier
> which(data$trestbps %in% c(bp2$out))
[1]  15  84 127 173 184 189 202 214 232
> which(data$chol %in% c(bp3$out))
[1]  49 122 153 174 182
> which(data$thalac %in% c(bp4$out))
[1] 246
> which(data$oldpeak %in% c(bp5$out))
[1]  92 124 184 192 286

> datanoout = data[-c(15, 84, 127, 173, 184, 189, 202, 214,
+       232, 49, 122, 153, 174, 182, 246, 92, 124, 192, 286),]
```

c. *Inconsistent*

Ketidakcocokan dalam penggunaan kode atau nama. Inkonsisten data terjadi ketika seseorang menyimpan file yang berisi data yang sama dengan format yang berbeda-beda. Beberapa inkonsisten data adalah duplikasi dalam format yang berbeda, kesalahan pada kode nama, dan lain sebagainya.

```
#inconsistenc
#tidak ada data yang inconsistenc
```

2. Data Integration

Data Integration merupakan penggabungan data dari berbagai Database ke dalam satu Database baru. Data Integration dapat membantu mengurangi dan menghindari redundansi data dan inkonsistensi dalam kumpulan data yang dihasilkan. Ini dapat membantu meningkatkan akurasi dan kecepatan proses data mining berikutnya.

Karena data yang kita gunakan didapat melalui 1 sumber, maka data tersebut tidak perlu di integrasi.

3. Data Reduction

Data Reduction dapat diterapkan untuk mengurangi ukuran data dengan menggabungkan atau menghilangkan data yang tidak dibutuhkan

Memilah kumpulan data yang berukuran besar baik secara manual maupun otomatis membutuhkan waktu yang cukup lama. Oleh karena itu, perlu adanya proses pengurangan data

untuk membatasi kumpulan data sehingga meningkatkan efisiensi penyimpanan sekaligus mengurangi biaya uang dan waktu. Proses mengurangi data atau disebut juga dengan reduksi data merupakan proses kompleks yang melibatkan beberapa langkah, yaitu Data Cube Aggregation, Attribute Subset Selection, Numerosity Reduction, dan Dimensionality Reduction. Dalam Data Cube Aggregation, data kubus merupakan array multidimensi yang dihasilkan dari data organization. Untuk mendapatkannya, kita bisa menggunakan operasi agregasi yang bisa memperoleh satu nilai untuk sekumpulan data. Attribute Subset Selection artinya memilih atribut yang paling relevan dalam sekumpulan data yang akan digunakan dan sisanya akan dibuang. Untuk memilih subset, kita bisa menggunakan batas minimum yang harus dicapai oleh semua atribut. Semua atribut yang di bawah ambang batas minimal akan secara otomatis dibuang. Teknik Numerosity Reduction adalah teknik reduksi data yang menggantikan data asli dengan representasi data yang lebih kecil. Teknik Dimensionality Reduction merupakan teknik reduksi data dengan mengurangi ukurannya.

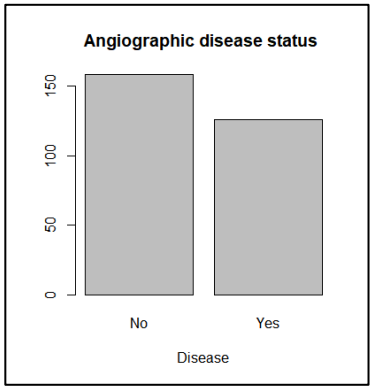
4. Data Tranformation and Data Discretization

Data Transformation berfungsi mengubah data ke dalam format yang sesuai untuk diproses dalam data mining. Beberapa metode data mining membutuhkan format data yang khusus sebelum bisa diaplikasikan Ada beberapa cara untuk melakukan transformasi data, yaitu normalization, attribute selection, discretization, dan konsep hierarchy generation. Normalization adalah proses menskalakan nilai data dalam rentang yang telah ditentukan sebelumnya. attribute selection merupakan prose yang menggunakan atribut untuk membuat data baru sehingga dapat mengatur kumpulan data dan membantu menganalisis data yang tersembunyi.

Teknik discretization merupakan proses transformasi data dengan mengganti nilai mentah atribut numerik dengan interval.

C. STATISTIKA DESKRIPTIF

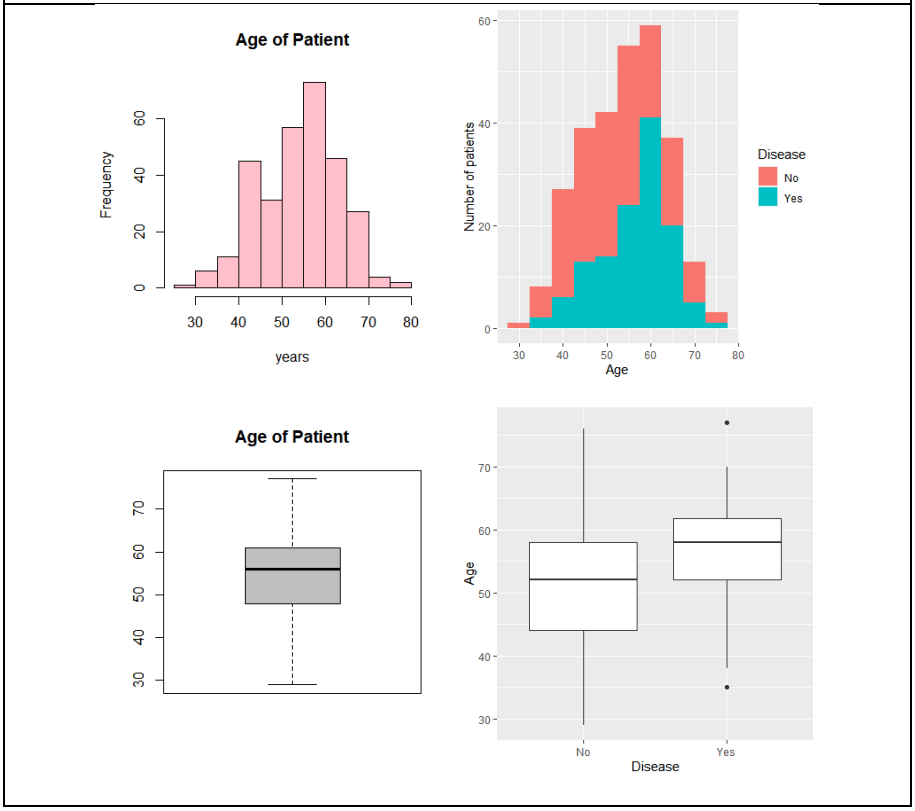
a. Variabel Angiographic disease status (0=no heart disease; more than 0=have heart disease)



Gambar diatas menunjukkan bahwa dalam data ini pasien yang terdiagnosis sehat atau tidak mengidap penyakit jantung lebih sedikit dibanding pasien yang sakit atau mengidap penyakit jantung.

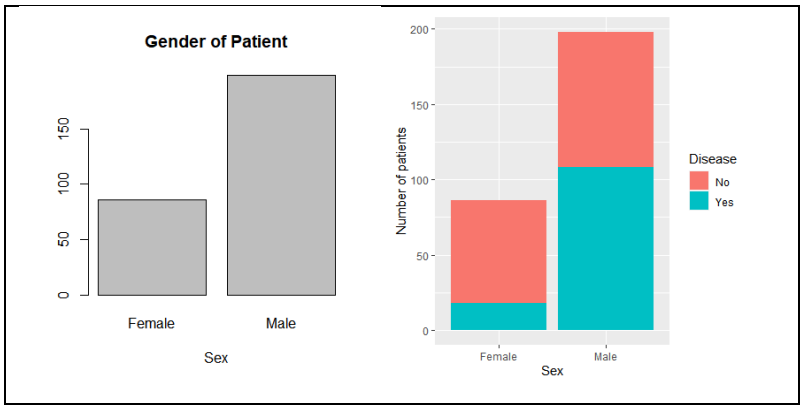
b. Variabel Age of Patient

```
> summary(data$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  29.00   47.00   55.00   54.07   60.00   77.00
```



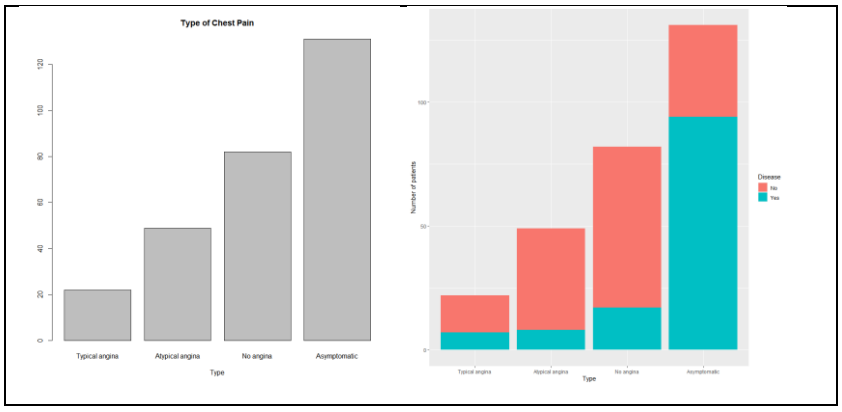
Usia rata – rata untuk pasian yang diperiksa dalam penelitian ini adalah 54 dnegan usia termudia dan tertua masing – masing adalah 29 dan 77 tahun. Histogram usia sedikit miring ke kiri, menunjukkan rata – rata usia sedikit lebih rendah dari median usia. Distribusi usia untuk kelompok penyakit jantung sedikit condong ke arah usia yang lebih tinggi sedangkan sebaliknya diamati untuk kelompok non-penyakit. Dengan kata lain, semakin tinggi usia, semakin besar kemungkinan pasien menderita penyakit jantung. Oleh karena itu, usia akan menjadi fitur prediktif. Secara keseluruhan, individu yang menunjukkan penyakit jantung memiliki median usia yang lebih tinggi dibandingkan dengan non-penyakit jantung, hal ini dapat dilihat pada *boxplot* yang dihasilkan.

c. Variabel Gender of Patient



Gambar diatas menunjukkan bahwa dalam data ini jenis kelamin laki – laki lebih banyak dibandingkan perempuan dan proporsi laki-laki yang didiagnosis dengan penyakit jantung secara signifikan lebih tinggi dibandingkan dengan perempuan. Perbandingan ini mengindikasikan bahwa pria lebih mungkin memiliki penyakit jantung daripada wanita.

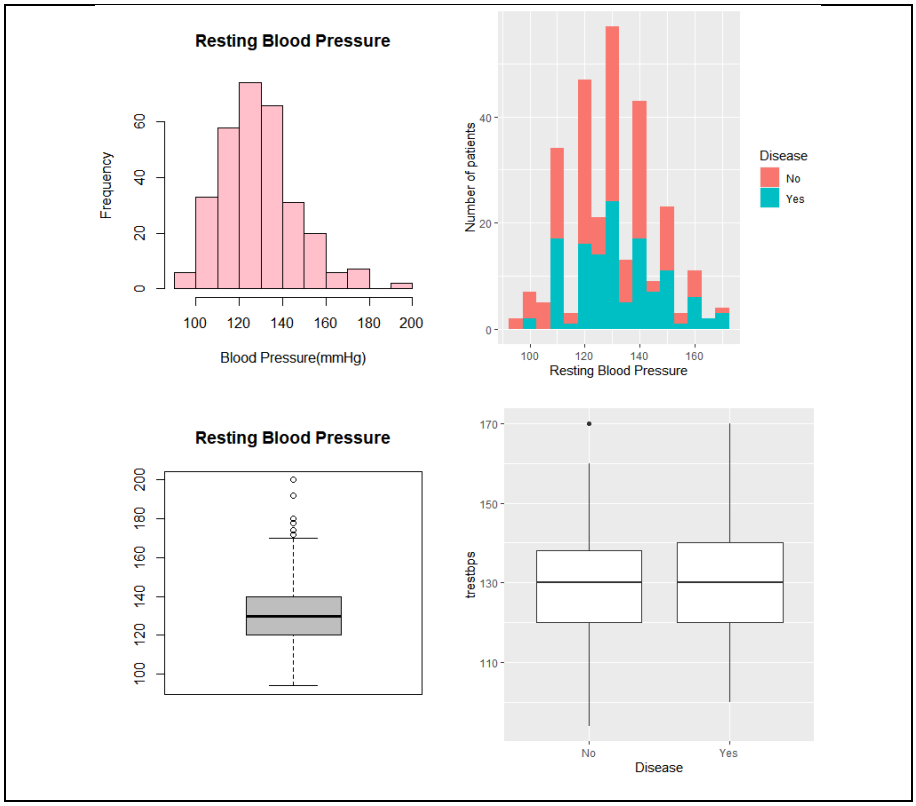
d. Variabel Type of Chest Pain (1: Typical Angina, 2: Atypical Angina, 3: Non-anginal Pain, 4: Asymptomatic)



Sebagian besar pasien yang terkena penyakit jantung mengalami nyeri dada tipe 4 yaitu *asymptomatic*. Namun, sebagian besar orang sehat mengalami nyeri tipe 3 yaitu *non angina.*

e. Variabel Resting Blood Pressure (in mm Hg on admission)

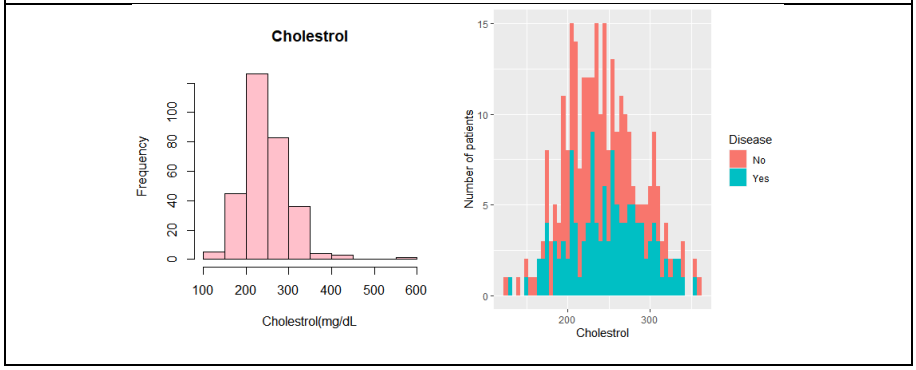
```
> summary(data$trestbps)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     94     120     130     130     140     170
```

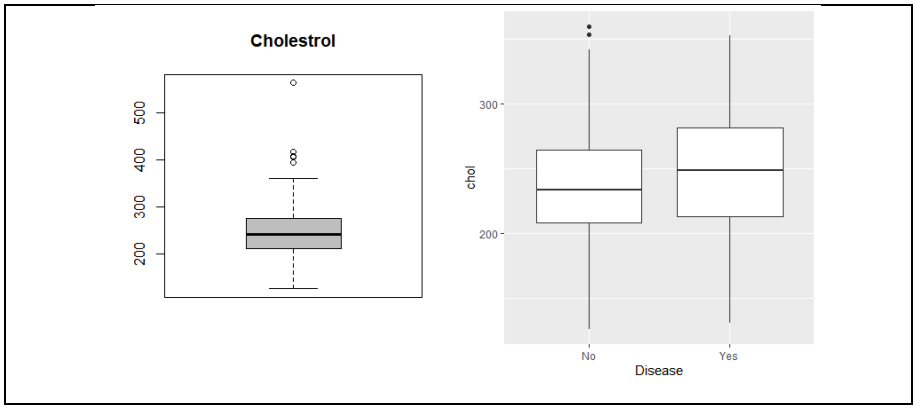


Resting blood pressure untuk keseluruhan menunjukkan nilai median 130 yang dimana mirip dnegan kelompok sakit dan tidak sakit, hal ini dapat dilihat pada posisi boxplot yang sejajar. Histogram tekanan darah istirahat miring ke kanan, menunjukkan bahwa beberapa pasien memiliki tekanan darah yang sangat tinggi. Ketika membandingkan Histogram secara terpisah untuk sakit dan tidak sakit, kita dapat melihat pasien yang memiliki penyakit jantung menunjukkan tekanan darah yang lebih tinggi dibandingkan dengan pasien yang tidak memiliki penyakit jantung.

f. Variabel Serum Cholesterol in mg/dl

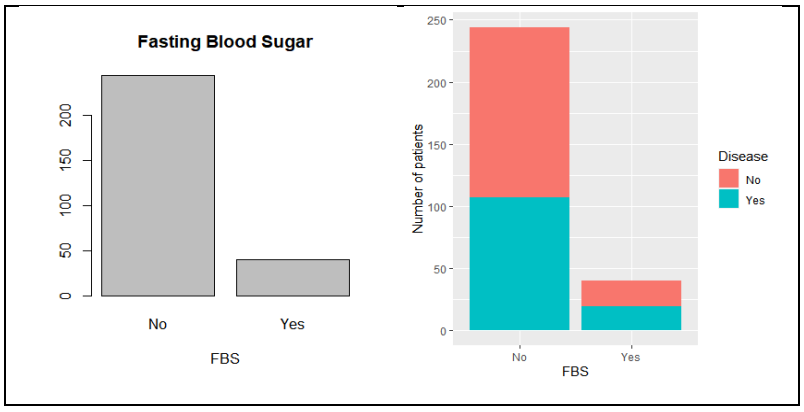
```
> summary(data$chol)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  126.0   210.8   239.5   242.5   271.0   360.0
```





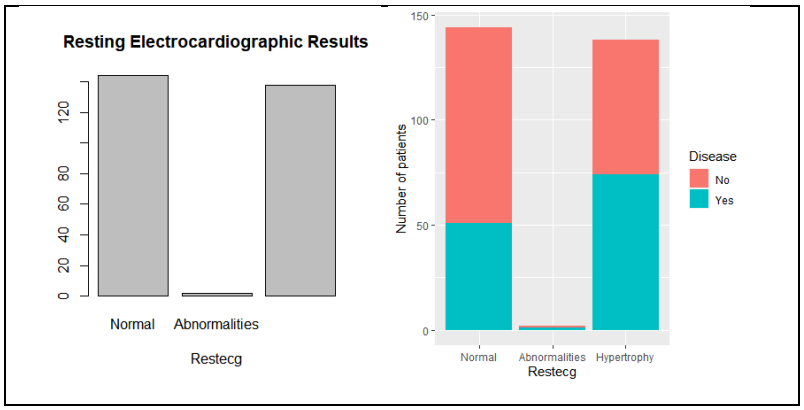
Rata – rata kadar kolesterol untuk pasian yang diperiksa dalam penelitian ini adalah 242.5 dnegan kadar terendah dan tertinggi masing – masing adalah 126mg/dL dan 360mg/dL. Distribusi kadar kolesterol pasien sangat miring ke kanan, menunjukkan bahwa hanya sedikit pasien yang memiliki kadar kolesterol sangat tinggi. Tingkat kolesterol untuk pasien tanpa penyakit jantung memiliki nilai median lebih rendah dibanding pasien dengan penyakit jantung, hal ini dapat diliat pada boxplot.

g. Variabel (Fasting Blood Sugar>120 mg/dl) 1=true; 0=false



Sebagian besar pasien memiliki kadar gula darah puasa < 120mg/dL. Hal ini tdak banyak berubah ketika dibagi menjadi kelompok dengan penyakit jantung dan tanpa penyakit jantung, karena pada gambar diatas menunjukkan pola yang sama untuk kedua kelompok pasien tersebut dan dapat diartikan bahwa kadar gula darah puasa mungkin bukan faktor penentu untuk memiliki penyakit jantung atau tidak.

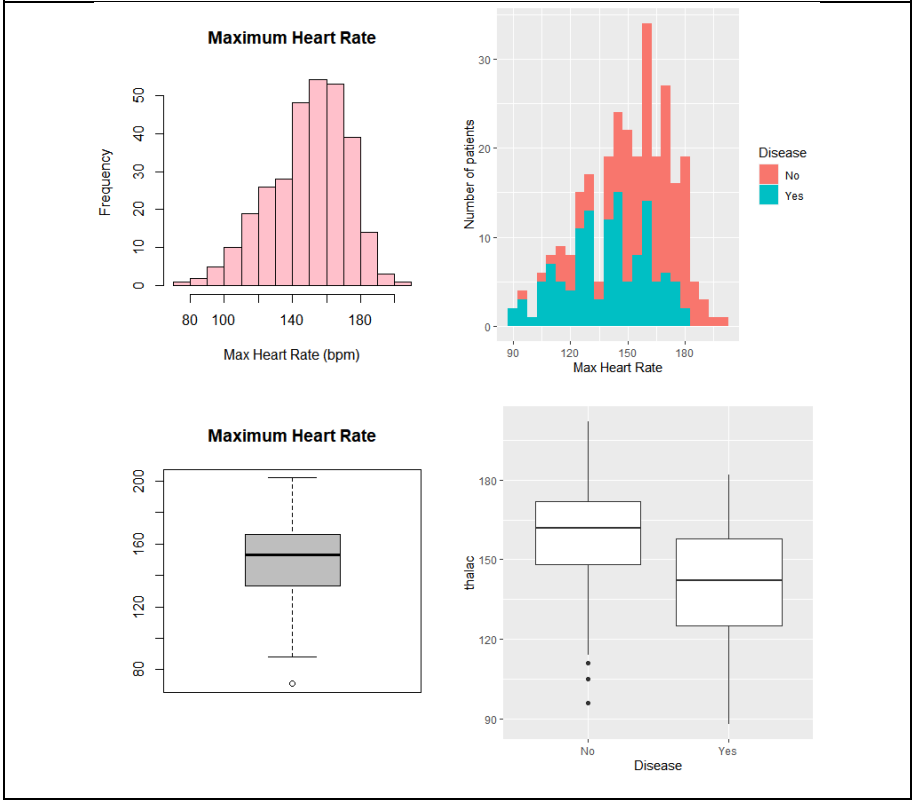
h. Variabel Resting ECG results (0=normal; 1 and 2 = abnormal)



Sebagian besar pasien menunjukkan hasil elektrokardiograf istirahat yang normal. Orang sehat atau tanpa penyakit janutng menunjukkan hasil ekg normal. Namun, proporsi yang lebih tinggi dari pasien dengan penyakit jantung memiliki pola gelombang abnormal yang menunjukkan fitur ini dapat berkontribusi pada beberapa kekuatan prediksi.

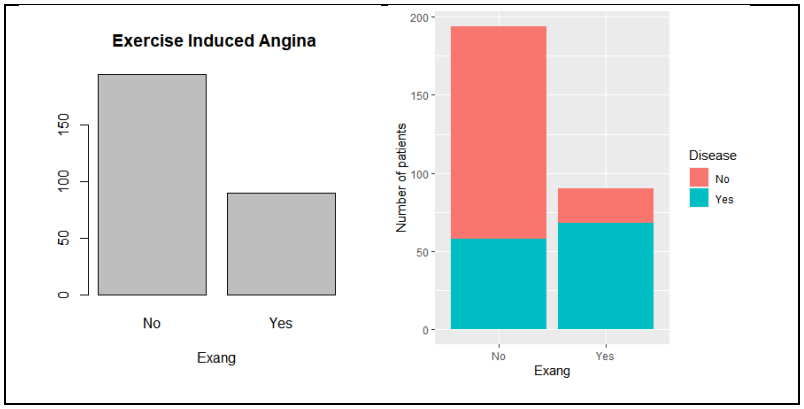
i. Variabel Maximum Heart Rate Achieved

```
> summary(data$thalac)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   88.0   133.8   153.0   150.0   168.0   202.0
```



Histogram untuk detak jantung maksimum yang dicapai oleh pasien miring ke kiri karena beberapa pasien menunjukkan detak jantung yang relatif rendah. Histogram terpisah untuk dua kelompok pasien menunjukkan orang sehat memiliki denyut jantung maksimum yang cukup tinggi (sekitar 160) dibandingkan dengan denyut jantung maksimum (150) pasien dengan penyakit jantung. Hal ini juga dapat terlihat melalui boxplot, dimana nilai median untuk denyut jantung maksimum pasien tanpa penyakit jantung lebih tinggi dibandingkan pasien dnegan penyakit jantung.

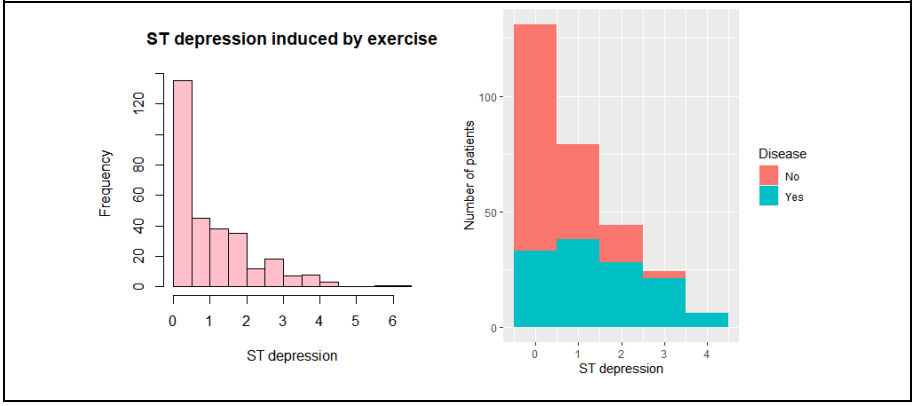
j. Variabel Exercise Induced Angina (1=yes; 0=no)

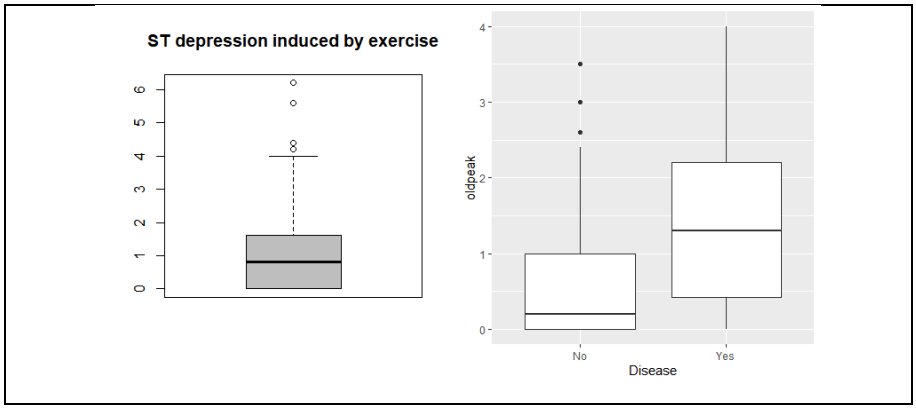


Dari seluruh pasien, mayoritas belum pernah mengalami angina yang diinduksi olahraga. Tetapi angina yang diinduksi olahraga untuk pasien dengan penyakit jantung lebih tinggi daripada pasien tanpa penyakit jantung. Ini menunjukkan bahwa angina yang diinduksi olahraga dapat menjadi faktor penentu untuk memiliki penyakit jantung.

k. Variabel ST Depression Induced by Exercise Relative to Rest

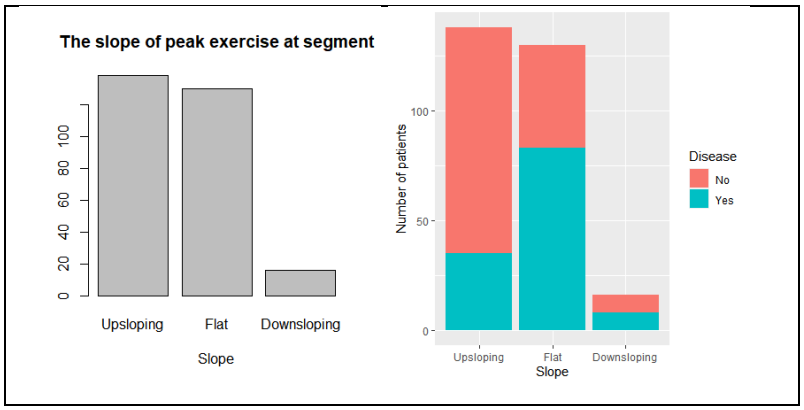
```
> summary(data$oldpeak)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.6000  0.9486  1.6000  4.0000
```





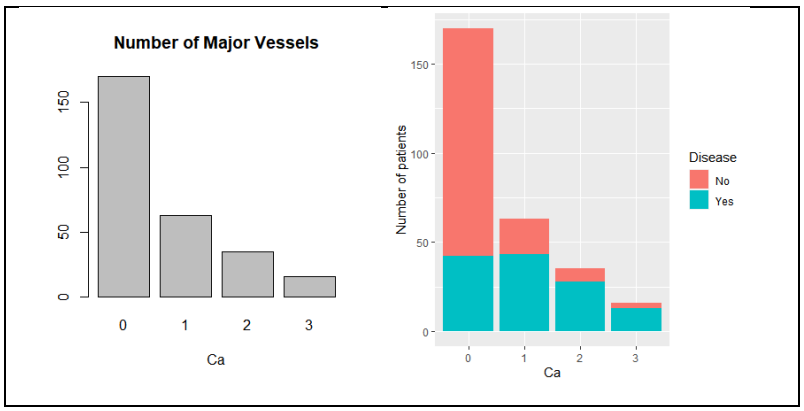
Histogram yang dihasilkan sangat miring ke kanan karena banyaknya angka *ST depression* yang rendah. Pasien tanpa penyakit jantung tercatat mayoritas sekitar 0, sedangkan pasien dengan penyakit jantung mayortias sekitar 0-4.

l. Variabel The slope of peak exercise at segment (1=upsloping; 2=flat; 3=down sloping)



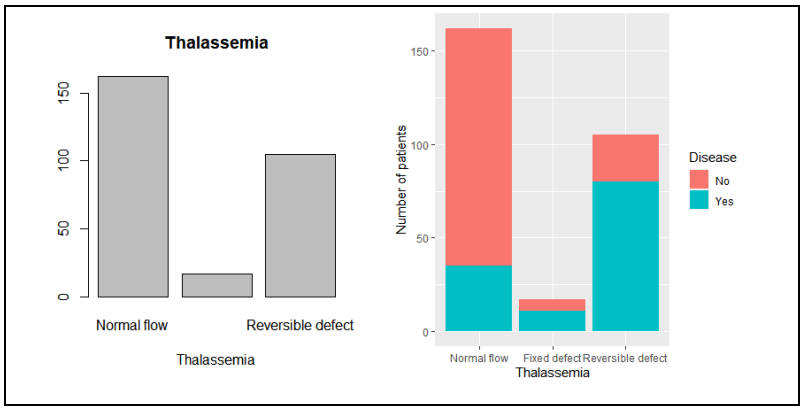
Up sloping dan *flat* adalah dua jenis utama segmen ST latihan untuk banyak pasien. Sebagian besar pasien tanpa penyakit jantung telah melakukan olahraga *upsloping* sedangkan sebagian besar pasien dengan penyakit jantung memiliki segmen ST olahraga *flat*.

m. Variabel Number of major vessels (0-3) coloured by fluoroscopy



Sebagian besar pasien melaporkan nilai ca sebagai nol dan distribusinya terlihat seperti eksponensial terbalik. Pola yang sama berlaku untuk nilai ca untuk orang sehat juga. Namun, bagi orang sakit nilai 0 dan 1 sama-sama penting.

n. Variabel Thalassemia (3=normal; 6=fixed defect; 7=reversable defect)



Pasien yang didiagnosis dengan *reversible β-Thalassemia defect* lebih banyak ditemukan pada kelompok pasien dengan penyakit jantung sementara sebagian pasien tanpa penyakit jantung menunjukkan *normal flow*. Perbedaan ini dapat berkontribusi pada prediksi penyakit jantung.

D. ANALISIS REGRESI LOGISTIK

Regresi Logistik adalah suatu metode analisis statistika untuk mendeskripsikan hubungan antara variabel terikat yang memiliki dua kategori atau lebih dengan satu atau lebih peubah bebas berskala kategori atau kontinu. Adapun regresi logistik dapat dibagi menjadi regresi logistik biner, regresi logistik multinomial dan regresi logistik ordinal.

Model regresi logistik biner digunakan untuk menganalisis hubungan antara satu variabel respon dan beberapa variabel prediktor, dengan variabel responnya berupa data kualitatif dikotomi yaitu bernilai 1 untuk menyatakan keberadaan sebuah karakteristik dan bernilai 0 untuk menyatakan ketidakberadaan sebuah karakteristik . Model regresi logistik biner digunakan jika variabel responnya menghasilkan dua kategori bernilai 0 dan 1. Bentuk model regresi logistik adalah sebagai berikut :

$$\pi(x) = \frac{\exp(\beta_0 + \beta_0 x_1 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_0 x_1 + \ldots + \beta_p x_p)}$$

Untuk mempermudah menaksir parameter regresi, maka $\pi(x)$ pada persamaan diatas ditransformasikan sehingga menghasilkan bentuk logit regresi logistik, sebagai berikut :

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_0 x_1 + \ldots + \beta_p x_p$$

PENJELASAN DATA

Data yang digunakan merupakan data sekunder yaitu data penyakit jantung yang diperoleh dari web UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/heart+disease). Dataset ini berisi 13 variabel dependen dan 1 variabel independen dengan jumlah observasi sebanyak 303 namun setelah dilakukan *pre-processing* berjumlah 284 karena terdapat 19 data *outliers.* Untuk rincian variabel dapat dilihat pada tabel berikut :

$\mathbf{X_1}$	Age	Age of patient
$\mathbf{X_2}$	Sex	Gender of Patient
$\mathbf{X_3}$	CP	Type of Chest Pain (1: Typical Angina, 2: Atypical Angina, 3: Non-anginal Pain, 4: Asymptomatic)
$\mathbf{X_4}$	Trestbps	Resting Blood Pressure (in mm Hg on admission)
$\mathbf{X_5}$	Chol	Serum Cholesterol in mg/dl
$\mathbf{X_6}$	Fbs	(Fasting Blood Sugar>120 mg/dl) 1=true; 0=false
$\mathbf{X_7}$	Restecg	Resting ECG results (0=normal; 1 and 2 = abnormal)
$\mathbf{X_8}$	Thalach	Maximum Heart Rate Achieved
$\mathbf{X_9}$	Exang	Exercise Induced Angina (1=yes; 0=no)
$\mathbf{X_{10}}$	Old peak	ST Depression Induced by Exercise Relative to Rest
$\mathbf{X_{11}}$	Slope	The slope of peak exercise at segment (1=upsloping; 2=flat; 3=down sloping)
$\mathbf{X_{12}}$	Ca	Number of major vessels (0-3) coloured by fluoroscopy
$\mathbf{X_{13}}$	Thal	Thalassemia (3=normal; 6=fixed defect; 7=reversable defect)
Y	Disease	Angiographic disease status (0=no heart disease; more than 0=have heart disease)

ANALISIS DATA

a. Stratifikasi Data

Stratifikasi data diperlukan untuk emmastikan setiap kelas respon (Disease) terambil dalam pengacakan sample. Startifikasi data dilakukan dengan fungsi filter() dalam *package* dplyr.

```
> library(dplyr)
> hn <- filter(data,Disease=="No")
> hd <- filter(data,Disease=="Yes")
```

b. Split Data

Data training digunakan untuk membangun model, sedangkan data testing digunakan untuk validasi dengan pembagian training:testing 70% : 30%.

```
> acak.hn <- sample(1:nrow(hn), 0.7*nrow(hn))
> acak.hd <- sample(1:nrow(hd), 0.7*nrow(hd))
> disease.train <- rbind(hn[acak.hn,],hd[acak.hd,])
> disease.test <- rbind(hn[-acak.hn,],hd[-acak.hd,])
```

c. Pembentukan Model

Data training digunakan untuk membangun model regresi logistik dengan menggunakan fungsi glm() karena regresi logistik termasuk ke dalam generalized linear model dengan family = binomial. Variabel yang digunakan adalah semua variabel, dimana variabel *disease* menjadi variabel responnya.

```
> logit1 <- glm(Disease ~
Age+Sex+CP+trestbps+chol+fbs+restecg+thalac+exang+oldpeak+slope+ca+t
hal, data=disease.train,family = binomial(link="logit"))
> summary(logit1)

Call:
glm(formula = Disease ~ Age + Sex + CP + trestbps + chol + fbs +
    restecg + thalac + exang + oldpeak + slope + ca + thal, family =
binomial(link = "logit"),
    data = disease.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8801  -0.4822  -0.1413   0.3532   2.5420

Coefficients:
                        Estimate Std. Error z value
(Intercept)           -7.801e+00  3.793e+00  -2.057
Age                   -7.254e-03  3.168e-02  -0.229
SexMale                1.597e+00  6.740e-01   2.370
CPAtypical angina      2.557e-01  9.396e-01   0.272
CPNo angina           -2.160e-01  8.371e-01  -0.258
CPAsymptomatic         1.887e+00  8.606e-01   2.192
trestbps               2.358e-02  1.669e-02   1.413
chol                   5.278e-03  5.888e-03   0.896
fbsYes                -7.529e-01  7.230e-01  -1.041
restecgAbnormalities   1.254e+01  1.455e+03   0.009
restecgHypertrophy     9.993e-01  5.084e-01   1.966
thalac                -8.710e-03  1.577e-02  -0.552
exangYes               1.182e+00  5.597e-01   2.111
oldpeak                5.262e-01  3.247e-01   1.621
slopeFlat              1.021e+00  5.664e-01   1.802
slopeDownsloping      -2.289e-01  1.235e+00  -0.185
ca1                    -1.622e+0  6.499e-01   2.496
```

```
ca2                    -2.489e+0  1.019e+00   2.442
ca3                    -1.642e+0  1.092e+00   1.504
thalFixed defect        3.437e-01  9.871e-01   0.348
thalReversible defect   1.491e+00  5.662e-01   2.633
                       Pr(>|z|)
(Intercept)             0.03970 *
Age                     0.81892
SexMale                 0.01779 *
CPAtypical angina       0.78551
CPNo angina             0.79640
CPAsymptomatic          0.02835 *
trestbps                0.15774
chol                    0.37002
fbsYes                  0.29775
restecgAbnormalities    0.99312
restecgHypertrophy      0.04934 *
thalac                  0.58080
exangYes                0.03474 *
oldpeak                 0.10511
slopeFlat               0.07149 .
slopeDownsloping        0.85302
ca1                     0.01255 *
ca2                     0.01460 *
ca3                     0.13251
thalFixed defect        0.72767
thalReversible defect   0.00847 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 272.04  on 197  degrees of freedom
Residual deviance: 122.78  on 177  degrees of freedom
AIC: 164.78

Number of Fisher Scoring iterations: 14
```

Berdasarkat output diatas, dapat dilihat bahwa dari tiga belas (13) variabel hanya terdapat enam (7) variabel yang signifikan variabel sex, cp, restecg, exang, slope, ca dan thal sedangkan enam (6) variabel lainnya tidak signifikan. Signifikan yang dimaksud adalah variabel yang memiliki nilai p-value < α . Kemudian akan dibentuk model kedua, dimana variabel yang tidak signifikan pada model 1 akan dikeluarkan dengan menggunakan *stepwise regression* metode *backward.*

```
> logit2 <- step(logit1,direction = "backward")
> summary(logit2)

Call:
glm(formula = Disease ~ Sex + CP + restecg + exang + oldpeak +
    slope + ca + thal, family = binomial(link = "logit"), data =
disease.train)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6013  -0.4668  -0.1685   0.3347   2.8980

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             -5.0337     1.1050  -4.555 5.23e-06 ***
SexMale                  1.2575     0.6015   2.091  0.03657 *
CPAtypical angina        0.2127     0.8793   0.242  0.80890
CPNo angina             -0.4080     0.7943  -0.514  0.60754
CPAsymptomatic           1.6299     0.7722   2.111  0.03480 *
restecgAbnormalities    12.4559  1455.3979   0.009  0.99317
restecgHypertrophy       1.1936     0.4903   2.435  0.01491 *
exangYes                -1.1857     0.5326   2.226  0.02601 *
oldpeak                 -0.5660     0.3197   1.770  0.07668 .
slopeFlat               -1.0792     0.5307   2.034  0.04198 *
slopeDownsloping         0.3816     1.2001  -0.318  0.75050
ca1                     -1.4473     0.5928   2.442  0.01462 *
ca2                     -2.3332     0.9571   2.438  0.01478 *
ca3                     -1.5207     0.9234   1.647  0.09959 .
thalFixed defect         0.5544     0.9043   0.613  0.53982
thalReversible defect    1.6698     0.5535   3.017  0.00256 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 272.04  on 197  degrees of freedom
Residual deviance: 126.81  on 182  degrees of freedom
AIC: 158.81

Number of Fisher Scoring iterations: 14
```

Berdasarkan output diatas, dapat dilihat bahwa semua variabel telah signifikan setelah dilakukan pengeluaran variabel yang tidak signifikan.

d. Menentukkan Model Terbaik

Langkah selanjutnya adalah menentukkan model terbaik dengan cara melihat nilai AIC yang paling kecil.

```
> AIC <- data.frame(logit1$aic,logit2$aic) ; AIC
  logit1.aic logit2.aic
1   164.7826   158.8127
```

Berdasarkan output diatas, diperoleh nilai AIC untuk masing - masing model, dimana model 1 memiliki nilai AIC sebesar 164.78 sedangkan model 2 memiliki nilai AIC sebesar 158.8127. Oleh karena itu dapat disimpulkan bahwa model 2 adalah model terbaik karena memiliki nilai AIC terendah.

e. Model Regresi Logistik

Berdasarkan pemodelan regresi logistik yang telah dilakukan pada data penyakit jantung diatas, diperoleh kesimpulan sebagai berikut :

1. Dalam analisis ini dibentuk dua (2) model, dimana model 1 merupakan model awal dan model 2 merupakan model setelah variabel yang tidak signifikan pada model 1 dihilangkan menggunakan *stepwise regression* metode *backward.* Didapat model 2 memiliki nilai AIC lebih rendah dibandingkan model 1. Oleh karena itu model 2 merupakan model terbaik.
2. Berdasarkan model yang dipleroleh, penyakit jantung pada pasien dipengaruhi secara signifikan oleh *sex, cp, restecg, exang, oldpeak, slope, ca,* dan *thal*. Peningkatan nilai variabel tersebut akan berdampak pada kinerja kardiovaskular secara keseluruhan dimana kinerja kardiovaskular akan menurun, sedangkan potensi penyakit kardiovaskular diprediksi akan meningkat. Variabel yang dipilih ada variabel yang berpengaruh secara signifikan terhadap variabel respon.
3. Diperoleh persamaan model rgeresi logistik pada model data penyakit jantung adalah sebagai berikut :

$$Logit[\pi(x)] = g(x) \ = -5.0337 + 1.2575\ SexMale + 1.6299\ CpAsymptomatic + 1.1936\ rectecgHypertrophy - 1.1857\ exangYes - 0.5660 Oldpeak - 1.0792\ slopeFlat - 1.4473\ ca1 - 2.3332\ ca2 - 1.5207\ ca3 + 1.6698\ thalReversibledefect$$

4. Variabel exang, oldpeak, slope, dan ca berpengaruh negatif terhadap respon, sedangkan variabel sex, restecg, dan thal berpengaruh positif terhadap respon.

f. Performance Matrics

Pada data mining ada bebeapa cara untuk mengukur kinerja dari model yang dihasilkan salah satunya menggunakan confusion matriks (akurasi). *Confusion matrix* adalah suatu metode yang digunakan untuk melakukan perhitungan akurasi pada konsep data mining**. *Accuracy*** menggambarkan seberapa akurat model dapat mengklasifikasikan dengan benar. Maka, *accuracy* merupakan rasio prediksi benar (positif dan negatif) dengan keseluruhan data. Dengan kata lain, *accuracy* merupakan tingkat kedekatan nilai prediksi dengan nilai aktual (sebenarnya). ***Precision*** menggambarkan tingkat keakuratan antara data yang diminta dengan hasil prediksi yang diberikan oleh model. Maka, *precision* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positf. Dari semua kelas positif yang telah di prediksi dengan benar, berapa banyak data yang benar-benar positif. ***Sensitivity*** menggambarkan keberhasilan model

dalam menemukan kembali sebuah informasi. Maka, *sensitivity* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif.

Prediksi respon pada data testing dilakukan menggunakan threshold = 0.5 dimana jika nilai probabilitas prediksi > 0.5 maka akan diprediksi di kelas event (disease) dan sebaliknya akan diprediksi masuk di kelas non event (no disease).

```
> # Prediksi
> prob.prediksi<-predict(logit2, disease.test, type="response")
> prediksi<-ifelse(prob.prediksi>0.5,"Yes","No")
> pred.aktual<-
data.frame("Prediksi"=prediksi,"Aktual"=disease.test$Disease)
> head(pred.aktual)
   Prediksi Aktual
2        No     No
3        No     No
9        No     No
23       No     No
24       No     No
32      Yes     No
```

Selanjutnya akan dibentuk *confusion matrix* untuk mengukur performa dari model yang telah dibentuk.

```
> library(caret)
> confusionMatrix(as.factor(prediksi), disease.test$Disease)
Confusion Matrix and Statistics

          Reference
Prediction No Yes
       No  36   5
       Yes 12  33

               Accuracy : 0.8023
                 95% CI : (0.7025, 0.8804)
    No Information Rate : 0.5581
    P-Value [Acc > NIR] : 1.773e-06

                  Kappa : 0.6068

 Mcnemar's Test P-Value : 0.1456

            Sensitivity : 0.7500
            Specificity : 0.8684
         Pos Pred Value : 0.8780
         Neg Pred Value : 0.7333
             Prevalence : 0.5581
         Detection Rate : 0.4186
   Detection Prevalence : 0.4767
      Balanced Accuracy : 0.8092

       'Positive' Class : No
```

Dari Output di atas, untuk hasil confusion matrix prediksi yang meleset cukup kecil yaitu 5 untuk *false positive* dan 12 untuk *false negative*. Berdasarkan hasil confusionMatrix dalam analisis ini, dapat kita ambil informasi bahwa:

- ***Accuracy***

  Kemampuan model dalam menebak target Y (**Disease** dan **No Disease**) sebesar 80.23%. Nilai akurasi sebesar 80.23% menunjukkan bahwa hasil klasifikasi dengan regresi logistik pada data ini sudah baik.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{36 + 33}{36 + 33 + 5 + 12} = 0.8023$$

- ***Sensitivity***

  Sedangkan dari keluruhan data aktual, orang yang **Health** atau **No Disease**, model dapat mampu menebak benar sebesar 75%.

$$Sensitivity = \frac{TP}{TP + FN} = \frac{36}{36 + 12} = 0.75$$

- ***Precision***

  Besar pasien yang benar terdeteksi penyakit jantung dari keseluruhan pasien yang diprediksi menderita penyakit jantung adalah 87.8%

$$Precision = \frac{TP}{TP + FP} = \frac{36}{36 + 5} = 0.878$$

- ***Specificity***

  Dari keseluruhan data aktual, orang yang **No Health** atau **Disease**, model mampu menebak dengan benar sebesar 86.84%.

$$Specificity = \frac{TN}{TN + FP} = \frac{33}{33 + 5} = 0.8684$$

Lampiran

IMPORT DATA

```
>  damin  <-  read.csv2("C:/Users/ASUS/Downloads/damin.csv",  na  =  "?",
stringsAsFactors=TRUE)
> data <- damin
> View(data)
> attach(data)
> dim(data)
[1] 303  14
> str(damin)
'data.frame':   303 obs. of  14 variables:
 $ Age     : int  63 67 67 37 41 56 62 57 63 53 ...
 $ Sex     : int  1 1 1 1 0 1 0 0 1 1 ...
 $ CP      : int  1 4 4 3 2 2 4 4 4 4 ...
 $ trestbps: int  145 160 120 130 130 120 140 120 130 140 ...
 $ chol    : int  233 286 229 250 204 236 268 354 254 203 ...
 $ fbs     : int  1 0 0 0 0 0 0 0 0 1 ...
 $ restecg : int  2 2 2 0 2 0 2 0 2 2 ...
 $ thalac  : int  150 108 129 187 172 178 160 163 147 155 ...
 $ exang   : int  0 1 1 0 0 0 0 1 0 1 ...
 $ oldpeak : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
 $ slope   : int  3 2 2 3 1 1 3 1 2 3 ...
 $ ca      : int  0 3 2 0 0 0 2 0 1 0 ...
 $ thal    : int  6 3 7 3 3 3 3 3 7 7 ...
 $ num     : int  0 2 1 0 0 0 3 0 2 1 ...
> #int/num = age, trestbps, chol, thalac, oldpeak
> #factor  = Sex, CP, fbs, restecg, exang, slope, ca, thal, num
> data$Sex <- factor(data$Sex)
> data$CP <- factor(data$CP)
> data$fbs <- factor(data$fbs)
> data$restecg <- factor(data$restecg)
> data$exang <- factor(data$exang)
> data$slope <- factor(data$slope)
> data$ca <- factor(data$ca)
> data$thal <- factor(data$thal)
> data$num <- factor(data$num)
> str(data) #cek tipe data lagi
'data.frame':   303 obs. of  14 variables:
 $ Age     : int  63 67 67 37 41 56 62 57 63 53 ...
 $ Sex     : Factor w/ 2 levels "0","1": 2 2 2 2 1 2 1 1 2 2 ...
 $ CP      : Factor w/ 4 levels "1","2","3","4": 1 4 4 3 2 2 4 4 4 4 ...
 $ trestbps: int  145 160 120 130 130 120 140 120 130 140 ...
 $ chol    : int  233 286 229 250 204 236 268 354 254 203 ...
 $ fbs     : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 1 2 ...
 $ restecg : Factor w/ 3 levels "0","1","2": 3 3 3 1 3 1 3 1 3 3 ...
 $ thalac  : int  150 108 129 187 172 178 160 163 147 155 ...
 $ exang   : Factor w/ 2 levels "0","1": 1 2 2 1 1 1 1 2 1 2 ...
 $ oldpeak : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
 $ slope   : Factor w/ 3 levels "1","2","3": 3 2 2 3 1 1 3 1 2 3 ...
 $ ca      : Factor w/ 4 levels "0","1","2","3": 1 4 3 1 1 1 3 1 2 1 ...
 $ thal    : Factor w/ 3 levels "3","6","7": 2 1 3 1 1 1 1 1 3 3 ...
 $ num     : Factor w/ 5 levels "0","1","2","3",..: 1 3 2 1 1 1 4 1 3 2
...
```

STATISTIKA DESKRIPTIF

```
> levels(data$Sex) <- c("Female", "Male")
> levels(data$CP) <- c("Typical angina", "Atypical angina", "No angina",
"Asymptomatic")
> levels(data$fbs) <- c("No", "Yes")
> levels(data$restecg) <- c("Normal", "Abnormalities","Hypertrophy")
> levels(data$exang) <- c("No", "Yes")
> levels(data$slope) <- c("Upsloping", "Flat", "Downsloping")
> levels(data$thal) <- c("Normal flow", "Fixed defect", "Reversible
defect")
> levels(data$Disease) <- c("No", "Yes")
>
> # STATISTIKA DESKRIPTIF ============================
> # QUANTITATIF ( stat desk, hist, boxplot)
> # AGE
> summary(data$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  29.00   47.00   55.00   54.07   60.00   77.00
> hist(Age,data = data, main="Age of Patient", xlab="years", col="pink")
> ggplot(data, aes(Age, fill=Disease)) +
+   geom_histogram(binwidth=5) +
+   labs(fill="Disease", x="Age", y="Number of patients")
> boxplot(Age,data = data, main="Age of Patient", col="grey")
> ggplot(data , aes(x = Disease, y = Age)) +
+   geom_boxplot()
>
> # TRESTBPS
> summary(data$trestbps)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     94     120     130     130     140     170
> hist(trestbps,data = data, main="Resting Blood Pressure", xlab="Blood
Pressure(mmHg)", col="pink")
> ggplot(data, aes(trestbps, fill=Disease)) +
+   geom_histogram(binwidth=5) +
+   labs(fill="Disease", x="Resting Blood Pressure", y="Number of
patients")
> boxplot(trestbps,data = data, main="Resting Blood Pressure", col="grey")
> ggplot(data , aes(x = Disease, y = trestbps)) +
+   geom_boxplot()
>
> # CHOL
> summary(data$chol)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  126.0   210.8   239.5   242.5   271.0   360.0
> hist(chol,data = data, main="Cholestrol", xlab="Cholestrol(mg/dL",
col="pink")
> ggplot(data, aes(chol, fill=Disease)) +
+   geom_histogram(binwidth=5) +
+   labs(fill="Disease", x="Cholestrol", y="Number of patients")
> boxplot(chol,data = data, main="Cholestrol", col="grey")
> ggplot(data , aes(x = Disease, y = chol)) +
+   geom_boxplot()
>
> # THALAC
> summary(data$thalac)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   88.0   133.8   153.0   150.0   168.0   202.0
> hist(thalac,data = data, main="Maximum Heart Rate", xlab="Max Heart Rate
(bpm)", col="pink")
> ggplot(data, aes(thalac, fill=Disease)) +
+   geom_histogram(binwidth=5) +
+   labs(fill="Disease", x="Max Heart Rate", y="Number of patients")
```

```
> boxplot(thalac,data = data, main="Maximum Heart Rate", col="grey")
> ggplot(data , aes(x = Disease, y = thalac)) +
+   geom_boxplot()
>
> # OLDPEAK
> summary(data$oldpeak)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.6000  0.9486  1.6000  4.0000
> hist(oldpeak,data = data, main="ST depression induced by exercise",
xlab="ST depression", col="pink")
> ggplot(data, aes(oldpeak, fill=Disease)) +
+   geom_histogram(binwidth=5) +
+   labs(fill="Disease", x="ST depression", y="Number of patients")
> boxplot(oldpeak,data = data, main="ST depression induced by exercise",
col="grey")
> ggplot(data , aes(x = Disease, y = oldpeak)) +
+   geom_boxplot()
>
> # QUALITATIF
> # Sex
> counts <- table(data$Sex)
> barplot(counts, main="Gender of Patient",
+         xlab="Sex")
> ggplot(data, aes(Sex, fill=Disease)) +
+   geom_bar() +
+   labs(fill="Disease", x="Sex", y="Number of patients")
>
> # CP
> counts <- table(data$CP)
> barplot(counts, main="Type of Chest Pain",
+         xlab="Type")
> ggplot(data, aes(CP, fill=Disease)) +
+   geom_bar() +
+   labs(fill="Disease", x="Type", y="Number of patients")
>
> # Fbs
> counts <- table(data$fbs)
> barplot(counts, main="Fasting Blood Sugar",
+         xlab="FBS")
> ggplot(data, aes(fbs, fill=Disease)) +
+   geom_bar() +
+   labs(fill="Disease", x="FBS", y="Number of patients")
>
> # restecg
> counts <- table(data$restecg)
> barplot(counts, main="Resting Electrocardiographic Results",
+         xlab="Restecg")
> ggplot(data, aes(restecg, fill=Disease)) +
+   geom_bar() +
+   labs(fill="Disease", x="Restecg", y="Number of patients")
>
> # exang
> counts <- table(data$exang)
> barplot(counts, main="Exercise Induced Angina ",
+         xlab="Exang")
> ggplot(data, aes(exang, fill=Disease)) +
+   geom_bar() +
+   labs(fill="Disease", x="Exang", y="Number of patients")
>
> # slope
> counts <- table(data$slope)
```

```
> barplot(counts, main="The slope of peak exercise at segment",
+         xlab="Slope")
> ggplot(data, aes(slope, fill=Disease)) +
+   geom_bar() +
+   labs(fill="Disease", x="Slope", y="Number of patients")
>
> # ca
> counts <- table(data$ca)
> barplot(counts, main="Number of Major Vessels",
+         xlab="Ca")
> ggplot(data, aes(ca, fill=Disease)) +
+   geom_bar() +
+   labs(fill="Disease", x="Ca", y="Number of patients")
>
> # thal
> counts <- table(data$thal)
> barplot(counts, main="Thalassemia",
+         xlab="Thalassemia")
> ggplot(data, aes(thal, fill=Disease)) +
+   geom_bar() +
+   labs(fill="Disease", x="Thalassemia", y="Number of patients")
>
> # Disease
> counts <- table(data$Disease)
> barplot(counts, main="Angiographic disease status ",
+         xlab="Disease")
> ggplot(data, aes(Disease, fill=Disease)) +
+   geom_bar() +
+   labs(fill="Disease", x="Disease", y="Number of patients")
>
```

ANALISIS REGRESI LOGISTIK

```
> # ANALISIS REGRESI LOGISTIK ====================
> # Stratifikasi Data
> library(dplyr)
> hn <- filter(data,Disease=="No")
> hd <- filter(data,Disease=="Yes")

> # Split Data
> acak.hn <- sample(1:nrow(hn), 0.7*nrow(hn))
> acak.hd <- sample(1:nrow(hd), 0.7*nrow(hd))
> disease.train <- rbind(hn[acak.hn,],hd[acak.hd,])
> disease.test <- rbind(hn[-acak.hn,],hd[-acak.hd,])

> #Pembentukan Model
> logit1 <- glm(Disease ~
Age+Sex+CP+trestbps+chol+fbs+restecg+thalac+exang+oldpeak+slope+ca+thal,
data=disease.train,family = binomial(link="logit"))
> summary(logit1)

Call:
glm(formula = Disease ~ Age + Sex + CP + trestbps + chol + fbs +
    restecg + thalac + exang + oldpeak + slope + ca + thal, family =
binomial(link = "logit"),
    data = disease.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8801  -0.4822  -0.1413   0.3532   2.5420
```

```
Coefficients:
                          Estimate Std. Error z value
(Intercept)             -7.801e+00  3.793e+00  -2.057
Age                     -7.254e-03  3.168e-02  -0.229
SexMale                  1.597e+00  6.740e-01   2.370
CPAtypical angina        2.557e-01  9.396e-01   0.272
CPNo angina             -2.160e-01  8.371e-01  -0.258
CPAsymptomatic           1.887e+00  8.606e-01   2.192
trestbps                 2.358e-02  1.669e-02   1.413
chol                     5.278e-03  5.888e-03   0.896
fbsYes                  -7.529e-01  7.230e-01  -1.041
restecgAbnormalities     1.254e+01  1.455e+03   0.009
restecgHypertrophy       9.993e-01  5.084e-01   1.966
thalac                  -8.710e-03  1.577e-02  -0.552
exangYes                 1.182e+00  5.597e-01   2.111
oldpeak                  5.262e-01  3.247e-01   1.621
slopeFlat                1.021e+00  5.664e-01   1.802
slopeDownsloping        -2.289e-01  1.235e+00  -0.185
ca1                      1.622e+00  6.499e-01   2.496
ca2                      2.489e+00  1.019e+00   2.442
ca3                      1.642e+00  1.092e+00   1.504
thalFixed defect         3.437e-01  9.871e-01   0.348
thalReversible defect    1.491e+00  5.662e-01   2.633
                        Pr(>|z|)
(Intercept)              0.03970 *
Age                      0.81892
SexMale                  0.01779 *
CPAtypical angina        0.78551
CPNo angina              0.79640
CPAsymptomatic           0.02835 *
trestbps                 0.15774
chol                     0.37002
fbsYes                   0.29775
restecgAbnormalities     0.99312
restecgHypertrophy       0.04934 *
thalac                   0.58080
exangYes                 0.03474 *
oldpeak                  0.10511
slopeFlat                0.07149 .
slopeDownsloping         0.85302
ca1                      0.01255 *
ca2                      0.01460 *
ca3                      0.13251
thalFixed defect         0.72767
thalReversible defect    0.00847 **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 272.04  on 197  degrees of freedom
Residual deviance: 122.78  on 177  degrees of freedom
AIC: 164.78

Number of Fisher Scoring iterations: 14

> logit2 <- step(logit1,direction = "backward")
Start:  AIC=164.78
Disease ~ Age + Sex + CP + trestbps + chol + fbs + restecg +
```

```
    thalac + exang + oldpeak + slope + ca + thal

           Df Deviance    AIC
- Age       1   122.83 162.84
- thalac    1   123.09 163.09
- chol      1   123.60 163.60
- fbs       1   123.90 163.90
<none>          122.78 164.78
- trestbps  1   124.81 164.81
- restecg   2   126.93 164.93
- slope     2   127.03 165.03
- oldpeak   1   125.55 165.55
- exang     1   127.33 167.33
- thal      2   130.37 168.37
- Sex       1   128.89 168.89
- ca        3   134.66 170.66
- CP        3   136.31 172.31

Step:  AIC=162.84
Disease ~ Sex + CP + trestbps + chol + fbs + restecg + thalac +
    exang + oldpeak + slope + ca + thal

           Df Deviance    AIC
- thalac    1   123.09 161.09
- chol      1   123.62 161.62
- fbs       1   123.99 161.99
<none>          122.83 162.84
- trestbps  1   124.86 162.86
- restecg   2   126.93 162.93
- slope     2   127.07 163.07
- oldpeak   1   125.64 163.64
- exang     1   127.49 165.49
- thal      2   130.46 166.46
- Sex       1   128.93 166.93
- ca        3   135.70 169.70
- CP        3   136.36 170.36

Step:  AIC=161.09
Disease ~ Sex + CP + trestbps + chol + fbs + restecg + exang +
    oldpeak + slope + ca + thal

           Df Deviance    AIC
- chol      1   123.77 159.77
- fbs       1   124.38 160.38
- trestbps  1   125.01 161.01
<none>          123.09 161.09
- restecg   2   127.23 161.24
- oldpeak   1   125.96 161.96
- slope     2   128.40 162.40
- exang     1   128.21 164.21
- thal      2   131.01 165.01
- Sex       1   129.03 165.03
- CP        3   137.14 169.14
- ca        3   137.21 169.21

Step:  AIC=159.77
Disease ~ Sex + CP + trestbps + fbs + restecg + exang + oldpeak +
    slope + ca + thal

           Df Deviance    AIC
- fbs       1   125.04 159.04
```

```
<none>           123.77 159.77
- trestbps  1    125.88 159.88
- restecg   2    128.23 160.23
- oldpeak   1    126.81 160.81
- slope     2    129.49 161.49
- exang     1    129.04 163.04
- Sex       1    129.19 163.19
- thal      2    131.61 163.61
- CP        3    138.21 168.21
- ca        3    138.49 168.49

Step:  AIC=159.04
Disease ~ Sex + CP + trestbps + restecg + exang + oldpeak + slope +
    ca + thal

           Df Deviance    AIC
- trestbps  1   126.81 158.81
<none>          125.04 159.04
- restecg   2   129.94 159.94
- slope     2   130.38 160.38
- oldpeak   1   128.47 160.47
- Sex       1   130.06 162.06
- exang     1   130.22 162.22
- thal      2   133.85 163.85
- ca        3   138.56 166.56
- CP        3   140.24 168.24

Step:  AIC=158.81
Disease ~ Sex + CP + restecg + exang + oldpeak + slope + ca +
    thal

          Df Deviance    AIC
<none>         126.81 158.81
- oldpeak  1   130.13 160.13
- slope    2   132.64 160.64
- restecg  2   133.31 161.31
- Sex      1   131.40 161.40
- exang    1   131.85 161.85
- thal     2   136.65 164.65
- ca       3   139.72 165.72
- CP       3   140.43 166.43
> summary(logit2)

Call:
glm(formula = Disease ~ Sex + CP + restecg + exang + oldpeak +
    slope + ca + thal, family = binomial(link = "logit"), data =
disease.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6013  -0.4668  -0.1685   0.3347   2.8980

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -5.0337     1.1050  -4.555 5.23e-06
SexMale                1.2575     0.6015   2.091  0.03657
CPAtypical angina      0.2127     0.8793   0.242  0.80890
CPNo angina           -0.4080     0.7943  -0.514  0.60754
CPAsymptomatic         1.6299     0.7722   2.111  0.03480
restecgAbnormalities  12.4559  1455.3979   0.009  0.99317
restecgHypertrophy     1.1936     0.4903   2.435  0.01491
```

```
exangYes                  1.1857     0.5326   2.226  0.02601
oldpeak                   0.5660     0.3197   1.770  0.07668
slopeFlat                 1.0792     0.5307   2.034  0.04198
slopeDownsloping         -0.3816     1.2001  -0.318  0.75050
ca1                       1.4473     0.5928   2.442  0.01462
ca2                       2.3332     0.9571   2.438  0.01478
ca3                       1.5207     0.9234   1.647  0.09959
thalFixed defect          0.5544     0.9043   0.613  0.53982
thalReversible defect     1.6698     0.5535   3.017  0.00256

(Intercept)            ***
SexMale                *
CPAtypical angina
CPNo angina
CPAsymptomatic         *
restecgAbnormalities
restecgHypertrophy     *
exangYes               *
oldpeak                .
slopeFlat              *
slopeDownsloping
ca1                    *
ca2                    *
ca3                    .
thalFixed defect
thalReversible defect  **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 272.04  on 197  degrees of freedom
Residual deviance: 126.81  on 182  degrees of freedom
AIC: 158.81

Number of Fisher Scoring iterations: 14

> # Memilih model terbaik dengan aic (model 2)
> AIC <- data.frame(logit1$aic,logit2$aic) ; AIC
  logit1.aic logit2.aic
1   164.7826   158.8127
>
> # Prediksi
> prob.prediksi<-predict(logit2, disease.test, type="response")
> prediksi<-ifelse(prob.prediksi>0.5,"Yes","No")
> pred.aktual<-
data.frame("Prediksi"=prediksi,"Aktual"=disease.test$Disease)
> head(pred.aktual)
   Prediksi Aktual
8        No     No
9        No     No
10       No     No
13      Yes     No
15       No     No
18       No     No
>
> # Akurasi dan Presisi
> library(caret)
> akurasi = confusionMatrix(as.factor(prediksi), disease.test$Disease)
> akurasi
```

```
Confusion Matrix and Statistics

          Reference
Prediction No Yes
       No  36   5
       Yes 12  33

               Accuracy : 0.8023
                 95% CI : (0.7025, 0.8804)
    No Information Rate : 0.5581
    P-Value [Acc > NIR] : 1.773e-06

                  Kappa : 0.6068

 Mcnemar's Test P-Value : 0.1456

            Sensitivity : 0.7500
            Specificity : 0.8684
         Pos Pred Value : 0.8780
         Neg Pred Value : 0.7333
             Prevalence : 0.5581
         Detection Rate : 0.4186
   Detection Prevalence : 0.4767
      Balanced Accuracy : 0.8092

       'Positive' Class : No
```