Jurnal Matematika Thales (JMT), 2024
Vol. 06 No. 02

BEBERAPA MODEL STATISTIK PENYAKIT JANTUNG (SOME STATISTICAL HEART DISEASE MODELS)

LIANA ISNAINI, MUH FADLAN, VINSENSIUS SETIA D. S. RUEK, ANISYA ANANDA PERMATASARI

Abstract. This article aims to develop and compare some statistical models to predict heart disease diagnosis. We conducted a literature study to find suitable statistical models to perform a binary classification for our Heart Disease dataset. This article analyzes common factors and parameters used by medical practitioners to see how big their impacts towards determining whether someone has a heart disease or not.
Keywords: Heart Disease, KNN, Logistic, Decision Tree.

Abstrak. Artikel ini bertujuan untuk mengembangkan dan membandingkan beberapa model statistik untuk memprediksi diagnosis penyakit jantung. Studi literatur dilakukan untuk memilih model statistika yang sesuai untuk melakukan klasifikasi biner untuk *dataset* penyakit jantung yang dimiliki. Artikel ini menganalisa fatkor-faktor sekaligus parameter-parameter yang umum digunakan oleh praktisi medis untuk melihat seberapa besar pengaruh mereka terhadap penentuan apakah seseorang memiliki penyakit jantung atau tidak.
Kata-kata kunci: Penyakit Jantung, KNN, Logistik, Pohon Keputusan.

1. PENDAHULUAN

Penyakit jantung adalah salah satu penyebab utama kematian di dunia. Para peneliti dan ahli medis telah berusaha mencari penyebab penyakit jantung selama bertahun-tahun. Selain itu, praktisi medis semakin banyak menggunakan variabel untuk mengklasifikasikan kondisi jantung seseorang. Misalnya, beberapa tes yang dikenal luas untuk mendiagnosis penyakit jantung adalah tes darah, ECG, dan tes stres.

2010 Mathematics Subject Classification: 62J12, 62H30, 62P10

Ada beberapa faktor yang dapat menentukan apakah seseorang diklasifikasikan memiliki penyakit jantung atau tidak, seperti usia, jenis kelamin, tipe nyeri dada, tekanan darah istirahat, kolesterol serum, gula darah puasa, hasil elektrokardiogram istirahat, detak jantung maksimal yang dicapai, angina yang diinduksi oleh latihan, depresi ST Oldpeak, dan kemiringan segmen ST puncak saat latihan.

Jenis kelamin dapat mempengaruhi risiko mengembangkan penyakit jantung dalam beberapa cara. Wanita memiliki perlindungan hormonal pramenopause (estrogen) yang mungkin melindungi dari penyakit jantung. Setelah menopause, perlindungan ini berkurang karena penurunan kadar estrogen. Ada perbedaan dalam prevalensi faktor risiko seperti tekanan darah tinggi, kadar kolesterol, dan diabetes antara pria dan wanita. Wanita dan pria bisa menunjukkan gejala yang berbeda saat mengalami serangan jantung. Wanita lebih mungkin mengalami gejala atipikal seperti nyeri dada yang kurang parah atau gejala lain seperti sesak napas, mual, atau nyeri punggung [5]. Wanita dan pria mungkin merespon pengobatan medis seperti obat-obatan dan intervensi bedah dengan cara yang berbeda. Hal ini dapat mempengaruhi hasil pengobatan dan prognosis setelah menderita penyakit jantung. Faktor seperti diet, tingkat aktivitas fisik, dan kebiasaan merokok bisa berbeda antara pria dan wanita, yang secara langsung dapat mempengaruhi risiko mereka terkena penyakit jantung.

Tekanan darah tinggi memaksa jantung bekerja lebih keras untuk memompa darah ke seluruh tubuh. Hal ini menyebabkan ruang jantung kiri bagian bawah, yang disebut ventrikel kiri, menebal dan membesar. Ventrikel kiri yang menebal dan membesar meningkatkan risiko serangan jantung dan gagal jantung [3]. Ini juga meningkatkan risiko kematian mendadak ketika jantung tiba-tiba berhenti berdetak, yang disebut kematian jantung mendadak. Tes glukosa puasa dilakukan dengan melihat konsentrasi glukosa dalam plasma darah yang didahului dengan puasa selama 8-12 jam, ini sering disebut tes Gula Darah Puasa (GDP) [4]. Kadar glukosa puasa diukur menggunakan metode kolorimetri enzimatik. Hasil gula darah puasa dinyatakan dalam mg/dL. Berpuasa sebelum tes darah berarti seseorang tidak boleh makan atau minum apapun, kecuali air, selama beberapa jam sebelum tes. Berpuasa sebelum tes darah juga tidak boleh mengunyah permen karet, merokok, dan berolahraga. Berpuasa sebelum melakukan tes darah bertujuan untuk memastikan bahwa hasil pemeriksaan tidak terpengaruh oleh konsumsi makanan terakhir. Dengan begitu, dokter dapat melakukan analisis dengan lebih akurat.

Dalam penelitian ini, beberapa model statistik untuk menentukan apakah seseorang memiliki penyakit jantung atau tidak. Metode seperti *logistic regression*, K-Nearest Neighbors (KNN), dan *decision tree* dipilih karena kemampuannya menangani variabel kategorik dan kontinu.

2. **Analisis Data Eksploratif**

Heart dataset diperoleh dari Kaggle. Data ini dikumpulkan dengan menggabungkan lima data berbeda yang sudah tersedia. Lima data tersebut berasal dari: Cleveland, Switzerland, Hungary, Long Beach VA, dan Statlog (*Heart*) Data Set. Data ini berisi

1190 data dengan 629 sampel penyakit jantung dan 561 sampel non-penyakit jantung (normal).

Data ini memiliki 11 variabel prediktor yang digunakan untuk menentukan kondisi jantung. Variabel responsnya berupa kategori biner. Variabel-variabel dari data ini dijelaskan lebih rinci dalam Tabel 1.

Tabel 1. Variabel pada *Heart Dataset*

No.	**Variabel**	**Kode**	**Unit**	**Tipe Data**
1	age	Age	tahun	numerik
2	sex	Sex	1,0	biner
3	chest pain type	chest pain type	1, 2, 3, 4	kategorik
4	resting blood pressure	resting bp s	mm Hg	numerik
5	serum cholesterol	cholesterol	mg/dl	numerik
6	fasting blood sugar	fasting blood sugar	1,0 > 120 mg/dl	biner
7	resting electrocardiogram results	resting ecg	0, 1, 2	kategorik
8	maximum heart rate achieved	max heart rate	71 - 202	numerik
9	exercise induced angina	exercise angina	0,1	biner
10	oldpeak = ST	oldpeak	depression	numerik
11	the slope of the peak exercise ST segmen	ST slope	1, 2, 3	kategorik
12	class	target	0,1	biner

Penjelasan lebih lanjut, sebagai berikut :

Tabel 2. Deskripsi Variabel Biner dan Kategorik

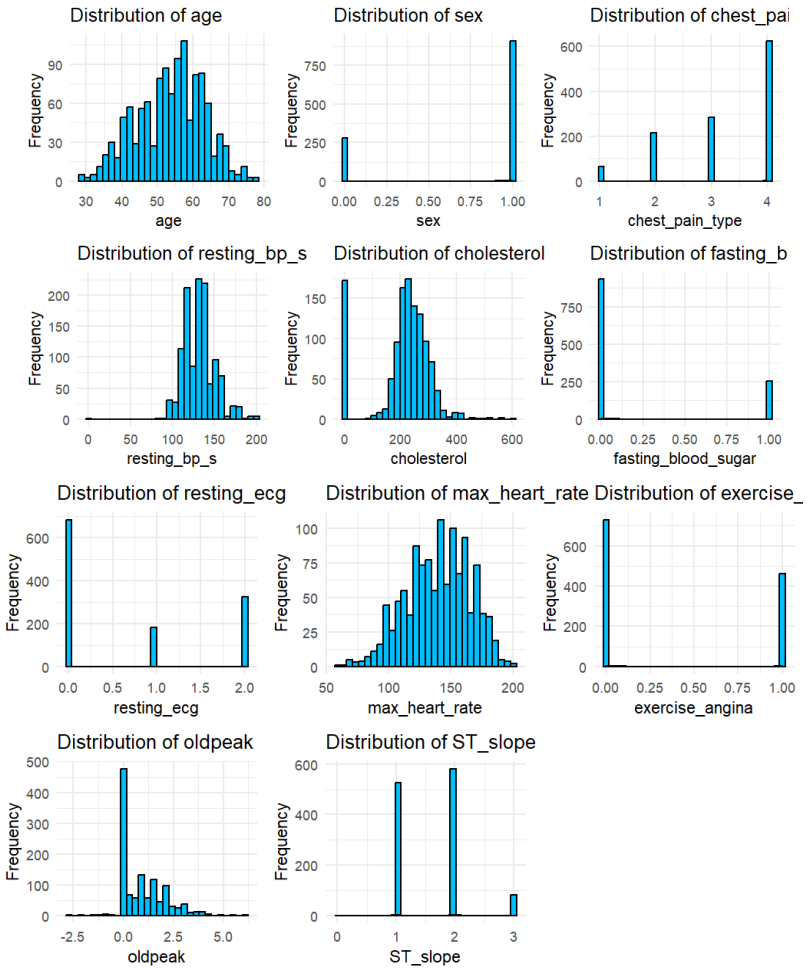
Sex	1 = Laki-laki, 0 = Perempuan;
Chest Pain Type	- Kode 1: typical angina - Kode 2: atypical angina - Kode 3: non-anginal pain - Kode 4: asymptomatic
Fating Blood Sugar	(gula darah > 120 mg/dl) 1 = benar; 0 = salah
Resting Electrocardiogram results	- Kode 0: normal - Kode 1: having ST-T wave abnomarlity (T wave inversions and/or ST elevation or depression of > 0.05 mV) - Kode 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
Exercise induced angina	1 = benar; 0 = salah
The slope of the peak exercise ST segment	- Kode 1: upsloping - Kode 2: flat - Kode 3: downsloping
class	1 = penyakit jantung, 0 = Normal

Statistik deskriptif dari masing-masing variabel prediktor, yaitu

Tabel 3. Statistik Deskriptif

Variabel	**Mean**	**StDev**	**Min**	**Max**
age	53.7202	9.3582	28.0	77.0
chest pain type	3.2328	0.9355	1.0	4.0
cholesterol	210.3639	101.4205	0.0	603.0
exercise angina	0.3874	0.4874	0.0	1.0
fasting blood sugar	0.2134	0.4099	0.0	1.0
max heart rate	139.7328	25.5176	60.0	202.0
oldpeak	0.9228	1.0863	-2.6	6.2
resting bp s (blood pressure)	132.1538	18.3688	0.0	200.0
resting ecg	0.6983	0.8704	0.0	2.0
sex	0.7639	0.4249	0.0	1.0
ST slope	1.6244	0.6105	0.0	3.0

Berikut ini, grafik distribusi dari semua variabel, yaitu



GAMBAR 1. Grafik Distribusi

Berdasarkan statistik deskriptif dan grafik distribusi di atas, rata-rata dari prediktor *age* (usia) adalah 53,7202, yang berarti rata-rata usia responden dalam dataset adalah 53,7202 tahun. Standar deviasi dari prediktor ini adalah 9,3582, yang menunjukkan bahwa distribusi usia responden cukup menyebar. Nilai minimum dan maksimum dari prediktor ini masing-masing adalah 28 tahun dan 77 tahun. Statistik deskriptif lainnya untuk setiap prediktor ditampilkan di Tabel 3.

Melihat grafik distribusi yang ditampilkan di atas, terdapat beberapa data yang tidak masuk akal, yaitu terdapat 172 data yang memiliki nilai *cholesterol* sebesar 0, ada 1 data yang memiliki nilai tekanan darah istirahat (*bp s*) sebesar 0, dan ada 1 data yang memiliki nilai kemiringan ST (*ST slope*) sebesar 0. Nilai-nilai yang tidak masuk

akal ini akan diubah menjadi nilai yang lebih masuk akal dengan menggunakan **Metode Imputasi K-nearest Neighbor (KNN)**. Langkah-langkahnya adalah: pertama, nilai 0 ini dinyatakan sebagai nilai yang hilang, kemudian dengan KNN, kami menetapkan jumlah tetangga terdekat yang dilambangkan dengan K. Di sini, kami memilih $K = 5$. Terakhir, jarak terpendek dari setiap observasi yang tidak mengandung nilai yang hilang dihitung. Setelah algoritma selesai, nilai yang hilang kemudian diisi dengan nilai yang lebih deskriptif.

Untuk menguji akurasi model, dataset dibagi menjadi data pelatihan (*training data*) dan data uji (*test data*) dengan rasio 7 : 3.

```
> intrain <- sample(nrow(df), nrow(df)*0.7)
> heart_train <- df[intrain,]
> heart_test <- df[-intrain,]
> dim(heart_test)
[1] 357  12
> dim(heart_train)
[1] 833  12
```

Jadi, terdapat 357 data uji dan 833 data *training* (pelatihan).

3. Regresi Logistik

3.1. **Pengertian.** Regresi linear tidak bisa digunakan untuk memodelkan klasifikasi biner (atau klasifikasi pada umumnya) karena sifat tak terbatas dari fungsi linear. Oleh karena itu, diperlukan cara lain untuk menjelaskan hubungan antara prediktor dan variabel respons.

Untuk klasifikasi biner secara spesifik, salah satu cara melakukannya adalah dengan mengaitkan X ke fungsi probabilitas $P(X) = Pr(Y = 1|X)$ (di sini, menggunakan pengkodean biner 0/1 untuk respons). Perhatikan bahwa dengan mendefinisikan $P(X)$ seperti ini, $P(X)$ hanya bisa memberikan nilai antara 0 dan 1. Ada banyak fungsi seperti itu, salah satunya disebut fungsi logistik, yang diberikan sebagai berikut [2].

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}. \tag{3.1}$$

Mirip dengan regresi linear, persamaan (3.1) dapat diperluas ke bentuk dengan banyak prediktor, seperti berikut

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}. \tag{3.2}$$

Estimasi parameter dari persamaan (3.2) menggunakan metode estimasi likelihood maksimum (MLE).

3.2. **Membuat Model Logistik.** Dengan menggunakan dataset pelatihan, regresi logistik untuk menganalisis hubungan antara setiap prediktor dan respon, serta signifikansinya sebagai berikut

```
    Call:
glm(formula = target ~ ., family = binomial(link = "logit"),
    data = heart_train)

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -5.7651414  1.6046671  -3.593 0.000327 ***
age                            0.0293033  0.0134102   2.185 0.028878 *
sexMale                        1.8633684  0.2896186   6.434 1.24e-10 ***
chest_pain_typeAtypical Angina 0.0752636  0.5283270   0.142 0.886719
chest_pain_typeNon Angina Pain 0.3711427  0.4841849   0.767 0.443360
chest_pain_typeAsymtomatic     1.9424215  0.4745168   4.093 4.25e-05 ***
resting_bp_s                  -0.0003377  0.0062981  -0.054 0.957243
cholesterol                    0.0026771  0.0021440   1.249 0.211792
fasting_blood_sugarTrue        1.0551581  0.2869399   3.677 0.000236 ***
resting_ecgHaving ST-T         0.0863443  0.3272725   0.264 0.791911
resting_ecgShowing Probable    0.1992474  0.2573135   0.774 0.438731
max_heart_rate                -0.0062964  0.0051371  -1.226 0.220330
exercise_anginaYes             0.9129386  0.2430423   3.756 0.000172 ***
oldpeak                        0.2758857  0.1189571   2.319 0.020384 *
ST_slopeFlat                   2.2500142  0.2558983   8.793  < 2e-16 ***
ST_slopeDownsloping            1.1123219  0.4749760   2.342 0.019188 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1152.35  on 832  degrees of freedom
Residual deviance:  571.91  on 817  degrees of freedom
AIC: 603.91

Number of Fisher Scoring iterations: 5
```

Model ini disebut sebagai Model 1. Cek kesesuaian Model 1, yaitu dengan membandingkan deviasi residu dengan distribusi *chi*-kuadrat (χ^2) yang sesuai untuk model ini. Hasilnya adalah sebagai berikut.

```
> with(fit1, cbind(res.deviance = deviance, df = df.residual, pvalue =
pchisq(deviance, df.residual, lower.tail=FALSE)))
         res.deviance  df     pvalue
   [1,]      571.912   817      1
```

Berdasarkan output R di atas, menunjukkan bahwa nilai *p.value* adalah 1 yang lebih besar dari 0,05. Oleh karena itu, tidak menolak H_0, yaitu model penuh (Model 1) cocok digunakan dalam penelitian ini. Namun, ada beberapa prediktor yang tidak signifikan karena nilai *p.value* lebih besar dari 0,05, sehingga model akan disederhanakan dengan

mengeliminasi beberapa prediktor menggunakan teknik *backward elimination*. Teknik ini dilakukan dengan menghapus prediktor dengan nilai *p.value* tertinggi dari model saat ini dan kemudian menjalankan regresi yang sama. Langkah-langkah ini diulang secara iteratif hingga mendapatkan model yang paling sesuai dan sederhana. Dalam Model 1, variabel "resting bps" dihilangkan terlebih dahulu, yang menghasilkan Model 2 berikut ini.

```
Call:
glm(formula = target ~ age + sex + chest_pain_type + cholesterol +
    fasting_blood_sugar + resting_ecg + max_heart_rate + exercise_angina +
    oldpeak + ST_slope, family = binomial(link = "logit"), data = heart_train)

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -5.803331   1.438298  -4.035 5.46e-05 ***
age                            0.029197   0.013261   2.202 0.027687 *
sexMale                        1.864423   0.289033   6.451 1.11e-10 ***
chest_pain_typeAtypical Angina 0.076491   0.527810   0.145 0.884773
chest_pain_typeNon Angina Pain 0.373082   0.482781   0.773 0.439655
chest_pain_typeAsymtomatic     1.944740   0.472516   4.116 3.86e-05 ***
cholesterol                    0.002665   0.002132   1.250 0.211247
fasting_blood_sugarTrue        1.055159   0.286978   3.677 0.000236 ***
resting_ecgHaving ST-T         0.085838   0.327177   0.262 0.793044
resting_ecgShowing Probable    0.199983   0.256945   0.778 0.436386
max_heart_rate                -0.006294   0.005137  -1.225 0.220505
exercise_anginaYes             0.910998   0.240318   3.791 0.000150 ***
oldpeak                        0.275205   0.118286   2.327 0.019986 *
ST_slopeFlat                   2.250710   0.255609   8.805  < 2e-16 ***
ST_slopeDownsloping            1.114343   0.473373   2.354 0.018570 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1152.35  on 832  degrees of freedom
Residual deviance:  571.91  on 818  degrees of freedom
AIC: 601.91

Number of Fisher Scoring iterations: 5
```

Kemudian, memeriksa deviasi residu dari Model 2 dengan membandingkannya dengan distribusi *chi*-kuadrat (χ^2) dari model ini. Hasilnya adalah sebagai berikut.

```
> with(fit2, cbind(res.deviance = deviance, df = df.residual, pvalue =
pchisq(deviance, df.residual, lower.tail=FALSE)))
      res.deviance  df    pvalue
[1,]     571.9149    818      1
```

Berdasarkan output R di atas, menunjukkan bahwa nilai *p.value* dari uji *chi*-kuadrat untuk deviasi residu adalah 1, yang lebih besar dari 0,05. Oleh karena itu, tidak menolak H_0, yaitu Model 2 dapat digunakan dalam penelitian ini. Selanjutnya, membandingkan nilai AIC dari Model 1 dan Model 2. *Akaike Information Criterion* (AIC) adalah estimator dari kesalahan prediksi dari model statistik untuk data yang diberikan. Semakin kecil nilai AIC, semakin baik model tersebut. Berdasarkan hasil di atas, nilai AIC dari model penuh = 603,91 dan nilai AIC dari Model 2 = 601,91, yang menunjukkan bahwa Model 2 dapat digunakan sebagai model untuk penelitian ini. Kemudian, proses tersebut diulangi, yaitu menghilangkan variabel prediktor dengan nilai *p.value* tertinggi pada Model 2, yaitu "resting ecg". Dengan demikian, diperoleh Model 3 sebagai berikut:

```
    Call:
glm(formula = target ~ age + sex + chest_pain_type + cholesterol +
    fasting_blood_sugar + max_heart_rate + exercise_angina +
    oldpeak + ST_slope, family = binomial(link = "logit"), data = heart_train)

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -5.951046   1.415650  -4.204 2.63e-05 ***
age                            0.031198   0.013024   2.395 0.016600 *
sexMale                        1.870151   0.288821   6.475 9.47e-11 ***
chest_pain_typeAtypical Angina 0.039461   0.522303   0.076 0.939776
chest_pain_typeNon Angina Pain 0.340398   0.478760   0.711 0.477085
chest_pain_typeAsymtomatic     1.927155   0.469476   4.105 4.04e-05 ***
cholesterol                    0.002865   0.002116   1.354 0.175888
fasting_blood_sugarTrue        1.049585   0.285637   3.675 0.000238 ***
max_heart_rate                -0.005732   0.004969  -1.154 0.248629
exercise_anginaYes             0.902834   0.239763   3.766 0.000166 ***
oldpeak                        0.277724   0.118147   2.351 0.018740 *
ST_slopeFlat                   2.248394   0.255546   8.798  < 2e-16 ***
ST_slopeDownsloping            1.126719   0.473388   2.380 0.017307 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1152.35  on 832  degrees of freedom
Residual deviance:  572.53  on 820  degrees of freedom
AIC: 598.53

Number of Fisher Scoring iterations: 5
```

Kemudian, cek kesesuaian model dengan memeriksa deviasi residu dari Model 3 dan membandingkannya dengan distribusi *chi*-kuadrat (χ^2) dari model ini. Hasilnya adalah sebagai berikut.

```
> with(fit3, cbind(res.deviance = deviance, df = df.residual, pvalue =
pchisq(deviance, df.residual, lower.tail=FALSE)))
      res.deviance    df     pvalue
[1,]     572.5292      820      1
```

Berdasarkan hasil dari output R di atas, nilai *p.value* dari uji *chi*-kuadrat untuk deviasi residu adalah 1, yang lebih besar dari 0,05. Oleh karena itu, tidak menolak H_0, yaitu Model 3 cocok digunakan dalam penelitian ini. Selanjutnya, membandingkan nilai AIC antara Model 2 dan Model 3. Berdasarkan hasil di atas, nilai AIC untuk Model 2 adalah 601,91 dan nilai AIC untuk Model 3 adalah 598,53, yang menunjukkan bahwa Model 3 lebih baik digunakan sebagai model untuk penelitian ini. Kemudian menghapus variabel prediktor yang memiliki nilai *p.value* tertinggi pada Model 3, yaitu "max heart rate", sehingga diperoleh Model 4 seperti berikut:

```
    Call:
glm(formula = target ~ age + sex + chest_pain_type + cholesterol +
    fasting_blood_sugar + exercise_angina + oldpeak + ST_slope,
    family = binomial(link = "logit"), data = heart_train)

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -7.100681   1.027212  -6.913 4.76e-12 ***
age                            0.036185   0.012350   2.930 0.003390 **
sexMale                        1.901495   0.288499   6.591 4.37e-11 ***
chest_pain_typeAtypical Angina 0.078993   0.525677   0.150 0.880552
chest_pain_typeNon Angina Pain 0.369192   0.482933   0.764 0.444582
chest_pain_typeAsymtomatic     2.007925   0.470023   4.272 1.94e-05 ***
cholesterol                    0.002622   0.002095   1.252 0.210652
fasting_blood_sugarTrue        1.047755   0.285441   3.671 0.000242 ***
exercise_anginaYes             0.942573   0.236747   3.981 6.85e-05 ***
oldpeak                        0.260812   0.116404   2.241 0.025053 *
ST_slopeFlat                   2.333250   0.245816   9.492  < 2e-16 ***
ST_slopeDownsloping            1.197456   0.466174   2.569 0.010208 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1152.35  on 832  degrees of freedom
Residual deviance:  573.86  on 821  degrees of freedom
AIC: 597.86

Number of Fisher Scoring iterations: 5
```

Kemudian, cek kesesuaian model dengan memeriksa deviasi residu dari Model 4 dan membandingkannya dengan distribusi *chi*-kuadrat (χ^2) dari model ini. Hasilnya adalah sebagai berikut.

```
> with(fit4, cbind(res.deviance = deviance, df = df.residual, pvalue =
pchisq(deviance, df.residual, lower.tail=FALSE)))
        res.deviance  df    pvalue
[1,]     573.8609    821      1
```

Berdasarkan output R di atas, menunjukkan bahwa nilai *p.value* dari uji *chi*-kuadrat untuk deviasi residu adalah 1, yang lebih besar dari 0,05. Dengan demikian, tidak menolak H_0, yaitu Model 4 cocok digunakan dalam penelitian ini. Kemudian, membandingkan nilai AIC dari Model 3 dan Model 4. Berdasarkan hasil di atas, nilai AIC Model 3 = 598,53 dan nilai AIC Model 4 = 597,86, yang menunjukkan bahwa Model 4 lebih baik digunakan sebagai model dalam penelitian ini. Selanjutnya, menghilangkan variabel prediktor yang memiliki nilai *p.value* tertinggi pada Model 4, yaitu "kolesterol". Dengan demikian, diperoleh Model 5 sebagai berikut:

```
    Call:
glm(formula = target ~ age + sex + chest_pain_type + fasting_blood_sugar +
    exercise_angina + oldpeak + ST_slope, family = binomial(link = "logit"),
    data = heart_train)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -6.43219    0.86324  -7.451 9.25e-14 ***
age                            0.03624    0.01233   2.938  0.00330 **
sexMale                        1.81171    0.27679   6.545 5.93e-11 ***
chest_pain_typeAtypical Angina 0.10730    0.52283   0.205  0.83739
chest_pain_typeNon Angina Pain 0.36733    0.48050   0.764  0.44459
chest_pain_typeAsymtomatic     2.04025    0.46695   4.369 1.25e-05 ***
fasting_blood_sugarTrue        1.07579    0.28371   3.792  0.00015 ***
exercise_anginaYes             0.97144    0.23533   4.128 3.66e-05 ***
oldpeak                        0.27421    0.11546   2.375  0.01755 *
ST_slopeFlat                   2.32938    0.24545   9.490  < 2e-16 ***
ST_slopeDownsloping            1.16596    0.46263   2.520  0.01173 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1152.35  on 832  degrees of freedom
Residual deviance:  575.45  on 822  degrees of freedom
AIC: 597.45
```

Kemudian, cek kesesuaian model dengan memeriksa deviasi residu dari Model 5 dan membandingkannya dengan distribusi *chi*-kuadrat (χ^2) dari model ini. Hasilnya adalah sebagai berikut.

```
> with(fit5, cbind(res.deviance = deviance, df = df.residual, pvalue =
pchisq(deviance, df.residual, lower.tail=FALSE)))
```

```
        res.deviance   df   pvalue
[1,]     575.4466     822       1
```

Melihat hasil output R di atas, nilai *p.value* dari uji c*chi*-kuadrat adalah 1, yang lebih besar dari 0,05. Oleh karena itu, tidak menolak H_0, yaitu Model 5 cocok untuk digunakan dalam penelitian ini. Kemudian, membandingkan nilai AIC antara Model 4 dan Model 5. Berdasarkan hasil di atas, nilai AIC Model 4 = 597.86 dan nilai AIC Model 5 = 575.45, yang menunjukkan bahwa Model 5 adalah model yang tepat untuk digunakan dalam penelitian ini. Karena semua prediktor pada Model 5 signifikan dan Model 5 memiliki nilai AIC terkecil di antara semua model yang telah dibuat, maka Model 5 dipilih sebagai model utama dalam penelitian ini.

Model logistik dari *heart dataset* adalah

$$p(X) = \frac{e^{-6.43+0.04X_1+1.81X_2(1)+0.11X_3(2)+0.37X_3(3)+2.04X_3(4)+1.08X_4(1)+0.97X_5(1)+0.27X_6+2.33X_7(2)+1.17X_7(3)}}{1+e^{-6.43+0.04X_1+1.81X_2(1)+0.11X_3(2)+0.37X_3(3)+2.04X_3(4)+1.08X_4(1)+0.97X_5(1)+0.27X_6+2.33X_7(2)+1.17X_7(3)}}$$

di mana

$$\begin{aligned}
X_1 &= \text{Age} \\
X_2(1) &= \text{Sex Male} \\
X_3(2) &= \text{Chest Pain Type Atypical Angina} \\
X_3(3) &= \text{Chest Pain Type Non Angina} \\
X_4(4) &= \text{Chest Pain Type Asymtomatic} \\
X_5(1) &= \text{Fasting Blood Sugar True} \\
X_6 &= \text{Oldpeak} \\
X_7(2) &= \text{ST Slope Flat} \\
X_7(3) &= \text{ST Slope Downsloping}
\end{aligned}$$

Kemudian menguji akurasi model dengan menggunakan data uji. Berikut ini adalah perbandingan antara hasil prediksi dan data aktual.

```
> heart_test$pred_heart <- factor(ifelse(heart_test$prob_heart > 0.5,
"Heart Disease", "Normal"))heart_test[1:10, c("pred_heart", "target")]
```

```
pred_heart      target
Heart Disease   Heart Disease
Normal          Normal
Normal          Normal
Normal          Normal
Heart Disease   Heart Disease
Normal          Normal
Heart Disease   Heart Disease
Normal          Normal
Normal          Normal
```

Kemudian melihat *confusion matrix* dari model logistik. *Confusion matrix* adalah tabel yang digunakan untuk mengevaluasi kinerja model klasifikasi. Berikut ini adalah *confusion matrix*:

```
> log_conf <- confusionMatrix(heart_test$pred_heart, heart_test$target,
positive = "Heart Disease")
    Confusion Matrix and Statistics
                          Reference
Prediction         Normal     Heart Disease
  Normal            140             40
  Heart Disease      27            150
```

Di bawah ini nilai presisi, *recall*, spesifisitas, dan akurasi dari model logistik.

```
Accuracy     Recall    Specificity   Precision
0.8123249   0.7894737   0.8383234    0.8474576
```

Interpretasinya, sebagai berikut:

- **Akurasi** sebesar 81,23% berarti bahwa 81,23% dari 357 sampel data terklasifikasi **benar** menggunakan Model 5.
- **Recall** sebesar 78,95% berarti bahwa dari 190 data uji yang diklasifikasikan **memiliki** penyakit jantung **dari set data uji**, hanya 78,95% yang benar-benar terklasifikasi memiliki penyakit jantung oleh Model 5 (proporsi *true positive*).
- **Spesifisitas** sebesar 83,83% berarti bahwa dari 167 data uji yang diklasifikasikan **tidak memiliki** penyakit jantung **dari set data uji**, hanya 83,83% yang benar-benar terklasifikasi tidak memiliki penyakit jantung oleh Model 5 (proporsi *true negative*).
- **Presisi** sebesar 84,74% berarti bahwa dari 177 data uji yang diklasifikasikan memiliki penyakit jantung **oleh Model kami**, hanya 84,74% yang sebenarnya memiliki penyakit jantung.

4. *K-NEAREST NEIGHBORS*

4.1. **Pengertian.** Seperti yang telah disinggung di bagian sebelumnya, dalam pengaturan klasifikasi biner (atau klasifikasi secara umum), tujuannya adalah untuk menjelaskan dan memperkirakan distribusi kondisional dari Y diberikan X. Lebih spesifiknya, ingin memperkirakan nilai $Pr(Y = 1|X)$ untuk semua nilai X.

Diberikan sebuah dataset dengan n prediktor dan 1 respon, setiap titik data dalam dataset dapat dipandang sebagai sebuah titik dalam bidang $\mathbb{R}^{n+1}$. Dengan memikirkan hal ini secara geometris, dapat menerapkan konsep "jarak" antara setiap titik data. Dalam bidang $\mathbb{R}^{n+1}$, maka menggunakan jarak *Euclidean* secara *default*.

Diberikan sebuah bilangan bulat positif K dan sebuah titik data x_0, klasifikator KNN pertama-tama mengidentifikasi K titik yang ada dalam data pelatihan yang paling dekat (dalam hal jarak *Euclidean*) dengan x_0, yang disebut sebagai $\mathcal{N}_0$. Untuk memperkirakan probabilitas kondisional untuk kelas 1, cukup menghitung jumlah titik data dalam $\mathcal{N}_0$ yang nilai responsnya sama dengan 1. Secara matematis, ini ditulis

sebagai berikut:

$$Pr(Y = 1|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = 1) \tag{4.1}$$

di mana $I(y = 1)$ adalah fungsi indikator dari $\{y = 1\}$, yaitu sebuah fungsi yang memberikan nilai 1 jika $y = 1$, dan memberikan nilai 0 jika $y \neq 1$ [1].

Pemilihan K sangat penting dalam algoritma ini. Nilai K yang kecil mengarah pada bias yang rendah tetapi varians yang tinggi, sementara nilai K yang lebih besar mengarah pada varians yang rendah tetapi bias yang tinggi. Hal lain yang perlu dicatat adalah bahwa, dalam pengaturan klasifikasi biner, lebih baik memilih K sebagai bilangan ganjil untuk menghindari hasil yang *ties*.

4.2. **Klasifikasi dengan *K-Nearest Neighbors*.** Seperti yang telah disebutkan, lebih baik memilih K sebagai angka ganjil. Oleh karena itu, kami akan mencoba dan membandingkan beberapa model KNN dengan $K = 1$, $K = 3$, $K = 5$, dan $K = 7$. Seperti yang diberikan dalam Model 5 dari bagian regresi logistik, variabel prediktor yang digunakan dalam penelitian ini adalah 'umur', 'jenis kelamin', 'jenis nyeri dada', 'gula darah puasa', 'angina akibat olahraga', 'oldpeak', dan 'kemiringan ST'. Kami juga menggunakan dataset pelatihan dan pengujian yang sama seperti pada model regresi logistik.

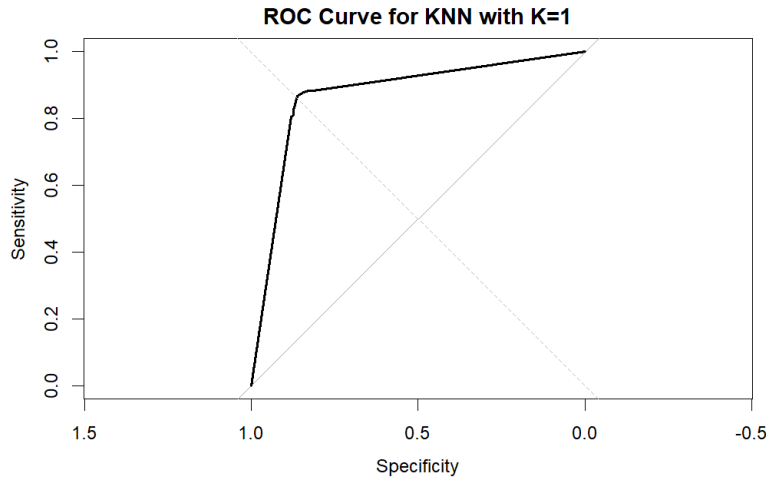
```
> print(tab1<-table(knn.k1,test.Y))
                          test.Y
knn.k1          Heart Disease Normal
  Heart Disease           160     22
  Normal                   30    145

> print(tab3<-table(knn.k3,test.Y))
                          test.Y
knn.k3          Heart Disease Normal
  Heart Disease           151     44
  Normal                   39    123

> print(tab5<-table(knn.k5,test.Y))
                          test.Y
knn.k5          Heart Disease Normal
  Heart Disease           153     43
  Normal                   37    124

> print(tab7<-table(knn.k7,test.Y))
                          test.Y
knn.k7          Heart Disease Normal
  Heart Disease           149     48
  Normal                   41    119
```

Berikut ini *overall error rate* dari setiap model KNN.

```
Overall error rate for k = 1: 0.1456583
Overall error rate for k = 3: 0.232493
Overall error rate for k = 5: 0.2240896
Overall error rate for k = 7: 0.2492997
```

Berdasarkan hasil yang ditunjukkan di atas, dipilih nilai K yang memberikan tingkat kesalahan keseluruhan terkecil, yaitu $K = 1$. Sekarang, memeriksa apakah model KNN ini layak digunakan untuk penelitian ini dengan mengevaluasinya menggunakan kurva *Receiver Operating Characteristic* (ROC). Kurva ROC menunjukkan hubungan antara *True Positive Rate* (TPR) dan *False Positive Rate* (FPR) yang sering digunakan untuk menilai kinerja pengklasifikasi biner.

Area di bawah kurva ROC (AUC) digunakan untuk mengukur kinerja keseluruhan model. Nilai AUC yang lebih tinggi menunjukkan kinerja model yang lebih baik. Berikut ini menunjukkan area di bawah kurva ROC untuk model KNN.



Gambar 2. ROC Curve for KNN with K = 1

```
[1] "AUC: 0.8725"
```

Ini menunjukkan bahwa area di bawah kurva ROC untuk model ini adalah 0,8725, yang cukup mendekati 1. Ini berarti bahwa model dapat mengklasifikasikan individu yang sehat ("Normal") dan individu dengan penyakit jantung ("Heart Disease") dengan cukup baik, yaitu, model KNN dengan $K = 1$ ini layak digunakan dalam penelitian ini.

Selanjutnya, menguji akurasi model menggunakan data uji. Berikut adalah beberapa contoh perbandingan antara prediksi dan data aktual.

```
> set.seed(1)
> knn.k1 <- knn(train.X, test.X, train.Y, k=1)
> cbind(data.frame(knn.k1), test.Y)[1:10,]
    knn.k1           target
```

```
1 Normal          Heart Disease
2 Normal          Normal
3 Normal          Normal
4 Normal          Normal
5 Heart Disease   Heart Disease
6 Normal          Normal
7 Heart Disease   Heart Disease
8 Normal          Normal
9 Normal          Normal
10 Heart Disease  Heart Disease
```

Di bawah ini nilai presisi, recall, spesifisitas, dan presisi model KNN.

```
Akurasi Recall Spesifisitas Presisi
0.8543417 0.8421053 0.8682635 0.8791209
```

Berdasarkan output di atas, diperoleh

- Akurasi model ini adalah 85,43%.
- Recall model ini adalah 84,21%.
- Spesifisitas model ini adalah 86,83%.
- Presisi model ini adalah 87,91%.

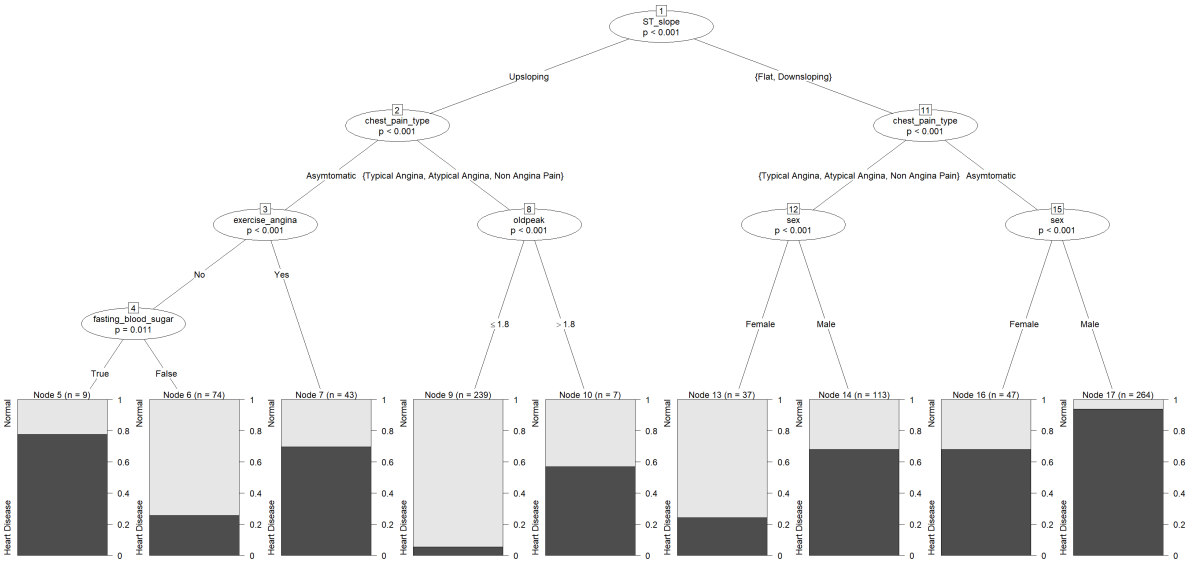
5. Pohon Keputusan (*Decision Tree*)

5.1. **Pengertian.** Pohon keputusan (*Decision Tree*) adalah struktur pohon yang terdiri dari simpul yang mewakili keputusan dan cabang-cabang yang mewakili konsekuensi dari sebuah keputusan. Cara untuk membangun pohon keputusan adalah dengan membagi data menjadi kelompok-kelompok kecil berdasarkan atribut-atribut yang ada dalam data tersebut. Pembagian kelompok ini dilakukan secara berulang hingga semua elemen data yang berasal dari kelas yang sama dapat dimasukkan dalam satu kelompok. Ada beberapa algoritma untuk membangun pohon keputusan, termasuk menetapkan atribut sebagai akar pohon, membangun cabang-cabang untuk setiap nilai, membagi kasus dalam setiap cabang, memverifikasi, dan mengulang proses tersebut dengan tujuan agar setiap cabang hanya memiliki kelas yang serupa. Metode ini memiliki beberapa keuntungan, termasuk kemampuannya untuk menginterpretasi data kategorikal dan numerik serta kecenderungannya yang tidak memerlukan normalisasi atau standardisasi data. Namun, terdapat juga beberapa kekurangan pada metode ini, seperti kualitas keputusan yang dihasilkan sangat bergantung pada desain pohon, terjadinya tumpang tindih terutama ketika terdapat banyak kelas dan kriteria yang digunakan, serta kesulitan dalam merancang pohon keputusan yang optimal [6].

5.2. **Membuat Klasifikasi dengan Pohon Keputusan.** Dengan menggunakan dataset pelatihan, diperoleh model pohon keputusan sebagai berikut.

```
heart.tree <- ctree(target~., data=heart_train) plot(heart.tree)
```

Pohon keputusan di atas terdiri dari beberapa bagian. Simpul di puncak gambar adalah ST Slope, yang juga disebut sebagai simpul akar (*root node*). Simpul akar ini dibagi menjadi beberapa simpul lainnya. Beberapa dari simpul-simpul ini kemudian



Gambar 3. Pohon Keputusan

akan dibagi lagi. Proses ini terus berlangsung hingga tidak ada simpul lagi yang dapat dibagi. Pada pohon keputusan di atas, simpul 5, 6, 7, 9, 10, 13, 14, 16, dan 17 adalah simpul terminal yang memberikan prediksi akhir apakah individu tersebut memiliki penyakit jantung ("*Heart Disease*") atau tidak ("Normal").

- **Root Node (Node 1).**
  Prediktor "ST slope" memiliki nilai $p < 0.01$. Ini menunjukkan bahwa prediktor "ST slope" sangat signifikan terhadap hasil. Prediktor ini adalah faktor utama yang menentukan percabangan.
  Pembagian berdasarkan ST slope:
  Jika prediktor ini adalah *Upsloping*, maka akan menuju ke Node 2.
  Jika prediktor ini adalah *Flat* atau *Downsloping*, maka akan menuju ke Node 11.
- **Node 2**
  Prediktor "Chest pain type" memiliki nilai $p < 0.001$.
  Pembagian:
  Jika jenis nyeri dada adalah Asimtomatik, maka akan menuju ke Node 3.
  Jika jenis nyeri dada adalah Angina Tipikal, Angina Atypikal, atau Nyeri Non-Angina, maka akan menuju ke Node 8.
- **Node 3**
  Prediktor "exercise angina" memiliki nilai $p < 0.001$.
  Pembagian:

Jika exercise angina adalah Tidak, maka akan menuju ke Node 4.
Jika exercise angina adalah Ya, maka akan menuju ke Node 7.

- **Node 4**
  Prediktor "fasting blood sugar" memiliki nilai p = 0.011.
  Pembagian:
  Jika fasting blood sugar adalah Benar, maka akan menuju ke Node 5.
  Jika fasting blood sugar adalah Salah, maka akan menuju ke Node 6.
- **Node 5**
  Jumlah data dengan tipe ST slope Upsloping, jenis nyeri dada Asimtomatik, tanpa exercise angina, dan fasting blood sugar lebih dari 120 mg/dl adalah 9 sampel. Dari 9 sampel ini, 22.2% memiliki jantung normal dan 77.8% memiliki penyakit jantung.
- **Node 6**
  Jumlah data dengan tipe ST slope Upsloping, jenis nyeri dada Asimtomatik, tanpa exercise angina, dan fasting blood sugar kurang dari 120 mg/dl adalah 74 sampel. Dari 74 sampel ini, 74.3% memiliki jantung normal dan 25.7% memiliki penyakit jantung.
- **Node 7**
  Jumlah data dengan tipe ST slope Upsloping, jenis nyeri dada Asimtomatik, dan exercise angina Ya adalah 43 sampel. Dari 43 sampel ini, 30.2% memiliki jantung normal dan 69.8% memiliki penyakit jantung.
- **Node 9**
  Jumlah data dengan tipe ST slope Upsloping, angina tipikal, angina atypical, atau nyeri dada non-angina, dan Oldpeak kurang dari 1.8 adalah 239 sampel. Dari 239 sampel ini, 94.6% memiliki jantung normal dan 5.4% memiliki penyakit jantung.
- **Node 10**
  Jumlah data dengan tipe ST slope Upsloping, angina tipikal, angina atypical, atau nyeri dada non-angina, dan Oldpeak lebih dari 1.8 adalah 7 sampel. Dari 7 sampel ini, 42.9% memiliki jantung normal dan 51.7% memiliki penyakit jantung.
- **Node 13**
  Jumlah data dengan tipe ST slope Flat atau Downsloping, angina tipikal, angina atypical, atau nyeri dada non-angina, dan jenis kelamin Perempuan adalah 37 sampel. Dari 37 sampel ini, 75.7% memiliki jantung normal dan 24.3% memiliki penyakit jantung.
- **Node 14**
  Jumlah data dengan tipe ST slope Flat atau Downsloping, angina tipikal, angina atypical, atau nyeri dada non-angina, dan jenis kelamin Laki-laki adalah 113 sampel. Dari 113 sampel ini, 31.9% memiliki jantung normal dan 68.1% memiliki penyakit jantung.
- **Node 15**
  Prediktor "Sex" memiliki nilai p < 0.001.
  Pembagian:

Jika Perempuan, maka akan menuju ke Node 16.
Jika Laki-laki, maka akan menuju ke Node 17.

- **Node 16**
  Jumlah data dengan tipe ST slope Flat atau Downsloping, jenis nyeri dada Asimtomatik, dan jenis kelamin Perempuan adalah 47 sampel. Dari 47 sampel ini, 31.9% memiliki jantung normal dan 68.1% memiliki penyakit jantung.
- **Node 17**
  Jumlah data dengan tipe ST slope Flat atau Downsloping, jenis nyeri dada Asimtomatik, dan jenis kelamin Laki-laki adalah 264 sampel. Dari 264 sampel ini, 6.1% memiliki jantung normal dan 93.9% memiliki penyakit jantung.

Kemudian menguji akurasi model menggunakan data uji. Berikut adalah beberapa contoh dari prediksi vs data aktual.

```
    prediksi.tree     test.Y
1 Heart Disease    Heart Disease
2 Normal           Normal
3 Normal           Normal
4 Normal           Normal
5 Heart Disease    Heart Disease
6 Normal           Normal
7 Heart Disease    Heart Disease
8 Normal           Normal
9 Normal           Normal
10 Heart Disease   Heart Disease
```

Kemudian *confusion matrix* dari model ini sebagai berikut.

```
> tree_conf <- confusionMatrix(prediksi.tree, heart_test$target,
positive = "Heart Disease")
Confusion Matrix and Statistics
                     Reference
Prediction       Normal    Heart Disease
  Normal           120            17
  Heart Disease     47           173
```

Hasil kinerja prediksi model ini menggunakan data uji diberikan sebagai berikut.

```
Accuracy     Recall    Specificity  Precision
0.8207283   0.9105263   0.7185629   0.7863636
```

Evaluasi kinerja model ini sebagai berikut.

- Akurasi model ini adalah 82.07%.
- Recall model ini adalah 91.05%.
- Spesifisiti model ini adalah 71.86%.
- Presisi model ini adalah 78.63%.

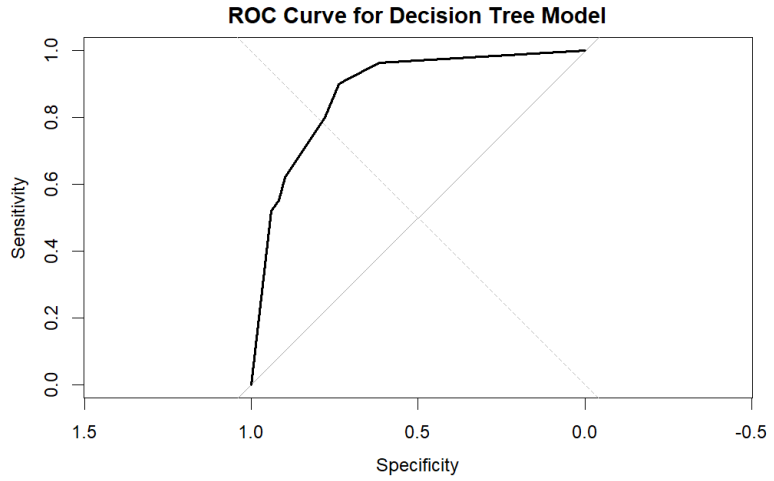
Kemudian memeriksa apakah model *decision tree* ini mengalami *overfitting* atau tidak dengan membandingkan hasil kinerja prediksinya menggunakan data pelatihan dan data

uji. Hasil kinerja prediksi model ini menggunakan data pelatihan diberikan sebagai berikut.

```
Accuracy     Recall     Specificity   Precision
0.8487395   0.9066059    0.784264     0.8240166
```

Karena hasil kinerja prediksi menggunakan dataset pelatihan vs dataset uji cukup mirip, model kami tidak menunjukkan tanda-tanda *overfitting* atau *underfitting* yang signifikan.

Kemudian cek apakah model *decision tree* ini layak digunakan untuk penelitian ini dengan menggunakan kurva *Receiver Operating Characteristic* (ROC), yang merupakan kurva yang digunakan untuk mengevaluasi kinerja model klasifikasi. Berikut menunjukkan area di bawah kurva ROC dari model *decision tree* ini.



Gambar 4. ROC Curve for Decision Tree

```
[1] "AUC: 0.8769"
```

Ini menunjukkan bahwa area di bawah kurva ROC dari model ini adalah 0.8769, yang cukup dekat dengan 1. Ini berarti bahwa model ini dapat mengklasifikasikan individu yang sehat ("Normal") dan individu dengan penyakit jantung ("*Heart Disease*") dengan cukup baik, yaitu model *decision tree* ini layak digunakan dalam penelitian ini.

6. **Membandingkan Metode Klasifikasi**

Pada bagian ini, membandingkan tiga metode klasifikasi yang berbeda, yaitu regresi logistik, *K-nearest neighbors* (KNN), dan metode *decision tree*. Evaluasi kinerja dari model-model ini sebagai berikut. Dari tabel ini, model terbaik untuk mengklasifikasikan dataset penyakit jantung ini adalah model KNN, karena memiliki nilai kinerja tertinggi dalam tiga aspek, yaitu akurasi, spesifisitas, dan presisi.

Tabel 4. Performance Measures of Four Models

No.	Model	Accuracy	Recall	Specificity	Precision
1	Logistic Regression	0.8123249	0.7894737	0.8383234	0.8474576
2	K-Nearest Neighbors	0.8543417	0.8421053	0.8682635	0.8791209
3	Decision Tree	0.8207283	0.9105263	0.7185629	0.7863636

7. PENUTUP

Dalam penelitian ini, telah disajikan studi komparatif dari tiga metode klasifikasi dan kinerja setiap metode telah diukur dalam hal akurasi, recall, spesifisitas, dan presisi. Dari hasil-hasil ini, dapat menyimpulkan bahwa model *K-Nearest Neighbors* (KNN) memiliki ukuran kinerja tertinggi dibandingkan dengan regresi logistik dan *decision tree* untuk dataset yang diberikan. Oleh karena itu, model KNN adalah model yang terbaik untuk data penyakit jantung dalam penelitian ini.

References

[1] James, G., Witten, D., Hastie, T., dan Tibshirani, R. 2017.*An Introduction to Statistical Learning with Applications in R*. New York: Springer Science.
[2] Hosmer, D.W. dan Lemeshow, S. 2000. *Applied Logistic Regression*, 2nd. Canada: John Wiley and Sons.
[3] Mayo Clinic Staff. 2023. High blood pressure dangers: Hypertension's effects on your body. Diakes pada 15 Juni 2024 dari https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/in-depth/high-blood-pressure/art-20045868.
[4] Medline Plus. 2023. https://medlineplus.gov/lab-tests/fasting-for-a-blood-test/ accessed 15 Juni 2024.
[5] Oona. 2024. Kenali Perbedaan Gejala Serangan Jantung pada Pria dan Wanita. Diakses pada 15 Juni 2024 dari https://myoona.id/blog/kesehatan/gejala-serangan-jantung/.
[6] Ramadhon, R.N., dkk. 2024. Implementasi Algoritma *Decision Tree* untuk Klasifikasi Pelanggan Aktif atau Tidak Aktif pada Data Bank. *Karimah Tauhid*, 3(2), 1860-1874

Liana Isnaini: Universitas Gadjah Mada.
E-mail: lianaisnaini@mail.ugm.ac.id

Muh. Fadlan: Universitas Gadjah Mada.
E-mail: muh.fadlan@mail.ugm.ac.id

Anisya Ananda Permatasari: Universitas Gadjah Mada.
E-mail: anisya.a@mail.ugm.ac.id

VINSENSIUS SETIA D. S. RUEK: Universitas Gajah Mada.
E-mail: vinsensiussetiadsruek@mail.ugm.ac.id