Classification

Muhammad Apriandito

5/23/2019

Pada Praktek kali ini kita akan membuat model klasifikasi dengan algoritma Decision Tree, Naive Bayes, dan K-NN menggunakan dataset insurance. Dataset ini merupakan dataset yang didapatkan dari kaggle, namun telah melalui tahapan pre-processing. sehingga data yang digunakan sudah dalam kondisi baik/siap digunakan.

Decision Tree

Import Library

```r
# Import Library
library(rpart)
library(rattle)
```

```
## Warning: package 'rattle' was built under R version 3.5.3
```

```
## Rattle: A free graphical interface for data science with R.
## Version 5.2.0 Copyright (c) 2006-2018 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.
```

```r
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 3.5.3
```

```r
library(RColorBrewer)
```

Import Dataset

```r
#Import Data
insurance <- read.csv("insurance.csv")
```

Data Exploration

```r
#Melihat Kondisi Data
dim(insurance)
```

```
## [1] 1338    8
```

```r
head(insurance,10)
```

```
##    Age    Sex    Bmi Children     Smoker    Region   Charges Claim
## 1   19 Female 27.900        0     Smoker Southwest 16884.924   Yes
## 2   18   Male 33.770        1 Non Smoker Southeast  1725.552   Yes
## 3   28   Male 33.000        3 Non Smoker Southeast  4449.462    No
## 4   33   Male 22.705        0 Non Smoker Northwest 21984.471    No
## 5   32   Male 28.880        0 Non Smoker Northwest  3866.855   Yes
## 6   31 Female 25.740        0 Non Smoker Southeast  3756.622    No
## 7   46 Female 33.440        1 Non Smoker Southeast  8240.590   Yes
## 8   37 Female 27.740        3 Non Smoker Northwest  7281.506    No
## 9   37   Male 29.830        2 Non Smoker Northeast  6406.411    No
## 10  60 Female 25.840        0 Non Smoker Northwest 28923.137    No
```

```
#Melihat Data Kosong
sum(is.na(insurance))
```

```
## [1] 0
```

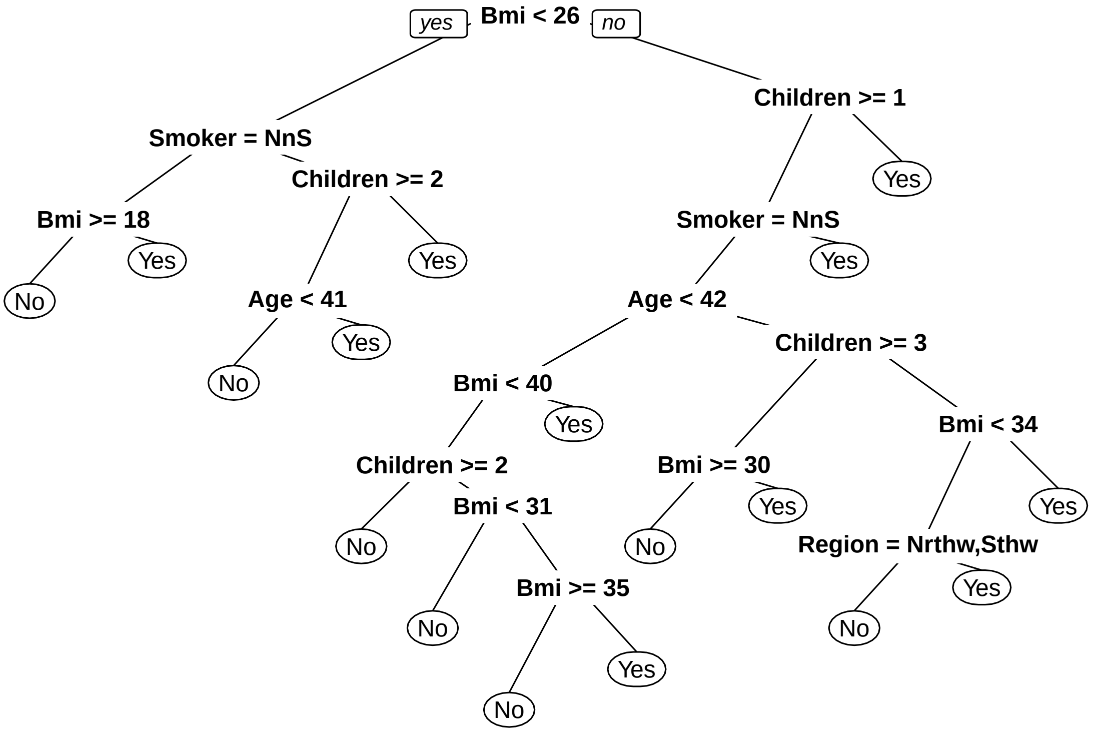
Data Preprocessing

```
#Membangi Data Ke Training dan Testing (70:30)
index_train <- sample(1:nrow(insurance), 0.7 * nrow(insurance))
train <- insurance[index_train, ]
test <- insurance[-index_train, ]
```

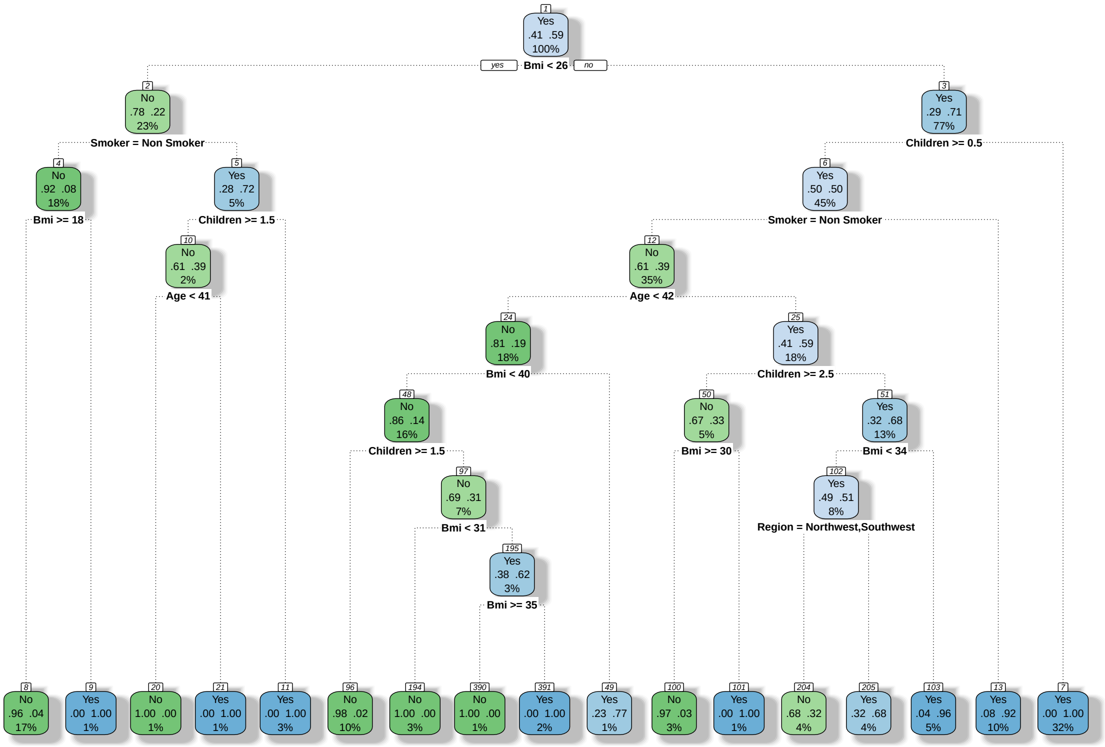
Model Building

```
#Membuat Model Decison Tree Untuk Mengklasifikasi Apakah Seseorang akan klaim Asuransi atau tidak.
tree <- rpart(Claim ~., train, method = "class")
```

```
#Memvisualisasikan Model Decision Tree
prp(tree)
```



```
#Memvisualisasikan Decison Tree dengan lebih informatif
fancyRpartPlot(tree)
```



Rattle 2019−May−23 02:12:06 Dhito

```
#Menggunakan Untuk Melakukan Prediksi Pada Data Testing
prediction <- predict(tree, test, type = "class")
```

Validation

```
#Validasi Menggunakan Confussion Matrix
conf <- table(test$Claim, prediction)
conf
```

```
##      prediction
##         No Yes
##   No   160  15
##   Yes   12 215
```

```
TP <- conf[1, 1]
FN <- conf[1, 2]
FP <- conf[2, 1]
TN <- conf[2, 2]
```

```
#Menghitung Nilai Akurasi
acc <- (TP + TN)/(TP + FN + FP + TN)
accdt <- acc
acc
```

```
## [1] 0.9328358
```

```
#Menghitung Nilai Precision
prec <- TP / (TP + FP)
```

```
prec
```

```
## [1] 0.9302326
```

```
#Menghitung Nilai Recall
rec <- TP / (TP + FN)
rec
```

```
## [1] 0.9142857
```

Naive Bayes

Import Library

```
#Import Library
library(naivebayes)
```

```
## Warning: package 'naivebayes' was built under R version 3.5.3
```

Model Building

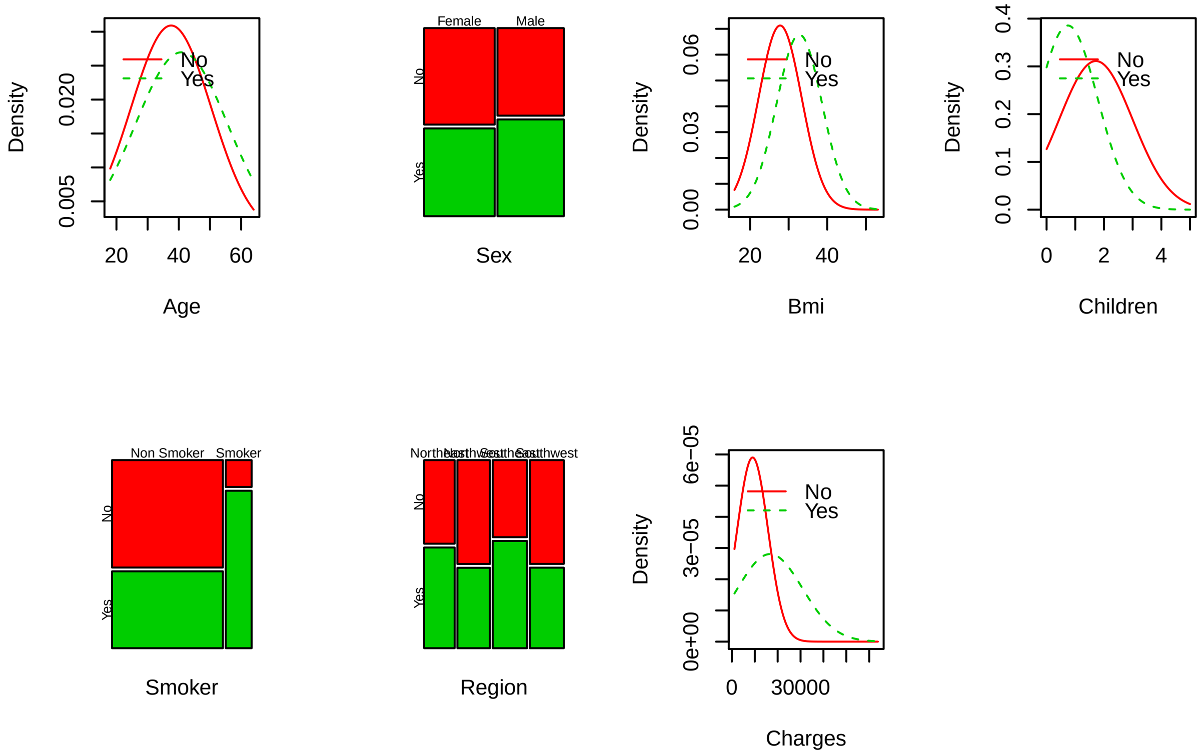
```
#Membuat model prediksi Naive Bayes
nb <- naive_bayes(Claim ~ ., data = train)

#Melihat model yang telah dibuat
nb
```

```
## ================================ Naive Bayes =================================
## Call:
## naive_bayes.formula(formula = Claim ~ ., data = train)
##
## A priori probabilities:
##
##        No       Yes
## 0.4059829 0.5940171
##
## Tables:
##
## Age           No      Yes
##   mean 37.53947 40.88669
##   sd   12.90481 14.79078
##
##
## Sex               No       Yes
##   Female 0.5394737 0.4910072
##   Male   0.4605263 0.5089928
##
##
## Bmi            No       Yes
##   mean 27.799618 32.775207
##   sd    5.597293  5.886824
##
##
## Children          No       Yes
##     mean 1.7184211 0.7428058
##     sd   1.2819747 1.0347313
```

```
##
##
## Smoker                 No        Yes
##   Non Smoker 0.94473684 0.67625899
##   Smoker     0.05526316 0.32374101
##
## # ... and 2 more tables
```

```
#Visualisasi Model
par(mfrow=c(2,4))
plot(nb)
```



```
#Melakukan prediksi dengan data testing
pred_nb <- predict(nb, as.data.frame(test))
```

Validation

```
#Membuat Confussion Matrix Naive Bayes
confnb <- table(test$Claim, pred_nb)
confnb
```

```
##      pred_nb
##        No Yes
##   No  151  24
##   Yes  58 169
```

```
TPn <- confnb[1, 1]
FNn <- confnb[1, 2]
```

```r
FPn <- confnb[2, 1]
TNn <- confnb[2, 2]
```

```r
#Menghitung Nilai Akurasi
accnb <- (TPn + TNn)/(TPn + FNn + FPn + TNn)
accnb
```

```
## [1] 0.7960199
```

```r
#Menghitung Nilai Precision
precnb <- TPn / (TPn + FPn)
precnb
```

```
## [1] 0.722488
```

```r
#Menghitung Nilai Recall
recnb <- TPn / (TPn + FNn)
recnb
```

```
## [1] 0.8628571
```

K-NN

Import Library

```r
#import library yang dibutuhkan
library(class)
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------------------

## v ggplot2 3.1.1       v purrr   0.3.2
## v tibble  2.1.1       v dplyr   0.8.0.1
## v tidyr   0.8.3       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0

## Warning: package 'ggplot2' was built under R version 3.5.3

## Warning: package 'tibble' was built under R version 3.5.3

## Warning: package 'tidyr' was built under R version 3.5.3

## Warning: package 'purrr' was built under R version 3.5.3

## Warning: package 'dplyr' was built under R version 3.5.3

## -- Conflicts ---------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Data Pre-Processing

```r
#Mengubah Data Ke Tipe Numerik
insurance1 <- insurance %>% mutate_if(is.factor, as.numeric)
```

```r
#Membuat fungsi Normalisasi
normalize<-function(x){
  temp<-(x-min(x))/(max(x)-min(x))
  return(temp)
}
```

```r
#Melakukan Normalisasi
kinsurance_n<-as.data.frame(lapply(insurance1[,c(1:7)],normalize))
```

```r
#Membagi ke Data Train dan Data Testing
index_train <- sample(1:nrow(kinsurance_n), 0.7 * nrow(kinsurance_n))
kinsurance_train <- kinsurance_n[index_train, ]
kinsurance_test <- kinsurance_n[-index_train, ]
```

```r
#Mengambil Label
kinsurance_train_target<-insurance1[index_train,8]
kinsurance_test_target<-insurance1[-index_train, 8]
```

Model Building

```r
#Membuat KNN-Model dengan Nilai K=2
knnmodel <-knn(train=kinsurance_train,test=kinsurance_test,cl=kinsurance_train_target,k=2)
```

Validation

```r
#Validasi Menggunakan Confussion Matrix
confknn <- table(kinsurance_test_target, knnmodel)
confknn
```

```
##                         knnmodel
## kinsurance_test_target   1   2
##                      1 143  32
##                      2  23 204
```

```r
TPk <- confknn[1, 1]
FNk <- confknn[1, 2]
FPk <- confknn[2, 1]
TNk <- confknn[2, 2]
```

```r
#Melihat Nilai Akurasi K-NN
acck <- (TPk + TNk)/(TPk + FNk + FPk + TNk)
acck
```

```
## [1] 0.8631841
```

```r
#Melihat Nilai Precision K-NN
preck <- TPk / (TPk + FPk)
preck
```

```
## [1] 0.8614458
```

```r
#Melihat Nilai Recall K-NN
reck <- TPk / (TPk + FNk)
reck
```

```
## [1] 0.8171429
```

Model Comparison

```
#Nilai Akurasi Decision Tree
accdt
```

```
## [1] 0.9328358
```

```
#Nilai Akurasi Naive Bayes
accnb
```

```
## [1] 0.7960199
```

```
#Nilai Akurasi K-NN
acck
```

```
## [1] 0.8631841
```