setwd("C:/Users/Matrix/Desktop/New folder")

> filenameo <- "heart.csv"

> heartds <- read.csv(filenameo, header=TRUE)

> dim(heartds)

[1] 289 14

> heartds$output<- as.factor(heartds$output)

>

>

>

> sapply(heartds, class)

age sex cp trtbps chol fbs restecg thalachh

"integer" "integer" "integer" "integer" "integer" "integer" "integer" "integer"

exng oldpeak slp caa thall output

"integer" "numeric" "integer" "integer" "integer" "factor"

> head(heartds)

age sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall output

1 60 1 3 145 233 1 0 150 0 2.3 0 0 1 yes

2 35 1 2 130 250 0 1 187 0 3.5 0 0 2 yes

3 41 0 1 130 204 0 0 172 0 1.4 2 0 2 yes

4 55 1 1 120 236 0 1 178 0 0.8 2 0 2 yes

5 56 0 0 120 354 0 1 163 1 0.6 2 0 2 yes

6 55 1 0 140 192 0 1 148 0 0.4 1 0 1 yes

> summary(heartds)

age sex cp trtbps chol

Min. :29.00 Min. :0.0000 Min. :0.000 Min. : 94.0 Min. :126

1st Qu.:47.00 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:120.0 1st Qu.:212

Median :54.00 Median :1.0000 Median :1.000 Median :130.0 Median :243

Mean :54.01 Mean :0.6782 Mean :1.021 Mean :131.4 Mean :248

3rd Qu.:60.00 3rd Qu.:1.0000 3rd Qu.:2.000 3rd Qu.:140.0 3rd Qu.:276

Max. :77.00 Max. :1.0000 Max. :3.000 Max. :200.0 Max. :564

fbs restecg thalachh exng

Min. :0.0000 Min. :0.0000 Min. : 71.0 Min. :0.0000

1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:136.0 1st Qu.:0.0000

Median :0.0000 Median :1.0000 Median :154.0 Median :0.0000

Mean :0.1453 Mean :0.5156 Mean :150.2 Mean :0.3183

3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:168.0 3rd Qu.:1.0000

Max. :1.0000 Max. :2.0000 Max. :202.0 Max. :1.0000

oldpeak slp caa thall output

Min. :0.000 Min. :0.000 Min. :0.0000 Min. :0.000 no :124

1st Qu.:0.000 1st Qu.:1.000 1st Qu.:0.0000 1st Qu.:2.000 yes:165

Median :0.600 Median :1.000 Median :0.0000 Median :2.000

Mean :1.008 Mean :1.419 Mean :0.7128 Mean :2.315

3rd Qu.:1.600 3rd Qu.:2.000 3rd Qu.:1.0000 3rd Qu.:3.000

Max. :6.200 Max. :2.000 Max. :4.0000 Max. :3.000

> filenametr <- "train.csv"

> trainds <- read.csv(filenametr, header=TRUE)

> dim(trainds)

[1] 231 14

> trainds$output<- as.factor(trainds$output)

>

> sapply(trainds, class)

age sex cp trtbps chol fbs restecg thalachh

"integer" "integer" "integer" "integer" "integer" "integer" "integer" "integer"

exng oldpeak slp caa thall output

"integer" "numeric" "integer" "integer" "integer" "factor"

> head(trainds)

age sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall output

1 60 1 3 145 233 1 0 150 0 2.3 0 0 1 yes

2 35 1 2 130 250 0 1 187 0 3.5 0 0 2 yes

3 41 0 1 130 204 0 0 172 0 1.4 2 0 2 yes

4 56 0 0 120 354 0 1 163 1 0.6 2 0 2 yes

5 55 1 0 140 192 0 1 148 0 0.4 1 0 1 yes

6 56 0 1 140 294 0 0 153 0 1.3 1 0 2 yes

> summary(trainds)

age sex cp trtbps chol

Min. :29.00 Min. :0.0000 Min. :0 Min. : 94.0 Min. :126.0

1st Qu.:47.00 1st Qu.:0.0000 1st Qu.:0 1st Qu.:120.0 1st Qu.:212.5

Median :54.00 Median :1.0000 Median :1 Median :130.0 Median :245.0

Mean :54.28 Mean :0.6623 Mean :1 Mean :131.9 Mean :249.7

3rd Qu.:61.00 3rd Qu.:1.0000 3rd Qu.:2 3rd Qu.:140.0 3rd Qu.:279.5

Max. :77.00 Max. :1.0000 Max. :3 Max. :200.0 Max. :564.0

fbs restecg thalachh exng

Min. :0.0000 Min. :0.0000 Min. : 71.0 Min. :0.000

1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:133.0 1st Qu.:0.000

Median :0.0000 Median :1.0000 Median :155.0 Median :0.000

Mean :0.1472 Mean :0.5368 Mean :150.1 Mean :0.316

3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:168.0 3rd Qu.:1.000

Max. :1.0000 Max. :2.0000 Max. :202.0 Max. :1.000

oldpeak slp caa thall output

Min. :0.0000 Min. :0.000 Min. :0.0000 Min. :0.000 no : 99

1st Qu.:0.0000 1st Qu.:1.000 1st Qu.:0.0000 1st Qu.:2.000 yes:132

Median :0.6000 Median :1.000 Median :0.0000 Median :2.000

Mean :0.9874 Mean :1.429 Mean :0.7056 Mean :2.338

3rd Qu.:1.6000 3rd Qu.:2.000 3rd Qu.:1.0000 3rd Qu.:3.000

Max. :5.6000 Max. :2.000 Max. :4.0000 Max. :3.000

> filenamete <- "test.csv"

> testds <- read.csv(filenamete, header=TRUE)

> dim(testds)

[1] 58 14

> testds$output<- as.factor(testds$output)

> sapply(testds, class)

age sex cp trtbps chol fbs restecg thalachh

"integer" "integer" "integer" "integer" "integer" "integer" "integer" "integer"

exng oldpeak slp caa thall output

"integer" "numeric" "integer" "integer" "integer" "factor"

> head(testds)

age sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall output

1 55 1 1 120 236 0 1 178 0 0.8 2 0 2 yes

2 44 1 1 120 263 0 1 173 0 0.0 2 0 3 yes

3 48 0 2 130 275 0 1 139 0 0.2 2 0 2 yes

4 59 1 2 150 212 1 1 157 0 1.6 2 0 2 yes

5 51 1 2 110 175 0 1 123 0 0.6 2 0 2 yes

6 44 1 1 130 219 0 0 188 0 0.0 2 0 2 yes

> summary(testds)

age sex cp trtbps

Min. :34.00 Min. :0.0000 Min. :0.000 Min. :100.0

1st Qu.:45.75 1st Qu.:0.2500 1st Qu.:0.000 1st Qu.:113.5

Median :55.00 Median :1.0000 Median :1.000 Median :125.5

Mean :52.95 Mean :0.7414 Mean :1.103 Mean :129.1

3rd Qu.:59.00 3rd Qu.:1.0000 3rd Qu.:2.000 3rd Qu.:138.0

Max. :69.00 Max. :1.0000 Max. :3.000 Max. :178.0

chol fbs restecg thalachh

Min. :157.0 Min. :0.0000 Min. :0.000 Min. :103.0

1st Qu.:209.0 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:138.2

Median :234.5 Median :0.0000 Median :0.000 Median :150.5

Mean :240.9 Mean :0.1379 Mean :0.431 Mean :150.6

3rd Qu.:264.5 3rd Qu.:0.0000 3rd Qu.:1.000 3rd Qu.:164.5

Max. :409.0 Max. :1.0000 Max. :1.000 Max. :188.0

exng oldpeak slp caa

Min. :0.0000 Min. :0.000 Min. :0.000 Min. :0.0000

1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:1.000 1st Qu.:0.0000

Median :0.0000 Median :0.700 Median :1.000 Median :0.0000

Mean :0.3276 Mean :1.088 Mean :1.379 Mean :0.7414

3rd Qu.:1.0000 3rd Qu.:1.800 3rd Qu.:2.000 3rd Qu.:1.0000

Max. :1.0000 Max. :6.200 Max. :2.000 Max. :4.0000

thall output

Min. :0.000 no :25

1st Qu.:2.000 yes:33

Median :2.000

Mean :2.224

3rd Qu.:3.000

Max. :3.000

>

> library(caret)

Loading required package: ggplot2

Loading required package: lattice

> library(ggplot2)

> library(knitr)

> library(tidyr)

> library(dplyr)

Attaching package: ‘dplyr’

The following objects are masked from ‘package:stats’:

filter, lag

The following objects are masked from ‘package:base’:

intersect, setdiff, setequal, union

>

>

> library(ROCR)

> set.seed(100)

> x = trainControl(method = "repeatedcv", number = 10, repeats = 3, classProbs = TRUE, summaryFunction = twoClassSummary)

> knn = train(output~. , data = trainds[,1:14], method = "knn",preProcess = c("center","scale"),trControl = x, metric = "ROC", tuneLength = 10)

> knn

k-Nearest Neighbors

231 samples

13 predictor

2 classes: 'no', 'yes'

Pre-processing: centered (13), scaled (13)

Resampling: Cross-Validated (10 fold, repeated 3 times)

Summary of sample sizes: 208, 208, 208, 207, 207, 209, ...

Resampling results across tuning parameters:

k ROC Sens Spec

5 0.8811427 0.7477778 0.9069597

7 0.8859290 0.7377778 0.9199634

9 0.8936121 0.7381481 0.9102564

11 0.9010969 0.7218519 0.9228938

13 0.9012098 0.7007407 0.9373626

15 0.9051272 0.6981481 0.9521978

17 0.9078449 0.6811111 0.9523810

19 0.9076496 0.6770370 0.9523810

21 0.9096988 0.6866667 0.9547619

23 0.9149746 0.6903704 0.9549451

ROC was used to select the optimal model using the largest value.

The final value used for the model was k = 23.

> plot(knn)

> testds$Predicted = predict(knn, testds, "prob")[,2]

> plot(performance(prediction(testds$Predicted, testds$output),"tpr", "fpr"))

> testds$Predicted = ifelse(testds$Predicted> 0.5, "no","yes")

> confusionMatrix(factor(testds$Predicted), factor(testds$output))

Confusion Matrix and Statistics

Reference

Prediction no yes

no 8 29

yes 17 4

Accuracy : 0.2069

95% CI : (0.1117, 0.3335)

No Information Rate : 0.569

P-Value [Acc > NIR] : 1.0000

Kappa : -0.5281

Mcnemar's Test P-Value : 0.1048

Sensitivity : 0.3200

Specificity : 0.1212

Pos Pred Value : 0.2162

Neg Pred Value : 0.1905

Prevalence : 0.4310

Detection Rate : 0.1379

Detection Prevalence : 0.6379

Balanced Accuracy : 0.2206

'Positive' Class : no

> testds$Predicted = ifelse(testds$Predicted> 0.5, "yes","no")

> confusionMatrix(factor(testds$Predicted), factor(testds$output))

Confusion Matrix and Statistics

Reference

Prediction no yes

no 0 0

yes 25 33

Accuracy : 0.569

95% CI : (0.4323, 0.6984)

No Information Rate : 0.569

P-Value [Acc > NIR] : 0.555

Kappa : 0

Mcnemar's Test P-Value : 1.587e-06

Sensitivity : 0.000

Specificity : 1.000

Pos Pred Value : NaN

Neg Pred Value : 0.569

Prevalence : 0.431

Detection Rate : 0.000

Detection Prevalence : 0.000

Balanced Accuracy : 0.500

'Positive' Class : no

Warning message:

In confusionMatrix.default(factor(testds$Predicted), factor(testds$output)) :

Levels are not in the same order for reference and data. Refactoring data to match.