Naive Bayes and k-fold crossvalidation

#Introduction

This tutorial will use a heart risk dataset. First we load the data set and briefly inspect it. Note: that it contains both qual and quant data.

```
#setwd("C:\\Users\\jerem\\Google Drive\\Online\\iCuse\\IST707\\Week7")

filename="LabeledDataRiskHeart.csv"
RiskDF <- read.csv(filename, header = TRUE, stringsAsFactors = TRUE)
(head(RiskDF))
```

```
##    Label Gender Cholesterol MaritalStatus Weight Height StressLevel
## 1   Risk      M         251             S    267     70           5
## 2 NoRisk      F         105             M    103     62           1
## 3 Medium      M         156             S    193     72           3
## 4 NoRisk      F         109             M    100     63           2
## 5   Risk      M         198             S    210     70           4
## 6   Risk      F         189             S    189     64           3
```

```
(str(RiskDF))
```

```
## 'data.frame':    32 obs. of  7 variables:
##  $ Label        : Factor w/ 3 levels "Medium","NoRisk",..: 3 2 1 2 3 3 2 1 3 2 ...
##  $ Gender       : Factor w/ 2 levels "F","M": 2 1 2 1 2 1 1 1 2 2 ...
##  $ Cholesterol  : int  251 105 156 109 198 189 121 134 250 118 ...
##  $ MaritalStatus: Factor w/ 2 levels "M","S": 2 1 2 1 2 2 2 1 2 1 ...
##  $ Weight       : int  267 103 193 100 210 189 105 125 156 190 ...
##  $ Height       : int  70 62 72 63 70 64 65 60 69 71 ...
##  $ StressLevel  : int  5 1 3 2 4 3 1 2 5 3 ...
```

```
## NULL
```

```
(nrow(RiskDF))
```

```
## [1] 32
```

```
RiskDF$StressLevel<-as.factor(RiskDF$StressLevel)
RiskDF$Cholesterol<-as.numeric(RiskDF$Cholesterol)
RiskDF$Weight<-as.numeric(RiskDF$Weight)
RiskDF$Height<-as.numeric(RiskDF$Height)
(str(RiskDF))
```

```
## 'data.frame':    32 obs. of  7 variables:
##  $ Label        : Factor w/ 3 levels "Medium","NoRisk",..: 3 2 1 2 3 3 2 1 3 2 ...
##  $ Gender       : Factor w/ 2 levels "F","M": 2 1 2 1 2 1 1 1 2 2 ...
##  $ Cholesterol  : num  251 105 156 109 198 189 121 134 250 118 ...
##  $ MaritalStatus: Factor w/ 2 levels "M","S": 2 1 2 1 2 2 2 1 2 1 ...
##  $ Weight       : num  267 103 193 100 210 189 105 125 156 190 ...
##  $ Height       : num  70 62 72 63 70 64 65 60 69 71 ...
##  $ StressLevel  : Factor w/ 5 levels "1","2","3","4",..: 5 1 3 2 4 3 1 2 5 3 ...
```

```
## NULL
```

Crossvalidation

Next we set up or experimental evaluation. We will use k-fold crossvalidation. The split function helps to facilitate the partitioning of the data set which determines the k folds.

```
#################################################################
#############  Create k-folds for k-fold validation ############
#################################################################


# Number of observations
N <- nrow(RiskDF)
# Number of desired splits
kfolds <- 2
# Generate indices of holdout observations
# Note if N is not a multiple of folds you will get a warning, but is OK.
holdout <- split(sample(1:N), 1:kfolds)
```

Experimental Validation

Running k-fold crossvalidation requires that we run k trials. This is facilitated using a for loop that iterates k times. During each iteration, k-1 partition are assigned to the training set and the remaining partition is the test set.

Naive Bayes

During each iteration a Naive Bayes model is trained and tested using naiveBayes and predict, respectively.

```
#####  Run training and Testing for each of the k-folds
AllResults<-list()
AllLabels<-list()
for (k in 1:kfolds){

  RiskDF_Test=RiskDF[holdout[[k]], ]
  RiskDF_Train=RiskDF[-holdout[[k]], ]
  ## View the created Test and Train sets
  (head(RiskDF_Train))
  (table(RiskDF_Test$Label))

  ## Make sure you take the labels out of the testing data
  (head(RiskDF_Test))
  RiskDF_Test_noLabel<-RiskDF_Test[-c(1)]
  RiskDF_Test_justLabel<-RiskDF_Test$Label
  (head(RiskDF_Test_noLabel))


  #### e1071
  ## formula is label ~ x1 + x2 + .  NOTE that label ~. is "use all to create model"
  NB_e1071<-naiveBayes(Label~., data=RiskDF_Train, na.action = na.pass)
  NB_e1071_Pred <- predict(NB_e1071, RiskDF_Test_noLabel)
  NB_e1071
```

```
  ## Accumulate results from each fold
  AllResults<- c(AllResults,NB_e1071_Pred)
  AllLabels<- c(AllLabels, RiskDF_Test_justLabel)

}
```

Results

Results are presented in tabular form below. You can easily create a confusion matrix from this data – try it!

```
### end crossvalidation -- present results for all folds
table(unlist(AllResults),unlist(AllLabels))
```

```
##
##      1  2  3
##   1  4  1  1
##   2  3 10  0
##   3  2  0 11
```

Another NB library

Below is another NB package you can try. It is similar but also has some fun visualizations.

Laplace Modeling. Try varying the laplace parameter ... what happens? Review the PPT to determine why and explain your results.

```
## using naivebayes package
## https://cran.r-project.org/web/packages/naivebayes/naivebayes.pdf

##Also see
##https://www.rdocumentation.org/packages/naivebayes/versions/0.9.2/topics/naive_bayes
## Try varying the Laplace value ... how does this affect the results???


#prior <- as.vector(c(0, .4, .6) )

NB_object<- naive_bayes(Label~., laplace = 0 , data=RiskDF_Train)
NB_prediction<-predict(NB_object, RiskDF_Test_noLabel , type = c("class"))
head(predict(NB_object, RiskDF_Test_noLabel, type = "prob"))
```

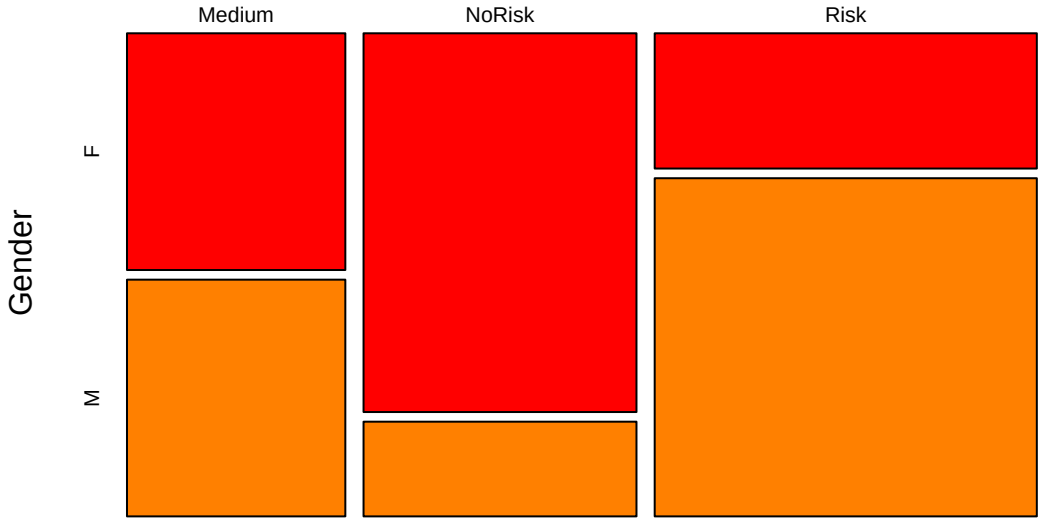
```
##            Medium       NoRisk         Risk
## [1,] 4.296970e-07 9.999994e-01 1.452974e-07
## [2,] 6.467453e-04 9.993502e-01 3.085562e-06
## [3,] 9.829678e-01 2.132800e-03 1.489936e-02
## [4,] 8.473189e-02 2.734732e-09 9.152681e-01
## [5,] 9.808211e-01 9.802094e-07 1.917797e-02
## [6,] 2.071298e-05 9.999793e-01 2.073118e-08
```

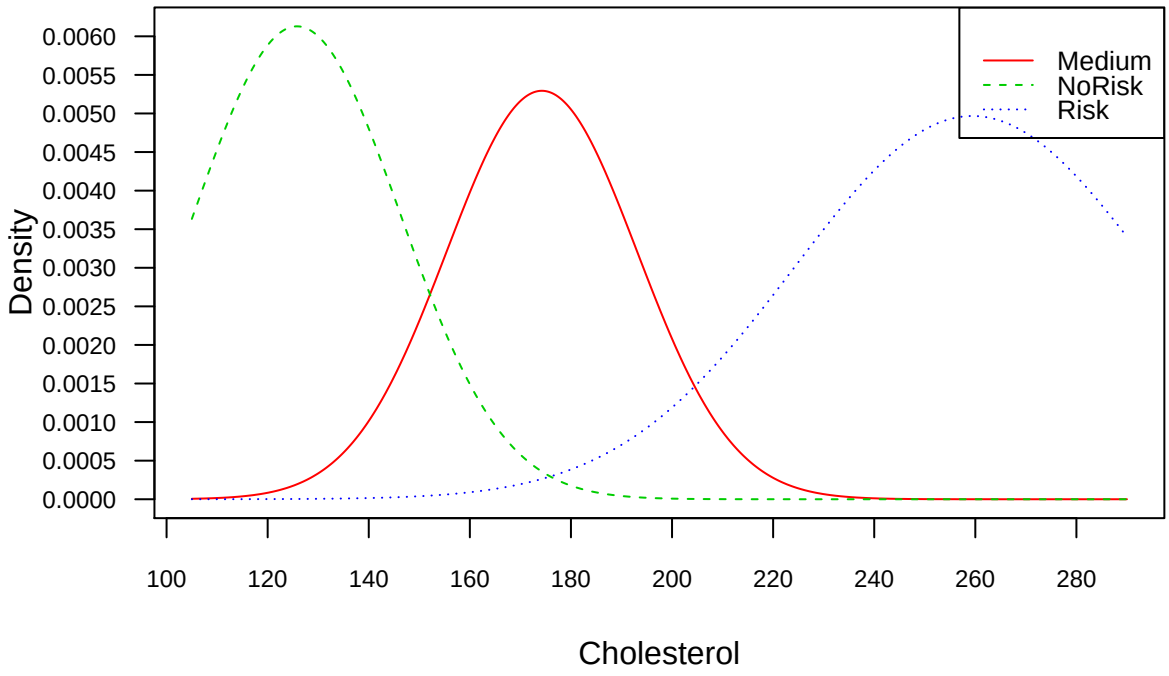
```
table(NB_prediction,RiskDF_Test_justLabel)
```

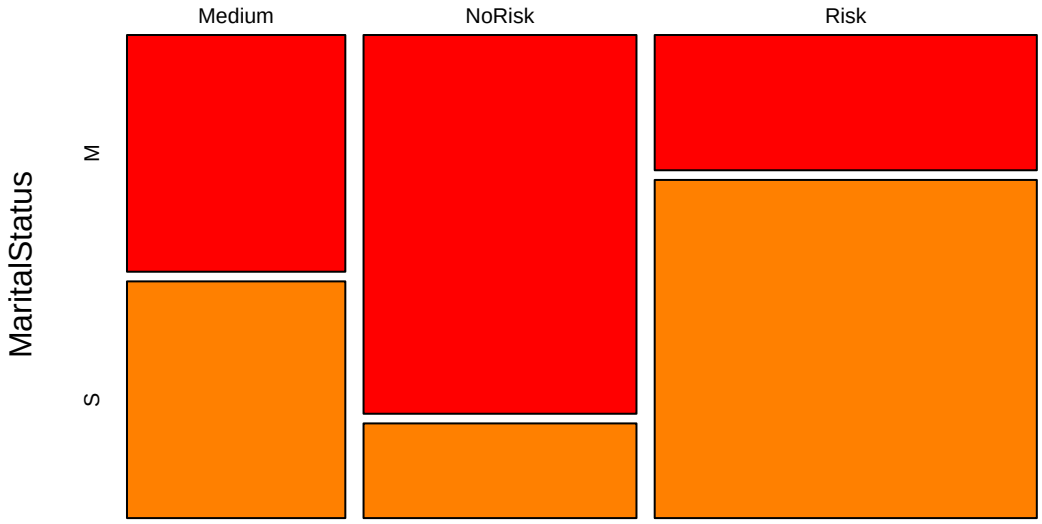
```
##              RiskDF_Test_justLabel
## NB_prediction Medium NoRisk Risk
##        Medium      2      1    1
```

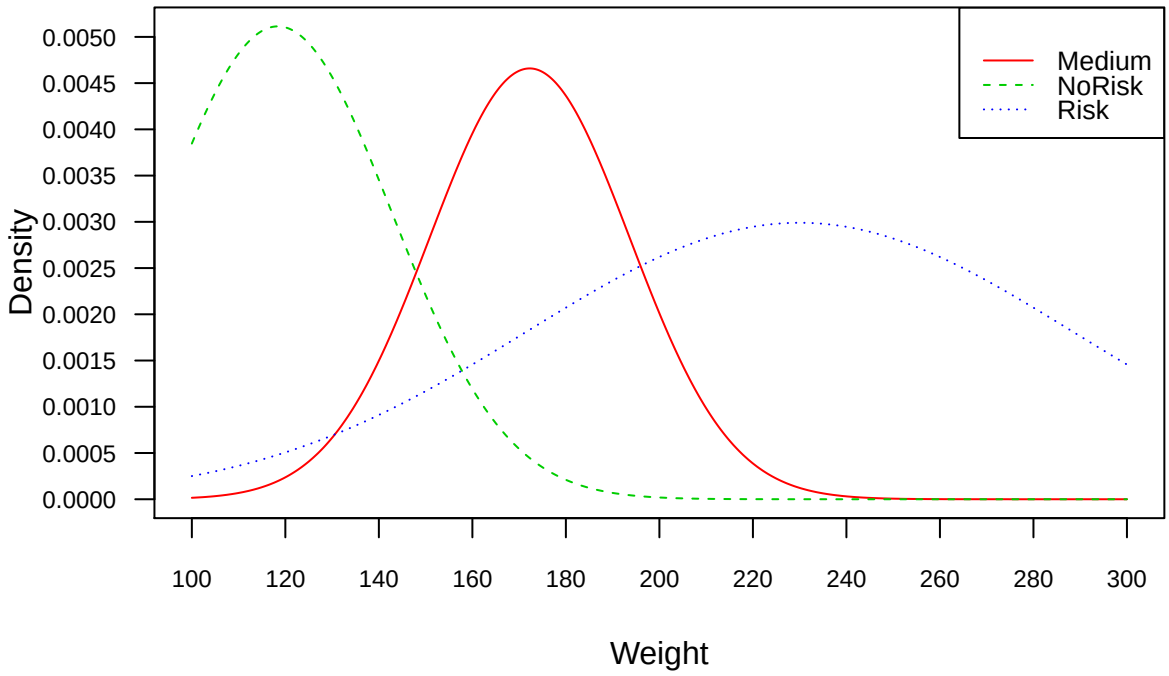
```
##         NoRisk      3      5    0
##         Risk        0      0    4
```

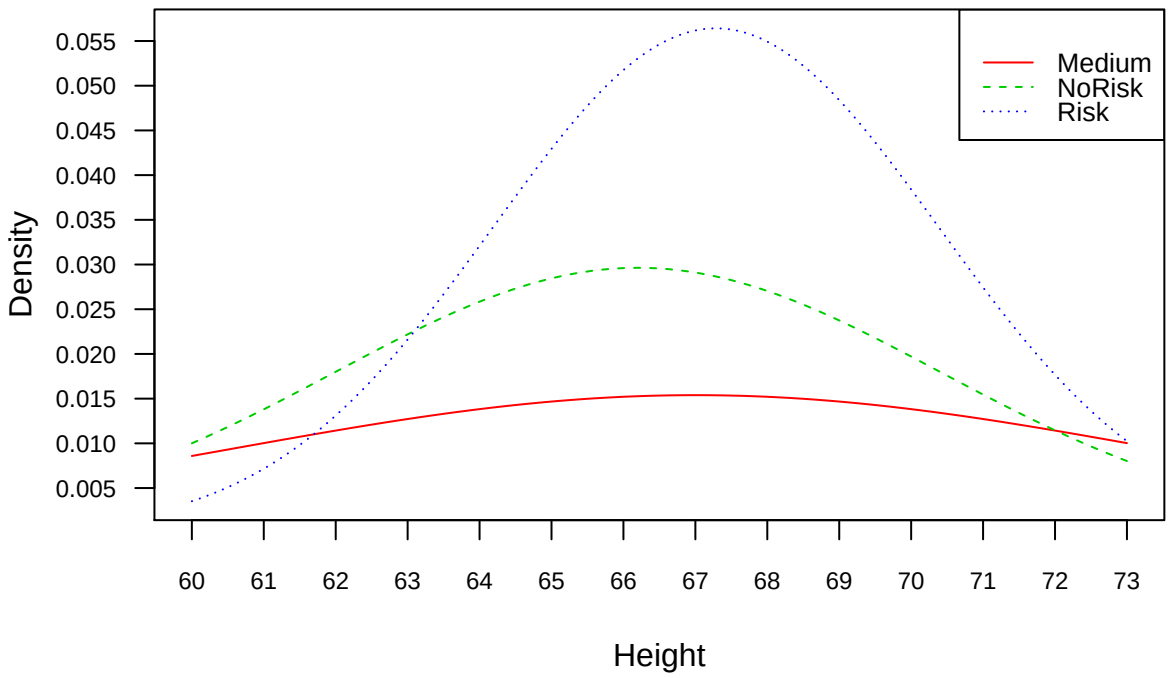
```
plot(NB_object, legend.box = TRUE)
```

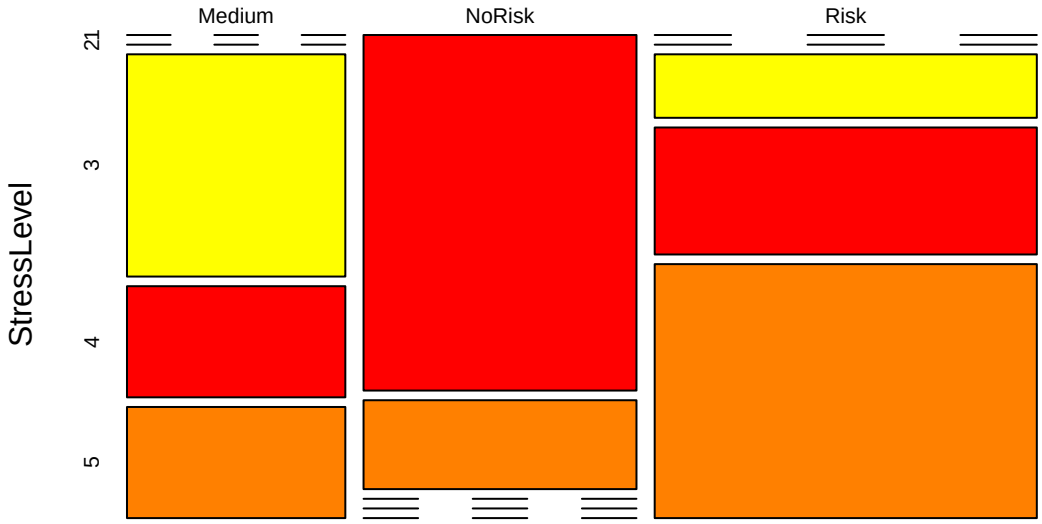



Density
0.0060
0.0055
0.0050
0.0045
0.0040
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
100
120
140
160
180
200
220
240
260
280
Cholesterol
Medium
NoRisk
Risk


Medium
NoRisk
Risk
MaritalStatus
M
S


Density
0.0050
0.0045
0.0040
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
100
120
140
160
180
200
220
240
260
280
300
Weight
Medium
NoRisk
Risk


Density
0.055
0.050
0.045
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
60
61
62
63
64
65
66
67
68
69
70
71
72
73
Height
Medium
NoRisk
Risk


Medium
NoRisk
Risk
StressLevel
21
3
4
5