

CS5525 Final Project Code Submission

Jennifer Appiah-Kubi, Rebecca DeSipio, Ajinkya Fotedar

12/11/2021

Contents

| | |
|-------------------------------|----------|
| Classification Methods | 1 |
| Decision Trees | 1 |
| KNN | 8 |
| SVM | 8 |
| Logistic Regression | 12 |

Classification Methods

First set the directory (path which contains the `heart.csv` data), and import any needed libraries

```
#setwd(" ") # uncomment to set working directory via code
library(tree)
library(randomForest) # bootstrap/bagging & random forest
library(class)        # KNN
library(caret)         # SVM
library(ggplot2)
library(cowplot)
```

Decision Trees

```
# Read in and organize data
## -- Read data
heart <- read.csv("heart.csv")
Target <- as.factor(heart$target) # target heart rate

## -- Split into training and test sets
train <- sample(1:nrow(heart), 0.75*nrow(heart))
heart.test <- heart[-train, ]
Target.test <- Target[-train]

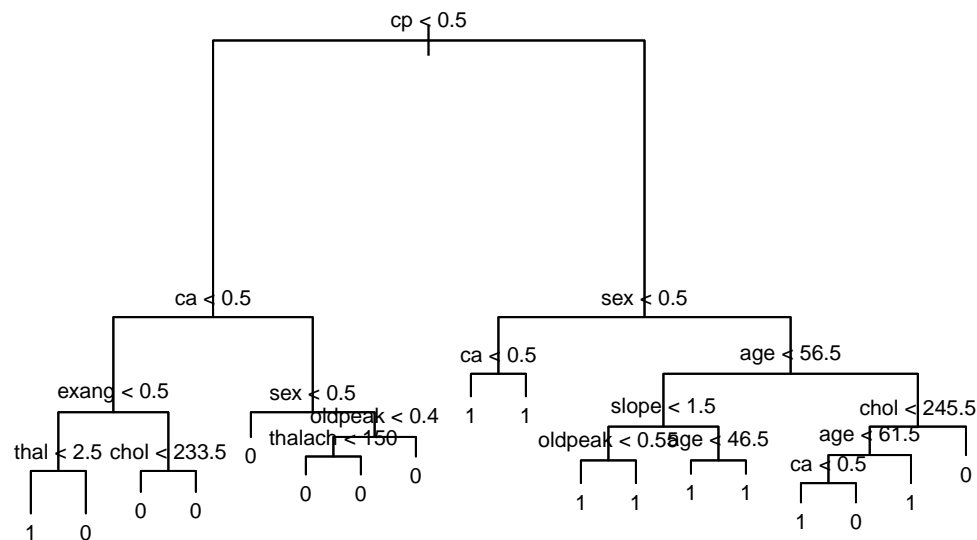
# ----- #
#                               Fit a Classification Tree                               #
# ----- #

# Fit a classification tree to the training data
set.seed(2441139)
tree.heart <- tree(Target~. -target, heart, subset=train)
summary(tree.heart)

##
```

```
## Classification tree:
## tree(formula = Target ~ . - target, data = heart, subset = train)
## Variables actually used in tree construction:
## [1] "cp"      "ca"      "exang"    "thal"    "chol"    "sex"      "oldpeak"
## [8] "thalach" "age"      "slope"
## Number of terminal nodes: 18
## Residual mean deviance: 0.5099 = 106.6 / 209
## Misclassification error rate: 0.1189 = 27 / 227

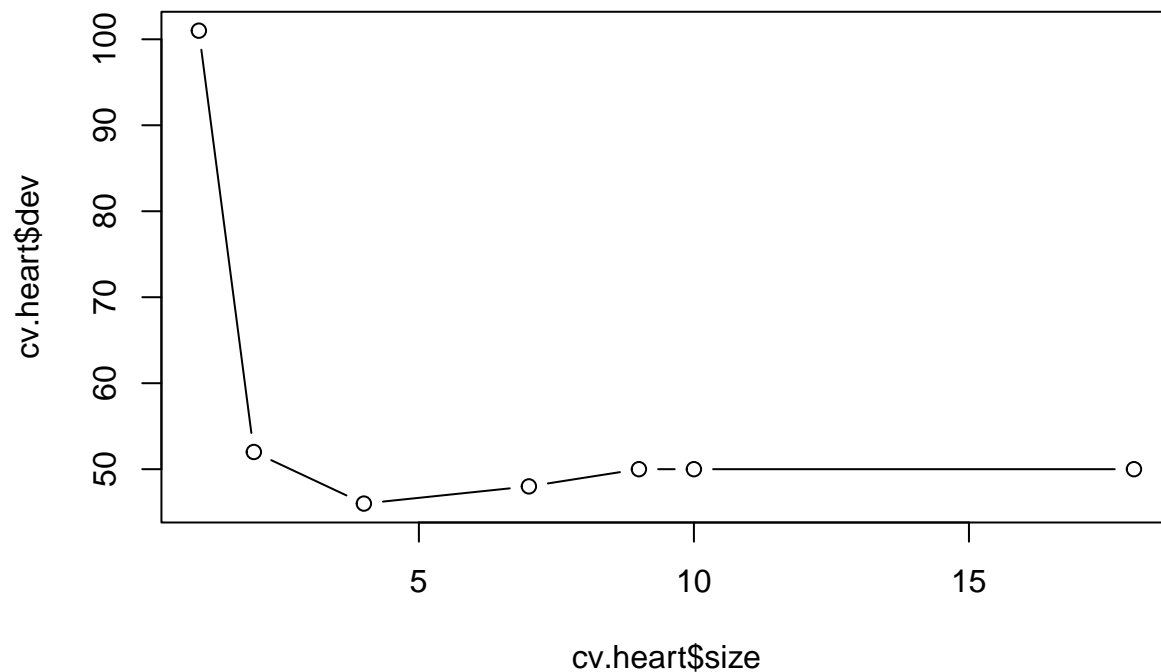
## -- Plot tree
plot(tree.heart)
text(tree.heart, pretty=1, cex=0.7)
```



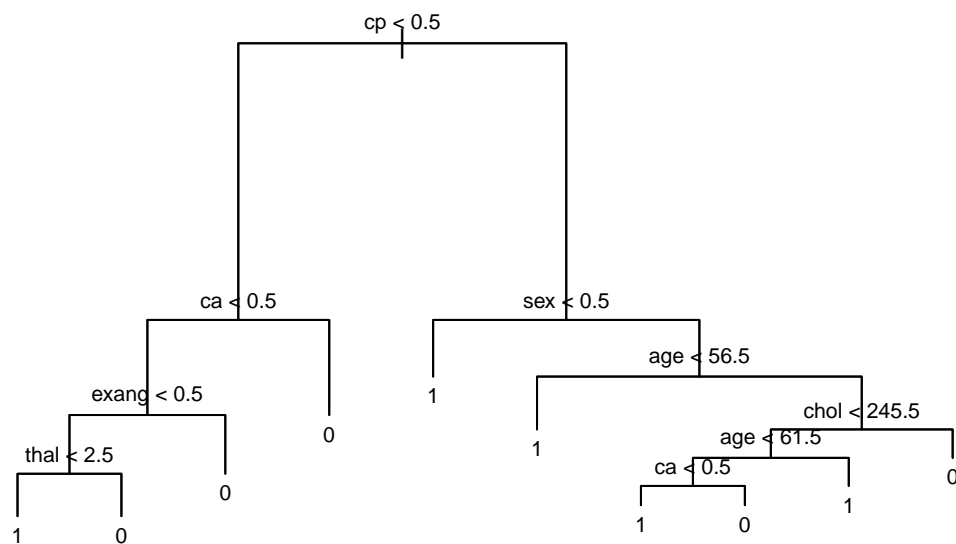
```
# Prune the classification tree
set.seed(2441139)
cv.heart <- cv.tree(tree.heart, FUN=prune.misclass)
cv.heart

## $size
## [1] 18 10 9 7 4 2 1
##
## $dev
## [1] 50 50 50 48 46 52 101
##
## $k
## [1] -Inf 0.000000 1.000000 2.000000 2.666667 6.000000 49.000000
##
## $method
## [1] "misclass"
##
## attr(,"class")
## [1] "prune" "tree.sequence"

plot(cv.heart$size, cv.heart$dev, type='b')
```



```
prune.heart <- prune.misclass(tree.heart, best=10)
plot(prune.heart)
text(prune.heart, pretty=1, cex=0.65)
```



```
# Predict using test set and pruned tree. Compare.
tree.pred <- predict(tree.heart, heart.test, type='class') # test tree
prune.pred <- predict(prune.heart, heart.test, type='class') # pruned tree

table(prune.pred, Target.test)
```

```
##           Target.test
## prune.pred  0  1
##           0 32  6
##           1  5 33
```

```
table(tree.pred, Target.test)
```

```
##           Target.test
## tree.pred 0  1
##           0 32  7
##           1  5 32
```

```
# -----#
#                               Bagging                               #
# -----#
```

```
set.seed(2441139)
```

```
# Perform bagging
```

```
bag.heart <- randomForest(as.factor(as.character(heart$target))~., data=heart,
                          subset=train, mtry=ncol(heart)-1,
                          importance=TRUE)
```

```
bag.heart
```

```
##
```

```
## Call:
```

```
## randomForest(formula = as.factor(as.character(heart$target)) ~ ., data = heart, mtry = ncol(heart))
```

```
##           Type of random forest: classification
```

```
##           Number of trees: 500
```

```
## No. of variables tried at each split: 13
```

```
##
```

```
##           OOB estimate of  error rate: 20.26%
```

```
## Confusion matrix:
```

```
##           0  1 class.error
```

```
## 0 74  27  0.2673267
```

```
## 1 19 107  0.1507937
```

```
# Predict on bagged tree
```

```
bag.pred <- predict(bag.heart, heart.test, type='class')
```

```
table(bag.pred, Target.test)
```

```
##           Target.test
```

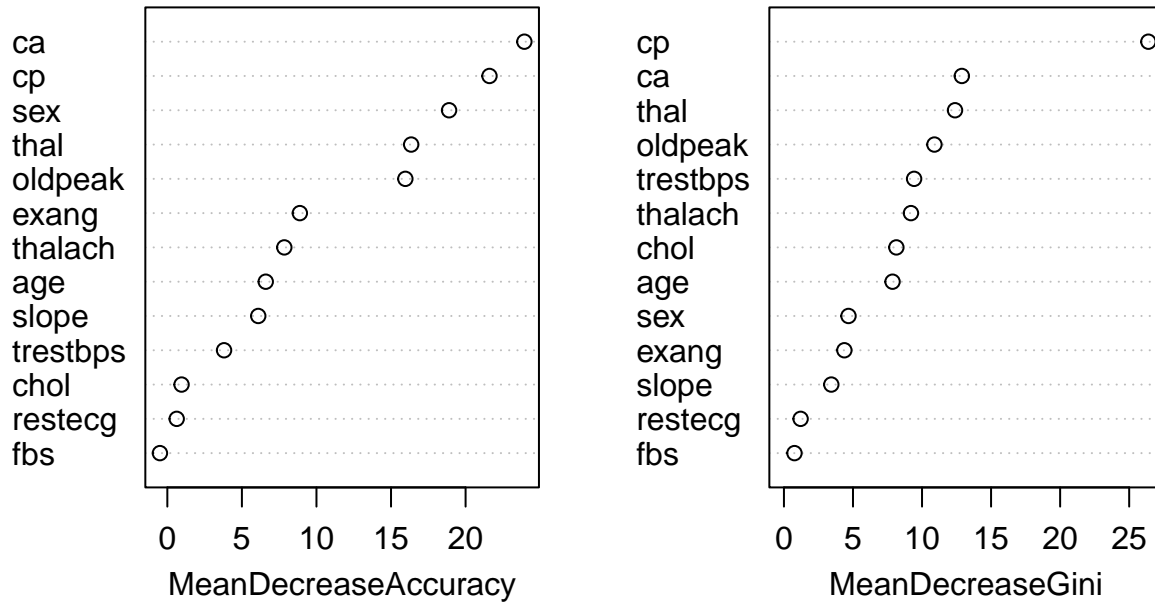
```
## bag.pred 0  1
```

```
##           0 28  2
```

```
##           1  9 37
```

```
varImpPlot(bag.heart)
```

bag.heart



```
# ----- #
#                               Random Forest                               #
# ----- #
set.seed(2441139)

# Perform Random Forest
rf.heart <- randomForest(as.factor(as.character(heart$target))~., data=heart,
                          subset=train, mtry=sqrt(ncol(heart)-1),
                          ntree=25, importance=TRUE)
rf.heart

##
## Call:
## randomForest(formula = as.factor(as.character(heart$target)) ~ ., data = heart, mtry = sqrt(ncol(heart)-1),
##               Type of random forest: classification
##               Number of trees: 25
##               No. of variables tried at each split: 4
##
##               OOB estimate of  error rate: 22.03%
## Confusion matrix:
##      0   1 class.error
## 0 69  32  0.3168317
## 1 18 108  0.1428571

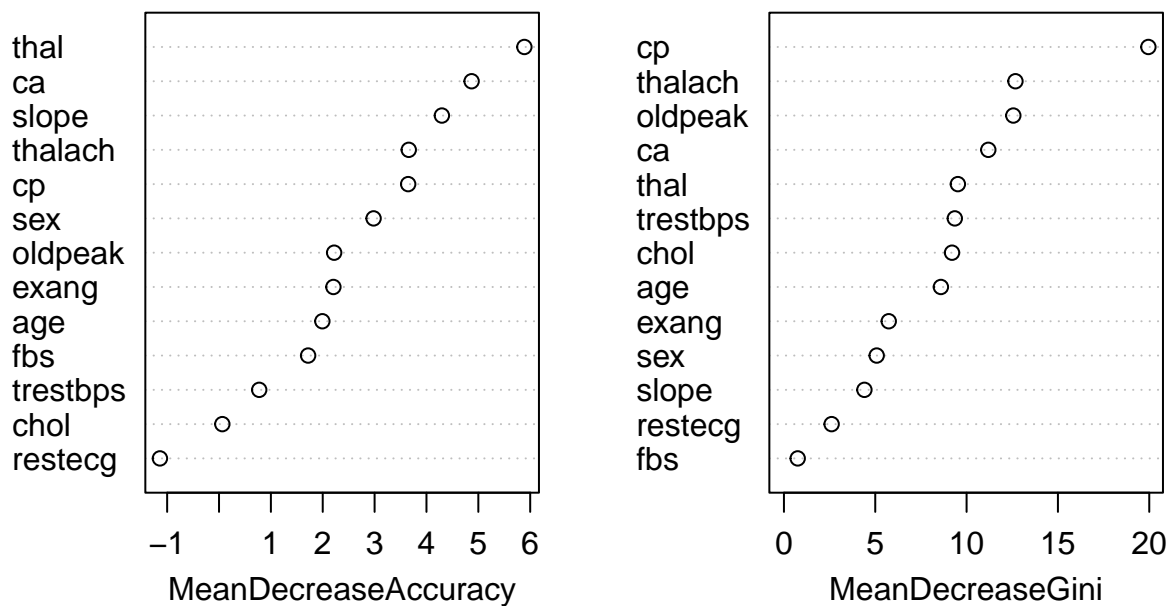
# Predict on the forest
rf.pred <- predict(rf.heart, heart.test, type='class')
table(rf.pred, Target.test)

##           Target.test
## rf.pred  0   1
```

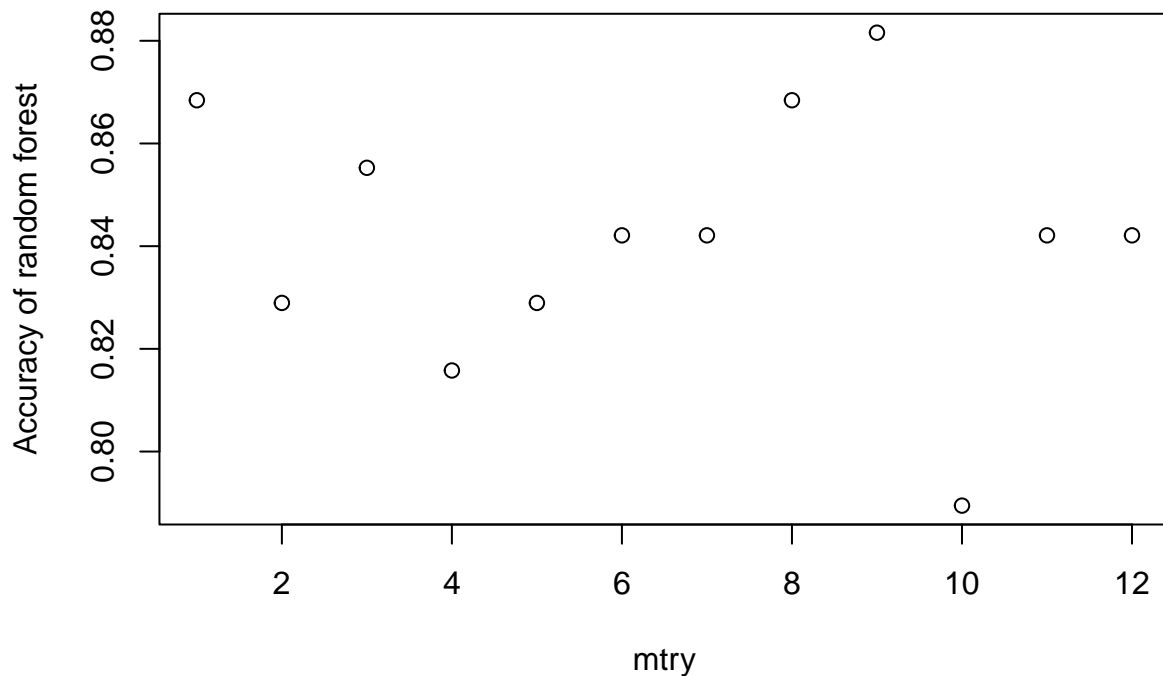
```
##      0 30  3
##      1  7 36
```

```
varImpPlot(rf.heart)
```

rf.heart



```
# ----- #
# ----- #
# Determine Best Model (Random Forest) #
# ----- #
# ----- #
# Investigate how mtry affect the accuracy
Acc <- rep(0,ncol(heart)-2)
for (m in 1:(ncol(heart)-2)){
  set.seed(2441139)
  rf.heart <- randomForest(as.factor(as.character(heart$target))~., data=heart,
                           subset=train, mtry=m,
                           ntree=25)
  rf.pred <- predict(rf.heart, heart.test, type='class')
  t <- table(rf.pred, Target.test)
  acc <- sum(diag(t))/sum(t)
  Acc[m] <- acc
}
mbest <- which(Acc==max(Acc))
plot(1:(ncol(heart)-2), Acc, xlab='mtry', ylab='Accuracy of random forest') # include plot in final sub
```



```
# Now use the best value of m for the random forest
```

```
set.seed(2441139)
```

```
rf.heart <- randomForest(as.factor(as.character(heart$target))~., data=heart,
                          subset=train, mtry=mbest,
                          ntree=25, importance=TRUE)
```

```
rf.heart
```

```
##
```

```
## Call:
```

```
## randomForest(formula = as.factor(as.character(heart$target)) ~ ., data = heart, mtry = mbest, ntree = 25, importance = TRUE)
```

```
##           Type of random forest: classification
```

```
##           Number of trees: 25
```

```
## No. of variables tried at each split: 9
```

```
##
```

```
##           OOB estimate of  error rate: 22.03%
```

```
## Confusion matrix:
```

```
##      0   1 class.error
```

```
## 0 72  29   0.2871287
```

```
## 1 21 105   0.1666667
```

```
# Predict on the forest
```

```
rf.pred <- predict(rf.heart, heart.test, type='class')
```

```
table(rf.pred, Target.test)
```

```
##           Target.test
```

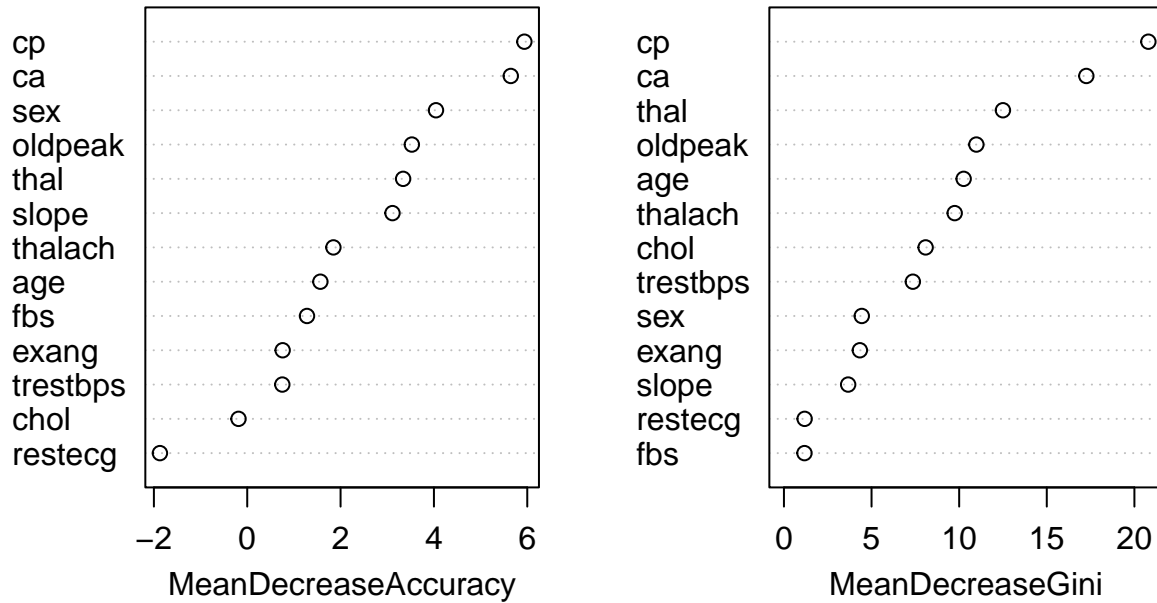
```
## rf.pred  0   1
```

```
##           0 27  3
```

```
##           1 10 36
```

```
varImpPlot(rf.heart)
```

rf.heart



KNN

```
# K-Nearest Neighbor
cl <- as.factor(heart$target[train])
knn.heart <- knn(heart[train,], heart.test, cl, k = 5, prob=TRUE)
table(knn.heart, Target.test)
```

```
##           Target.test
## knn.heart  0  1
##           0 16 12
##           1 21 27
```

SVM

```
set.seed(2441139)

# splitting data into test and train
intrain <- createDataPartition(y = heart$target, p = 0.7, list = F)

training <- heart[intrain,]
testing <- heart[-intrain,]
training[["target"]] <- as.factor(training[["target"]])

dim(training)

## [1] 213 14
```



```

dim(testing)

## [1] 90 14

# model training with svm
trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
svm.mod <- train(target ~ ., data = training, method = "svmLinear",
                  trControl = trctrl,
                  preProcess = c("center", "scale"),
                  tuneLength = 10)

svm.mod

## Support Vector Machines with Linear Kernel
##
## 213 samples
## 13 predictor
## 2 classes: '0', '1'
##
## Pre-processing: centered (13), scaled (13)
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 192, 191, 192, 192, 191, 191, ...
## Resampling results:
##
## Accuracy   Kappa
## 0.8266955   0.6511675
##
## Tuning parameter 'C' was held constant at a value of 1

# prediction using the above model
svm.pred <- predict(svm.mod, newdata = testing)
svm.pred

## [1] 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 0 0 0 1 0 1 1 1 1 1 1 0 1 1 0 0 1 1 1 1 1
## [39] 1 1 1 1 1 1 1 0 0 1 1 1 1 0 1 0 1 1 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
## [77] 0 0 0 1 1 1 0 0 1 0 1 0 0 0
## Levels: 0 1

# accuracy of the trained model
confusionMatrix(table(svm.pred, testing$target))

## Confusion Matrix and Statistics
##
##
## svm.pred  0  1
##          0 28 11
##          1  9 42
##
##              Accuracy : 0.7778
##              95% CI : (0.6779, 0.8587)
##      No Information Rate : 0.5889
##      P-Value [Acc > NIR] : 0.0001266
##
##              Kappa : 0.5448
##
##      Mcnemar's Test P-Value : 0.8230633
##

```

```

##          Sensitivity : 0.7568
##          Specificity : 0.7925
##          Pos Pred Value : 0.7179
##          Neg Pred Value : 0.8235
##          Prevalence : 0.4111
##          Detection Rate : 0.3111
##          Detection Prevalence : 0.4333
##          Balanced Accuracy : 0.7746
##
##          'Positive' Class : 0
##

# costs for further tuning with 10-fold cross-validation
grid <- expand.grid(C = c(0, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 5))

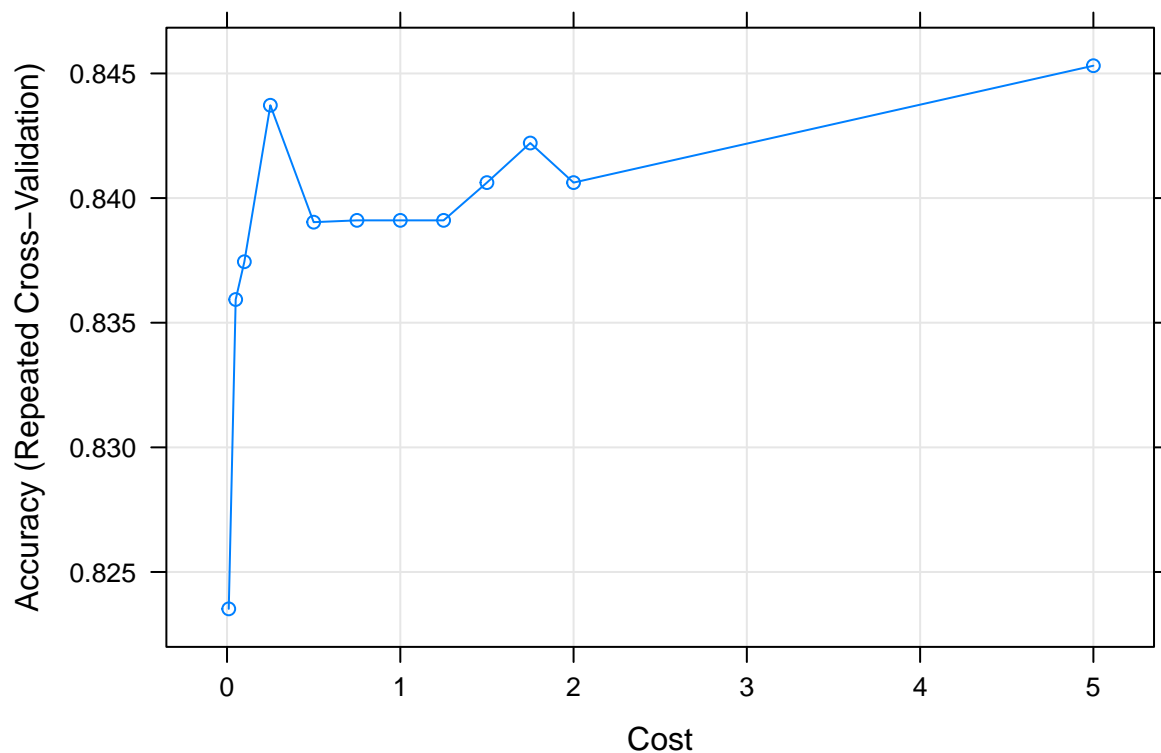
svm.mod.grid <- train(target ~ ., data = training, method = "svmLinear",
                      trControl = trctrl,
                      preProcess = c("center", "scale"),
                      tuneGrid = grid,
                      tuneLength = 10)

svm.mod.grid

## Support Vector Machines with Linear Kernel
##
## 213 samples
## 13 predictor
## 2 classes: '0', '1'
##
## Pre-processing: centered (13), scaled (13)
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 192, 191, 192, 192, 192, 192, ...
## Resampling results across tuning parameters:
##
##  C      Accuracy   Kappa
##  0.00      NaN      NaN
##  0.01  0.8235209  0.6414705
##  0.05  0.8359307  0.6677455
##  0.10  0.8374459  0.6712576
##  0.25  0.8437229  0.6843572
##  0.50  0.8390332  0.6749413
##  0.75  0.8391053  0.6752263
##  1.00  0.8391053  0.6753687
##  1.25  0.8391053  0.6753687
##  1.50  0.8406205  0.6783406
##  1.75  0.8422078  0.6815951
##  2.00  0.8406205  0.6785395
##  5.00  0.8453102  0.6879521
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was C = 5.

# accuracy plot of tuned model
plot(svm.mod.grid)

```



```
# prediction using tuned model
svm.pred.grid <- predict(svm.mod.grid, newdata = testing)
svm.pred.grid

## [1] 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 0 0 1 0 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1
## [39] 1 1 1 1 1 1 1 0 0 1 1 1 1 0 1 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
## [77] 0 0 0 1 1 1 0 0 1 0 1 0 0 0
## Levels: 0 1

# accuracy of the tuned model
confusionMatrix(table(svm.pred.grid, testing$target))

## Confusion Matrix and Statistics
##
##
## svm.pred.grid  0  1
##               0 28 10
##               1  9 43
##
##               Accuracy : 0.7889
##               95% CI : (0.6901, 0.8679)
##               No Information Rate : 0.5889
##               P-Value [Acc > NIR] : 4.918e-05
##
##               Kappa : 0.5658
##
## Mcnemar's Test P-Value : 1
##
##               Sensitivity : 0.7568
##               Specificity : 0.8113
##               Pos Pred Value : 0.7368
```

```
##          Neg Pred Value : 0.8269
##          Prevalence : 0.4111
##          Detection Rate : 0.3111
##    Detection Prevalence : 0.4222
##          Balanced Accuracy : 0.7840
##
##          'Positive' Class : 0
##
```

Logistic Regression

```
set.seed(2441139)
```

```
# pre-processing of data
str(heart)
```

```
## 'data.frame':  303 obs. of  14 variables:
## $ age      : int  63 37 41 56 57 57 56 44 52 57 ...
## $ sex      : int  1 1 0 1 0 1 0 1 1 1 ...
## $ cp       : int  3 2 1 1 0 0 1 1 2 2 ...
## $ trestbps : int  145 130 130 120 120 140 140 120 172 150 ...
## $ chol     : int  233 250 204 236 354 192 294 263 199 168 ...
## $ fbs      : int  1 0 0 0 0 0 0 0 1 0 ...
## $ restecg  : int  0 1 0 1 1 1 0 1 1 1 ...
## $ thalach  : int  150 187 172 178 163 148 153 173 162 174 ...
## $ exang    : int  0 0 0 0 1 0 0 0 0 0 ...
## $ oldpeak  : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
## $ slope    : int  0 0 2 2 2 1 1 2 2 2 ...
## $ ca       : int  0 0 0 0 0 0 0 0 0 0 ...
## $ thal     : int  1 2 2 2 2 1 2 3 3 2 ...
## $ target   : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
heart$sex <- ifelse(test = heart$sex == 0, yes = "F", no = "M")
heart$sex <- as.factor(heart$sex)
heart$cp <- as.factor(heart$cp)
heart$fbs <- as.factor(heart$fbs)
heart$restecg <- as.factor(heart$restecg)
heart$exang <- as.factor(heart$exang)
heart$slope <- as.factor(heart$slope)
heart$ca <- as.factor(heart$ca)
heart$thal <- as.factor(heart$thal)
heart$age <- as.numeric(heart$age)
heart$trestbps <- as.numeric(heart$trestbps)
heart$chol <- as.numeric(heart$chol)
heart$thalach <- as.numeric(heart$thalach)
heart$target <- ifelse(test = heart$target == 0, yes = "Healthy", no = "Unhealthy")
heart$target <- as.factor(heart$target)
```

```
str(heart)
```

```
## 'data.frame':  303 obs. of  14 variables:
## $ age      : num  63 37 41 56 57 57 56 44 52 57 ...
## $ sex      : Factor w/ 2 levels "F","M": 2 2 1 2 1 2 1 2 2 2 ...
## $ cp       : Factor w/ 4 levels "0","1","2","3": 4 3 2 2 1 1 2 2 3 3 ...
## $ trestbps : num  145 130 130 120 120 140 140 120 172 150 ...
```

```
## $ chol      : num  233 250 204 236 354 192 294 263 199 168 ...
## $ fbs       : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 2 1 ...
## $ restecg   : Factor w/ 3 levels "0","1","2": 1 2 1 2 2 2 1 2 2 2 ...
## $ thalach   : num  150 187 172 178 163 148 153 173 162 174 ...
## $ exang     : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 1 ...
## $ oldpeak   : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
## $ slope     : Factor w/ 3 levels "0","1","2": 1 1 3 3 3 2 2 3 3 3 ...
## $ ca        : Factor w/ 5 levels "0","1","2","3",...: 1 1 1 1 1 1 1 1 1 1 ...
## $ thal      : Factor w/ 4 levels "0","1","2","3": 2 3 3 3 3 2 3 4 4 3 ...
## $ target    : Factor w/ 2 levels "Healthy","Unhealthy": 2 2 2 2 2 2 2 2 2 2 ...
```

```
# getting the number of samples based on gender
xtabs(~ target + sex, data = heart)
```

```
##           sex
## target      F   M
## Healthy    24 114
## Unhealthy   72  93
```

```
# simple logistic model
logistic <- glm(target ~ sex, data = heart, family = "binomial")
summary(logistic)
```

```
##
## Call:
## glm(formula = target ~ sex, family = "binomial", data = heart)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.6651  -1.0923   0.7585   1.2650   1.2650
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.0986     0.2357   4.661 3.15e-06 ***
## sexM          -1.3022     0.2740  -4.752 2.01e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 392.80  on 301  degrees of freedom
## AIC: 396.8
##
## Number of Fisher Scoring iterations: 4
```

```
R_sq_1 <- 1 - logistic$deviance / logistic$null.deviance
R_sq_1
```

```
## [1] 0.05947945
```

```
BIC_1 <- logistic$deviance + 2 * log(dim(heart)[1])
BIC_1
```

```
## [1] 404.2246
```

```

# complex logistic model
logistic <- glm(target ~ ., data = heart, family = "binomial")
# summary(logistic)

R_sq_2 <- 1 - logistic$deviance / logistic$null.deviance
R_sq_2

## [1] 0.569889

BIC_2 <- logistic$deviance + 14 * log(dim(heart)[1])
BIC_2

## [1] 259.623

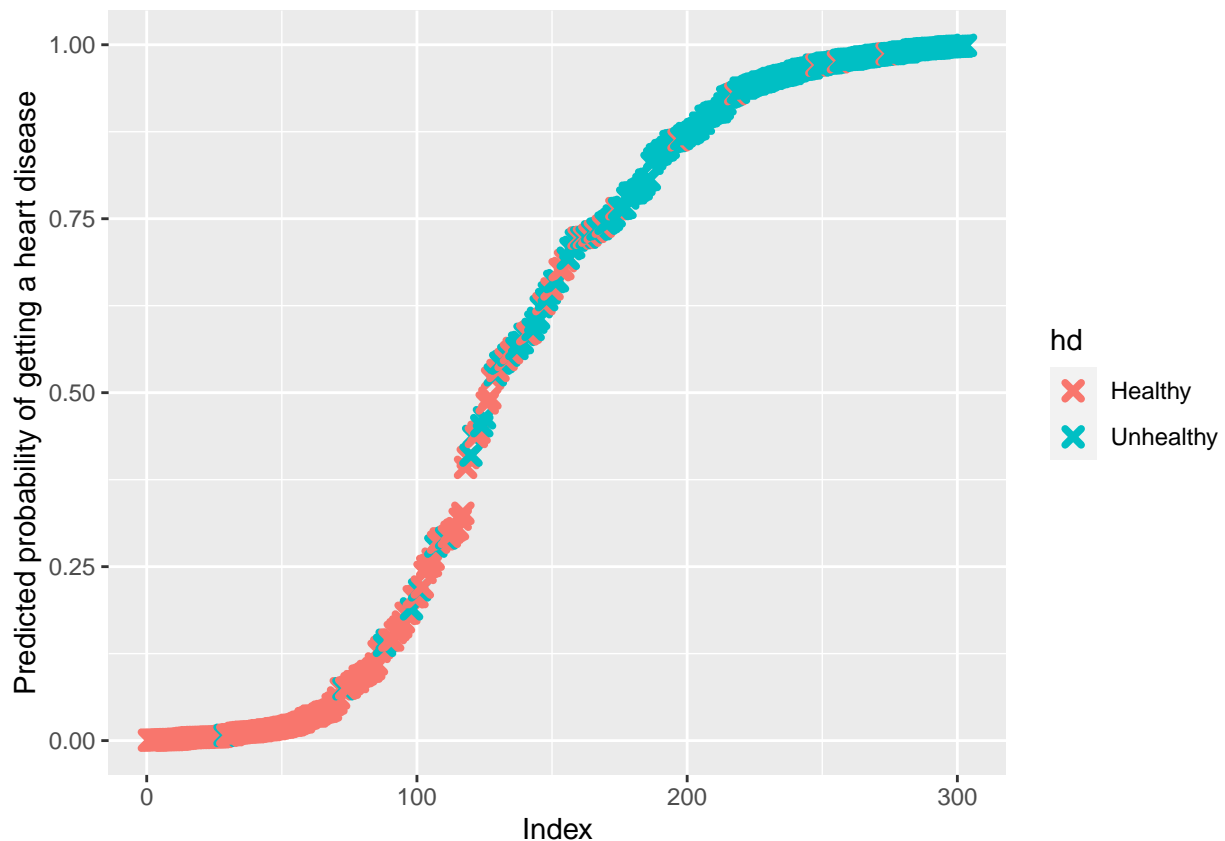
# why age isn't of statistical significance
median(heart$age)

## [1] 55

# plotting the probability of getting a heart disease
predict.hd <- data.frame(prob.of.hd = logistic$fitted.values, hd = heart$target)
predict.hd <- predict.hd[order(predict.hd$prob.of.hd, decreasing = FALSE), ]
predict.hd$rank <- 1:nrow(predict.hd)

ggplot(data = predict.hd, aes(x = rank, y = prob.of.hd)) +
  geom_point(aes(color = hd), alpha = 1, shape = 4, stroke = 2) +
  xlab("Index") +
  ylab("Predicted probability of getting a heart disease")

```



```

# predicting model accuracy
log.mod <- glm(target ~ ., data = training, family = "binomial")
log.pred <- predict(log.mod, newdata = testing, type = "response")
log.pred <- ifelse(log.pred > 0.5, 1, 0)
confusionMatrix(table(log.pred, testing$target))

```

```

## Confusion Matrix and Statistics
##
##
## log.pred  0  1
##           0 28 11
##           1  9 42
##
##               Accuracy : 0.7778
##               95% CI : (0.6779, 0.8587)
##      No Information Rate : 0.5889
##      P-Value [Acc > NIR] : 0.0001266
##
##               Kappa : 0.5448
##
##  Mcnemar's Test P-Value : 0.8230633
##
##      Sensitivity : 0.7568
##      Specificity : 0.7925
##      Pos Pred Value : 0.7179
##      Neg Pred Value : 0.8235
##      Prevalence : 0.4111
##      Detection Rate : 0.3111
##      Detection Prevalence : 0.4333
##      Balanced Accuracy : 0.7746
##
##      'Positive' Class : 0
##

```