

Machine learning models for cancer predictive analysis

Natalia

28 May 2019

```
data <- read.csv("C://Users//Natalia//Desktop//ITMO//R//R project//cancer data//breast cancer coimbra//  
View(data)
```

Analyse the dataset and tidy it up.

```
# Analyse the data - checking for values, NAs, data type.  
summary(data)
```

```
##      Age      BMI      Glucose      Insulin  
## Min.   :24.0   Min.   :18.37   Min.   : 60.00   Min.   : 2.432  
## 1st Qu.:45.0   1st Qu.:22.97   1st Qu.: 85.75   1st Qu.: 4.359  
## Median :56.0   Median :27.66   Median : 92.00   Median : 5.925  
## Mean   :57.3   Mean   :27.58   Mean   : 97.79   Mean   :10.012  
## 3rd Qu.:71.0   3rd Qu.:31.24   3rd Qu.:102.00   3rd Qu.:11.189  
## Max.   :89.0   Max.   :38.58   Max.   :201.00   Max.   :58.460  
##      HOMA      Leptin      Adiponectin      Resistin  
## Min.   : 0.4674   Min.   : 4.311   Min.   : 1.656   Min.   : 3.210  
## 1st Qu.: 0.9180   1st Qu.:12.314   1st Qu.: 5.474   1st Qu.: 6.882  
## Median : 1.3809   Median :20.271   Median : 8.353   Median :10.828  
## Mean   : 2.6950   Mean   :26.615   Mean   :10.181   Mean   :14.726  
## 3rd Qu.: 2.8578   3rd Qu.:37.378   3rd Qu.:11.816   3rd Qu.:17.755  
## Max.   :25.0503   Max.   :90.280   Max.   :38.040   Max.   :82.100  
##      MCP.1      Classification  
## Min.   : 45.84   Min.   :1.000  
## 1st Qu.:269.98   1st Qu.:1.000  
## Median :471.32   Median :2.000  
## Mean   :534.65   Mean   :1.552  
## 3rd Qu.:700.09   3rd Qu.:2.000  
## Max.   :1698.44   Max.   :2.000
```

```
str(data)
```

```
## 'data.frame': 116 obs. of 10 variables:  
## $ Age : int 48 83 82 68 86 49 89 76 73 75 ...  
## $ BMI : num 23.5 20.7 23.1 21.4 21.1 ...  
## $ Glucose : int 70 92 91 77 92 92 77 118 97 83 ...  
## $ Insulin : num 2.71 3.12 4.5 3.23 3.55 ...  
## $ HOMA : num 0.467 0.707 1.01 0.613 0.805 ...  
## $ Leptin : num 8.81 8.84 17.94 9.88 6.7 ...  
## $ Adiponectin : num 9.7 5.43 22.43 7.17 4.82 ...  
## $ Resistin : num 8 4.06 9.28 12.77 10.58 ...  
## $ MCP.1 : num 417 469 555 928 774 ...  
## $ Classification: int 1 1 1 1 1 1 1 1 1 1 ...
```

```
head(data)
```

```
##      Age      BMI Glucose Insulin      HOMA      Leptin Adiponectin Resistin  
## 1  48 23.50000      70    2.707 0.4674087    8.8071    9.702400    7.99585
```

```
## 2 83 20.69049 92 3.115 0.7068973 8.8438 5.429285 4.06405
## 3 82 23.12467 91 4.498 1.0096511 17.9393 22.432040 9.27715
## 4 68 21.36752 77 3.226 0.6127249 9.8827 7.169560 12.76600
## 5 86 21.11111 92 3.549 0.8053864 6.6994 4.819240 10.57635
## 6 49 22.85446 92 3.226 0.7320869 6.8317 13.679750 10.31760
```

```
## MCP.1 Classification
```

```
## 1 417.114 1
## 2 468.786 1
## 3 554.697 1
## 4 928.220 1
## 5 773.920 1
## 6 530.410 1
```

```
dim(data)
```

```
## [1] 116 10
```

```
library(tidyverse)
```

```
## -- Attaching packages ----- tidyverse 1.2
```

```
## 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
```

```
## -- Conflicts ----- tidyverse_conflicts
```

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

```
map_int(data, function(.x) sum(is.na(.x)))
```

```
##      Age      BMI      Glucose      Insulin      HOMA
##      0        0        0        0        0
##      Leptin Adiponectin Resistin MCP.1 Classification
##      0        0        0        0        0
```

```
library(plyr)
```

```
## -----
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## -----
```

```
##
```

```
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
```

```
##
```

```
##      arrange, count, desc, failwith, id, mutate, rename, summarise,
##      summarize
```

```
## The following object is masked from 'package:purrr':
```

```
##
```

```
##      compact
```

```
#change value names in "Classifications" for "healthy" and "cancer":
```

```
data$Classification <- factor(as.character(data$Classification))
data$Classification <- revalue(data$Classification, c("1"="healthy"))
data$Classification <- revalue(data$Classification, c("2"="cancer"))
head(data)
```

```
##      Age      BMI Glucose Insulin      HOMA  Leptin Adiponectin Resistin
## 1  48 23.50000      70   2.707 0.4674087  8.8071   9.702400  7.99585
## 2  83 20.69049      92   3.115 0.7068973  8.8438   5.429285  4.06405
## 3  82 23.12467      91   4.498 1.0096511 17.9393  22.432040  9.27715
## 4  68 21.36752      77   3.226 0.6127249  9.8827   7.169560 12.76600
## 5  86 21.11111      92   3.549 0.8053864  6.6994   4.819240 10.57635
## 6  49 22.85446      92   3.226 0.7320869  6.8317  13.679750 10.31760
##      MCP.1 Classification
## 1 417.114      healthy
## 2 468.786      healthy
## 3 554.697      healthy
## 4 928.220      healthy
## 5 773.920      healthy
## 6 530.410      healthy
```

```
data <- as.data.frame(data, stringsAsFactors=T)
data$Classification <- factor(as.character(data$Classification))
supply(data,mode)
```

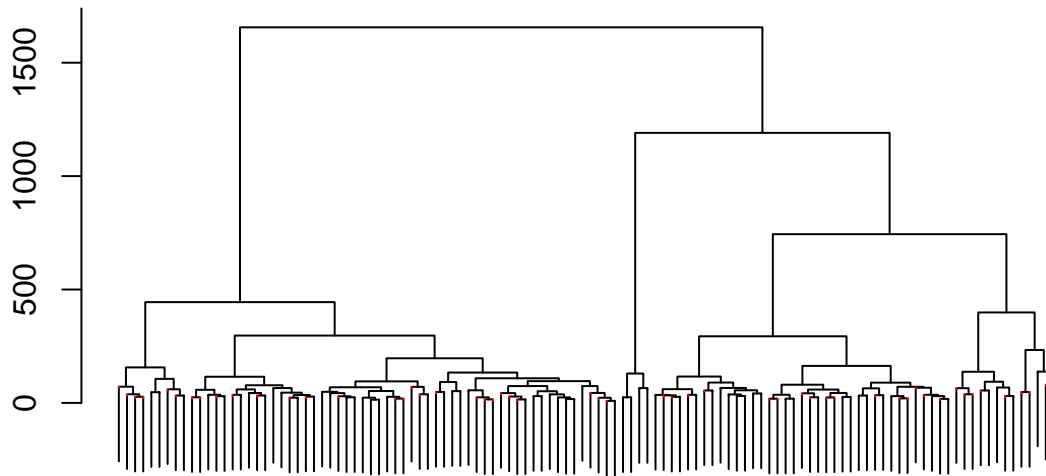
```
##      Age      BMI      Glucose      Insulin      HOMA
##      "numeric"      "numeric"      "numeric"      "numeric"      "numeric"
##      Leptin  Adiponectin  Resistin      MCP.1 Classification
##      "numeric"      "numeric"      "numeric"      "numeric"      "numeric"
```

DATA EXPLORATION

Hierarchical clustering

```
library(sparcl)
hc <- hclust(dist(data[,1:9]), method = "complete")
ColorDendrogram(hc,y=factor(data$Classification), main = "Hierarchical clustering", branchlength=5)
```

Hierarchical clustering



```
dist(data[, 1:9])  
hclust (*, "complete")
```

Most of the benign (black) and malignant (red) samples cluster together.

K-means clustering

```
fit <- kmeans(data[,c(1:9)], 2)  
names(fit)
```

```
## [1] "cluster"      "centers"      "totss"        "withinss"  
## [5] "tot.withinss" "betweenss"    "size"         "iter"  
## [9] "ifault"
```

#k-means did a fairly good job

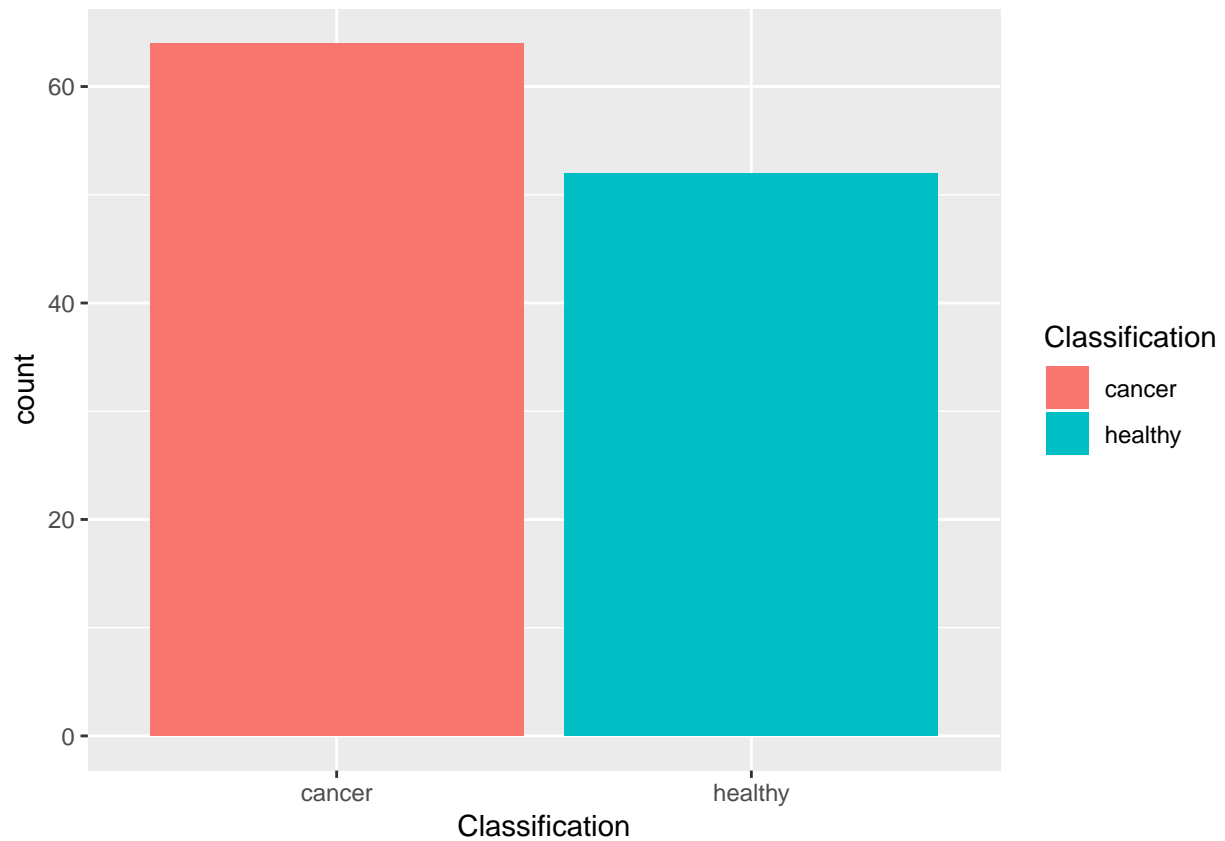
```
table(data.frame(fit$cluster, data[,10]))
```

```
##          data...10.  
## fit.cluster cancer healthy  
##          1      19      12  
##          2      45      40
```

Response variable for classification

```
library(ggplot2)
```

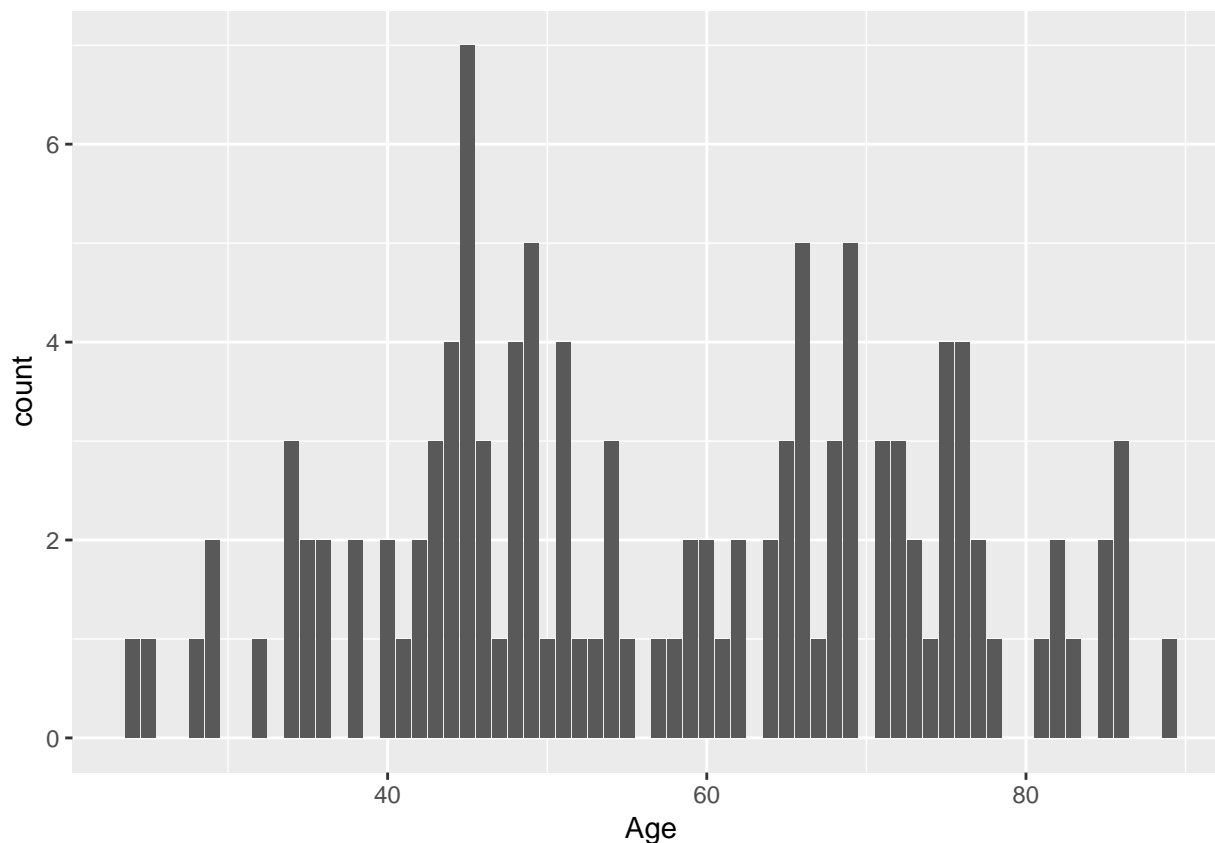
```
ggplot(data, aes(x = Classification, fill = Classification)) +  
  geom_bar()
```



Response variable for regression

```
ggplot(data, aes(x = Age)) +  
  geom_histogram(stat = "count")
```

Warning: Ignoring unknown parameters: binwidth, bins, pad



Principal Component Analysis

```
library(pcaGoPromoter)
```

```
## Loading required package: ellipse
##
## Attaching package: 'ellipse'
## The following object is masked from 'package:graphics':
##
##     pairs
## Loading required package: Biostrings
## Loading required package: BiocGenerics
## Loading required package: parallel
##
## Attaching package: 'BiocGenerics'
## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB
## The following objects are masked from 'package:dplyr':
```

```

##
##   combine, intersect, setdiff, union
## The following objects are masked from 'package:stats':
##
##   IQR, mad, sd, var, xtabs
## The following objects are masked from 'package:base':
##
##   anyDuplicated, append, as.data.frame, basename, cbind,
##   colMeans, colnames, colSums, dirname, do.call, duplicated,
##   eval, evalq, Filter, Find, get, grep, grepl, intersect,
##   is.unsorted, lapply, lengths, Map, mapply, match, mget, order,
##   paste, pmax, pmax.int, pmin, pmin.int, Position, rank, rbind,
##   Reduce, rowMeans, rownames, rowSums, sapply, setdiff, sort,
##   table, tapply, union, unique, unsplit, which, which.max,
##   which.min
## Loading required package: S4Vectors
## Loading required package: stats4
##
## Attaching package: 'S4Vectors'
## The following object is masked from 'package:plyr':
##
##   rename
## The following objects are masked from 'package:dplyr':
##
##   first, rename
## The following object is masked from 'package:tidyr':
##
##   expand
## The following object is masked from 'package:base':
##
##   expand.grid
## Loading required package: IRanges
##
## Attaching package: 'IRanges'
## The following object is masked from 'package:plyr':
##
##   desc
## The following objects are masked from 'package:dplyr':
##
##   collapse, desc, slice
## The following object is masked from 'package:purrr':
##
##   reduce
## The following object is masked from 'package:grDevices':
##
##   windows

```

```

## Loading required package: XVector
##
## Attaching package: 'XVector'
## The following object is masked from 'package:plyr':
##
##     compact
## The following object is masked from 'package:purrr':
##
##     compact
##
## Attaching package: 'Biostrings'
## The following object is masked from 'package:base':
##
##     strsplit
library(ellipse)

data <- na.omit(data)

# perform pca and extract scores
pcaOutput <- pca(t(data[, 1:9]), printDropped = FALSE, scale = TRUE, center = TRUE)
pcaOutput2 <- as.data.frame(pcaOutput$scores)

# define groups for plotting:

pcaOutput2$groups <- data$Classification

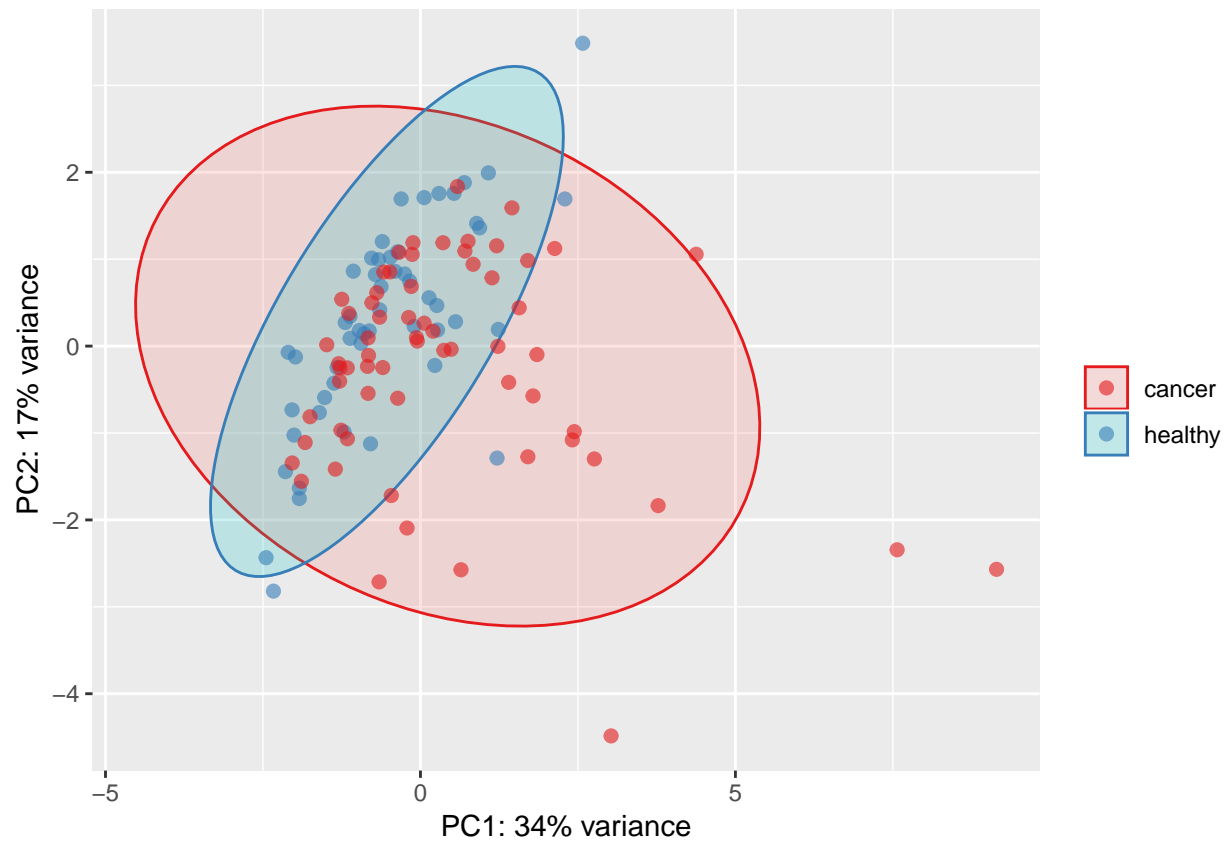
centroids <- aggregate(cbind(PC1, PC2) ~ groups, pcaOutput2, mean)

conf.rgn <- do.call(rbind, lapply(unique(pcaOutput2$groups), function(t)
  data.frame(groups = as.character(t),
    ellipse(cov(pcaOutput2[pcaOutput2$groups == t, 1:2]),
      centre = as.matrix(centroids[centroids$groups == t, 2:3]),
      level = 0.95),
    stringsAsFactors = FALSE)))

#Plot PCA with variance %:

ggplot(data = pcaOutput2, aes(x = PC1, y = PC2, group = groups, color = groups)) +
  geom_polygon(data = conf.rgn, aes(fill = groups), alpha = 0.2) +
  geom_point(size = 2, alpha = 0.6) +
  scale_color_brewer(palette = "Set1") +
  labs(color = "",
    fill = "",
    x = paste0("PC1: ", round(pcaOutput$pov[1], digits = 2) * 100, "% variance"),
    y = paste0("PC2: ", round(pcaOutput$pov[2], digits = 2) * 100, "% variance"))

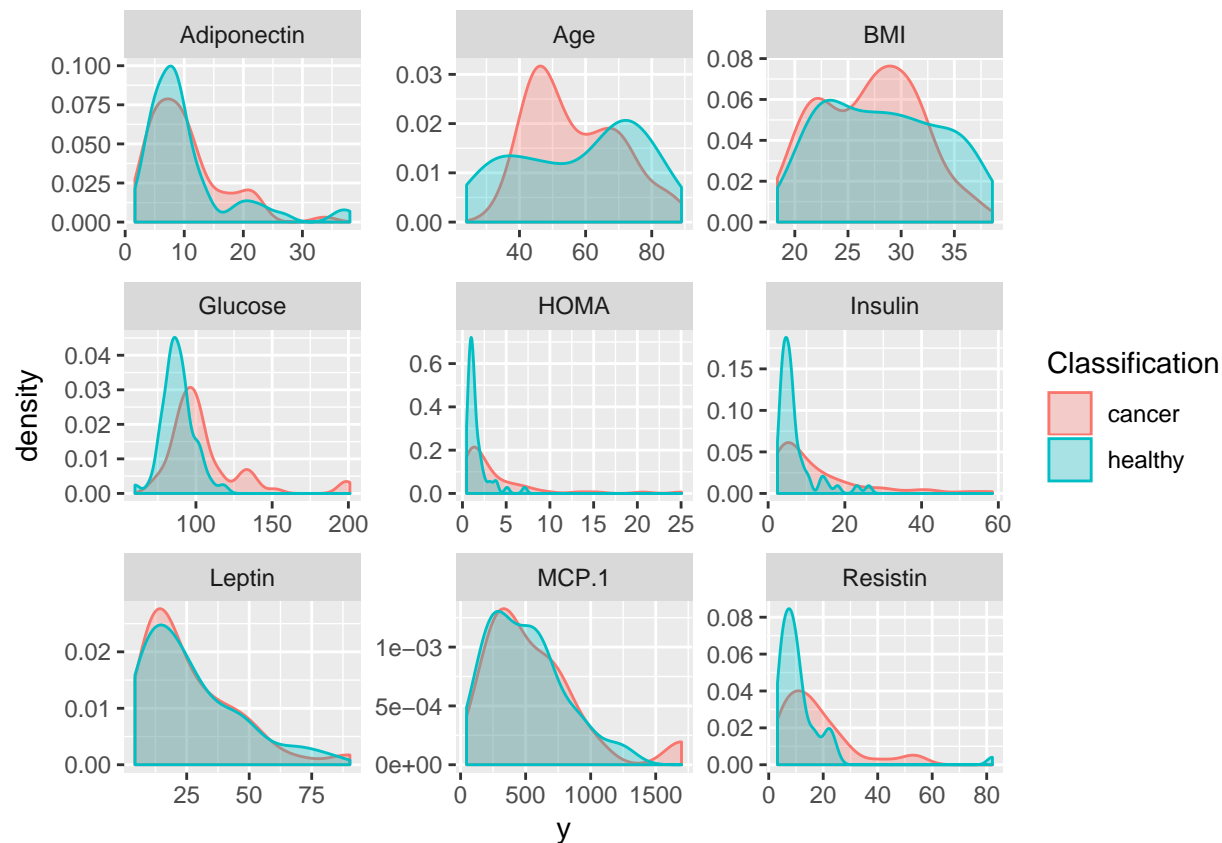
```

Features

```
library(tidyr)

gather(data, x, y, Age:MCP.1) %>%
  ggplot(aes(x = y, color = Classification, fill = Classification)) +
  geom_density(alpha = 0.3) +
  facet_wrap(~ x, scales = "free")
```



MACHINE LEARNING PACKAGES FOR R

caret

```
# configure multicore
library(doParallel)

## Loading required package: foreach
##
## Attaching package: 'foreach'
## The following objects are masked from 'package:purrr':
##
##   accumulate, when
## Loading required package: iterators
cl <- makeCluster(detectCores())
registerDoParallel(cl)

library(caret)

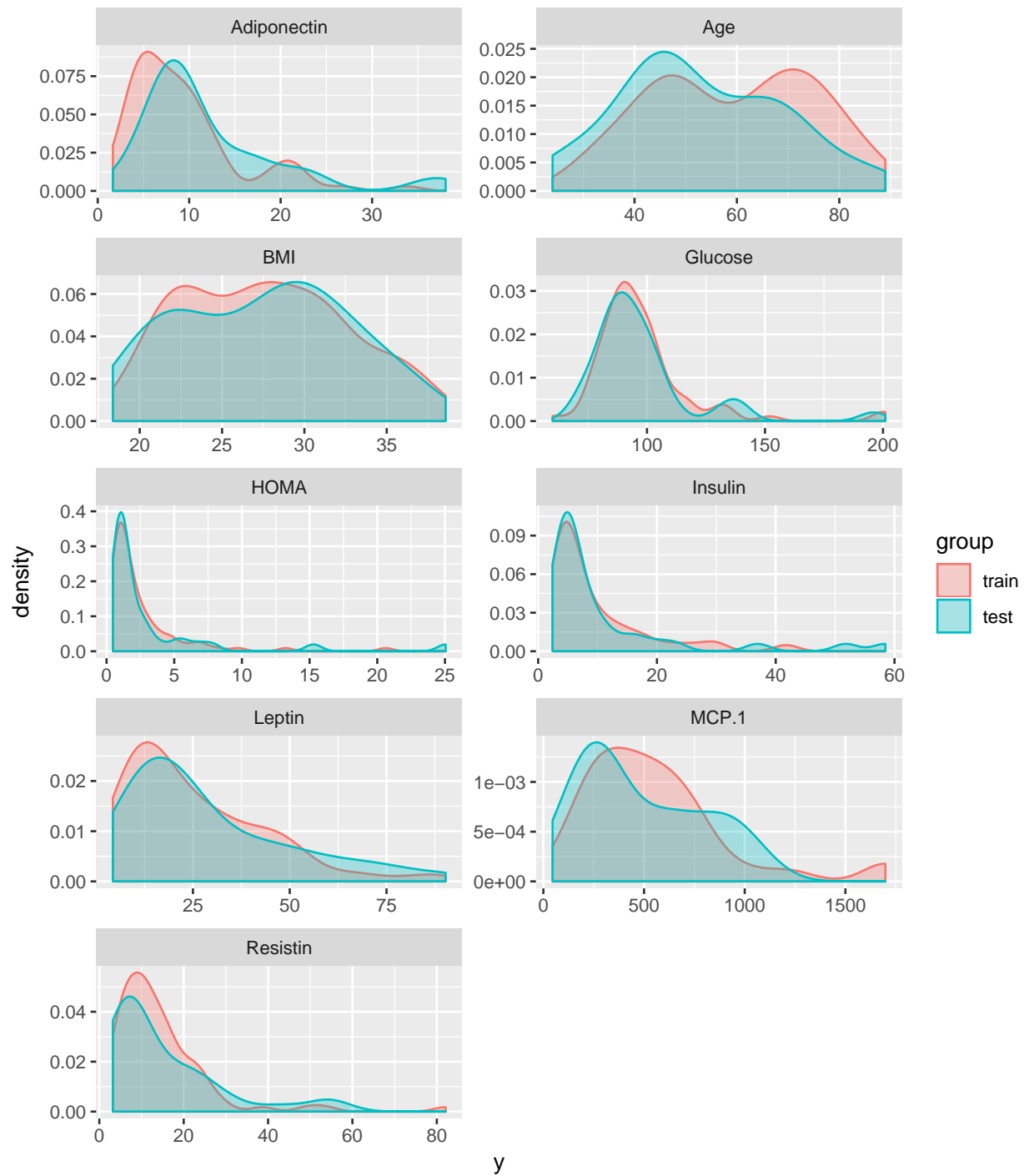
## Loading required package: lattice
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':  
##  
## lift
```

Training, validation and test data

```
set.seed(42)  
index <- createDataPartition(data$Classification, p = 0.7, list = FALSE)  
train_data <- data[index, ]  
test_data <- data[-index, ]
```

```
library(dplyr)  
  
rbind(data.frame(group = "train", train_data),  
      data.frame(group = "test", test_data)) %>%  
  gather(x, y, Age:MCP.1) %>%  
  ggplot(aes(x = y, color = group, fill = group)) +  
    geom_density(alpha = 0.3) +  
    facet_wrap(~ x, scales = "free", ncol = 2)
```



REGRESSION

```
set.seed(42)
model_glm <- caret::train(Age ~ .,
  data = train_data,
  method = "glm",
  preProcess = c("scale", "center"),
```

```

trControl = trainControl(method = "repeatedcv",
                           number = 10,
                           repeats = 10,
                           savePredictions = TRUE,
                           verboseIter = FALSE))

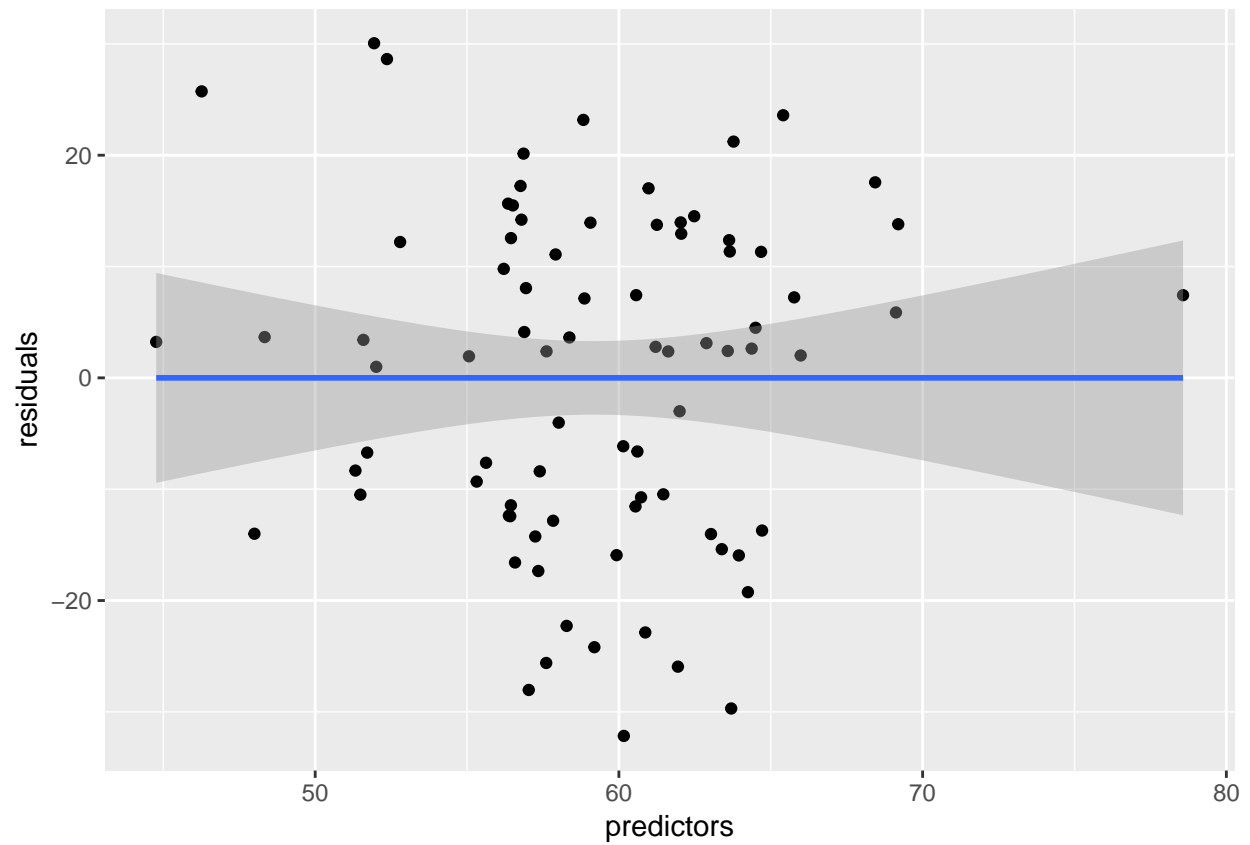
model_glm

## Generalized Linear Model
##
## 82 samples
## 9 predictor
##
## Pre-processing: scaled (9), centered (9)
## Resampling: Cross-Validated (10 fold, repeated 10 times)
## Summary of sample sizes: 74, 74, 74, 72, 74, 74, ...
## Resampling results:
##
##    RMSE      Rsquared    MAE
## 17.63297 0.1196448 14.88396

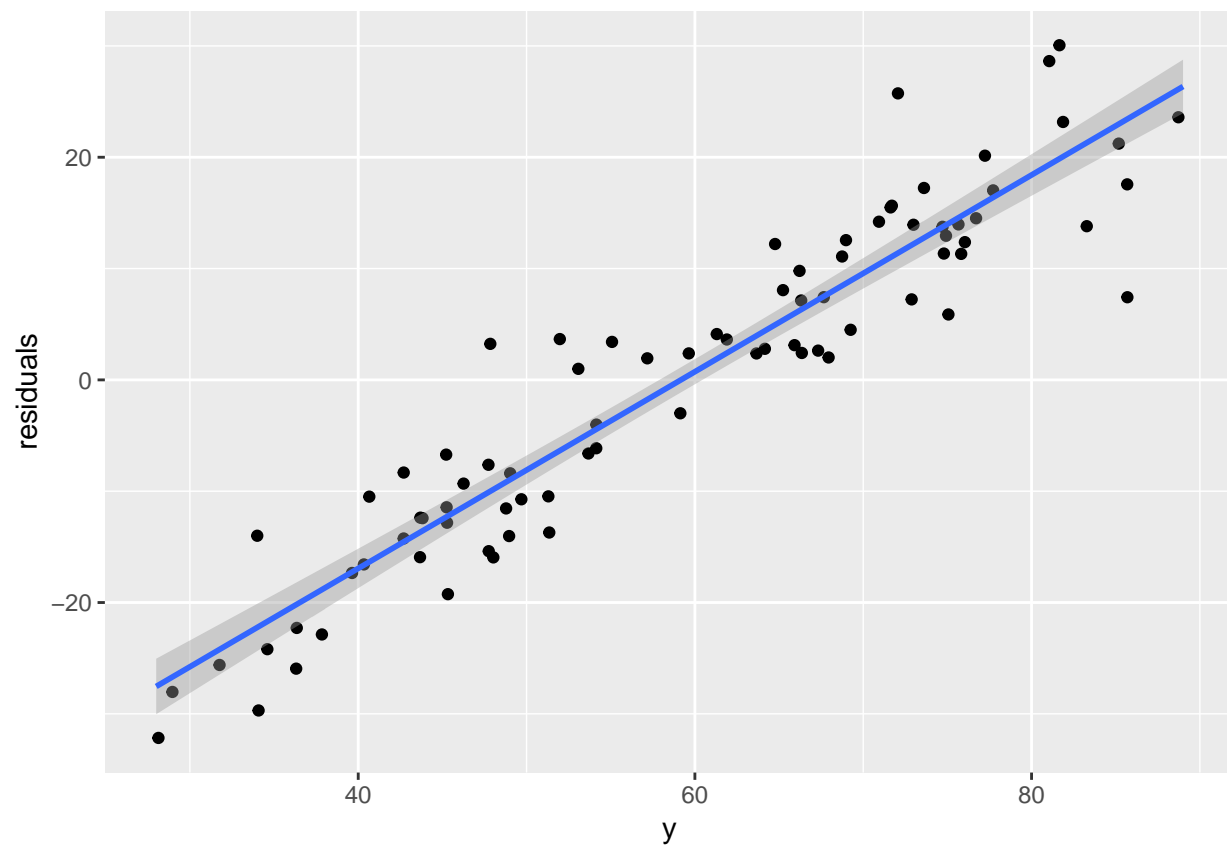
predictions <- predict(model_glm, test_data)

# model_glm$finalModel$linear.predictors == model_glm$finalModel$fitted.values
data.frame(residuals = resid(model_glm),
            predictors = model_glm$finalModel$linear.predictors) %>%
  ggplot(aes(x = predictors, y = residuals)) +
  geom_jitter() +
  geom_smooth(method = "lm")

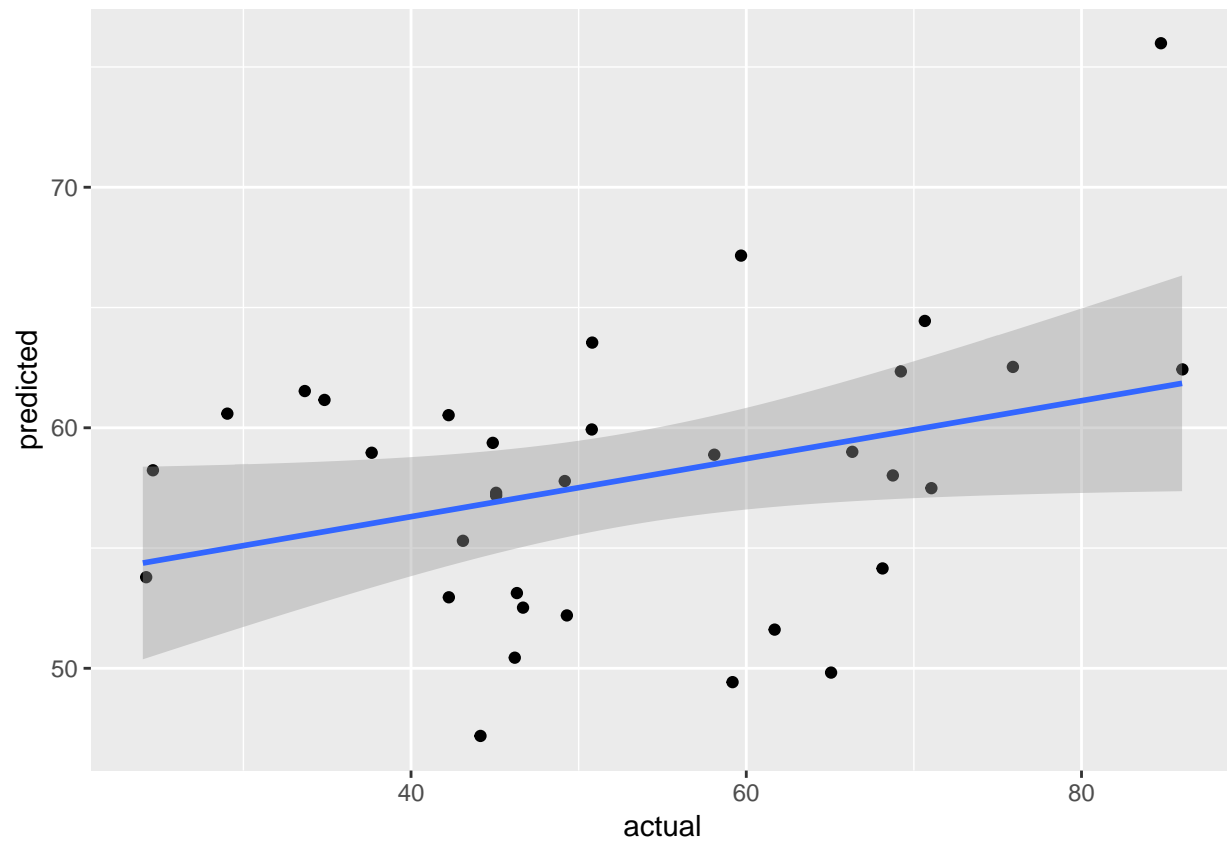
```



```
# y == train_data$Age
data.frame(residuals = resid(model_glm),
           y = model_glm$finalModel$y) %>%
  ggplot(aes(x = y, y = residuals)) +
    geom_jitter() +
    geom_smooth(method = "lm")
```



```
data.frame(actual = test_data$Age,  
            predicted = predictions) %>%  
ggplot(aes(x = actual, y = predicted)) +  
  geom_jitter() +  
  geom_smooth(method = "lm")
```



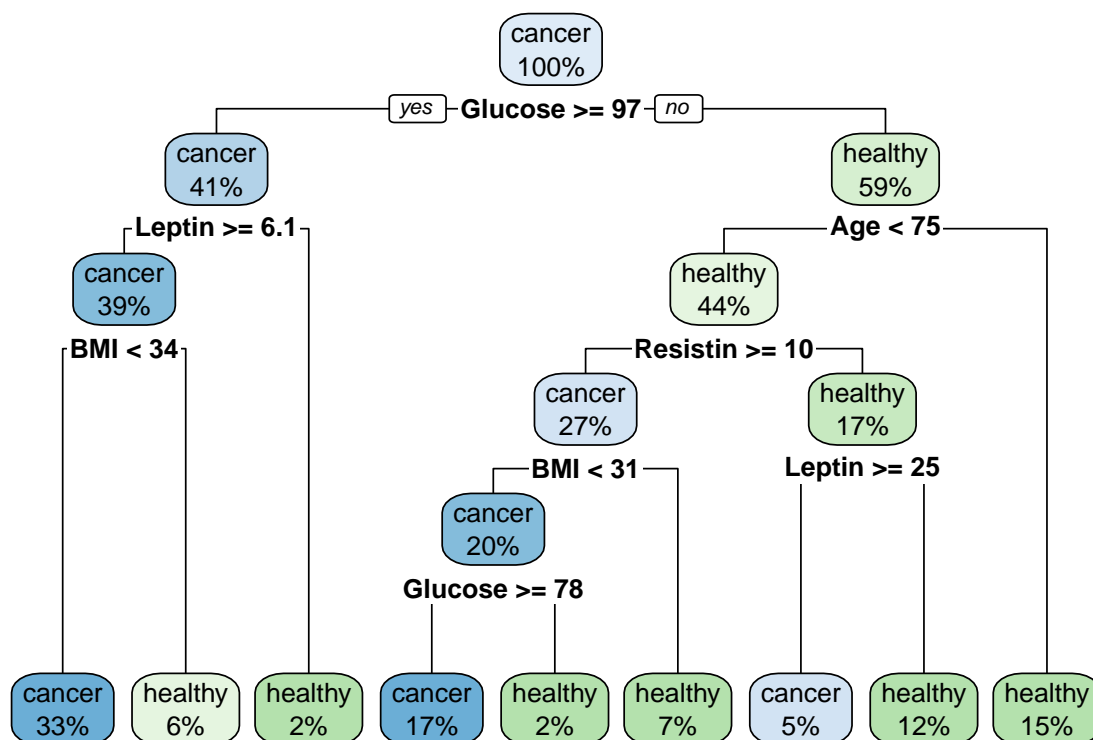
CLASSIFICATION

Decision trees

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

set.seed(42)
fit <- rpart(Classification ~ .,
             data = train_data,
             method = "class",
             control = rpart.control(xval = 10,
                                     minbucket = 2,
                                     cp = 0),
             parms = list(split = "information"))

rpart.plot(fit, extra = 100)
```

Random Forests

*#Random Forests predictions are based on the generation of
#multiple classification trees.
#They can be used for both, classification and regression tasks.
#Here, it is classification task.*

```

set.seed(42)
library(randomForest)

## randomForest 4.6-14
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
##
## The following object is masked from 'package:BiocGenerics':
##
##   combine
##
## The following object is masked from 'package:dplyr':
##
##   combine
##
## The following object is masked from 'package:ggplot2':
##
##   margin
  
```

```
model_rf <- caret::train(Classification ~ .,
  data = train_data,
  method = "rf",
  preProcess = c("scale", "center"),
  trControl = trainControl(method = "repeatedcv",
    number = 10,
    repeats = 10,
    savePredictions = TRUE,
    verboseIter = FALSE))
```

*#When savePredictions = TRUE is specified,
#can access the cross-validation results with model_rf\$pred.*

```
model_rf$finalModel$confusion
```

```
##          cancer healthy class.error
## cancer      35      10  0.2222222
## healthy     10      27  0.2702703
```

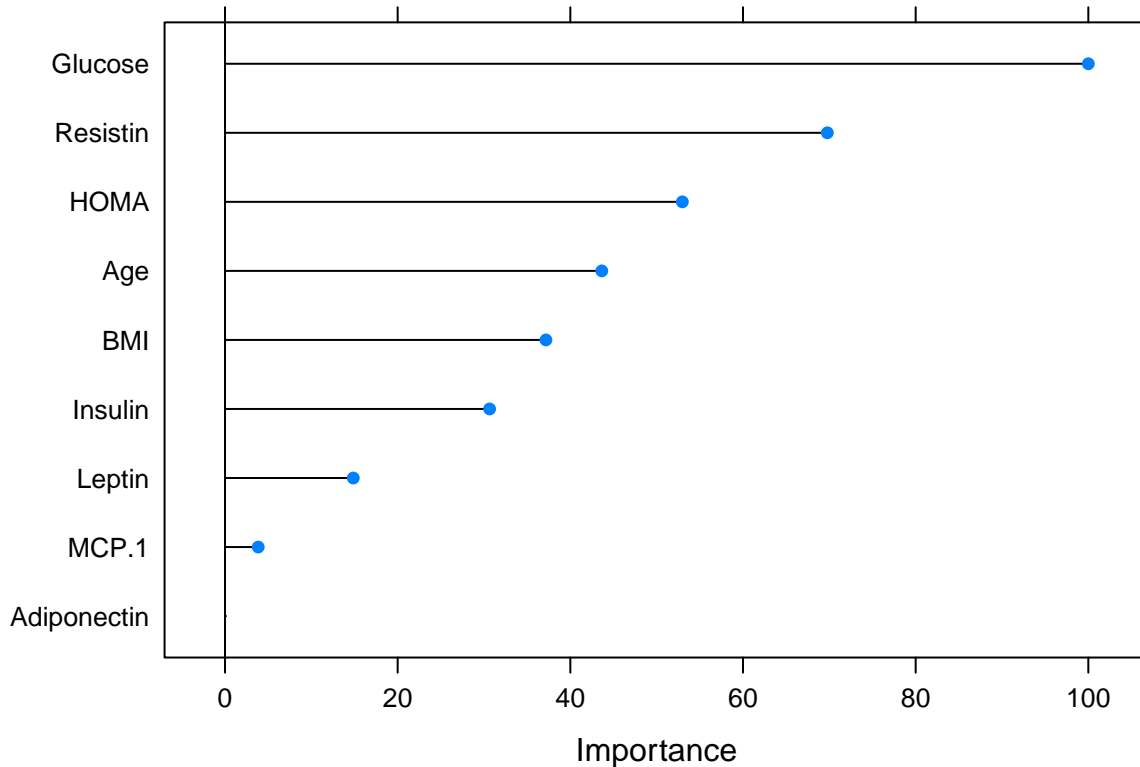
Feature Importance

```
imp <- model_rf$finalModel$importance
imp[order(imp, decreasing = TRUE), ]
```

```
##      Glucose      Resistin      HOMA      Age      BMI      Insulin
##      7.011906      5.740823      5.034404      4.642025      4.369878      4.095543
##      Leptin      MCP.1 Adiponectin
##      3.431462      2.968458      2.806573
```

estimate variable importance

```
importance <- varImp(model_rf, scale = TRUE)
plot(importance)
```



Predicting test data

```
confusionMatrix(predict(model_rf, test_data), test_data$Classification)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction cancer healthy
##   cancer      14      5
##   healthy      5     10
##
##           Accuracy : 0.7059
##           95% CI : (0.5252, 0.849)
##   No Information Rate : 0.5588
##   P-Value [Acc > NIR] : 0.0582
##
##           Kappa : 0.4035
##
## Mcnemar's Test P-Value : 1.0000
##
##           Sensitivity : 0.7368
##           Specificity : 0.6667
##   Pos Pred Value : 0.7368
##   Neg Pred Value : 0.6667
##           Prevalence : 0.5588
```

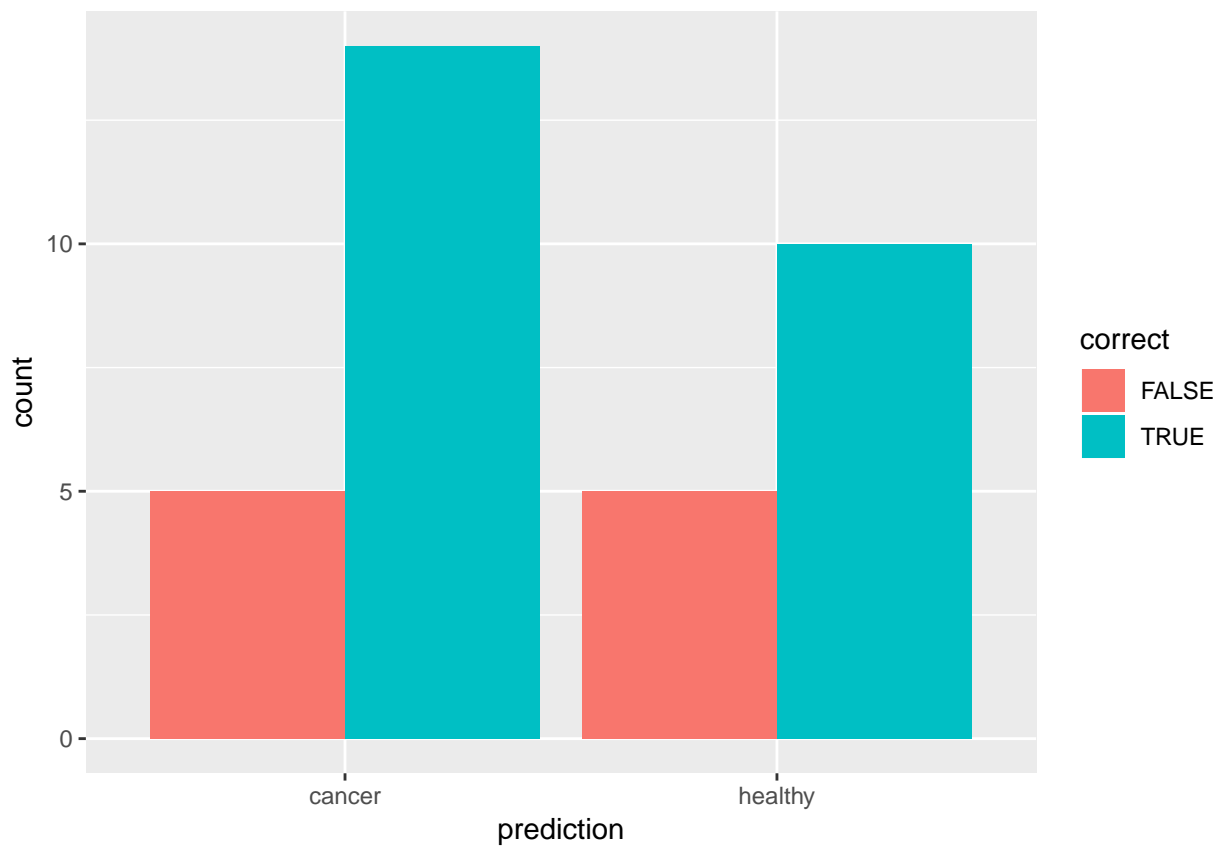
```
##      Detection Rate : 0.4118
##      Detection Prevalence : 0.5588
##      Balanced Accuracy : 0.7018
##
##      'Positive' Class : cancer
##
```

```
results <- data.frame(actual = test_data$Classification,
                      predict(model_rf, test_data, type = "prob"))

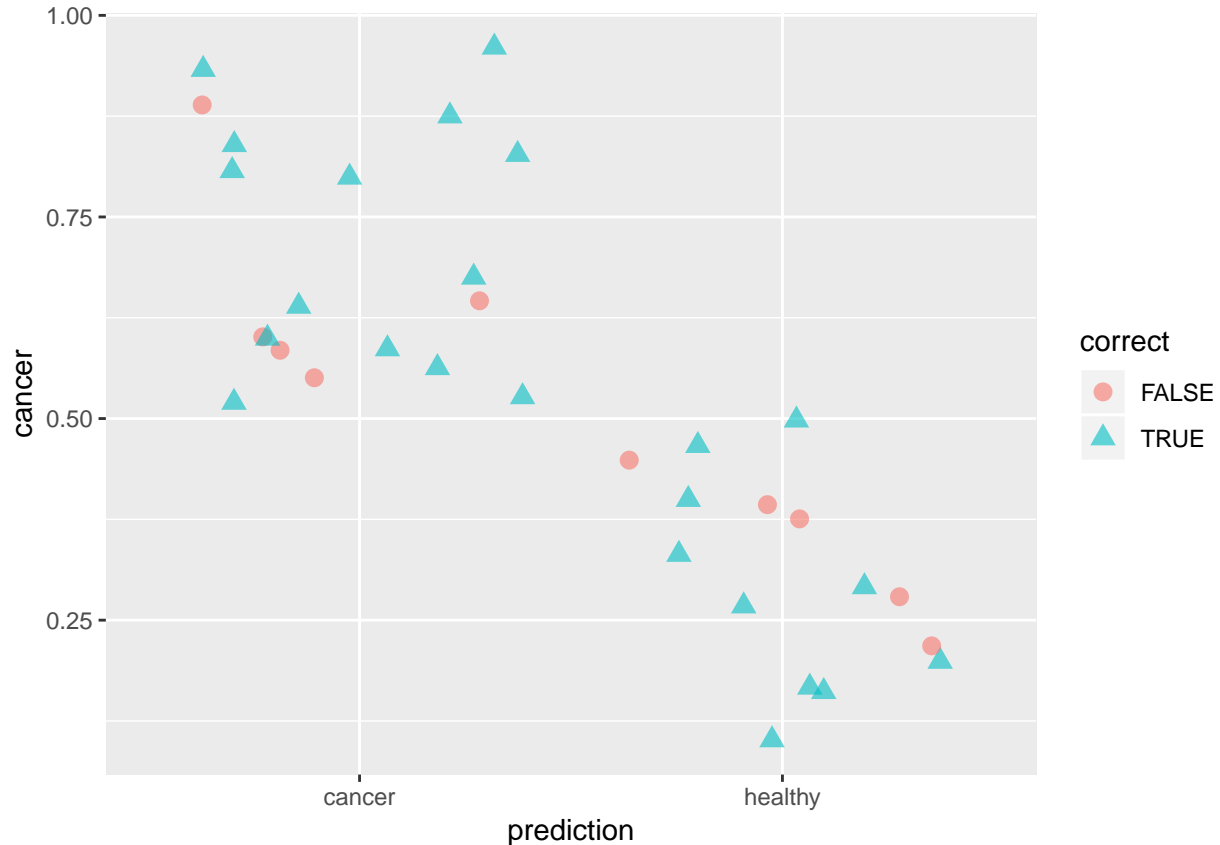
results$prediction <- ifelse(results$healthy > 0.5, "healthy",
                           ifelse(results$cancer > 0.5, "cancer", NA))

results$correct <- ifelse(results$actual == results$prediction, TRUE, FALSE)

ggplot(results, aes(x = prediction, fill = correct)) +
  geom_bar(position = "dodge")
```



```
ggplot(results, aes(x = prediction, y = cancer, color = correct, shape = correct)) +
  geom_jitter(size = 3, alpha = 0.6)
```



EXTREME GRADIENT BOOSTING.

Extreme gradient boosting (XGBoost) is a faster and improved implementation of gradient boosting for supervised learning.

```
set.seed(42)
library(xgboost)
```

```
##
## Attaching package: 'xgboost'

## The following object is masked from 'package:XVector':
##
##   slice

## The following object is masked from 'package:IRanges':
##
##   slice

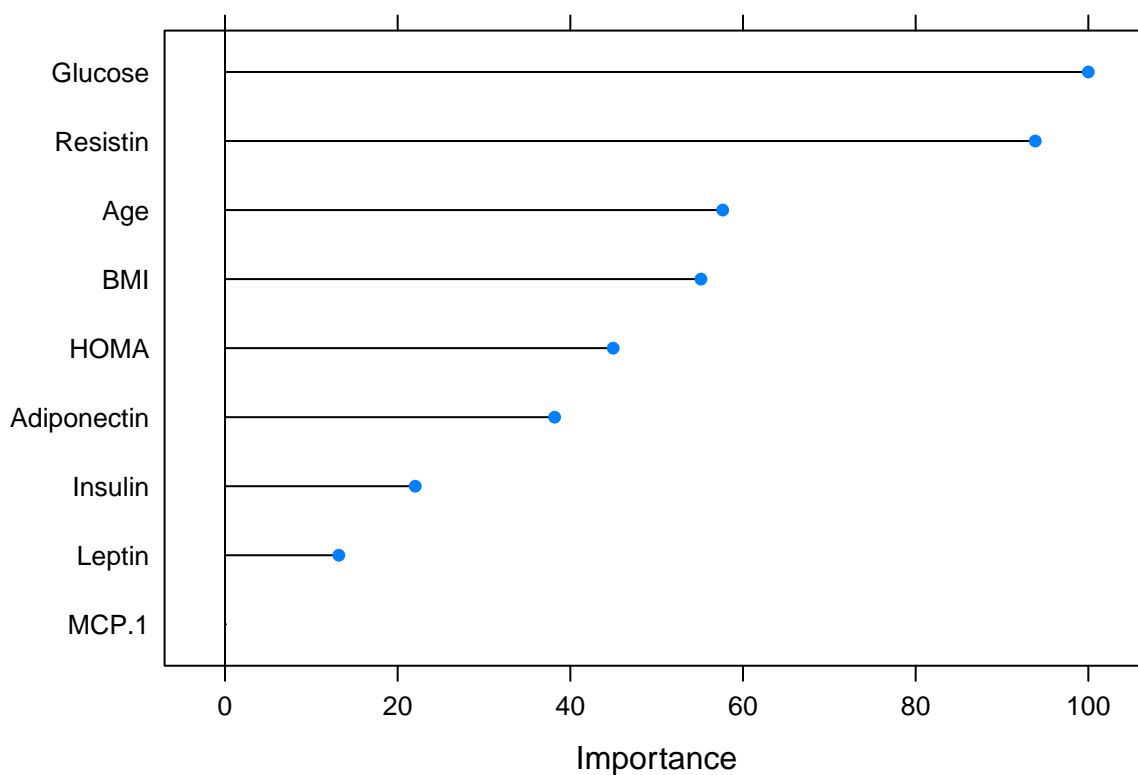
## The following object is masked from 'package:dplyr':
##
##   slice

model_xgb <- caret::train(Classification ~ .,
  data = train_data,
  method = "xgbTree",
  preProcess = c("scale", "center"),
  trControl = trainControl(method = "repeatedcv",
```

```
number = 10,
repeats = 10,
savePredictions = TRUE,
verboseIter = FALSE))
```

Feature Importance

```
importance <- varImp(model_xgb, scale = TRUE)
plot(importance)
```



#Predicting test data

```
confusionMatrix(predict(model_xgb, test_data), test_data$Class)
```

```
## Confusion Matrix and Statistics
```

```
##
```

```
##           Reference
```

```
## Prediction cancer healthy
```

```
##   cancer      14      5
```

```
##   healthy     5     10
```

```
##
```

```
##           Accuracy : 0.7059
```

```
##           95% CI : (0.5252, 0.849)
```

```
##   No Information Rate : 0.5588
```

```
##   P-Value [Acc > NIR] : 0.0582
```

```
##
```

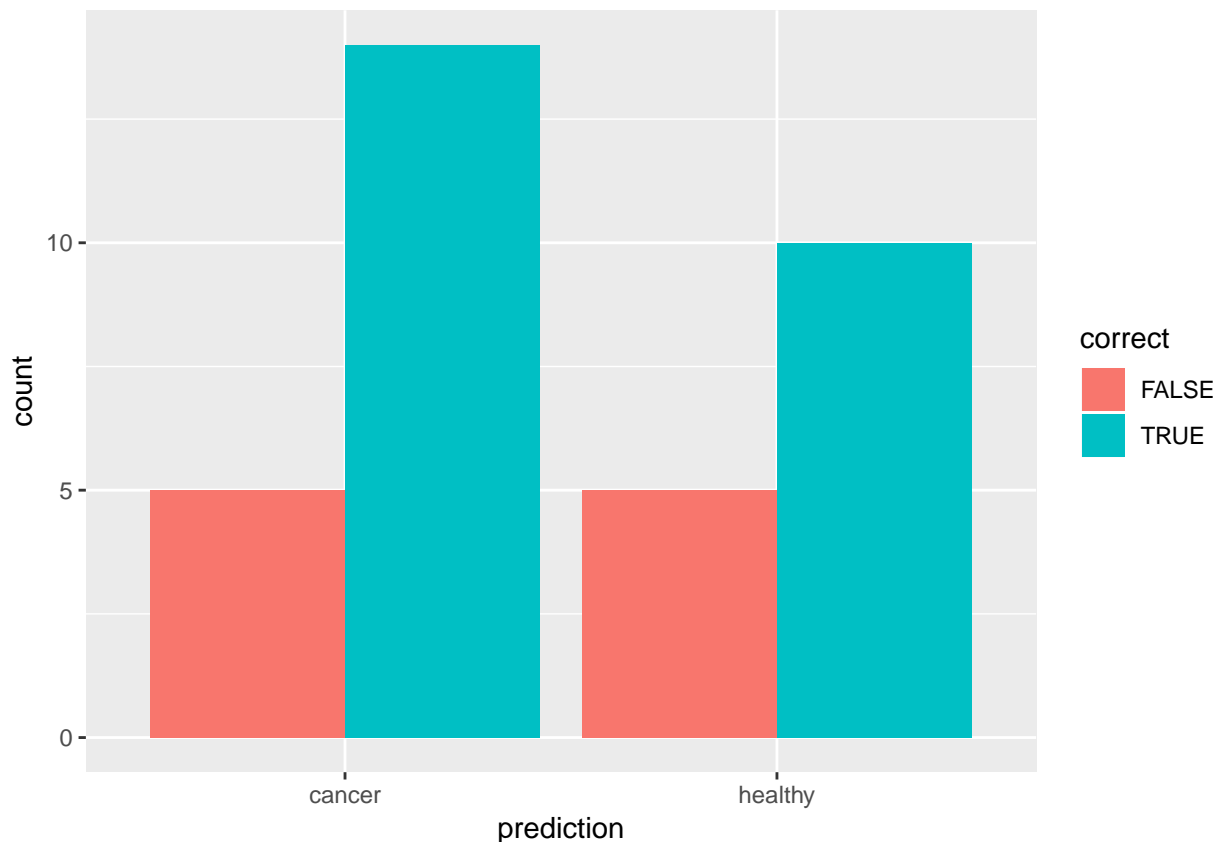
```
##           Kappa : 0.4035
##
## Mcnemar's Test P-Value : 1.0000
##
##           Sensitivity : 0.7368
##           Specificity : 0.6667
##           Pos Pred Value : 0.7368
##           Neg Pred Value : 0.6667
##           Prevalence : 0.5588
##           Detection Rate : 0.4118
##           Detection Prevalence : 0.5588
##           Balanced Accuracy : 0.7018
##
##           'Positive' Class : cancer
##
```

```
results <- data.frame(actual = test_data$Classification,
                      predict(model_xgb, test_data, type = "prob"))

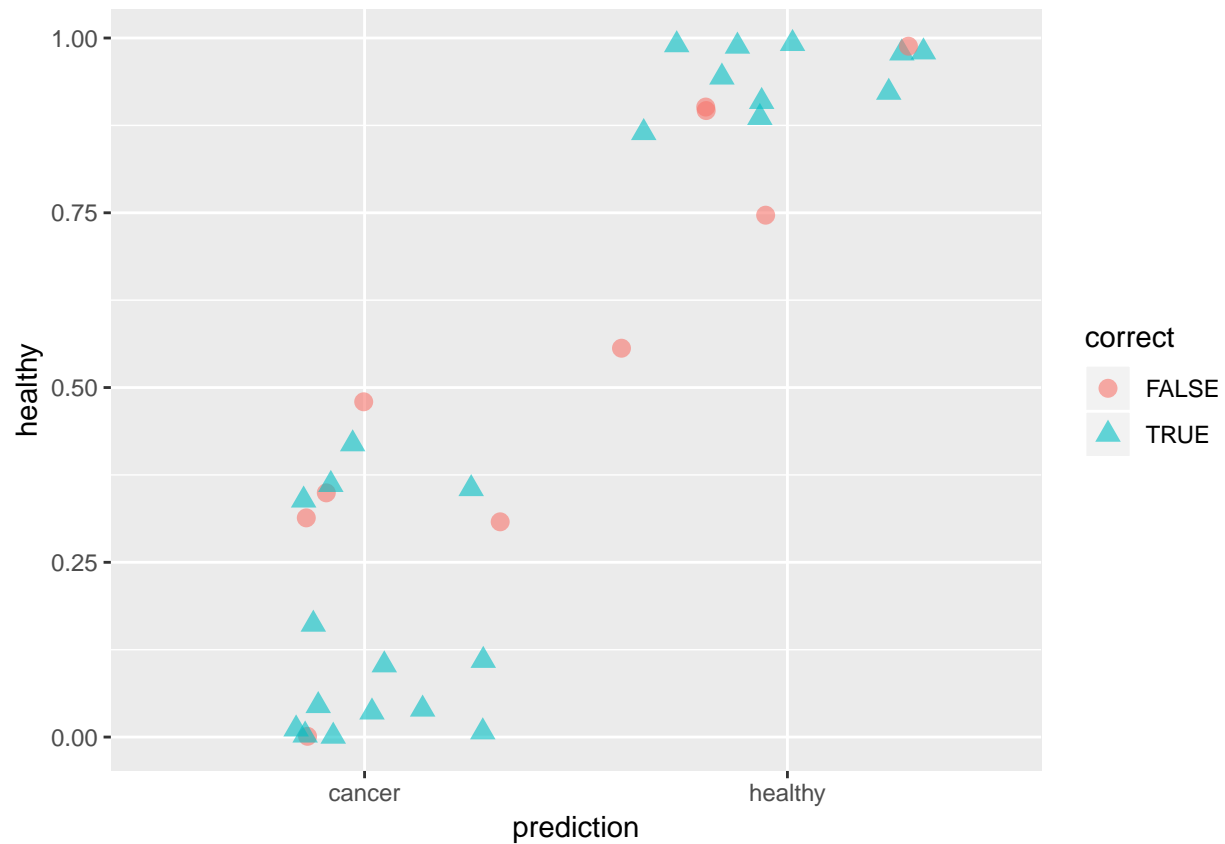
results$prediction <- ifelse(results$healthy > 0.5, "healthy",
                           ifelse(results$cancer > 0.5, "cancer", NA))

results$correct <- ifelse(results$actual == results$prediction, TRUE, FALSE)

ggplot(results, aes(x = prediction, fill = correct)) +
  geom_bar(position = "dodge")
```



```
ggplot(results, aes(x = prediction, y = healthy, color = correct, shape = correct)) +  
  geom_jitter(size = 3, alpha = 0.6)
```



Feature Selection

Correlation

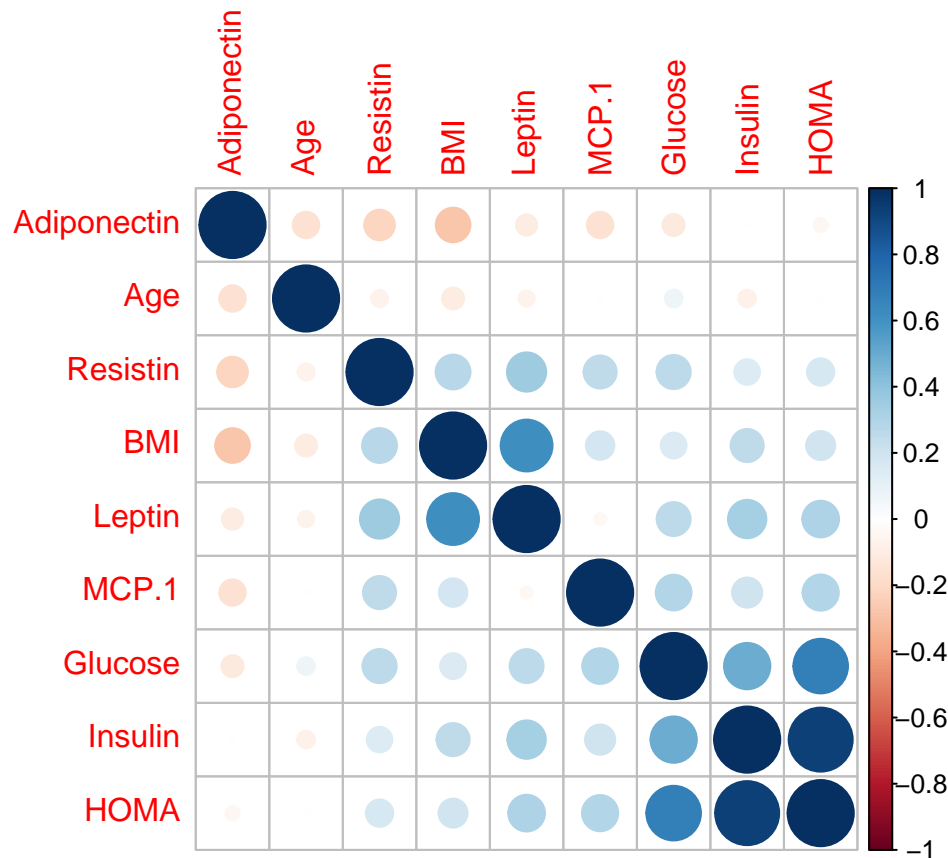
```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
# calculate correlation matrix
```

```
corMatMy <- cor(train_data[, 1:9])
```

```
corrplot(corMatMy, order = "hclust")
```

```
#Apply correlation filter at 0.70:
highlyCor <- colnames(train_data[, -1])[findCorrelation(corMatMy, cutoff = 0.7, verbose = TRUE)]

## Compare row 5 and column 4 with corr 0.935
## Means: 0.33 vs 0.217 so flagging column 5
## All correlations <= 0.7

# which variables are flagged for removal?
highlyCor

## [1] "Leptin"

#then we remove these variables
train_data_cor <- train_data[, which(!colnames(train_data) %in% highlyCor)]
```

GRID SEARCH WITH CARET

Automatic Grid

```
set.seed(42)
model_rf_tune_auto <- caret::train(Classification ~ .,
  data = train_data,
  method = "rf",
  preProcess = c("scale", "center"),
  trControl = trainControl(method = "repeatedcv",
    number = 10,
```

```

                                repeats = 10,
                                savePredictions = TRUE,
                                verboseIter = FALSE,
                                search = "random"),

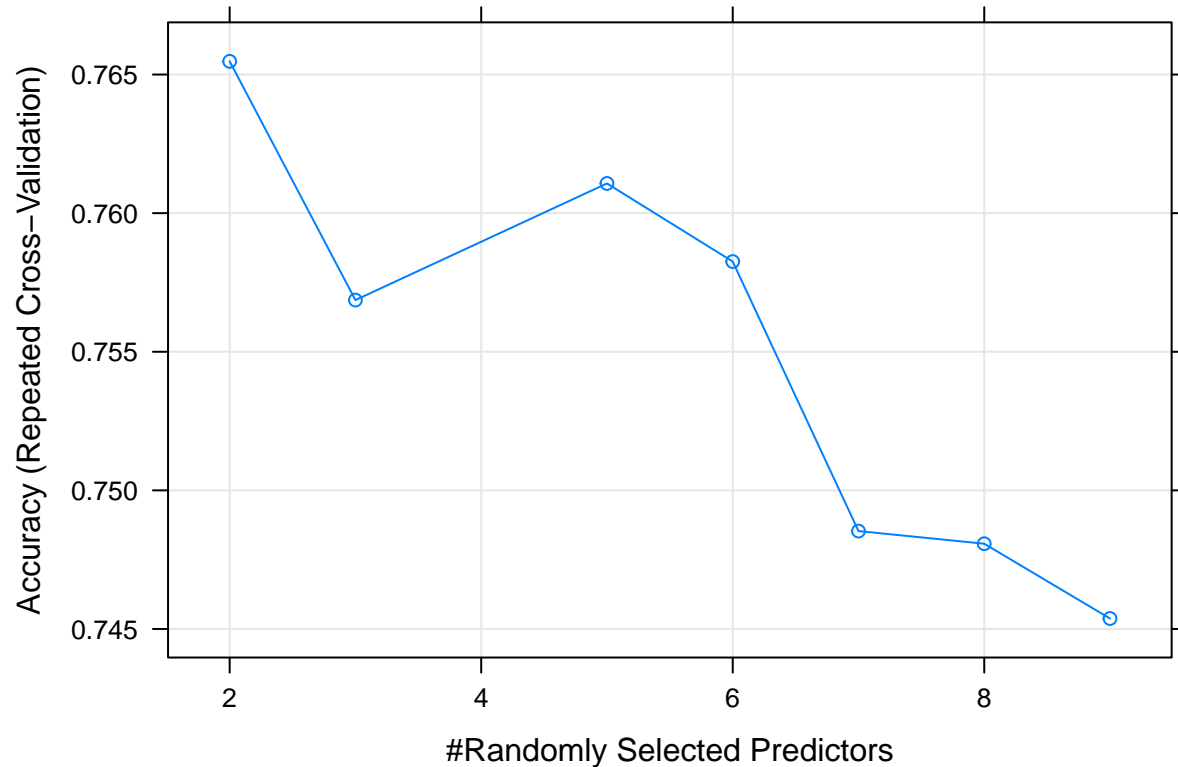
                                tuneLength = 15)

model_rf_tune_auto

## Random Forest
##
## 82 samples
## 9 predictor
## 2 classes: 'cancer', 'healthy'
##
## Pre-processing: scaled (9), centered (9)
## Resampling: Cross-Validated (10 fold, repeated 10 times)
## Summary of sample sizes: 74, 75, 73, 74, 74, 74, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##   2     0.7654762  0.5207469
##   3     0.7568651  0.5033515
##   5     0.7610714  0.5134804
##   6     0.7582540  0.5092464
##   7     0.7485317  0.4869195
##   8     0.7480754  0.4881144
##   9     0.7453770  0.4815800
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 2.

plot(model_rf_tune_auto)

```

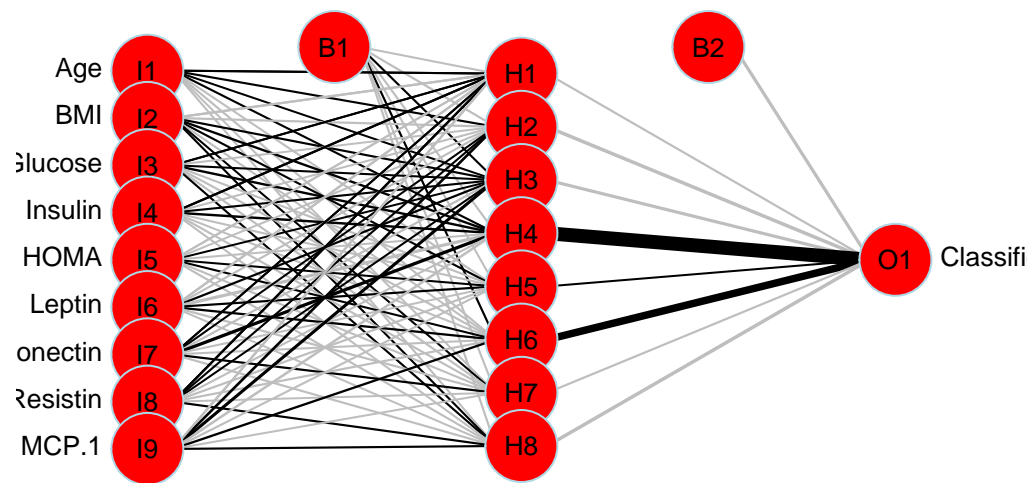


NEURAL NETWORK MODEL

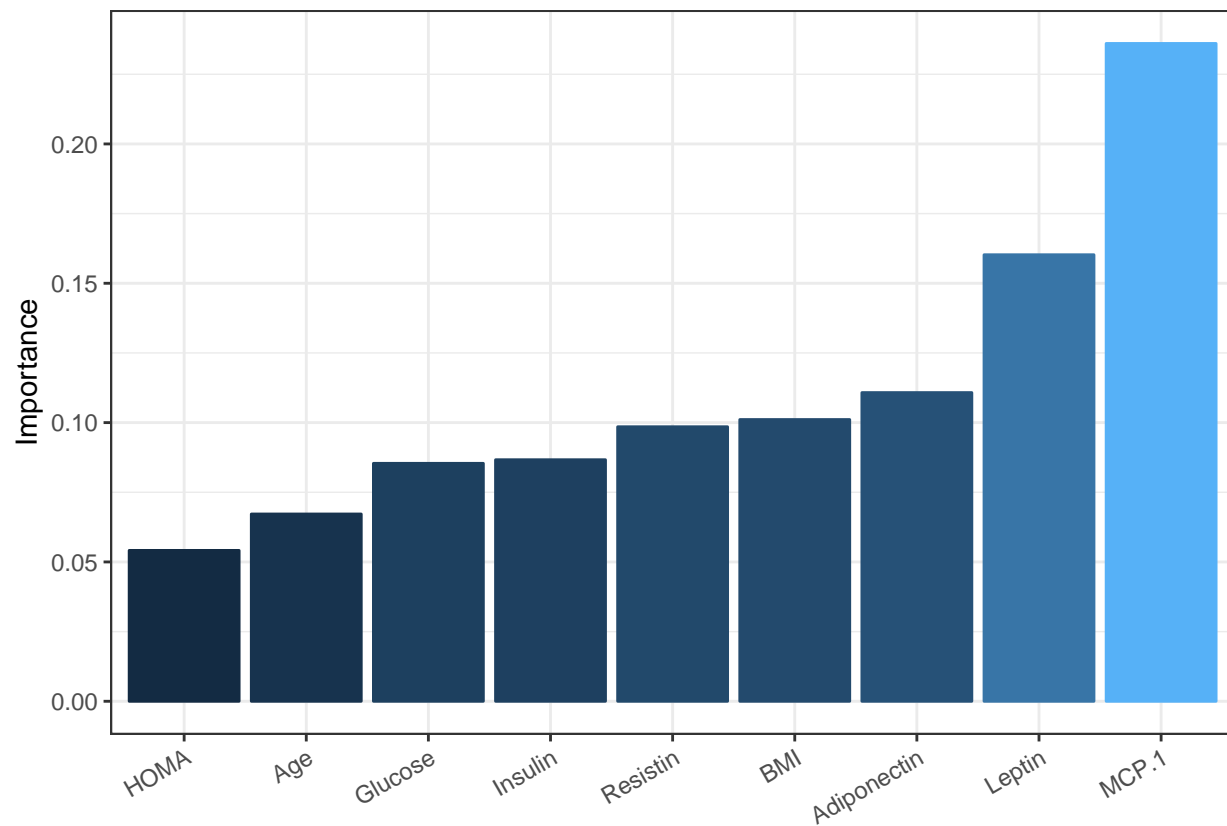
```
library(nnet)
model_nnet<-nnet(Classification ~. ,
                  data= train_data,
                  size=8
)
```

```
## # weights: 89
## initial value 58.952401
## iter 10 value 54.963035
## iter 20 value 54.245492
## iter 30 value 54.244855
## iter 30 value 54.244855
## iter 30 value 54.244855
## final value 54.244855
## converged
```

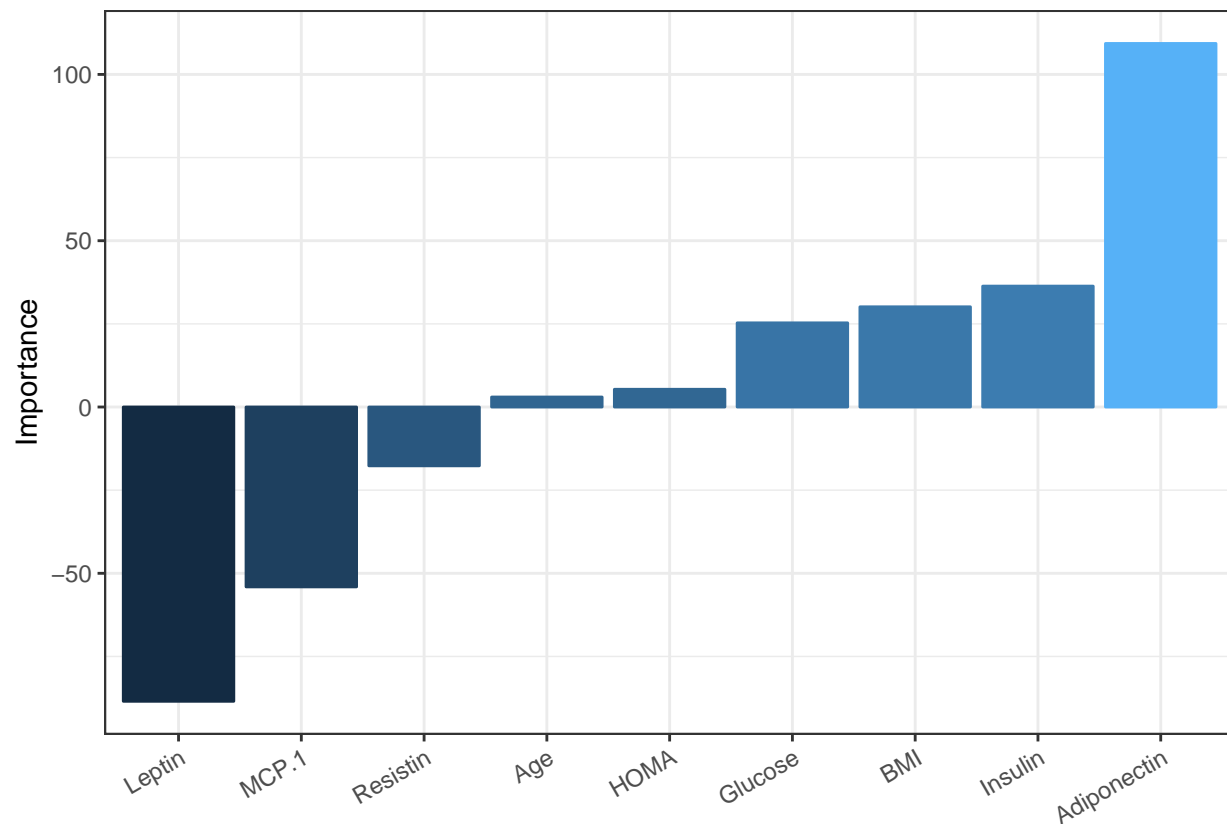
```
library(NeuralNetTools)
# Plot a neural interpretation diagram for a neural network object
plotnet(model_nnet, cex_val =.8,max_sp=T, circle_cex=5, circle_col = 'red')
```



```
#Relative importance of input variables in neural networks using Garson's algorithm:
garson(model_nnet) +
  theme(axis.text.x = element_text(angle = 30, hjust = 1))
```



```
olden(model_nnet) +  
  theme(axis.text.x = element_text(angle = 30, hjust = 1))
```



Here both the positive and negative value represents relative contributions of each connection weight among the variables

```
#Predict
predict_nnet <- predict(model_nnet,test_data, type = "class")
```

```
#Draw the crosstable
```

```
library(gmodels)
CrossTable(test_data$Class,predict_nnet,prop.chisq = F,prop.r = F,prop.c = F,dnn =c("Actual Diagnosis",
```

```
##
##
##   Cell Contents
## |-----|
## |                      N |
## |          N / Table Total |
## |-----|
##
##
## Total Observations in Table:  34
##
##
##               | predict_nnet
## test_data$Class |      cancer | Row Total |
## -----|-----|-----|
##           cancer |          19 |          19 |
##               |          0.559 |          |
```

##	-----	-----	-----
##	healthy	15	15
##		0.441	
##	-----	-----	-----
##	Column Total	34	34
##	-----	-----	-----
##			
##			