preliminaryAnalysis.R

riserate

Wed Aug 30 21:23:18 2017

source('helper\_functions.R')  
library(randomForest)

## randomForest 4.6-12

## Type rfNews() to see new features/changes/bug fixes.

library(e1071)  
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

##   
## Attaching package: 'ggplot2'

## The following object is masked from 'package:randomForest':  
##   
## margin

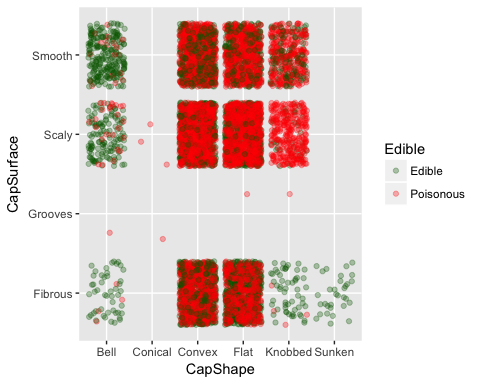
library(ggplot2)  
  
set.seed(123)   
  
data = prepareAndCleanData()  
data.backup = data  
  
head(data)

## Edible CapShape CapSurface CapColor Bruises Odor GillAttachment  
## 1 Poisonous Convex Smooth Brown True Pungent Free  
## 2 Edible Convex Smooth Yellow True Almond Free  
## 3 Edible Bell Smooth White True Anise Free  
## 4 Poisonous Convex Scaly White True Pungent Free  
## 5 Edible Convex Smooth Gray False None Free  
## 6 Edible Convex Scaly Yellow True Almond Free  
## GillSpacing GillSize GillColor StalkShape StalkRoot  
## 1 Close Narrow Black Enlarging Equal  
## 2 Close Broad Black Enlarging Club  
## 3 Close Broad Brown Enlarging Club  
## 4 Close Narrow Brown Enlarging Equal  
## 5 Crowded Broad Black Tapering Equal  
## 6 Close Broad Brown Enlarging Club  
## StalkSurfaceAboveRing StalkSurfaceBelowRing StalkColorAboveRing  
## 1 Smooth Smooth White  
## 2 Smooth Smooth White  
## 3 Smooth Smooth White  
## 4 Smooth Smooth White  
## 5 Smooth Smooth White  
## 6 Smooth Smooth White  
## StalkColorBelowRing VeilType VeilColor RingNumber RingType  
## 1 White Partial White One Pendant  
## 2 White Partial White One Pendant  
## 3 White Partial White One Pendant  
## 4 White Partial White One Pendant  
## 5 White Partial White One Evanescent  
## 6 White Partial White One Pendant  
## SporePrintColor Population Habitat  
## 1 Black Scattered Urban  
## 2 Brown Numerous Grasses  
## 3 Brown Numerous Meadows  
## 4 Black Scattered Urban  
## 5 Brown Abundnant Grasses  
## 6 Black Numerous Grasses

summary(data)

## Edible CapShape CapSurface CapColor   
## Edible :4208 Convex :3656 Scaly :3244 Brown :2284   
## Poisonous:3916 Flat :3152 Smooth :2556 Gray :1840   
## Knobbed: 828 Fibrous:2320 Red :1500   
## Bell : 452 Grooves: 4 Yellow :1072   
## Sunken : 32 f : 0 White :1040   
## Conical: 4 g : 0 Buff : 168   
## (Other): 0 (Other): 0 (Other): 220   
## Bruises Odor GillAttachment GillSpacing   
## f : 0 None :3528 a : 0 c : 0   
## t : 0 Foul :2160 f : 0 w : 0   
## True :3376 Fishy : 576 Attached : 210 Close :6812   
## False:4748 Spicy : 576 Descending: 0 Crowded:1312   
## Almond : 400 Free :7914 Distant: 0   
## Anise : 400 Notched : 0   
## (Other): 484   
## GillSize GillColor StalkShape StalkRoot   
## b : 0 Buff :1728 e : 0 Bulbous:3776   
## n : 0 Pink :1492 t : 0 Missing:2480   
## Broad :5612 White :1202 Enlarging:3516 Equal :1120   
## Narrow:2512 Brown :1048 Tapering :4608 Club : 556   
## Gray : 752 Rooted : 192   
## Chocolate: 732 ? : 0   
## (Other) :1170 (Other): 0   
## StalkSurfaceAboveRing StalkSurfaceBelowRing StalkColorAboveRing  
## Smooth :5176 Smooth :4936 White :4464   
## Silky :2372 Silky :2304 Pink :1872   
## Fibrous: 552 Fibrous: 600 Gray : 576   
## Scaly : 24 Scaly : 284 Brown : 448   
## f : 0 f : 0 Buff : 432   
## k : 0 k : 0 Orange : 192   
## (Other): 0 (Other): 0 (Other): 140   
## StalkColorBelowRing VeilType VeilColor RingNumber   
## White :4384 p : 0 White :7924 n : 0   
## Pink :1872 Partial :8124 Brown : 96 o : 0   
## Gray : 576 Universal: 0 Orange : 96 t : 0   
## Brown : 512 Yellow : 8 None: 36   
## Buff : 432 n : 0 One :7488   
## Orange : 192 o : 0 Two : 600   
## (Other): 156 (Other): 0   
## RingType SporePrintColor Population Habitat   
## Pendant :3968 White :2388 Several :4040 Woods :3148   
## Evanescent:2776 Brown :1968 Solitary :1712 Grasses:2148   
## Large :1296 Black :1872 Scattered:1248 Paths :1144   
## Flaring : 48 Chocolate:1632 Numerous : 400 Leaves : 832   
## None : 36 Green : 72 Abundnant: 384 Urban : 368   
## e : 0 Buff : 48 Clustered: 340 Meadows: 292   
## (Other) : 0 (Other) : 144 (Other) : 0 (Other): 192

#Odor and SporePrintColor are the best predictors  
p = ggplot(data,aes(x=CapShape, y=CapSurface, color=Edible))  
p + geom\_jitter(alpha=0.3) + scale\_color\_manual(breaks = c('Edible','Poisonous'),values=c('darkgreen','red'))



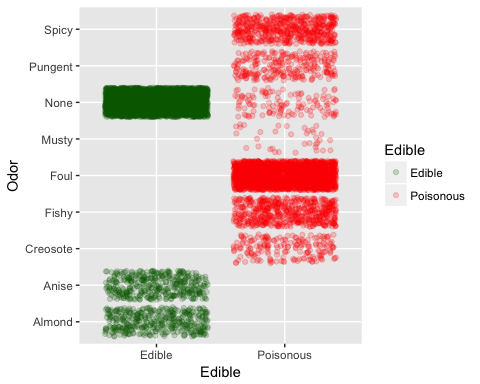
p = ggplot(data,aes(x=StalkColorBelowRing, y=StalkColorAboveRing, color=Edible))  
p + geom\_jitter(alpha=0.3) + scale\_color\_manual(breaks = c('Edible','Poisonous'),values=c('darkgreen','red'))



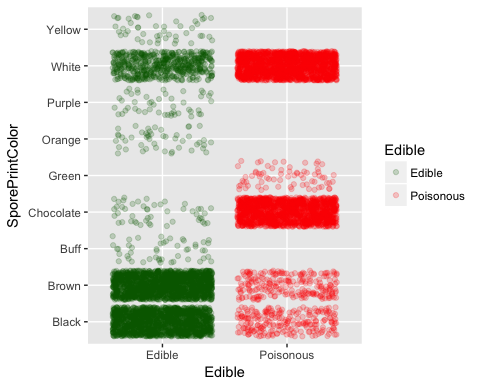
p = ggplot(data,aes(x=Odor, y=SporePrintColor, color=Edible))  
p + geom\_jitter(alpha=0.3) + scale\_color\_manual(breaks = c('Edible','Poisonous'),values=c('darkgreen','red'))



p = ggplot(data,aes(x=Edible, y=Odor, color = Edible))  
p + geom\_jitter(alpha=0.2) + scale\_color\_manual(breaks = c('Edible','Poisonous'),values=c('darkgreen','red'))



p = ggplot(data,aes(x=Edible, y=SporePrintColor, color = Edible))  
p + geom\_jitter(alpha=0.2) + scale\_color\_manual(breaks = c('Edible','Poisonous'),values=c('darkgreen','red'))



#Create data for training  
sample.ind = sample(2,   
 nrow(data),  
 replace = T,  
 prob = c(0.05,0.95))  
data.dev = data[sample.ind==1,]  
data.val = data[sample.ind==2,]  
  
#See how data sets look as edible vs poisonous proportion  
table(data$Edible)/nrow(data)

##   
## Edible Poisonous   
## 0.5179714 0.4820286

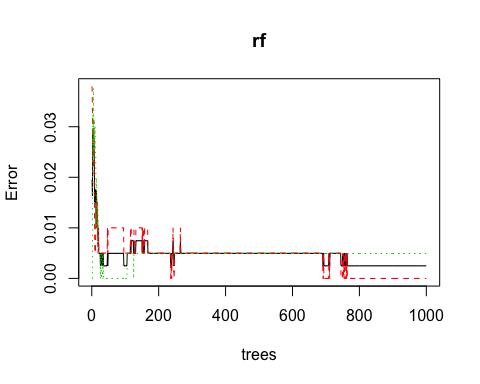
table(data.dev$Edible)/nrow(data.dev)

##   
## Edible Poisonous   
## 0.4962779 0.5037221

table(data.val$Edible)/nrow(data.val)

##   
## Edible Poisonous   
## 0.5191037 0.4808963

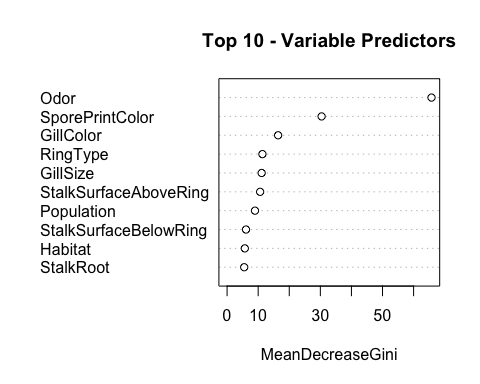
rf = randomForest(Edible ~ .,   
 ntree = 1000,  
 data = data.dev)  
plot(rf)



print(rf)

##   
## Call:  
## randomForest(formula = Edible ~ ., data = data.dev, ntree = 1000)   
## Type of random forest: classification  
## Number of trees: 1000  
## No. of variables tried at each split: 4  
##   
## OOB estimate of error rate: 0.25%  
## Confusion matrix:  
## Edible Poisonous class.error  
## Edible 200 0 0.000000000  
## Poisonous 1 202 0.004926108

varImpPlot(rf,  
 sort = T,  
 n.var=10,  
 main="Top 10 - Variable Predictors")



#Looks like Odor is the greatest indicator  
  
  
var.imp = data.frame(importance(rf,  
 type=2))  
# make row names as columns  
var.imp$Variables <- row.names(var.imp)  
var.imp[order(var.imp$MeanDecreaseGini,decreasing = T),]

## MeanDecreaseGini Variables  
## Odor 65.75846348 Odor  
## SporePrintColor 30.40327948 SporePrintColor  
## GillColor 16.41359952 GillColor  
## RingType 11.36965626 RingType  
## GillSize 11.12495013 GillSize  
## StalkSurfaceAboveRing 10.62786609 StalkSurfaceAboveRing  
## Population 8.97448108 Population  
## StalkSurfaceBelowRing 6.08896562 StalkSurfaceBelowRing  
## Habitat 5.70197687 Habitat  
## StalkRoot 5.50521017 StalkRoot  
## Bruises 4.98638699 Bruises  
## CapColor 4.85220998 CapColor  
## GillSpacing 4.18623088 GillSpacing  
## StalkColorBelowRing 4.05940595 StalkColorBelowRing  
## RingNumber 2.79844022 RingNumber  
## StalkShape 2.55983752 StalkShape  
## StalkColorAboveRing 2.20089911 StalkColorAboveRing  
## CapSurface 1.11693219 CapSurface  
## CapShape 1.11347913 CapShape  
## VeilColor 0.20768802 VeilColor  
## GillAttachment 0.06740127 GillAttachment  
## VeilType 0.00000000 VeilType

# Predicting response variable  
data.dev$predicted.response <- predict(rf , data.dev)  
  
  
## Loading required package: lattice  
## Loading required package: ggplot2  
# Create Confusion Matrix  
confusionMatrix(data = data.dev$predicted.response,  
 reference = data.dev$Edible,  
 positive = 'Edible')

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Edible Poisonous  
## Edible 200 0  
## Poisonous 0 203  
##   
## Accuracy : 1   
## 95% CI : (0.9909, 1)  
## No Information Rate : 0.5037   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 1   
## Mcnemar's Test P-Value : NA   
##   
## Sensitivity : 1.0000   
## Specificity : 1.0000   
## Pos Pred Value : 1.0000   
## Neg Pred Value : 1.0000   
## Prevalence : 0.4963   
## Detection Rate : 0.4963   
## Detection Prevalence : 0.4963   
## Balanced Accuracy : 1.0000   
##   
## 'Positive' Class : Edible   
##

# Predicting response variable  
data.val$predicted.response <- predict(rf ,data.val)  
  
# Create Confusion Matrix  
confusionMatrix(data=data.val$predicted.response,  
 reference=data.val$Edible,  
 positive='Edible')

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Edible Poisonous  
## Edible 3960 8  
## Poisonous 48 3705  
##   
## Accuracy : 0.9927   
## 95% CI : (0.9906, 0.9945)  
## No Information Rate : 0.5191   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.9855   
## Mcnemar's Test P-Value : 1.872e-07   
##   
## Sensitivity : 0.9880   
## Specificity : 0.9978   
## Pos Pred Value : 0.9980   
## Neg Pred Value : 0.9872   
## Prevalence : 0.5191   
## Detection Rate : 0.5129   
## Detection Prevalence : 0.5139   
## Balanced Accuracy : 0.9929   
##   
## 'Positive' Class : Edible   
##