Machine Learning Approach in Detecting Poisonous Mushrooms

Ewa Poczman

17/06/2019

Introduction

This report presents how machine learning algorithms can be used to detect poisonous mushrooms.

In my home country, Poland, wild mushroom picking is a national pastime. Despite the fact that the death rates from mushroom poisoning are decreasing from highs of 500 per year a few decades ago, nowadays nearly 100 people die every year because they consumed a poisonous mushroom.

The dataset used in this analysis was downloaded from Kaggle and was originally contributed to UCI Machine Learning repository in 1987. According to the dataset summary provided by the UCI, it “includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms (1981)”. While in the original dataset each species was identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended, this dataset divides the mushrooms into two categories: poisonous and edible, with the latter category combining cattegories of “poisonous” and of “unknown edibility”.

The aim of the machine learning exercise presented below is to predict which mushrooms are poisonous, with priority placed on accurately detecting poisonous mushrooms. While mushroom lovers wouldn’t want to throw away an edible mushroom, the consequences of consuming a poisonous mushroom can be deadly.

As the outcome variable - identifying mushroom as posionous or edible - is categorical and so are all predictors, the two machine learning approaches presented below are decision trees and random forests.

Finally, while the dataset includes a wide range of features, only those that can be inspected visually by a layman are included in the analysis.

In the preparation of this report the following steps have been performed: - inspection of the dataset, data cleaning and selection of predictors meeing the criteria of practicality - partitioning of the dataset into training and test set - development of machine learning algorithms on the training set - testing of each of the algorithms on the test set - comparison of the outcomes of the two approaches

Analysis

The following libraries will be used in the inspection and analysis of the dataset:

```
library(tidyverse)
library(caret)
library(rpart)
library(rpart.plot)
library(randomForest)
```

The dataset can be loaded from the following github location:

```
url <- "https://raw.githubusercontent.com/epoczman/createyourown/master/mushrooms.c
sv"
download.file(url, "mushrooms.csv")
mushrooms <- read.csv("mushrooms.csv")
```

Initial inspection of the dataset reveals that it is one data frame contaning 23 variables and 8124 observations. Apart from the edible vs. poisonous mushroom class (outcome variable), 22 features are recorded for each observation. Both the outcome variable and the features are categorical nominal factor variables.

One of the variables, “veil.type” has only one level, with all observations in the dataset described as having a “partial veil type”, coded as “p”:

```
## [1] "p"
```

This variable should be excluded from the dataset as it doesn’t contain any information that could be used in the prediction of variable class.

The dataset is inspected for existance of any missing values and there appear to be no missing values. However, further inspecton reveals that the variable stalk.root includes value “?”

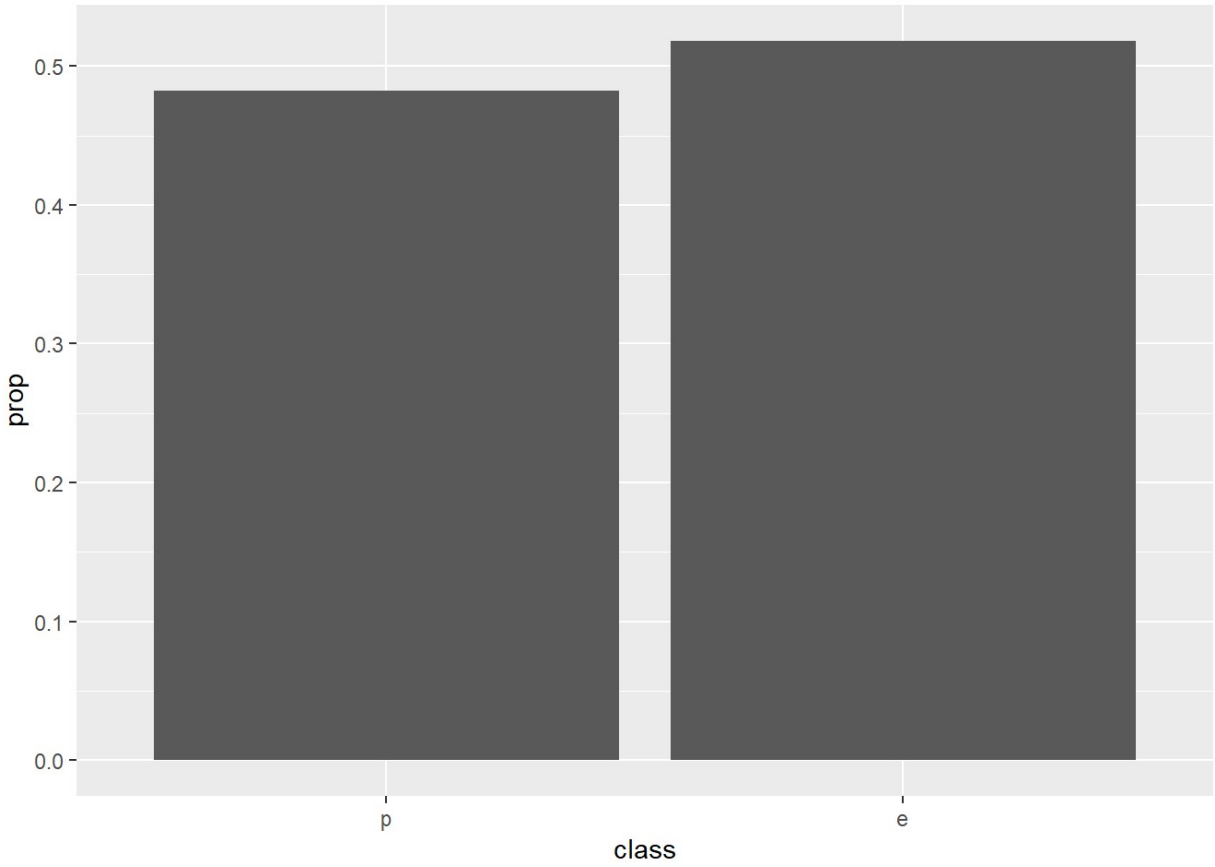
```
## [1] e c b r ?
## Levels: ? b c e r
```

This value is defined as a missing value in the information provided on the dataset by the UCI. As this value is assigned to a substantial portion of the dataset, 31%, the entire variable “stalk.root” should be removed.

Two other variables: “odor” and “spore.print.colour” cannot be inspected visually by a layman. “Odor” can be considered subjective and most people would not feel comfortable if their decision whether or not to consume a mushroom depended on accurate identification of its odor as e.g. “almond”. “Spore.print.colour” can be inspected visually but collecting spore prints might be too difficult for a layman. Those two variables should therefore be excuded from the dataset.

As detection of poisonous mushrooms takes priority before detection of edible mushrooms, the outcome variable should be releveled, to have level “p” (denoting poisonous mushrooms) as the first level.

Finally, before the data is partitioned into training and test set, the prevalence of the two classes is established, with poisonous mushrooms constituting nearly half of the entire new dataset.



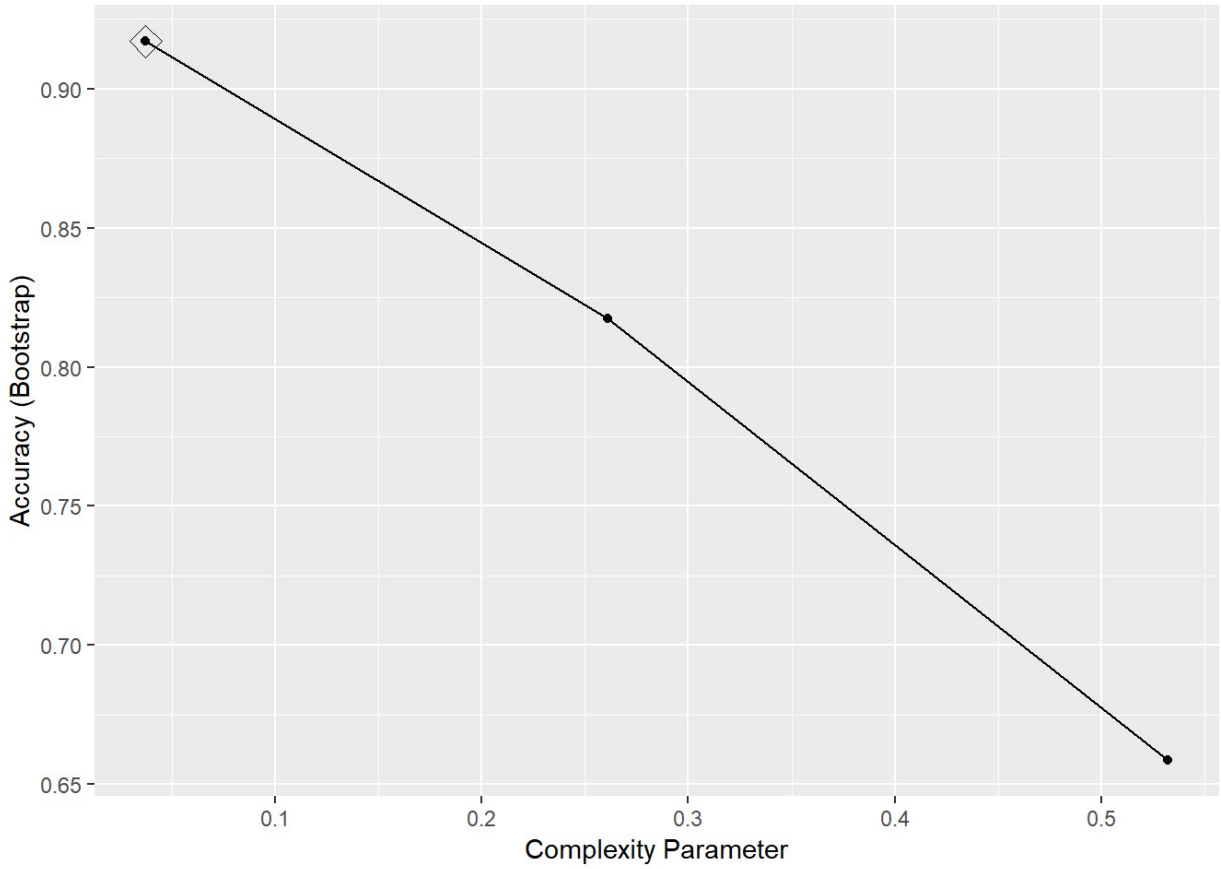
Once the dataset is pre-processed, it can be partitioned into training and test sets. As the objective of the analysis is to build machine learning algorithms that would identify poisonous mushrooms based on their features, the variable “class” will be defined as outcome variable (y).

The dataset is then split into two sets: training set with 7311 observations and test set with 813 observations.

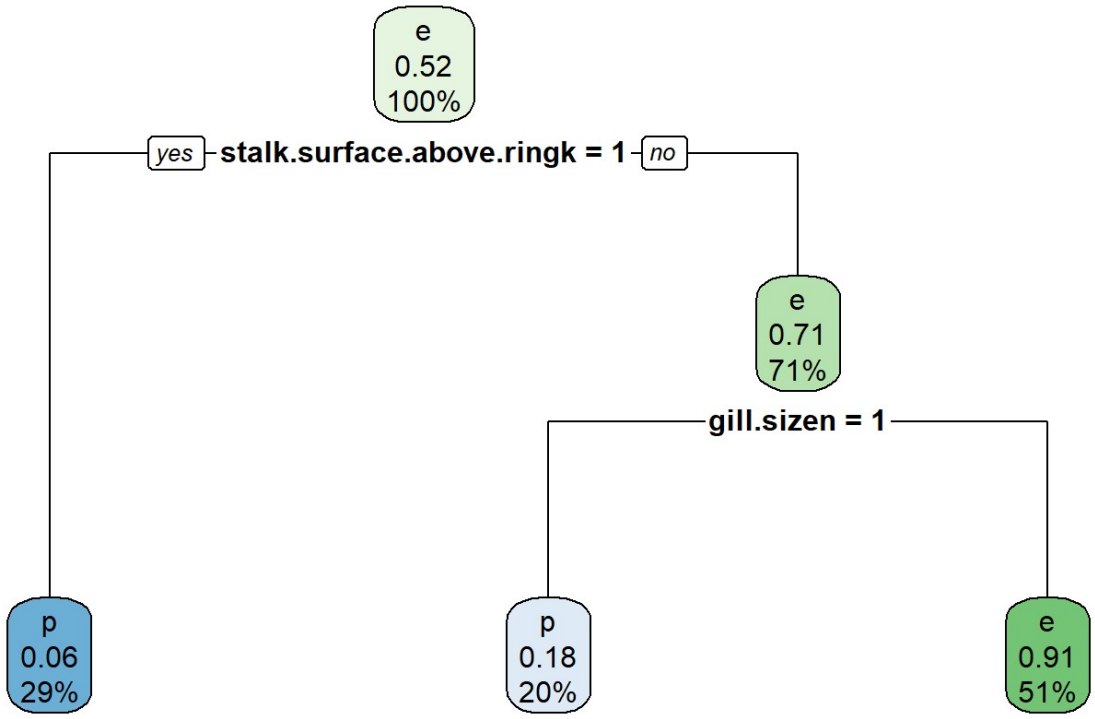
##Model 1 - Decision Trees

The first model is built with *train* function and *rpart* method, using all predictors in the new dataset.

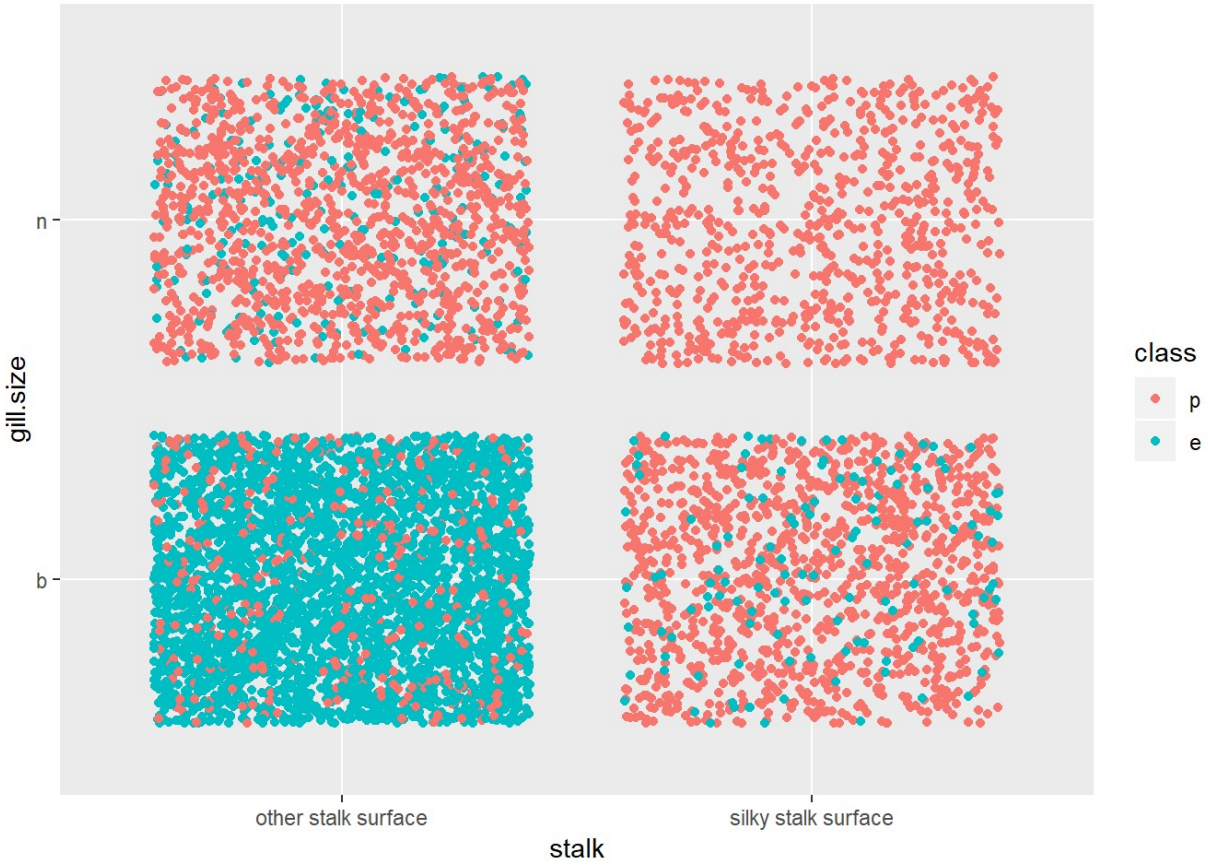
The chart below shows how the optimal model has relatively low complexity parameter and accuracy (as measured in the *training set*) of above 90%.



The decision tree plotted for this model shows that there are only two variables used in decision making: if stalk surface above mushroom ring is identified as “k” (which stands for “silky”), the mushroom will be predicted to be poisonous; if the surface is not silky, gill size will be considered, with gills identified as “n” (“narrow”) leading to a prediction of a poisonous mushroom and mushroom predicted to be edible otherwise.



The below chart shows visually how the two variables greatly improve prediction of mushroom class.

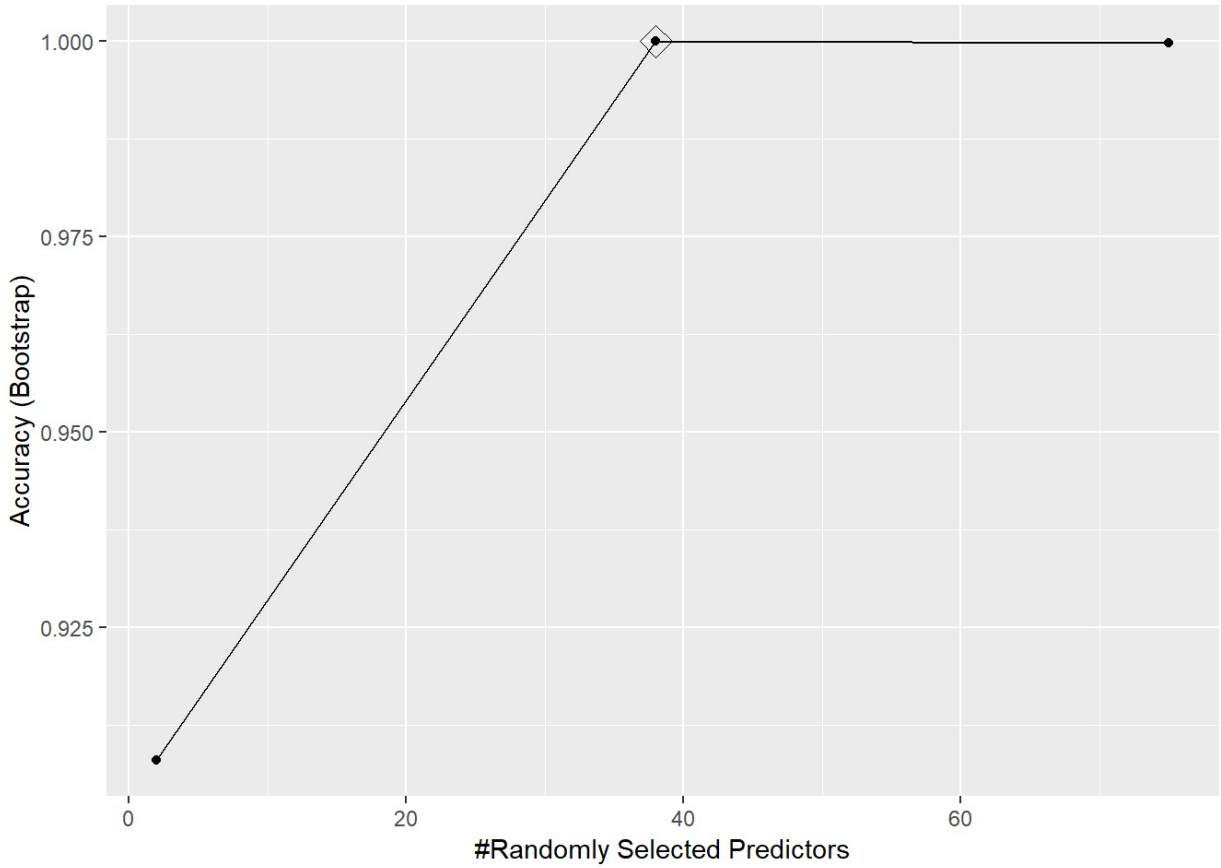


However, it appears that sensitivity of the algorithm is below 1. Those who follow the decision rule defined in the algorithm and expect the mushroom with other than silky stalk surface above ring and broad gills to be edible might still make an error (with cases of poisonous mushrooms depicted by the red dots in the bottom left corner quadrant).

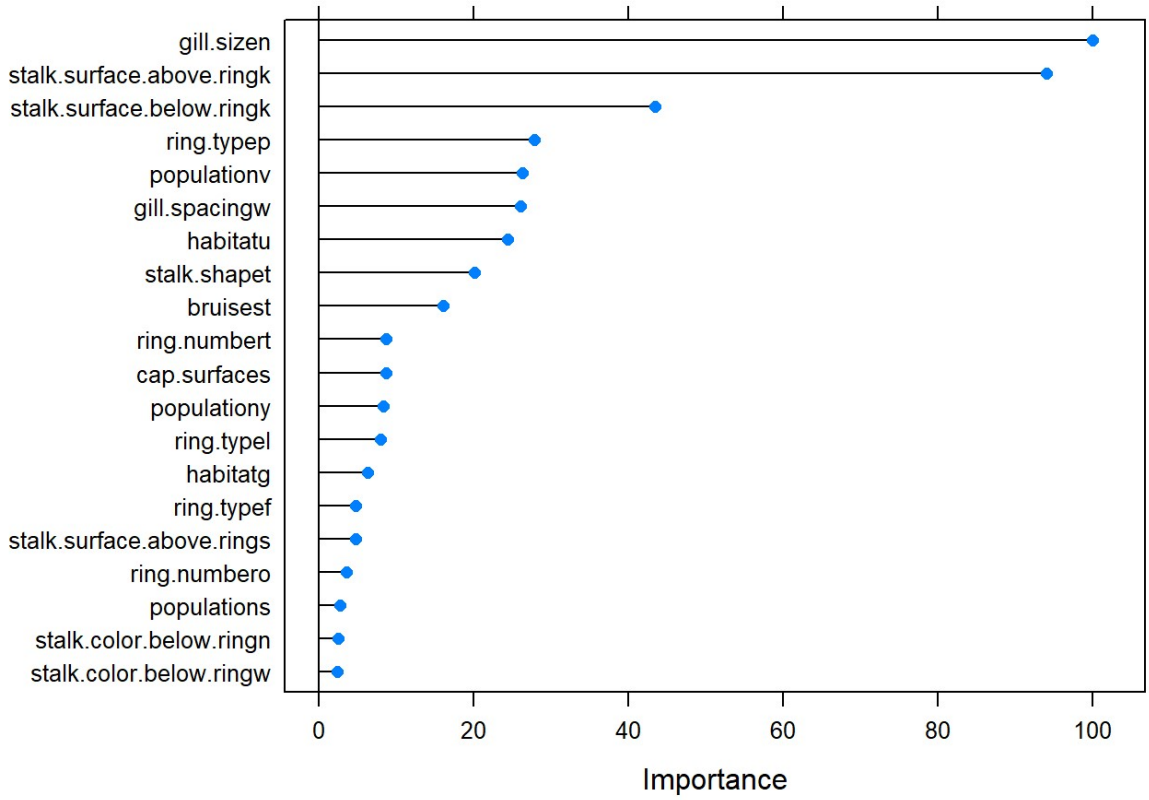
##Model 2 - Random Forests

Similarly to the first model, the second one was also built using function *train*. This time, *random forest (rf)* method was used.

The model obtained by training has the accuracy of 1 - better than Model 1.



Model generated using random forest method is difficult to interpret but through accessing the importance of predictor veriables we can show which of the features analysed should be prioritized. This time again narrow gill size and silky stalk surface above ring are the two variables of highest importance in the model.



Results

The two models are then tested on the test set to determine their ability to detect poisonous mushrooms.

Model 1

Accuracy of Model 1 (decision trees) is 0.9212792.

Closer inspection of the confusion matrix for this model reveals that model sensitivity is at a similar level:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   p   e
##          p 363  35
##          e  29 386
##
##                Accuracy : 0.9213
##                  95% CI : (0.9006, 0.9389)
##     No Information Rate : 0.5178
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.8424
##
##  Mcnemar's Test P-Value : 0.532
##
##             Sensitivity : 0.9260
##             Specificity : 0.9169
##          Pos Pred Value : 0.9121
##          Neg Pred Value : 0.9301
##              Prevalence : 0.4822
##          Detection Rate : 0.4465
##    Detection Prevalence : 0.4895
##       Balanced Accuracy : 0.9214
##
##        'Positive' Class : p
##
```

Model 2

Model 2 results in accuracy of 1.

This means that models has perfect sensitivity and specificity.

Conclusion

If the decision whether a mushroom is edible or poisonous had to be made based on the result of a fair coin toss, on average 24% of cases would end in consumption of a poisonous mushroom.

The two models greatly improve ability to detect poisonous mushrooms, with Model 1 (Decision Trees) detecting 9 in every 10 poisonous mushrooms and Model 2 (Random Forrests) accurately predicting whether a mushroom is poisonous or edible at all times.

Both models indicate that the most important features in prediction are gill size and stalk surface above ring.

As the random forest model might be difficult to apply in decision making of average mushroom pickers, further research should be conducted on development of a tool, such as e.g. an app, that could conduct assessment based on images of mushrooms taken by users.