Mushroom Classification

Rajesh Grandhi , Vivek Pruthi, Murali Korrapati and Jyothi Pulimamidi

July 22, 2017

A mushroom (or toadstool) is the fleshy, [**spore**](https://en.wikipedia.org/wiki/Spore)-bearing [**fruiting body**](https://en.wikipedia.org/wiki/Sporocarp_%28fungi%29) of a [**fungus**](https://en.wikipedia.org/wiki/Fungus), typically produced above ground on soil or on its [**food**](https://en.wikipedia.org/wiki/Food) source.

The standard for the name "mushroom" is the cultivated white button mushroom, [**Agaricus bisporus**](https://en.wikipedia.org/wiki/Agaricus_bisporus); hence the word "mushroom" is most often applied to those fungi ([**Basidiomycota**](https://en.wikipedia.org/wiki/Basidiomycota), [**Agaricomycetes**](https://en.wikipedia.org/wiki/Agaricomycetes)) that have a stem ([**stipe**](https://en.wikipedia.org/wiki/Stipe_%28mycology%29)), a cap ([**pileus**](https://en.wikipedia.org/wiki/Pileus_%28mycology%29)), and gills (lamellae, sing. [**lamella**](https://en.wikipedia.org/wiki/Lamella_%28mycology%29)) on the underside of the cap. These gills produce microscopic spores that help the fungus spread across the ground or its occupant surface.

"Mushroom" describes a variety of gilled fungi, with or without stems, and the term is used even more generally, to describe both the fleshy fruiting bodies of some [**Ascomycota**](https://en.wikipedia.org/wiki/Ascomycota) and the woody or leathery fruiting bodies of some Basidiomycota, depending upon the context of the word.

Identifying mushrooms requires a basic understanding of their [**macroscopic**](https://en.wikipedia.org/wiki/Macroscopic) structure.While modern identification of mushrooms is quickly becoming molecular, the standard methods for identification are still used by most and have developed into a fine art harking back to [**medieval**](https://en.wikipedia.org/wiki/Medieval) times and the [**Victorian era**](https://en.wikipedia.org/wiki/Victorian_era), combined with microscopic examination. The presence of juices upon breaking, bruising reactions, odors, tastes, shades of color, habitat, habit, and season are all considered by both amateur and professional mycologists. Tasting and smelling mushrooms carries its own hazards because of poisons and [**allergens**](https://en.wikipedia.org/wiki/Allergens). Chemical [**tests**](https://en.wikipedia.org/wiki/Chemical_test) are also used for some genera

It is better to have a look at the macroscopic structure of mushrooms . Our Project involves analysis of the data containing information about macroscopic properties of mushrooms.

## 

## About The Project :

This project is about identifying the macroscopic properties of mushrooms and analyzing data about those properties . We will try to fit Machine learning models to help us identify mushrooms, that are safe to eat. It will also help us uncovering few myths about mushrooms like:

* *All white mushrooms are safe to eat.* I have actually heard people say this and it may be the most erroneous myth of all. Not all poisonous mushrooms are brightly colored. The destroying angel from earlier in this page is just one glaring example of a hazardous white mushroom.
* *Heating a poisonous mushroom and stirring it with a silver spoon will turn the spoon black.* Some believe that the toxins will blacken silver when heated. This myth has been around for a long time with no basis in fact. To date no toxins are known to have this reaction with silver.
* *Any mushroom is safe to eat once thoroughly cooked.* This is an unreliable assumption. Most toxins are not broken down by heat and are not made safer by cooking.
* *Insects can tell which mushrooms are poisonous and will avoid them*. Not true! Just because it's deadly to us doesn't mean it's deadly to a bug. Some toxic species, such as the death cap, will still harbor insects and their larvae.
* *They taste bitter/sour/bad.* Don't rely on those tricky taste buds! I've read reports of people mistakenly eating amanitas and saying they tasted good.
* *All poisonous mushrooms have a pointed cap.* The shape of any part of a mushroom plays no role in its toxicity.

## Importing the data:

mushrooms\_data<-read.csv("C:\\vik\\2017\\personal\\DSLA\\course material\\project 1 files\\mushrooms.csv",header = TRUE,sep = ",")

## Exploring the data

### Dimensions of the mushroom datasets are :

8124 rows/observations 23 columns/features

### Fields in the dataset are:

Following are the fields in the dataset that are provided:

## [1] "class" "cap.shape"   
## [3] "cap.surface" "cap.color"   
## [5] "bruises" "odor"   
## [7] "gill.attachment" "gill.spacing"   
## [9] "gill.size" "gill.color"   
## [11] "stalk.shape" "stalk.root"   
## [13] "stalk.surface.above.ring" "stalk.surface.below.ring"  
## [15] "stalk.color.above.ring" "stalk.color.below.ring"   
## [17] "veil.type" "veil.color"   
## [19] "ring.number" "ring.type"   
## [21] "spore.print.color" "population"   
## [23] "habitat"

### Following are the definitions of these fields:

* Fields/Attributes/features of the dataframe are
  + classes: edible=e, poisonous=p
  + cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
  + cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s
  + cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r,pink=p,purple=u,red=e,white=w,yellow=y
  + bruises: bruises=t,no=f
  + odor: almond=a,anise=l,creosote=c,fishy=y,foul=f,musty=m,none=n,pungent=p,spicy=s
  + gill-attachment: attached=a,descending=d,free=f,notched=n
  + gill-spacing: close=c,crowded=w,distant=d
  + gill-size: broad=b,narrow=n
  + gill-color: black=k,brown=n,buff=b,chocolate=h,gray=g,green=r,orange=o,pink=p,purple=u,red=e,white=w ,yellow=y
  + stalk-shape: enlarging=e,tapering=t
  + stalk-root: bulbous=b,club=c,cup=u,equal=e,rhizomorphs=z,rooted=r,missing=?
  + stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s
  + stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s
  + stalk-color-above-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y
  + stalk-color-below-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y
  + veil-type: partial=p,universal=u
  + veil-color: brown=n,orange=o,white=w,yellow=y
  + ring-number: none=n,one=o,two=t
  + ring-type: cobwebby=c,evanescent=e,flaring=f,large=l,none=n,pendant=p,sheathing=s,zone=z
  + spore-print-color: black=k,brown=n,buff=b,chocolate=h,green=r,orange=o,purple=u,white=w,yellow=y
  + population: abundant=a,clustered=c,numerous=n,scattered=s,several=v,solitary=y
  + habitat: grasses=g,leaves=l,meadows=m,paths=p,urban=u,waste=w,woods=d

In the sample provided , we see that 51.79% of the mushrooms are edible , but this is a sample is not the representative of the whole population. That is why the stratified sampling and cross validation is even more important for this data

## e p   
## 51.79714 48.20286

## Machine Learning

We will follow following steps to decide about the classification model:

1. split the data in train set and test set
2. train the model on the train set
3. check the efficiency of the model on the train set
4. predict the classification of the test set data
5. check the efficiency of the model on the test set

We will iterate these steps for different models and then compare the efficiencies of different models to choose the best model. We chose caret package in r to do the modeling as it streamlines the model training process

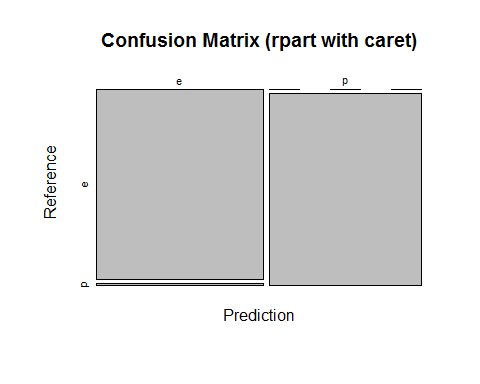
The caret package (short for classification and regression training) contains functions that help to streamline the model training process for complex regression and classification problems. The package utilizes a number of R packages but tries not to load them all at package start-up . We will also use the cross validation to try to do is balance the amount of bias in our models -- that's how well the model performs on the training data -- with the amount of variance in our model -- that's how well the model does on new data

## Model I : rpart:

We have trained the model on the trainset , we will also do the 10 fold cross validation and repeat 4 times. After the model is defined we find the predictions using the model for the test set and find out how well the model did. We follow the same process for all the models and thanks to caret package, training the models are more convenient than ever. Trees (also called decision trees, recursive partitioning) are a simple yet powerful tool in predictive statistics. The idea is to split the covariable space into many partitions and to fit a constant model of the response variable in each partition. In case of regression, the mean of the response variable in one node would be assigned to this node. The structure is similar to a real tree (from the bottom up): there is a root, where the first split happens. After each split, two new nodes are created (assuming we only make binary splits). Each node only contains a subset of the observations. The partitions of the data, which are not split any more, are called terminal nodes or leafs. This simple mechanism makes the interpretation of the model pretty easy.

Following is the performance statistics of this model :

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction e p  
## e 841 8  
## p 0 775  
##   
## Accuracy : 0.9951   
## 95% CI : (0.9903, 0.9979)  
## No Information Rate : 0.5179   
## P-Value [Acc > NIR] : < 2e-16   
##   
## Kappa : 0.9901   
## Mcnemar's Test P-Value : 0.01333   
##   
## Sensitivity : 1.0000   
## Specificity : 0.9898   
## Pos Pred Value : 0.9906   
## Neg Pred Value : 1.0000   
## Prevalence : 0.5179   
## Detection Rate : 0.5179   
## Detection Prevalence : 0.5228   
## Balanced Accuracy : 0.9949   
##   
## 'Positive' Class : e   
##

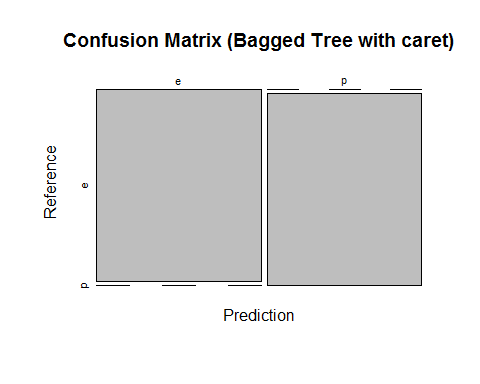


**##Model II : Decision Tree and Bagging:**

Bagging is one of the methods for improving the performance of weak learners such as Trees. Classification trees are adaptive and robust. So we create multiple trees by considering all variables at split and then averaging them for results. Training the model with the same cross validation options like last model and checking the performance :

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction e p  
## e 841 0  
## p 0 783  
##   
## Accuracy : 1   
## 95% CI : (0.9977, 1)  
## No Information Rate : 0.5179   
## 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 : 0.5179   
## Detection Rate : 0.5179   
## Detection Prevalence : 0.5179   
## Balanced Accuracy : 1.0000   
##   
## 'Positive' Class : e   
##

plot(mushroom\_tbl\_crt\_bag\_test$table,main="Confusion Matrix (Bagged Tree with caret)")

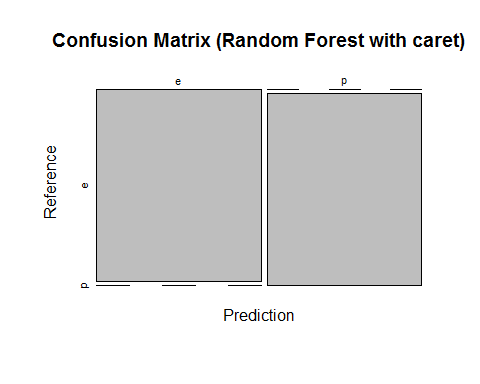


## Model III : RandomForest :

RandomForest is a specific kind of bagging trees , where trees are random and are not every split considers all the variables. Training the model with the same cross validation options as last models:

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction e p  
## e 841 0  
## p 0 783  
##   
## Accuracy : 1   
## 95% CI : (0.9977, 1)  
## No Information Rate : 0.5179   
## 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 : 0.5179   
## Detection Rate : 0.5179   
## Detection Prevalence : 0.5179   
## Balanced Accuracy : 1.0000   
##   
## 'Positive' Class : e   
##

plot(mushroom\_tbl\_crt\_rf\_test$table,main="Confusion Matrix (Random Forest with caret)")

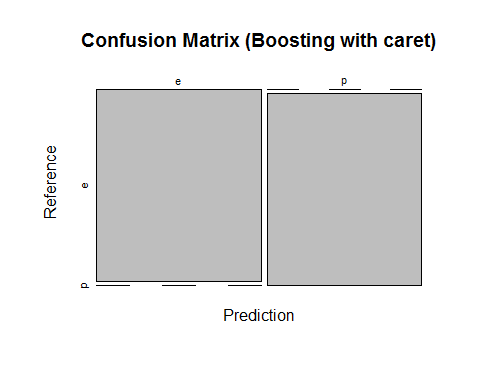


## Model III : Boosting :

Boosting is classification , in which trees are sequentially grown and every subsequent tree learns from the results of last tree. Training the model with the same cross validation options as last models:

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction e p  
## e 841 0  
## p 0 783  
##   
## Accuracy : 1   
## 95% CI : (0.9977, 1)  
## No Information Rate : 0.5179   
## 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 : 0.5179   
## Detection Rate : 0.5179   
## Detection Prevalence : 0.5179   
## Balanced Accuracy : 1.0000   
##   
## 'Positive' Class : e   
##

plot(mushroom\_tbl\_crt\_boost\_test$table,main="Confusion Matrix (Boosting with caret)")



## Model Comparison:

At this stage , we compare the models for their performance to decide about the models to follow:

##   
## Call:  
## summary.resamples(object = comparison)  
##   
## Models: rpart, bagging, randomforest, boosting   
## Number of resamples: 10   
##   
## Accuracy   
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  
## rpart 0.9892 0.9927 0.9946 0.9945 0.9965 0.9985 0  
## bagging 0.9985 1.0000 1.0000 0.9998 1.0000 1.0000 0  
## randomforest 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 0  
## boosting 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 0  
##   
## Kappa   
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  
## rpart 0.9784 0.9853 0.9892 0.9889 0.9931 0.9969 0  
## bagging 0.9969 1.0000 1.0000 0.9997 1.0000 1.0000 0  
## randomforest 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 0  
## boosting 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 0

dotplot(comparison)

