Data Science 'Mushrooms' Capstone Project

András Gelencsér

14 December 2019

Overview

This document describes the result of the capstone project for the HarvardX Data Science course based on the Mushrooms dataset avaiable in the UCI Machine Learning Repository (Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.)

The "Mushrooms" dataset was recorded down from The Audubon Society Field Guide to North American Mushrooms (1981). G. H. Lincoff (Pres.), New York: Alfred A. Knopf and is avaiable on the https://archive.ics.uci.edu/ml/datasets/mushroom url.

The goal of this project is to develop a model based on data science and machine learning techniques to predict if a mushroom is eatable based on the known attributes.

The dataset contains 8124 records and uses 22 attributes to describe the mushrooms. Each dataset contains additional attribute to flag if the mushroom is eatable or not.

We will use 80% of the avaiable data for train our model. The reamining 20% will be used as test dataset for evaluating the developed model.

For the developed model is very important to detect poisonus mushrooms to avoid possible health damages or death. Therefore we will evluate the model performance with the F1 score instead of the overal accuracy, beacuse it's very important to reach high sensitivity with our model.

The following steps were done during the project for developing the algorithm:

1. Download the data and prepare for the analysis
2. Analyze the data structure
3. Define the classification model based on the result the data analysis
4. Implement and train the model based on the train set
5. Review the model based on the test set

The final model performed an F1 score of 1 by using a knn prediction model.

Methods & analysis

Preparation

For the analysis the data we will use the following R packages: * tidyverse * caret * gridExtra * ggplot2 * Rborist

the following R code loads and install the required packages if needed:

```
if(!require(tidyverse)) install.packages("tidyverse",
                         repos = "http://cran.us.r-project.org", dependencies = TRUE)
if(!require(caret)) install.packages("caret",
                         repos = "http://cran.us.r-project.org", dependencies = TRUE)
if(!require(gridExtra)) install.packages("gridExtra",
                         repos = "http://cran.us.r-project.org", dependencies = TRUE)
if(!require(ggplot2)) install.packages("ggplot2",
                         repos = "http://cran.us.r-project.org", dependencies = TRUE)
```

```r
if(!require(Rborist)) install.packages("Rborist",
                         repos = "http://cran.us.r-project.org", dependencies = TRUE)
```

At first, we need to prepare the data set for the analysis. The analysis is based on the reduced movielens data set including approximately 1,000,000 rating data.

The following R code downloads the data from the https://archive.ics.uci.edu/ml/datasets/mushroom site. To avoid unnecessary data traffic, the data will be downloaded only if not done yet.

```r
#create directory for the data file if necessary
if (!dir.exists("mushrooms")){
  dir.create("mushrooms")
}
baseurl = "https://archive.ics.uci.edu/ml/machine-learning-databases/mushroom/"
files_to_download = c("agaricus-lepiota.data", "agaricus-lepiota.names")

for (f in files_to_download){
  #download the files only if not done yet
  if (!file.exists(paste("./mushrooms/",f,sep='')))
  {
    download.file(paste(baseurl,f,sep=''), paste("./mushrooms/",f,sep=''))
  }
}

fl <- file("mushrooms/agaricus-lepiota.data")
mushroom_data <- str_split_fixed(readLines(fl), ",", 23)
close(fl)
```

The dataset is stored in the "agaricus-lepiota.data" file. This file contains 23 columns separated via comma. There is no header line in the file. The description file "agaricus-lepiota.names" gives us information about the classes and the 22 attributes in the data file.

So we can extract the data by parsing the data file. The following R code reads the data and stores in a data frame:

```r
fl <- file("mushrooms/agaricus-lepiota.data")
mushroom_data <- str_split_fixed(readLines(fl), ",", 23)
close(fl)

#set the column names for the features
colnames(mushroom_data) <- c("classes", "cap-shape", "cap-surface", "cap-color",
                             "bruises?", "odor", "gill-attachment",
                             "gill-spacing", "gill-size", "gill-color", "stalk-shape",
                             "stalk-root", "stalk-surface-above-ring",
                             "stalk-surface-below-ring", "stalk-color-above-ring",
                             "stalk-color-below-ring", "veil-type", "veil-color",
                             "ring-number", "ring-type", "spore-print-color",
                             "population", "habitat")
#convert to data frame
mushroom_data <- as.data.frame(mushroom_data)

#remove temporary file variable
rm(fl)
```

For training the classification method we will use 80% of the avaiable data as training set. The remaining data will be used for evaluating the developed method. The following R code creates the train and test set:

```r
#initialize random sequenz
set.seed(1, sample.kind = "Rounding")
#create index for train and test set
#20% of the data will be used for the test set
test_idx = createDataPartition(y = mushroom_data$classes, times=1, p=0.2, list=FALSE)
train_data = mushroom_data[-test_idx,]
test_data = mushroom_data[test_idx,]
#remove temporary variables
rm(mushroom_data, test_idx)
```

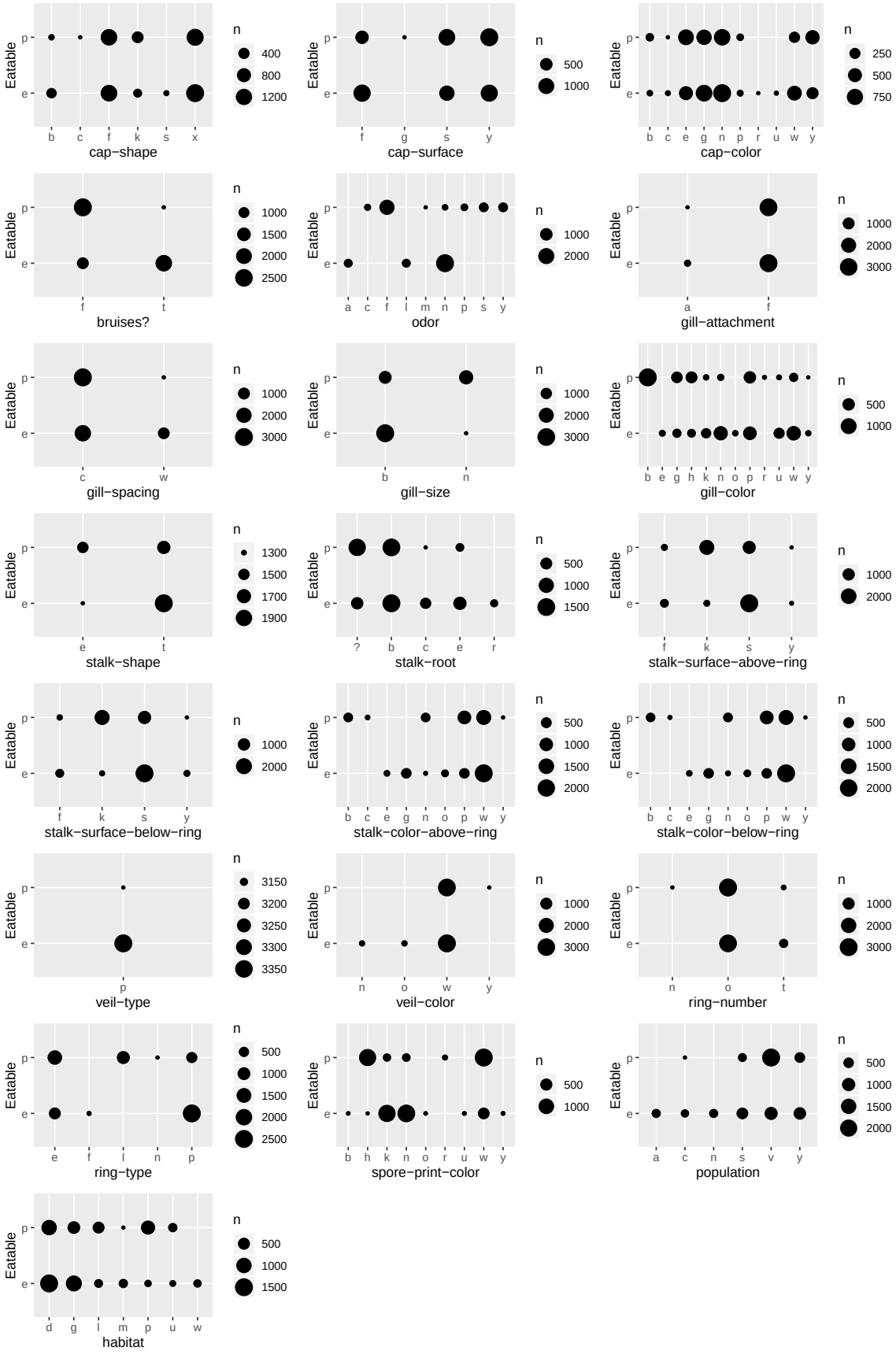
Available Data

For the analysis we have data about 7,000 mushrooms. All the features are categorical with different possible values.

1. cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
2. cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s
3. cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y
4. bruises?: bruises=t,no=f
5. odor: almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s
6. gill-attachment: attached=a,descending=d,free=f,notched=n
7. gill-spacing: close=c,crowded=w,distant=d
8. gill-size: broad=b,narrow=n
9. 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
10. stalk-shape: enlarging=e,tapering=t
11. stalk-root: bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=?
12. stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s
13. stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s
14. stalk-color-above-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
15. stalk-color-below-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
16. veil-type: partial=p,universal=u
17. veil-color: brown=n,orange=o,white=w,yellow=y
18. ring-number: none=n,one=o,two=t
19. ring-type: cobwebby=c,evanescent=e,flaring=f,large=l, none=n,pendant=p,sheathing=s,zone=z
20. spore-print-color: black=k,brown=n,buff=b,chocolate=h,green=r, orange=o,purple=u,white=w,yellow=y
21. population: abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y
22. habitat: grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d

Feature analysis

```r
plots <- list()
#Plots all the attributes for eatable <-> poisious
for (i in 1:(ncol(train_data)-1))
{
  summarized_data <- train_data %>% group_by(classes, .[,i+1]) %>% summarise(n = n())
  names(summarized_data)[2] <- "attr"
  plot <- summarized_data %>% ggplot(aes(attr , classes)) + geom_point(aes(size=n)) +
          xlab(names(train_data)[i+1]) + ylab("Eatable")
  plots[[i]] <- plot
}
grid.arrange(grobs=plots,ncol=3)
```


Eatable
p
e
n
cap−shape
b c f k s x
400 800 1200
cap−surface
f g s y
500 1000
cap−color
b c e g n p r u w y
250 500 750
bruises?
f t
1000 1500 2000 2500
odor
a c f l m n p s y
1000 2000
gill−attachment
a f
1000 2000 3000
gill−spacing
c w
1000 2000 3000
gill−size
b n
1000 2000 3000
gill−color
b e g h k n o p r u w y
500 1000
stalk−shape
e t
1300 1500 1700 1900
stalk−root
? b c e r
500 1000 1500
stalk−surface−above−ring
f k s y
1000 2000
stalk−surface−below−ring
f k s y
1000 2000
stalk−color−above−ring
b c e g n o p w y
500 1000 1500 2000
stalk−color−below−ring
b c e g n o p w y
500 1000 1500 2000
veil−type
p
3150 3200 3250 3300 3350
veil−color
n o w y
1000 2000 3000
ring−number
n o t
1000 2000 3000
ring−type
e f l n p
500 1000 1500 2000 2500
spore−print−color
b h k n o r u w y
500 1000
population
a c n s v y
500 1000 1500 2000
habitat
d g l m p u w
500 1000 1500