Predicting Mushroom Edibility from Physical Characteristics

Ina Ding

Introduction

The primary goal of this project is to best predict whether or not a mushroom is poisonous using various physical characteristics, rarity, and habitat of the fungus. The data set consists of 8124 hypothetical mushroom samples, constructed from the Audobon Society Field Guide. The samples correspond to 23 exisitng species of mushrooms from the Agaricus and Lepiota Families. Each mushroom is categorized as either poisonous or edible, with mushrooms 'not recommended for eating' or of unknown edibility counted as poisonous. Though the observations are hypothetical mushrooms, analyzing them can still provide beneficial results that can be applied to help identify the edibility of the near 14,000 existing species of mushrooms.

Some light data cleaning was done, including removing the veil-type variable because all observations had the same veil type. The "poisonous" variable was changed from from p/e to 1/0 for ease of fitting with models like logistic regression models.

Data Dictionary: 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
```

Citations Mushroom. (1987). UCI Machine Learning Repository. https://doi.org/10.24432/C5959T.

```r
library(ggplot2)
mushrooms <- read.csv("Mushrooms - Sheet1.csv")
mushrooms$poisonous[mushrooms$poisonous == 'p'] <- 'Poisonous'
mushrooms$poisonous[mushrooms$poisonous == 'e'] <- 'Edible'

no_veil <- subset(mushrooms, select = -c(veil))
no_veil$poisonous[no_veil$poisonous == 'Poisonous'] <- 1
no_veil$poisonous[no_veil$poisonous == 'Edible'] <- 0
no_veil$poisonous <- as.numeric((no_veil$poisonous))
```

```r
table(no_veil$poisonous)
```

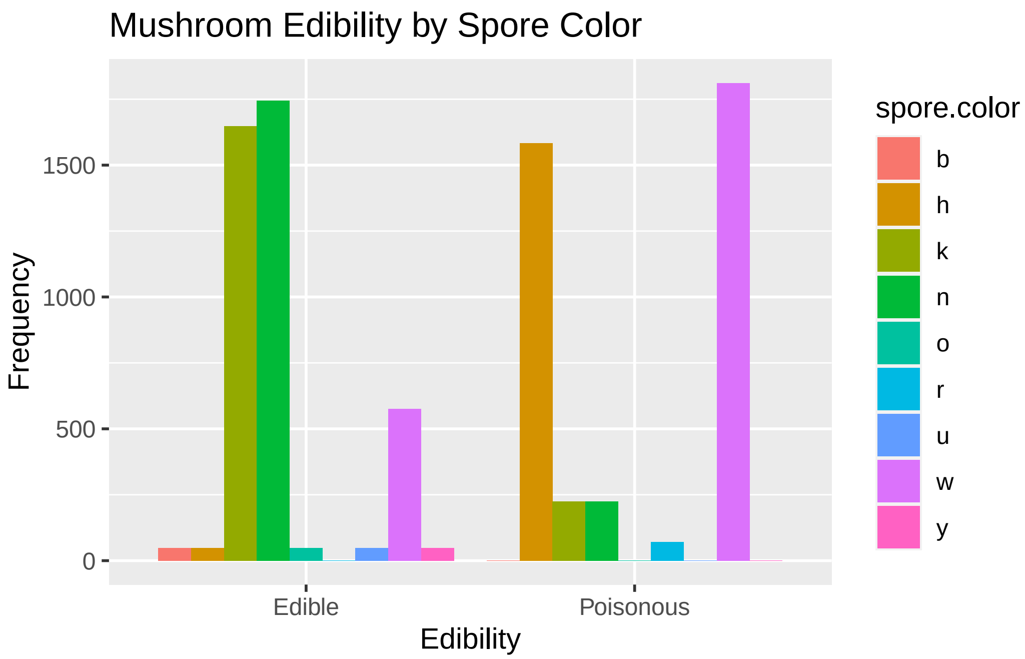
```
   0    1 
4208 3916 
```

Roughly half (48.2%) of the mushrooms in the data set are poisonous, providing motivation for our model, as otherwise the chances of predicting correctly would be roughly a 50/50 draw.

Certain variables are much more indicative of mushroom edibility than others.

```r
spore <- with(mushrooms, table(spore.color, poisonous))

ggplot(as.data.frame(spore), aes(factor(poisonous), Freq, fill = spore.color)) +
  geom_col(position = 'dodge') +
  labs(title="Mushroom Edibility by Spore Color", x="Edibility",y="Frequency")
```



Key:

buff=b

chocolate=h

black=k

brown=n

orange=o

green=r

purple=u

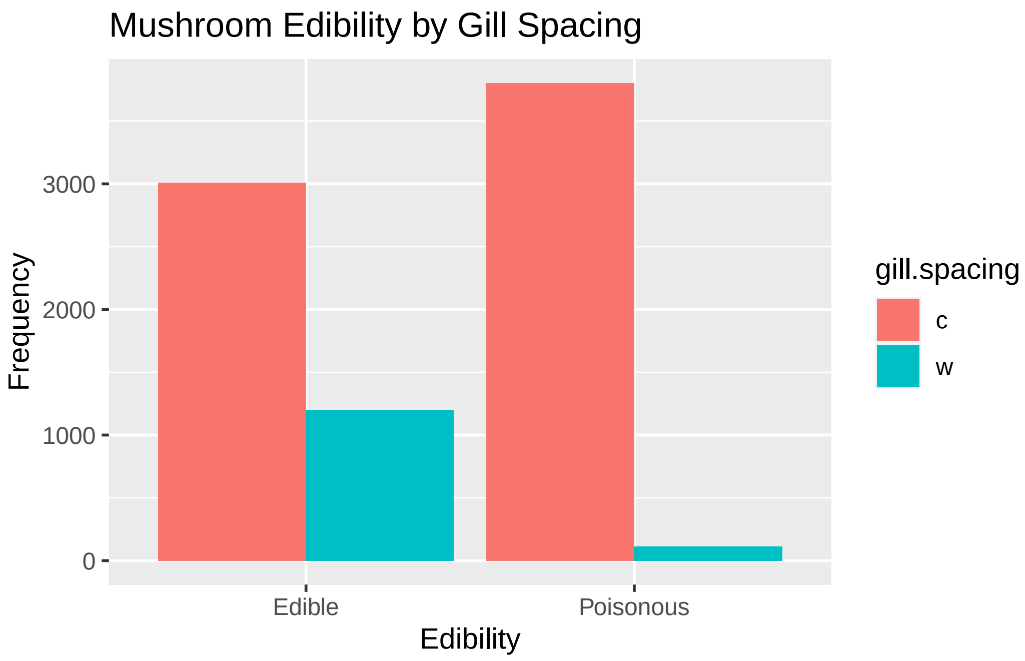
white=w

yellow=y

Mushrooms with buff, black, brown, orange, purple, and yellow spores are nearly all edible while mushrooms with chocolate, green,and white spores are mostly poisonous, making spore color a good solo indicator of edibility.

Mushroom odor is also a good predictor of edibility (graph not included as to conserve space). Mushrooms with an almond, anise and no odor are nearly all edible, while mushrooms smelling of creosote, foul, pungent, spicy, or fishy are nearly all poisonous.

```
gill_spacing <- with(mushrooms, table(gill.spacing, poisonous))
ggplot(as.data.frame(gill_spacing), aes(factor(poisonous),
                                        Freq, fill = gill.spacing)) +
  geom_col(position = 'dodge') +
  labs(title="Mushroom Edibility by Gill Spacing", x = "Edibility", y="Frequency")
```



Key:

close=c

crowded=w

While odor and spore color are both pretty good predictors of edibility all around, other variables like gill spacing are only edibility indicators for specific levels of the varaible. That is, mushrooms with wide gill spacing are a good indicator that a mushroom is edible, but knowing that a mushroom has close gill spacing is not really indicative of the mushroom's edibility.

Methodology

We were interested in including all the variables as predictors in our model. After some experimentation with LASSO and all subset variable selection models, there turned out to be so many variables and categories within each variable that any coefficients were shrunken to very small numbers—thus it seemed better to just keep all variables in.

The primary outcome of interest is whether or not a mushroom is poisonous or edible. Since this is a binary outcome using either a logistical regression model or classification tree seemed the most fitting.

For creating and testing the classification tree, the original data set was split into two halves (one for training and one for testing). With the training data set, multiple complexity parameter (cp) levels ranging from 0.01 to 0.0001 were tested. The cp level determines how "pruned" the classification tree is, where smaller cp levels render larger trees. cp levels between 0.04-0.07 seemed to render the most accurate classification trees, with sensitivity of 0.48, specificity of 0.55, positive predictive value of 0.5, and negative predictive value of 0.53. Seeing as there are only two possible classifications (poisonous or edible), however, the outcomes of our classification tree are not all that better than just randomly categorizing a mushroom as poisonous or edible.

On the other hand, the log-odds model was able to predict mushroom edibility with 100% accuracy when the discrimination threshold was set to 0.5, and was the final model chosen. In terms of meeting assumptions, the notion of linearity does not apply as all of our predictors are categorical. The independence assumption is not necessarily met. While the exact process of which the hypothetical mushroom observations were created is unclear, we do know that the mushrooms were designed after 23 existing mushrooms. Based on this information, it is likely that mushrooms designed off of the same real mushroom likely have some similar attributes thus violating the independence assumption. However, this violation of independence is not necessarily a bad thing——the whole point of the model (to predict edibility) is to test whether or not similarly looking mushrooms may have similar poisonous status', and it is likely that mushrooms based off the same mushroom will have the same poisonous/edible characteristic. For this reason the log-odds model was still chosen as the final model and is still believed to be able to provide useful information. No variable interactions or transformations were made in the final model.

Results

The following is our final logistical regression model that uses all variables in the data set as predictors (besides the veil-type variable, which as mentioned all mushrooms had the same characteristic for).

```
library(tidymodels)
library(tidyverse)
library(leaps)
library(glmnet)
```

```
logodds <- glm(poisonous ~ .,
          data = no_veil,
          family = "binomial")

tidy(logodds)
```

```
# A tibble: 96 x 5
   term          estimate std.error statistic p.value
   <chr>            <dbl>     <dbl>     <dbl>   <dbl>
 1 (Intercept)  -2.66e+ 1   308827. -8.60e- 5    1.00
 2 cap.shapec    1.94e- 7   206186.  9.43e-13    1.00
 3 cap.shapef    1.72e-10    23436.  7.35e-15    1.00
 4 cap.shapek   -5.10e-10    25480. -2.00e-14    1.00
 5 cap.shapes    1.48e-10    80322.  1.84e-15    1.00
 6 cap.shapex    1.24e-10    22453.  5.52e-15    1.00
 7 cap.surfaceg  2.28e- 9   252069.  9.06e-15    1.00
 8 cap.surfaces -9.70e-12    13416. -7.23e-16    1.00
 9 cap.surfacey  1.94e-11    11243.  1.72e-15    1.00
10 cap.colorc    1.38e- 9    77867.  1.77e-14    1.00
# ... with 86 more rows
```

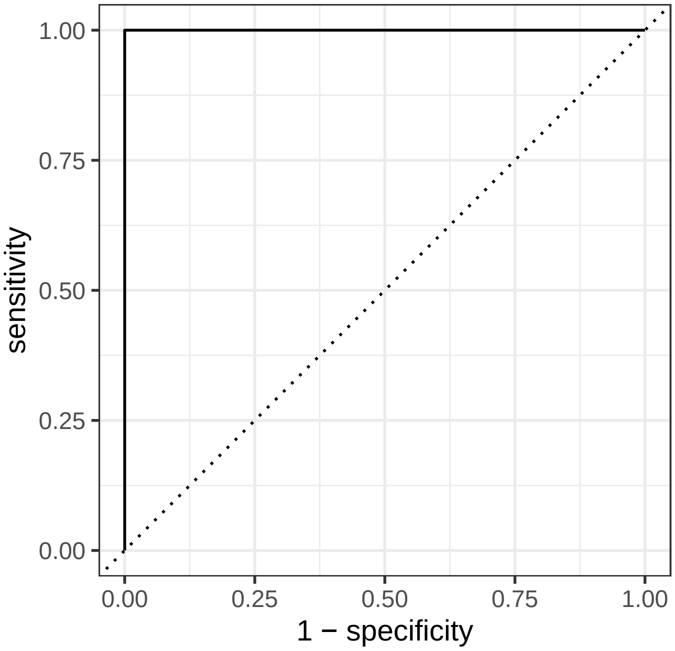
```
logodds_aug <- augment(logodds)

logodds_aug <- logodds_aug %>%
  mutate(prob = exp(.fitted)/(1 + exp(.fitted)),
         pred_pois = ifelse(prob > 0.5, "Poisonous", "Edible")) %>%
  select(.fitted, prob, pred_pois, poisonous)

table(logodds_aug$pred_pois, logodds_aug$poisonous)
```

```
               0    1
Edible      4208    0
Poisonous      0 3916
```

```
logodds_aug %>%
  roc_curve(
    truth = as.factor(poisonous),
    prob,
    event_level = "second"
  ) %>%
  autoplot()
```



```
logodds_aug %>%
  roc_auc(
    truth = as.factor(poisonous),
    prob,
    event_level = "second"
  )
```

```
# A tibble: 1 x 3
```

```
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary             1
```

Evident by both the table and ROC curve, our model can perfectly predict a mushroom's edibility. Consequently, the model's sensitivity, specificity, positive predictive value, and negative predictive value are all 1. Given the variables in the mushroom data set, there is reason to believe that a mushroom's edibility can be predicted by its physical attributes.

Discussion

It is likely that our model has 'perfect' predicting power due to the sheer number of predictors and characteristics, which allows each mushroom to be close to unique. Given this, there is reason to believe that by knowing a mushroom's physical characteristics, you can determine whether or not the mushrooms is poisonous or edible. There are, however, limitations to the conclusions that our model can make. For one, as discussed earlier, the independence assumption for our data set is likely violated due to the fact that the mushrooms are all generated based on 23 species of real mushrooms. Given this, if mushrooms based off the same species have similar characteristics same edibility, then this makes it extremely easy to predict the edibility of mushrooms in our data set, and our model is more using the species of mushroom as a predictor as opposed to its specific physical characteristics. The fact that the data set only represents 23 species of mushrooms when there are believed to be over 27,000 species in total also makes our model not equipped to predict the edibility of all mushrooms.

Ideas for future work include further using a larger data set that either includes real mushrooms (as opposed to hypothetical ones) and more importantly a much wider range of mushroom species. With a more diverse sample size, key attributes or combinations of attributes may be identified to predict edibility so that predicting models don't have to include so many predictors. More exhaustive/thorough variable interaction exploration may also be able to reveal key identifiers of edibility. Lastly, using separate training and testing data sets would better test the usefulness of a predictor model. Of these all, however, I believe that using a larger and more diverse data set that includes a wider range of mushroom species would be the most useful for future work in predicting mushroom edibility.