Mushroom Edibility Analysis

Tofu-FC - Huiwen Wang, Rocky Zhang, Darrick Zhang

2024-10-28

Introduction

Project Motivation / Background:

Mushrooms are vital to the general wellness of the ecosystem, decomposing and recycling the nutrients in the soil. Mushrooms also provide a valuable food source full of nutrients for human beings and other important organisms. However, some mushroom species can also be poisonous and harmful.

The importance of this research has been highlighted in a multitude of studies. Take this quote, for example:

> The ingestion of wild and potentially toxic mushrooms is common in the United States, with poison centers logging cases in the National Poison Data System (NPDS) for over 30 years. From 1999 to 2016, there were 133,700 reported cases of mushroom exposure, mostly unintentional and involving children under six years old. While the majority of cases resulted in no or minor harm, there were 704 instances of major harm and 52 fatalities, primarily due to cyclopeptide-producing mushrooms ingested unintentionally by older adults. Misidentification of edible mushroom species is a common cause of poisoning and may be preventable through education (Brandenburg and Ward 2018).

As shown by studies and other similar studies, accurate classification of mushrooms is crucial for preventing poisoning incidents. Many toxic mushroom species closely resemble edible varieties, making it easy for foragers to misidentify them. Thus, our research will focus on what physical features and environmental factors of mushrooms humans can use to identify toxic/poisonous mushrooms in the wild. By conducting a research study on how to distinguish between safe and dangerous species, we can mitigate the incidence of mushroom poisoning and ensure safer foraging practices.

Research Question:

What environmental factors and/or physical features of mushrooms indicate that a wild mushroom is poisonous?

Hypothesis:

Mushrooms in the wild with obvious physical features like white gills, white rings, red caps, or red stems tend to be poisonous. These obvious physical traits are more likely to be spotted by animals, which would provide an evolutionary disadvantage unless they contain certain self-defense mechanisms, such as poison or toxins. Additionally, the habitat and season in which mushrooms are planted and grow may also affect whether they're poisonous. Different temperatures, humidity, and light can affect the production of toxins, which may also affect the edibility of mushrooms.

Data Description:

The data was curated on April 26, 1987, and submitted to the UCI by the National Audubon Society Field Guide. The National Audubon Society conducted extensive field research throughout North America, recording their observations on various aspects of mushrooms. Their research incorporate a wide range of physical characteristics, including size, shape, color, and texture of the mushrooms. Additionally, they documented environmental factors such as the type of habitat and seasonal variations. Importantly, the study also focused on the toxicity of the mushrooms, noting which species were poisonous. This comprehensive dataset provides valuable insights into the relationship between mushrooms and their environments, contributing significantly to the understanding of the factors influencing mushroom toxicity.

Our response variable is `class`, which is a qualitative variable labeled "e" for edible or "p" for poisonous.

Key quantitative predictor variables include `stem.height` and `stem.width`, the height (cm) and width (cm) of the stem of the mushroom. We are also interested in `cap.diameter`, the diameter of the mushroom cap (cm).

Key qualitative predictor variables include `cap.shape`, the shape of the mushroom cap; `gill.color`, the color of the fungi gills, `stem.color`, the color of the mushroom stem; `habitat`, the habitat that the mushroom is grown/found; and `season`, the season that the mushroom is grown/found (spring=s, summer=u, autumn=a, winter=w).

The `cap.shape` codes are:

b	c	x	f	s	p	o
bell	conical	convex	flat	sunken	spherical	others

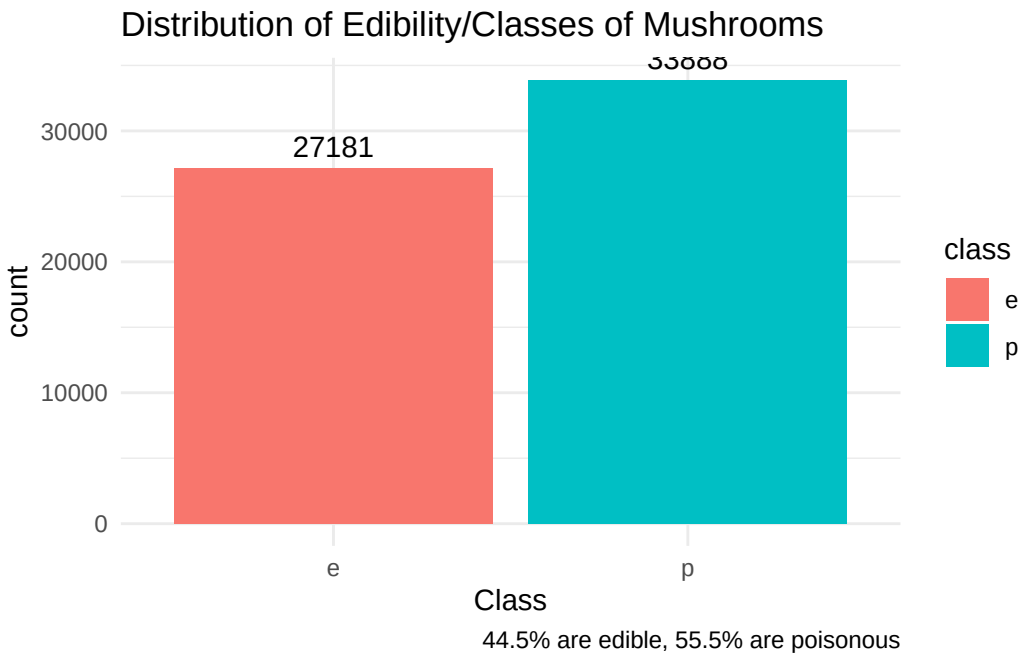
The `gill.color` and `stem.color` are:

n	b	g	r	p	u	e	w	y	l	o	k	f
brown	buff	gray	green	pink	purple	red	white	yellow	blue	orange	black	none

The `habitat` codes are:

g	l	m	p	h	u	w	d
grasses	leaves	meadows	paths	heaths	urban	waste	woods

Exploratory Data Analysis



Looking at the overall distribution of our response variable `class`, most of the mushrooms in our dataset seem to be poisonous ("p"). 33888 of the observations, or 55.5% of them are labeled poisonous, as opposed to 27181 (44.5%) of them as edible.

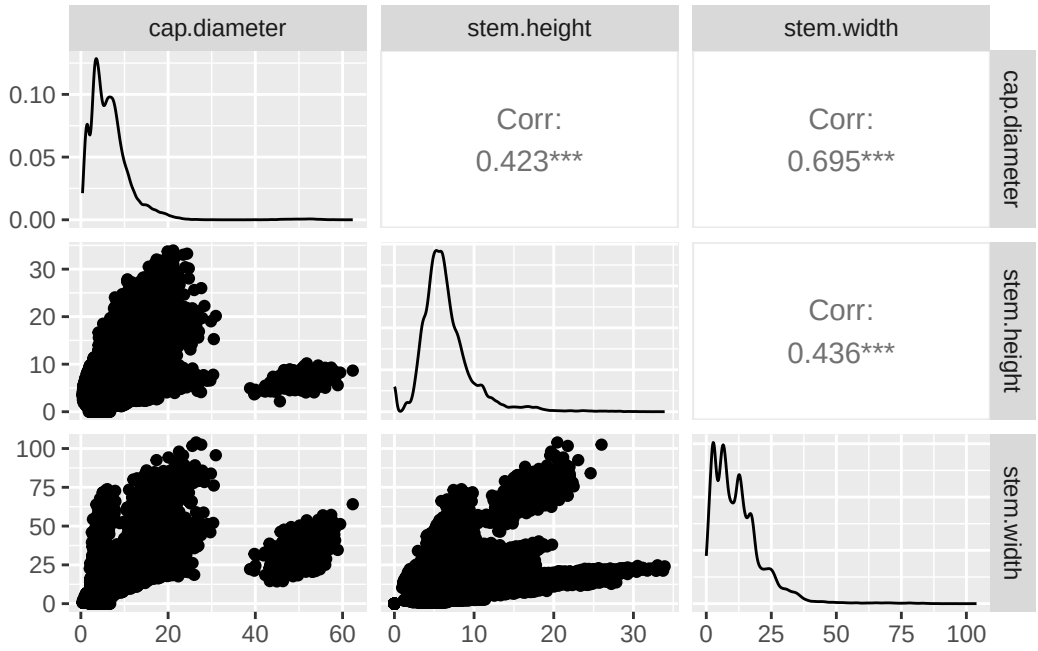


Table 4: cap diameter summary statistics

min	q1	median	q3	max	mean	sd
0.38	3.48	5.86	8.54	62.34	6.734	5.265

Table 5: stem height summary statistics

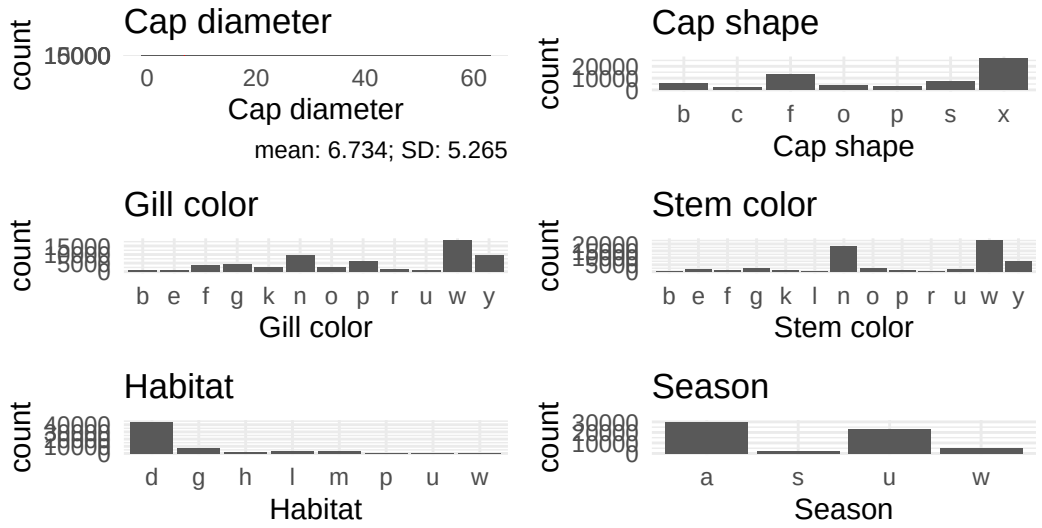
min	q1	median	q3	max	mean	sd
0	4.64	5.95	7.74	33.92	6.582	3.37

Table 6: stem width summary statistics

min	q1	median	q3	max	mean	sd
0	5.21	10.19	16.57	103.91	12.149	10.036

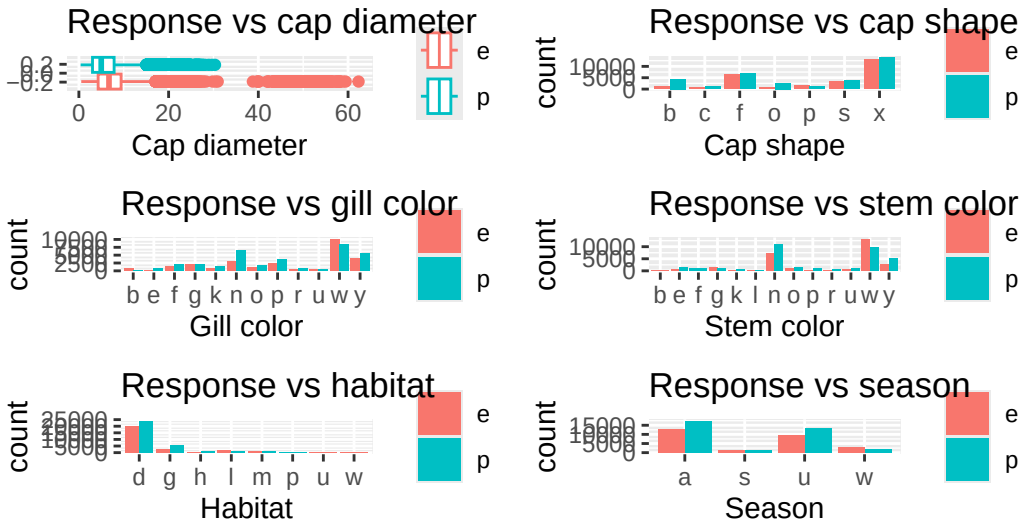
We were first interested in seeing the relationship between our continuous, quantitative variables. While not totally linear, with correlations of 0.423 and 0.436, they graphs seem to be somewhat linear, with some redundancy in their information. Thus, for the rest of the EDA we will focus mainly on visualizing `cap.diameter`. We may consider adding them back for the

final model, but were more interested in seeing some of the EDA with the categorical variables. The mean cap diameter is 6.734cm, with a SD of 5.265cm. The mean stem height is 6.582cm, with a SD of 3.37cm. The mean stem width is 12.149cm, with a SD of 10.036cm.



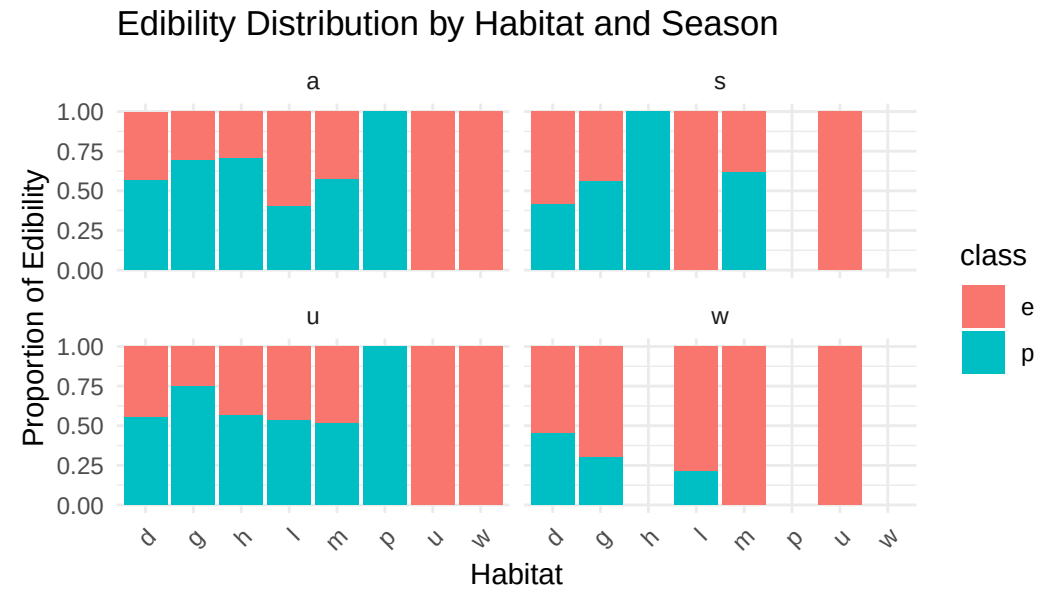
cap shape: bell=b, conical=c, convex=x, flat=f, sunken=s, spherical=p, others=o
colors: brown=n, buff=b, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y, blue=l, orange=o, black=k
habitat: grasses=g, leaves=l, meadows=m, paths=p, heaths=h, urban=u, waste=w, woods=d
season: spring=s, summer=u, autumn=a, winter=w

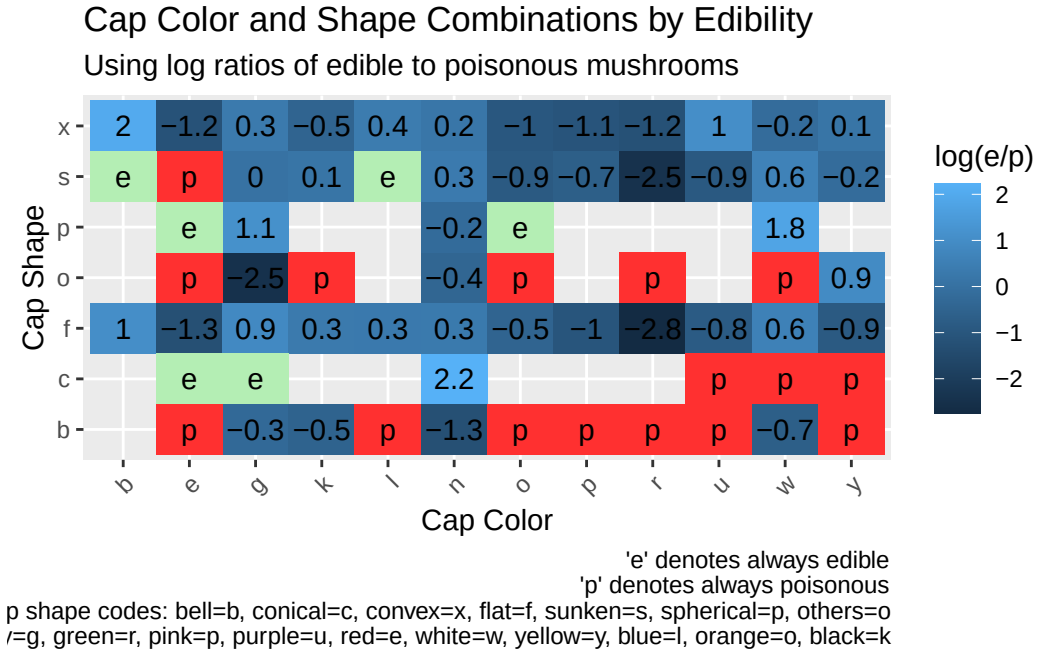
To visualize the distribution of some of our predictor variables, we use a histogram for our continuous variable `cap.diameter` and bar graphs for our categorical variables. The distribution of cap diameter seems to be unimodal, skewed right. From the For qualitative variables, there appears to be more common physical and environmental characteristics. For cap shape, flat and convex tends to be the most common; for stem root the most common was "missing data" (which likely suggests that this may be a variable to remove); for stem color the most common is white, yellow, and brown; for habitat, woods is the most common (see above for code meanings); for season, autumn and summer tends to be the most common.



cap shape: bell=b, conical=c, convex=x, flat=f, sunken=s, spherical=p, others=o
colors: brown=n, buff=b, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y, blue=l, orange=o, black=k
habitat: grasses=g, leaves=l, meadows=m, paths=p, heaths=h, urban=u, waste=w, woods=d
season: spring=s, summer=u, autumn=a, winter=w

The bivariate exploratory data analysis also shows some interesting findings for predictors. In particular, categories that have a disparity between the two different classes could offer potential modeling power in classification of the classes (toxicity). Larger, more extreme cap diameter is often linked to edibility. Cap shape of convex, bell, and others is more likely to be poisonous/toxic than edible. We also see that gill color of brown and yellow tends to be poisonous. For stem color, yellow and brown tends to be more poisonous than not. And lastly the habitat of woods and the season of autumn and summer also observes a similar observation.



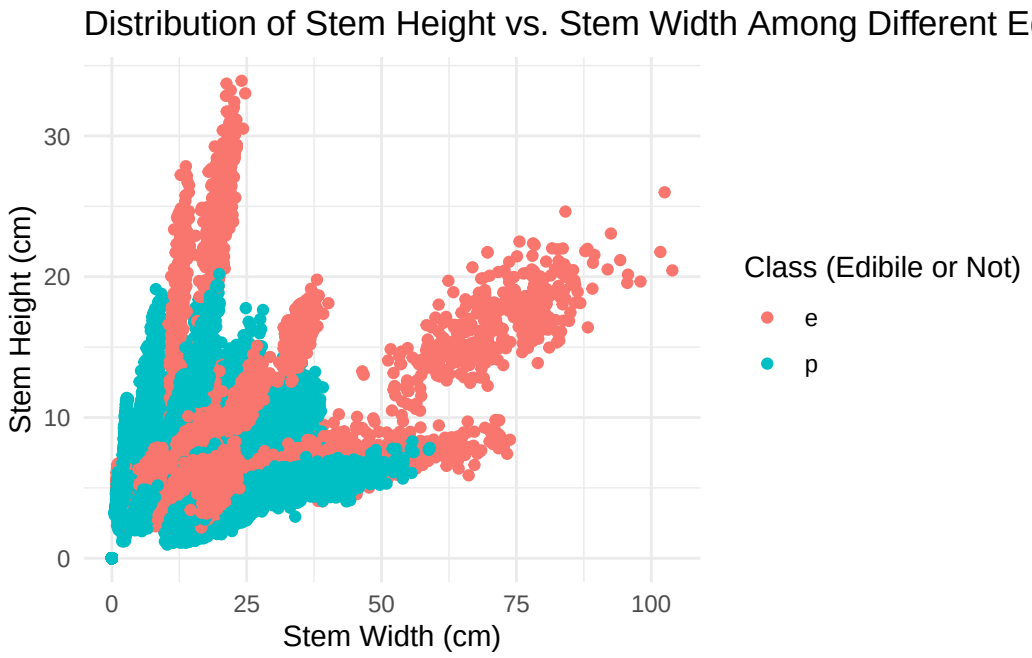


We believe that these predictors may have potential interactive effects that could help us with our model to predict edibility of mushrooms due to the fact that certain more specific characteristics tends to be poisonous.

Habitat $\times$ **Season**: Our EDA reveals that mushrooms in certain habitats might only be

edible during specific seasons. For example, mushrooms in the meadows are edible in the winter, but may be poisonous in other seasons.

Cap color × Cap shape: Certain combinations of cap color and cap shape are always edible or poisonous. Additionally, the log ratios across combinations of cap colors and cap sizes are varied with no pattern – for a mushroom with a sunken cap shape, it could be always edible (if the color is buff) to always poisonous (if the color is red). Similarly, if a mushroom is brown, it could be high likely it is edible (if the cap shape is conical) or likely it is poisonous (if the cap shape is bell).



Finally, we look at multivariate data analysis including 2 predictors and our response variable. Here, we visualize the effect of both stem width and stem height on the response variable, `class`. Interestingly, it seems like mushrooms with either high stem width or stem height seem to be edible. This suggests there may be some potential interaction effects between stem height and stem width – the low value of one alone does not seem to predict if the mushroom is poisonous, but requires the low value of both.

```
[1] -41960.76
```

```
[1] -33818.07
```

```
[1] 16285.38
```

```
[1] 0
```

The p-value is small, so we reject the H_0. The data provide evidence of at least one non-zero model coefficient in the model.

term	df.residual	residual.deviance	df	deviance	p.value
class_binary ~ cap.diameter + season + cap.shape + cap.color + gill.color + stem.color + habitat	61017	67636.13	NA	NA	NA
class_binary ~ cap.diameter + season + cap.shape + cap.color + gill.color + stem.color + habitat + habitat * season + cap.shape * cap.color	60953	62378.33	64	5257.798	0

Since the p-value is low, we decide to include these interaction terms.

```
[1] 67740.13
```

```
[1] 62610.33
```

```
[1] 68209.16
```

```
[1] 63656.62
```

For both AIC and BIC, the model with the interaction effect does much better in comparison to the main model.

Initial modeling

Simple Model

$$\text{logit}(Odds(\text{class} = p)) = \beta_0 + \beta_1 \cdot \text{cap.diameter} + \beta_2 \cdot \text{season}$$

term	estimate	std.error	statistic	p.value

Tidy Output

term	estimate	std.error	statistic	p.value
(Intercept)	0.856	0.021	39.927	0.000
cap.diameter	-0.083	0.002	-34.223	0.000
seasons	-0.599	0.050	-11.951	0.000
seasonu	0.013	0.022	0.610	0.542
seasonw	-0.699	0.037	-18.820	0.000

Confusion Matrix

class binary: edible = 0, poisonous = 1

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2639 1757
         1 5515 8409

               Accuracy : 0.6031
                 95% CI : (0.5959, 0.6102)
    No Information Rate : 0.5549
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.158

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.3236
            Specificity : 0.8272
         Pos Pred Value : 0.6003
         Neg Pred Value : 0.6039
             Prevalence : 0.4451
         Detection Rate : 0.1441
   Detection Prevalence : 0.2400
      Balanced Accuracy : 0.5754
```

```
'Positive' Class : 0
```

Next steps + questions

- Determine the need of interaction terms
- Test more specified models
- Evaluate an efficient parsimonious model
- Optimize on p cutoff
- Sensitivity vs. specificity

References

Brandenburg, William E., and Karlee J. Ward. 2018. "Mushroom Poisoning Epidemiology in the United States." *Mycologia* 110 (4): 637–41. https://doi.org/10.1080/00275514.2018.1479561.