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:

With this in mind, we are interested to see whether there are any environmental factors and/or physical features of mushrooms that can help curious human avoid poisonous mushrooms that may grow in their yard or in the wilderness.

Hypotheses:

1. Mushrooms in the wild with obvious physical features like white gills, white rings, red caps, or red stems tend to be poisonous. We want to explore this hypothesis because mushrooms with 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.
2. The habitat and season of the mushrooms would likely affect whether they're poisonous. We form this hypothesis because in specific habitats, mushrooms might face intense competition for resources such as nutrients, light, or space. Producing toxins can give them a competitive edge by inhibiting the growth of other fungi or microorganisms. In addition, the growing season can influence the chemical composition of mushrooms. Different temperature, humidity, and light during various seasons can affect the production of secondary metabolites, including toxins. For instance, some mushrooms might produce more toxins during wet seasons to fend off increased microbial activity.

Data Description:

The dataset we used can be found here: Mushroom Dataset

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.

The dataset includes a lot of different characteristics and predictor variables, but many are irrelevant to our research question or induce multicollinearity in our models (for more info please visit the dataset website above). The predictor variables of interest and their description can be found below:

Stem height and width, height or width in cm of the stem of the mushroom, both are quantitative variables of initial interest that may help with physical identification.

Cap diameter, diameter of the cap of the mushroom in cm, it is a quantitative variable that could also help with physical identification.

Cap shape, shape of the cap of the mushroom, it is a qualitative variable that may help in the physical identification of edibility.

Type of stem root and color, physical descriptions of the stem representing a qualitative variable of interest.

Ring type, qualitative variable representing the physical appearance of the ring (typically on the upper part of the stem) of a mushroom.

Habitat and season, qualitative variables representing the habitat and season in which that mushroom is grown and found. These variables are used for environmental predictors for edibility.

Our response variable is class, which is a qualitative variable of two levels representing the mushroom edibility.

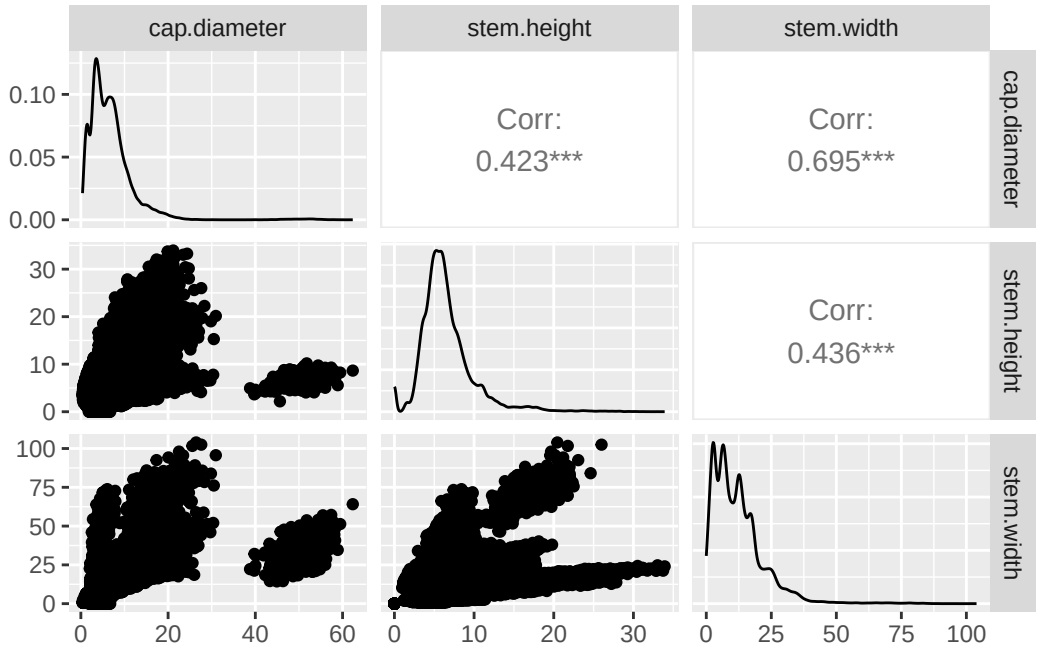
Exploratory Data Analysis

```
Number of rows with at least one missing value: 0
```

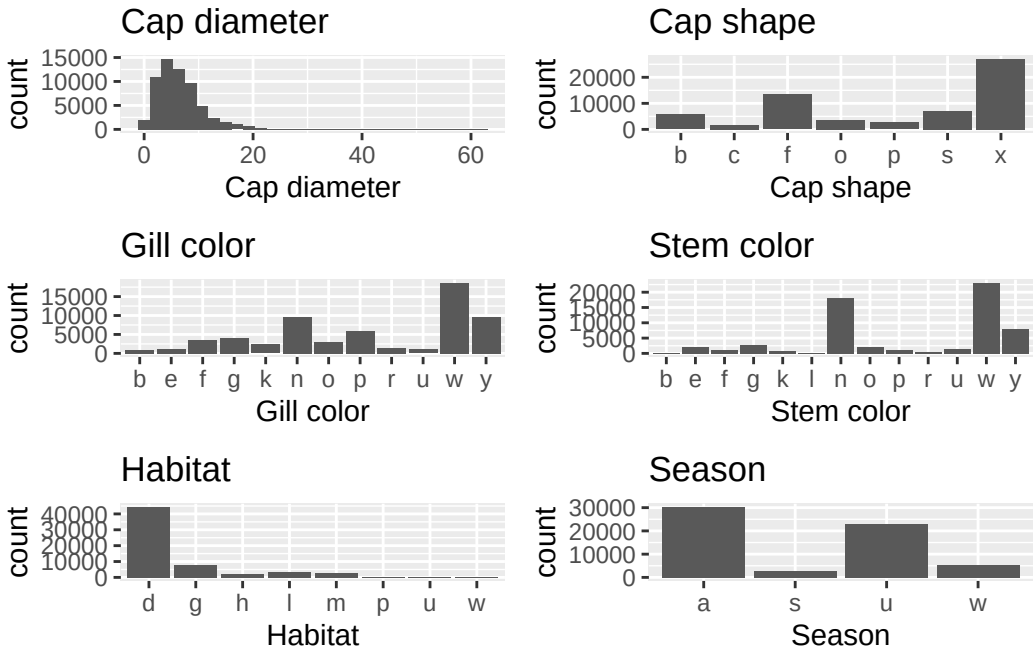
Before we do any EDA, we check that there are no rows with missing values.

class	count	percentage
e	27181	44.509
p	33888	55.491

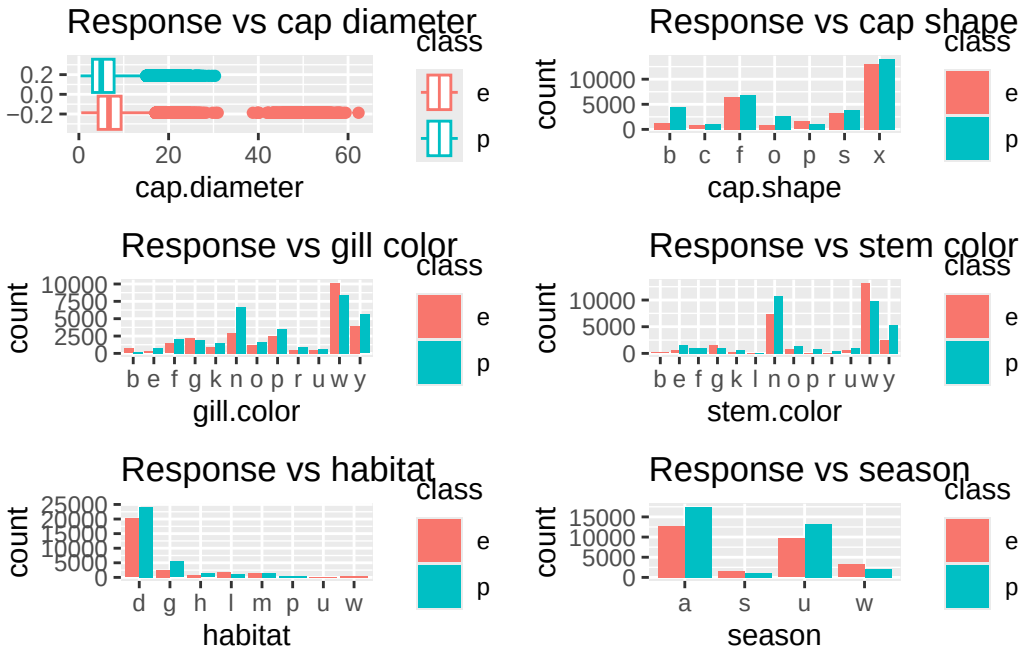
Looking at the overall distribution of our response variable `class`, most of the mushrooms in our dataset seem to be poisonous ("p"). 33888 of the observations, or 55.5% of them are labeled poisonous, as opposed to 27181 (44.5%) of them as edible.



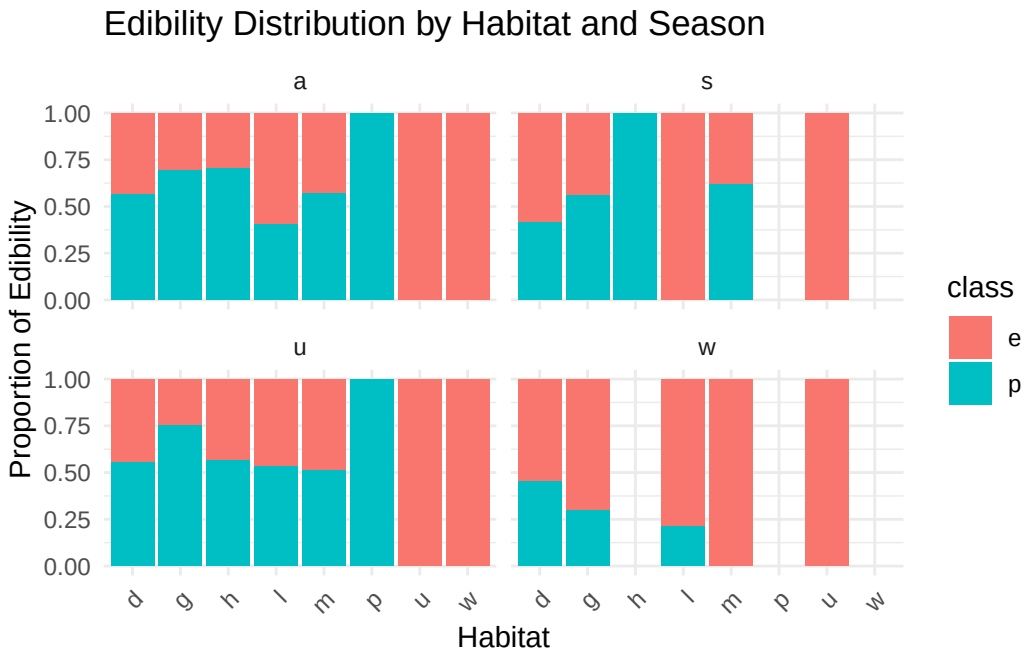
We were first interested in seeing the relationship between our continuous, quantitative variables. While not totally linear, with correlations of 0.423 and 0.436, they graphs seem to be somewhat linear, with some redundancy in their information. Thus, for the rest of the EDA we will focus mainly on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

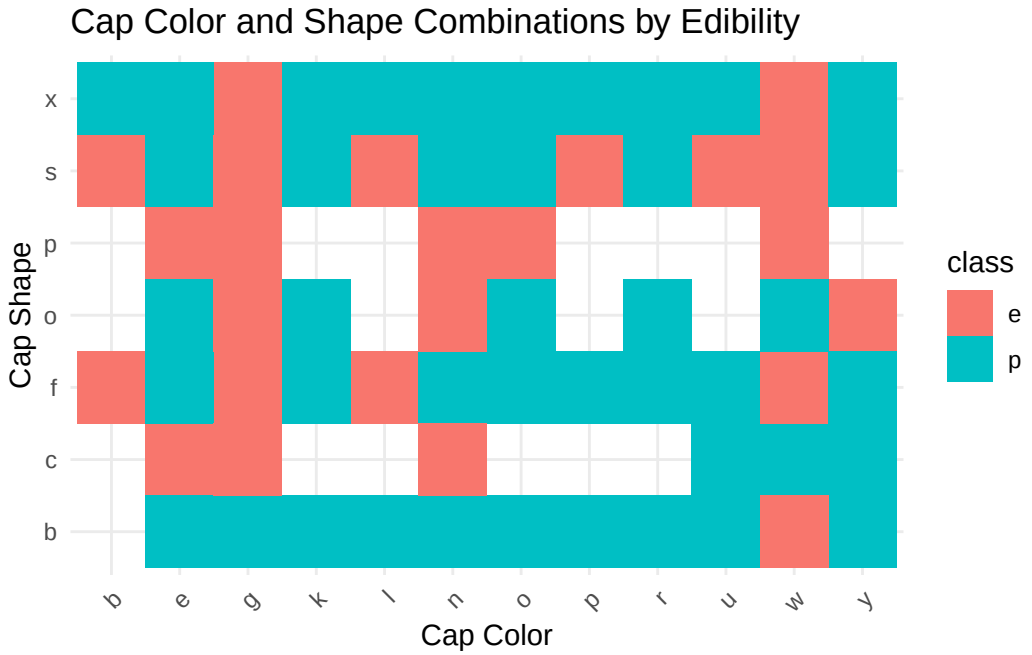


To visualize the distribution of some of our predictor variables, we use a histogram for our continuous variable `cap.diameter` and bar graphs for our categorical variables. The distribution of cap diameter seems to be unimodal, skewed right. From the For qualitative variables, there appears to be more 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; for season, autumn and summer tends to be the most common.



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 stem root of club and rooted tends to be poisonous, but there is relatively a low amount of observations. 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.

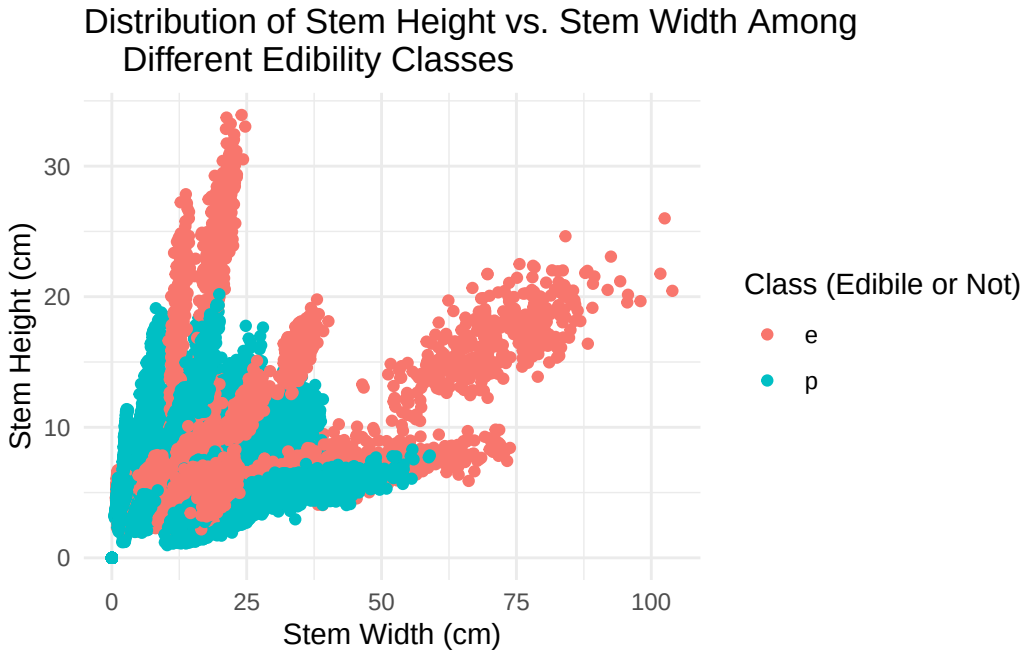




We believe that these predictors may have potential interactive effects that could help us with our model to predict edibility of mushrooms.

Habitat $\times$ **Season**: Mushrooms in certain habitats might only be edible during specific seasons.

Cap color × Cap shape: Some combinations of cap color and shape may signal edibility more strongly than either feature alone.



Finally, we look at multivariate data analysis including 2 predictors and our response variable. While the distribution of stem width and stem height seemed to be the same as cap diameter, and there may be some potential multicolinearity, we are still interested in visualizing their effect on the edibility of the mushroom. 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.

Data dictionary

The data dictionary can be found here

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.