Edible Mushrooms

Introduction

The topic of edible mushrooms is an in important one; any mushroom forum on the internet one can find is often filled 100 of post asking for identification to mushrooms that are edible. While growing safe edible mushrooms is an option of some species, that are many more that are just found to be too difficult. Foraging for mushrooms sometimes is the only way to get certain ones.

Mushrooms serve as a protein source and meat replacement for many and are often found to be a delicacy in their wide varieties. Mushrooms can in appear in vastly different ways from their shape, location, and textures. This variety and sometime similarity between edible and poisonous mushrooms can pose a challenge for those looking to expand their culinary experience into mushrooms outside of the grocery store.

My hope through this analysis is to look at a set of mushrooms and try to find if there are any better ways to go about mushroom hunting. While looking for common traits might be the go-to thing to first think of, it has been proven over many times there is no broad defining traits of mushrooms. Instead, I plan to take a look at the habitats these mushrooms come from and try to find if there are places and traits that lead to more edible mushrooms.

Explain Data

The data used for this analysis come from the UCI Mushroom Data Set. This data set is derived from The Audubon Society Field Guide to North American Mushrooms by G. H. Lincoff. This data contains 8124 observations with 22 different attributes describing the appearance and smell of the mushrooms as well as population size and habitat. Most attributes in this set contain more than a few different options. Gill color alone has twelve different possibilities.

The attributes included in this data is as followed:

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

Its important to note this data includes the descriptive data of 23 species of gilled mushrooms in the Agarucus and Lepiota Family. The Agarucus family include the mushroom type very common seen in supermarkets everywhere. This mushroom grows easily in cultivations and keeps well longer than most other mushrooms.

With data so vast, the process of running analysis on it will be challenging. My process will be involving cutting down the data to what I find to be vital. During my research I discovered categorical data like this can be a challenge to work with especially when it is multivariate.

Cleaning Process

The first step in the cleaning process was to decide how to handle missing data. Luckily, missing data was marked in the data with a '?' and only appeared in one attribute, stalk root. Stalk root's Cramer's V ended up falling in the .3 to .5 range and while the association was moderate, I had plenty of other variable with stronger association.

The missing value appeared in 2480 observations. Choosing to drop rows with a missing a missing value would have dropped observations from 8142 to 5644. I found this number too significant and decided to keep the rows. I instead opted to simply drop the attribute from the analysis since my focus would be on attributes with a strong association to the class type of edible versus poisonous.

Analysis

My first analysis of this data was to find what attributes had the strongest associations to the class type. I preformed a proc freq on all variables against class type and also generated frequency pot plots to get a fist look on how the frequencies were distributed amongst the different variables. From here I used the Cramer's V test and put the variables into three groups.

High association: Bruises? (-.5015), odor (.9710), gill size (.5400), gill color (.6808), stalk above surface ring (.5879), Stalk below surface ring (.5748), stalk color above ring (.5249), stalk color below ring (.5147), ring type (.6033), spore print by color (.7526).

Moderate association: Gill Spacing, stalk root, population, habitat.

Low association: Cap Shape, Cap Surface, cap color, gill attachment, stalk shape, veil color, ring number.

Nothing scored between a 0 and .1, which would have indicated little if any association.

After deciding high association, I decided to make a new data set with the top four scoring variables, odor, gill color, ring type and spore print by color.

I also ran prof freq to get a 2x2 frequency table and plot for class by habitat. This immediately showed there are 3 areas where poisonous mushrooms are more frequent, path, leaves and urban. This also showed in waste there were zero observations of poisonous mushrooms.

Modeling

I found modeling this massive amount of categorical data to be quite challenging. I tried several methods, and they often gave inconclusive results or did not work with the data presented. I opened to go with proc catmod to make a model. I decided a predictive model would be best since I was trying to make predictions based on traits if a mushroom was edible or not.

I made the model by setting the population to habitat since the population otherwise was too vast. This model converged and gave only maximum likelihood estimates for some of the odor and gill color options. The scents anise and almond both had very small p values while the rest were large.

The maximum predicted probabilities found the habitat with highest probabilities of edible mushrooms were grasses, meadows, and waste while the highest poisonous were leaves, paths, and urban habitat.

I ran the same catmod on odor and gill color to see which of these traits in the two variables leaned towards edible mushroom. It was found the scents almond, anise and none had high probabilities and the gill colors red, black, brown, orange, purple, white and yellow had high probabilities.

Further Analysis

With the results in hand, I decided to do further frequency analysis on the three highest probability habitats against odor and gill color based on the information gathered during the modeling step. These showed while the frequency matched most the time based on what was a high probability trait, it did not always line up such as with brown in grasses. It showed a much higher frequency of poisonous mushrooms and not the expected high probability edible mushrooms.

Results

The following results can be shown from the analysis of this mushroom data set. Meadows, grasses, and wasted have the highest probability of edible mushrooms. The scents almond, anise and none had high probabilities of edible mushrooms and the gill colors red, black, brown, orange, purple, white and yellow had high probabilities.

Crossing these finding it is found within meadows mushrooms with while, brown, black, or grey gills tend to edible and mushrooms with the smell of anise or almond tend to be edible. Within grasses mushrooms with brown or black gills or have no smell tend to be edible. Then finally it was found that mushrooms in waste always tend to be edible.

There are limitations to these finding, however. The first is, this fits only to two families of mushrooms. The amounts found in different habitats and with different traits varies wildly. This data is vast and has a lot of traits to consider in many different combinations that may need to varying results based on the focus of study.

Conclusion

In conclusion while these are some decencies and associations found with in the data it is challenging to derive a good pattern for edible mushroom. Often the best way to know if a mushroom is edible is to take in all diagnostic factors of the mushroom and identify it completely. A mushroom should never be consumed with out being 100% sure it is safe and that is best done by knowing exactly what type of mushroom you have in hand. While I would like to explore this topic further, I find this data may be best suited for AI applications.