Math 151: Report

Jacob McGraw, Julio Arciga

May 9th, 2022

1 Introduction

Interestingly, there has been a resurgence of popularity in mushroom hunting in recent years. All over North America, cooks and general mushroom enthusiasts have been taking to the outdoors to find the most rare and exotic mushrooms they can find. However, mushroom hunting can be a dangerous endeavor. While some mushrooms can be turned into delicious food, many types of mushrooms are poisonous and can be seriously deadly. In our project, our goal was to take a dataset on mushrooms provided via The Audubon Society Field Guide to North American Mushrooms (1981) to create a model to predict whether a mushroom is poisonous or not. Additionally, our goal was to determine the most effective model we could find with the least number of variables as well as exploratory analysis to discover all we could about the mushrooms provided in the dataset.

2 Data Cleaning

Before any work could be done in regards to model building, we had to clean the data first. While the data was delivered mostly clean with no errors, there was still work to be done before any progress could be made. In the "gil.attatchment" feature, approximately 98% of the gill attachments were free, so this feature was removed. Furthermore, the feature "viel.type" was removed because all of its entries were the same. The next substantial problem was that the "stalk.root" feature has over one-third of its data missing. For the missing values, the data instead shows a question mark. Upon researching the topic further, it was discovered that certain types of mushrooms grow without stalks. However, the dataset does not specify whether or not the mushrooms have missing stalks due to the type of mushroom or because the sample was gathered incorrectly (i.e., the field researcher accidentally picked the mushroom wrong). Additionally, the dataset does not specify the individual mushroom type of each observation; this is most likely because the sheer number of mushrooms in North America would make it entirely plausible that each of the 8124 observations could be a distinct type of mushroom. Because of the large number of unknowns when dealing with this feature, we felt it would be best to remove it. Another problem was that "stalk.surface" above and below were almost

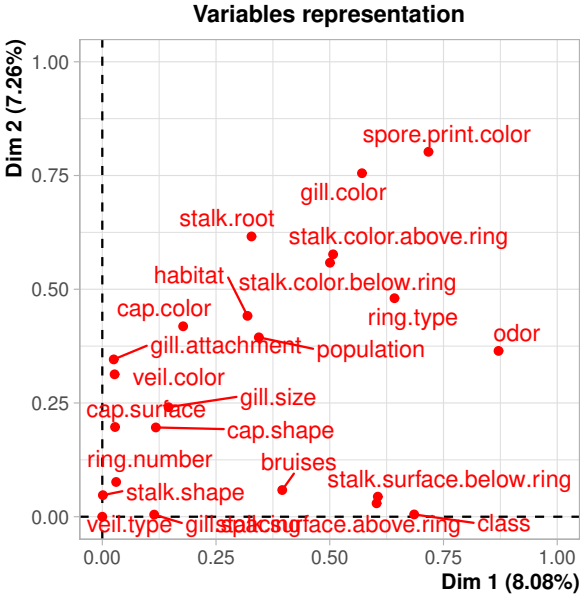

Figure 1: MCA Variable Representation

entirely the same; over 77% of the observations were identical. Therefore, we decided to remove one of the observations (specified later)

3 Exploratory Analysis

The first task our group did in terms of exploratory analysis was to use Multiple correspondence analysis to determine highly correlated variables to trim out of the overall dataset for simplicity's sake. Figure 1 shows the results of the multiple correspondence analysis.

3.1 Exploratory Analysis; variable selection

Figure 1 Shows a high level of correlation between the following variables: stalk.surface above and below the ring, stalk.color above and below the ring, gill attachment, and veil color. To simplify the results of our models, we will only use one of each; to create the best model, we will use the features that display the highest level of correlation with class, which are stalk.surface.above.ring, veil.color, and stalk.color.below.ring. Once these variables have been cleaned from the dataset we can create the most accurate and simple model possible.

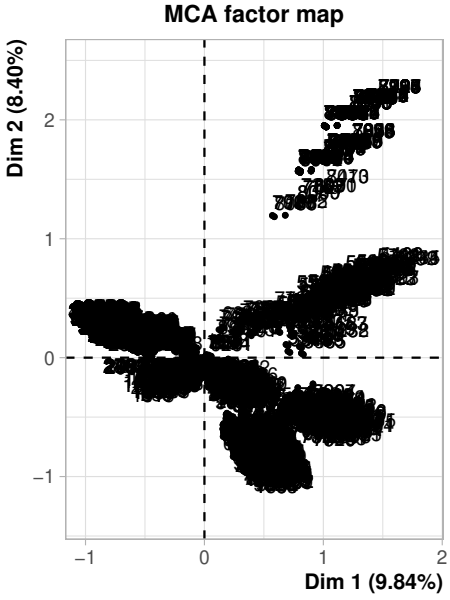

Figure 2: MCA Factor Map

3.2 Exploratory Analysis; Poison Clustering

One thing our group was part particularly interested in was clustering and possibly identifying the poison within a particular type of mushroom and identifying key traits within the different types of poisonous mushrooms. According to the North American Mycological Association, there are 10 distinct types of poison found in North American mushrooms (Mushroom Poisoning Syndromes, 2022). These poisons range in severity from immediately deadly to enjoyable (such as with psilocybin mushrooms). So the question remains, are there ways to cluster the mushrooms based on certain characteristics in an attempt to differentiate the mushrooms based on poison? A good way to start the analysis is with a multiple correspondence plot, as shown in Figure 2.

Figure 2 shows that the observations cluster into some very distinct patterns, so this is a good start. An unfortunate reality is that we will be unable to test our hypothesis in regards to clustering based on poison type because we don't know the type of poison within each mushroom. Furthermore, we won't be able to infer what type of poison is in each mushroom because we have no information on the type of mushroom of any of the observations. Therefore, we will proceed with caution and try to determine the best resulting clustering algorithm given our limitations.

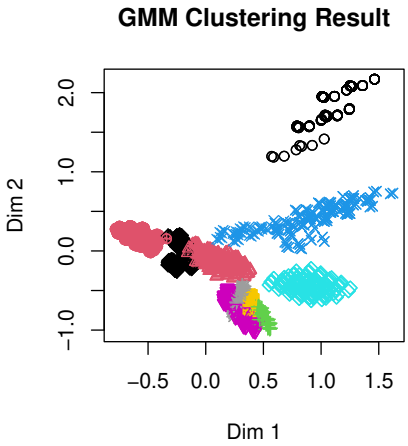

Figure 3: Clustering Result 1

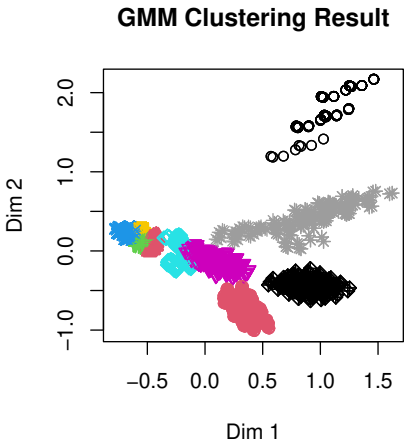

Figure 4: Clustering Result 2

3.2.1 Poison Clustering; GMM clustering

The first attempt at clustering the results into different poison types was by using GMM clustering on the results of the multiple correspondence analysis. The results can be seen in the following figures:

As we can see from Figures 3 and 4, the graph results do not converge to anything satisfactory.

3.2.2 Poison Clustering; K-Mediods

Another results we can try is to use k-mediods. the results from running this clustering algorithm can be see in Figure 5:

Interestingly, Figure 5 shows that the graph stabilizes to this result fairly nicely and has no noticeable alterations to its final output. It makes sense that k-modes would work better with this data because the model is designed for categorical data. The output gives us several interesting results; first, we note that a good majority of the clusters have elements that have at least one feature that is the same throughout the whole cluster, we can see this with the. following results:

- Cluster 1: stalk color is always white, veil color is always white
- Cluster 2: above stalk color is always white, ring type is always evanescent, habitat always grassy, veil color always white.
- Cluster 3: veil color always white, stalk ring below color always white, ring type always pendant, stalk surface above ring always smooth, always

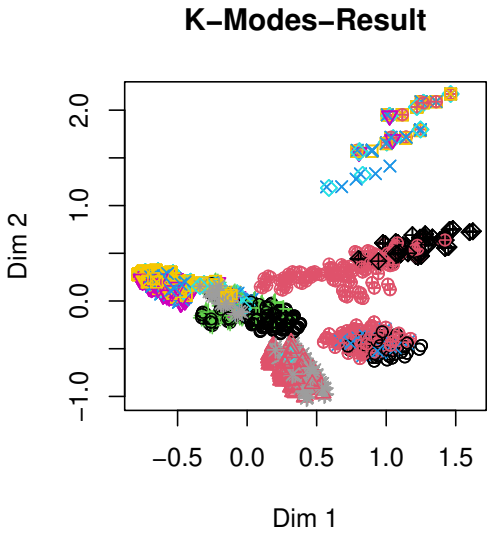

Figure 5: K Modes Clustering Result

bruises

- Cluster 4: ring type is always pendant, always odorless
- Cluster 5: stalk.surface.above.ring 99% smooth.
- Cluster 6: ring type always pendant, stalk surface above ring always smooth
- Cluster 7: stalk surface above ring always smooth, gill spacing always close
- Cluster 8: veil color always white, stalk color below always white, always odorless
- Cluster 9: always bruises, always no odor, gill spacing always close, gill size is always broad, gill color is red, stalk shape is enlarging, stalk surface above ring is smooth, veil color is always white, habitat is always waste, spore print color is white, population is clustered
- Cluster 10: stalk shape is always enlarging

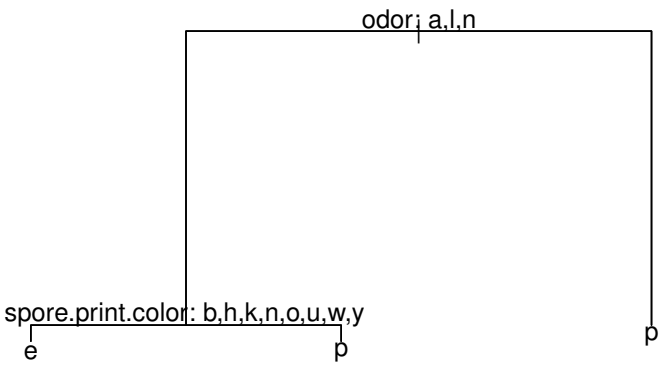

Figure 6: Clustering Mushrooms With Odor

From this result we see that there are many features that are fairly common across all clusters; for example, having a white veil color,pendant ring type, and odorless are found in many of the clusters. However, there are other clusters that have features that are completely unique. Interestingly, the clustering was able to isolate one type of poisonous mushroom that has appears to be highly identifiable. Cluster 9 is the smallest cluster and its observations have several distinct characteristics. This type of poisonous mushroom has a white veil color, no odor, red gill color, close gill spacing, easily bruised, smooth stalk surface, lives in waste, is clustered, and has a white spore color. If nothing else, this model does a good job of isolating this particular brand of trash-dwelling, clustered, red and white mushrooms.

4 Clustering Poisonous vs Edible Mushrooms

The most prescient task with the dataset was to try to classify the poisonous and the edible mushrooms as best as possible. To do this, our group used trees as an intuitive tool for mushroom hunters in the field. The results can be seen in Figure 6:

Figure 6 shows that if the Mushroom has a creosote, fishy, foul, musty, pungent, or spicy odor then it is a strong indication that the mushroom is poisonous. An almond, anise, or no smell at all it will have to be further

	Edible	Poison
Edible	1418	20
Poison	0	1406

Table 1: With Odor Clustering

	Edible	Poison
Edible	1357	27
Poison	111	1349

Table 2: Without Odor Clustering

determined through the spore print color. If it is green with these smells then it will be classified as poisonous. The accuracy of this model using a training set of 65 percent to a test set is a whopping 99 percent success rate of the predicted values to the expected values. Table 1 shows the exact results of the algorithm:

Table 1 shows the predicted against the actual values. We can see that our model correctly labels edible as edible and poison as poison to near perfection. The only problem is that in this model we have incorrectly labeled some mushrooms as edible when in fact they are poisonous. This of course will put a sense of fear in a lot of people that our model does not guarantee that a mushroom is edible. Nobody wants to take the risk of dying although it is theoretically a 0.7% chance of getting a poison mushroom and possibly dying.

4.1 Clustering Poisonous vs Edible Mushrooms Without Odor

In 2020, the worldwide pandemic of Covid-19 took over our ordinary lives. One of the consequences and long-term effects of catching Covid-19 is the loss of the ability to smell. Therefore, a growing number of individuals may need an alternate model that accounts for individual users that can't smell. Figure 7 shows the results of the updated model:

Figure 7 shows a model where we have excluded Odor as a variable and distinguish a poisonous mushrooms by other means. If the mushroom is buff, black, brown, orange, purple, yellow and has a broad gill size it is classified as edible. If it is chocolate, green, or white and 2 rings it will be classified as edible.

The accuracy rate is at 95% of our model predicting correctly. The exact results of this algorithm are shown via table 2, which shows more errors than the model with odor as a factor variable. But there are more errors in predicting whether the mushroom is edible, meaning that we would classify an edible mushroom as poisonous and will throw away some good mushrooms to eat. Which is fine as it only encounters 3% of the data. Now poisonous mushrooms that were classified as edible were our truly bad values but were not far away from the odor model. We encounter this dilemma 0.9% of the time as opposed

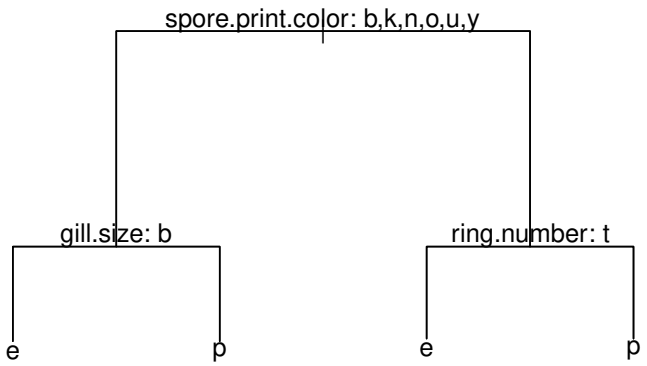

Figure 7: Clustering Mushrooms Without Odor

to 0.7% with having odor as a factor. We can see that in the case that we lost our sense of smell to certain circumstances then we can still get very positive results with just about 1% of having to pick a poisonous mushroom accidentally.

5 Conclusion

Contrary to our initial findings, there was a lot to discover in this mushroom dataset. At first glance, the data clusters to a highly accurate level even without the rigor of variable selection and data trimming. In the end, odor and spore print color were enough to have an accuracy level high enough to satisfy most mushroom hunters. interestingly, we lose a fair amount of accuracy with the loss of odor, but we can still make a model that's easy to use and will keep users safe 95% of the time. While our group made many interesting discoveries over the several weeks we spent analyzing the material, there was still more progress to be made. For example, given enough time we might have been able to investigate the clustering of the principal components to discover why there is such distinct grouping in the observations or more accurately gain meaning from the k-modes clustering result. In conclusion, mushrooms are fascinating and bizarre organisms that are an endlessly fascinating source of research.

6 References

Mushroom poisoning syndromes. North American Mycological Association. (n.d.). Retrieved May 6, 2022, from https://namyco.org/mushroom_poisoning_syndromes.php