Milestone Report 1

Mushroom Classification

Sergio Robledo

December 2, 2019

Mushroom Foraging (Preprocessing & Exploration)

Foraging for wild mushrooms, or mushroom picking, is the pastime of many people, and even though it does not seem like a dangerous hobby, consumption of a misidentified mushroom could cause death; it is therefore imperative for the person foraging for mushrooms to know the mushrooms' edibility. However, in order to become experienced enough to tell edible and poisonous mushrooms apart, a person would have to invest a good amount of time, because knowing the difference between an edible and fatal mushroom might be in the slightest detail. In order to make mushroom foraging more accessible to a wider range of people while preventing a possibly fatal intoxication, a reliable machine learning model capable of classifying mushroom edibility is needed. Below, the acquisition, cleaning and exploration of the mushroom dataset is explained.

Mushroom Dataset & Preprocessing

The mushroom dataset comes from the UCI Machine Learning repository and can be found here. The dataset records 22 characteristics of 8,124 mushrooms; the characteristics included range from mushroom odor to gill characteristics, and all mushrooms are labeled as either edible or poisonous. The only main preprocessing step the dataset required, was the renaming of the characteristics from single-letter representations to their full name; for example, for the possible types of surfaces seen on the mushrooms, an original value of 'f' was converted to 'fibrous', 'g' to 'grooves', 'y' to 'scaly', 's' to 'smooth' and 'k' to 'silky'.

Exploratory Analysis

The exploration and visualization phase of the project began by looking at the distribution of the observed mushrooms' edibility; knowing the distribution of the dataset adds proves useful when deciding what machine learning algorithm to use, because some models do not handle imbalanced datasets unless the data is balanced beforehand. The results can be seen below.

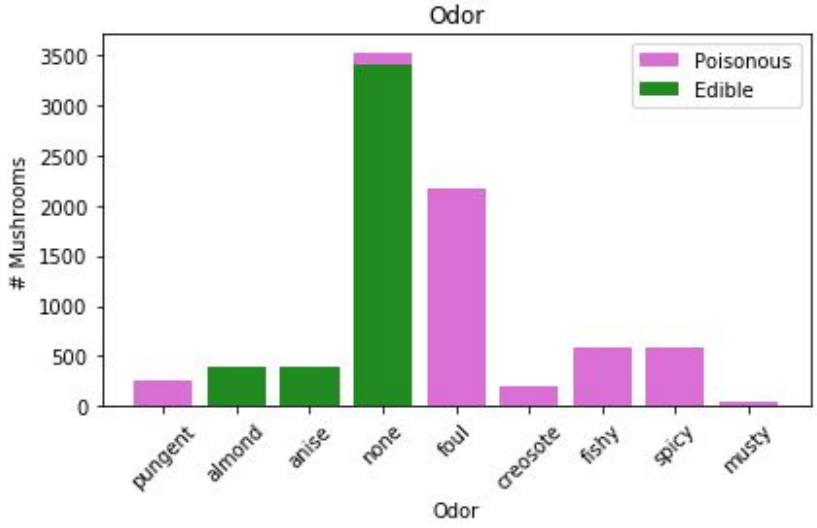
```
Total # of Observed Mushrooms: 8,124

                        Edible: 4,208 (52)%
          Poisonous Mushrooms: 3,916 (48)%
```

From above, we can see that there are a total of 8,124 mushrooms constituting the dataset, and about 52% are edible while the remaining 48% are poisonous; therefore, it can be concluded that the dataset is fairly balanced, so we do not need to do any further balancing.

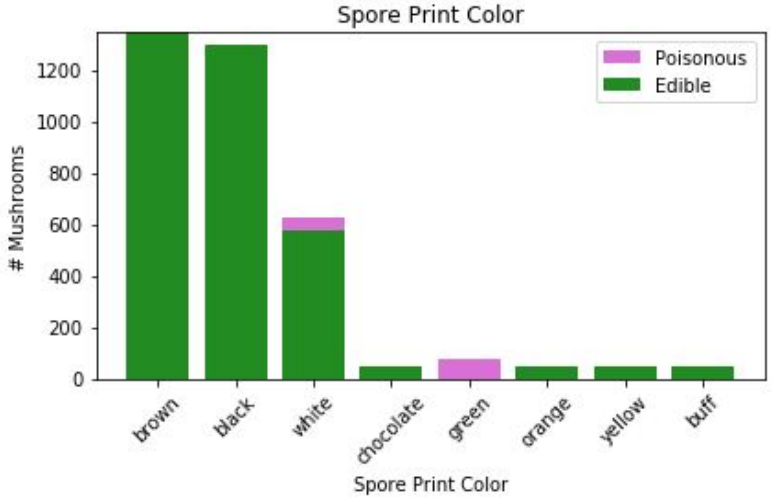
Next, using the Pearson correlation coefficient we are going to get the category with the highest correlation to the edibility classification of a mushroom and explore it further; below are the initial edibility correlations for all the mushroom properties.

```
Mushroom Edibility Correlations:
--------------------------------

odor_none                        0.785557
ring_type_pendant                0.540469
gill_size_broad                  0.540024
bruises_True                     0.501530
stalk_surface_above_ring_smooth  0.491314
                                     ...
gill_size_narrow                -0.540024
stalk_surface_below_ring_silky  -0.573524
stalk_surface_above_ring_silky  -0.587658
odor_foul                       -0.623842
veil_type_partial                     NaN
```

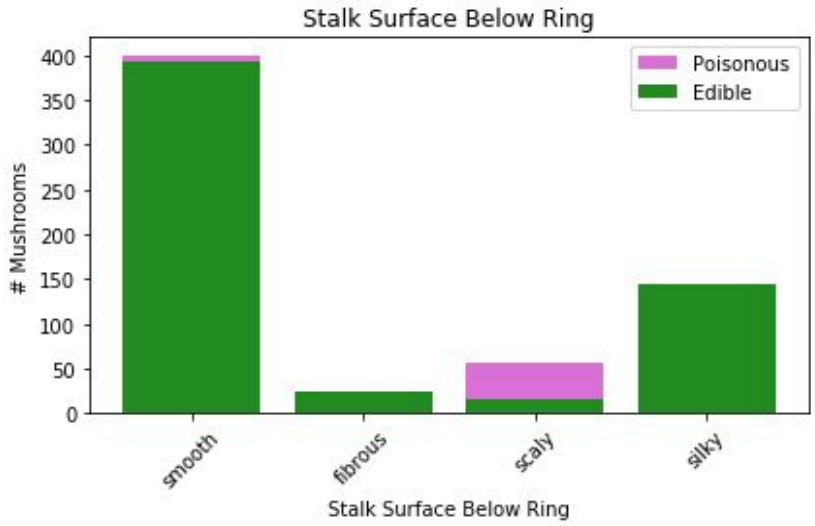
Characteristics with positive correlation values indicate that the probability of a mushroom being edible increases when it has the characteristic; on the other hand, negative correlations indicate the mushroom is more likely to be poisonous if it has such characteristic. Also, “veil_type_partial” has a value of NaN, because all mushrooms observed had partial veil types; this means that there is no correlation between the veil type of mushrooms observed and their edibility. We will now dig deeper into the dataset in an attempt to understand what rules govern the edibility of mushrooms; we will start by comparing the odors of edible and poisonous mushrooms.



The odor characteristic of the mushrooms turns out to be a very important starting feature of the mushrooms, since it can be used reliably to distinguish between edible and poisonous mushrooms; all edible mushrooms are either almond or anise smelling, and poisonous mushrooms smell pungent, foul, creosote, fishy, spicy or musty. However, mushrooms with no odor, while mostly edible, have a possibility of being poisonous, and because even one misclassification of edibility is unacceptable due to the possibly lethal consequences, we have to look at mushrooms with no odor more closely to see what sets them apart. After filtering the mushrooms to include only mushrooms with no odor, the new strongest correlation with edibility was the spore print color of the mushroom.

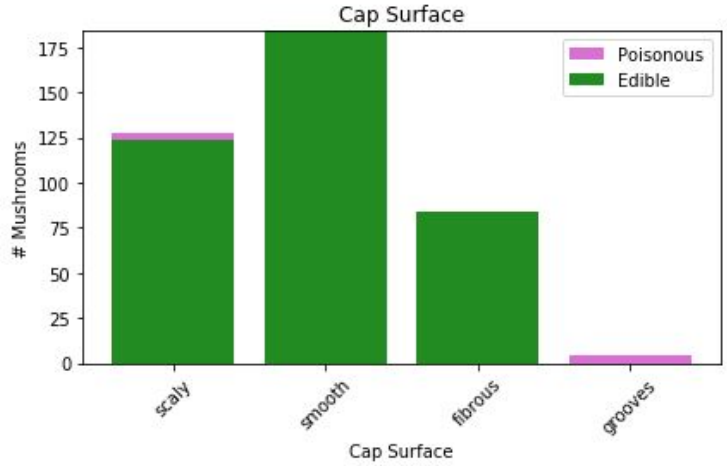


Of the mushrooms with no odor, the ones with a green spore print were all poisonous; using spore print color in conjunction with odor, we were able to reduce the number of misunderstood poisonous mushrooms from 120 to 48. However, we still have a mix of edible and poisonous mushrooms when an odorless mushrooms has a white spore print. Now, after filtering for odorless mushrooms with a white spore print, we got that the next strongest correlation to the edibility of a mushroom was the stalk surface below the ring.

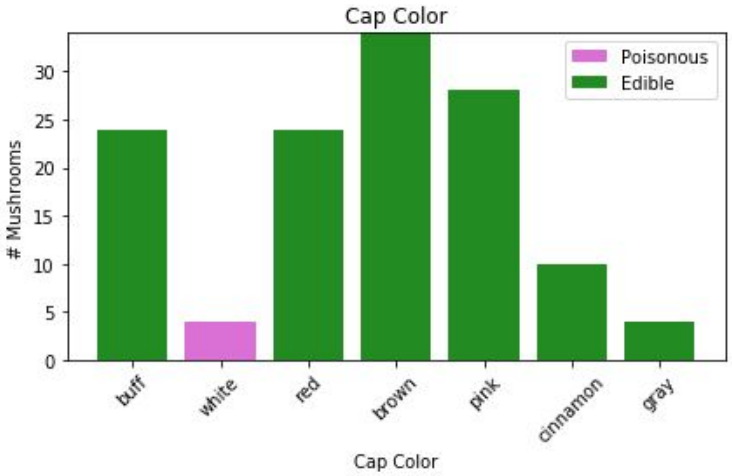


Even though this lead did not reduce the number of misunderstood poisonous mushrooms, it did reduce the pool of mushrooms that we have to further analyze by telling us that all odorless mushrooms with a white spore print and either fibrous or silky stalk surfaces below their rings are edible. Since the poisonous mushrooms mixed in with the edible mushrooms now fall under two categories, we will have to explore both separately; we will first filter the remaining mushrooms further to include ones with a smooth stalk surface below their rings. Once filtered, the strongest edibility correlation was the cap surface of the mushroom.

```
--------------------------------------------------

--- Stalk Surface Below Ring (Part 1) ---

--------------------------------------------------

Stalk Surface Below Ring (Smooth) Mushrooms: 400
Edible: 392
Poison: 8
```

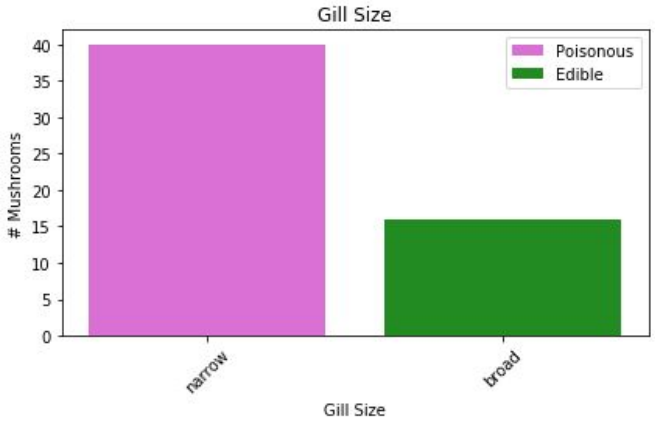


By filtering odorless mushrooms with white spore prints to smooth stalk surfaces before their rings, we can see that if their cap surface is groovy, they will all be poisonous; we are getting closer to learning the main differences between edible and poisonous mushrooms, but we still have mushrooms with a scaly cap surface to look at. Filtering by the cap color will give us the last characteristic needed to separate edible and poisonous mushrooms with smooth stalk surfaces and scaly cap surfaces.



After looking at the cap color of the remaining scaly cap mushrooms, we can finally confidently tell apart the remaining poisonous mushrooms from the other 124 edible mushrooms with the same previously looked at characteristics. The last thing we need to do is find the differences between odorless mushrooms with white spore prints and scaly stalk surfaces below their rings, since there is still a mixture of 56 edible and poisonous mushrooms with those characteristics.

```
------------------------------------------------

--- Stalk Surface Below Ring (Part 2) ---

------------------------------------------------

Stalk Surface Below Ring (Scaly) Mushrooms: 56
Edible: 16
Poison: 40
```



Finally, after filtering odorless mushrooms with white spore prints and scaly stalks below their rings by their gill size, we were able to separate the last 56 mushrooms with mixed edibility classifications into separate classifications; mushrooms with narrow gill are all poisonous, and mushrooms with broad gills are all edible.

Conclusion

Learning what the most important characteristics are when separating edible and poisonous mushrooms will prove very beneficial when constructing the final machine learning model, because we will be able to reduce the amount of data the model has to understand and in turn decrease the complexity of the model. The number of features

we will pass to the machine learning model was decreased from 22 to 6; these features include the mushrooms' odor, spore print color, stalk surface below rings, cap surface, cap color and gill size.