Deadly Mushroom Classification

Student: Rachel Nelson

Class: Bellevue University DSC680

Business Problem

There are over 10,000 known types of mushrooms in the world. Roughly 70-80 of them are poisonous and some even can even be fatal when eaten. How can we identify which mushrooms are safe to eat? What factors are important to identify such mushrooms? How accurate can we predict if a mushroom is safe or not? These are questions I will be looking to answer by comparing various classification algorithms on a data set of mushrooms characteristics.

Background/History

Mushroom foraging, also known as mushroom hunting, mushrooming and mushroom picking, is an activity where people gather mushrooms in the wild. While most mushrooms are edible, some can be deadly. In the United States, there are “around three deaths per year” due to people eating a poisonous mushroom (Keough, 2020).

Data Explanation (Data Prep/Data Dictionary/etc)

I will be getting the data set from Kaggle. This dataset was originally contributed to the UCI Machine Learning repository 30 years ago. This dataset from Kaggle “includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms (1981). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended” (2016, CI Machine Learning). The data includes 23 columns. All of the data is coded as a single character string for each attribute.

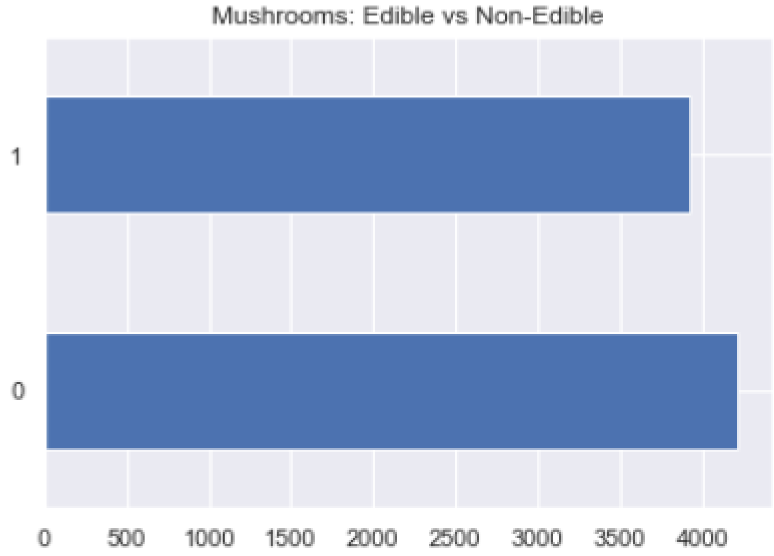
The mushroom data set contains the following attributes:

- **Class:** edible=e, poisonous=p
- **cap-shape:** bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
- **cap-surface:** fibrous=f, grooves=g, scaly=y, smooth=s
- **cap-color:** brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
- **bruises:** bruises=t, no=f
- **odor:** almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s
- **gill-attachment:** attached=a, descending=d, free=f, notched=n
- **gill-spacing:** close=c, crowded=w, distant=d
- **gill-size:** broad=b, narrow=n
- **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
- **stalk-shape:** enlarging=e, tapering=t
- **stalk-root:** bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?
- **stalk-surface-above-ring:** fibrous=f, scaly=y, silky=k, smooth=s

- **stalk-surface-below-ring:** fibrous=f, scaly=y, silky=k, smooth=s
- **stalk-color-above-ring:** brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
- **stalk-color-below-ring:** brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
- **veil-type:** partial=p, universal=u
- **veil-color:** brown=n, orange=o, white=w, yellow=y
- **ring-number:** none=n, one=o, two=t
- **ring-type:** cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z
- **spore-print-color:** black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y
- **population:** abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y
- **habitat:** grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d

In order to prepare the data for classification, I first had to take each of the attributes and re-code them into integers. After reviewing the data head types (which were all objects) and looking for any nulls (there were none), I reviewed the number of unique values per column. For attributes with only two factors, I recoded them into binary classifications (1 and 0). These attributes included class, bruises, gill attachment, gill spacing, gill size and stalk shape.

Next, I wanted to make sure that my data set has a good balance of both edible and non-edible for mushroom classification, which I was happy and surprised to see the data was well balanced:



Methods & Analysis

I will be using classification methods to determine if the mushroom is edible or poisonous. Models included logistic regression, decision trees and naïve bayes classifier. Models used will compare accuracy and look at type 1 and 2 errors to see which performs better. It is more important that we do not falsely identify deadly mushrooms as edible mushrooms, verses errors where we identify safe mushrooms as deadly (type 1 and type 2 "alpha and beta" errors).

I then created dummy variables from the remaining categorical variable and added them to the data frame.

Next, was to run a correlation analysis on the data set to see which attributes highly correlate to my target variable, Class. I made note of the variables with medium to high positive and negative correlation to my target variable.

Class	Pearsons Correlation
odor_n	-0.79
ring-type_p	-0.54
bruises	-0.50
gill-color_b	0.54
gill-size	0.54
stalk-surface-below-ring_k	0.57
stalk-surface-above-ring_k	0.59
odor_f	0.62

Next, it was time to throw them all into various algorithms and check out performance.

Naïve Bayes Classifier

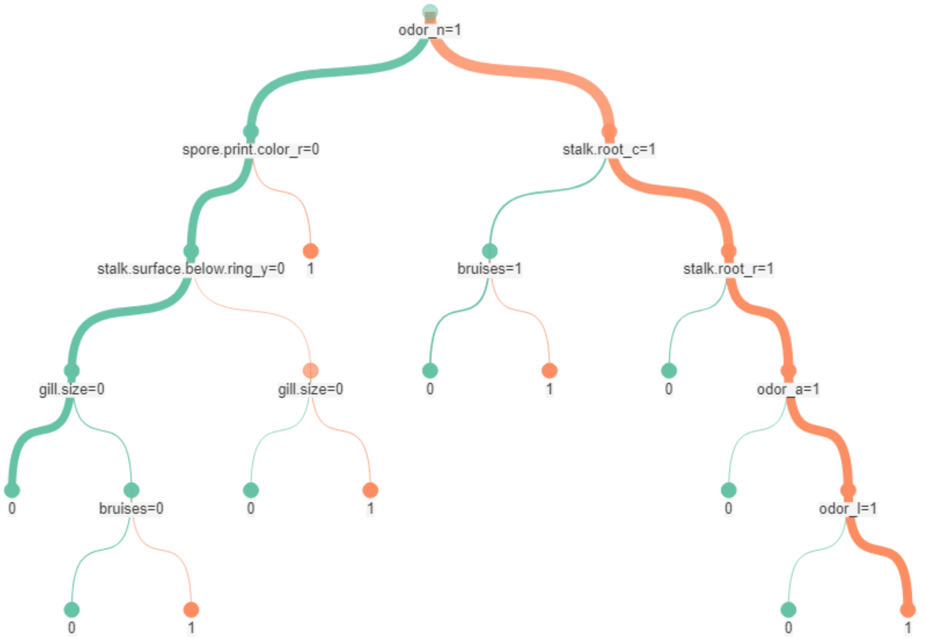
With the Naïve Bayes Classifier, our accuracy was 93%, which seems high, but when we look at the Confusion Matrix, we show that we predicted that 31 poisonous mushrooms were edible. So, although the accuracy is high, it is still unacceptable.

	Predicted Edible	Predicted Poisonous
Actual Edible	4177	524
Actual Poisonous	**31**	3392

Decision Tree

Next, we tried a decision tree, which gave us 100% accuracy! It also shows us that odor is the #1 identifier for if a mushroom is poisonous, which we also learned during the correlation analysis.

	Predicted Edible	Predicted Poisonous
Actual Edible	3916	0
Actual Poisonous	**0**	4208



Logistic Regression

The logistic regression performed the same as the decision tree, where we ended up with 100% accuracy.

	Predicted Edible	Predicted Poisonous
Actual Edible	3916	0
Actual Poisonous	**0**	4208

Conclusion

There are 6 main factors that you need to identify in a mushroom to tell if it’s poisonous:

1. Odor
2. Bruises
3. Gill size
4. Stalk surface below ring
5. Spor primary color
6. Stalk root

While I still recommend “if in doubt, throw it out” as a good practice for mushroom hunters, I would especially recommend this approach. I have also learned that odor is a main classifier for poisonous mushrooms. Throw out any that smell of almond and anise and beware of ones that are odorless.

Assumptions & Limitations

This dataset from Kaggle “includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms (1981). This only includes mushrooms from those families.

The assumption would be that this data is accurate and provides coverage of various mushroom types within the Agaricus and Lepiota mushroom families. There is also the assumption that this data logic would only be used to identify mushrooms in those families and not from other mushroom families.

Challenges

We had two models that performed the exact same, logistic regression and decision tree. While in most cases, I would go with the logistic regression model, in this case I decided to go with the decision tree because it provided a nice “Rule Tree” to follow. However, decision trees are notoriously known for overfitting data.

Future Uses/Additional Applications/Recommendations

A one pager guide for mushroom hunters could be created to show what to look for in poisonous mushrooms.

Ethical Assessment

There is a saying in the mushroom foraging community that “if in doubt, throw it out”, and even with 100% accuracy, I would not want to be responsible if there is a mushroom out there that does not fit this model but is still poisonous.

Resources

Keough, B. (2020, July 13). *Here's what you'll need to start foraging mushrooms*. The New York Times. Retrieved January 30, 2022, from https://www.nytimes.com/wirecutter/blog/how-to-hunt-mushrooms/