Mushroom Classification Problem

Nick Wetter
Damon Shorty

Intro to Problem

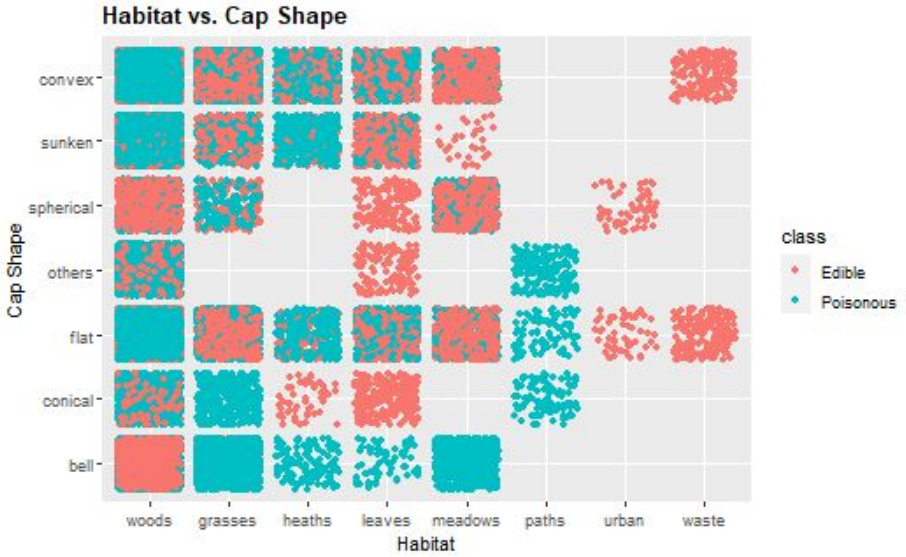
- Common to find mushrooms all over the world
- Determine whether mushrooms are edible or poisonous given a set of physical characteristics
- Dataset Source:
  - UCI Machine Learning Repository
- 1 Response variable (binary - edible or poisonous)
- 20 Predictor variables (3 quantitative and 17 qualitative with levels ranging from 2 to 13)

Summary of Statistics

Quantitative Variables

Correlation Values			
	Cap Diameter	Stem Height	Stem Width
Cap Diameter	1.000	0.423	0.695
Stem Height	0.423	1.000	0.436
Stem Width	0.695	0.436	1.000

Qualitative Variables



Early Challenges

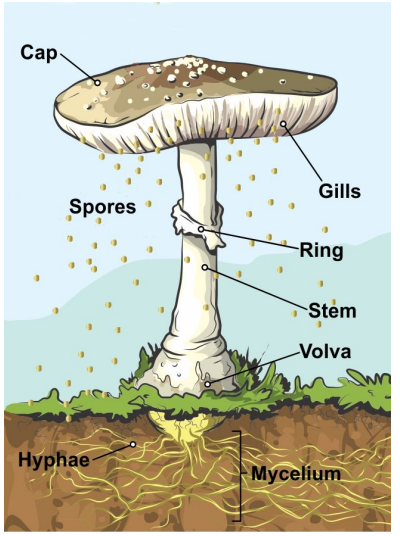
- Missing values in repository
- Large number of variables and mix of categorical/numerical values
    - SciKit is difficult about this

Models

- Unpruned Tree
  - 90.9% of test observations classified correctly
- Pruned Tree
  - Tree with 28 terminal nodes
  - 90.9% of test observations classified correctly
- Random Forest
  - 99.6% of test observations classified correctly
- Naive Bayes
  - 75.2% of test observations classified correctly

Ranking Variable Importance

1. Cap Surface
2. Gill Attachment
3. Gill Color
4. Stem Width (quantitative)
5. Stem Color
6. Stem Surface
7. Cap Color
8. Gill Spacing (qualitative)
9. Stem Root
10. Cap Shape



Models

- K Nearest Neighbors (Numerical features only)
  - Using 18 neighbors we got an accuracy score of 0.80613
  - Values from 1 to 99 were tested for k, our reported value was the most accurate though most values were very close
  - Both the mean and median missing value imputation strategies resulted in this being the value of k with the highest accuracy score
- Histogram-based Gradient Boosting Classification Tree (Num. features only)
  - No hyperparameters
  - Missing numerical value imputation strategies for this model gave the following scores:
    - Mean: 0.78786
    - Median: 0.78877

Results/Discussion

- Some of the models are quite good
- In this case, how good is good enough?
- Our current most accurate model is 99.6% accurate
- Is 0.4% still too big of a risk in a situation like this where the stakes are very high?
- Depends on your personal risk tolerance, would YOU eat the mushroom if a model this accurate said it was safe?

Further Research

- Determine whether mushrooms have medicinal properties based on similar dataset
- Create additional models with more variables included:
  - Histogram-based Gradient Boosting Classification Tree
  - K Nearest Neighbors
- Examine optimal hyperparameter values for models
- Cross validation to test for testing data leak