Identifying Features of Poisonous Mushrooms

Juliana McCausland

The goal of this project was to identify and explore features of mushrooms that are correlated with poisonousness. I found that using only a few features, you could confidently identify poisonous mushrooms (with 99.8% accuracy and perfect precision + recall). These features are odor, gill size, and spore print color. Additionally, odor alone can also produce a highly accurate model (with 98% accuracy and 98% ROC).

Design

There is an entire community of mushroom enthusiasts, many of whom enjoy scavenging for mushrooms. Consuming the wrong mushroom can lead to a long, miserable death. It can also lead to non-fatal but very uncomfortable digestive and neurological side effects. A company is developing an app to help users more safely scavenge by allowing users to identify poisonous mushrooms by looking at only a few features, making it more practical in real-world situations.

Data

I used the UCI mushroom dataset, which originally contained about 22 features along with the target classification of poisonous/edible. After converting the features to dummies, I had about 96 features. The features included information about mushroom shape, spore colors, odor, gill sizes, gill shapes, stalk shape, and more.

Algorithms

I performed an initial logistic regression and KNN implementation on the entire dataset (after converting to dummies) and saw that it produced perfect scores across the board. At this point, the goal was to eliminate as many features as possible while maintaining high accuracy so that users would only need to identify a few features to know with high confidence that a mushroom is poisonous. I created a baseline using odor features, which had 98% accuracy and high precision and recall. I then looked at other features using pairplots, boxplots, and feature importances (from both random forest and xgb). I played with combinations of features to see which would produce the highest-scoring models. KNN, logistic regression, and random forest were primarily used throughout the project.

Final models (KNN and LR produce same results):

- Features: odor, gill size, spore print color
- Accuracy: 99.88%
- Recall: 1.0
- Precision: 1.0
- F1: 1.0

Other findings:

- Highly predictive characteristics of poisonous mushrooms:
    - Foul odor
    - Narrow gill size
    - Chocolate-colored spore prints
    - White-colored spore prints

Tools

Numpy, pandas, sklearn, random forest classifier, xgb classifier, KNN, logistic regression, matplotlib, seaborn

Communication

A couple of my visualizations can be seen below: