Mushrooms can change your mind: A field guide for identifying edible mushrooms

By: Alex Ezazi

Abstract

The market for psychedelics is starting to legalize and there are start-ups scouring woods and forests for psychedelic mushroom species. The goal of this project is to provide a tool for field collectors to identify mushrooms as edible vs. poisonous as a preliminary filter to more extensive and expensive lab testing.

Design

The data was taken from the UCI repository (original source was the Audubon Society). Each observation is a single mushroom along with 21 physical characteristics as judged by the collector.

The main design objective is to create a model that is highly accurate, has as few features as possible, and the features are easily discernable in the field.

Data

The data consists of 8,480 rows and 22 columns. All data is categorical. The target variable is edible/poisonous. The 21 remaining feature columns have a total of 117 unique labels. The data was in very usable form already. 296 duplicates were eliminated.

The main challenge was how to best manage categorical data given my use case and, it turned out, computational limitations. After experimenting with generating dummy variables for all features (117), the best approach proved to be generating ordinal labels for categorical data. While this may not be a conventional approach, based on my research, use case, and this dataset, I believe it’s the best approach. See: https://www.kdnuggets.com/2015/12/beyond-one-hot-exploration-categorical-variables.html

Model

Preliminary logistic and random forest models indicated that the data is highly predictive with both models producing very accurate results. The challenge as to identify the best/fewest combination of features that still generated accurate predictions. F1 score, poisonous recall, and poisonous precision were used as scoring metrics.

All possible 3-feature combinations of the total 21 features were generated (1,330). Every combination was passed through both logistic and random forest models and scored; metrics were extracted and stored. Again, both models produced highly accurate results but with different feature combinations:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Feature Combination | F1 Score | Poison Recall | Poison Precision |
| Random Forest | Odor, stalk-surface-below-ring spore-print-color | .997 | .997 | .996 |
| Logistic | gill-spacing gill-size stalk-surface-above-ring | .946 | .932 | .952 |

The rf model feature combination was chosen for both accuracy and feature simplicity. A field guide was developed using a pair plot of the features and color coding.

Tools

* Pandas, numpy
* Seaborne, Matplotlib
* SciKit Learn
* Excel, Powerpoint