**Predicting The Edibility of *Agaricus* & *Lepiota* Mushrooms**

**Abstract**

The goal of this project was to use classification models to predict the edibility of mushrooms of both *Agaricus* and *Lepiota* genera, in order to find a good set of rules for identifying toxic mushrooms in the wild. I worked with data provided by the UCI Machine Learning Laboratory, applying purely categorical observations of mushrooms to four different classification models. All models proved to be highly accurate, with a small set of specific fungal features being the major determining factor.

**Design**

This project originates from a personal interest in mushroom hunting. The data, provided by the UCI Machine Learning Laboratory, classifies two very common genera of mushrooms as ‘**definitely poisonous**’, ‘d**efinitely edible**’, and ‘**edibility unknown and not recommended**’; this latter class is combined with the poisonous one in this analysis. Classifying edibility accurately with machine learning models may be able to shed light on core features of mushrooms that are hallmarks of being poisonous.

**Data**

The dataset contains 8,125 observations of mushrooms from either the *Agaricus* or *Lepiota* genus, with 23 features for each, including the target – all are categorical. A few feature highlights include a multitude of possible odor classifications, color types for various parts of the mushroom, habitat, and spore prints. Though these features could be grouped into more general categories, all features were applied to inform base models.

**Algorithms**

*Feature Engineering*

This data was remarkably clean to begin with, and minimal feature engineering was required. Categorical features were converted to binary dummy variables.

*Models*

Logistic regression, k-Nearest Neighbors, decision tree, and random forest classifiers were used before settling on the decision tree as the easiest model to interpret and the one most easily translated to action. SHAP values and histogram analysis, as well as domain knowledge, were used to guide the choice of included features as the model underwent refinement.

*Model Evaluation and Seletion*

The entire training dataset of 8,125 records was split into 80/20 train vs. test, and all scores reported below were calculated with 10-fold cross validation on the training portion only. This data is well-balanced according to the target, so accuracy was an appropriate choice of metric. However, due to the potentially serious consequences of a false negative, precision is a more practical choice – particularly precision on the edible class.

**Final decision tree 10-fold CV scores (training):**

* Accuracy: 0.925
* Recall: 0.92
* Precision: 0.94
* F1: 0.93
* AUC: 0.923

**Holdout:**

* Accuracy: 0.927
* Recall: 0.92
* Precision: 0.94
* F1: 0.93
* AUC: .924

**Tools**

* Numpy and Pandas for data manipulation
* Scikit-learn for modeling
* Matplotlib and Seaborn for plotting

**Communication**

In addition to the slides and visuals presented, *1UP or 6’ Under* will be available on my personal GitHub.