Preliminary Report - Edibility of Mushroom Species

Ishaan Saxena isaxena@purdue.edu
Nikita Rajaneesh nrajanee@purdue.edu
Swaraj Bhaduri sbhadur@purdue.edu
Utkarsh Jain jain192@purdue.edu

November 16, 2018

1 Introduction to the Problem

1.1 Definition of the Problem

Given a dataset $\mathcal{D}$ with $n = 8124$ samples where each sample represents a mushroom with features being the observations about the characterestics of the mushrooms such as odor, color, etc., we aim to test and compare various supervised learning models for the problem of classifying each sample into either poisonous or edible. Further, we will optimize the Hyperparameters of the model which initially performs the best on the dataset.[1]

1.2 Data Description

We are given $\mathcal{D}$ with $n = 8124$ samples wherein each sample has the following 22 features (excluding the class label).

1. cap-shape
2. cap-surface
3. cap-color
4. bruises
5. odor
6. gill-attachment
7. gill-color
8. stalk-root
9. stalk-surface-above-ring
10. stalk-surface-below-ring
11. stalk-color-above-ring
12. stalk-color-below-ring
13. veil-type
14. veil-color
15. ring-type
16. spore-print-color
17. habitat
18. gill-spacing
19. gill-size
20. stalk-shape
21. ring-number
22. population

These features have been further enumerated in Appendix A.

1.3 Encoding the Data

Note that all the features in our dataset are categorical variables. As a result, to proceed with evaluation of model performance, we must first encode these variables into numerical/binary values.

[1] This dataset can be found at https://www.kaggle.com/uciml/mushroom-classification.

Appendix A