Predicting Mushroom Class Based on Various Feature Sets Using Machine Learning

Andrew Jackson

Contents

Introduction

The goal of this project is to find which mushroom characteristics provide the most accurate results when determining whether or not that mushroom is safe to eat.

Data set

This project uses a single dataset which contains data on thousands of mushrooms. The first column in the dataset identifies whether the mushroom in each row is safe to eat or poisonous (p or e), referred to as the mushroom *class*. The rest of the columns contain categorical data ranging from physical appearance to habitat.

This dataset can be found here:
https://www.kaggle.com/uciml/mushroom-classification/data

The mushroom attributes used are:

CAPS

- cap-shape
- cap-surface
- cap-color

GILLS

- gill-attachment
- gill-spacing
- gill-size
- gill-color

STALKS

- stalk-shape
- stalk-root
- stalk-surface-above-ring
- stalk-surface-below-ring
- stalk-color-above-ring
- stalk-color-below-ring

Hypotheses

Running the predictor with the stalk feature set will provide the highest rate of accuracy because it has the greatest number of sub-categories. Running the predictor with the cap feature set will be the least accurate since there are very few “cap” sub-categories.

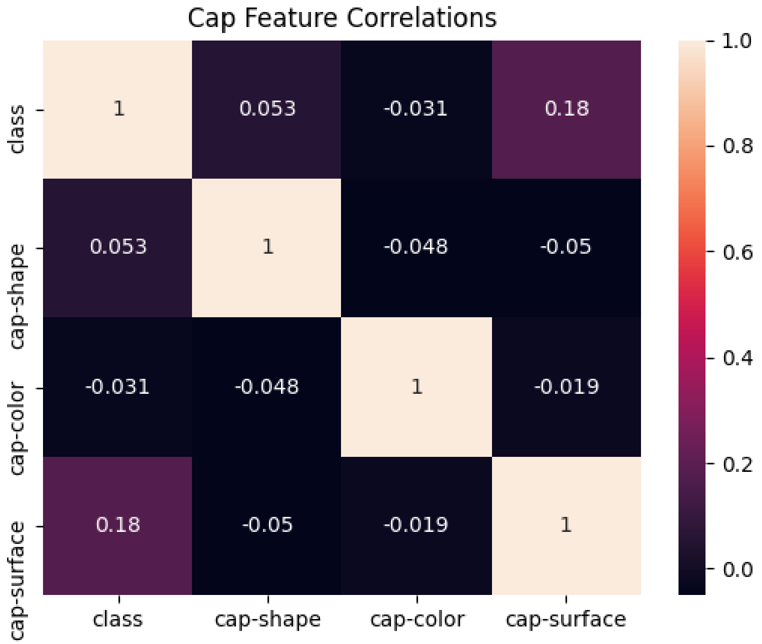
Test Plan

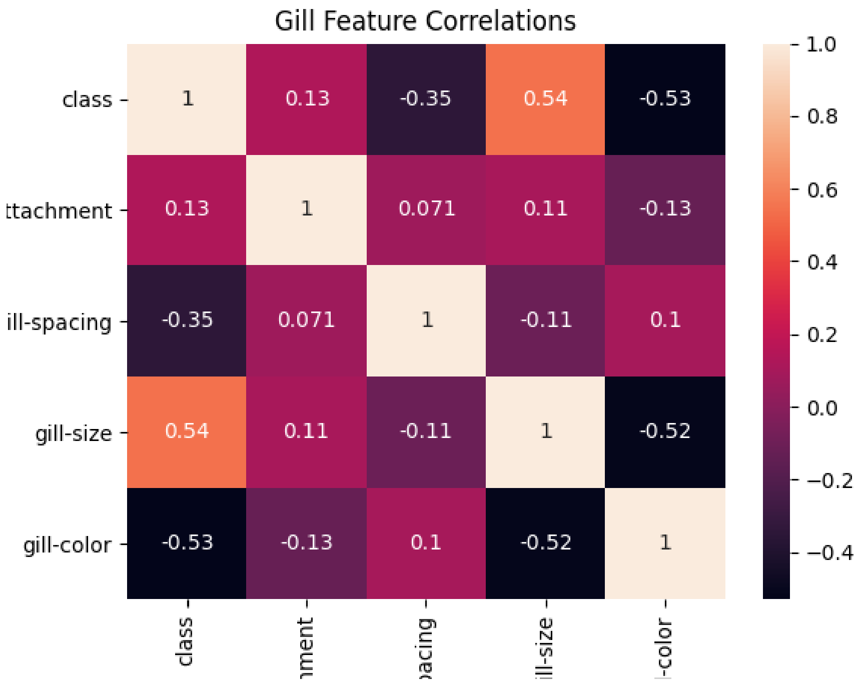
Two main steps will be taken to test this hypothesis:

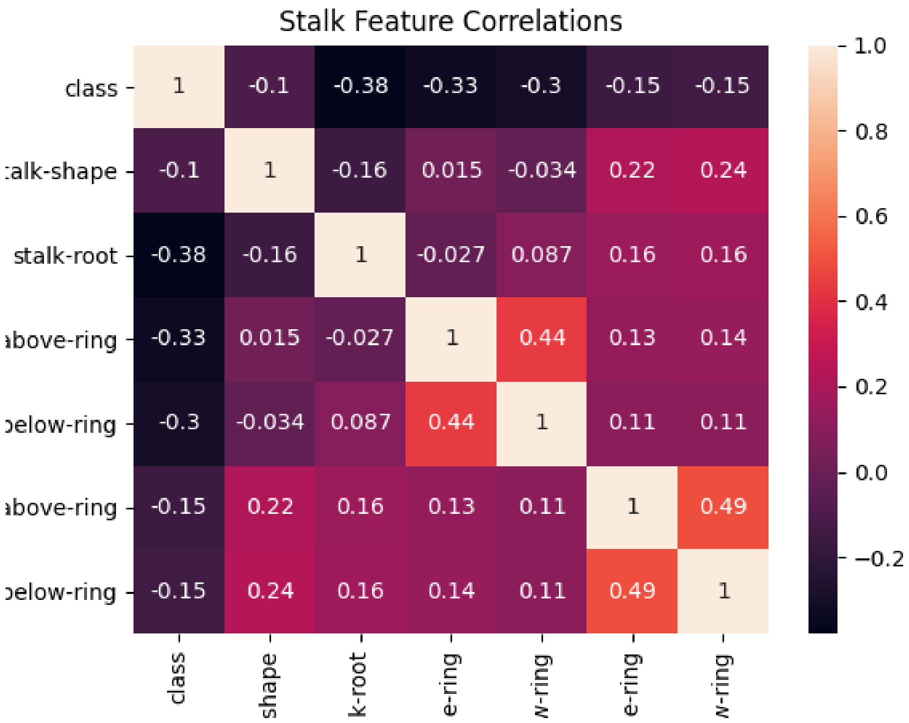
1. Correlation heat maps will be produced for each feature set in order to better understand the relationships between mushroom attributes

2. Three MLPClassifier models from the sklearn library will be trained based on the three different sets of features described above. For example, the first model will be trained on the whole dataset but only use the “cap-shape, cap-surface, and cap-color” features. The next will only use the “gill” features, etc. After training, the accuracy of each model will be determined by calling the mlp.predict() method and generating a confusion matrix.

Part 1. Correlation Heat Maps for Each Feature Set

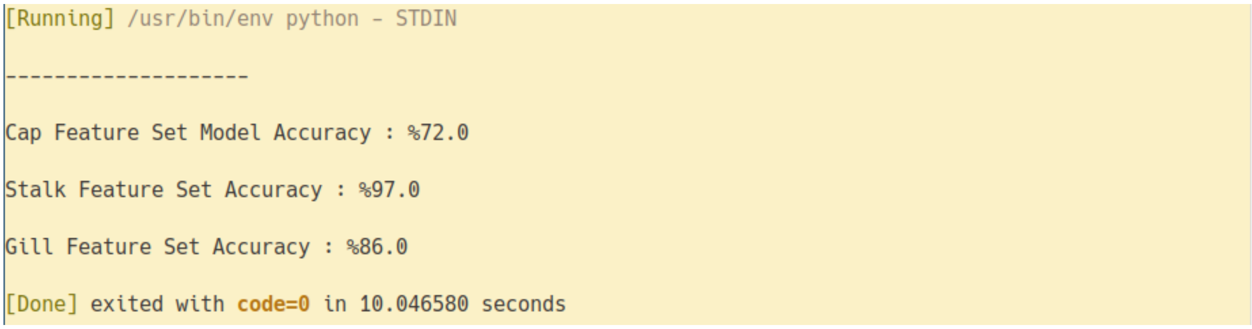



Gill Feature Correlations


Stalk Feature Correlations

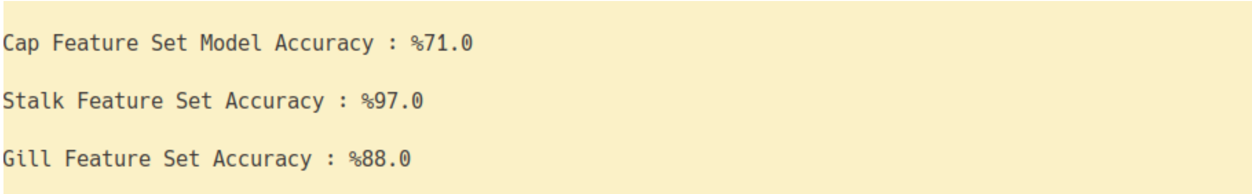
Part 2: Accuracies of Trained Model Predictions

Run #1



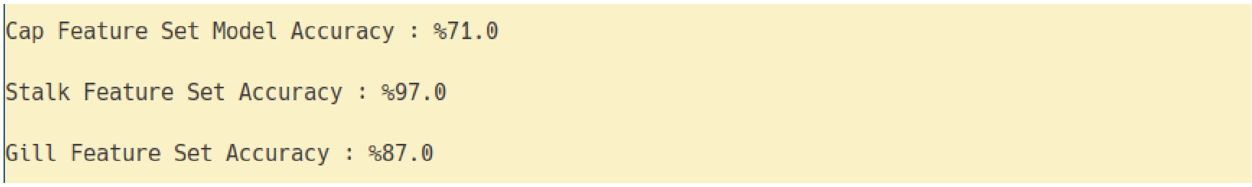
```
[Running] /usr/bin/env python - STDIN

--------------------

Cap Feature Set Model Accuracy : %72.0

Stalk Feature Set Accuracy : %97.0

Gill Feature Set Accuracy : %86.0

[Done] exited with code=0 in 10.046580 seconds
```

Run #2



```
Cap Feature Set Model Accuracy : %71.0

Stalk Feature Set Accuracy : %97.0

Gill Feature Set Accuracy : %88.0
```

Run #3



```
Cap Feature Set Model Accuracy : %71.0

Stalk Feature Set Accuracy : %97.0

Gill Feature Set Accuracy : %87.0
```

Results

1. Correlations

Across all the feature sets, the strongest correlations to class are shown to be gill-size (0.54) and gill-color (-0.53), making the gill feature set appear to be a strong candidate for creating a predictor with the greatest accuracy. The members of the cap feature set have the weakest relationship to class.

2. Predictions and Accuracies

Across the 1st, 2nd, and 3rd runs of the code, the feature sets consistently displayed the same order of accuracies, which was (from best to worst):

1. Stalk feature set
2. Gill feature set
3. Cap feature set

Conclusion

According to the results of the experiment, a mushroom's class (poisonous or safe) can be determined by stalk characteristics to a much higher degree of accuracy than other methods. An MLPClassifier trained on a mushroom dataset using only the stalk characteristics was able to consistently predict the mushroom's class at an accuracy of about 97%, while the cap feature set led to the worst accuracy. These results match the hypotheses made.

If I were to do this experiment again, I would test to see if mixing feature sets could allow for even greater accuracy. I would also look into using different machine learning models other than the MLPClassifier.