The Prediction of Mushroom Edibility Using Machine Learning

By: Ben Sturm

Project Summary

Project design

The goal of my project was to determine if machine learning could be used to predict whether a mushroom is poisonous or not based on some input parameters. These inputs consisted of key attributes for the mushroom such as color, odor, size, and habitat. In order to explore this subject further, I used the UCI mushroom data set. This data set consisted of 8127 individual mushroom samples from two different mushroom families, Agaricus and Lepiota.

Tools

Since my goal was to classify mushrooms into poisonous and nonpoisonous varieties, I used a variety of classification models from the sklearn library in Python. I also used other libraries such as pandas for data retrieval and preprocessing, matplotlib and seaborn for visualizations, and numpy for array manipulation and computations.

Data

As stated previously, the mushroom data set had 8127 instances of mushrooms consisting of a total of 23 individual mushroom species. The input variable consisted of 22 different features, which were all categorical. This meant that it was necessary to do one-hot-encoding in order to replace all categorical variables with one or more new features that have the values 0 and 1. After this step, I was left with 117 features, so my final input matrix was quite large. My target variable was edibility, were y=1 was for edible mushrooms and y=0 was for poisonous ones.

Algorithms

In my MVP Summary, I stated that I will be using logistic regression as my baseline model and then explore the performance of other models. However, to my surprise, logistic regression performed exceptionally well by just using the default parameters. The test set accuracy score with logistic regression was 0.999. I did also explore using 2 other models, in particular K-nearest neighbors and decision trees and achieved a perfect score of 1.000 with both models.

Results

Examining the logistic regression model further showed that the mushroom odor had a significant role in the prediction of edibility. We can observe that by exploring the beta coefficients of the logistic regression model provided below in Table 1.

Feature	β
odor_n	4.064
odor_l	2.656
odor_a	2.647
gill-size_b	1.959
gill-spacing_c	-1.522
stalk-surface-above-ring_k	-1.665
gill-color_b	-1.799
gill-size_n	-2.067
odor_p	-2.253
stalk-root_b	-2.287
odor_f	-2.608
odor_c	-2.638
spore-print-color_r	-3.331

Table 1. Logistic regression most important β-coefficients. Large positive β-coefficients are predictive of the mushroom being edible and large negative vales are predictive of the mushroom being poisonous.

In Table 1 we see that odor_n, odor_l and oder_a have relatively large positive β values whereas odor_p, odor_f, and odor_c have relatively large negative β values. Since for logistic regression $\log\left(\frac{p}{1-p}\right) = \sum_i^n \beta_i x_i$, features with larger positive β coefficients will be predictive of p>0.5 (i.e., edible mushrooms) and features with larger negative β coefficients will be predictive of p<0.5 (i.e., poisonous mushrooms). In fact, if we examine the distribution of mushroom edibility based on two features only, odor=none and odor=foul, we can already do a good job at classifying the mushrooms as shown in Figure 1.

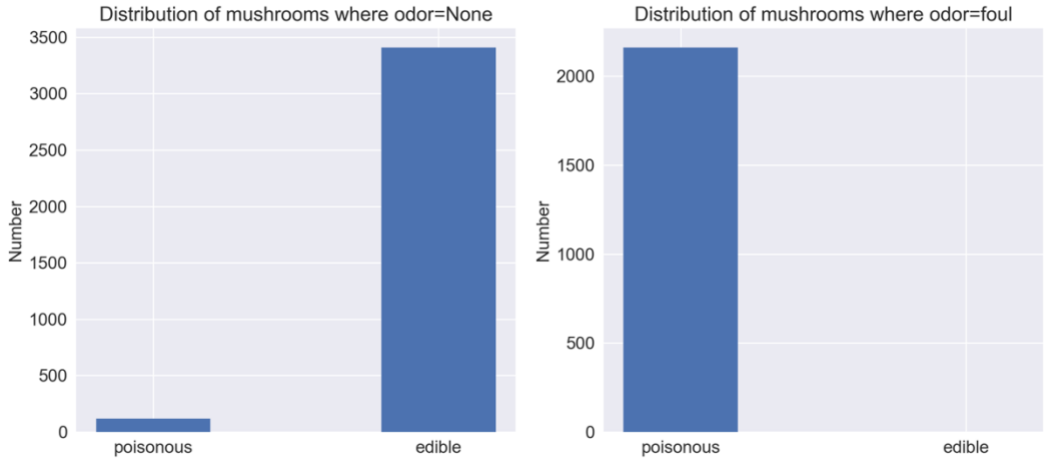

Figure1. Distribution of mushrooms for two features, odor=None (on left) and odor=foul (on right).

The final results for the three models that I ran are summarized in Table 2 below. Since we achieved an accuracy of 1.000 on our holdout data set, it wasn't necessary to explore other models or even to do model tuning.

Model	Precision	Recall	Accuracy	AUC
Logistic Regression	0.998	1.000	0.999	1.000
KNN	1.000	1.000	1.000	1.000
Decision Trees	1.000	1.000	1.000	1.000

Table 2. Hold out data set scores for the 3 different models explored.

A few important observations about this data set is that it consisted of 2 mushroom families in which odor plays a significant role in identification. However, this isn't always the case for other mushroom families, so it would be interesting to see how these models perform if we expanded our data set to include other mushrooms. In addition, one of the challenging aspects of mushroom identification is correctly identifying all of the attribute characteristics, which this data set has done for us. It could be interesting to see how robust our models are by leaving out key attributes or even swapping attributes, however that will be left for future work.