Predicting mushroom edibility

Abstract

The goal of my project was to accurately classify mushrooms into two different categories: either edible, or poison. I utilized UCI Machine Learning data to build a number of models, such as random forest, KNN, XGBoost, and logistic regression and eventually selected one for further testing.

Design

I identified this problem myself. After finding this dataset on Kaggle, I realized that foraging might be a fun activity that I, and many others, would like to attempt at some point in my life. However, without proper knowledge this simply would not be healthy or smart to do. I thought eventually a model like this could be combined with other methods to create a way of identifying mushrooms anywhere in North America without a guide's help.

Data

The dataset was made of 23 columns or features and 8124 individual data points. Pretty much every column was categorical so for certain models and methods some further modifications to the data was required.

Algorithms

Feature Engineering:

Convert much of the data to dummy variables using different methods such as OneHotEncode and get_dummies

Models

Logistic Regression, k-nearest neighbors (KNN), random forest, XGBoost with XGBoost being the final choice of model arbitrarily because the data was well balanced and different models could predict well.

Model Evaluation and Selection

The data was split into 80/20 for training and testing the XGBoost model that was selected. I used the score method in sklearn which evaluates accuracy, but also double checked this score with accuracy_score to be sure. Also F1 score which balances accuracy and precision

Final scores XGBoost model for test data: 95 features 1 predicted

Accuracy 1.0

F1 1.0

Precision 1.0

Recall 1.0

SHAP values were used to identify which features hold the most significance to the model.

Tools

Numpy and Pandas

Scikit-learn

Matplotlib and Seaborn

SHAP for evaluating feature importance

Communication

My findings and models will be posted on Github and presented using Google Slides and Google Docs.