Predicting which specific mushroom is poisonous or edible based on various features

Aalborg University Business School – Business Data Science

October 2023

Benjamin Ly, Camilla Dyg Hannesbo, Tobias Moesgård Jensen, Laura Veng Larsen

Submission date: 4th of October 2023

Key words: Python, Mushrooms, UML, SML

BUSINESS SCHOOL
AALBORG UNIVERSITY

Introduction

This analysis seeks to understand whether physical characteristics, namely the anatomy of the mushroom, as well as other features can be an indication of whether the particular mushroom is edible or poisonous.

"Mushroom poisoning is becoming one of the most serious food safety issues in China, which is responsible for nearly a half of all oral poisoning deaths … precise and timely species identification is of pivotal importance in mushroom incidents. More efforts and cooperation continue to be needed urgently for the governments, CDC staff, doctors, and mycologists in the future" (Li, H. et al. (2020)).

The overall motivation thus lies in making it simpler for the public to assess mushrooms and thereby independently, and correctly predict whether the mushroom is edible, which the interactive application is the fundamental solution to.

The structure follows the typical steps of initially analyzing the data exploratorily, followed by the step of processing- and cleaning the data. Then, the data was analyzed by way of unsupervised- and supervised learning. However, emphasizes supervised machine learning ("SML"), utilizing appropriate algorithms for predictive tasks. Furthermore, the SML model has been fitted, tested, and evaluated. Lastly, an interface was created to interact with the model.

Process

When conducting the analysis, we go through the structure step by step as we keep progressing when handling and analyzing the data.

Importing and loading the data

Initializing the process by importing the necessary libraries that are used throughout the analysis. Libraries such as pandas, sklearn, and xgboost were heavily used throughout the analysis.

Exploratory Data Analysis (EDA)

In the EDA phase, the data undergoes a thorough examination. Where we display and list the columns and rows for careful inspection as well as looking for missing values or anomalies. Visualization of the data helps us reveal some patterns, outliers, and potential relationships.

Data processing

Data processing involves cleaning the dataset of missing values, outliers, and inconsistencies. Feature engineering comes into play, where existing features are modified, or new ones are created to enhance the readability and the future model's capabilities. This will improve the dataset's quality and relevance.

Data Encoding

In the mushroom's dataset, the categorical variables need encoding to transform them into numerical format. Where OneHotEncode and LabelEncoder, were used to transform the categorical values into their representation in numerical value.

Fitting, Testing, and Evaluating the model

This pivotal stage involves the selection of a suitable machine-learning model. Initially, we chose to make an unsupervised learning where we will perform a principal component analysis to fit and transform our data. As well as reducing the dimensions for visualization purposes. The elbow method has been included to find the optimal number of clusters. For the clustering, we used KMeans as our unsupervised machine-learning algorithm.

For the supervised machine learning part, we have chosen to work with classification in a supervised machine learning model. We have fitted the training and testing data to see its performance on unseen data. We also used evaluation metrics such as accuracy, precision, recall, and F1 score to provide insights into the model's effectiveness.

Hyperparameter tuning using GridSearchCV

Hyperparameter tuning is used to optimize the model's performance as it systematically explores a range of hyperparameter values and identifies the combination that yields the best model.

Final evaluation

Here we have exclusively used Shap to understand the contribution of each feature to the model's predictions. As Shap values offer interpretability, revealing the impact of individual features on the model's output.

Creating a Data Pipeline

Creating a data pipeline ensures replicability and streamlines the workflow. The pipeline encapsulates the data processing and model training steps which provide an automated framework to work with.

Creating Gradio App

In the final step, the model is developed for practical use through a Gradio application. This allows users to interact with the model through their interface.

Main Findings

This section will highlight the main findings, as well as sub-findings. Ultimately, the dataset is unique, as it mainly has categorical values, which the analysis is characterized by.

In the Exploratory Data Analysis, it was found that 'veil-type' only had one unique value which is 'p', and secondly there were missing values, denoted with a question mark in column 'stalk-root', therefore these two columns were excluded. Also, the several bar charts depicted the target value and the features combined, which for instance in the Cap Shape revealed that

sunken (s) mushrooms are always edible, while conical (c) mushrooms are invariably poisonous, etc.

For our unsupervised machine learning model, we OneHotEncoded the data. We chose two components based on the explained variances which will also give us the possibility to visualize our PCA components after we performed the dimensionality reduction. The elbow method indicates that 3 clusters are the optimal number. Where the visualization interprets the 3 groups in clusters with similar features.

The supervised machine learning model was trained and tested, namely with a test size of 20%, deploying the Extreme Gradient Boosting Classifier, Random Forest Classifier, K-Nearest Neighbors Classifier, and Decision Tree Classifier, which revealed a 100% precision, recall and F1-score in all four models on predicting which mushrooms are edible or poisonous. This could indicate that the model might be overfitted and imbalanced, as the model may not generalize well to new and unseen data or the fact that the data is solely categorical which decreases the probability of false predictions. Though it can also indicate Appropriate Model Complexity. It's neither too simple (underfitting) nor too complex (overfitting). This means the model has the right balance of parameters and capacity to capture the relevant patterns in the data.

The Extreme Gradient Boosting Classifier was chosen as the model to proceed with, as the algorithm is good at predicting, performing and uses good parameters for finding relationships in data. The model revealed that the tested values for edible and poisonous were 845 and 780 respectively.

Furthermore, the top five most significant features are Spore Print Color, Odor, Stalk-Surface Above Ring, Ring Number, and Gill Spacing based on the SHAP values for the X data.

References

Li, H. *et al.* (2020) *Mushroom poisoning outbreaks - China, 2019*, *China CDC Weekly*. Available at: https://weekly.chinacdc.cn/en/article/doi/10.46234/ccdcw2020.005#:~:text=But%20with%20the%20utilization%20of,China%20(2%2D5 (Accessed: 01 October 2023).