Mushroom Classification

Poisonous vs Edible

[1]

CS 422 Term Project

Michael Evans & Grant Fitch
Professor Jiangwen Sun

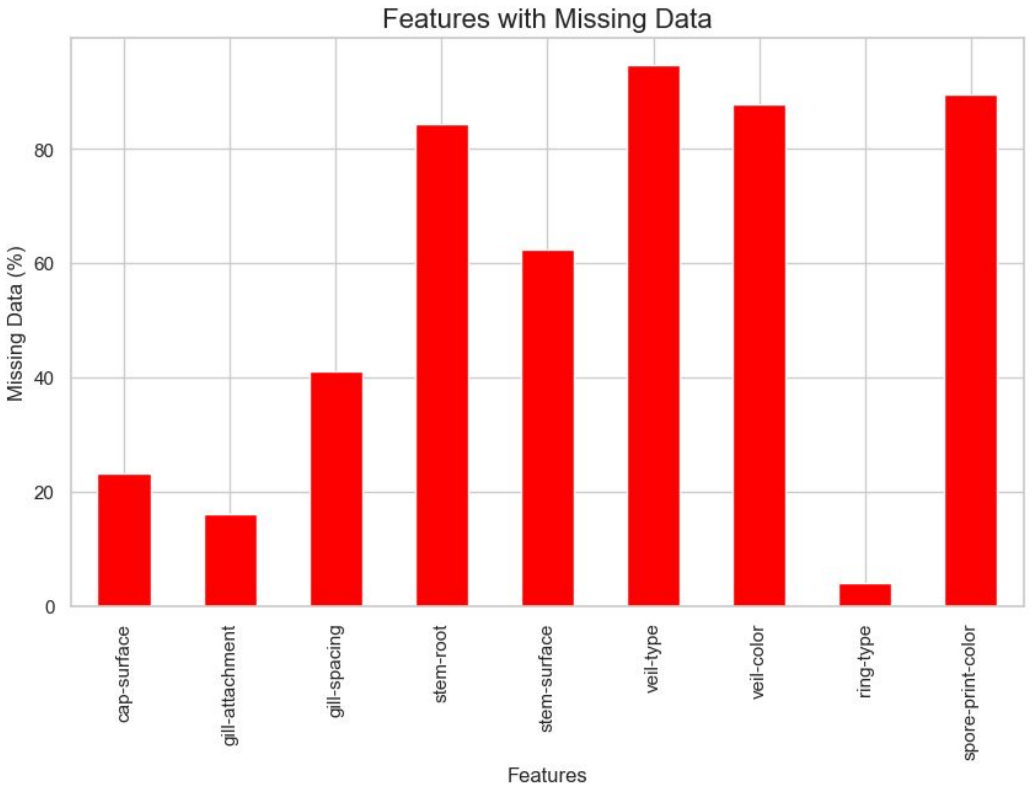
Mushroom Classification: Introduction

Background: With over 500 hospitalizations per year related to poisonous mushroom consumption in Germany [2], there is a demand for a robust and generalizable solution to automated mushroom classification.

Dataset: Our raw dataset comes from the UC Irvine Machine Learning Repository and was created based on 173 unique mushroom species. This set contains 61,069 samples with 20 features containing information about mushroom anatomy. This includes the color and shape of the mushroom cap, gills, stem, and veil, as well as the habitat, seasonal occurrence, and the target class of poisonous or edible . [4]

Mushroom Classification: Cleaning the Data

Missing Information: The first step of cleaning this data was to identify features that were missing data. 9 of the features in this table were missing information. Any feature with more than 20% missing data was removed from the set. After this step our dataset contained a target class with 13 features.



Mushroom Classification: Cleaning the Data

[6]

Missing Information: Next, samples were analyzed to determine where missing data persisted. It was determined that 12,000 of the samples were missing data from the 13 remaining features. These samples were dropped leaving 49,067 samples with complete information.

Mushroom Classification: Encoding the Data

Encoding Data: The 13 remaining features were divided into 3 categories. Categorical, Numerical, and Boolean.

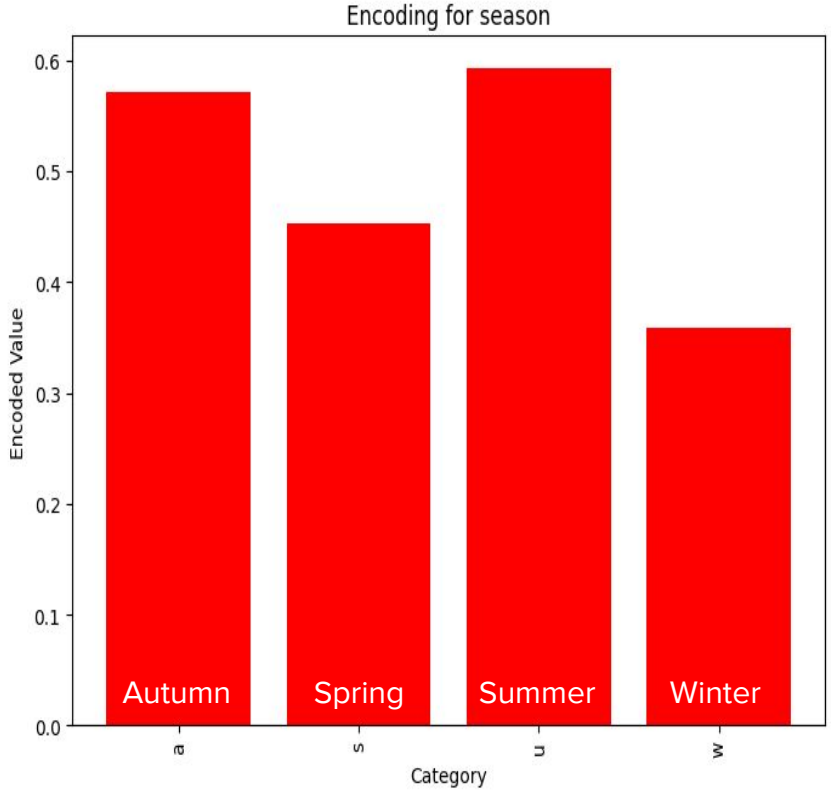
Boolean Encoding: The simplest of encoding, features such as ‘has ring’ are encoded with the method True = 1 and False = 0

Numerical Encoding: For numerical features such as stem height and width, additional cleaning can be performed. Using z scores, 2000 samples were identified as significant outliers and were removed from the dataset. The remaining numerical features were then normalized to provide consistency across the model.

Mushroom Classification: Encoding the Data

Categorical Encoding: Categorical Encoding is done for features such as cap, gill and stem colors, cap shape, and other non-numerical features. Target encoding is the process of replacing these features with the mean of the target variable, the likelihood that it is poisonous in this case.

Final Data: After preprocessing, our final dataset contains 47,051 samples with 13 features. This data is well split, containing 26,264 poisonous and 20,787 edible samples.



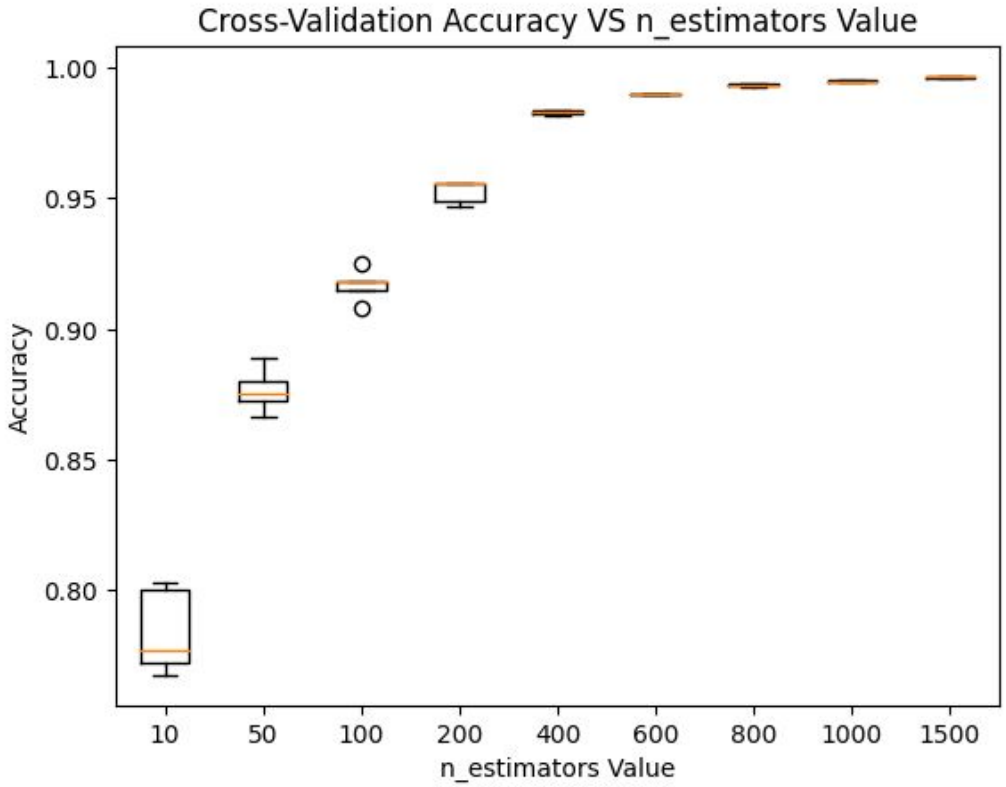
Mushroom Classification: Machine Learning Method

Gradient Boosting Classifier (GBC): This method is used for binary classification and is well suited for predicting if a mushroom is edible or poisonous. GBC builds a series of decision trees with each tree learns from the previous tree. This algorithm holds up well for large scale data containing a mix of numerical and encoded categorical variables.

K-Fold Cross-Validation: To evaluate the model's performance, the training data is split into 5 folds and the model is trained on each fold.

Mushroom Classification: Hyper Parameter Tuning

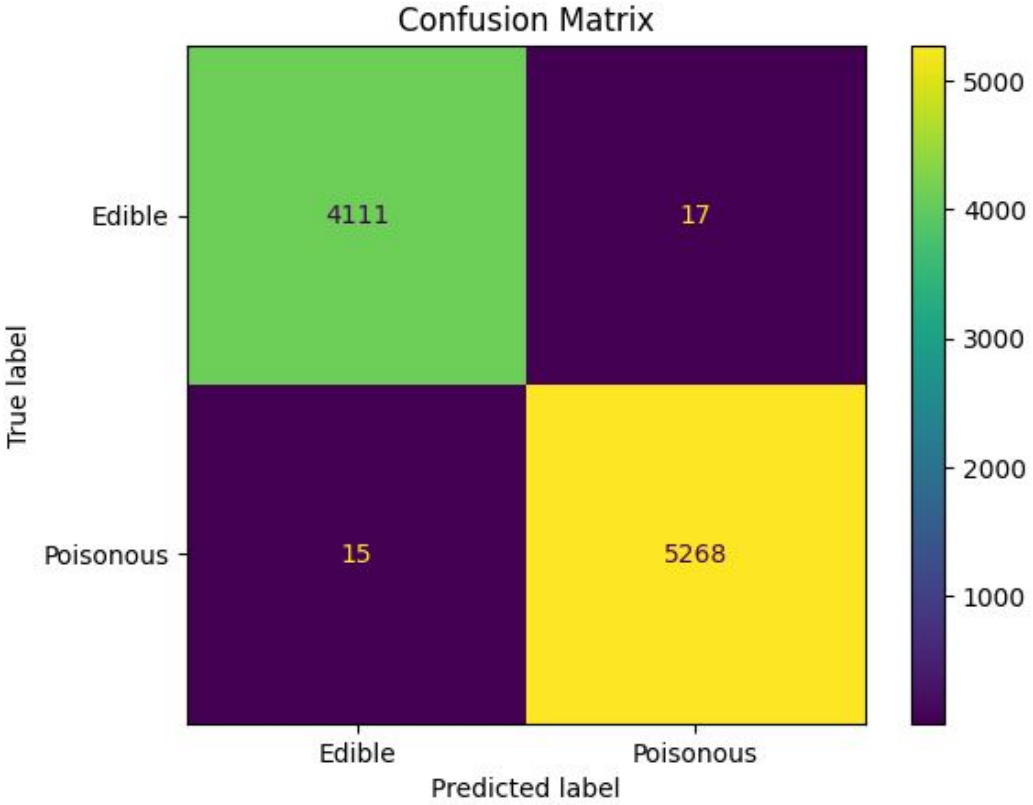
Hyperparameter Tuning: GBC uses n_estimators as a hyper parameter. By testing several values for n_estimators, an ideal fit can be determined.



Mushroom Classification: Model Evaluation

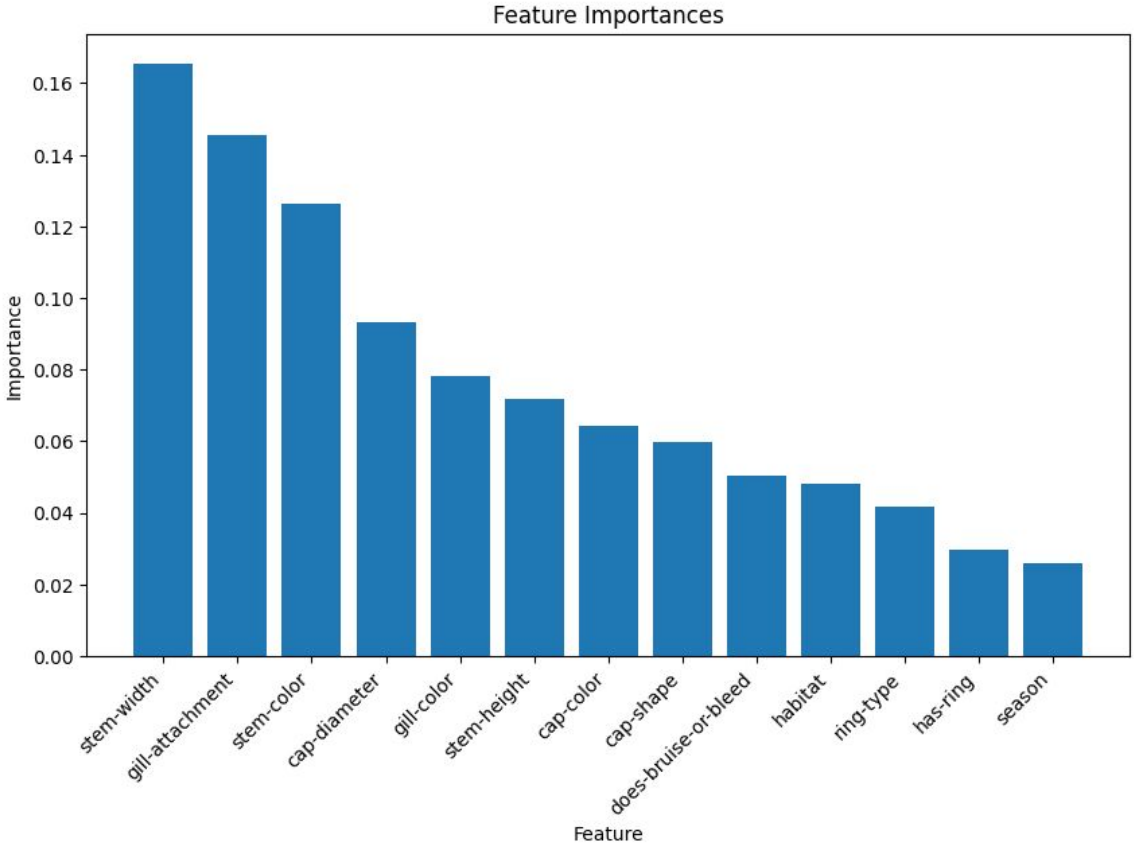
Confusion Matrix: Our testing data contains 4128 edible and 5283 poisonous mushrooms, for a total of 9411 training samples

```
Accuracy: 0.9965997237275529

Precision: 0.9967833491012299

Recall: 0.997160704145372

F1 Score: 0.9969719909159728
```



Mushroom Classification: Analyzing the Results

Importance Scores: These scores can offer a view into how our model was trained. While it is not a direct relation between feature and classification, it shows how our model trained. It does suggest that some physical features, such as the stem and gills, may have be better indicators for mushroom edibility than its habitat or season found.



References

[1] Image by Egor Kamelev from Pexels.
https://www.pexels.com/photo/closeup-photo-of-red-and-white-mushroom-757292/

[2] Wennig, R., Eyer, F., Schaper, A., Zilker, T., & Andresen-Streichert, H. (2020). Mushroom poisoning. Deutsches Ärzteblatt International, 117(42), 701.

[3] Tutuncu, K., Cinar, I., Kursun, R., & Koklu, M. (2022, June). Edible and poisonous mushrooms classification by machine learning algorithms. In 2022 11th Mediterranean Conference on Embedded Computing (MECO) (pp. 1-4). IEEE.

[4] Wagner, D., Heider, D., & Hattab, G. (2021). Secondary Mushroom [Dataset]. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/848/secondary+mushroom+dataset

[5] Image by Debivort on Wikipedia.
http://en.wikipedia.org/wiki/File:Mushroom_cap_morphology2.png

[6] Image by Andreas from Pixabay.
https://pixabay.com/photos/toadstool-mushrooms-mushroom-fall-4604694/

[7] Image by Andreas from Pixabay.
https://pixabay.com/photos/mushroom-moss-fall-forest-3816040/


Thank You For Listening
CS 422 - Mushroom Classification
Michael Evans & Grant Fitch