THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Classification of Mushrooms from the Agaricus and Lepiota Family

Aaron A. Gauthier and Pedro Uria Rodriguez

George Washington University – Machine Learning 1

27 November 2018

Problem Statement

Various guides clearly state that there is no simple rule for determining the edibility of a mushroom (some cites). Normally, one needs to identify the name of the mushroom to be able to tell if it is edible or not. However, this task is very difficult and time consuming for non-experts, so we thought that a machine learning approach on mushroom data could shed light on some of the features that almost guarantee a mushroom will be poisonous. This way, novices can avoid wasting time trying to identify a mushroom that will most likely end up being poisonous and focus on mushrooms that will most likely be edible instead.

Motivation

Realizing that many people in the USA do not know the difference between an edible mushroom and a poisonous one. We hope that through this project we can help educate people on the top significant features that determine whether a mushroom is safe for ingestion or not. We hope this will help save lives and avoid tragedy, especially with children who are curious.

Proposed Models

- Logistic Regression
- Decision Tree
- Random Forest
- K-Nearest Neighbors
- Gaussian Naïve Bayes
- Support Vector Machine
- K-Means Clustering

Exploratory Data Analysis

- Categorical data, 22 features
- 8,123 rows of observations – Large dataset!
- 2,480 missing values on *stalk-root* feature
  - **Imputed values as a percent of distribution (29% edible and 71% poisonous)**
  - **Imputed data with “most frequent” (mode)**

We can also impute the missing values by mode imputation

```
In [10]: dataset["stalk-root"].value_counts()
```

```
Out[10]: b    3776
         e    1119
         c     556
         r     192
         Name: stalk-root, dtype: int64
```

```
In [11]: # Imputes the missing values on the dataset
         from sklearn.impute import SimpleImputer
         imp = SimpleImputer(strategy="most_frequent")
         mushrooms_imputed = pd.DataFrame(imp.fit_transform(dataset), columns = columns)
```

NOTE: Uncomment and run the following cell if the above cell gives an error:

```
In [12]: #mushrooms_imputed = mushrooms.replace("?", "b")
```

```
In [13]: mushrooms_imputed["stalk-root"].value_counts()
```

```
Out[13]: b    6256
         e    1119
         c     556
         r     192
         Name: stalk-root, dtype: int64
```

but this is not such a good approach as the other values don't play any roll in the matter. A better approach would be to impute all classes on the same proportion as the non-missing data.

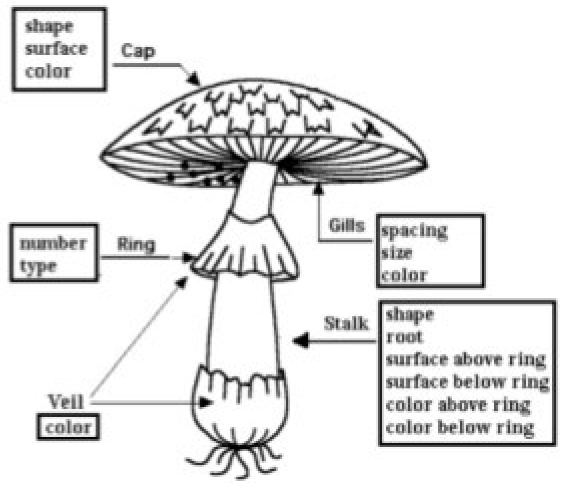
```
In [14]: # Subsets the dataset to only get the rows with missing values
         dataset_stalk_root_nan = dataset.iloc[
             dataset["stalk-root"][
             dataset["stalk-root"].isna()
         ].index
         ]
```

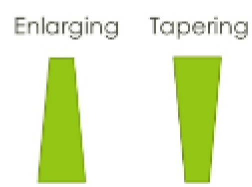
```
In [15]: dataset_stalk_root_nan["class"].value_counts()
```

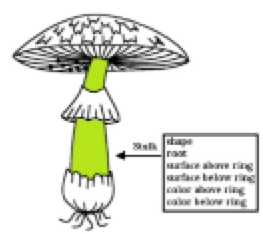
```
Out[15]: p    1760
         e     720
         Name: class, dtype: int64
```

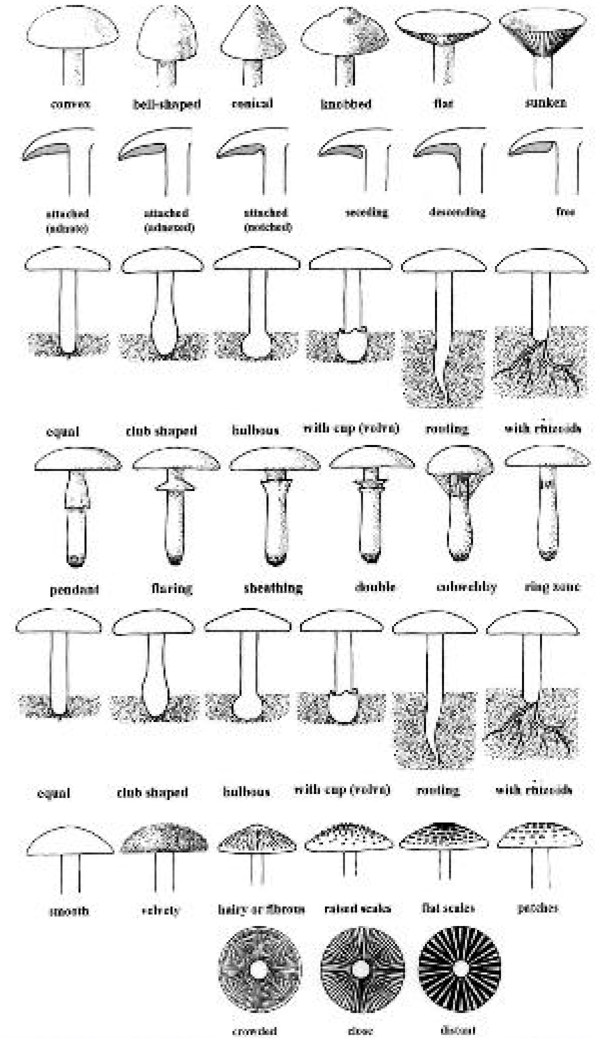
We can see that the rows with missing values are more of the poisonous class than of the edible... Thus, we should try to take this into account. Out of all the missing rows, 29 % would need to follow the dataset with only edible records distribution, while 71% would follow the poisonous dataset subset distribution, but the issue is that we don't know which ones are which, so the chance of imputting a wrong missing value is higher than with the mode, where we at least are guranteed to impute a percentage of values correctly (given they follow the same distribution). Thus, we will stick with the mode imputation approach.

Mushrooms









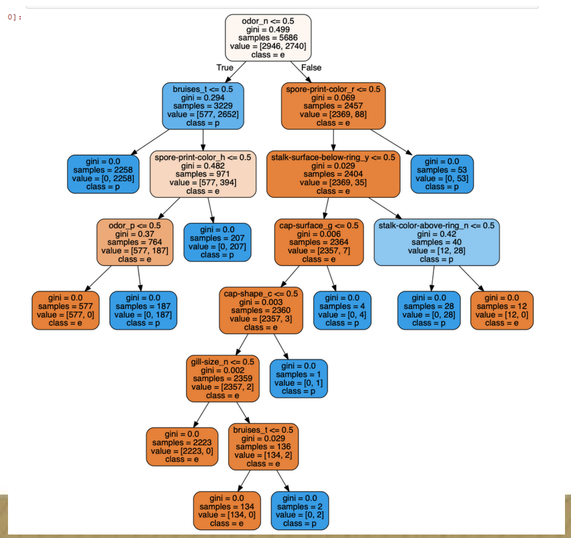
THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Logistic Regression

- Well suited for this type of problem
- Large dataset contributor to accuracy
- Strong linear relationship between features and edibility of mushrooms
- 100% accuracy score

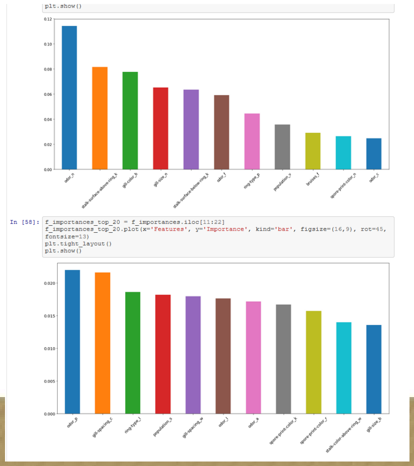
Decision Tree

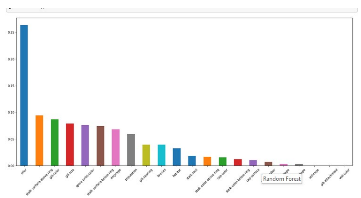
- Well suited for this type of problem – possibly best suited!
- Large dataset contributor to accuracy
- 100% accuracy score
- *Stalk-root* feature is a significant predictor
    - **Combination of other features can produce same accuracy for prediction**
- Decision tree changes based on the features that are utilized
    - **Result is unchanged – 100% accuracy score**



Random Forest

- Well suited for this type of problem – possibly best suited!
- Large dataset contributor to accuracy
- 100% accuracy score – no surprise!
  - **Ensemble of decision trees**
- 7-8 features required for 100% accuracy
  - **Mushroom explorers have to remember less – Good news!**
- Used to select feature importances!





K-Nearest Neighbors

- Well suited for this type of problem
- 100% accuracy score
- Large dataset contributor to accuracy

Gaussian Naïve Bayes

- Well suited for this type of problem using the smaller data set & hyperparameter tuning
- Full dataset
  - **Initial score – 98.75%, Improved score – 98.77%**
- Smaller dataset
  - **Initial score – 99.88%, Improved score – 100%**

- **Note: Improvements due to hyperparameter tuning

Support Vector Machine

- Suited for this type of problem after hyperparameter tuning
- Full dataset
  - **Initial score – 98.83%, Improved score – 100%**
- Smaller dataset
  - **Initial score – 99.79%, Improved score – 100%**

Note: Improvements due to hyperparameter tuning

K-Means Clustering

- Not well suited for this type of problem
- Accuracy score – 90.44%

Conclusion

- Experts have determined there are not specific set rules to classify mushrooms as edible or poisonous
  - **Algorithms can predict accurately**
  - **Improved false positives through imputation of data – 1 out of 1,175!**
- Acknowledgement: There is no substitute for domain expertise!
- Possible dangers despite high accuracy results:
  - **Mistakes made in assessments of mushrooms**
  - **Sample size could be too small**
  - **Not all inclusive – only for Agaricus and Leopiota**
  - **Different mushroom species could have identical traits**
    - **Same Traits = [could] different results**
- Models should only augment expertise
- Legal and ethical considerations – domain expertise always overrides model classification!

References

[1] Sebastian Raschka and Vahid Mirjalili. *Python Machine Learning*. Packt, Birmingham, 2017.

[2] Machine Learning I’s notes, slides, exercises and homework.

[3] Wikipedia’s entries for the various classifiers.

[4] https://arxiv.org/pdf/1410.5329v3.pdf

[5] A bunch of other articles and coding questions, all referenced in the *jupyter notebook*.