Mushroom Classification with Machine Learning

Adam Sansone

12/5/20

University of Illinois at Springfield

CSC 532 Introduction to Machine Learning

Abstract

This project tries to utilize a vast number of physical characteristics of mushrooms to determine if the mushroom is edible or poisonous. Several machine learning models were used to classify mushrooms into those two categories, edible or poisonous. The models that were used in this study include: Naïve Bayes, C5.0, Bagged CART, Random Forest, Ada Boost, Gradient Boost, and a convolutional Neural Network. These models were trained on 8124 observations of 23 variables that represented different physical attributes of the mushrooms. The study found that of the models used all but Naïve Bayes was able to perfectly classify the mushrooms observations in the test set.

Problem Definition and Project Goals

The purpose of this project is to classify mushrooms as either edible or poisonous based on a variety of attributes that were included in the dataset. The dataset was obtained from Kaggle.com who in turn obtained the dataset from the UCI Machine Learning repository. The data was originally found in a book called The Audubon Society Field Guide to North American Mushrooms (1981). The dataset consists of 23 features:

- Cap-shape
- Cap-surface
- Cap-color
- Bruises
- Odor
- Gill-attachment
- Gill-spacing
- Gill-size
- Gill-color
- Stalk-shape
- Stalk-root
- Stalk-surface-above-ring
- Stalk-surface-below-ring
- Stalk-color-above-ring
- Stalk-color-below-ring
- Veil-type
- Veil-color
- Ring-number
- Ring-type
- Spore-print-color
- Population
- Habitat

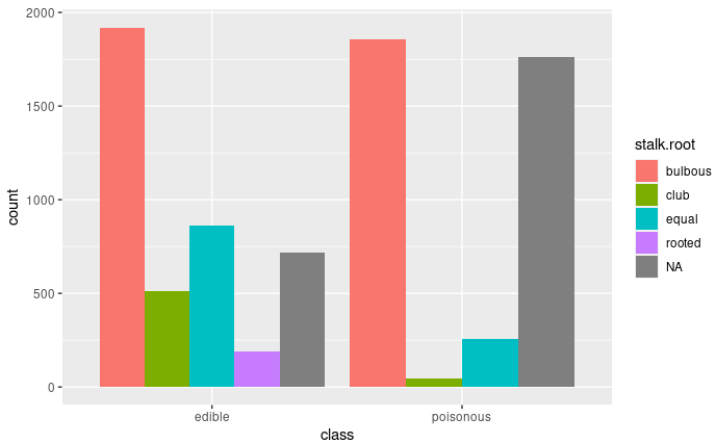
The features included are all categorical and the levels of each are represented as a single letter. The data will first be visualized using side-by-side bar graphs as well as tested against the class variable for significance using the chi-squared test. This will allow me to observe how the data relates and make necessary modifications before putting the data into the models.

Related Work

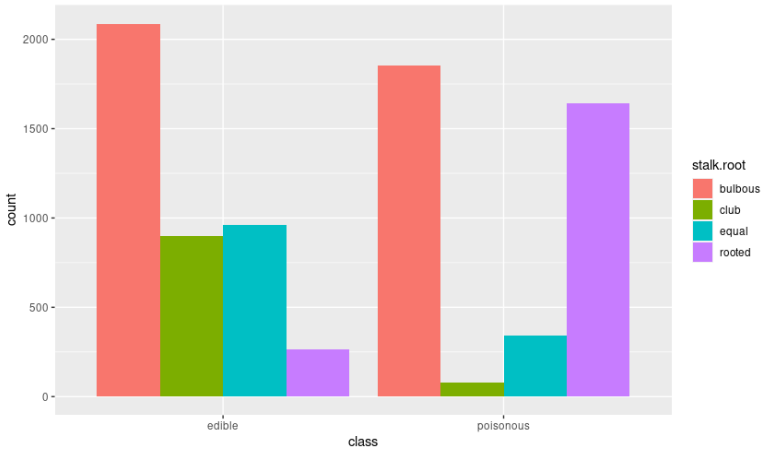
I was not able to find any scholarly papers or textbooks regarding classifying mushroom edibility.

Data Exploration and Preprocessing

The dataset contained 8124 observations each one coded as a single letter. The first step was to rewrite these letters into the actual name of the level. This would help to clarify the data for further data exploration and visualization. Next, I found that one variable contained quite a lot of missing values. The variable stalk root which contained 2480 “?”’s indicating a missing value. Since this was such a significant number rather than removing the feature, I decided to impute it. I utilized the mice library in R to impute the date.

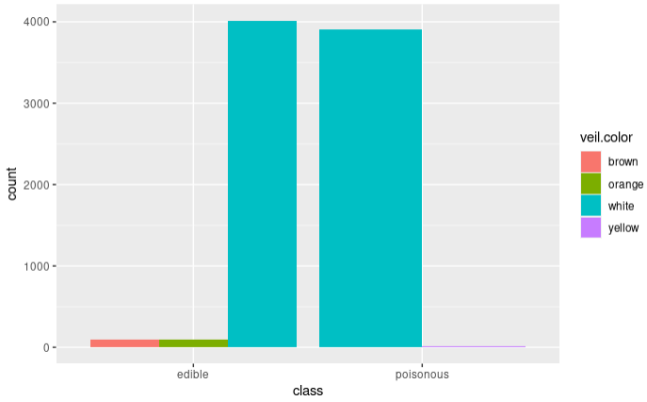


Before imputation

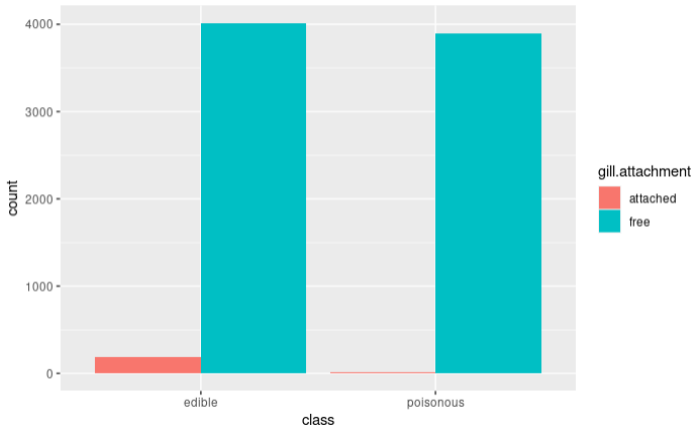


After Imputation

The imputation increased the count for most of the levels and also created a level in the poisonous mushroom class that was not there before. The next step I took was to remove the Gill Attachment and Veil Color features since they contained only 2 levels and 99% of the data was in one of those levels for both class variables.

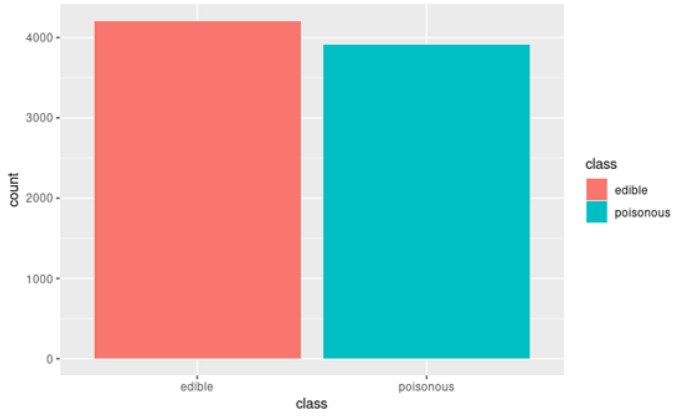


Veil Color

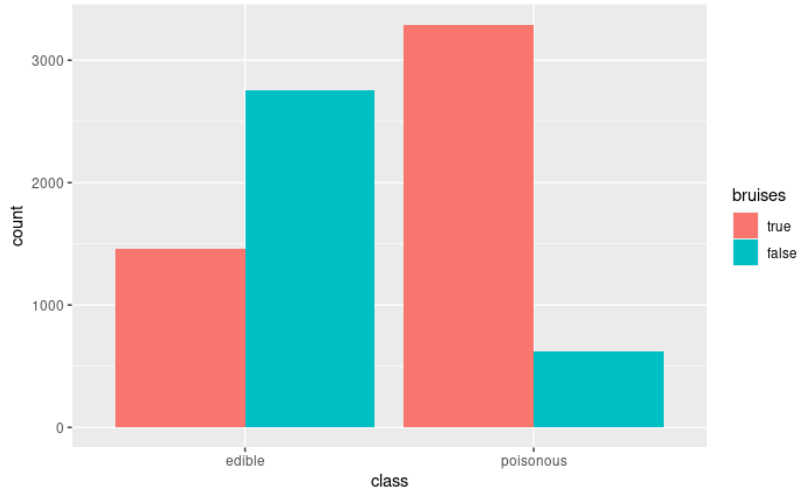


Gill Attachment

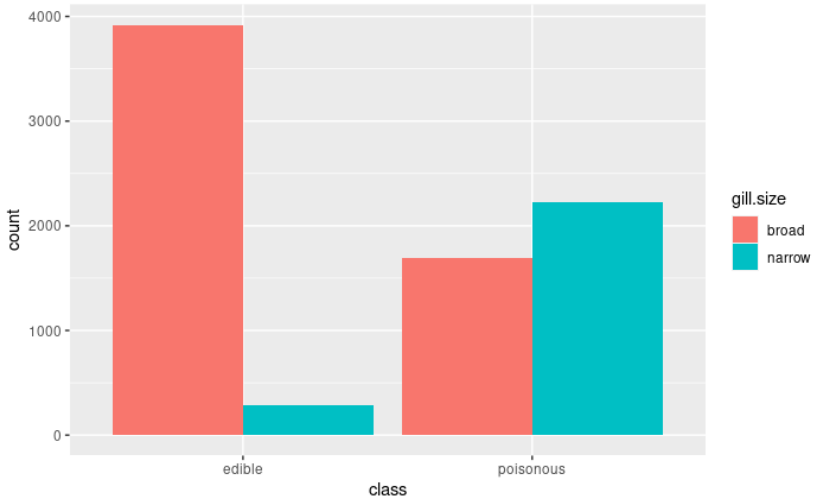
I also removed the Veil Type variable because it only contained a single level. I next combined levels in the following features into a single level called “other” since those levels did not have a significantly high count to being with, Stalk Surface Above Ring, Stalk Surface Below Ring, Stalk Color Above Ring, Stalk Color Below Ring, Ring Type, Spore Print Color. Since the models I will be using, except for the Neural Network, can take categorical features I will not one-hot-encode them until I get to the Neural Network model. The last observation I made was regarding the balance of the class variable which contained ~52% edible observations and ~48% poisonous observations.



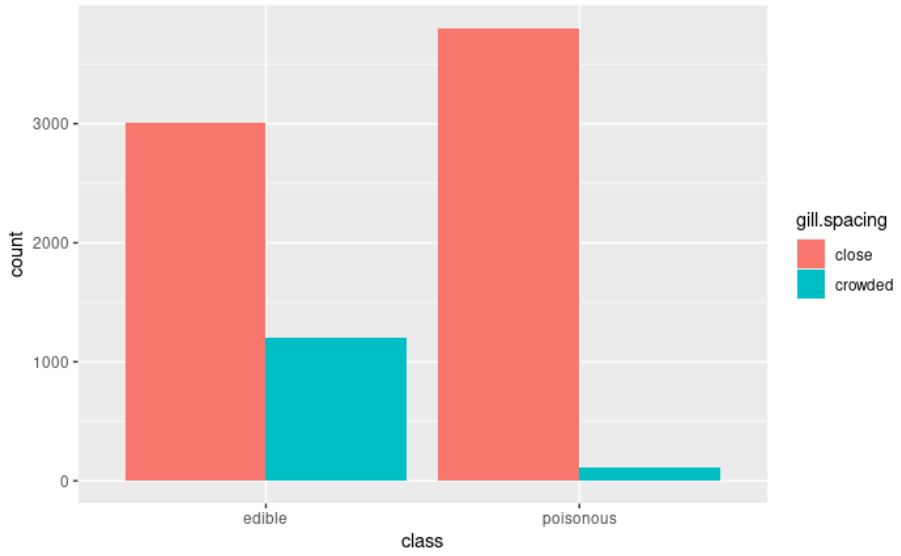
After observing the data, I ran each one through the chi-squared test which returned p-value’s less than 0.05 indicating a strong association between each one and the class variable. Let’s take a look at some of the more significant relations.



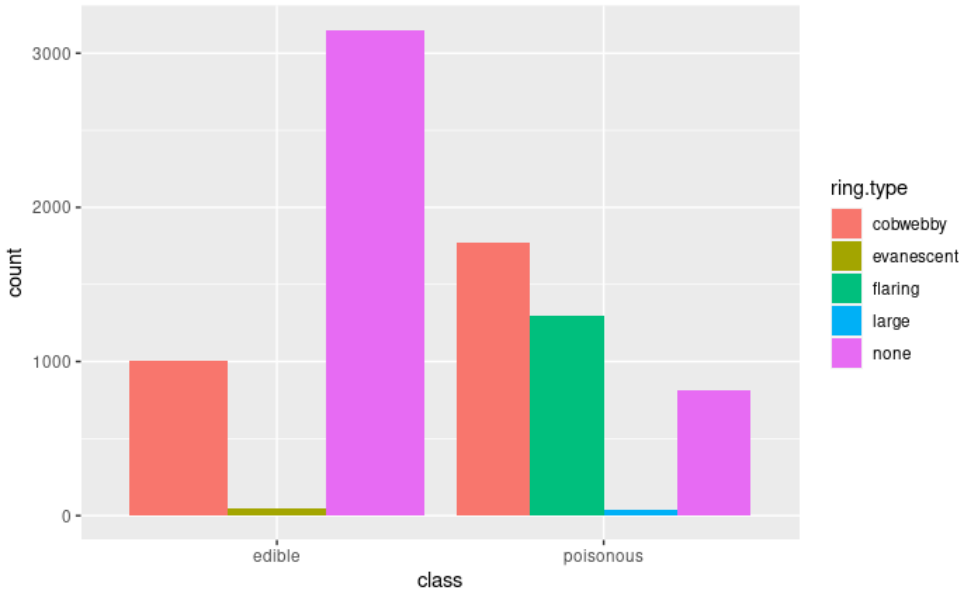
The Bruising graph shows a clear distinction between edible in poisonous mushrooms, in that poisonous mushrooms have a larger count of bruising while edible mushrooms don't.



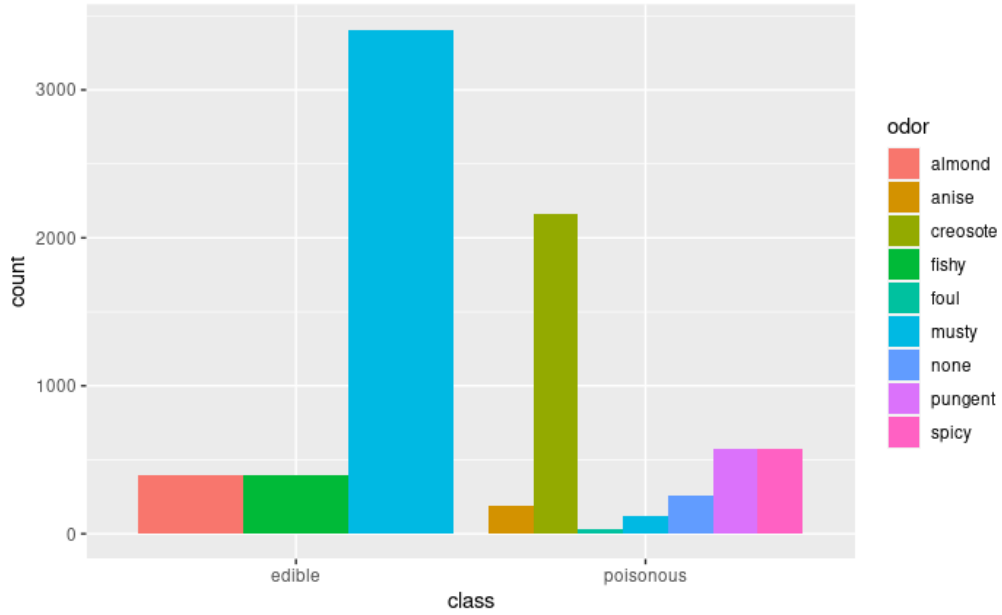
The Gill Size graph shows that edible mushrooms have far more broad gills than poisonous mushrooms. Poisonous mushrooms have an almost balanced count between the two.



The Gill Spacing graph shows that poisonous mushrooms have almost no observations that have crowded gill and are more likely to have gills that are close together.



The Ring Type is also quite interesting since the edible mushroom mostly consist of having no rings and poisonous mushrooms have a large number of flaring rings which edible mushrooms don't have at all.



Odor is another variable that shows a clear distinction between the two classes. The poisonous mushrooms have more of a creosote odor while edible is mustier. Edible mushrooms only have 3 distinct odors while poisonous can have seven!

Data Analysis and Results

The following models were used:

- Naïve Bayes
- C5.0
- Bagged CART
- Random Forest
- Ada Boost
- Gradient Boost
- Neural Network

The data was split first into a training set (80%) and a test set (20%) for the initial models. I split the training set further into a validation set once it came time to do the neural network. I let caret autotune all of the hyper-parameters except for the neural network which I created a tuning run for. I also used 10-fold cross-validation for each model. Every model returned a perfect accuracy of 1, except for Naïve Bayes which returned a 0.93. Although this does seem good it also had 60 false negatives, which when dealing with something deadly is not very good.

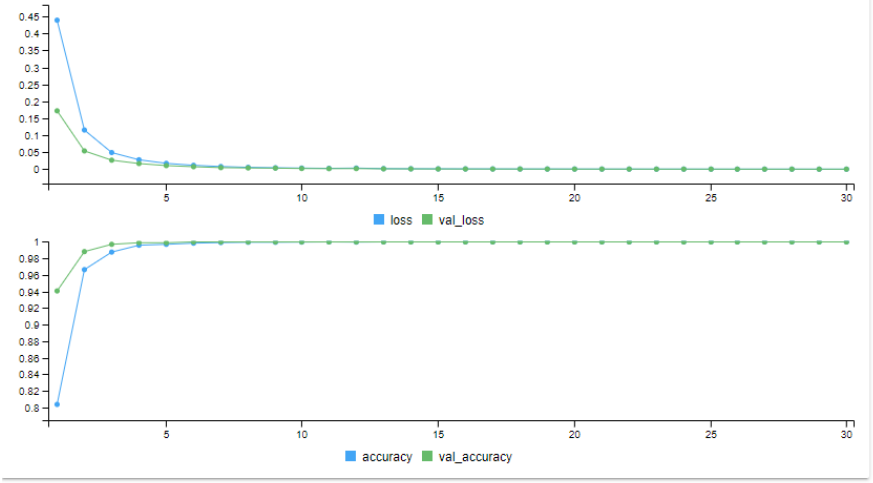
```
Call:
summary.resamples(object = compare)

Models: NB, C, B, RF, A, G
Number of resamples: 10

Accuracy
        Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
NB 0.9155146 0.9264531 0.9307159 0.9318432 0.9393241 0.9492308    0
C  1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000    0
B  1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000    0
RF 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000    0
A  1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000    0
G  1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000    0

Kappa
        Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
NB 0.8306892 0.8525433 0.8611896 0.8634561 0.8784246 0.8981955    0
C  1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000    0
B  1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000    0
RF 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000    0
A  1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000    0
G  1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000    0
```

The neural network was tuned using a tuning run, the majority of the models returned a very high accuracy so the first one in the list was chosen to use on the test set.

run_dir <chr>	metric_loss <dbl>	metric_accuracy <dbl>	metric_val_loss <dbl>	metric_val_accuracy <dbl>
1 runs/2020-12-02T05-17-35Z	4.0000e-04	1.0000	3.0000e-04	1.0000
2 runs/2020-12-02T05-17-32Z	4.3500e-02	0.9902	3.5100e-02	0.9961
3 runs/2020-12-02T05-17-18Z	6.9310e-01	0.5180	6.9310e-01	0.5360
4 runs/2020-12-02T05-17-10Z	3.2000e-03	1.0000	2.4000e-03	1.0000
5 runs/2020-12-02T05-16-59Z	1.1100e-02	0.9994	9.7000e-03	1.0000
6 runs/2020-12-02T05-16-45Z	3.3965e-06	1.0000	2.8623e-06	1.0000
7 runs/2020-12-02T05-16-40Z	1.5600e-01	0.9550	1.4140e-01	0.9698
8 runs/2020-12-02T05-16-35Z	1.5200e-02	0.9983	1.1900e-02	0.9990

The first run contained the following parameters:

History



Run

context	local
script	mushrooms.R
started	2020-12-04 21:52:00 GMT
time	00:00:04

Metrics

loss	0.0003
accuracy	1.0000
val_loss	0.0003
val_accuracy	1.0000

Flags

nodes	64
batch_size	200
activation	tanh
learning_rate	0.001
epochs	30

Optimization

loss	categorical_crossentropy
optimizer	<tensorflow.python.keras.optimizer_v2.adam.Adam>
lr	0.001

Training

epochs	30

Model

```
Model: "sequential_7"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_23 (Dense)             (None, 64)                6336
_________________________________________________________________
dense_22 (Dense)             (None, 64)                4160
_________________________________________________________________
dense_21 (Dense)             (None, 2)                 130
=================================================================
Total params: 10,626
Trainable params: 10,626
Non-trainable params: 0
```

Conclusion

Overall, all the models performed well. There is definitely opportunity with the Naïve Bayes model that could be due to some of the variables having a lot of levels. Since every model performed well continued research would be to try and find the least computationally intense model that can classify the data. It did seem that there were a few features that were marked as important in each of the models more than others, such as odor, bruising, gill-size, and stalk root. Using Principal Component Analysis or Embedding Vectors might improve the performance and alleviate some of the computation needed for all of the levels of data.

References

Mushroom Classification Safe to Eat or Deadly Poison. (n.d.). Kaggle. Retrieved 12/5/20 from https://www.kaggle.com/uciml/mushroom-classification.