Juliet Maharaj

Introduction to Data Science & Machine Learning

Professor Iordan Slavov

**Statistical Learning Techniques of Mushroom Data**

Introduction:

This dataset contains descriptions of assumed samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms (1981). The dataset was included in the UCI Machine Learning repository. Each species belongs to one of the categories, edible, definitely poisonous, or of unknown edibility and not recommended. Information such as edibility is important for people in the mushroom hunting world. It allows them to carefully collect more useful information on mushrooms. This dataset provides information that we can use to see which characteristics are commonly attributed to either edible, poisonous or unknown edibility mushrooms.

Our main focus is to create an accurate model with our machine learning tools that will conclude the correct type of edibility a mushroom can be classified as. We will be examining the logistic regressions, the random forest and the KNN. We will weigh the accuracy of these models by using the accuracy rate and the area under the curve (AUC).

Description of Data:

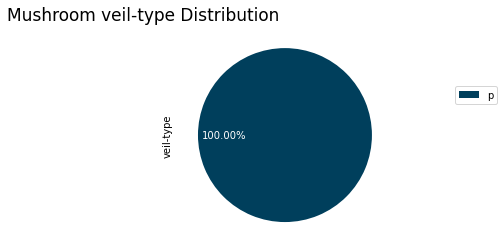
This dataset contains samples of mushrooms. Each column represents a new mushroom with its own features. Our main focus is the feature of whether or not the mushroom is edible or not. We will examine the other features of this dataset and see which features are significant and which are not. The dataset consists of 23 variables.

* class: edible=e, poisonous=p
* cap-shape: bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
* cap-surface: fibrous=f, grooves=g, scaly=y, smooth=s
* cap-color: brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
* bruises: bruises=t, no=f
* odor: almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s
* gill-attachment: attached=a, descending=d, free=f, notched=n
* gill-spacing: close=c, crowded=w, distant=d
* gill-size: broad=b, narrow=n
* gill-color: black=k, brown=n, buff=b, chocolate=h, gray=g, green=r, orange=o, pink=p, purple=u, red=e, white=w, yellow=y
* stalk-shape: enlarging=e, tapering=t
* stalk-root: bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?
* stalk-surface-above-ring: fibrous=f, scaly=y, silky=k, smooth=s
* stalk-surface-below-ring: fibrous=f, scaly=y, silky=k, smooth=s
* stalk-color-above-ring: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
* A picture containing text, linedrawing

  Description automatically generatedstalk-color-below-ring: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
* veil-type: partial=p, universal=u
* veil-color: brown=n, orange=o, white=w, yellow=y
* ring-number: none=n, one=o, two=t
* ring-type: cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z
* spore-print-color: black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y
* population: abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y
* habitat: grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d

First, we can examine the distributions of these variables. We can check for any patterns or abnormalities with these pie charts.

Chart, pie chart

Description automatically generated

Chart, pie chart

Description automatically generatedChart, pie chart

Description automatically generated

Chart, pie chart

Description automatically generatedChart, pie chart

Description automatically generated

Chart, pie chart

Description automatically generatedChart, pie chart

Description automatically generated

Chart, pie chart

Description automatically generatedChart, pie chart

Description automatically generated

Chart, pie chart

Description automatically generatedChart, pie chart

Description automatically generated

Chart, bar chart

Description automatically generatedChart, bar chart

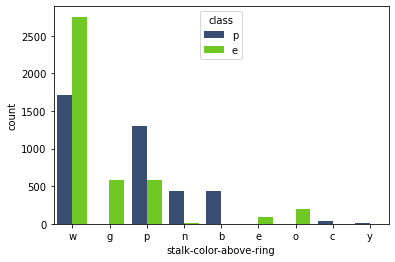
Description automatically generatedThere are many observations that we can construct from the distribution of these variables. In this entire dataset, there is only one mushroom veil-type, which is partial. We can also see that there are multiple distributions that only have two factors. This observation may be able to lead us to the conclusion that we want, whether or not the mushrooms are edible. Also, there exist a variety of mushroom gill colors. An example of a question we can try to answer with our dataset is, does any particular mushroom gill color correspond to the mushroom’s edibility? Next, we will examine some graphs that contain the variables and the count of their edibility.

Chart

Description automatically generated

Chart, bar chart

Description automatically generated



Chart

Description automatically generated

These bar charts are interesting to our initial interest in which characteristics of a mushroom are most significant in determining their edibility. The navy-blue bar represents poisonous. The green bar represents the mushroom being edible. You can see how each category of the variable displays the total number of edible and poisonous mushrooms. Let’s examine spore print color. The categories are black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y. We can see that black and brown spore print color are more likely to be edible. We can also see that chocolate and white are more likely to be poisonous, based on these bar charts. There are not many categories that have an equal amount of poisonous and edible mushrooms. One assumption we can say is that certain categories of these variables are attributed to their edibility. This can be helpful in determining the edibility of unknown mushrooms.

Next, we are going to examine more complex ways of predicting the edibility of a mushroom. First, we are going to modify the class variable. The category edible will know be represented by 1. The category poisonous will be represented by 0.

Chart, scatter chart

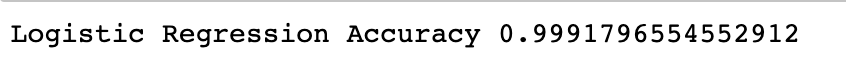
Description automatically generated

The correlation coefficient pair plot shows the correlation between all the variables. There are a few variables that we could say are highly correlated. The variable odor none seems to be highly correlated with a few variables, such as stalk surface below ring scaly, stalk surface above ring scaly, and ring type none. Also, gill attachment none is highly correlated with veil color white, stalk color below ring yellow, stalk color below ring brown, and stalk color above ring brown. On the other hand, there are many variables that have no correlation to the other variables. An example of a variable that has little to no correlation in this dataset is gill color green, black and orange. The correlation matrix helps us determine which variables would be most significant in a model. We can use the most significant variables in the process of predicting the edibility of mushrooms.

Analysis:

Now, we must randomly split the data into two sets, the test and training dataset. The training data will be randomly sampled from 70% of the initial dataset.

**Logistic Regression**

I started off with a logistic regression because it is most useful in probability prediction. I used a logistic regression because our main focus is to determine which variables are significant in a mushroom being either poisonous or edible. The logistic regression I created contained all of the initial variables because they were all significant enough to be included. The logistic regression accuracy came out to be 0.9991796554552912. A logistic regression accuracy value is going to be between zero and one. Zero being a badly fit model and one being a perfectly fit model. Accuracy is the proportion of correct predictions over total predictions. Being that the logistic regression accuracy came out to 0.9991796554552912, shows us that we have a model that fits our dataset significantly. 

Next, we can examine the confusion matrix associated with our logistic regression.

Text

Description automatically generated

We can see that the true negative, in this case would equate to edibility being poisonous given we predicted the edibility as poisonous, is entirely more significant than a false negative. There are actually no false negatives, which confirms the fact that we have a significantly fit model. Along with the fact that the true positive number is significantly higher than the false positive number. This means that the number of mushrooms that were predicted to be edible, were predicted edible correctly more often than not. There were only two cases of predicting a mushroom being edible when it actually was not.

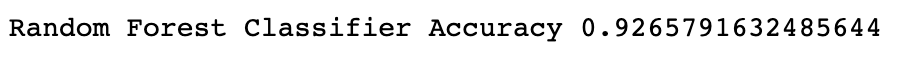
**Decision Tree**

A decision tree is a useful visual method in predicting the outcome of a dataset. A decision tree is a convenient tool because it can be used for both categorical and numerical data. And like all of the other machine learning tools we use, we can validate our model by checking the accuracy. Being able to see how reliable your data is, is important because if you have a great model, that is not accurate, it’s totally irrelevant. Each branch of a decision tree represents a possible outcome for our model. Our dataset has twenty-three variables which may make our decision tree a bit complex. But we will only check the accuracy of our model to determine if it is significant or not.

With our dataset, the decision tree accuracy value came out to one. This is the best outcome that we could receive. This means that the model containing all of our initial variables forms the best model in predicting the edibility of mushrooms. If our accuracy was lower, we would try to determine which variables were significant and which variables were not significant, and then form a new model with only significant variables.

**Random Forest**

We can run a random forest test to determine a significant model for our dataset. A random forest is an estimator that fits multiple decision trees in various subsets, then averages all of that data to improve the predictive accuracy of our model. We ran a random forest on our data and the random forest classifier accuracy came out to 0.9265791632485644.



The accuracy of our random forest test wasn’t as high as our first two test, but still significantly high. The random forest classifier was 92% accurate. If it were below 90%, I would have tried to determine which variables could be removed for being insignificant. Although 0.9265791632485644 is not as good as 1, it is still significant enough to keep all of our variables in this model for predicting edibility of mushrooms.

Conclusion:

To conclude, we used logistic regressions, decision trees and random forest machine learning methods to predict the edibility of mushrooms. By calculating these values, we created a significant model that can be used to predict the edibility of mushrooms. This is important for the shroom hunting world. This model could be the determining factor in whether or not someone dies from a poisonous mushroom. All of the test that we ran had an accuracy of above 90%, which is amazing. From these accuracy levels, we can say that the model with all of our variables is valid. If we didn’t initially receive such a high accuracy level, we would have determined which variables individually were most significant in our model. Once we found the variables most significant variables, they would be used in a new more reliable model. All of our variables were significant, so our model was accurate on the first try. The plot below displays the accuracy levels that we calculated for each test.

Chart, scatter chart

Description automatically generated

Chart

Description automatically generatedAnother way to compare the performance of our model is by using the AUC, area under the curve.

The AUC for the model we created came out to be one, which is as perfect as it could get. Being that the accuracy on all of the test we ran came out to above 90%, I would assume the AUC would also be high. This means that we have the best possible model for our dataset. The mushroom data that we started with must have been cleaned and regularized already because it is uncommon to have such a perfect model on the first try with all initial variables. Normally, we would have to clean the data, then run multiple machine learning test to see which variables are significant enough to be included in a perfect model.

Chart

Description automatically generatedLastly, another way to check the validity of our model is to examine the true positive and false positive rates. As expected, the graph shows that our model is as good as it can be.

In conclusion, there is nothing different that I would do to improve this model. We got lucky with a dataset that was in good shape for machine learning tests. Now, shroom hunters can use this model when they are trying to determine the edibility of their unknown mushroom. They can compare its characteristics to the model we produced to predict whether or not it is edible. Hopefully this will be a helpful tool for mushroom hunters in the future!