Cassy DeBlois
Data Science
15 May 2023

Machine Learning Mushroom Models

Summary

For this project, I analyzed the Mushroom Data Set from the UCI Machine Learning Repository. The data set contains 8124 data points of hypothetical samples of 23 species of gilled mushrooms in the *Agaricus* and *Lepiota* families. Each data point has 22 mushroom attribute fields that contained a single-letter indication of which category within the attribute the hypothetical mushroom had, along with one field for if it is edible or not. However, no specific identification was assigned to any of the data. After some light data cleaning, I used five different classification models to predict the edibility of the samples. Surprisingly, several of the models were able to perfectly predict the edible mushrooms, but I determined that the best model is the decision tree since it could potentially be used in the field like a dichotomous key.

Methods

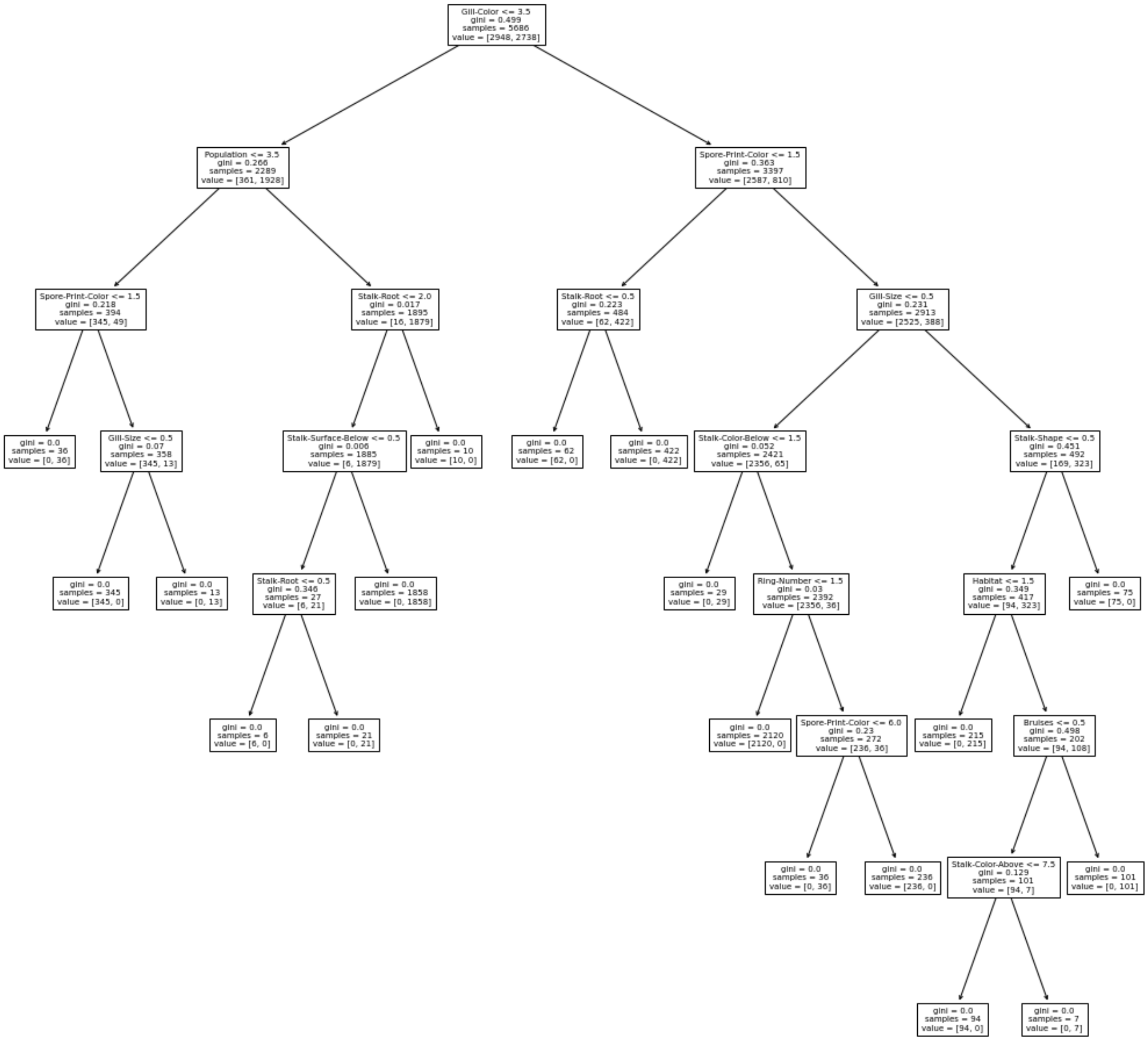
Most of the methods used in this project were from sklearn in which I frequently consulted documentation pages for assistance with using the various method calls. To start, I read the data csv into a pandas data frame and used sklearn's ordinal encoder to change all the data fields from categorical to numerical data so that the machine learning models could easily use them. I dropped the "veil-type" category since it had many missing data points and then tested correlations, but decided to use all the columns to make a model anyway since I could adjust later if I needed to. Then, I started working on the models which were of the types: linear regression, support vector machine, k nearest neighbors, decision tree, and Gaussian Bayes. For each one, I set several hyperparameters that could be tested and ran them through a grid search to find the best options (with the exception of Gaussian Bayes since it didn't have a hyperparameter that I could test). Lastly, I collected the confusion matrix, f1 score, and accuracy of each for analysis.

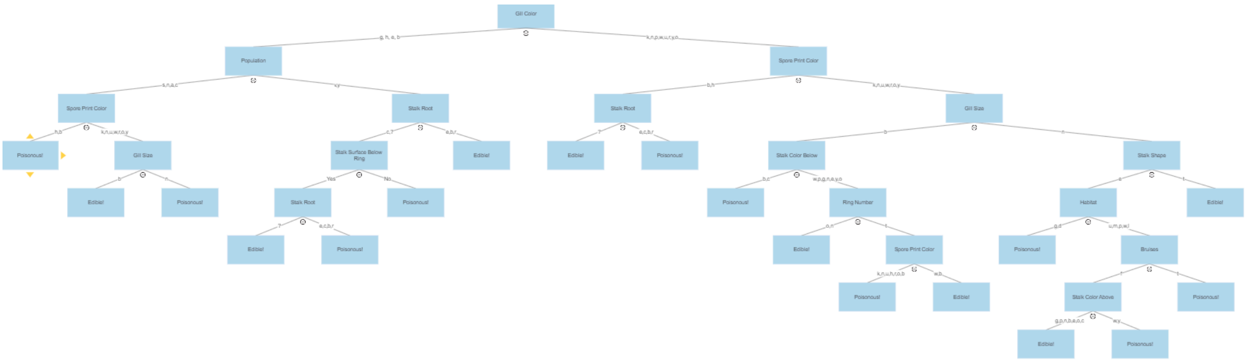
Results and Discussion

Figure 1. Results of each model selection grid search and the metrics of the best classifiers.

Classifier	Linear Regression	Support Vector Machine	K Nearest Neighbor	Decision Tree	Gaussian Bayes
Best Hyperparameters	fit_intercept = True C = 15 penalty = l2	kernel = poly C = 1 degree = 5	n_neighbors = 10 weights = distance	max_depth = 10 min_samples_leaf = 5	None
F1 Score	0.956	1.0	1.0	1.0	0.914
Accuracy	0.957	1.0	1.0	1.0	0.918
Confusion Matrix	1197 | 54 50 | 1137	1251 | 0 0 | 1187	1251 | 0 0 | 1187	1251 | 0 0 | 1187	1139 | 112 87 | 1100

I was quite surprised to see that three of the five classifiers perfectly matched which mushrooms were edible since I know this is a common issue among mushroom foragers as there are no general guidelines that can be followed to determine edibility. However, of the models that perfected the classification, I thought that the decision tree classifier would be most beneficial to use since someone in the field may not have access to a laptop, but could use a decision tree like a dichotomous key. However, printing out the decision tree (shown below) yields complicated results because all the data was read in as numerical values so I had to create my own version of the visualization that could be used. Unfortunately, after making the visualization with a website, I found out that saving it in a more legible format is paid content so I apologize for the poor quality. The first tree is from sklearn and the second is from the website.


Gill-Color <= 3.5
gini = 0.499
samples = 5686
value = [2948, 2738]
Population <= 3.5
gini = 0.266
samples = 2289
value = [361, 1928]
Spore-Print-Color <= 1.5
gini = 0.363
samples = 3397
value = [2587, 810]
Spore-Print-Color <= 1.5
gini = 0.218
samples = 394
value = [345, 49]
Stalk-Root <= 2.0
gini = 0.017
samples = 1895
value = [16, 1879]
Stalk-Root <= 0.5
gini = 0.223
samples = 484
value = [62, 422]
Gill-Size <= 0.5
gini = 0.231
samples = 2913
value = [2525, 388]
gini = 0.0
samples = 36
value = [0, 36]
Gill-Size <= 0.5
gini = 0.07
samples = 358
value = [345, 13]
Stalk-Surface-Below <= 0.5
gini = 0.006
samples = 1885
value = [6, 1879]
gini = 0.0
samples = 10
value = [10, 0]
gini = 0.0
samples = 62
value = [62, 0]
gini = 0.0
samples = 422
value = [0, 422]
Stalk-Color-Below <= 1.5
gini = 0.052
samples = 2421
value = [2356, 65]
Stalk-Shape <= 0.5
gini = 0.451
samples = 492
value = [169, 323]
gini = 0.0
samples = 345
value = [345, 0]
gini = 0.0
samples = 13
value = [0, 13]
Stalk-Root <= 0.5
gini = 0.346
samples = 27
value = [6, 21]
gini = 0.0
samples = 1858
value = [0, 1858]
gini = 0.0
samples = 29
value = [0, 29]
Ring-Number <= 1.5
gini = 0.03
samples = 2392
value = [2356, 36]
Habitat <= 1.5
gini = 0.349
samples = 417
value = [94, 323]
gini = 0.0
samples = 75
value = [75, 0]
gini = 0.0
samples = 6
value = [6, 0]
gini = 0.0
samples = 21
value = [0, 21]
gini = 0.0
samples = 2120
value = [2120, 0]
Spore-Print-Color <= 6.0
gini = 0.23
samples = 272
value = [236, 36]
gini = 0.0
samples = 215
value = [0, 215]
Bruises <= 0.5
gini = 0.498
samples = 202
value = [94, 108]
gini = 0.0
samples = 36
value = [0, 36]
gini = 0.0
samples = 236
value = [236, 0]
Stalk-Color-Above <= 7.5
gini = 0.129
samples = 101
value = [94, 7]
gini = 0.0
samples = 101
value = [0, 101]
gini = 0.0
samples = 94
value = [94, 0]
gini = 0.0
samples = 7
value = [0, 7]


Gill Color
g, h, e, b
k,n,p,w,u,r,y,o
Population
Spore Print Color
s,n,a,c
v,y
b,h
k,n,u,w,r,o,y
Spore Print Color
Stalk Root
Stalk Root
Gill Size
h,b
k,n,u,w,r,o,y
c,?
e,b,r
?
e,c,b,r
b
n
Poisonous!
Gill Size
Stalk Surface Below Ring
Edible!
Edible!
Poisonous!
Stalk Color Below
Stalk Shape
b
n
Yes
No
b,c
w,p,g,n,e,y,o
e
t
Edible!
Poisonous!
Stalk Root
Poisonous!
Poisonous!
Ring Number
Habitat
Edible!
?
e,c,b,r
o,n
t
g,d
u,m,p,w,l
Edible!
Poisonous!
Edible!
Spore Print Color
Poisonous!
Bruises
k,n,u,h,r,o,b
w,b
f
t
Poisonous!
Edible!
Stalk Color Above
Poisonous!
g,p,n,b,e,o,c
w,y
Edible!
Poisonous!

Reflection

I ran into two main issues with the project. First, the regression classifier would not converge and the support vector machine had difficulty running which meant that to test my code I has to wait a long time before any results came up. I was worried that it would never be complete at times or that I had some significant error that caused it difficulty. I should have specified in the grid search or the model when to cut off trying to converge to save time, but I instead just would wait each time. I also had difficulty with making a visualization for the decision tree since there was no obvious way for me to convert the numerical labels back into categorical labels.

We covered classification models in class, but I was grateful for this project because I was able to more specifically learn the differences between each classification type and make a decision on which was best for the context of my project. I definitely understand why many data science questions lead to the answer of “it depends” since there is no fit-all solution with data.