Mushroom Classification: a Decision Tree Model

November 02, 2021

A. Introduction

A1. Background

With over 50,000 mushroom species in North America, one can only imagine the great deal of usage this has. Whether used in foods as an ingredient or in medicine for its medicinal properties, mushrooms are a necessity. But if one were to consume it, how would they know if it is safe to do? Poisonous mushrooms can be hard to distinguish in the wild!

Predicting the class of a mushroom with great accuracy is an objective that a mushroom enthusiast might look into. This highlights the need for tools that can accurately predict mushroom classification using features/attributes about the given mushroom. In this project, we would take a data-driven approach for the prediction of mushroom class using Machine Learning Model.

A2. Question

For this project, the question that can be asked: “How Decision Trees can be used to predict whether a mushroom is edible or poisonous based on certain attributes?”

A3. Target

The intended audience for this project are researchers who are interested in classifying mushroom using Machine Learning models. As there are many features/attributes included in this dataset, they might be keen to know which specific features are necessary to label classification.

B. Data Acquisition and Cleaning

B1. Data Source

We need the following data:

1. Dataset contains different species of mushrooms:

Data Source = Kaggle: https://www.kaggle.com/uciml/mushroom-classification

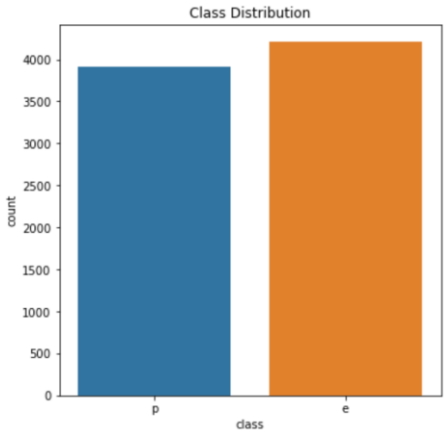
Description: The Kaggle Dataset consists of 8124 samples and 23 features/attributes.

The column features are:

- Classes: 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-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
- Stalk-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

B2. Data Distribution

Before doing the analysis, we thought it would be best to visualize the distribution of the “class” feature to get some insights.



From the image above, there are more mushrooms that were classified as “edible” than there are “poisonous”. Using a specific function, we know that out of 8124 samples, there are 4208 considered edible and 3961 being poisonous.

B3. Data Cleaning

The Kaggle Dataset initially have to be converted into a CSV file which was then downloaded in the data assets column in Watson Studio. The file has 8124 entries and 23 attributes.

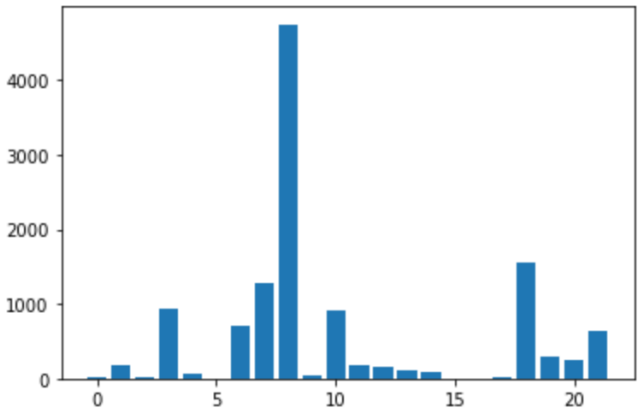
We initially took the information about the dataset. This is to see what type each attribute is and to see if there are attributes that have missing values. When ensured that there is none, the pre-processing stage is started.

The initial step taken was to separate the features and the target variables. The features included all attributes except the “class” column which is the target variable. After separating the dataset, we split both variables into the training and Test sets. By giving the Test set 20% of the 8124 samples, we ended up having 6499 samples under the training set and 1625 samples in the Test set.

Since the dataset is categorical in nature, in order to use these, we must prepare both input data and the target. To do this, we made functions that would encode categorical features as an integer array. We did this to the training and test sets.

B4. Feature Selection

As the feature variables consists of 22 attributes, some of these are actually unnecessary for the analysis of the project. To figure out which attributes are essential, we performed feature selection, specifically Chi-Square. Using the training sets of both feature and target variables, we took the scores for the features and plotted it to see which are relevant.



From the plot above, the features that stood out the most are Features 3, 6, 7, 8, 10, 18, and 21. The names of these features are: “bruises”, “gill-spacing”, “gill-size”, “gill-color”, “stalk-root”, “ring-type”, and “habitat” specifically. Hence, we built a model using these Chi-Squared features.

Now that we know which features to include in the analysis, we again separated the original dataset to only include the seven attributes in the features variable. Using the new feature variable, we split it into the training and test sets. This still gave the same number of samples as the initial split.

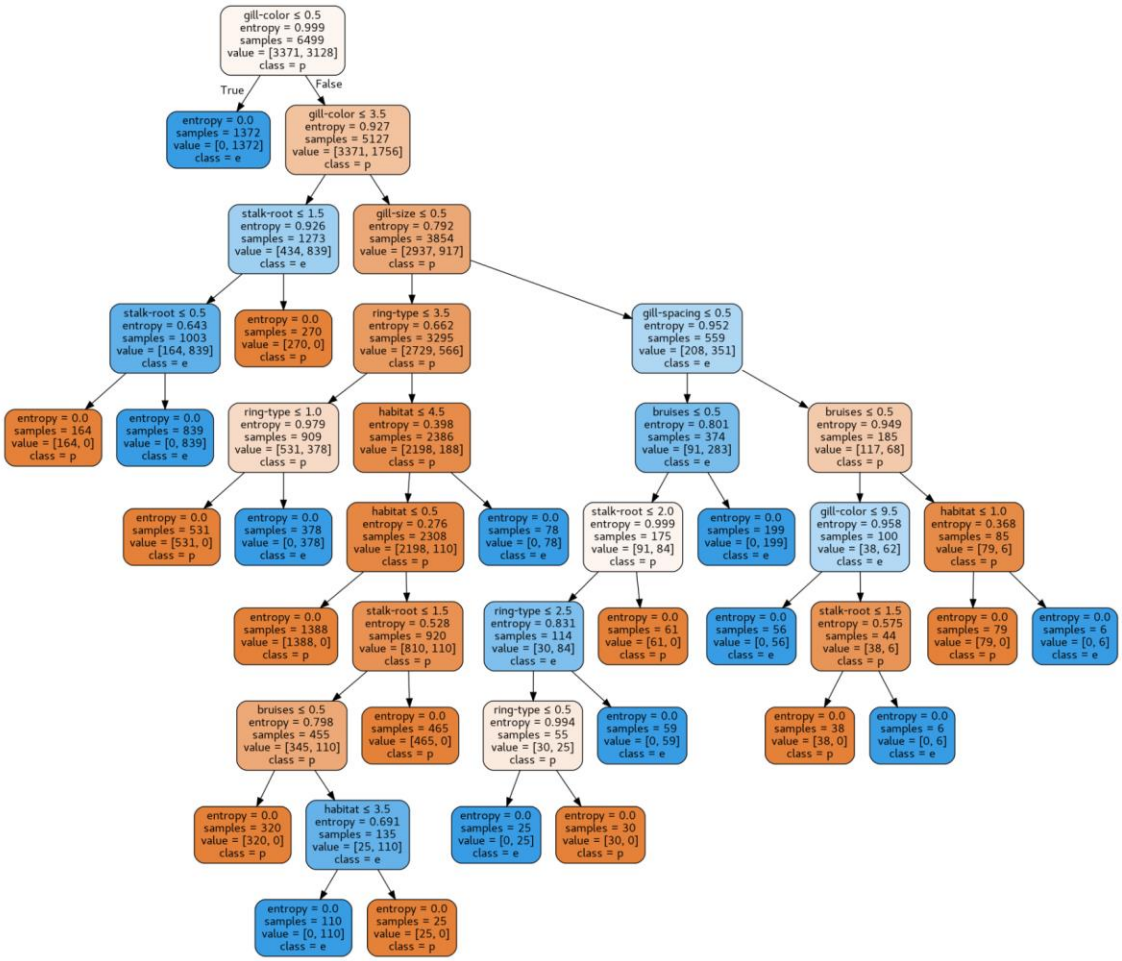
C. Exploratory Data Analysis

For this project, the aim is to use a Machine Learning model that would be able to accurately predict if a new mushroom is edible or not. Hence, we will specifically use a Decision Tree model. We want to create a training model that we can use to predict the class of the target variable by learning simple decision rules inferred from prior data.

To use this model, the DecisionTreeClassifier() function was imported from the sklearn library. Then the training sets were used to fit a model tree which is then used to predict the label of the data of the test set from the features variable.

After predicting the class of the features variable test set, the model is evaluated. To review how often the classifier is correct, the accuracy score of the predictions and the target variable test set was calculated. The prediction and accuracy score were done for both training and test sets.

Visualizing the tree, we got that there are nine branches in which the root node was the gill-color attribute.



Tuning the parameters of the tree can impact the model in terms of under-fitting and over-fitting. One of the ways to do this is to change the range of values for the maximum depth of the tree. To see the significance in the variation of values, the accuracy scores are then calculated for both the training and test sets. Below is a data frame of the accuracy scores for each maximum depth:

	Max Depth	Train	Test
0	1	0.729805	0.734154
1	2	0.792122	0.782769
2	3	0.855670	0.860923
3	4	0.888444	0.889846
4	5	0.963379	0.959385
5	6	0.977535	0.974154
6	7	0.979228	0.974769
7	8	0.996153	0.995692
8	9	1.000000	1.000000

It can be seen that as the number of branches increase, so does the accuracy of the model. The model performs well on the training set as well as the test set. Overfitting is not as issue in this case.

C1. Building a Predictive System

Now, this Decision Tree model can be of use to segregate a mushroom based on seven features. To do this, data from the original dataset was used and only included the values for the seven attributes. After, it the input data is then encoded and changed into a numpy array. The numpy array is then reshape so that the model can evaluate the data being fed to it. We then used the model to predict the class of the input data. As the class of the input data is already known, the predicted classification of the model can be immediately assessed.

D. Conclusion

Identifying whether a mushroom is edible or poisonous is a problem that many may not think of readily, especially if one does not mind it. However, for those who are mushroom enthusiasts or a professional who studies the usage of mushrooms in science, segregating the mushroom for its edibility is a needed examination. One of the ways to do this, is to use a Decision Tree model that would look at the attributes that are available and decide which are necessary for the assessment. From this model, we can create a predictive system that can be easily available to use and redo by the enthusiasts and professionals as a regulation for their own use.