Tyler Graham
CS5610
October 30, 2023

Project 3B

Project 3B utilized a dataset of 8124 mushrooms with 23 variables. These range from poisonous or edible to physical characteristics like cap shape and cap surface. Due to the large amount of features in the dataset, a decision tree was used to help create a model to predict whether the mushroom is poisonous or not.

This type of model could also be very useful in a professional environment to help predict fault detection. In embedded systems, there's numerous sensors being continuously monitored with different states and variables from their measurements. A decision tree could be trained on this and the model begins to detect the issue it could safely shutdown prior to becoming unsafe for the user.

On the initial decision tree of the mushroom data the following features were used: odor, stalk root, stalk color above ring, bruises, gill size, stalk surface below ring, cap surface, cap color, stalk color below ring and spore print color for a total of 14 features. The decision tree model does not use all 117 columns because it some of the datasets main contain similar information and are irrelevant to the model to make predictions.

The three largest importance features are odor_n, stalk_root_c and spore-print-color-r. These are the root of the decision tree and its immediate nodes. The vast majority of mushrooms can be split into two leafs of the tree accounting for >80% of the mushrooms in the data set. The boolean expressions of leaf 1: odor_n ->stalk-root_c ->stalk-surface-below-ring_y -> spore-print-color_u -> odor_a -> odor_l and for leaf 2: odor_n -> spore-print-color_r -> stalk-surface-below-ring_y-> cap-shape_c -> cap_-surface_g -> gill-size_n.

To put these in plain English for leaf 1: Start by smelling the mushroom. If it has a smell, take a look at the root of the stalk; if it looks like a club. Then, check the lower part of the stalk just below the ring if its yellow then check the color of the spore print if it is purple and finally, if it smells like almonds and anise, it might be potentially poisonous.

For Leaf 2: To identify an edible mushroom, first, inspect its odor; it should be odorless. Then, look at the spore print color; it should be red. Next, check the stalk surface below the ring, which should be yellow. Proceed to check the cap shape; it should be a club. The cap surface should be grooved. Finally, confirm the gill size, which should be narrow. If the mushroom exhibits all these traits, it has a good chance to be edible.

When increasing the ccp_alpha of the decision tree to .05 on the decision tree model, it reduces the tree to a single root with one node. However, the tradeoff for this reduction in nodes for the tree is an accuracy of .94%. That way to increase the simplicity to the user down to determine if it is poisonous by does it have no odor and a club stalk root it is poisonous otherwise it is edible.

To compare the usefulness of the decision tree, a logistic model was used on the same dataset to also predict if a mushroom is poisonous or edible. When using the logistic, the two largest coefficients were odor-n and spore-print-color-r. These were the root and one of the nodes of the decision tree. There is significant overlap in these models, however, the logistic gave a lot more weight to other odors than a decision tree. See the below table on the coefficients > abs(2.5).

Feature	Coefficient	Absolute Coefficient
odor_n	-4.488982	4.488982
spore-print-color_r	3.659015	3.659015
odor_f	3.105649	3.105649
odor_c	3.048699	3.048699
odor_a	-3.011739	3.011739
odor_l	-2.989874	2.989874
stalk-root_b	2.924122	2.924122
gill-size_b	-2.583478	2.583478
gill-size_n	2.539279	2.539279

Using the logistical models coefficient, the key elements to look for on edible mushroom or with a negative coefficient are. Having an odor of almond, anise and a gill size of board. When using the positive coefficient avoid the poisonous mushrooms that have a spore print color of red, odor of foul, odor of creosote, a stalk root of bulbous and a gill size of narrow.

Overall, the features considered by the two models do have significant overlap and characterize them in very similar ways. The decision tree and logistic regression can both provide good models for this type of dataset and machine learning problem. They both provided the user an accurate method to predict poisonous or edible mushrooms.