*Creates a categorical decision tree and analyzes the relationships between variables of interest*

**Assignment**

**2**

A2

ALY6040 Data Mining Applications

Assignment 2 – Decision Trees

**PREPERATION:**

By: John DiSessa

For: Professor Ellis

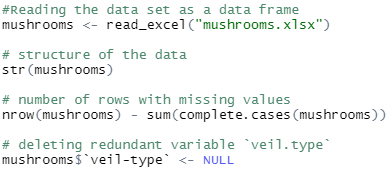
On: June 5th, 2022

Introduction

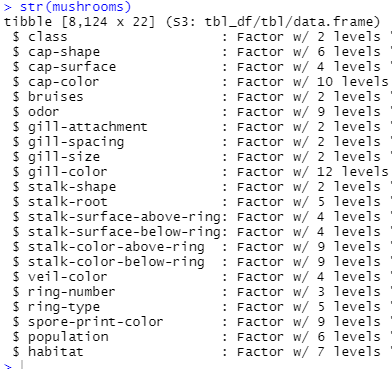
We were able to demonstrate our ability to identify, create, and analyze a categorical variable decision tree in R. Our dataset contained 8,124 rows of data for 23 variables regarding mushroom characteristics. It is important to identify which mushrooms are poisonous and which are edible for human consumption. Mushroom experts have the experience to easily identify edible and poisonous mushrooms, but novices struggle. The decision tree we created provides an accurate tool for people to identify mushrooms.

Dataset Exploration

Before conducting any analysis, we had to import our Mushrooms dataset into R. We then identified the structure of the data, the number of rows with missing values, and also noticed that every mushroom’s ‘veil type’ in our dataset contained ‘p’. Since this was redundant and did not provide any useful information into differentiating between edible and poisonous mushrooms, we removed it from the dataset.

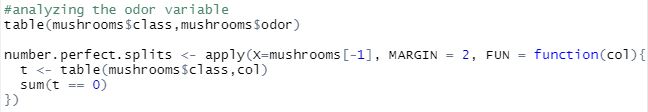


With the original structure of the dataset, each variable’s values were coded as individual characters. We decided to clean this up to have the dataset be coded as a better representation of the values than various characters. Even though this did not affect the calculations, it gave us a better idea of the ranges of each variable.

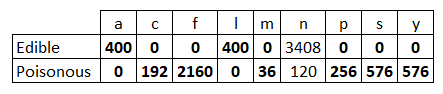


Decision Tree Setup

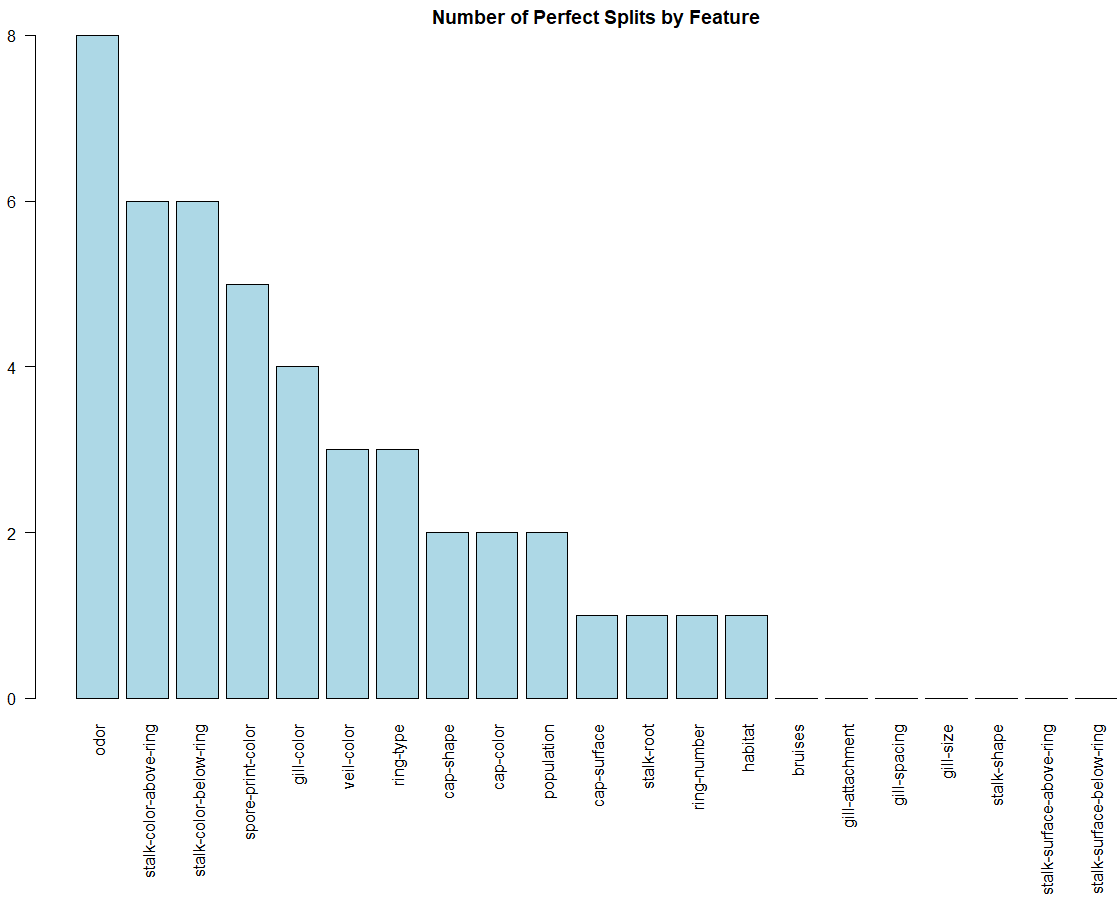
Since we identified our target variable as either edible or poisonous, we needed a root node in order to begin creating our decision tree. We used ‘odor’ as our root node because it gave a near perfect split between edible and poisonous mushrooms.



The table below shows how many edible and poisonous mushrooms there are for each of the 9 different odors. As you can see odors ‘a’ and ‘l’ automatically indicate that the mushroom is edible. An odor of ‘n’ indicates a 97% chance that the mushroom is edible. All other odors ‘c’, ‘f’, ‘m’, ‘p’, ‘s’, and ‘y’ show that the mushrooms are poisonous.



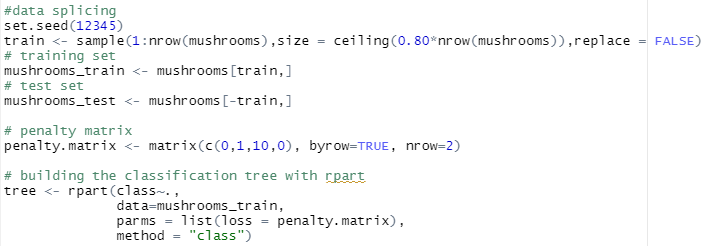
Even though we used odor as our root node, we created the following histogram to see which variables had perfect splits of edible and poisonous and which did not. This also let us confirm that our choice of odor was the correct choice for the root node since it had more perfect splits than any other mushroom feature.



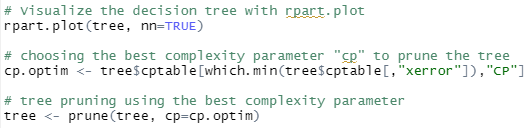
Even though many variables had a lot of perfect splits, that did not necessarily indicate that we could have used them as our decision nodes since many of them overlapped with each other. If variables are highly correlated with each other, using them as a decision node would not tell us new information to help us differentiate between edible and poisonous. For example, stalk-color-above-ring and stock-color-below-ring have many perfect splits, but have a correlation coefficient of .53. Additionally, many variables only have 2 or 3 different categories which means there are less possibilities for perfect splits. Gill-size, for example, only has 2 categories and neither of which indicate a perfect split between edible and poisonous. However, because it is uncorrelated with the other variables, it really helps us distinguish between our target variable since it tells us new information.

Decision Tree Creation

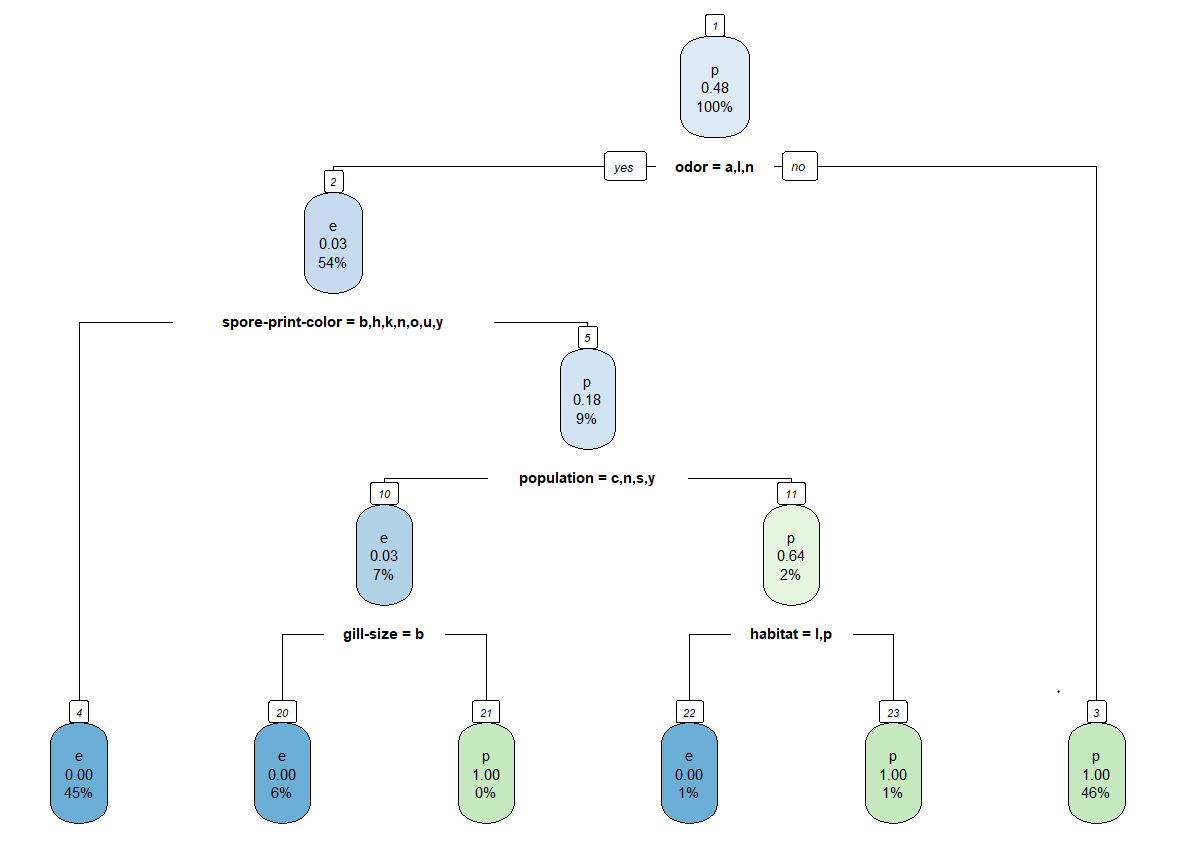
With our dataset being cleaned and analyzed, we started to create our decision tree model. In order to do so, we used the machine learning strategy of splitting our dataset into training and testing datasets. The training dataset was used to create our model and the testing dataset was used to evaluate the accuracy of our model. We randomly took 80% of our dataset and assigned it as our training data and the other 20% was assigned to our testing data. Another input to our model was a penalty for false positives. Our goal was to minimize Type 1 errors. There is infinitely more harm in falsely identifying a mushroom as edible than falsely identifying one as poisonous. Our penalty matrix for false positives was 10 times larger than the penalty for false negatives.



From the training data, we created our decision tree using the rpart.plot library. We pruned the tree by using the optimal complexity parameter as calculated below.

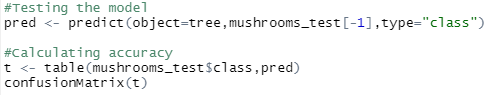


The following decision tree can be used to guide novices in the differentiation between edible and poisonous mushrooms. If the odor is ‘c’, ‘f’, ‘m’, ‘p’, ‘s’, or ‘y’, we can assume the mushroom to be poisonous and thus no need for more deliberation. If the odor is ‘a’, ‘l’, or ‘n’, we then look at the spore-print-color. If the color is ‘b’, ‘h’, ‘k’, ‘n’, ‘o’, ‘u’, or ‘y’, we can be confident the mushroom is edible. If the spore-print-color is ‘r’ or ‘w’, we enter a subtree that is more complicated than using just the previous variables. Our model predicts that only 9% of all mushrooms would enter this subtree since their toxicity could not be determined from odor or spore-print-color. From this subgroup, if the population is ‘c’, ‘n’, ‘s’, or ‘y’, then we look at the gill size. If the gill size is ‘b’, the mushroom is edible and poisonous if the gill size is not ‘b’. If the population is ‘a’ or ‘v’ we them look at the habitat. If the habitat is ‘l’ or ‘p’, the mushroom is edible and poisonous if not. From this decision tree, we are able to accurately identify all mushrooms.

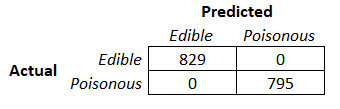


Model Accuracy

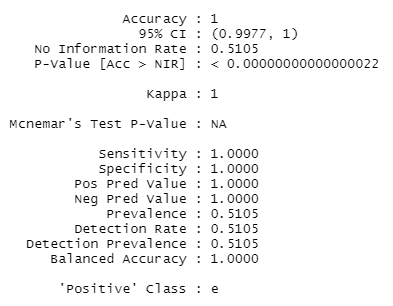
Now that our model created our decision tree, we need to evaluate just how accurate it actually is at predicting the mushroom class. We compare our model to the testing data to see how many mushrooms were edible that our model predicted would be edible and how many mushrooms were poisonous that our model predicted would be poisonous.



As you can see from the following confusion matrix, our model correctly predicted the class for every mushroom in our test dataset. This results in a 100% accuracy score.



From the advanced accuracy calculations below, we built an incredibly accurate model. We are 95% confident that the true accuracy score ranges between 99.77% and 100%. Our p-value is also almost 0 so there is virtually a 0% chance that our accuracy results could have been determined via chance. The sensitivity score of 1 means that we correctly identified all 829 edible mushrooms. The specificity score of 1 means that we correctly identified all 795 poisonous mushrooms. The prevalence score of 51.05% refers to the true rate of all edible mushrooms compared to the total amount of mushrooms in the dataset. Our detected prevalence of the same score means we got a balanced accuracy of 100%.



Summary

We successfully created a decision tree that accurately distinguished between edible and poisonous mushrooms. After cleaning the data and identifying the root node, we used the machine learning strategy of splitting the dataset into training and testing data. From this we were able to create the categorical variable decision tree and verify the accuracy of our model.

Citations

“Confusion Matrix in Machine Learning.” *GeeksforGeeks*, 18 Nov. 2021, https://www.geeksforgeeks.org/confusion-matrix-machine-learning/.

“Confusion Matrix.” *Confusion Matrix - an Overview | ScienceDirect Topics*, https://www.sciencedirect.com/topics/engineering/confusion-matrix#:~:text=A%20confusion%20matrix%20is%20a,malignant%20tissue%20is%20considered%20cancerous.

jhuno137. “Classification Tree Using RPART (100% Accuracy).” *Kaggle*, Kaggle, 11 Jan. 2017, https://www.kaggle.com/code/jhuno137/classification-tree-using-rpart-100-accuracy/report.

Kiely, Philip. “A Pirate's Guide to Accuracy, Precision, Recall, and Other Scores.” *FloydHub Blog*, FloydHub Blog, 25 May 2020, https://blog.floydhub.com/a-pirates-guide-to-accuracy-precision-recall-and-other-scores/.

Kuhn, Max. “The Caret Package.” *17 Measuring Performance*, 27 Mar. 2019, https://topepo.github.io/caret/measuring-performance.html.

Martin, Wendy. “Cohen's Kappa in R: Best Reference.” *Datanovia*, 7 Nov. 2019, https://www.datanovia.com/en/lessons/cohens-kappa-in-r-for-two-categorical-variables/#:~:text=kappa%20can%20range%20form%20%2D1,rater%20agreement%20exceeds%20chance%20agreement.

Narkhede, Sarang. “Understanding Confusion Matrix.” *Medium*, Towards Data Science, 15 June 2021, https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62.

Pearson, Ron. *The GOODMANKRUSKAL Package: Measuring Association between Categorical Variables*, 18 Mar. 2020, https://cran.r-project.org/web/packages/GoodmanKruskal/vignettes/GoodmanKruskal.html.

Swift, Amelia, et al. “What Are Sensitivity and Specificity?” *Evidence-Based Nursing*, Royal College of Nursing, 1 Jan. 2020, https://ebn.bmj.com/content/23/1/2.

“What Is Training Data in Machine Learning?” *MonkeyLearn Blog*, 2 Nov. 2020, https://monkeylearn.com/blog/training-data/.

