Mushroom Classification – Safe to Eat or Deadly to Poison?

Yun Xiao, Ruofei Li, Zilong Chen
IST 707 Final Project Report — Group 4

Abstract

We consider the problem of identifying poisonous and edible mushrooms by its features, e.g. gill color and ring type, using the dataset from UCI Machine Learning repository. After analyzing data distribution, we find some symbols may be the unique attribute indicating mushroom quality. To solve the problem, we use four data analysis methods: Association Rules, Decision Tree, Random Forest, Gradient Boosting Model and Bootstrap Aggregating. Then, after comparing the performance of models, we find out the best model and some interesting features for mushroom identification.

1 Introduction

Mushroom hunting (otherwise known as "shrooming") is enjoying new peaks in popularity these years. Mushroom foraging is the activity of gathering mushrooms in the wild, typically for eating. In the United States mushroom picking is popular in the Appalachian area and on the west coast in northern California, Oregon and Washington, and in many other regions. (dict.eudic.net).

Many mushroom species are favored for eating by mushroom hunters. However, a Czech adage warns that "every mushroom is edible, but some only once." (dict.eudic.net). Some mushrooms are deadly or extremely hazardous when consumed. Some that are not deadly can nevertheless cause permanent organ damage. A strongly advise is that only positively identified mushrooms should be eaten. (dict.eudic.net).

Much more care, education, and experience are typically required to make a positive identification of many species. Many field guides on mushrooms are available, but the ability to identify and prepare edible mushrooms is often passed down through generations, especially in the Slavic countries. Many mushroom guidebooks call attention to similarities between species, especially significant if an edible species is similar to, or commonly confused with, one that is potentially harmful. (dict.eudic.net).

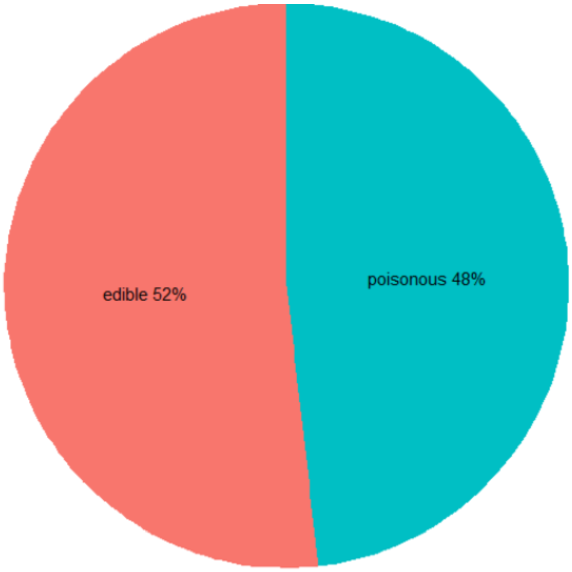
Since mushroom hunting is a such skillful activity, we are considering using data analysis to identify mushroom instead of using field guides or just passing down through generation. We try to figure out how to determine whether a mushroom is edible or poisonous. In this case, we put forward two main questions:

- What features are most indicative of whether a mushroom is poisonous or not?
- How to tell the differences between poisonous and edible mushrooms?

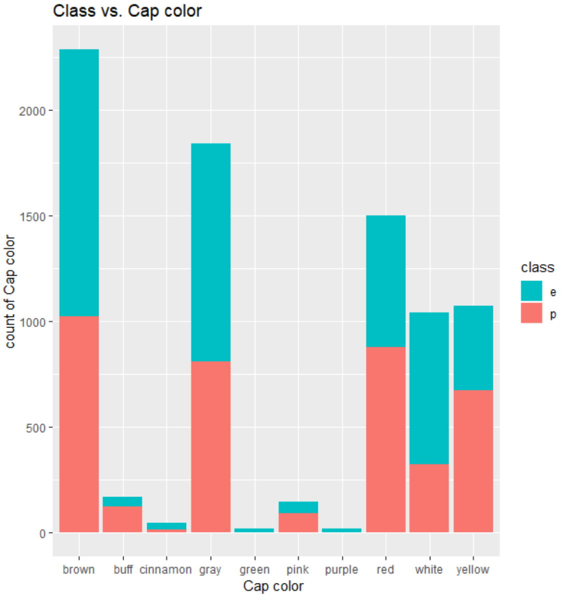
We are using the dataset from UCI machine learning database. The dataset contains 8,124 observations and 22 features of mushrooms such as cap shape, odor, etc. Other than the 22 features, each observation contains a column called “class” indicating whether the mushroom is edible (“e”) or poisonous (“p”). “spore print color”, "population", "habitat". Each species is identified as definitely edible, poisonous, or of unknown edibility and not recommended. We learn which features spell certain death and which are most palatable in this dataset of mushroom characteristics.

Here are the names of all columns: "class", "cap shape", "cap surface", "cap color", "bruises", "odor", "gill attachment", "gill spacing", "gill size", "gill color", "stalk shape", "stalk root", "stalk surface above ring", "stalk surface below ring", "stalk color above ring", "stalk color below ring", "veil type", "veil color", "ring number", "ring type".

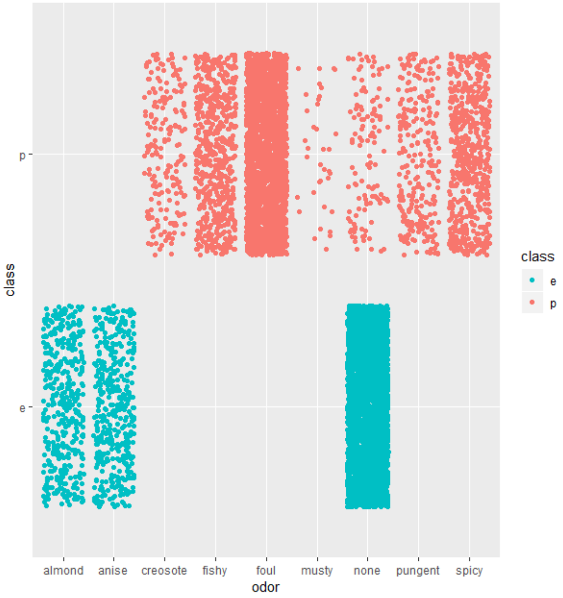
2 Data Exploration and Visualization

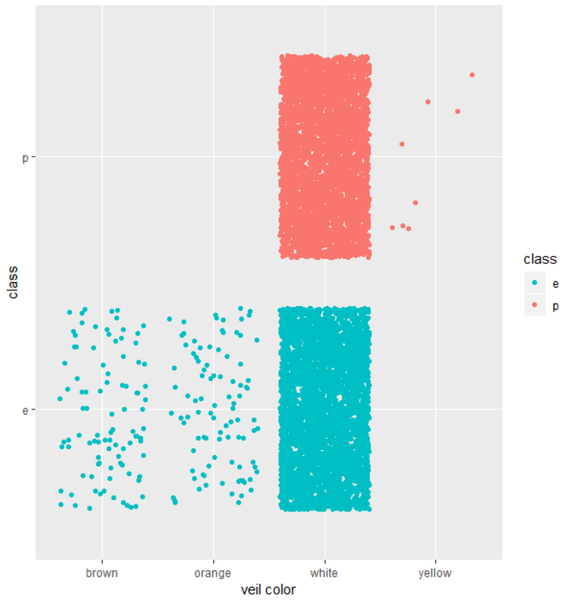


The above pie chart of dataset shows that 52% of mushroom class is edible and 48% of it is poisonous



The above analysis of cap color for mushrooms by its features indicates that it is always edible when the mushroom is green or purple. For most of colors, however, it is hard to identify because the proportions for edible and poisonous are both about 50 percent.





Next, we draw the above jitter charts for ‘odor’ and ‘veil color’ to analyze the distribution of edible and poisonous. We can see that when odor is ‘almond’ or ‘anlse’ the mushroom is all edible and almost all edible for ‘none’ class. For the other class like ‘foul’ and ‘fishy’, the mushroom is always poisonous. When we analyze the distribution of ‘veil color’, the brown and orange are the symbols for edible mushrooms and yellow is dangerous for people.

3 Data Preprocessing

The dataset is randomly divided into a training dataset which contains 70% of the data and a test dataset which contains the other 30% of the data. These two datasets are used in Association Rules but are further processed for other machine learning methods.

For other machine learning models, the attributes “veil type” and “spore print color” are removed because there are only one veil type and only one spore print color. Then the train dataset are separated to “train_x” which contains all the independent variables and “train_y” which contains the dependent variable “class”.

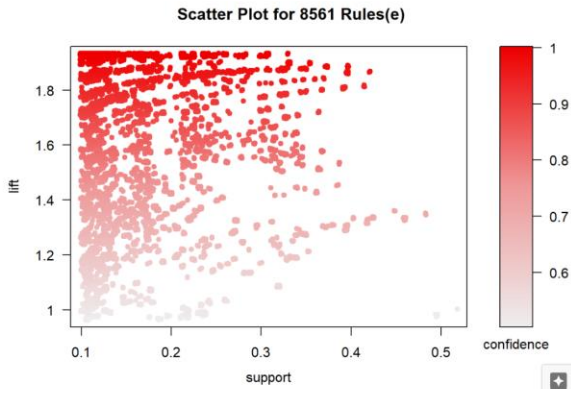
4 Machine Learning Models

4.1 Association Rules

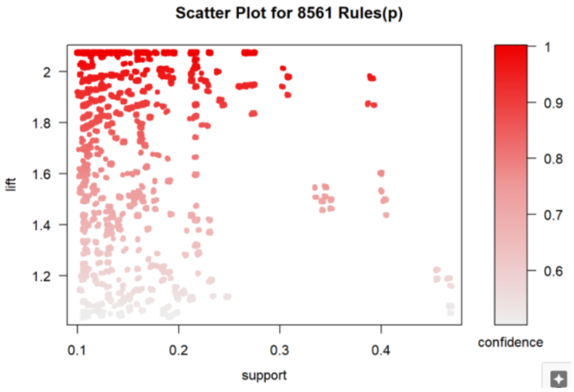
Association rules are basically if statements that show relationships between large data itemsets. This method can be used as unsupervised or supervised learning. Here, we use it as supervised learning to discover the patterns between independent variables and the dependent variable “class” by forced the rhs to be either class “e” as edible or class “p” as poisonous.

We first set support as 0.1, confidence level as 0.5, minlen as 3, and maxlen as 5, and inspect the rules by ordering them from largest “lift” to smallest. The top rules show that broad gill size and brown gill color, none odor and equal stalk root, and none bruises and equal stalk root will lead to edible mushrooms, and their lifts are 1.93 which indicate that those rules are positively related. The top rules also show that large ring type, buff gill color, and foul odor will lead to poisonous mushrooms, and their lifts are 2.075 which also indicate positive relationships.

Then we tune the rules by increasing support to 0.4. The top rules show that none odor, partial veil type, and broad gill size will lead to edible mushrooms (lifts are 1.86) and that none bruises, free gill attachment, and one ring will lead to poisonous mushrooms (lifts are 1.60).



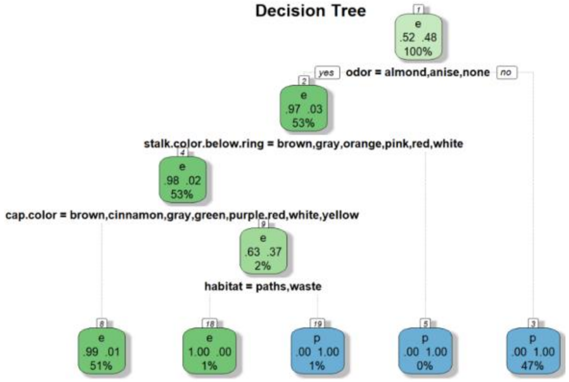
We visualize the rules when rhs is “e”. According to the above graph, the largest support is around 0.5. Most rules have support less than 0.3. The high confidence rules located at where the lifts are high. The highest lift is below 1.9.



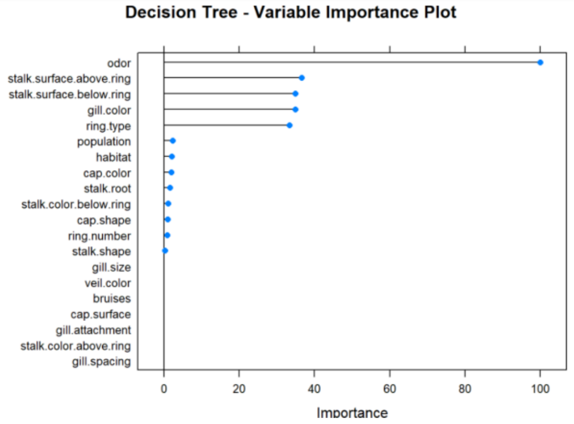
For the rules that have rhs as “p”, according to the above graph, no rule has support than 0.5. Most rules have support than 0.5. Most rules have support less than 0.2. Similar to “e”, the high confidence rules located at where the lifts are high. The highest lift is more than 2.

4.2 Decision Tree

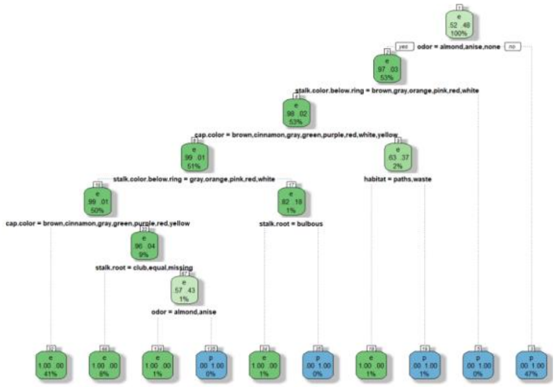
The first decision tree model is trained at default settings. The final rule is selected when cp=0.002479942. After applying this model to the test data, the accuracy is 0.9943, sensitivity is 1.00, and the specificity is 0.9881.



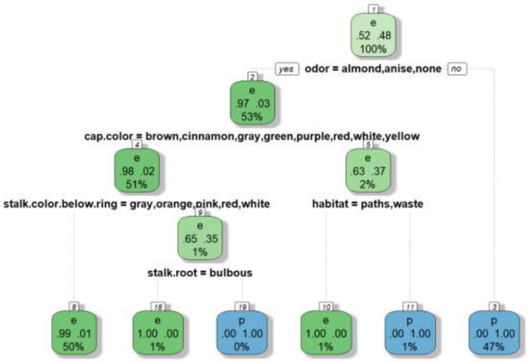
According to the above plot, if the odor is not almond, anise, and none, the mushroom is poisonous. If the odor is almond, anise, or none, then let’s look at the stalk color below ring. If the stalk color below ring is brown, gray, orange, pink, red, or white, then let’s look at the cap color; if not, the mushroom is poisonous.



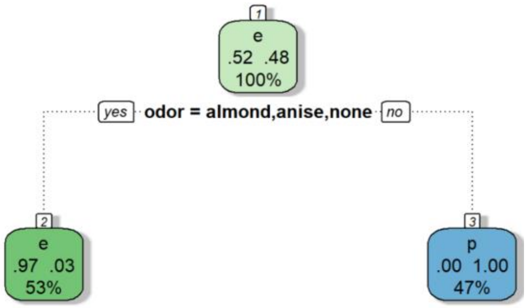
According to the above variable importance plot, the most determinable attributes are odor, stalk surface above and below the ring, gill color, and ring type.


e
.52 .48
100%
yes
odor = almond,anise,none
no
stalk.color.below.ring = brown,gray,orange,pink,red,white
cap.color = brown,cinnamon,gray,green,purple,red,white,yellow
stalk.color.below.ring = gray,orange,pink,red,white
habitat = paths,waste
cap.color = brown,cinnamon,gray,green,purple,red,yellow
stalk.root = bulbous
stalk.root = club,equal,missing
odor = almond,anise

Then we tune the model by setting tuneGrid as "(cp=seq(0,0.01,0.001)". The final model has cp=0. According to the above plot, it's very likely that this model has overfitting issues because it has many nodes.



So we preprune the tree by setting tuneLength as 8, minsplit as 50, minbucket as 20, and maxdepth as 5. Based on the above plot, the determinable attributes are somewhat different from the first default model.

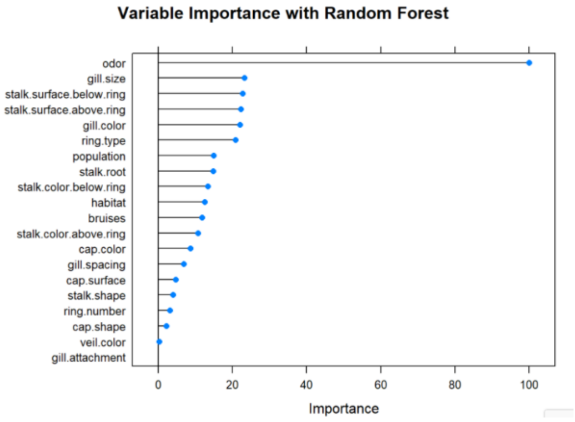


We also try postpruning the tree, and only one attribute is left based on the above plot.

4.3 Random Forest

"Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees." (Random Forest, Wikipedia). In this project, we first run random forest model with default settling.

The final model has mtry=2, and the model accuracy is 0.99. Then we also generate the importance of each features.



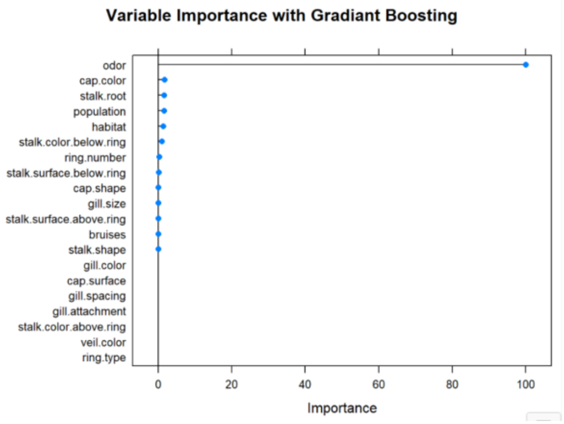
According to the figure above, we can find that 'odor' is the most important variable for classifying mushroom and its importance is 1. Next, we predict the model and make confusion matrix to see the accuracy of the model.

The accuracy of this model is 1, which is much high. Then we also make ROC curve to see the relationship between model specificity and sensitivity. After calculating the AUC of model, we can find the area under the curve is 1.

4.4 GBM

"Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function." (Gradient boosting, Wikipedia). In this project, we run GBM model with default setting.

We can see that the final values used for the model is n.trees=150, interaction depth=3, shrinkage=0.1, and n minobsinnode = 10. Then we use 'varImp' function to see the importance of each variable.

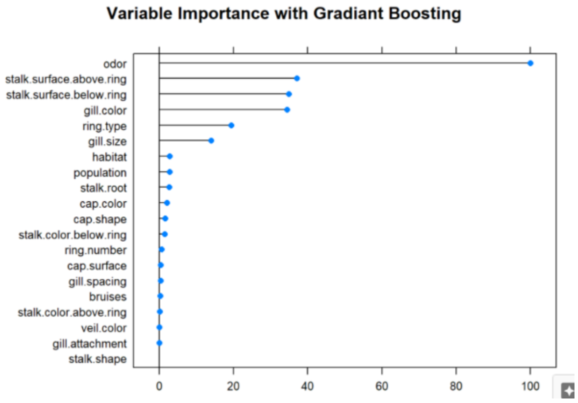


According to the figure above, 'odor' is the most important variable for identifying mushroom. Next, we check the accuracy of the model.

We can find the accuracy is 1, which means our model is perfectly predicted. All the data are correctly predicted by our model.

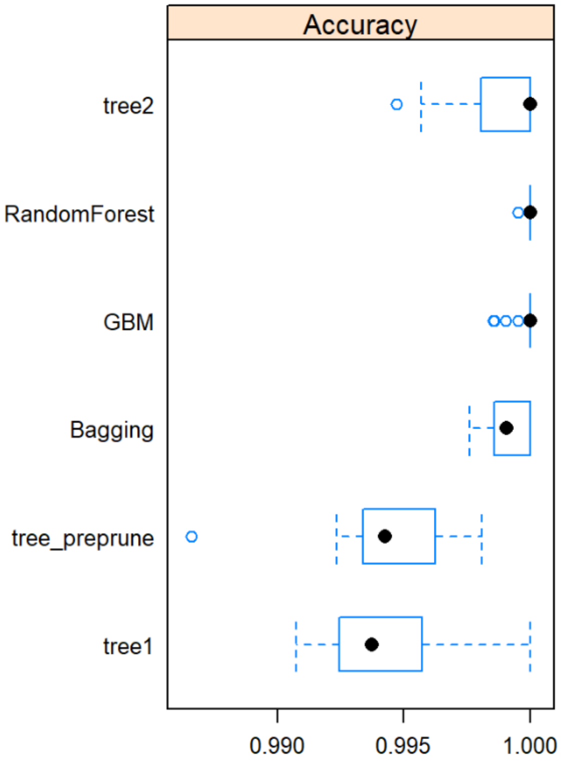
4.5 Bagging

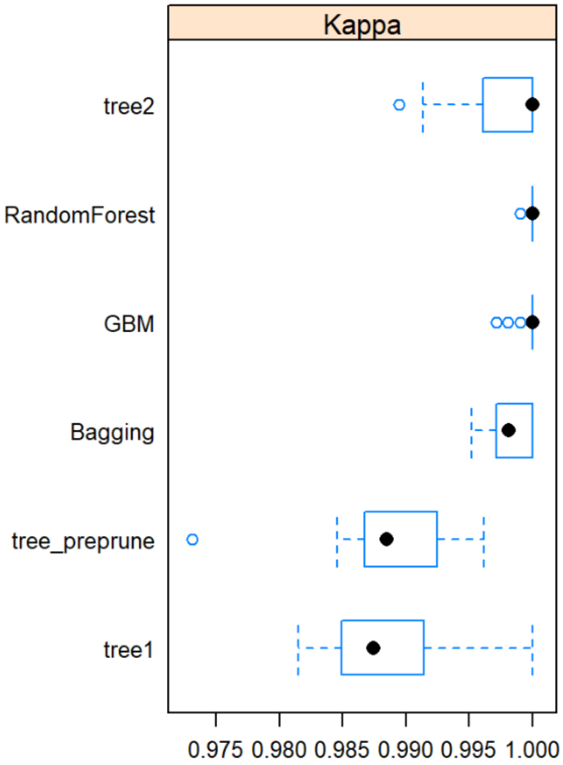
Bootstrap aggregating, also called bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting. In this case, we use bagging with default setting here and generate the importance of variables.



The most important variable is 'oder'. This result is also coincident with our conclusion above. Then we predict this model and we find the accuracy of this model is 1. We find all data are correctly predicted by our model.

5 Model Comparison and Conclusion





According to the above plots, the random forest and GBM models achieve the highest accuracies and vary the least.

All the variable importance plots show that "odor" is the most important variable the determine a mushroom is edible or poisonous. Based on the decision tree models, almond, anise and none odor lead to edible mushrooms.

Other features such as stalk surface above and below the ring, gill color, ring type, cap color, and so on are also indicative.

R Shiny

Association rules ("p" as rhs): https://zilong.shinyapps.io/IST707ARPoison/

Association rules ("e" as rhs): https://zilong.shinyapps.io/IST707Project/

Graphs: https://yxiao19.shinyapps.io/data2/

Decision Tree: https://yxiao19.shinyapps.io/data/

We also did R Shiny for other models but were not able to publish them. We showed them to the Professor during poster presentation.

References

Shrooming. *dict.eudic.net*. Retrieved from http://dict.eudic.net/dicts/en/shrooming

Random Forest. *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Random_forest

Gradient boosting. *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Gradient_boosting