Mushroom Classification

Safe to Eat or Deadly to Poison?

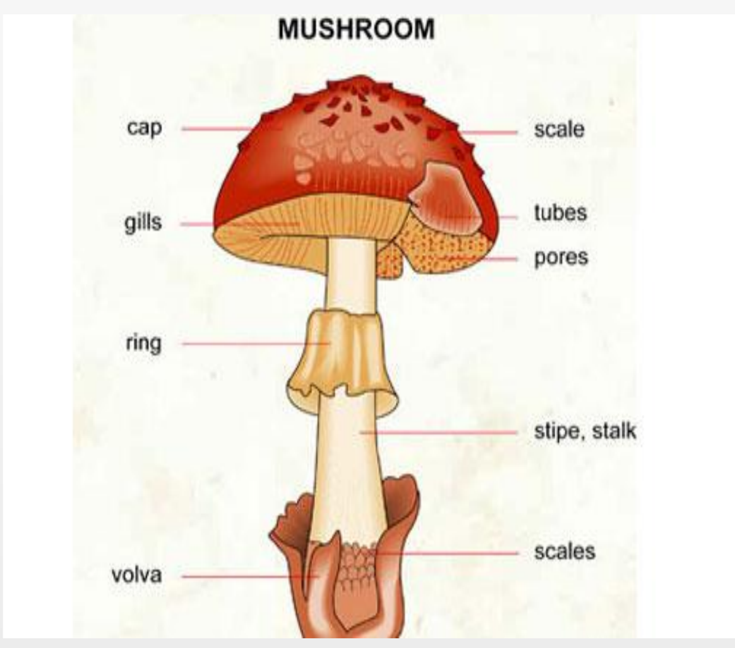
IST707 - M003 - Final Project - Group 4

Yun Xiao, Zilong Chen, Ruofei Li

THE iSCHOOL Syracuse University

Introduction

Learning repository nearly 30 years ago, mushroom hunting (otherwise known as "shrooming") is enjoying new peaks in popularity. In this case, we are interested in the mushroom classification.



In this project, 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?

Data Description

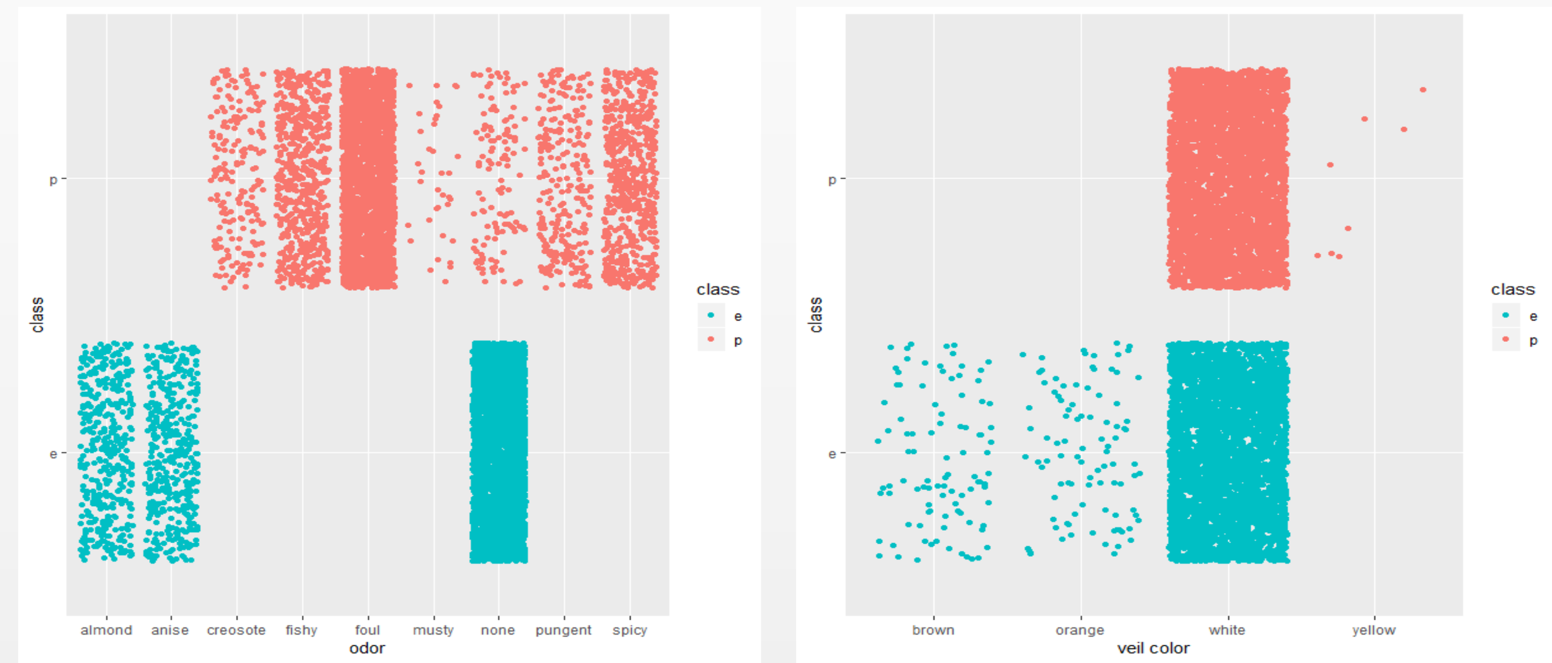
The dataset was originally from UCI Machine Learning database and is now available on Kaggle.
The dataset contains 8,124 observations and 22 features of mushrooms such as cap shape, odor, etc. Other than the 22 features, each observation contains a column called "class" indicating whether the mushroom is edible ("e") or poisonous ("p"). "spore.print.color", "population", "habitat".

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",

Data Visualization

We can see that 52% of our 'class' is edible and 48% of it is poisonous. Additionally, we can view our 'cap color' group by different features. We can see that, when the mushroom is green or purple, it is always edible. Next, we also draw jitter chart for 'odor' and 'veil color'.

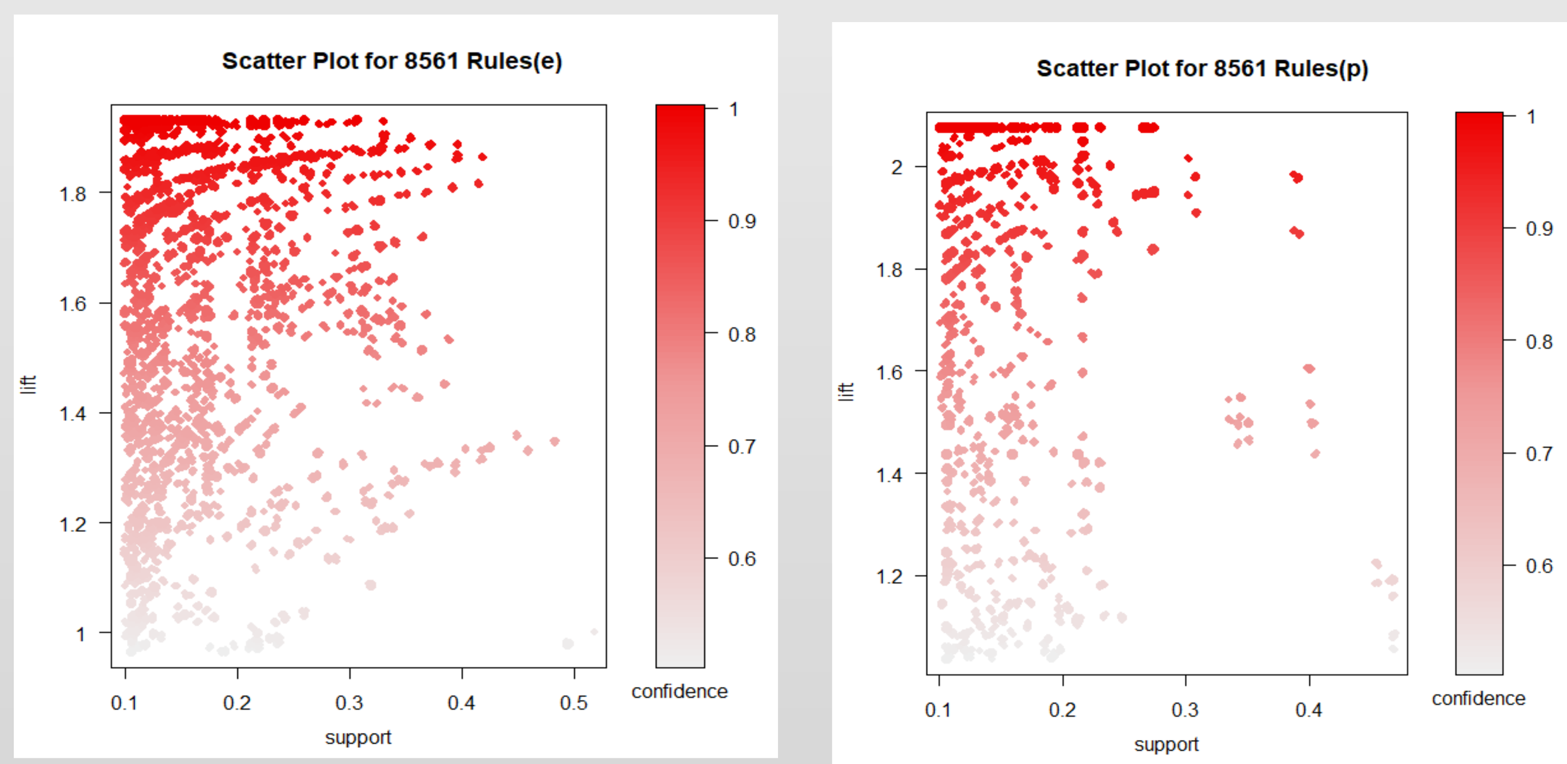
Data Visualization



We can see that when odor is 'almond', 'anlse', or 'none', the mushroom is all edible. However, when order is 'foul', the mushroom is all poisonous. At the same time, when the veil color is brown or orange, the mushroom is edible.

Association Rules

Edible Poisonous



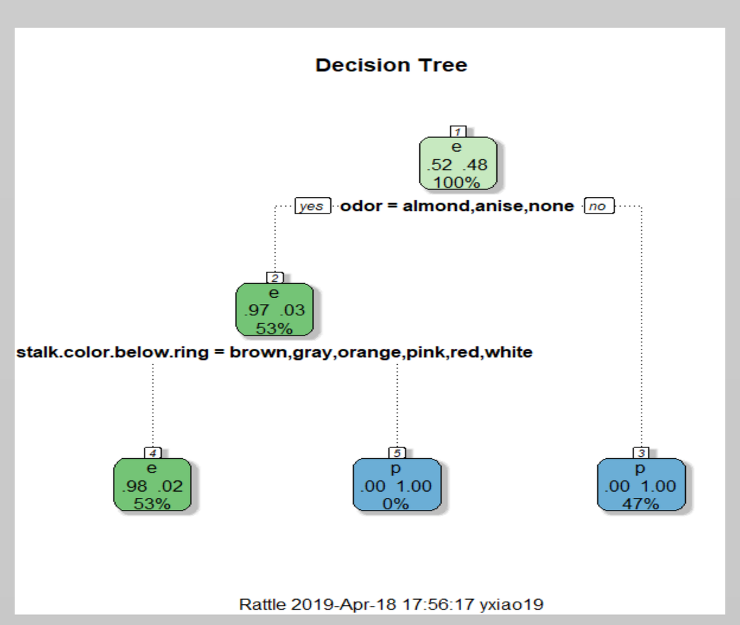
Edible

	lhs		rhs	support	confidence	lift	count
[1]	{gill.size=broad,gill.color=brown}	=>	{class=e}	0.1083210	1	1.930608	880
[2]	{odor=none,stalk.root=equal}	=>	{class=e}	0.1063516	1	1.930608	864
[3]	{bruises=no,stalk.root=equal}	=>	{class=e}	0.1063516	1	1.930608	864
[4]	{gill.spacing=crowded,habitat=grasses}	=>	{class=e}	0.1299852	1	1.930608	1056
[5]	{gill.spacing=crowded,stalk.shape=tapering}	=>	{class=e}	0.1063516	1	1.930608	864
[6]	{gill.spacing=crowded,gill.size=broad}	=>	{class=e}	0.1299852	1	1.930608	1056
[7]	{bruises=bruises,population=solitary}	=>	{class=e}	0.1201379	1	1.930608	976
[8]	{odor=none,population=solitary}	=>	{class=e}	0.1191531	1	1.930608	968

Poisonous

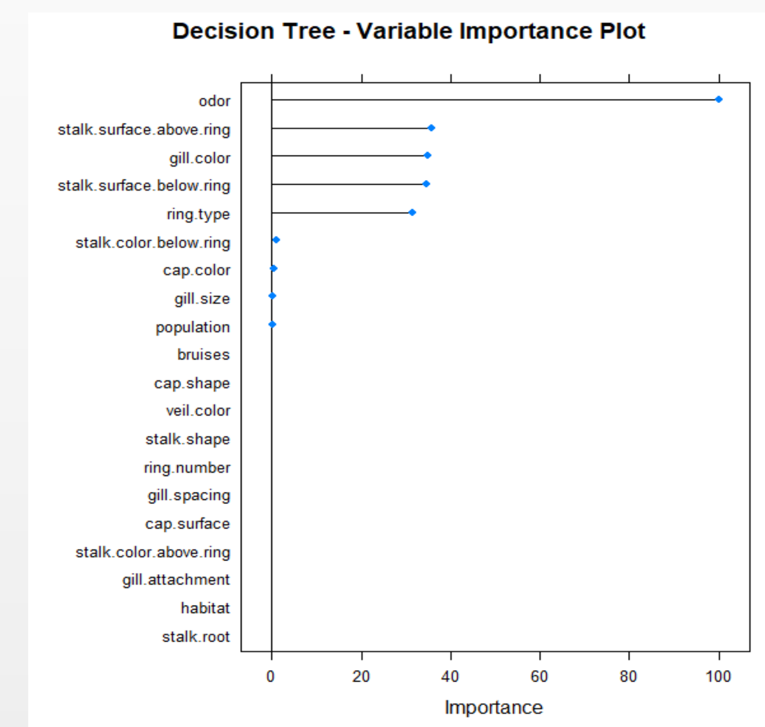
	lhs		rhs	support	confidence	lift	count
[1]	{ring.type=large}	=>	{class=p}	0.1595273	1	2.074566	1296
[2]	{gill.color=buff}	=>	{class=p}	0.2127031	1	2.074566	1728
[3]	{odor=foul}	=>	{class=p}	0.2658789	1	2.074566	2160
[4]	{odor=foul,ring.type=large}	=>	{class=p}	0.1595273	1	2.074566	1296
[5]	{stalk.surface.below.ring=silky,ring.type=large}	=>	{class=p}	0.1595273	1	2.074566	1296
[6]	{stalk.surface.above.ring=silky,ring.type=large}	=>	{class=p}	0.1595273	1	2.074566	1296
[7]	{stalk.shape=enlarging,ring.type=large}	=>	{class=p}	0.1595273	1	2.074566	1296
[8]	{stalk.root=bulbous,ring.type=large}	=>	{class=p}	0.1595273	1	2.074566	1296

Decision Tree

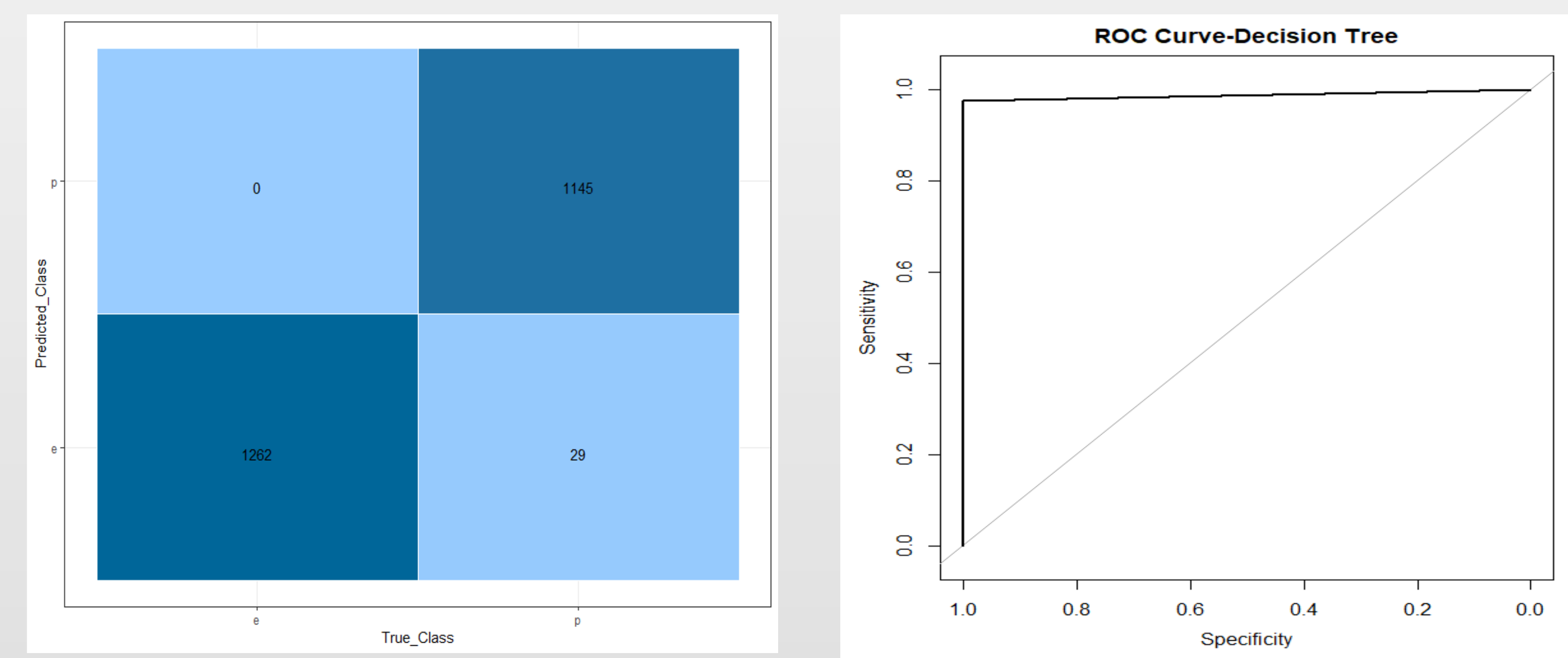


First, we run Decision Tree with default setting. 52% of the mushroom is edible, and 48% is poisonous. 53% has almond, anise, or none odor, and among it, 97% is edible. The other 47% has other odor, and all of it is poisonous.

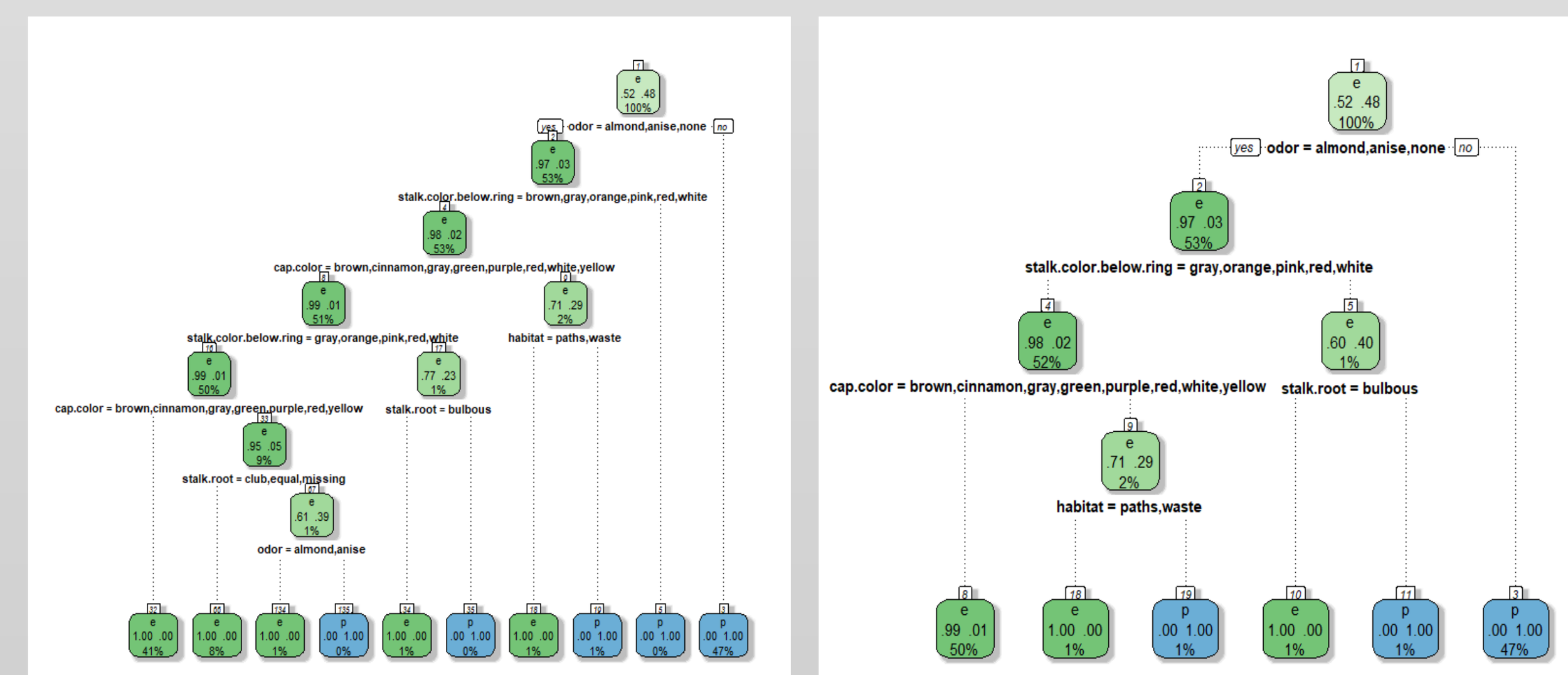
Decision Tree



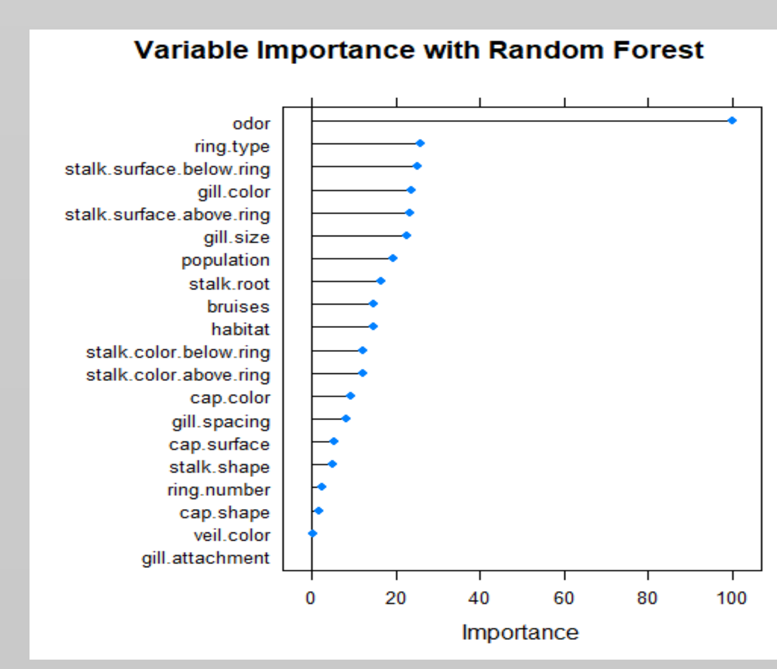
We can find that 'order' is the most important variable for mushroom classification. Additionally, 'stalk.surface. above. ring', 'gill color', 'stalk.surface. below. ring' and 'ring type' are also important variables for classifying mushroom. Next, we tune our decision tree model.

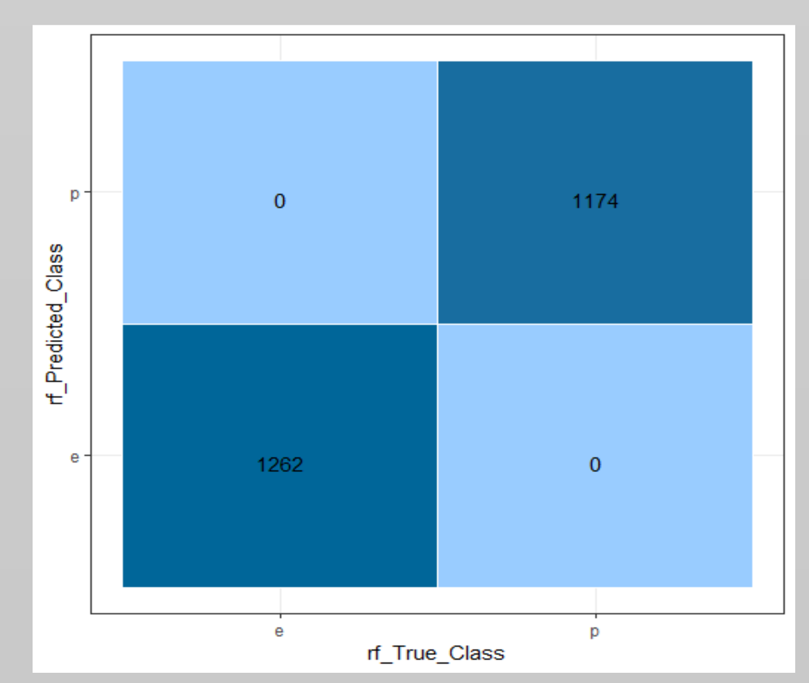


According to the ROC Curve-Decision Tree, our AUC under the curve is 0.9954. That is to say, our model accuracy is high. Then, we tune the decision tree model. The following are the pre pruning decision tree and post pruning decision tree model.



Random Forest



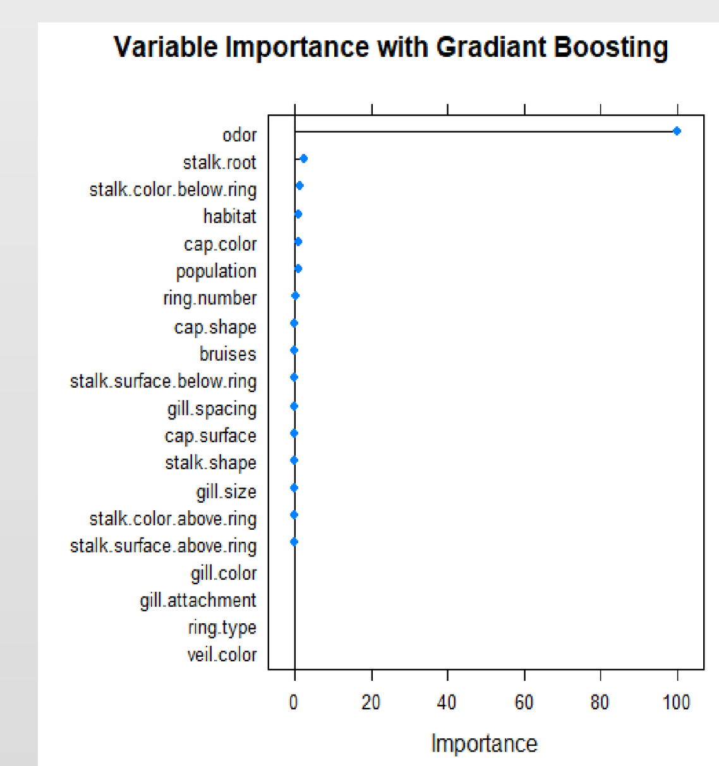


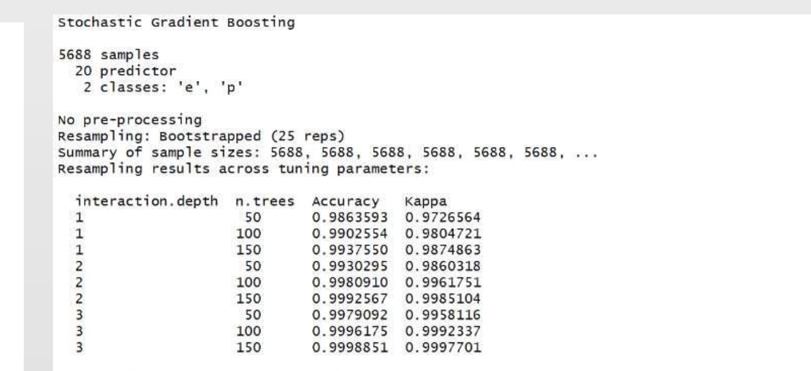
Here, we can see odor is the most important variable to classify mushroom. Additionally, our model is highly accurate.

Random Forest

```
Random Forest

5688 samples
  20 predictor
   2 classes: 'e', 'p'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 5688, 5688, 5688, 5688, 5688, 5688, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.9997711  0.9995414
  11    0.9996565  0.9993120
  20    0.9995606  0.9991201

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 2.
```

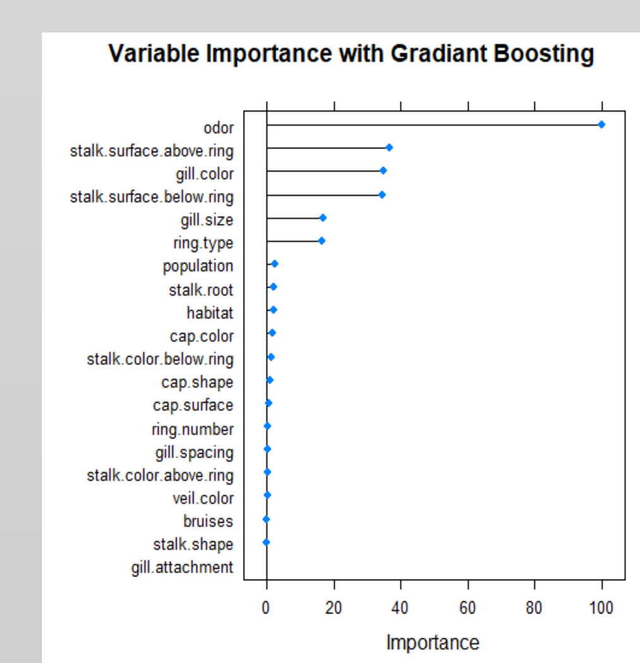
GBM





According to the Gradiant Boosting Model, we can find that odor is the most important variable to classify mushroom.

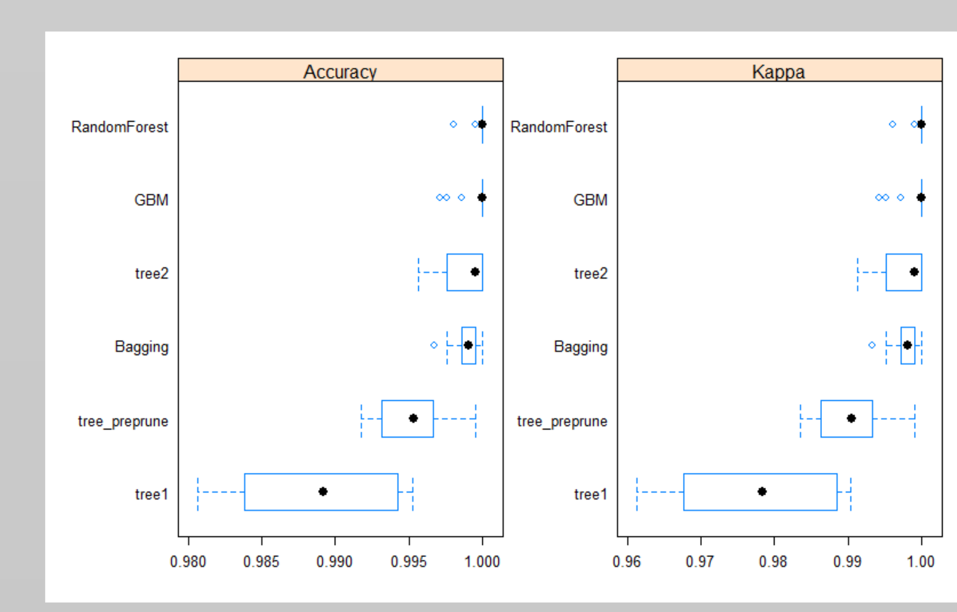
Bagging




```
Bagged CART

5688 samples
  20 predictor
   2 classes: 'e', 'p'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 5688, 5688, 5688, 5688, 5688, 5688, ...
Resampling results:

  Accuracy   Kappa
  0.9993857  0.9987687
```

According to the Bagging Model, we can find that odor is the most important variable to classify mushroom.

Conclusion



We can find that the Random Forest is the most accuracy model, which accuracy is nearly equal to 1. Moreover, odor is the most important variable for classifying mushroom.