ALY 6040  
Module 2: Mushroom Data

Professor: Justin Grosz

Chongfan Bao CPS in Informatics

Northeastern University

06/06/2021

**Abstract**

For this project, I will explore the mushroom dataset in order to learn how machine learn is contributed to make decision. I will develop two models: Decision Trees and Random Forest to predict and analyze which mushroom is edible or poisonous based on other variable characteristics. In this process, I will divide the dataset into two subsets: training dataset and test dataset. This is in order to test the accuracy of two model and help me to improve my model. Compared with these two models, they both are more than 99 percentage accuracy which indicates our models fits and predict the dataset. However, the accuracy of the decision trees model is better than the random forest model.

*Keywords:* machine learning, decision trees, random forest

**Introduction**

Mushroom is a gourmet dish and most people will forage for mushroom. However, some of the mushrooms are deathful if people eat by accident or get poisoning.Nobady can know every single of mushroom if it is safe to eat. Hence, it is necessary to identity the mushrooms if it is edible. To leverage data analysis to build a predict model is a good way. Before implement data analysis, I will convert file from “.xlsx” to “.csv”. This is because the R has a strict limitation for the attribute of data and variable name and can help me avoid unnecessary troubles. I will construct decision trees model and random forest model to estimate if this kind of mushroom is poisonous. In the end, I will compare with the two models to find the “best” predictive model.

# Data overview

This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms (UCI UCI Machine Learning, 2016). There is total 8124 objects with 22 variables. From the figure 1, I can notice the number of classifications of each variable. Meanwhile, I also find that the variable named “veil.type” only have one type. After I check the original source, I find that the type of variable should include two classifications: partial as the p and universal as the u in dataset. The other finding is that there is not any missing data. Thus, I don’t need to deal with the missing values.

# Data preprocessing and modeling

I find this data is relatively clean except the one single variable: veil.type. I will delete it. This is because it is meaningless for our data analysis. After data clean, I am ready to build the two models. First of all, I will look up the distribution of the odor variable under the circumstance of the class variable in order to evaluate which characteristic of variable is more identifiable. The more zeros that a variable appears in the case of the “class” variables, it means that the characteristics of this variable are used to determine whether mushrooms are poisonous. Given this, I will calculate the amount of missing type of each variable under the circumstance of the “class” variable. Then, to order the amount by decreasing order plots the bar chat and knows the importance of each variable based on identifying if mushrooms is poisonous or not.

After this, I divided the data into two subsets for the decision trees model and the random forest model individually. One sub dataset is training dataset to train the model. The other is testing dataset to test the accuracy of model. At the same time, I wanted to know the split of edible to poisonous mushrooms in the data set and compare it to the training and test data. So I calculate each dataset percentage of the class variable. On the one hand, I, in the process, build a loss matrix for decision tree model as a classification splitting parameter. Then, I select the best complexity parameter to prune the tree. On the other hand, I want to find the importance variable to the class variable in the random forest and order them by decreasing order. Finally, I am going to estimate these two models to find which one is better.

**Data Analysis**

From figure 2, I can know the edible state of type of odor mushroom and the distribution of the odor. Only the odorless mushrooms cannot be identified if the mushroom is edible or poisonous. So, according to the number of zero, I am able to estimate which characteristic is more important to know the toxicity of mushrooms. From the figure 3, I find that the “odor” variable is the best way to estimate the toxicity of mushrooms. And the “stalk.color.above.ring” variable is as important as “stalk.color.below.ring” to verify the toxicity of mushroom.

The picture 4 depicts the percentage of the “class” variable in each model dataset. From this picture, I notice that the random sample appears to have created roughly the same ratio of edible to poisonous upon creating train and test data compared with the original data, which also indicates that I create random sample don’t have bias and are good for varifying the model. Then. I start to draw the decision tree and show the situation in the picture 5. I can notice that when the odorless mushrooms appear, the 46 percentages probability prove the mushroom is toxic. When the “gill.size” variable is same as“b” alphabet after filtering the “ spore.print.color” and “ population” variables, people can guarantee the mushrooms are poisonous. Now, it is time to train the decision trees model via training data to optimize the decision trees model. After we prune the tree with training data set and leverage the test data set to estimate the model. I find the accuracy of model is 1 showed in figure 6 that means the model is good and accurate enough.

For the random forest model, I use the training data to fit the random forest model and print the random forest model. In the figure 8, I can know the number of variables tried at each split to be 4 and an OOB estimate of error rate 0.5%. Thus, I can ensure the training model fit the training data almost perfectly. Also, from the confusion matrix item, I also notice that the only two mushrooms are identified by mistake. The random regression model would have predicted 2 editable mushrooms as to be poisonous. Next, I would like to explore which variable is more important to identity the toxicity of mushroom. Based on figure 9, we can know that the odor is the most one of important factor to estimate the toxicity of mushrooms as the decision trees model indicate. Finally, I decide to use the testing data to estimate the accuracy of the random forest model. I find there is 99.9 percentage chance to guarantee the model is effective. It is almost close to 100 percentage accuracy.

**Conclusion**

In this module, I find both models are good enough to predict and analyze the toxicity of mushrooms. The decision trees model is 100 percentage accuracy when I use the testing data to estimate the model. Compared with decision trees model, the random forest model also has a high degree of accuracy. They both are almost same. Furthermore, I also notice that to smell the odor of mushroom is the best to know whether the mushrooms are poisonous or not. People is sensitive to the smell. Therefore, anyone can be easily to identify the toxicity of most of mushrooms.

**References**

GitHub. Retrieved from https://github.com/stoltzmaniac/Mushroom-Classification/blob/master/mushroomRandomForest-modified%20for%20ghost

Mushroom Classification Retrieved from https://www.kaggle.com/uciml/mushroom-classification

Scott, S., (2017). Random Forest Classification of Mushrooms. Retrieved from https://www.r-bloggers.com/2017/01/random-forest-classification-of-mushrooms/

UCI Machine Learning (2016). Mushroom Classification. Retrieved from https://www.kaggle.com/uciml/mushroom-classification

Appendix A

**Figure1:** *Data summary*

A picture containing text, receipt

Description automatically generated

**Figure2:***The odor variable distribution under the circumstance of the class variable A picture containing calendar

Description automatically generated*

**Figure3:***The importance to identify the toxicity of mushrooms*

*Chart, histogram

Description automatically generated*

**Figure4:** *Percentage of “class” variable in each dataset*

*Graphical user interface, text, application

Description automatically generated*

**Figure5:** *Decision tree*

*Timeline

Description automatically generated*

**Figure6:** *Calculating accuracy of decision tree model*

*Text

Description automatically generated with low confidence*

**Figure7:** *Random forest model*

**Chart, histogram

Description automatically generated with medium confidence**

**Figure8:** *Print Random forest model*

*Text, letter

Description automatically generated*

**Figure9:** *Identify the toxicity of mushrooms by mushroom characteristics*

*Chart

Description automatically generated*

**Figure10:** *The accuracy of random forest*

*Text

Description automatically generated*

Appendix B

R code

Text

Description automatically generated with medium confidence

Text

Description automatically generated with medium confidence

Text

Description automatically generated

Graphical user interface, application

Description automatically generated with medium confidenceGraphical user interface, text

Description automatically generated