Choose Your Own: Mushroom Classification Project

Min Zhou

03-08-19

Contents

1 Introduction

Vitamin D is one of the most important essential micronutrients with many known and unknown biological functions in the human body. Although sunlight is the most common source of vitamin D, due to our indoor sedentary lifestyle and the use of sunscreen for the prevention of skin cancer along with many other factors, vitamin D is one of the nutrients of public health concern (ref.1). Luckily, mushrooms with ample exposure of sunlight are a great source of vitamin D. Grocery store mushrooms without the exposure of sunlight are not a good source of vitamin D. You can place them under the sunlight to harvest vitamin D (ref.2), but it is much more fun to forage your own vitamin D rich mushrooms in the wild. Mushroom hunting if done wrong can be deadly. For example, mistaking baby death caps for white button mushrooms is an often fatal mistake as consuming only half of the death cap mushroom can kill an adult human (ref.3). This is why I decided to use the UCI agaricus-lepiota mushroom data set (ref.4) to study key visual characteristics of these gilled mushrooms to separate the poisonous ones from the edible ones.

Agaricus is a genus of mushrooms that contains the most widely known edible and poisonous mushrooms (ref.5) while Lepiota is a genus of gilled mushrooms containing lethally poisonous species and zero known recommended species for consumption (ref.6). This mushroom data set contains hypothetical samples based on 23 species of gilled Agaricus and Lepiota mushrooms. Although the data set only labels each mushroom as either p (poisonous) or e (edible), originally, each species is identified as definitely edible, definitely poisonous, and unknown edibility. For safety purposes, the unknown edibility is also labeled as p for poisonous. There are total 8124 observations with 22 features. Detailed information about each of the different features can be found in section 7 of the agaricus-lepiota.names file (ref.7). The key goals for this **Choose Your Own Project** is to **find the most important visual features for accurately distinguishing poisonous mushrooms from the edible ones when mushroom foraging in the wild**. **100% accuracy** will also be the goal for this project as the consequences of being wrong can be fatal. We choose to focus on visual characteristics because these features can be much easier to qualify compared to odor and other nonvisiual attributes, especially when reference images are available. For those who are interested in more details, key visual reference tutorials for mushrooms and a spore print color guide can be found in (ref.8,9).

2 Data Analysis and Feature Selection

The initial data set is split randomly into train (~ 80%) and test (~20%) sets. All data analysis and model fitting are performed on the train set and the test set is only used for assessing model performance (prediction accuracy). Below is a quick overview of the column variable types for the train set.

	var_type
class	factor
cap_shape	factor
cap_surface	factor
cap_color	factor
bruises	factor
odor	factor
gill_attachment	factor
gill_spacing	factor
gill_size	factor
gill_color	factor
stalk_shape	factor
stalk_root	factor
stalk_surface_above_ring	factor
stalk_surface_below_ring	factor
stalk_color_above_ring	factor
stalk_color_below_ring	factor
veil_type	factor
veil_color	factor
ring_number	factor
ring_type	factor
spore_print_color	factor
population	factor
habitat	factor

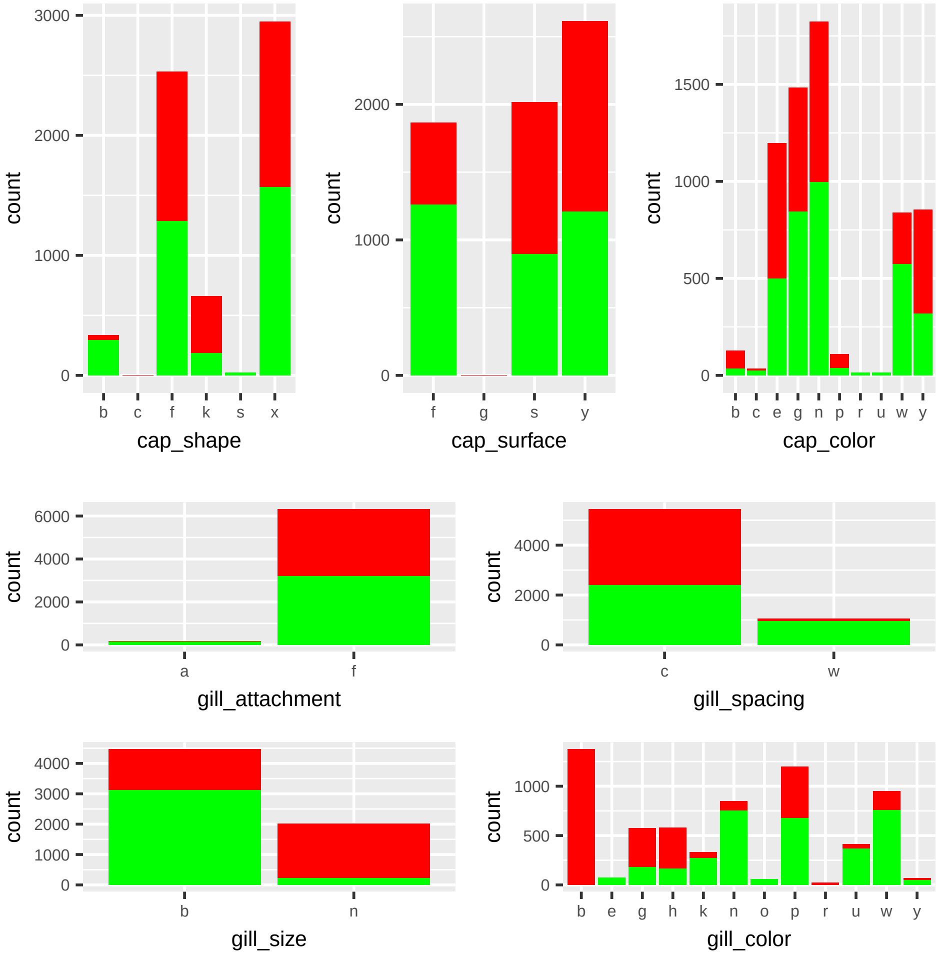
Since all columns of this data set are of variable type factor, we will take a look at the number of factor levels to check for column variabilities.

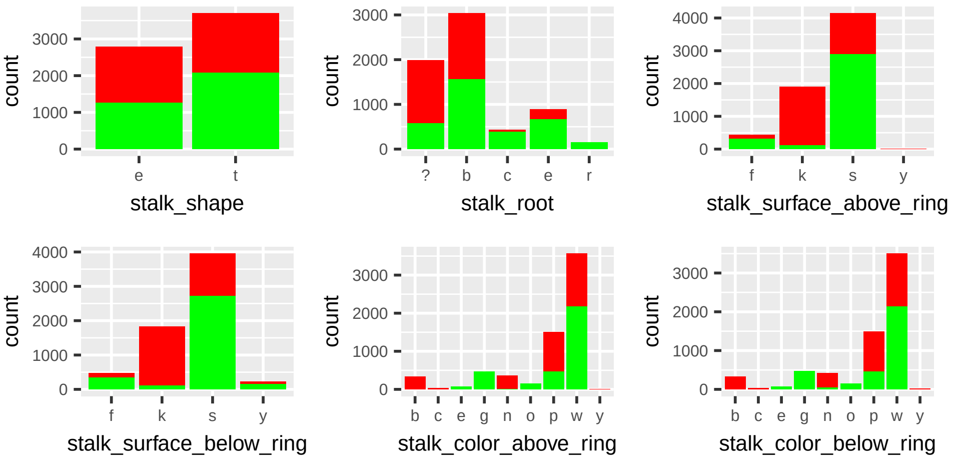
	factor_level_number
class	2
cap_shape	6
cap_surface	4
cap_color	10
bruises	2
odor	9
gill_attachment	2
gill_spacing	2
gill_size	2
gill_color	12
stalk_shape	2
stalk_root	5
stalk_surface_above_ring	4
stalk_surface_below_ring	4
stalk_color_above_ring	9
stalk_color_below_ring	9
veil_type	1
veil_color	4
ring_number	3
ring_type	5
spore_print_color	9
population	6
habitat	7

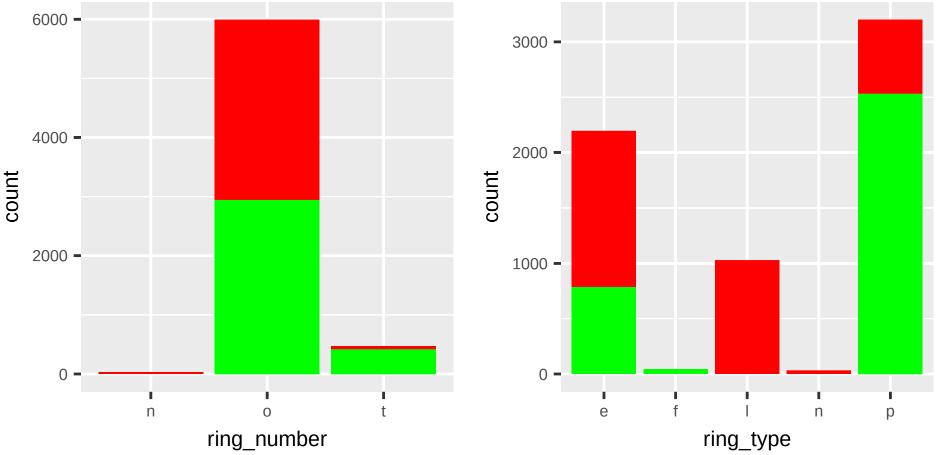
In addition to the removal of nonvisual features (`odor`, `population`, and `habitat`), the `veil_type` is removed since it has zero column variablity (only one level). Before further data exploration, let's check for any NAs in the remaining columns of the train set.

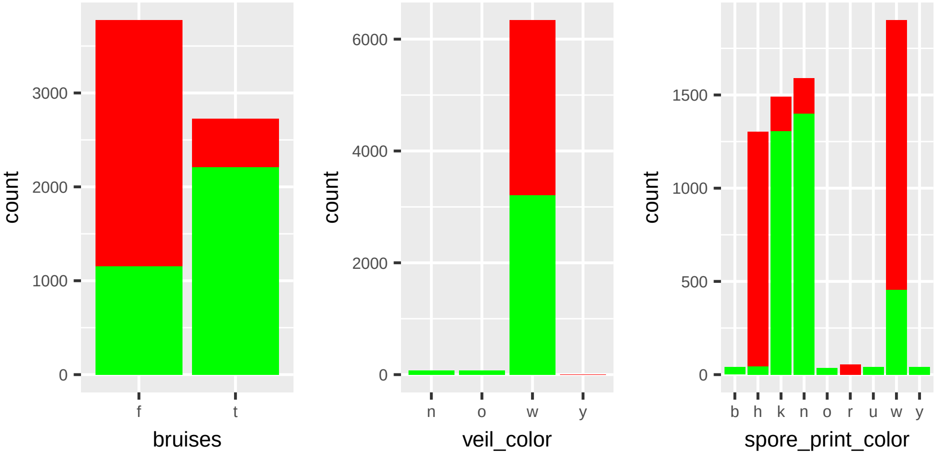
	Number_of_NAs
class	0
cap_shape	0
cap_surface	0
cap_color	0
bruises	0
gill_attachment	0
gill_spacing	0
gill_size	0
gill_color	0
stalk_shape	0
stalk_root	0
stalk_surface_above_ring	0
stalk_surface_below_ring	0
stalk_color_above_ring	0
stalk_color_below_ring	0
veil_color	0
ring_number	0
ring_type	0
spore_print_color	0

With zero NAs, we now look at the relationship between each remaining features and the mushroom class label visually first.









In the above plots, the red color designates poisonous mushrooms and the green color represents edible mushrooms. From the plots, we can see a number of features always indicate a poisonous mushroom (e.g, green `gill_color` or `spore_print_color`), while others always indicate an edible mushroom (e.g., green and purple `cap_color` and flaring `ring_type`). The plots definitely suggest strong correlation between each of the remaining features and the mushroom class label. To check the correlation statistically, we will perform a Chi-squared test and use a p-value of 0.01 to reject or accept the H_0 hypothesis (the null hypothesis: the two variables are independent). If the p-value is less than 0.01, we will reject H_0 and assume that there is a correlation between the feature and the class label (ref.10).

features	**correlation**
cap_shape	correlated
cap_surface	correlated
cap_color	correlated
bruises	correlated
gill_attachment	correlated
gill_spacing	correlated
gill_size	correlated
gill_color	correlated
stalk_shape	correlated
stalk_root	correlated
stalk_surface_above_ring	correlated
stalk_surface_below_ring	correlated
stalk_color_above_ring	correlated
stalk_color_below_ring	correlated
veil_color	correlated
ring_number	correlated
ring_type	correlated
spore_print_color	correlated

The Chi-squared test result suggests that all 18 remaining features are correlated with the class label. Before we move onto model fitting, let's use Chi-squared test to also check if the features are correlated with one another, again using p-value of 0.01.

	cap_shape	cap_surface	cap_color	bruises	gill_attachment	gill_spacing
cap_shape	correlated	correlated	correlated	correlated	correlated	correlated
cap_surface	correlated	correlated	correlated	correlated	correlated	correlated
cap_color	correlated	correlated	correlated	correlated	correlated	correlated
bruises	correlated	correlated	correlated	correlated	correlated	correlated
gill_attachment	correlated	correlated	correlated	correlated	correlated	correlated
gill_spacing	correlated	correlated	correlated	correlated	correlated	correlated
gill_size	correlated	correlated	correlated	correlated	correlated	correlated
gill_color	correlated	correlated	correlated	correlated	correlated	correlated
stalk_shape	correlated	correlated	correlated	correlated	correlated	correlated
stalk_root	correlated	correlated	correlated	correlated	correlated	correlated
stalk_surface_above_ring	correlated	correlated	correlated	correlated	correlated	correlated
stalk_surface_below_ring	correlated	correlated	correlated	correlated	correlated	correlated
stalk_color_above_ring	correlated	correlated	correlated	correlated	correlated	correlated
stalk_color_below_ring	correlated	correlated	correlated	correlated	correlated	correlated
veil_color	correlated	correlated	correlated	correlated	correlated	correlated
ring_number	correlated	correlated	correlated	correlated	correlated	correlated
ring_type	correlated	correlated	correlated	correlated	correlated	correlated
spore_print_color	correlated	correlated	correlated	correlated	correlated	correlated

	gill_size	gill_color	stalk_shape	stalk_root	stalk_surface_above_ring	stalk_surface_below_ring
cap_shape	correlated	correlated	correlated	correlated	correlated	correlated
cap_surface	correlated	correlated	correlated	correlated	correlated	correlated
cap_color	correlated	correlated	correlated	correlated	correlated	correlated
bruises	correlated	correlated	correlated	correlated	correlated	correlated
gill_attachment	correlated	correlated	correlated	correlated	correlated	correlated
gill_spacing	correlated	correlated	correlated	correlated	correlated	correlated
gill_size	correlated	correlated	correlated	correlated	correlated	correlated
gill_color	correlated	correlated	correlated	correlated	correlated	correlated
stalk_shape	correlated	correlated	correlated	correlated	correlated	correlated
stalk_root	correlated	correlated	correlated	correlated	correlated	correlated
stalk_surface_above_ring	correlated	correlated	correlated	correlated	correlated	correlated
stalk_surface_below_ring	correlated	correlated	correlated	correlated	correlated	correlated
stalk_color_above_ring	correlated	correlated	correlated	correlated	correlated	correlated
stalk_color_below_ring	correlated	correlated	correlated	correlated	correlated	correlated
veil_color	correlated	correlated	correlated	correlated	correlated	correlated
ring_number	correlated	correlated	correlated	correlated	correlated	correlated
ring_type	correlated	correlated	correlated	correlated	correlated	correlated
spore_print_color	correlated	correlated	correlated	correlated	correlated	correlated

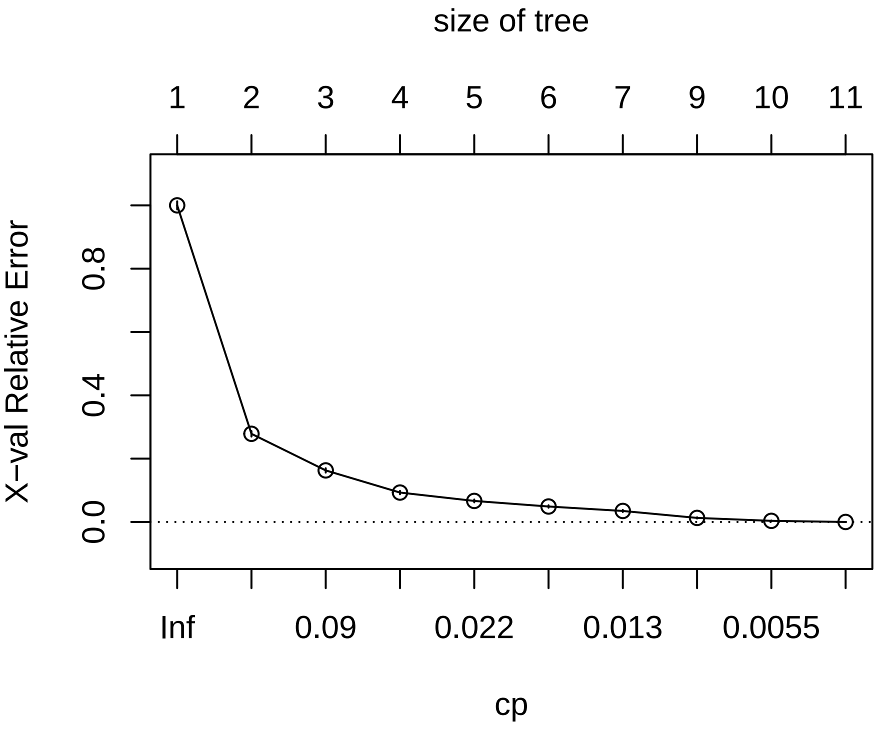
	stalk__color__above__ring	stalk__color__below__ring	veil__color	ring__number	ring__type	spore__print__color
cap__shape	correlated	correlated	correlated	correlated	correlated	correlated
cap__surface	correlated	correlated	correlated	correlated	correlated	correlated
cap__color	correlated	correlated	correlated	correlated	correlated	correlated
bruises	correlated	correlated	correlated	correlated	correlated	correlated
gill__attachment	correlated	correlated	correlated	correlated	correlated	correlated
gill__spacing	correlated	correlated	correlated	correlated	correlated	correlated
gill__size	correlated	correlated	correlated	correlated	correlated	correlated
gill__color	correlated	correlated	correlated	correlated	correlated	correlated
stalk__shape	correlated	correlated	correlated	correlated	correlated	correlated
stalk__root	correlated	correlated	correlated	correlated	correlated	correlated
stalk__surface__above__ring	correlated	correlated	correlated	correlated	correlated	correlated
stalk__surface__below__ring	correlated	correlated	correlated	correlated	correlated	correlated
stalk__color__above__ring	correlated	correlated	correlated	correlated	correlated	correlated
stalk__color__below__ring	correlated	correlated	correlated	correlated	correlated	correlated
veil__color	correlated	correlated	correlated	non__correlated	correlated	correlated
ring__number	correlated	correlated	non__correlated	correlated	correlated	correlated
ring__type	correlated	correlated	correlated	correlated	correlated	correlated
spore__print__color	correlated	correlated	correlated	correlated	correlated	correlated

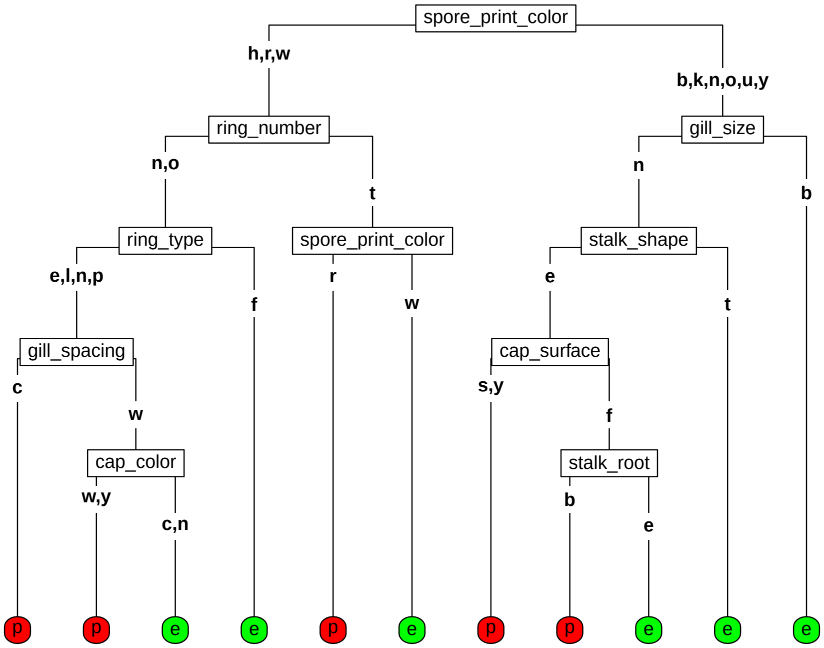
We can see that, with the exception of `ring_number` and `veil_color`, all the other features appear to be intercorrelated according to the Chi-squared test. For this reason, we will keep all 18 features for now and use tree based models for classification and feature importance ranking because tree based models don't require the attributes to be independent. Tree based models also have many other benefits especially their ease of use.

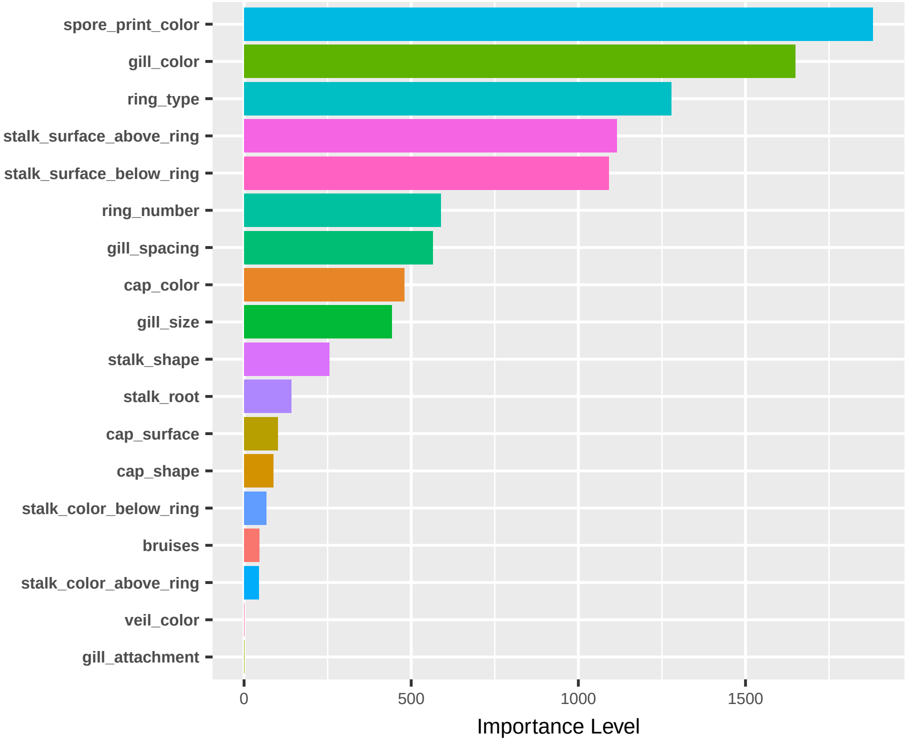
3 Method and Analysis

3.1 Decision Tree

rpart and *rpart.plot* libraries are used to perform decision tree classification. We will use `rpart.control` and `plotcp` to find the best Complexity Parameter, `cp` (ref.11).





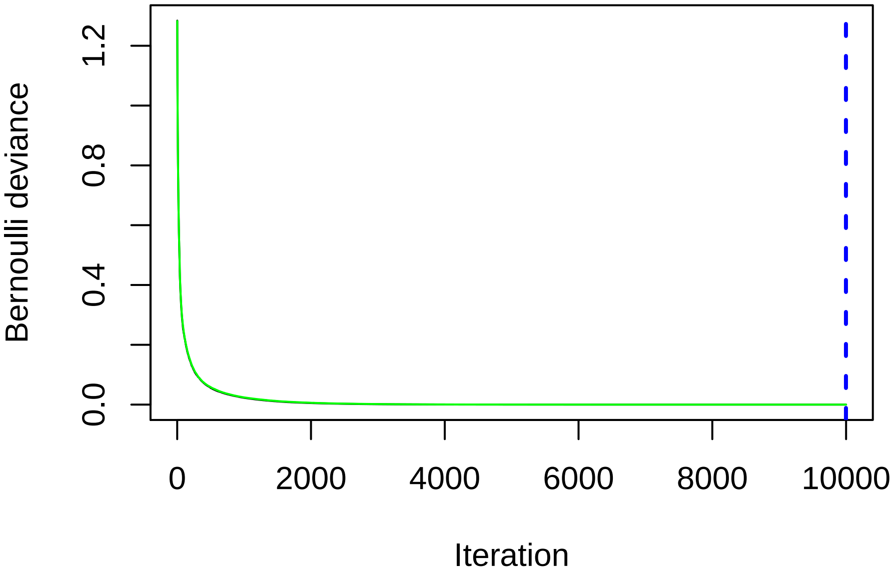


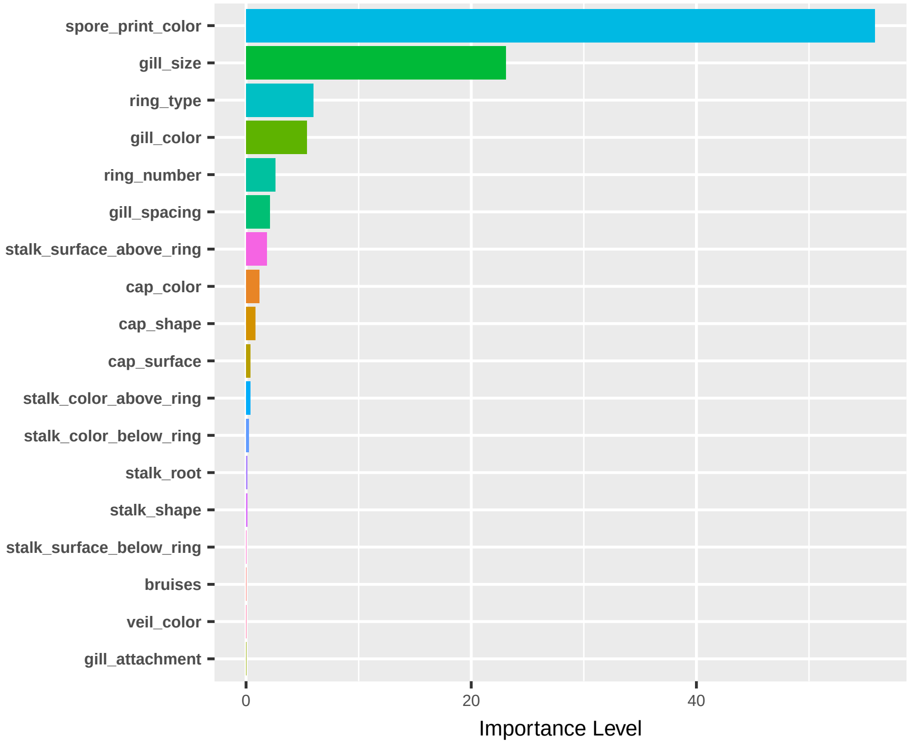
The `cp` plot shows that a `cp` of 0 gives the lowest error. Looking at the `Importance Level` plot, the top 5 features using *decision tree classification method* are `spore_print_color`, `gill_color`, `ring_type`, `stalk_surface_above_ring`, and `stalk_surface_below_ring` while `gill_attachment` and `veil_color` have almost zero importance. The *accuracy* using this simple method is 100%. The splitting process can be easily interpreted visually using the decision tree plot above where the red "p" stands for poisonous and the green "e" stands for edible.

One thing to note is that although `gill_color` is the second most important feature in the `Importance Level` plot, the feature is not present in the decision tree splitting plot. This is easily explained. First, the feature importance here is calculated based on the *Gini Importance* but the splitting of the tree is based on *Gini Impurity* (ref.12). The two criteria are not the same. In addition, if you review the mushroom spore print color guide link (ref.9) in the **References** section, the site mentions that the easiest way to check spore color is to look at the gill color if the mushrooms are mature. The `gill_color` is heavily correlated to the `spore_print_color` for mature mushrooms.

3.2 Gradient Boosted Machine (GBM)

Although the basic *decision tree classification method* gives an *accuracy* of 100% and provides a very easy to understand decision tree plot, the features are highly intercorrelated and decision tree splits can be highly variable with just slight changes in the observations (ref.13). To make the classification more robust, we will use *gradient boosted decision tree model* (ref.14). We will use the *gbm* library and use `gbm.perf` with a cross validation of 5 folds to test for the optimal number of trees for the classification prediction. The *gbm model* also provides information for feature importance (ref.15).



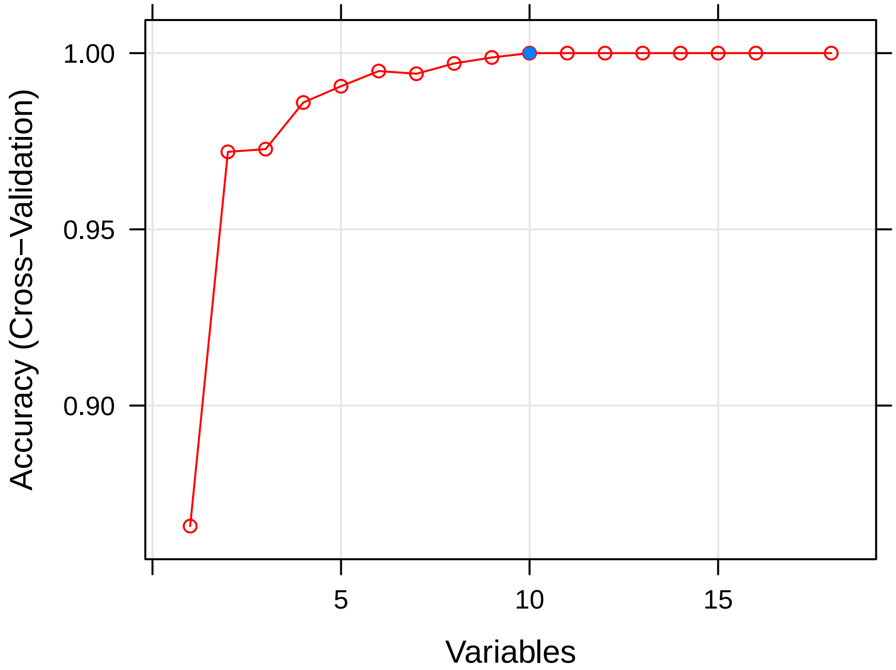


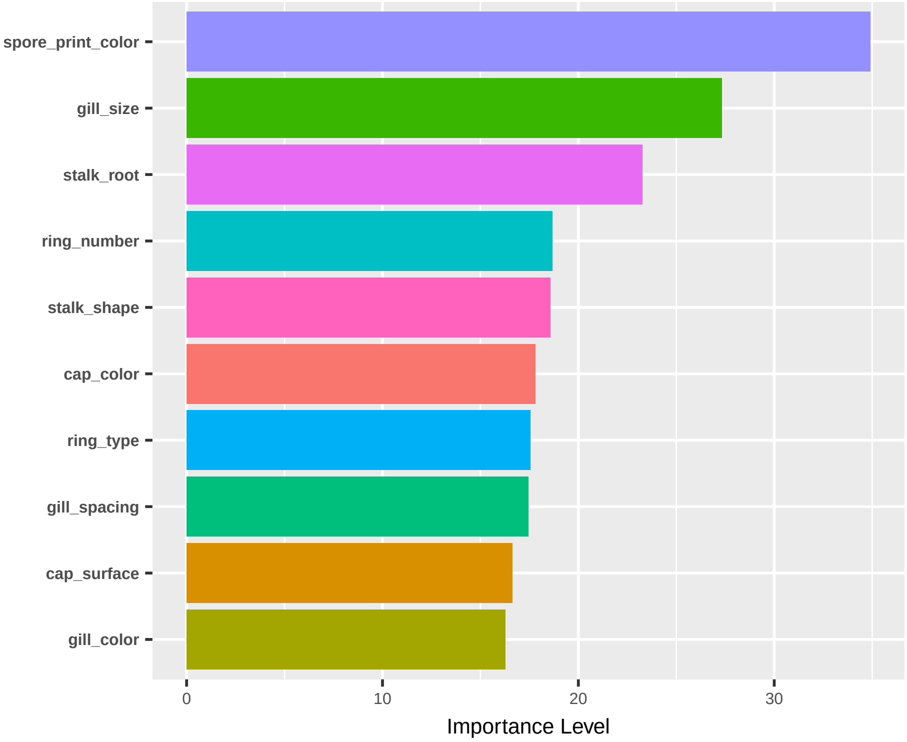
The `Iteration` (number of trees) plot shows that the optimal number of trees is close to 10000. The `Importance Level` plot shows the top 5 features using *gradient boosted decision tree method* are `spore_print_color`, `gill_size`, `ring_type`, `gill_color`, and `ring_number` while `gill_attachment`, `veil_color`, `bruises`, `stalk_surface_below_ring`, `stalk_shape`, and `stalk_root` have very tiny effects. The *accuracy* using this more sophisticated method is also 100%.

Both models appear to agree that `veil_color` and `gill_attachment` have very little effect on the classification while `spore_print_color`, `ring_type`, and `gill_color` tend to have high importance for distinguishing the poisonous mushrooms from the edible ones.

3.3 Random Forest with Recursive Feature Elimination (RF_RFE)

The first two models seem to suggest that we can remove `veil_color` and `gill_attachment` from the 18 features for 100% *accuracy* prediction. Before we test out the theory, we will use the *recursive feature elimination method* (RFE) from the **caret** package and use `rfFuncs` and a cross validation of 10 folds to find out the optimal number and combination of features. `rfFuncs` is one of the pre-defined sets of RFE functions in the **caret** package for the *random forest* (RF) model (ref.16). This method is especially useful for features that are intercorrelated (ref.17).





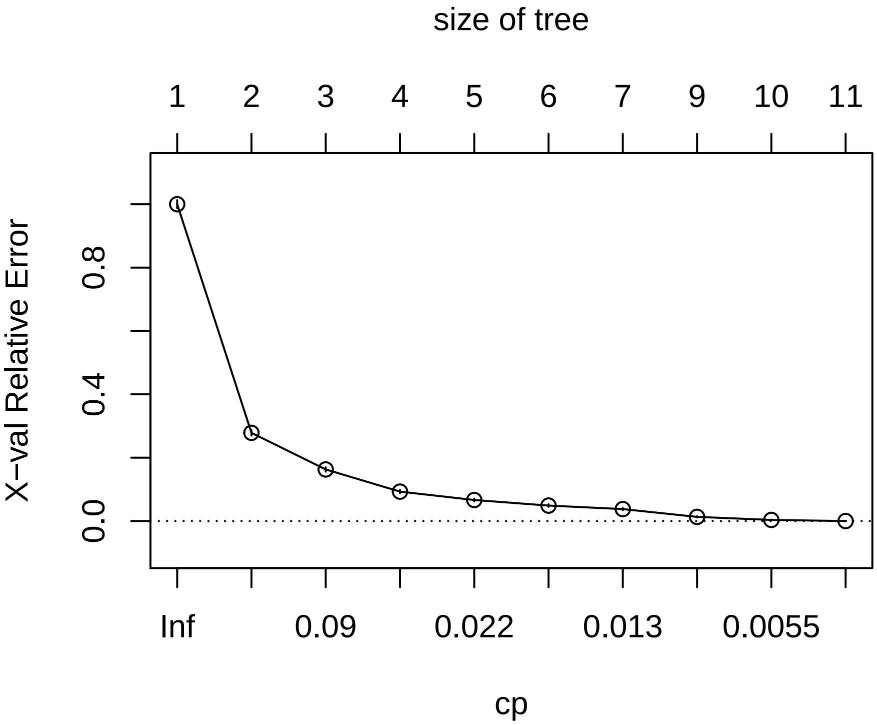
The `Variables` plot shows that only 10 out of the 18 features are needed for 100% *accuracy* and the `Importance Level` plot shows the 10 selected features with the most important attribute at the top and the least important at the bottom. The top 5 features are `spore_print_color`, `gill_size`, `stalk_root`, `ring_number`, and `stalk_shape`. As expected from the first two models, both `veil_color` and `gill_attachment` are not needed for accurate classification along with `cap_shape`, `bruises`, `stalk_surface_above_ring`, `stalk_surface_below_ring`, `stalk_color_above_ring`, and `stalk_color_below_ring`.

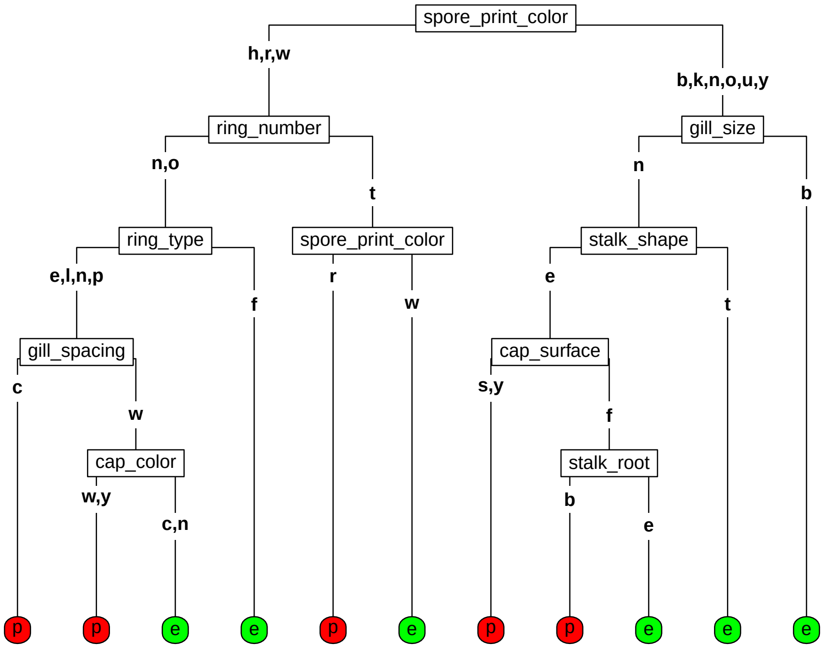
Although the top 5 features for the 3 models are not in agreement, `spore_print_color` and `ring_number` appear to be ranked high in all of them and `veil_color`, `gill_attachment`, `bruises`, `stalk_color_above_ring`, `cap_shape`, and `stalk_color_below_ring` are ranked low in all of them.

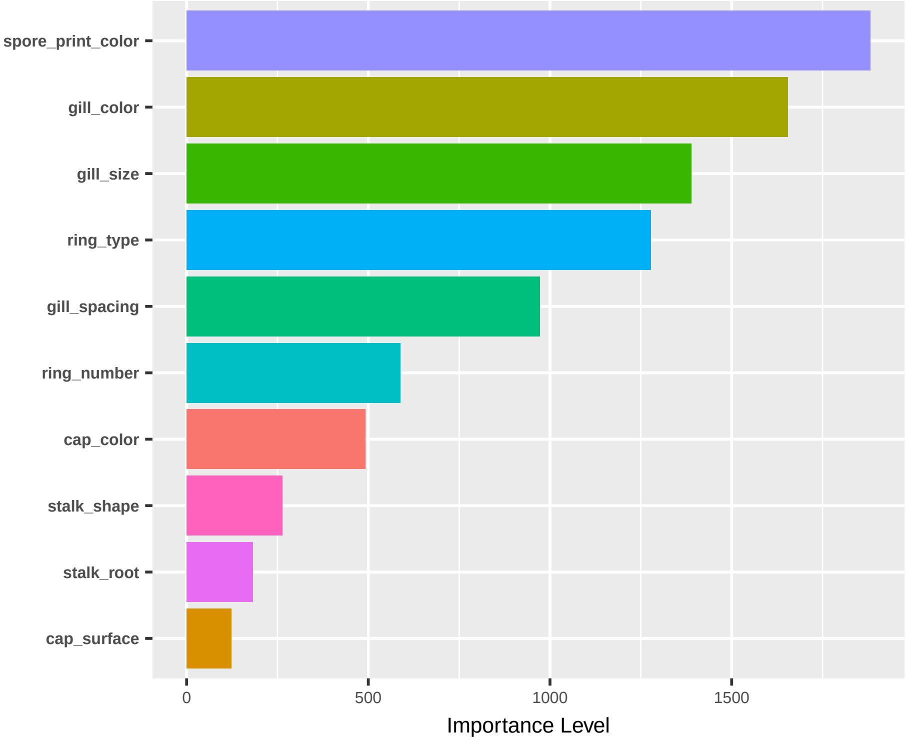
3.4 $Model_1$ and $Model_2$ Revisited

After finding out the 10 optimal features using *RF_RFE*, let's revisit both the *decision tree classification method* and *gradient boosted decision tree method* using only the 10 selected features.

3.4.1 Decision Tree Revisited

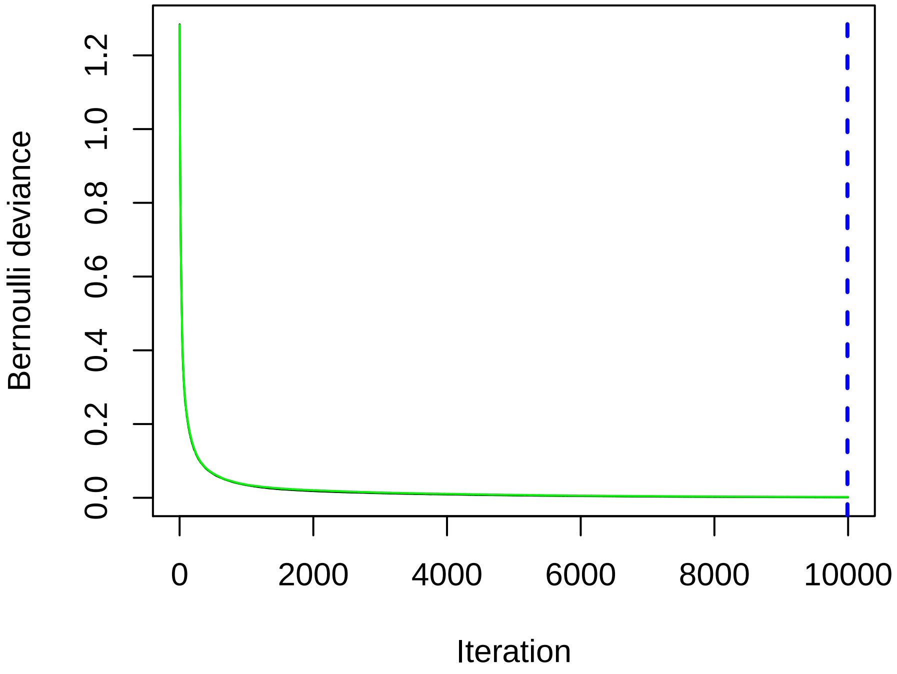


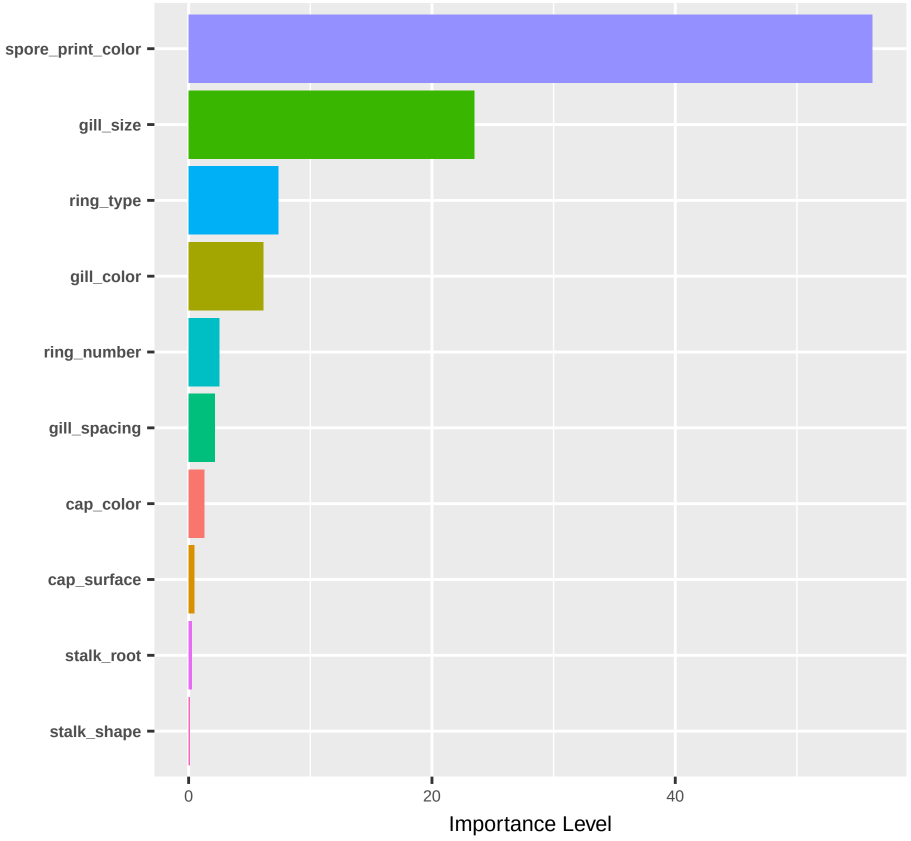




The `cp` plot again shows that a `cp` of 0 gives the lowest error and according to the `Importance Level` plot, the new top 5 features using only the 10 features are `spore_print_color`, `gill_color`, `gill_size`, `ring_type`, and `gill_spacing` while `cap_surface` and `stalk_root` have the lowest importance. Both *decision tree* models have `spore_print_color`, `gill_color` and `ring_type` as the most important attributes. Interestingly, the decision tree plots for both models are exactly the same. As expected, the *accuracy* again is 100% since the splitting process for both tree models are exactly the same as illustrated in the tree plots.

3.4.2 GBM Revisited





The Iteration plot again shows that the optimal number of trees is close to 10000 and the top 5 features using only the 10 features are the same as $model_2$. stalk_shape, stalk_root, and cap_surface remain as the least important attributes. Not surprisingly, the *accuracy* is still 100% since the overall importance ranking stays basically the same for both GBM models.

4 Results

models	**accuracy**
tree	100%
GBM	100%
RF_RFE	100%
tree revisit	100%
GBM revisit	100%

tree_top_5	**GBM_top_5**	**RF_RFE_top_5**	**tree_top_5_r**	**GBM_top_5_r**
spore_print_color	spore_print_color	spore_print_color	spore_print_color	spore_print_color
gill_color	gill_size	gill_size	gill_color	gill_size
ring_type	ring_type	stalk_root	gill_size	ring_type
stalk_surface_above_ring	gill_color	ring_number	ring_type	gill_color
stalk_surface_below_ring	ring_number	stalk_shape	gill_spacing	ring_number

Above are comparison tables for the 5 models. The second table lists the most important feature at the top and the 5th most important feature at the bottom. The *accuracy* remains 100% for all of the models and spore_print_color is consistently the most important attribute. gill_color, gill_size, and ring_type are listed among the top 5 features for 4 out of the 5 models. ring_number is listed in the top 5 features for GBM models and RF_RFE model and is the 6th most important feature in both tree models.

5 Conclusions

After using 5 tree based models, we are able to identify spore_print_color, gill_color, gill_size, ring_type, and ring_number as the overall 5 most important visual features. Out of the initial 22 features, 10 visual features are selected for the goal of 100% classification *accuracy*. The final 10 features are spore_print_color, gill_size, stalk_root, stalk_shape, gill_color, cap_color, gill_spacing, ring_type, cap_surface, ring_number. Since spore_print_color is the most important attribute, it is crucial to be patient and let mushrooms leave a thick enough deposit on a white paper overnight before consumption. That being said, as mentioned in the **Decision Tree** section, sometimes you can use gill_color to find out the spore_print_color for mature mushrooms because as more of the spores mature, the gill color changes closer to the color of the spores. The decision tree splitting plot maps out a very easy to understand process of classifying the mushrooms into poisonous and edible ones using 9 out of the 10 selected visual features (no gill_color). It is important to note that decision tree is not stable and any changes in the training set can change the splitting tree (ref.13). Although GBM and RF models are harder to interpret, both methods provide more robust information based upon many trees (ref.14) and we are still able to gain valuable information on the most important visual attributes. Finally, it needs to be noted that the mushroom data set only contains a small subset of the mushrooms (ref.18) and more valuable mycological data should be acquired for more robust mushroom classification.

6 References

1. Dietary Guidelines 2015-2010

2. Mushroom in Sunlight for vitamin D

3. Wikipedia-Amanita phalloides (death cap)

4. UCI agaricus-lepiota mushroom data set

5. Wikipedia-Agaricus

6. Wikipedia-Lepiota

7. UCI agaricus-lepiota.names

8. Mushroom Visual Reference Tutorials

9. Spore Print Color Guide

10. Chi-Squared Test

11. decision tree - rpart package

12. gini-impurity and gini-importance

13. drawbacks of decision tree

14. tree based methods

15. GBM package

16. caret Recursive Feature Elimination

17. Feature Selection Using Random Forest

18. Scientific and Common Names of Mushrooms