Mushroom Classification

Lauren Foltz (lvfoltz@syr.edu) &
Baskar Dakshin (bdakshin@syr.edu)

IST 707 Data Analytics Final Project
Syracuse University / December 2018

Table of Contents

Introduction

Mushrooms, the fruiting body of fungi, have been eaten by humans for thousands of years. All mushrooms contain protein, fiber, and the powerful antioxidant selenium, but specific types are sought after for specific health benefits. Shiitake mushrooms, for instance, contain all 8 essential amino acids, as well as eritadenine, a compound that reduces cholesterol. Reishi mushrooms are valued for their immune boosting effects, maitake for their stabilizing impact on blood sugar, and porcini for their anti-inflammatory properties. Initially, including mushrooms in the diet meant foraging, and came with a risk of ingesting poisonous mushrooms. However, beginning in the 1600s, many varieties of mushrooms have been successfully cultivated. *Agaricus bisporus* is one of the most consumed mushrooms in the world, and is cultivated in over 70 countries. The top mushroom producer in the world is China (5 million tons), followed by Italy (762K tons), and the United States (391 tons). Within the United States, the majority of mushrooms are grown in Pennsylvania.

Unlike plants, fungi do not get energy from sunlight, but from decomposing matter, and tend to grow well in moist environments. A shady environment is not a requirement, but it does help them retain their moisture. When the conditions are right (generally in autumn), the network of mycelium will produce fruiting bodies, which first look like pins, consisting of thin stalk and tiny cap. Although they start out small, the fruiting bodies quickly "mushroom." Once the cap, which looks like an umbrella, grows large enough, the veil (a thin membrane underneath the cap) ruptures, allowing the gills to drop spores. If the spores find their way to an appropriate growth substrate, they will germinate, and fungal filaments will appear. Some fungi require a certain amount of light before fruiting, while others can grow in dark caves.

Some mushrooms contain psilocybin, a psychedelic compound that has been illegal in the United States for over 40 years. However, attitudes may be changing, as different parties look into the possible therapeutic benefits and weigh them against the risks. Although psilocybin is illegal in the US, a company called COMPASS Pathways recently got their research on using psilocybin to treat depression fast tracked by the FDA. Another study is investigating the benefits of psilocybin in easing anxiety and depression among cancer patients. About 80% of the cancer patients in this study showed significant reductions in anxiety and depression after a single dose; the results lasted for months and there were minimal side effects. A study at Johns Hopkins, published in 2006, noted that study participants ranked their psilocybin experience as one of the most meaningful in their lives, comparable to the birth of a first child or death of a first parent. Trials are also underway for treatment of alcoholism and tobacco addiction.

Analysis

The Data

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 (1981). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. The data was retrieved from: https://www.kaggle.com/uciml/mushroom-classification

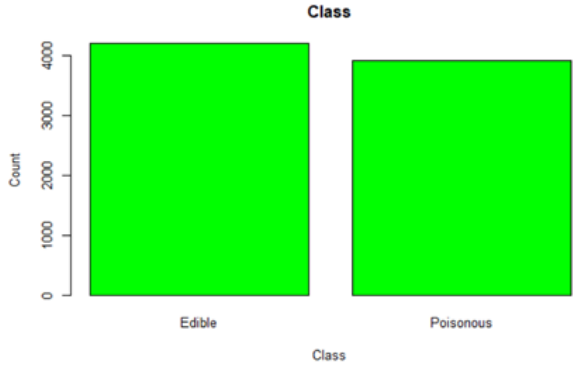
The data used contained 8,124 rows and 23 columns. Each row represented a mushroom, and it was labeled as edible or poisonous. There were no missing values, but a value of "?" was present under "stalk root" for many rows. This column was removed for many of the models. Additionally, veil type was found to only have one level, so it was removed in order to decrease dimensionality. Variable values were updated from single letters to full words using *recode*, so that the models and visuals would make sense to humans. A histogram was created for each variable, and distributions were observed. For some models, the data was separated into a training set and an unlabeled testing set. Labels were checked to ensure they were factors. No incorrect values or outliers were observed.

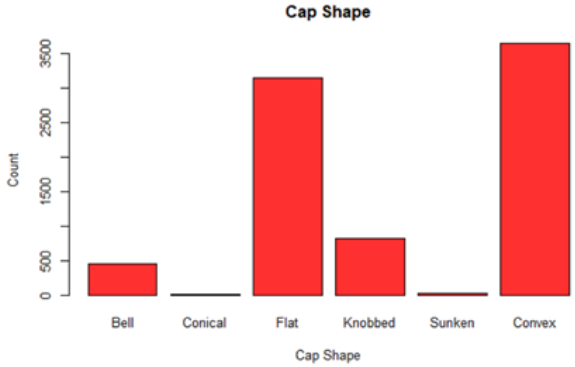
Mushroom Variables:

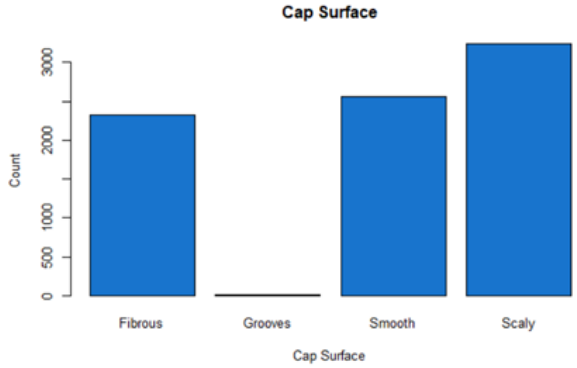
class
cap shape
cap surface
cap color
bruises
odor
gill attachment
gill spacing
gill size
gill color
stalk shape
stalk root

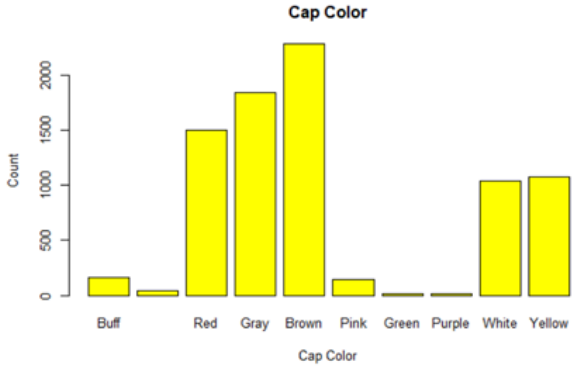
stalk surface (above the ring)
stalk surface (below the ring)
stalk color (above the ring)
stalk color (below the ring)
veil type
veil color
ring number
ring type
spore print color
population
habitat

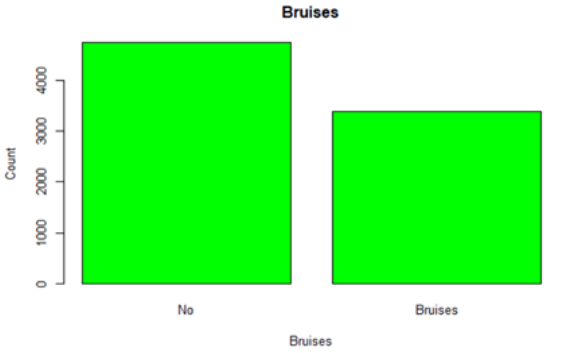
Exploratory Data Analysis

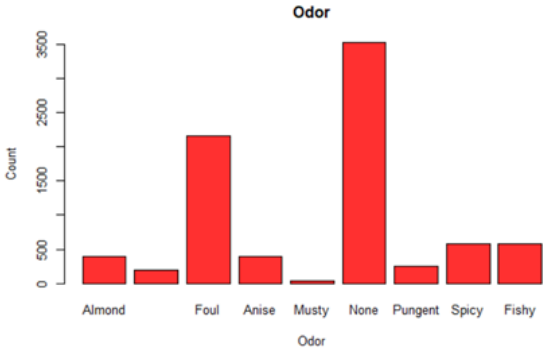


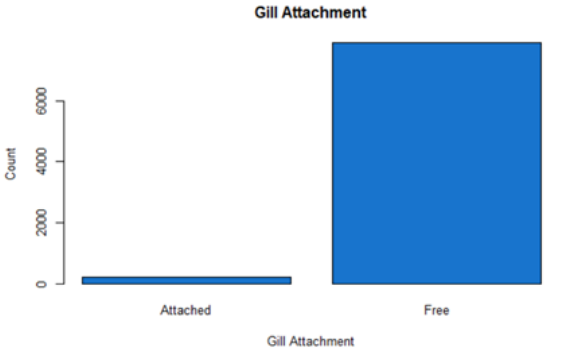


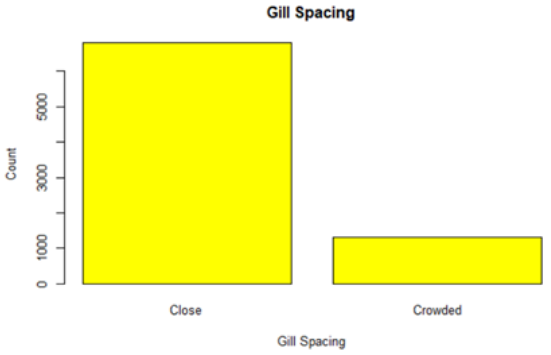


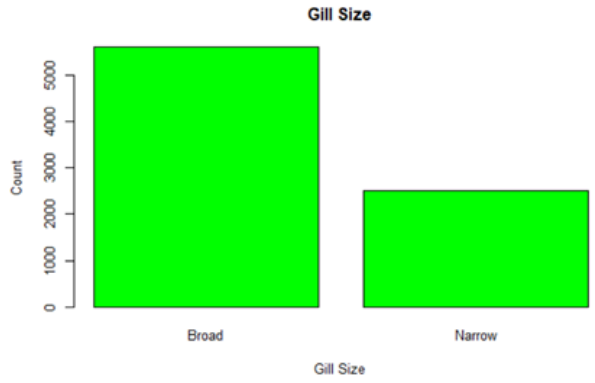


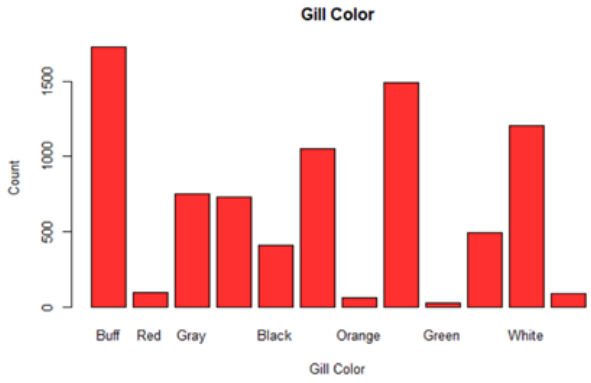


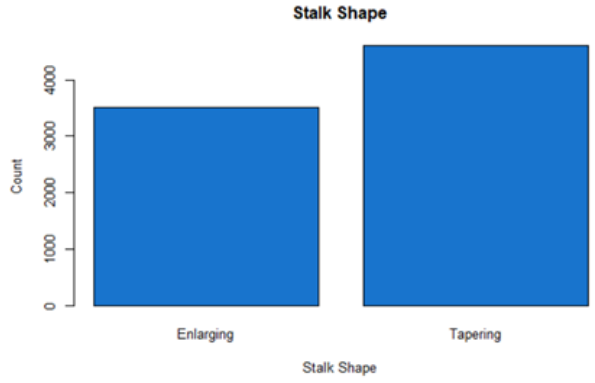


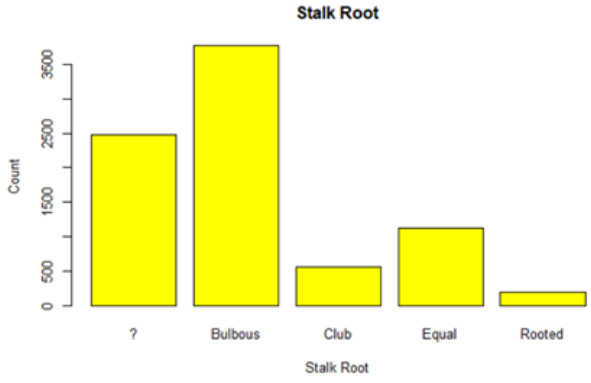


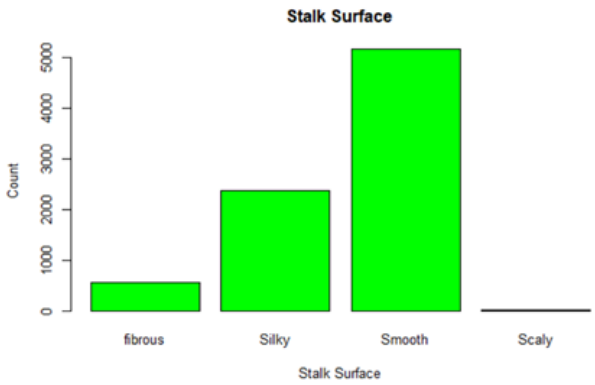


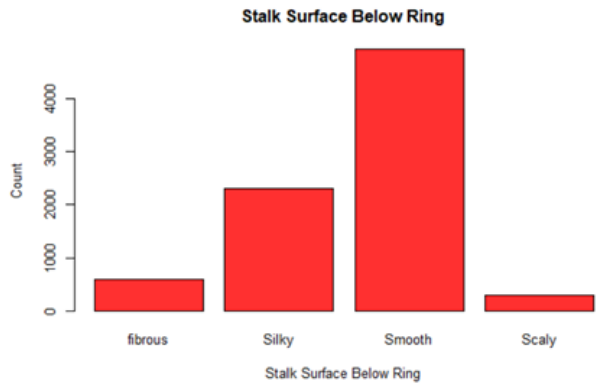

Gill Size
Count
0
1000
2000
3000
4000
5000
Broad
Narrow
Gill Size

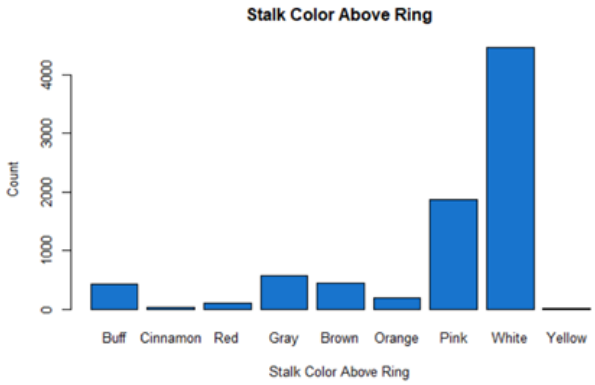

Gill Color
Count
0
500
1000
1500
Buff
Red
Gray
Black
Orange
Green
White
Gill Color

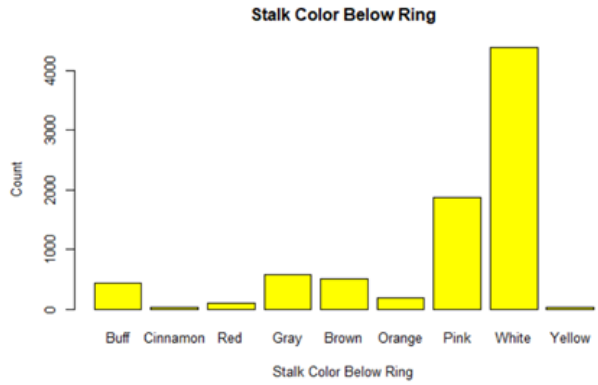

Stalk Shape
Count
0
1000
2000
3000
4000
Enlarging
Tapering
Stalk Shape

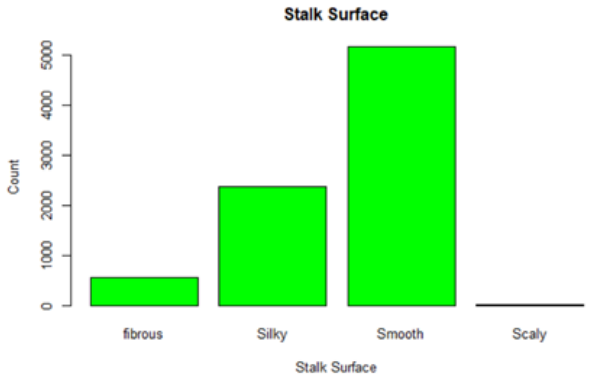

Stalk Root
Count
0
500
1500
2500
3500
?
Bulbous
Club
Equal
Rooted
Stalk Root

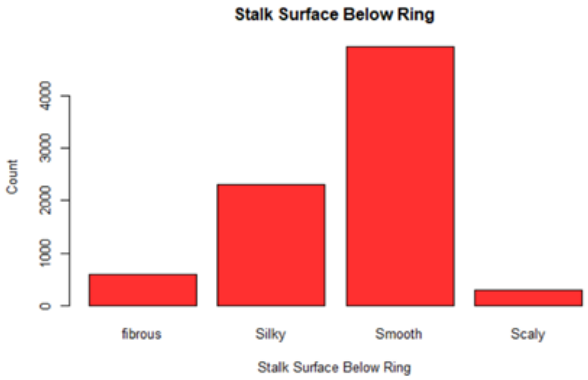

Stalk Surface
Count
0
1000
2000
3000
4000
5000
fibrous
Silky
Smooth
Scaly
Stalk Surface

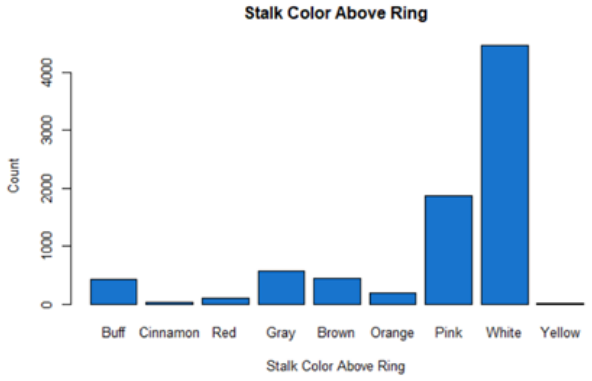

Stalk Surface Below Ring
Count
0
1000
2000
3000
4000
fibrous
Silky
Smooth
Scaly
Stalk Surface Below Ring

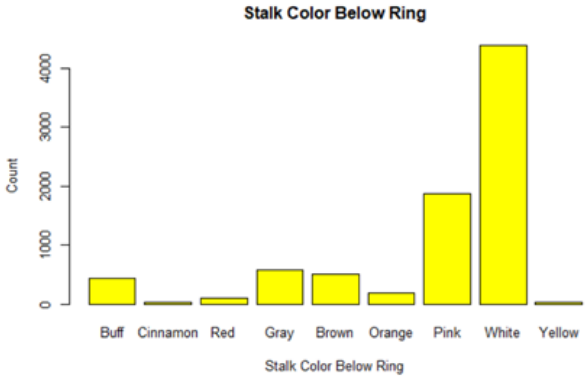

Stalk Color Above Ring
Count
0
1000
2000
3000
4000
Buff
Cinnamon
Red
Gray
Brown
Orange
Pink
White
Yellow
Stalk Color Above Ring

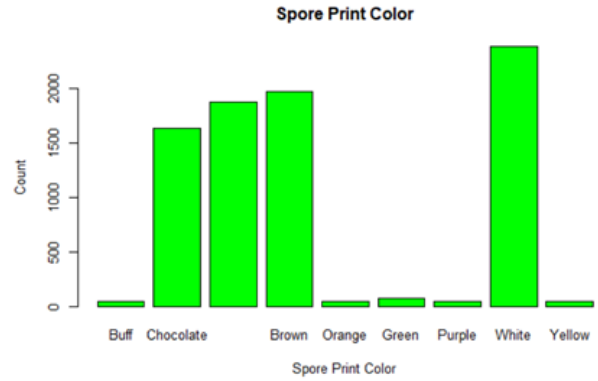

Stalk Color Below Ring
Count
0
1000
2000
3000
4000
Buff
Cinnamon
Red
Gray
Brown
Orange
Pink
White
Yellow
Stalk Color Below Ring

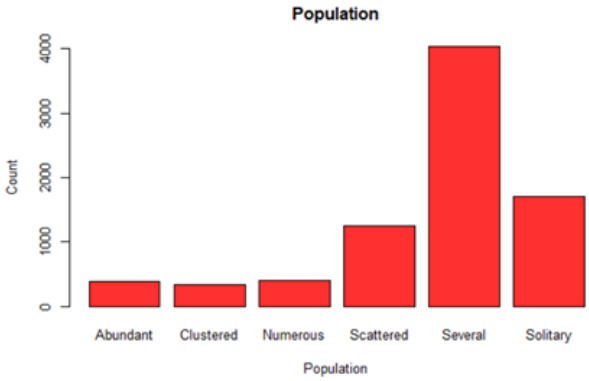

Stalk Surface
Count
0
1000
2000
3000
4000
5000
fibrous
Silky
Smooth
Scaly
Stalk Surface

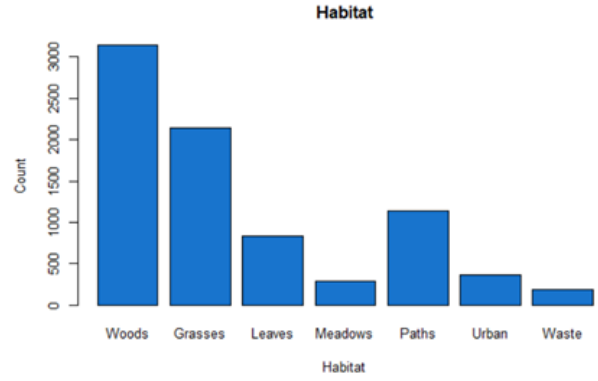

Stalk Surface Below Ring
Count
0
1000
2000
3000
4000
fibrous
Silky
Smooth
Scaly
Stalk Surface Below Ring


Stalk Color Above Ring
Count
0
1000
2000
3000
4000
Buff
Cinnamon
Red
Gray
Brown
Orange
Pink
White
Yellow
Stalk Color Above Ring


Stalk Color Below Ring
Count
0
1000
2000
3000
4000
Buff
Cinnamon
Red
Gray
Brown
Orange
Pink
White
Yellow
Stalk Color Below Ring


Spore Print Color
Count
0
500
1000
1500
2000
Buff
Chocolate
Brown
Orange
Green
Purple
White
Yellow
Spore Print Color


Population
Count
0
1000
2000
3000
4000
Abundant
Clustered
Numerous
Scattered
Several
Solitary
Population


Habitat
Count
0
500
1000
1500
2000
2500
3000
Woods
Grasses
Leaves
Meadows
Paths
Urban
Waste
Habitat

Models

Association Rule Mining Model

Association Rule Mining is a type of unsupervised learning that evaluates transactions for correlation/association. In R, the *arules* function from the "Apriori" package is used to get rules. This method allows for several parameters to be adjusted.

Support is the fraction of all transactions that contain both the left-hand side (LHS) and the right-hand side (RHS). A common setting is 20-40% of the transactions, but due to the sparse nature of some transaction data, support may not be that high. If a support value is chosen that is too high and no rules come back, it can be incrementally lowered until rules are generated. Conversely, a low initial value can be chosen, and it can be gradually increased to get a smaller set of rules.

Confidence represents the probability that items on the RHS appear in transactions with items on the LHS. A strong confidence is 0.90 or higher, but 0.80 may be reasonable.

Lift: is a measure of independence, much like correlation. If Lift is <1, then the items are independent and there is no correlation/association. Lift values should be over 1, and although higher values are better, seeing values like 1.3 and 1.5 is more common.

Rule Length can be set using *minlen* and/or *maxlen*. Minlen=2 is used to avoid getting rules with an empty LHS. Maxlen can be used to limit the number of items in a rule, providing simpler rules.

LHS & RHS: Finally, a particular item can be select to be on the LHS (the antecedent) or the RHS (the consequent). In transaction data, the LHS represents what a person has already placed in the cart, and the RHS represents what a person is likely to add to the cart before checkout.

Summary of ARM Models

Model #	Support	Confidence	LHS	RHS	Maximum Length?	# of Rules
1	0.50	0.95				76
2	0.40	0.90				283
3	0.30	0.90		class = poisonous		12
4	0.30	0.90		class = poisonous	3	2
5	0.30	0.90		class = edible	3	9
6	0.30	0.90		class = edible		17
7	0.01	0.60	class = poisonous			6
8	0.01	0.60	class = edible			13

Clustering Models

K-Nearest Neighbour (KNN):

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The *k*-NN algorithm is among the simplest of all machine learning algorithms.

Both for classification and regression, a useful technique can be used to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones

For the analysis,Mushroom dataset was split into training and Testing set with 70/30 split ratio.The data was scaled during training of the Knn Model. TrainControl was used to Control the computational nuances of the train function and repeatedCv was used as the resampling method.

K-Means Clustering:

The *k*-means clustering algorithm is a data mining and machine learning tool used to cluster observations into groups of related observations without any prior knowledge of those relationships. By sampling, the algorithm attempts to show in which category, or cluster, the data belong to, with the number of clusters being defined by the value *k.*

The advantage of *k*-means clustering is that it tells about your data (using its unsupervised form) rather than you having to instruct the algorithm about the data at the start (using the supervised form of the algorithm).

The Mushroom dataset has categorical attributes and to use K-means clustering,the categorical data is converted into numerical using One-hot encoding.This is basically creating dummy variables for each value of the category, for all the variables.
K-means model was run with nstart = 50, indicates R will run 50 different random starting assignments and selects the lowest within cluster variation and with different center values and then the purity of the clusters was calculated.

Enhanced K-Means Clustering:

In this algorithm, two methods are used, one method for finding the better initial centroids. And another method for an efficient way for assigning data points to appropriate clusters. Enhanced k-means is also used to compute hierarchical clustering and partitioning clustering in a single line function call. Compared to the standard k-means clustering method which requires the user to specify the optimal number of clusters, the function eclust() computes automatically the *gap statistic* for estimating the right number of clusters.

Similar to K-Means ,the enhanced K-means also requires numerical data and hence converted the categorical data into numeric with one-hot encoding. Enhanced K-means model was run with nstart = 25 and bootstrapping value of 5, indicates R will run 25 different random starting assignments and selects the lowest within cluster variation and with different center values and then the purity of the clusters was calculated.

K-Mode Clustering:

In the previous k-means clustering model and Enhanced K-means ,the algorithm uses numerical distances and our dataset consists of categorical attributes which got converted to numerical using one-hot encoding.so it could have considered close two really distant objects that merely have been assigned two close numbers and hence the low purity and clustering.

k-modes is an extension of k-means. Instead of distances it uses dissimilarities (that is, quantification of the total mismatches between two objects: the smaller this number, the more similar the two objects). And instead of means, it uses modes. A mode is a vector of elements that minimizes the dis-similarities between the vector itself and each object of the data. We will have as many modes as the number of clusters we required, since they act as centroids.

ROCK Clustering:

This model belongs to the class of agglomerative hierarchical clustering algorithms. The steps involved in clustering using ROCK are described in the following figure. After drawing a random sample from the database, a hierarchical clustering algorithm that employs links is applied to the sampled points. Finally, the clusters involving only the sampled points are used to assign the remaining data points on disk to the appropriate clusters.

The categorical dataset is convert to binary values using as.dummy function and ran the model with n=5 and theta=0.8 to understand the tightness of the clusters.

Decision Trees Model:

Decision Trees are a type of supervised learning used for classification problems. The first step was induction, where the machine was trained on labeled data using the C4.5 algorithm. The second step was deduction, where the model was tested against unlabeled samples that were held out of the training data set. Three algorithms can be used to measure information gain with Entropy: "InfGain," "Gini," and "GainRatio."

```
           CP nsplit   rel error      xerror         xstd
1 0.970072993      0 1.000000000 1.000000000 0.0137522314
2 0.016788321      1 0.029927007 0.029927007 0.0032809723
3 0.006204380      2 0.013138686 0.013138686 0.0021828391
4 0.002372263      3 0.006934307 0.006934307 0.0015881793
5 0.000000000      5 0.002189781 0.002189781 0.0008935026
```

```
Variable importance
odor   spore.print.color   gill.color   stalk.surface.above.ring   ring.type   stalk.color.below.ring
  26                  19           16                         14          14                       11
```

Random Forest Trees Model:

Random Forest is a type of ensemble learning, where a set of classifiers is constructed, then a committee is formed and a vote is taken to select the best classifier. In this case, randomized sets of Decision Trees are created, and one tree is selected per subset. This creates a "forest", which can then be combined into one best tree. Note that if it is determined that a particular variable is not contributing significantly to information gain, the user could choose to remove that variable in order to reduce the dimensionality of the data.

The training data set was passed to the model without the class variable at the same time calculating the variable importance.

Naïve Bayes Model:

Naïve Bayes is a type of supervised learning used for classification problems. This model works by calculating probabilities of events and assumes that each event is independent.

This model was used to classify each mushroom in the unlabeled test set as edible or poisonous, and the results were then checked against the labels. The model showed a 95% accuracy and gained the most information from the variable: odor. There are several visualizations in the results section showing which variable values were strong predictors for either edible or poisonous.

Support Vector Machine Model:

SVM is a type of supervised learning. It can solve both linear separable and non-linear separable classification problems. The goal is to find the widest margin between classes. The training examples closer to the margin are the support vectors, while the ones further from the margin are non-support vectors, and are not part of the model. Within the data set, each row is a vector, and the number of columns is the dimensionality. An SVM can use different kernels, such as polynomial, linear, radial and sigmoid.

Ran the model with linear, polynomial and radial kernel and tuned the model to get the best cost and Gamma values using 10-fold cross validation sampling method. TrainControl was used to Control the computational nuances of the train function and repeatedCv was used as the resampling method.

The cost, a penalty for mislabeling data, can be set to a higher or lower value. Additionally, the model can be scaled in R if it has not already been normalized. The strengths are a high tolerance to noise data, probabilistic prediction, flexibility in data representations (the ability to use discrete or continuous data), and scalability. The weaknesses are that each kernel type requires a number of parameters, and interpretability of non-linear kernels is a challenge.

The parameters of the three SVM models are shown below.

```
> print(svm_model3);

Call:
svm(formula = class ~ ., data = mush_train, type = "C-classification", kernel = "linear",
    cost = 1, gamma = 0.5)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  1
      gamma:  0.5

Number of Support Vectors:  176

> print(svm_model2);

Call:
svm(formula = class ~ ., data = mush_train, type = "C-classification", kernel = "polynomial",
    cost = 1, gamma = 0.5)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  polynomial
       cost:  1
     degree:  3
      gamma:  0.5
     coef.0:  0

Number of Support Vectors:  474

> print(svm_model1);

Call:
svm(formula = class ~ ., data = mush_train, type = "C-classification", kernel = "radial",
    cost = 1, gamma = 0.5)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.5

Number of Support Vectors:  5234
```

Text Mining Model:

Text mining is utilized by gathering text documents into a corpus, tokenizing the text, and counting word frequencies. To adjust for different document lengths, the word frequencies can be normalized.

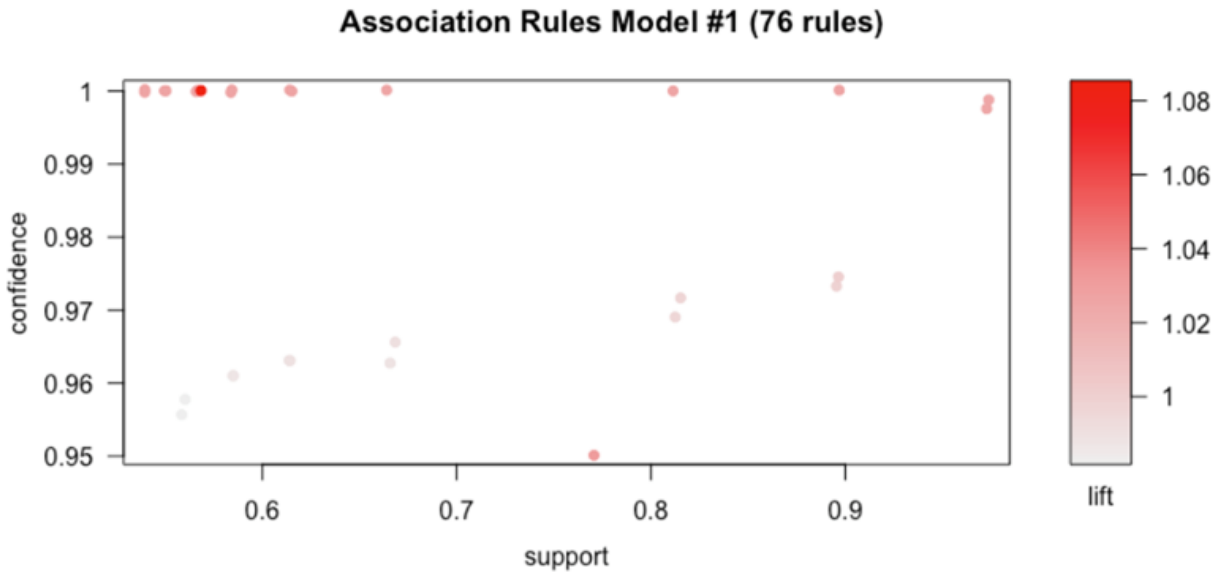
Two word-clouds were created from two different documents (written almost 100 years apart) and the frequencies of the top words were compared.

Results

Association Rule Mining Results:

ARM #1 Results:

- Support was set at 0.50 and confidence was set at 0.95. This produced 76 rules.
- The top ten rules for confidence all had a perfect confidence of 1.
- The top ten rules for support ranged from 0.667 to 0.973.
- The top ten rules for lift ranged from 1.03 to 1.08, which is quite low.



ARM ARM #1 Top 10 rules, sorted by Confidence:

	lhs		rhs	support	confidence	lift
[1]	{stalk.color.below.ring=White}	=>	{gill.attachment=Free}	0.540	1	1.03
[2]	{stalk.color.below.ring=White}	=>	{veil.color=White}	0.540	1	1.03
[3]	{stalk.color.above.ring=White}	=>	{gill.attachment=Free}	0.549	1	1.03
[4]	{stalk.color.above.ring=White}	=>	{veil.color=White}	0.549	1	1.03
[5]	{stalk.shape=Tapering}	=>	{ring.number=One}	0.567	1	1.08
[6]	{stalk.shape=Tapering}	=>	{gill.attachment=Free}	0.567	1	1.03
[7]	{stalk.shape=Tapering}	=>	{veil.color=White}	0.567	1	1.03
[8]	{gill.attachment=Free,stalk.surface.below.ring=Smooth}	=>	{veil.color=White}	0.584	1	1.03
[9]	{stalk.surface.below.ring=Smooth,veil.color=White}	=>	{gill.attachment=Free}	0.584	1	1.03
[10]	{gill.attachment=Free,stalk.surface.above.ring=Smooth}	=>	{veil.color=White}	0.613	1	1.03

ARM #1 Top 10 rules, sorted by Support:

	lhs		rhs	support	confidence	lift
[1]	{gill.attachment=Free}	=>	{veil.color=White}	0.973	0.999	1.024
[2]	{veil.color=White}	=>	{gill.attachment=Free}	0.973	0.998	1.024
[3]	{ring.number=One}	=>	{gill.attachment=Free}	0.898	0.974	1.000
[4]	{ring.number=One}	=>	{veil.color=White}	0.897	0.973	0.998
[5]	{veil.color=White,ring.number=One}	=>	{gill.attachment=Free}	0.897	1.000	1.027
[6]	{gill.spacing=Close}	=>	{veil.color=White}	0.815	0.972	0.996
[7]	{gill.spacing=Close}	=>	{gill.attachment=Free}	0.813	0.969	0.995
[8]	{gill.attachment=Free,gill.spacing=Close}	=>	{veil.color=White}	0.813	1.000	1.025
[9]	{gill.attachment=Free,gill.spacing=Close}	=>	{ring.number=One}	0.772	0.950	1.031
[10]	{gill.size=Broad}	=>	{veil.color=White}	0.667	0.966	0.990

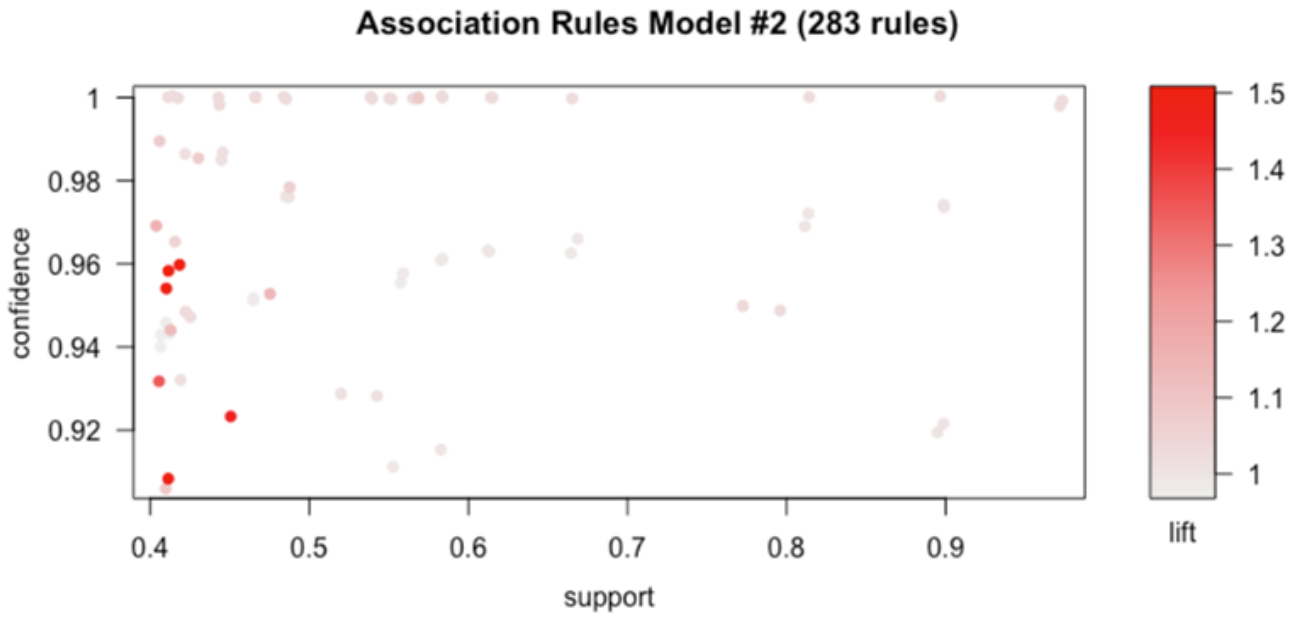
ARM #1 Top 10 rules, sorted by Lift:

	lhs		rhs	support	confidence	lift
[1]	{stalk.shape=Tapering}	=>	{ring.number=One}	0.567	1.00	1.08
[2]	{gill.attachment=Free,gill.spacing=Close}	=>	{ring.number=One}	0.772	0.95	1.03
[3]	{stalk.color.below.ring=White}	=>	{gill.attachment=Free}	0.540	1.00	1.03
[4]	{stalk.color.above.ring=White}	=>	{gill.attachment=Free}	0.549	1.00	1.03
[5]	{stalk.shape=Tapering}	=>	{gill.attachment=Free}	0.567	1.00	1.03
[6]	{stalk.surface.below.ring=Smooth,veil.color=White}	=>	{gill.attachment=Free}	0.584	1.00	1.03
[7]	{stalk.surface.above.ring=Smooth,veil.color=White}	=>	{gill.attachment=Free}	0.613	1.00	1.03
[8]	{veil.color=White,ring.number=One}	=>	{gill.attachment=Free}	0.897	1.00	1.03
[9]	{stalk.color.below.ring=White}	=>	{veil.color=White}	0.540	1.00	1.03
[10]	{stalk.color.above.ring=White}	=>	{veil.color=White}	0.549	1.00	1.03

ARM #2 Results:

With the hope of producing rules with higher lift, support was reduced to 0.40 and confidence was reduced to 0.90. This produced 283 rules.

- The top ten rules for confidence all had a perfect confidence of 1.
- The top ten rules for support ranged from 0.813 to 0.973.
- The top ten rules for lift ranged from 1.08 to 1.51, which is much better than in Model 1.



ARM #2 Top 10 rules, sorted by Confidence:

	lhs		rhs	support	confidence	lift
[1]	{bruises=Bruises}	=>	{gill.attachment=Free}	0.416	1	1.03
[2]	{bruises=Bruises}	=>	{veil.color=White}	0.416	1	1.03
[3]	{stalk.color.below.ring=White}	=>	{gill.attachment=Free}	0.540	1	1.03
[4]	{stalk.color.below.ring=White}	=>	{veil.color=White}	0.540	1	1.03
[5]	{stalk.color.above.ring=White}	=>	{gill.attachment=Free}	0.549	1	1.03
[6]	{stalk.color.above.ring=White}	=>	{veil.color=White}	0.549	1	1.03
[7]	{stalk.shape=Tapering}	=>	{ring.number=One}	0.567	1	1.08
[8]	{stalk.shape=Tapering}	=>	{gill.attachment=Free}	0.567	1	1.03
[9]	{stalk.shape=Tapering}	=>	{veil.color=White}	0.567	1	1.03
[10]	{odor=None,veil.color=White}	=>	{gill.attachment=Free}	0.410	1	1.03

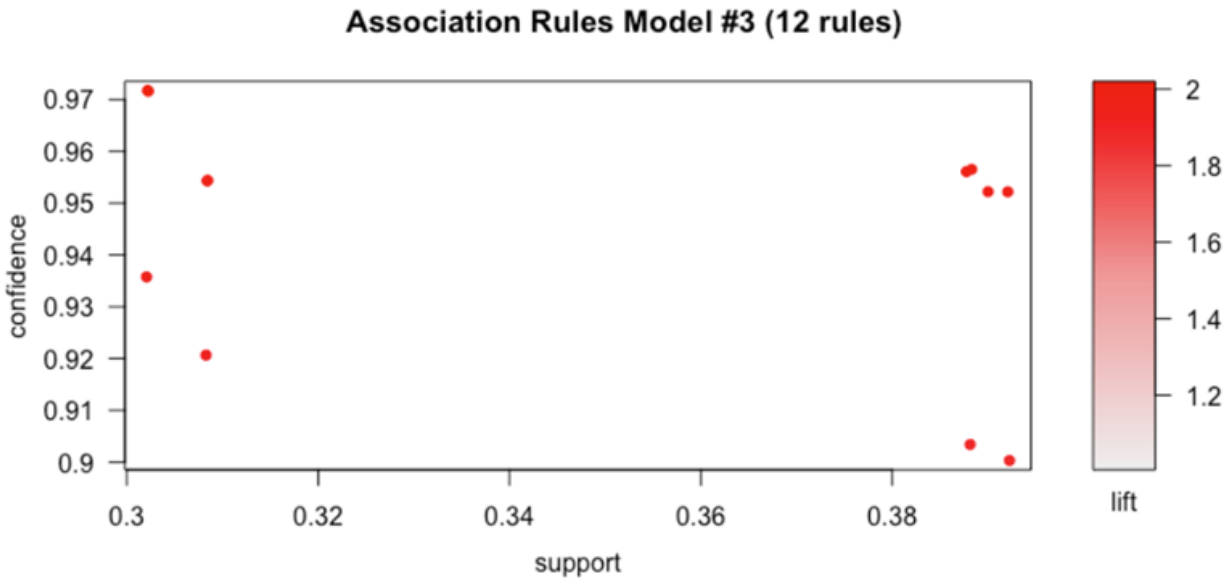
ARM #2 Top 10 rules, sorted by Support:

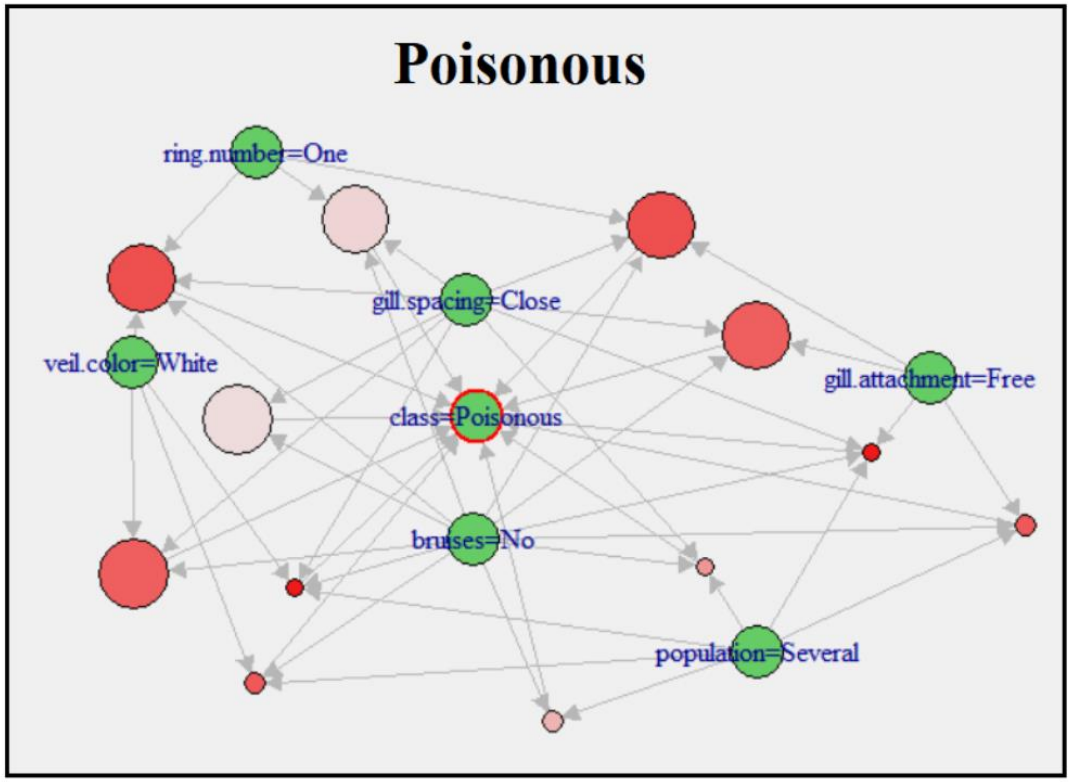
	lhs		rhs	support	confidence	lift
[1]	{gill.attachment=Free}	=>	{veil.color=White}	0.973	0.999	1.024
[2]	{veil.color=White}	=>	{gill.attachment=Free}	0.973	0.998	1.024
[3]	{ring.number=One}	=>	{gill.attachment=Free}	0.898	0.974	1.000
[4]	{gill.attachment=Free}	=>	{ring.number=One}	0.898	0.922	1.000
[5]	{ring.number=One}	=>	{veil.color=White}	0.897	0.973	0.998
[6]	{veil.color=White}	=>	{ring.number=One}	0.897	0.920	0.998
[7]	{veil.color=White,ring.number=One}	=>	{gill.attachment=Free}	0.897	1.000	1.027
[8]	{gill.spacing=Close}	=>	{veil.color=White}	0.815	0.972	0.996
[9]	{gill.spacing=Close}	=>	{gill.attachment=Free}	0.813	0.969	0.995
[10]	{gill.attachment=Free,gill.spacing=Close}	=>	{veil.color=White}	0.813	1.000	1.025

ARM #2 Top 10 rules, sorted by Lift:

	lhs		rhs	support	confidence	lift
[1]	{ring.number=One,ring.type=Pendant}	=>	{stalk.surface.above.ring=Smooth}	0.420	0.960	1.51
[2]	{stalk.surface.below.ring=Smooth,ring.type=Pendant}	=>	{stalk.surface.above.ring=Smooth}	0.410	0.959	1.50
[3]	{gill.spacing=Close,ring.type=Pendant}	=>	{stalk.surface.above.ring=Smooth}	0.409	0.954	1.50
[4]	{stalk.surface.above.ring=Smooth,ring.type=Pendant}	=>	{stalk.surface.below.ring=Smooth}	0.410	0.908	1.49
[5]	{ring.type=Pendant}	=>	{stalk.surface.above.ring=Smooth}	0.451	0.923	1.45
[6]	{odor=None}	=>	{gill.size=Broad}	0.405	0.932	1.35
[7]	{bruises=Bruises}	=>	{gill.spacing=Close}	0.403	0.969	1.16
[8]	{population=Several}	=>	{gill.spacing=Close}	0.474	0.952	1.14
[9]	{ring.number=One,ring.type=Pendant}	=>	{gill.spacing=Close}	0.414	0.944	1.13
[10]	{stalk.shape=Tapering}	=>	{ring.number=One}	0.567	1.000	1.08

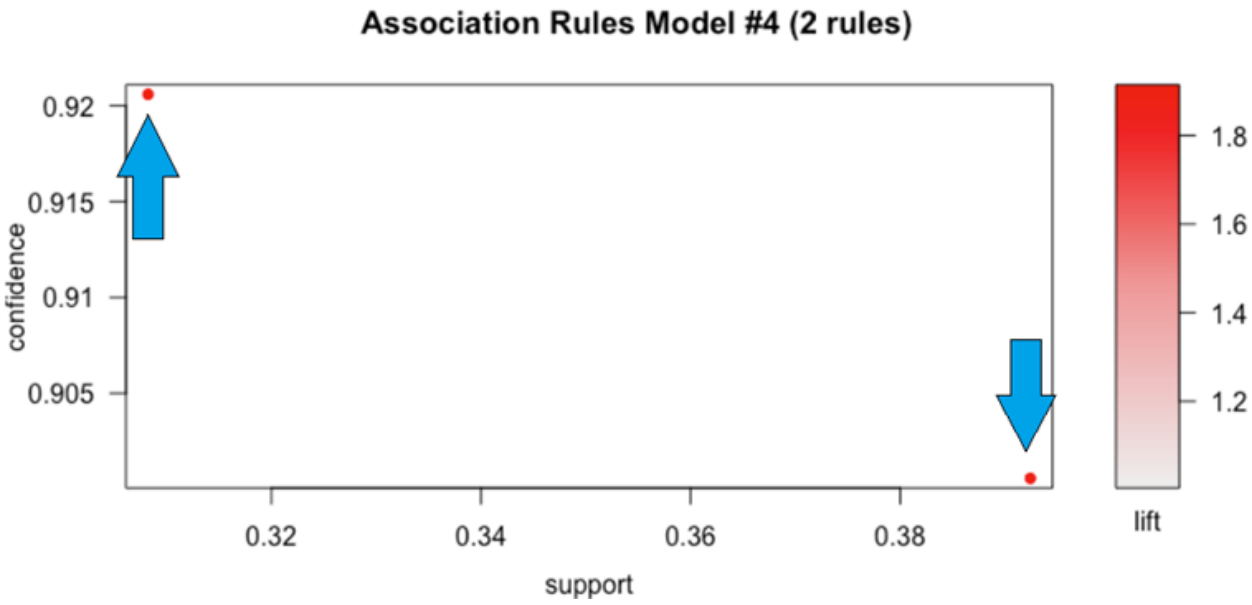
ARM #3 Results: The RHS was set to class = poisonous, support was set at 0.30, and confidence was set at 0.90. This produced 12 rules, but they were very complicated. The interactive plot shows that poisonous mushrooms are very closely associated with closely spaced gills and lack of bruises.





ARM #4 Results:

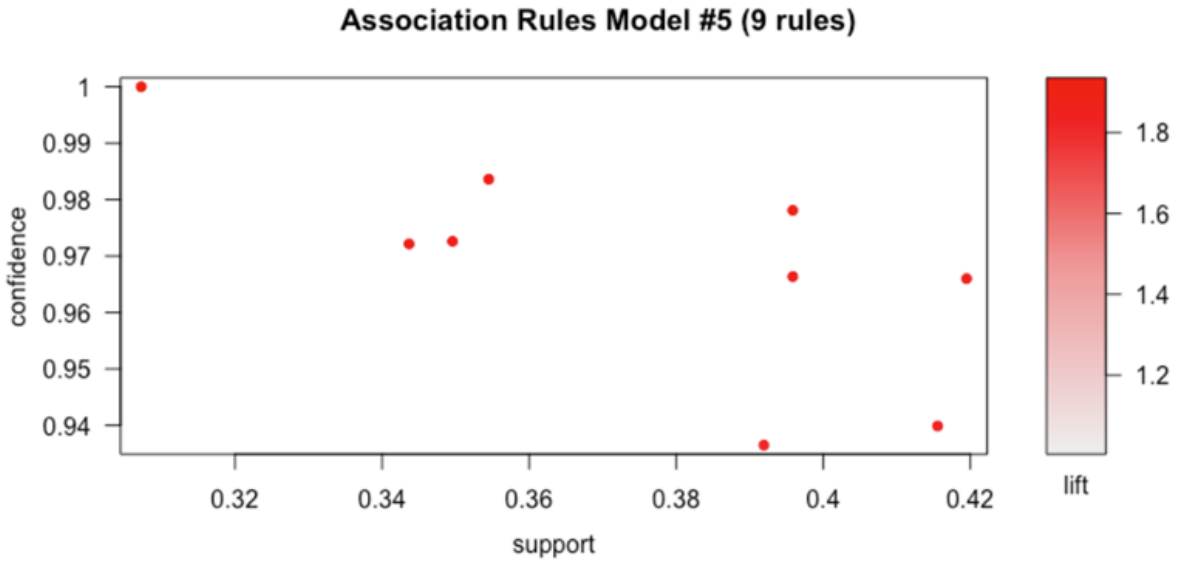
In an effort to produce simpler rules, the extra restraint of maximum length = 3 was added to Model 3, and this was run as Model 4. This produced 2 rules.




```
    lhs                                  rhs                 support confidence lift
[1] {bruises=No,population=Several} => {class=Poisonous} 0.308   0.921      1.91
[2] {bruises=No,gill.spacing=Close} => {class=Poisonous} 0.392   0.901      1.87
```

ARM #5 Results:

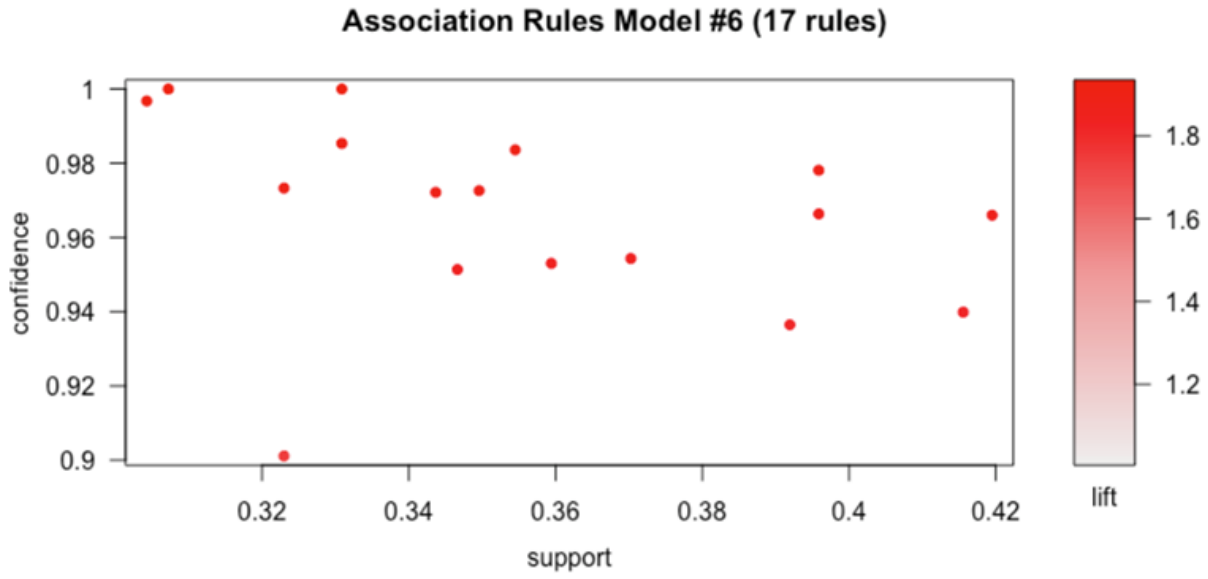
The RHS was set to class = edible, with a support of 0.03, confidence of 0.90, and maximum rule length of 3. This produced 9 rules with strong confidence and strong lift.

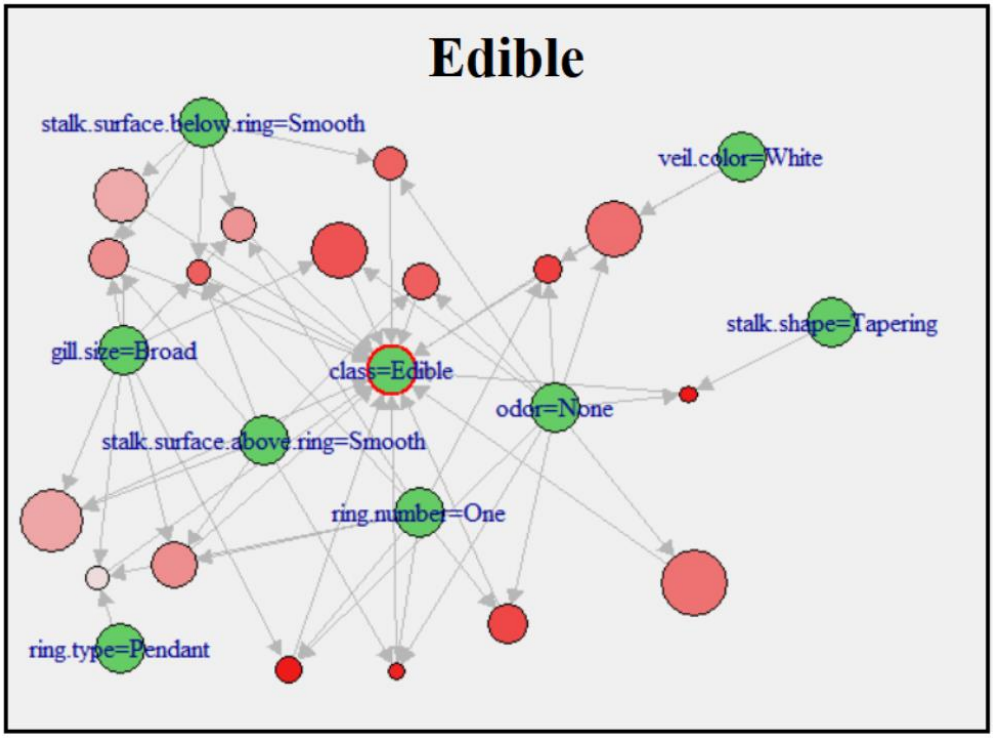


lhs		rhs	support	confidence	lift
{odor=None}	=>	{class=Edible}	0.419	0.966	1.86
{odor=None,stalk.shape=Tapering}	=>	{class=Edible}	0.307	1.000	1.93
{odor=None,stalk.surface.below.ring=Smooth}	=>	{class=Edible}	0.344	0.972	1.88
{odor=None,stalk.surface.above.ring=Smooth}	=>	{class=Edible}	0.350	0.973	1.88
{odor=None,gill.size=Broad}	=>	{class=Edible}	0.396	0.978	1.89
{odor=None,ring.number=One}	=>	{class=Edible}	0.355	0.984	1.90
{odor=None,veil.color=White}	=>	{class=Edible}	0.396	0.966	1.87
{gill.size=Broad,stalk.surface.below.ring=Smooth}	=>	{class=Edible}	0.392	0.936	1.81
{gill.size=Broad,stalk.surface.above.ring=Smooth}	=>	{class=Edible}	0.416	0.940	1.81

ARM #6 Results:

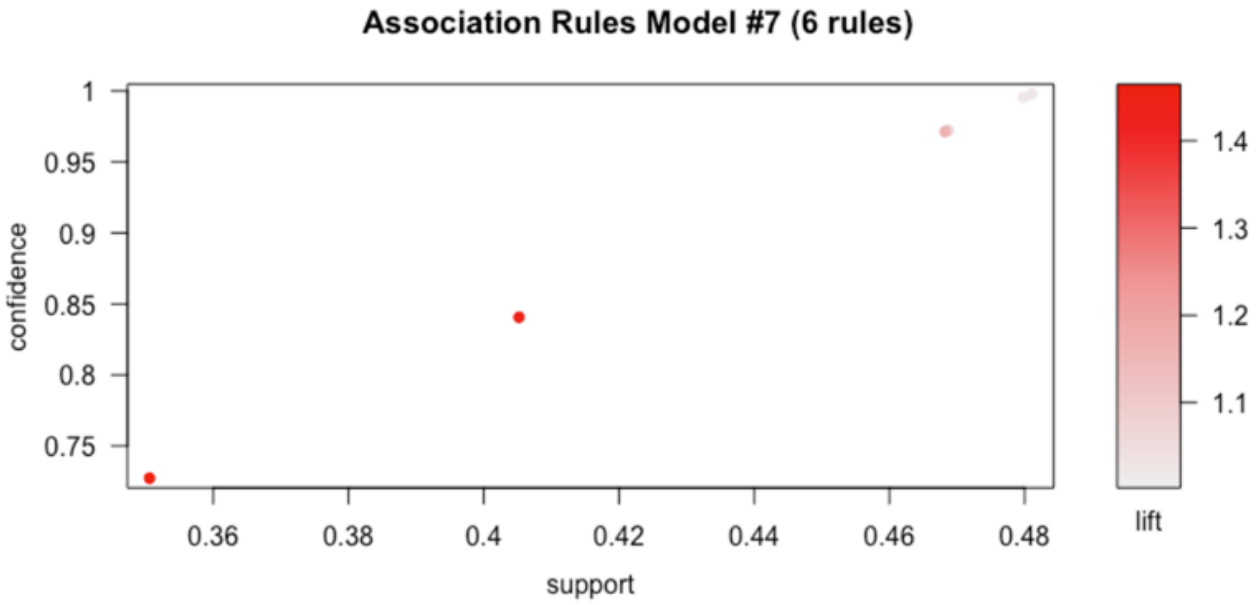
The maximum rule length of 3 was removed from Model 5, in order to produce more rules. This produced 17 rules. The interactive plot shows that no odor, one ring number, and smooth stalk surface above the ring are closely associated with being edible.





ARM #7 Results:

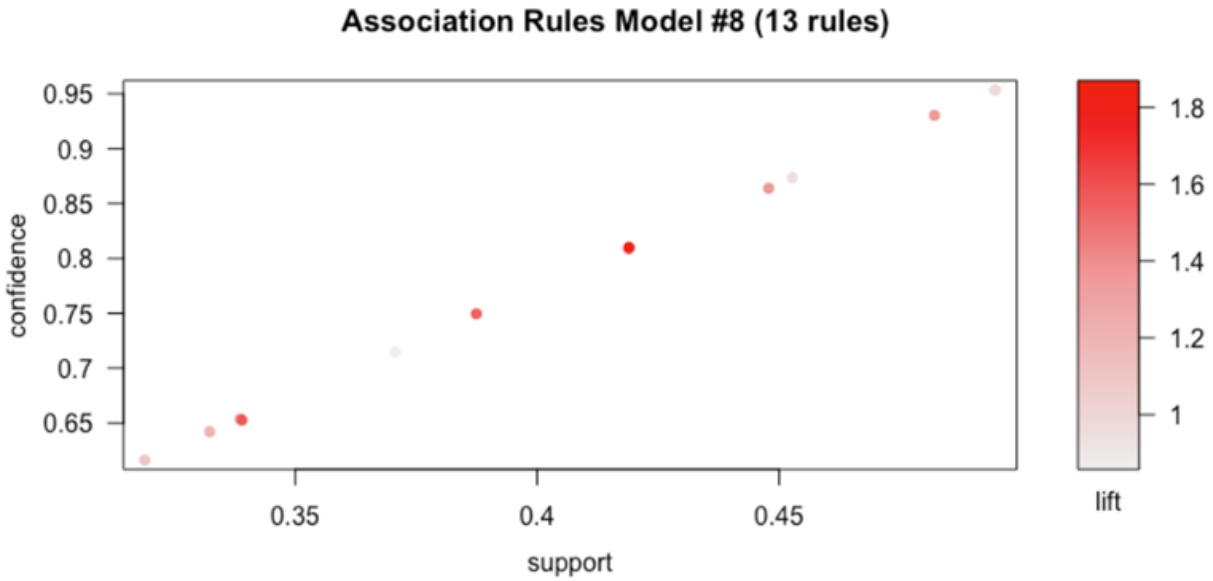
The LHS was set to class = poisonous, support was set at 0.01, and confidence was set at 0.60. This produced 6 rules.



	lhs		rhs	support	confidence	lift
[1]	{class=Poisonous}	=>	{veil.color=White}	0.481	0.998	1.02
[2]	{class=Poisonous}	=>	{gill.attachment=Free}	0.480	0.995	1.02
[3]	{class=Poisonous}	=>	{ring.number=One}	0.469	0.972	1.06
[4]	{class=Poisonous}	=>	{gill.spacing=Close}	0.468	0.971	1.16
[5]	{class=Poisonous}	=>	{bruises=No}	0.405	0.841	1.44
[6]	{class=Poisonous}	=>	{population=Several}	0.351	0.727	1.46

ARM #8 Results:

The LHS was set to class = edible, support was set at 0.01, and confidence was set at 0.60. This produced 13 rules.



ARM #8 Top 10 rules, sorted by Confidence:

	lhs		rhs	support	confidence	lift
[1]	{class=Edible}	=>	{gill.attachment=Free}	0.494	0.954	0.980
[2]	{class=Edible}	=>	{veil.color=White}	0.494	0.954	0.978
[3]	{class=Edible}	=>	{gill.size=Broad}	0.483	0.932	1.349
[4]	{class=Edible}	=>	{ring.number=One}	0.453	0.875	0.949
[5]	{class=Edible}	=>	{stalk.surface.above.ring=Smooth}	0.448	0.865	1.358
[6]	{class=Edible}	=>	{odor=None}	0.419	0.810	1.865
[7]	{class=Edible}	=>	{stalk.surface.below.ring=Smooth}	0.419	0.808	1.330
[8]	{class=Edible}	=>	{ring.type=Pendant}	0.388	0.749	1.534
[9]	{class=Edible}	=>	{gill.spacing=Close}	0.370	0.715	0.853
[10]	{class=Edible}	=>	{bruises=Bruises}	0.339	0.654	1.574

ARM #8 Top 10 rules, sorted by Support:

	lhs		rhs	support	confidence	lift
[1]	{class=Edible}	=>	{gill.attachment=Free}	0.494	0.954	0.980
[2]	{class=Edible}	=>	{veil.color=White}	0.494	0.954	0.978
[3]	{class=Edible}	=>	{gill.size=Broad}	0.483	0.932	1.349
[4]	{class=Edible}	=>	{ring.number=One}	0.453	0.875	0.949
[5]	{class=Edible}	=>	{stalk.surface.above.ring=Smooth}	0.448	0.865	1.358
[6]	{class=Edible}	=>	{odor=None}	0.419	0.810	1.865
[7]	{class=Edible}	=>	{stalk.surface.below.ring=Smooth}	0.419	0.808	1.330
[8]	{class=Edible}	=>	{ring.type=Pendant}	0.388	0.749	1.534
[9]	{class=Edible}	=>	{gill.spacing=Close}	0.370	0.715	0.853
[10]	{class=Edible}	=>	{bruises=Bruises}	0.339	0.654	1.574

ARM #8 Top 10 rules, sorted by Lift:

	lhs		rhs	support	confidence	lift
[1]	{class=Edible}	=>	{odor=None}	0.419	0.810	1.86
[2]	{class=Edible}	=>	{bruises=Bruises}	0.339	0.654	1.57
[3]	{class=Edible}	=>	{ring.type=Pendant}	0.388	0.749	1.53
[4]	{class=Edible}	=>	{stalk.surface.above.ring=Smooth}	0.448	0.865	1.36
[5]	{class=Edible}	=>	{gill.size=Broad}	0.483	0.932	1.35
[6]	{class=Edible}	=>	{stalk.surface.below.ring=Smooth}	0.419	0.808	1.33
[7]	{class=Edible}	=>	{stalk.color.below.ring=White}	0.333	0.643	1.19
[8]	{class=Edible}	=>	{stalk.color.above.ring=White}	0.339	0.654	1.19
[9]	{class=Edible}	=>	{stalk.shape=Tapering}	0.319	0.616	1.09
[10]	{class=Edible}	=>	{gill.attachment=Free}	0.494	0.954	0.98

<u>ARM Discussion:</u>

Setting a high confidence (0.95) gave rules with very low lift (1.08). Lowering confidence to 0.90 generated rules with higher support and lift. Models with a LHS or RHS set generated fewer rules with high lift. Restricting the rule length also generated fewer rules.

Poisonous mushrooms are associated with not bruising, and edible mushrooms are associated with no odor.

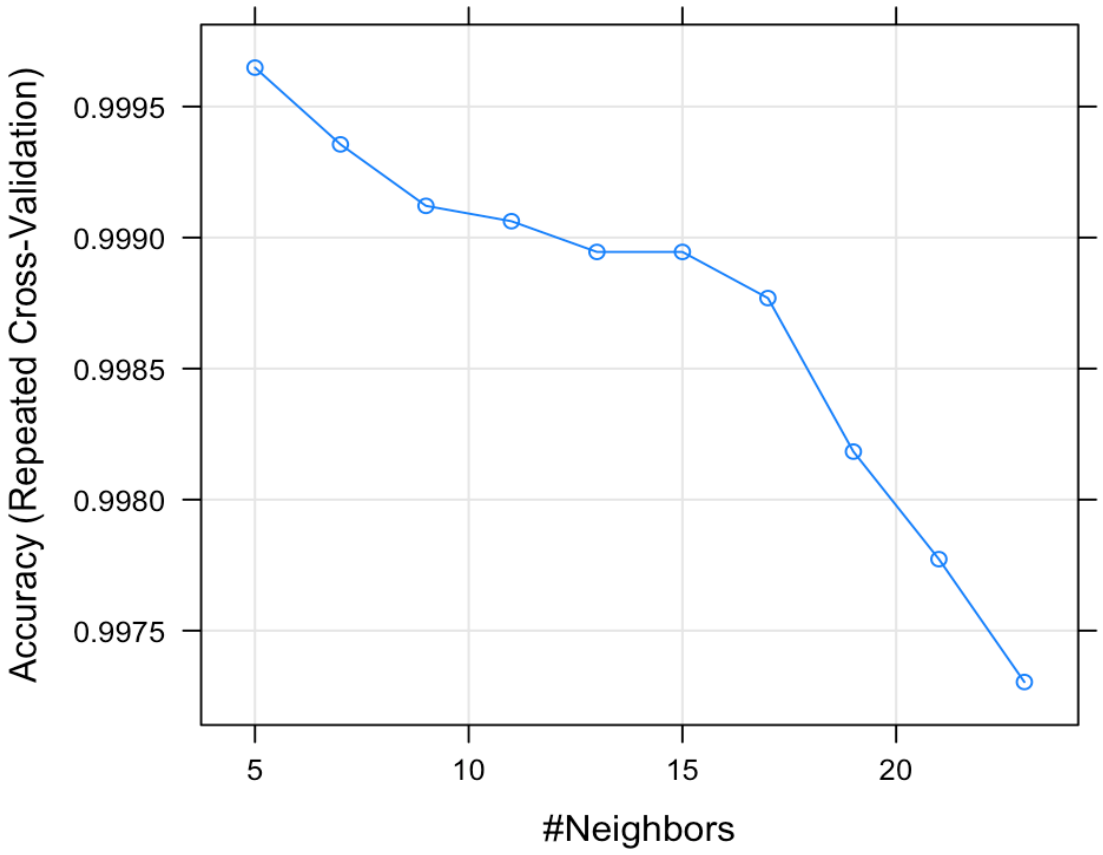
lhs		rhs	support	confidence	lift
{class=Poisonous}	=>	{bruises=No}	0.405	0.841	1.44
{class=Edible}	=>	{odor=None}	0.419	0.810	1.86

Clustering Results:

K-Nearest Neighbour Results:

Knn Model's Accuracy was used to select the optimal K-Value..Based on 10-fold Cross-Validation, a final value of 5 was selected as K and had a 99.9% accuracy in predicting edible vs poisonous mushrooms.

Plotting yields Number of Neighbours vs accuracy (based on repeated cross validation).



```
k-Nearest Neighbors

5687 samples
  21 predictor
   2 classes: 'Edible', 'Poisonous'

Pre-processing: centered (95), scaled (95)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 5118, 5119, 5119, 5119, 5118, 5119, ...
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   5  0.9994134  0.9988250
   7  0.9992960  0.9985899
   9  0.9992960  0.9985899
  11  0.9992960  0.9985899
  13  0.9992960  0.9985899
  15  0.9992960  0.9985899
  17  0.9991788  0.9983552
  19  0.9988856  0.9977678
  21  0.9981825  0.9963595
  23  0.9978894  0.9957722

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 5.
```

kNN Confusion Matrix results:

```
Confusion Matrix and Statistics

          Reference
Prediction  Edible Poisonous
  Edible      1264         2
  Poisonous      0      1171

               Accuracy : 0.9992
                 95% CI : (0.997, 0.9999)
    No Information Rate : 0.5187
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9984
 Mcnemar's Test P-Value : 0.4795

            Sensitivity : 1.0000
            Specificity : 0.9983
         Pos Pred Value : 0.9984
         Neg Pred Value : 1.0000
             Prevalence : 0.5187
         Detection Rate : 0.5187
   Detection Prevalence : 0.5195
      Balanced Accuracy : 0.9991

       'Positive' Class : Edible
```

K-Means Clustering Results:

K-means searches for the **minimum *sum of squares* assignment**, i.e. it minimizes unnormalized variance (=total_SS) by assigning points to cluster centers.

Cluster with Centroid value k=5

```
Within cluster sum of squares by cluster:
[1]  9814.935  4860.000  5401.455  7821.425 17844.026
 (between_SS / total_SS =  38.2 %)
```

Cluster with Centroid value k=4

```
Within cluster sum of squares by cluster:
[1]  4860.000 26679.507  9793.309  7816.111
 (between_SS / total_SS =  33.6 %)
```

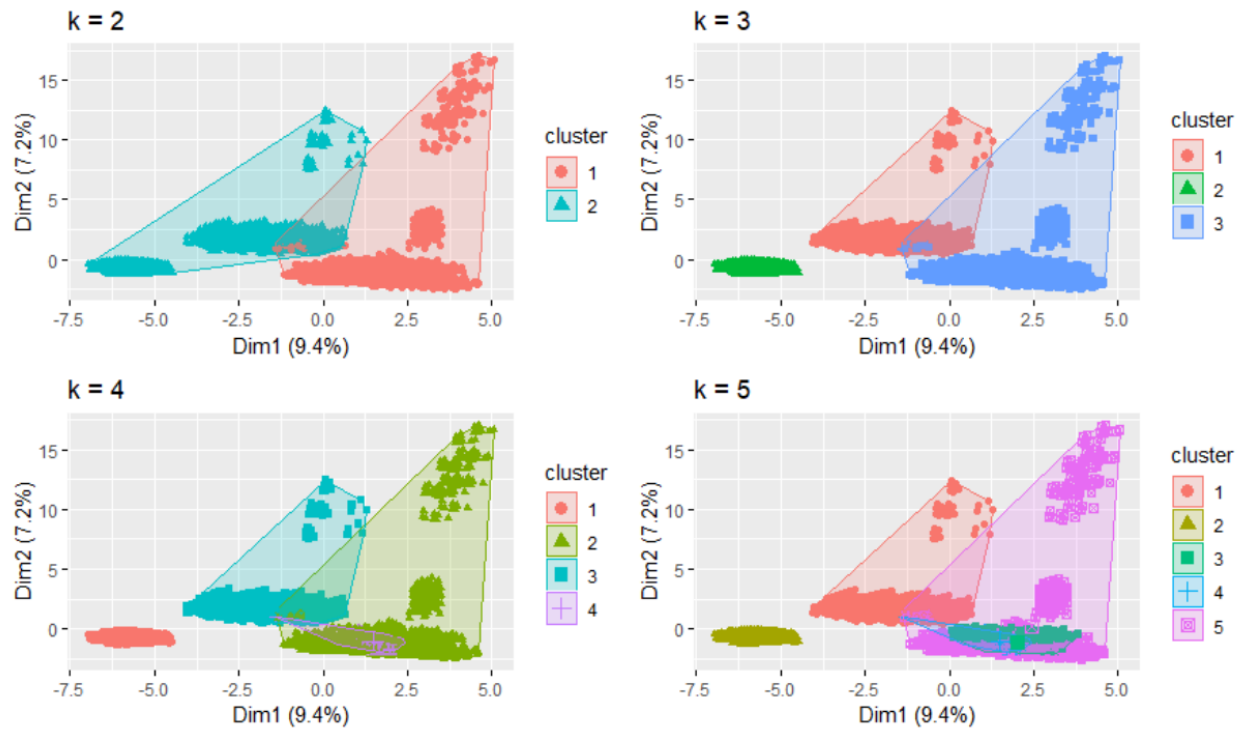
Cluster with Centroid value k=3

```
Within cluster sum of squares by cluster:
[1]  9741.063  4860.000 40054.921
 (between_SS / total_SS =  26.2 %)
```

Cluster with Centroid value k=2

```
Within cluster sum of squares by cluster:
[1] 40165.65 22522.34
 (between_SS / total_SS =  15.4 %)
```

Clusters shown with different K values:



Purity of clustering is a simple measure of the accuracy, which is between 0 and 1. 0 indicates poor clustering, and 1 indicates perfect clustering.

Purity of cluster with Centroids:

```
#Purity of Cluster with K=2 0.8953717
#Purity of Cluster with K=3 0.8945101
#Purity of Cluster with K=4 0.8938946
#Purity of Cluster with K=5 0.8936484
```

Based on the above results K-means clustering doesn't seem to be a good clustering model on this dataset.

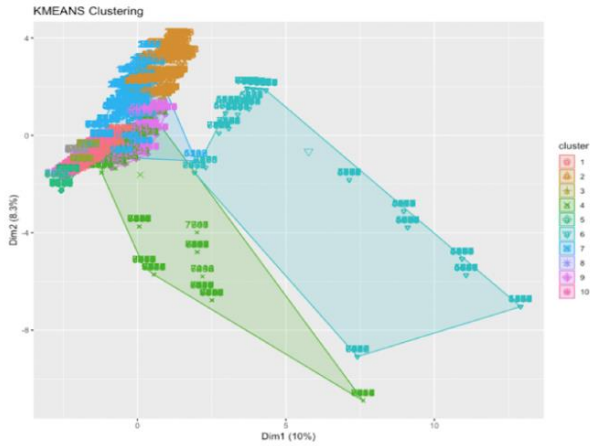
Enhanced K-Means Clustering Results:

Optimal number of clusters was determined as 10 by enhanced k-means. The function fviz_gap_stat() [in factoextra] is used to visualize the gap statistic plot. The average silhouette width measure to 0.17 and shows that the objects are poorly clustered. The silhouette measures (SiSi) how similar an object ii is to the the other objects in its own cluster versus those in the neighbor cluster. SiSi values range from 1 to - 1: **A value of SiSi close to 1 indicates that the object is well clustered. In the other words, the object ii is similar to the other objects in its group.** A value of SiSi close to -1 indicates that the object is poorly clustered, and that assignment to some other cluster would probably improve the overall results.

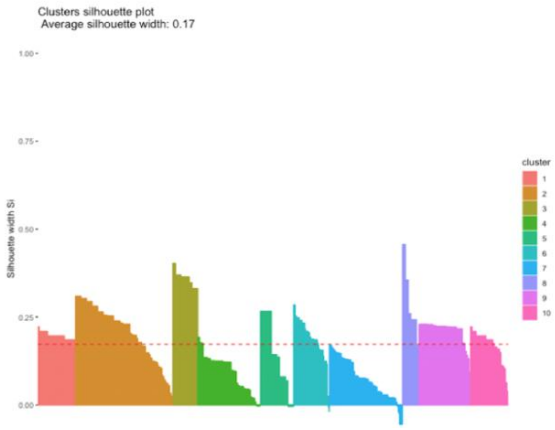
This shows that both k-means and Enhanced K-means clustering doesn't seems to be good clustering modelling on mushroom dataset.

	cluster	size	ave.sil.width
1	1	648	0.20
2	2	1680	0.21
3	3	432	0.36
4	4	1082	0.09
5	5	576	0.14
6	6	618	0.19
7	7	1264	0.07
8	8	288	0.29
9	9	888	0.21
10	10	648	0.17

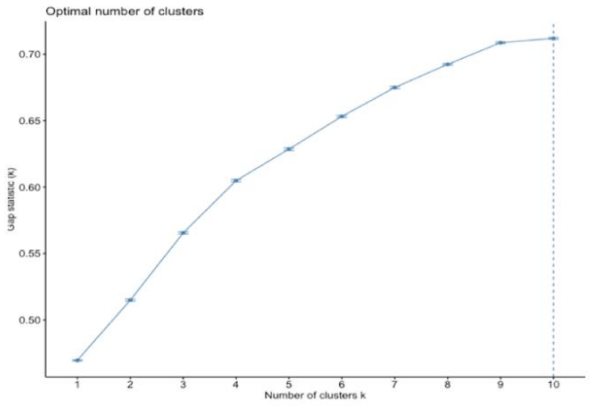
K-Means Clustering:



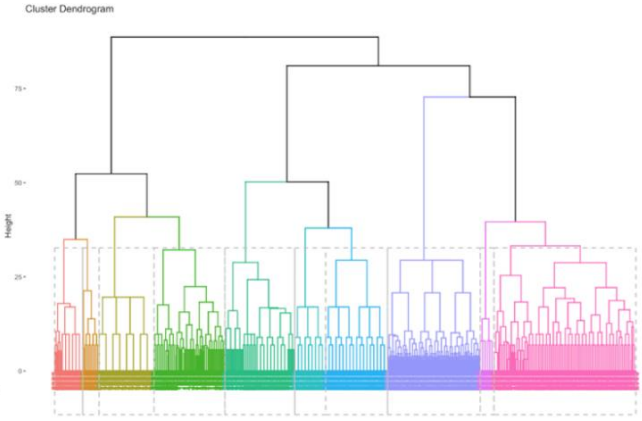
Clusters Silhouette Plot:



Optimal Number of Clusters:

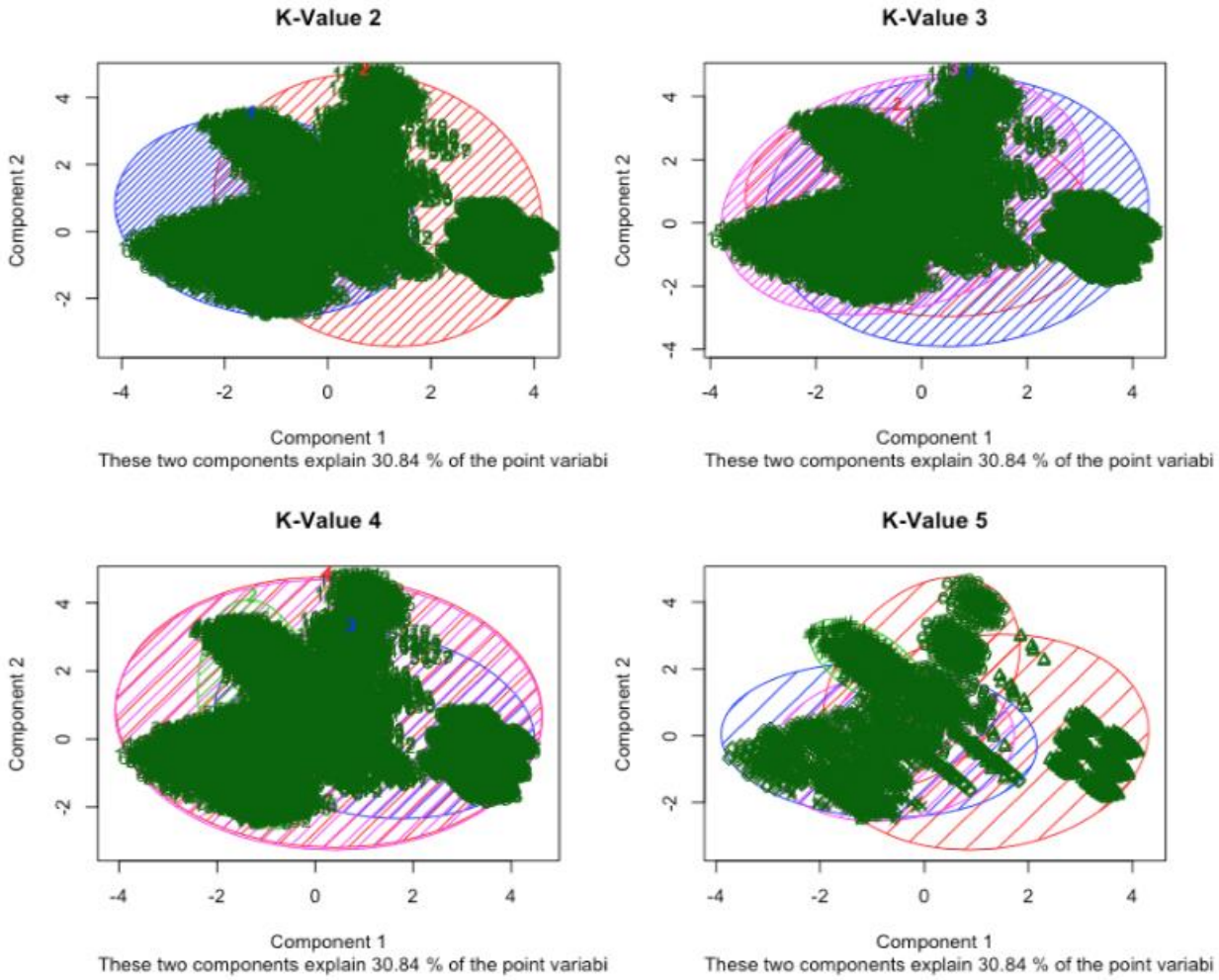


Cluster Dendrogram:



K-Mode Clustering Results:

K-mode clustering also gave a much lesser purity on the cluster and doesn't seems to a good clustering model fit.

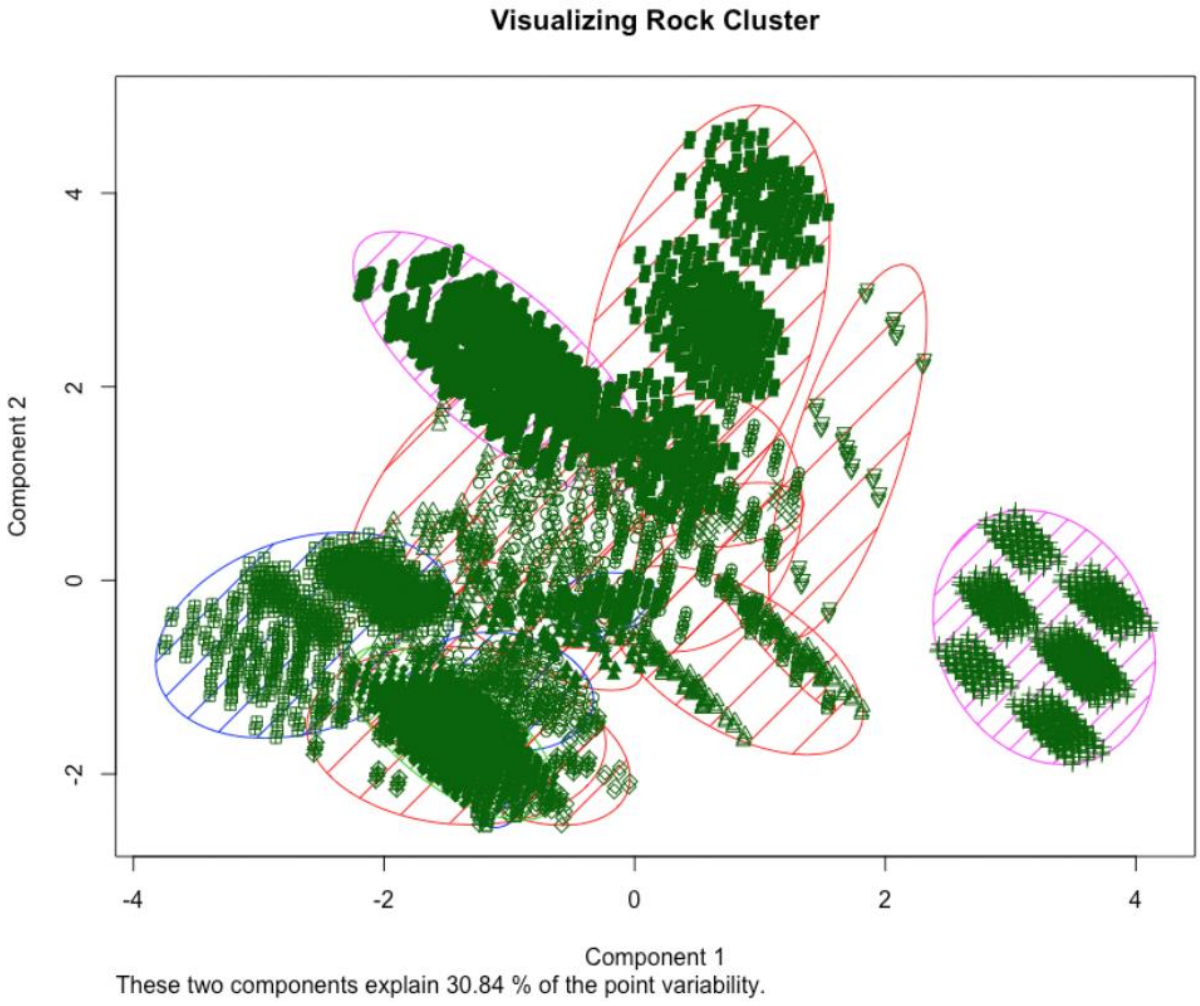


Purity of K-Mode Clusters:

```
#Purity of K-Mode Cluster with K=2 0.888971
#Purity of K-Mode Cluster with K=3 0.7548006
#Purity of K-Mode Cluster with K=4 0.8857706
#Purity of K-Mode Cluster with K=5 0.8872477
```

ROCK Clustering Results:

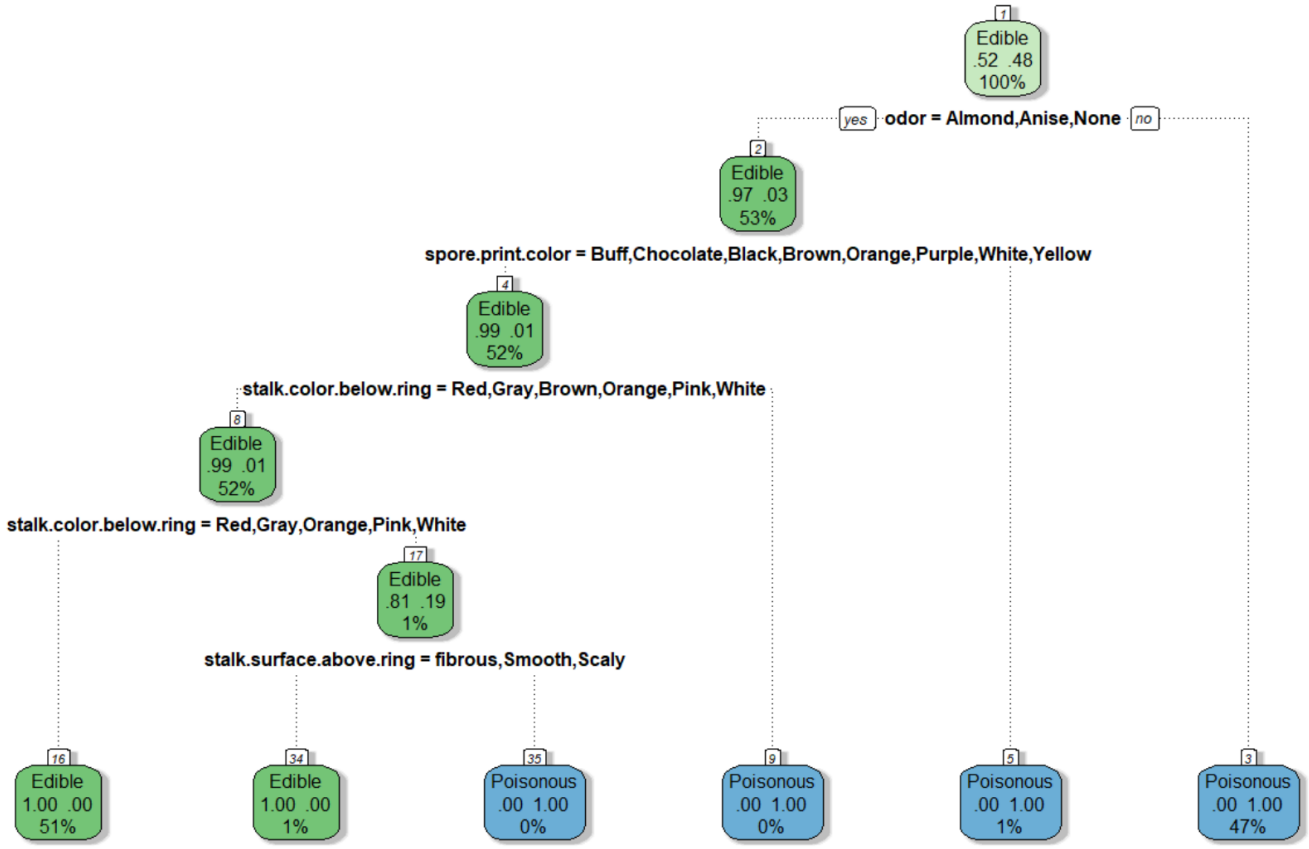
The rock clustering method had a cluster purity of 99.6%. Theta value of 0.8 was used to identify the number of clusters and was able to group the data into 10 clusters. Among all the clustering modell techniques Rock Clustering seems to be the best fit.



Decision Trees Results:

The decision tree model was created with Rpart with split as Gini that created a decision tree with variables that are important.Variables odor,spore print color,stalk color below and above ring are used in the tree construction.The accuracy of the model came in at 99.5%.

Different combination of Variables were tried against Decision Class to check for improvement in accuracy

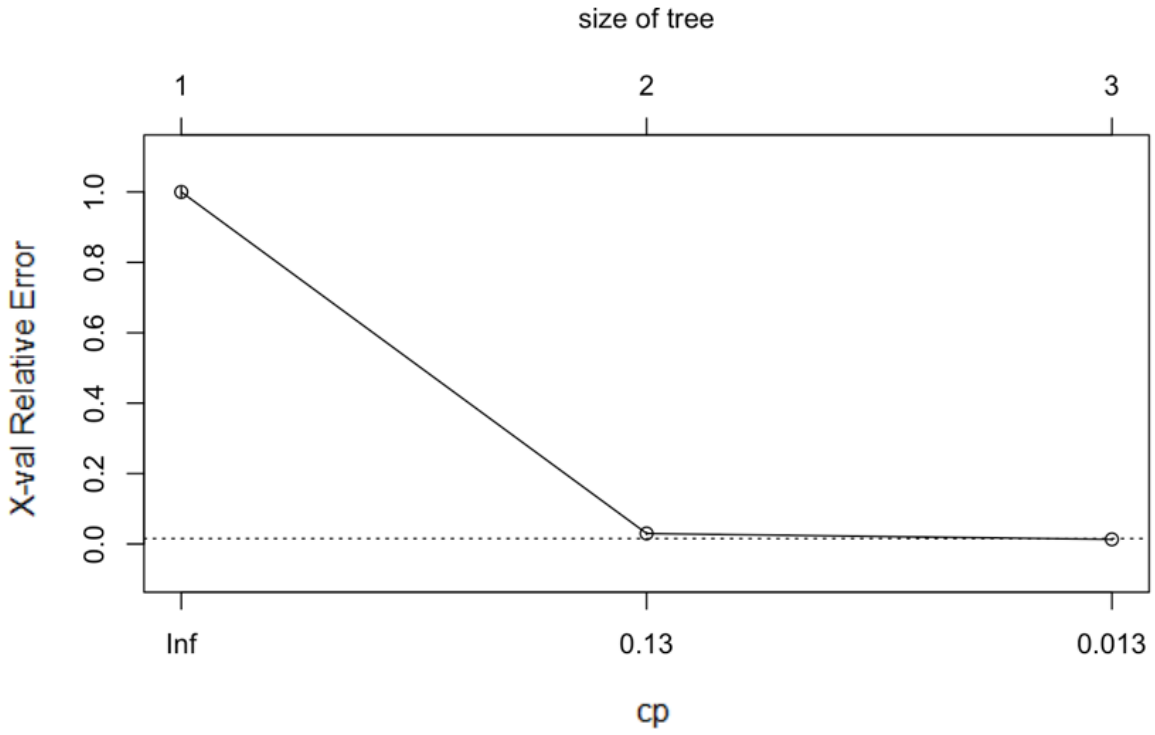



```
Variables actually used in tree construction:
[1] odor                    spore.print.color       stalk.color.below.ring   stalk.surface.above.ring

Root node error: 2740/5687 = 0.4818

n= 5687

         CP nsplit rel error    xerror      xstd
1 0.9700730      0 1.0000000 1.0000000 0.0137522
2 0.0167883      1 0.0299270 0.0299270 0.0032810
3 0.0062044      2 0.0131387 0.0131387 0.0021828
4 0.0023723      3 0.0069343 0.0069343 0.0015882
5 0.0000000      5 0.0021898 0.0021898 0.0008935
```

Decision Tree Confusion Matrix Results:

```
Confusion Matrix and Statistics

          Reference
Prediction  Edible Poisonous
  Edible      1261        12
  Poisonous      0      1164

               Accuracy : 0.9951
                 95% CI : (0.9914, 0.9975)
    No Information Rate : 0.5174
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9901
 Mcnemar's Test P-Value : 0.001496

            Sensitivity : 1.0000
            Specificity : 0.9898
         Pos Pred Value : 0.9906
         Neg Pred Value : 1.0000
             Prevalence : 0.5174
         Detection Rate : 0.5174
   Detection Prevalence : 0.5224
      Balanced Accuracy : 0.9949

       'Positive' Class : Edible
```

Plot Cp vs Relative Error:



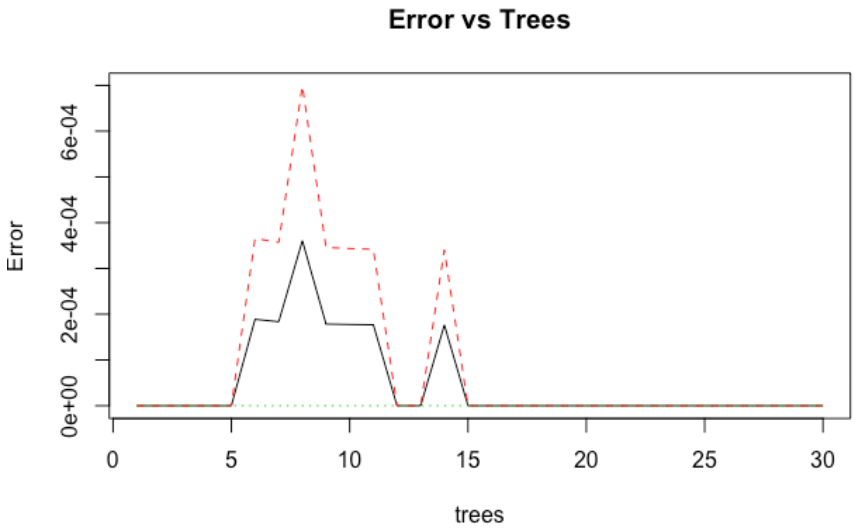
Random Forest Results:

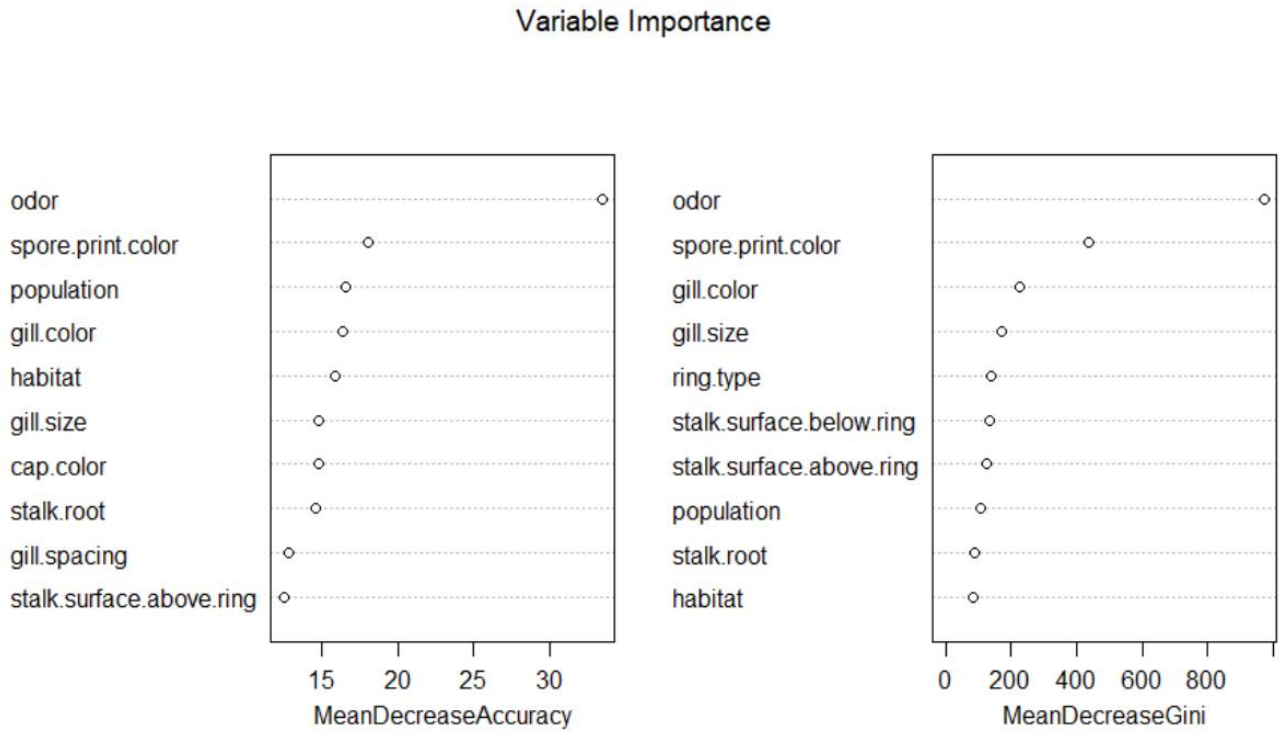
Random Forest has better predictive power and accuracy than the normal decision tree. Created a Random forest model by calling the rpart function. Created a base model with default parameters and value and then tuned the model with different mtrys and trees. The CP (complexity parameter) is used to control tree growth. If the cost of adding a variable is higher then the value of CP, then tree growth stops. Here the tree growth stopped at 15.The variable importance was calculated and plotted. Identified Odor and spore print color as the variables with highest importance.

Accuracy came in at 100% compared to 99.5% from normal decision Tree.

Random Forest Confusion Matrix Results:

```
Confusion Matrix and Statistics

          Reference
Prediction  Edible Poisonous
  Edible      1264         0
  Poisonous      0      1173

               Accuracy : 1
                 95% CI : (0.9985, 1)
    No Information Rate : 0.5187
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1
 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.5187
         Detection Rate : 0.5187
   Detection Prevalence : 0.5187
      Balanced Accuracy : 1.0000

       'Positive' Class : Edible
```





Naïve Bayes Results

This model had a 95% accuracy rate, but many poisonous mushrooms were classified as edible.

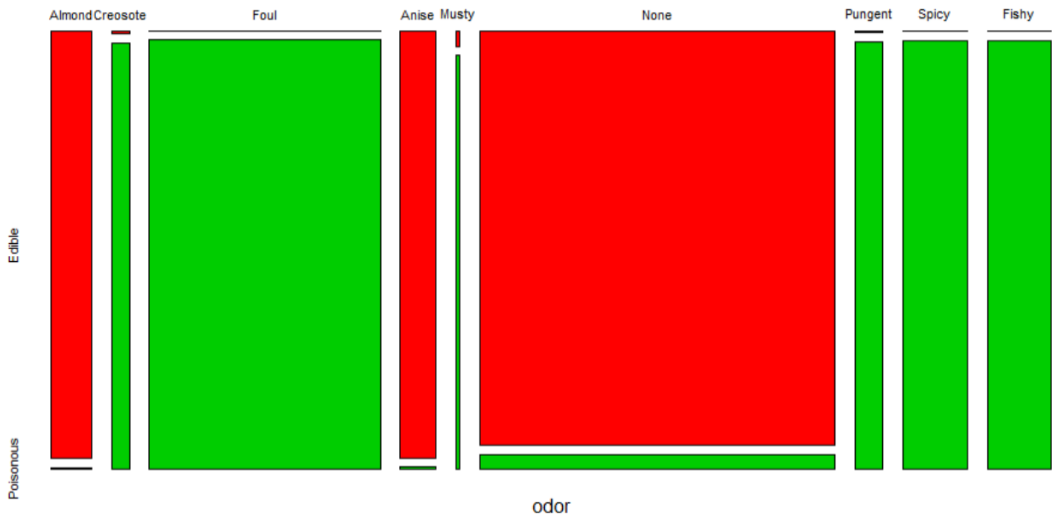
Model from Naïve Bayes Package:

nb_prediction	TestClassLabels: Edible	Poisonous
Edible	1244	113
Poisonous	1	1079

Accuracy: 95.32%
Most information gained from: Odor

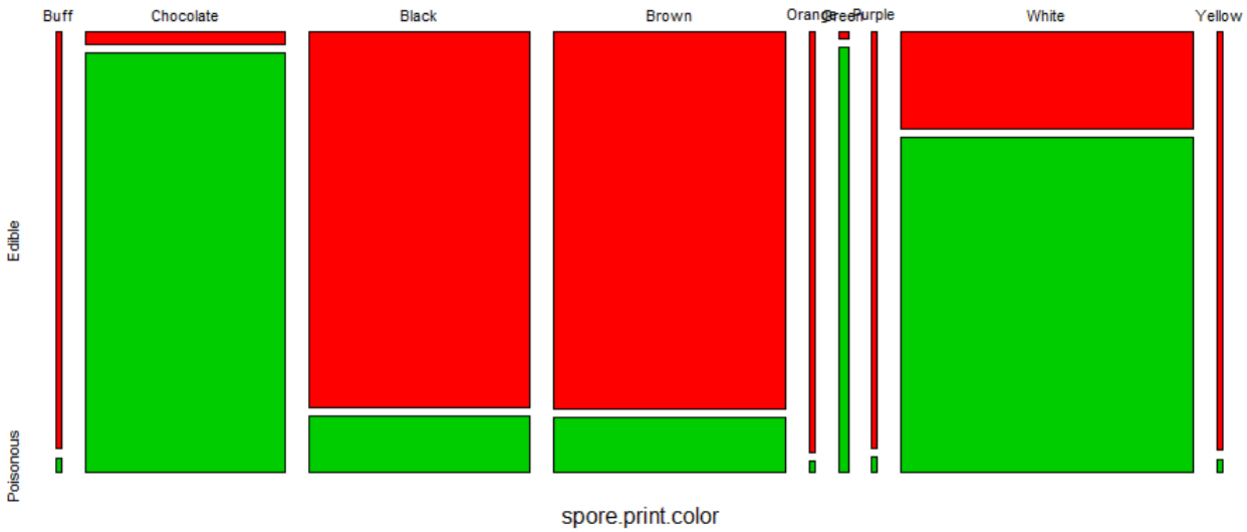
Naive Bayes Odor:

This model showed that odor is a strong predictor of whether a mushroom is edible or poisonous. Mushrooms with an odor of almond, anise, or none are likely to be edible, while mushrooms with an odor of creosote, foul, musty, pungent, spicy, or fishy are likely to be poisonous.



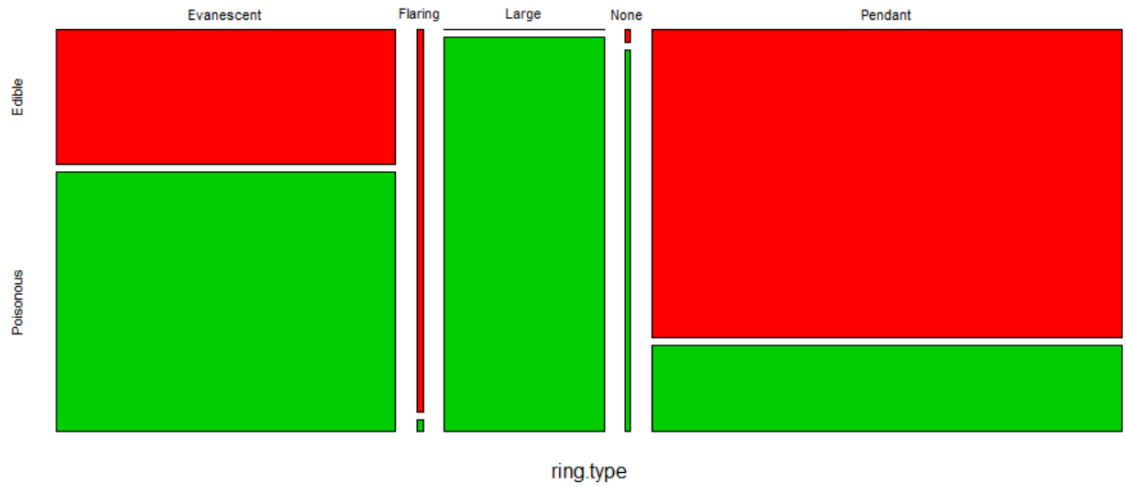
Naive Bayes Spore Print Color:

A spore print color of buff, orange, purple, or yellow is likely to be edible, and a spore print color that is chocolate or green in color is likely to be poisonous. Black, brown, and white were more ambiguous.



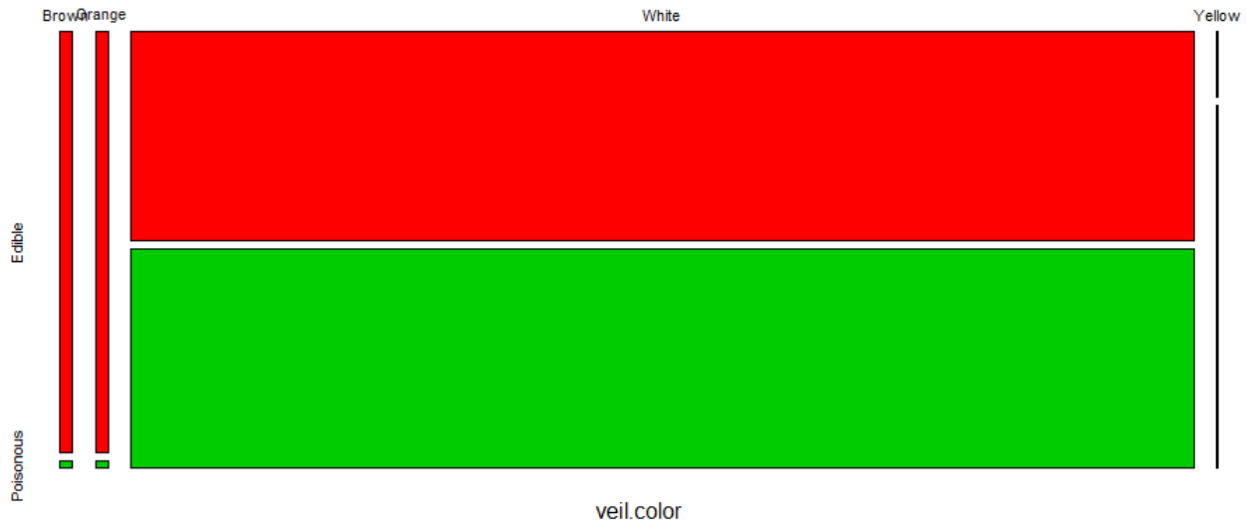
Naive Bayes Ring Type:

A large ring type or no ring is likely to be poisonous, while a flaring ring is likely to be edible.



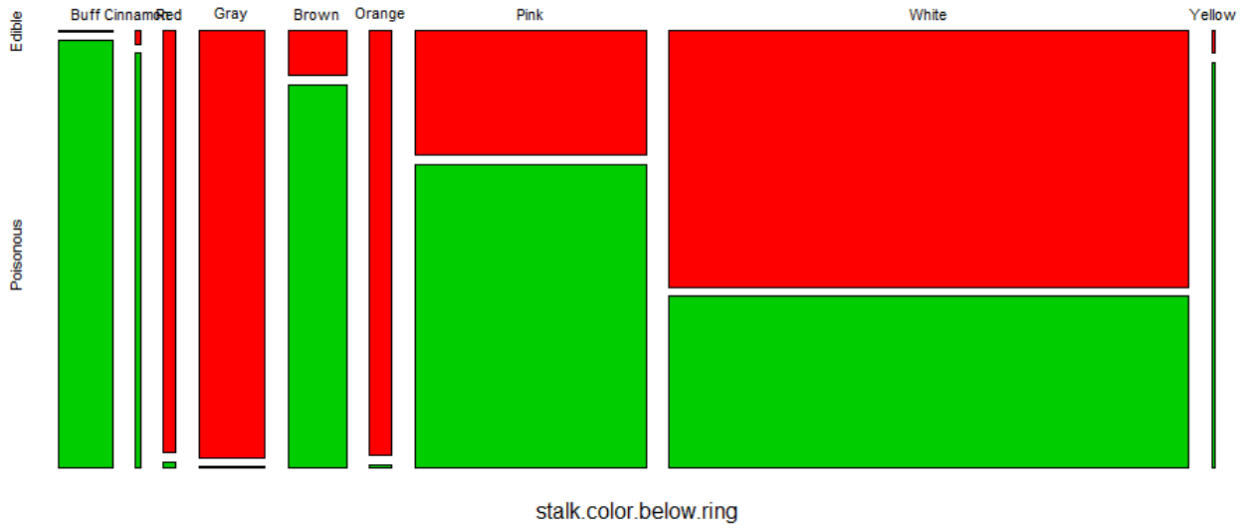
Naive Bayes Veil Color:

A brown or orange veil color is very likely to be edible.



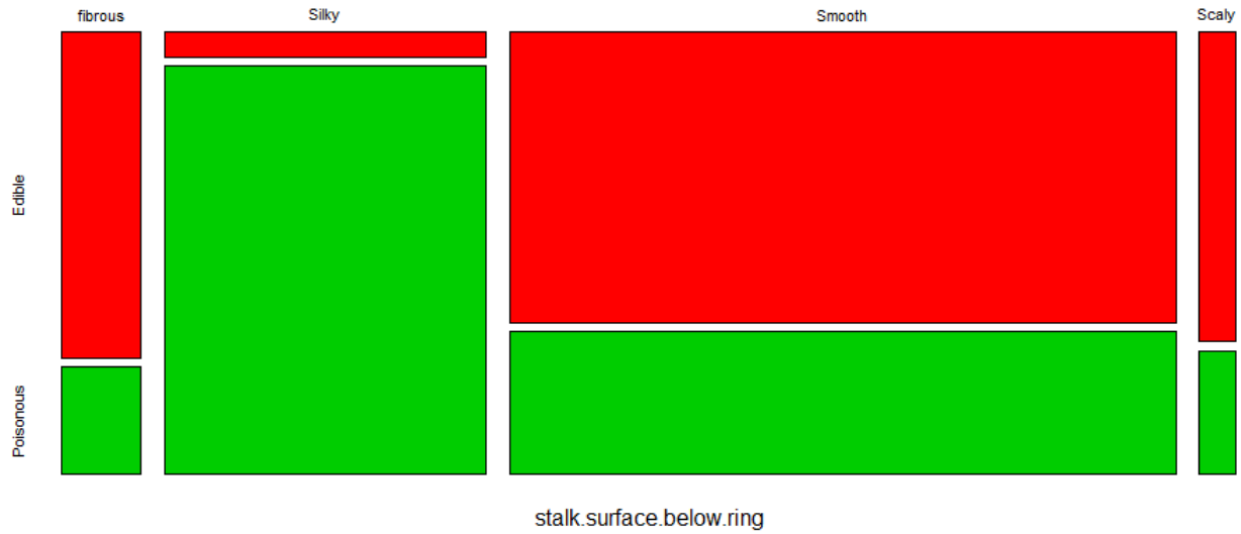
Naive Bayes Stalk Color:

Stalk color above the ring and stalk color below the ring showed almost identical results. In either case, buff, cinnamon, and yellow are likely to be poisonous, while red, gray, and orange are likely to be edible.



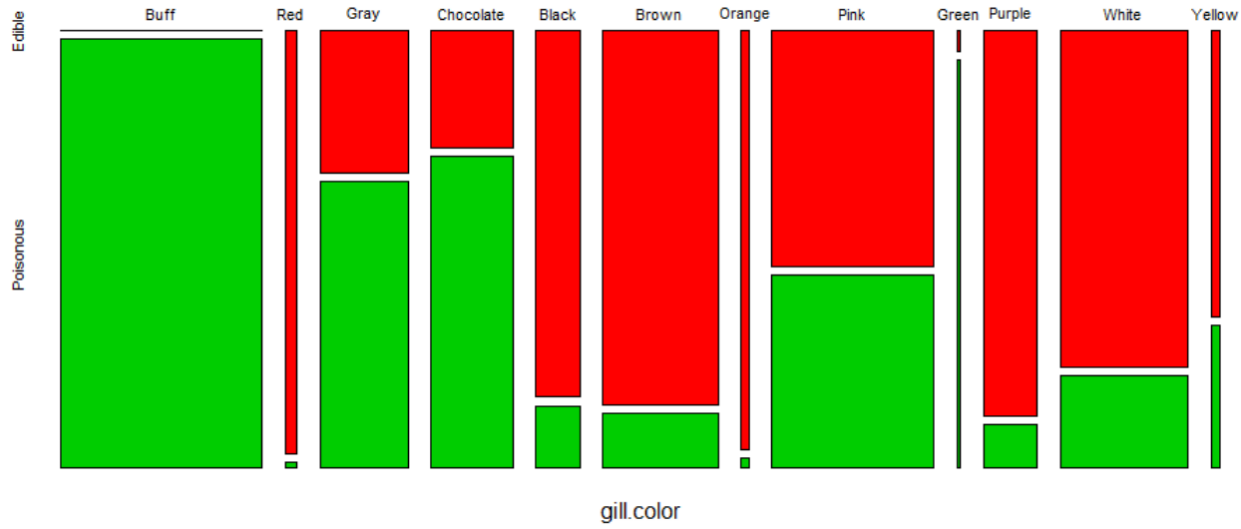
Naive Bayes Stalk Surface:

Stalk surface above the ring and stalk surface below the ring showed almost identical results. In either case, a silky stalk is likely to be poisonous.



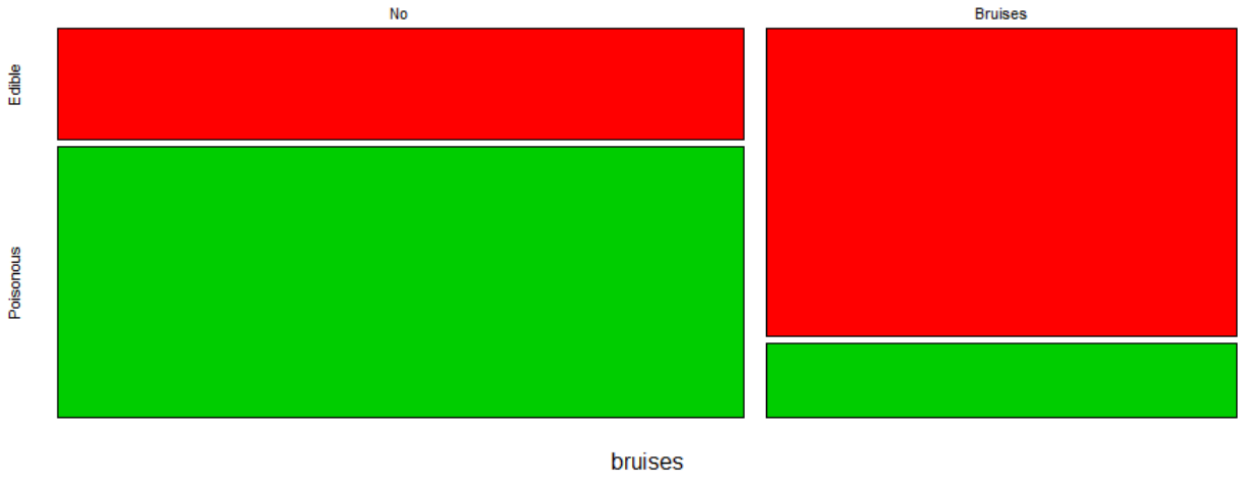
Naive Bayes Gill Color:

A buff or green gill color is likely to be poisonous, while a red or orange gill color is likely to be edible.



Naive Bayes Bruising:

Mushrooms that bruise are more likely to be edible than mushrooms that do not bruise.



Support Vector Machine Results:

Support Vector Machine achieved a 100% accuracy in predicting edible vs poisonous mushrooms. Except for sigmoid kernel, all other kernels did a perfect prediction.

Radial:

```
Confusion Matrix and Statistics

          Reference
Prediction  Edible Poisonous
  Edible      1264         0
  Poisonous      0      1173

               Accuracy : 1
                 95% CI : (0.9985, 1)
    No Information Rate : 0.5187
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1
 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.5187
         Detection Rate : 0.5187
   Detection Prevalence : 0.5187
      Balanced Accuracy : 1.0000

       'Positive' Class : Edible
```

Polynomial

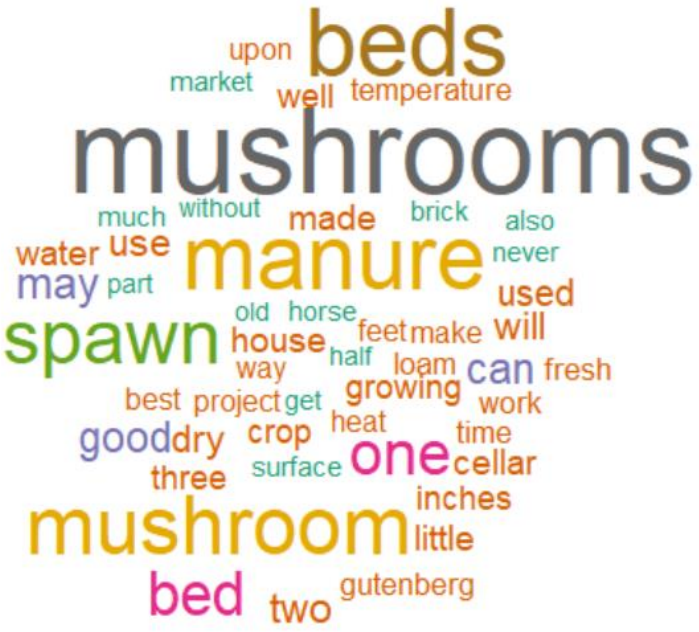
```
Confusion Matrix and Statistics

          Reference
Prediction  Edible Poisonous
  Edible      1264         0
  Poisonous      0      1173

               Accuracy : 1
                 95% CI : (0.9985, 1)
    No Information Rate : 0.5187
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1
 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.5187
         Detection Rate : 0.5187
   Detection Prevalence : 0.5187
      Balanced Accuracy : 1.0000

       'Positive' Class : Edible
```

Sigmoid

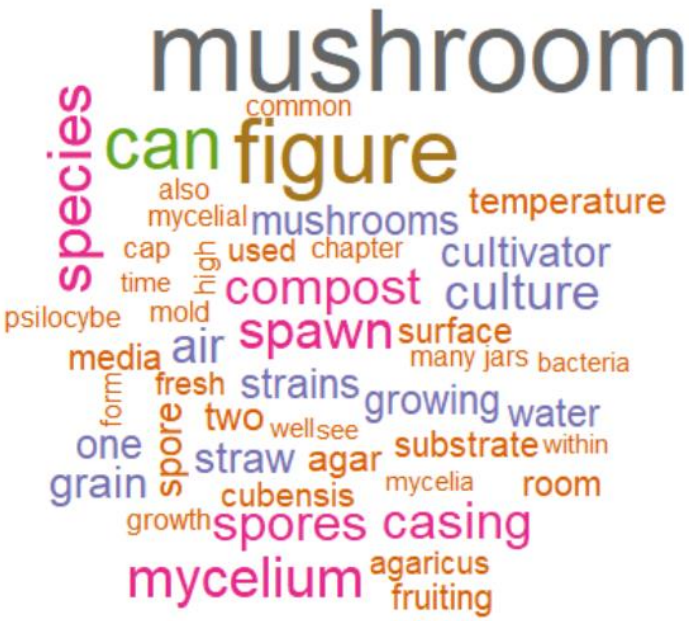
```
Confusion Matrix and Statistics

          Reference
Prediction  Edible Poisonous
  Edible       821       477
  Poisonous    440       699

               Accuracy : 0.6237

                 95% CI : (0.6041, 0.643)
    No Information Rate : 0.5174
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.2457
 Mcnemar's Test P-Value : 0.2345

            Sensitivity : 0.6511
            Specificity : 0.5944
         Pos Pred Value : 0.6325
         Neg Pred Value : 0.6137
             Prevalence : 0.5174
         Detection Rate : 0.3369
   Detection Prevalence : 0.5326
      Balanced Accuracy : 0.6227

       'Positive' Class : Edible
```

Text Mining Results

Word Cloud from *Mushrooms, How to Grow Them*, 1892



Word Cloud from *The Mushroom Cultivator A Practical Guide*, 1983



Changes in Word Frequencies:

	1892	1983
Beds	451	8
Manure	420	8
Spawn	343	82
Cellar	100	0
Loam	99	1
Temperature	92	46

	1892	1983
Compost	1	76
Straw	37	56
Agar	0	45
Species	7	95
	4708 terms	3199 terms

The second text has much less mention of beds, manure, cellar, and loam. The terms spawn, temperature, and straw were still being utilized. The terms compost, agar, and species became used more frequently.

Conclusions:

A Czech adage states that "every mushroom is edible, but some only once." There are about 100 species of mushrooms that are classified as poisonous, with some causing only gastrointestinal distress, and others causing death. One of the most deadly mushrooms is the species Amanita, also called "death cap" which is toxic to the liver. Mushroom hunting requires that the hunter be able to identify the mushroom species, and have an awareness of which species are edible. Hunters should avoid collecting mushrooms that are difficult to identify, and should be familiar with which poisonous mushrooms look very similar to edible ones (sometimes called "evil twins"). Mushrooms should be cooked unless it can be verified that the mushroom species can be eaten raw. Even if a mushroom is considered edible, it is wise to only consume a small amount if it is a species the person has not eaten before, in case the person is sensitive and has an adverse reaction.

In addition to being able to identify the species, several characteristics of mushrooms are clues as to whether they may be poisonous or edible. Mushrooms that bruise tend to be edible, but this is not always the case. Mushrooms with no odor, or an almond or anise odor are almost always edible. Mushrooms with a foul, creosote, pungent, spicy, fishy, or musty odor are almost always poisonous. A gill color of buff or green are usually poisonous, while red or orange are usually edible. A silky stalk surface is usually poisonous, as is a stalk color of buff, cinnamon, brown, or yellow (red, gray, or orange are usually safe to eat). A brown or orange veil color are usually safe to eat. A large or missing ring is likely to be poisonous, while a flaring ring is likely to be edible. Finally, mushrooms with chocolate or green spore print colors should be avoided, while the colors buff, orange, purple, and yellow are usually safe.

Finally, foragers need to know where and where not to look for mushrooms. The Chernobyl disaster caused a substantial amount of radioactivity to be released into the environment, and mushroom hunters need to consider the additional risk of radiation exposure that consuming some mushrooms may carry. Fungi, the structures that produce mushrooms, can have quite large underground growth, covering thousands of acres. Sometimes mushroom maps are available to assist hunters in finding safe and edible mushrooms. Several mushrooms, such as sulfur shelf, appear in bulk in the same location, year after year. In fact, a specimen weight 100 pounds was found in the United Kingdom in 1990. The popularity of mushroom picking has led to several annual festivals around September, such as Pennsylvania's mushroom festival at Kennett Square, which just had its 32nd annual festival.