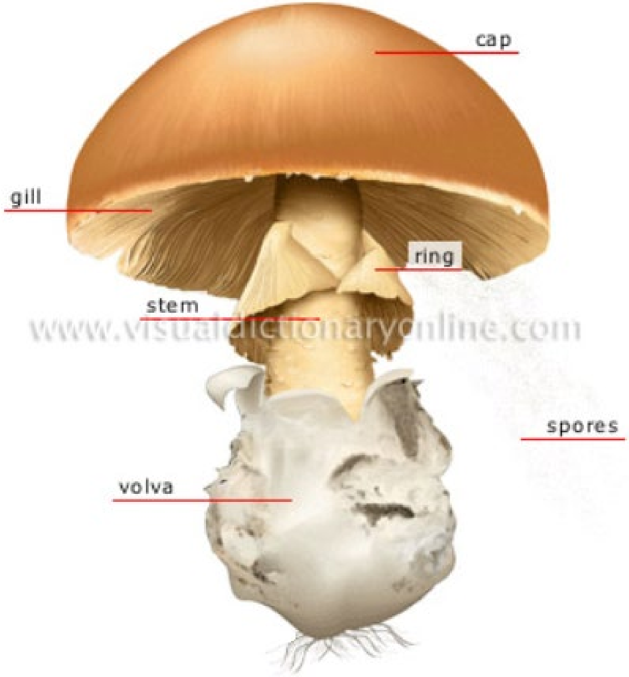
Data Mining Project

Baskar Dakshinamoorthy & Lauren Foltz

About the Data Set

- Mushroom Data set
- 8,124 rows x 23 columns
- The target variable is Edible / Poisonous
- Other variables include
  - Descriptions of the mushroom structure
    - Cap, Gills, Stalk, etc.
  - Population
    - Clustered, Numerous, Scattered, etc.
  - Habitat
    - Grasses, Leaves, Meadows, etc.
  - Odor
    - Almond, Fishy, Spicy, etc.

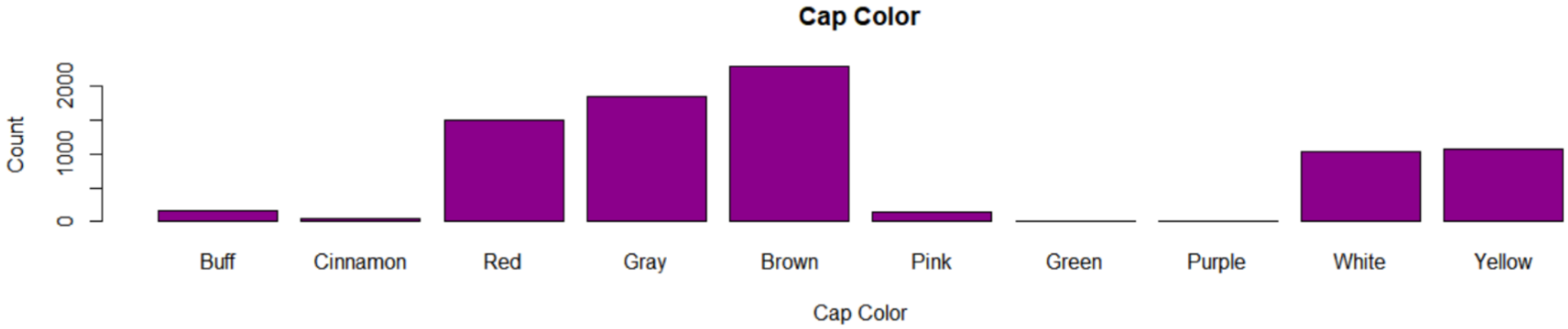


Goals & Questions

- The main goal is to predict “Edible” or “Poisonous” based on the other variables.
- Other questions include
    - Which variables are the most useful for prediction?
    - Are certain traits associated with each other?
    - Does this data set cluster well, and how many clusters would there be?

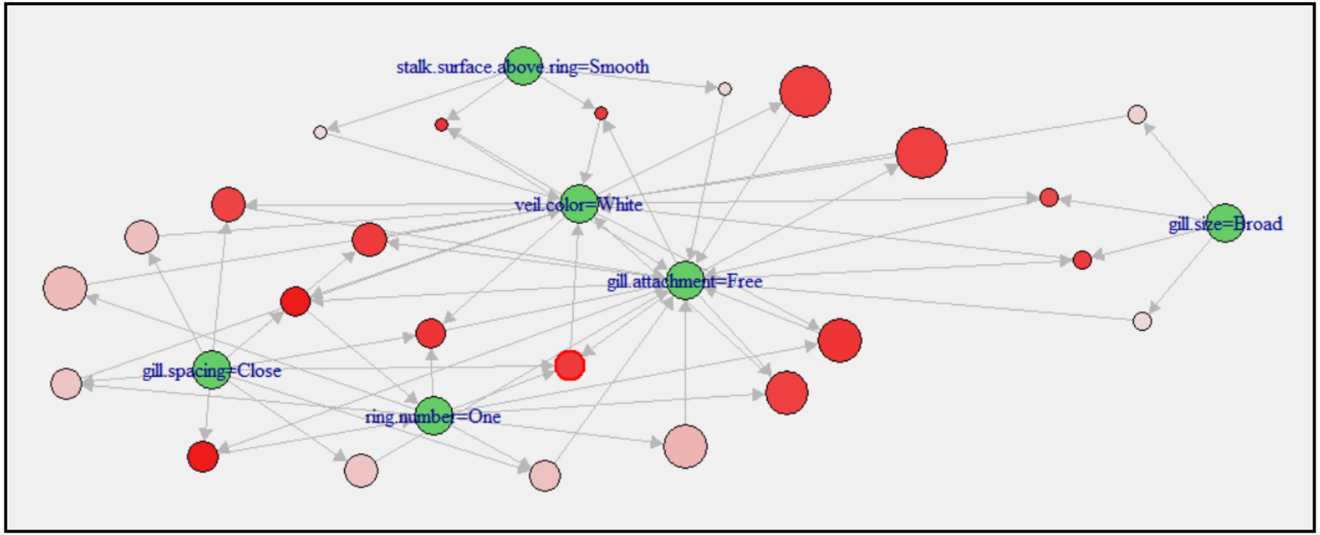
Data Cleaning and Visual EDA

- Missing Values: “Stalk Root” had a “?” for over 2,000 instances, and didn’t appear to be a strong predictor, so it was removed.
- Values were initially just letters, so “recode” was used to rename the variables into something more meaningful.
- Variables were visualized using bar charts.



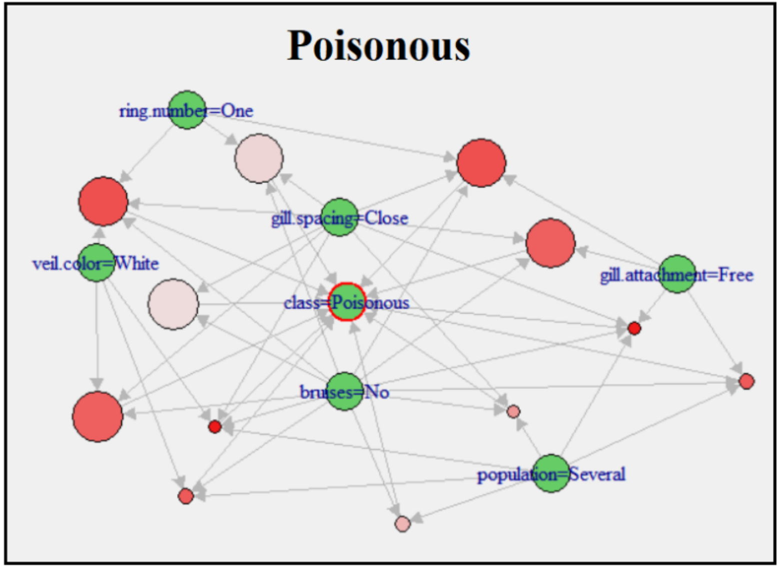
Association Rules Mining

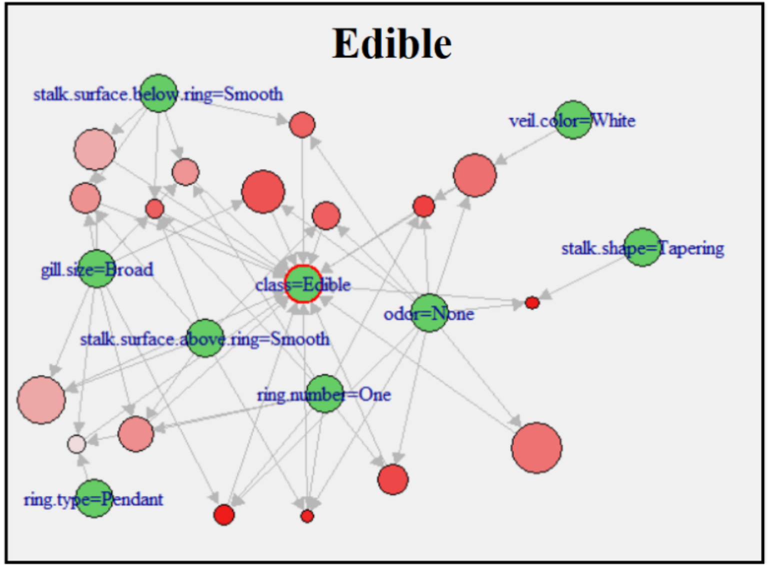
- First we ran models without LHS or RHS specified.
  - Confidence and Support were strong within the dataset, but Lift was not so strong.
  - A white veil color was highly associated with free gill attachment.



Association Rules Mining

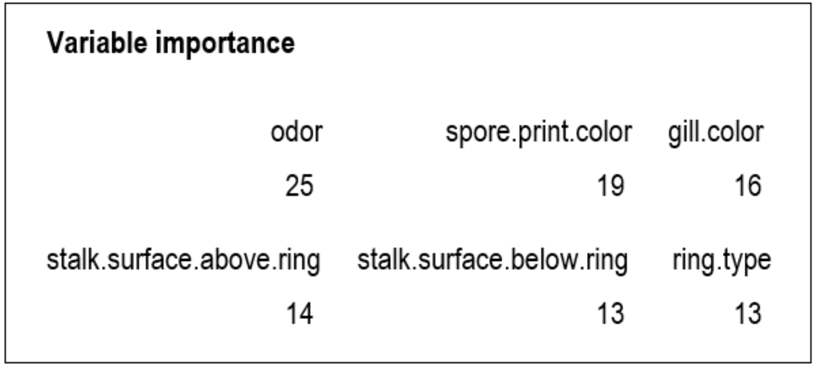
- We ran several models to determine associations with edibility.
  - Poisonous Mushrooms are associated with closely spaced gills and no bruises.
  - Edible mushrooms are associated with no odor and a smooth stalk surface above the ring.



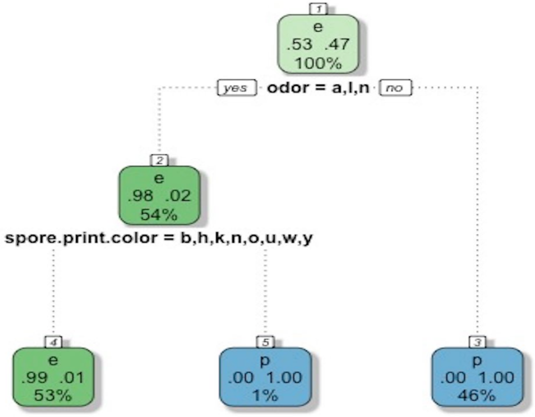


Decision Tree Classification

- Used 70:30 data split for creating training and test datasets.
- Prediction Accuracy came in at 99.3% With gini index as criteria.
- Complexity Parameter(CP) used 2 variables to construct the decision tree.

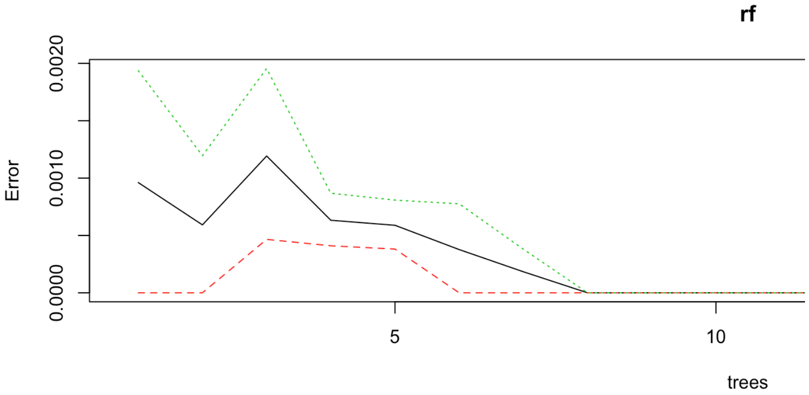


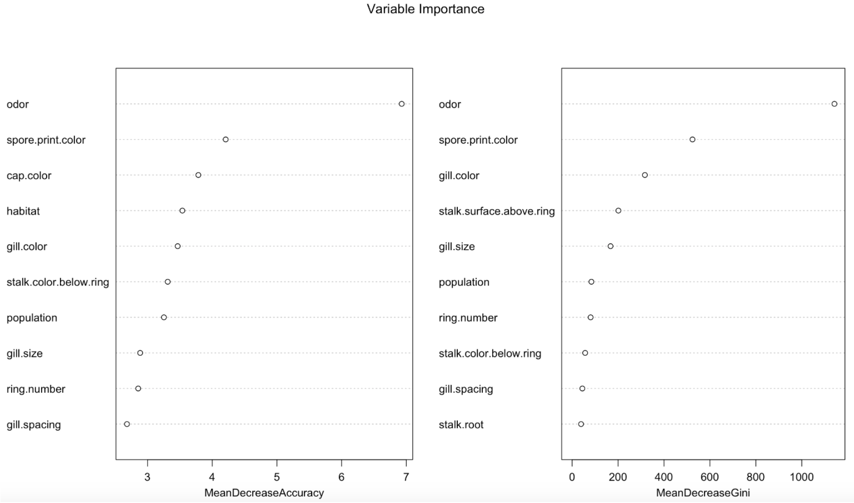
```
Variable importance

                   odor         spore.print.color  gill.color
                     25                        19          16
stalk.surface.above.ring stalk.surface.below.ring   ring.type
                     14                        13          13
```



Random Forest Classifier

- Used Random Forest Classifier to predict the importance of Variables and to avoid overfitting.
- Accuracy came in at 100% compared to Decision Tree Accuracy of 99.38%
  - Error Rate flatlined after constructing 8 decision trees
  - Mean Decrease Accuracy and Mean Decrease Gini showed odor and spore print color as the top variables used in prediction.





Looking Forward

- Methods still left to use:
  - Clustering
  - Naive Bayes
  - SVM
  - Text Mining
- Mushroom data set has no quant data
  - We may use the wine quality data set.