Mushroom Classification

Winfield Lai

November 26, 2019

Mushroom Data

- Mushrooms (5935 observations) are categorized as Edible (3767 observations) or Poisonous (2168 observations). [1]
- Covariate data is composed of 22 categorical factors (e.g. CapColor, VeilColor, CapShape,...etc) of a Mushroom.
    - e.g. CapColor is either brown, gray, red,... etc
- Objective: Accurately classify a mushroom with as few and as unambiguous categorical factors as possible.
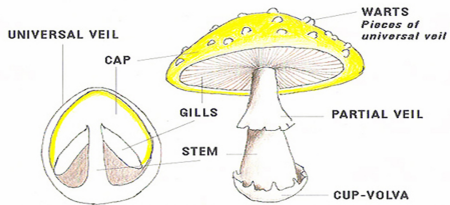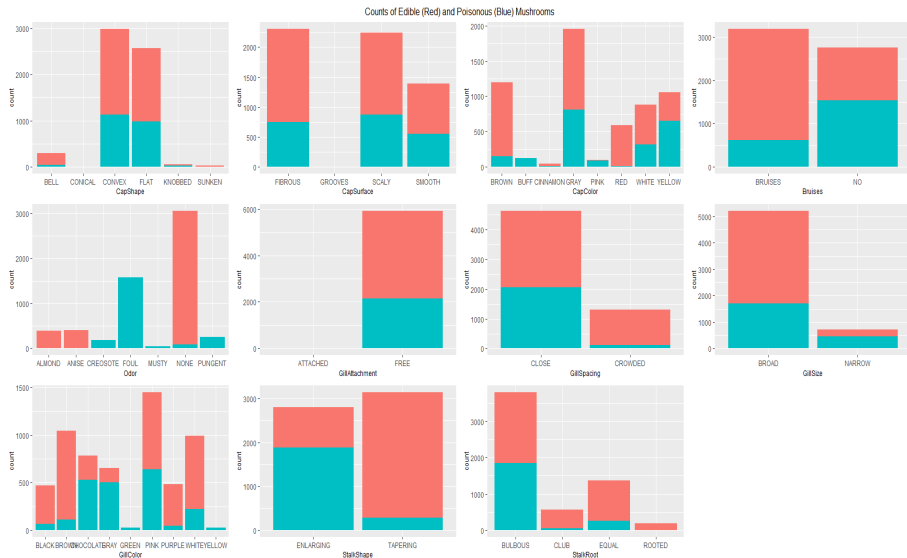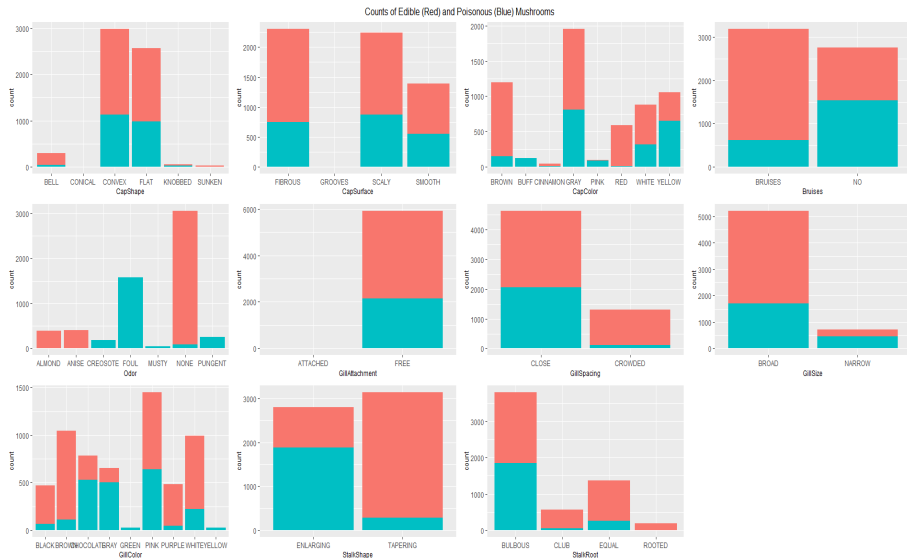

image from https://anandakalyani.org/blog/2015/12/07/discovering-mushrooms-at-master-unit/

Stacked Bar Graph of Mushroom Data

Counts of Edible (Red) and Poisonous (Blue) Mushrooms

Stacked Bar Graph of Mushroom Data

Counts of Edible (Red) and Poisonous (Blue) Mushrooms
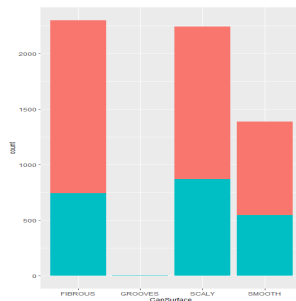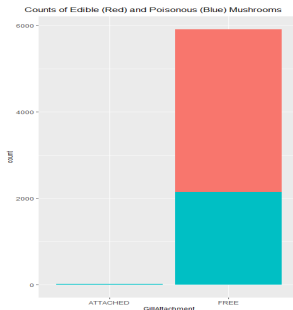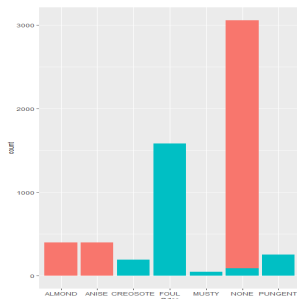
Stacked Bar Graph of Mushroom Data

- Some covariates (i.e GillAttachment) have no meaningful contribution
- Some covariates have good separation (i.e. Odor)
- Most covariates are mixed.

Counts of Edible (Red) and Poisonous (Blue) Mushrooms

Classification and Variable Reduction Methods

- Using the full set of covariates results in an average ARI of 1 for all classification methods listed below (fifteen different 50:50 train/test set).
- Association rules and stacked bar graphs will be used to choose a subset of covariates for classification. Average ARI will be used for comparisons.
- 2 Classification methods will be optimized on the chosen subset of covariates
  - Random Forests
  - Neural Networks

Classification and Variable Reduction Methods: Association Rules

- Association rules are statements wherein a set of non-empty covariates (A) is "associated" to a different non-empty set of Covariates (B).
- Certain measures have been used to characterize whether a rule is meaningful or interesting (lift, support, confidence, etc).
- Support: $s(A \Rightarrow B) = P(A, B)$
- Confidence: $c(A \Rightarrow B) = \frac{P(A,B)}{P(A)}$
- Lift: $L(A \Rightarrow B) = \frac{P(A,B)}{P(A)P(B)}$

Classification and Variable Reduction Methods: Random Forest

- Uses M classification trees to make a prediction based on the given predictor variables
- At each split, a random subset of the predictor variables are considered for the next split. Size of the subset can be controlled.
- M classification trees are produced via a re-sampling method of the training set.
- $\hat{f}$ Indicates the prediction. The predictions are hardened for classifications.

$$\hat{f}_{\text{forest}}(x) = \frac{1}{M} \sum_{m=1}^{M} \hat{f}^m(x)$$

Classification and Variable Reduction Methods: Neural Network

- Composed of a hidden layer, input covariates and an output
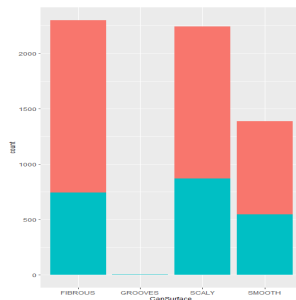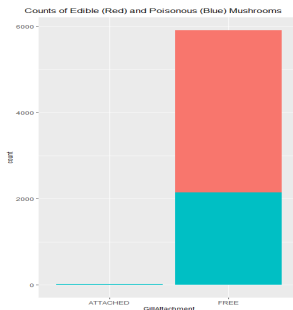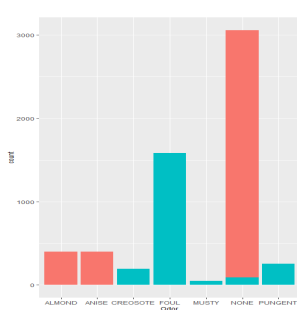- An activation function that controls the activation in the hidden layer

$$\sigma(v) = \frac{1}{1 - e^{-v}}$$

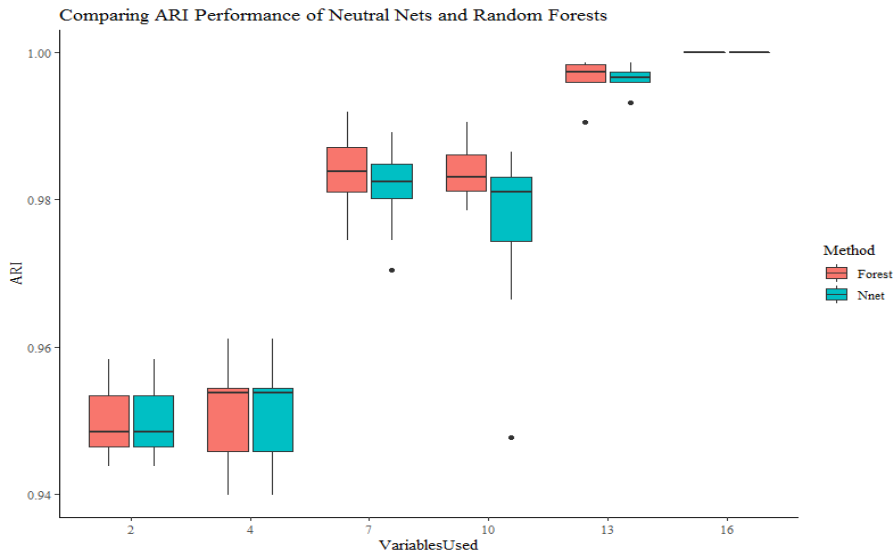- An output function that transforms the neural network output to something more usable

$$g_k(T) = \frac{E^{T_k}}{\sum_{h=1}^{K} e^{T_k}}$$

Variable Reduction

- The covariates used in the final model are based off of the associated rules with the highest standardized lift and on stacked bar graphs like the bottom left one

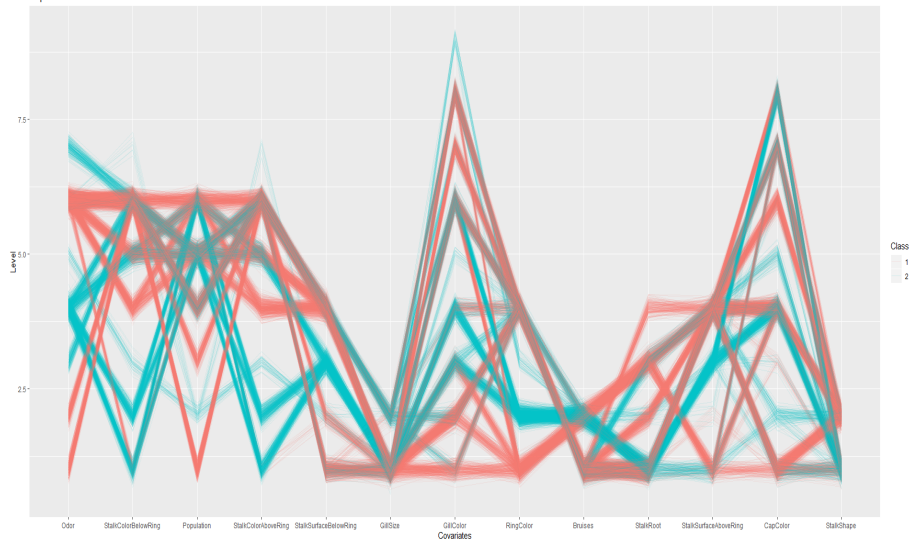- ARI's were calculated (on unoptimized methods) as covariates were added to the model

Counts of Edible (Red) and Poisonous (Blue) Mushrooms

Variable Reduction

Comparing ARI Performance of Neutral Nets and Random Forests

Variable Reduction

Classification

- 13 Variables will be used - We will see if parameter optimization can increase the ARI to 1.
- Parameters were optimized for Random Forests and Neural Network
- Random Forest: $mTry = 5, nTree = 300 \Rightarrow ARI = 0.9979$
- Neural Net: $size = 2, maxit = 100, decay = 06 \Rightarrow ARI = 0.9972$

Conclusion

- Depending on how high of an ARI desired, we can choose between 2 and 16 variables to get from near perfect classification to perfect classification.
- It is possible to classify whether a mushroom is poisonous or edible based on less than the 22 covariates.
- There might be an issue with over fitting the data.

References

[1] Dheeru Dua and Casey Graff. Uci machine learning repository, Access 2019.

[2] McNicholas. Mcmaster stats 780 data science lecture notes, 2019.