RUTGERS

PROJECT REPORT

FINANCIAL DATA MINING

Vishal Reddy Mekala-213007396
Fall 2022

INTRODUCTION

Mushrooms are reproductive structures produced by fungi that grow under various conditions worldwide. Mushrooms are eaten worldwide, and besides this, they have applications in the medicinal and beauty industries as well, making it essential to classify a poisonous one from a safe one. There are thousands of species of mushrooms, but only a few species of them cause the majority of cases of poisoning when eaten by humans, and only a handful of these species are potentially lethal when ingested.

DATA DESCRIPTION

The dataset is obtained from the UCI Machine Learning repository and contains samples of hypothetical mushrooms corresponding to 173 species. Every record belongs to a unique mushroom identified by various attributes. Most attributes are nominal variables, including our response, while a few are metric. Post data cleaning, the number of observations total to approximately 40,000 with 15 different variables.

PROBLEM STATEMENT

The problem at hand is a supervised machine learning- binary classification problem, with "class" as our response variable. This attribute essentially classifies a mushroom as either "edible" or "poisonous".

OBJECTIVE

The goal of this project is to conduct exploratory data analysis to make inferences based on the characteristics of mushrooms and then build a model that is accurate, interpretable, and consistent to predict edibility based on future data.

The primary metric used to compare models is Accuracy and Recall. I chose Recall as an essential metric because the cost of classifying a poisonous mushroom as an edible is very high and can potentially be life-threatening.

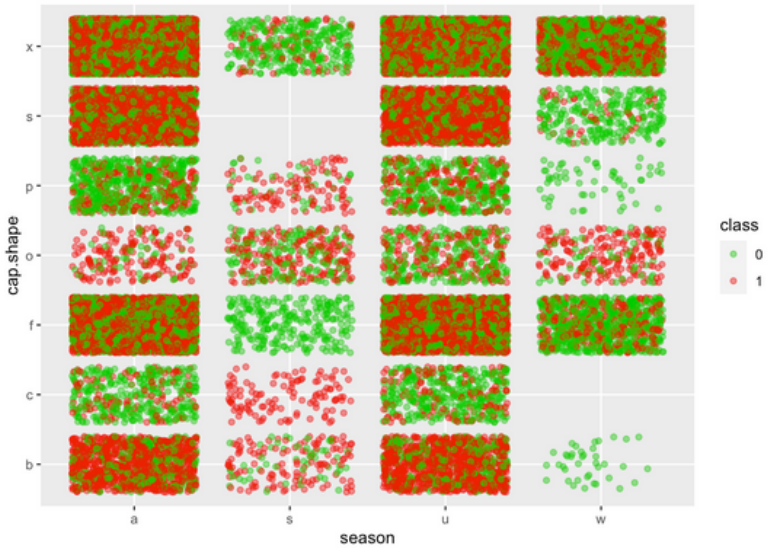
Tasks:

- Exploratory Data Analysis
- Logistic Regression
- Linear Discriminant Analysis
- Tree-based Algorithms

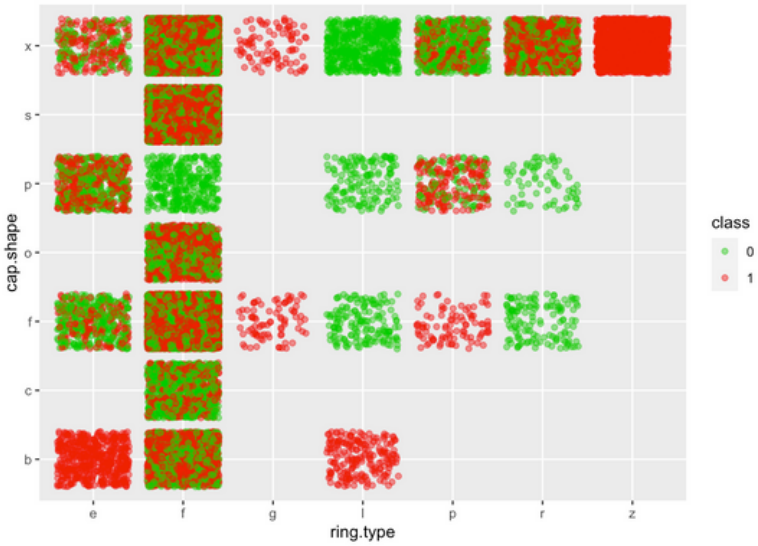
RUTGERS

EXPLORATORY DATA ANALYSIS

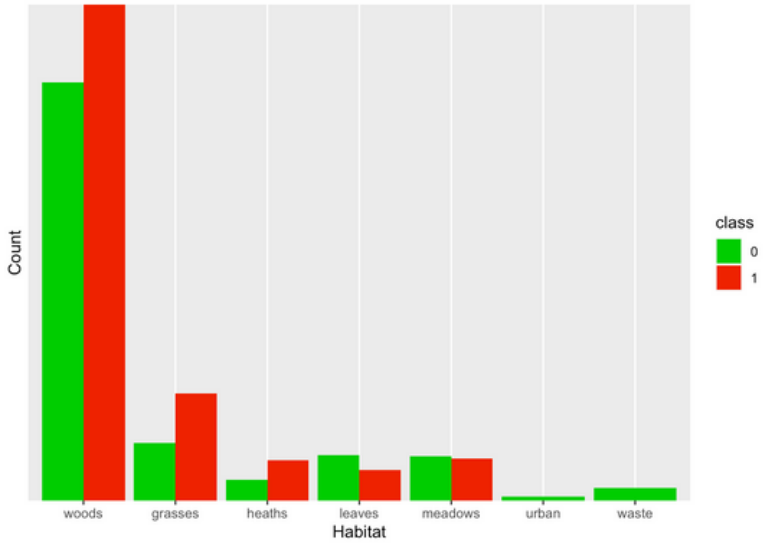
The green patches indicate that mushrooms in that region are safe to eat, while the red indicates poisonous.



- The first inference that can be made is that mushrooms in the season 'spring' and 'winter' that are 'flat cap' and 'bell-shaped' respectively are safe to consume.



- From the second plot, it looks like ring type and cap shape seem to give us much information on the edibility of the mushrooms. Convex shaped caps(denoted by 'x') seem to be the most challenging type to classify visually.

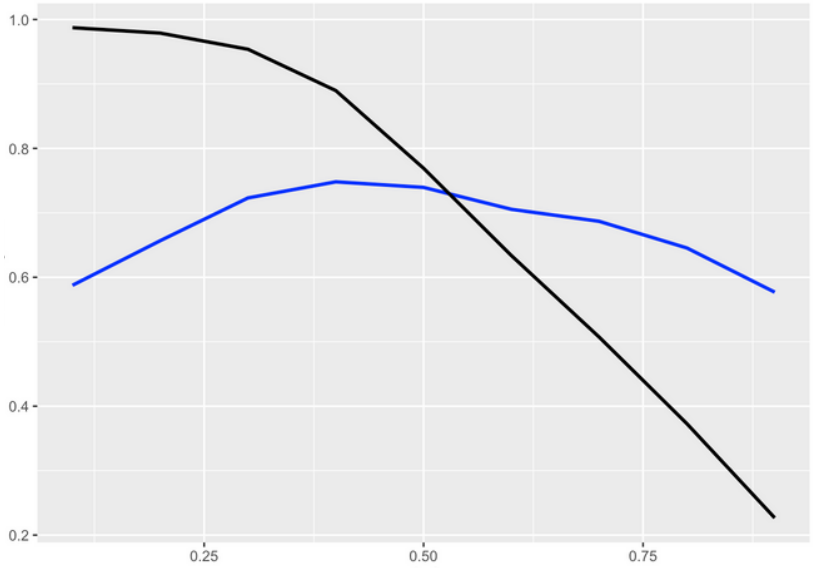


- Although the ones found in urban and waste habitats are classified as not poisonous, it is essential to consider other factors as well because it might be cross contaminated.

RUTGERS

LOGISTIC REGRESSION

On performing Logistic Regression on this data and validating it with the testing data, the accuracy and recall levels were roughly 79%. The threshold to classify an observation as 'edible' or 'poisonous' was chosen by hyperparameter tuning.
The blue line below shows us how the accuracy levels change with different thresholds, while the black represents recall.



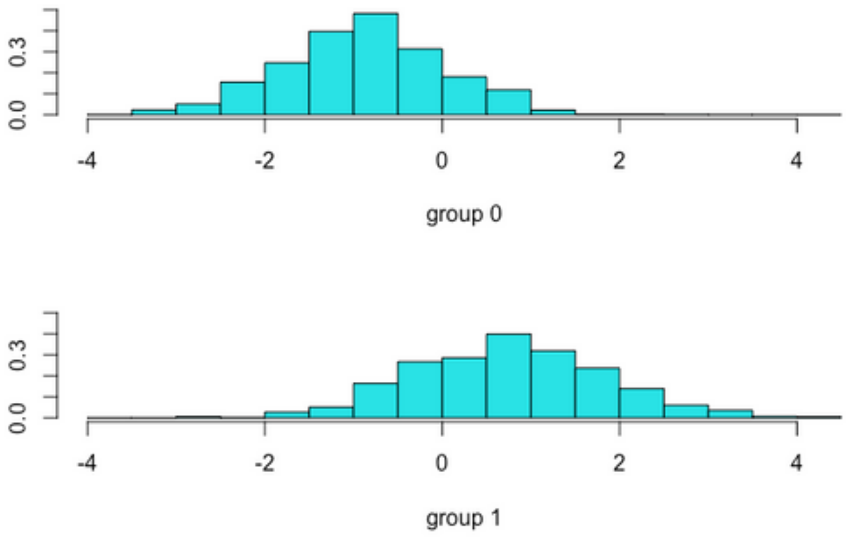
CONFUSION MATRIX		PREDICTED	
		0	1
ACTUAL	0	2699	689
	1	839	3185

The confusion matrix for the model looks like this. As seen, there are several misclassifications, making us want to try other models that could potentially reduce errors.

LINEAR DISCRIMINANT ANALYSIS

Although Logistic Regression and Linear Discriminant Analysis reveal the same pattern, they differ in how they operate. LDA requires following a particular set of assumptions, while the previous approach doesn't.

The stacked histogram plots can show us how well this model performs. As seen from the the plot, it is evident that from the region-[0-2], both the histograms have some values indicating misclassifications.



The linear discriminant analysis model also gives us similar results to Logistic Regression. Accuracy and Recall are 79% and 80%, respectively. The confusion matrix for this model is given below.

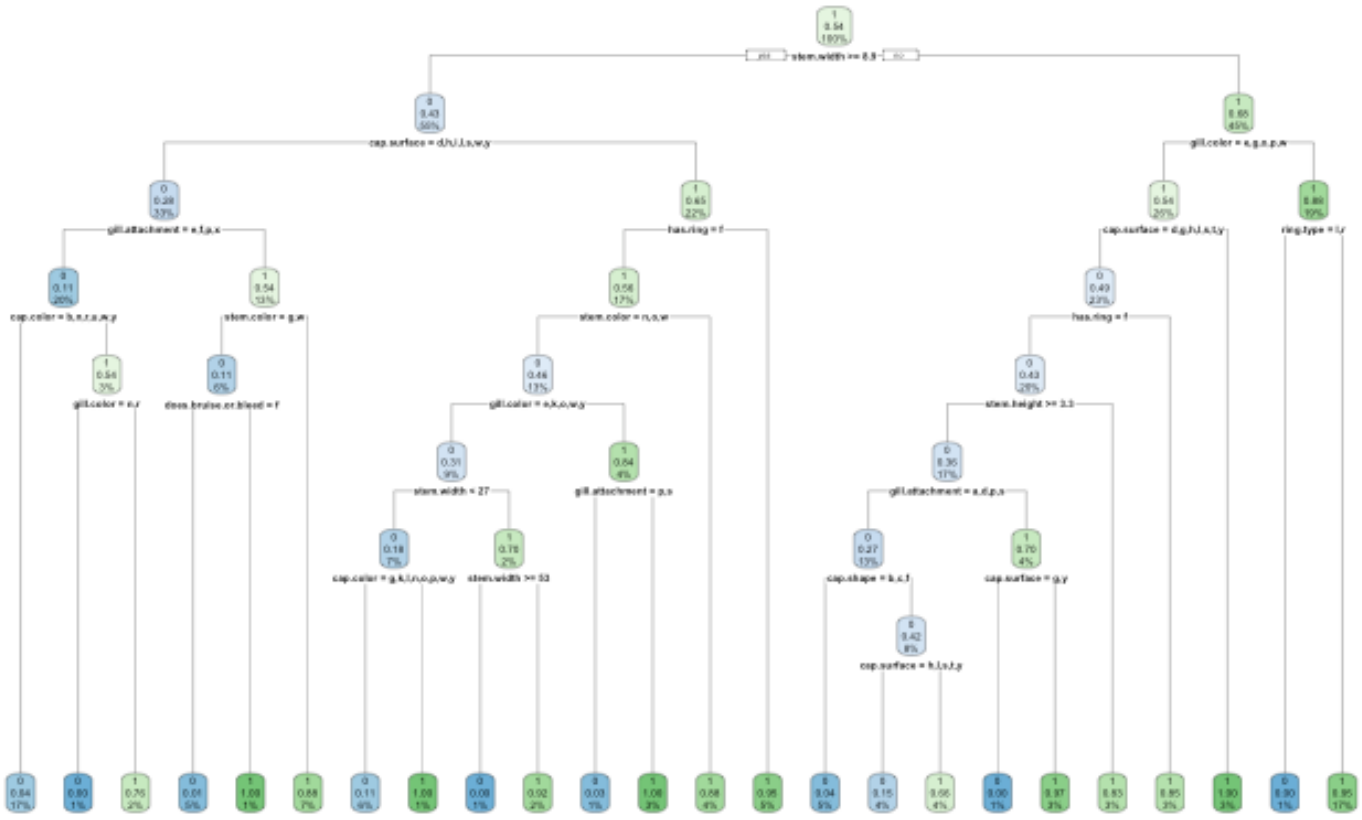
We are looking for more accurate predictions since we want to keep people healthy and away from hospitals. Trees could be a better way to model this data, possibly giving us better results.

CONFUSION MATRIX		PREDICTED	
		0	1
ACTUAL	0	2699	689
	1	839	3185

DECISION TREES

The main advantage of trees over other algorithms is their ease of interpretability. We have arrived at three different decision trees based on different complexity parameter(cp) values in this project.

The first tree was built using R's default cp value, '0.01'. The plot of the tree is given below. The accuracy of this model was 92%, while recall was 96%.



Later I started modeling trees from cp=0 and pruned the decision trees according to their bestcp. Surprisingly both, the pruned tree and the pre-pruned tree seem to perform similarly. Both of them have roughly 99% accuracy and recall. This leads us to more questions- are trees overfitting the data?
Although the depth of the pre-pruned and pruned tree differ a little, their metrics are pretty much similar. The confusion matrix for both trees is given below.

CONFUSION MATRIX (Pre-Pruned)		PREDICTED	
		0	1
ACTUAL	0	3378	10
	1	15	4009
Misclassification rate=0.3%			

CONFUSION MATRIX (Pruned)		PREDICTED	
		0	1
ACTUAL	0	3348	40
	1	45	3979
Misclassification rate=1.1%			

A possible explanation for such high accuracy levels could be overfitting. Maybe random forests can answer this question since it works on Bagging.

RANDOM FORESTS

Now if overfitting was the case for decision trees, random forests should not be giving us the same almost perfect predictions. But surprisingly, it does.

The output of random forests is shown here.
The confusion matrix is for the training data and as seen, it has an almost perfect prediction and OOB error rate is almost negligible. Just like decision trees, even after pruning random forests, there is approximately 99% accuracy levels in training and testing data.

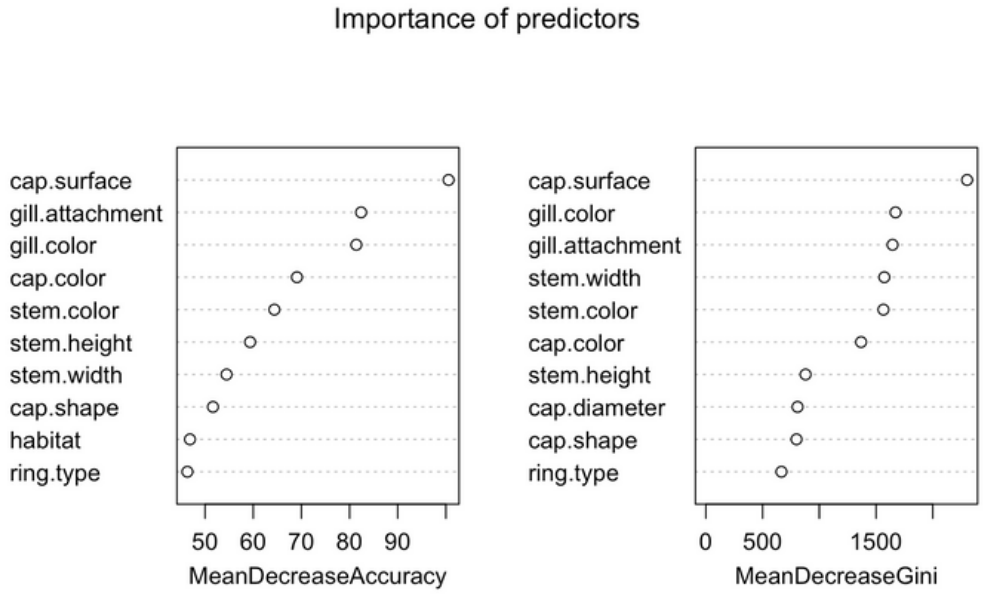
Now, why are we getting such high values?

- Is the data structured this way?
- Are just a few predictors sufficient to classify mushrooms instead of all the variables we used?

```
                Call:
randomForest(formula = class ~ .,
data = train, importance = TRUE)
               Type of random forest:
                     classification
                     Number of trees: 500
No. of variables tried at each split: 3

OOB estimate of  error rate: 0.01%
                     Confusion matrix:
      0     1 class.error
0 13556     0  0.0000000000
1     4 16093  0.0002484935
```

Let's check the importance of variables to get a deeper understanding. From the table below, we can see the top predictors contributing to random forests.



After rerunning the random forests on just the top 'six' predictors shown above, we have yet again arrived at an almost perfect prediction (99%) on the testing data, implying the other predictors are redundant for this data.

CONCLUSION AND SCOPE

Although in terms of pure metrics, tree-based algorithms definitely stand apart from logistic regression and linear discriminant analysis, giving almost perfect predictions, there is more scope for this problem. Based on my analysis and upon reading other blogs on similar datasets, it looks like the data is structured this way.

Since the dataset depicts hypothetical mushrooms involving mainly categorical variables, it would be sensible to try the Multiple Correspondence Analysis, just like PCA, but for ordinal variables. However, it was a great way to learn and apply various algorithms learned in this course, and my quest on this problem will go on.

ACKNOWLEDGEMENT

I want to thank my Professor, **Ying Hung,** for giving me this wonderful opportunity to work on this project. It was indeed a great learning experience.

SOURCES

Data: https://archive.ics.uci.edu/ml/datasets/Secondary+Mushroom+Dataset

Existing Analysis: https://www.kaggle.com/code/turksoyomer/classification-methods-on-mushroom-dataset