Data Mining Final Presentation

Kelly Hwang 09.20.2018

Objective Overview

Using a dataset that contains mushroom characteristics and traits, can we predict whether or not the mushroom is edible?

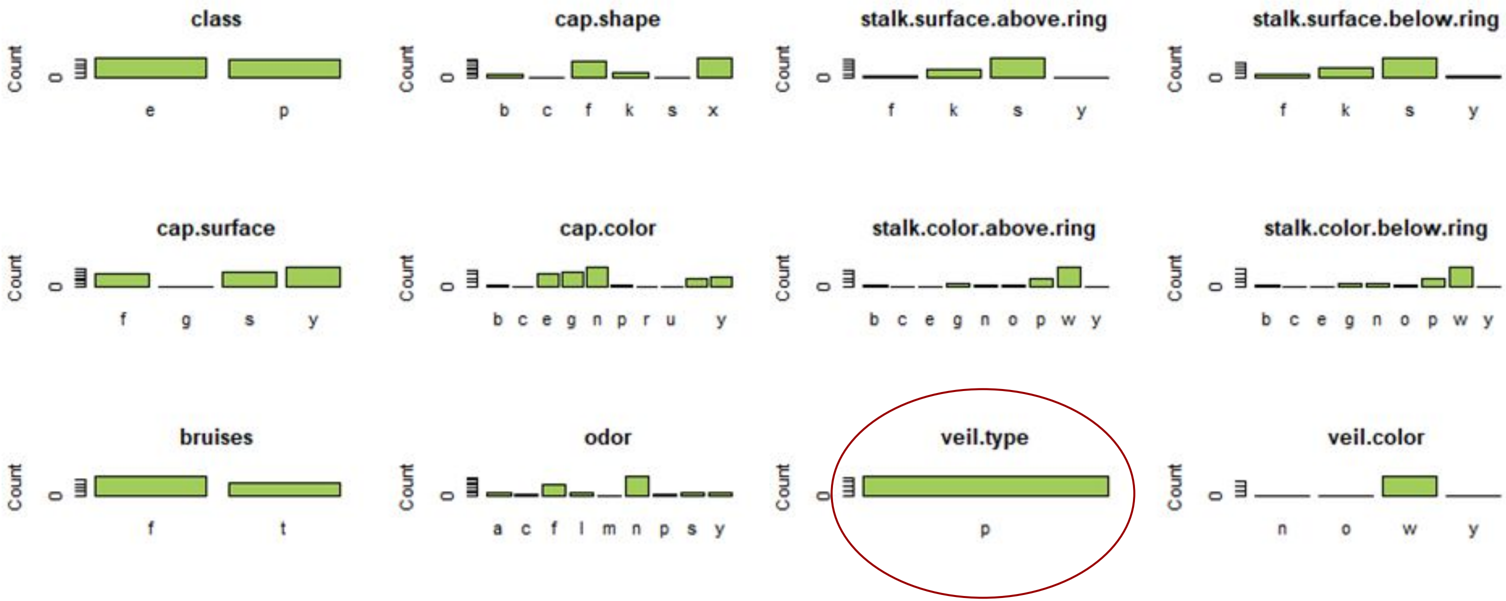
Data Attribute Information

Class: e (edible) | p (poisonous)

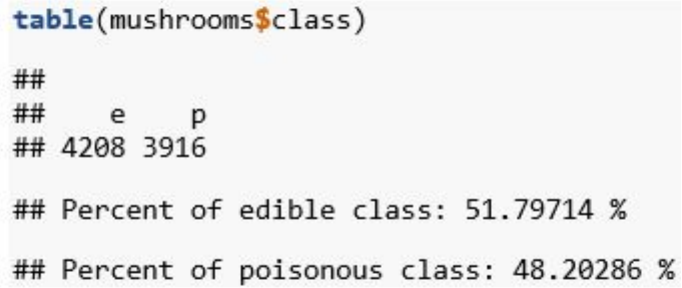
23 Variables:

Cap Shape/Surface/Color
Bruises
Odor
Gill Attachment/Spacing/Size/Color
Stalk Shape/Root/Surface/Color
Veil Type/Color
Ring Number/Type
Spore Color
Population
Habitat

Data Exploration



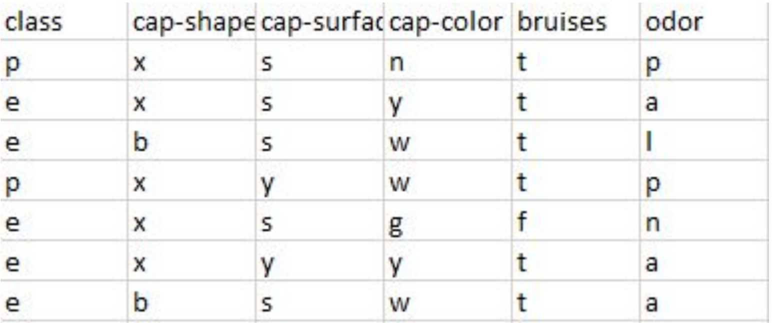
Data Exploration



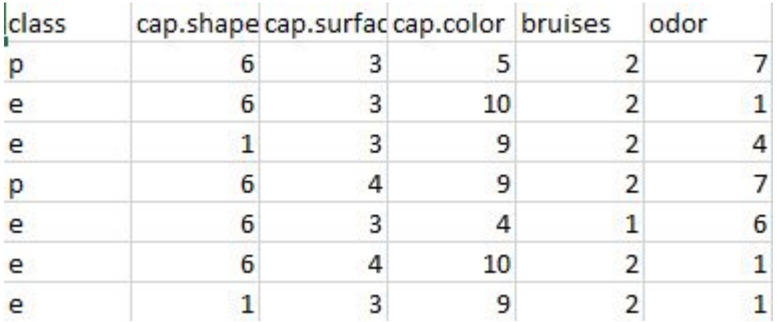
```
table(mushrooms$class)

## 
##    e    p 
## 4208 3916

## Percent of edible class: 51.79714 %

## Percent of poisonous class: 48.20286 %
```

Transforming Data

Change data from letters to numbers in order to feed into models



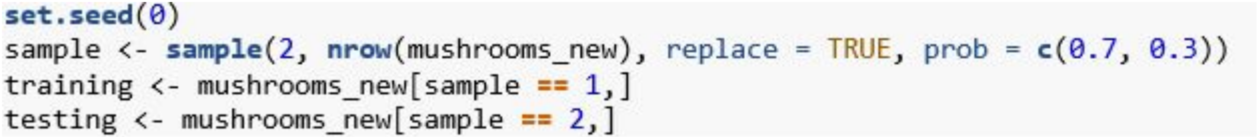
class	cap-shape	cap-surfac	cap-color	bruises	odor
p	x	s	n	t	p
e	x	s	y	t	a
e	b	s	w	t	l
p	x	y	w	t	p
e	x	s	g	f	n
e	x	y	y	t	a
e	b	s	w	t	a



class	cap.shape	cap.surfac	cap.color	bruises	odor
p	6	3	5	2	7
e	6	3	10	2	1
e	1	3	9	2	4
p	6	4	9	2	7
e	6	3	4	1	6
e	6	4	10	2	1
e	1	3	9	2	1

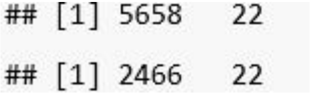
Training & Testing Sets

70/30 Split



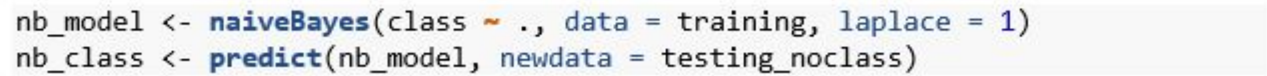
```
set.seed(0)
sample <- sample(2, nrow(mushrooms_new), replace = TRUE, prob = c(0.7, 0.3))
training <- mushrooms_new[sample == 1,]
testing <- mushrooms_new[sample == 2,]
```

Dimensions

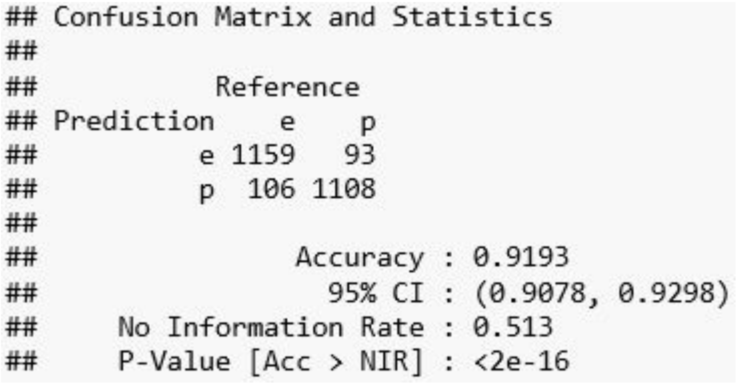


```
## [1] 5658   22

## [1] 2466   22
```

Model 1 - Naive Bayes

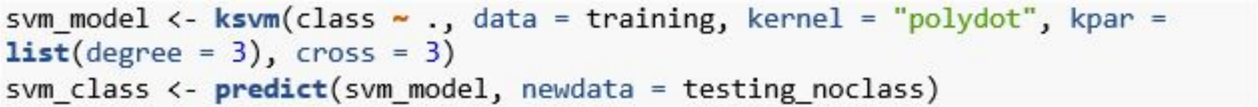


```
nb_model <- naiveBayes(class ~ ., data = training, laplace = 1)
nb_class <- predict(nb_model, newdata = testing_noclass)
```

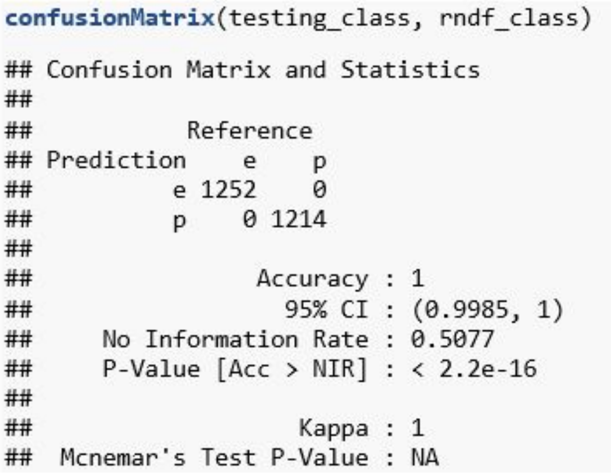


```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    e    p
##          e 1159   93
##          p  106 1108
##
##                Accuracy : 0.9193
##                  95% CI : (0.9078, 0.9298)
##     No Information Rate : 0.513
##     P-Value [Acc > NIR] : <2e-16
```

Model 2 - SVM

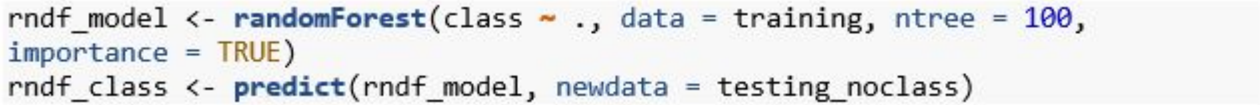


```
svm_model <- ksvm(class ~ ., data = training, kernel = "polydot", kpar =
list(degree = 3), cross = 3)
svm_class <- predict(svm_model, newdata = testing_noclass)
```

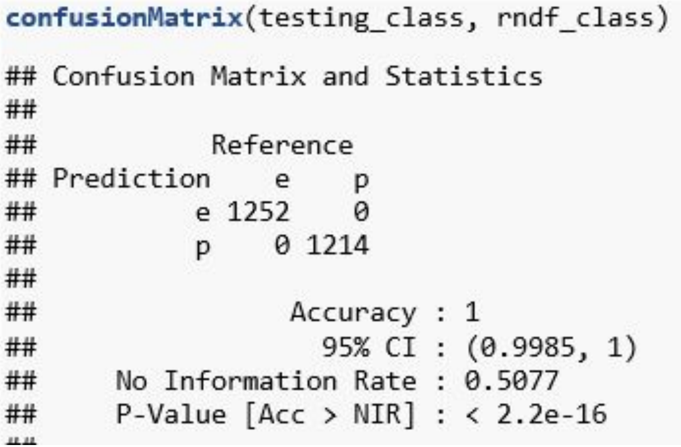


```
confusionMatrix(testing_class, rndf_class)

## Confusion Matrix and Statistics
##
##           Reference
## Prediction    e    p
##          e 1252    0
##          p    0 1214
##
##                Accuracy : 1
##                  95% CI : (0.9985, 1)
##     No Information Rate : 0.5077
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 1
##  Mcnemar's Test P-Value : NA
```

Model 3 - Random Forest

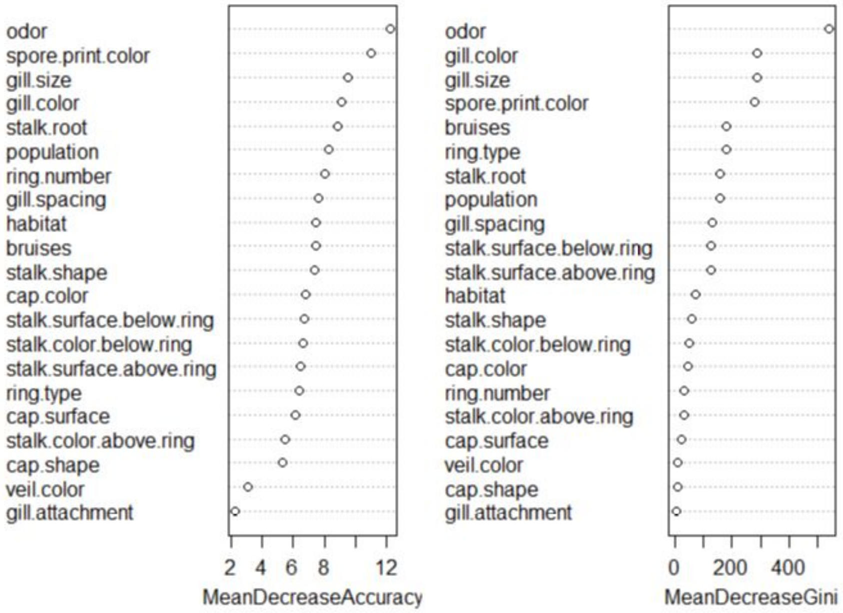


```
rndf_model <- randomForest(class ~ ., data = training, ntree = 100,
importance = TRUE)
rndf_class <- predict(rndf_model, newdata = testing_noclass)
```



```
confusionMatrix(testing_class, rndf_class)

## Confusion Matrix and Statistics
##
##           Reference
## Prediction    e    p
##          e 1252    0
##          p    0 1214
##
##                Accuracy : 1
##                  95% CI : (0.9985, 1)
##     No Information Rate : 0.5077
##     P-Value [Acc > NIR] : < 2.2e-16
##
```

Model 3 - Variable Importance



Findings

- Veil type was the same among the entire mushroom data population.
- Odor, Gill Size, and Spore Print Color are top indicators per Random Forest.
- Naive Bayes classification model produced an outcome of 92% accuracy. SVM and Random Forest models produced an outcome of 100% accuracy.

References

Image: VectorStock | https://www.vectorstock.com/royalty-free-vector/edible-mushrooms-vector-827231