Sean McLean

ALY 6040

Professor Reilly

Module 2

Technique Practice

Introduction

The dataset uploaded into R shows characteristics of mushrooms that people forage for consumption or to give to restaurants for purchase. The assignment's mission is to practice data mining of the Excel dataset and to study relationships in the data. After the analysis, the goal is to formulate outcomes and conclusions for stakeholders from the data. This will hopefully provide more detail into what mushrooms are okay to devour and sell and what mushrooms should be avoided.

Code Walk Through

The dataset consists of 23 attributes and 8,124 cells of data that describe the characteristics of mushrooms. All the elements in each attribute in the dataset are character strings. The str () function below displays the structure of the dataset:

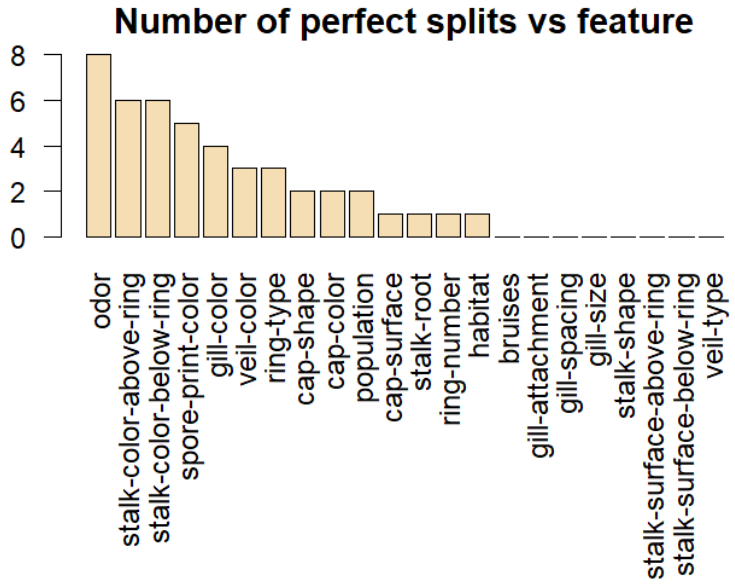
```
> str(mushrooms)
tibble [8,124 × 23] (S3: tbl_df/tbl/data.frame)
 $ class                   : chr [1:8124] "p" "e" "e" "p" ...
 $ cap-shape               : chr [1:8124] "x" "x" "b" "x" ...
 $ cap-surface             : chr [1:8124] "s" "s" "s" "y" ...
 $ cap-color               : chr [1:8124] "n" "y" "w" "w" ...
 $ bruises                 : chr [1:8124] "t" "t" "t" "t" ...
 $ odor                    : chr [1:8124] "p" "a" "l" "p" ...
 $ gill-attachment         : chr [1:8124] "f" "f" "f" "f" ...
 $ gill-spacing            : chr [1:8124] "c" "c" "c" "c" ...
 $ gill-size               : chr [1:8124] "n" "b" "b" "n" ...
 $ gill-color              : chr [1:8124] "k" "k" "n" "n" ...
 $ stalk-shape             : chr [1:8124] "e" "e" "e" "e" ...
 $ stalk-root              : chr [1:8124] "e" "c" "c" "e" ...
 $ stalk-surface-above-ring: chr [1:8124] "s" "s" "s" "s" ...
 $ stalk-surface-below-ring: chr [1:8124] "s" "s" "s" "s" ...
 $ stalk-color-above-ring  : chr [1:8124] "w" "w" "w" "w" ...
 $ stalk-color-below-ring  : chr [1:8124] "w" "w" "w" "w" ...
 $ veil-type               : chr [1:8124] "p" "p" "p" "p" ...
 $ veil-color              : chr [1:8124] "w" "w" "w" "w" ...
 $ ring-number             : chr [1:8124] "o" "o" "o" "o" ...
 $ ring-type               : chr [1:8124] "p" "p" "p" "p" ...
 $ spore-print-color       : chr [1:8124] "k" "n" "n" "k" ...
 $ population              : chr [1:8124] "s" "n" "n" "s" ...
 $ habitat                 : chr [1:8124] "u" "g" "m" "u" ...
```

To check for null values the nrow() and sum () functions are used to calculate the total which came up as zero. The column ‘veil.type’ is removed from the dataset because of its redundancy, and ‘NULL’ function is used to complete this task. The odor variable is then analyzed by using the table () function with the class variable added to show the difference of elements between the ‘e’ and ‘p’ elements in the class attribute. The results from the table are shown below:

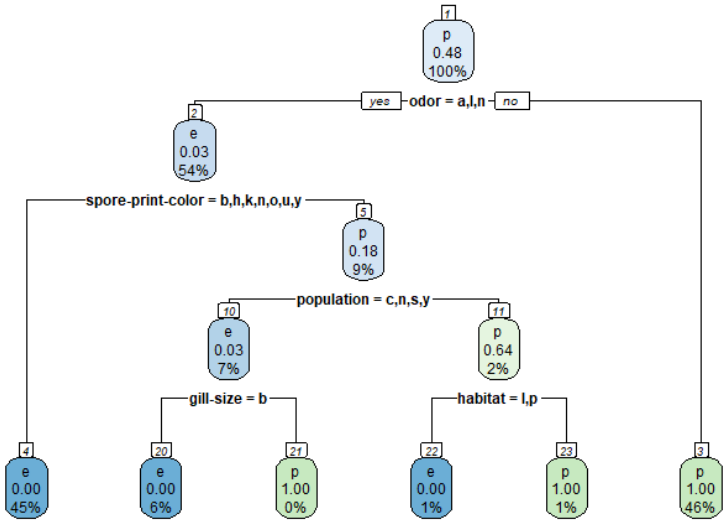
```
> table(mushrooms$class,mushrooms$odor)

     a    c    f    l    m    n    p    s    y
e  400    0    0  400    0 3408    0    0    0
p    0  192 2160    0   36  120  256  576  576
```

The apply () and order () functions are then utilized to select all the perfect splits for each attribute that could be used for a decision tree. A bar chart is then executed to show the perfect splits per variable in descending order.



Training and testing sets are used to divide the data to create a decision tree, and the set.seed() function is added to allow the algorithm to reproduce the values. The matrix () and rpart() functions are then utilized for building the decision tree and its classifications before the visualization is displayed.



To test the model and show how accurate the dataset is, the decision tree goes through the pruning process. This is performed using the cp () function, which is the complexity parameter,

and after that code has been executed the model can be tested by applying the predict () function before looking at the confusion matrix of the results.

```
> confusionMatrix(t)
Confusion Matrix and Statistics

   pred
      e   p
  e 829   0
  p   0 795

                Accuracy : 1
                  95% CI : (0.9977, 1)
    No Information Rate : 0.5105
    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 : 0.5105
         Detection Rate : 0.5105
   Detection Prevalence : 0.5105
      Balanced Accuracy : 1.0000

       'Positive' Class : e
```

Analysis

There is almost an even number of mushrooms in the two classes with the ‘e’ class having a few hundred more than the ‘p’ class, so there are less biases with the data being fairly equal in division. There are only three elements in the ‘e’ class with most of them being the ’n’ element. There are seven elements in the ‘p’ class with the most being the ‘f’ element.’ The proper code is used to provide a breakdown of the perfect splits per variable which is then shown in a bar chart in descending order. The odor attribute has the most perfect splits with eight and the stalk-color-above-ring and stock-color-below-ring both have six perfect splits. Of the 22 features in the dataset there are eight that do not have any perfect splits.

The decision tree is built from the matrix, classification system, and training and test sets and the odor variable is the beginning of the decision tree as the root node based off the algorithms. The two options from the variable are yes and no, and on the ‘yes’ side a decision node is created as the spore-print-color attribute which then based off its options have more decision nodes created with the variable population which then creates another decision node with the gill-size and habitat variables. The ‘no’ side of the odor variable ends up being a terminal node in the decision

tree. After pruning the tree, the confusion matrix statistics show that there are 829 true negatives, 795 true positives, and zero false negatives and false positives. The accuracy and confidence interval are close to one and contains an incredibly low p-value. The sensitivity, specificity, positive predictive value, negative predictive value, and balanced accuracy performed well with values at one, and the prevalence, detection rate, and detection prevalence metrics were all balanced at 0.5105.

Interpretations & Recommendations

The odor variable is the best root node to use from the computed algorithm and has the most perfect splits that makes it more suitable for the decision tree. There are slightly more non-poisonous mushrooms in the dataset, and almost all the odor non-poisonous mushrooms with spore-print-color are also non-poisonous. At 45 percent, it would be recommended to collect mushrooms with odor and have the spore-print-color because they are not poisonous to consume or sell. The mushrooms that do not have an odor are at 46 percent and are poisonous to consume, which serves as a major caution to people that are amateur mushroom collectors.

There are no false positives or false negatives according to the confusion matrix, so the dataset correctly predicted all its cells. The metrics that pertain to accuracy and prediction equaled one which indicate perfection to the dataset. The detection rates and prevalence were 0.5105, showing that the dataset was close to being balanced between both elements in the class variable. A low p-value below 0.05 means there is a difference between the elements in the class variable if hypothesis testing was incorporated into this dataset analysis. Overall, I would recommend using the decision tree as a tool to show the difference between edible and poisonous mushrooms and their characteristics, so it shows which ones to pick that are safe to consume and sell.

Conclusion

From uploading the data and then conducting data mining there are strong aspects of the dataset that will provide the necessary information about key elements of mushrooms. Through the usage of training and testing sets a decision tree is the best data visualization tool to show the major differences between mushrooms. Using a confusion matrix to verify this information is critical so it assures the consumer that they are choosing mushrooms that will not affect their health or someone else’s well-being.

References

Kabacoff, R. (2015). *R In Action. Data analysis and graphics with R. Second Edition.* (pp. 105, 394-395). Manning.