Module 2 – Technique Practice

Name: Aishwarya Doni

College of Professional Studies, Northeastern University

ALY6040: Data Mining
Professor: Yesha Vora

April 19, 2025

Mushroom Classification Using Decision Trees

Introduction

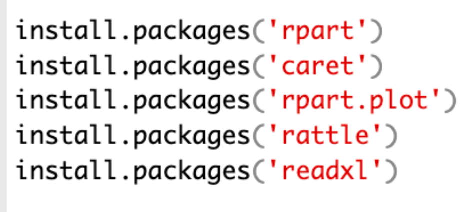
Mushroom foraging is both a recreational and commercial activity practiced globally. However, due to the subtle visual similarities between edible and poisonous mushrooms, foraging can pose serious health risks if misidentifications occur. While expert foragers rely on years of experience, data-driven tools can aid novices by identifying mushrooms' edibility based on observable features.

This report uses a decision tree algorithm implemented in R to classify mushrooms as either edible or poisonous using a dataset from UCI. We walk through the data mining process including loading, cleaning, splitting, modeling, and interpreting the results. The goal is to build an interpretable and accurate model and provide recommendations for practical mushroom classification systems.

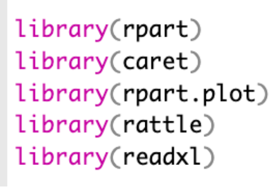
Code Walkthrough

This section explains how the R code transforms and analyzes the mushroom dataset.

1. **Installing and Loading Libraries:**

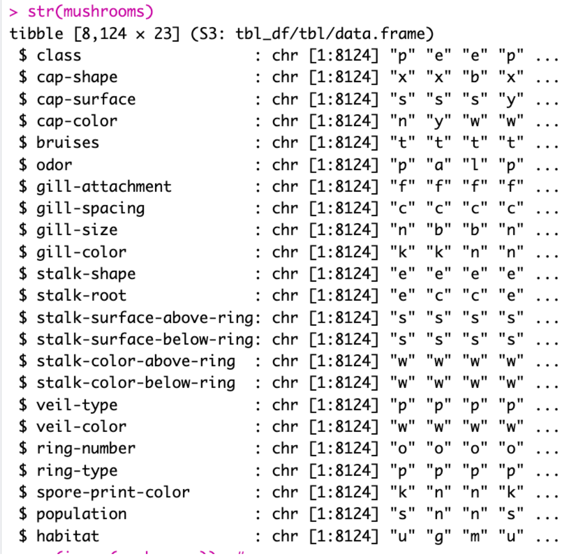


```r
install.packages('rpart')
install.packages('caret')
install.packages('rpart.plot')
install.packages('rattle')
install.packages('readxl')
```



```r
library(rpart)
library(caret)
library(rpart.plot)
library(rattle)
library(readxl)
```

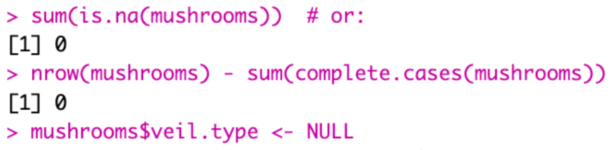
2. **Reading the Dataset:**



```
> 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" ...
```

The dataset has 8,124 rows and 23 columns. Each row represents a mushroom and the columns represent physical characteristics. The class column indicates whether the mushroom is edible (e) or poisonous (p).

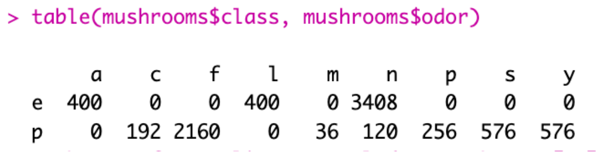
3. Cleaning the Dataset:



```
> sum(is.na(mushrooms))  # or:
[1] 0
> nrow(mushrooms) - sum(complete.cases(mushrooms))
[1] 0
> mushrooms$veil.type <- NULL
```

There are no missing values. The column veil.type was removed as it contains only one value and adds no variability.

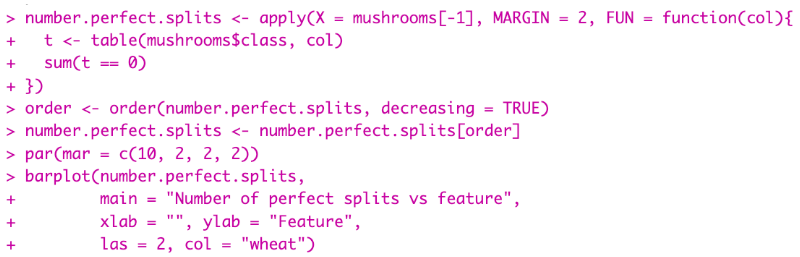
4. Exploratory Analysis:



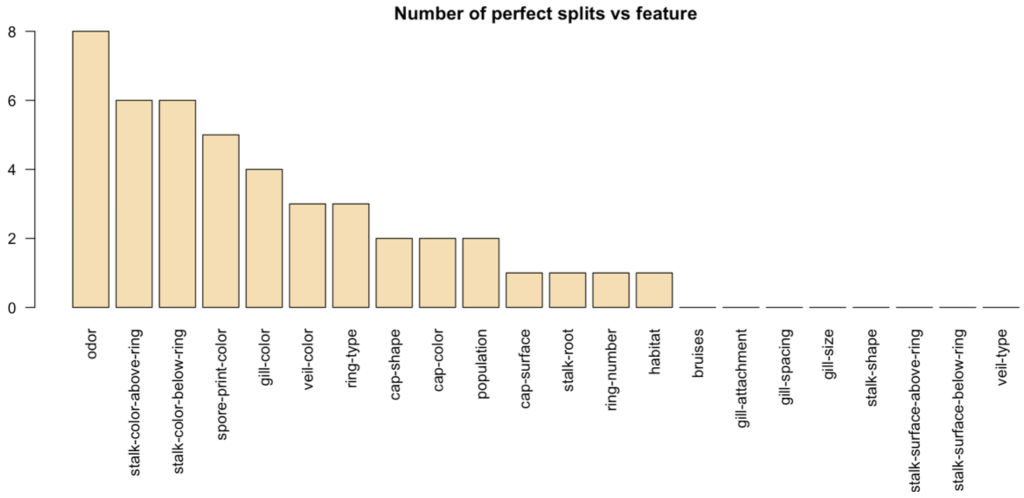
```
> table(mushrooms$class, mushrooms$odor)

       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 output clearly shows that odor is a strong indicator of mushroom edibility. For example, mushrooms with odor ‘p’(foul) are all poisonous, while those with 'a' (almond) are all edible.

5. Perfect Splits Calculation



```
> number.perfect.splits <- apply(X = mushrooms[-1], MARGIN = 2, FUN = function(col){
+   t <- table(mushrooms$class, col)
+   sum(t == 0)
+ })
> order <- order(number.perfect.splits, decreasing = TRUE)
> number.perfect.splits <- number.perfect.splits[order]
> par(mar = c(10, 2, 2, 2))
> barplot(number.perfect.splits,
+         main = "Number of perfect splits vs feature",
+         xlab = "", ylab = "Feature",
+         las = 2, col = "wheat")
```



This code calculates the number of perfect splits (where one feature value only occurs in one class) for each variable and visualizes it.

Analysis of Output

From the bar plot, we observe that odor, spore.print.color, and gill-color have the highest number of perfect splits, making them the most important predictors.

1. **The dataset was split:**

```
> set.seed(12345)
> train <- sample(1:nrow(mushrooms), size = ceiling(0.80 * nrow(mushrooms)), replace = FALSE)
> mushrooms_train <- mushrooms[train, ]
> mushrooms_test <- mushrooms[-train, ]
```

2. **A penalty matrix was introduced:**

```
> penalty.matrix <- matrix(c(0, 1, 10, 0), byrow = TRUE, nrow = 2)
```

This penalizes classifying a poisonous mushroom as edible more heavily.

3. **Decision tree was trained and pruned:**

```
> tree <- rpart(class ~ .,
+               data = mushrooms_train,
+               parms = list(loss = penalty.matrix),
+               method = "class")
> rpart.plot(tree, nn = TRUE)
> cp.optim <- tree$cptable[which.min(tree$cptable[,"xerror"]), "CP"]
> tree <- prune(tree, cp = cp.optim)
```

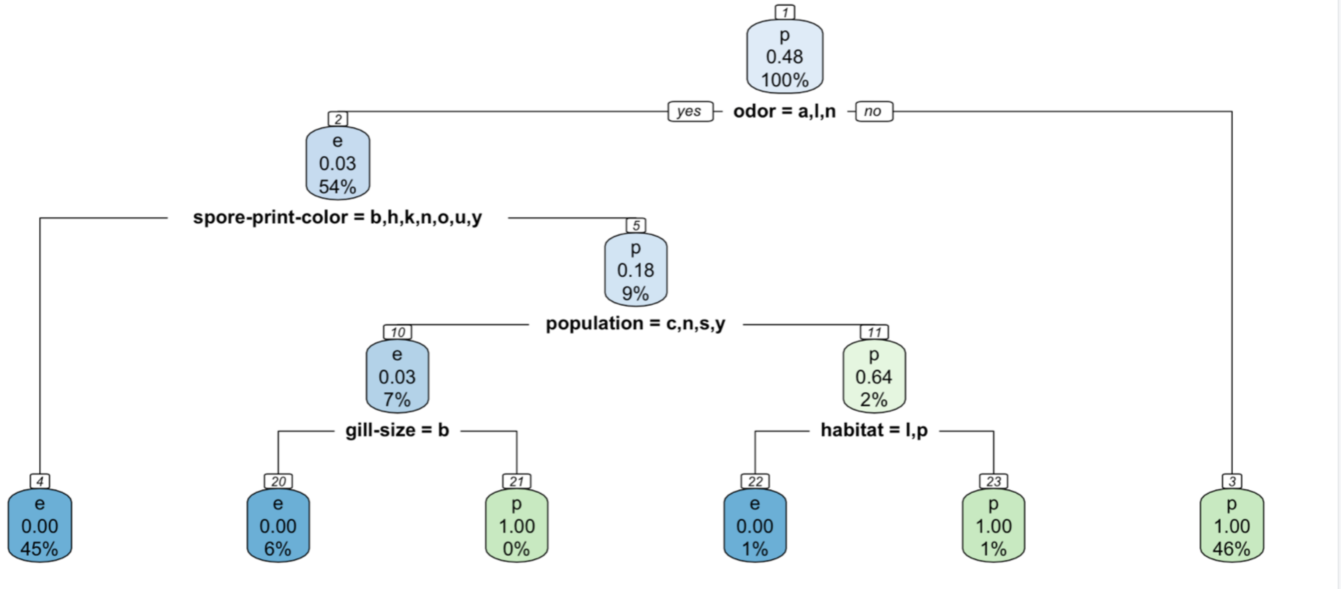


Figure 1: Pruned Decision Tree for Mushroom Classification

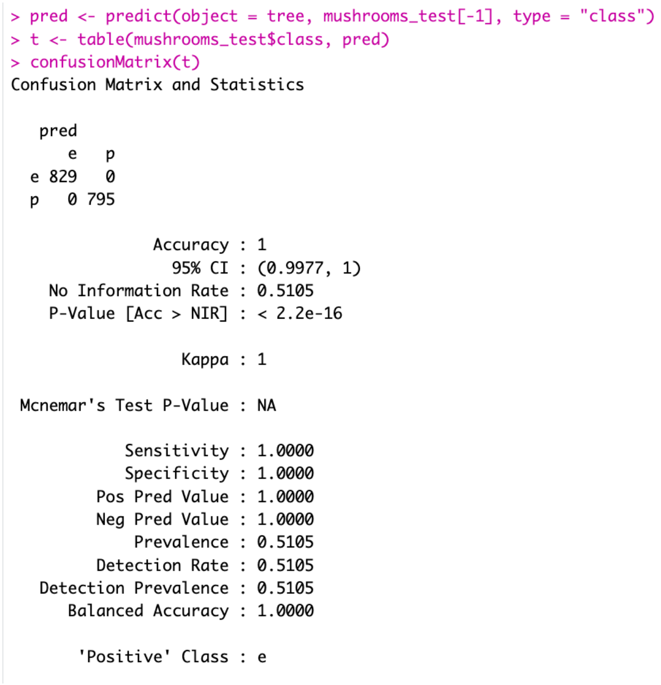
The pruned decision tree (Figure 1) demonstrates how the model uses hierarchical logic to classify mushrooms as edible (e) or poisonous (p). The most significant splitting variable is odor, which appears at the root node. If a mushroom's odor is almond (a), anise (l), or none (n), the tree classifies it as potentially edible and proceeds to evaluate other variables like spore-print-color, population, gill-size, and habitat.

Key insights:

- Odor = p, f, y, s, etc. → always poisonous (100%**).**
- For mushrooms with "safe" odors (a, l, n), additional checks are done. For example:
    a) If spore-print-color ∈ {b,h,k,n,o,u,y} and population ∈ {c,n,s,y}, the mushroom is classified based on gill-size and habitat.
    b) When gill-size = b, mushrooms are **poisonous**.
    c) When habitat ∈ {l, p}, mushrooms are **poisonous**.

This hierarchy provides a transparent, rule-based method that non-experts could follow.

4. Prediction and Evaluation:



```
> pred <- predict(object = tree, mushrooms_test[-1], type = "class")
> t <- table(mushrooms_test$class, pred)
> 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
```

Figure 2: Confusion Matrix and Performance Metrics of the Decision Tree Model

The confusion matrix and associated statistics (Figure 2) show that the trained and pruned decision tree model achieved perfect classification on the test data.

Summary of results:

- **Accuracy**: 100%
- **Sensitivity (Recall for edible mushrooms)**: 1.000
- **Specificity (Recall for poisonous mushrooms)**: 1.000
- **Kappa**: 1 (perfect agreement beyond chance)
- **No misclassifications**: 829 edible and 795 poisonous mushrooms correctly classified.

These metrics indicate that:

- The model perfectly distinguished between edible and poisonous mushrooms in the test set.
- There were no false positives or false negatives, which is critical when misclassifying a poisonous mushroom could have severe health consequences.

These results highlight the reliability and real-world applicability of the decision tree in mushroom classification tasks.

Conclusion

This study demonstrates the effectiveness of decision trees as a powerful and interpretable data mining tool for classifying mushrooms based on their physical characteristics. Using a well-structured dataset, we trained and pruned a decision tree model that achieved perfect accuracy on the test set, correctly classifying all mushrooms as either edible or poisonous.

The analysis revealed that **odor** is the most influential variable, followed by **spore-print-color**, **population**, and **gill-size**. The interpretability of the tree makes it particularly useful for real-world applications such as mobile apps or field guides for mushroom foragers. Furthermore, the use of a penalty matrix ensured that the model prioritizes safety by heavily penalizing the misclassification of poisonous mushrooms.

Future research could incorporate environmental and seasonal variables, allowing more refined classifications, such as differentiating between culinary and medicinal mushrooms. Overall, this exercise not only showcases the power of data mining but also contributes to public health and safety through practical application.

References

1. UCI Machine Learning Repository. (n.d.). *Mushroom Data Set*. Retrieved from https://archive.ics.uci.edu/ml/datasets/mushroom
2. Kuhn, M. (2022). *caret: Classification and Regression Training* (R package version 7.0-1). Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=caret
3. Therneau, T., Atkinson, B., & Ripley, B. (2023). *rpart: Recursive Partitioning and Regression Trees* (R package version 4.1-24). https://CRAN.R-project.org/package=rpart