# Assignment 2: Technique Practice on Mushroom Dataset

Yash S  
ALY6040 – Module 2  
Northeastern University: College of Professional Studies  
2025 – 04 – 23

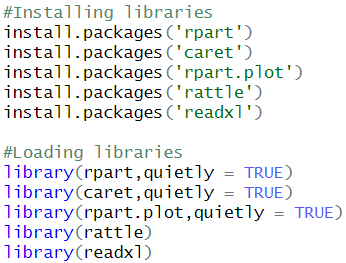
**Introduction**

**Introduction**  
This report outlines the development and deployment of a decision tree model aimed at classifying mushrooms as either edible (e) or poisonous (p). The model leverages 22 observable morphological features to enhance the identification process. The key motivation behind this initiative is to aid mushroom foragers in distinguishing between safe and toxic varieties, with a strong focus on minimizing false negatives—instances where a poisonous mushroom could be misclassified as edible, potentially leading to severe consequences.

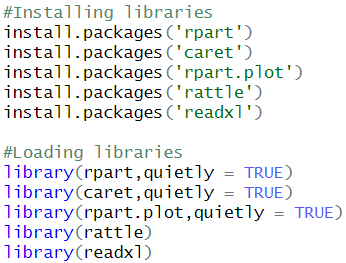
**Dataset Overview**  
The dataset utilized comprises 8,124 mushroom samples, each described by 22 categorical features capturing physical characteristics such as cap shape, gill dimensions, surface color, and odor. Every entry is clearly labeled as edible or poisonous, offering a well-structured foundation for classification modeling. The dataset is notably clean, with no missing values and entirely categorical variables, eliminating the need for numeric transformations. One feature, ‘veil.type’, was removed from the analysis due to its uniformity across all records, contributing no value to the model. This dataset presents a practical and critical use case for decision tree algorithms, emphasizing not only classification accuracy but also the imperative to reduce high-risk misclassifications that may impact human health.

## Data Preparation

**Cleaning and Preprocessing**  
The initial phase of the analysis involves setting up the R environment by installing and loading the required libraries. These libraries are fundamental for efficient data handling, model construction, and visual representation. They provide the core functionality needed to prepare the dataset and implement the decision tree approach adopted for this classification task.



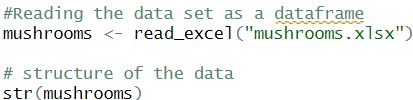
To maintain a clean and readable console output, the **quietly = TRUE** parameter is applied when loading essential libraries such as rpart and caret. This approach suppresses non-critical messages during the loading process, allowing for a more streamlined and focused analysis environment.



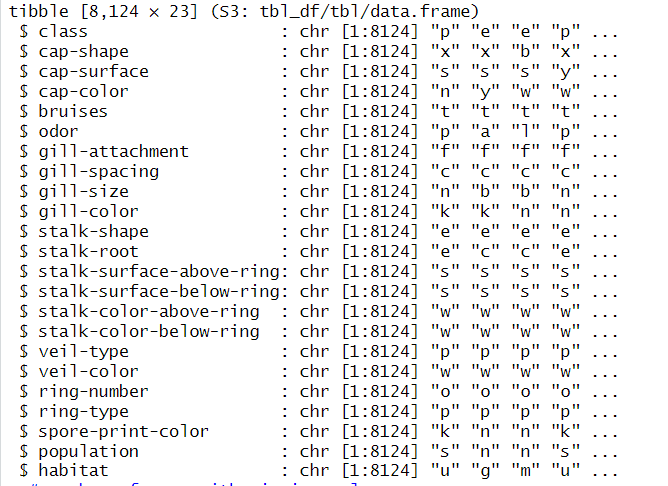
This methodical setup creates a robust foundation for conducting thorough data analysis and facilitates the efficient application of decision tree modeling techniques.

### Cleaning and Preprocessing

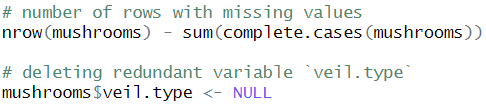
**Code:**



**Output:**



**Code:**

****

**Output:**

****

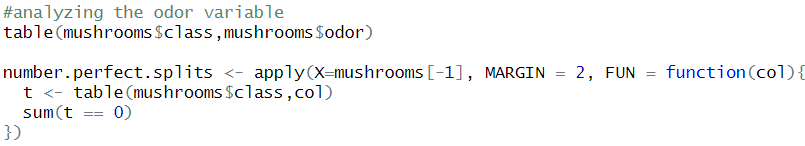
**Explanation:**

* The dataset comprises 8,124 entries and 23 features, each detailing specific characteristics of mushroom specimens.
* An inspection using the str(mushrooms) function reveals that all variables are categorical, capturing traits such as cap shape and odor.
* A completeness check using complete.cases() confirms the absence of missing values, indicating a fully intact dataset.
* The feature *veil.type* was removed from the analysis due to its uniformity, as it contained only a single repeated value ("p"), contributing no useful information for classification.

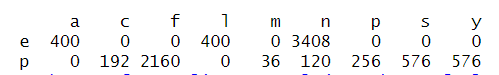
**Data Analysis**

## Explanatory Data Analysis (EDA)

**Code:**



**Output:**



**Explanation:**

The table above presents a cross-tabulation between mushroom odor (represented by single-letter codes) and their classification as edible or poisonous. Each cell indicates the number of mushrooms exhibiting a particular odor and their corresponding classification.

Edible mushrooms are exclusively associated with three odors:

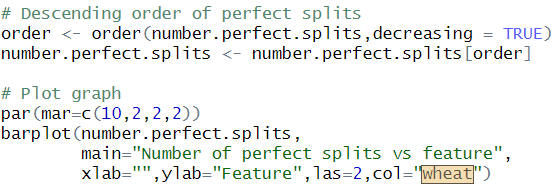
* **a** (almond): 400 instances
* **l** (anise): 400 instances
* **n** (none): 3,408 instances  
  These three odors collectively account for all edible samples in the dataset, with no instances appearing in the poisonous category. Their exclusivity makes them highly reliable indicators of edibility.

In contrast, poisonous mushrooms are linked to a broader range of odors, including:

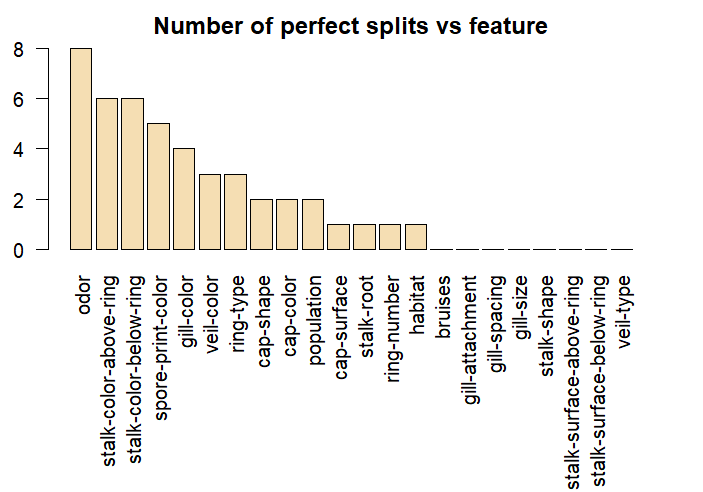
* **c** (creosote): 192 instances
* **f** (foul): 2,160 instances
* **m** (musty): 36 instances
* **n** (none): 120 instances
* **p** (pungent): 256 instances
* **s** (spicy): 576 instances
* **y** (fishy): 576 instances

This distinct separation highlights odor as a critical variable in the classification process. Specifically, the presence of almond, anise, or no odor strongly indicates that a mushroom is safe to consume, while any deviation from these odors is predominantly associated with toxicity.

**Code:**

****

**Output:**



**Explanation:**

**Perfect Splits Analysis: Feature Importance in Decision Tree Modeling**

* **X-axis (Features):** Represents the various categorical attributes in the dataset, such as odor, gill color, and stalk color characteristics.
* **Y-axis (Perfect Splits Count):** Denotes the frequency with which each feature created a perfect split in the decision tree—i.e., when the feature entirely separates the data into homogeneous class groups.

**Most Informative Features:**

* **Odor** emerges as the most influential variable, responsible for **8 perfect splits**, underscoring its strong discriminative power.
* **Stalk-color-above-ring** and **stalk-color-below-ring** also demonstrate significant impact, each contributing to **6 perfect splits**. These features are instrumental in enhancing classification accuracy.

**Least Informative Features:**

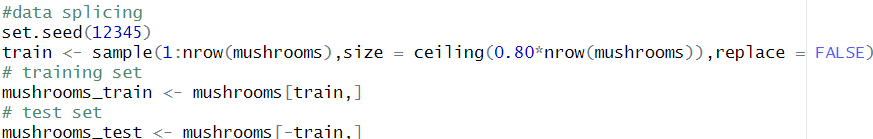
* Attributes such as **veil-type**, **stalk-surface-below-ring**, and **stalk-surface-above-ring** did not produce any perfect splits, indicating minimal relevance in separating edible and poisonous mushrooms within the model.

**Summary of Feature Influence (Bar Plot Interpretation):**

* **Top Contributors:** Odor, stalk-color-above-ring, spore-print-color, and gill-color are among the most effective in producing clean splits.
* **Excluded Feature:** *Veil-type* was removed due to its lack of variability, providing no meaningful distinction for classification.

### Model Training & Evaluation

### Code:

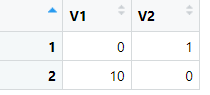


**Explanation:**

* The dataset is partitioned into 80% training and 20% testing subsets, using a fixed random seed (12345) to ensure consistency and reproducibility of results across multiple runs.

**Code:**

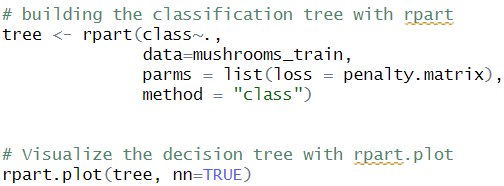
****



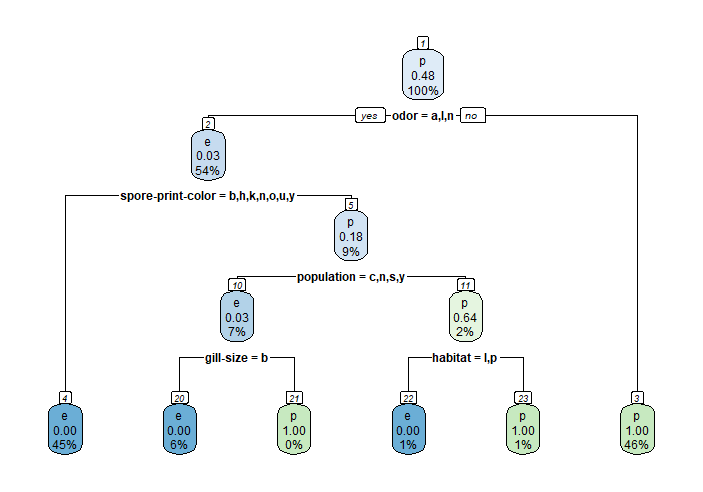
**Explanation:**

• A penalty cost of 10 is assigned to false negatives to emphasize the severity of misclassifying a poisonous mushroom as edible, reinforcing the model's focus on safety and risk mitigation.

**Code:**



**Output:**

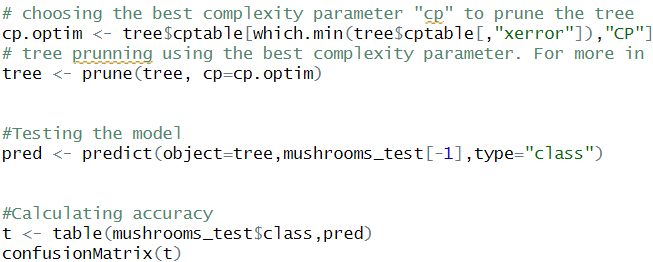


**Explanation:**

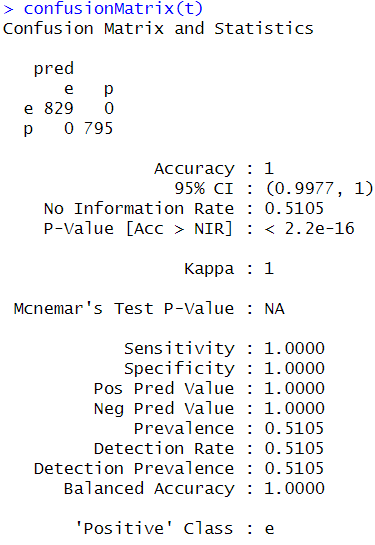
**Decision Tree Structure Overview**

* **Primary Split:** Mushrooms emitting almond, anise, or no odor are classified as edible. All other odors are treated as indicators of potential toxicity.
* **Secondary Split:** For mushrooms initially classified as edible, *spore print color* serves as a validation layer to reinforce or challenge the initial classification.
* **Tertiary Split:** Rare or atypical spore print colors prompt a shift in classification towards poisonous.
* **Quaternary Split:** The *population* attribute is then evaluated to further distinguish between safe and unsafe specimens.
* **Fifth Level:** A broad gill size generally supports an edible classification, whereas narrow gills are more commonly associated with poisonous varieties.
* **Final Split:** Habitat type plays a confirming role—mushrooms located in leafy or path-like environments are more frequently toxic, providing an additional cue for classification.

**Code:**

****

**Output:**

****

**Explanation:**

**Model Performance Summary**

* **Classification Results:** The model accurately classified all 829 edible and 795 poisonous mushrooms, achieving a perfect prediction rate with no errors.
* **Accuracy:** A flawless accuracy score of **1.00** was attained, demonstrating complete alignment between predicted and actual classes.
* **Confidence Interval:** The model’s performance is statistically robust, with a **95% confidence interval** ranging from **0.9977 to 1.00**.
* **Baseline Comparison:** A simplistic model predicting only the majority class would yield an accuracy of **51.05%**, highlighting the superior performance of the decision tree.
* **Statistical Significance:** A **p-value < 2.2e-16** confirms that the model's performance is significantly better than random chance.
* **Kappa Statistic:** A **kappa score of 1.0** indicates perfect agreement between observed and predicted classifications.
* **McNemar’s Test:** Not applicable, as there were no misclassifications.
* **Sensitivity (Recall):** A value of **1.0** indicates that all edible mushrooms were correctly identified.
* **Specificity:** A value of **1.0** confirms that every poisonous mushroom was accurately detected.
* **Precision (Positive Predictive Value):** All mushrooms predicted as edible were truly edible, resulting in **100% precision**.
* **Negative Predictive Value:** All mushrooms predicted to be poisonous were correctly classified, confirming the model’s reliability.
* **Prevalence:** Edible mushrooms represented **51.05%** of the test dataset.
* **Detection Rate:** All edible cases were accurately detected, consistent with their prevalence in the data.
* **Detection Prevalence:** The predicted proportion of edible mushrooms precisely matched their actual occurrence.
* **Balanced Accuracy:** The model achieved perfect balance, effectively identifying both edible and poisonous classes.
* **Target Class:** Performance metrics were calculated with edible mushrooms as the primary class of interest.

**Interpretation**

The decision tree model demonstrates exceptional precision in classifying mushrooms based on observable characteristics. Its performance suggests strong potential for practical use in identification tasks. However, due to the deterministic nature of decision trees, caution is advised when applying the model to novel or inconsistent input data, where generalizability may be limited.

**Recommendations**

* **Practical Application:** The model can be integrated into a mobile application for mushroom foragers. Enhancements such as image recognition and GPS functionality could increase identification reliability in field conditions.
* **Contextual Enhancement:** Incorporating additional environmental variables, such as seasonal growth patterns and regional habitat data from open sources, may improve the model’s adaptability and accuracy.

**Conclusion**

**Conclusion**

The decision tree model achieved perfect classification performance by leveraging key features such as odor and stalk coloration. While these results are promising, further validation in real-world settings is essential. In the short term, the model can serve as an educational tool for mushroom identification. In the long term, it holds potential for contributing to ecological research and biodiversity documentation.

**Works Cited**

## References

- Bluman, A. (2018). Elementary Statistics: A Step-by-Step Approach (10th ed.). McGraw-Hill.  
- Kabacoff, R. (2022). R in Action: Data Analysis and Graphics with R and tidyverse (3rd ed.). Manning Publications.  
- R Core Team (2024). R: A language and environment for statistical computing.  
- Therneau, T., & Atkinson, B. (2023). rpart: Recursive Partitioning and Regression Trees.  
- Dua, D., & Graff, C. (2019). *Mushroom Data Set*. UCI Machine Learning Repository. <https://archive.ics.uci.edu/ml/datasets/Mushroom>