INFS 5116 DATA VISUALISATION SP5 2023
Visualisation Project Plan

Visualizing and Classifying Mushroom Edibility: A Dual Approach Using Data Visualization and Machine Learning

Enna huahy057@mymail.unisa.edu.au

1. Introduction

High-dimensional data presents a paradox: while it confounds human cognitive faculties, it offers rich material for machine learning algorithms, often at the expense of interpretability. To tackle this, the project explores mushroom edibility classification through two distinct paradigms: direct data visualization and machine-driven analytical visualization. These paradigms diverge in two key aspects: speed of analysis and interpretability. Direct data visualization relies on innate human pattern recognition abilities but is limited in scalability. Conversely, machine-driven visualization allows for rapid, scalable analysis but may compromise on interpretability. This project aims to weigh the pros and cons of each approach and scrutinize how human cognition fares in comparison to algorithmic computations.

> While this project focuses primarily on visualization techniques for mushroom edibility, it's worth considering the broader questions surrounding human versus machine analysis. For instance, could a mycelium computer outperform a traditional silicon-based one, or is human thought inherently more nuanced than algorithmic computations? Though outside the scope of this research, these thoughts illuminate the tension between human cognition and machine learning, particularly in the context of text and image classification through large language models and computer vision algorithms. These questions underline the complexities and ethical considerations involved in data science projects that strive for both speed and interpretability.

---

2. Data Source

The project leverages two distinct mushroom datasets, each catering to a different visualization paradigm, to dissect edibility classification.

- **Primary Dataset**: This dataset consists of 177 observations featuring 24 nominal and metrical variables such as `cap.size` and `cap.shape`. Sourced from a textbook, it is less structured, with single cells containing multiple values and missing entries. These irregularities pose challenges for both visualization approaches and machine learning algorithms (Wagner et al., 2023).
- **Secondary Dataset**: Comprising 61,069 observations across 16 variables (Wagner et al., 2023). It is well-structured, thus ideal for machine learning and subsequent analytical visualizations. The distribution shape is similar to the primary data, with both datasets having roughly 44.5% edible and 55.5% poisonous mushrooms. This indicates that while the raw counts differ, the relative proportions of edible and poisonous mushrooms are similar in both datasets.
- **Significance for Visualization**: Each dataset serves a unique purpose in the dual approach to visualization. The secondary dataset, with its structured format, is primed for machine-driven analysis and performance visualization. In contrast, the primary dataset requires extensive data cleaning and transformation to be compatible with both human and machine analysis.

Data Source: Secondary Mushroom Dataset (Wagner et al., 2023)

3. Data Preparation

To ensure that the datasets are primed for advanced analysis and machine learning algorithms, both primary and secondary datasets underwent rigorous data cleaning and transformation processes.

Primary Dataset

- **Missing Values**: Columns with missing values exceeding a threshold of 25% were omitted. The threshold was set at 25% to balance the trade-off between data integrity and usability.
- **Range Variables**: The variables `cap.size`, `stem.height`, and `stem.width` were decomposed into `min` and `max` components.
- **Standardization**: All numeric variables were standardized to a common unit, millimeters (mm).
- **Imputation**: Specific columns, such as `Cap.surface`, had missing values imputed with the mode.
- **One-hot Encoding**: Categorical variables underwent one-hot encoding, excluding unique identifiers and the target variable for machine learning tasks. The engineered variables were given distinct names to avoid confusion with the original variables.

Secondary Dataset Unlike the primary dataset, the secondary dataset did not contain any missing values, thus simplifying the data preparation process.

- **Standardization**: Numeric variables were standardized to millimeters (mm).
- **One-hot Encoding**: Similar to the primary dataset, categorical variables were one-hot encoded.

The objective of these preparations is to create cleaned, structured datasets optimized for future analyses, visualizations, and machine learning tasks.

---

4. Project Aims

This section delineates not only the technical ambit but also the social and ethical facets integral to the study.

Objective Hierarchy

- **Fundamental**: Evaluate variables contributing to mushroom edibility through statistical mapping and relational analysis.
- **Intermediate**: Perform a comparative analysis of machine learning algorithms and data visualization methodologies.
- **Comprehensive**: Investigate machine learning efficacy vis-à-vis human intuition and scrutinize ethical ramifications.

Scope and Methodologies

- The project is circumscribed by its focus on data visualization and machine learning, predominantly utilizing Python or R libraries for these tasks.

Success Indicators

- Algorithmic accuracy in edibility classification and the interpretability of visual elements will serve as performance metrics.

Ethical and Practical Impediments

- **Computational Strain**: Handling the complexity arising from high-dimensional data.
- **Ethical Constraints**: Emphasis on the significance of false positives and negatives, owing to their potential harmful effects, particularly in edibility determination.

---

5. Data Exploration

This section engages in an initial assessment of the data, employing both statistical and visual techniques. It serves as a precursor to the in-depth analysis proposed in Section 4.

1. Bar Plots for Class Distribution To understand the distribution of classes (edible, poisonous), bar plots have been generated for both the primary and secondary datasets. These plots provide a baseline view of class distribution within the datasets.

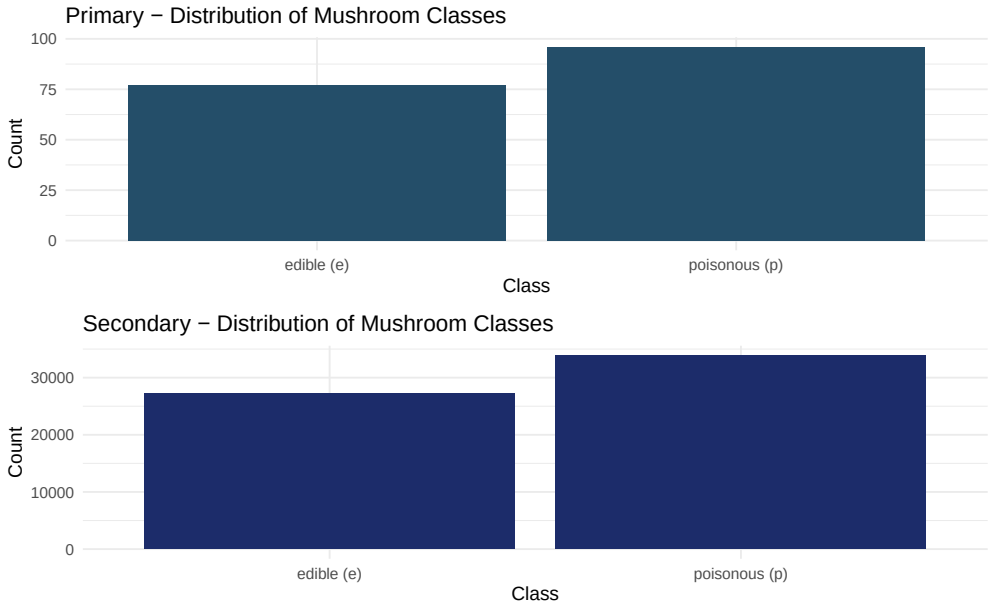

Figure 1. Barplot for class distribution in primary and secondary dataset.

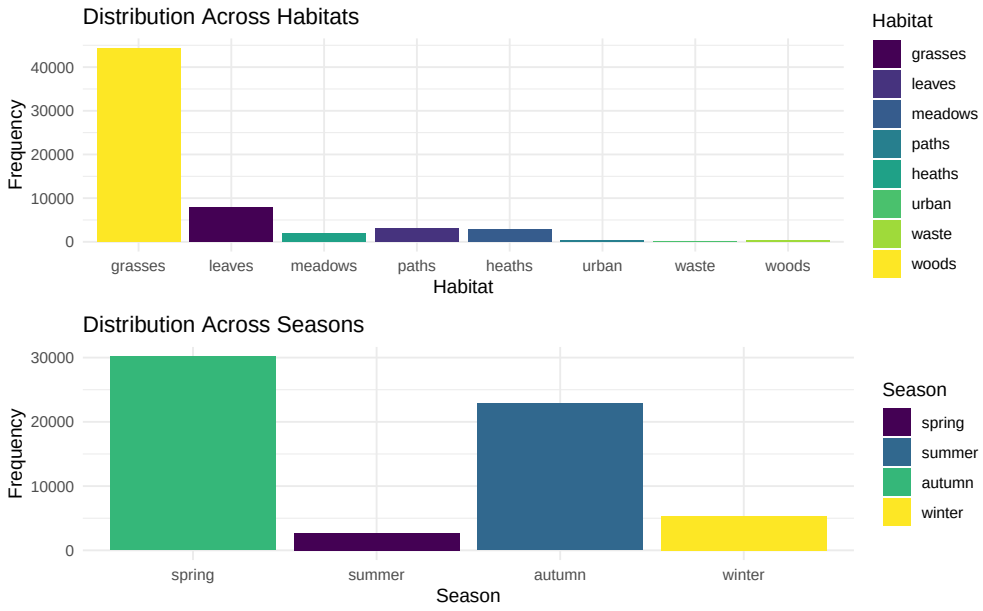

Figure 2. Distribution across habitat and season for the secondary dataset, these are the environmental attributes of the mushroom data.

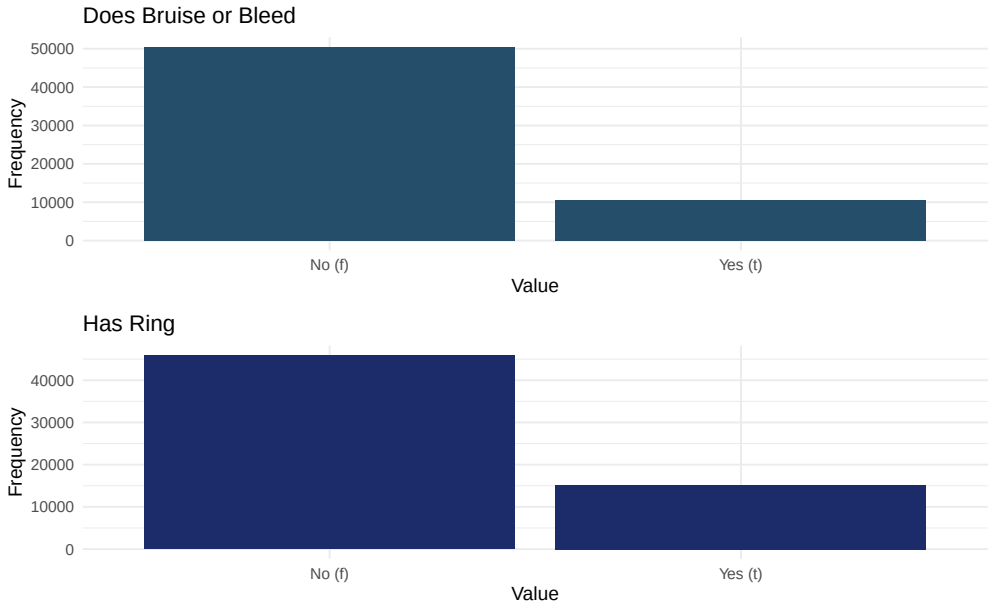

Figure 3. Binary morphological attributes for the secondary dataset.

2. Heatmap for Attribute Correlations A heatmap has been constructed to visualize the correlations between continuous features in the secondary dataset. This visualization aids in understanding how numerical attributes relate to each other.

Based on the correlation matrix and heatmap, it is observed that "cap.diameter" and "stem.width" have a strong positive correlation, suggesting that they tend to increase together. Additionally, "cap.diameter" and "stem.height" display a moderate positive correlation.

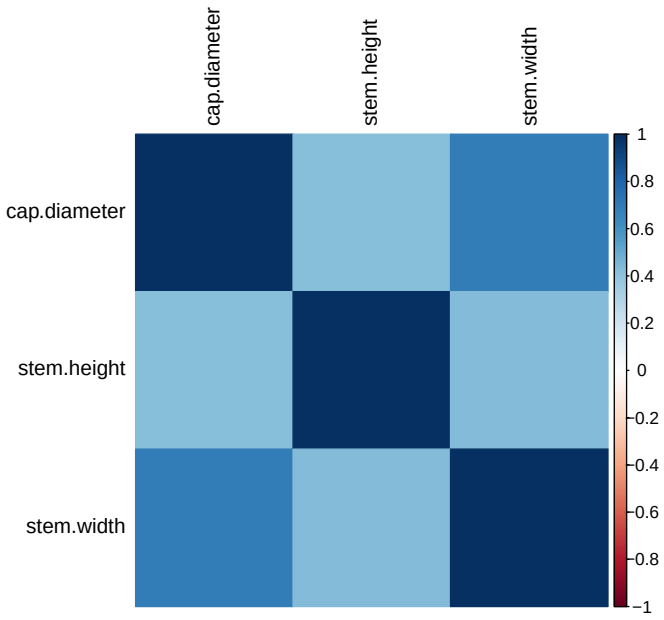

Figure 4. Heatmap for attribute correlations in the secondary dataset.

3. Feature Exploration

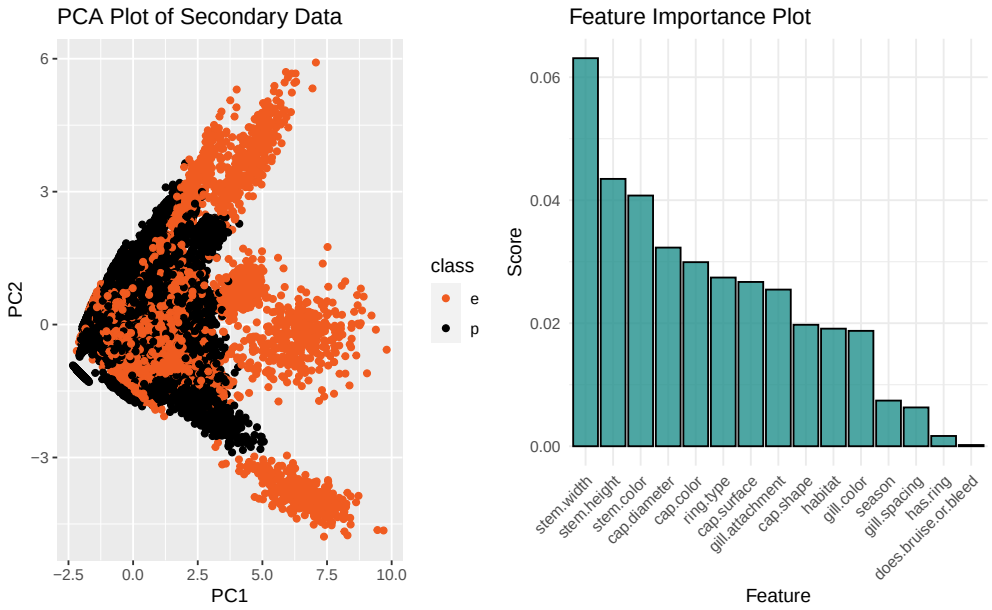

Figure 5. PCA plot and feature importance plot for the secondary dataset.

The Principal Component Analysis (PCA) plot visualizes the distribution of the mushroom dataset in a new feature space defined by principal components. These principal components are linear combinations of the original features, selected to capture the most variance in the data. In this specific plot, the first two principal components (PC1 and PC2) are displayed. Points in the plot represent individual mushroom samples and are colored based on their classes (edible or poisonous). A well-separated PCA plot could imply that the classes are distinguishable based on the derived principal components, which in turn suggests that a simpler model might suffice for classification.

The feature importance plot provides insights into which features are most informative for predicting the target variable, in this case, whether a mushroom is edible or poisonous. The method used here for feature selection is Information Gain, which quantifies how well a feature discriminates between the classes of the target variable. Features at the top of the plot are more important for classification than those at the bottom. In a machine learning pipeline, less important features might be removed to simplify the model without sacrificing performance significantly.

Both these plots serve complementary roles. While the PCA plot helps in reducing the dimensionality and visualizing the entire dataset, the feature importance plot helps to identify the most important individual features for classification.

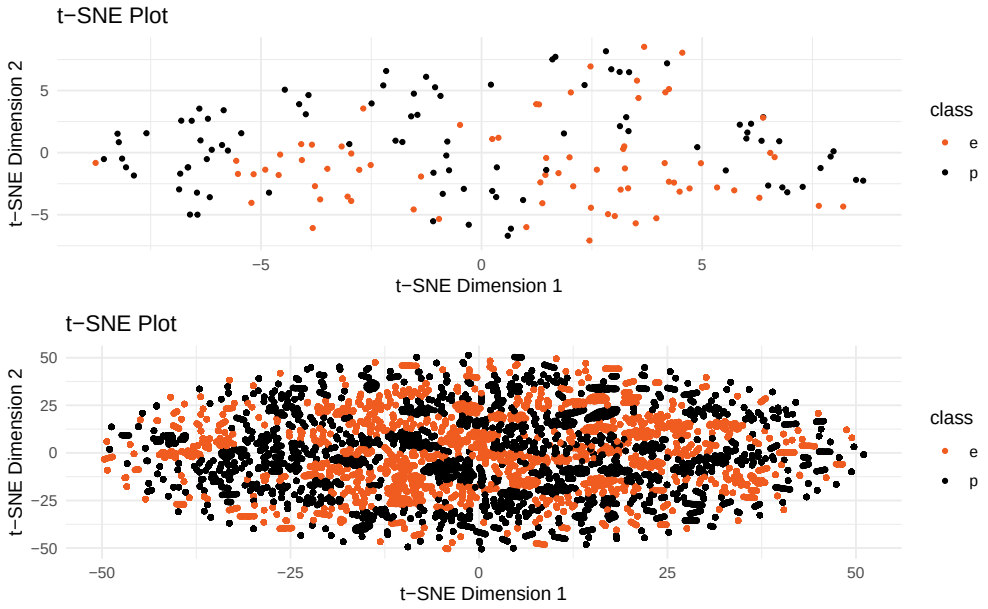

Figure 6. The t-SNE plot for the primary and secondary datasets.

To further explore features and patterns within the datasets, dimensionality reduction techniques like t-SNE have been applied. t-SNE plots are used to visualize the data in a reduced-dimensional space and identify potential clusters or separations among data points. There are visible clusters in the t-SNE plot for the secondary dataset. However, the t-SNE plot for the primary dataset does not show any clear clusters. This could be because the primary dataset is smaller and has fewer features than the secondary dataset.

Discuss:

Preliminary data exploration leverages descriptive statistics and visual elements:

- **Descriptive Statistics**: Primary and secondary datasets contain complete character variables, while some missing values exist in numeric variables of primary dataset.
- **Visual Elements**: Bar plots and heatmaps indicate class distribution and attribute correlations.
- **Observations**: High correlation between `cap.diameter` and `stem.width` suggests they may have a combined influence on mushroom edibility.
- **Feature Exploration**:Implement dimensionality reduction via t-SNE and temporal comparison with older datasets for a historical perspective.

6. References

Wagner, D., Heider, D., & Hattab, G. (2021). Mushroom data creation, curation, and simulation to support classification tasks. *Scientific Reports, 11*(1), 8134-12. https://doi.org/10.1038/s41598-021-87602-3

Wagner, D., Heider, D., and Hattab, G. (2023). Secondary Mushroom Dataset. UCI Machine Learning Repository. https://doi.org/10.24432/C5FP5Q.