INFS 5116 DATA VISUALISATION SP5 2023
Visualisation Project Plan

Edible or Not: Evaluating Data Visualization and Machine Learning Algorithms in Mushroom Edibility Classification

Enna H

1. Introduction

Explain the context for the proposed project and the main question(s) you hope to be able to answer using visualisations. One short paragraph (1/4 page) is sufficient.

High-dimensional data often exceeds the cognitive limits of human interpretation. Scientists may utilize data visualization techniques to distill complex patterns into understandable forms and facilitate informed decision-making. However, as dimensionality increases, the efficacy of human cognition declines. Conversely, machine learning (ML) algorithms excel at pattern recognition in high-dimensional data but often lack transparent interpretability. The project explores two different paradigms for understanding data: data visualization, which leverages graphical representations to enhance human cognition, and machine learning, which relies on algorithms for automated decision-making. Focusing on mushroom edibility classification, the project aims to visualize both the mushroom datasets and the performance metrics of machine learning algorithms. This offers a unique lens to compare the strengths and limitations of each approach. The research questions are:

1. How effectively can data visualization techniques elucidate the complexities of the primary and secondary mushroom datasets for edibility classification?
2. To what extent can machine learning algorithms accurately classify mushroom edibility based on these datasets?
3. How do visual representations of machine learning performance metrics compare with human interpretations of visualized data?
4. What are the implications of combining data visualization and machine learning techniques in edibility classification tasks?

---

2. Data Source

Describe the data source(s) that you plan on using for your visualisation project, where it comes from, how it was collected, its size, number and type of variables etc. The data source(s) you choose should be large and rich enough to allow for visual presentation of data features from multiple perspectives and at different levels of complexity.

One paragraph (up to 1/2 page) is sufficient. You can use dot points.

This project utilizes two mushroom datasets, each offering unique advantages for data visualization and machine learning approaches to mushroom edibility classification.

- Primary Dataset

The primary dataset consists of 177 observations and 24 variables of both nominal and metrical types. Variables include attributes like `cap.size`, `cap.shape`, and `habitat`. A balanced class distribution is maintained (Wagner et al., 2023).

- Secondary Dataset

In contrast, the secondary dataset encompasses 61,069 observations across 16 variables. It has an imbalanced class representation but offers a robust foundation for machine learning algorithms, as well as for visualizing these algorithms' performance metrics (Wagner et al., 2023).

- Significance for Visualization

Both datasets are versatile enough for a dual approach that involves data visualization and machine learning. Both dataset is suited for exploratory data visualization, while the secondary dataset lends itself to machine learning classification and the subsequent visualization of algorithmic performance. The pirmary dataset is more tricky for both tasks as it is less structured, have muitiple values in one cell, more missing values as a results extrated from a textbook using nature language processing.

---

3. Data Preparation

Perform and provide a summary of any data preparation tasks that are required before you embark on building your data visualisations, e.g. joining of files, identifying data problems (missing values or data errors) or creating new variables. One short paragraph is sufficient. You can use dot points.

In preparation for data visualization and machine learning model training, several data preparation tasks have been performed on both the primary and secondary datasets. In the primary dataset, columns with more than 25% missing values were removed, and range variables for cap size, stem height, and stem width were split into minimum and maximum values. Numeric variables were converted to millimeters (mm) and standardized. Missing values in "Cap.surface," "gill.attachment," and "ring.type" were imputed with the mode. One-hot encoding was applied to categorical variables except for the unique "name" and the target variable "class."

For the secondary dataset, no missing values were found in the selected version. Numeric variables were converted to millimeters (mm) and standardized. One-hot encoding was performed for categorical variables, excluding the target variable "class."

These data preparation steps ensure that both datasets are cleaned and four dataset(2 for visulization 2 for ml) are ready for further analysis, visualization, and machine learning model training.

---

4. Project Aims

Propose some elementary, intermediate and overall level questions that may be addressed using your chosen data set(s). Identify the scope of the proposed visualisation project and possible problems that may occur along the way. No more than one and a half page. You can use dot points.

---

1. To showcase how data visualization techniques translate high-dimensional data into forms that human cognition can grasp.
2. To demonstrate the efficacy of machine learning algorithms in identifying patterns within the same dataset.
3. To compare how data visualization and machine learning techniques interpret multi-dimensional attributes differently.

Potential challenges include high dimensionality that may lead to overfitting and the requirement for computational resources for complex models.

---

4. Project Aims

Elementary Level Questions:

1. Which variables in the primary and secondary datasets have the most impact on mushroom edibility?
2. What is the distribution of edible and poisonous mushrooms in both datasets?
3. How do individual variables correlate with edibility?

Intermediate Level Questions:

1. How does variable selection affect the performance of machine learning algorithms in classifying mushroom edibility?
2. What are the trade-offs between different data visualization techniques in representing high-dimensional data?
3. Can feature importance from machine learning algorithms align with human interpretation from visualizations?

Overall Level Questions:

1. How does the performance of machine learning algorithms compare to human interpretation for mushroom edibility classification?
2. Can combining machine learning and data visualization techniques provide a more robust approach for mushroom edibility classification?
3. What are the ethical considerations for deploying machine learning models based on these datasets, particularly concerning potential false positives or negatives?

Scope of the Project:

- The project aims to juxtapose two approaches: data visualization to enhance human cognition and machine learning for automated decision-making.
- It intends to explore elementary to advanced questions revolving around the efficacy, interpretability, and ethical implications of these approaches.
- While the focus remains on mushroom edibility classification, the methodologies could extend to other high-dimensional classification problems.

Potential Challenges:

1. High Dimensionality: Given the high-dimensional nature of the datasets, overfitting could become a potential issue, particularly for machine learning algorithms.
2. Computational Resources: Complex machine learning models and intricate visualizations may demand significant computational resources.
3. Interpretability: Balancing the comprehensibility of visualizations against the complexity of machine learning models could be challenging.
4. Ethical Concerns: Incorrect classification could have life-threatening implications; therefore, model interpretability and validation are crucial.

---

5. Data Exploration

Present results of preliminary exploration of your data, which can include descriptive statistics and graphs aimed at helping you to get to know and understand your data before you begin building your main visualisations. What do your preliminary results suggest in relation to the questions of interest identified in section 4? Comment briefly. Four to six graphics (at least two types) plus one paragraph is sufficient. You can use dot points if you wish.

In this section, we delve into the preliminary exploration of the data, employing various descriptive statistics and graphical representations to gain a better understanding of the datasets before embarking on the main visualizations. The objective is to shed light on the questions of interest identified in Section 4.

- Descriptive Statistics & Graphics Attribute-Based Distribution: One of the initial steps in data exploration is to understand the distribution of attributes within the datasets. Visual techniques, including heatmaps, are being employed to visually depict the frequency of mushroom attributes such as size and color.

The dataset “mushroom_1” consists of 173 rows and 20 columns, encompassing 14 character and 6 numeric variables. The character variables are complete with no missing values, while the numeric variables exhibit some missing values, particularly in “cap.diameter_max” and “stem.height_max.” These numeric variables have been standardized. The dataset appears to be well-structured for analysis.

Similarly, the dataset “mushroom_2” contains 61,069 rows and 16 columns, comprising 13 character and 3 numeric variables. All character and numeric variables are complete with no missing values, and the numeric variables have been standardized. This dataset is also well-prepared for analysis.

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.

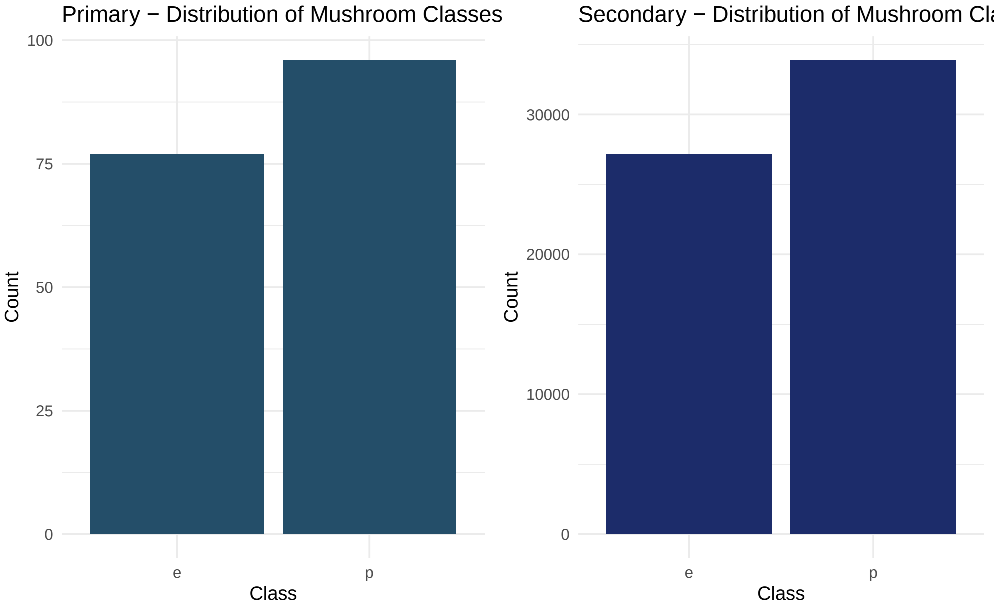

Figure 1. barplot for class distribution in primary and secondary dataset.

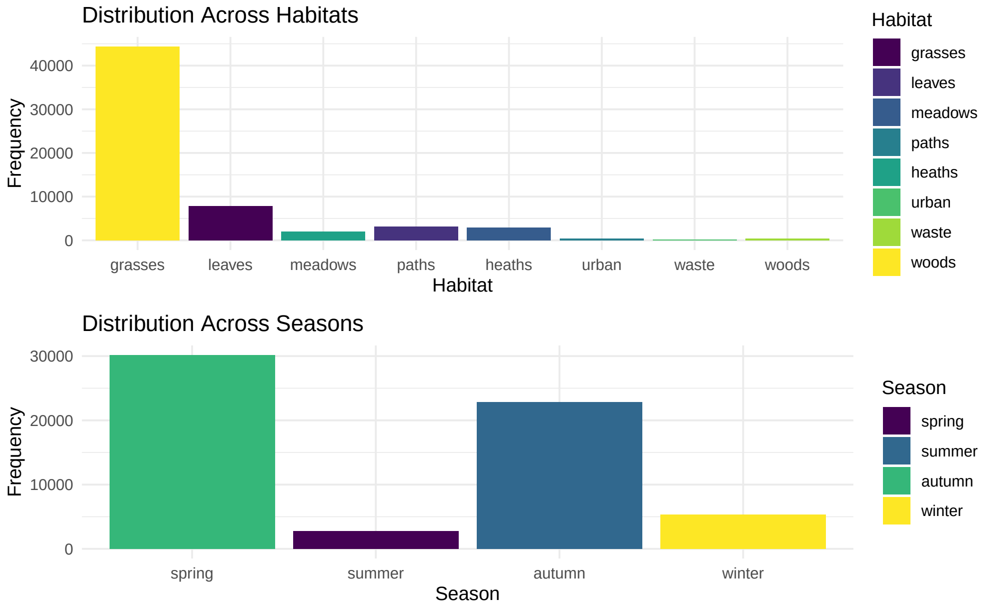

Figure 2. distribution across habitat and season for the secondary dataset.

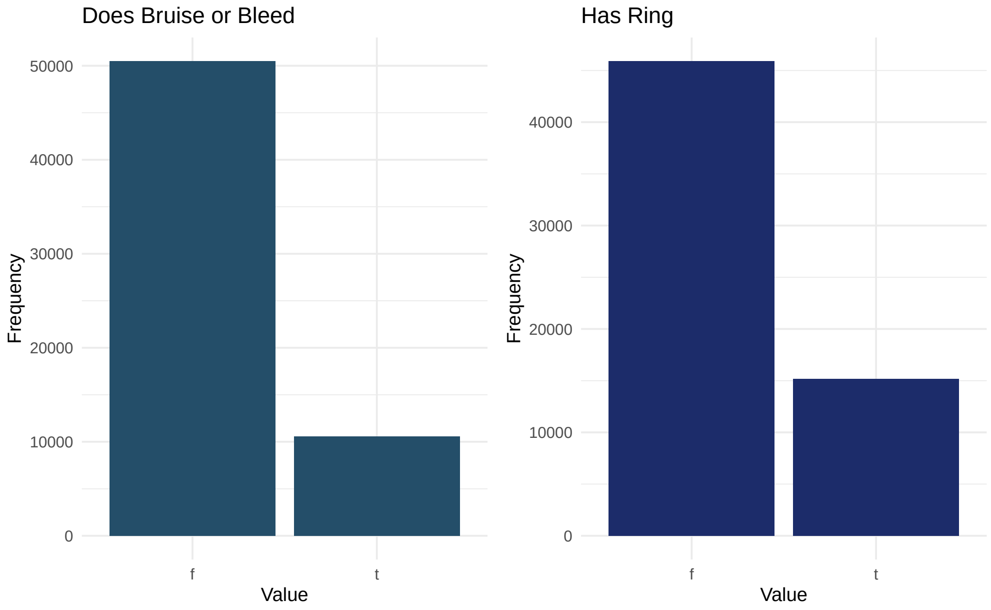

Figure 3. binary morphological attributes for the secondary dataset.

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. However, it's important to note that the correlations are not excessively strong, indicating that these variables provide distinct information.

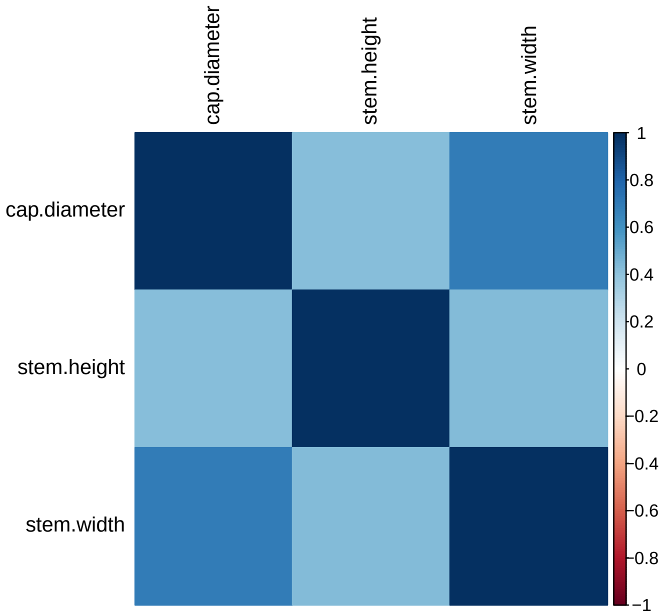

Figure 4. heatmap for attribute correlations in the secondary dataset.

Feature Exploration 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.

Additionally, a temporal comparison is planned, involving a contrast with a dated dataset (e.g., the 1987 UCI dataset) to understand how attribute importance has evolved over time.

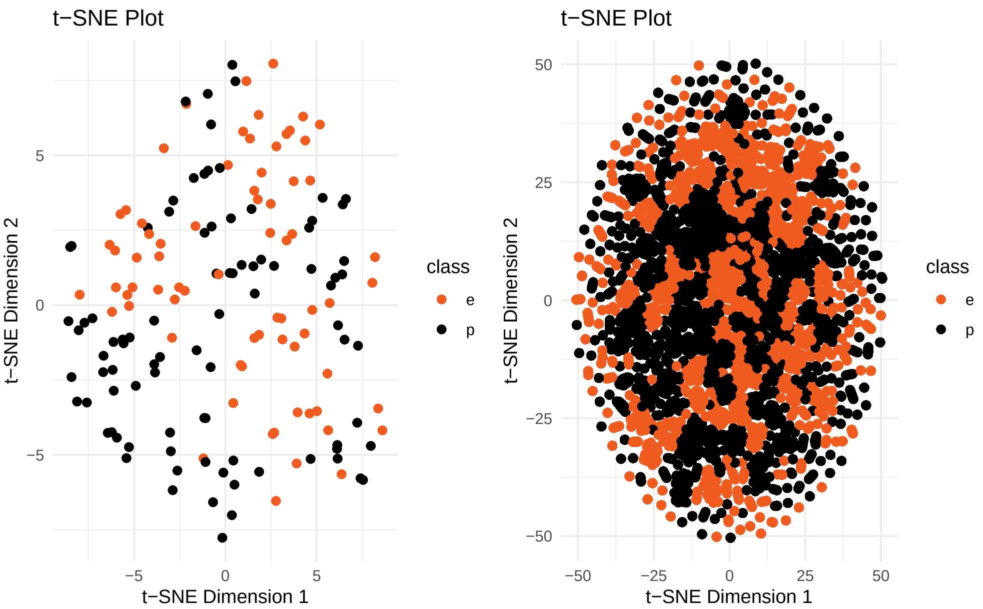

Figure 5. t-SNE plot for the primary and secondary datasets.

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.