Decision Tree Classification of Mushroom Edibility

Prepared by: Hemn Sheikholeslami

Executive Summary

This report presents the findings from a decision tree analysis conducted to classify mushrooms as either edible or poisonous. The key objectives were to familiarize with decision tree classification, to identify the most predictive features, and to evaluate the model's effectiveness. The model achieved high accuracy, and the feature importance analysis provided valuable insights. Recommendations include using the model for educational purposes and further research on additional data.

Introduction

Background

This project focuses on the necessity to classify mushrooms effectively, which is crucial for both consumer safety and educational purposes. The decision tree model offers a straightforward interpretative framework for such classifications.

Objectives

The primary goal was to develop a decision tree model to classify mushrooms based on certain features and to evaluate the model's performance and predictive power.

Data Description

Source

The data was sourced from the Mushroom Data Set available in the UCI Machine Learning Repository, which includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms.

Features

The dataset comprises 8124 samples with 22 features, including cap shape, odor, gill size, and color, crucial for predicting mushroom edibility.

Sample

```
class, cap-shape, cap-surface, cap-color, bruises, odor, ...
p, x, s, n, t, p, ...
e, x, s, y, t, a, ...
```

Methodology

Data Preprocessing

The data underwent preprocessing which included encoding categorical variables, handling missing values, and splitting the dataset into training, validation, and test sets.

Model Training

A decision tree classifier was employed, with hyperparameters optimized using Grid Search CV. The training involved not only the base model but also variations to assess feature impact.

Feature Analysis

Feature importance was assessed using the model’s built-in capabilities, supplemented by statistical tests and additional visualizations to understand feature interactions.

Results

Model Performance

The decision tree model demonstrated exceptional accuracy, precision, and recall across all data splits, indicating strong predictive performance.

Feature Importance

The analysis revealed that features such as odor and gill color were highly predictive of mushroom class. Refer to the visualizations in the accompanying Jupyter notebook for detailed charts.

Visualizations

Refer to Figures 1 through 4 in the Jupyter notebook for visualizations of feature importance and distribution plots that support these findings.

Discussion

Interpretation of Findings

The high accuracy of the decision tree model suggests it effectively captured the patterns necessary for classifying mushrooms, though with caution against potential overfitting.

Challenges and Limitations

One challenge was ensuring the model did not overfit given its high accuracy. Future models might include cross-validation techniques to mitigate this risk.

Conclusion and Recommendations

The decision tree model is recommended for educational and possibly commercial applications given its high accuracy and interpretability. Future work could explore ensemble methods to enhance model robustness.

Appendices

Code

```
# Sample Python code for model training
from sklearn.tree import DecisionTreeClassifier
model = DecisionTreeClassifier()
model.fit(X_train, y_train)
```

References

All data sourced from the UCI Machine Learning Repository. Model documentation and Python libraries such as scikit-learn were referenced.