MushroomEdibilityPrediction Documentation

version

2024, Juan José Borrero Mejía

November 08, 2024

Contents

MushroomEdibilityPrediction documentation

Bussiness Understanding

Project Objective

The goal of this project is to predict whether a mushroom is edible or poisonous based on its physical characteristics.

Problem Statement

Incorrect predictions could lead to serious health risks if poisonous mushrooms are misclassified as edible. Therefore, we aim to achieve a high level of accuracy and reliability in our model.

Success Criteria

- **Model Performance**: Achieve an accuracy of at least 80%.
- **Practicality**: The model should be easy to interpret and explain to non-technical stakeholders.

Constraints

- Limited dataset from Kaggle and UCI.
- Must operate within a reasonable time frame for data processing and model prediction.

Data Understanding

Data Sources

- Kaggle Playground Series S4E8
- Mushroom Dataset - UCI Machine Learning Repository

Data Dictionary

Feature	Description
cap-shape	Shape of the mushroom cap
cap-color	Color of the mushroom cap
gill-size	Size of the gills
gill-color	Color of the gills
…	…

For the complete list of features visit the UCI Machine Learning Repository.

Initial Observations

- The dataset contains mostly categorical features.
- Target variable: **edibility** (edible (e) or poisonous (p)).

Data Quality Issues

- Some missing values in the color attributes.
- Possible class imbalance between edible and poisonous mushrooms.

Data Preparation

Modeling

Model Selection

- Experiment with various classification algorithms:
    - Classification Tree
    - K Nearest Neighbors (KNN)
    - Support Vector Machine (SVM)
    - Neural Network
    - Bagging
    - Random Forest
    - Gradient Boosting

Model Training

- Train each model using cross-validation to optimize performance.
- Perform hyperparameter tuning for the best model 3 models.
- Save the best model for each of the 3 algorithms chosen before.
- Save the best overall model for deployment.

Evaluation Metrics

- **Accuracy**: Measure of correctly predicted instances.
- **ROC-AUC**: Area under the Receiver Operating Characteristic curve.
- **F1 Score**: Balances precision and recall, especially useful for imbalanced classes.
- **Recall**: For poisonous class. Measure of actual positive instances that were correctly predicted.

Model Results

- Summary of each model’s performance on training and validation sets.

Evaluation

Model Performance Summary

- **Chosen Model**: Random Forest (if selected based on performance).
- **Accuracy**: 96%
- **F1 Score**: 0.95

Business Goal Evaluation

- The model meets the accuracy and interpretability requirements set in the Business Understanding phase.

Limitations

- Possible overfitting if the model is too complex.
- Limited generalizability to other mushroom types not in the dataset.

Deployment

Deployment Plan

- **Platform**: Streamlit will be used to deploy the model.
- **Environment**: It will be temporarily hosted upon execution by localtunel.