Identifying Poisonous Mushrooms

Group Members

- Rachel Baryol / 75284798
- Luis Borruel / 99576876
- Heather Burke / 56136141
- Devon Chessher / 06662605

Synopsis

Identifying whether a mushroom is poisonous or not is not a trivial matter. It requires specific identification of various attributes to determine the lethality of a specimen. This brings us to our solution: leveraging machine learning algorithms to automate mushroom classification based on morphological characteristics such as cap shape, color, odor, gill properties, and habitat information. The project involves preprocessing mushroom attribute data, applying feature selection techniques, and training multiple classification algorithms including decision trees, random forests, and support vector machines. Model performance is evaluated using cross-validation with emphasis on minimizing false negatives given the potentially fatal consequences of misclassifying poisonous mushrooms. The final deliverable is a robust classification tool that assists foragers, researchers, and educators in making informed decisions about mushroom safety, reducing poisoning risks while advancing understanding of morphological patterns that distinguish toxic from non-toxic fungi.

Data Set

The data set is from the UC Irvine Machine Learning Repository describing mushrooms in terms of physical characteristics and classifying them as poisonous or edible. The data includes descriptions of samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota families. Each species is identified as edible, definitely poisonous, or of unknown edibility, and not recommended, which is combined with the poisonous identification.

The data itself comprises 8,124 instances with 22 categorical features.

https://archive.ics.uci.edu/dataset/73/mushroom

Problem Statement

Given a mushroom's physical attributes, predict whether it is edible or poisonous using a decision tree classifier.

Evaluation Metrics

To evaluate the performance of our mushroom classification model, we will primarily use random forest regression, which builds an ensemble of decision trees and aggregates their predictions to improve accuracy and reduce overfitting.

Techniques

As a starting point for our classification task, we will use a Decision Tree model due to its transparency, interpretability, and compatibility with categorical data, which makes it particularly well-suited for the mushroom dataset. We will be splitting the data based on attributes that maximize information gain, resulting in a model that can be easily visualized and understood. This is valuable in our case, as the ability to trace a prediction back through the tree allows users to understand why a mushroom is classified as edible or poisonous.