You will have to form a group of 3 to 4 students.

Choose a dataset from the UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php) with at least 5000 instances and 20 attributes for classification or regression. Compare how the different approaches seen in class perform on this dataset to predict accurately the classes or the values of the unlabeled data. You should determine what are the best hyper-parameters for each approach you are using. You could use any Python libraries.

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

First Project: Machine Learning Analysis

This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family (pp. 500-525).

Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended.

This latter class was combined with the poisonous one.

The Guide clearly states that there is no simple rule for determining the edibility of a mushroom; no rule like ``leaflets three, let it be'' for Poisonous Oak and Ivy.

Abstract

Have you ever been walking down the streets of Windsor and been plagued by the question of determining the edibility of a mushroom? [1] While mushroom foraging has a rich tradition and a community of enthusiasts, it also poses inherent risks. This comes as some species can be toxic and potentially fatal if consumed. Reliable identification of edible and non-edible mushrooms is crucial for the safety and enjoyment of foragers and nature enthusiasts alike. This research endeavor delves into the mixture of machine learning methods with the natural world. Therefore, the main goal of this research endeavor is to develop a robust and accurate method for classifying the edibility of gilled mushrooms, thereby, providing a valuable tool for those exploring the world of mycology.

1. Introduction

The root of the problem lies in the difficulty of reliably distinguishing between edible and non-edible mushrooms. This challenge is compounded by the wide variety of mushroom species with subtle visual differences. Machine learning presents a compelling opportunity for tackling the task of classifying the edibility of mushrooms. Its capacity to analyze extensive datasets, detect intricate patterns, and provide data-driven insights aligns seamlessly with the complex task of mushroom identification. Through this research, we intend to harness the power of machine learning to revolutionize the way we approach the age-old practice of mushroom foraging, making it safer, more accessible, and more enjoyable for all. As we examine the details of our dataset and data preprocessing in the following sections, we will highlight the steps taken to prepare the data for our machine learning models.

1. Dataset

The dataset encompasses various mushroom characteristics, including cap shape, cap surface, cap color, bruises, odor, gill attachment, gill spacing, gill size, gill color, stalk shape, stalk root, stalk surface above and below the ring, stalk color above and below the ring, veil type, veil color, ring number, ring type, spore print color, population, and habitat. These attributes provide a rich and diverse set of features for our machine learning models to analyze.

1. Data Preprocessing
2. Model Selection
3. Model Training and Evaluation
4. Hyper-Parameter Tuning
5. Comparison of Results
6. Conclusion and Findings
7. Description of Student Participation

References

[1] – fun guyz

dataset

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