Milestone 1 – Proposal

Deadly Mushrooms

Topic

The name of my project will be Deadly Mushroom Classification. I will be using factors about mushrooms to determine which are edible and which are poisonous.

Business Problem

There are over 10,000 known types of mushrooms in the world. Roughly 70-80 of them are poisonous and some even can even be fatal when eaten. How can we identify which mushrooms are safe to eat? What factors are important to identify such mushrooms? How accurate can we predict if a mushroom is safe or not? These are questions I will be looking to answer.

Datasets

I will be getting the data set from Kaggle. This dataset was originally contributed to the UCI Machine Learning repository 30 years ago.

Mushroom Dataset: https://www.kaggle.com/uciml/mushroom-classification?select=mushrooms.csv

This dataset from Kaggle "includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms (1981). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended" (2016, CI Machine Learning).

The data includes 23 columns. Most of the columns (21) are strings while the remaining (2) are Boolean operators. The attributes included are class, cap-shape, cap-surface, cap-color, bruises, odor, gill-attachment, gill-spacing, gill-size, gill-color, stalk-shape, stalk-root, stalk-surface-above-ring, stalk-surface-below-ring, stalk-color-above-ring, stalk-color-below-ring, veil-type, veil-color, ring-number, ring-type, spore-print-color, population, habitat,

Methods

I will be using classification methods to determine if the mushroom is edible or poisonous. Possible Models will include logistic regression, decision trees and clustering algorithms. Models used will compare accuracy and look at type 1 and 2 errors to see which performs better. It is more important that we do not falsely identify deadly mushrooms as edible mushrooms, verses errors where we identify safe mushrooms as deadly (type 1 and type 2 "alpha and beta" errors).

Ethical Considerations

Identifying a mushroom as edible when it is not, can be disastrous to the consumer. It can also open legal problems. Even if the accuracy is 99%, what happens when that 1% of poisonous mushrooms gets identified as safe? Is it better not to try and identify the mushroom at all? How much responsibility should the data scientist who develops a model have over the performance of the model or how it is used? These are ethical considerations of identify harmful substances.

Challenges/Issues

The data set are mostly string variables so I will need to cleanse and prep the data prior to throwing it into an algorithm. I also need to look at the distribution of data since there are a large variety of mushrooms, but only a few poisonous ones. If the data is imbalanced, I will need to oversample/under sample.

References

UCI Machine Learning. (2016, December 1). *Mushroom classification*. Kaggle. Retrieved January 16, 2022, from https://www.kaggle.com/uciml/mushroom-classification