CSE 446 Project Part 1: Choosing Datasets:

Group Details:

- Ahmed Mohammad Awwad (*sherman6*)
- Paul Michael Curry (*paulmc*)
- Sandip Samantaray (*sandip80*)

Dataset:

Name: mushroom_edibility

Overview:

This dataset contains data for the family of mushrooms. What makes a mushroom edible? Imagine you wake up one day and find yourself on an uninhabited island. All you can find on the island are mushrooms which is your only source of food. And of course you have a computer with you that is capable of running ML algorithms. Well, you are in luck. Using our dataset you can implement an ML algorithm (given you are 446 survivor) that would let you know if a mushroom you are about to eat is edible or not (good luck if you overfit the data XD). Humor aside, the classification of mushrooms furthers our understanding of the natural world and could even provide insights into the development of intelligence. After all, animal intelligence developed through the identification of what is dangerous and what is safe based off of contextual clues and features.

Logistics:

mushroom_edibility has a total of 143 attributes, 4784 training instances, 600 development instances and 600 test instances. This is roughly an 80-10-10 split. The initial dataset was a mixture of values for a given attribute (8 color values for gill_color). So we decided to separate each value in a given attribute into its own separate attribute such that, the only possible values for any given attribute would be 0 or 1. This increases the number of dimensions for an instance but also makes it easier to program with. Our dataset was not too hard because its categorical data was easily broken up into binary values. The dataset was not too easy because it did involve clean up and wasn't initially perfectly formatted.

Dataset format:

Attributes:

Please refer to the attributes.txt file bundled with the dataset.

Input:

Each attribute has either a value of 0 or 1. 0 means FALSE and 1 means TRUE.

Output:

The output is either 0 or 1. 0 means the mushroom is poisonous, 1 means it is edible.

How did we divide dataset to training and test data?

Given the number of attributes and total instances, we randomly chose 600 instances each for test and development data. We have made sure that there is no bias for instance selection by partitioning the instance sets based off of a random ordering via the excel random() function. Thus, there are no repeated instances in each set and their were partitioned randomly.

Why is our Dataset ethical?

Our dataset is ethical because it does not compromise anyone's privacy. The dataset was sourced from The Audubon Society Field Guide to North American Mushrooms, and was provided free and with open access.