ECS 171 Project Report

Steven Alvarado, Noor Ashour, Khaiber Amin, Dhruv Arora

Problem statement

The problem that our group addresses centers around building a classification model that can accurately predict whether or not a species of mushroom is safe for human consumption. Our goal in developing this model is to help automate the identification of poisonous mushrooms and reduce the risk of misclassification and potential harm. Therefore, our plan is to use machine learning to classify whether a mushroom is edible or not based on a description of its attributes, such as cap shape, gill size, odor, etc.

Dataset description

From the UCI Machine Learning Repository, we have determined that the Mushroom dataset contains a sufficient number of features to train and validate this model effectively. Within it, we will utilize the physical characteristics that are provided from a sample of 23 mushroom species to analyze any relationships between fruiting body features and human toxicity/edibility.

Exploratory data analysis

For dataset analysis and understanding, we imported our data as a data frame in a shared Jupyter notebook and dissected it with several functions to illustrate any significant details and trends. Through our code, we reformatted the attribute values for more verbosity and clarity and gathered statistics such as percentage breakdowns of each column to influence what variables we should keep or remove. For instance, we discovered that the mushroom dataset has a large majority of samples that have free gill attachments (97.42%), which signified a variable that did not provide enough unique information that would effectively train our model. The veil type attribute was the largest indicator of this, where 100% of the samples were all one type. As a result, such features were eliminated from the dataset to ensure that our model utilizes the most useful information. Some notes we made regarding the dataset is that all the data is categorical (i.e. in string format). Thus, we'll have to apply techniques like one-hot encoding or label encoding to transform the data into numerical format. There were also missing data values, all in the feature"stalk-root" (2480 out of 8124). We are considering ignoring "stalk-root" as part of our feature selection process.

Literature review

We read the relevant literature on mushroom classification to get a better understanding of how people generally approach this problem. For example, we learned that macroscopic identification methods are the ones that we should be

concerned with since these types of observations are what our given mushroom dataset consists of. We studied a handful of peer-reviewed papers that also gave us inspiration for different classification algorithms to experiment with and what we could prioritize moving forward. Many of the studies we researched compared multiple algorithms and/or models to see which ones were most effective, influencing our potential approaches. With that, we were able to form a good idea of the variety of methods we can choose from that

Model development

Model evaluation

Experiment results

Conclusion

Discussion