Scott Breitbach
DSC680, Weeks 1-4
12-June-2022

Project 1: Milestone 1

Topic:

Support rapid clinical diagnoses through automatic machine identification of microbes using measurements of cellular features generated from microscopic imagery.

Business Problem:

When dealing with an infection, time is an important factor because doubling times for many bacteria are often measured in hours and sometimes even minutes. Culturing bacteria for identification via genetic analysis and other methods is often time consuming and having technicians manually identify microbes can vary from person to person and is often subjective. Rapid and accurate identification of bacteria can help to support appropriate treatment options, saving lives and potentially limiting further development of antibiotic-resistant forms of bacteria.

Datasets:

To establish an initial proof of concept for microbe identification, I will utilize a dataset available on Kaggle, titled 'Microbes Dataset'. The dataset contains around 34,500 observations. In addition to unique serial numbers and the 'microorganisms' name target variable, each observation has 24 measured features, each dealing with various aspects of size and shape of the microorganism, all numeric.

Methods:

This is a classification problem and I plan to try several different methods to determine which is the most effective at predicting the microorganisms with this dataset. An initial look at the data suggests that there are no Null values, so I won't have to worry about imputing values, or removing rows with missing data. Because the target variables are not balanced, I may have to sample from the more represented organisms and/or do some resampling of the less represented organisms.

Depending on the learning algorithms I work with, I may need to do some feature scaling of features as well as dealing with outliers, if present. The target variable is categorical so these organism names will need to be encoded appropriately. I will do some feature selection as I believe some of the measurements may be redundant, like the area of the organism ('Area') and the diameter of a circle with equivalent area ('EquivDiameter'), though there may also be opportunities for feature creation.

Right now, percent accuracy seems like a decent metric for model evaluation.

Ethical Considerations:

One ethical concern is that the people tend to trust ML outputs without consideration of the probabilities present. The algorithm may be 73% sure that it identifies an organism "*x*" and the operator is likely to interpret that output as "this is organism *x*". The implication here is that it could lead to misdiagnosing an infection, which may lead to incorrect treatment, so the patient's health and well-being is at stake. Upon final implementation, it will be important to make it clear that this is a tool and perhaps show the top 2-3 organism matches, perhaps with probabilities accompanying each.

Another concern upon implementation is going to be patient anonymity and security of the patient's information. Perhaps this could be ameliorated by using serial numbers or sample IDs that can be matched up to patient records separately.

Challenges / Issues:

This is tough to predict because often I don't know what the issues are going to be until they pop up, but this class is going to be a real test of what I've learned and my ability to recall it. Things like which types of algorithms to use for which tasks, data preprocessing requirements for the different algorithms, even how to code a lot of it. Like learning anything new, if you don't use it a lot, you're going to start losing it so this will be a nice refresher for me. I hope to revisit some code I've written previously, and I anticipate doing a lot of Googling.

References:

Kaggle Microbes Dataset: https://www.kaggle.com/datasets/sayansh001/microbes-dataset

I feel the dataset I selected is large enough to work with for a classification task and the entire dataset is labeled with target variables. I will need to split the dataset into Training, Validation, and Testing sets.

The company I work for already has an established product in this field, which I have taken some inspiration from for this project: https://www.wired.com/sponsored/story/cloud-to-clinic-zoetis-vision-for-veterinary-practices/

Additionally, I found a resource 'Guide to any Classification Problem' on Kaggle that I will reference for an overall process outline to get me started and keep me on track: https://www.kaggle.com/code/durgancegaur/a-guide-to-any-classification-problem