Classifying microorganisms using machine learning
Project 4: Classification proposal
By: Jing Lee

I. Question/Need:

What is the question behind your analysis or model and what practical impact will your work have?

Microscope image analysis is a laborious and tedious job requiring skilled and trained personnel to perform the job. The analysis is prone to human error. The lab personnel need to spend hours a day to identify hundreds of cell images and researchers may be susceptible to decision fatigue and bias. Some microscope software has built in functions to measure the features of the images in an automated way, but the scientist still has to identify what the image is. Automating the classification of cell images by predicting the type such as fungi or algae, can help speed up identification of microorganisms while reducing human errors and bias.

I'm proposing to build a machine learning model that can be used to classify the type of a microorganism based on features extracted from microscope image analysis. As a proof of concept, my model will predict whether features of an algae are Ulothrix or not.

Who is your client and how will that client benefits from exploring this question or building this model/system?

My client will be microscope companies. The automated classification feature can be used alongside the microscope software to speed up overall image analyses while reducing bias and error.

II. Data Description:

What dataset(s) do you plan to use, and how will you obtain the data? Please include a link!

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

What is an individual sample/unit of analysis in this project? In other words, what does one row or observation of the data represent?

A row of data represents the features of a microscope image of a particular microorganism.

What characteristics/features do you expect to work with? In other words, what are your columns of interest?

The 15 columns of interest are: solidity, eccentricity, equivdiameter, extrema, filled area, extent, orientation, euler number, bounding box, convex hull, major axis length, minor axis length, perimeter, centroid, and area. These are numerical features extracted from microscope images of organisms.

If modeling, what will you predict as your target?

I will predict whether the algae is Ulothrix or not.

III. Tools:

How do you intend to meet the tools requirement of the project?

- Exploratory data analysis: Pandas

- Visualizations: matplotlib and seaborn
- Classification model: sklearn

Are you planning in advance to need or use additional tools beyond those required?

- Visualize results in Tableau dashboard

IV. MVP Goal:

- EDA, and results of baseline model such as knn and logistic regression.