Angela Zhu | MAIS202: Deliverable 1 | February 8th 2024

Choice of dataset

https://www.kaggle.com/datasets/nwheeler443/gono-unitigs/data
This dataset has DNA sequences of bacteria which might indicate that the bacteria has antibiotic resistance against three specific antibiotics. I chose this dataset because antibiotic resistance is a very real problem we are facing right now that is only going to get worse with time. Being able to predict antibiotic resistance from DNA sequencing results, which is becoming more accessible due to cheaper and more efficient technology, would mean better and more effective drug prescriptions.

a. Data Preprocessing

I don't think there is any preprocessing needed for the data provided in the dataset since it's meant as a tutorial. However, if the input is just a general DNA sequence, then it will have to be converted into unitigs before it can be used with the model, which I am not quite sure how to do as of yet, although the dataset does provide the paper that should contain the relevant methodology.

b. Machine Learning Model

The goal of the model would be to predict if a specific bacteria strain has antibiotic resistance against one of the three antibiotics from its DNA sequence. In the code section of Kaggle, the dataset provider also provided instructions on how they trained the model. Random forest yielded the highest accuracy while taking the least time which is why I will use that model.

c. Evaluation Metric

The confusion matrix can be used to evaluate the amount of false negative/positive, which in this case would be labeling a strain as having resistance when it does not for example. This will be particularly relevant if this model is to be used as a "diagnostic" tool.

Application

The user would have to input the sequencing data obtained from a bacteria strain. Then the unitigs would have to be extracted somehow. The output would be if the strain of bacteria of the patient has antibiotic resistance. With this knowledge, the clinicians will be able to make a more informed decision on the antibiotic to prescribe the patient.

Since I was not able to contact my TPM, an alternative, and probably simpler but less cool, project would be to predict the likelihood of someone having a stroke depending on features such as gender, age, work type etc., using this dataset https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data. Again, random forest will be used along with the confusion matrix as evaluation metric.