Angela Zhu | MAIS202: Deliverable 2 | February 18th 2024

Problem Statement

The prediction of antibiotic resistance of a bacterial strain from its DNA sequence reads.

Data Preprocessing

The dataset that will be used was obtained from:
https://www.kaggle.com/datasets/nwheeler443/gono-unitigs/data.
In this dataset's metadata file, there are 3786 bacterial samples each with its unique ID, the year it was sequenced, the country, continent, and whether it has antibiotic resistance against azithromycin, ciprofloxacin, ceftriaxone, cefixime, tetracycline or penicillin (0 is not resistance, 1 has resistance) amongst other information. There are three accompanying unitig files, each for an antibiotic agent of interest: azithromycin, ciprofloxacin and cefixime. An unitig is a short DNA sequence reading generated from next-generation sequencing techniques.

In each unitig file, the rows represent the unitigs which had the lowest P-value associated with resistance against the specific antibiotic. Each column is a sample. For each unitig and sample pair, a 0 means that the unitig was absent from the sample whereas a 1 means that it is present.

Therefore, the possible labels are

- Has resistance against the antibiotic of interest.
- Does not have resistance against the antibiotic of interest.

The features are the presence or absence of each unitig. For azithromycin, there are 515 possible unitigs, for cefixime, there are 384 possible unitigs and for ciprofloxacin, there are 8873 possible unitigs. For this preliminary result report, I will start with azithromycin. I started by removing all samples which had N/A as its label.

Machine Learning Model

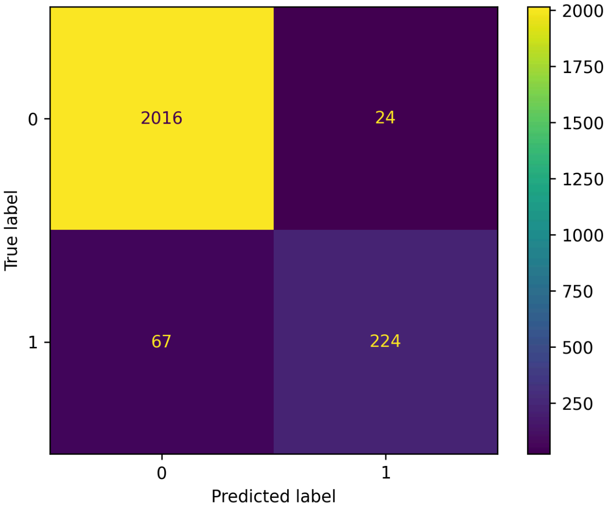
The random forest model will be used for classification: the dataset is divided into random subsets which are each used to train a single decision tree. Then, each decision tree analyzes the data to be classified and outputs a result. The most common result is taken to be the final classification. The Scikit-learn library can be used to implement this model [1].

Parameters that can be modified are the number of trees in the forest, the gini criterion which evaluates the quality of the split, the maximum depth of the tree, the minimum number of samples to split an internal node, the minimum number of samples for a leaf node, the leaf's minimum weight fraction, maximum number of features for splitting etc. [2] For now, the default values were taken and the training and testing/validation dataset were split 33/67.

The preprocessing part was the most challenging part. In theory, there should be the same samples in the .Rtab file and the metadata file. However, checking the size shows that metadata has 3786 samples and .Rtab has 3972 samples. This is concerning because the initial assumption was that they had the same samples in the same order. To make sure the feature values are associated with the right sample, the data frames had to be joined.

Preliminary Results

The model gave an accuracy of 0.96, precision of 0.9 and recall of 0.77 which means that while the model is good at only labeling a result as positive if it truly is resistant, it is less good at finding all resistant samples. The following figure shows the resulting confusion matrix:



Next Steps

The next steps would be to examine antibiotic resistance against the other drugs in the metadata and to tune the hyperparameters. Then, the impact of other features such as region could be examined. Another possible analysis would be to determine if there is any correlation between how having antibiotic resistance to one drug might make the strain more likely or less likely to be resistant against another drug.

References

[1] "Random Forest Classification with Scikit-Learn." Accessed: Feb. 18, 2024. [Online]. Available: https://www.datacamp.com/tutorial/random-forests-classifier-python

[2] "sklearn.ensemble.RandomForestClassifier," scikit-learn. Accessed: Feb. 17, 2024. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html