Daniele Bugatti, Nicole Dupertuis,
Amanda Giacomuzzi, Namdoel Oberholzer

Escherichia Coli Ceftriaxone Resistance

Overview of the Dataset:

Antibiotic resistance in bacteria is a critical issue in modern medicine. A study conducted from 2013 to 2015 revealed that an astonishing 87.5% of the bacteria sampled in North Iran were resistant to at least one antibiotic [1]. It is evident that this percentage has likely increased until now, underscoring the urgency of addressing this topic. Another study published in 2022 successfully compiled a dataset containing mass spectrometer results from more than 300'000 samples of bacteria and fungi. This dataset was then used to train a machine learning model capable of predicting whether a bacteria is resistant to a specific type of antibiotic; with the ultimate goal of prescribing the most effective treatment to a patient [2]. In this project, we only selected the samples used to train for the resistance to Ceftriaxone of Escherichia Coli, with the aim of building and testing different models onto real-world scenarios.

Data Preprocessing and Visualization:

The dataset consists of an integer label, either 1 or 0, indicating resistance to the antibiotic; mass spectrometry results, which are floats ranging from 0 to 1, which have been most likely normalized; and a column named "Unnamed : 0" which has elements of type string. This column contained a serial number ending with either "MALDI_1" or "MALDI_2", suggesting it might relate to the machines used for the observations. To prevent overfitting, we decided to remove this column. Additionally, we checked the dataset for NA and duplicates but found none. The reformatted dataset now consists in 5999 features and one label spanning over 1386 samples. As the dataset contains more features than samples, a feature selection must be performed before fitting with either a filter, wrapper or embedded method. To ensure that some features can be removed without losing too much information, we created a Pearson's correlation matrix of the features and plotted its heat map.

Pearson's Correlation Heatmap
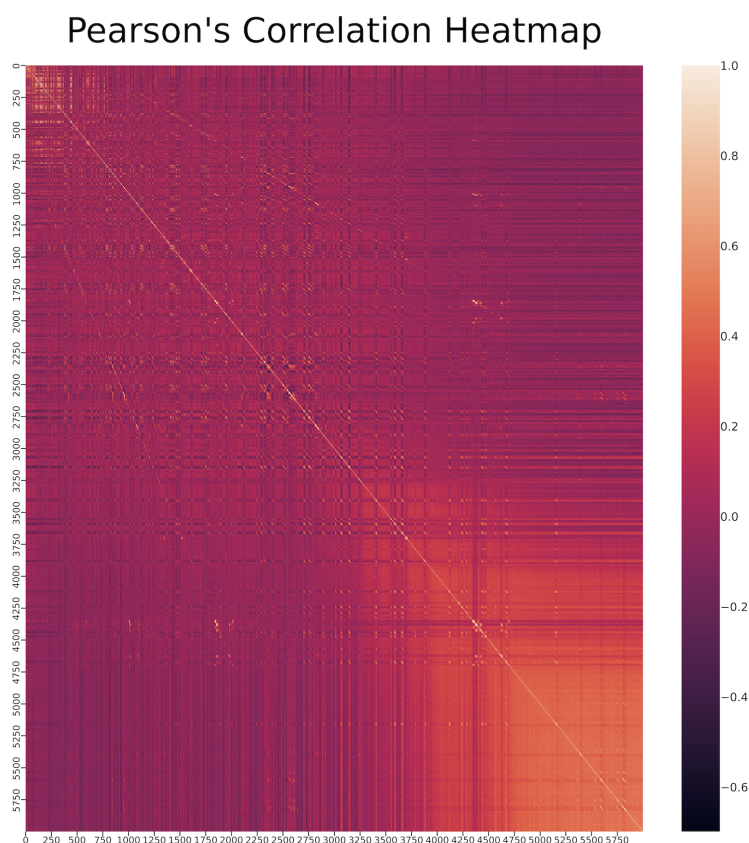
qei0gfnoiqnegfqefqgionwhrb0ioegqiapnbehtpogjrfirwnioaegbg wrg0jaev09bsonveiajpvads e90ajgf0'kAEVPOIMSJBPRGAEFKwopeavdsfbkpm

SVM:

aafef

Random Forest:

aafef

Logistic Regression:

aafef

KNN:

aafef

Results:

aafef

References

[1] Study on antibiotic resistance on bacteria: `https://pmc.ncbi.nlm.nih.gov/articles/PMC5629843/`
[2]