Checkpoint 1 Guide: Problem Definition, Data Gathering, KPIs

Problem Definition

Antibiotics are prescribed and given to patients to counter bacterial infection. In some circumstances, a patient's specific bacteria may be tested against antibiotics to help determine which antibiotics should be prescribed. Since lab tests for susceptibility take 24-48 hours for the test results to come in, having a predictive model that could help decide on the patient visit could be impactful. Additionally, some circumstances may not involve culture testing (like in smaller or rural clinics), so this information would be useful in those situations too. The goal is to identify patient-level features that may influence the resistance of a bacteria to a given antibiotic. This is to give doctors and patients both information that may influence their decision making when prescribing an antibiotic for identified organisms. It can also raise awareness of which factors affect antibiotic resistance the most for any healthcare provider or national/international health agency.

We are looking at the antibiotic resistance microbiology dataset (ARMD) found at https://www.nature.com/articles/s41597-025-05649-7 We use the April 11, 2025 version of the data.

This data was collected from the Stanford Hospital electronics records from two university-associated hospitals. This dataset includes data from 2007 to 2024. The data includes patient level information such as age, sex, socioeconomic disadvantage, labs done at time of visit, etc. Additionally, they have organism identified, and culture analysis of whether the identified organism was resistant or susceptible to tested antibiotics. We will not be addressing all antibiotics and organism combinations, to ensure that there are sufficient data amounts. We will also not be using all features from the dataset, either because we believe they may not be relevant, or because the datasets are too large to maneuver the data.

The data is contained in multiple related .csv files that can be indexed using their IDs and datetime of visit. It is a one-time download. The link above contains the article with the data description, and the link to the dryad website for csv downloads (https://datadryad.org/dataset/doi:10.5061/dryad.jq2bvq8kp). All data has been de-identified for use by other researchers.

The original dataset has over 280,000 unique patients that cover resistance data for over 2 million organism/antibiotic combinations. This dataset will be reduced based on early data analysis and visualizations. Given the data collection process, the data is representative of those who live near the Stanford Hospital system.

We will use different models for predicting classifiers. The best performing model will be used for feature selection, to determine which of the selected features could potentially predict antibiotic resistance outcome. For determining which model will be best, we'll consider accuracy, precision, recall and the F1 score. We will use 5-fold cross-validation.

Deliverables:
Read me file containing this information.
Data folder with original data, and data cleaning scripts, with the final data contained in csv files
EDA scripts
Model scripts
Feature importance script
Conclusions and presentation