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.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.

Data description:

Based on EDA, we have focused the dataset down to certain core features. First, we are only looking at E. Coli, which is the largest represented organism in the dataset, for simplicity of modeling and resistant mechanisms are organism dependent. We choose data from 2016 onward because we want more recent data, to more accurately represent today's resistant patterns and because 2016 had a shift in "standard" antibiotics tested (old vs. new) so should only be considering data from past 2016. Second, the target antibiotics that we would like to predict resistance of have been reduced to the antibiotics that were tested in at least 75% of the E. Coli data. This is to ensure there are appropriate data amounts for each antibiotic, and also to better focus the question on antibiotics that patients are more likely to receive for E. Coli.

The model input contains features that include, for a given antibiotic:

- Age of patient
- ADI score of patient
- Ward information antibiotic was given
- Previous exposure to an antibiotic - 0: not exposed, 1: exposed
- Previous exposure to any organism - 0: not exposed, 1: exposed to E. Coli, 2: exposed to something else
- Whether the patient has visited a nursing home in the last 90 days

The model output is the resistance of a given antibiotic for each patient (1: susceptible, 2: resistant).

Four model approach: Logistic regression, SVM Classifier, RandomForest, XGBoost
All four models are trained and evaluated using the same framework.
Their performance on the held-out test set is compared, and the model with the lowest FNR is selected as the final model. Feature selection is performed on the best model, to determine which features most contribute to the outcome.

Metrics for Best Model:

1) Misclassifying resistant patients as non-resistant is clinically critical.
   The baseline metric is the False Negative Rate (FNR):
   $FNR = FN / (TP + FN) = \#(\hat{y} = 0 \,|\, y = 1) / [\#(\hat{y} = 1 \,|\, y = 1) + \#(\hat{y} = 0 \,|\, y = 1)]$
   This measures the probability of predicting "susceptible" when the patient is actually

resistant.

The goal is to minimize the FNR.

2) Accuracy score for prediction (% of total correctly classified)
3) Precision score, out of all of the ones that were predicted resistant, how many were actually resistant (false positives score)
4) F1 Score (determined by precision and recall score)
5) 5-fold cross-validation (stratified by resistant/susceptible, random_state=, test_size = 0.2)

Train–Test Split:

The dataset of 24,000 rows is divided into 80% training (19,200) and 20% testing (4800, held out).

Within the training set, a 5-fold cross-validation is applied: four folds for training (≈45,611 rows) and one for validation (≈11,403), rotated iteratively.