Predicting Antibiotic Resistance

Chiara Mattamira, Dominique Hughes, Haejun (Stella) Oh, Mustafain Ali, Tinghao Huang

Executive Summary

Introduction

Infections that could once be easily cured by simple antibiotics are becoming harder to treat due to emerging antibiotic resistance in bacteria. In clinical practice, physicians typically prescribe medications based on their past experience and clinical guidelines while the laboratory runs diagnostic tests to determine antibiotic effectiveness. Our project aims to bridge this gap by aiding the decision making process through a predictive model that can anticipate antibiotic resistance using patient-level data.

Dataset:

We utilize the **Antimicrobial Resistance and Microbiology Dataset (ARMD)** that contains over 750,000 microbiological culture records from ~284,000 unique patients tested across 55 different antibiotics. For this study, we include records from 2016 onwards to ensure consistency in antibiotic testing standards and to ensure that our predictions reflect contemporary resistance patterns relevant to the current clinical practice.

Furthermore, we focus exclusively on E. coli isolates (the most common organism in the dataset) and restrict our analyses to nine antibiotics tested in all of E. coli samples. Based on published clinical guidelines, we select age, socioeconomic disadvantage (ADI score), hospital ward types, prior antibiotic exposure, prior infection types and recent nursing home visits as the key features for determining antibiotic resistance.

Methods:

After pre-processing our dataset to remove duplicate entries, one-hot encoding categorical variables and keeping only unique patient IDs, we train six different models (dummy classifier, logistic regression, random forest, XGboost, support vector machine and K-nearest neighbors) to predict whether an E.coli isolate is resistant or susceptible for the nine commonly used antibiotics.

The modeling pipeline uses an 80/20 train–test split with a nested cross-validation framework i.e. an outer 5-fold CV for model selection and an inner 3-fold CV with Randomized Search for hyperparameter tuning. StandardScaler is applied to distance-based models for better feature comparability, while SMOTE is used in the pipeline to address class imbalance by oversampling resistant cases and downsampling is used to reduce the susceptible(negative) cases. Afterwards, the best performing model is validated on the held-out test set.

Results:

Model performance was evaluated using accuracy, precision, recall, F-1 scores and the False Negative Rate (FNR: resistant cases misclassified as susceptible). From a clinical perspective, minimizing the FNR was considered a critical performance indicator.

Among all the models, **Logistic Regression** achieved the highest accuracy of 62.0%, the highest F1 weighted score of 75.4% and the lowest FNR score of 38.3%. It also demonstrated a clear interpretation of features, yielded consistent calibration curves and comparable results on the training and test datasets. Moreover, the hyperparameters were also rigorously tuned and evaluated on the test data as well.

Conclusions and Future Work

Our chosen model achieved reasonable accuracy but was limited by the inherent nature of the dataset used. Not all factors that contribute to resistance can be captured with limited patient-level features. Additionally, simpler models underfit complex patterns, while ensemble methods risk overfitting and metric bias, collectively constraining performance.

This work can be extended to study time-series modeling to track resistance trends, feature-importance exploration using richer datasets, and cross-antibiotic correlation analysis to uncover linked resistance mechanisms.