Report for “Applied Machine Learning for Biological Problems” project 2 - Predicting Antibiotic resistance

Battisti Nicola, Efthymiadis Georgios, Predl Michael, Trinh Philipp

Contents

Introduction

For Project 2, our goal was to predict antibiotic resistance of *Neisseria gonorrhoeae*, the causative agent of gonorrhea. The “antibio” dataset consists of genomic patterns of a subset of *Neisseria gonorrhoeae* strains, that are associated with resistance to the three different antibiotics: azithromycin, ciprofloxacin and cefixime. Specifically, we analyzed 3971 *Neisseria gonorrhoeae* strain genomes, focusing on the DNA segments statistically associated with resistance to the aforementioned antibiotics. These genome patterns were utilized to train our Machine Learning model.

We’ve devised a highly accurate machine learning approach built upon a robust pipeline covering all stages of the data science process. Our approach involved exploratory data analysis (EDA), data preprocessing, including duplicate detection, pinpointing of missing values, outlier observation, and handling low-information features.

The development of high accuracy models is crucial for guiding clinical decision making. Accurate predictions of antibiotic resistance enables fast and effective treatment plans, improving the management of gonorrhea infections. Moreover, the model supports surveillance efforts by detecting emerging antibiotic-resistant strains early.

Methods

The project was conducted with the utilization of Google Colab, where a collaborative Jupyter notebook was developed. The major Python packages utilized included Pandas, Scikit-learn, Numpy, Matplotlib, Seaborn, and Auto-Sklearn. Auto-Sklearn was employed as a tool of automated machine learning (AutoML). Initially, the first step involved data loading, exploratory data analysis, as well as preprocessing using the packages Pandas & Matplotlib. For more meaningful order, the data were transposed for downstream analysis. Subsequently duplicate observations and features were detected and addressed. Additionally, low-information features were identified and removed. Outliers were detected using statistical methods and removed. Moreover, feature correlation was analyzed through correlation matrices and visualizations.

Finally, the dataset was split into training and testing sets (70% train / 30% test), and a random forest model trained as baseline (on original data with only NaN values removed), as well as with the optimized dataset. Then, Auto-Sklearn was employed to automatically select and optimize machine learning models for each antibiotic.

Results

After a holistic data preprocessing, the results fine-tuned demonstrated significant improvements in the quality of the datasets for all three antibiotics. The observation of duplicates were successfully identified and removed, resulting in curated and more reliable datasets. Furthermore, low-information features were appropriately handled, for a more robust approach. Outliers were detected, ensuring the data's integrity. Feature correlation analysis provided insights into the relationships between different genomic patterns. The application of various classifiers yielded very promising results, with trained models exhibiting high accuracy (above 95% for all cases) in predicting antibiotic resistance.

Dataset description

The dataset contains 3971 genomes, for which the susceptibility to azithromycin (AM), ciprofloxacin (CF) and cefixime (CX) were measured in 3478, 3088, and 3401 cases respectively. Based on the genomes, 515 (AM), 8873 (CF) and 384 (CX) kmers were provided as features. While the number of genomes tested for resistance against each of the antibiotic compounds is similar, the fraction of resistant genomes is not (see table 1). Ciprofloxacin shows the highest fraction of resistant strains with almost 50%. Resistance to azithromycin is less prevalent with roughly 12% of strains. However, only 5 cases of resistance to cefixime were recorded. With such extensive target imbalance and almost no data points for one of the categories, we could not see a possibility to arrive at a robust and trustworthy model. Therefore we excluded cefixime resistance from the final modeling stage. The goal of predicting resistance to the three antibiotics can be achieved with two types of machine learning models: binary classification and multicategorical classification. As the feature-spaces of the three antibiotic compounds show almost no overlap, a multicategorical classifier could not benefit from their combination. Instead, we opted for training separate models to predict the resistance.

Resistance	Azithromycin	Ciprofloxacin	Cefixime
No	3031	1660	3396
Yes	447	1428	5

Table 1. Overview of resistance / susceptibility to the three antibiotics azithromycin, ciprofloxacin and cefixime in the genomes of the dataset.

Feature selection

Feature selection played an important role in the optimization of the datasets. The curation of the datasets from this point of view aimed to enhance model interpretability while preserving the essential genomic information required for accurate predictions. Duplicate features were identified and merged, retaining essential k-mer information. This process involved checking for features that were identical to the target value, in order to avoid potential leakage. Low-information features, i.e. those with all or all but one being absent or present, were identified and then removed, for the reduction of noise and model enhancement. The removal of observations affected both target categories roughly to the same extent (see figure 1).The correlation of all features to the target variable was calculated, which revealed some features with as much correlation as 72.1% (AM), 74.7% (CF) and 18.7% (CX) (see table S1).

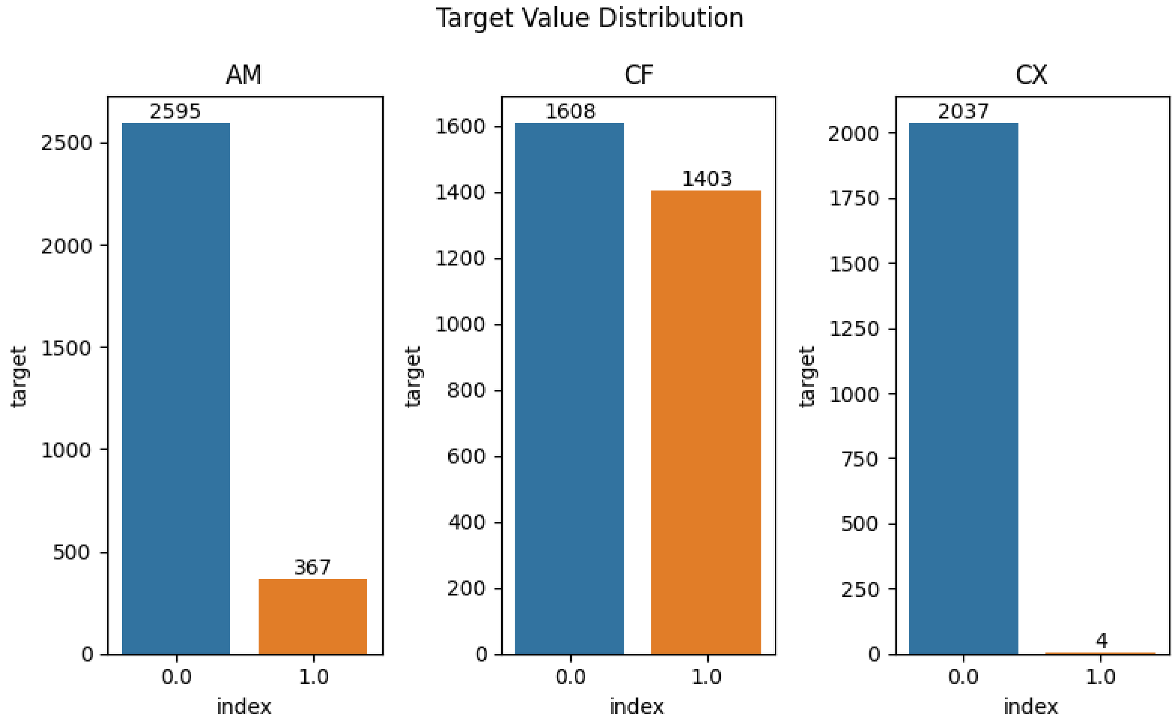

Figure 1. Comparison of target value distributions after data preprocessing. Encoding of the target variable antibiotic resistance is 1: resistant and 0: susceptible.

Model selection/optimization

For the baseline model, a random forest classifier was chosen. Our baseline model was constructed using balanced class weights, and 100 estimators. The baseline models were trained separately for each antibiotic (except cefixime), utilizing the corresponding datasets including only observations with resistance measurements. For the revised model, the random forest classifier was used again with the same hyperparameters, but on the fully preprocessed datasets. The training process was iterated 100 times, and error curves were plotted in order to assess model stability. Moreover, confusion matrices were generated for the evaluation of the model's performance. For the auto-Sklearn model (automated machine learning), the process included fitting the model, generating a confusion matrix, and finally evaluating the classification report. The AutoSklearn2Classifier was used with a memory limit of 12GB and a time limit of 10 minutes.

Performance evaluation

The baseline model resulted in very high accuracy of over 95% on both azithromycin and ciprofloxacin (see table 2). The optimized models performed slightly worse in terms of accuracy. The best accuracy was reached with models built with auto-sklearn. However, it should be noted that these are measurements based on a single model and train / test split, compared to the average over 100 models in the random forest models. An optimized random forest model was also built for cefixime. This model achieved an average accuracy of over 99%, but due to very low positive target observations and high target imbalance, this model is likely prone to overfitting and unfit to generalize to unseen resistant strains.

Baseline Models (RF)			
Antibiotic	Azithromycin	Ciprofloxacin	Cefixime
Average accuracy	0,9629	0,9652	NA
Optimized Models (RF)			
Antibiotic	Azithromycin	Ciprofloxacin	Cefixime
Average accuracy	0,9592	0,9632	0,9967*
Optimized Models (AS)			
Antibiotic	Azithromycin	Ciprofloxacin	Cefixime
Accuracy (1 measurement)	0,9696	0,9690	NA

Table 2. Comparison of model performance. Abbreviations: RF - random forest classifier, AS - auto-sklearn classifier. The average accuracy of the cefixime resistance model is given for completion’s sake, it needs to be noted that some train test splits used in training and evaluating the model did not contain cases of resistance in one of the splits.

Conclusion & further directions

In this report we could show high accuracy machine learning models for predicting antibiotic resistance in *Neisseria gonorrhoeae*. Of the three antibiotics, robust models could be built for azithromycin and ciprofloxacin. The dataset of cefixime contained only four observations of resistance, after deduplication of the data. More data collection, especially of strains resistant to cefixime, is required for robust predictions.

We curated the dataset rigorously, removing noise and potential data leakage. It is important to note that the revised random forest models, while maintaining similar accuracy, also showed stability improvements. Models built by auto-sklearn resulted in the best performance, but further evaluation over multiple train test splits is required to be certain.

When considering further improvement of the models, tuning them towards the specific application would be an important direction. Considering the already high accuracy, focusing the improvement of the model predictions on avoiding errors with bad impact in a clinical setting is advisable.

During data preparation, several kmers with high correlation to antibiotic resistance were identified. These are leads for further investigations of what conveys the antibiotic resistance in these strains. They could also be used as markers to trace the evolution of antibiotic resistance among the *Neisseria gonorrhoeae* strains.

This work shows that antibiotic resistance screening based on genomic features is well possible. For fruitful translation to medical applications, the limited scope of this dataset needs to be addressed. The dataset of this work contains data of only one species and strains of only one outbreak. Generalisability - if desired - needs to be tested with further data.

In clinical or time-sensitive settings, there is a need for a rapid method to extract information about the presence and absence of kmers for model input. The sequencing and assembly processes can be slow, making it essential to establish a fast method for extracting this information. If relevant kmers can be directly measured and confidently attributed to the pathogen, especially considering their potential high abundance as indicated by kmer occurrence, this approach could significantly speed up medical decision making.

With the increasing prevalence of antibiotic-resistant pathogen strains, machine learning will become an essential tool for rapid identification of the optimal antibiotic agents to determine life-saving medical interventions.

Supplements

Supplemental table S1 - Top 5 correlating features

Azithromycin	
Kmer	Correlation to resistance
GGGTTTAAAACGTCGTGAGACAGTTTGGTCCCTATCTGCAGTGGGCGTTGGAAGTTTGACG	0.7207139190182661
CATCTGCCTGGCAAACGCTTCCCCGTCGCCCTCGAA	0.4766997443941503
AACATCAGGAAAACGGCACATTCCACGCCGT	0.43074261195447505
TTTCAGACGGCATCTGCCTGGCAAACGCTTCCC	0.39926599909208876
AACTTTCTGAACATATTTGCCTTTGATTTCG	0.3111762285077589
Ciprofloxacin	
Kmer	Correlation to resistance
AAATTGCGGATCGATGCGCGAAGGGTCGAATGC	0.7470544207574276
CTTGGCACACAGTACCGAACCGGCGGCAATACCGATG	0.7171277632560408
TATACTGCCGTTCAAGTTACCTTTGGTCAGAAAAAA	0.6971484219184869
GTTGTGCCGAATATTGCACCGATGGAAAGGGGGAGGATGTT	0.6968027072259367
CAGGCGGAAATATAGTGGATTAAATTTAAAC	0.685167439702747
Cefixime	
Kmer	Correlation to resistance
CCTGATGGGGAATGCAAGTTATGCGATTGCC	0.18689331632415798
AAACGACTTGCCTTTCTGTTCGAGTTTGTTCCTCAAAACATC	0.10431690197017547
GTAATGGCGTTTTAATTCTTTTTCAAATACAAAGTT	0.1033411953657866
CAAACAACTTCTATTACGCCAAAAGTCAGGCGATGCTCTACACC	0.07856923208931954
TCTTCAATAAAGGACAAAGACAGGAGGCGCGA	0.07585324677713323