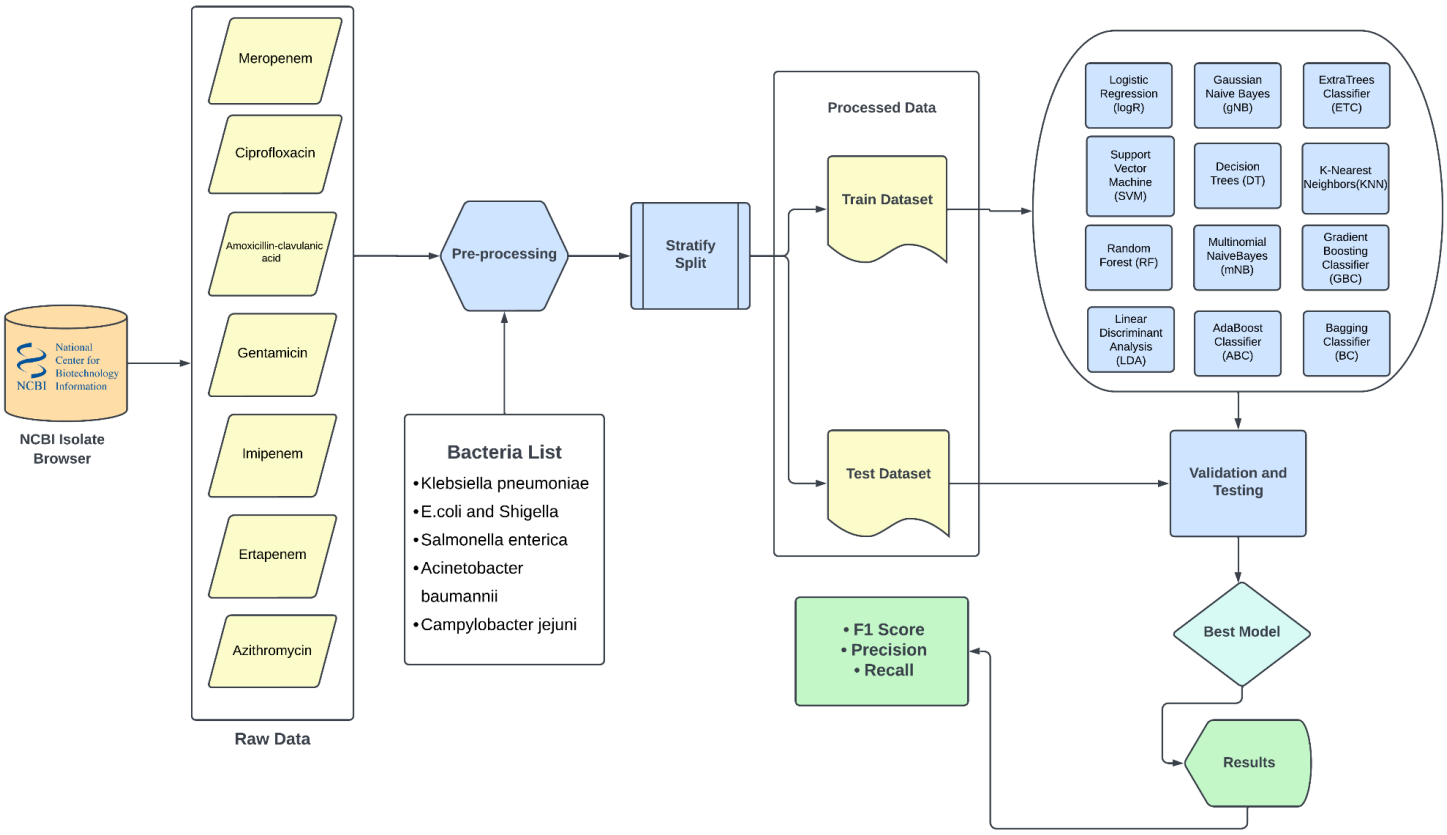
Antimicrobial Resistance Prediction Using Statistical ML Models



View Diagram here: Click Here

Datasets:

Bacterial strains, AMR genotypes and AST phenotypes

Website: Isolates Browser - Pathogen Detection - NCBI

Raw Datasets:
https://github.com/RAGUL1902/AMR_Prediction_Project/tree/master/raw_data_from_NCBI

Processed Datasets:
https://github.com/RAGUL1902/AMR_Prediction_Project/tree/master/datasets

Information about the datasets:
https://github.com/RAGUL1902/AMR_Prediction_Project/blob/master/datasets_info.csv

The NCBI Isolate Browser is a web-based tool developed by the National Center for Biotechnology Information (NCBI) that allows users to explore and analyze microbial genomes and their associated metadata.

Bacterial genus and species were selected in the 'organism group', with filters checked for 'has AMR genotypes' and 'has AST phenotypes'. Then the isolates were downloaded for different antibiotics separately.

The Collected Antibiotics are as follows:

- Meropenem
- Ciprofloxacin
- Amoxicillin-clavulanic acid
- Gentamicin
- Imipenem
- Ertapenem
- Azithromycin

Once this data was collected, the following list of bacteria isolates were segregated from each antibiotic dataset.

- Klebsiella pneumoniae
- E.coli and Shigella
- Salmonella enterica
- Acinetobacter baumannii
- Campylobacter jejuni

The combination of these bacteria and antibiotics were separated and made into different datasets. Based on the AMR genotypes and AST phenotypes, binary matrices of genotypes (0 for absence and 1 for presence of an AMR gene) and relevant antibiotics' phenotypes (0 for susceptibility and 1 for resistance to an antibiotic) were created.

Prediction - Machine Learning Models

Code for statistical models:
https://github.com/RAGUL1902/AMR_Prediction_Project/blob/master/Statistical_models_on_all_datasets.ipynb

Machine learning in Python with Scikit-learn (https://scikit-learn.org/stable/) was used to evaluate the performance of 12 machine learning algorithms, namely,

- Logistic Regression (logR),
- Gaussian Naive Bayes (gNB),
- Support Vector Machine (SVM),
- Decision Trees (DT),
- Random Forest (RF),
- K-Nearest Neighbors(KNN),
- Linear Discriminant Analysis (LDA),
- Multinomial NaiveBayes (mNB),
- AdaBoost Classifier (ABC),
- Gradient Boosting Classifier (GBC),
- ExtraTrees Classifier (ETC),
- Bagging Classifier (BC).

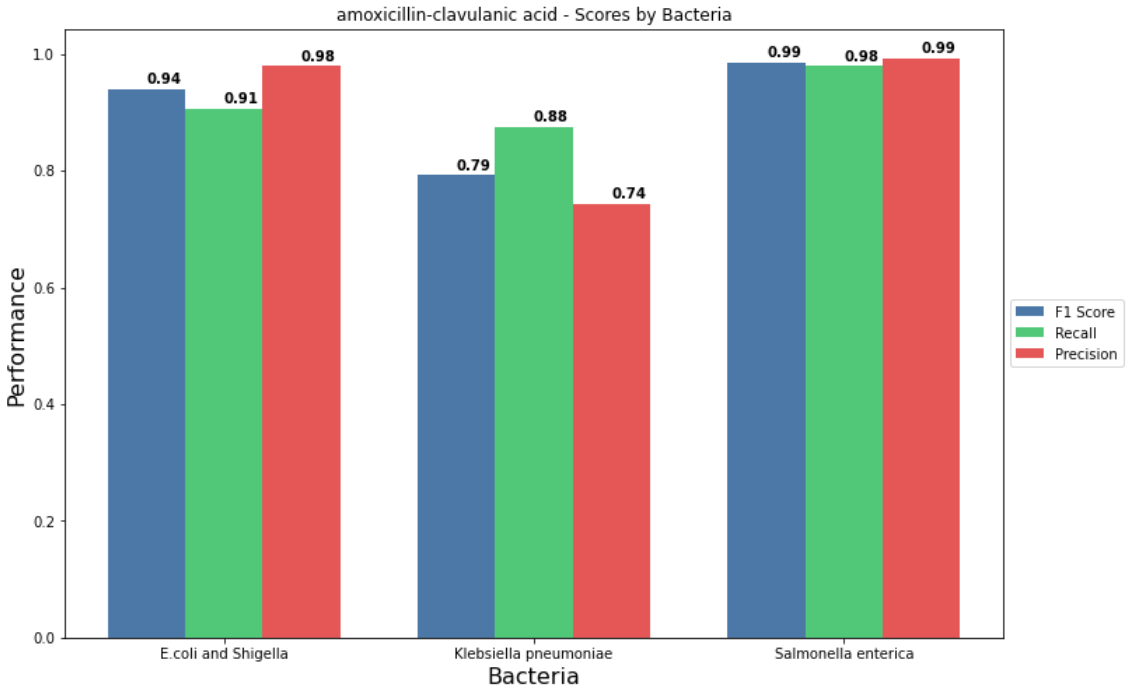
We used the bacterial strains' AMR genotype and AST phenotype data for training and testing the machine learning algorithms. The performance metrics include precision, recall, F1-score and Confusion Matrix.

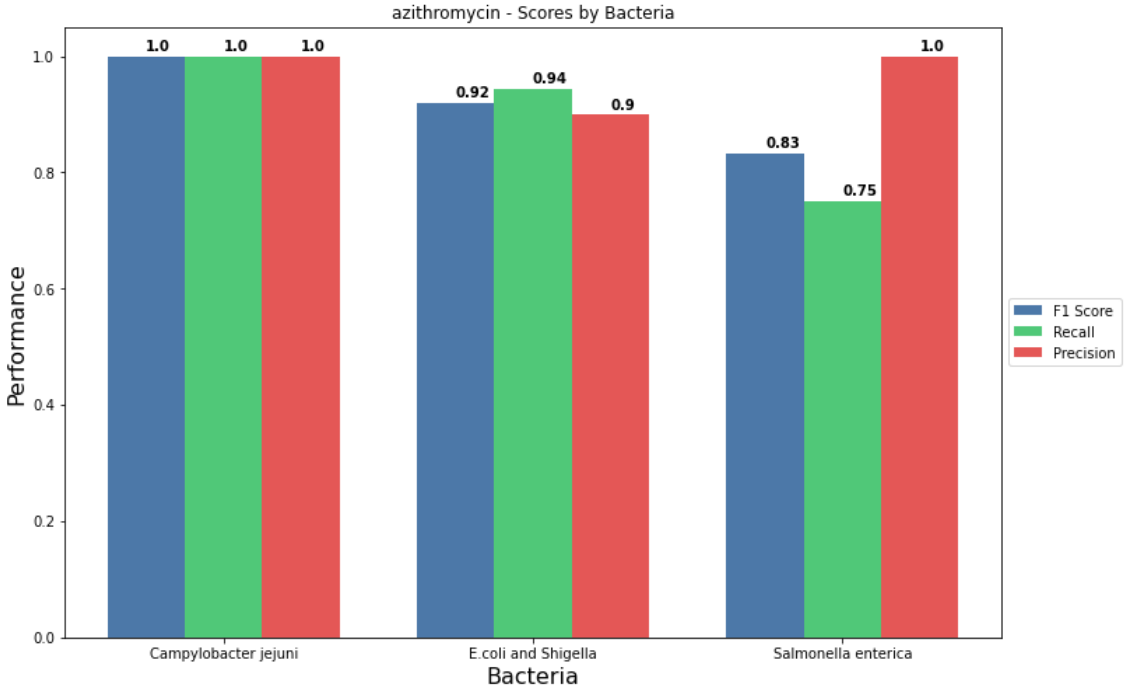
Results

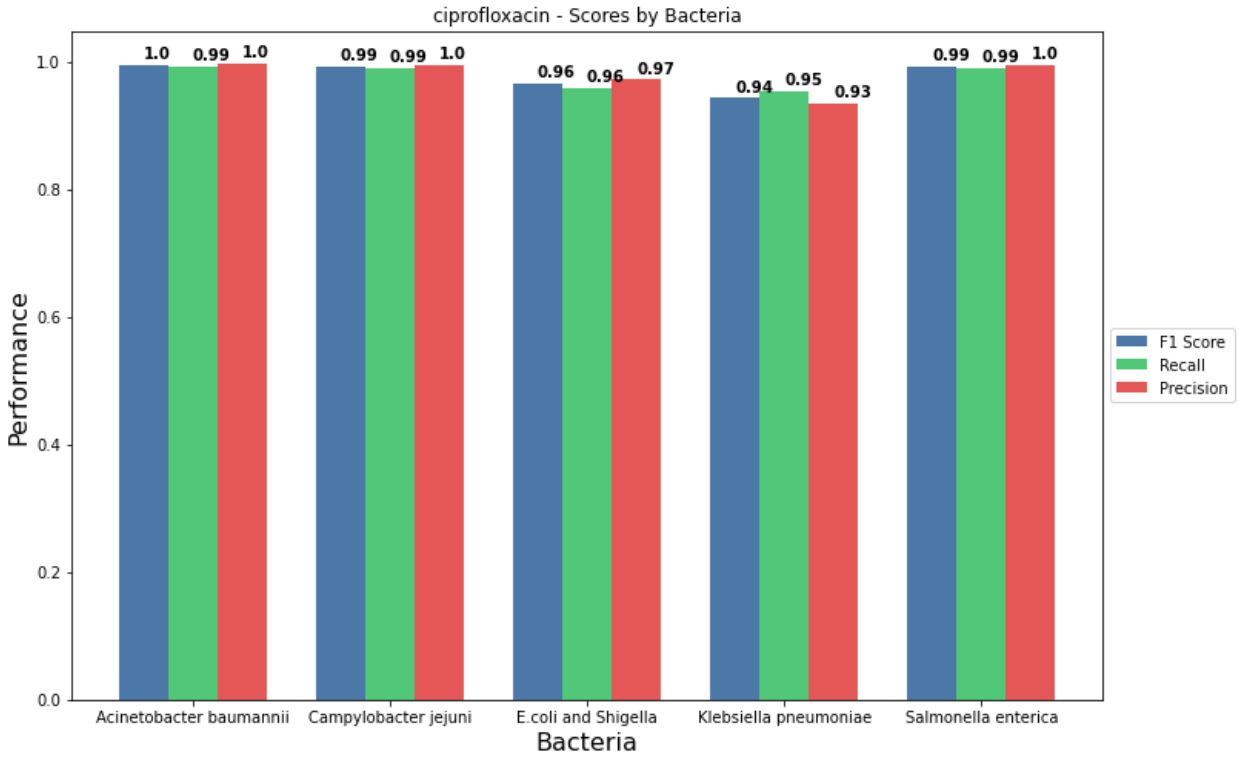
Results - Tabular format:
https://github.com/RAGUL1902/AMR_Prediction_Project/blob/master/Best_statistical_model_on_all_datasets.csv

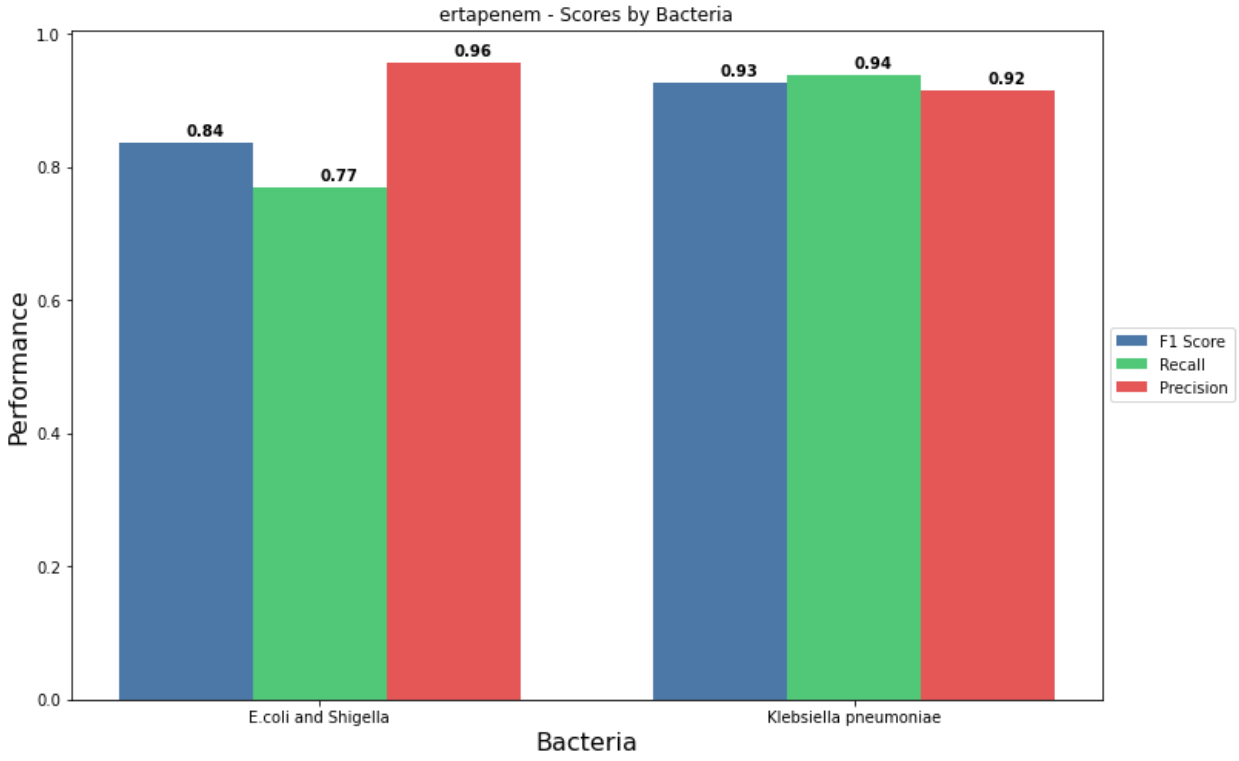
Results - Plots:
https://github.com/RAGUL1902/AMR_Prediction_Project/blob/master/plot_results.ipynb

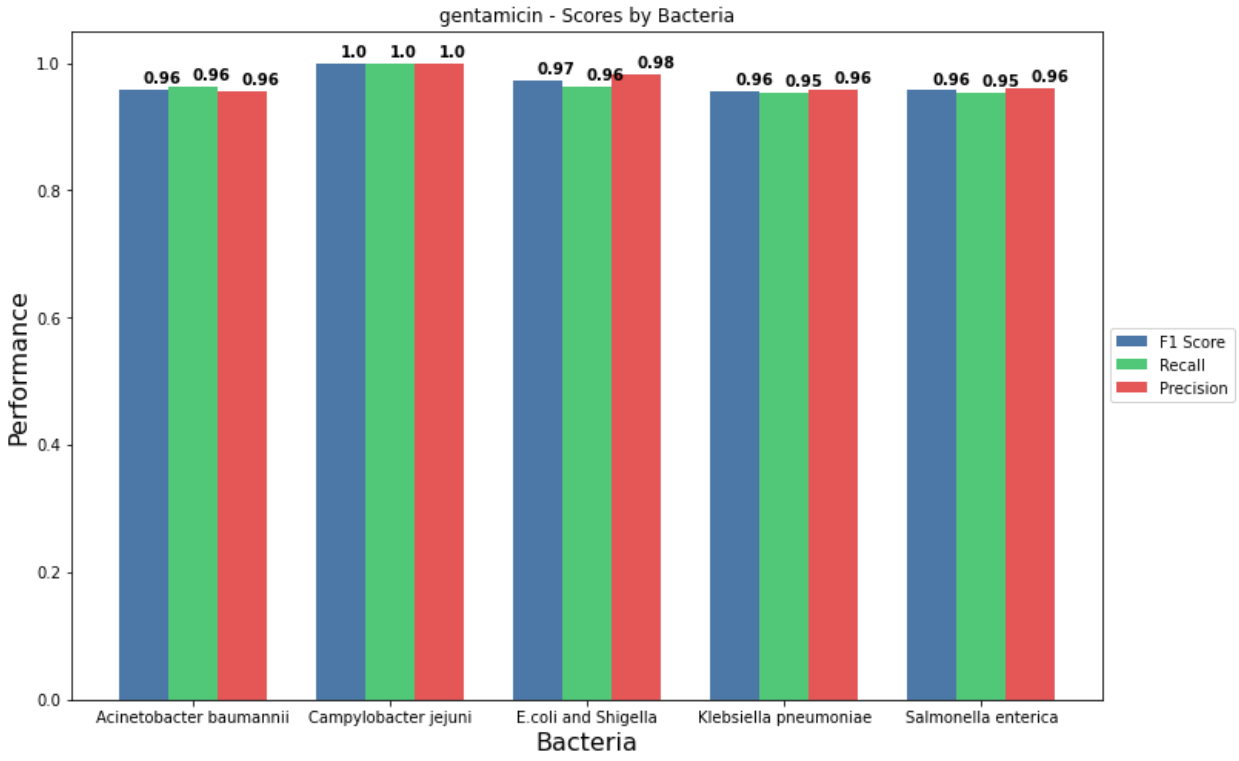
Assessment of different machine learning algorithms on the entire dataset of bacterial strains with known genotypes and phenotypes showed there is no single optimal machine learning model for predicting resistance phenotype across all bacterial species considered for the prediction. Added the Results as bar plots:

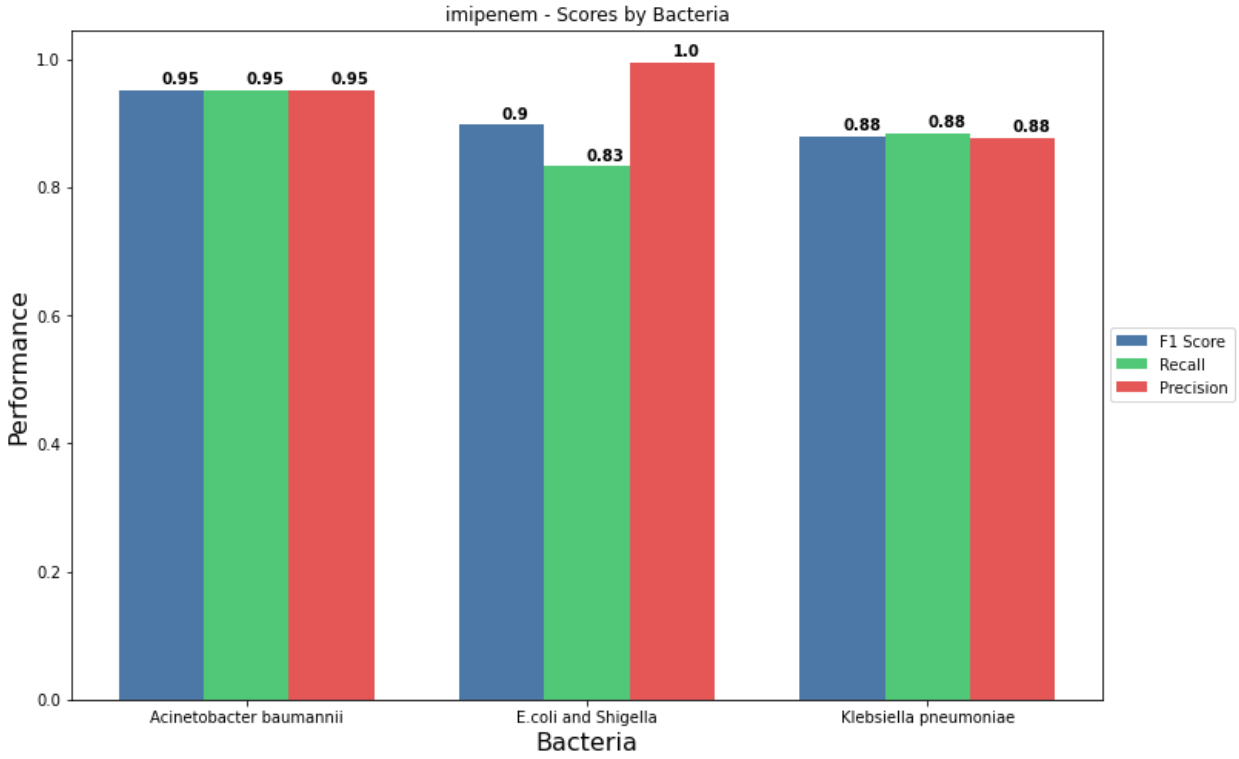

amoxicillin-clavulanic acid - Scores by Bacteria
Performance
Bacteria
E.coli and Shigella
Klebsiella pneumoniae
Salmonella enterica
0.94
0.91
0.98
0.79
0.88
0.74
0.99
0.98
0.99
F1 Score
Recall
Precision

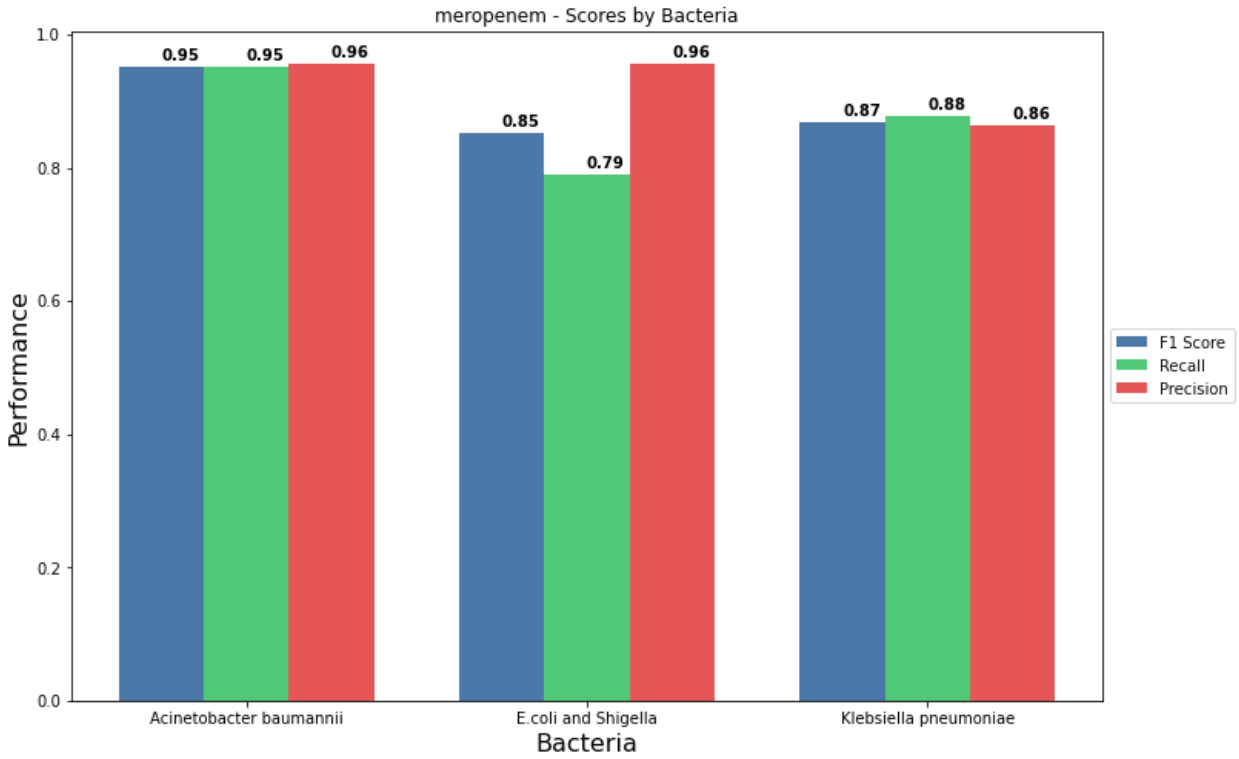

azithromycin - Scores by Bacteria
Performance
Bacteria
Campylobacter jejuni
E.coli and Shigella
Salmonella enterica
1.0
1.0
1.0
0.92
0.94
0.9
0.83
0.75
1.0
F1 Score
Recall
Precision


ciprofloxacin - Scores by Bacteria
Performance
Bacteria
Acinetobacter baumannii
Campylobacter jejuni
E.coli and Shigella
Klebsiella pneumoniae
Salmonella enterica
1.0 0.99 1.0
0.99 0.99 1.0
0.96 0.96 0.97
0.94 0.95 0.93
0.99 0.99 1.0
F1 Score
Recall
Precision


ertapenem - Scores by Bacteria
Performance
Bacteria
E.coli and Shigella
Klebsiella pneumoniae
0.84
0.77
0.96
0.93
0.94
0.92
F1 Score
Recall
Precision


gentamicin - Scores by Bacteria
Performance
Bacteria
F1 Score
Recall
Precision
0.96 0.96 0.96
1.0 1.0 1.0
0.97 0.96 0.98
0.96 0.95 0.96
0.96 0.95 0.96
Acinetobacter baumannii
Campylobacter jejuni
E.coli and Shigella
Klebsiella pneumoniae
Salmonella enterica


imipenem - Scores by Bacteria
Performance
Bacteria
F1 Score
Recall
Precision
0.95 0.95 0.95
0.9 0.83 1.0
0.88 0.88 0.88
Acinetobacter baumannii
E.coli and Shigella
Klebsiella pneumoniae


meropenem - Scores by Bacteria
1.0
0.8
0.6
0.4
0.2
0.0
Performance
0.95
0.95
0.96
0.85
0.79
0.96
0.87
0.88
0.86
F1 Score
Recall
Precision
Acinetobacter baumannii
E.coli and Shigella
Klebsiella pneumoniae
Bacteria