Predicting Antimicrobial Resistance for Common Pathogen-Antibiotic Pairs

AI - 395T | AI in Healthcare | Dr. Ying Ding

By- Aayushee Agarwal and Sharang Agarwal
(aa222655, sa62567)

Background and Motivation

Background:

- Antimicrobial resistance (AMR) is a major global health threat.
- AMR increases morbidity, mortality, and healthcare costs.
- Timely antibiotic administration is crucial in critical care.
- Current culture-based methods for determining susceptibility take 24-72 hours.

Motivation:

- Machine learning offers the potential for earlier AMR prediction.
- Early prediction can support targeted empirical therapy.
- This can improve patient outcomes and can mitigate the spread of resistance.

Research Question and Objectives

Research Question:

Can we accurately predict antimicrobial susceptibility for common pathogen-antibiotic pairs using early clinical data?

Objectives:

- Develop a machine learning pipeline
- Evaluate model performance on common pathogen-antibiotic pairs
- Identify important predictive factors

Dataset

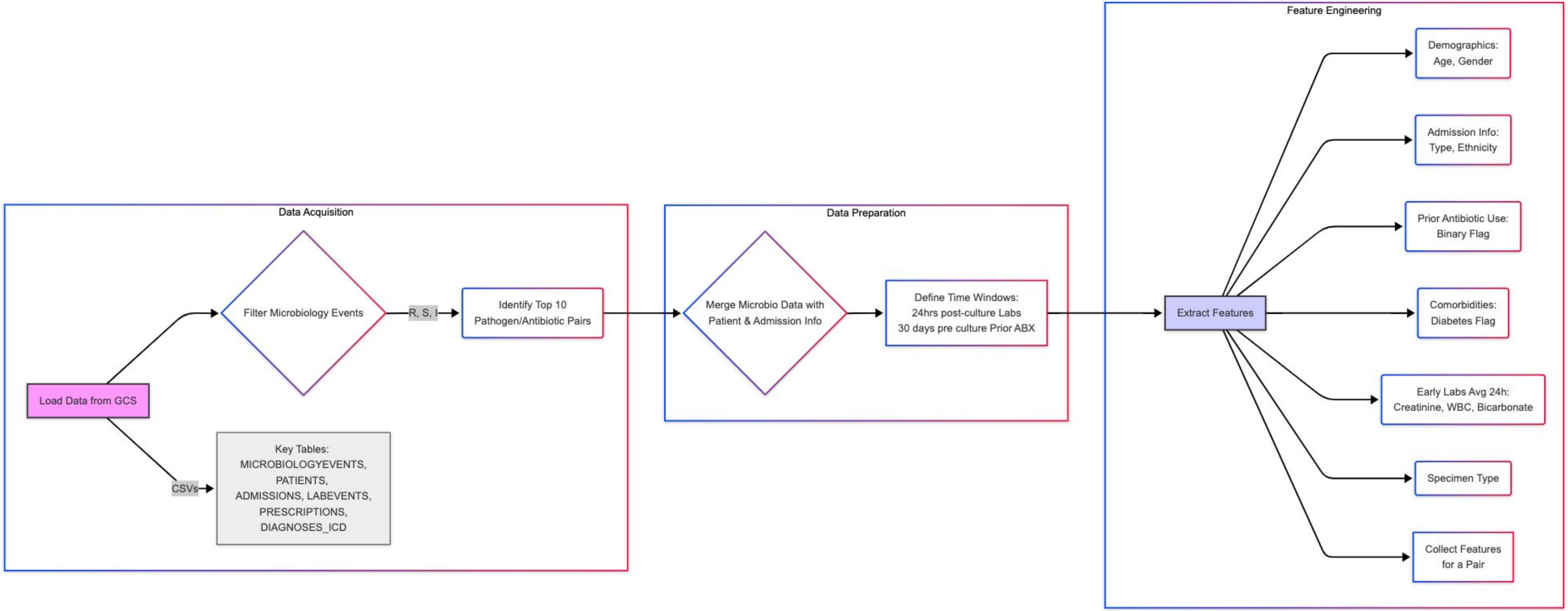
Data Source: MIMIC-III

Data tables used:

- MICROBIOLOGY EVENTS
- PATIENTS
- ADMISSIONS
- LABEVENTS
- PRESCRIPTIONS
- DIAGNOSES_ICD

Data access and storage: Google Cloud Storage

Flowchart of Data Preparation and Feature engineering



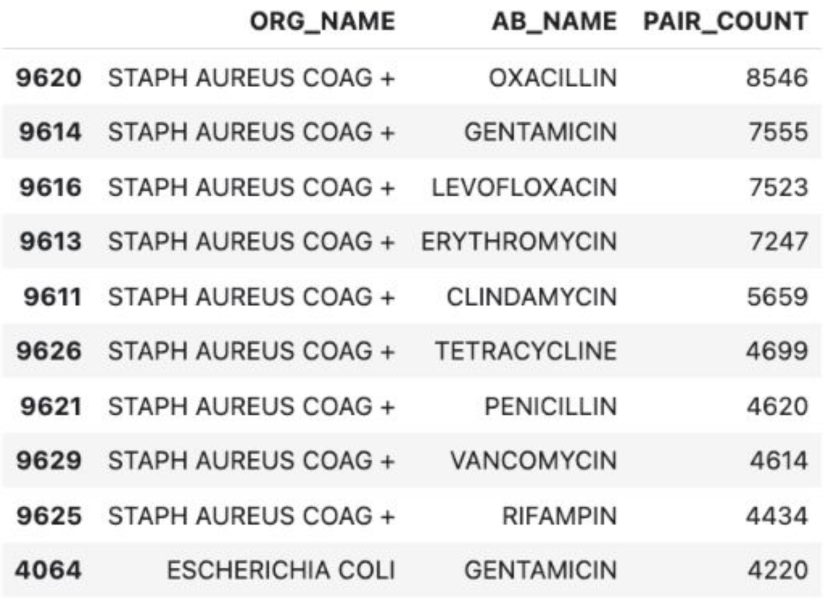
Data Acquisition
Load Data from GCS
Filter Microbiology Events
R, S, I
Identify Top 10 Pathogen/Antibiotic Pairs
CSVs
Key Tables: MICROBIOLOGYEVENTS, PATIENTS, ADMISSIONS, LABEVENTS, PRESCRIPTIONS, DIAGNOSES_ICD
Data Preparation
Merge Microbio Data with Patient & Admission Info
Define Time Windows: 24hrs post-culture Labs 30 days pre culture Prior ABX
Feature Engineering
Extract Features
Demographics: Age, Gender
Admission Info: Type, Ethnicity
Prior Antibiotic Use: Binary Flag
Comorbidities: Diabetes Flag
Early Labs Avg 24h: Creatinine, WBC, Bicarbonate
Specimen Type
Collect Features for a Pair

Data Acquisition and Preparation

- **Data Acquisition:**
  - Loading required CSVs from GCS using specified data types and parsing date columns.
  - Filtering microbiology events and identifying the top 10 organism-antibiotic pairs.

- **Data Preparation:**
  - Iterating through each pair, filtering microbiology events for the current pair, Merging microbiology data with admission and patient data.
  - Calculating patient age from admission and birth dates.
  - Defining time windows for feature extraction (24 hours post-culture for lab results, 30 days pre-culture for prior antibiotic prescriptions).

Top 10 organism - Antibiotic pair

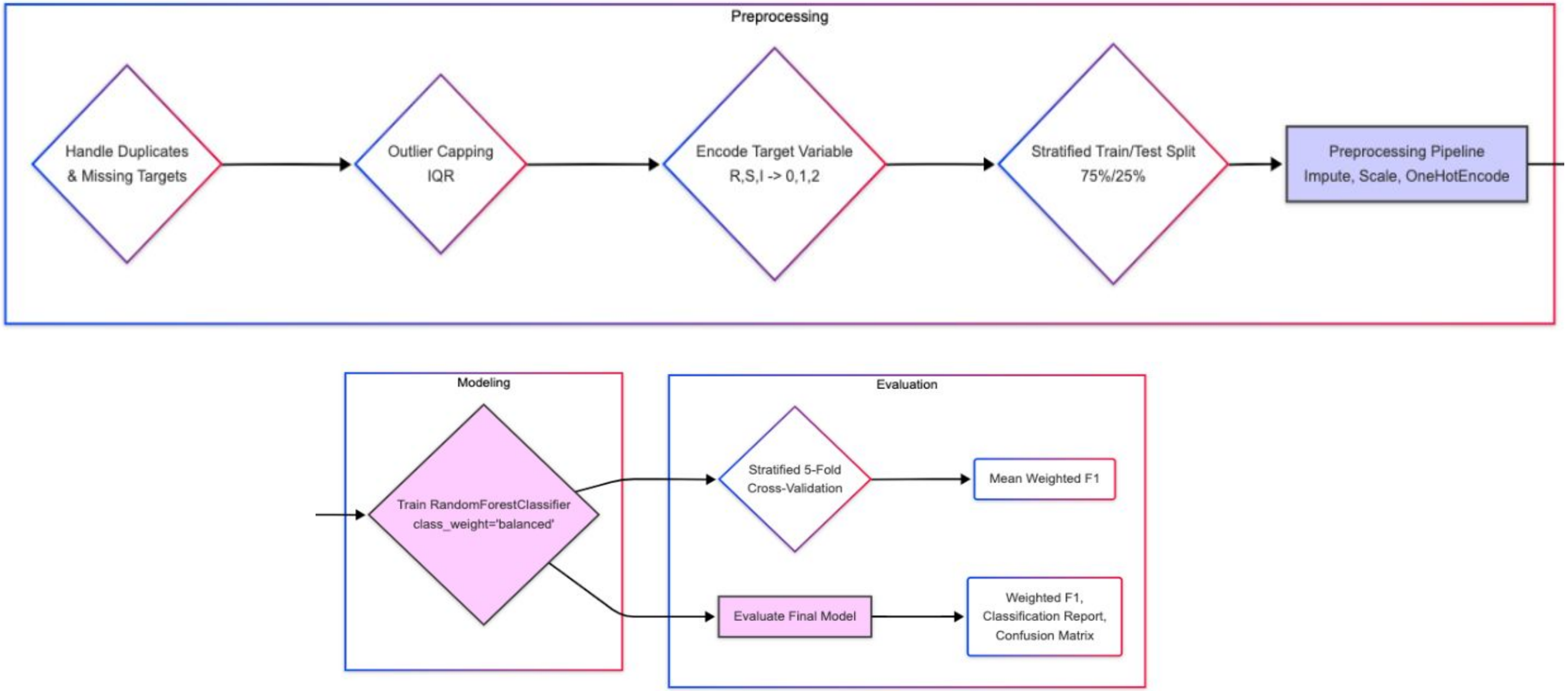


	ORG_NAME	AB_NAME	PAIR_COUNT
9620	STAPH AUREUS COAG +	OXACILLIN	8546
9614	STAPH AUREUS COAG +	GENTAMICIN	7555
9616	STAPH AUREUS COAG +	LEVOFLOXACIN	7523
9613	STAPH AUREUS COAG +	ERYTHROMYCIN	7247
9611	STAPH AUREUS COAG +	CLINDAMYCIN	5659
9626	STAPH AUREUS COAG +	TETRACYCLINE	4699
9621	STAPH AUREUS COAG +	PENICILLIN	4620
9629	STAPH AUREUS COAG +	VANCOMYCIN	4614
9625	STAPH AUREUS COAG +	RIFAMPIN	4434
4064	ESCHERICHIA COLI	GENTAMICIN	4220

Feature Engineering

- **Demographics**:
  - Age at admission (calculated from ADMITTIME and DOB, capped at 90 years)
  - Gender
- **Admission Information**:
  - Admission type
  - Ethnicity
- **Prior Antibiotic Exposure**:
  - Binary flag indicating if any antibiotic prescription started within 30 days prior to the culture time
- **Comorbidities**:
  - Binary flag indicating the presence of diabetes (derived from ICD9 codes starting with '250')
- **Early Lab Results**:
  - Average values of Creatinine, White Blood Cell Count (WBC), and Bicarbonate recorded within the first 24 hours following the CULTURE_CHARTTIME
- **Specimen Type**:
  - The type of specimen collected.
- **Time window for feature extraction** - first 24 hours

Flowchart for Data Preprocessing and Model Evaluation


Preprocessing
Handle Duplicates & Missing Targets
Outlier Capping IQR
Encode Target Variable R,S,I -> 0,1,2
Stratified Train/Test Split 75%/25%
Preprocessing Pipeline Impute, Scale, OneHotEncode
Modeling
Train RandomForestClassifier class_weight='balanced'
Evaluation
Stratified 5-Fold Cross-Validation
Mean Weighted F1
Evaluate Final Model
Weighted F1, Classification Report, Confusion Matrix

Data Preprocessing

- Missing values and duplicate values are handled
- Outlier handling: Capping numerical features at 1.5 times the interquartile range (IQR) below the first quartile (Q1) and above the third quartile (Q3).
- Target encoding: Encoding the categorical target variable (INTERPRETATION) into numerical labels.
- Train/test split: Splitting the dataset into 75% training and 25% testing, stratified by the encoded target variable. Pairs were skipped if the minimum class count was less than the number of CV splits (i.e. 5).
- Preprocessing Pipeline: Column Transformer for numerical (imputation with median, standardization) and categorical (imputation with most frequent, one-hot encoding) features.

Machine Learning Model, Training and Validation

- **Machine Learning Model:** Random Forest Classifier

  Chosen for its robustness and ability to provide feature importances.

- **Model Training and Validation:**
  - Pipeline Integration: Combining the preprocessor and classifier into a single Pipeline object.
  - Cross-validation strategy: Stratified 5-fold cross-validation on the training set.
- **Evaluation Metrics:**
  - Weighted F1-score.
  - Classification report (precision, recall, F1-score for each class).
  - Confusion matrix.

Table 1: Summary of Multiclass Classification Results for Top Organism/Antibiotic Pairs

Organism / Antibiotic Pair	Status	N Samples	Target Distr.	Mean CV F1 (W)	Test F1 (W)	Top 5 Features (Approx.)
STAPH AUREUS COAG+ / VANCOMYCIN	Completed	4123	I: 0.2%, S: 99.8%	0.9970	0.9966	AVG_CREATININE_FIRST24H, AVG_BICARBONATE_FIRST24H, AVG_WBC_FIRST24H, AGE_AT_ADMISSION, SPEC_TYPE_DESC_BLOOD CULTURE
STAPH AUREUS COAG+ / PENICILLIN	Completed	4069	R: 98.4%, S: 1.6%	0.9785	0.9833	AVG_WBC_FIRST24H, AGE_AT_ADMISSION, AVG_BICARBONATE_FIRST24H, AVG_CREATININE_FIRST24H, HAD_PRIOR_ANTIBIOTICS
STAPH AUREUS COAG+ / RIFAMPIN	Completed	3962	I: 0.7%, R: 2.2%, S: 97.2%	0.9647	0.9770	AVG_WBC_FIRST24H, AGE_AT_ADMISSION, AVG_BICARBONATE_FIRST24H, AVG_CREATININE_FIRST24H, HAS_DIABETES
STAPH AUREUS COAG+ / GENTAMICIN	Completed	6742	I: 0.3%, R: 3.0%, S: 96.7%	0.9666	0.9676	AVG_CREATININE_FIRST24H, AGE_AT_ADMISSION, AVG_WBC_FIRST24H, AVG_BICARBONATE_FIRST24H, HAD_PRIOR_ANTIBIOTICS
ESCHERICHIA COLI / GENTAMICIN	Completed	3505	I: 0.9%, R: 12.2%, S: 86.9%	0.8565	0.8573	AVG_WBC_FIRST24H, AGE_AT_ADMISSION, AVG_BICARBONATE_FIRST24H, AVG_CREATININE_FIRST24H, SPEC_TYPE_DESC_BLOOD CULTURE
STAPH AUREUS COAG+ / ERYTHROMYCIN	Completed	6479	I: 2.1%, R: 69.7%, S: 28.2%	0.6915	0.7262	AVG_WBC_FIRST24H, AGE_AT_ADMISSION, AVG_BICARBONATE_FIRST24H, AVG_CREATININE_FIRST24H, HAS_DIABETES
STAPH AUREUS COAG+ / OXACILLIN	Completed	7553	R: 59.6%, S: 40.4%	0.6938	0.7249	AGE_AT_ADMISSION, AVG_WBC_FIRST24H, AVG_CREATININE_FIRST24H, AVG_BICARBONATE_FIRST24H, HAS_DIABETES
STAPH AUREUS COAG+ / LEVOFLOXACIN	Completed	6717	I: 1.4%, R: 61.9%, S: 36.7%	0.7216	0.7177	AVG_WBC_FIRST24H, AGE_AT_ADMISSION, AVG_CREATININE_FIRST24H, AVG_BICARBONATE_FIRST24H, HAD_PRIOR_ANTIBIOTICS
STAPH AUREUS COAG+ / CLINDAMYCIN	Skipped - Min Class Count < 5 (Full Set)	NaN	N/A	N/A	N/A	N/A
STAPH AUREUS COAG+ / TETRACYCLINE	Skipped - Min Class Count < 5 (Full Set)	NaN	N/A	N/A	N/A	N/A

Overall Performance

- The predictive performance is measured by the weighted F1-score on the test set, varied considerably across the different pairs.
- High performance (Test F1 > 0.95) was observed for pairs where one class was extremely dominant, such as Vancomycin, Penicillin, Rifampin, and Gentamicin resistance prediction for Staphylococcus aureus. In these cases, the model primarily learned to predict the majority class ('S' or 'R').
- Moderate performance (Test F1 ≈ 0.86) was achieved for Escherichia coli / Gentamicin, which had a somewhat balanced distribution between 'S' and 'R' but still a small 'I' class.
- Lower performance (Test F1 ≈ 0.72) was seen for Staphylococcus aureus against Erythromycin, Oxacillin, and Levofloxacin. These pairs had more balanced distributions between 'R' and 'S', suggesting greater difficulty in distinguishing resistant and susceptible cases based solely on the included early clinical features.

Feature Importance

- Consistently, patient age (AGE_AT_ADMISSION) and early lab values (AVG_WBC_FIRST24H, AVG_CREATININE_FIRST24H, AVG_BICARBONATE_FIRST24H) were among the top predictors across most pairs.
- Prior antibiotic exposure (HAD_PRIOR_ANTIBIOTICS) and the presence of diabetes (HAS_DIABETES) also appeared in the top 5 for several pairs.
- The specific specimen type (e.g., SPEC_TYPE_DESC_BLOOD CULTURE) was important for some pairs.
- The relative ranking of these features varied depending on the specific organism-antibiotic combination.

Interpretation of Results

- The results indicate that prediction performance is highly dependent on the specific organism-antibiotic pair and the underlying distribution of resistance classes.
- For pairs with highly skewed distributions, high F1-scores are achieved, but performance is largely driven by correctly identifying the majority class. The clinical utility for these pairs might be limited unless the model shows high precision/recall for the rare resistant/susceptible class.
- For pairs with more balanced distributions, the lower F1-scores suggest that the included early clinical features may not be sufficient to reliably distinguish resistance patterns. More informative features might be necessary for these challenging cases.
- Key predictors consistently include patient age and early laboratory results (WBC, creatinine, bicarbonate), aligning with clinical intuition.

Conclusion

- We successfully developed and applied a pipeline to predict multi-class antimicrobial susceptibility for common pathogen-antibiotic pairs using early clinical data from MIMIC-III CSVs on GCS.
- The Random Forest models achieved variable performance, with weighted F1-scores ranging from ≈0.72 to ≈0.997 on the test set, heavily influenced by the class distribution of each pair.
- Key predictors consistently included age and early laboratory results (WBC, Creatinine, Bicarbonate).
- This work establishes a baseline and highlights the challenges of predicting AMR using limited early clinical data, especially for pairs with balanced resistance profiles. The pipeline developed provides a framework for future research and development in this area.

Future Work

- Incorporate additional data sources, such as vital signs and detailed patient history, to improve model accuracy.
- Explore advanced feature engineering techniques, including time-series analysis, to capture temporal patterns in the data.
- Evaluate the performance of other machine learning models, including gradient boosting machines and deep learning approaches, and perform systematic hyperparameter tuning to optimize model performance.
- Validate the models on external datasets from different hospitals or healthcare systems to assess their generalizability.
- Investigate methods to improve prediction for minority classes, such as specialized sampling techniques or cost-sensitive learning, to address class imbalance.
- Explore the potential for integrating genomic data into the model to further enhance predictive power.