Methodology Summary

1. Data Preprocessing and Feature Engineering
The dataset sourced from Pfizer ATLAS includes microbial details, such as isolate ID, species, family, and demographic details like patient information (age, gender, source), including patient classification (in/outpatient) and geographical context (country, state). To analyze the relationship between various features and antimicrobial resistance (AMR), it was necessary to restructure the dataset.

1.1 Reshaping the Dataset
In the original dataset, every row represented an individual bacterial isolate, and the antibiotic susceptibility results were spread across multiple columns. There were 134 total columns in the original dataset with 49 antibiotics. Each antibiotic had two columns, for instance, Amoxicillin I and Amoxicillin. The antibiotic columns ended with the suffix I, indicating that they contained interpretation results (Resistant, Intermediate, Susceptible). The other column had the Minimum Inhibitory Concentration value. The aim was to transform the dataset into a format where one row contains the result of one antibiotic susceptibility test performed on an isolate. This format is better suited to machine learning models that predict some target variable from multiple features. The column containing the results of antibiotic susceptibility was unpivoted into two new columns:
- Antibiotic: Contained the names of the antibiotics tested.
- AMR: Contained the corresponding antimicrobial resistance interpretations for each antibiotic-isolate pair.

The original dataset size of 917,049 rows was transformed into 4,493,540 rows.

1.2 Data Cleaning and Handling Missing Values
Once the data was unpivoted, the next step was to handle the missing values. Two distinct datasets were derived from the transformed atlas data:
- Phenotype-Only Dataset: This dataset included phenotypic data along with clinical and demographic features, intended to train machine learning models based solely on observable characteristics and patient information. In this dataset, rows where phenotype and AMR data were missing were eliminated. The size of the dataset was reduced to 2,760,037. This step ensured that all records used for modelling had complete phenotypic information, enhancing data quality and model reliability. The phenotype had the following information: ESBL (Extended-Spectrum Beta-Lactamase), BL Neg (Beta-Lactam Negative), MSSA (Methicillin-Sensitive Staphylococcus Aureus), MRSA (Methicillin-Resistant Staphylococcus Aureus), and BL Pos (BetaLactam Positive).
- Phenotype + Genotype Dataset: This dataset incorporated both phenotypic data and genetic markers to determine whether including genotypic information enhances the accuracy of the models. All features from the phenotype-only dataset were used, with the addition of genetic

markers. There were 23 genetic markers where NEG indicated absence while POS indicated presence. Unlike the phenotype data, there were numerous rows where only some genetic markers were present. Instead of eliminating these rows, which would have drastically reduced the dataset size, missing values in the genetic marker columns were replaced with "NA" placeholder values. This approach preserved the dataset's comprehensiveness while acknowledging the absence of certain genetic markers without introducing potentially inaccurate imputed values to prevent bias. The size of this dataset was 589,998.