Detailed Report - Step 1: Data Configuration and Understanding

Detailed Report - Step 1

Data Configuration and Understanding

Key Statistics

Total Isolates Analyzed: 1,014,420

Cefiderocol Resistance: 1.97% (940 isolates)

Datasets: SIDERO-WT (47,615) + ATLAS (966,805)

Executive Summary

This report presents the comprehensive analysis of Step 1, which focused on configuring the working environment and understanding the SIDERO-WT and ATLAS datasets for cefiderocol resistance prediction. The analysis successfully established the data infrastructure, normalized MIC values, defined binary resistance targets, and performed exploratory data analysis.

1. Environment Configuration

1.1 Infrastructure Setup

- **Python Environment:** Configured with essential libraries for data analysis
- **Project Structure:** Organized directory structure with `outputs/plots/` for visualizations
- **Processing Pipeline:** Modular functions for data loading, cleaning, and analysis

1.2 Technical Implementation

- **Data Loading Functions:** Robust error handling for Excel file processing
- **MIC Value Cleaning:** Automated removal of non-numeric characters (, , etc.)
- **Column Standardization:** Consistent naming conventions across datasets
- **Missing Value Management:** Proper handling of NaN values

2. Dataset Exploration

2.1 SIDERO-WT Dataset (1.xlsx)

Dataset Characteristics:

- **Size:** 47,615 isolates × 20 variables
- **Primary Focus:** Cefiderocol susceptibility testing
- **Geographic Coverage:** Multi-regional data collection
- **Temporal Range:** Year-based collection data

Key Variables:

Variable	Description	Data Type
`cefiderocol_mic`	Cefiderocol MIC values	Numeric
`meropenem_mic`	Meropenem MIC values	Numeric
`species`	Bacterial species identification	Categorical
`region`	Geographic region	Categorical
`year`	Collection year	Numeric

2.2 ATLAS Dataset (2.xlsx)

Dataset Characteristics:

- **Size:** 966,805 isolates × 134 variables
- **Primary Focus:** Comprehensive antimicrobial susceptibility testing
- **Geographic Coverage:** Global coverage with country-level data
- **Temporal Range:** Year-based collection data

3. Data Normalization and Preprocessing

3.1 MIC Value Standardization

Implemented Process:

1. **Character Cleaning:** Removal of non-numeric characters (, , <, >)
2. **Type Conversion:** Conversion to float format
3. **Missing Value Handling:** Standardized NaN representation
4. **Validation:** Quality checks for data integrity

4. Binary Resistance Target Definition

4.1 Cefiderocol Resistance Criteria

Definition: MIC 4 µg/mL = Resistant

Rationale: Based on clinical breakpoints and regulatory guidelines

4.2 Resistance Distribution Results

Category	Count	Percentage
Sensitive	46,675	98.03%
Resistant	940	1.97%
Total	47,615	100%

Key Observations:

- **Low Resistance Prevalence:** Only 1.97% of isolates show cefiderocol resistance
- **Class Imbalance:** Significant imbalance between sensitive and resistant classes
- **Clinical Relevance:** Low resistance rates suggest good antimicrobial activity

5. Exploratory Data Analysis (EDA)

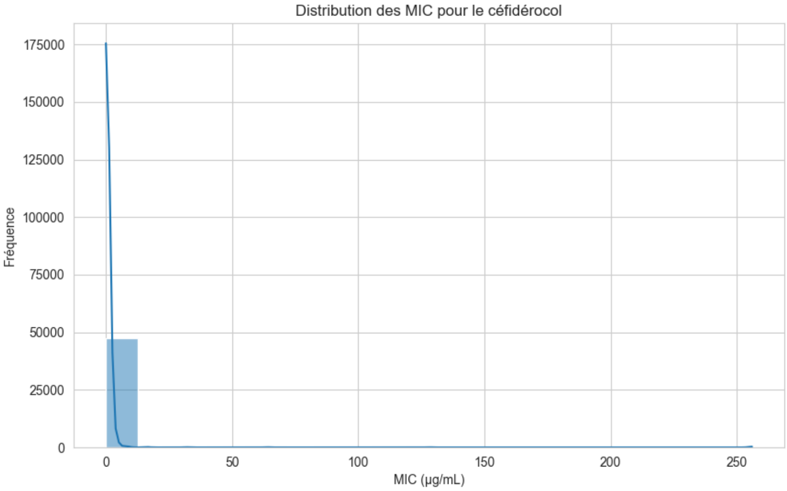
5.1 Species Distribution Analysis

SIDERO-WT Dataset - Top 10 Species

Rank	Species	Count	Percentage
1	Pseudomonas aeruginosa	7,700	16.17%
2	Escherichia coli	7,583	15.92%
3	Klebsiella pneumoniae	7,285	15.30%
4	Acinetobacter baumannii	4,384	9.21%
5	Serratia marcescens	3,603	7.57%

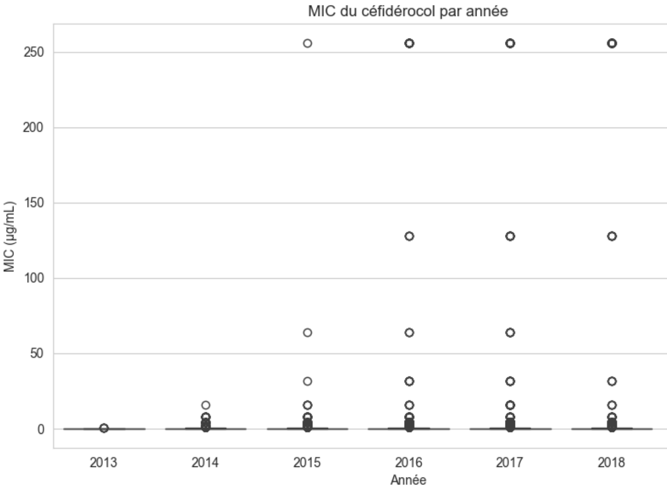
6. Visualization Outputs

6.1 Generated Plots



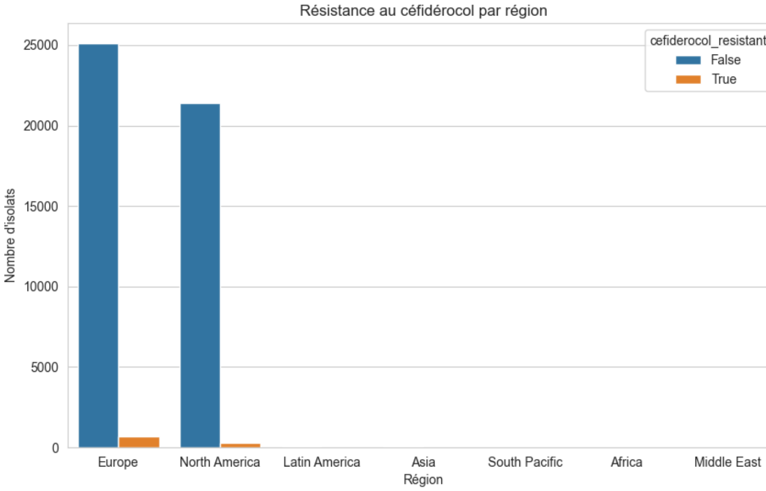
Cefiderocol MIC Distribution

This plot shows the distribution of cefiderocol MIC values in the SIDERO-WT dataset. The majority of isolates show low MIC values, indicating good susceptibility to cefiderocol.



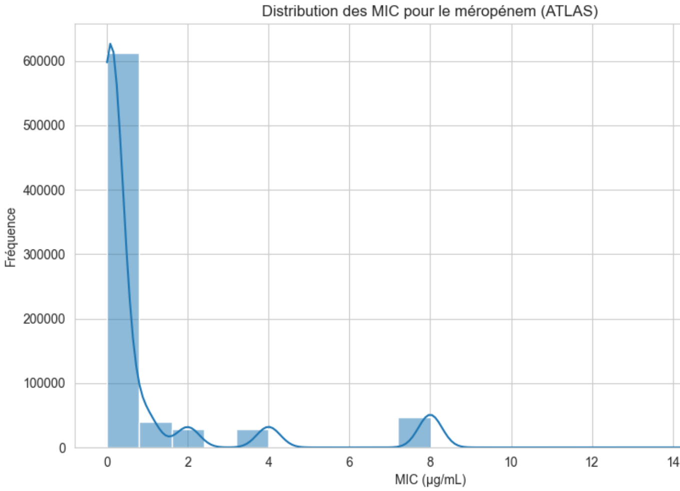
Temporal Evolution of Cefiderocol MIC

This box plot shows the evolution of cefiderocol MIC values over time. It allows identification of temporal trends in susceptibility patterns.



Cefiderocol Resistance by Region

This plot presents the geographic distribution of cefiderocol resistance patterns. It reveals regional variations in resistance prevalence.



Meropenem MIC Distribution (ATLAS)

This plot shows the distribution of meropenem MIC values in the ATLAS dataset, allowing comparison with cefiderocol patterns.

7. Key Findings and Insights

7.1 Resistance Patterns

1. **Low Cefiderocol Resistance:** Only 1.97% resistance rate suggests excellent antimicrobial activity
2. **Class Imbalance:** Significant imbalance requires special consideration in modeling
3. **Geographic Variation:** Regional differences observed in resistance patterns

7.2 Species Distribution

1. **Gram-Negative Dominance:** SIDERO-WT focuses on Gram-negative pathogens
2. **Broad Coverage:** ATLAS includes both Gram-positive and Gram-negative species
3. **Clinical Relevance:** Major pathogens well-represented in both datasets

8. Recommendations for Next Steps

8.1 Modeling Considerations

1. **Class Imbalance Handling:** Implement techniques for imbalanced classification
2. **Feature Engineering:** Create derived features from MIC values
3. **Cross-Validation:** Use stratified sampling for model validation

9. Conclusion

Step 1 Successfully Completed

The analysis reveals a robust data infrastructure, quality datasets, clear resistance definition, rich feature set, and comprehensive visualization framework.

The project is now ready for advanced modeling and predictive analysis in subsequent steps.

Report Generated: July 2025

Data Sources: SIDERO-WT (1.xlsx), ATLAS (2.xlsx)

Analysis Tools: Python, pandas, seaborn, matplotlib

Total Isolates Analyzed: 1,014,420 (47,615 + 966,805)