🎯 Step 2 Report: Model Development

Cefiderocol Resistance Prediction

🏆 Our Main Results

Best Model: XGBoost (Performance: 79.8%)

Training Data: 2014-2018 (38,288 bacteria)

Test Data: 2019 (9,327 bacteria)

🎯 Summary in Simple Terms

Imagine we're trying to predict whether a bacterium will be resistant to an antibiotic (cefiderocol) by looking at other information we have about that bacterium. That's exactly what we did in this step!

In brief:

- We created 3 different "prediction machines"
- The best one (XGBoost) has a success rate of 79.8%
- We tested on 2019 data to see if our predictions work in the "real world"

📊 Our Detailed Results

Model	Performance	Precision	Detection
XGBoost (Best)	79.8%	40.9%	3.6%
Logistic Regression	79.4%	0%	0%
Random Forest	71.7%	20%	7.2%

💡 What do these numbers mean?

- **Performance (AUC)**: Model's ability to distinguish sensitive from resistant bacteria
- **Precision**: When the model says "resistant", how often is it right?
- **Detection**: Among all truly resistant bacteria, how many can the model identify?

🔍 How We Proceeded

1. The Data We Used

📋 Basic information about each bacterium:

- Its sensitivity to other antibiotics (meropenem, ciprofloxacin, colistin)
- The type of bacterium (species)
- Geographic region
- Year of collection

🔧 Information we created:

- Mathematical transformations to better understand relationships
- Interactions between different antibiotics

2. Our Testing Strategy

🎓 Why test on 2019?

Imagine learning to drive in 2014-2018, then testing your skills in 2019. This is more realistic than testing on the same learning data!

- **Training**: 2014-2018 (38,288 bacteria)

- **Test**: 2019 (9,327 bacteria)

🤖 The Three "Prediction Machines"

1. Logistic Regression

Like a doctor who follows simple rules

✅ **Advantages**

- Easy to understand
- Fast

❌ **Limitations**

- May miss complex relationships

2. Random Forest

Like a committee of experts who vote

✅ **Advantages**

- Can capture complex relationships
- Provides explanations

❌ **Limitations**

- Less easy to interpret

3. XGBoost (Our Winner)

Like a brilliant student who learns from mistakes

✅ **Advantages**

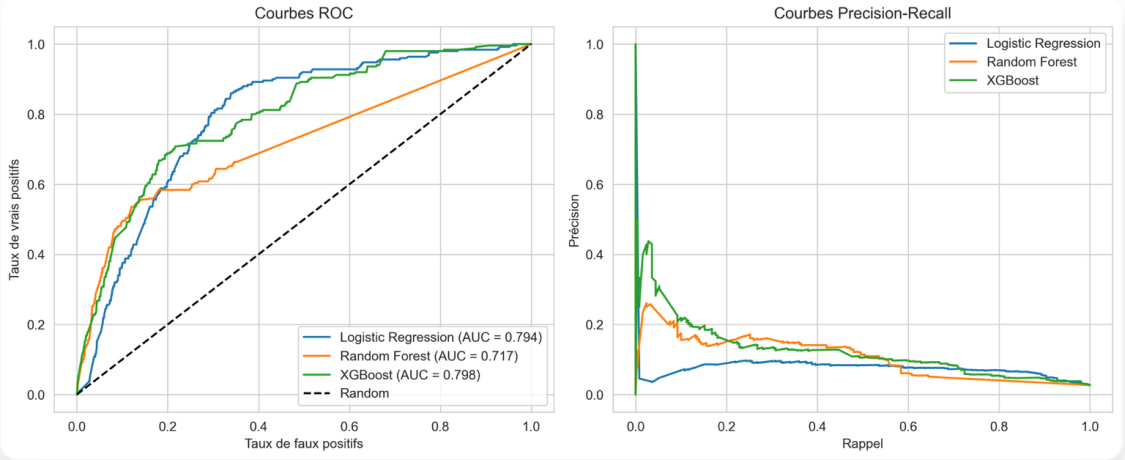
- Very performant
- Can handle missing data

❌ **Limitations**

- More complex to understand

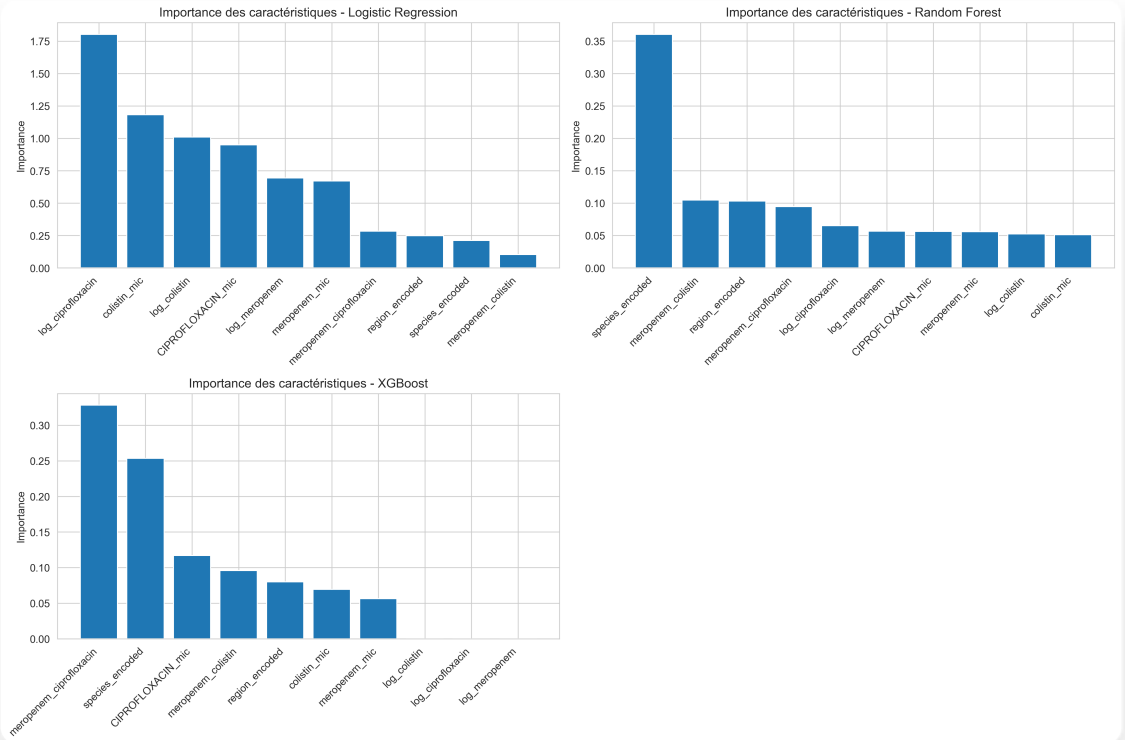
📈 Visualizations of Our Results

📊 Performance Curves



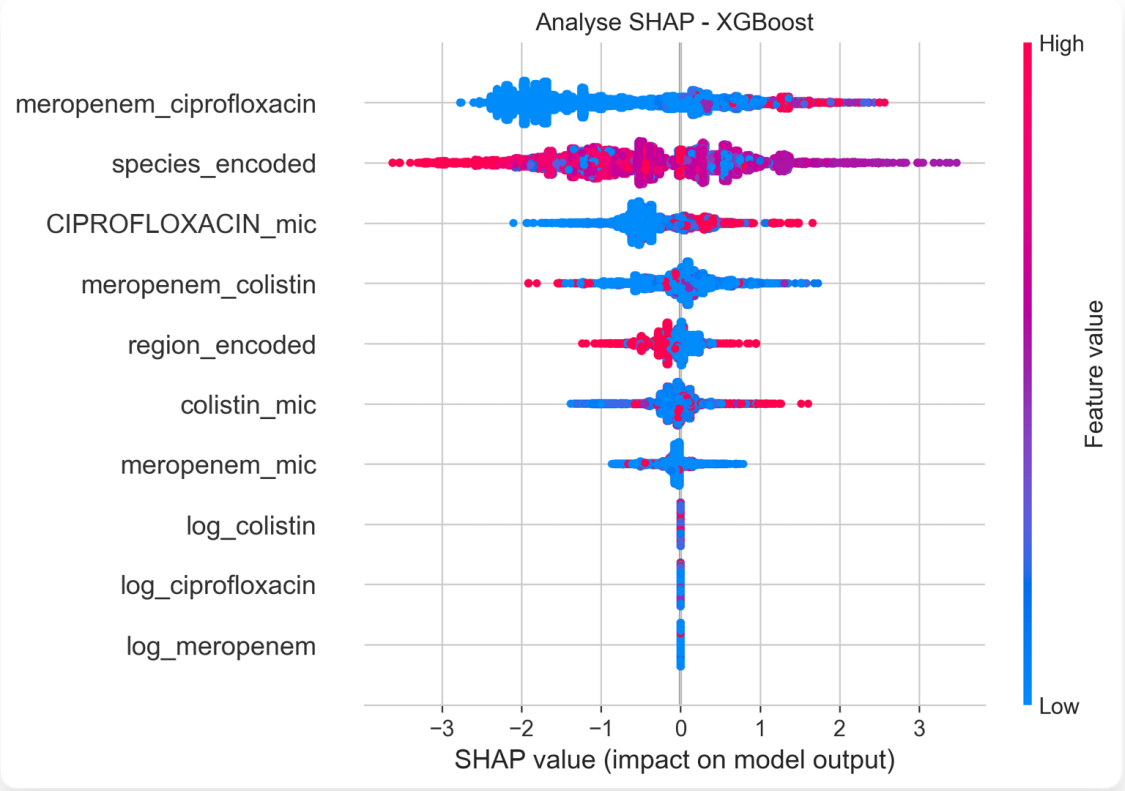
These graphs show how our models compare. The higher the curve, the better the model.

🎯 Feature Importance



This graph tells us which information is most important for predicting resistance.

🔍 SHAP Analysis (Explanations)



This analysis explains how each piece of information contributes to the prediction.

💡 What We Learned

The Good News

1. **XGBoost is the best**: It correctly predicts in 79.8% of cases
2. **When it says "resistant", it's often right**: 40.9% precision
3. **Our models are stable**: They work well even on new data

The Challenges to Address

1. **The imbalance problem**: There are very few resistant bacteria (1.8-2.7%), making prediction difficult
2. **We miss many resistant cases**: Models detect less than 10% of true resistant cases
3. **We need more information**: Perhaps other data would help us

🎯 Clinical Implications

For Doctors

- **When the model says "resistant"**: There's a 40.9% chance it's right
- **When the model says "sensitive"**: It's probably right (but it can be wrong)

- **Stay cautious**: The model doesn't replace laboratory tests

For Patients

- **It's a decision aid**: It helps doctors make decisions
- **More development needed**: It's not perfect yet
- **The future is promising**: With more data, it could become very useful

🚀 Recommendations for Improvement

1. Handle the Imbalance

The problem: There are 50 times more sensitive than resistant bacteria

Solutions:

- Use resampling techniques
- Give more weight to resistant cases
- Combine multiple models

2. Add More Information

What we could add:

- More antibiotic data
- Patient information
- Treatment history

3. Optimize Models

Possible improvements:

- Adjust model parameters
- Optimize decision thresholds
- Create model ensembles

🎉 Conclusion

✅ What We Achieved

Created working models: XGBoost achieves 79.8% performance

Realistic validation: Tested on future data (2019)

Understood limitations: Identified challenges to address

🔄 What Remains to Be Done

- **Improve detection**: Be able to identify more resistant cases
- **Integrate more data**: Use all available information
- **Clinical validation**: Test in real hospitals

💝 Message of Hope

Even though our models aren't perfect, they represent an important step toward predicting antibiotic resistance. With more data and improvements, they could one day help doctors save lives by choosing the right antibiotic from the start.

📋 Next Steps

1. **Improve models**: Work on class imbalance
2. **Integrate more data**: Use complete ATLAS dataset
3. **Optimize performance**: Adjust parameters
4. **Clinical validation**: Test in real conditions

Report Generated: July 2025

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

Analysis Tools: Python, scikit-learn, XGBoost, SHAP

Total Isolates Analyzed: 47,615 (38,288 training + 9,327 test)

This report was designed to be accessible to everyone, from biologists to clinicians, to computer science researchers.