

**Bioinformatics Computing II**

(Project Proposal)

**Submitted to:**

Dr. Adnan Ahmad Ansari

**Submitted by:**

Muhammad Ekramah

**Department:**

National Centre for Bioinformatics (NCB)

**Registration no.**

04282213040

**Date of Submission:**

### 06 October 2025

### 

**Machine Learning Prediction of Antibiotic Resistance in Neisseria gonorrhoeae**

**Background:**

Antimicrobial resistance (**AMR**) is a rapidly growing global health threat, compromising the effectiveness of current antibiotic therapies and increasing the risk of treatment failure.

*Neisseria gonorrhoeae* causes gonorrhea, the second most common sexually transmitted infection in Europe. The bacterium has developed resistance to multiple antibiotics including ciprofloxacin, azithromycin, and cefixime, making treatment increasingly difficult. Machine learning combined with genomic data offers the potential for rapid resistance prediction, enabling better treatment decisions and antimicrobial stewardship.

**Objectives**

Develop machine learning models to predict antibiotic resistance in N. gonorrhoeae using genomic unitig data for three antibiotics: ciprofloxacin, azithromycin, and cefixime.

1. predicting antibiotic resistance and Compare performance of multiple ML algorithms (Logistic Regression, Random Forest, Support Vector Machine (SVM), and Gradient Boosting algorithms (XGBoost/CatBoost)for predicting antibiotic resistance.)
2. Identify genomic features most predictive of resistance
3. Validate results using cross-validation techniques

**Methodology**

**Dataset:**

* **Source:** Kaggle dataset "antibiotic resistance in gonorrhoea”
* **Samples:** 3,478 N. gonorrhoeae isolates
* **Features:** 515 unitigs (DNA segments) associated with resistance
* **Labels:** Binary resistance/susceptible classifications for antibiotics on based of MIC values

**4. Approach**

### Step 1: Data Acquisition and Understanding

* The dataset **“Predicting Antibiotic Resistance in *Neisseria gonorrhoeae*”** was downloaded from **Kaggle (gono-unitigs)**.
* It includes **genomic unitigs** (binary genomic features), **metadata** (country, year, continent), and **phenotypic resistance labels** for three antibiotics based on MIC values: **Azithromycin (AZM), Cefixime (CFX), and Ciprofloxacin (CIP)**.
* Each record (isolate) represents a bacterial genome with its corresponding resistance phenotype.

**Step 2: Data Preprocessing**

* **Data cleaning:** Remove isolates with missing resistance labels or incomplete metadata; filter unitigs with extremely low or high frequency (rare or constant features).
* **Feature encoding:** Convert categorical metadata (e.g., country, continent) into numerical form using one-hot encoding. Represent each unitig as a binary variable (1 = present, 0 = absent).
* **Data splitting:** Divide data into training and testing sets, applying **Stratified K-Fold Cross-Validation (k = 5)** to maintain class balance during model training and validation.

**Step 3: Model Implementation**

* Implement multiple **supervised machine learning classifiers** to predict resistance (Resistant = 1, Sensitive = 0):
  + **Logistic Regression:** Linear baseline model for interpretable probability-based predictions.
  + **Random Forest:** Ensemble of decision trees using bagging, providing robustness and feature importance ranking.
  + **Support Vector Machine (SVM):** Margin-based classifier using the RBF kernel for non-linear data separation.
  + **Gradient Boosting Models (XGBoost / CatBoost):** Sequential boosting algorithms that iteratively correct previous errors; efficient and suitable for imbalanced, high-dimensional genomic data.
* All models will be implementedd in **Python** using **scikit-learn** and **XGBoost/CatBoost** libraries.

**Step 4: Model Evaluation and Comparison**

* Use **5-Fold Stratified Cross-Validation** to ensure consistent and reliable performance estimates.
* Evaluate models using multiple metrics:
  + **Accuracy:** Overall correctness of predictions.
  + **Precision:** Proportion of predicted resistant isolates that are truly resistant.
  + **Recall (Sensitivity):** Proportion of actual resistant isolates correctly identified.
  + **F1-score:** Harmonic mean of precision and recall, suitable for imbalanced datasets.
* Average results across folds for stable performance and compare models to identify the best-performing algorithm for each antibiotic.

### Step 5: Feature Importance and Biological Interpretation

* Extracte **top-ranked unitigs** contributing most to resistance prediction using feature importance from tree-based models.
* Map significant unitigs to known **antimicrobial resistance (AMR) genes** using **BLAST** for biological validtion and interpretation.

**Tools and Technologies**

* **Programming:** Python with scikit-learn, pandas, numpy,matplotlib, seaborn
* **ML Algorithms:** Logestic reggeression ,Random Forest, Support Vector Machine, gradient boost models (catboost orXGBoost)
* **Validation:** Cross-validation, confusion matrices, ROC curves

## Expected Outcomes

* Train models with >70% accuracy for resistance prediction.
* Performance comparison across different algorithms accuracy.
* Ranking of most predictive genomic features.
* Complete Python implementation with documented code
* Performance analysis report comparing all models
* Feature importance analysis identifying key resistance markers.
* Final project report with methodology and results

**Timeline (27 Sept – 29 Nov)**

| Week | Dates | Milestones |
| --- | --- | --- |
| Week 1-2 | 27 Sept – 11 Oct | Environment setup, Data acquisition, exploration, and preprocessing(cleaning) |
|  |  |  |
| Week 3-4 | 11 – 24 Oct | Implement baseline models ( Logistic Regression , Random Forest) |
|  |  |  |
| Week 5 | 25 – 31 Oct | Develop advanced models (SVM,gradiant boosting algo (XGBoost or CatBoost) with hyperparameter tuning |
| Week 6 | 1 – 7 Nov | Feature importance analysis and model comparison |
| Week 7 | 8 – 14 Nov | Validation, testing, and results compilation |
|  |  |  |
| Week 8-9 | 15– 29 Nov | Final report writing and presentation preparation |

### 

### Conclusion

This project will apply machine learning techniques to predict antibiotic resistance in *Neisseria gonorrhoeae* using genomic unitigs and associated metadata. Through systematic preprocessing, model training, and evaluation, several classifiers Logistic Regression, Random Forest, Support Vector Machine, and Gradient Boosting algorithms (XGBoost/CatBoost) will be implemented and compared using 5-fold cross-validation. The models are expected to achieve reliable performance across key evaluation metrics, demonstrating that genomic sequence data can be effectively utilized to infer resistance patterns. Feature importance analysis will further identify biologically meaningful unitigs linked to known antimicrobial resistance genes, providing both computational and biological insights. Overall, this study will establish a reproducible, data-driven framework that integrates bioinformatics and machine learning for rapid prediction of antibiotic resistance, contributing to the broader effort of combating antimicrobial resistance through genomic surveillance and informed therapeutic decision-making.