Antibiotic Resistance Mechanisms in Staphylococcus aureus based on MALDI-TOF
IA en Salud, MIAA

1. Dataset

For this laboratory exercise, the MALDI-TOF Mass Spectrograms (MS) dataset called DRIAMS [1] presented in Nature Medicine [2] will be utilized. This dataset consists of 800,000 MALDI-TOF MS obtained from four different institutions in Switzerland. However, for this laboratory exercise, a simplified version of the dataset will be used, which is available on Kaggle [3]. Only DRIAMS-D, obtained from the Viollier AG laboratory in Basel, Switzerland, will be used.

The simplified version of the dataset contains 1731 MALDI-TOF MS of *Staphylococcus aureus*, along with their resistance to two different antibiotics: **Erythromycin** and **Ciprofloxacin**.

The dataset is available on AulaGlobal, under the name ***practica-maldi.zip***.
In this scenario, the provided mass spectrograms are already preprocessed, and each MALDI-TOF MS is represented by a list of 6000 intensities. The mass spectra are binned into fixed bins of 3Da, ranging from 2,000Da to 20,000Da, resulting in a 6,000-dimensional vector representation for each sample. Note that the values of the MALDI-TOF MS are relatively small. Therefore, a preprocessing step may be necessary to train gradient-based methods effectively.

2. Objective

The objective is to predict the resistance to both Erythromycin and Ciprofloxacin based on the information contained in the MALDI-TOF MS spectra. It is necessary to explore various proposals for the analysis, while ensuring that the resulting models are interpretable. If dimensionality reduction techniques are employed, they must be accompanied by a clear interpretation of their outputs. In addition, it is also recommended to consider feature selection methods that do not involve dimensionality reduction, such as Random Forest or Logistic Regression. It is essential to evaluate the performance of different methods for dimensionality reduction and classification and provide a justification for the selection of the final analysis pipeline.

3. Evaluation

Each group must submit a **2-pages report** containing four blocks:

1. **Exploratory data analysis**: how did you balance the classes, missing data, categorization.
2. **Preprocessing**: which preprocess did you propose and why
3. **Models**: at least you have to try out 3 different models. Which works better? Why do you think it does?
4. **Explainability/Interpretability**: which peaks/features are most important to predict each resistance?

[1] *https://datadryad.org/stash/dataset/doi:10.5061/dryad.bzkh1899q*
[2] Weis, C. et al. Direct antimicrobial resistance prediction from clinical MALDI-TOF mass spectra using machine learning. Nat Med (2022). https://doi.org/10.1038/s41591-021-01619-9
[3] https://www.kaggle.com/datasets/drscarlat/driams