**Instructions for Course Project**

**Objective:**

The objective of this project is to investigate associations between beta-lactam resistance genes and beta-lactam resistance using machine learning techniques.

**Hypothesis:**

Genetic profiles consisting of multiple beta-lactamase genes are associated with MICs and resistance categories for beta-lactam drugs.

**Files**

1. The data set is “Data Set for Oliveira Machine Learning Project.csv”
2. A data dictionary explaining the variables is the file “Data Dictionary.xlsx”
3. The Excel file “Required Information for Analysis.xlsx” contains additional information necessary for you to complete analysis

**Data Cleaning Tasks**

1. Classify each isolate as containing an extended spectrum beta-lactamase (ESBL) enzyme, carbapenemase enzyme, both, or neither
   1. Genes included in these categories are provided in the “Resistance Gene Categories” tab of “Required Information for Analysis.xlsx”
2. You will need to categorize each isolate as Susceptible (S), Intermediate (I) or Resistant (R) to each drug (even if you are not assigned to that analysis group – it is good practice)
   1. Breakpoint values are provided in the ‘Breakpoints for Resistance’ tab of “Required Information for Analysis.xlsx”
   2. ‘Susceptible’ is always coded as *less than or equal to* the ‘Lower Breakpoint’
   3. ‘Resistant’ is always coded as *greater than* *or equal to* the ‘Upper Breakpoint’
   4. ‘Intermediate’ can be coded as an else statement
   5. Example logic:
      1. If Amikacin ≤16 then Amikacin.Category = “S”
      2. If Amikacin ≥64 then Amikacin.Category = “R”
      3. Else Amikacin.Category = “I”
   6. Note that not all drugs will have an Intermediate category
3. MIC values are tested on a fold-dilution scale. You will need to ‘log-transform’ MIC values by taking the log2 of the MIC value.
4. Some bacteria have MICs that are not represented on the panels we test. They are represented as either one digit below or one digit greater than the lowest or highest value on the panel, respectively. You will need to plot the MICs for each drug to determine what these values are for each drug.
   1. For values below the lowest dilution, treat them as equal to the lowest dilution and convert them to this value
   2. For values above the highest dilution (ending in .1) convert to the next highest dilution
   3. Refer to the ‘Standard MIC Dilution Values’ tab of “Required Information for Analysis.xlsx” for this step.
5. You will notice that many isolates are missing values for at least one variable.   
   **Do not impute missing values.**

**Machine Learning Analysis**

For your assigned outcome (MICs or Resistance Categories), consider which machine learning methods would be most appropriate for the data structure and for the research question. Think before you act. It is important to establish an analysis plan *a priori* to give the project direction and to avoid ‘fishing’ for associations.

The hypothesis of this study is that beta-lactamase genes cause resistance to beta-lactam antibiotics. You are provided with several drugs that are not beta-lactams. These are denoted in the ‘Drug Names’ tab of “Required Information for Analysis.xlsx”. These non-beta-lactam drugs provide a useful reference to demonstrate that the genetic profiles you identify are specific to beta-lactams.

In addition to antibiotic MICs and resistance genes, the data set contains many demographic variables. You are encouraged to include these variables in your analysis, as trends or genetic profiles may vary by demographic.