

 FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO   FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO	First Degree in Artificial Intelligence and Data Science  <b>Elements of Artificial Intelligence and Data Science</b>	2024/2025  1 <sup>st</sup> Year  2 <sup>nd</sup> Semester
TEACHERS: Luís Paulo Reis, Miriam Santos, Rita Ribeiro, Moisés Santos		

## ***Exercise Sheet 5: Data Cleaning and Preprocessing***

### **5. Breast Tissue Impedance Measurements**

Consider the dataset **breast.csv** available on Moodle (and [Kaggle](#)), which contains data from 106 samples related to breast tissue measurements. Each sample is described by 9 features (IO, PA500, HFS, DA, AREA, A.DA, MAX.IP, DR, P) and belongs to one of 6 identified classes (CAR, FAD, MAS, GLA, CON, ADI).

#### **5.1. Feature Assessment and Visualization**

a. Choose **two classes** and plot the **distribution histograms** for each of the 9 features. Draw your conclusions regarding the discriminative power of the features for separating the selected classes.

**Hint:** Use `seaborn.histplot()` and find out how to overlay histograms.

b. Calculate the correlation matrix for the features in this dataset. Identify which features are the most correlated and which are the least correlated and present a scatter plot for each case.

**Hint:** Use `pandas.DataFrame.corr()` and `seaborn.heatmap()` to compute and visualize the correlation matrix and `seaborn.scatterplot()` to produce the scatter plots for your exploration.

c. For the 3 most discriminative features, produce their boxplots and discuss their contribution to distinguish between the chosen classes. **Hint:** Use `seaborn.boxplot()`.

#### **5.2. Data Cleaning, Transformation, and Reduction**

d. Write a function `rmoutliers()` that returns the outliers that exist in a feature. A value should be considered an outlier if it is at least 3 standard deviations away from the mean. Use this function to detect the outliers of the features **PA500** and **HFS**. How many values were detected as outliers? Compare your results with their boxplots and comment on the results.

e. Write a function `random_oversampling()` that performs oversampling with replacement. The function should take as arguments the matrix of patterns from the class of interest and the number of samples to replicate, and it should return a matrix with the replicated patterns. Use the function to perform oversampling on the class CON so that it has the same number of patterns as the class ADI. **Tip:** Beyond a custom function, you could use [imbalanced-learn](#) to explore other methods, such as [RandomOverSampler](#), among others.

f. Write a function `norm_min_max()` that normalizes a feature according to the min-max transformation. Also write a function `norm_zscore()` that normalizes a feature according to the z-score transformation. Use these functions to normalize the features of the dataset and observe their effects. Compare your implementation with those provided in scikit-learn: [MinMaxScaler](#) and [StandardScaler](#).

g. Using [Principal Component Analysis](#), try to answer the following questions:

1. What is the variance associated with the first principal component?
2. How many components should you keep to keep at least 98% of the total variance?
3. How many components should you keep according to Kaiser's criterion?
4. How many components should you keep according to the Scree Test? (*you should create the Scree Plot to answer this question*).