LAB 1
Breast Cancer Survival Prediction using Multi-Omics Data

IA en Salud, MIAA

1 Dataset

In this project, we will use heterogeneous multi-omics data sourced from The Cancer Genome Atlas (TCGA) breast invasive carcinoma (BRCA) project[1]. The dataset contains 705 BRCA samples and incorporates four different omics data types, amounting to a total of 1936 features.

The dataset contains 806 features related to copy number variation (cn), encoded as follows: -2 for homozygous deletion, -1 for hemizygous deletion, 0 for neutral/no change, 1 for gain, and 2 for high-level amplification. It also includes 249 features indicating mutations (mu), with encoding for somatic (1) or germline (0). Furthermore, there are 604 features related to gene expression levels (rs) and 223 features related to phospho-protein levels (pp). Additionally, the dataset includes a feature that specifies the type of cancer, specifically distinguishing between ductal and lobular. The dataset is available on AulaGlobal, under the name `data-brca.zip`.

2 Objective

The objective is to predict the survival (or vital status) of the patients based on their multi-omics data.

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.

[1] https://portal.gdc.cancer.gov/projects/TCGA-BRCA

3 Submission and evaluation

This assignment comprises 25% of the course's overall grade. We will assess both the design process and the quality of the final solution, including its performance.

The only deliverable for this assignment is a Weights and Biases report.