Hackaton 2023
Detecting Attributes of Breast Cancer

Ruben Hoba 342797040 – Michael Wolhandler 342768587
Moshe Marciano 777902495 – Isaac Labé 66622765

9 June 2023

I Part 0: Pre-processing

The dataset used in this study consists of 49,351 records, where each record corresponds to a patient visit. The dataset includes 34 features that provide information about the patient and their sickness. The objective of this study is to preprocess the data, which involves transforming them into numerically computable quantities and addressing the issue of missing values. Additionally, the study aims to select only the relevant features that optimize the model's performance.

To begin the pre-processing stage, the features are categorized into five groups:

- **Useless data:** This category includes information such as the patient's name, ID, hospital location, and oncologist choice. These features are deemed irrelevant for the prediction task and are therefore removed from the dataset.
- **Datetime features:** This group comprises features such as the diagnosis date and the date of surgery relative to the diagnosis. These datetime features are engineered to represent the difference in days between the two dates. Subsequently, the values are normalized to account for variations in the timeframe between diagnosis and surgery.
- **Semi-numerical features:** These features involve tumor marker tests that determine sensitivity to estrogen, sensitivity to progesterone, presence of the HER2 gene, and the rate of cell multiplication in the tumor (KI67 protein). These data can be converted into percentages, which are then normalized to a range between 0 and 1.
- **Numerical features:** This category encompasses features that solely contain numerical values, such as age and surgery sum. Notably, the age feature is one-hot encoded into ten classes, representing age intervals of 10 years each (e.g., 0-10, 10-20, ..., 90-100).
- **Categorical features:** This group includes features with string categories, such as the metastases mark (TNM) and side of the body. These features require specific encoding techniques to handle them effectively.

Common functions are applied to each category of data to retain the relevant features:

- The first step involves keeping features where the number of non-null values exceeds a threshold $\theta = 30\%$ of all samples. Additionally, only categorical features with less than 10 unique values are retained.
- The second step addresses missing values by replacing them with randomly generated values from the probability distribution of the feature values across the samples.

Finally, all categorical features are one-hot encoded, and the numerical features are normalized to ensure consistency and comparability within the dataset.

II Part 1: Predicting Metastases

The objective of this section is to train a model for predicting the sites of metastases. The metastasis data are categorical and a patient can have multiple sites where metastases are detected. The label data are presented as a list indicating the "value" of the site region, such as

$$['LYM - Lymphnodes','HEP - Hepatic']$$

To utilize these labels numerically, we will binarize the list, representing all possible values of the categories. For example, the previous label will be encoded as

$$[0, 1, 0, 0, 0, 0, 1, 0, 0, 0],$$

where the index of the "1" corresponds to the true label.
To classify the site labels using the preprocessed features, we employ a Multi-label classifier. In this study, we use a decision tree classifier as it has been observed to be more powerful than Linear Regression or Logistic Regression for this particular task. The classification results show an accuracy rate of 0.78 for macro average and 0.84 for micro average.

III Part 2: Predicting Tumor Size

The second task involves predicting the size of the tumor using numerical features. The tumor size data is provided in centimeters (cm) and does not require preprocessing. In order to enhance the predictive performance, we selected a Gradient Boosting Regressor. This choice was based on the assumption of a more linear relationship between the features and tumor size. The mean squared value obtained using this regressor is 4.85.