# Predicting Benign and Malign Breast Cancer

##### Liam Phan, Michael Bigler & Manuela Giansante

##### 22th November 2022

Overview

**The goal of our project is to predict the benignity or malignity of breast tumour based on the collected data on patients.**

**The analysis will include:**

* **EDA**
* **Dimension Reduction, Transformation and Standardization if needed**
* **Unsupervised learning**
  + **Cluster analysis**
* **Supervised learning**
  + **Logistic Regression**
  + **Neural Nets**
  + **Discriminant analysis**
  + **K-Nearest neighbours**
  + **Classification trees**
  + **Ensemble methods of the above used methods**

###### Objectives

* **We want to predict whether the tumour is benign or malign.**
* **We want to cluster the data and see if there is a link to the classification of benignity and malignity.**
* **We want to find the most impactful metrics within the scope of our prediction.**

Description of Dataset

**The** [dataset](https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset) **contains 32 columns and 570 rows, describing 10 features (mean, SE, worst value (largest)).**

|  |  |  |
| --- | --- | --- |
| ****ID number**** | **Identification for the data points (per tumour, there is no replication)** | |
| ****Diagnosis**** | **M = malignant, B = benign** | |
| ****Radius**** | **Mean of distances from center to points on the perimeter** | |
| ****Texture**** | **Standard deviation of gray-scale values** | |
| ****Perimeter**** | **Outer perimeter of Lobes** | |
| ****Area**** | **Area of Lobes** | |
| ****Smoothness**** | **Local variation in radius lengths** | |
| ****Compactness**** | **Perimeter^2 / area - 1.0** | |
| ****Concavity**** | **Severity of concave portions of the contour** | |
| ****Concave Points**** | **Number of concave portions of the contour** | |
| ****Symmetry**** | **Symmetry of the breasts** | |
| ****Fractal Dimension**** | | **"Coastline approximation" - 1** |

# Structure of the final report

1. Introduction
2. Development
   1. Data analysis
      1. Structure of data
      2. Missing values
      3. Distribution of data
      4. Correlations
   2. Data preparation
      1. Transformations
      2. Standardization
      3. Partitioning
   3. Unsupervised learning
      1. Cluster analysis
   4. Supervised learning
      1. Logistic Regression
      2. Neural Nets
      3. Discriminant analysis
      4. K-Nearest neighbours
      5. Classification trees
      6. Ensemble methods of the above used methods
   5. Comparison of models and definition of best model
3. Conclusion