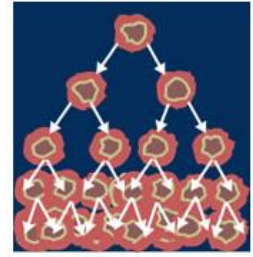


# Predicting breast cancer



## Overview

Breast Cancer (BC) is a common disease. However, the amount of available data has made it possible for business analyst and data scientist to design and deploy useful predictive models that are able to foresee the likelihood of breast cancer based on a variety of attributes related to patients and ultrasonic imaging.

## Your task

In this assignment, your task is to design a predictive model based on clustering and/or classification to predict the likelihood of existence of breast cancer in a patient. In particular, we ask you to apply the tools and techniques that can help you to predict patients with high likelihood of breast cancer. The final deliverable of your assignment task should be a report containing the following sections:

- Defining Business Objectives

The project report should start with the description of well-defined business objective. The model is supposed to address a business question. Clearly stating that objective will allow you to define the scope of your project and will provide you with the exact test to measure its success.

- Exploring data

Once you have addressed missing values and duplicate data problem you will need to explore inherent relationships between the different variables. The focus variable for this study is the Result column (since you are asked to predict it). So, this section should show your efforts to identify from the remaining columns in the dataset which are likely to have high predictive power on the 'Result' column. You may use both basic statistical analyses such as correlations and present them as visual graphs or tables (raw data).

- Preparing Data

You'll use historical data to train your model. Data may contain duplicate records and outliers; depending on the analysis and the business objective, you decide whether to keep or remove them. Also, the data could have missing values, may need to undergo some transformation, and may be used to generate derived attributes that have more predictive power for your objective. Overall, the quality of the data indicates the quality of the model. You need to provide a data dictionary of all data items used in your analysis and their justification to be included in your model.

## Assignment 3: INF30030 – Business Analytics

- Sampling Your Data

After preparing the data, the next step is Data Sampling. The data needs to be split into two sets: training and test datasets. While splitting, consider the % split between training and test data – It's always good to have more training data than test data (Rule of thumb – 70% training and 30% test data). Also make sure that the splitting process produces a stratified sample rather than a pure random sample. You need to build the model using the training dataset and the Test data set should be used to verify the accuracy of the model's output. Doing so is absolutely crucial. Otherwise you run the risk of overfitting your model — training the model with a limited dataset, to the point that it picks all the characteristics (both the signal and the noise) that are only true for that particular dataset.

- Building the Model

Sometimes the data or the business objectives lend themselves to a specific algorithm or model. Other times the best approach is not so clear-cut. As you explore the data, run as many algorithms as you can. Make sure you use techniques such as cross validation and ensembles as well to see if your modelling improves.

- Evaluating the Model

Each model iteration has to be evaluated and improved upon. To do the comparison models need to be evaluated based on model metrics such as confusion matrix, accuracy, precision, and recall. The final model should be the most optimized model based on the model metrics. Finally, you have to be smart how to present your results to the business stakeholders in an understandable and convincing way (such as reports, charts and/or dashboard) so they adopt your model.

### **Datasets**

To assist you with your assignment task, you are provided with a dataset (see Canvas > assignments folder) to help you build a model. At step 4 you should split your data into training and test or cross validation.

### **Deliverables**

Submit a softcopy of the project report including the six phases of model building. You will be required to present your project after the submission of your assignment.

Good luck