# Class Imbalance Problem

Project 5

## Exercises

**Exercise 5.1.** [15 p.]

Let's consider once again the Wisconsin Diagnostic Breast Cancer [https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29] data available at UCI Machine Learning Repository.[1]

The data consist of two classes, malignant and benign (let's denote them 1 and 0, respectively), of sizes around 37% and 63% of all observations.

We are interested in investigating various techniques design to deal with imbalanced data. The proposed study plan is as follows.

1. Consider logistic regression model.
2. Evaluate the model trained on the original dataset.
3. Reduce the number of observations in malignant class to be only 10% of the whole dataset.
4. Evaluate the model trained on the data created in step 2.
5. Use some variants of undersampling, oversampling and cost sensitive approach and calculate overall quality of logistic regression model used in within these contexts.

At each step models should be evaluated in terms of accuracy, recall and precision. Use crossvalidation or train-test split, whatever you decide to be the best.

**Results**

As usual, present your results and conclusions in `Jupyter Notebook` or `knitr` short report.

[1] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.