# 03. Classification – Homework

## Overview

* [Thyroid Disease dataset](http://archive.ics.uci.edu/ml/datasets/thyroid+disease) [[CSV](https://www.openml.org/data/get_csv/57/dataset_57_hypothyroid.arff)]
* Target – Class

## Evaluation Criteria

Scoring (10 max):

* EDA: exploration of variables and properties of data with conclusions - 1
* Data preprocessing *if needed - 2*
  + Handling missing values
  + Dropping unimportant features
  + Data normalization
  + Categorical encoding
* [Scikit pipeline](http://scikit-learn.org/stable/modules/classes.html#module-sklearn.pipeline) usage – 2
* Metrics and cross-validation chosen as well as reasoning behind – 1
* Different techniques for handling imbalance data (over\under sampling, class weight) - 2
* Feature engineering e.g. [transformations](http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.FunctionTransformer.html#sklearn.preprocessing.FunctionTransformer) (nonlinear) - 1
* Feature importance and [hyperparameters](http://scikit-learn.org/stable/modules/classes.html#hyper-parameter-optimizers) tuning – 1

Models: Logistic Regression w\wo regularization; K-nearest neighbors