# Credit project - guidelines

## Data

The data comes from the website (http://archive.ics.uci.edu/ml/datasets/Adult); this is a publicly available dataset used in machine-learning projects.

## Purpose

To develop a model (logistic regression) that predicts a person's income (under 50k per year or over 50k per year) from the data

## Tasks

1. In the type_employer column:
    1. Replace the entries "Federal-gov" and "Local-gov" with "SL-gov"
    2. Replace the entries "Self-emp-inc" and "Self-emp-not-inc" with "self-emp"
2. Reduce the number of entries in the "marital" column to three (Married;Not-Married;Never-Married)
3. Reduce the number of entries in the country column (e.g. grouping by continents? another approach?)
4. Replace "?" entries with NA values
5. Delete rows containing NA entries
6. Divide data into test and learning set
7. Build model using functions (glm)
8. Refine the model using step functions
9. Test model on test data (confusion table)
10. Calculate the F1-score for the developed model