# Project Assignment 1:
# Baseline Predictor

The goal of this assignment is to build a baseline system for automatic recognition of the age, gender, and personality of social media users. To implement the baseline, you only need to use the known labels (i.e., train data). A training dataset with 9500 labeled instances is provided.

Your software need to "predict" the following information about the user as **output**:

- **gender**, as either "male" or "female"
- **age**, as either "xx-24", "25-34", "35-49", or "50-xx"
- **personality**, as a score between $[1, 5]$ for each of the five traits of the Big Five personality model, namely Openness to experience, Conscientiousness, Extroversion, Agreeableness, and Emotional Stability (reversely referred to as Neuroticism).

Your project description contains a brief description of the personality traits of the Big Five Personality Model. Predicting gender is a binary classification or concept learning task. Predicting age is a multi-class classification task. Predicting the personality scores corresponds to solving five regression tasks.
For this assignment, you need to implement baseline predictors either *majority baseline* or *average baseline* to infer gender, age and personality traits.

- Classification Task: Implement *majority baseline* which predicts the label that is most common in the training dataset.
- Regression task: Implement *average baseline* which predicts the score that is mean value score of the training dataset.

The scores from these baseline algorithms provide the required point of comparison when evaluating all other machine learning algorithms that you will develop for the project.

## Required Submission Format

The testing command `ift6758` shall take as input (i) an absolute path to a test dataset with new instances (not containing the labels) and (ii) an absolute path to an empty output directory; the format looks like this:

```
ift6758 -i path/to/test/my-test-data/ -o path/to/output/directory/
```

The format of the test dataset is the same as the training dataset, with an identical internal directory structure. For each user of the test dataset, your software must output a corresponding XML file that looks like this:

```
<user
id="8157f43c71fbf53f4580fd3fc808bd29"
age_group="xx-24"
gender="female"
extrovert="2.7"
neurotic="4.55"
```

```
agreeable="3"
conscientious="1.9"
open="2.1"
/>
```

This file should be saved in the specified output directory and have the user's id value as the base file name and "xml" as its extension. For example, the file name of the XML file whose contents are shown above would be:

```
8157f43c71fbf53f4580fd3fc808bd29.xml
```

A public test dataset with data of 334 users (no labels though!) is available in the folder `/data/public-test-data`. Check that your software runs properly by executing the command:

```
ift6758 -i /data/public-test-data/ -o ~/results/
```

Using a similar command as the one above, the predictive capabilities of your software will be tested on a "hidden" test dataset of $n = 1334$ users. This set contains the 334 users from the public test dataset, as well as 1000 new users who are not in training dataset nor in the public test dataset. Your solutions for age and gender will be assessed based on **accuracy**, i.e. the number of correctly classified instances divided by the total number $n$ of instances. For personality identification, the average **Root Mean Squared Error (RMSE**) over all five personality traits will be used. RMSE is defined as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

with $y_i$ the actual value and $\hat{y}_i$ the predicted value.

## Due: Monday 30 September

**Your software will be tested at some time point between Tuesday 9am and Thu 9am. Make sure that on Monday before 11:59pm, the command `ift6758` invokes a stable version of your software on your account ( /submissions), and that this stays available until at least Tuesday 9am. We automatically collect your software on Monday at midnight.**