# Non-neural machine learning (MIB-ITLB358) Home Assignment

You will be working with a real-life dataset that contains user- and session-related features from a cursor flow dataset. Your task will be to predict whether a user clicked on a displayed ad (Group A, classification), or the binned income range of the user (Group B, regression).

*data source:* [*https://gitlab.com/iarapakis/the-attentive-cursor-dataset*](https://gitlab.com/iarapakis/the-attentive-cursor-dataset)

The raw datafiles:

* participants.tsv: <https://drive.google.com/uc?export=download&id=1U3h6wTWeG57gpVJsdgwrbJwYWxJHYv_6>
* groundtruth.tsv: <https://drive.google.com/uc?export=download&id=1wQ1PP0bXLfnhwZouOS_wuoCozSTpKR4H>

The prepared data for modelling (in case you cannot tackle the data preparation step):

* prepared\_data.csv: https://drive.google.com/uc?export=download&id=12VqL8hE6C-83Cu8SvX8uPAbMnFOLbiIH

## Task description

### First task for all groups: data preparation

* Load the participants.tsv and groundtruth.tsv files into **pandas dataframes**.
* Make sure that user\_id values are **unique** and that the set of user\_id's is the **same** in the two datafiles.
* **Merge** the ad\_clicked and attention columns of the groundtruth data into the participants data.
* Try to make numeric columns have **numeric** datatypes.
* Make the country column have the value **"non-USA"** in all cells in which it's not **"USA"**.
* **One-hot encode** the columns "country", "gender", "ad\_position", "ad\_type", "ad\_category", making sure
  + to handle cells that encode **undefined values**, and
  + to **break perfect multicollinearity**.

*If for some reason you cannot tackle this data preparation step, you can use the* [*prepared\_data.csv*](https://drive.google.com/uc?export=download&id=12VqL8hE6C-83Cu8SvX8uPAbMnFOLbiIH) *and move directly to the modelling task. Just load the prepared\_data.csv file into a dataframe, and go on to selecting the X, y (input, target) data from it.*

### Modelling task

#### Group A: Classification

* Prepare...
  + ...your **input data**, which should consist of the "education", "age", "income", "attention" column values and the one-hot encoded variables.
    - Explore and handle any missing values as you see fit.
  + ...your **target data**, which should be the "ad\_clicked" values.
    - Explore the class ratios.
* Create a **train-test split** with 33% of your data as test set. Consider best practices when creating such a split.
* Train a **linear model** that can predict the target from the independent variables.
* Train a **random forest model** that can predict the target from the independent variables.
* **Evaluate** and compare your models. Can you conclude anything about the relationship between the input and target features?

#### Group B: Regression

* Prepare...
  + ...your **input data**, which should consist of the "education", "age", "attention" column values and the one-hot encoded variables.
  + ...your **target data**, which should be the "income" values.
  + Explore and handle any missing values as you see fit.
* Create a **train-test split** with 33% of your data as test set. Consider best practices when creating such a split.
* Train a **linear model** that can predict the target from the independent variables.
* Train a **random forest model** that can predict the target from the independent variables.
* **Evaluate** and compare your models. Can you conclude anything about the relationship between the input and target features?

## Please observe the following

* You must use a single standalone Jupyter Notebook to solve the task and submit the .ipynb file. **Upload your .ipynb file to Moodle.**Note for those working on Google Colab: a link to your notebook will not suffice: you have to download and submit the file itself.
* Follow the **principle of literate programming**, and make use of the markdown cells of the notebook.
  + Don't just code, explain and motivate the steps you take.
  + Don't just display results and plots, interpret and evaluate them.
* For the **deadline**, please refer to the Moodle page of the module.

## Assessment

The assignment will be assessed based on the following criteria (see the marking grid on Moodle):

* Specification fulfillment and conceptual grounding (60%)
* Literate programming and markdown cells (20%)
* Coding (20%)

The **resit arrangement** for the assignment is the same as above; you may resubmit the same paper, with corrections, that you submitted by the original deadline. The resubmission deadline will be specified on Moodle after the grades for the original submission are published.

## Academic conduct notice

Where the Academic Conduct Officer has reason to suspect that a piece of work submitted by a student was wholly or in part written by someone other than the student who submitted it, and this has not been disclosed by the student, they may call for the student to defend the work in viva or a written comprehension test. The burden of proof in such a viva or test will be upon the student to demonstrate to the examination panel’s satisfaction his/her full comprehension of the work s/he has submitted. Failure to appear without satisfactory explanation will result in immediate failure of that assessment, with consequences of academic misconduct and application of sanctions.