

# **Home assignment - Data Analytics 2 (ITLB355, MIB)**

## **Task description**

You will process real-life data which includes historic US election results.

### **The tasks of both groups:**

Your task is as follows:

- What are the steps if the dataset were "big data"?
  - Theoretical question!
  - Please create a code in a markdown cell, which represents a basic big data pipeline (with explicative steps).
  - If you want use pseudo-code.
- Download the Election dataset from Github
  - Practical steps  
!wget "https://drive.google.com/uc?export=download&id=1aw0kM5a4Q-bhbHyqK4HI-tHrvxVpxpWm" -O 'election.csv'
- Modeling
  - **Team A**
    - Classification
      - Decision Tree (baseline)
      - Random Forest Classifier
      - Split data into inputs and outputs
      - Split data into train and test set
      - Input features:
        - year, state\_fips, state\_cen, state\_ic, candidatevotes, totalvotes, (one-hot encoded) writein
      - Output feature(s):
        - (one-hot encoded) party\_simplified

### **Please observe the following:**

- You must use a single standalone Jupyter Notebook to solve the task and submit
  - the .ipynb file, and
  - a .pdf file generated from the notebook.

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.

## **Deadline**

Please refer to the Moodle page of the module.

## **Assessment**

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

- Specification fulfillment (50%)
- Literate programming and markdown cells (20%)
- Execution performance (20%)
- Clean code (10%)

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.

Upload your files(.ipynb, .pdf) to Moodle.

## **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.