IA & SHS: Data challenge

March 7, 2022

1 Introduction: what the Data Challenge is about?

Inputs: You will work on data containing the compensations of state employees, along with other related information such as their gender, employment date, the agency they belong to, grades and so on. Specifically, you have a *dataset to train your model* (`train_val_data.csv`) with the true labels, and a *test dataset* (`test_data.csv`) without the labels, on which you are asked to predict the annual compensation[1].
Link to the data: `https://bit.ly/2022_DataChallenge`

You you will be asked to:

1. to start with an EDA (Exploratory Data Analysis)[2] (uni and multi-dimensional) to summarize, describe, and visualize data;
2. predict the annual compensation (annual) on the test set, based on all the other features;
3. elaborate on your strategy and why the methods you proposed performed better or/and are more interpretable than others;
4. *(optional)* analyze the effect of gender on salaries. Imagine you are a statistician in charge of the data analysis. You are asked by a journalist the question: *is there an effect of the gender on the annual compensation?* What answer would you provide?

Bonus: extra-points will be awarded for those who achieve the best results in any of the following areas:

- the EDA, that is the most ways of vizualizing the data;
- the prediction on the test set;
- the (optional) analysis of gender on salaries.

See section 2 for the instructions.

Figure 1: You probably know him. Just as Hercule Poirot's challenge is to *predict* who the murderer is based on the *evidences* he obtains, as a Data Scientist, you are challenged to make your best *prediction* from the *insights* you get from the *data*.

[1] In Python, using the pandas (pd) method `pd.read_csv(datasetname.csv)` will open the datasets correctly
[2] `https://en.wikipedia.org/wiki/Exploratory_data_analysis`

2 Expected outputs and instructions

2.1 Predictions and two-page summary

Before April 3, 12:59 (CET) you will send at

- `lorenzo.gasparollo@umontpellier.fr`,
- and `benedicte.colnet@inria.fr`,

the following documents,

1. the predicted values for the annual compensation on the test dataset in a form of a csv file called name_surname_predictions.csv;
2. a two-page summary note in PDF called name_surname_twopage.pdf following the same structure of your presentation but with more details (see 2.2).

2.2 Presentation

Furthermore, on April 4, you will be asked to present your approach and methodology. You will prepare the slides for a **5 minutes** speech, followed by 5 minutes of questions from the class and teachers.
We would like to see a clear description of your approach and quantitative results; therefore, it is helpful to organise the presentation in these three sections:

1. Method - *how did you approach the problem? why did you do the preprocessing that way, choose that particular model, etc.?*
2. Discussion on your best prediction results - *can you explain why?*
3. Conclusion - *what are the main takeaways of your analysis?*

We stress that you will have to **stick to a 5 minutes presentation**.

2.3 General guidelines

You are free to use the programming language and tools that work best for you and we do not ask you to provide code snippets.

Furthermore, we suggest you to be as much clear as possible when presenting your approach. Be critical. There is no one good answer and the evaluation is mostly done based on the method you propose. Feel free to discuss with other students about your approach and do not be afraid to ask questions during the labs.