FINAL PROJECT
GENERAL INFORMATION

• This assignment will be solved in groups of **five** students. There are 45 students registered for the course that makes for 9 groups of 5 students.
You have until December 5th at 23.55h to upload a document indicating the names of the students in each group. One document per group. After this date, all changes/incorporations will be handled by the teachers.
• **The data**: To complete the project, all groups will use Kaggle's AI/ML Salaries dataset. The data description and csv file are available at:
`https://www.kaggle.com/datasets/cedricaubin/ai-ml-salaries`
• **Final report due date**: February 1st, 2023 at 23.55h.

INSTRUCTIONS

- Each group must turn in a report explaining the work done and the results obtained. Maximum length: equivalent to 8 pages in pdf format plus cover page.
- The project must be done in R. Rmd files must be turned in together with its html/pdf output.
- For deadline and uploading instructions, please consult aula digital.

The project has two parts:

PART 1: Data analysis.
In this part, you can apply all concepts seen throughout the course. You may consider all features or a subset (only study US, ...) depending on the research questions asked in Part II. Select the most informative graphical representation.
If you must do significant work to get the data or convert it into the proper format, then describe the process and effort required. How many examples are in the data set? How many features? Which will be used?
PART 2: Context of the problem, models and evaluation.
You should start by giving a brief description of what you plan to do. What problems are you trying to solve? Be sure to formulate the problem as a data mining problem (is it a classification problem, a clustering problem, association rule mining, ...)? What exactly are you trying to predict (for prediction tasks), group (for clustering tasks), ...? How will you evaluate your results? How will you know if your results are good? It is critical that your problem is well-defined. What learning tools do you plan to use and what techniques/algorithms do you plan to use (decision trees, Apriori, ...)?

Depending on how complex your project is, you should consider including parameter tuning, key features of the data distribution, dimensionality reduction, ...