**CSCI - 4146 - 6409 - The Process of Data Science - Fall 2022**

**Assignment 1**

**The submission must be done through Brightspace. Due date and time as shown on Brightspace under Assignments.**

* To prepare your assignment solution, use the assignment template notebook available on Brightspace.
* The detailed requirements for your writing and code can be found in the evaluation rubric document on Brightspace.
* Questions will be marked individually with a letter grade. Their weights are shown in parentheses after the question.
* Assignments can be done by a pair of students or individually. If the submission is by a pair of students, only one of the students should submit the assignment on Brightspace.
* We will use plagiarism tools to detect any type of cheating and copying (your code and PDF).
* Your submission is a single Jupyter notebook and a PDF (With the compiled results generated by your Jupyter notebook). The PDF is the printout of the final notebook with all the code and results. File names should be:
  + **A1-<your\_name1>-<your\_name2>.ipynb**
  + **A1-<your\_name1>-<your\_name2>.pdf**
* **Forgetting to submit both files results in 0 markings for both students.**

**In this assignment, you will use the Rating of European Restaurants dataset.**

**Link to the dataset**<https://www.kaggle.com/stefanoleone992/tripadvisor-european-restaurants>

You will go through the first three stages of a data science project defined in CRISP-DM and derive some data insights.

1. [0.3] **Business understanding**
   1. Formulate a business problem that can be solved with the dataset.
   2. For the business problem, propose 3 data science solutions and assess their feasibility (required data, required business capacity). Select the final solution and explain your decision.
   3. For the final solution, identify the prediction subject, its domain concepts and sub-concepts (if there are any). Draw a hierarchical graph of the concepts.
   4. For each domain concept, design descriptive features that best describe a concept using data from the dataset. Summarize the resulting ABT in a table with the following columns:
      1. Feature Name
      2. Domain Concept
      3. Feature Description
      4. Feature Type
      5. Data Type
2. [0.3] **Data Exploration**
   1. Build the data quality report of the resulting ABT.
   2. Identify data quality issues and build the data quality plan.
3. [0.2] **Data Preparation**
   1. Preprocess your data according to the data quality plan.
4. [0.2] **Data Insights**
   1. Build the correlation heatmap of the features in the ABT. Derive insights from it, and relate it to the business problem being addressed.
      1. What are the descriptive features that highly correlate with the target feature? Propose some hypotheses explaining the correlation.
      2. What are the domain concepts that highly correlate with each other? Propose some hypotheses explaining the correlation.
      3. Are there any features that are useless for a predictive model?