MIDTERM PROJECTS (Groups of 3 / 4 people)

The midterm project consists in 2 parts, which different deadlines.

PART 1: DATASET GENERATION

Each group will generate a dataset to be used in the following machine learning tasks:

1. **Multivariate Linear Regression**
2. **Supervised Learning** (Classification with up to 10 classes; must provide a classification algorithm achieving at least 80% accuracy)
3. **Unsupervised Learning** (Ensure that clustering is not too trivial but still solvable)

Requirements for Dataset Creation

- The datasets should be realistic, well-structured, and grounded in a meaningful **business context**. Avoid overly simplistic or artificial data; the dataset should simulate a real-world scenario with clear relevance and purpose.
- Provide a brief description of the dataset and the problem it addresses.
- Ensure enough samples (at least 500 observations per dataset).
- For classification, the dataset must allow a model to achieve at least 80% accuracy.
- Save datasets in CSV format.

DEADLINE PART 1 -> 28 Nov

PART 2: DATASET ANALYSIS

After dataset creation, all datasets will be made visible, and each group will evaluate all other groups' datasets (you must use the provided ***evaluation_criteria_datasets.*xlsx** to evaluate other people's work).
Then, you will select one of the datasets (your choice) and perform an analysis of the data, following the steps below.

Analysis Questions

Each group must analyze the dataset assigned to them using a systematic approach. Consider addressing the following:

1. Exploratory Data Analysis (EDA)
2. Model Selection
3. Evaluation & Validation
4. Interpretation
5. Challenges & Improvements

Before submitting, you must complete your own self-evaluation form (template available in the same Excel file).

The objective is to apply critical thinking to your own work and reflect on what strategies you can adopt next time to produce an even more effective and insightful analysis.

DEADLINE PART 2 -> 14 Dec