Project Specification

Group 1:

- CAI, EDISON
- MEREACRE, SERGIU
- OBrien, Jack
- WALSH, DAVID

---

1. Dataset: steel_industry_data_excerpt.csv

---

2. Project Teams

This is a group project. Each group must consist of 3-5 students, and typically 4 students.

---

3. Tasks

3.1. Dataset and Predictive Analytics Task

In this project, you must use the dataset provided above. It is an excerpt from the Steel Industry Energy Consumption dataset described at https://archive.ics.uci.edu/dataset/851/steel+industry+energy+consumption. Please refer to this link for the meaning of the columns. In your project, please make sure you use the excerpt with 2190 data rows provided here and NOT the original dataset with 35040 data rows.

3.2. Use the lab exercises as the basis for creating one or multiple Jupyter notebooks that contain:

a. The names and IDs of the students in your project group, which parts of the project each student has worked on, and what percentage of the whole work on the project is their contribution. If you submit multiple notebooks, include this information in all notebooks. If the percentage of work done by each student is not specified, it will be assumed that all team members contributed equally to the project.

b. ***Exploratory data analysis:*** Perform exploratory data analysis of the dataset. Comment on your observations.

c. ***Clustering:*** Construct and train clustering pipelines for at least two clustering algorithms. Visualise the clusterings and discuss their usefulness for a better understanding of the underlying patterns in the dataset. Please note that in the lab exercise on clustering, we did not use pipelines, while here you are asked to use them.

d. ***Regression:*** Attempt at least three different regression algorithms to train regression models for predicting the value of attribute **'Usage_kWh'**. Construct a pipeline for each of the algorithms you have chosen. Evaluate the trained models and select the best model. Explain your choice in a markdown cell. Aim at training a model that is as good as possible. In this process, you may attempt various data preparation and possibly dimensionality reduction strategies.

e. ***Classification:*** Let ***m*** be the mean for column **'Usage_kWh'**. Replace column **'Usage_kWh'** with the categorical column **'Usage_kWh_categorical'** with two categorical values: **'Low'** for **'Usage_kWh'** $\leq$ ***m*** and **'High'** for **'Usage_kWh'** $\geq$ **m**. Attempt at least three different classification algorithms for training models that can be used for predicting the value of **'Usage_kWh_categorical'**. Construct a pipeline for each of the algorithms you have chosen. Evaluate the trained models and select the best model. Explain your choice in a markdown cell. Aim at training a model that is as good as possible. In this process, you may attempt various data preparation and possibly dimensionality reduction strategies.

---

4. Submission

a. **Deadline:** Sunday, 28th of April, 23:55
b. Submit your Jupyter Notebooks on Brightspace.

---

5. Marking Scheme/Rubric

3.2b EDA For higher marks: Aim at utilising various plotting techniques. Comment on the insights provided by each plot. Draw overall conclusions from the EDA.	10 marks
3.2c Clustering For higher marks: Fine-tune some of the parameters of the estimators in the pipeline with grid search. Attempt to describe the clusters you have found with words.	10 marks

3.2d Regression For higher marks: Fine-tune some of the parameters of the estimators in the pipeline with grid search.	10 marks
3.2e Classification For higher marks: Fine-tune some of the parameters of the estimators in the pipeline with grid search.	10 marks