GREEN PRAXIS
Environmental Data Scientist - Take Home Test

Exercise 1 - Descriptive

Based on a data science project you have performed in the past

- Describe how you went about the project in some detail: what were the goals, what were the data sources available, what were the process steps you took, what techniques did you use, why did you use these techniques, what did not work as expected, why not, what were your final conclusions, what did you learn?
- Write 1-2 pages max

Exercise 2 - Coding

Exercise 2a - Soil Carbon:

- Take the two RMQS excel sheets that have been shared to accompany this exercise. These are measures of carbon sequestration in the soil based on different soil characteristics. Some key details:
    - These are soil samples from 'Réseau de Mesures de la Qualité des Sols' and we are interested in learning about the relationship between soil characteristics and carbon sequestration. Specifically, we want to see if we can use this dataset to create a model to enable us to predict the carbon content of soil based on soil and plant characteristics.
    - The unique key for each sample is the combination of `id_site` and `no_couche` (layer number)
    - Our target variable (dependent variable) is `carbone_16_5_1`, column AC of the analyses_composites sheet. This is the % of carbon per unit of soil.
    - The features we will consider for this exercise are:
        - Soil composition - this is defined as the % of clay and silt, and in this sheet is defined as the sum of `argile`, `limon_fin` & `limon_grossier / 1000` (from `RMQS1_analyses_composites`)
        - Soil ph: `ph_eau_6_1`
        - Vegetation cover - this is defined in `desc_code_occupation1 and desc_code_occupation3` (from `RMQS1_occupation_nommatp_nomsol`).
        - Location - this is defined in `x_theo y_theo` in the Lambert 93 coordinate system. If you choose you can use the 3rd excel file `SiteTempPrec` which has a mapping of sites to average precipitation and temperature which can be used as alternative features.
- Use this data to create a model with the goal of predicting how much carbon we could expect to be stored in soil based on an input set of feature values

- Perform an assessment on the quality of the model and comment on what you see. Do you see a possibility of being able to use this for predictions? What options do we have to make the model perform better?
- We want to perform our own sampling to build our own dataset with elements that we expect to add predictive power. If we capture 7 numeric and 3 categorical variables at each site, how many samples should we take? What other information would you need to know to make this assessment. Please explain your reasoning.

Exercise 2b - Leaf health (bonus extra):

- Take this leaves dataset
- Create a model that predicts if a leaf is healthy or diseased
- Perform an assessment on the quality of the model. What do you see? How might it perform better?