Data Science Competition 2023

Introduction

Context

You're a group of young data scientists, freshly graduated from the Department of Informatics of the University of Bergen. A biology lab has contacted you to uncover the mysteries of the measurements they made in their last experiments. They had some vague objectives when doing the experiments, but don't have the expertise to analyse the dataset themselves and struggle to find the right words when trying to explain to you what to look for in the data. So here you are, ready to accomplish your first task as a data scientist!

You are allowed to work in groups (up to 3 people in a group) or alone and you can ask questions to your fellow data scientists on Discord, server: "Data Science Competition 2023" (invite link). You are also allowed to ask questions to the lab that contacted you, represented here by Natacha Galmiche (`natacha.galmiche@uib.no`) either on the "general-questions" and "technical-issues" channels or by direct messages on Discord if your question really needs to be private.

Tasks

In the OneDrive folder, you're given a zip file `competition_dataset.zip` containing

- a dataset that consists of two files `abundance_table.csv` and `env_parameter_sample.csv`.
- a technical description of this dataset `dataset_description.pdf`.
- a jupyter notebook `competition-getting_started.ipynb` to help you get a better understanding of the dataset with some basic operations.

You're encouraged to read `dataset_description.pdf` first and then the notebook `competition-getting_started.ipynb`. Once you are more familiar with the dataset, you can start exploring it yourself or in groups. You have until Friday 23.59 to make groups. (see the `groups.xlsx` file in the OneDrive folder)

Note that the dataset is a part of an ongoing research project (in real life too!)! So **we kindly ask you not to publish the data on the internet nor to send it to anyone outside this data science competition**. Thank you for your understanding!

Potential tasks biologists mentioned to you

Once you have a rough overview of the dataset, you may start taking on one (or some) of the following tasks:

- Study potential relationships between a given sample environment or specific condition within an environment and the composition.
- Study the impact of the different aquaculture systems, treatments on the abundance.
- Predict the composition given the metadata. It could be done for the entire taxonomy or only at a specific taxonomy rank (e.g. "R") or sub-rank (e.g. "R1") (regression). This could be useful because

doing the actual sampling has a cost, especially for lines that are classified "No_sample" in the `env_parameter_sample.csv` file.
- Find anomalies in replicates (each "sample" XX_X has 3 replicates XX_XA, XX_XB, XX_XC to check for experimental or procedural error) (anomaly detection, visualisation)
- Study the evolution of a given sample with time. Each "sample" XX has 3 time steps (XX_1X, XX_2X, XX_3X). (Regression, statistical inference, visualisation)
- Predict the evolution of the abundance for future time steps.
- Predict single values (could be useful in case something couldn't be measured or if there is an anomaly in one of the replicates of a sample)
- Dimensionality reduction/feature selection, each sample has a dimension of roughly 23500...
- Study the potential relationships between the different organisms.
- **And many more! Don't hesitate to surprise them, after all, *you are the data scientist*!**

You don't know where to start?

Then, first imagine you are trying to explain the dataset and its properties to yourself. This should already give you some good content to share with the biology lab. Then, explore the path that is most interesting to you, anything would be valuable to your contractor, who has no mean to explore the dataset themselves!

General remarks on the tasks

Note that if you are handling too many datapoints at a time, it is always possible to

- ignore the "time series" aspect of the data (e.g. look at a specific time step, or look at the abundance, regardless of the time step).
- look at certain taxonomic ranks (e.g. "R") or sub-ranks (e.g. "R1").
- look at specific aquaculture system / replicate / treatment.

Finally, remember that this is a perfect example of a really life setting.... and in real life datasets are not always perfect and clean and the results are not always astonishing. Learning how to overcome this is a part of the key objectives of this competition!

Delivery

In research everything should be reproducible, therefore, the biology lab wants to get your source code. However, the biology researchers are not experts in programming, so they expect you to share your results and explain your methodology without having to run your code. Thus, rather than a simple python file (`.py`) they would prefer a **jupyter notebook** (`.ipynb`) file (using python or R, and potentially using Google Colab) **with the output of each cell visible and comments on your results and methodology throughout the notebook**. If you can only provide a python/R script without using a notebook, then they would need a report (as a `.pdf` file) explaining your general approach, and with your results and comments on your results.

Once all the deliveries are received, the lab will give you feedback! There will be winners in each category. The category of a group is defined as the highest category among its members (e.g. a group of 2 bachelor and 1 master students is in the master category).

It is planned to **announce the winners of the competition on Friday, November 17, from 14:00 in the Big Auditorium**. We will keep you updated on that.

What to submit:

A `.zip` file containing:

- a jupyter notebook (`.ipynb`) (+ optionally a `.pdf`)
- **OR** a python/R script + pdf.

Where to submit: by email to `natacha.galmiche@uib.no`. In the filename of your submission, please include:

- Your category ("bachelor", "master" or "PhD")
- Your group number as specify in the `groups.xlsx` file (Ex: "Gr1") or your name (E.g. "Natacha_Galmiche")
- Example `PhD-Natacha_Galmiche.zip` or `Master-Gr12.zip`.

Deadline: Sunday, 5th November at 23:59.