**Dataset :**<https://archive.ics.uci.edu/dataset/913/forty+soybean+cultivars+from+subsequent+harvests>

**Tasks :**

1. Create a **public** Github / Gitlab repository. **Update your repo regularly** as you complete the following tasks.
2. You are allowed to structure the project as you wish, however we recommend the following project structure:
3. **data/** (directory where datasets will be stored)
4. **models/** (directory where models will be stored)
5. **main.ipynb** (main file used to review your project)
6. **other.py modules**can be placed in the root of the project. These modules should serve to structure classes / functions that are impractical to be placed within the notebook itself. Make sure to give these modules suggestive names.
7. **requirements.txt**(should contain required libraries that you use within your project)
8. Create a short introduction where you present your table of contents and approach to solving the problem(s).
9. Perform data cleaning, data preprocessing and feature engineering tasks if necessary
10. Exploratory Data Analysis **(EDA)**
11. What are the most important factors in determining MHG (Thousand seed weight)?
12. What are the most important factors in determining GY (Grain yield)?
13. What is the delta in MHG and GY from Season 1 to Season 2?
14. You are allowed to further investigate the dataset based on your own curiosities
15. **Predict MHG** based on sintethic data you will create for a new cultivar; **Create clusters** to showcase similar cultivars. Include:
16. Experimentation
17. Choosing the models
18. Hyper-parameter tuning
19. Results
20. Conclusions
21. Send an email to [**kaili.lian@vodafone.com**](http://mailto:kaili.lian@vodafone.com/) with the link to your repository after you are finished and prior to 11*th* April 2024 EOD. **Make sure your repository is public and accessible.**

**Criteria :**

* **Fairness** - Based on your implementation and Github / Gitlab activity some assumptions will be made about your level of implication in the project vs assistance from 3rd parties. This project needs to reflect your personality and thought-process.
* **Presentation** - Ensure there is a logical flow that creates a story related to your findings, decision-making and conclusion.
* **Readability** - Use Jupyer's built-in markup functionality to add your thought-process, should you consider it necessary, as well as factors that influenced your decision making.
* **100% Completion is NOT a requirement.**
* **Completeness** - Be thorough in completing each task, make sure you accomplished all practical steps you considered necessary. Due to time constraints, we understand you won't be able to go through everything in high detail. However try to implement all major sub-tasks you will identify per task. Rather than prioritising completion of all tasks, focus on doing a good job on the tasks you will manage to complete.
* **Quality** - High quality code will be appreciated as it showcases your knowledge of the language standards and best practices as well as your experience with the programming language. It will also make it easier for us to follow your implementation.