**EXAM DOCUMENTATION.**

Dear sigmaritan,

Hope you enjoyed the learning and exploration of the Machine Learning field and you found this field as attractive as we all do. But now it’s time to put all the knowledge accumulated by you into practice. This examination was created with the aim to test your knowledge and to find the level of understanding of the Machine Learning and Data Analysis principles.

The exam will consist of two parts:

1. The exploration of the data set and training of a Machine Learning model on it.
2. The presentation of the insights and results obtained from the work you have done on this data set in a PowerPoint presentation in front of the jury.

The first part should be submitted by loading all the notebooks into a GitHub repository and sending the link to that repository to Vladimir Stojoc by the 27th of March 2022, 00.00.

The second part will be presented individually by each sigmaritans that is examined by the jury consisting of the members of the castor and pollux teams. The date of the presentation will be agreed upon personally with each person and the jury.

**TASK#1.**

In the archive, you have a CSV file. This file is the data that you should use during the exam passing.

In your case, it is a datasetcontaining data collected from various files belonging to UNICEF India. These data allow you to compare the quality of water from over 3200 water sources from India.

The Data Set has the following fields:

* ph - ph of the drinking water which must be lie between 6.5-8.5.
* Hardness - hardness is present in water like magnesium and calcium salts.
* Solids - Solid particles present in drinking water in Microns.
* Conductivity - basically drinking water acts as an Insulator since it doesn't contain any dissolved Salt, the lower the conductivity, the higher the potable.
* Carcinogenic - These are harmful Cancer-Causing agents that are present in Drinking waters that cause Cancer and are not potable.
* Sulfates - sulfate particles are present in the water.
* TriHalomethanes - Halomethanes are solvents present in water which made up of three of the four hydrogen atoms of methane are replaced by halogens.
* Turbidity - t is a measure of the relative clarity of a liquid.
* Organic\_compounds - H, CO2, HCN, NaCl compounds present in water.
* Potability - A binary column that shows if water is potable or not,

In this task, you should:

1. Load and analyze the data set, and try to extract insights from the data.
2. Prepare the data to machine learning algorithm ready format,
3. Handle the NaN values in the data set.
4. Apply some post-processing and post-analysis techniques such as Feature Selection, Feature Engineering, or others and see if they can help you.
5. Create a model or a Pipeline that will show the best performance in predicting the target column.

The target column is **Potability**. Try to create models that will predict this value.

Some more advice:

* You can split different steps of your work into separated notebooks.
* Commend the code and use the markdown cell type to explain what or why you are doing.
* Explain in a written form the choices and decisions that you made during the processing of the data.
* Use plots, but don’t abuse them.
* Don’t chase a perfect model, better show us why you chose a specific model as the best one.

**TASK#2.**

Your second task is to create and prepare a presentation on the work that you have done. In the presentation that you should explain the work that you have done and why did you make the choices that you made during the analysis and Machine Learning model development. Some more pieces of advice to the presentation are:

* KISS (Keep It Simple Stupid) - keep the presentation as simple as possible.
* Give some context about the field on which your data set is about.
* Try to avoid as much as possible having a lot of text on your slides.
* Try to tell a story.
* Don’t show the whole process, only the most important findings, and decisions.
* Finish with a conclusion.
* Don’t make lots of slides.

Good luck!