Welcome to Practical Lab 1 - Foundations

The submission should be done on the course website. Links to two files should be provided:

1. .ipynb file in the repository

2. .html file on GitHub Pages (that is, the rendered version, not the source code in the repository).

**How to Write a Good ML Report**

The notebook should be divided into sections, each one with its own title. Follow the process and structure of the Machine Learning workflow. Format your HTML as a short report. It should be divided into short sections, each one with a Markdown title. Note that your textual insights, conclusions clarity of exposition are an important part of the grade.

Notice to follow a [good notebook structure](https://github.com/CSCN8010/CSCN8010/blob/main/class_notebooks/notebook_structure/notebook_structure.ipynb), and focus on readability and clarity of the code, insights and conclusions.

In this lab you should be able to demonstrate the following:

* Structure a notebook according to the ML workflow
* Have a clear notebook, with explanations, graphs and insights directed at non-ML-experts (e.g. Product Managers).
* Run an efficient **EDA** (Exploratory Data Analysis)
* Fit multiple models and choose the best one using **cross-validation?**
* Relevant and clear model evaluation
* Provide clear conclusions, insights and recommendations on the best model to use

Write the notebook as if you are writing it to your peers at work, and specifically to your product manager. This is the person who provided you with the task at hand. That person has only high-level knowledge about the problem and of the techniques you use. They are the decision makers reviewing your suggested solution (model), so make sure that it is written such that they can correctly and quickly understand and assess your analysis in terms of its steps and conclusions and solution.

**[ ] Setting up your personal practical lab GitHub repository (1 point)**

The first step is to set up the groundwork for this lab, and the following ones for the rest of the course. We will use GitHub, Git, the Command Line Interface (CLI), and Visual Studio Code (VSCode) to create a repository. This repository will be used to hold all your course's practical labs. Steps:

1. Create a new public repository in your GitHub account.
2. Enable GitHub Pages.
3. Clone this repository onto your laptop.

**[ ] Univariate Linear Regression on the California Housing Prices Dataset (8 points)**

[Link to Data](https://www.kaggle.com/datasets/camnugent/california-housing-prices)

In this lab you will train your first machine learning model in our course - a univariate linear regression model. You are asked to run a linear regression between the *median house value*(dependent variable) and each of the following independent variables. You should **run three models**, one for each independent variable:

1. *median income*
2. *population*
3. *number of households*

Steps:

1. [ ] Framing the Problem - Describe the goal of this report (0.5 point).
2. [ ] Getting the Data - hyperlink to the source (0.25 point) and load into Pandas (0.25 point).
3. [ ] Exploratory Data Analysis (EDA) (2 points):
   1. Explore statistics (e.g. using Pandas `describe` function)
   2. Create three Seaborn scatter plots of *median house value*vs. *median income*, *population*and *number of households*. There should be one scatter plot per each independent variable. Add a title to each graph (using Seaborn). Provide textual interpretation of the graphs, and insights specific to the linear regression you are going to conduct next. (1 point).
   3. Based on the above, describe the 4 variables used in this lab. Observations such as their meaning, range and additional characteristics that you notice. This should be concise, and relevant (1 point).
4. [ ] Run three linear regressions (fitting) between (1.5 point):
   1. *median house value*vs. *median income*
   2. *median house value*vs. *population*
   3. *median house value*vs. *number of households*.
5. [ ] In a single table for all three linear regressions, provide per regression model (1.5 points):
   1. intercept and slope (0.5 point)
   2. Mean Squared Error (0.5 point)
   3. Mean Absolute Error (0.5 point)
6. [ ] Plot the resulting line on top of the scatter data for each of the three models (three separate graphs), add the line parameters (intercept and slope), MSE and MAE to the graph (you can use a text box). You can use any plotting library for that (Matplotlib, Seaborn or Plotly)  (2 point)
7. [ ] Summary: provide a conclusion. Compare the models in terms of their goodness-of-fit, and add additional insights you observed (1 point)

**[ ] Publishing your Notebook (1 point)**

Publish the Notebook as HTML using GitHub Pages. Convert your notebook into an HTML file and publish it using GitHub Pages. You can find step-by-step instructions on how to do this in the course repository, under the 'github\_pages' folder.