HW 1: Warm-up, Python and Jupyter Notebook

Due date: Jan. 31, 2023

1. Student Background Information

Send an e-mail message to me (yz22@nyu.edu) with the subject heading “Machine Learning in Molecular Science 2023, your name”. The message body should state your name, student status, course status (registered or auditing), and background information. The background could contain your course history, programming experience in python, background in math, your expectation, interest, concern or suggestions about the course, , and so on.

1. To go through the resources (Python Youtube Tutorials) on Brightspace website to install/learn/refresh Anaconda, jupyter notebook, and basic python programming. **If you plan to use Jupyter Hub to finish your homework, you can skip the Step2. You should be able to get access to Jupyter hub through the following web address:**

**https://chemga-1500001-spring.rcnyu.org**

1. (optional ) When using Anaconda on your own computer, you can create an isolated Python environment dedicated to this machine learning course. This is recommended as it makes it possible to have a different environment for each project (e.g. one for this project), with potentially different libraries and library versions. You can download **enviroment\_ML2023.yml**, and set up the **mlms2023** environment:

$ conda env create -f enviroment\_ML2023.yml

$ python -m ipykernel install --user --name=mlms2023

This environment contains most scientific libraries that are required for this course. This includes all the libraries we will need (NumPy, Matplotlib, Pandas, Jupyter, rdkit, pytorch and a few others), and **you should be able to get access to this mlms2023 kernel through Jupyter Notebook**.

**TASKS:** I. To go through examples 1 of Chapter 1 of Hands-on book {the example code can be found at: <https://github.com/ageron/handson-ml>3 }, tools\_matplotlib.ipynb, tools\_numpy.ipynb, and tools\_pandas.ipynb, and submit one jupyter notebook (HW1a-Yourname.ipynb), in which should finish the following tasks:

1. To check the data ( life satisfaction, GDP per capita ) regarding United States (5 points)
2. What is your fitted linear model ? Make a prediction for Unite States given its GDP per capita. ( 15 points)
3. Please summarize 5 key functions each for matplotlib, numpy and pandas that you have explored/learned. (30 points )
4. Following example 1 of Chapter 1, please use what you learned to make a linear model between measured solubility and molecular weight ( ESOL dataset delaney-processed.csv at : <https://moleculenet.org/datasets-1>) . Submit one jupyter notebook. (HW1b-Yourname.ipynb ) ( 50 Points)