Dataset

- Go to the web site: https://archive.ics.uci.edu/ml/datasets/Molecular+Biology+%28Splice-junction+Gene+Sequences%29
- Summary of the dataset:
    - This is a dataset of DNA sequences that define the boundaries between regions that are spliced out during protein creation (introns) or retained for protein creation (exons).
    - There are 3 classes: boundaries between exon and intron sequences (EI), boundaries between intron and exon sequences (IE), or sequences that are neither.
    - The dataset has 3190 examples of DNA sequences.
    - Each sequence is 60 base-pairs long.

Data challenge

- Generate a model in any ML/DL framework and use this data to learn to classify a 60 element DNA sequence into the categories of IE, EI or neither.
- Download the data at Data Folder (file name splice.data). Ingest the data into the ML/DL framework of your choice and build and train a model to learn this classification task. (No need to follow any modeling instructions on the webpage, if there is any)
- Compare performance of classical machine learning models and deep learning models.

What to submit

- In a Jupyter notebook or a similar format,
    - Document your model development thought process and the reason for various choices made in generating the model and training it.
    - For example, describe how you choose your network or hyperparameters and give justification for those choices.
    - While model performance is not critical, please quantify model performance and describe how you evaluated you model performance.