COS30018 - Option B - Task 2: Data processing 1

The current project code (**v0.1**) has many issues and limitations. Some of them are already discussed by the comments in the code. In particular, the data processing is not very good as the user must manually choose the start date and end date for the training data and then start date and end date for the test data. Furthermore, even though we can obtain a dataset with multiple features (e.g., Open, High, Low, Volume, AdjClose), the current version **v0.1** just ignores them all and use the feature Close only. There are several ways you can rectify these issues. In particular, you can learn better data processing methods from project (**P1**)

(**P1**) https://github.com/x4nth055/pythoncode-tutorials/tree/master/machine-learning/stock-prediction

Clearly, it is not good enough to just copy-and-paste the code from (**P1**) without understanding what it does.

Your tasks this week:

1. Write a function to load and process a dataset with multiple features with the following requirements:
    a. This function will allow you to specify the start date and the end date for the whole dataset as inputs.
    b. This function will allow you to deal with the NaN issue in the data.
    c. This function will also allow you to use different methods to split the data into train/test data; e.g. you can split it according to some specified ratio of train/test and you can specify to split it by date or randomly.
    d. This function will have the option to allow you to store the downloaded data on your local machine for future uses and to load the data locally to save time.
    e. This function will also allow you to have an option to scale your feature columns and store the scalers in a data structure to allow future access to these scalers.
2. Most of the above requirements have already been fulfilled by the code in the project (**P1**). Feel free to learn from it. But you will have to explain what their code does using detailed comments (the same way we commented the code in **v0.1**)
3. Upload your Task 2 Report (as a PDF file) to the project Wiki before the deadline and email your project leader to notify that it is ready for viewing and feedback.

Your Task 2 Report will contain the following details:

- Summary of your effort to explain the less straightforward lines of code, focusing especially on those lines that require you to do some research on the Internet. Note that, if you choose not to explain a particular line of code or explain it with too few details, a teaching staff can ask you to explain it and you fail to explain properly, you'll receive a low mark for this task.

Due date: 11:59pm Friday 25 August 2023

Assessment Criteria:
You can get up to 10 marks for successfully completing Task B.2.