Practical-2

Introduction to reproducible Machine Learning Operations

The aim of the practial is to get the hands-on experience of reproducing the machine learning operations at each stage. Student needs to apply the following steps in the practical.

1. Ensure that the numpy, scikit learn, and matplotlib libraries are available in your system. Create the requirements.txt file and make a note of the versions of these libraries.

2. Write a python code to import the Sample.txt data. Further, apply the following processes on the imported data.
   a) Scale the dataset. Use the StandardScalar function of scikit learn to normalize the dataset. Ensure reproduciblity: Store the standard scalar object into your local file system, so that the same data normalization can be applied to the other data during the deployement.
   b) Split the data: Write the python code to split the normalized data into randomly selected proportion as per the constant ratio. For e.g., if the constant ratio is 0.8, then the code must randomly select the 80 % proportion of the data as the training dataset and remaining 20 % as the testing dataset.
   c) Store the snapshot of the data as the numpy file. Ensure the same dataset could be loaded into separate python code. (Extra question: how to ensure that the same random generation could be achieved on each execution.)
3. Apply the linear regression algorithm on the dataset and assess the prediction on the test dataset.
   a) Store the trained model into the local file system to ensure the reproducibility of the prediction. Import the model and the test dataset into other python file. Check whether the same prediction is obtained in the latter case.