ML For Biopharma – Assignment 1

Name – Shrestha Srivastava
Roll no – 20260

Problem Statement – Predicting Bioactivity, Solubility and Molecuar weight of a compounds with respect to a biological target of any disease. Comapring deep learning techniques with machine learning techniques

Disease chosen – Enchephalitis and nervous disorders caused by Nipah virus

Biological target- Endothelial cells (Nipah virus also targets endothelial cells, which line the blood vessels. This can lead to vascular damage and a range of clinical manifestations, including respiratory and systemic symptoms.)

Solution:

Data collection- Data collection was done with the help of Chembl website. I downloaded a CSV file from chembl containing the information for endothelial cells and its solubility, bioactivity and molecular weight.

Data cleaning – To clean the data first of all I droped all the NAN values from "Smiles", "Bioactivity (or Standard Value)", "Molecular Weight", "AlogP (measure of solubility)", "Standard Units (measurement unit for bioactivity)"

The units of bioactivity were present in different units such as 'nM' '%' 'hr' 'uM' 'ug.mL-1' "10'-4No_unit" 'degrees C' "10'-3/s" "10'5/M/s" '% ID/g' 'µM' 'equiv' out of these I could only use nM, uM, µM. I dropped all the rows from the dataframe that contained other units. I then converted the uM and µM values to nM

Data Visualization – I plotted the data of bioactivity, solubility, and molecular weight.

Data Preprocessing – I used SMILE strings as input features, so to feed it to my model I converted the SMILE strings to one-hot encoded vectors. I also used solubility, bioactivity and molecular weight as output values.I also normalized the solubility, bioactivity and molecular weight using MinMax Scaler. Then I created the torch dataloader for train, test and validation data.

Model- I created a LSTM model using pytorch which contains one LSTM layer and 2 fully connected layers. I passed the last hidden state of LSTM to FC layers. I use validation data after each epoch and at the end I use the test dataset

Traditional machine learning approach – I use the random forest regressor for prediciting solubility, bioactivity and molecular weight. I follow the same steps for data cleaning and data preprocessing here as well. Since the random regressor only accepts two dimensional input. I flattend the encoded SMILE strings into one dimension (the other dimension being batch size).

Conclusion – Between traditional machine learning and deep learning tehnique, deep learning clearly outperforms the ML models.