**Bioinformatics Project – SARS Coronavirus Bioactivity Prediction**

**Defining the Problem:** The problem was to predict Bioactivity values using ML models, on the SARS Coronavirus data

**Our Approach and Strategy:**

We downloaded the ChEMBL data of target SARS Coronavirus (Replicase polyprotein 1ab single-protein database) and filtered out the IC50 type molecules from the database.

After filtering data, we removed all the data records of molecules which had the IC50 value missing. We then reindexed and saved the unprocessed, filtered data into a .csv file.

We classified the molecules based on their bioactivity and we learnt that the bioactivity data of each compound Is in the IC50 units. We classified the molecules as follows:

1. IC50 value <= 1000 : active
2. IC50 value >= 10000 : inactive
3. IC50 value between 1000 and 10000 : intermediate

(NOTE: IC50 value is same as standard value that is given in the databases)

We then made a data frame consisting of only relevant data features and bioactivity class of each molecules. This data frame was saved as the pre-processed data into a .csv file.

We made the descriptors for the molecules from the SMILES notations of the molecules using rdkit.

We essentially made Lipinski Descriptors (i.e. Molecular Weight, ALogP, Number of H acceptors and donors).

**Lipinski’s Rule :**

Christopher Lipinski, a scientist at Pfizer, came up with a set of rule-of-thumb for evaluating the **drug likeness** of compounds. Such drug likeness is based on the Absorption, Distribution, Metabolism and Excretion (ADME) that is also known as the pharmacokinetic profile. Lipinski analysed all orally active FDA-approved drugs in the formulation of what is to be known as the **Rule-of-Five** or **Lipinski's Rule**.

The Lipinski's Rule stated the following:

* Molecular weight < 500 Dalton
* Octanol-water partition coefficient (LogP) < 5
* Hydrogen bond donors < 5
* Hydrogen bond acceptors < 10

Since these are the molecular descriptors which determine the drug-likeness of compounds, we took them as our descriptors for our project. We took PSA as another descriptor as it was mentioned as an example descriptor by Parthiban Sir.

Total Molecular Descriptors = 5

Also, to allow the data to be uniformly distributed, we normalized and converted IC50 to pIC50 (the reasons have been explained in the project notebook). We saved the data frames having Data + Descriptors and Data only, as two .csv files.

We decided to process the data through three ML models: Multiple Linear Regression(MLR), Random Forest Model and SVM Classification Model. This is the summary of the procedure and approach we had towards the project.

**Observations:**

We observed that the R2 score for MLR as around 0.30, the accuracy score for Random Forest as around 85% and the accuracy with SVM Classification around 70%.

**Conclusions:**

We, collectively, learned about how data is retrieved and processed for real-life applications and also how data descriptors are found out. We also understood about the importance of model selection and the various options and types of models we can apply. The suitability of ML models was also understood in a basic way.

**Acknowledgements:**

A huge thank you to Data Professor's Bioinformatics YouTube tutorial series. Link:https://www.youtube.com/playlist?list=PLtqF5YXg7GLlQJUv9XJ3RWdd5VYGwBHrP

This series helped us in acquisition and preparation of data used in this project.

**Team Members:**

Sreevathsa Golla [Dept. of EES, IISER Bhopal, India],

Desaraju Harsha Vardhan [Dept. of EECS, IISER Bhopal, India],

RVRS Mithilesh [Dept. of Mathematics, IISER Bhopal, India]

Thank you having a look at our work!