**PROJECT 1: DRUG DISCOVERY**

PROBLEM STATEMENT-

The project was based upon predicting the IC50 value(the concentration of a drug that is required for 50% inhibition in vitro) for various drugs manually curated by chEMBL database to fight SARS CoV-2.

APPROACH-

We downloaded the dataset from the official website of chEMBL database and performed some initial cleaning activities on it to get the chembl.csv dataset.

Over this dataset we performed many layers of pre-processing to get a final\_draft.csv database that contains the feature matrix (containing the descriptor information) and the prediction matric (containing the pIC50 values of the molecules).

After pre-processing of the dataset, we than finally applied two of the ML models namely- Random Forest Regression and Simple Linear Regression to predict the right molecule with potential drug capability against the virus.

OBSERVATION-

1. Descriptors are nothing but the chemical properties of the molecules. And the rdkit package does this task of assigning each molecule characterised by their “Smiles” notation certain physical, chemical properties in terms of numerical values.
2. There is a set rule-of-thumb for evaluating the **druglikeness** of a compound.  Such druglikeness is based on the Absorption, Distribution, Metabolism and Excretion (ADME) that is also known as the pharmacokinetic profile. Lipinski analyzed all orally active FDA-approved drugs in the formulation of what is to be known as the **Rule-of-Five** or **Lipinski's Rule**.
3. To allow **IC50** data to be more uniformly distributed, we can convert **IC50** to the negative logarithmic scale which is **-log10(IC50) or pIC50 data**.
4. Taking a look at pIC50 values, the **actives** and **inactives** displayed **statistically significant difference**, which is to be expected since threshold values (IC50 < 1,000 nM = Actives while IC50 > 10,000 nM = Inactives, corresponding to pIC50 > 6 = Actives and pIC50 < 5 = Inactives) were used to define actives and inactives.
5. Of the 4 Lipinski's descriptors (MW, LogP, NumHDonors and NumHAcceptors), only LogP exhibited **no difference** between the **actives** and **inactives** while the other 3 descriptors (MW, NumHDonors and NumHAcceptors) shows **statistically significant difference** between **actives** and **inactives**.

CONCLUSION-

1. Pre-processing is perhaps the most challenging portion of Machine Learning. Modelling is just the final step of the whole process.
2. Mordred-Descriptor can calculate an entire molecular descriptor twice as fast as PaDEL-Descriptor. As a result, in our model we have used Mordred instead of PaDEL calculator.

Submitted By-

Sarthak Mishra

Kasi V

Abee Nelson