Report

By using the beautifulSoup library of python I have scrapped the data for different websites like Drugbank,and some research papers. I have also taken data directly from chembl_webresource_client library of python, some data from Pubchempy and other data from clinical trials database (tdcommons.ai).
I have combined all the data in two classes as positive dataset – the drugs which are in their phase 3 or phase 4 and marked them as positive and then for rejected drugs in clinical phases mark them as negative.

After getting the drugs ids and name I have converted the drug name to its Chembl or Pubchem id after converting it into smiles I have then found their descriptors.

For Descriptors, I have used Mordred, Rdkit libraries, Mordred is used to generate 2D descriptors and Rdkit is used to generate some 2D and Rdkit's 3D descriptors.

After doing this procedure we got drugs with their descriptors value and made a csv file('2d_3D_descriptors.csv') and then we segregated positive and negative descriptors and put them in their respective files(positive_dataset.csv and neagtive_dataset.csv).
Then we have gathered the efficacy value from Chembl_Webresource_Client, and appended it with the respective drug descriptors in 2D_3D_descriptors.csv file.

Efficacy Model:
Input Features: descriptors
Label: efficacy values

1. Preprocessing techniques:

- Normalization
- Robust Scaler
- Spearman Correlation
- SMOTE synthetic dataset generator

2. Models used:
   - K nearest neighbor
   - RandomForestClassifier
   - MLP Classifier
   - SVC
   - Decision Tree
3. After this we performed stacking ensemble method which was giving the best results, results are:
   - Accuracy testing data = 81.9%
   - Matthews Correlation = 0.64
   - Sensitivity: 0.81
   - Specificity:0.82
   - Precision: 0.81
   - F1 score: 0.81

PIC50 Regression Model:
Input Features: descriptors
Label: Pic50 values

1. Preprocessing Technique:
   - Spearman's Correlation
   - 5 fold cross validation for XGBRegressor,RandomForestRegressor and DecisionTreeRegressor and GradientBoostingRegressor.
2. Results from the cross validation of RandomForestRegressor were not dispersed, that's why we chose RandomForestRegressor.
3. Results:
   - Testing R2 score: 93.8%

Receptors Regression Model:
Input Features: descriptors
Label: Active/Inactive(0/1)

1. Preprocessing Technique:
    - Spearman's Correlation
    - Robust Scaler
    - ENN synthetic data generator
2. Models used:
    - XGBClassifier
    - RandomForestClassifier
3. RandomForestRegressor was giving the best results among the two as data we get from Tox21 was much lower, then we trained on the same regressor model.

4. Average R2 score for training sample: 87.6%.

Then, we implemented the model in the backend using the "joblib" library and predicted the values on the models we used.