COVID-19 Drug Discovery (2019201049)

Aim
The objective is to prepare a machine learning model that can be used to propose potential novel effective drugs to fight SARS-CoV-2, the virus responsible for COVID-19.

Why SVR over linear regression?

In earlier assignment of house prediction we tried linear regression but we come to know by observation that SVR gives better result if tuned hyperparameter than simple linear regression though ohter regressor also can give better result but due to eaiser implementaion we choose SVR

SVR is a powerful algorithm that allows us to choose how tolerant we are of errors, both through an acceptable error margin(ϵ) and through tuning our tolerance of falling outside that acceptable error rate.

Major challege choosing hyperameter?

To get good hyperparameter I tried different values of C,gamma ,kernel and epsilon but still choosing to get manually optimized value is tough task
 then I choose Optunity and GridCVSearch to find good hyperparameter

after running C=[1-30], gamma=[0.001,0.0007,0.0006,0.1,1,0.0005], epsilon=[1-2] & kernel[‘rbf’,’poly’,’sigmoid’]

GridCVSearch

selected C=25 , gamma=0.0007 , kernel=rbf , epsilon=’1.3’

similar we get equal rmseusing optunity on

c=18 , gamma=’0.001’,epsilon=1.3,kernel=’rbf’

after this still accuracy is not upto the mark due to precision setted by default to 8 by libreoffice

over above hyperparameter we get lowest Rmse

Data preprocessing

All things about data preprocessing mentioned in references given by TA
we decide to use Mol2Vec bcause on kaggle notebook it is clearly mentioned that it will give better result