Rishabh Bhonsle (17217) | Prithvi Poddar (17191)

PROJECT ASSIGNMENT I

ML MODEL TO PREDICT BIOACTIVITY VALUES OF INHIBITORS OF CORONA VIRUS

INDEX

1. AIM

The aim of this project is to predict the following with sufficient accuracy:

i. **Bioactivity** of a new drug compounds in terms of its inhibition percentage.
ii. **QED Weighted** (quantitative estimation of drug-likeness) as drug likeness is a key consideration when selecting compounds during the early stages of drug discovery.

2. OVERVIEW

The Coronaviridae are a family of positive single stranded encapsulated viruses. They typically cause mild respiratory diseases, but infections with the ß-coronavirus SARS-CoV, MERS and SARS-CoV-2 can lead to acute respiratory diseases and high mortality, particularly in individuals with underlying health conditions. Multiple interventional clinical trials have been initiated in the search for effective pharmacological treatments against SARS-CoV-2 infection and the related disease Covid-19. Bioinformatics analyses have proposed repurposed drugs based on the interactome between viral encoded proteins and host-cell pathways. In the absence of safe and effective vaccines against SARS-CoV-2, repurposing of existing drugs represents a first pragmatic strategy for the treatment of Covid-19 patients. We thus identify potential inhibitors of SARS-CoV-2 in-vitro cellular toxicity in human (Caco-2) cells by predicting its bioactivity using ML after training the data on a large scale drug repurposing collection (ref 3). Bioactivity describes the characteristic of an implant material to interact with or initiate a specific reaction of living tissue upon exposure. The biochemical systems encountered by a drug molecule (implant material) are extremely complex. The factors affecting the bioactivity (ref 4) may be divided into three categories:

I. Physicochemical properties such as solubility, partition coefficients, and ionization.
II. Chemical structure parameters such as resonance, inductive effect, oxidation potentials, types of bonding, and isosterism.
III. Spatial considerations such as molecular dimensions, interatomic distances, and stereochemistry.

3. DATA

To identify possible candidates for progression towards clinical studies against SARS-CoV-2, the authors of the paper (ref 1) screened a well-defined collection of compounds. We obtained this data via the data base - ChEMBL_27 SARS-CoV-2 release under the title Identification of inhibitors of SARS-CoV-2 in-vitro cellular toxicity in human (Caco-2) cells using a large scale drug repurposing collection (ref 3). We obtain this data in two parts as follows:

I. **inhibition_1.csv** – Data in this data set is used for stating the bioactivity measure which are trained against the corresponding attributes from dataset_1.csv to predict bioactivity values in terms of inhibition percentage for new compounds.

II. **dataset_1.csv** – Data in this data set is used for stating factors that affect towards bioactivities as stated previously. The factors selected from the given set of attributes are Molecular Weight, AlogP, PSA, HBA, HBD, CX ApKa, CX BpKa & CX LogD which we shall use as our descriptors. Here we also extract the QED Weighted which are trained against the attributes to predict QED Weighted values for new compounds.

Here is a rough overview of our descriptors and targets (via Wiki and ref_5). Here we see the reason as to why the following descriptors are chosen by comparing their definitions to the factors that affect bioactivity of a compound.

A. DESCRIPTORS

- *Molecular Weight* – Measure of the mass of a given molecule
- *AlogP* – Measure of lipophilicity which is a key physicochemical property that plays a crucial role in determining ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties and the overall suitability of drug candidates.
- *PSA* – Measure of the polar surface area (PSA) of a molecule is defined as the surface sum over all polar atoms or molecules, primarily oxygen and nitrogen, also including their attached hydrogen atoms.
- *HBA* – Hydrogen Bond acceptor atoms
- *HBD* – Hydrogen Bond donor atoms
- *CX ApKa, CX BpKa* – Measure of pH
- *Log D* – Measure of the distribution coefficient. It is the ratio of the sum of the concentrations of all forms of the compound (ionized plus un-ionized) in each of the two phases, one essentially always aqueous; as such, it depends on the pH of the aqueous phase, and log D = log P for non-ionizable compounds at any pH.

B. TARGETS

- *QED Weighted* – Measure of drug likeness and is a key consideration when selecting compounds during the early stages of drug discovery.
- *Standard Value* – Measure of bioactivity of a drug compounds in terms of its inhibition percentage.

We further use Pandas Profiling to see a brief visualization of our data against the following parameters:

- **Overview** – Summary of our dataset
- **Variables** – A brief overview of our descriptors and targets under the heading
- Interactions – An interactive graph to compare trends of targets vs descriptors individually
- **Correlations** – The Pearson's correlation coefficient (r) is a measure of linear correlation between two variables. Its value lies between -1 and +1, -1 indicating

total negative linear correlation, 0 indicating no linear correlation and 1 indicating total positive linear correlation.

4. APPROACH

As stated in the paper, it will be key to determine whether any clinical-stage compounds or related molecules could safely achieve active concentrations at the targeted site, human lung epithelia.

Compounds in our data sets are screened for their inhibition of viral induced cytotoxicity using the human epithelial colorectal adenocarcinoma cell line Caco-2 and a SARS-CoV-2 isolate obtained from an individual originally exposed to the virus in the Wuhan region of China.

We thus have to identify inhibitors of SARS-CoV-2 in-vitro cellular toxicity in human (Caco-2) cells using a large scale drug repurposing collection for progression towards clinical studies against SARS-CoV-2 by predicting Percent Inhibition and QED Weighted values of compounds.

5. STRATEGY

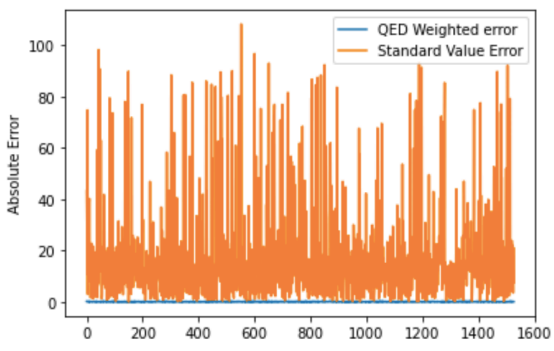
We implement ML techniques on our data in two parts:

A. LINEAR REGRESSION

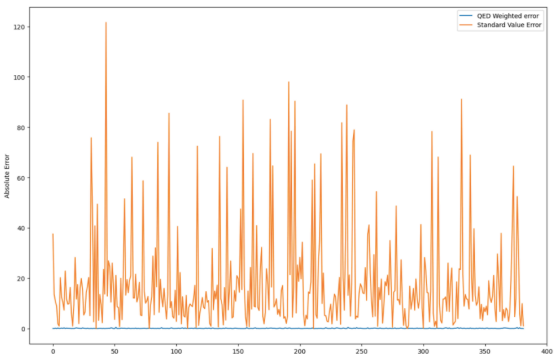
We apply linear regression on the data set with an 80:20 split for training data and the test data respectively. The regularization scores obtained are as follows:

- QED Weighted regularization score: 0.5361050635235873
- Standard Value regularization score: 0.03089728635307798

We can see that QED Weighted is more linearly related to the features than Standard Values. Linear Regression thus seems a good model for QED Weighted Predictions. We thus need to find a different model for our Standard Value data.



Absolute Error in each **training data** point



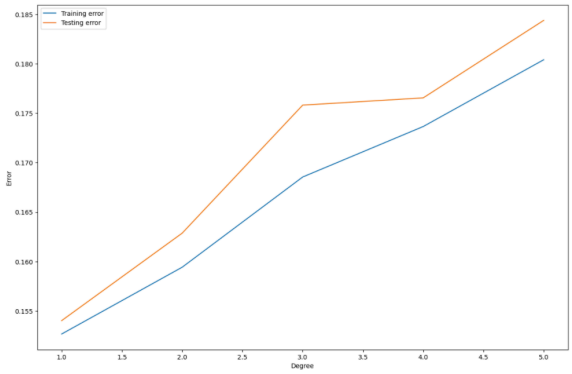
Absolute Error in each **test data** point

The Root Mean Squared Error Obtained for our data set are as follows:

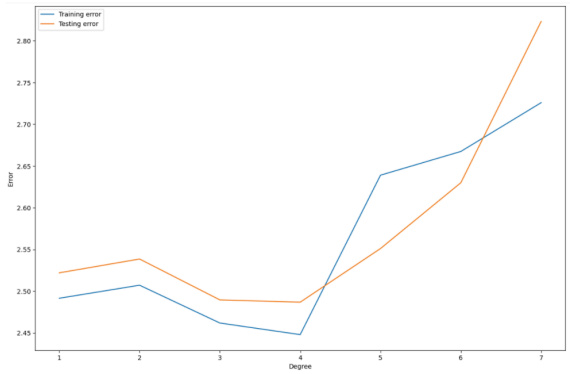
- RMSE for QED weighted in training set: 0.1457422561447954
- RMSE for Standard Values in training set: 23.646267681886698
- RMSE for QED weighted in test set set: 0.15251513968147162
- RMSE for Standard Values in test set: 26.642521574238412

B. SVM

Here we apply polynomial regression with SVM, for both the target features and try to figure out which polynomial curve fits the best.



Polynomial Regression on **QED Weighted** values



Polynomial Regression on **Standard** values

In polynomial regression on QED Weighted values, we see that degree 1 gives the least error. Hence the QED Weighted depends linearly with the input parameters. This was confirmed in our linear regression model above as well.

In polynomial regression on Standard values, we see that degree 4 gives the least error. Although, the standard values are not in any polynomial relation with any of the parameters, for regression purpose, a degree 4 polynomial is the best fit.

The Root Mean Squared Error Obtained for our data set are as follows:

- RMSE for QED weighted in training set: 0.15267284356929767
- RMSE for Standard Values in training set: 2.4478688987176187
- RMSE for QED weighted in test set: 0.15402692332804294
- RMSE for Standard Values in test set: 2.486751415809695

6. CONCLUSION

We observe that the RMSE has significantly reduced for the polynomial regression with degree 4, as compared to linear regression. Whereas RMSE value is optimum for linear regression.

Thus, we predict Bioactivity values via a degree 4 polynomial regression using SVM whereas QED Weighted values are predicted using linear regression. QED Weighted was

expected to be a linear fit as the QED values are in fact a measure of the descriptors we considered in some sense.

7. REFERENCES

- **ref_1** - https://www.researchsquare.com/article/rs-23951/v1
- **ref_2** - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3524573/
- **ref_3** - https://www.ebi.ac.uk/chembl/document_report_card/CHEMBL4303101/
- **ref_4** - https://www.drugtimes.org/how-drugs-act/factors-affecting-bioactivity.html
- **ref_5** - https://www.researchgate.net/figure/Values-MW-clogP-HBA-HBD-PSA-logBB-and-logP-e-a-for-ML-and-1_tbl1_259626002