Covid19 Drug Discovery

Devbrat Anuragi
17078

The objective is to build conventional machine learning models such as Random Forest, Linear Regression etc (and NOT neural networks) to predict Bioactivity values (y=bioactivity).

Covid19 Drug Discovery

Devbrat Anuragi
17078

Brief

This project really focuses on the data collection part. In the following section I have explained how get to know how to collect relevant data for you model from Chembl database using Chembl API. How to use Lipinsk's descriptors, PaDEL . In This project I have used the Libraries like rdkit, chembl_webresource_client extensively.

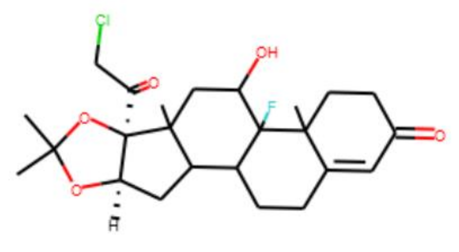
Data Collection

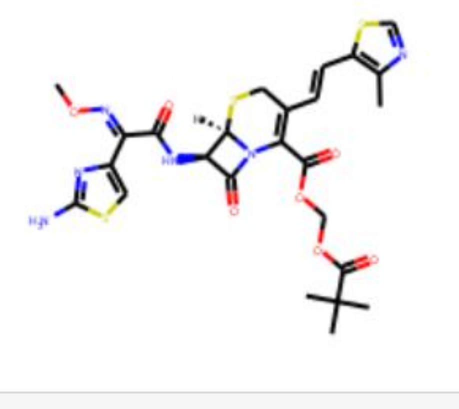
Downloaded the this cvs file from chembl database. This file contain 6900 molecules and their intrinsic properties like 'Name','Synonyms','Type','Max Phase','#RO5 Violations','#Rotatable Bonds','CX ApKa','CX BpKa','Structure Type','Inorganic Flag','#RO5 Violations (Lipinski)','Molecular Weight (Monoisotopic)','Molecular Species','Molecular Formula','Passes Ro3','Molecular Weight','Targets','Bioactivities','QED Weighted','CX LogP','CX LogD','Aromatic Rings','Heavy Atoms','HBA Lipinski','HBD Lipinski'.

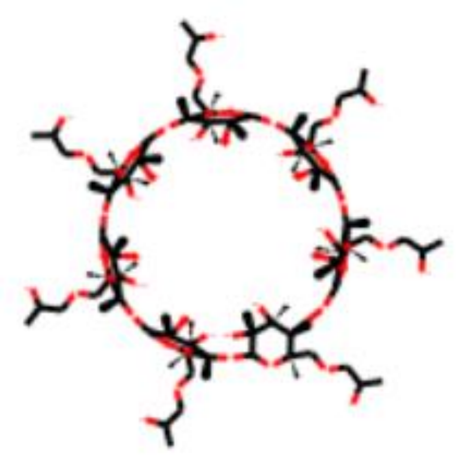
But not all these properties are of our use. Only 'AlogP','HBD', 'HBA', 'PSA', 'MW' are for our use. I will discuss them in the following text shortly.

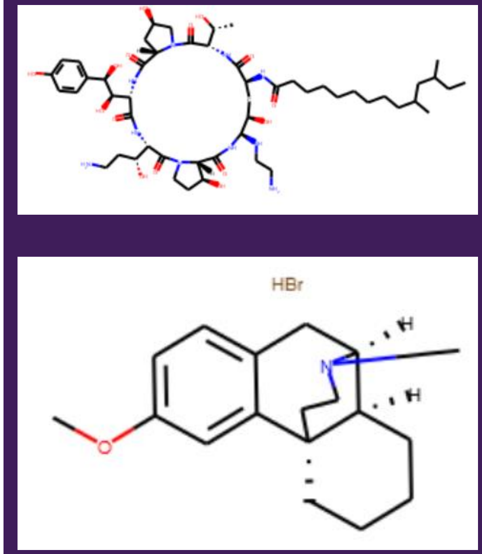
Our target variable was "bioactivity". And there is the column of this in the given data set also but Actually this is not the bioactivity value, These are just the frequency of the number of bioactivity a molecule has. So our main aim to fetch the bioactivity values of each from the chemlb database where standard type is ic50 , standard unit is nM. Earlier we were asked to do this work manually, but thanks to chembl API, using this API I managed to fetch the bioactivity value from the Database. Out of 6900 molecules I have got 2000 molecule's bioactivity in the desired units.

Using rdkit for getting Molecular Structure from Smiles



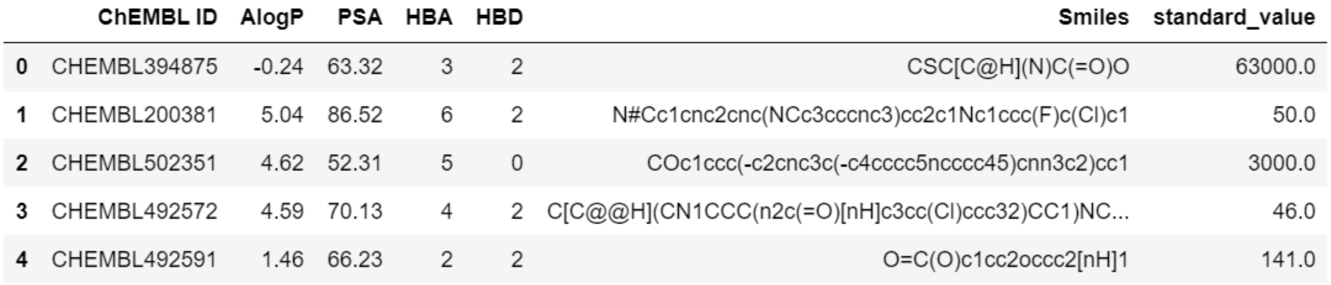






The above figure are some of the molecule's structure from the csv file.

Data set formed is as follow:



	ChEMBL ID	AlogP	PSA	HBA	HBD	Smiles	standard_value
0	CHEMBL394875	-0.24	63.32	3	2	CSC[C@H](N)C(=O)O	63000.0
1	CHEMBL200381	5.04	86.52	6	2	N#Cc1cnc2cnc(NCc3cccnc3)cc2c1Nc1ccc(F)c(Cl)c1	50.0
2	CHEMBL502351	4.62	52.31	5	0	COc1ccc(-c2cnc3c(-c4cccc5ncccc45)cnn3c2)cc1	3000.0
3	CHEMBL492572	4.59	70.13	4	2	C[C@@H](CN1CCC(n2c(=O)[nH]c3cc(Cl)ccc32)CC1)NC...	46.0
4	CHEMBL492591	1.46	66.23	2	2	O=C(O)c1cc2occc2[nH]1	141.0

Data Preprocessing

Now form the above dataset I check if any column contain a cell which Is empty or has value nan or NONE, If there is I have dropped the corresponding row.

One this is done I have used Lipinski's Rule stated as follow:

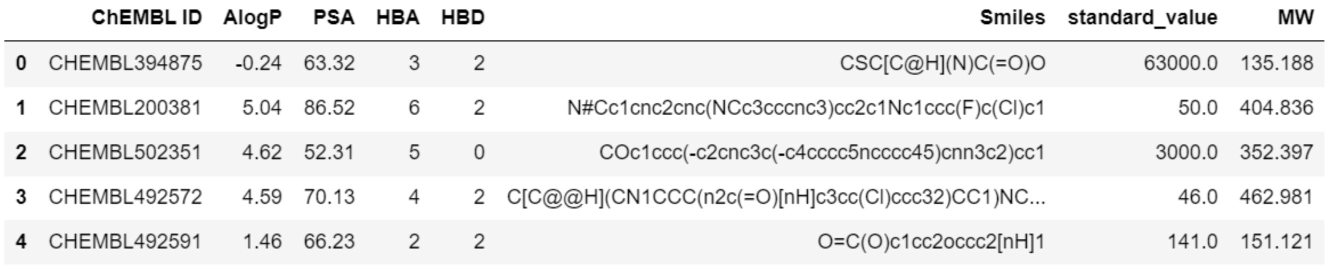
Christopher Lipinski, a scientist at Pfizer, came up with a set of rule-of-thumb for evaluating the **druglikeness** of compounds. Such druglikeness is based on the Absorption, Distribution, Metabolism and Excretion (ADME) that is also known as the pharmacokinetic profile. Lipinski analyzed all orally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

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

Using the smiles notation I have calculated the Molecular weight of the compound, If the weight is greater than 500 Dalton I have removed the corresponding row from the dataset.

After this step data set looks like this:



	ChEMBL ID	AlogP	PSA	HBA	HBD	Smiles	standard_value	MW
0	CHEMBL394875	-0.24	63.32	3	2	CSC[C@H](N)C(=O)O	63000.0	135.188
1	CHEMBL200381	5.04	86.52	6	2	N#Cc1cnc2cnc(NCc3cccnc3)cc2c1Nc1ccc(F)c(Cl)c1	50.0	404.836
2	CHEMBL502351	4.62	52.31	5	0	COc1ccc(-c2cnc3c(-c4cccc5ncccc45)cnn3c2)cc1	3000.0	352.397
3	CHEMBL492572	4.59	70.13	4	2	C[C@@H](CN1CCC(n2c(=O)[nH]c3cc(Cl)ccc32)CC1)NC...	46.0	462.981
4	CHEMBL492591	1.46	66.23	2	2	O=C(O)c1cc2occc2[nH]1	141.0	151.121

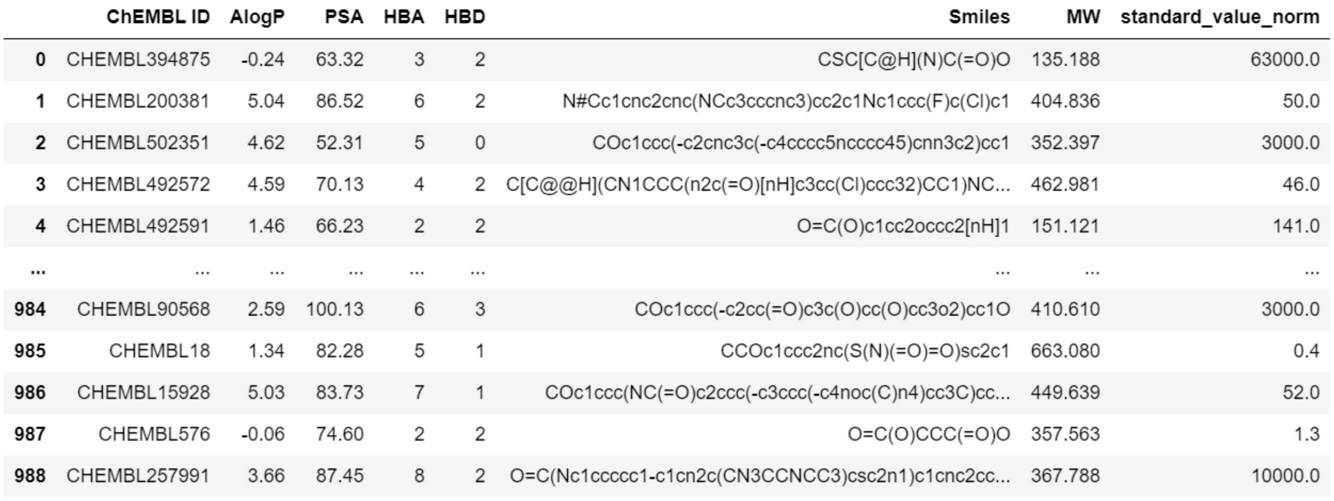
Notice that a new column has been added to the data set.

Now another step for data preprocessing is to convert the IC50 value to the pIC50 value. To allow **IC50** data to be more uniformly distributed, we will convert **IC50** to the negative logarithmic scale which is essentially **-log10(IC50)**.

This custom function pIC50() will accept a DataFrame as input and will:

- Take the IC50 values from the standard_value column and converts it from nM to M by multiplying the value by 10−9−9
- Take the molar value and apply -log10
- Delete the standard_value column and create a new pIC50 column

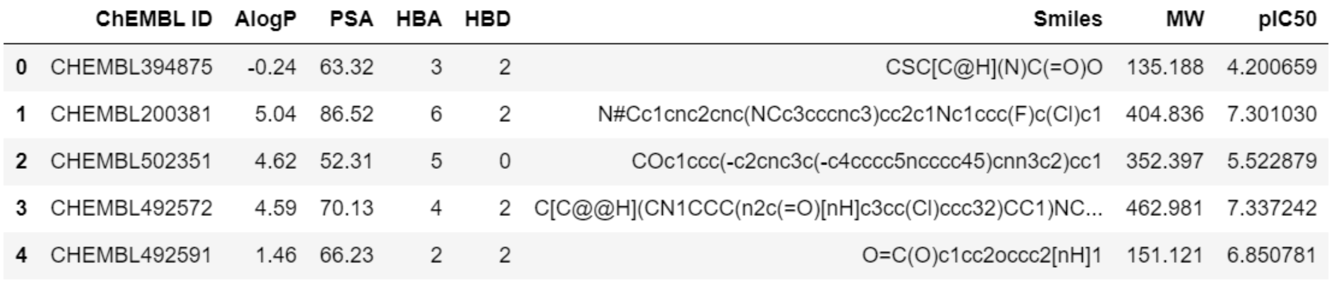
We will first apply the norm_value() functions so that the values in the standard_value column is normalized.



	ChEMBL ID	AlogP	PSA	HBA	HBD	Smiles	MW	standard_value_norm
0	CHEMBL394875	-0.24	63.32	3	2	CSC[C@H](N)C(=O)O	135.188	63000.0
1	CHEMBL200381	5.04	86.52	6	2	N#Cc1cnc2cnc(NCc3cccnc3)cc2c1Nc1ccc(F)c(Cl)c1	404.836	50.0
2	CHEMBL502351	4.62	52.31	5	0	COc1ccc(-c2cnc3c(-c4cccc5ncccc45)cnn3c2)cc1	352.397	3000.0
3	CHEMBL492572	4.59	70.13	4	2	C[C@@H](CN1CCC(n2c(=O)[nH]c3cc(Cl)ccc32)CC1)NC...	462.981	46.0
4	CHEMBL492591	1.46	66.23	2	2	O=C(O)c1cc2occc2[nH]1	151.121	141.0
...	...	...	...	...	...	...	...	...
984	CHEMBL90568	2.59	100.13	6	3	COc1ccc(-c2cc(=O)c3c(O)cc(O)cc3o2)cc1O	410.610	3000.0
985	CHEMBL18	1.34	82.28	5	1	CCOc1ccc2nc(S(N)(=O)=O)sc2c1	663.080	0.4
986	CHEMBL15928	5.03	83.73	7	1	COc1ccc(NC(=O)c2ccc(-c3ccc(-c4noc(C)n4)cc3C)cc...	449.639	52.0
987	CHEMBL576	-0.06	74.60	2	2	O=C(O)CCC(=O)O	357.563	1.3
988	CHEMBL257991	3.66	87.45	8	2	O=C(Nc1ccccc1-c1cn2c(CN3CCNCC3)csc2n1)c1cnc2cc...	367.788	10000.0

This contain the normalized value

Now calculating the the pIC50 value from the above table

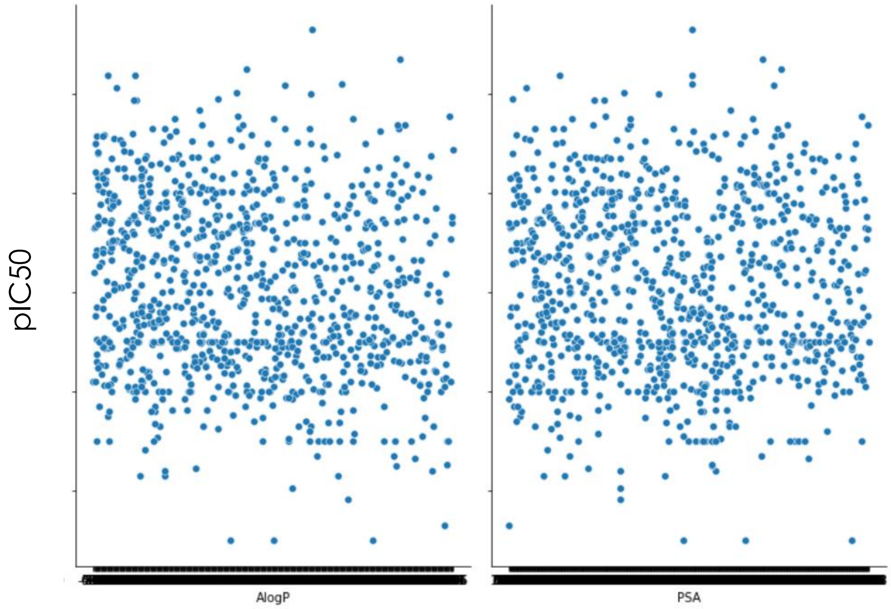


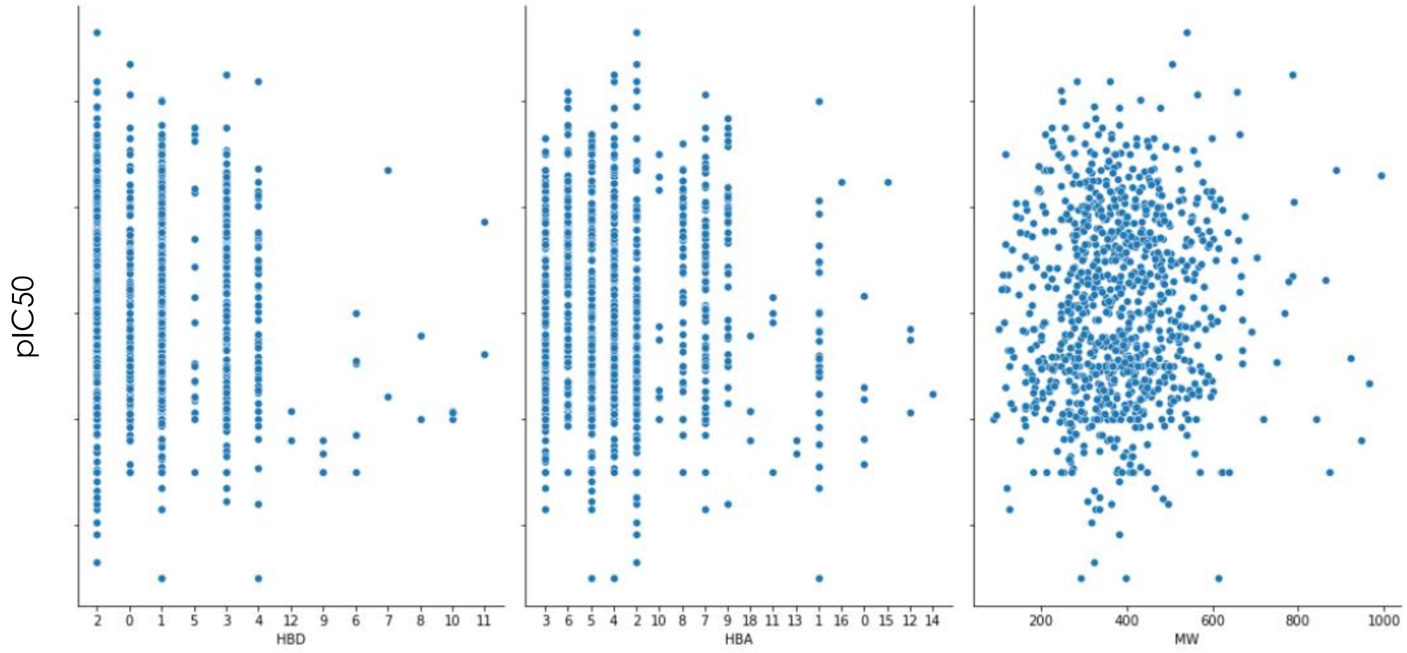
	ChEMBL ID	AlogP	PSA	HBA	HBD	Smiles	MW	pIC50
0	CHEMBL394875	-0.24	63.32	3	2	CSC[C@H](N)C(=O)O	135.188	4.200659
1	CHEMBL200381	5.04	86.52	6	2	N#Cc1cnc2cnc(NCc3cccnc3)cc2c1Nc1ccc(F)c(Cl)c1	404.836	7.301030
2	CHEMBL502351	4.62	52.31	5	0	COc1ccc(-c2cnc3c(-c4cccc5ncccc45)cnn3c2)cc1	352.397	5.522879
3	CHEMBL492572	4.59	70.13	4	2	C[C@@H](CN1CCC(n2c(=O)[nH]c3cc(Cl)ccc32)CC1)NC...	462.981	7.337242
4	CHEMBL492591	1.46	66.23	2	2	O=C(O)c1cc2occc2[nH]1	151.121	6.850781

Notice that the IC50 colum is replaced by the pIC50 column.

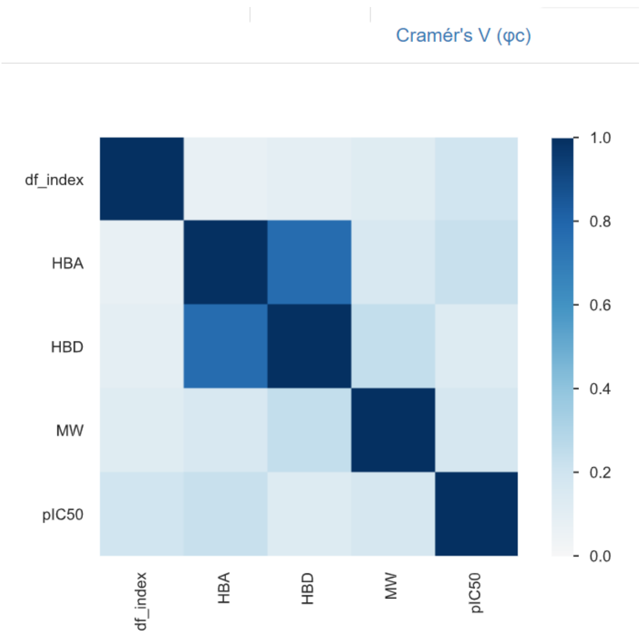
Data Visualisation

Following are the graph for the each feature vs. pIC50





Also folloeing is the Phik Correlation



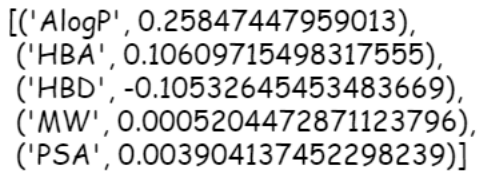
Please check The Data Report.html file I have submiited for the detailed Data report.

Form all the above graph It is too easy to see that there is no clear linear, polynomial, relation between pIC50 and other features.

Different Models

Multiple Linear Regression

I have tried to use Multiple linear Regression on this data set to estimate the bioactivities. In this part I have estimated the coefficient for each features as follows



```
[('AlogP', 0.25847447959013),
 ('HBA', 0.10609715498317555),
 ('HBD', -0.10532645453483669),
 ('MW', 0.0005204472871123796),
 ('PSA', 0.003904137452298239)]
```

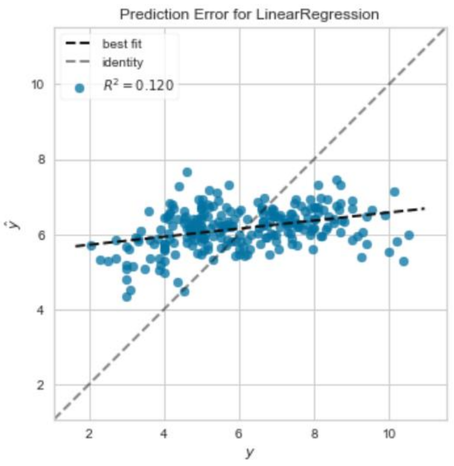
This means

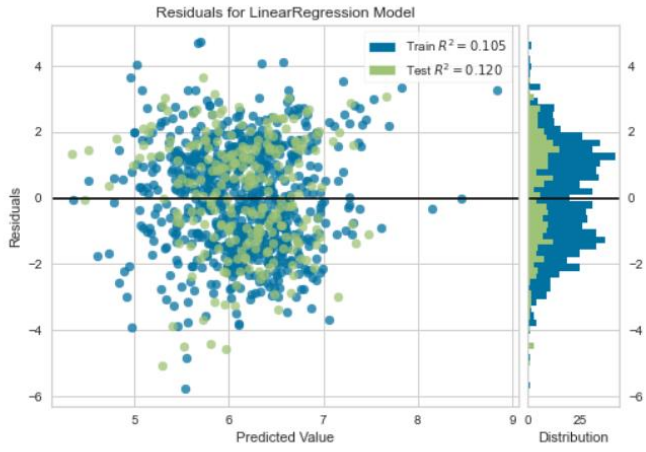
bioactivity = (0.25 * AlogP) + (0.106 * HBA) – (0.105 * HBD) + (0.00052 * MW) + (0.0039 * PSA)

But This approach is not much efficient on the train Data set itself, which is somewhat same for our intuition as we saw in the above graphs.

Linear Regression

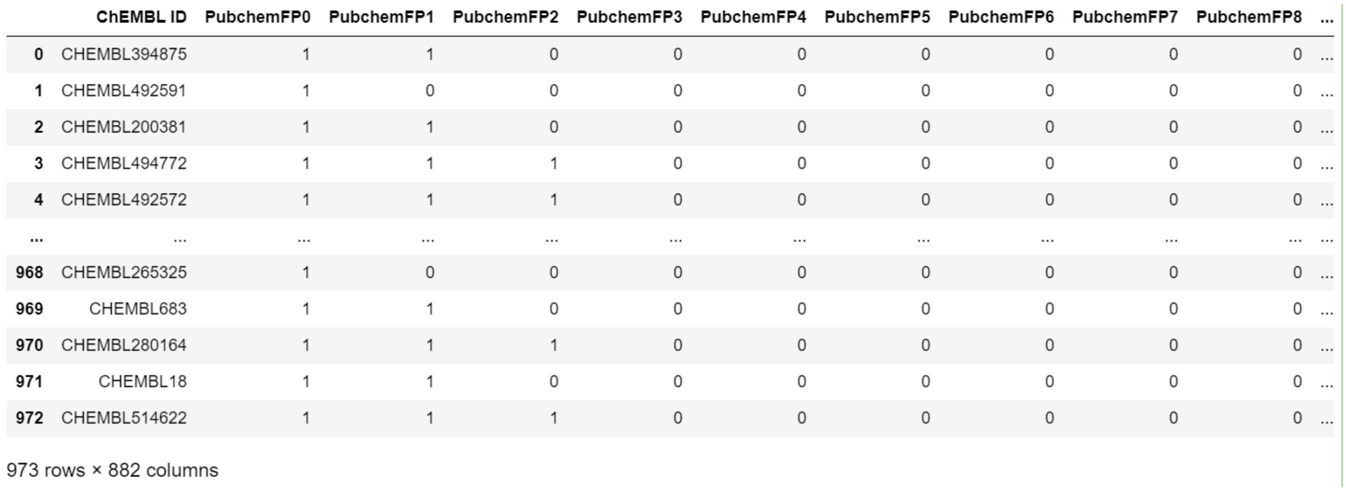
This time I tried to split the dataset into train and test data set and Tried to fit linearly. But even In this approach the r2 score on the test data set was 0.12 which confirms that a linear model can not fit this data set. Following are the some graph related to these approach





PaDEL Discriptors

I have calculated the PaDEL descriptor using padel.sh and a zip folder. In order to get the PaDEL Descriptors first you have to create .smi file containing Smile and CHEMBL_id In this order only. Onceyou have the this file you need to have padel.sh and the unzipped padel folder in you working directory, Then Run the padel.sh terminal. This will take some time, one completed you will find a 'descriptors_output.csv' in your working directory, It will look something like this.



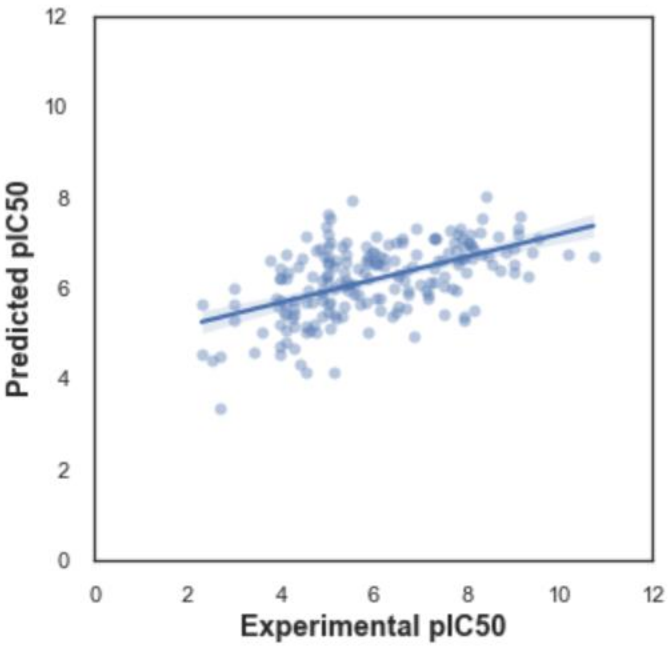
	ChEMBL ID	PubchemFP0	PubchemFP1	PubchemFP2	PubchemFP3	PubchemFP4	PubchemFP5	PubchemFP6	PubchemFP7	PubchemFP8	...
0	CHEMBL394875	1	1	0	0	0	0	0	0	0	...
1	CHEMBL492591	1	0	0	0	0	0	0	0	0	...
2	CHEMBL200381	1	1	0	0	0	0	0	0	0	...
3	CHEMBL494772	1	1	1	0	0	0	0	0	0	...
4	CHEMBL492572	1	1	1	0	0	0	0	0	0	...
...	...	...	...	...	...	...	...	...	...	...	...
968	CHEMBL265325	1	0	0	0	0	0	0	0	0	...
969	CHEMBL683	1	1	0	0	0	0	0	0	0	...
970	CHEMBL280164	1	1	1	0	0	0	0	0	0	...
971	CHEMBL18	1	1	0	0	0	0	0	0	0	...
972	CHEMBL514622	1	1	1	0	0	0	0	0	0	...

973 rows × 882 columns

Now I have splited it into train data and test data.

RandomForestRegressor

The test score is now 0.62 which is fairly good.



Conclusion:

Random Forest Regressor work fine in comparisons to the other methods. But as there is no clear relationship between descriptors and bioactivity, The conventional Methods are not enough to predict bioactivity with high accuracy, we need something more powerful than conventional machine learning method like we need CNN, neural network or GANs to predicts these value more accurately