**Machine Learning Project in Drug Discovery**

|  |  |  |
| --- | --- | --- |
| Ayush Mathur (18065) |  | Instructor name: Dr. Parthiban Srinivasan |
| Rohit Jain (18...) |  |  |
| Vatsalya Sharan (18...) |  |  |

# **Introduction**

A few decades back it was impossible to even imagine that one day computers would be revolutionizing medicine. The disbelief was not unfounded as one could not fathom the idea of a computer performing expensive lab experiments that needed human expertise. However, the advancement of SAR and QSAR models has changed the whole paradigm, birthing the field of bioinformatics. The idea is simple: molecules with similar structures have similar properties. Thus, a regression method is employed between these chemical properties and the targeted biological activity values (which include both the therapeutic effect and side effects). Mathematically,

Activity = f(physiochemical properties and/or structural properties) + error.

While these models may not tell us what our desired drug is, they provide a direction and help us write-off compounds from potential drugs’ list. While at a small scale this can feel like a ruse, but for pharmaceuticals researching thousands of compounds, this can save millions.

Given today’s scenario, the pandemic struck world we are living in, these models find their relevance in everyone’s life.

## **Target Choice**

Our group worked on almost all targets with tags “coronavirus” or “coronavirus-2” or “Covid”. Other than the following targets, data was insufficient:

1. “CHEMBL5118” : Replicase polyprotein 1ab
2. “CHEMBL3927” : SARS coronavirus 3C-like proteinase
3. “CHEMBL4303835” : SARS-CoV-2

We did our analysis and modelling on all three of them. For sake of this report, however, we’ll be reporting only the results of “Replicase polyprotein 1ab”.

## **Data**

We extracted data on “CHEMBL4303835” using the chembl\_web\_client library on python. The inspiration for this came from a video on YouTube by “Dataprofessor”.

**Aim**

The aim of this project is to determine a suitable conventional machine learning model using the ChEMBL dataset to predict bioactivity values of different compounds on a SARS-coronavirus replicase polyprotein 1ab target using the following descriptors and analyze the results.

1. Mol Wt
2. AlogP
3. PSA
4. HBD Lipinski
5. HBA Lipinski
6. aromatic rings
7. num of rotatable bonds
8. QED

# Data Preprocessing

From the source itself, we filtered “standard\_values” for IC50 standard, which are our desired bioactivity values in this dataset. Initially the data had a lot of useless columns, but we extracted only the “canonical\_smiles” and bioactivity values for each compound. We cleaned the data in two stages from here. First, we removed any compound for which “standard\_values” was not given. Then we used rdkit to check whether the SMILES are valid and represent an actual molecule. Once this was done, we used rdkit functions to generate the desired chemical descriptors. Once the whole exercise was done, we had data of 214 compounds.

We saved this data into a csv file and normalized it using the StandardScaler() in sklearn package to use in our modelling exercise.

# Building up the model

As we learnt from practice, the choice of model was very consequential here. For our project, we tried three different conventional machine leaning models:

1. Linear Regression
2. Random Forest Regression, and
3. Support Vector Machine

Once we had constructed these models, we calculated the R2 scores, which we used a s the parameter of performance of these models.

## Linear Regression

We started our analysis with the linear regression model. For this purpose, we converted our data into the array format and divided it into train and test sets using the train\_test\_split() function in sklearn. From here, we ran the regressions multiple time, using different random states in the said function, along with trying different test set sizes.

## Random Forest Regression

After our analysis of different targets, we believed that random forest would give best results. Therefore, we repeated the same process we had used in linear regression to run our tests. Since more estimators meant more computational time, we used a standard of 100 estimators while looping through the random states to quickly analyze its effects on the results. Our test size was fixed at 20%.

## Support Vector Machine Regression

Our experience on the coronavirus related targets had familiarized with the fact that this was the worst performing model of the three. Anyways, we looped through the different random states and different SVM kernels available in sci-kit learn.

# Observations

Building up these models, we observed the following:

1. Bioactivities data had a large range and had some very big outliers.
2. The Linear Regression model, while quick and useful in finding useful descriptors from the list of several we had initial access to, produces the worst results on normalized data.
3. The Random Forest Regression method yields the best R2 scores, with SVM with polynomial kernel being the second best.
4. However, the R2 greatly varied with choice of random states in selection of training data.
5. Exclusion of any of the descriptor had negative effects on the R2 value. The effects were especially large on linear regression results.

# Results

The results of our project can be summarized as follows:

1. Random Forest Regression method produced the largest R2 values. For a particular random state, it was 0.78
2. Support Vector Machine regression produced a maximum R2 of 0.4 using the polynomial kernel.
3. Linear regression produced a maximum R2 of 0.17.
4. The performance of these methods showed variance in performance based on the choice of random state, kernel (for svm) or estimators (in random forest)

## Further experimenting with Random Forest Regression

By the end of our project, we could clearly see that Random Forest was producing the best results. Therefore, we increased the number of estimators to 1000 while looping to see if the performance has improved. To our surprise, the model not only improved its performance in general, different random states now had a more consistent performance of 70%+ R2 . This motivated us to invest more energy in it.

# Conclusion

By the end of our project, we had realized the following:

1. Random forest regression was the was most suitable model in our scenario.
2. It produced the largest R2 values. At 1000 estimators, we reached a maximum of 0.8

Therefore, we have successfully created a conventional machine learning based model which quantitatively predicts the IC50 bioactivity values of different compounds based on their chemical and structural properties.