Problem:

I selected a protein data corresponding to SARS Coronavirus. Data contained information of different drugs that can inhibit the effect of that protein. Effectiveness of the drug can be estimated by Bio-activity value for that drug. In the information for each drug given to us we were supplemented with the Chemical Formula of each drug in **Smile** form. Our problem is to calculate different properties of Drug using its Molecular Formula in **Smile** format and applying the RdKit Descriptor function on it. Using that descriptor value we have to calculate the Bioactivity of Drug (that is effectiveness) and comparing it with actual Bioactivity value we need to train our Machine Learning Model.

Approach, Strategy and Procedure:

My whole strategy and procedure are described in steps:

1. To extract data from Chembl website.
2. Then I looked at data using pandas functions and decided what features of data can be useful in data analysis.
3. Then I decided to discard those training data which had some null values in their features, but fortunately this was not the case with Chembl data I decided to work on, so I wasn't required to execute this step.
4. Looking at data I decided to extract different descriptor values from Molecular Formula (in Smiles format), which I can use as features to train my Machine Learning Model.
5. I found correlation between each feature and discarded those features which were having very low correlation values with Standard Value of Bioactivity.
6. Also I applied filters on training data so that only Drugs with Mol. weight less than 500 and bioactivity values less than 1000000 were considered, because some bioactivity values were abnormally very high in number.
7. I normalized/rescaled the remaining features to make them compatible for Machine Learning training.
8. Then I splitted rescaled features into training dataset and testing dataset.
9. Finally I trained my Machine Learning Model with 4 different regression models which are: Linear Regression, Polynomial Regression, Tree Regression, Support Vector Regression.
10. Then I calculated predicted bioactivities values by the model and compared it with actual bioactivity values.
11. I calculated the r2 score using predicted value and actual value of bioactivity of each regression model to estimate robustness of the model.
12. For real life purposes we can select a model which will give us the best r2 score. Closer the r2 score to 1 better is the performance of our model.

Observations and Conclusions:

1. **Observation:** Number of training data was 133 which is very less for real life machine learning applications.
   Conclusion: Hence the predictions of models may not be accurate, also fitting might not be as good as it should be.
2. **Observation:** Correlation of any of the features with Bioactivity was not greater than 0.37.
   Conclusion: Because features were not highly correlated with bioactivity values they may not predict the Bioactivity values to good accuracy level.
3. **Observation:** r2 score obtained for Linear Regression, Tree Regression, Polynomial regression and SVM regression were 0.149, -2.055, -2.982 and -0.144 respectively.
   Conclusion: Because r2 score is very less for all the models it justifies observation and conclusions pointed out in point 1 and point 2, that is there is lot if mismatch between predicted and actual values of Bioactivity, hence the model I obtained is not very robust for predictions of Bioactivities.