|  |  |
| --- | --- |
|  |  |
| MLProject: Covid data analysis |  |
|  |  |
|  | Team:Anushk Naval (18046)Subhajit Pramanik (18398) |
|  | Subhojit Pal (18400)Submitted to:Dr. Parthiban Srinivasan |

### Abstract:

For this project, we were asked to use and experiment with the drug dataset available at Chembl site, and to explore how to use different machine learning algorithm in a project and find pattern in data.

We were expected gain insights and experience in data cleaning and processing using common data-mining and machine learning library.

We were expected to submit a report about the dataset and different algorithms that were used.

During the project the use of neural network or deep learning techniques were restricted.

This is the report on the data and the techniques used in the project with interpretations.

Keywords: Machine learning, Data cleaning, Regression Modeling, Bioinformatics.

### Opening and running the file:

#### Use Google Colab to run the file.

#### Program collects data directly from Chembl Database at the runtime.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | Introduction: In this project our team worked on Chembl drug data, we selected the compounds that target SARS coronavirus having type – single protein from the database.  The bioactivity of the data we are calculating potency of the drug that’s the amount required to produce an effect of given intensity lower the value the more the potency.  We calculated the descriptors and fingerprint for the compounds imported and then ran different regression models to compute the bioactivity of the compounds Methods and Data: The dataset we used could be found: <https://www.ebi.ac.uk/chembl/>  The database contains the details of compounds with their drug-like properties, bringing together chemical, bioactivity and genomic data to aid the translation of genomic information into effective drugs.  We imported the data of our interest target (SARS coronavirus).  The imported dataset has information of molecules like chembl id, canonical smiles, bio-activity values for our project we took its type to be “IC50” (make data more uniform).  Data cleaning and processing:  From the data imported any missing values for our project has been removed and only the features required like the molecule chembl id, the smiles and their bioactivity values were taken.  The data is then categorizing into three class – active (< 1000 nm), intermediate (1001nm-10000nm) and inactive (above 10000nm) based on their bioactivity values.  Then to uniformly distribute the activity values the standard IC50 values were converted to pIC50 values.  The intermediate bioactivity class was removed to easily compare between inactive and active compounds. | |  |

|  |  |  |
| --- | --- | --- |
|  | The following data frame is obtained after cleaning the data.  Molecular descriptors:  The molecule as a real object contains all the chemical information, by theoretical pathways the information encoded in the molecular structure into one or more numbers is molecular descriptors, used to establish relationship between structures and properties, biological activities and other experimental properties.  Molecular fingerprints:  Fingerprint representations of molecular structure and properties are a particularly complex forms of descriptors, encoded as binary bits “patterns” characteristic of a given molecule. |  |
|  | The graph shows the frequency distribution of the two bioactivity classes of our data.  In the program we calculted Mordred descriptors and PaDEL fingerprints for the molecules.  In both the processes the smiles of the molecules are used to generate the data further.  As mordred gives us about 1500 molecular descriptor of the compound we used  Mann-Whitney U test which test the equality of means in two independent samples.  The test is used here to see difference between the two bioactivity classes, this will test the statistical significance of difference.  The code for that is taken and modified from: <https://machinelearningmastery.com/nonparametric-statistical-significance-tests-in-python/>  By the test where the distribution was same it fails to reject the descriptor and where the distribution is different it rejected it giving us Descriptors of some significance to the prediction.  Similarly PaDEL fingerprints were calculated, we download the PaDEL-Descriptor software from: <https://github.com/dataprofessor/bioinformatics/raw/master/padel.zip>  <https://github.com/dataprofessor/bioinformatics/raw/master/padel.sh>  Then it took our data and calculated PaDEL fingerprints.  And finally each data is concatinated with pIC50 values. |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | The Mordred data:    The PaDEL data:    After the generation of datasets both sets are processed and split into 80:20 ration that is 80% data for training and 20% data for test.  This is done using train\_test\_split from sklearn library.  Then the models are trained using the train data and are tested for their r2 values.  r2 value –  The values range from 0 to 1 commonly stated as percentages from 0% to 100%,  It’s the statistical measure of how close the data are to the fitted regression lines. Also known as coefficient of determination.   * 0% indicates that the model explains none of the variability of the response data around its mean. * 100% indicates that the model explains all the variability of the response data around its mean.   In our program we computed r2 between predicted and actual values of bioactivity to get the performance of model, more techniques like mean square error, mean absolute error, variance score is also computed. | |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | Model: In the project we applied 6 different regression models using:   * Random forest regressor * Support vector regressor * Decision tree regressor * k-Neighbour regressor * Adabooster regressor * Voting regressor   All the models are applied on both the datasets (Mordred and PaDEL data). Regression using Random Forest: Statistics for mordred data:    Statistics for PaDEL data:    Regression plots for both the data: | |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | Regression using Support Vector Regressor: Statistics for mordred data:    Statistics for PaDEL data:    Regression plots for both the data: | |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | Regression using Decision Tree Regressor: Statistics for mordred data:    Statistics for PaDEL data:    Regression plots for both the data: | |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | Regression using KNeighbour Regressor: Statistics for mordred data:    Statistics for PaDEL data:    Regression plots for both the data: | |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | Regression using Adabooster Regressor: Statistics for mordred data:    Statistics for PaDEL data:    Regression plots for both the data: | |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | Regression using Voting Regressor: Statistics for mordred data:    Statistics for PaDEL data:    Regression plots for both the data: | |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | All Model Statistics: These are the best statistics obtained running the program:  For Mordred Data:    For PaDEL Data:   Conclusion:  * The PaDel fingerprint calculations were faster and better for predicting the bioactivity values as compared to Mordred descriptors. * After running and observing different models to fit the data the decision tree regressor was better in most of the iterations for Mordred data. * After running and observing different models to fit the data the voting regressor was better in most of the iterations for PaDEL data. * The max r2 achieved for Mordred data was: 0.395 * The max r2 achieved for PaDEL data was: 0.642   **\*\* The values of r2 can change at the time of run.** | |  |

##### References:

* <https://www.ebi.ac.uk/chembl/>
* <https://machinelearningmastery.com/nonparametric-statistical-significance-tests-in-python/>
* <https://github.com/dataprofessor/bioinformatics/raw/master/padel.zip>
* <https://github.com/dataprofessor/bioinformatics/raw/master/padel.sh>
* <https://www.youtube.com/watch?v=plVLRashaA8>
* <https://www.youtube.com/watch?v=qWVTxfLq2ak>
* <https://www.youtube.com/watch?v=zD2focOkQ48>
* <https://www.youtube.com/watch?v=wGaGm0sj04M>

Go corona go

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Thank You. |