GLA UNIVERSITY
ESTD. 2010
ऋते ज्ञानान्न मुक्तिः
GLA
UNIVERSITY
MATHURA
Recognised by UGC Under Section 2(f)
Accredited with A+ 3.46 Score Grade by NAAC
12-B Status from UGC

Report

On

Drug Discovery And Classification

Submitted By:

Saksham Saxena 2AE 2215500130

Vanshika Raghav 2AE 2215500169

Suparshava Bhatnagar 2AE 2215500157

Risabh Bhaghel 2AE 221500121

Submitted To:

Mr. Sayantan Sinha

Technical Trainer

Dep. CEA

Drug Repurposing

Objective:

The main objective of the project is to address the challenges in drug discovery by repurposing existing drugs for new therapeutic purposes i.e. to build a machine learning model which finds similarities in protein binding properties between molecules using the number of different functional groups for our purpose.

Scope:

The scope of the project encompasses various stages of drug repurposing, focusing primarily on predicting drug-target interactions (DTIs) using machine learning techniques

Methodology:

We are using Drug Repurposing method for prediction with tensor, Standard Scaler ,numpy and pandas.

Proposed System:

- Drug repurposing is a strategy for identifying new uses for approved or investigational drugs that are outside the scope of the original medical indication.
- The whole idea is based on the idea that similar molecules are usually associated with similar protein targets. Thus, these approaches predict interactions based on similarities between connections protein.
- This method was overtly popular in 2020 when researchers were rushing to find out drugs for CoVID-19. Designing a completely new drug and getting it approved through several rounds of clinical trials and approval agencies is time taking and not worth it as it has been seen that less than 10% of interesting candidates make it to market.
- For proteins, conversion regulates amino acids in 7 groups according to their physical chemical properties. In case of Smiles strings, an encoding technique was used to convert each character as an integer to be used as a feature of the drug.

Features:

- Reduce time for new drug development.

Team Members:

Saksham Saxena

Vanshika Raghav

Suparshava Bhatnagar

Risabh Bhaghel

Output and Discussion:

First step:

- We needed a dataset that contains the number of functional groups in different molecules along with the information that they are active or inactive in inhibiting a particular protein or protein substance that is in turn responsible for some disease.
- That seemed so easy, until we did not find any. Our primary source Drug Bank, was huge and inaccessible. And finally when we gained academic license to access the database it was mostly composed pharmacological data that we had no use of rather than cheminformatics data and 3D structures that we actually needed.
- serendipitously, we came across a research paper by debankit which has all the dataset we needed for our prediction model on E.COLI virus and HIV.

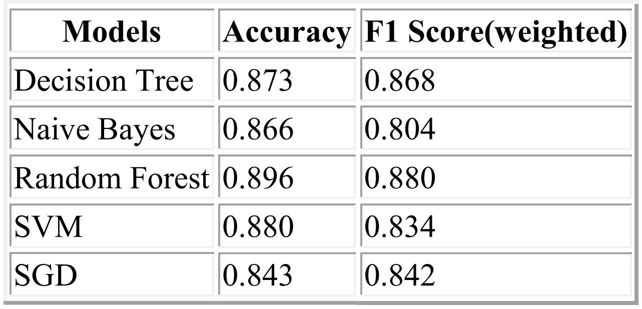
Second step:

We have taken the prepared datasets as input and converted into two lists containing the features and labels.

We have split the dataset into 30 percent testing data and 70 percent training data.

We have implemented Decision tree classifier, Naive Bayes classifier, Random Forest classifier, SVM and SGD models.

On E-coli inhibitor data:



Models	Accuracy	F1 Score(weighted)
Decision Tree	0.873	0.868
Naive Bayes	0.866	0.804
Random Forest	0.896	0.880
SVM	0.880	0.834
SGD	0.843	0.842

On HIV inhibitor data:

Models	Accuracy	F1 Score(weighted)

Decision Tree	0.890	0.885
Naive Bayes	0.860	0.801
Random Forest	0.920	0.908
SVM	0.896	0.866
SGD	0.843	0.849

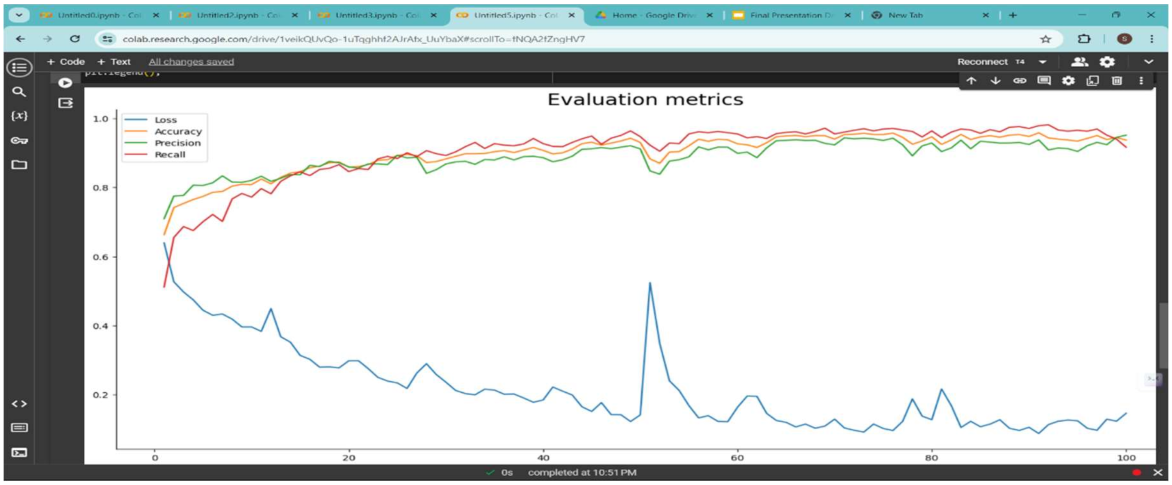
We have tried to mix the data of both the E-Coli and HIV so that we can classify the different molecules to check if they are helpful in these two diseases. Our model holds and thus our plan has been quite successful as of yet.

Models	Accuracy	F1 Score(weighted)
Decision Tree	0.876	0.871
Naive Bayes	0.893	0.842
Random Forest	0.924	0.915
SVM	0.908	0.877
SGD	0.874	0.868

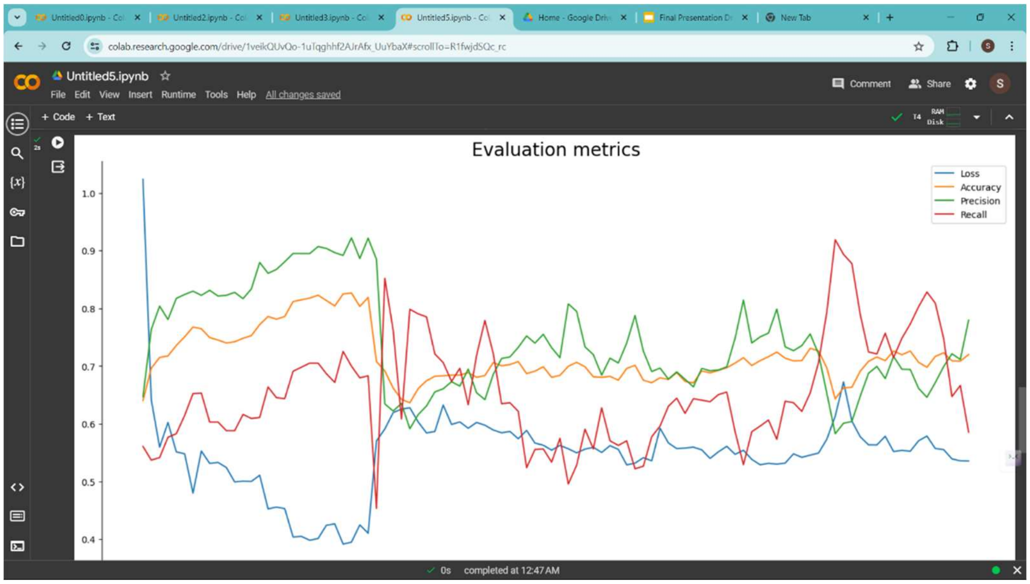
Deep Learning Model For the HIV Dataset

- We have split the dataset containing around 3000 rows into 20 percent test data and 80 percent training data.
- We have used the Standard Scaler so as to not confuse the neural network regarding the weights of the features.
- We have used the tensorflow library to train our model. There are 4 Dense layers with 3 Relu Activation function and 1 last Sigmoid activation function.
- We have taken two different learning rate cases. In case of 0.03 learning rate,
- In 0.01 learning rate, we get a good result after 100 epochs. (Training Accuracy = 0.93 in 100 epochs)

With 0.01 learning rate



With 0.03 learning rate:



The confusion matrix for the first case with learning rate of 0.03

[[224 58]

[147 160]]

The confusion matrix for the first case with learning rate of 0.01

[[250 32]

[129 178]]

As is observable, the diagonal elements representing True Positives and True Negatives are better for the 0.01 learning rate case. This is the test data result we got with the learning rate of 0.01.

Accuracy : 0.73

Precision : 0.85

Recall : 0.58

Ideas to Improve:

We believe, the output we get from our model can again be used in the model created in our reference papers, thereby increasing the accuracy of getting a viable new drug.

If we can increase the number of datapoints, then we can get a better trained model.

References:

- El-Behery, H., Attia, A. F. (2021).

- https://www.sciencedirect.com/science/article/pii/S1476927121001031?via%3Dihub
- https://www.niser.ac.in/~smishra/teach/cs460/2021/project/21cs460_group14/
- https://github.com/debankit/21cs460_group14

Conclusion:

- Although these results might seem very promising and give the conception that now given any new small molecule our model can predict whether it can be used as a drug for some disease, that is misleading.
- Drug designing is a very complicated process(what we learned in the hard way) and this model, if extended to all major diseases, given their databases are publicly available(which isn't, all due credit to big pharma), will definitely help drug designers to make better predictions.
- A small molecule being active for a particular disease causing protein doesn't imply it can be used as a drug. 'Active' is a broad term as interpreted by Mohapatra et al. includes even those which just interact with the protein(not necessarily block the foreign disease causing element get attached).
- Other factors like Docking efficiency, Orientation of the molecule, tautomerism or conjugation if present also comes into play. But, there are softwares like AdMET and Vina AutoDOCK to do that. Our model acts as an enabling technology and helps make researchers make better predictions by narrowing down their search significantly.