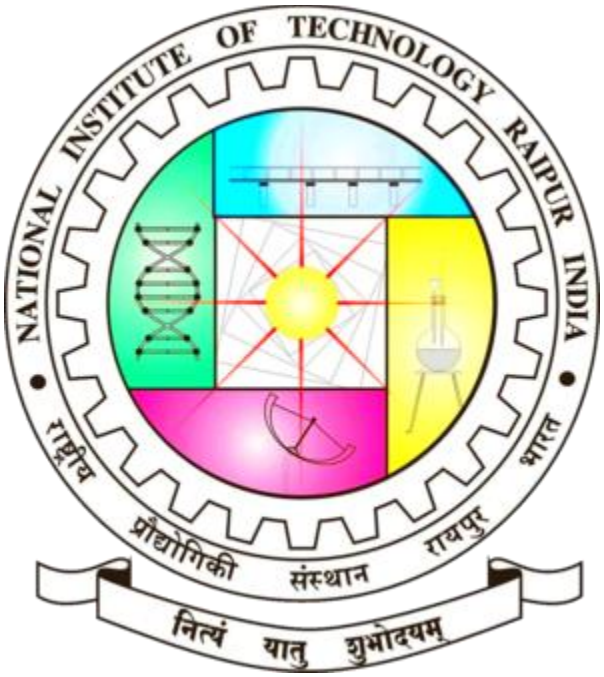
Machine Learning approach to model an antidote against Snake Venom

Project Report Submitted

in partial fulfillment of the requirement for the degree of

BACHELOR OF TECHNOLOGY



By

Ankeet Singh - 19112001

Brijesh Painkra - 19112009

DEPARTMENT OF BIOTECHNOLOGY

NATIONAL INSTITUTE OF TECHNOLOGY RAIPUR

(May 2022)

CERTIFICATE

It is to certify that the work contained in the project report titled **"Machine Learning approach to model antidote against Snake Venom"** by **"Ankeet Singh"** and **"Brijesh Painkra"** has been carried out under my supervision and this work has not been submitted elsewhere for a degree.

Dr. D.N. ROY
Assistant Professor
Department of Biotechnology
NIT Raipur
November 2022

DECLARATION

I declare that this written submission represents my ideas in my own words and where others ideas have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

(Signature)

Ankeet Singh and Brijesh Painkra

(Name of the Student)

19112001 and 19112009

(Roll no.)

Date: 28/11/2022

ACKNOWLEDGEMENT

We would like to thank our Project Guide **Dr. D.N. Roy**, Department of Biotechnology, NIT Raipur for their constant support, help, motivation and guidance. Their constant encouragement, constructive criticism and valuable guidance have been the cause for successful completion of the project. We gratefully acknowledge **Dr. Lata Upadhaya**, Head of the Department of Biotechnology, NIT Raipur, for allowing us to work on this field and project in particular. Finally we wish to convey our warmest and deepest gratitude to our family members, to whom we owe all our achievements, and to our friends for their endless support and encouragement.

Ankeet Singh - 19112001

Brijesh Painkra -19112009

Date - 28/11/2022

Contents:

List of Figures

Abstract

Snake Bites globally account for several hundred thousands of deaths from millions of bite cases reported and is also therefore recognized by WHO as one of the neglected tropical diseases in 2009. The enzymes mostly commonly found in deadly snake venoms are Acetylcholinesterase, Phospholipases A(2), Serine Proteinases, Metalloproteinases and l-amino acid oxidases. But according to numerous researches Phospholipase A2 or PLA2 seems to be a promising target among these to make a broad spectrum of antivenom[1]. A sizable non-redundant dataset of 615 molecules with the reported IC50 values against PLA2 was obtained from ChEMBL and employed to build a QSAR model using Machine Learning. The ML model built was to predict the bioactivity of compounds or IC50 against PLA2 in an effort to produce a broad range antivenom. Instead of directly predicting the IC50 the model predicted pIC50 or -log[IC50] to reduce the range of prediction. After cleaning of data that includes steps like removing duplicates and checking for missing values, the molecular descriptors were calculated using the PubChem library[2]. The data was then split into training and testing sets for our primary model which was built on Random Forest Algorithm. Further models were also built using different algorithms to test their efficiency from where it came out that our primary model was among the best performing with an RMSE error of 0.7. Some statistical analysis was also done to perform exploratory data analysis or to see the chemical space of compounds using the PaDEL-descriptor software[3] calculating molecules lipinski's descriptors from their chemical structure represented by SMILES notation. The information gathered from these analyses will help build guidelines for the design of novel and robust PLA2 inhibitors.

Abbreviation

1) PLA2 - Phospholipase A2 ,
2) sPLA2 - secreted Phospholipase A2,
3) MW - Molecular Weight,
4) LogP - Partition coefficient
5) numHBD - number of Hydrogen Bond donor sites
6) numHBA - number of Hydrogen Bond acceptor sites
7) ML – Machine learning

Objective

To build a ML model that will predict pIC50(i.e. -log[IC50]) or bioactivity data of a compound against the target of our interest, in this case an enzyme from snake venom.

To make a web app out of the model so a layman can also use it for prediction.

Introduction

In India alone deaths due to snake bite is several thousands per year from among over million cases reported as per a recent report of WHO[4]. While in a global perspective according to an extrapolation by Chippaux, each year 5,400,000 persons are bitten by snakes globally, from these 2,500,000 are envenomed, 125,000 pass away, and more than 100,000 experience severe aftereffects [5]. Unfortunately, governments and international health organisations have long ignored snakebite, despite the fact that the fatality rate from snakebites is equal to one-fifth of malaria fatalities globally and half of HIV/AIDS deaths in India [6].

Antivenin is till now the sole methodical treatment for snake envenomation cases mentioned above. Although currently employed immunised animal sera (mostly sheep or horses) are quite successful, they are constrained by a few limitations [7]. First, antivenin cannot undo local tissue damage brought on by snake venom exposure, which frequently results in amputation [7]. Additionally, anaphylaxis, pyrogenic responses, and serum sickness are examples of early and late adverse reactions to antivenin [8]. Additionally, antivenins are frequently difficult to get. Due to a shortage of cold chain storage and other intricate political factors, certain isolated, rural areas that need antivenom the most are unable to obtain appropriate supply. Finally, in low-income nations, the majority of antivenoms are out of reach from the patient's well wishers or family [9].

Recently, the most effective antivenin against African vipers, mambas, and cobras, Fav-Afrique, was no longer being produced by the nonprofit French pharmaceutical company Sanofi Pasteur. As a result, rural Africa is currently experiencing a severe snakebite problem [10]. This worrying condition together with all the drawbacks of traditional antivenom mentioned above highlights the need for novel antivenom medication candidates and antivenin substitutes. The ML model built focuses in particular on the chemical family of Group IIA of snake venom phospholipase A2s (PLA2s), which is also found in many venomous snake species. Here, we discuss these PLA2s inhibitors' by

categorizing them in active, intermediate or inactive classes on the basis of their IC50 values found in the ChEMBL database, and the possibility that they could serve as a common drug for broad-spectrum antivenom medications.

Literature Review

Snake venoms are complex mixtures made up of L-amino acid oxidases, disintegrins, phospholipase A2s, metalloproteases, C-lectins, serine proteases, and a few other substances [11]. The choice of PLA2 was made due to various studies concluded whose highlights have been explained in detail in this section. Among which one of the main reasons can be attributed to the availability of a wide range of PLA2 inhibitors both artificial and lab made[12]. Also the activity of PLA2 of the snake enzyme is considered as a rate limiting step. The PLA2 is also seen as a cause of neurotoxic, cytotoxic and myotoxic effect causing substance. Due to which targeting this enzyme can help build a broad range of antivenom.

Materials and Methodology

Materials

As our model is virtual so there is no physical material required but apart from this the required things are :-

- Python Language
    - Python libraries
    - ML algorithms
- ChEMBL database
- SMILES structure
- PubChem

Methodology

We used Python language to build our model. The data was collected from the ChEMBL database that contains bioactivity data of over 2 million compounds. We used the ChEMBL web resource client to collect data natively from our python code on Google Collab. The data

collected included the IC50 values against PLA2, their chemical structure represented by SMILES notation and their ChEMBL ID that represented the molecule uniquely. The data was then cleaned like removing salts from chemical structure and other redundant functional groups from the structure using PubChem molecular fingerprints library. But Prior to that the data gathered was pre-processed in the sense that empty fields and redundant data were found out and removed. The primary target of removing redundant data is to have data normalized so our model can only learn from highly indicative features. Further, PubChem molecular descriptor was only used to do feature engineering of our input features into a vector space of [0, 1] that basically represented the chemical structure. In which the PubChem molecular descriptor primarily used Canonical SMILES structure which was collected through ChEMBL as mentioned above.

Then classification of the molecule was done on the basis of their IC50 value; active if their IC50 values were less than 1000nM, intermediated if their IC50 values were between 1000nM - 10,000nM and inactive if they were beyond the 10,000nM. The prediction range would have been very large if the original scale of IC50 had been chosen. In such a case the model would not have properly learned the huge scale of chaotic data from its representation graph. Therefore conversion of IC50 to pIC50 was done to access a small subset of data which made the category of active compounds fall above 6, intermediate between 5-6 and inactive below 5. Then this data of over 600 compounds received after preprocessing was split into 80/20 ratio. The 80% was used to train our Random Forest Algorithm from Python's Scikit learn library and after training the 20% was used to test our model’s accuracy.

We also used Python’s lazy predict library to compute over 30 different ML algorithms and their performance from our collected dataset of compounds. After analysis of which we computed their training time and decided to built a web app that can predict the activity of compounds for our target molecule(PLA2) based on the learning of our ML model. The web app was built using Streamlit library and to manage different package redundancy we used a natively built Conda environment.

Results and Discussion

Before building our model we did an exploratory data analysis on our training data using the Rdkit library to calculate the Lipinksi descriptors. The Lipinski rule is used to evaluate the likelihood of a molecule to be a drug.

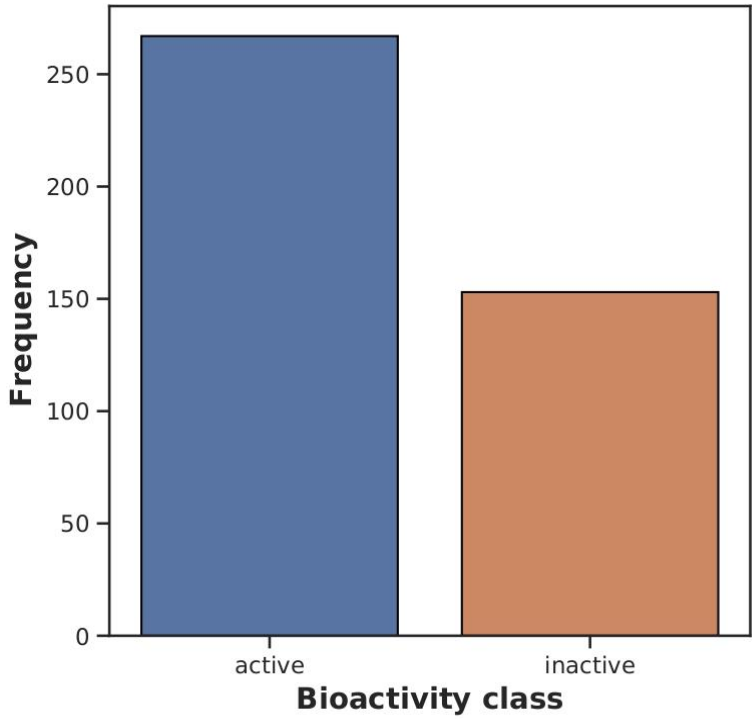


Fig 1 – Frequency

Fig- 1 bar graph represented the classification of our molecules; 250 were active, 150 were inactives and the rest 200 were intermediate molecules. The exploratory data analysis was not done on these intermediate molecules.

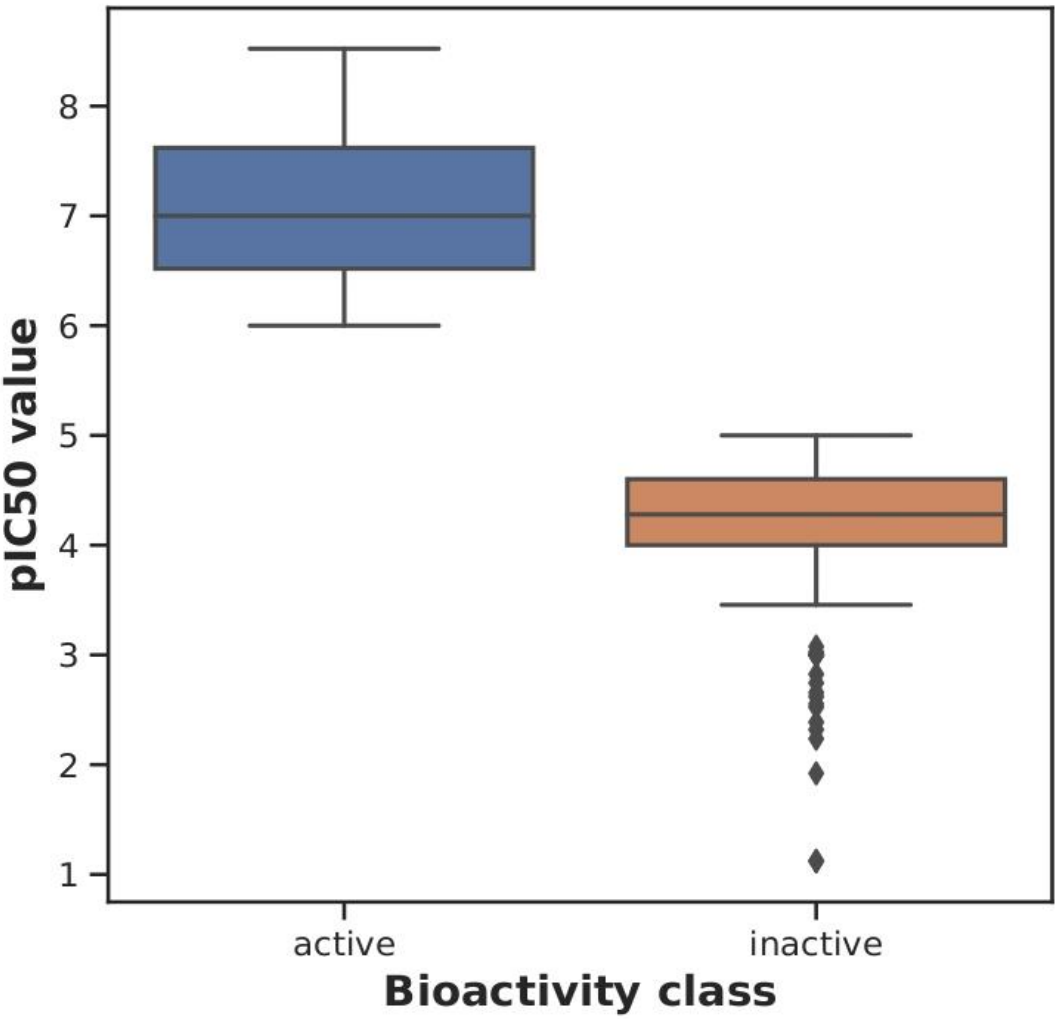


Fig 2 – pIC50 value

The pIC50 values distribution of the active and inactive molecules was also plotted and as obvious we see a distinction which is the primary basis on which we classified the compounds. The active molecules showed activity below 8 and little close to 6.5 while the inactive molecules were in the range of 4 to slightly less than 5.

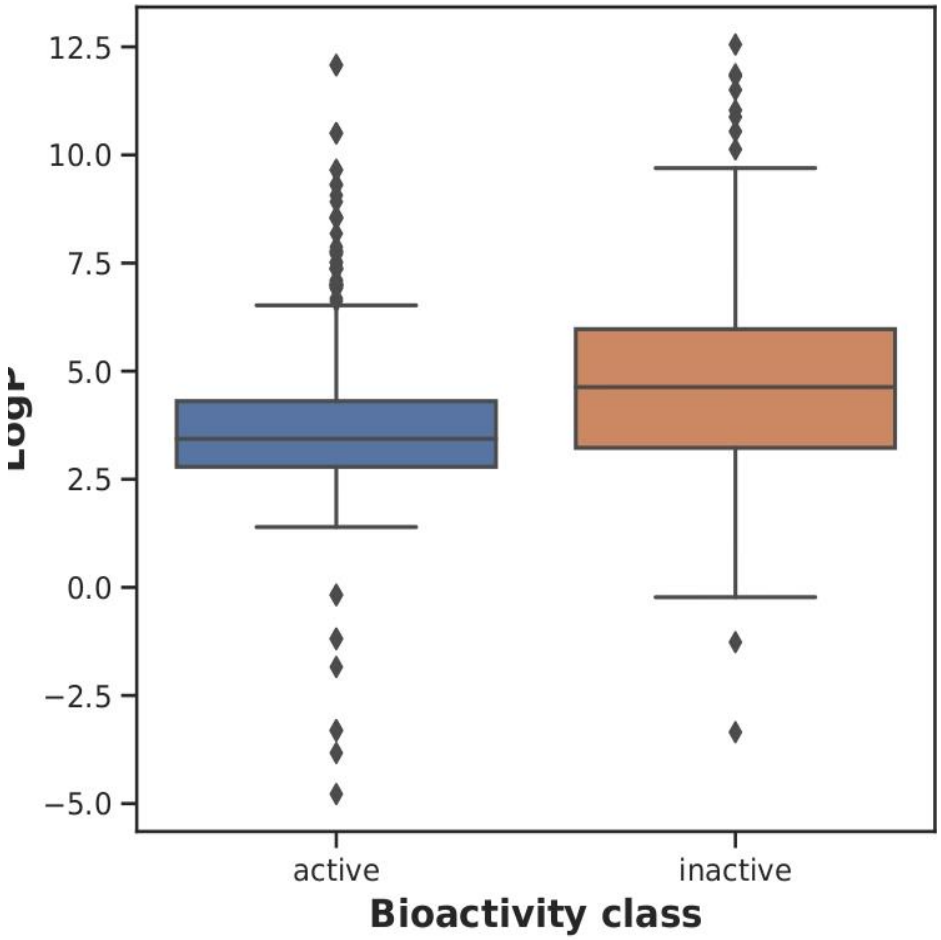


Fig 3 - LogP

This bar plot represented the logP values distribution of active and inactive molecules. The logP value tells the property of compound to dissolve in lipophilic solvent over polar solvent. The active molecules fell between 2.5 to 5.0 and inactive molecules fell between 2.5 - 7.5. According to the Lipinski rule that states the value of LopP to be under 5 actually indicates the likeness of our active molecule to be a good drug candidate.

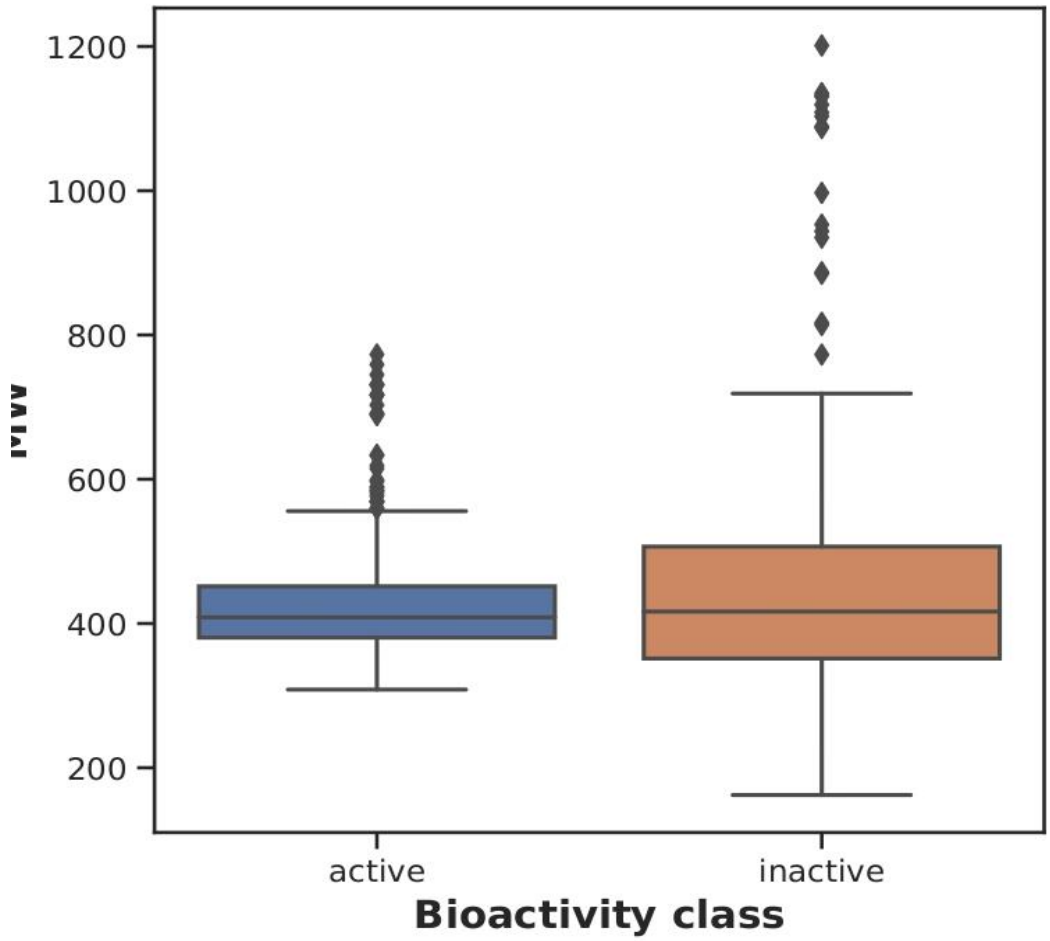


Fig 4 – Molecular Weight (MW)

The above plot is the representation of MW distribution of active and inactive compounds in our training dataset. As is evident the active compounds are showing a range of around 400 dalton while inactive compounds are giving a range of 500 to little less than 400 dalton. These results states that our active molecules are yet again at a better position to be drug candidate according to Lipinski rule of five which puts an upper limit of 500 Dalton for a molecule to show good drug likeness.

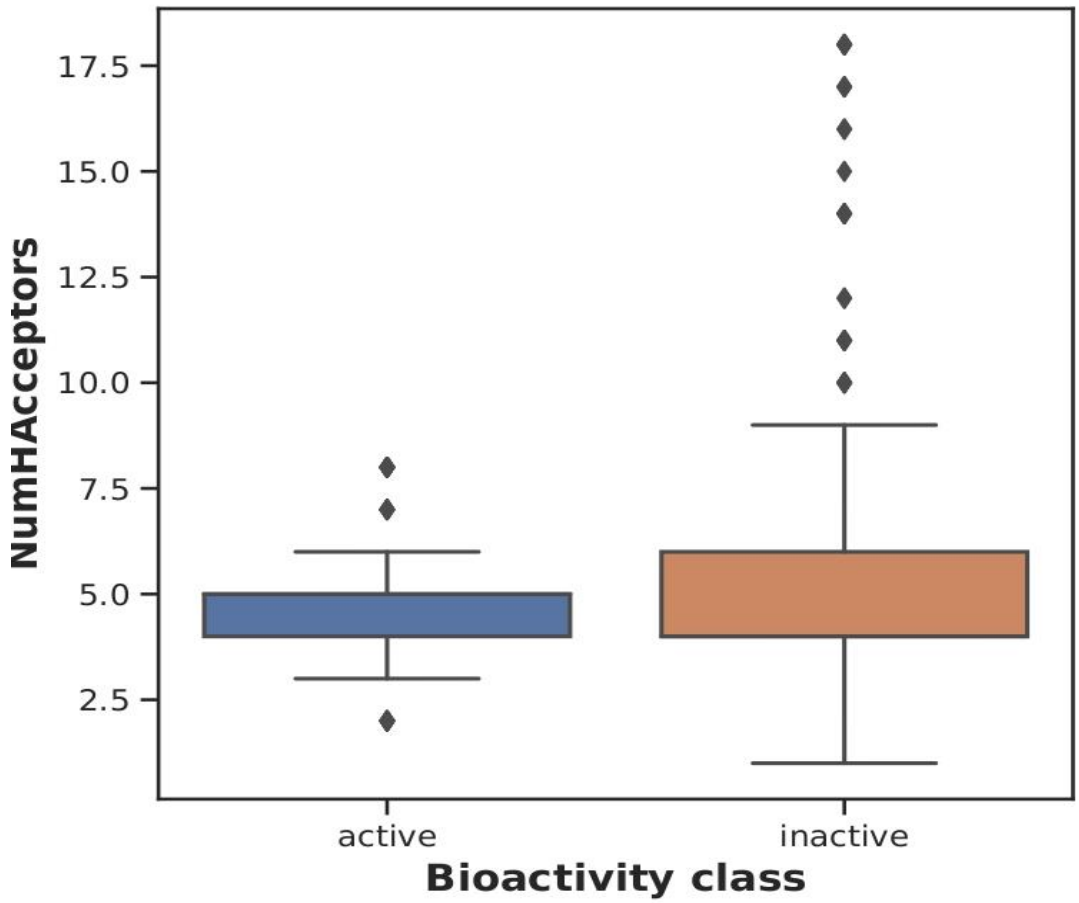


Fig 5 - NumHAcceptors

This above plot is a representation of the number of Hydrogen Bond Acceptors of active and inactive molecules which for active molecules is less than 5 and for inactive molecules is greater than 5. And according to the Lipinski rule this value should be less than 10 for a molecule to be classified as a likely molecule for a drug. So both active and inactive molecules satisfy this condition.

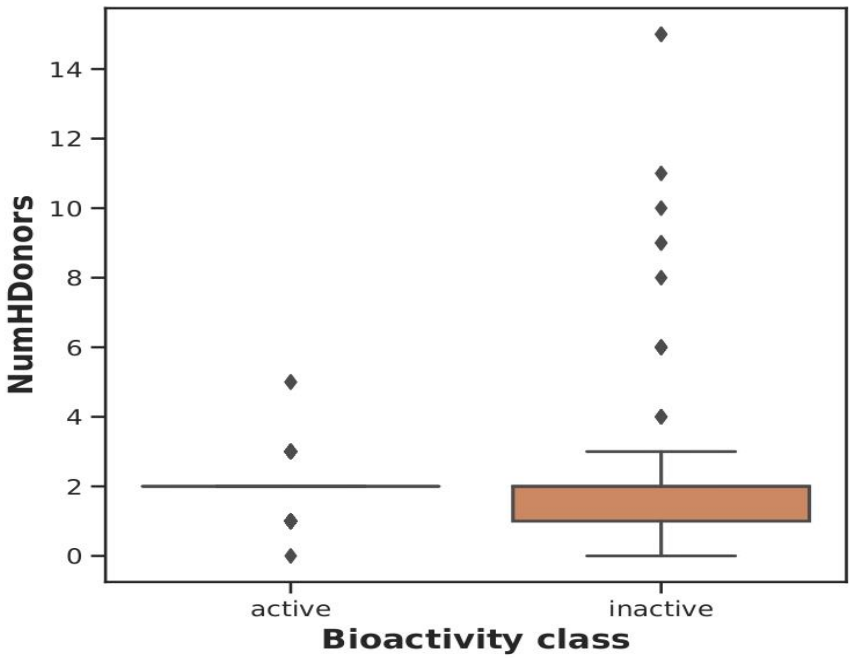


Fig 6 - NumHDonors

The above plot represents the distribution of the number of Hydrogen Bond Donors between active and inactive molecules. According to Lipinski rule this value should be less than 5 for a molecule to show likelihood characteristics of a drug. The plot above indicates our active compounds to fall under such categories along with inactive compounds.

As we found many properties of active and inactive to be different and at the same time some similar too so we performed a statistical analysis called Mann-Whitney U test on the generated Lipinski descriptor data to see if there is some statistical difference between the two classes. The results were 1.21e-05 for LogP, 0.658 for MW, 0.910 for NumHAcceptors, 0.0247 for NumHDonors. From which we can conclude that LogP and NumHDonors are statistically different as they are less than 0.05 required to reject the null hypothesis. The Activity data is also obviously different as classified earlier.

Then we build our model using Random Forest Algorithm and its prediction over our test data set can be seen here:

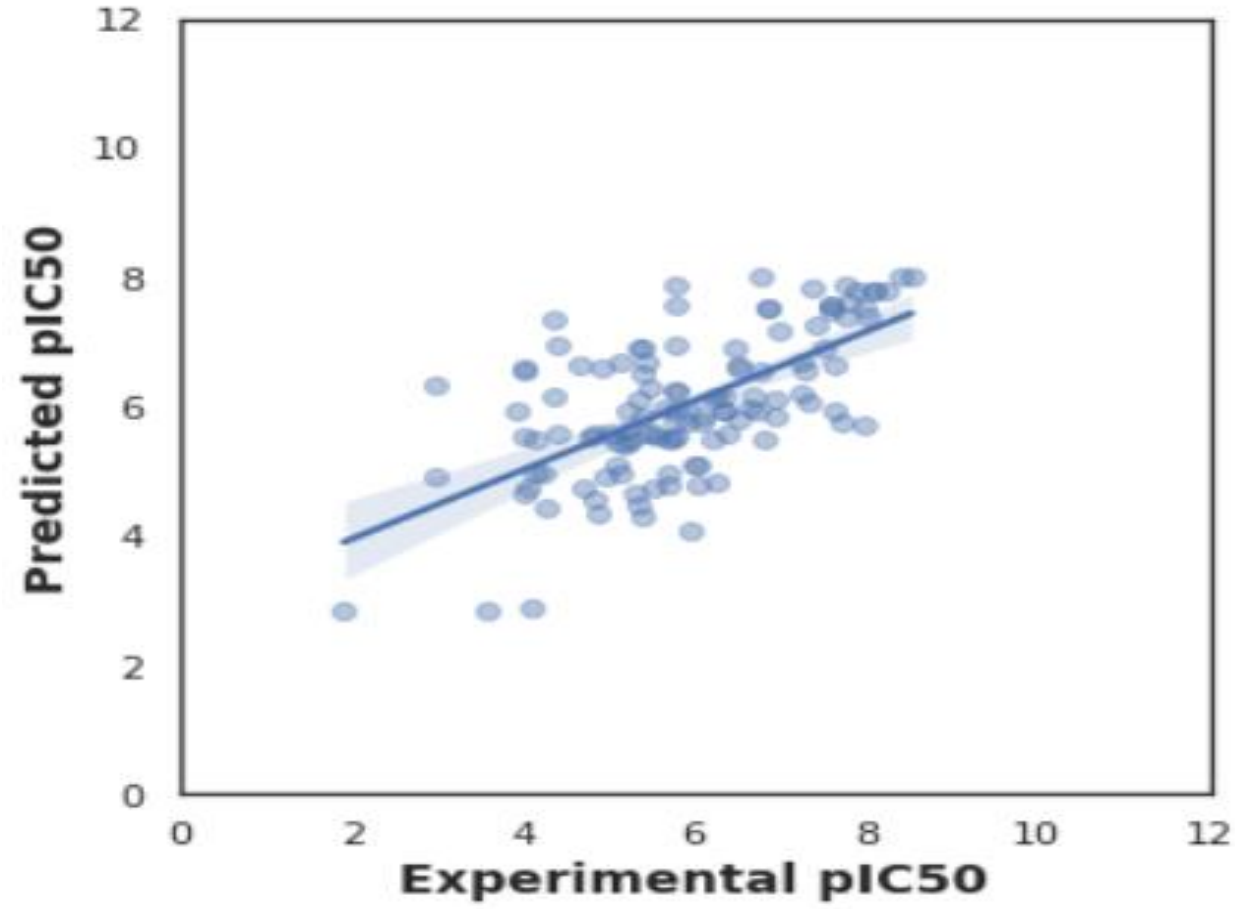


Fig 7 – Predicted pIC50 value Vs Experimental pIC50 value

The experimental and predicted values for the compounds are very close as seen from the plot which was also confirmed from the RMSE or root mean square error of the model that came out to be 0.3.

Lastly using lazy predict library we built ML model of over 30 different algorithms and calculated their efficiency:

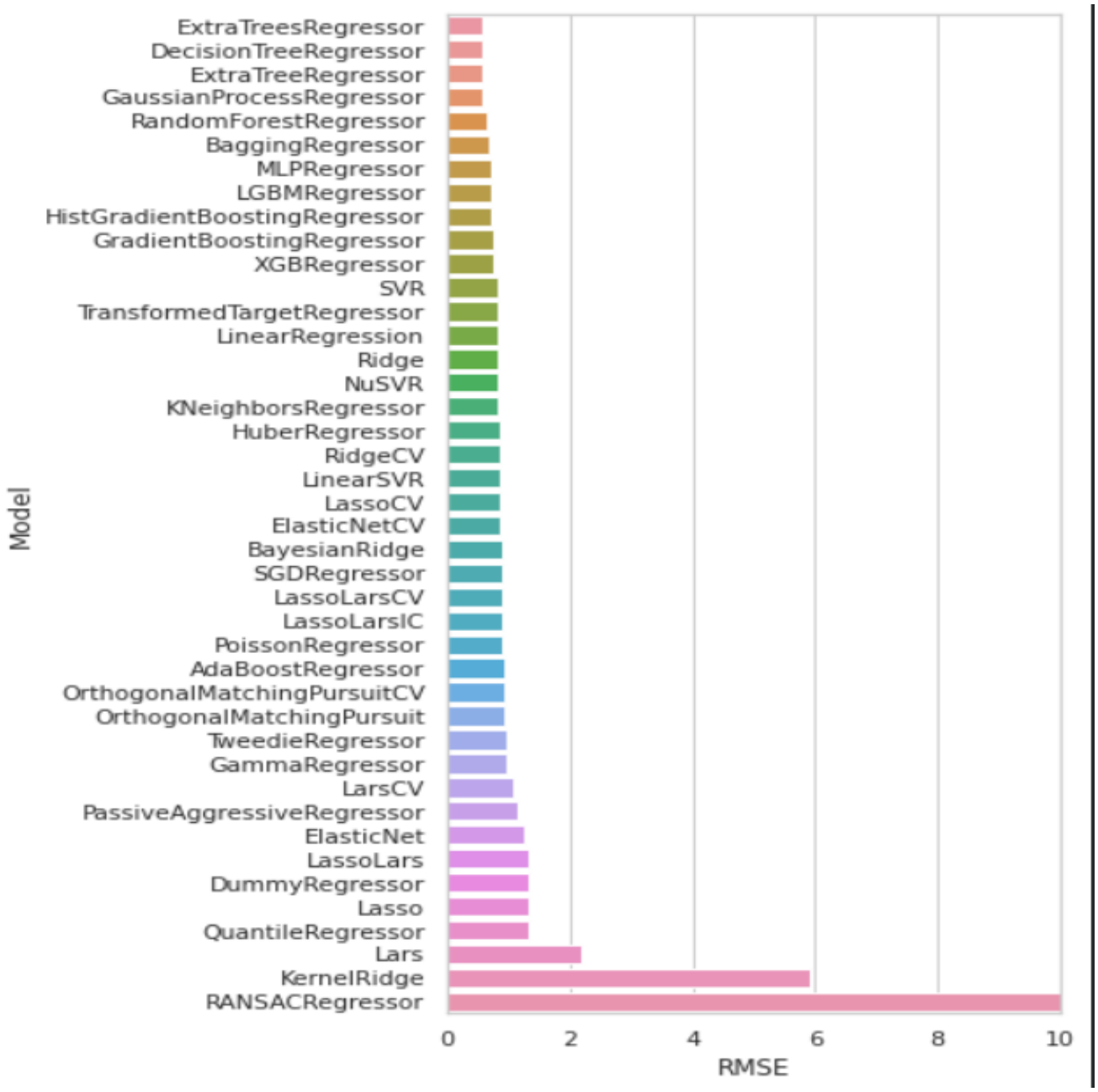


Fig 8 – RMSE value

This plot shows that the ExtraTreesRegressor model has the least RMSE error although our random forest model is also very close to it. But due to the popularity and reliability of Random Forest Algorithm we decided to build our web app using that only. The below plot then shows the time taken by all the models to train which is also an important task to consider during a ML project.

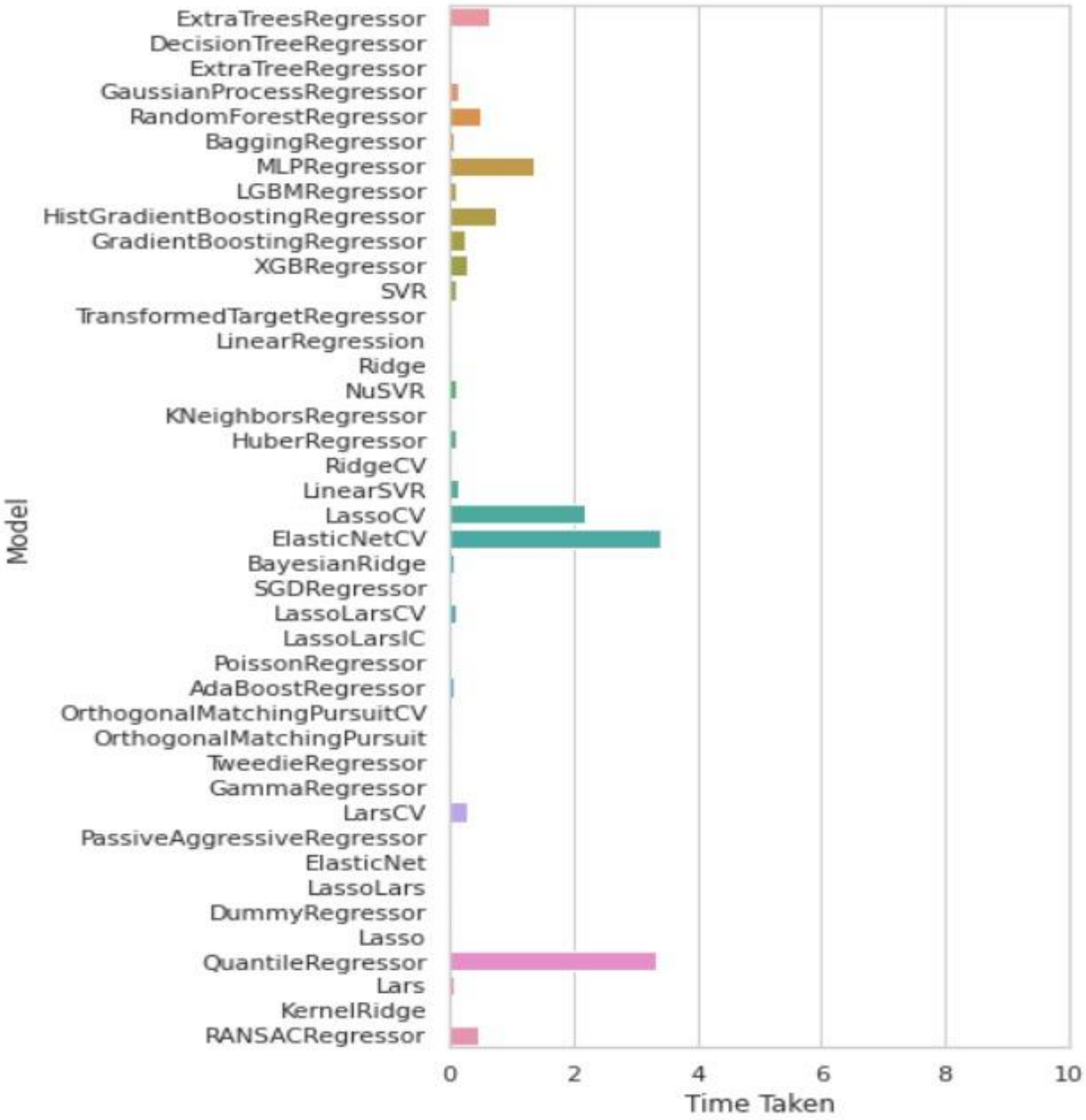


Fig 9 – least time taking

Again in this our Random forest algorithm is among the one taking least time.

Conclusion

Machine learning technology is a boon to modern society. Machine learning technology is growing so fast that it enters in nearly every field of work and becomes a very important asset to them. Machine learning techniques help decision-makers in the pharmaceutical industry, and allows them to make good decisions in a variety of applications, including QSAR analysis, hit discoveries, and de novo drug designs.

The key reason for inadequate interpretability outcomes, which may result in limiting the applications in drug development, must be applied to ML difficulties and the data available. In the pharmaceutical industry, ML algorithms and deep learning approaches are used on a regular basis. Particularly in image analysis and omics data, ML approaches have run into resolving various conflicts in drug development areas and health service hubs.

In the field of medicine, machine learning (ML) models make predictions based on trained data inside a known framework, i.e., the compound structure, which can be used in place of more conventional techniques like PPT inhibitors and macrocycles.

In our project “Machine Learning approach to model antidote against Snake Venom” we use ML technology and it has a very important role or we can say that our model is entirely dependent on the ML. ML has different algorithms, all are unique in their own way and has different usability and efficiency. We have used the random forest algorithm which is the third most efficient in our case but its ease to use and reliability have motivated us to choose this algorithm. This algorithm also made the work easy and hassle free by getting trained in quite a fast duration. Then after training it started predicting quite well and even accurately too.

In modern times there are many ML models in the pharmaceutical field, taking inspiration from them we designed our model and we hope it will help in finding the appropriate drug in less time making the work easy for others to develop the drugs.

Reference

1) Huixiang Xiao # 1, Hong Pan # 1, Keren Liao 1, Mengxue Yang 1, Chunhong Huang: "Snake Venom PLA2, a Promising Target for Broad-Spectrum Antivenom Drug Development"; Pubmed; https://pubmed.ncbi.nlm.nih.gov/29318152/
2) Fernández-de Gortari, E., García-Jacas, C.R., Martinez-Mayorga, K. *et al.* Database fingerprint (DFP): an approach to represent molecular databases. *J Cheminform* 9, 9 (2017). https://doi.org/10.1186/s13321-017-0195-1
3) Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011 May;32(7):1466-74. doi: 10.1002/jcc.21707. Epub 2010 Dec 17. PMID: 21425294.
4) WHO: "SnakeBite Envenoming in India"; https://www.who.int/news-room/fact-sheets/detail/snakebite-envenoming
5) Chippaux J.-P: "Snake-bites: appraisal of the global situation."; *Bulletin of the World Health Organization.* 1998;76(5):515–524.
6) Editorial, "Snake bite—the neglected tropical disease," *Lancet*, vol. 386, no. 9999, pp. 1110, 2015.
7) Gutiérrez J. M., Theakston R. D. G., Warrell D. A. Confronting the neglected problem of snake bite envenoming: the need for a global partnership. *PLoS Medicine.* 2006;3(6):0727–0731. doi: 10.1371/journal.pmed.0030150.
8) De Silva H. A., Ryan N. M., De Silva H. J. Adverse reactions to snake antivenom, and their prevention and treatment. *British Journal of Clinical Pharmacology.* 2016;81(3):446–452. doi: 10.1111/bcp.12739.
9) Gutiérrez J. M., Williams D., Fan H. W., Warrell D. A. Snakebite envenoming from a global perspective: towards an integrated approach. *Toxicon.* 2010;56(7):1223–1235. doi: 10.1016/j.toxicon.2009.11.020.
10) Schiermeier Q. Africa braced for snakebite crisis. *Nature.* 2015;525(7569):p. 299. doi: 10.1038/525299a.

11) Warrell D. A. Snake bite. *The Lancet.* 2010;375(9708):77–88. doi: 10.1016/S0140-6736(09)61754-2.

12) Karakas M., Koenig W. Varespladib methyl, an oral phospholipase A2 inhibitor for the potential treatment of coronary artery disease. *IDrugs.* 2009;12(9):585–592.