Bioinformatics : Drug Discovery

By: Anagha Acharya - 1BM19CS224
Ramya Ramesh - 1BM19CS227
Tasmiya Fathima - 1BM19CS172
Trisha Lakhani - 1BM19CS214

Guide: Dr. Asha G R
Assistant Professor
Department of Computer Science & Engineering
B.M.S. College of Engineering

Outline

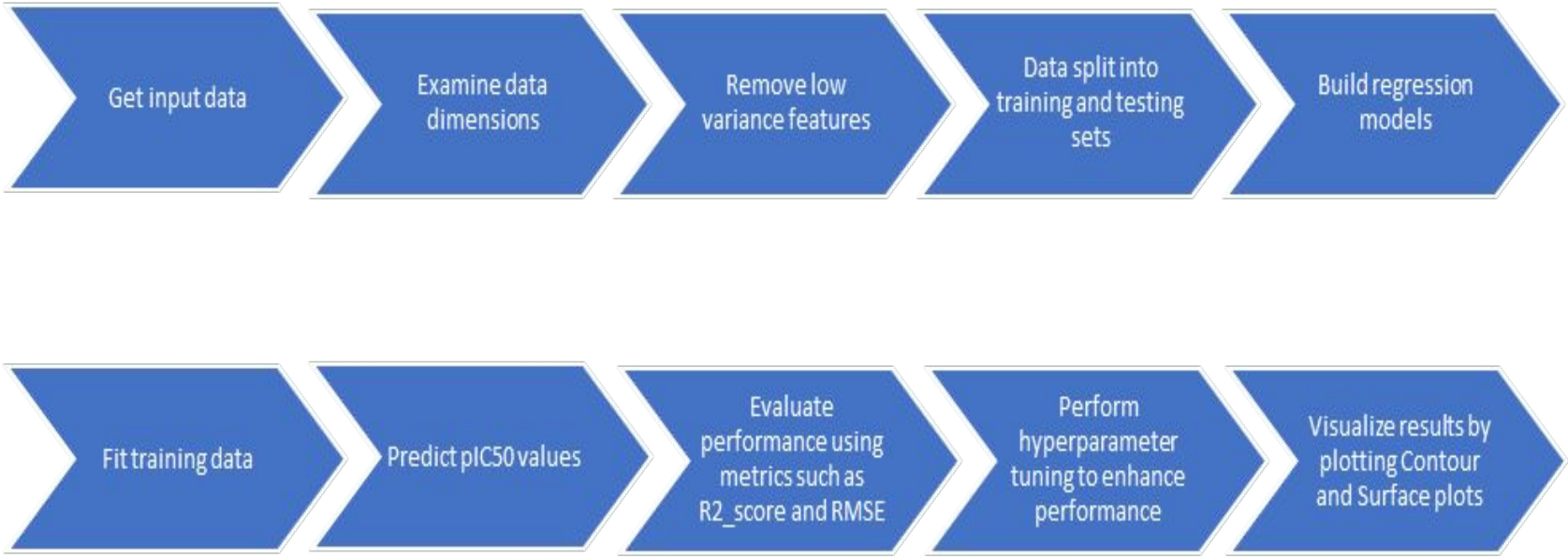
1. Problem Statement
2. Proposed Methodology
3. Tools used for implementation
4. Implementation
5. Testing
6. Results and Discussion
7. Applications / Relevance
8. References

Problem Statement

The long development pipeline faces increasing costs and additional challenges, including the lack of predictive validity of current animal models, insufficient knowledge regarding underlying mechanisms of disease, patient heterogeneity, lack of targets and biomarkers, a high rate of failed clinical trials etc. Bioinformatics analyses provide key information throughout the entire drug discovery and development process, from aiding the identification and validation of drug targets and leads through to helping assess the outcomes of phase 1, 2 and 3 clinical trials; as well as supporting drug repurposing efforts.

The aim of the project is to learn about Bioinformatics through Drug Discovery. The chosen protein target is Anaplastic Lymphoma Kinase(ALK). Various drug-target interactions are studied to make predictions on their action. This is done by searching for the target ALK enzyme and acquiring curated bioactivity data, performing exploratory data analysis and building regression models and scatter plots to compare Predicted and Actual values of drug action on target following which Model comparisons based on their performances can be made.

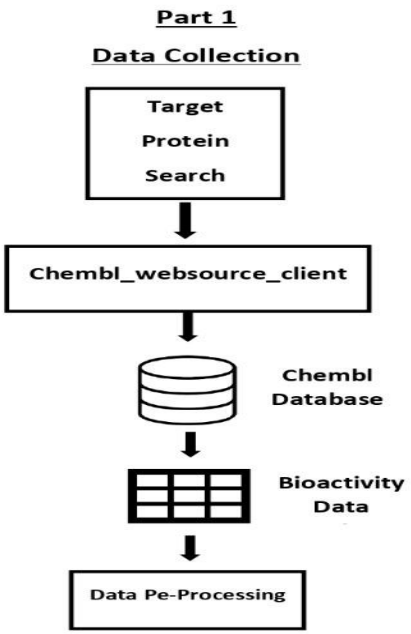
Proposed Methodology

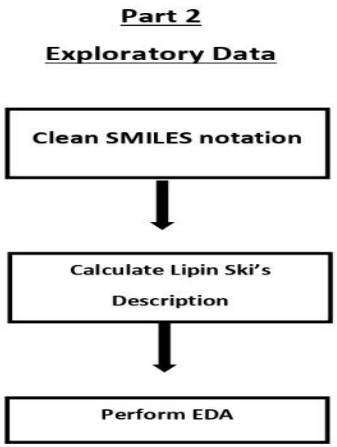


Tools used for Implementation

- Jupyter Notebook / Google colab
- Conda
- Python libraries (numpy, pandas, matplotlib, scikit-learn)
- RDkit
- PaDEL Descriptors
- Lazy predict
- Streamlit (for Web App)

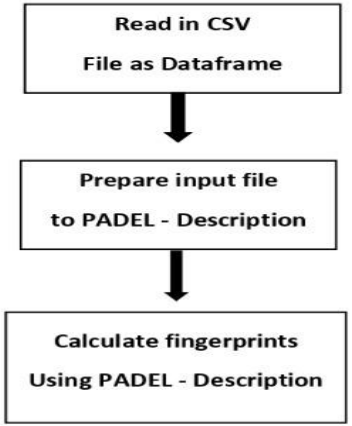
Implementation





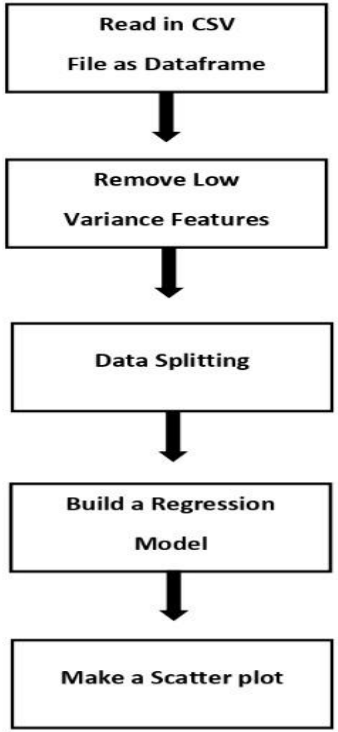
Part 3

Descriptor Calculation



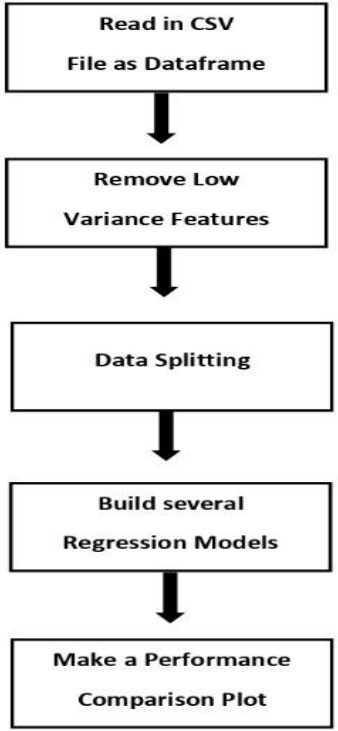
Part 4

Model Building



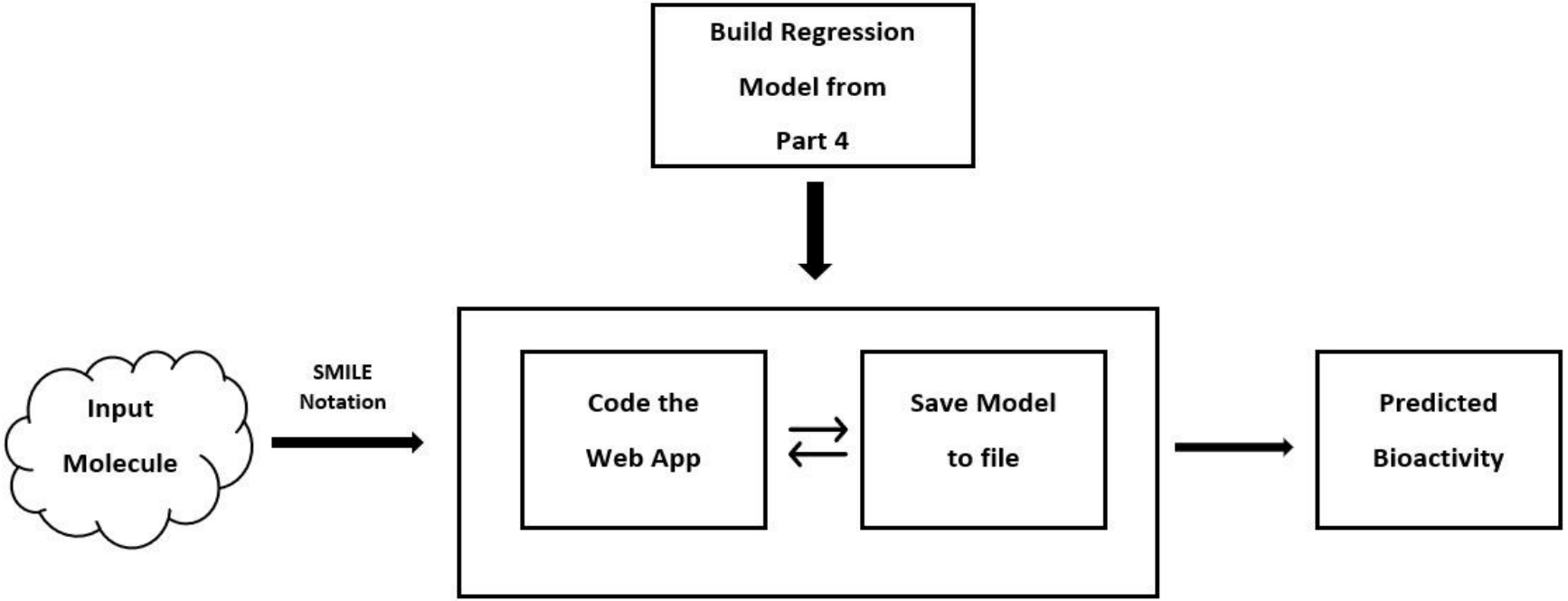
Part 5

Model Comparison



Part – 6

Deploy Model

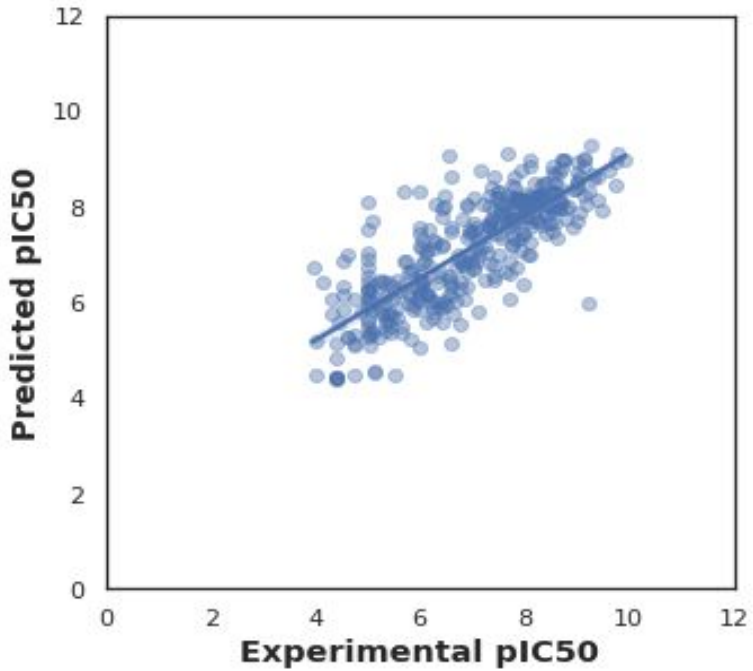


Testing

```
pd.DataFrame(data={'Actuals': Y_test, 'Predictions': Y_pred})
```

	Actuals	Predictions
23	9.301030	8.657282
29	5.080922	5.380648
30	5.031517	5.293212
32	4.744727	5.153417
44	4.301030	5.746938
...	...	...
1937	5.337242	5.409336
1946	5.267606	5.756796
1947	5.443697	5.756796
1950	5.408935	5.756796
1951	5.086186	7.682671

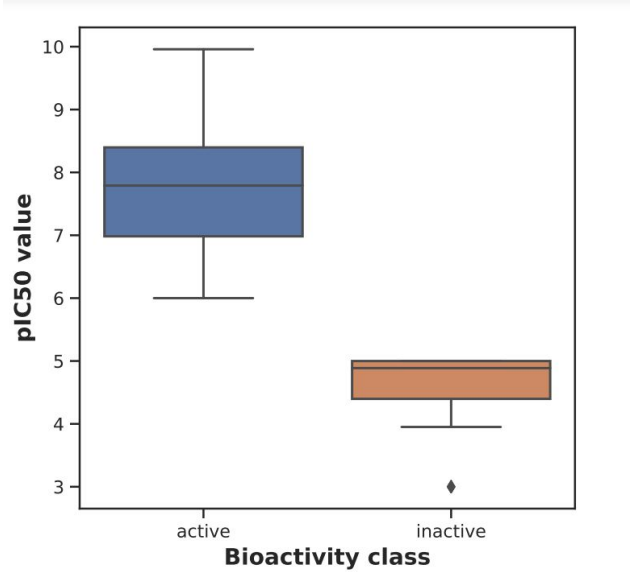
392 rows × 2 columns



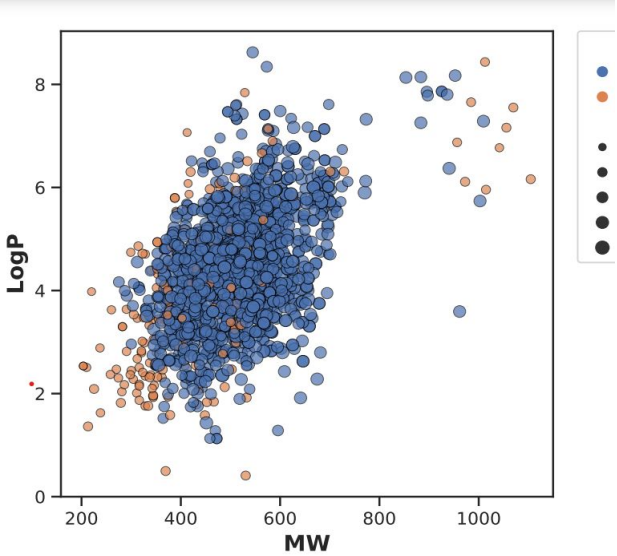
Results and Discussion

Exploratory Data Analysis plots

Bioactivity vs pIC50 plot



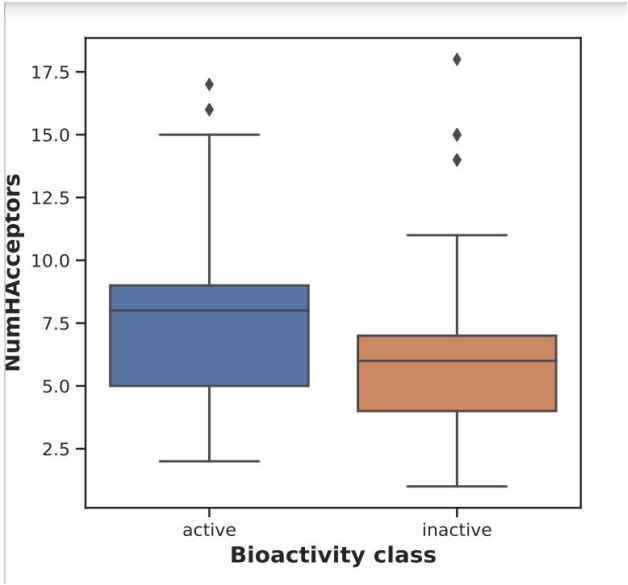
MW vs LogP plot



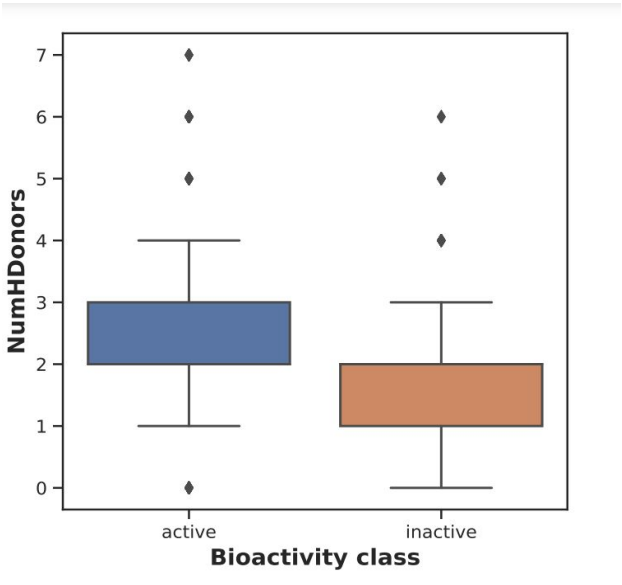
Results and Discussion

Exploratory Data Analysis plots

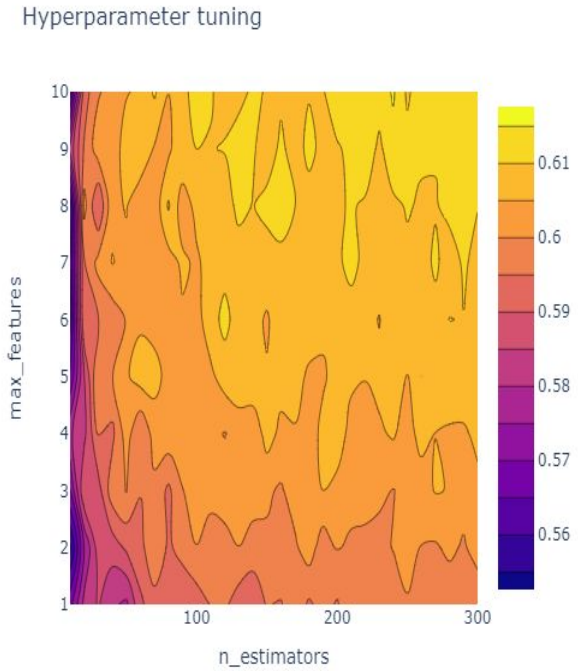
Bioactivity vs NumHAcceptors plot



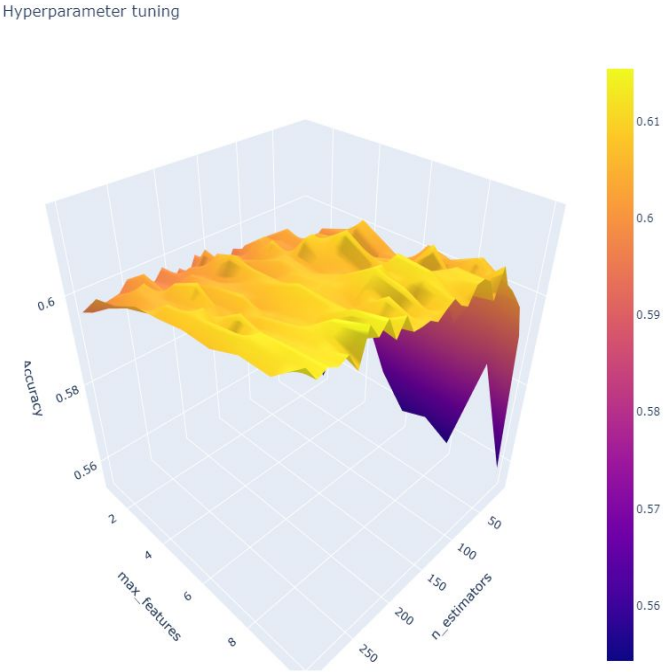
Bioactivity vs NumHDonors plot



Results and Discussion



2D Contour Plot



3D Surface Plot

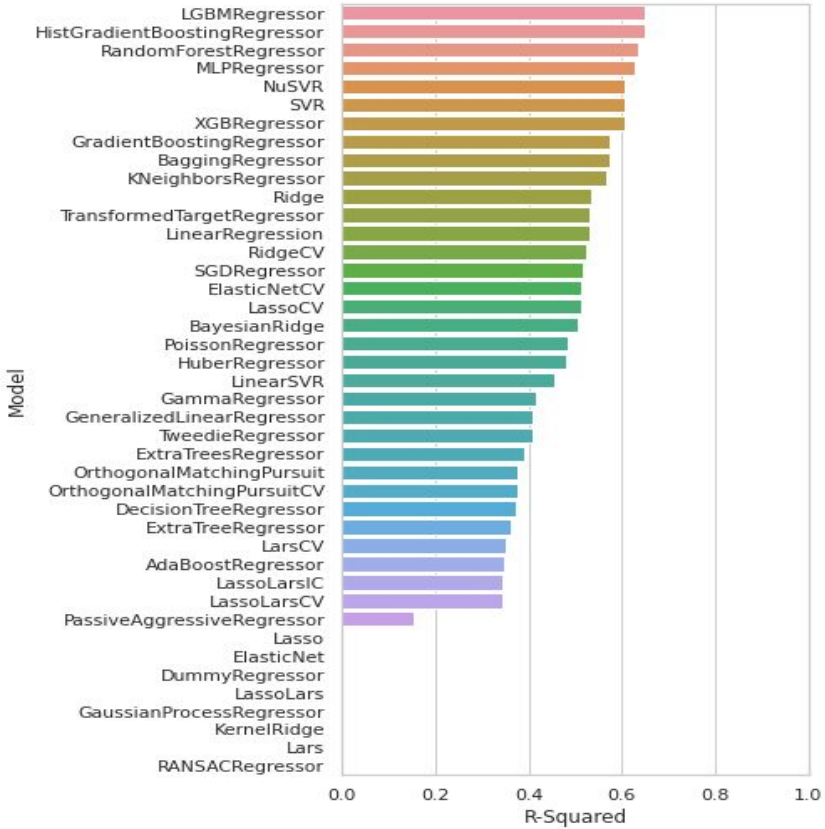
Prediction output

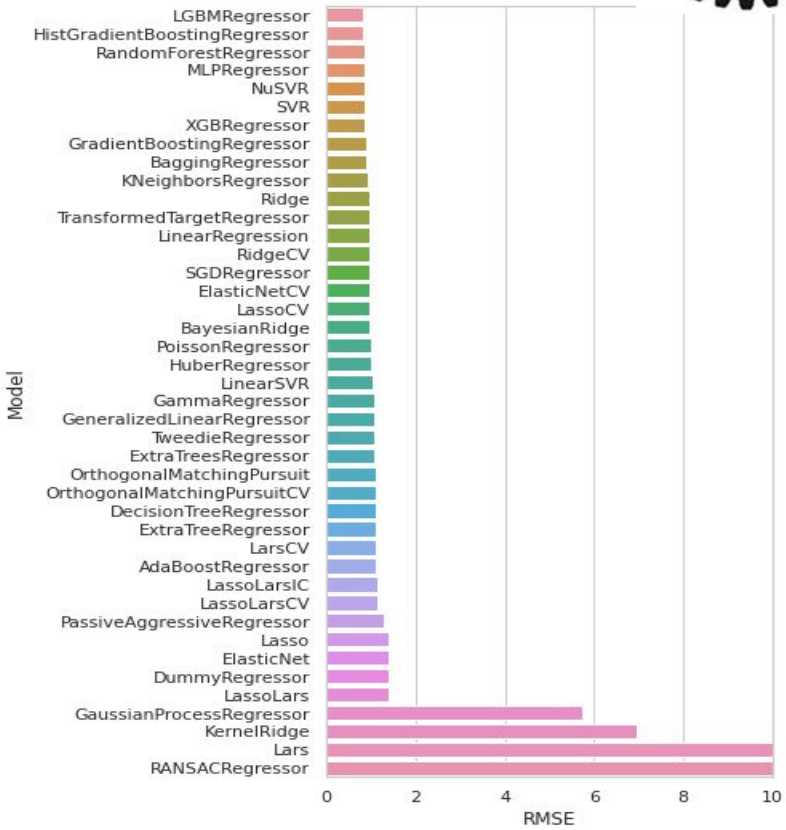
	Molecule_name	pIC50
0	CHEMBL867052	5.5006
1	CHEMBL867052	5.0719
2	CHEMBL867052	5.1835
3	CHEMBL867052	4.8830
4	CHEMBL867052	5.7715
5	CHEMBL867052	4.6718
6	CHEMBL867052	5.2923
7	CHEMBL867052	6.0569
8	CHEMBL867052	5.4651
9	CHEMBL867052	5.3369

Download Predictions

pIC50 values of drugs predicted

Results and Discussion





Application/Relevance

Anaplastic Lymphoma Kinase is an enzyme encoded by the ALK gene. Mutated forms of the ALK gene and protein have been found in non-small cell lung cancer, anaplastic large cell lymphoma and neuroblastoma. Alk-positive lung cancer occurs in about **5%** of all lung cancer patients and are usually metastatic. Without treatment, the expectancy of patients is within 12 months. With research, improved treatments and better medicines are being found out.

Various ML algorithms are used to find the best model for predicting the best chemical compound for acting against ALK. Bioinformatic analysis can not only accelerate drug target identification and drug candidate screening and refinement, but also facilitate characterization of side effects and predict drug resistance. High-throughput data such as genomic, epigenetic, transcriptomic, proteomic, and ribosome profiling data have all made significant contribution to mechanism-based **drug discovery** and **drug repurposing**. Moreover, bioinformatics has also innovated personalised medicine research thus bringing new discoveries in terms of drugs that can be personalized to someone's genetic pattern.

Conclusion and Future Work

The methodical process through which new candidate medications are found is known as drug discovery. It is a difficult, risky, time-consuming, yet potentially very profitable process. Bioinformatic analysis can speed up the identification of therapeutic targets, the screening of drug candidates, and the refinement of those candidates. It can also make it easier to characterize side effects and anticipate drug resistance.

To create a model, various ML techniques are applied to the training set of data. The best chemicals (based on pIC50) acting on ALK to inhibit cancer growth are then found using the model.

Prediction, identification, and storage of data relating to physiologically active candidates are the main focuses of current bioinformatics techniques. To locate targets for drug repurposing and identify new therapeutics, we use data analysis and machine learning.

The process can be shortened by conducting additional research on potential therapeutic targets, resulting in speedier and more effective pharmaceuticals entering the market and saving lives.

Better Deep learning methods can be used for increasing the efficiency of the model. The performance of the rf value can be enhanced by taking more parameters for hyperparameter tuning.

References

1. Fabio Boniolo, Emilio Dorigatti, Alexander J. Ohnmacht, Dieter Saur, Benjamin Schubert & Michael P. Menden (2021), “Artificial intelligence in early drug discovery enabling precision medicine”, Expert Opinion on Drug Discovery, 2021 Sep; 16(9): 991-1007.

2. https://www.datacamp.com/tutorial/streamlit

3. Saw Simeon, Nuttapat Anuwongcharoen, Watshara Shoombuatong1 Aijaz Ahmad Malik, Virapong Prachayasittikul, Jarl E.S. Wikberg, Chanin Nantasenamat, “Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking”, PeerJ, August 9, 2016

4. Yasen Jiao and Pufeng Du, “Performance measures in evaluating machine learning based bioinformatics predictors for classifications”, Higher Education Press and Springer-Verlag Berlin Heidelberg 2016,Quantitative Biology 2016, 4(4): 320–330

5. H.C. Stephen Chan, Hanbin Shan, Thamani Dahoun, Horst Vogel, Shuguang Yuan, “Advancing Drug Discovery via Artificial Intelligence”, Trends in Pharmacological Sciences, Volume 40, Issue 8, August 2019, Pages 592-604.

6. Choudhury C, Arul Murugan N, Priyakumar UD, “Structure-based drug repurposing: Traditional and advanced AI/ML-aided methods”, Drug Discovery Today. 2022 Mar 14:S1359-6446(22)00112-X.

7. Pillai N, Dasgupta A, Sudsakorn S, Fretland J, Mavroudis PD, “Machine Learning guided early drug discovery of small molecules” Drug Discovery Today. 2022 Mar 29:S1359-6446(22)00127-1

8. Chandak T, Wong CF, “EDock-ML: A web server for using ensemble docking with machine learning to aid drug discovery”, Protein Science, 2021, 30:1087–1097.

9. Leonardo L.G.Ferreira, Adriano D.Andricopulo, “ADMET modeling approaches in drug discovery”, Drug Discovery Today, Volume 24, Issue 5, May 2019, Pages 1157-1165.

10. Olayan, Rawan S., Haitham Ashoor, and Vladimir B. Bajic. "DDR: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches." Bioinformatics 34.7 (2018): 1164-1173.

11. Sullivan, Ivana, and David Planchard. "ALK inhibitors in non-small cell lung cancer: the latest evidence and developments." *Therapeutic advances in medical oncology* 8.1 (2016): 32-47.

12. https://medlineplus.gov/genetics/gene/alk/

“Thank You”