Machine Learning in Breast Cancer Drug Discovery

Presented By - VAIBHAV RAMAKRISHNAN

PROJECT BACKGROUND

Context:

- Addressing challenges in breast cancer medication discovery with bioinformatics and machine learning.

Objectives:

- Develop a tailored QSAR model for medication research.
- Targeted data gathering on the aromatase protein.
- Create a user-friendly web application for model deployment.
- Use Random Forest algorithm for simple and interpretable candidate identification.

Project Duration and Deliverables

October 2023 – December 2023

- Data Collection using ChEMBL bioactivity database
- QSAR Modeling
- Model Comparison
- Web Application Development

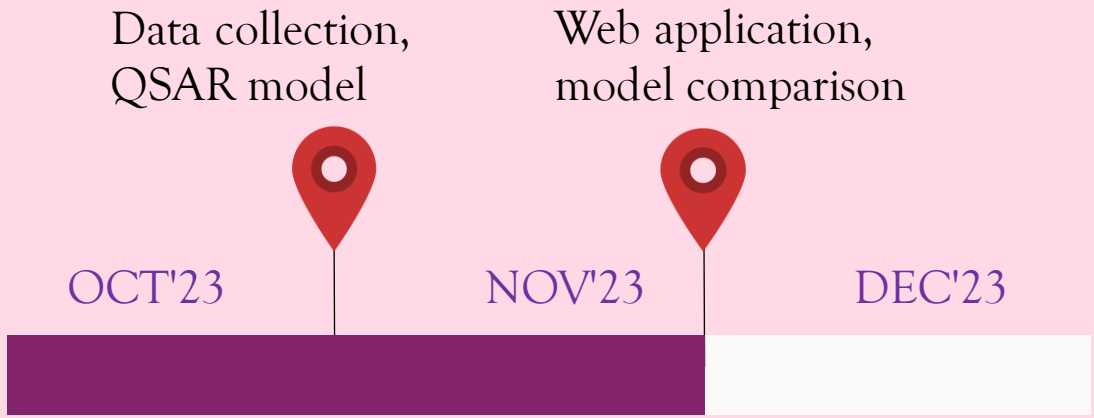

Fig. 1 – Project Timeline in 2023

My Role and Contributions

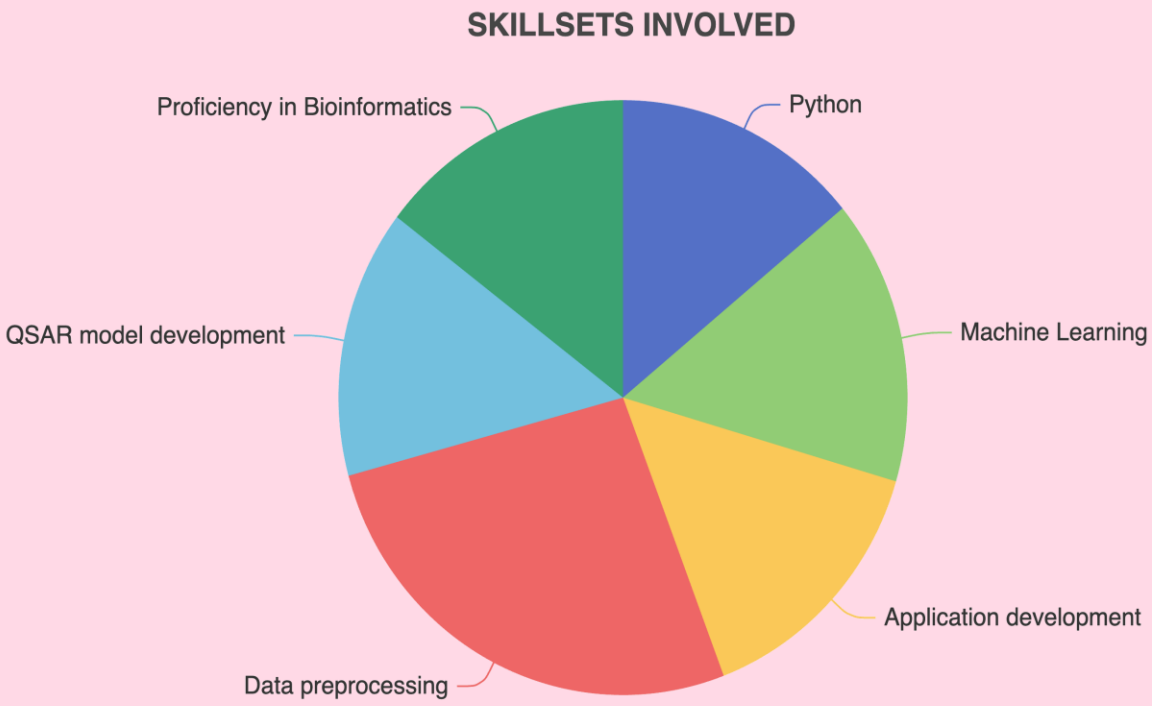

Fig. 2 – Overview of Skills developed

- Conducted data gathering efforts focusing on the aromatase protein.
- Developed and fine-tuned the QSAR model using Random Forest.
- Designed a user-friendly web application.
- Conducted comparative analysis of machine learning techniques.

Results

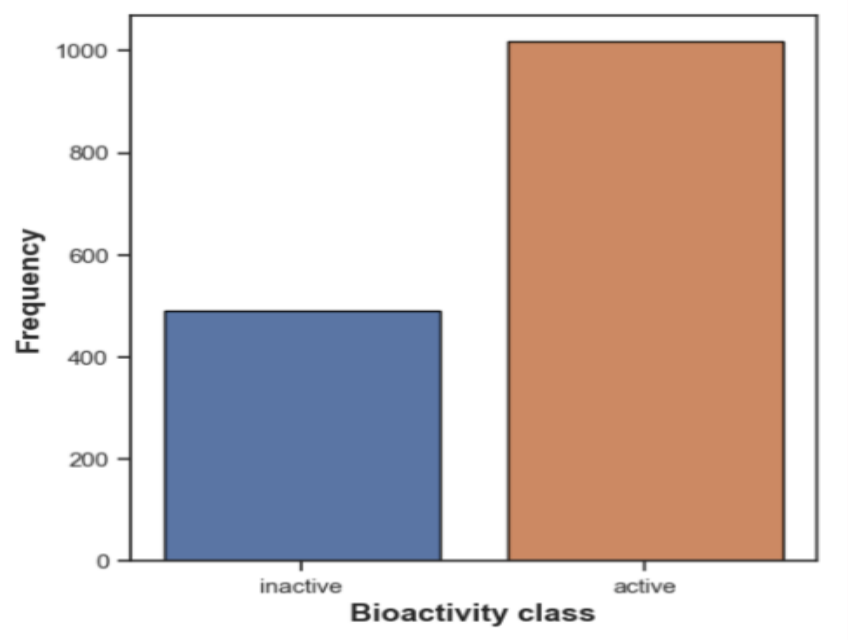

Fig. 3 – Bioactivity class distribution (Active v/s Inactive)

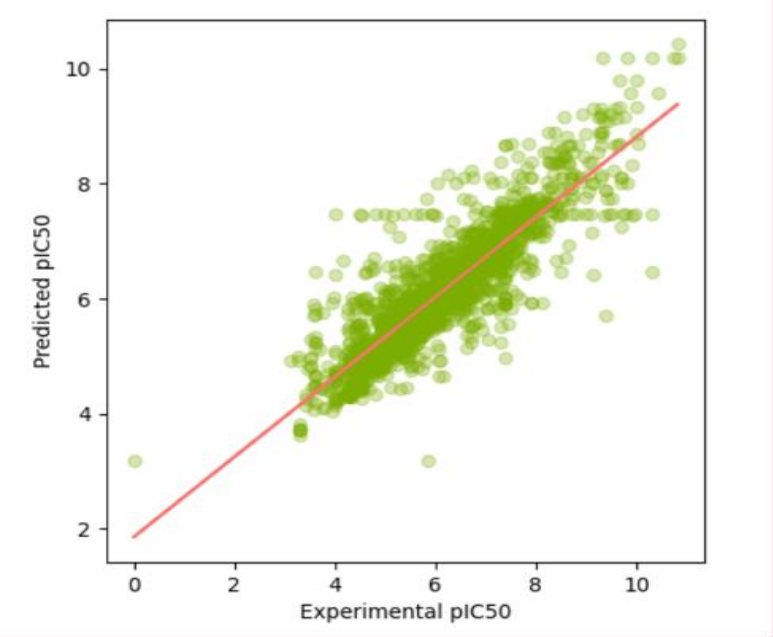

Fig. 4 – Scatterplot of Predicted v/s Experimental pIC50 values for the Random Forest Model

- Random Forest model achieved an impressive R-squared value of 76%.
- Model exhibited competitive performance with an MSE score of 0.40.
- Mastering SMILES notation facilitated model development.
- Our user-friendly application revolutionizes breast cancer drug discovery.

Key Lessons

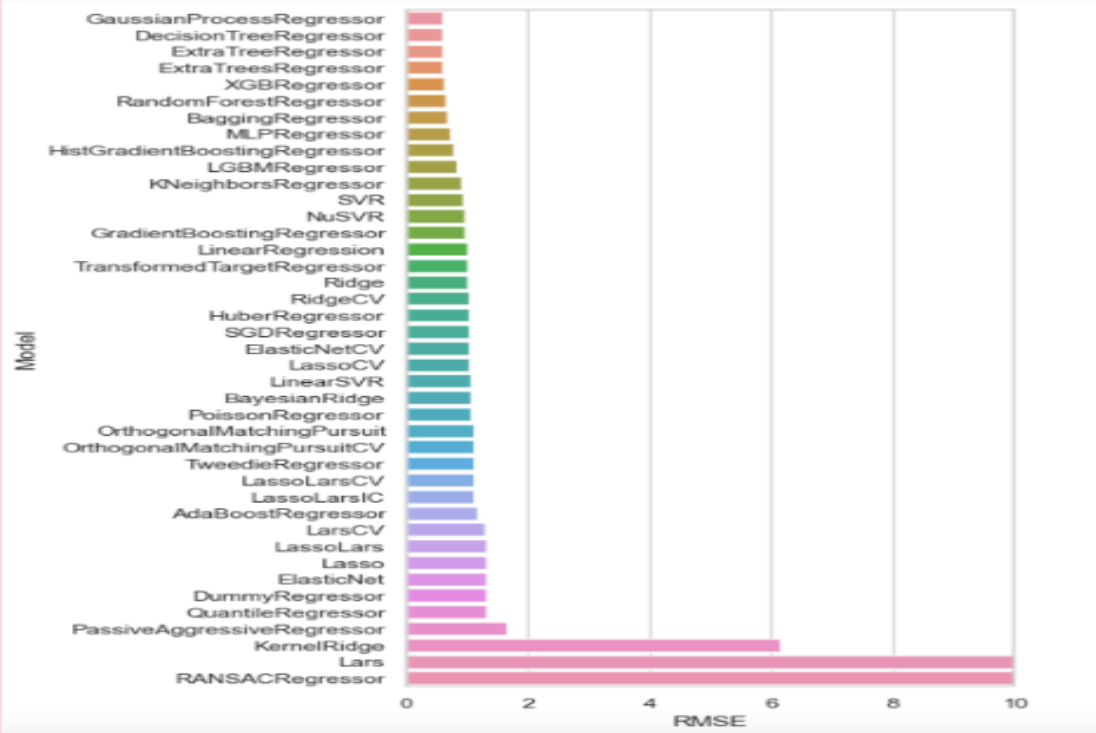

Fig. 5 – Comparing ML models: RMSE Performance

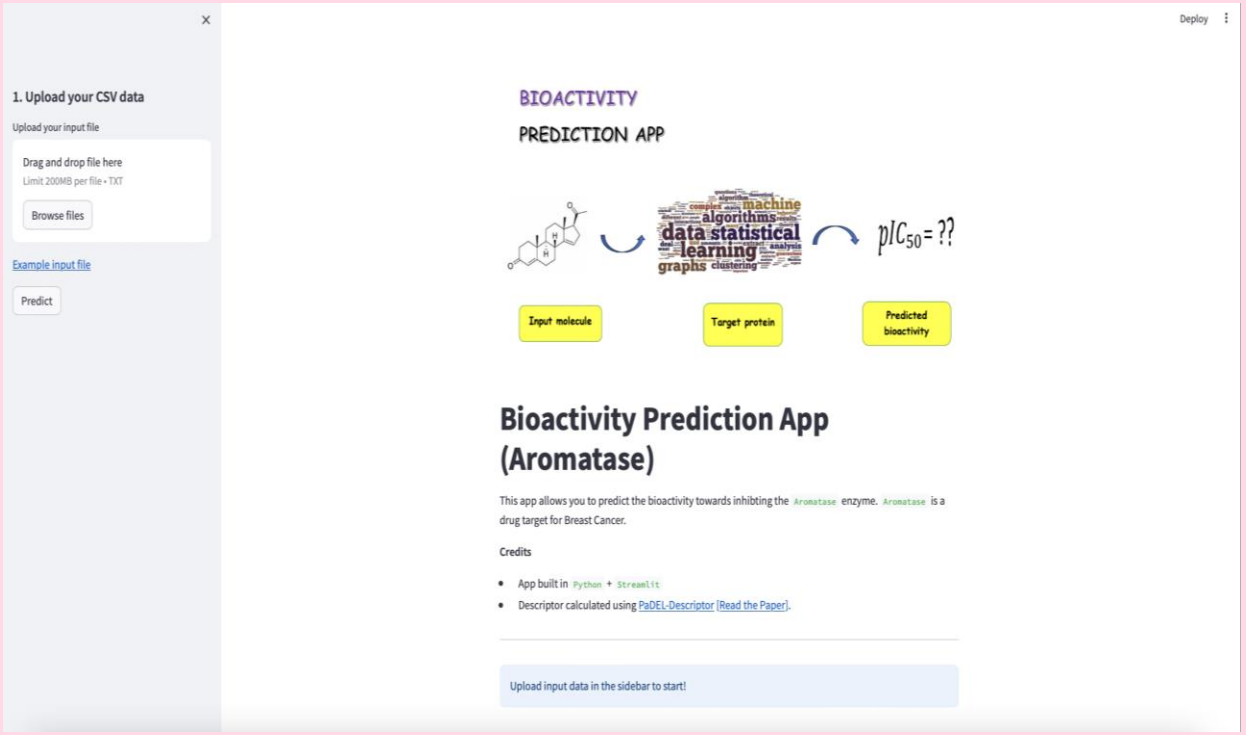

Fig. 6 – Bioinformatics web application homepage

- **Creative Model Development**: Overcoming dataset complexities led to innovative outcomes.
- **Adaptable Problem-Solving**: Flexibility in addressing challenges was key to success.
- **Insightful Algorithm Analysis**: Strategic comparisons guided effective model selection (see Fig. 5).
- **User-Focused Design**: Prioritizing user needs ensured an intuitive bioactivity prediction tool (see Fig. 6).

Relevance to Focus Area: Information Technology

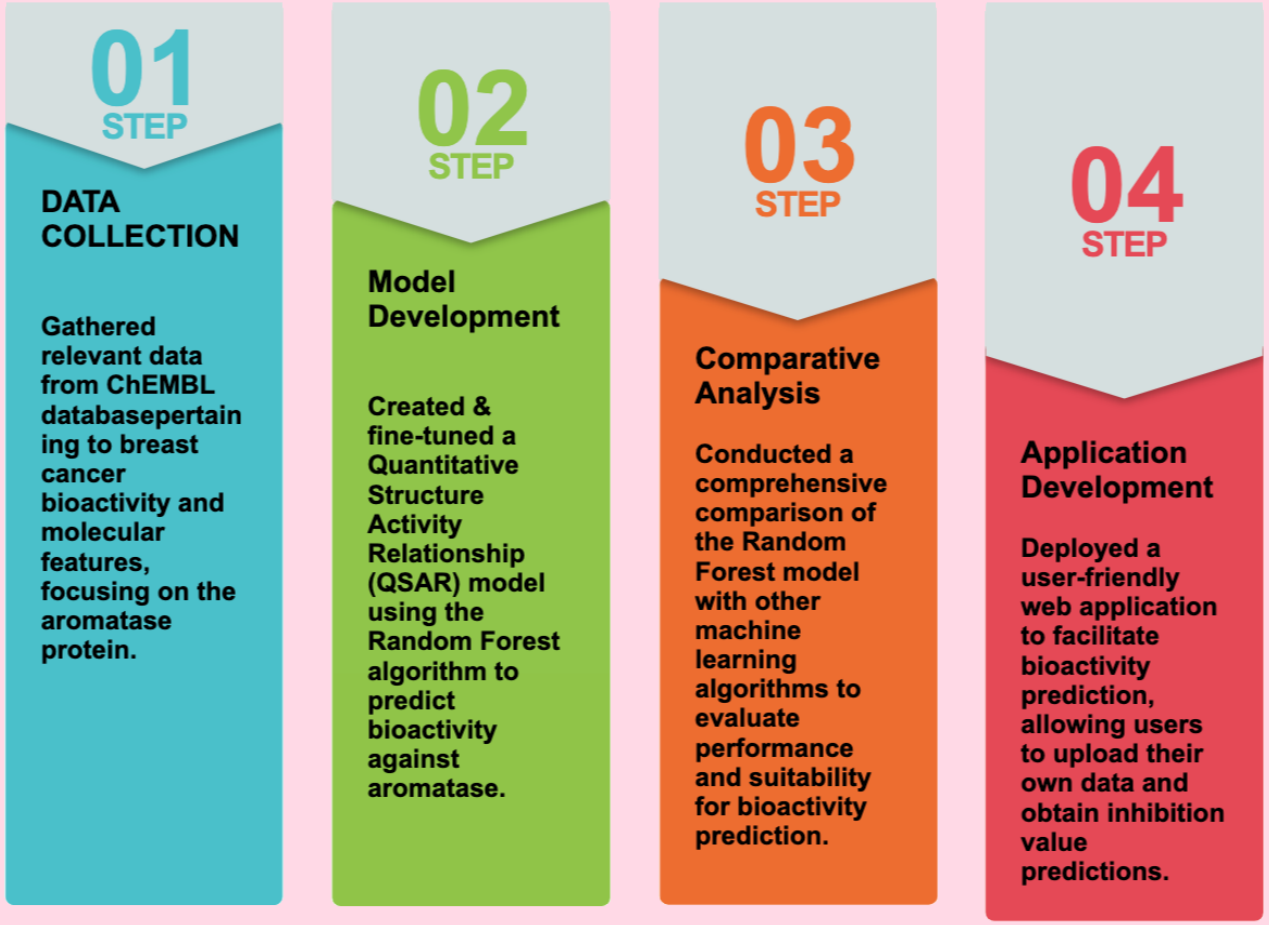

Fig. 7 – Project Workflow

- **Integration of IT and Data Science:** This project merges IT principles with advanced data science methods.
- **Role of Data Analytics:** Heavy reliance on data analytics techniques for preprocessing, modeling & analysis.
- **Skills Development:** Enhanced proficiency in machine learning, stats, and predictive modeling.
- **Data-Driven Decision-Making:** Emphasis on leveraging data for effective decision-making in IT.

Appendix

example_aromatase.txt

```
CCn1c(C(c2ccc(F)cc2)n2ccnc2)c(C)c2cc(Br)ccc21 CHEMBL431859
CCn1cc(C(c2ccc(F)cc2)n2ccnc2)c2ccccc21 CHEMBL113637
Clc1ccccc1Cn1cc(Cn2ccnc2)c2ccccc21 CHEMBL112021
CCn1ccc2cc(C(c3ccc(F)cc3)n3ccnc3)ccc21 CHEMBL41761
Cn1cc(C(c2ccc(F)cc2)n2ccnc2)c2cc(Br)ccc21 CHEMBL111868
```

Fig. 8 – Input File Contents for Application

GitHub Project Link
https://github.com/vaibhavramakrishnan/bioinformatics

Calculated molecular descriptors

	Name	PubchemFP0	PubchemFP1	PubchemFP2	PubchemFP3	PubchemFP4	PubchemF
0	CHEMBL113637	1	1	1	0	0	
1	CHEMBL111868	1	1	0	0	0	
2	CHEMBL41761	1	1	1	0	0	
3	CHEMBL112021	1	1	1	0	0	
4	CHEMBL431859	1	1	1	0	0	

(5, 882)

Fig. 9 – Molecular descriptors calculated using the Application

	molecule_name	pIC50
0	CHEMBL431859	6.0124
1	CHEMBL113637	5.8233
2	CHEMBL112021	5.1389
3	CHEMBL41761	5.2041
4	CHEMBL111868	5.5643

Download Predictions

Fig. 10 – Prediction output from the Application

Appendix (Continued..)

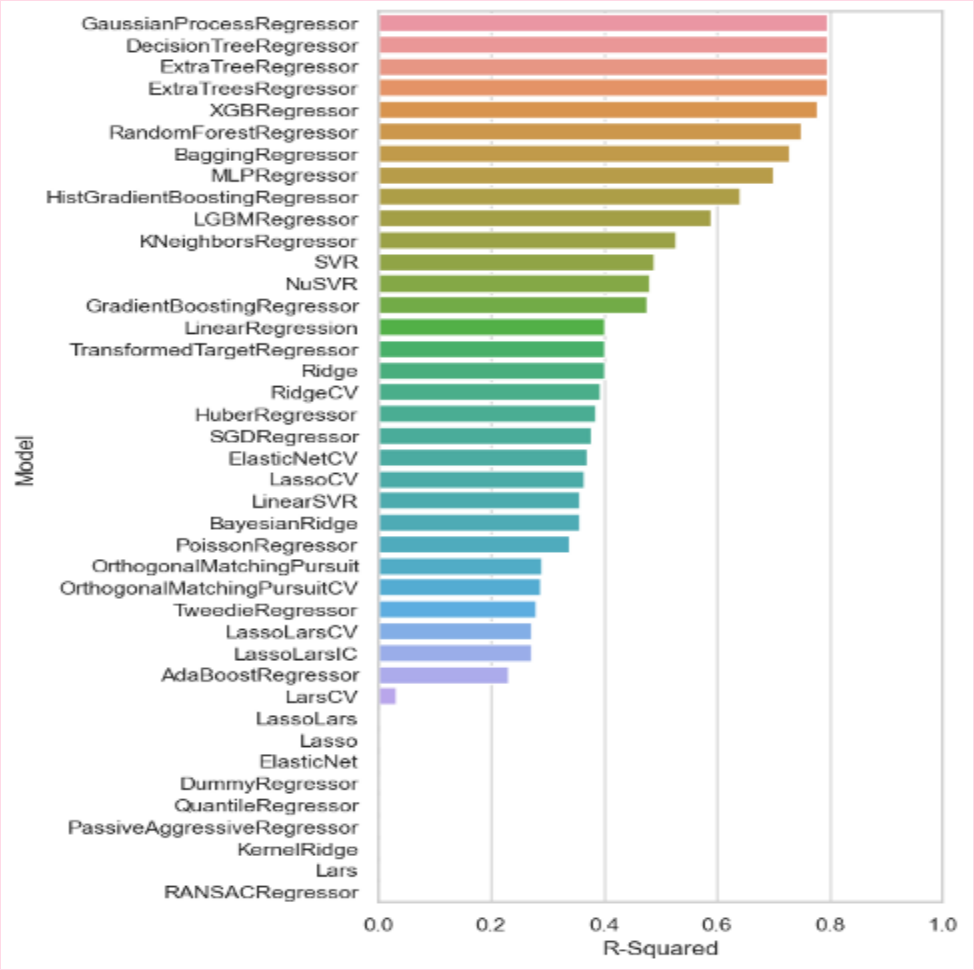

Fig. 11 – Training Set R-Squared Values

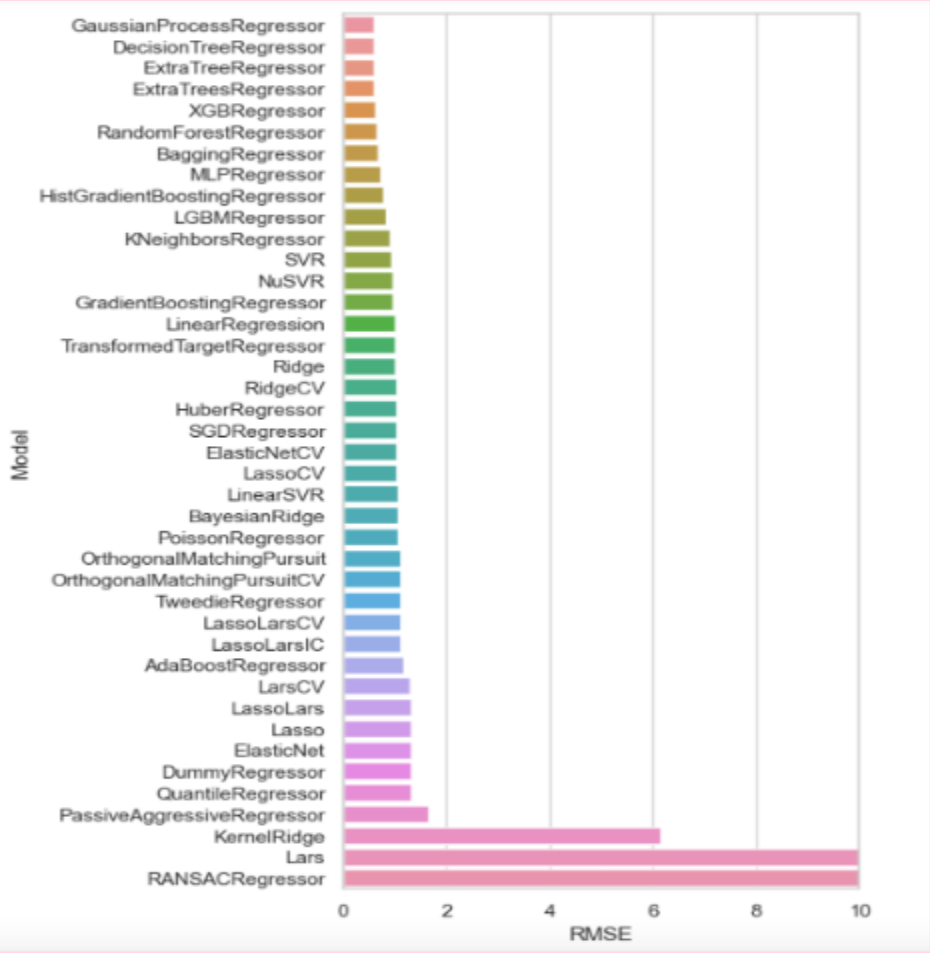

Fig. 12 – Training Set RMSE Values

Appendix (Continued..)

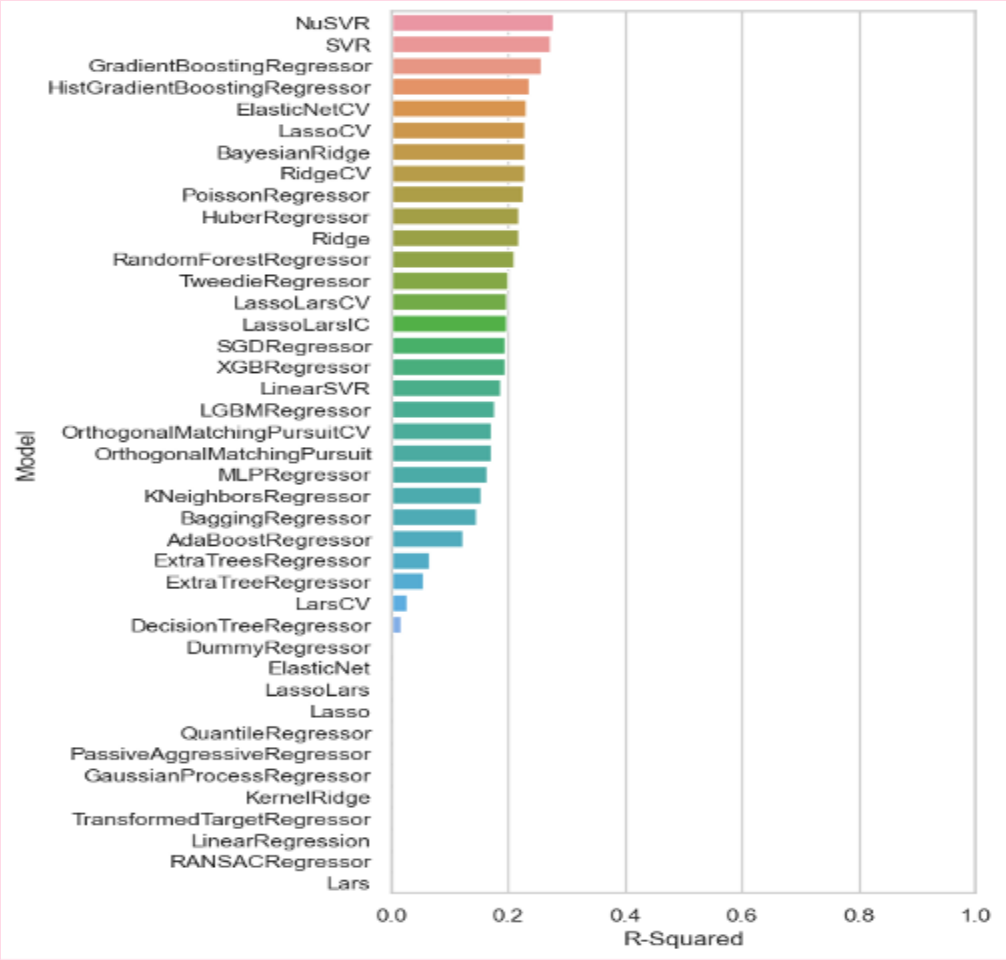

Fig. 13 – Test Set R-Squared Values

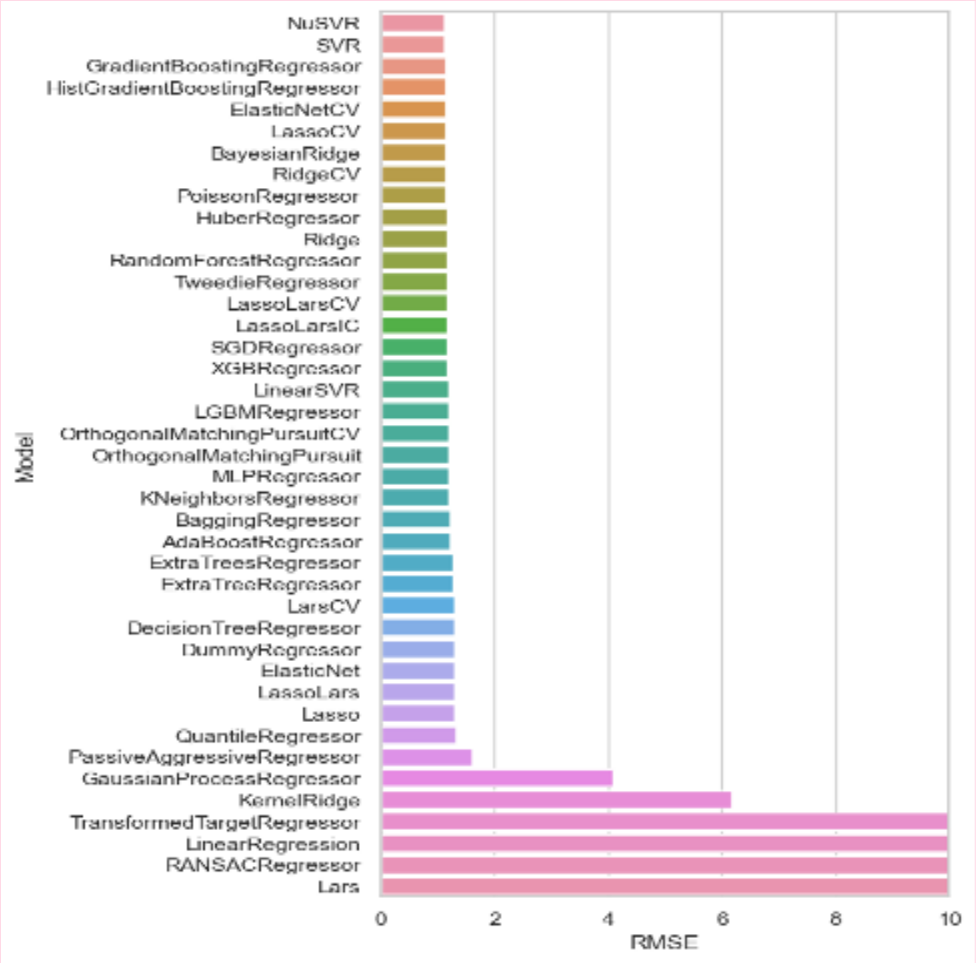

Fig. 14 – Test Set RMSE Values

Appendix (Continued..)

jupyter CDD_ML_Part_1_Aromatase_Bioactivity_Data_Concised Last Checkpoint: 30/12/2023 (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Markdown

Search for Target protein

Target search for Aromatase

In [4]:
```
# Target search for coronavirus
target = new_client.target
target_query = target.search('aromatase')
targets = pd.DataFrame.from_dict(target_query)
targets
```

Out[4]:

	cross_references	organism	pref_name	score	species_group_flag	target_chembl_id	target_components	target_type	tax_id
0	[{'xref_id': 'P11511', 'xref_name': None, 'xre...	Homo sapiens	Cytochrome P450 19A1	20.0	False	CHEMBL1978	[{'accession': 'P11511', 'component_descriptio...	SINGLE PROTEIN	9606
1	[{'xref_id': 'P22443', 'xref_name': None, 'xre...	Rattus norvegicus	Cytochrome P450 19A1	20.0	False	CHEMBL3859	[{'accession': 'P22443', 'component_descriptio...	SINGLE PROTEIN	10116

Select and retrieve bioactivity data for *Human Aromatase* (first entry)

Fig. 15 – ChEMBL dataset accessed via Jupyter Notebook

Handling missing data

If any compounds has missing value for the **standard_value** and **canonical_smiles** column then drop it.

In [9]:
```
df2 = df[df.standard_value.notna()]
df2 = df2[df.canonical_smiles.notna()]
df2
```

```
/var/folders/v0/fdzcqv555678rbr0psghgcnm0000gn/T/ipykernel_87683/3852201246.py:2: UserWarning: Boolean Series key will be reindexed to match DataFrame index.
  df2 = df2[df.canonical_smiles.notna()]
```

Out[9]:

	action_type	activity_comment	activity_id	activity_properties	assay_chembl_id	assay_description	assay_type	assay_variant_accession	assay_variant_
0	None	None	82585	[]	CHEMBL666794	Inhibition of Cytochrome P450 19A1	B	None	
1	None	None	94540	[]	CHEMBL666794	Inhibition of Cytochrome P450 19A1	B	None	
2	None	None	112960	[]	CHEMBL661700	In vitro inhibition of human Cytochrome P450 19A1	B	None	
3	None	None	116766	[]	CHEMBL661700	In vitro inhibition of human Cytochrome P450 19A1	B	None	
4	None	None	118017	[]	CHEMBL661700	In vitro inhibition of human Cytochrome P450 19A1	B	None	
...	...	...	...	...	...	...	...	...	
2961	{'action_type': 'INHIBITOR', 'description': 'N...	None	24742461	[{'comments': None, 'relation': '=', 'result_f...	CHEMBL5118295	Inhibition of aromatase in human JEG-3 cells u...	B	None	
2962	None	None	24783443	[]	CHEMBL5130158	Inhibition of human placental microsome CYP19 ...	A	None	
2963	{'action_type': 'INHIBITOR', 'description':	None	24886565	[]	CHEMBL5157477	Inhibition of aromatase	B	None	

Fig. 16 – ChEMBL dataset preprocessing

Appendix (Continued..)

Descriptor	Statistics	p	alpha	Interpretation
MW	284791.0	0.000018	0.05	Different distribution (reject H0)

Fig. 17 – Mann-Whitney U Test for active v/s inactive class for molecular weight (MW)

Descriptor	Statistics	p	alpha	Interpretation
LogP	252241.5	0.843704	0.05	Same distribution (fail to reject H0)

Fig. 18 – Mann-Whitney U Test for active v/s inactive class for solubility (LogP)

Descriptor	Statistics	p	alpha	Interpretation
NumHDonors	234631.0	0.023067	0.05	Different distribution (reject H0)

Fig. 19 – Mann-Whitney U Test for active v/s inactive class for Hydrogen Bond Donors (NumHDonors)

Descriptor	Statistics	p	alpha	Interpretation
NumHAcceptors	270837.0	0.009792	0.05	Different distribution (reject H0)

Fig. 20 – Mann-Whitney U Test for active v/s inactive class for Hydrogen Bond Acceptors (NumHAcceptors)

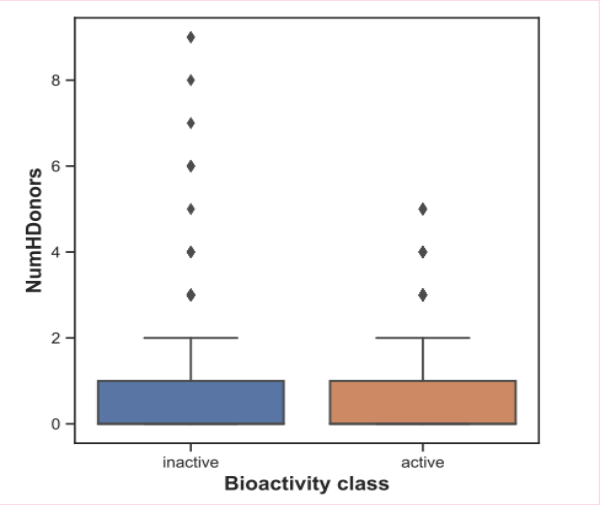

Fig. 21 – Boxplot of active v/s inactive class for NumHDonors