PREDICTION OF BIOLOGICAL ACTIVITY OF DRUG COMPOUNDS USING MACHINE LEARNING TECHNIQUES

CP302 - CAPSTONE PROJECT

Supervisor:
Dr. Vishwajeet Mehandia

Presented by:
Rahul Kumar Saw (2021CHB1052)

Introduction to Drug Bioactivity Prediction

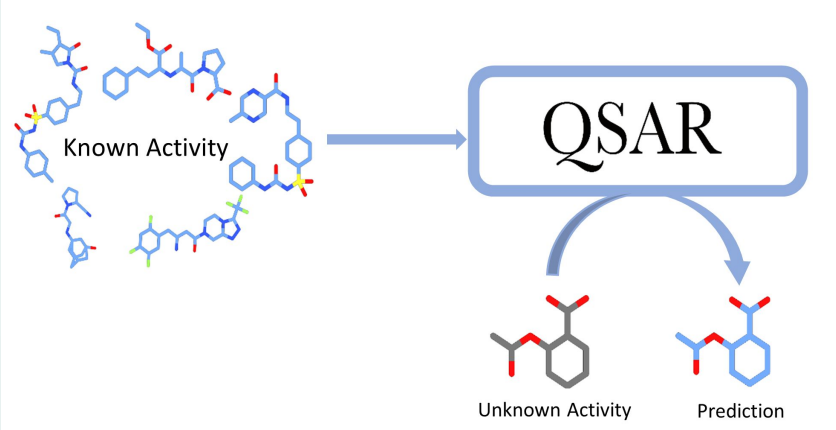
- Drug discovery & development - requires effective methods for screening of drug candidates for a given target.
- QSAR modeling - reveals the relationship between the structural properties of chemical compounds and biological activities.
- Molecular Descriptors/Fingerprints - mathematical representations of molecules' properties that are generated by algorithms. Eg. - **MW**, $\mathbf{N_{atoms}}$, $\mathbf{N_{H\text{-}Donors}}$

Objective - To predict the biological activity of drug compounds from molecular descriptors for the target protein - 'Prostaglandin E Synthase'

https://i.pinimg.com/originals/aa/12/c8/aa12c854b2f964aff88a00b11ab59178.jpg

QSAR Modeling

- SMILES - encodes the structural information of chemical compounds using text strings. (Ethanol - ‘CCO’)
- Lipinski Descriptors & PaDEL Descriptor
- Statistical Modeling Approaches - regression analysis, machine learning algorithms (random forest, decision trees, neural network, deep learning etc)



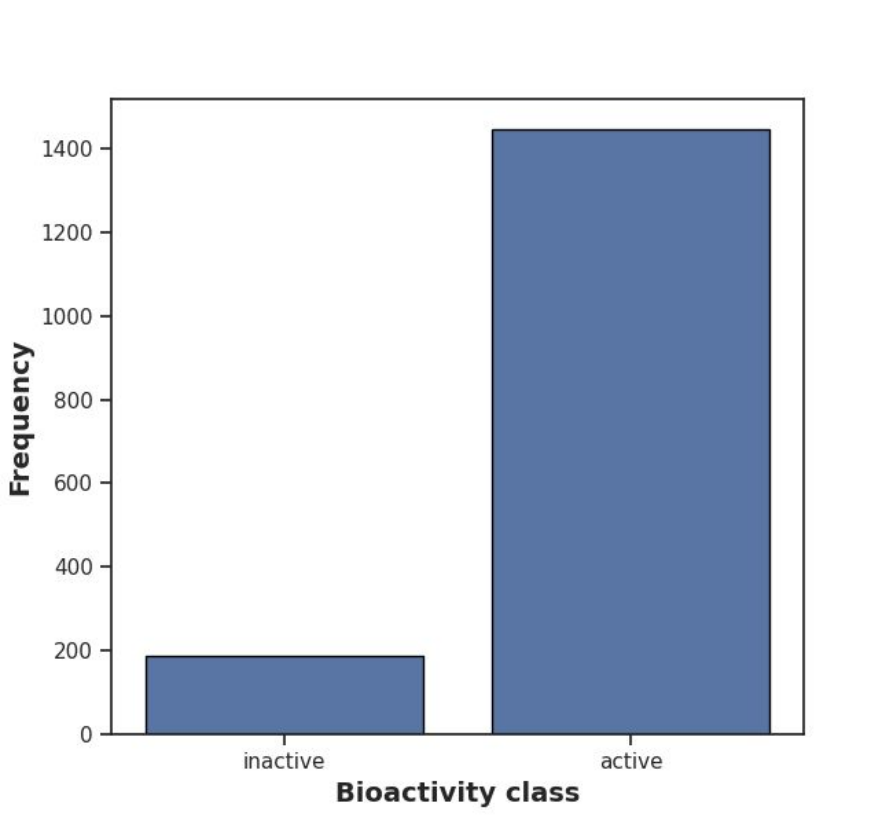
https://protoqsar.com/wp-content/uploads/2021/10/headerQSARen.png

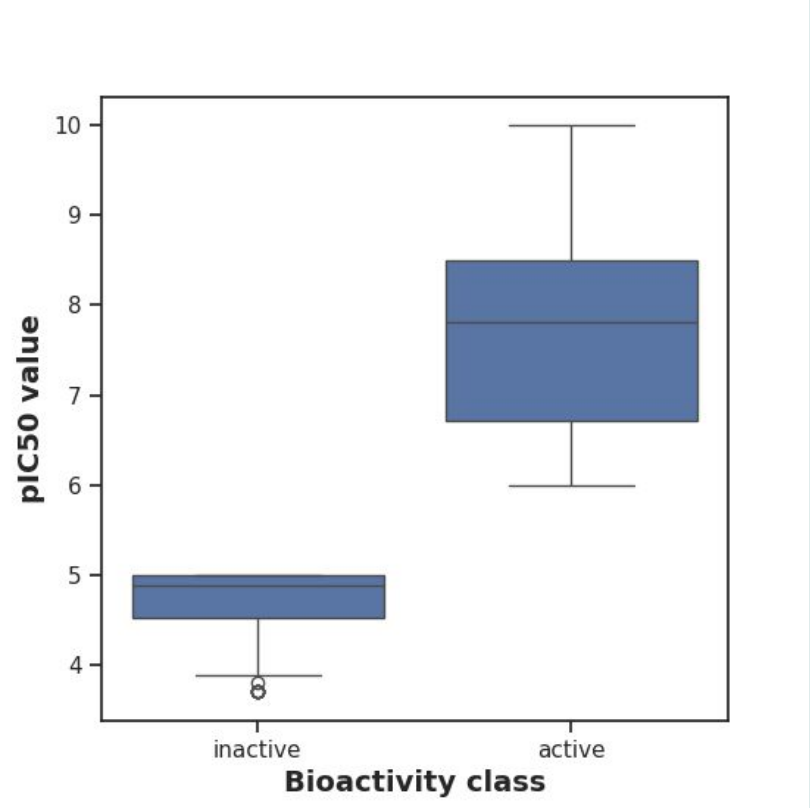
Data Collection & Preprocessing

Measure of drug Bioactivity: **IC50 value**
- Half-maximal inhibitory concentration (IC50) is the most widely used and informative measure of a drug's efficacy

- Source of Bioactivity Data - ChEMBL, Pubchem
- Selection of Target Protein and Compounds - ‘Prostaglandin E Synthase’
- Data Cleaning
- Calculation of Lipinski Descriptors
- Conversion of IC50 to pIC50 Values: pIC50 = -log(IC50)
- Generation of PubChem Fingerprints using PaDEL Descriptor

Data Analysis





Random Forest Regression: Methodology

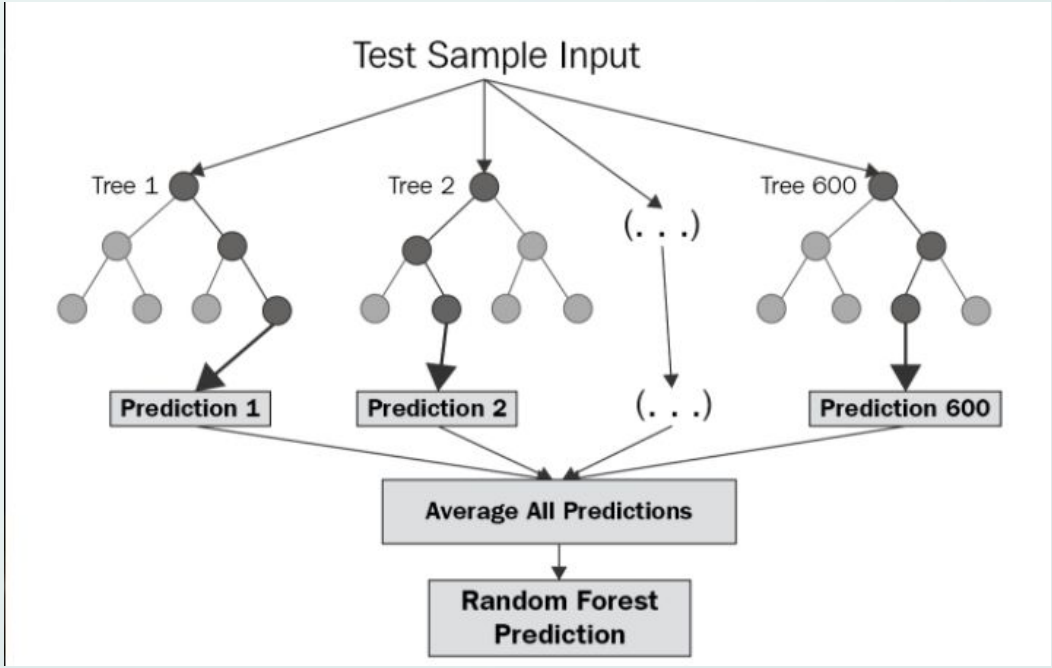
RandomForestRegressor (n_estimators, random_state)

- Ensemble Learning Method

Parameters:

n_estimators

random_state



https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

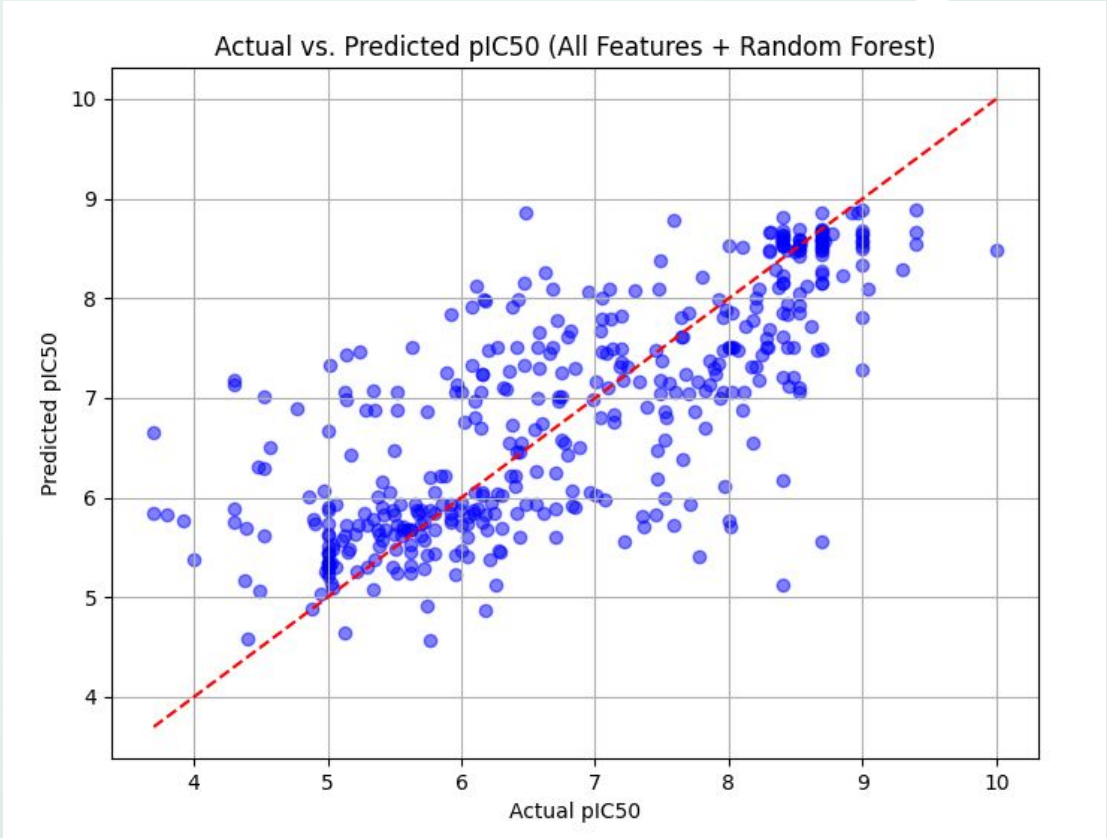
Random Forest Regression

No. of features = **882**

MSE: 0.89590

R2: 0.527949

AIC: 1762.28796



Input Feature Selection

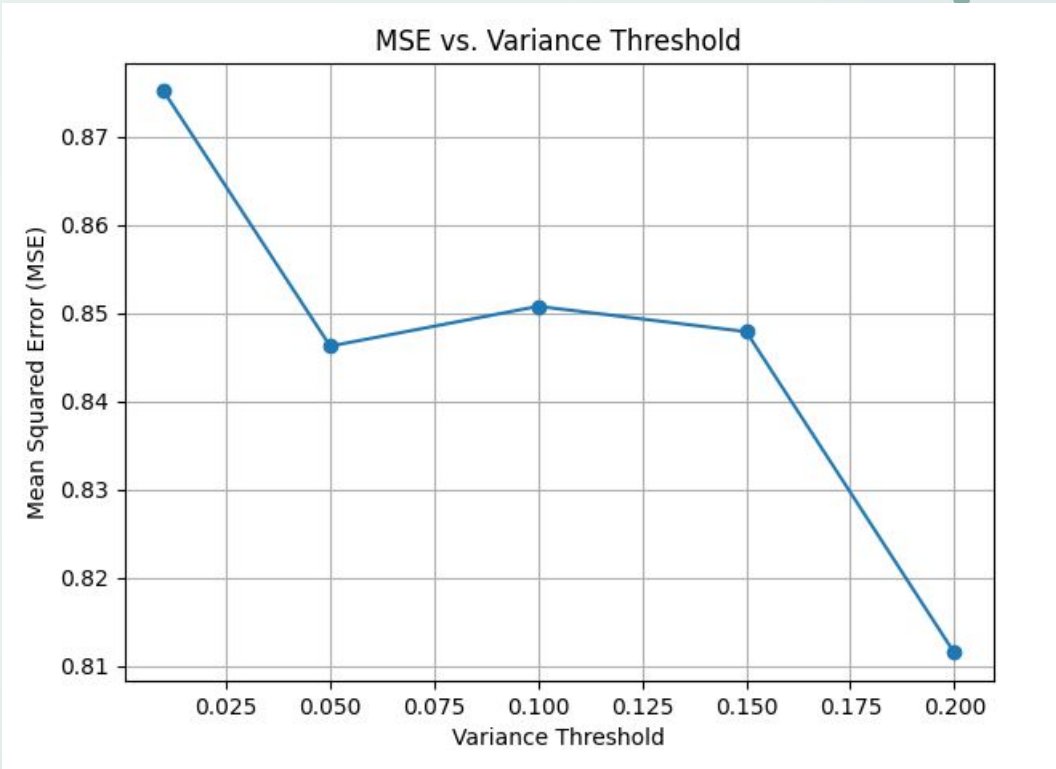
Initial no. of input features = 882

Removing the low variance features.

Best Threshold: 0.2

Best MSE: 0.81151667

Final number of input features: 131



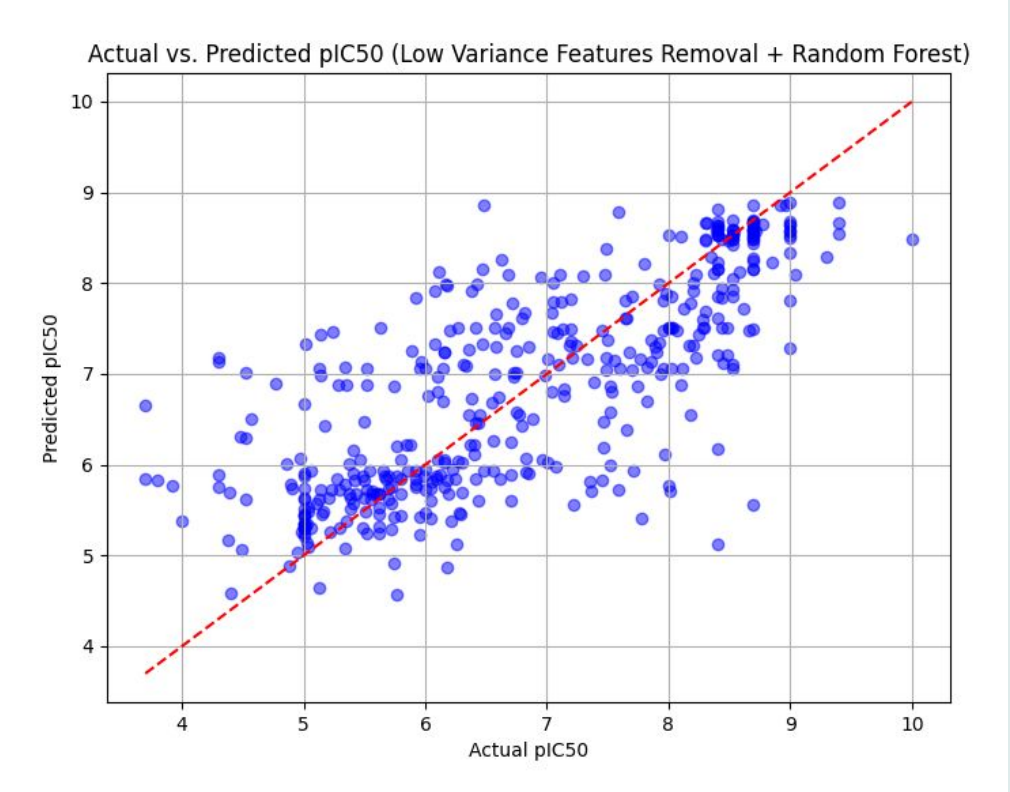
Predictions on Selected Features

No. of features = **131**

MSE: 0.813661

R2: 0.57522

AIC: 262.41242



Feature Selection Using Lasso

Initial no. of features= 131

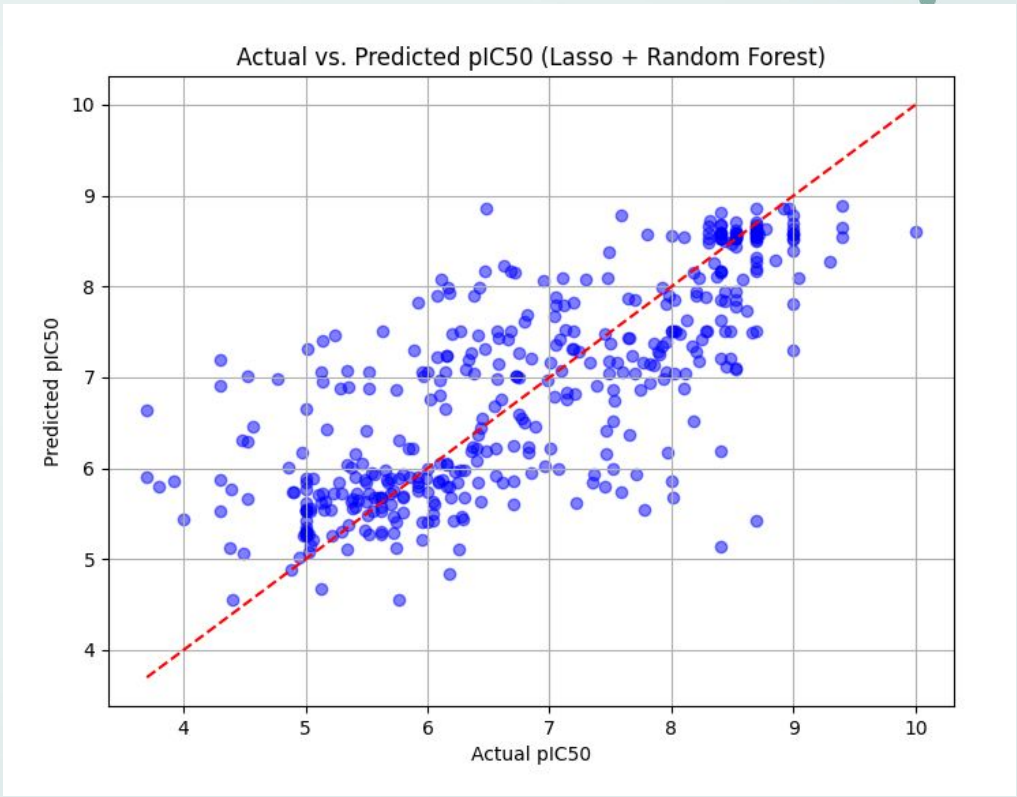
After Lasso Regression:

Final no. features = **92**

MSE: 0.805105

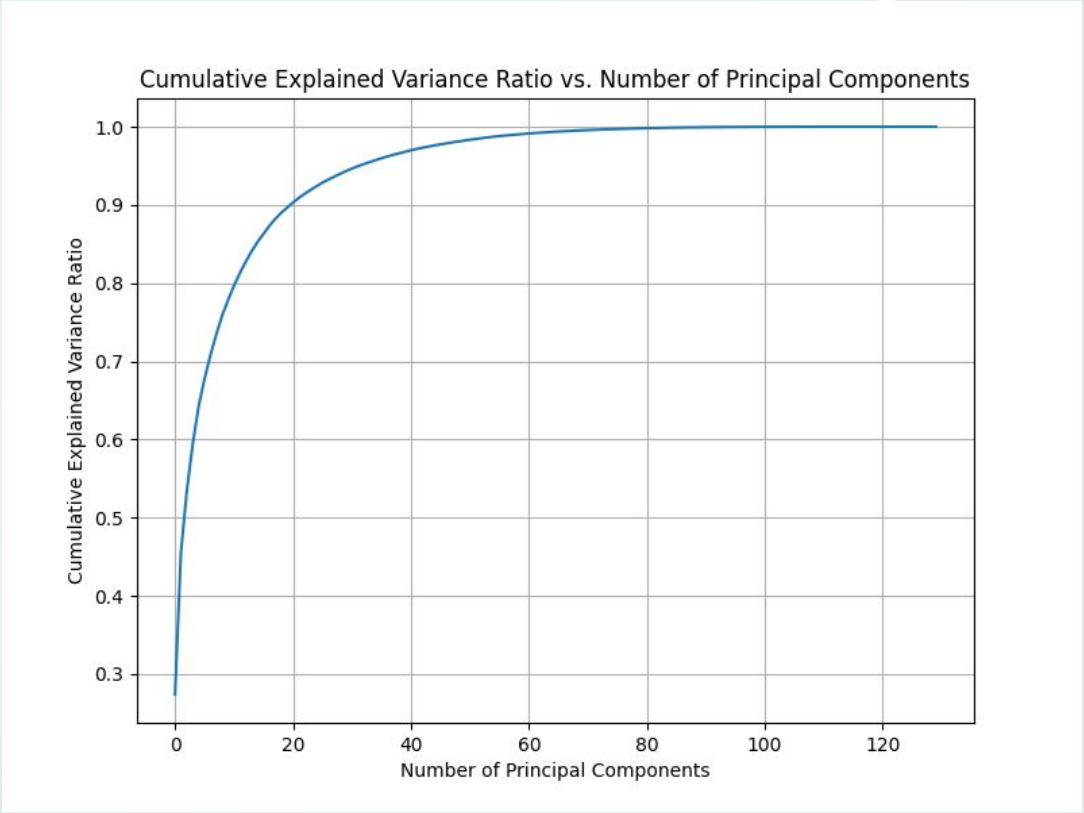
R2 Score = 0.599688

AIC = 184.43356



Feature Selection Using PCA

Number of principal components to explain the maximum variance in data: **50**



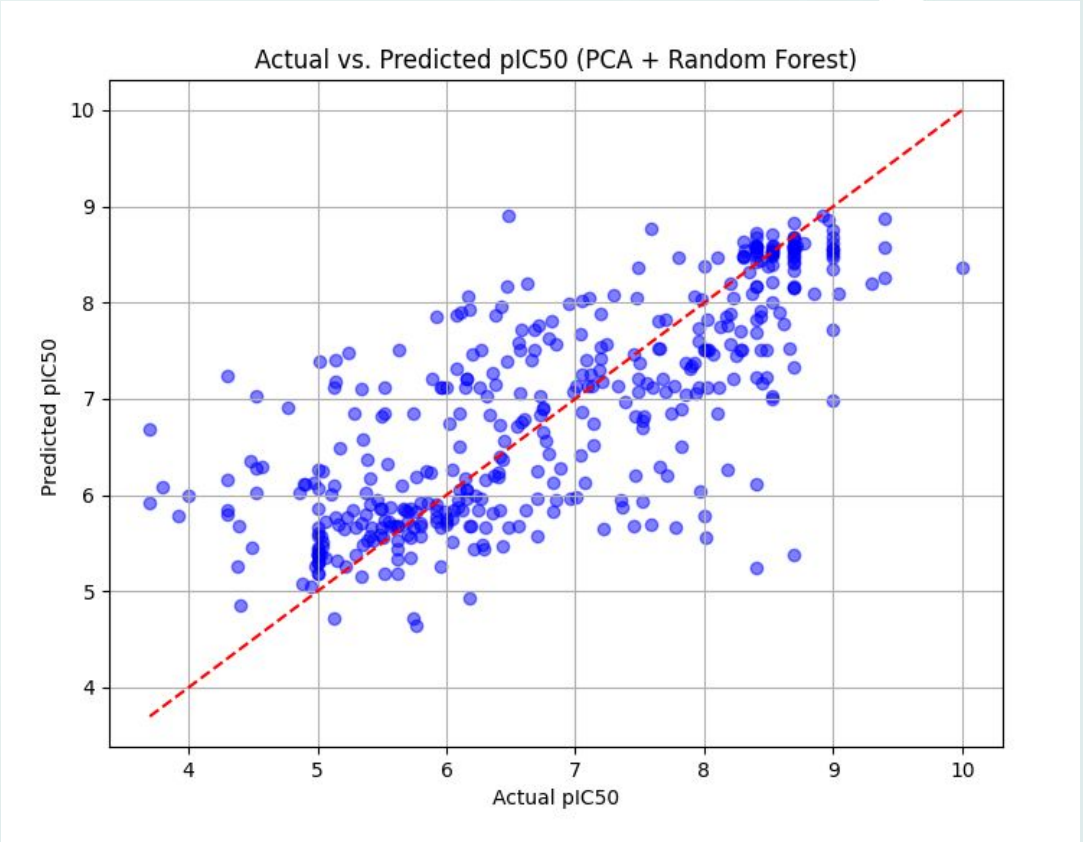
Predicting Using the Selected Features

Principal Components: **50**

MSE: 0.824266

R2 Score: 0.569685

AIC: 100.386522



Comparison of Performance

Model	MSE	R2 Score	AIC
RF with 882 Features	0.89590	0.52794	1762.28796
RF with 131 Features	0.81366	0.57522	262.41242
RF with 92 Features (Lasso)	0.80510	0.59968	184.43356
RF with 50 Features(PCA)	**0.80426**	**0.56968**	**100.38652**

Comparing the Performance of Different Algorithms

Model	MSE	R2 Score	AIC
Ridge Regression	0.8024	0.5811	100.4404
Lasso Regression	1.2424	0.3514	99.5659
Random Forest Regression	0.8154	0.5743	100.4082
Gradient Boosting	0.7564	0.6051	100.5583
K-Nearest Neighbors	0.7562	0.6052	100.5590
Decision Tree	1.1085	0.4213	99.7940
Support Vector Regression	**0.7136**	**0.6275**	**100.6750**

Conclusion

In conclusion, this project demonstrates the importance of QSAR in modeling in drug discovery.

Feature selection played a crucial role in optimizing model performance by identifying the most relevant input features. Random Forest was selected as the primary algorithm

Performance of the different algorithms were compared, SVR performed best amongst them.

Further research can be done using the combination methods like model stacking or advanced deep learning to uncover the complex non-linear relationships, and improve the prediction performance.

References

- https://levelup.gitconnected.com/random-forest-regression-209c0f354c84
- Bio-activity prediction of drug candidate compounds targeting sars-cov-2 using machine learning approaches. PLOS ONE, 18:e0288053, 9 2023.
- Senem Aykul and Erik Martinez-Hackert. Determination of half-maximal inhibitory concentration using biosensor-based protein interaction analysis. Analytical Biochemistry, 508:97–103, 9 2016.
- Brogi. Computational approaches for drug discovery. Molecules, 24:3061, 8 2019.
- A machine learning (ml) driven web-app for bioactivity prediction of sars-cov-2 main protease (mpro) antagonists. PLOS ONE, 18:e0287179, 6 2023.
- Chembl: a large-scale bioactivity database for drug discovery. Nucleic Acids Research, 40:D1100–D1107, 1 2012.
- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

THANK YOU