CRITERIA

Dataset :
This dataset contains SMILES Notation, pIC50 (Including Blinded) along with 2D Descriptors.
Steps Involved :

- Importing necessary libraries like pandas, sklearn, rdkit etc. for this model.
- Read train and test csv files from previous datasets & Join (Including predicted values of binded) them together using concat. function.
- After shuffling them, Save this file as merged_file.csv.
- This file contains 2D molecular descriptors like logP Values, Molecular Weight, No. of Rotateable Bonds, Aromatic Proposition, Topological Polar Surface, Ring Count of molecules, No. of Hydrogen Acceptor, No. of Hydrogen Donor.

Internal Validation:
Internal validation is a process of evaluating the performance of a QSAR model on the same dataset it was trained on. The main goal of internal validation is to estimate the predictive ability of the model and to identify any potential overfitting or underfitting issues.
STEPS:

- Dividing the dataset into a training set and a test set.
- Random Forest Regressor is then trained on the training dataset and the number of trees in the forest is set to 100.
- Trained model is then used to make predictions on test datasets.
- Performance of the model is then evaluated using different evaluation metrics such as mean absolute error, mean squared error, R-squared score, root mean squared error and mean absolute percentage error. These metrics are used to check the accuracy of the model and how well it can predict the values of the dependent variable.
- Comparing the performance of the model on the test set to the performance on the training set.
- Repeat the above steps multiple times, with different partitions of the dataset to minimize the possibility of overfitting.
- Once the model performs well on the whole dataset, it is considered to be internally validated.
- Finally, the model can be externally validated by applying it to new, independent data sets to estimate its predictive ability.

```
Mean Absolute Error            :  0.35748352040555603
Mean Squared Error             :  0.18490135203406802
R-squared scorer               : 0.8284425191637621
Root mean squared Error        :  0.4300015721297633
Mean absolute percentage error :  0.3503740967947357
```

To pass Golbraikh and Tropsha criteria :
The Golbraikh and Tropsha acceptable model criteria for QSAR models includes several statistical measures that are used to evaluate the performance of the model.
I have used leave-one-out cross-validation (LOOCV) Q2, coefficient of determination (R2).
For a model to pass the Golbraikh and Tropsha criteria, it should have a LOOCV Q2 value greater than 0.5 and an R2 value greater than 0.7. Additionally, the model should have a low standard deviation of the residuals and a good correlation between observed and predicted values.

Results:
R2 value is : 0.83 and have a very low mae, mse score.
Hence it passes the required criteria.

Google Drive Link for Code :
https://drive.google.com/drive/folders/1tcxypmZY_uRuwljDkDfkHyT1zXYGVWbe?usp=sharing