Building QSAR Model

QSAR (Quantitative Structure-Activity Relationship) modeling is a computational technique used to predict the biological activity of chemical compounds based on their structural features. The main goal of QSAR modeling is to establish a mathematical relationship between the structural features of a chemical compound and its biological activity. This relationship is established by comparing the structural features of a set of known active compounds with a set of known inactive compounds, and using this information to predict the activity of new compounds. QSAR models can be used in a variety of applications, such as drug discovery, toxicity prediction, and environmental risk assessment. There are several different QSAR modeling techniques, including multiple linear regression, artificial neural networks, and support vector machines. Each technique has its own strengths and weaknesses, and the choice of technique will depend on the specific application and the available data.

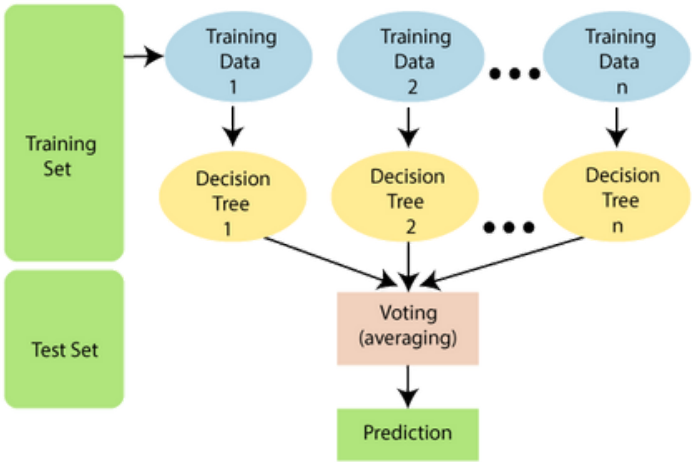
Choosen Model : Random Forest Regression Model

I have been working on a machine learning project in which my aim to predict the activity of a protein called 3CLpro using various chemical compounds. One of the methods that I have implemented is Random Forest Regression, which is a type of ensemble learning method that utilizes multiple decision trees to make predictions. The process begins by generating multiple decision trees using a random subset of the data and features. These decision trees are then combined to make a final prediction by averaging or majority voting.

Random Forest is a popular machine learning algorithm used for both classification and regression problems. It is an ensemble method, meaning that it combines multiple decision trees to make predictions. The basic idea behind this algorithm is to create a large number of decision trees and then combine their predictions to make a more accurate and stable prediction.

The decision trees used in Random Forest are created using a technique called bootstrap aggregating, or bagging, which involves randomly selecting a subset of the data with replacement to train each tree. This helps to reduce the variance in the predictions made by each tree and thus improves the overall performance of the model.

Another key aspect of Random Forest is that it also uses a technique called feature randomness, which randomly selects a subset of the features to consider when making splits in each tree. This helps to decorrelate the trees and further reduce the variance in the predictions.



In our case, I used RF algorithm to predict the 3clpro values of a set of compounds. We trained the model on a dataset that contains information about these compounds such as their SMILES notation and their corresponding 3clpro values. I used the RF algorithm to learn the relationship between these compounds and their 3clpro values, and then used this trained model to make predictions on a set of unseen compounds.

In summary, I have used Random Forest algorithm in this project to predict the 3clpro values of compounds. I have trained the model on a dataset and made predictions on a set of unseen compounds. This algorithm is powerful and widely used in various fields, and it works well with large datasets and non-linear relationships. I believe that this algorithm will be effective in predicting 3clpro values of compounds and will help in drug discovery.

HOW THIS MODEL WORKS

In Random Forest Algorithm, a large number of decision trees are created by randomly selecting a subset of the data and a subset of the features at each node. These decision trees are then combined to make a final prediction. The final prediction is made by taking the average or the mode of the predictions of all the decision trees. This combination of decision trees helps to reduce the overfitting problem that can occur in a single decision tree.

Dataset :

This dataset contains Compound No., SMILES Notation, pIC50 (Including Blinded).

Steps Involved :

- Importing necessary libraries like pandas, sklearn, rdkit etc. for thsi model.
- Used rdkit library to read the smile notation of the compounds from the csv file.
- Calculate various molecular descriptors from the smile notation of the compounds using rdkit library. These molecular descriptors are then used as features for the Random Forest model.
- Molecular Descriptors : 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.
- Saved these descriptors in another csv file.

Regression Model:

- The original dataset is splitted into 3 datasets & read from three csv files: "Training_Dataset_csv.csv", "Validation_Dataset_csv.csv", "Test_Dataset_csv.csv".
- Three datasets are then separated into independent variables (X_train, X_valid, X_test) and dependent variable (y_train, y_valid, y_test) respectively.
- Independent variables are the 8 features: "logP", "MolecularWeight", "RotatableBonds", "AromaticProportion", "TPSA", "HBA", "HBD", "RingCount"
- Dependent variable is "pIC50 (IC50 in microM)".
- Random Forest Regressor is then trained on the training dataset and the number of trees in the forest is set to 90.
- Trained model is then used to make predictions on validation and 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.
- Finally Prediction is made on the Test Dataset and Evaluated the values for Blinded Compounds.

CHOOSING THE MOST PROMISING COMPOUND FOR DRUG DESIGN.

Steps Involved :

- **Preparing the protein and ligand structures:**
  Main protease of SARS-CoV-2 (the virus that causes COVID-19), is 6LU7 which we can download this from the PDB (Protein Data Bank). Once having the PDB file, This file can be used for molecular docking with ligand compounds to predict their binding affinity to the 3CLpro enzyme.
  To convert the ligand structure from SMILES notation to PDB format, We can use a tool OpenBabel.
- **Grid generation & Energy minimization:**
  A grid is generated around the protein active site, where the ligand is expected to bind. The grid represents the space where the ligand can explore and dock.
  The ligand and protein structures are energy minimized to remove any high-energy conformations and to ensure stability.
- **Docking:**
  Molecular Docking can be done with Autodocks software. The ligand is docked into the active site of the protein using a molecular docking
  algorithms. The algorithm will explore different binding poses of the ligand and evaluate the binding energy of each pose.
- **Scoring:**
  The docked poses are scored based on the binding energy and other factors such as ligand protein interactions and ligand flexibility.
- **Post-processing:**
  The top-scoring poses are analyzed and visualized to identify the most promising binding pose.

Binding affinity refers to the strength of the interaction between a drug compound (ligand) and its target protein (receptor).
For choosing best drugs we predicted the binding affinity and selectivity of the compound. Out of the 10 compounds, one with the higher binding affinity, the stronger the interaction and the more likely the drug will be effective.

Golden Rule:

Based on 3CLpro50 value...

Choosing the best drug compound based on 3CLpro50 values can be done by selecting the compound with the lowest 3CLpro50 value. 3CLpro50 is an IC50 measurement, which is a measure of the concentration of a drug required to inhibit an enzyme by 50%. Therefore, a lower IC50 value indicates a more potent drug, and thus the compound with the lowest 3CLpro50 value would be the best choice.
Best Drug design compound would be :

CC(SC1=NC(C2=CC=CC=C2)=C(C#N)C(=O)N1)C(=O)NC1=CC=C(Cl)C=C1