Dhynasah Cakir

Drug Discovery project

1/25/2022

Bioactivity of Estrogen Receptor Alpha Compounds Report

Objective:

The purpose of this project is to construct a machine learning model in the process of studying quantitative structure relationships (QSR models). QSR models are important for drug discovery research. They help us understand the biological activity of molecules and the interpretation of the model can help researchers understand how to design a better drug. This project specifically investigates the activity of Molecules associated with the Estrogen receptor alpha.

Introduction:

Estrogen receptor alpha (Erα) is a nuclear transcription factor that plays a critical role in regulating many complex physiological processes in the human body. Modulation of this receptor by different pharmaceutical drugs or therapeutic agents is being studied for conditions such as cancer, metabolic disease, inflammation, osteoporosis, and neurodegeneration (Paterni et. al). This project looks at the activity of many of the compounds that are reported as modulators of Erα and creating a machine learning model to predict the activity of these compounds. Python is used to construct machine learning models in the process of studying quantitative structure relationships (QSR models). QSR are important for drug discovery research. They help us understand the biological activity of the molecules and the interpretation of the model can help researchers understand how to design a better drug.

The data for this is from the ChEMBL database. The ChEMBL Database is a database that contains curated bioactivity data of more than 2 million compounds. It is compiled from more than 76,000 documents, 1.2 million assays and the data spans 13,000 targets and 1,800 cells and 33,000 indications.

This document gives instructions for parts 1-4 of the code and explain their purpose and results, as well as the overall conclusion of this project.

Part 1:

In part 1 data was downloaded from the ChEMBL database and processing it to so we can do exploratory analysis in the next part. To retrieve the data, we want we need to use target search. Targets refers to the target protein or target organism that the drug will act on. Biologically, these compounds will encounter the protein and induce a modulatory effect. either activate it, or it can inhibit it. The data is cleaned, and a new column is added called bioactivity classification based on the activity of the compound. This will be used later in the machine learning model. Lastly the dataset is saved to a csv file. Either download and save to your computer or save to google drive. The file is already saved in the github folder if you want to add it to your google drive.

Part 2:

This part is exploratory analysis. Upload csv from part 1. Or use path from saved file to load data into data frame.

First the Lipinski descriptors are calculated.

Christopher Lipinski, a scientist at Pfizer, came up with a set of rule-of-thumb for evaluating the drug likeness of compounds. Such drug likeness is based on the Absorption, Distribution, Metabolism and Excretion (ADME) that is also known as the pharmacokinetic profile.

Second: IC50 is converted to pIC50. To allow IC50 data to be more uniformly distributed, IC50 is converted to the negative logarithmic scale which is essentially -log10(IC50). IC50 is the half maximal inhibitory concentration. It is a measure of potency of a substance in inhibiting a specific biological or biochemical function.

Exploratory analysis is performed. See notebook for interpretation of results.

Part 3:

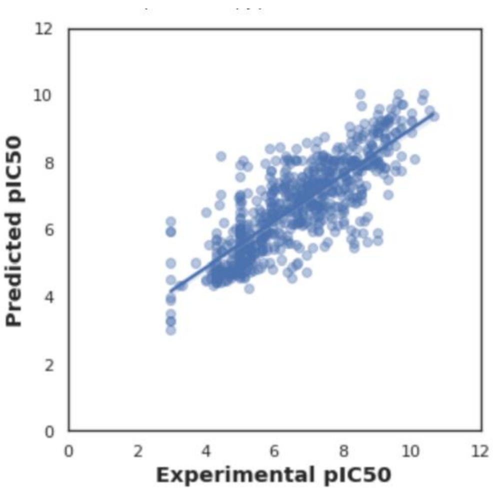
Part 3 focuses on preparing the dataset for model building. Here we will be calculating molecular descriptors and fingerprints. The padel software is installed and downloaded. The descriptors and fingerprints. Some of the descriptors include atom type electrotopological state descriptors, Crippen's logP and MR, extended topochemical atom (ETA) descriptors, molecular linear free energy relation descriptors, and ring counts. Although this is not an exhaustive list.

For more information on descriptors:

http://www.yapcwsoft.com/dd/padeldescriptor/

Download and name file. It will be used in the next section.

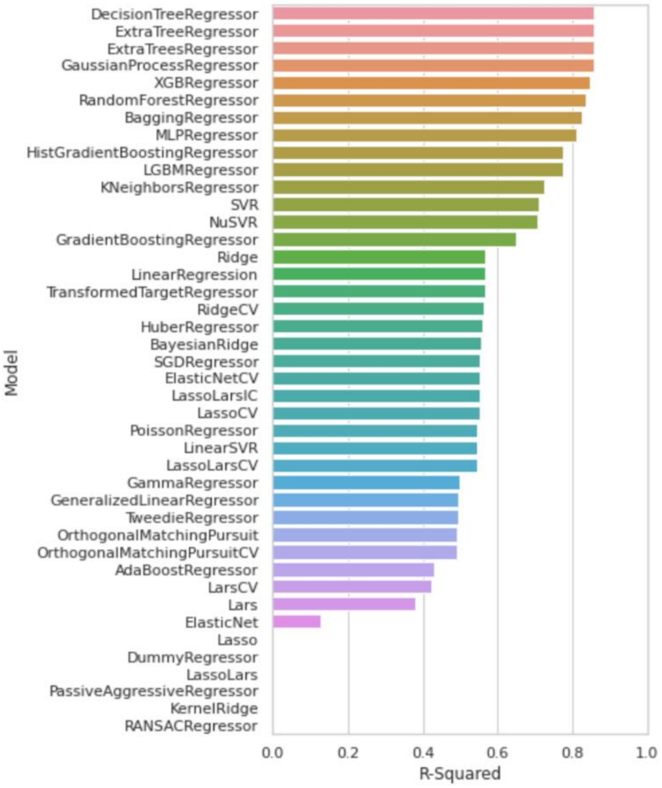
Part 4:



Load dataset from part 3. A random forest regression model was created.

From the scatter plot of predicted vs experimental pIC50, the values are highly correlated. This means the model was very accurate.

A set of models were created to compare the accuracy between them. This was done using a lazy classifier. Plotting the R values of all the models, its apparent that decision tree classifiers and, random forest, and other Gaussian classifiers with an R-squared value of .8 are more accurate than any other type of classification model.


DecisionTreeRegressor
ExtraTreeRegressor
ExtraTreesRegressor
GaussianProcessRegressor
XGBRegressor
RandomForestRegressor
BaggingRegressor
MLPRegressor
HistGradientBoostingRegressor
LGBMRegressor
KNeighborsRegressor
SVR
NuSVR
GradientBoostingRegressor
Ridge
LinearRegression
TransformedTargetRegressor
RidgeCV
HuberRegressor
BayesianRidge
SGDRegressor
ElasticNetCV
LassoLarsIC
LassoCV
PoissonRegressor
LinearSVR
LassoLarsCV
GammaRegressor
GeneralizedLinearRegressor
TweedieRegressor
OrthogonalMatchingPursuit
OrthogonalMatchingPursuitCV
AdaBoostRegressor
LarsCV
Lars
ElasticNet
Lasso
DummyRegressor
LassoLars
PassiveAggressiveRegressor
KernelRidge
RANSACRegressor
Model
R-Squared