PERMEABILITY PREDICTRION OF ORGANICS USING MACHIENE LEARNING

SUBMITTED BY:

HRITIKA SHAH (MT22101)

HELI SHAH (MT22100)

DEVANG PRAJAPATI (MT22050)

GNANESH PATEL (MT22030)

PROJECT GUIDE:

DR. N ARUL MURGAN

INTRODUCTION:

Project is about developing various machine learning model for permeability prediction of organics, where we had a data set that contains smiles of organic compound. We are expected to train machine learning model for permeability of organics using various machine learning techniques.

We have been assigned research paper “QSPR Model for Caco-2 Cell Permeability Prediction using a Combination of HQPSO and Dual-RBF Neural Network” for the reference, all the dataset is provided from this research paper.

TECHNOLOGY AND SOFTWARE USED:

Programming Language: Python

Libraries: Sklearn (For building ML models),

Pandas (For Pre-processing of Dataset),

Numpy (For data analysis),

Lazy predict (To apply 27 machine Learning models),

Rdkit (For conversion from smiles to mol2),

Mordred (For descriptors).

DATASET:

Dataset is given in research paper [1].

Dataset Link: https://drive.google.com/drive/u/2/folders/1LlyJZUSII3BszkRxJT\_-51XSlaCHaKYT

We have been assigned 3 dataset files:

Initial\_modeling\_dataset

QPSR\_modeling\_dataset

Training\_Testing\_dataset

Training\_Testing\_dataset is dataset used in research paper for outcome. We have used QPSR dataset smiles and generate descriptors for those smiles.

PROJECT OUTLINE:

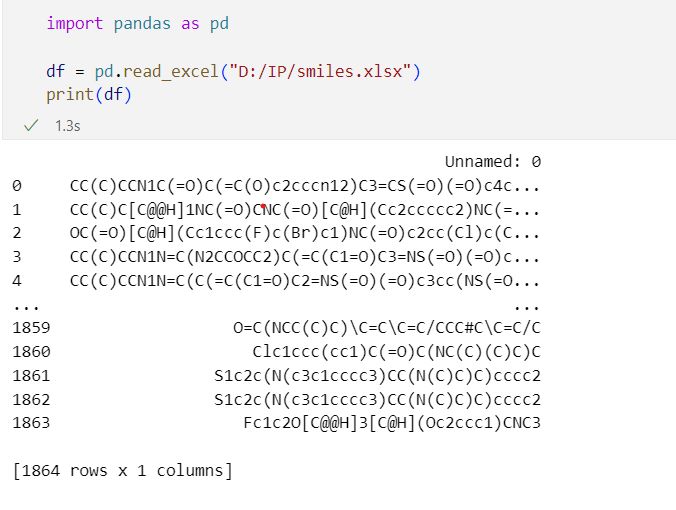
* Smiles extraction from dataset
* Conversion to canonical smile
* Canonical to mol and mol2 conversion
* Descriptor generation using Rdkit
* Descriptor generation using Mordred
* Removed non-numeric descriptors from dataset
* Scale down using standard scalar
* PCA for dimension reduction and feature selection
* Developed Machine Learning models on obtained descriptors.
* Lazy Predict results

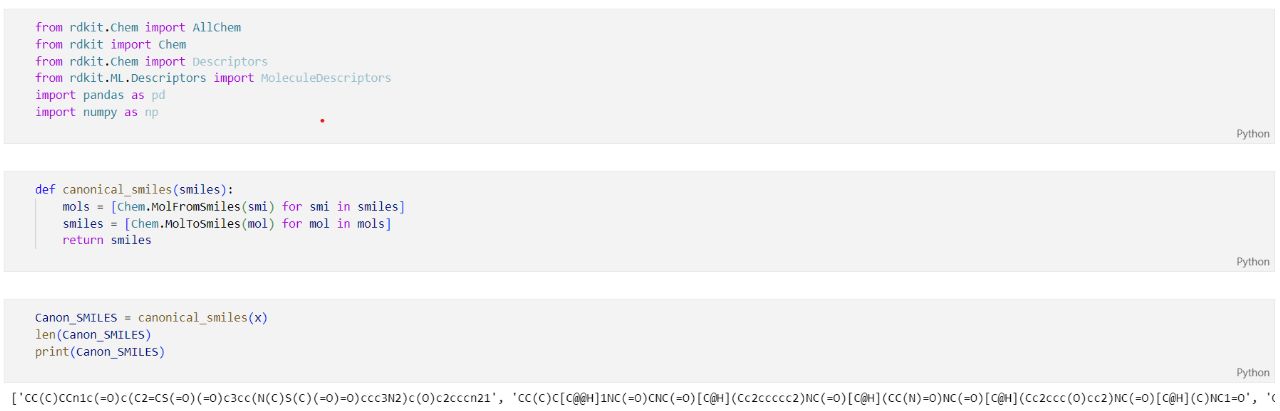
1. Smile Extraction from dataset





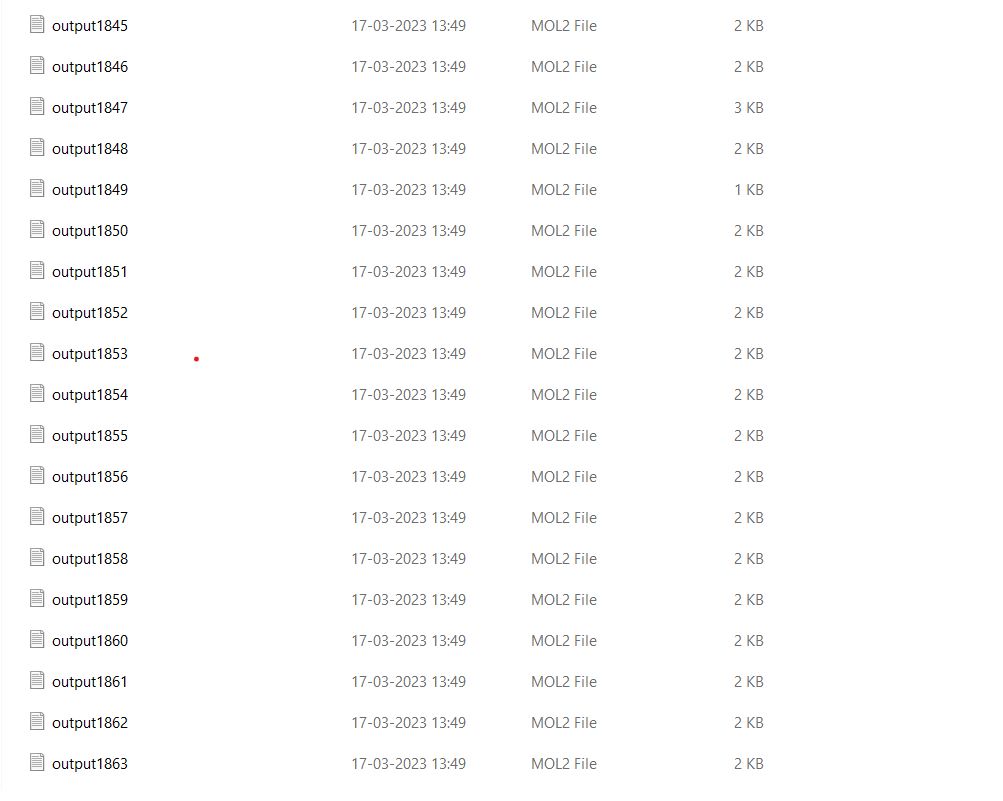
1. Conversion to canonical smile

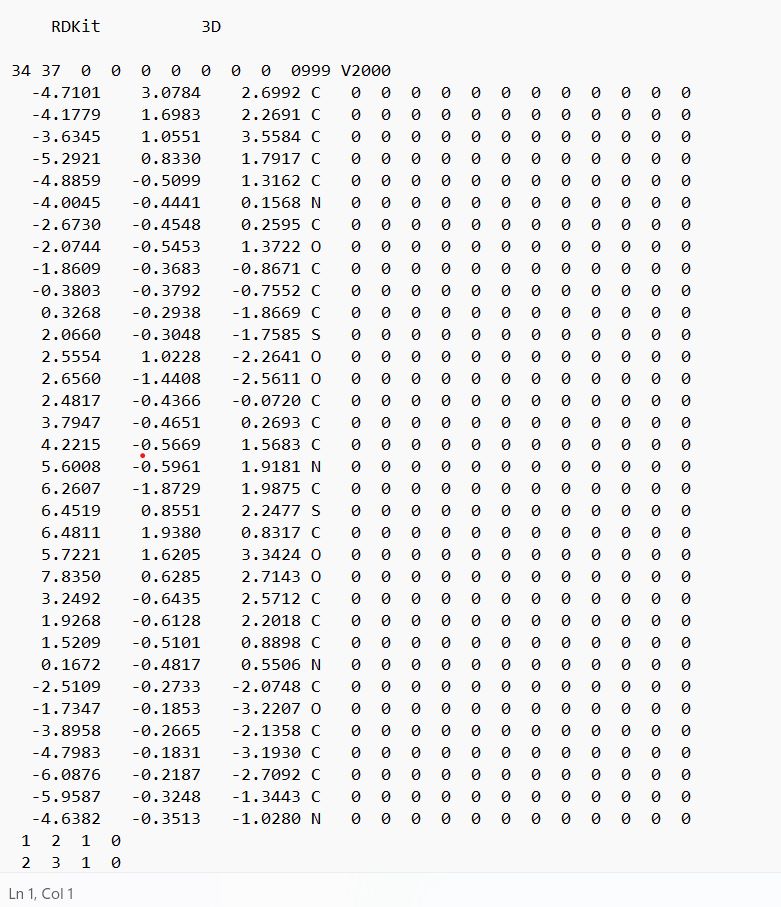




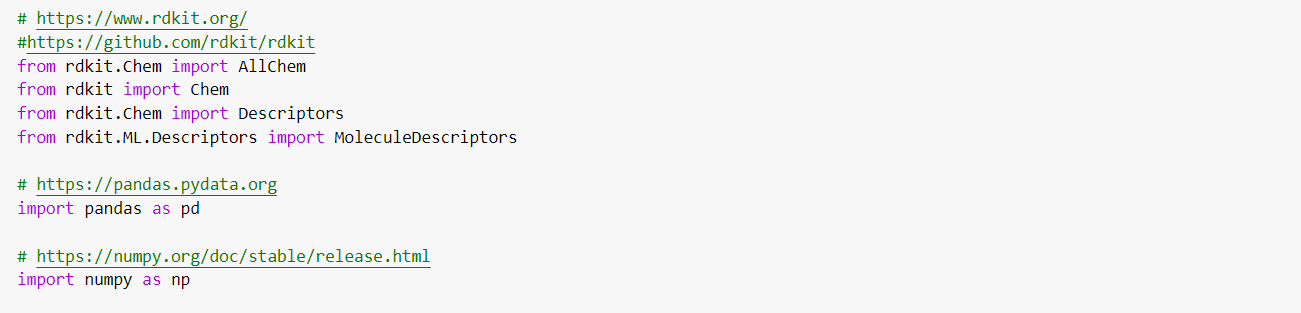
1. Canonical to mol and mol2 conversion





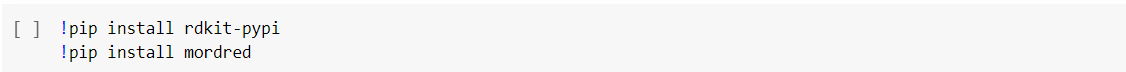


1. Descriptor generation using Rdkit

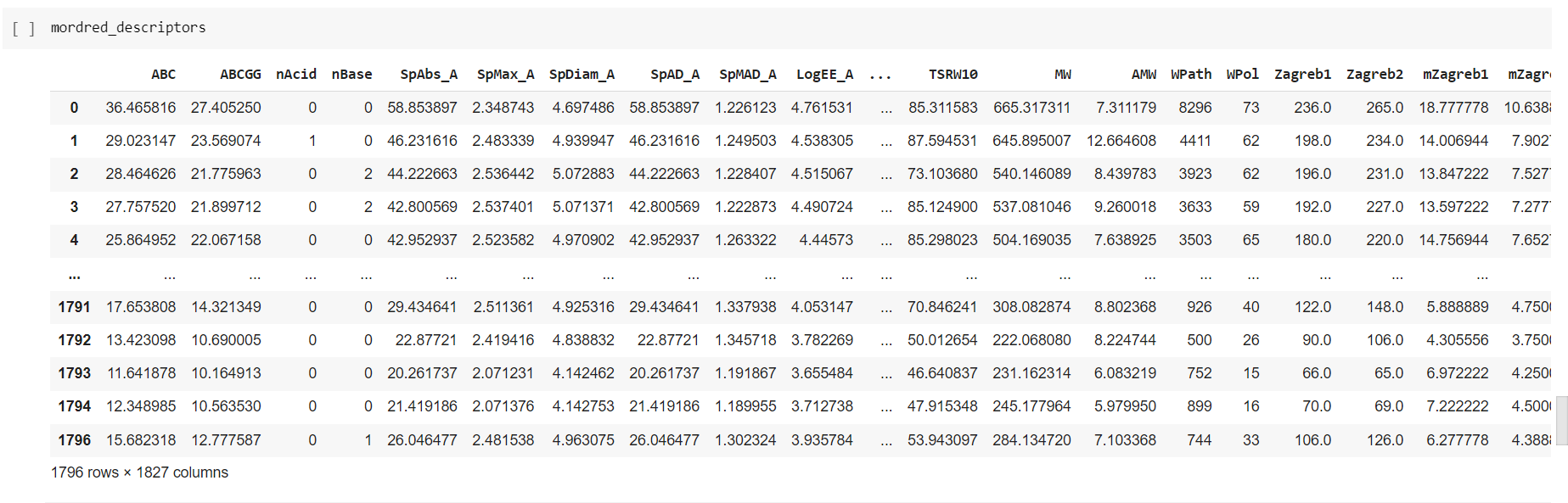




1. Descriptor generation using Mordred

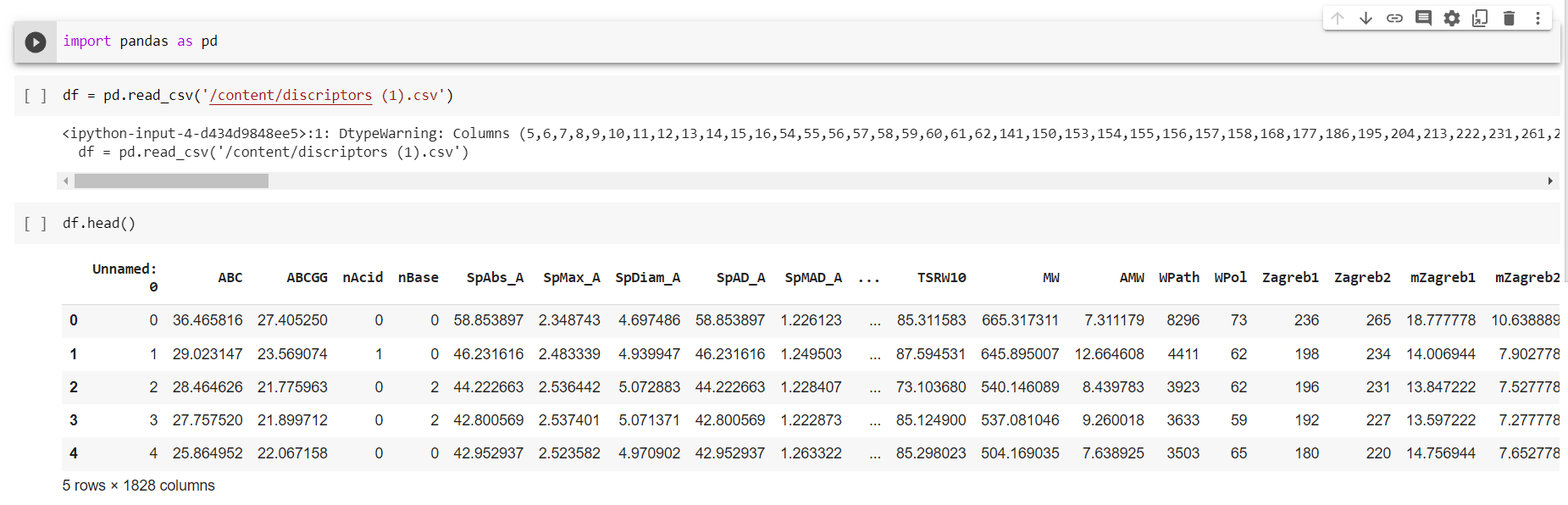


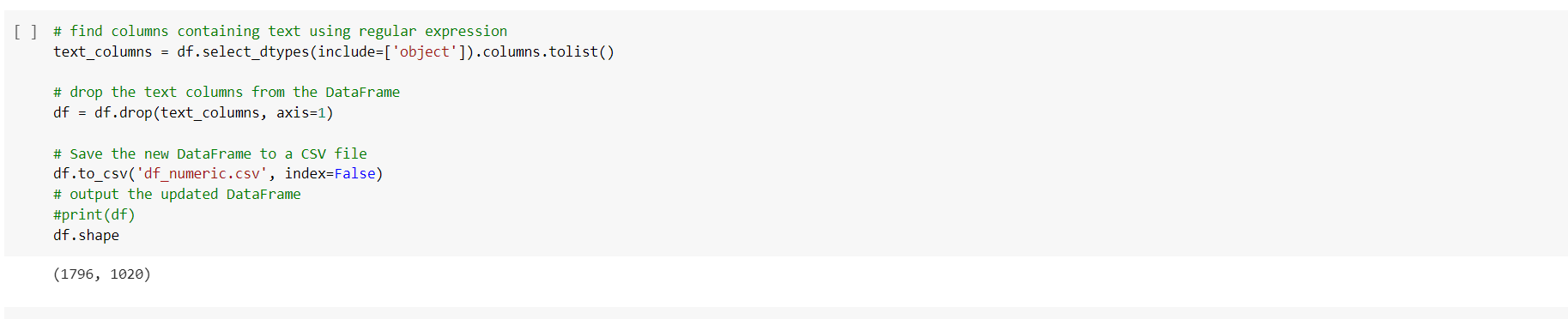




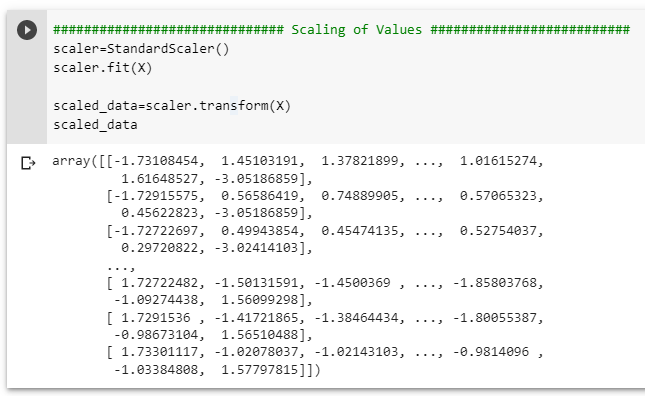


1. Removed non-numeric descriptors from dataset

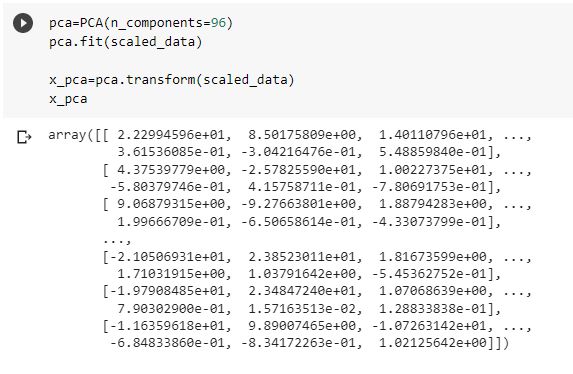




1. Scale down using standard scalar

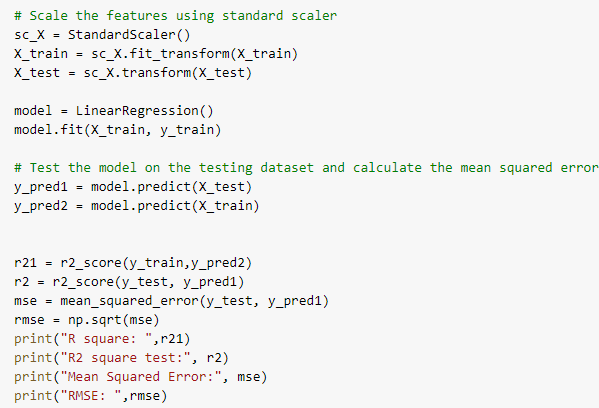


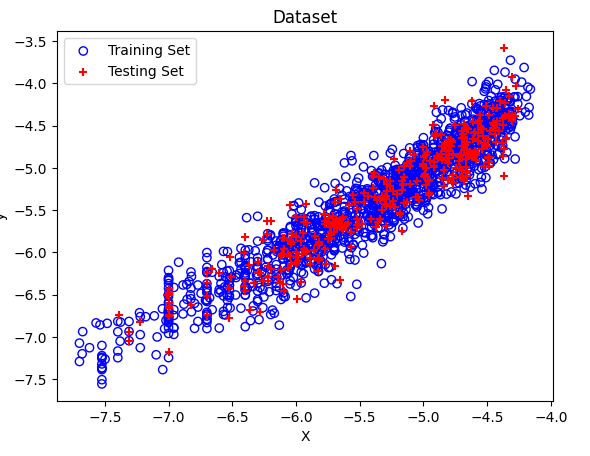
1. PCA for dimension reduction

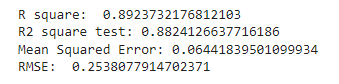


1. Developed Machine Learning models on obtained descriptors

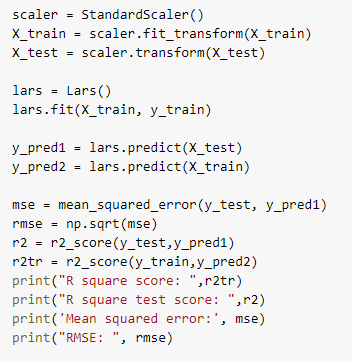
-Linear Regression:

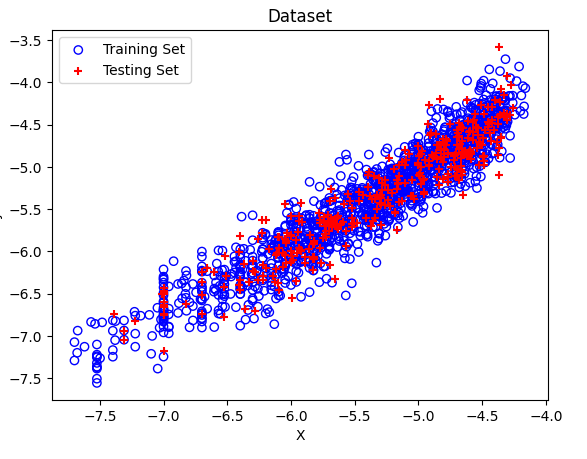


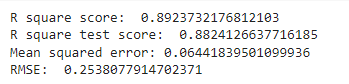




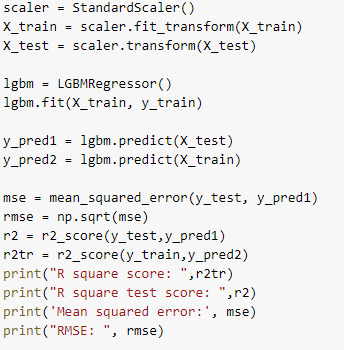
* Lars:

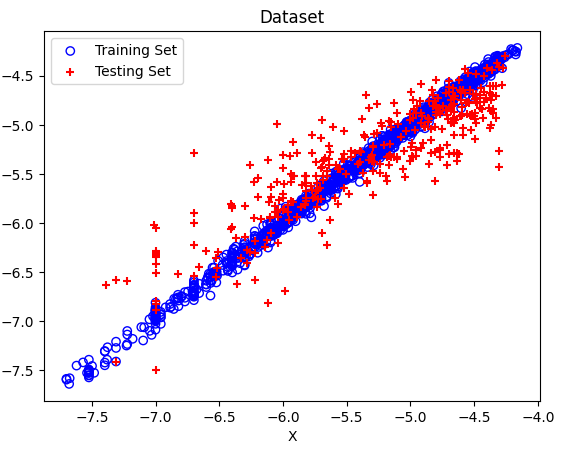


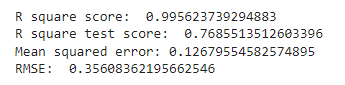




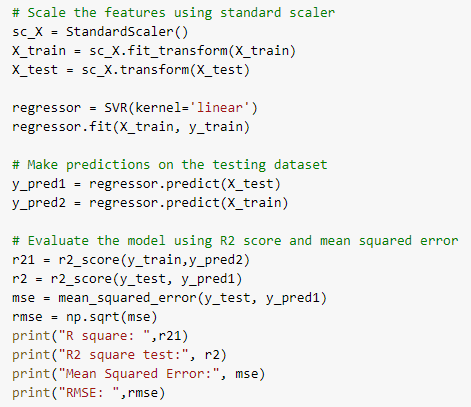
* LGBMRegressor:

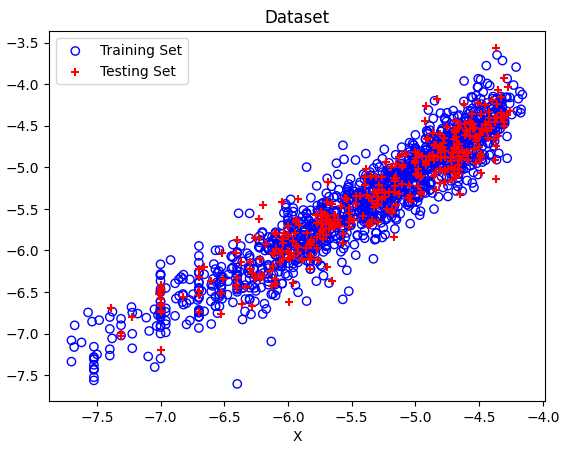


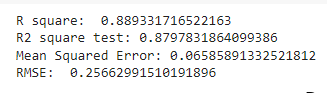




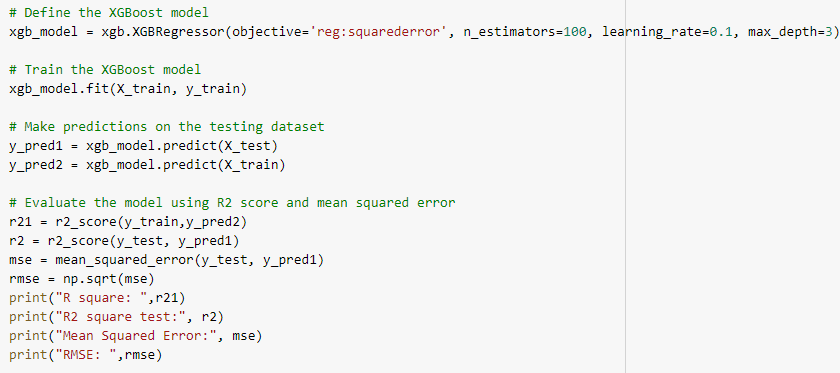
* SVR:

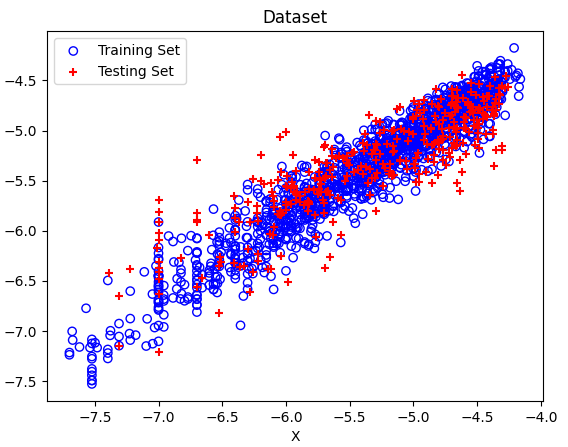


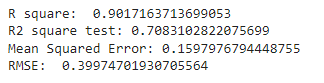




-xgBoost:

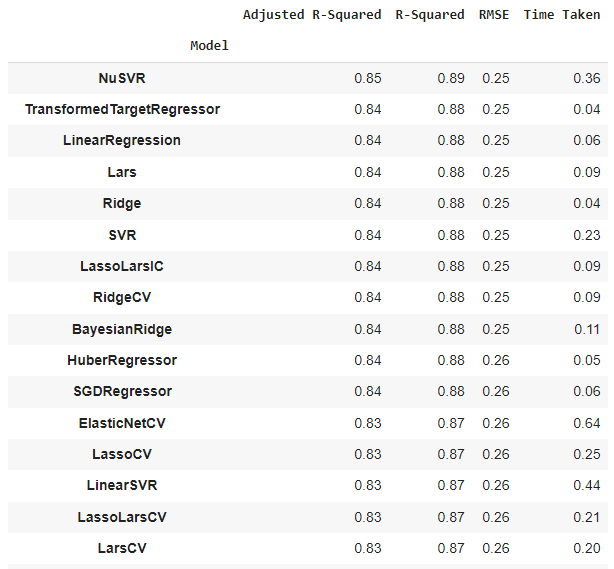


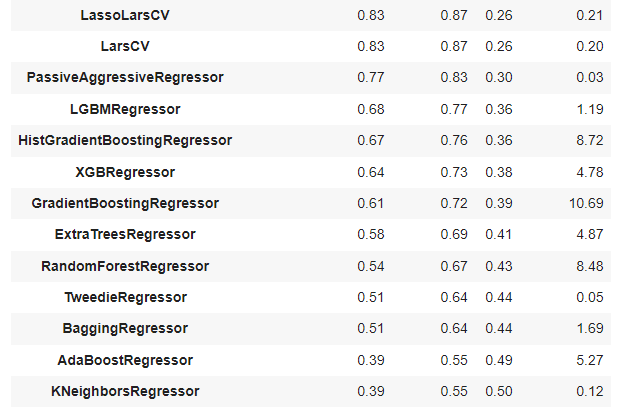




* Lazy Predict results:







Conclusion:

From the results we obtained, we conclude that some of our models are getting better results than results obtained in current research paper on testing dataset.

We have also used models such as LARS and LGBMregressor which are not shown in the paper and they are giving equivalent results. Other models which we have used are SVM, Linear Regressor and XgBoost. We have also used LazyPredict which gives us results from 25 models approximately.

Reference:

[1] QSPR Model for Caco-2 Cell Permeability Prediction using a Combination of HQPSO and Dual-RBF Neural Network