(Protocosas) D:\scripts\generate\_models\ML\_pipeline>python ML\_pipeline.py

#########################################################################

##################### WELCOME TO ML\_pipeline script #####################

#########################################################################

This script will allow you to:

- generate ML models for regression or classification

Please input your PATH (enter to: "../data/Af\_MIC80\_definitva/no3D/OWNdesc/"): C:\Users\Vicente\Desktop\model herg\

Please input your MODEL NAME (enter to: Af\_MIC80): TOX\_hERGinh

######################### MAIN MENU #########################

Please select what do you want to do:

[1] Perform a MODEL for Regression

[2] Perform a MODEL for Classification

[3] Exit

Your choice: 2

Please select the method used for DESCRIPTOR SELECTION:

[1] RFE

[2] GBM

[3] PI

[4] OWN

Your choice: 2

You want to perform a Classification model, by using descriptors selected by GBM.

Is that correct?(Y/n):

The following files located in "C:\Users\Vicente\Desktop\model herg\" folder are needed:

TOX\_hERGinh-train\_reduction\_GBM.csv

TOX\_hERGinh-test\_reduction\_GBM.csv

Continue (Y/n)?

PARAMETERS:

train molecules: 15830

test molecules: 6785

total molecules: 22615

DESCRIPTORS: ( 20 )

['SLogP', 'PEOE\_VSA7', 'JhetZ', 'SpMAD\_Z', 'SssCH2', 'JGI3', 'SdssC', 'SMR\_VSA3', 'IC3', 'GATS5Z', 'ATSC1i', 'EState\_VSA9', 'VSA\_EState4', 'PEOE\_VSA8', 'S-108', 'GATS4s', 'D/Dr6', 'H-047', 'SMR\_VSA7', 'VSA\_EState2']

mols/descriptor ratio: 791.5

PLEASE, ENSURE SEED HYPEPARAM IF USING XGB MODELS AND ENSURE REPRODUCIBILITY!!

PLEASE, ENSURE SEED HYPEPARAM IF USING XGB MODELS AND ENSURE REPRODUCIBILITY!!

PLEASE, ENSURE SEED HYPEPARAM IF USING XGB MODELS AND ENSURE REPRODUCIBILITY!!

[+] Training the model...

Best parameters for LGBMClassifier:

{'class\_weight': {0: 1, 1: 0.63}, 'random\_state': 9}

Grid test score: 0.8134554643082754

------------------------------

Best model and parameters:

LGBMClassifier(class\_weight={0: 1, 1: 0.63}, random\_state=9)

Best model and parameters:

{'class\_weight': {0: 1, 1: 0.63}, 'random\_state': 9}

GridSearch parameters:

GridSearchCV(cv=5, estimator=LGBMClassifier(), n\_jobs=-1,

param\_grid={'class\_weight': [{0: 1, 1: 2}, {0: 1, 1: 0.63}],

'random\_state': [9]},

return\_train\_score=True, scoring='f1\_micro')

-------------------------------

A JSON file with the descriptors has been saved as C:\Users\Vicente\Desktop\model herg\TOX\_hERGinh.json

######################################################

################### Model results ####################

######################################################

############### Training set results ###############

precision recall f1-score support

0.0 0.85 0.87 0.86 6441

1.0 0.91 0.89 0.90 9389

accuracy 0.88 15830

macro avg 0.88 0.88 0.88 15830

weighted avg 0.89 0.88 0.89 15830

Confussion matrix:

OBS\PRED 0 1

0 5609 832

1 991 8398

############### Test set results ###############

precision recall f1-score support

0.0 0.77 0.78 0.77 2761

1.0 0.84 0.84 0.84 4024

accuracy 0.81 6785

macro avg 0.81 0.81 0.81 6785

weighted avg 0.81 0.81 0.81 6785

Confussion matrix:

OBS\PRED 0 1

0 2140 621

1 653 3371

################ Global results ###############

|Train| Test

acc |0.88 |0.81

prec\_PPV |0.91 |0.84

recall |0.89 |0.84

f1 |0.90 |0.84

auc |0.88 |0.81

sensit\_TPR |0.89 |0.84

spec\_TNR |0.87 |0.78

NPV |0.85 |0.77

FNR |0.11 |0.16

FPR |0.13 |0.22

FDR |0.09 |0.16

FOR |0.15 |0.23

F\_score |0.90 |0.84

MCC |0.76 |0.61

CSI |0.82 |0.73

######################################################

############## Cross-validation results ##############

######################################################

|Train | Test

acc |0.88 +/- 0.002 |0.82 +/- 0.006

prec |0.90 +/- 0.001 |0.85 +/- 0.004

recall |0.89 +/- 0.002 |0.84 +/- 0.010

f1 |0.89 +/- 0.001 |0.85 +/- 0.006

auc |0.95 +/- 0.000 |0.90 +/- 0.002

sensit |0.89 +/- 0.002 |0.84 +/- 0.010

spec |0.86 +/- 0.002 |0.78 +/- 0.006

NPV |0.84 +/- 0.003 |0.77 +/- 0.011

FNR |0.11 +/- 0.002 |0.16 +/- 0.010

FPR |0.14 +/- 0.002 |0.22 +/- 0.006

FDR |0.10 +/- 0.001 |0.15 +/- 0.004

FOR |0.16 +/- 0.003 |0.23 +/- 0.011

F\_score |0.89 +/- 0.001 |0.85 +/- 0.006

MCC |0.74 +/- 0.003 |0.62 +/- 0.011

CSI |0.81 +/- 0.002 |0.73 +/- 0.008

Please, close the plots to continue...

TRAIN experimental Positive: 9389

TRAIN experimental Negative: 6441

TEST experimental Positive: 4024

TEST experimental Negative: 2761

Do you want to perform any other step?(y/n): n

Thanks for using ML\_pipeline!

(Protocosas) D:\scripts\generate\_models\ML\_pipeline>