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

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

######################### WELCOME TO NEO script #########################

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

This script will allow you to:

- eliminate 3D descriptors

- "y" transformation

- perform the initial unsupervised feature reduction

- perform the train/test split based on kmeans

- descriptor standarization

- select the relevant features based on:

· Recursive feature elimination (RFE)

· Feature importance (FI) based on Ligth gradient boosting machine (LGBM)

· Permutation importance (PI)

- select your own features features

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\_no3D): TOX\_hERGinh

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

Please select what do you want to do:

[01] Elimination of 3D descriptors [your dataset will be saved as [Name]\_no3D]

[1] "y" transformation + dataset random order + Knn imputation

[2] Initial feature reduction: infinite, correlated, constant and empty values

[3] Generation of train and test sets based in kmeans

[4] Descriptor standarization

[5] Feature selection by RFE

[6] Feature selection by FI based on LGBM

[7] Feature selection by Permutation importance

[8] Select own features (inside the script)

[0] Exit NEO

Your choice: 1

This part of the code will do the y" transformation, randomization of the dataset order and Knn imputation.

From this version of NEO is its compulsory to perform the inputation here, as this will create the un-imputed file needed for reimputation.

[+] "y" transformation

A file located in "C:\Users\Vicente\Desktop\model herg\" folder is needed

This file must be called: "TOX\_hERGinh-paralel\_calculated\_with\_y.csv"

Continue (Y/n)?y

Please select your type of model:

[1] Regression

[2] Classification

Your choice: 2

I am so sorry, there is nothing yet for your request. Please try tomorrow with more coffee and cookies.

[+] dataset random sort

The following file has been created (save it as you will need it for feature reduction):

C:\Users\Vicente\Desktop\model herg\TOX\_hERGinh-calculated\_preimputation.csv

[+] dataset imputation

Size of the database, preimputation: (22615, 3722)

Warning! Some of descriptors have too many NaN values and have been removed: ['LPRS', 'VDA', 'MDDD', 'MAXsLi', 'MAXssBe', 'MAXssssBe', 'MAXssBH', 'MAXsssB', 'MAXssssB', 'MAXsCH3', 'MAXdCH2', 'MAXtCH', 'MAXdsCH', 'MAXsssCH', 'MAXddC', 'MAXtsC', 'MAXdssC', 'MAXaaaC', 'MAXssssC', 'MAXsNH3', 'MAXsNH2', 'MAXssNH2', 'MAXdNH', 'MAXssNH', 'MAXaaNH', 'MAXtN', 'MAXsssNH', 'MAXdsN', 'MAXaaN', 'MAXsssN', 'MAXddsN', 'MAXaasN', 'MAXssssN', 'MAXsOH', 'MAXdO', 'MAXssO', 'MAXaaO', 'MAXsF', 'MAXsSiH3', 'MAXssSiH2', 'MAXsssSiH', 'MAXssssSi', 'MAXsPH2', 'MAXssPH', 'MAXsssP', 'MAXdsssP', 'MAXsssssP', 'MAXsSH', 'MAXdS', 'MAXssS', 'MAXaaS', 'MAXdssS', 'MAXddssS', 'MAXsCl', 'MAXsGeH3', 'MAXssGeH2', 'MAXsssGeH', 'MAXssssGe', 'MAXsAsH2', 'MAXssAsH', 'MAXsssAs', 'MAXsssdAs', 'MAXsssssAs', 'MAXsSeH', 'MAXdSe', 'MAXssSe', 'MAXaaSe', 'MAXdssSe', 'MAXddssSe', 'MAXsBr', 'MAXsSnH3', 'MAXssSnH2', 'MAXsssSnH', 'MAXssssSn', 'MAXsI', 'MAXsPbH3', 'MAXssPbH2', 'MAXsssPbH', 'MAXssssPb', 'MINsLi', 'MINssBe', 'MINssssBe', 'MINssBH', 'MINsssB', 'MINssssB', 'MINsCH3', 'MINdCH2', 'MINtCH', 'MINdsCH', 'MINsssCH', 'MINddC', 'MINtsC', 'MINdssC', 'MINaaaC', 'MINssssC', 'MINsNH3', 'MINsNH2', 'MINssNH2', 'MINdNH', 'MINssNH', 'MINaaNH', 'MINtN', 'MINsssNH', 'MINdsN', 'MINaaN', 'MINsssN', 'MINddsN', 'MINaasN', 'MINssssN', 'MINsOH', 'MINdO', 'MINssO', 'MINaaO', 'MINsF', 'MINsSiH3', 'MINssSiH2', 'MINsssSiH', 'MINssssSi', 'MINsPH2', 'MINssPH', 'MINsssP', 'MINdsssP', 'MINsssssP', 'MINsSH', 'MINdS', 'MINssS', 'MINaaS', 'MINdssS', 'MINddssS', 'MINsCl', 'MINsGeH3', 'MINssGeH2', 'MINsssGeH', 'MINssssGe', 'MINsAsH2', 'MINssAsH', 'MINsssAs', 'MINsssdAs', 'MINsssssAs', 'MINsSeH', 'MINdSe', 'MINssSe', 'MINaaSe', 'MINdssSe', 'MINddssSe', 'MINsBr', 'MINsSnH3', 'MINssSnH2', 'MINsssSnH', 'MINssssSn', 'MINsI', 'MINsPbH3', 'MINssPbH2', 'MINsssPbH', 'MINssssPb'] The limit is marked to the 15.0 %

[+] fitting

[+] transforming

Size of the database, postimputation: (22615, 3567)

C-001 C-002 C-003 C-004 C-005 C-006 C-007 ... MAXaaCH MAXaasC MINssCH2 MINaaCH MINaasC SLogP SMR

0 0.0 7.0 2.0 0.0 1.0 3.0 0.0 ... 2.28414 1.45828 0.78125 2.00446 0.96090 4.46290 124.2678

1 1.0 1.0 0.0 0.0 2.0 3.0 0.0 ... 1.88100 0.90848 0.27206 1.84188 0.68199 4.07230 132.6877

2 0.0 2.0 0.0 0.0 0.0 5.0 0.0 ... 1.99602 1.06667 0.43825 1.51726 0.56664 3.89110 113.9820

3 5.0 4.0 1.0 0.0 1.0 2.0 0.0 ... 0.69609 -0.52782 -0.07548 0.30364 -1.29347 4.26600 140.3420

4 0.0 4.0 0.0 0.0 0.0 4.0 0.0 ... 2.02883 1.32681 0.76460 1.90268 1.13241 0.23470 68.2510

... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...

22610 0.0 5.0 1.0 0.0 0.0 0.0 0.0 ... 2.05860 1.20134 0.78946 1.71537 0.53479 3.18550 125.5066

22611 2.0 2.0 0.0 0.0 1.0 2.0 0.0 ... 1.97362 0.96874 0.59793 1.07251 -0.53001 5.90004 146.4605

22612 3.0 2.0 0.0 1.0 2.0 0.0 0.0 ... 2.15430 1.18796 0.75349 2.02918 1.18796 4.65460 107.2750

22613 1.0 0.0 0.0 0.0 2.0 2.0 0.0 ... 2.15542 1.26964 0.72311 2.03991 1.20885 3.66262 83.9700

22614 0.0 0.0 0.0 0.0 2.0 1.0 0.0 ... 2.01633 1.14482 0.67329 1.90957 0.72636 3.77290 94.3407

[22615 rows x 3567 columns]

The following files have been created:

C:\Users\Vicente\Desktop\model herg\TOX\_hERGinh-calculated\_imputed\_ytransformed.csv

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

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

Please select what do you want to do:

[01] Elimination of 3D descriptors [your dataset will be saved as [Name]\_no3D]

[1] "y" transformation + dataset random order + Knn imputation

[2] Initial feature reduction: infinite, correlated, constant and empty values

[3] Generation of train and test sets based in kmeans

[4] Descriptor standarization

[5] Feature selection by RFE

[6] Feature selection by FI based on LGBM

[7] Feature selection by Permutation importance

[8] Select own features (inside the script)

[0] Exit NEO

Your choice: 2

A file located in "C:\Users\Vicente\Desktop\model herg\" folder is needed

This file must be called: "TOX\_hERGinh-calculated\_imputed\_ytransformed.csv"

Continue (Y/n)?y

[1] Initial feature reduction: infinite, correlated, constant and empty values

0 infinite values

0 features with greater than 0.00 missing values.

1559 features with a correlation magnitude greater than 0.90.

819 features with a single unique value.

Data has not been one-hot encoded

Removed 2378 features including one-hot features.

The following files have been created:

C:\Users\Vicente\Desktop\model herg\TOX\_hERGinh-initial\_reduction.csv

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

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

Please select what do you want to do:

[01] Elimination of 3D descriptors [your dataset will be saved as [Name]\_no3D]

[1] "y" transformation + dataset random order + Knn imputation

[2] Initial feature reduction: infinite, correlated, constant and empty values

[3] Generation of train and test sets based in kmeans

[4] Descriptor standarization

[5] Feature selection by RFE

[6] Feature selection by FI based on LGBM

[7] Feature selection by Permutation importance

[8] Select own features (inside the script)

[0] Exit NEO

Your choice: 3

Please select your type of model:

[1] Regression

[2] Classification

Your choice (1/2)?: 2

Please input your desired TEST SIZE (enter to: "0.25"): 0.3

A file located in "C:\Users\Vicente\Desktop\model herg\" folder is needed

This file must be called: "TOX\_hERGinh-initial\_reduction.csv"

Continue (Y/n)?

[+] Generation of train and test sets based in kmeans

OPTIMAL NUMBER OF CLUSTERS: 4

NUMBER OF CLUSTERS: 4

SETS: {2, 12427, 8788, 1398}

ALERTS!!

compound number: 13354

SMILE: CSCC[C@@H]1NC(=O)CNC(=O)[C@H](CCC(=O)O)NC(=O)[C@@H]2CSSC[C@H](NC(=O)[C@@H](N)Cc3ccc(O)cc3)C(=O)N[C@@H](CCC(N)=O)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](Cc3c[nH]c4ccccc34)C(=O)N[C@@H](CCSC)C(=O)N[C@@H](Cc3c[nH]c4ccccc34)C(=O)N[C@@H]([C@@H](C)O)C(=O)N[C@H]3CSSC[C@H](NC(=O)[C@H](C(C)C)NC1=O)C(=O)N[C@@H](CCCNC(=N)N)C(=O)N[C@@H](CC(C)C)C(=O)N[C@@H](Cc1c[nH]c4ccccc14)C(=O)N[C@H](C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](CC(C)C)C(=O)N[C@@H](Cc1c[nH]c4ccccc14)C(N)=O)CSSC[C@H](NC(=O)[C@H](CCCCN)NC(=O)[C@H](CCCNC(=N)N)NC(=O)[C@H](CCC(=O)O)NC(=O)[C@H](CO)NC(=O)[C@H](CC(=O)O)NC3=O)C(=O)N2 1.0

compound number: 20801

SMILE: CC(C)C[C@@H]1NC(=O)CNC(=O)[C@H](CCC(=O)O)NC(=O)[C@@H]2CSSC[C@H](NC(=O)[C@@H](N)Cc3ccc(O)cc3)C(=O)N[C@@H](CCC(N)=O)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](Cc3ccccc3)C(=O)N[C@@H](CC(C)C)C(=O)N[C@@H](C)C(=O)N3CCC[C@H]3C(=O)N[C@H]3CSSC[C@H](NC(=O)[C@H](C(C)C)NC1=O)C(=O)N[C@@H](CCCNC(=N)N)C(=O)N[C@@H](CC(C)C)C(=O)N[C@@H](Cc1c[nH]c4ccccc14)C(=O)N[C@H](C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](CC(C)C)C(=O)N[C@@H](Cc1c[nH]c4ccccc14)C(N)=O)CSSC[C@H](NC(=O)[C@H](CCCCN)NC(=O)[C@H](CCCNC(=N)N)NC(=O)[C@H](CCC(=O)O)NC(=O)[C@H](CO)NC(=O)[C@H](CC(=O)O)NC3=O)C(=O)N2 1.0

[13354, 20801]

you have some molecular alerts. It means that these molecules are quite dissimilar

You can (1) eliminate them or (2) maintain them

What is your choice (1/2)?2

Ok, continue with entire dataframe.

NUMBER OF CLUSTERS: 3

SETS: {2, 6858, 15755}

ALERTS!!

compound number: 13354

SMILE: CSCC[C@@H]1NC(=O)CNC(=O)[C@H](CCC(=O)O)NC(=O)[C@@H]2CSSC[C@H](NC(=O)[C@@H](N)Cc3ccc(O)cc3)C(=O)N[C@@H](CCC(N)=O)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](Cc3c[nH]c4ccccc34)C(=O)N[C@@H](CCSC)C(=O)N[C@@H](Cc3c[nH]c4ccccc34)C(=O)N[C@@H]([C@@H](C)O)C(=O)N[C@H]3CSSC[C@H](NC(=O)[C@H](C(C)C)NC1=O)C(=O)N[C@@H](CCCNC(=N)N)C(=O)N[C@@H](CC(C)C)C(=O)N[C@@H](Cc1c[nH]c4ccccc14)C(=O)N[C@H](C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](CC(C)C)C(=O)N[C@@H](Cc1c[nH]c4ccccc14)C(N)=O)CSSC[C@H](NC(=O)[C@H](CCCCN)NC(=O)[C@H](CCCNC(=N)N)NC(=O)[C@H](CCC(=O)O)NC(=O)[C@H](CO)NC(=O)[C@H](CC(=O)O)NC3=O)C(=O)N2 1.0

compound number: 20801

SMILE: CC(C)C[C@@H]1NC(=O)CNC(=O)[C@H](CCC(=O)O)NC(=O)[C@@H]2CSSC[C@H](NC(=O)[C@@H](N)Cc3ccc(O)cc3)C(=O)N[C@@H](CCC(N)=O)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](Cc3ccccc3)C(=O)N[C@@H](CC(C)C)C(=O)N[C@@H](C)C(=O)N3CCC[C@H]3C(=O)N[C@H]3CSSC[C@H](NC(=O)[C@H](C(C)C)NC1=O)C(=O)N[C@@H](CCCNC(=N)N)C(=O)N[C@@H](CC(C)C)C(=O)N[C@@H](Cc1c[nH]c4ccccc14)C(=O)N[C@H](C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](CC(C)C)C(=O)N[C@@H](Cc1c[nH]c4ccccc14)C(N)=O)CSSC[C@H](NC(=O)[C@H](CCCCN)NC(=O)[C@H](CCCNC(=N)N)NC(=O)[C@H](CCC(=O)O)NC(=O)[C@H](CO)NC(=O)[C@H](CC(=O)O)NC3=O)C(=O)N2 1.0

[13354, 20801]

you have some molecular alerts. It means that these molecules are quite dissimilar

You can (1) eliminate them or (2) maintain them

What is your choice (1/2)?2

Ok, continue with entire dataframe.

NUMBER OF CLUSTERS: 2

SETS: {2, 22613}

ALERTS!!

compound number: 13354

SMILE: CSCC[C@@H]1NC(=O)CNC(=O)[C@H](CCC(=O)O)NC(=O)[C@@H]2CSSC[C@H](NC(=O)[C@@H](N)Cc3ccc(O)cc3)C(=O)N[C@@H](CCC(N)=O)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](Cc3c[nH]c4ccccc34)C(=O)N[C@@H](CCSC)C(=O)N[C@@H](Cc3c[nH]c4ccccc34)C(=O)N[C@@H]([C@@H](C)O)C(=O)N[C@H]3CSSC[C@H](NC(=O)[C@H](C(C)C)NC1=O)C(=O)N[C@@H](CCCNC(=N)N)C(=O)N[C@@H](CC(C)C)C(=O)N[C@@H](Cc1c[nH]c4ccccc14)C(=O)N[C@H](C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](CC(C)C)C(=O)N[C@@H](Cc1c[nH]c4ccccc14)C(N)=O)CSSC[C@H](NC(=O)[C@H](CCCCN)NC(=O)[C@H](CCCNC(=N)N)NC(=O)[C@H](CCC(=O)O)NC(=O)[C@H](CO)NC(=O)[C@H](CC(=O)O)NC3=O)C(=O)N2 1.0

compound number: 20801

SMILE: CC(C)C[C@@H]1NC(=O)CNC(=O)[C@H](CCC(=O)O)NC(=O)[C@@H]2CSSC[C@H](NC(=O)[C@@H](N)Cc3ccc(O)cc3)C(=O)N[C@@H](CCC(N)=O)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](Cc3ccccc3)C(=O)N[C@@H](CC(C)C)C(=O)N[C@@H](C)C(=O)N3CCC[C@H]3C(=O)N[C@H]3CSSC[C@H](NC(=O)[C@H](C(C)C)NC1=O)C(=O)N[C@@H](CCCNC(=N)N)C(=O)N[C@@H](CC(C)C)C(=O)N[C@@H](Cc1c[nH]c4ccccc14)C(=O)N[C@H](C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](CC(C)C)C(=O)N[C@@H](Cc1c[nH]c4ccccc14)C(N)=O)CSSC[C@H](NC(=O)[C@H](CCCCN)NC(=O)[C@H](CCCNC(=N)N)NC(=O)[C@H](CCC(=O)O)NC(=O)[C@H](CO)NC(=O)[C@H](CC(=O)O)NC3=O)C(=O)N2 1.0

[13354, 20801]

you have some molecular alerts. It means that these molecules are quite dissimilar

You can (1) eliminate them or (2) maintain them

What is your choice (1/2)?2

Ok, continue with entire dataframe.

NUMBER OF CLUSTERS: 1

SETS: {22615}

ALERTS!!

[]

index SMILES y C-001 ... MINaaCH MINaasC SLogP cluster

0 0 CN1CCC(c2ccccc2CC2CCN(c3ccc(C4(O)CCC4)cc3)C2=O... 1.0 0.0 ... 2.00446 0.96090 4.46290 0

1 1 C=CCn1c(SCCOc2ccc(OC)cc2)nnc1C(C)NC(=O)Cc1ccc(... 0.0 1.0 ... 1.84188 0.68199 4.07230 0

2 2 O=c1cc(OCc2ccccc2)ccn1-c1ccc(OCCN2CCCC2)cc1 1.0 0.0 ... 1.51726 0.56664 3.89110 0

3 3 CC(C)N(C)[C@@H]1CC[C@H](N2CCC(NC(=O)c3cc(F)cc(... 0.0 5.0 ... 0.30364 -1.29347 4.26600 0

4 4 C1CCc2c(nc3nnnn3c2N2CCOCC2)C1 0.0 0.0 ... 1.90268 1.13241 0.23470 0

... ... ... ... ... ... ... ... ... ...

22610 22610 Nc1nc2c(s1)CCC2C(=O)Nc1ccc(CC2CCC(C(O)c3cccnc3... 0.0 0.0 ... 1.71537 0.53479 3.18550 0

22611 22611 Cc1cc(Cl)cc(N2C=Nc3cc(C)c(-c4cc(O)cc(F)c4)cc3C... 1.0 2.0 ... 1.07251 -0.53001 5.90004 0

22612 22612 CCC(OC(C)=O)C(CC(C)N(C)C)(c1ccccc1)c1ccccc1 1.0 3.0 ... 2.02918 1.18796 4.65460 0

22613 22613 Cc1ccccc1C(OCCN(C)C)c1ccccc1 1.0 1.0 ... 2.03991 1.20885 3.66262 0

22614 22614 COc1ccc(CNc2ccc(-c3ccccc3OC)nn2)cc1 0.0 0.0 ... 1.90957 0.72636 3.77290 0

[22615 rows x 1193 columns]

0

cluster0

index SMILES y C-001 ... MINaaCH MINaasC SLogP cluster

0 0 CN1CCC(c2ccccc2CC2CCN(c3ccc(C4(O)CCC4)cc3)C2=O... 1.0 0.0 ... 2.00446 0.96090 4.46290 0

1 1 C=CCn1c(SCCOc2ccc(OC)cc2)nnc1C(C)NC(=O)Cc1ccc(... 0.0 1.0 ... 1.84188 0.68199 4.07230 0

2 2 O=c1cc(OCc2ccccc2)ccn1-c1ccc(OCCN2CCCC2)cc1 1.0 0.0 ... 1.51726 0.56664 3.89110 0

3 3 CC(C)N(C)[C@@H]1CC[C@H](N2CCC(NC(=O)c3cc(F)cc(... 0.0 5.0 ... 0.30364 -1.29347 4.26600 0

4 4 C1CCc2c(nc3nnnn3c2N2CCOCC2)C1 0.0 0.0 ... 1.90268 1.13241 0.23470 0

... ... ... ... ... ... ... ... ... ...

22610 22610 Nc1nc2c(s1)CCC2C(=O)Nc1ccc(CC2CCC(C(O)c3cccnc3... 0.0 0.0 ... 1.71537 0.53479 3.18550 0

22611 22611 Cc1cc(Cl)cc(N2C=Nc3cc(C)c(-c4cc(O)cc(F)c4)cc3C... 1.0 2.0 ... 1.07251 -0.53001 5.90004 0

22612 22612 CCC(OC(C)=O)C(CC(C)N(C)C)(c1ccccc1)c1ccccc1 1.0 3.0 ... 2.02918 1.18796 4.65460 0

22613 22613 Cc1ccccc1C(OCCN(C)C)c1ccccc1 1.0 1.0 ... 2.03991 1.20885 3.66262 0

22614 22614 COc1ccc(CNc2ccc(-c3ccccc3OC)nn2)cc1 0.0 0.0 ... 1.90957 0.72636 3.77290 0

[22615 rows x 1193 columns]

index SMILES y C-001 ... MINaaCH MINaasC SLogP cluster

18080 18080 Cc1ccc2c(c1)Cc1nc3ccccc3nc1-2 0.0 1.0 ... 2.00834 1.05898 3.50942 0

6389 6389 COC(=O)c1c(NC(C)=O)csc1C 0.0 2.0 ... 1.72519 0.43769 1.80152 0

89 89 NC1=NC(=O)C2CCCN2c2ccc(cc2)OC/C=C/CNCC(=O)Nc2c... 1.0 0.0 ... 1.66881 0.28984 3.07000 0

10768 10768 O=C(Nc1ccc(Cl)c(Cl)c1)N(Cc1cccnc1)Cc1ccco1 1.0 0.0 ... 1.57784 0.37899 5.21570 0

17167 17167 O=C(CNc1n[nH]c2ccc(C(F)(F)F)cc12)NC1CN([C@H]2C... 0.0 0.0 ... 1.00465 -0.75192 3.91540 0

[5 rows x 1193 columns]

index SMILES y C-001 ... MINaaCH MINaasC SLogP cluster

18080 18080 Cc1ccc2c(c1)Cc1nc3ccccc3nc1-2 0.0 1.0 ... 2.00834 1.05898 3.50942 0

6389 6389 COC(=O)c1c(NC(C)=O)csc1C 0.0 2.0 ... 1.72519 0.43769 1.80152 0

89 89 NC1=NC(=O)C2CCCN2c2ccc(cc2)OC/C=C/CNCC(=O)Nc2c... 1.0 0.0 ... 1.66881 0.28984 3.07000 0

10768 10768 O=C(Nc1ccc(Cl)c(Cl)c1)N(Cc1cccnc1)Cc1ccco1 1.0 0.0 ... 1.57784 0.37899 5.21570 0

17167 17167 O=C(CNc1n[nH]c2ccc(C(F)(F)F)cc12)NC1CN([C@H]2C... 0.0 0.0 ... 1.00465 -0.75192 3.91540 0

[5 rows x 1193 columns]

Train set contains:

6441 negative values

9389 positive values

ratio neg / pos: 0.686015550111833

Test set contains:

2761 negative values

4024 positive values

ratio neg / pos: 0.6861332007952287

If you find this imbalanced, try to decomment line 44 of split\_by\_kmeans.py module. It can give an error!

The following files have been created:

C:\Users\Vicente\Desktop\model herg\TOX\_hERGinh-cleaned\_from\_kmeans.csv

C:\Users\Vicente\Desktop\model herg\TOX\_hERGinh-train\_set.csv

C:\Users\Vicente\Desktop\model herg\TOX\_hERGinh-test\_set.csv

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

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

Please select what do you want to do:

[01] Elimination of 3D descriptors [your dataset will be saved as [Name]\_no3D]

[1] "y" transformation + dataset random order + Knn imputation

[2] Initial feature reduction: infinite, correlated, constant and empty values

[3] Generation of train and test sets based in kmeans

[4] Descriptor standarization

[5] Feature selection by RFE

[6] Feature selection by FI based on LGBM

[7] Feature selection by Permutation importance

[8] Select own features (inside the script)

[0] Exit NEO

Your choice: 4

[+] Descriptor standarization

Please select the method to standarize the descriptors:

[1] StandardScaler

[2] MinMaxScaler

Your choice (1/2)?: 1

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

These files must be called:

"TOX\_hERGinh-train\_set.csv"

"TOX\_hERGinh-test\_set.csv"

Continue (Y/n)?

The following files have been created:

C:\Users\Vicente\Desktop\model herg\TOX\_hERGinh-stand\_train\_set.csv

C:\Users\Vicente\Desktop\model herg\TOX\_hERGinh-stand\_test\_set.csv

C:\Users\Vicente\Desktop\model herg\TOX\_hERGinh-alldataset.sca

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

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

Please select what do you want to do:

[01] Elimination of 3D descriptors [your dataset will be saved as [Name]\_no3D]

[1] "y" transformation + dataset random order + Knn imputation

[2] Initial feature reduction: infinite, correlated, constant and empty values

[3] Generation of train and test sets based in kmeans

[4] Descriptor standarization

[5] Feature selection by RFE

[6] Feature selection by FI based on LGBM

[7] Feature selection by Permutation importance

[8] Select own features (inside the script)

[0] Exit NEO

Your choice: 6

[+] Reduction by FI based on lgbm

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

These files must be called:

"TOX\_hERGinh-stand\_train\_set.csv"

"TOX\_hERGinh-stand\_test\_set.csv"

Continue (Y/n)?

Please define if your model is for [1] classification or [2] regression:1

Please define your parameters for lgbm selection for classification parameters:

eval\_metric (l2/auc/binary\_logloss):auc

Training Gradient Boosting Model

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[341] valid\_0's auc: 0.912385 valid\_0's binary\_logloss: 0.369165

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[447] valid\_0's auc: 0.916708 valid\_0's binary\_logloss: 0.365176

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[407] valid\_0's auc: 0.920817 valid\_0's binary\_logloss: 0.353278

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[310] valid\_0's auc: 0.914322 valid\_0's binary\_logloss: 0.360494

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[330] valid\_0's auc: 0.910234 valid\_0's binary\_logloss: 0.369414

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[295] valid\_0's auc: 0.912314 valid\_0's binary\_logloss: 0.366377

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[442] valid\_0's auc: 0.917628 valid\_0's binary\_logloss: 0.357028

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[427] valid\_0's auc: 0.919505 valid\_0's binary\_logloss: 0.358075

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[394] valid\_0's auc: 0.922172 valid\_0's binary\_logloss: 0.346532

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[340] valid\_0's auc: 0.915009 valid\_0's binary\_logloss: 0.36855

356 features with zero importance after one-hot encoding.

Please, close the plots to continue...

420 features required for 0.90 of cumulative importance

Please, close the plots to continue...

419 features required for cumulative importance of 0.90 after one hot encoding.

770 features do not contribute to cumulative importance of 0.90.

feature importance normalized\_importance cumulative\_importance

0 SLogP 100.1 0.008938 0.008938

1 PEOE\_VSA7 77.9 0.006956 0.015894

2 JhetZ 73.0 0.006518 0.022413

3 SpMAD\_Z 69.6 0.006215 0.028628

4 SssCH2 66.0 0.005893 0.034521

... ... ... ... ...

946 B03[C-Si] 0.0 0.000000 1.000000

945 B03[N-Br] 0.0 0.000000 1.000000

944 B03[O-P] 0.0 0.000000 1.000000

964 Cl-087 0.0 0.000000 1.000000

1188 B03[Si-N] 0.0 0.000000 1.000000

[1189 rows x 4 columns]

The following files have been created:

C:\Users\Vicente\Desktop\model herg\TOX\_hERGinh-train\_featured\_importances.csv

Now you can select a number of features.

Please, indicate the number of selected features: 20

Selected features:

['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']

Size of the database, preimputation: (15830, 20)

[+] fitting

[+] transforming

Size of the database, postimputation: (15830, 20)

The following files have been created:

C:\Users\Vicente\Desktop\model herg\TOX\_hERGinh-train\_reduction\_GBM.csv

C:\Users\Vicente\Desktop\model herg\TOX\_hERGinh-test\_reduction\_GBM.csv

C:\Users\Vicente\Desktop\model herg\TOX\_hERGinh-train\_set.sca

train\_set (15830, 22)

test\_set (6785, 22)

Do you agree with that number of features?(y/n): y

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

Thanks for using NEO!

(Protocosas) D:\scripts\generate\_models\NEO>