# Protocolo seguido para generar el modelo de Inhibición CYP2C9

## Endpoint

Inhibidores o no inhibidores para CYP2C9

## Origen de los datos

Los datos vienen de la herramienta Deep-PK: <https://biosig.lab.uq.edu.au/deeppk/data>

## Tratamiento de los datos

Se obtienen un total de tres archivos csv, separados en test, train y validation. Para procesar estos archivos se ha empleado el script “prepare\_preprocessed\_and\_Hygieia”. El archivo final contiene un total de 19274 moléculas, de las cuales 6919 formar parte de la clase inhibidores.

## Transformación de la “y”

No hay transformación

## Train/test ratio

Se ha realizado una partición del 80% para el train y del 20% para el test. Durante el proceso de partición se han mantenido los compuestos señalados como *quite dissimilar*.

## Scaler

Se ha usado el *Standard Scaler*

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

######################### 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/Enrique/Documents/GitHub/IRB/Models/CYP2C9 Inhibitor/

Please input your MODEL NAME (enter to: Af\_MIC80\_no3D): IRB\_CYP2C9inh\_DeepPK\_final

######################### 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/Enrique/Documents/GitHub/IRB/Models/CYP2C9 Inhibitor/" folder is needed

This file must be called: "IRB\_CYP2C9inh\_DeepPK\_final-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/Enrique/Documents/GitHub/IRB/Models/CYP2C9 Inhibitor/IRB\_CYP2C9inh\_DeepPK\_final-calculated\_preimputation.csv

[+] dataset imputation

Size of the database, preimputation: (19274, 4676)

Warning! Some of descriptors have too many NaN values and have been removed: ['LPRS', 'VDA', 'MDDD', 'MAXsLi', 'MAXssBe', 'MAXssssBe', 'MAXssBH', 'MAXsssB', 'MAXssssB', 'MAXsCH3', 'MAXdCH2', 'MAXssCH2', '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', 'MINssCH2', '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: (19274, 4519)

C-001 C-002 C-003 C-004 C-005 C-006 C-007 C-008 ... SsssPbH SssssPb MAXaaCH MAXaasC MINaaCH MINaasC SLogP SMR

0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 ... 0.0 0.0 1.42457 0.35019 1.05272 -0.46721 2.29888 77.4637

1 2.0 3.0 0.0 0.0 0.0 4.0 0.0 0.0 ... 0.0 0.0 1.96025 1.04003 1.65293 0.42938 3.21910 116.5367

2 1.0 0.0 0.0 0.0 0.0 1.0 0.0 1.0 ... 0.0 0.0 1.66751 0.94736 1.45403 0.08217 -0.38768 63.0851

3 0.0 3.0 0.0 0.0 0.0 4.0 0.0 1.0 ... 0.0 0.0 2.03201 1.10717 1.94030 0.92966 2.71790 88.1698

4 0.0 0.0 0.0 0.0 1.0 4.0 0.0 0.0 ... 0.0 0.0 2.13837 1.07850 1.64480 0.66778 2.92028 99.3990

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

19269 0.0 0.0 0.0 0.0 1.0 4.0 0.0 0.0 ... 0.0 0.0 2.15078 1.10117 1.63482 0.86261 3.14200 94.5580

19270 0.0 5.0 1.0 0.0 2.0 0.0 0.0 0.0 ... 0.0 0.0 1.70963 1.02529 1.65005 0.48802 3.85530 93.2342

19271 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 ... 0.0 0.0 1.91965 0.98986 1.73712 0.15065 4.10412 83.2430

19272 0.0 6.0 0.0 0.0 1.0 0.0 0.0 4.0 ... 0.0 0.0 2.28761 1.45796 2.16812 1.38771 4.51640 100.8150

19273 0.0 0.0 0.0 0.0 2.0 1.0 0.0 3.0 ... 0.0 0.0 1.33893 0.15622 1.23097 0.15622 -1.36210 86.2083

[19274 rows x 4519 columns]

The following files have been created:

C:/Users/Enrique/Documents/GitHub/IRB/Models/CYP2C9 Inhibitor/IRB\_CYP2C9inh\_DeepPK\_final-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/Enrique/Documents/GitHub/IRB/Models/CYP2C9 Inhibitor/" folder is needed

This file must be called: "IRB\_CYP2C9inh\_DeepPK\_final-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.

2187 features with a correlation magnitude greater than 0.90.

718 features with a single unique value.

Data has not been one-hot encoded

Removed 2905 features including one-hot features.

The following files have been created:

C:/Users/Enrique/Documents/GitHub/IRB/Models/CYP2C9 Inhibitor/IRB\_CYP2C9inh\_DeepPK\_final-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"): 0.2

A file located in "C:/Users/Enrique/Documents/GitHub/IRB/Models/CYP2C9 Inhibitor/" folder is needed

This file must be called: "IRB\_CYP2C9inh\_DeepPK\_final-initial\_reduction.csv"

Continue (Y/n)?y

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

OPTIMAL NUMBER OF CLUSTERS: 4

NUMBER OF CLUSTERS: 4

SETS: {19256, 1, 3, 14}

ALERTS!!

compound number: 1915

SMILE: O=C1c2ccccc2C(=O)c2c(Nc3cc4c5c(ccc6c7ccc8c9c(cc(Nc%10cccc%11c%10C(=O)c%10ccccc%10C%11=O)c(c3c56)c97)-c3ccccc3C8=O)C(=O)c3ccccc3-4)cccc21 0.0

compound number: 3186

SMILE: C=C(NC(=O)C(=C)NC(=O)c1csc(C2=N[C@@H]3c4csc(n4)[C@@H]4NC(=O)c5csc(n5)[C@@H]([C@](C)(O)[C@H](C)O)NC(=O)C5CSC(=N5)C(=CC)NC(=O)C([C@H](C)O)NC(=O)c5csc(n5)[C@]3(CC2)NC(=O)C(C)NC(=O)C(C)=NC(=O)C(C)NC(=O)C(C(C)CC)N[C@H]2C=Cc3c([C@H](C)O)cc(nc3[C@@H]2O)C(=O)O[C@H]4C)n1)C(N)=O 0.0

compound number: 6477

SMILE: C=C(NC(=O)C(C)=NC(=O)c1csc(C2=NC3c4csc(n4)C4NC(=O)c5csc(n5)C(C(C)(O)C(C)O)NC(=O)C5CSC(=N5)C(=CC)NC(=O)C(C(C)O)NC(=O)c5csc(n5)C3(CC2)NC(=O)C(C)NC(=O)C(C)=NC(=O)C(C)NC(=O)C(C(C)CC)NC2C=Cc3c(C(C)O)cc(nc3C2O)C(=O)OC4C)n1)C(N)=O 0.0

compound number: 14298

SMILE: CC(C)(C)c1cc2c(O)c(c1)Cc1cc(C(C)(C)C)cc(c1O)Cc1cc(C(C)(C)C)cc(c1O)Cc1cc(C(C)(C)C)cc(c1O)Cc1cc(C(C)(C)C)cc(c1O)Cc1cc(C(C)(C)C)cc(c1O)Cc1cc(C(C)(C)C)cc(c1O)Cc1cc(C(C)(C)C)cc(c1O)C2 0.0

[1915, 3186, 6477, 14298]

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: {19256, 17, 1}

ALERTS!!

compound number: 1915

SMILE: O=C1c2ccccc2C(=O)c2c(Nc3cc4c5c(ccc6c7ccc8c9c(cc(Nc%10cccc%11c%10C(=O)c%10ccccc%10C%11=O)c(c3c56)c97)-c3ccccc3C8=O)C(=O)c3ccccc3-4)cccc21 0.0

[1915]

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: {19264, 10}

ALERTS!!

[]

index SMILES y C-001 C-002 ... MAXaasC MINaaCH MINaasC SLogP cluster

0 0 N#CC(=Cc1ccc(O)c(O)c1)C(=O)c1ccc(O)c(O)c1 0.0 0.0 0.0 ... 0.35019 1.05272 -0.46721 2.29888 0

1 1 CCOc1ccc(CCNC(=O)CCCN2C(=O)c3ccccc3C2=O)cc1OCC 1.0 2.0 3.0 ... 1.04003 1.65293 0.42938 3.21910 0

2 2 Cc1ccc(S(=O)(=O)NN=CC(O)CO)cc1 0.0 1.0 0.0 ... 0.94736 1.45403 0.08217 -0.38768 0

3 3 O=S1(=O)CC(N2CCCCC2)C2=C(O1)c1ccccc1SC2 0.0 0.0 3.0 ... 1.10717 1.94030 0.92966 2.71790 0

4 4 CN1CCN(c2ncnc3ccc(-c4cccc(C#N)c4)cc23)CC1 0.0 0.0 0.0 ... 1.07850 1.64480 0.66778 2.92028 0

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

19269 19269 COc1ccccc1-c1ccc2ncnc(N3CCOCC3)c2c1 0.0 0.0 0.0 ... 1.10117 1.63482 0.86261 3.14200 0

19270 19270 COc1ccc(C(=O)Nc2nnc(C3CCCCC3)s2)cc1OC 0.0 0.0 5.0 ... 1.02529 1.65005 0.48802 3.85530 0

19271 19271 Cc1cccc(C(=O)n2c(Cl)c(C=O)c3ccccc32)c1 1.0 1.0 0.0 ... 0.98986 1.73712 0.15065 4.10412 0

19272 19272 CN1C2CCC1CC(OC1c3ccccc3CCc3ccccc31)C2 0.0 0.0 6.0 ... 1.45796 2.16812 1.38771 4.51640 0

19273 19273 CN(C)OP(=O)(O)OCC1OC(n2cnc3c(N)ncnc32)C(O)C1O 0.0 0.0 0.0 ... 0.15622 1.23097 0.15622 -1.36210 0

[19274 rows x 1618 columns]

0

1

cluster0

index SMILES y C-001 C-002 ... MAXaasC MINaaCH MINaasC SLogP cluster

0 0 N#CC(=Cc1ccc(O)c(O)c1)C(=O)c1ccc(O)c(O)c1 0.0 0.0 0.0 ... 0.35019 1.05272 -0.46721 2.29888 0

1 1 CCOc1ccc(CCNC(=O)CCCN2C(=O)c3ccccc3C2=O)cc1OCC 1.0 2.0 3.0 ... 1.04003 1.65293 0.42938 3.21910 0

2 2 Cc1ccc(S(=O)(=O)NN=CC(O)CO)cc1 0.0 1.0 0.0 ... 0.94736 1.45403 0.08217 -0.38768 0

3 3 O=S1(=O)CC(N2CCCCC2)C2=C(O1)c1ccccc1SC2 0.0 0.0 3.0 ... 1.10717 1.94030 0.92966 2.71790 0

4 4 CN1CCN(c2ncnc3ccc(-c4cccc(C#N)c4)cc23)CC1 0.0 0.0 0.0 ... 1.07850 1.64480 0.66778 2.92028 0

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

19269 19269 COc1ccccc1-c1ccc2ncnc(N3CCOCC3)c2c1 0.0 0.0 0.0 ... 1.10117 1.63482 0.86261 3.14200 0

19270 19270 COc1ccc(C(=O)Nc2nnc(C3CCCCC3)s2)cc1OC 0.0 0.0 5.0 ... 1.02529 1.65005 0.48802 3.85530 0

19271 19271 Cc1cccc(C(=O)n2c(Cl)c(C=O)c3ccccc32)c1 1.0 1.0 0.0 ... 0.98986 1.73712 0.15065 4.10412 0

19272 19272 CN1C2CCC1CC(OC1c3ccccc3CCc3ccccc31)C2 0.0 0.0 6.0 ... 1.45796 2.16812 1.38771 4.51640 0

19273 19273 CN(C)OP(=O)(O)OCC1OC(n2cnc3c(N)ncnc32)C(O)C1O 0.0 0.0 0.0 ... 0.15622 1.23097 0.15622 -1.36210 0

[19264 rows x 1618 columns]

index SMILES y C-001 C-002 ... MAXaasC MINaaCH MINaasC SLogP cluster

9976 9976 CC(C)=C1C(=O)C(c2ccccc2)=C2CN3C(=O)N(CCc4ccccc... 1.0 2.0 2.0 ... 0.967530 0.99351 -0.793220 6.4963 0

7540 7540 O=C(CCCN1CC=C(n2c(=O)[nH]c3ccccc32)CC1)c1ccc(F... 0.0 0.0 3.0 ... 0.556420 1.31631 -0.331700 3.6784 0

3444 3444 O=C1C=CC(=O)c2c1c(O)c1c(NCCNCCO)ccc(NCCNCCO)c1c2O 0.0 0.0 0.0 ... 0.486260 1.71664 -0.371880 0.1738 0

17403 17403 COc1c(O)cc2oc3cc(O)c(CC=C(C)C)c(O)c3c(=O)c2c1C... 1.0 4.0 1.0 ... 0.370530 1.29852 -0.315760 5.2495 0

2776 2776 O=C1CN(CN2CCOCC2)C(=O)CN1CN1CCOCC1 0.0 0.0 0.0 ... 0.804027 1.89946 0.551593 -1.7634 0

[5 rows x 1618 columns]

cluster1

index SMILES y C-001 C-002 ... MAXaasC MINaaCH MINaasC SLogP cluster

1915 1915 O=C1c2ccccc2C(=O)c2c(Nc3cc4c5c(ccc6c7ccc8c9c(c... 0.0 0.0 0.0 ... 0.80409 1.67448 0.24075 13.1980 1

3186 3186 C=C(NC(=O)C(=C)NC(=O)c1csc(C2=N[C@@H]3c4csc(n4... 0.0 11.0 3.0 ... 0.22176 1.24879 -0.45472 1.1315 1

6225 6225 Oc1ccc2c3c1OC1c4[nH]c5c(c4CC4(O)C(C2)N(CC2CC2)... 0.0 0.0 10.0 ... 1.18666 1.77258 0.14163 3.8241 1

6477 6477 C=C(NC(=O)C(C)=NC(=O)c1csc(C2=NC3c4csc(n4)C4NC... 0.0 12.0 3.0 ... 0.22265 1.24983 -0.45358 1.4891 1

9286 9286 CNC(CC(C)C)C(=O)NC1C(=O)NC(CC(N)=O)C(=O)NC2C(=... 0.0 4.0 3.0 ... -0.16226 0.76920 -0.88220 0.1062 1

12227 12227 CNC(CC(C)C)C(=O)NC1C(=O)NC(CC(N)=O)C(=O)NC2C(=... 0.0 4.0 3.0 ... -0.16226 0.76920 -0.88220 0.1062 1

13769 13769 Oc1ccc2c3c1O[C@@H]1c4[nH]c5c(c4C[C@@]4(O)[C@@H... 0.0 0.0 10.0 ... 1.18666 1.77258 0.14163 3.8241 1

13952 13952 Oc1ccc2c3c1O[C@H]1c4[nH]c5c(c4C[C@@]4(O)[C@@H]... 0.0 0.0 10.0 ... 1.18666 1.77258 0.14163 3.8241 1

14298 14298 CC(C)(C)c1cc2c(O)c(c1)Cc1cc(C(C)(C)C)cc(c1O)Cc... 0.0 24.0 8.0 ... 0.97946 2.03815 0.07130 20.7512 1

16993 16993 CNC(CC(C)C)C(=O)NC1C(=O)NC(CC(N)=O)C(=O)NC2C(=... 0.0 4.0 3.0 ... -0.16226 0.76920 -0.88220 0.1062 1

[10 rows x 1618 columns]

index SMILES y C-001 C-002 ... MAXaasC MINaaCH MINaasC SLogP cluster

13952 13952 Oc1ccc2c3c1O[C@H]1c4[nH]c5c(c4C[C@@]4(O)[C@@H]... 0.0 0.0 10.0 ... 1.18666 1.77258 0.14163 3.8241 1

3186 3186 C=C(NC(=O)C(=C)NC(=O)c1csc(C2=N[C@@H]3c4csc(n4... 0.0 11.0 3.0 ... 0.22176 1.24879 -0.45472 1.1315 1

14298 14298 CC(C)(C)c1cc2c(O)c(c1)Cc1cc(C(C)(C)C)cc(c1O)Cc... 0.0 24.0 8.0 ... 0.97946 2.03815 0.07130 20.7512 1

1915 1915 O=C1c2ccccc2C(=O)c2c(Nc3cc4c5c(ccc6c7ccc8c9c(c... 0.0 0.0 0.0 ... 0.80409 1.67448 0.24075 13.1980 1

6225 6225 Oc1ccc2c3c1OC1c4[nH]c5c(c4CC4(O)C(C2)N(CC2CC2)... 0.0 0.0 10.0 ... 1.18666 1.77258 0.14163 3.8241 1

[5 rows x 1618 columns]

index SMILES y C-001 C-002 ... MAXaasC MINaaCH MINaasC SLogP cluster

9976 9976 CC(C)=C1C(=O)C(c2ccccc2)=C2CN3C(=O)N(CCc4ccccc... 1.0 2.0 2.0 ... 0.967530 0.99351 -0.793220 6.4963 0

7540 7540 O=C(CCCN1CC=C(n2c(=O)[nH]c3ccccc32)CC1)c1ccc(F... 0.0 0.0 3.0 ... 0.556420 1.31631 -0.331700 3.6784 0

3444 3444 O=C1C=CC(=O)c2c1c(O)c1c(NCCNCCO)ccc(NCCNCCO)c1c2O 0.0 0.0 0.0 ... 0.486260 1.71664 -0.371880 0.1738 0

17403 17403 COc1c(O)cc2oc3cc(O)c(CC=C(C)C)c(O)c3c(=O)c2c1C... 1.0 4.0 1.0 ... 0.370530 1.29852 -0.315760 5.2495 0

2776 2776 O=C1CN(CN2CCOCC2)C(=O)CN1CN1CCOCC1 0.0 0.0 0.0 ... 0.804027 1.89946 0.551593 -1.7634 0

[5 rows x 1618 columns]

Train set contains:

9884 negative values

5535 positive values

ratio neg / pos: 1.7857271906052394

Test set contains:

2471 negative values

1384 positive values

ratio neg / pos: 1.7854046242774566

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/Enrique/Documents/GitHub/IRB/Models/CYP2C9 Inhibitor/IRB\_CYP2C9inh\_DeepPK\_final-cleaned\_from\_kmeans.csv

C:/Users/Enrique/Documents/GitHub/IRB/Models/CYP2C9 Inhibitor/IRB\_CYP2C9inh\_DeepPK\_final-train\_set.csv

C:/Users/Enrique/Documents/GitHub/IRB/Models/CYP2C9 Inhibitor/IRB\_CYP2C9inh\_DeepPK\_final-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/Enrique/Documents/GitHub/IRB/Models/CYP2C9 Inhibitor/" folder are needed

These files must be called:

"IRB\_CYP2C9inh\_DeepPK\_final-train\_set.csv"

"IRB\_CYP2C9inh\_DeepPK\_final-test\_set.csv"

Continue (Y/n)?y

The following files have been created:

C:/Users/Enrique/Documents/GitHub/IRB/Models/CYP2C9 Inhibitor/IRB\_CYP2C9inh\_DeepPK\_final-stand\_train\_set.csv

C:/Users/Enrique/Documents/GitHub/IRB/Models/CYP2C9 Inhibitor/IRB\_CYP2C9inh\_DeepPK\_final-stand\_test\_set.csv

C:/Users/Enrique/Documents/GitHub/IRB/Models/CYP2C9 Inhibitor/IRB\_CYP2C9inh\_DeepPK\_final-alldataset.sca

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