# Protocolo seguido para generar el modelo de CYP3A4

## Endpoint

Inhibition of CYP3A4 (Classification)

## Origen de los datos

Los datos vienen del repositorio disponible en la web de Deep-PK:

* <https://biosig.lab.uq.edu.au/deeppk/data>

## Tratamiento de los datos

Del Excel original nos quedamos con las columnas “SMILES\_standarized” y “label”.

**Interpretation:**A compound is considered to be a cytochrome P450 inhibitor (Class 1) if the concentration required to inhibit P450 activity by 50% is less than 10 uM.

## Transformación de la “y”

No se ha llevado a cabo

## 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*

(base) C:\Users\Enrique>activate Protocosas

(Protocosas) C:\Users\Enrique>cd Documents\GitHub\generate\_models\NEO

(Protocosas) C:\Users\Enrique\Documents\GitHub\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/Enrique/Documents/GitHub/IRB/Models/CYP3A4 Inhibitor/

Please input your MODEL NAME (enter to: Af\_MIC80\_no3D): CYP3A4\_Inhibitor

######################### 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: 01

This part will eliminate 3D descriptors from your dataset.

The output dataset will be renamed as "[The\_name\_of\_your\_dataset]\_no3D-paralel\_calculated\_with\_y.csv".

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

A file located in "C:/Users/Enrique/Documents/GitHub/IRB/Models/CYP3A4 Inhibitor/" folder is needed

This file must be called: "CYP3A4\_Inhibitor-paralel\_calculated\_with\_y.csv"

Continue (Y/n)?y

Initial number of descriptors: 4678

Final number of descriptors: 3729

The following files have been created:

C:/Users/Enrique/Documents/GitHub/IRB/Models/CYP3A4 Inhibitor/CYP3A4\_Inhibitor\_no3D-paralel\_calculated\_with\_y.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: 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/CYP3A4 Inhibitor/" folder is needed

This file must be called: "CYP3A4\_Inhibitor\_no3D-paralel\_calculated\_with\_y.csv"

Continue (Y/n)?n

######################### 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/CYP3A4 Inhibitor/" folder is needed

This file must be called: "CYP3A4\_Inhibitor\_no3D-paralel\_calculated\_with\_y.csv"

Continue (Y/n)?n

######################### 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: 0

Thanks for using NEO!

(Protocosas) C:\Users\Enrique\Documents\GitHub\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/Enrique/Documents/GitHub/IRB/Models/CYP3A4 Inhibitor/

Please input your MODEL NAME (enter to: Af\_MIC80\_no3D): CYP3A4\_Inhibitor

######################### 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/CYP3A4 Inhibitor/" folder is needed

This file must be called: "CYP3A4\_Inhibitor-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/CYP3A4 Inhibitor/CYP3A4\_Inhibitor-calculated\_preimputation.csv

[+] dataset imputation

Size of the database, preimputation: (23015, 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', 'ITH', 'ISH', 'HIC', 'HGM', 'H0u', 'H1u', 'H2u', 'H3u', 'H4u', 'H5u', 'H6u', 'H7u', 'H8u', 'HTu', 'HATS0u', 'HATS1u', 'HATS2u', 'HATS3u', 'HATS4u', 'HATS5u', 'HATS6u', 'HATS7u', 'HATS8u', 'HATSu', 'H0m', 'H1m', 'H2m', 'H3m', 'H4m', 'H5m', 'H6m', 'H7m', 'H8m', 'HTm', 'HATS0m', 'HATS1m', 'HATS2m', 'HATS3m', 'HATS4m', 'HATS5m', 'HATS6m', 'HATS7m', 'HATS8m', 'HATSm', 'H0v', 'H1v', 'H2v', 'H3v', 'H4v', 'H5v', 'H6v', 'H7v', 'H8v', 'HTv', 'HATS0v', 'HATS1v', 'HATS2v', 'HATS3v', 'HATS4v', 'HATS5v', 'HATS6v', 'HATS7v', 'HATS8v', 'HATSv', 'H0e', 'H1e', 'H2e', 'H3e', 'H4e', 'H5e', 'H6e', 'H7e', 'H8e', 'HTe', 'HATS0e', 'HATS1e', 'HATS2e', 'HATS3e', 'HATS4e', 'HATS5e', 'HATS6e', 'HATS7e', 'HATS8e', 'HATSe', 'H0p', 'H1p', 'H2p', 'H3p', 'H4p', 'H5p', 'H6p', 'H7p', 'H8p', 'HTp', 'HATS0p', 'HATS1p', 'HATS2p', 'HATS3p', 'HATS4p', 'HATS5p', 'HATS6p', 'HATS7p', 'HATS8p', 'HATSp', 'H0i', 'H1i', 'H2i', 'H3i', 'H4i', 'H5i', 'H6i', 'H7i', 'H8i', 'HTi', 'HATS0i', 'HATS1i', 'HATS2i', 'HATS3i', 'HATS4i', 'HATS5i', 'HATS6i', 'HATS7i', 'HATS8i', 'HATSi', 'H0s', 'H1s', 'H2s', 'H3s', 'H4s', 'H5s', 'H6s', 'H7s', 'H8s', 'HTs', 'HATS0s', 'HATS1s', 'HATS2s', 'HATS3s', 'HATS4s', 'HATS5s', 'HATS6s', 'HATS7s', 'HATS8s', 'HATSs', 'RCON', 'RARS', 'REIG', 'R1u', 'R2u', 'R3u', 'R4u', 'R5u', 'R6u', 'R7u', 'R8u', 'RTu', 'R1u+', 'R2u+', 'R3u+', 'R4u+', 'R5u+', 'R6u+', 'R7u+', 'R8u+', 'RTu+', 'R1m', 'R2m', 'R3m', 'R4m', 'R5m', 'R6m', 'R7m', 'R8m', 'RTm', 'R1m+', 'R2m+', 'R3m+', 'R4m+', 'R5m+', 'R6m+', 'R7m+', 'R8m+', 'RTm+', 'R1v', 'R2v', 'R3v', 'R4v', 'R5v', 'R6v', 'R7v', 'R8v', 'RTv', 'R1v+', 'R2v+', 'R3v+', 'R4v+', 'R5v+', 'R6v+', 'R7v+', 'R8v+', 'RTv+', 'R1e', 'R2e', 'R3e', 'R4e', 'R5e', 'R6e', 'R7e', 'R8e', 'RTe', 'R1e+', 'R2e+', 'R3e+', 'R4e+', 'R5e+', 'R6e+', 'R7e+', 'R8e+', 'RTe+', 'R1p', 'R2p', 'R3p', 'R4p', 'R5p', 'R6p', 'R7p', 'R8p', 'RTp', 'R1p+', 'R2p+', 'R3p+', 'R4p+', 'R5p+', 'R6p+', 'R7p+', 'R8p+', 'RTp+', 'R1i', 'R2i', 'R3i', 'R4i', 'R5i', 'R6i', 'R7i', 'R8i', 'RTi', 'R1i+', 'R2i+', 'R3i+', 'R4i+', 'R5i+', 'R6i+', 'R7i+', 'R8i+', 'RTi+', 'R1s', 'R2s', 'R3s', 'R4s', 'R5s', 'R6s', 'R7s', 'R8s', 'RTs', 'R1s+', 'R2s+', 'R3s+', 'R4s+', 'R5s+', 'R6s+', 'R7s+', 'R8s+', 'RTs+', 'PMI1', 'PMI2', 'PMI3', 'NPR1', 'NPR2', 'RDF010u', 'RDF015u', 'RDF020u', 'RDF025u', 'RDF030u', 'RDF035u', 'RDF040u', 'RDF045u', 'RDF050u', 'RDF055u', 'RDF060u', 'RDF065u', 'RDF070u', 'RDF075u', 'RDF080u', 'RDF085u', 'RDF090u', 'RDF095u', 'RDF100u', 'RDF105u', 'RDF110u', 'RDF115u', 'RDF120u', 'RDF125u', 'RDF130u', 'RDF135u', 'RDF140u', 'RDF145u', 'RDF150u', 'RDF155u', 'RDF010m', 'RDF015m', 'RDF020m', 'RDF025m', 'RDF030m', 'RDF035m', 'RDF040m', 'RDF045m', 'RDF050m', 'RDF055m', 'RDF060m', 'RDF065m', 'RDF070m', 'RDF075m', 'RDF080m', 'RDF085m', 'RDF090m', 'RDF095m', 'RDF100m', 'RDF105m', 'RDF110m', 'RDF115m', 'RDF120m', 'RDF125m', 'RDF130m', 'RDF135m', 'RDF140m', 'RDF145m', 'RDF150m', 'RDF155m', 'RDF010v', 'RDF015v', 'RDF020v', 'RDF025v', 'RDF030v', 'RDF035v', 'RDF040v', 'RDF045v', 'RDF050v', 'RDF055v', 'RDF060v', 'RDF065v', 'RDF070v', 'RDF075v', 'RDF080v', 'RDF085v', 'RDF090v', 'RDF095v', 'RDF100v', 'RDF105v', 'RDF110v', 'RDF115v', 'RDF120v', 'RDF125v', 'RDF130v', 'RDF135v', 'RDF140v', 'RDF145v', 'RDF150v', 'RDF155v', 'RDF010e', 'RDF015e', 'RDF020e', 'RDF025e', 'RDF030e', 'RDF035e', 'RDF040e', 'RDF045e', 'RDF050e', 'RDF055e', 'RDF060e', 'RDF065e', 'RDF070e', 'RDF075e', 'RDF080e', 'RDF085e', 'RDF090e', 'RDF095e', 'RDF100e', 'RDF105e', 'RDF110e', 'RDF115e', 'RDF120e', 'RDF125e', 'RDF130e', 'RDF135e', 'RDF140e', 'RDF145e', 'RDF150e', 'RDF155e', 'RDF010p', 'RDF015p', 'RDF020p', 'RDF025p', 'RDF030p', 'RDF035p', 'RDF040p', 'RDF045p', 'RDF050p', 'RDF055p', 'RDF060p', 'RDF065p', 'RDF070p', 'RDF075p', 'RDF080p', 'RDF085p', 'RDF090p', 'RDF095p', 'RDF100p', 'RDF105p', 'RDF110p', 'RDF115p', 'RDF120p', 'RDF125p', 'RDF130p', 'RDF135p', 'RDF140p', 'RDF145p', 'RDF150p', 'RDF155p', 'RDF010i', 'RDF015i', 'RDF020i', 'RDF025i', 'RDF030i', 'RDF035i', 'RDF040i', 'RDF045i', 'RDF050i', 'RDF055i', 'RDF060i', 'RDF065i', 'RDF070i', 'RDF075i', 'RDF080i', 'RDF085i', 'RDF090i', 'RDF095i', 'RDF100i', 'RDF105i', 'RDF110i', 'RDF115i', 'RDF120i', 'RDF125i', 'RDF130i', 'RDF135i', 'RDF140i', 'RDF145i', 'RDF150i', 'RDF155i', 'RDF010s', 'RDF015s', 'RDF020s', 'RDF025s', 'RDF030s', 'RDF035s', 'RDF040s', 'RDF045s', 'RDF050s', 'RDF055s', 'RDF060s', 'RDF065s', 'RDF070s', 'RDF075s', 'RDF080s', 'RDF085s', 'RDF090s', 'RDF095s', 'RDF100s', 'RDF105s', 'RDF110s', 'RDF115s', 'RDF120s', 'RDF125s', 'RDF130s', 'RDF135s', 'RDF140s', 'RDF145s', 'RDF150s', 'RDF155s', 'L1u', 'L1m', 'L1v', 'L1e', 'L1p', 'L1i', 'L1s', 'L2u', 'L2m', 'L2v', 'L2e', 'L2p', 'L2i', 'L2s', 'L3u', 'L3m', 'L3v', 'L3e', 'L3p', 'L3i', 'L3s', 'P1u', 'P1m', 'P1v', 'P1e', 'P1p', 'P1i', 'P1s', 'P2u', 'P2m', 'P2v', 'P2e', 'P2p', 'P2i', 'P2s', 'G1u', 'G1m', 'G1v', 'G1e', 'G1p', 'G1i', 'G1s', 'G2u', 'G2m', 'G2v', 'G2e', 'G2p', 'G2i', 'G2s', 'G3u', 'G3m', 'G3v', 'G3e', 'G3p', 'G3i', 'G3s', 'E1u', 'E1m', 'E1v', 'E1e', 'E1p', 'E1i', 'E1s', 'E2u', 'E2m', 'E2v', 'E2e', 'E2p', 'E2i', 'E2s', 'E3u', 'E3m', 'E3v', 'E3e', 'E3p', 'E3i', 'E3s', 'Tm', 'Tv', 'Te', 'Tp', 'Ti', 'Ts', 'Tu', 'Am', 'Av', 'Ae', 'Ap', 'Ai', 'As', 'Gu', 'Gm', 'Ku', 'Km', 'Kv', 'Ke', 'Kp', 'Ki', 'Ks', 'Du', 'Dm', 'Dv', 'De', 'Dp', 'Di', 'Ds', 'Vu', 'Vm', 'Vv', 'Ve', 'Vp', 'Vi', 'Vs', 'TDB01u', 'TDB02u', 'TDB03u', 'TDB04u', 'TDB05u', 'TDB06u', 'TDB07u', 'TDB08u', 'TDB09u', 'TDB10u', 'TDB01m', 'TDB02m', 'TDB03m', 'TDB04m', 'TDB05m', 'TDB06m', 'TDB07m', 'TDB08m', 'TDB09m', 'TDB10m', 'TDB01v', 'TDB02v', 'TDB03v', 'TDB04v', 'TDB05v', 'TDB06v', 'TDB07v', 'TDB08v', 'TDB09v', 'TDB10v', 'TDB01e', 'TDB02e', 'TDB03e', 'TDB04e', 'TDB05e', 'TDB06e', 'TDB07e', 'TDB08e', 'TDB09e', 'TDB10e', 'TDB01p', 'TDB02p', 'TDB03p', 'TDB04p', 'TDB05p', 'TDB06p', 'TDB07p', 'TDB08p', 'TDB09p', 'TDB10p', 'TDB01i', 'TDB02i', 'TDB03i', 'TDB04i', 'TDB05i', 'TDB06i', 'TDB07i', 'TDB08i', 'TDB09i', 'TDB10i', 'TDB01s', 'TDB02s', 'TDB03s', 'TDB04s', 'TDB05s', 'TDB06s', 'TDB07s', 'TDB08s', 'TDB09s', 'TDB10s', 'TDB01r', 'TDB02r', 'TDB03r', 'TDB04r', 'TDB05r', 'TDB06r', 'TDB07r', 'TDB08r', 'TDB09r', 'TDB10r', 'Mor01u', 'Mor02u', 'Mor03u', 'Mor04u', 'Mor05u', 'Mor06u', 'Mor07u', 'Mor08u', 'Mor09u', 'Mor10u', 'Mor11u', 'Mor12u', 'Mor13u', 'Mor14u', 'Mor15u', 'Mor16u', 'Mor17u', 'Mor18u', 'Mor19u', 'Mor20u', 'Mor21u', 'Mor22u', 'Mor23u', 'Mor24u', 'Mor25u', 'Mor26u', 'Mor27u', 'Mor28u', 'Mor29u', 'Mor30u', 'Mor31u', 'Mor32u', 'Mor01m', 'Mor02m', 'Mor03m', 'Mor04m', 'Mor05m', 'Mor06m', 'Mor07m', 'Mor08m', 'Mor09m', 'Mor10m', 'Mor11m', 'Mor12m', 'Mor13m', 'Mor14m', 'Mor15m', 'Mor16m', 'Mor17m', 'Mor18m', 'Mor19m', 'Mor20m', 'Mor21m', 'Mor22m', 'Mor23m', 'Mor24m', 'Mor25m', 'Mor26m', 'Mor27m', 'Mor28m', 'Mor29m', 'Mor30m', 'Mor31m', 'Mor32m', 'Mor01v', 'Mor02v', 'Mor03v', 'Mor04v', 'Mor05v', 'Mor06v', 'Mor07v', 'Mor08v', 'Mor09v', 'Mor10v', 'Mor11v', 'Mor12v', 'Mor13v', 'Mor14v', 'Mor15v', 'Mor16v', 'Mor17v', 'Mor18v', 'Mor19v', 'Mor20v', 'Mor21v', 'Mor22v', 'Mor23v', 'Mor24v', 'Mor25v', 'Mor26v', 'Mor27v', 'Mor28v', 'Mor29v', 'Mor30v', 'Mor31v', 'Mor32v', 'Mor01e', 'Mor02e', 'Mor03e', 'Mor04e', 'Mor05e', 'Mor06e', 'Mor07e', 'Mor08e', 'Mor09e', 'Mor10e', 'Mor11e', 'Mor12e', 'Mor13e', 'Mor14e', 'Mor15e', 'Mor16e', 'Mor17e', 'Mor18e', 'Mor19e', 'Mor20e', 'Mor21e', 'Mor22e', 'Mor23e', 'Mor24e', 'Mor25e', 'Mor26e', 'Mor27e', 'Mor28e', 'Mor29e', 'Mor30e', 'Mor31e', 'Mor32e', 'Mor01p', 'Mor02p', 'Mor03p', 'Mor04p', 'Mor05p', 'Mor06p', 'Mor07p', 'Mor08p', 'Mor09p', 'Mor10p', 'Mor11p', 'Mor12p', 'Mor13p', 'Mor14p', 'Mor15p', 'Mor16p', 'Mor17p', 'Mor18p', 'Mor19p', 'Mor20p', 'Mor21p', 'Mor22p', 'Mor23p', 'Mor24p', 'Mor25p', 'Mor26p', 'Mor27p', 'Mor28p', 'Mor29p', 'Mor30p', 'Mor31p', 'Mor32p', 'Mor01i', 'Mor02i', 'Mor03i', 'Mor04i', 'Mor05i', 'Mor06i', 'Mor07i', 'Mor08i', 'Mor09i', 'Mor10i', 'Mor11i', 'Mor12i', 'Mor13i', 'Mor14i', 'Mor15i', 'Mor16i', 'Mor17i', 'Mor18i', 'Mor19i', 'Mor20i', 'Mor21i', 'Mor22i', 'Mor23i', 'Mor24i', 'Mor25i', 'Mor26i', 'Mor27i', 'Mor28i', 'Mor29i', 'Mor30i', 'Mor31i', 'Mor32i', 'Mor01s', 'Mor02s', 'Mor03s', 'Mor04s', 'Mor05s', 'Mor06s', 'Mor07s', 'Mor08s', 'Mor09s', 'Mor10s', 'Mor11s', 'Mor12s', 'Mor13s', 'Mor14s', 'Mor15s', 'Mor16s', 'Mor17s', 'Mor18s', 'Mor19s', 'Mor20s', 'Mor21s', 'Mor22s', 'Mor23s', 'Mor24s', 'Mor25s', 'Mor26s', 'Mor27s', 'Mor28s', 'Mor29s', 'Mor30s', 'Mor31s', 'Mor32s', 'PBF', 'RadiusGyration', 'Spherocity', 'InertialShapeFactor', 'Eccentricity', 'Asphericity', 'PNSA1', 'PNSA2', 'PNSA3', 'PNSA4', 'PNSA5', 'PPSA1', 'PPSA2', 'PPSA3', 'PPSA4', 'PPSA5', 'DPSA1', 'DPSA2', 'DPSA3', 'DPSA4', 'DPSA5', 'FNSA1', 'FNSA2', 'FNSA3', 'FNSA4', 'FNSA5', 'FPSA1', 'FPSA2', 'FPSA3', 'FPSA4', 'FPSA5', 'WNSA1', 'WNSA2', 'WNSA3', 'WNSA4', 'WNSA5', 'WPSA1', 'WPSA2', 'WPSA3', 'WPSA4', 'WPSA5', 'RNCG', 'RPCG', 'RNCS', 'RPCS', 'TASA', 'TPSA', 'RASA', 'RPSA'] The limit is marked to the 15.0 %

[+] fitting

[+] transforming

Size of the database, postimputation: (23015, 3565)

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

0 2.0 5.0 0.0 0.0 0.0 1.0 0.0 ... 0.0 2.03287 1.188210 1.861030 0.716780 4.44424 112.1751

1 1.0 0.0 0.0 0.0 0.0 0.0 0.0 ... 0.0 2.07870 1.195560 2.008740 1.144680 3.57671 59.2170

2 0.0 0.0 0.0 0.0 1.0 1.0 0.0 ... 0.0 1.91859 0.974630 1.442610 -0.253460 2.07870 79.3389

3 0.0 3.0 0.0 0.0 0.0 1.0 0.0 ... 0.0 1.88097 1.059987 1.676067 1.059987 0.21310 33.2775

4 1.0 2.0 0.0 0.0 0.0 1.0 0.0 ... 0.0 2.02814 0.359590 1.686260 0.359590 3.67300 78.7902

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

23010 2.0 1.0 2.0 0.0 1.0 3.0 0.0 ... 0.0 1.93691 0.968170 1.920520 0.968170 3.46150 108.9590

23011 2.0 0.0 0.0 0.0 2.0 1.0 0.0 ... 0.0 1.81376 0.678700 1.680980 0.473240 1.51730 128.5644

23012 1.0 1.0 0.0 0.0 1.0 2.0 0.0 ... 0.0 2.01582 1.102540 1.778300 0.550720 4.62839 126.4957

23013 1.0 0.0 0.0 0.0 1.0 2.0 0.0 ... 0.0 1.88658 0.774080 1.773630 0.581320 4.40169 108.7947

23014 2.0 0.0 1.0 0.0 0.0 1.0 0.0 ... 0.0 1.88425 0.911980 1.884250 0.607260 3.15620 103.8322

[23015 rows x 3565 columns]

The following files have been created:

C:/Users/Enrique/Documents/GitHub/IRB/Models/CYP3A4 Inhibitor/CYP3A4\_Inhibitor-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/CYP3A4 Inhibitor/" folder is needed

This file must be called: "CYP3A4\_Inhibitor-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.

1560 features with a correlation magnitude greater than 0.90.

689 features with a single unique value.

Data has not been one-hot encoded

Removed 2249 features including one-hot features.

The following files have been created:

C:/Users/Enrique/Documents/GitHub/IRB/Models/CYP3A4 Inhibitor/CYP3A4\_Inhibitor-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/CYP3A4 Inhibitor/" folder is needed

This file must be called: "CYP3A4\_Inhibitor-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: {1, 20, 4, 22990}

ALERTS!!

compound number: 22049

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

[22049]

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: {24, 1, 22990}

ALERTS!!

compound number: 22049

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

[22049]

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: {8, 23007}

ALERTS!!

[]

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

0 0 Cc1ccc(CNC(=O)C2(NC(=O)Nc3ccc(C)cc3)CCCCC2)cc1 1.0 2.0 5.0 ... 1.188210 1.861030 0.716780 4.44424 0

1 1 Cc1csc(=S)n1-c1ccccc1 0.0 1.0 0.0 ... 1.195560 2.008740 1.144680 3.57671 0

2 2 CNC1NC(=O)c2ccccc2N1Cc1ccc(F)cc1 0.0 0.0 0.0 ... 0.974630 1.442610 -0.253460 2.07870 0

3 3 O=C(O)CNC1CCC1 1.0 0.0 3.0 ... 1.059987 1.676067 1.059987 0.21310 0

4 4 CCCCOC(=O)c1cc2c(cn1)[nH]c1ccccc12 1.0 1.0 2.0 ... 0.359590 1.686260 0.359590 3.67300 0

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

23010 23010 CO[C@H]1COC(=O)CC=C[C@@H](C)COC(=O)C(OCc2ccccc... 1.0 2.0 1.0 ... 0.968170 1.920520 0.968170 3.46150 0

23011 23011 COc1cccc(C2Oc3ccc(OC)cc3C(=NOCC(O)C3OC4OC(C)(C... 1.0 2.0 0.0 ... 0.678700 1.680980 0.473240 1.51730 0

23012 23012 COc1cc(C=C2C(=N)N3C=CSC3=NC2=O)ccc1OCCCOc1ccccc1C 1.0 1.0 1.0 ... 1.102540 1.778300 0.550720 4.62839 0

23013 23013 CCn1c(-c2ccccc2Cl)nn(CC(=O)Nc2ccccc2OC)c1=S 1.0 1.0 0.0 ... 0.774080 1.773630 0.581320 4.40169 0

23014 23014 CC(=O)N(CC(C)C(=O)O)c1c(I)cc(I)c(N)c1I 0.0 2.0 0.0 ... 0.911980 1.884250 0.607260 3.15620 0

[23015 rows x 1320 columns]

0

1

cluster0

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

0 0 Cc1ccc(CNC(=O)C2(NC(=O)Nc3ccc(C)cc3)CCCCC2)cc1 1.0 2.0 5.0 ... 1.188210 1.861030 0.716780 4.44424 0

1 1 Cc1csc(=S)n1-c1ccccc1 0.0 1.0 0.0 ... 1.195560 2.008740 1.144680 3.57671 0

2 2 CNC1NC(=O)c2ccccc2N1Cc1ccc(F)cc1 0.0 0.0 0.0 ... 0.974630 1.442610 -0.253460 2.07870 0

3 3 O=C(O)CNC1CCC1 1.0 0.0 3.0 ... 1.059987 1.676067 1.059987 0.21310 0

4 4 CCCCOC(=O)c1cc2c(cn1)[nH]c1ccccc12 1.0 1.0 2.0 ... 0.359590 1.686260 0.359590 3.67300 0

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

23010 23010 CO[C@H]1COC(=O)CC=C[C@@H](C)COC(=O)C(OCc2ccccc... 1.0 2.0 1.0 ... 0.968170 1.920520 0.968170 3.46150 0

23011 23011 COc1cccc(C2Oc3ccc(OC)cc3C(=NOCC(O)C3OC4OC(C)(C... 1.0 2.0 0.0 ... 0.678700 1.680980 0.473240 1.51730 0

23012 23012 COc1cc(C=C2C(=N)N3C=CSC3=NC2=O)ccc1OCCCOc1ccccc1C 1.0 1.0 1.0 ... 1.102540 1.778300 0.550720 4.62839 0

23013 23013 CCn1c(-c2ccccc2Cl)nn(CC(=O)Nc2ccccc2OC)c1=S 1.0 1.0 0.0 ... 0.774080 1.773630 0.581320 4.40169 0

23014 23014 CC(=O)N(CC(C)C(=O)O)c1c(I)cc(I)c(N)c1I 0.0 2.0 0.0 ... 0.911980 1.884250 0.607260 3.15620 0

[23007 rows x 1320 columns]

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

3048 3048 CN(Cc1ccco1)c1ncnc2ccc(-c3ccccc3C(F)(F)F)cc12 1.0 0.0 0.0 ... 0.757250 1.088260 -0.669730 5.54500 0

2964 2964 CC1(O)CC2C(CC1O)C2(C)C 0.0 3.0 2.0 ... 1.213347 1.700357 0.633873 1.16430 0

8003 8003 COc1ccccc1CN1CC[C@@]2(CCCN(C(=O)c3cccn3C)C2)C1 1.0 0.0 3.0 ... 1.240370 1.922860 0.787890 3.16200 0

20579 20579 Cc1noc(C)c1-c1nccc(NCCN2CCOCC2)n1 0.0 2.0 0.0 ... 0.872060 1.755240 0.645950 1.49254 0

4390 4390 COC(=O)C1=C(C)N(Cc2ccccc2)[C@]23O[C@](C(=O)OC)... 0.0 1.0 0.0 ... 0.690650 1.674880 0.448460 1.77110 0

[5 rows x 1320 columns]

cluster1

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

8445 8445 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

9743 9743 C=C(NC(=O)C(=C)NC(=O)c1csc(C2=NC3c4csc(n4)C4NC... 0.0 11.0 3.0 ... 0.22176 1.24879 -0.45472 1.1315 1

9898 9898 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

13314 13314 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

19669 19669 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

21025 21025 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

22049 22049 O=C1c2ccccc2C(=O)c2c(Nc3cc4c5c(ccc6c7ccc8c9c(c... 0.0 0.0 0.0 ... 0.80409 1.67448 0.24075 13.1980 1

22235 22235 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

[8 rows x 1320 columns]

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

19669 19669 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

9743 9743 C=C(NC(=O)C(=C)NC(=O)c1csc(C2=NC3c4csc(n4)C4NC... 0.0 11.0 3.0 ... 0.22176 1.24879 -0.45472 1.1315 1

22235 22235 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

9898 9898 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

8445 8445 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

[5 rows x 1320 columns]

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

3048 3048 CN(Cc1ccco1)c1ncnc2ccc(-c3ccccc3C(F)(F)F)cc12 1.0 0.0 0.0 ... 0.757250 1.088260 -0.669730 5.54500 0

2964 2964 CC1(O)CC2C(CC1O)C2(C)C 0.0 3.0 2.0 ... 1.213347 1.700357 0.633873 1.16430 0

8003 8003 COc1ccccc1CN1CC[C@@]2(CCCN(C(=O)c3cccn3C)C2)C1 1.0 0.0 3.0 ... 1.240370 1.922860 0.787890 3.16200 0

20579 20579 Cc1noc(C)c1-c1nccc(NCCN2CCOCC2)n1 0.0 2.0 0.0 ... 0.872060 1.755240 0.645950 1.49254 0

4390 4390 COC(=O)C1=C(C)N(Cc2ccccc2)[C@]23O[C@](C(=O)OC)... 0.0 1.0 0.0 ... 0.690650 1.674880 0.448460 1.77110 0

[5 rows x 1320 columns]

Train set contains:

11038 negative values

7373 positive values

ratio neg / pos: 1.4970839549708395

Test set contains:

2761 negative values

1843 positive values

ratio neg / pos: 1.4981009224091155

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/CYP3A4 Inhibitor/CYP3A4\_Inhibitor-cleaned\_from\_kmeans.csv

C:/Users/Enrique/Documents/GitHub/IRB/Models/CYP3A4 Inhibitor/CYP3A4\_Inhibitor-train\_set.csv

C:/Users/Enrique/Documents/GitHub/IRB/Models/CYP3A4 Inhibitor/CYP3A4\_Inhibitor-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/CYP3A4 Inhibitor/" folder are needed

These files must be called:

"CYP3A4\_Inhibitor-train\_set.csv"

"CYP3A4\_Inhibitor-test\_set.csv"

Continue (Y/n)?y

The following files have been created:

C:/Users/Enrique/Documents/GitHub/IRB/Models/CYP3A4 Inhibitor/CYP3A4\_Inhibitor-stand\_train\_set.csv

C:/Users/Enrique/Documents/GitHub/IRB/Models/CYP3A4 Inhibitor/CYP3A4\_Inhibitor-stand\_test\_set.csv

C:/Users/Enrique/Documents/GitHub/IRB/Models/CYP3A4 Inhibitor/CYP3A4\_Inhibitor-alldataset.sca