Report

ROBERT v 0.0.1 2023/05/02 20:47:18
Citation: ROBERT v 0.0.1, Dalmau, D.; Alegre-Requena, J. V., 2023. https://github.com/jvalegre/robert

Command line used in ROBERT: robert --ignore ['solvent','sample'] --y dG --csv_name Bandar_db.csv --epoch 10 --pfi_max 2 --train [60,70,80]

CURATE

o Starting data curation with the CURATE module

o Database Bandar_db.csv loaded successfully, including:
- 20 datapoints
- 18 accepted descriptors
- 2 ignored descriptors
- 0 discarded descriptors

o Analyzing categorical variables
- No categorical variables were found.

o Duplication filters activated
Excluded datapoints:

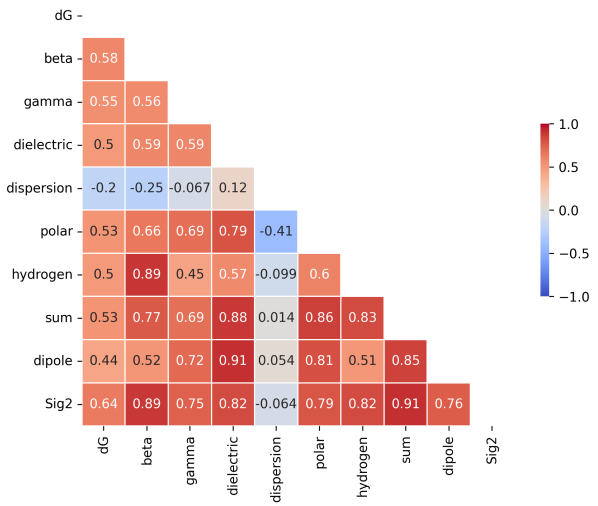
o Correlation filter activated with these thresholds: thres_x = 0.85, thres_y = 0.02
Excluded descriptors:
- n: R**2 = 0.02 with the dG values
- Sig3: R**2 = 0.89 with beta
- Hbond_acc: R**2 = 0.91 with beta
- B: R**2 = 0.99 with beta
- MV_boltz: R**2 = 0.0 with the dG values
- area: R**2 = 0.02 with the dG values
- volume: R**2 = 0.01 with the dG values
- V: R**2 = 0.01 with the dG values

o 12 columns remaining after applying correlation filters:
- solvent
- dG
- beta
- gamma
- dielectric
- dispersion
- polar
- hydrogen
- sum
- dipole
- Sig2
- sample

o The Pearson heatmap was stored in C:\Users\David\Desktop\Artculo ROBERT\Pruebas_ROBERT\Regression\Bandar\CURATE\Pearson_heatmap.png.

o The curated database was stored in C:\Users\David\Desktop\Artculo ROBERT\Pruebas_ROBERT\Regression\Bandar\CURATE\Bandar_db_CURATE.csv.

Time CURATE: 0.78 seconds



GENERATE

o Starting generation of ML models with the GENERATE module

o Database C:\Users\David\Desktop\Art&#65533;culo ROBERT\Pruebas_ROBERT\Regression\Bandar\CURATE\Bandar_db_CURATE.csv loaded successfully, including:
- 20 datapoints
- 10 accepted descriptors
- 2 ignored descriptors
- 0 discarded descriptors

o Starting heatmap scan with 4 ML models ['RF', 'GB', 'NN', 'VR'] and 3 training sizes [60, 70, 80]
.
Heatmap generation:
- 1/12
- 2/12
- 3/12
- 4/12
- 5/12

x The PFI filter was disabled for model GB_70 (no variables passed)
- 6/12
- 7/12
- 8/12

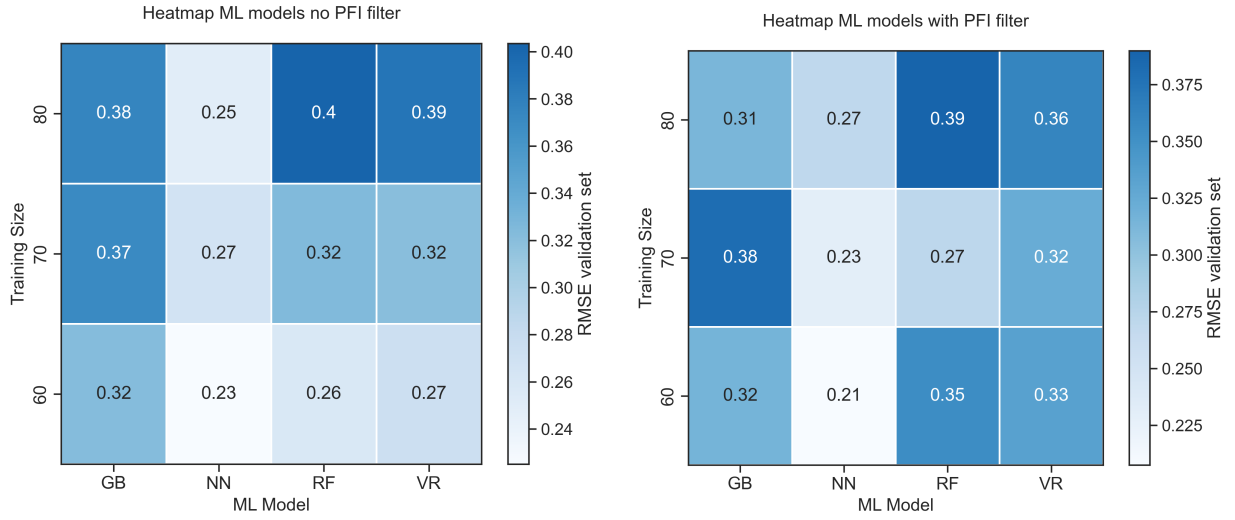
x The PFI filter was disabled for model RF_80 (no variables passed)
- 9/12
- 10/12
- 11/12

x The PFI filter was disabled for model VR_80 (no variables passed)
- 12/12

o Heatmap ML models no PFI filter succesfully created in C:\Users\David\Desktop\Art&#65533;culo ROBERT\Pruebas_ROBERT\Regression\Bandar\GENERATE\Raw_data

o Heatmap ML models with PFI filter succesfully created in C:\Users\David\Desktop\Art&#65533;culo ROBERT\Pruebas_ROBERT\Regression\Bandar\GENERATE\Raw_data

Time GENERATE: 25.03 seconds



PREDICT

o Representation of predictions and analysis of ML models with the PREDICT module

o ML model NN_60.csv (with no PFI filter) and its corresponding Xy database were loaded successfully, including:
- Target value: dG
- Model: NN
- Descriptors: ['beta', 'gamma', 'dielectric', 'dispersion', 'polar', 'hydrogen', 'sum', 'dipole', 'Sig2']
- Training points: 12
- Validation points: 8
  - Train set with predicted results: NN_60_train_No_PFI.csv
  - Validation set with predicted results: NN_60_valid_No_PFI.csv

o Saving graphs and CSV databases in C:\Users\David\Desktop\Art♠culo ROBERT\Pruebas_ROBERT\Regression\Bandar\PREDICT:
  - Graph in: C:\Users\David\Desktop\Art♠culo ROBERT\Pruebas_ROBERT\Regression\Bandar\PREDICT/Results_NN_60_No_PFI.png

o Results saved in C:\Users\David\Desktop\Art♠culo ROBERT\Pruebas_ROBERT\Regression\Bandar\PREDICT/Results_NN_60_No_PFI.dat:
  - Points Train:Validation = 12:8
  - Proportion Train:Validation = 60:40
  - Train : R2 = 0.51, MAE = 0.32, RMSE = 0.4
  - Validation : R2 = 0.51, MAE = 0.18, RMSE = 0.23

o SHAP plot saved in C:\Users\David\Desktop\Art♠culo ROBERT\Pruebas_ROBERT\Regression\Bandar\PREDICT/SHAP_NN_60_No_PFI.png
o SHAP values saved in C:\Users\David\Desktop\Art♠culo ROBERT\Pruebas_ROBERT\Regression\Bandar\PREDICT/SHAP_NN_60_No_PFI.dat:
  - gamma = min: -0.24, max: 0.21
  - beta = min: -0.12, max: 0.19
  - Sig2 = min: -0.043, max: 0.13
  - dispersion = min: -0.052, max: 0.13
  - polar = min: -0.088, max: 0.081
  - dipole = min: -0.029, max: 0.063
  - dielectric = min: -0.071, max: 0.059
  - sum = min: -0.045, max: 0.042
  - hydrogen = min: -0.014, max: 0.0088

o PFI plot saved in C:\Users\David\Desktop\Art♠culo ROBERT\Pruebas_ROBERT\Regression\Bandar\PREDICT/PFI_NN_60_No_PFI.png
o PFI values saved in C:\Users\David\Desktop\Art♠culo ROBERT\Pruebas_ROBERT\Regression\Bandar\PREDICT/PFI_NN_60_No_PFI.dat:
  Original score (from model.score, R2) = 0.5
  - gamma = 1.2 +- 0.6
  - beta = 0.19 +- 0.22
  - dielectric = 0.14 +- 0.24
  - dipole = 0.13 +- 0.13

- Sig2 = 0.098 +- 0.051
- sum = 0.065 +- 0.081
- polar = 0.011 +- 0.18
- hydrogen = 0.0082 +- 0.022
- dispersion = -0.11 +- 0.09

o Outliers plot saved in C:\Users\David\Desktop\Art&#65533;culo ROBERT\Pruebas_ROBERT\Regression\Bandar\PREDICT/Outliers_NN_60_No_PFI.png
o Outlier values saved in C:\Users\David\Desktop\Art&#65533;culo ROBERT\Pruebas_ROBERT\Regression\Bandar\PREDICT/Outliers_NN_60_No_PFI.dat:
x No variable names found or names option not specified! Outliers will be printed with no names
Train: 1 outliers out of 12 datapoints (8.3%)
Validation: 0 outliers out of 8 datapoints (0.0%)

o ML model NN_60_PFI.csv (with PFI filter) and its corresponding Xy database were loaded successfully, including:
- Target value: dG
- Model: NN
- Descriptors: ['beta', 'gamma']
- Training points: 12
- Validation points: 8
- Train set with predicted results: NN_60_train_PFI.csv
- Validation set with predicted results: NN_60_valid_PFI.csv

o Saving graphs and CSV databases in C:\Users\David\Desktop\Art&#65533;culo ROBERT\Pruebas_ROBERT\Regression\Bandar\PREDICT:
- Graph in: C:\Users\David\Desktop\Art&#65533;culo ROBERT\Pruebas_ROBERT\Regression\Bandar\PREDICT/Results_NN_60_PFI.png

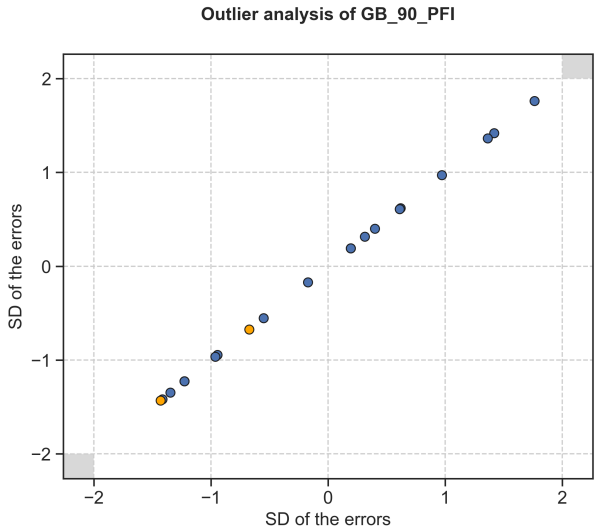
o Results saved in C:\Users\David\Desktop\Art&#65533;culo ROBERT\Pruebas_ROBERT\Regression\Bandar\PREDICT/Results_NN_60_PFI.dat:
- Points Train:Validation = 12:8
- Proportion Train:Validation = 60:40
- Train : R2 = 0.47, MAE = 0.28, RMSE = 0.39
- Validation : R2 = 0.6, MAE = 0.18, RMSE = 0.21

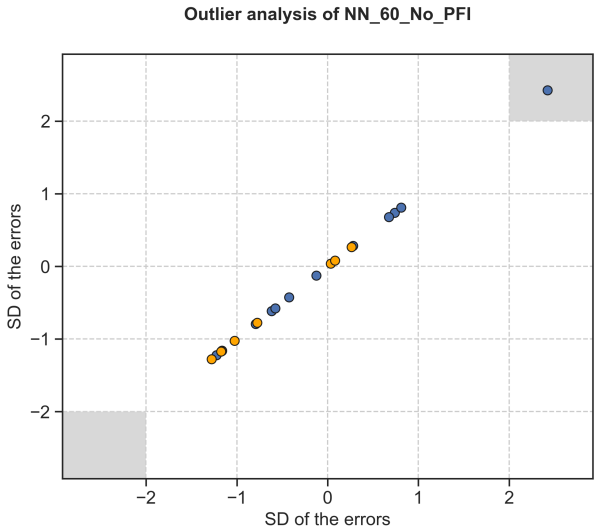
o SHAP plot saved in C:\Users\David\Desktop\Art&#65533;culo ROBERT\Pruebas_ROBERT\Regression\Bandar\PREDICT/SHAP_NN_60_PFI.png
o SHAP values saved in C:\Users\David\Desktop\Art&#65533;culo ROBERT\Pruebas_ROBERT\Regression\Bandar\PREDICT/SHAP_NN_60_PFI.dat:
- beta = min: -0.21, max: 0.4
- gamma = min: -0.22, max: 0.24

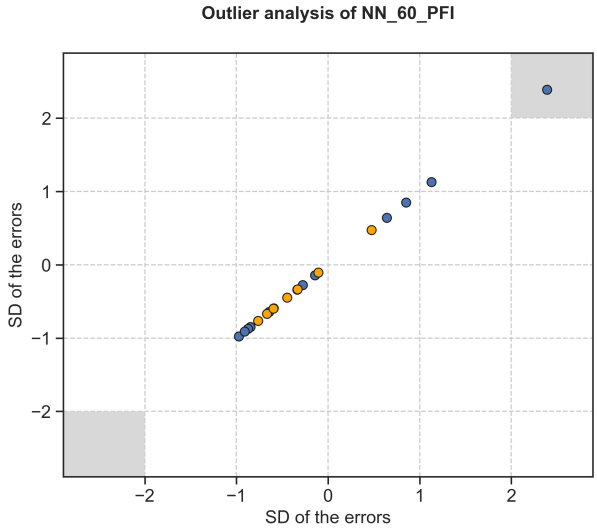
o PFI plot saved in C:\Users\David\Desktop\Art&#65533;culo ROBERT\Pruebas_ROBERT\Regression\Bandar\PREDICT/PFI_NN_60_PFI.png
o PFI values saved in C:\Users\David\Desktop\Art&#65533;culo ROBERT\Pruebas_ROBERT\Regression\Bandar\PREDICT/PFI_NN_60_PFI.dat:
Original score (from model.score, R2) = 0.58
- beta = 0.55 +- 0.25
- gamma = 0.51 +- 0.3

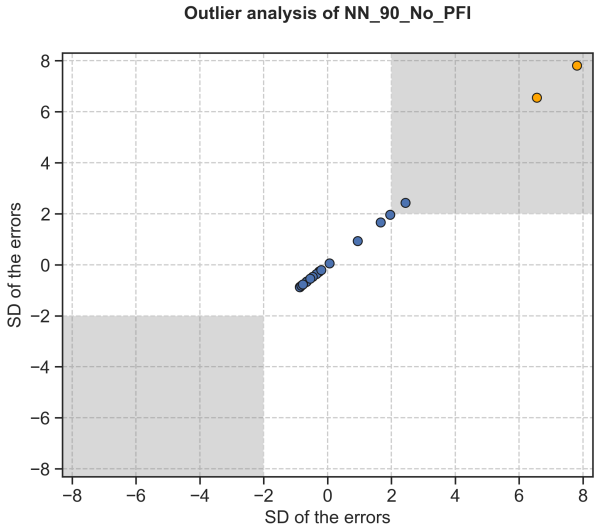
o Outliers plot saved in C:\Users\David\Desktop\Art&#65533;culo ROBERT\Pruebas_ROBERT\Regression\Bandar\PREDICT/Outliers_NN_60_PFI.png
o Outlier values saved in C:\Users\David\Desktop\Art&#65533;culo ROBERT\Pruebas_ROBERT\Regression\Bandar\PREDICT/Outliers_NN_60_PFI.dat:
x No variable names found or names option not specified! Outliers will be printed with no names
Train: 1 outliers out of 12 datapoints (8.3%)
Validation: 0 outliers out of 8 datapoints (0.0%)

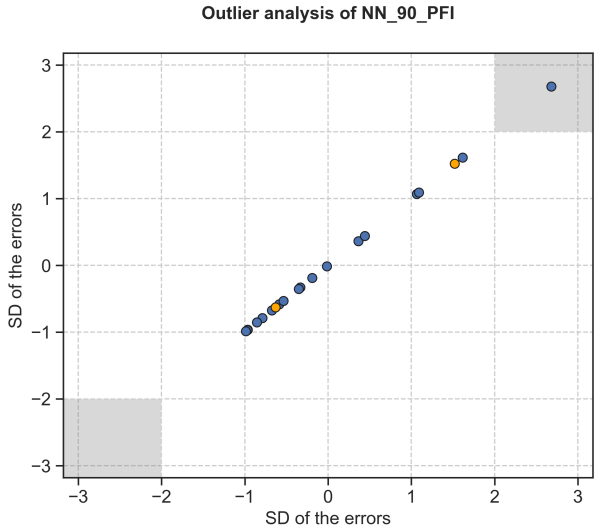
Time PREDICT: 8.72 seconds

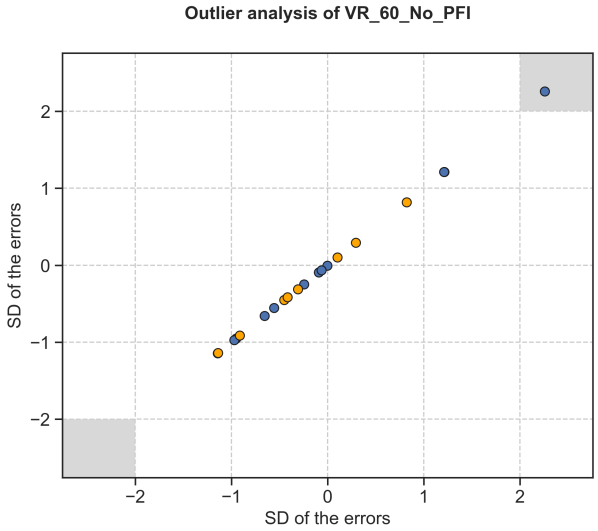

Outlier analysis of GB_90_PFI
SD of the errors
SD of the errors


Outlier analysis of NN_60_No_PFI
SD of the errors
SD of the errors

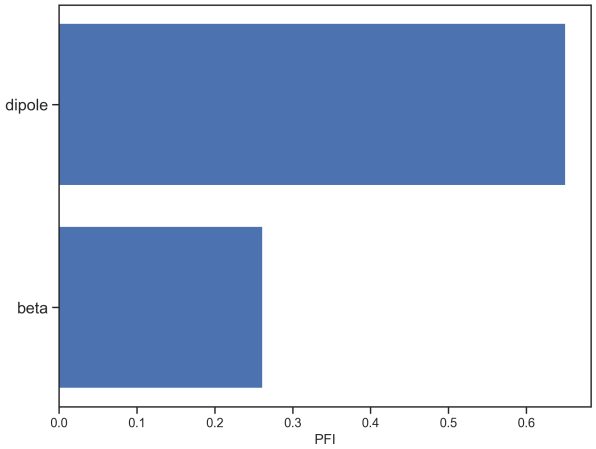

Outlier analysis of NN_60_PFI
SD of the errors
SD of the errors


Outlier analysis of NN_90_No_PFI
SD of the errors
SD of the errors

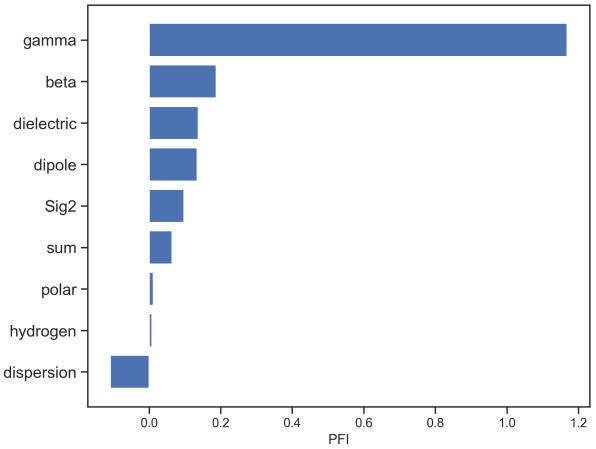

Outlier analysis of NN_90_PFI
SD of the errors
SD of the errors


Outlier analysis of VR_60_No_PFI
SD of the errors
SD of the errors

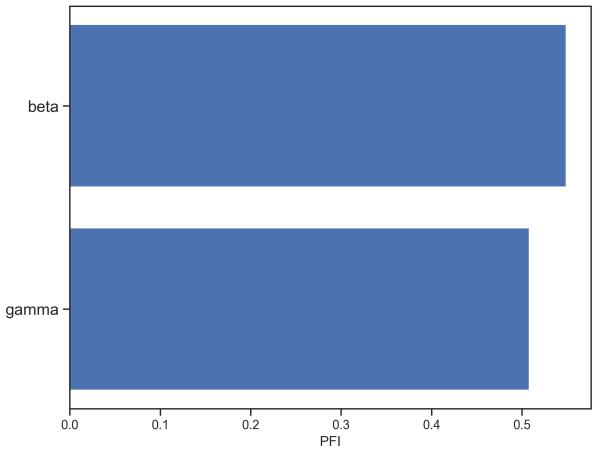
Permutation feature importances (PFIs) of GB_90_PFI



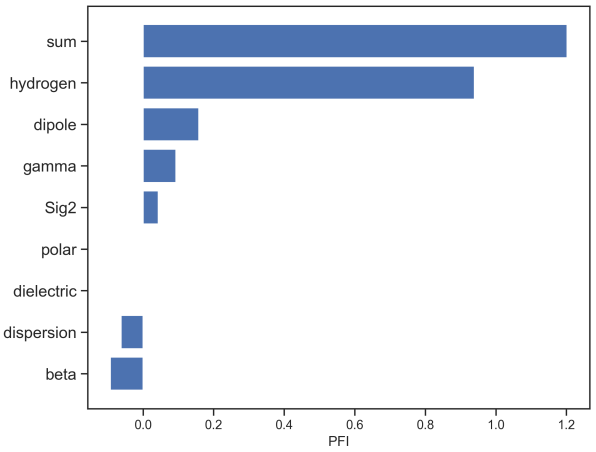
Permutation feature importances (PFIs) of NN_60_No_PFI



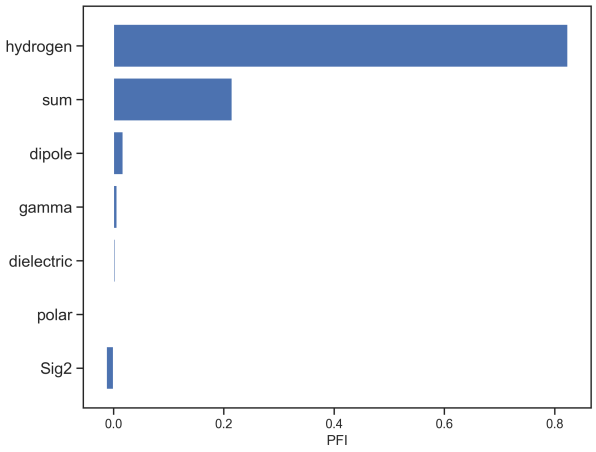
Permutation feature importances (PFIs) of NN_60_PFI



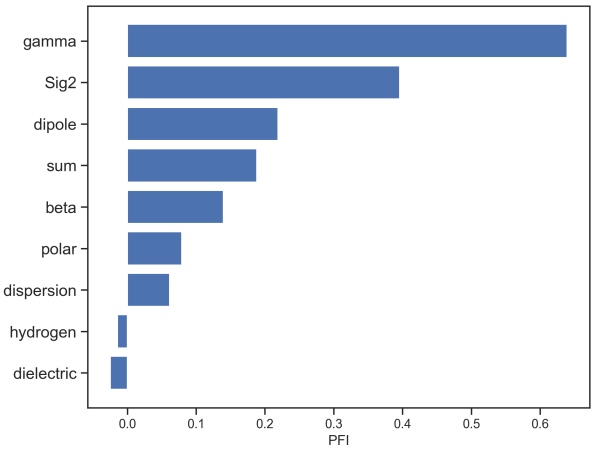
Permutation feature importances (PFIs) of NN_90_No_PFI



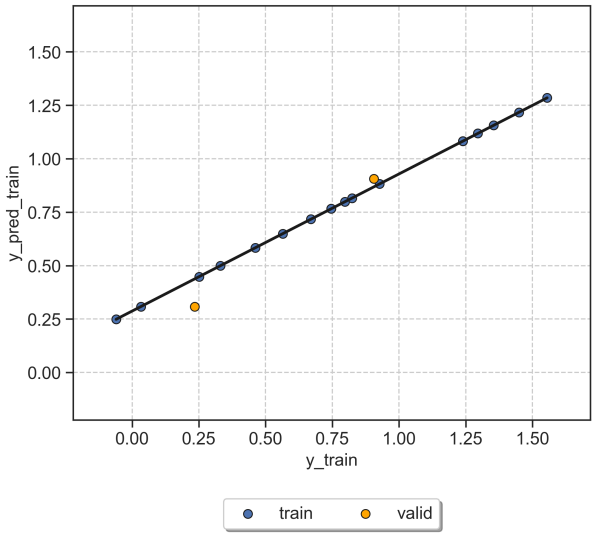
Permutation feature importances (PFIs) of NN_90_PFI



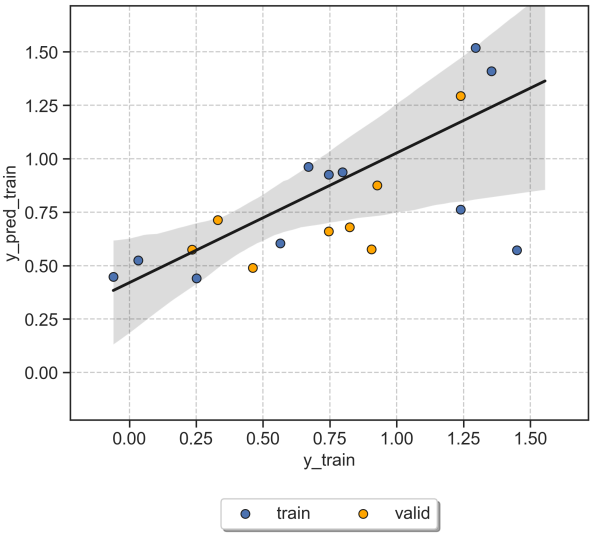
Permutation feature importances (PFIs) of VR_60_No_PFI



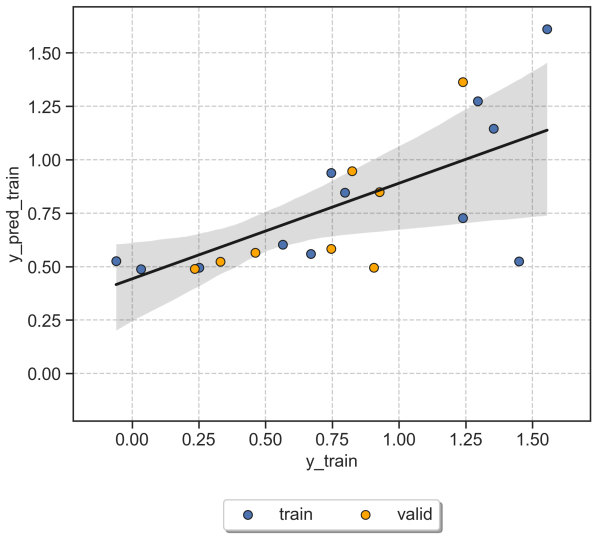
Predictions_train_valid of GB_90_PFI



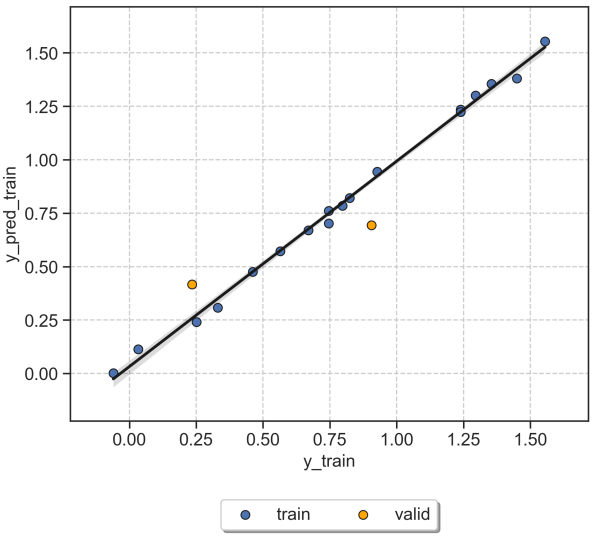
Predictions_train_valid of NN_60_No_PFI



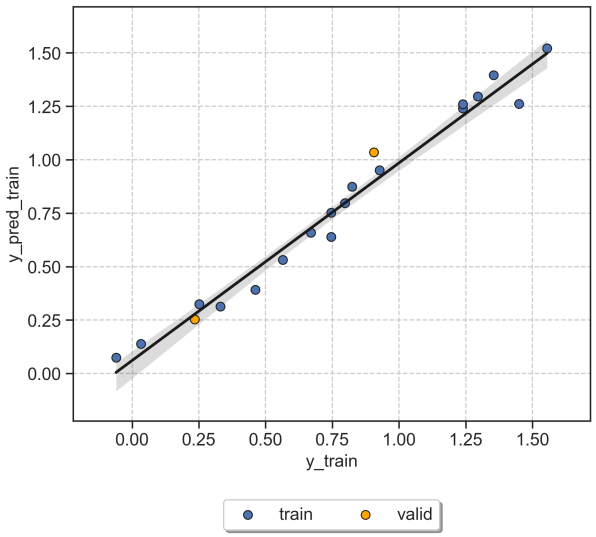
Predictions_train_valid of NN_60_PFI



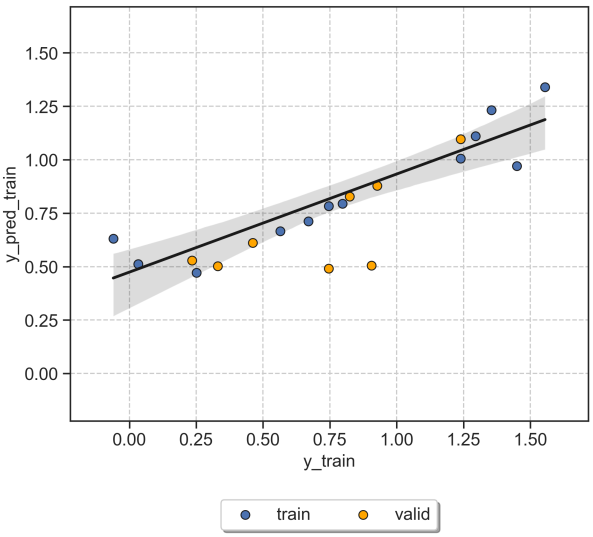
Predictions_train_valid of NN_90_No_PFI

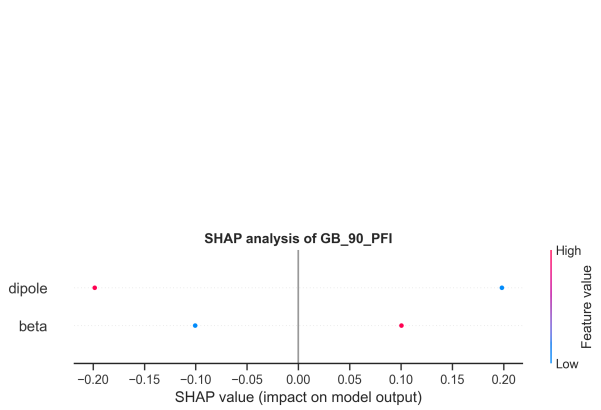


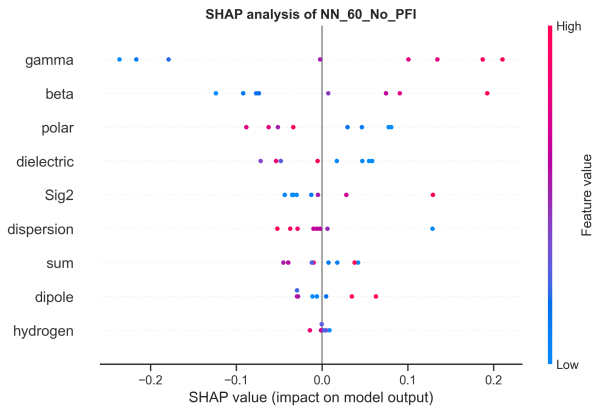
Predictions_train_valid of NN_90_PFI

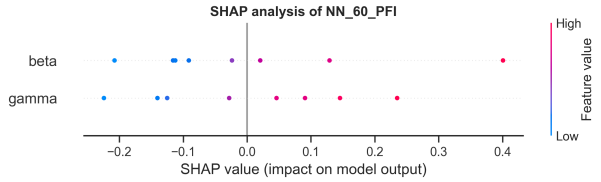


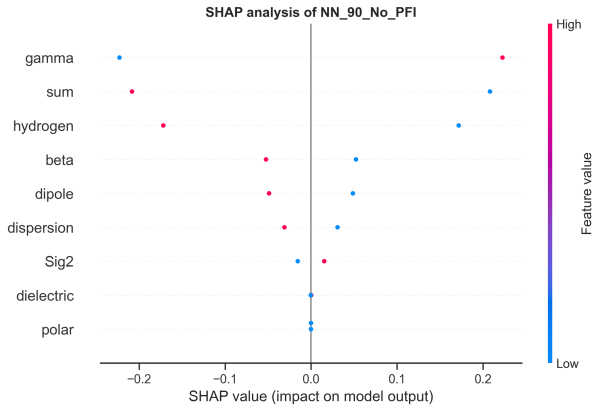
Predictions_train_valid of VR_60_No_PFI

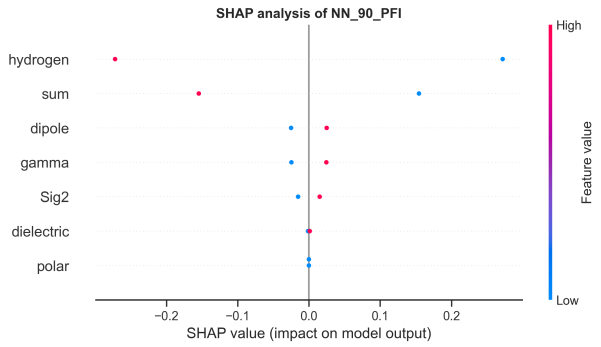


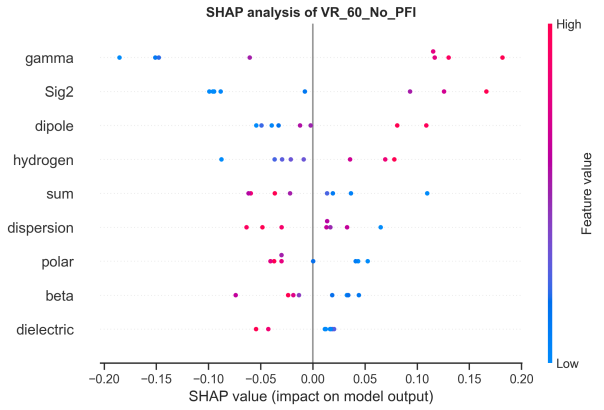












VERIFY

o Starting tests to verify the prediction ability of the ML models with the VERIFY module

o ML model NN_60.csv (with no PFI filter) and its corresponding Xy database were loaded successfully, including:
- Target value: dG
- Model: NN
- Descriptors: ['beta', 'gamma', 'dielectric', 'dispersion', 'polar', 'hydrogen', 'sum', 'dipole', 'Sig2']
- Training points: 12
- Validation points: 8

o ML model NN_60.csv (with no PFI filter) and its corresponding Xy database were loaded successfully, including:
- Target value: dG
- Model: NN
- Descriptors: ['beta', 'gamma', 'dielectric', 'dispersion', 'polar', 'hydrogen', 'sum', 'dipole', 'Sig2']
- Training points: 12

- Validation points: 8

o ML model NN_60.csv (with no PFI filter) and its corresponding Xy database were loaded successfull
y, including:
- Target value: dG
- Model: NN
- Descriptors: ['beta', 'gamma', 'dielectric', 'dispersion', 'polar', 'hydrogen', 'sum', 'dipole'
, 'Sig2']
- Training points: 12
- Validation points: 8

o ML model NN_60.csv (with no PFI filter) and its corresponding Xy database were loaded successfull
y, including:
- Target value: dG
- Model: NN
- Descriptors: ['beta', 'gamma', 'dielectric', 'dispersion', 'polar', 'hydrogen', 'sum', 'dipole'
, 'Sig2']
- Training points: 12
- Validation points: 8

o VERIFY donut plots saved in C:\Users\David\Desktop\Art�culo ROBERT\Pruebas_ROBERT\Regression\B
andar\VERIFY/VERIFY_tests_NN_60_No_PFI.png
o VERIFY test values saved in C:\Users\David\Desktop\Art�culo ROBERT\Pruebas_ROBERT\Regression\B
andar\VERIFY/VERIFY_tests_NN_60_No_PFI.dat:
Results of the VERIFY tests:
Original score (train set for CV): RMSE = 0.4, with a +- threshold (thres_test option) of 20.0
%:
- 5-fold CV: NOT DETERMINED, data splitting was done with KN. CV result: RMSE = 0.61
Original score (validation set): RMSE = 0.23, with a +- threshold (thres_test option) of 20.0%
:
o X_shuffle: PASSED, RMSE = 0.51 is higher than the threshold (0.27)
o y_shuffle: PASSED, RMSE = 0.43 is higher than the threshold (0.27)
o onehot: PASSED, RMSE = 0.46 is higher than the threshold (0.27)

o ML model NN_60_PFI.csv (with PFI filter) and its corresponding Xy database were loaded successful
ly, including:
- Target value: dG
- Model: NN
- Descriptors: ['beta', 'gamma']
- Training points: 12
- Validation points: 8

o ML model NN_60_PFI.csv (with PFI filter) and its corresponding Xy database were loaded successful
ly, including:
- Target value: dG
- Model: NN
- Descriptors: ['beta', 'gamma']
- Training points: 12
- Validation points: 8

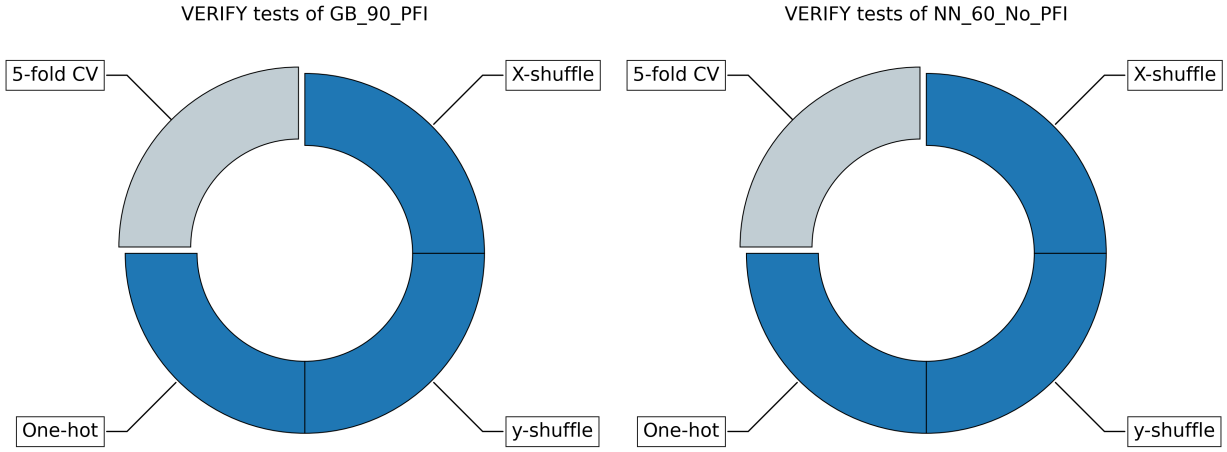
o ML model NN_60_PFI.csv (with PFI filter) and its corresponding Xy database were loaded successful
ly, including:
- Target value: dG
- Model: NN
- Descriptors: ['beta', 'gamma']
- Training points: 12
- Validation points: 8

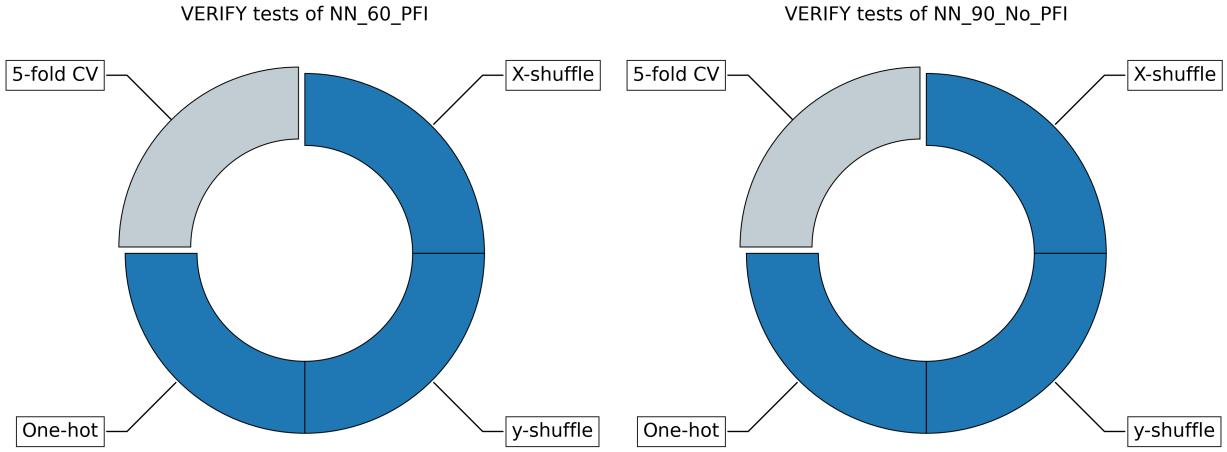
o ML model NN_60_PFI.csv (with PFI filter) and its corresponding Xy database were loaded successful
ly, including:
- Target value: dG
- Model: NN
- Descriptors: ['beta', 'gamma']
- Training points: 12
- Validation points: 8

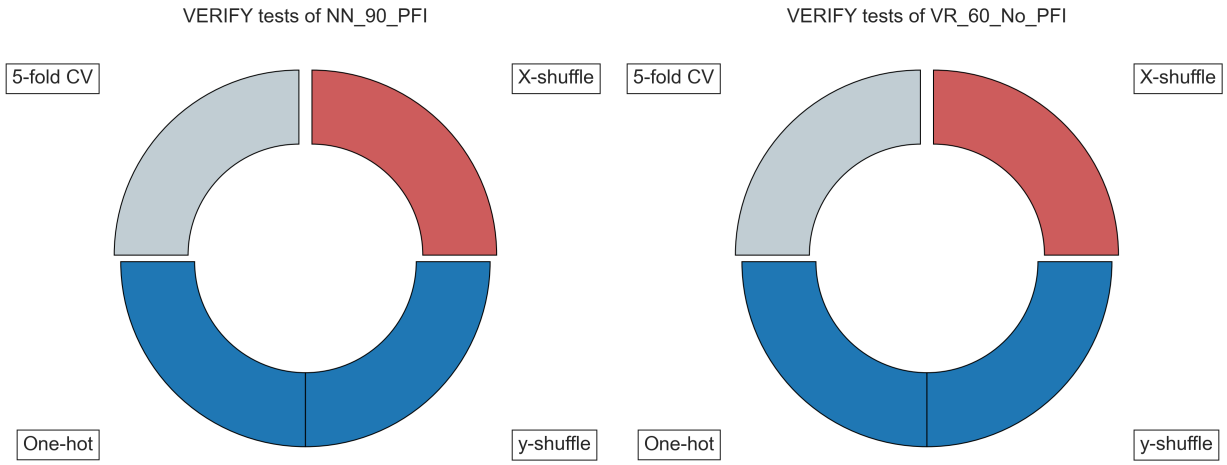
o VERIFY donut plots saved in C:\Users\David\Desktop\Art�culo ROBERT\Pruebas_ROBERT\Regression\B
andar\VERIFY/VERIFY_tests_NN_60_PFI.png
o VERIFY test values saved in C:\Users\David\Desktop\Art�culo ROBERT\Pruebas_ROBERT\Regression\B
andar\VERIFY/VERIFY_tests_NN_60_PFI.dat:
Results of the VERIFY tests:
Original score (train set for CV): RMSE = 0.39, with a +- threshold (thres_test option) of 20.

0%:
- 5-fold CV: NOT DETERMINED, data splitting was done with KN. CV result: RMSE = 0.5
Original score (validation set): RMSE = 0.21, with a +- threshold (thres_test option) of 20.0%
:
- o X_shuffle: PASSED, RMSE = 0.39 is higher than the threshold (0.25)
- o y_shuffle: PASSED, RMSE = 0.31 is higher than the threshold (0.25)
- o onehot: PASSED, RMSE = 0.36 is higher than the threshold (0.25)

Time VERIFY: 1.56 seconds







AQME-ROBERT