ROBERT v 1.0.2 2023/07/15 12:11:59
Citation: ROBERT v 1.0.2, Dalmau, D.; Alegre-Requena, J. V., 2023. https://github.com/jvalegre/robert

Command line used in ROBERT: robert --aqme --y solubility --csv_name solubility_short.csv

Total execution time: 1196.3 seconds

AQME

o Starting the generation of AQME descriptors with the AQME module

Time AQME: 243.72 seconds

CURATE

o Starting data curation with the CURATE module

o Database AQME-ROBERT_solubility_short.csv loaded successfully, including:
- 200 datapoints
- 221 accepted descriptors
- 2 ignored descriptors
- 0 discarded descriptors

o Analyzing categorical variables
- No categorical variables were found

o Duplication filters activated
Excluded datapoints:
- No datapoints were removed

o Correlation filter activated with these thresholds: thres_x = 0.9, thres_y = 0.001
Excluded descriptors:
- total energy: R**2 = 1.0 with electronic energy
- HOMO-LUMO gap/eV: R**2 = 0.94 with LUMO
- electronic energy: R**2 = 0.98 with NumValenceElectrons
- Total charge: error in R**2 with the solubility values (are all the values the same?)
- HOMO: R**2 = 0.0 with the solubility values
- Fermi-level/eV: R**2 = 0.94 with LUMO
- Total dispersion C6: R**2 = 0.96 with Total polarizability alpha
- Chi0v: R**2 = 0.92 with Total dispersion C8
- LabuteASA: R**2 = 0.96 with Total polarizability alpha
- Total polarizability alpha: R**2 = 1.0 with MolMR
- MaxEStateIndex: R**2 = 1.0 with MaxAbsEStateIndex
- HeavyAtomMolWt: R**2 = 1.0 with MolWt
- MolWt: R**2 = 1.0 with ExactMolWt
- HeavyAtomCount: R**2 = 0.95 with NumValenceElectrons
- NumRadicalElectrons: error in R**2 with the solubility values (are all the values the same?)
- MaxAbsPartialCharge: R**2 = 0.94 with MinPartialCharge

- BCUT2D_CHGHI: R**2 = 0.93 with BCUT2D_LOGPHI
- Chi0n: R**2 = 0.97 with Chi0
- Chi0: R**2 = 0.96 with Chi1
- Chi1n: R**2 = 0.93 with Chi1
- Chi3n: R**2 = 0.96 with Chi2n
- PEOE_VSA8: R**2 = 0.0 with the solubility values
- SMR_VSA10: R**2 = 0.97 with SlogP_VSA12
- SMR_VSA4: R**2 = 0.0 with the solubility values
- SlogP_VSA6: R**2 = 0.91 with SMR_VSA7
- SMR_VSA8: error in R**2 with the solubility values (are all the values the same?)
- SlogP_VSA10: R**2 = 0.0 with the solubility values
- SlogP_VSA11: R**2 = 0.0 with the solubility values
- SlogP_VSA9: error in R**2 with the solubility values (are all the values the same?)
- EState_VSA11: error in R**2 with the solubility values (are all the values the same?)
- EState_VSA4: R**2 = 0.0 with the solubility values
- EState_VSA7: R**2 = 0.0 with the solubility values
- VSA_EState8: R**2 = 0.0 with the solubility values
- NumSaturatedHeterocycles: R**2 = 1.0 with NumAliphaticHeterocycles
- fr_benzene: R**2 = 1.0 with NumAromaticCarbocycles
- NumHeteroatoms: R**2 = 0.0 with the solubility values
- NumSaturatedRings: R**2 = 0.0 with the solubility values
- fr_Al_COO: error in R**2 with the solubility values (are all the values the same?)
- fr_Al_OH_noTert: error in R**2 with the solubility values (are all the values the same?)
- fr_Ar_COO: error in R**2 with the solubility values (are all the values the same?)
- fr_Ar_NH: error in R**2 with the solubility values (are all the values the same?)
- fr_COO: error in R**2 with the solubility values (are all the values the same?)
- fr_COO2: error in R**2 with the solubility values (are all the values the same?)
- fr_C_O_noCOO: R**2 = 1.0 with fr_C_O
- fr_HOCCN: error in R**2 with the solubility values (are all the values the same?)
- fr_guanido: R**2 = 1.0 with fr_Imine
- fr_N_O: error in R**2 with the solubility values (are all the values the same?)
- fr_Ndealkylation1: error in R**2 with the solubility values (are all the values the same?)
- fr_Ndealkylation2: error in R**2 with the solubility values (are all the values the same?)
- fr_Nhpyrrole: error in R**2 with the solubility values (are all the values the same?)
- fr_SH: R**2 = 0.0 with the solubility values
- fr_alkyl_carbamate: error in R**2 with the solubility values (are all the values the same?)
- fr_allylic_oxid: R**2 = 0.0 with the solubility values
- fr_amidine: error in R**2 with the solubility values (are all the values the same?)
- fr_aryl_methyl: error in R**2 with the solubility values (are all the values the same?)
- fr_azide: error in R**2 with the solubility values (are all the values the same?)
- fr_azo: error in R**2 with the solubility values (are all the values the same?)
- fr_barbitur: error in R**2 with the solubility values (are all the values the same?)
- fr_benzodiazepine: error in R**2 with the solubility values (are all the values the same?)
- fr_diazo: error in R**2 with the solubility values (are all the values the same?)
- fr_dihydropyridine: error in R**2 with the solubility values (are all the values the same?)
- fr_hdrzone: error in R**2 with the solubility values (are all the values the same?)
- fr_imidazole: error in R**2 with the solubility values (are all the values the same?)
- fr_imide: error in R**2 with the solubility values (are all the values the same?)
- fr_isocyan: error in R**2 with the solubility values (are all the values the same?)
- fr_isothiocyan: error in R**2 with the solubility values (are all the values the same?)
- fr_ketone_Topliss: R**2 = 1.0 with fr_ketone
- fr_lactam: error in R**2 with the solubility values (are all the values the same?)
- fr_lactone: error in R**2 with the solubility values (are all the values the same?)

- fr_morpholine: error in R**2 with the solubility values (are all the values the same?)
- fr_nitro_arom: error in R**2 with the solubility values (are all the values the same?)
- fr_nitro_arom_nonortho: error in R**2 with the solubility values (are all the values the same?)
- fr_nitroso: error in R**2 with the solubility values (are all the values the same?)
- fr_oxazole: error in R**2 with the solubility values (are all the values the same?)
- fr_oxime: error in R**2 with the solubility values (are all the values the same?)
- fr_phenol: R**2 = 0.0 with the solubility values
- fr_phenol_noOrthoHbond: R**2 = 0.0 with the solubility values
- fr_phos_acid: error in R**2 with the solubility values (are all the values the same?)
- fr_phos_ester: error in R**2 with the solubility values (are all the values the same?)
- fr_piperdine: error in R**2 with the solubility values (are all the values the same?)
- fr_piperzine: error in R**2 with the solubility values (are all the values the same?)
- fr_prisulfonamd: error in R**2 with the solubility values (are all the values the same?)
- fr_quatN: error in R**2 with the solubility values (are all the values the same?)
- fr_sulfide: R**2 = 0.0 with the solubility values
- fr_sulfonamd: error in R**2 with the solubility values (are all the values the same?)
- fr_sulfone: error in R**2 with the solubility values (are all the values the same?)
- fr_tetrazole: error in R**2 with the solubility values (are all the values the same?)
- fr_thiocyan: error in R**2 with the solubility values (are all the values the same?)
- fr_thiophene: R**2 = 0.0 with the solubility values
- fr_unbrch_alkane: error in R**2 with the solubility values (are all the values the same?)

o 134 columns remaining after applying duplicate and correlation filters:
- code_name
- smiles
- solubility
- Dipole module/D
- LUMO
- Total dispersion C8
- Total FOD
- MaxAbsEStateIndex
- MinAbsEStateIndex
- MinEStateIndex
- qed
- ExactMolWt
- NumValenceElectrons
- MaxPartialCharge
- MinPartialCharge
- MinAbsPartialCharge
- FpDensityMorgan1
- FpDensityMorgan2
- FpDensityMorgan3
- BCUT2D_MWHI
- BCUT2D_MWLOW
- BCUT2D_CHGLO
- BCUT2D_LOGPHI
- BCUT2D_LOGPLOW
- BCUT2D_MRHI
- BCUT2D_MRLOW
- AvgIpc
- BalabanJ
- BertzCT
- Chi1

- Chi1v
- Chi2n
- Chi2v
- Chi3v
- Chi4n
- Chi4v
- HallKierAlpha
- Ipc
- Kappa1
- Kappa2
- Kappa3
- PEOE_VSA1
- PEOE_VSA10
- PEOE_VSA11
- PEOE_VSA12
- PEOE_VSA13
- PEOE_VSA14
- PEOE_VSA2
- PEOE_VSA3
- PEOE_VSA4
- PEOE_VSA5
- PEOE_VSA6
- PEOE_VSA7
- PEOE_VSA9
- SMR_VSA1
- SMR_VSA2
- SMR_VSA3
- SMR_VSA5
- SMR_VSA6
- SMR_VSA7
- SMR_VSA9
- SlogP_VSA1
- SlogP_VSA12
- SlogP_VSA2
- SlogP_VSA3
- SlogP_VSA4
- SlogP_VSA5
- SlogP_VSA7
- SlogP_VSA8
- TPSA
- EState_VSA1
- EState_VSA10
- EState_VSA2
- EState_VSA3
- EState_VSA5
- EState_VSA6
- EState_VSA8
- EState_VSA9
- VSA_EState1
- VSA_EState10
- VSA_EState2
- VSA_EState3
- VSA_EState4

- VSA_EState5
- VSA_EState6
- VSA_EState7
- VSA_EState9
- FractionCSP3
- NHOHCount
- NOCount
- NumAliphaticCarbocycles
- NumAliphaticHeterocycles
- NumAliphaticRings
- NumAromaticCarbocycles
- NumAromaticHeterocycles
- NumAromaticRings
- NumHAcceptors
- NumHDonors
- NumRotatableBonds
- NumSaturatedCarbocycles
- RingCount
- MolLogP
- MolMR
- fr_Al_OH
- fr_ArN
- fr_Ar_N
- fr_Ar_OH
- fr_C_O
- fr_C_S
- fr_Imine
- fr_NH0
- fr_NH1
- fr_NH2
- fr_aldehyde
- fr_alkyl_halide
- fr_amide
- fr_aniline
- fr_bicyclic
- fr_epoxide
- fr_ester
- fr_ether
- fr_furan
- fr_halogen
- fr_hdrzine
- fr_ketone
- fr_methoxy
- fr_nitrile
- fr_nitro
- fr_para_hydroxylation
- fr_priamide
- fr_pyridine
- fr_term_acetylene
- fr_thiazole
- fr_urea

x The Pearson heatmap was not generated because the number of features and the y value (132) is higher than

30.

o The curated database was stored in CURATE/AQME-ROBERT_solubility_short_CURATE.csv.

Time CURATE: 1.84 seconds

------- Images generated by the CURATE module -------

GENERATE

o Starting generation of ML models with the GENERATE module

o Database AQME-ROBERT_solubility_short_CURATE.csv loaded successfully, including:
- 200 datapoints
- 131 accepted descriptors
- 2 ignored descriptors
- 0 discarded descriptors

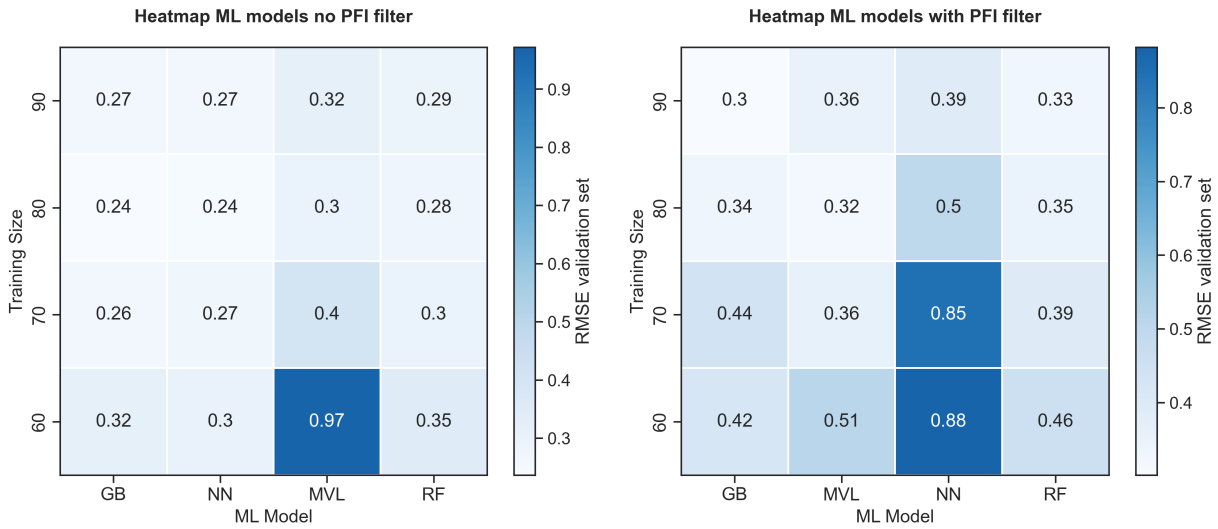
o Starting heatmap scan with 4 ML models (['RF', 'GB', 'NN', 'MVL']) and 4 training sizes ([60, 70, 80, 90]).
- 96 models were tested, for more information check the GENERATE_data.dat file in the GENERATE folder

o Heatmap ML models no PFI filter succesfully created in GENERATE/Raw_data

o Heatmap ML models with PFI filter succesfully created in GENERATE/Raw_data

Time GENERATE: 938.27 seconds

------- Images generated by the GENERATE module -------



VERIFY

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

------- Starting model with all variables (No PFI) -------

o ML model GB_80 (with no PFI filter) and Xy database were loaded, including:
- Target value: solubility
- Model: GB
- Descriptors: ['Dipole module/D', 'LUMO', 'Total dispersion C8', 'Total FOD', 'MaxAbsEStateIndex', 'MinAbsEStateIndex', 'MinEStateIndex', 'qed', 'ExactMolWt', 'NumValenceElectrons', 'MaxPartialCharge', 'MinPartialCharge', 'MinAbsPartialCharge', 'FpDensityMorgan1', 'FpDensityMorgan2', 'FpDensityMorgan3', 'BCUT2D_MWHI', 'BCUT2D_MWLOW', 'BCUT2D_CHGLO', 'BCUT2D_LOGPHI', 'BCUT2D_LOGPLOW', 'BCUT2D_MRHI', 'BCUT2D_MRLOW', 'AvgIpc', 'BalabanJ', 'BertzCT', 'Chi1', 'Chi1v', 'Chi2n', 'Chi2v', 'Chi3v', 'Chi4n', 'Chi4v', 'HallKierAlpha', 'Ipc', 'Kappa1', 'Kappa2', 'Kappa3', 'PEOE_VSA1', 'PEOE_VSA10', 'PEOE_VSA11', 'PEOE_VSA12', 'PEOE_VSA13', 'PEOE_VSA14', 'PEOE_VSA2', 'PEOE_VSA3', 'PEOE_VSA4', 'PEOE_VSA5', 'PEOE_VSA6', 'PEOE_VSA7', 'PEOE_VSA9', 'SMR_VSA1', 'SMR_VSA2', 'SMR_VSA3', 'SMR_VSA5', 'SMR_VSA6', 'SMR_VSA7', 'SMR_VSA9', 'SlogP_VSA1', 'SlogP_VSA12', 'SlogP_VSA2', 'SlogP_VSA3', 'SlogP_VSA4', 'SlogP_VSA5', 'SlogP_VSA7', 'SlogP_VSA8', 'TPSA', 'EState_VSA1', 'EState_VSA10', 'EState_VSA2', 'EState_VSA3', 'EState_VSA5', 'EState_VSA6', 'EState_VSA8', 'EState_VSA9', 'VSA_EState1', 'VSA_EState10', 'VSA_EState2', 'VSA_EState3', 'VSA_EState4', 'VSA_EState5', 'VSA_EState6', 'VSA_EState7', 'VSA_EState9', 'FractionCSP3', 'NHOHCount', 'NOCount', 'NumAliphaticCarbocycles', 'NumAliphaticHeterocycles', 'NumAliphaticRings', 'NumAromaticCarbocycles', 'NumAromaticHeterocycles', 'NumAromaticRings', 'NumHAcceptors', 'NumHDonors', 'NumRotatableBonds', 'NumSaturatedCarbocycles', 'RingCount', 'MolLogP', 'MolMR', 'fr_Al_OH', 'fr_ArN', 'fr_Ar_N', 'fr_Ar_OH', 'fr_C_O', 'fr_C_S', 'fr_Imine', 'fr_NH0', 'fr_NH1', 'fr_NH2', 'fr_aldehyde', 'fr_alkyl_halide', 'fr_amide', 'fr_aniline', 'fr_bicyclic', 'fr_epoxide', 'fr_ester', 'fr_ether', 'fr_furan', 'fr_halogen', 'fr_hdrzine', 'fr_ketone', 'fr_methoxy', 'fr_nitrile', 'fr_nitro', 'fr_para_hydroxylation', 'fr_priamide', 'fr_pyridine', 'fr_term_acetylene', 'fr_thiazole', 'fr_urea']
- Training points: 160
- Validation points: 40

o VERIFY donut plots saved in VERIFY/VERIFY_tests_GB_80_No_PFI.png
o VERIFY test values saved in VERIFY/VERIFY_tests_GB_80_No_PFI.dat
Results of the VERIFY tests:
Original score (train set in CV): RMSE = 0.028, +- 20% threshold (thres_test option):
- 5-fold CV: NOT DETERMINED, data splitting was done with KN. CV result: RMSE = 0.54
Original score (validation set): RMSE = 0.24, +- 20% threshold (thres_test option):
o y_mean: PASSED, RMSE = 1.2 is higher than the threshold (0.28)
o y_shuffle: PASSED, RMSE = 2.0 is higher than the threshold (0.28)
o onehot: PASSED, RMSE = 0.43 is higher than the threshold (0.28)

------- Starting model with PFI filter (only important descriptors used) -------

o ML model GB_90_PFI (with PFI filter) and Xy database were loaded, including:
- Target value: solubility
- Model: GB
- Descriptors: ['Total dispersion C8', 'MolLogP', 'MolMR']
- Training points: 180
- Validation points: 20

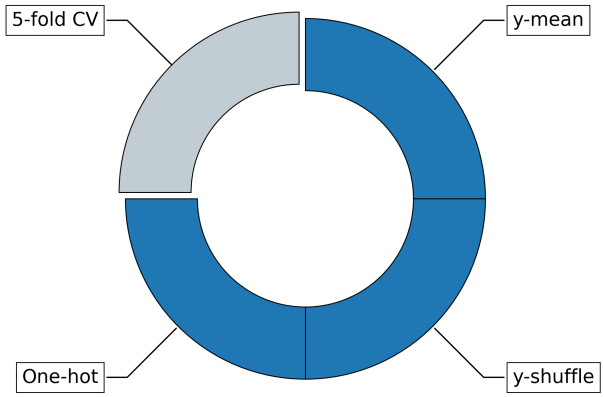
o VERIFY donut plots saved in VERIFY/VERIFY_tests_GB_90_PFI.png
o VERIFY test values saved in VERIFY/VERIFY_tests_GB_90_PFI.dat
Results of the VERIFY tests:
Original score (train set in CV): RMSE = 0.0027, +- 20% threshold (thres_test option):
- 5-fold CV: NOT DETERMINED, data splitting was done with KN. CV result: RMSE = 0.64
Original score (validation set): RMSE = 0.3, +- 20% threshold (thres_test option):
o y_mean: PASSED, RMSE = 0.88 is higher than the threshold (0.36)

o y_shuffle: PASSED, RMSE = 0.97 is higher than the threshold (0.36)
o onehot: PASSED, RMSE = 0.82 is higher than the threshold (0.36)

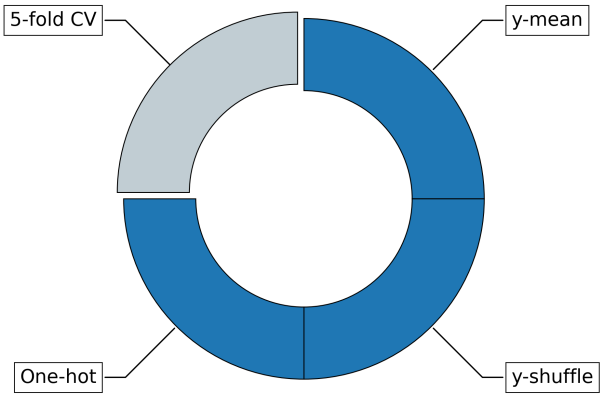
Time VERIFY: 2.42 seconds

------- Images generated by the VERIFY module -------

VERIFY tests of GB_80_No_PFI



VERIFY tests of GB_90_PFI



PREDICT

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

------- Starting model with all variables (No PFI) -------

o ML model GB_80 (with no PFI filter) and Xy database were loaded, including:
- Target value: solubility
- Model: GB
- Descriptors: ['Dipole module/D', 'LUMO', 'Total dispersion C8', 'Total FOD', 'MaxAbsEStateIndex', 'MinAbsEStateIndex', 'MinEStateIndex', 'qed', 'ExactMolWt', 'NumValenceElectrons', 'MaxPartialCharge', 'MinPartialCharge', 'MinAbsPartialCharge', 'FpDensityMorgan1', 'FpDensityMorgan2', 'FpDensityMorgan3', 'BCUT2D_MWHI', 'BCUT2D_MWLOW', 'BCUT2D_CHGLO', 'BCUT2D_LOGPHI', 'BCUT2D_LOGPLOW', 'BCUT2D_MRHI', 'BCUT2D_MRLOW', 'AvgIpc', 'BalabanJ', 'BertzCT', 'Chi1', 'Chi1v', 'Chi2n', 'Chi2v', 'Chi3v', 'Chi4n', 'Chi4v', 'HallKierAlpha', 'Ipc', 'Kappa1', 'Kappa2', 'Kappa3', 'PEOE_VSA1', 'PEOE_VSA10', 'PEOE_VSA11', 'PEOE_VSA12', 'PEOE_VSA13', 'PEOE_VSA14', 'PEOE_VSA2', 'PEOE_VSA3', 'PEOE_VSA4', 'PEOE_VSA5', 'PEOE_VSA6', 'PEOE_VSA7', 'PEOE_VSA9', 'SMR_VSA1', 'SMR_VSA2', 'SMR_VSA3', 'SMR_VSA5', 'SMR_VSA6', 'SMR_VSA7', 'SMR_VSA9', 'SlogP_VSA1', 'SlogP_VSA12', 'SlogP_VSA2', 'SlogP_VSA3', 'SlogP_VSA4', 'SlogP_VSA5', 'SlogP_VSA7', 'SlogP_VSA8', 'TPSA', 'EState_VSA1', 'EState_VSA10', 'EState_VSA2', 'EState_VSA3', 'EState_VSA5', 'EState_VSA6', 'EState_VSA8', 'EState_VSA9', 'VSA_EState1', 'VSA_EState10', 'VSA_EState2', 'VSA_EState3', 'VSA_EState4', 'VSA_EState5', 'VSA_EState6', 'VSA_EState7', 'VSA_EState9', 'FractionCSP3', 'NHOHCount', 'NOCount', 'NumAliphaticCarbocycles', 'NumAliphaticHeterocycles', 'NumAliphaticRings', 'NumAromaticCarbocycles', 'NumAromaticHeterocycles', 'NumAromaticRings', 'NumHAcceptors', 'NumHDonors', 'NumRotatableBonds', 'NumSaturatedCarbocycles', 'RingCount', 'MolLogP', 'MolMR', 'fr_Al_OH', 'fr_ArN', 'fr_Ar_N', 'fr_Ar_OH', 'fr_C_O', 'fr_C_S', 'fr_Imine', 'fr_NH0', 'fr_NH1', 'fr_NH2', 'fr_aldehyde', 'fr_alkyl_halide', 'fr_amide', 'fr_aniline', 'fr_bicyclic', 'fr_epoxide', 'fr_ester', 'fr_ether', 'fr_furan', 'fr_halogen', 'fr_hdrzine', 'fr_ketone', 'fr_methoxy', 'fr_nitrile', 'fr_nitro', 'fr_para_hydroxylation', 'fr_priamide', 'fr_pyridine', 'fr_term_acetylene', 'fr_thiazole', 'fr_urea']
- Training points: 160

- Validation points: 40
  - Train set with predicted results: GB_80_train_No_PFI.csv
  - Validation set with predicted results: GB_80_valid_No_PFI.csv

o Saving graphs and CSV databases in:
  - Graph in: PREDICT/Results_GB_80_No_PFI.png

o Results saved in PREDICT/Results_GB_80_No_PFI.dat:
  - Points Train:Validation = 160:40
  - Proportion Train:Validation = 80:20
  - Number of descriptors = 131
  - Proportion points:descriptors = 200:131
  - Train : R2 = 1.0, MAE = 0.023, RMSE = 0.028
  - Validation : R2 = 0.96, MAE = 0.16, RMSE = 0.24

o SHAP plot saved in PREDICT/SHAP_GB_80_No_PFI.png
o SHAP values saved in PREDICT/SHAP_GB_80_No_PFI.dat:
  - MolLogP = min: -0.45, max: 0.71
  - Total dispersion C8 = min: -0.4, max: 0.3
  - MinPartialCharge = min: -0.21, max: 0.27
  - NumHAcceptors = min: -0.079, max: 0.23
  - TPSA = min: -0.053, max: 0.2
  - NOCount = min: -0.055, max: 0.16
  - Chi1v = min: -0.21, max: 0.11
  - MolMR = min: -0.26, max: 0.1
  - PEOE_VSA1 = min: -0.014, max: 0.1
  - Chi1 = min: -0.034, max: 0.088
  - SlogP_VSA2 = min: -0.067, max: 0.085
  - Chi2n = min: -0.05, max: 0.081
  - PEOE_VSA6 = min: -0.058, max: 0.08
  - VSA_EState1 = min: -0.1, max: 0.065
  - ExactMolWt = min: -0.024, max: 0.062
  - Chi2v = min: -0.048, max: 0.05
  - MinEStateIndex = min: -0.079, max: 0.05
  - Dipole module/D = min: -0.083, max: 0.047
  - Chi3v = min: -0.052, max: 0.047
  - MaxPartialCharge = min: -0.028, max: 0.046
  - SlogP_VSA5 = min: -0.068, max: 0.043
  - MaxAbsEStateIndex = min: -0.027, max: 0.042
  - NumValenceElectrons = min: -0.057, max: 0.04
  - BertzCT = min: -0.018, max: 0.039
  - BalabanJ = min: -0.052, max: 0.036
  - MinAbsPartialCharge = min: -0.025, max: 0.035
  - BCUT2D_MRLOW = min: -0.029, max: 0.033
  - Kappa1 = min: -0.022, max: 0.033
  - AvgIpc = min: -0.021, max: 0.032
  - SMR_VSA1 = min: -0.038, max: 0.032
  - EState_VSA1 = min: -0.075, max: 0.031
  - Kappa3 = min: -0.031, max: 0.03
  - Kappa2 = min: -0.046, max: 0.03
  - PEOE_VSA11 = min: -0.0026, max: 0.027
  - FpDensityMorgan1 = min: -0.075, max: 0.026
  - Ipc = min: -0.012, max: 0.025

- MinAbsEStateIndex = min: -0.013, max: 0.025
- LUMO = min: -0.034, max: 0.021
- EState_VSA9 = min: -0.01, max: 0.02
- BCUT2D_LOGPLOW = min: -0.014, max: 0.02
- BCUT2D_MWLOW = min: -0.014, max: 0.016
- SMR_VSA6 = min: -0.0042, max: 0.014
- BCUT2D_MWHI = min: -0.0088, max: 0.014
- PEOE_VSA7 = min: -0.026, max: 0.013
- SMR_VSA5 = min: -0.017, max: 0.013
- FpDensityMorgan3 = min: -0.021, max: 0.013
- FpDensityMorgan2 = min: -0.016, max: 0.013
- BCUT2D_LOGPHI = min: -0.0024, max: 0.012
- qed = min: -0.0064, max: 0.012
- HallKierAlpha = min: -0.028, max: 0.012
- Total FOD = min: -0.0068, max: 0.011
- SlogP_VSA4 = min: -0.025, max: 0.011
- PEOE_VSA4 = min: -0.0006, max: 0.01
- SMR_VSA7 = min: -0.013, max: 0.0098
- NHOHCount = min: -0.0013, max: 0.0098
- VSA_EState6 = min: -0.0082, max: 0.0096
- Chi4v = min: -0.025, max: 0.0088
- VSA_EState7 = min: -0.014, max: 0.0084
- Chi4n = min: -0.01, max: 0.0082
- VSA_EState3 = min: -0.0027, max: 0.0077
- BCUT2D_MRHI = min: -0.0087, max: 0.0071
- NumHDonors = min: -0.0013, max: 0.007
- VSA_EState2 = min: -0.00098, max: 0.0069
- BCUT2D_CHGLO = min: -0.021, max: 0.0069
- PEOE_VSA5 = min: -0.016, max: 0.0067
- FractionCSP3 = min: -0.0041, max: 0.0066
- SlogP_VSA3 = min: -0.0011, max: 0.0063
- PEOE_VSA13 = min: -0.00053, max: 0.0061
- EState_VSA8 = min: -0.0052, max: 0.006
- NumRotatableBonds = min: -0.00095, max: 0.006
- PEOE_VSA9 = min: -0.015, max: 0.0037
- VSA_EState10 = min: -0.0051, max: 0.0031
- SlogP_VSA12 = min: -0.0039, max: 0.003
- fr_Ar_OH = min: -0.00027, max: 0.003
- PEOE_VSA3 = min: -0.0017, max: 0.003
- PEOE_VSA10 = min: -0.0012, max: 0.0029
- EState_VSA2 = min: -0.0088, max: 0.0027
- VSA_EState5 = min: -0.0045, max: 0.0023
- VSA_EState4 = min: -0.0051, max: 0.0022
- SlogP_VSA7 = min: -0.0019, max: 0.0019
- fr_alkyl_halide = min: -0.00028, max: 0.0016
- fr_C_O = min: -0.00014, max: 0.001
- fr_halogen = min: -0.00099, max: 0.001
- NumSaturatedCarbocycles = min: -0.02, max: 0.00091
- fr_ether = min: -0.008, max: 0.00076
- PEOE_VSA14 = min: -5.1e-05, max: 0.00062
- EState_VSA10 = min: -0.00019, max: 0.00058
- PEOE_VSA2 = min: -0.00066, max: 0.00044
- RingCount = min: -0.00015, max: 0.00042

- PEOE_VSA12 = min: -0.0055, max: 0.00033
- VSA_EState9 = min: -0.0021, max: 0.0002
- fr_methoxy = min: -0.0019, max: 0.00017
- EState_VSA3 = min: -0.00016, max: 0.00016
- fr_nitro = min: -0.0022, max: 0.00011
- SMR_VSA9 = min: -0.00031, max: 6e-05
- fr_nitrile = min: -0.0013, max: 4.3e-05
- NumAromaticRings = min: -5.2e-05, max: 2.6e-05
- fr_aldehyde = min: -2.5e-05, max: 2.5e-05
- fr_NH0 = min: -1.1e-05, max: 1.6e-05
- fr_urea = min: 0.0, max: 0.0
- fr_thiazole = min: 0.0, max: 0.0
- fr_term_acetylene = min: 0.0, max: 0.0
- fr_pyridine = min: 0.0, max: 0.0
- fr_priamide = min: 0.0, max: 0.0
- fr_para_hydroxylation = min: 0.0, max: 0.0
- fr_ketone = min: 0.0, max: 0.0
- fr_hdrzine = min: 0.0, max: 0.0
- fr_furan = min: 0.0, max: 0.0
- fr_ester = min: 0.0, max: 0.0
- fr_epoxide = min: 0.0, max: 0.0
- fr_bicyclic = min: 0.0, max: 0.0
- fr_aniline = min: 0.0, max: 0.0
- fr_amide = min: 0.0, max: 0.0
- fr_NH2 = min: 0.0, max: 0.0
- fr_NH1 = min: 0.0, max: 0.0
- fr_Imine = min: 0.0, max: 0.0
- fr_C_S = min: 0.0, max: 0.0
- fr_Ar_N = min: 0.0, max: 0.0
- fr_ArN = min: 0.0, max: 0.0
- SlogP_VSA8 = min: 0.0, max: 0.0
- SlogP_VSA1 = min: 0.0, max: 0.0
- SMR_VSA3 = min: 0.0, max: 0.0
- NumAromaticHeterocycles = min: 0.0, max: 0.0
- NumAromaticCarbocycles = min: 0.0, max: 0.0
- NumAliphaticRings = min: 0.0, max: 0.0
- NumAliphaticHeterocycles = min: 0.0, max: 0.0
- NumAliphaticCarbocycles = min: 0.0, max: 0.0
- EState_VSA5 = min: 0.0, max: 0.0
- fr_Al_OH = min: -6.2e-06, max: 0.0
- SMR_VSA2 = min: -1.2e-05, max: 0.0
- EState_VSA6 = min: -0.001, max: 0.0

o PFI plot saved in PREDICT/PFI_GB_80_No_PFI.png
o PFI values saved in PREDICT/PFI_GB_80_No_PFI.dat:
Original score (from model.score, R2) = 0.96

- Total dispersion C8 = 0.086 +- 0.02
- Dipole module/D = 0.0044 +- 0.002
- MinEStateIndex = 0.0035 +- 0.0013
- ExactMolWt = 0.0025 +- 0.0025
- LUMO = 0.002 +- 0.00083
- NumValenceElectrons = 0.0012 +- 0.0013
- MaxAbsEStateIndex = 0.0003 +- 0.0012

- Total FOD = 0.00021 +- 0.0012
- MinAbsEStateIndex = 8.3e-05 +- 0.00043
- qed = -5.8e-05 +- 0.00022

o Outliers plot saved in PREDICT/Outliers_GB_80_No_PFI.png
o Outlier values saved in PREDICT/Outliers_GB_80_No_PFI.dat:
Train: 4 outliers out of 160 datapoints (2.5%)
- mol_1101 (2.1 SDs)
- mol_147 (2.9 SDs)
- mol_149 (3.2 SDs)
- mol_606 (3.8 SDs)
Validation: 1 outliers out of 40 datapoints (2.5%)
- mol_253 (4.4 SDs)

------- Starting model with PFI filter (only important descriptors used) -------

o ML model GB_90_PFI (with PFI filter) and Xy database were loaded, including:
- Target value: solubility
- Model: GB
- Descriptors: ['Total dispersion C8', 'MolLogP', 'MolMR']
- Training points: 180
- Validation points: 20
  - Train set with predicted results: GB_90_train_PFI.csv
  - Validation set with predicted results: GB_90_valid_PFI.csv

o Saving graphs and CSV databases in:
  - Graph in: PREDICT/Results_GB_90_PFI.png

o Results saved in PREDICT/Results_GB_90_PFI.dat:
  - Points Train:Validation = 180:20
  - Proportion Train:Validation = 90:10
  - Number of descriptors = 3
  - Proportion points:descriptors = 200:3
  - Train : R2 = 1.0, MAE = 0.0022, RMSE = 0.0027
  - Validation : R2 = 0.89, MAE = 0.19, RMSE = 0.3

o SHAP plot saved in PREDICT/SHAP_GB_90_PFI.png
o SHAP values saved in PREDICT/SHAP_GB_90_PFI.dat:
  - MolLogP = min: -0.65, max: 1.5
  - MolMR = min: -0.4, max: 0.63
  - Total dispersion C8 = min: -0.2, max: 0.47

o PFI plot saved in PREDICT/PFI_GB_90_PFI.png
o PFI values saved in PREDICT/PFI_GB_90_PFI.dat:
Original score (from model.score, R2) = 0.88
  - MolLogP = 0.58 +- 0.061
  - Total dispersion C8 = 0.24 +- 0.09
  - MolMR = 0.24 +- 0.054

o Outliers plot saved in PREDICT/Outliers_GB_90_PFI.png
o Outlier values saved in PREDICT/Outliers_GB_90_PFI.dat:
Train: 7 outliers out of 180 datapoints (3.9%)

- mol_122 (2.2 SDs)
- mol_123 (3.1 SDs)
- mol_147 (2.1 SDs)
- mol_269 (4.6 SDs)
- mol_532 (2.4 SDs)
- mol_606 (2.1 SDs)
- mol_907 (2.1 SDs)

Validation: 1 outliers out of 20 datapoints (5.0%)

- mol_253 (3.2 SDs)

Time PREDICT: 10.05 seconds

------- Images and summary generated by the PREDICT module -------

No PFI (all descriptors):

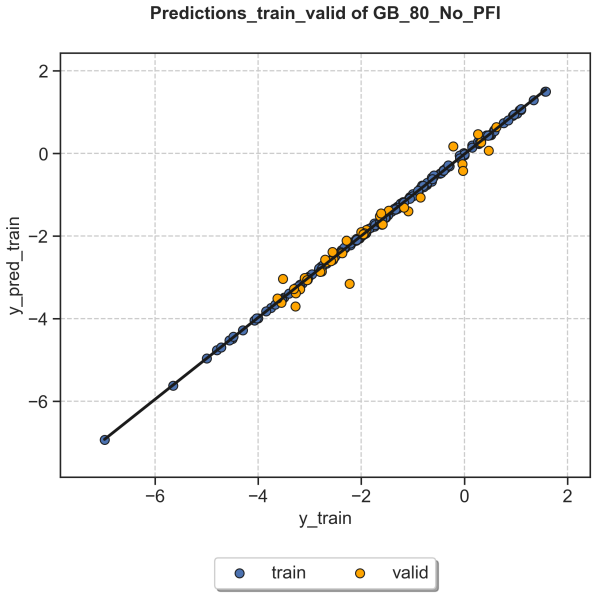
Results_GB_80_No_PFI.dat:

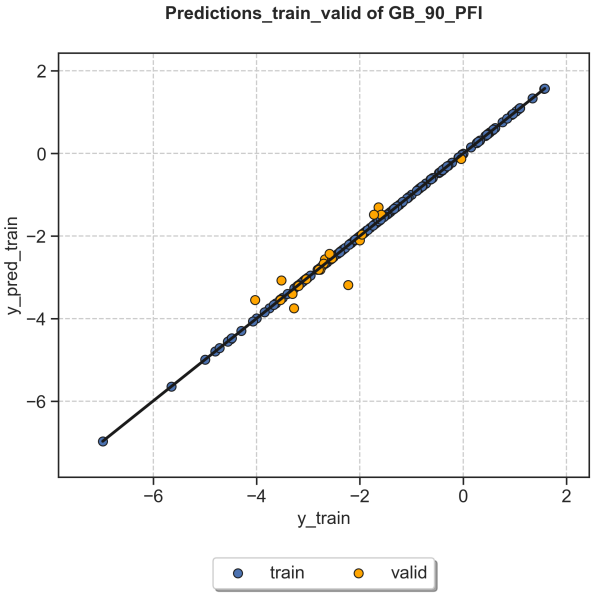
- Points Train:Validation = 160:40
- Proportion Train:Validation = 80:20
- Number of descriptors = 131
- Proportion points:descriptors = 200:131
- Train : R2 = 1.0, MAE = 0.023, RMSE = 0.028
- Validation : R2 = 0.96, MAE = 0.16, RMSE = 0.24

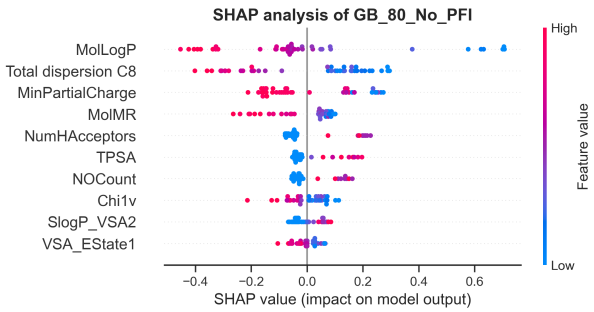
PFI (only important descriptors):

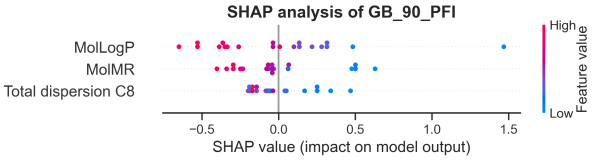
Results_GB_90_PFI.dat:

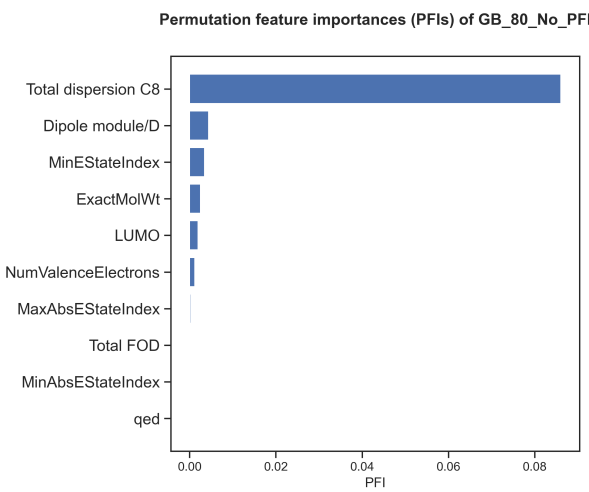
- Points Train:Validation = 180:20
- Proportion Train:Validation = 90:10
- Number of descriptors = 3
- Proportion points:descriptors = 200:3
- Train : R2 = 1.0, MAE = 0.0022, RMSE = 0.0027
- Validation : R2 = 0.89, MAE = 0.19, RMSE = 0.3

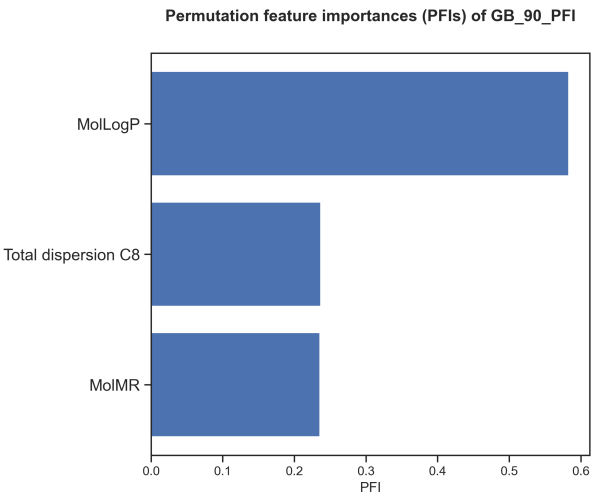


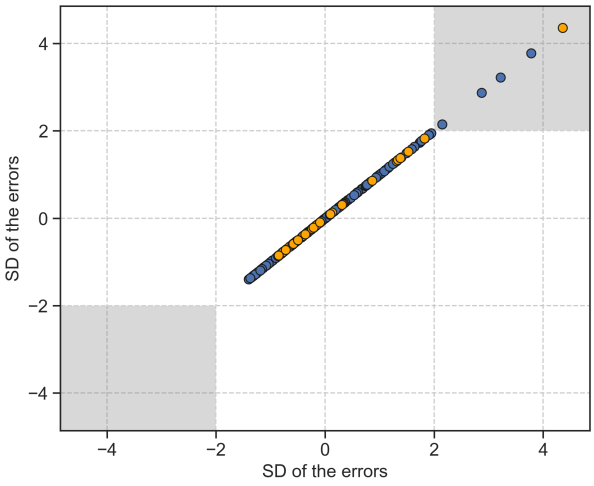


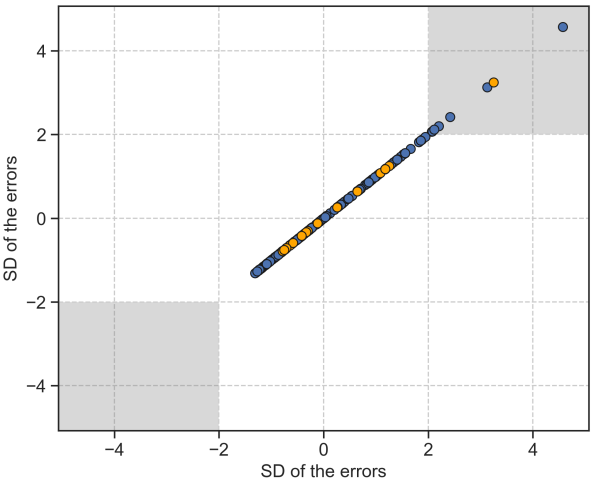





Permutation feature importances (PFIs) of GB_80_No_PFI
Total dispersion C8
Dipole module/D
MinEStateIndex
ExactMolWt
LUMO
NumValenceElectrons
MaxAbsEStateIndex
Total FOD
MinAbsEStateIndex
qed
0.00
0.02
0.04
0.06
0.08
PFI


Permutation feature importances (PFIs) of GB_90_PFI
MolLogP
Total dispersion C8
MolMR
0.0
0.1
0.2
0.3
0.4
0.5
0.6
PFI


Outlier analysis of GB_80_No_PFI
SD of the errors
4
2
0
−2
−4
SD of the errors


Outlier analysis of GB_90_PFI
SD of the errors
4
2
0
−2
−4
SD of the errors