Supplementary Materials

Application of Machine Learning for Predicting G9a Inhibitors

Mariya L. Ivanova , Nicola Russo, Nadia Djaid and Konstantin Nikolic
School of Computing and Engineering, University of West London, London W5 5RF, UK

Contents

Figures

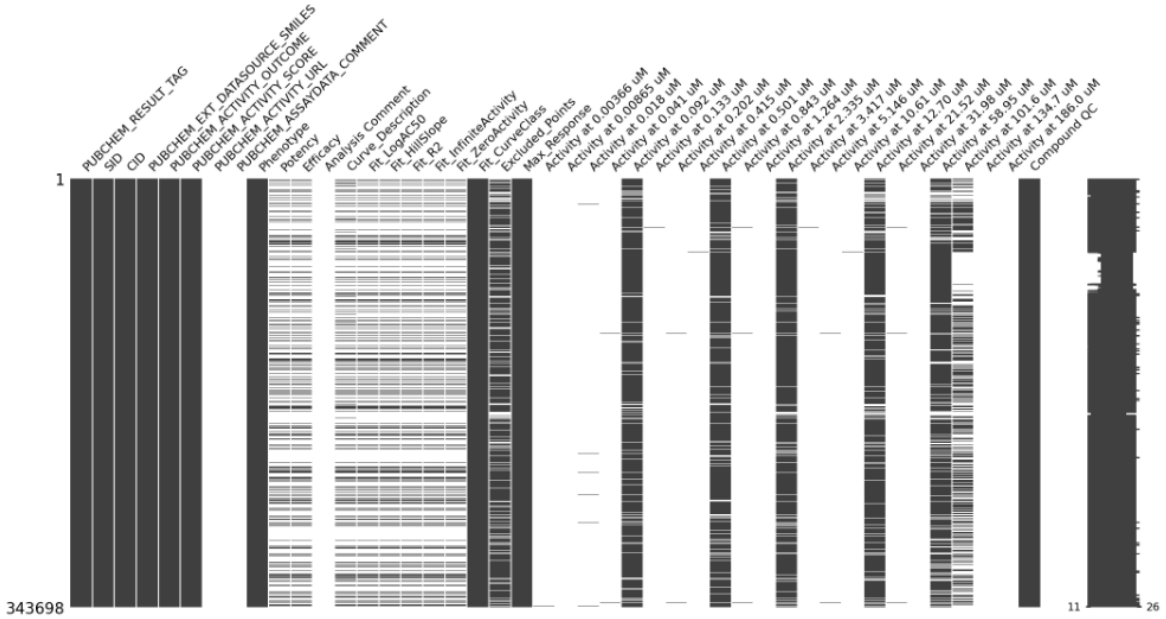

Figure 1. Visualisation of the AID 504332 PubChem BioAssay,

which describes activity of G9a enzyme in the presence of various compounds and substances. The assay results are represented as a matrix with 343,698 rows (where each row represents one compound or substance), and 56 columns (where each column represents a relevant attribute for each compound). Only compounds and substances IDs and the activity label are taken from this table. The IDs were then used for collecting information about the compounds from PubChem.

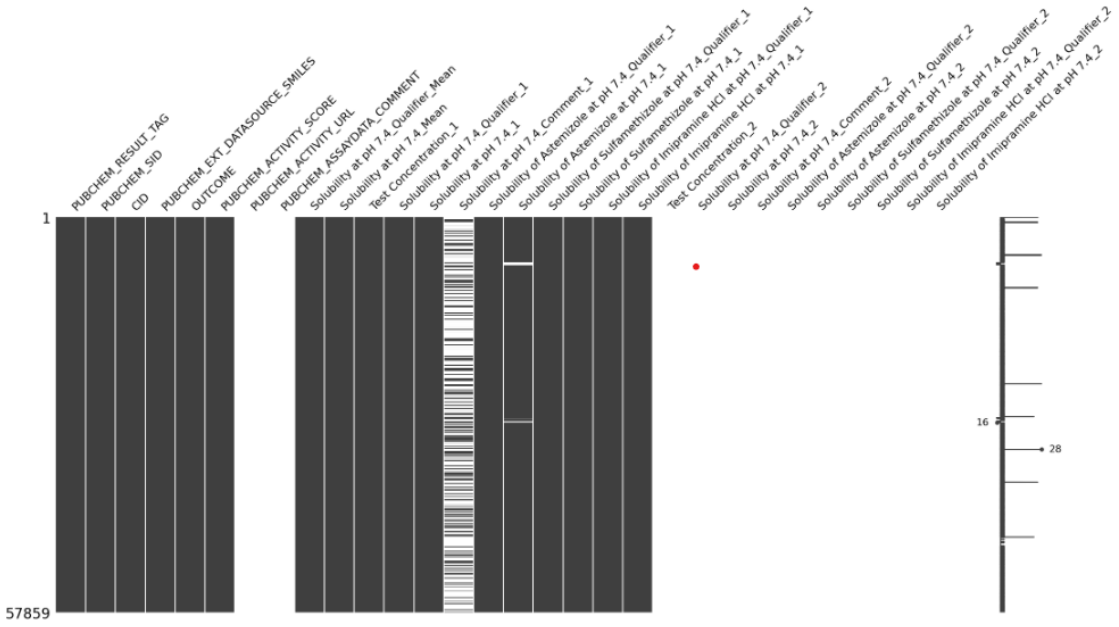

Figure 2. Visualisation of the AID 1996 PubChem BioAssay,

which describes solubility of different compounds. The matrix consists of 57,859 rows and 30 columns. Only compounds and substances IDs and Solubility (in water) at 7.4pH Mean are taken from this table.

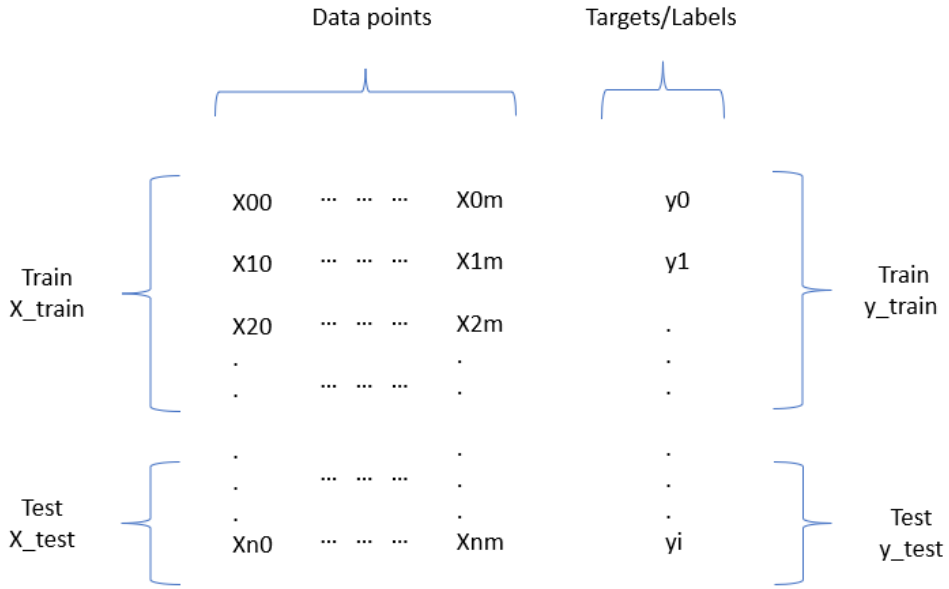

Figure 3. Mechanism of data partitioning for ML

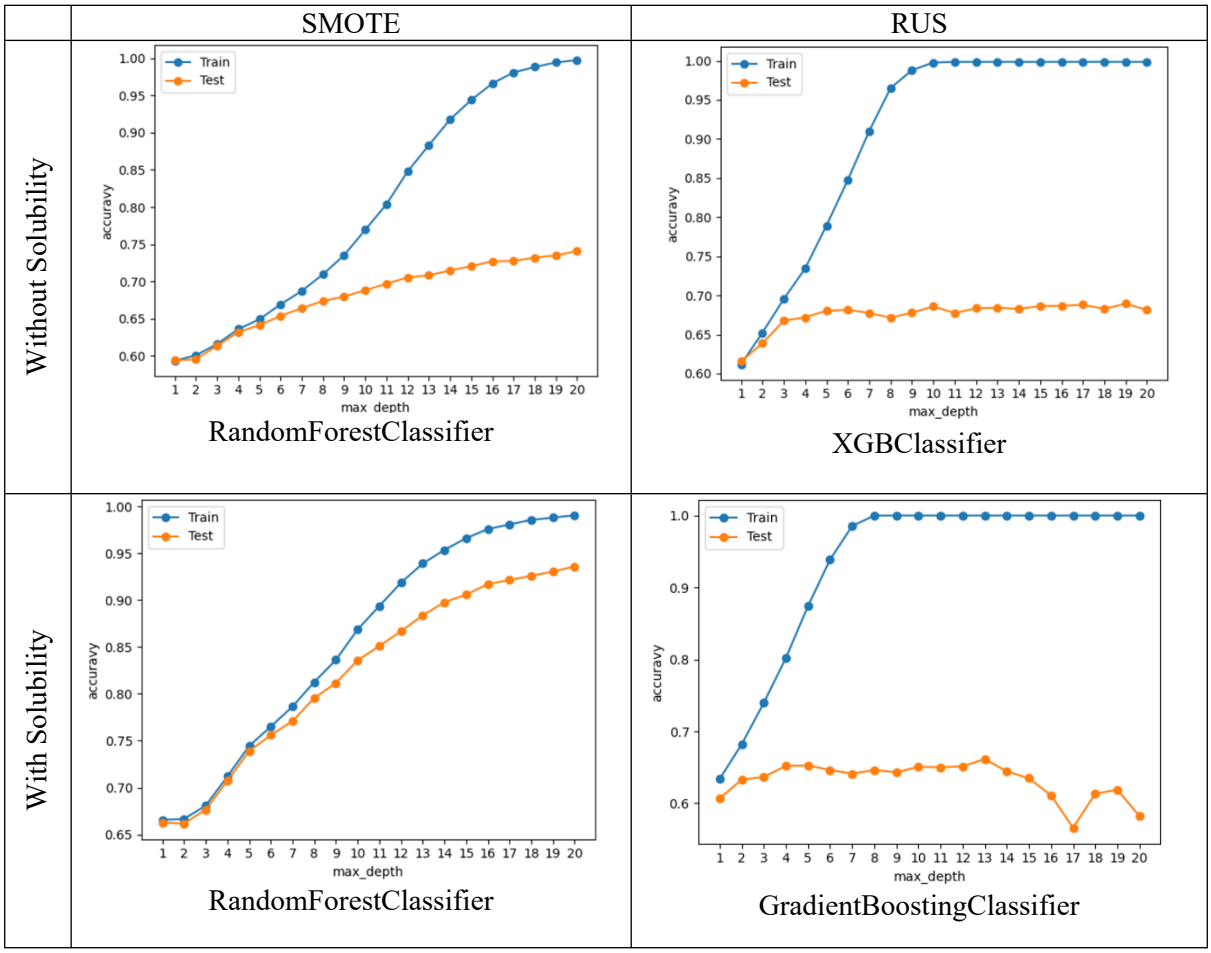

Figure 4. Comparison of the training and testing performance of the model that performed best in cross-validation, for each of the four datasets.

This analysis indicates when the overfitting starts, which is typically at the point just before the divergence between the two curves exceeds 5%.

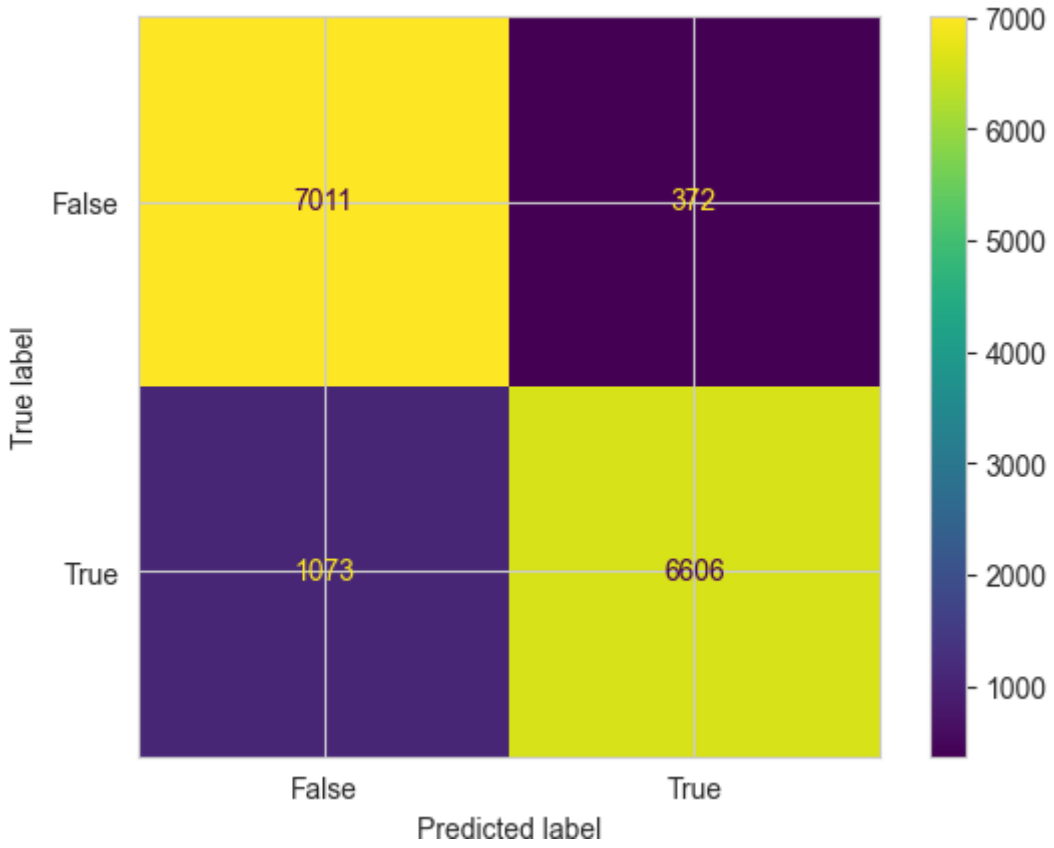

Figure 5. Confusion matrix of the five-feature ML with Random Forest Classifier

Tables

Table 1. Cross-validation (CV) scores for Accuracy, for four datasets and five classifiers.

Without Solubility data — RUS

Cross-validation results

	1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
4	XGBoost	0.6790	0.0044	[0.6816, 0.678, 0.6835, 0.6808, 0.6709]
2	RandomForest	0.6757	0.0031	[0.6778, 0.6738, 0.6802, 0.6755, 0.6714]
3	GradientBoost	0.6494	0.0068	[0.6515, 0.6471, 0.6616, 0.6422, 0.6445]
1	Decision	0.5844	0.0059	[0.588, 0.5748, 0.5924, 0.5819, 0.5847]
0	SVM	0.5756	0.0029	[0.5738, 0.5747, 0.5766, 0.5807, 0.5723]

Without Solubility data — SMOTE

	1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
2	RandomForest	0.7456	0.0016	[0.7469, 0.7449, 0.7469, 0.7465, 0.7428]
4	XGBoost	0.7406	0.0031	[0.7418, 0.7368, 0.7459, 0.7399, 0.7389]
3	GradientBoost	0.6862	0.0032	[0.6881, 0.6836, 0.6892, 0.689, 0.6812]
1	Decision	0.6360	0.0032	[0.6386, 0.6333, 0.6393, 0.6377, 0.6312]
0	SVM	0.5782	0.0030	[0.5841, 0.577, 0.577, 0.5771, 0.5758]

With Solubility data — RUS

	1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
3	GradientBoost	0.6475	0.0096	[0.6565, 0.6329, 0.6582, 0.6403, 0.6499]
4	XGBoost	0.6445	0.0043	[0.6444, 0.6456, 0.639, 0.6518, 0.6416]
2	RandomForest	0.6426	0.0109	[0.6418, 0.6444, 0.6518, 0.6524, 0.6224]
1	Decision	0.5618	0.0060	[0.5551, 0.5602, 0.5644, 0.5574, 0.5721]
0	SVM	0.5437	0.0065	[0.5341, 0.5398, 0.5536, 0.5446, 0.5466]

With Solubility data — SMOTE

	1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
2	RandomForest	0.9516	0.0024	[0.9507, 0.9487, 0.9545, 0.9499, 0.9543]
4	XGBoost	0.9458	0.0007	[0.9468, 0.9462, 0.9458, 0.9445, 0.946]
1	Decision	0.8712	0.0065	[0.8703, 0.8648, 0.8683, 0.8689, 0.8836]
3	GradientBoost	0.8680	0.0031	[0.8685, 0.8693, 0.8632, 0.8665, 0.8727]
0	SVM	0.6071	0.0046	[0.6064, 0.6071, 0.6107, 0.5989, 0.6122]

Table 2: Parameters used for Hyperparameter tuning of the RFC with Dataset 4 with five features.

Hyperparameters	**Values**	**Default**
'n_estimators'	50, 100, 200	100
criterion	'gini', 'entropy', 'log_loss'	‘gini’
'max_depth'	None, 2, 5, 10, 20	None
'min_samples_split'	2, 10, 20, 50	2
'min_samples_leaf'	1, 2, 10, 20, 50	1
'oob_score'	False, True	

Table 3. Classification report of Random Forest Classifier with the five feature Dataset 4.

	precision	recall	f1-score	support
Active (target 1)	0.87	0.95	0.91	7383
Inactive (target 0)	0.94	0.86	0.90	7679
accuracy			0.90	15062
macro avg	0.91	0.90	0.90	15062
weighted avg	0.91	0.90	0.90	15062

Table 4. Six architectures of ANN tuned by Optuna

ANN hyperparameters tuned by Optuna					
run	n_layers	neurons	drop_out	optimiser	learn. rate
		67	0.499		
1	2	17	0.296	SGD	0.0009
		23	0.305		
2	2	27	0.296	Adam	0.00003
		28	0.294		
		74	0.489		
3	3	80	0.251	SGD	0.00004
		74	0.353		
		34	0.33		
4	3	45	0.232	SGD	0.0001
		122	0.318		
		49	0.467		
5	3	50	0.329	Adam	0.00005
		114	0.371		
6	2	67	0.227	Adam	0.0005

Table 5. Accuracy results from six runs of RFC and ANN

Accuracy		
	RFC	ANN
	0.9	0.66
	0.902	0.6
	0.9	0.62
	0.9	0.68
	0.898	0.64
	0.9	0.7
	sklearn RFC	ANN
mean	0.900	0.650
stand dev	0.001	0.037

Table 6 Area under the curve results from six runs of sklearn RFC with dataset 4 and PySpark RFC with dataset 5

Area Under the Curve (AUC)		
	sklearn RFC	PySpark RFC
	0.892	0.654
	0.903	0.657
	0.9	0.657
	0.901	0.653
	0.9	0.67
	0.901	0.677
	sklearn RFC	PySpark RFC
mean	0.900	0.661
stand dev	0.004	0.010

Feature Description

A list with the features used for the final datasets (on GitHub)

#	Column
0	Solubility_at_pH_7_4
1	MW
2	TPSA
3	XL
4	HAC
5	HBDC
6	HBAC
7	RBC
8	CBUC
9	MMX6
10	MMX
11	SX6
12	SX
13	MMY6
14	MMY
15	SY6
16	SY
17	Volume_1
18	Volume_2
19	MMX6_3D
20	MMX_3D
21	SX6_3D
22	SX_3D
23	MMY6_3D
24	MMY_3D
25	SY6_3D
26	SY_3D
27	MMZ6_3D
28	MMZ_3D
29	SZ6_3D
30	SZ_3D
31	Volume_1_3D
32	XY_3D_volume
33	XZ_3D_volume
34	YZ_3D_volume
35	C_relative
36	H_relative
37	O_relative
38	S_relative
39	N_relative
40	Br_relative
41	Cl_relative
42	F_relative
43	C
44	H
45	O
46	S
47	N
48	Br
49	Cl
50	F
51	C_rel_2D
52	allAtoms_rel_2D
53	C_rel_XY_3D
54	allAtoms_rel_XY_3D
55	C_rel_XZ_3D
56	allAtoms_rel_XZ_3D
57	C_rel_YZ_3D
58	allAtoms_rel_YZ_3D
59	Similarity
60	target

Features imported from the PubChem database.

1. **Molecular weight** [feature 1 in the list above] is calculated as the sum of the mass of each constituent atom multiplied by the number of atoms of that element in the molecular formula.
2. **The topological polar surface area** [feature 2] is an estimate of the polar surface area of a molecule, computed as the surface sum over polar atoms in the molecule[11].
3. **XLogP3** [3] is a predicted octanol-water partition coefficient measuring the hydrophilicity or hydrophobicity of a molecule[12].
4. **Heavy Atom Count** [4] is the number of heavy atoms (i.e., non-hydrogen atoms) in the given compound.
5. **Hydrogen Bond Donor Count** [5] in the compound
6. **Hydrogen Bond Acceptor Count** [6] in the compound.
7. **Rotatable Bond Count** [7] is defined as any single-order non-ring bond, where atoms on either side of the bond are in turn bound to nonterminal heavy (i.e., non-hydrogen) atoms. That is, where rotation around the bond axis changes the overall shape of the molecule and generates conformers.
8. **Covalently Bonded Unit Count** [8] is the number of groups of atoms connected by covalent bonds.
9. **The atomic coordinates**[13] provided by PubChem carry information about the two and three-dimensional presentation of the molecular structure of the substances and compounds. These coordinates were not used directly as attributes in the new data set but to calculate new features later.
10. **SMILES** Simplified Molecular-Input Line-Entry System[14] that a line notation describes the three-dimensional structure of Chemical formulations was used to calculate the attribute Similarity described below.

11. **Molecular formulas** contain information about the type of atoms and their number in molecules. These data were used to calculate the mass fraction and the relative fraction of atoms in the molecule for each compound under consideration.

Features derived by additional calculations.

13. **Difference between Min and Max of the atoms' coordinates** [features 9-10; 13-14; 19-20; 23-24; 27-28] of the compounds. Based on the PubChem database, in other to distinguish the isomers from each other, the minima of the 2D coordinates of the substances were subtracted from the maxima along the x and y axes. Initially, this was performed for all atoms, but given that carbon atoms create the "skeleton" of the organic compound, similar calculations were also done for carbon atoms only. Thus, four new features were added to the new data set. The same was done with the 3D coordinates of the compounds, but this time six new features were added as z-axis calculations were also considered.
14. **Skewness of the atom coordinate distribution** [features 11-12; 15-16; 21-22; 25-26; 29-30]. The next four features derived from 2D coordinates and six new features from 3D coordinates are related to the distribution of the atoms along the x, y and z axes. Following the logic of the min-max difference explained above, the values of skewness were calculated first for all atoms and then only for the carbon atoms.
15. **Two types of hypothetical volumes of the molecules** [features 17-18, 31-34]. The first type was calculated based on the 2D coordinates that carry information about the isomers. Since atoms, respectively molecules are not flat, the 2D data of the atomic coordinates of the substances were used to obtain a hypothetical 3D volume of the compounds. These computations were based on the previously obtained min-max values of the atoms' coordinates. The second variant of the imaginary volumes was calculated using the already computed skewness of the atoms` coordinates.
16. **The relative proportion of the atoms in the molecules of the considered compounds** [features 35-42], are features that carry information about the proportion of the number of a given type of atom to the total number of atoms in the molecules of the given compound. This added eight features where each one of them was named after the corresponding chemical element.
17. **The mass proportion of the atoms in the molecules of the considered compound** [features 43-50] was calculated as the proportion of the mass of all identical atoms in the molecule to the mass of all atoms in the molecule of the given compound. These calculations added eight features to the new data set.
18. **The size ratio of the molecules of a considered compound** [features 51-58] was calculated along the x, y and z axes, creating six new columns.
19. **SMILES similarity** [feature 59] was based on the comparison between SMILES of the compounds and Lysine. Lysine was chosen because the methylation has been performed in the Lysine residue. The “Similarity” feature was generated using the toolkit for cheminformatics RDKit‘s packages. For more information, see the code in GitHub, file “data_sets_generation_with_SMOTE. ipynb “, [In 87] “Adding similarity between the compounds based on their SMILES.

Packages

- PyPl https://pypi.org/

- missingno https://pypi.org/project/missingno/

- Chemformula https://pypi.org/project/chemformula/

- Tables https://pypi.org/project/tables/

- Jupyter Notebook How to cite Jupyter notebook - Cite Bay
- NumPy - Citing NumPy
- Matplotlib https://matplotlib.org/
- Seaborn https://seaborn.pydata.org/
- pandas https://pandas.pydata.org/
- Rdkit. https://www.rdkit.org https://doi.org/10.5281/zenodo.591637
- imbalanced-learn https://imbalanced-learn.org/stable/
- Xgboost https://xgboost.readthedocs.io/en/stable/