KNIMEを使った材料探索
基本操作(3)

早稲田大学 応用化学科
講師(任期付) 畠山 歓
https://github.com/KanHatakeyama/
satokan@toki.早稲田.jp

今回扱う内容

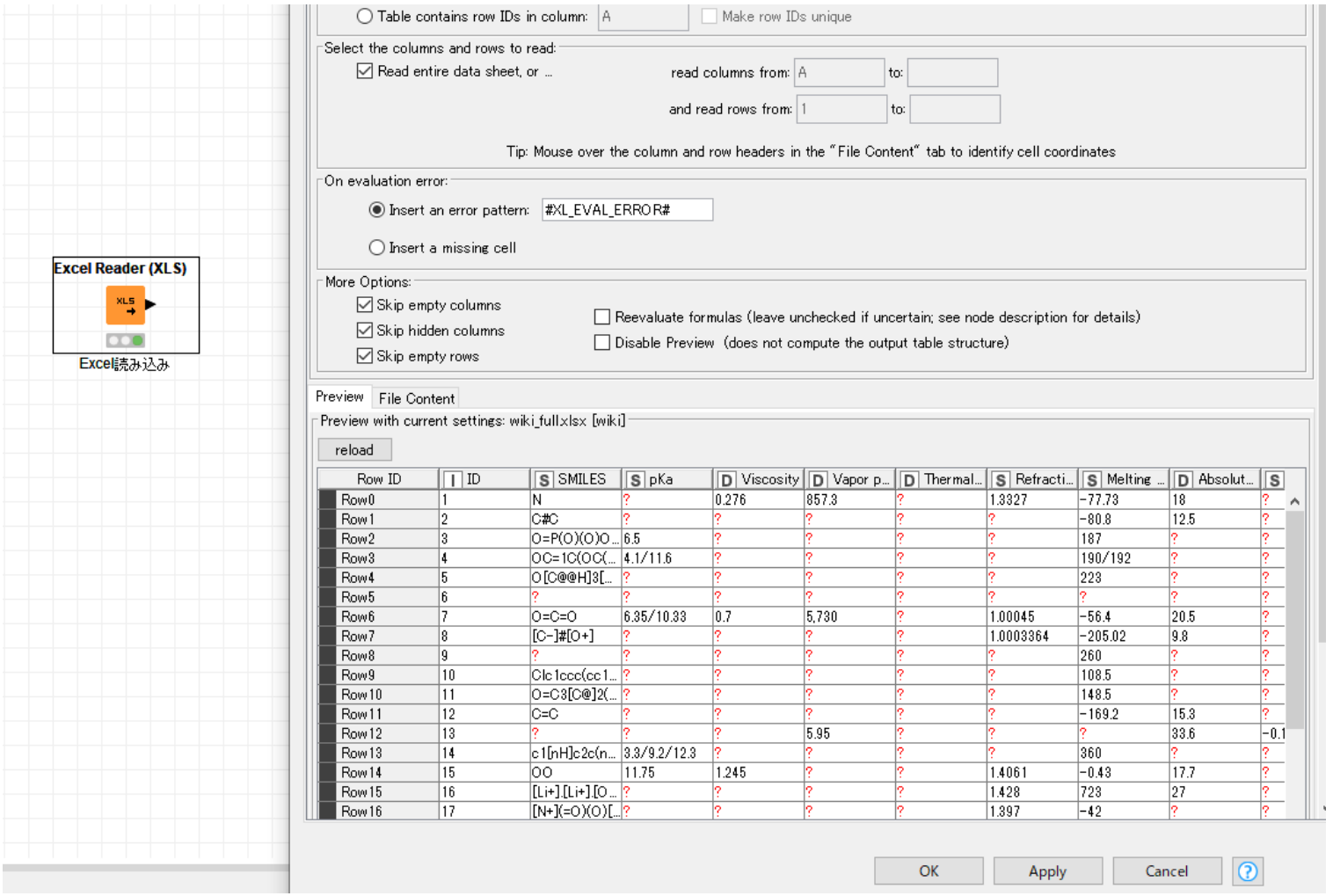
- 欠損データの処理
- 記述子の利用
- 線形モデルの回帰例

	A	B	C	D	E	F	G	H	I	J
	ID	SMILES	pKa	Viscosity	Vapor pre	Thermal C	Refractive	Melting te	Absolute	Partitic
	1	N		0.276	857.3		1.3327	-77.73	18	
	2	C#C						-80.8	12.5	
	3	O=P(O)(O	6.5					187		
	4	OC=1C(OC	4.1/11.6					190/192		
	5	O[C@@H]3[C@@H](O)[C@H](O)[C@@H](CO)O[C@H]3OC[C@						223		
	6									
	7	O=C=O	6.35/10.33	0.7	5730		1.00045	-56.4	20.5	
	8	[C-]#[O+]					1.000336	-205.02	9.8	
	9							260		
	10	O=C(c1cc(Cl)c(Cl)cc1Cl)c2c(Cl)cc(Cl)cc2Cl						108.5		
	11	O=C3[C@]2(CC[C@@H]1[C@@]4(C(=C/C[C@H]1[C@@H]2CC3						148.5		
	12	C=C						-169.2	15.3	
	13				5.95				33.6	-0.
	14	c1[nH]c2c	3.3/9.2/12.3					360		
	15	OO	11.75	1.245			1.4061	-0.43	17.7	
	16	[Li+].[Li+]	[O-]C([O-])=O				1.428	723	27	

今回のデータベース

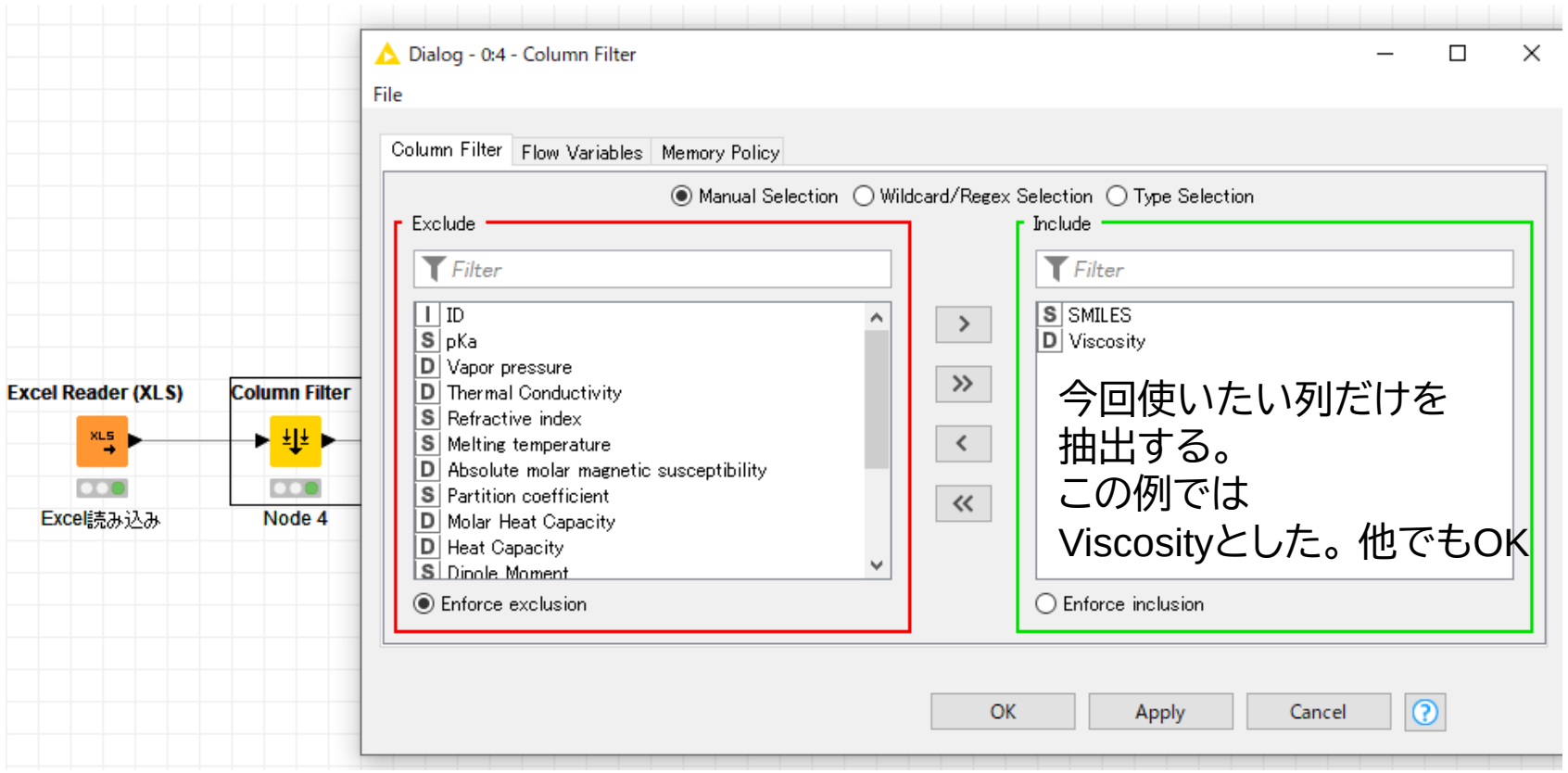
wiki_full.xlsx

欠損データが非常に多い



Row ID	ID	SMILES	pKa	Viscosity	Vapor p...	Thermal...	Refracti...	Melting ...	Absolut...	
Row0	1	N	?	0.276	857.3	?	1.3327	-77.73	18	?
Row1	2	C#C	?	?	?	?	?	-80.8	12.5	?
Row2	3	O=P(O)(O)O...	6.5	?	?	?	?	187	?	?
Row3	4	OC=1C(OC(...	4.1/11.6	?	?	?	?	190/192	?	?
Row4	5	O[C@@H]3[...	?	?	?	?	?	223	?	?
Row5	6	?	?	?	?	?	?	?	?	?
Row6	7	O=C=O	6.35/10.33	0.7	5,730	?	1.00045	-56.4	20.5	?
Row7	8	[C-]#[O+]	?	?	?	?	1.0003364	-205.02	9.8	?
Row8	9	?	?	?	?	?	?	260	?	?
Row9	10	Clc1ccc(cc1...	?	?	?	?	?	108.5	?	?
Row10	11	O=C3[C@]2(...	?	?	?	?	?	148.5	?	?
Row11	12	C=C	?	?	?	?	?	-169.2	15.3	?
Row12	13	?	?	?	5.95	?	?	?	33.6	-0.1
Row13	14	c1[nH]c2c(n...	3.3/9.2/12.3	?	?	?	?	360	?	?
Row14	15	OO	11.75	1.245	?	?	1.4061	-0.43	17.7	?
Row15	16	[Li+].[Li+].[O...	?	?	?	?	1.428	723	27	?
Row16	17	[N+](=O)(O)[...	?	?	?	?	1.397	-42	?	?

Wiki_full の読み込み



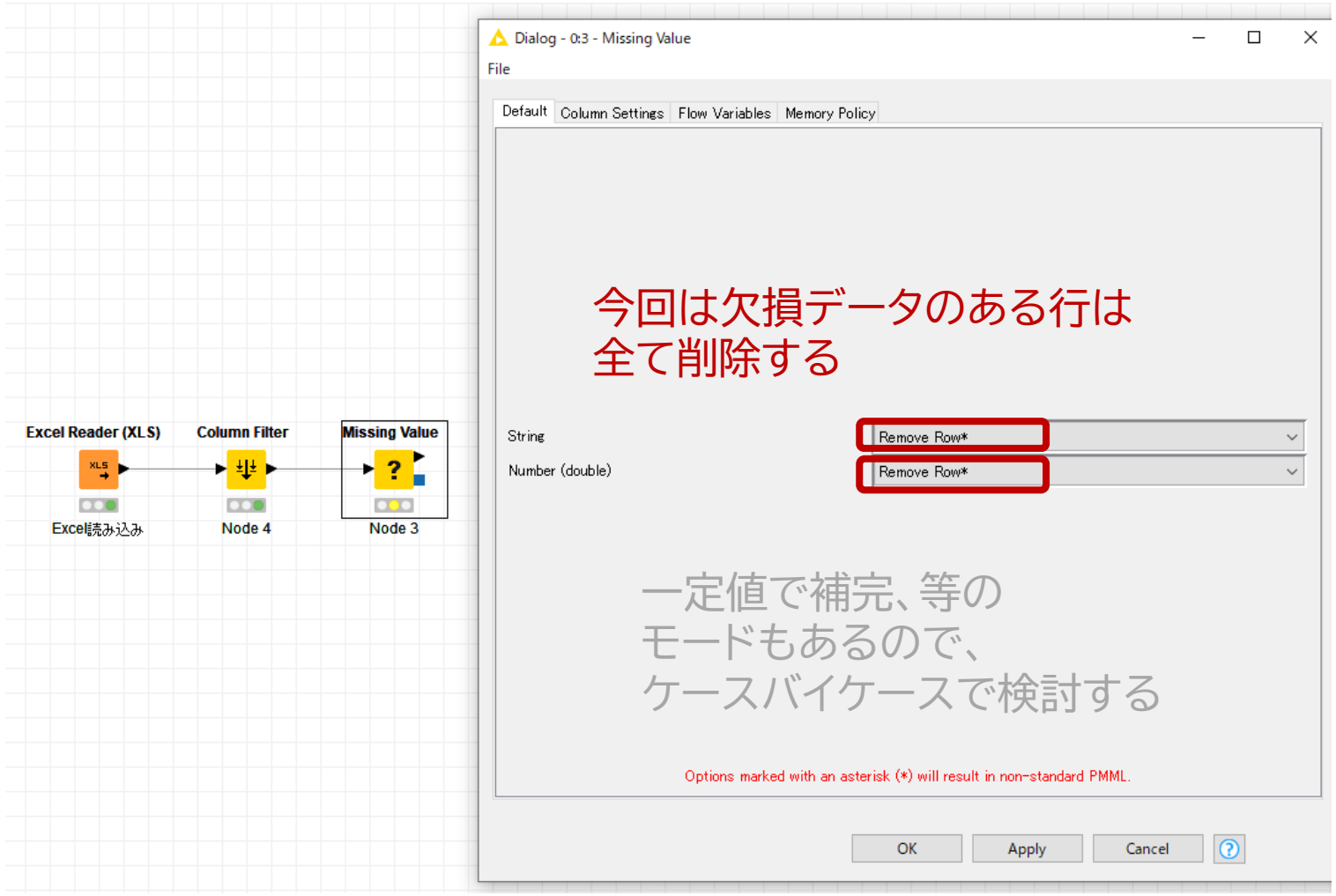
Column filterを設置&設定

Row10	O=C3[C@]2(...	?
Row11	C=C	?
Row12	?	?
Row13	c1[nH]c2c(n...	?
Row14	OO	1.245
Row15	[Li+].[Li+].[O ...	?
Row16	[N+](=O)(O)[...	?
Row17	c1c2c(nc[nH...	?
Row18	c1cncnc1	?
Row19	?	?
Row20	c1ccc(cc1)O	?
Row21	C1OCNCC1	1.573

欠損データばかりで学習出来ない

Column filterの Filtered table*

*ノードを右クリックしてfiltered tableを選択

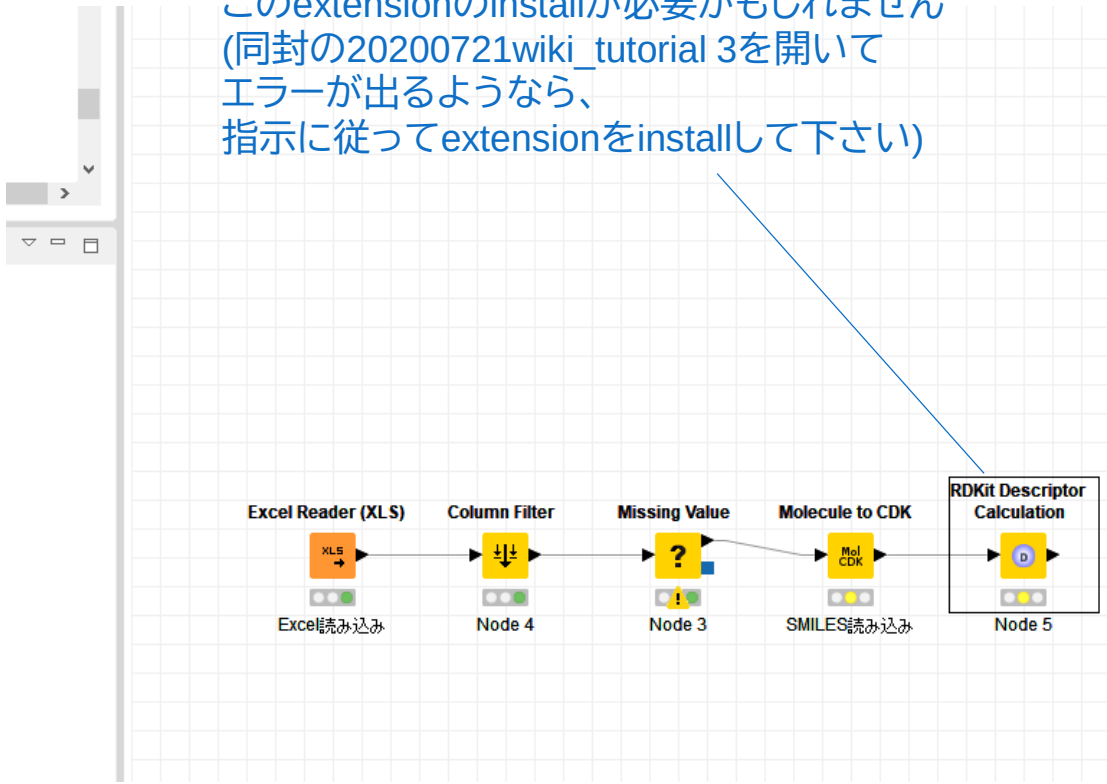


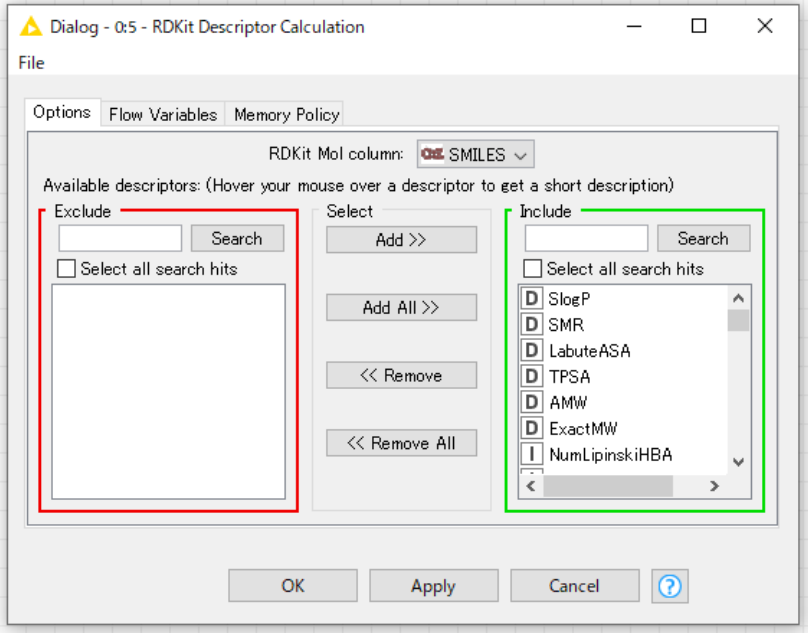
Missing valueの設置&設定

Row61	O=CO	1.57
Row68	ClC(Cl)Cl	0.563
Row84	O=C(C)CC	0.43
Row88	c1ccccc1C=C	0.762
Row91	N1C=CC=C1	1.225
Row108	[O+]#C[Ni-4...	0.305
Row109	CC(O)CO	42
Row115	C1CCOC1	0.48
Row118	Cl/C(Cl)=C(...	0.89
Row119	C(=S)=S	0.436
Row121	C[N+](=O)[O ...	0.63
Row173	Cl[Fe](Cl)Cl	12

Missing value のoutput table

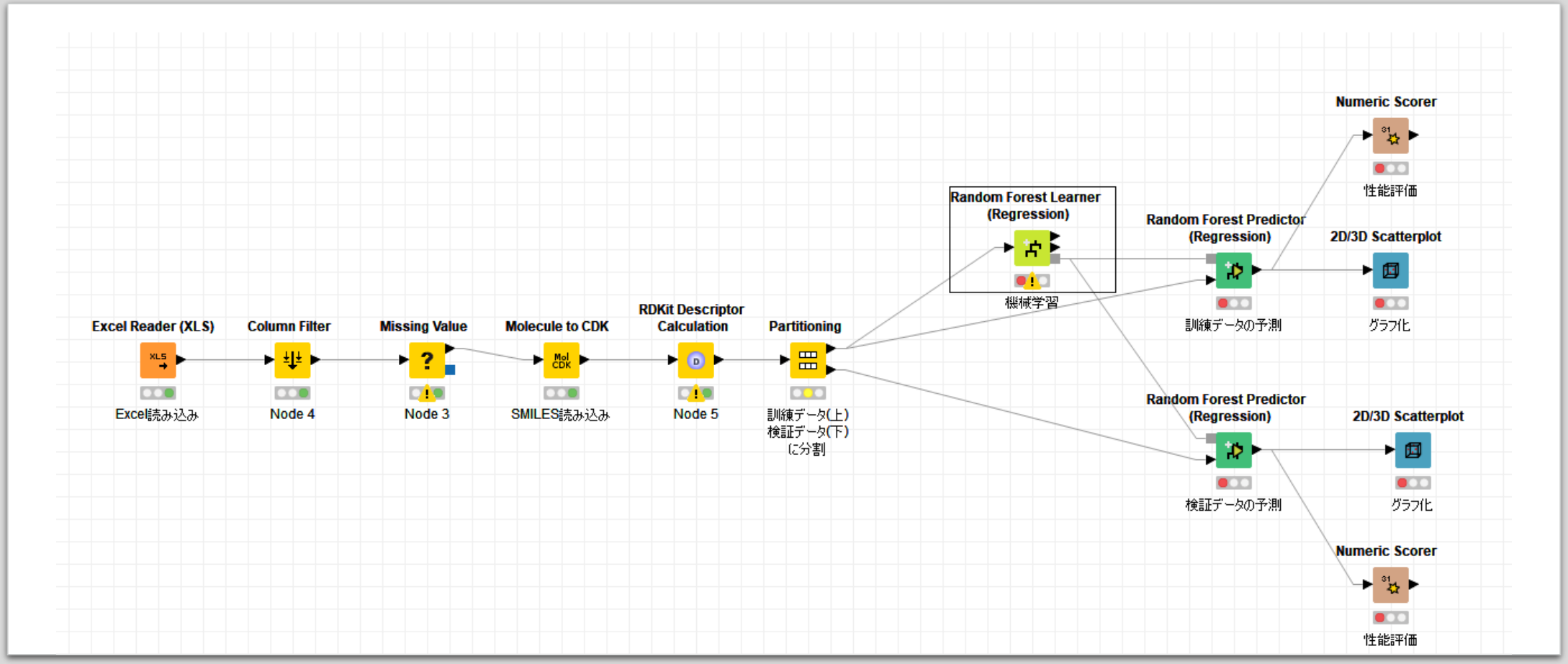
このextensionのinstallが必要かもしれません
(同封の20200721wiki_tutorial 3を開いて
エラーが出るようなら、
指示に従ってextensionをinstallして下さい)



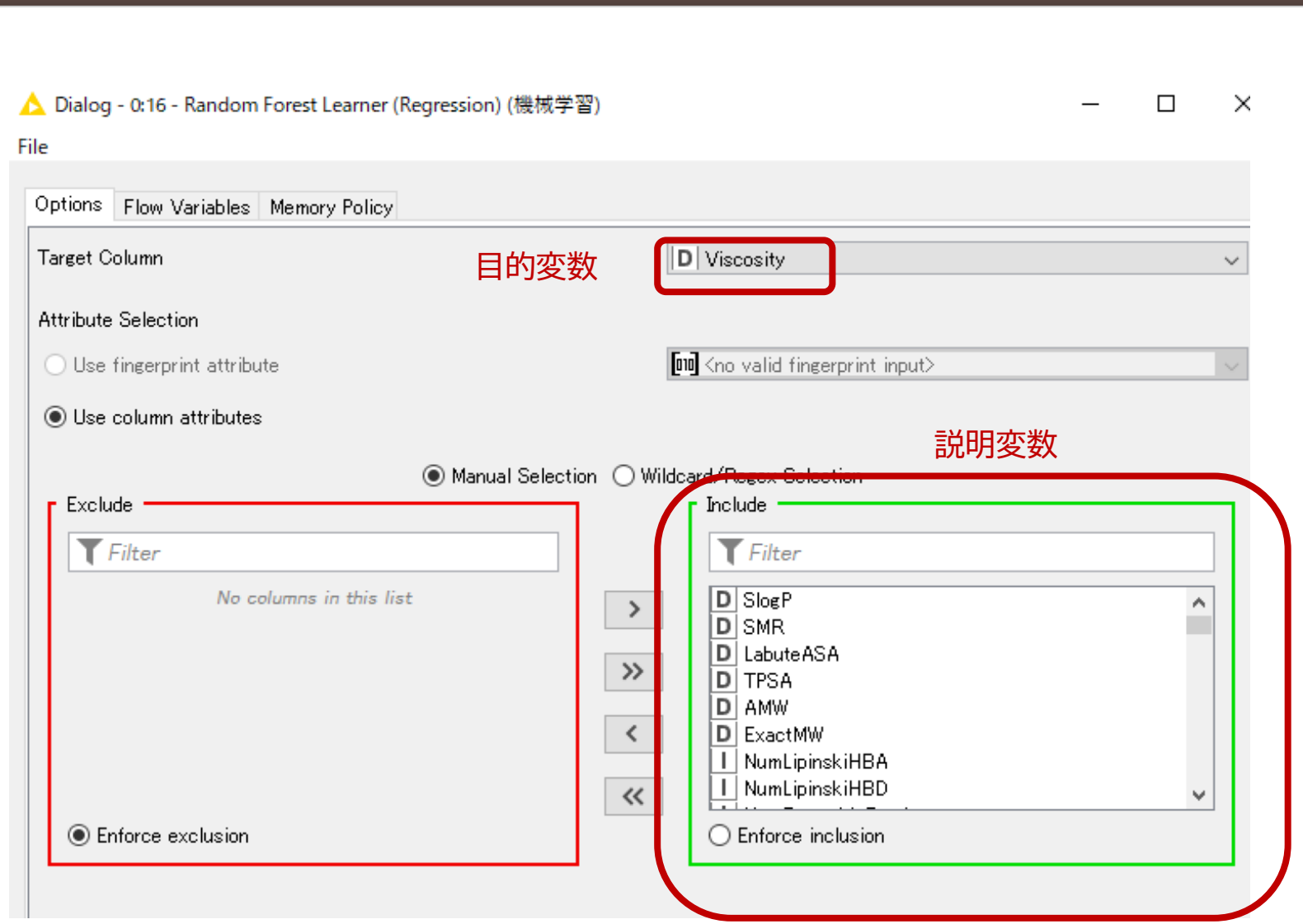


対象分子の分子量、極性etcを
自動で計算してくれる

今回はDescriptorと呼ばれる手法で分子構造を数値化

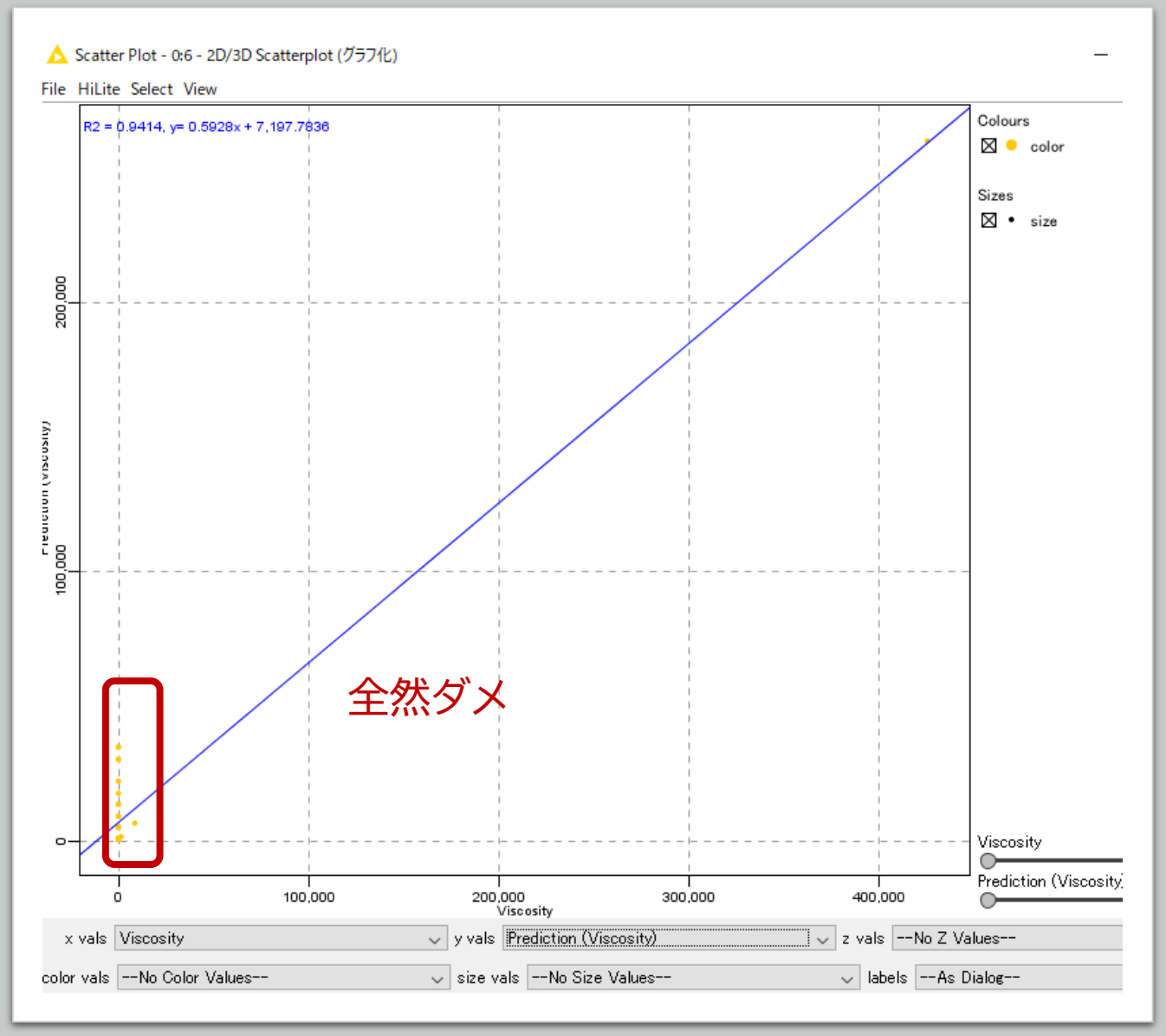


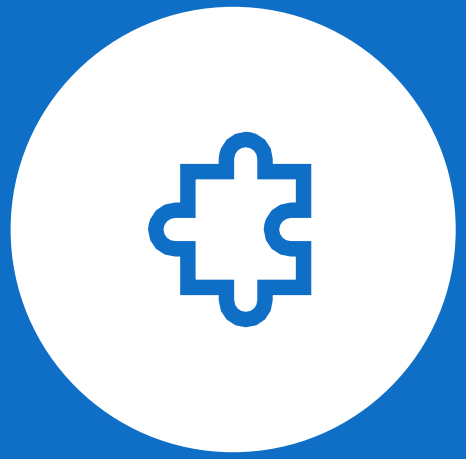
ここから先は前回と同じ



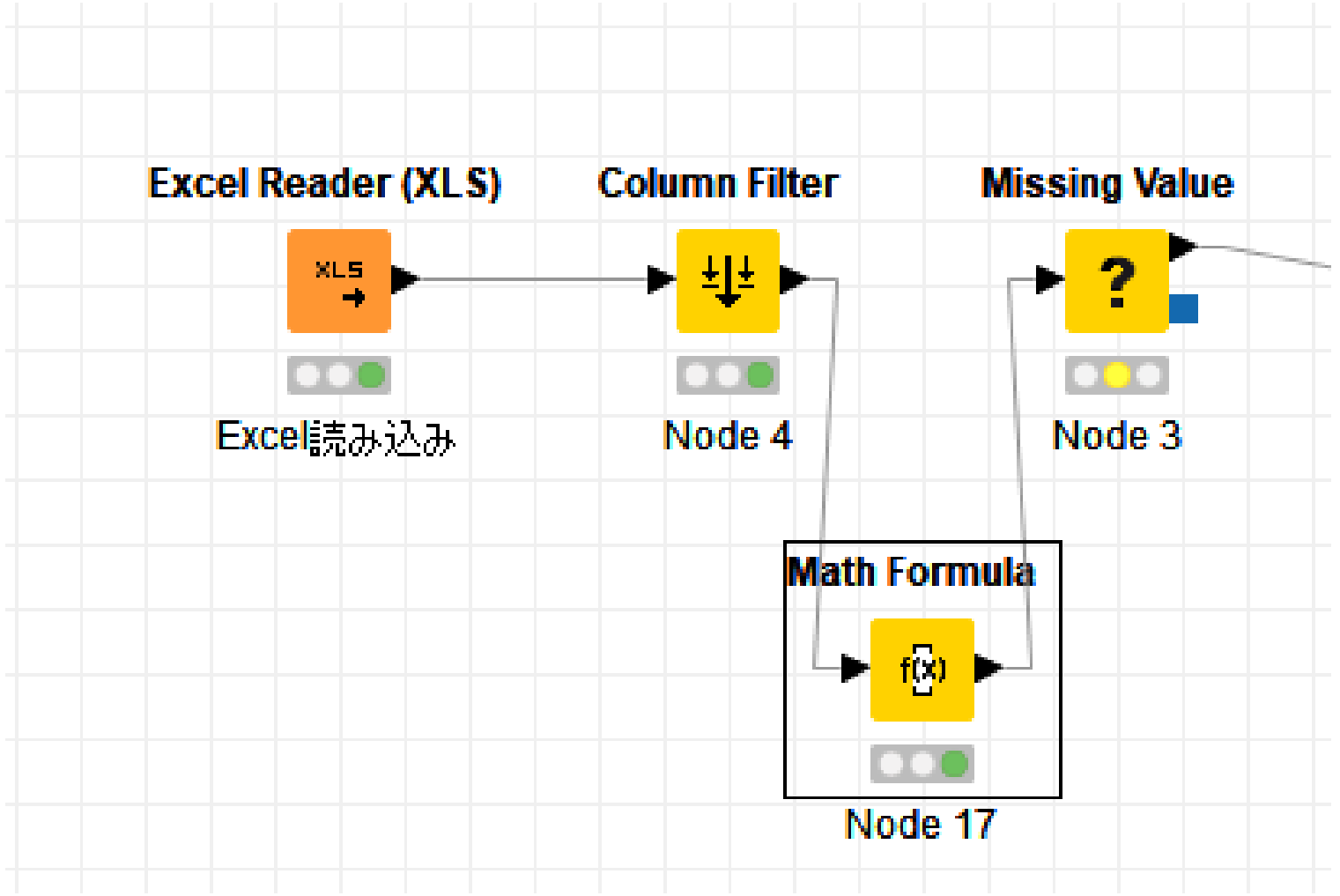
Learnerの設定を間違えないこと

訓練結果

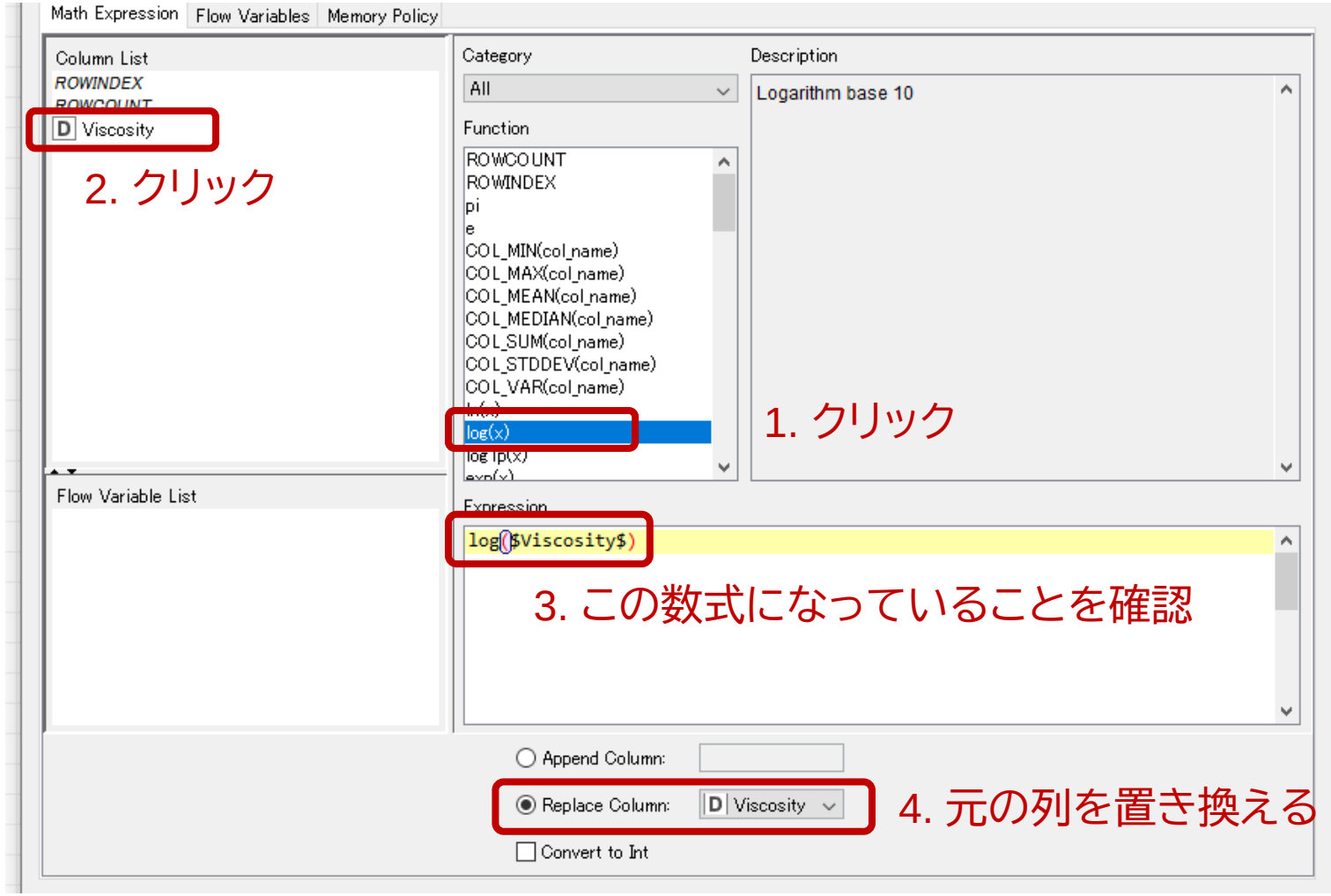

Scatter Plot - 0:6 - 2D/3D Scatterplot (グラフ化)
File HiLite Select View
R2 = 0.9414, y= 0.5928x + 7,197.7836
Colours
color
Sizes
size
200,000
100,000
0
0
100,000
200,000
300,000
400,000
Viscosity
全然ダメ
Viscosity
Prediction (Viscosity)
x vals Viscosity
y vals Prediction (Viscosity)
z vals --No Z Values--
color vals --No Color Values--
size vals --No Size Values--
labels --As Dialog--



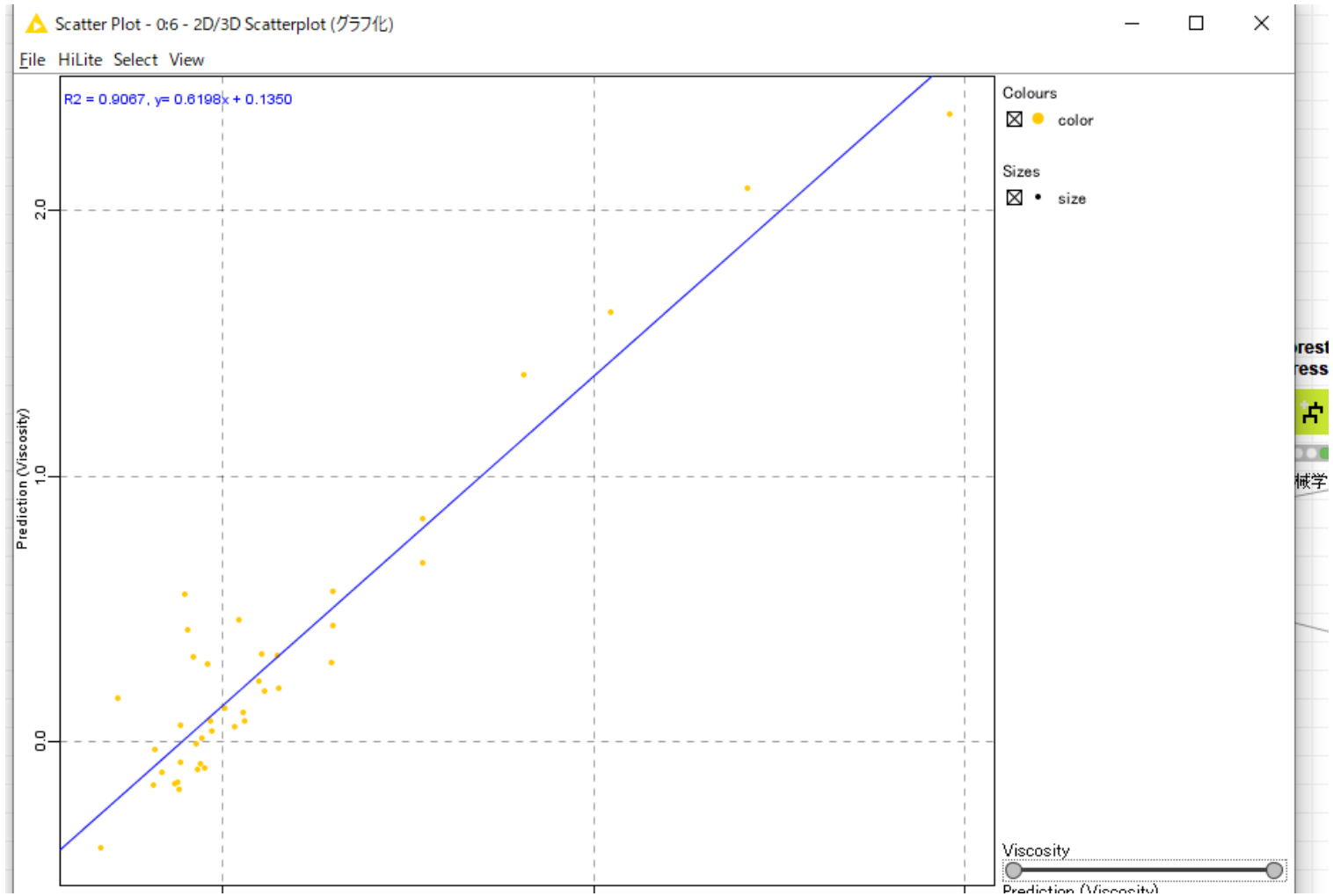
どうしたら良いか?



Math formulaを追加してみる

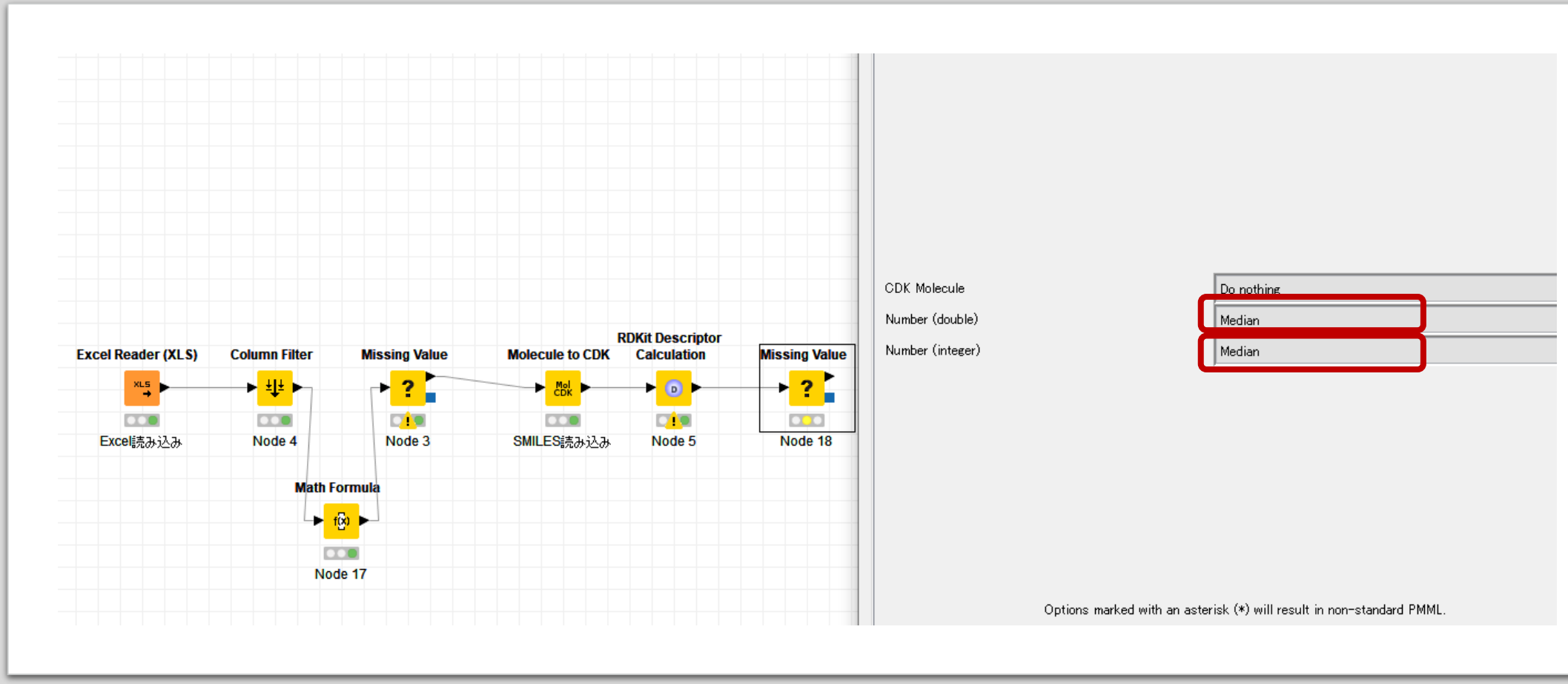


Viscosityを logに変換

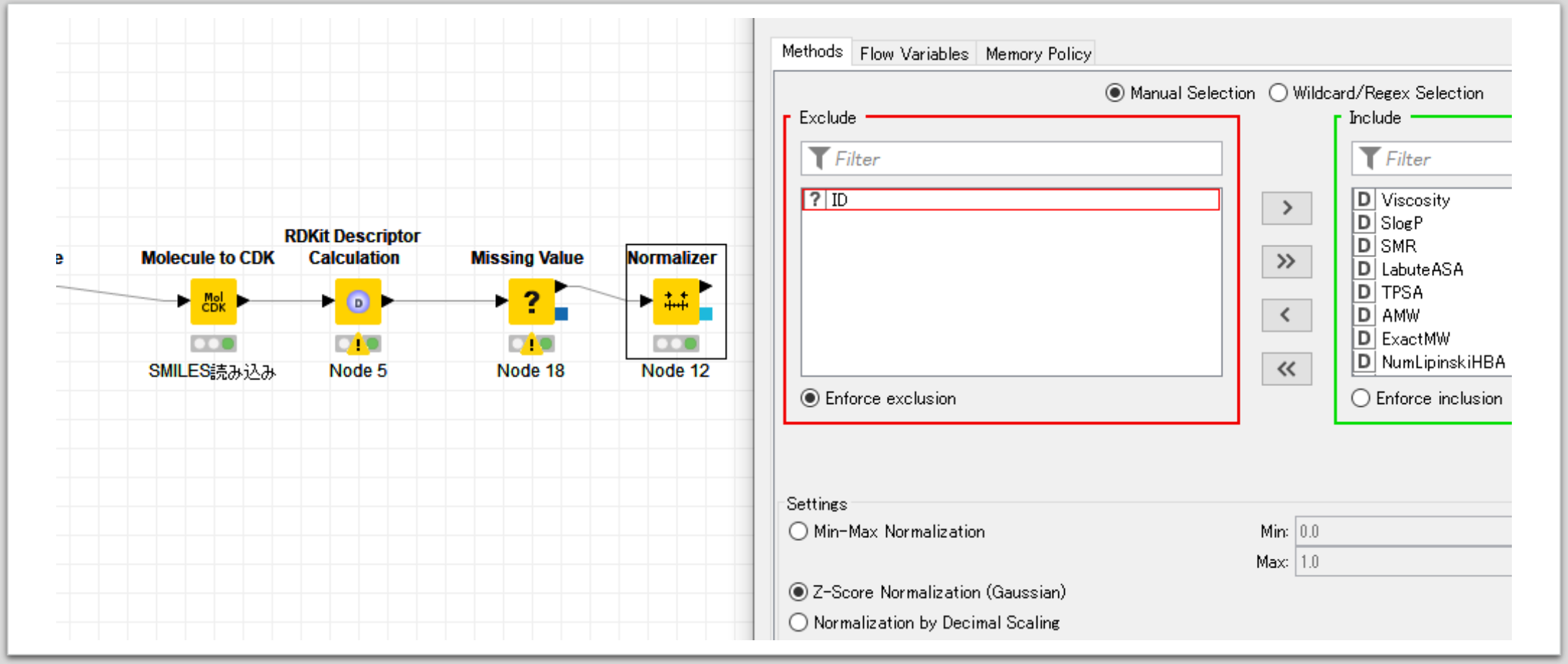

Scatter Plot - 0:6 - 2D/3D Scatterplot (グラフ化)
File HiLite Select View
R2 = 0.9067, y= 0.6198x + 0.1350
Prediction (Viscosity)
2.0
1.0
0.0
Colours
color
Sizes
size
Viscosity

精度が向上

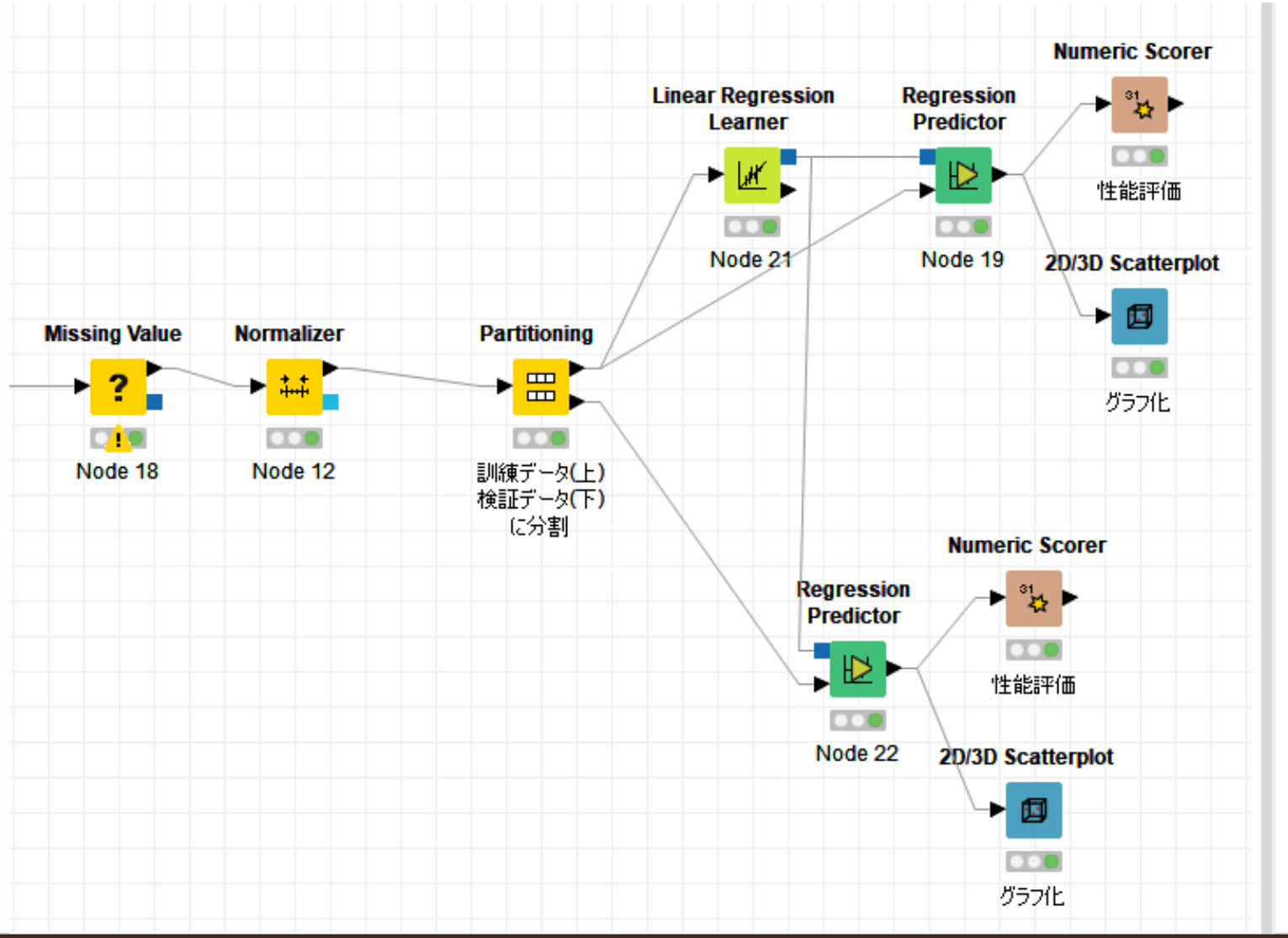
参考1: 線形モデルの検討



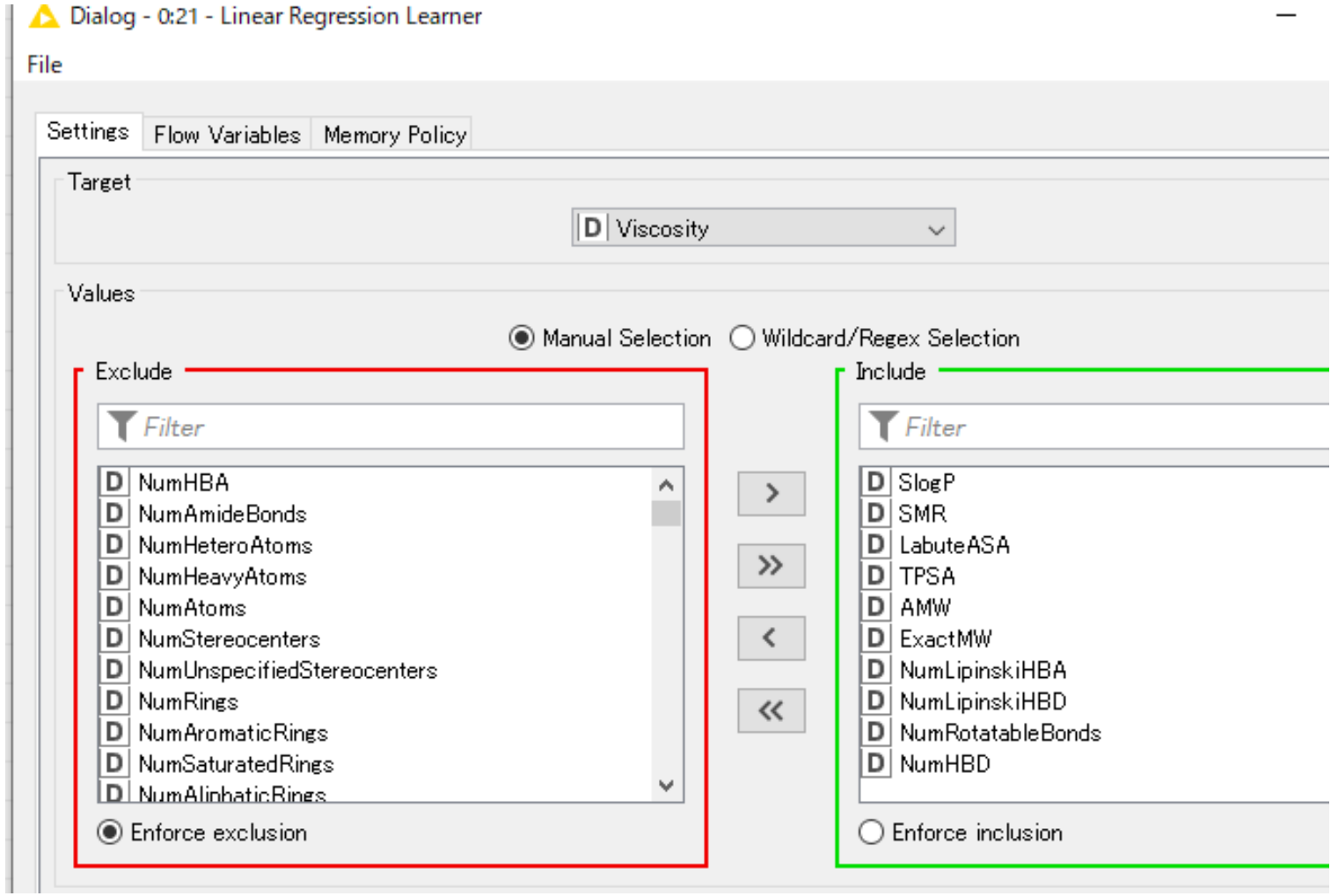
Descriptorは地味に欠損値があるので、中央値(median)で補填する



説明変数、目的変数全てを正規化する



Random forestと同じ要領で回帰する



回帰時の説明変数を自分で色々と選んでみる

説明変数の選択が悪いと、上手くいかない場合もある

ficients and Statistics" - Rows: 6 | Spec - Columns: 5 | P

	S Variable	D Coeff.	D Std. Err.	D t-v
	SlogP	-0.599	0.376	-1.595
	SMR	0.314	0.914	0.343
	LabuteASA	0.538	1.073	0.501
	TPSA	-0.144	0.191	-0.754
	ExactMW	0.061	0.385	0.159
	Intercept	-0.083	0.139	-0.599

分子量(MW)が上がると、粘度が上がることを示唆??

回帰モデルの係数

モデルを右クリック→Coefficients and statisticsをクリック

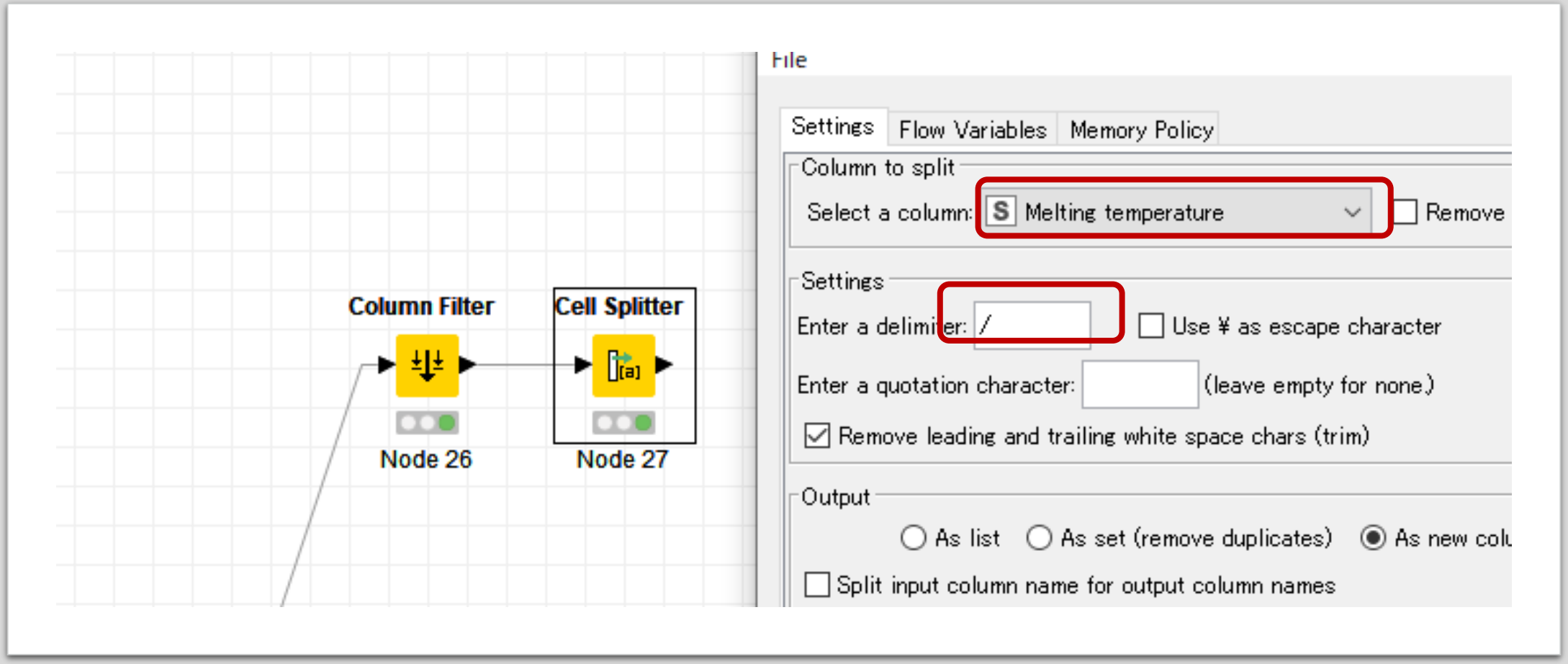
参考2: データに不要な文字列が入っている場合

融点

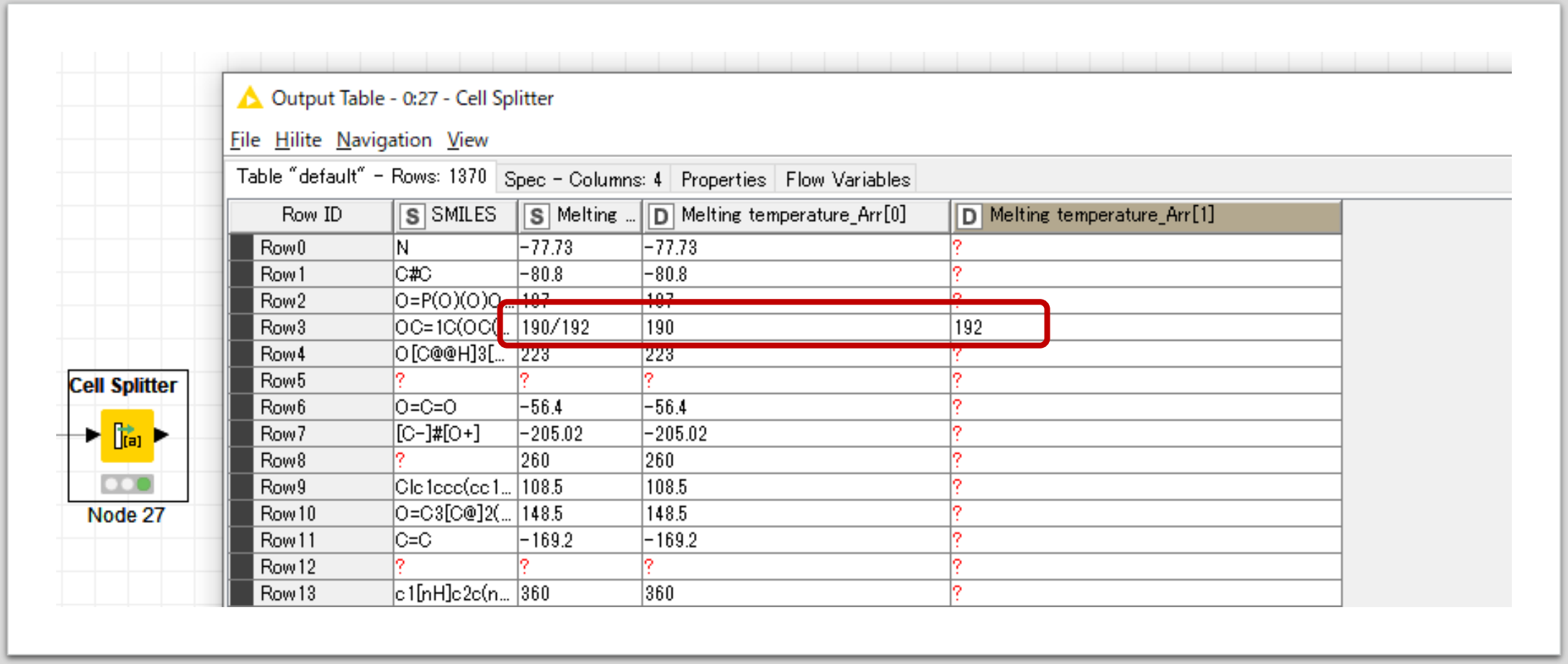
Table default Rows: 1079 Spec - Columns: 2

Row ID	SMILES	Melting ...
Row0	N	-77.73
Row1	C#C	-80.8
Row2	O=P(O)(O)O...	187
Row3	OC=1C(OC(...	190/192
Row4	O[C@@H]3[...	223
Row5	?	?
Row6	O=C=O	-56.4
Row7	[C-]#[O+]	-205.02
Row8	?	260
Row9	Clc1ccc(cc1...	108.5
Row10	O=C3[C@]2(...	148.5
Row11	C=C	-169.2
Row12	?	?
Row13	c1[nH]c2c(n...	360
Row14	OO	-0.43
Row15	[Li+].[Li+].[O...	723
Row16	[N+](=O)(O)[...	-42
Row17	c1c2c(nc[nH...	214

スラッシュ区切りなので学習出来ない!



Cell splitter



Row ID	SMILES	Melting ...	Melting temperature_Arr[0]	Melting temperature_Arr[1]
Row0	N	-77.73	-77.73	?
Row1	C#C	-80.8	-80.8	?
Row2	O=P(O)(O)O...	197	197	?
Row3	OC=1C(OC(...	190/192	190	192
Row4	O[C@@H]3[...	223	223	?
Row5	?	?	?	?
Row6	O=C=O	-56.4	-56.4	?
Row7	[C-]#[O+]	-205.02	-205.02	?
Row8	?	260	260	?
Row9	Clc1ccc(cc1...	108.5	108.5	?
Row10	O=C3[C@]2(...	148.5	148.5	?
Row11	C=C	-169.2	-169.2	?
Row12	?	?	?	?
Row13	c1[nH]c2c(n...	360	360	?

Output table

二つの列に分割されたので、とりあえず機械学習は可能になった

その他 困ったときの処理

String manipulation

Math formula

Rule engine

これらのノードを使えば、大半の細かなデータ処理は可能 (詳細は割愛)
それでも難しい時は、Python script ノードを利用する

TODO

もっと色々なモデルを試したい