

# 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	Partitio
	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	
0	9							260		
1	10	ClC(Cl)(Cl)C(Cl)(Cl)Cl						108.5		
2	11	O=C3[C@]2(CC[C@@H]1[C@@]4(C(=C/C[C@H]1[C@@H]2CC3)						148.5		
3	12	C=C						-169.2	15.3	
4	13				5.95				33.6	-0.
5	14	c1[nH]c2c	3.3/9.2/12.3					360		
5	15	OO	11.75	1.245			1.4061	-0.43	17.7	
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



☐ Table contains row IDs in column:  ☐ Make row IDs unique

Select the columns and rows to read:

☒ Read entire data sheet, or ... read columns from:  to:

and read rows from:  to:

Tip: Mouse over the column and row headers in the "File Content" tab to identify cell coordinates

On evaluation error:

☒ Insert an error pattern:

☐ Insert a missing cell

More Options:

☒ Skip empty columns ☐ Reevaluate formulas (leave unchecked if uncertain; see node description for details)

☒ Skip hidden columns ☐ Disable Preview (does not compute the output table structure)

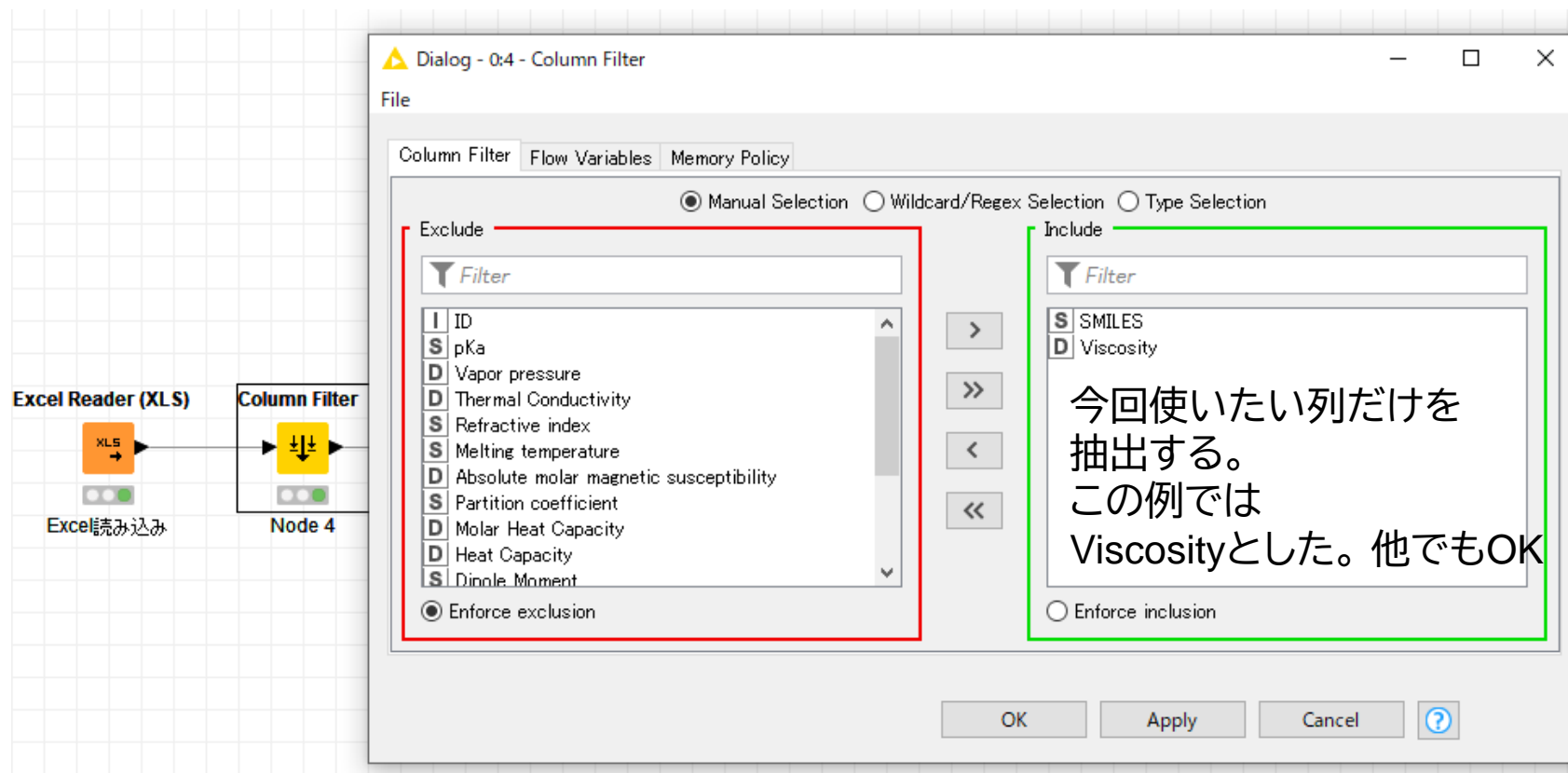
☒ Skip empty rows

Preview File Content

Preview with current settings: wiki\_full.xlsx [wiki]

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



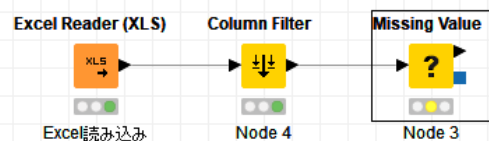
# Column filterを設置&設定

Row 9	O=C1CCC(=C1)C	?
Row 10	O=C3[C@]2(...	?
Row 11	C=C	?
Row 12	?	?
Row 13	c1[nH]c2c(n...	?
Row 14	OO	1.245
Row 15	[Li+].[Li+].[O ...	?
Row 16	[N+](=O)(O)[...	?
Row 17	c1c2c(nc[nH...	?
Row 18	c1cnncnc1	?
Row 19	?	?
Row 20	c1ccc(cc1)O	?
Row 21	O1CCNCC1	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



Dialog - 0:3 - Missing Value

File

Default Column Settings Flow Variables Memory Policy

今回は欠損データのある行は  
全て削除する

String Remove Row\*

Number (double) Remove Row\*

一定値で補完、等の  
モードもあるので、  
ケースバイケースで検討する

Options marked with an asterisk (\*) will result in non-standard PMML.

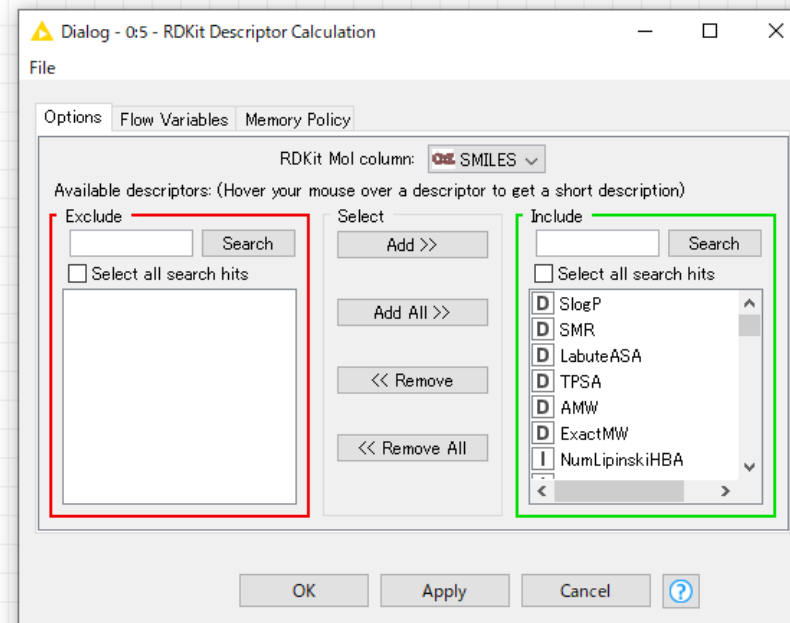
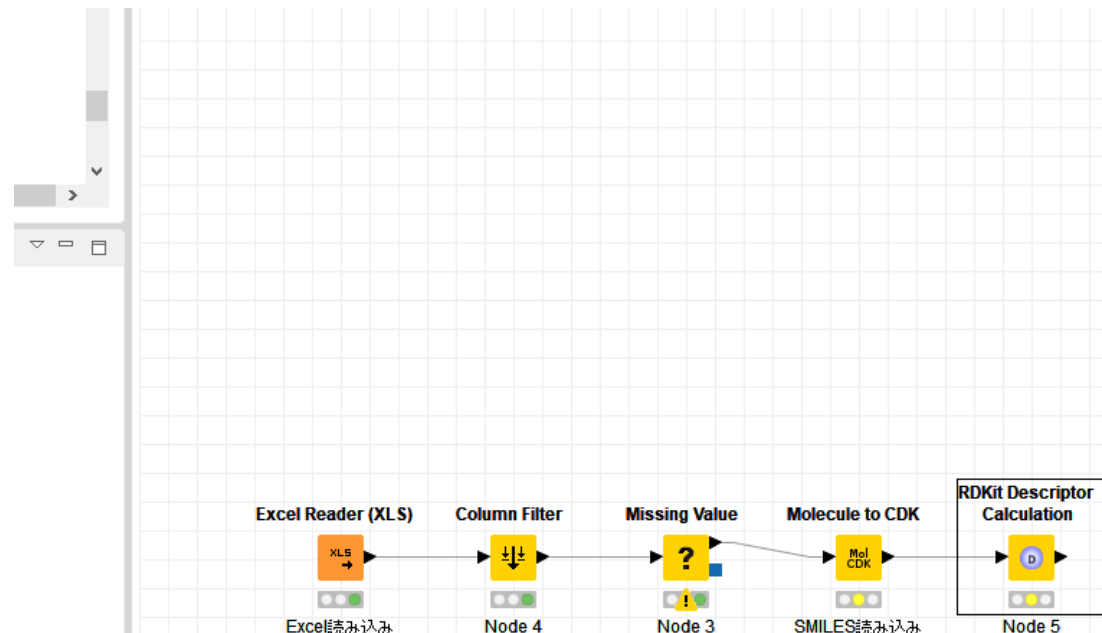
OK Apply Cancel ?

# Missing valueの 設置&設定

Row	SMILES	Value
Row 61	<chem>O=CO</chem>	1.57
Row 68	<chem>ClC(Cl)Cl</chem>	0.563
Row 84	<chem>O=C(C)CC</chem>	0.43
Row 88	<chem>c1ccccc1C=O</chem>	0.762
Row 91	<chem>N1C=CC=C1</chem>	1.225
Row 108	<chem>[O+]#C[Ni-4...]</chem>	0.305
Row 109	<chem>CC(O)CO</chem>	42
Row 115	<chem>C1CCOC1</chem>	0.48
Row 118	<chem>Cl/C(C)=C(...)</chem>	0.89
Row 119	<chem>C(=S)=S</chem>	0.436
Row 121	<chem>C[N+](=O)[O...]</chem>	0.63
Row 173	<chem>Cl[Fe](Cl)Cl</chem>	12

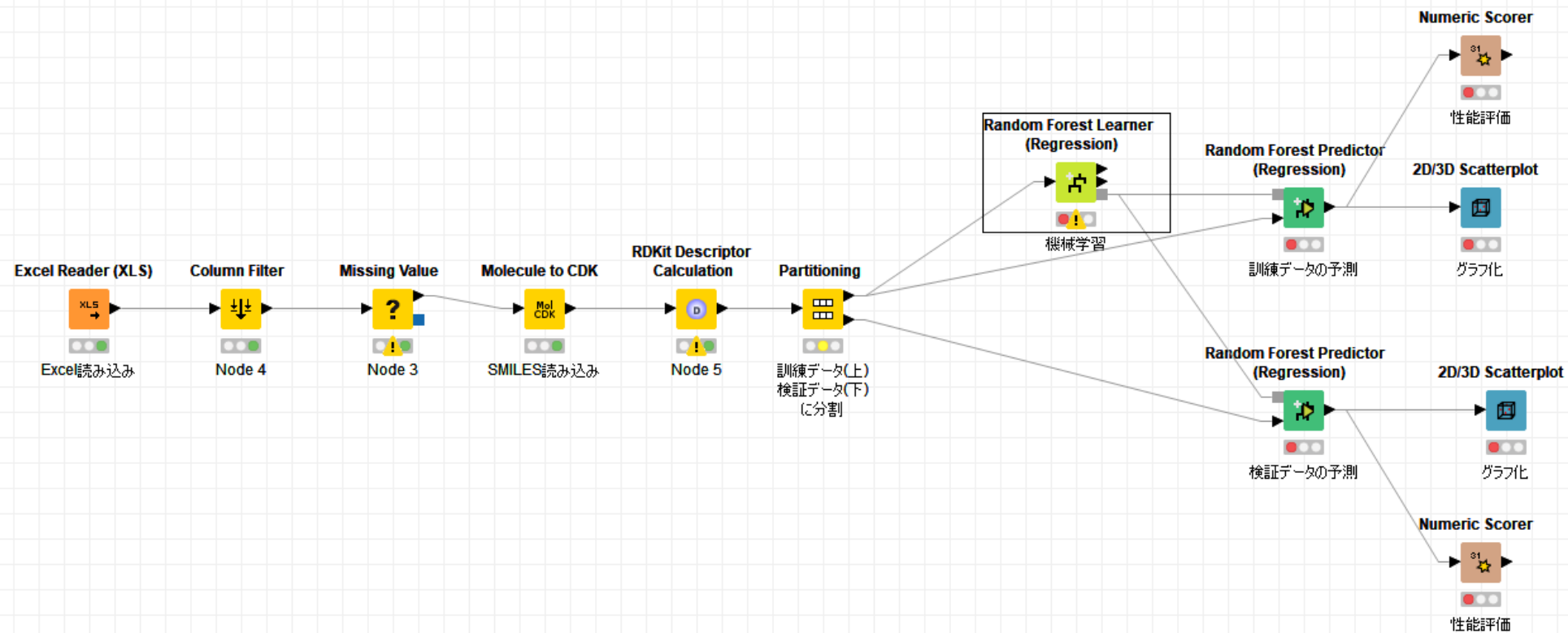
Missing  
value  
Output  
table





対象分子の分子量、極性等を  
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

Target Column

目的変数

D Viscosity

Attribute Selection

☐ Use fingerprint attribute

[010] <no valid fingerprint input>

☒ Use column attributes

☒ Manual Selection

☐ Wildcard/Regex Selection

説明変数

Exclude

Filter

No columns in this list

☒ Enforce exclusion

Include

Filter

D SlogP

D SMR

D LabuteASA

D TPSA

D AMW

D ExactMW

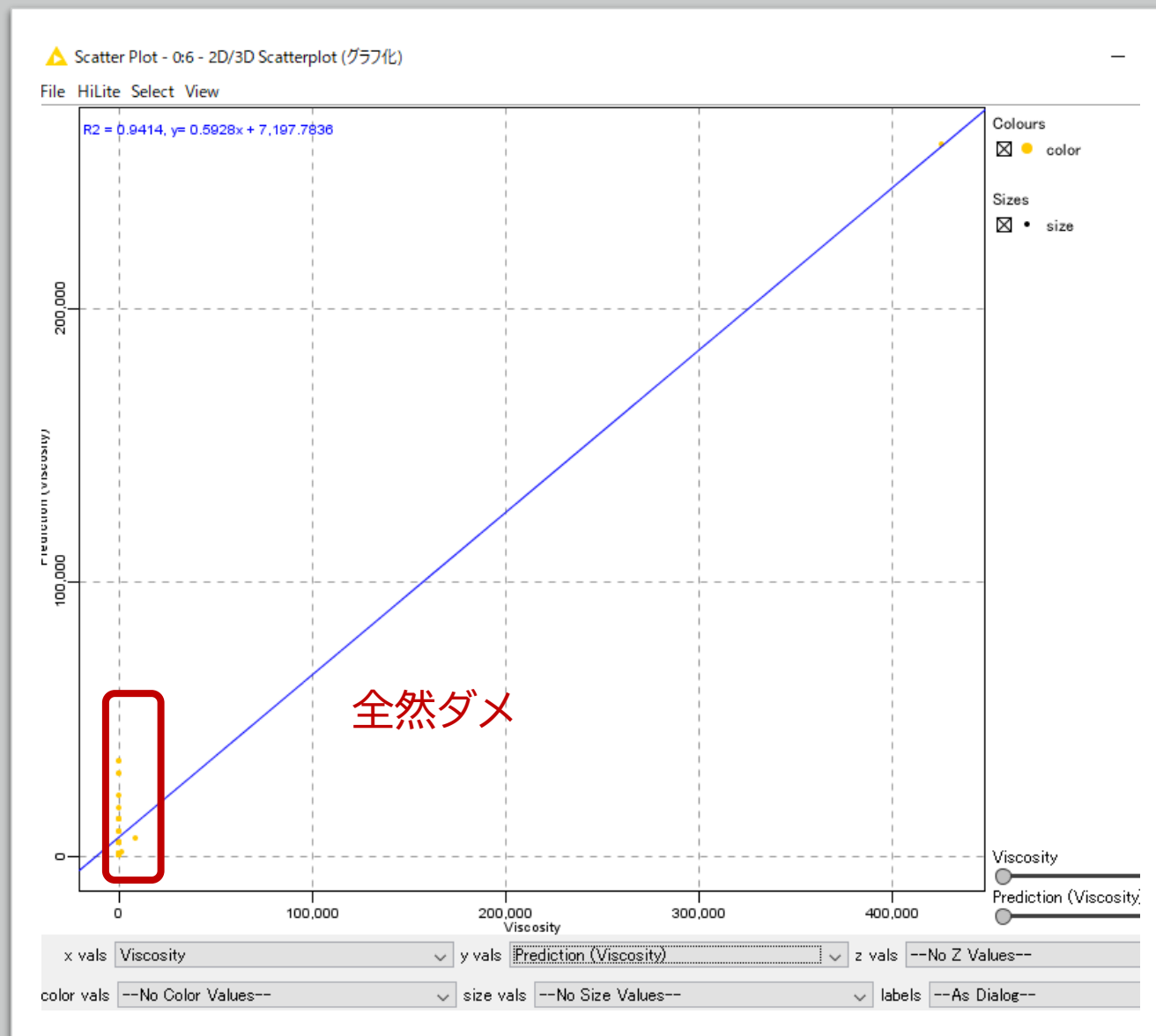
I NumLipinskiHBA

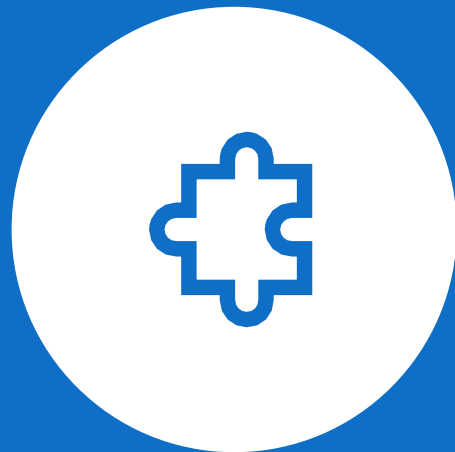
I NumLipinskiHBD

☐ Enforce inclusion

Learnerの  
設定を  
間違えない  
こと

# 訓練結果





どうしたら良いか？

Excel Reader (XLS)



Excel読み込み

Column Filter



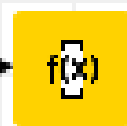
Node 4

Missing Value



Node 3

Math Formula



Node 17

Math  
formulaを  
追加  
してみる

# Viscosityをlogに変換

Math Expression   Flow Variables   Memory Policy

Column List  
ROWINDEX  
ROWCOUNT  
**D Viscosity**

2. クリック

Category  
All

Description  
Logarithm base 10

Function  
ROWCOUNT  
ROWINDEX  
pi  
e  
COL\_MIN(col\_name)  
COL\_MAX(col\_name)  
COL\_MEAN(col\_name)  
COL\_MEDIAN(col\_name)  
COL\_SUM(col\_name)  
COL\_STDDEV(col\_name)  
COL\_VAR(col\_name)  
**log(x)**

1. クリック

Flow Variable List

Expression  
**log(\$Viscosity\$)**

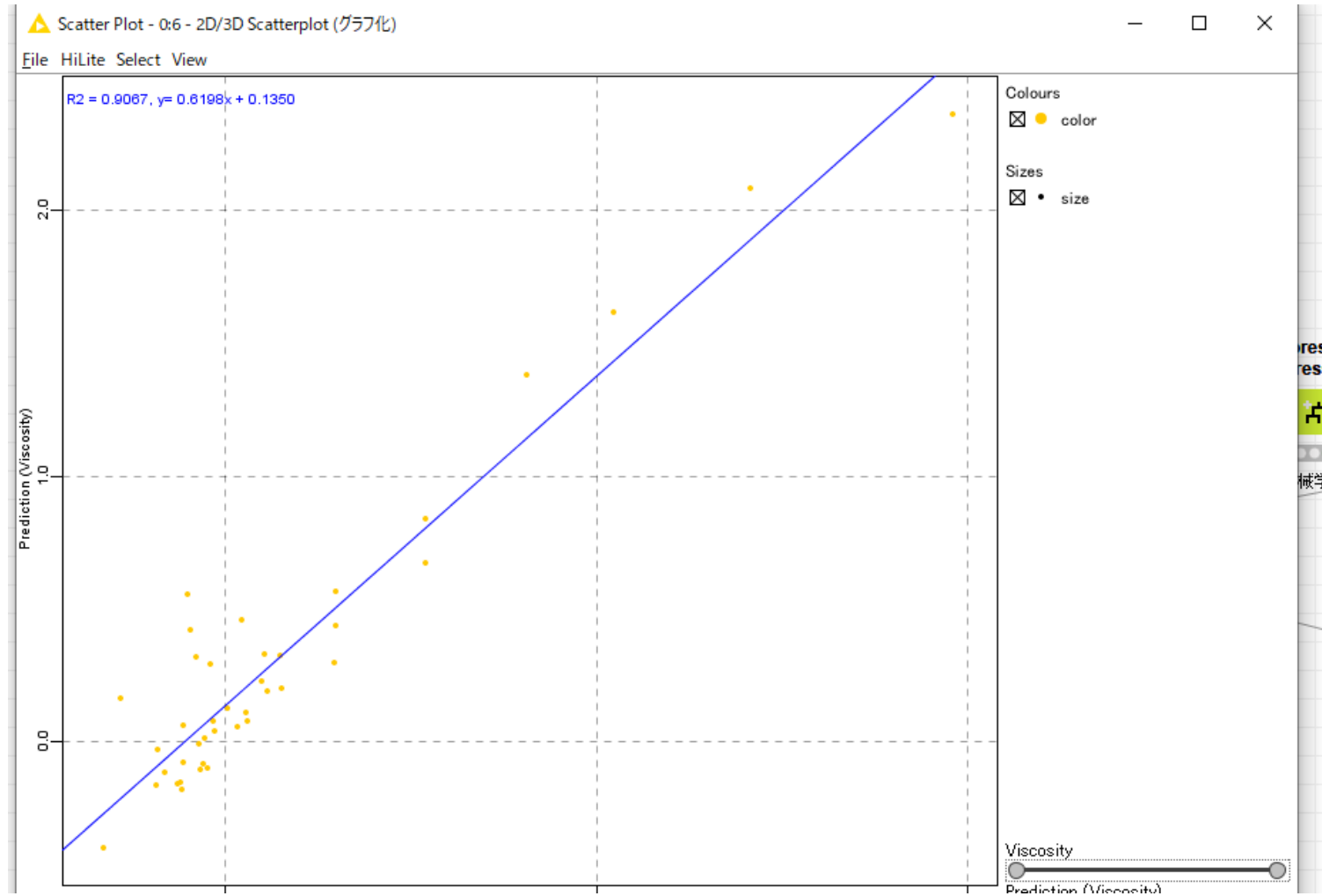
3. この数式になっていることを確認

☐ Append Column:

☒ Replace Column: **D Viscosity**

4. 元の列を置き換える

☐ Convert to Int

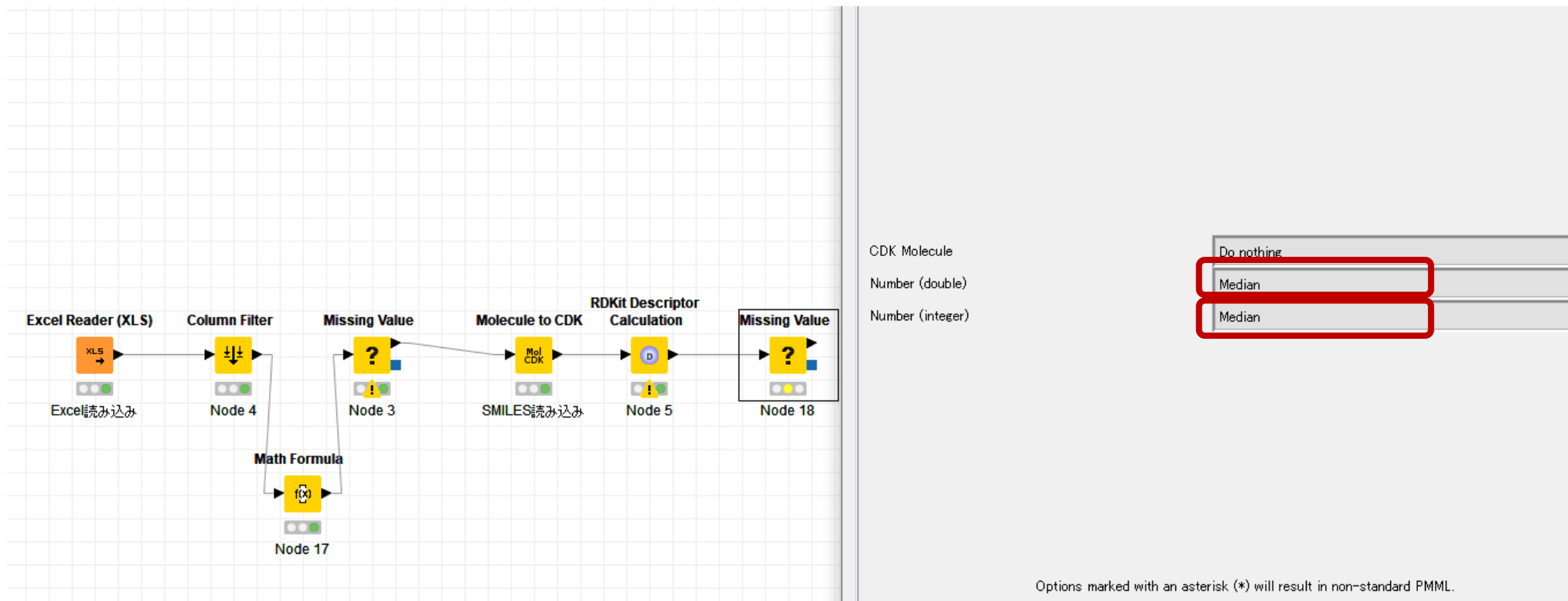


精度が向上

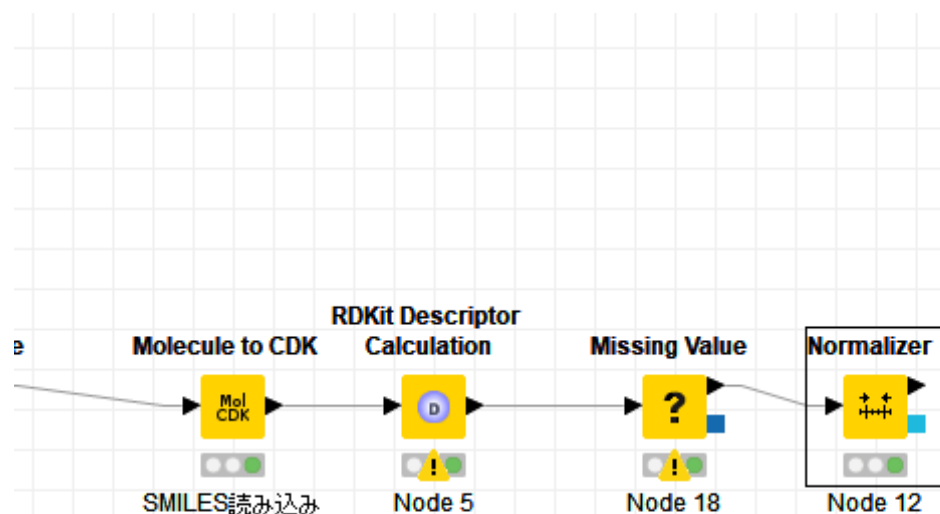


# 参考1: 線形モデル の検討

---



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



Methods | Flow Variables | Memory Policy

☒ Manual Selection ☐ Wildcard/Regex Selection

Exclude

Filter

? ID

☒ Enforce exclusion

Include

Filter

☐ Enforce inclusion

- ☐ Viscosity
- ☐ SlogP
- ☐ SMR
- ☐ LabuteASA
- ☐ TPSA
- ☐ AMW
- ☐ ExactMW
- ☐ NumLipinskiHBA

Settings

☐ Min-Max Normalization

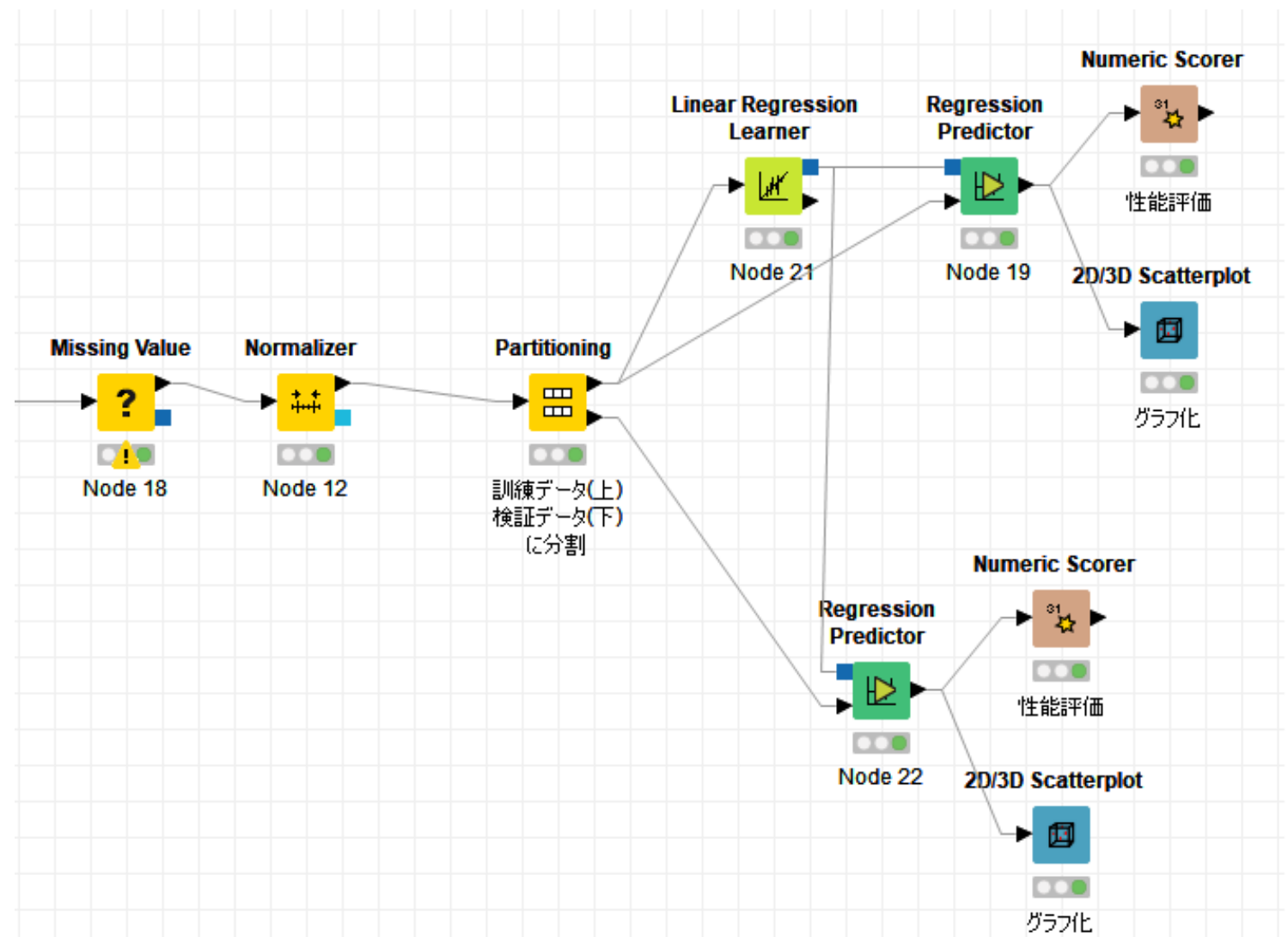
☒ Z-Score Normalization (Gaussian)

☐ Normalization by Decimal Scaling

Min: 0.0

Max: 1.0

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選んでみる

Dialog - 0:21 - Linear Regression Learner

File

Settings Flow Variables Memory Policy

Target

D Viscosity

Values

☒ Manual Selection ☐ Wildcard/Regex Selection

Exclude

Filter

- D NumHBA
- D NumAmideBonds
- D NumHeteroAtoms
- D NumHeavyAtoms
- D NumAtoms
- D NumStereocenters
- D NumUnspecifiedStereocenters
- D NumRings
- D NumAromaticRings
- D NumSaturatedRings
- D NumAliphaticRings

☒ Enforce exclusion

>

>>

<

<<

Include

Filter

- D SlogP
- D SMR
- D LabuteASA
- D TPSA
- D AMW
- D ExactMW
- D NumLipinskiHBA
- D NumLipinskiHBD
- D NumRotatableBonds
- D NumHBD

☐ Enforce inclusion

説明変数の選択が悪いと、上手くいかない場合もある

navigation view

Coefficients and Statistics" - Rows: 6

Spec - Columns: 5

P

	<b>S</b> Variable	<b>D</b> Coeff.	<b>D</b> Std. Err.	<b>D</b> 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

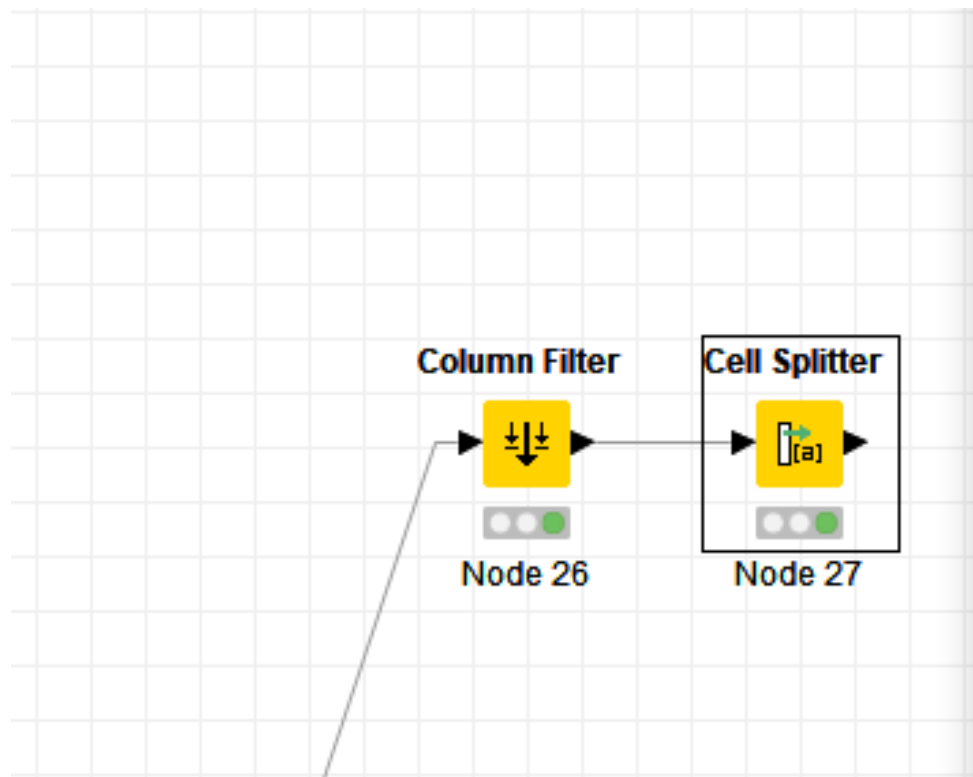
---

Table: default    Rows: 18/0    Spec: Columns: 2

Row ID	<b>S</b> SMILES	<b>S</b> 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!!





File

Settings

Flow Variables

Memory Policy

Column to split

Select a column:

S

Melting temperature



☐ Remove

Settings

Enter a delimiter:

/



Use ¥ as escape character

Enter a quotation character:

(leave empty for none)



Remove leading and trailing white space chars (trim)

Output



As list



As set (remove duplicates)



As new col



Split input column name for output column names

# Cell splitter

Output Table - 0:27 - Cell Splitter

File Hilite Navigation View

Table "default" - Rows: 1370 Spec - Columns: 4 Properties Flow Variables

Row ID	[S] SMILES	[S] Melting ...	[D] Melting temperature_Arr[0]	[D] Melting temperature_Arr[1]
Row0	N	-77.73	-77.73	?
Row1	C#C	-80.8	-80.8	?
Row2	O=P(O)(O)O...	187	187	?
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	?

Cell Splitter

Node 27

# 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