

# KNIMEを使った 材料探索 基本操作(2)

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

訓練、検証データセットの作成

結果の定量評価

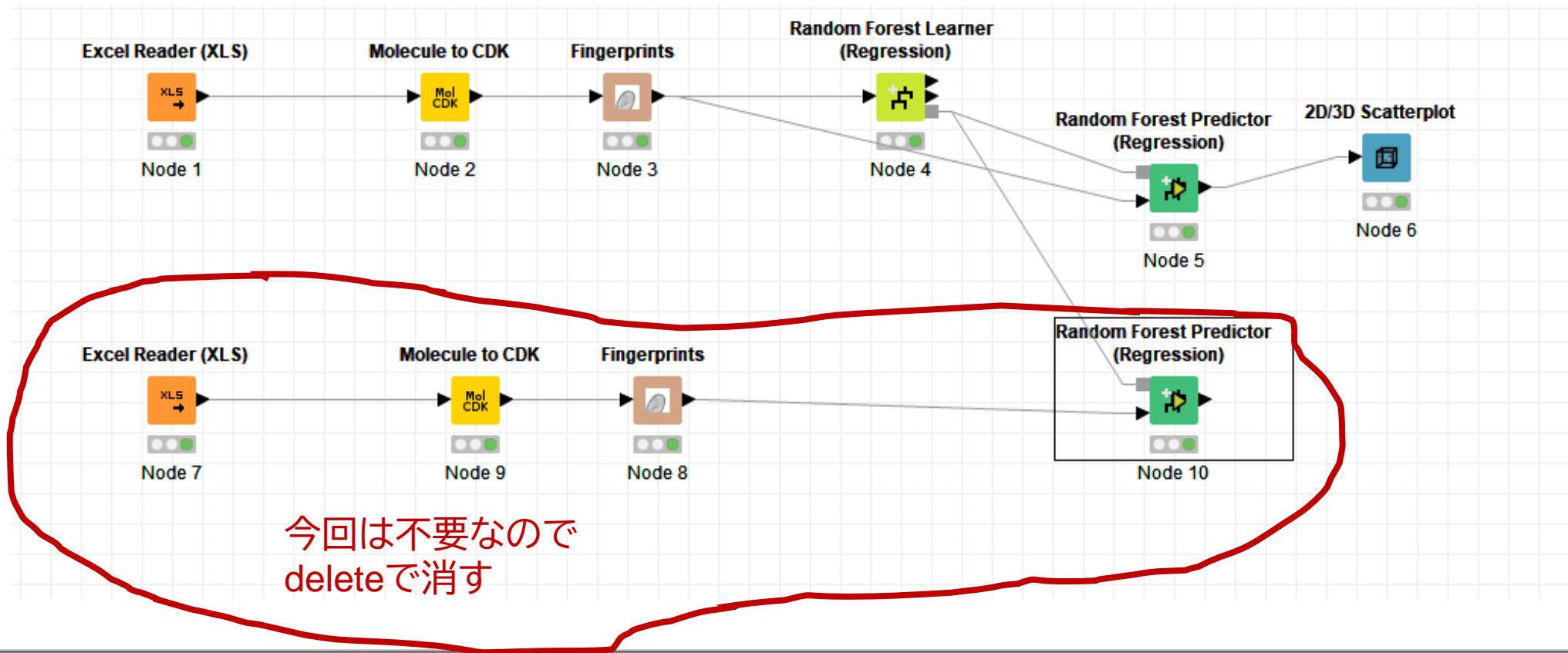
正規化

各種モデルの利用

# 訓練・検証データセットの作成

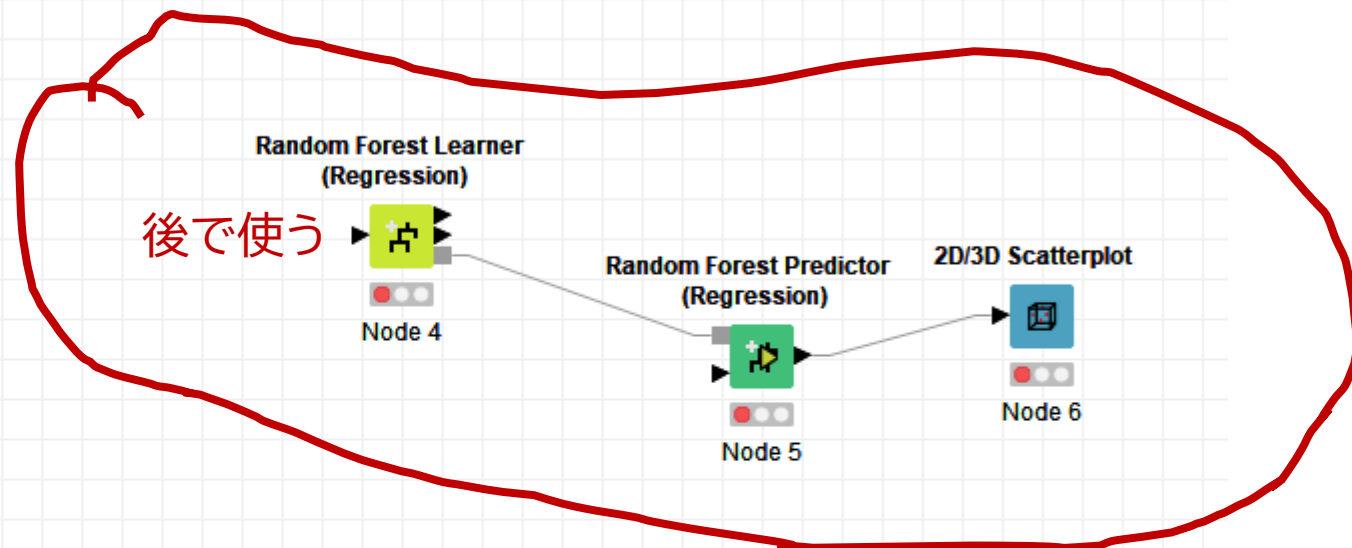
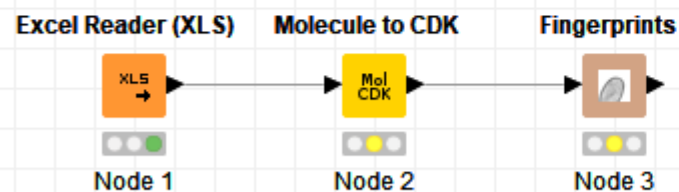
- 何故分けるのか?
  - Googleや書籍などで調べて下さい
  - 試験 検証 データ 分割 理由





前回の画面

Deleteでノード間の結合は切れます

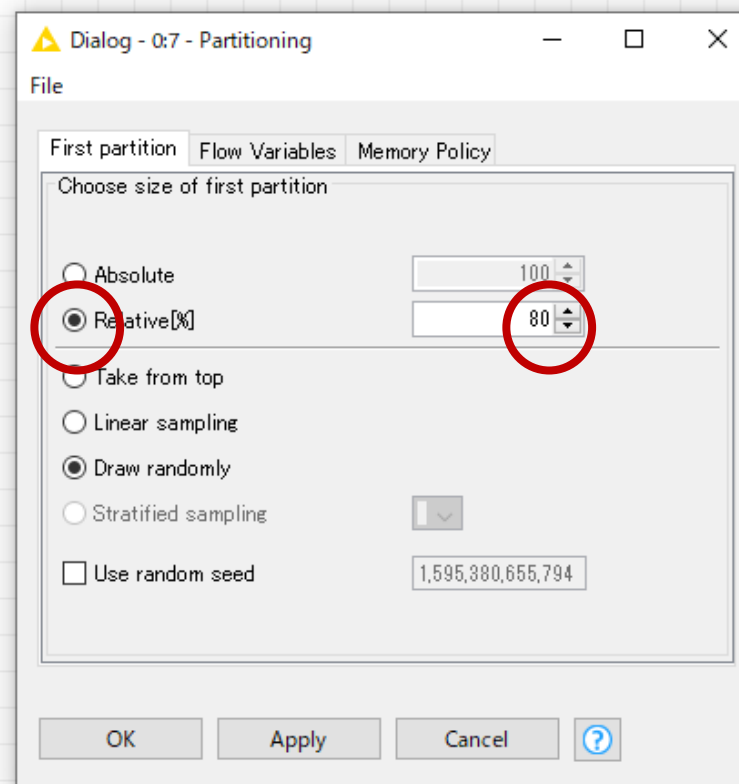
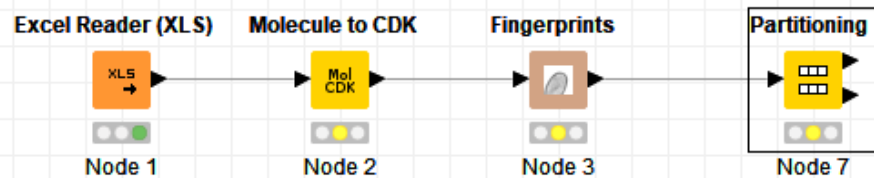


# ノードをつなぎ替える

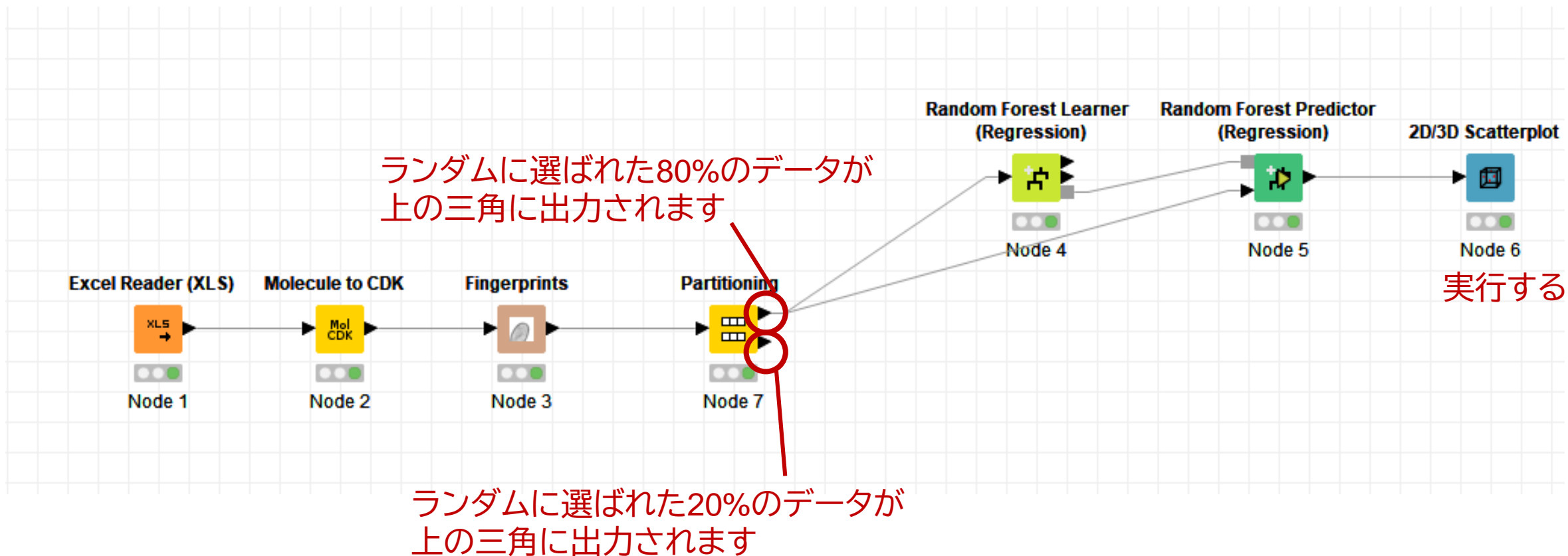
The screenshot displays the KNIME software interface. On the left, the 'Node Repository' pane is visible, showing a tree structure of nodes. The search bar at the top of the repository contains the text 'parti'. Under the 'Manipulation' category, the 'Transform' sub-category is expanded, and the 'Partitioning' node is highlighted. A red circle is drawn around the 'Partitioning' node, and a red arrow points from it towards the workflow area on the right. The workflow area shows a sequence of three nodes: 'Excel Reader (XLS)' (Node 1), 'Molecule to CDK' (Node 2), and 'Fingerprints' (Node 3). The nodes are connected by arrows, indicating a sequential flow. The 'Excel Reader (XLS)' node is orange, 'Molecule to CDK' is yellow, and 'Fingerprints' is brown. The background of the workflow area is a light gray grid.

Partitioningを設置

データの80%を訓練  
20%を検証用に  
ランダム抽出する、という指示

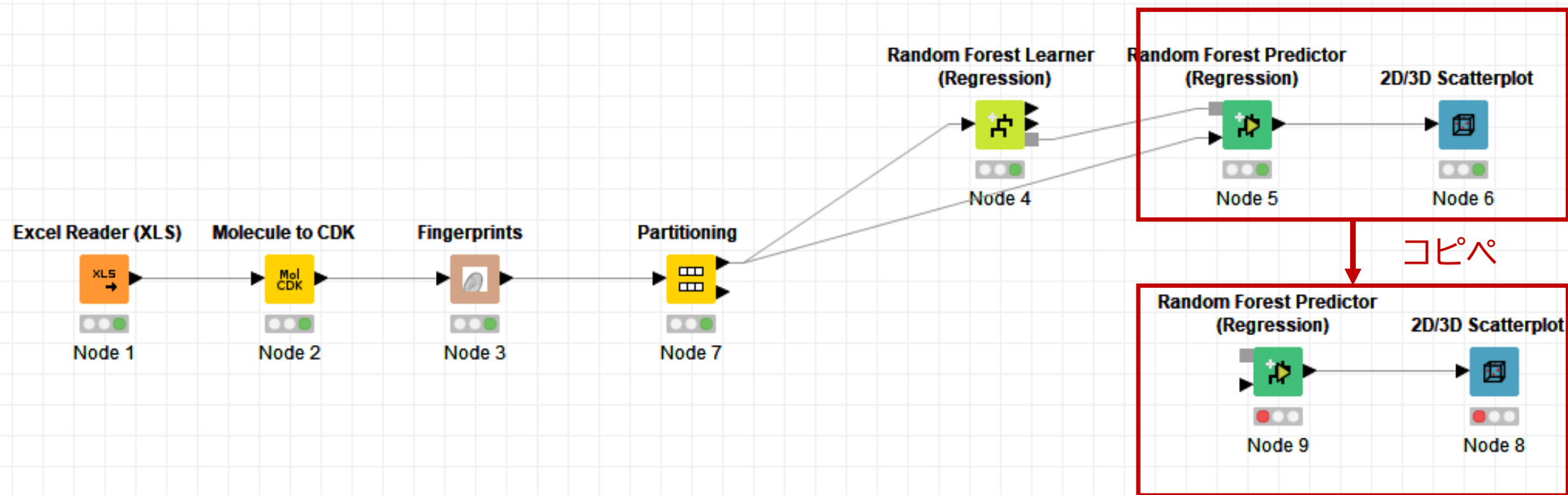


Partitioning  
を設定

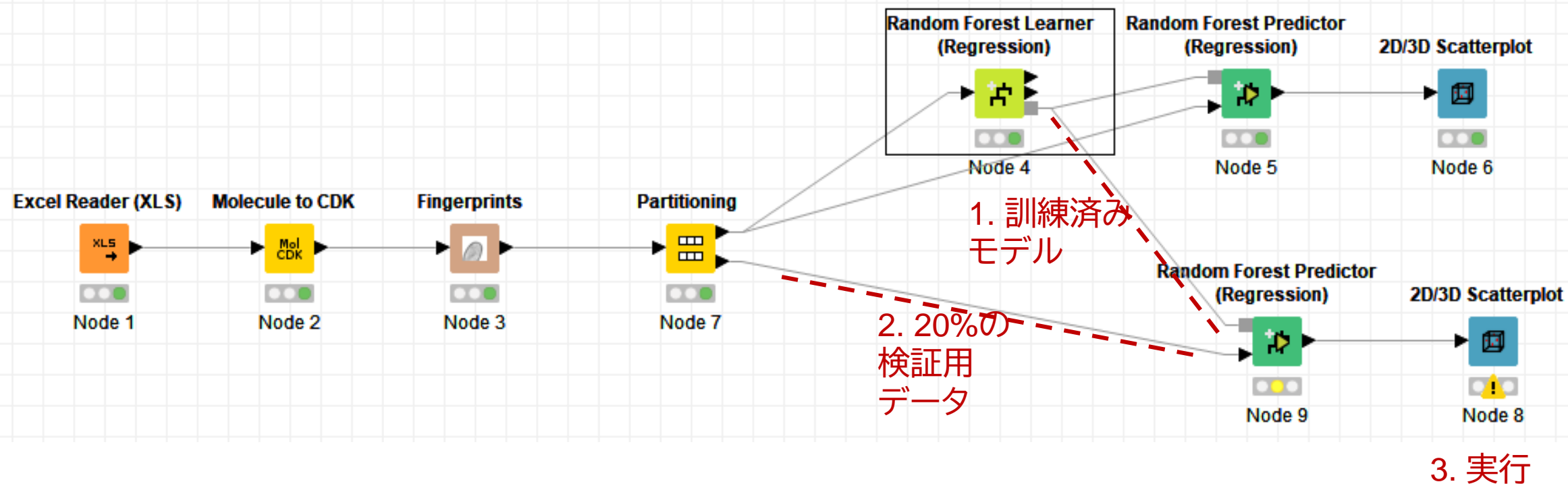


# 訓練データを学習

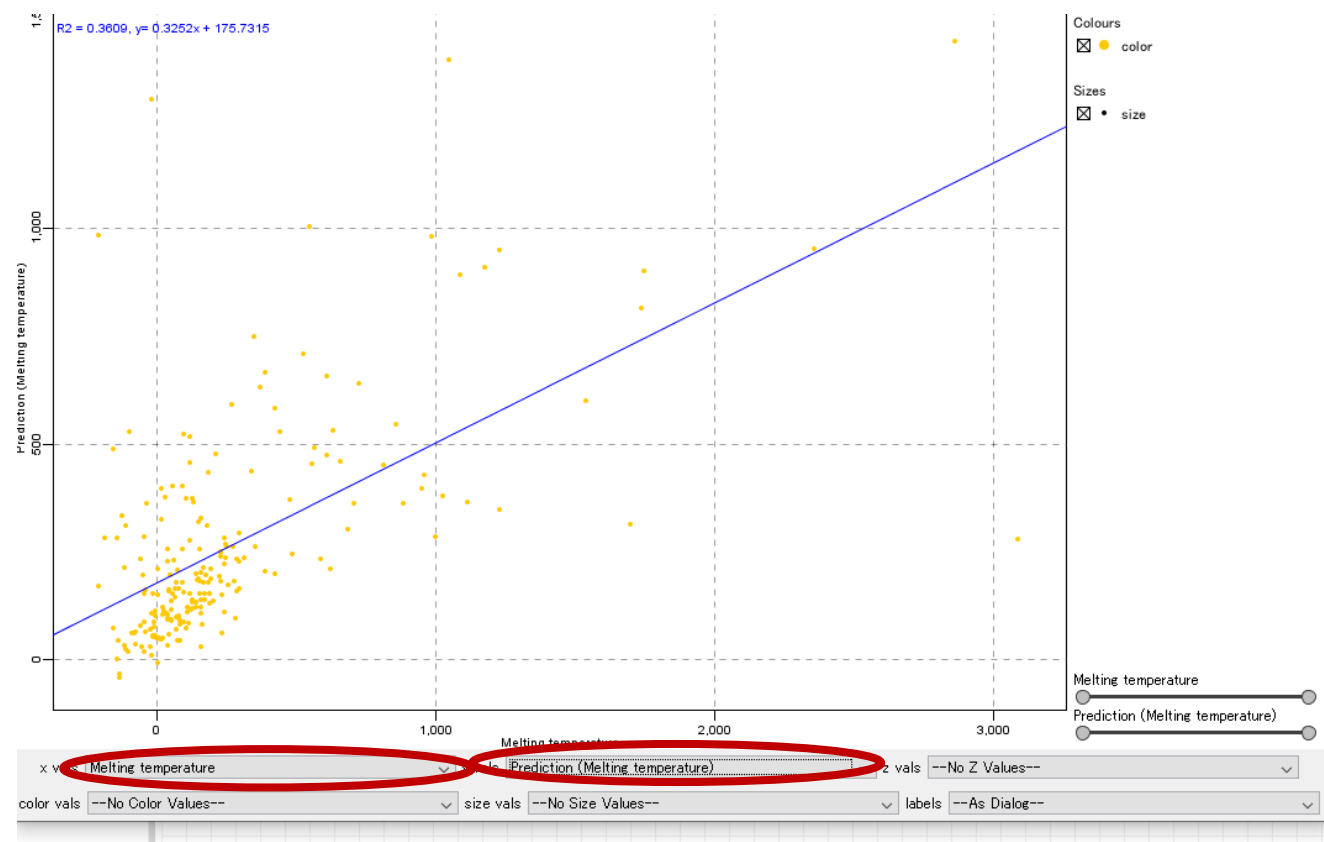




# 検証用Predictorを準備



繋ぐ



右クリック  
→ View: scatter plot

結果確認

# 結果の定量評価

$R^2$ , MAE, MSE, ...

意味は他所で調べて下さい

0422refindex  
0422test  
0704\_cheminfo\_basic  
0921miの会作業  
2019003pythonTest  
2019003pythonTest\_  
20190819ion\_database  
20190819ion\_database 1  
20190819solarcell  
20190828pythonTest  
20190829ion  
20190829ion 1  
20190829loopTest  
20190830ion  
20190830ion2  
20190831heatmap  
20190904PrepRandomMolecules

Repository

score

Analytics

Mining

Scoring

Scorer

Numeric Scorer

Entropy Scorer

Scorer (JavaScript)

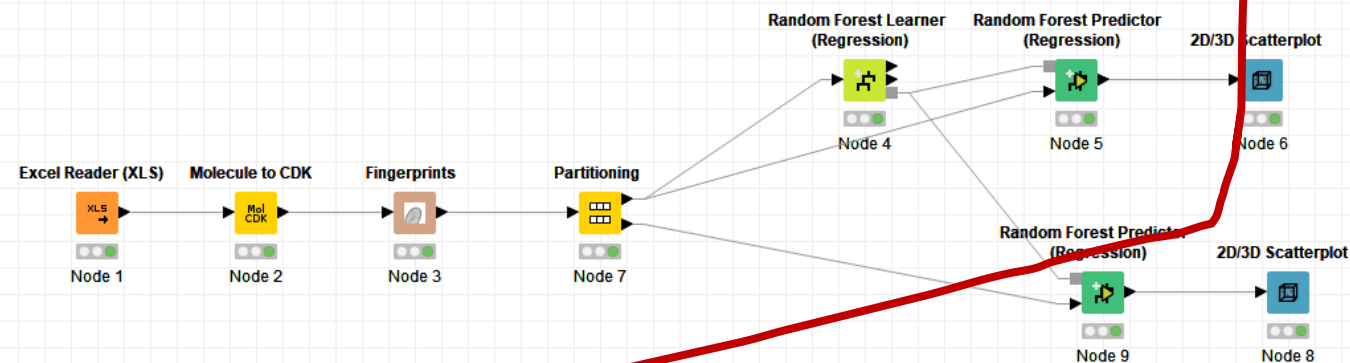
KNIME Labs

H2O Machine Learning

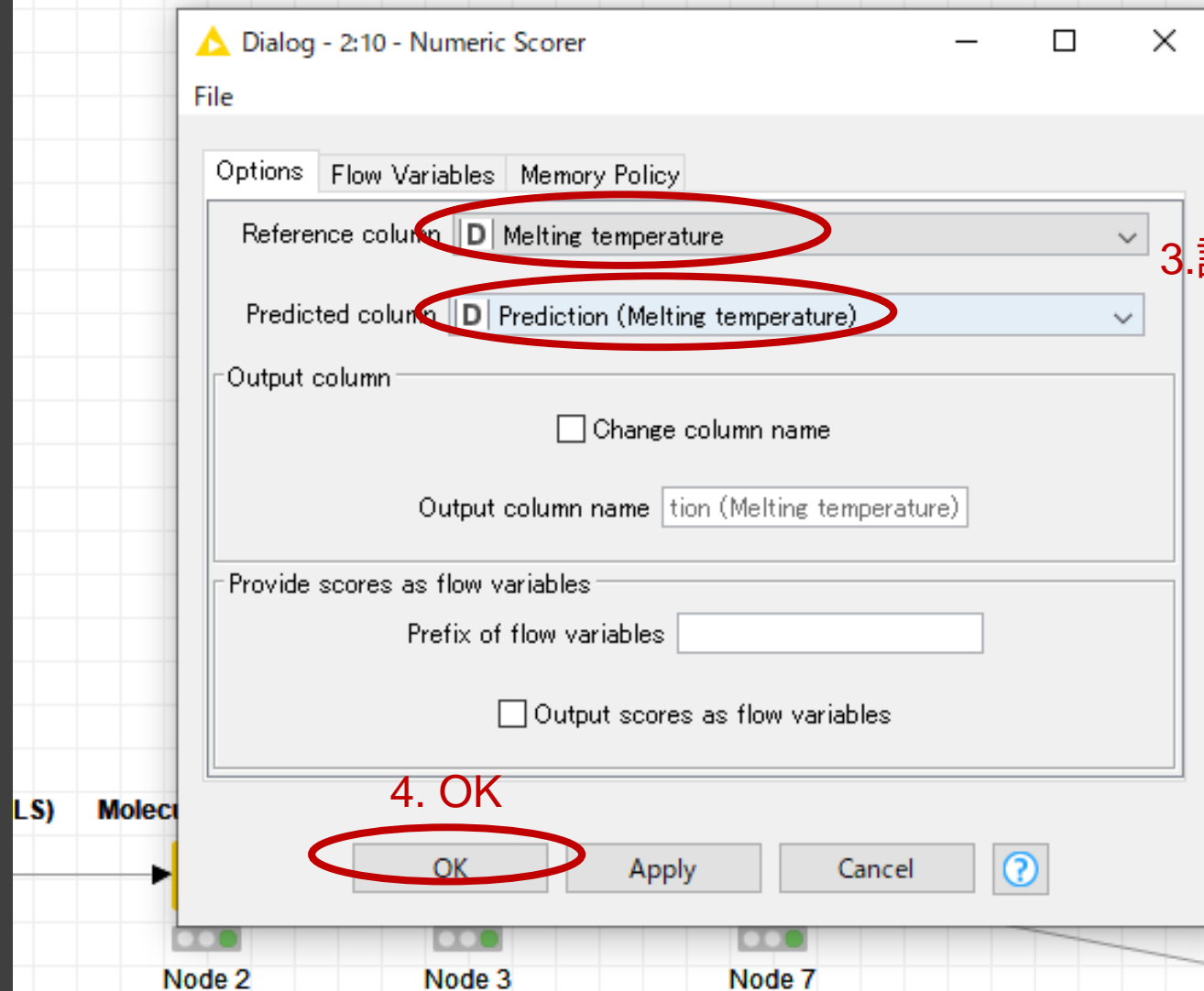
Scoring

H2O Binomial Scorer

H2O Multinomial Scorer



# Numeric Scorerを配置

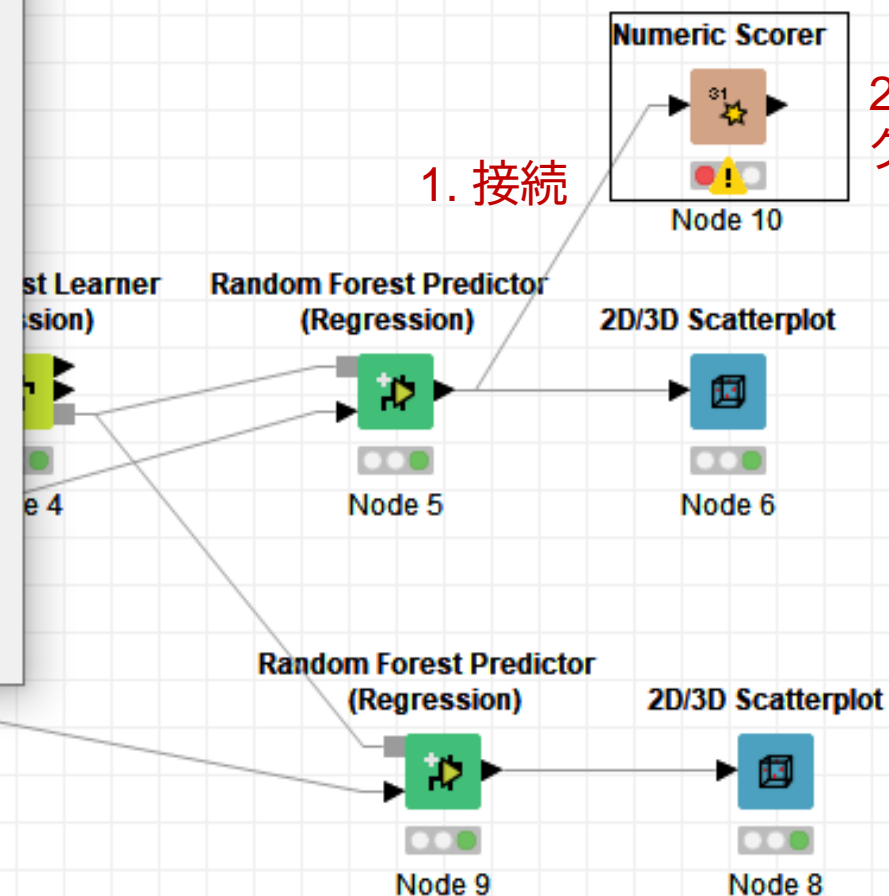


3. 評価項目を設定

1. 接続

2. ダブル  
クリック

5. 実行



# Statistics - 2:10 - Numeric Scorer

File Hilite Navigation View

Table "Scores" - Rows: 5 Spec - Column: 1 Properties Flow Variables

Row ID	Predicti...
R^2	0.801
mean absolute error	111.926
mean squared error	43,819.422
root mean squared deviation	209.331
mean signed difference	-4.919

結果が出てきました!

右クリック  
→Statistics

Numeric Scorer



Node 10

Random Forest Learner  
(Regression)



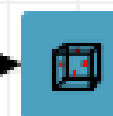
Node 4

Random Forest Predictor  
(Regression)



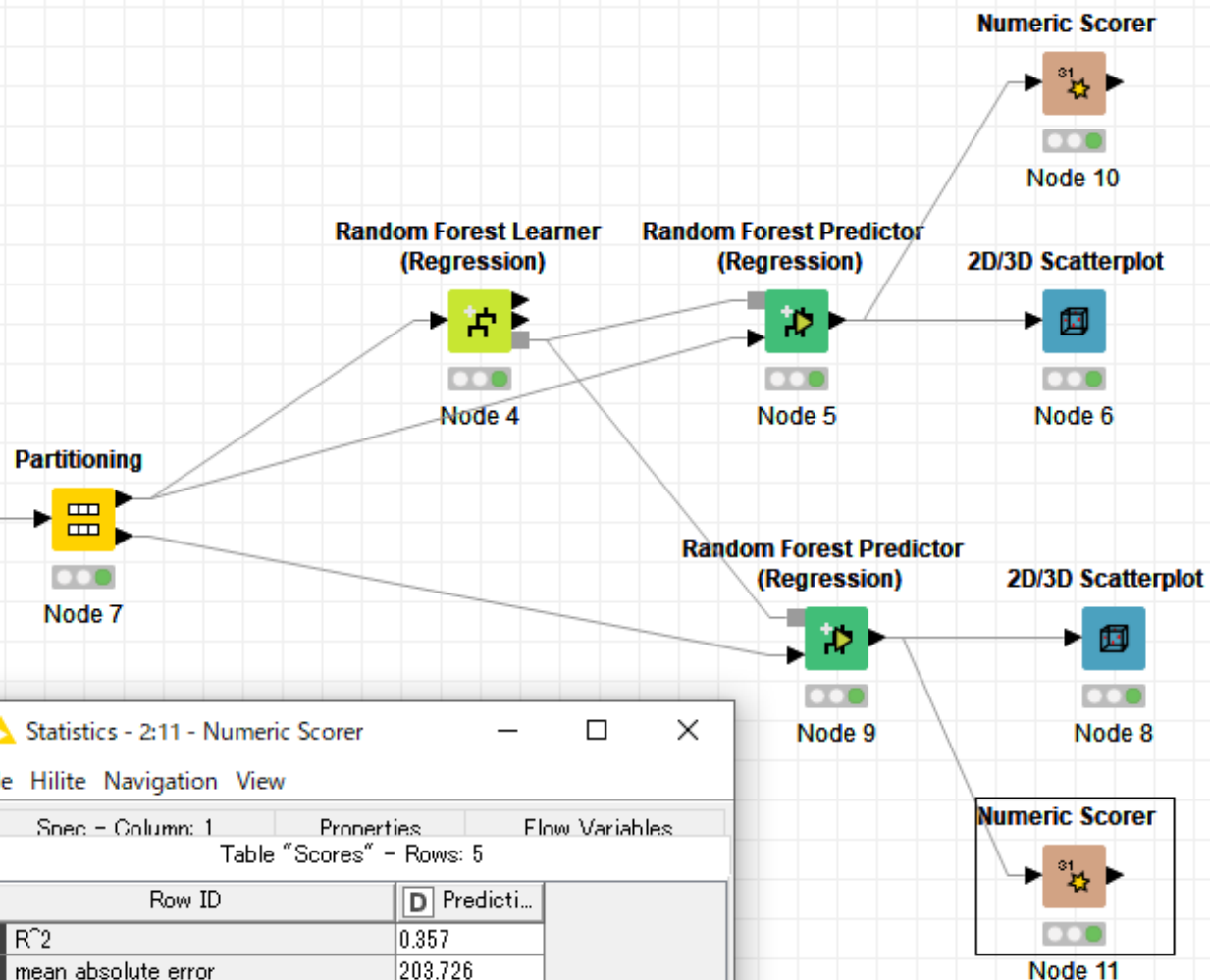
Node 5

2D/3D Scatterplot



Node 6

Numeric  
scorerを  
コピペし、  
検証データの  
結果も見る



Statistics - 2:11 - Numeric Scorer

File Hilite Navigation View

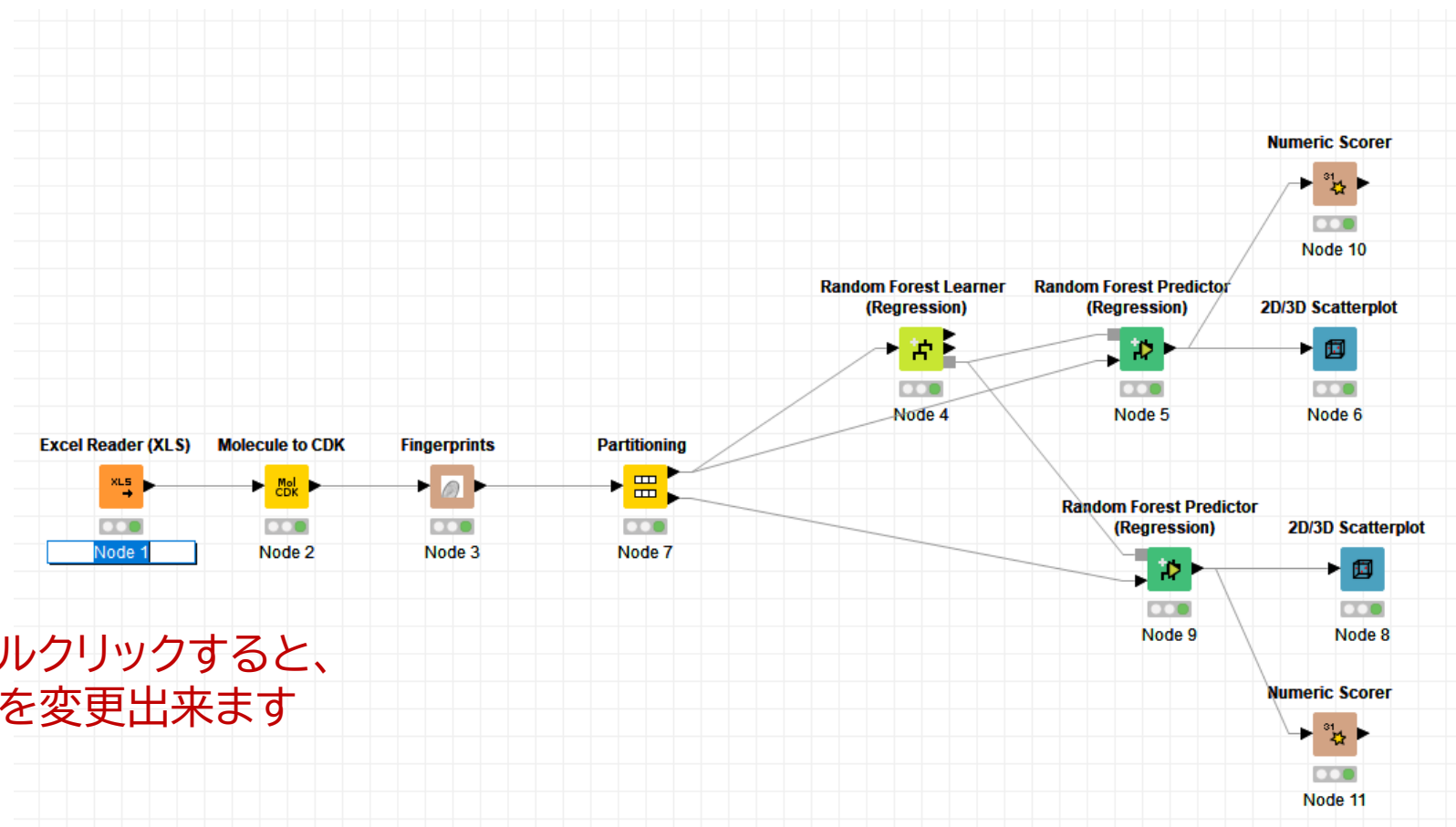
Spec: - Column: 1 Properties Flow Variables

Table "Scores" - Rows: 5

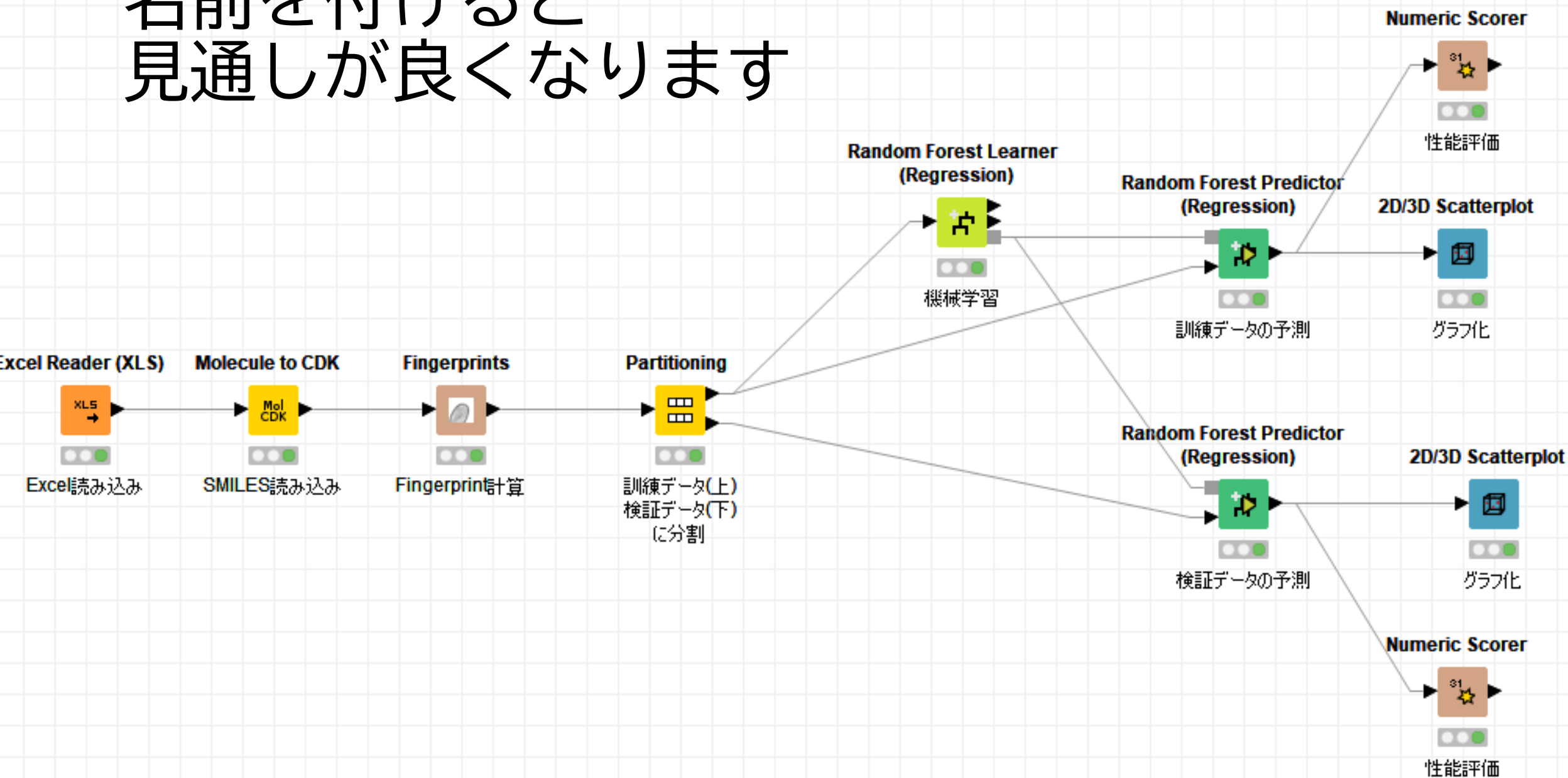
Row ID	Predicti...
R <sup>2</sup>	0.357
mean absolute error	203.726
mean squared error	138,910.585
root mean squared deviation	372.707
mean signed difference	1.734



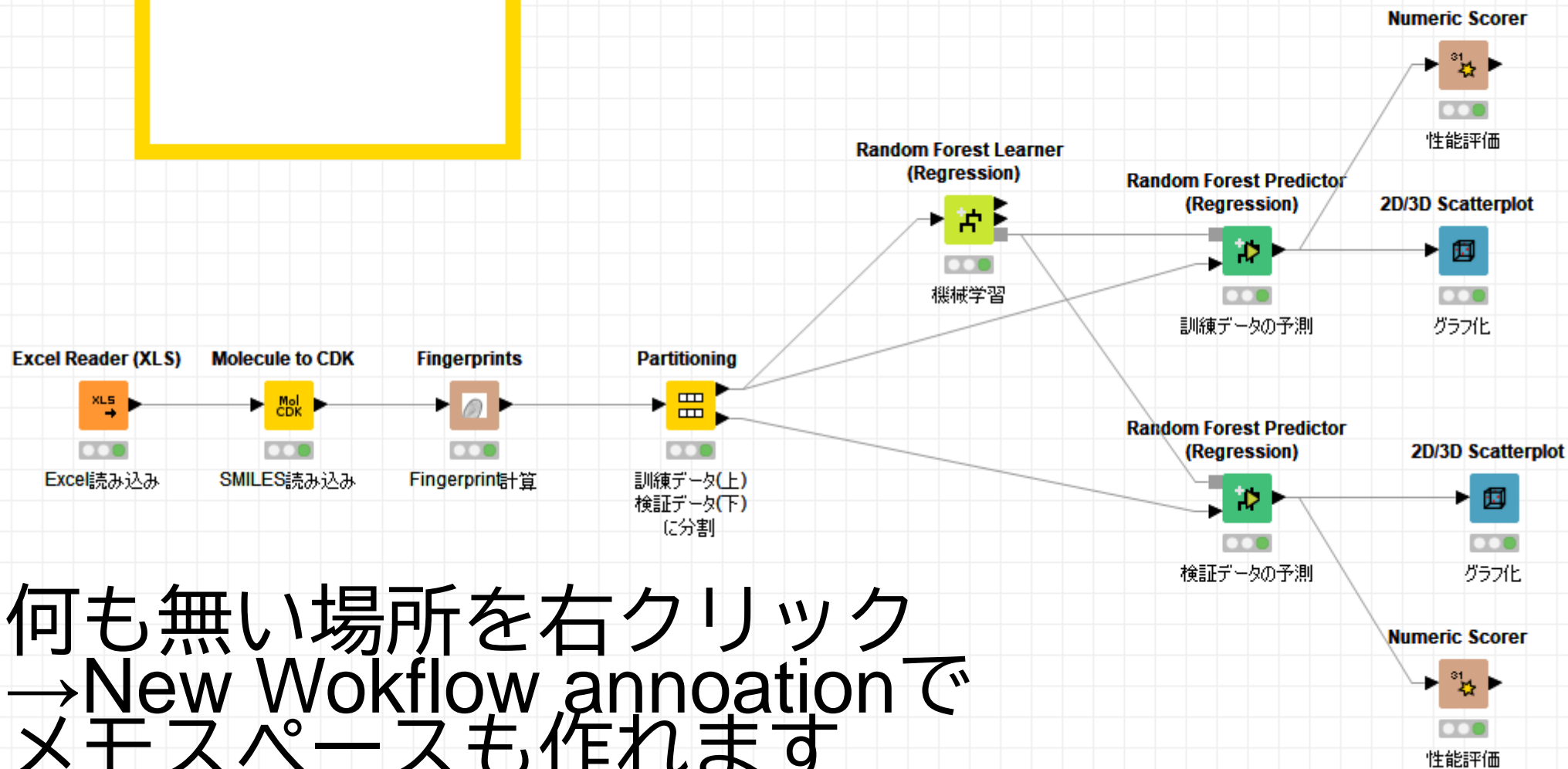
# 全体図



# 名前を付けると 見通しが良くなります



Wikipediaのデータを処理する



何も無い場所を右クリック  
→ New Workflow annotationで  
メモスペースも作れます

# 正規化

いわゆる標準得点 (z-score)を計算します  
操作の理由は他所で調べて下さい

- 20190819ion\_database
- 20190819ion\_database 1
- 20190819solarcell
- 20190828pythonTest
- 20190829ion
- 20190829ion 1
- 20190829loopTest
- 20190830ion
- 20190830ion2
- 20190831heatmap
- 20190904PrepRandomMolecules

Repository

normal

Manipulation

Column

Transform

Denormalizer

Normalizer

Excel Reader (XLS)



Excel読み込み

Molecule to CDK



SMILES読み込み

Fingerprints



Fingerprint計算

Partitioning



訓練データ(上)  
検証データ(下)  
に分割

## Normalizerの設置

Excel Reader (XLS)



Excel読み込み

Molecule to CDK



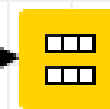
SMILES読み込み

Fingerprints



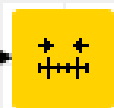
Fingerprint計算

Partitioni



訓練データ  
検証データ  
に分割

Normalizer



Node 12

ノードのつ  
なぎ替え

# Normalizer 設定

Dialog - 2:12 - Normalizer

File

Methods | Flow Variables | Memory Policy

☒ Manual Selection ☐ Wildcard/Regex Selection

Exclude

Filter

☒ ID

Normalizeしたくない  
パラメータはこっち

☒ Enforce exclusion

Include

Filter

☒ Melting temperature

Normalizeしたい  
パラメータはこっち

☐ Enforce inclusion

> >> < <<

Settings

☐ Min-Max Normalization

☒ Z-Score Normalization (Gaussian)

☐ Normalization by Decimal Scaling

Min: 0.0

Max: 1.0

Normalizeの方法を選べます

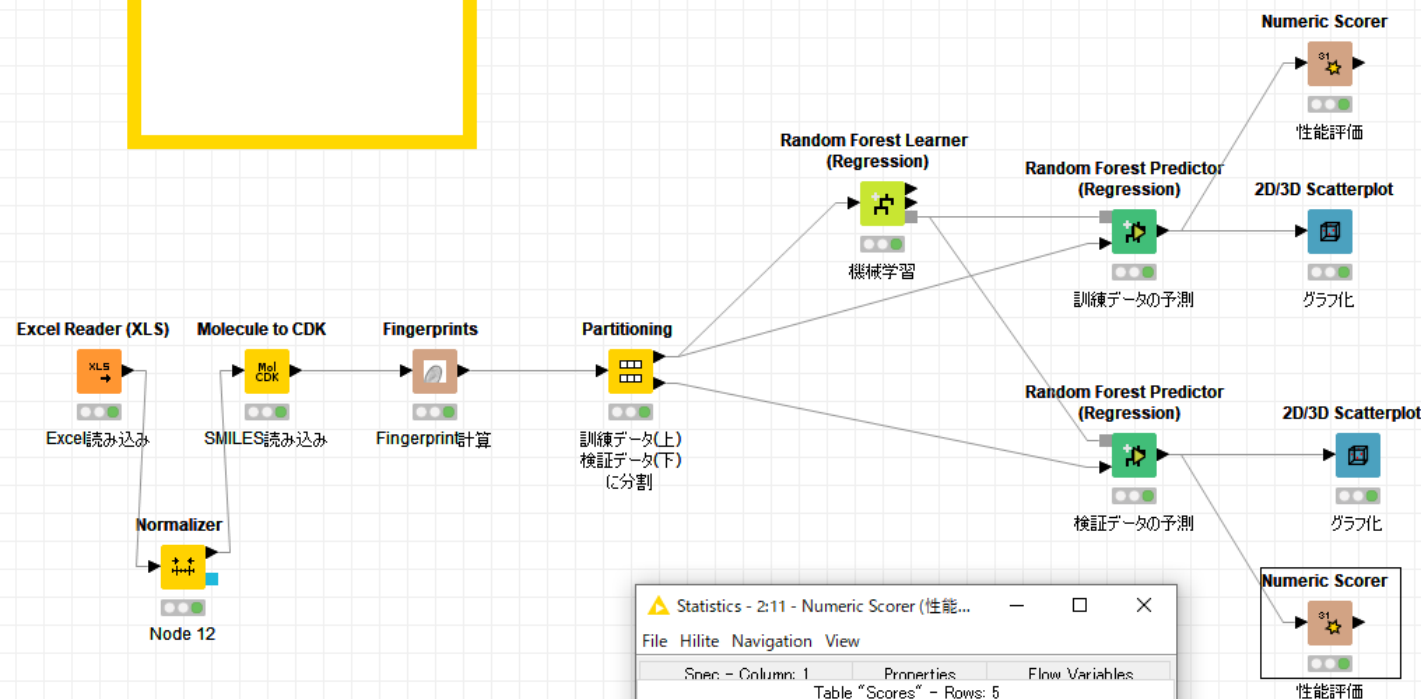
Row ID	ID	SMILES	Melting ...
Row0	1	[Cu]=S	0.518
Row1	2	c1cc2ccc3cc...	-0.299
Row2	3	O1[Fe]2O[F...	2.736
Row3	4	O=C1NC(=O...	-0.026
Row4	5	P#[Y]	-0.12
Row5	6	C1=CC=C(C...	0.07
Row6	7	ClC(C)C(=O...	-0.34
Row7	8	FC(F)F	-0.88
Row8	9	O=[N+](O-]	-0.318
Row9	10	CCC[C@@H]	-0.365
Row10	11	c1ccc2c(c1)	-0.613
Row11	12	O=C(O)[C@]	0.059
Row12	13	F[Co](F)F	1.43
Row13	14	[Cs+].[I-]	0.8
Row14	15	C1CCC(CC1...	-0.54
Row15	16	O=C2c3c(O[...	-0.013
Row16	17	BrC(F)(F)F	-0.007

正規化  
されました

Normalizer  
右クリック  
→  
Normalized  
table



Wikipediaのデータを処理する



Statistics - 2:11 - Numeric Scorer (性能...

File Hilite Navigation View

Spec - Column: 1 Properties Flow Variables

Table "Scores" - Rows: 5

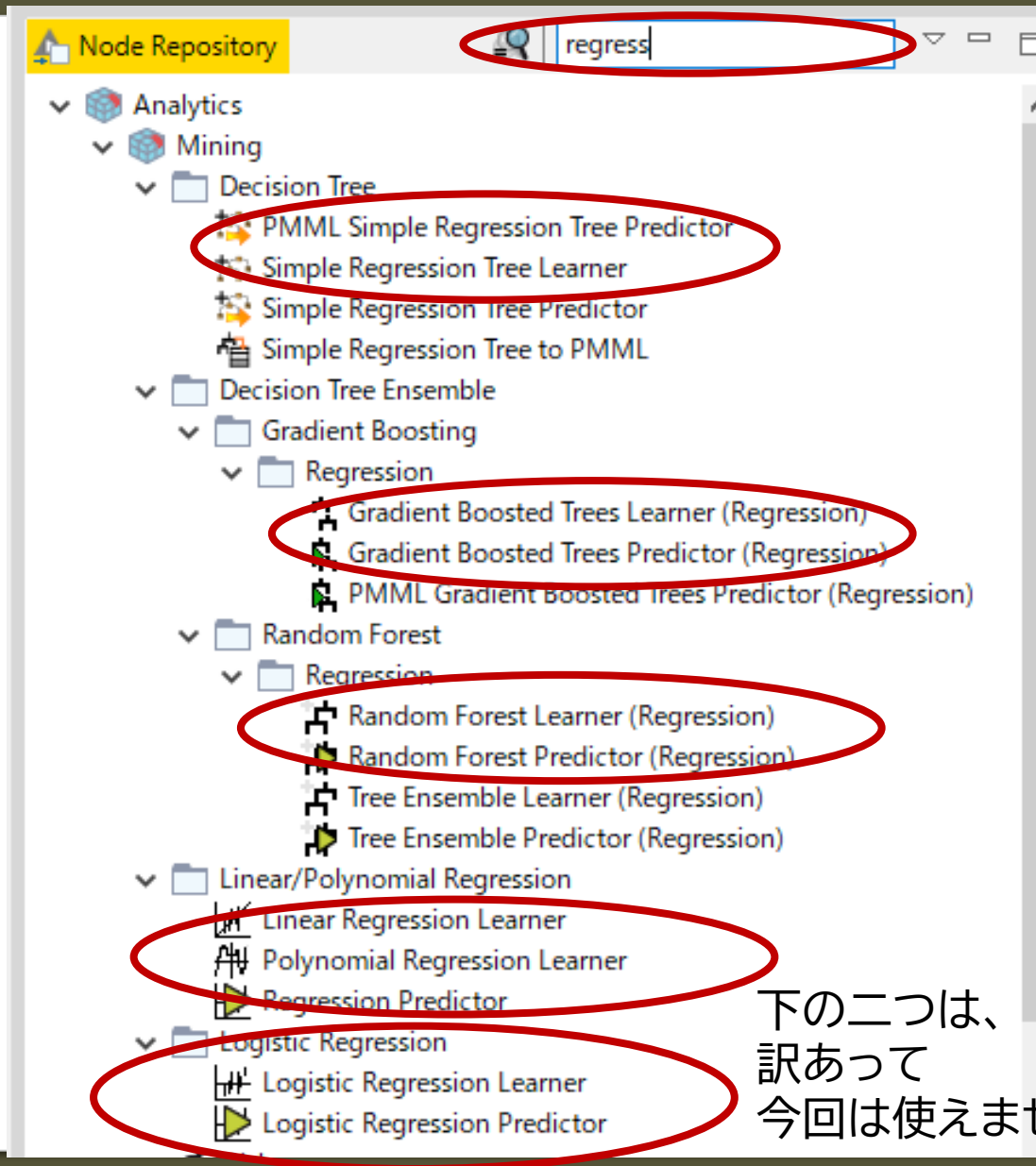
Row ID	Predicti...
R <sup>2</sup>	0.351
mean absolut...	0.472
mean squared...	0.829
root mean squ...	0.911
mean signed ...	-0.039

全て実行

Random forestは正規化不要なので  
基本的に予測性能は変わりません

# 各種モデルの利用

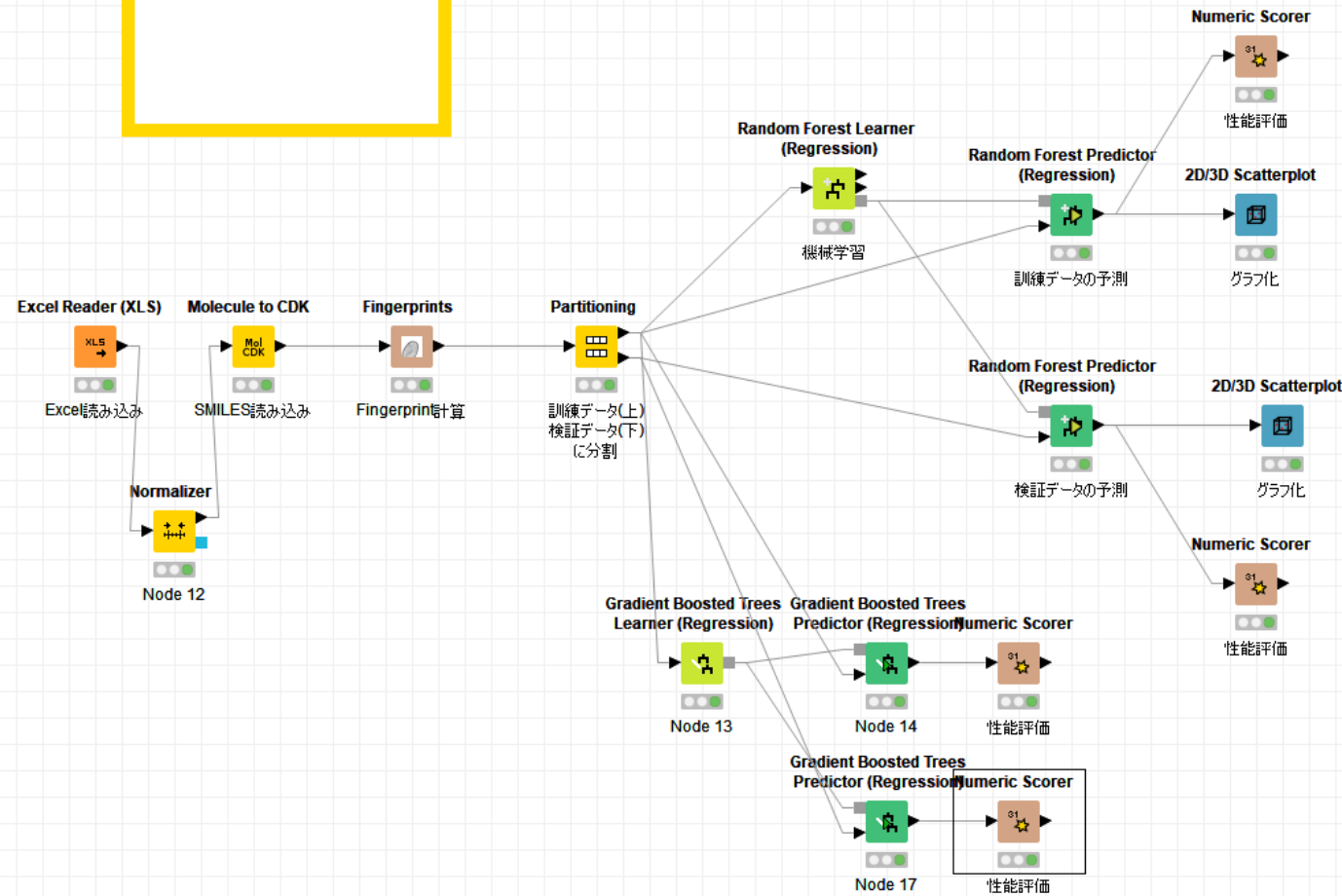
実はKNIMEの対応モデルは標準では  
あまり多くありません



利用可能な  
回帰モデル

下の二つは、  
訳あって  
今回は使えません(Fingerprint)

Wikipediaのデータを処理する



Gradient  
boostingを  
追加してみ  
た例

いわゆるハイパーパラメータ  
色々いじると  
性能が上がるかも？

# Gradient boostingの Advanced options

Options **Advanced Options** Flow Variables Memory Policy

## Tree Options

☒ Use mid point splits (only for numeric attributes)

☒ Use binary splits for nominal columns

Missing value handling

XGBoost

## Boosting Options

Alpha (percentage of the data that are not treated as outlier)

0.95

## Bagging Options

Data Sampling (Rows)

☐ Fraction of data to learn single model

1

☐ With replacement

☒ Without replacement

Attribute Sampling (Columns)

☒ All columns (no sampling)

☐ Sample (square root)

☐ Sample (linear fraction)

1

☐ Sample (absolute value)

10

Attribute Selection

☒ Use same set of attributes for entire tree

☐ Use different set of attributes for each tree node

☒ Use static random seed

1595481521713

New

今後の  
TODO

---

Fingerprint以外の方法

---

欠損データの処理

---

多彩なモデルの利用