ノンコーディングで行う機械学習勉強会

1回目：8/27, 30　機械学習概要、ソフトウェアRapidMiner Studio概要と事例デモ・実習
2回目：9/10, 11　分類1：主要なアルゴリズム説明と応用事例デモ・実習
3回目：9/24, 25　分類2：データ前処理と後処理、教師データと
テストデータの分割による分類問題の実習
4回目：10/8, 9　分類3：交差検証、最適アルゴリズム探索の実習
5回目：10/23, 25　回帰：主要なアルゴリズム説明と実習
6回目：11/5, 6　（応用）時系列データの機械学習
7回目：11/19, 20（応用）Extensionによる機能拡張と画像の分類
8回目：12/3, 5　もう一つの選択肢「Mathematica – Wolfram Engine」

今回取り扱うデータセット

天体の構成がパルサー星かどうかを判別する → True or False の**2値問題**

(1) Mean of the integrated profile
(2) Standard deviation of the integrated profile
(3) Excess kurtosis of the integrated profile
(4) Skewness of the integrated profile
(5) Mean of the DM-SNR curve
(6) Standard deviation of the DM-SNR curve
(7) Excess kurtosis of the DM-SNR curve
(8) Skewness of the DM-SNR curve

上記8つの説明変数を基に、パルサーがどうか（目的変数）の判別を行う。

→ 機械学習を行う上では**その分野の専門知識は一切なくても行える**。
（つまり、経験や勘に頼らなくてもよい。時には専門知識による常識が、機械学習の妨げになることもある）

機械学習プロセス

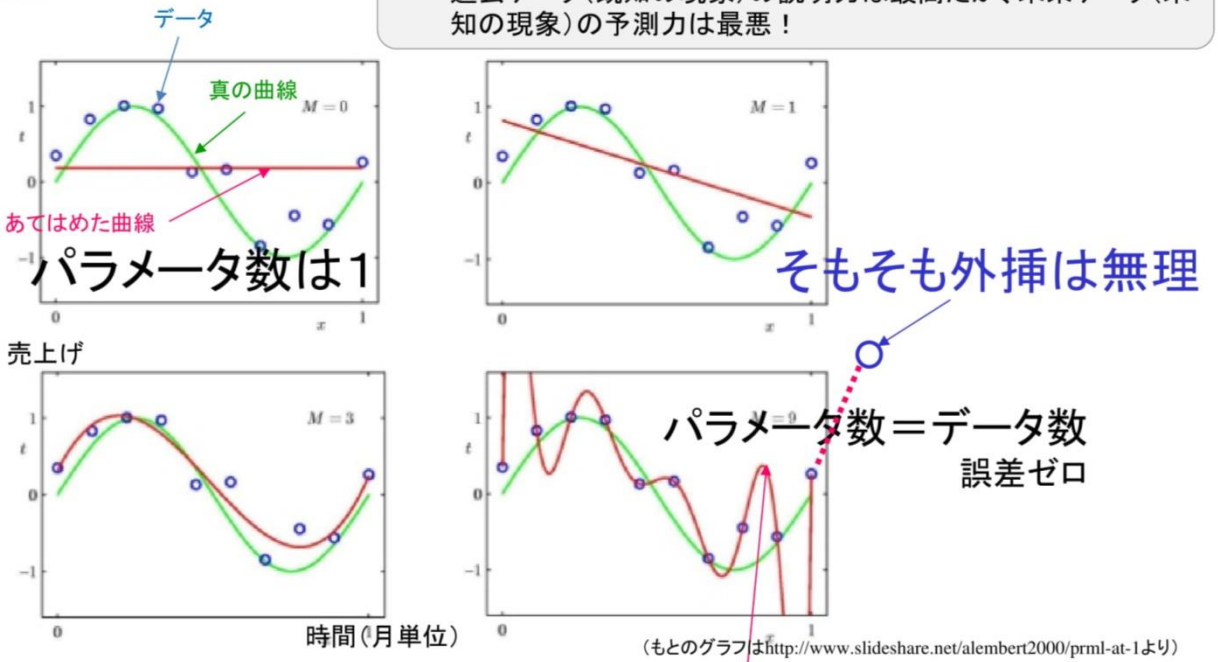

課題設定
データ取得
データ前処理
手法選択
モデル学習
モデル評価
・分析から得たいものを決定
・ログ
・アンケート
・WebAPI
etc..
・欠損値処理
・特徴量選択
・次元削除
etc..
・SVM
・決定木
・k-NN
etc..
・手法選択で選択をしたモデルを使用し学習させる
・正解率
・適合率
・再現率
・汎化性能の評価
etc..

（復習）過学習

表現能力の高いモデルには過学習（over fitting）が避けられない

- 与えられたデータにモデル（回帰曲線）が完全に一致すること
- 過去データ（既知の現象）の説明力は最高だが、未来データ（未知の現象）の予測力は最悪！

過学習

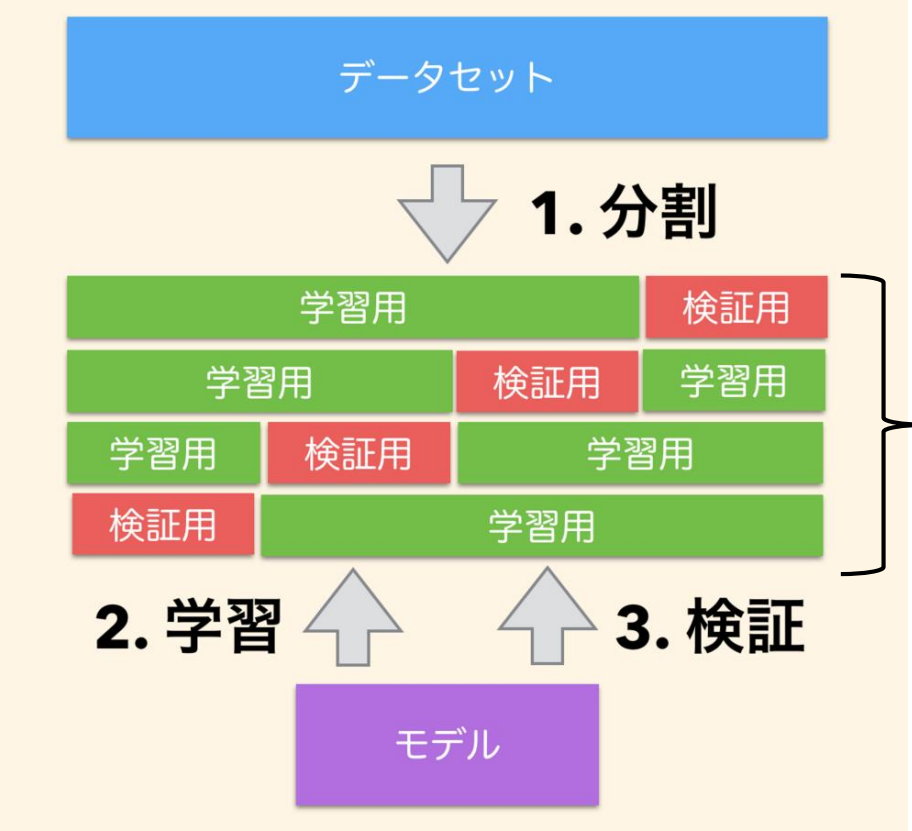


内挿の失敗：こんな予測を信じますか？

回帰曲線で起こる過学習は、モデルを複雑にすると、他の分類器でも起きうる。

過学習と実際の精度検証

交差検証（Cross Validation） ＝ k-分割交差検証



元のデータセットを
K回分割して、繰り返し
学習・検証を行う。

過学習を抑える方法

1. データ数を増やす　⇒　画像データならデータの水増し

2. 良い特徴量を設計する（良いモデルを選択する）

3. ドロップアウト　⇒　学習データを一部破棄する　（ディープラーニングでは定番）

4. 正則化（Regularization）

L2 正則化：

正則化項：特徴量の重みを 0 にしようとする力

負の対数尤度：データを正しく分類しようとする力

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1).$$

(https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression より引用)

コスト：両者のバランスを決めるパラメータ

L1 正則化

疎な解を求めたいときに利用する

L2 正則化 (特徴量の二乗和)：

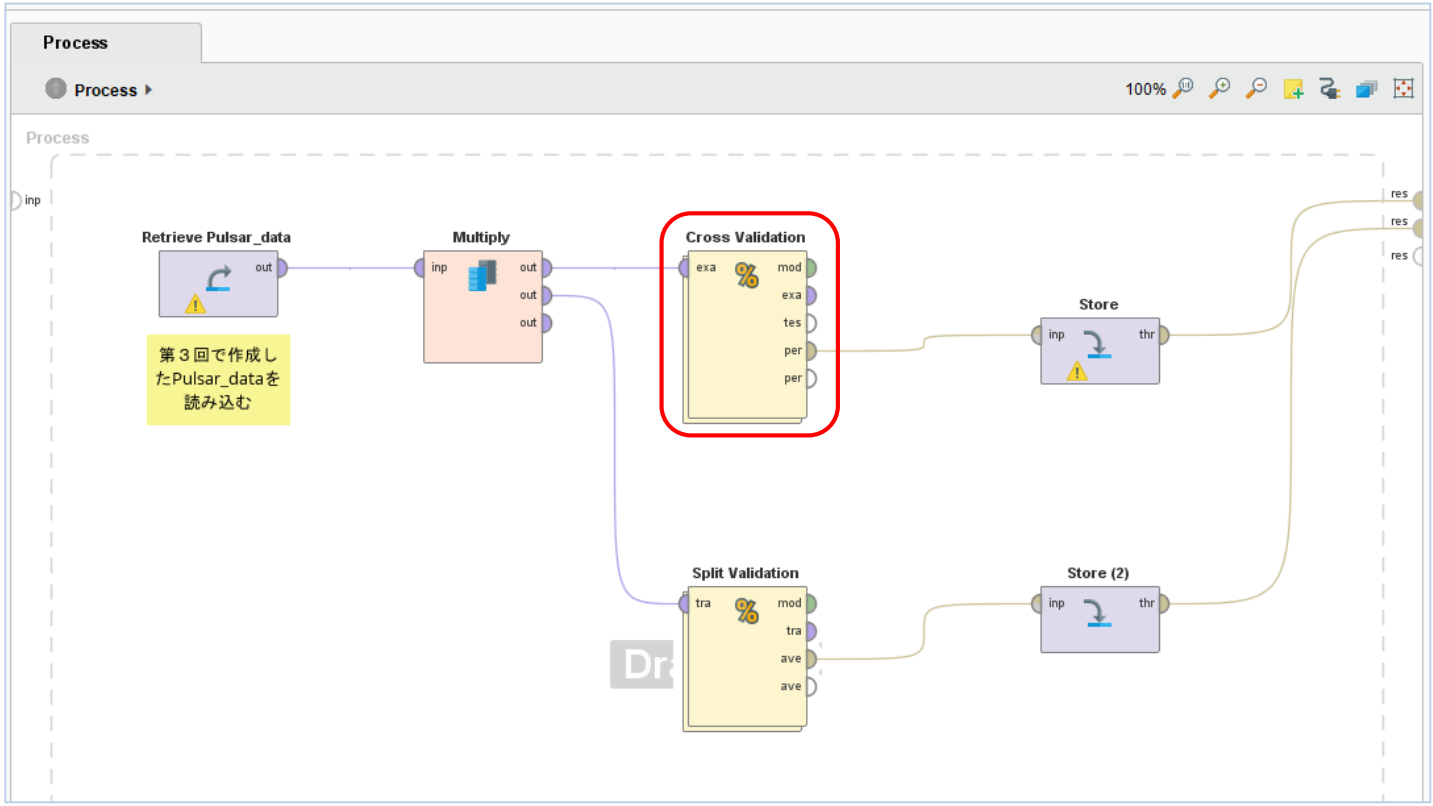
$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1).$$

L1 正則化 (特徴量の絶対値の和)：

$$\min_{w,c} \|w\|_1 + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1).$$

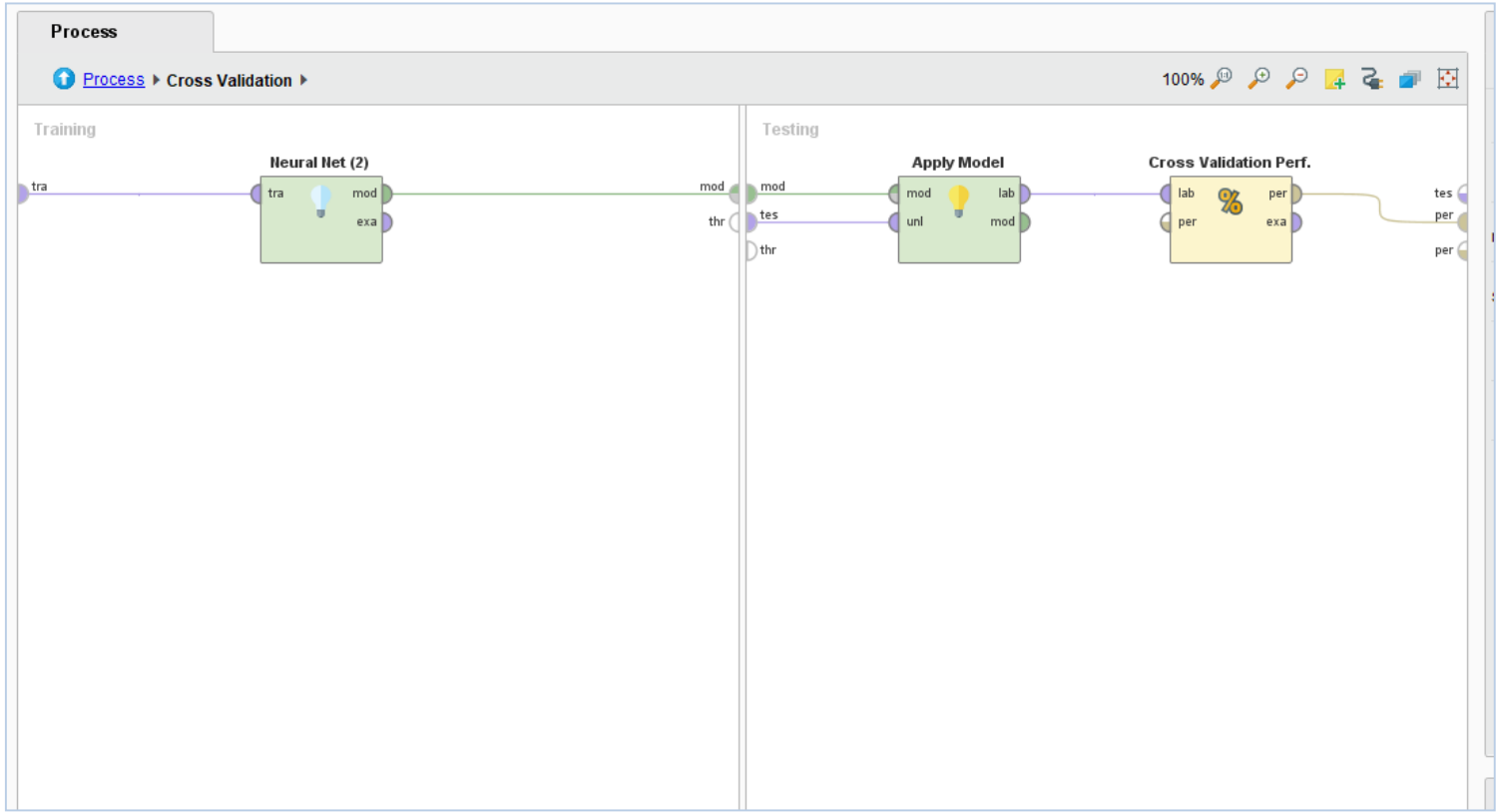
機械学習デモ＋実習

1．交差検証（NN, RF）



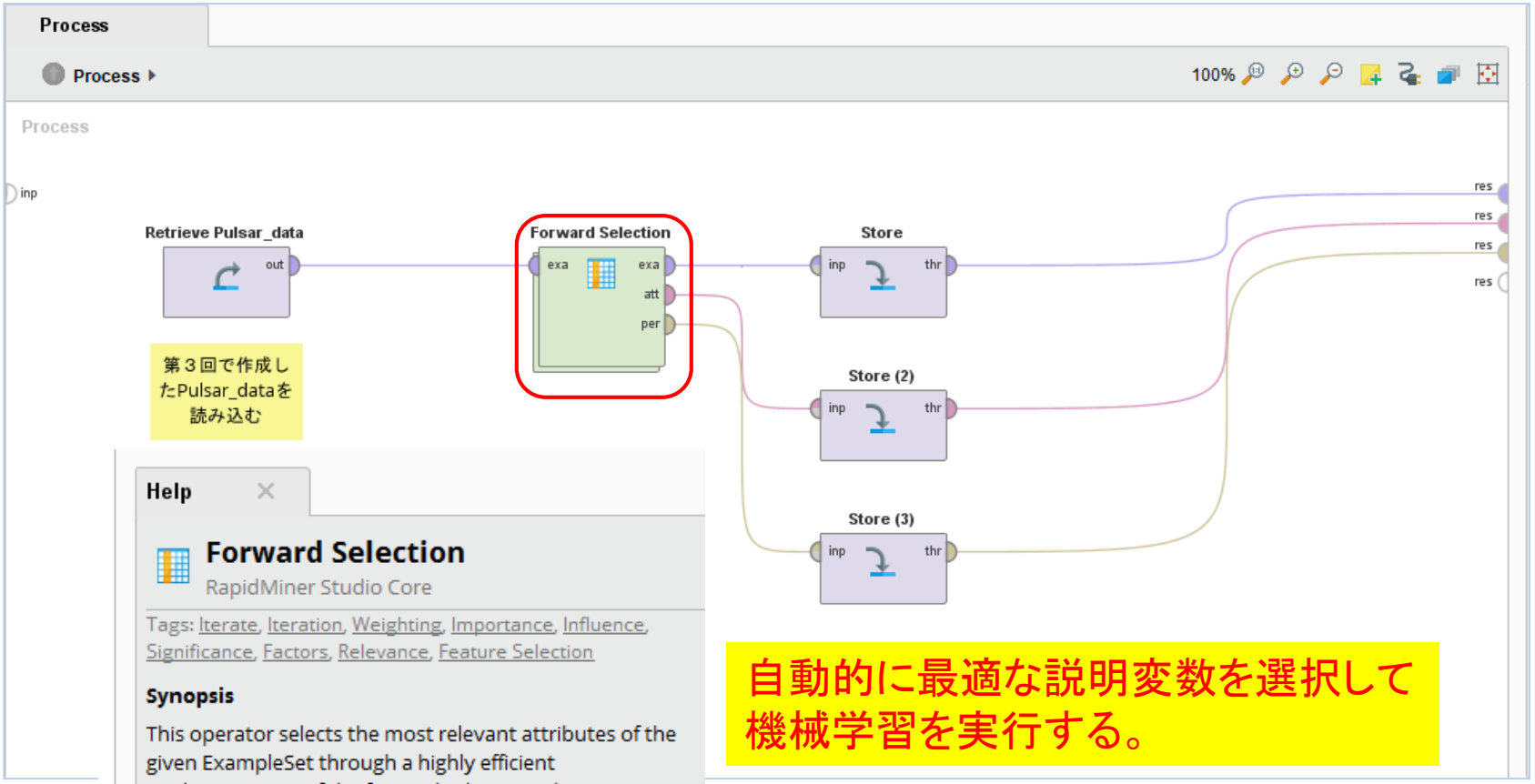
機械学習デモ＋実習

1．交差検証（NN, RF）



機械学習デモ＋実習

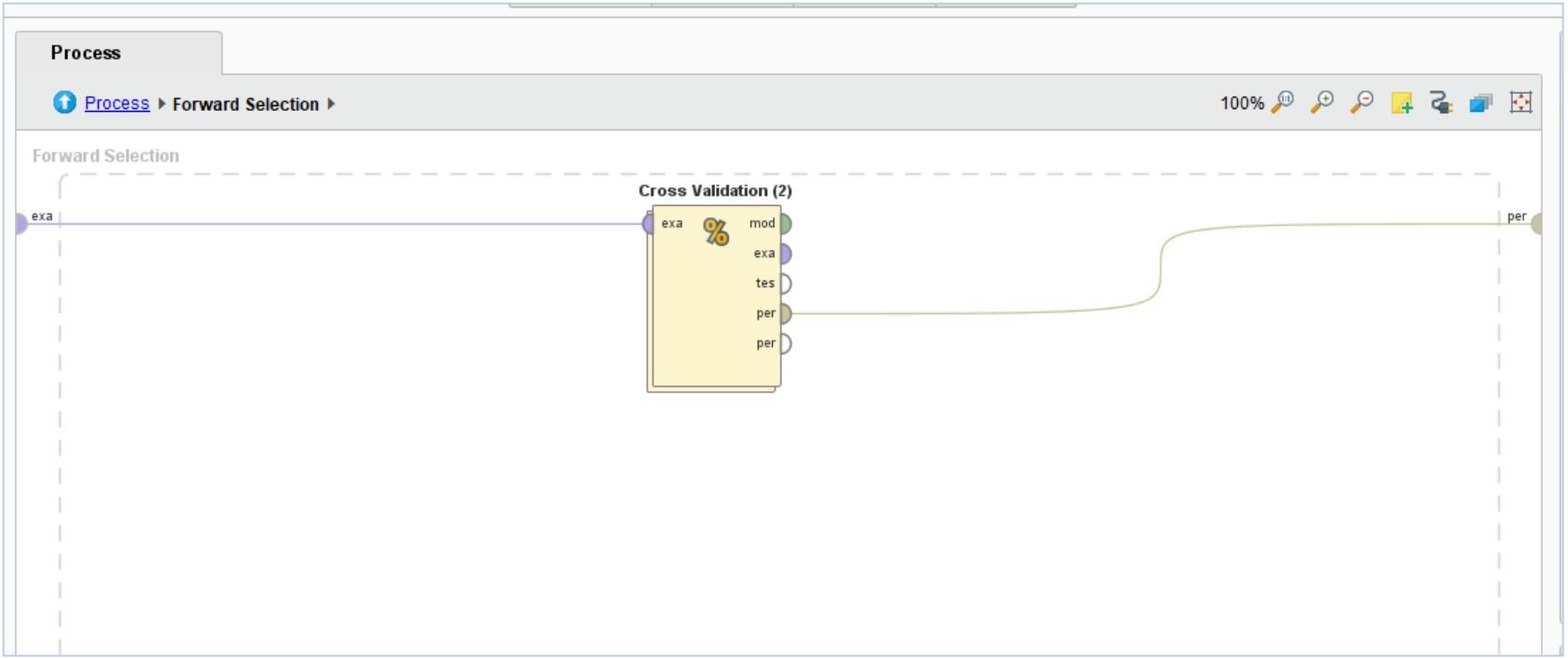
2．Forward Selection



自動的に最適な説明変数を選択して機械学習を実行する。

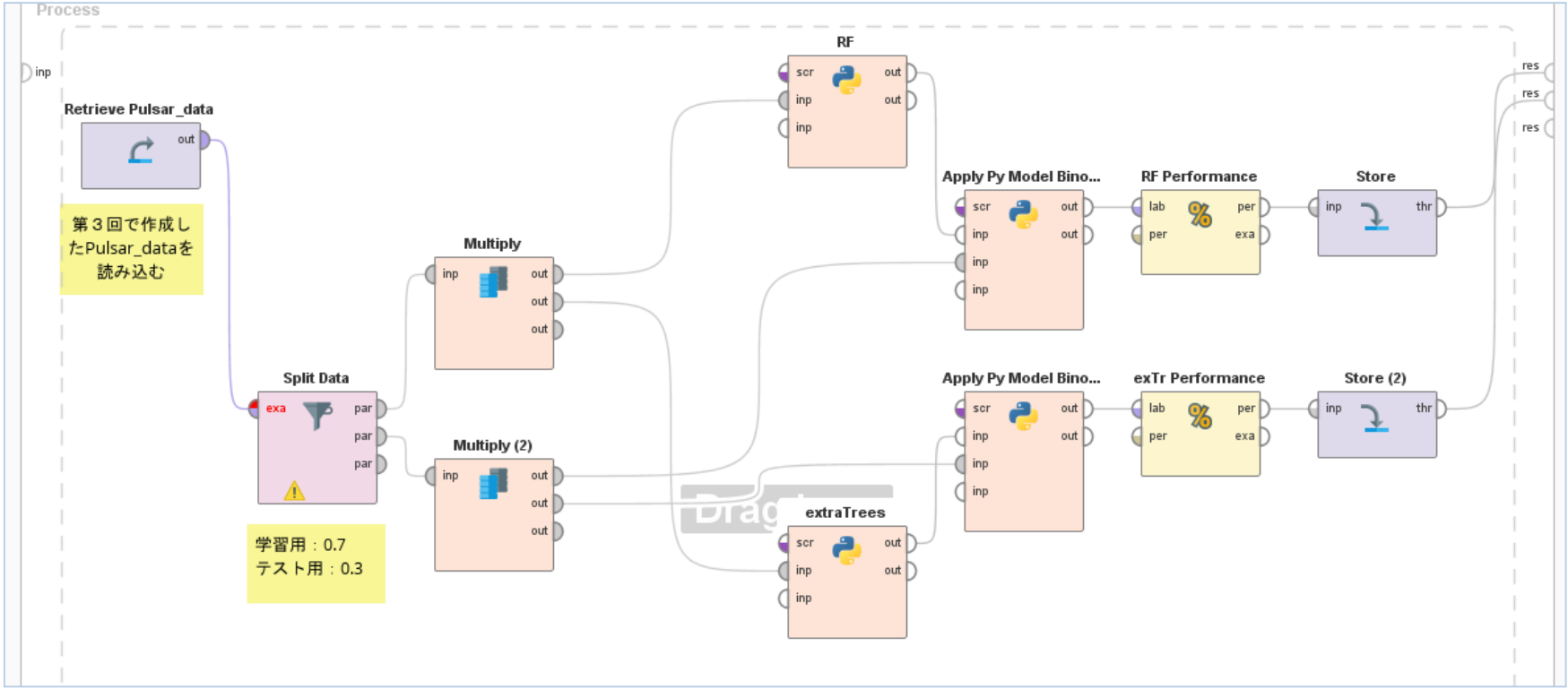
機械学習デモ＋実習

2．Forward Selection



機械学習デモ＋実習

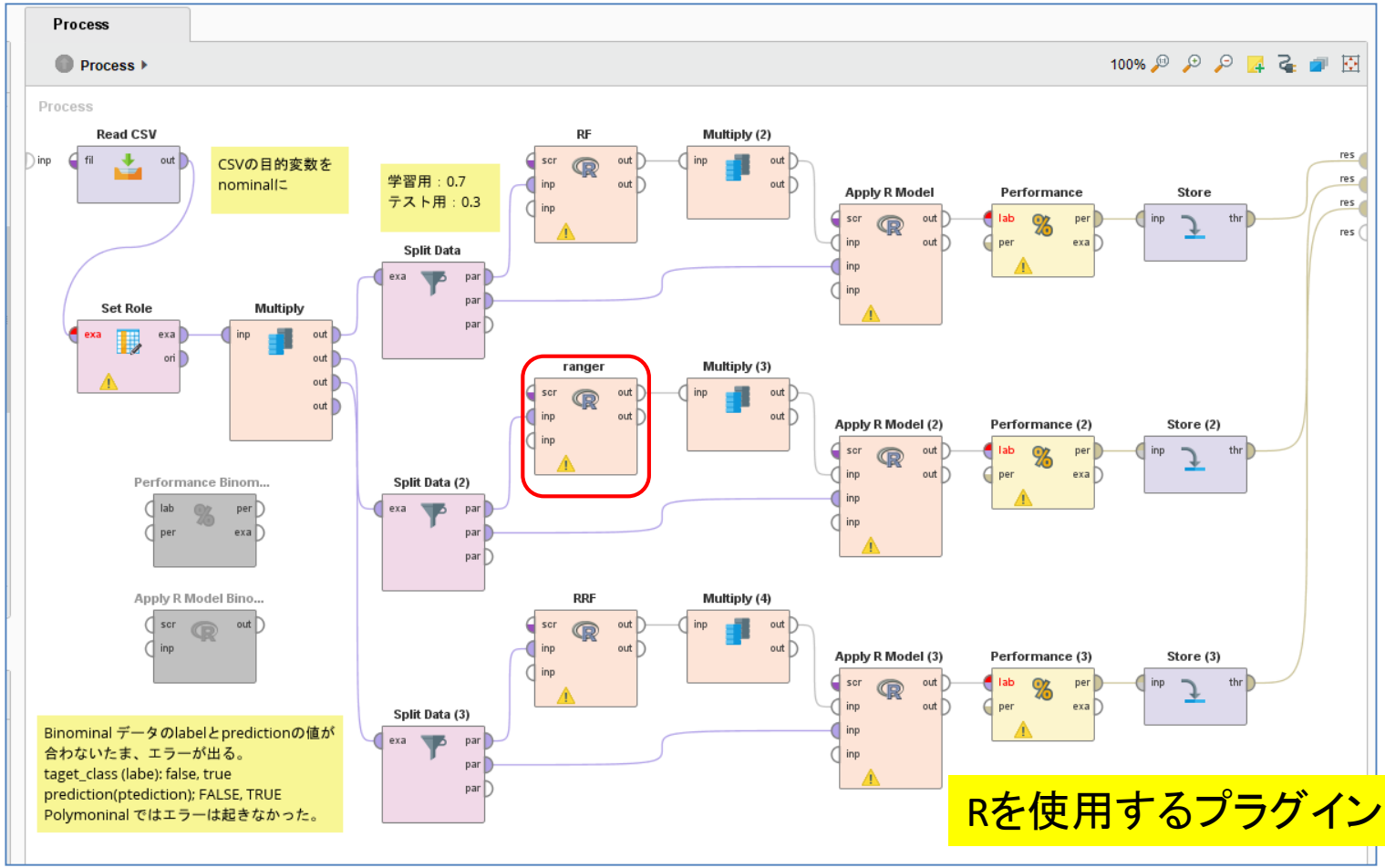
3．Python Script



Python Script を使用するプラグインインストール

機械学習デモ＋実習

4. R Script



機械学習デモ＋実習

4．R Script 結果

R Script ranger

accuracy: 98.42%

	true nonPulsar	true Pulsar	class precision
pred. nonPulsar	4851	58	98.82%
pred. Pulsar	27	434	94.14%
class recall	99.45%	88.21%	

RapidMiner Random Forest
（3回目）

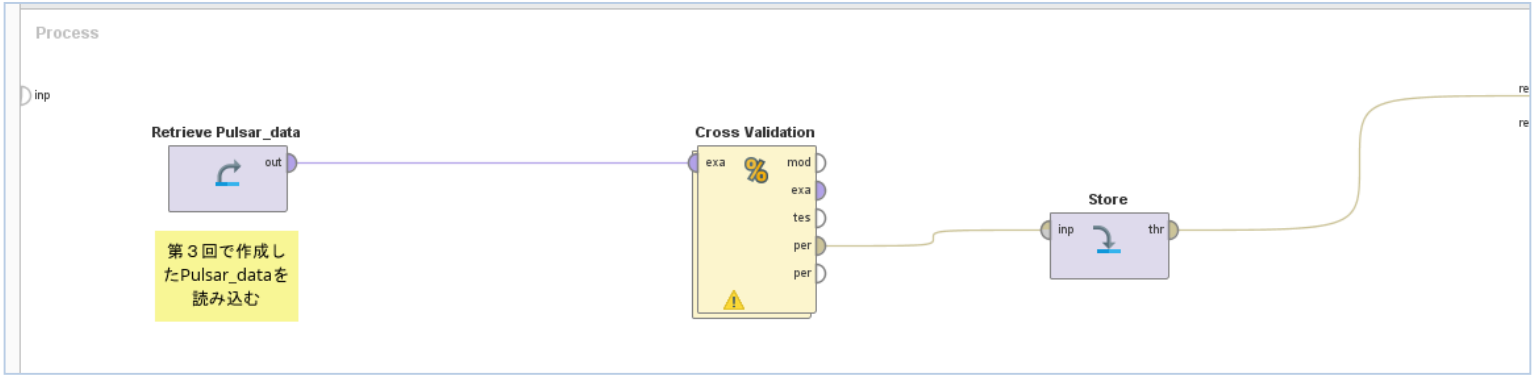
accuracy: 98.16%

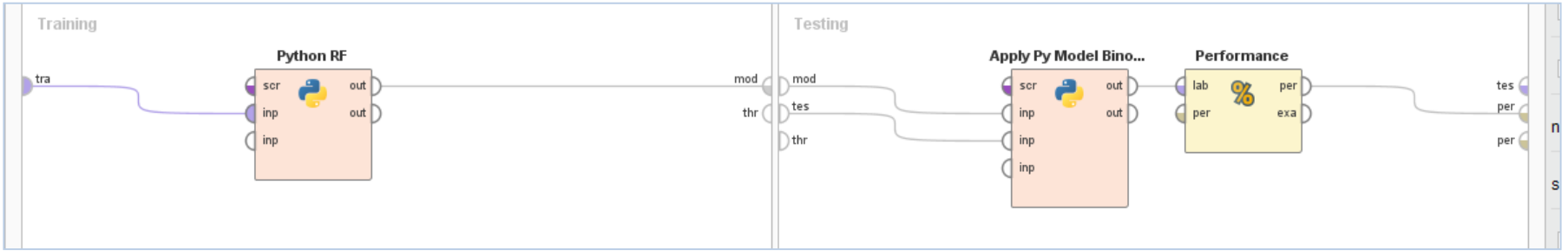
	true false	true true	class precision
pred. false	4872	81	98.36%
pred. true	18	398	95.67%
class recall	99.63%	83.09%	

計算速度：　RapidMiner > Python >> R

機械学習デモ＋実習

5. Cross Validation of Python Script

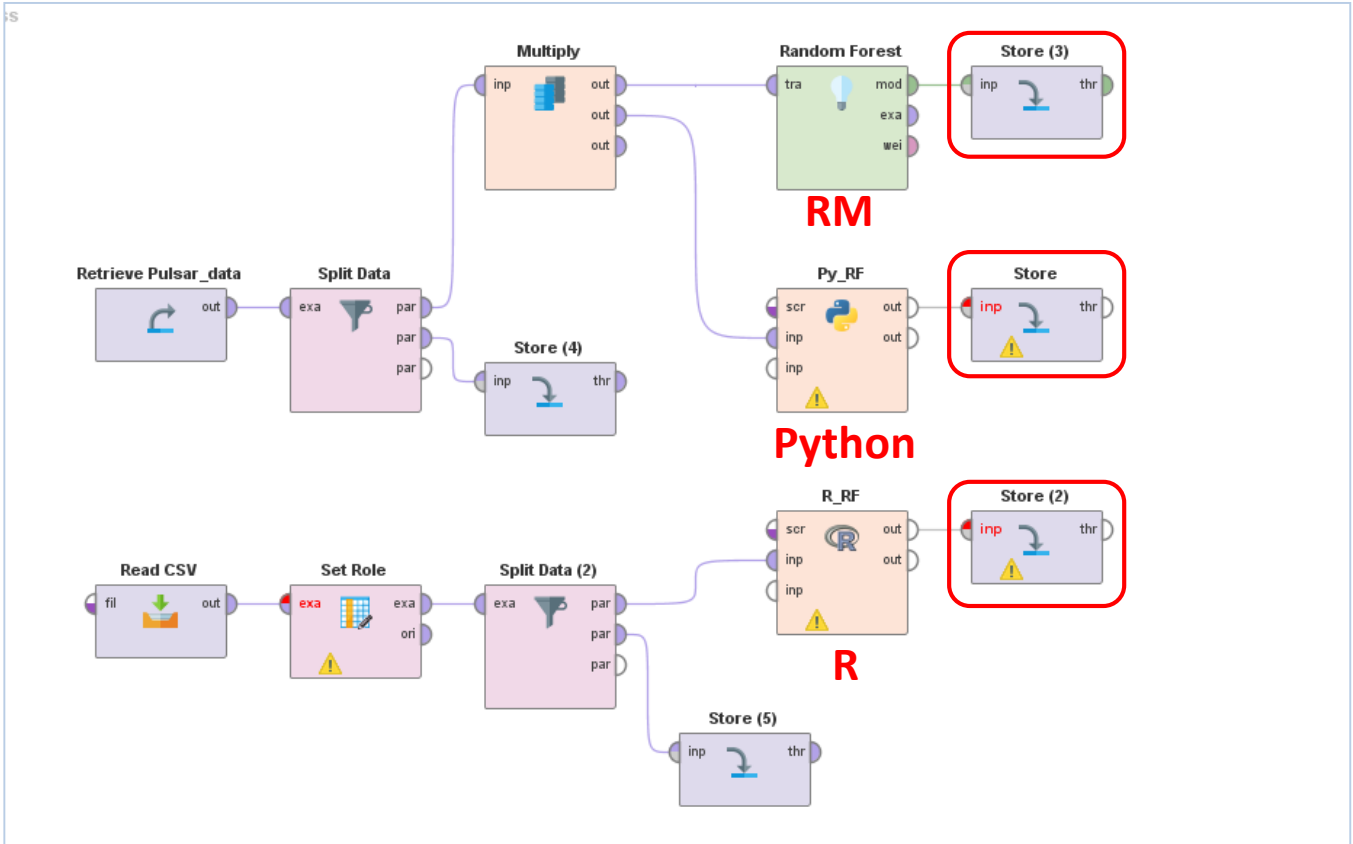




PythonやRのスクリプトもRMの分類器と同様に交差検証ができる

機械学習デモ＋実習

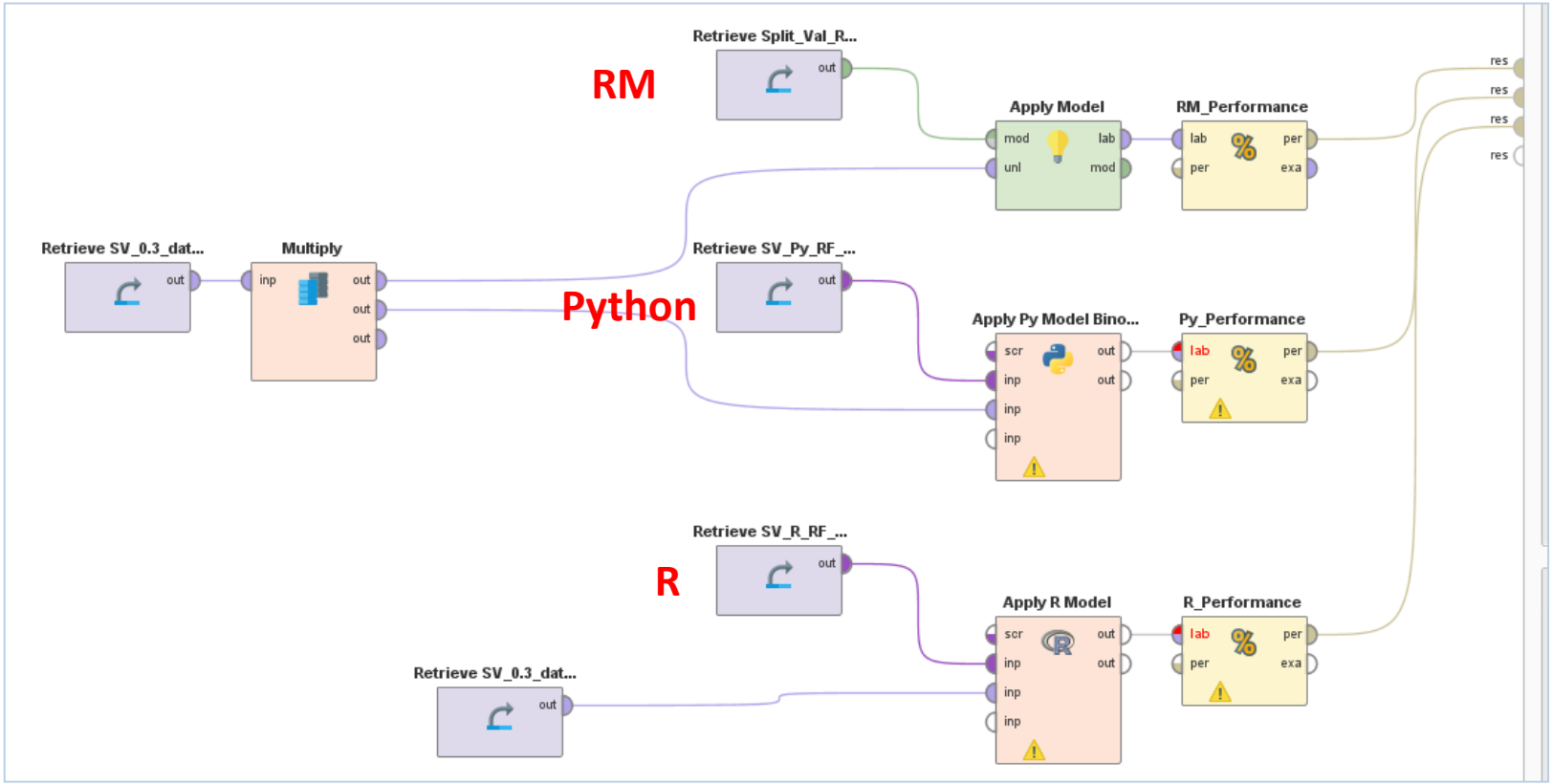
6. Modelの保存



学習したモデル（RM, Python, R script）はメタデータとして保存できる

6. Modelの再利用

機械学習デモ＋実習



メタデータで保存したモデルでの予想は高速で行える

7. 分類問題の評価指数（2値問題）

混合行列(Confusion Matrix)とTP, FP, FN, TN

	実際は正 (Positive)	実際は負 (Negative)
予測が正 (Positive)	**TP** True Positive	**FP** False Positive 第1種の誤り
予測が負 (Negative)	**FN** Flase Negative 第2種の誤り	**TN** True Negative

適合率と再現率はトレードオフの関係
両者のバランスを評価するのがF1値

参考：
https://qiita.com/FukuharaYohei/items/be89a99c53586fa4e2e4

指標	意味	式
正解率(Accuracy)	全予測 正答率	$\frac{TP + TN}{TP + FP + FN + TN}$
適合率(Precision)	正予測 の正答率	$\frac{TP}{TP + FP}$
再現率(Recall)	正に対する 正答率	$\frac{TP}{TP + FN}$
特異率(Specificity)	負に対する 正答率	$\frac{TN}{FP + TN}$
F値(F-measure)	適合率と 再現率の 調和平均	$\frac{2 \times 適合率 \times 再現率}{適合率 + 再現率}$

機械学習デモ＋実習

正則化に関連した自主勉強：

RapidMiner のHPからダウンロードしたプロセスを添付PDFと以下のビデオを参照して解読

https://www.youtube.com/watch?v=oSLASLV4cTc

使用するデータセット（プロセスではSampleから読み込み） https://www.kaggle.com/adx891/sonar-data-set

Data Set Information:

The file "sonar.mines" contains 111 patterns obtained by bouncing sonar signals off a metal cylinder at various angles and under various conditions. The file "sonar.rocks" contains 97 patterns obtained from rocks under similar conditions. The transmitted sonar signal is a frequency-modulated chirp, rising in frequency. The data set contains signals obtained from a variety of different aspect angles, spanning 90 degrees for the cylinder and 180 degrees for the rock.

Each pattern is a set of 60 numbers in the range 0.0 to 1.0. Each number represents the energy within a particular frequency band, integrated over a certain period of time. The integration aperture for higher frequencies occur later in time, since these frequencies are transmitted later during the chirp.

The label associated with each record contains the letter "R" if the object is a rock and "M" if it is a mine (metal cylinder). The numbers in the labels are in increasing order of aspect angle, but they do not encode the angle directly.