

ノンコーディングで行う機械入門

Lesson 1 機械学習概要、ソフトウェアRapidMiner Studio概要と事例デモ・実習

Lesson 2 分類1: 主要なアルゴリズム説明と応用事例デモ・実習

Lesson 3 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

Lesson 4 分類3: 交差検証、最適アルゴリズム探索の実習

Lesson 5 回帰: 主要なアルゴリズム説明と実習

Lesson 6 (応用) 時系列データの機械学習

Lesson 7 (応用) Extensionによる機能拡張と画像の分類

Lesson 8 自ら学ぶ: RapidMiner のウェブサイトの活用

機械学習プロセス



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



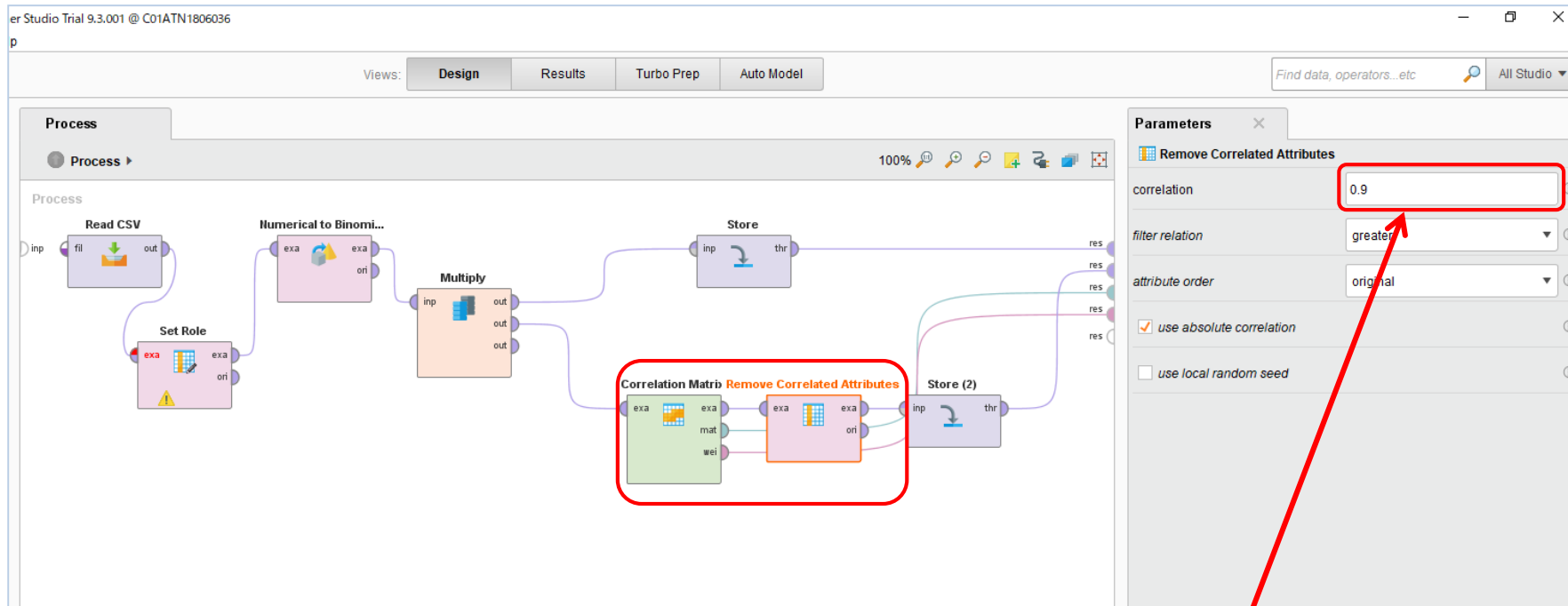
上記8つの説明変数を基に、パルサーかどうか(目的変数)の判別を行う。

→ 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)

機械学習デモ＋実習

1. データ前処理



相関係数が設定値以上の項目(attribute) は
集約して1つにする。

機械学習デモ+実習

1. データ前処理 項目の相関結果

//Local Repository/勉強会用/第3回/processes/1_データ前処理 - RapidMiner Studio Trial 9.3.001 @ C01ATN1806036

File Edit Process View Connections Settings Extensions Help

Views: Design

Result History AttributeWeights (Correlation Matrix) Correlation Matrix (Correlation Matrix)

Data

Pairwise Table

Matrix Visualization

Annotations

Attribut...	Mean of...	Standar...	Excess ...	Skewne...	Mean of...	Standar...	Excess ...	Skewne...
Mean of t...	1	0.547	-0.874	-0.739	-0.299	-0.307	0.234	0.144
Standard...	0.547	1	-0.521	-0.540	0.007	-0.048	0.029	0.028
Excess k...	-0.874	-0.521	1	0.946	0.414	0.433	-0.341	-0.214
Skewne...	-0.739	-0.540	0.946	1	0.412	0.415	-0.329	-0.205
Mean of t...	-0.299	0.007	0.414	0.412	1	0.797	-0.616	-0.354
Standard...	-0.307	-0.048	0.433	0.415	0.797	1	-0.810	-0.576
Excess k...	0.234	0.029	-0.341	-0.329	-0.616	-0.810	1	0.924
Skewne...	0.144	0.028	-0.214	-0.205	-0.354	-0.576	0.924	1

機械学習デモ＋実習

1. データ前処理

項目の相関係数の閾値以上の項目削除結果

Result History: AttributeWeights (Correlation Matrix) × Correlation Matrix (Correlation Matrix) × ExampleSet (Remove Correlated Attributes) × ExampleSet (Multiply)

Views: Design Results Turbo Prep Auto Model

Open in: Turbo Prep Auto Model

Row No.	target_class	Mean of the integrated...	Standard deviation of t...	Excess kurtosis of t...	Mean of the DM-SNR...	Standard deviation of ...	Excess kurtosis of the DM-S...
1	false	140.562	55.684	-0.235	3.200	19.110	7.976
2	false	102.508	58.882	0.465	1.677	14.860	10.576
3	false	103.016	39.342	0.323	3.121	21.745	7.736
4	false	136.750	57.178	-0.068	3.643	20.959	6.896
5	false	88.727	40.672	0.601	1.179	11.469	14.270
6	false	93.570	46.698	0.532	1.636	14.545	10.622
7	false	119.484	48.765	0.031	0.999	9.280	19.206
8	false	130.383	39.844	-0.158	1.221	14.379	13.539
9	false	107.250	52.627	0.453	2.332	14.487	9.001
10	false	107.258	39.496	0.466	4.079	24.980	7.397
11	false	142.078	45.288	-0.320	5.376	29.010	6.076
12	false	133.258	44.058	-0.081	1.632	12.008	11.972
13	false	134.961	49.554	-0.135	10.696	41.342	3.894
14	false	117.945	45.507	0.325	2.836	23.118	8.943
15	false	138.180	51.524	-0.032	6.330	31.576	5.156
16	false	114.367	51.946	-0.094	2.738	17.192	9.051
17	false	109.641	49.018	0.138	1.508	12.073	13.368
18	false	100.852	51.744	0.394	2.841	21.636	8.302
19	false	136.094	51.691	-0.046	9.343	38.096	4.345
20	true	99.367	41.572	1.547	27.555	61.719	2.209
21	false	100.891	51.890	0.627	3.884	23.045	6.953

説明変数が8個から6個に減少

機械学習デモ＋実習

2. 欠損値の処理

///Local Repository/勉強会用/第3回/processes/2_欠損値処理 - RapidMiner Studio Trial 9.3.001 @ C01ATN1806036

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model

Result History ExampleSet (Multiply) ExampleSet (Guess Types) ExampleSet (Impute Missing Values) ExampleSet (Filter Examples)

Data

Statistics

Visualizations

Annotations

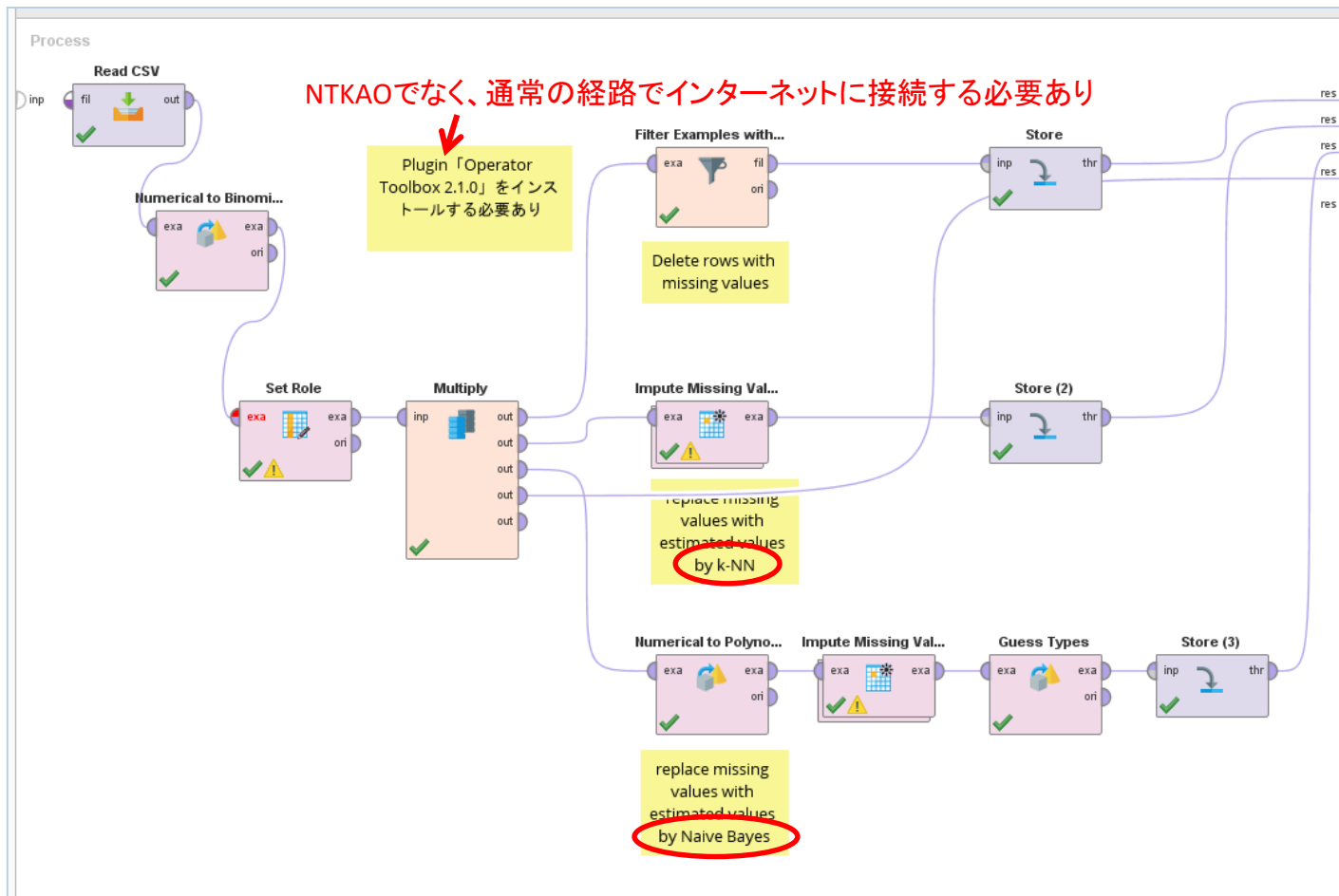
Name	Type	Missing	Statistics
Label target_class	Binominal	0	
Mean of the integrated profile	Real	4	Min: 5.812, Max: 184.297, Average: 110.317
Standard deviation of the integ...	Real	4	Min: 28.029, Max: 76.784, Average: 46.360
Excess kurtosis of the integrat...	Real	4	
Skewness of the integrated pr...	Real	2	
Mean of the DM-SNR curve	Real	7	
Standard deviation of the DM-...	Real	6	Min: 7.805, Max: 108.078, Average: 26.463
Excess kurtosis of the DM-SNR...	Real	6	Min: -2.637, Max: 29.765, Average: 8.306
Skewness of the DM-SNR curve	Real	6	Min: -1.946, Max: 949.700, Average: 105.946

目的変数の欠損値は0でなければならない

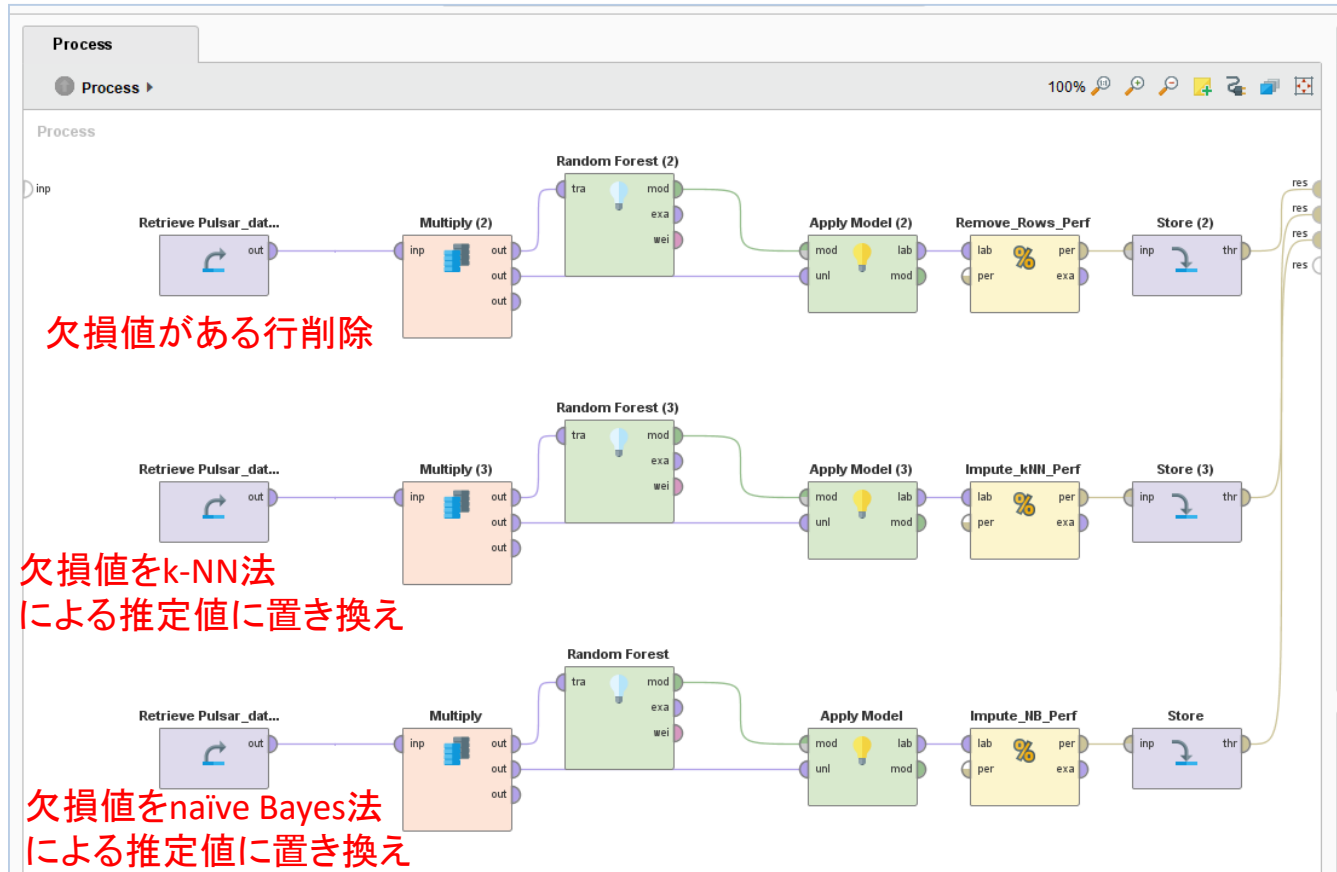
説明変数に欠損値がある場合は
(1)削除する
(2)欠損値を推定値で補う
のいずれかの処理をする

機械学習デモ＋実習

2. 欠損値の処理



3. 欠損値のテスト



結論だけ書くと、
今回の条件では
k-NN法での欠損
値の置き換えが
一番精度が高った



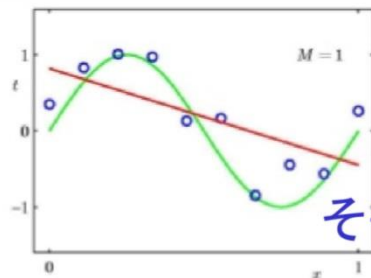
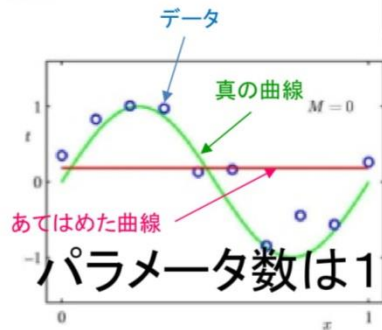
データセットや機械
学習アルゴリズムに
より、結果は違う

過学習

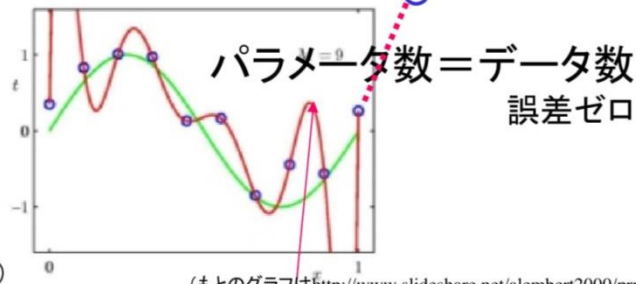
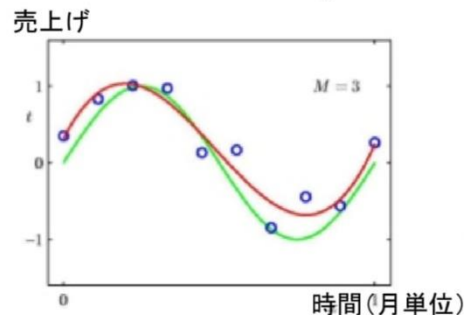
表現能力の高いモデルには過学習 (over fitting) が避けられない

過学習

- 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！



そもそも外挿は無理



(もとのグラフは<http://www.slideshare.net/alembert2000/prml-at-1>より)

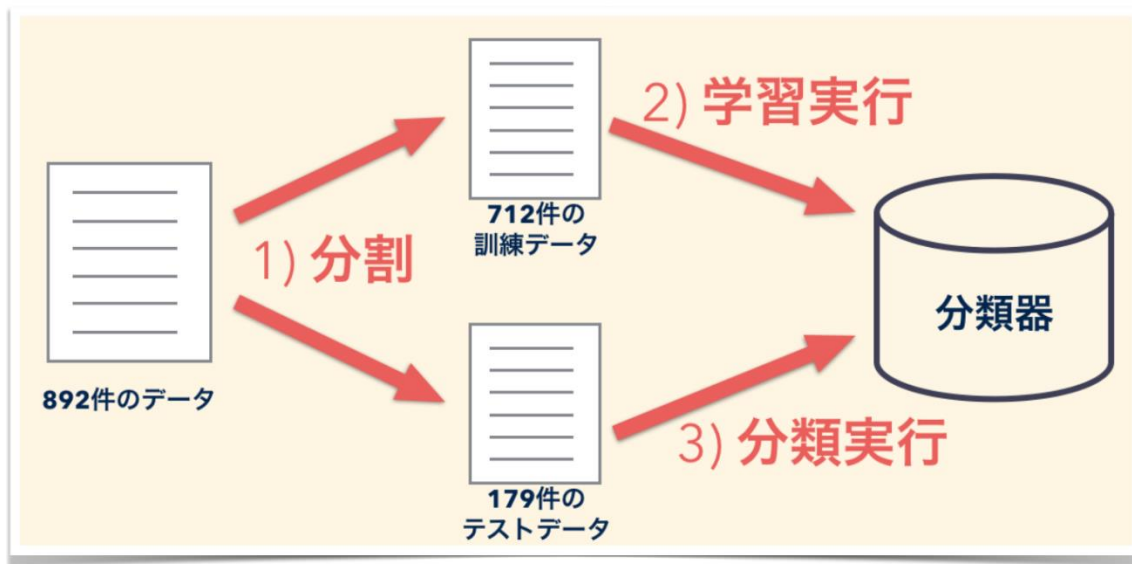
内挿の失敗:こんな予測を信じますか？

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

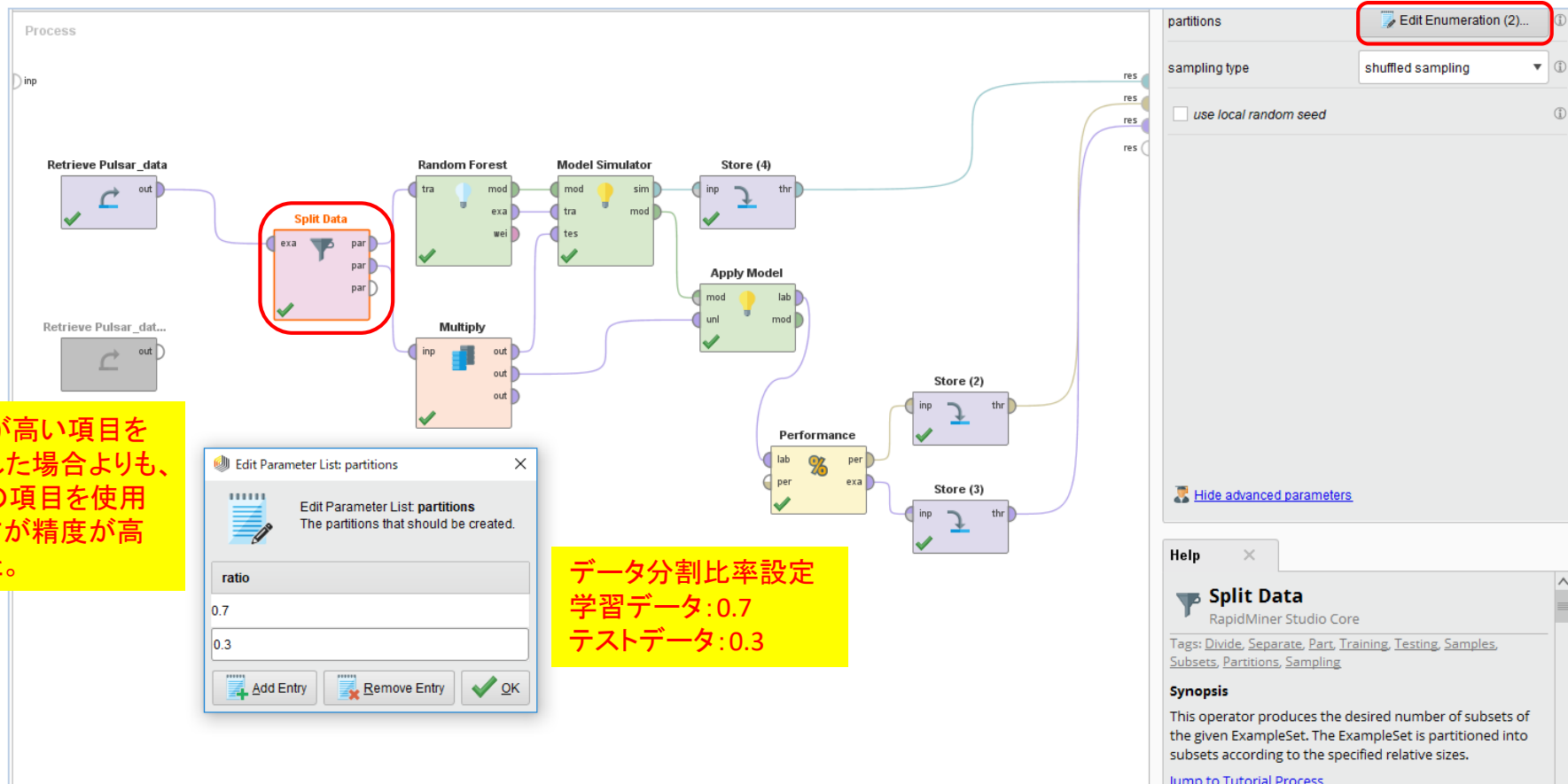
本勉強会では、2つの手法を取り扱います。

(1) 教師(訓練)データとテストデータに**分割**



(2) 交差検証 (Cross Validation) → 次回の勉強会で取り扱います。

3. データ分割による検証



3. データ分割による検証

機械学習デモ＋実習

学習データ:テストデータ= 7 : 3

☒ Table View ☐ Plot View

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%	

学習データ = テストデータ (2回目で実習)

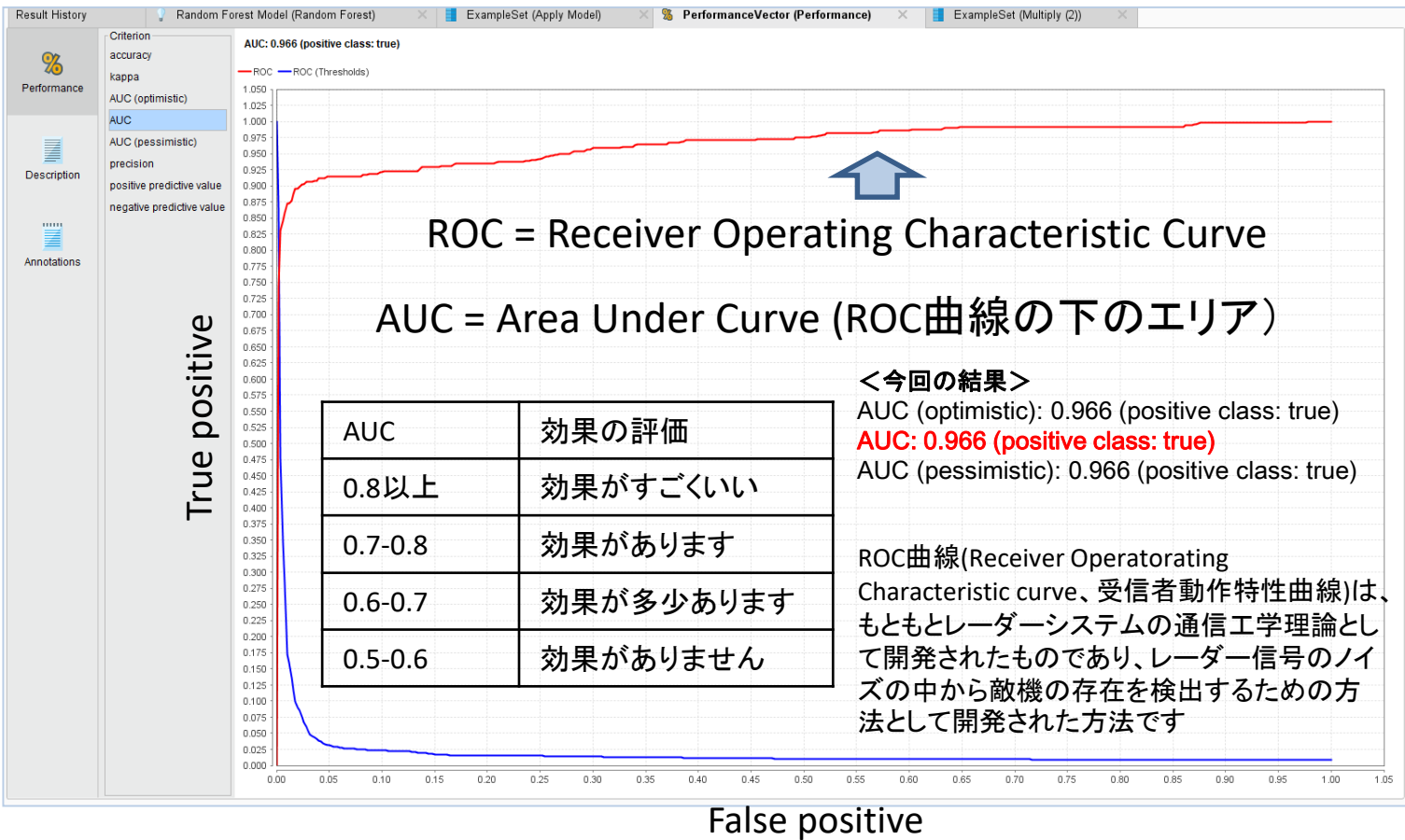
☒ Table View ☐ Plot View

accuracy: 99.99%

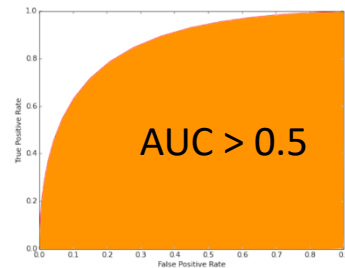
	true false	true true	class precision
pred. false	16259	2	99.99%
pred. true	0	1637	100.00%
class recall	100.00%	99.88%	

3. データ分割による検証

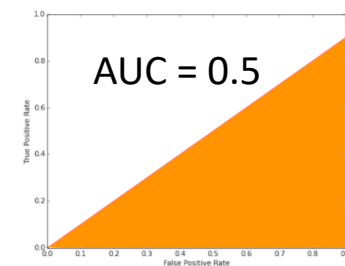
ROC曲線とAUC



分類器の効果



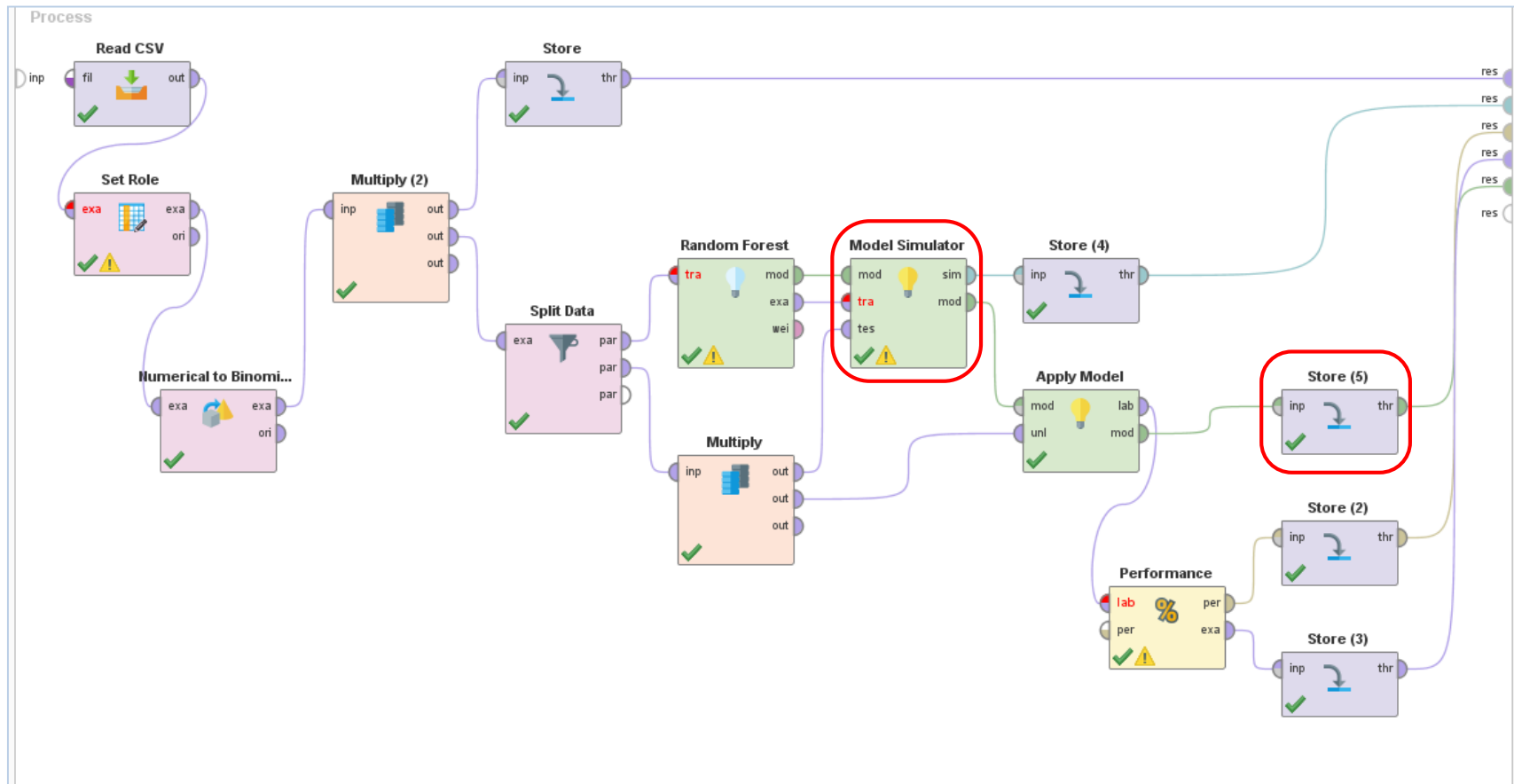
効果がある分類器



効果がない分類器

機械学習デモ＋実習

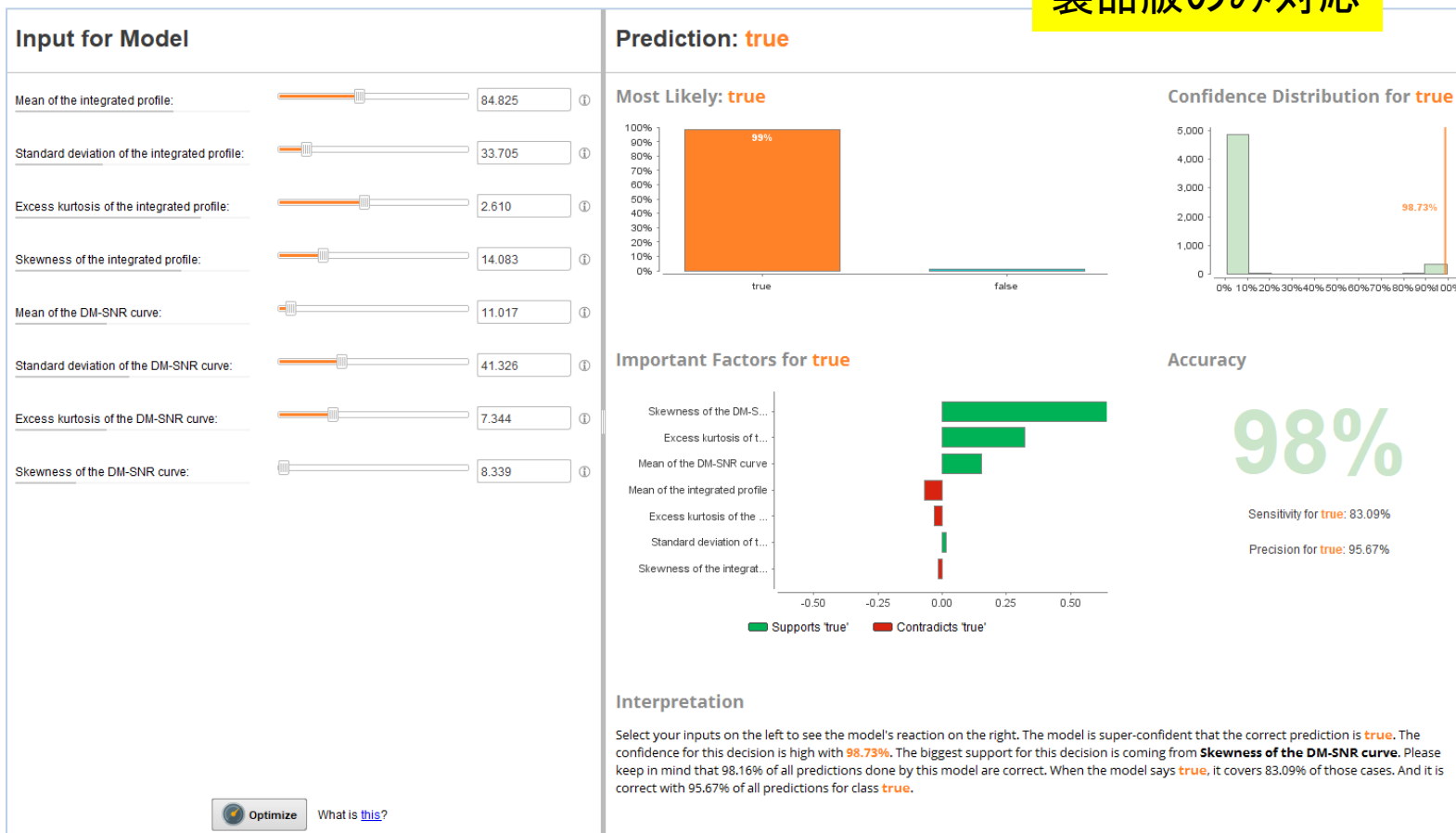
4. Model Simulator とModel保存



4. Model Simulator

機械学習デモ+実習

製品版のみ対応



機械学習デモ＋実習

- Compare ROCs
- Vote → 複数のアルゴリズムで行うアンサンブル学習
- Optimize Parameters
→ ハイパーパラメータのグリッドサーチ