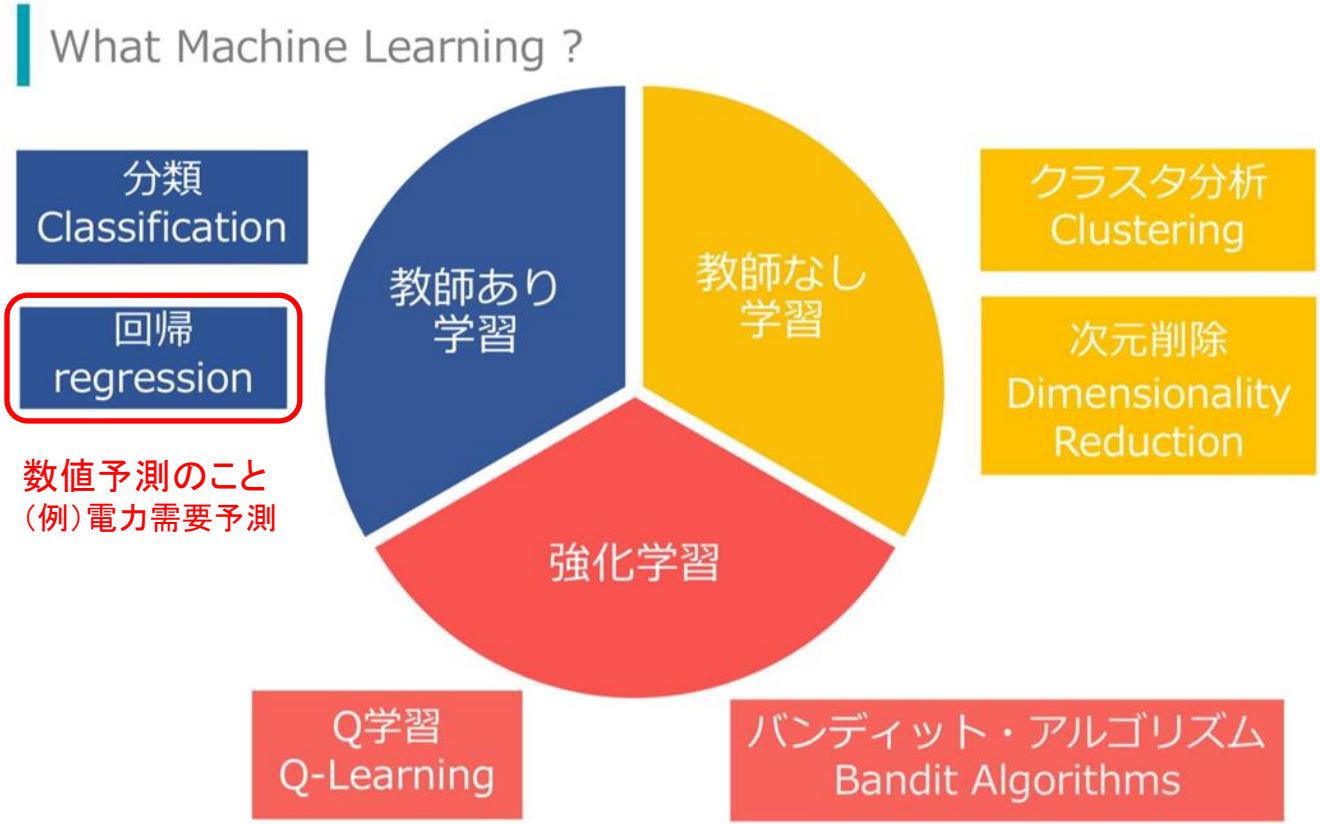
Confidential
期限：2025年12月31日
ヘアケア4研2Gr

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　（応用）テキストマイニング入門

機械学習概要



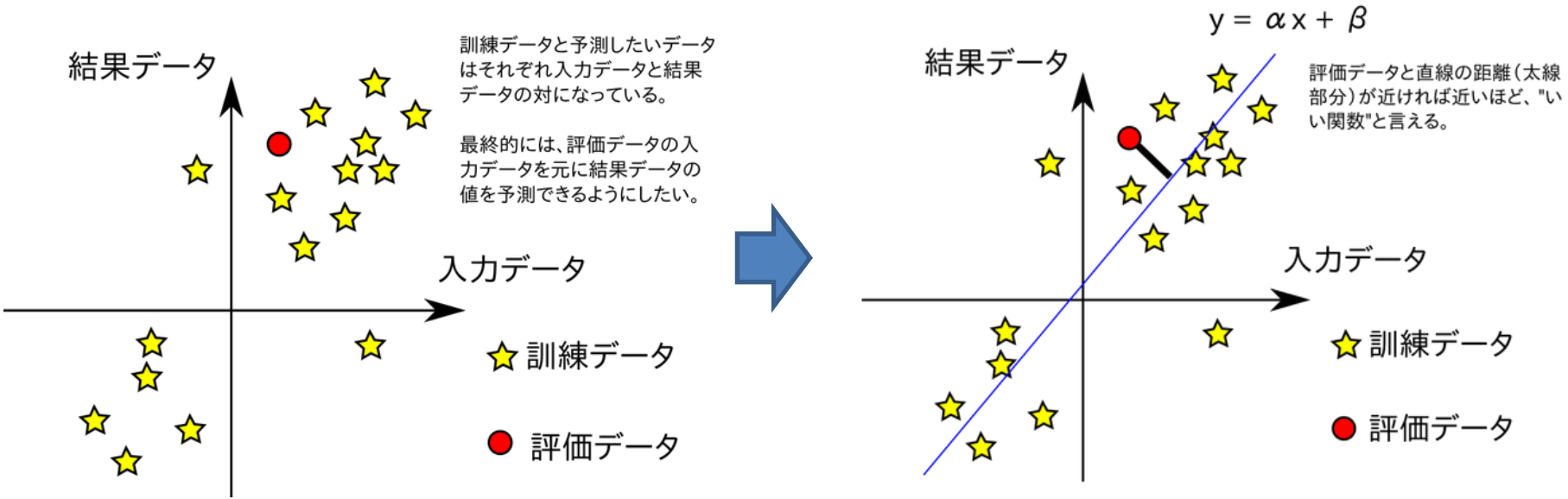
今回は回帰を学びます

数値予測のこと
（例）電力需要予測

Copyright 2019 KSK Analytics Inc.

Confidential
期限：2025年12月31日
ヘアケア4研2Gr

回帰とは？



教師（訓練）データをもとに学習した数式で、未知のデータの数値を予測する

Confidential
期限：2025年12月31日
ヘアケア4研2Gr

回帰に用いられる主なアルゴリズム

・関数で回帰を行う
　　・線形回帰：1次関数で回帰を行う　→　（例）一般化線形回帰

・分類と同じアルゴリズムでも回帰にも使えるものが多くある
　　・ツリー系　→　（例）ランダムフォレスト
　　・SVM
　　・Neural Network系　→（例）Neural Networks, Deep Learning

Confidential
期限：2025年12月31日
ヘアケア4研2Gr

KaggleとGitHub

Kaggle（カグル）は企業や研究者がデータを投稿し、世界中の統計家やデータ分析家がその最適モデルを競い合う、予測モデリング及び分析手法関連プラットフォーム及びその運営会社である。モデル作成にクラウドソーシング手法が採用される理由としては、いかなる予測モデリング課題には無数の戦略が適用可能であり、どの分析手法が最も効果的であるか事前に把握することは不可能であることに拠る。2017年3月8日、**Google**はKaggle社を買収すると発表した。（Wikipedia）

GitHub（ギットハブ）は、ソフトウェア開発のプラットフォームであり、ソースコードをホスティングする。コードのバージョン管理システムにはGitを使用する。Ruby on RailsおよびErlangで記述されており、アメリカのカリフォルニア州サンフランシスコ市に拠点を置くGitHub社によって保守されている。2009年のユーザー調査によると、GitHubは最もポピュラーなGitホスティングサイトとなった。2018年に**マイクロソフト**による買収が発表されている。（Wikipedia)

今回用いるデータセット（Kaggle Datasetsより）

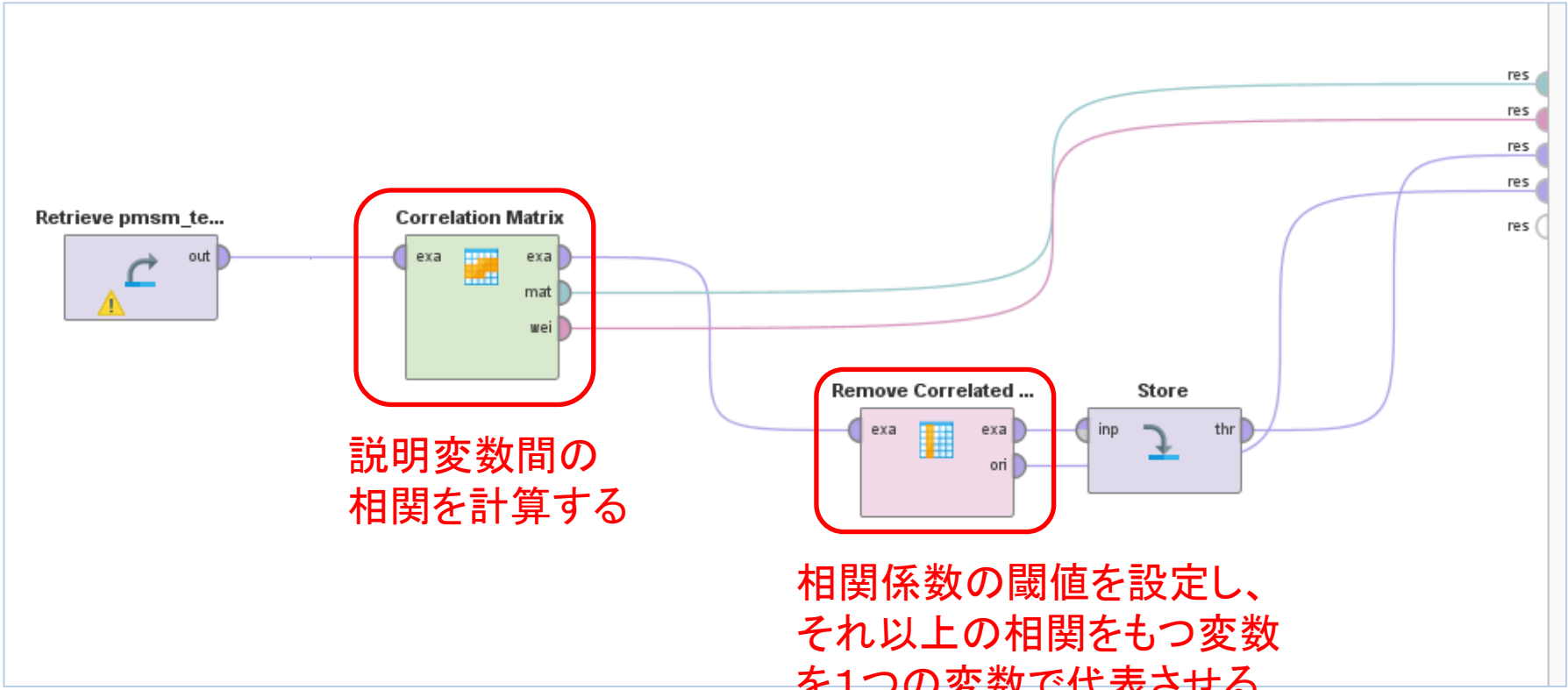
https://www.kaggle.com/wkirgsn/electric-motor-temperature

ambient	coolant	u_d	u_q	motor_speed	torque	i_d	i_q	pm	stator_yoke	stator_tooth	stator_winding
−0.75214297	−1.1184461	0.3279352	−1.2978575	−1.2224282	−0.2501821	1.0295724	−0.24586003	−2.522071	−1.8314217	−2.0661428	−2.0180326
−0.77126324	−1.1170206	0.3296648	−1.2976865	−1.2224293	−0.2491333	1.029509	−0.24583231	−2.5224178	−1.8309687	−2.0648587	−2.0176313
−0.78289163	−1.1166813	0.3327715	−1.3018217	−1.2224278	−0.24943107	1.0294477	−0.24581794	−2.5226731	−1.8304	−2.064073	−2.0173435
−0.78093535	−1.1167642	0.3336999	−1.301852	−1.2224301	−0.24863635	1.0328449	−0.2469548	−2.521639	−1.8303328	−2.0631368	−2.0176322
−0.7740426	−1.116775	0.3352061	−1.303118	−1.2224286	−0.24870083	1.0318071	−0.24660969	−2.5219002	−1.8304977	−2.0627947	−2.0181448
−0.7629362	−1.1169548	0.33490124	−1.3030168	−1.2224286	−0.248197	1.0310309	−0.24634062	−2.522203	−1.8319309	−2.0625494	−2.017884
−0.74922806	−1.1161705	0.33501354	−1.3020816	−1.2224296	−0.24791418	1.0304929	−0.24616154	−2.5225377	−1.8330117	−2.0621152	−2.0172427
−0.7384499	−1.1139864	0.3362563	−1.3051548	−1.2224321	−0.24832098	1.0301074	−0.24603489	−2.5228438	−1.8321822	−2.0619526	−2.0172133
−0.7309097	−1.1118276	0.3349053	−1.3037902	−1.2224315	−0.24778472	1.0298513	−0.24598087	−2.5228078	−1.8315759	−2.062443	−2.0177386
−0.72712964	−1.1094859	0.3359881	−1.3056333	−1.2224314	−0.24829444	1.0296358	−0.24588774	−2.5226767	−1.8314383	−2.062317	−2.0181801
−0.72371286	−1.1082836	0.33539978	−1.3045602	−1.2224283	−0.24791297	1.029509	−0.24583231	−2.5226302	−1.8314928	−2.0620575	−2.0176919
−0.71775365	−1.1085879	0.33443072	−1.3043374	−1.2224288	−0.24772124	1.0293863	−0.24580358	−2.5226388	−1.8318189	−2.0622487	−2.017435

目的変数：**ambient**
説明変数：**11項目**
今回用いたデータセットの行数：**33,426**

Confidential
期限：2025年12月31日
ヘアケア4研2Gr

データ前処理



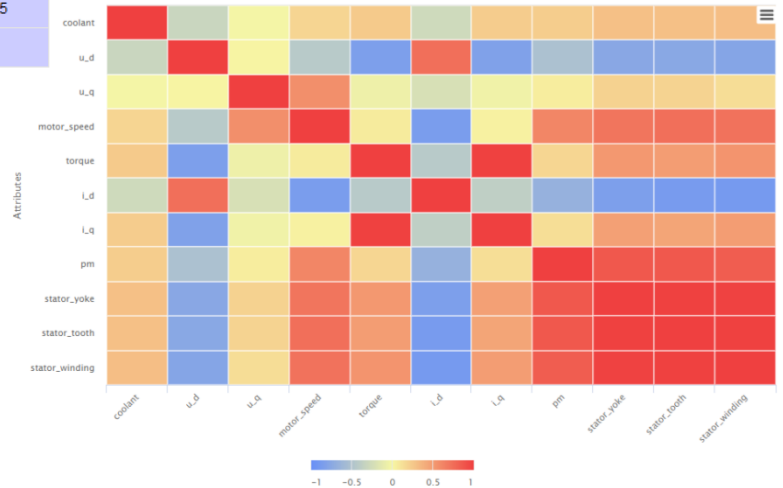
説明変数間の
相関を計算する

相関係数の閾値を設定し、
それ以上の相関をもつ変数
を1つの変数で代表させる
→　次元削減

Confidential
期限：2025年12月31日
ヘアケア4研2Gr

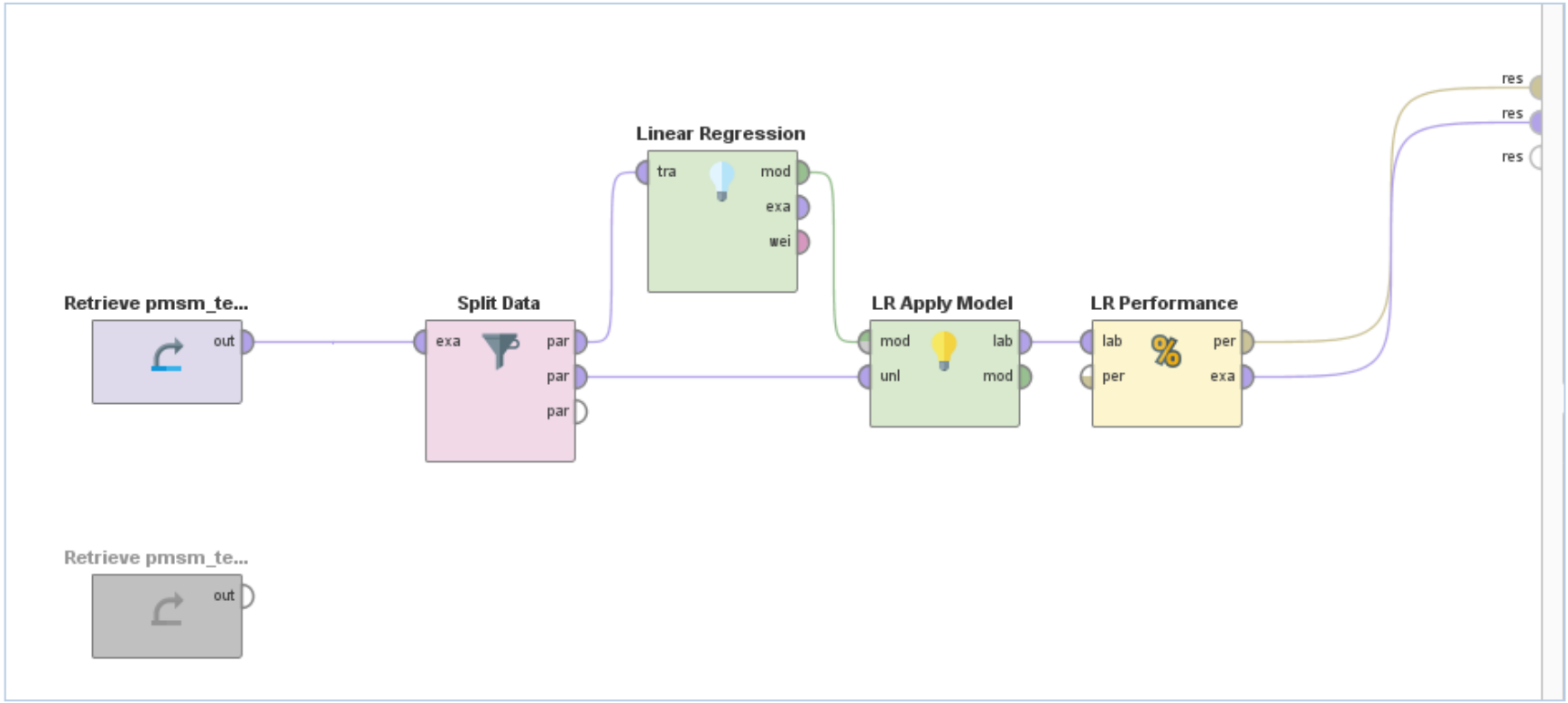
データ前処理 （相関マトリックス）

Attribut...	coolant	u_d	u_q	motor_...	torque	i_d	i_q	pm	stator_...	stator_t...	stator_...
coolant	1	-0.320	-0.021	0.184	0.248	-0.283	0.233	0.229	0.301	0.301	0.310
u_d	-0.320	1	0.015	-0.437	-0.850	0.742	-0.826	-0.519	-0.766	-0.764	-0.791
u_q	-0.021	0.015	1	0.567	-0.063	-0.219	-0.049	0.056	0.201	0.195	0.142
motor_s...	0.184	-0.437	0.567	1	0.067	-0.867	0.031	0.618	0.709	0.744	0.725
torque	0.248	-0.850	-0.063	0.067	1	-0.437	0.997	0.182	0.517	0.494	0.539
i_d	-0.283	0.742	-0.219	-0.867	-0.437	1	-0.388	-0.675	-0.848	-0.880	-0.897
i_q	0.233	-0.826	-0.049	0.031	0.997	-0.388	1	0.138	0.474	0.449	0.493
pm	0.229	-0.519	0.056	0.618	0.182	-0.675	0.138	1	0.869	0.871	0.841
stator_yo...	0.301	-0.766	0.201	0.709	0.517	-0.848	0.474	0.869	1	0.996	0.989
stator_to...	0.301	-0.764	0.195	0.744	0.494	-0.880	0.449	0.871	0.996	1	0.995
stator_wi...	0.310	-0.791	0.142	0.725	0.539	-0.897	0.493	0.841	0.989	0.995	1



Confidential
期限：2025年12月31日
ヘアケア4研2Gr

回帰：プロセス

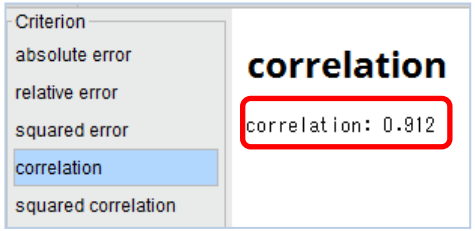


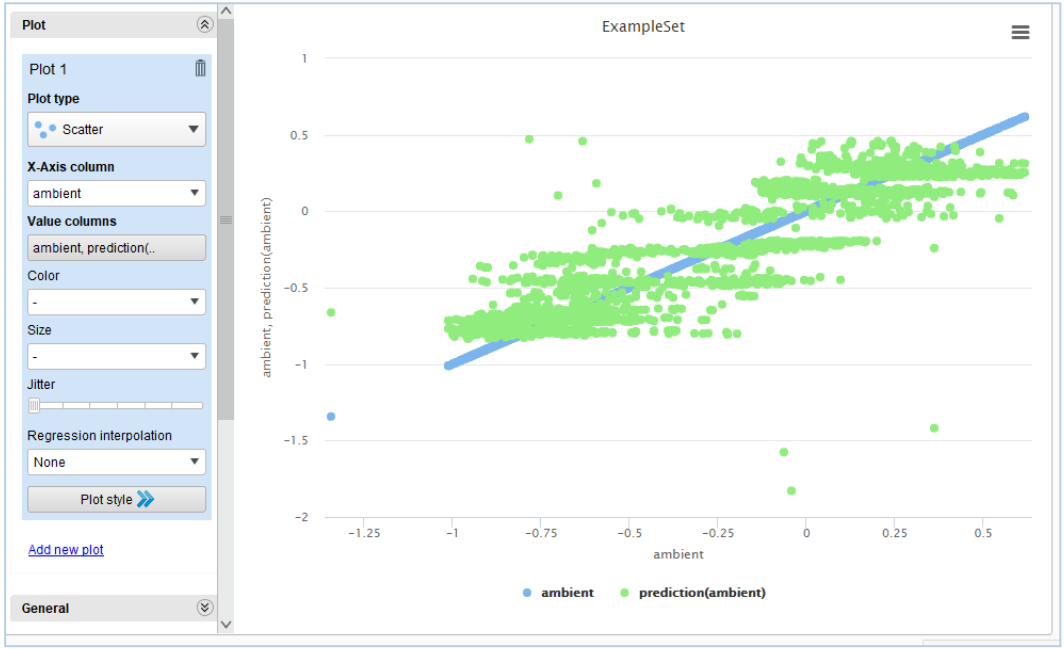
アイコンの配置は、分類問題と同じ

Confidential
期限：2025年12月31日
ヘアケア4研2Gr

回帰：結果

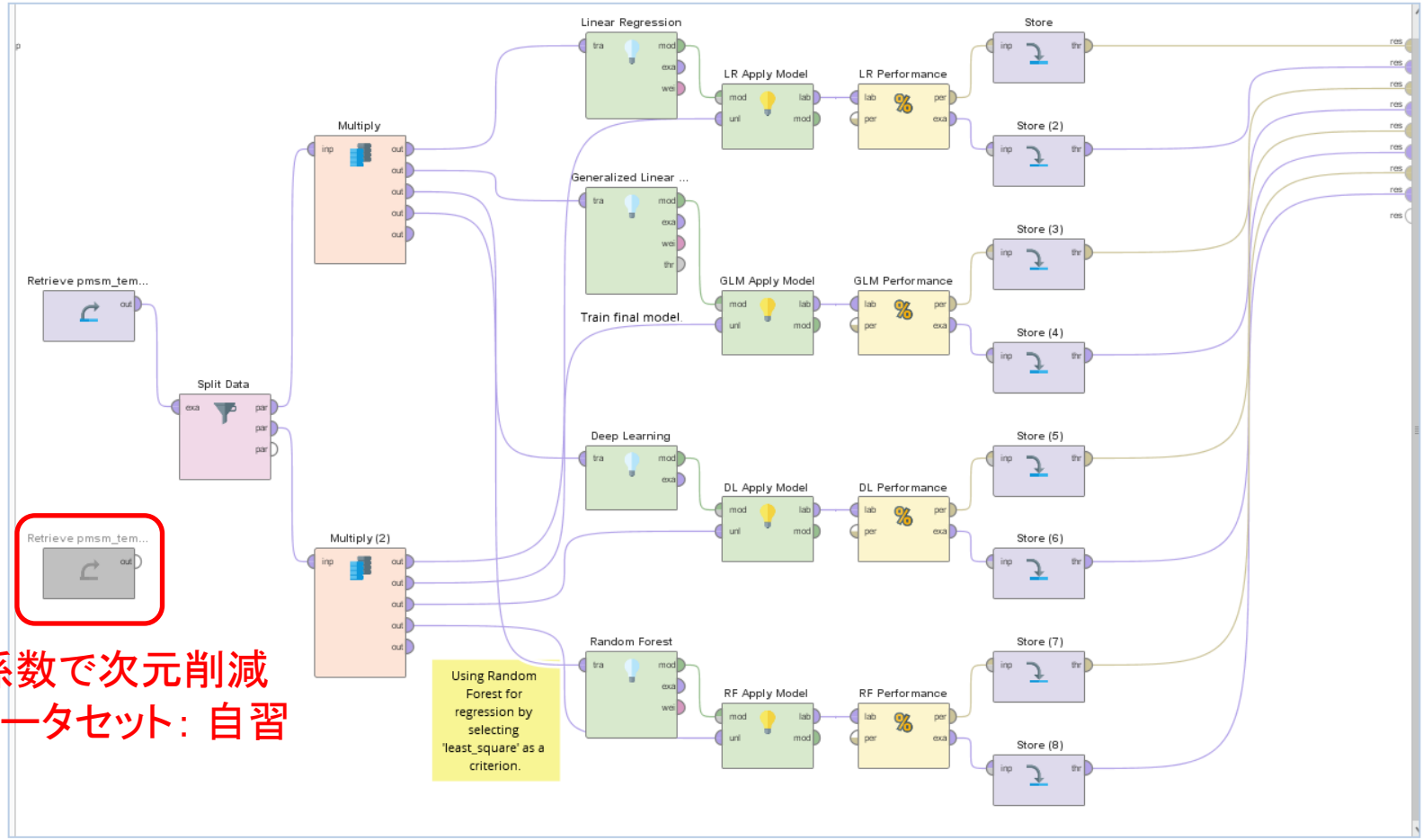
ambient	prediction(a...	coolant	u_d
-0.783	0.472	-1.117	0.333
-0.631	0.458	-1.114	0.335
-0.604	-0.124	-1.107	0.296
-0.593	0.182	-1.100	0.261
-0.703	0.103	-1.097	0.237
-0.678	-0.394	-1.096	0.191
-0.718	-0.698	-1.096	0.188
-0.717	-0.718	-1.095	0.191
-0.759	-0.768	-1.095	0.189
-0.830	-0.789	-1.095	0.188
-0.908	-0.768	-1.095	0.188
-0.958	-0.765	-1.095	0.190
-0.854	-0.760	-1.094	0.189
-0.896	-0.747	-1.093	0.189
-0.906	-0.757	-1.094	0.188





Confidential
期限：2025年12月31日
ヘアケア4研2Gr

回帰：種々のアルゴリズム



相関係数で次元削減したデータセット：自習

Confidential
期限：2025年12月31日
ヘアケア4研2Gr

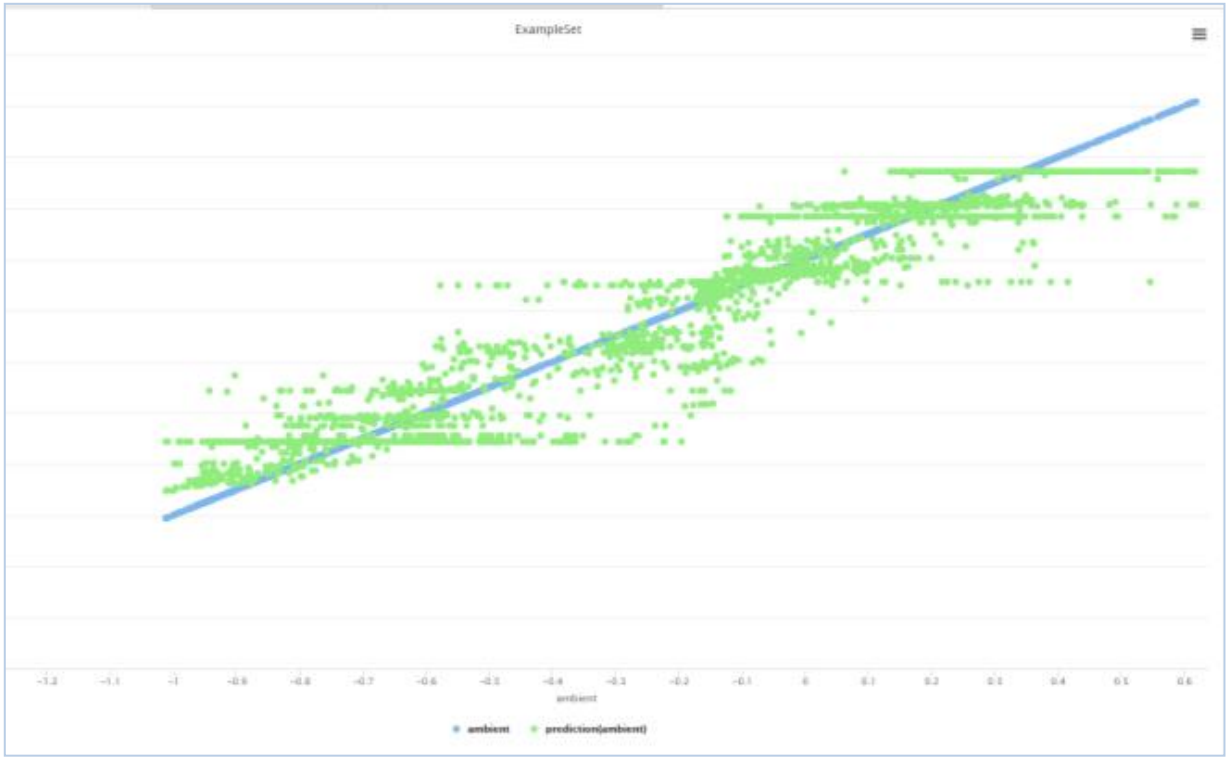
回帰：種々のアルゴリズム結果

Criterion
- absolute error
- relative error
- squared error
- correlation
- squared correlation

correlation

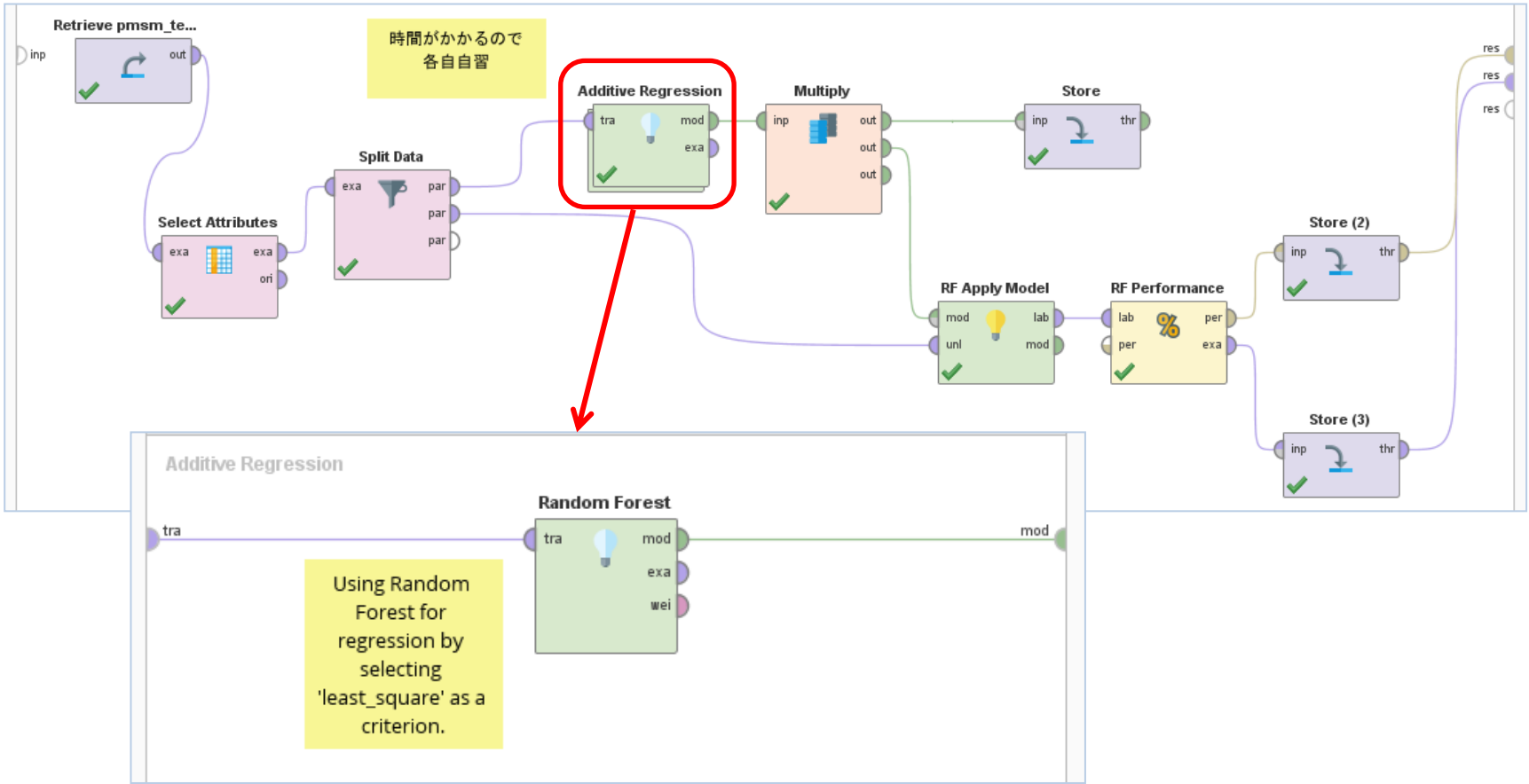
correlation: 0.962

Random Forest



Confidential
期限：2025年12月31日
ヘアケア4研2Gr

回帰：Additive Regression



Confidential
期限：2025年12月31日
ヘアケア4研2Gr

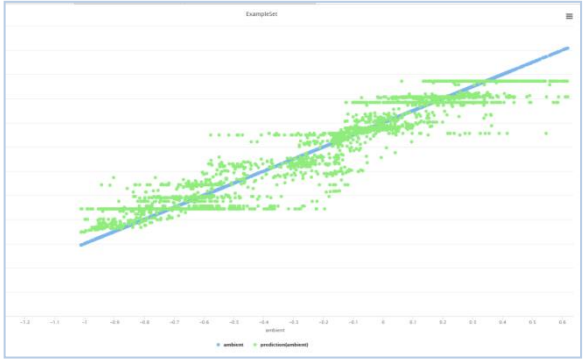
回帰：Additive Regression RF 結果

Criterion
- absolute error
- relative error
- squared error
- correlation
- squared correlation

correlation

correlation: 0.988

RF, correlation: 0.962



Additive Regression RF (10 times)

Confidential
期限：2025年12月31日
ヘアケア4研2Gr

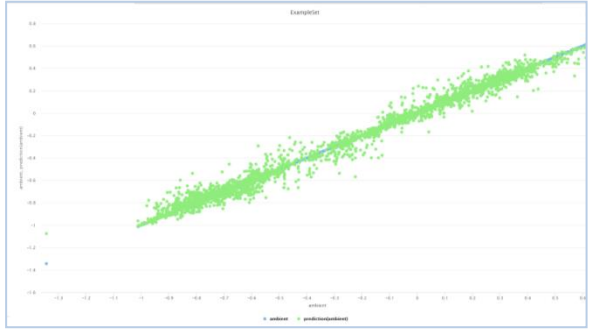
回帰：Additive Regression k-NN 結果

Criterion
- absolute error
- relative error
- squared error
- correlation
- squared correlation

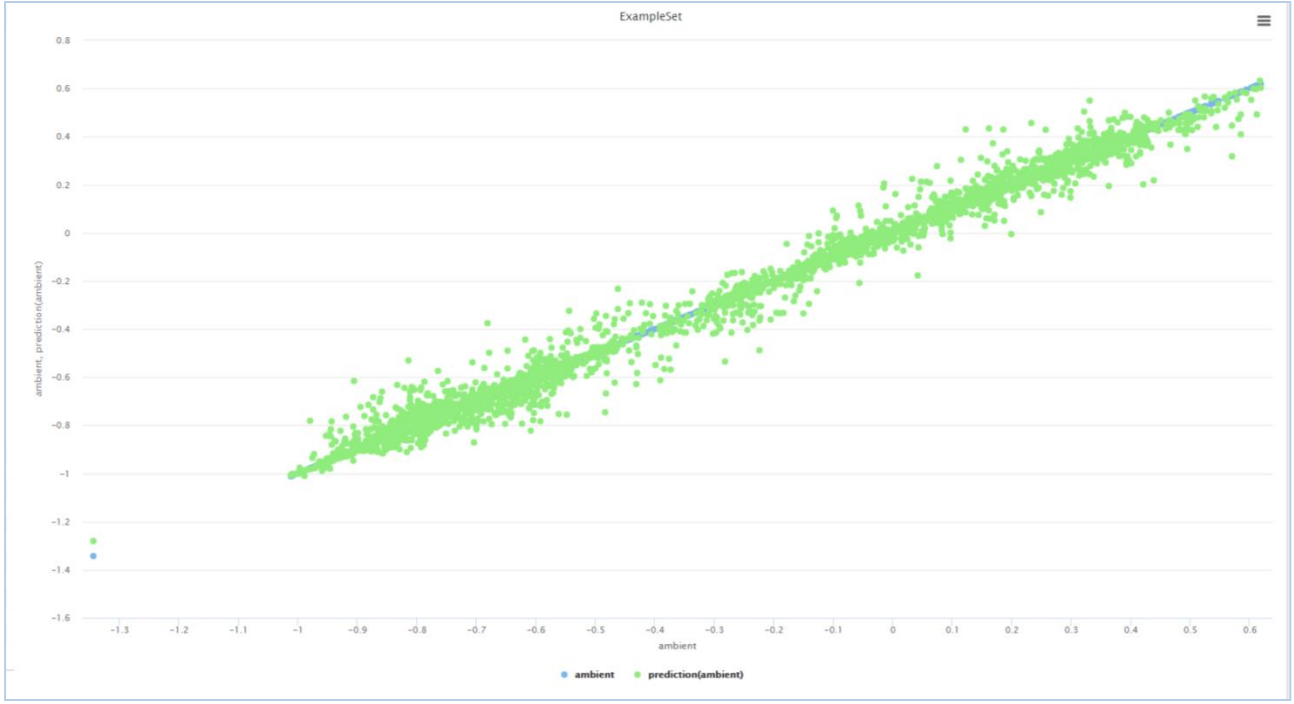
correlation

correlation: 0.994

K-NN, correlation: **0.991**



Additive Regression k-NN (2 times)



Confidential
期限：2025年12月31日
ヘアケア4研2Gr

k-NN による回帰分析

- ✓目的変数の値を推定したいサンプル $\mathbf{x}_{\text{new}}$ について、すべての
  モデル構築用サンプルとの間でユークリッド距離を計算する

- ✓最も距離の近い k 個のサンプルを選択する

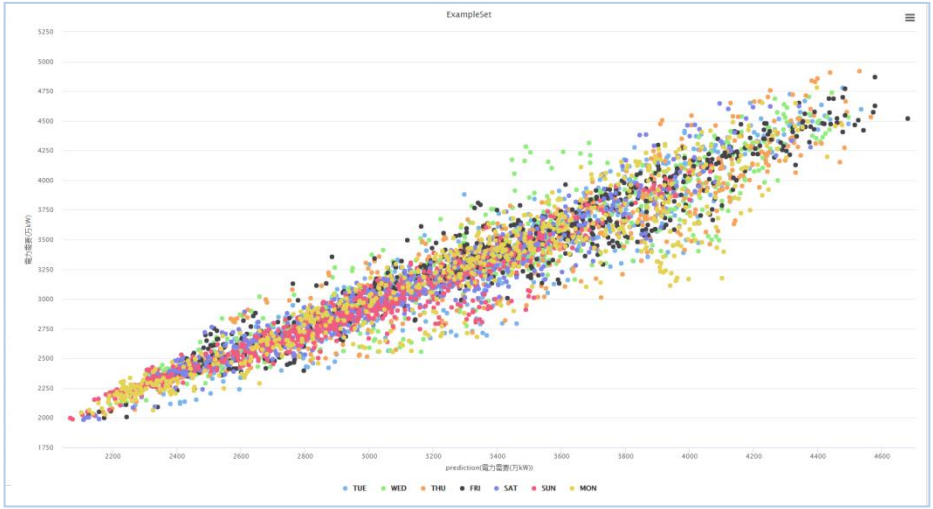
- ✓k 個の目的変数の値の平均値を、$\mathbf{x}_{\text{new}}$ の推定された値とする

- ✓k 個の目的変数の値の標準偏差で推定値の信頼度を検討できる
  - 標準偏差が小さい (k 個の値がばらついていない) 方が、
    標準偏差が大きい (k 個の値がばらついている) 方より
    目的変数の推定値を信頼できる

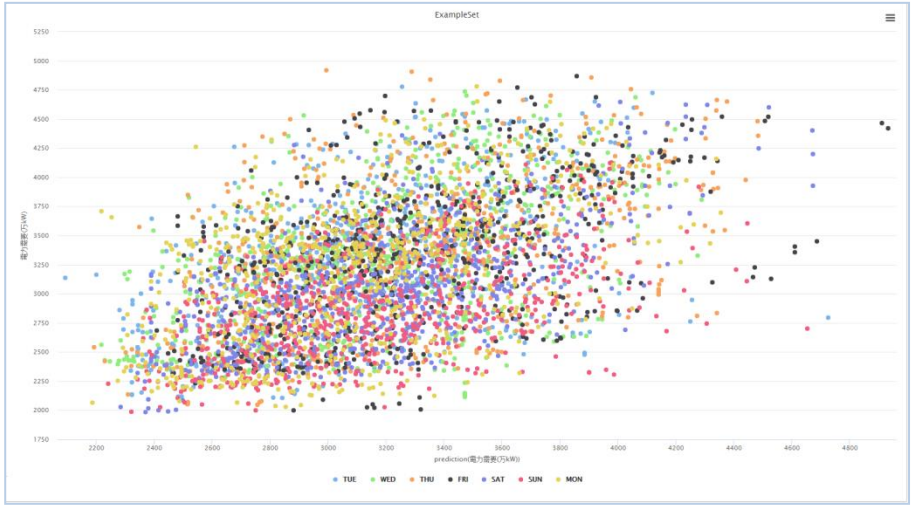
Confidential
期限：2025年12月31日
ヘアケア4研2Gr

K-NN法は回帰問題で
いつも高い精度を出すか？

東京の電力需要予測



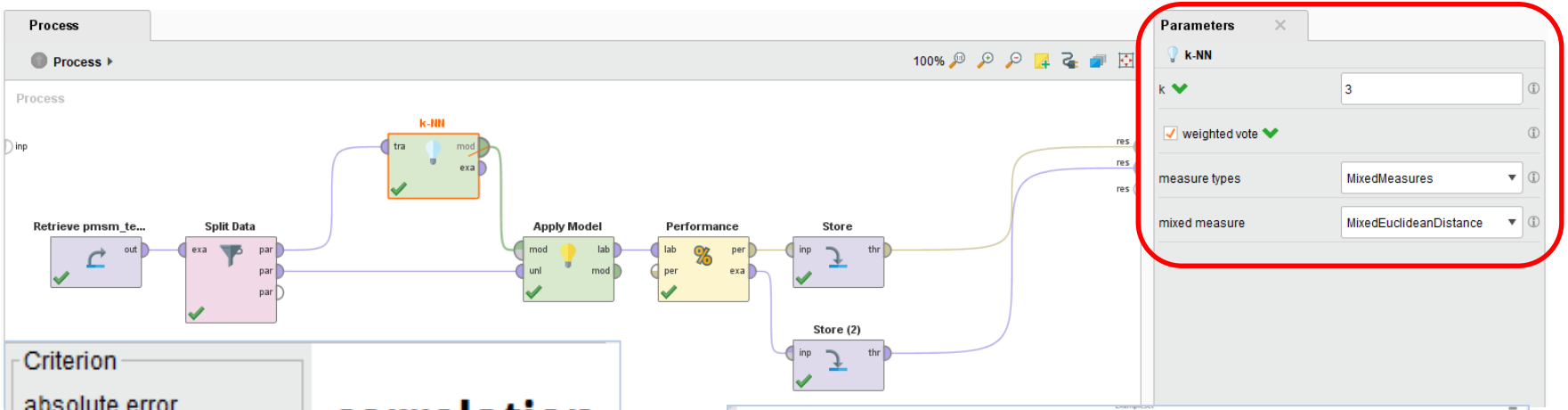
Random Forest (RF)
R = 0.946

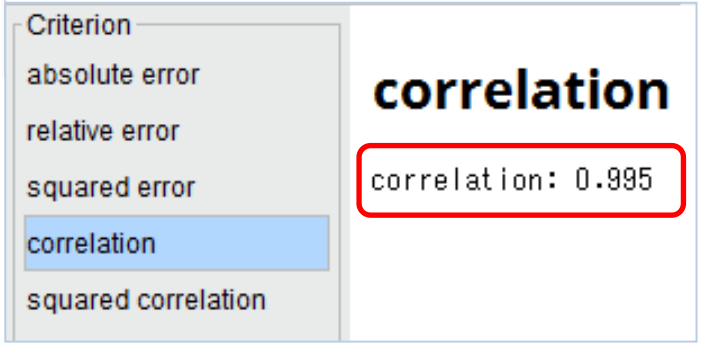


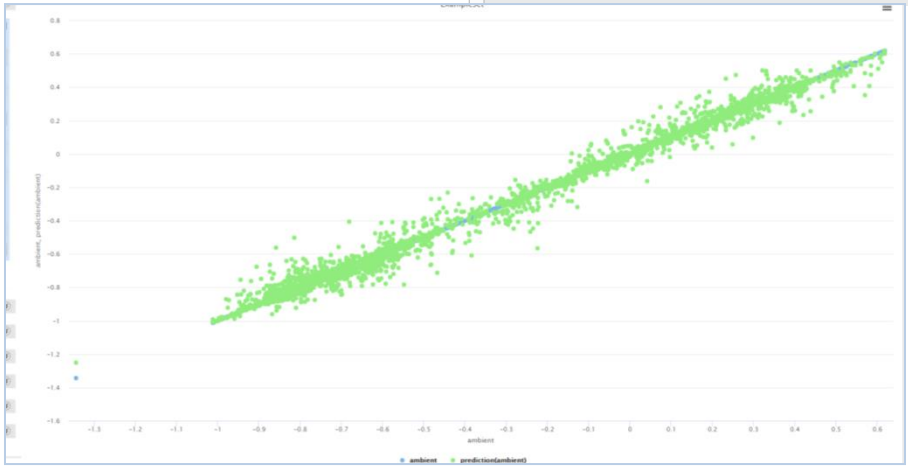
K-NN
R = 0.511

RFは分類・回帰問題共に、常に一定の高い精度を出すが、他のアルゴリズムは
データセットとの相性があって、試してみないとわからない。

回帰：k-NN Hyper parameters 最適化結果







Hyper parametersを最適化することで、
この場合は非常に単純なモデルで
一番高い精度が出た。