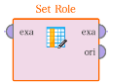
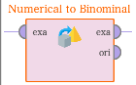
前処理（データ加工・クレンジング）Cheat Sheet

列名の変更と列の役割

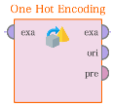
列名の変更を行います。複数の列に対して列名を変更することもできます。Rename by Replacing, Rename by Generic namesも列名を変更するオペレーター。



列の役割を定義します。id, label, prediction, batch, weight, clusterなどを定義できます。Exchange Roleで列の役割を入れ替えることもできます。

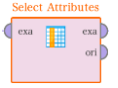


列の型を変更します。Numeric, Integer, Real, Binominal, Polynominal, Nominal, Date, Textなどの型に変更します。

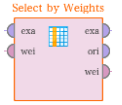


Nominal列（カテゴリ列）を0,1の形に置き換えます。パラメーターのremove with too many valuesにチェックを入れるとNominal列に複数のカテゴリ（例えば10以上）がある列は対象から外します。

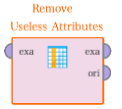
列選択



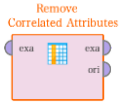
列選択を行います。invert selectionで選択していない列が残ります。include special attributesでlabelなど特別に定義された列が外されます。



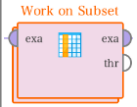
Weight byで算出されたWeight値に基づいて列を絞り込みます。top k, top p%, lessなどの条件で列選択できます。同類のSelect by Randomではランダムで列を選択します。



モデリングに使えそうにない列を削除します。具体的には、データの散らばりを示す標準偏差が小さい値やIDのような役割を持っているnominal値（カテゴリ）です。

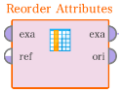


多重共線性を考慮するために相関の強い列同士の一方を削除します。削除する際の敷居値はユーザー側で指定することができます。



サブフローの形になっており、サブフローの中で指定した列の加工を行うことができます。サブフローの中で処理が完了すれば元の残された列に結合され、出力されます。

列の並び替え

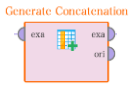


列順をユーザー指定で並び替えることができるオペレーターです。アルファベット順や降順、昇順で並び替えることができます。

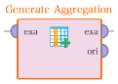
列追加

数式を用いて列同士の計算を行った結果を新しい列として追加します。if関数を用い、条件に合致するもの"T"、合致しないものを"F"とし、新しい列を作ることができます。

ID列を新たに作成します。offsetは開始番号より1つの前の数字をセットします。同類のGenerate Empty Attributeでは、空の列を生成します。



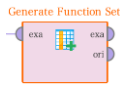
列同士を結合して新たな列を生成します。X列の都道府県名、Y列に市区町村名が入っている場合、X列とY列を結合し新たなZ列ができます。例:(大阪府大阪市西区)



行（列方向）計算の実行結果を新たに追加します。average:平均、SD:標準偏差、max:最大値、min:最小値、sum:合計などを列方向に計算します。

対象列の数値（＋、－を含む）の絶対値を返します。対象列に上書きされます。

first attribute と second attributeを掛算したものが列として追加されます。それぞれ2列ずつ選択すると、2×2＝4列が新たに追加されます。

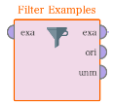


列同士の計算結果（和、差、乗算、除算、逆数、平方根、乗算、sin、cos、接線、円弧接線、絶対値、最小、最大、非可換環など）を新しい列に追加します。

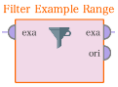
単語の出現頻度から対象単語の重要度を指標化します。各ドキュメントのある単語の出現回数/全体の単語の出現回数で計算されています。（０＜ｘ＜１）

Nominal（カテゴリ）列に対して、あるカテゴリの出現割合から重み（weight）を算出します。total weightはユーザー側で設定でき、その値をカテゴリの頻度に応じて割り振ります。

行選択

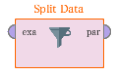


条件に合致する行（or条件, and条件）をフィルタリングします。（例えば、男性かつ30歳以上などの場合）condition classを変更すると、欠損値を含まない行のみを選択したりすることもできます。



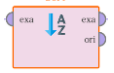
指定行（例えば、5行目～10行目）のみを取り出します。invert filterを選択すると、指定行に合致しない行が取り出されます。

サンプリングを行います。サンプリング手法は、stratified, bootstrapping, Kennard-Stone, Model Basedがあり、不均衡データの調整にも使用されます。



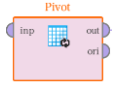
データを分割を行います。partitionsでデータを分割する割合をしています。sampling typeも4つの中から選択することができます。

集計・並び替え



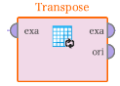
指定した対象列を昇順（increasing）、降順（decreasing）で並び替えます。同類にパレートランキングやシャッフルするオペレーターがあります。

グループごとに集計を行います。（例えば、性別・業種別の年収平均、年収中央値、年収標準偏差など）集計方法は、集計対象ごとで指定します。



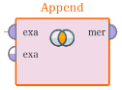
ピボットテーブルを作成します。グループ行、グループ列を選択し、集計方法を選択します。

横持ち型のデータのセットを縦持ち型のデータセットへと変形します。

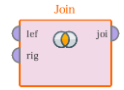


行と列の入れ替えを行います。

結合



データセットA、データセットBがあった場合、行方向に結合を行います。上のポートに繋げたデータセットが上になり、下のポートに繋げたデータセットが下になります。



key列を指定して2つのデータセットを列方向に結合します。結合方法は下記の4パターンがあります。

inner_join

left_join

right_join

outer_join