

Aprendizagem e Extração do Conhecimento

PL03 · CRISP – DM | Preparação dos Dados

Mestrado em Engenharia Biomédica

Hugo Peixoto – hpeixoto@di.uminho.pt

Conteúdo autoral; design e estruturação suportados por IA.

PARTE 1 • PREPARAÇÃO DOS DADOS

CRISP-DM

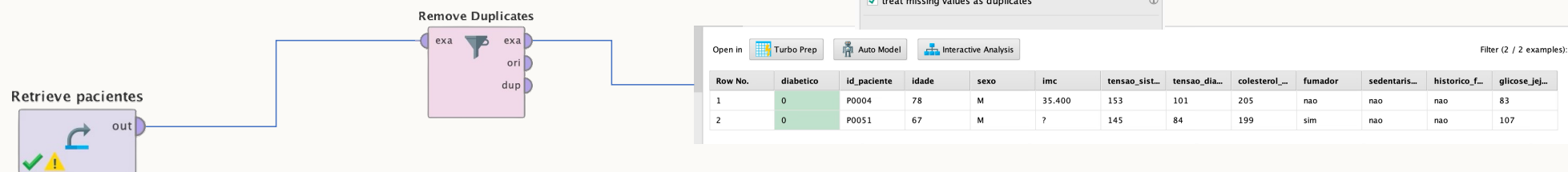


PREPARAÇÃO DOS DADOS

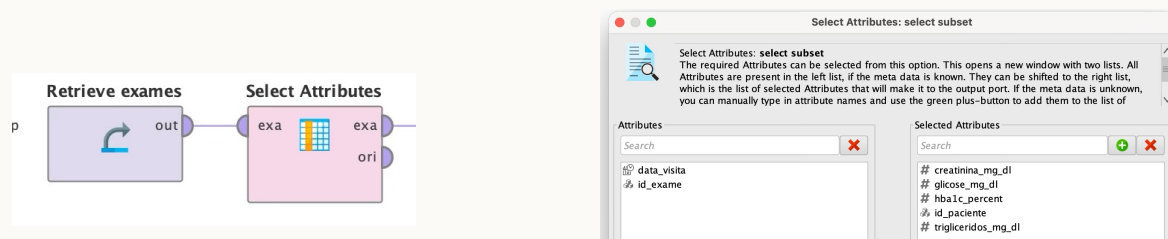
- 01.** Carregar os dados e remover duplicados
- 02.** Normalizar categorias inconsistentes
- 03.** Tratar valores impossíveis
- 04.** Lidar com valores em falta
- 05.** Normalizar o dataset
- 06.** Detetar e remover outliers
- 07.** Agregar e juntar os dois datasets
- 08.** Guardar o dataset final

PREPARAÇÃO DOS DADOS

01. Carregar os dados e remover duplicados



Nos exames não é necessária a data da visita nem o id_exame



PREPARAÇÃO DOS DADOS

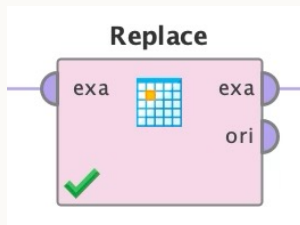
02. Normalizar categorias inconsistentes

- Verificar nas Estatísticas:
 - sexo tem 4 valores (M, F, Masculino, Feminino);
 - fumador tem 3 valores (sim, nao, Nao).
- Adicionar " Map" sobre sexo: Masculino → M, Feminino → F.
- Adicionar outro "Replace " ou "Map" sobre fumador: Nao → nao.
- Confirmar nas Estatísticas que cada atributo ficou só com 2 valores.

PREPARAÇÃO DOS DADOS

02. Normalizar categorias inconsistentes

- Verificar nas Estatísticas:
 - fumador tem 3 valores (sim, nao, Nao).



Parameters ✕

 **Replace**

attribute filter type single ⓘ

attribute fumador ⓘ

☐ invert selection ⓘ

☐ include special attributes ⓘ

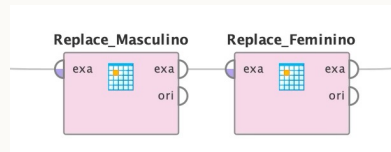
replace what Nao ⓘ

replace by nao ⓘ

PREPARAÇÃO DOS DADOS

02. Normalizar categorias inconsistentes

- Verificar nas Estatísticas:
 - sexo tem 4 valores (M, F, Masculino, Feminino);



Parameters ×

Replace_Masculino (Replace)

attribute filter type all ⓘ

☐ invert selection ⓘ

☐ include special attributes ⓘ

replace what Masculino ⓘ

replace by M ⓘ

Parameters ×

Replace_Feminino (Replace)

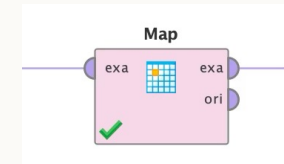
attribute filter type all ⓘ

☐ invert selection ⓘ

☐ include special attributes ⓘ

replace what Feminino ⓘ

replace by F ⓘ



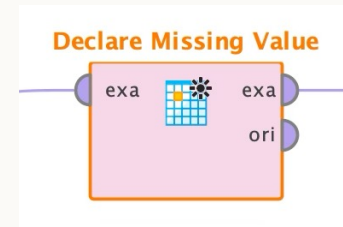
Edit Parameter List: value mappings

Edit Parameter List: value mappings
Multiple mappings can be specified through this parameter. If only a single mapping is required. It can be done using the parameters *replace what* and *replace by* as in the *Replace* operator. Old values and new values can be easily specified through this parameter. Multiple mappings can be defined for the same old value but only the new value corresponding to the first mapping is taken as replacement.

old values	new value
Masculino	M
Feminino	F

PREPARAÇÃO DOS DADOS

03. Tratar Valores Impossíveis



Idade

Parameters

Declare Missing Value

attribute filter type: single

attribute: idade

☐ invert selection

☐ include special attributes

mode: expression

expression value: idade > 120

imc

Parameters

Declare Missing Value (2) (Declare Missing Value)

attribute filter type: single

attribute: imc

☐ invert selection

☐ include special attributes

mode: expression

expression value: imc < 1

creatinina

Parameters

Declare Missing Value (3) (Declare Missing Value)

attribute filter type: single

attribute: creatinina

☐ invert selection

☐ include special attributes

mode: expression

expression value: creatinina < 0

hba1c

Parameters

Missing_hba1c (Declare Missing Value)

attribute filter type: single

attribute: hab1c

☐ invert selection

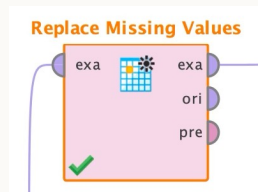
☐ include special attributes

mode: expression

expression value: hab1c > 15

PREPARAÇÃO DOS DADOS

04. Lidar com valores nulos



Parameters ✕

 **Replace Missing Values**

attribute filter type all ⓘ

☐ invert selection ⓘ

☐ include special attributes ⓘ

default average ⓘ

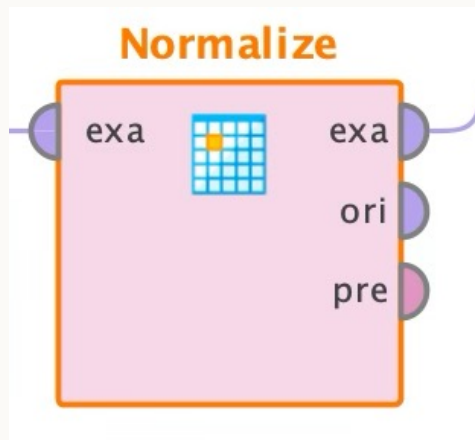
columns  Edit List (0)... ⓘ

Name	Type	Missing
Label ▼ diabetico	Binominal	0
▼ fumador	Nominal	0
▼ sexo	Nominal	0
▼ idade	Integer	0
▼ imc	Real	0
▼ tensao_sistolica	Real	0
▼ tensao_diastolica	Real	0
▼ colesterol_total	Real	0
▼ sedentarismo	Nominal	0
▼ historico_familiar_diabetes	Nominal	0
▼ glicose_jejum	Real	0
▼ creatinina	Real	0
▼ glicose	Real	0
▼ hab1c	Real	0
▼ trigliceridos	Real	0

PREPARAÇÃO DOS DADOS

05. Normalizar

Valores numéricos usado o z-transform



Parameters ✕

 **Normalize**

attribute filter type all ⓘ

☐ invert selection ⓘ

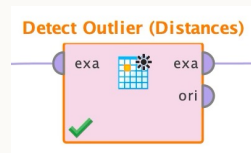
☐ include special attributes ⓘ

method Z-transformation ⓘ

PREPARAÇÃO DOS DADOS

06. Detetar e remover outliers

Coloca 10 outliers por defeito. Porquê?



Parameters 

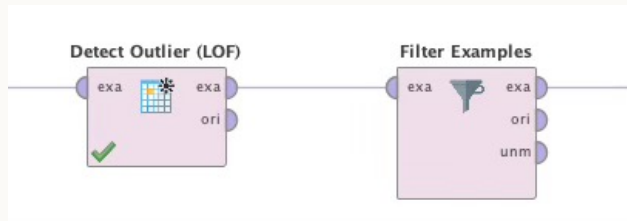
 **Detect Outlier (Distances)**

number of neighbors	10	
number of outliers	10	
distance function	euclidian distance	

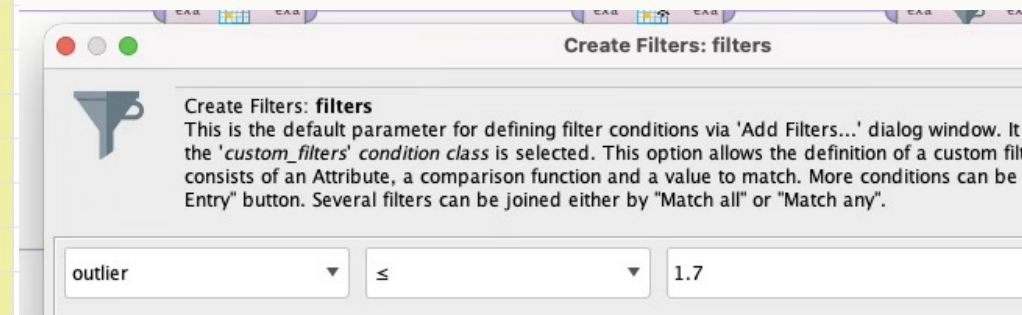
PREPARAÇÃO DOS DADOS

06. Detetar e remover outliers

Avaliar a distância dos outliers!



3	0	1.206
299	0	1.269
124	0	1.288
2	0	1.303
204	0	1.337
1	0	1.347
111	1	1.381
300	0	1.402
205	0	1.546
157	0	1.752
275	0	2.103
8	0	2.343
113	0	2.970



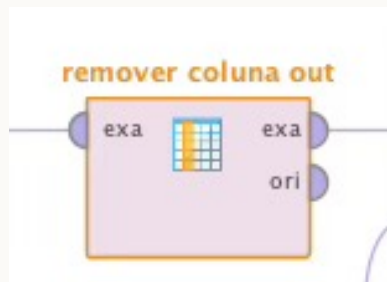
Mais fácil perceber o salto e as distâncias!

Em vez dos 10 por defeito do "*Detect Outlier (Distances)*", os dados aqui apoiam um corte por score de *LOF*, não por contagem fixa. "outlier se $LOF > 1,7$ " apanharia só os últimos 4 valores (1,752, 2,103, 2,343, 2,970), que são precisamente os candidatos a ser outlier!

PREPARAÇÃO DOS DADOS

06. Detetar e remover outliers

Depois de detetar e remover os outliers devemos remover a coluna outlier!
Operador “Select Attributes”!



Parameters ✕

 **remover coluna out (Select Attributes)**

type exclude attributes ▼ ⓘ

attribute filter type one attribute ▼ ⓘ

select attribute outlier ▼ ⓘ

☒ also apply to special attributes (id, label..) ⓘ

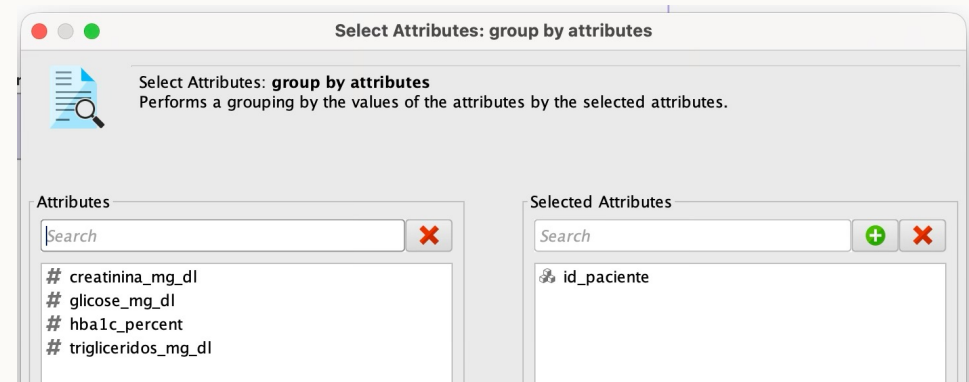
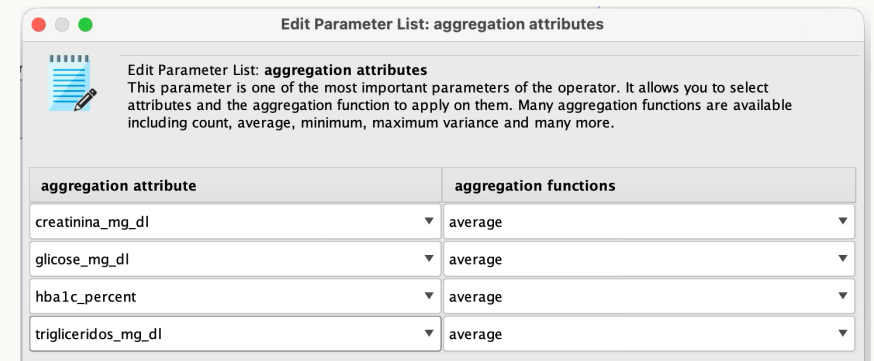
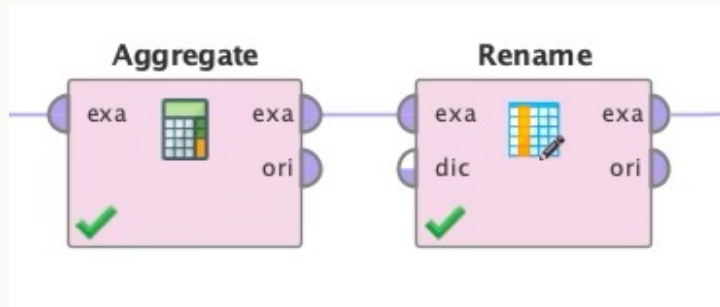
PREPARAÇÃO DOS DADOS

07. Agregar e juntar os dois datasets

- É necessário um conjunto de passos antes de fazer o join dos 2 datasets!
 - **Como lidar com vários resultados por paciente?**
 - **Existem atributos desnecessários?**

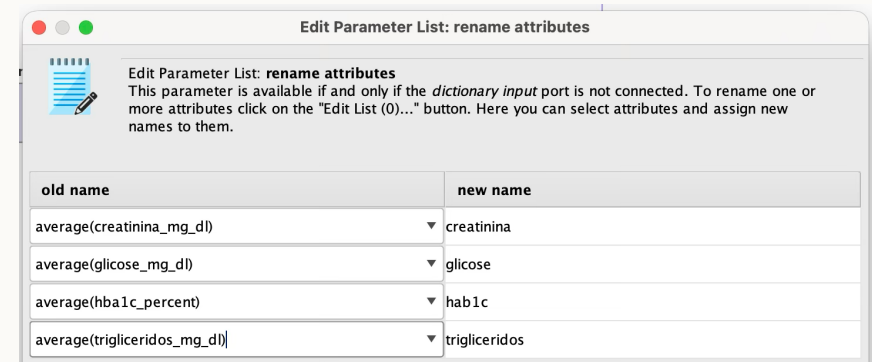
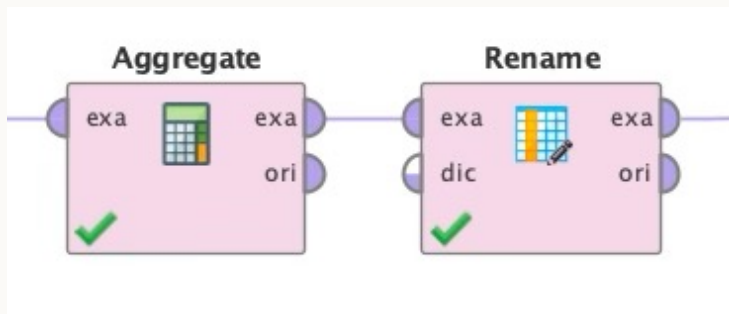
PREPARAÇÃO DOS DADOS

07. Agregar e juntar os dois datasets



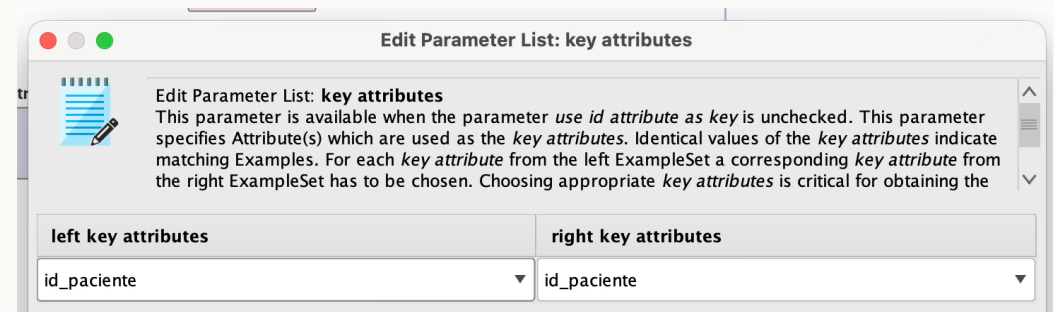
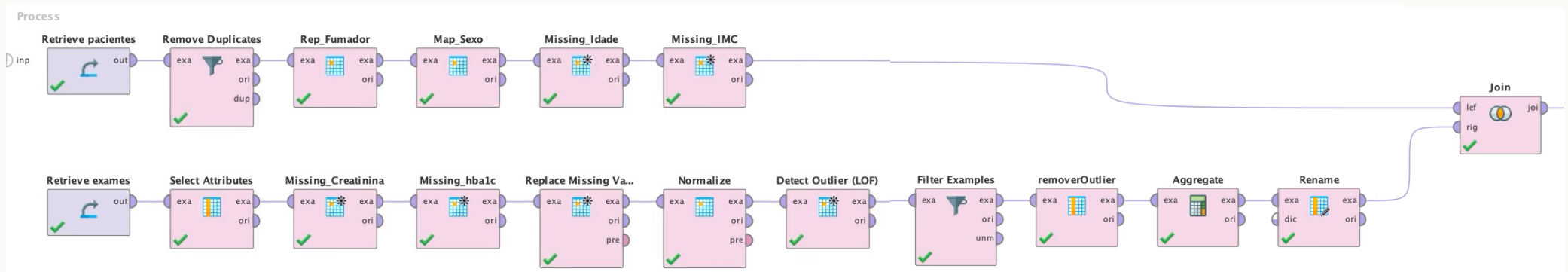
PREPARAÇÃO DOS DADOS

07. Agregar e juntar os dois datasets



PREPARAÇÃO DOS DADOS

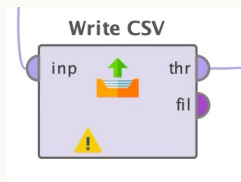
07. Agregar e juntar os dois datasets



PREPARAÇÃO DOS DADOS

08. Guardar o dataset final

- Podemos exportar o dataset!



Parameters ✕

 **Write CSV**

csv file  ⓘ

column separator ⓘ

☒ write attribute names ⓘ

☒ quote nominal values ⓘ

☒ format date attributes ⓘ

date format  ⓘ

☐ append to file ⓘ

encoding ⓘ

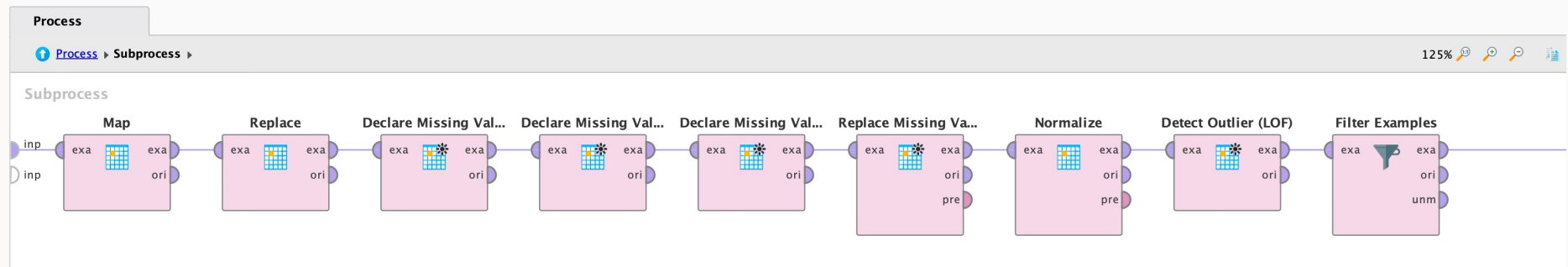
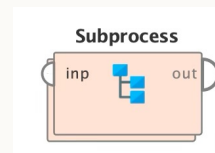
 [Hide advanced parameters](#)

 [Change compatibility \(12.0.001\)](#)

PREPARAÇÃO DOS DADOS

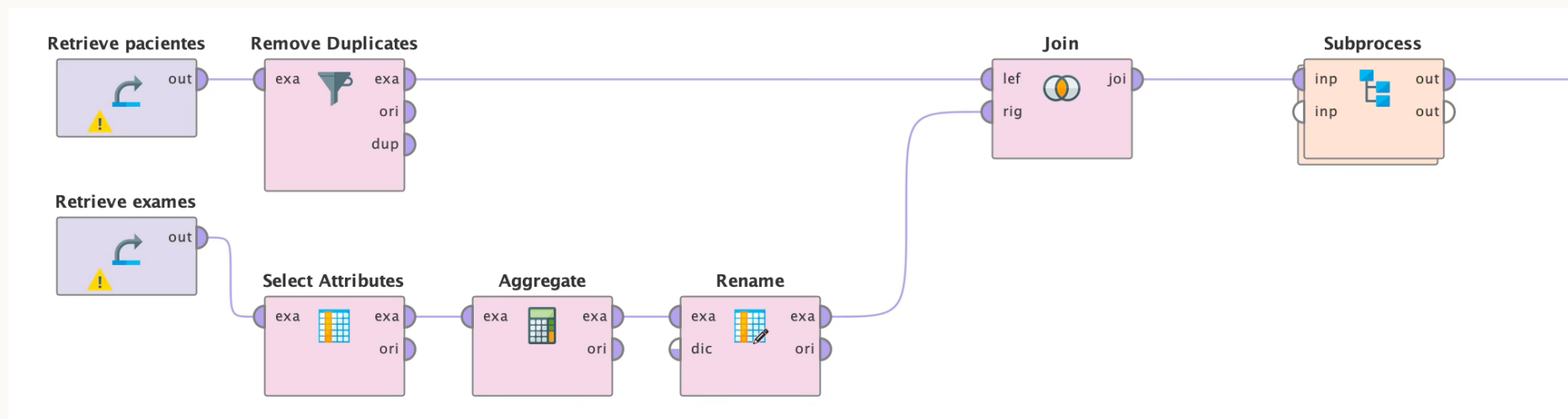
08. Guardar o dataset final

- Ou criar um subprocesso que englobe a preparação dos dados!



PREPARAÇÃO DOS DADOS

PROCESSO FINAL



NÃO ESQUECER DE GUARDAR O PROCESSO NO AI STUDIO!

Aprendizagem e Extração do Conhecimento

PL03 · CRISP – DM | Preparação dos Dados

Mestrado em Engenharia Biomédica

Hugo Peixoto – hpeixoto@di.uminho.pt

Conteúdo autoral; design e estruturação suportados por IA.