# Aprendizagem e Extração do Conhecimento

*PL04* · CRISP – DM | Construção de Modelos

Mestrado em Engenharia Biomédica

Hugo Peixoto – hpeixoto@di.uminho.pt

*Conteúdo autoral; design e estruturação suportados por IA.*

# O caminho de hoje

**PARTE 1 – CONSTRUÇÃO DE MODELOS**

- Correlações
- Regras de Associação
- Regressão Linear

# PARTE I · CONSTRUÇÃO DE MODELOS

# CRISP-DM


01
Compreensão do Negócio
✓ concluído
02
Compreensão dos Dados
✓ concluído
03
Preparação dos Dados
✓ concluído
04
Construção dos Modelos
← estamos aqui
05
Avaliação de Performance

# CORRELAÇÕES

# Correlações

**Antes de começar:**

- Criar 2 datasets para o pacientes.csv:

Diabético = inteiro

Diabético = binominal

Porquê ?!

# Correlações

## Coeficientes de correlação:

]0, 1] - Correlações Positivas

[-1, 0[ - Correlações Negativas

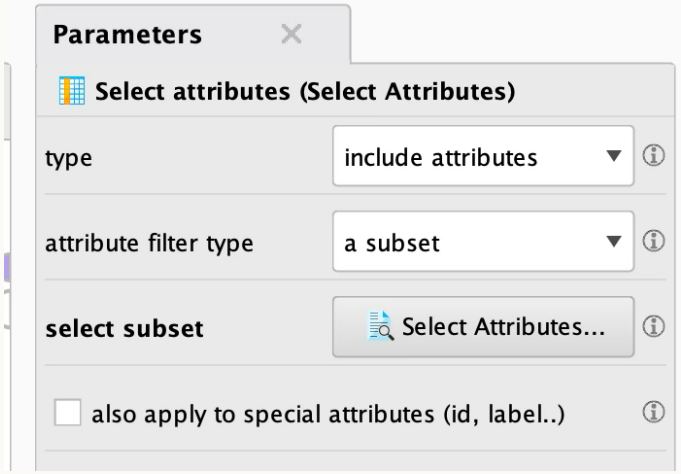
Correlações positivas significam, por um lado, que à medida que o valor de um atributo aumenta, o valor do outro atributo também aumenta.

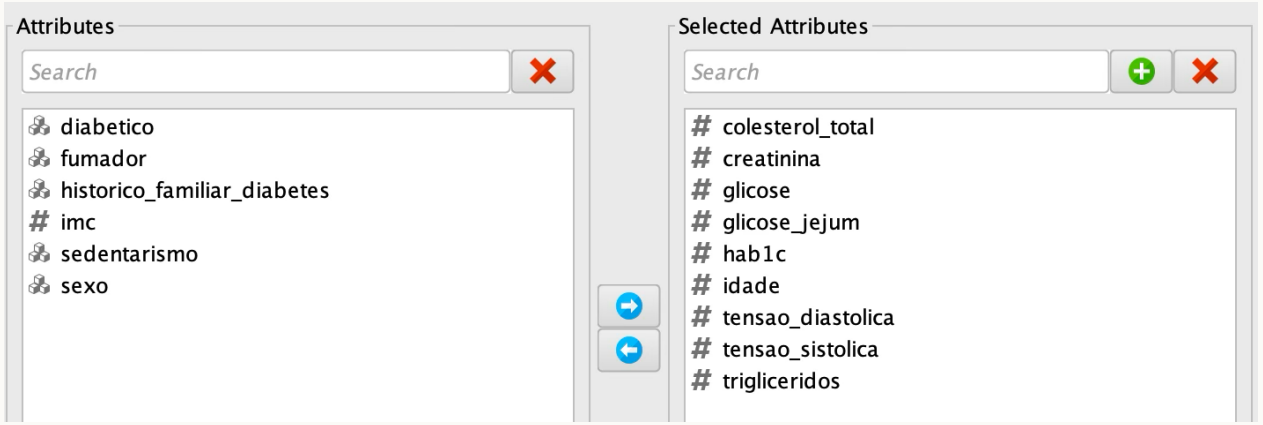
Por outro lado, uma correlação positiva também pode ser encontrada quando à medida que o valor de um atributo diminui, o valor do outro atributo também diminui.

# Correlações

## Preparar o ambiente:

- Apenas atributos numéricos!

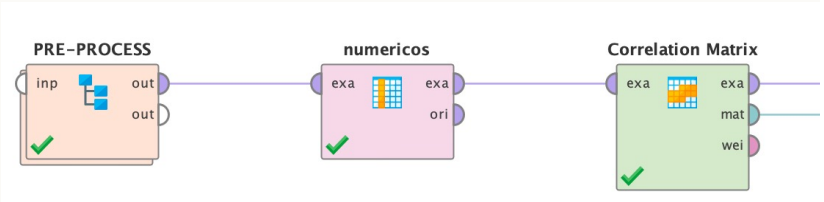


# Correlações

## Modelo:

Selecionar o operador ”*Correlation Matrix*” e correr!




| Attribu... | idade | imc | tensao... | tensao... | coleste... | glicose... | diabeti... | averag... | averag... | averag... | averag... |
|---|---|---|---|---|---|---|---|---|---|---|---|
| idade | 1 | -0.023 | 0.056 | 0.048 | -0.032 | -0.035 | -0.056 | -0.021 | -0.029 | -0.054 | 0.044 |
| imc | -0.023 | 1 | -0.038 | -0.022 | -0.008 | 0.103 | 0.105 | -0.023 | 0.098 | 0.026 | 0.061 |
| tensao_... | 0.056 | -0.038 | 1 | 0.081 | 0.003 | 0.080 | -0.010 | 0.012 | 0.081 | 0.054 | -0.091 |
| tensao_... | 0.048 | -0.022 | 0.081 | 1 | -0.086 | -0.059 | -0.072 | 0.019 | -0.053 | -0.052 | 0.058 |
| colester... | -0.032 | -0.008 | 0.003 | -0.086 | 1 | 0.025 | -0.018 | -0.013 | 0.026 | 0.027 | 0.062 |
| glicose_... | -0.035 | 0.103 | 0.080 | -0.059 | 0.025 | 1 | 0.654 | -0.015 | 0.932 | 0.710 | -0.077 |
| diabetico | -0.056 | 0.105 | -0.010 | -0.072 | -0.018 | 0.654 | 1 | 0.019 | 0.609 | 0.447 | -0.010 |
| average... | -0.021 | -0.023 | 0.012 | 0.019 | -0.013 | -0.015 | 0.019 | 1 | -0.035 | -0.025 | 0.062 |
| average... | -0.029 | 0.098 | 0.081 | -0.053 | 0.026 | 0.932 | 0.609 | -0.035 | 1 | 0.778 | -0.017 |
| average... | -0.054 | 0.026 | 0.054 | -0.052 | 0.027 | 0.710 | 0.447 | -0.025 | 0.778 | 1 | 0.023 |
| average... | 0.044 | 0.061 | -0.091 | 0.058 | 0.062 | -0.077 | -0.010 | 0.062 | -0.017 | 0.023 | 1 |

# Correlações

**Correlações de Pearson:**

As correlações de Pearson são estatisticamente corretas para valores binominais de 0 ou 1.



**Não são corretas variáveis categóricas com mais de 2 categorias sem ordem natural !**

Um coeficiente de 0,65 entre *glicose_jejum* e *diabetico* significaria que valores mais altos de glicose em jejum tendem a estar associados ao grupo *diabetico = 1*.

# Correlações

Quando os valores dos atributos se movem na mesma direção, a correlação é positiva.

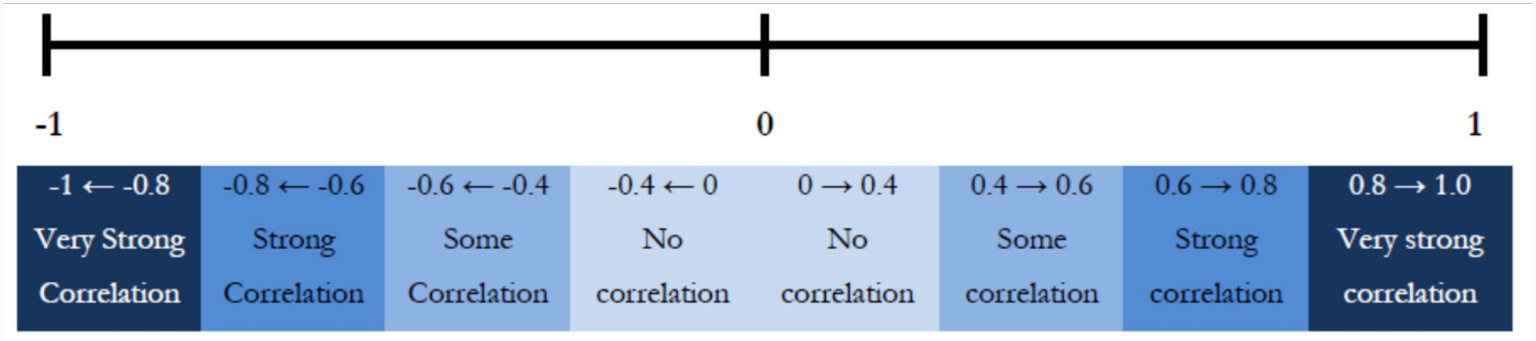
| | | | | |
|---|---|---|---|---|
|  Glicose Sobe |  Glicose em jejum sobe | |  Glicose desce |  Glicose em jejum desce |

Quando os valores dos atributos se movem em direções opostas, a correlação é negativa.

| | | | | |
|---|---|---|---|---|
|  Tensao diastólica desce |  Colestrol total sobe | |  Tensao diastólica sobe |  Colestrol total desce |

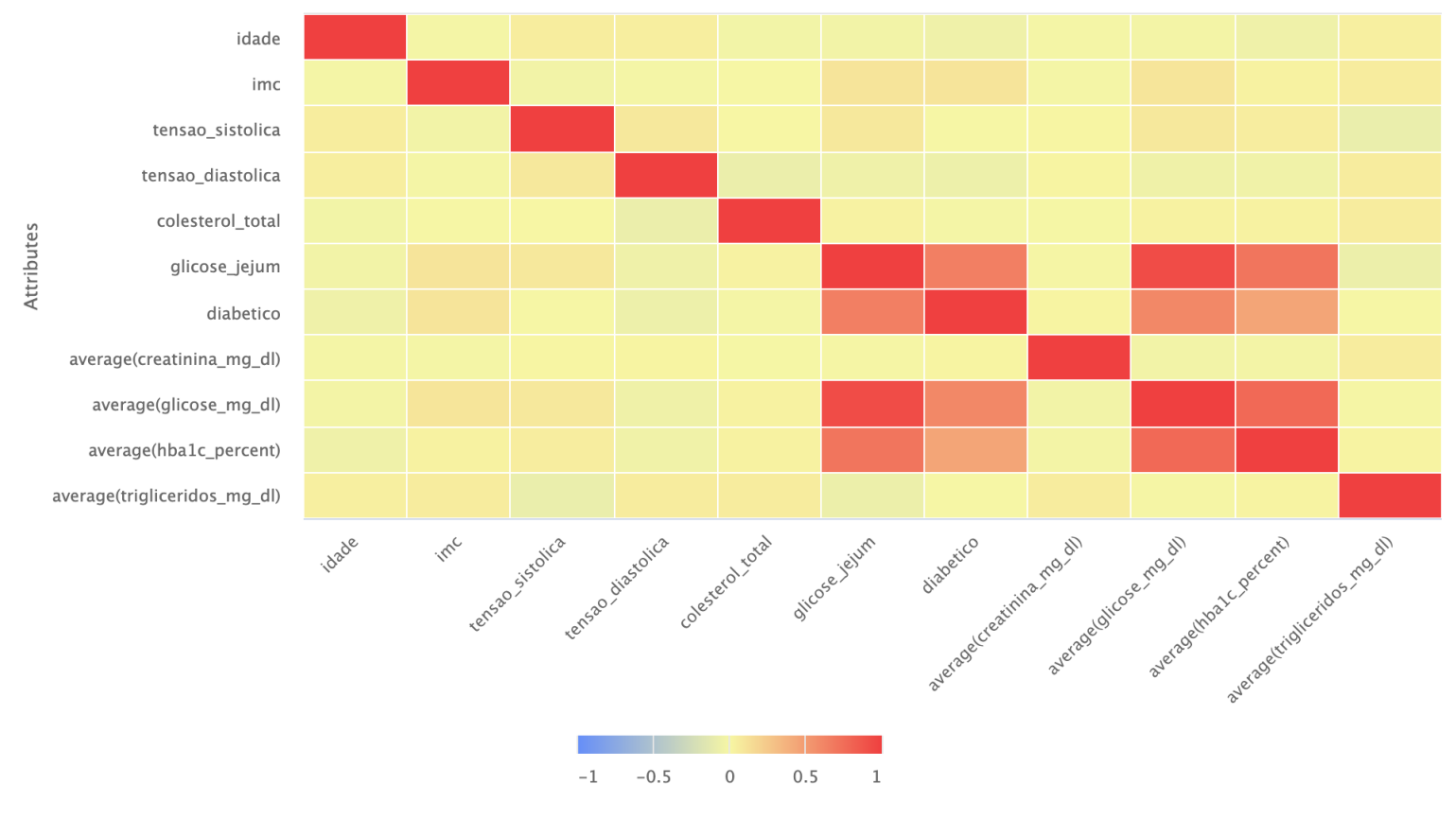
# Correlações

Os coeficientes de correlação não permitem apenas determinar a relação entre atributos, mas também nos dizem algo sobre a **força** da correlação




Quanto mais próximo um coeficiente de correlação estiver de 1 ou de -1, mais forte é a correlação dos atributos.

# Correlações



# Correlações

| Atributo | Correlação com `diabetico` |
|---|---|
| `glicose_jejum` | **0,654** |
| `average(glicose_mg_dl)` | **0,609** |
| `average(hba1c_percent)` | **0,447** |
| `imc` | 0,105 |
| `average(creatinina_mg_dl)` | 0,019 |
| `average(trigliceridos_mg_dl)` | -0,010 |
| `tensao_sistolica` | -0,010 |
| `colesterol_total` | -0,018 |
| `idade` | -0,056 |
| `tensao_diastolica` | -0,072 |

- Os três atributos de glicose dominam, e por uma boa razão clínica.
- encontrar uma correlação alta entre *glicose_jejum* e *diabetico* não é propriamente "Data Mining a encontrar um padrão escondido". É o modelo a redescobrir o critério clínico que já existe!
- Os restantes 7 atributos têm correlação quase nula com *diabetico*. Isto não significa que sejam inúteis - só significa que, isoladamente e de forma linear, não se relacionam com o diagnóstico!

# Correlações

- *glicose_jejum* (0,654) e *glicose* (0,609) e *hab1c* (0,447) estão também muito correlacionados entre si

- Os três medem essencialmente o mesmo fenómeno!

# Correlações



**A suposição de que uma correlação prova causalidade é perigosa e muitas vezes falsa**

Outra falsa interpretação é que os coeficientes de correlação são percentagens(%).

Um coeficiente de correlação de 0,930 ≠ 93,0% de variabilidade entre esses atributos.

A fórmula matemática subjacente ao cálculo dos coeficientes de correlação mede apenas a força, como indicado pela proximidade de 1 ou -1, da interação entre os atributos.

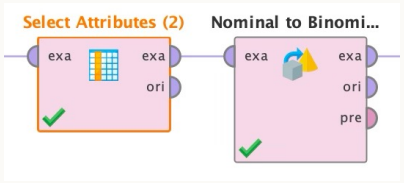
# Correlações – Quizz time!

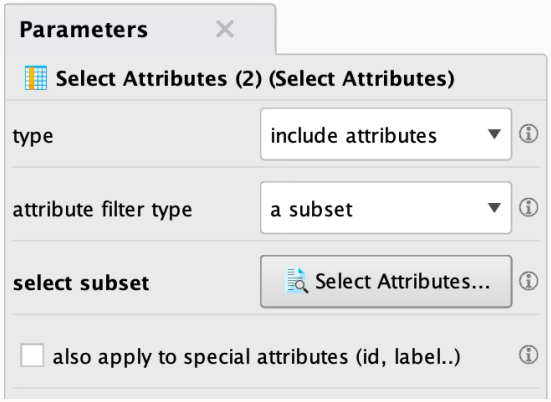
**Se um modelo de classificação usar só glicose_jejum, que accuracy seria de esperar, e isso seria um bom modelo clínico ou só estaria a replicar a definição do diagnóstico?**

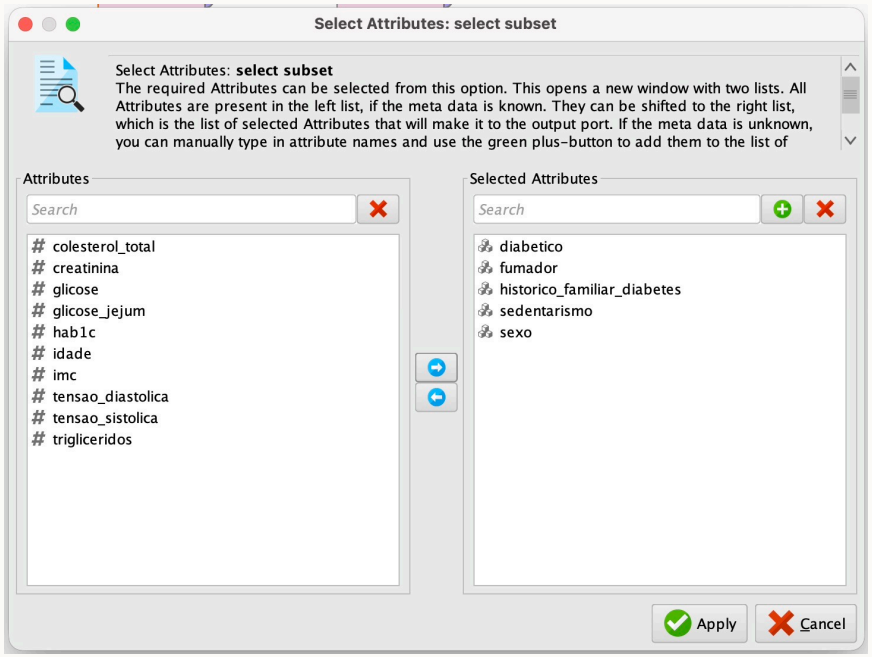
# REGRAS DE ASSOCIAÇÃO

# Regras de Associação

- As Regras de Associação funcionam com valores binominais (Sim/Não).
- Selecionar os atributos necessários e converter nominais para binominais (caso necessário!)

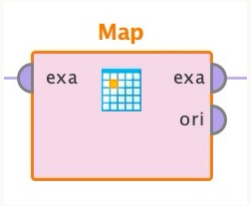


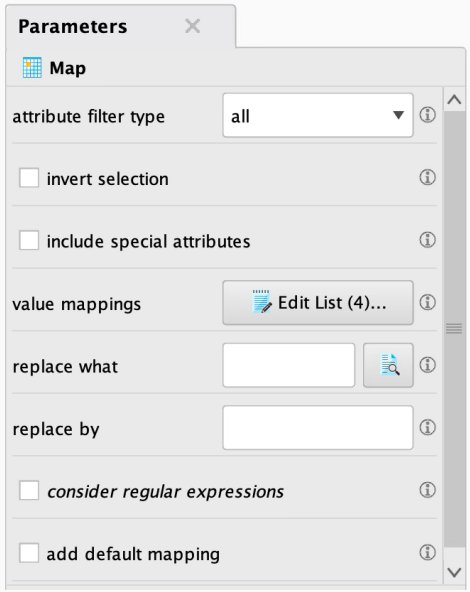



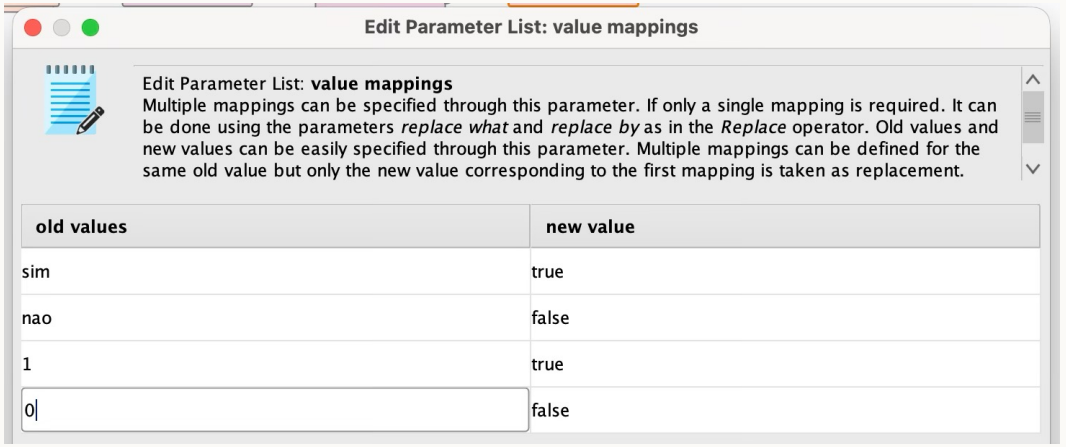



# Regras de Associação

- Mapear os “sim” -> true, “não” -> false, ”1” –>true, ”0” -> false

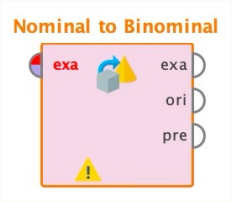


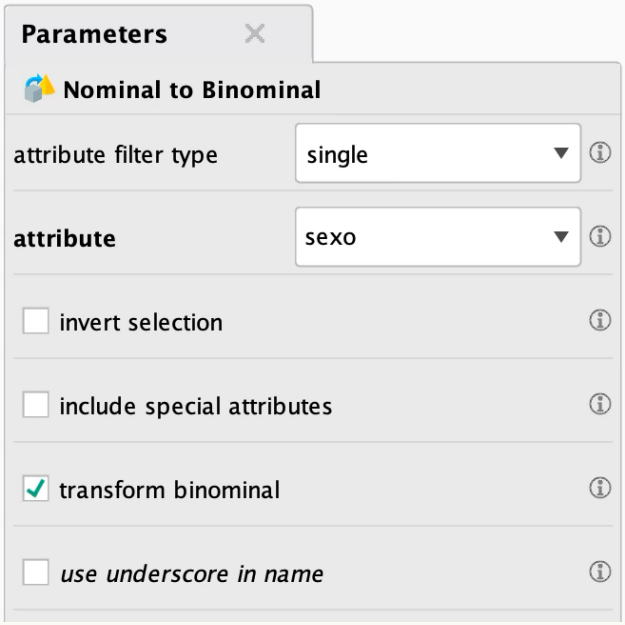



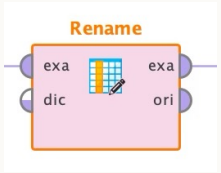
# Regras de Associação

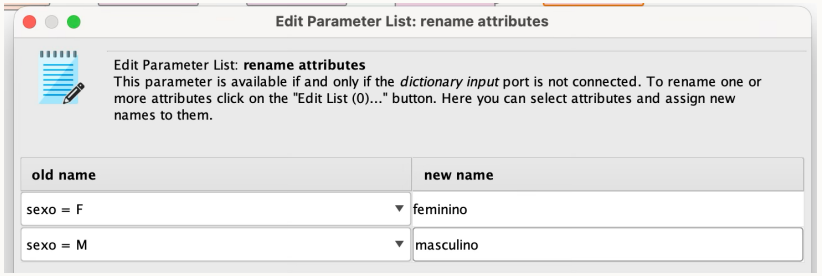
- Converter o sexo para 2 colunas! Masculino e Feminino (Transform binominal)
- Colocar false e true se necessário!



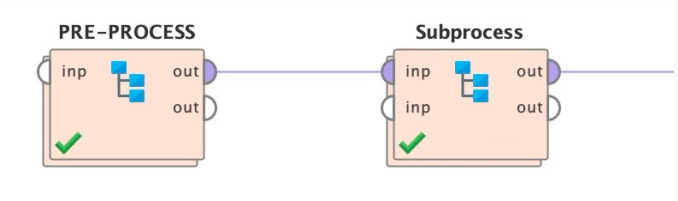

- Renomear as colunas



# Regras de Associação

- Criar um subprocesso para armazenar o dataset pronto!



# Regras de Associação

## O que vamos encontrar? Itemsets:

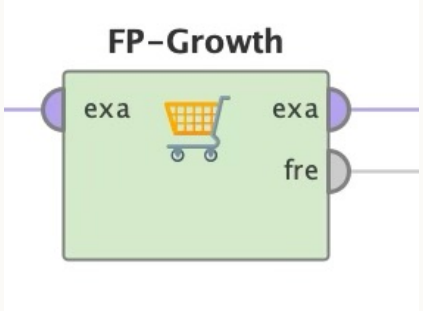
- **O que é um itemset?**

Um conjunto de itens (atributos com um valor específico) que aparecem juntos no mesmo registo. No nosso caso, um item é, por exemplo, sedentarismo = true ou diabetico = true; um itemset é uma combinação destes, como {sedentarismo = true, diabetico = true}.

**SUPORTE:** A proporção de registos (pacientes) onde esse itemset aparece por completo.

- Um itemset de tamanho 1 (um único item) é só a frequência desse item isolado;
- Um itemset de tamanho 2 ou 3 mede quantas vezes essa combinação específica acontece junto.

# Regras de Associação




- Percorre o dataset e encontra todos os itemsets **frequentes -** ou seja, todas as combinações de itens cujo suporte ultrapassa um limiar mínimo definido (min support).
- Em vez de testar todas as combinações possíveis uma a uma (o que seria muito lento), o FP-Growth organiza os dados numa estrutura compacta (uma "FP-Tree") que lhe permite encontrar essas combinações de forma muito mais eficiente.

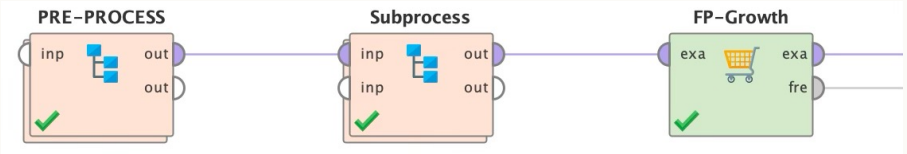
# Regras de Associação

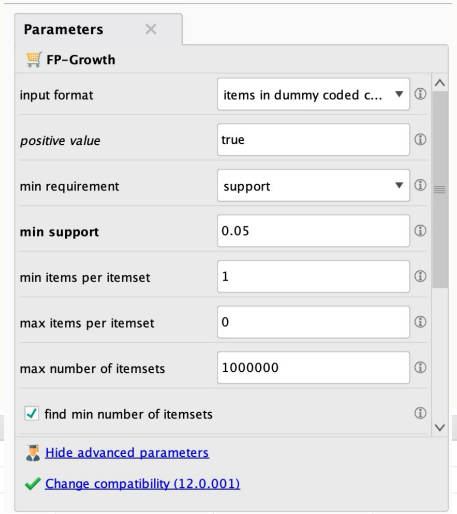
- **Porque é só o primeiro passo, não o resultado final?**

O FP-Growth dá os itemsets frequentes mas não diz "se A, então B". Essa transformação, de itemsets frequentes para regras com direção, confiança e lift, é feita a seguir pelo **"Create Association Rules"**, que usa os itemsets do FP-Growth como matéria-prima.

# Regras de Associação

- Encontrar os itemSets mais frequentes!!



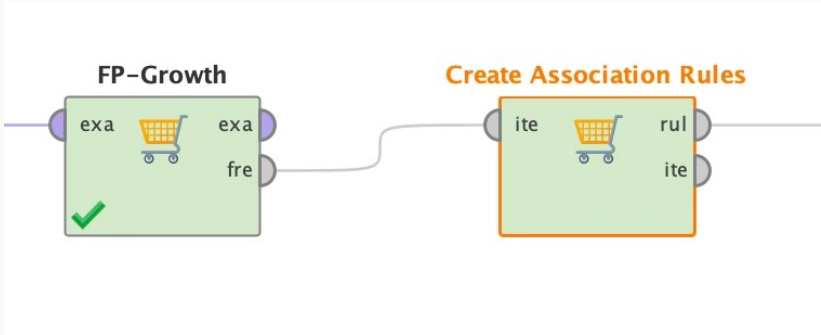

| Size | Support | Item 1 | Item 2 | | | |
|---|---|---|---|---|---|---|
| 1 | 0.517 | feminino | | | | |
| 1 | 0.483 | masculino | | | | |
| 1 | 0.412 | sedentarismo | | | | |
| 1 | 0.297 | fumador | | | | |
| 1 | 0.260 | historico_familiar_diabetes | | | | |
| 1 | 0.132 | diabetico | | | | |
| 2 | 0.216 | feminino | sedentarismo | | | |
| 2 | 0.159 | feminino | fumador | | | |
| 2 | 0.149 | feminino | historico_familiar_diabetes | | | |
| 2 | 0.078 | feminino | diabetico | | | |
| 2 | 0.196 | masculino | sedentarismo | | | |
| 2 | 0.139 | masculino | fumador | | | |
| 2 | 0.111 | masculino | historico_familiar_diabetes | | | |
| 2 | 0.054 | masculino | diabetico | | | |
| 2 | 0.115 | sedentarismo | fumador | | | |
| 2 | 0.125 | sedentarismo | historico_familiar_diabetes | | | |
| 2 | 0.084 | sedentarismo | diabetico | | | |
| 2 | 0.091 | fumador | historico_familiar_diabetes | | | |
| 2 | 0.088 | fumador | diabetico | | | |
| 2 | 0.071 | historico_familiar_diabetes | diabetico | | | |
| 3 | 0.064 | feminino | sedentarismo | fumador | | |
| 3 | 0.074 | feminino | sedentarismo | historico_familiar_diabetes | | |
| 3 | 0.044 | feminino | sedentarismo | diabetico | | |
| 3 | 0.044 | feminino | fumador | historico_familiar_diabetes | | |
| 3 | 0.054 | feminino | fumador | diabetico | | |
| 3 | 0.044 | feminino | historico_familiar_diabetes | diabetico | | |
| 3 | 0.051 | masculino | sedentarismo | fumador | | |
| 3 | 0.051 | masculino | sedentarismo | historico_familiar_diabetes | | |

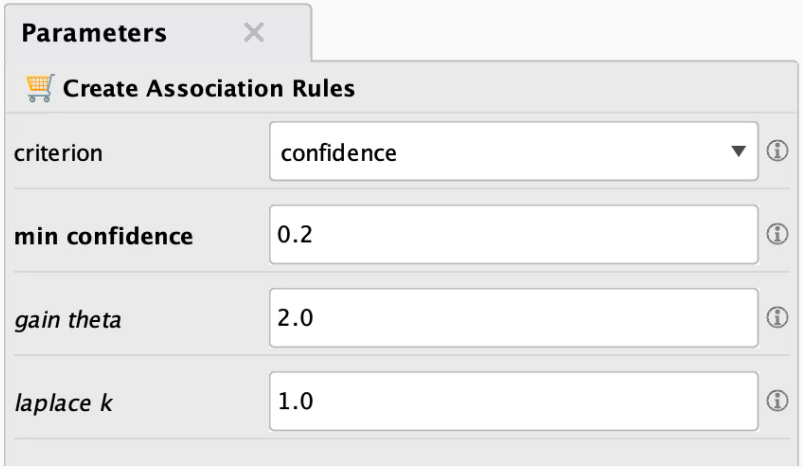
# Regras de Associação – Quizz time!

**Encontram alguns itemsets mais frequentes? O que quer isso dizer?**

# Regras de Associação

Para investigar estas relações podemos usar o operador *"Create Association Rules"*. Este operador usa os dados da matriz de frequência de padrões e procura quaisquer padrões que ocorram com frequência suficiente para que possam ser considerados regras. Procure este operador, arraste-o para o processo (tal como na imagem) e corra. Corra o modelo e selecione o separador dos resultados.



# Regras de Associação

- **O que é a confiança?**

Confiança de 20% numa regra A → C significa que só em 1 de cada 5 vezes que A acontece, B também acontece.

Isso é uma associação fraca, mesmo que apareça na lista.

| id | Items |
|---|---|
| 1 | A, B, C |
| 2 | A, C |
| 3 | A, D |
| 4 | B, E, F |

| Itemset freq. | support |
|---|---|
| {A} | 75% |
| {B} | 50% |
| {C} | 50% |
| {A,C} | 50% |

**Para a regra A => C:**

suporte = suporte(A ∪ C) = 50%
confiança = suporte(A ∪ C) / suporte (A) = 50 / 75 = 66.6%

# Regras de Associação

- **Resultados:**

| No. | Premises | Conclusion | Support | Confidence | LaPlace | Gain | p-s | Lift ↓ |
|---|---|---|---|---|---|---|---|---|
| 72 | sedentarismo, fumador, historico_familiar_diabetes | masculino, diabetico | 0.014 | 0.364 | 0.977 | -0.061 | 0.012 | 6.727 |
| 159 | masculino, sedentarismo, fumador, historico_familiar_diabetes | diabetico | 0.014 | 0.667 | 0.993 | -0.027 | 0.011 | 5.060 |
| 153 | sedentarismo, fumador, historico_familiar_diabetes | diabetico | 0.024 | 0.636 | 0.987 | -0.051 | 0.019 | 4.830 |
| 123 | masculino, historico_familiar_diabetes, diabetico | sedentarismo, fumador | 0.014 | 0.500 | 0.987 | -0.041 | 0.010 | 4.353 |
| 132 | feminino, fumador, historico_familiar_diabetes | diabetico | 0.024 | 0.538 | 0.981 | -0.064 | 0.018 | 4.087 |
| 129 | masculino, sedentarismo, historico_familiar_diabetes | diabetico | 0.027 | 0.533 | 0.977 | -0.074 | 0.020 | 4.048 |
| 120 | masculino, diabetico | sedentarismo, historico_familiar_diabetes | 0.027 | 0.500 | 0.974 | -0.081 | 0.020 | 4 |
| 63 | masculino, sedentarismo, diabetico | fumador, historico_familiar_diabetes | 0.014 | 0.333 | 0.974 | -0.068 | 0.010 | 3.654 |
| 112 | feminino, sedentarismo, fumador | diabetico | 0.030 | 0.474 | 0.968 | -0.098 | 0.022 | 3.595 |
| 83 | feminino, diabetico | sedentarismo, fumador | 0.030 | 0.391 | 0.956 | -0.125 | 0.021 | 3.407 |
| 43 | masculino, fumador, historico_familiar_diabetes | sedentarismo, diabetico | 0.014 | 0.286 | 0.968 | -0.081 | 0.010 | 3.383 |
| 80 | diabetico | sedentarismo, fumador | 0.051 | 0.385 | 0.928 | -0.213 | 0.036 | 3.348 |
| 105 | sedentarismo, fumador | diabetico | 0.051 | 0.441 | 0.942 | -0.179 | 0.036 | 3.348 |
| 53 | feminino, diabetico | fumador, historico_familiar_diabetes | 0.024 | 0.304 | 0.950 | -0.132 | 0.017 | 3.337 |
| 75 | masculino, diabetico | sedentarismo, fumador | 0.020 | 0.375 | 0.968 | -0.088 | 0.014 | 3.265 |

# Regras de Associação: Resultados

Lift = (probabilidade de A e B acontecerem juntos) ÷ (probabilidade esperada se A e B fossem independentes um do outro)

**Como interpretar o valor:**

- **Lift = 1** → não há associação real. A e B acontecem juntos exatamente tanto quanto seria de esperar só pelo acaso (independência estatística pura).
- **Lift > 1** → associação positiva. A presença de A torna B **mais provável** do que seria sozinho. Quanto mais alto, mais forte a associação.
- **Lift < 1** → associação negativa. A presença de A torna B **menos provável**.

# Regras de Associação

**Regra 159** (masculino, sedentarismo, fumador, historico_familiar_diabetes → diabetico): lift 5,06, confiança 66,7% — o número mais impressionante da tabela. **Mas o suporte é só 0,014**, ou seja, isto baseia-se em **cerca de 4 pacientes** (0,014 × 296). Com uma amostra tão pequena, este resultado pode não generalizar — é fácil 4 pacientes coincidirem por acaso.

**Regra 153** (sedentarismo, fumador, historico_familiar_diabetes → diabetico, sem exigir masculino): lift 4,83, confiança 63,6%, suporte 0,024 (~7 pacientes). Ainda pequeno, mas já mais robusto que a anterior, e mais simples (3 fatores em vez de 4).

**Regra 105** (sedentarismo, fumador → diabetico): lift 3,348, confiança 44,1%, **suporte 0,051 (~15 pacientes)**.
Tem o melhor equilíbrio entre força da associação (lift > 3, claramente acima do acaso) e tamanho de amostra suficiente para confiar no resultado.

# Regras de Associação

Praticamente todas as regras fortes que terminam em diabetico envolvem a mesma combinação de 3 ingredientes:

- sedentarismo, fumador, historico_familiar_diabetes, sozinhos ou em pares.

sexo (masculino/feminino) aparece em quase todas as regras só porque tem suporte alto (>0,48 cada), não porque acrescente força real à associação.

**Regra 153** (sem sexo) tem lift quase tão alto como a 159 (com sexo), confirmando que o sexo é "ruído" a inflar artificialmente o número de premissas.

# Regras de Associação: Resultados

- A combinação sedentarismo + fumador + historico_familiar_diabetes é a que mais se associa a diabetico, com lift entre 3,3 e 4,8 segundo o tamanho do subconjunto.
- Regra mais robusta (maior suporte): sedentarismo, fumador → diabetico — lift 3,35, confiança 44%, ~15 pacientes.
- Regras com lift > 5 existem, mas com suporte muito baixo (<5% dos pacientes) — força estatística forte, mas amostra pequena demais para generalizar com confiança.
- sexo não acrescenta valor real às regras — aparece por ter alta frequência base, não por associação genuína.
- Os 3 fatores de risco comportamentais/familiares justificam ser usados como atributos prioritários na Classificação.

# REGRESSÃO LINEAR

# Regressão Linear

- Modelo preditivo que usa datasets de treino e scoring para gerar **previsões numéricas**.
- Cada atributo no dataset é avaliado estatisticamente pela sua capacidade de prever o atributo do tipo label.
- Os atributos com fraca capacidade de previsão são removidos do modelo.
- Depois das previsões de regressão linear serem calculadas, os resultados podem ser resumidos para determinar se há diferenças nas previsões em subconjuntos dos dados de scoring.
- À medida que mais dados são recolhidos, estes podem ser adicionados dataset de treino para o tornar mais robusto ou expandir os intervalos de alguns atributos para incluir ainda mais valores.
- É muito importante lembrar que os intervalos para os atributos de scoring devem estar dentro dos intervalos dos atributos de treino para garantir previsões válidas.

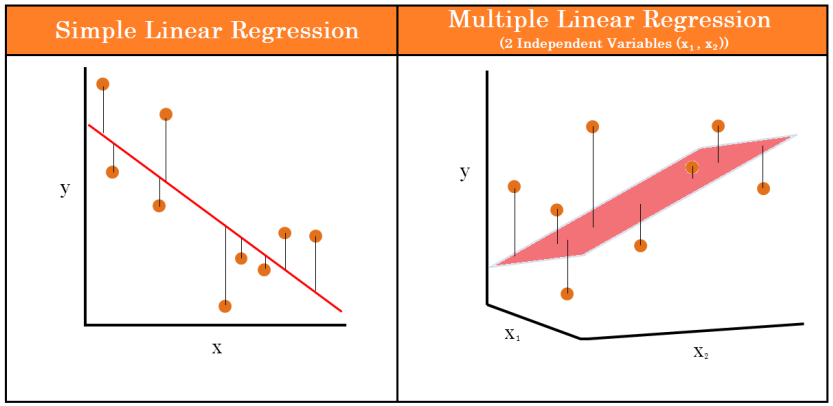
# Regressão Linear

Regressão Linear só tem uma variável independente:

$$y = \beta_0 + \beta_1 x$$

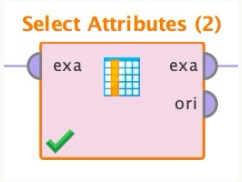
Regressão Linear Múltipla tem várias variáveis independentes

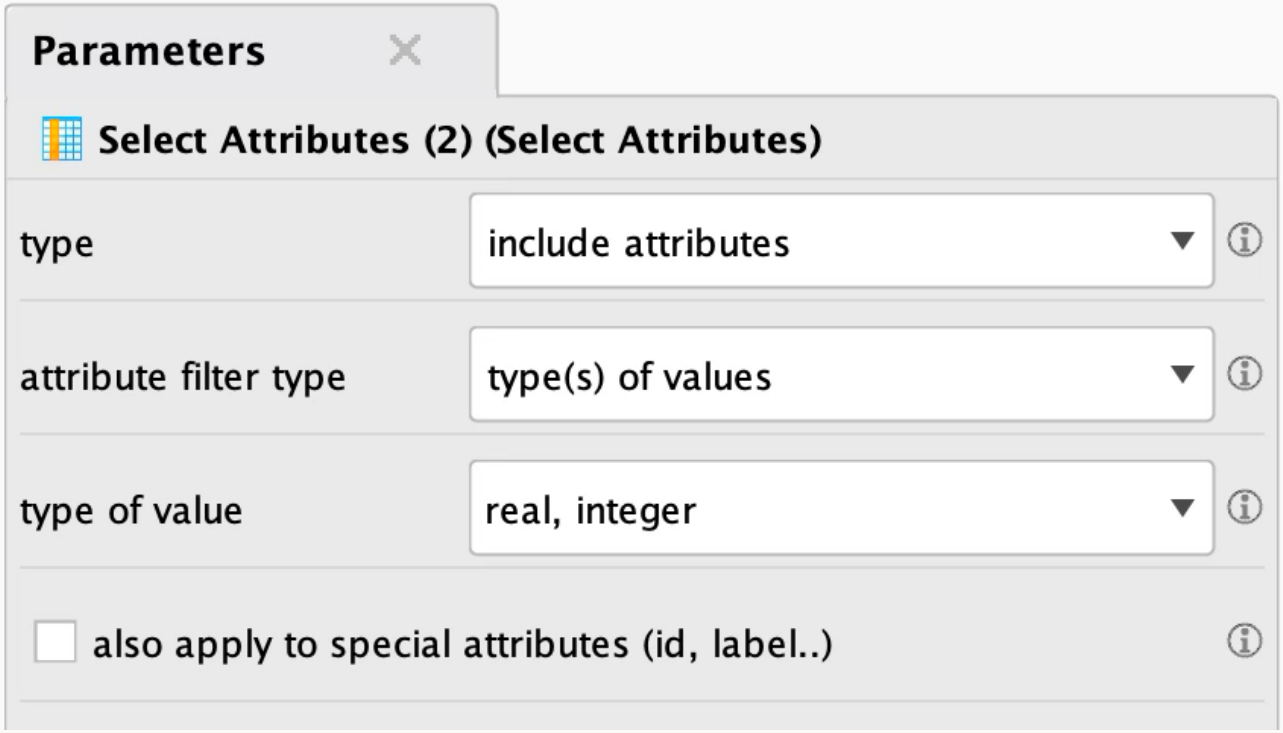
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$



# Regressão Linear – Atributos

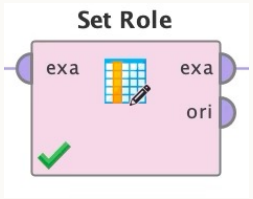
Apenas atributos numéricos devem ser considerados. Utilizamos o operador Select Attributes.

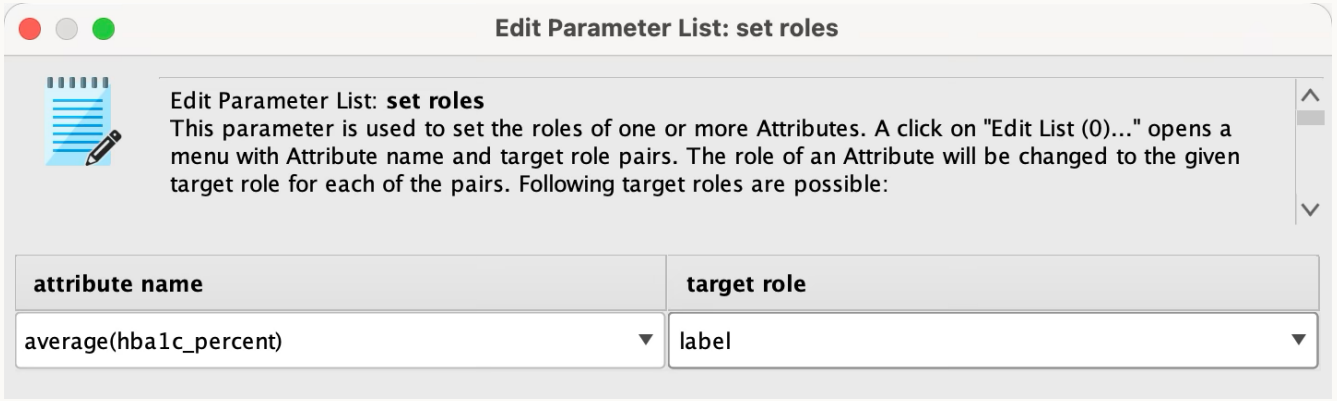


# Regressão Linear – Target

A regressão linear procura encontrar uma formula para determinar um valor numérico. Para tal temos de definir qual é o nosso target (label). Neste caso será o hab1c! Usamos o operador “Set Role”.

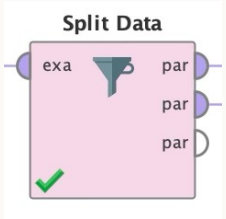


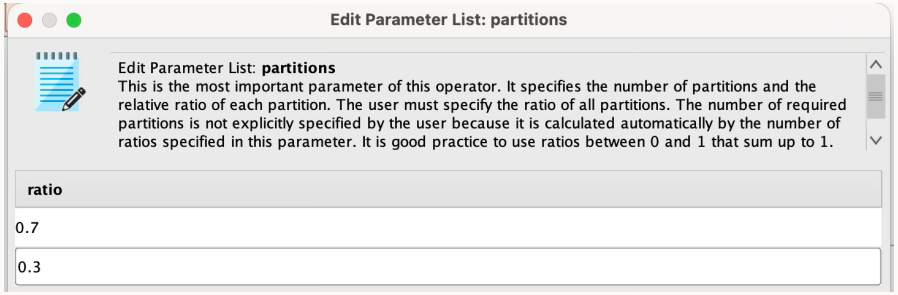



# Regressão Linear – Treino vs Teste

O próximo passo é definir os dois conjuntos de treino e teste. Para tal usamos o operador “split data”.

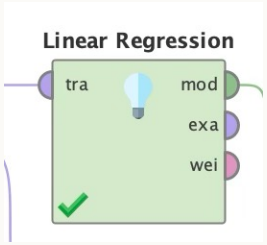
Neste operador podemos definir as percentagens a utilizar, neste caso usando as definições base. Para a definição das percentagens vamos usar 70% para treinar, 30% para testar.

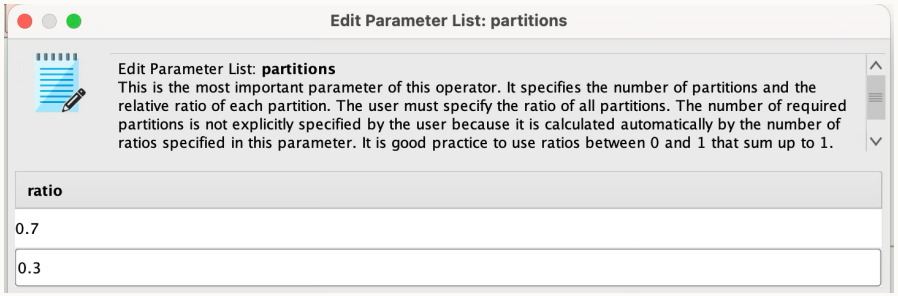


# Regressão Linear

Encontrar o operador ‘Linear Regression’ e arraste-o para a janela do processo.

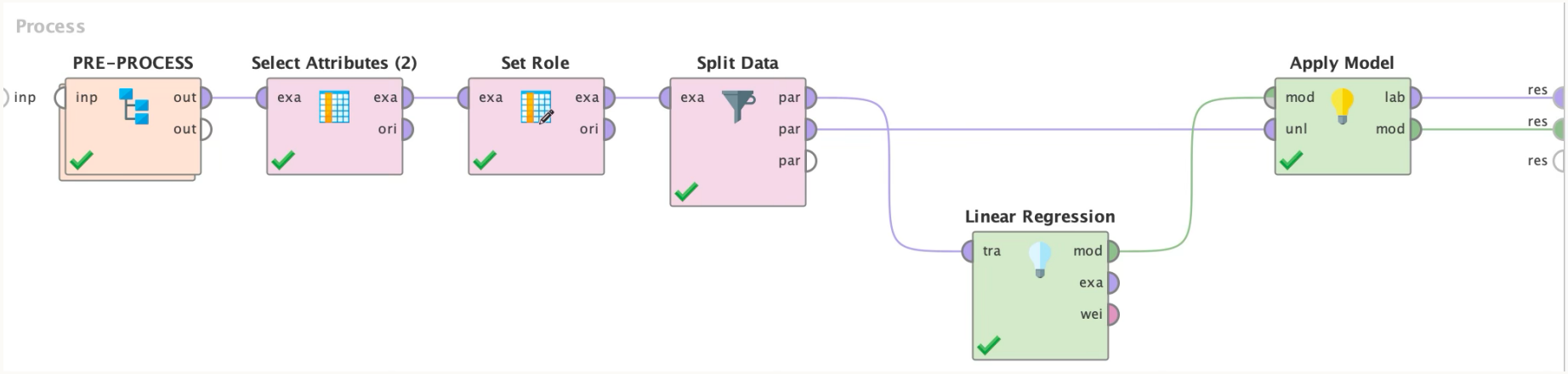


# Regressão Linear

O passo seguinte na modelação é usar um operador do tipo ‘Apply Model’ para ligar o fluxo de treino ao fluxo de teste. Procure este operador e arraste-o para a janela do processo.

Certifique-se de conectar as portas lab e mod às portas res como ilustrado na figura. De seguida devemos colocar também a criação do modelo como modelo de saída para procedermos depois à avaliação.



# Regressão Linear - Modelo

Corra o modelo. O facto de existirem duas saídas do operador 'Apply Model' conectadas às portas res, resultará em dois separadores na perspetiva de resultados. Vamos examinar primeiro o separador LinearRegression.

| Attribute | Coefficient | Std. Error | Std. Coefficient | Tolerance | t-Stat | p-Value | Code |
|---|---|---|---|---|---|---|---|
| idade | -0.029 | 0.038 | -0.033 | 0.983 | -0.760 | 0.448 | |
| imc | -0.036 | 0.044 | -0.036 | 0.975 | -0.822 | 0.412 | |
| tensao_sistolica | -0.001 | 0.042 | -0.001 | 0.981 | -0.033 | 0.974 | |
| tensao_diastolica | 0.006 | 0.036 | 0.008 | 0.972 | 0.173 | 0.863 | |
| average(glicose_mg_dl) | 0.759 | 0.041 | 0.797 | 0.968 | 18.316 | 0 | **** |
| average(trigliceridos_mg_dl) | 0.043 | 0.076 | 0.024 | 0.982 | 0.559 | 0.576 | |
| (Intercept) | -0.031 | 0.038 | ? | ? | -0.814 | 0.416 | |

**Coeficiente:**

Valor positivo tem impacto positivo no valor;
Valor negativo Tem impacto negativo.

**p-Value: < 0.05** Define quão importante um atributo é.

# Regressão Linear - Modelo

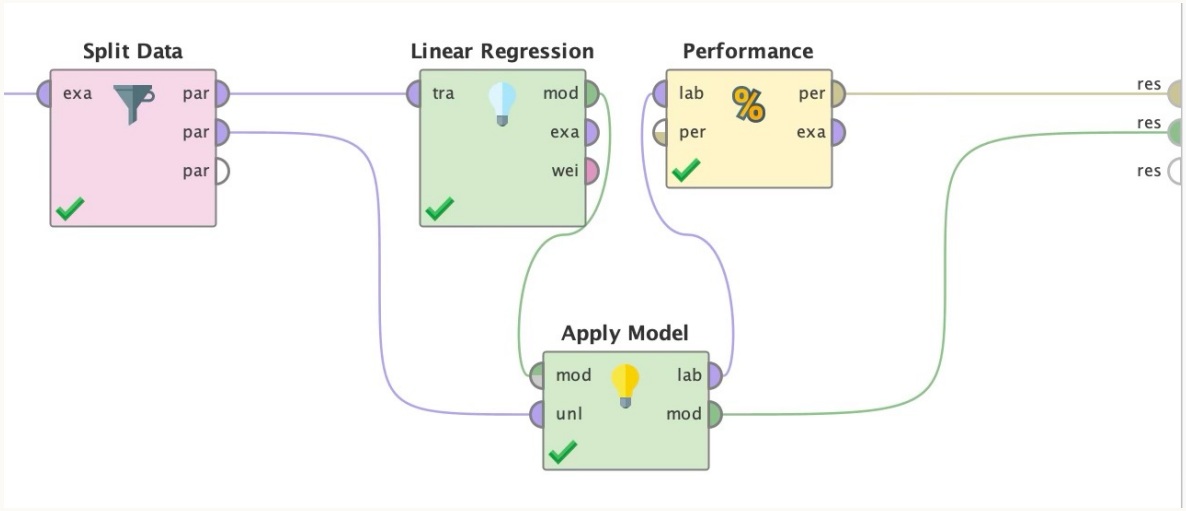
**hab1c** =

- 0.029 * idade
- 0.036 * imc
- 0.001 * tensao_sistolica
+ 0.006 * tensao_diastolica
+ 0.759 * average(glicose_mg_dl)
+ 0.043 * average(trigliceridos_mg_dl)
- 0.031

| average(hba1c_percent) | prediction(average(hba1c_percent)) |
|---|---|
| -1.448 | -1.241 |
| 0.000 | -0.427 |
| 1.789 | 0.802 |
| -1.009 | -0.223 |
| 0.747 | 0.283 |
| 1.551 | 1.196 |
| -0.241 | 0.106 |
| -0.406 | -0.357 |
| -1.009 | -0.828 |
| 0.143 | 0.477 |
| 0.417 | -0.194 |
| 0.454 | -0.171 |
| -1.832 | -1.326 |
| 0.088 | 0.410 |

# Regressão Linear - Avaliação

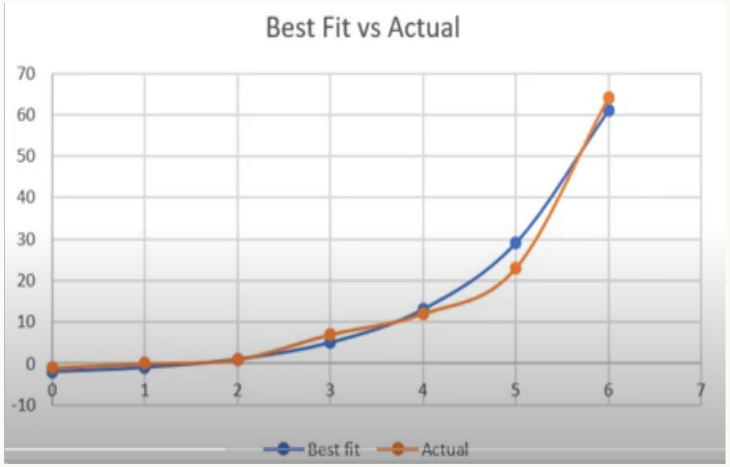
Volte para o separador Design e procure o operador “Performace” e adicione-o.
O que vamos avaliar é o “root mean squared error”.



# Regressão Linear - Avaliação

Root Mean Square Error (RMSE): Desvio padrão entre os valores previstos e os valores observados.

root_mean_squared_error: 0.564 +/- 0.000



# Regressão Linear – Quizz time!

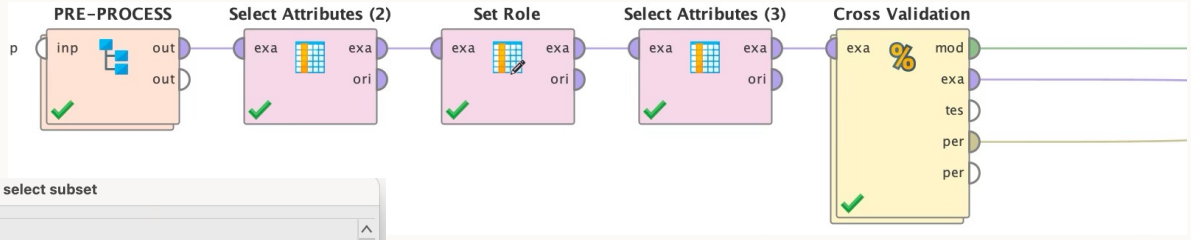
root_mean_squared_error: 0.564 +/- 0.000

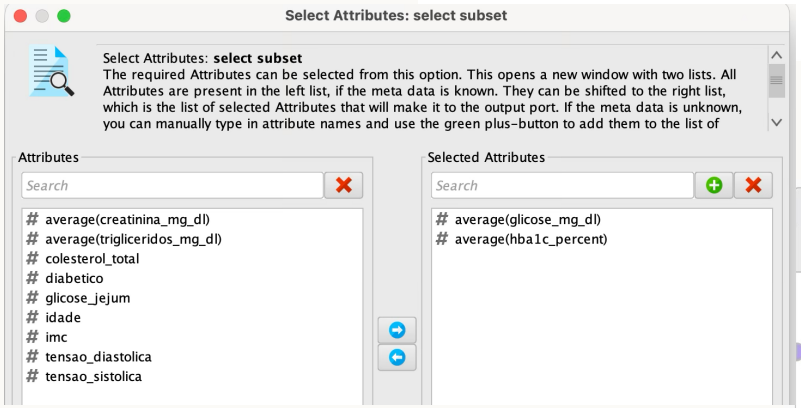


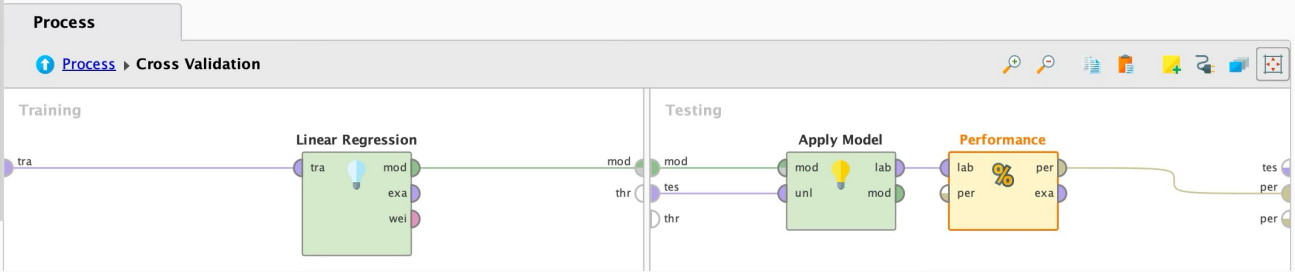
**Será este um bom resultado??**

# Regressão Linear – Quizz time!

Vamos tentar refazer o processo escolhendo uma forma de sampling diferente e apenas o atributo *average(glicose_mg_ml)*!



# Regressão Linear – Quizz time!

| Attribute | Coefficient | Std. Error | Std. Coefficient | Tolerance | t-Stat | p-Value | Code |
|---|---|---|---|---|---|---|---|
| average(glicose_mg_dl) | 0.736 | 0.035 | 0.778 | 1 | 21.256 | 0 | **** |
| (Intercept) | -0.017 | 0.032 | ? | ? | -0.519 | 0.604 | |

root_mean_squared_error: 0.544 +/- 0.091



**Será este um bom resultado??**

# Regressão Linear – Quizz time!

Vamos tentar refazer o processo escolhendo outros atributos!

**$R^2$ (coeficiente de determinação): a proporção da variação do alvo que o modelo consegue explicar**, numa escala de 0 a 1 (ou 0% a 100%).
**Como interpretar:**

**$R^2 = 0$** → o modelo não explica nada; é tão bom como prever sempre a média.
**$R^2 = 1$** → o modelo explica perfeitamente todos os dados (sem erro nenhum).

**($R^2 \approx 0{,}778^2 = 0{,}605$)** → o modelo explica **60% da variação** do hba1c; os restantes 40% ficam por explicar - por ruído, por atributos que não tens no dataset, ou por variação que não segue uma relação linear simples.

# Regressão Linear – Quizz time!

**A lição mais importante de toda esta sequência:** o mesmo par de atributos (glicose e hba1c) deu resultados ligieramente diferentes, e a diferença não estava no modelo nem no algoritmo. Estava sempre em **como os dados foram preparados antes de chegarem ao modelo**.

Preparação dos dados não é burocracia antes da parte “interessante”! É a parte que decide se o resultado final é bom ou não!

# Aprendizagem e Extração do Conhecimento

*PL04* · CRISP – DM | Construção de Modelos

Mestrado em Engenharia Biomédica

Hugo Peixoto – hpeixoto@di.uminho.pt

*Conteúdo autoral; design e estruturação suportados por IA.*