



PL09 – RapidMiner: Regras de Associação

AEC - Mestrado em Engenharia Biomédica

<https://hpeixoto.me/class/aec>

Plano de Aula - PL09



Regras de Associação: Exemplo



Ficha Exercícios (fe06)





Regras de Associação: Exemplo

Contexto e Perspectiva



O Pedro é gerente municipal de uma cidade de médio porte, mas que está em constante crescimento. Como a maioria dos municípios, a cidade tem recursos limitados face às necessidades que encontra.

O Pedro sabe que os cidadãos da comunidade são ativos em várias **organizações comunitárias** como **igrejas**, **clubes sociais** e entusiastas de **passatempos**, e acredita que estes grupos possam trabalhar juntos para atender algumas necessidades da comunidade.

Antes de começar a pedir às organizações comunitárias que comecem a trabalhar em conjunto, o Pedro precisa de descobrir se existem associações naturais entre os diferentes tipos de grupos

O Data Mining pode ajudá-la a compreender estas associações



Regras de Associação

Business Understanding

O objetivo do Pedro é identificar e tirar proveito das conexões existentes na sua comunidade local para realizar algum trabalho que beneficie toda a comunidade.

O Pedro e a sua família estão envolvidos num grupo amplo de organizações comunitárias, por isso ele está ciente, num sentido mais geral, da diversidade dos grupos assim como dos seus interesses, objetivos e potenciais contribuições.

Identificar indivíduos com quem trabalhar em cada igreja, clube social ou organização política será esmagador sem primeiro categorizar as organizações em grupos e procurar associações entre eles.

As **regras de associação** são uma metodologia de Data Mining que procura encontrar ligações frequentes entre os atributos de um data set.



Regras de Associação

Business Understanding

As **regras de associação** são comuns quando se faz análise de cestos de compras. Comerciantes e fornecedores em vários setores usam esta abordagem de Data Mining para tentar encontrar quais os produtos que são frequentemente comprados em conjunto.

Por exemplo, quando se compra um smartphone, acessórios como protetores de ecrã, carregadores ou auriculares são frequentemente recomendados. Os itens recomendados são identificados por técnicas de **regras de associação** entre itens que clientes anteriores compraram em conjunto com o item que você comprou.

Isto acontece quando a associação é tão frequente no conjunto de dados, que a **associação** pode ser considerada uma **regra**. Assim nasce o nome desta abordagem de Data Mining: "regras de associação".



Data Understanding

Usando o conhecimento do Pedro sobre a comunidade local foi criado um **pequeno questionário** que foi executado online através de um site. Os líderes de cada organização convidada a participar no estudo receberam uma password única. Cada líder compartilhou com os membros do seu grupo a password. Após o término do questionário, foi criado um data set composto pelos seguintes atributos:

- **Elapsed_Time:** tempo que a pessoa gastou para completar o questionário. Expresso em minutos decimais (4,5 neste atributo seriam quatro minutos e trinta segundos).
- **Time_in_Community:** tempo que a pessoa viveu na área por 0-2 anos, 3-9 anos ou 10+ anos. Está registado no data set como “Short”, “Medium”, ou “Long”, respetivamente.
- **Gender:** sexo da pessoa.
- **Working:** resposta do tipo sim/não indicando se a pessoa tem ou não um emprego remunerado atualmente.



Regras de Associação

Data Understanding

- **Age:** idade da pessoa em anos.
- **Family:** resposta do tipo sim/não indicando se a pessoa é ou não membro de uma organização comunitária orientada para a família, como ligas recreativas ou desportivas para crianças, grupos de genealogia, etc.
- **Hobbies:** resposta do tipo sim/não indicando se a pessoa é ou não atualmente membro de uma organização comunitária orientada a hobbies, como rádio amadora, recreação ao ar livre, motocicletas ou passeios de bicicleta.
- **Social_Club:** resposta do tipo sim/não indicando se a pessoa é ou não membro de uma organização social comunitário.
- **Political:** resposta do tipo sim/não indicando se a pessoa é ou não membro de uma organização política com reuniões regulares na comunidade, como um partido político.

Regras de Associação

Data Understanding

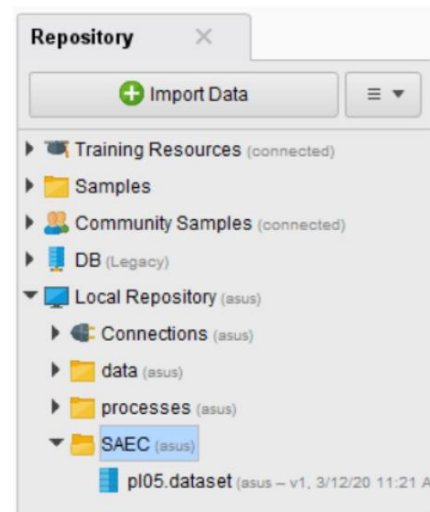
- **Professional:** resposta do tipo sim/não indicando se a pessoa é ou não membro de uma organização profissional com reuniões de comitês locais, como um comitê de uma lei ou sociedade médica, um grupo de pequenos empresários.
- **Religious:** resposta do tipo sim/não indicando se a pessoa é ou não atualmente membro de uma igreja na comunidade.
- **Support_Group:** resposta do tipo sim/não indicando se a pessoa é ou não membro de uma organização comunitária orientada para o apoio, como Alcoólicos Anónimos.

Regras de Associação

Data Preparation

Download do dataset: pl09-dataset.csv

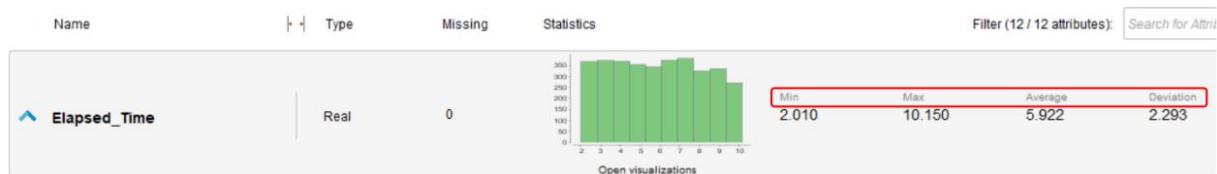
1. Importar o CSV para o repositório rapidminer (ImportData -> MyComputer)
2. Verificar a view dos resultados e inspecionar os dados CSV importados (Data,Statistics)



Regras de Associação

Data Preparation

3. Arraste o dataset pl09.dataset para uma nova janela de processo no RapidMiner
4. Execute o modelo para inspecionar os dados e salve o processo como pl09_processo.
5. Seleccione a vista “Results” e escolha a opção “Statistics”. Note que:
 - Não existe nenhum missing value para nenhum dos 12 atributos.
 - Para os dados numéricos, o RapidMiner apresenta o valor mínimo, o valor máximo, a média e o desvio padrão para cada atributo.



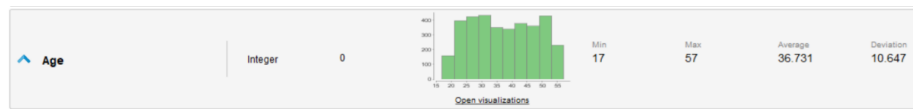
Data Preparation

Qualquer valor inferior a dois desvios padrão abaixo da média ou dois desvios padrão acima da média, é estatisticamente considerado como outlier.

Por exemplo, no atributo “Age”, a idade média é 36,731, enquanto o desvio padrão é 10,647.

Dois desvios padrão acima da média seriam 58,025 ($36,731 + (2 \cdot 10,647)$), e dois desvios padrão abaixo da média seriam 15,437 ($36,731 - (2 \cdot 10,647)$).

- Ao observar o valor Min e Max, é possível perceber que o atributo “Age” tem um intervalo de 17 a 57, por isso todas as instâncias estão dentro de dois desvios padrão acima e abaixo da média, ou seja, não existem outliers.



É importante saber que embora dois desvios padrão sejam uma diretriz, não é uma regra universal.



Data Preparation

Notas:

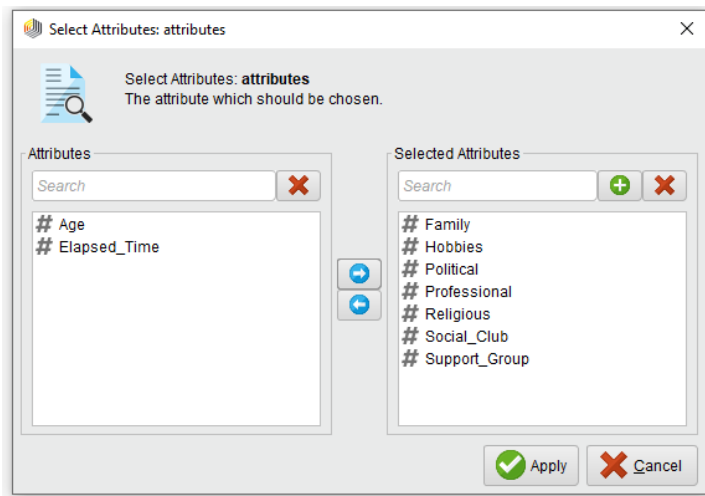
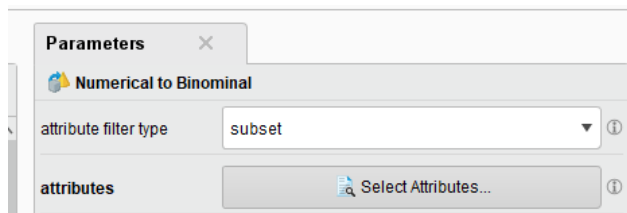
- Os atributos do tipo sim/não foram registados como 0 ou 1 e importados como 'integer'.
 - Os operadores de regras de associação do RapidMiner requerem que os atributos sejam do tipo de dados 'binominal'.
6. Volte para a view “Design”. Na caixa Operadores, pesquise “Numerical to Binomial” e adicione esse operador na janela de processo.



Data Preparation

7. Na janela do processo, clique em cima do operador “Numerical to Binomial”.

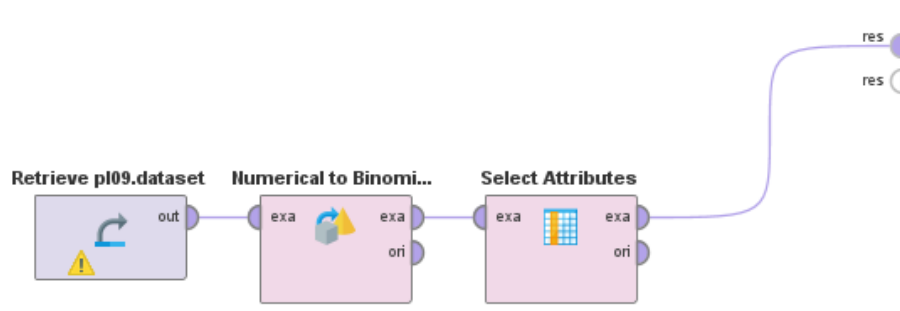
No painel lateral direito intitulado Parameters, mude o "attribute filter type" para "subset" e depois selecione a opção "Select Attributes".
Selecione os seguintes atributos para inclusão: *Family, Hobbies, Social_Club, Political, Professional, Religious, Support_Group*.



Regras de Associação

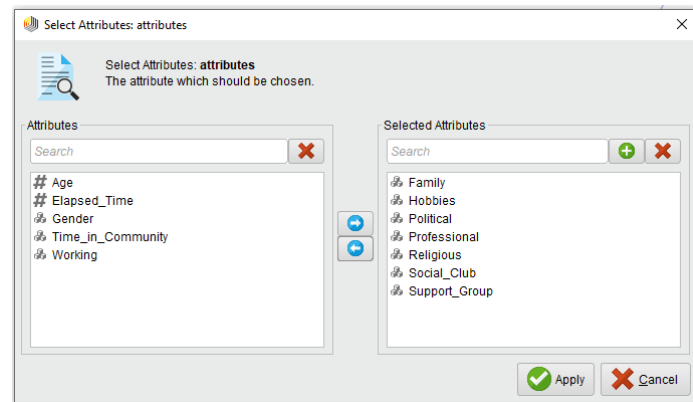
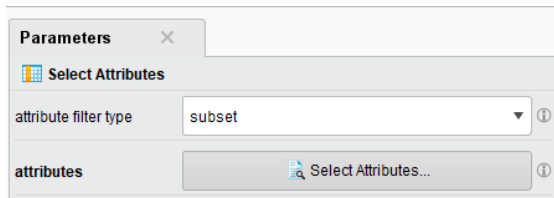
Data Preparation

8. É necessário reduzir o número de atributos no nosso conjunto de dados. O tempo que cada pessoa demorou para completar o questionário não é relevante no contexto do nosso problema, assim como outros atributos como o sexo e a idade. Adicione um operador do tipo *Select Attributes* e arraste para a janela do processo.



Data Preparation

9. Na janela do processo, clique em cima do operador Select Attributes. No painel lateral direito intitulado Parameters, mude o attribute filter type para “subset” e depois selecione a opção “Select Attributes”. Selecciona os seguintes atributos para inclusão: *Family, Hobbies, Social_Club, Political, Professional, Religious, Support_Group*.





Data Preparation

10. Clique no botão 'play' para correr o modelo.

Row No.	Family	Hobbies	Social_Club	Political	Professional	Religious	Support_Gr...
1	true	false	false	false	false	false	false
2	false	false	false	false	false	true	true
3	true	true	false	false	true	false	false
4	false	false	false	false	false	false	false
5	false	false	false	true	true	false	true
6	false	false	false	false	true	false	false

Os valores de 1 ou 0 são agora refletidos como 'verdadeiro' ou 'falso'.

No RapidMiner, o tipo de dados 'binominal' é usado em vez de 'binomial'. **Binomial** significa um de dois números (geralmente 0 e 1). Binominal, por outro lado, significa um de dois valores que podem ser tanto numéricos como baseados em caracteres.



Modeling

Encontrar os item sets mais frequentes:

- Apriori
- FP-Growth

Encontrar as regras de associação mais fortes de acordo com os itemsets mais frequentes encontrados:

- Devem satisfazer um mínimo de confiança e suporte.

Modeling

O RapidMiner apresenta vários operadores de regras de associação. Neste exemplo será usado o operador FP-Growth.

FP (Frequent Pattern)

Sem ter frequências de combinações de atributos, não poderíamos determinar se algum dos padrões nos dados ocorre com frequência suficiente para ser considerado regra.

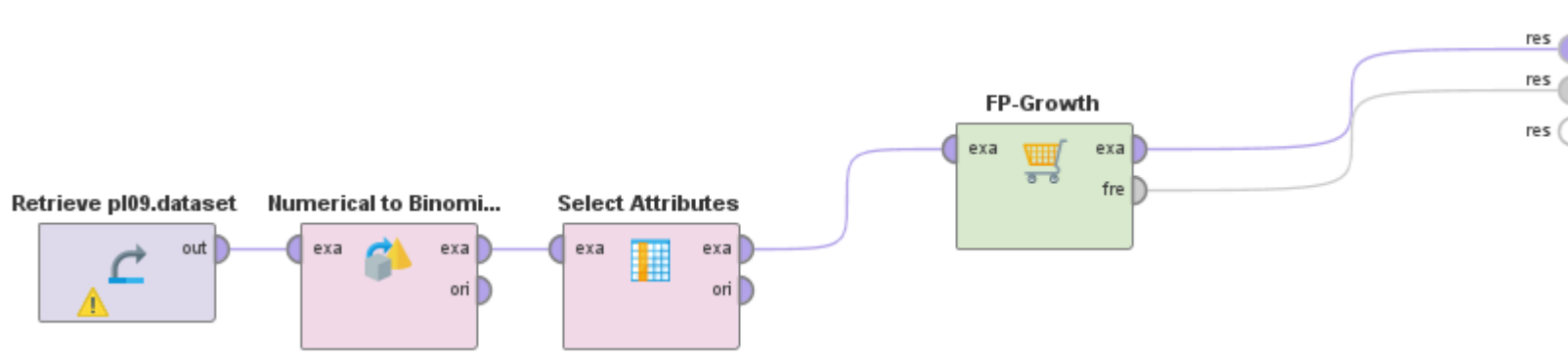
<https://www.softwaretestinghelp.com/fp-growth-algorithm-data-mining/>



Regras de Associação

Modeling

10. Arraste operador FP-Growth para o processo. Anote o parâmetro min support no lado direito. Certifique-se de que as portas **exa** e **fre** estão conectadas às portas **res**.



Porta **exa** irá gerar um separador de exemplo (observações e estatísticas do dataset)

Porta **fre** irá gerar uma matriz de qualquer padrão frequente que o operador possa encontrar nos dados



Modeling

11. Corra o modelo e selecione o separador dos resultados.

No. of Sets: 6

Total Max. Size: 2

Min. Size: 1

Max. Size: 2

Contains Item:

Update View

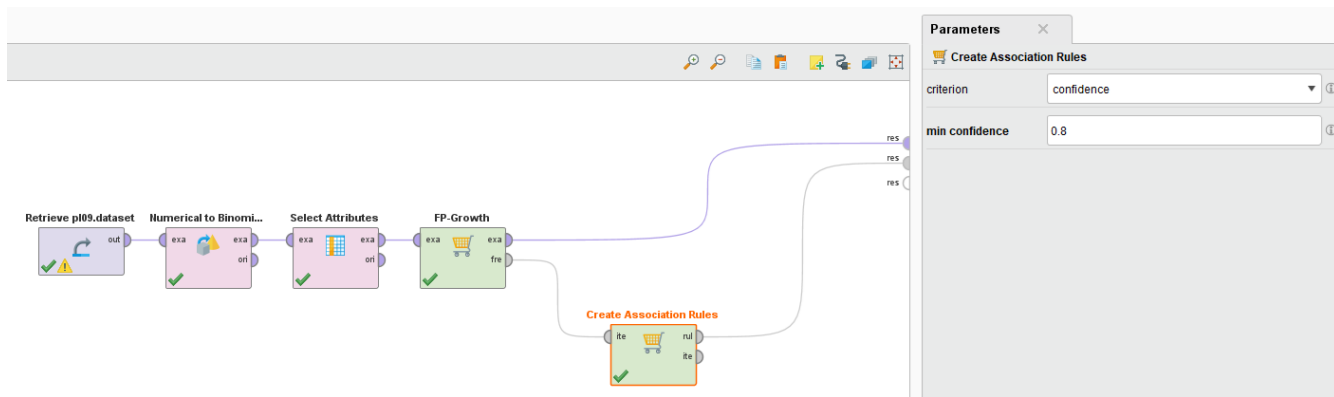
Size	Support	Item 1	Item 2
1	0.419	Religious	
1	0.390	Family	
1	0.324	Professional	
1	0.300	Hobbies	
2	0.225	Religious	Family
2	0.239	Religious	Hobbies

As organizações religiosas podem ter algumas conexões naturais com as organizações Família e Hobbies.

Regras de Associação

Modeling

12. 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.

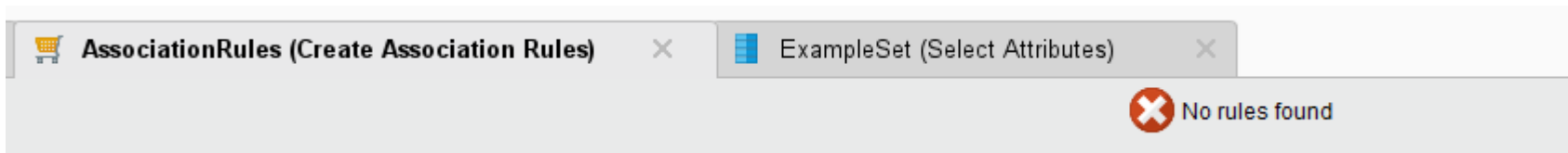


As organizações religiosas podem ter algumas conexões naturais com as organizações Família e Hobbies.

Regras de Associação

Modeling

13. Resultado: Não foram encontradas regras de associação.



O processo CRISP-DM é de natureza cíclica e, às vezes, é necessário voltar atrás entre as etapas antes de criar um modelo que produza resultados.



Evaluation - exemplo

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

Suporte:

$$\text{Sup}(A,B) = 1$$

$$\text{Sup}(A,C) = 2$$

Frequência de um padrão:

min. sup = 2, {A,C} é um padrão frequente

Para min. sup = 50% e min. conf. = 50%:

A \Rightarrow C com 50% suporte e 66% confiança (2 em 3)

C \Rightarrow A com 50% suporte e 100% confiança (2 em 2)



Evaluation - exemplo

Para min. sup = 50% e min. conf. = 50%:

id	Items comprados
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%

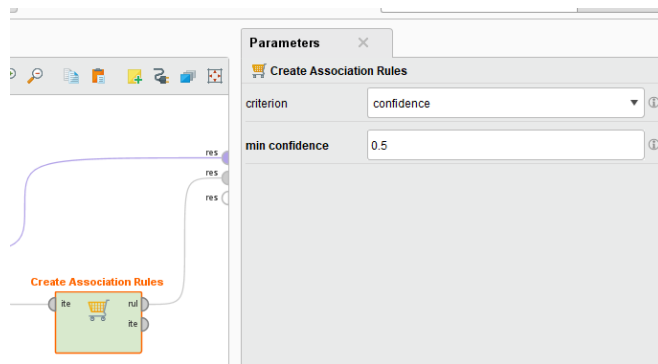
Para a regra $A \Rightarrow C$:

$\text{support} = \text{support}(\{A, C\}) = 50\%$

$\text{confidence} = \text{support}(\{A, C\}) / \text{support}(\{A\}) = 66.6\%$

Evaluation

No separador de Design, clique no operador Create Association Rules e mude o parâmetro min confidence para 0.5 -> qualquer associação com pelo menos 50% de confiança deve ser exibida como regra.





Regras de Associação

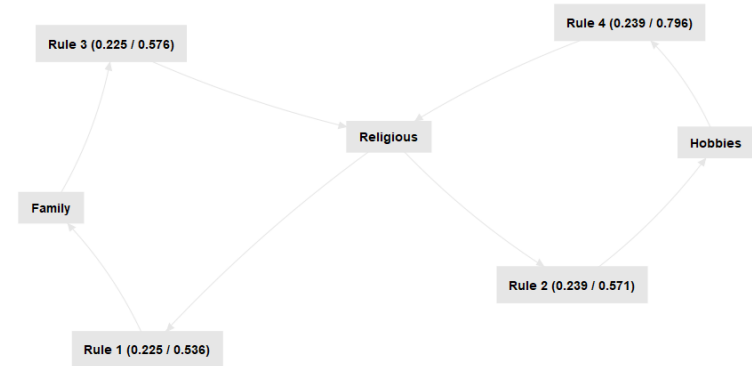
Evaluation

Min. Criterion:

confidence ▼

Min. Criterion Value:

No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Convicti...
1	Religious	Family	0.225	0.536	0.863	-0.613	0.061	1.376	1.316
2	Religious	Hobbies	0.239	0.571	0.873	-0.598	0.113	1.902	1.630
3	Family	Religious	0.225	0.576	0.881	-0.555	0.061	1.376	1.371
4	Hobbies	Religious	0.239	0.796	0.953	-0.361	0.113	1.902	2.852



Regras de Associação

Evaluation

- O palpite de que as organizações religiosas, familiares e de hobby estão relacionadas estava correto;
- A regra número 4 apresenta uma percentagem de confiança de quase 80%;
- As outras associações têm percentagens de confiança mais baixas, mas ainda assim são muito boas;
- Podemos observar que cada uma das quatro regras são suportadas por mais de 20% das observações no dataset;
- % de suporte: regra 1 = regra 3 e regra 2 = regra 4
- % de confiança: regra 1 = regra 2 = regra 3 = regra 4



Evaluation

Existem ligações entre os tipos de grupos comunitários?



Sim, as organizações de igreja, família e hobby da comunidade têm alguns membros em comum.



Parece que Pedro terá mais sorte em encontrar grupos que queira colaborar em projetos pela cidade, envolvendo organizações relacionadas com igrejas, hobbies e família.



Ficha de Exercícios 06



PL09 – RapidMiner: Regras de Associação

AEC - Mestrado em Engenharia Biomédica

<https://hpeixoto.me/class/aec>