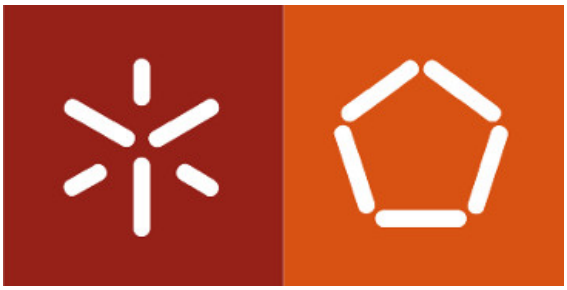

Universidade do Minho

Mestrado Integrado em Engenharia Informática

Engenharia do Conhecimento

Descoberta do Conhecimento

Resolução da Ficha de Exercícios 06

Gonçalo Pinto, A83732

João Diogo Mota, A80791

José Gonçalo Costa, PG42839

José Nuno Costa, A84829

30 de abril de 2021

Conteúdo

Lista de Figuras

Capítulo 1

Introdução

No 2º semestre do 1º ano do Mestrado em Engenharia Informática da Universidade do Minho, existe uma unidade curricular enquadrada no perfil de Engenharia do Conhecimento denominada por Descoberta do Conhecimento, que tem como objetivo a introdução ao conceito de descoberta do conhecimento.

Nesta ficha pretende-se que cada grupo execute os processos de *Business Understanding, Data Understanding, Data Preparation, Modeling e Evaluation* enquadrados na metodologia do *CRISP-DM* que suporta processos de *Data Mining* num determinado caso de estudo. A metodologia utilizada de *Data Mining* foi baseada em regras de associação que procura encontrar ligações frequentes entre os atributos de um *dataset*.

Para este efeito, foi fornecido um conjunto de dados pela equipa docente sobre os tipos de artigos que são comprados por um determinado cliente, desta forma pretende-se perceber as ligações entre artigos.

Capítulo 2

Parte 1

1. ***O que são regras de associação? Para que servem?***
   Regras de associação, tal como o nome sugere, são regras que permitem descobrir associações ou relações entre um conjunto de dados. Estas procuram descobrir padrões frequentes entre estes dados de forma a dar a entender possíveis benefícios em várias áreas de negócio. Estas regras são utilizadas por exemplo na saúde, no sentido que qualquer doença possui sintomas em comum, é possível uma maior aproximação do diagnóstico à causa real verificando e relacionando os sintomas do paciente; no entretenimento, por exemplo com o sistema de recomendação da plataforma *Netflix*, que permite apresentar os filmes e séries que mais poderão agradar ao cliente tendo em consideração o seu histórico; e sobretudo na área das compras e vendas de forma a apresentar produtos que poderão interessar aos clientes tendo por base um histórico de preferências destes após a compra de determinados produtos.
   De uma forma mais técnica, estas regras de associação fazem uso de modelos de *Machine Learning* para determinar padrões nos dados fornecidos reconhecendo padrões de *if-then*. Uma regra de associação, contém duas partes: um antecedente (*if)*; e um consequente (*then*). Ao encontrar estes padrões de *if-then*, são usados critérios de suporte (frequência com que um item aparece entre os dados) e confiança (indica o número de vezes que os *if-thens* são encontrados a *true*) para identificar os relacionamentos mais fortes.

2. ***Quais são as duas principais métricas calculadas nas regras de associação e como são calculadas?***
   As duas principais métricas calculadas nas regras de associação são:

   - **Suporte**: indica a percentagem de frequência com que os relacionamentos *if-then* aprecem no conjunto de dados;
   - **Confiança**: indica o número de vezes que estes relacionamentos foram verificados como verdade.

   De forma a entender melhor como estas métricas podem ser calculadas, é apresentado um exemplo relacionado com uma das áreas nas quais estas regras de associação têm mais impacto, que é as compras em que o objectivo é identificar se a localização de dois produtos, traz benefícios ao estarem localizados próximos numa loja ou não.
   Para este exemplo serão considerados 100 clientes de um supermercado e analisada a relação entre cereais e leite. Como tal considera-se que em 60 destes clientes compraram ambos os produtos, 30 compraram apenas leite e os restantes 10 apenas cereais. Desta forma, para ser calculado o suporte é necessário verificar nos 100 clientes em estudo, a quantidade (neste caso percentagem) de clientes que compraram estes dois produtos em simultâneo, sendo este valor:

$$suporte(\%) = 60/100 * 100 = 60\%$$

Em relação à confiança, este valor é dado pela quantidade de pessoas que compraram os dois produtos sobre o número de pessoas que compraram esse produto mesmo sendo apenas esse. Sendo assim, a confiança para o leite (em percentagem) é dada por:

$$confianca(\%) = 60/90 * 100 = 66.67\%$$

e a confiança para os cereais é de:

$$confianca(\%) = 60/70 * 100 = 85.71\%$$

3. ***De que tipo de dados devem ser os atributos de um dataset para usar os operadores Frequent Pattern no RapidMiner?***
   Os dados devem ser do tipo binomial, isto é, significa um de dois números (geralmente 0 e 1). Binomial, por outro lado, significa um de dois valores que podem ser tanto numéricos como baseados em caracteres.

4. ***Como é que os resultados das regras de associação são interpretados? No exemplo dos slides desta aula, qual foi a regra mais forte e como se sabe?***
   Os resultados das regras de associação são verificados através das colunas *Premises, Conclusion, Support e Confidence*, onde as regras são interpretadas no sentido *if Premises then Conclusion* e posteriormente de forma a descobrir a regra mais forte é usado os valores das colunas *Support e Confidence*.

No.	Premises	Conclusion	Support	Confiden... ↓	LaPlace	Gain	p-s	Lift	Convicti...
16	Family, Hobbies	Religious	0.155	0.828	0.973	-0.219	0.077	1.978	3.379
15	Hobbies	Religious	0.239	0.796	0.953	-0.361	0.113	1.902	2.852
14	Social_Club	Religious	0.147	0.783	0.966	-0.229	0.069	1.871	2.682
13	Religious, Family	Hobbies	0.155	0.689	0.943	-0.294	0.087	2.297	2.253
12	Religious, Hobbies	Family	0.155	0.648	0.932	-0.323	0.062	1.662	1.732
11	Hobbies	Family	0.187	0.623	0.913	-0.413	0.070	1.598	1.618
10	Family	Religious	0.225	0.576	0.881	-0.555	0.061	1.376	1.371
9	Religious	Hobbies	0.239	0.571	0.873	-0.598	0.113	1.902	1.630
8	Religious	Family	0.225	0.536	0.863	-0.613	0.061	1.376	1.316
7	Hobbies	Religious, Family	0.155	0.516	0.888	-0.445	0.087	2.297	1.602
6	Family	Hobbies	0.187	0.479	0.854	-0.593	0.070	1.598	1.344
5	Professional	Family	0.130	0.402	0.853	-0.519	0.004	1.030	1.020
4	Family	Religious, Hobbies	0.155	0.397	0.831	-0.625	0.062	1.662	1.262

Figura 2.1: Regras de associação, ordenadas de forma descendente pelo grau de confiança.

Como se pode observar pela figura 2.1 do exemplo dos slides desta aula, a regra com maior grau de confiança é a número 16, contudo a regra número 15 é regra que apresenta maior equilíbrio entre suporte e confiança.

Capítulo 3

Parte 2

1. ***Download do dataset order.csv, importação para o RapidMiner e colocação na janela de processo. Etapa de Data Understanding.***
   A primeira etapa do processo de *Data Mining* passa pelo processo de *Data Understanding*, que tem como objectivo a familiarização com os dados. O objectivo deste trabalho é identificar e tirar proveito das conexões existentes no *dataset* sobre *artigos* para realizar algum trabalho que beneficie a empresa. O conjunto de dados utilizado possui 138 atributos, grande parte deles representa a quantidade de um determinado produto que foi comprado no pedido, por uma questão de elevada extensão apresentamos apenas a descrição de alguns atributos, os restantes que não são referidos tem um significado semelhante apenas o tipo de produto comprado diverge:

   - **order_id**, representa o id do pedido.
   - **order_dow**, representa o dia da semana em que o pedido foi feito.
   - **order_hour_of_day**, representa a hora do dia em que o pedido foi feito.
   - **days_since_prior_order**, representa o número de dias passados desde o pedido principal.
   - **asian foods**, representa quanta comida asiática foi comprada no pedido.
   - **bread**, indica a quantidade de pão comprada no pedido.
   - **butter**, indica a quantidade de manteiga comprada no pedido.
   - **candy chocolate**, indica a quantidade de chocolate comprada no pedido..
   - **cat food care**, indica a quantidade de comida de gato comprada no pedido..
   - **cereal**, indica a quantidade de cereais comprada no pedido.
   - **coffee**, indica a quantidade de café comprada no pedido.
   - **dog food care**, indica a quantidade de comida de cão comprada no pedido.
   - **eggs**, indica a quantidade de ovos comprados no pedido.
   - **frozen pizza**, indica a quantidade de pizzas congeladas comprados no pedido.
   - **milk**, indica a quantidade de leite comprado no pedido.
   - **red wines**, indica a quantidade de vinho tinto comprado no pedido.
   - **soft drinks**, indica a quantidade de refrigerantes comprados no pedido.
   - **white wines**, indica a quantidade de vinho branco comprado no pedido.
   - **yogurt**, indica a quantidade de iogurtes comprados no pedido.

2. ***Execução dos passos referentes à etapa de Data Preparation no seu dataset.***
   A segunda etapa do processo de *Data Mining* passa pelo processo de *Data Preparation*, que diz respeito ao processamento dos dados de modo a que os dados transformados estejam numa forma mais adequada para os algoritmos de *Data Mining*. Assim, o primeiro passo foi verificar se existem valores em falta no conjunto de dados fornecido, o que não se verificou. Nesta fase também foi possível verificar algumas estatísticas interessantes relativas aos dados como: o valor mínimo, máximo, médio e desvio padrão, como se apresenta na figura 3.1.

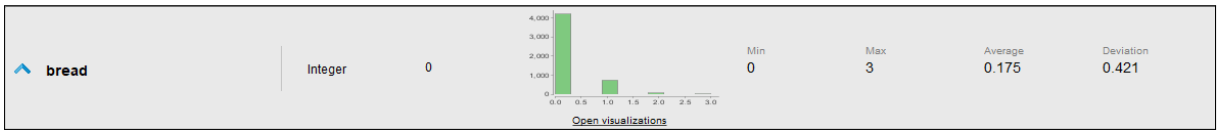

Figura 3.1: Estatísticas da coluna "*Bread*".

Nesta fase também tem como objectivo processar os dados transformando-os num formato mais adequado para os algoritmos a utilizar, para isso foi efectuado a conversão dos valores de numéricos para binomiais, utilizando o operador "*Numeric to Binominal*"do *RapidMiner*, nesta conversão foi necessário selecionar os atributos que fazem sentido para atingir o objetivo deste trabalho, assim foram removidos os seguintes atributos: *order_id; order_dow; order_hour_of_day; days_since_prior_order*. Este conjunto de atributos foram removidos uma vez não representam produtos que os clientes possam comprar, não se encaixando na análise pretendida.

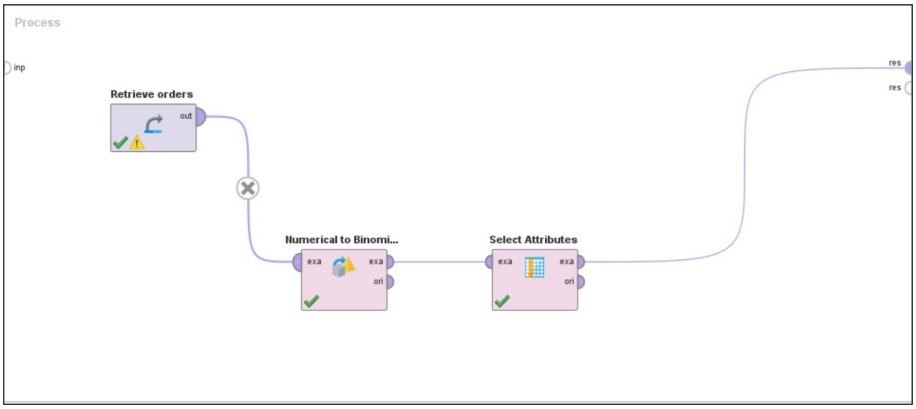

Figura 3.2: Processo de transformação dos dados para binomiais e sua selecção.

Row No.	air freshene...	asian foods	baby acces...	baby bath b...	baby food fo...	bakery dess...	baking ingre...	baking supp...	beauty	beers coolers	bo
1	false	false	false	false	false	false	false	false	false	false	fal
2	false	false	false	false	false	false	false	false	false	false	fal
3	false	false	false	false	true	false	false	false	false	false	fal
4	false	false	false	false	false	false	true	false	false	false	fal
5	false	false	false	false	false	false	false	false	false	false	fal
6	false	false	false	false	false	false	false	false	false	false	fal
7	false	false	false	false	false	false	false	false	false	false	fal
8	false	false	false	false	false	false	false	false	false	false	fal
9	false	true	false	false	false	false	false	false	false	false	fal
10	false	false	false	false	false	false	false	false	false	false	fal
11	false	false	false	false	false	false	false	false	false	false	fal
12	false	false	false	false	false	false	false	false	false	false	fal
13	false	false	false	false	false	false	false	false	false	false	fal

Figura 3.3: Resultado obtido após o processo de transformação dos dados para binomiais e sua selecção.

Na figura 3.3 podemos observar apenas os atributos seleccionados com valores a variar entre *true ou false*, isto é, *true* representa a compra do produto e *false* que o produto não foi comprado.

3. ***Geração das regras de associação para o dataset. Modificação dos valores de confiança (min confidence) e suporte (min support) para identificar os níveis ideais, de modo a obter regras interessantes com valores de confiança e suporte razoáveis. Análise de outras medidas de força das regras, como LaPlace ou Conviction.***
   A terceira etapa do processo de *Data Mining* passa pelo processo de *Modeling*, nesta fase pretende-se usar algoritmos de associação disponíveis no *RapidMiner* para descobrir associações nos dados. De forma a gerar regras de associação para o *dataset* foi utilizado o algoritmo *FP-Growth* com a opção *find min number of itemsets* desactivada, como se pode observar na figura 3.4.

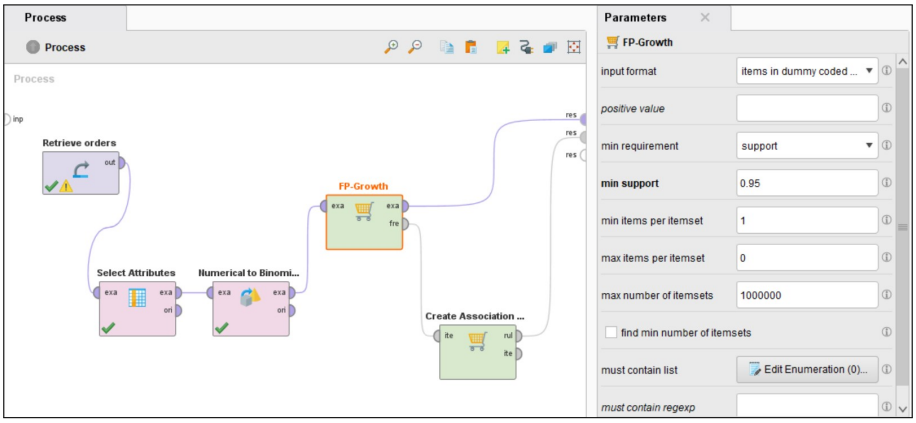

Figura 3.4: Geração das regras de associação para o dataset usando o operador *FP-Growth*.

Numa primeira fase foi definido no algoritmo *FP-Growth* o suporte (min support) de 0.2 e na criação das regras de associação uma confiança (min confidence) de 0.75, assim foram obtidos os resultados que se apresentam na figura 3.5.

No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Conviction
1	fresh vegetables, packaged vegetables fruits	fresh fruits	0.205	0.809	0.961	-0.301	0.067	1.482	2.375

Figura 3.5: Regras de associação usando o operador *FP-Growth* com suporte de 0.2 e com confiança de 0.75.

Posteriormente com o mesmo algoritmo foi modificado os valores de suporte e de confiança para 0.18 e 0.65, respectivamente. Os resultados obtidos após esta mudança estão presentes na figura 3.6.

No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Conviction
2	fresh vegetables	fresh fruits	0.328	0.716	0.911	-0.589	0.078	1.312	1.600
3	yogurt	fresh fruits	0.184	0.717	0.942	-0.330	0.044	1.313	1.603
4	fresh fruits, packaged vegetables fruits	fresh vegetables	0.205	0.719	0.938	-0.364	0.074	1.569	1.930
5	packaged vegetables fruits	fresh fruits	0.284	0.739	0.928	-0.485	0.075	1.355	1.744
6	fresh vegetables, packaged vegetables fruits	fresh fruits	0.205	0.809	0.961	-0.301	0.067	1.482	2.375

Figura 3.6: Regras de associação usando o operador *FP-Growth* com suporte de 0.18 e com confiança de 0.65.

Partindo das regras de associação geradas na figura 3.6, de forma a perceber a influência de outras medidas de força das regras, como *LaPlace ou Conviction* alteraram-se os valores desses parâmetros. Assim, modifcou-se o *gain theta* de 2.0 para 85.0 e o *laplace k* de 1.0 para 65.0, contudo as regras foram as mesmas apesar de os valores serem suficientes elevados comparativamente ao valor por *default*. Apenas a coluna *LaPlace e Gain* alteram os seus valores como era expectável uma vez que variou-se esses valores, na figura 3.7 é possível observar os resultados obtidos após esta alteração. Assim é possível concluir que estes dois parâmetros têm pouca força na descoberta de novas regras de associação.

No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Conviction
2	fresh vegetables	fresh fruits	0.328	0.716	0.020	-38.636	0.078	1.312	1.600
3	yogurt	fresh fruits	0.184	0.717	0.018	-21.644	0.044	1.313	1.603
4	fresh fruits, packaged vegetables fruits	fresh vegetables	0.205	0.719	0.018	-23.969	0.074	1.569	1.930
5	packaged vegetables fruits	fresh fruits	0.284	0.739	0.020	-32.407	0.075	1.355	1.744
6	fresh vegetables, packaged vegetables fruits	fresh fruits	0.205	0.809	0.018	-21.300	0.067	1.482	2.375

Figura 3.7: Regras de associação usando o operador *FP-Growth* com *gain theta* de 85.0 e *laplace k* de 65.0.

De seguida apresentamos algumas descobertas após a utilização do algoritmo *FP-Growth*:

(a) As regras de associação encontradas usando o operador *FP-Growth* com suporte de 0.15 e com confiança de 0.5 encontram-se presentes na figura 3.8.

No.	Premises	Conclusion	Support	Conf... ↓	LaPlace	Gain	p-s	Lift	Conviction
13	fresh vegetables, packaged vegetables fruits	fresh fruits	0.205	0.809	0.961	-0.301	0.067	1.482	2.375
12	packaged vegetables fruits	fresh fruits	0.284	0.739	0.928	-0.485	0.075	1.355	1.744
11	fresh fruits, packaged vegetables fruits	fresh vegetables	0.205	0.719	0.938	-0.364	0.074	1.569	1.930
10	yogurt	fresh fruits	0.184	0.717	0.942	-0.330	0.044	1.313	1.603
9	fresh vegetables	fresh fruits	0.328	0.716	0.911	-0.589	0.078	1.312	1.600
8	milk	fresh fruits	0.156	0.687	0.942	-0.298	0.032	1.258	1.450
7	packaged cheese	fresh fruits	0.158	0.684	0.941	-0.303	0.032	1.254	1.438
6	packaged vegetables fruits	fresh vegetables	0.253	0.658	0.905	-0.516	0.077	1.435	1.583
5	fresh fruits, fresh vegetables	packaged vegetables fruits	0.205	0.623	0.907	-0.452	0.078	1.621	1.634
4	fresh fruits	fresh vegetables	0.328	0.602	0.859	-0.763	0.078	1.312	1.359
3	fresh vegetables	packaged vegetables fruits	0.253	0.552	0.859	-0.664	0.077	1.435	1.373

Figura 3.8: Regras de associação usando o operador *FP-Growth* com suporte de 0.15 e com confiança de 0.5.

(b) Os atributos que estão mais fortemente associados são *fresh vegetables fresh fruits* pois apresenta maior equilíbrio entre suporte e confiança.

(c) Não existem relações entre dois produtos que se desviem muito do normal, contudo a única associação menos lógica que se observa acontece com *package cheese* e *fresh fruits* onde o grupo não consegue observar uma relação comparativamente à de *milk* com *fresh fruits* que pode ser explicada pela preferência dos clientes em fazerem *milkshakes*.

(d) Foi necessário algumas tentativas diferentes para encontrar um valor de suporte valores de suporte adequado ao dataset, uma vez que para encontrar uma regra foi necessário estabelecer um suporte de 0.2 um diferença considerável tendo em conta que o valor *default* para este campo é 0.95, o que não aconteceu no parâmetro da confiança uma vez que com o valor *default* já se encontrou resultados.

(e) Uma regra de associação que se considerou boa o suficiente ao ponto de ser possível basear nela para se tomar decisões, é a associação entre *fresh vegetables fresh fruits*, uma vez que apresenta um dos valores mais altos em termos de suporte, bem como um bom valor de confiança.

4. ***Utilização do algoritmo WFPGrowth no RapidMiner.***
   Enquadrado também no processo de *Modeling* foi criado um novo modelo de regras de associação usando o mesmo conjunto de dados, mas desta vez, usando o operador *W-FPGrowth no RapidMiner*. Desta forma, os resultados obtidos utilizando este algoritmo com suporte de 0.5 (U) e com confiança de 0.65 (C) são apresentados na figura 3.10.

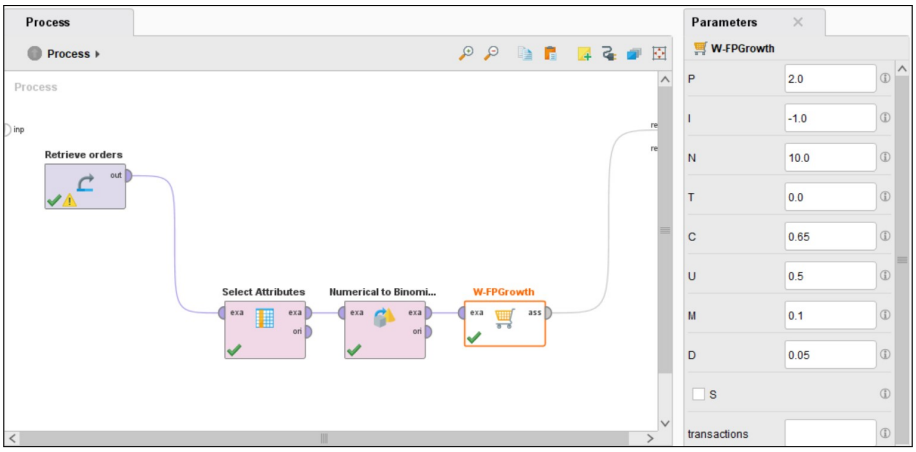

Figura 3.9: Criação de um novo modelo de regras de associação usando o operador *W-FPGrowth*.

W-FPGrowth

```
FPGrowth found 1 rules (displaying top 1)

1. [packaged vegetables fruits=true]: 1923 ==> [fresh vegetables=true]: 1265   <conf:(0.66)> lift:(1.44) lev:(0.08) conv:(1.58)
```

Figura 3.10: Regras de associação obtidas com *W-FPGrowth*.

5. ***Utilização do algoritmo Apriori (W-Apriori) no dataset num novo processo.***
   Um outro algoritmo frequentemente usado no processo de *Data Mining* para associações é o algoritmo *Apriori*, enquadrado mais uma vez na fase de *Modeling* foi criado um novo processo com este algoritmo. Assim, foi possível obter os resultados que se apresenta na figura 3.12 com um suporte de 0.25 (M) e confiança de 0.65 (C).

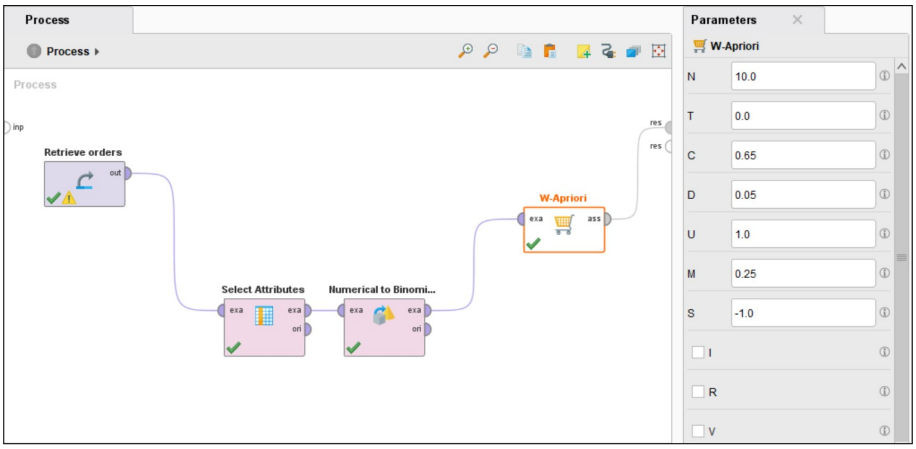

Figura 3.11: Criação de um novo modelo de regras de associação usando o operador *W-Apriori*.

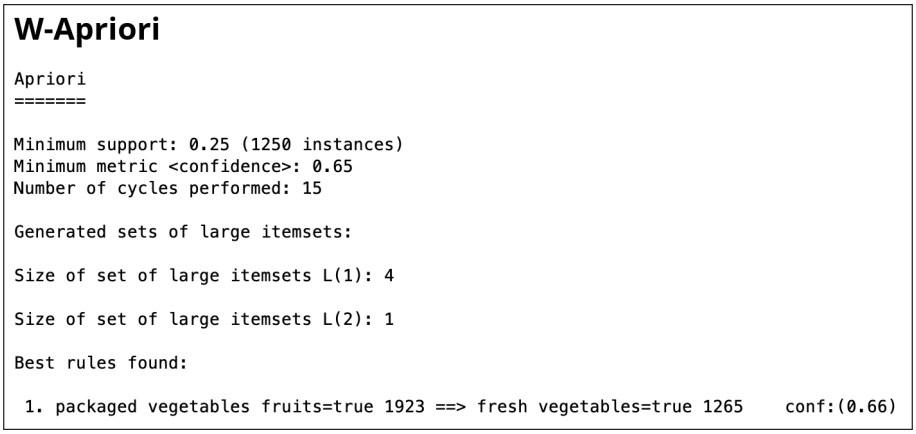

Figura 3.12: Regras de associação obtidas com *W-Apriori*.

6. ***Conclusões globais, comparando os resultados obtidos através de cada uma das técnicas utilizadas.***
   Na última etapa do processo de *Data Mining* consiste no processo de *Evaluation*, que passa por comparar os diferentes modelos obtidos no processo de *Modeling* e obter algumas conclusões. Assim, constatou-se que para a mesma relação de associação, a associação entre "se o cliente compra *package vegetables fruits*, então compra *fresh vegetables*", os valores de confiança obtidos com o algoritmo *W-FPGrowth* e *W-Apriori* foram os mesmos (0.66), por outro lado com o algoritmo *FP-Growth* também foi obtido um valor de confiança próximo mas superior, 0.739.

Capítulo 4

Conclusão

O presente relatório descreveu, de forma sucinta, o processo de exploração e estudo das regras de associação no *RapidMiner* num caso de estudo específico.

Após a realização deste problema, foi possível compreender o conceito de regras de associação, bem como, desenvolver um modelo de regras de associação no *RapidMiner*.

Por fim, esperamos que os conhecimentos obtidos e consolidados sejam de enorme utilidade tendo uma perspetiva futura.