REVIEW QUESTIONS

1. What are association rules? What are they good for?

   R.: As regras de associação são utilizadas para encontrar elementos que implicam a presença de outros elementos num determinado conjunto de dados, ou seja, encontrar relacionamentos/associações entre atributos ou padrões. Por exemplo, com isto e com acesso a uma base de dados de um supermercado, é possível saber o que costuma ser comprado "junto", como o exemplo das fraldas e cervejas, e tirar proveito destas descobertas.

2. What are the two main metrics that are calculated in association rules and how are they calculated?

   R.: As duas principais métricas utilizadas são a percentagem de confiança e de suporte. O suporte é o número de instâncias que satisfazem uma regra, ou seja, a frequência dos items na base de dados. Esta métrica é calculada dividindo o número de transações em que, por exemplo, foi comprado café pelo número de transações total.
   A confiança é a proporção de regras que satisfazem o lado esquerdo, para o qual o lado direito também suporta. Para mostrar como esta é calculada, apresenta-se o seguinte exemplo:

   SE café ENTÃO pão

   Assim, vê-se quantas vezes aparece "café" e "pão" na base de dados (na mesma transação), e divide-se pelo que está antes do ENTÃO, isto é, número de transações em que aparece "café".

3. What data type must a data set's attributes be in order to use Frequent Pattern operators in RapidMiner?

   R.: Todos os valores dos items devem ser nominais.

4. How are rule results interpreted? In this chapter's example, what was our strongest rule? How do we know?

   R.: Uma regra tem um antecedente e um consequente. Desta forma, qualquer regra é lida da forma **SE antecedente(s) ENTÃO consequente(s)**. Por exemplo, voltando à base de dados do supermercado, SE pão ENTÃO café, que indica que se um cliente compra pão, tipicamente compra café. Claramente que esta regra seria "classificada" pela confiança e pelo suporte, que indicariam as regras que poderiam ser classificadas como "boas regras".
   Assim, encontraram-se as seguintes regras:

No.	Premises	Conclusion	Support	Confid... ↓	LaPlace	Gain	p-s	Lift	Conviction
13	Hobbies, Social_Club	Religious	0.110	0.888	0.988	-0.137	0.058	2.122	5.208
12	Family, Hobbies	Religious	0.155	0.828	0.973	-0.219	0.077	1.978	3.379
11	Hobbies	Religious	0.239	0.796	0.953	-0.361	0.113	1.902	2.852
10	Social_Club	Religious	0.147	0.783	0.966	-0.229	0.069	1.871	2.682
9	Religious, Social_Club	Hobbies	0.110	0.745	0.967	-0.185	0.065	2.482	2.741
8	Religious, Family	Hobbies	0.155	0.689	0.943	-0.294	0.087	2.297	2.253
7	Social_Club	Family	0.129	0.685	0.950	-0.247	0.056	1.758	1.940
6	Social_Club	Hobbies	0.123	0.656	0.946	-0.253	0.067	2.188	2.038
5	Religious, Hobbies	Family	0.155	0.648	0.932	-0.323	0.062	1.662	1.732
4	Hobbies	Family	0.187	0.623	0.913	-0.413	0.070	1.598	1.618

Com isto, podemos identificar as 3 primeiras regras como as melhores. A regra mais forte talvez seria a número 11, por ter um maior compromisso entre confiança e suporte.

EXERCISE

3. As necessary, conduct your Data Understanding and Data Preparation activities on your data set. Ensure that all of your variables have consistent data and that their data types are appropriate for the FP-Growth operator.

   R.: Foram escolhidos todos os atributos, com a exceção de “days_since_pior_order”, “order_dow”, “order_hour_of_day” e “order_id”. Todos os atributos restantes são referentes ao tipo dos produtos comprados.

5. Document your findings. What rules did you find? What attributes are most strongly associated with one another. Are there products that are frequently connected that surprise you? Why do you think this might be? How much did you have to test different support and confidence values before you found some association rules? Were any of your association rules good enough that you would base decisions on them? Why or why not?

   R.: Para chegar a estas regras, foi necessário variar (descer) bastante o valor de confiança e suporte. Não se encontrou nenhuma regra com elevado suporte e elevada confiança, sendo que, tentando encontrar um compromisso entre estas duas métricas, tem-se, por exemplo, a regra:

   SE fresh vegetables e packed vegetables fruits ENTÃO fresh fruits

   Isto faz sentido, já que é costume fazerem-se as compras de produtos frescos conjuntamente.

No.	Premises	Conclusion	Support	Confidence	LaPlace	.	.	.	Conviction
2	packaged cheese	fresh fruits	0.158	0.684	0.941	.	.	.	1.438
3	milk	fresh fruits	0.156	0.687	0.942	.	.	.	1.450
4	fresh vegetables	fresh fruits	0.328	0.716	0.911	.	.	.	1.600
5	yogurt	fresh fruits	0.184	0.717	0.942	.	.	.	1.603
6	fresh fruits, packaged vegetables fruits	fresh vegetables	0.205	0.719	0.938	.	.	.	1.930
7	packaged vegetables fruits	fresh fruits	0.284	0.739	0.928	.	.	.	1.744
8	fresh vegetables, packaged vegetables fruits	fresh fruits	0.205	0.809	0.961	.	.	.	2.375

CHALLENGE STEP!

6. Build a new association rule model using your same data set, but this time, use the WFPGrowth operator. (Hints for using the W-FPGrowth operator: (1) This operator creates its own rules without help from other operators; and (2) This operator's support and confidence parameters are labeled U and C, respectively.
   R.: Resultado da execução do algoritmo (confiança e suporte mínimo= 65%):

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)
```

EXPLORATION!

7. The Apriori algorithm is often used in data mining for associations. Search the RapidMiner Operators tree for Apriori operators and add them to your data set in a new process. Use the Help tab in RapidMiner's lower right hand corner to learn about these operators' parameters and functions (be sure you have the operator selected in your main process window in order to see its help content).
   R.: Resultado da execução do algoritmo (confiança = 65% e suporte mínimo = 100%):

W-Apriori

```
Apriori
=======

Minimum support: 0.1 (500 instances)
Minimum metric <confidence>: 0.65
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 19

Size of set of large itemsets L(2): 10

Best rules found:

 1. packaged vegetables fruits=true 1923 ==> fresh vegetables=true 1265    conf:(0.66)
```