

Business Intelligence e Mineração de Dados

Marco Rodrigues nº4652

Mestrado em Engenharia de Software

2016-2017

**Índice**

[1. Introdução 3](#_Toc488182669)

[1.1. Estrutura do Documento 3](#_Toc488182670)

[2. Data Mining 3](#_Toc488182671)

[3. Regras de Associação 4](#_Toc488182672)

[4. Trabalhos Relacionados 5](#_Toc488182673)

[5. Casos Práticos 6](#_Toc488182674)

[5.1. Dataset Groceries 6](#_Toc488182675)

[5.1.1. Instalação 6](#_Toc488182676)

[5.1.2. Criação das Regras de Associação 10](#_Toc488182677)

[5.1.3. Análise de Resultados 13](#_Toc488182678)

[5.2. Dataset Titanic 14](#_Toc488182681)

[5.2.1. Instalação 14](#_Toc488182682)

[5.2.2. Criação das Regras de Associação 15](#_Toc488182683)

[5.2.3. Análise de Resultados 16](#_Toc488182684)

[6. Conclusão 17](#_Toc488182685)

[7. Referências 18](#_Toc488182686)

# Introdução

O presente documento tem como objetivo explorar a técnica de data mining Classificação, fazendo um enquadramento teórico e seguindo depois exemplos práticos de forma a consolidar conhecimento e aplicar os conceitos teóricos.

# Estrutura do Documento

Este documento está estruturado em seis capítulos. O primeiro refere-se á introdução do documento. O segundo faz uma pequena introdução ao tema “data mining”. O terceiro capitulo faz um enquadramento ás Regras de Associação. O quarto capítulo contém os casos práticos realizados com Regras de Associação.

# Data Mining

O termo Data Mining, que traduzindo para português resulta em Mineração de Dados, consiste num conjunto de técnicas avançadas para utilizar sobre grandes quantidades de dados á procura de padrões consistentes, de forma que das mesmas seja possível extrair conhecimento [1].

Esse conjunto de técnicas inclui as seguintes:

* Árvores de Decisão (Classificação)
* Regras de Associação
* Redes Neuronais Artificiais
* Algoritmos Genéticos
* Avaliação de Técnicas e Algoritmos

# Árvores de Decisão

A Classificação é uma técnica de mineração de dados usada para classificar uma coleção de dados através da determinação de uma categoria para os mesmos. A utilização desta técnica permite, por exemplo, classificar entidades ou tipificar cenários, tarefas que se tornaram bastante comuns em soluções de suporte á decisão.

Para classificar são utilizadas Árvores de Decisão que correspondem a uma forma de representação de um conjunto de regras e seguem uma hierarquia de classes ou valores, tornando o esquema representado graficamente semelhante a uma árvore. Encaixam-se no tipo de aprendizagem supervisionada.

De forma resumida, as Árvores de Decisão classificam instâncias desde o nó raíz até aos terminais (folhas) e cada nó da árvore especifica um teste para os atributos da instância e cada ramo que descende desse nó corresponde a um dos valores possíveis para esse atributo.

Para construir uma árvore de decisão é necessário um conjunto de dados de treino e um conjunto de dados de teste. O conjunto de dados de treino é utilizado para identificar um modelo que classifica os dados considerando a variável de saída. Este conjunto deve ter casos perfeitamente definidos e corretamente classificados, dado que servem para que o algoritmo possa aprender como se deverá comportar perante situações idênticas. O conjunto de dados de teste serve para verificar o desempenho do algoritmo e avaliar a sua utilidade em tarefas de decisão. Representam a única forma de garantir que a estrutura resultante será bem sucedida em previsões de casos no futuro.

Do ponto de vista de quem vai utilizar o sistema o objetivo será encontrar uma árvore que melhor se adapte ao problema, ou seja, a que melhor classifique as instâncias do domínio em questão.

No caso de uma árvore de decisão não classificar todos os casos de forma correta, as exceções são adicionadas ao conjunto de treino de forma que o algoritmo possa “aprender” essas exceções e considerá-las no futuro.

Há dois tipos de árvores de decisão, as Árvores de Classificação e as Árvores de Regressão.

As Árvores de Classificação servem para qualificar os registos e associá-los com a classe determinada e garantir que essa mesma classificação esteja correta.

As Árvores de Regressão realizam estimativas do valor de uma determinada variável.

Há vários algoritmos de implementação de árvores de decisão que seguem uma metodologia top-down como o CART, CHAID, ID3 ou o C4.5.

# Trabalhos Relacionados

Em vários setores da indústria tem sido cada vez mais aposta a análise avançada de dados, e as Regras de Associação são técnicas utilizadas neste tipo de análises que têm contribuído bastante para aplicabilidades em casos reais.

Segundo artigo relacionado [2], em fábricas de químicos, alarmes podem disparar quando variáveis dos processos ultrapassam determinados limites, no entanto devido á interação de variáveis operacionais de processo é possível que um determinado número de alarmes seja disparado desnecessariamente e consequentemente provoque tempo perdido por parte dos operadores das fábricas. Motivado por esta necessidade foi publicado um artigo que através de mineração de dados é descoberta uma correlação entre as variáveis de processos de sequências de alarmes e desta forma consegue eliminar alarmes redundantes e otimizar o tempo de trabalho dos operadores.

Outro caso que demonstra a utilidade das Regras de Associação é o da aplicação das mesmas para análise de dados clínicos [3]. Neste caso foi feita uma análise a um dataset público e utilizada uma extensão do algoritmo FP-Growth, que permite encontrar padrões frequentes em bases de dados de grande dimensão. Com a utilização desta técnica foi possível retirar conclusões de valor do ponto de vista clínico, como por exemplo o facto de mais de 30% dos casos presentes no dataset que sofrem de hipertensão também apresentam problemas de alta pressão sistólica e de fígado.

Um terceiro caso é o da utilização das Regras de Associação para análise de dados de acidentes rodoviários [4], de forma a melhorar as políticas de segurança rodoviária por parte das autoridades. Foi utilizado um dataset de dados reais de uma das estradas mais utilizadas em Marrocos, de onde se pôde retirar regras como por exemplo se a nacionalidade do condutor era Marroquina e o acidente foi fatal, a causa do acidente foi “Adormecimento” com um grau de confiança de 88%, ou se a nacionalidade era Marroquina e a causa do acidente foi excesso de velocidade, então o acidente não foi fatal, provocando apenas feridos, com um grau de confiança também de 88%. Este tipo de regras pode ser utilizado para efetuar melhorias nas políticas de controlo dessa estrada por parte das autoridades, ou até mesmo para efetuar alterações na própria estrada.

# Casos Práticos

# Dataset Groceries

Este exemplo prático será realizados com a ferramenta de data mining **R** [5]. Esta fornece alguns conjuntos de dados (datasets) que seriam facilmente aplicáveis a cenários reais. Um desses datasets é o da Mercearia (Grocerie) e representa um conjunto de dados comprados ao longo do tempo numa mercearia. Sâo dados transacionais, onde tem apenas uma linha com os produtos separados por vírgulas (ver imagem abaixo), o que significa que os dados não possuem qualquer estrutura.

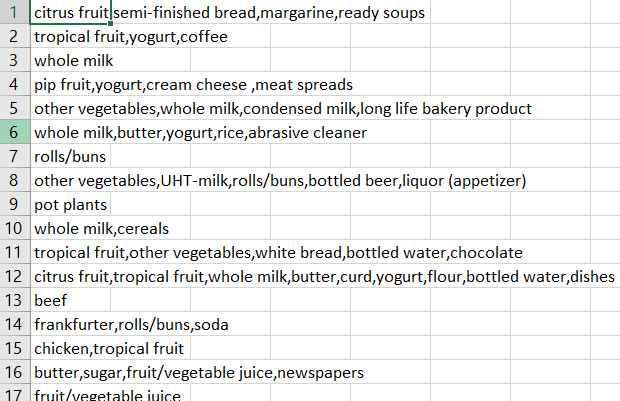


Figura 1 - Groceries Dataset

Com este exemplo prático vou demonstrar como é possível utilizar regras de associação para encontrar relações entre os dados adquiridos na mercearia com um determinado grau de confiança.

# Instalação

O primeiro passo é garantir que temos o pacote de regras de associação da ferramenta R. Para tal, é necessário executar o comando de instalação:

* **install.packages(“arules”)**

De seguida selecionar um “mirror” (ver imagem abaixo) e continuar a instalação até recebermos uma mensagem de instalação completa.

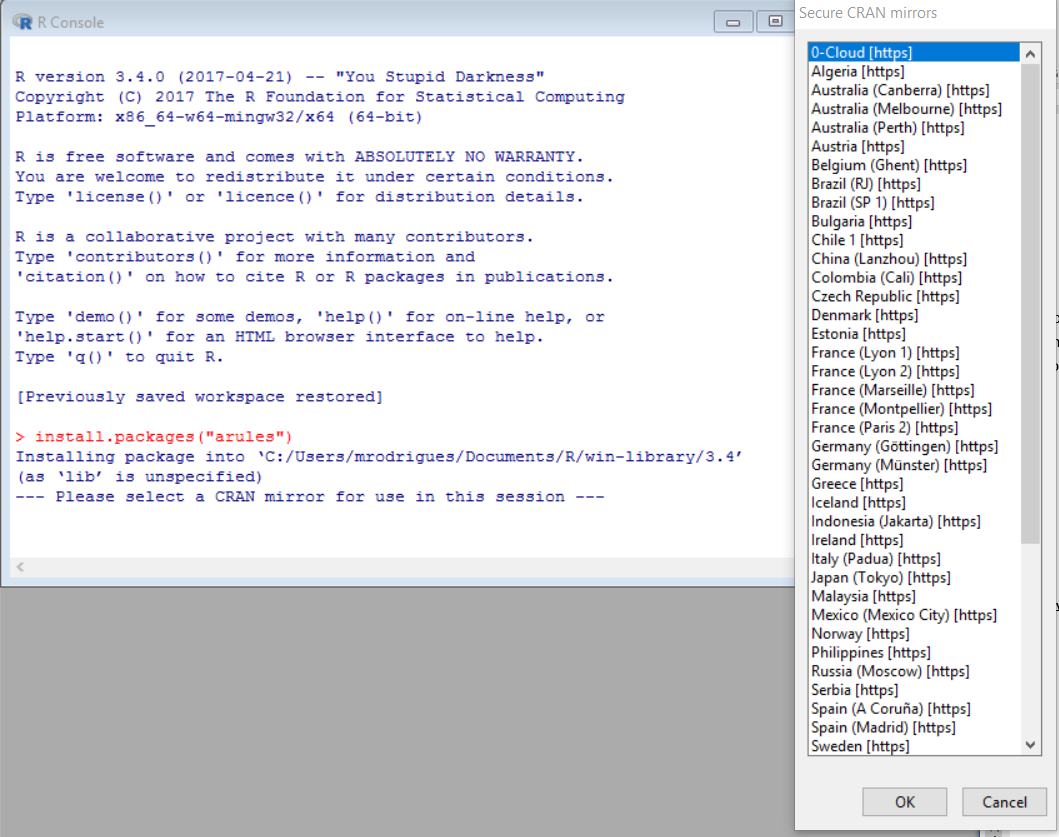


Figura 2 - Instalação do ARULES

De seguida é necessário carregar o pacote para o ambiente de desenvolvimento com o comando:

* **require(arules)**

Para visualizarmos todos os datasets que estão disponíveis para as Regras de Associação, executar o comando:

* **data(package=”arules”)**

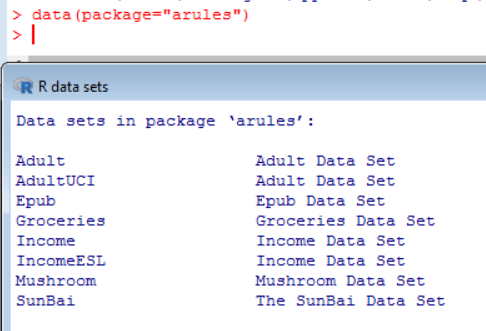


Figura 3 - Datasets para ARULES

Neste exemplo prático vamos instalar o dataset Grocerie, com o comando:

* **data(Groceries)**

De seguida podemos visualizar alguns detalhe acerca do dataset Groceries com o comando:

* **Groceries**

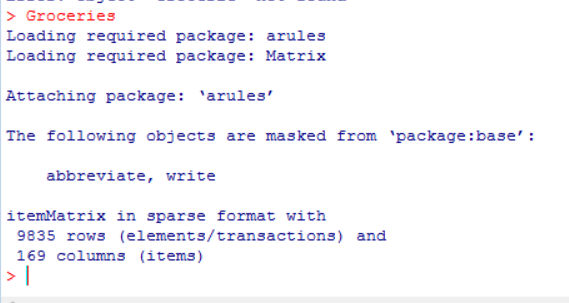


Figura 4 – Groceries

Na imagem podemos constatar que o dataset Groceries tem 9835 registos (linhas ou transações) e 169 colunas (itens). No entanto, através de outro comando conseguimos obter mais alguns detalhes que podem ser bastante úteis numa análise:

* **Summary(Groceries)**

Com este comando podemos constatar algumas coisas interessantes como:

* + Densidade = 0.026 . Sublinhada a vermelho na imagem, a densidade representa a quantidade de informação realmente existente a dividir pela quantidade de informação que poderia existir, isto é, a ferramenta R para ter uma melhor capacidade de análise dos dados e melhor desempenho ao carregar um dataset fá-lo para uma matriz virtual. Cria o número de linhas que o dataset tem e o número de colunas (itens diferentes) necessário. Em grande parte da matriz esta é vazia, porque há bastantes transações com poucos itens, e esta densidade representa a quantidade de informação que existe na matriz.
  + Most Frequent Items: aqui são nos fornecidos quais os itens que mais constam na lista e quantas vezes aparecem. No nosso exemplo em primeiro lugar (a azul) está “whole milk” (leite de vaca normal) que aparece 2513 vezes, seguido de “other vegetables (outros legumes) e rolls/buns (pães).
  + Element length distribution: este detalhe (a preto) agrega-nos a informação por tamanhos de transações e pelo número de vezes que consta no dataset. Isto é, por baixo de “sizes” podemos ver o número 1 e por baixo deste o número de vezes que constam transações feitas para apenas 1 artigo (2159), de seguida o número 2 e o número de vezes com transações de 2 artigos (1643). A transação com mais artigos é de 32 artigos e consta apenas uma vez. Com esta análise é fácil perceber que grande parte das transações são para apenas 1 a 12 artigos. A partir daí é mais raro acontecer. A média exata é de 4.409 itens por transação.

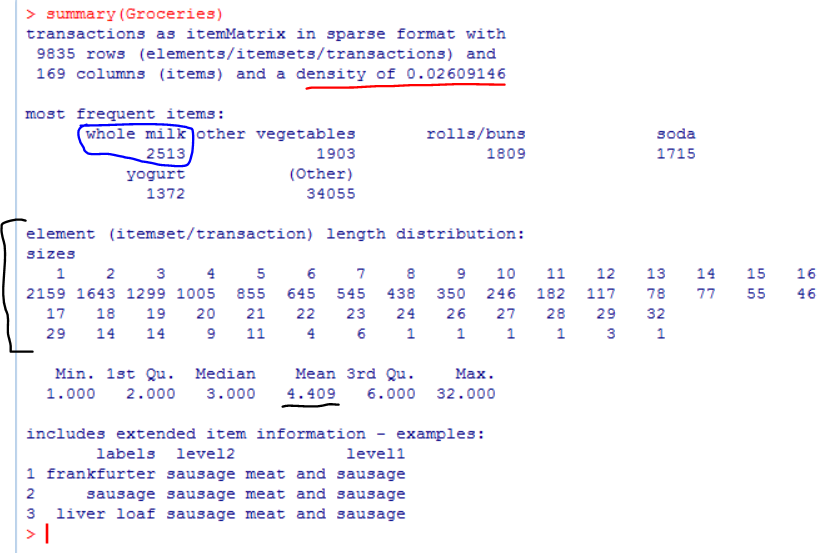


Figura 5 - Sumário do dataset Groceries

Para consultar o conteúdo do dataset devemos utilizar o comando:

* **inspect(Groceries)**: no exemplo vou apenas consultar as primeiras 5 transações por isso acrescento ao comando o intervalo e o comando fica inspect(Groceries[1:5])

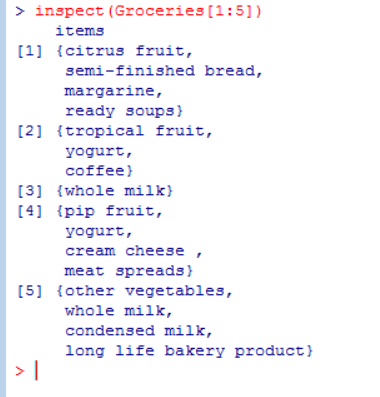


Figura 6 - Inspecionar datasets

# Criação das Regras de Associação

De seguida vamos ver qual a frequência com que um item aparece dado o total de transações. Nas Regras de Associação a frequência corresponde ao conceito de **suporte**. Na ferramenta R executamos o comando:

* **itemFrequency[,1]**: neste caso estamos a fazer a verificação para o primeiro item que aparece na lista, e o resultado é 0.058% para o item frankfurter. Este resultado diz-nos que este item aparece em 580 transações (calculado através da multiplicação pelo número total de transações).

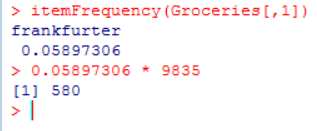


Figura 7 - Frequência de um item

Podemos executar o mesmo comando para um conjunto de itens.

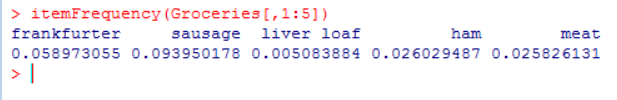


Figura 8 - Frequência de itens

Podemos também executar o mesmo comando e obter o resultado em formato de um gráfico, através de:

* **itemFrequencyPlot(Groceries,support=0.10)**: com a indicação de support=0.10 estou a indicar a fazer o pedido para apenas serem disponibilizados os itens que têm uma frequência igual ou superior a 10%.

No exemplo verificamos que sã vários os itens que respeitam essa restrição, sendo que os mais frequentes são o “whole milk” com pouco mais de 25% de frequência (suporte), seguido de “other vegetables” e “rolls/buns”.

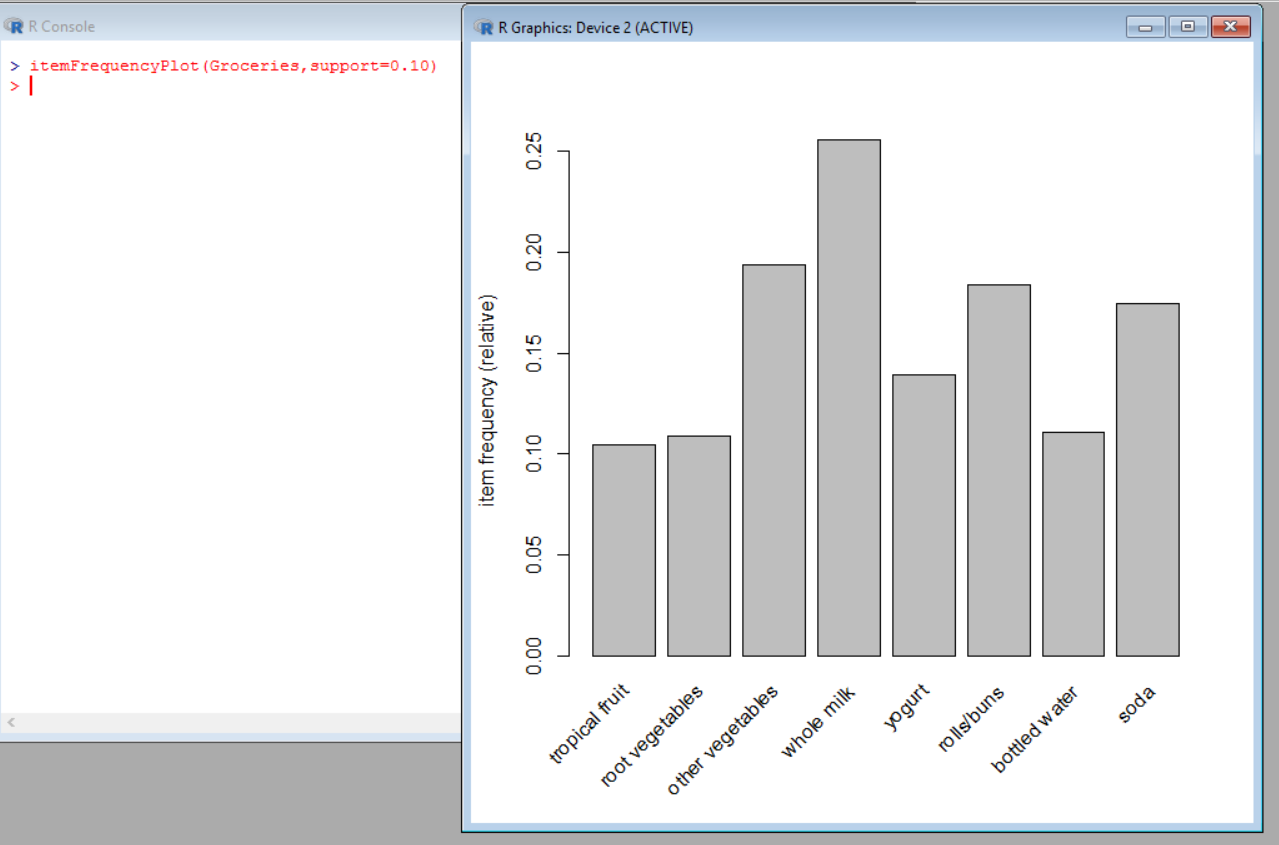


Figura 9 - Gráfico de frequência dos itens

De seguida vamos fazer uma análise aos dados mediante um conceito fundamental nas Regras de Associação, a **confiança**.

A ferramenta R fornece comandos que nos permitem analisar este tipo de associações com os respetivos detalhes. Para começar, devemos criar um modelo obtido com as regras identificadas no dataset, com um determinado grau de suporte e de confiança.

O seguinte comando permite atingir esse objetivo:

* **model1 <- apriori(Groceries,parameter=list(support=0.007,confidence=0.25,minlen=2))**

Analisando o commando, consegue-se perceber que o “model1” é o nome do modelo. Para aplicar as restrições ao dataset e daí obter o conjunto de regras é utilizada a função “apriori”.

Na função “apriori” incluímos o dataset Groceries e os parâmetros:

* + support=0.007: apenas estamos a considerar regras com um suporte mínimo de 0.007
  + confidence=0.25: grau de confiança mínimo de 25%
  + minlen=2: para regras que considerem pelo menos 2 itens distintos

De seguida para obtermos um sumário das regras que agregamos devemos executar o seguinte comando:

* **summary(model1)**

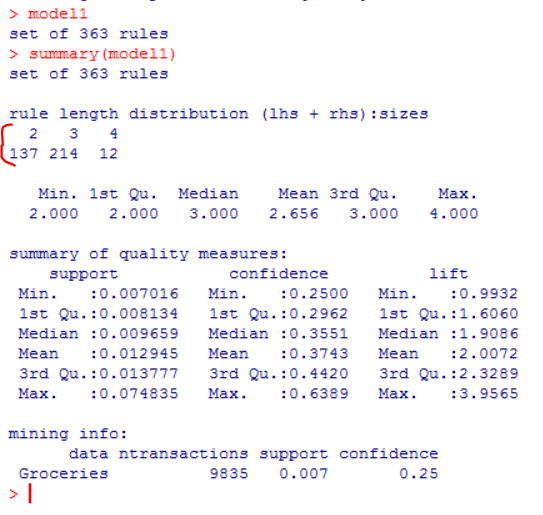


Figura 10 - Sumário das Regras de Associação

Aqui podemos perceber que temos um total de 363 regras de associação agregadas, sendo que 137 delas tem associação entre 2 itens, 214 têm 3 itens e 12 regras têm 4 itens.

Esta informação é interessante, no entanto nós agora queremos saber QUAIS são essas regras, e para tal devemos executar o seguinte comando:

* **inspect(model1[1:5])**: neste exemplo vou visualizar as 5 primeiras regras.

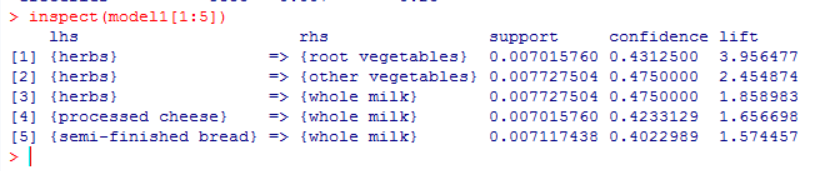


Figura 11 - 5 regras de associação identificadas

Olhando para o exemplo, o “Lift” representa a probabilidade do item “root vegetables” ser adquirido em conjunto com o item “herbs” em relação á probabilidade geral do item “root vegetables” ser adquirido de forma independente.

Dado que o “Lift” do “root vegetables” com o item “hebrs” é 3.95, significa que o item “herbs” aumenta 3.95 vezes a probabilidade de ser adquirido “root vegetables”.

Para pedir á ferramenta que nos retorne as regras com o “Lift” mais alto, podemos utilizar o comando:

* **inspect(sort(model1,by=”lift”)[1:5])**: desta forma obtemos as 5 primeiras regras.

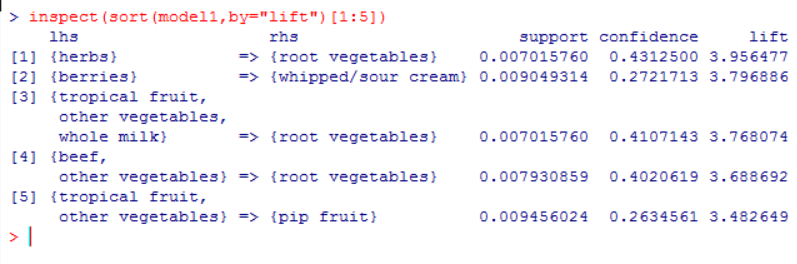


Figura 12 - Regras de Associação por Lift

# Análise de Resultados

Olhando para as Regras de Associação que obtivemos (visível na figura anterior), podemos observar que:

* Quem compra “herbs” aumenta 3.95 vezes a probabilidade de comprar “root vegetables”
* Quem compra “berries” (frutos vermelhos) aumenta 3.79 vezes a probabilidade de comprar “whipped/sour cream” (natas)
* Quem compra “other vegetables”, “tropical fruit” e “whole milk” aumenta 3.76 vezes a probabilidade de comprar “root vegetables”
* Quem compra “beef” e “other vegetables” aumenta 3.68 vezes a probabilidade de comprar “root vegetables”
* Quem compra “tropical fruit” e “other vegetables” aumenta 3.48 vezes a probabilidade de comprar “pip fruit”

Posto isto numa situação real, para a gestão de um supermercado físico, poderia fazer sentido por exemplo colocar os frutos vermelhos junto das natas de forma a potenciar a venda desse conjunto (segunda regra de associação), ou se fosse um hipermercado online poderia ser uma sugestão automática para quem comprasse frutos vermelhos.



# Dataset Titanic

Este exemplo prático será realizado com a ferramenta de data mining **Rstudio** [6].

Como fonte de informação será utilizado um dataset do desastre do Titanic [7] .

# Instalação

Em primeiro lugar devemos carregar o dataset com o comando:

* load("...caminho...\\titanic.raw.rdata")

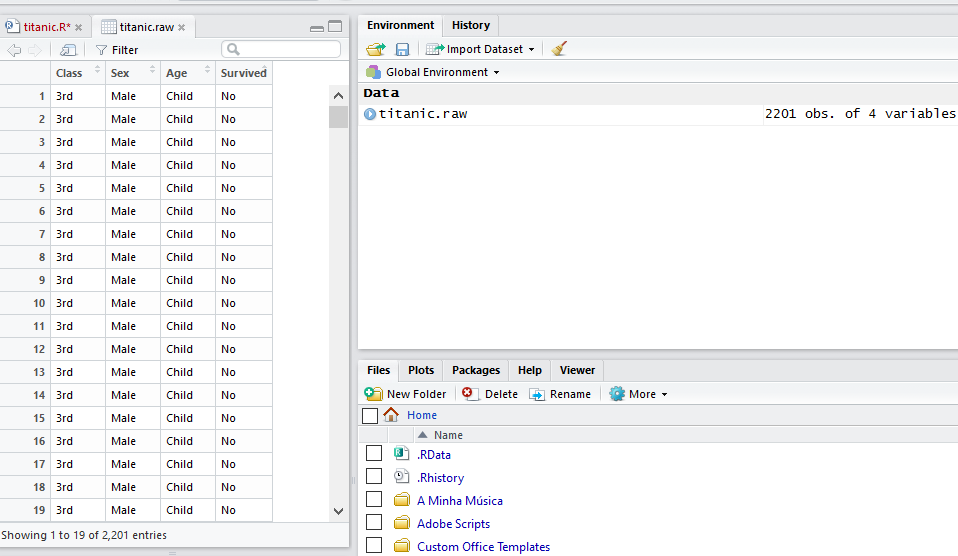


Figura 13 - Carregar Titanic Dataset

Podemos constatar que do lado direito, na janela “Environment” aparecem o número de observações verificadas no dataset e o número de variáveis. Do lado esquerdo vemos o conteúdo do dataset.

De seguida é necessário correr o comando:

* **attach(titanic.raw)**

Isto irá anexar a base de dados titanic ao R, que significa que a ferramenta poderá fazer pesquisas pelos objetos que constam na base de dados.

Os próximos passos são manuais, é necessário ir ao menu “Tools”->”Install Packages” e instalar o pacote “**Matrix**” e “**arules**”. Para ser possível utilizar a o pacote “arules” de forma a executar comandos de regras de associação, ainda é necessário invocar a livraria com:

* **library(arules)**

# Criação das Regras de Associação

O algoritmo **apriori** utiliza pesquisas avançadas para encontrar conjuntos de itens frequentes.

Nesta fase já estamos em condições de encontrar as regras de associação na ferramenta Rstudio com o dataset carregado. Para tal devemos executar os comandos:

* **rules=apriori(titanic.raw)**
* **inspect(rules)**

Com o inspect podemos visualizar as regras encontradas (ver imagem seguinte).

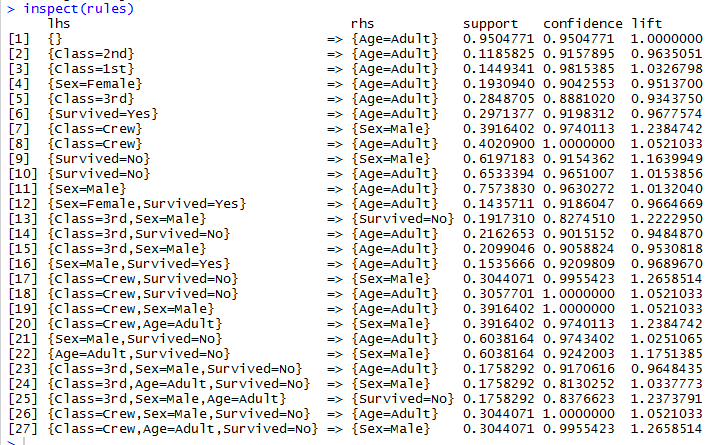


Figura - Regras de Associação encontradas no Titanic dataset

Como é possível constatar, foram encontradas 27 regras de associação.

De seguida vamos procurar apenas regras que nos dêm informação acerca da sobrevivência. Para tal, devemos executar os seguintes comandos:

* **rules <- apriori(titanic.raw, parameter = list(minlen=2, supp=0.005, conf=0.8), appearance = list(rhs=c("Survived=No", "Survived=Yes"), default="lhs"), control = list(verbose=F))**
* **rules.sorted <- sort(rules, by="lift")**
* **inspect(rules.sorted)**

Com estes comandos estamos a introduzir a aparência da condição “Survived” nas regras de associação. O resultado é o seguinte:

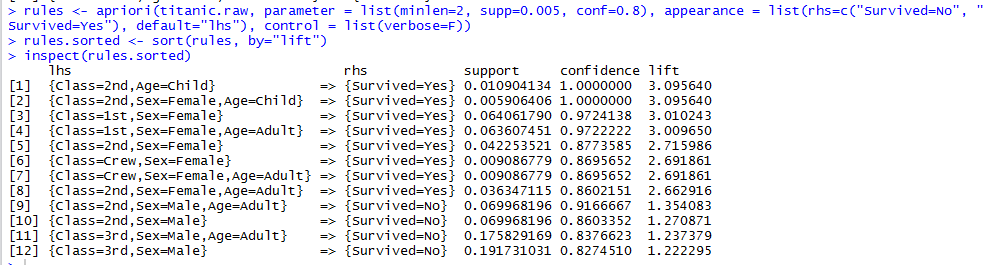
****

Figura - Regras de Associação com parâmetro "Survived"

# Análise de Resultados

Após a obtenção das Regras de Associação, podemos constatar que:

* Quem viajou em segunda classe e era criança (1%), sobreviveu com 100% de grau de confiança. E se era segunda classe e criança, a probabilidade de ter sobrevivido aumentava 3.09 vezes. (Figura 15 – Regra 1)
* Quem viajou em primeira classe e era do sexo feminino (6.4%), sobreviveu com grau de confiança de 97%, sendo que viajar em primeira classe e ser do sexo feminino aumentava em 3.01 vezes a probabilidade de sobreviver. (Figura 15 – Regra 3)
* 19% (suporte) dos casos totais são do sexo feminino e da faixa etária adulta, com um grau de confiança de 90% (confidence) um lift que nos diz que 95% dos casos de sexo feminino são adultos. (Figura 14 – Regra 4)
* 40% (suporte) dos casos totais pertenciam á classe “Crew” (tripulação) e são adultos. O grau de confiança é de 100%, e o lift diz-nos que se for da classe “Crew” é altamente provável (105%) de ser da faixa etária adulta. (Figura 14 – Regra 8)

# Conclusão

Neste trabalho foram realizados dois casos práticos utilizando as Regras de Associação como técnica de mineração de dados a dois datasets diferentes.

No primeiro caso foi utilizado o dataset “Grocerie” que representava um conjunto de dados de compras numa mercearia. Após a aplicação de técnicas para a descoberta de Regras de Associação, os resultados obtidos foram os seguintes: quem compra “herbs” aumenta 3.95 vezes a probabilidade de comprar “root vegetables”, quem compra “berries” (frutos vermelhos) aumenta 3.79 vezes a probabilidade de comprar “whipped/sour cream” (natas), quem compra “other vegetables”, “tropical fruit” e “whole milk” aumenta 3.76 vezes a probabilidade de comprar “root vegetables”, quem compra “beef” e “other vegetables” aumenta 3.68 vezes a probabilidade de comprar “root vegetables” e quem compra “tropical fruit” e “other vegetables” aumenta 3.48 vezes a probabilidade de comprar “pip fruit”.

No segundo caso foi utilizado o dataset relativo á tragédia do Titanica e após a aplicação de técnicas para a descoberta de Regras de Associação, foi possível concluir que: quem viajou em segunda classe e era criança (1%), sobreviveu com 100% de grau de confiança. E se era segunda classe e criança, a probabilidade de ter sobrevivido aumentava 3.09 vezes, quem viajou em primeira classe e era do sexo feminino (6.4%), sobreviveu com grau de confiança de 97%, sendo que viajar em primeira classe e ser do sexo feminino aumentava em 3.01 vezes a probabilidade de sobreviver, 19% (suporte) dos casos totais são do sexo feminino e da faixa etária adulta com um grau de confiança de 90% (confidence) um lift que nos diz que 95% dos casos de sexo feminino são adultos, 40% (suporte) dos casos totais pertenciam á classe “Crew” (tripulação) e são adultos, o grau de confiança é de 100%, e o lift diz-nos que se for da classe “Crew” é altamente provável (105%) de ser da faixa etária adulta.

# Referências

[1] – Bibliografia sobre Data Mining. Disponibilizada pelo Prof. Jorge Ribeiro na página do Moodle da disciplina de Business Intelligence e Mineração de Dados do Mestrado em Engenharia de Software.

[2] – Artigo sobre a utilização de Regras de Associação em fábrica de químicos. J. Wang, Li, Huang, & Su, “Association rules mining based analysis of consequential alarm

sequences in chemical processes”, Mar.2016

[3] – Artigo sobre a utilização de Regras de Associação para análise de dados clínicos. B. Wang et al., “Comprehensive Association Rules Mining of Health

Examination Data with an Extended FP-Growth Method”, Jan.2017

[4] – Artigo sobre a utilização de Regras de Associação para análise de dados sobre acidentes rodoviários numa estrada em Marrocos. Addi, Tarik, & Fatima, “An Approach Based On Association Rules Mining To Improve Road Safety In Morocco”, Apr.2016

[5] – Sítio da ferramenta de Data Mining R. Disponível em: <https://www.r-project.org/>

[6] – Sítio da ferramenta de Data Mining Rstudio. Disponível em: <https://www.rstudio.com/>

[7 ] – Sítio dedicado a Data Mining e que disponibiliza o dataset Titanic. Disponível em: <http://www.rdatamining.com/data>