

Analytics e Inteligência Artificial Data Science

Tema da aula
Análise de Cesta de Compras



BUSINESS SCHOOL

Graduação, pós-graduação,
MBA, Pós- MBA, Mestrado
Profissional, Curso In
Company e EAD



CONSULTING

Consultoria personalizada
que oferece soluções
baseadas em seu
problema de negócio



RESEARCH

Atualização dos
conhecimentos e do material
didático oferecidos nas
atividades de ensino



Líder em Educação Executiva, referência de ensino nos cursos de graduação, pós-graduação e MBA, tendo excelência nos programas de educação. Uma das principais **escolas de negócio do mundo**, possuindo convênios internacionais com Universidades nos EUA, Europa e Ásia. +8.000 **projetos de consultorias** em organizações públicas e privadas.



Único curso de graduação em administração a receber as notas máximas



A primeira escola brasileira a ser finalista da maior competição de MBA do mundo



Única *Business School* brasileira a figurar no *ranking* LATAM



Signatária do Pacto Global da ONU



Membro fundador da ANAMBA - Associação Nacional MBAs



Credenciada pela AMBA - Association of MBAs



Credenciada ao Executive MBA Council



Filiada a AACSB - Association to Advance Collegiate Schools of Business



Filiada a EFMD - European Foundation for Management Development



Referência em cursos de MBA nas principais mídias de circulação



O **Laboratório de Análise de Dados** – LABDATA é um Centro de Excelência que atua nas áreas de ensino, pesquisa e consultoria em análise de informação utilizando técnicas de **Big Data, Analytics** e **Inteligência Artificial**.



Profª Drª Alessandra Montini

O LABDATA é um dos pioneiros no lançamento dos cursos de *Big Data* e *Analytics* no Brasil. Os diretores foram professores de grandes especialistas do mercado.

- +10 anos de atuação.
- +9.000 alunos formados.

Docentes

- Sólida formação acadêmica: doutores e mestres em sua maioria;
- Larga experiência de mercado na resolução de *cases*;
- Participação em congressos nacionais e internacionais;
- Professor assistente que acompanha o aluno durante todo o curso.

Estrutura

- 100% das aulas realizadas em laboratórios;
- Computadores para uso individual durante as aulas;
- 5 laboratórios de alta qualidade (investimento +R\$2MM);
- 2 unidades próximas à estação de metrô (com estacionamento).



PROFA. DRA. ALESSANDRA DE ÁVILA MONTINI

Diretora do LABDATA-FIA, apaixonada por dados e pela arte de lecionar. Tem muito orgulho de ter criado na FIA cinco laboratórios para as aulas de Big Data e Inteligência Artificial. Possui mais de 20 anos de trajetória nas áreas de Data Mining, Big Data, Inteligência Artificial e Analytics. Cientista de dados com carreira realizada na Universidade de São Paulo. Graduada e mestra em Estatística Aplicada pelo IME-USP e doutora pela FEA-USP. Com muita dedicação chegou ao cargo de professora e pesquisadora na FEA-USP, ganhou mais de 30 prêmios de excelência acadêmica pela FEA-USP e mais de 30 prêmios de excelência acadêmica como professora dos cursos de MBA da FIA. Orienta alunos de mestrado e de doutorado na FEA-USP. Parecerista da FAPESP e colunista de grandes portais de tecnologia.





PROF. ÂNGELO CHIODE, MSc

Bacharel, mestre e candidato ao PhD em Estatística (IME-USP), atua como professor de Estatística Aplicada para turmas de especialização, pós-graduação e MBA na FIA. Trabalha como consultor nas áreas de Analytics e Ciência de Dados há 13 anos, apoiando empresas na resolução de desafios de negócio nos contextos de finanças, aquisição, seguros, varejo, tecnologia, aviação, telecomunicações, entretenimento e saúde. Nos últimos 5 anos, tem atuado na gestão corporativa de times de Analytics, conduzindo projetos que envolviam análise estatística, modelagem preditiva e *machine learning*. É especializado em técnicas de visualização de dados e design da informação (Harvard) e foi indicado ao prêmio de Profissional do Ano na categoria Business Intelligence, em 2019, pela Associação Brasileira de Agentes Digitais (ABRADi).



Conteúdo Programático

6



DISCIPLINAS



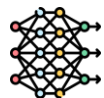
**IA E TRANSFORMAÇÃO
DIGITAL**



ANALYTICS



**INTELIGÊNCIA ARTIFICIAL:
MACHINE LEARNING**



**INTELIGÊNCIA ARTIFICIAL:
DEEP LEARNING**



**EMPREENDEDORISMO E
INOVAÇÃO**



**COMPORTAMENTO
HUMANO E SOFT SKILLS**

TEMAS: ANALYTICS E MACHINE LEARNING

ANÁLISE EXPLORATÓRIA DE DADOS

INFERÊNCIA ESTATÍSTICA

TÉCNICAS DE PROJEÇÃO

TÉCNICAS DE CLASSIFICAÇÃO

TÓPICOS DE MODELAGEM

TÉCNICAS DE SEGMENTAÇÃO

TÓPICOS DE ANALYTICS

MANIPULAÇÃO DE BASE DE DADOS

AUTO ML

TEMAS: DEEP LEARNING

REDES DENSAS

REDES CONVOLUCIONAIS

REDES RECORRENTES

MODELOS GENERATIVOS

FERRAMENTAS

LINGUAGEM R

LINGUAGEM PYTHON

DATABRICKS



Conteúdo da Aula

- 1. Introdução
- 2. Objetivo
- 3. Indicadores Gerais
- 4. Algoritmo Apriori
- Referências Bibliográficas



1. Introdução



Case: Supermercado

1. INTRODUÇÃO | ANÁLISE DE CESTA DE COMPRAS

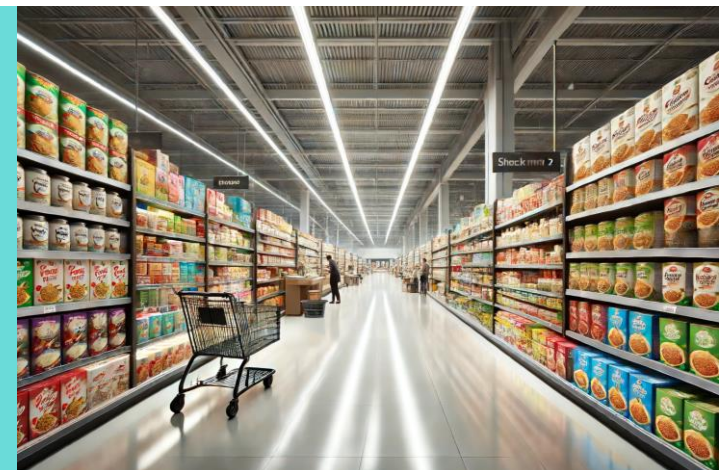
9

Exemplo:

Analisar padrões de compra de produtos alimentícios para otimizar o layout de supermercados, colocando itens frequentemente comprados juntos em locais adjacentes para aumentar a conveniência do cliente e as vendas.

Aplicação:

Varejo alimentar



Case: Higiene Pessoal

1. INTRODUÇÃO | ANÁLISE DE CESTA DE COMPRAS

10

Exemplo:

Estudar a associação entre diferentes produtos higiene pessoal comercializados em uma farmácia, para criar promoções combinadas (“combos”) e aumentar a atratividade das ofertas.

Aplicação:

Varejo farmacêutico



Case: Eletrônicos e Acessórios

1. INTRODUÇÃO | ANÁLISE DE CESTA DE COMPRAS

11

Exemplo:

Avaliar a relação entre produtos eletrônicos e acessórios para identificar oportunidades de *cross-selling*, sugerindo acessórios complementares para o cliente no ato da compra e impulsionando a receita.

Aplicação:

Varejo de eletrônicos



2. Objetivo





Objetivo

2. OBJETIVO | ANÁLISE DE CESTA DE COMPRAS

13

A **análise de cesta de compras** (*market basket analysis*, ou MBA) é uma técnica estatística que visa identificar padrões de **presença concomitante** de itens (ou grupos de itens) em transações de vendas, especialmente em varejo.

Exemplos:

- Quando os clientes de um *e-commerce* de moda esportiva compram **tênis de corrida**, eles também levam **garrafas reutilizáveis** e **fones de ouvido sem fio**?
- Quando os clientes de um restaurante pedem *delivery* de **pratos veganos**, eles costumam pedir também **sucos detox** e **sobremesas sem glúten**?
- Quando os clientes de uma loja de eletrônicos compram **câmeras fotográficas**, eles são suscetíveis a comprar itens como **tripés** e **cartões de memória**?

Nesta aula, vamos estudar os principais indicadores utilizados em análise de cestas de compras, bem como um algoritmo que identifica associações frequentes e gera regras de associação.





Objetivo

2. OBJETIVO | ANÁLISE DE CESTA DE COMPRAS

14

As análises de cesta de compras lidam com dados a nível das **transações** realizadas no histórico, e **não** com dados em visão consolidada de **clientes únicos**. Dessa forma, identifica-se oportunidades de venda de itens associados (*cross-sell*) em uma mesma ocasião, e não potenciais itens que serão comprados no futuro.

Exemplo: Amazon

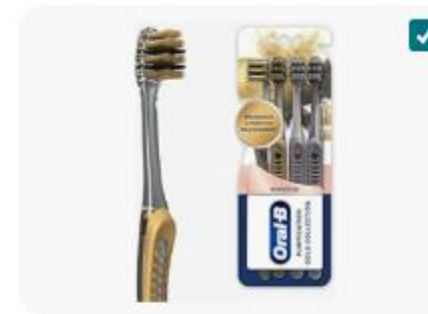
Frequentemente comprados juntos



Este item: Oral-B Creme Dental
Clareador 3D White Perfection
102G

R\$ 14⁷² (R\$ 0,14/Grama)

Item visualizado/
colocado na cesta



Oral-B Escova Dental Purification
Gold Collection - 4 Unidades

R\$ 29⁹⁰ (R\$ 7,48/Unidade)

Item recomendado
na mesma compra



Objetivo

2. OBJETIVO | ANÁLISE DE CESTA DE COMPRAS

Já os **modelos de recomendação** são algoritmos mais complexos que se baseiam no histórico de transações anteriores a nível de indivíduos, a fim de prever por quais outros itens/produtos eles podem se interessar futuramente.

Exemplo: Spotify



Novos itens recomendados com
base no histórico



3. Indicadores Principais





Indicadores Principais

3. INDICADORES GERAIS | ANÁLISE DE CESTA DE COMPRAS

17

Os três indicadores mais utilizados em análise de cesta de compras são:

- **Suporte**
- **Confiança**
- **Lift**





Suporte

3. INDICADORES GERAIS | ANÁLISE DE CESTA DE COMPRAS

18

O indicador de **suporte** consiste na proporção de transações em que houve compra concomitante de itens específicos.

O suporte de uma combinação de **dois itens** (A, B) é calculado como:

$$\text{Suporte (A, B)} = \frac{\# \text{ de transações com os itens A + B}}{\# \text{ total de transações}}$$

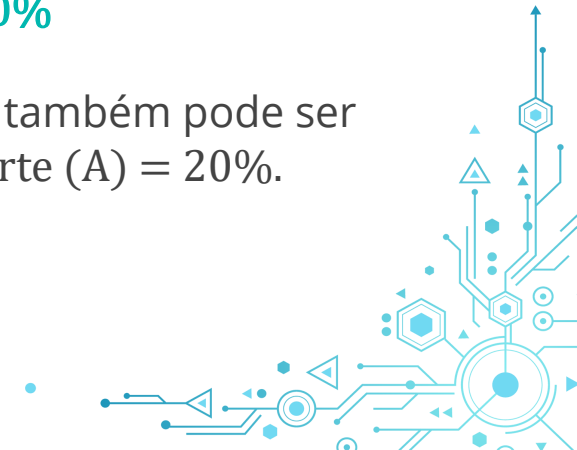
Exemplo:

De um total de 1.000 transações em uma empório de alimentos:

- 200 englobavam algum tipo de queijo (A)
- 150 englobavam algum item de charcutaria – presunto, salame etc. (B)
- 100 englobavam tanto algum queijo quanto algum item de charcutaria (A e B)

$$\text{Suporte (A, B)} = 100 / 1.000 = 0,10 = \mathbf{10\%}$$

Obs.: A frequência relativa de ocorrência de um item isolado também pode ser escrita em termos da medida de suporte. Por exemplo, Suporte (A) = 20%.





Suporte

3. INDICADORES GERAIS | ANÁLISE DE CESTA DE COMPRAS

19

O indicador de **suporte** consiste na proporção de transações em que houve compra concomitante de itens específicos.

O suporte de uma combinação de **dois ou mais itens** (A, B, ...) é calculado como:

$$\text{Suporte (A, B, ...)} = \frac{\text{\# de transações com os itens A + B + ...}}{\text{\# total de transações}}$$





Confiança

3. INDICADORES GERAIS | ANÁLISE DE CESTA DE COMPRAS

20

O indicador de **confiança** consiste na proporção de transações em que houve compra de um determinado item em concomitância com outros, em relação ao total de transações que incluíam apenas estes outros itens.

A confiança associada ao item B dada a presença **de um item** A é dada por:

$$\text{Confiança } (A \Rightarrow B) = \frac{\# \text{ de transações com os itens } A + B}{\# \text{ de transações com os itens } A}$$

Exemplo:

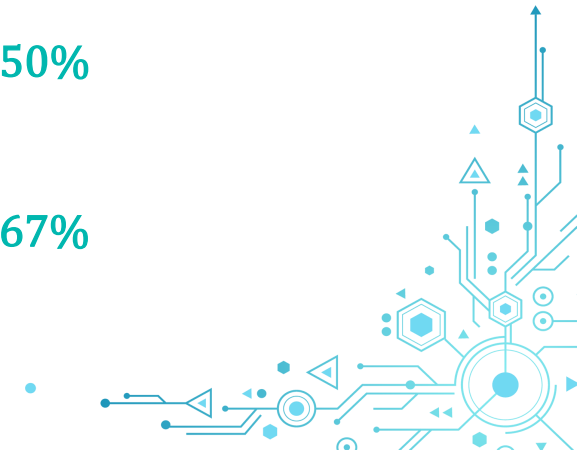
De um total de 1.000 transações em uma empório de alimentos:

- 200 englobavam algum tipo de queijo (A)
- 150 englobavam algum item de charcutaria – presunto, salame etc. (B)
- 100 englobavam tanto algum queijo quanto algum item de charcutaria (A e B)

$$\text{Confiança } (A \Rightarrow B) = 100 / 200 = 0,50 = \mathbf{50\%}$$

Por outro lado, note que:

$$\text{Confiança } (B \Rightarrow A) = 100 / 150 \approx 0,67 = \mathbf{67\%}$$





Confiança

3. INDICADORES GERAIS | ANÁLISE DE CESTA DE COMPRAS

21

O indicador de **confiança** consiste na proporção de transações em que houve compra de um determinado item em concomitância com outros, em relação ao total de transações que incluíam apenas estes outros itens.

A confiança associada ao item B dada a presença **dos itens** A1, A2, ... é dada por:

$$\text{Confiança } (A1, A2, \dots \Rightarrow B) = \frac{\# \text{ de transações com os itens } A1 + A2 + \dots + B}{\# \text{ de transações com os itens } A1 + A2 + \dots}$$



Lift

3. INDICADORES GERAIS | ANÁLISE DE CESTA DE COMPRAS

O indicador de **lift** consiste no incremento associado à proporção de transações históricas em que houve compra concomitante de itens específicos, em relação à proporção esperada caso esses itens fossem comprados de forma totalmente independente (ou não associada).

O *lift* de uma combinação de **dois itens** (A, B, ...) é calculado como:

$$\text{Lift (A, B)} = \frac{\text{Suporte (A, B)}}{\text{Suporte (A)} \cdot \text{Suporte (B)}}$$

Exemplo:

De um total de 1.000 transações em uma empório de alimentos:

- 200 englobavam algum tipo de queijo (A)
- 150 englobavam algum item de charcutaria – presunto, salame etc. (B)
- 100 englobavam tanto algum queijo quanto algum item de charcutaria (A e B)

$$\text{Lift (A, B)} = 0,10 / (0,20 \cdot 0,15) = 0,10 / 0,03 \approx \mathbf{3,3}$$

Ou seja, os itens A e B aparecem juntos com uma frequência que corresponde a **3,3 vezes** a frequência esperada caso não houvesse associação entre eles.



Lift

3. INDICADORES GERAIS | ANÁLISE DE CESTA DE COMPRAS

23

O indicador de **lift** consiste no incremento associado à proporção de transações históricas em que houve compra concomitante de itens específicos, em relação à proporção esperada caso esses itens fossem comprados de forma totalmente independente (ou não associada).

O *lift* de uma combinação de **dois ou mais itens** (A, B, ...) é calculado como:

$$\text{Lift (A, B, ...)} = \frac{\text{Suporte (A, B, ...)}}{\text{Suporte (A)} \cdot \text{Suporte (B)} \cdot \dots}$$





Embasamento: Probabilidade

3. INDICADORES GERAIS | ANÁLISE DE CESTA DE COMPRAS

Os cálculos de suporte, confiança e *lift* têm como origem a **teoria das probabilidades**, que estabelece algumas notações e relações úteis.

Eventos

- Evento A : O item A estar presente em uma transação
- Evento B : O item B estar presente em uma transação
- Evento $A \cap B$: Os itens A e B estarem presentes em uma transação
Também chamado de **intersecção** entre os eventos A e B

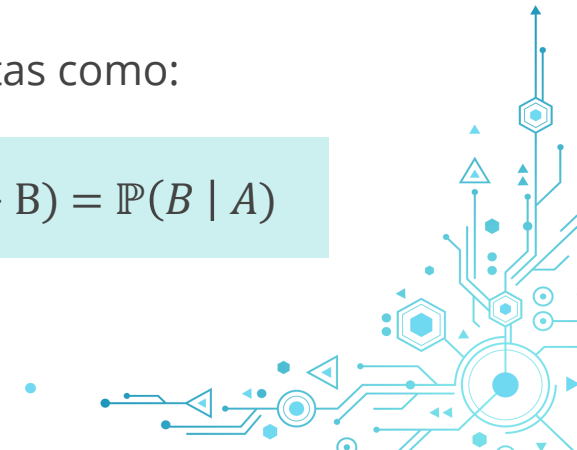
Probabilidades

- $\mathbb{P}(A)$: Probabilidade de o evento A ocorrer
- $\mathbb{P}(B)$: Probabilidade de o evento B ocorrer
- $\mathbb{P}(A \cap B)$: Probabilidade de a intersecção entre os eventos A e B ocorrer
- $\mathbb{P}(B | A)$: Probabilidade de o evento B ocorrer, dado que o evento A ocorreu
Também chamada de **probabilidade condicional** de B dado A

Então, as fórmulas de **suporte** e **confiança** podem ser escritas como:

$$\text{Suporte } (A, B) = \mathbb{P}(A \cap B)$$

$$\text{Confiança } (A \Rightarrow B) = \mathbb{P}(B | A)$$





Embasamento: Probabilidade

3. INDICADORES GERAIS | ANÁLISE DE CESTA DE COMPRAS

25

Os cálculos de suporte, confiança e *lift* têm como origem a **teoria das probabilidades**, que estabelece algumas notações e relações úteis.

Lei da Multiplicação

- Quando dois eventos A e B são independentes (ou seja, a ocorrência de um não altera a perspectiva a respeito da ocorrência do outro), a probabilidade associada à intersecção $A \cap B$ é igual ao **produto** das probabilidades de A e B .

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

Com base nesta lei, a fórmula do ***lift*** poderia ser escrita como:

$$\text{Lift}(A, B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A \cap B) \text{ sob independência}}$$



4. Algoritmo Apriori





Algoritmo Apriori

4. ALGORITMO APRIORI | ANÁLISE DE CESTA DE COMPRAS

27

O **algoritmo apriori** é uma técnica de mineração de dados utilizada para encontrar regras de associação em grandes conjuntos de dados transacionais.

O funcionamento do algoritmo envolve duas etapas principais:

- Primeiro, o algoritmo identifica todos os conjuntos de itens que aparecem com determinada **frequência mínima** (ou seja, **suporte mínimo**) no banco de dados.
- Em seguida, os conjuntos identificados no primeiro passo são utilizados para gerar **regras de associação** entre itens, com determinada **confiança mínima**.

Ao final do processo, recomenda-se **filtrar** apenas as regras com **lift maior que 1**, ou seja, cujos itens aparecem juntos com maior frequência do que seria esperado caso fossem independentes.

Por fim, **ordena-se** as regras identificadas da maior para a menor confiança, ou do maior para o menor *lift*.



Case: Pet Shop

4. ALGORITMO APRIORI | ANÁLISE DE CESTA DE COMPRAS

28

Uma rede comercializadora de produtos para pets gostaria de explorar a associação existente entre as principais categorias de produtos nas cestas de compras de seus clientes. Para isso, selecionou uma base de dados com 74.285 compras realizadas nos últimos 6 meses. Estas compras possuíam itens de 2 a 4 categorias distintas, de um total de 25 categorias.



Categorias de produtos

- Ração seca
- Ração úmida
- Petiscos
- Brinquedos
- Camas
- Casinhas
- Tapetes higiênicos
- Coleiras
- Peitorais
- Guias
- Comedouros
- Bebedouros
- Shampoos
- Condicionadores
- Higiene bucal
- Vitaminas e suplementos
- Antipulgas e carrapatos
- Escovas e pentes
- Caixas de transporte
- Cobertores
- Roupas
- Fraldas
- Cortadores de unha
- Desinfetantes de ambientes
- Perfumes

Arquivo: Pet_Shop.txt

@LABDATA FIA. Copyright all rights reserved.



Case: Pet Shop

4. ALGORITMO APRIORI | ANÁLISE DE CESTA DE COMPRAS

29

Uma rede comercializadora de produtos para pets gostaria de explorar a associação existente entre as principais categorias de produtos nas cestas de compras de seus clientes. Para isso, selecionou uma base de dados com 74.285 compras realizadas nos últimos 6 meses. Estas compras possuíam itens de 2 a 4 categorias distintas, de um total de 25 categorias.



ID_TRANSACAO	GRUPO_ITENS
1	racao_umida
1	guias
1	caixas_transporte
2	racao_seca
2	racao_umida
2	brinquedos
3	petiscos
3	roupas
3	vitaminas_suplementos
4	guias
...	...

Arquivo: Pet_Shop.txt

@LABDATA FIA. Copyright all rights reserved.



Preparação da base

Antes de iniciar as análises, precisamos transformar a base de dados da visão granular (cada categoria comprada em cada cesta) para uma visão **consolidada por cesta**.

Em linguagem R:

```
transacoes <- as(split(dados_pet_shop$GRUPO_ITENS, dados_pet_shop$ID_TRANSACAO), "transactions")
```

Arquivo: Pet_Shop.txt

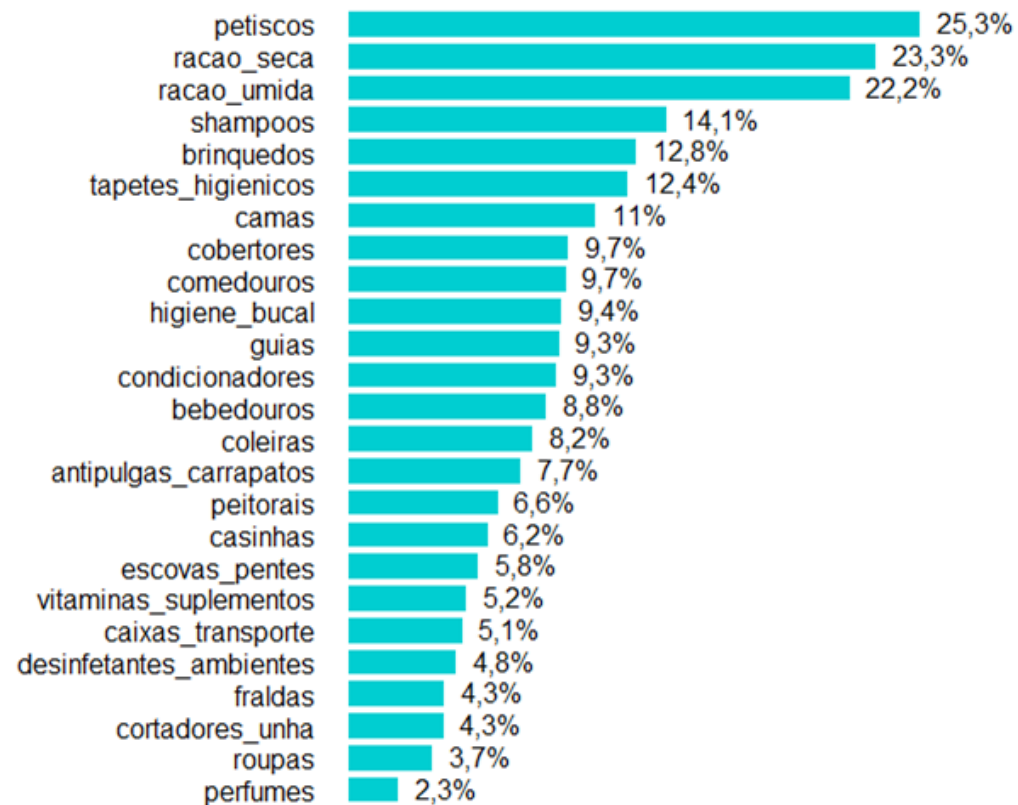
@LABDATA FIA. Copyright all rights reserved.



Análise exploratória

Qtde. de categorias	Qtde. de transações
2	50.397 (68%)
3	17.113 (23%)
4	6.775 (9%)
Total	74.285 (100%)

Distribuição de frequências das categorias de produtos



Arquivo: Pet_Shop.txt



Algoritmo apriori

Vamos utilizar o algoritmo apriori para identificar regras de associação com suporte mínimo de **0,01%** e confiança mínima de **10%**. Isso significa que:

- *Suporte*: Para que $A \rightarrow B$ seja uma regra, as categorias A e B devem aparecer juntas em ao menos **0,01%** do total de cestas – ou seja, no mínimo, 74 cestas.
- *Confiança*: Para que $A \rightarrow B$ seja uma regra, a proporção de cestas com as categorias A e B deve corresponder a, pelo menos, **10%** do total de cestas com a categoria A.

Em linguagem R:

```
rules <- apriori(transacoes, parameter = list(supp = 0.001, conf = 0.01, minlen = 2))
rules <- rules[!is.redundant(rules)] # Remoção de regras redundantes*
rules <- subset(rules, lift > 1)    # Seleção de regras com lift maior do que 1
summary(rules)                    # Resumo das regras identificadas
```

* Observação: Dada uma regra de associação $A + B \rightarrow X$, diz-se que a regra $A + B + C \rightarrow X$ é **redundante** se tiver confiança menor ou igual à da primeira regra.

Arquivo: Pet_Shop.txt



Algoritmo apriori

set of 20 rules

rule length distribution (lhs + rhs):sizes

2
20

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2	2	2	2	2	2

summary of quality measures:

support	confidence	coverage	lift	count
Min. :0.005937	Min. :0.1052	Min. :0.04251	Min. :1.833	Min. : 441
1st Qu.:0.010285	1st Qu.:0.1521	1st Qu.:0.06428	1st Qu.:2.028	1st Qu.: 764
Median :0.025005	Median :0.2517	Median :0.09389	Median :2.509	Median :1858
Mean :0.028580	Mean :0.2449	Mean :0.10343	Mean :2.470	Mean :2123
3rd Qu.:0.027152	3rd Qu.:0.2823	3rd Qu.:0.12483	3rd Qu.:2.856	3rd Qu.:2017
Max. :0.102591	Max. :0.4631	Max. :0.25259	Max. :3.304	Max. :7621

mining info:

data	ntransactions	support	confidence
transacoes	74285	0.001	0.01

Arquivo: Pet_Shop.txt



Case: Pet Shop

4. ALGORITMO APRIORI | ANÁLISE DE CESTA DE COMPRAS

34

Algoritmo apriori

set of 20 rules

20 regras de associação identificadas

```
rule length distribution (lhs + rhs):sizes
  2
20
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2	2	2	2	2	2

summary of quality measures:

support	confidence	coverage	lift	count
Min. :0.005937	Min. :0.1052	Min. :0.04251	Min. :1.833	Min. : 441
1st Qu.:0.010285	1st Qu.:0.1521	1st Qu.:0.06428	1st Qu.:2.028	1st Qu.: 764
Median :0.025005	Median :0.2517	Median :0.09389	Median :2.509	Median :1858
Mean :0.028580	Mean :0.2449	Mean :0.10343	Mean :2.470	Mean :2123
3rd Qu.:0.027152	3rd Qu.:0.2823	3rd Qu.:0.12483	3rd Qu.:2.856	3rd Qu.:2017
Max. :0.102591	Max. :0.4631	Max. :0.25259	Max. :3.304	Max. :7621

mining info:

data	ntransactions	support	confidence
transacoes	74285	0.001	0.01

Arquivo: Pet_Shop.txt

@LABDATA FIA. Copyright all rights reserved.



Case: Pet Shop

4. ALGORITMO APRIORI | ANÁLISE DE CESTA DE COMPRAS

35

Algoritmo apriori

set of 20 rules

rule length distribution (lhs + rhs):sizes

2
20

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2	2	2	2	2	2

Todas as regras são de comprimento 2, ou seja, envolvem apenas 2 categorias de produtos (do tipo A → B)

summary of quality measures:

support	confidence	coverage	lift	count
Min. :0.005937	Min. :0.1052	Min. :0.04251	Min. :1.833	Min. : 441
1st Qu.:0.010285	1st Qu.:0.1521	1st Qu.:0.06428	1st Qu.:2.028	1st Qu.: 764
Median :0.025005	Median :0.2517	Median :0.09389	Median :2.509	Median :1858
Mean :0.028580	Mean :0.2449	Mean :0.10343	Mean :2.470	Mean :2123
3rd Qu.:0.027152	3rd Qu.:0.2823	3rd Qu.:0.12483	3rd Qu.:2.856	3rd Qu.:2017
Max. :0.102591	Max. :0.4631	Max. :0.25259	Max. :3.304	Max. :7621

mining info:

data	ntransactions	support	confidence
transacoes	74285	0.001	0.01

Arquivo: Pet_Shop.txt



Algoritmo apriori

set of 20 rules

rule length distribution (lhs + rhs):sizes

2
20

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2	2	2	2	2	2

summary of quality measures:

support	confidence	coverage	lift	count
Min. :0.005937	Min. :0.1052	Min. :0.04251	Min. :1.833	Min. : 441
1st Qu.:0.010285	1st Qu.:0.1521	1st Qu.:0.06428	1st Qu.:2.028	1st Qu.: 764
Median :0.025005	Median :0.2517	Median :0.09389	Median :2.509	Median :1858
Mean :0.028580	Mean :0.2449	Mean :0.10343	Mean :2.470	Mean :2123
3rd Qu.:0.027152	3rd Qu.:0.2823	3rd Qu.:0.12483	3rd Qu.:2.856	3rd Qu.:2017
Max. :0.102591	Max. :0.4631	Max. :0.25259	Max. :3.304	Max. :7621

Medidas resumo acerca do suporte, confiança, *lift* e *coverage*

Em uma regra do tipo $A \rightarrow B$, o *coverage* é a medida de suporte restrita apenas ao item A

mining info:

data	ntransactions	support	confidence
transacoes	74285	0.001	0.01

Arquivo: Pet_Shop.txt



Case: Pet Shop

4. ALGORITMO APRIORI | ANÁLISE DE CESTA DE COMPRAS

37

Algoritmo apriori

set of 20 rules

rule length distribution (lhs + rhs):sizes
2
20

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2	2	2	2	2	2

summary of quality measures:

support	confidence	coverage	lift	count
Min. :0.005937	Min. :0.1052	Min. :0.04251	Min. :1.833	Min. : 441
1st Qu.:0.010285	1st Qu.:0.1521	1st Qu.:0.06428	1st Qu.:2.028	1st Qu.: 764
Median :0.025005	Median :0.2517	Median :0.09389	Median :2.509	Median :1858
Mean :0.028580	Mean :0.2449	Mean :0.10343	Mean :2.470	Mean :2123
3rd Qu.:0.027152	3rd Qu.:0.2823	3rd Qu.:0.12483	3rd Qu.:2.856	3rd Qu.:2017
Max. :0.102591	Max. :0.4631	Max. :0.25259	Max. :3.304	Max. :7621

mining info:

data	ntransactions	support	confidence
transacoes	74285	0.001	0.01

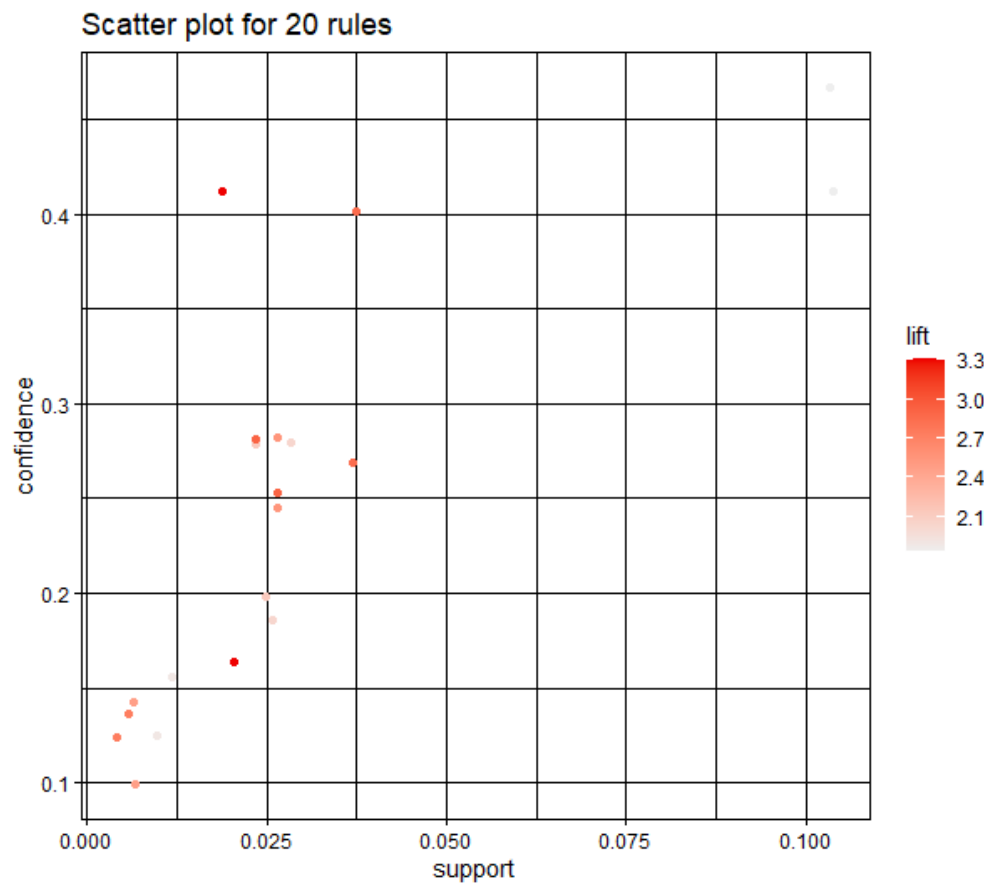
Informações gerais da execução do algoritmo

Arquivo: Pet_Shop.txt

@LABDATA FIA. Copyright all rights reserved.



Algoritmo apriori



Representação visual das regras identificadas, com base em suporte, confiança e *lift*

Arquivo: Pet_Shop.txt

@LABDATA FIA. Copyright all rights reserved.



Case: Pet Shop

4. ALGORITMO APRIORI | ANÁLISE DE CESTA DE COMPRAS

39

Algoritmo apriori

	lhs		rhs	support	confidence	coverage	lift	count
[1]	{racaio_umida}	=>	{petiscos}	0.102591371	0.4631138	0.22152521	1.833426	7621
[2]	{desinfetantes_ambientes}	=>	{tapetes_higienicos}	0.019667497	0.4093584	0.04804469	3.303910	1461
[3]	{petiscos}	=>	{racaio_umida}	0.102591371	0.4061501	0.25259474	1.833426	7621
[4]	{condicionadores}	=>	{shampoos}	0.037140742	0.4013091	0.09254897	2.855757	2759
[5]	{higiene_bucal}	=>	{shampoos}	0.026923336	0.2850221	0.09446052	2.028247	2000
[6]	{bebedouros}	=>	{comedouros}	0.024688699	0.2814178	0.08772969	2.914012	1834
[7]	{cobertores}	=>	{camas}	0.027152184	0.2785912	0.09746248	2.543026	2017
[8]	{guias}	=>	{brinquedos}	0.025321397	0.2713503	0.09331628	2.126068	1881
[9]	{shampoos}	=>	{condicionadores}	0.037140742	0.2642973	0.14052635	2.855757	2759
[10]	{comedouros}	=>	{bebedouros}	0.024688699	0.2556454	0.09657401	2.914012	1834
[11]	{camas}	=>	{cobertores}	0.027152184	0.2478496	0.10955105	2.543026	2017
[12]	{brinquedos}	=>	{guias}	0.025321397	0.1983968	0.12763007	2.126068	1881
[13]	{shampoos}	=>	{higiene_bucal}	0.026923336	0.1915892	0.14052635	2.028247	2000
[14]	{tapetes_higienicos}	=>	{desinfetantes_ambientes}	0.019667497	0.1587353	0.12390119	3.303910	1461
[15]	{peitorais}	=>	{coleiras}	0.010284714	0.1549381	0.06637948	1.893334	764
[16]	{cortadores_unha}	=>	{escovas_pentes}	0.006098136	0.1434452	0.04251195	2.474647	453
[17]	{fraldas}	=>	{caixas_transporte}	0.005936596	0.1384181	0.04288887	2.731046	441
[18]	{coleiras}	=>	{peitorais}	0.010284714	0.1256786	0.08183348	1.893334	764
[19]	{caixas_transporte}	=>	{fraldas}	0.005936596	0.1171315	0.05068318	2.731046	441
[20]	{escovas_pentes}	=>	{cortadores_unha}	0.006098136	0.1052020	0.05796594	2.474647	453

Arquivo: Pet_Shop.txt

@LABDATA FIA. Copyright all rights reserved.



lab.data



Algoritmo apriori

	lhs		rhs	support	confidence	coverage	lift	count
[1]	{racaio_umida}	=>	{petiscos}	0.102591371	0.4631138	0.22152521	1.833426	7621
[2]	{desinfetantes_ambientes}	=>	{tapetes_higienicos}	0.019667497	0.4093584	0.04804469	3.303910	1461
[3]	{petiscos}	=>	{racaio_umida}	0.102591371	0.4061501	0.25259474	1.833426	7621
[4]	{condicionadores}	=>	{shampoos}	0.037140742	0.4013091	0.09254897	2.855757	2759
[5]	{higiene_bucal}	=>	{shampoos}	0.026923336	0.2850221	0.09446052	2.028247	2000
[6]	{bebedouros}	=>	{comedouros}	0.024688699	0.2814178	0.08772969	2.914012	1834
[7]	{cobertores}	=>	{camas}	0.027152184	0.2785912	0.09746248	2.543026	2017
[8]	{guias}	=>	{brinquedos}	0.025321397	0.2713503	0.09331628	2.126068	1881
[9]	{shampoos}	=>	{condicionadores}	0.037140742	0.2642973	0.14052635	2.855757	2759
[10]	{comedouros}	=>	{bebedouros}	0.024688699	0.2556454	0.09657401	2.914012	1834
[11]	{camas}	=>	{cobertores}	0.027152184	0.2478496	0.10955105	2.543026	2017
[12]	{brinquedos}	=>	{guias}	0.025321397	0.1983968	0.12763007	2.126068	1881
[13]	{shampoos}	=>	{higiene_bucal}	0.026923336	0.1915892	0.14052635	2.028247	2000
[14]	{tapetes_higienicos}	=>	{desinfetantes_ambientes}	0.019667497	0.1587353	0.12390119	3.303910	1461
[15]	{peitorais}	=>	{coleiras}	0.010284714	0.1549381	0.06637948	1.893334	764
[16]	{cortadores_unha}	=>	{escovas_pentes}	0.006098136	0.1434452	0.04251195	2.474647	453
[17]	{fraldas}	=>	{caixas_transporte}	0.005936596	0.1384181	0.04288887	2.731046	441
[18]	{coleiras}	=>	{peitorais}	0.010284714	0.1256786	0.08183348	1.893334	764
[19]	{caixas_transporte}	=>	{fraldas}	0.005936596	0.1171315	0.05068318	2.731046	441
[20]	{escovas_pentes}	=>	{cortadores_unha}	0.006098136	0.1052020	0.05796594	2.474647	453

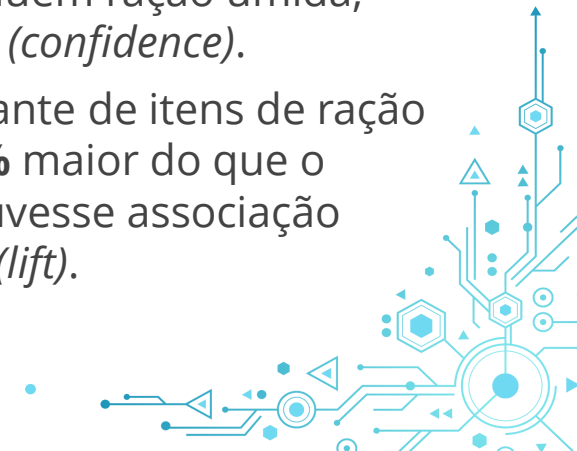
Ordenando as regras da maior para a menor **confiança**, a regra mais forte é:

ração úmida → petiscos

Ou seja, quem compra ração úmida tende a comprar petiscos na mesma ocasião. Note que:

- ~**10%** das cestas incluem itens de ração úmida e petiscos (*suporte*).
- ~**22%** das cestas incluem itens de ração úmida (*coverage*).
- Entre as cestas que incluem ração úmida, ~**46%** incluem petiscos (*confidence*).
- A frequência concomitante de itens de ração úmida e petiscos é **83%** maior do que o esperado caso não houvesse associação entre essas categorias (*lift*).

Arquivo: Pet_Shop.txt



Algoritmo apriori

	lhs		rhs	support	confidence	coverage	lift	count
[1]	{racaio_umida}	=>	{petiscos}	0.102591371	0.4631138	0.22152521	1.833426	7621
[2]	{desinfetantes_ambientes}	=>	{tapetes_higienicos}	0.019667497	0.4093584	0.04804469	3.303910	1461
[3]	{petiscos}	=>	{racaio_umida}	0.102591371	0.4061501	0.25259474	1.833426	7621
[4]	{condicionadores}	=>	{shampoos}	0.037140742	0.4013091	0.09254897	2.855757	2759
[5]	{higiene_bucal}	=>	{shampoos}	0.026923336	0.2850221	0.09446052	2.028247	2000
[6]	{bebedouros}	=>	{comedouros}	0.024688699	0.2814178	0.08772969	2.914012	1834
[7]	{cobertores}	=>	{camas}	0.027152184	0.2785912	0.09746248	2.543026	2017
[8]	{guias}	=>	{brinquedos}	0.025321397	0.2713503	0.09331628	2.126068	1881
[9]	{shampoos}	=>	{condicionadores}	0.037140742	0.2642973	0.14052635	2.855757	2759
[10]	{comedouros}	=>	{bebedouros}	0.024688699	0.2556454	0.09657401	2.914012	1834
[11]	{camas}	=>	{cobertores}	0.027152184	0.2478496	0.10955105	2.543026	2017
[12]	{brinquedos}	=>	{guias}	0.025321397	0.1983968	0.12763007	2.126068	1881
[13]	{shampoos}	=>	{higiene_bucal}	0.026923336	0.1915892	0.14052635	2.028247	2000
[14]	{tapetes_higienicos}	=>	{desinfetantes_ambientes}	0.019667497	0.1587353	0.12390119	3.303910	1461
[15]	{peitorais}	=>	{coleiras}	0.010284714	0.1549381	0.06637948	1.893334	764
[16]	{cortadores_unha}	=>	{escovas_pentes}	0.006098136	0.1434452	0.04251195	2.474647	453
[17]	{fraldas}	=>	{caixas_transporte}	0.005936596	0.1384181	0.04288887	2.731046	441
[18]	{coleiras}	=>	{peitorais}	0.010284714	0.1256786	0.08183348	1.893334	764
[19]	{caixas_transporte}	=>	{fraldas}	0.005936596	0.1171315	0.05068318	2.731046	441
[20]	{escovas_pentes}	=>	{cortadores_unha}	0.006098136	0.1052020	0.05796594	2.474647	453

A segunda regra de maior confiança identificada é também a de maior *lift*:

desinfetantes → tapetes higiênicos

Ou seja, quem compra desinfetantes tende a comprar tapetes higiênicos na mesma ocasião.

- ~**2%** das cestas incluem itens de desinfetantes e tapetes higiênicos (*suporte*).
- ~**5%** das cestas incluem itens de desinfetantes (*coverage*).
- Entre as cestas que incluem desinfetantes, ~**41%** incluem tapetes higiênicos (*confidence*).
- A frequência concomitante de itens de desinfetantes e tapetes higiênicos é **230%** maior do que o esperado caso não houvesse associação entre essas categorias (*lift*).

Arquivo: Pet_Shop.txt

Algoritmo apriori

	lhs		rhs	support	confidence	coverage	lift	count
[1]	{raca_umida}	=>	{petiscos}	0.102591371	0.4631138	0.22152521	1.833426	7621
[2]	{desinfetantes_ambientes}	=>	{tapetes_higienicos}	0.019667497	0.4093584	0.04804469	3.303910	1461
[3]	{petiscos}	=>	{raca_umida}	0.102591371	0.4061501	0.25259474	1.833426	7621
[4]	{condicionadores}	=>	{shampoos}	0.037140742	0.4013091	0.09254897	2.855757	2759
[5]	{higiene_bucal}	=>	{shampoos}	0.026923336	0.2850221	0.09446052	2.028247	2000
[6]	{bebedouros}	=>	{comedouros}	0.024688699	0.2814178	0.08772969	2.914012	1834
[7]	{fraldas}	=>	{caixas_transporte}	0.005936596	0.1171315	0.05068318	2.731046	441
[8]	{coleiras}	=>	{peitorais}	0.010284714	0.1256786	0.0610333	1.764647	764
[9]	{caixas_transporte}	=>	{fraldas}	0.005936596	0.1171315	0.05068318	2.731046	441
[10]	{escovas_pentes}	=>	{cortadores_unha}	0.006098136	0.1052020	0.05796594	2.474647	453

Note que é mais propício de se obter valores altos de *lift* para combinações de baixo suporte na base de dados.

Exemplo: é mais fácil que uma frequência relativa dobre de 1% para 2% do que de 30% para 60%.

Por esse motivo, recomenda-se selecionar as regras de **maior confiança**, tal como fizemos, e não necessariamente as de maior *lift*, cuja relevância pode ser pequena, na prática.

A segunda regra de maior confiança identificada é também a de maior *lift*:

desinfetantes → tapetes higiênicos

Ou seja, quem compra desinfetantes tende a comprar tapetes higiênicos na mesma ocasião.

- ~**2%** das cestas incluem itens de desinfetantes e tapetes higiênicos (*suporte*).
- ~**5%** das cestas incluem itens de desinfetantes (*coverage*).
- Entre as cestas que incluem desinfetantes, ~**41%** incluem tapetes higiênicos (*confidence*).
- A frequência concomitante de itens de desinfetantes e tapetes higiênicos é **230%** maior do que o esperado caso não houvesse associação entre essas categorias (*lift*).

Arquivo: Pet_Shop.txt

Algoritmo apriori

	lhs		rhs	support	confidence	coverage	lift	count
[1]	{racaos_umida}	=>	{petiscos}	0.102591371	0.4631138	0.22152521	1.833426	7621
[2]	{desinfetantes_ambientes}	=>	{tapetes_higienicos}	0.019667497	0.4093584	0.04804469	3.303910	1461
[3]	{petiscos}	=>	{racaos_umida}	0.102591371	0.4061501	0.25259474	1.833426	7621
[4]	{condicionadores}	=>	{shampoos}	0.037140742	0.4013091	0.09254897	2.855757	2759
[5]	{higiene_bucal}	=>	{shampoos}	0.026923336	0.2850221	0.09446052	2.028247	2000
[6]	{bebedouros}	=>	{comedouros}	0.024688699	0.2814178	0.08772969	2.914012	1834
[7]	{cobertores}	=>	{camas}	0.027152184	0.2785912	0.09746248	2.543026	2017
[8]	{guias}	=>	{brinquedos}	0.025321397	0.2713503	0.09331628	2.126068	1881
[9]	{shampoos}	=>	{condicionadores}	0.037140742	0.2642973	0.14052635	2.855757	2759
[10]	{comedouros}	=>	{bebedouros}	0.024688699	0.2556454	0.09657401	2.914012	1834
[11]	{camas}	=>	{cobertores}	0.027152184	0.2478496	0.10955105	2.543026	2017
[12]	{brinquedos}	=>	{guias}	0.025321397	0.1983968	0.12763007	2.126068	1881
[13]	{shampoos}	=>	{higiene_bucal}	0.026923336	0.1915892	0.14052635	2.028247	2000
[14]	{tapetes_higienicos}	=>	{desinfetantes_ambientes}	0.019667497	0.1587353	0.12390119	3.303910	1461
[15]	{peitorais}	=>	{coleiras}	0.010284714	0.1549381	0.06637948	1.893334	764
[16]	{cortadores_unha}	=>	{escovas_pentes}	0.006098136	0.1434452	0.04251195	2.474647	453
[17]	{fraldas}	=>	{caixas_transporte}	0.005936596	0.1384181	0.04288887	2.731046	441
[18]	{coleiras}	=>	{peitorais}	0.010284714	0.1256786	0.08183348	1.893334	764
[19]	{caixas_transporte}	=>	{fraldas}	0.005936596	0.1171315	0.05068318	2.731046	441
[20]	{escovas_pentes}	=>	{cortadores_unha}	0.006098136	0.1052020	0.05796594	2.474647	453

Temos também uma regra de caminho inverso, que aponta que:

tapetes higiênicos → desinfetantes

Ou seja, quem compra tapetes higiênicos tende a comprar tapetes higiênicos na mesma ocasião.

- **~2%** das cestas incluem itens de desinfetantes e tapetes higiênicos (*suporte*).
- **~12%** das cestas incluem itens de tapetes higiênicos (*coverage*).
- Entre as cestas que incluem tapetes higiênicos, **~16%** incluem desinfetantes (*confidence*).
- A frequência concomitante de itens de desinfetantes e tapetes higiênicos é **230%** maior do que o esperado caso não houvesse associação entre essas categorias (*lift*).

Arquivo: Pet_Shop.txt

Algoritmo apriori

O suporte e o *lift* associados a esta regra não sofreram alterações em relação à regra anterior, por conta de sua fórmula de cálculo, mas a **confiança** diminuiu bastante.

Ou seja, entre quem compra desinfetantes, muitos compram tapetes higiênicos, mas não há uma associação tão forte no sentido oposto.

Em resumo, ter alta confiança na regra $A \rightarrow B$ **não implica** alta confiança na regra $B \rightarrow A$.

	lhs		rhs	support	confidence	coverage	lift	count
								7621
								1461
								7621
								2759
								2000
								1834
								2017
								1881
								2759
								1834
								2017
[12]	{brinquedos}	=>	{guias}	0.025321397	0.1983566	0.0000000	0.0000000	1881
[13]	{shampoos}	=>	{higiene_bucal}	0.026923336	0.1915892	0.14052635	2.028247	1834
[14]	{tapetes_higienicos}	=>	{desinfetantes_ambientes}	0.019667497	0.1587353	0.12390119	3.303910	1461
[15]	{peitorais}	=>	{coleiras}	0.010284714	0.1549381	0.06637948	1.893334	764
[16]	{cortadores_unha}	=>	{escovas_pentes}	0.006098136	0.1434452	0.04251195	2.474647	453
[17]	{fraldas}	=>	{caixas_transporte}	0.005936596	0.1384181	0.04288887	2.731046	441
[18]	{coleiras}	=>	{peitorais}	0.010284714	0.1256786	0.08183348	1.893334	764
[19]	{caixas_transporte}	=>	{fraldas}	0.005936596	0.1171315	0.05068318	2.731046	441
[20]	{escovas_pentes}	=>	{cortadores_unha}	0.006098136	0.1052020	0.05796594	2.474647	453

Temos também uma regra de caminho inverso, que aponta que:

tapetes higiênicos → desinfetantes

Ou seja, quem compra tapetes higiênicos tende a comprar tapetes higiênicos na mesma ocasião.

- ~**2%** das cestas incluem itens de desinfetantes e tapetes higiênicos (*suporte*).
- ~**12%** das cestas incluem itens de tapetes higiênicos (*coverage*).
- Entre as cestas que incluem tapetes higiênicos, ~**16%** incluem desinfetantes (*confidence*).
- A frequência concomitante de itens de desinfetantes e tapetes higiênicos é **230%** maior do que o esperado caso não houvesse associação entre essas categorias (*lift*).

Arquivo: Pet_Shop.txt

Case: Patologias

4. ALGORITMO APRIORI | ANÁLISE DE CESTA DE COMPRAS

45

Um grupo de pesquisa em doenças crônicas gostaria de compreender as múltiplas associações entre 10 diferentes patologias de natureza metabólica e cardiovascular. Para isso, organizaram uma base de dados a partir de prontuários recentes de 1.010 pacientes que eram portadores de ao menos duas das patologias investigadas.



Patologias

- Hipertensão
- Diabetes tipo 2
- Doença renal crônica
- Insuficiência cardíaca
- Apneia do sono
- Obesidade
- Doença arterial coronariana
- Dislipidemia
- Síndrome metabólica
- Esteatose hepática

Arquivo: Patologias.txt

@LABDATA FIA. Copyright all rights reserved.



Case: Patologias

4. ALGORITMO APRIORI | ANÁLISE DE CESTA DE COMPRAS

46

Um grupo de pesquisa em doenças crônicas gostaria de compreender as múltiplas associações entre 10 diferentes patologias de natureza metabólica e cardiovascular. Para isso, organizaram uma base de dados a partir de prontuários recentes de 1.010 pacientes que eram portadores de ao menos duas das patologias investigadas.



Tarefa: Identifique as principais regras de associação entre as patologias dos pacientes, considerando suporte mínimo de 5% e confiança mínima de 70%.

Arquivo: Patologias.txt

@LABDATA FIA. Copyright all rights reserved.



lab.data



Referências Bibliográficas

ANÁLISE DE CESTA DE COMPRAS

47

- Dietrich, D., Heller B., Yang B. *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. Wiley, 2015.
- Blattberg R. C., Kim B., Neslin, S. A. *Database Marketing: Analyzing and Managing Customers*. Springer, 2008.





lab.data

<http://labdata.fia.com.br>
Instagram: @labdatafia
Facebook: @LabdataFIA

