

**MBA  
USP  
ESALQ**

# **Associação de dados aplicados à área de varejo online**

Rodrigo Garcia Zaroni  
Nuno Manoel Martins Dias Fouto



## Resumo

Este estudo de caso aplicou a metodologia CRISP-DM em um contexto de ecommerce, com foco na implementação do **algoritmo de associação de dados “Apriori”**, objetivando aplicar a metodologia e o **algoritmo para otimizar estratégias de “cross-selling” no “e-commerce”**.

A pesquisa foi realizada com base nas etapas do **CRISP-DM, desde o entendimento do negócio até a disponibilização do modelo, utilizando dados de um site de revenda de produtos militares**. O artigo detalha os passos desde as etapas de preparação e modelagem com o **algoritmo “Apriori”**, incluindo o desenvolvimento e disponibilização de uma **API em “Python Flask”**, para a integração à plataforma de e-commerce.

Os resultados obtidos demonstraram a eficácia da aplicação da metodologia CRISP-DM, bem como do algoritmo Apriori, na **identificação de regras de sugestão entre os produtos oferecidos**, permitindo sua utilização para as estratégias de marketing no setor de varejo online.



“O varejo restrito (bens de consumo, exceto automóveis e materiais de construção) fechou 2022 com uma expansão nominal de 7,7%, movimentando **R\$ 2,14 trilhões** e respondendo por **21,4% do PIB brasileiro**. Porém o ticket médio do “e-commerce” recuou 7,5% em relação a 2021, enquanto o número de **compradores subiu 24%** e a quantidade de pedidos avançou 7,9%.”

Terra, E.; Besnosoff, F.; Muller, R. 2023. Estudo O Papel do Varejo na Economia Brasileira 2023 – SBVC. Disponível em: <https://sbvc.com.br/10aed-estudo-o-papel-do-varejo-naeconomia-brasileira-2023-sbvc/>. Acesso em: 22 out. 2023.

## Desafio (Motivação)

Aumentar a quantidade de  
itens vendidos e/ou a  
migração para produtos  
com maior preço.

O “cross-selling” denominado  
combo, consiste em oferecer  
aos clientes, produtos ou  
serviços adicionais, com base  
em itens já comprados, ou  
que tiveram interesse prévio  
pelos clientes (Fadillah et al., 2021).

Desafios nesse contexto  
são frequentes para os  
gerentes de projetos de  
dados em diversos setores  
do varejo.

Fadillah, A.R.; Nurma Yulita, I.; Pradana, A.; Suryani, M. 2021. Data Mining Implementation Using Frequent Pattern Growth on Transaction Data for Determining Cross-selling and Upselling (Case Study: Cascara Coffee). Em: 2021 International Conference on Artificial Intelligence and Big Data Analytics, ICAIBDA 2021. Institute of Electrical and Electronics Engineers Inc., páginas 272–277



## Material e Métodos

A metodologia científica adotada neste estudo foi o estudo de caso, Yin (2001).  
**Analizou-se um website de varejo com vendas diretas ao consumidor.**

**CRISP-DM**, Schröer et al.(2021)

- I) Entendimento do Negócio
- II) Entendimento dos Dados
- III) Preparação dos Dados
- IV) Modelagem
- V) Avaliação
- VI) Implantação do modelo

Yin, Robert K. Estudo de caso: planejamento e métodos/Robert K. Yin; trad. Daniel Grassi – 2.ed. – Porto Alegre: Bookman, 2001

Schröer, C.; Kruse, F.; Gómez, J.M. 2021. A systematic literature review on applying CRISPDM process model. Em: Procedia Computer Science. Elsevier B.V., volume 181, páginas 526–534.



# Fases I e II - Entendimento : Negócio e Dados



**Produtos mais vendidos**  
**(22161200-green e 22161200-black)**  
**Relógio Militar Paracord de Sobrevivência À Prova D'água**

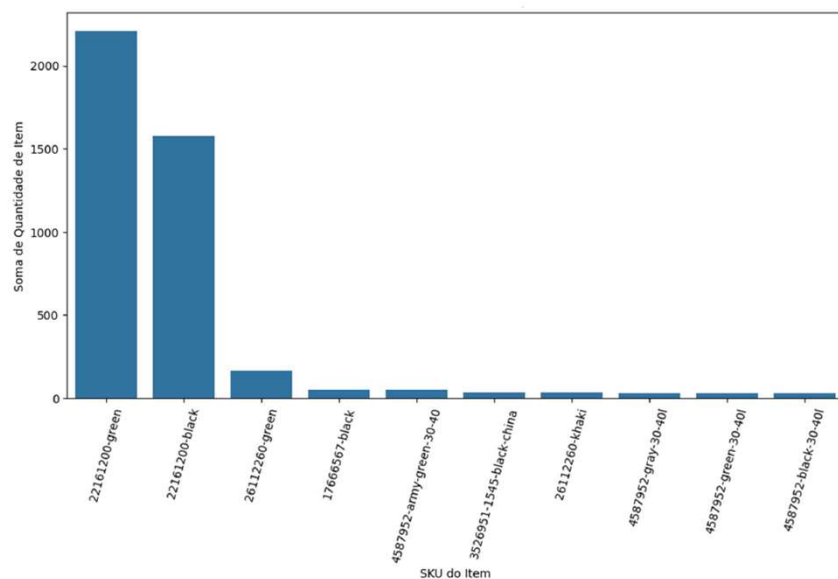


Figura 1. Principais produtos (SKUs) vendidos (TOP 10)  
Fonte: Dados originais da pesquisa<sup>9</sup>

**Ticket Médio R\$ 141,76.**

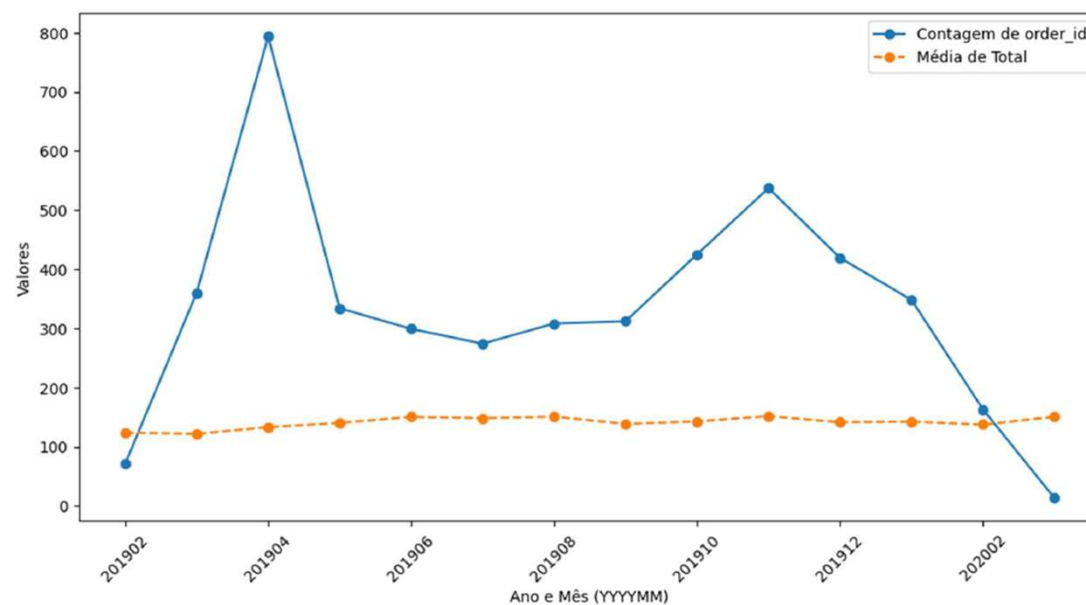


Figura 2. Série temporal da quantidade de pedidos (order\_id) e média do ticket (total) disponível no dataset  
Fonte: Dados originais da pesquisa<sup>10</sup>

## Fase III – Preparação dos Dados

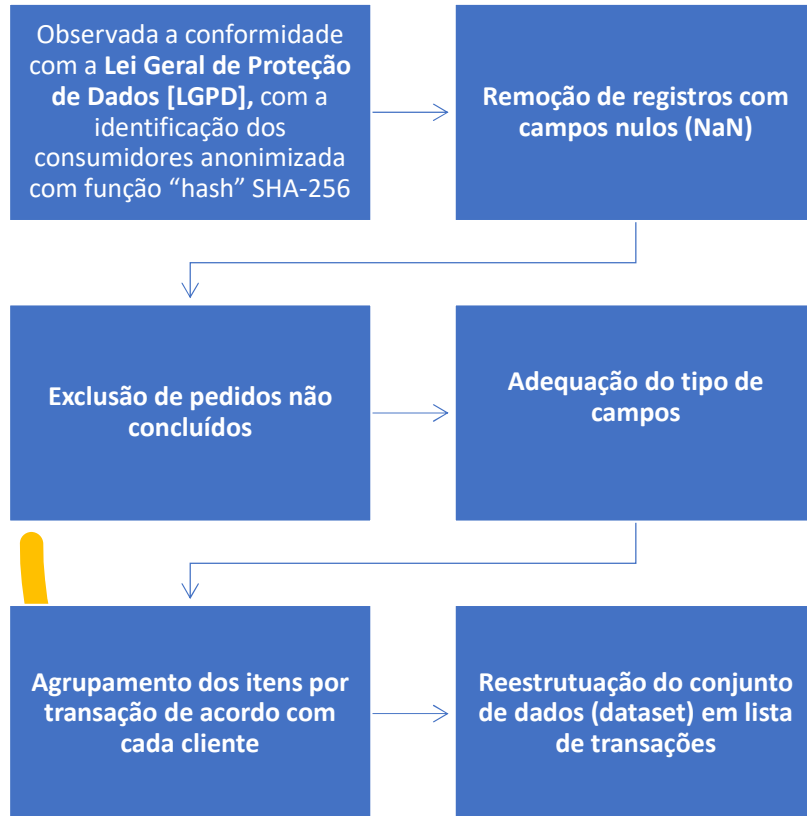


Tabela 2. Campos do dataset selecionados para a realização da análise, com dados exemplo

Email	Order_id	Item_sku	Item_name	Item_quantity	Item_total	Subtotal	Total	Order_status	Created_at
0f73...	5498	22161200-black	Relógio Militar	1	127,90	127,90	127,90	Completed	07/03/2020
3cd1...	5497	22161200-green	Relógio Militar	1	127,90	127,90	127,90	Completed	06/03/2020
Efte...	5495	22161200-green	Relógio Militar	1	127,90	127,90	127,90	Pending	05/03/2020
7a03...	5494	22161200-black	Relógio Militar	1	127,90	127,90	127,90	Completed	04/03/2020

Fonte: Dados originais da pesquisa<sup>12</sup>

# Fase IV– Modelagem

- Técnica de modelagem utilizada

O “Apriori” explora a propriedade de que todos os subconjuntos não vazios de um conjunto de itens frequentes também são frequentes. Essa característica é essencial para reduzir eficientemente o espaço de busca e melhorar o desempenho do algoritmo em grandes bases de dados (Agrawal et al., 1993).

Agrawal, R.; Imielinski, T.; Swami, A. 1993. Mining associations between sets of items in large databases. In Proc. of the ACM SIGMOD Int'l Conference on Management of Data, páginas 207-216, Washington D.C., maio de 1993.

## Mining Association Rules between Sets of Items in Large Databases

Rakesh Agrawal    Tomasz Imielinski\*    Arun Swami

IBM Almaden Research Center  
650 Harry Road, San Jose, CA 95120

### Abstract

We are given a large database of customer transactions. Each transaction consists of items purchased by a customer in a visit. We present an efficient algorithm that generates all significant association rules between items in the database. The algorithm incorporates buffer management and novel estimation and pruning techniques. We also present results of applying this algorithm to sales data obtained from a large retailing company, which shows the effectiveness of the algorithm.

### 1 Introduction

Consider a supermarket with a large collection of items. Typical business decisions that the management of the supermarket has to make include what to put on sale, how to design coupons, how to place merchandise on shelves in order to maximize the profit, etc. Analysis of past transaction data is a commonly used approach in order to improve the quality of such decisions. Until recently, however, only global data about the cumulative sales during some time period (a day, a week, a month, etc.) was available on the computer. Progress in bar-code technology has made it possible to store the so called *basket* data that stores items purchased on a per-transaction basis. Basket data type transactions do not necessarily consist of items bought together at the same point of time. It may consist of items bought by a customer over a period of time. Examples include monthly purchases by members of a book club or a music club.

Several organizations have collected massive amounts of such data. These data sets are usually stored on tertiary storage and are very slowly migrating to database systems. One of the main reasons for the limited success of database systems in this area is that current database systems do not provide necessary functionality for a user interested in taking advantage of this information.

This paper introduces the problem of “mining” a large collection of basket data type transactions for association rules between sets of items with some minimum specified confidence, and presents an efficient algorithm for this purpose. An example of such an association rule is the statement that 90% of transactions that purchase bread and butter also purchase milk. The antecedent of this rule consists of bread and butter and the consequent consists of milk alone. The number 90% is the confidence factor of the rule.

The work reported in this paper could be viewed as a step towards enhancing databases with functionalities to process queries such as (we have omitted the confidence factor specification):

- Find all rules that have “Diet Coke” as consequent. These rules may help plan what the store should do to boost the sale of Diet Coke.
- Find all rules that have “bagels” in the antecedent. These rules may help determine what products may be impacted if the store discontinues selling bagels.



# Métricas Apriori

**Suporte:** frequência de transações contém o item X. O "mínimo de suporte" é o valor mínimo definido pelo analista para determinar se um item é considerado frequente.

**Confiança:** probabilidade de encontrar o item Y em transações que contêm X.

**Lift:** frequência do item Y em transações que contêm o item X, com a frequência de ocorrência de Y em todas as transações.

"lift" maior que 1  $\rightarrow$  Y tem maior probabilidade de ocorrer nas transações que contêm X, valores inferiores a 1 demonstram uma não associação entre X e Y.

$$\text{Suporte}(X) = \frac{\text{freq}(X)}{N} \quad (1)$$

onde,  $\text{Suporte}(X)$ : é a frequência (freq) deste item X dividido pela quantidade de transações N do conjunto de dados utilizado (dataset).

$$\text{Confiança}(X \rightarrow Y) = \frac{\text{Suporte}(X \cap Y)}{\text{Suporte}(X)} \quad (2)$$

onde,  $\text{Suporte}(X \cap Y)$ : é o suporte conjunto de X e Y, que é a proporção de transações no conjunto de dados que contêm tanto X quanto Y.

$\text{Suporte}(X)$ : é o suporte de X, que é a proporção de transações no conjunto de dados que contêm X.

$$\text{Lift}(X \rightarrow Y) = \frac{\text{Confiança}(X \rightarrow Y)}{\text{Suporte}(Y)} \quad (3)$$

onde,  $\text{Confiança}(X \rightarrow Y)$ : é a confiança da regra, como discutido anteriormente.

$\text{Suporte}(Y)$ : é o suporte do item Y no conjunto de dados.

## Fase V – Avaliação

O protótipo, criado com a regra de associação “apriori”, sustentou a hipótese do “cross-sell”,

Tabela 3. Regras e medidas encontradas após a aplicação do algoritmo “Apriori”

Rule	Support	Confidence	Lift
13199838-black -> 3526951-1545-black-silver-china	0,033%	0,13	127,21
26112260-green -> 16058326-army-camouflage	0,033%	1,00	29,08
26256057-green -> 17666567-black	0,033%	0,13	9,79
4587952-army-green-30-40 -> 18379073-army-green-l	0,033%	1,00	101,77
18379073-black-l -> 4587952-40l-army-green-30-40l-china	0,033%	0,33	339,22
4587952-black-30-40l -> 18379073-black-m	0,033%	0,17	26,78
4587952-gray-30-40l -> 18379073-black-m	0,033%	0,17	28,27
19040980-sunglasses -> preto	0,033%	0,33	78,28
19768167-od -> 19768167-tan	0,066%	0,20	87,23
4587952-black-30-40l -> 19768167-white	0,033%	0,25	40,17
verde-oliva -> 22471597-01-3l	0,033%	0,50	101,77
4587952-green-30-40l -> 22491228-armygreen	0,033%	0,17	28,27
23784889-grey-11 -> 4587952-40l-gray-30-40l-china	0,033%	1,00	1526,50
26112260-green -> 4587952-40l-gray-30-40l-china	0,033%	0,50	14,54
3526951-1545-black-china -> 8055270-army-green	0,033%	0,20	23,48

Fonte: Resultados originais da pesquisa<sup>13</sup>

100% Confiança

Probabilidade de encontrar o item Y em transações que contêm X.

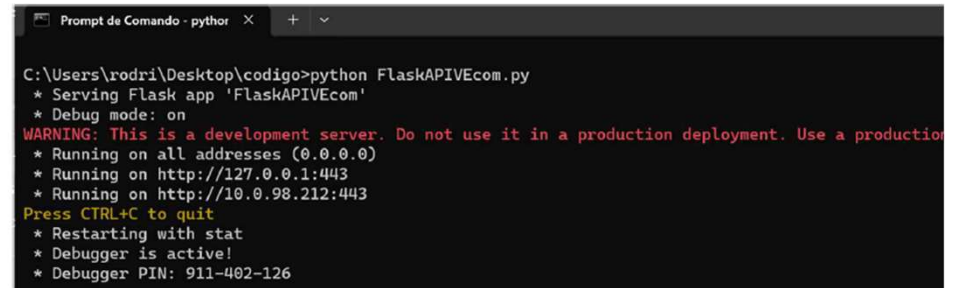
## Fase V – Implantação do Modelo

→ PYTHON

→ FLASK

→ API

→ ROTAS API (MÉTODO GET)  
/Swagger /listaskus /recomendacao



```
Prompt de Comando - pythor X + v

C:\Users\rodri\Desktop\codigo>python FlaskAPIEcom.py
* Serving Flask app 'FlaskAPIEcom'
* Debug mode: on
WARNING: This is a development server. Do not use it in a production deployment. Use a production
* Running on all addresses (0.0.0.0)
* Running on http://127.0.0.1:443
* Running on http://10.0.98.212:443
Press CTRL+C to quit
* Restarting with stat
* Debugger is active!
* Debugger PIN: 911-402-126
```

Figura 4. Execução da API do modelo realizada em ambiente teste local  
Fonte: Resultados originais da pesquisa<sup>14</sup>



```
127.0.0.1:443/recomendacao?sku=23784889-grey-11

1 // 20240129151428
2 // http://127.0.0.1:443/recomendacao?sku=23784889-grey-11
3
4 [
5   "4587952-401-grey-30-401-china"
6 ]
```

Figura 6. Resposta da API consulta GET em 127.0.0.1:443/recomendacao?sku=23784889-grey-11  
Fonte: Resultados originais da pesquisa<sup>16</sup>

## Resultados e Discussão

- **Estudo limitado ao tamanho do Dataset utilizado**

➔ Métricas encontradas:

- suporte baixo, indicando uma ocorrência infrequente das combinações no conjunto total de transações.
- confiança, Alguns produtos apresentaram uma confiança abaixo de 1.0, sinalizando uma associação moderada entre os itens.
- “lift”, valores superiores a 1, destacando-se a relação:

23784889-grey-11 e 4587952-40l-gray-30-40l-china, que apresentou um “lift” de 1526.5

Figura 7. Site do E-commerce com os produtos disponíveis incluindo o Tênis Tático Militar de SKU: 23784889-grey-11

Figura 8. Sugestão adicional da Mochila Tática Militar 40L 3 (SKU: 4587952-40l-gray-30-40l-china)

Fonte: Resultados originais da pesquisa



Figura 7.



Figura 8.

## Considerações Finais

- Implementou-se o algoritmo de associação de dados "Apriori".
- As fases da metodologia CRISP-DM foram seguidas: entendimento do negócio, dos dados, preparação, modelagem, avaliação e implantação.
- Após a geração dos resultados das associações, os dados juntamente com os produtos recomendados para "cross-sell" via e-commerce foram disponibilizados por meio de uma API em Flask Python.
- O tamanho reduzido da base de dados impactou nas métricas do algoritmo. Sugere-se para estudos futuros a utilização de conjuntos de dados mais amplos e a aplicação de outros algoritmos como Eclat e FP-Growth, além do treinamento de modelos de redes neurais.
- Conclusão: Demonstração da relevância da metodologia CRISP-DM e do algoritmo de associação "Apriori" para os departamentos de tecnologia e marketing em varejo online, com potencial de aplicação em outras indústrias que possuam volumes de dados maiores.



# Considerações Finais

- Implementou-se o algoritmo de associação de dados "Apriori".
- As fases da metodologia CRISP-DM foram seguidas: entendimento do negócio, dos dados, preparação, modelagem, avaliação e implantação.
- Após a geração dos resultados das associações, os dados juntamente com os produtos recomendados para venda cruzada via e-commerce foram disponibilizados por meio de uma API em Flask Python.
- O tamanho reduzido da base de dados impactou nas métricas do algoritmo. Sugere-se para estudos futuros a utilização de conjuntos de dados mais amplos e a aplicação de outros algoritmos como Eclat e FP-Growth, além do treinamento de modelos de redes neurais.
- Conclusão: Demonstração da relevância da metodologia CRISP-DM e do algoritmo de associação "Apriori" para os departamentos de tecnologia e marketing em varejo online, com potencial de aplicação em outras indústrias que possuam volumes de dados maiores.

- 
- **Referências Bibliográficas e Código-fonte em Python utilizado, consulte no Trabalho de Conclusão de Curso (TCC) disponibilizado.**