



**UNIVERSIDADE FEDERAL DE SANTA MARIA**  
**CURSO DE CIÊNCIA DA COMPUTAÇÃO**

**Trabalho Prático 1**

**Disciplina:** Mineração de Dados

**Docente:** Dr. Joaquim Assunção

Davi de Castro Machado  
João Pedro Righi

**Santa Maria, RS**

## 1. Resumo do Pré-Processamento

- **Dados Originais:** O formato inicial dos dados (JSON) continha os dados de uma padaria com objetos de cada compra, contendo os produtos comprados em cada uma. O arquivo estava com um erro de sintaxe (vírgula faltando) e com erros na formatação de texto devido ao uso de acentos.

```
{
  "compra":999,
  "produtos": [
    "Queijo Mussarela",
    "Caf  Igua\u00e7\u00f3",
    "Refri - Pepsi"
  ]
},
{
  "compra":999,
  "produtos": [
    "Queijo Minas",
    "P\u00e3o Gajeta",
    "Doce Leite"
  ]
},
{
  "compra":999,
  "produtos": [
    "Presunto Seara",
    "P\u00e3o Cabrito",
    "Queijo Mussarela"
  ]
},
{
```

- **Remo  o de Acentos e Convers  o para Min  sculas:** Fizemos uso da fun   o `remover_acentos` para padronizar os nomes dos produtos, convertendo todo texto para texto sem acento e em letra min  scula.
- **Elimina  o de Chaves N  o Relevantes:** Decidimos remover as chaves “compra” visto que estavam repetindo o valor erroneamente e n  o seriam pertinentes para o processo de minera   o e extra   o das regras de associa   o.

- **Transformação para Formato Hot Encoding:** Após carregar os dados JSON das compras, as transações precisaram ser representadas de maneira que pudessem ser processadas pelo algoritmo Apriori. Para isso, aplicamos uma técnica chamada **hot encoding** para transformar as listas de produtos em um **DataFrame binário** onde cada coluna representa um produto, e cada linha indica uma transação (compra) específica.

#### **Estrutura dos Dados Codificados:**

- Cada linha do DataFrame representa uma transação completa, e cada coluna representa um produto único.
- Um valor **True** ou **1** indica que o produto foi comprado naquela transação, enquanto **False** ou **0** indica ausência.
- Este formato permite que o Apriori identifique facilmente as combinações de produtos que ocorrem com frequência, simplificando o cálculo de métricas de suporte e confiança.

- **Filtragem dos Dados para Associação:**
- Algoritmo Apriori
- Esse algoritmo foi utilizado para encontrar subconjuntos frequentes nos dados, ou seja, combinações de produtos que aparecem juntos em transações frequentemente.
- O **suporte** de um itemset é a proporção de transações que contêm esse conjunto de produtos. Definimos um suporte mínimo de **0.04** para filtrar apenas os conjuntos mais recorrentes.

#### **Regras de Associação:**

- Após identificar os conjuntos de produtos frequentes, aplicamos as regras de associação para encontrar relações de "Se X, então Y".
- A **confiança** de uma regra reflete a frequência com que uma transação contendo o antecedente também contém o consequente. Utilizamos uma confiança mínima de **0.4** para identificar regras fortes e confiáveis.
- A métrica de **lift** também foi calculada para indicar a força de cada regra em comparação com a ocorrência aleatória dos produtos.

## 2. Respostas e Justificativas

- Top 5 Regras de Associação:

### Top 5 Regras de Associação

Regra	Associação	Suporte	Confiança	Lift
12	pastel presunto e queijo → refri - pepsi	0.0407	1.0000	7.6875
16	presunto perdigao, queijo mussarela → pao frances	0.0488	0.8571	3.7653
15	pao frances, presunto perdigao → queijo mussarela	0.0488	0.7500	2.3654
13	pao frances, presunto sadia → queijo mussarela	0.0407	0.7143	2.2527
14	presunto sadia, queijo mussarela → pao frances	0.0407	0.6250	2.7455

Aqui vemos a relação forte de produtos como pão, presunto e queijo.

O que está relacionado com a cultura de fazer sanduíches para café da manhã e da tarde.

- Regra Mais Influente (1 para 1):

### Regra Mais Influente 1 para 1

Regra	Associação	Suporte	Confiança	Lift
12	pastel presunto e queijo → refri - pepsi	0.0407	1.0000	7.6875

A relação mais forte foi a de pastel de presunto e queijo com refri Pepsi.

Onde todas as compras que incluíam pastel de presunto e queijo também envolviam refri Pepsi.

- Regras que Implicam na Compra de "Doce":

### Regras que Implicam na Compra de 'Doce'

Regra	Associação	Suporte	Confiança	Lift
10	refri - fanta → doce	0.0813	0.5882	1.5729
3	queijo minas → doce	0.0569	0.5385	1.4398
11	cafe melita → doce	0.0732	0.5294	1.4156
8	pao gajeta → doce	0.0569	0.5000	1.3370
2	pastel frango → doce	0.0650	0.4706	1.2583

Os resultados das associações envolvendo "Doce" apresentam uma maior diversidade devido à nossa escolha de generalizar a categoria "Doce".

Sem essa generalização, seria necessário utilizar um limite de confiança muito baixo (cerca de 0,17 ou menos) para obtermos associações relevantes, o que reduziria a robustez das regras geradas.

Com a generalização, conseguimos identificar associações mais consistentes e com confiança mais elevada, proporcionando insights mais significativos.

A conclusão principal é de que os clientes compram doce em conjunto de salgados e café, provavelmente para consumir após ou ao mesmo tempo que os mesmos.