ANÁLISE DE CESTO DE COMPRAS EM UM MERCADO VAREJISTA DE PEQUENO PORTE
MARKET BASKET ANALYSIS IN A RETAIL STORE

Daniel Lucas da Silveira Souza
Universidade Federal do Paraná, Curitiba

Silvana Pereira Detro
Universidade Federal do Paraná, Curitiba

Resumo: A análise de cesta de compras não é recente, contudo, por ser tratada como não trivial, iterativa e capaz de prover insights de potencial valor para tomada de decisão, esta análise prossegue sendo utilizada para controle de estoques, gestão de promoções e decisão sobre posicionamento de itens em lojas. Neste trabalho, foram apresentadas pesquisas no que tange o assunto mineração de dados, com foco na associação de dados e no método Apriori. A intenção foi a de conceituar métodos que verifiquem a interação entre variáveis através de indicadores como suporte e confiança, neste caso, a interação entre produtos comprados em conjunto. Com a utilização do KDD como processo de descobrimento de conhecimento, pode ser verificado um modelo ao qual analistas podem utilizar para criar eventos, promoções ou posicionar seus produtos em loja. Com aplicações do método, obteve-se que produtos de panificação foram mais vendidos em conjunto com frios e laticínios, assim como, o salame de 100g esteve na mesma lista de compras que limão e cervejas de 260ml e 350ml. Dificuldades técnicas como manipulação e modelagem dos dados foram pontuadas, porém houve consenso com a literatura que a análise de cesto de compras é uma análise de grande valor para tomada de decisão.

Palavras-chave: Associação de dados, Apriori, análise de cesto de compras

Abstract: Market basket analysis is shown by literature as a powerful tool for decision making. Even though it isn't new, its capability to provide insights on stock management, advertising and product placement in stores still brings some attention to analysts. In this research, data mining was presented through literature review as mean for discovering implicit patterns in data. Focus was given to the Apriori method as it can measure the relationship between variables, thorough support and confidence indicators. Knowledge Discovery in Databases (KDD) was also explored as a data assessment process so it could assure that analysts could apply the method presented on their product advertisement and placement strategies . After applying Apriori method, the results showed that Bakery products were mostly sold together with Dairy and Cold Cuts, while clients preferred to purchase Sliced salami 100g with Lemon, 269ml and 350ml Beers. Some technical difficulties, such as data classification and modeling were pointed at the end. Nevertheless, there was consent with authors in literature that Market basket analysis provide value to decision making.

Key-words: Market Basket Analysis, Apriori, data mining, Knowledge Discovery in Database

1 INTRODUÇÃO

Apesar das organizações gerarem e armazenarem grandes quantidades de dados, apenas isto não traz benefícios diretos. O valor real para a tomada de decisões e realizar melhorias se encontra no entendimento da informação obtida através de

análises destes dados (SILVA, 2019). Neste contexto, Carvalho (2018) comenta que a abordagem clássica de análise é lenta, cara e altamente subjetiva. Com isto, o autor sugere utilizar métodos de análise de dados, como KDD e mineração de dados, para reduzir custos, aumentar ganhos e melhorar o desempenho do negócio.

O KDD(*Knowledge Discovery in Databases*) apresentado por Anselmo (2017) como um processo estabelecido de descoberta de informações, iterativo e interativo, permite aos analistas descobrirem padrões intrínsecos aos dados. Em uma das etapas deste processo, na mineração de dados, informações implícitas aos dados podem ser traduzidas para que tomada de decisões sejam feitas. Tais decisões podem ser encontradas em diversas áreas como produção agrícola de grãos (MIRHASHEMI, 2021), gestão de UTIs em hospitais (GUIMARÃES, 2018), entre vários outras.

Nesta pesquisa, será demonstrada a aplicação da Análise de Cesto de Compras (MBA - *Market Basket Analysis*). Esta análise consiste na aplicação de métodos de mineração de dados sobre base de dados transacionais de compras com intenção de descobrir grupos de produtos mais vendidos em conjunto (PRAWIRA, 2020)

E apesar desta problemática de agrupar produtos em cestas de compra mais prováveis não ser recente (ANSELMO, 2017), isto garante a oportunidade para analistas iniciantes focar nos detalhes técnicos relativos à aplicação de métodos de mineração de dados, assim como na conceituação de termos que facilitem a escolha de método para resolução de seus problemas específicos.

Com isto em mente, este trabalho apresentará a definição de mineração de dados, principalmente, no que tange seu motivo de aplicação. Em seguida, serão apresentados resultados de análises de cesto de compras encontrados na literatura. Neste momento serão apresentados os conceitos de suporte, confiança e alavancagem. Conceitos de associação de dados, uma das tarefas de mineração de dados apresentadas inicialmente. Mais adiante, o processo de descobrimento de conhecimento (KDD) será descrito com objetivo de apresentar uma metodologia consistente de abordagem sobre dados reais obtidos a partir de bancos de dados. Por fim, será feita a análise de uma base de dados retirada do ERP de um mercado varejista de pequeno porte através das etapas do KDD como demonstração.

2 REVISÃO DE LITERATURA

A mineração de dados é descrita como um processo de descobrimento de padrões sobre uma base de dados formatada. Estes padrões podem ser utilizados para prever ou descrever contextos (CARVALHO, 2018), e são obtidos através de tarefas, as quais são resumidas no Quadro 1:

Quadro 1 – Tarefas da mineração de dados

Tarefa	Definição
a) Classificação	Descoberta de uma função que mapeie (classifique) um item de dados em um conjunto de classes pré-definidas;
b) Regressão	Descoberta de uma função que mapeie um item de dados em uma variável de predição de valor real;
c) Agrupamento (clusterização)	Identificação de um conjunto finito de categorias (clusters) que descrevam os dados;
d) Sumarização	Busca de uma descrição compacta para um subconjunto de dados;
e) Modelagem de dependência	Busca de um modelo que descreva as dependências significativas entre as variáveis;

Fonte: adaptado de CARVALHO(2018)

O Quadro 1 mostra que cinco dos principais resultados das aplicações de tais ferramentas são classificações, agrupamentos ou modelos.

A classificação é sugerida para segmentação de dados com atributo de classe já definido. Isto é caracterizado como processo supervisionado onde os dados apresentam uma coluna com a classificação existente. Mirhashemi(2021), por exemplo, propõe a previsão de necessidade de água na produção de grãos através de árvores de decisão onde, a partir de medições relativas a chuvas, ventos, luminosidade solar, dentre outros, cenários foram obtidos.

A regressão funciona de forma similar para atributos com formatação numérica. Retas são traçadas e ajustadas no espaço de forma a classificar medições dos dados (WANG, ZHAO, 2020).

No caso do agrupamento, centroides são criados no espaço de forma a relacionar dados com maior similaridade. Geralmente, essa junção é feita através de um cálculo de distância onde a medida do dado é subtraída de determinado valor e

subsequentemente comparada às medidas dos centroides. O número desses centroides corresponde ao número de grupos (BEHESHTIAN-ARDAKANI, 2018).

A sumarização, de certa forma, é o resultado obtido através das ferramentas de mineração de dados. No resultado, poderão ser identificados métricas de inicialização, número de instâncias, índices de assertividade, agrupamentos, entre outros.

A modelagem de dependência é relativamente diferente das classificações e agrupamentos anteriores. Segundo o Quadro 1, esta tarefa é responsável por descrever dependência entre variáveis, em outras palavras, identificar o quão relacionado um atributo está em relação a outro (WANDERLEY, 2020). Na prática, os algoritmos ou métodos desta tarefa realizam contagens em conjuntos criados a partir da base de dados.

A modelagem de dependência verifica a associação de dados. Esta associação é criada através da criação e contagem de regras. A regra, geralmente representada como X -> Y, corresponde a um grupo de itens ser escolhido anteriormente a outro(s) como demonstrado por (WANDERLEY, 2020). Em seu estudo sobre a adoção de um programa de lealdade em supermercados, o autor apresenta a relação entre categorias de produtos de maneira bem resumida, conforme apresentado na Tabela 1.

Tabela 1 – Regras de associação de produtos de um supermercado

Regra	Regras de Associação
1	[PROMOÇÃO, CAFÉ] -> [PADARIA] (confiança: 0.643)
2	[PROMOÇÃO, CARNES, FRANGOS] -> [HORTIFRUTI] (confiança: 0.606)
3	[PROMOÇÃO, CARNES, PÃES DE PACOTE] -> [HORTIFRUTI] (confiança: 0.900)
4	[PROMOÇÃO, CARNES] -> [HORTIFRUTI] (confiança: 0.557)
5	[PROMOÇÃO, FRANGOS] -> [CARNES] (confiança: 0.623)

Fonte: adaptado de Wanderley(2020)

A regra 1 de associação da Tabela 1 apresenta a confiança com que a categoria PADARIA é selecionada sendo que PROMOÇÃO e CAFÉ já tenham sido selecionadas. A confiança de 0,643. A confiança pode ser calculada, segundo Suprianto (2019), através da Equação 1:

(1) Confiança = Transações que contenham X e Y / Transações que contenham X

Esta equação mostra o cálculo da confiança para transações que contenham X e Y sendo que X já tenha sido escolhido. Em outras palavras, a categoria PADARIA existe em 64,3% das ocorrências de PROMOÇÃO e CAFÉ simultaneamente (Tabela 1) .

Outros autores também utilizam o índice suporte para verificar a relevância da regra, ou seja, a probabilidade de uma regra ocorrer. De acordo com Suprianto (2019), o suporte pode ser calculado pela Equação 2:

(2) Suporte = Transações que contenham X e Y / Todas Transações

A Tabela 2 apresenta cinco regras referentes à venda conjunta de produtos com seus respectivos suportes.

Tabela 2 – Regras de associação de uma cafeteria

Regra	Regras de Associação	Quantidade	Suporte(%)
1	Café 500g -> Açúcar 1kg	19	38
2	Café 500g -> Açúcar 500g	18	36
3	Café 500g -> Galão de água	5	10
4	Café 500g -> Arroz barato 5kg	12	24
5	Café 500g -> Arroz premium 5kg	16	32

Fonte: adaptado de Anselmo(2017)

De acordo com as regras apresentadas na Tabela 2, pode-se verificar que os produtos que mais foram vendidos juntos se referem à primeira, à segunda e à quinta regra, as quais mostram que:

a) 19 clientes que compraram Café de 500g também compraram 1 kg açúcar;

b) 18 clientes que compraram Café de 500g também compraram pacote de 500g de açúcar;

c) 16 clientes que compraram Café de 500g também compraram Arroz premium de 5kg.

O suporte mostra que estas regras apresentam probabilidade de 38%, 36% e 32% de ocorrer, respectivamente. Neste caso, as regras identificadas mostram que seria interessante observar o estoque de Açúcar de 1kg, Açúcar de 500g e de Arroz premium de 5kg a partir da venda de Café de 500g.

A Tabela 3 apresenta regras obtidas considerando o suporte e a confiança, os quais devem ser analisadas em conjunto.

Tabela 3 – Padrão de consumo através de regras de associação

Regras de Associação	Suporte	Confiança(%)
Se comprar Arroz frito, irá comprar Cappuccino	7/14	50
Se comprar Arroz frito, irá comprar Sopa de macarrão	4/14	28,57
Se compra Arroz frito, irá comprar Chá doce	6/14	42,85
Se comprar Arroz frito, irá comprar Chocolate	5/14	35,71
Se comprar Arroz frito,irá comprar Chá empurrado	4/14	28,57

Fonte: adaptado de Suprianto (2019)

A Tabela 3 apresenta cinco regras, bem como a confiança e o suporte de cada uma delas. A primeira regra mostra que a venda de Arroz frito está relacionada com a venda de Cappuccino. Esta regra tem uma confiança de 50% e o índice de suporte corresponde a quantas vezes esta regra ocorreu, ou seja, demonstra a dimensão da amostra selecionada.

Outro índice de resultado é *lift* ou alavancagem, o qual é apresentado na Equação 3 (ANSELMO, 2017):

(3) *Lift*(X=>Y) = suporte(X e Y) / (suporte(X) * suporte(Y))

O *lift* é considerado por Anselmo(2017) como um índice de correlação que mostra o quão relevante é uma escolha em relação à outra. Por exemplo, se o valor de *lift* for maior que 1, a escolha de X implica positivamente na escolha de Y; se *lift* for igual a 1, pode-se desconsiderar a regra; e se *lift* menor que 1, logo a escolha de X implica negativamente na escolha de Y (COSTA, 2019).

A regra de associação é usualmente utilizada para analisar o comportamento de vendas, ou seja, quais produtos são vendidos em conjunto. Por isso, normalmente

ela é chamada de Análise de cesta de compra (*Market Basket Analysis* - MBA). Os exemplos apresentados nas Tabelas 1, 2 e 3 são resultados de aplicação do MBA. Por meio deste tipo de análise, é possível identificar a relação entre produtos específicos, categorias de produtos e produtos em promoção, o que permite definir a melhor localização dos produtos nas lojas, facilitando a compra para os clientes, bem como o controle de estoque. Possibilita ainda conseguir vantagens relacionadas à promoção de produtos, à aplicação de design em catálogos, no layout da loja ou em campanhas promocionais (ANSELMO, 2017, SUPRIANTO, 2019 E WANDERLEY, 2020).

3 MATERIAIS E MÉTODOS

Neste trabalho, serão observadas compras coletadas de um banco de dados de um mercado varejista de pequeno porte. Isto caracteriza a natureza da pesquisa como aplicada por “gerar conhecimentos para aplicação prática dirigidos à solução de problemas específicos” (PRODANOV, FREITAS. 2013). Hipóteses serão formuladas com objetivo de descrever e generalizar o comportamento de compra dos clientes. Para isto será utilizado o método de associação de dados da mineração de dados para calcular probabilidades e graus de confiança para regras que associam os produtos de uma compra. O método será classificado como estatístico-indutivo por determinar, em termos numéricos, a probabilidade de acerto de determinada conclusão (PRODANOV, FREITAS. 2013) ao mesmo tempo em que “as constatações particulares levarão à elaboração de generalizações” (SILVA, MENEZES. 2001).

O procedimento adotado será o de pesquisa de campo, visto que trará análise de fatos, seguida da pesquisa bibliográfica do tema Análise de Cesto de Compras (Market Basket Analysis) para conceituação, definição de variáveis relevantes, definição de amostra e técnicas de análise (PRODANOV, FREITAS. 2013). Cabe informar que os artigos utilizados como embasamento foram retirados dos portais ScienceDirect e Periódicos Capes com uso das palavras-chave e data de publicação nos últimos 5 anos.

Na seção de desenvolvimento, será necessária uma metodologia para seleção, exploração e modelagem de uma grande quantidade dados para que padrões desconhecidos sejam encontrados (SILVA, 2019). Isto será feito através de um

processo consistente conhecido como Knowledge Discovery in Database, o KDD. De acordo com Wanderley (2020), o KDD é descrito como um processo com propósito de identificar, validar e criar padrões com potencial de utilidade para análise de dados.

Figura 1 – As etapas do KDD

Fonte: Carvalho (2018)

De acordo com a Figura 1, o KDD é composto por 5 etapas (CARVALHO, 2018)

- Seleção é colocada como etapa de definição de objetivo e seleção de dados. Hermaliani et al(2020), em sua pesquisa, busca entender quais frutas são vendidas em conjunto com maior frequência e grau de confiança. Para isto, os autores selecionam 1 ano de vendas. Dentro deste ano, apenas as vendas que contenham alguma das 3 frutas mais vendidas em cada mês foram selecionadas. Inicialmente, cada mês apresentava 2000 vendas com cerca de 150 frutas diferentes. O motivo desta seleção foi a de reduzir o número de instâncias e atributos e assim facilitar a criação de regras.

  Anselmo(2017) indica ainda que o suporte obtido no resultado pode ser relativamente baixo graças à “grande diversidade e importância repartida dos produtos vendidos pela empresa”.
- Na etapa de Pré-processamento, podem ocorrer remoção de ruído, decisão de estratégias para lidar com dados ausentes, criação de colunas, agrupamento de tabelas;

- Na etapa de Transformação, o dado é transformado para assumir o modelo do método escolhido.

  A associação de dados exige que os dados assumam o formato nominal. Em outras palavras, os valores devem ter valores binários ou categóricos. Mirhashemi (2021), por exemplo, classifica seus atributos (horas de sol, velocidade de vento, humidade do ar) em faixas Baixa, Médio e Alta. Enquanto isto, Prawira (2020) mostra o identificador da compra no índice das linhas e distribui seus produtos nas colunas de sua tabela. A Tabela 4 mostra que os valores assumem apenas Sim (“Y”) e Não (“N”) para produtos contidos ou não na compra de cada linha. Na venda PEMB01, todos produtos foram comprados, menos a garrafa de água de 600ml, enquanto na venda PEMB03, apenas açúcar de 1kg foi comprado (Tabela 4).

Tabela 4 – Modelo tabular com valores nominas

ID Venda	Café 500g	Açucar 1kg	Açucar 500g	Galão de água	Garrafa de água 600ml	Arroz barato 5kg
PEMB01	Y	Y	Y	Y	N	Y
PEMB02	N	Y	N	N	Y	N
PEMB03	N	Y	N	N	N	N
PEMB04	Y	Y	Y	N	N	Y
PEMB05	Y	N	N	Y	N	N

Fonte: adaptado de Prawira(2020)

- Na mineração de dados, a tarefa escolhida será executada para serem encontrados os padrões de interesse.

  No caso da associação de dados, nesta etapa serão criadas as regras de associação, assim como cálculo do suporte, confiança e alavancagem.

- Por fim, a Interpretação dos resultados para identificar presença de padrões através de visualizações e modelos. Guimarães, Coelho e Nunes (2018) comentam sobre a falta de um especialista e como isto interfere na avaliação dos dados e resultados obtidos dificultando a tomada de decisão.

4 DESENVOLVIMENTO

A análise será feita sobre uma base de dados retirada do ERP de vendas de um mercado varejista de pequeno porte. Cada instância ou linha corresponde à passagem de um produto comprado em um determinado caixa. As colunas são atributos deste evento e correspondem ao horário, à data, ao valor de compra, à quantidade, entre outras, conforme resumida na Tabela 5.

Tabela 5 – Primeiras transações de vendas em Excel pós formatação

Código NF	17683	17684	17684	17685
Caixa	2	2	2	2
Data da venda	01/04/2020	01/04/2020	01/04/2020	01/04/2020
Hora da venda	07:23:00	07:24:00	07:24:00	07:25:00
Barra do produto	9974	78600010	9974	3
Código do produto	2325	4815	2325	2397
Descrição	PAO FRANCES C/ SAL	TIC TAC MENTA 16G	PAO FRANCES C/ SAL	PAO DE QUEIJO CASEIRO KG
Quantidade	0,605	1	0,08	0,306
Valor Bruto	4,83	1,75	0,64	4,9
Valor Venda	4,83	1,75	0,64	4,9
Custo Unitário	2,12	1,09	2,12	8,08
ICMS	7	18	7	12
Custo sem ICMS	1,97	0,89	1,97	7,11
Valor ICMS	0,13	0,32	0,02	0,39

Fonte: Elaborado pelo autor

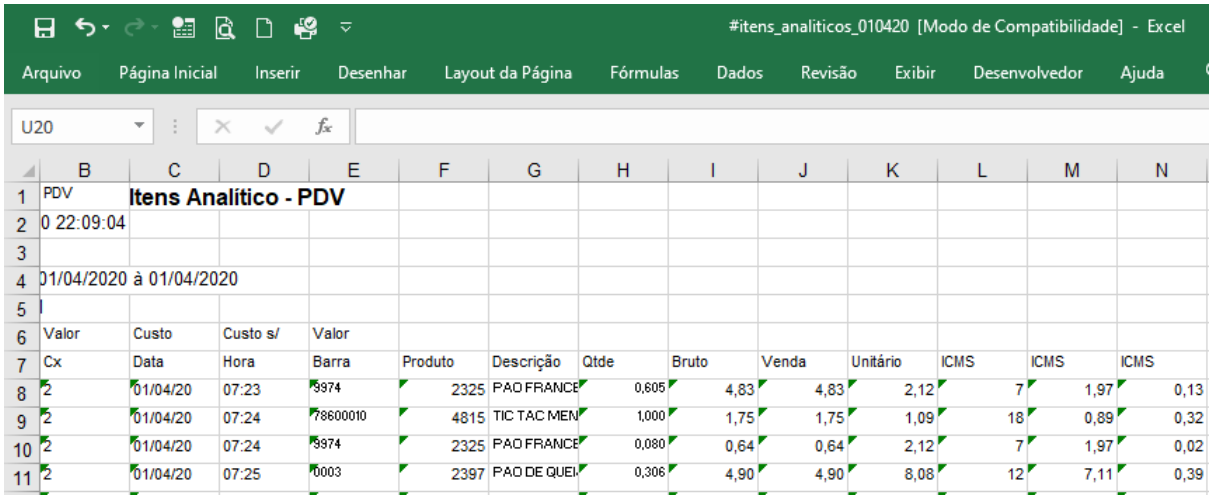
1.1. Seleção dos dados

A seleção ocorreu através do download manual de tabelas de venda diárias correspondente ao período de 1 ano (entre 1/10/2019 e 31/10/2020). Os arquivos diários de venda precisaram ser transformados em 1 único arquivo Excel. Isto foi feito com utilização do VBA (*Virtual Basic for Applications*) que é a linguagem de programação embutida ao Excel. Neste momento foram retiradas linhas não conformes (vazias ou que continham valores totais ou que não tratavam do assunto).

Após formatação, a tabela Excel apresentou 2.446.524 transações ou linhas de um total de 410.987 vendas.

A Figura 2 mostra trecho da tabela Excel de vendas previamente à correção. Pode-se observar, por exemplo, que as linhas 1 a 5 não apresentam utilidade para análise e precisaram ser excluídas, assim como, as linhas 6 e 7 apresentaram o nome das colunas de forma separada e precisaram de correção.

Figura 2 – Vendas em Excel retirada do ERP



#itens_analiticos_010420 [Modo de Compatibilidade] - Excel

Arquivo | Página Inicial | Inserir | Desenhar | Layout da Página | Fórmulas | Dados | Revisão | Exibir | Desenvolvedor | Ajuda

U20

	B	C	D	E	F	G	H	I	J	K	L	M	N
1	PDV	Itens Analítico - PDV											
2	0 22:09:04												
3													
4	01/04/2020 à 01/04/2020												
5													
6	Valor	Custo	Custo s/	Valor									
7	Cx	Data	Hora	Barra	Produto	Descrição	Qtde	Bruto	Venda	Unitário	ICMS	ICMS	ICMS
8	2	01/04/20	07:23	9974	2325	PAO FRANCE	0,605	4,83	4,83	2,12	7	1,97	0,13
9	2	01/04/20	07:24	78600010	4815	TIC TAC MEN	1,000	1,75	1,75	1,09	18	0,89	0,32
10	2	01/04/20	07:24	9974	2325	PAO FRANCE	0,080	0,64	0,64	2,12	7	1,97	0,02
11	2	01/04/20	07:25	0003	2397	PAO DE QUEI	0,306	4,90	4,90	8,08	12	7,11	0,39

Fonte: Elaborado pelo autor

1.2. Pré-processamento dos dados

Neste momento, foi utilizada a linguagem *python* para manipulação dos dados. O ambiente de trabalho utilizado foi *jupyter notebook* por apresentar facilidade de uso assim como de transferência e salvamento automático.

Para iniciar o processamento, a tabela de vendas teve a remoção de colunas não relevantes a esta análise como quantidade, valor de compra, valor de venda. Em seguida, houve a necessidade de agrupar a tabela de vendas a outra da tabela de produtos a qual continha a categoria de cada produto. Este agrupamento foi feito para que a descrição dos produtos fosse substituída pela categoria e, assim, o número total de atributos fosse reduzido (total de 5.371 descrições únicas para cerca de 178 categorias).

Também foram formatadas as colunas de data para facilitar filtros de data como dias da semana e mês. A partir disto, a base foi segmentada em 13 tabelas. Cada

tabela representa vendas de 1 mês no modelo transacional, em outras palavras, com a transação de um único produto em cada linha.

Tabela 6 – Trecho de tabela de vendas pós inserção de categoria

Código NF	Descrição	Categoria
50	PAO FRANCES C/ SAL	PADARIA PROD PROPRIA
51	MOCA FIESTA NESTLE BEIJINHO	DOCES EM GERAL
52	BROA FUBA KG	PADARIA PROD PROPRIA
53	MISTURA P/ EMPANAR HIKARI	OUTRAS FARINHAS
53	PAO FRANCES C/ SAL	PADARIA PROD PROPRIA
53	DOCE PE DE MOCA 50G	DOCES EM GERAL
53	PAO DOCE AMANTEGADO	PADARIA PROD PROPRIA
53	MINI SALSICHA UN	PADARIA PROD PROPRIA
53	PAO FORMA VISCONTI	COMPLEM. ALIMENTAR

Fonte: Elaborado pelo autor

A Tabela 6 mostra como produtos com descrição diferentes como pão francês com sal e broa de fubá kg apresentam categorias semelhantes e, portanto, possibilitarão a redução na dimensão da tabela em modelo tabular ao transformar categorias em colunas (como será visto logo em seguida).

1.3. Transformação dos dados

Transformar os dados trata-se de modificar a tabela transacional de vendas mensal para assumir o modelo tabular em que o indicador da venda esteja no índice das linhas e a categoria de cada produto não duplicado nas colunas. O valor 1 representa que o produto da coluna estava contido na venda da linha, enquanto 0 representa que o produto não tenha sido comprado. Por exemplo, a Tabela 7 mostra que a venda 36.840 contem frios e ceras, mas não contem hortifruti, café em pó, leites de saquinho ou linha diet/light.

Tabela 7 – Vendas em modelo tabular

Código NF	HORTIFRUTI	CAFE EM PO	FRIOS	CERAS	LEITES SAQUINHO	LINHA DIET/LIGHT
36836	0	0	0	0	0	0
36837	0	1	0	0	0	0
36840	0	0	1	1	0	0
36841	0	0	0	0	0	0
36842	0	0	0	0	1	1
36843	0	0	0	0	0	0
36844	1	0	0	0	1	0
36845	1	0	0	0	0	0

Fonte: Elaborado pelo autor

1.4. Mineração de dados

Para realizar a associação dos dados, foi utilizado o software gratuito Weka. A tabela de vendas salva como arquivo .csv foi importada ao software e formatada para ter valores nominais. Com isto, o método Apriori pode ser utilizado.

Dois indicadores foram informados antes da inicialização: o suporte mínimo e a confiança mínima. Ambos indicadores servem como filtros para o modelo para que regras com valor de suporte e confiança inferiores aos informados sejam desconsideradas.

Neste caso, a confiança mínima foi arbitrada em 50% enquanto o suporte mínimo foi de 0,5%. Visto que o número de instâncias e atributos foi considerado relativamente alto, tais valores foram arbitrados para que nenhuma regra fosse desconsiderada previamente à observação do analista.

Figura 3 – Resultado do Apriori do software Weka

```
Apriori
=======

Minimum support: 0 (242 instances)
Minimum metric <confidence>: 0.5
Number of cycles performed: 20

Generated sets of large itemsets:

Size of set of large itemsets L(1): 92

Size of set of large itemsets L(2): 197

Size of set of large itemsets L(3): 69

Size of set of large itemsets L(4): 4

Best rules found:

 1. FRIOS=1 LEITES SAQUINHO=1 509 ==> PADARIA PROD PROPRIA=1 429    <conf:(0.84)> lift:(1.87) lev:(0) [199] conv:(3.46)
 2. HORTIFRUTI=1 MOLHOS E POLPA TOMAT=1 453 ==> ACOUGUE=1 356    <conf:(0.79)> lift:(3.07) lev:(0) [240] conv:(3.44)
 3. FRIOS=1 LINHA DIET/LIGHT=1 733 ==> PADARIA PROD PROPRIA=1 572    <conf:(0.78)> lift:(1.73) lev:(0) [242] conv:(2.49)
 4. FRIOS=1 LEITE LONGA VIDA=1 510 ==> PADARIA PROD PROPRIA=1 396    <conf:(0.78)> lift:(1.73) lev:(0) [166] conv:(2.44)
 5. FRIOS=1 5289 ==> PADARIA PROD PROPRIA=1 4062    <conf:(0.77)> lift:(1.71) lev:(0.03) [1681] conv:(2.37)
 6. FRIOS=1 REFRIG PET SABORES=1 636 ==> PADARIA PROD PROPRIA=1 486    <conf:(0.76)> lift:(1.7) lev:(0) [199] conv:(2.32)
 7. FRIOS=1 REFRI 2,25L E 2,5LT=1 374 ==> PADARIA PROD PROPRIA=1 269    <conf:(0.72)> lift:(1.6) lev:(0) [100] conv:(1.94)
 8. HORTIFRUTI=1 OLEOS DE SOJA=1 359 ==> ACOUGUE=1 256    <conf:(0.71)> lift:(2.78) lev:(0) [164] conv:(2.57)
 9. IOGURTE SAQUINHO=1 1028 ==> PADARIA PROD PROPRIA=1 722    <conf:(0.7)> lift:(1.56) lev:(0.01) [259] conv:(1.84)
10. PADARIA PROD PROPRIA=1 MOLHOS E POLPA TOMAT=1 504 ==> ACOUGUE=1 349    <conf:(0.69)> lift:(2.7) lev:(0) [219] conv:(2.4)
```

Fonte: Elaborado pelo autor

A Figura 3 mostra a síntese da análise oferecida pelo software Weka. Pode-se dar foco à quantidade de instâncias, à confiança mínima, ao número de ciclos e às 10 regras criadas a partir dos dados selecionados. Neste software, as regras aparecem em ordem decrescente em relação à métrica escolhida, neste caso, a confiança. A partir disto, poderia ser afirmado que 429 vendas de produtos de padaria ocorrem dentro de 509 vendas com leites de saquinho e frios (Regra 1). Isto apresenta uma confiança de 84%, uma confiança considerada relativamente alta. Contudo, estas 429 vendas representam apenas 0,88% das 48.463 vendas ocorridas em outubro de 2020.

1.5. Interpretação

Após aplicação do método Apriori sobre os 13 meses de venda, os resultados foram sintetizados na Tabela 8. Nesta, são apresentadas as 25 regras criadas dentro das análises junto à soma de suas ocorrências e à confiança média. Inicialmente, pode-se perceber que 5 regras estiveram em todas as 13 análises (regras 1, 2, 3, 4 e 6). 4 destas regras envolvem a compra de um produto de Padaria sendo que um

produto de Frios foi comprado. Em 2 destas 4, laticínios são comprados em conjunto. A regra 3, também é interessante por mostrar uma ação que pode ser considerada comum que é a compra de um produto de açougue junto à hortifruti e molhos. A regra 4 chama ainda mais atenção. Ela nos mostra que em todos os meses, a compra de produtos da linha diet/light esteve associada a compra de padaria. Em uma revisão rápida pode ser notado que diversos produtos de bebida como sucos (sucos de 1L, refrescos e outros) estavam inseridos nesta categoria, o que demonstra que o cadastro foi feito de maneira errônea, visto que já existe uma categoria para estes produtos.

Tabela 8 – Regras de associação para os 13 meses de venda

Regra	Regras de associação	Ocorrências	Confiança média
1	FRIOS | LEITES SAQUINHO ==> PADARIA PROD PROPRIA	13	84%
2	LEITE LONGA VIDA |FRIOS ==> PADARIA PROD PROPRIA	13	80%
3	HORTIFRUTI | MOLHOS E POLPA TOMAT ==> ACOUGUE	13	77%
4	FRIOS | LINHA DIET/LIGHT ==> PADARIA PROD PROPRIA	13	75%
5	REFRIG PET SABORES | FRIOS ==> PADARIA PROD PROPRIA	5	75%
6	FRIOS ==> PADARIA PROD PROPRIA	13	75%
7	TABACARIA E FUMOS | FRIOS ==> PADARIA PROD PROPRIA	1	75%
8	REFRIG PET SABORES | FRIOS ==> PADARIA PROD PROPRIA	8	73%
9	HORTIFRUTI | FEIJAO ==> ACOUGUE	4	73%
10	SALGADINHOS | FRIOS ==> PADARIA PROD PROPRIA	5	72%
11	FRIOS | CHOCOLATES TABLETES ==> PADARIA PROD PROPRIA	2	71%
12	FRIOS | REFRI 2,25L E 2,5LT ==> PADARIA PROD PROPRIA	6	71%
13	PADARIA PROD PROPRIA | MOLHOS E POLPA TOMAT ==> ACOUGUE	7	71%
14	HORTIFRUTI |OLEOS DE SOJA ==> ACOUGUE	4	70%
15	PADARIA PROD PROPRIA | MOLHOS E POLPA TOMAT ==> ACOUGUE	2	70%
16	HORTIFRUTI | CAFE EM PO ==> ACOUGUE	2	70%
17	HORTIFRUTI | CONDIMENTOS ==> ACOUGUE	1	70%
18	IOGURTE SAQUINHO ==> PADARIA PROD PROPRIA	6	70%
19	LINHA DIET/LIGHT | LEITES SAQUINHO ==> PADARIA PROD PROPRIA	3	69%
20	COZINHA | HORTIFRUTI ==> ACOUGUE	1	69%
21	HORTIFRUTI |SALGADINHOS ==> ACOUGUE	1	69%
22	HORTIFRUTI | REFRI 2,25L E 2,5LT ==> ACOUGUE	2	69%
23	PADARIA PROD PROPRIA | HORTIFRUTI | REFRIG PET SABORES ==> ACOUGUE	1	68%
24	LEITES SAQUINHO ==> PADARIA PROD PROPRIA	3	68%
25	HORTIFRUTI | MASSAS ==> ACOUGUE	1	67%

Fonte: Elaborado pelo autor

A Tabela 8 ainda mostra que 16 das 25 regras contem produtos de padaria. As outras 9 restantes demonstram que produtos de açougue serão comprados, quando produtos hortifruti estão na cesta de compras. A frequência quase unânime destas categorias indica a necessidade de separá-las da base para investigações mais aprofundadas em outros produtos. E neste momento observa-se a iteratividade da mineração dados, a necessidade de utilizar filtros e/ou de modificar a base para contestar hipóteses criadas ou para verificar a existência de novas hipóteses potencialmente relevantes.

Outras 2 análises serão apresentadas de maneira sucinta, a próxima terá como restrição a presença de um produto único e a última, a presença de uma categoria.

Nesta segunda análise, na seleção dos dados, foram escolhidas apenas vendas que continham um produto específico, o salame fatiado de 100g. O produto foi escolhido arbitrariamente, porém pode ser especulado que este produto geralmente não é consumido sozinho, além de que sua embalagem é pequena e lhe permite ser pendurado facilitando posicionamento estratégico. Na etapa de pré-processamento, foram apenas mantidas as colunas de Descrição e Código NF. Na etapa de transformação, as descrições dos produtos foram colocadas nas colunas e os códigos de nota fiscal nos índices das linhas para obtermos a tabela em modelo tabular. Após a mineração destes dados, obteve-se o resultado da Figura 4.

Figura 4 – Resultado Apriori para vendas de salame fatiado de 100g no Weka

```
Apriori
=======

Minimum support: 0.06 (125 instances)
Minimum metric <confidence>: 0.5
Number of cycles performed: 19

Generated sets of large itemsets:

Size of set of large itemsets L(1): 22

Size of set of large itemsets L(2): 15

Size of set of large itemsets L(3): 1

Best rules found:

 1. LIMAO=1 350 ==> SALAME FATIADO AURORA 100G=1 294    <conf:(0.84)> lift:(1.4) lev:(0.04) [84] conv:(2.46)
 2. CERV SKOL LATA 269ML=1 163 ==> SALAME FATIADO AURORA 100G=1 127    <conf:(0.78)> lift:(1.3) lev:(0.01) [29] conv:(1.76)
 3. CERV SKOL LATA 350ML=1 176 ==> SALAME FATIADO AURORA 100G=1 132    <conf:(0.75)> lift:(1.25) lev:(0.01) [26] conv:(1.57)
 4. QUEIJO MUSSARELA=1 405 ==> SALAME FATIADO AURORA 100G=1 296    <conf:(0.73)> lift:(1.22) lev:(0.02) [53] conv:(1.47)
 5. PAO DOCE VARIADOS=1 160 ==> PAO FRANCES C/ SAL=1 114    <conf:(0.71)> lift:(1.74) lev:(0.02) [48] conv:(2.01)
 6. REFRIG COCA COLA 2,25LT=1 194 ==> SALAME FATIADO AURORA 100G=1 137    <conf:(0.71)> lift:(1.18) lev:(0.01) [20] conv:(1.34)
 7. PAO FRANCES C/ SAL=1 QUEIJO MUSSARELA=1 237 ==> SALAME FATIADO AURORA 100G=1 166    <conf:(0.7)> lift:(1.17) lev:(0.01) [23] conv:(1.32)
 8. TOMATE=1 195 ==> SALAME FATIADO AURORA 100G=1 131    <conf:(0.67)> lift:(1.12) lev:(0.01) [14] conv:(1.2)
 9. CEBOLA=1 193 ==> SALAME FATIADO AURORA 100G=1 128    <conf:(0.66)> lift:(1.11) lev:(0.01) [12] conv:(1.17)
10. CARNE ACEM BOVINO=1 217 ==> SALAME FATIADO AURORA 100G=1 143    <conf:(0.66)> lift:(1.1) lev:(0.01) [12] conv:(1.16)
```

Fonte: Elaborado pelo autor

Como interpretação do resultado, a Figura 4 mostra que produtos como limão e cervejas lata de 269ml e 350ml foram comprados em conjunto com salame com grau de confiança superior a 75%. Vale observar que o suporte ainda pode ser considerado baixo, visto que as regras ocorreram em cerca de 5 a 10% das vendas totais de salame (2.281 instâncias ou vendas). Outro ponto importante é o número de atributos, nesta seleção foram utilizadas as descrições dos produtos, logo o valor foi bem superior ao de categorias. Neste modelo, houveram 2.435 atributos ou colunas.

Como última análise, a categoria selecionada para restringir o número de instância foi a de molhos. Na tabela transacional, o número de linhas foi de 201.633. Ao ser transformada para modelo tabular, mantendo como coluna as descrições dos produtos, a dimensão da tabela passou a ser de 13.882 instâncias e 4.107 atributos. Ao importar esta tabela para o software Weka e aplicar o Apriori, houve a espera de cerca 25 minutos de processamento até que um erro fosse lançado. O erro representa que a memória para processamento foi excedida. Isto indica que 1) outra filtragem deva ser feita, 2) uma máquina mais robusta (com maior memória) seja utilizada e/ou 3) o processo de análise deva ser feito de outra forma.

5 CONSIDERAÇÕES FINAIS

Buscando entender como produtos vendidos se relacionam para poder identificar potenciais promoções ou melhorar a compra por parte do cliente, seja no posicionamento do produto ou em facilitar sua disposição, pode-se perceber que a compra de produtos diários como os de panificação, frios e laticínios se mostraram os mais relevantes. Em 13 meses observados, a compra em conjuntos destas categorias apresentou cerca de 80% de confiança. Para análise de um produto único, o salame fatiado de 100g esteve associado a venda de limões e cervejas de 269ml e 350ml com confiança de cerca de 75%.

A análise de cesto de compras se mostrou não trivial como visto na revisão bibliográfica. Apesar da simplicidade na interpretação dos resultados, houveram certas dificuldades na manipulação e na modelagem dos dados. Algumas dificuldades podem ser listadas como: 1) modelos não podem conter muitos atributos ao serem inseridos no software Weka, pois a memória do sistema pode chegar ao seu limite; 2) a classificação dos dados pode ter sido feita de maneira incorreta ou seguido um

raciocínio não identificado previamente à análise; 3) a seleção dos dados pode exigir habilidades técnicas de programação e/ou permissões de acessos, caso a base de dados se encontre em ERPs que não permitam acesso direto à fonte dos dados; e 4) o número elevado de dados exige uma noção de segmentação de dados relativamente aprofundada para que não haja duplicação ou replicação de um contexto.

Apesar das dificuldades encontradas, insights dos resultados obtidos permitiram que ideias da gerência sobre consumo dos clientes tivessem embasamento. Desta forma, um método de abordagem perante a criação de promoções e posicionamento de produtos na loja pode ser estabelecido, visto que foram identificados como dados devem ser selecionados, processados, transformados, analisados e interpretados.

Em pesquisas futuras, seria interessante a verificação da análise de cesta compras como forma de indicador na realocação de produtos. Realizar a mudança de produtos em uma loja exige definição de estratégia, organização e treinamento de pessoal para execução. Utilizar os resultados do Apriori como indicadores de performance na mudança seria de grande utilidade para tornar todo o processo de gestão de produtos mais robusta e independente.

REFERÊNCIAS

SUPRIANTO, Panjaitan et al. **Implementation of Apriori Algorithm for Analysis of Consumer Purchase Patterns**. Journal of Physics. Conference Series 1255.1, 2019.

HERMALIANI, Eni Heni, et. al. **Data Mining Technique to Determine the Pattern of Fruits Sales & Supplies Using Apriori Algorithm**. Journal of Physics. Conference Series 1641.1, 2020.

ANSELMO, Filomena Clara Gouveia. **Regras de Associação-Market Basket Analysis-Items Frequentes e Itens Raros**. FEP Economia e Gestão. U. Porto, 2017.

PRAWIRA, Tresna Yudha; SUNARDI, Sunardi; FADLIL Abdul. **Market Basket Analysis To Identify Stock Handling Patterns & Item Arrangement Patterns Using Apriori Algorithms**. Khazanah Informatika 6.1, 2020. p 33-41.

CARVALHO, Marcelo Batista De; TSUNODA, Denise Fukumi. **Análise De Dados Em Artigos Recuperados Da Web of Science (WoS)**. Encontros Bibli, 2018. p 112-25.

GUIMARÃES, Norton Coelho; NUNES, Rosângela Da Silva. **Descoberta De Regras De Associação Com Algoritmo Tertius**. Multi-Science Journal 1.7, 2018. p 41-45.

SOUZA, Wanderley De, Junior; SARFATI, Gilberto. **Adoption of a Loyalty Program through a Mobile Application: Analysis of the Supermercados Meu Econômico Case Using the Basket Analysis**. Revista Brasileira De Marketing 19.2, 2020. p 287.

SILVA, Jesus, et al. **Association rules extraction for customer segmentation in the SMEs sector using the apriori algorithm**. Procedia Computer Science 151, 2019. p 1207-1212.

COSTA, Bernardo. **Uma Introdução ao Algoritmo Apriori.** Medium, 2019. Disponível em: https://medium.com/@bernardo.costa/uma-introdu%C3%A7%C3%A3o-ao-algoritmo-apriori-60b11293aa5a. Acesso em: setembro de 2022.

BEHESHTIAN-ARDAKANI, Arash; FATHIAN, Mohammad; GHOLAMIAN, Mohammad Reza. **A Novel Model for Product Bundling and Direct Marketing in E-commerce Based on Market Segmentation**. Decision Science Letters 7.1, 2018. p 39-54.

WANG, Diya; ZHAO, Yixi**. Using News to Predict Investor Sentiment: Based on SVM Model**. Procedia Computer Science, 2020. p. 191-199 Disponível em: (https://www.sciencedirect.com/science/article/pii/S187705092031588X).

PRODANOV, Cleber Cristiano; FREITAS, Ernani Cesar de. **Metodologia do Trabalho Científico: Métodos e Técnicas da Pesquisa e do Trabalho Acadêmico**. 2ª ed. Universidade Feevale. Nova Hamburgo, RS, 2013.

SILVA, Edna Lúcia da; MENEZES, Estera Muszkat. **Metodologia da pesquisa e elaboração** de dissertação. Revista Atual 3ª ed. Florianópolis: Laboratório de Ensino a Distância da UFSC, 2001. p 121.