UNIVERSIDADE FEDERAL DE PELOTAS
Centro de Desenvolvimento Tecnológico
Programa de Pós-Graduação em Computação

Dissertação

Recomendação de Produtos através do método de Market Basket Analysis Aplicado ao Cenário de Big Data

Francine Machado Moraes

Pelotas, 2024

Francine Machado Moraes

Recomendação de Produtos através do método de Market Basket Analysis Aplicado ao Cenário de Big Data

Dissertação apresentada ao Programa de Pós-Graduação em Computação do Centro de Desenvolvimento Tecnológico da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Tiago Thompsen Primo

Pelotas, 2024

Universidade Federal de Pelotas / Sistema de Bibliotecas
Catalogação da Publicação

M827r Moraes, Francine Machado

Recomendação de produtos através do método de Market Basket Analysis aplicado ao cenário de Big Data [recurso eletrônico] / Francine Machado Moraes ; Tiago Thompsen Primo, orientador. — Pelotas, 2023.
65 f. : il.

Dissertação (Mestrado) — Programa de Pós-Graduação em Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, 2023.

1. Recomendação de Produtos. 2. Sistemas de Recomendação. 3. Regras de Associação. 4. Análise de Cesta de Mercado. I. Primo, Tiago Thompsen, orient. II. Título.

CDD 005

Elaborada por Maria Inez Figueiredo Figas Machado CRB: 10/1612

Francine Machado Moraes

Recomendação de Produtos através do método de Market Basket Analysis Aplicado ao Cenário de Big Data

Dissertação aprovada, como requisito parcial, para obtenção do grau de Mestre em Ciência da Computação, Programa de Pós-Graduação em Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas.

Data da Defesa: 11 de dezembro de 2023

Banca Examinadora:

Prof. Dr. Tiago Thompsen Primo (orientador)
Doutor em Computação pela Universidade Federal do Rio Grande do Sul.

Prof. Dr. Marilton Sanchotene de Aguiar
Doutor em Computação pela Universidade Federal do Rio Grande do Sul.

Prof. Dr. Ulisses Brisolara Corrêa
Doutor em Computação pela Universidade Federal de Pelotas.

Prof. Dr. Emanuel Marques Queiroga
Doutor em Computação pela Universidade Federal de Pelotas.

AGRADECIMENTOS

Primeiramente, quero expressar minha profunda gratidão ao meu orientador, Tiago Primo, pela orientação excepcional, paciência e pelo constante apoio ao longo deste processo que foram cruciais para a conclusão desta dissertação.

Agradeço aos professores e membros da banca por dedicarem seu tempo à avaliação e discussão deste trabalho durante a defesa e seminários.

Expresso também minha profunda gratidão ao meu noivo pelo apoio constante e pela força fundamental que me proporcionou ao longo de todo esse processo desafiador.

Agradeço sinceramente à empresa envolvida neste trabalho pela sua notável disponibilidade e colaboração.

Finalmente, gostaria de agradecer a todos que, de alguma forma, contribuíram para este projeto, direta ou indiretamente. Este trabalho não teria sido possível sem o apoio de cada um de vocês.

...

RESUMO

MORAES, Francine Machado. **Recomendação de Produtos através do método de Market Basket Analysis Aplicado ao Cenário de Big Data**. Orientador: Tiago Thompsen Primo. 2024. 66 f. Dissertação (Mestrado em Ciência da Computação) – Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, 2024.

O Market Basket Analysis (MBA), ou análise de cesta de mercado em português, consiste em extrair uma grande quantidade de associações de produtos a partir da base de dados de transações. Originalmente, era utilizada para analisar cestas de compras em mercados, mas não se limita a este cenário, podendo ser aplicada em qualquer negócio que venda produtos. Os sistemas de recomendação, amplamente empregados em e-commerces destacados, como Amazon e Mercado Livre, são comumente fundamentados em técnicas de Mineração de Dados e MBA.

Este estudo apresenta uma proposta de sistema de recomendação adaptável, desenvolvido para operar tanto em ambientes digitais quanto em lojas físicas de uma extensa rede varejista.

Este trabalho propõe um sistema de recomendação adaptável tanto para plataformas digitais quanto para lojas físicas de uma grande rede varejista utilizando a técnica MBA. O diferencial desta abordagem reside na capacidade de re- alizar recomendações de produtos em larga escala, otimizando o processamento de grandes volumes de dados por meio de ferramentas e estratégias de processamento distribuído os quais foram essenciais para a resolução dos desafios encontrados relacionados a big data. A motivação para tal desenvolvimento decorre da observação de que a maioria das pesquisas existentes concentra-se em sistemas de recomendação ade- quados apenas para pequenas quantidades de dados além de auxiliar outros pesquisadores na resolução dos desafios existentes neste cenário. Foi possível entender as relações entre os produtos recomendados separados por região, setores e estações do ano onde foi possível entender os comportamentos entre regiões norte e sul do país, produtos do mesmo setor se relacionam fortemente, eventos climáticos influenciam na recomendação além do recomendador entender as variações do produto e recomenda-los entre si. A eficácia do sistema proposto foi avaliada com uma estimativa de precisão de 22,53% para o canal digital e 31,28% para as lojas físicas considerando dezembro/2023.

Palavras-chave: Recomendação de Produtos. Sistemas de Recomendação. Regras de Associação. Análise de Cesta de Mercado.

ABSTRACT

MORAES, Francine Machado. **Product Recommendation through the Market Basket Analysis Method Applied to the Big Data Scenario**. Advisor: Tiago Thompsen Primo. 2024. 66 f. Dissertation (Masters in Computer Science) – Technology Development Center, Federal University of Pelotas, Pelotas, 2024.

Market Basket Analysis (MBA), or Análise de Cesta de Mercado in Portuguese, involves extracting a large number of product associations from transactional databases. Originally used to analyze shopping baskets in markets, it is not limited to this scenario and can be applied to any business that sells products. Recommendation systems, widely employed in prominent e-commerce platforms like Amazon and Mercado Livre, are commonly grounded in Data Mining and MBA techniques.

This study presents a proposal for an adaptable recommendation system, developed to operate in both digital and physical environments of an extensive retail network.

This work proposes an adaptable recommendation system for both digital platforms and physical stores of a large retail network using the MBA technique. The differentiating factor of this approach lies in its ability to make product recommendations on a large scale, optimizing the processing of large volumes of data through tools and distributed processing strategies, which were essential for resolving challenges related to big data. The motivation for this development stems from the observation that most existing research focuses on recommendation systems suitable only for small data quantities, in addition to aiding other researchers in addressing challenges in this scenario. It was possible to understand the relationships between recommended products based on region, sectors, and seasons, where behaviors between the northern and southern regions of the country were observed, products from the same sector strongly related, weather events influencing recommendations, and the recommender understanding product variations and recommending them accordingly. The effectiveness of the proposed system was assessed with an accuracy estimate of 22.53% for the digital channel and 31.28% for physical stores considering december/2023.

Keywords: Product Recommendation. Recommendation System. Association Rule. Market Basket Analysis.

SUMÁRIO

1 INTRODUÇÃO

E-commerce, ou comércio eletrônico em português, refere-se às transações comerciais realizadas por meio de equipamentos eletrônicos, como computadores e *smartphones*. Essa modalidade de comércio chegou ao Brasil em meados dos anos 2000, conforme indica (CRUZ, 2021).

Nos primeiros dez anos após sua introdução, o *e-commerce* começou a ganhar fôlego no Brasil. O estudo de (CRUZ, 2021) aponta que, entre 2011 e 2020, houve um aumento de 259,2% no número de usuários de comércio eletrônico e um aumento de 351,3% no faturamento dessas empresas. Segundo o autor, esse crescimento não se trata apenas de uma evolução do *e-commerce*, mas também de uma mudança no comportamento de compra dos consumidores. Ele observa: *"No passado, as pessoas não se sentiam confortáveis em fazer compras pelo celular ou em digitar o número do cartão de crédito em um site. Hoje, muitas têm o cartão cadastrado em vários sites, buscando praticidade e comodidade."*

Com a migração crescente para plataformas de compras *online*, uma quantidade massiva de transações passou a ser armazenada em bancos de dados, tornando a capacidade de processar esse volume tão crucial quanto a de produzir esses dados.

No contexto do *e-commerce*, oferecer uma experiência personalizada aos clientes é vital. Isso pode ser alcançado por meio de sistemas de recomendação, como os empregados pela Amazon, que sugerem produtos relacionados ao que o cliente tem no carrinho, incentivando-o a considerar itens complementares.

Uma estratégia eficaz para melhorar as vendas é a mineração de dados, um processo que, segundo (IBM, 2023a), identifica padrões em grandes volumes de informações. Ele destaca sua aplicação em projetos de aprendizado de máquina, especialmente na identificação de relações entre variáveis. Uma técnica específica, a MBA, é descrita por (BLATTBERG, 2008) como um meio de entender as associações entre produtos vendidos juntos ou em compras subsequentes.

O autor (GURUDATH, 2020) conduziu um estudo de caso para uma empresa de entrega rápida de produtos de supermercado, realizando uma análise descritiva dos padrões de compra do cliente. O objetivo era facilitar a reposição e manter um estoque

adequado de produtos, utilizando a análise de cesta de mercado e o algoritmo Apriori em aproximadamente 3 milhões de registros. Por sua vez, o autor (KAMEPALLI, 2019) propôs um novo modelo de mineração de padrões frequentes e infrequentes baseado em pesos para bancos de dados de comércio eletrônico em tempo real, visando identificar padrões em grandes conjuntos de dados.

Os sistemas de recomendação, amplamente usados em *e-commerces* como Amazon e Mercado Livre, baseiam-se frequentemente em técnicas de Mineração de Dados e MBA. Esses sistemas oferecem sugestões com base no comportamento de navegação do usuário e em produtos complementares, como explica (RICCI, 2011). O MBA tem sido uma ferramenta valiosa para esses sistemas, auxiliando na identificação de associações entre produtos e no processamento de grandes volumes de dados.

Este estudo apresenta uma proposta de sistema de recomendação adaptável, desenvolvido para operar tanto em ambientes digitais quanto em lojas físicas de uma extensa rede varejista. O diferencial desta abordagem reside na capacidade de realizar recomendações de produtos em larga escala, otimizando o processamento de grandes volumes de dados por meio de ferramentas e estratégias de processamento distribuído. A motivação para tal desenvolvimento decorre da observação de que a maioria das pesquisas existentes concentra-se em sistemas de recomendação adequados apenas para pequenas quantidades de dados além de auxiliar outros pesquisadores na resolução dos desafios existentes neste cenário. Foi possível entender as relações entre os produtos recomendados separados por região, setores e estações do ano onde foi possível entender os comportamentos entre regiões norte e sul do país, produtos do mesmo setor se relacionam fortemente, eventos climáticos influenciam na recomendação além do recomendador entender as variações do produto e recomenda-los entre si. A eficácia do sistema proposto foi avaliada com uma estimativa de precisão de 22,53% para o canal digital e 31,28% para as lojas físicas considerando dezembro/2023.

Este documento está estruturado de forma a fornecer uma compreensão abrangente sobre diversos tópicos relacionados a recomendação de produtos aplicado ao cenário de big data utilizando MBA. Inicia-se com uma introdução, seguida por uma revisão do referencial teórico, abordando conceitos essenciais, como o contexto de aplicação, MBA e sistemas de recomendação. A seção subsequente detalha a metodologia da revisão sistemática da literatura, delineando perguntas de pesquisa, strings de busca, fontes de pesquisa, critérios de seleção e a análise dos resultados.

A proposta de solução é apresentada em detalhes, destacando as tecnologias utilizadas, infraestrutura e a limpeza e organização do dataset. Os desafios enfrentados e as estratégias adotadas são discutidos na seção seguinte. A análise de resultados é realizada, abordando aspectos como sazonalidade, volumetria de dados e distribuição de execução.

A seção subsequente elabora sobre a solução proposta, incluindo aspectos de recomendação e a estrutura geral do Dag. A discussão aprofundada dos resultados abrange a performance do modelo, análise por produto, setores, estações e produtos mais vendidos por região. O documento encerra com conclusões finais e referências bibliográficas para orientar futuras pesquisas nesse campo dinâmico do e-commerce.

2 REFERENCIAL BIBLIOGRÁFICO

Neste capítulo, serão abordados o contexto de aplicação desta tese e os conceitos fundamentais, incluindo *MBA*, sistemas de recomendação, bem como suas técnicas específicas, como filtragem colaborativa, filtragem baseada em conteúdo, filtragem híbrida e regras de associação.

2.1 Contexto de Aplicação

Canais de venda são os meios pelos quais empresas oferecem produtos ou serviços ao público-alvo, independentemente da área de atuação. São estratégias essenciais para que as empresas alcancem seus clientes de acordo com (SEBRAE, 2023a). Estes podem ser categorizados principalmente em canais físicos e canais online. Os canais físicos referem-se a locais como lojas, mercados e quiosques, onde as transações (vendas) são realizadas presencialmente, permitindo que o cliente interaja diretamente com o vendedor ou prestador de serviço. Por outro lado, o canal online, frequentemente referido como *e-commerce* ou comércio eletrônico, refere-se à compra e venda de produtos ou serviços através da internet, utilizando dispositivos eletrônicos como computadores e *smartphones* (CRUZ, 2021).

Historicamente, as transações em canais físicos têm sido a norma por séculos, desde os primeiros mercados e feiras. No entanto, o comércio eletrônico só se tornou possível com o advento da tecnologia da informação, especialmente após a invenção dos primeiros computadores. A análise indica que as forças da globalização e as grandes revoluções em Tecnologias da Informação e Comunicação (TIC) estão impulsionando o rápido crescimento do e-commerce global (KSHETRI, 2001). Apesar de sua introdução recente, levou cerca de uma década para que o *e-commerce* ganhasse tração e se tornasse lucrativo. De acordo com (CRUZ, 2021), o número de usuários de *e-commerce* cresceu em impressionantes 259,2% entre 2011 e 2020, enquanto o faturamento das empresas nesse segmento aumentou em 351,3% durante o mesmo período.

A combinação de transações em canais digitais e físicos resulta na geração de

uma quantidade massiva de dados. Por exemplo, considerando uma grande rede de lojas de departamento com 173 lojas físicas, além de um site e aplicativo, com uma média de 755 transações por filial diariamente, média de 6 itens por transação por filial diariamente e uma variedade total de 100 mil produtos, isso resulta em cerca de 130 mil transações por dia. Levando-se em conta a variedade de produtos vendidos, o volume de dados armazenados pode crescer exponencialmente. Essa grande quantidade de dados, juntamente com sua variedade e a velocidade com que são gerados, é comumente referida como *Big Data*.

Uma característica distintiva do setor de varejo é a sazonalidade, que o (SEBRAE, 2023b) define como períodos em que há uma variação significativa na demanda ou no volume de transações. Por exemplo, para a rede de lojas em questão, os períodos sazonais são influenciados por datas comemorativas mensais, levando a picos de vendas em certos momentos do ano.

2.2 MBA

O *MBA*, ou análise de cesta de mercado em português, consiste em extrair uma grande quantidade de associações de produtos a partir da base de dados de transações, como afirma (BLATTBERG, 2008). Originalmente, era utilizada para analisar cestas de compras em mercados, mas não se limita a este cenário, podendo ser aplicada em qualquer negócio que venda produtos. Assim, a análise de cesta de mercado permite entender o comportamento das relações (associações) entre os produtos e automatizar o processo para identificar quais produtos tendem a ser comprados juntos ou em compras subsequentes.

(BLATTBERG, 2008) afirma que as regras de associação nos permitem identificar relações negativas, positivas e complementares entre os elementos em uma base de dados. No cenário da empresa estudada neste trabalho, podemos exemplificar que, se os clientes compram travesseiros, eles também tendem a comprar fronhas, e o cliente que comprou travesseiro e fronha tende, em uma compra posterior, a adquirir o produto Jogo de Lençol.

Além das estratégias de *cross-sell* e *marketing*, (BLATTBERG, 2008) destaca o uso da técnica para redução de preços e promoções, afirmando: *“A redução de preço de um produto não só aumenta sua própria demanda, mas também a demanda de seu produto complementar. Ou seja, se dois produtos são complementares um ao outro, suas demandas tendem a estar positivamente associadas. Por outro lado, se dois produtos são substitutos, suas demandas tendem a ser negativamente correlacionadas.”*

2.3 Sistemas de recomendação

De acordo com (RICCI, 2011), Sistemas de Recomendação são ferramentas e técnicas de *software* que fornecem sugestões de itens úteis a um usuário. Estas sugestões, apresentadas pelo sistema, têm como objetivo apoiar os usuários em diversos processos de tomada de decisão, como escolher itens para comprar, músicas para escutar, entre outros. (BOBADILLA, 2013) comenta que o objetivo dos sistemas de recomendação é transformar os dados do cliente e seus interesses em previsões sobre seus futuros interesses. No contexto da recomendação de produtos, isso significa que, com base nos produtos que um cliente já comprou e demonstrou interesse, o sistema pode sugerir outros produtos que também possam ser do interesse desse cliente, considerando suas compras anteriores, características e comportamentos.

(BOBADILLA, 2013) também destaca os chamados métodos de similaridade para a implementação do sistema de recomendação, que podem ser baseados em similaridades entre usuários ou entre itens. O método de similaridade entre usuários consiste em encontrar usuários com características de perfil ou de compra semelhantes e recomendar itens bem avaliados por esses usuários. Já o método de similaridade entre itens se baseia na semelhança de outros itens com o item escolhido pelo usuário.

Além disso, com base nos métodos de similaridade, existem alguns algoritmos principais para desenvolver sistemas de recomendação: filtragem colaborativa, filtragem baseada em conteúdo e regras de associação.

A filtragem colaborativa, por exemplo, é amplamente utilizada em sistemas onde as recomendações são baseadas nas preferências de usuários similares (GOLDBERG, 1992). Já a filtragem baseada em conteúdo utiliza características dos itens para fazer recomendações, como visto em sistemas de recomendação de notícias (PAZZANI, 2007). As regras de associação, utilizadas em análises de cesta de mercado, identificam padrões de itens frequentemente comprados juntos (AGRAWAL, 1993).

2.3.1 Filtragem colaborativa

Os algoritmos de filtragem colaborativa preveem os possíveis interesses de um agente baseado em outros agentes com características similares às suas, como explica (BOBADILLA, 2013) e (PRIMO, 2013).

Para melhor entendimento, podemos dizer que, a filtragem colaborativa funciona da seguinte forma, por exemplo, um usuário A está comprando o produto P e este usuário A tem características parecidas com o usuário B. O usuário B já comprou o produto P em algum momento e também comprou o produto Z, logo, devido a semelhança dos usuários podemos recomendar o produto Z para o usuário A. Para gerar estas recomendações utilizando filtragem colaborativa o autor (JUNIOR, 2017) afirma que pode-se partir de dois tipos de técnicas: a *model-based* ou a *memory-based*.

A técnica *memory-based*, ou em português, baseada em memória computa as sugestões em memória utilizando funções matemáticas que calculam as similaridades em matrizes *user-user* ou *item-item*. Pode-se utilizar algoritmos de mercado baseado em vizinhança ou realizar o cálculo de similaridade utilizando por exemplo, distância euclidiana entre outras como explica o autor (UFG, 2017).

Na abordagem *user-user* monta-se uma matriz onde cada usuário é um vetor de tamanho N e após isto é computada a similaridade entre todos os usuários. Quanto mais similar (próximo) estiver um usuário ao usuário alvo da predição sendo calculada, maior será sua influência no resultado final. A figura 1 demonstra o processo de similaridade.

$$\begin{array}{c} \\ U_1 \\ U_2 \\ U_3 \\ U_N \end{array}\!\!\begin{array}{c} \begin{array}{cccc} I_1 & I_2 & I_3 & I_N \end{array} \\ \begin{bmatrix} 1 & & 4 & .. \\ & 3 & & .. \\ & 4 & 5 & .. \\ .. & .. & .. & .. \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ U_1 \\ U_2 \\ U_3 \\ U_N \end{array}\!\!\begin{array}{c} \begin{array}{cccc} U_1 & U_2 & U_3 & U_N \end{array} \\ \begin{bmatrix} 1 & Sim_{12} & Sim_{13} & Sim_{1N} \\ Sim_{21} & 1 & Sim_{23} & Sim_{2N} \\ Sim_{31} & Sim_{32} & 1 & Sim_{3N} \\ .. & .. & .. & .. \end{bmatrix} \end{array}$$

Figura 1 – Matriz user-item

Na abordagem *item-item* é feito o mesmo processo da user-user primeiro, monta-se a matriz considerando que cada item é um vetor de tamanho n. e depois, computa-se a similaridade entre todos os itens. A figura 2 demonstra o processo.

$$\begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_N \end{array}\!\!\begin{array}{c} \begin{array}{cccc} U_1 & U_2 & U_3 & U_N \end{array} \\ \begin{bmatrix} 1 & & & .. \\ & 3 & 4 & .. \\ 4 & & 5 & .. \\ .. & .. & .. & .. \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_N \end{array}\!\!\begin{array}{c} \begin{array}{cccc} I_1 & I_2 & I_3 & I_N \end{array} \\ \begin{bmatrix} 1 & Sim_{12} & Sim_{13} & Sim_{1N} \\ Sim_{21} & 1 & Sim_{23} & Sim_{2N} \\ Sim_{31} & Sim_{32} & 1 & Sim_{3N} \\ .. & .. & .. & .. \end{bmatrix} \end{array}$$

Figura 2 – Matriz item-item

A técnica *model-based* utiliza as avaliações dos usuários para o treinamento do modelo que gerará as recomendações. Os algoritmos de mercado comumente utilizados neste cenário utilizam o método SVD (*Singular Value Decomposition*) da álgebra linear.

Segundo (JUNIOR, 2017), o SVD “tenta agregar uma grande quantidade de features em um conjunto reduzido de conceitos” (atribuindo-lhes também algum parâmetro de intensidade), “acepções” mais abrangentes das features de quais foram derivados, conseguindo dessa forma generalizar o conhecimento obtido e torná-los aplicáveis a

casos mais variados". O SVD auxilia também na remoção de *outliers* e na redução da dimensionalidade.

2.3.2 Filtragem baseada em conteúdo

Os algoritmos de filtragem baseada em conteúdo preveem os possíveis interesses de um agente baseado nas características similares dos itens/interesses a serem recomendados, sem depender necessariamente de uma interação com um usuário. (PRIMO, 2013).

A base da similaridade dos itens nesta técnica é identificar as palavras-chave e a importância delas nos documentos ou textos analisados, uma forma de realizar isto é utilizando o modelo estatístico TF-IDF.

O TF-IDF é o produto da Frequência do Termo e da Frequência Inversa do Documento e tem como objetivo atribuir pesos às palavras-chave que aparecem muito em um documento, mas que não aparecem em poucos documentos criando assim associações. O cálculo do TD-IDF é realizado através do cálculo: TF-IDF = *Term Frequency* (TF) * *Inverse Document Frequency* (IDF). As equações 1 e 2 demonstram o cálculo do TF e do IDF separadamente.

$$tf = \frac{n_i}{\sum_k n_k} \tag{1}$$

Cálculo *Term Frequency*

$$idf = \log \frac{|D|}{|(d_i \supset t_i)|} \tag{2}$$

Cálculo de *Invert Document Frequency*

A filtragem baseada em conteúdo é uma técnica amplamente utilizada em sistemas de recomendação, como aqueles para livros ou filmes, onde as características dos itens (como gênero, autor ou diretor) são utilizadas para fazer recomendações personalizadas ao usuário (LOPS, 2011).

2.3.3 Filtragem híbrida

A filtragem híbrida busca combinar as vantagens da filtragem baseada em conteúdo e da filtragem colaborativa tornando possível diversos tipos de abordagens para a realização da combinação de métodos. Segundo (BARBOSA, 2014) existem quatro principais estratégias de combinação, sendo elas, ponderada, mista, combinação sequencial e comutação, já o autor (PRIMO, 2013) apresenta os métodos balanceado, permuta, mesclado, cascata entre outros.

Filtragem híbrida é uma técnica em sistemas de recomendação que combina as características dos itens (filtragem baseada em conteúdo) e as preferências dos usuários (filtragem colaborativa) para oferecer recomendações mais precisas e diversifica-

das (BURKE, 2002).

2.3.4 Regras de associação

Como pontua (VASCONCELOS, 2004) e (RICCI, 2011) os algoritmos de regras de associação tem como objetivo encontrar padrões frequentes de determinados elementos em um conjunto de dados de transações. Este tipo de algoritmo facilita as grandes bases de dados por serem rápidos e eficientes.

Por exemplo, dado um conjunto de dados de transações de uma loja de departamentos, é possível identificar os padrões de produtos destas transações. A regra de associação camiseta, calça -> cinto indica que o cliente que compra camisa e calça, com um determinado grau de certeza obtido através de métricas, também compra cinto.

A seção 2.3.4.1 apresenta as métricas utilizadas nesta técnica em maiores detalhes.

2.3.4.1 Métricas de associação

2.3.4.1.1 Support

Como explica (RICCI, 2011), o *support* indica a frequência em que um item aparece em um determinado conjunto de itens. Os seus valores estão em um intervalo de 0 a 1. Se for igual a 1 temos que o item aparece em todas as transações do *dataset*, se não, quanto menor o *support* menor é a ocorrência dele nos conjuntos de transações. A equação 3 mostra o cálculo para obtenção do *support*.

$$Support = \frac{number\ of\ transactions\ with\ items(s)}{number\ of\ transactions} \tag{3}$$

2.3.4.1.2 Confidence

O (RICCI, 2011) afirma que o *confidence* da regra é a frequência com que os itens em Y aparecem nas transações que contém X. Esta métrica indica a probabilidade de comprar o item Y dado a compra do item X. Seu intervalo vai de 0 à 1, se for igual a 1 temos que toda vez que um cliente compra um item Y ele também compra X, já se for 0 diz que X e Y não são comprados juntos. A equação 4 mostra o cálculo da métrica *confidence*.

$$Confidence = \frac{SupportXY}{SupportX} \tag{4}$$

2.3.4.1.3 Lift

É a probabilidade de levar os itens X e Y na mesma compra, seu intervalo vai de 0 ao infinito. Quando o *lift* é igual a 1 temos que X e Y são independentes e nenhuma

ligação é feita entre os dois quando, por exemplo, é comprado X. Quando *lift* é maior do que 1 quer dizer que X será provavelmente comprado junto com Y. A equação 5 mostra o cálculo da métrica *lift*.

$$Lift = \frac{Support(XY)}{Support(X)Support(Y)} \tag{5}$$

2.3.4.1.4 Leverage

A métrica *leverage* é similar ao *lift*, porém seu intervalo vai de -1 à 1. Quando comparamos o *leverage* e o *lift* podemos ter uma precisão de qual é o melhor item da lista de melhores itens. Por exemplo, há 5 pares de itens com *leverage* = 1, ou seja, a probabilidade do cliente levar cada item de cada par juntos é alta. Mas qual desses 5 pares têm a maior probabilidade de serem levados juntos? Como por exemplo, fronha e lençol, calça e camiseta, *notebook* e *mouse*, jogo de copo e jogo de talheres ou caderno e lápis? Para isso olhamos para o *lift* que vai de 0 ao infinito (ultrapassa o 1). A equação 6 mostra o cálculo da métrica *leverage*.

$$Leverage(X \rightarrow Y) = Support(XY) - Support(X)Support(Y) \tag{6}$$

2.3.4.1.5 Conviction

É a probabilidade de levar um outro item Z além dos itens X e Y na mesma compra. Seu intervalo vai de 0 ao infinito. Quanto maior o valor, maior é a probabilidade de levar o item Y junto com o item X e Y. Por exemplo, a probabilidade de levar X (Lençol), Y (Fronha), e Z (Travesseiro) são altas. A equação 7 mostra o cálculo da métrica *conviction*.

$$Conviction(X \rightarrow Y) = \frac{Support(X)Support(\overline{Y})}{Support(X\overline{Y})} \tag{7}$$

No presente capítulo, introduzimos o cenário da aplicação desenvolvida no contexto desta tese. Apresentamos, inicialmente, o conceito de análise de cestas de mercado, seguido pelos fundamentos dos Sistemas de Recomendação, englobando suas técnicas de filtragem colaborativa, baseada em conteúdo, híbrida e regras de associação. No próximo capítulo, aprofundaremos a revisão da literatura referente ao sistema em questão.

3 REVISÃO SISTEMÁTICA DA LITERATURA

A revisão sistemática da literatura (RSL) é um estudo secundário que utiliza uma série de procedimentos e técnicas em cima de artigos relacionados a um estudo primário com o objetivo de encontrar o estado da arte como explica o autor (DERMEVAL, 2020).

Para esta RSL foi utilizada a ferramenta parsifal para o cumprimento e organização de todos os procedimentos que estarão descritos nas próximas seções.

3.1 Pergunta de pesquisa

Conforme aponta (GRAZIOSI, 2011) *"Uma pergunta de pesquisa é a declaração de uma indagação específica que o pesquisador deseja responder para abordar o problema de pesquisa. A pergunta ou as perguntas de pesquisa orientam os tipos de dados a serem coletados e o tipo de estudo a ser desenvolvido."*

A pergunta de pesquisa que norteia este trabalho é: "Como estão sendo aplicados algoritmos de *frequent pattern mining* com análise de cesta de mercado para a recomendação de produtos?".

3.2 String de busca

As palavras chaves são essenciais para formar a *string* de busca que buscará os artigos a serem analisados na RSL. Abaixo podemos ver a listagem das palavras chaves escolhidas baseadas na pergunta de pesquisa e seus respectivos sinônimos.

1. *cross-selling* e *cross-sell*
2. *market* e *commerce*
3. *market basket analysis* e *mba*
4. *product recommendation* e *product recommendations*

A *string* de busca utiliza as palavras chaves e sinônimos em conjunto com operadores lógicos para filtrar a pesquisa por artigos. A *string* de busca visa encontrar artigos para responder a pergunta de pesquisa.

Neste trabalho foi elaborada a *string* em múltiplos níveis com a finalidade de encontrar no mínimo artigos relacionados a “recomendação de produtos e venda cruzada utilizando a técnica de *market basket analysis*” até “recomendação de produtos e venda cruzada utilizando a técnica de *market basket analysis* no comércio”.

A string de busca deste trabalho foi definida *“(("product recommendation"OR "product recommendations")) AND ("market basket analysis") AND ("cross-selling"OR "cross-sell")) OR (("product recommendation"OR "product recommendations") AND ("market basket analysis") AND ("cross-selling"OR "cross-sell") OR ("market"OR "commerce"))”*

3.3 Fontes de pesquisa

A busca por artigos foi feita nos repositórios listados abaixo.

1. ACM Digital Library - https://dl.acm.org/
2. Google Scholar - http://scholar.google.com.br
3. IEEE - https://ieeexplore.ieee.org/
4. ISI Web of Science - http://www.isiknowledge.com
5. Science@Direct - http://www.sciencedirect.com
6. SCOPUS - https://www.scopus.com
7. Springer Link - https://link.springer.com/

3.4 Critérios de seleção

Os critérios de seleção são divididos em dois tipos, critérios de inclusão e exclusão. Estes critérios são definidos para a aceitação ou rejeição de um artigo na RSL. Isto é feito por que quando aplicado a *string* de busca alguns artigos podem conter as palavras chaves mapeadas em seu texto mas não necessariamente abordam o tópico como assunto principal, as vezes as palavras são apenas mencionadas no texto.

3.4.1 Critérios de inclusão

1. Falam sobre recomendação de produtos e MBA
2. Falam sobre recomendação de produtos e MBA com *big data*

3.4.2 Critérios de exclusão

1. Não falam sobre recomendação de produtos e mba
2. Não foi possível acessar o artigo

3.5 Verificação de qualidade dos artigos

Neste passo, após a importação dos artigos nas fontes escolhidas com a *string* de busca aplicada e feita a seleção dos artigos é iniciado uma série de processos de verificação se os artigos são adequados aos critérios estabelecidos abaixo.

3.5.1 Perguntas

1. Foi possível acessar o artigo?
2. O acesso ao artigo é Gratuito?
3. O Artigo é Completo?
4. O Artigo é um Resumo?
5. O artigo tem h5?

3.5.2 Respostas

1. Sim
2. Não

3.6 Análise dos resultados

Foi aplicada a *string* de busca nas fontes de pesquisas presente na seção 3.3 e foram importados um total de 96 artigos entre os anos de 2017 e 2022. Para determinar se um artigo se encaixa em algum critério de exclusão ou inclusão foi lido primeiramente o título e depois o *abstract* e/ou introdução para atribuir a ele um dos critérios.

A tabela 1 apresenta os dados de importação dos artigos e a aplicação dos critérios de exclusão e inclusão.

Tabela 1 – Importação de arquivos e aplicação de critérios (continua)

Fontes	Importados	Rejeitados	Aceitos	Duplicados
ACM Digital Library	5	5	0	0
Google Schoolar	89	78	9	2

Tabela 1 – Importação de arquivos e aplicação de critérios (conclusão)

Fontes	Importados	Rejeitados	Aceitos	Duplicados
IEEE Digital Library	0	0	0	0
ISI Web Of Science	0	0	0	0
Science Direct	0	0	0	0
Scopus	0	0	0	0
Springer Link	2	0	0	2

A figura 3 mostra o percentual de artigos por fonte.

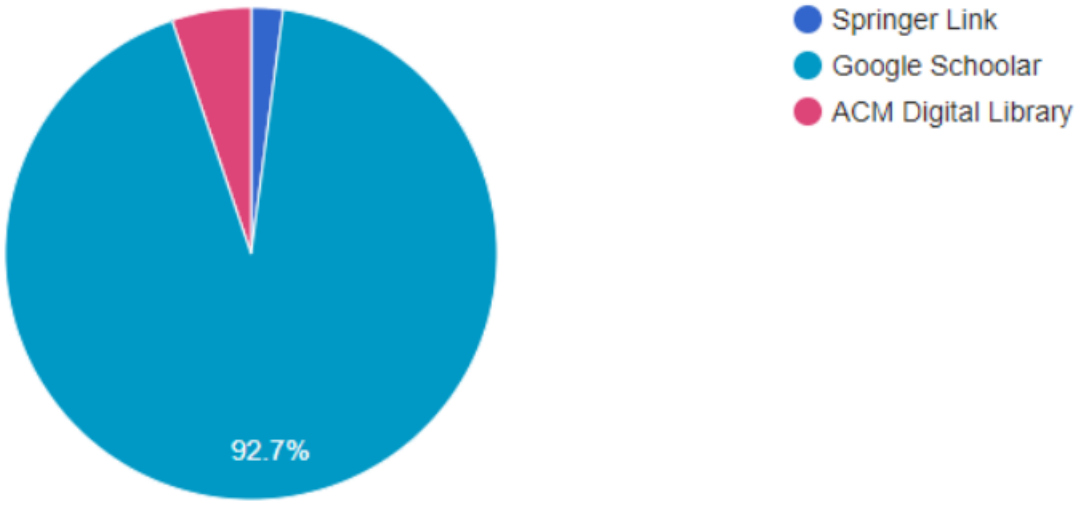

Figura 3 – Percentual de artigos por fonte

A tabela 2 apresenta a quantidade de artigos rejeitados, aceitos e duplicados por fonte e critério de seleção.

Tabela 2 – Artigos aceitos, rejeitados e duplicado

Fontes	Status	Critério de Seleção	Quantidade
ACM Digital Library	Rejeitado	Não falam sobre recomendação de produtos e MBA	5
Google Schoolar	Aceito	Falam sobre recomendação de produtos e MBA	9
Google Schoolar	Rejeitado	Falam sobre recomendação de produtos e MBA	1
Google Schoolar	Rejeitado	Não falam sobre recomendação de produtos e MBA	55
Google Schoolar	Rejeitado	Não foi possível acessar o artigo	14
Google Schoolar	Rejeitado	Não implementam o recomendador e a análise	8
Google Schoolar	Duplicado	Falam sobre recomendação de produtos e MBA com Big Data	1
Google Schoolar	Duplicado	Não falam sobre recomendação de produtos e MBA	1
Springer Link	Duplicado	Falam sobre recomendação de produtos e MBA	1
Springer Link	Duplicado	Não falam sobre recomendação de produtos e MBA	1

A figura 4 apresenta uma linha do tempo com a quantidade de artigos aceitos por ano de publicação.

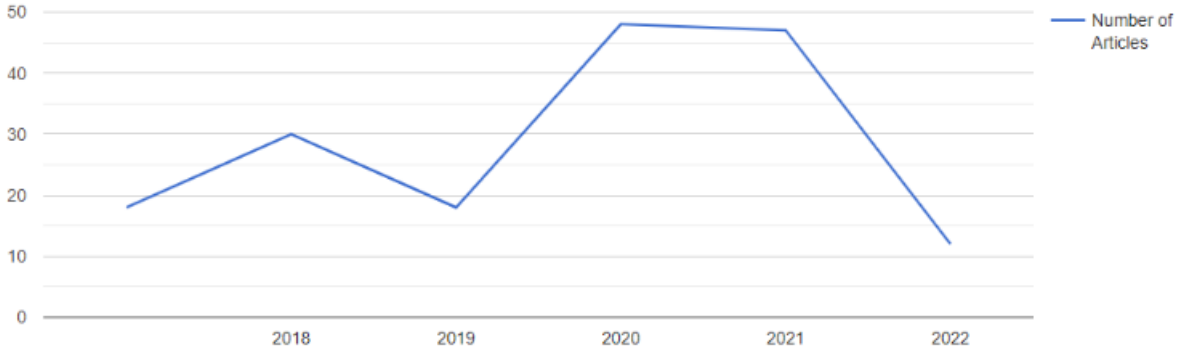

Figura 4 – Quantidade de artigos aceitos por ano de publicação

A tabela 3 apresenta os artigos aceitos e seus respectivos identificadores únicos (id).

Tabela 3 – Artigos selecionados

Id	Título
1	Basket market analysis using r-based apriori algorithm to find information from sales data
2	Customer segmentation and personalized marketing using k-means and apriori algorithm
3	Market basket analysis and recommendation system using association rules
4	A market basket analysis and high utility itemset minning in a retail company
5	Market basket analysis of basket data with demographics: a case study in e-retailing
6	The beauty or the beast inside retail stores? a market basket analysis of a cosmetic company
7	Unsupervised learning and market basket analysis in market segmentation
8	Using RFM model and market basket analysis for segmenting customers and assigning marketing strategies to resulted segments
9	Weighted based frequent and infrequent pattern mining model for real-time e-commerce databases

A tabela 4 apresenta os identificadores dos artigos aceitos e as respostas do questionário de qualidade apresentado na seção 3.5.

Tabela 4 – Identificadores e respostas do questionário de qualidade

Id	Foi possível acessar o artigo?	O artigo é gratuíto?	O artigo é completo?	O artigo é periódico?	O Artigo é um resumo?	O artigo tem índice h5?
1	Sim	Sim	Sim	Sim	Não	Sim
2	Sim	Sim	Sim	Não	Não	Sim
3	Sim	Sim	Sim	Não	Não	Não
4	Sim	Sim	Sim	Não	Não	Não
5	Sim	Sim	Sim	Sim	Não	Não
6	Sim	Sim	Sim	Não	Não	Não
7	Sim	Sim	Sim	Sim	Não	Não
8	Sim	Sim	Sim	Sim	Não	Não
9	Sim	Sim	Sim	Sim	Não	Não

Todos os artigos aceitos foram lidos em busca de encontrar as respostas para as seguintes perguntas:

1. Qual o objetivo do recomendador?
2. Qual algoritmo de recomendação foi utilizado?
3. Qual o método de validação dos resultados?
4. Qual o tamanho do *dataset*?
5. Quais resultados foram obtidos?

A tabela 5 mostra os objetivos e seus respectivos identificadores.

Tabela 5 – Objetivos e seus identificadores

Id Objetivo	Objetivo
1	Recomendar produtos para alavancar vendas e estratégias de marketing
2	Oferecer tratamento diferencial para o cliente
3	Estudo de CASO
4	Recomendação de produtos para clusters de cliente e afins
5	Entender padrões de compra

A tabela 6 mostra os ids dos artigos e as respostas para as perguntas apresentadas nesta seção.

Tabela 6 – Comparacao entre algoritmos

Id Artigo	Id Objetivo	Algoritmo	Quantidade de linhas do dataset	Resultados
1	1	Apriori	Aproximadamente 3450	Satisfatórios
2	1 e 4	Comparação entre apriori e uma extensão ECLAT	Aproximadamente 5000	Suficiente/Satisfatórios
3	1	Apriori e fpgrowth	Aproximadamente 3 milhões	Satisfatórios
4	1 e 2	Frequent itemsets mining (FIM)	Aproximadamente 966	Satisfatórios
5	1	Apriori	Aproximadamente 3160	Satisfatórios
6	3	Apriori	Aproximadamente 34	Satisfatórios
7	4	Apriori	Aproximadamente 541	Satisfatórios
8	4	Trabalho futuro	-	-
9	4	Apriori	Aproximadamente 119578	Satisfatórios

Conforme a pesquisa realizada os resultados encontrados levaram a escolha dos autores o algoritmo *Apriori* tratando de pouca volumetria de dados. Quando comparado ao cenário de *big data* e que uma parte deles ligam os projetos de clusterização de clientes, rfm e recomendação de produtos para potencializar os seus resultados.

Autores (MYTHILI, 2013) e (PATIL, 2022), mostram a diferença entres os algoritmos *Apriori* e *FP-Growth* que quando pensado em aplicar estes algoritmos ao cenários e *big data* é necessário atentar-se ao uso de memória e tempo de processamento.

O autor (MYTHILI, 2013) traz a tabela 7 que faz uma comparação entre os algoritmos.

Tabela 7 – Modelos

Parameters	Apriori	FP-Growth
Storage Strucuture	Array based	Tree based
Search Type	Breadth First Search	Divide and conquer
Technique	Join and Prune	Constructs conditional frequency pattern tree which satisfy minimum support
Number database scans	K+1 scans	2 scans
Memory utilization	Large memory (candidate generation)	Less memory (no candidate generation)
Database	Sparse/dense datasets	Large and medium datasets
Run time	More time	Less time

O autor (MYTHILI, 2013) traz também gráficos comparativos entre o tempo de execução dos dois algoritmos presentes nas figuras 5, 6, 7 e 8.

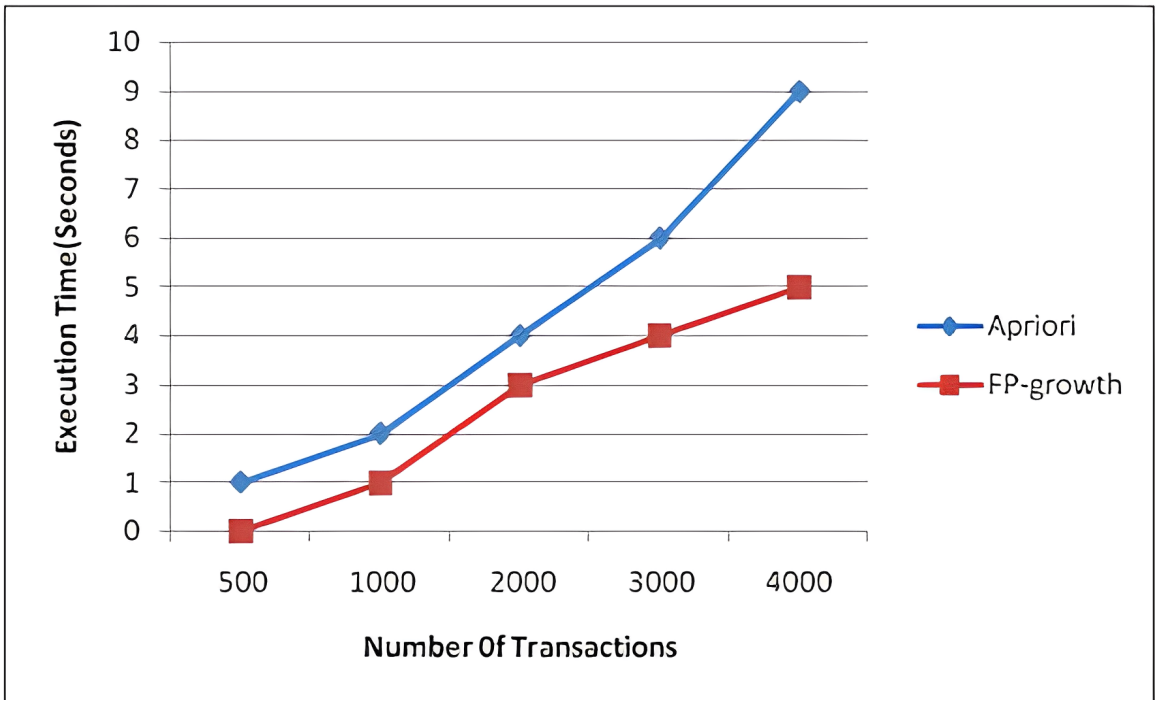

Figura 5 – Fonte: (MYTHILI, 2013). Tempo de execução com aumento de transações (vendas) utilizando support=0,1 com dados do dataset Mushroom

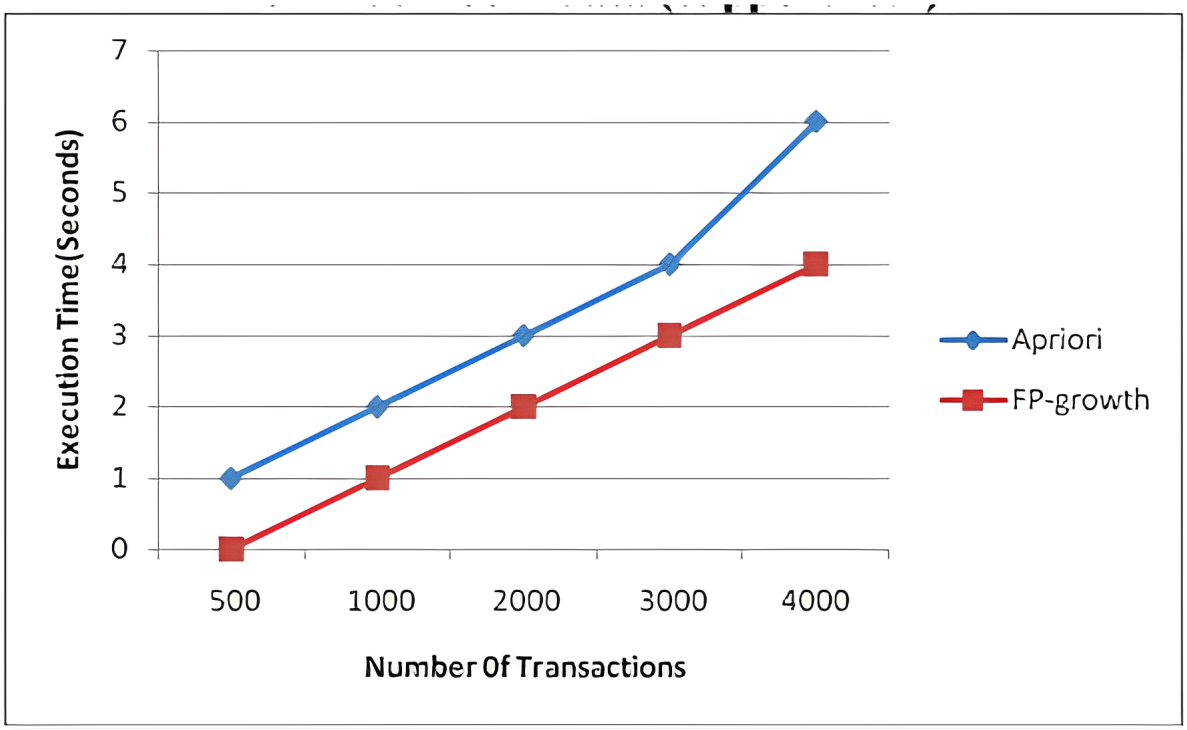

Figura 6 – Fonte: (MYTHILI, 2013). Tempo de execução com aumento de transações (vendas) utilizando support=0,5 com dados do dataset Mushroom

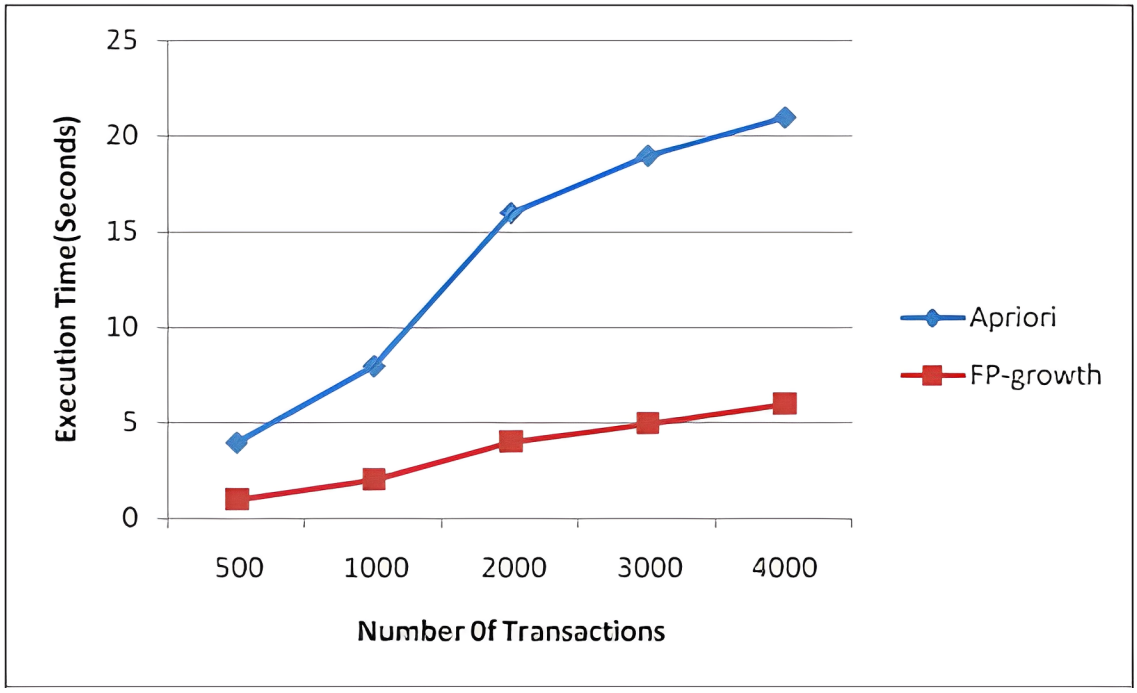

Figura 7 – Fonte: (MYTHILI, 2013). Tempo de execução com aumento de transações (vendas) utilizando support=0,1 com dados de supermercado

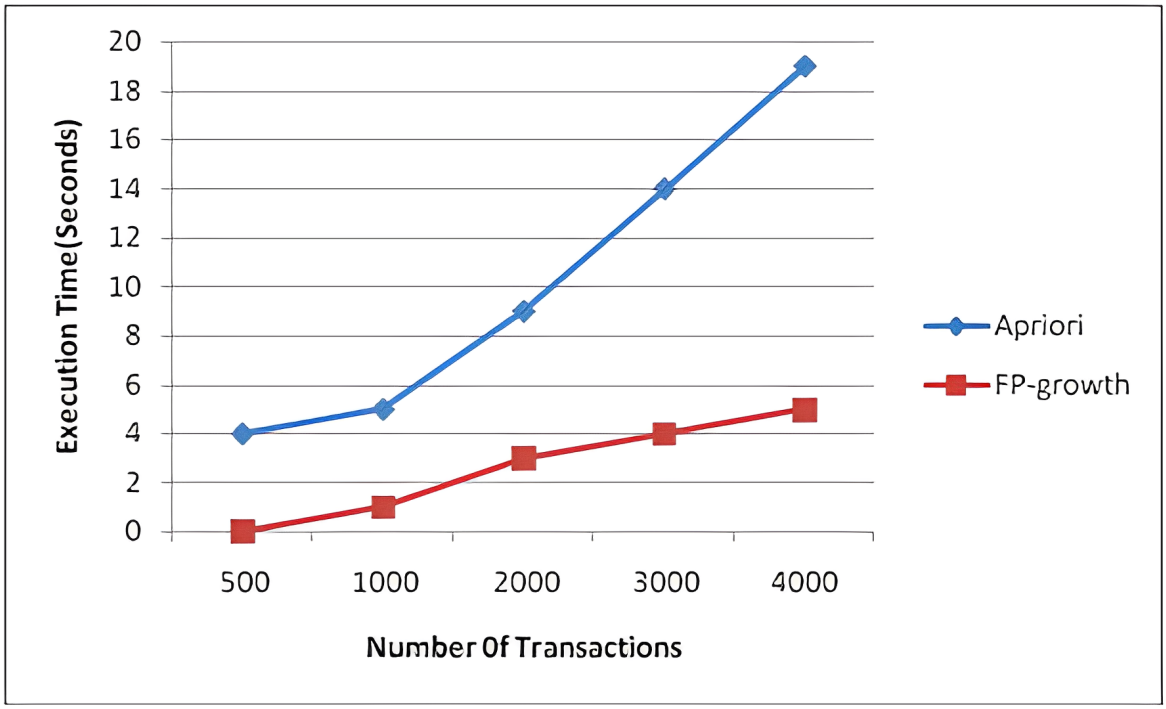

Figura 8 – Fonte:(MYTHILI, 2013). Tempo de execução com aumento de transações (vendas) utilizando support=0,5 com dados de supermercado

Além da afirmação dos autores, podemos observar na tabela de comparação dos algoritmos que há uma inviabilidade de aplicar o *Apriori* no cenário de *big data* por que ele não é escalável nem eficiente devido a quantidade de vezes que ele varre o *dataset* para obter seus resultados, ou seja, quanto maior o *dataset*, maior a quantidade de

vezes que o algoritmo irá realizar o *scan* e maior o número de registros a serem observados. Com isto, concluímos que não há ou há poucos recomendadores com regra de associação no cenário de *big data*.

Neste capítulo, apresentamos a Revisão Sistemática da Literatura desenvolvida para fundamentar esta tese e entender a área de aplicação em questão. Abordamos a formulação da pergunta de pesquisa, a elaboração da *string* de busca, as fontes de pesquisa consultadas, os critérios utilizados para a seleção dos artigos, o processo de verificação da qualidade desses artigos e, por fim, a análise dos resultados obtidos. Na próxima seção será apresentada a solução proposta.

4 SOLUÇÃO PROPOSTA

A presente proposta utiliza tecnologias e ferramentas específicas para grande volumetria de dados, ferramenta de orquestração de tarefas para melhor organização além de unir os ambientes *on-premise* e *cloud computing* para lidar com o processamento massivo dos dados.

4.1 Tecnologias Utilizadas

Para que seja possível manipular facilmente a grande volumetria de dados foi utilizada a linguagem de programação python, a ferramenta de processamento de dados massivos spark em sua versão para python chamada, pyspark, formato de dados parquet e tabelas no banco de dados postgres e como orquestrador e organizador de processos o Airflow.

4.1.1 Spark

Spark é uma ferramenta para processamento de dados em grande escala desenvolvida pela (APACHE, 2023a). Ela provê APIs (*Application Programming Interface*) nas linguagens de programação Java, Scala, Python e R. De acordo com a (AWS, 2023a) APIs são mecanismos que permitem a comunicação entre dois sistemas de computadores através de um conjunto de definições e protocolos.

Esta ferramenta atua distribuindo os dados de entrada em memória através de um *cluster* para processá-los paralelamente em diversas partes, isso permite um ganho significativo de velocidade de processamento e com isso ela torna-se extremamente útil para grandes volumes de dados.

4.1.2 FP-Growth

Com base nos resultados da revisão sistemática da literatura que serviu como fundamento para este trabalho, foi escolhido o algoritmo *FP-Growth* disponibilizado pela ferramenta Spark para realizar a obtenção dos resultados das regras de associação aplicadas ao contexto de *big data*.

O algoritmo *FP-Growth* é uma melhoria do *Apriori* e este algoritmo representa a *database*/*dataset* em formato de árvore chamado *Frequent Pattern Tree* ou *FP Tree* com o propósito de minerar o padrão mais frequente.

4.1.2.1 FP Tree

A árvore *Frequent Pattern Tree* (*FP Tree*) é uma representação compacta do *dataset* de transações (vendas), seu propósito é minerar o padrão mais frequente de itens do *dataset*. Como exemplifica (CHONNY, 2020), a *FP Tree* é construída mapeando cada conjunto de itens das vendas com o objetivo de identificar os itens que ocorrem com mais frequência para identificar quais outros itens têm grandes chances de serem compartilhados. A figura 9 exemplifica um mapeamento.

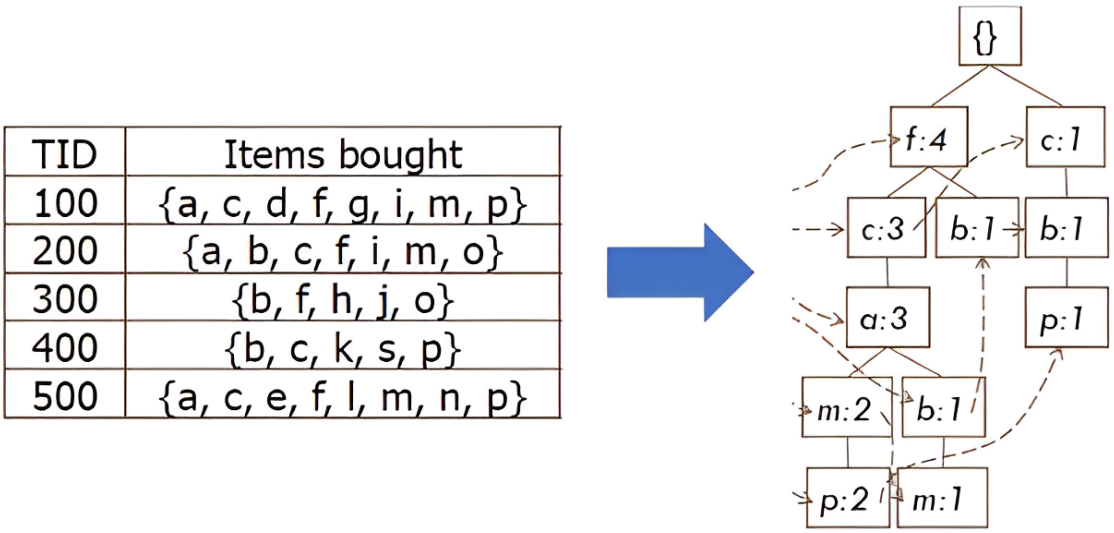

Figura 9 – Mapeamento FP Tree

4.1.2.2 Algoritmo

Dado o *dataset* de vendas para o algoritmo *FP-Growth* pode-se definir duas métricas de corte, sendo elas, *support* e *confidence* mínimos. Estas métricas são de extrema importância para o desempenho e resultados a serem retornados pelo algoritmo, isto porque, para cada transação de venda o algoritmo remove os itens que estão abaixo destas métricas filtrando e refinando parcialmente os resultados. Tendo as métricas de filtragem declaradas é possível executar o algoritmo e obter as métricas de associação dos conjuntos de itens do *dataset*.

4.1.2.3 FP-Growth vs Apriori

O *Apriori* é um método fundamental, já o *FP-Growth* é uma melhoria além de mais complexo com o poder de retornar seus resultados de uma maneira eficiente. A tabela 8 traz um comparativo entre os algoritmos.

Tabela 8 – Comparativo

Apriori	FP-Growth
Realiza pesquisa em largura	Realiza pesquisa em profundidade
É mais lento, o tempo aumenta exponencialmente conforme aumenta o número de conjuntos	É mais rápido, o tempo aumenta linearmente conforme aumentao número de conjuntos
Exige muita memória, todos os candidatos interligados são armazenados na memória	Exige pouca memória, armazena uma versão compactada da base
Realiza uma varredura da base de dados toda vez que vai encontrar o padrão frequente de um item	É necessário ler apenas duas vezes a base de dados

4.1.3 Simple Storage Service (S3)

O S3 é um serviço fornecido pela (AWS, 2023b) que tem a capacidade de armazenar e recuperar objetos com qualquer volume de dados sendo muito utilizado para a construção de *data lakes* devido sua escalabilidade, disponibilidade e segurança.

4.1.4 Parquet

Segundo (APACHE, 2023b), o parquet é um formato de armazenamento colunar que fornece esquemas eficientes de compactação e codificação de dados com desempenho aprimorado para lidar com dados complexos em massa.

4.1.5 PostgresSQL

Segundo a documentação oficial, PostgresSQL é um banco de dados relacional de código aberto que estende a linguagem SQL combinando diversos recursos que permitem com que o mecanismo armazene grandes quantidades de dados complexos de forma inteligente.

4.1.6 Airflow

Airflow é uma plataforma que permite o gerenciamento e orquestração de tarefas como apresentado por (APACHE, 2023c). A escolha desta plataforma permite o agendamento do processo da recomendação de ponta a ponta e também a organização de pequenas tarefas. O processo que contém o agendamento e a orquestração das tarefas é chamado de DAG (*Directed Acyclic Graph*)

Neste capítulo foi apresentada a solução proposta seguido das tecnologias utilizadas para o desenvolvimento do recomendador. No próximo capitulo serão apresentados os resultados da solução proposta.

4.2 Infraestrutura

Neste projeto, aproveitamos a infraestrutura já existente da empresa para executar tarefas em um *cluster* Spark on-premise. Esse ambiente consiste em 8 *workers*, cada um equipado com 80GB de memória e 10 cores. Além disso, exploramos as capacidades da AWS ao configurar um *cluster* EMR com características idênticas ao *cluster on-premise*. Adicionalmente, visando abranger uma variedade de cenários, incorporamos um computador com configuração de *cluster* de nó único, contendo 10 cores e 60GB de memória. Para uma análise comparativa, também incluímos um *cluster standalone on-premise* composto por 1 *worker*, notável por suas 120GB de memória e 20 cores. Essa abordagem diversificada permitiu uma análise e comparativa dos erros e resultados obtidos em diferentes nos ambientes de execução.

4.3 Limpeza e organização do dataset

Esta etapa tem como objetivo analisar e investigar conjuntos de dados para entender o padrão e comportamento dos mesmos, além de detectar possíveis anomalias, como aponta (IBM, 2023b).

A empresa na qual foi desenvolvido este trabalho possui uma grande velocidade no recebimento dos dados, variedade e volumetria, definido como big data por (ORACLE, 2023), por isso, serão utilizadas somente informações de venda do ano de 2022 da matriz da grande rede de lojas de departamentos.

No conjunto de dados de vendas, foram aplicados diversos filtros para refinar a análise. Excluímos produto que não são promoções, que não foram cancelados posteriormente a venda desfazendo assim a intenção de compra e sua relação com os demais itens e não são kits, por exemplo, kit de ar condicionado, onde a unidade condensadora e a unidade evaporadora são dois itens distintos com seu próprio código, porém, não podem ser vendidos separadamente ao cliente, ou seja, compra é sempre feita conjunta. A minimização de produtos de kits e promoções é de extrema importância pois, a nível de cliente, por exemplo, o produto A e B são vendidos juntos sem a possibilidade de serem adquiridos separadamente, porém a nível de sistema, são dois produtos diferentes com códigos distintos.

Após estes filtros as informações foram agregadas por mês formando 12 *datasets* para que assim os recursos de *hardware* e o algoritmo conseguissem lidar de forma saudável com a volumetria de dados.

Além desta divisão dos *datasets*, para melhores resultados e organização os dados foram divididos em dois grupos sendo eles geral e sazonais. Estas condições foram postas para que determinados produtos que só aparecem em um período específico do ano e com muita ênfase não tenham relevância na recomendação geral ou estes mesmo produtos sobreponham outros períodos sazonais. Neste trabalho, será abordada a recomendação geral.

Neste capítulo, abordamos as tecnologias empregadas, ferramentas utilizadas, infraestrutura necessária e o processo de limpeza e organização do dataset, elementos cruciais para o desenvolvimento deste trabalho. Na próxima seção, serão delineados os desafios enfrentados e as estratégias adotadas para a resolução destes obstáculos.

5 DESAFIOS E ESTRATÉGIAS

Neste capitulo serão apresentados os desafios e estratégias de resolução aplicados neste trabalho.

5.1 Sazonalidade

A sazonalidade é um aspecto crucial no varejo, influenciando os padrões de compra ao longo do ano. O algoritmo é projetado para ser executado mensalmente, abordando desafios como treinamento exaustivo, processamento massivo de dados e variações sazonais. Por exemplo, considerando a matriz da grande loja de departamentos no sul do Brasil, observa-se que nos meses de dezembro a fevereiro, há uma inclinação para a compra de itens de verão, enquanto junho a agosto vê uma demanda por produtos de inverno. Esta observação é corroborada por estudos que mostram como o clima e a localização geográfica influenciam os padrões de compra (JOHNSON, 2015).

5.2 Volumetria de dados

O processamento de grandes volumes de dados é uma tarefa desafiadora. (BAGUI, 2020). destacam que, em certos cenários, o algoritmo FP-Growth pode não ser executado eficientemente em sistemas distribuídos, mesmo com ferramentas como Spark (SIKHA, 2018). (HAOSONG, 2019). reiteram essa preocupação, apontando problemas de escalabilidade, especialmente quando o conjunto de dados é vasto e o suporte mínimo é definido como um valor baixo (próximo de zero).

5.3 Distribuição de execução

O algoritmo *fpgrowth* foi testado nos 3 ambientes diferentes presentes na seção 4.2.

No primeiro teste foi utilizado o *cluster* spark *on-premise* com 8 *workers*. Para a execução do algoritmo foi disparado o *job* spark distribuindo seus dados nos 8 *workers*

e diminuindo conforme os erros de execução surgiam fornecidos pelo spark. Na ocorrência dos erros, os *workers* que estavam executando o *job* morriam e após diversas tentativas o *job* falhava.

No segundo teste foi utilizado o *cluster* EMR na AWS [1] *on-premise* na intenção de validar se não havia algum problema no *setup* do ambiente *on-premise*. Foi executado o mesmo procedimento nos mesmos cenários e o comportamento do erro persistiu.

Foi observado que o algoritmo não executava nos *clusters* mencionados acima porém na máquina local a execução era completada e apenas o que divergia era a quantidade de dados. Com isso, foi realizado um terceiro teste em no *cluster* standalone onde o algoritmo executou o processo completo com sucesso.

A característica inerente do algoritmo *FP-Growth*, que se fundamenta na estrutura de árvores ou grafos para relacionar itens de cestas distintas, pode ser identificada como a causa subjacente dos problemas observados nos clusters distribuídos. A figura 10 ilustra visualmente a teoria proposta, destacando a relação entre as árvores ou grafos gerados pelo algoritmo *FP-Growth*.

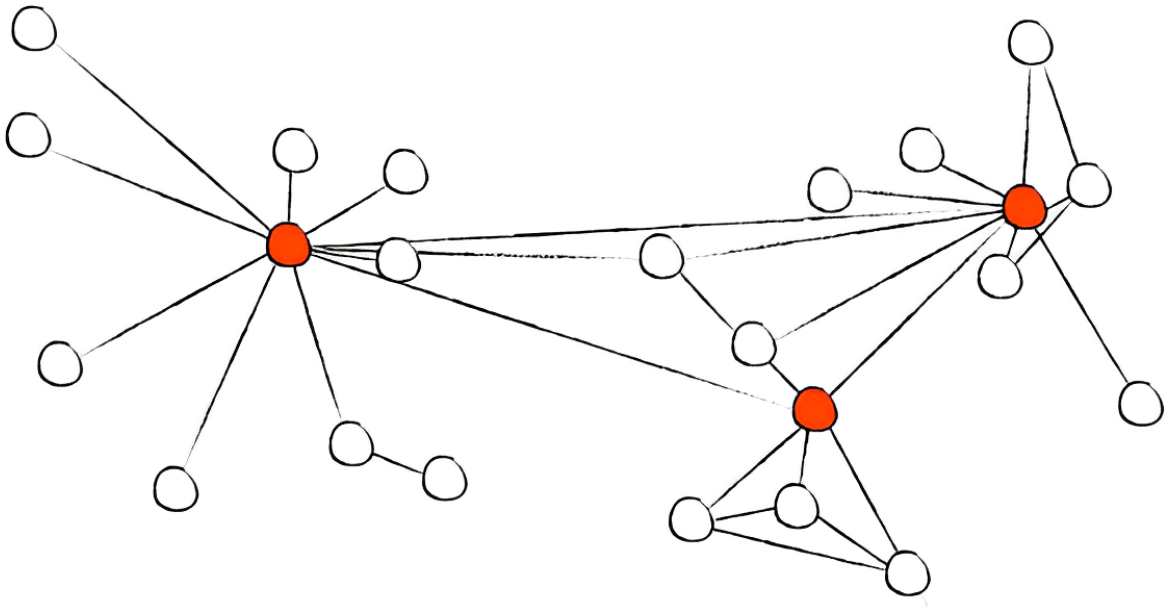

Figura 10 – Demonstração ligação de árvores/grafos

Ao distribuir o processo do *FP-Growth* em várias máquinas ou *workers*, observa-se que as árvores resultantes tornam-se separadas, perdendo a conexão entre si e na hora de conectar os resultados dos *workers* e um só esta conexão não é encontrada e é por isso que após um determinado tempo de execução o algoritmo falha e as máquinas do *cluster* spark morrem o que já não acontece quando se trata de um *cluster* de nó único.

Neste capítulo foram abordados os desafios enfrentados e a estratégia de resolu-

[1] https://aws.amazon.com/pt/emr/ Serviço oferecido pela AWS para processamento distribuído

ção dos mesmos. Na próxima seção, será apresentado o sistema de recomendação desenvolvido.

6 SOLUÇÃO PROPOSTA

Neste capítulo será apresentada o sistema de recomendação desenvolvido.

6.1 Recomendação

A revisão sistemática da literatura, juntamente com pesquisas e estudos práticos, informou a escolha do algoritmo *FP-Growth* e *MBA* para este trabalho. Utilizando o Airflow, foi possível automatizar o processo de recomendação de produtos, desde a coleta de dados até a implementação das recomendações em diferentes plataformas (APACHE, 2023c).

A análise feita sobre revisão sistemática da literatura complementar a este trabalho, pesquisas e alguns estudos práticos permitiram a assertividade na escolha do algoritmo e análise utilizadas neste trabalho, sendo eles, *FP-Growth* e *MBA*. Com o auxílio da ferramenta airflow foi possível organizar e automatizar o processo da recomendação de produtos de ponta a ponta para a dag geral.

O processo inicia-se no envio dos dados para um banco de dados centralizado contendo as transações do aplicativo, site e cada um dos pdvs de cada filial da rede de lojas. A cada dia primeiro do mês é realizada a execução do processo de recomendação que é disparado pelo Airflow utilizando a ferramenta de processamento massivo Spark e o S3 como *data lake* para que no fim do processo as sugestões de recomendação de produtos sejam salvas em um *data warehouse* Postgres para serem utilizadas no *site*, aplicativo e lojas físicas. A figura 11 ilustra o processo.

6.2 Dag Geral

Os grupos e seus respectivos subgrupos comentados na seção “Limpeza e organização do dataset” foram organizados na ferramenta Airflow onde cada grupo corresponde a uma DAG e cada subgrupo é um *task group* definido por (ASTRONOMER, 2023) como organizadores de tarefas.

A dag de itens gerais é organizada em grupos de tarefas referente às filiais/lojas da grande rede de departamentos, esta divisão é feita para que seja gerada a recomen-

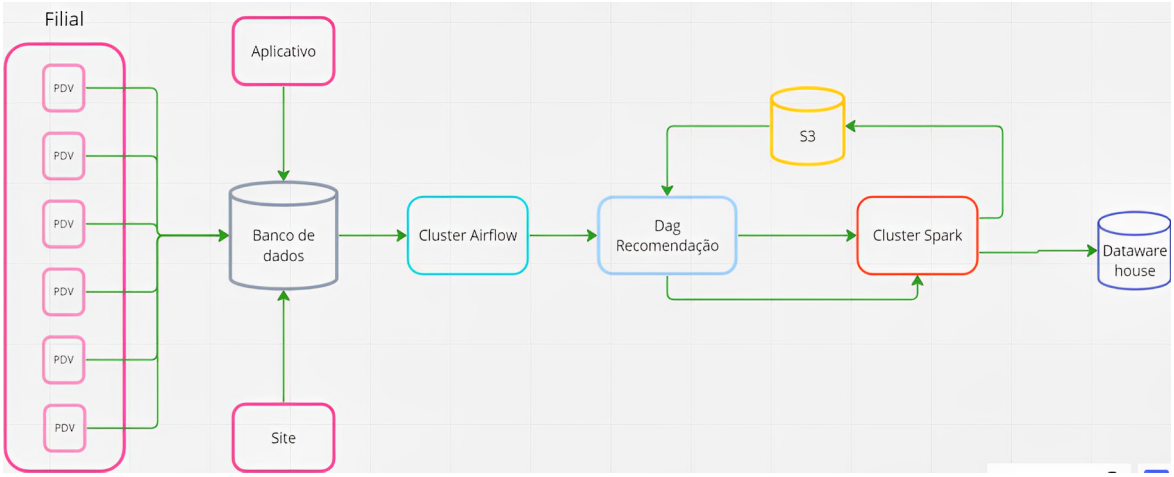

Figura 11 – Processo completo

dação de produtos para uma loja a qual é representante de cada região do país como apresentado na seção 4.3. Cada grupo corresponde a uma DAG e cada subgrupo é um *task group* definido por (ASTRONOMER, 2023) como organizadores de tarefas.

Cada região tem comportamentos diferentes devido ao clima e costumes, por exemplo, clientes do sul do brasil tendem a comprar roupas de inverno e cobertores mais pesados do que os clientes do norte do país. As figuras 12 e 13 ilustra a dag geral.

6.2.1 Grupo de tarefas vendas

Utilizando o agregador de tarefas, organizou-se o processo de recomendação dos produtos utilizando a base de vendas (transações) para cada filial referência. Brevemente, as tarefas consistem em extrair os dados, gerar as métricas de associação para recomendação de produtos, transformar os dados retornados do modelo e criar a tabela de recomendações final no banco de dados Postgres com as top 15 recomendações. As próximas seções consistem em um detalhamento de cada tarefa do grupo de vendas.

6.2.1.1 Extração dos dados

A primeira tarefa consiste na obtenção do *dataset* com os dados do mês referência de vendas, filtrados e agrupados como comentado na seção 4.3. Como os grupos de tarefas têm a mesma estrutura, o que difere-as são os filtros de filial.

Neste procedimento é utilizada inteiramente a ferramenta Spark o *dataset* é salvo em formato Parquet na plataforma para armazenamento de objetos de alta disponibilidade S3 (*Amazon Simple Storage Service*).

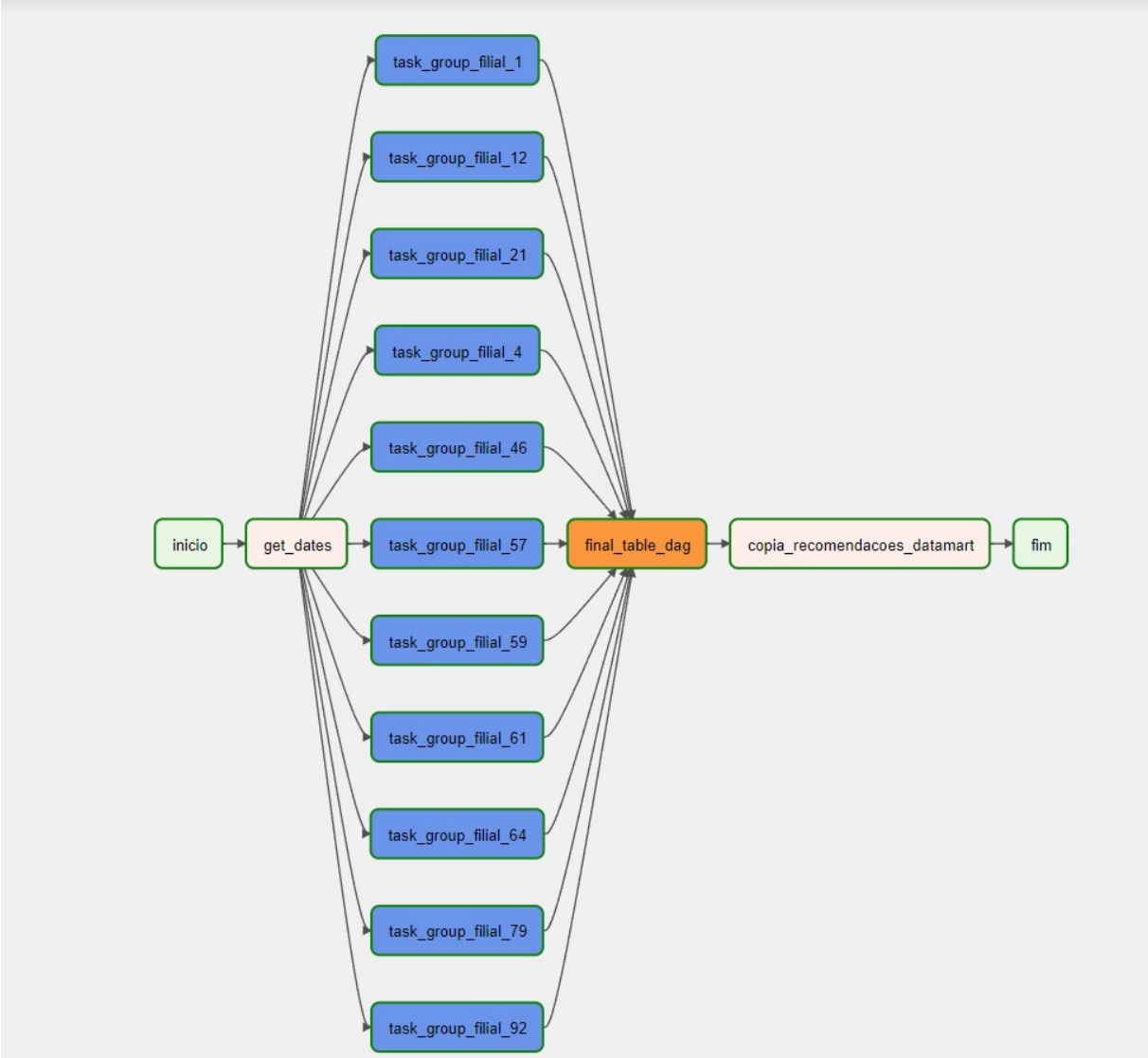

Figura 12 – Demonstração fluxo da DAG

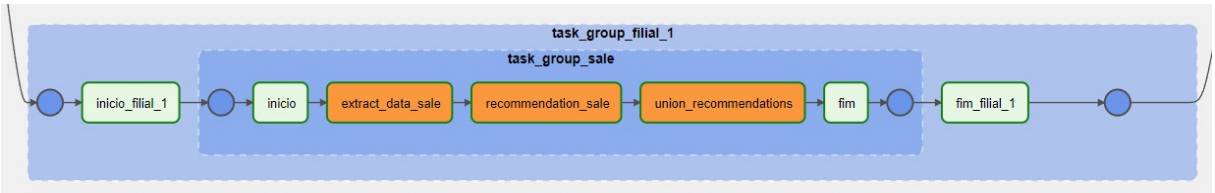

Figura 13 – Demonstração fluxo do grupo de tarefas

6.2.1.2 Recomendação de produtos

A segunda tarefa chamada de "*recommendation_sale*" utiliza como entrada os *dataset* da filial e do mês referência disponibilizados no AWS S3 pela tarefa anterior explicada na seção 6.2.1.1. Para cada *dataset* é executado o algoritmo *FP-Growth* disponibilizado pela ferramenta Apache Spark e são aplicadas estratégias de tuning de modelo e *tuning* de hiperparâmetros. O código abaixo mostra o código do algoritmo executado.

```
fp = FPGrowth(minSupport=min_support, minConfidence=min_confidence,
 itemsCol='BASKET', predictionCol='prediction')

model = fp.fit(df_transactions_model)
df_final = model.associationRules
```

6.2.1.3 *Otimização de Hiperparâmetros*

A escolha dos hiperparâmetros *minSupport* e *minConfidence* do modelo são dadas de acordo com mês referência, filial e a quantidade de itens máximos que existe nas transações da base em questão. Estas configurações foram definidas pois sempre que há uma maior quantidade de itens na cesta do cliente maior é a quantidade de associações retornada e faz-se necessário o corte de *min_support* e *min_confidence* com um número maior para evitar sobrecarga de recursos de hardware conseguindo assim rodar o modelo e obter os resultados. O código abaixo demonstra a lógica da escolha dos parâmetros.

```
# Definindo valor dos hiperparametros do modelo
min_confidence = 0.0001 if 50 >= num_max_cesta else 0.001
min_support = 0.0001 if 50 >= num_max_cesta else 0.001
```

6.2.1.4 *Tunning do modelo*

O *tuning* do modelo é feito através do número máximo de itens na cesta em um *loop* para a filial e o mês referência em questão. A código abaixo mostra os valores de início e fim do *loop*.

```
# Definindo numero maximo de itens na cesta:
num_max_cesta = 100 if 100 <= max_item else max_item
num_inicio = floor(num_max_cesta/2)
num_inicio = num_inicio if num_inicio > 2 else 3
num_max_cesta = (num_max_cesta + 2) if num_max_cesta <= num_inicio
 else num_max_cesta
```

Dentro deste *loop* é executado o algoritmo do *fpgrowth* aplicando uma estratégia de poda aleatória na quantidade de itens das transações. Por exemplo, dado uma transação com 200 itens, o valor atual iterado pelo for de itens máximos é 70 então serão filtradas todas as transações que contém 70 ou menos itens e este dataset filtrado será a entrada de dados do modelo.

A cada iteração do *loop* é contada a quantidade de associações retornadas para este valor de poda e o modelo que retornar a maior quantidade de associações será o resultado final e futuras recomendações. A figura 14 mostra o formato da saída dos dados.

antecedent	consequent	confidence	lift	support
{18811,435133,484361}	{484830}	1	922.5555555555555	0.0001204383957605684
{18811,435133,484361}	{464935}	1	2767.6666666666665	0.0001204383957605684
{18811,435133,484361}	{364004}	1	87.4	0.0001204383957605684
{18811,435133,484361}	{466487}	1	30.52573529411765	0.0001204383957605684
{18811,435133,484361}	{200295}	1	58.88652482269504	0.0001204383957605684
{18811,435133,484361}	{36627}	1	259.46875	0.0001204383957605684
{485508,432978,76198}	{363018}	1	2767.6666666666665	0.0001204383957605684

Figura 14 – Formato da saída dos dados

6.2.1.5 União das recomendações

Esta tarefa é responsável por ler os *datasets* armazenados no S3 da AWS, gerados pela tarefa anterior Recomendação de produtos, unificá-los em um *dataset*, adicionar as colunas mês e ano referente às recomendações, filtrar somente registros com a coluna *consequent* de tamanho 1 e salvar a tabela previamente tratada em parquet no Amazon S3 novamente. A figura 15 mostra os dados de saída da tarefa em questão.

antecedent	consequent	confidence	lift	support	mes	ano
{510174}	{439697}	0.3333333333333333	2599.3333333333335	0.00012823800974608873	2	2023
{510174}	{439677}	0.3333333333333333	1299.6666666666667	0.00012823800974608873	2	2023
{510174}	{439685}	0.3333333333333333	1299.6666666666667	0.00012823800974608873	2	2023
{510174}	{495143}	0.3333333333333333	1299.6666666666667	0.00012823800974608873	2	2023
{496120}	{521248}	0.5	354.45454545454544	0.00012823800974608873	2	2023
{496120}	{412484}	0.5	205.21052631578948	0.00012823800974608873	2	2023
{496120}	{535650}	0.5	278.5	0.00012823800974608873	2	2023
{496120}	{455193}	0.5	1299.6666666666665	0.00012823800974608873	2	2023
{496120}	{532983}	0.5	299.9230769230769	0.00012823800974608873	2	2023
{498158}	{498131}	0.5	3899.0000000000005	0.00012823800974608873	2	2023
{498158}	{536950}	0.5	779.8000000000001	0.00012823800974608873	2	2023
{498158}	{536549}	0.5	243.68750000000003	0.00012823800974608873	2	2023
{498158}	{497980}	0.5	324.91666666666663	0.00012823800974608873	2	2023
{498158}	{490052}	0.5	557	0.00012823800974608873	2	2023

Figura 15 – Formato da saída dos dados da tarefa união das recomendações

6.2.1.6 Tabela final

Esta tarefa é responsável por ler as tabelas previamente tratadas que contém as recomendações 1:1 (1 *antecedent* para 1 *consequent*) de todas as filiais, elencar as top 15 recomendações e normalizar a nomenclatura das colunas.

Na etapa de top recomendações, é elencado os 15 produtos antecedentes e consequentes com o maior valor de *confidence* e *lift* separadamente, tendo esse *ranking* de 15 produtos é feito a união dos resultados e um último *ranking* em cima dessa união gerando os resultados finais. Na parte de normalização é ocorre a substituição dos termos técnicos das colunas por nomenclaturas mais acessíveis a usuários de negócio. Os nomes das colunas passam de *confidence* e *lift* para “rankingmesmacompra” e ”rankingcompraposterior”. Essa abordagem visa facilitar a compreensão do propósito de cada coluna de maneira mais intuitiva.

A figura 16 mostra a tabela final.

id	idprodutoorigem	idprodutorecomendado	codigofilial	me... ▲ 1	rankingmesmacompra	rankingcompraposterior
3461	474246	361983	21	10/2022	15	13
3533	445298	172358	21	10/2022	2	1
3560	445298	531342	21	10/2022	13	12
3626	483330	320540	21	10/2022	5	5
3669	483538	530359	21	10/2022	3	2
3670	483538	496051	21	10/2022	4	3
3675	483538	464001	21	10/2022	9	8
3679	483538	385755	21	10/2022	12	12
3718	22950	389197	21	10/2022	7	7
4264	488623	412537	21	10/2022	7	7
4186	487680	524838	21	10/2022	4	4
4295	489590	495977	21	10/2022	9	5
4318	489590	442473	21	10/2022	14	10
4083	119137	388592	21	10/2022	8	6
4095	119137	494251	21	10/2022	16	15
4212	165626	518687	21	10/2022	5	5
4248	182027	513431	21	10/2022	2	2
4266	182027	393749	21	10/2022	8	8
4278	182027	477678	21	10/2022	12	12

Figura 16 – Formato da saída dos dados na tabela final

Neste capítulo foi apresentado o sistema de recomendação desenvolvido, destacando seu fluxo de tarefas e fornecendo um entendimento detalhado das funcionalidades de cada uma delas. Na próxima seção, abordaremos as discussões acerca do desempenho do recomendador, bem como os resultados obtidos, acompanhados das análises correspondentes.

7 DISCUSSÃO

Neste capítulo, serão analisados os resultados do recomendador por meio de uma análise que inclui a avaliação por produto, setor, estações do ano e os produtos mais vendidos em cada região. A seleção específica dos produtos mencionados na seção 7.2 baseou-se em sua popularidade e presença comum na grande maioria dos lares. Além disso, optou-se por realizar uma análise sazonal, abordando as estações de verão e inverno, nas regiões sul e norte, aproveitando suas marcantes diferenças climáticas.

7.1 Performance do Modelo

A partir da análise da DAG na ferramenta Airflow, conseguimos extrair os tempos médios de execução para tarefas como extração de dados, recomendação, fusão de recomendações e criação da tabela final.

O tempo médio geral para as 11 filiais está detalhado na Tabela 9. Nota-se variações significativas em alguns meses, indicando picos no tempo médio de execução das tarefas, conforme evidenciado na Tabela 9. Essas variações podem ser atribuídas a fatores como concorrência no ambiente *on-premise* da grande rede varejista, resultando tanto em aumentos quanto em reduções no tempo de execução.

Ano	Mês	Extração	Recomendação	Transformações
2023	4	01:29	04:40	01:24
2023	5	00:29	01:46	04:22
2023	6	01:18	05:20	06:00
2023	7	00:44	01:26	02:27
2023	8	00:29	02:50	03:33
2023	9	00:32	02:06	04:10
2023	10	00:33	02:26	04:43

Tabela 9 – Tempo médio em horas de execução das tarefas por mês no ano de 2023

Além disso, percebeu-se que quando o algoritmo *FP-Growth* lida com conjuntos de dados bem reduzidos, especialmente junto com um valor de suporte mínimo próximo de zero, há o risco de o algoritmo entrar em um ciclo infinito, gerando uma sobrecarga.

Nesse cenário, é essencial buscar um equilíbrio apropriado no tamanho do conjunto de dados. É aconselhável evitar conjuntos excessivamente pequenos, sem, contudo, optar por conjuntos muito extensos. Vale ressaltar a importância de realizar testes para definir os valores ideais de suporte mínimo e confiança mínima nos parâmetros do modelo, garantindo, assim, um desempenho eficaz do algoritmo.

Foi conduzida também uma comparação entre os pares de produtos recomendados no mês anterior (dezembro/2023) e os produtos efetivamente vendidos no mês atual (janeiro/2024), agrupados por venda e cliente. Os resultados indicaram uma estimativa de precisão 22,53% para o canal digital que consiste no aplicativo e *site*. Já para as lojas físicas foi calculada a estimativa de precisão para a loja matriz situada na região sul do país resultando em 31,28%

7.2 Análise por Produto

Na Figura 17, apresentamos as recomendações para o item "Smart TV LED 32 Polegadas". O item com a maior correlação é o suporte de parede universal, seguido pela antena digital. Esses produtos estão fortemente relacionados a qualquer cliente que adquire uma Smart TV. A métrica *confidence* indica a probabilidade de comprar o suporte universal junto com a Smart TV, enquanto o *lift* sugere que a compra subsequente mais indicada após a Smart TV seria a antena interna. Além disso, observamos a relação dos produtos de eletrodomésticos com itens dos setores de decoração e utilidades domésticas.

Antecedente	Consequente	Lift	Confidence
SMART TV LED 32" PTV32D10N5SKH HD PHILCO	SUPORTE UNIVERSAL UNI100 14" 84" ELG	86,07	0,37
SMART TV LED 32" PTV32D10N5SKH HD PHILCO	ANTENA INTERNA DIGITAL AI2021 INTELBRAS	512,33	0,15
SMART TV LED 32" PTV32D10N5SKH HD PHILCO	ASSADEIRA RETANGULAR LIBBEY 3,2L	81,67	0,14
SMART TV LED 32" PTV32D10N5SKH HD PHILCO	CORTINA RUSTICA VENEZA ESTAMPADA 260X170 HAVAN	87,11	0,14
SMART TV LED 32" PTV32D10N5SKH HD PHILCO	PAPEL DE PAREDE 9,5M X 53CM HAVAN	43,56	0,14
SMART TV LED 32" PTV32D10N5SKH HD PHILCO	POTE VIDRO HERMÉTICO RETANGULAR 1,5L	130,67	0,14
SMART TV LED 32" PTV32D10N5SKH HD PHILCO	SUPORTE INCLINAVEL SBRUB910 10" 85" BRASFORMA	326,68	0,14
SMART TV LED 32" PTV32D10N5SKH HD PHILCO	SUPORTE TRI-ARTICULADO 1040 10" 55" BRASFORMA	1.306,71	0,14
SMART TV LED 32" PTV32D10N5SKH HD PHILCO	TRAVESSEIRO 50X70 SUPORTE FIRME MAQUINETADO	46,67	0,14

Figura 17 – Smart TV LED 32 Polegadas x Recomendações

Para o produto "Faqueiro de 42 Peças", os itens recomendados pertencem aos setores de utilidades domésticas e eletrodomésticos. Os produtos consequentes indicados pelo algoritmo e a análise da cesta de mercado possuem a mesma métrica de *confidence*, exceto para a caneca de porcelana, sugerindo que esses itens, apesar de serem de setores distintos, possuem a mesma correlação com o antecedente em análise. A Figura 18 ilustra essa análise.

Antecedente	Consequente	Lift	Confidence
FAQUEIRO LAGUNA TRAMONTINA 42PC	CANECA PORCELANA 350ML	4.378,00	1,00
FAQUEIRO LAGUNA TRAMONTINA 42PC	BATEDEIRA PLANETARIA BBP760VM 700W BRITANIA	135,17	0,17
FAQUEIRO LAGUNA TRAMONTINA 42PC	CAIXA ORGANIZADORA M 5L	231,71	0,17
FAQUEIRO LAGUNA TRAMONTINA 42PC	FRITADEIRA AIR FRY INOX BFR11PI PT BRITANIA	49,15	0,17
FAQUEIRO LAGUNA TRAMONTINA 42PC	GARRAFA RETRO 500ML	540,67	0,17
FAQUEIRO LAGUNA TRAMONTINA 42PC	GRILL PRESS DIAMANTE PGR07P PHILCO	811,00	0,17
FAQUEIRO LAGUNA TRAMONTINA 42PC	JOGO FACAS INOX PLENUS TRAMONTINA 9PC	14,07	0,17
FAQUEIRO LAGUNA TRAMONTINA 42PC	JOGO JANTAR BIONA COLB 20PC	405,50	0,17
FAQUEIRO LAGUNA TRAMONTINA 42PC	JOGO PANELAS LIFE EASY 5PC	270,33	0,17
FAQUEIRO LAGUNA TRAMONTINA 42PC	LÂMPADA LED TKL60 9W 6500K TASCHIBRA	11,75	0,17
FAQUEIRO LAGUNA TRAMONTINA 42PC	PASSADEIRA OVAL ROMANO 40X120	1.622,00	0,17
FAQUEIRO LAGUNA TRAMONTINA 42PC	TAPETE FINE 60X100	811,00	0,17
FAQUEIRO LAGUNA TRAMONTINA 42PC	TORRADEIRA OTOR500 INOX DAY LIGHT OSTER	811,00	0,17

Figura 18 – Faqueiro de 42 Peças x Recomendações

7.3 Análise de Setores

Nesta seção, exploramos a análise detalhada da quantidade de recomendações com base nos setores da loja, revelando *insights* sobre as relações entre diferentes categorias de produtos. Observamos que, em sua maioria, as recomendações concentram-se em produtos do mesmo setor, indicando uma preferência consistente dos consumidores. No entanto, também identificamos correlações interessantes entre setores distintos. As Figuras 19, 20, 21, e 22 visualizam essas relações de maneira gráfica e informativa.

Setor Antecendente	Setor Consequente	Qtde Recomendações
DECORAÇÃO	DECORAÇÃO	1.290
DECORAÇÃO	CAMA MESA E BANHO	477
DECORAÇÃO	UTILIDADES DOMESTICAS	450
DECORAÇÃO	TAPETES	256
DECORAÇÃO	ELETRO-PORTÁTIL	229
DECORAÇÃO	FERRAMENTAS	218
DECORAÇÃO	BRINQUEDOS	161
DECORAÇÃO	CORTINAS	152
DECORAÇÃO	LINGERIE	141
DECORAÇÃO	CONF FEMININA	99
DECORAÇÃO	ESPORTIVO	49
DECORAÇÃO	CONF INFANTO	48
DECORAÇÃO	CONF MASCULINA	44
DECORAÇÃO	BEBE	40
DECORAÇÃO	PRAIA E CAMPING	21
DECORAÇÃO	CAMPING E PESCA	17
DECORAÇÃO	COSMÉTICOS E PERFUMARIA	3
DECORAÇÃO	PNEU	3
Total		**3.698**

Figura 19 – Relação entre o Setor Origem "Decoração"e os Setores Recomendados.

A Figura 19 apresenta a conexão entre o setor de origem "Decoração"e os setores recomendados. Nesta visualização, destacam-se padrões específicos de recomendação que indicam associações notáveis entre produtos de decoração e outros setores. Entre os setores mais destacados com produtos recomendados, sobressaem-se especialmente os segmentos de cama, mesa e banho, além de utilidades domésticas.

Setor Antecendente	Setor Consequente	Qtde Recomendações
CAMA MESA E BANHO	CAMA MESA E BANHO	24.520
CAMA MESA E BANHO	UTILIDADES DOMESTICAS	4.471
CAMA MESA E BANHO	TAPETES	1.398
CAMA MESA E BANHO	LINGERIE	1.294
CAMA MESA E BANHO	ELETRO-PORTÁTIL	1.215
CAMA MESA E BANHO	BRINQUEDOS	1.155
CAMA MESA E BANHO	FERRAMENTAS	1.075
CAMA MESA E BANHO	CONF FEMININA	779
CAMA MESA E BANHO	CORTINAS	768
CAMA MESA E BANHO	BEBE	534
CAMA MESA E BANHO	CONF INFANTO	502
CAMA MESA E BANHO	CONF MASCULINA	490
CAMA MESA E BANHO	DECORAÇÃO	477
CAMA MESA E BANHO	ESPORTIVO	343
CAMA MESA E BANHO	PRAIA E CAMPING	203
CAMA MESA E BANHO	CAMPING E PESCA	126
CAMA MESA E BANHO	COSMÉTICOS E PERFUMARIA	43
CAMA MESA E BANHO	MALAS	26
CAMA MESA E BANHO	PNEU	25
Total		**39.444**

Figura 20 – Relação entre o Setor Origem "Cama, Mesa e Banho"e os Setores Recomendados.

Na Figura 20, analisamos a conexão entre o setor de origem "Cama, Mesa e Banho"e os setores recomendados. Notamos que o setor antecedente, relacionado a cama, mesa e banho, apresenta uma correlação significativa com os setores de utilidades domésticas e tapetes, sendo estes os principais destinos das recomendações.

Essa observação reforça a tendência identificada anteriormente na análise do setor "Decoração"(Figura 19), onde também se destacaram as associações com produtos de utilidades domésticas revelando a interconectividade entre os setores.

Na Figura 21, a interação entre o setor de origem "Eletro-Portátil"e os setores alvos das recomendações são evidenciadas. Nessa análise, observamos que o setor de utilidades domésticas e ferramentas emerge como o principal destinatário das recomendações, estabelecendo uma conexão notável entre produtos eletro-portáteis e essas categorias específicas.

Essa constatação complementa a análise anterior dos setores "Decoração"e "Cama, Mesa e Banho"(Figuras 19 e 20), onde utilidades domésticas foram destacadas como setores correlacionados.

Setor Antecendente	Setor Consequente	Qtde Recomendações
ELETRO-PORTÁTIL	ELETRO-PORTÁTIL	7.594
ELETRO-PORTÁTIL	UTILIDADES DOMESTICAS	2.112
ELETRO-PORTÁTIL	FERRAMENTAS	1.498
ELETRO-PORTÁTIL	CAMA MESA E BANHO	1.215
ELETRO-PORTÁTIL	BRINQUEDOS	783
ELETRO-PORTÁTIL	LINGERIE	529
ELETRO-PORTÁTIL	CONF FEMININA	402
ELETRO-PORTÁTIL	CORTINAS	373
ELETRO-PORTÁTIL	TAPETES	360
ELETRO-PORTÁTIL	CONF MASCULINA	330
ELETRO-PORTÁTIL	ESPORTIVO	297
ELETRO-PORTÁTIL	CONF INFANTO	282
ELETRO-PORTÁTIL	BEBE	239
ELETRO-PORTÁTIL	DECORAÇÃO	229
ELETRO-PORTÁTIL	PRAIA E CAMPING	172
ELETRO-PORTÁTIL	CAMPING E PESCA	96
ELETRO-PORTÁTIL	PNEU	63
ELETRO-PORTÁTIL	COSMÉTICOS E PERFUMARIA	26
ELETRO-PORTÁTIL	MALAS	19
Total		**16.619**

Figura 21 – Relação entre o Setor Origem "Eletro-Portátil"e os Setores Recomendados.

Setor Antecendente	Setor Consequente	Qtde Recomendações
UTILIDADES DOMESTICAS	UTILIDADES DOMESTICAS	22.968
UTILIDADES DOMESTICAS	CAMA MESA E BANHO	4.471
UTILIDADES DOMESTICAS	ELETRO-PORTÁTIL	2.112
UTILIDADES DOMESTICAS	FERRAMENTAS	1.669
UTILIDADES DOMESTICAS	TAPETES	1.476
UTILIDADES DOMESTICAS	BRINQUEDOS	1.349
UTILIDADES DOMESTICAS	LINGERIE	1.345
UTILIDADES DOMESTICAS	CONF FEMININA	1.060
UTILIDADES DOMESTICAS	CONF INFANTO	777
UTILIDADES DOMESTICAS	CONF MASCULINA	659
UTILIDADES DOMESTICAS	CORTINAS	614
UTILIDADES DOMESTICAS	ESPORTIVO	492
UTILIDADES DOMESTICAS	DECORAÇÃO	450
UTILIDADES DOMESTICAS	BEBE	400
UTILIDADES DOMESTICAS	PRAIA E CAMPING	157
UTILIDADES DOMESTICAS	CAMPING E PESCA	135
UTILIDADES DOMESTICAS	COSMÉTICOS E PERFUMARIA	44
UTILIDADES DOMESTICAS	MALAS	37
UTILIDADES DOMESTICAS	PNEU	22
UTILIDADES DOMESTICAS	LINHA PET	1
Total		**40.238**

Figura 22 – Relação entre o Setor Origem "Utilidades Domésticas"e os Setores Recomendados.

Concluindo, a Figura 22 investiga a interação entre o setor de origem "Utilidades Domésticas"e os setores recomendados. Nesta análise, é notável que o setor de utilidades domésticas se destaca como uma fonte predominante de recomendações para os setores de cama, mesa e banho, bem como para o setor de eletro-portátil.

Esta observação estabelece uma conexão significativa com as análises anteriores dos setores "Decoração,Cama, Mesa e Banho,"e "Eletro-Portátil"(Figuras 19, 20, e 21), onde utilidades domésticas foram identificadas como um setor correlacionado.

7.4 Análise de Estações: Verão e Inverno

Nesta seção, conduzimos uma análise dos produtos recomendados nas estações de inverno e verão, nas regiões sul e norte do país.

A Figura 23 apresenta uma lista de produtos sugeridos considerando a compra do produto "Caixa Térmica"em junho na região norte. Notamos recomendações de produtos relacionados a praia, camping, brinquedos e utilidades domésticas, correlacionadas com as temperaturas mais elevadas da região norte em comparação com a região sul.

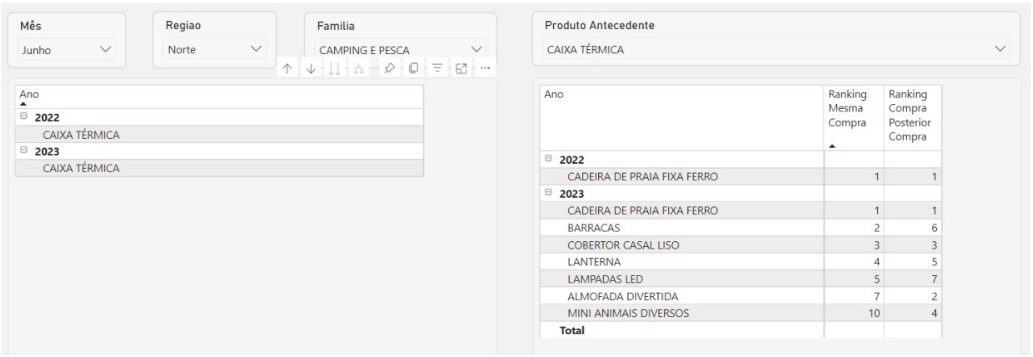

Figura 23 – Produto "Caixa Térmica" x Lista de Produtos Recomendados (Junho - Norte)

Na Figura 24, não observamos produtos relacionados a camping, pesca e praia para junho na região sul, devido às suas baixas temperaturas nos meses de inverno. No entanto, em janeiro na região sul, vemos uma lista de produtos recomendados após a compra de uma caixa térmica, com a maioria sendo itens de praia.

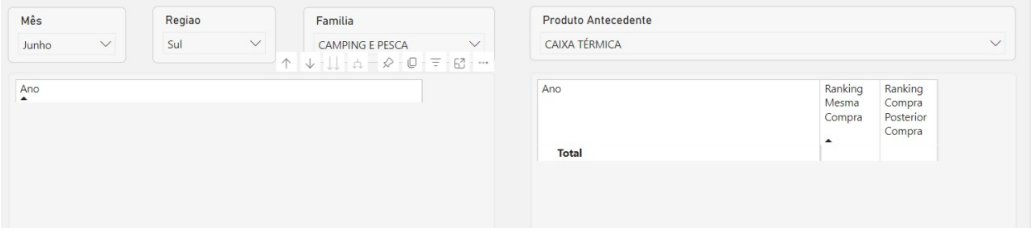

Figura 24 – Produto "Caixa Térmica" x Lista de Produtos Recomendados (Junho - Sul)

As Figuras 26 e 27 mostram uma lista de produtos recomendados para a mesma compra e para uma compra posterior após a escolha do produto "Edredom" em junho.

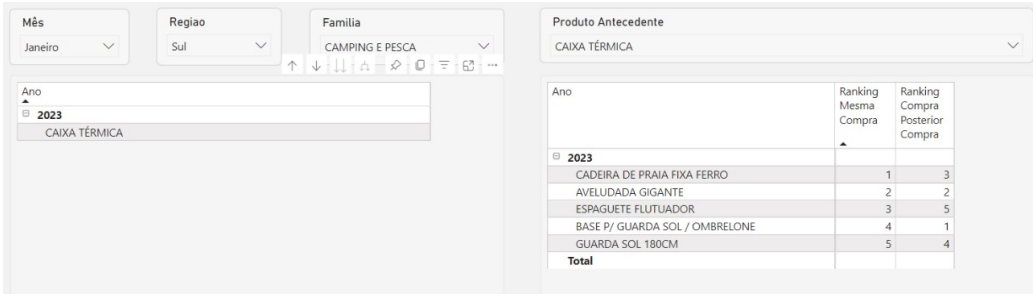

Figura 25 – Produto "Caixa Térmica" x Lista de Produtos Recomendados (Janeiro - Sul)

Observamos que, para a região norte, há mais produtos leves de cama, mesa e banho, enquanto na região sul há recomendações de produtos mais pesados, adequados para climas mais frios.

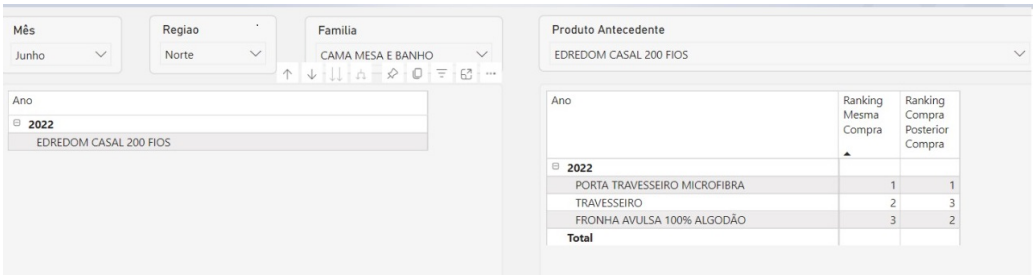

Figura 26 – Produto "Edredom" x Lista de Produtos Recomendados (Junho - Norte)

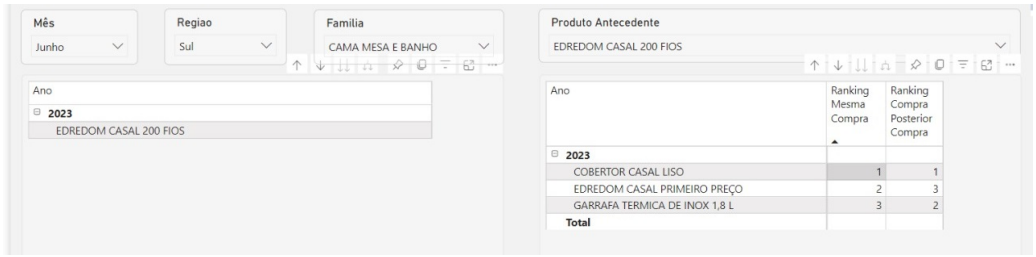

Figura 27 – Produto "Edredom" x Lista de Produtos Recomendados (Junho - Sul)

Observamos comportamento semelhante para o produto "Saia" nas regiões norte e sul, conforme evidenciado nas Figuras 28 e 29. Na região sul, são recomendadas roupas mais adequadas a temperaturas mais baixas, enquanto na região norte, as recomendações são para produtos mais leves, associados a climas mais quentes.

O mesmo padrão é observado para o produto "Guarda Sol" em junho. Não há recomendações de produtos relacionados a praia na região sul, como mostrado na Figura 30. No entanto, para a região norte, há sugestões de produtos de praia, conforme ilustrado na Figura 31. Em janeiro, na região sul, há recomendações para o produto "Guarda Sol", devido às altas temperaturas neste período, como evidenciado na Figura 32.

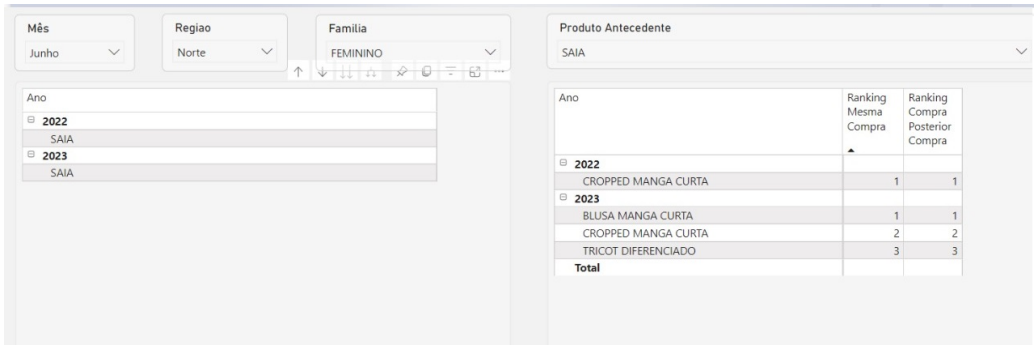

Figura 28 – Produto "Saia" x Lista de Produtos Recomendados (Junho - Norte)

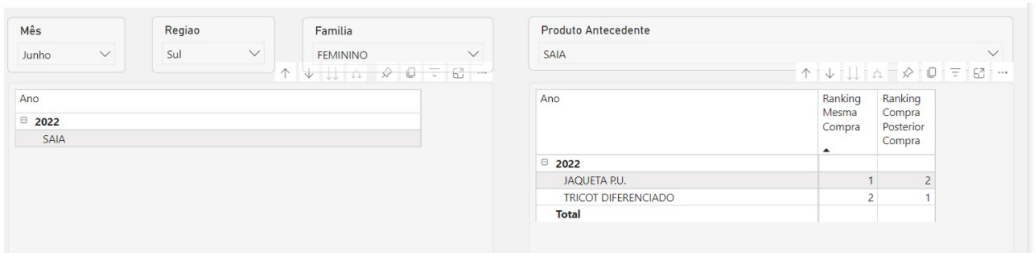

Figura 29 – Produto "Saia" x Lista de Produtos Recomendados (Junho - Sul)

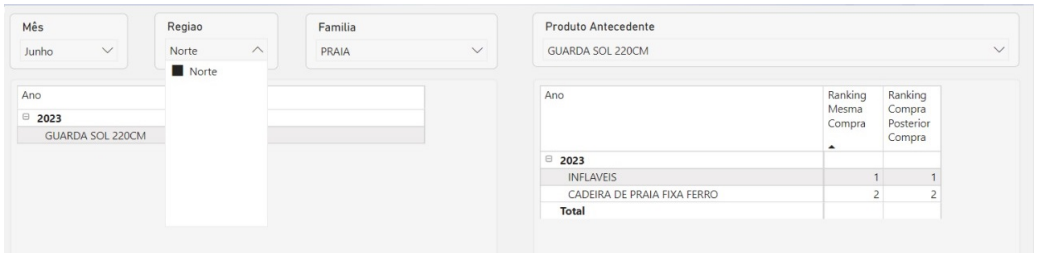

Figura 30 – Produto "Guarda Sol" x Lista de Produtos Recomendados (Junho - Sul)

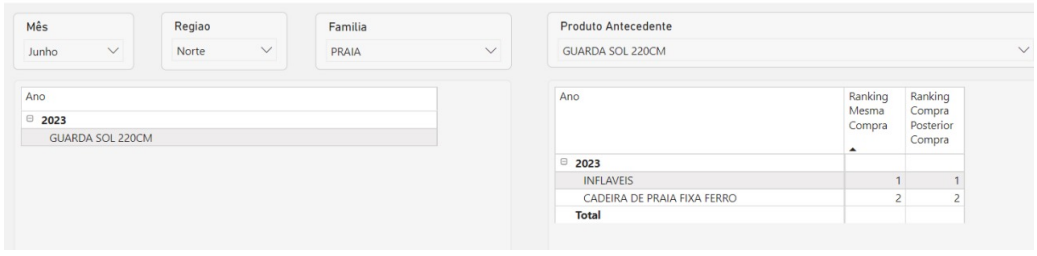

Figura 31 – Produto "Guarda Sol" x Lista de Produtos Recomendados (Junho - Norte)

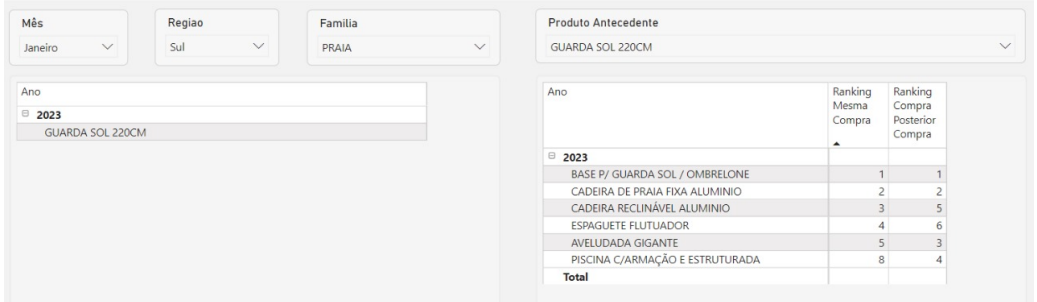

Figura 32 – Produto "Guarda Sol" x Lista de Produtos Recomendados (Janeiro - Sul)

7.5 Produtos Mais Vendidos por Região

Ao explorarmos a categoria de produtos infantis, identificamos um total de 15.867 itens vendidos no mês de referência, junho de 2022. Os cinco produtos mais destacados nessa categoria, apresentados na Figura 33, totalizaram 6.250 unidades vendidas, representando 39,39% do volume total. Esses produtos estão fortemente associados ao padrão de demanda observado durante a estação de inverno na região sul. Em contraste, na região norte, conforme ilustrado na Figura 34, a família de produtos infantis registrou um volume de vendas menor, com um comportamento mais alinhado a temperaturas mais elevadas.

2022/06	**6.250**		**15.867**	**39,39%**
CALÇA MOLETOM	1.756	1		11,07%
JAQUETA TACTEL/MICROFIBRA	1.423	2		8,97%
BLUSÃO MOLETOM	1.347	3		8,49%
CALÇA JEANS	908	4		5,72%
CALÇA SARJA	816	5		5,14%

Figura 33 – Produtos mais vendidos na região sul em junho/2022 categoria infantil

2022/06	**2.858**		**7.999**	**35,73%**
CAMISETA MANGA CURTA	728	1		9,10%
BLUSA M.C	650	2		8,13%
CALÇA MOLETOM	542	3		6,78%
BERMUDA MOLETOM	490	4		6,13%
CALÇA JEANS	448	5		5,60%

Figura 34 – Produtos mais vendidos na região norte em junho/2022 categoria infantil

Em setembro de 2023, de acordo com (TORTELLA, 2023), uma onda de ar seco estacionou no país, resultando em um bloqueio atmosférico que elevou as temperaturas. Esse fenômeno impactou diretamente nas vendas da categoria de produtos de praia, como evidenciado na comparação entre os meses de setembro de 2022 e 2023 na região sul, durante a estação de inverno. A Figura 35 ilustra o aumento significativo no volume de vendas desses produtos.

Neste capítulo, foram analisados os resultados do recomendador por meio de análises sobre produto, setor, estações do ano e os produtos mais vendidos em cada região além da análise de tempo de execução e estimativa de precisão do modelo. Na próxima seção será abordada as considerações finais deste trabalho.

2022/09	**195**		**232**	**84,05%**
CADEIRA DE PRAIA FIXA FERRO	90	1		38,79%
ESPAGUETE FLUTUADOR	32	2		13,79%
CADEIRA DE PRAIA FIXA ALUMINIO	25	3		10,78%
GUARDA SOL 180CM	18	4		7,76%
INFLAVEIS	15	5		6,47%
PISCINA INFLAVEL INFANTIL	15	5		6,47%
2022/10	**635**		**949**	**66,91%**
2022/11	**1.591**		**2.702**	**58,88%**
2022/12	**4.280**		**7.995**	**53,53%**
2023/01	**1.998**		**3.918**	**51,00%**
2023/02	**884**		**1.563**	**56,56%**
2023/03	**363**		**513**	**70,76%**
2023/04	**188**		**224**	**83,93%**
2023/05	**114**		**138**	**82,61%**
2023/06	**127**		**151**	**84,11%**
2023/07	**119**		**143**	**83,22%**
2023/08	**116**		**132**	**87,88%**
2023/09	**403**		**578**	**69,72%**
CADEIRA DE PRAIA FIXA FERRO	157	1		27,16%
INFLAVEIS	73	2		12,63%
CADEIRA DE PRAIA FIXA ALUMINIO	68	3		11,76%
ESPAGUETE FLUTUADOR	57	4		9,86%
GUARDA SOL 200CM	48	5		8,30%
2023/10	**446**		**701**	**63,62%**

Figura 35 – Quantidade de produtos vendidos setembro/2023 e 2022 categoria praia

8 CONSIDERAÇÕES FINAIS

O propósito central desta pesquisa consistiu no desenvolvimento de um sistema de recomendação abrangente, alinhando-se às exigências de uma extensa rede varejista e ao ambiente de pesquisa identificado na RSL. Considerando tanto o cenário do comércio eletrônico quanto das lojas físicas, e enfrentando os desafios apresentados pelo *big data*, a singularidade desta abordagem reside na viabilização eficiente da recomendação de produtos em larga escala, dada a imensidão dos conjuntos de dados.

Para atingir esse propósito, uma RSL foi conduzida na Seção 3, buscando responder à pergunta "Como os algoritmos de *frequent pattern mining* com análise de cesta de mercado estão sendo aplicados na recomendação de produtos?". A resposta a essa pergunta orientou a escolha do algoritmo *Fp-Growth*.

A solução proposta adotou uma abordagem integrativa, empregando tecnologias e ferramentas específicas para lidar com a vasta quantidade de dados da grande rede varejista. A combinação de ambientes *on-premise* e *cloud computing*, associada ao uso de ferramentas como Spark e S3, revelou-se uma estratégia eficaz para enfrentar o desafio do processamento massivo de dados, conforme detalhado na Seção 4.1.

A etapa de limpeza e organização do dataset proporcionou *insights* valiosos sobre padrões de compra, comportamentos sazonais e desafios inerentes ao processamento de grandes volumes de dados.

Apesar dos avanços, alguns desafios relacionados à sazonalidade e volumetria de dados foram identificados. Testes em diferentes ambientes evidenciaram dificuldades na execução do algoritmo FP-Growth em *clusters* distribuídos, apontando para a necessidade de aprimoramentos nesse aspecto.

A implementação bem-sucedida do algoritmo *FP-Growth*, aliado ao MBA, demonstrou ser uma escolha acertada, gerando recomendações relevantes e significativas. Destaca-se não apenas a precisão em sugerir variações de produtos, mas também a habilidade de ajustar as recomendações com base nas diversas culturas e climas das regiões do Brasil. Além disso, o sistema demonstra capacidade para impulsionar a venda de produtos complementares em uma mesma compra ou para compras

futuras, conforme evidenciado na Seção 7.

No entanto, há espaço para melhorias, como aprimorar a capacidade de generalização, demandando uma recomendação por categoria de produto. Além disso, como perspectiva para trabalhos futuros, seria valioso obter os *clusters* de clientes da grande rede varejista e realizar recomendações específicas para esses *clusters*. Isso permitiria identificar e sugerir produtos de forma mais precisa para clientes com características semelhantes, proporcionando uma compreensão mais profunda do comportamento de compra.

REFERÊNCIAS

AGRAWAL. Mining association rules between sets of items in large databases. **SIGMOD Rec.**, New York, NY, USA, v.22, n.2, p.207–216, jun 1993.

APACHE. **Apache Spark - A Unified engine for large-scale data analytics**. Disponível em: <https://spark.apache.org/docs/latest/>. Acesso em: 2023-09-17.

APACHE. **Apache Parquet**. Disponível em: <https://parquet.apache.org/>. Acesso em: 2023-09-17.

APACHE. **Apache Airflow**. Disponível em: <https://airflow.apache.org/>. Acesso em: 2023-09-17.

ASTRONOMER. **Airflow task groups**. Disponível em: <https://docs.astronomer.io/learn/task-groups>. Acesso em: 2023-09-17.

AWS. **O que é uma API?** Disponível em: <https://aws.amazon.com/pt/what-is/api/>. Acesso em: 2023-09-17.

AWS. **Amazon S3**. Disponível em: <https://aws.amazon.com/pt/s3/>. Acesso em: 2023-09-17.

BAGUI. A Heuristic Approach for Load Balancing the FP-Growth Algorithm On MapReduce. **Array**, Pensacola, Florida, v.7, p.100035, 08 2020.

BARBOSA. **Estudo de Técnicas de Filtragem Híbrida em Sistemas de Recomendação de Produtos**. 2014. 99p. Trabalho de Conclusão (Curso de Ciência da Computação) — Universidade Federal de Pernambuco, Recife.

BLATTBERG. **Database Marketing. International Series in Quantitative Marketing**. Austin, TX, U.S.A: Springer, 2008. 875p.

BOBADILLA. **Recommender systems survey**. Bruxelles, Bélgica: Elsevier, 2013. 109-132p. v.46.

BURKE. Hybrid Recommender Systems: Survey and Experiments. , USA, v.12, n.4, p.331–370, nov 2002.

CHONNY. **FP Growth — Frequent Pattern Generation in Data Mining with Python Implementation**. Disponível em: <https://towardsdatascience.com/fp-growth-frequent-pattern-generation-in-data-mining-with-python-implementation-244e561ab1c3>. Acesso em: 2023-09-17.

CRUZ. Crescimento do e-commerce no Brasil: desenvolvimento, servicos logisticos e o impulso da pandemia de Covid 19. **GeoTextos**, Bahia, 2021.

DERMEVAL. **Metodologia de Pesquisa Científica em Informática na Educação**: Abordagem Quantitativa (Volume 2). [S.l.]: https://metodologia.ceie-br.org/, 2020. 26p.

GOLDBERG. Using collaborative filtering to weave an information tapestry. In: COMMUNICATIONS OF THE ACM, 1992, Boston, MA. **Anais...** Association for Computing Machinery, 1992. v.35, n.12, p.61–70.

GRAZIOSI. **Elaboração da Pergunta Norteadora de Pesquisa**. Sao Paulo, Brazil: Universidade Federal de São Paulo, 2011. 243p.

GURUDATH. **Market Basket Analysis & Recommendation System Using Association Rules**. 2020. 59p. Mestrado em Ciência em Gerenciamento e Análise de Big Data. — Griffith College Dublin, Dublin.

HAOSONG. Scalability Challenges in Big Data Analytics: A Survey. **Big Data Research**, USA, v.17, p.33–41, 2019.

IBM. **What is data mining?** Disponível em: <https://www.ibm.com/topics/data-mining>. Acesso em: 2023-10-01.

IBM. **What is exploratory data analysis?** Disponível em: <https://www.ibm.com/topics/exploratory-data-analysis>. Acesso em: 2023-09-17.

JOHNSON. Climate and Consumer Behavior: Direct and Indirect Effects. **Journal of Marketing Research**, USA, v.52, n.6, p.820–835, 2015.

JUNIOR. **Avaliação de Técnicas de Filtragem Colaborativa para Sistemas de Recomendação**. 2017. 33p. Trabalho de Conclusão (Curso de Ciência da Computação) — Universidade Federal de Pernambuco, Recife.

KAMEPALLI. Weighted Based Frequent and Infrequent Pattern Mining Model for Real-time E-Commerce Databases. **Advancing the World of Information and Engineering**, [S.l.], v.62, p.8, 2019.

KSHETRI. Determinants of the Locus of Global E-Commerce. **Electron. Mark.**, [S.l.], v.11, p.250–257, 2001.

LOPS. Content-based Recommender Systems: State of the Art and Trends. **Recommender Systems Handbook**, Boston, MA, p.73–105, 2011.

MYTHILI. Performance Evaluation of Apriori and FP-Growth Algorithms. **International Journal of Computer Applications**, India-+-, 2013.

ORACLE. **O que é Big Data?** Disponível em: <https://www.oracle.com/br/big-data/what-is-big-data/>. Acesso em: 2023-09-17.

PATIL. Apriori Algorithm against Fp Growth Algorithm: A Comparative Study of Data Mining Algorithms. **SSBT's College of Engineering and Technology**, India, 2022.

PAZZANI. **Content-based recommendation systems**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. 325–341p.

PRIMO. **Método de representação de conhecimento baseado em ontologias para apoiar sistemas de recomendação educacionais**. 2013. 120p. Tese de doutorado (Ciência da computação) — Universidade Federal do Rio Grande do Sul, Porto Alegre.

RICCI. **Recommender Systems Handbook**. New York, USA: Springer, 2011.

SEBRAE. **O que são canais de venda? Entenda para ter os melhores resultados**. Disponível em: <https://sebrae.com.br/sites/PortalSebrae/ufs/pe/artigos/o-que-sao-canais-de-venda-entenda-para-ter-os-melhores-resultados,cb59b39cba395710VgnVCM1000004c00210aRCRD>. Acesso em: 2024-01-28.

SEBRAE. **Como se preparar para lidar com a sazonalidade de vendas**. Disponível em: <https://sebrae.com.br/sites/PortalSebrae/artigos/como-se-preparar-para-lidar-com-a-sazonalidade-de-vendas,b207c4ec9b805810VgnVCM100000d701210aRCRD>. Acesso em: 2023-02-03.

SIKHA. Big Data Processing with Spark in E-commerce. **International Journal of Computer Applications**, Pensacola, Florida, v.180, n.21, p.11–15, 2018.

TORTELLA. **Por que está tão quente no brasil no final do inverno?** Disponível em: <https://www.cnnbrasil.com.br/nacional/por-que-esta-tao-quente-no-brasil-no-final-do-inverno/>. Acesso em: 2023-09-19.

UFG. **Medidas de Semelhanca**. Disponível em: <https:files.cercomp.ufg.brwebyup417oAula_2_Medidas_de_Semelhan%C3%A7a_EE_II_2017.p
Acesso em: 2024-01-03.

VASCONCELOS. **Aplicação de Regras de Associação para Mineração de Dados na Web**. 2004. 20p. Relatório Técnico — Instituto de Informática Universidade Federal de Goiás, Goiás.