DOI: 10.5748/20CONTECSI/PSE/DSC/7336

eLocator: e207336

DISCOVERING SHOPPING PATTERNS AND TRENDS IN SUPERMARKETS USING SHOPPING BASKET ANALYSIS ; DESCOBERTA DE PADRÕES E TENDÊNCIAS DE COMPRAS EM SUPERMERCADOS UTILIZANDO ANÁLISE DE CESTAS DE COMPRAS

Helder Mateus Dos Reis Matos – https://orcid.org/0000-0002-5632-7948
Universidade Federal Do Pará

Wilton Freitas Ribeiro – https://orcid.org/0000-0002-7689-8991
Universidade Federal Do Pará

Reginaldo Cordeiro Dos Santos Filho – https://orcid.org/0000-0002-0456-8547
Universidade Federal Do Pará

João Crisóstomo Weyl Albuquerque Costa – https://orcid.org/0000-0003-4482-6886
Universidade Federal Do Pará

DISCOVERING OF SHOPPING PATTERNS AND TRENDS IN SUPERMARKETS USING MARKET BASKET ANALYSIS

ABSTRACT

This article addresses the application of Market Basket Analysis (MBA) to improve marketing strategies, promotional campaigns, and strategic decisions such as inventory management and shelf organization. Causal relationships were identified through the application of associative rules with high confidence. It is concluded that, through data analysis, MBA enables supermarkets to adjust their approaches to meet customer needs, enhancing profitability, and fostering customer loyalty.

Keywords: Market Basket Analysis, Supermarket, Data Mining, Associative Rule Learning, ABC analysis.

DESCOBERTA DE PADRÕES E TENDÊNCIAS DE COMPRAS EM SUPERMERCADOS UTILIZANDO ANÁLISE DE CESTAS DE COMPRAS

RESUMO

Este artigo aborda a aplicação de Market BasketAnalysis (MBA) para melhoria de estratégias de marketing, campanhas de promoção e decisões estratégicas, como gerenciamento de estoque e organização de prateleiras. Foi possível extrair relações causais por meio da aplicação de regras associativas com alta confiança. Conclui-se que, por meio da análise de dados, o MBA capacita os supermercados a ajustar suas abordagens para atender às necessidades dos clientes, aprimorando a rentabilidade e a fidelização.

Palavras-chave: Análise de Cesta de Compras, Supermercado, Mineração de Dados, Aprendizado de Regras Associativas, Análise ABC.

1. Introdução

Atualmente, vivemos na era do *Big Data* pautada primordialmente nos cinco Vs, onde a sociedade gera um Volume de dados demasiadamente grande diariamente, com uma Velocidade exponencial, em Variedades de formatos, com Valor intrinseco e com Veracidade. Neste contexto, as ferramentas de análise oferecidas pelas áreas conhecida como *ciência e mineração de dados* tornaram-se fundamentais na exploração da enorme quantidade de dados gerados a cada dia.

Dessa forma, estamos cercados por uma enorme quantidade de dados, e isso é particularmente verdadeiro para as empresas. Os supermercados são um exemplo de negócio que lida com grandes quantidades de dados, e para eles, entender o comportamento do cliente é fundamental para manter as vendas em alta. A análise de cestas de compras, também conhecida como *Market BasketAnalysis* (MBA), é uma técnica que ajuda os supermercados a entender melhor os padrões de compras dos clientes, bem como as relações entre os produtos comprados. Compreender essas informações é essencial para otimizar a distribuição de produtos e aumentar as vendas. Neste contexto, a aplicação de MBA em supermercados pode ser considerada uma ferramenta poderosa para melhorar a eficácia do marketing e das campanhas de promoção, além de ajudar na tomada de decisões estratégicas, como o gerenciamento de estoque e a organização das prateleiras. Por meio da análise de dados, o MBA permite que os supermercados ajustem suas estratégias de marketing e vendas para atender às necessidades de seus clientes e, assim, melhorar sua rentabilidade e fidelização.

O presente trabalho visa descrever a aplicação de um processo de mineração de dados sobre os dados transacionais de um supermercado, com foco no processamento de grandes volumes de dados gerados diariamente. O processo visa gerar uma ferramenta inteligente de análise de cestas de compras que possa ser utilizada pela rede de supermercados para melhoria de sua logística e gestão interna em relação ao gerenciamento dos produtos.

Este artigo é organizado da seguinte forma: a Seção 2 investiga trabalhos relacionados ao escopo da pesquisa atual. A Seção 3 descreve a metodologia proposta. A Seção 4 apresenta os resultados obtidos e suas discussões. A Seção 5 finaliza o artigo com as considerações finais e as perspectivas de trabalhos futuros.

2. Trabalhos relacionados

Essa seção explora a literatura relacionada ao escopo de mineração de dados para análise de cestas de compras, o que auxiliou na determinação dos conceitos e ferramentas a serem utilizadas ao longo da pesquisa.

Na literatura pode-se encontrar aplicações sobre MBA, como em (Schonrost, G. B. et al., 2020)que trata da aplicação de técnicas de mineração de dados na análise de transações em supermercados no Brasil. O estudo tem como objetivo identificar padrões de comportamento de compra dos consumidores a partir dos dados transacionais coletados e, com isso, otimizar a estratégia de marketing dos supermercados. Para isso, foram utilizadas técnicas de Regras Associativas (*AssociationRules*) para identificar os padrões de compra. O estudo foi baseado em dados reais de transações de um supermercado no Brasil e apresentou resultados promissores para a detecção de padrões.

Outro trabalho que utilizou técnicas de Regras Associativas está disponível em (Djukanovicet al., 2022), onde os autores tratam da aplicação do algoritmo Apriori para aprimorar a gestão de relacionamento com o cliente (do inglês,

CustomerRelationshipManagement - CRM) em uma empresa de telecomunicações de Montenegro. O estudo utiliza técnicas de mineração de dados para identificar associações entre os produtos e serviços contratados pelos clientes, a fim de melhorar a oferta de produtos e serviços personalizados e aumentar a fidelidade dos clientes. O artigo apresenta a metodologia utilizada, os resultados obtidos e discute as implicações práticas da aplicação do algoritmo Apriori na área de CRM.

A segmentação de clientes também é uma técnica muito utilizada para agrupar clientes com perfis de compra parecidos. Os pesquisadores em (Yosephet al., 2020)abordam a importância do uso de técnicas de mineração de dados e clusterização na segmentação de mercado, visando uma maior compreensão dos comportamentos e preferências dos consumidores. Os autores realizaram um estudo de caso em uma empresa de telecomunicações, aplicando técnicas de agrupamento, como os algoritmos *Expectation-Maximization* (EM) e K-Means++, em grandes volumes de dados para identificar perfis de consumidores e suas necessidades específicas. O artigo discute os resultados obtidos e como eles podem ser utilizados para aprimorar estratégias de marketing e aumentar a satisfação dos clientes.

Pode-se perceber a importância do uso da segmentação de clientes também em (Tavakoliet al., 2018), onde os pesquisadores tratam de como a análise de comportamento do usuário pode ser usada para segmentação de clientes e desenvolvimento de estratégias de marketing mais eficazes. O estudo usa técnicas de mineração de dados, trazendo um novo modelo R+FM baseado no modelo tradicional RFM (*Recency, Frequency, Monetary*), para analisar o comportamento dos usuários e segmentar os clientes com base em seus padrões de compra. Os resultados mostraram que a segmentação de clientes com base no modelo R+FM pode ser uma estratégia eficaz para o desenvolvimento de campanhas de marketing personalizadas e aumento das vendas, onde foi feita uma campanha por SMS de acordo com essas estratégias. Os resultados da campanha mostraram que o Modelo de Segmentação proposto melhorou o número de compras e a média monetária das cestas. O estudo foi realizado em uma empresa de comércio eletrônico no Iran.

3. Metodologia

Como facilitador na organização dos processos de mineração de dados a serem aplicados, utilizou-se o modelo *CRoss-Industry Standard Process for Data Mining* (CRISP-DM), uma abordagem direcionada a indústria que segmenta a extração de conhecimento em seis etapas flexíveis e personalizáveis, adaptáveis de acordo com as necessidades do projeto (Martínez-Plumed et al., 2021). A seguir serão descritos os detalhes metodológicos tomados ao longo de cada etapa.

3.1. Entendimento do Negócio

O **entendimento do negócio**explora as expectativas da organização interessada na mineração de dados para formalizar o planejamento das demais etapas.

A pesquisa está centrada na exploração dos dados transacionais de compras no setor do varejo de uma rede de supermercados de grande relevância no estado do Pará, sendo firmada uma parceria entre o grupo de pesquisa e os representantes do negócio para o desenvolvimento de ferramentas computacionais de gestão, monitoramento e apoio à decisão. Desta forma, foi redigido um plano de trabalho que expõe as potenciais soluções de mineração de dados a serem abordadas incluindo:

- **Análises de vendas conjuntas de produtos:** as regras associativas permitem descobrir relações entre produtos vendidos em conjunto dentro de um mesmo cupom

de venda, sendo possível agir sobre o posicionamento de produtos com grande relação, assim como no redirecionamento de marketing e promoções.

- **Análise de sazonalidade de compras:** as séries temporais consideram campanhas de longo prazo para encontrar padrões de compras relevantes ao longo do ano, considerando períodos como datas comemorativas, feriados, férias, perfis mensais, eventos regionais, etc.
- **Controle otimizado de estoque com análise de curvas ABC:** a análise de curvas ABC é uma técnica de classificação utilizada para identificar e priorizar itens com base na sua importância relativa. A análise permite uma melhor alocação de recursos, focalizando a atenção nos itens mais relevantes, otimizando a gestão de estoque e o planejamento de compras.

3.2. Entendimento dos Dados

O **entendimento dos dados** permite compreender o estado dos dados disponíveis para mineração.

A tarefa de *coleta* decide quais os dados utilizados para a análise, baseados na relevância destes para os objetivos do projeto, delimitando quais as tabelas de banco de dados, os atributos e as amostras de interesse. Foi utilizado um conjunto de dados fornecido pela rede de supermercados que corresponde ao recorte dos cupons de vendas dos meses entre julho e novembro de 2023, para a filial mais proeminente da cidade de Belém, em termos de quantidade de cupons emitidos, receita gerada e localização vantajosa em um dos bairros nobres da capital paraense. O escopo do atual trabalho foi delimitado ao uso das seguintes tabelas:

- Capa: armazena dados gerais sobre os cupons dos clientes. Os atributos incluem identificadores dos cupons, data de emissão do cupom e valor total do cupom.
- Item: armazena dados detalhados sobre os cupons, com ênfase nos itens individuais comprados pelos clientes. Os atributos incluem identificadores dos cupons e dos produtos, a quantidade em que um mesmo produto é comprado, o valor unitário do produto e o valor do desconto.
- Produto: armazena dados detalhados sobre os produtos únicos presentes para o recorte utilizado. Os atributos incluem categorizações internas da empresa a nível de departamento, seção, grupo e subgrupo, código EAN (*EuropeanArticleNumber*), descrição textual do produto e tipo da embalagem.

A tarefa de exploração realiza uma análise exploratória dos dados, onde são gerados tabelas, gráficos e outras ferramentas de visualização que permitam entender o cenário exposto no entendimento do negócio e que ajudem a formular hipóteses e transformações necessárias durante a preparação dos dados.

3.3. Preparação dos Dados

A **preparação dos dados** realiza a transformação dos dados originais em um formato que esteja apto a servir como entrada às ferramentas de modelagem de dados.

A tarefa de construção realiza a criação de novos atributos que sejam de interesse para a mineração, através da derivação de atributos existentes ou da geração de atributos completamente novos. Para efeitos de categorização dos produtos, serão utilizadas as análises ABC e XYZ, que respectivamente adicionam atributos que descrevem a capacidade de geração de receita, a estabilidade de vendas de um produto.

A análise ABC ordena os produtos pela sua contribuição cumulativa no valor total dos produtos vendidos no período analisado, onde é possível observar o desdobramento do Princípio de Pareto (80% dos resultados são gerados por 20% das causas). Assim, produtos categorizados como de Classe A geram a maior parte da receita, enquanto os produtos das demais classes geram pouco valor.

A análise XYZ ordena os produtos pela variabilidade em sua demanda, destacando os produtos cujas vendas apresentam um comportamento constante ao longo do período como pertencentes à classe X, e distribuindo aqueles cujos estoques são mais difíceis de prever nas demais classes.

É comum combinar os resultados das análises ABC e XYZ, a fim de obter as classes de produtos com maior geração de valor e estabilidade de estoque em comparação com aqueles que geram pouca renda e tem um comportamento de demanda imprevisível. A categorização obtida pelo atributo ABC-XYZ permite uma seleção de amostras que serão exploradas pela ferramenta de modelagem de dados, o que também reduz a dimensionalidade da quantidade de produtos únicos a serem analisados.

3.4. Modelagem

A etapa de **modelagem** utiliza os dados preparados para responder aos problemas propostos na etapa de entendimento do negócio de forma iterativa, sendo possível fazer múltiplas execuções da ferramenta de modelagem e retornar a qualquer uma das etapas anteriores para realização de ajustes.

As técnicas de modelagem utilizadas são derivadas do conceito de *AssociativeRule Learning* (ARL), um método de aprendizado capaz de gerar regras que descrevem padrões de relacionamento entre os atributos.

3.5. Avaliação

A etapa de avaliação verifica se os critérios de sucesso estabelecidos no planejamento foram atendidos pelos resultados da modelagem. Os resultados incluem tanto os modelos e ferramentas construídos quanto às conclusões e inferências levantadas pela pesquisa e validadas junto aos representantes da rede de supermercados.

Sendo o principal produto da pesquisa, as regras associativas, precisam ser armazenadas, resumidas e facilitadas, considerando o grande volume de regras que pode ser gerado ao explorar diversos cenários de compras em conjunto. Métricas de probabilidade como suporte, confiança e lift podem ser usadas como atributos de seleção para as regras mais importantes para a análise dos representantes do supermercado.

3.6. Implantação

Uma vez validada a utilização e a praticidade dos modelos de aprendizado, a implantação faz uso das descobertas para gerar melhorias dentro da organização.

De forma geral, a implantação prevê o planejamento e monitoramento dos resultados obtidos ao longo de novas iterações e dos impactos do uso da mineração de dados.

4. Análise dos Resultados

4.1. Análise Exploratória dos Dados

A primeira análise busca encontrar padrões quantitativos da emissão de cupons ao longo de diferentes disposições temporais. O mapa de calor da Figura 1 organiza essas quantidades ao longo das 24 horas do dia e dos 7 dias da semana, usando uma escala de cores que mapeia valores altos para cores quentes e valores baixos para cores frias. Dessa forma, é possível encontrar os horários e dias de pico de movimentações na filial mencionada.

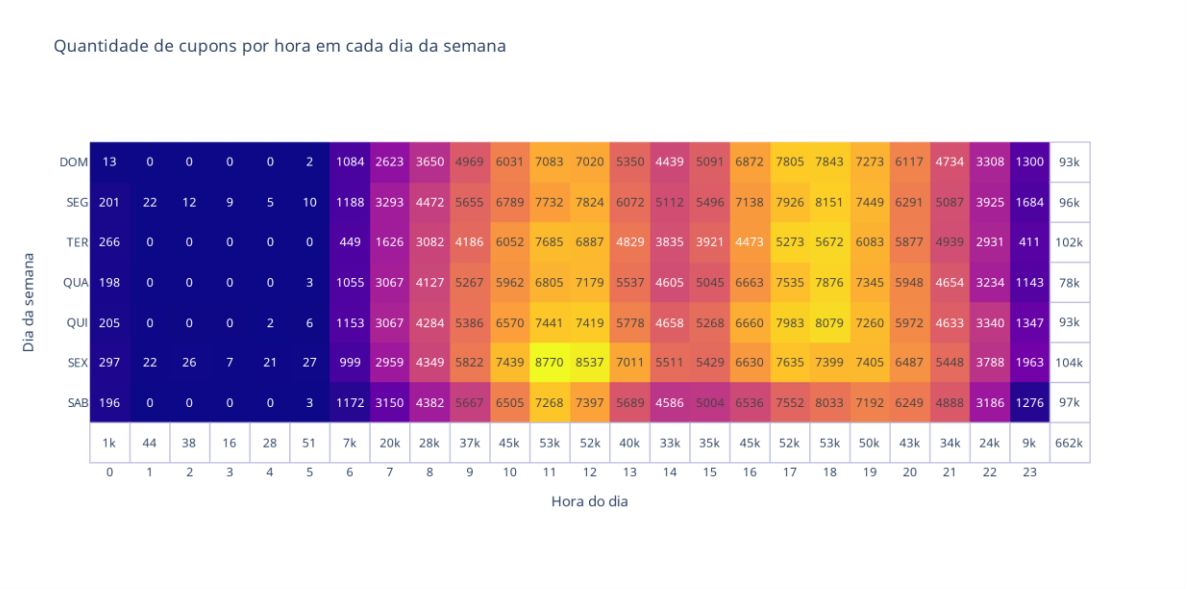

Figura 1. Mapa de calor da quantidade de cupons emitidos por hora do dia e por dia da semana.

Em relação aos horários, entre 10 e 13 horas e entre 17 e 20 horas foi encontrada a maior quantidade de cupons emitidos diariamente, horários estes relacionados com momentos de refeições importantes, como almoço e jantar, assim como com momentos de saída dos clientes de suas atividades diárias e consequente maior disponibilidade de realizar as compras. Assim, os turnos da tarde e noite são consideravelmente mais movimentados, em comparação ao turno da manhã.

Foram encontradas movimentações inesperadas no turno da madrugada, visto que a filial em questão tem horário de funcionamento previsto entre 6 e 23 horas. Uma grande parte destas compras, executadas durante as primeiras horas da madrugada, são referentes aos clientes que entraram no supermercado próximo ao seu fechamento, logo é permitido a estes um período adicional para finalizarem as compras. As demais movimentações podem ser repassadas aos representantes do supermercado para confirmação de suas validades, uma vez que podem representar anomalias que atrapalhem as subsequentes análises.

Em relação aos dias da semana, as movimentações são distribuídas de forma mais uniforme, com as sextas-feiras e terças-feiras representando a maior parte dos cupons emitidos, com uma queda entre estes dias. Durante o fim de semana, o comportamento dos horários de pico do resto da semana é atenuado, espalhando o quantitativo de cupons de forma mais balanceada ao longo do dia.

Em seguida, é observado como os produtos vendidos pela filial do supermercado no período são categorizados internamente, permitindo assim agrupamentos que serão utilizados principalmente na geração das regras associativas. A Figura 2 é um gráfico *sunburst* que trata da quantidade de produtos únicos distribuídos ao longo das categorias

hierárquicas do supermercado. Categorias com muitos produtos únicos recebem um arco de circunferência maior, partindo do centro, onde estão as categorias de Departamento, passando pelas categorias de Seção, até chegar nas categorias de Grupos localizadas nas bordas. Uma última categoria, de Subgrupo, foi removida do gráfico a fim de reduzir a quantidade de informações, visto que existem muitas ramificações a este nível.

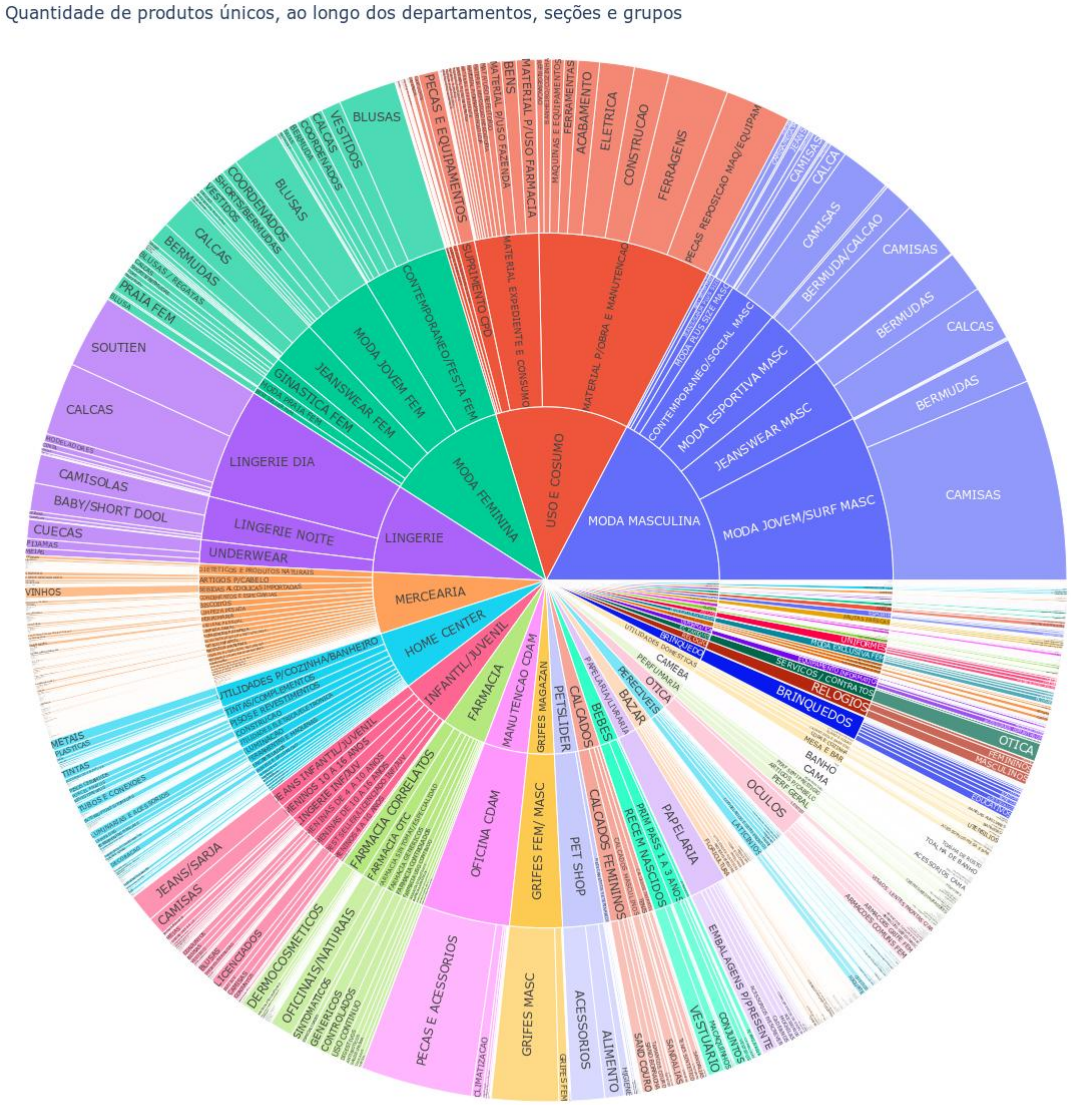

Figura 2. Gráfico *sunburst* para a quantidade de produtos únicos para as categorias de departamento, seção e grupo.

A partir do ângulo de grau 0 deste círculo, em sentido anti-horário, podemos destacar os quatro departamentos que possuem o maior número de produtos únicos, sendo estes de moda masculina, de moda feminina, de uso e consumo, e de lingerie. De fato, estes departamentos possuem uma alta variabilidade de produtos e que são substituídos diversas vezes ao longo de um período, especialmente em relação à moda como um todo. Este é um forte indicativo de que estes produtos estão presentes nas classes mais baixas da análise ABC-XYZ, pois não apenas a receita gerada é extremamente segmentada ao longo dos diversos produtos únicos, mas também sua demanda é consideravelmente difícil de prever. Por outro lado, os departamentos dispostos nos ângulos seguintes possuem distribuições de quantidades mais uniformes. Este comportamento é importante para os algoritmos de extração de frequências, como os que são usados na geração de regras associativas, pois a dimensionalidade do conjunto de dados gerados por estas extrações depende diretamente da

quantidade de produtos únicos, que é diretamente impactada pelo nível de segmentação utilizado para especificar cada produto.

4.2. Análise de Séries Temporais

A análise de séries temporais foi gerada com o intuito de destacar as tendências gerais de compras para os atributos quantitativos da quantidade de cupons emitidos por dia e do percentual de receita diária para o período de estudo de 5 meses mencionado nos entendimentos do negócio e dos dados. Foram utilizadas as técnicas de Método dos Mínimos Quadrados (MMQ), de Médias Móveis (MM), e de Ajuste Exponencial (EXP).

Para a quantidade de cupons emitidos diariamente, a Figura 3 ilustra as representações das séries temporais por MMQ, MM e EXP, respectivamente. Há uma tendência geral de queda ao longo dos 5 meses, que se torna bastante evidente após a segunda semana de agosto. Apenas um dos picos aparenta fazer referência a períodos próximos à feriados e pontos facultativos, na semana do dia dos pais e do dia da adesão do Pará à independência do Brasil. Em relação aos dias da semana, os sábados costumam representar a maioria dos vales semanais. Além disso, existem momentos de regularidade curtos entre os dias 6 e 10 de cada mês, seguidos das quedas bruscas esperadas, ditando assim o comportamento a longo prazo desta série. Os erros de cada variação das 3 técnicas apresentadas estão expostos na Tabela 1, com os valores mínimos de cada erro em cada técnica destacado em negrito.

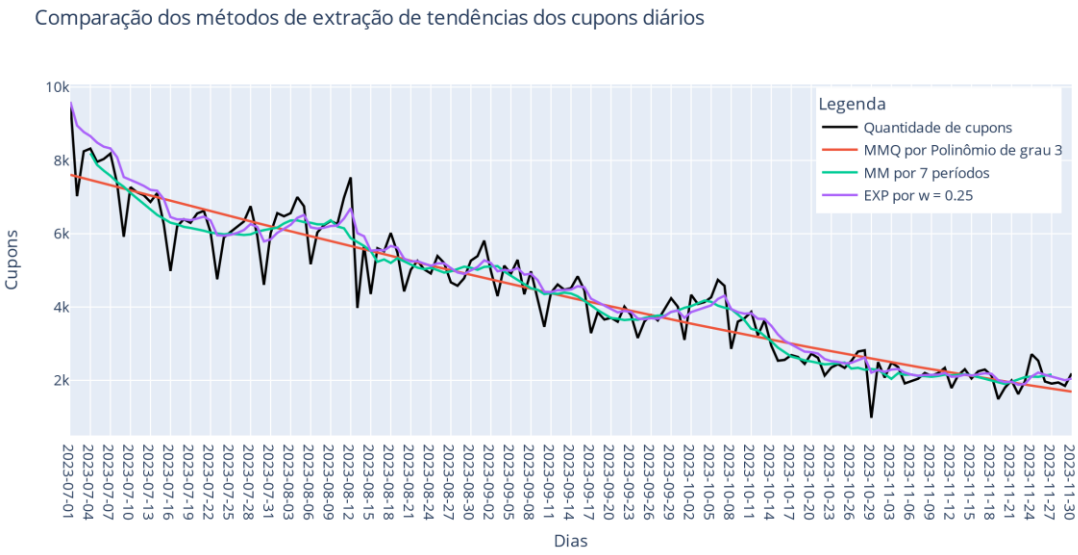

Figura 3. Tendências das quantidades de cupons diários por 3 técnicas diferentes.

Neste ponto, é importante destacar que a representação de MM descreve o comportamento da série no período de estudo, sem a possibilidade de realizar predições como nas outras duas representações. Assim, caso o intuito seja extrair a representação que melhor descreve a série, podemos escolher entre MM e EXP, enquanto que se o objetivo for extrair a representação com a maior capacidade preditiva, escolhemos entre MMQ e EXP, sendo que para esta última, quanto menor o valor de w, mais fiel à curva original será a curva de predição. Assim, o gráfico da Figura3 foi montado para destacar o comportamento de queda com menor erro obtido para MMQ polinomial de grau 3, em comparação com as variações de MM para 7 períodos e EXP para $w = 0.25$ que geram curvas mais suaves, a fim de evitar seguir com muita rigidez o comportamento do gráfico original.

Tabela 1. Erros calculados para as 4 técnicas de extração de tendências das quantidades de cupons e suas variações

Técnica	Variação	Erro	MAE	MSE	MAPE
MMQ	Linear	0	70479.09465	58917513.41	15223.50833
	Polinomial de grau 3	0	**68445.03148**	**56712279.12**	15106.4089
	Logaritmo	**0**	106808.4201	114147259.7	**14850.36**
	Exponencial	4939.715371	73376.1722	65384262.79	15291.46069
MM	3 períodos	**-970.666667**	**39146**	**22547615.78**	14875.32852
	5 períodos	216.8	46007.2	30035498.64	14695.60924
	7 períodos	148.142857	48583	34632872.31	**14497.64017**
EXP	$w = 0.25$	**-22636.83303**	48944.45743	35524327.58	**14608.72824**
	$w = 0.5$	-7539.130949	31094.11605	15176859.64	14904.66047
	$w = 0.75$	-2494.966321	**16829.39915**	**4158059.014**	15009.90346

Para a receita percentual gerada diariamente (normalizada entre 0 e 1, a fim de preservar a anonimidade destes dados), a Figura 4 ilustra as respectivas representações por MMQ, MM e EXP, cujos erros estão expostos na Tabela 2. Não foi observada uma tendência geral de crescimento ou decrescimento da receita, mas se tornam evidentes os dias de pico para 12 de agosto, 6 de outubro e 24 de novembro, além dos vales para 8 e 29 de outubro. Portanto, a predição da tendência por MMQ não segue o comportamento de nenhuma equação proposta. As médias móveis indicam um aumento no valor arrecadado por volta do início de cada mês, além de quedas nas segundas e terceiras semanas.

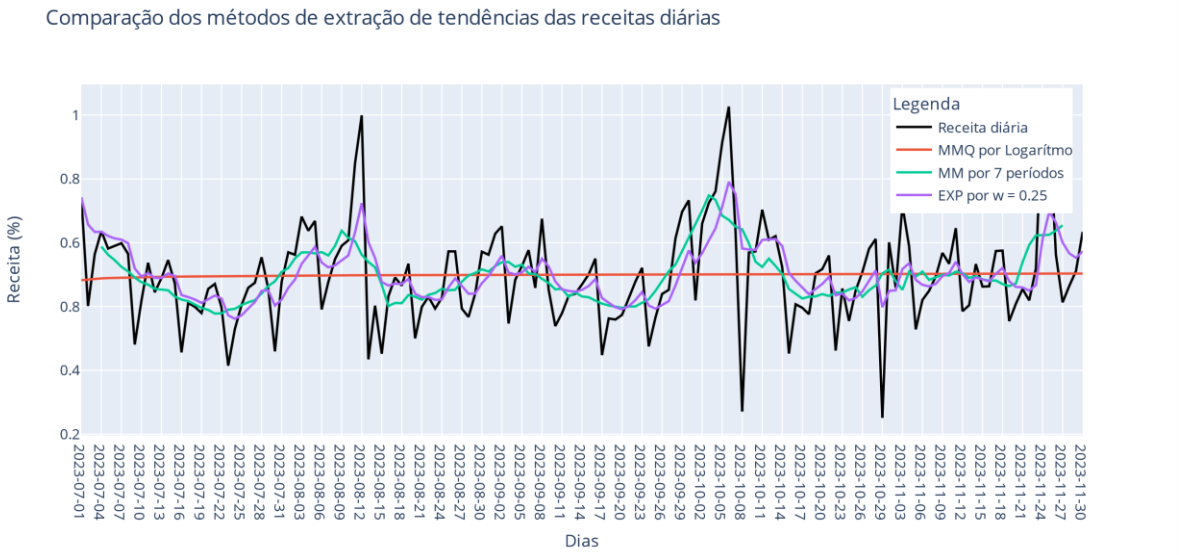

Figura 4. Tendências das receitas diárias por 3 técnicas diferentes.

Novamente foram escolhidas as representações com menor erro de MMQ para uma função polinomial, e as variações de MM para 7 períodos e EXP para $w = 0.25$.

Tabela 2. Erros calculados para as 3 técnicas de extração de tendências das receitas diárias e suas variações.

Técnica	Variação	Erro	MAE	MSE	MAPE
MMQ	Linear	0	70479.09465	58917513.41	15223.50833
	Polinomial de grau 3	0	**68445.03148**	**56712279.12**	15106.4089
	Logaritmo	**0**	106808.4201	114147259.7	**14850.36**
	Exponencial	4939.715371	73376.1722	65384262.79	15291.46069
MM	3 períodos	**-970.666667**	**39146**	**22547615.78**	14875.32852
	5 períodos	216.8	46007.2	30035498.64	14695.60924
	7 períodos	148.142857	48583	34632872.31	**14497.64017**
EXP	$w = 0.25$	**-22636.83303**	48944.45743	35524327.58	**14608.72824**
	$w = 0.5$	-7539.130949	31094.11605	15176859.64	14904.66047
	$w = 0.75$	-2494.966321	**16829.39915**	**4158059.014**	15009.90346

4.3. Análise ABC-XYZ

O modelo de classificação ABC aplicado no conjunto de dados fornecido segue a regra de Pareto esperada, como mostra a distribuição cumulativa de produtos da figura 5. Considerando cerca de 100 mil produtos únicos, a classe A abrange 12% dos itens e 80% da receita, representando a maior parte da contribuição total de valor gerado dentre todos os produtos. Por outro lado, cerca de 74% dos produtos únicos contribuem para apenas 10% do valor arrecadado, sendo postos à disposição dos gestores de estoque do supermercado para avaliação de sua rotatividade.

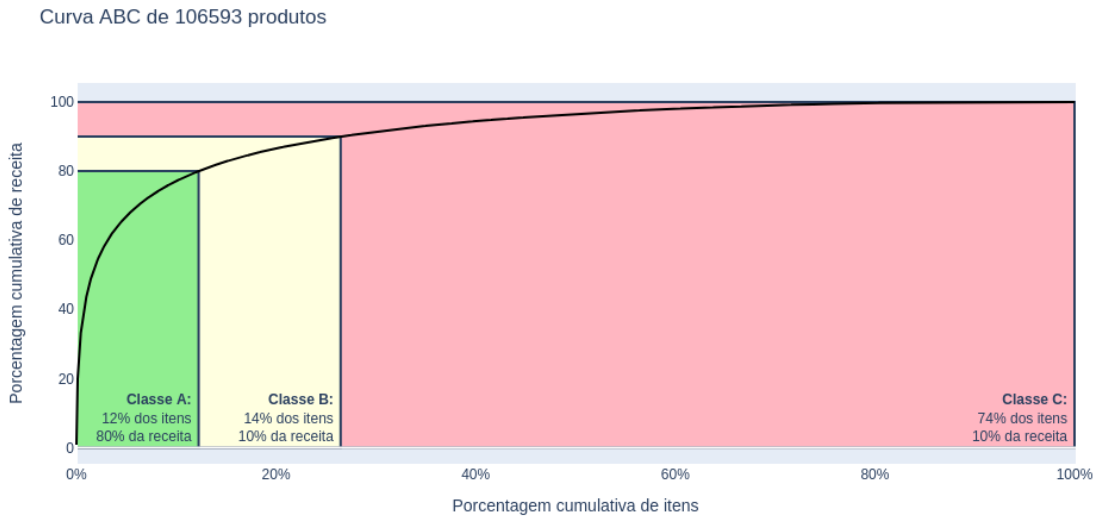

Figura 5. Curva ABC e a proporção de itens por quantidade e receita

A Figura 6 é um gráfico Treemap para organização hierárquica dos produtos que mais geram receita. A escala de cores dos quadrados é mapeada para os valores de SKU percentual, e, visto que o SKU percentual é a variável determinante da classe ABC, e que quanto menor for seu valor mais lucrativo será o produto, foi necessário o cálculo do complemento deste valor, $1 - SKU\%$, correspondente às áreas dos quadrados. Dessa forma, os quadrados com as maiores áreas e que estão alocados no canto superior esquerdo são os produtos com menor SKU percentual e maior geração de receita, cujos valores decrescem em direção ao canto inferior direito. Foi possível notar assim que a maior parte dos produtos de classe A pertencem ao departamento de *fast-food*, o que é diretamente relacionado com a pouca variabilidade de produtos únicos e grande demanda dos mesmos.

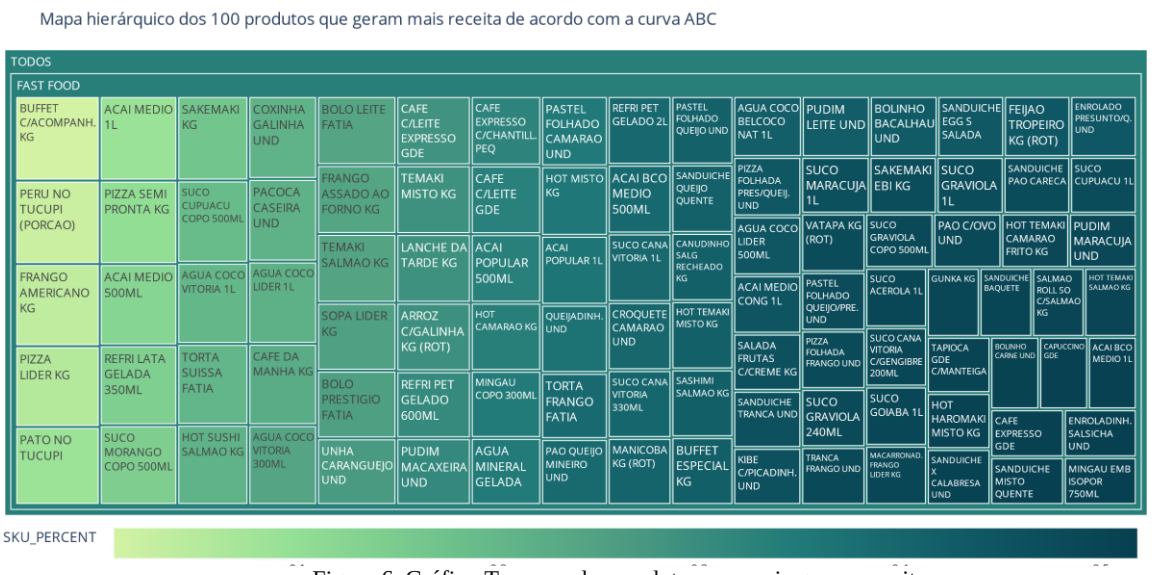

Figura 6. Gráfico Treemap dos produtos que mais geram receita.

O Treemap da Figura 7 seleciona 3 produtos de cada departamento, de forma que seja possível observar os produtos que geram mais receita para os principais departamentos. Além do fast-food, os produtos perecíveis, de mercearia e os hortifrutigranjeiros se revelam como os mais lucrativos, seguindo a mesma lógica da baixa variabilidade e alta demanda. Em seguida, departamentos de eletrodomésticos, celular, farmácia, móveis e informática possuem produtos de alto valor individual e baixa variabilidade, mas que não são vendidos com tanta frequência em relação aos anteriores, contribuindo em pequenas partes de alto valor na soma cumulativa.

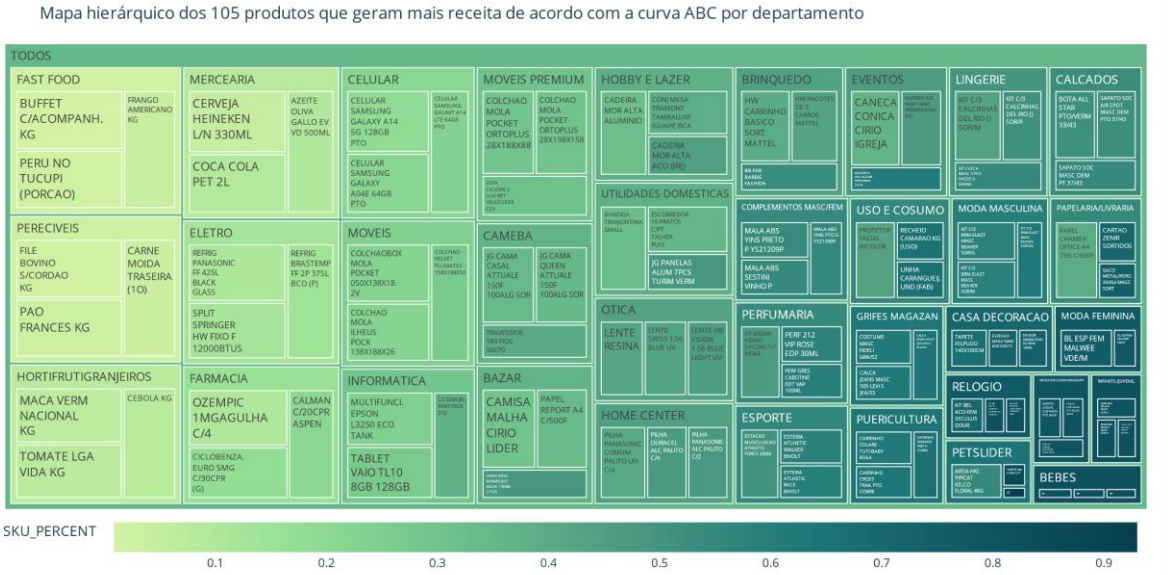

Figura 7. Gráfico Treemap dos produtos que mais geram receita por departamento.

Esta relação entre variabilidade, valor individual e demanda dos produtos únicos pode ser melhor compreendida por uma atualização do gráfico de *sunburst*, exposto na Figura 8, desta vez organizando apenas os produtos de classe A e separando maiores arcos do círculo para as categorias com menor SKU percentual. Apenas seis departamentos são responsáveis por pouco mais de 75% da receita gerada. O departamento de *fast-food*, outrora dominante no topo da hierarquia do Treemap sem restrições, ocupa um arco muito menor em relação à visão geral dos produtos de classe A. Apesar dos produtos individuais de*fast-food* gerarem muita receita, existem poucas variações em comparação com os produtos de mercearia, que individualmente geram menos receita, mas que têm uma maior variabilidade.

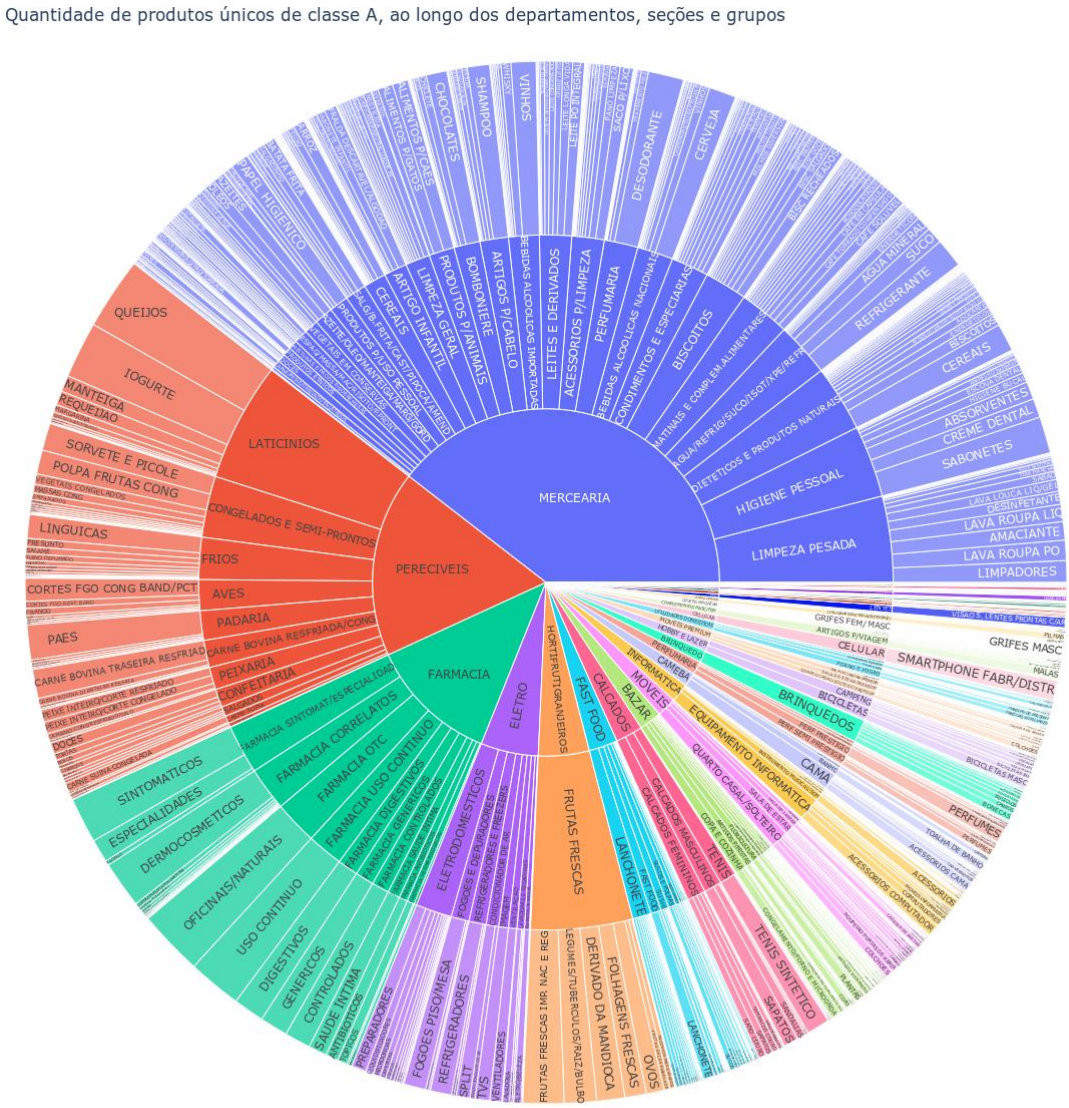

Figura 8. Gráfico *sunburst* para a quantidade de produtos únicos de classe A para as categorias de departamento, seção e grupo.

O SKU percentual da análise ABC e o coeficiente de covariância da análise XYZ podem ser usados em conjunto para a descoberta de classes de produtos com alta rentabilidade e alta estabilidade de vendas ao longo do período, além daqueles que por outro lado pouco contribuem para a receita total e que tem vendas extremamente difíceis de prever. A figura 8 utiliza as métricas propostas para distribuir espacialmente os produtos únicos, onde podemos observar esta relação entre concentração de valor e previsibilidade.

Os produtos que pertencem às classes A, B, X ou Y somam cerca de 22% do inventário, em comparação aos 78% daqueles de classes C ou Z. Esta informação tem impactos importantes na redução de dimensionalidade de dados passados para algoritmos de aprendizado, como será abordado nas regras associativas, já que o comportamento geral de compras pode ser analisado sem precisar considerar produtos pouco importantes para a venda total e que são altamente imprevisíveis.

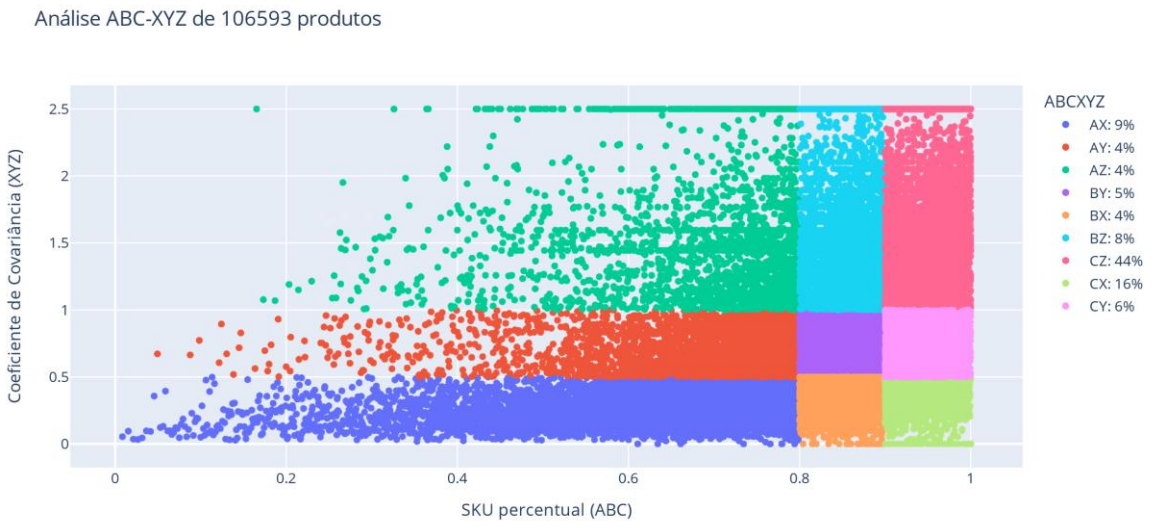

Figura 9. Distribuição dos produtos nas classes ABC-XYZ

4.4. Aprendizado por Regras Associativas

As regras associativas revelaram os padrões de compras de itens em conjunto para determinados subconjuntos de dados. Considerando apenas aqueles produtos que não pertencem às classes C ou X das respectivas classificações ABC e XYZ, foram criadas cestas de compras compostas pelas transações ocorridas no período especificado. Como a ramificação dos produtos únicos é demasiadamente extensa, optou-se por utilizar a categoria mais baixa de especificação dos produtos, os subgrupos, compondo assim regras que observam as compras em conjunto ao longo destes subgrupos.

A Tabela 3 mostra as 20 melhores regras extraídas a partir dos itens frequentes computados pelo algoritmo Apriori, para um suporte acima de 0.1. Apesar do baixo suporte, considerando o alto volume de compras diárias, a confiança destas primeiras regras é acima de 80%, com destaque para a presença de produtos do departamento de hortifrutigranjeiros, perecíveis e *fast-food*. Estes produtos não apenas estão presentes em diversas combinações culinárias, mas também são altamente perecíveis, comprados em poucas quantidades de cada vez e em frequências variadas. Dessa forma, existem muitos cupons que possuem produtos destes departamentos e que acabam sendo combinados com todo e qualquer tipo de produto, o que acaba gerando muitas regras compostas destes produtos.

Tabela 3. Regras geradas para suporte mínimo de 0.05.

Antecedentes	Consequentes	Suporte	Confiança	Lift
LEGUMES, FOLHAGENS FRESCAS REGIONAIS	TUBERCULOS/RAIZ/BULBOS	0.100011	0.815026	3.382152
LEGUMES, FRUTAS FRESCAS NACIONAIS	TUBERCULOS/RAIZ/BULBOS	0.106773	0.811366	3.366967
LEGUMES, TUBERCULOS/RAIZ/BULBOS	FRUTAS FRESCAS NACIONAIS	0.106773	0.7719	2.212685
TUBERCULOS/RAIZ/BULBOS, FOLHAGENS FRESCAS REGIONAIS	FRUTAS FRESCAS NACIONAIS	0.110792	0.762022	2.18437
FRUTAS FRESCAS REGIONAIS	FRUTAS FRESCAS NACIONAIS	0.103751	0.760948	2.18129
LEGUMES	TUBERCULOS/RAIZ/BULBOS	0.138325	0.759805	3.152998
FRUTAS FRESCAS NACIONAIS, FOLHAGENS FRESCAS REGIONAIS	TUBERCULOS/RAIZ/BULBOS	0.110792	0.74021	3.071687
PRATO REGIONAL	FRUTAS FRESCAS NACIONAIS	0.152907	0.725237	2.078924
LEGUMES, TUBERCULOS/RAIZ/BULBOS	FOLHAGENS FRESCAS REGIONAIS	0.100011	0.723016	3.290623
LEGUMES	FRUTAS FRESCAS NACIONAIS	0.131596	0.722846	2.07207

TUBERCULOS/RAIZ/BULBOS	FRUTAS FRESCAS NACIONAIS	0.167318	0.694327	1.990318
TUBERCULOS/RAIZ/BULBOS, FOLHAGENS FRESCAS REGIONAIS	LEGUMES	0.100011	0.68787	3.778404
FOLHAGENS FRESCAS REGIONAIS	FRUTAS FRESCAS NACIONAIS	0.149677	0.681215	1.952733
LEGUMES	FOLHAGENS FRESCAS REGIONAIS	0.122709	0.674029	3.067671
FRUTAS FRESCAS NACIONAIS, TUBERCULOS/RAIZ/BULBOS	FOLHAGENS FRESCAS REGIONAIS	0.110792	0.662167	3.013681
FOLHAGENS FRESCAS REGIONAIS	TUBERCULOS/RAIZ/BULBOS	0.145392	0.661716	2.745956
FRUTAS FRESCAS NACIONAIS, TUBERCULOS/RAIZ/BULBOS	LEGUMES	0.106773	0.638145	3.505265
TUBERCULOS/RAIZ/BULBOS	FOLHAGENS FRESCAS REGIONAIS	0.145392	0.603342	2.745956
LEGUMES	FRUTAS FRESCAS NACIONAIS, TUBERCULOS/RAIZ/BULBOS	0.106773	0.586493	3.505265
TUBERCULOS/RAIZ/BULBOS	LEGUMES	0.138325	0.574013	3.152998

Como proposta de visualização das regras extraídas, considere o grafo tripartido das 10 melhores regras obtidas, exposto na Figura 10. As regras são representadas pelos vértices laranjas no subconjunto central, enquanto que os produtos são representados pelos vértices azuis, sendo os antecedentes presentes no subconjunto superior e os consequentes no subconjunto inferior. As relações de antecedência partem de um ou mais produtos do subconjunto superior em direção à uma regra, enquanto que as relações de consequência partem de uma regra para um ou mais produtos no subconjunto inferior. Assim, podemos atentar para produtos com grande concentração de causas ou efeitos, como no caso exposto do subgrupo de frutas frescas nacionais que é consequente de 6 regras.

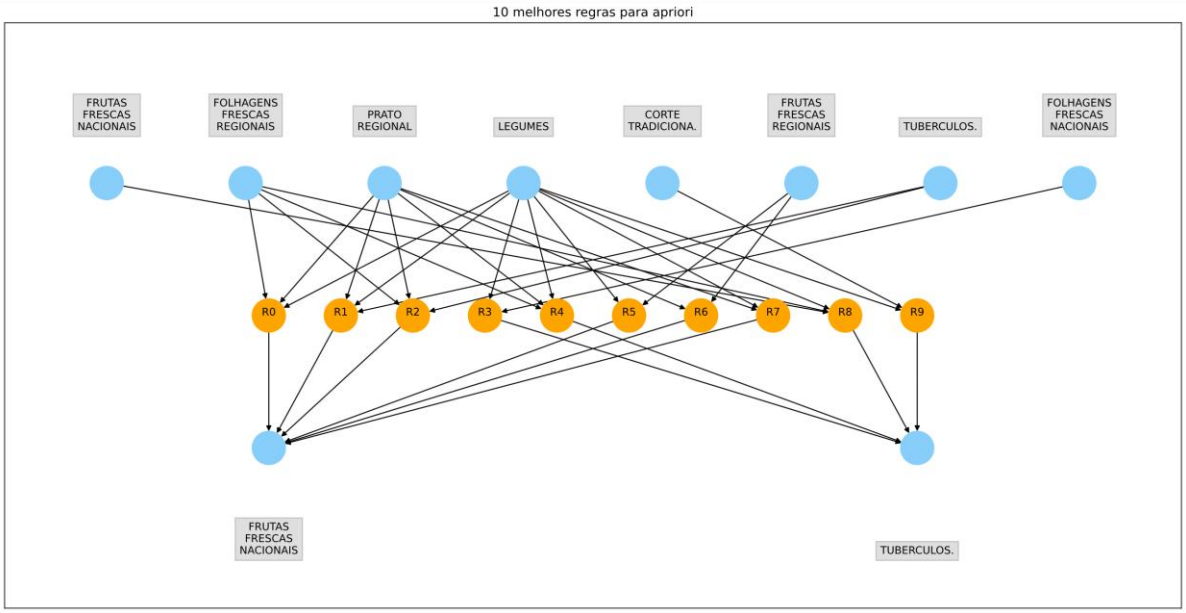

Figura 10. Grafo bipartido das 10 melhores regras geradas pelo algoritmo Apriori.

Foi criada uma segunda cesta, desconsiderando os produtos dos departamentos de hortifrutigranjeiros, perecíveis e *fast-food*, a fim de obter regras com produtos mais variados e que apresentem compras mais estáveis. A Tabela 4 mostra as novas relações de causa e efeito, com suportes e confianças muito menores, mas que revelam compras casadas esperadas como: leite condensado e creme de leite e vice-versa, usados no preparo de doces e sobremesas; filtro de papel e café, assim como açúcar e café; feijão e arroz, componentes indispensáveis na culinária brasileira; guardanapo e toalha de papel e vice-versa, usados em restaurantes e cozinhas; creme dental e sabonete, itens de limpeza básica; e itens de limpeza geral como sabão em pó e detergentes que levam à compra de água sanitária. Apesar da evidente relação entre estes itens estar confirmada nas regras, as regras não-usuais encontradas são o resultado principal a ser apresentado aos representantes do supermercado, a fim de verificar a validade destas mediante as variáveis de tempo, local e contexto

socioeconômico que levam a compra destas combinações, a saber: molho de tomate e creme de leite; arroz e café; detergente e café; óleo de soja e café. Uma vez verificada a relevância destas regras no contexto apresentado, tais informações favorecem a logística de reposicionamento de produtos na prateleira, o levantamento de produtos com desconto, e a combinação de regras com a análise ABC-XYZ para classificação das regras que favorecem a receita e a estabilidade de venda dos produtos.

Tabela 4. Regras sem os produtos perecíveis geradas para suporte mínimo de 0.05.

Antecedentes	Consequentes	Suporte	Confiança	Lift
LEITE CONDENSADO T/P ATE 400G	CREME LEITE T/P ATE 200G	0.02961	0.563838	6.486151
FILTRO PAPEL	CAFE TORRADO COMPENS.ATE 250G	0.010174	0.458172	5.09106
FEIJAO PRETO	ARROZ TIPO-1	0.012118	0.422928	5.592183
GUARDANAPO DE PAPEL	TOALHA DE PAPEL	0.023749	0.39624	5.082534
ALCOOL LIQ NEUTRO	TOALHA DE PAPEL	0.011818	0.349309	4.48056
CREME DENTAL ATE 90G	SABONETE 01G ATE 99G	0.012558	0.347164	4.00507
CREME LEITE T/P ATE 200G	LEITE CONDENSADO T/P ATE 400G	0.02961	0.340617	6.486151
ACUCAR TRITURADO ATE 1KG	CAFE TORRADO COMPENS.ATE 250G	0.020877	0.321621	3.573749
LIMPADOR ATE 500ML	TOALHA DE PAPEL	0.012455	0.320847	4.115477
MOLHO TOMATE PCTE	CREME LEITE T/P ATE 200G	0.010154	0.317144	3.648292
PAPEL HIG NEUTRO F.DUPLA C/4	TOALHA DE PAPEL	0.011176	0.313956	4.027088
TOALHA DE PAPEL	GUARDANAPO DE PAPEL	0.023749	0.304624	5.082534
LAVA ROUPA PO PCTE ATE 500G	AGUA SANITARIA ATE 1L	0.010066	0.296278	5.034847
LIMPADOR ATE 500ML	AGUA SANITARIA ATE 1L	0.011412	0.293972	4.995672
LAVA ROUPA PO PCTE ATE 1KG	AGUA SANITARIA ATE 1L	0.010073	0.281801	4.788832
BATATA FRITA PALHA ATE 190G	CREME LEITE T/P ATE 200G	0.012691	0.273663	3.1481
ARROZ TIPO-1	CAFE TORRADO COMPENS.ATE 250G	0.020462	0.270553	3.006304
OLEO SOJA	ARROZ TIPO-1	0.010954	0.269687	3.565951
LIMPADOR ATE 500ML	CAFE TORRADO COMPENS.ATE 250G	0.010403	0.267985	2.977765
OLEO SOJA	CAFE TORRADO COMPENS.ATE 250G	0.010617	0.261388	2.904466

As 10 melhores regras da cesta livre de produtos perecíveis podem ser visualizadas no grafo tripartido da Figura 11. É possível observar que com um aumento do número de antecedentes e consequentes, a poluição visual também aumenta, tornando cada vez mais difícil encontrar o caminho de partida e destino das retas que compõem as regras. É importante que as próximas interações de melhoria desta visualização otimizem a distribuição espacial dos vértices, organizando grupos de produtos comprados em conjunto de forma que fiquem próximos, alterando a forma como as retas se unem para compor uma regra, e utilizando outras variáveis visuais para trazer informações importantes, como as métricas de confiança e lift, e a categorização dos produtos.

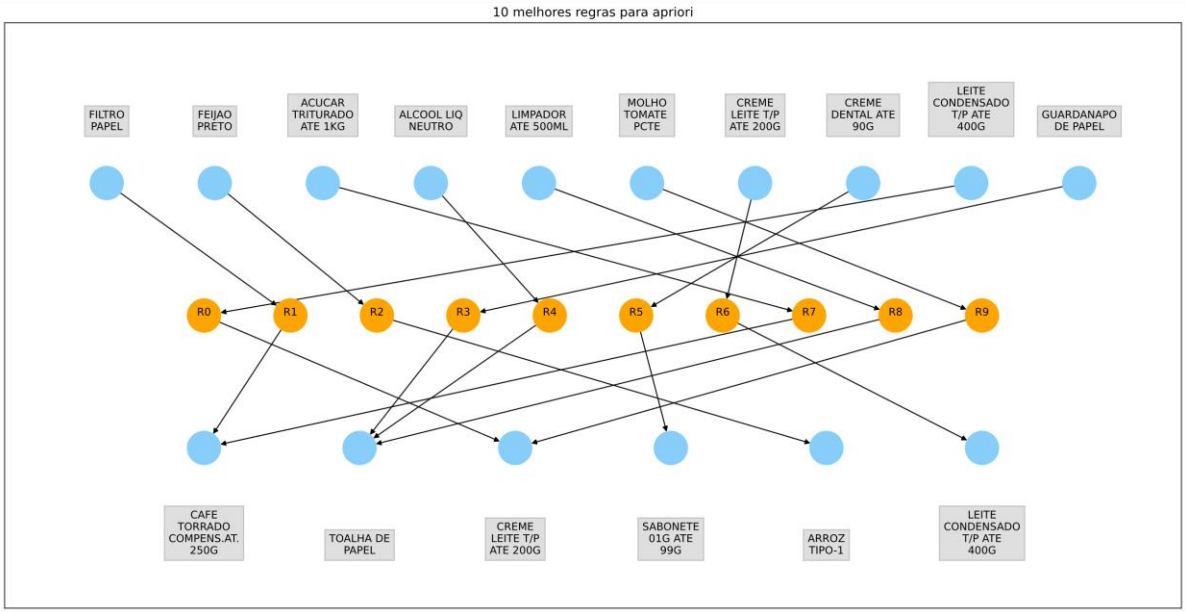

Figura 11. Grafo tripartido das 10 melhores regras geradas pelo algoritmo Apriori sem os produtos perecíveis.

5. Conclusões

A pesquisa desenvolvida focou no desenvolvimento de soluções de mineração de dados personalizadas, capazes de melhorar a eficiência operacional, a tomada de decisões estratégicas e a experiência do cliente na empresa varejista, o que tem o potencial de impulsionar o progresso tecnológico e econômico do estado do Pará. Há uma elevada qualidade na disposição dos dados, uma vez que guardam padrões de consumo de clientes de Belém e das adjacências. É notória a qualidade da base de dados transacional da rede de supermercados em questão, o que facilitou a aplicação das técnicas previstas e acelerou a geração dos resultados.

Como trabalhos futuros, devem ser apresentados os impactos na metodologia de logística e gestão de empresas de supermercados acerca da implantação de ferramentas de mineração de dados em suas rotinas, comparando fatores de impacto como receita, emissão de cupons e rotatividade dos produtos no antes e depois da aplicação destas técnicas. O destaque de tendências de sazonalidade e de irregularidades nas séries temporais de emissão de cupons e receita gerada também deve ser abordada.

Em relação à aplicação de outros algoritmos, *fpgrowth* e *fpmax* não foram capazes de gerar nenhuma regra, utilizando os mesmos parâmetros utilizados na execução do algoritmo *apriori*. A principal premissa destes algoritmos é otimizar a extração dos itens frequentes em cestas que possuem muitos produtos únicos, o que leva um tempo de execução considerável para o *apriori*(Grahne e Zhu, 2003). Mesmo com os pré-processamentos realizados e a substituição da descrição dos produtos únicos pela descrição dos subgrupos únicos, apenas este último algoritmo foi capaz de montar os subconjuntos de itens frequentes necessários para geração das regras associativas. Desta forma, outra importante contribuição para este trabalho seria obtida pelo estudo da extensibilidade destes algoritmos sob a premissa do *Big Data*, explorando as estruturas de dados construídas no processo e otimizando o uso de recursos computacionais.

Referências Bibliográficas

Djukanović, M., Rogic, S., Novicevic, L., Vesna, P.-B., &Jovanovic, M. (2022). *Application of Apriori Algorithm for CRM Improvement - Case Study from Montenegro*. 48–56. https://doi.org/10.1145/3543712.3543733

Grahne, G., & Zhu, J. (2003, January). Efficiently Using Prefix-trees in Mining Frequent Itemsets. *FIMI'03 Workshop on Frequent Itemset Mining Implementations: 2003*.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M. J., &Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8), 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

Schonhorst, G., Paes, V., Balestrassi, P., Paiva, A., & Campos, P. H. s. (2017, July). *Data Mining Association Rules Applied to Supermarket Transactional Data Modeling: a case study in Brazil*.

Tavakoli, M., MolaviHajiagha, M., Masoumi, V., Mobini, M., Etemad, S., &Rahmani, R. (2018). *Customer Segmentation and Strategy Development Based on User Behavior Analysis, RFM Model and Data Mining Techniques: A Case Study*. 119–126. https://doi.org/10.1109/ICEBE.2018.00027

Yoseph, F., Malim, N., Heikkilä, M., Brezulianu, A., Geman, O., &Rostam, N. A. (2020). The impact of big data market segmentation using data mining and clustering techniques. *Journal of Intelligent & Fuzzy Systems*, *38*, 1–15. https://doi.org/10.3233/JIFS-179698