Sistemas de Recomendação Aplicados a Dados de Notas Fiscais

Allan Amorim Cardoso

Universidade Federal do Rio de Janeiro
Instituto de Matemática
Departamento de Métodos Estatísticos

2024

CIP - Catalogação na Publicação

A524s

Amorim, Allan
Sistemas de recomendação aplicados a dados de notas fiscais / Allan Amorim. -- Rio de Janeiro, 2024.
54 f.

Orientador: João Batista de Morais Pereira.
Trabalho de conclusão de curso (graduação) - Universidade Federal do Rio de Janeiro, Instituto de Matemática, Bacharel em Estatística, 2024.

1. Sistemas de Recomendação. 2. Filtragem Colaborativa. 3. Regras de Associação. 4. Supermercados. 5. Comportamento do consumidor. I. Pereira, João Batista de Morais, orient. II. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

Universidade Federal do Rio de Janeiro

Bacharelado em Ciências Atuariais

Rio de Janeiro, 04 de abril de 2024

Ata de defesa de projeto final em Ciências Atuariais

Aluno: Allan Amorim Cardoso **DRE:** 112023722

Banca examinadora:

João Batista de Morais Pereira (orientador) – DME – UFRJ

Hugo Tremonte de Carvalho – DME – UFRJ

Marina Silva Paez – DME - UFRJ

Título do Projeto: *Sistemas de recomendação aplicados a dados de notas fiscais*

Horário da defesa: 13h00

Local: Laboratório de Sistemas Estocásticos (LSE), sala I-044-B, subsolo do bloco I do CT

Após apresentação do projeto final do candidato, a banca examinadora o arguiu, tendo decidido pela sua **APROVAÇÃO** e atribuído grau **10,0** (dez e zero).

Orientador:

_____

João Batista de Morais Pereira

Agradecimentos

Agradeço sinceramente a todas as pessoas que estiveram ao meu lado durante essa jornada de pesquisa e aprendizado. Este trabalho só foi possível graças ao apoio e incentivo de muitos indivíduos.

Em especial, gostaria de expressar minha profunda gratidão ao meu orientador, João Batista de Morais Pereira, que não apenas me orientou de forma excepcional, mas também foi meu professor durante a graduação. Sua dedicação, sabedoria e apoio constante foram fundamentais para o sucesso deste trabalho.

Agradeço também aos demais professores e colaboradores do Instituto de Matemática da Universidade Federal do Rio de Janeiro, que proporcionaram um ambiente acadêmico enriquecedor e desafiador.

Gratidão à minha namorada, Melanie, que foi uma fonte constante de apoio emocional e encorajamento ao longo dessa jornada de pesquisa e aprendizado. Agradeço também aos meus amigos e familiares, cuja presença e incentivo foram fundamentais em todos os momentos. Suas palavras de apoio foram um alicerce sólido durante os desafios mais difíceis, contribuindo significativamente para o sucesso deste trabalho.

Aos colegas de classe e colegas de estudos, agradeço pela troca de ideias, debates enriquecedores e pela atmosfera colaborativa que tornaram esta jornada universitária mais enriquecedora.

Àqueles que forneceram dados cruciais para minha pesquisa, expresso minha sincera gratidão. Suas contribuições foram essenciais para a realização das análises e conclusões deste estudo.

Obrigado a todos que tornaram este projeto possível.

Resumo

Este estudo aborda a análise de recomendações de categorias de produtos em supermercados, explorando a importância da personalização de recomendações no contexto altamente competitivo do varejo de supermercados. A eficácia dessas recomendações desempenha um papel crucial na melhoria da experiência do consumidor e nas estratégias de vendas adotadas por estabelecimentos desse setor.

O estudo concentra-se na análise detalhada das recomendações de categorias de produtos. A coleta de dados inclui informações sobre interações dos consumidores com diversas categorias de produtos de supermercados. Diferentes modelos de recomendação foram aplicados e testados utilizando uma amostra representativa. A investigação do comportamento desses modelos e a escolha do mais apropriado visam proporcionar recomendações mais relevantes e direcionadas, aumentando a probabilidade de compra e a satisfação do cliente.

As implicações práticas das descobertas destacam a relevância dessas estratégias na conquista e retenção de clientes, ressaltando a importância de abordagens personalizadas no setor de supermercados. Além disso, as limitações identificadas durante a pesquisa indicam áreas potenciais para futuras investigações, contribuindo para o contínuo aprimoramento dessas estratégias em ambientes de varejo.

Palavras-Chaves: Sistema de recomendação, Filtragem Colaborativa, Regras de Associação, Supermercados, Comportamento do consumidor.

Abstract

This study addresses the analysis of product category recommendations in supermarkets, exploring the importance of personalized recommendations in the highly competitive context of supermarket retail. The effectiveness of these recommendations plays a crucial role in enhancing the consumer experience and influencing the sales strategies adopted by establishments in this sector.

The study focuses on the detailed analysis of product category recommendations. Data collection includes information on consumer interactions with various supermarket product categories. Different recommendation models were applied and tested using a representative sample. Investigating the behavior of these models and choosing the most appropriate one aims to provide more relevant and targeted recommendations, increasing the likelihood of purchase and customer satisfaction.

The practical implications of the findings underscore the relevance of these strategies in acquiring and retaining customers, emphasizing the importance of personalized approaches in the supermarket sector. Additionally, the limitations identified during the research indicate potential areas for future investigations, contributing to the ongoing improvement of these strategies in retail environments.

Keywords: Recommendation System, Collaborative Filtering, Association Rules, Supermarkets, Consumer Behavior.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

A indústria de supermercados desempenha um papel vital na vida cotidiana das pessoas, fornecendo uma ampla gama de produtos essenciais. À medida que os supermercados expandem suas ofertas para atender às crescentes demandas dos consumidores, surge a necessidade de uma gestão eficaz da maneira como os produtos são apresentados aos clientes. Nesse cenário competitivo, a personalização de recomendações torna-se estratégia-chave para melhorar a experiência do consumidor e impulsionar as vendas.

A personalização de recomendações é uma estratégia bastante útil para os supermercados. À medida que os supermercados coletam dados sobre o comportamento de compra de seus clientes, eles têm a oportunidade de oferecer recomendações personalizadas, sugerindo produtos que são relevantes para os gostos e necessidades individuais de cada cliente.

Muitas empresas fora do setor de supermercados já perceberam os benefícios personalização de recomendações. Gigantes da tecnologia como *Amazon* e *Netflix* são exemplos notáveis de empresas que implementaram com sucesso sistemas de recomendação para melhorar a experiência do usuário e impulsionar as vendas.

No contexto dos supermercados, a aplicação eficaz de sistemas de recomendação pode resultar em uma série de vantagens, incluindo o aumento da satisfação do cliente. Uma dessas vantagens é o estímulo às vendas cruzadas, que se referem à prática de sugerir produtos complementares ou relacionados aos itens que o cliente está comprando, incentivando a compra adicional e aumentando o valor total do carrinho. Outra vantagem é a simplificação do processo de compra, que ocorre ao apresentar ao cliente produtos de interesse de forma personalizada e organizada, reduzindo o tempo e o esforço necessários

para encontrar e selecionar produtos, tornando a experiência de compra mais conveniente e eficiente.

O objetivo geral deste estudo consiste em comparar e fornecer justificativas para as diferenças nos resultados obtidos entre alguns algoritmos de Filtragem Colaborativa e Regras de Associação em sistemas de recomendação. Esse estudo visa avaliar a qualidade das recomendações em um cenário binário, onde há apenas a informação de presença de categorias de produtos em transações (notas fiscais).

Nesse contexto, as categorias de produtos, como Pão, Feijão, Arroz e Refrigerante, são utilizadas como agrupamentos que englobam todos os itens relacionados a cada uma delas. Por exemplo, a categoria Pão abrangerá diferentes tipos e marcas de pães disponíveis no supermercado. O mesmo princípio se aplica às demais categorias. A avaliação visa entender como os algoritmos de Filtragem Colaborativa e Regras de Associação lidam com a presença dessas categorias nas transações, influenciando as recomendações resultantes.

Os objetivos específicos deste estudo são:

- Explorar os conceitos fundamentais de sistemas de recomendação, com foco nas abordagens de Regras de Associação e Filtragem Colaborativa baseada em memória, com estratégias de implementação mais relevantes para o contexto de dados binários.

- Avaliar a eficácia das metodologias na previsão precisa das preferências dos usuários. Isso envolverá a utilização de conjuntos de treinamento e teste.

Ao longo deste trabalho, serão abordadas as metodologias utilizadas, os resultados obtidos e as implicações práticas dessas estratégias. Espera-se que este estudo não apenas esclareça a importância dessas estratégias, mas também forneça orientações valiosas para a escolha de modelos mais adequados a diferentes contextos.

No Capítulo 2, são explorados diferentes tipos de sistemas de recomendação, incluindo filtragem baseada em conteúdo, filtragem colaborativa e suas variantes, como filtragem colaborativa baseada em modelo, memória, usuário e item. O capítulo também discute aspectos como avaliação de modelos, utilizando técnicas como treinamento e teste, validação cruzada e métricas de desempenho.

No Capítulo 3, são apresentadas aplicações práticas dos conceitos discutidos, incluindo a descrição dos dados utilizados, uma análise exploratória desses dados e a implementação de sistemas de recomendação.

Por fim, no Capítulo 4, são apresentadas as considerações finais, enfocando a conclusão dos resultados obtidos pelos sistemas de recomendação aplicados aos dados de notas fiscais. Esta seção aborda não apenas os resultados observados, mas também identifica e explora as limitações identificadas durante o estudo. Além disso, são oferecidas sugestões para possíveis direções de pesquisa futura, visando aprimorar e estender a eficácia dos sistemas de recomendação neste contexto específico.

Capítulo 2

Sistemas de Recomendação

Sistemas de recomendação são mecanismos de filtragem de informações projetados para prever a classificação ou preferência que um usuário atribuiria a um item, como música, livros ou filmes, ou a um elemento social, como pessoas ou grupos, que ainda não tenham considerado (Sharma e Gera (2013)). Essa previsão é realizada por meio de modelos que levam em conta as características do item (abordagens baseadas em conteúdo) ou o ambiente social do usuário (abordagens de filtragem colaborativa). Utilizando técnicas estatísticas e descoberta de padrões, esses sistemas sugerem produtos com base em dados de uso anteriores (Sarwar et al. (2000)). Tais recomendações não apenas auxiliam os clientes a descobrir produtos de interesse, mas também impulsionam vendas adicionais, fortalecendo assim o relacionamento com o cliente.

2.1 Filtragem Baseada em Conteúdo

A abordagem baseada em conteúdo focaliza a análise das características intrínsecas dos itens, utilizando informações específicas sobre cada item para oferecer recomendações personalizadas.

Suponha que um usuário de um serviço de *streaming* de música tenha demonstrado interesse em músicas do gênero pop com vocais femininos. Este usuário, ao explorar o catálogo de músicas disponíveis na plataforma, interage com várias faixas e suas características, como gênero musical e tipo de vocais. Essas interações podem ser capturadas e utilizadas para personalizar recomendações futuras, fornecendo ao usuário uma experiência mais personalizada e relevante.

Para ilustrar essa estratégia, consideremos uma plataforma de streaming de música. Nessa plataforma, temos um conjunto de músicas com diferentes características, como gênero musical e tipo de vocais, e as preferências de um usuário específico são analisados. Na Tabela 2.1, um conjunto fictício de músicas com suas respectivas características é apresentado.

Tabela 2.1: Exemplo de características de itens e preferências do usuário

Música	**Gênero**	**Vocais**
Música A	Pop	Femininos
Música B	Rock	Masculinos
Música C	Pop	Femininos

Utilizando a abordagem baseada em conteúdo, o sistema identificaria músicas com características semelhantes, como gênero e a presença de vocais femininos. Se a Música A é uma escolha anterior do usuário, caracterizada por esses atributos, o sistema poderia recomendar outras músicas com características alinhadas, como a Música C, que compartilha o gênero pop e também apresenta vocais femininos.

Esta estratégia permite ao sistema fornecer recomendações personalizadas com base nas preferências específicas do usuário, utilizando as características das músicas previamente apreciadas.

2.2 Filtragem Colaborativa

A filtragem colaborativa se baseia na ideia de que as preferências e escolhas dos consumidores podem ser inferidas a partir das preferências de outros usuários com interesses semelhantes. Isso possibilita prever avaliações que um usuário atribuiria a itens desconhecidos ou criar listas personalizadas de itens recomendados, conhecidas como listas “top-N” (Sarwar et al. (2001); Deshpande e Karypis (2004)). A ideia principal é que usuários que concordam em avaliações de alguns itens geralmente concordam em avaliações de outros itens.

Os algoritmos de filtragem colaborativa são divididos em dois grupos: filtragem colaborativa baseada em memória e filtragem colaborativa baseada em modelo, conforme descrito por Breese et al. (1998).

As filtragens colaborativas, seja baseada em memória ou em modelo, são abordagens fundamentais em sistemas de recomendação, cada uma com características distintas. A seguir, serão apresentadas essas abordagens e suas peculiaridades.

2.3 Filtragem Colaborativa Baseada em Modelo

A filtragem colaborativa baseada em modelo é uma abordagem que utiliza técnicas estatísticas e algoritmos para analisar padrões nos dados de interação entre usuários e itens. Ao construir um modelo, o objetivo é generalizar as preferências dos usuários e itens, inferindo características a partir dos comportamentos observados dos usuários em relação aos itens.

Um modelo genérico dessa abordagem envolve a utilização de características latentes. Essas características são atributos não diretamente observáveis dos usuários e itens, que são inferidos a partir dos dados disponíveis. Um modelo genérico para predição pode ser expresso pela fórmula

$$\hat{R}_{u,i} = \mu + b_u + b_i + \sum_{f=1}^{k} (p_{u,f} \cdot q_{i,f}), \tag{2.1}$$

em que

$\hat{R}_{u,i}$ é a predição da avaliação do usuário u para o item i.

μ é a média global das avaliações.

b_u e b_i são os termos de viés do usuário e do item, respectivamente.

$p_{u,f}$ e $q_{i,f}$ são as características latentes dos usuários e itens.

k é o número de características latentes.

A Equação 2.1 ilustra como o modelo combina diferentes componentes para gerar uma predição personalizada para cada usuário-item. As características latentes representam atributos abstratos que encapsulam padrões complexos presentes nos dados de interação.

Os dados utilizados no processo de estimativa dos parâmetros são as interações passadas entre usuários e itens, ou seja, as avaliações que os usuários deram aos itens. Essas avaliações podem ser representadas por números, como uma escala de 1 a 5, indicando o grau de preferência ou satisfação do usuário com o item.

O parâmetro $\hat{R}_{u,i}$ representa a predição da avaliação do usuário u para o item i. Esse valor é uma estimativa da classificação que o usuário daria ao item, com base nas características latentes do usuário e do item, bem como nos termos de viés do usuário e do item, e na média global das avaliações.

O número de características latentes (k) no modelo pode afetar sua capacidade de capturar padrões complexos nos dados. Valores menores de k podem levar a uma simplificação do modelo, enquanto valores maiores podem permitir uma representação mais detalhada das interações entre usuários e itens, porém, valores muito altos de k podem levar a ao chamado "*overfitting*", onde o modelo se ajusta muito bem aos dados de treinamento, mas tem dificuldade em generalizar para novos dados de teste.

Ao utilizar métodos como fatoração de matriz, a filtragem colaborativa baseada em modelo consegue lidar eficazmente com a escassez de dados e capturar relações sutis entre usuários e itens. No entanto, a interpretabilidade dessas características latentes pode ser desafiadora, uma vez que elas representam combinações ponderadas de atributos que podem não ter uma correspondência direta com conceitos compreensíveis para os usuários. Mais detalhes podem ser vistos em Koren et al. (2009).

2.4 Filtragem Colaborativa Baseada em Mémoria

A filtragem colaborativa baseada em memória, por sua vez, utiliza diretamente as interações e preferências dos usuários para fazer recomendações. Essa técnica identifica usuários ou itens semelhantes com base nas interações passadas, construindo uma matriz de similaridade. Para recomendar itens a um usuário específico, a abordagem identifica usuários semelhantes e sugere itens que esses usuários semelhantes gostaram, mas que o usuário-alvo ainda não interagiu. A principal vantagem é sua interpretabilidade, já que as recomendações são explicáveis com base nas preferências de usuários semelhantes. No entanto, pode enfrentar desafios em cenários de dados esparsos, onde a matriz de similaridade pode ter muitos valores ausentes.

A principal distinção entre a abordagem baseada em memória e os métodos baseados em modelos é que, nos últimos, um modelo é construído a partir dos dados, com o uso de técnicas como árvores de decisão, modelos de regressão, classificadores de Bayes e modelos de fatores latentes.

Este estudo tratará, em filtragem colaborativa, um dos métodos baseados em memória, especificamente nos métodos de Filtragem Colaborativa Baseada em Usuário e Filtragem Colaborativa Baseada em Item.

2.4.1 Avaliações

No contexto dos sistemas de recomendação, as avaliações desempenham um papel crucial. Essas avaliações referem-se às interações registradas entre os usuários e os itens disponíveis em um sistema, como classificações, pontuações, avaliações numéricas ou feedback qualitativo, e são a base sobre a qual os sistemas de recomendação operam, pois fornecem informações essenciais sobre as preferências dos usuários em relação a itens específicos.

Seja $\mathcal{U} = \{u_1, u_2, \ldots, u_m\}$ o conjunto de usuários e $\mathcal{I} = \{i_1, i_2, \ldots, i_n\}$ o conjunto de itens. As avaliações são armazenadas em uma matriz de avaliações usuário-item de dimensões $m \times n$, denotada como $\mathcal{R} = (r_{jl})$, onde r_{jl} representa a avaliação do usuário u_j para o item i_l. Normalmente, apenas uma pequena fração das avaliações é conhecida, e muitos valores nas células de $\mathcal{R}$ estão ausentes.

A matriz de avaliações $\mathcal{R}$ pode ser expressa como

$$\mathcal{R} = \begin{bmatrix} r_{11} & r_{12} & \ldots & r_{1n} \\ r_{21} & r_{22} & \ldots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \ldots & r_{mn} \end{bmatrix}. \tag{2.2}$$

As avaliações geralmente variam em termos de escala, onde os usuários podem atribuir valores, como classificações de 1 a 5 estrelas, notas de 1 a 10 ou qualquer outra escala apropriada. Além disso, as avaliações podem ser explícitas, ou seja, os usuários fornecem feedback direto sobre um item, ou implícitas, onde as preferências são inferidas com base no comportamento do usuário, como visualizações, compras, cliques, tempo gasto em uma página, entre outros.

Avaliações em dados binários

Em alguns cenários, torna-se útil a possibilidade de adaptar algoritmos de filtragem colaborativa para lidar com dados binários, que representam apenas possíveis interesses

por itens (Mild e Reutterer (2003) e Lee et al. (2005)). Com as metodologias apropriadas, os algoritmos de filtragem colaborativa podem aprimorar a precisão e a relevância das recomendações, mesmo quando os dados são limitados. A representação binária, indicando a presença ou ausência de interesse, oferece uma simplificação na modelagem das preferências do usuário, facilitando a análise e interpretação dos dados.

Poucos estudos estão disponíveis para cenários em que informações de avaliações não estão amplamente acessíveis. Isso é comum quando os usuários evitam expressar avaliações diretas, muitas vezes devido à inconveniência envolvida. Por exemplo, em plataformas de comércio eletrônico, muitos usuários compram produtos, mas não deixam avaliações ou *feedbacks* devido ao tempo e ao esforço necessários para preencher formulários de avaliação detalhados. Nesses casos, as preferências precisam ser deduzidas a partir do comportamento de utilização. Um caso típico ocorre quando se pode rastrear quais produtos são adquiridos por um cliente em um supermercado, mas não se dispõe de informações sobre as razões que levam à não aquisição de outros produtos.

Existem diversas explicações para a ausência de compra por parte do cliente. Pode ocorrer que o cliente não tenha uma necessidade imediata para o produto em questão. Outra possibilidade é que o cliente simplesmente não conheça o produto, embora pudesse apreciar uma recomendação se fosse feita. Além disso, há casos em que o cliente pode não se interessar ou até mesmo não gostar do produto, justificando a eficácia da possível não recomendação do mesmo. Esses diferentes cenários evidenciam a complexidade das decisões de compra, que podem ser influenciadas por fatores como necessidades imediatas, conhecimento do produto e preferências individuais.

Dado $r_{jl} \in \{0, 1\}$, definimos:

$$r_{jl} = \begin{cases} 1, & \text{se o usuário } u_j \text{ tem preferência pelo item } i_l, \\ 0, & \text{caso contrário.} \end{cases} \tag{2.3}$$

Neste trabalho, o foco será dado à análise deste tipo específico de avaliação. Ao compreender as preferências dos usuários expressas por r_{jl}, busca-se o desenvolvimento de estratégias eficazes de recomendação que levem em consideração a complexidade e subjetividade dessas decisões.

2.4.2 Medidas de Similaridade

Antes de explorar as técnicas de filtragem colaborativa baseadas em usuário e item, é crucial que o conceito fundamental de medidas de similaridade entre usuários/itens em um contexto de recomendação seja compreendido. Essas medidas constituem métodos que quantificam a proximidade entre diferentes entidades, como usuários ou itens, com base em características ou comportamentos. A seguir, serão examinadas em detalhes algumas das medidas mais comuns, tais como a Correlação de Pearson, a Similaridade de Cosseno e o Coeficiente de Jaccard, que desempenham um papel vital na identificação de padrões de preferência e correlação entre usuários e itens.

Correlação de Pearson

A Correlação de Pearson é uma medida que avalia a correlação linear entre as avaliações de dois usuários ou itens. Ela fornece uma medida de quão bem as avaliações se ajustam a uma linha reta. A fórmula da Correlação de Pearson entre dois usuários, com índices i e j, é a seguinte:

$$\rho(i,j) = \frac{\sum (r_{ik} - \overline{r}_i)(r_{jk} - \overline{r}_j)}{\sqrt{\sum (r_{ik} - \overline{r}_i)^2 \sum (r_{jk} - \overline{r}_j)^2}}, \tag{2.4}$$

em que

- i e j são índices de dois usuários distintos,
- r_{ik} e r_{jk} representam as avaliações dos usuários de índices i e j para o item de índice k,
- $\overline{r}_i$ e $\overline{r}_j$ representam as médias das avaliações dos usuários de índices i e j.

As definições são análogas para a Correlação de Pearson entre itens.

Similaridade de Cosseno

A Similaridade de Cosseno mede o ângulo entre dois vetores de avaliação, indicando a direção em que eles estão orientados no espaço de avaliação. A fórmula da Similaridade de Cosseno entre dois usuários, com índices i e j, é a seguinte:

$$\cos(i, j) = \frac{\sum r_{ik} \cdot r_{jk}}{\sqrt{\sum r_{ik}^2 \cdot \sum r_{jk}^2}}, \tag{2.5}$$

em que

- i e j são índices de dois usuários,
- r_{ik} e r_{jk} representam as avaliações dos usuários de índices i e j para o item de índice k.

As definições são análogas para a Similaridade de Cosseno entre itens.

Coeficiente de Jaccard

O Coeficiente de Jaccard é uma medida que avalia a sobreposição relativa entre os gostos dos usuários ou itens, indicando o grau de proximidade de suas preferências em relação aos itens. Essa medida é particularmente útil para dados binários, onde os usuários ou itens expressam escolhas (1) ou não escolhas (0). A fórmula do Coeficiente de Jaccard entre dois usuários ou itens, i e j, é a seguinte:

$$Jaccard(i, j) = \frac{|A(i) \cap A(j)|}{|A(i) \cup A(j)|}, \tag{2.6}$$

em que

- i e j são índices de dois usuários ou itens distintos.
- $A(i)$ representa o conjunto de itens avaliados positivamente pelo usuário de índice i ou o conjunto de usuários que avaliaram positivamente o item de índice i.
- $A(j)$ representa o conjunto de itens avaliados positivamente pelo usuário de índice j ou o conjunto de usuários que avaliaram positivamente o item de índice j.
- $|A(i) \cap A(j)|$ representa o tamanho da interseção entre os conjuntos de itens avaliados positivamente pelos usuários de índices i e j ou o tamanho da interseção entre os conjuntos de usuários que avaliaram positivamente os itens de índices i e j.
- $|A(i) \cup A(j)|$ representa o tamanho da união desses conjuntos.

Um valor mais elevado de $Jaccard(i, j)$ indica uma maior similaridade entre os usuários ou itens i e j, sugerindo que eles compartilham preferências semelhantes para os itens.

Em resumo, essas medidas de similaridade podem ser aplicadas tanto entre usuários quanto entre itens, dependendo do contexto da recomendação.

Neste estudo, os cálculos de similaridade serão feitos usando o coeficiente de Jaccard, e portanto, será mais aprofundado a seguir.

Matriz de Similaridade

A matriz de similaridade é uma representação útil para medir a similaridade entre pares de usuários ou itens. Ela pode ser representada como uma matriz em que as linhas e colunas correspondem aos usuários ou itens, e os valores nas células representam as medidas de similaridade entre os pares.

Na Tabela 2.2, vista abaixo, pode ser visto um exemplo para representar uma matriz de similaridade.

Tabela 2.2: Matriz de similaridade

	Usuário 1	Usuário 2	Usuário 3	Usuário 4	Usuário 5
Usuário 1	1.0	0.3	0.2	0.4	0.1
Usuário 2	0.3	1.0	0.5	0.2	0.4
Usuário 3	0.2	0.5	1.0	0.3	0.2
Usuário 4	0.4	0.2	0.3	1.0	0.5
Usuário 5	0.1	0.4	0.2	0.5	1.0

Nesta matriz, os valores nas células representam as medidas de similaridade entre os pares de usuários. Quanto maior o valor, maior a similaridade entre os usuários.

Esparsidade

A esparsidade refere-se à condição em que a maioria dos elementos em uma matriz ou conjunto de dados é representada por valores nulos. Em sistemas de filtragem colaborativa baseada em memória, a presença significativa de valores nulos em matrizes de interações, indicando a falta de avaliações entre usuários e itens, pode ser um desafio. Isso pode impactar a precisão das recomendações, pois a similaridade entre usuários é calculada com base nessas avaliações.

Ao lidar com dados esparsos em um contexto de filtragem colaborativa com métricas como o Coeficiente de Correlação de Pearson, a Similaridade de Cosseno e a Similaridade de Jaccard, é crucial considerar como cada uma delas trata a ausência de interações. O Coeficiente de Correlação de Pearson, por sua sensibilidade a valores ausentes e a necessidade de dados completos para cálculos robustos, pode enfrentar desafios significativos em ambientes esparsos, onde a sobreposição entre avaliações pode ser limitada. A Similaridade de Cosseno, embora seja mais robusta em relação à escala absoluta das avaliações, pode subestimar a similaridade devido à falta de sensibilidade a itens não avaliados.

Em contraste, a Similaridade de Jaccard, projetada especificamente para dados binários, destaca-se na lidança com dados esparsos. Ao focar na presença ou ausência de interação, a Similaridade de Jaccard não é afetada pela escala absoluta das avaliações, proporcionando uma abordagem mais robusta quando a quantidade de interações é limitada.

O objetivo de introduzir a medida de Jaccard é mitigar o problema de "início frio" ("*cold start*") no método de filtragem colaborativa. Um usuário é considerado como "início frio" se uma parcela muito pequena dos itens foi avaliada por ele. Em outras palavras, o termo "início frio" refere-se a um usuário que tem uma participação muito limitada no sistema. A medida de Jaccard é uma ferramenta que pode ser empregada para lidar com essa situação, proporcionando uma abordagem mais robusta ao lidar com usuários que têm interações limitadas no sistema de recomendação.

O desenvolvimento de abordagens de filtragem colaborativa mais eficientes, que consideram todos os vetores de classificação dos usuários, é essencial para melhorar a precisão das recomendações e reduzir o tempo de computação. Como destacado por Bag et al. (2019), a utilização de métricas de similaridade relevantes, como o coeficiente de Jaccard, pode resultar em recomendações mais precisas e eficazes, especialmente em conjuntos de dados relativamente esparsos como as presenças de categorias de produtos em notas fiscais, A esparsidade ocorre neste contexto porque, em um grande catálogo de produtos, a maioria dos clientes compra apenas uma pequena fração dos produtos disponíveis. Isso resulta em uma matriz de dados onde a maioria das entradas são zeros, indicando a ausência de compras para muitos produtos.

2.4.3 Filtragem Colaborativa Baseada em Usuário (UBCF)

Na Filtragem Colaborativa Baseada no Usuário, nossa principal estratégia é identificar usuários semelhantes, com base na sobreposição de itens que eles avaliaram positivamente. Isso nos permite prever as preferências de um usuário com base nas avaliações de usuários semelhantes. A sigla UBCF, do inglês *User-Based Collaborative Filtering*, descreve essa abordagem de recomendação.

Considerando a matriz $\mathcal{R} = (r_{ij})$, que representa as interações de m usuários com n itens, nesta abordagem, define-se A_i como o conjunto de itens avaliados positivamente pelo i-ésimo usuário e $A_i \cap A_j$ como o conjunto de itens avaliados positivamente tanto pelo usuário u_i quanto pelo usuário u_j.

Por exemplo, se o usuário Maria avaliou positivamente os itens A, B e C, representados pelos índices 1, 6 e 12, respectivamente, então $A_{\text{Maria}} = \{1, 6, 12\}$. Suponha que outro usuário, André, tenha avaliado positivamente os itens A, D e E, representados pelos índices 1, 3 e 8, respectivamente. Nesse caso, $A_{\text{André}} = \{1, 3, 8\}$. A interseção entre esses conjuntos, $A_{\text{Maria}} \cap A_{\text{André}}$, conteria apenas o índice 1, indicando que ambos os usuários avaliaram positivamente o item A.

É importante observar que, em cenários de recomendação, é comum que a interseção entre os conjuntos seja vazia para muitas combinações de usuários, devido à natureza esparsa dos dados de avaliação. Isso ocorre principalmente porque os usuários tendem a avaliar apenas um subconjunto limitado dos inúmeros itens disponíveis.

Nesse contexto, tanto as avaliações positivas quanto a ausência de avaliações podem fornecer informações valiosas sobre as preferências dos usuários. A Filtragem Colaborativa Baseada no Usuário leva em consideração essa esparsidade e busca identificar a similaridade entre usuários, levando em conta a sobreposição relativa dos itens avaliados positivamente, bem como a ausência de avaliações em determinados itens.

Após o cálculo das similaridades entre os usuários, o sistema de recomendação pode sugerir categorias de produtos que foram apreciadas por outros usuários com históricos semelhantes. Isso é feito identificando as categorias de produtos que usuários semelhantes avaliaram positivamente e que ainda não foram exploradas pelo usuário alvo. Essas categorias são classificadas e recomendadas com base na ordem de relevância.

Para ilustrar o processo, considere a seguinte matriz de rating binário representada na tabela 2.3, onde os valores indicam se um usuário avaliou positivamente (1) ou não (0)

um determinado item:

Tabela 2.3: Exemplo de Matriz de Preferências de Itens por Usuários

	Item 1	Item 2	Item 3	Item 4	Item 5
Usuário A	0	1	0	1	1
Usuário B	1	1	0	1	0
Usuário C	1	0	1	0	0
Usuário D	0	1	1	0	1
Usuário E	1	0	1	0	0

Suponha, na tabela 2.4, que um novo usuário surge neste sistema, com as seguintes avaliações:

Tabela 2.4: Preferências de Itens pelo Novo Usuário

	Item 1	Item 2	Item 3	Item 4	Item 5
Usuário U	1	0	0	1	0

As similaridades de Jaccard com o novo usuário U serão:

$$Jaccard(U, A) = \frac{|A(U) \cap A(A)|}{|A(U) \cup A(A)|} = \frac{1}{4},$$

$$Jaccard(U, B) = \frac{|A(U) \cap A(B)|}{|A(U) \cup A(B)|} = \frac{2}{3},$$

$$Jaccard(U, C) = \frac{|A(U) \cap A(C)|}{|A(U) \cup A(C)|} = \frac{1}{3},$$

$$Jaccard(U, D) = \frac{|A(U) \cap A(D)|}{|A(U) \cup A(D)|} = \frac{0}{5},$$

$$Jaccard(U, E) = \frac{|A(U) \cap A(E)|}{|A(U) \cup A(E)|} = \frac{1}{3}.$$

Tem-se, assim, as seguintes similaridades apresentadas na tabela 2.5:

Tabela 2.5: Similaridades com o Novo Usuário

	Usuário A	Usuário B	Usuário C	Usuário D	Usuário E
Usuário U	0,25	0,66	0,33	0	0,33

Neste exemplo, o usuário U é o novo usuário para o qual se deseja fazer recomendações. Foram calculadas as similaridades de Jaccard entre o usuário U e outros usuários (A, B,

C, D, E). Os valores representam o quão semelhantes são os conjuntos de itens avaliados positivamente pelo usuário U em relação a esses outros usuários.

O método mais comum para prever as avaliações é focado nos k usuários mais próximos do novo usuário U. Para isso, os usuários (A, B, C, D, E) foram classificados em ordem decrescente de similaridade de Jaccard com o usuário U. Por exemplo, se escolhermos $k = 3$, os três usuários mais similares a U são B, C e E, com similaridades de $\frac{2}{3}$, $\frac{1}{3}$ e $\frac{1}{3}$, respectivamente.

Agora, é possível determinar os itens que esses usuários similares avaliaram positivamente e que ainda não foram explorados pelo usuário U. Esses itens são recomendados a U com base na ordem de relevância.

Dessa forma, o sistema de recomendação aproveita a sabedoria coletiva dos usuários semelhantes, conforme representada pela matriz de similaridade de Jaccard, para enriquecer a experiência de compras do usuário U.

Após o isolamento desses usuários mais similares, é feito o cálculo de previsão dos ratings de categorias. O rating aproximado $\hat{r}_{aj}$ de uma categoria de produto específica para o usuário a é usualmente calculado da seguinte forma:

$$\hat{r}_{aj} = \frac{1}{\sum_{i \in N(a)} s_{ai}} \sum_{i \in N(a)} s_{ai} r_{ij}, \tag{2.7}$$

em que

- $\hat{r}_{aj}$: Rating aproximado da categoria de produto j para o usuário a.
- $N(a)$: Conjunto de usuários mais similares ao usuário a.
- s_{ai}: Medida de similaridade entre o usuário a e o usuário i.
- r_{ij}: Rating real da categoria de produto j pelo usuário i.

Essa fórmula permite calcular uma previsão razoável do rating de uma categoria de produto específica para o usuário a com base nas avaliações dos usuários mais similares. Essas previsões podem então ser usadas para recomendar categorias de produtos ao usuário a.

Voltando ao exemplo, deseja-se calcular as estimativas de avaliação para os itens que não foram avaliados pelo usuário U, ou seja, os itens 2, 3 e 5.

Tem-se então:

$$\hat{r}_{u,2} = \frac{\left(\frac{1}{4} \times 1\right) + \left(\frac{2}{3} \times 1\right) + \left(\frac{1}{3} \times 0\right) + (0 \times 1) + \left(\frac{1}{3} \times 0\right)}{\frac{1}{4} + \frac{2}{3} + \frac{1}{3} + 0 + \frac{1}{3}} \approx 0,58,$$

$$\hat{r}_{u,3} = \frac{\left(\frac{1}{4} \times 0\right) + \left(\frac{2}{3} \times 0\right) + \left(\frac{1}{3} \times 1\right) + (0 \times 1) + \left(\frac{1}{3} \times 1\right)}{\frac{1}{4} + \frac{2}{3} + \frac{1}{3} + 0 + \frac{1}{3}} \approx 0,42,$$

$$\hat{r}_{u,5} = \frac{\left(\frac{1}{4} \times 1\right) + \left(\frac{2}{3} \times 0\right) + \left(\frac{1}{3} \times 0\right) + (0 \times 1) + \left(\frac{1}{3} \times 0\right)}{\frac{1}{4} + \frac{2}{3} + \frac{1}{3} + 0 + \frac{1}{3}} \approx 0,16.$$

Neste caso, portanto, o item 2 se mostrou o mais relevante para o usuário U, e caso apenas um item fosse recomendado, este o seria.

2.4.4 Filtragem Colaborativa Baseada em Item (IBCF)

Nas abordagens de Filtragem Colaborativa Baseada em Item, a similaridade entre os itens desempenha um papel crucial na identificação de itens semelhantes, com base nas avaliações e preferências dos usuários. A ideia é prever as preferências dos usuários em relação a categorias de produtos com base na similaridade entre os itens. A sigla IBCF, do inglês *Item-Based Collaborative Filtering*, descreve essa abordagem de recomendação.

Considere a matriz $\mathcal{R} = (r_{ij})$ que representa as interações entre m usuários e n itens. Nessa abordagem, a similaridade é calculada entre as categorias de produtos, levando em consideração os usuários que avaliaram positivamente essas categorias. Cada linha da matriz deve ser normalizada pela média para calcular a similaridade entre as categorias de produtos.

Para ilustrar esse conceito, considere a matriz de avaliações de categorias de produtos de supermercado, na qual diferentes usuários avaliaram positivamente várias categorias de produtos. Suponha que os usuários Maria, João e Ana avaliaram positivamente a categoria "Arroz". Nesse caso, o conjunto $U_{Arroz} = \{4, 13, 27\}$ representa os índices dos usuários que avaliaram positivamente essa categoria. Além disso, se somente Pedro e Ana avaliaram "Arroz" e "Refrigerante" em conjunto, então $U_{Arroz} \cap U_{Refrigerante} = \{13, 27\}$. É importante observar que, devido à natureza esparsa dos dados de avaliação, a interseção entre conjuntos pode ser vazia em muitos casos.

Após o cálculo da similaridade entre as categorias de produtos, a previsão das avaliações faltantes da categoria de produto de índice i para o usuário de índice a, denotada

como $\hat{r}_{ai}$, é tipicamente baseada nos k itens mais semelhantes à categoria de produto j. O conjunto desses k itens é representado por $\mathcal{S}(i)$ e consiste nas categorias de produtos mais semelhantes à categoria de produto de interesse.

A fórmula para calcular $\hat{r}_{ai}$ é dada por:

$$\hat{r}_{ai} = \frac{1}{\sum_{j \in \mathcal{S}(i)} |s_{ij}|} \sum_{j \in \mathcal{S}(i)} s_{ij} r_{aj}, \tag{2.8}$$

em que

- $\hat{r}_{ai}$: Rating aproximado da categoria de produto i para o usuário a.
- $\mathcal{S}(i)$: Conjunto de itens mais similares ao item i.
- s_{ij}: Medida de similaridade entre o item i e o item j.
- r_{aj}: Rating real da categoria de produto j pelo usuário a.

Nesta fórmula, o somatório ocorre sobre as categorias de produtos que estão presentes no conjunto $\mathcal{S}(i)$. O conjunto $\mathcal{S}(i)$ representa as categorias de produtos mais similares à categoria de produto i com base na similaridade de Jaccard. Essa similaridade possibilita que as recomendações sejam feitas com base no histórico do próprio usuário para categorias de produtos semelhantes.

2.5 Regras de Associação (AR)

As regras de associação são uma técnica de mineração de dados que visa identificar padrões de coocorrência entre itens em conjuntos de dados de transações. Essa técnica é amplamente utilizada em sistemas de recomendação para entender as associações entre itens consumidos por usuários e, assim, gerar recomendações mais precisas e personalizadas.

Sistemas de recomendação que utilizam regras de associação produzem recomendações com base em um modelo de dependência entre itens, definido por um conjunto de regras de associação (Fu et al. (2000); Mobasher et al. (2001); Geyer-Schulz et al. (2002)). A matriz de avaliação binária $\mathcal{R}$ é vista como um banco de dados em que cada usuário é tratado como uma transação que contém o subconjunto de itens em $\mathcal{I}$ com uma classificação de 1. Portanto, a transação $\mathcal{T}_k$ é o conjunto de todos os itens avaliados positivamente pelo

usuário de índice k, e o banco de dados completo de transações é $\mathcal{D} = \{\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_m\}$, onde m é o número de usuários.

Para construir o modelo de dependência, um conjunto de regras de associação é minerado a partir de $\mathcal{R}$. As regras de associação são no formato $X \rightarrow Y$, onde $X, Y \subseteq \mathcal{I}$ e $X \cap Y = \varnothing$. Para selecionar um conjunto de regras de associação úteis, são aplicados limiares em medidas de significância e interesse. Duas medidas amplamente aplicadas são **Suporte** e **Confiança**. Dado Freq($\mathcal{I}$) como o número de transações no banco de dados $\mathcal{D}$ que contêm todos os itens em $\mathcal{I}$:

- **Suporte**:

  O suporte mede a frequência relativa com que um conjunto de itens X e Y aparece nas mesmas transações. O suporte é calculado pela divisão do número de transações (frequência absoluta) que contêm X e Y pelo número total de transações. A fórmula do suporte é:

$$\text{Suporte}(X \rightarrow Y) = \text{Suporte}(X \cup Y) = \frac{\text{Freq}(X \cup Y)}{|\mathcal{D}|}. \tag{2.9}$$

- **Confiança**:

  A confiança mede a probabilidade relativa de que os itens no conjunto Y também estejam presentes quando o conjunto X está presente em uma transação. A confiança é calculada pela divisão do suporte de X e Y pelo suporte de X. A fórmula da confiança é:

$$\text{Confiança}(X \rightarrow Y) = \frac{\text{Suporte}(X \rightarrow Y)}{\text{Suporte}(X)} = \frac{\text{Freq}(X \cup Y)}{\text{Freq}(X)}. \tag{2.10}$$

2.5.1 Algoritmo Apriori

Para desenvolver as regras de associação, existem alguns métodos, onde pode-se destacar um usual com sendo o algoritmo Apriori (Agrawal e Srikant (1994)). Ele utiliza uma abordagem de busca em profundidade para gerar conjuntos de itens candidatos com k elementos a partir de conjuntos de itens com $k-1$ elementos. Em seguida, ele elimina os conjuntos de itens que não são frequentes, ou seja, com baixo suporte.

O algoritmo segue as seguintes etapas:

Etapa 1: Geração de itens de tamanho k

1. Gerar todos os conjuntos de itens L_k de tamanho k, inicialmente $k = 1$, a partir das transações.

Etapa 2: Eliminação de Conjuntos de Itens Infrequentes

1. Para cada conjunto de itens L_k:
   (a) Calcular o suporte do conjunto de itens nas transações.
   (b) Se o suporte for menor que o suporte mínimo, descartar o conjunto de itens.
2. Formar os conjuntos de itens frequentes de tamanho k que atende o suporte mínimo.

Etapa 3: Repetição do Processo

1. Repetir as etapas 1 e 2 para k = 1, 2, 3, ..., até que não seja possível gerar novos conjuntos de itens candidatos ou um limite de tamanho seja alcançado.

Etapa 4: Geração de Regras de Associação

1. Para cada conjunto de itens frequentes:
   (a) Gerar todas as regras de associação possíveis ($X \rightarrow Y$) onde $X \cup Y$ é o conjunto de itens frequentes.

Etapa 5: Cálculo de Suporte e Confiança

1. Para cada regra de associação X $\rightarrow$ Y:
   (a) Calcular o suporte: Suporte($X \cup Y$) / Número total de transações.
   (b) Calcular a confiança: Suporte($X \cup Y$) / Suporte(X).

Etapa 6: Seleção de Regras de Interesse

1. Filtrar as regras de associação com base nos limiares mínimos de suporte e confiança para identificar as regras mais interessantes.

2.5.2 Recomendação

Para fazer uma recomendação para um usuário ativo (u_a) com base no conjunto de itens ($\mathcal{I}_a$) que o usuário gosta e no conjunto de regras de associação ($\mathcal{R}$), os seguintes passos são necessários:

1. Encontrar todas as regras correspondentes $X \rightarrow Y$ para as quais $X \subseteq \mathcal{I}_a$ em R.
2. Recomendar n distintos itens contidos nos conjuntos (Y) das regras correspondentes com as maiores confianças.

Como forma de auxílio na compreensão do processo de criação das regras de associação, considere a representação da figura 2.1 em forma de grafo como a base de dados:

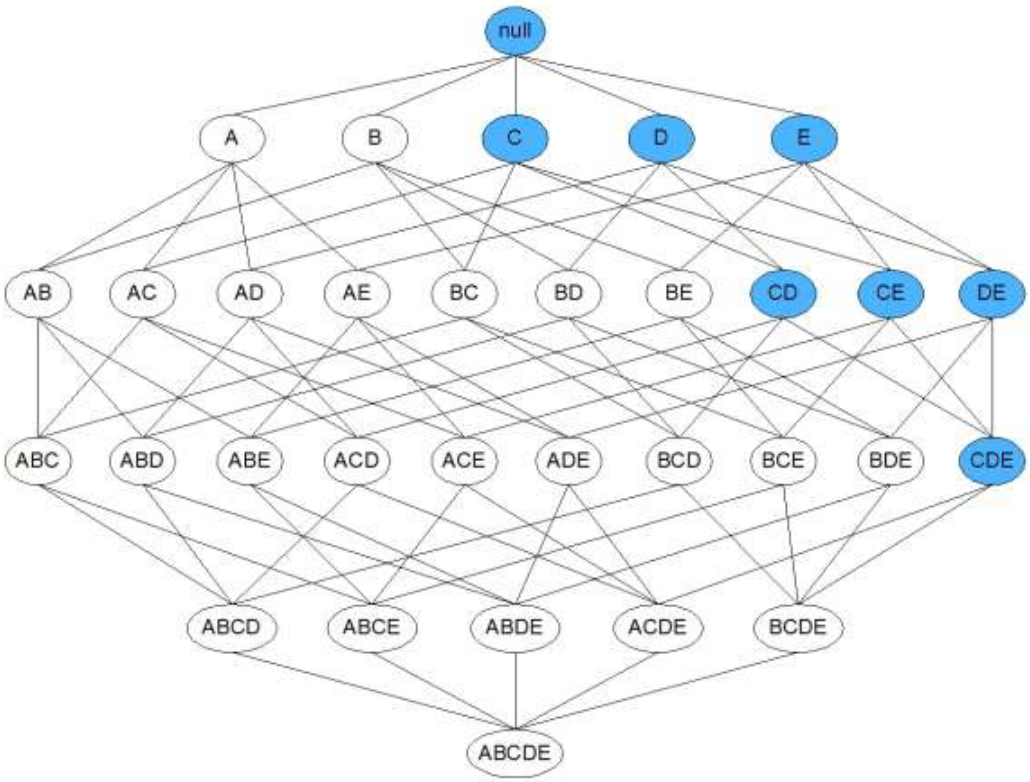

Figura 2.1: Grafo do algoritmo Apriori (itens frequentes).
Fonte: Adaptado de *https://diegonogare.net/2020/05/algoritmo-apriori-para-sistemas-de-recomendacao/* (2020)

Na figura 2.1, A, B, C, D, E são itens e cada nó é um conjunto de itens. Os nós destacados em azul indicam os conjuntos de itens que atenderam aos limiares de suporte e confiança de regras de associação, incluindo C, D, E, CD, CE, DE, CDE, e o conjunto vazio, definido como *null* na imagem.

É importante ter em mente que, se um item é classificado como frequente, isso implica que todos os seus itens associados também são considerados frequentes. No exemplo, o item CDE é considerado frequente, e, como resultado, todos os itens que o precedem nessa sequência também são considerados frequentes.

Por outro lado, se um item não é considerado frequente, todos os itens que o sucedem também não serão considerados frequentes, como pode ser visto na figura 2.2:

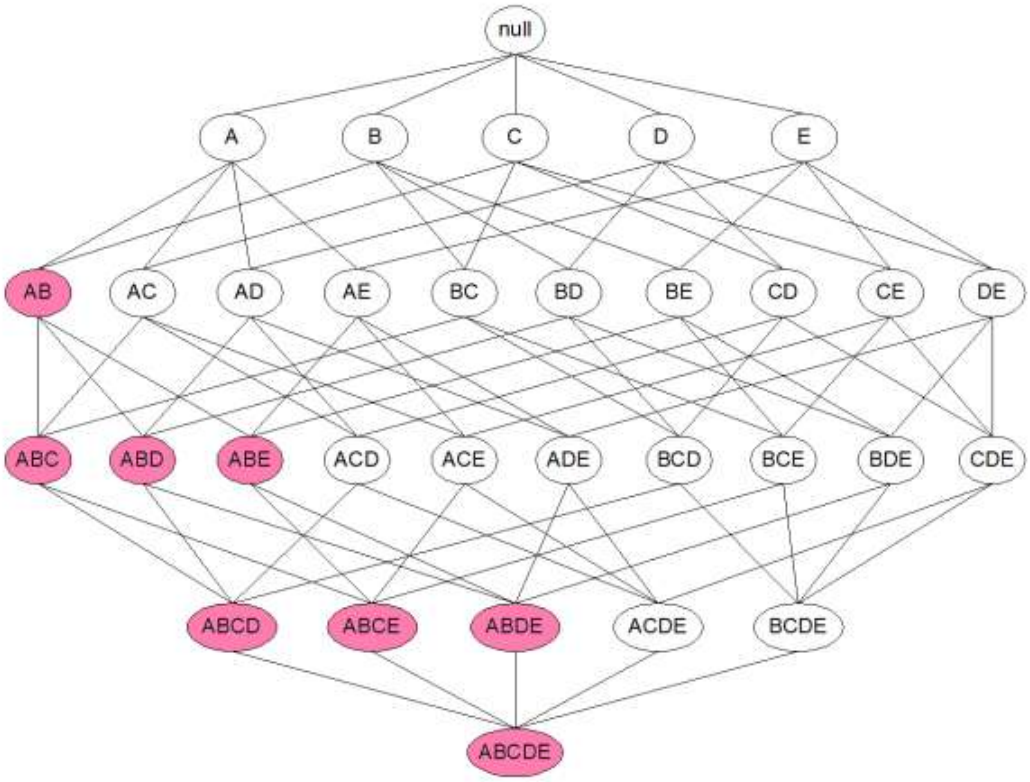

Figura 2.2: Grafo do algoritmo Apriori (itens não frequentes).
Fonte: Adaptado de *https://diegonogare.net/2020/05/algoritmo-apriori-para-sistemas-de-recomendacao/* (2020)

De maneira análoga à figura 2.1, na figura 2.2 tem-se que A, B, C, D, E são itens e cada nó é um conjunto de itens. Os nós destacados em rosa indicam os conjuntos de itens que não atenderam aos limiares de suporte e confiança de regras de associação, incluindo AB, ABC, ABD, ABE, $ABCD$, $ABCE$, $ABDE$ e $ABCDE$.

Seguindo uma abordagem semelhante, mas com uma lógica inversa, a figura 2.2 ilustra o conjunto AB, que não foi classificado como frequente. Portanto, todos os itens derivados de AB também são desconsiderados.

Exemplo de aplicação

Para ilustrar as etapas na construção de regras de associação, um exemplo envolvendo transações de compras de produtos será apresentado na tabela 2.6. Este exemplo contém uma lista de transações com informações de presença de quatro diferentes categorias de produtos. O objetivo é identificar conjuntos de itens frequentemente comprados juntos, o que permitirá a criação de regras de associação úteis para recomendações de categorias produtos.

Tabela 2.6: Exemplo de Transações de Produtos

Transação	Açúcar	Arroz	Feijão	Pão
1	0	1	0	0
2	0	1	0	1
3	0	1	1	1
4	1	1	1	0
5	1	0	0	0
6	1	0	1	1
7	0	1	1	1
8	0	1	1	1
9	1	1	1	0
10	0	1	0	1

A tabela 2.7 todos os conjuntos possíveis para X, juntamente com seus respectivos suportes:

Tabela 2.7: Suportes de X

Conjunto X	Suporte de X
{Arroz}	8/10
{Feijão}	6/10
{Pão}	6/10
{Açúcar}	4/10
{Pão, Arroz}	5/10
{Arroz, Feijão}	5/10
{Pão, Feijão}	4/10
{Feijão, Açúcar}	3/10
{Arroz, Açúcar}	2/10
{Pão, Açúcar}	1/10
{Pão, Arroz, Feijão}	3/10
{Arroz, Feijão, Açúcar}	2/10
{Pão, Feijão, Açúcar}	1/10
{Pão, Arroz, Açúcar}	0/10
{Pão, Arroz, Feijão, Açúcar}	0/10

Agora, é possível calcular a confiança de cada regra $X \to Y$, com $X \cap Y = \varnothing$. Abaixo, as 6 maiores confianças:

1. {Açúcar, Pão} $\to$ {Feijão}:

   Suporte de Açúcar e Pão (suporte(X)): $\frac{1}{10}$

   Suporte de Açúcar, Pão e Feijão (suporte($X \cup Y$)): $\frac{1}{10}$

   Confiança da regra: suporte($X \cup Y$)/suporte(X) $= \frac{1/10}{1/10} = 1$

2. {Açúcar, Arroz} $\to$ {Feijão}:

   Suporte de Açúcar e Arroz (suporte(X)): $\frac{2}{10}$

   Suporte de Açúcar, Arroz e Feijão (suporte($X \cup Y$)): $\frac{2}{10}$

   Confiança da regra: suporte($X \cup Y$)/suporte(X) $= \frac{2/10}{2/10} = 1$

3. {Pão} → {Arroz}:

$$\text{Suporte de Pão (suporte}(X)\text{): } \frac{6}{10}$$
$$\text{Suporte de Pão e Arroz (suporte}(X \cup Y)\text{): } \frac{5}{10}$$
$$\text{Confiança da regra: suporte}(X \cup Y)/\text{suporte}(X) = \frac{5/10}{6/10} = \frac{5}{6}$$

4. {Feijão} → {Arroz}:

$$\text{Suporte de Feijão (suporte}(X)\text{): } \frac{6}{10}$$
$$\text{Suporte de Feijão e Arroz (suporte}(X \cup Y)\text{): } \frac{5}{10}$$
$$\text{Confiança da regra: suporte}(X \cup Y)/\text{suporte}(X) = \frac{5/10}{6/10} = \frac{5}{6}$$

5. {Açúcar} → {Feijão}:

$$\text{Suporte de Açúcar (suporte}(X)\text{): } \frac{4}{10}$$
$$\text{Suporte de Açúcar e Feijão (suporte}(X \cup Y)\text{): } \frac{3}{10}$$
$$\text{Confiança da regra: suporte}(X \cup Y)/\text{suporte}(X) = \frac{3/10}{4/10} = \frac{3}{4}$$

6. {Feijão, Pão} → {Arroz}:

$$\text{Suporte de Feijão e Pão (suporte}(X)\text{): } \frac{4}{10}$$
$$\text{Suporte de Feijão, Pão e Arroz (suporte}(X \cup Y)\text{): } \frac{3}{10}$$
$$\text{Confiança da regra: suporte}(X \cup Y)/\text{suporte}(X) = \frac{3/10}{4/10} = \frac{3}{4}$$

Suponha que uma nova transação, contendo {Açúcar, Pão} esteja em curso, e que se deseja recomendar uma nova categoria. O subconjunto de maior confiança é o próprio {Açúcar, Pão}, com confiança 1, cuja regra contém como consequente o item Feijão. Portanto, comprar feijão será a recomendação.

Considere agora, na mesma situação, que a modelagem definiu como 0,2 mínimo de suporte para as regras de associação. Como a regra {Açúcar, Pão} → {Feijão} tem suporte abaixo do suporte mínimo definido, então essa regra será excluída da busca e a nova regra com maior confiança será {Pão} → {Arroz}, com confiança $\frac{5}{6}$.

Dessa forma, a categoria recomendada será "Arroz", considerando a alta confiança associada à regra {Pão} → {Arroz}.

A definição de um suporte mínimo para recomendações de itens por meio de regras de associação é uma prática benéfica, pois ajuda na redução de regras irrelevantes, evitando a geração de recomendações com base em padrões excepcionais que podem não ser representativos da maioria dos casos. Além disso, ao focar em conjuntos de itens frequentes, a eficiência computacional é aprimorada, o que é particularmente relevante em conjuntos de dados extensos.

2.6 Avaliação de Modelos

2.6.1 Treinamento e Teste

O processo de avaliação de modelos em sistemas de recomendação geralmente envolve dividir o conjunto de dados em conjuntos de treinamento e teste. Essa divisão é fundamental para verificar o desempenho do modelo e sua capacidade de fazer previsões precisas. Os conjuntos de treinamento e teste são criados da forma descrita a seguir.

Conjunto de Treinamento

O conjunto de treinamento desempenha um papel fundamental no desenvolvimento e treinamento de sistemas de recomendação. Este conjunto de dados, que é uma parte dos dados originais, é utilizado para capacitar os modelos a reconhecerem padrões e interações entre usuários e itens.

Durante o processo de treinamento, os modelos UBCF e IBCF utilizam o conjunto de treinamento para compreender as preferências dos usuários em relação aos itens, seja considerando a semelhança entre usuários (UBCF) ou entre itens (IBCF). Por outro lado, o modelo baseado em Regras de Associação (AR) utiliza esses dados para identificar associações frequentes entre itens e, assim, gerar regras que indicam correlações de preferência.

Conjunto de Teste

Este conjunto é composto pela parte restante dos dados originais e é usado para avaliar o desempenho do modelo. O modelo faz previsões com base no que aprendeu no conjunto de treinamento e essas previsões são comparadas com as avaliações reais no conjunto de teste para medir sua precisão.

O processo de divisão entre treinamento e teste deve ser realizado de forma aleatória e estratificada para evitar viés.

Matriz de Treinamento e Teste

Considere o seguinte exemplo hipotético para ilustrar a divisão de dados de treinamento e teste em um cenário de sistema de recomendação. Suponha que uma tabela de interações entre usuários e itens seja utilizada, na qual os itens são representados nas colunas e os usuários nas linhas. Cada valor atribuído às colunas de itens é binário, indicando a presença (1) ou ausência (0) de uma interação entre um usuário e um item.

A tabela 2.8 apresenta um exemplo de uma divisão entre conjuntos de treinamento e de teste.

Tabela 2.8: Tabela de Avaliações de Usuários para Itens

Tipo	Usuário	Item 1	Item 2	Item 3	Item 4	Item 5
Treinamento	Usuário 1	0	0	0	1	0
	Usuário 2	1	0	1	0	1
	Usuário 3	0	1	1	0	0
	Usuário 4	1	0	0	1	1
Teste	Usuário 5	1	1	0	0	1

Nesse cenário, uma divisão a ser feita pode ser separando 80% para treinamento e 20% para teste, onde 1 usuário é destinado ao conjunto de teste, enquanto os outros 4 usuários são fornecidos para o conjunto de treinamento.

A seguir, será discutido um dos métodos de avaliação, que é a validação cruzada, e como ela é usada para avaliar o desempenho dos modelos.

2.6.2 Validação Cruzada

A validação cruzada é uma técnica crucial para avaliar o desempenho de modelos de recomendação. Este método envolve dividir o conjunto de dados em várias partições e executar repetidamente o processo de treinamento e teste em diferentes partições.

A abordagem da validação cruzada *k-fold* (Geisser (1975)) é um dos métodos mais comuns para avaliar o desempenho de modelos de recomendação. No entanto, vale destacar que existem outras técnicas de validação cruzada, como *Leave-One-Out Cross-Validation*

(Stone (1974)) e validação cruzada de Monte Carlo (Xu e Liang (2001)), cada uma com suas vantagens e desvantagens.

Validação Cruzada *k-fold*

Na validação cruzada *k-fold*, os dados são divididos em k partições. O modelo de recomendação é treinado k vezes, onde em cada iteração, $k-1$ partições são utilizadas para treinamento, enquanto a partição restante é reservada para teste. A alternância dos conjuntos de treinamento e teste permite que cada partição atue como conjunto de teste em algum momento do processo.

Essa prática é crucial para avaliar o desempenho do modelo de recomendação de forma justa e abrangente em diferentes tipos de usuários. A utilização de conjuntos de teste específicos para cada iteração evita influências indesejadas e garante que o modelo seja robusto e versátil em cenários diversos.

A validação cruzada *k-fold* proporciona avaliações mais confiáveis, pois aborda a variabilidade nos dados e reduz a possibilidade de resultados enviesados por uma única divisão específica dos dados. Essa abordagem é especialmente útil na identificação de eventuais problemas no modelo, contribuindo para aprimorar sua generalização e desempenho em condições variadas.

2.6.3 Métricas de Desempenho

Métricas de desempenho são parâmetros ou critérios que são usados para avaliar o quão bem um sistema, modelo ou processo está funcionando. No contexto de sistemas de recomendação, medidas de desempenho são usadas para avaliar o quão bem um modelo de recomendação está fazendo previsões precisas e relevantes para os usuários. Essas métricas ajudam a determinar o quão eficaz é o sistema de recomendação na tarefa de sugerir itens aos usuários.

Existem várias métricas comumente usadas, incluindo as descritas a seguir.

Classificações e Valores TP, FP, TN e FN

A classificação de recomendações envolve a avaliação de quatro tipos de classificações, cada um com seu significado:

- **TP (Verdadeiro Positivo)**: Representa os casos em que o modelo classificou corretamente uma recomendação como relevante.
- **FP (Falso Positivo)**: Refere-se aos casos em que uma recomendação foi classificada incorretamente como relevante pelo modelo.
- **TN (Verdadeiro Negativo)**: Indica os casos em que o modelo classificou corretamente uma recomendação como não relevante.
- **FN (Falso Negativo)**: Corresponde aos casos em que uma recomendação foi classificada incorretamente como não relevante pelo modelo.

A tabela 2.9 apresenta um exemplo que demonstra os quatro tipos de classificação em uma matriz.

Tabela 2.9: Classificações de recomendações de filmes

Filme	**Classificação Real**	**Classificação do Modelo**	**Classificação Resultante**
Filme 1	R (Relevante)	R (Relevante)	TP
Filme 2	NR (Não Relevante)	R (Relevante)	FP
Filme 3	NR (Não Relevante)	NR (Não Relevante)	TN
Filme 4	R (Relevante)	NR (Não Relevante)	FN
Filme 5	R (Relevante)	R (Relevante)	TP
Filme 6	NR (Não Relevante)	NR (Não Relevante)	TN

No exemplo da tabela 2.9, os seguintes valores são apresentados:

- TP (Verdadeiro Positivo) = 2: Isso ocorre porque dois filmes foram classificados corretamente como "Relevante" pelo modelo.
- FP (Falso Positivo) = 1: Isso ocorre porque um filme foi erroneamente classificado como "Relevante" pelo modelo, quando na verdade não era.
- TN (Verdadeiro Negativo) = 2: Isso ocorre porque dois filmes foram classificados corretamente como "Não Relevante" pelo modelo.
- FN (Falso Negativo) = 1: Isso ocorre porque um filme foi erroneamente classificado como "Não Relevante" pelo modelo, quando era, na verdade, relevante.

Essas métricas são essenciais para avaliar o desempenho de modelos de recomendação e entender como eles classificam as recomendações em relação às classificações reais.

Taxa de Verdadeiros Positivos (TPR ou Revocação)

A Taxa de Verdadeiros Positivos (TPR), também conhecida como Revocação (*Recall*), é uma métrica que mede a proporção de recomendações relevantes corretamente classificadas como relevantes pelo modelo. Ela é calculada como:

$$TPR = \frac{\text{Recomendações Corretas}}{\text{Total de Itens Positivos}} = \frac{TP}{TP + FN}. \tag{2.11}$$

Taxa de Falsos Positivos (FPR)

A Taxa de Falsos Positivos (FPR) é uma métrica que mede a proporção de recomendações não relevantes erroneamente classificadas como relevantes pelo modelo. Ela é calculada como:

$$FPR = \frac{\text{Recomendações Incorretas}}{\text{Total de Itens Negativos}} = \frac{FP}{FP + TN}. \tag{2.12}$$

Precisão

A Precisão avalia o número de recomendações corretas (TP) feitas pelo modelo em relação ao total de recomendações feitas (TP + FP). É útil quando se lida com conjuntos de classificação de recomendação binária.

$$\text{Precisão} = \frac{\text{Recomendações Corretas}}{\text{Total de Recomendações}} = \frac{TP}{TP + FP}. \tag{2.13}$$

F1-Score

O *F1-Score* desempenha um papel crucial na busca pelo equilíbrio entre Precisão e Revocação, dois aspectos essenciais do desempenho do modelo. O F1-Score é calculado como a média harmônica da Precisão e da Revocação, destacando a importância de identificar corretamente os verdadeiros positivos e encontrar todas as instâncias positivas.

$$\textit{F1-Score} = \frac{2 \times \text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}}. \tag{2.14}$$

Raiz do Erro Quadrático Médio (RMSE)

A Raiz do Erro Quadrático Médio (RMSE) é calculada como a raiz quadrada da média das diferenças ao quadrado entre as previsões do modelo ($\hat{r}_i$) e as avaliações reais (r_i). Quanto menor o RMSE, melhor o modelo, pois as previsões do modelo estão mais próximas dos valores reais, o que indica maior precisão e melhor desempenho do modelo. Em resumo, um RMSE menor significa previsões mais precisas.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{r}_i - r_i)^2}, \tag{2.15}$$

em que i representa o índice de cada recomendação e n representa o número total de recomendações no conjunto de dados de teste.

O RMSE não é a métrica mais adequada para avaliar modelos de filtragem colaborativa em dados binários. O RMSE é frequentemente usado em problemas de regressão, onde as previsões são valores contínuos. Em contraste, a filtragem colaborativa em dados binários geralmente envolve a previsão de classificações binárias (como 0 ou 1), o que torna o RMSE menos apropriado.

Para avaliar modelos de filtragem colaborativa em dados binários, investigar a Curva ROC e a precisão é uma prática mais apropriada.

Curva ROC (*Receiver Operating Characteristic*)

A Curva ROC é uma representação gráfica que avalia a capacidade de um classificador binário, como um modelo de recomendação, em distinguir entre duas classes, por exemplo, as classificações positivas e negativas, em diversas configurações de decisão. Ela é criada plotando a Taxa de Falsos Positivos (também conhecida como Especificidade) no eixo horizontal (eixo x) e a Taxa de Verdadeiros Positivos (ou Sensibilidade) no eixo vertical (eixo y) para diferentes pontos de corte de decisão.

A Curva ROC oferece uma visão visual de como o desempenho do classificador varia quando diferentes critérios de decisão são aplicados. Isso ajuda os desenvolvedores e analistas a entender como o modelo lida com a distinção entre as classes e a escolher o ponto de corte que melhor se ajusta às necessidades do sistema de recomendação. Quanto mais a curva se aproxima do canto superior esquerdo, melhor é o desempenho do modelo na diferenciação das classes.

A figura 2.3 apresenta um exemplo hipotético de uma curva ROC para dois modelos distintos de sistemas de recomendação, nomeados A e B.

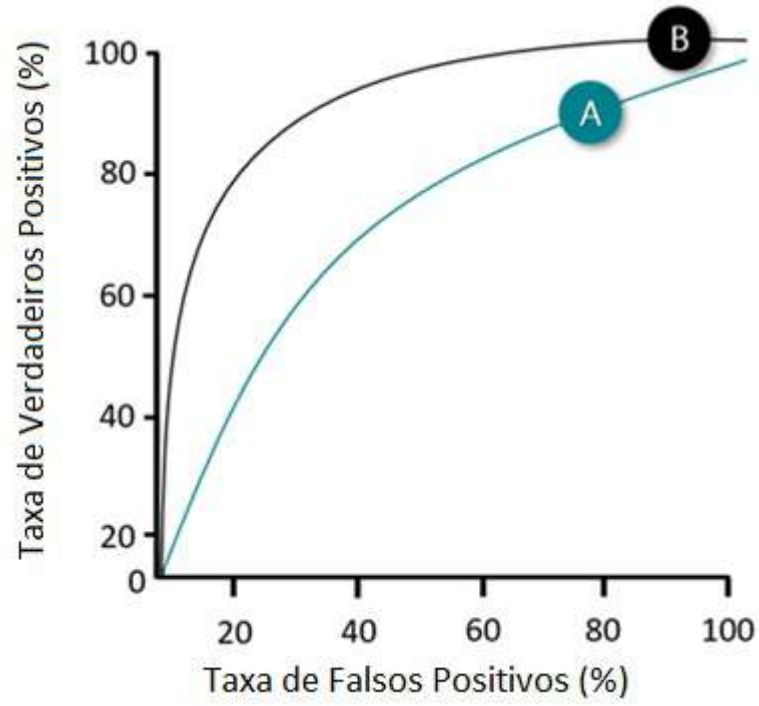

Figura 2.3: Exemplo de Curva ROC para modelos hipotéticos de sistemas de recomendação A e B.

Fonte: Adaptado de *https://med.estrategia.com/public/questoes/Observe-imagem-com19c8279637/*

Capítulo 3

Aplicações

Neste capítulo, os resultados obtidos a partir da aplicação de alguns modelos de Sistemas de Recomendação, discutidos anteriormente, serão apresentados. Esses resultados surgiram da análise do desempenho desses modelos em um conjunto de dados reais, que foi coletado e processado de forma específica para cada abordagem. A maneira como esses dados foram coletados e escolhidos, bem como a forma como foram utilizados para a comparação dos modelos, será explorada em detalhes. Este capítulo concentra-se na aplicação e avaliação desses modelos, destacando seu desempenho nas tarefas específicas de recomendação. Além disso, uma síntese visual das principais descobertas e conclusões extraídas ao longo desta dissertação será fornecida.

3.1 Dados

Nesta pesquisa, os dados foram obtidos em colaboração com uma empresa de inteligência de mercado, reunindo informações sobre presenças de categorias de produtos em notas fiscais não identificadas de supermercados presentes no Estado do Rio de Janeiro. A partir de uma base de dados contendo aproximadamente 60000 notas fiscais, foi realizada uma amostragem aleatória simples (AAS) resultando em um conjunto de 2500 notas fiscais.

Os dados, armazenados em um arquivo de tamanho 1 MB aproximadamente, contém duas colunas com a informação do número de identificação de uma nota fiscal e também a categoria de produto presente na nota.

A amostra contém 335 categorias de produto distintas, e todo a análise foi realizada

através da linguagem de programação estatística e gráfica *R*, com o uso do ambiente de desenvolvimento integrado *RStudio*. O principal pacote usado foi o *RecommenderLab*, um pacote criado para desenvolver e testar algoritmos de recomendação.

3.2 Análise Exploratória de Dados

A análise exploratória dos dados de categorias presentes em notas de supermercado é proporcionada por uma visão aprofundada dos padrões de compra, preferências do consumidor e dinâmicas do mercado varejista. Por meio dessa investigação, busca-se compreender a distribuição e a frequência das diferentes categorias de produtos em notas fiscais, identificando quais itens são mais prevalentes nas compras dos clientes.

A análise exploratória será iniciada examinando a distribuição do número de categorias distintas por nota.

A Figura 3.1 apresenta o histograma correspondente.

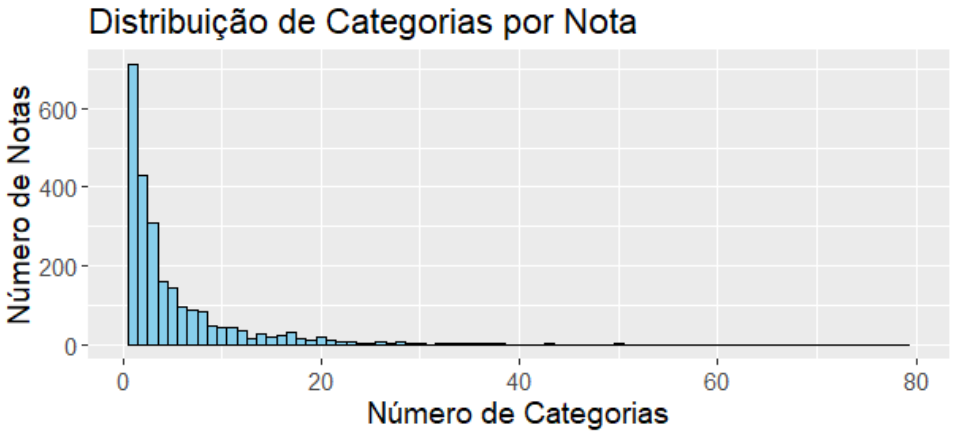

Figura 3.1: Distribuição do número de categorias por nota

A concentração da maioria das notas em faixas menores de itens em supermercados reflete um comportamento de compra caracterizado por necessidades essenciais e hábitos específicos. Um exemplo desse comportamento são as compras matinais, que frequentemente incluem apenas itens básicos, como pães ou leite. Em supermercados, os consumidores frequentemente buscam produtos fundamentais, como alimentos básicos e itens de higiene, resultando em compras mais direcionadas. A natureza prática e eficiente das compras em supermercados, onde os consumidores tendem a repor itens conforme necessário, contribui para a observada concentração em faixas menores de itens.

Tabela 3.1: Dez Categorias mais Populares

Categoria	**Frequência Absoluta**	**Frequência Relativa**
Pão	766	30,6%
Biscoitos	455	18,2%
Refrigerante	382	15,3%
Chocolate	368	14,7%
Banana	290	11,6%
Cebola	285	11,4%
Bovino	265	10,6%
Leite	240	9,6%
Tomate	237	9,5%
Frango	232	9,3%

A tabela 3.1 revela *insights* valiosos sobre os padrões de compra dos consumidores. A categoria "Pão" se destaca como a mais popular, representando cerca de 30,6% das notas, indicando sua presença significativa nas compras. No entanto, a diversidade é evidente, já que categorias como "Biscoitos", "Refrigerante", e "Chocolate" também têm frequências expressivas. Itens essenciais, como "Leite", "Tomate" e "Frango", figuram entre os mais frequentes, sugerindo que produtos básicos são comuns nas compras, apesar de deixar evidente um padrão de consumo que privilegia alimentos ultraprocessados.

3.3 Sistemas de Recomendação

Nesta seção, os resultados obtidos pelos modelos de filtragem colaborativa UBCF e IBCF, juntamente com as Regras de Associação (AR), todos focados em dados binários, serão apresentados. Além disso, as metodologias propostas serão discutidas por meio da comparação do desempenho em dados reais divididos em conjuntos de treinamento e teste. A natureza dos dados será detalhadamente explicada, assim como sua utilização na avaliação dos modelos de recomendação e das regras de associação. O processo de seleção e processamento dos dados será descrito, proporcionando uma compreensão completa de como os modelos UBCF, IBCF e as regras de associação (AR) foram comparados neste contexto. Por fim, um resumo abrangente e visual de todas as descobertas e discernimen-

tos obtidos ao longo desta pesquisa em sistemas de recomendação será apresentado.

Dados de transações com conjuntos de categorias de produtos foram selecionados para demonstrar o desempenho dos modelos e das regras de associação.

Os valores para o número de recomendações, n, a serem geradas nos métodos foram definidos como 1, 3, 5, 10, 15, 20, 30 e 50.

Além disso, a validação cruzada *k-fold* foi usada para treinar todos os modelos, com cinco partições, reservando uma delas (20%) para fins de teste.

Entre as métricas que compõem essa análise estão indicadores essenciais que incluem TPR, FPR, Precisão, F1-Score, a Curva ROC. Como essas métricas são calculadas para cada uma das 5 etapas da validação cruzada, então tirou-se a média aritmética simples dos valores obtidos.

Essas métricas fornecerão uma ampla visão do desempenho dos modelos, permitindo uma compreensão abrangente de como eles realizam previsões e classificações em relação aos dados reais, e ajudarão a avaliar quão bem esses modelos e regras cumprem seus propósitos em termos de recomendações e previsões.

3.3.1 Resultados

Para avaliar o desempenho dos modelos UBCF (Filtragem Colaborativa Baseada em Usuário) e IBCF (Filtragem Colaborativa Baseada em Item), a similaridade de Jaccard foi utilizada para calcular a proximidade entre os usuários e itens, respectivamente. No caso do UBCF, os 50 vizinhos mais próximos foram considerados com base na métrica de similaridade de Jaccard. Quanto ao IBCF, as 50 maiores similaridades para cada item foram levadas em consideração, também utilizando a métrica de similaridade de Jaccard. A escolha por esses valores vem com o objetivo de possibilitar também a recomendação de categorias com poucas aparições e considerar diversas outras transações na modelagem.

Para avaliar o desempenho do modelo AR (Regras de Associação), os limiares para suporte e confiança foram definidos ambos como 0,01. O motivo de utilizar valores baixos para suporte e confiança foi pela alta esparsidade dos dados, o que implica em baixos suportes.

Os valores apresentados nas figuras 3.2 e 3.3 representam as métricas de desempenho dos modelos AR, IBCF e UBCF obtidas através de testes utilizando a técnica de validação cruzada k-fold com 5 partições. Cada valor apresentado é a média das métricas calculadas

em cada uma das 5 partições. Por exemplo, o valor de TPR representa a média das taxas de verdadeiro positivo sobre as 5 partições.

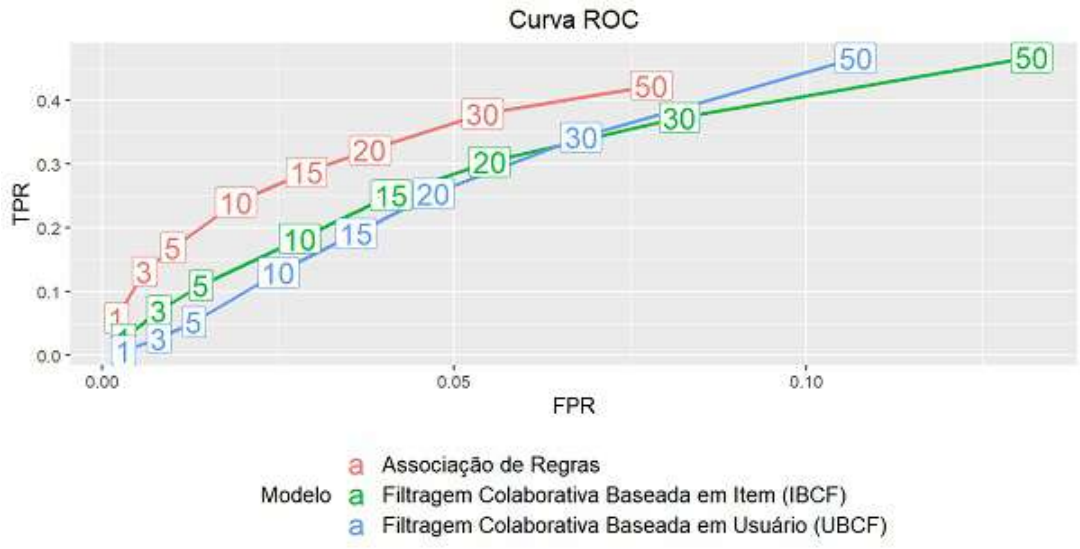

Figura 3.2: Curva ROC dos resultados obtidos pelos modelos AR, IBCF e UBCF

Observando a Curva ROC na figura 3.2, pode-se notar que há diferenças na capacidade dos modelos de distinguir entre exemplos positivos e negativos.

A escolha por não apresentar o gráfico com os eixos exibido até o valor 1 foi feita para que a visualização das curvas se tornasse mais destacada.

Todos os modelos exibem um aumento na TPR à medida que n aumenta. Isso sugere que, independentemente do modelo, à medida que mais recomendações são feitas, eles têm um melhor desempenho na identificação de itens relevantes, o que é um resultado positivo. O TPR do modelo UBCF é menor para pequenas recomendações, porém tem taxa de variação mais acentuada que os demais, fazendo com que o TPR supere o dos demais modelos para um alto número de recomendações. O modelo AR possui valores TPR superiores em pequenas e médias recomendações.

O modelo AR demonstra um desempenho notavelmente superior em termos de FPR, mantendo uma taxa consideravelmente baixa de falsos positivos em comparação com os modelos IBCF e UBCF, e com isso, se torna um modelo eficiente quando for importante não recomendar itens irrelevantes. Em contrapartida, o modelo IBCF tem um alto crescimento de FPR à medida que a quantidade de recomendações aumenta.

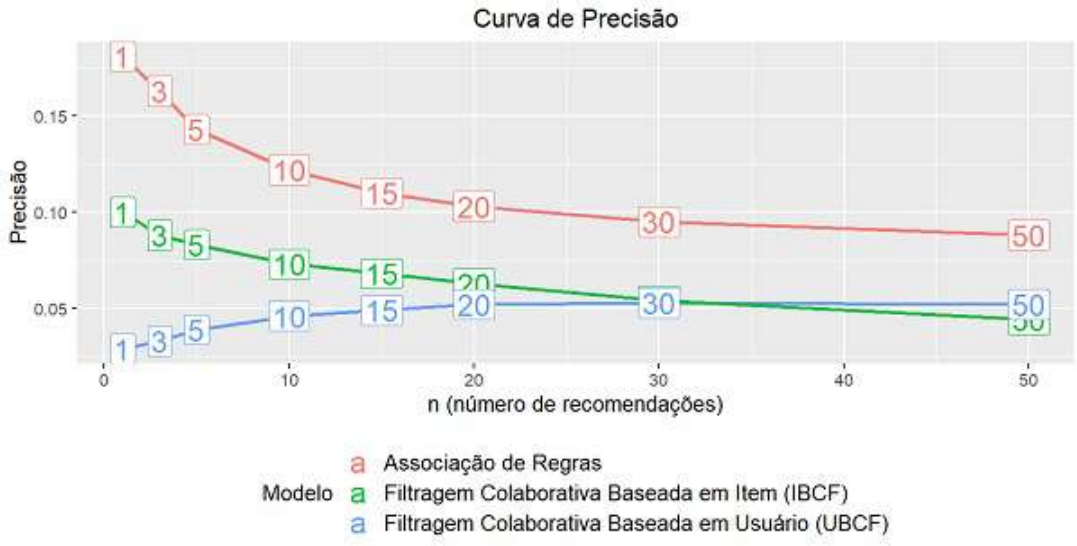

Figura 3.3: Curva de Precisão dos resultados obtidos pelos modelos AR, IBCF e UBCF

A Curva de Precisão revela que, nas recomendações, os modelos AR e IBCF apresentam declínios na precisão a taxas semelhantes à medida que a quantidade de recomendações aumenta. No entanto, o modelo AR mantém uma proporção significativamente maior de itens relevantes nas recomendações. É crucial notar que, ao analisar a curva de precisão, a precisão do modelo UBCF permanece em ascensão, permitindo que ultrapasse a precisão de UBCF para mais de 30 recomendações. No entanto, ao considerar todo o histórico utilizado, o modelo AR demonstrou ser mais preciso.

3.3.2 Exemplos de recomendação para casos específicos

Considere a seguir dois usuários fictícios, Flávio e Marta, fazendo compras online em uma rede de supermercados.

Usuário Flávio

Flávio fez uma compra online em um supermercado, adquirindo produtos das categorias alho e arroz.

As recomendações das principais 3 categorias de produtos para cada método de recomendação são apresentadas na tabela 3.2.

Tabela 3.2: Recomendações para Flávio

IBCF	UBCF	AR
Cebola	Feijão	Feijão
Sabão em Pó	Banana	Cebola
Açúcar	Cenoura	Biscoitos

Ao comparar os resultados dos diferentes métodos de recomendação para Flávio, podemos observar que o método IBCF recomenda categorias como cebola, sabão em pó e açúcar. Essas recomendações podem refletir associações entre categorias frequentemente adquiridas juntas por outros clientes do supermercado.

O método UBCF recomenda categorias como feijão, banana e cenoura, que podem indicar uma tendência geral de compra ou uma preferência comum entre os clientes, possivelmente carregando um perfil de compra para almoço.

O método AR oferece recomendações de feijão, cebola e biscoitos. Percebe-se que o método AR foi capaz de recomendar categorias também sugeridas pelos outros métodos, evidenciando sua capacidade de considerar padrões de compra mais complexos e associações entre produtos. Assim, o método AR pode fornecer recomendações mais diversificadas e adaptadas aos interesses e preferências de Flávio.

Usuária Marta

Marta fez uma compra online no mesmo supermercado, adquirindo produtos das categorias iogurte e leite fermentado.

As recomendações das principais 3 categorias de produtos para cada método de recomendação são apresentadas na tabela 3.3.

Tabela 3.3: Recomendações para Marta

IBCF	UBCF	AR
Biscoitos	Pão	Biscoitos
Creme de Leite	Queijos	Pão
Café em Pó	Biscoitos	Cebola

Ao analisar as recomendações para Marta, podemos observar que o método IBCF

sugere categorias como biscoitos, creme de leite e café em pó. Estes itens podem estar associados às categorias de iogurte e leite fermentado adquiridos por Marta, sugerindo uma tendência de compra comum entre essas categorias.

O método UBCF recomenda categorias como pão, queijos e biscoitos, que podem ser possíveis compras comuns para refeições matinais.

O método AR recomenda biscoitos, pão e cebola. Neste caso, vemos que a categoria biscoitos está presente em todas as recomendações, o que a faz ser uma categoria bem relacionada com as categorias adicionadas ao carrinho, comum em compras similares à de Marta e em compras contendo uma ou mais categorias selecionadas por Marta.

Os métodos UBCF e AR apresentaram recomendações bastante similares, e que possivelmente poderiam atender bem aos interesses de Marta.

Capítulo 4

Considerações Finais

A presente pesquisa teve como objetivo aprimorar a compreensão e a eficácia dos modelos de recomendação, combinando Filtragem Colaborativa Baseada em Usuário (UBCF), Filtragem Colaborativa Baseada em Item (IBCF) e Regras de Associação (AR). A análise das Curvas ROC e de Precisão revelou nuances nas capacidades dos modelos em distinguir exemplos positivos e negativos. Embora o UBCF tenha inicialmente uma Taxa de Verdadeiros Positivos (TPR) menor, sua taxa de variação acentuada o destacou para um alto número de recomendações. No entanto, considerando a natureza binária das avaliações, o AR mostrou-se mais adaptado, mantendo uma proporção significativamente maior de itens relevantes nas recomendações.

Esses resultados sublinham a importância do modelo AR, especialmente para reduzir recomendações irrelevantes e garantir alta precisão em contextos de avaliações binárias. A incorporação do AR não apenas minimiza falsos positivos, mas também maximiza a entrega de recomendações relevantes. Essa abordagem é valiosa ao evitar itens indesejados e identificar eficazmente aqueles alinhados com as preferências dos usuários.

Por fim, a análise abrangente dos modelos UBCF, IBCF e AR ressalta a influência positiva da incorporação de regras de associação no desempenho da filtragem colaborativa. Essa abordagem não apenas proporciona recomendações mais precisas, mas também oferece maior flexibilidade e robustez ao considerar diferentes padrões de preferência dos usuários. Ao capturar padrões específicos de coocorrência entre itens preferidos pelos usuários, o AR se adapta a preferências variadas, proporcionando flexibilidade na representação dos gostos individuais. Essa flexibilidade permite ao modelo AR fornecer recomendações mais ajustadas às nuances das escolhas dos usuários, resultando em uma

abordagem mais precisa e adaptável a diferentes cenários.

4.1 Limitações e Futuras Pesquisas

4.1.1 Limitações do Estudo

As limitações do estudo atual devem ser reconhecidas para proporcionar uma visão precisa das contribuições:

1. **Foco em Filtragem Colaborativa Baseada em Memória**: O escopo do estudo concentrou-se exclusivamente em métodos de recomendação baseados em memória, como a Filtragem Colaborativa Baseada em Usuário (UBCF) e a Filtragem Colaborativa Baseada em Item (IBCF). Portanto, métodos baseados em modelo ou outras abordagens inovadoras não foram considerados, apesar de poderem contribuir para uma visão mais abrangente e diversificada das técnicas disponíveis.

2. **Escalabilidade**: Questões de escalabilidade, uma preocupação importante em sistemas de recomendação, podem não ter sido abordadas completamente, especialmente ao lidar com grandes conjuntos de dados e um grande número de usuários e itens.

4.1.2 Futuras Pesquisas

Com base nas limitações identificadas e no foco em métodos baseados em memória, algumas sugestões para futuras pesquisas nesta área específica são apresentadas:

1. **Melhorias nos Métodos Baseados em Memória**: Um foco nas melhorias dos métodos de recomendação baseados em memória é sugerido, incluindo a exploração de diferentes estratégias de cálculo de similaridade, ponderação de itens e ajustes de hiperparâmetros para otimizar o desempenho dessas abordagens.

2. **Integração de Abordagens Híbridas**: A combinação de métodos baseados em memória com métodos baseados em modelo pode ser considerada para criar sistemas de recomendação mais robustos e precisos. Essas abordagens híbridas podem aproveitar o melhor dos dois mundos para melhorar a qualidade das recomendações.

3. **Consideração de Contexto**: A inclusão de informações contextuais, como preferências temporais e sazonalidade, pode ser explorada para tornar as recomendações

mais relevantes. A pesquisa futura pode focar na incorporação efetiva do contexto nas estratégias de recomendação baseadas em memória.

Referências Bibliográficas

Agrawal, R. e Srikant, R. (1994) Fast algorithms for mining association rules in large databases. Em *Proceedings of the 20th International Conference on Very Large Data Bases*, 487–499. Santiago de Chile.

Bag, S., Kumar, S. K. e Tiwari, M. K. (2019) An efficient recommendation generation using relevant Jaccard similarity. *Information Sciences*, **483**, 53–64.

Breese, J. S., Heckerman, D. e Kadie, C. (1998) Empirical analysis of predictive algorithms for collaborative filtering. Em *Uncertainty in Artificial Intelligence. Proceedings of the Fourteenth Conference*, 43–52.

Deshpande, M. e Karypis, G. (2004) Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems*, **22**, 143–177.

Fu, X., Budzik, J. e Hammond, K. (2000) Mining navigation history for recommendation. Em *Proceedings of the 5th international conference on Intelligent user interfaces*, 106–112. ACM.

Geisser, S. (1975) The predictive sample reuse method with applications. *Journal of the American Statistical Association*, **70**, 320–328.

Geyer-Schulz, A., Hahsler, M. e Jahn, M. (2002) A customer purchase incidence model applied to recommender systems. Em *WEBKDD 2001 - Mining Log Data Across All Customer Touch Points, Third International Workshop*, vol. 2356 de *Lecture Notes in Computer Science LNAI*, 25–47. Springer-Verlag.

Koren, Y., Bell, R. e Volinsky, C. (2009) Matrix factorization techniques for recommender systems. *Computer (Long. Beach. Calif).*, 30–37.

Lee, J., Jun, C., Lee, J. e Kim, S. (2005) Classification-based collaborative filtering using market basket data. *Expert Systems with Applications*, **29**, 700–704.

Mild, A. e Reutterer, T. (2003) An improved collaborative filtering approach for predicting cross-category purchases based on binary market basket data. *Journal of Retailing and Consumer Services*, **10**, 123–133.

Mobasher, B., Dai, H., Luo, T. e Nakagawa, M. (2001) Effective personalization based on association rule discovery from web usage data. Em *Proceedings of the ACM Workshop on Web Information and Data Management (WIDM01)*. Atlanta, Georgia.

Sarwar, B., Karypis, G., Konstan, J. e Riedl, J. (2000) Analysis of recommendation algorithms for e-commerce. Em *Proceedings of the 2nd ACM conference on Electronic commerce (EC '00)*, 158–167. ACM.

— (2001) Item-based collaborative filtering recommendation algorithms. Em *Proceedings of the 10th International Conference on World Wide Web*, 285–295. ACM.

Sharma, L. e Gera, A. (2013) A survey of recommendation system: Research challenges. *International Journal of Engineering Trends and Technology (IJETT)*, **4**, 1989–1992.

Stone, M. (1974) Cross-validatory choice and the assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, **36**, 111–147.

Xu, Q.-S. e Liang, Y.-Z. (2001) Monte carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, **56**, 1–11.