UNIVERSIDADE FEDERAL DE SANTA CATARINA
CTC - CENTRO TECNOLÓGICO
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO E SISTEMAS

Léo Carlos Michel Neto

Desenvolvimento de artefatos para um Sistema de Recomendação de produtos utilizando filtragem híbrida

Florianópolis
2024

Léo Carlos Michel Neto

Desenvolvimento de artefatos para um Sistema de Recomendação de produtos utilizando filtragem híbrida

Trabalho de Conclusão de Curso submetido ao Departamento de Engenharia de Produção e Sistemas da Universidade Federal de Santa Catarina para a obtenção do título de Graduado em Engenharia de Produção Elétrica.
Orientador: Prof. Ricardo Villarroel Dávalos, Dr.

Florianópolis
2024

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

Michel Neto, Léo Carlos
Desenvolvimento de artefatos para um Sistema de Recomendação de produtosutilizando filtragem híbrida / Léo Carlos Michel Neto ; orientador, Ricardo Villarroel Dávalos, 2024.
111 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Engenharia de Produção Elétrica, Florianópolis, 2024.

Inclui referências.

1. Engenharia de Produção Elétrica. 2. Sistemas de Recomendação. 3. Análise de Negócios. 4. Design Science Research. I. Villarroel Dávalos, Ricardo. II. Universidade Federal de Santa Catarina. Graduação em Engenharia de Produção Elétrica. III. Título.

Léo Carlos Michel Neto

Desenvolvimento de artefatos para um Sistema de Recomendação de produtos utilizando filtragem híbrida

O presente trabalho em nível de Graduação foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Guilherme Ernani Vieira, Dr.
Universidade Federal de Santa Catarina - UFSC

Prof. Maurício Uriona Maldonado, Dr.
Universidade Federal de Santa Catarina - UFSC

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de Graduado em Engenharia de Produção Elétrica.

Documento assinado digitalmente
Monica Maria Mendes Luna
Data: 19/02/2024 14:27:57-0300
CPF: ***.104.993-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Mônica Maria Mendes Luna, Dra.
Coordenadora do Curso de Engenharia de Produção

Prof. Maurício Uriona Maldonado, Dr.
Avaliador

Prof. Guilherme Ernani Vieira, Dr.
Avaliador

Documento assinado digitalmente
Ricardo Villarroel Davalos
Data: 19/02/2024 12:05:41-0300
CPF: ***.328.619-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Ricardo Villarroel Dávalos, Dr.
Orientador

Florianópolis, 2024.

AGRADECIMENTOS

Gostaria de expressar minha profunda gratidão a todos que tornaram possível a realização deste trabalho de conclusão de curso.

Primeiramente, quero agradecer à Universidade Federal de Santa Catarina e a todos os professores e servidores que contribuíram para a minha formação. Em especial, aos professores dos Departamentos de Elétrica e Engenharia de Produção e Sistemas, cujo conhecimento e orientação foram fundamentais para o meu crescimento acadêmico.

Ao meu dedicado orientador, Ricardo Dávalos, expresso minha sincera gratidão por guiar meus conhecimentos, compartilhar sua expertise e proporcionar valiosas orientações ao longo deste trabalho. Sua contribuição foi essencial para o meu desenvolvimento acadêmico.

Aos meus pais, Léo F. e Simone, e ao meu irmão, Breno, que sempre me apoiaram e proporcionaram a oportunidade de estudar na UFSC. Não tenho palavras para expressar minha gratidão, seu amor e apoio incondicional tornaram este percurso possível.

À minha companheira, Carolina, dedico um agradecimento especial. Ela esteve ao meu lado durante todos esses anos, compartilhando as vitórias e sendo um apoio constante nos momentos difíceis. Sua presença foi o alicerce que sustentou minha jornada pessoal, acadêmica e profissional.

Às amizades que me estimularam a ingressar na UFSC, como Pedro M., Nevton, Otávio, Pedro G. e Bruno A. Seus incentivos foram o primeiro passo em direção a esta conquista. Também quero agradecer aos amigos que fiz durante minha trajetória na Universidade, como Lucca, André, Bruno P., Thales, Maria Eduarda e Eduardo, cuja amizade e apoio foram inestimáveis.

Por fim, aos colegas de trabalho da Indicium, Igor, Lucas, Bernardo e Luanna, sou grato por me incentivarem a concluir minha trajetória acadêmica e por compartilharem comigo essa jornada.

A todos vocês, meu sincero obrigado por fazerem parte desta jornada e por contribuírem para o meu crescimento acadêmico e pessoal. Este trabalho é dedicado a cada um de vocês, que desempenharam um papel fundamental na minha vida.

“Não é o conhecimento, mas o ato de aprender, não é a posse, mas o ato de chegar lá, que nos dá a maior satisfação.”

(Carl Friedrich Gauss, 1811)

RESUMO

A utilização de Sistemas de Recomendação desempenha um papel importante no contexto do comércio eletrônico, onde a vasta quantidade de dados disponíveis torna desafiador para os usuários encontrar produtos que atendam às suas preferências específicas. Nesse cenário, a implementação eficaz de Sistemas de Recomendação não apenas aprimora a experiência do usuário, mas também desempenha um papel significativo no aumento das taxas de conversão e na fidelização do cliente. O presente trabalho de conclusão de curso explora os conceitos e as técnicas relacionados à mineração de dados para *Business Analytics* e Sistemas de Recomendação, com ênfase na solução de um problema externo da organização *Olist*. O objetivo principal é o de personalizar a experiência dos usuários por meio da instanciação de um modelo de recomendação de produtos. Inicialmente, são discutidas as etapas da mineração de dados, desde a coleta até a interpretação das informações, com foco na conscientização do problema e no levantamento de sugestões. Em seguida, concentra-se na análise de Sistemas de Recomendação, examinando estratégias de coleta de informações do usuário e técnicas de recomendação, incluindo filtragem colaborativa e filtragem baseada em conteúdo. Além disso, a pesquisa aborda a avaliação de Sistemas de Recomendação, considerando métricas como *precision@k*, *recall@k* e pontuação *AUC ROC*. O modelo *LightFM* Híbrido - Produto foi, então, selecionado para compor o artefato de pesquisa uma vez que obteve os resultados mais satisfatórios nas etapas de Avaliação do Artefato. Assim, o estudo estabelece uma base sólida para o desenvolvimento e a avaliação de sistemas de recomendação, a fim de compreender as melhores práticas no campo e resolver o problema da ausência de processo de personalização da experiência do usuário da *Olist*.

Palavras-chave: Sistemas de Recomendação. *Business Analytics*. *Design Science Research*.

ABSTRACT

The use of Recommendation Systems plays an important role in the context of e-commerce, where the vast amount of available data makes it challenging for users to find products that match their specific preferences. In this scenario, the effective implementation of Recommendation Systems not only enhances the user experience but also plays a significant role in increasing conversion rates and fostering customer loyalty. The present undergraduate thesis explores the concepts and techniques related to data mining for Business Analytics and Recommendation Systems, with an emphasis on solving an external problem for the organization Olist. The main objective is to personalize the user experience by instantiating a product recommendation model. Initially, the steps of data mining are discussed from data collection to information interpretation, focusing on problem awareness and suggestion gathering. Then, it focuses on the analysis of Recommendation Systems, examining strategies for user information collection and recommendation techniques, including collaborative filtering and content-based filtering. Furthermore, the research addresses the evaluation of Recommendation Systems, considering metrics such as precision@k, recall@k, and AUC ROC score. The LightFM Hybrid - Product model was then selected to compose the research artifact because of it's most satisfactory results obtained in the Artifact Assessment stages. Thus, the study establishes a solid foundation for the development and the evaluation of recommendation systems aiming to understand best practices in the field and address the issue of the lack of user experience personalization in Olist.

Keywords: *Recommendation Systems*. *Business Analytics*. *Design Science Research*.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Nos últimos anos o comércio eletrônico tem experimentado crescimento acelerado, proporcionando aos consumidores uma ampla variedade de produtos e serviços disponíveis na *internet*. No entanto, a abundância de opções pode levar à sobrecarga de informações, tornando difícil para os clientes encontrarem os itens que realmente desejam.

O advento desse ambiente de comércio do tipo *Business to Customer* (B2C) de forma *on-line*, ou seja, da venda direta do varejo para o consumidor em plataforma eletrônica, representa uma revolução significativa no cenário varejista, alterando a dinâmica tradicional do comércio. Antes da incorporação do *e-commerce*, as operações de varejo eram caracterizadas por um inventário limitado, favorecendo produtos de alto giro e o potencial lucrativo associado ao excesso de estoque.

No entanto, com a introdução do comércio eletrônico, o paradigma do varejo evoluiu para abranger um inventário ilimitado, permitindo a disponibilidade de uma ampla variedade de produtos, inclusive aqueles de nichos específicos que anteriormente poderiam ter enfrentado dificuldades para se estabelecerem no mercado convencional. Esta transição não apenas redefine a abordagem tradicional de estoque, mas também reconfigura fundamentalmente a estratégia de vendas, transformando a forma como os produtos são comercializados e consumidos (CATES, 2019).

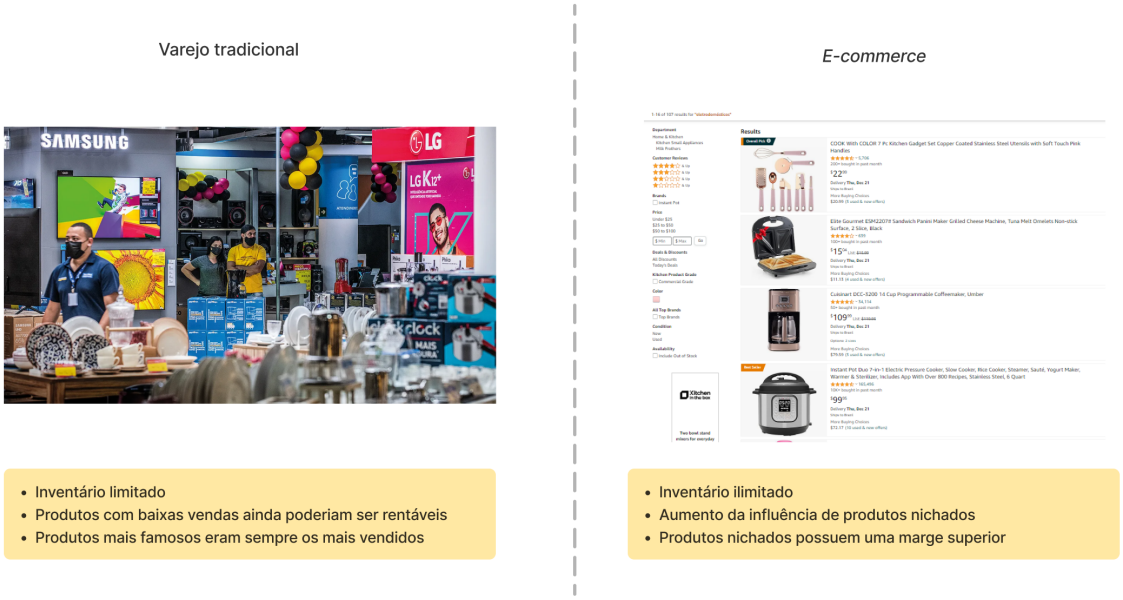
A ascensão do comércio eletrônico não apenas amplia a oferta de produtos disponíveis, mas também reestrutura a dinâmica do mercado, possibilitando a localização facilitada de produtos de nicho. Assim, produtos especializados, anteriormente limitados em visibilidade nas lojas físicas, agora podem ser prontamente encontrados e adquiridos por consumidores interessados. A Figura 1 ilustra as principais diferenças entre as configurações do Varejo tradicional e as do comércio eletrônico.

Em suma, a transição para o comércio eletrônico transpassa a simples alteração de canais de vendas ao representar uma transformação profunda na maneira como os consumidores interagem e transacionam. Assim, redefinem-se as estratégias de *marketing* e de vendas em um ambiente cada vez mais virtualizado e interconectado. Nesse sentido, os Sistemas de Recomendação surgem como uma solução eficaz para ajudar companhias de comércio eletrônico a filtrarem e recomendarem produtos a seus clientes em meio à diversidade de escolhas possíveis (RESNICK; VARIAN, 1997).

Em ambientes de comércio eletrônico, frequentemente é apresentado aos usuários a seção "*Usuários que compram esse produto também compram: [. . .]*", que mostra produtos relacionados ao atual produto que um usuário pode decidir se deseja

realizar a compra ou não. Nessa funcionalidade de endereços eletrônicos de vendas de produtos reside a aplicação de Sistemas de Recomendação.

Figura 1 – Diferenças entre o Varejo Tradicional e Comércio Eletrônico



Fonte: Elaborado pelo Autor

O tema dos Sistemas de Recomendação ganhou relevância no final dos anos noventa à medida que a *internet* se tornava uma ferramenta cada vez mais utilizada para a realização de negócios e compras de alta escala. As partes, muitas vezes, se encontravam em diferentes localidades. Ademais, o comércio eletrônico facilita o processo de coleta de dados ao passo que integra uma interface de usuário que poderia ser utilizada para recomendar itens de forma não invasiva (AGGARWAL, 2016).

Desde então, os Sistemas de Recomendação se tornaram mais reconhecidos como uma ferramenta de personalização da experiência do cliente em um ambiente de compra *on-line*. O estudo desses sistemas é diverso, uma vez que o mesmo permite a utilização de diferentes tipos de dados de preferência e a necessidade de usuários para realizar as recomendações.

Entre as técnicas mais comumente utilizadas nestes sistemas, a filtragem colaborativa é uma das mais amplamente adotada. Essa abordagem se baseia na coleta e na análise de dados sobre as preferências e os comportamentos dos usuários para identificar padrões e semelhanças entre eles. Com base nessas informações, o sistema é capaz de fazer recomendações personalizadas, sugerindo itens que outros usuários com gostos semelhantes adquiriram ou avaliaram positivamente. Logo, a filtragem colaborativa utiliza o conhecimento coletivo dos usuários para melhorar a

precisão e a relevância das recomendações (RESNICK; VARIAN, 1997).

Para executar recomendações, os algoritmos de recomendação do tipo filtragem colaborativa não trabalham com a matriz esparsa em si, mas sim com uma representação em vetores através de fatoração matricial. Assim, é possível obter uma matriz esparsa por meio do produto de duas matrizes de menor grau, sendo uma para os usuários e outra para os produtos, nas quais as linhas destes vetores possuem o fator latente de cada interação vista na matriz esparsa.

Assim, esses vetores são treinados, aproximados das interações originais ao máximo possível por meio do algoritmo de Gradiente Estocástico Descendente e então obtêm-se valores para campos vazios. Ou seja, previsões de aquisição para usuários que ainda não compraram determinado produto. Devido à simplicidade de entendimento e aplicação, esse tipo de técnica foi um fator determinante para a construção do algoritmo de recomendação vencedor do *Netflix Prize*, uma competição de algoritmos criada pela empresa *Netflix* (SARDANA, 2016).

Como será denotado no Capítulo 2, entende-se que em Sistemas de Recomendação, técnicas de filtragem colaborativa são responsáveis por incoporar importantes técnicas à resolução de problemas deste tipo.

Apesar destes pontos positivos, esse tipo de técnica não costuma obter bons resultados quando o inventário de produtos é relativamente grande, quando poucos usuários compraram produtos diferentes ou quando existe uma grande gama de novos usuários. Ou seja, a filtragem colaborativa tende a não ser uma técnica adequada quando a matriz de interação produtos x usários é muito esparsa.

Na contramão desse problema, os modelos de recomendação baseados em conteúdo se destacam. Neles, metadados sobre os usuários e/ou itens disponíveis são coletados - são referidos como *features* de metadados de usuários/itens - e então são gerados aprendizados a partir dessas informações. À título de exemplo, é possível diferenciar os produtos e os usuários por sua categoria, por seu nome, por endereço de residência e por outros atributos encontrados nos dados coletados.

Em suma, esse aprendizado é gerado pela construção de modelos de classificação para cada usuário/item. Por exemplo, se determinado usuário costuma comprar produtos da seção de “Eletrônicos” do comércio eletrônico, o algoritmo de classificação tende a prever recomendações para itens que também pertencem a esta seção. Assim, os modelos de recomendação baseados em conteúdo tendem a empacotar esse conjunto de classificadores para cada usuário.

Apesar disso, esse tipo de técnica apresenta algumas desvantagens. Uma delas é a necessidade de se haver uma grande quantidade de dados para cada usuário. Logo, se um usuário é novo ou possui poucas interações com os produtos da loja, é difícil extrair informações das interações históricas deste cliente para que o modelo possa estimar algo para este mesmo usuário. Outro ponto negativo é

relacionado à impossibilidade de compartilhamento de informações entre os usuários, uma vez que os classificadores são treinados apenas usuário a usuário.

Conforme será extensivamente analisado no Capítulo 2 do presente trabalho, é inviável a aplicação isolada de cada uma dessas técnicas em contextos de poucas informações e interações por usuário ou produto. Para endereçar este problema, emergem os Sistemas de Recomendação híbridos. O principal deles e mais utilizado nesses tipos de situações, foi definido por Maciej Kula no modelo, e pacote *Python*, *LightFM* (KULA, 2015).

Tal modelo é uma variação da fatoração matricial, uma vez que ao invés de estimar fatores latentes para cada usuário e para cada item, os fatores latentes da filtragem colaborativa são estimados com uma agregação dos metadados de usuários e de itens, advindos da filtragem baseada em conteúdo.

Por exemplo, em um método de fatoração matricial comum, estima-se o fator latente para o produto “Televisão”, para então, talvez, recomendar esse produto ao cliente. Já em como foi definido o modelo *LightFM*, tem-se uma estimação de fator latente para a combinação “Televisão”, categoria “Artigos de Esportes” e localidade “São Paulo” do usuário, ou outras combinações relevantes.

O benefício de utilizar essa abordagem reside no fato de que, em diversas situações, existirão poucas distintas *features* de metadados nos contextos acima abordados. Dessa forma, a matriz de interação, ou seja, os dados coletados das interações dos usuários ao longo do tempo nos produtos disponíveis, será usada para estimar um modelo com menos parâmetros.

Outro ponto positivo é que os próprios identificadores de usuários e de itens também podem ser incluídos como metadados. Por exemplo, para o valor” Televisão” do respectivo exemplo, o qual possui um identificador único “16688974”, é possível estimar uma representação para uma soma desses parâmetros.

Neste contexto de transição do varejo e de problemas de recomendação de produtos a clientes, a empresa *Olist*, uma companhia de comêrcio eletrônico de eletroeletrônicos e artigos variados, deseja implementar um sistema de recomendação de produtos na sua plataforma de vendas *on-line,* com o objetivo central de aumentar as vendas de seus produtos. Isso pode ser possível a partir da personalização, atração e fidelização de clientes nas quais os Sistemas de Recomendação visam proporcionar.

Uma de suas soluções, a *Olist Store*, conecta consumidores e revendores de produtos diversos em um *marketplace*. Em resumo, os revendores da *Olist* cadastram produtos na plataforma que encarrega-se de disponibilizá-los a potenciais consumidores. Entende-se que a proposição de valor ao cliente através de relacionamentos individuais não é uma tarefa simples e exige profissionais capacitados, tecnologia da informação e algoritmos sofisticados. Nota-se tal escalabilidade na empresa em questão, que possui em seu portfolio mais de 10.000

itens disponíveis aos milhares de clientes que acessam a plataforma em rempo real.

Frente ao crescimento do *e-commerce* e ao respectivo aumento da concorrência, ou seja, novas companhias que cada vez mais aderem a esse canal, se mostra essencial atender as expectativas da base de clientes. Frequentemente, profissionais de *marketing* visam atender grupos mais específicos e para isso, seus planos de ação eventualmente estão relacionados aos desenvolvimentos de sistemas de Gestão Estratégica de Tecnologia da Informação. Nesse contexto, as empresas optam, portanto, por adotar soluções personalizadas em detrimento de soluções padronizadas (ZENONE, 2007).

Finalmente, fica evidente que as empresas nascidas digitais necessitam de apoio em sistemas que sejam capazes de automatizar atividades de atração e retenção dos clientes, de modo a oferecer tratamento personalizado, destacando-se frente à concorrência (VENSON, 2002).

Diante da complexidade da tarefa de implementação de um Sistema de Recomendação personalizado no ambiente de comércio eletrônico da empresa *Olist*, mostra-se adequado adotar uma abordagem de pesquisa robusta e estruturada. A *Design Science Research* apresenta-se como um método apropriado para abordar tal problema, uma vez que combina princípios científicos com técnicas de *design* para criar soluções satisfatórias em um contexto empresarial (DRESCH, 2015).

A primeira etapa da *Design Science Research* envolve a análise do contexto e dos requisitos do sistema. Nesse caso, a empresa *Olist* precisa entender as preferências de seus clientes, os tipos de produtos que estão disponíveis em sua plataforma e os desafios específicos que enfrentam no fornecimento de recomendações personalizadas.

Em seguida, durante a fase de *design*, o objetivo é aplicar algoritmos de recomendação personalizados com base nas informações coletadas na fase de análise. Esses algoritmos podem ser adaptados para levar em consideração as preferências individuais dos clientes, o histórico de compras e outros dados relevantes, geralmente a partir de um modelo matemático já construído e disponível para a comunidade acadêmica.

À medida são que implementados, a *Design Science Research* permite a coleta de dados sobre o desempenho das recomendações e o *feedback* dos usuários, por meio de técnicas de avaliação de artefato. Essa abordagem, baseada em evidências, garante que as soluções sejam iteradas e validadas por um viés quantitativo.

Logo, é possível denotar que a aplicação *Science Design Research* deve resultar na criação e na utilização de diversos tipos de artefatos que podem variar desde modelos conceituais a conceitos do campo de conhecimento de Sistemas de Recomendação. Assim, a instanciação de um Sistema de Recomendação de produtos surge como o artefato-chave do presente estudo. Através desse artefato, será possível

traduzir os princípios da metodologia em uma solução tangível e funcional que atenda às necessidades da empresa *Olist* e de seus clientes.

Portanto, a metodologia emerge como uma solução que embasa a resolução da falta de personalização das recomendações no comércio eletrônico da empresa *Olist*. Através da análise, do *design*, da implementação e da avaliação estruturados da *Design Science Research*, a empresa pode alcançar seus objetivos de proporcionar uma experiência personalizada aos clientes e aumentar as vendas de seus produtos.

Por fim, o presente estudo visa solucionar satisfatoriamente o problema da ausência de personalização da experiência de usuários no comércio eletrônico da empresa *Olist.* Tal resolução se dará a partir do desenvolvimento de um artefato de pesquisa do tipo instanciação de um Sistema de Recomendação de produtos.

1.1 PROBLEMÁTICA DO ESTUDO

Nas técnicas de personalização de oferta de produtos aos usuários, as organizações utilizam o conhecimento que possuem sobre o cliente para proporcionar uma melhor experiência de compra, oferecendo itens que de fato são relevantes a determinado usuário que adentra ao *website* da loja (VENSON, 2002).

A ausência de personalização nas recomendações de produtos da loja *Olist* representa um obstáculo que pode afetar diretamente o sucesso e a competitividade do negócio. No contexto atual do comércio eletrônico, os consumidores estão expostos a um volume diverso de opções, o que torna a personalização das recomendações um elemento-chave para aumentar o engajamento do cliente e a conversão de vendas.

A falta desse aspecto pode resultar em experiências de compra menos satisfatórias para os usuários, uma vez que estes são abordados com produtos que não são relevantes para suas preferências e necessidades. Esse desafio não apenas impacta negativamente a satisfação do cliente, mas também pode resultar em perda de oportunidades de vendas, já que os consumidores podem optar por procurar produtos em outras plataformas de comércio eletrônico (ZENONE, 2007).

Portanto, a questão da personalização das recomendações se mostra como uma problemática crítica que exige atenção e soluções eficazes para impulsionar o sucesso e a competitividade da loja *Olist*. Assim, Sistemas de Recomendação são utilizados pelos comerciantes a fim de melhorar seus resultados financeiros.

Entretanto, tais sistemas só conseguem reter um potencial cliente se as suas recomendações possuírem, de fato, itens relevantes. Dessa forma, recomendar um produto adequadamente aumenta as chances de retenção de clientes, bem como o faturamento da empresa.

Outro ponto essencial a ser endereçado no presente trabalho está relacionado à capacidade humana de tomada de decisão frente a uma gama diversa de opções a serem escolhidas. Com o objetivo de analisar o comportamento dos usuários diante de

diferentes configurações de *mix* de produtos, um estudo realizado pela Universidade de *Harvard* organizou um experimento científico com duas barracas de venda de geléia para torradas, abertas ao público (IYENGAR; LEPPER, 2000).

A primeira barraca continha 24 opções de produtos de cada tipo, como dietéticos, reduzidos em calorias totais, com mistura de sabores, entre outros. Já a segunda barraca continha apenas seis variações distintas de tipos de geléia. Em um primeiro momento, ficou nitído que a barraca com mais opções atraiu mais clientes para dentro da loja.

Todavia, os autores conseguiram demonstrar que uma alta disponibilidade de produtos não significa um potencial de vendas superior. Apesar da barraca com maior disponibilidade de geleias ter atraído mais clientes, a taxa de conversão de clientes, ou seja, o percentual de clientes que entraram na barraca e compraram uma geleia, foi de 30% para a barraca com apenas 6 opções, enquanto que a barraca diversa obteve apenas 3% de conversão dos clientes

Analogamente, ao se adaptar o resultado desse experimento para o contexto de vendas de produtos em meio eletrônico, é possível notar que uma elevada variedade de produtos ao consumidor pode ser, na verdade, um fator de desmotivação para a aquisição de um produto.

Assim, entende-se que os dois grandes pilares de resolução do problema de pesquisa estão pautados na retenção de compras e na customização individual do *mix* de produtos ofertados aos usuários no ambiente de compra on-line. Com isso, o presente trabalho de conclusão de curso deve buscar solucionar satisfatoriamente o problema de ausência de personalização da experiência dos usuários de comércio eletrônico, utilizando-se das metodologias e técnicas mais apropriadas ao contexto da organização Olist.

Embora o objetivo principal de um sistema de recomendação seja aumentar a receita do comerciante, isso é frequentemente alcançado através de métodos não triviais em um primeiro momento. Para alcançar o objetivo empresarial de aumento de receita, Aggarwal (2016) lista os principais objetivos operacionais e técnicos comuns dos Sistemas de Recomendação, os quais também são inerentes à este estudo:

a) Relevância: Recomendar itens que sejam de fato relevantes para determinado cliente;

b) Novidade: O SIstema de Recomendação é pouco efetivo se o usuário já teve contato com o produto recomendado;

c) Serendipidade: Gerar um efeito-surpresa no usuário, estimulando um sentimento de que determinado produto já era querido ou necessário;

d) Aumento da diversidade de recomendação: Gerar recomendações que sejam diversas em sua composição;

Nesse sentido, a instanciação do artefato de Sistema de Recomendação permitirá definir um modelo para a customização das recomendações de produtos de acordo com as preferências dos usuários, contribuindo assim para a personalização da experiência no comércio eletrônico.

Assim, o presente trabalho de conclusão de curso visa desenvolver um artefato de pesquisa do tipo instanciação, na forma de aplicação de algoritmo em modelo de sistema de recomendação. Esta solução deve ser adaptada à estrutura de processos e de propósito da organização em questão, contribuindo para a geração de sugestões de compras personalizadas para os clientes de uma loja virtual, e portanto, para o aumento da receita e do volume de vendas dos produtos comercializados.

1.2 OBJETIVOS

Nestas subseções, são estabelecidos os propósitos e as metas a serem alcançados no presente trabalho. O Objetivo Geral está alinhado com a problemática identificada anteriormente e reflete o impacto que a pesquisa pode ter no campo de estudo. Enquanto isso, os Objetivos Específicos detalham as etapas, as tarefas e as submetas necessárias para atingir o Objetivo Geral. Nas seções posteriores são elencados tais elementos.

1.2.1 Objetivo Geral

Desenvolver artefato tipo instanciação de Sistema de Recomendação personalizada de produtos para clientes em comércio eletrônico.

1.2.2 Objetivos Específicos

a) Descrever os processos de desenvolvimento e avaliação de Sistemas de Recomendação;
b) Aplicar técnicas de recomendação em um *dataset* de feedback implícito;
c) Avaliar a performance dos modelos matemáticos de recomendação de produtos;
d) Gerar recomendações personalizadas de compra de produtos para clientes da base de dados da empresa;

1.3 JUSTIFICATIVA DO ESTUDO

Nos últimos anos, devido ao crescente aumento do volume de dados armazenados, os sistemas e de competências de gestão das organizações têm sido alvo de novos desafios relacionados com a criação de conhecimento específico,

tornando-se imprescindível o recurso às TI no processo de transformação dos dados em conhecimento (DUARTE, 2018).

Os novos desenvolvimentos em TI são importantes para todas as disciplinas de negócios, pois estas desencadeiam mudanças no mercado, nas operações, no comércio eletrônico, na logística, nos recursos humanos, no setor financeiro, na contabilidade e no relacionamento com consumidores e de parceiros de negócios. Dessa forma, identificar quais fontes de dados estão disponíveis, que tipo de dados eles fornecem e como tratar esses dados é fundamental para gerar o máximo valor possível para a empresa (TURBAN; VOLONINO, 2013).

É possível evidenciar a relevância da utilização de Sistemas de Recomendação em empresas nascidas no meio digital ao analisar o estudo de caso da *Netflix*, que contribuiu significativamente para a comunidade de pesquisa com resultado do concurso *Netflix Prize* (AGGARWAL, 2016). Este concurso foi projetado para fornecer um fórum de competição entre vários algoritmos de filtragem colaborativa contribuídos pelos participantes. Um conjunto de dados com classificações de filmes da *Netflix* foi disponibilizado e a tarefa era prever as classificações de combinações específicas de usuário e item.

Os prêmios foram concedidos com base na melhoria do próprio algoritmo de recomendação da *Netflix*, conhecido como *Cinematch*, ou pela melhoria da pontuação anteriormente alcançada acima de um certo limite. Muitos algoritmos de recomendação conhecidos, como modelos de fatores latentes, que serão avaliados no presente trabalho, foram popularizados pelo concurso da *Netflix*.

Pro fim, de acordo com Venson (2002), os algoritmos de sistema de recomendação surgem nesse contexto como uma importante contribuição na personalização da experiência de clientes no comércio eletrônico. Por meio da eloboração de *websites* mais representativos e personalizados ao usuário frente à tela do dispositivo, as possibilidades de conversão desse potencial cliente para comprador podem aumentar significativamente.

A justificativa para a utilização de diferentes técnicas de recomendação, como a colaborativa baseada em modelo, baseada em conteúdo e híbrida, está relacionada à necessidade de abordar os desafios complexos envolvidos na recomendação de produtos ou de conteúdos em ambientes empresariais.

Cada uma dessas abordagens oferece vantagens específicas que podem ser relevantes para atender às demandas e às expectativas dos usuários em diferentes cenários (AGGARWAL, 2016). Isso porque tais técnicas possuem a capacidade de proporcionar personalização de recomendações com base em interações passadas, considerar características específicas dos itens e usuários e combinar essas duas componentes a fim de oferecer recomendações, atendendo as necessidades dos usuários, especialmente em cenários *de e-commerce*.

Portanto, ao testar e comparar essas diferentes técnicas de recomendação, o presente estudo pode identificar qual delas se adapta melhor ao contexto específico da empresa, levando em consideração fatores como a disponibilidade de dados, a diversidade de itens e as preferências dos usuários. Assim, evidencia-se que a utilização de abordagens variadas objetiva buscar uma solução eficaz e personalizada para atender às demandas do comércio eletrônico e melhorar a experiência do cliente.

1.4 ESTRUTURA

A estrutura do trabalho é composta de 5 capítulos. O primeiro capítulo é destinado à introdução e à apresentação do trabalho. Neste, o tema é contextualizado, os objetivos da pesquisa são apresentados e são fornecidas as devidas justificativas quanto à relevância do trabalho. O segundo capítulo reúne definições e demais informações que compõem a fundamentação teórica do trabalho.

O terceiro capítulo é dedicado a classificar metodologicamente a pesquisa e relatar as etapas realizadas no desenvolvimento do trabalho, explicando os procedimentos utilizados. Em seguida, no quarto capítulo, são apresentados os resultados obtidos durante o desenvolvimento do artefato e são aplicadas as técnicas para avaliar o seudesempenho. Finalmente, no quinto e último capítulo, são expostas as principais considerações a respeito do estudo.

2 REFERENCIAL TEÓRICO

Neste capítulo será explorado o referencial teórico que fundamenta o presente trabalho, abordando três campos de conhecimento interconectados. Inicialmente, ocorrerá uma especificação das técnicas de *Data Mining* (Mineração de Dados) voltadas para *Business Analytics* (Análise de Negócios), destacando a adaptação da metodologia *SEMMA*. Este enfoque oferece uma estrutura robusta para a extração de conhecimentos valiosos a partir de conjuntos de dados, proporcionando aprendizados importantes para o desenvolvimentos de análises e de modelagens estatísticas.

Em seguida, os conhecimentos de Sistemas de Recomendação serão levantados em diversas dimensões, desde os tipos de coleta de informações dos usuários até as estratégias e técnicas de recomendação. Serão detalhadas as principais técnicas, incluindo a Filtragem Colaborativa, a Filtragem Baseada em Conteúdo e a abordagem Híbrida, com ênfase na implementação do modelo *LightFM*. Este segmento visa fornecer uma compreensão aprofundada de como personalizar a experiência do usuário por meio de recomendações eficazes.

Finalmente, será abordada a *Design Science Research* (DSR), um campo de estudo que integra a criação de soluções inovadoras com o rigor metodológico da pesquisa científica. Serão evidenciadas as etapas de desenvolvimento do DSR, desde a identificação do problema até a construção e avaliação de artefatos. Esta abordagem orientada à prática se mostra adequada diante da necessidade de criação de soluções eficazes que atendam às necessidades específicas da organização, contribuindo, assim, para o avanço do conhecimento e da prática nas áreas de Mineração de Dados e Sistemas de Recomendação. Dessa forma, este capítulo visa proporcionar uma base teórica sólida para a compreensão e implementação desses conceitos no contexto do trabalho.

2.1 MINERAÇÃO DE DADOS PARA *BUSINESS ANALYTICS*

Cunhado pelo cientista da computação da *IBM* H. P. Luhn, o termo *Business Analytics* representa funcionalmente as habilidades, tecnologias, práticas e algoritmos relacionados à coleta e à mineração de dados com o objetivo de gerar informações relevantes e apresentáveis a interessados em determinado contexto, como por exemplo, executivos de uma organização ou usuários de uma plataforma (LEE, 2013).

Já do ponto de vista de Shmueli (2017), Análise de Negócios equivale às aplicações e à arte de levantar dados quantitativos a fim de embasar a tomada de decisão em organizações, contando com métodos de análises de dados e modelagem estatística. Evidencia-se em ambas as literaturas referenciadas que a finalidade de todos esses elementos desse campo de estudo é fornecer informações assertivas para agentes e/ou usuários de organização, embasando uma próxima ação ou decisão.

A Análise de Negócios é dimensionada de três maneiras, de acordo com Sapple, Lee e Pakath (2014):

a) Quanto ao domínio - Campo de atuação no qual o aspecto de Análise de Negócios são aplicados. Tais domínios incluem as tradicionais e modernas unidades de negócio de organizações: *Maketing Analytics, Supply Chain Analytics*, *Financial Analytics*, *Human Resource Analytics*, entre outras;

b) Quanto à orientação - Direção do pensamento da unidade de negócio frente à aplicação de *Business Analytics,* sendo considerado uma representação de seu respectivo elemento central. A orientação mais em voga recentemente é a de *Predictive Analytics*, ou seja, quais os meios necessários para predizer se algo pode ocorrer ou não;

c) Quanto à técnica - Forma na qual uma atividade de Análise de Negócios é realizada, podendo ser diferenciada em relação aos mecanismos utilizados. Mineração de Dados, *Online Analytical Processing* (*OLAP*), *Data Warehousing* e *Text Mining* são alguns dos exemplos elencados pelo autor.

Enquanto isso, Lee (2013) diferencia os termos Análise de Negócios*, Mineração de Dados* e *Business Intelligence* de acordo com suas funções, interações e objetivos dentro das organizações. Para o autor, Mineração de Dados é um termo coletivo utilizado para descrever diversas técnicas de análise, como estatísticas, inteligência artificial e aprendizado de máquina, que são empregadas para examinar grandes volumes de dados presentes nos bancos de dados da organização. Seu propósito é identificar padrões no conjunto de dados.

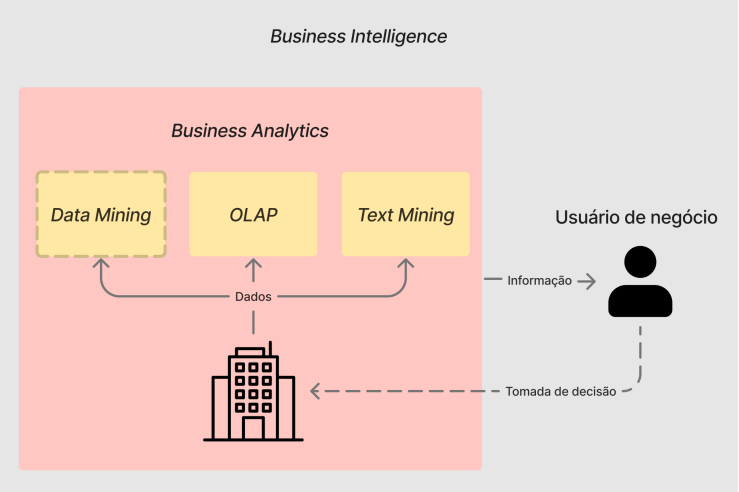
Já a Análise de Negócios é empregada para descrever toda a função de aplicar habilidades, tecnologias, práticas padrão e algoritmos relacionados à Mineração de Dados e métodos de compilação de dados para gerar informações valiosas, geralmente apresentadas em um formato altamente compreensível, de modo que os gestores possam tomar decisões de negócios e controlar e gerenciar as operações empresariais. Quando a função de Análise de Negócios é executada de maneira eficiente e eficaz, ela pode se tornar uma competência central para a organização na forma de uma *Business Intelligence* (Inteligência de Negócios) valiosa que apoiará as ações estratégicas empreendidas pela organização.

A partir dessa visão do autor, entende-se que procedimentos de Mineração de Dados costumam ser aplicados como plano de fundo para as funções de *Business Analytics*, enquanto que as funções via interface são realizadas por relatórios e plataformas de visualizações de dados e métricas do negócio em questão.

Nesse contexto, evidenciam-se as vantagens da aplicação das práticas de Mineração de Dados*,* uma vez que tais questões dificilmente seriam identificadas por

meio de uma análise qualitativa e sem a utilização de recursos computacionais. A Figura 2 mostra, em forma de diagrama, a abrangência de todos esses termos e como estes interagem no ambiente no qual se inserem, de acordo com a bibliografia.

Figura 2 – Domínios da Business Intelligence, Análise de Negócios e Mineração de Dados



Fonte: Elaborado pelo autor

Logo, Mineração de Dados é um termo que se refere aos procedimentos para análise e exploração em amplos volumes de dados que têm como objetivo buscar padrões, previsões, associações, erros e dentre outros fatores. Incluem, dessa forma, métodos estatísticos e de aprendizagem de máquina como suporte para a geração de informações de negócio, e, portanto, para tomada de decisão orientada a dados (SHMUELI *et al.*, 2017).

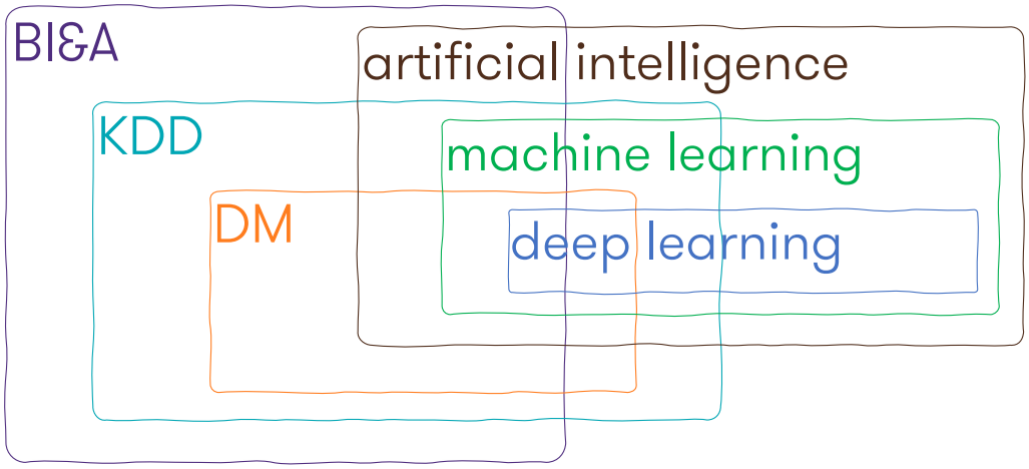
Geralmente aplicam-se tais métodos de maneira automatizada e a um nível individual, como por exemplo, ao possibilitar prever o valor a ser gasto por um cliente da loja. A partir das informações geradas pela mineração de padrões nos bancos de dados de organizações, é possível desenvolver novas estratégias para o negócio, validar a necessidade e perfil dos consumidores, prever resultados financeiros da organização ou personalizar a experiência dos usuários em plataformas *on-line*, por exemplo.

Em termos de técnicas e áreas do conhecimentos abordadadas, a Mineração de Dados combina conceitos de aprendizado de máquina e inteligência artificial, reconhecimento de padrões, estatística e sistemas de bancos de dados, bem como entendimento de regras e problemas de negócio de uma organização. A Figura 3 ilustra a conexão e sobreposição de alguns desses campos de pesquisa.

Então, por meio da aplicação de práticas de Mineração de Dados, é possível identificar possíveis padrões que auxiliam na previsão, associação e agrupamento de

eventos, produtos ou clientes de maneira mais assertiva. Assim, permite-se que as organizações forneçam melhores produtos ou serviços aos clientes, por exemplo.

Figura 3 – Sobreposição dos conhecimentos relacionados à Mineração de Dados



Fonte: CASTELO (2019)

2.1.1 Etapas da Mineração de Dados

Existem diversas metodologias diferentes de Mineração de Dados, mas não há uma em formato padrão aceita pela comunidade acadêmica para aplicar tal técnica. Dessa forma, diferentes fornecedores de soluções de Mineração de Dados criaram suas próprias metodologias proprietárias, que foram adaptadas por cientistas de dados e profissionais da área de forma a permitir novos desenvolvimentos sem obrigatorieade de utilização de ferramenta específica (SOLARTE, 2002).

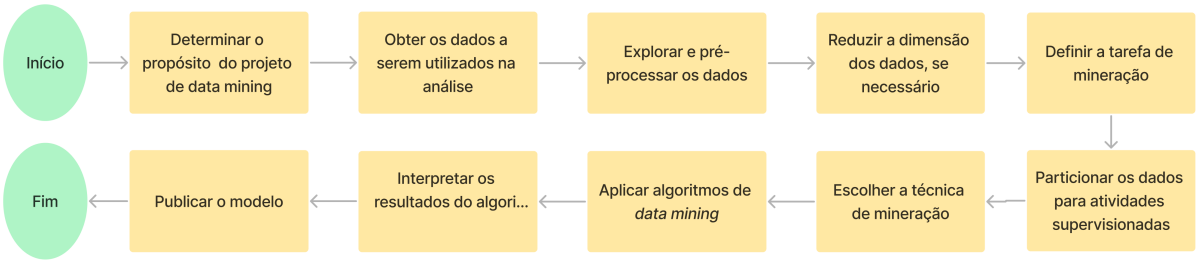
Uma delas, a metodologia *SEMMA*, é proposta pelo *SAS Institute*, uma das empresas mais importantes no desenvolvimento de aplicativos de *software* estatístico, e licenciada dentro da solução de *software Enterprise Miner*. Na *SEMMA*, o processo de mineração consiste em cinco etapas: amostragem, exploração, modificação, modelagem e avaliação.

Apesar da metodologia em questão conter alguns dos pilares de projetos de Mineração de Dados, ela aborda apenas as partes de estatística, de modelagem e de manipulação de dados do processo. Assim, a mesma não inclui algumas partes genéricas de projetos de sistemas de informação em um contexto de uma organização, como as fases de análise, *design* e implementação, tampouco considera os papéis da organização e dos *stakeholders* durante o projeto.

As etapas de Mineração de Dados especificadas a seguir são extraídas de Shmueli (2016), que adaptou a metodologia de forma a corrigir tal ausência de

etapas de entendimento e validação das regras de negócio de uma organização. O fluxograma da Figura 4 mapeia a sequência de etapas elaborada pelo autor.

Figura 4 – Fluxograma das etapas de Mineração de Dados



Fonte: Elaborado pelo autor

a) Compreender projeto de Mineração de Dados: Esta etapa permite desenvolver uma compreensão do objetivo do projeto em questão. Deve, ainda, responder perguntas relacionadas ao problema de mineração como: Como o interessado utilizará os resultados? Quem será afetado pelos resultados? A análise será um esforço único ou um procedimento contínuo? Assim, o objetivo desta etapa é ambientar o *Business Analyst* no contexto de solução do problema da organização. Em Sistemas de Recomendação, por exemplo, os interessados principais são os usuários da plataforma que entrega continuamente as recomendações de produto, os afetados pelos resultados são estes usuários bem como a organização que potencialmente terá aumento em suas receitas devido às recomendações assertivas;

b) Obter conjunto de dados: A obtenção do conjunto de dados geralmente envolve amostragem de um grande banco de dados da organização em questão, a fim de capturar registros a serem utilizados na análise. A precisão dessa amostra em relação aos registros de interesse afeta a capacidade dos resultados da Mineração de Dados de generalizar para registros fora dessa amostra. Vale ressaltar que é possível haver a combinação de dados de diferentes bancos de dados ou fontes, sendo que este podem ser podem ser internos, como compras anteriores feitas pelos clientes e avaliações de produtos, ou externos, por exemplo coletando classificações de crédito e dados demográficos do governo. Embora a Mineração de Dados opere com bancos de dados relativamente grandes, geralmente a análise a ser realizada se restringe a apenas milhares ou dezenas de milhares de registros;

c) Explorar, limpar e pré-processar os dados: Esta etapa envolve verificar se os dados estão em condições adequadas para análise. Ainda, se preocupa

com questões de qualidade e representatividade de dados, como por exemplo, validar e corrigir dados ausentes, fora de escala razoàvel (*outliers*) ou não representativos à realidade. Os dados também são revisados graficamente, por exemplo, por meio de uma matriz de gráficos de dispersão mostrando a relação de cada variável com todas as outras variáveis. Ademais, se faz necessário garantir consistência nas definições de campos, unidades de medida, períodos de tempo e tipos de dados. Nesta etapa, novas variáveis podem ser criadas a partir das já existentes;

d) Reduzir a dimensão dos dados, se necessário: A redução de dimensão envolve operações como eliminar variáveis desnecessárias, transformar variáveis já existentes, remover caracteres específicos de valores de nomes de produtos, e, por fim, criar novas variáveis. Por exemplo, uma que registra se pelo menos um de vários produtos foi comprado. Nesta etapa, deve-se entender o significado de cada variável e se é adequado incluí-la no modelo. Em Sistemas de Recomendação, é nesta etapa que são definidas as matrizes de interação entre usuário, produtos e conteúdos relevantes;

e) Determinar a tarefa de Mineração: A determinação da tarefa de mineração envolve traduzir a pergunta ou problema geral da primeira etapa em uma pergunta de Mineração de Dados mais específica, definindo uma atividade de modelagem estatística que trará as informações necessárias para solucionar o problema de Análise de Negócio. Do ponto de vista de Ricci, Rokach e Shapira (2011), as tarefas de Mineração de Dados mais comumente utilizadas em Sistemas de Recomendação podem ser categorizadas como: Classificação, Clusterização ou Criação de Regras de Associação;

f) Particionar os dados: Algoritmos de aprendizado supervisionado são aqueles utilizados em classificação e previsão e recebem essa nomenclatura porque são necessários dados os quais os valores do resultado de interesse seja conhecido. Tais dados podem também ser chamados de dados rotulados, uma vez que contém o valor do resultado para cada registro. Esses dados de treinamento são os quais o algoritmo de classificação ou previsão aprende ou é treinado acerca da relação entre as variáveis independentes e a variável dependente. A partir do momento que o algoritmo passa pelos dados de treinamento, ele é aplicado a outra amostra de dados rotulados, conhecidos por os dados de validação, onde o resultado é conhecido, mas inicialmente oculto, a fim de medir o quão bem ele se compara a outros modelos. Essa técnica é popularmente conhecida como Validação Cruzada e suas especificações para Sistemas de Recomendação são descritas na subseção 2.2.4.1.

g) Definir a técnica de mineração: A escolha da técnica está diretamente relacionada ao problema/projeto de mineração em questão e a tarefa de mineração pode ser, eantão, definida. Cada técnica de algoritmo possui suas particularidades e podem ser aplicadas de maneira simultânea em um contexto de desenvolvimento *offline*, uma vez que diferentes contextos e *datasets* geram resultados diferentes para um mesmo tipo de algoritmo, por exemplo;

h) Aplicar algoritmos para realizar a tarefa: A etapa de aplicação dos algoritmos para realização da tarefa é um processo tipicamente iterativo, com a realização de testes a partir de diferentes variantes e utilizando múltiplas variantes do mesmo algoritmo. Quando apropriado, o feedback do desempenho do algoritmo nos dados de validação é utilizado para refinar as configurações.

i) Interpretar os resultados dos algoritmos: O processo de avaliação dos resultados dos algoritmos envolve determinar o melhor algoritmo a ser implementado a partir da avaliação de métricas de erro e precisão relativos à técnica adota e, quando possível, testar a escolha final nos dados de teste a fim de entender a sua performance em casos específicos. Vale ressaltar que cada algoritmo também pode ser testado nos dados de validação para fins de ajuste. Dessa forma, os dados de validação se tornam parte do processo de ajuste e provavelmente subestimam o erro na implantação do modelo finalmente escolhido;

j) Implantar o modelo: Finalmente, após a escolha do modelo e de suas configurações adequadas, a última etapa de Mineração de Dados envolve integrar o modelo em sistemas operacionais e executá-lo em registros reais para produzir decisões ou ações. À título de exemplo, um modelo de recomendação de produtos pode ficar disponível à aplicação de um *website* por meio de uma *API*. Assim, no momento que um usuário visualiza um produto na plataforma, produtos recomendados são oferecidos a este potencial cliente.

2.2 SISTEMAS DE RECOMENDAÇÃO

A abundância e disponibilidade de informações por meio da *Internet* configura um meio com alta diversidade de opções de navegação e experiência ao usuário. Frequentemente, este possui poucas habilidades para realizar escolhas dentre as várias alternativas que lhe são apresentadas (CAZELLA, 2010).

Emergem nesssas situações os Sistemas de Recomendação, os quais auxiliam no aumento da capacidade e eficácia deste processo de escolha por parte dos usuários. Sucintamente, esses tipos de sistema fornecem recomendações (*output*) para usuários interessados a partir do processamento de dados de entrada (*input*) de

utilização ou preferência de outros indivíduos. Em um contexto de comércio eletrônico, um dos grandes desafios deste tipo de sistema é combinar as expectativas dos usuários frente à grande disponibilidade de produtos a serem recomendados aos mesmos.

De acordo com os dados disponíveis no *Dashboard* do Comércio Eletrônico Nacional, elaborado pelo Ministério do Desenvolvimento, Indústria, Comércio e Serviços, o comércio eletrônico brasileiro movimentou cerca de R$ 450 bilhões, sendo superior ao dobro da soma dos valores registrados nos anos anteriores à pandemia do *Covid-19* (DO DESENVOLVIMENTO, INDÚSTRIA, COMÉRCIO E SERVIÇOS, 2023).

É possível validar portanto, a representatividade do *e-commerce* não apenas em um contexto empresarial, mas também frente à toda sociedade. Naturalmente, o aquecimento do setor costuma ser convertido em investimento em novas tecnologias e processos de personalização da experiência de clientes.

Os Sistemas de Recomendação configuram-se como importantes ferramentas para a personalização de sistemas computacionais, principalmente na *web*. Estes possuem as funcionalidades de captar preferências e sugerir itens relevantes para cada usuário, de acordo com a análise de seu comportamento de navegação, compra, preferências, entre outros aspectos.

O conceito base de Sistemas de Recomendação é que existem relações significativas entre as atividades centradas nos usuários e nos itens os quais estes interegem. À título de exemplo, um usuário interessado em produtos veganos é mais provável de se interessar por outro produto desta categoria ou por produtos orgânicos, em comparação à produtos de origem animal.

As diversas categorias de itens tendem a apresentar correlações significativas, que podem ser aproveitadas para fazer recomendações mais precisas, como também as dependências podem estar presentes em menor granularidade, através dos itens individuais. Tais relacionamentos podem ser aprendidos de maneira orientada por dados a partir da matriz de *feedbacks*, e assim, o modelo resultante é usado para fazer previsões para os usuários-alvo (AGGARWAL, 2016).

O autor ainda afirma que quanto maior o número de itens avaliados, os quais podem ser ofertados para um usuário, mais robustas serão as previsões feitas sobre o comportamento futuro do usuário. Existem diversos modelos de aprendizado que podem ser usados para realizar essa tarefa.

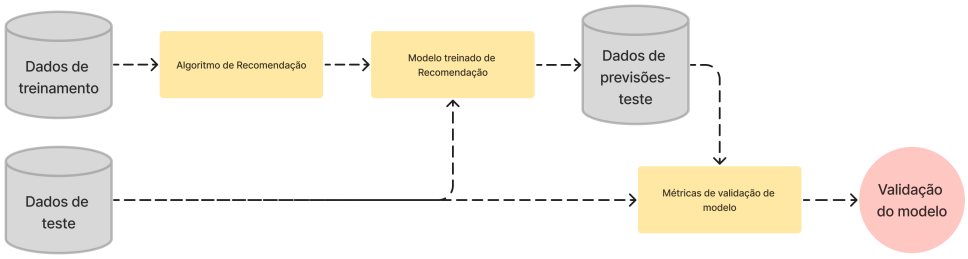
Do ponto de vista de Aggarwal (2016), o fluxo de trabalho genérico de desenvolvimento de Sistemas de Recomendação que utilizam técnicas de recomendação baseadas em aprendizado de máquina para a geração das recomendações devem ocorrer de acordo com a Figura 5, que traz um diagrama de fluxo deste processo.

Primeiramente, um conjunto de treinamento é fornecido a um algoritmo de

recomendação que produz um modelo de recomendação que pode ser usado para gerar novas previsões. Para avaliar o modelo, um conjunto de teste separado é fornecido ao modelo treinado, onde previsões são geradas para cada par de usuário-item.

Previsões com rótulos conhecidos, os valores reais e concretizados, são então usados como entrada para o algoritmo de avaliação para produzir resultados de avaliação .Nota-se que essas etapas serão, portanto, executadas ao longo das etapas de Mineração de Dados, as quais são definidas na seção 2.1 do presente trabalho.

Figura 5 – Fluxo de Trabalho de desenvolvimento de Sistemas de Recomendação



Fonte: Adaptado de Aggarwal (2016)

Por fim, Cazella, Nunes e Raetegui (2010) bem como Aggarwal (2016) são alinhados ao fato que os Sistemas de Recomendação podem ser formalmente classificados quanto às possiveis tipos de coletas de informação dos usuários, às estratégias e técnicas de recomendação, e adicionalmente, os métodos para avaliação destes sistemas. Cada classificação é detalhada nos seguintes subtópicos.

2.2.1 Coleta de Informações do Usuário

O *design* dos algoritmos de recomendação possui relação direta pela forma utilizada para rastrear as interações usuário~item. A metrificação dessa interação pode ser baseada em uma escala que indica o nível específico de afinidade ou não do item em questão, normalmente definida em conjuntos de intervalos discretos.

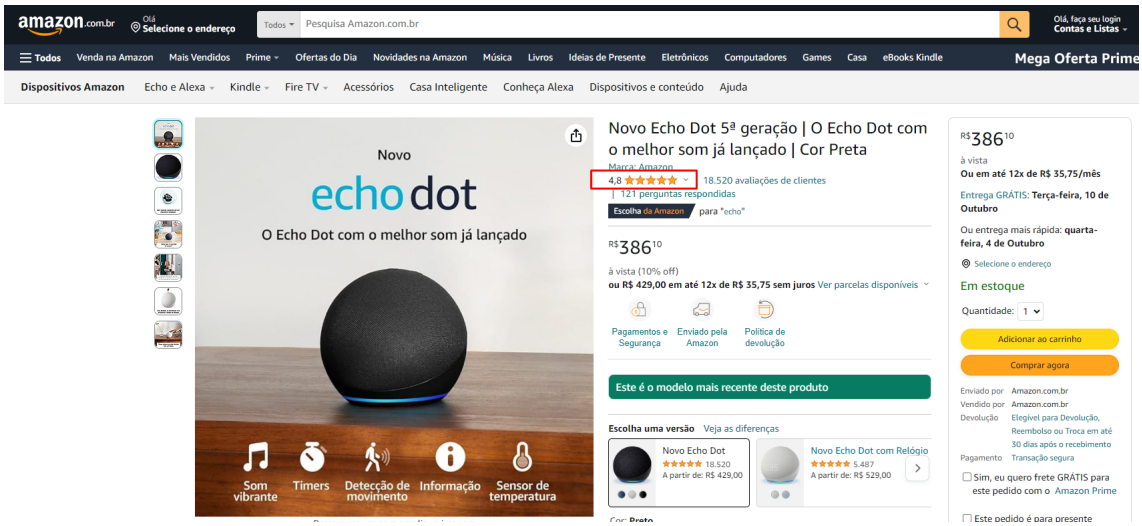
Uma escala de avaliação de 5 pontos, por exemplo, pode ser retirada do conjunto {1, 2, 3, 4, 5}, onde uma avaliação de 1 indica um extremo desagrado, e uma avaliação de 5 indica uma forte afinidade com o item. Essas são as características conferidas à uma coleta de informações de *feedback* explícito dos usuários (RICCI; ROKACH; SHAPIRA, 2011). A Figura 6 ilustra um exemplo de uma média de avaliações explícitas para um produto vendido na comércio eletrônico *Amazon*

Já para o caso das avaliações unárias, as preferências dos clientes são derivadas de suas atividades frente à determinado conjunto de itens, em contraste às avaliações explicitamente especificadas. Por exemplo, o comportamento de compra

de um cliente pode ser convertido como uma preferência pelo item. No entanto, o ato de não comprar um item em um grande universo de possibilidades nem sempre indica um desagrado por parte do usuário (RESNICK; VARIAN, 1997).

Desse modo, as avaliações unárias frequentemente representam conjuntos de dados de *feedback* implícito, ou seja, aqueles que não possuem mecanismo para especificar a afinidade do usuário por um item, mas sim inferir certa relação entre este par a partir de comportamentos do usuário.

Figura 6 – Exemplo de Feedback explícito na loja Amazon



Fonte: Amazon (2023)

A principal motivação para utilizar *ratings* implícitos é a de desprender do usuário a necessidade de verificar e analisar cada item disponível. Alguns exemplos desses comportamentos são: histórico de compra, histórico de navegação, padrões de pesquisa, movimentos do mouse, tempo que o usuário passou assistindo um vídeo, quanto tempo passou lendo um artigo, entre outros.

Assim, de maneira sucinta, as preferências do usuário são especificadas por meio de *reviews* ou avaliações na coleta de dados de *feedback* explícito, enquanto que para o *feedback* implícito, tais preferências são mais genéricas, representando por exemplo a visita de um potencial cliente à pagina de um produto ou se o mesmo realizou uma compra.

De acordo com o que foi levantado em Steck (2010) e validado em Kula (2018), modelos matemáticos que utilizam filtragem explicíta possuem desempenho consideravelmente inferior quando comparado a um mesmo *dataset* de *feedback implícito.* Assim, se mostra inapropriado o desenvolvimento de um sistema com este tipo de coleta de dados. Vale ressaltar que diversos projetos de Sistemas de Recomendação disponíveis na *internet* são, em sua maioria, baseados em *feedback* implícito.

2.2.2 Estratégias de Recomendação

Os Sistemas de Recomendação a serem implementados em organizações devem buscar a fidelidade, para, então, visarem o aumento da lucratividade das empresas, uma vez que existem diferentes estratégias de personalização de experiência de usuário, cada qual com a sua complexidade de execução e efetividade em seus resultados (CAZELLA, 2010).

Uma estratégia utilizada em Sistemas de Recomendação é baseada no uso das avaliações dos usuários para estabelecer a reputação de um item ou de um produto. Após a aquisição de determindado produto por parte de usuário, este ganha a possibilidade de adicionar um *review* sobre o produto adquirido. Assim, as avaliações de clientes se mostram eficazes como uma ferramenta que objetiva assegurar outros consumidores a respeito da qualidade e utilidade dos produtos comercializados (CAZELLA, 2010).

Já na estratégia de Recomendação por Associação, aplicam-se técnicas capazes de construir associações entre produtos que foram avaliados por usuários ao banco de dados da organização. Dessa forma, o sistema pode recomendar algum item que possa complementar o produto que está sendo comprado ou recomendar itens semelhantes. Esse é o tipo mais complexo de recomendação, uma vez que exige uma análise mais profunda dos hábitos do usuário para a identificação de padrões e, então, para a recomendação de itens (MIRANDA BORGES; LUIZ DE OLIVEIRA, 2010).

Por meio da estratégia de Associação por Conteúdo é possível fazer recomendações com base no conteúdo de determinado item, por exemplo, dimensões do produto, categoria, valor de venda, entre outros. Assim, é necessário que sejam definidas associações num escopo mais restrito, como por exemplo, se determinado produto está na mesma faixa de preço a de outros produtos oferecidos ou avaliados (CAZELLA, 2010).

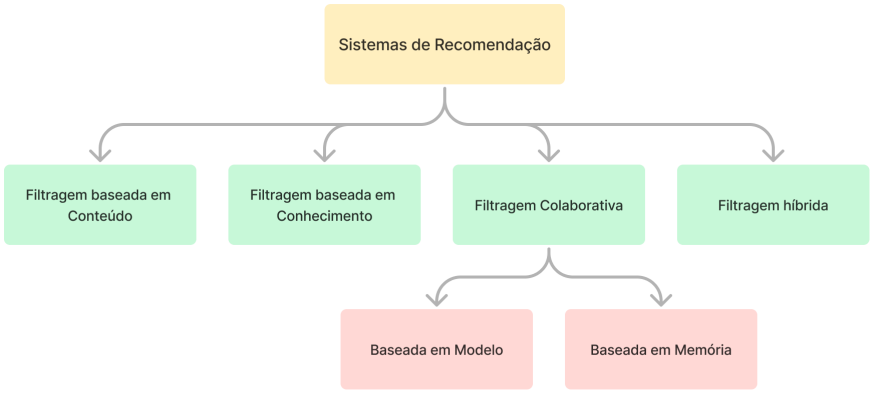
Por fim, a estratégia de Análise de Sequência de Ações estuda seqüências que são utilizadas para capturar o comportamento de usuários através de históricos de atividade temporal, ao longo de sua experiência com uma plataforma de comércio eletrônico, por exmplo. Algumas dessas técnicas frequentemente utilizadas são Clusterização e *SPADE* (*Sequential PAttern Discovery using Equivalence classes*), que, inclusive, podem alimentar sistemas baseados em outras estratégias de recomendação (MIRANDA BORGES; LUIZ DE OLIVEIRA, 2010).

2.2.3 Técnicas de Sistemas de Recomendação

Para que ocorra um certo nível de personalização das recomendações pelos sistemas, estes frequentemente utilizam os padrões de comportamento dos usuários para fundamentar suas recomendações.

Afunilando os conhecimentos de Estratégias de Recomendação por Associação e por Conteúdo, as técnicas mais relevantes de acordo com o contexto da construção do artefato e mais aplicadas de acordo com a biliografia são listadas a seguir. O diagrama da Figura 7 ilustra uma árvore de relacionamentos de alguns destes conceitos.

Figura 7 – Tipos de Sistemas de Recomendação



Fonte: Adaptado de Kapadia (2020)

2.2.3.1 Filtragem manual

A filtragem manual é uma técnica de recomendação que consiste em manter listas de produtos a recomendar organizadas por parâmetros arbitrários, muito relacionados ao contexto da organização. Esta estratégia não exige uma análise profunda, pois não é uma recomendação personalizada. Tais listas podem ser criadas baseadas nos produtos mais populares, mais vendidos, com maior média de avaliação de usuários, entre outros fatores (MIRANDA BORGES; LUIZ DE OLIVEIRA, 2010).

Já que se configura como uma técnica de geração de recomendações não personalizadas, todos os usuários recebem as mesmas recomendações, o que levanta certos questionamentos a despeito de sua precisão e acurácia.

2.2.3.2 Filtragem colaborativa baseada em modelo

Em Sistemas de Recomendação que utilizam filtragem colaborativa, os usuários recebem sugestões de itens com base em outros usuários que possuam afinidades semelhentes em itens já consumidos e avaliados. De acordo com Kapadia (2020), a filtragem colaborativa é centrada na coleta e na análise de dados sobre o comportamento dos usuários, suas atividades usuais e valiações a fim de viabilizar a previsão de suas afinidades por itens a serem adquiridos.

Este tipo de técnica leva em consideração que indivíduos mantêm as suas afinidades ao longo do tempo imutáveis. Por exemplo, se um indivíduo “A” tem afinidade pelos produtos 1, 2 e 3 e, o indivíduo “B”, pelos produtos 2, 3 e 4, conclue-se que os mesmos possuem interesses semelhantes e se torna razoável afirmar que “A” terá afinidade com o produto 4 e “B” com o produto 1.

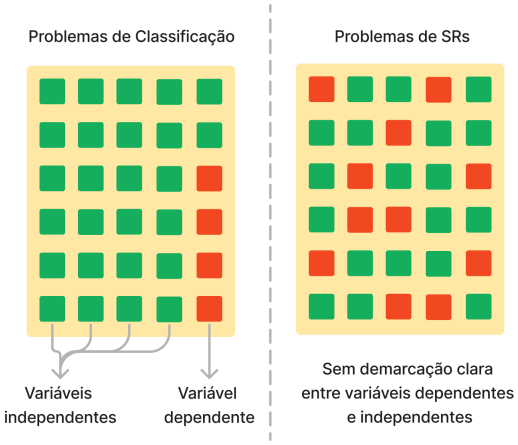
De maneira sucinta, Sistemas de Recomendação modernos que utilizam filtragem colaborativa são baseados em memória ou em modelo, em termos de solução de seus respectivos algortimos de otimização. No primeiro caso, são previstas similaridades usuário-usuário ou item-item de acordo com classificações de k-vizinhos, por exemplo, podendo ser analisadas em usuários ou itens (AGGARWAL, 2016).

Nos métodos de filtragem colaborativa baseados em modelo, um modelo resumido dos dados é criado anteriormente à previsão da matriz de avaliações, de forma semelhante como ocorre com métodos de aprendizado de máquina supervisionado ou não supervisionado. Assim, a fase de treinamento é separada da fase de previsão. Classificação Bayesiana, árvores de regressão e modelos de fatores latentes são alguns exemplos desses métodos.

Devida a extensa aplicação em trabalhos acadêmicos, como por exemplo no artigo do modelo *LightFM,* a solução por modelos de fatores latentes será detalhada na subseção 2.2.3.2.1. Ademais, a partir desses exemplos é possível notar que uma grande parte desses modelos podem ser generalizados para um contexto de filtragem colaborativa, uma vez que problemas tradicionais de regressão e classificação são casos especiais de compleção matricial, ou seja, filtragem colaborativa.

A Figura 8 denota tal relação, com a matriz da esquerda sendo um problema clássico de classificação: Previsão de uma variável dependente, que pode ser dividida em dados teste e dados treinamento, em uma matriz m x n a partir de variáveis independentes conhecidas. Ao lado direito, um problema de filtragem colaborativa em uma matriz de avaliações, onde não é clara a demarcação entre as linhas teste e linhas treinamento. Ademais, a subseção 2.2.4.1 detalhará como é aplicada técnica de Validação Cruzada para um contexto de Sistemas de Recomendação, garantindo confiabilidade aos procedimentos de treinamento e teste.

Figura 8 – Comparação entre problemas de recomendação e classificação



Fonte: Adaptado de Aggarwal (2016)

2.2.3.2.1 Fatoração Matricial

A filtragem colaborativa baseada em modelo, resolvida por fatoração matricial, é uma abordagem poderosa para Sistemas de Recomendação que lida com a tarefa de prever interações desconhecidas entre usuários e itens em uma matriz esparsa. O cerne dessa técnica reside na criação de representações latentes para usuários e itens, buscando capturar características não observáveis diretamente.

A fatoração matricial permite que o modelo aprenda automaticamente essas representações latentes durante o treinamento, decompondo a matriz esparsa original em dois conjuntos de matrizes, uma para usuários e outra para itens. Essas matrizes capturam a essência dos padrões de interação, permitindo uma reconstrução eficaz da matriz completa, inclusive preenchendo valores faltantes (SARDANA, 2016).

Ao criar representações latentes para usuários e itens, a fatoração matricial resolve o problema da alta dimensionalidade e esparsidade dos dados, transformando o espaço de interações em um espaço latente de dimensões menores. Esse espaço latente representa de maneira mais eficiente as relações entre usuários e itens, permitindo a generalização para itens não observados.

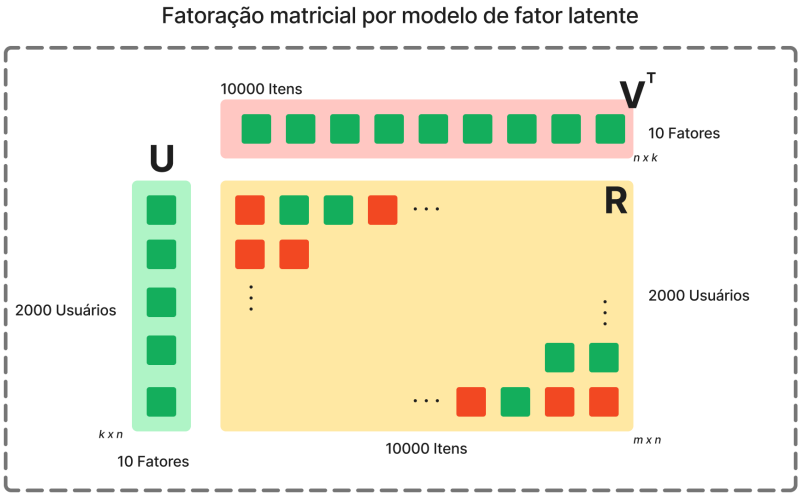
Durante o processo de treinamento, a fatoração matricial ajusta iterativamente os parâmetros do modelo, incluindo os fatores latentes, para minimizar a discrepância entre as previsões e as interações reais. Essa abordagem não só melhora a capacidade do modelo de fazer recomendações precisas, mas também supera o desafio da esparsidade inerente aos dados de interação do mundo real (LEE, D., 2018).

Na Figura 9, pode-se observar que a matriz de interações, de cor amarela, é decomposta em duas outras matrizes de representação de fatores latentes, uma de

usuários e outra de itens. Os quadrados da cor verde indicam valores preenchidos pelo usuário quando o mesmo adquire um produto, enquanto que os quadrados da cor vermelha indicam valores faltantes.

O objetivo, portanto, é gerar estimações das matrizes de fatores latentes, a partir da resolução de um problema de otimização, como a técnica do Gradiente Estocástico Descendente, por exemplo. A partir disso, é possivel realizar o produto interno destas duas matrizes e definir uma nova matriz de interações, agora com os valores faltantes, ou seja, aqueles que devem ser previstos, devidamente preenchidos.

Figura 9 – Filtragem colaborativa baseada em modelo de fatoração matricial



Fonte: Adaptado de Lee (2018)

A fatoração matricial é um processo iterativo que envolve várias etapas para otimizar a representação latente de usuários e itens, prever valores faltantes na matriz esparsa e melhorar continuamente a precisão do modelo. Tais etapas são descritas nas alíneas a seguir, conforme Aggarwal (2016).

a) Definir modelo matemático de otimização: Neste passo é crucial estabelecer um modelo matemático de otimização que descreva a relação entre os valores conhecidos e desconhecidos na matriz esparsa. Isso inclui a formulação de uma função de perda, que quantifica a discrepância entre as previsões do modelo e os valores reais. A escolha adequada dessa função de perda é fundamental para guiar o treinamento do modelo na direção certa.

b) Resolver problema de Otimização com Gradiente Estocástico Descendente (GED): O segundo passo envolve a resolução do problema de otimização usando técnicas como o GED. Durante o treinamento, o GED itera sobre os valores conhecidos na matriz esparsa, ajustando os parâmetros do modelo, incluindo

os fatores latentes para usuários e itens. Os valores preenchidos da matriz esparsa são utilizados para atualizar as representações latentes, refinando-as para melhor capturar os padrões nas interações. Por se tratar do algoritmo central de otimização do presente trabalho, este será detalhadamente descrito na subseção 2.2.3.2.2. O problema de otimização a ser resolvido pelo Gradiente estocástico é definido na Equação 2.1.

$$Minimize \sum_{i,j} \parallel R_{i,j} - U_i(V_j)^T \parallel^2 s.t.\ R_{i,j}\ is\ not\ empty \tag{2.1}$$

c) Prever valores não preenchidos com o produto escalar: Após a convergência do SGD, as matrizes de representações latentes para usuários e itens foram otimizadas. Agora, o modelo pode prever os valores não preenchidos na matriz esparsa realizando o produto escalar dessas matrizes otimizadas. Essa etapa é essencial para criar uma matriz completa que representa as estimativas do modelo para todas as interações possíveis entre usuários e itens. A Figura 10 ilustra todo este processo de preenchimento dos valores ausentes a partir do produto escalar das matrizes latentes obtidas com a solução do problema de otimização.

d) Comparar e reiterar até convergência: A última etapa consiste em comparar as matrizes esparsas originais (com valores conhecidos) e a matriz esparsa preenchida com as previsões do modelo. Essa comparação fornece uma medida de quão bem o modelo está performando. O processo é iterado até que a convergência seja alcançada, ou seja, até que as previsões do modelo estejam suficientemente próximas dos valores reais. A reiteração permite que o modelo ajuste continuamente as representações latentes para refletir padrões mais refinados presentes nos dados de interação.

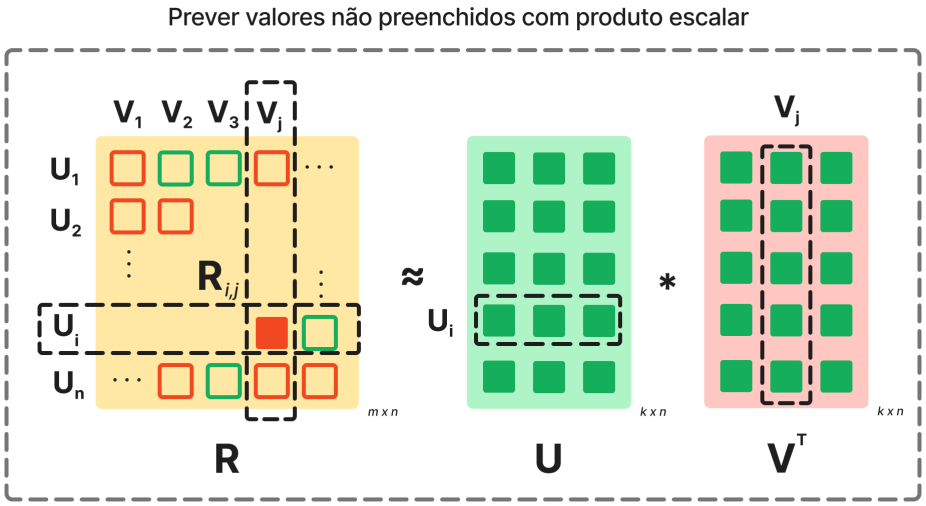
Conforme já mencionado, um dos algoritmos de aprendizagem disponíveis para situação de filtragem colaborativa baseada em modelo de fatoração matricial é o *Stochastic Gradient Descent*. Tal algoritmo faz parte dos modelos de fatores latentes, uma vez que o seu objetivo é o de utilizar métodos de redução dimensional para estimar indiretamente os dados da matriz de avaliações (RESNICK; VARIAN, 1997).

Partindo para uma definição formal em termos de álgebra linear, a intuição modelos de fatores latentes é definida por fatoração matricial. As definições matemáticas a seguir foram aplicadas em Sales (2019) a partir dos conhecimentos verificados em Aggarwal (2016) . Em modelos básicos desse tipo, uma matriz R de notas é fatorada em uma matriz $m \times k$ U e, $n \times k$ V de acordo com o que já foi ilustrado nas Figuras 9 e 10, e também é algebricamente definida na Equação 2.2.

$$R \approx UV^T \tag{2.2}$$

As colunas de U ou V representam um vetor latente, ao passo que cada linha representam um fator latente. A i-ésima linha u_i de U é corresponde ao fator do usuário e contém as *k* entradas relacionadas à afinidade do usuário *i* em relação aos *k* valores na matriz de avaliações. Dessa forma, o valor de R pode ser expresso como o produto escalar do i-ésimo fator do usuário com o j-ésimo fator do item, conforme a Equação 2.3.

Figura 10 – Previsão de valores não preenchidos com produto escalar



Fonte: Adaptado de Lee (2018)

$$r_{ij} \approx \bar{u_i} \cdot \bar{v_j} \tag{2.3}$$

Assim, a partir da equação 2.3, é possível extrapolhar a matriz de interações a ser preenchida como um somatório do produto cartesiano entre a afinidade do usuário i e a afinidade do item j, ambos no intervalo das k entradas definidas. Portanto, evidencia-se que métodos de fatoração de matrizes oferecem uma maneira sofisticada de aproveitar todas as correlações entre linhas e colunas de uma só vez para estimar toda a matriz de dados.

Finalmente, as aplicações de aprendizado de máquina em filtragem colaborativa, no formato do emprego do algoritmo do Gradiente Estocástico Descendente, buscam gerar e otimizar os fatores de itens e de usuarios a partir da minimização do erro quadrado do conjunto de interações estimados, advindos da Equação 2.1 do problema de otimização. A equação 2.4 é obtida a partir das derivadas

parciais das somas dos erros quadráticos das representações latentes em termos do erro de estimação, que λ controla o nivel de regularização e κ se refere ao conjunto de pares (u, i) das avaliações conhecidas da matriz de entrada.

$$Error(U_k \cdot \sqrt{S_k}^T, \sqrt{S_k} \cdot V_k^T) = \sum_{(u,i) \in \kappa} (r_{u,i} - p_u q_i^T) + \lambda(\|p_u\|^2 + \|q_i\|^2) \tag{2.4}$$

2.2.3.2.2 Gradiente Estocástico Descendente

Um dos métodos que buscam realizar a minimização do erro da Equação 2.4, baseado em técnica de aprendizagem de máquina, é chamado de *Stochastic Gradient Descent*. Este realiza uma iteração em todas as notas dos dados de treinamento, de forma que para cada entrada, a avaliação r_{ij} tem seu valor previsto pelo algoritmo e o seu respectivo erro é calculado na equação 2.5.

$$r_{ij} = r_{ui} - \hat{r}_{ui} \tag{2.5}$$

A aplicação eficaz de um algoritmo de filtragem colaborativa baseado em fatoração matricial, como o Gradiente Estocástico Descendente (SGD), envolve uma série de etapas fundamentais para ajustar as representações latentes de usuários e itens. Inicialmente, o algoritmo calcula a estimativa da matriz interativa por meio do produto das representações latentes. Em seguida, o erro entre os valores observados na matriz esparsa e as estimativas é calculado.

As matrizes de representação latentes para usuários e itens são inicializadas com valores arbitrários e os ajustes necessários são determinados através do cálculo das derivadas parciais da função de erro. A atualização dos valores a serem estimados nas próximas iterações é então realizada, incorporando os resultados das derivadas parciais. Importante mencionar também a inclusão da variável taxa de aprendizado γ que desempenha um papel crucial no controle da convergência do algoritmo. A descrição detalhada dessa sequência de etapas é realizada nas alíneas a seguir:

a) Definir o resultado do produto das representações latentes: A primeira etapa consiste em calcular o produto das representações latentes para usuários U e itens V para obter uma estimativa da matriz interativa original UV_T;

b) Definir o erro do valor visto da matriz esparsa menos o resultado do produto das representações latentes: Calcular o erro entre os valores conhecidos da matriz esparsa R e a estimativa do produto das representações latentes UV_T. O erro é dado pela diferença entre a matriz de interações original e o produto interno das matrizes de representação latente;

c) Iniciar as matrizes de representação latentes com valores arbitrários: Inicializar as matrizes de fatores latentes de itens e usuários com valores arbitrários. Esses valores serão ajustados iterativamente durante o processo de treinamento;

d) Calcular os ajustes nas matrizes de representações latentes: Para ajustar as representações latentes, calcular a derivada parcial da função de erro em relação a cada variável a ser prevista (para u usuários e v produtos). Assim são obtidas nas Equações 2.6 e 2.7 definidas. Tal etapa determina a direção e a magnitude do ajuste necessário para minimizar o erro;

e) Calcular o valor a ser estimado nas próximas iterações: Ainda nas Equações 2.6 e 2.7, o valor a ser estimado nas próximas iterações é atualizado usando o valor original da variável a ser prevista, somado ao resultado da derivada parcial da função de erro. Isso é feito para cada variável, tanto para usuários quanto para itens;

f) Ajustar com a variável taxa de aprendizado γ: Para controlar a taxa de convergência do algoritmo, multiplicar o ajuste das variáveis pelas derivadas parciais pelo valor da variável taxa de aprendizado γ. A taxa de aprendizado é um hiperparâmetro que determina o tamanho dos passos tomados em direção à otimização.

Portanto, após cada iteração, os parâmetros são alterados proporcionalmente à taxa de aprendizado γ . Ao passo que ocorrem as iterações a partir das entradas observadas na matriz, que atualizam as matrizes de fatores, eventualmente a convergência será alcançada. Um valor típico para a taxa de aprendizado é um pequeno valor constante, como $\gamma = 0,005$ (AGGARWAL, 2016).

$$q_i \leftarrow q_i + \gamma \cdot (e_{ui} \cdot p_u - \lambda \cdot q_i) \tag{2.6}$$

$$p_u \leftarrow q_i + \gamma \cdot (e_{ui} \cdot q_i - \lambda \cdot p_u) \tag{2.7}$$

Em geral, o método de descida de gradiente estocástico é preferível quando o tamanho dos dados é muito grande e o tempo de computação é o principal gargalo. Vale ressaltar que a execução destes modelos até a convergência por muitas iterações pode reduzir qualidade da solução nas entradas não observadas.

2.2.3.3 Filtragem baseada em conteúdo

As técnicas de recomendação colaborativas vistas na subseções anteriores utilizam as correlações nos padrões de avaliações entre os usuários para fazer

recomendações. Entretanto, tais métodos não levam em conta os atributos dos itens à disposição, como faixa de preço, tamanho e categoria, o que não se mostra ideal, uma vez que recomendações com pouca significância podem ser realizadas.

Para solucionar esse problema, os Sistemas de Recomendação baseados em conteúdo buscam corresponder os usuários a itens que são semelhantes, baseando-se nos atributos dos objetos que o usuário demonstrou afinidade. Diferentemente dos sistemas colaborativos, os sistemas baseados em conteúdo são orientados pelas próprias avaliações do usuário e pelos atributos dos itens que o mesmo avaliou como uma afinidade positiva (SERRANO, 2018).

A fim de realizar recomendações baseadas em conteúdo, um modelo é construído a partir da descrição dos itens avaliados por um usuário, sendo uma representação d e seus interesses. Para recomendar novos itens, os atributos destes são comparados com os atributos do modelo do usuário alvo. O processo de recomendação é baseado em três passos, definidos nas alíneas a seguir (AGGARWAL, 2016).

a) Pré-processamento e extração de *features*: Esta etapa consiste em extrair as informações dos contextos internos da organização, em cada um de seus itens, e armazená-las em um modelo de espaço vetorial. O conteúdo das informações dependende do domínio no qual o sistema de recomendação irá se inserir. Portanto, na primeira etapa são coletadas informações dos usuários referentes aos itens de sua preferência, criando, assim, o chamado *Item Profiler* (Perfil dos Itens);

b) Aprendizagem de perfil dos usuários: Nesta etapa, um modelo especifíco de usuários é determinado por meio de técnicas tradicionais de classificação ou regressão, por exemplo, a fim de predizer os interesses deste usuário a partir de itens previamente avaliados. O resultado dessa etapa é a inferência do chamado de *User Profiler* (Perfil dos Usuários), uma vez que relaciona conceitualmente os interesses dos usuários a partir atributos dos itens;

c) Filtragem e recomendação: Finalmente, o modelo criado no passo anterior é usado para fazer recomendações ao usuário. Assim, nesta etapa ocorre a correspondência entre o "*User Profile*" inferido e o catálogo de itens disponíveis. Essa equiparação é importante para realizar recomendações precisas, uma vez que permite identificar itens no catálogo que são alinhados com as preferências do usuário.

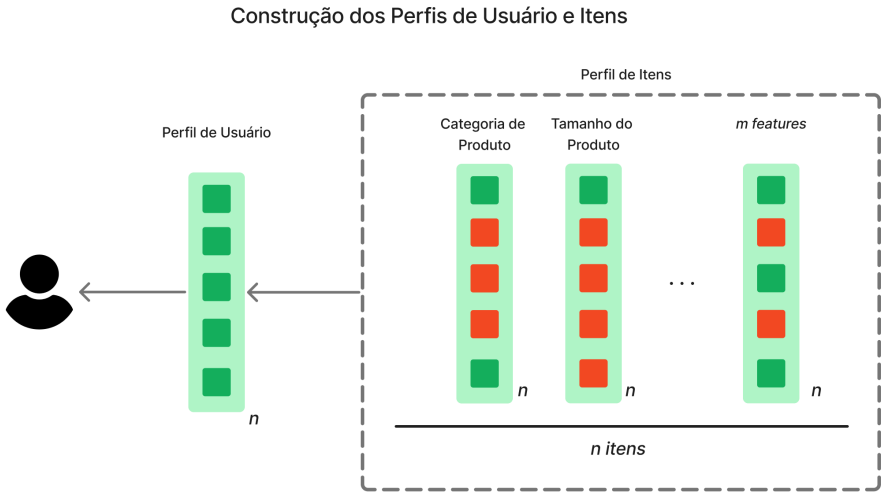
Como é possível perceber a partir da conceituação das etapas acima, sistemas baseados em filtragem de conteúdo possuem a vantagem de que a recomendação para um usuário independe de outras avaliações de outros usuários, uma vez que

apenas os itens avaliados pelo usuário a receber a recomendação são levados em consideração. Entretanto, o sistema fica dependente das informações disponíveis dos itens, o que geralmente leva à uma análise pouco complexa da base de dados.

Em Sistemas de Recomendação baseados em conteúdo, o usuário tende receber recomendações muito semelhantes com o que já foi adquirido e dificilmente algo novo é sugerido, uma vez que o sistema sugere items os quais possuem altas pontuações quando confrontadas com o *user profiler*. Tal comportamento, chamado de superespecialização, é ferido o príncipio da serendipidade defendido anteriomente no começo deste capiítulo (RICCI; ROKACH; SHAPIRA, 2011) .

Uma técnica simples utilizada em diversos modelos para criar *User Profiles,* que acoplam Sistemas de Recomendação baseada em conteúdo, é a da média ponderada. Para um dado usuário, é definida uma média ponderada de todos os vetores que expressam a relação entre as *features* capturadas. Tal abstração pode ser visualizada na Figura 11.

Figura 11 – Criação de Perfis de Usuários e Itens



Fonte: Elaborado pelo Autor

A partir desse vetor de Perfil de Usuário, técnicas de classificação são utilizadas para estimar a preferência ou indiferença do usuário perante outros produtos do catálogo. Uma delas é a estimação a partir do cosseno dos vetores de Perfil de Usuário e Perfil de Itens. Como será discorrido na subseção 2.2.3.4.1, tal estimação propriamente dita não será realizada, mas ocorrerá um acoplamento de tais vetores na matriz de fatores latentes de usuários e na de itens, utilizando a mesma analogia para construção desses vetores.

Com todas essas lacunas em técnicas que analisam o comportamento de usuários semelhantes (Filtragem coloborativa) ou que identificam os principais atributos dos itens preferidos pelos usuários (Filtragem baseada em conteúdo), urge

a necessidade da combinação desses métodos. Nesse sentido, são definidas as técnicas de filtragem híbrida em Sistemas de Recomendação.

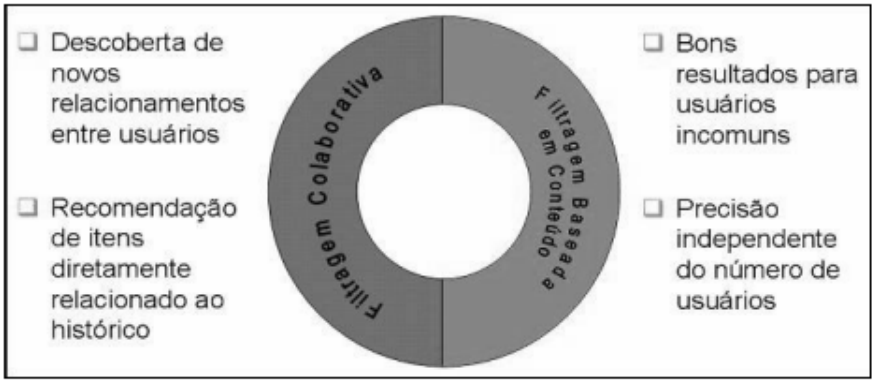
2.2.3.4 Filtragem híbrida

No presente estudo, a filtragem híbrida configura-se como a junção das técnicas definidas nas seções 2.2.3.2 e 2.2.3.3, combinando métodos colaborativos com métodos baseados em conteúdo. A técnica híbrida permite realizar recomendações mais precisas, pois serão baseadas nos conteúdos com maior afinidade dos usuários, e mais variadas, uma vez que será levada em consideração a atividade em itens entre usuários com características semelhantes (RESNICK; VARIAN, 1997).

Sistemas de Recomendação híbridos, especificamente aqueles que combinam técnicas de filtragem colaborativa baseada em modelo de fatoração matricial com técnicas de filtragem baseada em conteúdo, surgem como uma abordagem eficaz para superar limitações decorrentes da aplicação isolada dessas técnicas, especialmente em cenários de conjuntos de dados escassos em informações e interações por usuário ou produto.

A Figura 12 ilustra a junção das vantagens de filtragens colaborativas e baseadas em conteúdo, de forma que compensam as desvantagens de ambas, evidenciando a filtragem híbrida como uma otimização da aplicação de ambas as técnicas. Ademais, um dos principais problemas nos Sistemas de Recomendação é que o número de interações inicialmente disponíveis é relativamente pequeno. Em tais casos, torna-se mais difícil aplicar modelos tradicionais de filtragem colaborativa e filtragem colaborativa baseada em modelo.

Figura 12 – Características de sistemas de filtragem híbrida



Fonte: Extraído de Cazella (2016)

Os métodos baseados em conteúdo são mais robustos do que os modelos colaborativos na presença de poucas avaliações, o chamado *start-cold problem*,

porém, esse conteúdo nem sempre está disponível. Dessa forma, evidencia-se a importância desse tipo de filtragem para a definição de soluções satistórias em problemas de recomendação.

Essa abordagem se torna valiosa diante da inviabilidade de extrair benefícios significativos a partir da utilização isolada de filtragem colaborativa ou baseada em conteúdo em tais contextos. Um dos principais Sistemas de Recomendação híbridos, proposto em Kula (2018) e implementado no pacote *Python*, o modelo *LightFM*, apresenta uma solução inovadora ao introduzir uma variação da fatoração matricial que estima fatores latentes não apenas para usuários e itens, mas também para os metadados associados a eles.

2.2.3.4.1 Modelo LightFM

Como já brevemente denotado, os Sistemas de Recomendação híbridos combinam mais de um tipo de técnica com o objetivo de ampliar as vantagens de cada uma e diminuir as desvantagens associadas a cada método. O modelo *LightFM* híbrido conteúdo-colaborativo, empacota uma solução que visa solucionar todos esses problemas.

Neste modelo, ao invés de se encontrar representações latentes de usuários e itens diretamente, é definida uma representação latente de cada característica para cada combinação usuário x item. Nesse sentido, a representação latente de um item é apenas a soma das representações latentes das características do mesmo. Da mesma forma para os usuários, são somadas as representações latentes das características do usuário para obter a representação latente de um cliente (KULA, 2018).

O processo de aprendizado dos fatores latentes durante o treinamento do modelo *LightFM* envolve otimizar os parâmetros do modelo, incluindo os fatores latentes, para minimizar uma função de perda. Essa função de perda compara as previsões do modelo com as interações reais (ou classificações) na matriz de interações entre usuários e itens.

A matriz de interações entre usuários e itens é o principal dado de entrada para o modelo *LightFM*. Esta representa as relações conhecidas entre usuários e itens, indicando quais itens foram interagidos por quais usuários. Esta é a mesma matriz R que é estudada na seção 2.2.3.2.2 sobre fatoração matricial. Durante o treinamento, o modelo ajusta seus parâmetros, incluindo os fatores latentes, para melhor se adequar a esses padrões de interação observados. O processo geral de treinamento do modelo pode ser resumido nos seguintes passos, conforme visto em (CATES, 2019):

a) Inicialização dos Fatores Latentes: Os fatores latentes para usuários e itens são inicializados com valores aleatórios;

b) Previsão: O modelo usa esses fatores latentes e outros parâmetros para fazer previsões sobre as interações entre usuários e itens;

c) Cálculo da Perda: A função de perda é calculada comparando as previsões do modelo com as interações reais na matriz de interações. Essa função de perda é uma medida da discrepância entre as previsões do modelo e os dados reais. A definição da função de perda já foi definida na seção 2.2.3.2.2, sendo a mesma utilizada neste modelo;

d) Otimização: Otimizadores, como o gradiente descendente estocástico, são aplicados para ajustar os parâmetros do modelo, incluindo os fatores latentes, de maneira a minimizar a função de perda. Durante esse processo, os valores dos fatores latentes são atualizados iterativamente para melhorar a capacidade do modelo de fazer previsões precisas;

e) Iteração: Os passos b-d são repetidos por várias iterações do treinamento, permitindo que o modelo refine gradualmente seus parâmetros e aprenda padrões mais complexos nos dados.

Ao final do treinamento, espera-se que os fatores latentes aprendidos capturem padrões relevantes nas interações entre usuários e itens, permitindo ao modelo fazer recomendações precisas, mesmo para itens não interagidos previamente. Esse processo é uma forma de aprendizado por filtragem colaborativa, onde o modelo aprende com o comportamento passado dos usuários e itens para fazer previsões futuras.

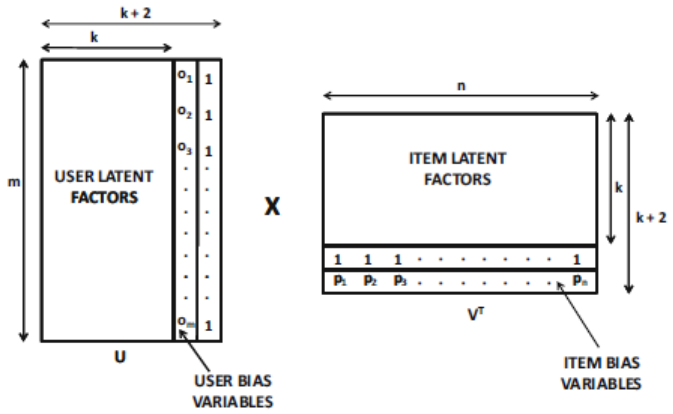
Adicionalmente, o modelo em questão permite embutir aos fatores latentes da filtragem colaborativa as estimações com uma agregação dos metadados de usuários e itens, advindos da filtragem baseada em conteúdo. Dessa forma, é possível gerar recomendações significativas a partir da estimação de menos parâmetros para itens e usuários novos, reduzindo a fatoração matricial a um caso especial. Tal agregação pode ser visualizada na Figura 13 e evidenciada na Equação 2.12.

Vale ressaltar que a relação entre o *LightFM* e o modelo de filtragem colaborativa é governada pela estrutura dos conjuntos de *features* de usuário e item. Logo, se os conjuntos de características consistirem apenas de variáveis indicadoras para cada usuário e item, ou seja, se contiverem apenas os fatores latentes aprendidos pelo Gradiente Estocástico, o *LightFM* se reduz ao modelo padrão de filtragem colaborativa.

Porém, se os conjuntos de características também incluírem *features* de metadados compartilhadas por mais de um item ou usuário, o LightFM estende o modelo de Filtragem Colaborativa, permitindo que os fatores latentes das

características expliquem parte da estrutura das interações do usuário, permitindo o acoplamento de tais informações, como explicado anteriormente.

Figura 13 – Introdução da componente baseada em conteúdo aos fatores latentes



Fonte: Extraído de Aggarwal (2016)

Ainda, é possível solucionar o problema de falta de compartilhamento de informações entre os classificadores baseados em conteúdos, conforme visto anteriormente, uma vez que as representações latentes são combinadas para cada previsão de um par usuário x item.

Em suma, o modelo *LightFM* é o mais adequado para situações de diversos novos usuários e novos itens, além de possuir implementação simplificada. De acordo com seu próprio autor, a adição dos metadados pode influenciar o modelo a recomendar itens já consumidos pelo usuários. Porém, se o conjunto de dados possui as características alta esparsividade e poucas informações históricas, ainda é possível obter resultados satisfatórios com tal aplicação (KULA, 2018).

A fim de se realizar uma descrição formal deste modelo, considera-se U o conjunto de usuários, I o conjunto de itens, F^U o conjunto de características do usuário e F^I, o conjunto de características do item. Cada usuário interage com um número de itens, sendo interações positivas (favoráveis) ou negativas (desfavoráveis). O conjunto de todos os pares de interações entre usuário e item $(u, i) \in U \times I$ representa a união de ambas as interações positivas S^+ e negativas S^-, caso existam

Usuários e itens são completamente descritos por suas *features*, de forma que os conjuntos de *features* de usuários são representados por $f^u \subset F^U$. O mesmo vale para cada item i, cujas características são dadas por $f^i \subset F^I$. O modelo é então parametrizado em termos de incorporações de fatores de usuário e item, e_f^U e e_f^I respectivamente, para cada *feature* f. Cada uma destas são descritas por um fator escalar, sendo b_f^U para *features* do usuário e b_f^I para *features* do item.

Conforme mencionado anteriormente, a representação latente do usuário u é

dada pela soma dos vetores latentes de suas características (q_u), o que analogamente se aplica ao item i, (p_i). Enquanto isso, o fator de *bias* para o usuário u é dado pela soma dos viéses das características (b_u), e, novamente, o mesmo vale para o item i ao finalmente definir a soma de características de itens (b_i). Tais definições podem ser respectivamentes evidenciadas nas Equações 2.8, 2.9, 2.10 e 2.11.

$$q_u = \sum_{j \in f_u} e_j^U \tag{2.8}$$

$$p_i = \sum_{j \in f_i} e_j^I \tag{2.9}$$

$$b_u = \sum_{j \in f_u} b_j^U \tag{2.10}$$

$$b_i = \sum_{j \in f_i} b_j^I \tag{2.11}$$

A previsão do modelo para o usuário u e o item i é então dada pelo produto escalar das representações do usuário e do item, ajustado pelos *biases* das *features* do usuário e do item ($\hat{r}_{ui}$), conforme Equação 2.12. É possível perceber que, com exceção da incorporação dos vetores de *features* de metadados extraídos de usuários e itens, o modelo *LightFM* performa as mesmas operações de fatoração matricial baseada em modelo vistas nas subseções 2.2.3.2.1 e 2.2.3.2.2.

Ademais, o autor define em seu trabalho que o interesse do modelo é o de prever dados binários, ou seja, classificar os resultados. Dessa forma, a função preditora ($f(x)$) utilizada foi a sigmoide, como pode ser observado na Equação 2.13.

$$\hat{r}_{ui} = f(q_u \cdot p_i + b_u + b_i) \tag{2.12}$$

$$f(x) = \frac{1}{1 + exp(-x)} \tag{2.13}$$

Por fim, o objetivo de otimização do modelo consiste em maximizar a probabilidade dos dados condicionais aos parâmetros estabelecidos, dada pela Equação 2.14. Finalmente, o modelo é então treinado por gradiente estocástico, o mesmo visitado na seção 2.2.3.3 do presente trabalho.

$$L(e^U, e^I, b^U, b^I) = \prod_{(u,i \in S^+)} \hat{r}_{ui} \times \prod_{(u,i \in S^-)} (1 - \hat{r}_{ui}) \tag{2.14}$$

2.2.4 Avaliação de Sistemas de Recomendação

Sistemas de Recomendação são avaliados usando métodos *online* ou *offline*. Em um sistema *online*, as reações do usuário são medidas em relação às recomendações apresentadas. Assim, métodos *online* de avaliação levam em consideração que o usuário participará desse processo ao interagir com as recomendações previstas pelo sistema. A título de exemplo, testes *A/B* e testes de usabilidade são frequentemente aplicados nesse contexto (AGGARWAL, 2016).

Devida a essa característica das avaliações do tipo *on-line*, de dependência de contato com os usuários do sistema recomendador, os métodos *offline* são mais rotineiramente utilizados para avaliar Sistemas de Recomendação do ponto de vista de pesquisa acadêmica. Isso porque tais métodos utilizam os dados históricos do problema de recomendação para realizar a avaliação dos modelos.

Vale ressaltar que todas as entradas observadas de uma matriz de avaliações não podem ser usadas tanto para treinar o modelo quanto para avaliar a precisão do mesmo. Isso porque acarretaria em uma superestimação por parte do modelo devido ao *overfitting*. É necessário, portanto, usar conjuntos de dados diferentes para avaliação e treinamento do modelo, o que ja é comum para ténicas de Mineração de Dados. Para tanto, será aplicada a técnica de Validação Cruzada para a divisão do conjunto de dados de treinamento do modelo e dados de usados para seu teste. Adicionalmente, serão abordadas as métricas de avaliação mais adequadas para uma situação de ranqueamento de produtos.

2.2.4.1 Validação Cruzada

A aplicação da técnica de Validação Cruzada para avaliação de métricas em Sistemas de Recomendação, apresenta particularidades em comparação com seu uso em aplicações tradicionais de *Machine Learning*. A principal distinção reside na natureza dos dados de interação entre usuários e itens. Em aplicações tradicionais, é comum realizar uma separação completa entre conjuntos de treino e teste. No entanto, nos Sistemas de Recomendação, essa abordagem é ajustada para acomodar a natureza implícita e esparsa das interações.

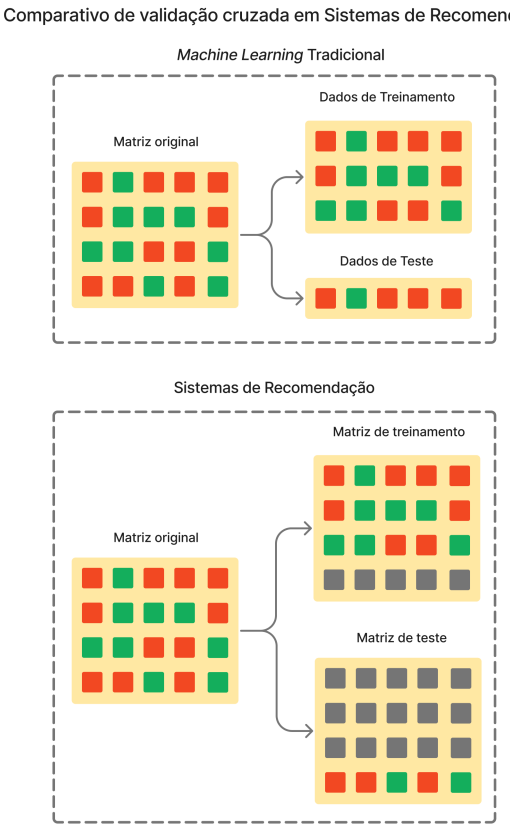
Em Sistemas de Recomendação, a Validação Cruzada é implementada criando duas matrizes de interações a partir da matriz original, com a diferenciação de que as interações destinadas ao teste são ocultadas no conjunto de treino e vice-versa. Tal fato reflete a realidade de que, na prática, os Sistemas de Recomendação precisam lidar com conjuntos de dados incompletos e interações não observadas, exigindo uma abordagem adaptada para avaliação. A *precision@k* e a *recall@k,* por exemplo, são métricas importantes nesse contexto, pois medem a precisão e a abrangência das recomendações, respectivamente, considerando apenas os *top-k*

itens recomendados.

Dessa forma, a Validação Cruzada em Sistemas de Recomendação permite uma avaliação robusta e realista do desempenho do modelo, considerando as nuances inerentes às interações esparsas e à natureza implícita desses dados. Ao ocultar interações de teste durante o treinamento, a validação reflete melhor a capacidade do modelo de fazer recomendações precisas para itens não observados anteriormente. A Figura 14 ilustra essas diferenças para os contextos de aplicações tradicionais de *Machine Learning* em comparação à Sistemas de Recomendação

Nesse sentido, métodos de avaliação por ranqueamento são frequentemente usados na avaliação do consumo real de itens. Portanto, esses métodos são bem adequados para conjuntos de dados de feedback implícito, como vendas, cliques ou visualizações de filmes. Os itens que são eventualmente consumidos também são referidos como verdadeiros positivos ou positivos verdadeiros (KULA, 2015).

Figura 14 – Comparativo de validação cruzada em Sistemas de Recomendação



Fonte: Adaptado de Cates (2019)

2.2.4.2 Métricas de avaliação por Ranqueamento

A arquitetura da avaliação dos modelos por ranaqueamento é baseada na ideia de que o algoritmo de recomendação deve gerar uma lista ranqueada de qualquer

número de itens, ou seja, uma lista dos *top-t* produtos, a partir de um valor numérico atribuído a todos os itens produzido pelo modelo. Para este trabalho, serão utilizadas as métricas de avaliação definidas por Aggarwal (2016) e implementadas por Kula (2015).

Nesse sentido, entende-se que o a relevância dos itens recomendados é baseada no tamanho da lista criada. Logo, alterar o número de itens recomendados na lista classificada tem um efeito direto na relação dos itens recomendados que são realmente consumidos e os itens consumidos que são capturados pelo sistema de recomendação.

Variando o tamanho da lista recomendada, pode-se examinar a fração de itens relevantes na lista e a fração de itens relevantes que são perdidos pela lista. Se a lista recomendada for muito pequena, o algoritmo deixará de capturar itens relevantes, gerando os chamados falsos negativos.

Alternativamente, se uma lista muito grande for recomendada, isso resultará em muitas recomendações equivocadas que nunca serão concretizadas em ações pelo usuário, configurando-se nos chamados falsos positivos. Estes fatores levam a noção de que a avaliação de Sistemas de Recomendação devem buscar compensar as ocorrências de falsos positivos e falsos negativos (KULA, 2015).

Todavia, em um contexto real, o tamanho correto da lista de recomendação não é precisamente conhecido. Mesmo assim, curvas de compensação podem ser quantificadas usando uma variedade de métricas. Assim, no presente trabalho, essa relação pode ser medida por meio da análise das métricas de *precision@k* e *recall@k*.(AGGARWAL, 2016).

Assumindo que se selecione o conjunto dos principais itens classificados para se recomendar ao usuário. Para qualquer valor dado de t produtos a serem recomendados, o que equivale ao tamanho da lista, o conjunto de itens recomendados é denotado por $S(t)$, e representa o conjunto verdadeiro de itens relevantes consumidos pelo usuário.

Ainda, o conjunto G representa os itens relevantes que de fato foram consumidos pelo usuário, ou seja, os casos de verdadeiros-positivos. Assim, para qualquer tamanho t da lista recomendada, a *precision@k* é definida como a porcentagem de itens recomendados que realmente foram consumidos pelo usuário. A Figura 15 ilustra a cobertura de avaliação de ambas as métricas nos espectros dos dados reais em relação aos dados reais.

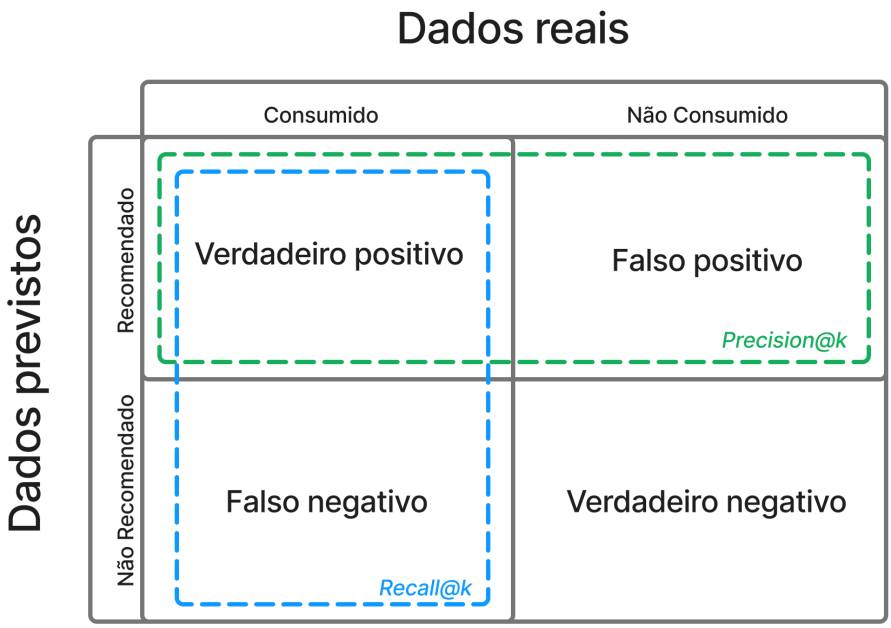
Finalmente, a métrica de *recall@k* é definida de maneira análoga como a porcentagem de verdadeiros-positivos que foram recomendados como positivos para uma lista de tamanho t. A definição de ambas métricas pode ser denotada pelas Equações 2.15 e 2.16, respectivamente.

$$Precision(t) = 100 \cdot \frac{|S(t) \cap G|}{|S(t)|} \tag{2.15}$$

$$Recall(t) = 100 \cdot \frac{|S(t) \cap G|}{|G|} \tag{2.16}$$

Adicionalmente, com o objetivo de avaliar a capacidade preditora do modelo de uma forma mais genérica, as implementações também serão avaliadas pela média da *Receiveing Operator Curve Area Under the Curve* (ROC AUC). Tal métrica não avalia os resultados por meio de geração de listas, mas pode ser útil para entender o comportamento dos dados frente às diferentes configurações do modelo.

Figura 15 – Cobertura das métricas de avaliação de Sistemas de Recomendação



Fonte: Adaptado de Cates (2019)

Os eixos que compõem a *ROC* são a taxa de verdadeiros-positivos, $TRP(t)$, e a taxa de falsos negativos, $FPR(t)$, para o conjunto completo dos dados de teste, ou seja, não se refere a presença nas listas geradas. A primeira delas é definida como a porcentagem de verdadeiros-positivos que foram incluídos na lista de recomendação de tamanho t. Já a taxa de falsos positivos equivale a porcentagem de falso-positivos relatados na lista recomendada em relação aos verdadeiros-negativos, ,ou seja, itens irrelevantes não consumidos pelo usuário.

Portanto, assumindo que v representa o conjunto de todos os itens disponíveis, o conjunto dos valores verdadeiros-negativos é dado por $(v - G)$ e o conjunto dos itens erroneamente recomendados por $(S(t) - G)$, a *False Positive Rate* (taxa de falsos-positivos) é definida na Equação 2.17.

$$FPR(t) = 100 \cdot \frac{|S(t) - G|}{|\upsilon - G|} \tag{2.17}$$

Portanto, a taxa de falsos positivos pode ser entendida como um *recall* invertido, onde o conjunto de verdadeiros-negativos que são capturados incorretamente na lista recomendada, são evidenciados. Finamente, a *ROC* é definida traçando o $FPR(t)$ no eixo X e $FPR(t)$ no eixo Y para diferentes valores de t. Em suma, a *ROC* compara o *recall* desejado contra o *recall* indesejado (AGGARWAL, 2016).

Uma estratégia quantitativa de avaliar a *ROC* se dá pelo cálculo da área sob a curva. Assim, para cada usuário definem-se as curvas com o conjuntos de dados teste, e a média da área de todos os usuários fornece a pontuação *ROC AUC* do modelo. Assim, entende-se que tal métrica está mais relacionada a avaliar o poder de classificação de modelo, do que propriamente se as recomendações realizadas são relevantes. A Equação 2.18 defini o cálculo de tal métrica, na qual N representa o conjunto dos valores de teste.

$$mean\ ROC\ AUC = \frac{1}{N} \sum_{n=1}^{N} ROC\ AUC_n \tag{2.18}$$

O pacote *LightFM* disponibiliza funções para o cálculo dessas três métricas a partir do valor t desejado de itens, do modelo inicial fitado e dos valores recomendados do conjunto de dados teste. Essa funcionalidade do pacote será aplicada no Capítulo 4 do presente trabalho.

2.3 *DESIGN SCIENCE RESEARCH*

O *Design* está relacionado ao ato de inventar e fazer uma uma ideia ou projeto se tornar em algo, envolvendo, portanto, a criação de algo que talvez ainda não exista. Segundo Stevenson (2010), o termo *design* tem seu primeiro registro descrito como um plano ou projeto a ser realizado por um *designer*.

Formalmente, o *design* deve resultar na geração de artefatos de pesquisa, podendo estes serem construtos, modelos, métodos e instanciações. Tais artefatos devem ser elaborados levando em consideração a necessidade de solução de determinado problema ou atingimento de objetivos específicos. Os artefatos fazem a conexão entre um contexto externo, o qual o artefato busca resolver o problema inserido, e um contexto interno, onde a organização do artefato propriamente dita é definida (SIMON, 1996).

Já na visão de Manson (2006), o *design* remete ao processo de construção de um ambiente interno que satisfatoriamente consegue atingir os objetivos definidos pelo *designer* ao ser aplicado à um contexto externo. Tal processo é caracterizado por

rodadas iterativas de tentativa-e-erro até que os resultados do artefato possam ser considerados satisfatórios e resolverem o problema identificado no contexto externo.

Finalmente, a aplicação de conhecimentos para criar e desenvolver um artefato e a avaliação sistemática dos seus resultados frente o problema a ser solucionado configura-se como uma pesquisa, neste caso, a *Design Science Research* (VAISHNAVI; KUECHLER, 2021). Dessa forma, essa abordagem de pesquisa é definida como um conjunto de práticas criativas e analíticas a serem aplicadas em determinado contexto, geralmente em Engenharia de Produção (LACERDA *et al.*, 2013).

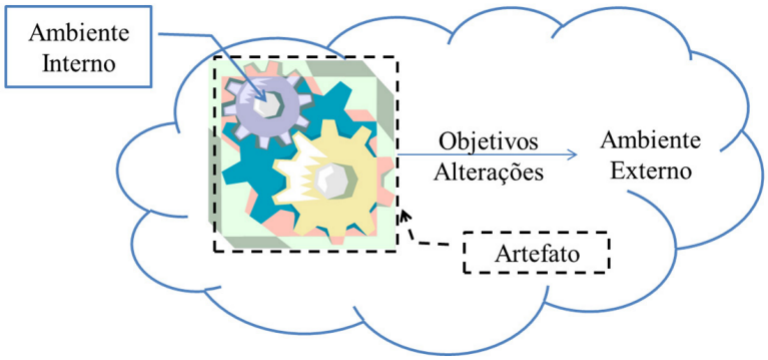
No contexto de solução de problemas em Sistemas da Informação, a *design science research* aplica duas atividades primárias (VAISHNAVI; KUECHLER, 2021):

a) A criação de novos conhecimentos a partir do desenvolvimento de artefatos de pesquisa;

b) A análise da aplicação do artefato e sua performance em consoância com os objetivos a serem atingidos.

A Figura 16 ilustra o artefato de pesquisa como um processo interativo de construção elementos inseridos em um contexto interno a fim de solucionar problemas alocados em um contexto externo. Os objetivos do artefato devem, portanto, nortear a solução do problema de pesquisa.

Figura 16 – Caracterização de um artefato de pesquisa



Fonte: Extraído de Lacerda (2013)

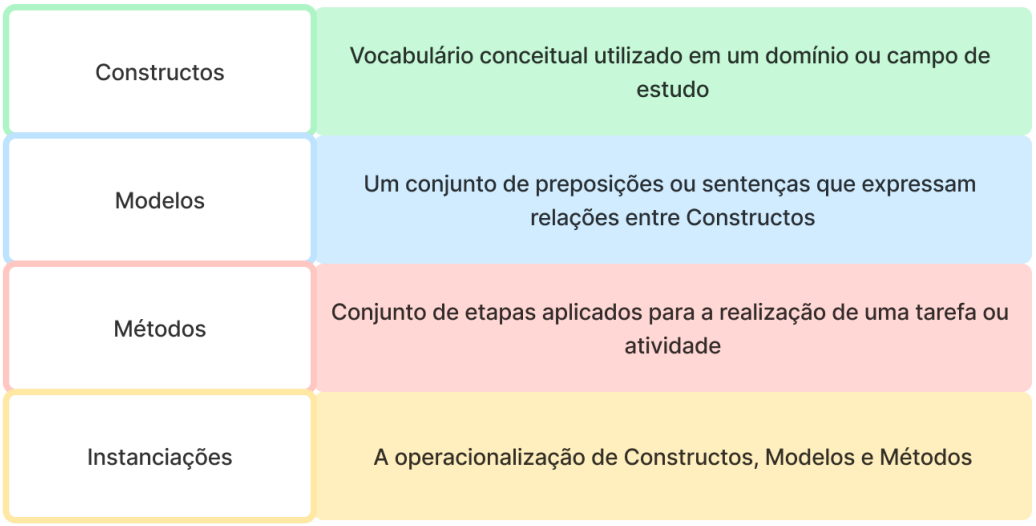
Os artefatos podem ser classficados em 4 tipos: Constructos, Modelos, Métodos e Instanciações. Construtos referem-se à formalização de conceitos dentro docampo de pesquisa do *designer* e surgem justamente durante a formulação e refinamento de um problema. Definem, portanto, as técnicas e as conceitualizações que o artefato se propõe a resolver (MARCH; STOREY, 2008).

Um artefato de tipo modelo é um conjunto de proposições e sentenças que relacionam diversos construtos. Modelos são definidos em termos da sua utilização e os conceitos aplicados ao construtos que são relacionados. Métodos são conjutos de etapas, algoritmos ou protocolos a serem seguidos para que determinada tarefa seja executada. Assim, métodos devem ter o suporte de construtos e e modelos que representam inseridos no ambiente da solução (MARCH; STOREY, 2008).

Uma instanciação equivale a articulação de modelos, métodos e construtos que conceretiza um artefato em seu ambiente. Assim, esse tipo de artefato costuma demonstrar a viabilidade e a eficácia dos modelos e métodos conteplados. A Figura 17 sintetiza os principais tipos de artefatos de acordo com a bibliografia.

A aplicação da *Science Design Research* requer a execução 6 etapas. Na Conscientização do problema, identifica-se claramente um problema a ser solucionado por meio de um artefato de pesquisa. Na próxima etapa, ocorre a o levantamento de alternativas viáveis a serem adotadas como solução do problema identificado (MARCH; STOREY, 2008).

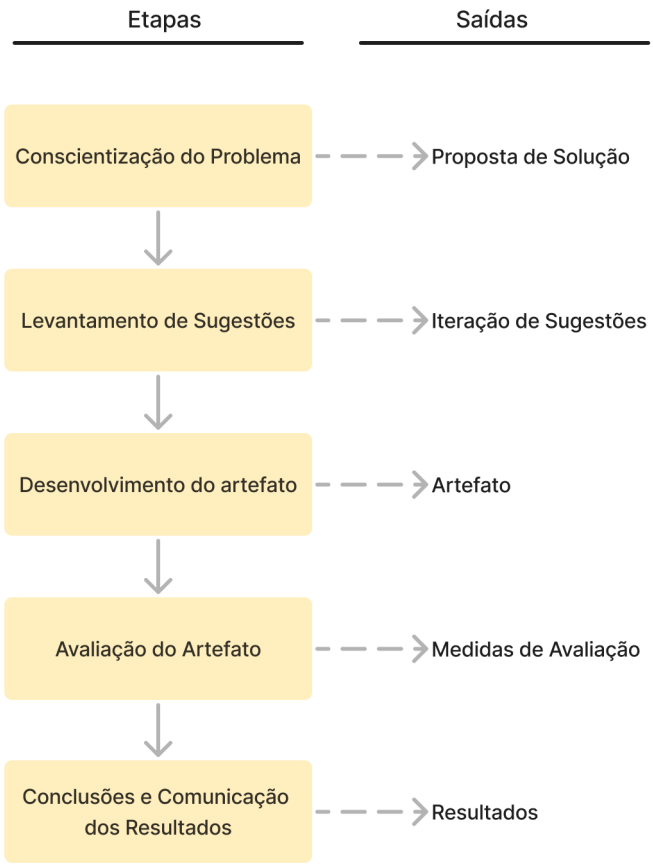
Figura 17 – Tipos de artefatos de pesquisa



Fonte: Adaptado de Manson (2006)

Na fase de desenvolvimento de artefato, definem-se os construtos, modelos, métodos ou instâncias que trazem uma solução satisfatória ao problema, para em seguida, serem avaliados em função da sua utilidade e resultados da aplicação. Por fim, a Conclusão indica o fim de um ciclo de pesquisa ou o encerramento de um projeto específico de pesquisa em ciência do design, e os resultados alcançados devem ser concluidos e comunicados à comunidade acadêmica. A Figura 18 elenca as etapas da Science Design Research bem como as respectivas saídas obtidas ao final da aplicação da etapa, que serão detalhadamente definidos nas subseções a seguir.

Figura 18 – Método da Design Science Research e suas saídas



Fonte: Adaptado de Vaishnavi (2021)

2.3.1 Conscientização do Problema

A detecção de um problema de pesquisa pode surgir de diferentes fontes, como a partir de problemas identificados dentro de organizações, abrindo um contexto de aplicação, ou situações contextualizadas por áreas de pesquisas acadêmicas. Ademais, a etapa de conscientização do problema também deve buscar compreender a problemática detectada (VAISHNAVI; KUECHLER, 2021).

Ainda, a revisão bibliográfica em discipllinas relacionadas ao problema também pode fornecer a oportunidade de aplicar novas descobertas no campo do pesquisador. O resultado da Conscientização do Problema é, portanto, a definição e a formalização do problema a ser solucionado, suas fronteiras e delimitações, bem como as soluções satisfatórias necessárias (LACERDA *et al.*, 2013).

2.3.2 Levantamento de Sugestões

A próxima etapa da metodologia da *design science research* está vinculada às atividades de levantar alternativas de artefato adequadas para a solução

dos problemas. Dessa forma, o *output* da fase de Levantamento de Sugestões são análises relativamente estruturadas sobre alternativas apropriadas a serem desenvolvidas e avaliadas.

Do ponto de vista de Manson (2006), tal etapa configura-se como essencialmente criativa na qual diferentes *designers* tendem a gerar entendimentos distintos frente a um mesmo conjunto de observações. Dessa forma, este o levantamento das alternativas de artefato pode ser considerado relativamente subjetivo e difícil de padronizar.

Portanto, a fim de garantir validade e rastreabilidade para a pesquisa, faz-se necessário justificar criteriosamente a escolha das alternativas de artefato. Essa justificativa pode ocorrer via protocolo pré-estabelecido, geralmente pertencente à área específica de pesquisa do projeto de *design science research* em questão, ou a partir de fundamentações vistas em levantamentos bibliográficos (LACERDA *et al.*, 2013).

2.3.3 Desenvolvimento do artefato

O desenvolvimento do artefato equivale ao processo de construção do artefato da pesquisa propriamente dito. Para Simon (1996), nesta etapa o *designer* define o contexto interno do artefato a partir das configurações do contexto externo e dos objetivos que foram previamente explicitados na fase de Conscientização do Problema.

As técnicas aplicadas durante o desenvolvimento podem variar de acordo com a área de pesquisa específica, a depender dos artefatos em construção. Alguns exemplos de artefatos de pesquisa são algoritmos computacionais, representações gráficas, protótipos, maquetes em escala, entre outros. Por fim, o resultado da etapa do Desenvolvimento é o artefato de pesquisa em estado funcional.

2.3.4 Avaliação do artefato

Após o desenvolvimento do artefato, o mesmo deve ser validado por critérios implícitos ou explícitos que podem estar formalizados na proposta de problema realizada na primeira etapa, sendo que qualquer resultado diferente do esperado requer uma análise e avaliação dos motivos. As hipóteses iniciais que são formuladas durante a formalização do problema e desenvolvimento do artefato raramente são validadas em sua totalidade (MARCH; STOREY, 2008).

A validação do artefato deve seguir um procedimento extensivo de verificação dos resultados obtidos a partir do desenvolvimento amparado na filosofia pragmática. Para isso, se faz necessário formalizar os seguintes elementos para a garantia de um processo de avaliação adequado (LACERDA *et al.*, 2013):

a) O ambiente interno, o ambiente externo e os objetivos explicitamente;

b) Como o artefato pode ser testado;

c) Os mecanismos de medição dos resultados.

Os critérios pelos quais o artefato é avaliado são estabelecidos pelo *designer* de acordo com o ambiente na qual o artefato irá operar. De maneira análoga a outras abordagens de pesquisa, a avaliação de um artefato projetado requer a definição de métricas apropriadas e possivelmente a coleta e análise de dados adequados.

Os artefatos podem ser avaliados em termos de funcionalidade, completude, consistência, precisão, desempenho, confiabilidade, usabilidade, adequação à organização e outros atributos de qualidade relevantes. Um resumo das possíveis técnicas de avaliação é fornecido na Figura 19.

Figura 19 – Tipos de avaliação de artefatos de pesquisa

Classe	Técnicas
Observacional	Estudo de Caso: Avaliar profundamente artefato em um contexto de negócio Estudo em Campo: Monitorar utilização do artefato em diversos projetos
Analítica	Análise Estática: Avaliar atributos estáticos do artefato, como complexidade Análise estrutural: Avaliar a arquitetura técnica do artefato Análise Dinâmica: Avaliar atributos dinâmicos do artefato, como performance
Experimental	Experimento Controlado: Avaliar artefato em um ambiente controlado Simulação: Executar artefato com dados artificiais
Por testes	Funcionais: Utilizar informações vistas em bibliografia para construir um argumento de utilidade ao artefato Estrututurais: Realizar testes de cobertura em métricas de implantação do artefato

Fonte: Adaptado de Vaishnavi (2021)

2.3.5 Conclusão do artefato e comunicação dos resultados

Por fim, a última etapa da *Design Science Research*, consiste na conclusão do processo de pesquisa e sua comunicação às comunidades de interesse, sendo sintetizados todos os seus procedimentos e métodos de condução, além justificar as escolhas realizadas pelo *designer* (LACERDA *et al.*, 2013).

Proposta por Hevner (2004), tal etapa deve apresentar os resultados da pesquisa para as comunidades acadêmica e empresariais, sendo uma componente importante para o desenolvimento das áreas de pesquisa. Assim, a pesquisa em

Design Science deve ser apresentada tanto para o público mais orientado à tecnologia quanto para aquele mais orientado à gestão.

Em publicações acadêmicas, a estrutura dessa etapa é geralmente aplicada na forma da estruturação do trabalho acadêmico em si, assim como uma visto em processos de pesquisa empírica, com definição do problema, revisão da literatura, geração de hipóteses, coletas de dados, resultados e conclusão (PEFFERS *et al.*, 2007).

2.4 CONSIDERAÇÕES FINAIS DO REFERENCIAL TEÓRICO

Os conhecimentos brevemente visitados ao longo do Capítulo 2 serão aplicados para estabelecer os procedimentos metodológicos bem como permitir o correto desenvolvimento do trabalho no contexto da empresa *Olist.* Neste capítulo, a pesquisa explorou conceitos essenciais relacionados à Mineração de Dados para Análise de Negóciose Sistemas de Recomendação.

Inicia-se com a análise das diversas etapas da Mineração desde a coleta até a interpretação de informações. Essa análise de dados assume maior relevância quando aplicada a Sistemas de Recomendação, que vem desempenhado importante papel de personalização da experiência de usuários em contextos de comércio eletrônico.

Nestes sistemas, foram abordadas as estratégias de coleta de informações do usuário, que direcionam a compreensão de suas preferências. O estudo explora as estratégias de recomendação, destacando a associação por conteúdo, por Sequência de Ações e por avaliações. Cada uma dessas estratégias apresenta suas respectivas vantagens e desafios, influenciando diretamente na qualidade das recomendações fornecidas aos usuários.

Além disso, a pesquisa aprofunda-se nas técnicas de Sistemas de Recomendação, elencando as suas características e fatores de aplicabilidade em diferentes cenários. Em particular, foram analisadas a filtragem colaborativa baseada em modelo e a filtragem baseada em conteúdo, duas relevantes técnicas no campo das recomendações. Enquanto a filtragem colaborativa baseada em modelo utiliza padrões de comportamento de usuários para gerar recomendações personalizadas, a filtragem baseada em conteúdo emprega características específicas dos itens recomendados.

Por fim, a pesquisa aborda a avaliação de Sistemas de Recomendação, sendo uma etapa que busca assegurar a eficácia das recomendações. A avaliação desses sistemas objetiva mensurar o desempenho dos modelos, garantindo que os usuários recebam recomendações relevantes. Diversas métricas, como *Precision*, *Recall* e a área sob a curva *ROC (AUC)*, foram levantadas no tópico de avaliação de Sistemas de Recomendação.

Em síntese, o referencial teórico estabelecido neste capítulo fornece uma base

sólida para a pesquisa em questão, oferecendo certa compreensão dos conceitos, estratégias e técnicas relacionados a Sistemas de Recomendação. Novamente, tais conhecimentos serão aplicados durante o desenvolvimento e avaliação dos artefatos propostos nesta pesquisa em *Design Science Research*.

3 METODOLOGIA

Neste capítulo, é realizada a classificação da pesquisa conforme seus objetivos, sua natureza, abordagem metodológica e os procedimentos técnicos utilizados. Ainda, são descritas as etapas realizadas, bem como as técnicas e as ferramentas utilizadas no desenvolvimento da pesquisa.

3.1 CLASSIFICAÇÃO DA PESQUISA

A pesquisa pode ser classificada conforme diferentes perspectivas, segundo a natureza, abordagem metodológica, objetivos e procedimentos técnicos utilizados. Quanto à natureza, a pesquisa desenvolvida é classificada como aplicada, definida por Edna e Menezes (2005) como aquela que, por meio da geração de conhecimentos práticos, busca elaborar soluções para problemas específicos.

Do ponto de vista da forma de abordagem ao problema essa pesquisa caracteriza-se como quantitativa, uma vez que devemos utilizar dados e informações para realizar análises e classificar os objetos de pesquisa (EDNA; MENEZES, 2005).

Em relação aos objetivos, essa pesquisa define-se como exploratória, pois visa explicitar e se aproximar ao problema de personalização de recomendação de produtos em empresas de comércio eletrônico. Segundo Gil (1981), o seu planejamento pode ser flexível, devendo elencar e levar em conta os diversos fatores que podem influenciar os resultados da pesquisa.

Por fim, do ponto de vista dos procedimentos técnicos a pesquisa classifica-se como *design science research,* uma vez que visa solucionar de maneira satisfatória a ausência de recomendações personalizadas de produtos aos clientes da loja por meio de um artefato de pesquisa, possuindo, assim, um caráter prescritivo (DRESCH, 2015).

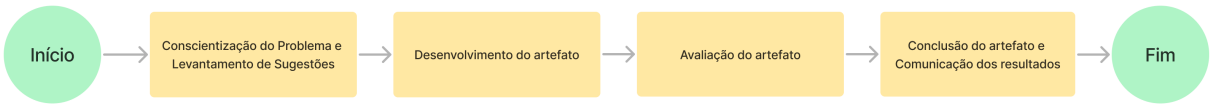
Ademais, o resultado da pesquisa pode ser classificado como instanciação de acordo com Lacerda (2013), uma vez que, quando se há o objetivo de se resolver determinado problema, conceitos, modelos e métodos de Sistemas de Recomendação e análise de dados são articulados de maneira conjunta em um só artefato, que enderece os contextos e objetivos estabelecidos.

3.2 ETAPAS DA METODOLOGIA

As etapas da Metodologia adotadas no trabalho baseiam-se nas etapas de condução da *Design Science Research* vistos em Dresch, Lacerda, Miguel (2015). Para os autores, tal pesquisa deve resultar em um artefato que permita uma solução satisfatória para um dado problema, sendo um processo composto por 6 etapas: conscientização do problema, levantamento de sugestões, desenvolvimento, avaliação

e conclusão do artefato, e por fim a comunicação dos resultados. A sequência destas etapas pode ser verificada no fluxograma da Figura 20.

Figura 20 – Etapas Metodológicas

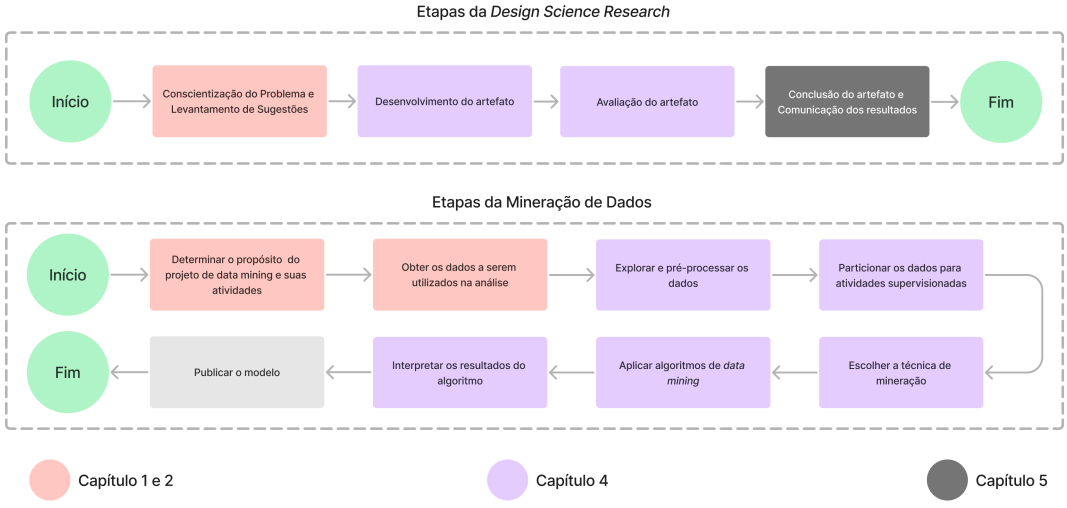


Fonte: Elaborado pelo autor

Levando em consideração que o contexto de solução de problema do presente trabalho é intríseco ao campo de estudo da Gestão Estratégica de Tecnologia da Informação, bem como a *Business Analytics*, se faz necessário expandir os procedimentos de execução de etapas da *Design Science Research.* Isso derá pela adaptação da referência vista na subseção 2.1.1 frente sucessão de ações definidas pelase *DSR.*

Ou seja, dentro de cada execução da *DSR*, técnicas estruturadas de *DM* serão aplicadas para que se obtenham as saídas desejadas, conforme a Figura 19 da subeção 2.3. Dessa forma, as etapas da pesquisa de *Design Science Research* ainda direcionam as etapas técnicas de mineiração de dados, como se observa na Figura 21, o que garante que a metodologia do estudo em questão seja a *Design Science Research.*

Figura 21 – Etapas da Pesquisa e Procedimentos técnicos por Capítulo



Fonte: Elaborado pelo autor

É notável ressaltar que por meio da revisão da literatura, é possível denotar uma sobreposição de algumas atividades propostas, como por exemplo a Conscientização

do Problema em *Design Science Research* frente à Determinação do propósito de projeto de *Data Mining.* Desse modo, algumas dessas etapas não serão vistas de maneira em cascata ao longo desta pesquisa.

3.2.1 Conscientização do Problema e Levantamento de Sugestões

Nesta etapa apresenta-se o contexto do problema de pesquisa, justifica-se a relevância do assunto e define-se uma oportunidade a ser resolvida, para então direcionar a criação do artefato. No presente trabalho, os contextos interno e externo do problema são apresentados na seção introdutória, que expõe a motivação por trás deste desenvolvimento. Em seguida, na seção 1.1 realiza-se a descrição do problema, a apresentação da empresa cujos dados são objeto de estudo, suas limitações e qual uma solução satisfatória para o problema de pesquisa.

Ademais, na segunda seção do Captítulo 2 são elencadas as alternativas de artefato para a solução satisfatória do problema em questão. A escolha das mesmas foi baseada, principalmente, pela presença destas alternativas em outras produções acadêmicas que aplicaram procedimentos experimentais com algoritmos de sistemas de recomendação.

3.2.2 Desenvolvimento do artefato

O desenvolvimento do artefato em si será baseado nos procedimentos técnicos de Shmueli (2017) , para aplicação de métodos de manipulação, partição e Análise Exploratória de Dados. Ainda, os conhecimentos formalizados em Aggarwal (2016) de sequências de atividades que realizam o pré-processamento de dados, modelagem estatística e aplicação dos algoritmos de recomendação também serão aplicados.

A partir da combinação dos conhecimentos acumulados ao longo do Capítulo 2, principalmente daqueles vistos dos autores acima citados, foram estabelecidas as etapas técnicas de Desenvolvimento do artefato de pesquisa, que são a realização da Análise Exploratório dos Dados, o Pré-processamento e Partição dos dados, e finalmente, a Execução dos modelos de Recomendação. Tais procedimentos são definidos nas subseções a seguir e executados ao longo das subseções 4.1.1, 4.1.2 e 4.1.3.

3.2.2.1 Análise Exploratória dos Dados

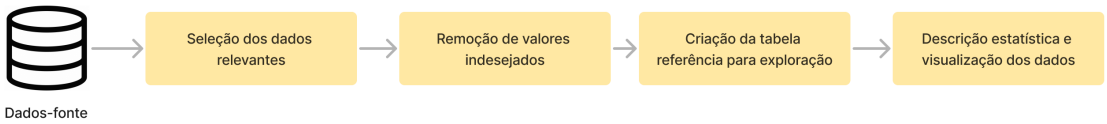
A realização de uma análise exploratória de dados é válida para compreender profundamente o conjunto de dados antes de aplicar modelos de recomendação. Para garantir o sucesso dessa análise, são definidos quatro procedimentos técnicos sequenciais vistos na Figura 22. Inicialmente, a seleção dos dados relevantes direciona o foco da análise para as informações mais pertinentes ao contexto do

problema. Isso ajuda a reduzir a complexidade do conjunto de dados e a concentrar esforços nas variáveis mais impactantes.

Em seguida, a remoção de valores indesejados lida com inconsistências e garante a integridade dos dados. Tal procedimento inclui o tratamento de valores ausentes, *outliers* ou quaisquer outras irregularidades que possam distorcer a análise. A criação de uma tabela de referência para exploração é outra etapa importante, pois organiza as informações de maneira estruturada, facilitando a análise comparativa entre diferentes variáveis e identificação de padrões.

Finalmente, a descrição estatística e visualização dos dados oferece aprendizados e entendimentos mais aprofundados sobre as características do conjunto de dados. Esses procedimentos revelam padrões, tendências e relações entre as variáveis, proporcionando uma compreensão mais holística. A combinação desses quatro passos sequenciais contribui para uma Análise Exploratória robusta, fornecendo a base necessária para o desenvolvimento de modelos de recomendação mais eficazes e adaptados ao contexto específico do conjunto de dados.

Figura 22 – Procedimentos da análise exploratória



Fonte: Elaborado pelo autor

3.2.2.2 Pré-Processamento e Partição dos dados

A aplicação dos processos de pré-processamento de dados e partição dos dados é importante para a utilização de modelos de recomendação, especialmente quando é utilizada a biblioteca *LightFM*. Inicialmente, se obtem matriz esparsa de interação usuários *x* itens, que representa o comportamento histórico dos usuários em relação aos itens.

A partir desta, é necessário criar uma matriz esparsa de interação usuários *x* itens para testes e outra para treinamento, utilizando as técnicas vistas de Validação Cruzada da subseção 2.2.4.1. Essa partição permite avaliar o desempenho do modelo em dados não vistos durante o treinamento, garantindo uma avaliação mais robusta e realista. Felizmente, o pacote *LightFM* disponibiliza uma função que a partir da matriz de interação e da proporção de partição, entrega os conjuntos de treino e teste.

Além disso, na fase de pré-processamento, é necessário obter a matriz de *features* de produtos e usuários, cujas características foram identificadas durante a Análise Exploratória dos Dados. Essas informações adicionais enriquecem

a representação do modelo, permitindo uma personalização mais refinada das recomendações.

Para garantir a compatibilidade com a biblioteca *LightFM*, se faz necessário aplicar uma técnica de mapeamento que converta identificadores de tipo *string* para inteiros em todas as matrizes, pois o pacote não aceita identificadores no formato *string*. Esse processo de mapeamento facilita a manipulação eficiente dos dados e garante a correta interpretação pelo modelo durante o treinamento e avaliação.

3.2.2.3 Execução dos modelos de recomendação

A implementação de modelos de recomendação usando o pacote *LightFM* em *Python* oferece uma abordagem versátil e eficiente para sistemas de recomendação. Inicialmente, são utilizadas as matrizes de interação usuário *x* item como ponto de partida para criar três abstrações distintas que representam estratégias diferentes para recomendações personalizadas. A primeira abstração concentra-se exclusivamente na filtragem colaborativa, utilizando dados da matriz de interação para fornecer recomendações iniciais.

As duas abstrações subsequentes incorporam progressivamente mais informações. O segundo modelo combina filtragem colaborativa com a inclusão de características de produto nos fatores latentes do produto, aprimorando ainda mais a personalização das recomendações. O terceiro modelo aumenta a complexidade ao integrar não apenas características de produto, mas também características de item, oferecendo uma abordagem abrangente que leva em consideração múltiplos aspectos do comportamento do usuário e das características dos itens.

É importante destacar que, em cada modelo, os dados adicionais são incorporados, expandindo as colunas das matrizes de fatores latentes. Posteriormente, a Etapa de Avaliação do artefato analisará cada modelo individualmente, sendo então definidos nas alíneas a seguir.

a) *LightFM* - Filtragem Colaborativa;
b) *LightFM* - Híbrido (Produto);
c) *LightFM* - Híbrido (Produto + Cliente).

Também é possível notar que cada configuração do modelo agrega as diferentes técnicas de Sistemas de Recomendação vistos na subseção 2.2.3. O modelo *LightFM* - Filtragem Colaborativa utiliza os as noções de fatoração matricial baseada em problema de otimização por Gradiente Estocástico Descendente.

Os modelos *LightFM* - Híbrido (Produto) e *LightFM* - Híbrido (Produto + Cliente) adicionam às matrizes latentes vetores com valores das *features* obtidas dos dados de produtos e clientes. Dessa forma, cada distinta *feature* definida é vista como um vetor

a ser incorporado nas matrizes de fatores latentes, conforme denotado na subseção 2.2.3.4.5.

3.2.3 Avaliação do artefato

A avaliação dos resultados do artefato perante o contexto da empresa de comércio eletrônico ocorrerá na subseção 4.2 deste trabalho, por meio da análise de métricas de desempenho dos modelos estatísticos desenvolvidos. Essa avaliação fornecerá aprendizados sobre o desempenho e a eficácia de cada abordagem obtidas com a etapa de Desenvolvimento do artefato, permitindo ajustes e otimizações para garantir recomendações precisas e relevantes para os usuários.

Com objetivo de se realizar uma análise estruturada e criteriosa, de acordo com que se recomenda em estudos de *Design Science Research*, a avaliação do artefato ocorrerá seguindo os procedimentos técnicos vistos em Cremonesi, Koren e Turin (2010), os quais defendem que a avaliação dos modelos de recomendação devem ocorrer por metodologias que combinem avaliações por Método de Ranqueamento e métricas de de acurácia.

Adicionalmente, será aplicada uma Avaliação por comparativo de recomendações a clientes selecionados. Ou seja, para um um mesmo usuário, serão comparadas as recomendações feitas pelos dois modelos com melhor desempenho identificados na etapa de Avaliação por Método de Ranqueamento. Esse processo será aplicado em outro usuário com perfil de compra distinto para que a haja uma avaliação completa em relação às diferentes categorias de produtos consumidas. O detalhamento de ambos os métodos de avaliação é realizado nas próximas subseções.

3.2.3.1 Avaliação por Método de Ranqueamento

A avaliação do Sistema de Recomendação por meio do Método de Ranqueamento é necessária para medir sua eficácia em fornecer recomendações relevantes, principalmente quando se deseja gerar listas de recomendações ao usuários. Utilizando as métricas *precision@k* e *recall@k*, onde *k* irá representar os primeiros 5, 10 e 15 produtos recomendados, é possível avaliar a qualidade das sugestões em diferentes pontos de corte.

Assim, será possível entender a proporção de itens relevantes entre os *k* primeiros recomendados e a proporção de itens relevantes que foram corretamente recuperados até o *k*-ésimo ponto. Tais métricas fornecem aprendizados sobre a precisão e a abrangência das recomendações, considerando diferentes níveis de relevância para os usuários.

Adicionalmente, a avaliação será enriquecida ao incorporar a média da Área Sob a Curva ROC de todos os usuários. Desse modo, é possível entender a

capacidade dos modelos de distinguirem entre itens relevantes e não relevantes. Ao calcular a média da ROC AUC para todos os usuários, se obtem uma medida agregada do desempenho do sistema de recomendação em termos de discriminação entre itens.

3.2.3.2 Avaliação por comparativo de recomendações a clientes selecionados

Para viabilizar uma análise comparativa de recomendações geradas ao mesmos clientes, serão escolhidos 2 usuários, um deles com diversas compras em seu histórico e outro com apenas uma. Para estes usuários, serão previstas 5 recomendações em forma de lista de produtos, e os resultados serão comparados, destacando disparidades e semelhanças entre as categorias dos produtos recomendados pelos modelos.

Juntamente com as previsões de recomendação serão visualizados produtos, e suas respectivas categorias, que de fato foram adquiridas pelos usuário, ou seja, os verdadeiro-positivos do conjunto de dados. A estratégia de apresentar tanto os positivos conhecidos quanto as recomendações sugeridas pelos modelos permite uma comparação direta entre as escolhas dos modelos e os comportamentos reais de compra dos usuários.

Em suma, tal abordagem permitirá avaliar como os modelos se comportam em uma variedade de cenários, abordando a diversidade de comportamentos de compra dos usuários. Ainda, se mostra adequado esse tipo de validação no contexto dos dados coletados, com poucos usuários com mais de um pedido realizado na plataforma de comércio eletrônico.

3.2.4 Conclusão do artefato e Comunicação dos resultados

Neste trabalho o processo conclusão do artefato ocorre na subseção 5.1, ao descrever as considerações finais de cada capítulo de maneira resumida, e também na subseção 5.2, na qual são elencados os principais pontos de atenção observados ao longo do desenvolvimento do trabalho de conclusão de curso, os quais podem ser endereçados em futuros trabalhos acadêmicos.

Quanto à comunicação dos resultados, é assumido que o próprio trabalho derivado deste processo de desenvolvimento de artefato de pesquisa, materializa-se como a etapa de comunicação. Para elucidar a adequação do método frente ao problema, a Tabela 1 elenca especificamente quais os objetivos de pesquisa que deverão ser abordados ao longo das etapas de atividade da *Design Science Research,* que por fim corroborará para o desenvolvimento do artefato de pesquisa.

Tabela 1 – Etapas por Objetivos de Pesquisa

Etapas por Objetivos de Pesquisa	Descrever os processos de desenvolvimento e avaliação.	Aplicar técnicas de recomendação.	Avaliar a performance dos modelos matemáticos	Gerar recomendações personalizadas de compra de produtos para clientes.
Conscientização do Problema	X			
Levantamento de Sugestões	X			
Desenvolvimento do artefato		X	X	
Avaliação do artefato		X	X	
Conclusão do artefato		X	X	X
Comunicação dos resultados				X

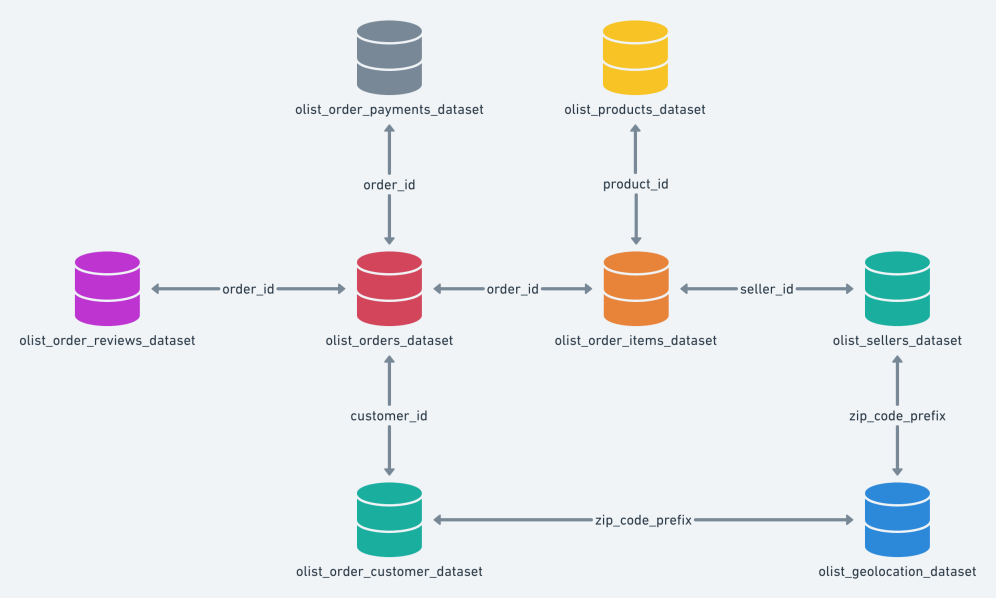
3.3 DADOS COLETADOS

A amostra de dados utilizadas neste trabalho, que pertence a plataforma de comércio eletrônico *Olist*, pode ser encontrada publicamente no portal da *Kaggle*, uma empresa que realiza competições e treinamentos na área de análise de dados e *Data Science* (OLIST; SIONEK, 2018).

Este conjunto de dados é frequentemente utilizado na academia como uma amostra para resolução de diversos problemas de ciência de dados, como pode ser observado em Fortunato (2022), um trabalho de conclusão de curso centrado na segmentação de clientes da base por meio de métodos particionais. Na Figura 23 são apresentadas as tabelas e seus respectivos relacionamentos entre si.

Composto por diversas tabelas inter-relacionadas, o conjunto de dados pode proporcionar aprendizados valiosos sobre transações, produtos, vendedores e clientes associados ao Olist. A tabela *olist_orders* centraliza informações sobre pedidos, como identificação do pedido, datas, status, e custos associados. Outra tabela importante é a *olist_customers*, que fornece detalhes sobre os clientes, incluindo identificação, localização e data de cadastro.

Figura 23 – Esquema dos dados da Olist



Fonte: SIONEK (2018)

Além disso, há tabelas específicas que detalham informações sobre produtos (*olist_products*), vendedores (*olist_sellers*), e revisões dos clientes (*olist_order_reviews*). Essas tabelas agregam detalhamento ao conjunto de dados de pedidos, incluindo categorias de produtos, informações de frete, e avaliações dos clientes, respectivamente. A relação entre essas tabelas possibilita a análise de diversas perspectivas, desde o desempenho dos vendedores até a satisfação do cliente.

O conjunto de dados *Olist* também contém informações geográficas, como estados e cidades, o que permite análises regionais. A qualidade e a variedade dos dados oferecem oportunidades para a Análise Exploratória, modelagem preditiva e o desenvolvimento de sistemas de recomendação personalizados.

Ademais, tal conjunto de dados é significativo para estudos que buscam compreender o ecossistema do comércio eletrônico no Brasil e desenvolver soluções assertivas para aprimorar operações e experiência do cliente. Os arquivos de cada conjunto de dados são disponibilizados na plataforma *Kaggle* no formato *.csv* (Valores Separados por Vírgula), logo, os mesmos podem ser carregados para análise por diferentes ferramentas de manipulação de dados, como *Python* e *R*. As informações contidas em cada um dos conjuntos de dados é mostrado na Tabela 2.

Tabela 2 – Documentação das tabelas fonte a serem utilizadas pelo artefato

Nome da Tabela	Descrição
olist_customers_dataset	Informações do cliente que efetuou o pedido, bem como seu endereço e chave estrangeira para relacionamento com os dados de pedidos.
olist_geolocation_dataset	Conjunto de CEPs brasileiros.
olist_sellers_dataset	Dados geográficos dos fornecedores dos produtos pedidos pelos clientes da loja.
olist_orders_dataset	Registro dos pedidos feitos na Olist entre 2016 e 2018, contendo detalhes como data do pedido, identificador do pedido, identificador do cliente e quantidade de produtos pedidos.
olist_order_items_dataset	Registros dos produtos transacionados em cada um dos pedidos, incluindo peso e valor monetário.
olist_order_reviews_dataset	Conjunto das avaliações realizadas pelos clientes frente à sua experiência com os pedidos feitos na plataforma *Olist*.
olist_products_dataset	Detalhamento dos produtos disponíveis na loja, como tamanho, modelo, cor e categoria.
category_name_translation	Tabela suporte para tradução das categorias dos produtos entre Inglês e Português.

3.3.1 Ferramentas de manipulação dos dados

Este estudo utilizará a linguagem de programação *Python* como ferramenta de suporte para as atividades de manipulação, exploração e modelagem dos dados da loja de comércio eletrônico. A adoção desta linguagem se mostra vantajosa para elaboração da pesquisa, uma vez que a mesma dispõe de diversos pacotes da comunidade para aplicação em contexto de *Data Science*. Neste trabalho, os pacotes utilizados são delimitados na alínea a seguir.

a) *LightFM*: é um pacote de algoritmos de recomendação frequentemente utilizado para implementação e avaliação de sistemas de recomendação, adotado em meio acadêmico e por organizações de forma geral (KULA, 2015);

b) *Pandas*: O *pandas* facilita procedimentos de manipulação de dados como *merges*, *join* e tratamento de dados faltantes (MCKINNEY, 2012);

c) *Matplotlib*: Biblioteca aplicada para criar gráficos estáticos, mas também visualizaçãoes animadas e interativas em *Python* (ARI; USTAZHANOV, 2014);

d) *Surprise*: Pacote desenvolvido para facilitar o procedimento de aplicação de algoritmos de predição de *ratings*, possuindo também ferramentas baseadas

em *sci-kit development* de avaliação dos modelos, como validação cruzada e métricas nativas (HUG, 2020).

3.4 DELIMITAÇÕES

O artefato não inclue a disponibilização das recomendações em um ambiente de utilização e integração contínua, como por exemplo, os procedimentos necessários para hospedar a execução do modelo e entregar a recomendação à um cliente específico autenticado no site da loja. O processo de *deployment* e *Machine Learning Ops* pode ser constituido em um novo artefato de pesquisa, sequencial ao presente estudo.

Isso se deve à complexidade de operação e manutenção de sistemas de aprendizado de máquina, bem como à estrutura e processos relacionados a automatização dos *inputs* (coleta e tratamento dos dados de *reviews* de pedidos da loja) e dos *outputs* (Geração da recomendação a nível de cliente e disponibilização da informação no *website*) de modelos de aprendizagem-máquina (OLGA; MONTEIRO, 2021).

4 DESENVOLVIMENTO E AVALIAÇÃO DO ARTEFATO

A metodologia da *Design Science Research* define etapas de desenvolvimento e avaliação de artefatos, na qual deve resultar em uma solução satisfatória ao problema de personalização da recomendação de produtos da *Olist,* através de um Sistema de Recomendação.

A etapa de desenvolvimento envolve, portanto, a implementação do Sistema de Recomendação. Uma vez que o sistema é construído, é necessário avaliá-lo de maneira criteriosa. As métricas de desempenho, como *precision*, *recall* e *AUC*, são utilizadas para medir a eficácia do sistema. Adicionalmente, um comparativo de recomendação entre diferentes clientes agregará aprendizados sobre a tendência de previsão de cada modelo.

Essa avaliação contínua permite ajustar e otimizar o artefato, garantindo que ele atenda aos objetivos estabelecidos inicialmente. A *Design Science Research* fornece uma estrutura sólida para abordar o problema de personalização da Olist por meio da criação e aperfeiçoamento do Sistema de Recomendação, garantindo que o mesmo seja eficaz, relevante e capaz de melhorar a experiência do usuário.

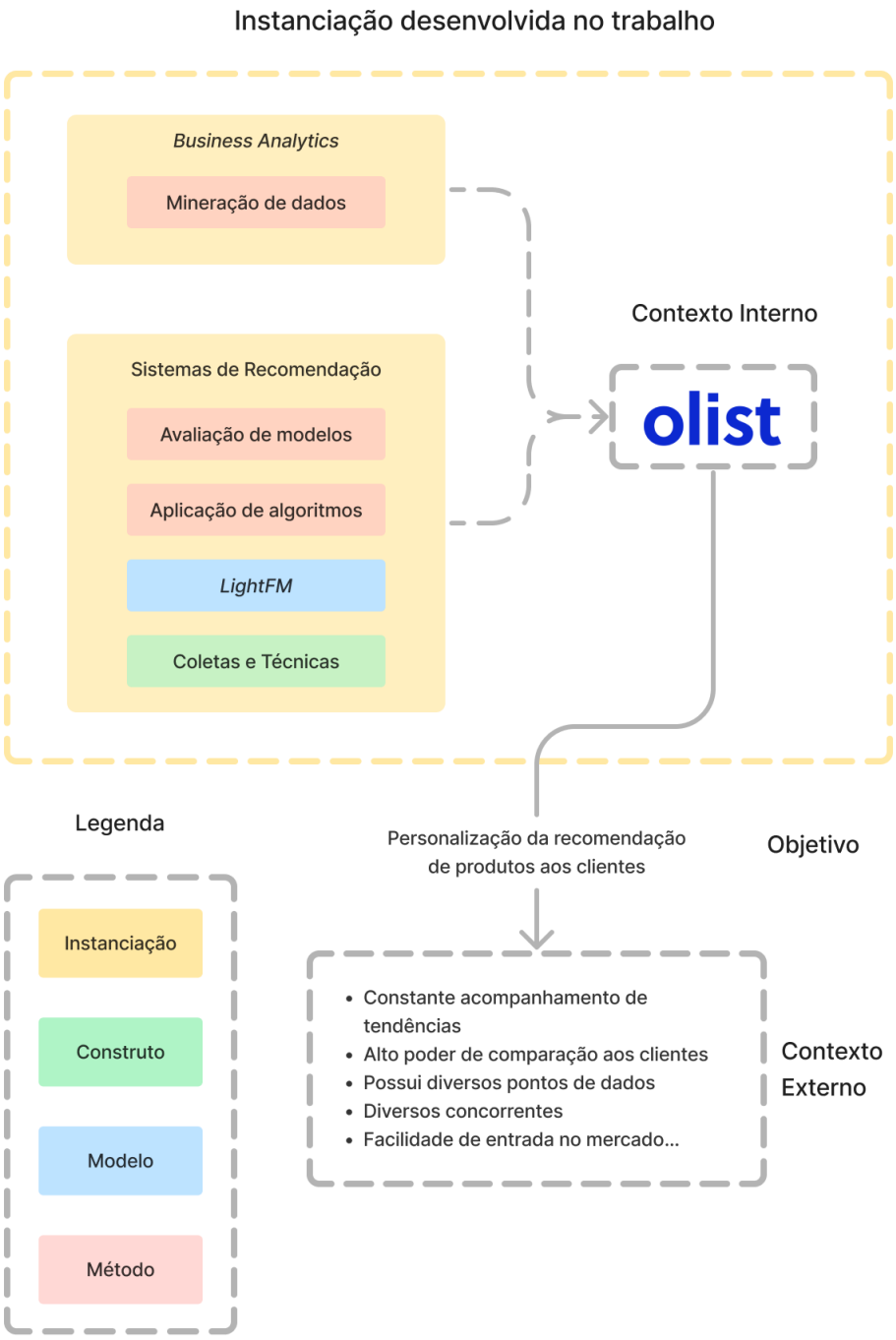
Ademais, também é possível definir a instanciação em construção no presente estudo, ao agregar os diferentes tipos de artefato, como modelos, construtos e métdos, que acabam por trazer uma solução satisfatória ao problema de personalização de recomendação de produtos na empresa *Olist.*

Avaliando a Figura 24, entende-se que a instanciação deve combinar artefatos das áreas de pesquisa da *Business Analytics* e de Sistemas de Recomendação. Do primeiro campo de conhecimento serão utilizados os métodos da Mineração de Dados durante a etapa de Desenvolvimento do Artefato.

Já no segundo, métodos de Aplicação e Avaliação de modelos de recomendação serão seguidos de forma a possibilitar a geração de listas de recomendação satisfatoriamente precisas aos clientes da base. Para isso, foi necessário o entendimento dos tipos de coletas, estratégias e ténicas de recomendação que foram adotadas ao se utilizar o modelo *LightFM.* O agrupamento destes campos de conhecimento na instanciação a ser desenvolvida objetivou estabelecer um sistema que personalize a recomendação de produtos aos clientes inseridos no contexto interno da empresa *Olist*.

Esse problema a ser satisfatoriamente solucionado pelo artefato, está imerso em um contexto externo de comércio eletrônico, no qual é caracterizado por constante acompanhamento das tendências do usuário, necessidade de personalização da experiência do usuário na loja e também pelo alto poder de comparação de preços e atributos de produtos conferida aos potenciais clientes.

Figura 24 – Instanciação desenvolvida no presente trabalho



Fonte: Elaborado pelo autor

4.1 DESENVOLVIMENTO DO ARTEFATO

De acordo com o que foi definido no Capítulo 3 do presente trabalho, o desenvolvimento do artefato consistirá em resolver o problema de ausência de personalização de recomendação de produtos para os clientes da empresa *Olist*, a

partir da aplicação dos conceitos vistos no Capítulo 2.2.

Ainda, a sequência de etapas da Mineração de Dados vistas no Capítulo 2.1 foram adaptadas de forma a serem aplicadas em um contexto de desenvolvimento de Sistemas de Recomendação em um estudo da *Design Science Research*. Nesse sentido, o resultado prático do Desenvolvimento do artefato será apresentado nas três seções a seguir.

4.1.1 Análise Exploratória dos Dados

Conforme definido na subseção 3.2.3, os procedimentos técnicos empregados para a realização de uma Análise Exploratória de Dados são executados nas subseções a seguir. A partir disso, é possível se obter aprendizados sobre as caraterísticas dos dados coletados bem como facilitar o posterior processo de aplicação do modelo de recomendação.

4.1.1.1 Seleção dos dados relevantes

A primeira etapa de inciação da Análise Exploratória de dados é realizar uma prévia inspeção nos dados, que conforme já mencionado na subseção 3.3, são obtidos na plataforma *Kaggle* e então carregados pela linguagem *Python*. Tal análise deve ser feita em conjuntos de dados que meçam de alguma forma a experiência de um usuário na plataforma ou em uma de suas dimensões, sendo portanto, relevantes para a realização de tal análise.

Por exemplo, informações sobre os produtos e a localização dos revendedores não se conectam diretamente com um usuário, mas certamente uma compra feita pelo mesmo está conectada à um produto que foi vendida pelo revendedor.

Como este trabalho é focado na avaliação da base de clientes, nem todos os campos são relevantes para análise. Desta forma, utilizando uma abordagem adotada em Fortunato (2022), as colunas serão classificadas em 3 grupos quanto ao seu valor analítico, como pode ser observado nas Tabelas 3, 4, 5, 6 e 7.

a) Alta relevância: O valor dessas colunas é essencial para a análise;

b) Baixa relevância: A ausência dessas colunas não compromete a realização da análise. No entanto, sua inclusão pode oferecer intuições quanto à resposta buscada;

c) Indiferente: Não tem relevância dado o contexto do problema.

Assim, os *datasets* a passarem pela Análise Exploratória de dados serão os de Pedidos, Clientes, Avaliações dos Pedidos, Produtos e Pedidos-Item e as próximas etapas considerarão que apenas as colunas que possuem alta relevância serão efetivamente analisadas.

A Tabela 3 descreve a relevância das colunas no conjunto de dados de pedidos da Olist. A relevância é categorizada em “Alta Relevância” para colunas fundamentais, como *order_id*, *customer_id*, *order_status*, *product_id*, *seller_id*, e *review_id*, e “Baixa Relevância” para colunas menos críticas, como *order_purchase_timestamp* e *shipping_limit_date*.

Tabela 3 – Definição de relevância das colunas da tabela *olist_orders_dataset*

Coluna	Descrição
order_id	Alta Relevância
customer_id	Alta Relevância
order_status	Alta Relevância
order_purchase_timestamp	Baixa Relevância
product_id	Alta Relevância
seller_id	Alta Relevância
shipping_limit_date	Baixa Relevância
review_id	Alta Relevância

Na Tabela 4, são definidas as colunas com suas respectivas relevâncias no conjunto de dados de clientes da *Olist*. Colunas como *customer_id*, *customer_unique_id*, *customer_city* são consideradas de “Alta Relevância”, enquanto *customer_zip_code* é classificado como de “Baixa Relevância”.

Tabela 4 – Definição de relevância das colunas da tabela *olist_customers_dataset*

Coluna	Descrição
customer_id	Alta Relevância
customer_unique_id	Alta Relevância
customer_city	Indiferente
customer_zip_code	Baixa Relevância

A Tabela 5 destaca a relevância das colunas no conjunto de dados de revisões. Colunas como *review_id*, *order_id*, *review_score*, e *review_comment_text* são de “Alta Relevância”. Por outro lado, *review_creation_date* é considerada de “Baixa Relevância”.

Tabela 5 – Definição de relevância das colunas da tabela *review_dataset*

Coluna	Descrição
review_id	Alta Relevância
order_id	Alta Relevância
review_score	Alta Relevância
review_comment_text	Alta Relevância
review_creation_date	Baixa Relevância

Na Tabela 6 são apresentadas as colunas e suas respectivas relevâncias no conjunto de dados de produtos da *Olist*. Colunas como

product_id, *product_category_name*, *product_weight_g*, *product_length_cm* são classificadas como de “Alta Relevância”, enquanto *product_name_length* e *product_description_length* têm uma classificação de “Baixa Relevância”.

Tabela 6 – Definição de relevância das colunas da tabela *olist_products_dataset*

Coluna	Descrição
product_id	Alta Relevância
product_category_name	Alta Relevância
product_name_lenght	Baixa Relevância
product_description_lenght	Baixa Relevância
product_weight_g	Alta Relevância
product_lenght_cm	Alta Relevância

Finalmente, a Tabela 7 fornece informações sobre a relevância das colunas no conjunto de dados de itens de pedidos da Olist. Colunas como *order_id*, *order_item_id*, *product_id*, *seller_id*, *price*, e *freight_value* são consideradas de “Alta Relevância”, com atenção especial às colunas que refletem os itens comprados e suas informações financeiras. A coluna *freight_value* é classificada como de “Baixa Relevância”.

Tabela 7 – Definição de relevância das colunas da tabela *olist_order_items_dataset*

Coluna	Descrição
order_id	Alta Relevância
order_item_id	Alta Relevância
product_id	Alta Relevância
seller_id	Alta Relevância
price	Alta Relevância
freight_value	Baixa Relevância

Esta será a abordagem utilizada uma vez que a ausência de tais colunas seria problemática para a definição do *dataset* da Análise Exploratória ou aplicação dos algoritmos de recomendação. Colunas de baixa relavância poderão ser usadas para complementar a análise exploratória enquanto que as colunas indiderentes não serão analisadas, já que não contribuirão de forma alguma para o estudo.

4.1.1.2 Remoção de valores indesejados

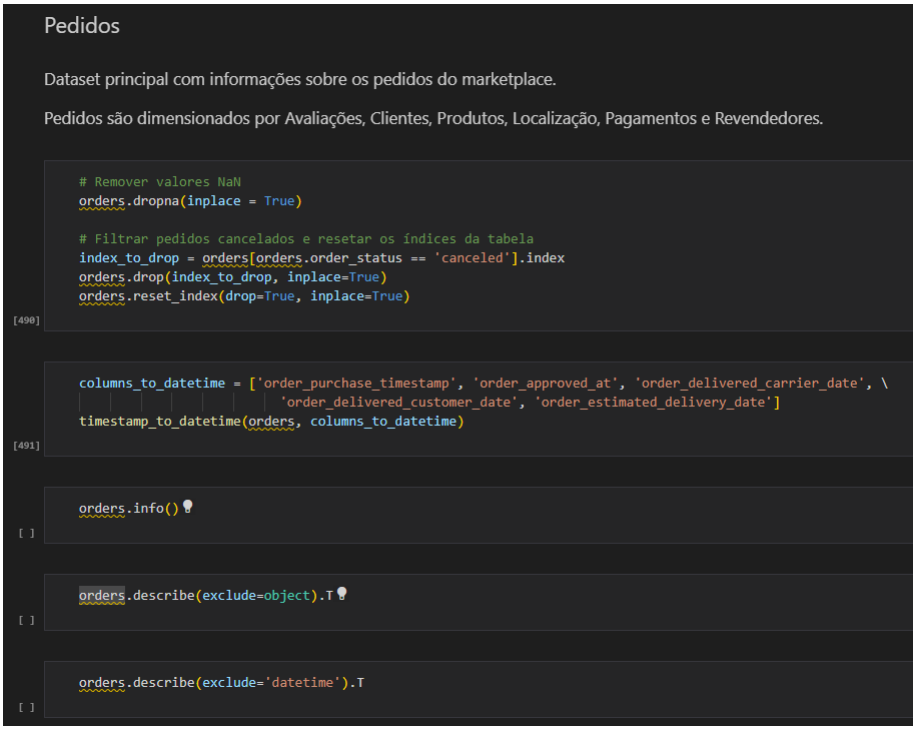
Antes de se iniciar procedimentos de sumarização estatística e visualização dos dados, é necessário avaliar a presença de ruído e definir os *datasets* agregados a partir dos dados relevantes disponíveis, a serem analisados. De maneira geral, esse ruído pode ser entendido como dados faltantes, por exemplo ausência de chave primária de cliente em um pedido, ou ainda dados duplicados, na qual a chave primária da tabela não é respeitada.

Ademais, mesmo que as avaliaçõesn explícitas não sejam *inputs* de modelos de recomendação, pedidos que que não possuíam tal descrição receberam um valor “Sem Avaliação do Pedido”. Isso foi feito com o objetivo de justamente permitir posteriormente uma análise mais aprofundada sobre a atual experiência dos usuários na *Olist*.

Ainda, os pedidos cancelados tamém foram removidos da análise, uma vez que o modelo não pode considerar interações que efetivamente não foram concretizadas. Todas as manipulações acima mencionadas foram realizadas em um *Jupyter Notebook* por meio da ferramenta de linguagem de programação *Python* e seu pacote de manipulação de *dataframes*, *Pandas*.

A figura 25 ilustra as pré-inspeções e transformações nos dados de Pedidos, que foram aplicadas de maneira semelhante aos outros *datasets*, de acordo com o caso específico também ja exemplificados. Adicionalmente, tipos de dados são alterados para permitirem análises temporais.

Figura 25 – Inspeção dos conjunto de dados relevantes



Fonte: Elaborado pelo autor

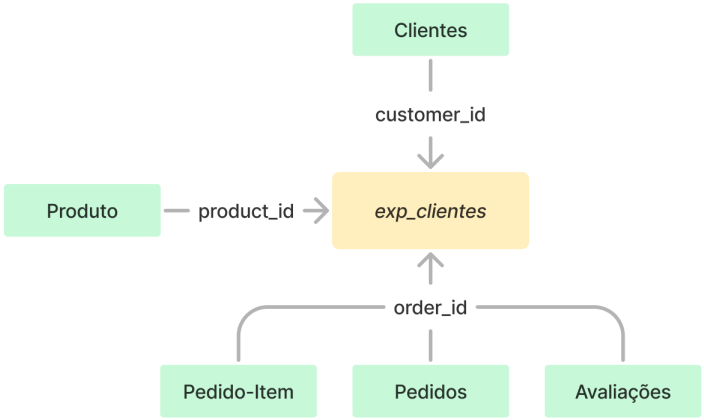
4.1.1.3 Criação da tabela referência para exploração

Em sequência, os datasets são combinados para permitir uma análise abrangente das avaliações dos clientes e seus comportamentos de compra. A integração desses conjuntos de dados visa entender os fatores que podem impactar

o modelo de recomendação e por quê.

A Análise Exploratória busca esclarecimentos sobre as características dos usuários, fornecendo aprendizados valiosos para prever os hábitos de consumo na plataforma Olist, especialmente entre os revendedores cadastrados. A Figura 26 inicialmente ilustra os conjuntos de dados utilizados nessa fase da análise exploratória, oferecendo uma visão geral dos dados a serem considerados.

Figura 26 – Conjuntos de dados a serem utilizados na Análise Exploratória



Fonte: Elaborado pelo autor

A mesma Figura também apresenta o *dataset* criado, que será explorado mais a fundo na análise. Os valores entre as setas conectoras representam as chaves utilizadas como parâmetros para a função *merge()* do pacote *Pandas*. Isso resulta no conjunto de dados *exp_clientes*, onde a coluna indicando a chave primária da tabela é *order_id*. Por fim, a Tabela 8 fornece uma descrição detalhada de cada coluna que compõe a tabela *exp_clientes*, oferecendo uma visão mais clara das informações que serão utilizadas na análise exploratória, enriquecendo a compreensão dos padrões e comportamentos dos clientes na plataforma Olist.

Tal abordagem estruturada na criação da tabela de referência para exploração permite uma análise mais eficiente e facilita a interpretação dos resultados obtidos. A combinação de dados relevantes proporciona uma visão holística, permitindo que o processo de análise exploratória seja mais direcionado e informativo para o desenvolvimento e avaliação do modelo de recomendação LightFM.

Tabela 8 – Descrição das colunas da tabela *exp_clientes*

Coluna	Descrição
order_id	Identificador único para cada pedido
customer_unique_id	Identificador único para cada cliente
review_id	Identificador único para cada revisão de produto
order_purchase_timestamp	Data e hora em que o pedido foi realizado
product_id	Identificador único para cada produto
seller_id	Identificador único para cada vendedor
shipping_limit_date	Data limite de entrega do pedido
customer_id	Identificador único para cada cliente
review_creation_date	Data em que a revisão foi criada
review_score	Pontuação atribuída à revisão
review_comment_text	Texto de comentário da revisão
product_category_name	Nome da categoria do produto
product_name_length	Comprimento do nome do produto
product_description_length	Comprimento da descrição do produto
product_weight_g	Peso do produto em gramas
product_length_cm	Comprimento do produto em centímetros
freight_value	Valor do frete do pedido
price	Preço do produto

4.1.1.4 Descrição estatística e visualização dos dados

Com o intuito de direcionar a análise exploratória, algumas perguntas do negócio da *Olist* são levantadas. Ainda, tais questionamentos devem buscar entender padrões ou hipóteses que podem ser decisivos para a performance dos modelos colaborativos baseados em modelo. As perguntadas levantadas são descritas a seguir.

a) Das Categorias de Produto:

   1. Quais obtiveram maior Receita?
   2. Quais foram as mais bem avaliadas?
   3. Quais possuem o maior volume de vendas?

b) Dos Clientes:

   1. Qual frequência de compra de produtos na loja?

c) Das Avaliações:

   1. Como a média das avaliações se altera ao longo do tempo?
   2. Existem fatores que influenciam a percepção de satisfações dos clientes frente às suas compras?

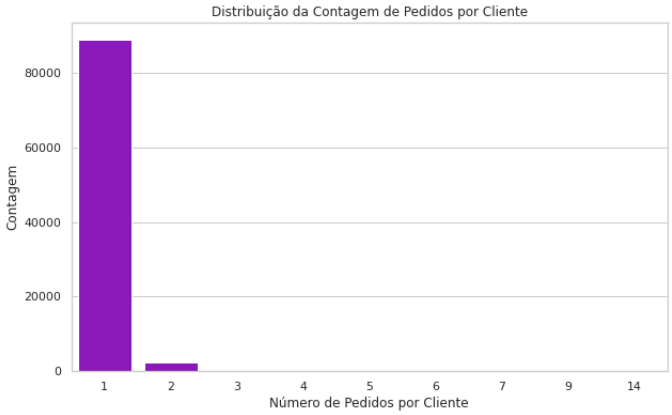
Ademais, por meio de técnicas de visualização de dados e aplicação de estatísticas descritivas ao conjuntos de dados *exp_clientes* será viabilizada a análise

exploratória. Para isso, será utilizado o pacote *Seaborn* da Linguagem *Python* bem como o pacote *Pandas*, nas explorações em *plots* de gráficos e cálculos das estatísticas, respectivamente.

Uma prévia inspeção da contagem de pedidos por cliente, conforme Figura 27, já demonstra que a *Olist* possui a caraterística baixa frequência de compra entre a sua clientela. Isso leva a dois entendimentos. O primeiro é que fica evidente a necessidade da empresa por adotar estratégias que aumentam a fidelização de clientes, uma vez que dos mais de 92.000 pedidos, 88.908 foram realizados por clientes distintos.

A segunda conclusão deste gráfico é que o modelo a ser desenvolvido deverá contar com uma filtragem baseada em conteúdo robusta, uma vez que a quantidade de clientes distintos com apenas uma compra é alta, o que torna mais difícil assimilar suas semelhanças por meio da informação de histórico compra de produtos ou avaliação de um pedido. Dessa forma, é possível responder a pergunta b.1 da alínea de perguntas de negócio, na qual essencialmente não existe taxa de recompra para os clientes da loja.

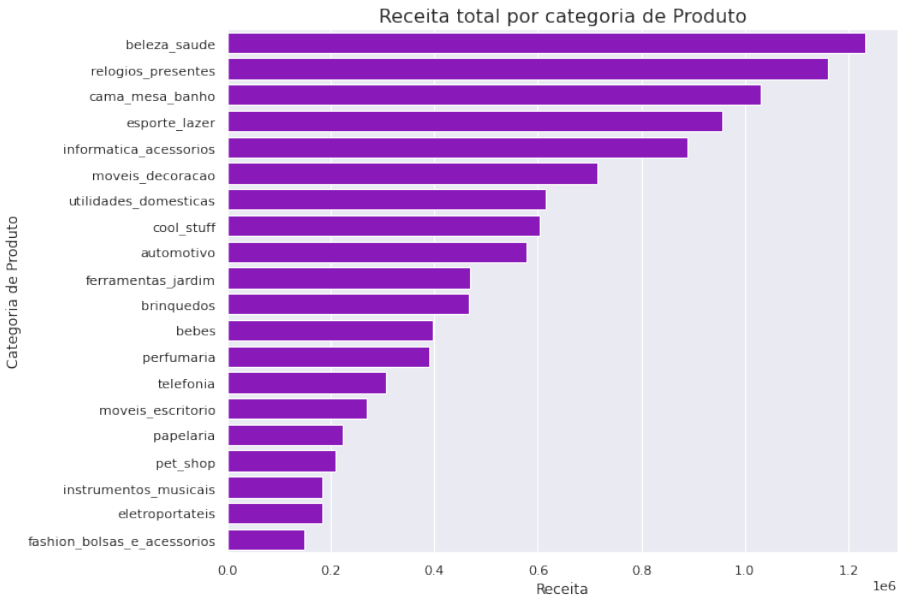
Figura 27 – Gráfico de Distribuição da Contagem de Pedidos por Cliente



Fonte: Elaborado pelo autor

Ao verificar os produtos com maior receita de vendas pela Figura 28, é possível notar que as categorias Cama, Mesa e Banho, Beleza e Saúde, Esporte e Lazer e Relógios e Presentes são as mais significativas dentre as 71 categorias existentes. Adicionalmente, apenas 11 itens obtiveram resultados de receita superior a R$ 400.000,00. Assim, obtem-se a resposta para a alínea a.1 da lista de perguntas de negócio.

Figura 28 – Gráfico de Receita Total por Categoria de Produto



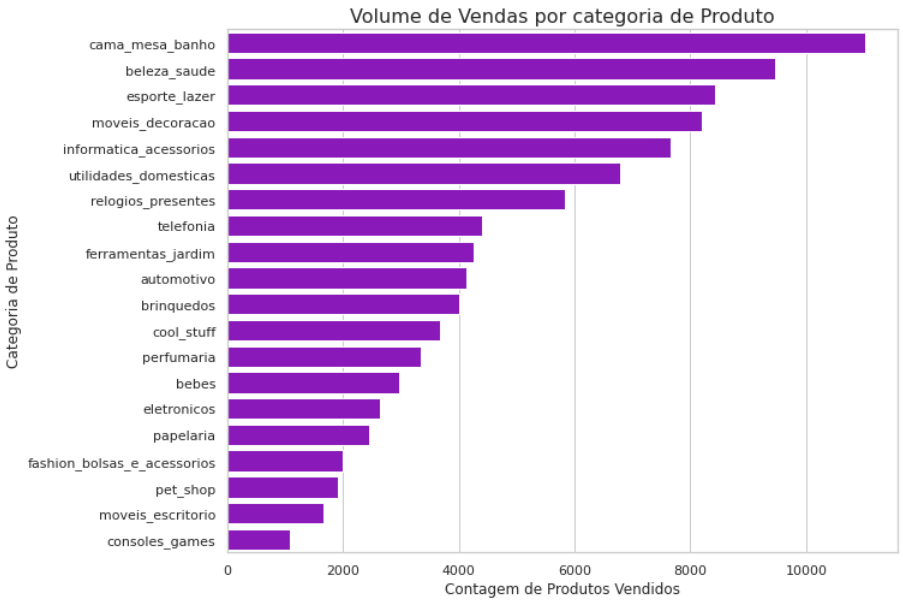
Fonte: Elaborado pelo autor

Todavia, tal relação de produtos com maior receita não é exatamente igual aos produtos com maior Volume de Vendas ao se observar a Figura 29. Mesmo assim, é possível evidenciar que os produtos mais frequentemente comprados possuem certa correlação com produtos que trazem a maior fatia de caixa para a empresa.

A categoria de Relógios e Presentes não entra no grupo das *top* 5 categorias com maior volume, o que implica que o preço médio destes itens é superior ao das outras categorias, e logo, o seu valor agregado é maior. Finalmente, é possível definir que as categorias Cama, Mesa e Banho e Beleza e Saúde são as mais financeiramentes importantes para a empresa, uma vez que ambas se posicionam nos *top* 3 produtos de maior Receita e Volume de Vendas. Logo, obtem-se a resposta para a alínea a.3 da lista de perguntas de negócio.

Outra prévia inspeção dos valores médios das avaliações pelas categorias de produtos ilustra o fato de que diversos produtos de maior faturamento e volume de vendas estão na extremidade inferior da pontuação média das análises, com exceção das categorias Saúde e Beleza e Esportes e Lazer, que mesmo assim não figuram entre as *top* 5 categorias. Finalmente, Assim, obtem-se a resposta para a alínea a.2 da lista de perguntas de negócio.

Figura 29 – Gráfico de Volume de Vendas Total por Categoria de Produto



Fonte: Elaborado pelo autor

Adicionalmente, vale ressaltar que as categorias melhor avaliadas pelos clientes da *Olist* não representam os produtos mais frequentemente comprados ou com maior relevância financeira para a companhia. Tal fato pode levantar a hipótese de que talvez tais produtos não são devidamente ofertados ou recomendados ao clientes, por exemplo. Essa noção pode ser obtida analisando a Figura 30.

Com essa disparidade, percebe-se a necessidade de um entendimento aprofundado das interações entre clientes e vendedores da *Olist*, na qual será realizada uma breve análise das compras dos clientes ao longo do tempo, com o objetivo de agregar as tais interações aos clientes, produtos e avaliações. Com isso, será possível inferir se existe uma relação entre a venda dos produtos e a satisfação dos clientes que deve ser incorporada ou não ao modelo de recomendação.

O gráfico da Figura 31 ilustra uma linha temporal do Volume de Pedidos da base dados em comparação com a Média das Avaliações dos clientes no mesmo período, de 15/09/2016 até 29/08/2018. Neste gráfico é possível denotar uma correlação diretamente proporcional entre estes dois indicadores. As linhas verticais pontilhadas de cor vermelha indicam os principais pontos de correlação.

O intervalo entre a semana 45 e 49 de 2017 mostra que em situações de grande volume de pedidos, a satisfação dos clientes com relação aos produtos consumidos tende a cair. O mesmo pode ser observado nas semanas 29 e 37 de 2017 e ainda nas semanas 7, 23 e 31 do ano de 2018. Alternativamente, a queda das vendas na semana 53 de 2017 não indicou piora nas avaliações, mas sim, relativa melhora em sua média.

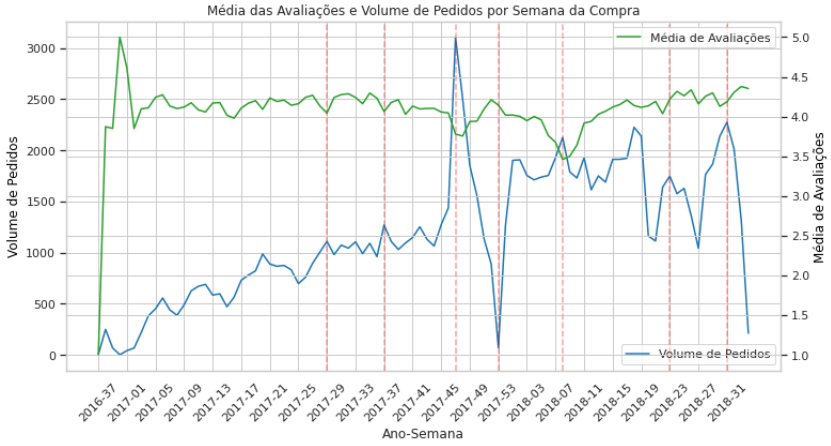
Figura 30 – Gráfico da Média das Avaliações dos Clientes por Categoria de Produto

Fonte: Elaborado pelo autor

Ao realizar uma equiparação destas semanas com períodos de alta demanda do comércio no calendário nacional, é possível entender que o pico observado na semana 47 de 2017 corresponde ao dia da *Black Friday*, que promove uma temporada de compras pré-Natal com significativas promoções em muitas lojas varejistas e de Comércio Eletrônico.

Por fim, é possível denotar que a base de clientes da *Olist* aumentou significativamente apos tal evento, já que no período pós-*Black Friday*, que corresponde a menos da metade das semanas de toda a base de dados, o número de pedidos é quase que o dobro quando comparado ao período anterior à este evento. Esse entendimento pode ser visualizado na Figura 32.

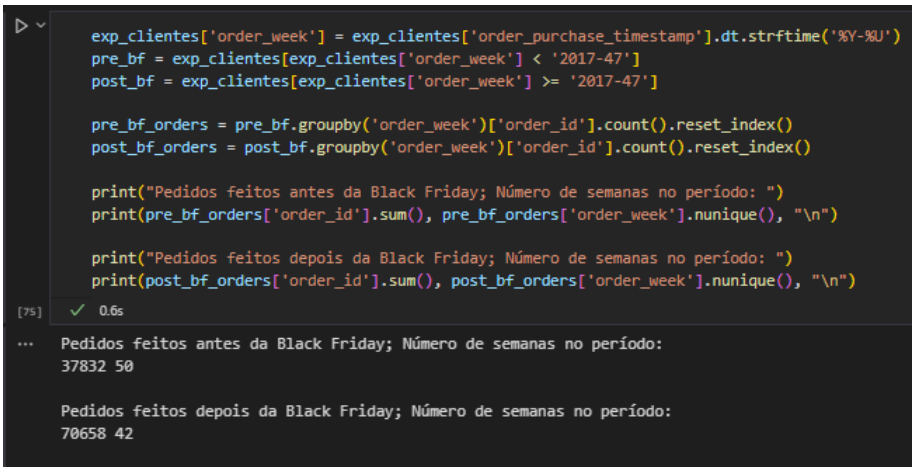
Figura 31 – Gráfico da Média das Avaliações dos Clientes e Volume de Pedidos por Semana de Compra



Fonte: Elaborado pelo autor

Pensando em determinar um fator que conecte o alto volume de vendas com a instatisfação dos clientes com seus pedidos, a Figura 33 adiciona à análise temporal uma linha que corresponde ao total de pedidos atrasados por semana. Com essa visualização, é possível perceber que a ocorrência de atrasos nos pedidos é uma componente totalmente relacionada à experiência do cliente com a loja, uma vez que tais eventos coincididem com quedas da média das avaliações dos pedidos.

Figura 32 – Total de Pedidos antes e depois da Black Friday



```
exp_clientes['order_week'] = exp_clientes['order_purchase_timestamp'].dt.strftime('%Y-%U')
pre_bf = exp_clientes[exp_clientes['order_week'] < '2017-47']
post_bf = exp_clientes[exp_clientes['order_week'] >= '2017-47']

pre_bf_orders = pre_bf.groupby('order_week')['order_id'].count().reset_index()
post_bf_orders = post_bf.groupby('order_week')['order_id'].count().reset_index()

print("Pedidos feitos antes da Black Friday; Número de semanas no período: ")
print(pre_bf_orders['order_id'].sum(), pre_bf_orders['order_week'].nunique(), "\n")

print("Pedidos feitos depois da Black Friday; Número de semanas no período: ")
print(post_bf_orders['order_id'].sum(), post_bf_orders['order_week'].nunique(), "\n")
```

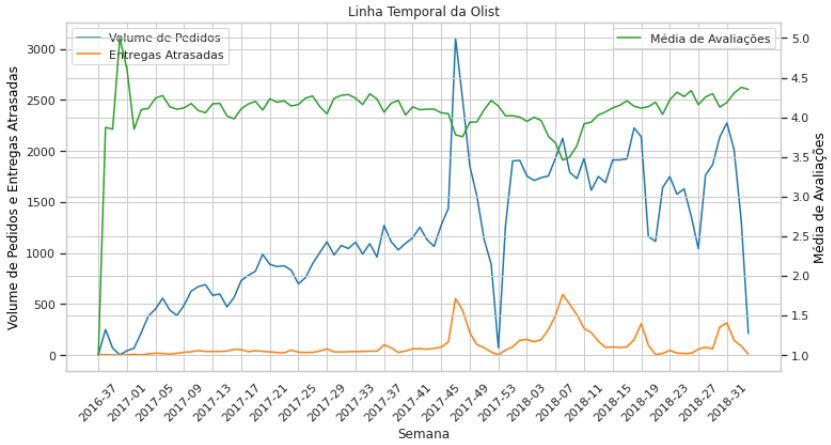
```
[75]  ✓ 0.6s

...  Pedidos feitos antes da Black Friday; Número de semanas no período:
     37832 50

     Pedidos feitos depois da Black Friday; Número de semanas no período:
     70658 42
```

Fonte: Elaborado pelo autor

Finalmente, conclui-se que a análise exploratória revelou diversos fatores de atenção que tendem a contribuir para a satisfação do cliente. Uma das correlações investigadas de alta intensidade é tempo de entrega do pedido. Ademais, foi observado que muitas das categorias mais populares também eram algumas das pior avaliadas. Entretanto, é provável que isso ocorra devido a itens que foram comprados com mais frequência, levando a mais chances de avaliações mais baixas aparecerem. Dessa forma, esclarecem-se as dúvidas levantadas nas alíneas c.1 e c.2 da lista de perguntas de negócio que direcionou o processo de análise.

Conforme mencionado anteriormente no Capítulo 2, mesmo que a satisfação do cliente não contribua diretamente para nosso sistema de recomendação, é possível obter um panorama geral do que a Olist oferece e o que os clientes estão comprando mais. Novamente, as categorias mais populares em termos de Volume de Vendas foram encontradas nos setores de Cama, Mesa e Banho, Decorações de Móveis, Lazer Esportivo e acessórios de computador. Tais itens serão algumas das principais conexões entre os clientes e como serão recomendados os produtos.

Figura 33 – Gráfico da Média das Avaliações dos Clientes, Total de Entregas Atrasadas e Volume de Pedidos por Semana de Compra



Fonte: Elaborado pelo autor

Por fim, esclarecem-se os principais desafios a serem enfrentados na modelagem estatística do Sistema de Recomendação. Um deles é a de recomendar produtos a um cliente em um contexto no qual mal foram consumidos por outros usuários. Para endereçar este problema, será essencial o papel da filtragem hibrída, que leva em consideração não somente o que outros usários compraram, mas também as características dos produtos adquiridos.

Ademais, a filtragem hibrída também irá lidar com um problema comum dos sistemas de recomendação, que é o problema de *Cold Start* que ocorre com uma base de clientes novos, que também já foi endereçada no Capítulo 2. No contexto da *Olist*, mais de 90% do total de clientes realizou apenas uma compra, o que portanto, será um fator determinante para a definição do modelo de Recomendação.

4.1.2 Pré-Processamento e Partição dos dados

O artefato de recomendação de produtos a ser desenvolvido deve ser capaz de recomendar produtos para os clientes com base em semelhanças entre outros usuários (componente colaborativa), nas características dos itens que tais usuários adquiriram e nas características dos usuários dos usuários da base (componente baseada em conteúdo). Portanto, o pré-processamento dos dados deve resultar em três matrizes esparsas, que serão aplicadas como *inputs* para os algoritimos de recomendação.

Apenas um destes conjuntos de dados irá passar pelo processo de Partição dos dados, que é aquele quepossui as interações de itens e usuários. Isso se deve ao fato de que os próprios valores faltantes dessa matriz deverão ser previstos pelo modelo. Conforme explicitado na subseção 2.2.3.2.1. As caraterísticas de itens e

usuários, da componente de conteúdo, serão acoplados ao problema de resolução por Fatoração Matricial, com o objetivo de se obterem previsões mais assertivas na matriz de interações.

4.1.2.1 Pré-Processamento dos dados

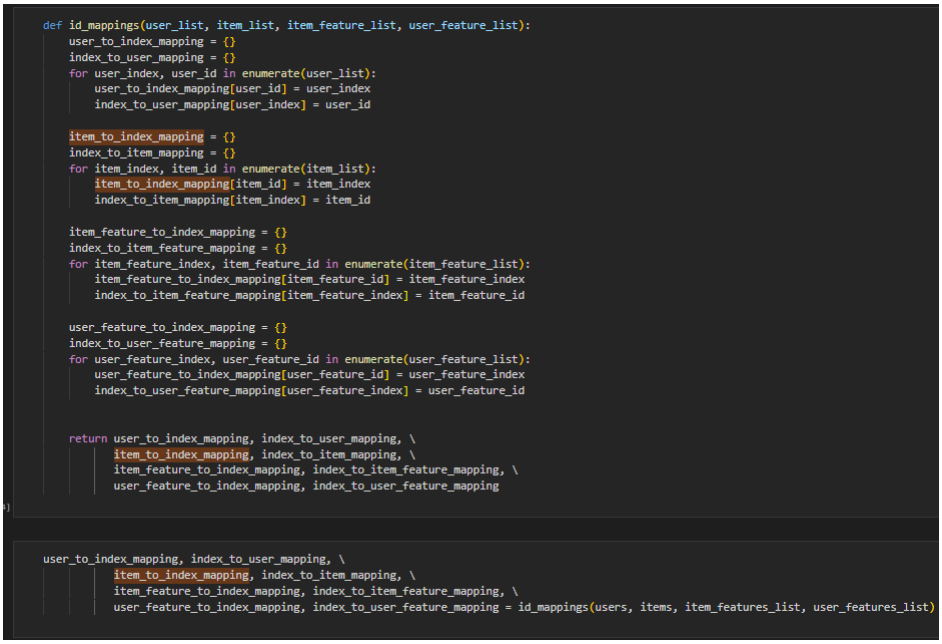
Um processamento comum para obtenção de tais matrizes esparsas vem da característica do pacote *LightFM* de não aceitação de *strings* (Caracteres) em seus *inputs* de modelo, portanto, o primeiro passo é realizar um *mapping* (Mapeamento) dos identificadores de produto e usuário, para que essas informações sejam convertidas em variáveis do tipo *integer*. Assim, chaves identificadoras são transformadas de uma estrutura de combinação de caracteres e números para apenas números. Essa atividade deve ser realizada para os conjuntos de itens, usuários e a matriz de interações entre itens e usuários e foi operacionalizada por meio de uma função *Python* conforme Figura 34.

A função *id_mapping()* tem como objetivo criar mapeamentos entre identificadores de usuários, itens e características associadas a itens e usuários. Para usuários, por exemplo, ela gera mapeamentos bidirecionais entre os identificadores originais e índices numéricos correspondentes, facilitando a referência eficiente durante o processo de modelagem. De maneira análoga, para itens e suas características, a função cria mapeamentos semelhantes, permitindo uma representação numérica eficaz no contexto do algoritmo de recomendação.

Logo, para cada conjunto mencionado anteriomente são criados 2 variáveis: Uma de Mapeamento de índice *x* usuário e outra no caminho oposto, do usuário *x* Mapeamento de índice, no exemplo das listas de usuários. Logo, as listas obtidas na Figura x serão utilizadas para substituir as atuais chaves identificadoras em caracteres, o que ainda permite identificar os produtos recomendados na fase de Avaliação do Artefato. Os argumentos passados a função são listas *Python* contendo os valores em caracteres de cada abstração a ser obtida (usuário, item, *feature* de usuário e *feature* de item).

Para se obter a matriz exparsa de interações dos usuários com os produtos, são selecionadas, a partir do mesmo *dataset* obtido na subseção 4.1.1.3, as colunas *customer_unique_id*, *product_id*, *product_count*, na qual a terceira representa a presença (valor 1) ou ausência (valor 0) de histórico de produtos comprados pelo cliente de determinado produto. Também é necessário acoplar o Mapeamento de Indíces de usuários (variável *user_to_index_mapping*) e itens (variável *item_to_index_mapping*) obtidos anteriormente.

Figura 34 – Função de Mapeamento de caracteres para valores inteiros

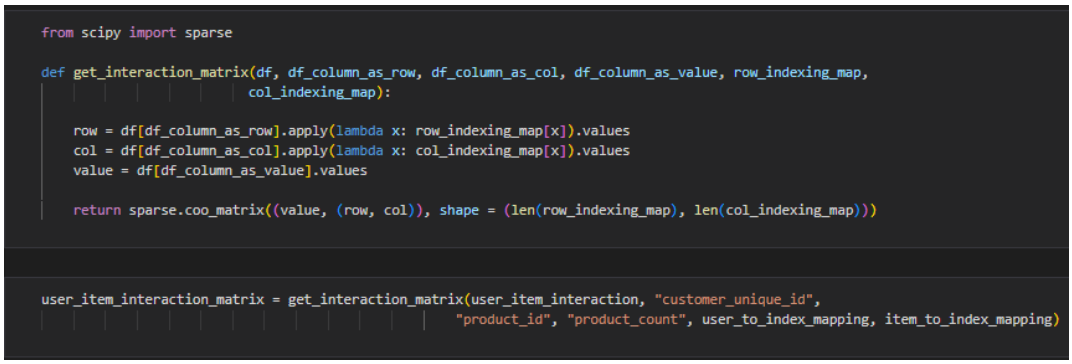


```
def id_mappings(user_list, item_list, item_feature_list, user_feature_list):
    user_to_index_mapping = {}
    index_to_user_mapping = {}
    for user_index, user_id in enumerate(user_list):
        user_to_index_mapping[user_id] = user_index
        index_to_user_mapping[user_index] = user_id

    item_to_index_mapping = {}
    index_to_item_mapping = {}
    for item_index, item_id in enumerate(item_list):
        item_to_index_mapping[item_id] = item_index
        index_to_item_mapping[item_index] = item_id

    item_feature_to_index_mapping = {}
    index_to_item_feature_mapping = {}
    for item_feature_index, item_feature_id in enumerate(item_feature_list):
        item_feature_to_index_mapping[item_feature_id] = item_feature_index
        index_to_item_feature_mapping[item_feature_index] = item_feature_id

    user_feature_to_index_mapping = {}
    index_to_user_feature_mapping = {}
    for user_feature_index, user_feature_id in enumerate(user_feature_list):
        user_feature_to_index_mapping[user_feature_id] = user_feature_index
        index_to_user_feature_mapping[user_feature_index] = user_feature_id


    return user_to_index_mapping, index_to_user_mapping, \
            item_to_index_mapping, index_to_item_mapping, \
            item_feature_to_index_mapping, index_to_item_feature_mapping, \
            user_feature_to_index_mapping, index_to_user_feature_mapping


user_to_index_mapping, index_to_user_mapping, \
        item_to_index_mapping, index_to_item_mapping, \
        item_feature_to_index_mapping, index_to_item_feature_mapping, \
        user_feature_to_index_mapping, index_to_user_feature_mapping = id_mappings(users, items, item_features_list, user_features_list)
```

Fonte: Elaborado pelo autor

A operacionalização destas operações para a obtenção da matriz esparsa de interações entre usuários e itens foi definida por meio de uma função *Python* que pode ser observada na Figura 35. Finalmente, com a aplicação desta, é obtida tal matriz que fica armazenada na variável *user_item_interaction_matrix*. Vale relembrar que estes dados representarão a contribuição colaborativa dos modelos de recomendação a serem construídos.

Figura 35 – Matriz esparsa de interações entre usuários - itens



```
from scipy import sparse

def get_interaction_matrix(df, df_column_as_row, df_column_as_col, df_column_as_value, row_indexing_map,
                           col_indexing_map):

    row = df[df_column_as_row].apply(lambda x: row_indexing_map[x]).values
    col = df[df_column_as_col].apply(lambda x: col_indexing_map[x]).values
    value = df[df_column_as_value].values

    return sparse.coo_matrix((value, (row, col)), shape = (len(row_indexing_map), len(col_indexing_map)))


user_item_interaction_matrix = get_interaction_matrix(user_item_interaction, "customer_unique_id",
                                                      "product_id", "product_count", user_to_index_mapping, item_to_index_mapping)
```

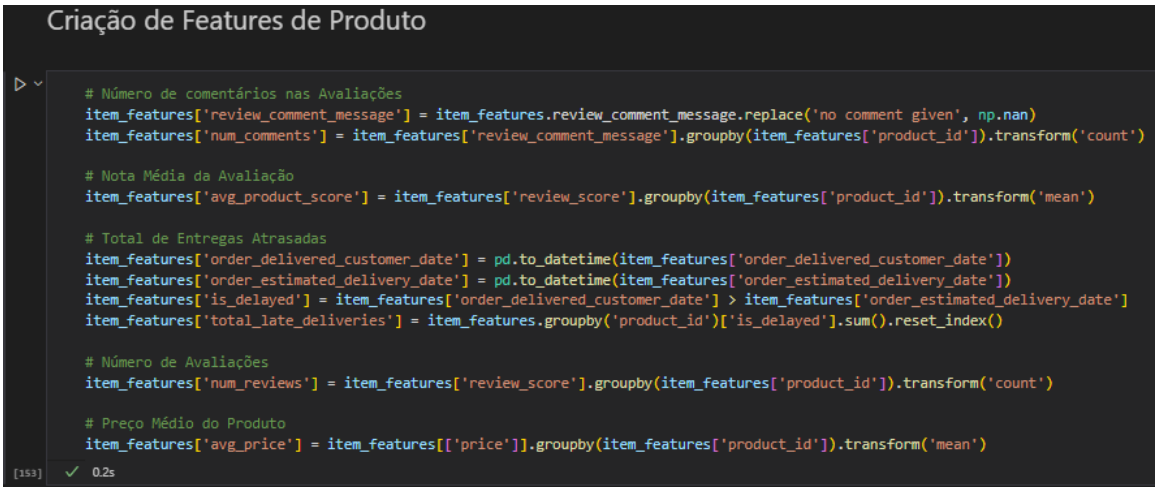
Fonte: Elaborado pelo autor

Já para os dados de entrada de *features* de produtos, advindos da componente baseada em contéudo, é necessário criar uma matriz produto ~ *feature* do produto, na qual os valores são prenchidos de maneira binária em relação à presença ou

à ausência de determinada *feature*. Para isso, são criadas novas *features* para os produtos, de acordo com resultados da Análise Exploratória.

A figura 36 ilustra a criação destas *features* de produto. Tais colunas, e outras que trazem informações relevantes para o modelo, como categoria do produto e revendedor, são agrupadas em um *dataframe* no qual sua chave primária equivale ao identificador do produto. Logo, cada linha deste conjunto de dados corresponde à uma métrica ou informação a respeito de um produto. A alínea a seguir elenca as *features* criadas e a respectiva motivação de escolha para cada uma delas.

Figura 36 – Features de Produto Criadas



Fonte: Elaborado pelo autor

a) *review_comment_message* (Mensagem de avaliação do cliente) - A inclusão das mensagens de *review* pode enriquecer o modelo, permitindo que ele compreenda nuances qualitativas e subjetivas do *feedback* dos clientes frente a determinado produto. O modelo pode aprender a associar certas palavras ou tópicos presentes nas mensagens de review com preferências dos usuários, proporcionando recomendações mais alinhadas com suas expectativas;
b) *num_comments* (Número de comentários da avaliação do produto) - A quantidade de comentários reflete a popularidade e a interação dos usuários com um produto. Produtos populares tendem a atrair mais atenção. Ao incorporar o número de comentários, o modelo pode dar mais peso a produtos com maior engajamento, melhorando a precisão das recomendações para produtos que estão sendo discutidos ou avaliados ativamente pelos clientes;
c) *avg_product_score* (Score médio das avaliações) - A avaliação média de um produto é um indicador chave da satisfação geral dos clientes e da qualidade percebida. Essa feature pode influenciar as preferências do modelo, priorizando produtos com pontuações mais altas, o que geralmente reflete uma melhor

aceitação pelos consumidores;

d) *is_delayed* (Binário indicando se o produto já foi atrasado alguma vez) - A pontualidade na entrega é crucial para a satisfação do cliente, conforme visto na subseção 4.1.1. Produtos que frequentemente sofrem atrasos podem impactar negativamente a experiência do usuário. Essa feature binária pode ser utilizada para destacar produtos que têm um histórico de entregas pontuais, melhorando a confiança do usuário nas recomendações;

e) *total_late_deliveries* (Número total de atrasos) - Além de indicar se um produto foi atrasado, o número total de atrasos oferece uma perspectiva mais granular do histórico de entrega. Pode ser utilizado para ajustar a relevância de produtos com base na frequência de atrasos, tornando as recomendações mais adaptadas às expectativas de entrega dos usuários;

f) *num_reviews* (Número de avaliações do produto) - Similar ao número de comentários, o número de avaliações reflete a interação dos usuários e pode indicar a confiabilidade das opiniões sobre um produto. Pode ser utilizado para priorizar produtos com um grande volume de avaliações, fornecendo recomendações mais robustas para itens que foram extensivamente avaliados;

g) *avg_price* (Preço médio do produto) - O preço médio do produto é um fator crítico nas decisões de compra dos consumidores e pode influenciar as preferências. Essa feature pode ser usada para ajustar as recomendações com base no perfil de preço do produto, garantindo que as sugestões estejam alinhadas com seu orçamento e preferências financeiras.

Os resultados da tabela agregada são pivotados para que em seguida seja aplicado o mesmo processo de *mapping* de valores, no qual cada *feature* seja identificada por um número inteiro. Por fim, uma matriz esparsa de *features* do produto é criada e está pronta para servir de entrada ao modelo de recomendação. A Figura 37 ilustra a tabela agregada de *features* produto criada.

Figura 37 – Tabela agregada de features por produto

```
# São atribuídas 8 features por produto
product_features.head(10)
```

✓ 0.0s

	product_id	feature	feature_count
0	87285b34884572647811a353c7ac498a	utilidades_domesticas	1
1	87285b34884572647811a353c7ac498a	3504c0cb71d7fa48d967e0e4c94d59d9	1
2	87285b34884572647811a353c7ac498a	4	1
3	87285b34884572647811a353c7ac498a	3	1
4	87285b34884572647811a353c7ac498a	4.0	1
5	87285b34884572647811a353c7ac498a	0	1
6	87285b34884572647811a353c7ac498a	29,99	1
7	87285b34884572647811a353c7ac498a	maua	1
8	87285b34884572647811a353c7ac498a	SP	1
9	6cc44821f36f3156c782da72dd634e47	cama_mesa_banho	1

Fonte: Elaborado pelo autor

Por fim, o processo descrito acima é analogamente aplicado com o objetivo de obter a matriz de *features* dos usuários. Conforme já denotado na Análise Exploratória, a base de clientes da *Olist* é composta majoritariamente por usuários com baixa frequência de pedidos. Portanto, as *features* selecionadas estão relacionadas apenas à cidade e estado de cadastro deste cliente, e se o mesmo teve o primeiro pedido antes ou depois da *Black Friday* de 2017. Este processo pode ser visto no código *Python* da Figura 38.

Figura 38 – Features de Usuário Criadas

```python
user_features['order_purchase_timestamp'] = pd.to_datetime(user_features['order_purchase_timestamp'])

# Calcule a data da primeira compra de cada cliente.
first_purchase_date = user_features.groupby('customer_unique_id')['order_purchase_timestamp'].min()

# Crie uma nova coluna 'primeira_compra_black_friday' com valores 0 ou 1.
black_friday_date = pd.to_datetime('2017-11-24')
first_purchase_date = first_purchase_date.reset_index()
first_purchase_date['first_purchase_black_friday'] = (first_purchase_date['order_purchase_timestamp'] >= black_friday_date).astype(int)


# Combine a nova coluna com o DataFrame original.
user_features = user_features.merge(first_purchase_date[['customer_unique_id', 'first_purchase_black_friday']], on='customer_unique_id', how='left')

user_features['total_orders'] = user_features['order_id']\
                                .groupby(user_features['customer_unique_id'])\
                                .transform('nunique')
user_features = user_features[['customer_unique_id', 'customer_state', 'total_orders', 'first_purchase_black_friday']]
```

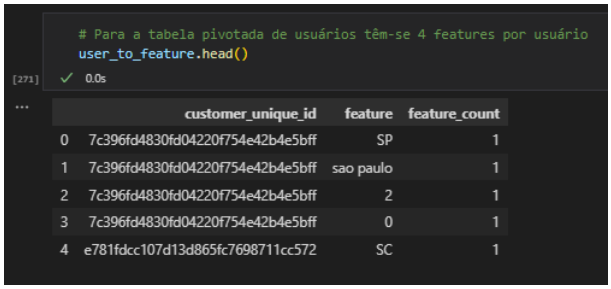
Fonte: Elaborado pelo autor

As alíneas a seguir descrevem com detalhes as *features* de usuários introduzidas ao modelo. Finalmente, na Figura 39, é possível encontrar a tabela agregada contendo a descrição das *features* e seus respectivos valores por usuário. Conforme já descritos, esses dados passam pelo processo de pivotação e *mapping* de colunas para que seja possível o seu *input* na abstração de modelo do pacote *LightFM.*

a) *customer_state* (Estado de residência do cliente) - A inclusão da *feature* permite ao modelo considerar a localização geográfica dos usuários, o que pode impactar as preferências de compra devido a variações regionais em comportamentos de consumo e disponibilidade de produtos. Isso pode levar a recomendações mais relevantes e adaptadas a padrões específicos de cada região;
b) *total_orders* (Total de pedidos realizados por cliente) - Tal *feature* oferece uma visão abrangente da atividade de compra de cada usuário. Integrar essa informação no modelo permite priorizar usuários com históricos mais extensos, resultando em recomendações mais personalizadas para consumidores frequentes. Essa abordagem reforça a fidelização e engajamento do usuário, promovendo uma experiência de compra mais atraente;

c) *first_purchase_black_friday* (Primeira compra na Black Friday) - A criação desta coluna surgiu da observação de que o atraso de pedidos, especialmente durante picos de vendas como a Black Friday, influencia a satisfação do cliente. Ao incluir essa *feature*, o modelo pode ajustar suas recomendações com base na experiência inicial do usuário, considerando se a primeira compra ocorreu em momentos de alta demanda e, portanto, se há maior probabilidade de o usuário ter enfrentado atrasos. Isso contribui para um modelo mais sensível à satisfação do cliente, adaptando-se a contextos específicos de compra e otimizando as sugestões para usuários que possam ter experimentado eventos atípicos em seus primeiros pedidos.

Figura 39 – Tabela agregada de features por usuário



```
# Para a tabela pivotada de usuários têm-se 4 features por usuário
user_to_feature.head()
```

[271] ✓ 0.0s

	customer_unique_id	feature	feature_count
0	7c396fd4830fd04220f754e42b4e5bff	SP	1
1	7c396fd4830fd04220f754e42b4e5bff	sao paulo	1
2	7c396fd4830fd04220f754e42b4e5bff	2	1
3	7c396fd4830fd04220f754e42b4e5bff	0	1
4	e781fdcc107d13d865fc7698711cc572	SC	1

Fonte: Elaborado pelo autor

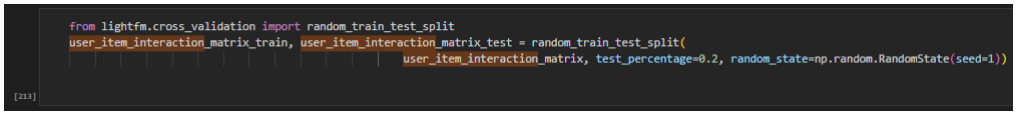
Portanto, ao final desta sequência de operações, são obtidos 3 conjuntos de dados, sendo que ambos os conjuntos referentes à *features* de produto e usuário já servirão como dados de entrada para a execução do modelo de recomendação. O conjunto de dados de interações entre usuários e itens passará por um procedimento, a ser detalhado na próxima subseção, para então estar adequado para o treinamento do modelo na subseção 4.1.3 e avaliação do modelo na subseção 4.2. Cada conjunto de dado obtido nesta subseção é definido nas alíneas a seguir.

a) *user_item_interaction_matrix* - Matriz esparsa contendo as interações históricas dos usuários frente aos itens da *Olist*;
b) *user_feature_interaction_matrix* - Matriz contendo a presença ou ausência de todas as *features* de usuário definidas no Pré-Processamento para cada identificador de usuário;
c) *product_to_feature_interaction_matrix* - Matriz contendo a presença ou ausência de todas as *features* de produto definidas no Pré-Processamento para cada identificador de produto.

4.1.2.2 Partição dos dados

A etapa de partição de dados, anterior à execução dos modelos de recomendação, é abordada com a aplicação da Técnica de Validação Cruzada, desempenhando um papel importante na na avaliação eficaz do desempenho do modelo. Neste estudo, optou-se por utilizar a função *random_train_test_split()* fornecida pelo pacote *LightFM*, configurada com uma taxa de teste de 20%. A aplicação da função pode ser observada na figura 40.

Figura 40 – Partição da matriz de interação de usuários e itens



Fonte: Elaborado pelo autor

Esse processo de particionamento dos dados garante a generalização do modelo, uma vez que permite avaliar o quão bem o modelo se comporta em dados não utilizados durante o treinamento. A escolha de uma porcentagem de teste de 20% é uma prática comum, equilibrando a necessidade de dados suficientes para treinamento com a necessidade de dados independentes para avaliação. Os conjuntos de dados de treino e teste são salvos nas variáveis *user_item_interaction_matrix_train* e *user_item_interaction_matrix_test*, respectivamente, a partir da entrada da matriz esparsa salva na variável *user_item_interaction_matrix*.

Conforme visto na subseção 2.2.4.1, a Técnica de Validação Cruzada contribui significativamente para mitigar problemas de *overfitting* (sobreajuste) ou *underfitting* (subajuste), promovendo uma avaliação mais representativa do desempenho do modelo. Ao dividir os dados em conjuntos de treinamento e teste, múltiplas vezes, a validação cruzada oferece uma visão mais abrangente da capacidade do modelo de generalizar padrões, reduzindo a dependência da aleatoriedade associada a uma única divisão de dados. Essa abordagem reforça a confiabilidade dos resultados obtidos, garantindo que o modelo seja avaliado em diferentes contextos e cenários, contribuindo para uma interpretação mais sólida dos resultados.

A escolha da taxa de 20% para o conjunto de teste foi ponderada considerando a necessidade de manter uma quantidade substancial de dados para o treinamento, enquanto ainda se assegura que uma parte significativa dos dados seja reservada para a avaliação. Essa prática equilibrada permite uma validação cruzada eficaz, proporcionando insights valiosos sobre o desempenho do modelo em dados não observados. Em suma, a aplicação da Técnica de Validação Cruzada nesta etapa de particionamento dos dados visa garantir a confiabilidade dos resultados obtidos durante a análise de sistemas de recomendação.

4.1.3 Execução dos modelos de Recomendação

Os modelos do pacote *LightFM* requerem interações entre usuário e produto, o que o torna capaz de lidar como um sistema de recomendação colaborativo básico, por meio do método de gradiente estocástico definido no Capítulo 2. Em seguida, será introduzida a matriz esparsa para informações do produto e novamente o modelo será testado. É esperado que isso aumente a complexidade do modelo, mas que melhore os resultados.

A última abordagem para comparação será adicionar uma segunda componente baseada em conteúdo, ao modelo interior, que se refere aos usuários. Devido ao fato do conjunto de dados ser composto principalmente por compradores de uma única vez, o desafio principal é fazer com que o sistema de recomendação sugira itens relevantes com base em um pequeno conjunto de dados. Nesse sentido, espera-se que as informações-chave sobre o produto melhorem os resultados.

As Figuras 41, 42 e 43 apresentam o código executado para a aplicação dos algoritmos de recomendação, sendo eles o modelo apenas colaborativo do pacote *LightFM* com previsão das avaliações por gradiente estocástico, modelo híbrido com agregação de componente de produto apenas, e finalmente modelo híbrido com componente de produto e usuário, respectivamente.

Figura 41 – Modelo de Filtragem Colaborativa por Gradiente Estocástico

```
# Definição da Função de Perda como WARP
only_colab_model = LightFM(loss = "warp")

# Treinamento do Modelo Utilizando apenas Filtragem Colaborativa
start = time.time()

only_colab_model.fit(user_item_interaction_matrix_train,
          item_features=None,
          sample_weight=None,
          epochs=1,
          num_threads=4,
          verbose=False)

end = time.time()
print("Tempo de Execução do Modelo = {0:.{1}f} ".format(end - start, 2))

###
# Pré-avaliação do modelo pelo cálculo da pontuação na curva ROC
start = time.time()

auc_only_colab_model = auc_score(model = only_colab_model,
                    test_interactions = user_item_interaction_matrix_test,
                    num_threads = 4, check_intersections = False)

end = time.time()

####

print("Tempo de Execução do Cálculo = {0:.{1}f} ".format(end - start, 2))
print("Pontuação na ROC = {0:.{1}f}".format(auc_only_colab_model.mean(), 2))
[69]  ✓ 14.9s
...  Tempo de Execução do Modelo = 0.18
     Tempo de Execução do Cálculo = 14.75
     Pontuação na ROC = 0.68
```

Fonte: Elaborado pelo autor

Através do método *.fit(),* pertencente à classe da representação latente do modelo de recomendação, são passados os parâmetros de treinamento. Para o primeiro caso (*LightFM* - Filtragem Colaborativa), apenas a matriz de interação usuários e itens de treinamento é fornecida. Ademais são passados parâmetros adicionais ao treinamento, como o número de *epochs* e *threads* a serem utilizados.

Em um segundo modelo (*LightFM* - Híbrido (Produto)), a mesma lógica acima de código é aplicada, porém neste momento, a matriz esparsa de *features* do produto é passada como argumento. Desta forma, são introduzidas tais características dos produtos como fatores latentes no problema de Fatoração Matricial. A aplicação do método para este modelo pode ser visto na Figura x, e o resultado dos parâmetros obtidos com o treinamento são salvos na variável *hybrid_model*.

Figura 42 – Modelo de Filtragem Híbrida com apenas componente de produto

```
# Definição da Função de Perda como WARP
hybrid_model = LightFM(loss = "warp")

# Treinamento do Modelo Utilizando Filtragem Colaborativa e Filtragem Baseada em Conteúdo
start = time.time()

hybrid_model.fit(user_item_interaction_matrix_train,
                 item_features=product_to_feature_interaction_matrix,
                 sample_weight=None,
                 epochs=1,
                 num_threads=4,
                 verbose=False)

end = time.time()
print("Tempo de Execução do Treinamento do Modelo = {0:.{1}f} ".format(end - start, 2))

####

# Pré-avaliação do modelo pelo cálculo da pontuação na curva ROC
start = time.time()

auc_hybrid_model = auc_score(model = hybrid_model,
                             test_interactions = user_item_interaction_matrix_test,
                             train_interactions = user_item_interaction_matrix_train,
                             item_features = product_to_feature_interaction_matrix,
                             num_threads = 4, check_intersections=False)
end = time.time()

####

print("Tempo de Execução do Cálculo = {0:.{1}f} ".format(end - start, 2))
print("Pontuação na ROC = {0:.{1}f}".format(auc_hybrid_model.mean(), 2))
[70]  ✓ 33.4s

...  Tempo de Execução do Treinamento do Modelo = 0.31
     Tempo de Execução do Cálculo = 33.16
     Pontuação na ROC = 0.79
```

Fonte: Elaborado pelo autor

Finalmente, o terceiro modelo (*LightFM* - Híbrido (Produto + Cliente)) recebe a matriz esparsa de usuários e itens de treinamento, as matriz esparsas de *features* dos produtos e também a matriz esparsa de *features* dos clientes. Assim, são introduzidas tais características dos produtos e usuários como fatores latentes no problema de Fatoração Matricial. A aplicação do método para este modelo pode ser visto na Figura x, e o resultado dos parâmetros obtidos com o treinamento são salvos na variável *hybrid_users_products*.

Adicionalmente, para cada construção de modelo, é realizada uma prévia avaliação dos resultados, ao aplicar-se a função *auc_score()* disponível no pacote *LightFM*, e obter o seu valor médio por meio do método *.mean()*. Em um primeiro momento, é possível entedender que o modelo híbrido com *features* de clientes e produtos obteve o melhor resultado, entretanto, tal conclusão só pode ser retirada após uma análise mais aprofundada das métricas de *precision* e *recall,* a qual será realizada na próxima seção.

Figura 43 – Modelo de Filtragem Híbrida componente de produto e usuário

```
# Definição da Função de Perda como WARP
hybrid_users_products = LightFM(loss = "warp")

# Treinamento do Modelo Utilizando Filtragem Colaborativa e Filtragem Baseada em Conteúdo
start = time.time()

hybrid_users_products.fit(user_item_interaction_matrix_train,
          user_features=user_feature_interaction_matrix,
          item_features=product_to_feature_interaction_matrix,
          sample_weight=None,
          epochs=1,
          num_threads=4,
          verbose=False)

end = time.time()
print("Tempo de Execução do Treinamento do Modelo = {0:.{1}f} ".format(end - start, 2))

####

# Pré-avaliação do modelo pelo cálculo da pontuação na curva ROC
start = time.time()
auc_hybrid_users_products = auc_score(model = hybrid_users_products,
                    test_interactions = user_item_interaction_matrix_test,
                    train_interactions = user_item_interaction_matrix_train,
                    user_features=user_feature_interaction_matrix,
                    item_features = product_to_feature_interaction,
                    num_threads = 4, check_intersections=False)
end = time.time()

####

print("Tempo de Execução do Cálculo = {0:.{1}f} ".format(end - start, 2))
print("Pontuação na ROC = {0:.{1}f}".format(auc_hybrid_users_products.mean(), 2))

✓ 40.4s

Tempo de Execução do Treinamento do Modelo = 0.34
Tempo de Execução do Cálculo = 40.13
Pontuação na ROC = 0.80
```

Fonte: Elaborado pelo autor

4.2 AVALIAÇÃO DO ARTEFATO

De acordo com os conceitos definidos no Capítulo 2 sobre a *Design Science Research*, a avaliação do artefato se dará ao validar experimentalmente o impacto da solução, a qual objetiva aumentar a personalização das recomendações de produtos à clientes. Assim, por meio de simulações, o artefato será executado em dados de teste, de forma a reproduzir artificialmente os resultados em um ambiente *offline.*

Os procedimentos técnicos da avaliação da instanciação de Sistema de Recomendação se dará de acordo com o definido na seção 2.2.4, Avaliação de Sistemas de Recomendação. Logo, simulações de geração de listas de produtos

recomendados aos clientes serão avaliadas por meio das métricas *precision@k*, *recall@k* e média da *ROC AUC*.

Adicionalmente, será realizada uma análise comparativa dos dois modelos mais performáticos identificados na fase anterior, a partir de uma seleção de clientes com históricos de compra variados. Essa abordagem teve como objetivo avaliar o desempenho de cada modelo em diferentes cenários, uma vez que para cada cliente, serão geradas recomendações de ambos os modelos, as quais serão comparadas com os verdadeiro-positivos identificados.

4.2.1 Avaliação por Método de Ranqueamento

Com o objetivo de se obter dados de performance dos modelos em listas de recomendação, de acordo com o tamanhos das mesmas, é criada a função em *Python* da Figura 44. Assim, para cada modelo, para cada tamanho da lista, são calculadas as três métricas utilizando as funções *precision_at_k()*, *recall_at_k()* e *auc_score()*, disponíveis no pacote *LightFM*, conforme estipulado na subseção 3 e baseado nos conhecimentos vistos no Capítulo 2.

Figura 44 – Função para obtenção das métricas de cada modelo treinado

```
# Função para calcular métricas para diferentes modelos e tamanhos de lista (k)
def calculate_metrics(model, train, test, user_features=None, item_features=None, k_values=[5, 10, 20]):
    metrics = {'Modelo': [], 'K': [], 'Precision': [], 'Recall': [], 'AUC': []}
    for k in k_values:
        precision = precision_at_k(model, test, train_interactions=train
                                   , k=k
                                   , user_features=user_features
                                   , item_features=item_features
                                   , check_intersections=False
                                  ).mean()
        recall = recall_at_k(model, test, train_interactions=train
                                    , k=k, user_features=user_feature
                                    , item_features=item_features
                                    , check_intersections=False
                                   ).mean()
        auc = auc_score(model, test, train_interactions=train
                        , user_features=user_features
                        , item_features=item_features
                        , check_intersections=False
                        )
        auc = auc.mean()
        metrics['Modelo'].append(model)
        metrics['K'].append(k)
        metrics['Precision'].append(precision)
        metrics['Recall'].append(recall)
        metrics['AUC'].append(auc)
    return pd.DataFrame(metrics)
✓ 0.0s
```

Fonte: Elaborado pelo autor

Ainda, as comparações das métricas de precisão e recall serão feitas nos *cutoffs* de 5, 10 e 20. Os *cutoffs* descrevem onde interromper a lista de produtos classificados previstos. Tais valores são definidos pois este é um intervalo de recomendação frequentemente utilizados por plataformas de *E-commerce*, conforme revisitado no Capítulo 2. A função definida é então chamada e os seus resultados são

guardados em um *dataframe*, o qual pode ser observado na Tabela 9. Para efeitos de visualização, os valores de *precision* e *recall* são plotados nas Figuras 45 e 46, respectivamente.

Tabela 9 – Métricas de Desempenho dos Modelos

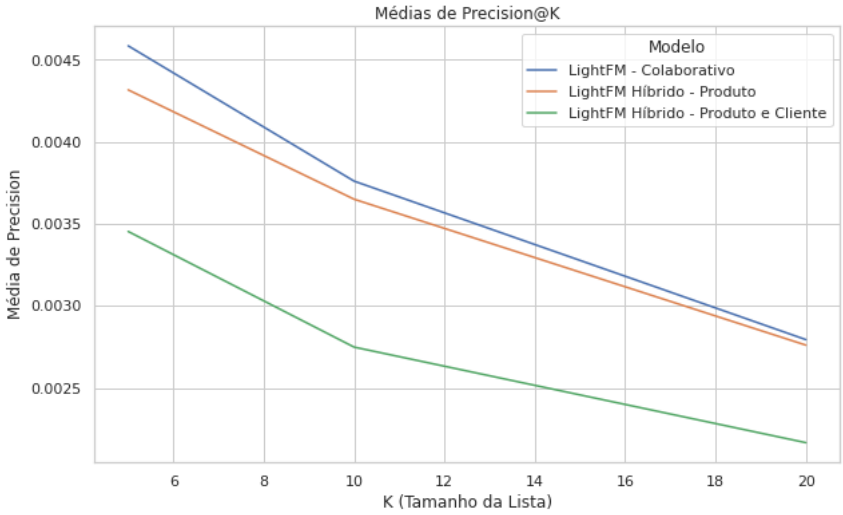
Modelo	*K*	*Precision*	*Recall*	AUC
LightFM - Colaborativo	5	0.004586	0.021672	0.679444
	10	0.003761	0.035855	
	20	0.002794	0.053558	
LightFM Híbrido - Produto	5	0.004317	0.020098	0.788643
	10	0.003650	0.034427	
	20	0.002760	0.052457	
LightFM Híbrido - Produto e Cliente	5	0.003454	0.016836	0.797165
	10	0.002748	0.026829	
	20	0.002166	0.042382	

De acordo com os conceitos vistos no Capítulo 2 do trabalho, foi entendido que a precisão das recomendações pode impactar diretamente as decisões de compra dos usuários. Os resultados mostram que o modelo *LightFM* híbrido, que combina filtragem colaborativa e baseada em conteúdo, superou os modelos puramente colaborativos em termos da métricas *AUC*, o que indica uma melhor capacidade de classificar as preferências dos usuários.

No entanto, os valores de *precision* e *recall* foram relativamente mais baixos, sugerindo que, embora o modelo híbrido possa classificar bem, ele pode não ser tão eficaz em sugerir produtos relevantes. Portanto, a inclusão de recursos de usuários no modelo *LightFM* híbrido não mostrou-se totalmente benéfica, em comparação com o modelo puramente colaborativo. Mesmo que por uma diferença de poucos centésimos, é possível entender que não foi possível coletar *features* de usuário que de fato foram relevantes para o aprendizado do modelo.

Por outro lado, o modelo *LightFM* com componente apenas colaborativa possui *precision* e *recall* mais altos, mesmo nas listas maiores. Isso indica que as recomendações são mais precisas e abrangentes, cobrindo uma parte maior dos itens relevantes. No entanto, o *AUC* desse modelo é ligeiramente mais baixo em comparação com ambos os modelos com *features* de usuários e produtos.

Figura 45 – Gráfico das médias de Precision por Tamanho k das Listas



Fonte: Elaborado pelo autor

Essa diferença nos resultados destaca a importância de escolher um modelo com base nas necessidades e objetivos específicos do sistema de recomendação. Se a ênfase estiver na precisão das recomendações, o modelo de filtragem apenas colaborativa pode ser a escolha preferida, uma vez que fornece valores mais altos de *precision* e *recall*. No entanto, se obter poder de classificação de todo o conjunto de itens for mais importante, o modelo hibrido pode ser a opção mais adequada devido ao seu AUC mais alto.

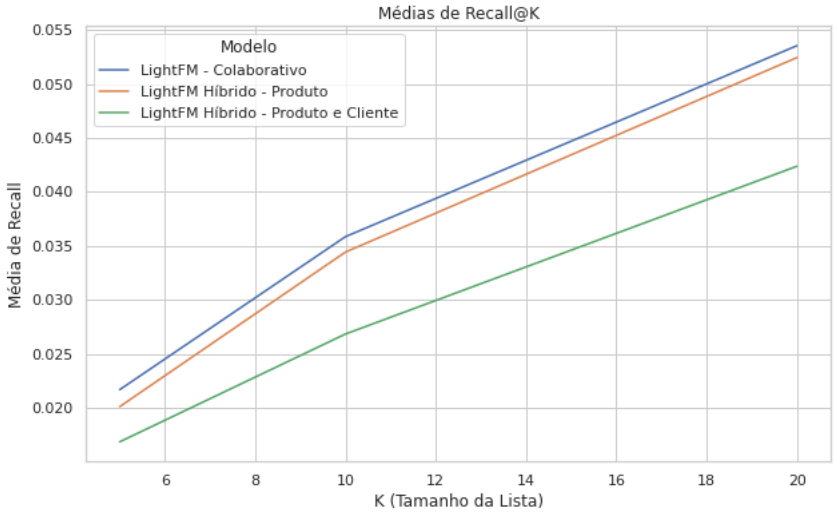
Em última análise, a escolha entre os modelos deve ser orientada pelas metas e requisitos do sistema de recomendação, considerando a importância relativa da *precision*, *recall* e *AUC* no contexto específico do comércio eletrônico e as necessidades dos usuários. Tal fato destaca a importância de considerar características adicionais ao criar Sistemas de Recomendação, especialmente em ambientes onde os produtos e as preferências dos usuários podem variar amplamente.

Finalmente, levando em consideração os pontos levantados nessam etapa avaliação do artefato, os modelos a serem escolhidos para a próxima etapa será o *LightFM* Híbrido - Produto e o *LightFM* Colaborativo. O primeiro pois este consegue equilibrar as características vistas nos modelos LightFM Híbrido - Produto e Cliente e LightFM - Colaborativo, cumprindo então com objetivos do artefato e objetivos gerais de aplicação de Sistemas de Recomendação. Já o segundo modelo, obteve um desempenho ligeiramente inferior ao primeiro modelo escolhido.

É possível entender que tal modelo possui métricas de precisão muito próximas às obtidas pelo modelo *LightFM* - Colaborativo, porém sua pontuação *ROC AUC* é quase tão elevada quanto a observada no modelo *LightFM Híbrido* - Produto e Cliente.

Assim, pode se dizer que o modelo *LightFM* Híbrido - Produto é um bom classificador e relativamente preciso dentre as listas gerados, dado o contexto da empresa *Olist* que possui uma cartela de clientes extremamente diversa.

Figura 46 – Gráfico das médias de Recall por Tamanho k das Listas



Fonte: Elaborado pelo autor

4.2.2 Avaliação por comparativo de recomendações a clientes selecionados

A análise comparativa dos dois modelos, *LightFM* Híbrido - Produto e o *LightFM* Colaborativo, envolve primeiramente a seleção de dois clientes com históricos de compra variados, um com certa frequência de pedidos e outro usuário com apenas um pedido realizado. Com isso, são previstas recomendações de produtos as quais são comparadas com valores de interações conhecidas dos usuários. Essa abordagem teve como objetivo avaliar o desempenho de cada modelo em diferentes cenários, identificar padrões consistentes e avaliar a confiabilidade global dos modelos. Para gerar

Para o usuário *c8ed31310fc440a3f8031b177f9842c3*, que adquiriu 10 produtos, como o produto *4a5c3967bfd3629fe07ef4d0cc8c3818*, sendo todos estes da categoria de Construção, os resultados do *LightFM* Colaborativo não eram de certa forma os esperados. Enquanto os positivos conhecidos incluem ferramentas de construção, as recomendações apresentam itens da mesma categoria em sua maioria, a qual não pertence ao conjunto de categorias de produtos mais vendidos. Esse resultado inesperado de recomendar categorias, que mesmo já adquiridas pelo cliente anteriormente, não possuem alto valor de volume de vendas, sugere que o LightFM Colaborativo pode estar detectando padrões de clientes com compras semelhantes que também adquiriram esses itens diversos.

Por outro lado, para o mesmo usuário, *LightFM* Híbrido - Produto forneceu recomendações como Videogames, Bolsas e Acessórios e arte. Essas sugestões não parecem estar alinhadas com o comportamento inicial de compra de ferramentas de construção. Notavelmente, as recomendações deste modelo exibem uma tendência a oferecer uma variedade de itens dentro de outras categorias, indicando um potencial para ajudar os clientes a descobrir produtos relacionados às suas preferências.

a) Modelo *LightFM* - Colaborativo

1. *18fa9cc25ea8b54f32d029f261673c0f* - Ferramentas para Construção;
2. *97d94ffa4936cbc2555e83aefc1f427b* - Ferramentas para Construção;
3. *cd46a885543f0e169a49f1eb25c04e43* - Acessórios de Informática;
4. *556702ebb73d3786a852ad3d5a8ad268*- Ferramentas para Construção;
5. *43724b27731595d954e911f443fb1cc4*- Ferramentas para Construção.

b) Modelo *LightFM* - Híbrido (Produto);

1. *0aabfb375647d9738ad0f7b4ea3653b1* - Videogames;
2. *d017a2151d543a9885604dc62a3d9dcc* - Bolsas e Acessórios;
3. *4fe644d766c7566dbc46fb851363cb3b* - Artes;
4. *3519403062e217f433e0bbdc52e0b19f* - Videogames;
5. *18fa9cc25ea8b54f32d029f261673c0f* - Ferramentas para Construção.

Já para o usuário *698e1cf81d01a3d389d96145f7fa6df8*, que comprou apenas item o *9571759451b1d780ee7c15012ea109d4*, relacionado a automóveis, o modelo *LightFM* Colaborativo recomendou produtos em diversas categorias, como Cama, Mesa & Banho, Acessórios de Esporte e Telefonia, sendo estas posicionadas entre as 5 categorias de produtos mais vendidos da loja. Em contraste, o *LightFM* Híbrido - Produto sugeriu itens adicionais relacionados a automóveis, mantendo uma consistência temática com os positivos conhecidos do usuário.

Os itens recomendados pelo *LightFM* Híbrido - Produto para o usuário mencionado anteriormente, destacam sua inclinação para propor itens dentro da mesma categoria, contribuindo para uma experiência do usuário mais relevante, principalmente quando se detêm poucas informações históricas dos usuários.As alíneas a seguir apresentam os resultados das recomendações, bem como suas respectivas categorias, para o usuário *698e1cf81d01a3d389d96145f7fa6df8*.

a) Modelo *LightFM* - Colaborativo

1. *b59fb744c6f3cd1dc23b10f760848d98* - Cama, Mesa & Banho;
2. *dca18a6e2fb6da75092ed874094ed7b6* - Acessórios de Esporte;
3. *cd46a885543f0e169a49f1eb25c04e43* - Telefonia;
4. *c78b767da00efb70c1bcccab87c28cd5* - Acessórios de Informática;

5. *b8dac5113b06a97e64943234522572b9* - Móveis Decoração.

b) Modelo *LightFM* - Híbrido (Produto)

1. *9ddc4249779322828f89d2a9c04f7ee1* - Automóveis;
2. *629e019a6f298a83aeecc7877964f935* - Automóveis;
3. *a659cb33082b851fb87a33af8f0fff29* - Automóveis;
4. *334479e94cba98064050db1c9636e244* - Automóveis;
5. *b8dac5113b06a97e64943234522572b9* - Automóveis.

Resumidamente, o *LightFM* - Colaborativo tende a sugerir itens populares, potencialmente influenciado por tendências mais amplas no conjunto de dados. Enquanto isso, o *LightFM* Híbrido - Produto demonstra uma propensão a recomendar itens intimamente alinhados com as compras anteriores dos usuários, alternando uma abordagem mais personalizada quando existem informações históricas dos usuários.

Em contrapartida, o aspecto negativo do modelo *LightFM* - Colaborativo reside em sua limitação de sugerir produtos mais vendidos dentro da loja como um todo, oferecendo aos usuários os mesmos produtos o qual este já esta acostumado a visualizar ao adentrar no *website*. Isso porquê plataformas de comércio eletrônico já possuem uma prática de ofertar os produtos mais vendidos ou em promoção na página geral do endereço eletrônico.

Dessa forma, nesta etapa de avaliação dos modelos de recomendação é possível que afirmar que a aplicação do modelo *LightFM Híbrido - Produto* se mostra a mais adequada, principalmente quando se leva em consideração a composição da carteira de clientes *Olist*, que possui muitos clientes cadastrados com baixíssima taxa de recompra.

4.3 CONSIDERAÇÕES FINAIS DO CAPÍTULO

O presente capítulo executou metodologicamente o desenvolvimento do artefato de pesquisa, seguindo procedimentos técnicos de Mineração de Dados e Sistemas de Recomendação baseados nas diretrizes de Shmueli (2017) e Aggarwal (2016). Ao combinar conhecimentos adquiridos no Capítulo 2, foram estabelecidas etapas técnicas específicas, centradas na Análise Exploratória dos Dados, Pré-processamento e Partição dos Dados, e na Execução dos Modelos de Recomendação. Cada procedimento foi definido nas subseções correspondentes, promovendo uma estrutura c para o desenvolvimento do artefato.

A Análise Exploratória dos Dados, como primeiro passo, revelou-se importante para compreender a fundo o conjunto de dados. A aplicação de quatro procedimentos sequenciais, desde a seleção de dados relevantes até a descrição estatística e visualização, proporcionou uma compreensão holística das características do

conjunto de dados. Essa abordagem robusta estabeleceu a base necessária para o desenvolvimento posterior dos modelos de recomendação, assegurando uma análise informada e significativa.

O segundo bloco de desenvolvimento do artefato, o Pré-processamento e Partição dos Dados, destacou a importância de preparar adequadamente os dados para a aplicação dos modelos de recomendação, especialmente ao utilizar a biblioteca *LightFM*. A obtenção de matrizes esparsas de interação e a criação de conjuntos de treino e teste foram passos que garantiram a posterior avaliação do modelo em dados não vistos durante o treinamento. Além disso, o pré-processamento envolveu a obtenção de matrizes de *features* de produtos e usuários, permitindo o enriquecimento da representação do modelo.

A terceira etapa, Execução dos Modelos de Recomendação, destacou a versatilidade e eficiência da biblioteca *LightFM* em implementar diferentes estratégias de recomendação. Os modelos de Filtragem Colaborativa e Híbridos foram configurados de maneira a incrementalmente integrar características específicas de produtos e clientes. A descrição detalhada desses modelos proporcionou uma compreensão clara de como cada abordagem considera diferentes aspectos do comportamento do usuário e características dos itens.

Finalmente, a seção de Avaliação do Artefato executou a estratégia de avaliação que permitiu a escolha do melhor modelo de maneira quantitativa e qualitativa. A combinação de avaliações por Método de Ranqueamento e métricas de acurácia, seguindo os princípios de Cremonesi, Koren e Turin (2010), possibilitou medir o desempenho e a eficácia dos modelos.

A inclusão de uma avaliação comparativa entre recomendações a clientes selecionados proporcionou aprendizados sobre a adequação dos modelos a diferentes perfis de usuários e padrões de compra. Essa abordagem estruturada e abrangente atende aos requisitos de *Design Science Research*, garantindo uma avaliação criteriosa e orientada aos resultados do artefato de pesquisa desenvolvido.

Com base nas análises dos três modelos de recomendação - *LightFM* Colaborativo, *LightFM* Híbrido - (Produto) e *LightFM* Híbrido - (Produto + Cliente) - os resultados indicam que o modelo *LightFM* Híbrido - (Produto) apresentou um desempenho satisfatório nas métricas de Método de Ranqueamento, destacando-se em proporcionar recomendações relevantes. Além disso, a Avaliação Comparativa revelou uma característica distintiva desse modelo: a propensão para recomendar produtos com categorias já conhecidas pelo cliente, especialmente quando o histórico de compras é limitado. Essa característica é especialmente relevante para a *Olist*, onde muitos clientes possuem poucas informações de histórico, tornando o modelo *LightFM* Híbrido (Produto) uma escolha adequada para a instanciação proposta nesta pesquisa de *DSR*.

Ao considerar o modelo LightFM - Colaborativo, observou-se que ele tende a recomendar os produtos mais vendidos da loja para clientes com históricos de compras reduzidos. Dado o grande número de clientes com essa característica na Olist, o *LightFM* -Colaborativo não foi escolhido para compor a instanciação desenvolvida neste trabalho de DSR. Em vez disso, a preferência recai sobre o *LightFM* Híbrido - (Produto), que se destaca por sua capacidade de oferecer recomendações mesmo para clientes com poucas informações de histórico, contribuindo para uma experiência de recomendação mais eficaz e alinhada às características da plataforma de comércio eletrônico da *Olist*.

5 CONCLUSÕES E TRABALHOS FUTUROS

5.1 CONCLUSÕES

No decorrer deste trabalho, buscou-se solucionar um problema crítico na área de comércio eletrônico: a falta de personalização nas recomendações de produtos aos clientes da *Olist*. No primeiro capítulo, foi definido o cenário e os desafios enfrentados pela empresa em questão, uma plataforma de comércio eletrônico que deseja proporcionar uma experiência de compra mais personalizada aos seus clientes.

Neste mesmo capítulo, foram definidos os objetivos específicos e o objetivo geral do trabalho, destacando a necessidade de implementar o sistema de recomendação de produtos. Neste ponto, também contextualizou-se a importância da pesquisa no contexto mais amplo de comércio eletrônico.

O segundo capítulo foi dedicado à fundamentação teórica, no qual foi explorado três tópicos essenciais que fornecem a base para a pesquisa. Primeiramente, formalizou-se a adaptação das etapas da mineração de dados no campo de estudo da *Business Analytics*, seguindo o modelo *SEMMA*.

Em seguida, revisaram-se os conceitos de Sistemas de Recomendação, abordando suas classificações quanto à coleta de dados, estratégias, técnicas e métodos de avaliação. Dentre as métricas de avaliação, destacaram-se as métricas de avaliação por Método de Ranqueamento e a pontuação *ROC AUC*. Por fim, apresentou-se a *Design Science Research*, que formaliza as etapas do trabalho e exige a geração de um artefato concreto.

O terceiro capítulo estabeleceu a metodologia adotada. Enquadrando a pesquisa como *Design Science Research*, as etapas de *Data Mining* e o *workflow* de desenvolvimento de Sistemas de Recomendação foram adaptados e sobrepostos, de forma a estruturar os desenvolvimentos do Capítulo 4. Além disso, foram definidas as ferramentas utilizadas, como *Python*, o pacote *Seaborn* e *LightFM*, e também fornecida uma descrição detalhada dos dados coletados da *Olist*.

No capítulo anterior à conclusão, o trabalho foi divido em duas seções principais. A primeira, Desenvolvimento do Artefato, compreendeu as etapas de Análise Exploratória dos Dados, Pré-processamento e Partição dos Dados e Aplicação de Algoritmos de Recomendação. A segunda, Avaliação do Artefato, buscou simular recomendações ao analisar as métricas previamente discutidas no Capítulo 2 bem como realizar uma análise comparativa de recomendações para um mesmo usuário.

Nesta última subseção, foi possível identificar que o modelo que melhor atendeu às necessidades da *Olist* foi o *LightFM* Híbrido - Produto, uma vez que demonstrou um relativo equilíbrio entre as métricas de avaliação discutidas. Adicionalmente, também demonstrou a tendência de recomendar produtos similiares aos já consumidos pelo usuário, ao prever recomendações a partir de poucas informações históricas de

usuário, o que se mostra extremamente adequado no contexto de conjunto de dados da *Olist*. Este modelo de recomendação foi capaz de oferecer recomendações relativamente personalizadas aos clientes da plataforma, atendendo aos objetivos estabelecidos.

Mais especificamente, ao longo deste trabalho foi possível compreender que a aplicação de modelos de recomendação desempenha um papel importante em uma variedade de contextos, desde o comércio eletrônico até a personalização de conteúdo, melhorando a experiência do usuário e aumentando a eficácia das recomendações.

Neste estudo, foram explorados três modelos do pacote *LightFM* para avaliar seu desempenho em cenários diferentes. Os resultados obtidos revelam esclarecimentos importantes que podem orientar a implementação prática desses modelos. Tais resultados possuem implicações significativas para a pesquisa em Sistemas de Recomendação e para a implementação prática em contextos do mundo real.

Isso porquê entende-se que foi demonstrado a importância de se considerar abordagens híbridas que combinam filtragem colaborativa e baseada em conteúdo, bem como a inclusão de recursos adicionais, para melhorar o desempenho da recomendação.

Além disso, foi ressaltada a necessidade de se avaliar o desempenho de modelos sob várias métricas para obter uma compreensão mais abrangente de seu comportamento. Essas confirmações se mostraram relevantes na etapa de Avaliação de artefatos de pesquisa da *Design Science Research,* pois orientam o refinamento contínuo dos artefatos.

Concluí-se, portanto, que o presente trabalho representa um avanço significativo no campo de Sistemas de Recomendação, especialmente no contexto de comércio eletrônico. Através da combinação de técnicas de mineração de dados, Sistemas de Recomendação e metodologia *Design Science Research*, foi possível desenvolver um artefato com resultados satisfatórios que proporciona uma experiência de compra mais personalizada aos clientes da *Olist*.

No entanto, faz-se necessário reconhecer que o campo de Sistemas de Recomendação é dinâmico e sujeito a constante evolução. Portanto, são sugeridos estudos futuros que busquem aprimorar ainda mais o desempenho do sistema de recomendação, explorando novos algoritmos e técnicas de avaliação. A personalização contínua é essencial para atender às crescentes demandas e expectativas dos consumidores no mercado de comércio eletrônico.

Por fim, este trabalho demonstra a importância de abordagens interdisciplinares na busca por soluções inovadoras e relevantes. Ao unir conhecimentos de mineração de dados, sistemas de recomendação e metodologia *Design Science Research*,

foi possível contribuir para a melhoria das experiências de compra *online* e, consequentemente, para o sucesso da *Olist* no mercado.

5.2 TRABALHOS FUTUROS

Com o objetivo de se obter uma melhoria contínua do Sistema de Recomendação da *Olist*, algumas áreas de pesquisa e desenvolvimento podem ser exploradas. Uma delas diz respeito à criação de novas *features* de produtos e usuários por meio de outras técnicas avançadas de mineração de dados, como clusterização e classificação, as quais podem enriquecer o conjunto de dados e aprimorar a capacidade do sistema de recomendação em identificar padrões de preferência. Essa abordagem poderia envolver a extração de informações não triviais dos dados disponíveis, aumentando a qualidade das recomendações.

A melhoria contínua do modelo de recomendação é fundamental. Isso implica a exploração de algoritmos mais avançados, como aprendizado profundo e redes neurais, bem como a otimização dos modelos existentes. A pesquisa pode se concentrar em desenvolver modelos que sejam mais eficazes na previsão das preferências dos clientes, resultando em recomendações mais precisas.

Ainda, a implantação do sistema em um ambiente de produção em tempo real é um passo importante para a empresa. Isso permitiria que as previsões fossem entregues aos usuários de forma imediata, melhorando a experiência de compra e a eficácia das recomendações. A implantação desse processo envolve desafios técnicos de computação em nuvem, como escalabilidade e baixa latência, que requerem um estudo dedicado a esta frente de trabalho.

A consideração da diversidade nas recomendações é relevante para atender a diferentes perfis de clientes. Pesquisas futuras podem se concentrar em estratégias para aumentar a diversidade das recomendações, sem comprometer a relevância. Isso pode envolver a criação de algoritmos que equilibrem a diversidade e a personalização.

Além das áreas de pesquisa mencionadas anteriormente, o ajuste de hiperparâmetros dos modelos de recomendação também é um aspecto relevante a ser explorado. O chamado *tuning* de hiperparâmetros envolve a otimização de configurações como, a taxa de aprendizado, número de fatores latentes, pesos de regularização e outros parâmetros específicos dos algoritmos.

A precisão na seleção e ajuste dos hiperparâmetros pode ter um impacto significativo no desempenho dos modelos, tornando-os mais eficazes na geração de recomendações personalizadas. Portanto, uma pesquisa futura nessa área pode contribuir para o refinamento e aprimoramento contínuo do Sistema de Recomendação da *Olist*.

Por fim, a avaliação do desempenho do sistema deve ser realizada de maneira

contínua. Pesquisas futuras podem se concentrar no desenvolvimento de métricas personalizadas que reflitam os objetivos e as necessidades específicas da *Olist*. Tal estudo poderia garantir que o sistema esteja alinhado com os indicadores de sucesso da empresa e permitiria ajustes contínuos para aprimorar o desempenho do sistema.

REFERÊNCIAS

01 AGGARWAL, C. C. *Recommender systems*. [S.l.]: Springer, 2016. v.1

ARI, N.; USTAZHANOV, M. 2014 11th International Conference on Electronics, Computer and Computation (ICECCO). set. 2014, Abuja, Nigeria: IEEE, set. 2014. p. 1–6. Disponível em: <http://ieeexplore.ieee.org/document/6997585/>.

ASAPANNA, R. *Evaluating Recommendation Systems — Part 2*. [S.l: s.n.]., 2019

CATES, J. How to Design and Build a Recommendation System Pipeline in Python (Jill Cates). 2019. Disponível em: <https://www.youtube.com/watch?v=v_mONWiFv0k&t=188s&ab_channel=PyConCanada>.

CÉSAR CAZELLA, S.; S. N. NUNES,. M. A.; BERNI REATEGUI, E. André Ponce de Leon F. de Carvalho. *A Ciência da Opinião: Estado da arte em Sistemas de Recomendação*, 2010. Disponível em: <https://almanaquesdacomputacao.com.br/gutanunes/publications/JAI4.pdf>.

CREMONESI, P.; KOREN, Y.; TURRIN, R. Proceedings of the 2010 ACM Conference on Recommender Systems. 26 out. 2010, [S.l: s.n.], 26 out. 2010. Disponível em: <https://dl.acm.org/doi/10.1145/1864708.1864721>.

DO DESENVOLVIMENTO, INDÚSTRIA, COMÉRCIO E SERVIÇOS, M. Dashboard do Comércio Eletrônico Nacional. 2023.

DRESCH, A.; PACHECO LACERDA, D.; CAUCHICK MIGUEL, P. A. A Distinctive Analysis of Case Study, Action Research and Design Science Research. *Review of Business Management*, p. 1116–1133, 24 nov. 2015. Disponível em: <http: //rbgn.fecap.br/RBGN/article/view/2069>.

DUARTE, T. M. G. M. *Implementação de um Sistema de Business Intelligence*. 2018. tese de doutorado – 2018.

FORTUNATO, L. Segmentação de clientes de um e-commerce brasileiro utilizando RFV e métodos de clusterização particionais. 2022.

GIL, A. Como Elaborar Projetos de Pesquisa. SÃO PAULO: ATLAS S.A, 1991. Disponível em: <https://wwwp.fc.unesp.br/Home/helber-freitas/tcci/gil_como_elaborar_projetos_de_pesquisa_-anto.pdf>.

HEVNER *et al.* Design Science in Information Systems Research. *MIS Quarterly*, v. 28, n. 1, p. 75, 2004. Disponível em: <https://www.jstor.org/stable/10. 2307/25148625>.

HOLSAPPLE, C.; LEE-POST, A.; PAKATH, R. A unified foundation for business analytics. *Decision Support Systems*, v. 64, p. 130–141, ago. 2014. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0167923614001730>.

HUG, N. Surprise: A Python library for recommender systems. *Journal of Open Source Software*, v. 5, n. 52, p. 2174, 5 ago. 2020. Disponível em: <https://joss.theoj.org/papers/10.21105/joss.02174>.

IYENGAR, S. S.; LEPPER, M. R. When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, v. 79, n. 6, p. 995–1006, dez. 2000. Disponível em: <http://doi.apa.org/getdoi.cfm?doi=10. 1037/0022-3514.79.6.995>.

KAPADIA, S. *Recommendation System in Python: LightFM*. [S.l: s.n.]., 2020 KULA, M. *Explicit vs implicit feedback models*. [S.l: s.n.]. Disponível em: <https://github.com/maciejkula/explicit-vs-implicit>., 2018

KULA, M. Metadata Embeddings for User and Item Cold-start Recommendations. Publisher: arXiv Version Number: 1, 2015. Disponível em: <https://arxiv.org/ abs/1507.08439>.

LACERDA, D. P. *et al.* Design Science Research: método de pesquisa para a engenharia de produção. *Gestão & Produção*, v. 20, n. 4, p. 741–761, 26 nov. 2013. LEE, D. Matrix Factorization via stochastic gradient descent. 2018. Disponível em:<https://www.youtube.com/watch?v=qlyTyNREROE&ab_channel=DanielLee>.

LEE, P. M. Use Of Data Mining In Business Analytics To Support Business Competitiveness. *Review of Business Information Systems (RBIS)*, v. 17, n. 2, p. 53–58, 7 maio de 2013. Disponível em: <https://clutejournals.com/index.php/RBIS/article/ view/7843>.

LÚCIA DA SILVA EDNA; MUSZKAT MENEZES. *Metodologia da Pesquisa e Elaboração de Dissertação*. [S.l.]: Universidade Federal de Santa Catarina - UFSC, 2005. (, 4).

MANSON, N. Operations Research Society of South Africa. *Is operations research really research?*, 2006. Disponível em: <https://www.ajol.info/index.php/orion/article/view/34262>.

MARCH; STOREY. Design Science in the Information Systems Discipline: An Introduction to the Special Issue on Design Science Research. *MIS Quarterly*, v. 32, n. 4, p. 725, 2008. Disponível em: <https://www.jstor.org/stable/10.2307/25148869>.

MCKINNEY, W. pandas: powerful Python data analysis toolkit. 12 nov. 2012.

Disponível em:
<https://pandas.pydata.org/pandas-docs/version/0.7.3/pandas.pdf>.

MIRANDA BORGES, D.; LUIZ DE OLIVEIRA, F. Congresso de Computação e Tecnologias da Informação. 2010, Palmas-TO: [s.n.], 2010.

OLGA, A. Q. DE M.; MONTEIRO, G. L. MLOps - Transformando Teoria em Prática. 2021. Disponível em:
<https://repositorio.insper.edu.br/handle/11224/3723>.

OLIST; SIONEK, A. *Brazilian E-Commerce Public Dataset by Olist*. [S.l: s.n.]., [S.d.]

PEFFERS, K. *et al.* A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, v. 24, n. 3, p. 45–77, dez. 2007. Disponível em:
<https://www.tandfonline.com/doi/full/10.2753/MIS0742-1222240302>.

RESNICK, P.; VARIAN, H. R. Recommender systems. *Communications of the ACM*, Publisher: ACM New York, NY, USA, v. 40, n. 3, p. 56–58, 1997.

RICCI, F.; ROKACH, L.; SHAPIRA, B. Introduction to Recommender Systems Handbook. In: RICCI, F. *et al.* (Org.). Boston, MA: Springer US, 2011. p. 1–35. Disponível em:
<https://link.springer.com/10.1007/978-0-387-85820-3_1>.

SALES, D. Sistema de recomendação baseado em filtragem colaborativa em uma base de dados de feedback implícito. 2019. Disponível em:
<https://www.cin. ufpe.br/~tg/2019-1/TG_CC/tg_dso.pdf>.

SARDANA, D. Divya Sardana | Building Recommender Systems Using Python. 2016. Disponível em:
<https://www.youtube.com/watch?v=39vJRxIPSxw&ab_channel= PyData>.

SERRANO, L. How does Netflix recommend movies? Matrix Factorization. 2018. Disponível em:
<https://www.youtube.com/watch?v=ZspR5PZemcs&ab_channel=Serrano.Academy>.

SHMUELI, G. *et al. Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. [S.l.]: Wiley, 2017. Disponível em:
<https://books.google.com. bo/books?id=ETwuDwAAQBAJ>.

SIMON, H. A. *The sciences of the artificial*. 3rd ed ed. Cambridge, Mass.: MIT Press, 1996.

SOLARTE, J. *A Proposed Data Mining Methodology and its Application to Industrial Engineering*. 2002. tese de doutorado – 2002. Disponível em:
<https://trace. tennessee.edu/utk_gradthes/2172>.

STECK, H. KDD ’10: The 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 25 jul. 2010, Washington DC USA: ACM, 25 jul. 2010. p. 713–722. Disponível em: <https://dl.acm.org/doi/10.1145/1835804.1835895>.

STEVENSON, A. (Org.). *Oxford dictionary of English*. 3rd ed ed. New York, NY: Oxford University Press, 2010.

TURBAN, E.; VOLONINO, L. *Tecnologia da Informação para Gestão-: Em Busca de um Melhor Desempenho Estratégico e Operacional*. [S.l.]: Bookman Editora, 2013.

VAISHNAVI, V.; KUECHLER, B. Association for information systems. *DESIGN SCIENCE RESEARCH IN INFORMATION SYSTEMS*, 2021. Disponível em:<http://www.desrist.org/design-research-in-information-systems/>.

VENSON, E. Um modelo de sistema de recomendação baseado em filtragem colaborativa e correlação de itens para personalização no comércio eletrônico. Publisher: Florianópolis, SC, 2002.

ZENONE, L. C. *CRM - Customer Relationship Management: Gestão do Relacionamento com o Cliente e a Competitividade Empresarial*. [S.l.]: Novatec Editora, 2007. Disponível em: <https://books.google.com.bo/books?id=I0DdXsZ3U0wC>.