UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS DE RUSSAS

CURSO DE GRADUAÇÃO EM ENGENHARIA DE SOFTWARE

BRENO OLIVEIRA MAURÍCIO LIMA

USO DE TÉCNICAS DE MODELAGEM PREDITIVA PARA PREVER O COMPORTAMENTO DE COMPRAS EM E-COMMERCES

RUSSAS

2026

BRENO OLIVEIRA MAURÍCIO LIMA

USO DE TÉCNICAS DE MODELAGEM PREDITIVA PARA PREVER O COMPORTAMENTO DE COMPRAS EM E-COMMERCES

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Software do Campus de Russas da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de Software.

Orientadora: Profa. Drª. Tatiane Fernandes Figueiredo

RUSSAS

2026

BRENO OLIVEIRA MAURÍCIO LIMA

USO DE TÉCNICAS DE MODELAGEM PREDITIVA PARA PREVER O COMPORTAMENTO DE COMPRAS EM E-COMMERCES

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Software do Campus de Russas da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de Software.

Aprovada em: 29/01/2026

BANCA EXAMINADORA

Profa. Drª. Tatiane Fernandes Figueiredo (Orientadora)
Universidade Federal do Ceará (UFC)

Prof. Dr. Prof. Dr. Eurinardo Rodrigues Costa
Universidade Federal do Ceará (UFC)

Prof. Dr.Pablo Luiz Braga Soares
Universidade Federal do Ceará (UFC)

AGRADECIMENTOS

Primeiramente agradeço a mim mesmo, por nunca ter desistido, depois de tantas e tantas coisas que eu passei e que só eu sei, mantive a força, coragem e perseverança para concluir essa minha graduação, sendo uma pessoa vindo lá de um pequeno arquipélago na costa ocidental africana, saindo de casa sozinho e tão novo e com pouca experiência de vida. Mas durante todo esse tempo aprendi muito e sou grato pela pessoa que me tornei e ainda continuo me transformando.

Em segundo lugar, mas não menos importante, agradeço aos meus pilares, meus pais, Alícia Maria dos Santos Oliveira e José Calazans Maurício Lima, pela confiança, por acreditarem em mim e por me ter dado essa oportunidade para estudar fora do meu país onde que, por enquanto, são poucas as pessoas que tem possibilidade de sair e realizar os seus sonhos mundo a fora. Agradeço também a minha amada irmã Zelcia dos Santos Oliveira Lima à todos os meus familiares e amigos.

De igual modo, deixo a minha gratidão à minha professora e orientadora Dra Tatiane Fernandes Figueiredo, uma excelente profissional e ainda melhor um ótimo ser humano, por ter aceitado me orientar e me guiar durante essa trajetória acadêmica nessa pesquisa.

Allahu Akbar

()

RESUMO

Com o avanço do comércio eletrônico, as empresas passaram a enfrentar desafios cada vez mais relevantes relacionados à retenção de clientes, sobretudo devido à facilidade com que consumidores podem migrar entre fornecedores. A rotatividade de clientes, ou churn, configurou-se como um fator crítico para o desempenho financeiro, uma vez que implicou perda potencial de receita e aumento dos custos associados à aquisição de novos consumidores. Nesse contexto, a aplicação de técnicas de aprendizado de máquina apresentou-se como uma abordagem promissora para prever o comportamento dos usuários e gerar informações estratégicas capazes de subsidiar ações de mitigação do churn. Com esse propósito, o presente trabalho avaliou modelos preditivos destinados a estimar a probabilidade de um produto visualizado em um ambiente de comércio eletrônico ser efetivamente adquirido pelo cliente. Os experimentos foram realizados utilizando uma base de dados pública disponibilizada na plataforma Kaggle. O estudo abrangeu etapas de pré-processamento, engenharia de atributos, treinamento e validação dos modelos. Os resultados obtidos foram comparados com aqueles reportados na literatura, visando contribuir para o aprimoramento de soluções preditivas aplicadas ao setor de comércio eletrônico.

Palavras-chave: comércio eletrônico; rotatividade de clientes; *churn*; aprendizado de máquina

ABSTRACT

With the rapid expansion of electronic commerce, companies have faced increasingly significant challenges related to customer retention, particularly due to the ease with which consumers can switch between providers. Customer turnover, or churn, has emerged as a critical factor affecting financial performance, as it entails potential revenue loss and higher costs associated with acquiring new customers. In this context, the application of machine learning techniques proved to be a promising approach for predicting user behavior and generating strategic insights to support churn-mitigation actions. With this purpose, the present study evaluated predictive models designed to estimate the probability that a product viewed in an e-commerce environment would be effectively purchased by the customer. The experiments were conducted using a public dataset available on the Kaggle platform. The study encompassed preprocessing, feature engineering, model training, and validation stages. The results obtained were compared with those reported in the literature, aiming to contribute to the improvement of predictive solutions applied to the electronic commerce sector.

Keywords: e-commerce; customers churn; churn prediction; machine learning

SUMÁRIO

1 INTRODUÇÃO

Nos últimos anos, o comércio tem se consolidado de forma predominante no ambiente digital, e os consumidores passaram a demonstrar preferência pela aquisição de produtos e serviços por meios digitais, em detrimento das lojas físicas (Farooqi *et al.* (2022)) . De acordo com Albertin (2010), um e-commerce corresponde a uma modalidade de comércio realizada eletronicamente que, devido ao amplo alcance proporcionado pela internet, permite a conexão entre varejistas em escala global. Estruturado sobre tecnologias de comunicação e informação, esse modelo busca atender a diversos objetivos de negócio, sendo caracterizado por seu fácil acesso e por apresentar custos operacionais relativamente reduzidos.

Um dos grandes problemas enfrentados pelas empresas de e-commerce é a rotatividade de clientes ou simplesmente rotatividade, do inglês churn. A rotatividade de clientes pode ser considerada uma oportunidade perdida de lucro. Os custos para conquistar novos clientes costumam ser de cinco a até seis vezes maiores do que os custos para reter um cliente existente. Como resultado, os esforços dos especialistas para manter a participação de mercado deixaram de se concentrar na aquisição de novos clientes e passaram a se concentrar na retenção dos clientes existentes. Por esse motivo, a rotatividade de clientes, também conhecida como atrito ou desvio de clientes, é uma grande preocupação para diversos setores. Isso é particularmente importante no contexto do comércio eletrônico, onde os consumidores podem comparar produtos ou serviços e trocar de fornecedor com o mínimo de esforço (Pondel *et al.* ()).

Diante desse problema, foco deste estudo, o uso de técnicas de modelagem preditiva para prever o comportamento dos clientes é de suma importância para diminuir essa rotatividade que é umas das principais dificuldades enfrentadas pelas grandes empresas do comércio eletrônico. De acordo com (Shobana *et al.* (2023)) as empresas de comercio eletrônico, especialmente as do segmento B2C (Business-to-Consumer) estão envolvidas em uma competição acirrada pela sobrevivência, tentando obter acesso às bases de clientes de seus rivais e, ao mesmo tempo, evitar que os clientes atuais desertem. Também segundo os autores, o custo de aquisição de novos clientes está aumentando à medida que mais concorrentes entram no mercado com gastos iniciais e estratégias de penetração de ponta, tornando a retenção de clientes essencial para essas organizações.

Shobana *et al.* (2023) também abordaram a problemática da rotatividade de clientes em ambientes de e-commerce, empregando uma base de dados privada composta por informações como histórico de pesquisas, compras realizadas, frequência de aquisição, avaliações e feedbacks

dos consumidores. Para a tarefa de predição, os autores utilizaram dois modelos de aprendizado de máquina, *Support Vector Machine* (SVM) e uma Rede Neural Artificial (RNA), alcançando acurácias de 77,36% e 82,64%, respectivamente. Já (Farooqi *et al.* (2022)) discutiram possíveis melhorias para um sistema de recomendação em uma plataforma de e-commerce. Para isso, utilizaram técnicas como filtragem colaborativa, filtragem baseada em conteúdo, sistema híbrido e pré-processamento com *Bag of Words*. Entre as abordagens avaliadas, o sistema híbrido apresentou o melhor desempenho, alcançando precisão de 42,06%.

Os autores (Falqueto e Cezar (2022)) empregaram técnicas de clusterização, utilizando o algoritmo K-Means, para aprimorar a compreensão do comportamento dos clientes e, consequentemente, otimizar as estratégias voltadas à atração e retenção do público-alvo de cada organização. Como resultado, foram identificados quatro *clusters* distintos. De forma semelhante ao estudo de (Farooqi *et al.* (2022)),(Loukili *et al.* (2023)) discutiram o desenvolvimento de um sistema de recomendação baseado em técnicas de mineração de dados, como o algoritmo FP-Growth, obtendo aproximadamente 4.970 regras de associação e uma probabilidade média de 69,3%, indicando um relevante potencial de conversão nas recomendações geradas.

Nesse contexto, o presente trabalho propôs a construção e a avaliação de modelos preditivos baseados em algoritmos de *machine learning*, com o objetivo de estimar se um cliente, após visualizar um produto em uma plataforma de e-commerce, efetivamente realizará a compra. A pesquisa utilizou uma base de dados pública disponibilizada na plataforma *Kaggle* e contemplou as etapas de exploração e preparação dos dados, engenharia de atributos, treinamento e validação dos modelos, com ênfase na comparação de desempenho entre diferentes abordagens preditivas. O propósito foi contribuir para o desenvolvimento de soluções capazes de auxiliar empresas a compreenderem de forma mais precisa o comportamento de seus usuários e, a partir disso, implementarem estratégias mais eficazes de retenção de clientes.

A organização deste trabalho é apresentada da seguinte forma. Neste capítulo são descritos a problemática e os objetivos, divididos em geral e específicos. **O Capítulo 2** expõe os principais conceitos que fundamentam a pesquisa. O **Capítulo 3** apresenta os estudos relacionados que contribuíram para a definição da problemática investigada. O **Capítulo 4** detalha a metodologia empregada, o desenvolvimento da pesquisa e a comparação dos resultados obtidos com aqueles reportados na literatura. Por fim, o **Capítulo 5** apresenta a conclusão do estudo, uma análise geral dos modelos implementados e possíveis direções para trabalhos futuros.

1.1 Objetivos

1.1.1 Objetivo geral

Criar e avaliar modelos preditivos para determinar se um produto de um *e-commerce* será comprado ou não após sua visualização pelo cliente.

1.1.2 Objetivos específicos

- Exploração a base de dados existente na literatura e disponibilizada na plataforma *Kaggle* relacionada ao problema em estudo.
- Limpar, padronizar e criar novas carateristicas relevantes para resolução da problemática discutida neste trabalho;
- Treinar, validar e testar ao menos 2 modelos preditivos baseados em algoritmos de aprendizado de máquina;
- Comparar os resultados obtidos por cada um dos modelos.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão abordados os conceitos fundamentais para compreensão da problemática até a solução proposta. No tópico 2.1 há uma breve explicação das áreas de conhecimento relacionadas.

2.1 Aprendizado de Máquina

O Aprendizado de Máquina (AM), ou *Machine Learning* (ML),é um subcampo da *Artificial Intelligence* que estuda métodos computacionais capazes de identificar padrões nos dados e realizar previsões ou classificações sem depender de instruções explicitamente programadas. A essência do ML está na capacidade de generalização: um modelo deve ser capaz de aprender a partir de exemplos e aplicar esse conhecimento a novas situações (Paixão *et al.* (2022)).

A relevância do ML aumentou substancialmente com o crescimento do volume de dados digitais, especialmente em plataformas de e-commerce, onde ele é aplicado em diferentes contextos, como detecção de fraude, previsão de demanda, recomendação personalizada, análise de comportamento e otimização da experiência do usuário. Neste trabalho, a técnica é aplicada ao problema de prever se a visualização de um produto resultará ou não em uma compra, caracterizando um típico caso de classificação binária.

O ML pode ser dividido em três grandes abordagens. A primeira delas, o aprendizado supervisionado, utiliza conjuntos de dados rotulados, nos quais cada exemplo possui uma classe associada. O objetivo é aprender uma função que relacione as variáveis explicativas aos rótulos, de forma a generalizar para novos casos. Esta é a abordagem adotada no presente estudo.

A segunda abordagem é o aprendizado não supervisionado, no qual não há rótulos associados às observações. O objetivo é identificar estruturas latentes nos dados, como agrupamentos ou padrões de similaridade — recurso frequentemente usado em segmentação de clientes. Por fim, o aprendizado por reforço, ou *Reinforcement Learning*, consiste em treinar um agente que interage com um ambiente dinâmico e aprende a realizar ações sequenciais com base em recompensas e penalidades recebidas (Ris-Ala (2023)). Embora amplamente utilizado em diversos contextos, este último não é empregado na metodologia deste estudo.

2.1.1 Regressão Logística

A Regressão Logística é um dos métodos mais consolidados para resolver problemas de classificação binária. Seu objetivo é modelar a probabilidade de ocorrência de um evento categórico em função de um conjunto de variáveis independentes. Ao contrário da regressão linear, que pode produzir valores fora do intervalo [0,1], a Regressão Logística utiliza uma transformação sigmoide que garante que a saída seja interpretável como probabilidade (Hosmer *et al.* (2013)). A probabilidade de ocorrência do evento (Y = 1) pode ser representada por:

$$P(Y=1) = \frac{1}{1+e^{-g(x)}}$$

onde a função linear g(x) é dada por:

$$g(x) = B_0 + B_1X_1 + \cdots + B_pX_p$$

Essa transformação produz uma curva em formato de "S", que comprime o valor linear g(x) dentro do intervalo probabilístico desejado. A razão de chances, ou *odds*, é definida por:

$$odds(Y=1) = \frac{P(Y=1)}{1-P(Y=1)}$$

O logaritmo dessa razão, conhecido como *logit*, pode ser escrito como:

$$\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = B_0 + B_1X_1 + \cdots + B_pX_p$$

Os coeficientes B_i são estimados por uma técnica, denominada em inglês por *Maximum Likelihood Estimation* (MLE), que busca os valores mais prováveis de gerar os dados observados. O termo e^{Bi} permite interpretar diretamente o impacto de cada variável sobre a razão de chances, um aspecto especialmente útil para entender quais características aumentam ou reduzem a probabilidade de compra. Devido à sua interpretabilidade, eficiência e estabilidade, a Regressão Logística é frequentemente utilizada para análise de comportamento em e-commerce. Neste trabalho, ela permite identificar fatores que influenciam a decisão de compra de produtos visualizados.

Além dessas propriedades básicas, é importante destacar algumas características adicionais para compreender completamente o funcionamento do algoritmo. Primeiramente, a escolha da função sigmoide não é arbitrária: trata-se de uma função diferenciável e monotônica,

o que facilita o processo de otimização durante o treinamento. Como consequência, pequenas alterações em g(x) produzem mudanças graduais na probabilidade estimada, o que torna o modelo estável e menos sujeito a flutuações abruptas. Outro ponto relevante é o papel das variáveis independentes $X_1, \ldots, X_p$. Cada uma delas contribui linearmente para o valor de g(x), antes da transformação sigmoide. Isso significa que a Regressão Logística pressupõe que a relação entre as variáveis explicativas e o *logit* da probabilidade seja linear. Caso existam relações não lineares, essas podem ser incorporadas criando novas variáveis, como interações ou termos quadráticos, tornando o modelo mais expressivo sem perder sua interpretabilidade.

A própria razão de chances fornece uma interpretação intuitiva do modelo: enquanto probabilidades limitam-se ao intervalo [0,1], as *odds* variam de 0 a ∞, permitindo quantificar o quão mais provável é um evento ocorrer em relação a não ocorrer. O logaritmo dessa razão garante simetria, possibilitando valores negativos (quando o evento é improvável), positivos (quando é provável) ou zero (quando ambas as chances são iguais).

O processo de treinamento por MLE também merece esclarecimento adicional. Em vez de minimizar distâncias, como ocorre na regressão linear com mínimos quadrados, o algoritmo maximiza a verossimilhança de observar os dados reais. Isso é realizado por meio de métodos numéricos como *Gradient Descent* ou *Newton-Raphson*. Na prática, o algoritmo ajusta os coeficientes de forma iterativa até que o modelo encontre uma combinação de valores que melhor represente o relacionamento entre variáveis e rótulos.

Por fim, vale ressaltar que, apesar de sua natureza simples, a Regressão Logística possui desempenho competitivo e é amplamente utilizada em aplicações do mundo real. Em contextos de e-commerce, ela permite identificar padrões relevantes, como tempo de navegação,número de visualizações, tipo de produto ou características de navegação do cliente, que podem estar diretamente associados à probabilidade de conversão. Sua interpretabilidade é um diferencial importante quando é necessário explicar o comportamento do modelo ou justificar decisões baseadas em suas predições.

2.1.2 Random Forest

A técnica *Random Forest* (RF) é um método de *ensemble learning* baseado na combinação de múltiplas árvores de decisão independentes. Cada árvore é construída a partir de subconjuntos aleatórios do conjunto de dados e de subconjuntos aleatórios de atributos, criando diversidade entre os modelos individuais e reduzindo o risco de sobreajuste (Breiman (2001)).

O processo de votação entre as árvores resulta em um modelo robusto, com precisão elevada e baixa variância. O objetivo da RF é aproximar uma função *f(X)* que minimize o erro esperado:

$$E_{XY}[L(Y, f(X))]$$

Em problemas de regressão, aplica-se tipicamente a função de perda quadrática:

$$L(Y, f(X)) = (Y - f(X))^2$$

Já em tarefas de classificação utiliza-se o erro *zero-one*:

$$L(Y, f(X)) = I(Y \neq f(X)) = \begin{cases} 0, & \text{se } Y = f(X) \\ 1, & \text{se } Y \neq f(X) \end{cases}$$

Cada árvore construída pela RF pode ser representada por:

$$h_j(X, \Theta_j)$$

onde Θ_j contém as escolhas aleatórias feitas durante sua construção. A predição final é obtida por votação majoritária:

$$f(x) = \arg\max_{y \in \mathscr{Y}} \sum_{j=1}^{J} I\big(y = h_j(x)\big)$$

Esse mecanismo de agregação confere ao método propriedades desejáveis, como resistência a ruído, capacidade de lidar com dados de alta dimensionalidade e boa performance mesmo sem ajustes complexos de hiperparâmetros (Cutler *et al.* (2012)).

No contexto deste trabalho, a *Random Forest* é especialmente útil por capturar relações não lineares entre as características do usuário e do produto exibido. Sua robustez e desempenho tornam-na uma ferramenta adequada para modelar a probabilidade de conversão a partir de dados comportamentais de visualização.

Além dessas características, é importante ressaltar que o uso de múltiplas árvores decorre de duas fontes de aleatoriedade distintas: a seleção de amostras com reposição *(bootstrap sampling)* e a escolha aleatória de subconjuntos de atributos em cada divisão das árvores. Essas duas camadas de variabilidade fazem com que cada árvore aprenda padrões diferentes a partir dos mesmos dados, aumentando substancialmente a diversidade do conjunto. Tal diversidade é o

principal elemento que reduz a variância do modelo final, amortecendo o impacto de ruídos ou amostras atípicas que, em uma árvore isolada, poderiam produzir resultados distorcidos. Outro aspecto relevante é que cada árvore individual tende a apresentar alto erro quando analisada separadamente, mas o conjunto como um todo apresenta desempenho muito superior. Isso ilustra o princípio fundamental dos métodos de ensemble: modelos fracos e instáveis podem se tornar extremamente fortes quando combinados adequadamente. A votação majoritária na RF amplia os acertos e dilui erros individuais, produzindo uma predição mais estável.

A RF também oferece, de forma natural, medidas de importância das variáveis utilizadas. Ao analisar como cada atributo contribui para a redução da impureza durante a construção das árvores, é possível identificar quais características possuem maior influência na decisão final do modelo. Em contextos de e-commerce, isso significa identificar quais fatores do comportamento do usuário, como tempo de navegação, número de visualizações anteriores, tipo de produto consultado ou horário de acesso, exercem maior impacto sobre a probabilidade de compra. Essas informações são valiosas para empresas que desejam ajustar estratégias de recomendação ou otimizar campanhas de marketing. Outro mecanismo particularmente útil da *Random Forest* é o erro *out-of-bag* (OOB). Como parte dos dados não é incluída no treinamento de cada árvore individual (devido ao uso de amostragem com reposição), essas observações podem ser utilizadas como validação imediata para estimar o desempenho generalizado do modelo. Isso significa que a RF é capaz de avaliar seu próprio desempenho sem a necessidade de uma etapa de validação separada, tornando o processo mais eficiente.

Por fim, vale destacar que a RF é capaz de capturar relações complexas e interações entre variáveis sem que essas relações precisem ser especificadas manualmente no modelo. Em problemas de comportamento de compra, como o tratado neste trabalho, tais interações são frequentes e, muitas vezes, não triviais de serem detectadas por modelos lineares.

3 TRABALHOS RELACIONADOS

3.1 *Prevenção da rotatividade de clientes de comércio eletrônico usando a estratégia de inteligência de negócios baseada em aprendizado de máquina*

O trabalho desenvolvido por Shobana *et al.* (2023) tem como objetivo apresentar estratégias de retenção de clientes em sistemas de comércio eletrônico, por meio da aplicação de inteligência de negócios baseada em algoritmos de aprendizado de máquina. A abordagem visa não apenas reter consumidores, mas também prevenir a rotatividade futura.

A base de dados utilizada pelos autores é composta por informações reais de uma empresa de e-commerce, incluindo dados como: pesquisas realizadas, compras efetuadas, valores das compras, pontuações dos produtos, datas da primeira e última compra, frequência de compras, avaliações e feedbacks dos clientes. Por motivos de privacidade e segurança, o conjunto completo de dados é mantido exclusivamente pela empresa, sendo disponibilizado apenas mediante solicitação.

A problemática central abordada é a rotatividade de clientes, também chamada de atrito, que ocorre quando um cliente encerra sua relação com a empresa, deixando de consumir seus produtos ou serviços. Para lidar com essa questão, o estudo propõe o uso da Máquina de Vetores de Suporte (SVM, do inglês Support Vector Machine), uma técnica de aprendizado supervisionado que busca encontrar o hiperplano ótimo de separação entre classes em um espaço de múltiplas dimensões. Além do SVM, os autores também empregaram uma Rede Neural do tipo Backpropagation (BP) para modelar a probabilidade de rotatividade.

Os resultados mostraram que o modelo baseado em SVM obteve uma taxa de acurácia de 77,36%, enquanto a rede neural BP superou esse desempenho, atingindo 82,64%. Com isso, os autores reforçam a importância de estratégias de retenção no comércio eletrônico, dado que o custo de aquisição de novos clientes pode ser até cinco vezes maior do que o custo de retenção. Assim, é recomendável que os gestores se concentrem em aspectos como o volume e a frequência das compras, adotando medidas efetivas para reduzir o índice de rotatividade e promover o crescimento sustentável da empresa.

3.2 *Aprimoramento de aplicativos de comércio eletrônico com sistemas de recomendação baseados em aprendizado de máquina*

O estudo de Farooqi *et al.* (2022) aborda estratégias para o aprimoramento de plataformas de e-commerce por meio da aplicação de aprendizado de máquina e análise de dados. Para isso, os autores simularam um conjunto de dados contendo informações sobre produtos e avaliações de usuários. O pré-processamento foi realizado utilizando a técnica *Bag-of-Words*, que extrai características de dados textuais para posterior aplicação em algoritmos de aprendizado de máquina e recuperação da informação.

Três métodos principais de recomendação foram avaliados:

- **Filtragem colaborativa**: identifica grupos de usuários com padrões de comportamento semelhantes para prever preferências futuras.
- **Filtragem baseada em conteúdo**: utiliza as características dos itens previamente avaliados pelo usuário para sugerir novos produtos.
- **Sistema híbrido**: combina os dois métodos anteriores com o objetivo de superar suas limitações e aumentar a precisão das recomendações.

Os resultados apontaram que o modelo híbrido foi o mais eficaz, com uma taxa de acurácia de 42,06%, superior aos modelos colaborativo (37,28%) e baseado em conteúdo (37,32%). Assim, conclui-se que sistemas híbridos são mais eficazes para recomendação em ambientes de comércio eletrônico, embora haja espaço para melhorias nos algoritmos empregados.

3.3 *Segmentação via Machine Learning: Proposta de Clusterização de Consumidores do E-Commerce de uma empresa multinacional do varejo esportivo*

O trabalho apresentado por Falqueto e Cezar (2022) propõe uma abordagem de segmentação de consumidores em ambientes de e-commerce, com o intuito de identificar perfis de compra semelhantes e otimizar estratégias de retenção e atração de clientes. O estudo utilizou um banco de dados com informações de 526.686 clientes, com histórico de vendas entre outubro de 2019 e setembro de 2020. A análise foi realizada com o uso da técnica de clusterização K-Means, que permite agrupar dados com base em similaridade — utilizando, neste caso, a distância euclidiana como métrica.

Inicialmente, os autores avaliaram a segmentação vigente com base no valor gasto

por cliente nos últimos 12 meses. Posteriormente, realizaram uma reconstrução da base de dados, inserindo variáveis como demanda (valor total gasto), número de pedidos e tempo de inatividade. Com base na clusterização dos dados, o valor de K=4 foi considerado o mais adequado para representar os diferentes grupos de consumidores. Os resultados revelaram que a nova segmentação foi mais eficaz do que a anterior, permitindo insights mais precisos sobre os padrões de comportamento dos clientes.

3.4 *Sistema de recomendação baseado em aprendizado de máquina para comércio eletrônico*

O estudo conduzido por Loukili *et al.* (2023) tem como foco o desenvolvimento de um sistema de recomendação para plataformas de e-commerce, baseado em técnicas de mineração de dados. O objetivo é sugerir produtos com base no histórico de compras dos usuários, combatendo o excesso de informação e a dificuldade de personalização nas interfaces dessas plataformas.

Para a implementação, os autores utilizaram o dataset "Online Retail"disponibilizado pela UCI Machine Learning Repository. O conjunto de dados inclui variáveis como: *StockCode*, *InvoiceNo*, *Description*, *Quantity*, *InvoiceDate*, *UnitPrice*, *CustomerID* e *Country*. O algoritmo adotado foi o FP-Growth, amplamente utilizado para geração de regras de associação. O sistema resultante produziu 4.970 regras, com uma probabilidade média de compra (Paverage) de 69,3%, indicando alto potencial de conversão. Apesar dos bons resultados, os autores sugerem como trabalho futuro a adoção de algoritmos mais avançados de aprendizado de máquina, bem como a realização de testes A/B em ambientes reais para validação prática.

3.5 Comparativo entre atividades do estado da arte

A Tabela 1 apresenta um comparativo entre os principais trabalhos do estado da arte discutidos anteriormente e as atividades que serão desenvolvidas nesta pesquisa.

Tabela 1 – Comparação entre esta monografia e o estado da arte.

Lista de Atividades	Trabalhos Comparados				
	Shobana (2023)	Farooqi (2022)	Falqueto e Cezar (2022)	Loukili (2023)	TCC Breno
Prevenção da rotatividade de clientes usando um modelo de BP	X				
Aprimoramento de aplicativos e-commerce usando um modelo de filtragem híbrida		X			
Clusterização de consumidores usando o algoritmo K-Means			X		
Desenvolvimento de um sistema de recomendação usando o algoritmo Growth FP				X	
Uso de regressão logística e floresta aleatória					X

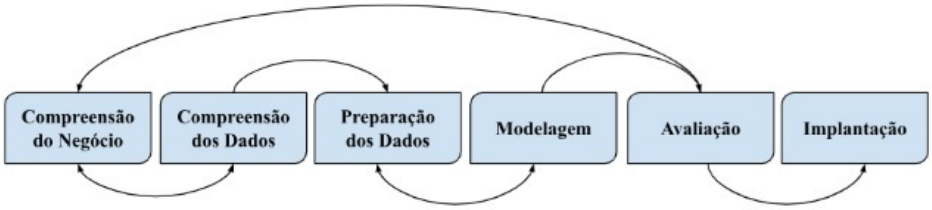
Fonte: Elaborado pelo Autor (2025).

4 PROCEDIMENTOS METODOLÓGICOS

A presente pesquisa caracteriza-se como quantitativa, de natureza descritiva e explicativa, utilizando técnicas de aprendizado de máquina como a Regressão Logística e Random Forest. O principal desafio é prever se uma sessão terminará em uma compra com base nos padrões de navegação do usuário. O estudo utiliza dados secundários, obtidos a partir de um conjunto de dados público disponibilizado na plataforma Kaggle, denominado ***E-Commerce Customer Journey – Click to Conversion***, que reúne informações sobre a jornada do consumidor desde a navegação até a conversão.

Como abordagem metodológica, foi adotada a CRISP-DM (Cross Industry Standard Process for Data Mining), uma metodologia amplamente utilizada em projetos de mineração de dados e ciência de dados, por fornecer um processo estruturado e iterativo para a construção de modelos preditivos. A CRISP-DM é composta por seis fases principais: compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e implementação.

Figura 1 – Fases interativas do CRISP-DM.



Fonte: Adaptado de Plotnikova *et al.* (2022)

Inicialmente, na fase de compreensão do negócio, definiu-se o problema de pesquisa, bem como os objetivos relacionados à previsão do comportamento de compra dos usuários em plataformas de comércio eletrônico. Em seguida, na etapa de compreensão dos dados, realizou-se a exploração do conjunto de dados, identificando suas principais características, variáveis disponíveis e possíveis inconsistências.

A fase de preparação dos dados envolveu procedimentos de limpeza, transformação e seleção de atributos relevantes, tornando o conjunto de dados adequado para a aplicação de algoritmos de aprendizado de máquina. Posteriormente, na etapa de modelagem, foram aplicadas técnicas de modelagem preditiva por meio de algoritmos supervisionados, visando à construção de modelos capazes de estimar a probabilidade de conversão dos usuários.

Na fase de avaliação, os modelos desenvolvidos foram analisados com base em

métricas de desempenho adequadas, permitindo comparar os resultados obtidos e verificar sua aderência aos objetivos propostos. Por fim, a etapa de implementação foi abordada de forma conceitual, discutindo-se a aplicabilidade dos modelos preditivos em cenários reais de e-commerce como apoio à tomada de decisão estratégica.

4.1 Compreensão do Negócio

Na fase de compreensão do negócio, conforme a metodologia CRISP-DM, buscou-se compreender o contexto do comércio eletrônico e a importância da análise do comportamento do consumidor para a tomada de decisões estratégicas. O crescimento do e-commerce tem aumentado a competitividade entre empresas, tornando essencial a utilização de dados para compreender os fatores que influenciam a conversão de usuários em compradores.

O problema de negócio abordado neste estudo refere-se à dificuldade de prever o comportamento de compra dos usuários durante sua navegação em plataformas de comércio eletrônico, uma vez que grande parte das sessões não resulta em conversão. Diante desse cenário, torna-se relevante o uso de técnicas analíticas capazes de antecipar a probabilidade de compra.

Como critérios de sucesso, considera-se a capacidade dos modelos em apresentar desempenho satisfatório na previsão da variável alvo, bem como em fornecer ***insights*** relevantes sobre os fatores que influenciam o comportamento de compra. O estudo está sujeito a restrições relacionadas ao uso de dados secundários e à ausência de implementação em ambiente produtivo, sendo conduzido com finalidade acadêmica.

4.2 Compreensão dos Dados

Essa etapa inicial dessa pesquisa consiste na exploração de uma base de dados no Kaggle intitulado de "E-Commerce Customer Journey: Click to Conversion"com o intuito de compreender e resumir as principais caraterísticas dos dados a serem analisados. Para isso, realizou-se o ***download*** da base de dados referida acima e a análise dos dados utilizando o ***Google Collab*** que é um serviço gratuito na nuvem que permite a escrever e executar código ***Python*** diretamente no navegador sem necessidade de instalar nada, usando o ***Jupyter Notebok***. Os dados estão armazenados em um único arquivo **.csv(comma-separated values, ou valores separados por vírgulas)** e possui 12719 **linhas** e 10 **colunas**. Informações adicionais sobre cada uma das *features* disponíveis na base são apresentadas no Quadro 1.

Quadro 1 – Descrição das variáveis do conjunto de dados

Variável	**Tipo**	**Descrição**
SessionID	Categórica	Identificador único para cada sessão de usuário.
UserID	Categórica	Identificador único para cada usuário.
Timestamp	Temporal	Data e hora em que o evento foi registrado.
PageType	Categórica	Tipo de página visitada pelo usuário, podendo ser: home, product_page, carrinho, checkout ou confirmação.
DeviceType	Categórica	Tipo de dispositivo utilizado na sessão, como Desktop, Mobile ou Tablet.
Country	Categórica	País de origem do usuário.
ReferralSource	Categórica	Fonte de referência que direcionou o usuário ao site, como Google, Redes Sociais, Direto ou Email.
TimeOnPage_seconds	Numérica	Tempo gasto pelo usuário na página específica, medido em segundos.
ItemsInCart	Numérica	Quantidade de itens presentes no carrinho do usuário no momento do evento.
Purchased	Binária	Variável alvo do estudo, assumindo valor 1 quando ocorre uma compra durante a sessão e 0 caso contrário.

Fonte: Dataset Kaggle.

Foi realizada uma análise exploratória com o objetivo de identificar padrões iniciais, tendências e possíveis relações entre as variáveis do conjunto de dados. Nessa etapa, investigaram se tanto o comportamento das variáveis explicativas quanto sua associação com a variável alvo, permitindo uma compreensão preliminar dos fatores que mais influenciaram a conversão. A análise exploratória permitiu identificar padrões relevantes no comportamento dos usuários ao longo das sessões de navegação. Inicialmente, observou-se que o conjunto de dados apresentou 12.719 registros e três variáveis numéricas principais *(TimeOnPage_seconds*, *ItemsInCart* e *Purchased*). A taxa média de conversão, *mean (Purchased)* foi de aproximadamente 30,7%, indicando que a maioria das sessões não resultou em compra, conforme evidenciado pela distribuição da variável alvo apresentada na Figura 2.

Figura 2 – Descrição estatística da distribuição dos dados da base

	TimeOnPage_seconds	ItemsInCart	Purchased
count	12719.000000	12719.000000	12719.000000
mean	97.427707	1.138533	0.397044
std	48.120729	1.689954	0.489304
min	15.000000	0.000000	0.000000
25%	56.000000	0.000000	0.000000
50%	98.000000	0.000000	0.000000
75%	139.000000	2.000000	1.000000
max	180.000000	5.000000	1.000000

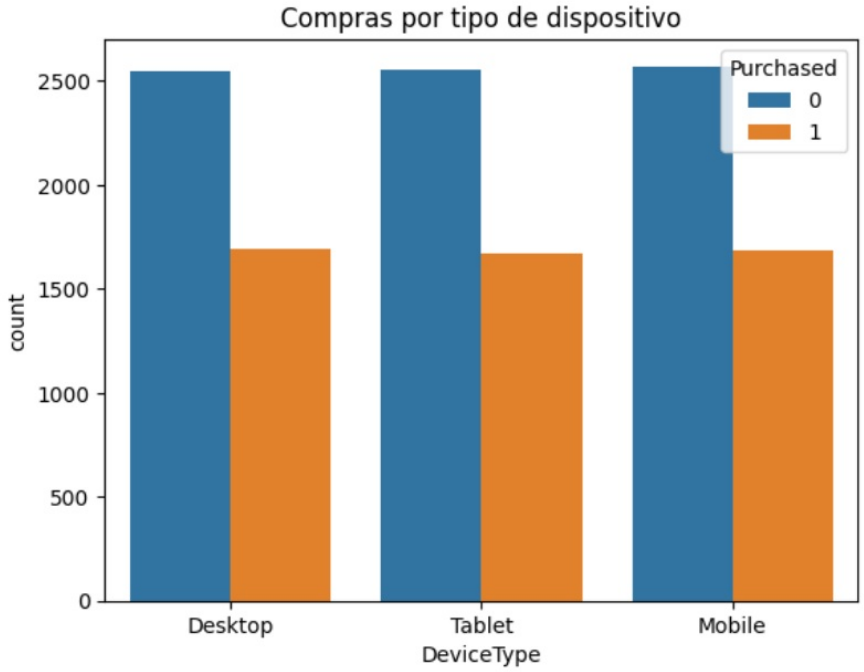
Fonte: Elaborada pelo autor.

No que se refere ao tipo de dispositivo utilizado, verificou-se que os volumes de acesso foram semelhantes entre *desktop*, *tablet* e *mobile*. Em todos os casos, o número de sessões sem compra superou o número de sessões convertidas, como ilustrado na Figura 3. Esse comportamento sugeriu que o tipo de dispositivo, isoladamente, não se apresentou como um discriminador forte da probabilidade de compra.

A análise da fonte de referência (origem do tráfego) revelou padrão semelhante: todas as fontes apresentaram maior volume de sessões sem compra. Entre elas, o Google destacou-se com o maior número absoluto de conversões, seguido por e-mail, enquanto o tráfego direto e a navegação oriunda de redes sociais apresentaram as menores quantidades de compras, conforme mostrado na Figura 4.

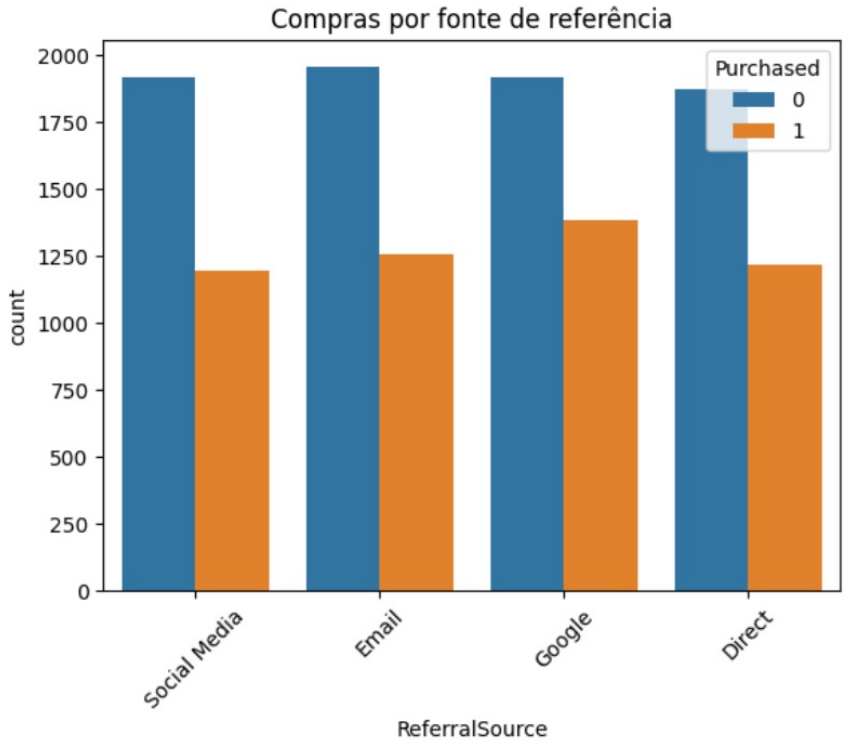
A relação entre o tempo gasto na página e o status de compra foi examinada por meio de boxplots. As distribuições analisadas apresentaram medianas e amplitudes interquartílicas praticamente idênticas para os grupos que converteram e não converteram, indicando que o tempo de permanência não diferiu de maneira significativa entre eles, como pode ser observado.

Figura 3 – Distribuição de sessões com e sem compra por tipo de dispositivo. Observa-se que os três tipos de dispositivo (*desktop*, *tablet* e *mobile*) apresentaram maior volume de sessões sem compra.



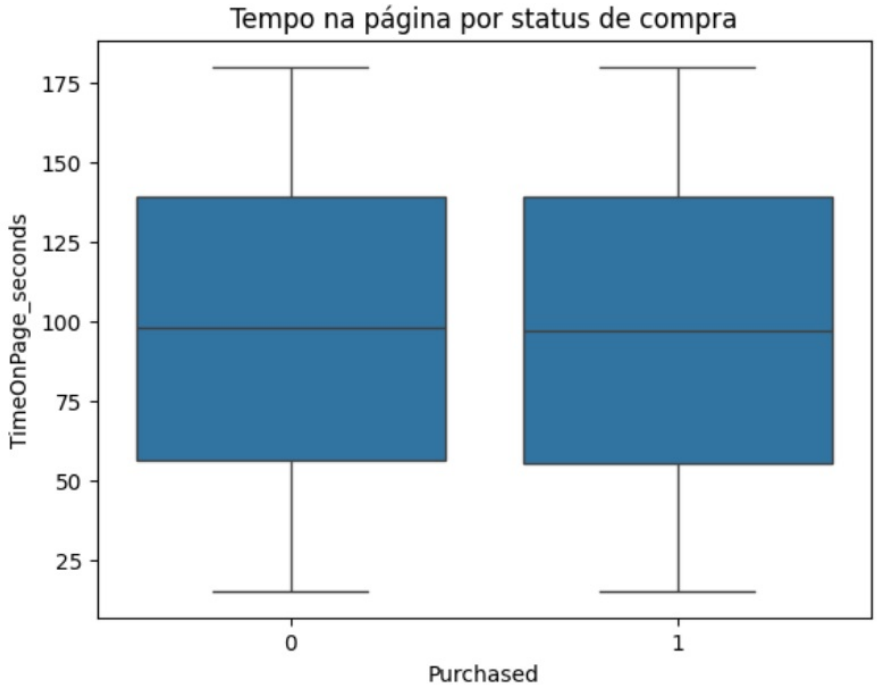
Fonte: Elaborada pelo autor.

Figura 4 – Distribuição de sessões com e sem compra por fonte de referência. As sessões provenientes do Google apresentaram o maior número de conversões.



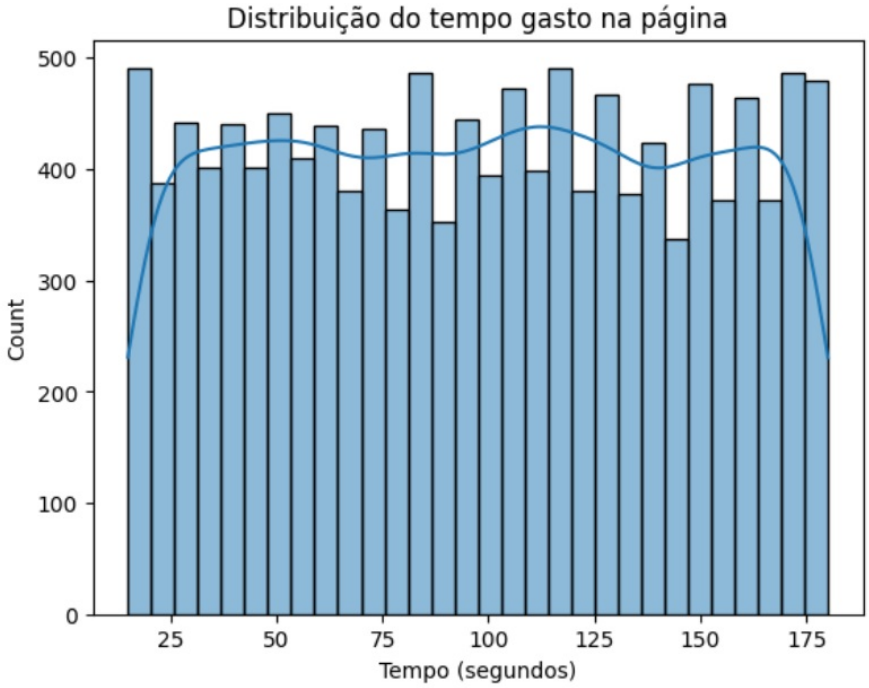
Fonte: Elaborada pelo autor.

Figura 5 – Distribuição do tempo na página para sessões com e sem compra. As medianas e dispersões são semelhantes, indicando pouca diferenciação entre os grupos.



Fonte: Elaborada pelo autor.

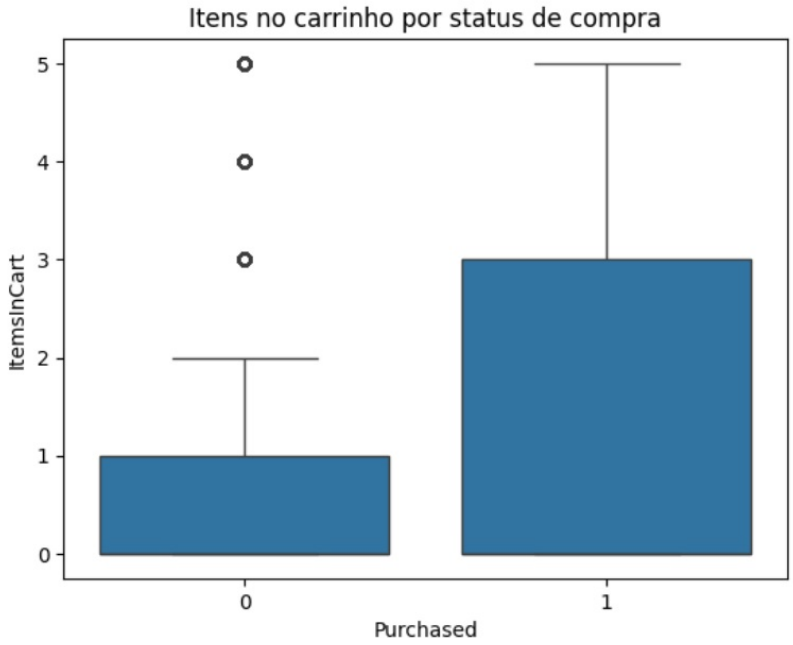
Figura 6 – Distribuição do tempo gasto na página. Observa-se uma dispersão relativamente uniforme no intervalo entre 15 e 180 segundos.



Fonte: Elaborada pelo autor.

Por outro lado, a variável *ItemsInCart* demonstrou maior poder de diferenciação entre os grupos. Sessões que resultaram em compra apresentaram maior número de itens adicionados ao carrinho, bem como maior dispersão de valores, enquanto sessões sem compra concentraram-se predominantemente em valores próximos de zero, como ilustrado na Figura 7. Esse comportamento sugeriu que a quantidade de itens no carrinho esteve mais fortemente associada à probabilidade de conversão em comparação ao tempo de permanência na página.

Figura 7 – Distribuição da quantidade de itens no carrinho por status de compra. Sessões que resultaram em compra apresentaram maior número de itens adicionados.



Fonte: Elaborada pelo autor.

De forma geral, os resultados da análise exploratória forneceram uma compreensão inicial consistente sobre a dinâmica da jornada do usuário na plataforma, permitindo identificar quais variáveis apresentaram maior potencial preditivo para as etapas subsequentes de construção e avaliação dos modelos de aprendizado de máquina.

4.3 Preparação dos dados

Na fase de preparação dos dados, foram realizadas etapas de seleção, limpeza e transformação do conjunto de dados, com o objetivo de torná-lo adequado à aplicação de técnicas de modelagem preditiva. Inicialmente, foram selecionadas as seguintes variáveis: ***PageType***, ***DeviceType***, ***Country***, ***ReferralSource***, ***TimeOnPageSeconds***, ***ItemsInCart***, ignorando ***SessionId***, ***UserId***, ***TimeStamp*** visto que não ajudam na previsão por serem identificadores de alta cardinalidade que não carregam padrões de comportamentais generalizáveis. Mantê-las induziria o modelo ao ***overfitting***, que é quando o algoritmo adapta praticamente quase da mesma forma aos dados de treinamento, levando a um modelo que não consegue fazer previsões ou conclusões precisas com outros dados que não sejam os do treinamento. Então se mantivéssemos essas

variáveis citadas acima, o algoritmo ia memorizar registros históricos específicos ao em vez de aprender as correlações universais entre navegação e conversão, o que seria ruim para a pesquisa. Em seguida, procedeu-se à limpeza dos dados, incluindo a remoção de registros duplicados e os valores atípicos identificados foram analisados e mantidos, por representarem comportamentos reais dos usuários.

As variáveis categóricas foram transformadas em representações numéricas por meio da codificação ***one-hot-encoding***, enquanto as variáveis numéricas passaram por padronização. Foi feita a padronização das variáveis numéricas para evitar influência desproporcional na modelagem.

4.4 Modelagem

Foram aplicados algoritmos de aprendizado de máquina supervisionado com o objetivo de prever o comportamento de compra dos usuários em ambientes de e-commerce. A variável alvo considerada foi a ocorrência de compra, enquanto as variáveis explicativas foram compostas por informações de navegação, tipo de dispositivo, fonte de referência e demais características associadas à sessão do usuário.

Para a construção dos modelos preditivos, foram selecionados dois algoritmos: Regressão Logística e *Random Forest*. A Regressão Logística foi empregada como modelo base devido à sua simplicidade, eficiência e interpretabilidade, possibilitando identificar de forma direta a influência de cada variável sobre a probabilidade de conversão. Já a *Random Forest* foi escolhida por sua capacidade de capturar relações não lineares e interações complexas entre variáveis, além de apresentar robustez em cenários com dados heterogêneos e padrões comportamentais mais difíceis de serem modelados linearmente.

Os dados foram divididos em conjuntos de treinamento e teste, usando a proporção clássica de 80% para os dados de treino e 20% para os dados de teste sendo os modelos treinados a partir do conjunto de treinamento. Para o algoritmo *Random Forest*, foi aplicado um ajuste de hiperparâmetros focado na técnica de poda *(pruning)* e estabilidade do conjunto. Estabeleceu-se o parâmetro ***max_depth*** em 10 níveis, visando controlar a expansão das árvores de decisão e mitigar o risco de sobreajuste aos dados de treinamento.

Adicionalmente, o número de estimadores ***(n_estimators)*** foi expandido para 200, proporcionando uma floresta mais robusta na captura de padrões comportamentais não-lineares.

4.5 Avaliação

Conforme preconiza a metodologia CRISP-DM, a fase de avaliação buscou analisar se o modelo desenvolvido atendeu aos objetivos de negócio estabelecidos inicialmente. Nessa etapa, os modelos de Regressão Logística e *Random Forest* foram submetidos a testes de validação utilizando dados não vistos durante o treinamento, com o intuito de mensurar sua capacidade de generalização no contexto de predição de conversão em e-commerce.

Os resultados obtidos demonstraram que ambos os algoritmos alcançaram níveis de desempenho satisfatórios e relativamente equilibrados. A Regressão Logística apresentou acurácia de 75,12%, superando levemente o Random Forest, que obteve 74,88%. Esse equilíbrio indicou que o problema de predição em estudo possui características de separabilidade linear que o modelo de Regressão Logística conseguiu capturar de maneira eficiente. Por outro lado, o *Random Forest* mostrou-se igualmente robusto, sendo capaz de lidar com a variabilidade dos dados comportamentais sem apresentar sinais de *overfitting* após a limpeza de variáveis de alta cardinalidade, como *SessionIDs*.

Um aspecto crucial dessa fase foi a identificação de um fenômeno de *data leakage* (vazamento de dados) nas iterações preliminares, situação em que colunas associadas a eventos finais da jornada (como páginas de confirmação de compra) elevaram artificialmente a acurácia para valores próximos a 98%. Em conformidade com as boas práticas de modelagem, optou-se pela remoção desses atributos e de identificadores únicos de sessão, resultando em modelos com acurácia numericamente menor, porém metodologicamente válidos e capazes de realizar predições baseadas exclusivamente em variáveis com real poder preditivo, como *TimeOnPageSeconds*, *ItemsInCart* e características demográficas.

Com base nas métricas avaliadas, concluiu-se que a Regressão Logística foi o modelo mais adequado ao cenário analisado, tanto pela acurácia ligeiramente superior quanto pela facilidade de interpretação de seus coeficientes. Essa interpretabilidade permite identificar de forma direta quais fatores, como fontes de referência *(ReferralSource)* e tipos de dispositivo *(DeviceType)*, estiveram mais associados à probabilidade de conversão, oferecendo subsídios valiosos para a gestão estratégica de plataformas de e-commerce.

5 CONCLUSÃO

5.1 Conclusão

O presente trabalho cumpriu o objetivo de desenvolver e avaliar modelos preditivos para o comportamento de compra em ambientes de e-commerce, utilizando a metodologia CRISP-DM como guia estrutural. A transição entre as fases de compreensão dos dados e modelagem evidenciou a importância crítica do tratamento adequado das variáveis, especialmente no processo de identificação e mitigação de *data leakage*, que inicialmente distorcia os resultados e comprometia a validade das predições.

A análise comparativa entre os algoritmos de Regressão Logística e *Random Forest* demonstrou que, após a remoção de identificadores únicos e variáveis indicativas de fechamento de jornada, ambos apresentaram desempenho equilibrado, com acurácias em torno de 75%. A Regressão Logística mostrou-se ligeiramente superior e destacou-se pela interpretabilidade dos coeficientes, permitindo identificar que fatores como tempo de permanência na página e quantidade de itens no carrinho foram os principais preditores da conversão.

Em síntese, o estudo validou a aplicabilidade de técnicas de aprendizado de máquina como ferramenta estratégica no apoio à tomada de decisão em sistemas de e-commerce. O modelo desenvolvido não apenas classificou usuários com base em seu comportamento, como também forneceu subsídios para que gestores pudessem intervir de forma proativa na jornada do cliente, otimizando recursos, aprimorando a experiência de navegação e potencializando as taxas de conversão.

5.2 Trabalhos Futuros

A partir dos resultados alcançados e das limitações observadas ao longo da pesquisa, identificaram-se oportunidades relevantes para aprofundamento e continuidade deste estudo. Os seguintes pontos mostram-se especialmente promissores para expandir a qualidade preditiva dos modelos e ampliar sua aplicabilidade prática no contexto de e-commerce:

(1) **Engenharia de atributos temporais:** Recomenda-se explorar informações contidas no campo Timestamp, transformando-o em variáveis cíclicas, como hora do dia, dia da semana, período do mês ou sazonalidade. Essa abordagem permitiria verificar se padrões temporais influenciam a probabilidade de conversão, possibilitando intervenções estratégicas em períodos

de maior propensão à compra.

(2) **Exploração de algoritmos mais avançados:** Embora os modelos empregados tenham apresentado desempenho satisfatório, algoritmos como *Gradient Boosting Machines* (GBM), *XGBoost*, *LightGBM* ou *CatBoost* podem ser analisados futuramente, principalmente pela capacidade desses métodos em lidar com interações complexas e variações sutis nos dados.

(3) **Estratégias de balanceamento de classes:** Considerando-se o desbalanceamento observado na variável alvo, abordagens como *SMOTE*, *ADASYN* ou ajuste de pesos de classe podem ser avaliadas em estudos posteriores, a fim de investigar se tais técnicas aumentam a estabilidade ou o desempenho dos modelos.

Em conjunto, esses desdobramentos oferecem caminhos consistentes para aprimorar a capacidade preditiva, interpretabilidade e aplicabilidade dos modelos no contexto real de plataformas de e-commerce, fortalecendo a integração entre ciência de dados e tomada de decisão estratégica.

REFERÊNCIAS

ALBERTIN, A. **Comércio Eletrônico: Modelo, Aspectos e Contribuições de sua Aplicação.** [S.l.]: Editora Atlas, 2010. ISBN 9788522456857.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, p. 5–32, 10 2001.

CUTLER, A.; CUTLER, D. R.; STEVENS, J. R. Random forests. In: **Ensemble machine learning**. [S.l.]: Springer, 2012. p. 157–175.

FALQUETO, A. A.; CEZAR, L. C. Segmentação via machine learning: Proposta de clusterização de consumidores do e-commerce de uma empresa multinacional do varejo esportivo. **HOLOS**, v. 4, 2022.

FAROOQI, R.; KESARWANI, S.; SHAKEEB, M.; SHARMA, N.; BHATNAGAR, I. Enhancing e-commerce applications with machine learning recommendation systems. **International Journal of Scientific Research in Science, Engineering and Technology**, p. 85–90, 05 2022.

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied logistic regression**. [S.l.]: John Wiley & Sons, 2013.

LOUKILI, M.; MESSAOUDI, F.; GHAZI, M. E. Machine learning based recommender system for e-commerce. **IAES International Journal of Artificial Intelligence**, v. 12, n. 4, p. 1803–1811, 2023.

PAIXÃO, G. M. de M.; SANTOS, B. C.; ARAUJO, R. M. de; RIBEIRO, M. H.; MORAES, J. L. de; RIBEIRO, A. L. Machine learning na medicina: Revisão e aplicabilidade. **Arquivos brasileiros de cardiologia**, SciELO Brasil, v. 118, n. 1, p. 95, 2022.

PLOTNIKOVA, V.; DUMAS, M.; MILANI, F. Applying the crisp-dm data mining process in the financial services industry: Elicitation of adaptation requirements. **Data Knowledge Engineering**, v. 139, p. 102013, 04 2022.

PONDEL, M.; WUCZYŃSKI, M.; GRYNCEWICZ, W.; ŁYSIK, Ł.; HERNES, M.; ROT, A.; KOZINA, A. Deep learning for customer churn prediction in e-commerce decision support. In: **Business Information Systems**. [S.l.: s.n.]. p. 3–12. 2021.

RIS-ALA, R. **Fundamentos de Aprendizagem por Reforço**. [S.l.]: Rafael Ris-Ala, 2023.

SHOBANA, J.; GANGADHAR, C.; ARORA, R. K.; RENJITH, P.; BAMINI, J.; CHINCHOLKAR, Y. devidas. E-commerce customer churn prevention using machine learning-based business intelligence strategy. **Measurement: Sensors**, Elsevier, v. 27, p. 100728, 2023.