REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

MACHINE LEARNING APLICADO À GESTÃO DE ESTOQUE: UM ESTUDO DE CASO DO SETOR DE SAÚDE E BELEZA

MACHINE LEARNING APPLIED TO INVENTORY MANAGEMENT: A CASE STUDY IN THE HEALTH AND BEAUTY SECTOR

MACHINE LEARNING APLICADO A LA GESTIÓN DE INVENTARIOS: UN ESTUDIO DE CASO EN EL SECTOR DE SALUD Y BELLEZA

Paulo Henrique Martins[1], Felipe Diniz Dallilo[2]

https://doi.org/10.47820/recima21.v7i12.9750

PUBLICADO: 12/2026

RESUMO

O crescimento acelerado do comércio eletrônico exige operações logísticas cada vez mais eficientes, especialmente no setor de Beleza e Saúde, cujos produtos apresentam prazos de validade rigorosos e forte volatilidade de consumo. Este trabalho tem como objetivo desenvolver e avaliar algoritmos preditivos de *Machine Learning* capazes de estimar a demanda futura dessa categoria, visando mitigar os riscos de ruptura e o excesso de mercadorias no armazém. A pesquisa caracteriza-se como aplicada e quantitativa, utilizando dados reais de transações extraídos do *Brazilian E-Commerce Public Dataset* (Olist). O processamento computacional foi realizado na linguagem Python, na qual os registros brutos foram tratados e transformados em uma série temporal diária. Para fins de validação comparativa, o estudo avaliou o desempenho do modelo estatístico tradicional ARIMA contra o algoritmo de aprendizado supervisionado *Random Forest*, utilizando as métricas matemáticas MAE, RMSE e WAPE. Nos experimentos realizados, o algoritmo Random Forest apresentou melhor desempenho preditivo nas métricas avaliadas, obtendo WAPE de 31,73%, frente a 37,00% do ARIMA, além de maior aderência visual às oscilações sazonais da série. Conclui-se que a adoção de modelos orientados a dados oferece suporte analítico mais consistente ao planejamento de compras, com potencial para auxiliar na mitigação de rupturas e excessos de estoque no setor.

PALAVRAS-CHAVE: Machine Learning. Previsão de Demanda. Comércio Eletrônico. Gestão de Estoque. Séries Temporais.

ABSTRACT

The accelerated growth of e-commerce requires increasingly efficient logistics operations, especially in the Beauty and Health sector, where products have strict expiration dates and strong consumption volatility. This study aims to develop and evaluate predictive Machine Learning algorithms capable of estimating the future demand for this category, aiming to mitigate the risks of stockouts and excess warehouse inventory. The research is characterized as applied and quantitative, using real transaction data extracted from the Brazilian E-Commerce Public Dataset (Olist). Computational processing was performed in Python, where raw records were cleaned and transformed into a daily time series. For comparative validation purposes, the study evaluated the performance of the traditional statistical model ARIMA against the supervised learning algorithm Random Forest, using the mathematical metrics MAE, RMSE, and WAPE.

[1] Universidade de Araraquara, Graduando em Sistemas de Informação.
[2] Universidade de São Paulo, Doutor em Engenharia de Software.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

In the experiments conducted, the Random Forest algorithm showed better predictive performance across the evaluated metrics, achieving a WAPE of 31.73%, compared to 37.00% for ARIMA, as well as greater visual alignment with the seasonal fluctuations in the time series. It is concluded that the adoption of data-driven models provides more consistent analytical support for purchasing planning, with the potential to help mitigate stockouts and excess inventory in the sector.

KEYWORDS: *Machine Learning. Demand Forecasting. E-commerce. Inventory Management. Time Series.*

RESUMEN
El crecimiento acelerado del comercio electrónico exige operaciones logísticas cada vez más eficientes, especialmente en el sector de Belleza y Salud, cuyos productos presentan fechas de caducidad rigurosas y fuerte volatilidad de consumo. Este trabajo tiene como objetivo desarrollar y evaluar algoritmos predictivos de Machine Learning capaces de estimar la demanda futura de esta categoría, buscando mitigar los riesgos de ruptura y el exceso de mercancías en el almacén. La investigación se caracteriza como aplicada y cuantitativa, utilizando datos reales de transacciones extraídos del Brazilian E-Commerce Public Dataset (Olist). El procesamiento computacional se realizó en el lenguaje Python, donde los registros brutos fueron tratados y transformados en una serie temporal diaria. Para fines de validación comparativa, el estudio evaluó el rendimiento del modelo estadístico tradicional ARIMA frente al algoritmo de aprendizaje supervisado Random Forest, utilizando las métricas matemáticas MAE, RMSE y WAPE. En los experimentos realizados, el algoritmo Random Forest presentó un mejor desempeño predictivo en las métricas evaluadas, con un WAPE del 31,73%, frente al 37,00% del ARIMA, además de una mayor correspondencia visual con las oscilaciones estacionales de la serie temporal. Se concluye que la adopción de modelos basados en datos ofrece un apoyo analítico más consistente para la planificación de compras, con potencial para contribuir a la mitigación de las rupturas de stock y los excesos de inventario en el sector.

PALABRAS CLAVE*: Machine Learning. Previsión de Demanda. Comercio Electrónico. Gestión de Inventarios. Series Temporales.*

INTRODUÇÃO

O crescimento acelerado do *e-commerce* brasileiro trouxe desafios complexos para a gestão logística, em especial o controle de estoque (Michel, 2022). A manutenção de estoques inadequados gera custos de armazenagem para produtos encalhados ou perda de receita devido à ruptura (falta de produtos). O trabalho delimita-se a investigar a previsão de demanda no comércio eletrônico, com foco específico na categoria de "Beleza e Saúde", utilizando algoritmos de *Machine Learning* (Aprendizado de Máquina).

Apesar de muitos varejistas utilizarem médias históricas simples para prever vendas, produtos de beleza possuem alta recorrência de compra, sazonalidade forte e prazos de validade rigorosos, o que exige métodos mais sofisticados. Segundo Martins (2023), a aplicação de modelos de aprendizado de máquina em séries temporais surge como uma solução robusta para

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

lidar com essas variáveis, preenchendo a lacuna entre a gestão tradicional de estoque e a aplicação prática de inteligência artificial preditiva baseada em dados reais.

O objetivo geral deste trabalho é desenvolver um modelo preditivo baseado em *Machine Learning* capaz de estimar a demanda futura de produtos da categoria de Beleza e Saúde, utilizando dados históricos de vendas de um *e-commerce* brasileiro. Para alcançar essa meta principal, foram estabelecidos os seguintes objetivos específicos: realizar o pré-processamento e a limpeza de um conjunto de dados público de *e-commerce*, isolando a categoria de interesse; identificar padrões de sazonalidade e tendências temporais nas vendas desses itens; treinar algoritmos de Aprendizado de Máquina (como Regressão Linear, Árvores de Decisão ou modelos de Séries Temporais) para a previsão de saída diária e mensal; e, por fim, comparar o desempenho dos modelos treinados por meio de métricas de erro estatístico, determinando a abordagem mais assertiva para o cenário proposto.

Diante da complexidade do gerenciamento logístico no varejo digital, a questão de pesquisa que norteia este estudo consiste em: como modelos de *Machine Learning* podem ser aplicados para prever com precisão a demanda de produtos de beleza e saúde em um *e-commerce*, auxiliando efetivamente na mitigação da ruptura e do excesso de estoque?

Para responder a esse questionamento, levanta-se a hipótese de que a aplicação de algoritmos preditivos em séries temporais de dados históricos de vendas permite identificar padrões de consumo complexos e sazonais que geralmente não são percebidos por métodos gerenciais tradicionais. Dessa forma, pressupõe-se que a adoção dessas tecnologias resulte em uma estimativa de estoque significativamente mais assertiva, validando o uso da inteligência artificial como uma ferramenta essencial para a otimização logística e a tomada de decisão no setor.

Foi realizada uma pesquisa bibliográfica e explicativa, adotando uma abordagem estritamente quantitativa para a análise estatística de um grande volume de dados numéricos e métricas de desempenho computacional. O procedimento técnico configura-se como um estudo de caso experimental, fundamentado em pesquisa documental e bibliográfica para a sustentação teórica do projeto. O universo da pesquisa compreende o comércio eletrônico brasileiro, tendo como amostra específica o *Brazilian E-Commerce Public Dataset by Olist*. Para a realização do experimento, os dados foram filtrados para abranger exclusivamente as transações da categoria de "Beleza e Saúde".

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

1. REFERENCIAL TEÓRICO

1.1. Gestão de estoque

A gestão de estoques é um dos elementos centrais para o sucesso operacional e financeiro de qualquer empresa varejista. O estoque atua, na prática, como uma reserva estratégica para alinhar a oferta e a demanda, garantindo que o consumidor final encontre o produto desejado no exato momento da compra. No entanto, manter mercadorias guardadas em galpões exige espaço físico, controle de segurança e, principalmente, capital imobilizado.

De acordo com Ballou (2006), o grande objetivo da administração de estoques é encontrar o ponto de equilíbrio financeiro da operação. A empresa não pode manter produtos em excesso, sob o risco de comprometer seu caixa e arcar com altos custos de armazenagem, e também não pode operar com mercadorias de menos, para evitar a ruptura de estoque, que resulta em perda direta de vendas. Historicamente, o controle desses níveis era realizado de forma manual e fundamentado na experiência dos gestores. Contudo, o aumento da complexidade do mercado moderno exige processos de controle muito mais dinâmicos, forçando as organizações a abandonarem modelos reativos, nos quais se compra apenas quando o produto falta, para adotarem estratégias preventivas de reposição.

1.2. Desafios do *e-commerce*

Se a gestão de estoques já é uma tarefa complexa no varejo físico convencional, o cenário se torna ainda mais desafiador no ambiente digital. O comércio eletrônico transformou as dinâmicas de consumo, eliminando fronteiras geográficas e elevando significativamente a expectativa dos clientes por disponibilidade imediata e entregas rápidas. Conforme explicam Chopra e Meindl (2016), a facilidade que o consumidor tem de comparar preços e trocar de fornecedor com apenas um clique faz com que a falta de um item resulte não apenas na perda imediata da venda, mas muitas vezes na perda do cliente para a concorrência.

No cenário brasileiro, as distâncias e os custos logísticos exigem operações altamente eficientes para equilibrar a disponibilidade dos produtos e as despesas de armazenagem (Michel, 2022). Dentro desse ecossistema digital, o segmento específico de Beleza e Saúde apresenta particularidades que agravam o controle logístico. Produtos cosméticos e de cuidados pessoais não podem permanecer armazenados por tempo indeterminado. Como destacam Shikuma, Leme e Lopes (2025), esses itens possuem prazos de validade rigorosos e são fortemente influenciados por variações de tendências e sazonalidades, como datas comemorativas e mudanças de estação. Consequentemente, tentar gerenciar o estoque deste nicho utilizando

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

métodos tradicionais frequentemente resulta no descarte de mercadorias vencidas ou em prateleiras virtuais vazias, gerando prejuízos para a operação.

1.3. Previsão de demanda

Para mitigar os riscos associados ao controle de estoque, a previsão de demanda atua como um mecanismo estratégico e indispensável. Prever a demanda significa tentar reduzir a incerteza sobre o futuro, permitindo que a empresa planeje a sua capacidade de armazenagem e de compra com antecedência (Slack; Brandon-Jones; Johnston, 2018). No varejo, antecipar o que o cliente vai adquirir evita tanto o excesso de produtos parados quanto a falta deles.

Historicamente, as organizações utilizam métodos qualitativos, fundamentados na intuição e na experiência dos gestores, e métodos quantitativos, que se apoiam puramente em dados numéricos do passado. No ambiente do comércio eletrônico, devido ao alto volume de transações diárias, a abordagem quantitativa tornou-se o padrão operacional. Tradicionalmente, o varejo tem se apoiado em métodos gerenciais simples dentro dessa categoria, como calcular a média de vendas dos meses anteriores para estimar o mês seguinte. Contudo, depender exclusivamente de médias simples em um mercado dinâmico resulta em previsões engessadas, que não acompanham as rápidas mudanças no comportamento do consumidor.

1.4. Séries temporais

Dentro do campo das previsões quantitativas, a abordagem mais estruturada é o uso de séries temporais. Essa técnica consiste na organização, observação e análise de dados históricos coletados em intervalos de tempo regulares, como horas, dias ou semanas, com o intuito de projetar comportamentos futuros (Hyndman; Athanasopoulos, 2021).

De acordo com Morettin e Toloi (2018), uma série temporal não é apenas um conjunto de números soltos; ela é composta por componentes matemáticos específicos. Os principais são a tendência, que indica se as vendas estão crescendo ou caindo a longo prazo, e a sazonalidade, que representa os padrões que se repetem em períodos conhecidos, como o aumento nas vendas de protetor solar no verão.

Embora o uso de séries temporais seja a base estatística do varejo, a literatura aponta que os métodos tradicionais desse campo apresentam limitações ao tentar capturar padrões não lineares e lidar com múltiplos fatores simultâneos. O mercado real é afetado diariamente por variáveis externas, como o impacto de campanhas promocionais, mudanças repentinas de preço e ações da concorrência. Nesses cenários complexos, a estatística clássica demonstra

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

fragilidade, evidenciando uma lacuna tecnológica na gestão logística convencional e justificando a necessidade de ferramentas computacionais mais avançadas para o processamento de dados.

1.5. Fundamentos de *machine learning*

O Aprendizado de Máquina, ou *Machine Learning*, representa uma subárea da Inteligência Artificial dedicada ao desenvolvimento de algoritmos capazes de extrair padrões de dados e melhorar seu desempenho por meio da experiência. Diferentemente da programação computacional tradicional, na qual o desenvolvedor precisa escrever regras lógicas manuais e estáticas para instruir o sistema sobre como resolver um problema, os algoritmos de *Machine Learning* analisam grandes conjuntos de dados históricos para identificar essas regras de forma matemática e automática (Mitchell, 1997; Géron, 2021).

Na literatura científica, a forma como os algoritmos aprendem a partir dos dados costuma ser dividida em três grandes vertentes principais: o aprendizado não supervisionado, que busca identificar agrupamentos ou estruturas escondidas em dados sem rótulos prévios; o aprendizado por reforço, em que o agente computacional aprende por meio de tentativas, erros e penalidades; e o aprendizado supervisionado (Faceli et al., 2021).

No contexto específico deste estudo, cujo propósito é projetar a quantidade de produtos que saem do estoque, adota-se o Aprendizado Supervisionado. Essa metodologia opera fornecendo ao algoritmo um conjunto de dados de treinamento contendo tanto as variáveis explicativas de entrada, chamadas de atributos, como a data, o mês e o dia da semana, quanto os desfechos reais de mercado conhecidos como rótulos, que representam a quantidade vendida. A partir dessa estrutura, a inteligência artificial analisa o histórico para mapear a relação entre as variáveis de entrada e a variável-alvo, ajustando seus parâmetros internos autonomamente para conseguir realizar previsões quando receber novos dados (Faceli et al., 2021).

Dentro do aprendizado supervisionado, os problemas dividem-se essencialmente em tarefas de classificação, quando a resposta esperada é uma categoria discreta, e tarefas de regressão, quando o objetivo é projetar um valor numérico contínuo. Como a estimativa de estoque busca prever o volume exato de itens comercializados em um determinado intervalo de tempo, este estudo enquadra-se como um problema de regressão. Durante o treinamento, o algoritmo calcula o desvio entre a estimativa gerada e o valor real retido no histórico, reduzindo essa margem de erro gradativamente para capacitar o sistema a projetar os cenários de abastecimento futuros com maior precisão.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

1.5.1. Machine learning aplicado à gestão de varejo

A transição para o varejo digital gerou um volume de informações sem precedentes. Nesse cenário, a previsão de demanda tem apontado cada vez mais para a adoção de técnicas de Inteligência Artificial, especificamente o *Machine Learning* (Aprendizado de Máquina). Diferentemente das abordagens estatísticas clássicas, os modelos de aprendizado de máquina possuem a capacidade de processar grandes volumes de informações (*Big Data*) de forma simultânea e identificar padrões complexos de consumo de maneira automatizada (Tamachiro et al., 2024). No ambiente do *e-commerce*, isso significa transformar milhares de registros soltos de transações financeiras em conhecimento aplicável para a operação logística.

Uma das maiores vantagens da aplicação dessas tecnologias no gerenciamento de estoques é a habilidade de lidar com múltiplas variáveis ao mesmo tempo. Enquanto os métodos tradicionais analisam o histórico de vendas de forma linear e isolada, o *Machine Learning* consegue cruzar o comportamento passado com diversos fatores externos. Algoritmos estruturados, como Árvores de Decisão, *Random Forest* e Redes Neurais, vêm ganhando destaque mercadológico e acadêmico justamente por essa flexibilidade. Essas técnicas conseguem ponderar o impacto simultâneo de campanhas promocionais, variações de preço, dias da semana e feriados, gerando estimativas preditivas com taxas de acurácia muito superiores, mesmo em categorias voláteis como a de cosméticos e cuidados pessoais.

Além dos ganhos técnicos em precisão, a incorporação de modelos preditivos computacionais na rotina do *e-commerce* vai além da simples automação do cálculo de compras. Conforme argumentam Provost e Fawcett (2016), o uso inteligente da tecnologia altera a estratégia da organização, consolidando uma cultura corporativa orientada a dados (*data-driven*). Ao apoiar as decisões logísticas em previsões matemáticas robustas em vez de intuições ou médias empíricas, o varejista otimiza todo o seu fluxo de suprimentos. Essa abordagem moderna permite reduzir drasticamente os custos operacionais com armazenagem desnecessária e, simultaneamente, mitigar as perdas financeiras causadas pela falta de produtos nas prateleiras virtuais.

1.6. Ferramentas e tecnologias de suporte

Para que a previsão de demanda saia do campo teórico e seja aplicada na prática, é indispensável estruturar um ambiente de desenvolvimento adequado. A base de dados do *e-commerce* possui milhares de registros diários, o que torna inviável qualquer tentativa de análise manual. Segundo Grus (2021), a resolução de problemas complexos na área de dados exige o

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

uso de tecnologias consistentes, para que o pesquisador consiga focar na análise dos resultados em vez de gastar tempo construindo sistemas básicos do zero. Com base nisso, o projeto definiu um conjunto de recursos tecnológicos específicos para processar o histórico de vendas e treinar os algoritmos.

1.6.1. Python

Neste estudo, a linguagem de programação Python foi escolhida como a base tecnológica. Segundo VanderPlas (2016), o Python consolidou-se nas últimas décadas como a linguagem padrão na comunidade de ciência de dados e inteligência artificial. Essa liderança ocorreu porque a linguagem combina uma sintaxe simples e legível, que facilita a escrita e a manutenção do código, com uma alta capacidade de processar grandes conjuntos de informações de forma eficiente.

Além de sua estrutura acessível, o Python conta com o suporte de uma comunidade global ativa, que desenvolve continuamente um vasto conjunto de bibliotecas complementares. Esses pacotes funcionam como blocos de código pré-construídos, viabilizando o uso de algoritmos preditivos avançados sem a necessidade de programar todas as regras estatísticas manualmente. Um exemplo fundamental no escopo deste trabalho é a biblioteca Scikit-Learn, que fornece os algoritmos de *Machine Learning* já desenvolvidos e validados, permitindo que o foco da pesquisa seja a aplicação prática dos modelos na gestão de estoque e a análise dos erros.

1.6.2. Pandas

Dentro da linguagem Python, o projeto utiliza a biblioteca Pandas como ferramenta central para lidar com as informações brutas do *e-commerce*. Antes que a inteligência artificial possa identificar qualquer padrão de compra com precisão, os dados de vendas precisam passar por um processo de preparação. Segundo McKinney (2018), criador da própria biblioteca, o Pandas foi projetado justamente para manipular estruturas de dados relacionais de forma rápida, utilizando formatos tabulares que funcionam de maneira semelhante a planilhas.

Na prática operacional do projeto, essa ferramenta é responsável por realizar a limpeza de registros inconsistentes, executar o cruzamento das tabelas de produtos e pedidos e organizar as transações financeiras em formato de séries temporais.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

1.6.3. Google Colab

O processamento de algoritmos preditivos, especialmente quando se trabalha com milhares de linhas de um histórico de vendas, costuma exigir bastante do computador. Muitas vezes, máquinas de uso pessoal não possuem memória RAM ou processadores suficientes para rodar esses testes sem apresentar lentidão ou travamentos. Para solucionar essa limitação física, o desenvolvimento prático do projeto foi realizado no Google Colab.

De acordo com Bisong (2019), o Colab é um ambiente de programação que funciona inteiramente na nuvem (*cloud computing*). A principal vantagem de adotar essa plataforma é a dispensa de configurações complexas. Toda a escrita do código em Python e o treinamento dos modelos ocorrem diretamente pelo navegador de internet, utilizando a infraestrutura de servidores fornecida gratuitamente pela própria Google.

1.7. Definição das métricas de avaliação

Para saber qual dos modelos apresentou o melhor resultado, escolheram-se três fórmulas matemáticas. Elas servem para comparar o volume real de vendas no dia i (representado por y_i) com o valor previsto pelo modelo (representado por $\hat{y}_i$), considerando a quantidade total de dias avaliados (n).

Para exemplificar o funcionamento das equações, considere, de maneira hipotética, que o *e-commerce* tenha vendido 12 unidades de um cosmético em um determinado dia, mas o modelo computacional tenha previsto uma saída de apenas 10 unidades; nesse caso, o sistema registra uma falha.

A primeira métrica a avaliar esse cenário é o Erro Médio Absoluto (MAE). Essa equação calcula a margem de erro direta em unidades físicas. No exemplo citado, o sistema extrai a diferença sem considerar sinais negativos, registrando uma falha simples de 2 produtos (|12 - 10| = 2). A equação abaixo processa essa conta para todos os dias do período de teste e extrai a média:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

A segunda métrica é a Raiz do Erro Quadrático Médio (RMSE). O objetivo dessa fórmula é evidenciar falhas críticas. No mesmo cenário hipotético, o erro inicial de 2 produtos seria elevado ao quadrado ($2^2 = 4$). Para a logística, olhar para esse resultado é essencial, pois errar

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

a previsão de forma brusca durante um pico de demanda gera um impacto operacional muito pior do que errar duas ou três unidades em dias normais.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Por fim, o estudo utiliza o Erro Percentual Absoluto Ponderado (WAPE). Enquanto o MAE e o RMSE demonstram a falha em unidades físicas de estoque, o WAPE traduz esse impacto para o fluxo de caixa. A fórmula abaixo processa a somatória de todos os desvios diários e a divide pelo volume total de cosméticos faturados no período, entregando a taxa percentual global de erro da operação logística.

$$WAPE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\sum_{i=1}^{n} y_i}$$

2. METODOLOGIA

2.1. Planejamento do projeto

O desenvolvimento prático desta pesquisa foi estruturado em quatro grandes etapas. O objetivo dessa divisão é garantir que os dados brutos de vendas sejam transformados corretamente em informações úteis para a inteligência artificial. O ciclo de vida do projeto, conforme ilustrado na Figura 1, seguirá a seguinte ordem:

1. Coleta e Filtragem dos Dados: A primeira etapa consiste em carregar as planilhas do *e-commerce* Olist para dentro do sistema. Como o escopo do trabalho é focado em um nicho específico, o código fará uma varredura para selecionar apenas as compras relacionadas à categoria de beleza e saúde, descartando o restante.
2. Preparação e Organização (Engenharia de Atributos): Agrupamento das vendas por dia para criar uma linha do tempo e extração de novas colunas indicativas de sazonalidade, como dias da semana e meses.
3. Treinamento dos Modelos: Aplicação dos dados estruturados em dois algoritmos distintos para fins de comparação: o modelo estatístico tradicional ARIMA (Séries Temporais) e o algoritmo de aprendizado de máquina *Random Forest* (Árvores de Decisão).

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

4. Avaliação de Resultados: Por fim, o sistema testará se a inteligência artificial realmente aprendeu de forma correta. As previsões geradas pelo algoritmo serão comparadas com o histórico real de vendas utilizando métricas matemáticas de erro (MAE, RMSE e WAPE), definindo assim qual modelo é o mais seguro para aplicação logística.

Figura 1. Fluxo das etapas de planejamento e desenvolvimento do projeto.



Fonte: Autoria própria, 2026.

2.2. Modelagem do diagrama de entidade-relacionamento (DER)

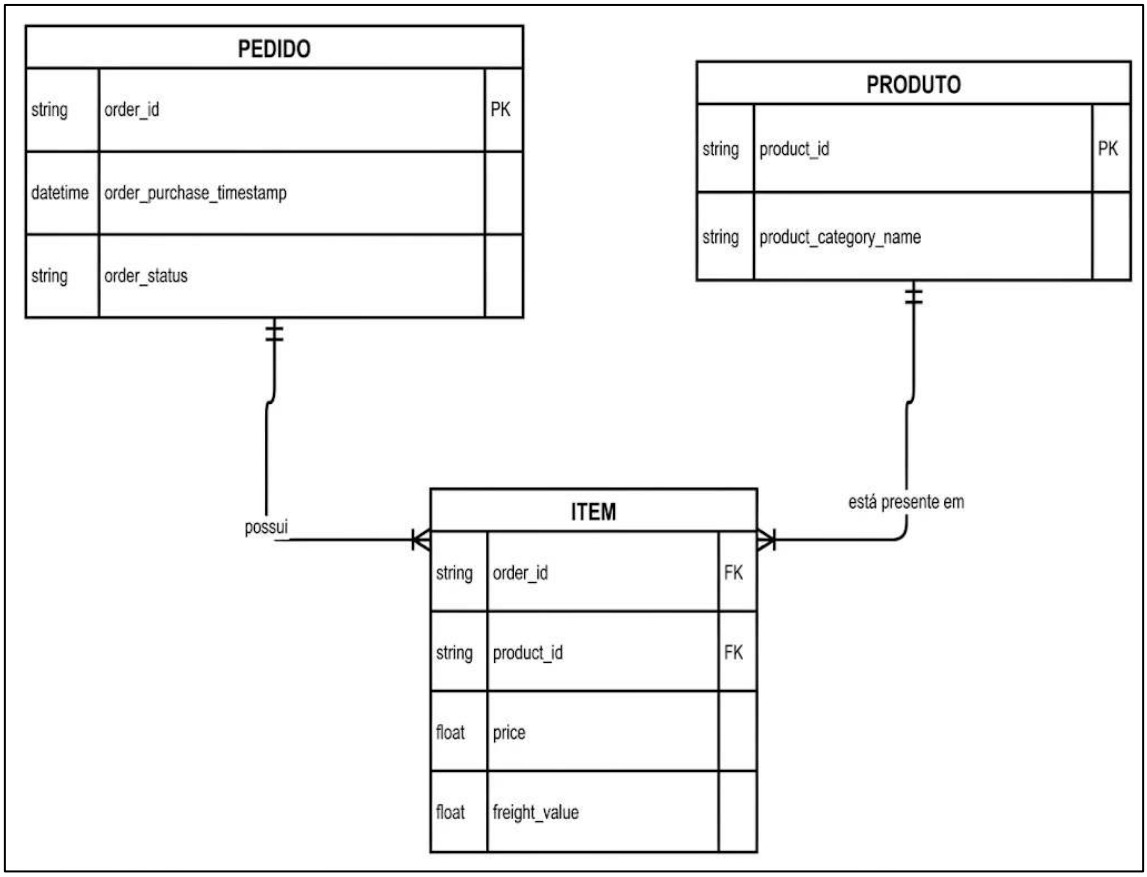
Antes de processar as informações com programação, é fundamental compreender como os dados de vendas do *e-commerce* estão armazenados e conectados. Para isso, utiliza-se o Diagrama de Entidade-Relacionamento (DER), que funciona como um mapa visual do banco de dados relacional (Elmasri; Navathe, 2011).

No escopo deste trabalho, o foco é a interação de três tabelas principais, conforme ilustrado na Figura 2. A tabela de PEDIDOS armazena o momento exato em que o cliente finalizou a compra. Ela se conecta com a tabela de ITENS, que guarda o preço e a quantidade de produtos vendidos, formando um relacionamento de "um para muitos" (pois um único pedido pode conter vários itens diferentes). Por fim, a tabela de ITENS se liga à tabela de PRODUTOS. É exatamente essa última conexão que permite ao sistema ler a descrição do produto e filtrar a base de dados, garantindo que o estudo analise apenas as mercadorias de saúde e beleza.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

Figura 2. Diagrama de Entidade-Relacionamento



Fonte: Autoria própria, 2026.

2.3. Modelagem do fluxo do sistema (UML)

Além de mapear a origem dos dados, também é necessário documentar a lógica de funcionamento do código que será escrito em Python. Para ilustrar o passo a passo que o algoritmo deverá seguir, o projeto adota o Diagrama de Atividades da UML (Fowler, 2005).

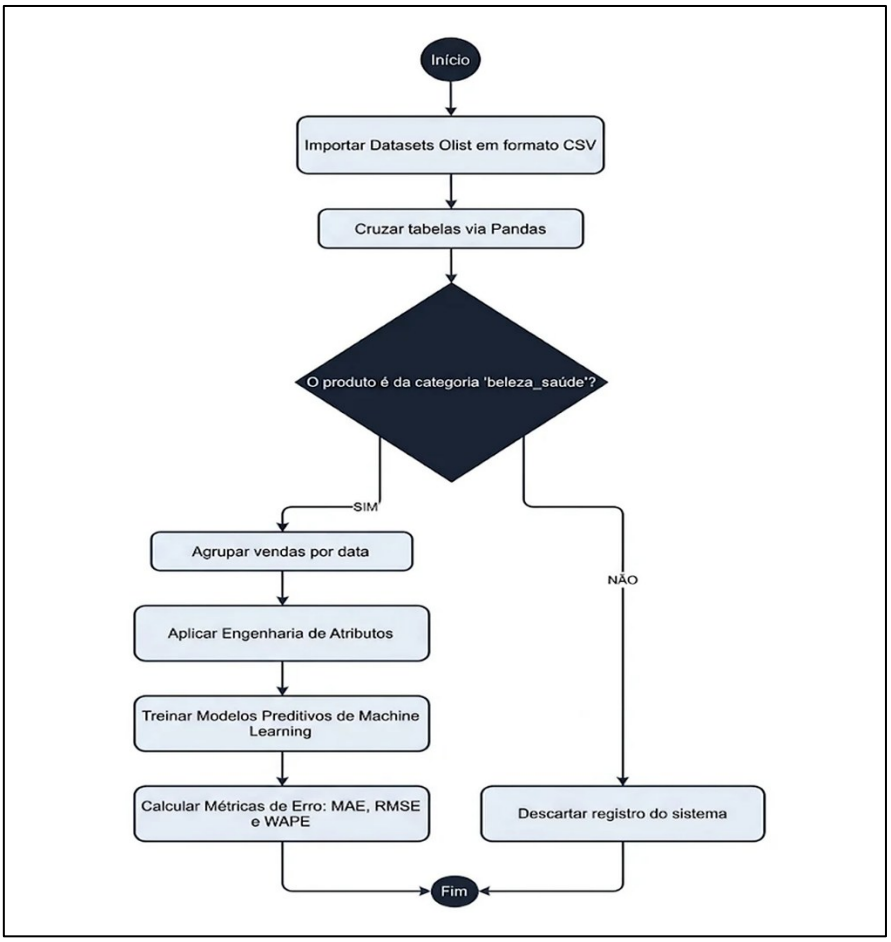
O fluxo de processamento, apresentado na Figura 3, começa com a leitura dos arquivos originais. Em seguida, o sistema realiza uma verificação condicional essencial: se o produto vendido pertencer ao segmento estipulado, a informação avança para a próxima etapa; caso contrário, ela é imediatamente descartada. Uma vez que os dados estão validados, o fluxo os encaminha para o agrupamento por datas e, posteriormente, para o treinamento dos modelos

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

preditivos. O processo computacional se encerra com a geração de gráficos e o cálculo das taxas de erro, que indicarão o nível de acerto do algoritmo na previsão de demanda.

Figura 3. Diagrama de atividades do fluxo de processamento em UML



Fonte: Autoria própria, 2026.

2.4. Pré-processamento e engenharia de atributos

A etapa inicial do desenvolvimento prático consistiu na importação das bibliotecas e no carregamento dos arquivos brutos para o ambiente de execução. Utilizando a biblioteca Pandas, os documentos contendo os registros de pedidos, itens e produtos foram lidos e armazenados em memória. O cruzamento dessas informações foi realizado por meio das chaves de identificação mapeadas na modelagem relacional. Esse processo devolveu os dados do *e-commerce* em 112.650 linhas.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

Com a base unificada, a etapa seguinte focou na limpeza e delimitação do escopo da pesquisa, aplicando-se um filtro condicional estrutural desenhado para reter estritamente as transações categorizadas como "beleza_saude". Essa operação reduziu a amostra para 9.670 registros relevantes. No mesmo bloco de execução, colunas irrelevantes foram descartadas e a data da compra foi ajustada para um formato adequado à análise, mantendo apenas o dia, o mês e o ano. Essa etapa facilitou a organização dos dados e permitiu realizar as análises estatísticas de forma mais consistente.

Como o objetivo do trabalho é estimar a necessidade futura de estoque, o formato de compras individuais dispersas inviabiliza a análise dos algoritmos preditivos. Dessa forma, esses registros são transformados em uma série temporal diária. A rotina de agrupamento contabilizou o volume total de itens cosméticos comercializados a cada dia, ordenando os resultados cronologicamente. No final da execução, é exibida a nova estrutura dos dados, apresentada na Figura 4, estabelecendo a base numérica que alimentará as fases subsequentes de treinamento computacional.

Figura 4. Agrupamento diário dos dados e formação da série temporal estruturados em tabela

	data_venda	quantidade_vendida
0	2016-09-15	3
1	2016-10-04	11
2	2016-10-05	4
3	2016-10-06	8
4	2016-10-07	7

Fonte: Autoria própria, 2026.

2.5. Divisão de dados e treinamento dos modelos

Na modelagem preditiva aplicada a séries temporais, a separação dos dados exige critérios específicos para garantir que a avaliação seja realista e respeite a ordem cronológica dos eventos, evitando o vazamento de informações futuras (*data leakage*). Após o agrupamento diário das transações da categoria de beleza e saúde, a série temporal consolidou um total de 602 dias com registros de vendas. Aplicou-se um corte temporal sequencial (*hold-out* cronológico) na proporção de 80% para treinamento e 20% para teste: os primeiros 481 dias foram destinados ao ajuste dos algoritmos, enquanto os 121 dias finais foram isolados exclusivamente para a validação preditiva.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

Para a modelagem supervisionada com o algoritmo *Random Forest Regressor*, definiram-se três variáveis preditoras (*features*) derivadas do calendário: um índice temporal contínuo (dia_sequencial) para capturar a tendência de longo prazo, o mês do ano (mes, variando de 1 a 12) e o dia da semana (dia_da_semana, codificado numericamente de 0 a 6, de segunda-feira a domingo), tendo como variável-alvo (y) a quantidade diária de itens vendidos. Na instanciação do modelo por meio da biblioteca *Scikit-Learn*, configuraram-se 100 árvores de decisão (n_estimators = 100) e fixou-se a semente aleatória em random_state = 42 para assegurar a reprodutibilidade determinística dos experimentos, mantendo-se os demais hiperparâmetros em seus valores padrão.

Em paralelo, para fins de comparação com a abordagem estatística clássica, ajustou-se um modelo univariado ARIMA (*AutoRegressive Integrated Moving Average*) utilizando a biblioteca *Statsmodels* diretamente sobre a série histórica de vendas diárias do conjunto de treino. O modelo foi parametrizado com ordem (p=5, d=1, q=0), incorporando cinco defasagens autorregressivas (p=5) para capturar a dependência recente de curto prazo, uma diferenciação de primeira ordem (d=1) para estabilizar a tendência da série e zero termos de média móvel (q=0). Após a etapa de treinamento (*fit*), ambos os modelos geraram projeções para o período de 121 dias do conjunto de teste para posterior avaliação comparativa.

3. RESULTADOS E DISCUSSÃO

3.1. Comportamento histórico da demanda

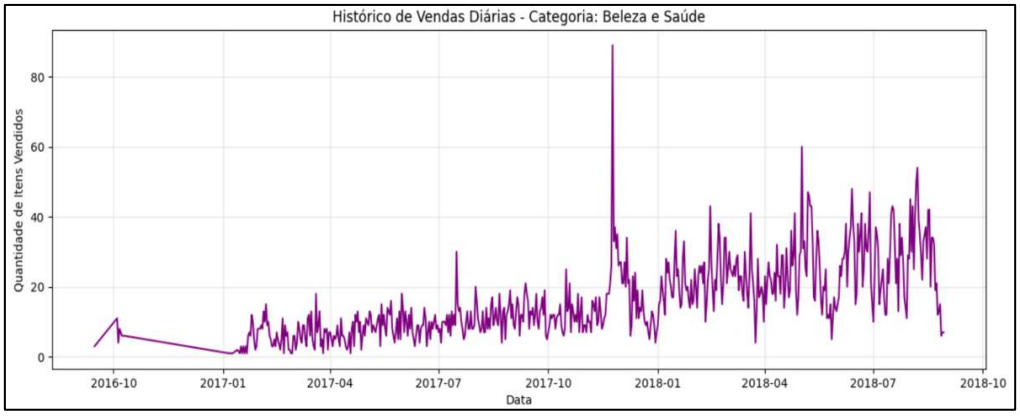
Antes de colocar a base de dados nos cálculos preditivos dos algoritmos, é necessário compreender a natureza das informações que alimentarão o sistema. Na análise de dados, a visualização exploratória constitui uma fase indispensável. Ela permite identificar rapidamente se o consumo da categoria é estável ou se possui anomalias e picos sazonais que exigirão um processamento computacional mais rigoroso.

Para realizar esse diagnóstico, o projeto utilizou a biblioteca *Matplotlib*, uma ferramenta do ecossistema Python dedicada à construção de gráficos estatísticos. A rotina de visualização permitiu plotar o histórico diário das transações da categoria de Beleza e Saúde desde os primeiros registros, no final de 2016, até meados de 2018, conforme apresentado na Figura 5.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

Figura 5. Histórico de vendas diárias da categoria de saúde e beleza



Fonte: Autoria própria, 2026.

A visualização do gráfico evidencia a alta complexidade do desafio logístico proposto. A linha de demanda da categoria não segue um padrão de crescimento contínuo e previsível, sendo marcada por uma volatilidade extrema no curto prazo. É possível notar oscilações bruscas no cotidiano e um pico acentuado de vendas no final de 2017, período que tradicionalmente coincide com datas promocionais agressivas do varejo eletrônico brasileiro, como a Black Friday.

Esse cenário não linear justifica tecnicamente a adoção de algoritmos de aprendizado de máquina. Se o controle de estoque dependesse apenas de projeções matemáticas simples ou do julgamento humano intuitivo, seria impossível prever variações tão agressivas. A incapacidade de acompanhar essas flutuações resultaria, inevitavelmente, na ruptura de estoque nos dias de pico ou na perda de mercadorias perecíveis nos períodos de baixa procura.

3.2. Avaliação quantitativa dos modelos

Após compreender o comportamento histórico da base de dados, os 20% finais da série temporal (121 dias) foram isolados para avaliar o desempenho preditivo dos algoritmos. O cálculo comparou as projeções geradas pelos modelos com o volume real de itens comercializados no período de teste, consolidando as métricas de erro MAE, RMSE e WAPE apresentadas na Tabela 1.

Tabela 1. Comparativo de desempenho preditivo

Modelo Preditivo	MAE (unidades)	RMSE	WAPE (%)

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

Série Temporal (ARIMA)	10,05	12,00	37,00%
Árvore de Decisão (*Random Forest*)	8,62	10,94	31,73%

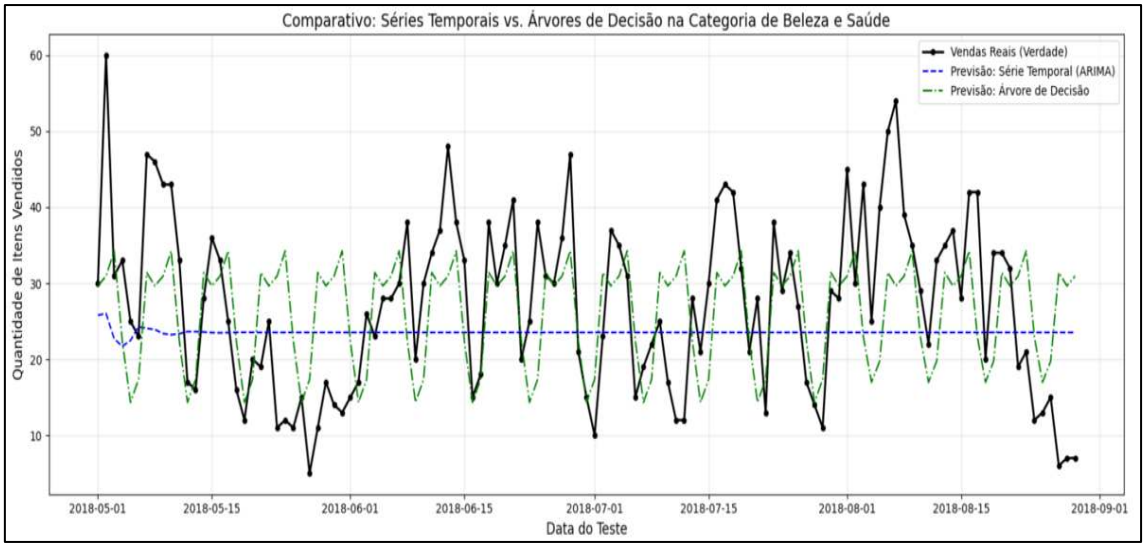
Fonte: Autoria própria, 2026.

A análise dos indicadores revela que o modelo de aprendizado de máquina (*Random Forest*) obteve menores índices de erro em todas as métricas avaliadas. No MAE, o algoritmo apresentou menor desvio médio diário em unidades em comparação ao modelo ARIMA, comportamento acompanhado pelo RMSE nos dias de maior oscilação. Por fim, o WAPE indicou uma redução de 5,27 pontos percentuais no erro relativo acumulado no conjunto de teste, sugerindo maior aderência preditiva para apoiar o planejamento de reposição.

3.3. Análise visual e impacto logístico

Para compreender a diferença estatística apontada na tabela, elaborou-se uma representação gráfica do período de testes. A Figura 6 sobrepõe as estimativas computacionais geradas pelos dois algoritmos à linha real de vendas do *e-commerce*.

Figura 6. Comparativo visual das previsões contra as vendas reais



Fonte: Autoria própria, 2026.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

O gráfico ilustra claramente a limitação técnica da abordagem estatística tradicional no varejo. A linha azul tracejada, referente ao modelo ARIMA, mantém-se praticamente plana. O sistema calculou apenas uma média contínua, ignorando completamente as oscilações diárias. Em contrapartida, a linha verde do *Random Forest* consegue acompanhar a volatilidade do mercado, curvando-se para acompanhar os picos de alta demanda e retraindo-se na baixa procura.

No contexto operacional da categoria de Beleza e Saúde, essa maior capacidade de acompanhar variações de demanda fornece subsídios mais precisos para o dimensionamento de pedidos. Embora este estudo não tenha mensurado diretamente os custos financeiros de armazenagem ou perdas por vencimento, previsões mais aderentes tendem a contribuir para políticas de estoque de segurança mais equilibradas.

4. CONSIDERAÇÕES FINAIS

O crescimento acelerado do comércio eletrônico exige operações logísticas cada vez mais eficientes. Este estudo cumpriu o seu objetivo principal ao desenvolver e aplicar modelos preditivos de *Machine Learning* para otimizar o gerenciamento de estoque da categoria de Beleza e Saúde, um setor desafiador devido à forte volatilidade de consumo e à perecibilidade dos itens.

Os resultados práticos extraídos da base de dados confirmaram a hipótese inicial da pesquisa. Constatou-se que o uso de algoritmos de aprendizado supervisionado, especificamente o *Random Forest*, apresentou menor erro preditivo em comparação ao modelo clássico ARIMA na amostra analisada. O algoritmo demonstrou capacidade para capturar variações temporais da demanda com maior aderência do que a abordagem autorregressiva linear utilizada como referência.

Conclui-se que o uso de modelos preditivos orientados a dados (data-driven) constitui uma ferramenta promissora de apoio à decisão no planejamento de compras, com potencial para auxiliar na redução de rupturas e excessos de estoque. Como trabalhos futuros, recomenda-se a inclusão de variáveis exógenas (como preço, frete e campanhas promocionais) e a simulação financeira direta dos custos de armazenagem e ruptura.

REFERÊNCIAS

BALLOU, Ronald H. Gerenciamento da cadeia de suprimentos/logística empresarial. 5. ed. Porto Alegre: Bookman, 2006.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

BISONG, Ekaba. Building Machine Learning and Deep Learning Models on Google Cloud Platform: a comprehensive guide for beginners. New York: Apress, 2019.

CHOPRA, Sunil; MEINDL, Peter. Gestão da cadeia de suprimentos: estratégia, planejamento e operações. 6. ed. São Paulo: Pearson Education do Brasil, 2016.

ELMASRI, Ramez; NAVATHE, Shamkant B. Sistemas de banco de dados. 6. ed. São Paulo: Pearson, 2011.

FACELI, Katti *et al*. Inteligência artificial: uma abordagem de aprendizado de máquina. 2. ed. Rio de Janeiro: LTC, 2021.

FOWLER, Martin. UML destilado: um breve guia para a linguagem padrão de modelagem de objetos. 3. ed. Porto Alegre: Bookman, 2005.

GÉRON, Aurélien. Mãos à obra: aprendizado de máquina com Scikit-Learn, Keras & TensorFlow. 2. ed. Rio de Janeiro: Alta Books, 2021.

GRUS, Joel. Data science do zero: primeiras regras com o Python. 2. ed. Rio de Janeiro: Alta Books, 2021.

HYNDMAN, Rob J.; ATHANASOPOULOS, George. Forecasting: principles and practice. 3. ed. Melbourne: OTexts, 2021. Disponível em: https://otexts.com/fpp3/. Acesso em: 20 maio 2026.

MARTINS, Emerson. Uma aplicação de machine learning na previsão de vendas no segmento de varejo. 2023. 168 f. Dissertação (Mestrado Profissional em Gestão e Tecnologia em Sistemas Produtivos) – Centro Estadual de Educação Tecnológica Paula Souza, São Paulo, 2023. Disponível em: https://ric.cps.sp.gov.br/handle/123456789/26730. Acesso em: 7 abr. 2026.

MCKINNEY, Wes. Python para análise de dados: tratamento de dados com Pandas, NumPy e IPython. São Paulo: Novatec, 2018.

MICHEL, Júlia de Souza Silva. Previsão de demanda no mercado online brasileiro: estudo comparativo entre dois modelos lineares. 2022. Dissertação (Mestrado em Gestão de Organizações, Liderança e Decisão) – Universidade Federal do Paraná, Curitiba, 2022. Disponível em: https://acervodigital.ufpr.br/xmlui/handle/1884/79637. Acesso em: 13 maio 2026.

MITCHELL, Tom M. Machine learning. New York: McGraw-Hill, 1997.

MORETTIN, Pedro A.; TOLOI, Clélia M. C. Análise de séries temporais: modelos lineares univariados. 3. ed. São Paulo: Blucher, 2018.

PROVOST, Foster; FAWCETT, Tom. Data science para negócios: o que você precisa saber sobre mineração de dados e pensamento analítico de dados. Rio de Janeiro: Alta Books, 2016.

SHIKUMA, Diogo Tsuyoshi; LEME, Luiz Gustavo Cardoso; LOPES, Silvio Cesar. Estoqueprev: sistema preditivo de demanda com aprendizado de máquina. 2025. Trabalho de Conclusão de Curso (Tecnologia) – Faculdade de Tecnologia, Centro Estadual de Educação Tecnológica Paula Souza, São Paulo, 2025. Disponível em: https://ric.cps.sp.gov.br/handle/123456789/39857. Acesso em: 10 abr. 2026.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

SLACK, Nigel; BRANDON-JONES, Alistair; JOHNSTON, Robert. Administração da produção. 8. ed. São Paulo: Atlas, 2018.

TAMACHIRO, Thiago Shoji Obi *et al*. Modelos de aprendizado de máquina para previsão da demanda por séries temporais em cadeias de suprimentos: uma revisão sistemática e análise bibliométrica da literatura. *In*: CONGRESSO BRASILEIRO DE ENGENHARIA DE PRODUÇÃO (CONBREPRO), 14., 2024, Ponta Grossa. Anais [...]. Ponta Grossa: APREPRO, 2024. Disponível em: https://aprepro.org.br/conbrepro/anais/2024/arquivos/10282024_161017_671fdf41e3594.pdf. Acesso em: 5 mar. 2026.

VANDERPLAS, Jake. Python data science handbook: essential tools for working with data. Sebastopol: O'Reilly Media, 2016.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.