

Universidade do Minho
Escola de Economia, Gestão e Ciência Política

Bruna Silva

Previsão da Procura com *Machine Learning* para um *e-Commerce* mais sustentável

junho 2026

Previsão da Procura com *Machine Learning* para um *e-Commerce* mais sustentável.
UMinho | 2026 Bruna Silva



Universidade do Minho
Escola de Economia, Gestão e Ciência Política

Bruna Silva

Previsão da Procura com *Machine Learning* para um *e-Commerce* mais sustentável

Dissertação de Mestrado
Mestrado em Gestão
Especialização em Logística

Trabalho efetuado sob a orientação do
Professor Doutor José António de Almeida Crispim

junho 2026

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.

Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

Licença concedida aos utilizadores deste trabalho



Atribuição-NãoComercial-SemDerivações

CC BY-NC-ND

https://creativecommons.org/licenses/by-nc-nd/4.0/

Agradecimentos

Em primeiro lugar, gostaria de expressar os meus agradecimentos ao Professor Doutor José António Almeida Crispim pelo apoio, atenção meticulosa aos detalhes e revisão rigorosa ao longo de todo o desenvolvimento deste trabalho.

Um agradecimento especial à minha família e ao meu namorado. Obrigada por toda a paciência, por estarem sempre lá quando eu preciso e por acreditarem em mim.

Declaração de integridade

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

DECLARAÇÃO DE USO DE FERRAMENTAS DE INTELIGÊNCIA ARTIFICIAL (IA)

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri a qualquer forma de utilização indevida de ferramentas de Inteligência Artificial (IA), nem à falsificação ou omissão de informações, fontes, dados, resultados ou processos em nenhuma das etapas conducentes à sua elaboração.

Declaro ainda que qualquer utilização de ferramentas de IA foi feita de forma transparente e responsável, encontrando-se devidamente assinalada/descrita no trabalho, e que procedi à verificação crítica da exatidão do conteúdo produzido com o seu apoio, assumindo integralmente a responsabilidade pelo que submeto.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho, bem como as normas aplicáveis de integridade académica e de proteção de dados.

Previsão da Procura com *Machine Learning* para um *e-Commerce* mais sustentável

Resumo

A volatilidade do mercado de retalho, acentuada por mudanças recentes no comportamento do consumidor e instabilidades globais, exige uma gestão de inventário cada vez mais precisa. Esta dissertação propõe uma abordagem integrada de Ciência de Dados para a previsão da procura e otimização logística, utilizando o *dataset* "Online Retail II". A metodologia seguiu um fluxo rigoroso de preparação de dados, resultando na criação de dois *sub-datasets* estratégicos: um para a previsão de vendas e outro para a análise de sustentabilidade. Foram implementados e comparados modelos estatísticos e de aprendizagem automática, com o *Extreme Gradient Boost* (XGBoost) a demonstrar uma performance superior através do ajuste seletivo de hiperparâmetros, permitindo um equilíbrio entre a complexidade do modelo e a precisão da previsão. Os resultados das previsões foram integrados num modelo de simulação no software AnyLogic, permitindo avaliar o impacto das decisões de inventário (ponto de encomenda) nos custos operacionais e na pegada de carbono (CO_2). O estudo conclui que a integração de modelos de Inteligência Artificial (IA) com ferramentas de simulação permite não só reduzir ruturas de *stock*, mas também quantificar os *trade-offs* entre eficiência financeira e sustentabilidade ambiental, oferecendo uma vantagem competitiva no setor do *e-Commerce*.

Palavras-chave: Previsão da Procura, Inteligência Artificial, Machine Learning, *e-Commerce*, Simulação, Sustentabilidade.

Demand Forecasting with Machine Learning for a more Sustainable E-commerce.

Abstract

The volatility of the retail market, heightened by recent shifts in consumer behavior and global instabilities, demands increasingly precise inventory management. This dissertation proposes an integrated Data Science approach for demand forecasting and logistics optimization, utilizing the 'Online Retail II' dataset. The methodology followed a rigorous data preparation pipeline, resulting in the creation of two strategic sub-datasets: one for sales forecasting and another for sustainability analysis. Statistical and machine learning models were implemented and compared, with Extreme Gradient Boosting (XGBoost) demonstrating superior performance through selective hyperparameter tuning, enabling a balance between model complexity and predictive accuracy. The forecasting results were integrated into a simulation model using AnyLogic software, allowing for the assessment of inventory decisions (reorder point) on operational costs and carbon footprint (CO_2). The study concludes that the integration of Artificial Intelligence (AI) models with simulation tools not only reduces stockouts but also quantifies the trade-offs between financial efficiency and environmental sustainability, providing a competitive advantage in the e-commerce sector.

Keywords: Demand Forecasting, Artificial Intelligence, Machine Learning, E-commerce, Simulation, Sustainability.

ÍNDICE

Agradecimentos ii
Declaração de integridade iii
Resumo iv
Abstract v
Índice de tabelas viii
Índice de figuras viii
Lista de abreviaturas e siglas x

1. INTRODUÇÃO 11

1.1. Enquadramento e justificação do tema 11
1.2. Objetivos da Investigação 12
1.3. Estrutura da dissertação 12

2. REVISÃO DE LITERATURA 14

2.1. Enquadramento teórico da previsão da procura 14
2.2. Inteligência Artificial e *Machine Learning* no *e-Commerce* 15
2.2.1. Aplicações e benefícios 16
2.2.2. Desafios e Limitações 16
2.3. Modelos Aplicados à Previsão da Procura 17
2.3.1. Modelos De Previsão Estatísticos 18
2.3.2. Modelos de *Machine Learning* 20
2.3.3. Modelos de *Deep Learning* 21
2.3.4. Métricas de Avaliação de Desempenho 24
2.4. Sustentabilidade e logística no *e-Commerce* 25
2.4.1. Dimensões da sustentabilidade (económica, ambiental e social) 26
2.4.2. O papel da previsão na sustentabilidade operacional 26
2.5. Simulação e *Machine Learning* 27
2.5.1. Simulação de eventos discretos e baseada em agentes 27
2.5.2. Modelos híbridos preditivo-simulativos 28
2.6. Falhas Identificadas e oportunidades de investigação 28
2.7. Síntese de Literatura e Taxonomia de Investigação 29

3. METODOLOGIA 31

3.1. Desenho Metodológico 31
3.2. Estratégias de Investigação e Etapas de Estudo 32

3.3. Base de Dados e Preparação 33
3.3.1. Origem e Natureza dos Dados 33
3.3.2. Descrição das Variáveis e Indicadores 33
3.3.3. Protocolo de Limpeza e Tratamento de Discrepâncias 34
3.3.4. Caracterização Estatística da Amostra 39
3.4. Modelação Preditiva 42
3.4.1. Engenharia de Atributos (*Feature Engineering*) 43
3.4.2. Arquiteturas e Parametrização dos Modelos 44
3.4.3. Métricas de Avaliação de Desempenho 46
3.5. Simulação e Dinâmica Operacional 47
3.5.1. Arquitetura e Lógica do Modelo de Simulação 47
3.5.2. Integração e Fluxo de Dados 48
3.5.3. Dinâmica de Gestão de Inventários e Cenários de Decisão 48
3.5.4. Modelo de Sustentabilidade e Cálculo de Emissões ($\boldsymbol{CO2}$) 49
3.5.5. Parametrização da Simulação 49
3.6. Ferramentas e tecnologias utilizadas 53
4. ANÁLISE DE DADOS 55
4.1. Desempenho dos Modelos de Previsão 55
4.1.1. Resultados Quantitativos 55
4.2. Interpretação e Explicabilidade dos Modelos 60
4.2.1. Importância das Variáveis 60
4.3. Simulação Operacional 62
4.3.1. Configuração dos Cenários Simulados 63
4.3.2. Resultados Operacionais 63
4.4. Avaliação do Impacto Sustentável 64
4.4.1. Redução Estimada de Emissões de $\boldsymbol{CO2}$ 65
5. DISCUSSÃO DOS RESULTADOS 66
6. CONTRIBUTOS TEÓRICOS E PRÁTICOS DA INVESTIGAÇÃO 66
7. LIMITAÇÕES E SUGESTÕES PARA INVESTIGAÇÃO FUTURA 67
REFERÊNCIAS BIBLIOGRÁFICAS 68
Apêndice 72
Índice de equações

Equação 1 - Erro Quadrático Médio 24
Equação 2 - Raiz do Erro Quadrático Médio..... 24
Equação 3 - Erro Absoluto Médio..... 25
Equação 4 - Erro Médio Absoluto Percentual 25

Índice de tabelas

Tabela 1 - Conclusões e Falhas Identificadas na Literatura 28
Tabela 2 - Etapas do Tratamento dos Dados 35
Tabela 3 - Separação final dos dados em dois datasets..... 38
Tabela 4 - Métricas das variáveis, Quantidade e Preço Unitário..... 39
Tabela 5 - Resumo da Parametrização dos Modelos Preditivos 46
Tabela 6 - Notação matemática e variáveis do modelo de simulação..... 50
Tabela 7 - Parâmetros e Pressupostos Matemáticos do modelo de simulação..... 51
Tabela 8 - Métricas de Desempenho dos Modelos Preditivos 55
Tabela 9 - Comparação de custos operacionais entre os cenários simulados. Fonte: Elaboração própria com base nos resultados da simulação. 63
Tabela 10 - Valores de kg de CO2 na simulação operacional. Fonte: Elaboração Própria com base nos resultados da simulação..... 65
Tabela 11 - Tabela de Extração de Dados. Fonte: Elaboração Própria. 72

Índice de figuras

Figura 1 - Mecanismo interno de uma unidade LSTM..... 22
Figura 2 - Arquitetura de um modelo híbrido (CNN-LSTM). 23
Figura 3 - Taxonomia da Investigação 30
Figura 4 - Fluxograma da arquitetura metodológica..... 32
Figura 5 - Relatório de Discrepâncias. 37
Figura 6 - Decomposição da série temporal das vendas reais. 40
Figura 7 - Autocorrelação e Autocorrelação Parcial..... 41
Figura 8 - Diagrama lógico da simulação de gestão de inventário no AnyLogic. 48
Figura 9 - Esquema conceptual e integração das ferramentas e tecnologias utilizadas. 54

Figura 10 - Comparação entre a procura real e a previsão do modelo ARIMA(1,1,1), evidenciando a incapacidade de captura de picos de vendas. 57
Figura 11 - Comparação gráfica entre a procura real e as previsões do modelo XGBoost durante o período de teste. 59
Figura 12 - Gráfico de importância das variáveis (Feature Importance), destacando os preditores temporais e históricos com maior impacto nas previsões de procura do modelo XGBoost. Nota: A variável 'trimestre' foi automaticamente omitida pelo algoritmo por apresentar um score de importância nulo. 60
Figura 13 - Gráfico de resumo SHAP indicando o impacto e a direção das variáveis na previsão do modelo XGBoost. Nota: Ao contrário da Figura 12, a variável 'trimestre' é aqui incluída para fins de exaustividade metodológica, demonstrando graficamente uma distribuição de impacto estritamente nula. 62
Figura 14 - Evolução de stock do armazém de IA. 64
Figura 15 - Evolução de stock do armazém do cenário tradicional. 64
Figura 16 - Esquema PRISMA. 89
Figura 17 - Comparação gráfica entre a procura real e as previsões do modelo estatístico SARIMA no período de teste. 97
Figura 18 - Comparação gráfica entre a procura real e as previsões do modelo estatístico SARIMA Automatizado no período de teste. 98
Figura 19 - Comparação gráfica entre a procura real e as previsões do modelo LSTM (Deep Learning) no período de teste. 99

Lista de abreviaturas e siglas

ABS – *Agent-Based Simulation*

ACF – *Autocorrelation Function*

ADF – *Augmented Dickey-Fuller*

API – *Application Programming Interface*

ARIMA – *Autoregressive Integrated Moving Average*

CNN – *Convolutional Neural Network*

CO_2 – Dióxido de Carbono

DES – *Discrete Event Simulation*

DL – *Deep Learning*

GBM – *Gradient Boosting Machine*

IA – Inteligência Artificial

LLM – *Large Language Model*

LSTM – *Long Short-Term Memory*

M2TFM – *Multi-modal Transform-based Fusion Model*

MAE – *Mean Absolute Error*

MAPE – *Mean Absolute Percentage Error*

ML – *Machine Learning*

MSE – *Mean Squared Error*

PACF – *Partial Autocorrelation Function*

RMSE – *Root Mean Square Error*

ROP – *Reorder Point*

SARIMA – *Seasonal Autoregressive Integrated Moving Average*

STGNN – *Spatio-Temporal Graph Neural Network*

XGBoost – *Extreme Gradient Boosting*

1. INTRODUÇÃO

1.1. Enquadramento e justificação do tema

O presente estudo responde a desafios reais relacionados com a previsão da procura e a sustentabilidade operacional. Existe um crescimento exponencial do *e-Commerce* que transforma o paradigma das cadeias de abastecimento globais. Ao contrário do retalho físico tradicional, o ambiente digital é extremamente volátil, utiliza campanhas promocionais agressivas e tem expectativas de entrega cada vez mais exigentes por parte dos consumidores. Desta forma, a gestão de inventário torna-se um núcleo crítico da viabilidade financeira e operacional de qualquer empresa do setor.

A justificação para o presente estudo nasce de um dilema logístico, onde os modelos estatísticos tradicionais tornaram-se incapazes de antecipar com precisão os picos e quedas abruptas da procura. Devido a esta falha preditiva as empresas ficam perante dois cenários altamente prejudiciais. Por um lado, a rutura de *stock* resulta em vendas perdidas, penalizações e perda da confiança do cliente. Por outro lado, o excesso de *stock* preventivo implica um aumento dos custos de armazenagem. Para além disso, a falha neste planeamento alimenta um problema ambiental muitas vezes oculto, o aumento da logística inversa e dos envios fragmentados.

Apesar da literatura recente demonstrar um interesse crescente na aplicação de Inteligência Artificial (IA) e do *Machine Learning* (ML) para otimizar as previsões, grande parte dos estudos foca-se na minimização do erro matemático e coloca de parte a simulação do impacto dessas previsões num ambiente operacional real. No entanto, existem alguns estudos que consideram esta ligação. Por exemplo, segundo Cavalcante et al. (2019), a integração entre simulação e ML permite decisões mais resilientes e fundamentadas em dados, o que é essencial em ambientes de incerteza (Cavalcante et al., 2019).

Além disso, Badakhshan et al. (2024) destacam que a combinação de simulação com ML está ainda numa fase inicial, sobretudo em domínios como a sustentabilidade, o que representa uma oportunidade de inovação metodológica e prática (Badakhshan et al., 2024).

Assim, este trabalho contribui para um campo emergente que alia a IA a decisões sustentáveis no *e-Commerce*, com impacto potencial na redução de desperdícios, na otimização de recursos e na melhoria do serviço ao cliente. Para além da implementação e comparação de modelos de previsão da procura, também foi quantificado o seu impacto real financeiro e de sustentabilidade carbónica através da simulação digital.

1.2. Objetivos da Investigação

Esta dissertação tem como objetivo central responder a seguinte questão de investigação: Como podem as previsões de procura geradas por modelos de ML apoiar decisões práticas em contextos reais?

Consiste em avaliar a eficácia da transição de modelos de previsão tradicionais para algoritmos de ML e *Deep Learning* (DL) na gestão de inventário de um *e-Commerce*, medindo o seu impacto direto nos custos operacionais e nas emissões de CO_2.

De forma a atingir este propósito, definiram-se objetivos específicos. O primeiro objetivo passa por realizar uma análise exploratória e tratamento de dados históricos de vendas e sustentabilidade de uma operação real de retalho *online*. Posteriormente, desenvolver, treinar e comparar modelos preditivos tradicionais (como o *Autorregressive Integrated Moving Average model* (ARIMA)) com algoritmos avançados (como o *Extreme Gradient Boosting* (XGBoost)), identificando o modelo com maior precisão e capacidade de adaptação à volatilidade da procura. O objetivo seguinte passa por construir um modelo de simulação operacional no *software* Anylogic para testar as diferentes políticas de inventário e quantificar a poupança financeira gerada pela antecipação de ruturas de *stock*. O objetivo final passa por avaliar o *trade-off* ambiental, medindo o impacto das decisões de reabastecimento nas emissões diretas de carbono (CO_2).

1.3. Estrutura da dissertação

O presente documento encontra-se organizado em sete capítulos. O Capítulo 1 introduz o tema, contextualizando a problemática do *e-Commerce* e definindo os objetivos da investigação. O Capítulo 2 apresenta a revisão da literatura, explorando o estado da arte das metodologias de previsão da procura e a sua relação com a sustentabilidade na cadeia de abastecimento. O Capítulo 3 detalha a metodologia adotada, descrevendo o processo de extração e tratamento de dados, bem como o desenvolvimento e treino dos modelos preditivos em ambiente Python. Adicionalmente, demonstra o processo utilizado na simulação operacional. O Capítulo 4 foca-se na apresentação dos resultados, ilustrando a superioridade do modelo XGBoost e detalhando os *outputs* da simulação operacional no AnyLogic, com ênfase na análise de custos e emissões de CO_2. O Capítulo 5 promove a discussão dos resultados obtidos, validando as hipóteses da investigação face ao comportamento do simulador. O Capítulo 6 expõe os principais contributos teóricos e práticos gerados por este estudo para a academia e para a indústria. Por fim, o Capítulo

7 aponta as limitações encontradas ao longo da investigação e sugere linhas de desenvolvimento para trabalhos futuros.

2. REVISÃO DE LITERATURA

Nos últimos anos, a aplicação de técnicas de ML para a previsão da procura no contexto do *e-Commerce* tem vindo a ganhar destaque na literatura recente, com uma atenção crescente à sustentabilidade. Esta revisão apresenta uma síntese do tema, identificando as principais contribuições, modelos e metodologias utilizados, bem como as tendências emergentes que relacionam a previsão da procura, a IA e a sustentabilidade logística.

A pesquisa bibliográfica foi conduzida com base em artigos publicados entre 2023 e 2025 nas bases de dados Scopus e ScienceDirect, utilizando os termos de pesquisa "demand forecast", "e-Commerce", "machine learning" e "sustainable logistics", de acordo com as diretrizes do método PRISMA, ver Tabela 11 e Figura 16, Apêndice A. As fontes analisadas incluem artigos científicos e de revisão, com o intuito de compreender as contribuições mais recentes e de identificar as principais falhas relevantes para investigação futura.

2.1. Enquadramento teórico da previsão da procura

A importância da previsão da procura no retalho foi recentemente elevada por fatores como a instabilidade pós-pandemia e as mudanças drásticas no comportamento do consumidor, que geram flutuações anormais nas vendas (Vhatkar et al., 2025). Conforme sublinhado na literatura mais recente, a incapacidade de prever estas variações resulta em custos operacionais elevados e ineficiências na cadeia de abastecimento. Assim, a aplicação de ferramentas de IA não é apenas uma melhoria técnica, mas uma necessidade estratégica para garantir a resiliência operacional e a sustentabilidade das empresas num mercado global volátil.

A previsão é sobre prever o futuro com a maior precisão possível, tendo em conta toda a informação disponível, incluindo dados históricos e o conhecimento de quaisquer eventos futuros que possam impactar a previsão (Hyndman & Athanasopoulos, 2018).

A previsão da procura é um elemento crucial na gestão da cadeia de abastecimento, permitindo alinhar as decisões de produção, inventário e distribuição com as necessidades reais do mercado. No entanto, a cadeia de abastecimento tornou-se cada vez mais complexa e incerta devido à expansão da produção em massa. Atualmente, os métodos predominantes de previsão da procura da cadeia de abastecimento baseiam-se principalmente em modelos estatísticos tradicionais ou modelos de escala temporal única, resultando num desempenho de previsão inadequado. Atualmente, existem dois tipos principais de abordagens para problemas de previsão

de vendas ou de procura: os métodos tradicionais de previsão de séries temporais e os modelos baseados em ML (Chen et al., 2025).

Os sistemas de IA usam várias técnicas, incluindo ML para analisar grandes quantidades de dados e fazer previsões ou decisões com base nessa análise (Malik et al., 2024).

O ML, enquanto subcampo da IA, utiliza ferramentas e técnicas críticas para recolher e analisar grandes volumes de dados, implementa modelos preditivos para estimar a procura de produtos e serviços e permite o desenvolvimento de recursos eficientes para que as empresas respondam eficazmente às crescentes necessidades dos clientes. Ao aplicar o ML, as empresas tendem a ser capazes de prever a procura com base em diversas variáveis para melhorar a interação com o cliente, o comportamento de compra e as vendas cruzadas (Leung, 2025).

2.2. Inteligência Artificial e *Machine Learning* no *e-Commerce*

Numa era pré-digital, a maioria das empresas operava através de estruturas físicas, vendendo produtos a partir de um local fixo, ou entregando-os através de serviços postais. Com o avanço da tecnologia e de forma a alcançar um maior número de clientes em diversas partes do mundo, as empresas passaram a adotar ferramentas digitais para a comercialização dos seus produtos e serviços *online* (Shaikh et al., 2022).

O setor do *e-Commerce* envolve a compra e venda de bens e serviços através da internet, abrangendo uma vasta gama de atividades, incluindo retalho *online*, *marketplaces* digitais e pagamentos eletrónicos. A IA revolucionou este setor ao permitir recomendações personalizadas, análises preditivas e estratégias de marketing direcionadas. Estas tecnologias potenciam a experiência de compra dos consumidores, aumentam as vendas das empresas e otimizam os processos de gestão da cadeia de abastecimento (Malik et al., 2024).

Através do uso de técnicas de ML, as empresas são agora capazes de prever a procura com base em várias variáveis, o que lhes permite melhorar a sua interação com o cliente, compreender os seus hábitos de compra e analisar o controlo do comportamento do consumidor. Atualmente, os consumidores recorrem a diversos dispositivos para adquirir produtos e serviços *online* e, por conseguinte, as empresas utilizam o *e-Commerce* como forma de beneficiar das oportunidades de mercado. Posto isto, é essencial que as empresas antecipem a procura de produtos novos e existentes, analisem os hábitos de compra dos clientes e reforcem o seu envolvimento com estes, de modo a promover a lealdade à marca (Shaikh et al., 2022).

2.2.1. Aplicações e benefícios

O serviço de IA proporciona à cadeia de abastecimento da plataforma de *e-Commerce* uma nova oportunidade de adotar mais efetivamente o modo de revenda, o que aumenta potencialmente a lucratividade e a eficiência na cadeia de abastecimento (He et al., 2024).

O uso de novas e avançadas técnicas de ML oferece às organizações mais benefícios, como o suporte na previsão precisa da procura do cliente, o reforço do envolvimento do consumidor e a venda cruzada de produtos ou serviços. Para além disso, uma análise orientada através do ML dá suporte na análise do comportamento do cliente e rastreia o produto, desde o momento em que está nos fornecedores, até chegarem às mãos do consumidor final. No setor do *e-Commerce* estas novas técnicas de ML são aplicadas, de forma geral, para analisar o comportamento do cliente e envolvê-lo de forma eficaz para oferecer produtos e serviços que atendam as suas necessidades (Shaikh et al., 2022).

A relação entre as abordagens de ML e a melhoria do setor *e-Commerce* é predominantemente positiva, isto é, as empresas que incorporam técnicas de ML têm uma probabilidade significativamente maior de crescimento e de aumentar a sua produtividade e rendimento. Diversas abordagens do ML aplicadas à previsão da procura, à análise do comportamento do consumidor em relação às compras e à venda cruzada de produtos têm contribuído para o fortalecimento deste setor. Ao antecipar a procura de diversos produtos e serviços, avaliar os hábitos de compra do cliente e apoiar a adaptação de produtos, há uma contribuição para um crescimento e desenvolvimento sustentável (Shaikh et al., 2022).

2.2.2. Desafios e Limitações

A utilização de abordagens de ML oferece, de facto, benefícios às empresas, porém, existem alguns desafios e limitações que surgem com o uso destas técnicas.

Por exemplo, *algorithmic bias*, ou seja, o ato de algoritmos tomarem atitudes discriminatórias em relação a grupos específicos de seres humanos, especialmente quanto a minorias, é um dos maiores desafios relacionados com o uso de ML. Para além disso, existem as questões de privacidade de dados que surgem com a utilização da IA no *e-Commerce*. Apesar destes desafios, a IA contribui para o progresso social, criando oportunidades de emprego, promovendo o crescimento económico e ampliando o acesso a bens e serviços (Malik et al., 2024).

Alguns estudos destacam desafios na coexistência entre serviços manuais e automatizados. He et al. (2024), introduzem o conceito de atenuação da eficiência do serviço manual, reconhecendo que a eficiência dos serviços humanos diminui com o aumento da procura. Verifica-se também que a adoção de serviços baseados em IA não substitui automaticamente o serviço manual, sendo esta decisão influenciada por fatores como a sensibilidade do consumidor, a eficiência de custos da IA e a redução da eficiência dos serviços humanos (He et al., 2024).

Tendo em conta os desafios e limitações identificados na adoção de técnicas de ML, é essencial compreender quais os modelos de ML mais aplicados à previsão da procura e de que forma estes superam as limitações dos métodos tradicionais. A secção seguinte apresenta uma análise detalhada das abordagens mais relevantes identificadas na literatura recente.

2.3. Modelos Aplicados à Previsão da Procura

No que diz respeito ao trabalho desenvolvido na área, a literatura evoluiu de modelos estatísticos simples para arquiteturas complexas de ML e DL. Segundo a síntese de Vhatkar et al. (2025), estudos anteriores focaram-se na aplicação de algoritmos como *Random Forest*, XGBoost e Redes Neuronais (LSTM) para capturar padrões não-lineares. No entanto, a tendência atual (conforme a Tabela 1 do estudo mencionado) revela que o foco da investigação mudou da simples aplicação de modelos para a otimização de hiperparâmetros (via *Grid Search*). Este avanço permite que modelos mais 'leves' de *Gradient Boosting* consigam superar, em muitos casos, modelos de DL mais pesados, desde que as suas configurações sejam ajustadas com precisão ao contexto do retalho.

Segundo Zietsman & van Vuuren (2023), o campo de pesquisa sobre previsão de séries temporais é bem explorado e existem diversas referências tanto para modelos de previsão estatísticos, como para modelos de previsão com ML. Os modelos estatísticos incluem modelos simples de referência (*naive models*), médias móveis (*moving averages*), modelos de regressão, modelos de suavização exponencial, modelos autorregressivos integrados de médias móveis (*autorregressive integrated moving average* (ARIMA) *models*) e os modelos autorregressivos de heterocedasticidade condicional (*autoregressive conditional heteroscedastic* (ARCH) *models*), entre outros. Os modelos de ML, por sua vez, incluem modelos baseados em árvores, como por exemplo, o algoritmo XGBoost e o algoritmo LightGBM, redes neuronais, como por exemplo, *Recurrent Neural Netmworks* (RNNs), e *Convolutional Neural Networks* (CNNs), e arquiteturas *Transformers*, entre outros. (Zietsman & van Vuuren, 2023).

De seguida, apresentam-se os principais modelos identificados na literatura, com destaque nas suas aplicações, benefícios e limitações, e no modo como contribuem para previsões mais precisas e sustentáveis no contexto do *e-Commerce*.

2.3.1. Modelos De Previsão Estatísticos

Como referido anteriormente, existem diversos tipos de modelos que são incluídos nos modelos estatísticos de previsão. Os modelos de suavização exponencial e os modelos ARIMA são as duas abordagens mais utilizadas para a previsão de séries temporais e oferecem abordagens complementares ao problema. Por um lado, os modelos de suavização exponencial baseiam-se na descrição da tendência e da sazonalidade presentes nos dados. Por outro lado, os modelos ARIMA têm como principal objetivo caracterizar as autocorrelações existentes na série temporal (Hyndman & Athanasopoulos, 2018).

As previsões obtidas através dos métodos de suavização exponencial equivalem a médias ponderadas das observações passadas. Neste método considera-se que quanto mais recente é a observação, maior é o peso que lhe é atribuído. A suavização exponencial gera previsões fiáveis de forma rápida e aplicável numa grande variedade de séries temporais, tornando-se uma abordagem com uma relevância significativa para aplicações industriais (Hyndman & Athanasopoulos, 2018).

Num modelo de autorregressão prevê-se a variável de interesse através da combinação linear de valores passados da variável. Assim, um modelo autorregressivo de ordem p pode ser escrito da seguinte forma:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$$

Onde ε_t é o erro aleatório. Estes modelos são bastante flexíveis na forma como lidam com uma vasta variedade de padrões em séries temporais.

Por outro lado, o modelo das médias móveis utiliza erros de previsão passados num modelo semelhante ao da regressão:

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q},$$

Onde ε_t é o erro aleatório. Como não se observa diretamente os valores de ε_t, não é considerado um modelo de regressão tradicional. É de notar que cada y_t pode ser interpretado como uma média móvel ponderada dos erros de previsão passados.

Ao combinar a diferenciação com o modelo de autoregressão e com o modelo das médias móveis, obtemos um modelo ARIMA não sazonal. O modelo completo pode ser escrito da seguinte forma:

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

Onde y'_t é a série diferenciada. Chama-se a isto o modelo ARIMA(p, d, q), onde p é a ordem da parte autorregressiva, d é o grau de diferenciação de primeira ordem envolvido e q é a ordem da parte da média móvel.

Para clarificar a notação matemática apresentada, detalham-se de seguida as variáveis e os parâmetros que compõem as equações:

- y_t: Representa o valor real observado da variável de interesse (neste caso, a procura) no período atual t.
- y_{t-1}, ..., y_{t-p}: Correspondem aos valores reais observados em períodos passados, até à ordem autorregressiva p.
- c: Corresponde à constante do modelo, representando a tendência base da série temporal.
- ϕ_1, ..., ϕ_p: São os parâmetros da componente Autorregressiva (AR). Quantificam a influência direta que cada observação passada tem no valor presente.
- ε_t: É o termo de erro aleatório no momento t. Representa a parte da variação da procura que é completamente imprevisível e que o modelo não consegue capturar.
- ε_{t-1}, ..., ε_{t-q}: Representam os erros de previsão cometidos nos momentos anteriores, até à ordem de q.
- θ_1, ..., θ_q: São os parâmetros da componente de médias móveis. Indicam o peso de ajustamento que os erros de previsão passados exercem sobre o valor atual.
- y'_t: Representa a série temporal após a aplicação do processo de diferenciação (de ordem d), um passo estatístico necessário para remover a tendência e tornar a série estacionária antes da modelação.

Os modelos ARIMA também são capazes de modelar uma grande variedade de dados sazonais. Para tal basta incluir no modelo os termos sazonais $(P, D, Q)_m$, onde m é o número de observações por ano (Hyndman & Athanasopoulos, 2018).

Em síntese, ambos os modelos representam técnicas diferentes, mas complementares, para a modelação e previsão de séries temporais. Os modelos de suavização exponencial destacam-se pela capacidade de captar de forma eficiente padrões de tendência e sazonalidade, enquanto os modelos ARIMA permitem modelar dependências temporais complexas através da combinação de componentes autorregressivas e de médias móveis. Posto isto, a escolha entre estes dois modelos deve ser feita tendo em conta as características da série temporal e os objetivos analíticos e de previsão definidos. A comparação entre ambos pode conduzir a uma compreensão mais abrangente e a previsões mais robustas (Hyndman & Athanasopoulos, 2018).

2.3.2. Modelos de *Machine Learning*

Enquanto os modelos estatísticos, como o ARIMA, assumem relações lineares, os modelos de ML permitem a integração de um conjunto alargado de variáveis, como, preço, promoções, sazonalidade, clima e campanhas de marketing, realizando um desempenho preditivo superior aos modelos lineares (Zietsman & van Vuuren, 2023).

A base teórica dominante neste campo é o *Gradient Boosting Machine* (GBM), proposto originalmente por Friedman (2001), onde a estimação da função é vista de uma perspetiva de otimização numérica no espaço da função, em vez do espaço dos parâmetros. A otimização no espaço da função constrói o modelo final somando sucessivas funções mais simples, sendo que cada função adicionada é criada especificamente para corrigir os erros deixados pela soma das funções anteriores (Friedman, 2001).

Posto isto, o principal objetivo do GBM é a construção de um modelo aditivo numa abordagem por etapas, onde cada novo modelo é adicionado para corrigir os erros dos anteriores, minimizando uma função de perda esperada (Friedman, 2001).

Entre as diversas aplicações do GBM, destaca-se o modelo XGBoost, desenvolvido por T. Chen & Guestrin (2016), um sistema escalável e eficiente de *tree boosting*, amplamente utilizado para alcançar os melhores resultados possíveis até ao momento em diversos desafios de ML.

Este modelo diferencia-se do GBM tradicional através de diversas inovações algorítmicas e de sistema, como por exemplo, a regularização. O termo de regularização (Ω) utilizado na função objetivo do XGBoost penaliza a complexidade do modelo, sendo crucial para evitar *overfitting* e simplificar os modelos preditivos. Para além disso, o modelo inclui um método sensível à esparsidade, que lida com os dados em falta, aprendendo uma direção padrão para os dados ausentes durante o treino da árvore. De forma a reduzir o custo de ordenação dos dados,

processando milhões de exemplos de forma eficiente, o XGBoost tem uma estrutura de blocos em colunas que permite a computação paralela.

O modelo XGBoost tem a seguinte função objetivo:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

Onde $\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2$

O termo de perda ($\sum_i l(\hat{y}_i, y_i)$) mede o quão bem o modelo se ajusta aos dados de treino, é a diferença entre a previsão ($\hat{y}_i$) e o valor real (y_i). A função l deve ser diferenciável e convexa. O termo de regularização ($\sum_k \Omega(f_k)$), como dito anteriormente, serve para penalizar a complexidade do modelo e evitar *overfitting*. A regularização é definida como $\Omega(f)$, onde T é o número de folhas na árvore e w são os pesos nas folhas (T. Chen & Guestrin, 2016).

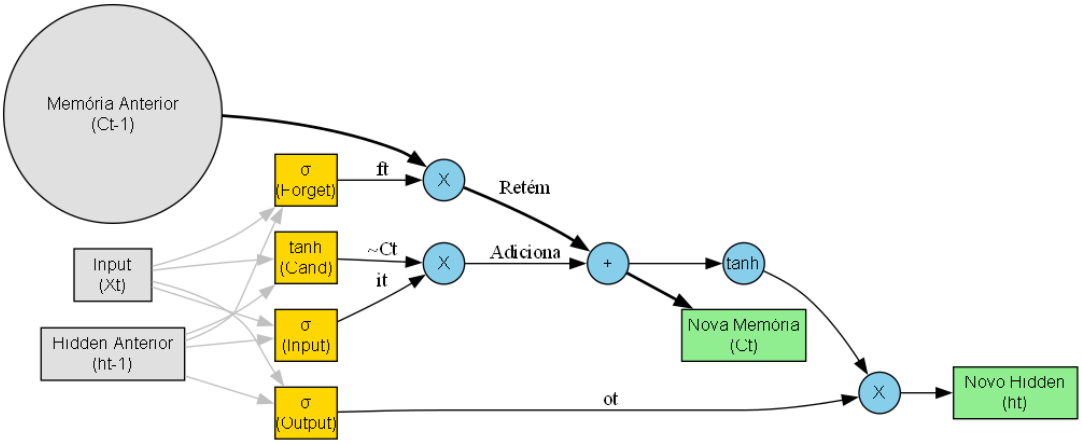
2.3.3. Modelos de *Deep Learning*

Os modelos de DL, ao contrário dos modelos estatísticos clássicos, distinguem-se pela capacidade de modelar relações não-lineares e capturar padrões em cadeias de abastecimento cada vez mais incertas. Em relação à previsão da procura, estas arquiteturas têm demonstrado superioridade no tratamento de grandes volumes de dados e na integração de variáveis exógenas heterogéneas (X. Chen et al., 2025).

As redes LSTM (*Long Short-Term Memory*) são um modelo de DL e têm sido muito utilizadas na previsão de séries temporais, devido à sua capacidade de reter informações sequenciais. Contudo, as arquiteturas LSTM tradicionais poderão apresentar limitações na explicação da relação entre inputs e outputs, bem como dificuldades em capturar dependências em sequências muito longas (Peng et al., 2022).

A estrutura interna de uma célula LSTM, ilustrada na Figura 1, utiliza portões lógicos de informação, decidindo o que memorizar e o que esquecer ao longo do tempo.

Figura 1 - Mecanismo interno de uma unidade LSTM.



Fonte: Elaboração Própria, adaptado de Peng et al. (2022).

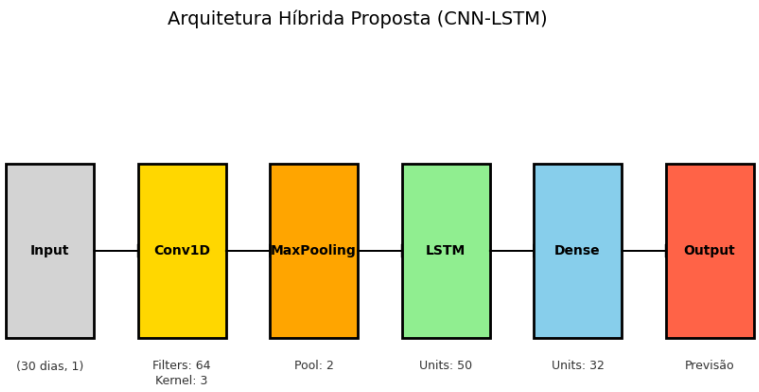
O diagrama detalha o fluxo de informação e as operações de portão (*gating*) que permitem à rede aprender dependências de longo prazo. O esquema ilustra como o fluxo de informação através dos portões de esquecimento (f_t) entrada (i_t) e saída (o_t) regulam a memória da célula. A estrutura concetual apresentada foi adaptada da arquitetura descrita por (Peng et al., 2022).

A lógica de funcionamento, baseada em Peng et al. (2022) demonstra como o *Forget Gate* (σ) decide a informação a descartar do estado da célula anterior (C_{t-1}), enquanto o *Input Gate* atualiza a memória com novos dados relevantes (X_t). A Figura 1 foi gerada através da linguagem Python, com a biblioteca Graphviz, utilizando o código de programação do Apêndice B.

De forma a diminuir as limitações de memória em sequências extensas, a literatura recente destaca a utilização de mecanismos de atenção. Como notado por Peng et al. (2022) a junção de módulos de atenção com as redes LSTM permite atribuir pesos diferentes às características de entrada, destacando os fatores mais críticos que influenciam a procura num determinado momento. Esta evolução é a base para arquiteturas mais modernas (Peng et al., 2022).

A pesquisa atual tende para a combinação de diferentes arquiteturas para maximizar a precisão. Para tal, são aproveitados os pontos fortes de cada modelo. Uma abordagem comum é a fusão de Redes Conolucionais (CNN) – para a extração de padrões locais – com redes LSTM, que processam as dependências temporais (Joseph et al., 2022). A Figura 2 exemplifica uma arquitetura híbrida típica neste contexto.

Figura 2 - Arquitetura de um modelo híbrido (CNN-LSTM).



Fonte: Elaboração Própria adaptado de Joseph et al. (2022).

O diagrama demonstra o fluxo sequencial de processamento, onde as camadas CNN (retângulo amarelo) são utilizadas para a extração de características e as camadas LSTM (retângulo verde) para análise temporal. A estrutura foi baseada na abordagem proposta por Joseph et al (2022). Tal como na figura anterior, a Figura 2 foi gerada através da linguagem Python, com a biblioteca Matplotlib, utilizando o código de programação do Apêndice C.

Esta combinação (CNN-LSTM) melhora a precisão ao capturar simultaneamente características espaciais e temporais, sendo eficaz na previsão de procura ao nível do item (Joseph et al., 2022).

Adicionalmente, Dasoomi et al. (2023) exploraram a otimização de hiperparâmetros, com a integração de algoritmos meta-heurísticos (*Gray Wolf Optimization*) com as redes convolucionais (GWO-CNN). Este estudo conclui que esta otimização permitiu superar a precisão de modelos convencionais como as redes LSMT isolada. (Dasoomi et al., 2023).

A literatura mais recente ultrapassa os dados puramente numéricos. Li et al. (2024), introduziram um modelo de fusão multimodal, M2TFM (*Multi-modal Transform-based Fusion Model*), baseado na arquitetura *Transformers,* que permite prever a procura de novos produtos através da utilização de imagens de produtos, texto (descrições) e dados contextuais. Esta abordagem é essencial para o retalho eletrónico, onde existe um histórico de vendas curto ou, por vezes, inexistente. (Li et al., 2024). Para além disso, Nie et al. (2025), demonstraram a eficácia da integração de *Large Language Models (LLMs)*

com redes espácio-temporais, STGNN (*Spatio-Temporal Graph Neural Network*). A consolidação destes dois métodos permite que o modelo integre ventos externos, como notícias, crises climáticas ou tendências sociais, através de processamento de linguagem natural, ajustando assim as previsões logísticas em tempo real. (Nie et al., 2025).

2.3.4. Métricas de Avaliação de Desempenho

De acordo com a literatura recente, a avaliação da precisão dos modelos de procura é feita através de métricas que quantificam o desvio entre os valores reais e os valores de previsão. No estudo de Chen et al (2025), são utilizadas as métricas MSE (*Mean Square Error*), MAE (*Mean Absolute Error*) e MAPE (*Mean Absolute Percentage Error*), pela capacidade de estas garantirem uma avaliação abrangente do desempenho dos algoritmos. (Chen et al., 2025).

De forma complementar Dasoomi et al. (2023), utiliza as seguintes métricas, acurácia, precisão, revocação, *F1-score*, MSE e RMSE (*Root Mean Squared Error*). (Dasoomi et al., 2023).

As referidas métricas têm a seguinte formulação matemática:

- **Erro Quadrático Médio (MSE)**

O MSE, apresentado na Equação 1, mede a variância do erro e penaliza desvios de grande magnitude. É uma métrica central na otimização de modelos de aprendizagem profunda, sendo utilizada por Dasoomi et al. (2023) e Chen et al (2025).

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

Equação 1 - Erro Quadrático Médio

- **Raiz do Erro Quadrático Médio (RMSE)**

O RMSE, apresentado na Equação 2, é classificado por Dasoomi et al. (2023) como uma métrica padrão e amplamente estabelecida para avaliar o desempenho de abordagens de ML. Por apresentar o erro na mesma unidade de medida dos dados originais, permite uma interpretação direta do desvio médio da previsão em relação à procura real.

$$RMSE = \sqrt{MSE}$$

Equação 2 - Raiz do Erro Quadrático Médio

- **Erro Absoluto Médio (MAE)**

O MAE, apresentado na Equação 3, fornece a média da magnitude absoluta dos erros. Ao contrário das métricas quadráticas, oferece uma visão linear da precisão e é menos sensível a valores externos (*outliers*). (Chen et al., 2025).

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

Equação 3 - Erro Absoluto Médio

- **Erro Médio Absoluto Percentual (MAPE)**

O MAPE, apresentado na Equação 4, exprime o erro em termos percentuais, o que facilita a comparação do desempenho entre diferentes escalas de procura e categorias de produtos. (Chen et al., 2025).

$$MAPE = \frac{1}{N}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100$$

Equação 4 - Erro Médio Absoluto Percentual

Onde N é o número de amostras, y_i o valor real e $\hat{y}_i$ o valor previsto.

2.4.Sustentabilidade e logística no *e-Commerce*

O crescimento acelerado das aplicações de Internet e do *e-Commerce* têm levado a um aumento constante do número de consumidores online. (Gao et al., 2024). Contudo, é gerada uma explosão no volume de resíduos de embalagens, provocando desafios ambientais significativos. (Xu et al., 2025). Torna-se então necessário adotar estruturas de cadeias de abastecimento sustentáveis que equilibrem a eficiência económica com a preservação ambiental. (Gao et al., 2024).

Nos dias de hoje, a redução, a reciclagem e a reutilização são as três estratégias mais comuns para mitigar o desperdício de embalagens no setor. (Xu et al., 2025). Entre estas, a reutilização de embalagens destaca-se por permitir uma redução drástica nas emissões de carbono e no consumo de recursos naturais após múltiplos ciclos de utilização, promovendo o conceito de "desperdício zero" ao fim da sua vida útil. (Xu et al., 2025). No entanto, estes sistemas dependem da gestão da incerteza da procura e de modelos de suporte à decisão. Modelos que integram a teoria dos jogos com abordagens *data-driven* demonstram que a coordenação entre decisões de encomenda e precificação é essencial para que a sustentabilidade não comprometa a rentabilidade. (Xu et al., 2025).

2.4.1. Dimensões da sustentabilidade (económica, ambiental e social)

A análise da logística no *e-Commerce* deve ser feita sob o prisma do *Triple Bottom Line*, integrando três dimensões fundamentais, de forma a garantir a sustentabilidade de uma rede logística a longo prazo (Cano et al., 2022):

- Dimensão Económica: Foca-se na estabilidade financeira e no crescimento do negócio. A sustentabilidade aqui é vista como uma vantagem competitiva, onde a otimização de recursos reduz custos operacionais e abre novas oportunidades de mercado.
- Dimensão Ambiental: Refere-se à conservação de recursos naturais e à redução da pegada de CO_2. No *e-Commerce*, isto traduz-se em rotas de entrega mais curtas, consolidação de pedidos para evitar viagens redundantes e a adoção de materiais de baixo impacto.
- Dimensão Social: Envolve a satisfação do cliente e o impacto na comunidade urbana. Inclui desde a redução do ruído e congestionamento nas cidades até à criação de valor para todos os *stakeholders* através de práticas transparentes e éticas.

Pode ser ainda considerada uma dimensão tecnológica, onde inovações como veículos autónomos, *drones* e cacifos inteligentes funcionam como facilitadores destas três dimensões, otimizando a entrega ao mesmo tempo que melhoram o bem-estar social e reduzem o impacto ambiental. (Cano et al., 2022).

2.4.2. O papel da previsão na sustentabilidade operacional

A ligação entre a eficiência logística e o impacto ambiental é facilitada através da precisão preditiva. A previsão da procura deixa de ser apenas uma ferramenta de gestão de inventário para se tornar um motor de sustentabilidade. Previsões rigorosas permitem que haja uma redução de desperdícios, evitando o excesso de *stock* que frequentemente resulta em produtos obsoletos ou descartados. Para além disso, permitem a otimização de transportes, ou seja, minimiza envios urgentes e fracionados, que são altamente poluentes, e, ainda, a gestão de embalagens reutilizáveis, como sugerido por Xu et al. (2025), prever o fluxo de retorno de embalagens é crítico para equilibrar o inventário.

A integração de modelos avançados, como os propostos por Datta et al. (2025), demonstra que é possível utilizar processos estocásticos e ML para implementar precificação dinâmica baseada no "nível ecológico" do produto (β). Ao ajustar o preço conforme o impacto ambiental e a previsão de procura, as empresas não só maximizam o lucro esperado ($Lucro =$

$\sum Receita(p, nível_{eco}) - Custos)$, como também induzem comportamentos de consumo mais conscientes. (Datta et al., 2025).

Apesar deste potencial, a literatura de Badakhshan et al. (2024) e Surur et al. (2025) destaca que, embora o ML seja amplamente utilizado para melhorar a precisão métrica (como o MAE ou RMSE), existe ainda uma escassez de estudos que quantifiquem diretamente o benefício ambiental (redução de kg de CO_2 ou toneladas de resíduos) derivado dessas previsões. (Badakhshan et al., 2024; Surur et al., 2025).

A precisão preditiva é hoje inseparável da Economia Circular. Segundo Mohammed et al. (2025), o erro de previsão é o maior gerador de logística inversa evitável. Através da utilização da IA, otimizando o *stock*, evita-se o transporte desnecessário de devoluções e o desperdício de produtos perecíveis, impactando diretamente os indicadores de sustentabilidade. (Mohammed et al., 2025).

2.5.Simulação e *Machine Learning*

A combinação entre simulação e ML constitui uma tendência emergente na literatura, com potencial para transformar a gestão da cadeia de abastecimento. Segundo Badakhshan et al. (2024), a utilização de modelos híbridos, que unem simulação e algoritmos de ML, possibilita a avaliação de múltiplos cenários operacionais antes da sua implementação real, reduzindo riscos em ambientes de alta incerteza. (Badakhshan et al., 2024).

2.5.1. Simulação de eventos discretos e baseada em agentes

As técnicas de modelação estocástica, tais como a Simulação de Eventos Discretos (DES) e a Simulação Baseada em Agentes (ABS), são amplamente utilizadas na gestão da cadeia de abastecimento para capturar as incertezas do mundo real. (Badakhshan et al., 2024). Enquanto a DES se foca na sequência de processos operacionais, a ABS permite modelar a interação dinâmica entre diversos intervenientes. De acordo com Xu et al. (2025), estas interações são fundamentais para compreender o equilíbrio entre agentes (como retalhistas e plataformas) na adoção de práticas sustentáveis, sendo a simulação a ferramenta ideal para testar políticas de reposição e transporte em contextos de procura volátil. (Xu et al., 2025).

2.5.2. Modelos híbridos preditivo-simulativos

A convergência entre análise preditiva e simulação constitui a fronteira atual da investigação. Cavalcante et al. (2019) demonstraram que esta integração aumenta a resiliência e a capacidade de tomada de decisão baseada em dados, permitindo avaliar de forma abrangente os impactos económicos e ambientais. Em contextos de *e-Commerce*, esta abordagem permite quantificar o impacto de previsões geradas por ML em termos de desperdício e emissões de carbono em cenários virtuais controlados. (Cavalcante et al., 2019). Contudo, Badakhshan et al. (2024) reforçam que, embora o campo esteja a crescer, ainda há uma carência de estudos que apliquem estes modelos híbridos em contextos reais de retalho sustentável. (Badakhshan et al., 2024).

2.6.Falhas Identificadas e oportunidades de investigação

A revisão de literatura realizada permite identificar as seguintes tendências e falhas de investigação relevantes:

Tabela 1 - Conclusões e Falhas Identificadas na Literatura

:

Área Temática	Principais Conclusões	Falhas Identificadas
Previsão da procura	Modelos de DL (LSTM, CNN) superam modelos de previsão estatísticos e modelos de ML.	Apesar do surgimento de modelos multimodais (Li et al., 2024), existe uma carência na aplicação destas múltiplas fontes de dados (imagens, texto, clima) diretamente na otimização de métricas de sustentabilidade (como a pegada de carbono).
ML no *e-Commerce*	Forte impacto da IA em personalização e previsão.	Pouca exploração dos riscos éticos e do *algorithmic bias*.
Sustentabilidade	Ligação entre previsão e redução de desperdício comprovada teoricamente.	Falta de quantificação empírica de indicadores ambientais (CO_2, devoluções).

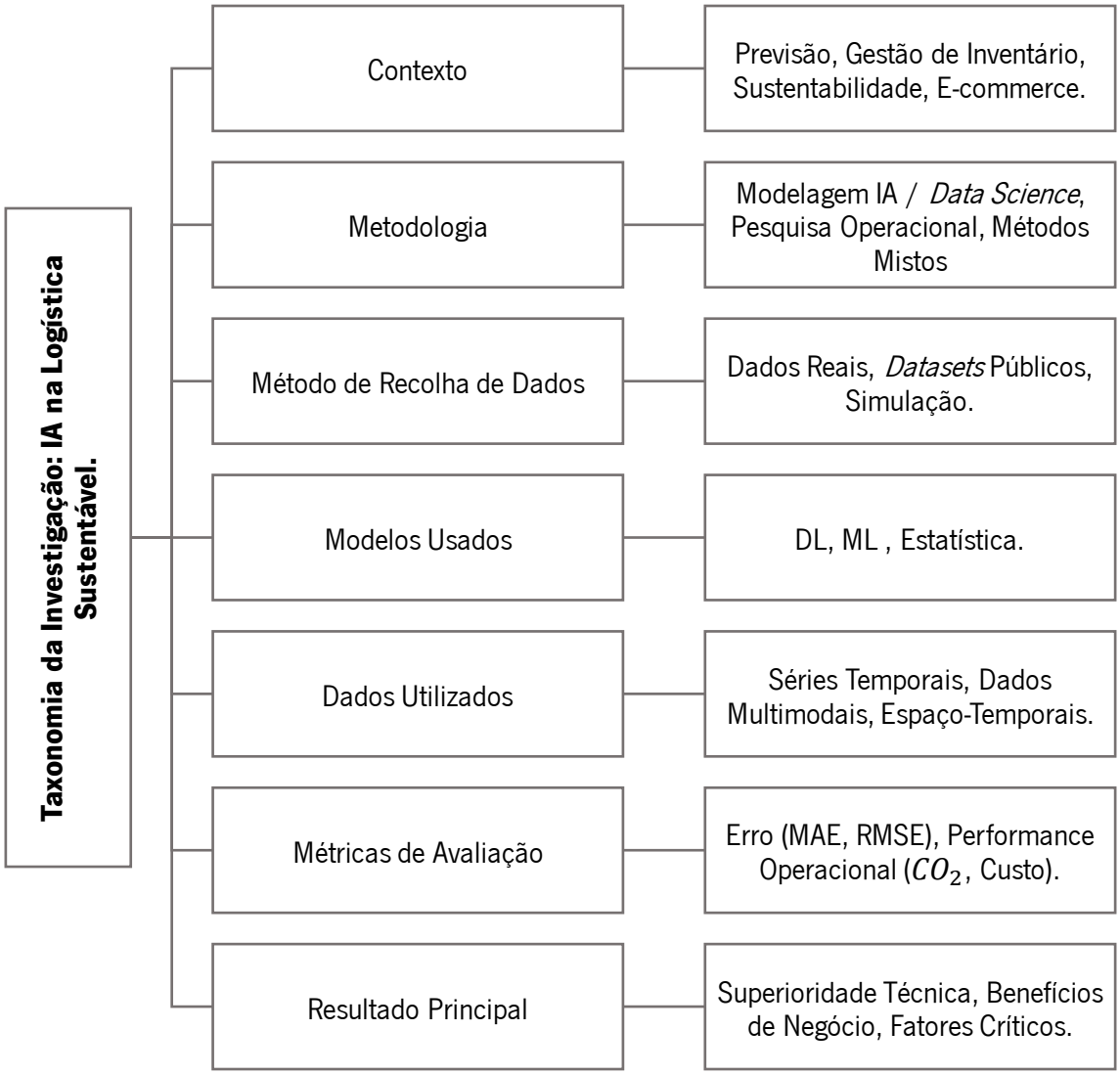
Simulação + ML	Combinação promissora para suporte à decisão.	Carência de estudos aplicados em contextos reais de *e-Commerce*.

Apesar do avanço nos modelos de ML, existe uma falha na literatura entre a previsão estática e a execução dinâmica (simulação). Esta dissertação tem como objetivo preencher esse vazio, através da utilização da simulação para validar como as previsões se comportam num ambiente de inventário com incerteza.

2.7.Síntese de Literatura e Taxonomia de Investigação

Com o intuito de sistematizar o conhecimento recolhido na revisão bibliográfica, apresenta-se nesta secção uma taxonomia (Figura 3) que classifica as principais dimensões da investigação em Previsão de Procura e Sustentabilidade Logística. Esta estrutura, baseada nos sete eixos críticos identificados (Contexto, Metodologia, Dados, Modelos, Tipologia, Métricas e Resultados), servirá de base para o desenvolvimento do modelo proposto no próximo capítulo.

Figura 3 - Taxonomia da Investigação



Fonte: Elaboração Própria

3. METODOLOGIA

Ao longo deste capítulo, é descrito o caminho metodológico usado para alcançar os objetivos de investigação. É detalhado o desenho de estudo, a natureza dos dados utilizados, as técnicas de processamento, as arquiteturas dos modelos de previsão e a lógica da integração com as métricas de sustentabilidade através de simulação.

3.1. Desenho Metodológico

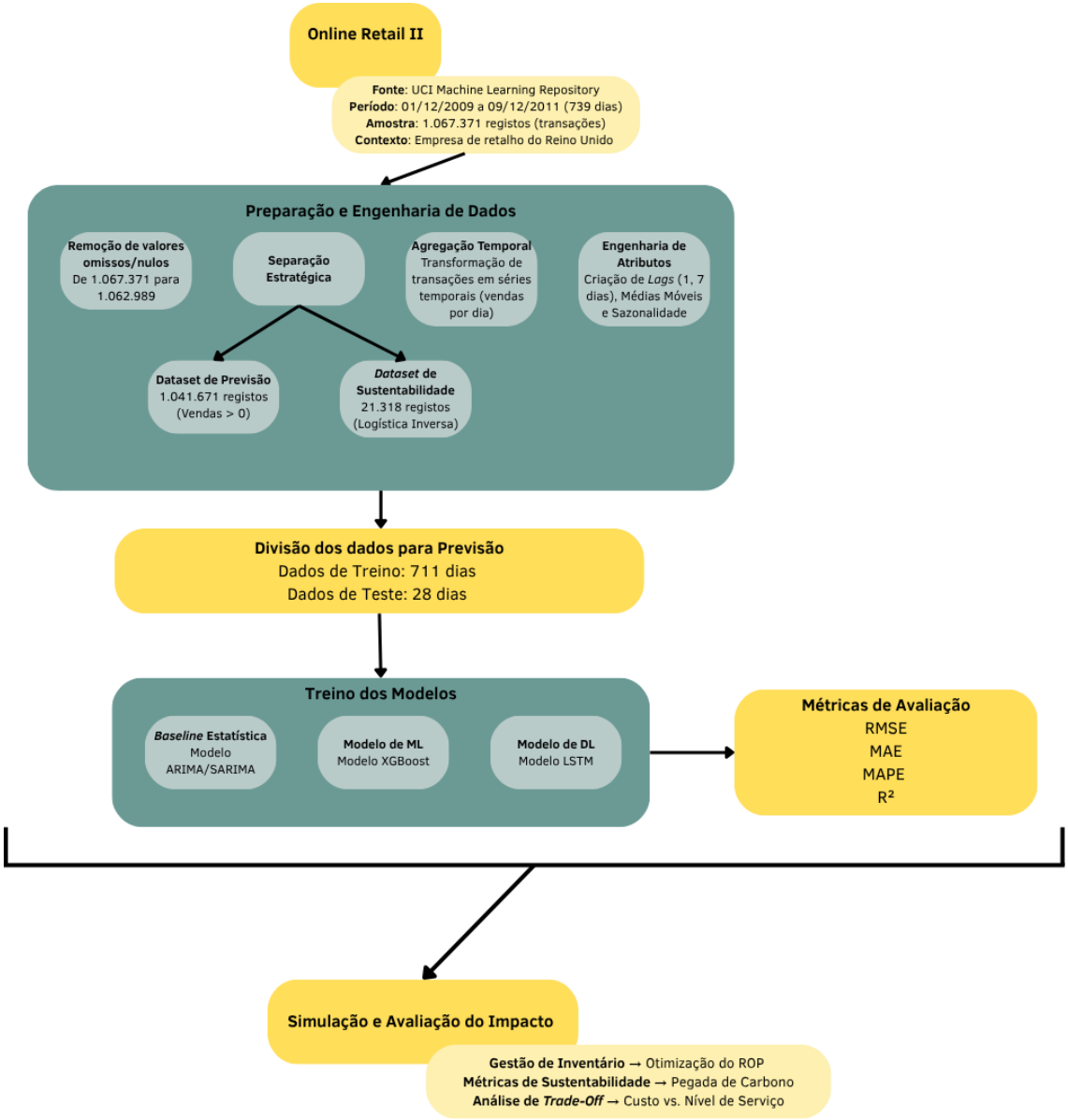
Nesta investigação é adotada uma estratégia experimental, com desenvolvimento e comparação de modelos preditivos e com a simulação de cenários operacionais. O estudo é guiado por uma filosofia pragmática, que valoriza a utilidade prática dos resultados para problemas reais de gestão logística e sustentabilidade.

A abordagem é dedutiva, fundamentada no teste de modelos e hipóteses com base em teorias e evidência empírica. Usa-se um método misto, combinando técnicas quantitativas, como a modelagem preditiva e a avaliação estatística de desempenho, com uma componente exploratória e interpretativa baseada na simulação de cenários práticos. Embora não envolva recolha qualitativa tradicional, a simulação permite compreender os impactos operacionais e sustentáveis de forma integrada, apoiando a tomada de decisão em contextos reais.

O desenho do estudo estrutura-se através de um horizonte transversal, centrando-se na análise de dados históricos num período delimitado para fins comparativos e preditivos. Os dados são obtidos a partir de bases de dados históricas de vendas reais, contemplando variáveis fundamentais como quantidades, preços, datas de faturação e cancelamentos. A análise envolve a modelagem estatística e preditiva, seguida de simulações para a avaliação dos impactos práticos e sustentáveis das previsões geradas.

A arquitetura metodológica desta investigação, detalhada na Figura 4, estrutura-se num fluxo lógico que percorre desde a extração de dados brutos do repositório UCI até à tomada de decisão operacional.

Figura 4 - Fluxograma da arquitetura metodológica



Fonte: Elaboração Própria, adaptado de (Z. Mohammed et al., 2026).

3.2. Estratégias de Investigação e Etapas de Estudo

A estratégia de investigação foi estruturada de forma a garantir uma transação fluída entre a análise preditiva e a aplicação prática. Primeiramente realizou-se a preparação e exploração de dados, onde se concretizou a aquisição, limpeza (tratamento de cancelamentos e valores omissos) e a análise exploratória para identificar padrões de sazonalidade e tendências de venda. De seguida, realizou-se o desenvolvimento e treino de modelos, implementando em Python o modelo ARIMA (referência estatística), o XGBoost (ML) e o LSTM (DL). Nesta fase, os modelos foram ajustados para minimizar os erros de previsão. Posteriormente, fez-se uma avaliação comparativa,

utilizando as métricas estatísticas (RMSE e MAE) para validar qual o modelo com melhor desempenho. Por fim, simularam-se cenários através do programa Anylogic, onde as previsões geradas foram introduzidas num ambiente simulado para testar o impacto de políticas de *stock* e sustentabilidade, como a redução de embalagens não sustentáveis e a otimização de recursos face a variações da procura.

3.3. Base de Dados e Preparação

3.3.1. Origem e Natureza dos Dados

Os dados utilizados para esta investigação foram obtidos a partir do repositório público *UCI Machine Learning Repository*, mais precisamente o conjunto de dados "Online Retail II" [1]. Este *dataset* inclui todas as transações efetuadas por um retalhista online sediado no Reino Unido, entre o período de 1 de dezembro de 2009 e 9 de dezembro de 2011. Este conjunto de dados foi escolhido por constituir uma base de dados real, extensa e adequada ao objetivo desta investigação, uma vez que reúne mais de um milhão de transações de um retalhista online do Reino Unido ao longo de aproximadamente dois anos, permitindo observar padrões temporais relevantes como tendência, sazonalidade, volatilidade e picos de procura (sobretudo de artigos de oferta). A sua natureza transacional, com variáveis como quantidade, preço unitário, data da fatura, cliente e país, oferece informação suficiente para o desenvolvimento e comparação de modelos preditivos aplicados à previsão da procura em contexto de *e-Commerce*. Adicionalmente, a presença de cancelamentos, devoluções e registos de desperdício torna esta base particularmente útil para uma análise que não se limita à precisão estatística, mas que procura também relacionar a previsão da procura com decisões operacionais e implicações de sustentabilidade. Apesar de se tratar de um *dataset* histórico, a sua dimensão, granularidade e origem real justificam a sua utilização como suporte empírico robusto para esta dissertação. (Chen, 2019).

3.3.2. Descrição das Variáveis e Indicadores

O conjunto de dados é composto por variáveis nominais e numéricas que permitem caracterizar o comportamento transacional da empresa. Abaixo, descrevem-se as variáveis do conjunto de dados e a sua relevância para a modelação:

- ***InvoiceNo* (Número da fatura):** Identificador nominal único de 6 dígitos atribuído a cada transação. No tratamento de dados, identificaram-se os códigos iniciados pela letra

[1] Disponível em: https://archive.ics.uci.edu/dataset/502/online+retail+ii

'c', que indicam cancelamentos, permitindo a limpeza de registos que não representam procura real.

- ***StockCode* (Código do produto):** Identificador nominal de 5 dígitos atribuído exclusivamente a cada produto distinto, essencial para a agregação da procura.
- ***Description* (Descrição):** Nome nominal do produto, utilizado para a compreensão qualitativa do inventário.
- ***Quantity* (Quantidade):** Variável numérica que representa as unidades de cada produto por transação. Esta foi a variável alvo para o treino dos modelos preditivos.
- ***InvoiceDate* (Data e hora da fatura):** Variável numérica que regista o momento exato da transação. Foi fundamental para a estruturação das séries temporais e para a criação de variáveis de sazonalidade (dia, mês e hora).
- ***UnitPrice* (Preço unitário):** Valor numérico do produto por unidade em libras esterlinas (£), utilizado para calcular o impacto económico das previsões.
- ***CustomerID* (ID do cliente):** Número nominal de 5 dígitos que identifica cada cliente de forma exclusiva, permitindo distinguir entre clientes retalhistas e grossistas.
- ***Country* (País):** Nome nominal do país de residência do cliente, útil para análises de distribuição logística.

3.3.3. Protocolo de Limpeza e Tratamento de Discrepâncias

O processo de preparação de dados transformou uma base bruta e ruidosa num conjunto de dados fidedignos para a investigação. A amostra inicial do *dataset* contava com 1 067 371 registos. Através de uma análise exploratória rigorosa, procedeu-se uma limpeza dos dados, de forma a garantir a integridade dos modelos de previsão e da simulação. Na tabela 2 são detalhadas as etapas de limpeza efetuadas.

Tabela 2 - Etapas do Tratamento dos Dados

Etapa de Tratamento	**Impacto na Amostra**	**Justificação Técnica**
Amostra Inicial Bruta	1 067 371 linhas e 8 colunas	Dataset Online Retail II (2009-2011).
Remoção de Nulos (Descrição)	- 4 382 linhas	Registos com Preço 0.0 sem descrição (ruído de sistema).
Ajuste de Clientes Omissos	243 007 linhas	Substituição por ID '99999' para evitar perda de volume de vendas.
Criação de uma nova coluna – "Total Vendas"	1 062 989 linhas e 9 colunas	De forma a saber o número total de vendas ou faturação, criou-se a coluna "Total Vendas" através da multiplicação da "Quantidade" com o "Preço Unitário"
Remoção de Ajustes Contabilísticos	- 5 linhas	Exclusão de "*Adjust bad debt*" (impacto de – 158 676,14 £ sem valor logístico).
Amostra Final Limpa	1 062 984 linhas	Base para separação entre Procura e Sustentabilidade.

Ao longo do tratamento dos dados, foram detetados 4 382 valores omissos na coluna "Descrição" e 243 007 valores omissos na coluna "ID do Cliente". Através de uma análise exploratória foi revelado que a totalidade dos registos sem descrição apresentavam um preço

unitário de 0.0, confirmando tratar-se de erros de sistema ou transações inválidas. Consequentemente, estes registos foram excluídos da amostra final.

Em relação aos valores em falta na coluna "ID do Cliente", optou-se por substituir os valores em falta por valores genéricos. Caso fossem removidas estas linhas, seriam perdidos 243 007 registos, perdendo, consequentemente, dados sobre as vendas desses mesmos clientes, que poderia levar os modelos a prever uma procura muito menor do que o que realmente aconteceu. No caso de remover a coluna inteira do "ID do Cliente" seria perdida a capacidade de distinguir se as encomendas são de clientes habituais ou se são várias encomendas pequenas.

Como se pode verificar na Figura 5, foi realizada uma análise dos valores das vendas e das quantidades. Verificou-se a existência de valores negativos no total das vendas e fez-se a comparação com o valor da quantidade, verificando de este era ou não negativo.

Figura 5 - Relatório de Discrepâncias.

RELATÓRIO DE DISCREPÂNCIAS

- Linhas com Total Negativo mas Quantidade Positiva: 5
- Linhas com Quantidade Negativa mas Total Zero: 768

Exemplo do Caso 1 (Ajustes Manuais):

	Descrição	Quantidade	Preço Un.	Total
179403	Adjust bad debt	1	-53594.36	-53594.36
276274	Adjust bad debt	1	-44031.79	-44031.79
403472	Adjust bad debt	1	-38925.87	-38925.87
825444	Adjust bad debt	1	-11062.06	-11062.06
825445	Adjust bad debt	1	-11062.06	-11062.06

Exemplo do Caso 2 (Danos/Perdas):

	Descrição	Quantidade	Preço Un.	Total
263	85123a mixed	-96	0.0	-0.0
283	short	-240	0.0	-0.0
284	21733 mixed	-192	0.0	-0.0
3162	lost	-1043	0.0	-0.0
3168	damages	-117	0.0	-0.0

Fonte: Elaboração Própria.

Foram descobertas duas discrepâncias, o caso um, onde são encontradas 5 linhas do *dataset* em que o valor total é negativo, mas a quantidade é positiva e, o caso dois, onde o valor total é negativo devido aos valores da quantidade serem negativos.

No caso 1 é de denotar a descrição do produto, "Ajuste de dívida incobrável", que por sua vez significa que nenhum produto foi enviado, foi um acerto contabilístico, logo, estas linhas tiveram de ser eliminadas de forma a não causar um prejuízo logístico gigante, sendo este na verdade um prejuízo financeiro.

No caso 2 são encontradas 768 linhas com a descrição "*short*" (falta de material), "*lost*" (perdido), "*damages*" (danos). Neste caso, considerou-se desperdício físico, ou seja, produtos que existiam, que gastaram recursos a serem produzidos, e foram para o lixo sem nunca serem vendidos. Estes registos não foram eliminados, optou-se pela criação de um *dataset* de sustentabilidade, necessário para posteriormente justificar a necessidade de melhorar a gestão de *stock*.

Posto isto, finalizou-se a limpeza e tratamento dos dados com a criação de dois conjuntos de dados, como é possível observar na tabela 3, os produtos que realmente foram vendidos e as devoluções e desperdícios. Para os modelos de previsão de procura é necessário saber quanto é que os clientes pediram, a procura bruta, e para a parte da sustentabilidade é importante saber quanto é que voltou para trás, logística inversa/desperdício. No Apêndice D é visível partes do código de programação utilizado no tratamento dos dados.

Tabela 3 - Separação final dos dados em dois datasets

Indicador Estatístico	**Valor Final**	**Observação Técnica**
Amostra de Vendas Limpas	1.041.671	Base para o treino de modelos de previsão.
Amostra de Sustentabilidade	21.313	Registos de devoluções (C) e desperdício (*Lost/Damaged*).

3.3.4. Caracterização Estatística da Amostra

Ao longo desta secção é analisado o perfil estatístico dos dados após o processo de limpeza, com ênfase na distribuição temporal e no comportamento das variáveis críticas para a previsão da procura.

A base de dados final da amostra de previsão compreende um total de 1 041 671 registos, cobrindo o período de dezembro de 2009 a dezembro de 2011. A Tabela 4 resume as métricas centrais das principais variáveis numéricas.

Tabela 4 - Métricas das variáveis, Quantidade e Preço Unitário

Variável	Média	Desvio Padrão	Mínimo	Máximo
Quantidade	10.96	126.51	1	80995
Preço Unitário (£)	4.08	51.45	0	25111.09

De forma a compreender a estrutura subjacente da procura, realizou-se a decomposição da série temporal de vendas (Figura 6). Esta análise permite isolar os seus diferentes componentes fundamentais:

- **Vendas Totais (£):** Representa o volume diário de transações brutas tal como ocorreram, caracterizando-se por uma elevada volatilidade e picos de consumo acentuados (especialmente nos períodos que antecedem as épocas festivas).
- **Tendência:** Observa-se uma trajetória de crescimento consistente no volume de transações ao longo do horizonte temporal de dois anos. Indica uma expansão contínua da operação logística do retalhista, evitando flutuações do dia a dia.
- **Sazonalidade:** O gráfico demonstra um padrão cíclico perfeitamente repetitivo e constante. Este comportamento visualiza a flutuação rítmica do consumo, confirmando que a procura não é linear, mas sim governada por ciclos temporais curtos.
- **Resíduos (Ruído):** Representa as variações irregulares que não seguem o padrão sazonal ou de tendência. A presença de forte ruído residual em determinados períodos comprova a extrema volatilidade do *e-Commerce*.

Figura 6 - Decomposição da série temporal das vendas reais.

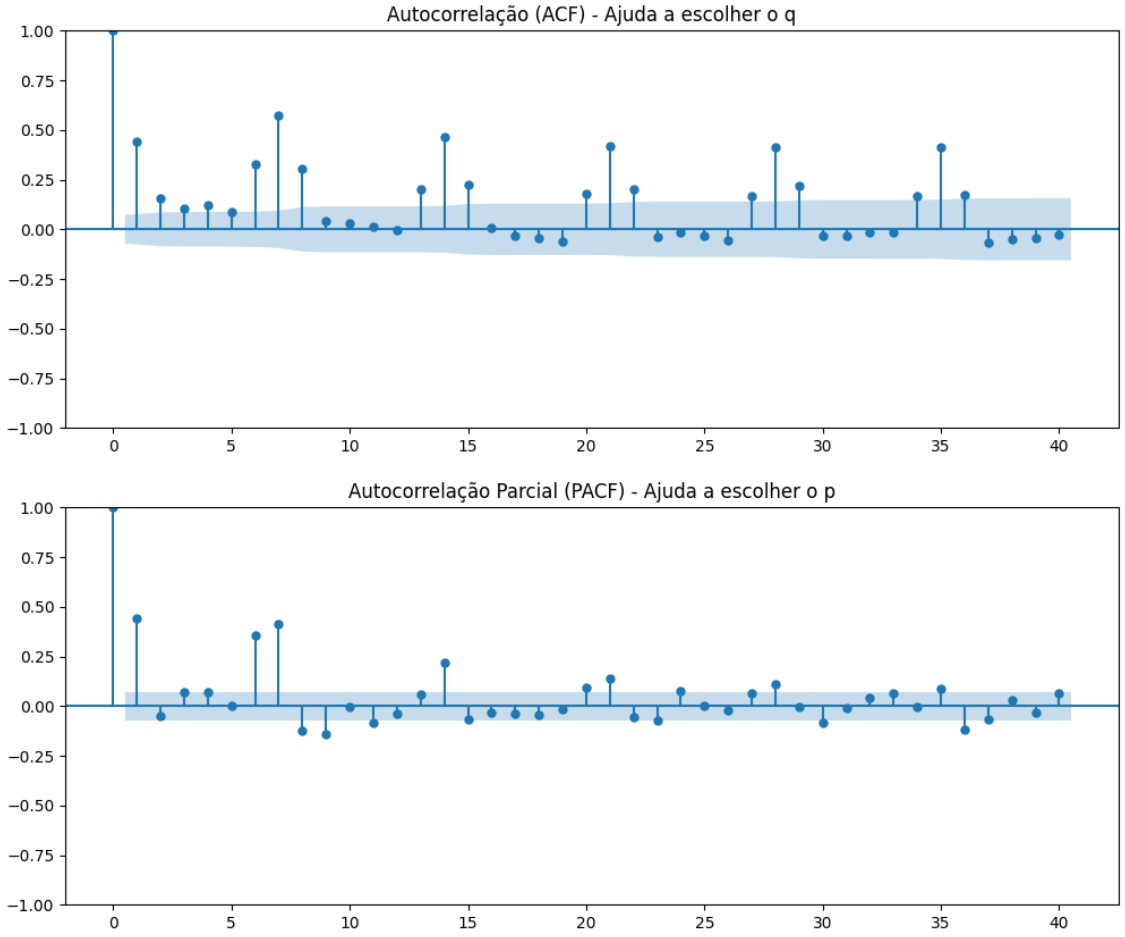


Fonte: Elaboração Própria.

De forma a validar a adequabilidade dos modelos de previsão, aplicou-se o teste *Augmented Dickey-Fuller* (ADF) à série temporal. Este teste verifica a presença de uma raiz unitária (*unit root*) na amostra, funcionando com base em duas hipóteses estatísticas, a hipótese nula, onde a série não é estacionária e a hipótese alternativa, onde a série é estacionária. A regra geral do teste é, se o valor-p for maior que 0.05, a hipótese nula não é rejeitada, logo a série não é estacionária, se o valor-p for menor ou igual que 0.05, a hipótese nula é rejeitada, sendo a série estacionária. No caso da série não ser estacionária, o parâmetro d de diferenciação no modelo ARIMA terá de ser utilizado. (Brooks, 2019; Dickey & Fuller, 1979).

O teste resultou num valor-p de 0.2836, superior ao nível de significância de 0.05. Com isto, a hipótese nula não é rejeitada, sendo a série não-estacionária. O resultado confirmou a

Figura 7 - Autocorrelação e Autocorrelação Parcial.



Fonte: Elaboração Própria.

presença de tendência e sazonalidade nos dados, justificando a necessidade de diferenciação da série para a aplicação do modelo ARIMA.

Os gráficos representados na Figura 7 revelam picos significativos de correlação a cada 7 períodos (dias). Esta evidência estatística confirma a existência de uma sazonalidade semanal forte, indicando que o comportamento de compra de segunda-feira, por exemplo, é um forte preditor para a segunda-feira seguinte.

Esta descoberta foi crucial para a parametrização dos modelos, definindo o valor de s=7 nas arquiteturas sazonais.

Recorde-se que, no modelo ARIMA (p,d,q), o parâmetro p representa a ordem autorregressiva, d o grau de diferenciação da série e q a ordem da componente de média móvel. Na identificação empírica destes parâmetros, a Função de Autocorrelação Parcial (PACF) foi utilizada como apoio à escolha de p, enquanto a Função de Autocorrelação total (ACF) serviu de apoio à escolha de q. Para maior detalhe, consultar a secção 2.3.1.

No caso do gráfico PACF, observou-se um pico significativo no primeiro e segundo desfasamentos, seguido de um corte abruto onde as correlações subsequentes decaem para dentro do intervalo de confiança de 95%. Segundo a metodologia de *Box-Jenkins*, este comportamento sugere um processo Auto-Regressivo de ordem 1 ou 2. Assim, foram testados modelos candidatos ARIMA(1,1,1) e ARIMA(2,1,1) para avaliar qual apresenta o melhor ajustamento aos dados. (Tunnicliffe Wilson, 2016).

A aplicação inicial de um modelo ARIMA(1,1,1) não-sazonal resultou numa previsão linear incapaz de capturar a volatilidade diária. Isto demonstra empiricamente que a série temporal possui padrões cíclicos de curto prazo (sazonalidade semanal) que um modelo ARIMA simples não consegue resolver. Tornou-se imperativo evoluir para um modelo SARIMA (*Seasonal* ARIMA).

3.4. Modelação Preditiva

Ao longo deste subcapítulo, é detalhada a fase de construção e de treino dos modelos de previsão de procura. A abordagem seguida baseia-se numa lógica comparativa, testando diferentes paradigmas matemáticos. Desde a estatística clássica ao DL, de forma a identificar qual a arquitetura que melhor captura a volatilidade e a sazonalidade do *e-Commerce*.

O principal objetivo é converter os dados históricos em previsões de alta-fidelidade, que por sua vez, alimentarão a simulação logística no Anylogic.

3.4.1. Engenharia de Atributos (*Feature Engineering*)

Após a fase de limpeza, os dados foram transformados de forma a garantir que os modelos de ML e DL capturassem com precisão as variações temporais do retalho *online*. Para tal, o processo de engenharia de atributos foi dividido em quatro partes:

- **Agregação e Decomposição Temporal**: Extraíram-se atributos específicos de tempo, da variável *InvoiceDate,* para capturar a sazonalidade, criando atributos correspondentes ao mês, ao dia do mês, ao dia da semana e à hora da transação. Isto permitiu que algoritmos como o XGBoost identificassem, de forma autónoma, picos de procura recorrentes, como épocas festivas mensais ou variações de volume a longo da semana.
- **Criação de Variáveis de Atraso (*Lags*):** De forma a capturar a inércia da procura e as tendências de curto prazo, foram criadas variáveis de atraso temporal. Estas variáveis representam a procura observada em períodos anteriores (t-1, t-7), servindo como um contexto histórico que permite aos modelos compreender a evolução dinâmica das vendas. No modelo LSTM, os dados foram estruturados em janelas temporais (*look-back windows*), transformando a série plana num formato tridimensional (amostras, janelas e variáveis).
- **Normalização e Escalonamento:** Dada a sensibilidade das redes neuronais a grandes variações de magnitude, aplicou-se a técnica de *MinMaxScaler* no modelo LSTM. Isto normalizou os valores de procura para um intervalo entre 0 e 1, garantindo a estabilidade da convergência do algoritmo durante o treino. No modelo XGBoost as variáveis mantiveram a escala original para preservar a interpretação das árvores de decisão.
- **Segmentação para Métricas de Sustentabilidade:** Criação de indicadores específicos de desperdício, para tal criou-se a variável *Total_Venda* que resulta da multiplicação da variável *Quantity pela variável Unit_Price*. Através então do cruzamento das variáveis Total_Venda e *Quantity* foram segmentados dois fluxos. O fluxo de procura real, focado em transações de venda efetiva, usada para treinar os modelos, e, o fluxo de desperdício operacional, onde estão isolados os registos em que a variável *UnitPrice* é igual a zero (identificados como *damages* ou *lost*) e as devoluções (faturas iniciadas por "C"). Isto permitiu criar a base de dados necessária para a posterior simulação do impacto ambiental e cálculo de emissões de CO_2 no Anylogic.

3.4.2. Arquiteturas e Parametrização dos Modelos

De forma a garantir uma análise comparativa abrangente, foram selecionados três modelos com fundamento teóricos diferentes. O modelo SARIMA, estatístico clássico, um algoritmo de ML baseado em árvores de decisão, o XGBoost e um modelo de DP especializado em sequências temporais, o LSTM.

I. SARIMA

O modelo ARIMA foi utilizado como uma linha de base estatística. Sendo reconhecido pela sua eficácia em séries temporais lineares, através da autocorrelação dos dados históricos. Este modelo foi utilizado de forma a capturar componentes de sazonalidade e tendência, exigindo uma fase prévia de verificação de estacionaridade de séries, através de diferenciação. Após esta verificação, optou-se por um modelo SARIMA (1,1,1) x (1,1,1,7), incorporando a sazonalidade semanal detetada na análise exploratória. A escolha manual dos parâmetros, em detrimento de algoritmos automatizados (Auto-ARIMA), visou garantir a robustez técnica face à volatilidade do *dataset*. No Apêndice E apresenta-se parte do código de programação utilizado para este modelo.

A inclusão deste modelo na investigação permitiu avaliar se modelos mais complexos oferecem um ganho de precisão significativo face às abordagens tradicionais.

II. XGBoost

Elevada performance em tarefas de regressão e capacidade de lidar com relações não lineares e dados tabulares complexos foram as razões que levaram à implementação do modelo XGBoost nesta investigação. Ao contrário do modelo estatístico tradicional, o XGBoost permitiu a integração de múltiplas variáveis, como o preço unitário e indicadores temporais. O algoritmo foi configurado para construir um conjunto de árvores de decisão de forma sequencial, onde cada árvore corrigia os erros anteriores, utilizando o método de *Gradient Boosting* para minimizar a função de perda.

De forma a otimizar a performance e prevenir o *overfitting,* foram definidos os seguintes hiperparâmetros*:*

Estimadores (*n_estimators*): 100 - É o número de árvores que o modelo vai criar.

Taxa de aprendizagem (*learning_rate*): 0.1 – Controla o quão rápido o modelo aprende.

Profundidade máxima (*max_depth*): 5 – Define a profundidade máxima de cada árvore.

Early_stopping_rounds: 50 – Paragem antecipada, ou seja, monitoriza o desempenho do modelo num conjunto de validação independente, interrompendo o treino caso não se verifique uma melhoria no erro durante 50 iterações consecutivas.

Estes valores foram selecionados através de um processo experimental de sintonização. Testaram-se diferentes configurações, tendo-se optado por estes valores por apresentarem o melhor equilíbrio entre a precisão preditiva, medida pelo MAE, e, a capacidade de generalização do modelo, evitando *overfitting*.

Dado a natureza de série temporal dos dados, não foi utilizada uma divisão aleatória comum (*shuffling*), pois iria afetar a cronologia dos eventos. Em vez de baralhar os dados aleatoriamente, utilizou-se a técnica *Walk-forward*. Esta técnica consiste em treinar o modelo num período passado e testá-lo no período imediatamente seguinte, movendo a janela de treino sucessivamente para a frente. Isto garante que a previsão para o dia n seja feita apenas com base na informação disponível até ao dia n-1, replicando fielmente o cenário real de tomada de decisão na gestão de inventários do *e-Commerce*.

No Apêndice F apresenta-se parte do código de programação utilizado para este modelo.

III. Redes Neuronais LSTM

A rede neural LSTM foi utilizada para tratar as dependências temporais de longo prazo e a volatilidade extrema do retalho online. Este modelo foi selecionado pela sua arquitetura específica, composta por portas de entrada, saída e esquecimento, que permitem reter informações relevantes durante longos períodos e descartar ruído.

A arquitetura implementada consistiu numa rede *Stacked* LSTM (camadas empilhadas), permitindo capturar representações de dados em múltiplos níveis de abstração. De forma a otimizar a aprendizagem, a primeira camada foi configurada com 50 unidades de memória e a segunda com 30 unidades. Para aumentar a capacidade de generalização e prevenir *overfitting,* integraram-se camadas de *Dropout* (0.2), que desativam aleatoriamente 20% dos neurónios durante o treino, forçando a rede a aprender padrões mais robustos. O algoritmo opera sobre uma janela temporal de 30 dias (*look-back*), capturando ciclos mensais de consumo. Por fim, utilizou-se o optimizador *Adam* para o ajuste dos pesos, permitindo capturar padrões cíclicos que modelos estatísticos tradicionais frequentemente ignoram.

No Apêndice G apresenta-se parte do código de programação utilizado para este modelo.

Na tabela 5 segue o resumo da parametrização dos modelos preditivos, analisando as componentes e parâmetros principais de cada modelo de previsão.

Tabela 5 - Resumo da Parametrização dos Modelos Preditivos

Modelo	**Componentes/Parâmetros Principais**	**Função/Justificação**
SARIMA	(1,1,1) x (1,1,1,7)	Captura de sazonalidade semanal (s=7).
XGBoost	*n_estimators*=100, *max_depth*=5, LR=0.1	Equilíbrio entre complexidade e generalização.
XGBoost	*early_stopping_rounds*=50	Prevenção de *overfitting* durante o treino.
LSTM	2 Camadas (50 e 30 unidades)	Aprendizagem de padrões em múltiplos níveis.
LSTM	*Dropout*=0.2, *Look-back*=30	Robustez contra ruído e memória mensal.

3.4.3. Métricas de Avaliação de Desempenho

Para avaliar a precisão preditiva e selecionar o modelo mais robusto para a simulação, foram utilizadas três métricas estatísticas complementares, MAE, RMSE e o MAPE. O RMSE foi priorizado por penalizar erros de grande magnitude, críticos para evitar ruturas de *stock* na operação logística. A fundamentação teórica e as respetivas formulações matemáticas destas métricas encontram-se detalhadas nas Equações 1, 2, 3 e 4 da secção 2.3.4.

3.5. Simulação e Dinâmica Operacional

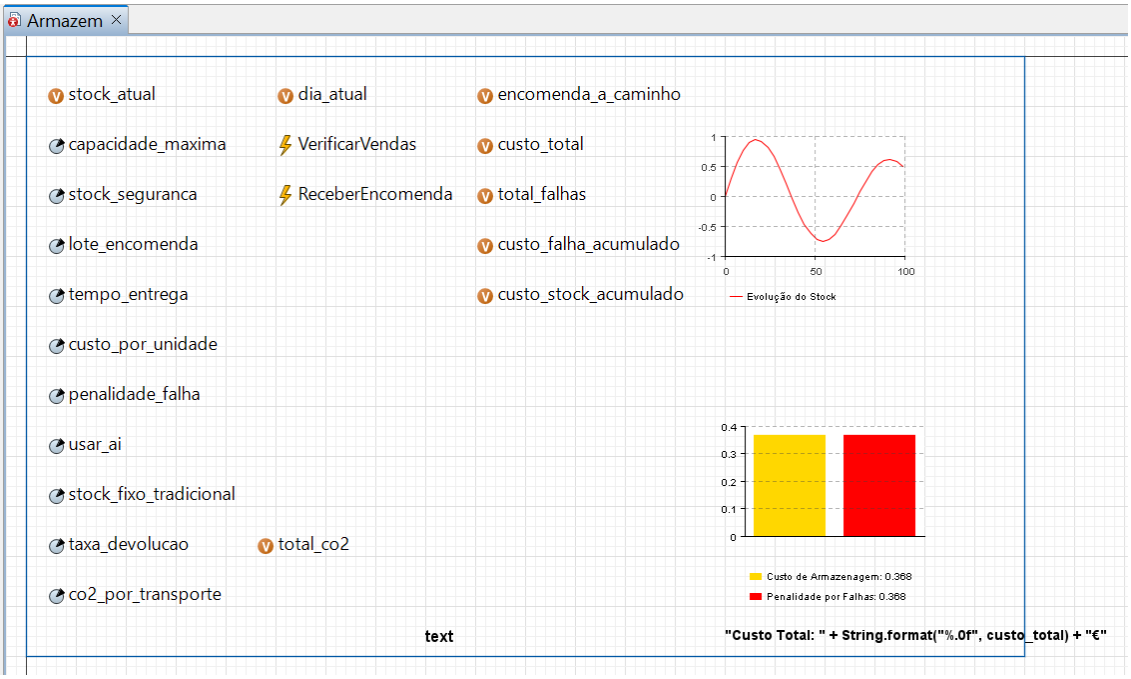
A simulação operacional foi desenvolvida com o objetivo de avaliar o impacto prático das previsões de procura na gestão diária de inventário de um armazém de *e-Commerce*, comparando uma política tradicional de reabastecimento com uma política dinâmica suportada por previsões geradas por ML. Desenvolveu-se um ambiente de simulação no *software* AnyLogic*,* onde as previsões gradas são testadas em cenários dinâmicos de gestão de inventário e simulação.

3.5.1. Arquitetura e Lógica do Modelo de Simulação

A simulação foi desenvolvida em ambiente Anylogic*,* utilizando modelação ABS. A Figura 8 apresenta o diagrama lógico do modelo de simulação operacional desenvolvido em AnyLogic para representar a gestão diária de inventário de um armazém de *e-Commerce*. O modelo encontra-se centrado no agente Armazem, que concentra a lógica de decisão e controlo do sistema, funcionando como a entidade responsável por monitorizar o *stock* disponível, processar a procura observada, emitir ordens de reabastecimento e acumular os indicadores de desempenho económico e ambiental. A dinâmica do modelo decorre em tempo discreto, com periodicidade diária, sendo acionada pelo evento cíclico VerificarVendas, executado a cada 24 horas. Em cada ciclo, o simulador começa por verificar a existência de encomendas pendentes cujo *lead time* tenha sido concluído, atualizando o *stock* com as respetivas entradas de mercadoria. De seguida, é processada a procura do dia, descontando-se ao inventário disponível a quantidade efetivamente vendida. Quando a procura excede o *stock* disponível, o sistema regista uma rutura de *stock*, gerando simultaneamente a penalização financeira associada à falha de serviço. Após a atualização das saídas do armazém, o modelo verifica a necessidade de reabastecimento com base no ponto de encomenda definido para cada cenário. É neste momento que a Figura 8 representa a principal lógica de decisão do estudo: no cenário tradicional, o sistema opera com um limiar fixo de segurança, simulando uma política estática e reativa de inventário; no cenário com IA, o ponto de encomenda é calculado dinamicamente em função da previsão diária gerada pelo modelo XGBoost, permitindo ajustar o *stock* de segurança à procura esperada. Sempre que o *stock* disponível é inferior ou igual ao ponto de encomenda, é emitida uma nova ordem ao fornecedor, com quantidade fixa previamente parametrizada. Esta ordem não entra imediatamente em *stock*, ficando sujeita ao atraso de abastecimento assumido no modelo. Paralelamente, o simulador acumula os custos de armazenagem, os custos de rutura e as emissões de CO_2 associadas quer ao transporte de abastecimento, quer à logística inversa. Desta forma, a Figura 8

não representa apenas um fluxograma técnico, mas sim a tradução visual da lógica através da qual as previsões de procura são convertidas em decisões operacionais, permitindo comparar, num ambiente controlado, o impacto económico e ambiental de diferentes políticas de inventário.

Figura 8 - Diagrama lógico da simulação de gestão de inventário no AnyLogic.



Fonte: Elaboração Própria.

3.5.2. Integração e Fluxo de Dados

A ponte tecnológica entre a fase analítica (Modelação Preditiva) e a fase operacional (Simulação) foi estabelecida através de ficheiros de intercâmbio de dados. No arranque da simulação (*Startup Code*), o AnyLogic executa um *script Java* que limpa as tabelas internas e importa vetores de venda reais e previsões geradas pelo modelo preditivo. Esta integração permite que o simulador teste a eficácia de decisões baseadas em dados previstos contra o comportamento real do mercado.

3.5.3. Dinâmica de Gestão de Inventários e Cenários de Decisão

A essência da investigação reside na comparação entre dois cenários operacionais através da variável booleana "usar_ai":

- **Cenário Tradicional ("usar_ai" = *false*):** O modelo opera com um *Stock* de Segurança Fixo, independente da variação da procura, simulando uma gestão de inventário estática.

- **Cenário de IA ("usar_ai" = *true*):** O modelo implementa um *Stock* de Segurança Dinâmico, calculado como o dobro da previsão gerada pela IA para o dia em curso ($Stock_{Seguranca} = Prevsao_{IA} \times 2.0$). Esta lógica permite que o armazém se adapte proactivamente a picos de procura, reduzindo simultaneamente o excesso de *stock* em dias de baixa atividade.

3.5.4. Modelo de Sustentabilidade e Cálculo de Emissões (CO_2)

O modelo quantifica o impacto ambiental de operação logística através de dois fluxos de emissões:

- **Emissões de Abastecimento:** Soma de emissões fixas ("co2_por_transporte") sempre que ocorre a chegada de uma encomenda de fornecedores ao armazém.
- **Emissões de Logística Inversa:** Implementação de uma lógica de devoluções baseada numa taxa de devolução de 5%, um valor extraído da análise histórica do *dataset* de sustentabilidade (faturas de devolução e perdas). Cada unidade devolvida gera um custo ambiental de transporte de retorno (calculado como 0.1kg de CO_2 por produto), permitindo medir como a precisão da IA pode mitigar as emissões associadas a movimentos logísticos desnecessários. O valor 0.1kg de CO_2 é de referência baseado em estimativas de mercado para a pegada de carbono logística urbana de última milha, adaptado para refletir o impacto unitário de cada item devolvido.

3.5.5. Parametrização da Simulação

O modelo compara uma política de revisão contínua tradicional com uma abordagem dinâmica baseada nas previsões do modelo com melhor resultado. A Tabela 6 define a formulação matemática do modelo. Para a execução da simulação no AnyLogic, estas variáveis assumiram os valores empíricos e lógicas de cálculo detalhadas na Tabela 7, refletindo a realidade operacional do caso de estudo. Os valores definidos visam representar de forma fidedigna as restrições físicas do armazém e as penalizações financeiras aplicáveis aos cenários testados.

Como os dados reais da capacidade física não estavam disponíveis, os limites de armazenamento foram inferidos com base no volume médio da procura diária, garantindo a viabilidade matemática da simulação.

Tabela 6 - Notação matemática e variáveis do modelo de simulação

Símbolo	Tipo	Definição	Unidade
t	Índice temporal	Dia da simulação	dias
T	Parâmetro	Horizonte total da simulação	dias
S_t	Variável de estado	*Stock* disponível no armazém no dia (t)	unidades
D_t	Entrada	Procura real observada no dia (t)	unidades
$\widehat{D}_t$	Entrada	Previsão da procura no dia (t)	unidades
ROP_t	Decisão	Ponto de encomenda no dia (t)	unidades
Q	Parâmetro	Lote fixo de encomenda	unidades
L	Parâmetro	Lead time de abastecimento	dias
O_t	Decisão	Quantidade encomendada no dia (t)	unidades
h	Parâmetro	Custo de posse por unidade e por dia	€/unidade/dia
c_t	Parâmetro	Custo de rutura por unidade não satisfeita	€/unidade
e_t	Parâmetro	Emissão de CO_2 por transporte de abastecimento	kg CO_2/viagem
e_r	Parâmetro	Emissão de CO_2 por unidade devolvida	kg CO_2/unidade
C^{total}	Saída	Custo total acumulado da operação	€
E^{total}	Saída	Emissões totais acumuladas	kg CO_2

Tabela 7 - Parâmetros e Pressupostos Matemáticos do modelo de simulação

Parâmetro Operacional	**Valor / Lógica Aplicada**	**Justificação no Contexto do Estudo**
Unidade de Tempo	Dias	Frequência de atualização de vendas e recálculo de inventário.
Capacidade Máxima do Armazém	500 000 unidades	Valor assumido para garantir a viabilidade logística, suportando os picos de procura diária (que podem ultrapassar as 100 000 unidades) e a receção dos lotes de encomenda sem rutura física do espaço.
Lote de Encomenda (Q)	200 000 unidades	Quantidade fixa encomendada ao fornecedor sempre que o *stock* desce do limite mínimo. Cobre aproximadamente 2 a 4 dias de procura padrão.
Ponto de Encomenda (Cenário IA)	ROP (*Reorder Point*) = Previsão diária x 2	Calculado dinamicamente: a previsão gerada pelo modelo com melhor resultado é multiplicada pelos 2 dias de *Lead Time*, garantindo que o *stock* de segurança cobre exatamente a procura esperada até à chegada do camião.
Custo de Posse (*Holding Cost*)	0.50 € / unidade / dia	Custo assumido por manter o produto armazenado. Otimizar este custo é um dos objetivos da IA.

Parâmetro Operacional	**Valor / Lógica Aplicada**	**Justificação no Contexto do Estudo**
Custo de Rutura (*Stockout Cost*)	10 € / unidade	Penalização financeira aplicada quando o *stock* é zero. O valor elevado reflete a perda de margem de lucro e a insatisfação do cliente.
Emissões de CO_2 por transporte (Pegada de Carbono)	50 kg de CO_2 por viagem/encomenda	Representa o impacto ambiental de cada deslocação do camião de abastecimento. Ao utilizar o modelo com melhor resultado para prever a procura com maior precisão, o sistema evita encomendas desnecessárias. A redução do número total de viagens traduz-se diretamente na minimização da pegada de carbono da cadeia de abastecimento, promovendo a sustentabilidade logística.
Taxa de Devolução	5% (0.05)	Modela a percentagem de inventário que é devolvida, que expira ou que sofre danos no armazém. Numa ótica de sustentabilidade, manter níveis de *stock* excessivos (como no cenário tradicional) gera um maior volume absoluto de desperdício físico. A IA, ao ajustar o *stock* à procura real minimiza o desperdício de materiais, alinhando a operação com práticas logísticas mais circulares.

A dinâmica diária do modelo de simulação decorre de forma iterativa, seguindo sequencialmente os seguintes passos lógicos:

1. No início do dia, verifica-se a chegada de encomendas pendentes (cumprimento do *lead time* de abastecimento).
2. Atualiza-se o *stock* disponível com as novas entradas de mercadoria.
3. Desconta-se a procura real observada nesse dia.
4. Caso o *stock* seja insuficiente para cobrir a procura, regista-se a rutura de inventário e calcula-se a respetiva penalidade financeira.
5. Verifica-se se o nível de *stock* atual desceu abaixo do Ponto de Encomenda (ROP). Em caso afirmativo, emite-se uma nova encomenda ao fornecedor.
6. Por fim, atualizam-se os indicadores globais, nomeadamente os custos acumulados (armazenagem e penalizações) e as emissões de CO2 resultantes da operação.

O cálculo do ROP e do *Stock* de Segurança foi adaptado dinamicamente para cada cenário. No Cenário de IA, o ROP é definido pela fórmula $ROP = \widehat{D}_t \times L$, onde $\widehat{D}_t$ representa a previsão diária gerada pelo modelo e L corresponde ao *lead time* de 2 dias. Esta lógica garante que o *stock* de segurança cobre exatamente a procura esperada até à chegada do novo lote. No Cenário Tradicional, por oposição, este valor permanece estático, baseando-se em médias históricas fixas, o que impossibilita a adaptação a picos repentinos de procura.

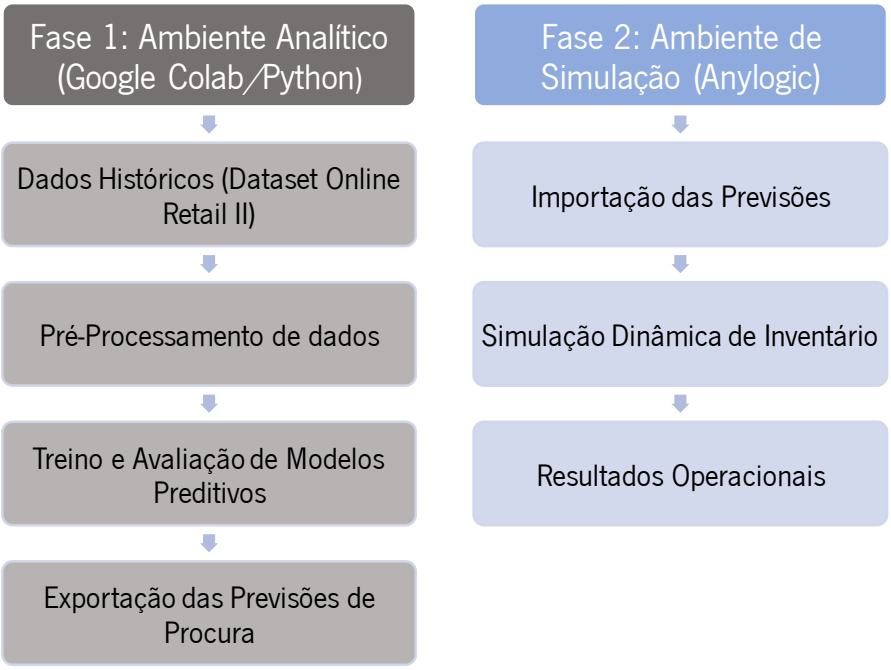
3.6. Ferramentas e tecnologias utilizadas

Nesta investigação foi utilizada a linguagem Python, para toda a fase de limpeza, análise exploratória e implementação dos modelos preditivos (ARIMA, XGBoost e LSTM), recorrendo a bibliotecas como Pandas para manipulação de dados e TensorFlow/Keras para as redes neurais. Foi usado Google Colab, ambiente de desenvolvimento em nuvem que permitiu o processamento computacional necessário para o treino dos modelos.

Para além disso, utilizou-se o programa AnyLogic, software de simulação utilizado para modelar o ambiente logístico e testar o impacto das previsões em cenários de *stock* real e sustentabilidade. Foi utilizada a versão "*Professional/University*" do programa Anylogic e integrou-se a linguagem *Java* para o processamento de dados externos.

Para clarificar a integração entre as diferentes tecnologias adotadas, a Figura 9 ilustra a arquitetura metodológica do sistema desenvolvido.

Figura 9 - Esquema conceptual e integração das ferramentas e tecnologias utilizadas.



Fonte: Elaboração Própria.

4. ANÁLISE DE DADOS

4.1. Desempenho dos Modelos de Previsão

Ao longo desta secção, apresenta-se a avaliação comparativa dos modelos de previsão de procura desenvolvidos. Para tal, o desempenho dos modelos estatísticos tradicionais (ARIMA e SARIMA) foi contrastado com as abordagens de ML (XGBoost) e DL (LSTM). A avaliação suportou-se em três métricas principais, RMSE, que penaliza grandes desvios e picos falhados, MAE, que quantifica a margem de erro diária média, e, o Coeficiente de Determinação (R^2), que indica a proporção de variância explicada pelo modelo.

4.1.1. Resultados Quantitativos

Os resultados obtidos de cada modelo de previsão estão sintetizados na tabela 8. É de denotar que o modelo XGBoost destacou-se como mais robusto globalmente, apresentando melhor capacidade de adaptação à variância, refletida no valor de R^2.

Tabela 8 - Métricas de Desempenho dos Modelos Preditivos

MODELO	RMSE (£)	MAE (£)	R^2
ARIMA (1,1,1)	41 163.88	28 875.07	-0.0909
SARIMA (SAZONAL)	31 603.31	19 013.36	0.3570
SARIMA AUTO	32 872.41	17 240.06	0.3043
LSTM	31 055.55	**15 193.00**	0.3791
XGBOOST	**28 287.72**	15 774.57	**0.4848**

Neste contexto logístico e de *e-Commerce*, importa salientar que erros de previsão de grande magnitude são significativamente mais críticos para a operação do que pequenos desvios médios diários. Uma falha considerável na previsão pode resultar numa rutura de *stock* severa (com perda de vendas e insatisfação do cliente) ou num excesso de inventário dispendioso. Por este motivo, a métrica RMSE foi privilegiada na avaliação e seleção do modelo ótimo, uma vez que

a sua formulação matemática eleva os erros ao quadrado, penalizando de forma mais acentuada e rigorosa os desvios de maior dimensão.

A análise dos resultados anteriores demonstra a existência de diferenças substanciais entre a capacidade preditiva das abordagens clássicas e dos algoritmos avançados, validando a necessidade de incorporar técnicas de IA na gestão de inventário do *e-Commerce*.

O modelo ARIMA (1,1,1) obteve o pior desempenho de todos os modelos testados, com o erro mais elevado (RMSE de £41 163.88) e um R^2 negativo. Este último valor indica que o modelo conseguiu ser pior do que simplesmente prever a média constante de vendas. O resultado deste modelo deve-se à incapacidade de capturar a não-linearidade e os picos abruptos típicos do *e-Commerce*.

Ao introduzir a componente sazonal no modelo ARIMA (SARIMA) houve uma melhoria considerável do seu desempenho, passando a um R^2 de 0.3570, provando que os ciclos temporais têm um peso significativo no comportamento dos consumidores. Contudo, o erro manteve-se elevado.

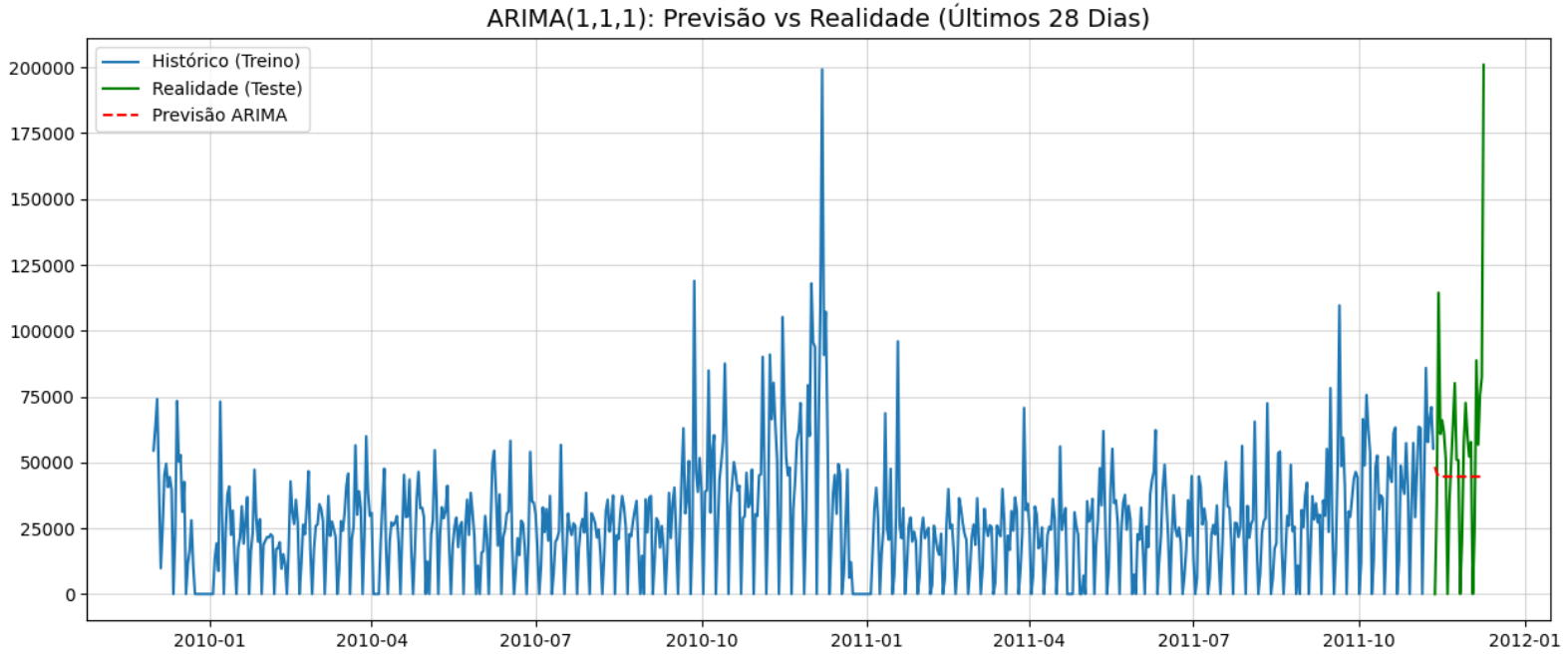
Ao transitar para os modelos de IA, verificou-se uma redução dos erros. O modelo LSTM obteve um MAE de £15 193.00, o que significa que, num dia "normal", as suas previsões estão ligeiramente mais próximas do valor real.

No entanto, no contexto de gestão de cadeia de abastecimento, a penalização de grandes erros preditivos é crítica, pois resulta em rutura de *stock*. É aqui que o modelo XGBoost vence, registando um RMSE mais baixo e, provando ser muito mais resistente e preciso nos picos de vendas face à rede neuronal. Adicionalmente, conseguiu o melhor ajuste global da experiência, explicando quando 50% de toda a volatilidade do sistema (R^2 de 0.4848). Dada a sua superioridade global na mitigação de erros graves e captura de variância, o modelo XGBoost foi o algoritmo selecionado para alimentar a simulação de inventário no Anylogic.

Para complementar a avaliação quantitativa das métricas de erro, procedeu-se à análise visual do ajustamento dos modelos através de gráficos de séries temporais.

A Figura 10 ilustra o comportamento do modelo ARIMA(1,1,1) face à procura real nos últimos 28 dias de teste. Como é possível observar, a linha de previsão apresenta um comportamento linear, falhando por completo na antecipação dos picos de vendas abruptos.

Figura 10 - Comparação entre a procura real e a previsão do modelo ARIMA(1,1,1), evidenciando a incapacidade de captura de picos de vendas.



Fonte: Elaboração Própria.

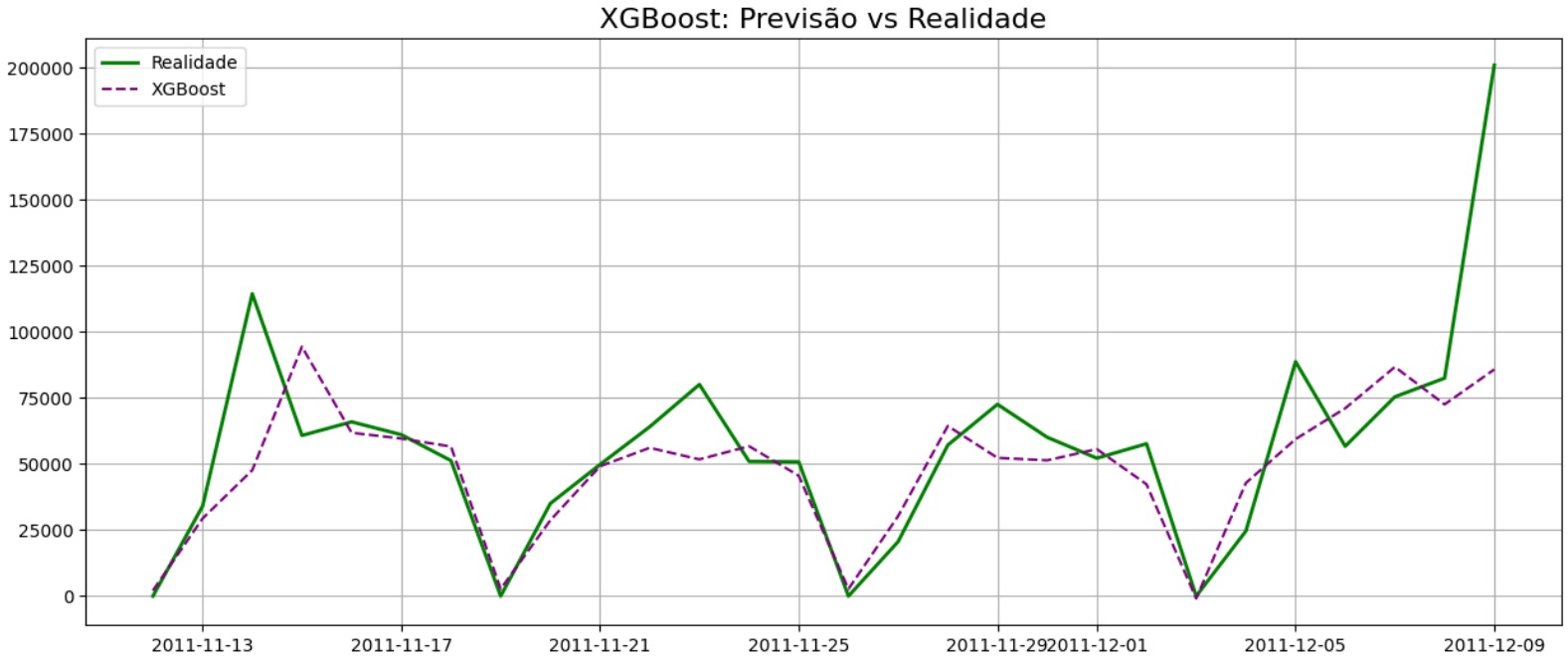
Numa perspetiva de operação logística, a adoção destas previsões resultaria numa subavaliação grave do *stock* de segurança, provocando ruturas de inventário nos dias de maior procura.

A Figura 11 ilustra o comportamento das previsões geradas pelo modelo vencedor (XGBoost) em sobreposição com a procura real observada durante o período de teste.

É possível observar que a linha de previsão acompanha de forma fiel a tendência central e a sazonalidade dos dados reais. Embora existam desvios naturais em dias de picos extremos de vendas (anomalias de procura), o modelo demonstra uma elevada capacidade de adaptação rápida, não desfasando a previsão nos dias subsequentes. Esta validação visual reforça a fiabilidade do modelo para ser integrado como motor de decisão no planeamento logístico, uma vez que a margem de erro visualizada não compromete a gestão macro do inventário.

(Nota: Os gráficos de ajustamento visual para os restantes modelos testados (LSTM e SARIMA) encontram-se disponíveis para consulta, ver Figuras 17, 18 e 19 no Apêndice H.)

Figura 11 - Comparação gráfica entre a procura real e as previsões do modelo XGBoost durante o período de teste.



Fonte: Elaboração Própria.

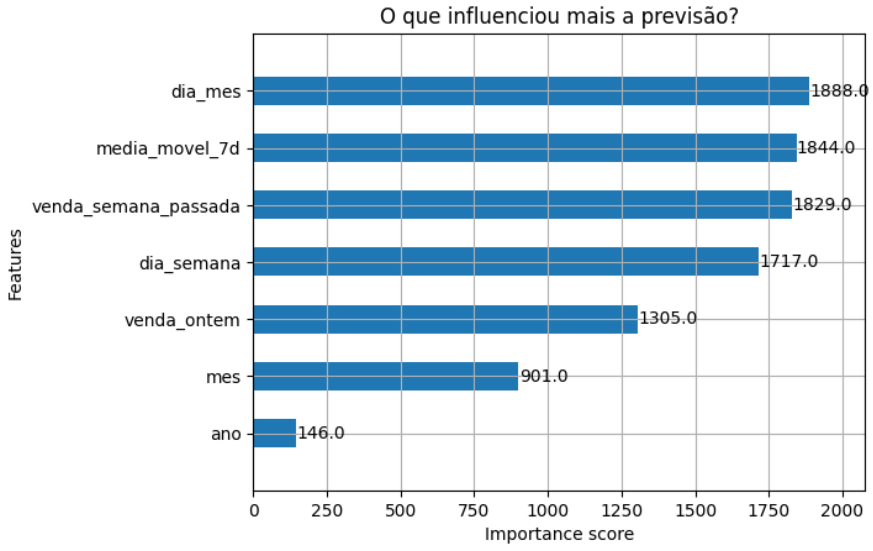
4.2. Interpretação e Explicabilidade dos Modelos

Apesar de algoritmos de ML como o XGBoost apresentarem uma capacidade preditiva superior, são frequentemente criticados por funcionarem como "caixas negras" (*black boxes*), ou seja, sabe-se que dados entram (*inputs*) e vê-se a previsão final que sai (*outputs*), mas não se consegue compreender o processo matemático interno que gerou essa resposta. No contexto da gestão da cadeia de abastecimento, a transparência na tomada de decisão é fundamental, pois os gestores logísticos precisam de compreender os motivos que levam o sistema a sugerir um determinado nível de *stock*. Para garantir esta interpretabilidade, analisou-se a importância das variáveis do modelo vencedor.

4.2.1. Importância das Variáveis

A avaliação da importância das características (*features*) permite identificar quais as métricas que tiveram o maior peso matemático nas decisões do XGBoost. A análise do modelo revelou que o algoritmo priorizou uma combinação de fatores cíclicos temporais e métricas de tendência histórica. O gráfico abaixo (Figura 12) ilustra o *ranking* de importância, destacando-se quatro variáveis principais na base da construção das previsões, o dia do mês, a média de vendas dos últimos sete dias, o valor de vendas no mesmo dia da semana anterior e o dia da semana.

Figura 12 - Gráfico de importância das variáveis (Feature Importance), destacando os preditores temporais e históricos com maior impacto nas previsões de procura do modelo XGBoost. Nota: A variável 'trimestre' foi automaticamente omitida pelo algoritmo por apresentar um score de importância nulo.



Fonte: Elaboração Própria.

Ao perceber quais as variáveis que mais influenciam a previsão do modelo, comprova-se que a procura no *e-Commerce* não é um fenómeno aleatório, sendo fortemente governada por padrões de comportamento do consumidor a curto e médio prazo. Ao analisar o impacto destas variáveis é justificável a superioridade do modelo de IA face aos métodos estatísticos clássicos.

O dia do mês é a variável mais impactante no modelo, revelando a existência de um forte ciclo mensal no poder de compra e no comportamento dos clientes. Esta variável permite ao modelo antecipar flutuações relacionadas com a injeção de liquidez no mercado, por exemplo, a receção de salários no final ou no início do mês, ou, com o pagamento de despesas fixas das famílias, adaptando os níveis de inventário para os picos de consumo mensais.

A segunda variável mais relevante é a média móvel de 7 dias. Ao contrário de modelos lineares que demoram a reagir a mudanças súbitas, a média móvel garante que o modelo XGBoost entenda rapidamente de a procura global da loja está a aumentar ou a diminuir. Por exemplo, se uma campanha de marketing ou um evento externo aumentar subitamente as vendas numa terça-feira, a média móvel informa o modelo para manter as previsões altas nos dias seguintes.

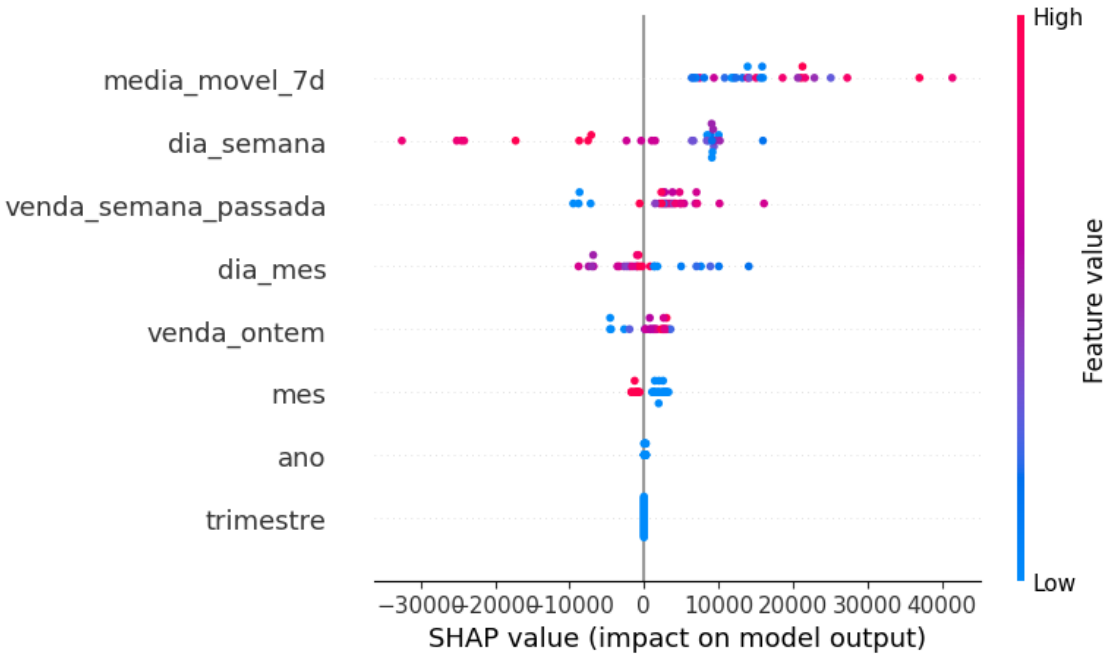
Ao incluir a variável da venda da semana passada, o modelo consegue capturar de forma pura a sazonalidade intra-semanal. O setor do retalho online possui padrões de repetição muito fiéis, ou seja, as segundas-feiras comportam-se tipicamente como as segundas-feiras anteriores. Isto dá ao modelo uma base preditiva sólida e segura, atuando como uma âncora histórica.

A distinção do dia da semana justifica-se pelas diferentes dinâmicas de navegação e compra entre os dias úteis e o fim de semana. Compreender matematicamente esta diferença é vital não só para a previsão teórica, mas para a operação logística, permitindo otimizar o planeamento de alocação de recursos em armazém e a programação de transportes de acordo com o dia específico.

Resumindo, esta capacidade de cruzar a sazonalidade (dia da semana e a venda na semana passada) com a flexibilidade da tendência atual (média móvel) e os ciclos de rendimento mensal (dia do mês) são o que tornam o modelo XGBoost mais eficaz.

Para conferir interpretabilidade ao modelo XGBoost, aplicou-se a metodologia SHAP (Figura 13). A análise deste gráfico permite perceber não só a importância, mas também a direção do impacto de cada variável. Observa-se que a "*media_movel_7d*" exerce o maior impacto positivo na procura: os seus valores mais altos (a vermelho) empurram a previsão significativamente para cima. Em contraste, a variável "*dia_semana*" apresenta um forte efeito negativo nos seus valores mais elevados (que correspondem ao final da semana), reduzindo a previsão. Este comportamento, aliado ao impacto positivo do histórico de vendas recentes, valida que o algoritmo captou com sucesso as dinâmicas operacionais e de sazonalidade do negócio.

Figura 13 - Gráfico de resumo SHAP indicando o impacto e a direção das variáveis na previsão do modelo XGBoost. Nota: Ao contrário da Figura 12, a variável 'trimestre' é aqui incluída para fins de exaustividade metodológica, demonstrando graficamente uma distribuição de impacto estritamente nula.



Fonte: Elaboração Própria.

4.3. Simulação Operacional

De forma a quantificar o impacto prático, tanto financeiro como ambiental, da adoção do modelo de IA na gestão diária de um armazém de *e-Commerce,* procedeu-se à transição das previsões teóricas para um ambiente de simulação logística.

4.3.1. Configuração dos Cenários Simulados

A simulação foi configurada para correr durante um período de teste de 28 dias. Estabeleceram-se dois cenários de atuação, o cenário tradicional ("usar_ia" = *false*) e o cenário de IA ("usar_ia" = *true*). O cenário tradicional representa a abordagem clássica, onde a política de inventário não utiliza os modelos preditivos avançados, baseando-se em parâmetros estáticos e reativos face à procura. No cenário de IA, a política de inventário utiliza as previsões diárias do modelo vencedor (XGBoost), antecipando ativamente os picos e quedas de vendas.

4.3.2. Resultados Operacionais

Procedeu-se à extração das métricas de desempenho utilizadas na simulação operacional. Na tabela 9 estão sintetizados os custos operacionais incorridos em cada cenário.

Tabela 9 - Comparação de custos operacionais entre os cenários simulados.

Métrica	Cenário 1 (Tradicional)	Cenário 2 (IA - XGBoost)	Variação
Custo de Armazenagem (Posse)	628 262,56 €	972 017,55 €	+ 54,7%
Custo de Rutura (Penalidade)	8 257 993,39 €	3 133 093,21 €	- 62,0%
Custo Total da Operação	**8 886 256,00 €**	**4 105 111,00 €**	**- 53,8%**

Fonte: Elaboração própria com base nos resultados da simulação.

Os resultados evidenciam um *trade-off* clássico da gestão de inventário: a redução do risco de rutura exige maior disponibilidade preventiva de *stock*. No entanto, no contexto analisado, o custo adicional de armazenagem revelou-se substancialmente inferior ao custo evitado em penalizações por falhas de serviço, conduzindo a uma redução líquida do custo total da operação.

No cenário 1, o armazém é incapaz de prever os picos de procura, consequentemente gerou ruturas catastróficas de *stock,* que por sua vez implicou um custo de penalidade superior a 8,2 milhões de euros. Por outro lado, no cenário 2, o modelo de ML, XGBoost, antecipa os picos de procura. O sistema encomendou mais inventário preventivo, o que gerou um aumento dos

custos de armazenagem em 54,7%. Contudo, ao investir em mais inventário, houve uma redução drástica de 62% nas penalidades por falhas de serviço.

A Figura 15 evidencia uma trajetória de stock fortemente instável no cenário tradicional, com múltiplos episódios de aproximação ou descida a níveis críticos, traduzindo a incapacidade do sistema em responder atempadamente aos picos de procura. Em contraste, a Figura 14 mostra uma evolução mais estável do inventário no cenário com IA, resultado da antecipação gerada pelas previsões do modelo XGBoost. Embora esta política implique níveis médios de *stock* mais elevados, reduz significativamente a frequência e severidade das ruturas.

Figura 14 - Evolução de stock do armazém de IA.

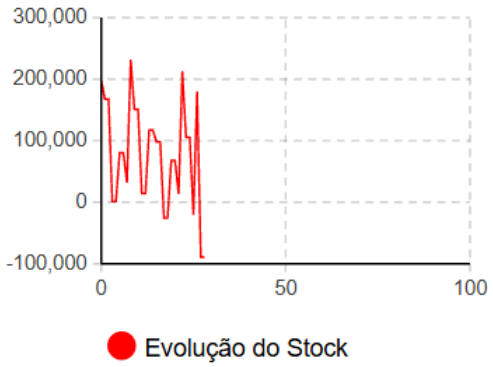
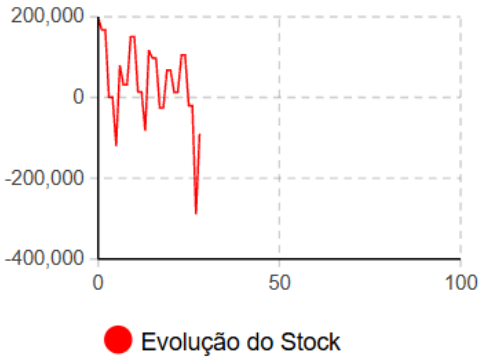


Figura 15 - Evolução de stock do armazém do cenário tradicional.



Fonte: Elaboração Própria.

De uma forma geral, ao introduzir as previsões do modelo de ML reduziu-se o custo total da operação para menos de metade, implicando uma poupança de aproximadamente 4,7 milhões de euros num único mês.

4.4. Avaliação do Impacto Sustentável

Apesar do impacto económico ser necessário para avaliação, a logística moderna exige que as operações sejam avaliadas pelo seu impacto ambiental e eficiência no uso de recursos. Para que tal fosse avaliado, a simulação foi programada para cruzar os indicadores da pegada de carbono com a análise histórica do desperdício logístico.

4.4.1. Redução Estimada de Emissões de CO_2

Na tabela 10 estão registados os valores gerados por cada de cenário de kg de CO_2. Esta análise revelou um cenário de estabilidade, sem penalização significativa na pegada de carbono direta da operação modelada.

Tabela 10 - Valores de kg de CO2 na simulação operacional.

	Cenário 1 (Tradicional)	**Cenário 2 (IA-XGBoost)**
Kg de CO_2	8 139,0	8 189,0

Fonte: Elaboração Própria com base nos resultados da simulação.

Através da análise destes valores, verificou-se um agravamento residual de apenas 0,6%, que correspondem a um aumento de 50 kg de CO_2 no cenário preditivo. Este aumento correlaciona-se diretamente com o aumento do volume de inventário movimentado e armazenado de forma preventiva para mitigar as ruturas. O modelo de simulação adotou um custo carbónico fixo por transporte, o que significa que o ligeiro agravamento reflete a maior frequência de reabastecimento preventivo ordenada pelo modelo XGBoost.

Concluindo, a transição para uma política preventiva permite alcançar uma eficiência económica massiva, uma poupança superior a 53% dos custos operacionais, mantendo a neutralidade nos indicadores ambientais diretos da instalação logística.

5. DISCUSSÃO DOS RESULTADOS

Este capítulo sintetiza e interpreta os principais achados desta dissertação, validando as hipóteses iniciais. A primeira fase do estudo confirmou que os modelos estatísticos clássicos, como o ARIMA, são insuficientes para capturar as flutuações abruptas da procura no retalho *online*. Em contraste, o modelo XGBoost demonstrou uma capacidade superior de ajustamento à volatilidade real, alcançando o menor erro de previsão.

A transição destas previsões para a simulação operacional no *software* AnyLogic confirmou o impacto prático do modelo. Provou-se que a adoção da IA transforma a gestão de inventário de um paradigma reativo para um sistema preventivo. No cenário tradicional, a incapacidade de antecipar a procura gerou múltiplas ruturas de *stock*, elevando os custos operacionais para mais de 8,8 milhões de euros. Com a integração do XGBoost, o armazém encomendou inventário de forma preditiva, estabilizando a operação. Esta abordagem reduziu drasticamente as penalidades por falhas de serviço, culminando numa poupança massiva de 4,7 milhões de euros (uma redução de 53,8% nos custos totais), alcançada com um *trade-off* carbónico estatisticamente negligenciável (um aumento residual de apenas 50 kg de CO2).

6. CONTRIBUTOS TEÓRICOS E PRÁTICOS DA INVESTIGAÇÃO

Ao provar a viabilidade no mundo real, através de uma simulação operacional, dos algoritmos matemáticos, é gerado valor académico e também para a indústria. Do ponto de vista académico, esta investigação tenta combater a lacuna existente entre a modelação preditiva e a simulação. Demonstra-se que a avaliação de um modelo preditivo não se deve limitar à minimização do erro matemático, mas sim estender-se à quantificação do seu impacto financeiro e sustentável num ambiente logístico dinâmico.

A nível prático e empresarial, prova-se que a IA não é apenas um custo tecnológico, mas um investimento com retorno financeiro imediato e massivo. Ao mitigar as ruturas de *stock*, a implementação de IA estabiliza a cadeia de abastecimento, protege a imagem da marca junto do consumidor e reduz o desperdício oculto associado a ineficiências logísticas.

7. LIMITAÇÕES E SUGESTÕES PARA INVESTIGAÇÃO FUTURA

Apesar do rigor aplicado, a investigação apresenta algumas limitações que devem ser consideradas na interpretação dos resultados.

O conjunto de dados "*Online Retail II*", embora real e extenso, refere-se ao período de 2009-2011. Desta forma, não captura alterações recentes no comportamento do consumidor pós-pandemia ou o impacto de novas tecnologias logísticas.

O estudo focou-se em variáveis endógenas ao conjunto de dados (preço, quantidade, tempo), não integrando variáveis exógenas como condições meteorológicas ou dados de redes sociais, que poderiam aumentar a precisão dos modelos de DL em cenários de extrema volatilidade.

Para além disso, o modelo de simulação assume condições controladas de transporte e prazos de entrega constantes, o que pode variar em cenários de disrupção logística real. O modelo assumiu no cálculo estático das emissões de CO_2 um custo fixo por transporte, não sendo diretamente sensível a variáveis como o peso exato de carga expedida, as rotas específicas dos veículos ou a eficiência energética do armazém real.

Para futuras investigações, sugere-se a integração destes modelos preditivos com APIs (*Application Programming Interfaces*) de dados dinâmicos, incorporando variáveis como tráfego do *website*, tendências de redes sociais ou campanhas de *marketing* no treino do XGBoost. Recomenda-se, ainda, a expansão da simulação para a fase de *Last-Mile Delivery* (entregas de última milha), de modo a quantificar com maior exatidão o impacto ambiental e financeiro da roteirização com base em IA.

REFERÊNCIAS BIBLIOGRÁFICAS

Badakhshan, E., Mustafee, N., & Bahadori, R. (2024). Application of simulation and machine learning in supply chain management: A synthesis of the literature using the Sim-ML literature classification framework. *Computers & Industrial Engineering*, *198*, 110649. https://doi.org/10.1016/j.cie.2024.110649

Brooks, C. (2019). *INTRODUCTORY ECONOMETRICS FOR FINANCE* (4th ed.). https://doi.org/10.1017/9781108524872

Cano, J. A., Londoño-Pineda, A., & Rodas, C. (2022). Sustainable Logistics for E-Commerce: A Literature Review and Bibliometric Analysis. *Sustainability*, *14*(19), Artigo 19. https://doi.org/10.3390/su141912247

Cavalcante, I. M., Frazzon, E. M., Forcellini, F. A., & Ivanov, D. (2019). A supervised machine learning approach to data-driven simulation of resilient supplier selection in digital manufacturing. *International Journal of Information Management*, *49*, 86–97. https://doi.org/10.1016/j.ijinfomgt.2019.03.004

Chen, D. (2019). *Online Retail II* [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5CG6D

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, X., He, Y., Hooshmand Pakdel, G., Liu, X., & Wang, S. (2025). A comprehensive multi-stage decision-making model for supplier selection and order allocation approach in the digital economy. *Advanced Engineering Informatics*, *63*, 102961. https://doi.org/10.1016/j.aei.2024.102961

Chen, Y., Zhang, H., Yan, X., & Miao, Q. (2025). Supply chain demand forecasting based on multi-time scale data fusion network. *Computers & Industrial Engineering*, *207*, 111324. https://doi.org/10.1016/j.cie.2025.111324

Dasoomi, M., Naderan, A., & Allahviranloo, T. (2023). A novel hybrid machine learning model for shopping trip estimation: A case study of Tehran, Iran. *Transportation Engineering*, *14*, 100218. https://doi.org/10.1016/j.treng.2023.100218

Datta, A., Dey, B. K., Bhuniya, S., Sangal, I., Mandal, B., Sarkar, M., Guchhait, R., Sarkar, B., & Ganguly, B. (2025). Adaptation of e-commerce retailing to enhance customer satisfaction within a dynamical system under transfer of risk. *Journal of Retailing and Consumer Services*, *84*, 104129. https://doi.org/10.1016/j.jretconser.2024.104129

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association*, *74*(366), 427–431. https://doi.org/10.2307/2286348

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Gao, Y., Lu, S., Cheng, H., & Liu, X. (2024). Data-driven robust optimization of dual-channel closed-loop supply chain network design considering uncertain demand and carbon cap-and-trade policy. *Computers & Industrial Engineering*, *187*, 109811. https://doi.org/10.1016/j.cie.2023.109811

He, P., Wang, T.-Y., Mardani, A., Wang, X.-J., & Chen, Z.-S. (2024). Selling mode selection and AI service strategy in an E-commerce platform supply chain. *Computers & Industrial Engineering*, *197*, 110560. https://doi.org/10.1016/j.cie.2024.110560

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice (2nd ed)*. https://otexts.com/fpp2/

Joseph, R. V., Mohanty, A., Tyagi, S., Mishra, S., Satapathy, S. K., & Mohanty, S. N. (2022). A hybrid deep learning framework with CNN and Bi-directional LSTM for store item demand forecasting. *Computers and Electrical Engineering*, *103*, 108358. https://doi.org/10.1016/j.compeleceng.2022.108358

Leung, E. K. H. (2025). Total fulfillment management: Principles, practices and use cases. *Transportation Research Part E: Logistics and Transportation Review*, *194*, 103888. https://doi.org/10.1016/j.tre.2024.103888

Li, X., Shen, J., Wang, D., Lu, W., & Chen, Y. (2024). Multi-modal transform-based fusion model for new product sales forecasting. *Engineering Applications of Artificial Intelligence*, *133*, 108606. https://doi.org/10.1016/j.engappai.2024.108606

Malik, S., Muhammad, K., & Waheed, Y. (2024). Artificial intelligence and industrial applications-A revolution in modern industries. *Ain Shams Engineering Journal*, *15*(9), 102886. https://doi.org/10.1016/j.asej.2024.102886

Mohammed, A., Lopes de Sousa Jabbour, A. B., Zubairu, N., Chiappetta Jabbour, C. J., & Al Naabi, H. (2025). Food waste in the era of e-commerce: A novel farm-to-fork management methodology. *International Journal of Production Economics*, *285*, 109603. https://doi.org/10.1016/j.ijpe.2025.109603

Nie, T., He, J., Mei, Y., Qin, G., Li, G., Sun, J., & Ma, W. (2025). Joint estimation and prediction of city-wide delivery demand: A large language model empowered graph-based learning approach. *Transportation Research Part E: Logistics and Transportation Review*, *197*, 104075. https://doi.org/10.1016/j.tre.2025.104075

Peng, L., Wang, L., Xia, D., & Gao, Q. (2022). Effective energy consumption forecasting using empirical wavelet transform and long short-term memory. *Energy*, *238*, 121756. https://doi.org/10.1016/j.energy.2021.121756

Shaikh, A. A., Lakshmi, K. S., Tongkachok, K., Alanya-Beltran, J., Ramirez-Asis, E., & Perez-Falcon, J. (2022). Empirical analysis in analysing the major factors of machine learning in enhancing the e-business through structural equation modelling (SEM) approach. *International Journal of System Assurance Engineering and Management*, *13*(S1), 681–689. https://doi.org/10.1007/s13198-021-01590-1

Surur, F. M., Mamo, A. A., Gebresilassie, B. G., Mekonen, K. A., Golda, A., Behera, R. K., & Kumar, K. (2025). Unlocking the power of machine learning in big data: A scoping survey. *Data Science and Management*, *8*, 519–535. https://doi.org/10.1016/j.dsm.2025.02.004

Tunnicliffe Wilson, G. (2016). Time Series Analysis: Forecasting and Control,5th Edition, by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung, 2015. Published by John Wiley and Sons Inc., Hoboken, New Jersey, pp. 712. ISBN: 978-1-118-67502-1. *Journal of Time Series Analysis*, *37*, n/a-n/a. https://doi.org/10.1111/jtsa.12194

Vhatkar, M. S., Mahajan, P. S., Raut, R. D., Cheikhrouhou, N., & Ghoshal, S. (2025). Optimized hyperparameters for retail sales forecasting using grid search. *Engineering Applications of Artificial Intelligence*, *158*, 111472. https://doi.org/10.1016/j.engappai.2025.111472

Xu, X., Yue, R., Yang, B., & Li, Z. (2025). Integrating game theory and data-driven optimization models for online retailers with reusable packaging adoption. *Journal of Retailing and Consumer Services*, *84*, 104222. https://doi.org/10.1016/j.jretconser.2025.104222

Zietsman, H. J., & van Vuuren, J. H. (2023). A generic decision support framework for inventory procurement planning in distribution centres. *Computers & Industrial Engineering*, *177*, 109057. https://doi.org/10.1016/j.cie.2023.109057

Apêndice

A – PRISMA e Tabela de Extração de Dados

Tabela feita no Excel – "Tabela_Extração_Dados"

Tabela 11 - Tabela de Extração de Dados.

Autor	Título do Artigo	Ano	DOI	Contexto	Metodologia	Método de recolha de dados	Modelo(s) Usado(s)	Dados Utilizados	Métricas de Avaliação	Resultado Principal
Y. Chen et al.	A comprehensive multi-stage decision-making model for supplier selection and order allocation approach in the digital economy	2025	10.1016/j.aei.2024.102961	Previsão da procura para a alocação ótima de encomendas e a seleção de fornecedores	Comparação empírica de modelos de previsão baseados em séries temporais e machine learning	Dados históricos de vendas de produtos (lubrificantes) do setor automóvel	ARIMA, LSTM, CNN, BP, ELM, RF, SVM	Série temporal de vendas, preço, tipo de produto, obsolescência (modelo Weibull), etc.	MAPE, RMSE	LSTM superou ARIMA em precisão; inclusão da obsolescência via Weibull melhorou a previsão
Dasoomi et al.	A novel hybrid machine learning model for shopping trip estimation: A case study of Tehran, Iran	2023	10.1016/j.cie.2023.109057	Previsão do comportamento de compras (online/offline) para otimização de transporte e de sustentabilidade urbana	Modelo híbrido usando Gray Wolf Optimization (GWO) + CNN	Inquérito a 1000 consumidores de e-commerce em Teerão	CNN, LSTM, MLP, DT, KNN	Dados de hábitos de compra online/offline, variáveis demográficas, localização,	Accuracy, MSE	GWO-CNN superou todos os modelos comparados, com precisão de 97,81% e MSE = 0.325

							frequência de compras		
Zietsman & van Vuuren	A generic decision support framework for inventory procurement planning in distribution centres	2023	10.1016/j.treng.2023.100218	Planeamento de inventário em centros de distribuição com muitos SKUs e procura incerta	Framework genérico com previsão de procura e classificação de SKUs	Dados históricos de vendas, atributos de SKUs e encomendas anteriores	ARIMA, Exponential Smoothing, Croston, XGBoost, LightGBM, RNN, CNN, Transformers	Séries temporais de vendas por SKU, características do produto, dados de fornecedores	Não especificadas com valores, mas refere o RMSE, a MAE, e as métricas de serviço	Framework permite comparação de modelos (estatísticos vs ML) e seleção de métodos adequados por SKU
Datta et al.	Adaptation of e-commerce retailing to enhance customer satisfaction within a dynamical system under transfer of risk	2025	10.1016/j.jretconser.2024.104129	Otimização de preços e níveis de inventário em e-commerce com procura aleatória e canais duplos (online/offline)	Modelo matemático baseado no problema do vendedor de jornais com precificação dinâmica (Black-Scholes)	Modelagem analítica e simulação	Newsboy Problem, Geometric Brownian Motion, Black-Scholes	Procura estocástica, preço, nível ecológico do produto, esforço de vendas	Lucro esperado, custo, análise de sensibilidade	Estratégia combinada (preço dinâmico + produto verde + esforço de vendas) aumentou lucros em 34,75%

Badakhshan et al.	Application of simulation and machine learning in supply chain management: A synthesis of the literature using the Sim-ML literature classification framework	2024	10.1016/j.cie.2024.110649	Revisão sistemática da literatura sobre modelos híbridos que combinam simulação e machine learning na gestão da cadeia de abastecimento (SCM), com destaque nas falhas e oportunidades para futura pesquisa.	Revisão semi-sistemática baseada na abordagem PRISMA, apoiada por um framework próprio (Sim-ML).	Pesquisa em bases de dados como Scopus, Web of Science e IEEE Xplore, com critérios de inclusão e exclusão definidos.	Classificação de modelos Sim-ML, com destaque para modelos híbridos, como DES, ABS, SD e ML (SL, UL, RL).	99 artigos selecionados da literatura científica entre 2000 e 2023.	Análise bibliométrica, classificação por tipos de modelo, fluxo de dados, aplicações e cobertura de tecnologias da Indústria 4.0.	A integração de simulação e ML é promissora, mas subexplorada em áreas como otimização, sustentabilidade e tecnologias como digital twins. Falhas identificadas orientam direções de pesquisa futura.
Malik et al.	Artificial intelligence and industrial applications – A revolution in modern industries	2024	10.1016/j.asej.2024.102886	Revisão ampla sobre o impacto da inteligência artificial em diferentes setores industriais, destacando oportunidades e desafios da Indústria 4.0.	Revisão narrativa e conceitual com base em literatura dos últimos cinco anos.	Pesquisa em Google Scholar, PubMed, NIH, Web of Science, Springer, Embase.	Não aplicável (revisão descritiva sem modelagem matemática).	Cerca de 100 estudos revisados cobrindo diferentes indústrias e aplicações de IA.	Avaliação qualitativa dos impactos, benefícios e limitações do uso da IA em diferentes setores.	A IA tem potencial revolucionário em setores como saúde, alimentos, comércio eletrônico e farmacêutico, mas a sua adoção é limitada em

									países em desenvolvimento; há necessidade de regulamentação e políticas de incentivo.
Gao et al.	Data-driven robust optimization of dual-channel closed-loop supply chain network design considering uncertain demand and carbon cap-and-trade policy	2024	10.1016/j.cie.2023.109811	Proposta de um modelo robusto para redes de cadeia de abastecimento em circuito fechado (CLSC) com canais online e offline, incorporando incerteza na procura e políticas ambientais.	Modelagem matemática (MILP), otimização robusta orientada por dados e algoritmo genético adaptativo de dois estágios.	Dados simulados baseados em análise de componentes principais e kernel smoothing para construção de conjuntos de incerteza.	Modelo de otimização robusta baseado em MILP; Algoritmo Genético Adaptativo com integração do solver GUROBI.	Procuras incertas simuladas para rede de fornecimento com seis níveis e múltiplos períodos.	Custo total, desempenho do algoritmo, análise de sensibilidade a parâmetros.	A abordagem proposta melhora a robustez e a viabilidade económica e ambiental da rede CLSC, reduzindo faltas de *stock* e aumentando a eficiência frente à incerteza da procura.
Shaikh et al.	Empirical analysis in analysing the major factors of machine learning in enhancing the	2022	10.1007/s13198-021-01590-1	O estudo analisa os principais fatores do machine learning (ML) que impulsionam o e-business, com foco na previsão da	Quantitativa, utilizando modelagem de equações estruturais (SEM) e análise de mínimos	Questionários fechados aplicados a 155 gestores de empresas de e-commerce na Índia.	Structural Equation Modeling (SEM); Partial Least	Dados primários coletados via questionários (escala	Correlação entre variáveis; Confiabilidade (Alpha de Cronbach	ML influencia positivamente o e-business, com destaque para a previsão de procura, o

	e-business through structural equation modelling (SEM) approach		procura, no comportamento de compra, no compromisso do cliente e no cross-selling.	quadrados parciais (PLS).		Squares (PLS)	Likert de 5 pontos).	> 0.7); Significância estatística (valores-p)	comportamento de compra e o cross-selling. O compromisso do cliente não mostrou impacto significativo.
Mohammed et al.	Food waste in the era of e-commerce: A novel farm-to-fork management methodology	2025	https://doi.org/10.1016/j.ijpe.2025.109603	Investiga o desperdício de alimentos (FW) em cadeias de abastecimento eletrônicas (e-FSC), com foco no Omã, identificando causas e estratégias de mitigação.	Mista (qualitativa e quantitativa), com eum caso de estudo em duas redes de revendedores.	Questionários para especialistas da cadeia de abastecimento e consumidores;Reuniões com gestores	Analytic Hierarchy Process (AHP); DEMATEL (Decision Making Trial and Evaluation Laboratory)	Dados primários de 14 especialistas e 171 consumidores, além de dados secundários de literatura.	Relações causa-efeito (DEMATEL)	Principais causas de FW: falta de planeamento de produção (fornecedores), excesso de *stock* (distribuição) e armazenamento inadequado (revenda). Consumidores são os maiores geradores de FW. Estratégias eficazes incluem legislação,

										tecnologia e conscientização.
Xu et al.	Integrating game theory and data-driven optimization models for online retailers with reusable packaging adoption	2025	10.1016/j.jretconser.2025.104222	Com o crescimento do e-commerce, cresce também o problema dos resíduos de embalagens. Este estudo aborda a adoção de embalagens reutilizáveis por revendedores online, considerando tanto embalagens descartáveis como reutilizáveis, além de políticas de impostos sobre o plástico e subsídios governamentais.	Modelagem baseada em teoria dos jogos combinada com métodos de otimização orientados por dados (data-driven).	Análise de um conjunto real de dados com 456.548 registos de transações ao longo de 145 semanas de um revendedor online de entrega de alimentos.	Modelo de otimização separado (separated model); Modelo de otimização integrado (integrated model)	Dados reais de transações de uma plataforma de entrega de alimentos, abrangendo volume de pedidos e padrões de procura.	Adoção de embalagem reutilizável; Lucro do revendedor e da plataforma; Sensibilidade a impostos e subsídios	O modelo integrado supera o separado na promoção do uso de embalagens reutilizáveis e na melhoria do lucro de ambas as partes. Subsídios ajudam, mas em excesso causam pedidos desnecessários; impostos funcionam, mas com efeito decrescente.

Nie et al.	Joint estimation and prediction of city-wide delivery demand: A large language model empowered graph-based learning approach	2025	10.1016/j.tre.2025.104075	A crescente urbanização e e-commerce intensificam a procura de entregas urbanas. Este artigo procura prever e estimar essa procura de forma conjunta e transferível entre cidades, especialmente em regiões novas sem histórico.	Machine Learning com redes neurais de passagem de mensagens (message-passing) e codificação espacial baseada em LLMs (Large Language Models).	Dois conjuntos de dados reais de pedidos de entrega de oito cidades na China e nos EUA	Individual-collective message-passing neural network; STGNN (Spatiotemporal Graph Neural Network) com codificações de localização extraídas de LLMs	Dados de pedidos de entrega urbana de alimentos e pacotes, incluindo a localização, o tempo e a frequência.	Precisão da previsão; Eficiência computacional; Capacidade de transferência entre cidades	O modelo proposto superou os métodos de referência em precisão, eficiência e generalização para novas cidades, mesmo sem dados históricos.
Li et al.	Multi-modal transform-based fusion model for new product sales forecasting	2024	10.1016/j.engappai.2024.108606	Previsão de vendas de novos produtos no e-commerce, com desafios de ausência de histórico de dados e alta incerteza.	Desenvolvimento de um modelo de fusão multi-modal baseado em Transformer.	Dataset de e-commerce com mais de 10.000 itens de moda; integração de imagens, textos, atributos e fatores contextuais.	M2TFM (Multi-modal Transform-based Fusion Model), incluindo CNNs, BLIP, BERT, e Transforme	Imagens de produtos, descrições textuais, atributos dos produtos, fatores contextuais (feriados, clima, tendências),	Comparação de acurácia com modelos de ponta; não especificadas, mas com validação empírica em	O modelo M2TFM superou todos os métodos de comparação em previsão de vendas de novos produtos.

						rs com atenção.	e séries temporais.	dataset real.	
He et al.	Selling mode selection and AI service strategy in an E-commerce platform supply chain	2024	10.1016/j.cie.2024.110560	O estudo investiga a interação estratégica entre modos de venda (revenda vs. agência) e estratégias de serviço ao cliente (manual vs. IA) em cadeias de abastecimento de plataformas de e-commerce.	Modelagem analítica baseada em jogos de Stackelberg para analisar decisões de precificação, nível de serviço e escolha de modo de venda.	Modelagem teórica; uso de simulações numéricas para validação dos resultados.	Modelos de jogo de Stackelberg entre fabricante e plataforma de e-commerce; comparação entre quatro cenários combinando modos de venda e estratégias de serviço.	Parâmetros simulados em experimentos numéricos (não empíricos).	Preço de venda, nível de serviço de IA, procura, lucro da cadeia de abastecimento.	A eficiência do serviço de IA e o custo da IA influenciam fortemente a escolha do modo de venda e estratégia de serviço. IA tende a favorecer o modo de revenda ao aumentar lucros e reduzir perdas por eficiência decrescente do atendimento manual.

Surur et al.	Unlocking the power of machine learning in big data: a scoping survey	2025	10.1016/j.dsm.2025.02.004	Explora como o machine learning (ML) é usado para extrair valor de big data (BD) em diversos setores.	Revisão do objetivo com a análise de 150 estudos selecionados de mais de 2.500.	Pesquisa em bases acadêmicas (Scopus, IEEE, etc.) com critérios de inclusão/exclusão.	Deep Learning, Transfer Learning, Distributed Computing, Ensemble Learning.	Estudos acadêmicos e casos reais de aplicação industrial.	Volume, variedade, velocidade e veracidade dos dados (os 4 Vs do big data).	ML é essencial para lidar com os desafios do big data e gerar insights em tempo real, mas ainda enfrenta limitações técnicas em escalabilidade, qualidade dos dados e integração.
Cavalcante et al.	A supervised machine learning approach to data-driven simulation of resilient supplier selection in digital manufacturing	2019	10.1016/j.ijinfomgt.2019.03.004	O estudo responde à necessidade crescente de resiliência nas cadeias de abastecimento digitais, propondo uma abordagem baseada em machine learning supervisionado e simulação para apoiar a seleção de fornecedores resilientes em	Híbrida, combinando simulação com algoritmos de machine learning (k-NN e Regressão Logística)	Os dados foram gerados por simulação (modelo make-to-order), que forneceu uma base sintética para treinar e testar os modelos de machine learning.	Algoritmos de machine learning supervisionado – k-Nearest Neighbors (k-NN), Regressão Logística, e dois modelos híbridos (Hybrid A e Hybrid B).	Dados simulados sobre desempenho de fornecedores (entregas no prazo vs. atrasadas), com variáveis como data do pedido e quantidade	Área sob a curva ROC (AUC) para regressão logística, acurácia de classificação para k-NN; taxa de entregas no prazo como	A combinação de simulação com machine learning supervisionado (especialmente os modelos híbridos) melhorou a confiabilidade das entregas e permite identificar fornecedores críticos para

				ambientes de manufatura digital.				encomenda da.	métrica de desempenho global.	mitigar riscos e aumentar a resiliência da cadeia de abastecimeto.
Y. Chen et al.	Supply chain demand forecasting based on multi-time scale data fusion net- work	2025	10.1016/j.cie.2025.111324	O artigo aborda o desafio de prever a procura nas cadeias de abastecimento sob cenários complexos e incertos, propondo um modelo de rede neural baseado em fusão de dados em múltiplas escalas temporais.	Proposta de um modelo de previsão, com ênfase em redes neurais e fusão de dados multiescalares.	Utilização de dados históricos de séries temporais de procura	Multi-Time Scale Data Fusion Network (modelo de deep learning)	Séries temporais com diferentes resoluções (horária, diária, semanal), integradas via fusão de dados.	RMSE, MAE, MAPE	O modelo proposto apresentou melhor desempenho de previsão em comparação com abordagens tradicionais de previsão da procura, ao capturar padrões de diferentes escalas temporais.

Cano et al.	Sustainable Logistics for E-Commerce: A Literature Review and Bibliometric Analysis	2022	10.3390/su141912247	O artigo trata do crescimento do e-commerce e os desafios de sustentabilidade na logística, especialmente após a pandemia de COVID-19.	Revisão sistemática da literatura e análise bibliométrica com base na diretriz PRISMA.	Pesquisa em bases de dados (Scopus e Web of Science) com critérios específicos (2021–2022, palavras-chave como “logistics”, “e-commerce”, “sustainability”	Não aplicável (revisão de literatura, sem modelagem)	105 artigos científicos extraídos de Scopus e Web of Science.	Número de publicações, citações, coocorrência de palavras-chave, análise de redes de coautoria.	Identificação de tendências em sustentabilidade na logística do e-commerce, com destaque para last-mile delivery, impactos ambientais e o papel da tecnologia e políticas públicas.
Vhatkar, M. S., et al.	Optimized hyperparameters for retail sales forecasting using grid search.	2025	10.1016/j.engappai.2024.111472	Previsão de vendas no retalho para diversos produtos, considerando flutuações pós-pandemia.	Pipeline de Ciência de Dados com foco na otimização de hiperparâmetros.	Repositórios de dados públicos de retalho.	GBR (Gradient Boosting Regression), LSTM, GRU e RNN.	Séries temporais de vendas diárias.	RMSE, MAE e R^2.	A otimização por *Grid Search* permitiu que modelos de ML (como GBR) superassem modelos de *Deep Learning* em termos de precisão e eficiência.

Mohammed, Z., et al.	A hybrid learning framework for forecasting uncertainty and adaptive inventory planning in retail supply chains.	2026	10.1016/j.sca.2024.100180	Gestão de inventário e planeamento adaptativo em cadeias de abastecimento de retalho.	Estrutura de aprendizagem híbrida combinando modelos de variância condicional e meta-aprendizagem.	Dataset M5 Forecasting (Walmart) via Kaggle.	Híbrido: XGBoost, LightGBM, LSTM-GRU e GARCH.	Dados de vendas de retalho, preços e calendários promocionais.	WRMSSE (Weighted Root Mean Squared Scaled Error) e custos de inventário.	A integração de previsões de incerteza no planeamento de inventário reduz custos operacionais e melhora o nível de serviço ao cliente através de decisões de *stock* baseadas em risco.
Chen, T., & Guestrin, C.	XGBoost: A Scalable Tree Boosting System.	2016	10.1145/2939672.2939785	Desenvolvimento de um sistema de *tree boosting* escalável e eficiente para desafios de *Machine Learning* em grande escala.	Proposta de um novo algoritmo de aprendizagem para árvores esparsas e um sistema de computação distribuída para acelerar o treino.	Utilização de datasets públicos de referência (como AllFlight, Yahoo! Learning to Rank, e Higgs boson).	XGBoost (e comparação com Gradient Boosting e Random Forest).	Vários datasets de larga escala para testar a escalabilidade e velocidade.	Erro de classificação, Tempo de processamento (Velocidade) e Utilização de memória.	O XGBoost provou ser significativamente mais rápido que outras implementações de *boosting* e capaz de lidar com milhares de milhões de exemplos em sistemas reais,

										tornando-se o estado da arte para dados estruturados.
Brooks, C.	Introductory Econometrics for Finance.	2019	10.1017/9781108524872	Fornecer uma introdução teórica e prática aos métodos econométricos aplicados a séries temporais e dados financeiros.	Revisão teórica de modelos matemáticos e estatísticos para análise de dados históricos.	-	Modelos Univariados de Séries Temporais (ARIMA, SARIMA) e modelos de volatilidade	Exemplos teóricos e práticos de preços de ativos e séries temporais financeiras.	Critérios de Informação (AIC, BIC), significância estatística e testes de resíduos.	Estabelece os fundamentos de que séries temporais devem ser testadas quanto à sua estacionaridade e que modelos lineares (como ARIMA) servem como base comparativa essencial para previsões de dados históricos.

Dickey, D. A., & Fuller, W. A.	Distribution of the Estimators for Autoregressive Time Series With a Unit Root.	1979	10.2307/2286348	Investigação estatística sobre a presença de "raízes unitárias" em séries temporais, que causam não-estacionaridade.	Desenvolvimento de testes de hipóteses estatísticas e tabelas de distribuição para coeficientes de modelos autorregressivos.	Análise teórica e simulações de Monte Carlo.	Modelos Autorregressivos (AR) de primeira ordem.	Séries temporais simuladas matematicamente.	Estatística t (t-statistic) e valores críticos de probabilidade.	Criou-se o teste padrão (Teste de Dickey-Fuller) para determinar se uma série temporal precisa de ser diferenciada para se tornar estacionária, um passo obrigatório para a aplicação de modelos ARIMA.
Friedman, J. H.	Greedy function approximation: A gradient boosting machine.	2001	10.1214/aos/1013203451	Desenvolvimento de uma estrutura teórica para a estimação de funções estatísticas através de algoritmos de *boosting*.	Proposta do algoritmo *Gradient Boosting Machine* (GBM), que utiliza a descida do gradiente (*gradient descent*) em espaços de funções.	Demonstração teórica e testes computacionais em datasets de regressão e classificação.	Gradient Boosting Machines (GBM).	Datasets sintéticos e reais para testar a robustez da aproximação de funções.	Log-likelihood, Erro Quadrático Médio (MSE) e Desvio (Deviance).	Provou que é possível construir modelos preditivos altamente precisos através da combinação sequencial de modelos simples ("fracos"),

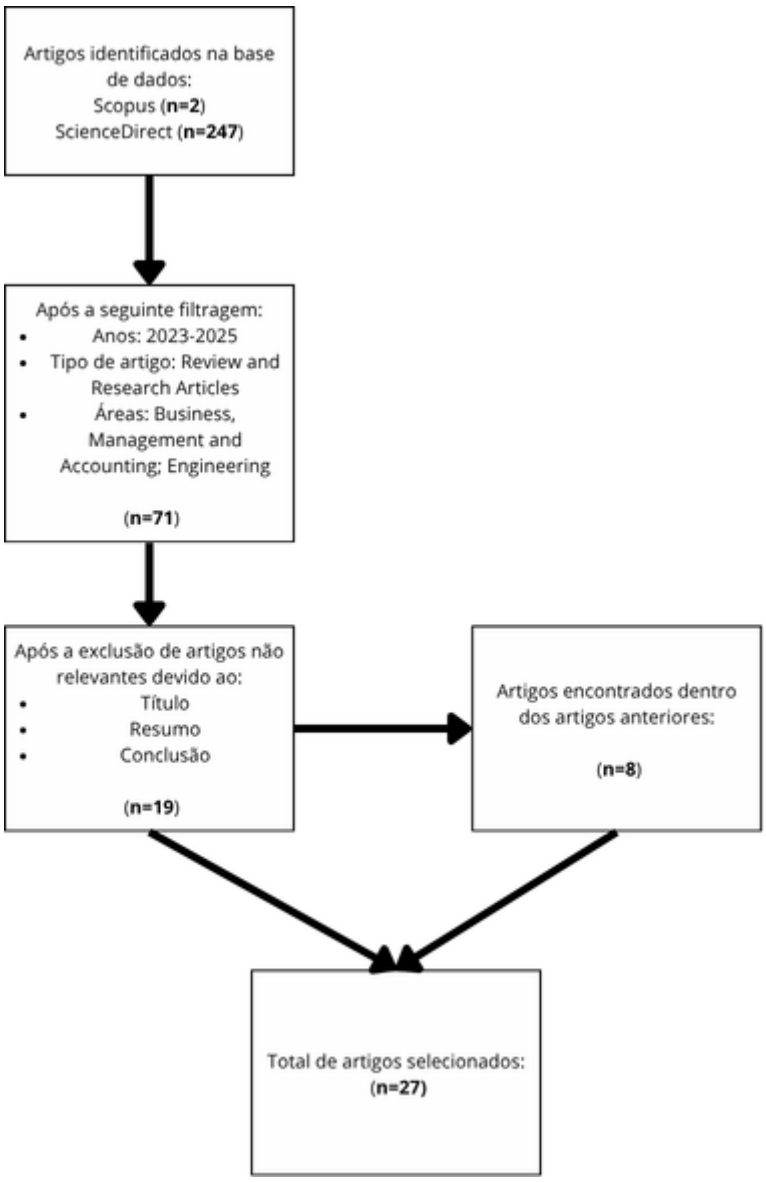
									onde cada novo modelo foca-se em corrigir os erros residuais do modelo anterior.
Hyndman, R. J., & Athanasopoulos, G.	Forecasting: Principles and Practice (2nd ed.).	2018	https://otexts.com/fpp2/	Fornecer uma introdução abrangente aos métodos de previsão, focando tanto na teoria estatística como na aplicação prática.	Guia sistemático sobre o ciclo de previsão: preparação de dados, visualização, escolha de modelos e avaliação de precisão.	-	ETS (Suavização Exponencial), ARIMA, Regressão Dinâmica e Redes Neuronais para séries temporais.	Diversos datasets reais de retalho, economia e turismo (como as vendas de farmácias na Austrália).	MAE, RMSE, MAPE e MASE (Mean Absolute Scaled Error).	Define as melhores práticas para a previsão de séries temporais, enfatizando que a escolha do modelo deve basear-se na minimização de métricas de erro em dados de teste (out-of-sample) e não apenas no ajuste aos dados de treino.

Joseph, R. V., et al.	A hybrid deep learning framework with CNN and Bi-directional LSTM for store item demand forecasting.	2022	10.1016/j.compeleceng.2022.108358	Previsão de procura ao nível de item/loja, lidando com a complexidade de padrões sazonais e tendências em dados de retalho.	Proposta de um modelo híbrido que utiliza Redes Neuronais Convolucionais (CNN) para extrair características e LSTMs Bidirecionais (Bi-LSTM) para capturar dependências temporais.	Utilização de datasets de competições de retalho (como o Favorita Grocery Sales Forecasting).	Híbrido CNN-BiLSTM (comparado com ARIMA, Random Forest e LSTM simples).	Dados históricos de vendas de diversos produtos em várias lojas.	RMSE (Root Mean Squared Error) e MAE (Mean Absolute Error).	O modelo híbrido superou significativamente os modelos tradicionais e de ML simples, provando que a estrutura Bi-LSTM é mais eficaz a capturar o contexto das séries temporais do que a LSTM convencional.
Peng, L., Wang, L., Xia, D., & Gao, Q.	Effective energy consumption forecasting using empirical wavelet transform and long short-term memory.	2022	10.1016/j.energy.2021.121756	Previsão de consumo de energia, lidando com a natureza não linear e altamente volátil das séries temporais.	Proposta de um modelo híbrido que utiliza a Transformada de Wavelet Empírica (EWT) para decompor a série temporal em componentes mais estáveis antes da previsão.	Monitorização de redes elétricas e bases de dados de consumo energético.	EWT-LSTM (Comparado com LSTM simples, RNN e modelos estatísticos).	Séries temporais de consumo de energia elétrica de larga escala.	RMSE, MAE, MAPE e R2.	A decomposição da série temporal através de Wavelets (EWT) reduziu significativamente o ruído dos dados, permitindo que a LSTM

									aprendesse padrões de longo prazo com muito maior precisão do que quando aplicada aos dados brutos.
Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (referenciado por Tunnicliffe Wilson, G.).	Time Series Analysis: Forecasting and Control.	2016	0.1111/jtsa.12194	Definição do quadro teórico e prático para a modelação, identificação, estimação e diagnóstico de séries temporais.	Metodologia Box-Jenkins, que utiliza um ciclo iterativo de identificação de modelos (baseado em correlogramas), estimação de parâmetros e verificação de diagnóstico.	-	Modelos Estocásticos, ARIMA, SARIMA e Modelos de Transferência.	Séries temporais clássicas (ex: dados de passageiros de companhias aéreas) para demonstração de processos estocásticos.	Funções de Autocorrelação (ACF), Funções de Autocorrelação Parcial (PACF) e testes de ruído branco (White Noise).	Estabeleceu o padrão mundial para a previsão de séries temporais, introduzindo o conceito de que séries não-estacionárias podem ser modeladas através de diferenciação (o "I" do ARIMA).

Fonte: Elaboração Própria.

Figura 16 - Esquema PRISMA.



Fonte: Elaboração Própria.

Bases de dados

As bases de dados utilizadas para a pesquisa de artigos foram as seguintes: *Scopus* e *ScienceDirect*.

Termos de pesquisa

Os termos de pesquisa foram definidos com base nos principais conceitos envolvidos na questão de investigação. Utilizou-se combinações com operadores booleanos (AND/OR) para ampliar os resultados relevantes:

- “previsão de procura” OR “demand forecasting”
- “e-Commerce” OR “comércio eletrónico”
- “machine learning” OR “aprendizado de máquina”
- “logística sustentável” OR “sustainable logistics”
- “ARIMA”
- “XGBoost”
- “redução de desperdícios” OR “waste reduction”

Scopus: “TITLE-ABS-KEY (''demand forecast*'' AND ''e-commerce'') AND TITLE-ABS-KEY (''machine learning'' OR ''ARIMA'') AND TITLE-ABS-KEY (''sustainab*'' OR ''waste reduction'') AND PUBYEAR > 2021 AND PUBYEAR < 2026 AND (LIMIT-TO (SUBJAREA , ''BUSI'') OR LIMIT-TO (SUBJAREA , ''ENGI''))”

ScienceDirect: (''demand forecasting'' AND ''e-commerce'') AND (''machine learning'' OR ARIMA) AND (sustainability OR ''waste reduction'')

Critérios de elegibilidade

Filtros: 2023-2025; Review and Research articles; Subject areas: Business, Management and Accounting; Engineering.

Após revisão (considerei alguns não relevantes, pelo título, resumo e conclusão).

B – Código de Programação utilizado para criar a figura 1

```
import matplotlib.pyplot as plt
import matplotlib.patches as patches
def draw_architecture():
    fig, ax = plt.subplots(figsize=(12, 6))
    # Definição das camadas (Nome, Cor, Descrição)
    layers = [("Input", "#D3D3D3", "(30 dias, 1)"), ("Conv1D", "#FFD700", "Filters: 64\nKernel: 3"), ("MaxPooling", "#FFA500", "Pool: 2"), ("LSTM", "#90EE90", "Units: 50"), ("Dense", "#87CEEB", "Units: 32"), ("Output", "#FF6347", "Previsão"]
    start_x = 0.1
    gap = 0.15
    box_width = 0.1
    box_height = 0.4
    y_pos = 0.3
    # Desenhar cada camada como uma caixa
```

```
    for i, (name, color, desc) in enumerate(layers):
        x = start_x + i * gap
        # Criar retângulo (Caixa da camada)
        rect = patches.Rectangle((x, y_pos), box_width, box_height,
                                 linewidth=2, edgecolor='black', facecolor=color)
        ax.add_patch(rect)
        # Adicionar Nome da Camada (Centro da caixa)
        ax.text(x + box_width/2, y_pos + box_height/2, name,
                ha='center', va='center', fontsize=10, fontweight='bold', color='black')
        # Adicionar Descrição (Em baixo da caixa)
        ax.text(x + box_width/2, y_pos - 0.05, desc,
                ha='center', va='top', fontsize=9, color='#333333')
        # Desenhar seta para a próxima camada (exceto na última)
        if i < len(layers) - 1:
            arrow_start = x + box_width
            arrow_len = gap - box_width
            ax.arrow(arrow_start, y_pos + box_height/2, arrow_len, 0,
                     head_width=0.03, head_length=0.02, fc='k', ec='k')
    # Configurações do gráfico
    ax.set_xlim(0, 1)
    ax.set_ylim(0, 1)
    ax.axis('off') # Esconder eixos
    plt.title("Arquitetura Híbrida Proposta (CNN-LSTM)", fontsize=14)
    # Guardar imagem
    plt.savefig('diagrama_arquitetura_manual.png', dpi=300, bbox_inches='tight'
    plt.show()
if __name__ == "__main__":
    draw_architecture()
```

C – Código de Programação utilizado para criar a figura 2

```
import graphviz
import os
def draw_lstm_cell():
    dot = graphviz.Digraph(comment='LSTM Cell', format='png'
    # --- Configurações Visuais ---
    dot.attr(rankdir='LR') # Esquerda para a Direita
    dot.attr(splines='spline') # Linhas curvas suaves
    dot.attr('node', fontname='Helvetica', fontsize='12')
    # --- 1. Entradas (Inputs) ---
    with dot.subgraph(name='cluster_inputs') as c:
        c.attr(style='invis')
        c.node('Xt', 'Input\n(Xt)', shape='box', color='black', style='filled', fillcolor='#E0E0E0')
        c.node('Ht_prev', 'Hidden Anterior\n(ht-1)', shape='box', color='black', style='filled', fillcolor='#E0E0E0')
        c.node('Ct_prev', 'Memória Anterior\n(Ct-1)', shape='circle', style='filled', fillcolor='#E0E0E0')
    # --- 2. Portões ---
    # As redes neuronais (amarelo/laranja)
    dot.attr('node', shape='rect', style='filled', fillcolor='#FFD700', width='0.6', fixedsize='true')
    dot.node('Sig_Forget', 'σ\n(Forget)')
    dot.node('Sig_Input', 'σ\n(Input)')
```

```
    dot.node('Tanh_Cand', 'tanh\n(Cand)')
    dot.node('Sig_Output', 'σ\n(Output)')
    # -- 3. Operações Matemáticas --
    # Círculos para operações (azul claro)
    dot.attr('node', shape='circle', style='filled', fillcolor='#87CEEB', width='0.4', fixedsize='true')
    dot.node('Mul_Forget', 'X')
    dot.node('Mul_Input', 'X')
    dot.node('Add_Cell', '+')
    dot.node('Tanh_Output', 'tanh')
    dot.node('Mul_Output', 'X')
    # -- 4. Saídas (Outputs) --
    dot.attr('node', shape='box', style='filled', fillcolor='#90EE90', width='1.2')
    dot.node('Ct_next', 'Nova Memória\n(Ct)')
    dot.node('Ht_next', 'Novo Hidden\n(ht)')
    # -- 5. Ligações --
    # Fluxo da Memória (A linha de cima "Conveyor Belt")
    dot.edge('Ct_prev', 'Mul_Forget', penwidth='2')
    dot.edge('Mul_Forget', 'Add_Cell', penwidth='2', label='Retém')
    dot.edge('Add_Cell', 'Ct_next', penwidth='2')
    dot.edge('Add_Cell', 'Tanh_Output') # Cópia para o output
    # Inputs a fluir para os portões
    # Para simplificar o diagrama, assume-se que Xt e Ht-1 se juntam
    dot.edge('Xt', 'Sig_Forget', color='grey')
    dot.edge('Xt', 'Sig_Input', color='grey')
    dot.edge('Xt', 'Tanh_Cand', color='grey')
    dot.edge('Xt', 'Sig_Output', color='grey')
    dot.edge('Ht_prev', 'Sig_Forget', color='grey')
    dot.edge('Ht_prev', 'Sig_Input', color='grey')
    dot.edge('Ht_prev', 'Tanh_Cand', color='grey')
    dot.edge('Ht_prev', 'Sig_Output', color='grey')
    # Lógica do Forget Gate
    dot.edge('Sig_Forget', 'Mul_Forget', label=' ft')
    # Lógica do Input Gate
    dot.edge('Sig_Input', 'Mul_Input', label=' it')
    dot.edge('Tanh_Cand', 'Mul_Input', label=' ~Ct')
    dot.edge('Mul_Input', 'Add_Cell', label='Adiciona')
    # Lógica do Output Gate
    dot.edge('Sig_Output', 'Mul_Output', label=' ot')
    dot.edge('Tanh_Output', 'Mul_Output')
    dot.edge('Mul_Output', 'Ht_next')
    file_name = 'diagrama_lstm_interno'
    dot.render(file_name, view=True)
if __name__ == "__main__":
    draw_lstm_cell()
```

Nota: Por questões de formatação e extensão, os apêndices seguintes apresentam apenas os blocos de código fundamentais desenvolvidos no âmbito desta dissertação. O código-fonte integral, incluindo as análises exploratórias completas e visualizações geradas em Python, encontram-se disponíveis para consulta pública através dos seguintes links:

Limpeza e tratamento de dados:

- https://colab.research.google.com/drive/1wlSpp65g63zbFC3EbnI_wII1EjN68fE-?usp=sharing

Modelo ARIMA:

- https://colab.research.google.com/drive/1Et_0ZbrXTa0xyxsNXt_7sgxvKj2s0VQd?usp=sharing

Modelo XGBoost:

- https://colab.research.google.com/drive/1AvBKWTbJDLhz0Y6RhlSym0Xjbh7sMC1Z?usp=sharing

Modelo LSTM:

- https://colab.research.google.com/drive/1JDS1BuXiuK0xIG05uQIrwJeyqxfj0H3I?usp=sharing

D – Código de Programação utilizado para tratamento de dados

```
#Vamos verificar se nas linhas que não tem descrição se o preço é ou não igual a 0
# 1. Criar um filtro: Quero ver apenas as linhas onde a Descrição está vazia (NaN)
linhas_sem_descricao = total_retail_data[total_retail_data['Descricao'].isnull()]
# 2. Dessas linhas sem descrição, mostra-me as colunas 'Descricao' e 'Preco_Unitario'
print(linhas_sem_descricao[['Descricao', 'Preco_Unitario']].head(10))
# 3. A Pergunta de Ouro: Quantas destas têm Preço = 0?
sem_desc_e_preco_zero = linhas_sem_descricao[linhas_sem_descricao['Preco_Unitario'] == 0]
print(f"\nTotal de linhas sem descrição: {len(linhas_sem_descricao)}")
print(f"Total destas que têm Preço 0.0: {len(sem_desc_e_preco_zero)}")
#Vamos então eliminar as linhas que tem valores omissos na descrição
total_retail_data = total_retail_data.dropna(subset=['Descricao'])

# -- SEPARAÇÃO ESTRATÉGICA PARA A TESE --
# 1. Dataset para SUSTENTABILIDADE (Logística Inversa + Desperdício Interno)
# Apanha tudo o que é <= 0.
# Isto inclui: Devoluções de Clientes (Negativos) E Quebras/Perdas em Armazém (Zero)
df_sustentabilidade = total_retail_data[total_retail_data['Total_Venda'] <= 0].copy()
# 2. Dataset para PREVISÃO DA PROCURA (Vendas Limpas)
# Apanha apenas as transações que geraram receita real
# É este que vamos usar no ARIMA e XGBoost
df_previsao = total_retail_data[total_retail_data['Total_Venda'] > 0].copy()
# -- RELATÓRIO FINAL DA SEPARAÇÃO --
print(f"Total Original: {len(total_retail_data)}")
print("-" * 30)
print(f"Registos para Análise de Sustentabilidade: {len(df_sustentabilidade)}")
print(f"Registos para Previsão da Procura: {len(df_previsao)}")
print("-" * 30)
# Verificação de segurança (A soma dos dois deve ser igual ao total)
if len(df_sustentabilidade) + len(df_previsao) == len(total_retail_data):
    print("Cálculo perfeito! Nenhum dado foi perdido.")
else:
    print("Atenção: Algo não bate certo nas contas.")
```

E – Código de Programação utilizado para modelo ARIMA

```
# 1. TREINAR SARIMA (Seasonal ARIMA)
# order=(1, 1, 1) -> O ARIMA normal que já tinha
# seasonal_order=(1, 1, 1, 7) -> (P, D, Q, s)
# O '7' é o segredo: dizemos que o ciclo é semanal (7 dias)
print("A treinar SARIMA... (Isto pode demorar 30 segundos, ele está a pensar muito!)")
modelo_sarima = ARIMA(treino, order=(1, 1, 1), seasonal_order=(1, 1, 1, 7))
modelo_sarima_fit = modelo_sarima.fit()
# 2. FAZER PREVISÃO
previsao_sarima = modelo_sarima_fit.forecast(steps=dias_teste)
previsao_sarima.index = teste.index
# 3. AVALIAR
rmse_sarima = np.sqrt(mean_squared_error(teste, previsao_sarima))
mae_sarima = mean_absolute_error(teste, previsao_sarima)
mape_sarima = mean_absolute_percentage_error(teste, previsao_sarima) * 100 # Multiplicado por 100 para dar %
r2_sarima = r2_score(teste, previsao_sarima)
print("-" * 30)
print(f"ARIMA Simples - RMSE: £41163.88 (Aborrecido)")
print(f"SARIMA (Sazonal) - RMSE: £{rmse_sarima:.2f}")
print(f"SARIMA (Sazonal) - MAE (Erro Médio Absoluto): £{mae_sarima:.2f}")
print(f"SARIMA (Sazonal) -  MAPE: {mape_sarima:.2f}%")
print(f"SARIMA (Sazonal) -  R2: {r2_sarima:.4f}")
print("-" * 30)
# 4. GRÁFICO
plt.figure(figsize=(15, 6))
plt.plot(teste.index, teste, label='Realidade (Teste)', color='green')
plt.plot(teste.index, previsao, label='ARIMA Simples', color='red', linestyle='--', alpha=0.3)
plt.plot(teste.index, previsao_sarima, label='SARIMA (Sazonal)', color='blue', linewidth=2)
plt.title('Duelo: Realidade vs SARIMA (Com Sazonalidade Semanal)', fontsize=14)
plt.legend()
plt.grid(True, alpha=0.5)
plt.show()
```

F – Código de Programação utilizado para modelo XGBoost

```
import xgboost as xgb
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score, mean_absolute_percentage_error
import numpy as np
# 1. DIVIDIR TREINO vs TESTE (Os mesmos 28 dias finais para ser justo)
test_days = 28
train_size = len(df_ml) - test_days
X = df_ml.drop('Total_Venda', axis=1) # As Pistas
y = df_ml['Total_Venda']              # O Alvo
```

```
X_train, X_test = X.iloc[:train_size], X.iloc[train_size:]
y_train, y_test = y.iloc[:train_size], y.iloc[train_size:]
# 2. CONFIGURAR E TREINAR O MODELO
reg = xgb.XGBRegressor(
    n_estimators=1000,    # Número de árvores
    learning_rate=0.01,   # Velocidade de aprendizagem
    max_depth=5,          # Profundidade da árvore
    early_stopping_rounds=50
)
reg.fit(X_train, y_train,
        eval_set=[(X_train, y_train), (X_test, y_test)],
        verbose=False)
# 3. PREVER E AVALIAR
previsao_xgb = reg.predict(X_test)
rmse_xgb = np.sqrt(mean_squared_error(y_test, previsao_xgb))
mae_xgb = mean_absolute_error(y_test, previsao_xgb)
mape_xgb = mean_absolute_percentage_error(y_test, previsao_xgb) * 100 # Multiplicado por 100 para dar %
r2_xgb = r2_score(y_test, previsao_xgb)
print("="*40)
print(f"SARIMA Manual (Base) - RMSE: £ 31603.31")
print(f"XGBoost - MAE: {mae_xgb}")
print(f"XGBoost - RMSE: {rmse_xgb}")
print(f"XGBoost - MAPE: {mape_xgb:.2f}%")
print(f"XGBooost - R2: {r2_xgb:.4f}")
print("="*40)
if rmse_xgb < 31603.31:
    diff = 31603.31 - rmse_xgb
    print(f"🎉 VITÓRIA! O XGBoost reduziu o erro em £{diff:.2f}!")
else:
    print("🤖 O XGBoost ainda precisa de afinação...")
# 4. GRÁFICO FINAL
plt.figure(figsize=(15, 6))
plt.plot(y_test.index, y_test, label='Realidade', color='green', linewidth=2)
plt.plot(y_test.index, previsao_xgb, label='XGBoost', color='purple', linestyle='--')
plt.title('XGBoost: Previsão vs Realidade', fontsize=16)
plt.legend()
plt.grid(True)
plt.show()
# 5. IMPORTÂNCIA DAS VARIÁVEIS
xgb.plot_importance(reg, height=0.5, title="O que influenciou mais a previsão?")
plt.show()
```

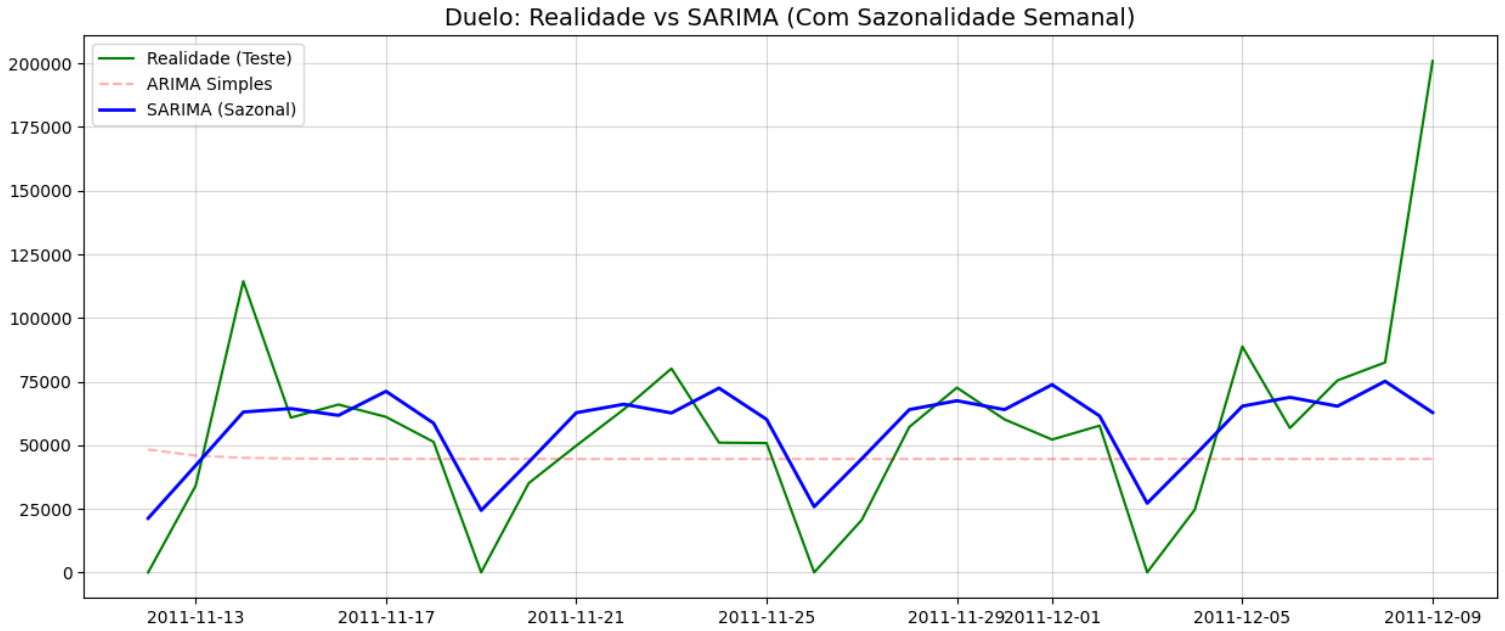
G – Código de Programação utilizado para modelo LSTM

```
from tensorflow.keras.layers import Dropout
# Arquitetura mais robusta
modelo = Sequential()
# 1ª Camada LSTM (return_sequences=True porque vamos ter outra LSTM a seguir)
modelo.add(LSTM(units=50, return_sequences=True, input_shape=(look_back, 1)))
modelo.add(Dropout(0.2)) # Desliga 20% dos neurónios para evitar "viciar"
# 2ª Camada LSTM
```

```
modelo.add(LSTM(units=30, return_sequences=False))
modelo.add(Dropout(0.2))
# Camada de Saída
modelo.add(Dense(1))
modelo.compile(optimizer='adam', loss='mean_squared_error')
print("A treinar a Rede Neuronal Robusta... (Isto demora um pouco mais)")
history = modelo.fit(X_train, y_train, epochs=150, batch_size=16, verbose=1)
```

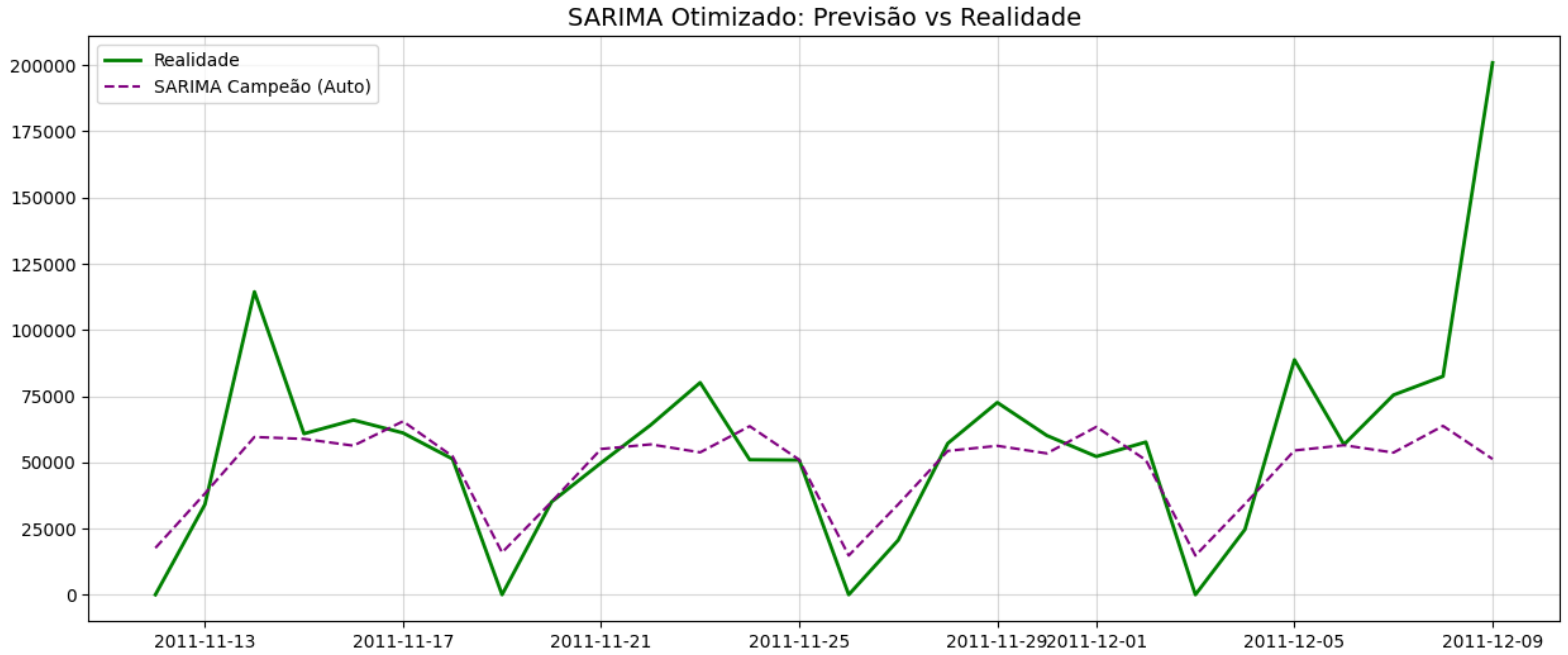
H – Análise Gráfica do Ajustamento dos Modelos Secundários

Figura 17 - Comparação gráfica entre a procura real e as previsões do modelo estatístico SARIMA no período de teste.



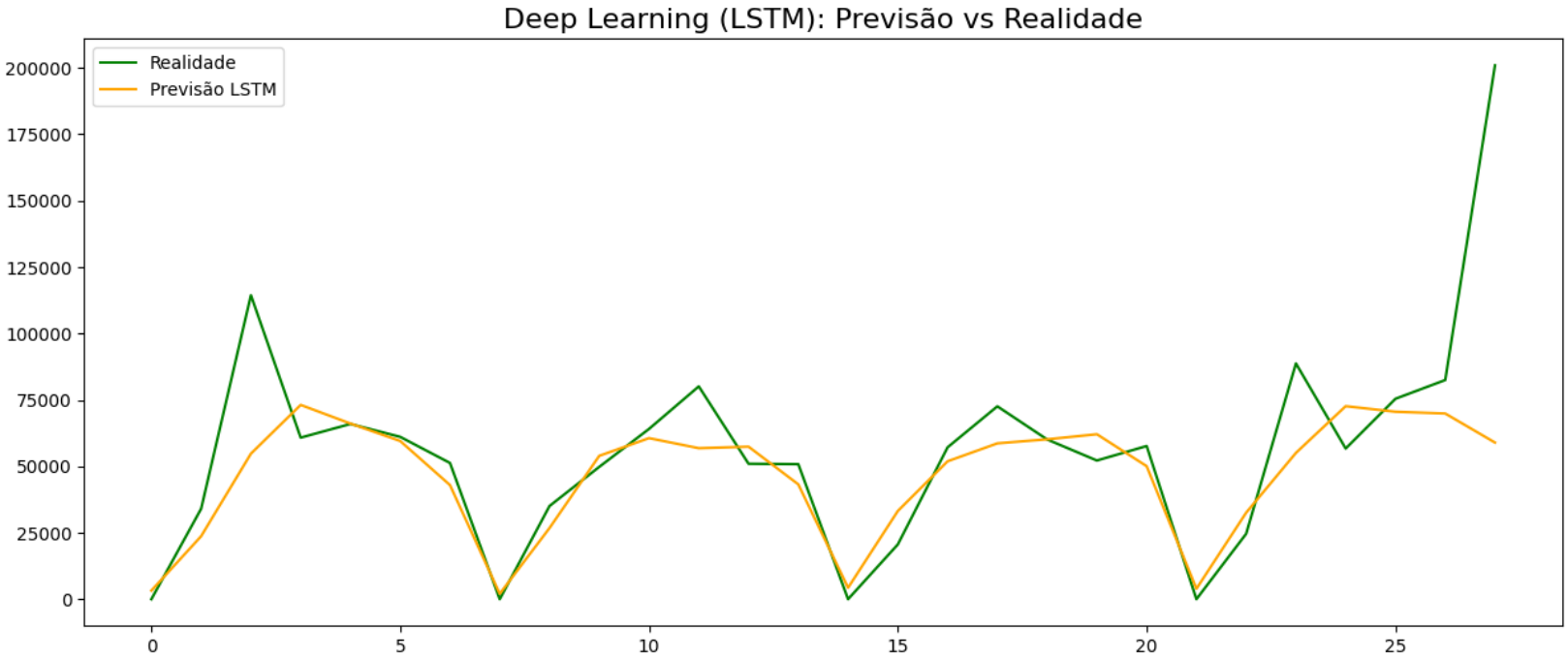
Fonte: Elaboração Própria.

Figura 18 - Comparação gráfica entre a procura real e as previsões do modelo estatístico SARIMA Automatizado no período de teste.



Fonte: Elaboração Própria.

Figura 19 - Comparação gráfica entre a procura real e as previsões do modelo LSTM (Deep Learning) no período de teste.



Fonte: Elaboração Própria.