MIGUEL SACADURA LUCAS

INTELIGÊNCIA ARTIFICIAL NA CADEIA DE ABASTECIMENTO: ANÁLISE DE SÉRIES TEMPORAIS PARA PREVISÃO DA PROCURA

Projeto Aplicado de Mestrado em Ciência de Dados para Empresas

ORIENTADORES

Professor Doutor David Alexandre Mendes Silva Simões

Professor Doutor Pedro Fernandes da Anunciação

SUPERVISOR

Carlos Pinto, Sovena Group – Responsável de Logística Interna

setembro de 2025

MIGUEL SACADURA LUCAS

INTELIGÊNCIA ARTIFICIAL NA CADEIA DE ABASTECIMENTO: ANÁLISE DE SÉRIES TEMPORAIS PARA PREVISÃO DA PROCURA

JÚRI

Presidente: Prof. Coordenador Ana de Jesus Pereira Barreira Mendes

Vogal Arguente: Profesor Ayudante Ricardo Teruel Gutiérrez

Orientador: Prof. Coordenador Pedro Fernandes da Anunciação

11 de novembro de 2025

Resumo

Numa era onde a componente digital penetra todo e qualquer setor de atividade, a cadeia de abastecimento é naturalmente uma área que não foge à regra. Processos de transformação digital ou digitalização oferecem maior capacidade às organizações de reformular áreas de gestão e subprocessos na perspetiva de acrescentar valor às respetivas atividades. Um dos temas mais sonantes da atualidade é o avanço significativo da inteligência artificial e como cada vez mais, essas descobertas e implementações oferecem a várias áreas, nomeadamente, à cadeia de abastecimento, novas formas de analisar dados, prever cenários futuros e formular novos tipos de ação/reação de acordo com as diferentes situações. A tomada de decisão é assim suportada por dados relevantes, de qualidade superior, previamente trabalhados para determinado objetivo e é através dessa premissa que se desenvolve todo este projeto.

Inevitavelmente, uma das áreas de maior influência no fluxo das cadeias de abastecimento é o planeamento da procura e como tal, torna-se imperativo apontar à máxima eficiência possível dos métodos e modelos utilizados para a prever. Sabendo que haverá sempre constrangimentos diretos ou indiretos que podem impactar negativamente estes fluxos, o esforço é feito fundamentalmente para minimizar esses mesmos estragos. A relação entre a existência de constrangimentos e a eficiência da cadeia de abastecimento será sempre uma de proporcionalidade inversa, uma vez que, quanto melhor ou mais precisa for a previsão calculada da procura, menor serão os constrangimentos causados às operações a jusante.

Neste seguimento, entram em ação os modelos preditivos desenvolvidos através de aprendizagem automática e como estes podem oferecer novos padrões de qualidade ao processo de previsão da procura e planeamento de necessidades. Modelos de análise e previsão em séries temporais são um dos exemplos utilizados para prever níveis de procura e será sobre esse que incidirá o foco neste projeto.

Palavras-chave: cadeia de abastecimento, planeamento da procura, inteligência artificial, previsão de séries temporais.

Abstract

In a digital era that permeates every sector of activity, the supply chain is naturally an area that does not escape this trend. Digital transformation processes offer every kind of organization greater capacity to reformulate management areas and subprocesses with the main goal being adding value to their respective activities. One of the most prominent topics nowadays is the remarkable evolution of artificial intelligence and how these breakthroughs and implementations offer various areas, namely, the supply chain, new ways to analyze data, predict possible scenarios and manage new types of action taking accordingly. Decision-making is thus supported by relevant, high-quality data, previously processed for a specific purpose creating a new concept of data-driven supply chains.

Inevitably, one of the areas with the most influence on the supply chain is demand planning and, as such, it becomes imperative to aim for the maximum efficiency possible when methods/models used to predict it are applied. Knowing that there will always be direct and/or indirect constraints that can negatively impact these flows, the effort is fundamentally made to minimize these unpredictable damages. The relationship between the existence of constraints and the efficiency of the supply chain will always be one of inverse proportionality, since the better or more accurate the calculated demand is, the fewer constraints caused to downstream operations such as inventory management or warehousing.

In this context, predictive models developed through machine learning come into play and specifically, how can they offer new quality standards to the demand forecasting and planning processes. Time series analysis and forecasting models are one of the examples used to predict demand levels, and this will be the focus of the project.

Keywords: supply chain, demand planning, artificial intelligence, time-series forecasting.

Índice

Lista de Figuras

Lista de Tabelas

Lista de Gráficos

Lista de Blocos

Lista de Abreviaturas

ADF – Teste Dickey-Fuller Aumentado

ADI – Análise de dados Inicial

AE – Análise exploratória

AIC – Akaike's Information Criterion

BDA – Big Data Analytics

CA – Cadeia de Abastecimento

CAD – Cadeia de Abastecimento Digital

CRISP-DM – Cross-Industry Standard Process for Data Mining

DP – Demand Planning

ERP – Enterprise Resource Planning

GCA – Gestão da Cadeia de Abastecimento

GR – Guia de remessa

IA – Inteligência Artificial

ICCE – International Center for Competitive Excellence

IoT – Internet of Things

MAE – Mean Absolute Error

MAPE – Mean Absolute Percentage Error

ML – Machine Learning

RPA – Robotic Process Automation

SI – Sistema de Informação

SJR – SCImago Journal Rank

ST – Séries Temporais

UM – Unidade medida

Introdução

O presente relatório foi realizado no âmbito do Mestrado em Ciência de Dados para Empresas, na Escola Superior de Ciências Empresariais – Instituto Politécnico de Setúbal. Trata-se de um documento elaborado fundamentalmente para detalhar e acompanhar o desenvolvimento do Projeto Aplicado numa organização do setor alimentar, Sovena Group.

A componente empresarial a nível global tem vindo a envidar esforços na tentativa de acompanhar aquilo que são as mais recentes tendências e inovações de mercado. Este é um fenómeno que se verifica pela grande vertente concorrencial que se tem vindo a impor nos mercados ano após ano. Atualmente, não é de todo descabido criar uma certa analogia entre aquilo que é a lei da seleção natural lançada no século XIX por Charles Darwin e a sobrevivência das empresas que praticam as suas atividades nos mais diferentes mercados. O mesmo cenário se aplica aos termos atuais, onde as organizações que melhor se adaptam às tendências, e neste caso realçamos as abruptas mudanças para uma era cada vez mais digital, serão aquelas que apresentarão maiores probabilidades de prevalecer de entre tantas outras.

Facto é que a mais recente situação pandémica provocada pelo vírus da COVID19 veio acelerar toda estra transição. Desta forma, a transformação digital deixou de ser apenas uma oportunidade para as empresas de aprimorarem os seus fluxos e processos, para passar a ser um fator de adoção quase obrigatório para a sua sobrevivência (Kraus et al., 2022). Esta afirmação pode ser explicada facilmente, uma vez que o objetivo primário da adoção de ferramentas digitais, centra-se principalmente na melhoria e automatização de processos, de forma que seja possível às empresas, ganhar maior flexibilidade e eficiência nos mesmos. Assim, para as empresas que não conseguirem concretizar este tipo de transformação, o futuro poderá revelar-se menos favorável, antevendo-se dificuldades em acompanhar a evolução do mercado e em manter a competitividade no mercado (Kraus et al., 2022).

Precisamente, num cenário onde cada vez mais a disponibilização e consequente utilização da informação em tempo-real é vista como o fator determinante na gestão de riscos e tomada de decisões, os termos referidos anteriormente ganham um peso ainda maior quando os abordamos no contexto de determinada cadeia de abastecimento (CA) (Zouari et al., 2021). Este processo de digitalização nas cadeias de abastecimento é atualmente algo cada vez mais presente, e muitas organizações investem neste sentido com o objetivo de conseguir responder de forma mais eficaz à crescente procura dos mercados e às necessidades dos consumidores, cada vez mais específicas (Zouari et al., 2021). Um dos pontos fundamentais para tornar este cenário possível, jaz precisamente num dos temas

mencionados anteriormente, a gestão da informação, algo que possibilita a desmaterialização dos processos (Zouari et al., 2021). Desta forma, é possível definir cadeias de abastecimento digitais (CAD) quando estamos perante sistemas que têm uma elevada capacidade de processar dados em quantidades massivas, aliando a esse fator a eficácia e velocidade na forma em como a informação flui de um interveniente ao outro (Büyüközkan & Göçer, 2018). Este tipo de cadeia, quando comparada com a convencional, oferece maior flexibilidade na análise de dados pois permite não só analisá-los com maior rapidez, mas também de forma mais fidedigna e em volumes consideravelmente mais amplos. Através dessa mesma análise de dados, é possível ganhar visibilidade de possíveis cenários futuros para elaboração de previsões ou ativação de planos preventivos (Zouari et al., 2021).

Desta feita, no sentido de enquadrar o tema, o idealizado para este trabalho de projeto vem no seguimento daquilo que foram alguns dos conteúdos lecionados no Mestrado em Ciências de dados para Empresas na perspetiva da manipulação e utilização dos dados com o intuito de desenvolver uma ferramenta que incida sobre o processo de planeamento de necessidades do departamento de Logística Interna da unidade fabril do Barreiro – Sovena Group – e que permita, principalmente, aumentar o nível de precisão na previsão da procura mas também, consequentemente, incrementar a eficiência de processos a jusante como o planeamento de produção ou a gestão de stocks e armazéns.

Assim sendo, o passo-a-passo para a elaboração deste relatório será, numa primeira fase efetuar uma breve revisão de literatura, conduzida numa perspetiva científica, através da pesquisa, análise e descrição de artigos e outras tipologias de publicações de carácter científico. Nesta, foi priorizada principalmente a utilização de artigos de revistas científicas de primeiro ou segundo quartil – de acordo com o indicador SCImago Journal Rank (SJR) – e está intencionalmente direcionada para o que será o desenvolvimento do projeto propriamente dito. Centra-se fundamentalmente em três pontos principais: cadeia de abastecimento, previsão da procura e inteligência artificial. Neste seguimento, serão introduzidos alguns conceitos e definições em torno destes respetivos temas iniciando na ciência de dados e inteligência artificial, e nessa mesma direção, tipos e modelos de aprendizagem automática, com foco a recair fundamentalmente em modelos de análise de séries temporais (ST), e como estes podem influenciar o processo de planeamento da procura e consequentemente, a eficiência de toda a CA. Posteriormente, passar para a cadeia de abastecimento, os seus conceitos fundamentais, desafios e como as novas tecnologias associadas ao termo oferecem novas perspetivas à sua gestão. Finalmente, será abordado o conceito de planeamento da procura e os modelos utilizados durante o desenvolvimento do projeto, e consequentemente como este processo pode ou não influenciar a eficiência na

cadeia de abastecimento, não deixando de parte a introdução a algumas das limitações e desafios adjacentes a este mesmo tema.

Por fim, o capítulo final estará destinado à apresentação e discussão dos resultados obtidos onde o foco principal incidirá sobre os outputs gerados pela ferramenta desenvolvida e onde será feita uma análise a algumas das métricas de *performance* escolhidas para avaliar o modelo preditivo e o seu comportamento comparativamente ao modelo utilizado anteriormente.

Capítulo 1 – Revisão de literatura

1.1. Ciência de dados e inteligência artificial

Pode-se considerar que os termos ciência de dados e inteligência artificial estão intrinsecamente relacionados por uma componente – os próprios dados. O processo de análise pode assumir diferentes formas consoante as técnicas e ferramentas utilizadas, contudo, o objetivo final mantém-se inalterado, a obtenção de conhecimento através da utilização dessas mesmas aplicações e modelos a dados existentes (Raman et al., 2018).

1.1.1. Ciência de dados

No estudo realizado por Jahani et al. (2023) os autores identificam que as tendências de pesquisa para os termos de *"data science"* e *"big data"* têm revelado um crescimento exponencial ao longo das últimas duas décadas. Cada vez mais são também os estudos relacionados com estes conceitos, nomeadamente quando relacionados em cenários da cadeia de abastecimento. A ciência de dados é um campo de estudo que assenta fundamentalmente na criação de ferramentas preditivas e estatísticas para prestar suporte à tomada de decisão, mais que isso, oferece também soluções que possibilitam a gestão orientada por dados – *"data driven"* Jahani et al. (2023). A combinação destas técnicas assumem um papel preponderante naquilo que é garantir a competitividade ao avaliar a integração passada e futura dos processos de negócio, bem como os níveis de serviço e custos. É então que surge o termo de analítica avançada que cada vez mais se perceciona como um ativo competitivo decisivo em várias áreas de negócio e que se define como um elemento fulcral para as organizações melhorarem o seu desempenho através de princípios de ciência de dados (Raman et al., 2018).

1.1.2. Inteligência artificial

Ainda que o projeto não se centre única e exclusivamente em métodos de inteligência artificial, nem aprofunde em detalhe as componentes associadas ao seu conceito, é importante introduzir alguns conceitos que facilitem o entendimento desta abordagem. Na sua origem, o conceito de IA remete para um espectro da ciência que cria máquinas inteligentes, através de poder computacional, capazes de captar e traduzir padrões que serão utilizados para simular a inteligência humana (de Mattos et al., 2024). O avanço tecnológico, em certa parte propulsionado por tecnologias desta natureza (inteligência artificial e aprendizagem automática), demonstram melhorias significativas naquilo que é a capacidade de execução na análise e previsão de séries temporais (Yadav et al., 2024). Este conceito representa a visão macro de outros termos subjacentes, e a relação entre cada um deles é demonstrada através da Figura 6.

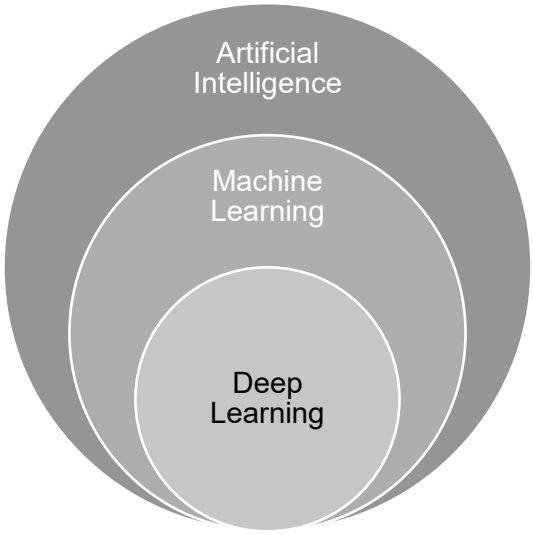

Figura 1 - Relação entre conceitos de inteligência artificial

Fonte: Balan et al. (2025)

Percetível é a relação entre o conceito mãe de inteligência artificial, que vê a aprendizagem automática como uma ramificação e consequentemente o conceito de aprendizagem profunda como um tipo de aprendizagem automática. Associadas a todos estes conceitos, surgem os conhecimentos estatísticos que estarão sempre relacionados com os mesmos.

1.1.2.1. Aprendizagem automática

Tal como demonstrado previamente, a aprendizagem automática (*machine learning* – ML), é uma *subset* do conceito de IA, e que se foca fundamentalmente no desenvolvimento de algoritmos capazes de aprender com os dados e gerarem previsões/tomarem decisões sem estarem necessariamente programadas para desempenhar determinadas tarefas (Ben Hamou et al., 2025). Dentro deste âmbito, existem diferentes técnicas que se adaptam de forma distinta a diferentes cenários em conformidade com o seu objetivo e principalmente natureza dos dados (Kühl et al., 2022):

- **Aprendizagem supervisionada:** as observações históricas são utilizadas para construir o conhecimento acerca de determinada tarefa que servirá para prever uma variável de natureza previamente conhecida – regressão, classificação.
- **Aprendizagem não-supervisionada:** compreende métodos e algoritmos que revelam padrões de dados previamente não identificados, consequentemente, não geram nenhuma solução correta uma vez que não existirá termo comparativo – agrupamento; associação.
- **Aprendizagem por reforço:** um sistema que combina punições e recompensas, de acordo com a situação, e que permite ao modelo aprender ao longo do tempo a forma como interpretar os dados.

Entre as três tipologias, a mais abrangente e utilizada é a aprendizagem supervisionada (Kühl et al., 2022), técnica que acaba também por ser utilizada no desenvolvimento deste projeto através da implementação dos modelos escolhidos para tal.

1.1.2.2. Processo de modelação

Um modelo de aprendizagem automática é bem mais do que simplesmente escolher que algoritmo utilizar, todo o processo inicia-se bem antes de colocar em prática qualquer técnica ou algoritmo. Zolbanin & Aubert (2025) introduzem uma nova perspetiva do processo de modelação em projetos de aprendizagem automática que combina vários elementos distintos, entre eles, a metodologia CRISP-DM, para dar origem a um passo-a-passo daquilo que será o processo como um todo. Também Bhatsada et al. (2025), ainda que num contexto diferente, identificam aquelas que consideram as etapas relevantes a assinalar e detalhar no mesmo processo. Para além dos acima, novamente numa área totalmente distinta, An et al. (2025) apresentam aquilo que consideram ser as etapas principais no processo de modelação em aprendizagem automática, ainda que neste caso, o estudo mencione a utilização de uma ferramenta de automatização que permite a otimização de algumas das etapas.

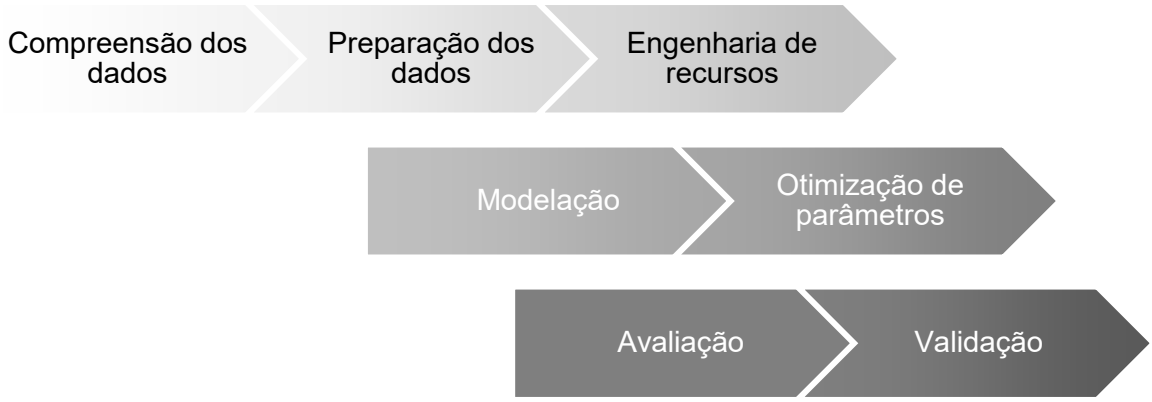

Figura 2 - Proposta de processo de modelação

Fonte: Adaptado de Zolbanin & Aubert (2025), Bhatsada et al. (2025) e An et al. (2025)

De forma geral, se realizarmos uma análise comparativa entre todos estes estudos, é possível verificar que existem bastantes semelhanças no que diz respeito à identificação de cada etapa e do carácter das respetivas. Assim sendo, assumindo que a problemática já está identificada e os objetivos do projeto delineados, não introduzindo assim estas duas fases iniciais do processo, o framework proposto na Figura 7 acaba por ser uma combinação dos três estudos mencionados, numa tentativa de incorporar os passos que estão presentes em simultâneo em cada um deles numa proposta de fluxograma capaz de sintetizar o processo.

1.1.3. Aplicações na cadeia de abastecimento

Na sua revisão sistemática da literatura, Yadav et al. (2024) enumeram algumas das técnicas de IA e as respetivas possíveis aplicações no âmbito da cadeia de abastecimento. Os autores indicam ainda que a gestão da cadeia de abastecimento pode ser significativamente melhorada através da implementação deste tipo de aplicações que albergam com elas vários pontos de reforma, nomeadamente a redução do desperdício através de processos aprimorados para a previsão da procura e gestão de inventário:

Tabela 1 - Técnicas de inteligência artificial e respetivas aplicações

Técnica	Aplicações
Aprendizagem automática	Previsão da procura; otimização de inventário
Automação robótica	Automação de tarefas; monitorização de conformidade
Análise preditiva	Otimização de rotas; manutenção preventiva
Internet of Things	Armazém inteligente; monitorização em tempo-real

Fonte: Adaptado de Yadav et al. (2024)

Um exemplo da aplicação deste tipo de ferramentas ao nível da cadeia de abastecimento está descrito no estudo publicado por Feizabadi (2022), que enfatiza que as organizações deveriam recorrer a uma abordagem baseada em aprendizagem automática para executar o processo de previsão da procura pois este permite alcançar um nível superior de precisão quando comparado aos métodos tradicionais. Com isto, demonstrou-se ainda que uma maior precisão no cálculo dos níveis de procura traduz-se em melhorias tanto a nível operacional como a nível financeiro.

1.1.3.1. Desafios na adoção de aplicações de IA

Por muito que estas tecnologias emergentes possibilitem às empresas elevar cada vez mais os níveis de eficiência, não só das suas cadeias de abastecimento, mas também de toda a organização de forma transversal, existem sempre limitações, obstáculos e desafios que desacelerem esse mesmo efeito. Nesse mesmo sentido, Yadav et al. (2024) enumera uma lista extensa de desafios que podem colocar algum entrave nos processos de adoção deste tipo de ferramentas, entre as quais as que foram consideradas para este relatório:

- **Integridade e qualidade dos dados:** necessidade de assegurar a consistência e integridade dos dados para garantir qualidade na tomada de decisão.
- **Custos de implementação:** dependendo da aplicação, podem existir custos elevados na adoção deste tipo de ferramentas.
- **Governança:** gestão dos dados e da sua confidencialidade de acordo com a natureza e propósitos das aplicações.

- **Parceiros:** garantir que toda a estrutura da cadeia de abastecimento está capaz, ou seja, que todos os intervenientes ao longo da mesma estão devidamente preparados.

No mesmo sentido, também de Mattos et al. (2024) apresentam o conceito de capacidade de IA como a habilidade que determinada organização tem para selecionar, organizar e alavancar recursos para a adoção de novas ferramentas neste âmbito. Para os autores, existem três componentes a considerar quando se considera este tipo de processos, uma componente tangível (vertente tecnológica), uma componente humana (recursos humanos), e por fim, uma componente intangível, de carácter mais organizacional. O estudo apresenta uma tabela descritiva para cada um dos componentes quando aplicável a um processo de planeamento da procura – os mesmos estão descritos na Tabela 7.

Tabela 2 - Recursos para planeamento da procura

Componente	Recurso
Tangível	**1.** Recolha e preparação de dados
(tecnológica)	**2.** Implementação e integração
	3. Algoritmos e modelos
	4. Treino e avaliação
Humana	**5.** Capacidade de negócio
(recursos humanos)	**6.** Equipa dedicada (ciência de dados)
	7. Promoção e adoção de formação em IA
	8. Confiança nos resultados obtidos
	9. Capital relacional
Intangível	**10.** Melhoria na coordenação interdisciplinar
(organizacional)	**11.** Reformulação de área e departamentos
	12. Mudança de gestão

Fonte: de Mattos et al. (2024)

1.2. Cadeia de abastecimento

Mentzer et al. (2001) observaram que era mais comum encontrar definições para o conceito de cadeia de abastecimento do que propriamente para como essa mesma cadeia era gerida. No entanto, de forma genérica e bebendo um pouco de cada uma das definições prévias, a cadeia de abastecimento pode ser considerada como um grupo de três ou mais entidades, diretamente envolvidas nos fluxos a montante e a jusante de produtos, serviços, finanças e informação desde a fonte primária, até ao consumidor final. Considerando a definição acima, e como demonstrado na Figura 1, os autores identificam três tipos diferentes de cadeias de

abastecimento, distinguidas fundamentalmente pelo número de intervenientes que nela atuam, ou por outras palavras, pelo seu nível de complexidade.

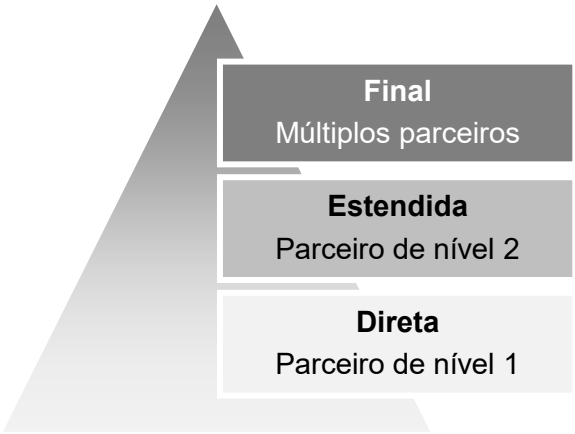

Figura 3 - Níveis de complexidade da cadeia de abastecimento

Fonte: Mentzer et al. (2001)

Por ordem crescente, primeiro surge a CA direta que consiste apenas em três intervenientes sendo estes uma empresa, um fornecedor e um consumidor, todos eles envolvidos de alguma forma nos diferentes fluxos ao longo da cadeia. De seguida, surge a CA estendida que envolve para além dos acima identificados, intervenientes intermédios, sejam eles fornecedores ou clientes dos respetivos (Mentzer et al., 2001). Por fim, e com o maior nível de complexidade, encontramos a CA final que é definida de ponta a ponta, abrangendo todos os intervenientes ao longo da cadeia.

1.2.1. Gestão da cadeia de abastecimento

Para Mentzer et al. (2001), compilando algumas das definições utilizadas em publicações prévias, a GCA pode ser definida como *"a coordenação estratégica e sistemática das funções organizacionais tradicionais e das táticas entre essas funções, dentro de uma determinada empresa e entre empresas ao longo cadeia, com o objetivo de melhorar, a longo prazo, o desempenho dos parceiros de forma individual e da cadeia de abastecimento como um todo"*.

Esta afirmação segue o pensamento lógico de que, para minimizar custos ao longo da cadeia, o fluxo informacional entre parceiros tem de acompanhar o mesmo nível de excelência que os restantes. Anos antes já Cooper et al. (1997) tinham introduzido uma definição geral para aquilo que seria a estrutura da GCA e onde a respetiva assenta principalmente em três elementos que se relacionam entre eles: processos de negócio, componentes de gestão e estrutura da cadeia de abastecimento.

1.2.1.1. Processos de negócio

De acordo com Cooper et al. (1997), estes processos de negócio podem adotar várias formas e atuar em várias direções, numa perspetiva de relacionamento entre parceiros. Tanto podem ser desenhados e executados totalmente de forma interna, como ter algumas vertentes relacionadas com um interveniente externo. Ainda assim, numa primeira fase, os autores apontam que a identificação inicial destes processos associados foi introduzida pelo International Center of Competitive Excellence (ICCE). A Tabela 1 apresenta sete processos, e entre eles, o foco para este estudo estará no processo de gestão da procura, onde a previsão e a redução da variabilidade atuam como fatores fundamentais.

Tabela 3 - Processos de negócio na cadeia de abastecimento

Processo	Descrição
Gestão da relação com o cliente	Envolve identificar mercados chave de acordo com as necessidades do consumidor e desenvolver/implementar programas para tal
Gestão do serviço ao cliente	Oferece um contato ao cliente e mantém-no informado acerca do estado da sua encomenda
Gestão da procura	Previsão e redução da variabilidade são tarefas chave neste processo
Atendimento de pedidos	Gere o processo de entrega ao consumidor de forma precisa e atempada
Gestão do fluxo de produção	Relacionado com a preocupação em produzir produtos com as especificações requeridas pelos consumidores
Gestão de compras	Gere as relações entre parceiros ao longo da cadeia
Desenvolvimento de produto e comercialização	Requer integração de parceiros e clientes chave para garantir o correto desenvolvimento de novos produtos

Fonte: Adaptado de Lambert & Enz (2017)

1.2.1.2. Componentes de gestão

Os componentes de gestão definem-se através dos elementos que auxiliam a estruturação e gestão desses mesmos processos. De acordo com o estudo de Cooper et al. (1997), apesar de reconhecerem a possível existência de mais, os componentes de gestão sugeridos são dez. Os primeiros seis são de influência mais facilmente mensurável uma vez que serão aqueles que têm maior capacidade de impactar diretamente a organização e a respetiva CA. Ainda assim, os quatro seguintes também detêm a sua quota parte de importância com a

atenuante de serem um pouco mais complexos e morosos aquando da necessidade de qualquer alteração a curto prazo.

Tabela 4 - Componentes de gestão na cadeia de abastecimento

Componente	Descrição
Planeamento e controlo	Planeamento mover a cadeia na direção correta e controlo para medir o desempenho
Estrutura de trabalho	Organização das suas tarefas e atividades diárias
Estrutura organizacional	Integração entre as organizações presentes na estrutura da cadeia de abastecimento
Estrutura do fluxo de produto	Estrutura da rede implementada na cadeia para os processos como sourcing, produção ou distribuição
Estrutura do fluxo de informação	Forma como a informação percorre os fluxos entre parceiros
Estrutura de produto	Coordenação necessária ao longo da cadeia para garantir o desenvolvimento de novos produtos
Métodos de gestão	filosofia corporativa da organização e técnicas de gestão
Estrutura de poder e liderança	Presença de liderança pode afetar a forma estrutural da cadeia e a maneira como ela está comprometida com os parceiros
Estrutura de risco e recompensa	Níveis de risco/recompensa partilhados entre parceiros
Cultura e valores	Cultura e valores da organização dizem muito da forma como os seus colaboradores são ou não valorizados

Fonte: Adaptado de Garay-Rondero et al. (2020)

Para este estudo, importa mencionar principalmente a relevância da componente relacionada com o fluxo de informação e com o planeamento e controlo, pois são essas que vão suportar aquilo que serão os níveis de procura e o planeamento de necessidades.

1.2.1.3. Estrutura da cadeia de abastecimento

Finalmente, a estrutura da própria cadeia é a configuração final em torno da empresa que abrange parceiros intervenientes ao longo de todos os fluxos (Cooper et al., 1997). Isto, naturalmente, envolve todas as empresas que participam nessa mesma cadeia, desde o fornecedor inicial até ao seu consumidor final. A Figura 3 demonstra um diagrama adaptado que identifica todas as áreas abordadas anteriormente, numa perspetiva macro em entendimento de como será a estrutura final da GCA suportada pelo fluxo de informação.

Figura 4 - Estrutura final da gestão da cadeia de abastecimento

Fonte: Adaptado de Cooper et al. (1997)

Como já referido anteriormente quando abordado o conceito de CA e os respetivos níveis de complexidade, também a estrutura cadeia será tanto ou mais complexa quanto maior for o número de parceiros envolvidos, mas não só, fatores como a complexidade do produto ou até mesmo a disponibilidade de matéria-prima podem influenciar esse mesmo nível de complexidade. De notar apenas que, no diagrama apresentado, para além dos sete processos de negócio introduzidos inicialmente, surge também um oitavo referente ao canal de devoluções, conceito relacionado com o termo logística inversa que abrange as todas as atividades logísticas desde a recolha de produtos usados já não necessários ao consumidor até produtos novamente utilizáveis no mercado por via de retrabalho ou reciclagem (Fleischmann et al., 1997).

1.2.2. Desafios na gestão da cadeia de abastecimento

Grande parte dos desafios encontrados ao longo das cadeias de abastecimento está relacionado com fatores externos, ou seja, não necessariamente gerados pelas próprias cadeias ou pelos parceiros que nelas estão inseridos (Choi et al., 2016). No entanto, de forma geral, a eficiência da CA é medida através da capacidade que a própria demonstra para ultrapassar estes obstáculos. Neste caso, abordaremos dois dos maiores desafios associados ao conceito de GCA, principalmente pois se enquadram naquilo que é o desenvolvimento do projeto – a gestão de risco e partilha de informação e a variabilidade da procura.

1.2.2.1. Gestão de risco e partilha de informação

Na perspetiva de organizações que tenham um elevado grau de complexidade na sua CA, é essencial que o fluxo de informação esteja bem estruturado e que os dados percorram os pontos necessários de forma rápida e coesa. A partilha de informação torna-se assim um fator fundamental nas cadeias de abastecimento e na forma como as mesmas são geridas uma vez que pode impactar não só o processo de tomada de decisão como toda eficiência da própria cadeia (Han & Dong, 2015). Também Choi et al. (2016) afirmam que esta partilha de informação é uma das principais formas de garantir melhoria na eficiência da CA, muito porque preparam os diferentes parceiros antecipadamente acerca dos mais variados cenários possíveis e porque potenciam uma gestão de risco mais eficaz e distribuída entre os mesmos.

Os autores mencionam ainda que é imperativa a partilha de informação em organizações que subcontratem processos da cadeia, algo que pode ser considerado de elevado risco pois teoricamente, num cenário desta natureza, pode-se perder um pouco daquilo que seria o controlo ou autonomia nos respetivos caso se mantivessem como gestão interna (Choi et al., 2016). Por outro lado, surgem novas oportunidades, nomeadamente através da utilização de mão-de-obra especializada, que à partida, estará mais bem preparada para realizar determinada tarefa, e mais importante que isso, promove a concorrência entre parceiros da mesma natureza potencializando uma possível redução de custos operacionais a médio-longo prazo (Choi et al., 2016). De forma geral, para o bom funcionamento da CA, é fundamental o desenvolvimento de relações próximas de confiança entre os parceiros. Han & Dong (2015) introduzem um modelo matemático para o conceito de confiança que alberga dois tipos de fatores: pré-determinados e instantâneos. Os autores entendem que por um lado, existem fatores já estipulados antes de qualquer transação entre parceiros que podem englobar subfatores como a reputação, histórico de relações ou até mesmo recomendações, e por outro lado, existem fatores instantâneos que estarão efetivamente relacionados a subfatores provenientes de transações a decorrer. Naturalmente, quando maior for os coeficientes destes dois fatores, maior será o grau geral de confiança atribuído a determinado parceiro.

1.2.2.2. Variabilidade da procura

O conceito de variabilidade da procura pode estar diretamente associado à variabilidade da CA como um todo. Para Germain et al. (2008) este conceito espelha o grau de capacidade que determinada empresa possui para prever com precisão os níveis de procura e é um componente que pode influenciar diretamente toda a cadeia e consequentemente o seu desempenho financeiro. Este mesmo estudo introduz uma estrutura que exemplifica como a variabilidade da procura pode afetar diretamente o desempenho financeiro.

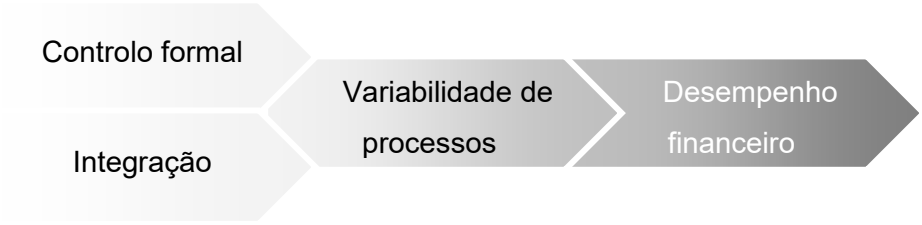

Figura 5 - Impacto da variabilidade da procura no desempenho financeira

Fonte: Adaptado de Germain et al. (2008)

Para os autores, o foco principal acaba por ser alinhar a estratégia de acordo com o cenário encontrado e apresentam dois mecanismos distintos para atuar e gerir a variabilidade da CA:

- **Controlo formal:** destaca a utilização de sistemas standardizados de monitorização como auxílio necessário para manter a consistência e previsibilidade nas operações.
- **Integração interfuncional:** implica a coordenação entre as diferentes áreas e departamentos da organização para que facilitem a troca de informação e priorizem a colaboração.

O diagrama representado na Figura 3 demonstra que o controlo formal será mais benéfico em ambientes onde a procura é mais previsível pois permite mais facilmente a deteção de desvios dos indicadores operacionais e consequentemente tomar medidas corretivas, enquanto a integração interfuncional é crucial em ambientes imprevisíveis pois permite dar resposta de forma mais ágil e coordenada às alterações nos padrões de consumo e nas condições de mercado (Germain et al., 2008).

1.2.3. Tecnologias na gestão da cadeia de abastecimento

Atualmente, é impossível dissociar o conceito de GCA do seu enquadramento tecnológico. A tecnologia associada ao termo tem vindo a consolidar-se como elemento central para os avanços e inovações constantes neste âmbito. Estas transformações visam, de forma sistemática, melhorar a eficiência operacional e a redução de custos, através da automatização e digitalização dos processos ao longo da cadeia.

1.2.3.1. Enterprise Resource Planning

A utilização de sistemas integrados de gestão suportados por plataformas tecnológicas surge no seguimento da crescente necessidade de as organizações integrarem os subsistemas de informação correspondentes às diversas áreas funcionais e partilharem informação não só internamente, mas como também com os seus fornecedores e restantes parceiros ao longo da cadeia (Umble et al., 2003). Para Chopra et al. (2022), este tipo de sistema oferece benefícios a vários níveis: operacionais, estratégicos, de gestão, infraestrutura tecnológica e organizacionais.

Este é um sistema que evoluiu bastante até chegar à sua versão mais atual desde os seus primórdios na década de 1960 onde os primeiros sistemas tinham como foco principal, e praticamente único, a gestão de inventário das organizações. Segundo Umble et al. (2003) com o passar dos anos, as empresas foram-se apercebendo que estava a tornar-se praticamente impossível gerir grandes quantidades de stock sem grande previsão de utilização futura e nesse sentido, surgem os primeiros sistemas de planeamento de requerimentos de material (MRP) que identificavam os materiais necessários à produção de determinados produtos. Continuando a evolução e alavancando o crescimento deste tipo de ferramentas surge uma nova versão do sistema durante a década de 1980 onde seria introduzido o MRP II, neste caso, denominado de sistemas de planeamento de recursos de produção que vem potencializar a integração dos sistemas de produção com o financeiro e contabilístico (Umble et al., 2003). Já bem próximo da viragem de século, surge então o primeiro sistema ERP como ferramenta capaz de integrar todo o planeamento de recursos através da inclusão de módulos de gestão para as diferentes áreas como as de produção, marketing, recursos humanos entre outros. Esta nova fórmula foi acumulando atualizações potenciadas fundamentalmente pela crescente adoção das organizações a este tipo de sistemas e atualmente já é possível encontrar fornecedores que ofereçam este sistema como um serviço em *cloud*, oferecendo desta forma novas vantagens como redução de custos operacionais e maior elasticidade na escalabilidade, permitindo às organizações ajustar os recursos (reduzir ou incrementar) conforme as necessidades do respetivo momento (Abd Elmonem et al., 2016). Na perspetiva da GCA, os sistemas ERP contribuem de forma positiva para determinadas tendências, nomeadamente na personalização em massa, na implementação de sistemas ou na normalização dos mesmos. Ainda assim, o seu contributo assume fundamentalmente uma vertente mais técnica do que propriamente estratégica (Akkermans et al., 2003).

1.2.3.2. Tecnologias emergentes

Segundo Manavalan & Jayakrishna (2019), o conceito de indústria 4.0 é uma mistura de tecnologia digital que vem transformar a indústria e oferecer novos caminhos e formas de adotar diferentes modelos de negócio. O essencial é revolucionar através da tecnologia possibilitando a integração ao nível da maquinaria industrial. Este conceito está aliado a um novo termo de *smart factory*, ou fábrica inteligente, que se serve fundamentalmente da sua capacidade de unir a vertente tangível (produtos/máquinas) à intangível (informação/digitalização). De acordo com o Manavalan & Jayakrishna (2019) o design deste tipo de indústria centra-se em seis pressupostos aliados entre si visando automatizados e digitalizar processos:

- **Virtualização:** SI que criam réplicas virtuais da informação acerca do mundo físico em dados digitais.
- **Interoperabilidade:** capacidade dos equipamentos e componentes se interligarem entre si e humanos para comunicar.
- **Descentralização:** capacidade de os sistemas digitais tomarem decisões autónomas para manter os fluxos imaculados e a decorrer.
- **Capacidade em tempo real:** capacidade de o sistema atualizar e comunicar a informação de forma constante para auxiliar a tomada de decisão.
- **Orientação ao serviço:** capacidade de o sistema servir as organizações e respetivos utilizadores.
- **Modularidade:** sistema que adere às constantes modificações nos requisitos adicionando ou substituindo diferentes módulos.

No mesmo sentido, e aliado também aos conceitos de fábrica inteligente e indústria 4.0, surge a internet das coisas ou de seu termo original, Internet of Things (IoT). De acordo com Bi et al. (2014), o conceito é uma extensão à internet que conhecemos, possibilitando a conexão entre *"coisas"*, sendo estas habitualmente sensores, máquinas, terminais ou outras tecnologias inteligentes. Todos estes equipamentos são capazes de gerar e captar dados passíveis de serem utilizados para análise. O importante é que, existindo, poderá ser utilizada para prestar suporte à tomada de decisão. Bi et al. (2014) apontam este como um fator a ter em conta nas empresas das próximas gerações, tomando a descentralização do processo de tomada de decisão como algo a ter em conta para oferecer maior capacidade de lidar com esta tipologia de sistemas mais dinâmicos e complexos. Também a adoção de técnicas de Inteligência Artificial (IA) podem gerar disrupções e impactos significativos na GCA, tanto positivos quanto negativos – novos riscos e desafios de integração (Culot et al., 2024).

1.3. Previsão da procura

Dado que o projeto é conduzido fundamentalmente na direção de desenvolver um modelo preditivo, é essencial virar as atenções ao processo de previsão da procura. Este processo baseia-se na capacidade de calcular os níveis de procura e para Zhu et al. (2021) é a base de toda a eficiência da CA dado que suporta essencialmente todo o tipo de decisões a nível operacional ao longo da mesma. Este é um processo que é habitualmente realizado através de modelos estatísticos que trabalham dados históricos. Esse mesmo estudo indica, com base no questionário desenvolvido por Weller & Crone (2012), que os métodos estatísticos são utilizados como forma de previsão da procura em cerca de 82% das empresas de diferentes indústrias. Este pipeline será naturalmente alimentado com os dados recolhidos previamente, e uma vez identificado e desenvolvido o modelo de previsão ideal, é imprudente não ir

constantemente avaliando a sua performance e realizado pequenos aperfeiçoamentos ao respetivo. Para Croxton et al. (2002) este é um processo de aprendizagem importante pois permite analisar os erros do modelo e ir corrigindo-os à medida que for necessário para garantir o nível de precisão dos métodos preditivos.

1.3.1. Gestão da procura como processo

O termo planeamento da procura pode ser considerado algo subjacente ao conceito de gestão da procura introduzido por Croxton et al. (2002). Para os respetivos autores, o processo de gestão da procura implica fundamentalmente atingir o equilíbrio entre aquilo que serão as necessidades dos consumidores e as capacidades instaladas em determinada CA. Ainda que já anteriormente Cooper et al. (1997) identificassem o processo como uma componente válida a considerar na GCA, para os autores este é de facto um processo fulcral pois quando gerido de forma correta e eficiente pode oferecer à organização maior flexibilidade de resposta aos níveis reais de procura. Em seguimento, Croxton et al. (2002) afirmam que este é um conceito que alberta também subprocessos de duas tipologias distintas: estratégicos e operacionais. Como é possível observar através da síntese ilustrada na Figura 4, cada um destes subprocessos apresenta atividades relacionadas com o processo de planeamento e gestão da procura.

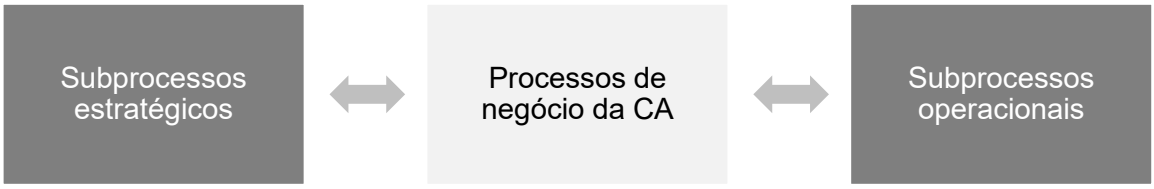

Figura 6 - Gestão da Procura

Fonte: Croxton et al. (2002)

Enquanto a vertente estratégica estará mais associada ao desenvolvimento e preparação de um sistema operacional capaz de acertar na sincronização entre os níveis de procura e oferta, a vertente operacional trata efetivamente de levar às equipas esta arquitetura para que as respetivas coloquem em prática o idealizado visando a tal sincronização através da mesma.

1.3.1.1. Sincronização da cadeia de abastecimento

Este é um conceito que está presente tanto na vertente estratégica como na vertente operacional devido à preponderância que tem naquilo que é a eficiência da GCA. Croxton et al. (2002) apontam esta componente como a atividade que permite combinar as quantidades previstas da procura com as capacidades instaladas ao longo da CA, componente habitualmente denominada de S&OP (*Sales and Operations Planning*). Esta determinação de

procedimentos afetos à sincronização surge como subprocesso estratégico na framework apresentada pelos autores, envolvendo a definição das práticas a adotar, a identificação dos requisitos de planeamento num horizonte temporal a longo prazo, a análise das capacidades de parceiros e por fim, a definição dos critérios de alocação ao longo da cadeia.

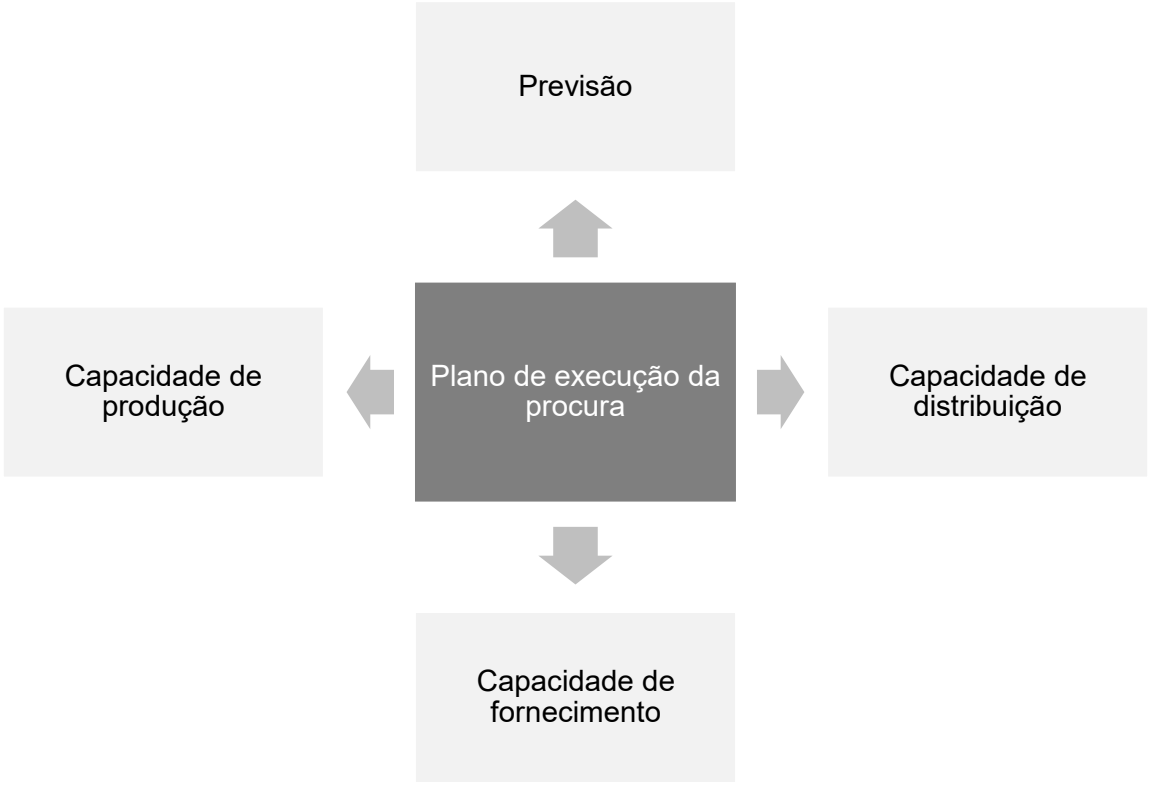

Figura 7 - Sincronização da cadeia de abastecimento

Fonte: Croxton et al. (2002)

Como demonstrado na Figura 5, esta gestão exige a coordenação entre as diferentes áreas de uma organização combinando, inicialmente, as informações provenientes da previsão inicialmente calculada, ajustando-a posteriormente em função das limitações da oferta e produção. Numa fase mais adiantada, as restrições ao nível da armazenagem e distribuição também serão mantidas em consideração.

1.3.2. Métodos de previsão

Como referido anteriormente, o grande objetivo a alcançar com a implementação de modelos preditivos é fundamentalmente prestar suporte à tomada de decisão, num cenário aplicado à GCA, servirá para análise de vendas, compras ou processos operacionais. Desta forma, a existência de alternativas não será um problema, mas sim como optar pela alternativa que melhor se adeque a determinado cenário. Neste sentido, Hofmann & Rutschmann (2018) identificam duas categorias principais no que diz respeito a técnicas/métodos de previsão:

Tabela 5 - Técnicas de previsão da procura

Qualitativas	Quantitativas
Estudo de mercado	Previsão de séries temporais
Estimação especialista	Previsão casual da procura

Fonte: Hofmann & Rutschmann (2018)

Como o próprio nome indica, técnicas qualitativas servem-se de dados e variáveis da mesma tipologia, na maior parte dos casos através de opiniões de especialistas nas respetivas áreas. Por outro lado, quando os dados existentes são de carácter quantitativo, o foco passa por conduzir a análise numa base da mesma vertente normalmente na procura de encontrar padrões, tendências ou correlações entre variáveis.

1.3.2.1. Séries temporais

Denomina-se de série temporal um conjunto de observações registadas ao longo de um intervalo de tempo. De acordo com Chatfield (2000), caracterizando-as a nível temporal, podemos identificar duas categorias, séries contínuas, quando os dados são registados sem interrupções, ou discretas, quando as observações são registadas apenas em certos momentos. Por outro lado, caracterizando-as através da quantidade de variáveis presentes nos dados, podemos encontrar séries univariadas quando existe apenas uma variável, ou multivariadas sempre que exista mais que uma. Chatfield (2000) aponta ainda que a análise a efetuar irá depender fundamentalmente do propósito a que pretende dar resposta:

Tabela 6 - Principais objetivos na análise de séries temporais

Objetivo	Descrição
Descrição	Descrever os dados utilizando medidas estatísticas
Modelação	Encontrar um modelo para descrever o processo de criação de dados
Previsão	Estimar possíveis valores futuros da série
Controlo	Suportar o analista na tomada de decisão

Fonte: Chatfield (2000)

Relativamente às componentes que podem caracterizar uma série temporal, o autor defende que, antes de implementar qualquer método, é aconselhável realizar uma análise prévia aos dados. Esta análise ajudará o analista a obter melhor perceção do comportamento dos dados podendo dar origem a alguns *insights* iniciais proveitosos. Para Chatfield (2000) esta análise pode ser denominada de Análise de Dados Inicial (ADI), processo que habitualmente antecede a atual popular Análise de Dados Exploratória (ADE). Desta feita, o

autor identifica quatro componentes principais que podem ser detetadas logo à partida, quando conduzida esta primeira fase analítica:

- **Sazonalidade:** refere-se à identificação de períodos sazonais, habitualmente detetados anualmente, através de padrões de comportamento que se repetem ao longo da série durante uma época específica do ano.
- **Tendência:** apresenta-se na forma de uma variação estável ao longo da série, de carácter positivo ou negativo, conforme o comportamento dos dados. Para existir tendência é necessário que a mesma se verifique ao longo de um período considerável.
- **Variação cíclica:** padrão de variações cíclicas que não apresentam sazonalidade.
- **Irregularidade:** descreve qualquer variação presente nos dados sem contabilizar as componentes anteriormente descritas.

Como exemplo, é possível verificar através do Gráfico 1 que, apesar das irregularidades presentes na linha de observações, é percetível identificar uma tendência de carácter positivo ao longo dos cerca de 90 dias de negociação.

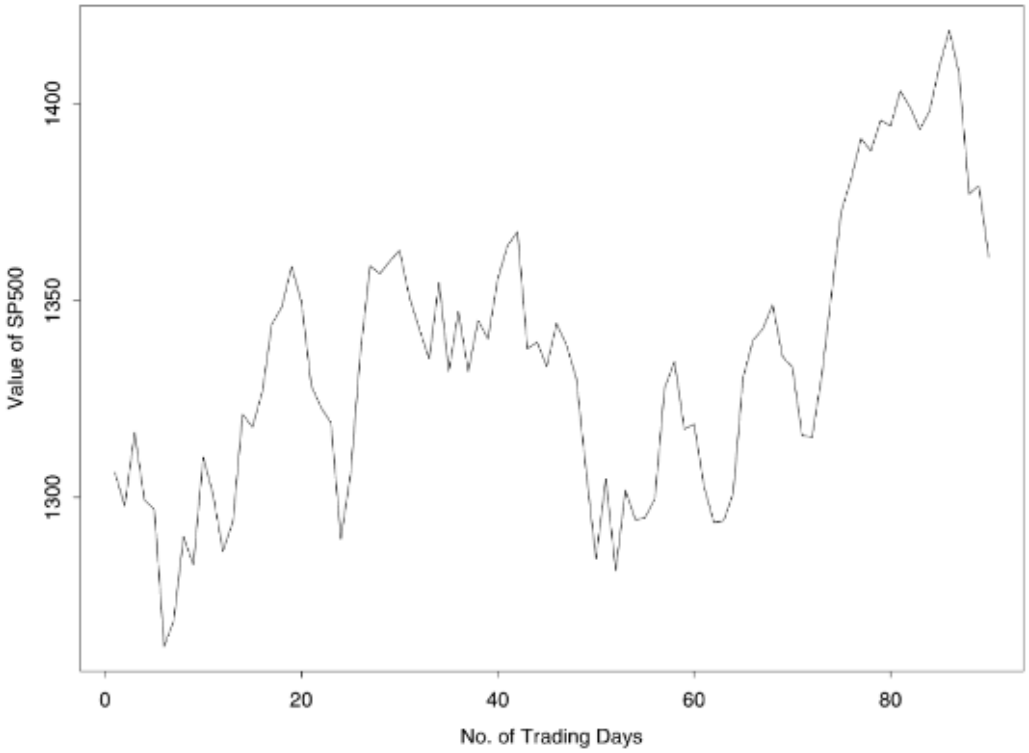

Gráfico 1 - Valor do índice S&P500 ao longo de 90 dias

Fonte: Chatfield (2000)

Esta afirmação é facilmente confirmada uma vez que é percetível que o valor inicial da série é inferior ao final, demonstrando assim uma tendência crescente na valorização do respetivo índice.

1.3.3. Modelo ARIMA

O modelo ARIMA (Autoregressive Integrated Moving Average), introduzido inicialmente por Box & Jenkins (1970) amplamente aceite pelos autores ao longo destes largos anos muito por força da sua simplicidade de compreensão e aplicação (Fattah et al., 2018). Este foi um modelo selecionado para o projeto pela sua capacidade de captar com precisão padrões de tendência e sazonalidade. Os modelos ARIMA partilham algumas semelhanças com os modelos de alisamento exponencial, uma vez que conseguem modelar tendências e padrões sazonais e ainda ser automatizados (de Oliveira & Cyrino Oliveira, 2018). Dentro desta tipologia, é possível identificar variantes mais simples como o modelo autorregressivo (1) que assume que a evolução da série pode ser descrita a partir dos valores das observações passadas.

$$y_t = \alpha + \sum_{i=1}^{p} \phi_i y_{t-i} + \omega_t \tag{1}$$

Também é possível encontrar o modelo de médias móveis (2), que interpreta a série temporal como o resultado de uma sequência de choques aleatórios, e por fim, o modelo autorregressivo e de médias móveis (3) que acaba por ser definido pela combinação de ambos os apresentados anteriormente (Fattah et al., 2018).

$$y_t = \alpha + \omega_t \sum_{i=1}^{q} \theta_j \omega_{t-j} \tag{2}$$

$$y_t = \alpha + \sum_{i=1}^{p} \phi_i y_{t-i} + \omega_t + \sum_{i=1}^{q} \theta_j \omega_{t-j} \tag{3}$$

Neste tipo de modelos, a abordagem mais clássica pode ser considerada relativamente limitada principalmente quando a ordem de ajuste sazonal é elevada, ou até mesmo quando a série não apresenta estacionariedade mesmo após o respetivo ajuste. Em situações desta natureza, os parâmetros do modelo tornam-se eles próprios um obstáculo à previsão de valores para a série em questão devido à elevada variabilidade sazonal Fattah et al. (2018).

Cheng et al. (2024) identificam os parâmetros estáticos seguintes:

- p, para a ordem do componente autorregressivo
- d, para a ordem do componente de diferenciação
- q, para a ordem do componente de média móvel

Uma série é considerada estacionária quando as suas funções de média, variância e autocorrelação são constantes ao longo da mesma, ou seja, não aumentam nem diminuem com o passar do tempo. Para esses casos, a opção deverá recair sobre modelos integrados

ARIMA (4) ou, para casos onde exista componente sazonal, os equivalentes sazonais SARIMA (Seasonal Autoregressive Integrated Moving Average).

$$\Delta^d y_t = \alpha + \sum_{i=1}^{p} \phi_i \Delta^d y_{t-i} + \omega_t + \sum_{i=1}^{q} \theta_j \omega_{t-j} \quad (4)$$

$$\Phi_P = (B^m)\phi(B)\Delta^d \Delta_m^D y_t = \Theta_Q(B^m)\theta(B)\omega_t \quad (5)$$

O modelo SARIMA (5) acaba então por ser uma nova ramificação que fundamentalmente consiste num modelo multiplicativo que combina os parâmetros não sazonais com quatro novos de carácter sazonal, onde estes últimos são definidos da seguinte forma (Negre et al., 2024):

- P, para a ordem do componente sazonal autorregressivo
- D, para a ordem do componente sazonal de diferenciação
- Q, para a ordem do componente sazonal de média móvel
- m, para o número de períodos por ciclo sazonal

As sugestões para as denotações matemáticas encontram-se descritas na Tabela 5:

Tabela 7 - Descrição de modelos autorregressivos e de médias móveis

Modelo	**Descritivo**
Autorregressivo	AR(p)
Médias móveis	MA(q)
Autorregressivo e de médias móveis	ARMA(p, q)
Autorregressivo e de médias móveis integrado	ARIMA(p, d, q)
Autorregressivo e de médias móveis integrado sazonal	SARIMA$(p, d, q) \times (P, D, Q)_m$

Fonte: Agyemang et al. (2023)

1.3.3.1. Construção do modelo

O processo de modelação para modelos ARIMA é definido através de uma abordagem que consiste em três etapas, conforme proposto por Box & Jenkins (1970). Esta metodologia consiste num processo de modelação iterativo que tipicamente é repetido determinado número de vezes até que o modelo seja suficientemente bom para ser selecionado. Assim sendo, estas três fases são definidas como (Agyemang et al., 2023):

- **Identificação:** consiste na identificação de um modelo que melhor descreva a série. Esta etapa avalia se a série temporal é estacionária, caso não o seja, é necessário aplicar o processo de diferenciação até que o mesmo lhe atribua tal característica. De seguida, determinam-se as ordens das componentes autorregressiva e de média

móvel, que habitualmente podem ser definidas através da análise dos gráficos de autocorrelação e autocorrelação parcial.

- **Estimação:** consiste na obtenção das melhores estimativas possíveis para os respetivos parâmetros do modelo. Nesta etapa, habitualmente utiliza-se uma de duas metodologias, a estimação por máxima verosimilhança (MLE) ou o mais comum método de estimação pelos mínimos quadrados.
- **Diagnóstico:** uma vez estimados os parâmetros, avalia-se a adequação do modelo ajustado e caso existam falhas no ajuste, o processo reinicia-se até que seja adequado.

1.3.3.2. Teste de Dickey-Fuller Aumentado

Uma das formas mais comuns de analisar a estacionariedade de uma série temporal passa por representar graficamente os valores das suas observações ao longo do tempo. Ainda assim, esta análise pode revelar-se suficientemente subjetiva para justificar a utilização de outros métodos capazes de verificar esta componente com maior precisão. É aqui que surgem testes como o de Dickey-Fuller Aumentado (ADF) apresentado por Said & Dickey (1984), que emerge como uma inovação ao inicialmente proposto por Dickey & Fuller (1979) e onde principal objetivo passa por testar a presença de uma raiz unitária, ou seja, a não estacionariedade da série temporal, através do estimador de mínimos quadrados ordinários (Org et al., 2018). Este é um teste estatístico onde a hipótese nula diz que a série tem uma raiz unitária, e a hipótese alternativa diz precisamente o contrário, isto é, afirmando a presença de estacionariedade na série (Fattah et al., 2018). Tal como habitualmente ocorre nos testes de hipótese, na eventualidade do valor de p ser superior ao nível de significância, a hipótese nula não será rejeitada, por outro lado, o inverso acontece caso se verifique o contrário.

1.3.4. Modelo Facebook Prophet

O modelo Prophet, desenvolvido pelo Facebook, é uma ferramenta *open-source* desenhada para possibilitar a modelação e previsão de séries temporais. O modelo utiliza um modelo de decomposição de séries temporais (Harvey & Peters, 1990) com três componentes principais: tendência, sazonalidade e feriados. Estes três componentes são combinados numa equação, onde a componente $g(t)$ é a função da tendência que captura as variações não periódicas, $s(t)$ descreve as alterações periódicas de sazonalidade e $h(t)$ adota o que será o efeito dos feriados, que tanto podem ocorrer em datas irregulares de forma isolada ou sequencial. Por fim, o termo ε_t corresponde à componente de erro proveniente de todas as variações não explicadas pelo modelo (Taylor & Letham, 2018):

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t \quad (6)$$

Este é um modelo que apresenta capacidade naquilo que é a gestão de dados em falta, mudanças estruturais nos padrões da respetiva série bem como gestão de valores fora da escala regular, habitualmente denominados de *outliers* (Cheng et al., 2024). Tal como no modelo ARIMA, a seleção deste modelo deve-se fundamentalmente à sua capacidade de captar padrões de tendência e sazonalidade bem como à facilidade de implementação. Para além disso, a flexibilidade deste permite a modelação de séries com múltiplos períodos, sem necessariamente exigir que as observações estejam espaçadas entre si de forma regular ao longo do tempo e da mesma forma que a gestão dos dados em falta é uma vantagem, essa mesma traduz-se na facilidade de implementação e consequentemente de otimização (Negre et al., 2024). O modelo tende ainda a apresentar níveis superiores de *performance* em bases de dados com padrões sazonais bem definidos, ainda assim, para aplicá-lo de forma eficaz, o modelo necessita obrigatoriamente de definir as variáveis corretamente, isto quer dizer, definir a variável alvo de previsão como y e a variável de data como ds (Agyemang et al., 2023).

1.3.4.1. Teste não paramétrico de Kruskal-Wallis

Um dos argumentos que pode ser inserido na instância do modelo Prophet é o de sazonalidade, e uma das formas de testar a presença da componente sazonal é precisamente aplicando o teste de Kruskal-Wallis. Este, é um método não paramétrico utilizado fundamentalmente para verificar se existem diferenças estatisticamente significativas na distribuição das várias amostras independentes (Bai et al., 2024):

- **Hipótese nula:** não existem diferenças significativas na distribuição entre as amostras
- **Hipótese alternativa:** existem diferenças significativas na distribuição entre as amostras

Este processo consiste fundamentalmente em combinar todas as amostras existentes, ordenar os dados de forma crescente e posteriormente atribuir a cada observação um *ranking* de classificação. De seguida, é analisada a média dos respetivos rankings por grupo até que seja possível afirmar se existem ou não diferenças significativas entre as médias das diferentes amostras. Recorre-se à análise da variância para estudar a diferença de *rankings* entre os diferentes grupos onde a soma total dos quadrados se decompõe em duas componentes distintas: a soma entre grupos e a soma dentro dos grupos. Se esta proporção, da soma dos quadrados entre grupos for de magnitude elevada, existirá diferença estatisticamente significativa entre as amostras, caso contrário, considera-se que as distribuições entre os grupos não apresentam diferenças estatisticamente significativas (Bai et al., 2024).

1.3.5. Modelo XGBoost

Recentemente as técnicas de Gradient Boosting Machine (GBM), e em particular o modelo XGBoost (Extreme Gradient Boosting), têm-se afirmado como ferramentas preponderantes na modelação preditiva muito devido à sua versatilidade e desempenho superior (Yu et al., 2025). Este, um modelo típico de aprendizagem automática, é um algoritmo pertencente à família dos modelos Gradient Boosted Decision Trees (GBDT), que se destaca fundamentalmente pelo seu elevado grau de eficiência computacional, flexibilidade na capacidade de regularização e poder de generalização de modelos e é atualmente, um dos algoritmos mais utilizados para solucionar problemas tanto de regressão como de classificação (Nie et al., 2025). Estas foram características tomadas em consideração aquando da seleção deste modelo para o desenvolvimento do projeto, fundamentalmente pois consegue lidar com grandes volumes de dados.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{7}$$

O resultado do modelo corresponde à soma ponderada de todos os modelos base onde $\hat{y}_i$ representa o valor previsto para a amostra i, o número de modelos base é dado por k, a previsão do k-ésimo modelo corresponde à componente $f_k(x_i)$ e onde o set de todos os modelos base possíveis é representado por F. O seu processo de otimização ocorre de forma similar a um problema de programação linear, onde a missão passa por minimizar aquilo que será a função objetivo do modelo. Esta função objetivo é composta por duas componentes, uma de perda, que de forma prática representará o erro do modelo, e uma segunda de regularização que atua principalmente como um regulador da complexidade do modelo prevenindo desta feita problemas de *overfitting* (Nie et al., 2025). Naturalmente, o valor da função objetivo pode ser constituído como um critério de avaliação do desempenho do modelo onde valores mais reduzidos da respetiva função, indicarão um modelo que desempenha de forma mais eficiente o seu papel. Este é ainda um modelo que apresenta diferentes parâmetros, intrínsecos à sua instância, passíveis de serem utilizados para melhorar a sua *performance* e neste âmbito, a tentativa de melhoria na eficiência do modelo pode ser realizada de forma manual, ou automática, através de algoritmos desenhados para esse mesmo objetivo. Abaixo, os principais parâmetros identificados no estudo de Nie et al. (2025):

- **max_depth:** limita a profundidade de cada árvore integrante do conjunto.
- **n_estimators**: determina a quantidade de árvores que compõem o modelo.
- **learning_rate:** regula a influência de cada árvore, determinando o contributo incremental de cada iteração para o modelo final.

1.3.5.1. Características do modelo XGBoost

Este, uma vez que necessita de um conjunto de dados já rotulado, é um modelo que encaixa na área de aprendizagem supervisionada e que tanto pode ser utilizado para resolver problemas de regressão como de classificação. No seu estudo, Yu et al. (2025) identificam quatro características fundamentais naquilo que é a aplicação de modelos desta natureza. De forma geral, estas notas relacionam-se como vantagens do próprio algoritmo que o caracterizam e distinguem dos restantes:

- **Precisão preditiva superior:** demonstra de forma consistente desempenho superior quando comparado a métodos estatísticos tradicionais através da eficácia a modelação de relações não lineares, tratamento de dados em falta e a gestão de interações complexas entre variáveis.
- **Mitigação de *overfitting*:** integra técnicas de regularização e de paragem antecipada para lidar com ruído e padrões complexos nos dados – regularização L1 e L2 também na penalização à complexidade do modelo.
- **Importância de características:** identifica e seleciona as variáveis que mais impactam o modelo e os seus outputs garantindo uma compreensão do peso de cada uma e a facilidade na tomada de decisão – remoção de ramos que não contribuem de forma significativa para o modelo.
- **Eficiência computacional e escalabilidade**: oferece bom grau de paralelização e garante a capacidade para trabalhar volumes de dados bastante elevados – processamento em paralelo garante eficiência em ambientes de computação distribuída.

Não sendo um modelo muito avançado como aqueles possíveis de encontrar no âmbito do *deep learning*, já é um modelo que aporta alguma capacidade na análise de séries temporais.

1.3.6. Métricas de avaliação de desempenho

Em contexto de análise de previsões, as métricas de avaliação de desempenho assumem o papel principal naquilo que será ultimar o modelo ou método preditivo utilizar. Estas medidas permitem aferir a eficácia do modelo através de uma avaliação à sua capacidade de gerar níveis de previsão, onde estes se tentam aproximar tanto quanto possível dos valores reais observados. Fundamentalmente, estas medidas são calculadas entre aquilo que foram as previsões geradas através do modelo utilizado, e os valores reais das respetivas observações. Para tal, é necessário dividir a base de dados em duas ramificações distintas. De acordo com Gholamy et al. (2018), estudos empíricos mostram que os melhores resultados são obtidos através de um *split* que divida a base de dados original entre 70-80% para a ramificação de treino, e os restantes 20-30% para a ramificação de teste. Desta forma, é possível garantir

que tanto os valores estimados como as métricas que os descrevem serão válidos e evitam resultados enviesados. No que diz respeito aos tipos de medidas existentes, Hyndman & Koehler (2006) identificaram quatro tipologias distintas:

- **Dependentes de escala:** utilizadas para comparar métodos distintos à mesma base de dados – as mais comuns são baseadas nos erros quadrados e absolutos. Este projeto utiliza o Erro Absoluto Médio como uma das medidas para avaliar o desempenho de cada modelo principalmente pois é uma medida de fácil interpretação que calcula a média das diferenças absolutas entre os valores previstos e os observados.

$$Mean\ Absolute\ Error\ (MAE) = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t| \quad (8)$$

- **Baseadas em percentagem:** utilizadas para comparar desempenho de previsões entre diferentes bases de dados – novamente, as mais comuns são baseadas nos erros quadrados e absolutos. Também selecionada como medida a utilizar para a avaliação do desempenho de cada modelo o Erro Percentual Absoluto Médio combina a medida apresentada anteriormente numa perspetiva percentual.

$$Mean\ Absolute\ Percentage\ Error\ (MAPE) = \frac{1}{n}\sum_{t=1}^{n}\frac{|y_t - \hat{y}_t|}{y_t} \times 100\% \quad (9)$$

- **Baseadas em erros relativos:** divisão de cada erro pelo valor do erro obtido através de outro método preditivo.
- **Medidas relativas:** comparação entre uma medida e a mesma obtida através de um método de referência, a que os autores denominam de método *benchmark*.

Mais à frente se verá, qua na fase de avaliação de cada um dos modelos utilizados, as métricas utilizadas foram o erro absoluto médio, para medir a grandeza absoluta do erro de previsão e da mesma forma o percentual, para possibilitar um termo comparativo entre as escalas dos diferentes artigos analisados.

Capítulo 2 – Objetivos e metodologia

O segundo capítulo está reservado à apresentação dos objetivos e à respetiva metodologia onde se aborda de forma sucinta que motivações originaram o nascimento desta ideia, e de forma teórica, que metodologia foi utilizada para conduzir o desenvolvimento do projeto, o porquê e quais os respetivos benefícios e limitações que a mesma pode albergar na elaboração de um trabalho desta natureza.

2.1. Objetivo geral

Este estudo, como um projeto de ciência de dados centrado na análise e previsão de séries temporais, aplicado a uma organização e ao seu cenário, tem como objetivo principal desenvolver um modelo capaz de gerar previsões para os níveis de procura em contexto multiproduto. Tal como mencionado, trata-se de um projeto aplicado diretamente numa organização real – Sovena Consumer Goods Portugal – e desenvolvido em consonância com o departamento de Logística Interna da unidade fabril sediada no Barreiro, Setúbal. Simultaneamente, o projeto assume essencialmente uma abordagem quantitativa, centrada fundamentalmente na análise e modelação de dados para previsão da procura. Ainda assim, numa parte final, o estudo complementa esta mesma vertente com uma interpretação qualitativa dos resultados, visando identificar oportunidades de melhoria face ao método atualmente utilizado pela organização.

2.1.1. Objetivos específicos

Garantir um rigoroso nível de eficiência no processo de previsão da procura é cada vez mais importante nas empresas de carácter industrial com realidades altamente competitivas e sensíveis às diferentes variações de mercado, espera-se, portanto, com este projeto, desenhar um modelo capaz de possibilitar esse mesmo cenário. Para tal, e em sintonia com aquilo que é o objetivo geral, foram definidos vários objetivos de escala mais reduzida:

- Aplicar diferentes modelos de previsão na série temporal em estudo – histórico de vendas da organização.
- Comparar o desempenho entre os diferentes modelos aplicados.
- Propor uma abordagem híbrida para contrastar com a abordagem atual.
- Avaliar se a proposta de uma abordagem híbrida é capaz de melhorar a eficiência do processo de planeamento da procura pelo menos 15%.

Desta feita, a grande questão de investigação associada ao desenvolvimento deste projeto prende-se fundamentalmente no seguinte: será ou não possível melhorar significativamente um processo de previsão dos níveis de procura através da utilização de uma metodologia híbrida quando comparada a uma metodologia de modelo único?

2.2. Metodologia

Para a condução do projeto em si como um projeto no âmbito da ciência de dados, a metodologia escolhida foi CRISP-DM (Cross-Industry Standard Process for Data Mining), uma metodologia que viu a sua primeira versão ser introduzida em 1999, mais focada em projetos alinhados com mineração de dados, mas que apesar de tudo, continua a ser reconhecida como válida e importante no desenvolvimento de projetos em ciência de dados, fundamentalmente pelo seu tradicional formato de etapas sequenciais que guiam os seus aplicantes desde a forma inicial dos dados até ao conhecimento obtido através dos mesmos (Martinez-Plumed et al., 2021).

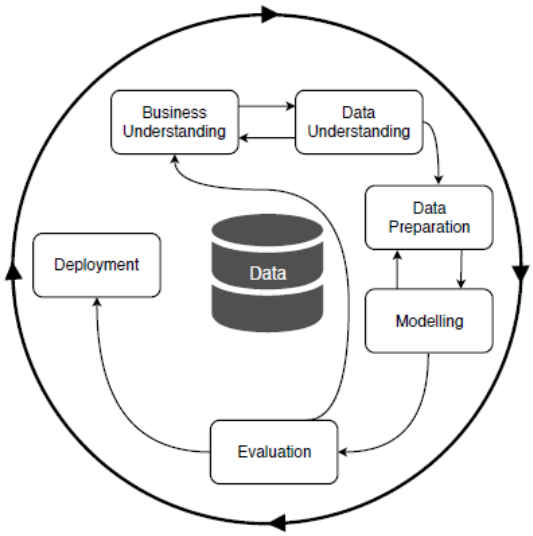

Figura 8 - Framework CRISP-DM

Fonte: Martinez-Plumed et al. (2021)

Esta sequência iterativa de etapas, mais concretamente seis, centram-se fundamentalmente na tradução dos dados em algo palpável para o objetivo comum. Esta metodologia inicia-se com a fase de compreensão de negócio, onde se identifica a problemática e o objetivo principal. As primeiras interações com os dados seguem-se nas duas etapas seguintes, primeiro a compreensão de dados, onde inicialmente se recolhem os dados e se efetua uma análise inicial aos mesmos e posteriormente a preparação de dados, onde se adaptam estruturas e dados propriamente ditos para deixar a base sobre a qual se vai trabalhar o mais limpa e pronta possível. As fases seguintes já estão relacionadas com a aplicação de modelos, onde surge o processo de modelação que dita o algoritmo a utilizar, seguida da fase de avaliação onde é executada uma análise ao desempenho do modelo através de determinadas métricas utilizadas para tal com o objetivo de escolher o melhor modelo a aplicar. Por fim, a etapa de implementação surge o final do ciclo, onde é operacionalizado o modelo escolhido em ambiente produtivo (Getachew et al., 2024).

2.2.1. Vantagens e limitações

No estudo conduzido por Aishah et al. (2018), esta metodologia apresenta diversas vantagens quando abordado o tema da sua aplicabilidade. Para os autores, o modelo não só é versátil, pois é possível aplicar em diferentes cenários e indústrias, como também oferece simultaneamente uma relação de equilíbrio entre a existência de uma estrutura sólida para auxiliar o desenvolvimento de projetos, e a flexibilidade na utilização de ferramentas de obtenção e tratamento de dados.

Ainda assim, na outra face da moeda, surgem algumas das limitações inerentes à metodologia em si e à sua aplicabilidade. Saltz (2021) afirma que apesar da sua popularidade inegável, o modelo CRISP-DM como apoio a projetos de ciência de dados falha em alguns aspetos ao longo dos respetivos ciclos. Apontamentos maioritariamente sobre detalhar a forma como os seus intervenientes devem priorizar tarefas, colaborar e comunicar entre si são alguns dos aspetos a melhorar, algo que para o autor poderá vir a ser colmatado através da combinação entre esta e as emergentes metodologias de coordenação de equipas em projetos de natureza similar como por exemplo, Scrum.

2.2.2. Recolha e tratamento de informação

As técnicas de recolha de informação foram bastante diretas e práticas já que os dados utilizados no desenvolvimento do projeto são dados estruturados que foram extraídos diretamente do ERP da organização, nomeadamente, SAP, na sua versão R/3. Desta forma, garante-se a autenticidade e qualidade da informação através de uma base de dados que contém informação relacionada com todas as vendas efetuadas entre Agosto de 2022 e Agosto de 2024.Já no que diz respeito às técnicas de tratamento desta mesma informação, o core do projeto foi desenvolvido recorrendo à linguagem de programação Python, na sua versão 3.10.9, suportada através da ferramenta Jupyter Notebook, onde foi redigido todo o código original para o desenvolvimento do projeto, incluindo claro, a geração dos seus outputs como valores, tabelas e gráficos.

Detalhando finalmente como foi delimitada a amostra a utilizar, ainda que a base de dados contenha informação sobre mais de 500 produtos, para facilitar o desenvolvimento de um projeto desta natureza, a amostra utilizada foi reduzida para dez produtos de acordo com um critério escolhido propositadamente para tal, neste caso, os dez produtos que acumulavam maior número de transações entre todos os presentes na base de dados original.

Capítulo 3 – Apresentação e discussão dos resultados

Tal como detalhado no capítulo anterior, o desenvolvimento do projeto foi conduzido de acordo com a metodologia CRISP-DM naquilo que será a sua adaptação ao espectro da ciência de dados. Assim sendo, numa primeira fase, e antes mesmo de passar a tudo o que esteja relacionado com dados e ao que os mesmos nos podem oferecer, importa compreender o negócio e como este está inserido no mercado e sociedade atual.

3.1. Compreensão do negócio

Este subcapítulo estará organizado em três fases distintas, onde iremos abordar a organização e posicionamento da empresa, os desafios e obstáculos principais que deram origem ao desenvolvimento deste projeto e por fim apresentar de forma sucinta de que maneira o processo de planeamento da procura decorre atualmente.

3.1.1. A organização e o seu posicionamento

O grupo Sovena, organização portuguesa e uma das empresas com maior relevância no setor agroindustrial, não só a nível nacional como também a nível internacional, teve a sua origem no grupo CUF – Companhia União Fabril. Atualmente, posiciona-se como um dos principais intervenientes no setor dos óleos alimentares, particularmente, óleo de girassol e azeite com as principais marcas de destaque como Fula e Oliveira da Serra respetivamente.

Descritos na Tabela 8, grupo apresenta fundamentalmente quatro segmentos de atuação ainda que a sua principal missão se centre no desenvolvimento de produtos alimentares de elevada qualidade, com grande enfoque naquilo que são as práticas sustentáveis de maior relevo em âmbito industrial, através de uma gestão eficiente de todos os seus recursos e intervenientes ao longo da cadeia de valor, cadeia esta que no caso do azeite está totalmente integrada, do olival ao lagar e do embalamento à distribuição.

Tabela 8 – Segmentos de atuação

Segmento	Descrição
Agricultura/sourcing	Campos agrícolas de cultivo (olival, girassol) próprios
Transformação	Transformação de óleos alimentares
Embalamento	Unidades fabris em diferentes países (PT, ES, IT, EUA)
Venda	Comercialização em dez países e mais de 70 mercados

Fonte: Sovena Group (2025)

Esta abordagem de integração vertical permite à organização assegurar os seus padrões de qualidade e garantir principalmente a rastreabilidade do produto e a eficiência e controlo operacional.

3.1.2. Definição da problemática

Tal como abordado em certa parte da revisão de literatura, o processo de previsão da procura constitui um pivot para o bom funcionamento de toda a cadeia de abastecimento, sobretudo em contextos em que os produtos a comercializar são de carácter alimentar/perecível. Nesse mesmo sentido, o desenvolvimento deste projeto visa redefinir e otimizar o processo utilizado atualmente para estimar os níveis de procura dos diferentes produtos da empresa. Esta necessidade nasce fundamentalmente pois o processo atual apresenta algumas limitações que colocam em causa a eficiência dos processos consequentes, tais como o planeamento de produção ao nível da unidade fabril e a própria gestão de inventário.

A decisão para a formulação da problemática recaiu sobre a framework SMART (*specific, measurable, achievable, relevant & time-bound*), pois acarreta vários benefícios como a permissão de definir claramente o objetivo do projeto em si e de que forma o mesmo é importante, permite também acompanhar e monitorizar o progresso do projeto de acordo com os objetivos inicialmente propostos e traz bastante clareza na comunicação para todos os envolvidos, uma vez que as bases iniciais estarão todas focadas e bem definidas desde a sua fase mais embrionária. Desta feita, a definição da problemática foi formulada da seguinte maneira:

- **Específica:** existe uma necessidade real de aprimorar o processo atual, que será saciada através do desenvolvimento de um modelo de ML focado na análise de séries temporais e capaz de gerar previsões para os níveis de procura de vários produtos;
- **Mensurável:** o desempenho do novo modelo a desenvolver será avaliado através de métricas quantitativas como, o erro absoluto médio (MAE) e a percentagem de erro médio absoluto (MAPE), permitindo com isto também comparações ao método atual;
- **Atingível:** através da análise, tratamento e modelação de dados históricos de vendas reais durante um período de dois anos que será estudado entre Agosto de 2022 e Agosto de 2024;
- **Relevante:** a otimização deste processo através de níveis de previsão mais fidedignos terá como consequência direta melhorias no planeamento de produção e consequentemente, gestão de inventário e cadeia de abastecimento de forma geral;
- **Temporal:** o desenvolvimento do projeto está estruturado para decorrer ao longo da duração do curso e estar concluído até ao final do primeiro semestre de 2025.

3.1.3. Processo atual

Na unidade fabril do Barreiro, onde o foco do desenvolvimento do projeto está centrado, o processo de previsão da procura assume duas abordagens diferentes de acordo com a

tipologia do produto. Para produtos em que a própria empresa é detentora da marca, a caracterização é de marca de fabricante (MDF) enquanto para produtos de marcas externas ao grupo a tipologia é denominada de marca de distribuição (MDD). Embora ambas as tipologias possam ser embaladas na mesma unidade fabril, as especificidades dos canais de venda e das relações contratuais determinam algumas diferenças na maneira como o processo de previsão é conduzido.

Para as MDF, o processo tem origem num orçamento geral executado anualmente que servirá de base para a estimativa da procura ao longo do respetivo ano. A partir dessa base anual é realizado um alisamento semanal que consiste na repartição da quantidade mensal de acordo com o número de semanas, tentando garantir dessa forma uma repartição consistente ao longo do mesmo período. Por fim, o processo fica completo através da introdução de inputs extra provenientes da área comercial, sejam elas acrescentos às insígnias dos diferentes canais de distribuição (HORECA, grossistas, retalhistas etc.), ou apenas ações promocionais programadas antecipadamente ou de última instância. Esta informação é compilada num ficheiro único antes de ser lançada para a respetiva unidade fabril, permitindo desta forma suportar os processos seguintes de planeamento de necessidades (a quatro semanas) e planeamento de produção (sequenciamento diário).

Figura 9 - Processo de previsão marca de fabricante

Fonte: Adaptação de manuais internos (Sovena Group, 2025)

Por outro lado, para as MDD, o processo inicia-se habitualmente com a criação de um contrato com o respetivo cliente. Este contrato será o vínculo que irá definir as componentes base tais como a duração e os volumes previstos que mais tarde alimentarão os níveis de previsão semanal. O cálculo desta, é realizado através da divisão do volume total contratado pela duração do contrato em semanas, assegurando desta forma, uma vez mais, uma projeção consistente ao longo do mesmo período.

Figura 10 - Processo de previsão marca de distribuição

Fonte: Adaptação de manuais internos (Sovena Group, 2025)

Apesar das diferenças existentes entre ambas as tipologias, ambos os processos culminam num mesmo ficheiro de objetivo comum, assegurar o envio da informação à unidade fabril para suportar a tomada de decisão naquilo que será a gestão dos processos de planeamento de necessidades e produção.

3.2. Compreensão dos dados

Antes de passar à compreensão dos dados propriamente dita, o primeiro ponto deste subcapítulo irá apresentar uma breve abordagem à linguagem de programação, ambiente e consequentes bibliotecas utilizadas. Posteriormente, iniciar-se-á a fase mais prática do projeto com a importação dos dados e o tratamento que antecede a análise exploratória e a fase de criação e aplicação dos modelos escolhidos.

3.2.1. Linguagem, ambiente e bibliotecas

O projeto, no seu carácter prático, foi todo ele desenvolvido com recurso à linguagem de programação Python (versão 3.10.9), não só pois foi uma das linguagens abordadas em diferentes unidades curriculares do mestrado, mas também pois é amplamente reconhecida na área da ciência de dados pela sua versatilidade e fundamentalmente pela sua vasta comunidade de utilizadores ativos, componente essa que continua a apresentar uma tendência crescente. A versão utilizada foi selecionada principalmente para assegurar a compatibilidade estável com todas as bibliotecas usadas no decorrer do projeto, e o ambiente escolhido foi o Jupyter Notebook, uma aplicação que pode ser utilizada no navegador e que permite a criação e partilha de documentos com diferentes blocos de código. Por fim, as bibliotecas utilizadas para suportar e conduzir todas as operações de importação, tratamento, análise e modelação foram as abaixo:

Bloco 1 - Importação de bibliotecas

```
# Import libraries
import pandas as pd
import numpy as np
import plotly.graph_objects as go
import re
import os
import scipy.stats as stats
import plotly.express as px
import xgboost as xgb
import warnings
import logging
from statsmodels.tsa.stattools import adfuller
from pmdarima import auto_arima
from prophet import Prophet
from pandas.api.types import is_string_dtype
from pandas.api.types import is_numeric_dtype
from sklearn.metrics import mean_absolute_error, mean_absolute_percentage_error
from statsmodels.tools.sm_exceptions import ConvergenceWarning
from sklearn.preprocessing import MinMaxScaler
```

3.2.2. Importação e estrutura inicial dos dados

O primeiro passo a dar foi o de importação do ficheiro Excel (.xlsx) extraído diretamente do SI da organização. Relativamente às suas características iniciais, os dados em estudo são estruturados, apresentam-se em formato tabular, ou seja, com linhas (observações) e colunas (variáveis) e contêm dados de vendas reais de múltiplos produtos entre Agosto de 2022 e Agosto de 2024, comportando, portanto, a nível temporal, um intervalo completo de dois anos. Esta primeira importação de dados foi executada através da função $read\_excel()$ que nos permite criar uma tabela com a respetiva informação, e o output de uma segunda função $shape$ gera uma tupla que identifica o número de observações e variáveis existentes:

- **Número de observações:** 123 681
- **Número de variáveis:** 12

Através da função $dtypes$, conhecemos os tipos de dados que se irão trabalhar, no entanto, para melhor entendimento dos mesmos, convém introduzir teoricamente a sua natureza:

- **Float64:** valores numéricos reais (decimais de precisão dupla)
- **Int64:** valores numéricos reais (inteiros)
- **Object:** dados heterogéneos (habitualmente texto/string)
- **Datetime64[ns]:** dados temporais (data/hora de precisão a nanossegundos)

Tabela 9 - Tipologia inicial de dados

Variável	Carácter prático	Tipologia
Referência	Identificador de guia de remessa	float64
Material	Identificador de produto	float64
Texto breve material	Descrição de produto	object
Qtd. UM registo	Quantidade em unidade de medida	int64
UM registo	Unidade de medida	object
Lote	Lote de produção	object
Doc.material	Identificador de movimento	float64
Data	Data de movimento	datetime64[ns]
Cliente	Identificador de cliente	float64
Quantidade	Quantidade em litros	float64
Texto cabeçalho documento	Notas a acrescentar	float64
Montante em MI	Valorização de venda em euros	float64

Posto isto, a Tabela 9 apresenta o nome de cada variável, a que se refere cada uma delas a nível prático e a respetiva tipologia de dados de acordo com a perspetiva inicial de importação crua dos mesmos.

3.3. Preparação dos dados

Esta secção dedica-se à preparação dos dados, fase fundamental para garantir a qualidade dos mesmos e a fiabilidade das análises e tarefas seguintes. De forma geral, será a área dedicada ao tratamento de valores nulos/duplicados e à transformação de dados.

3.3.1. Identificação e tratamento de valores nulos

Na etapa de preparação dos dados, o passo inicial foi realizar o tratamento de valores nulos que pudessem existir em cada uma das colunas da base de dados. Para tal, aplicou-se uma função que nos permite somar os valores nulos por variável, permitindo assim identificar quais os campos mais afetados e em que dimensão.

Bloco 2 - Verificação e remoção de valores nulos

```
# Checking presence of null values
with pd.option_context('display.max_rows', None):
    print(df_sales.apply(lambda x: sum(x.isnull()), axis = 0))

# Drop null column and values
df_sales = df_sales.drop('Texto cabeçalho documento', axis = 1)
df_sales = df_sales.dropna()
```

A Tabela 10 demonstra a frequência de valores nulos para cada variável constituinte da base de dados a analisar.

Tabela 10 - Valores nulos

Variável	Quantidade	Variável	Quantidade
Referência	4	Doc.material	4
Material	4	Data	4
Texto breve material	4	Cliente	4
Qtd. UM registo	0	Quantidade	4
UM registo	0	Texto cabeçalho documento	123 681
Lote	4	Montante em MI	4

Desta feita, foi possível constatar que a variável "Texto cabeçalho documento" apresentava valores nulos em todas as observações, e por esse mesmo motivo, foi eliminada da base de dados. Adicionalmente a esta ação, foram eliminadas também todas as linhas que continham valores em falta em qualquer uma das restantes onze variáveis, assegurando desta forma a integridade dos dados para as análises seguintes.

3.3.2. Deteção e tratamento de valores duplicados

A próxima etapa na preparação dos dados passa por detetar possíveis valores duplicados que possam ferir a integridade da base de dados. Para conduzir este passo de forma mais fidedigna possível sem afetar a qualidade dos dados foi necessário transformar antecipadamente todos os valores negativos a positivos nas variáveis "Qtd. UM registo", "Quantidade" e "Montante em MI" pois, em caso de estorno de determinado movimento no SI, o registo surgiria duplicado com valores a positivo, a negativo e posteriormente positivo novamente. Para tal, foi utilizado um $for\ loop$ com a função $abs()$ para que convertesse todos os valores negativos em positivos nas respetivas três colunas mencionadas anteriormente.

Bloco 3 - Conversão de negativos e remoção de duplicados

```
# Convert negative values
neg_val = ['Qtd.  UM registo', 'Quantidade', 'Montante em MI']

for c in neg_val:
    df_sales[c] = df_sales [c].abs()

# Drop duplicates
df_sales.drop_duplicates()
```

De seguida, e para finalizar este passo, foi utilizada a função $drop\_duplicates()$ para remover todas as observações duplicadas da base de dados, o que levou a uma alteração na dimensão para novos valores:

- **Número de observações:** 123 614
- **Número de variáveis:** 11

3.3.3. Transformação de dados

Finalmente, e mesmo antes de realizar qualquer parte da análise exploratória (AE) que se possa querer incluir, foram realizadas algumas ações ao nível da transformação e conversão de dados para preparar não só essa AE, mas também deixar a base de dados totalmente preparada para aquilo que será a aplicação dos modelos escolhidos para este caso.

3.3.3.1. Conversão de tipos de dados

Recuando um pouco para analisar novamente os dados presentes na Tabela 9, podemos verificar que determinadas colunas apresentavam tipos de dados inconsistentes com o seu conteúdo e finalidade.

Bloco 4 - Conversão de tipos de dados

```
# Convert data types
int_var = ['Qtd.  UM registo', 'Cliente', 'Quantidade']
str_var = ['Referência', 'Material', 'Doc.material']

for c in df_sales:
    if c in int_var:
        df_sales[c] = df_sales[c].astype(int)
    if c in str_var:
        df_sales[c] = df_sales[c].astype(str).str[:-2]
```

Para garantir a coerência e evitar erros em possíveis operações futuras a executar, o Bloco 4 detalha a conversão das colunas "Qtd. UM registo", "Cliente" e "Quantidade" para números inteiros e das colunas "Referência", "Material" e "Doc.material" para texto.

3.3.3.2. Pré-processamento e criação de variáveis

Depois de executados os processos de limpeza e tratamento inicial dos dados nos pontos acima descritos, procedeu-se a uma fase um pouco mais avançada de pré-processamento dos dados. Esta fase visa fundamentalmente preparar e estruturar os dados de forma a possibilitar numa fase inicial a condução de uma análise exploratória e numa fase posterior, de modelação e aplicação dos modelos. O primeiro passo desta etapa foi eliminar variáveis sem relevância para o pretendido, ou seja, colunas consideradas não essenciais para aquilo que seria a análise exploratória e a aplicação dos modelos.

Bloco 5 - Remoção de colunas dispensáveis

```
# Drop unecessary columns
nec_col = ['Material', 'Texto breve material', 'Data', 'Quantidade']

for c in df_sales:
    if c not in nec_col:
        df_sales = df_sales.drop(c, axis = 1)
```

Posto isto, das onze colunas existentes, removeram-se todas com exceção das variáveis "Material", "Texto breve material", "Data" e "Quantidade", deixando assim a base de dados com uma estrutura de 123 614 linhas e 4 colunas. As variáveis de calendário foram geradas através da variável "Data" e deram origem a novos períodos temporais possibilitando obter uma perspetiva inicial sobre possíveis tendências e padrões que possam estar representados nos dados:

Tabela 11 - Criação de variáveis temporais

Variável	Função utilizada
Ano-Semana	$dt.year.astype(str) + dt.isocalendar().week.astype(str)$
Ano-Mês	$dt.to\_period('M').astype(str)$
Ano-Trimestre	$dt.to\_period('Q').astype(str)$
Ano-Semestre	$dt.year.astype(str) + (dt.month - 1)//6 + 1).astype(str)$

Uma vez concluída a criação das variáveis de calendário, foram criadas duas colunas extra de caracterização de produto – "Categoria" e "Formato". Ambas foram geradas através da variável "Texto breve de material" que corresponde à descrição do produto. Para a primeira, retiveram-se os primeiros dois caracteres desta variável de texto correspondendo assim à categoria pretendida, para segunda a extração ocorreu através da execução de uma função que utiliza expressões regulares para captar as diferentes terminações dos formatos.

Bloco 6 - Criação de variáveis: categoria e formato

```
# Create product category
df_sales['Categoria'] = df_sales['Texto breve material'].str[:2]

# Create product format
def get_format(mat_desc):
    match = re.search(r'(\d+L|\d+ML)', mat_desc.upper())
    return match.group(1) if match else 'Unknown'

df_sales['Formato'] = df_sales['Texto breve material'].apply(get_format)
```

Após a criação desta coluna, restringiu-se o conjunto de dados apenas às categorias pretendidas para a condução e desenvolvimento do projeto: "AZ" e "OL" para produtos de azeite ou óleo respetivamente. Esta análise foi feita através da função $value\_counts()$ que permite efetuar uma verificação de quantas categorias foram geradas através do passo executado anteriormente, e desta forma, possibilitar a identificação dessas categorias potencialmente não pretendidas para a análise.

Tabela 12 - Categoria de produto e frequência

Categoria	Frequência
OL	63 757
AZ	56 209
>A	2 773
>O	530
(O	378
OB	24
>	6

A remoção das outras categorias ocorre fundamentalmente pois os produtos marcados com o símbolo “>” e “OBS” são considerados pré-obsoletos e obsoletos, respetivamente, e optou-se por não analisar esse tipo de produtos durante este estudo. Os valores para a categoria “Formato”, criada com o objetivo principal de categorizar os produtos de acordo com o seu formato ou capacidade, estão descritos na Tabela 13. É possível observar a existência de vários tipos de formatos, com frequências variadas ao longo da base de dados, inclusive a presença de 281 artigos cujo formato não foi possível de extrair através da função escrita para tal.

Tabela 13 - Formatos de produto e frequência

Formato	Frequência
1L	35 188
750ML	21 105
3L	19 649
10L	13 682
500ML	12 225
5L	6 211
250ML	5 816
200ML	3 613
2L	1 871
Unknown	281
7ML	164
75L	100
1000L	17
1000ML	15
20ML	14
12ML	8
75ML	7

Desta feita, optou-se por manter apenas os formatos com um mínimo de 300 observações numa tentativa, não só, de assegurar uma representatividade estatística mínima das classes geradas, mas também de manter a coerência para aquilo que seriam os produtos embalados na unidade fabril em questão, uma vez que a maioria dos formatos abaixo do limite de

ocorrências definido, não é embalado na unidade fabril sediada no Barreiro e por esse mesmo motivo, não deverão ser considerados.

Tabela 14 - Estrutura final da base de dados

Variável	Carácter prático	Tipologia
Data	Data de movimento em SI	datetime64[ns]
Ano-Semana	Variável semanal	object
Ano-Mês	Variável mensal	object
Ano-Trimestre	Variável trimestral	object
Ano-Semestre	Variável semestral	object
Material	Identificador de produto	float64
Texto breve material	Descrição de produto	object
Categoria	Categoria de produto	object
Formato	Formato de produto	object
Quantidade	Quantidade em litros	float64

Por fim, a base de dados foi reorganizada, modificando alguns dos índices através da aplicação da função $loc[]$, deixando desta feita, a estrutura final com 119 360 observações e 10 variáveis e ordenada de acordo com o exposto na Tabela 14.

3.3.4. Análise exploratória

Já em posse de uma base de dados preparada, é tempo de dar lugar a um novo processo, a análise exploratória de dados. Esta fase constitui uma etapa preponderante no desenvolvimento do projeto e visa fundamentalmente a compreensão de forma gráfica e visual das principais características dos dados, identificando desta feita tendências, padrões e até possíveis anomalias que possam influenciar as análises seguintes como a presença de outliers. Para nota, a criação das visualizações necessárias a esta fase foi toda ela conduzida depois da definição da biblioteca Plotly como *backend* preferencial.

3.3.4.1. Artigos mais transacionados

A primeira análise a executar neste ponto teve como intuito entender como os produtos se comportam ao longo do período em estudo, mais concretamente, identificar quais eram os produtos mais transacionados durante o mesmo, isto é, com o maior número de vendas ao longos dos dois anos em análise. Para tal, recorreu-se ao novamente método $value\_counts()$ para calcular a frequência de cada artigo e consequentemente à função $nlargest(10)$ para limitar esse output apenas a 10 resultados, obtendo desta forma um gráfico capaz de identificar os dez artigos mais transacionados ao longo desta série temporal.

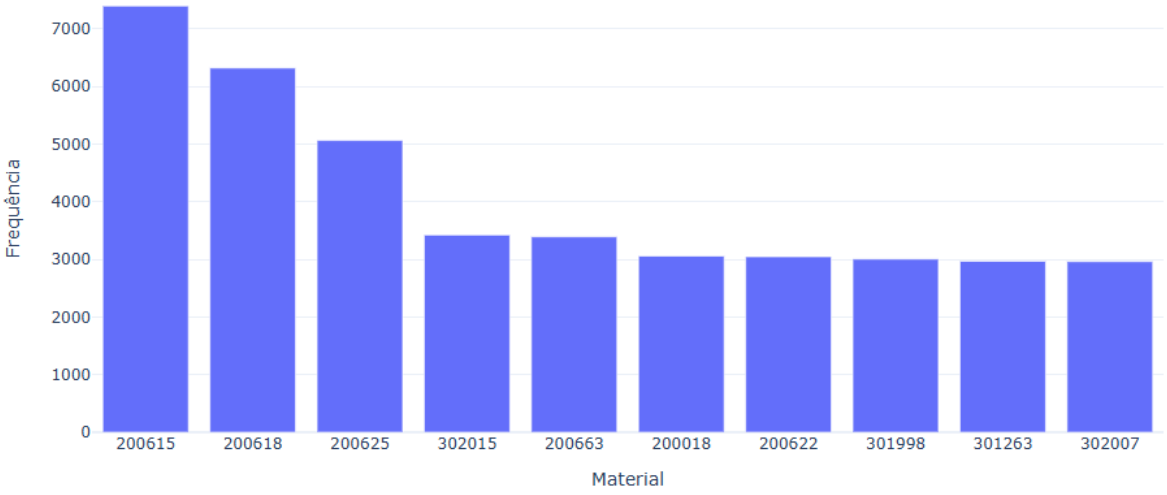

Gráfico 2 - Produtos mais transacionados

O resultado é representado através do Gráfico 2 – identificação de produto no eixo das abcissas e número total de transações no eixo das ordenadas – onde facilmente se destacam os três primeiros produtos (200615, 200618 e 200625), todos eles acima das 5 000 transações enquanto os restantes apresentam uma frequência relativamente equilibrada entre todos.

3.3.4.2. Artigos com maior volume de vendas

Depois de analisar a frequência de transações para cada produto, realizou-se uma segunda abordagem, ainda que relativamente no mesmo sentido, onde se tentaram identificar os produtos com maior volume de vendas. Uma vez mais, utilizou-se um gráfico de barras (Gráfico 3 – identificação de produto no eixo das abcissas e quantidade vendida em litros no eixo das ordenadas) compilando a função $groupby()$ associada à variável "Material" e a função $sum()$ associada à variável "Quantidade", para que o resultado devolvido originasse a soma da quantidade vendida para cada produto durante o intervalo de tempo em estudo.

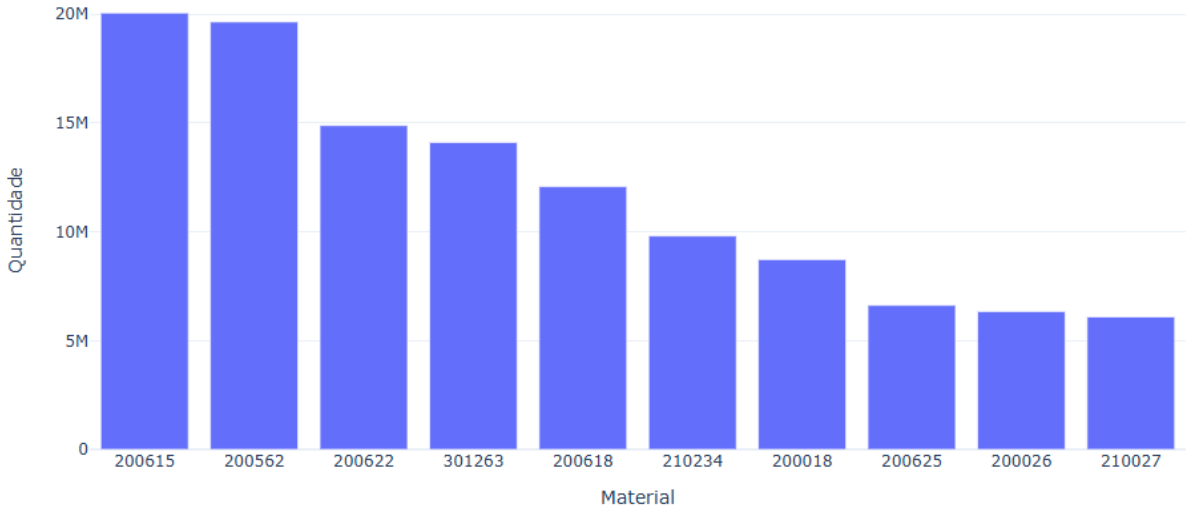

Gráfico 3 - Produtos com maior volume de vendas

O resultado apresenta novamente na primeira posição como artigo com maior volume de vendas a referência 200615 praticamente em paralelo com um outro artigo com o código 200562, ambos com cerca de vinte milhões de litros vendidos. A partir dessa segunda posição os produtos aproxima-se entre si ainda que com uma redução gradual até ao último artigo.

3.3.4.3. Tendência

De modo a possibilitar a análise da evolução do volume de vendas ao longo do período em estudo, procedeu-se à adaptação da série temporal para um intervalo mensal. Isto foi possível devido ao trabalho previamente feito onde foram geradas novas variáveis de calendário.

Tendência: geral

No estudo da tendência, optou-se por subdividir as análises em três fases distintas, esta primeira a destacar o comportamento geral do volume de vendas ao longo do tempo. Para tal, aplicou-se novamente a função $groupby()$, desta feita à variável "Ano-Mês", em conjunto com $sum()$ da variável "Quantidade", para agrupar as quantidades totais vendidas em litros a cada um dos meses presentes na série. A partir desta operação, é possível gerar um gráfico de linhas (Gráfico 4 – identificação da data no formato mês-ano no eixo das abcissas e quantidade vendida em litros no eixo das ordenadas) que demonstra a evolução mensal do volume total de vendas ao longo destes dois anos. Complementarmente, para facilitar aquilo que possa ser a compreensão da linha de vendas real, adicionou-se ao gráfico uma linha de tendência linear (tracejado vermelho). Esta adição foi possível através da função $np.polyfit()$ que assume como argumentos as datas existentes e a quantidade vendida, e que permite, através de uma estimativa dos coeficientes, evidenciar a natureza da tendência.

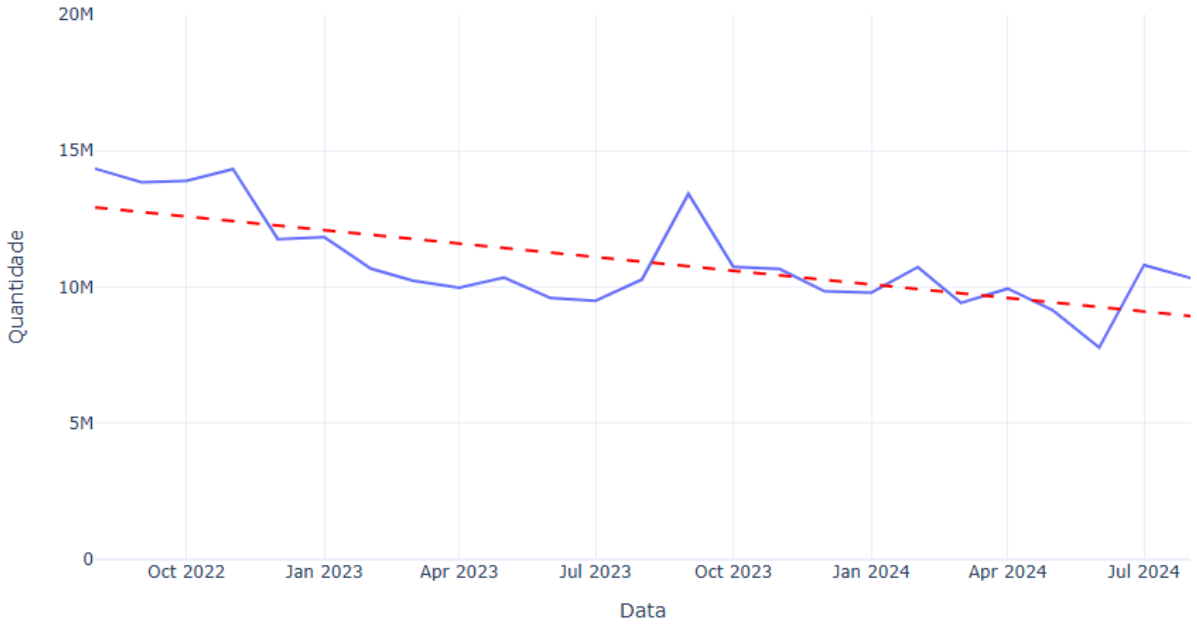

Gráfico 4 - Tendência global

Como é possível observar pela visualização acima, a tendência é ligeiramente decrescente, algo possível de confirmar através da presença de uma inclinação negativa da linha de tendência gerada no passo anterior, algo que pode evidenciar uma redução da procura ao longo do tempo.

Tendência: categoria

A criação de variáveis num dos pontos anteriores permite também nesta fase analisar aquilo que é a tendência do volume de cada categoria. Esta análise tem como base essa mesma variável previamente gerada através das duas primeiras letras da descrição do produto (variável "Texto breve material") e que distingue fundamentalmente aquilo que são azeites de óleos. Aqui, recorreu-se novamente ao agrupamento dos dados, utilizando a função $groupby()$, de acordo com a variável temporal ("Ano-Mês") e adicionalmente, pela variável categoria. Esta nova estrutura de dados foi resultado da aplicação de uma derradeira função $pivot()$ que permitiu converter a categoria de cada produto em duas colunas diferentes, uma para os produtos categorizados por "OL" e outra para "AZ".

No Gráfico 5, podemos descortinar que, apesar de demonstrar claramente uma predominância da categoria "OL" (linha vermelha) no que diz respeito ao volume transacionado, ditando ao longo de todo o período a transação de praticamente o dobro da quantidade transacionada quando comparada à categoria rival ("OL" a azul), esta apresenta uma ligeira tendência decrescente ao longo dos dois anos em análise. As oscilações, mais acentuadas para os óleos, podem indicar a presença de padrões de consumo e sazonalidade em alguns casos, e ainda que estejam também presentes alguns picos e quebras na tendência dos azeites, essa série em particular apresenta de algum modo um comportamento relativamente mais constante ao longo do mesmo período.

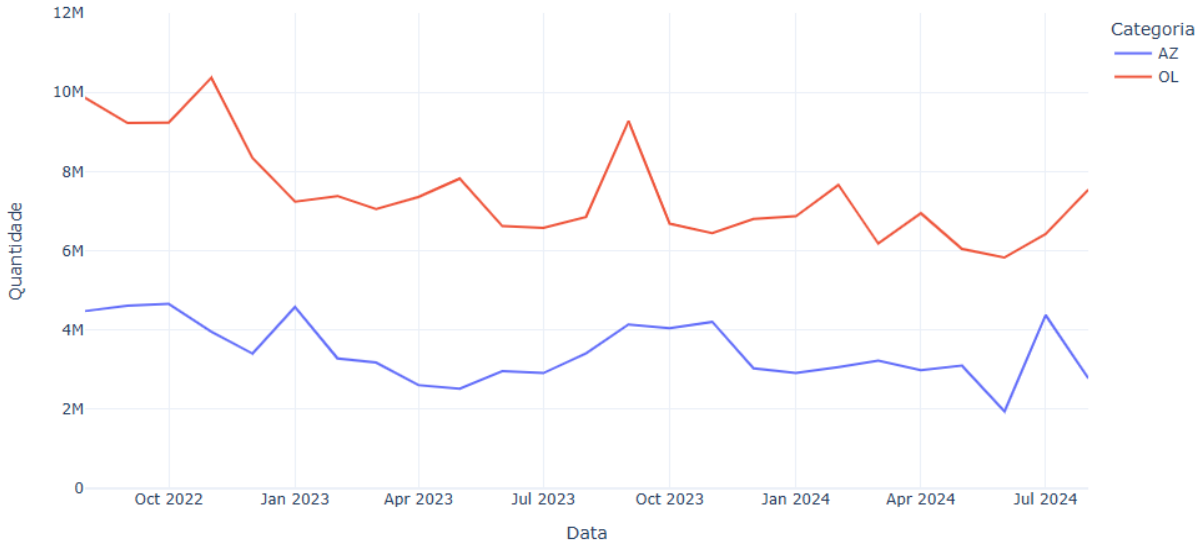

Gráfico 5 - Tendência por categoria

Tendência: formato

Tal como no ponto anterior, onde foi analisada a tendência de acordo com uma das variáveis anteriormente criadas, este último visa fundamentalmente executar o mesmo tipo de análise, alterando apenas a variável complementar em estudo da categoria para o formato. Esta variável, também ela gerada a partir da descrição de produto e criada através da utilização de expressões regulares, permite distinguir qual o formato do produto no que diz respeito ao seu volume (1L, 2L, 5L etc.). O procedimento de criação do Gráfico 6 foi exatamente o mesmo que o utilizado para a análise da tendência por categoria, desta feita, é possível identificar, para além dos comportamentos de tendência para cada um dos formatos presentes na base de dados, identificar também quais aqueles que representam maior volume transacionado durante o intervalo de tempo em estudo.

Analisando o conteúdo do respetivo gráfico, é possível perceber que o formato de 1 litro domina significativamente no que diz respeito ao volume transacionado ainda que com uma tendência ligeiramente decrescente. Esta realidade está diretamente relacionada com a análise anterior, isto é, a par do que tinha sido percecionado anteriormente na análise por categoria este formato é na esmagadora maioria, um formato presente em óleos. Os restantes formatos aparentam estar todos eles bem próximos uns dos outros e com linhas de tendência relativamente constantes apenas com superioridade para os formatos de 3L e 500ML, este último com maior predominância no mercado de azeites.

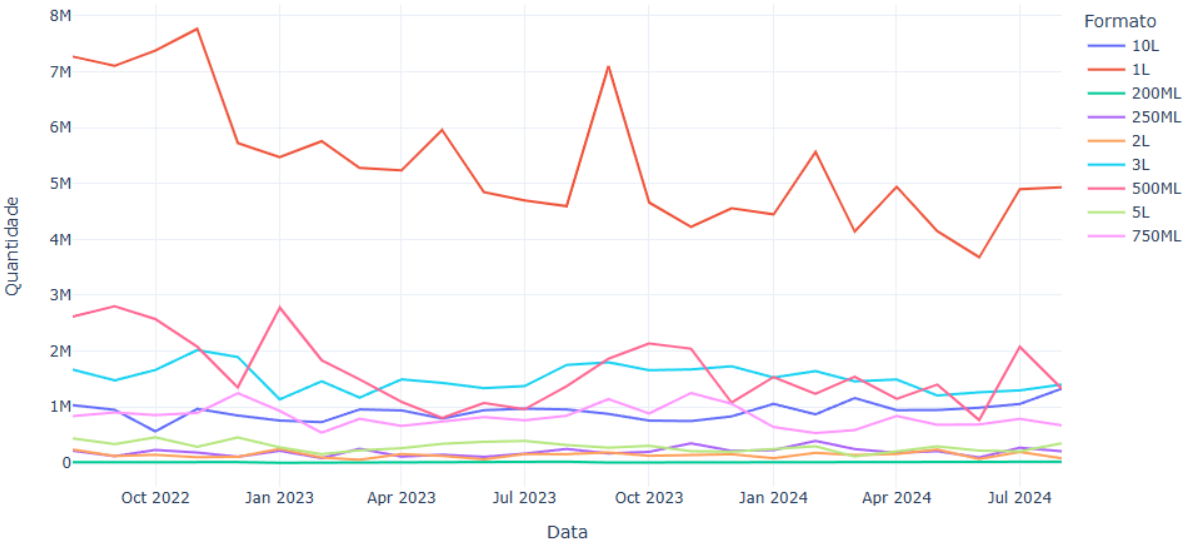

Gráfico 6 - Tendência por formato

3.3.4.4. Distribuição

O ponto seguinte visa fundamentalmente compreender a distribuição da variável "Quantidade", ou seja, qual o carácter das diferentes transações em termos de volume transacionado em cada uma delas.

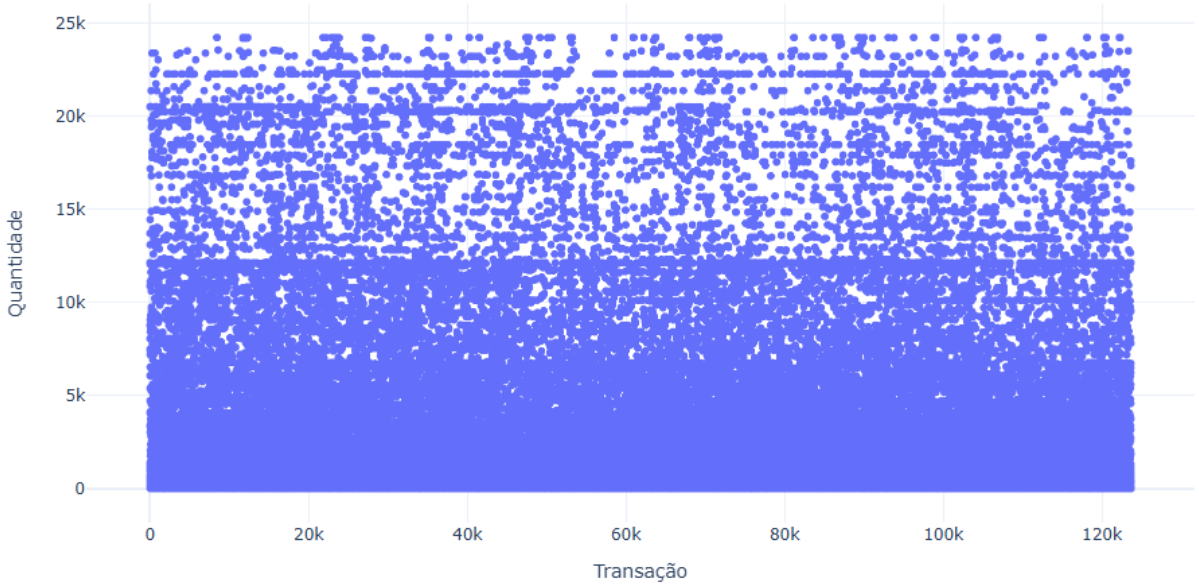

Gráfico 7 - Distribuição da quantidade

O Gráfico 7 surge nesse mesmo sentido, identificar a dispersão das diferentes quantidades vendidas presentes em cada transação da base de dados. Assim sendo, é possível observar uma elevada concentração nas quantidades mais reduzidas, e essa mesma concentração vai sendo reduzida à medida que as quantidades se tornam maiores. Apesar da concentração ser significativamente mais reduzida nas quantidades maiores, é possível também identificar um teto a rondar os 25 mil litros, retirando desta feita a possibilidade da existência de outliers na base de dados e como consequência, eliminando a necessidade de executar qualquer tipo de tratamento ou processamento de dados nesse mesmo sentido.

3.4. Modelação

Concluída a fase de preparação dos dados, etapa onde foi possível obter um entendimento inicial sobre o comportamento dos mesmos e proceder ao tratamento necessário para a preparação das próximas fases, a etapa seguinte centra-se na modelação das séries temporais. O objetivo principal desta fase é garantir a criação de modelos que garantam o lançamento de níveis de procura que se aproximem mais da realidade observada e nesse sentido, para além do modelo atualmente em utilização pela empresa, foram considerados três modelos: ARIMA, Facebook Prophet e XGBoost. Uma vez que a base de dados é de natureza multiproduto, ou seja, contém observações de vendas para vários artigos, os modelos foram aplicados individualmente a cada um dos artigos selecionados para análise. A seleção deste leque de produtos baseia-se na utilização dos dez artigos com o maior número de transações ao longo do período em estudo, removendo desta seleção as seguintes referências: 301263, por ser um produto *make-to-order*, 302015, por ser um artigo produzido fora da unidade do Barreiro e 200663, por ser um produto com dados de vendas incompletos.

3.4.1. Funções de apoio

Para conduzir o desenvolvimento da fase de modelação, foram escritas algumas funções de apoio que permitem transformar a base de dados de acordo com os objetivos pretendidos em cada modelo.

3.4.1.1. Adaptação da base de dados

A primeira função a ser desenvolvida visa alcançar uma adaptação inicial da base de dados às características necessárias. Esta função atua como uma primeira filtragem que garante que cada série contém apenas as variáveis e observações necessárias e recebe como argumentos a base de dados final originada pelos processos de pré-processamento e tratamento de dados, e o número identificador do material (ex. 200615). Desta feita, apresenta na sua estrutura três passos sequenciados:

- Seleção das colunas relevantes (data, material e quantidade);
- Agregação das quantidades por data e material
- Filtragem da base de dados de acordo com o produto selecionado

Bloco 7 - Adaptação da base de dados

```
def adapt_df(df, material):
    # Duplicate DataFrame
    df = df[['Data, 'Material', 'Quantidade']].copy()
    # Groupby 'Data' & 'Material'
    df = df.groupby(['Data', 'Material'])['Quantidade'].sum().reset_index()
    # Filter DataFrame by 'Material'
    df = df.loc[df['Material'] == material]
    return df
```

3.4.1.2. Verificação da componente sazonal

A próxima função é possível de implementar na sequência da apresentada anteriormente e tem como principal objetivo avaliar a presença de sazonalidade na série em estudo, ou seja, a análise é executada através da identificação, ou não, de padrões sistemáticos ou recorrentes ao longo dos doze meses do ano. Para tal, utiliza-se o teste não paramétrico de Kruskal-Wallis e o mesmo é conduzido com um nível de significância de cinco pontos percentuais ($\alpha = 0{,}05$). Caso o valor de p seja inferior ao nível de significância, rejeita-se a hipótese nula, que neste caso dirá que as distribuições da variável "Quantidade" são iguais para todos os meses, ou por outras palavras, que não existe sazonalidade, caso contrário, existirão diferenças estatisticamente significativas entre os meses do ano ditando assim a presença de sazonalidade. Esta função será particularmente útil nos passos seguintes pois os modelos ARIMA e Facebook Prophet contém argumentos para a componente sazonal e comportam-se de forma distinta de acordo com o valor do mesmo.

Assim sendo, a função de verificação da sazonalidade apresenta a seguinte estrutura:

- Transformação da coluna da data no formato "Ano-Mês"
- Extrai o valor numérico do respetivo mês
- Agrupa os valores da variável “Quantidade” por mês
- Aplica o teste de Kruskal-Wallis (nível de significância 5%)
- Retorna um booleano de valor $True$ caso exista sazonalidade

Bloco 8 - Verificação de sazonalidade

```
def check_seasonality(df):
    # Generate 'Ano-Mês' column to check seasonality
    df['Ano-Mês'] = df['Data'].dt.to_period('M').astype(str)
    # Ensure string format for column 'Ano-Mês'
    df['Ano-Mês'] = pd.to_datetime(df['Ano-Mês'], format = '%Y-%m')
    # Extracting month value from column 'Ano-Mês'
    df['Mês'] = df['Ano-Mês'].dt.month
    # Groupby 'Quantidade' for each month
    months = [df[df['Mês'] == month]['Quantidade'] for month in range(1, 13)]
    # Kruskal-Wallis test
    test_stat, p_value = stats.kruskal(*months)
    # Set significance level
    alpha = 0.05
    return p_value < alpha
```

3.4.1.3. Verificação da estacionariedade

A função seguinte permite determinar se a série temporal associada à variável “Quantidade” apresenta ou não comportamento estacionário. Esta avaliação é fundamental uma vez que constitui um dos pressupostos para a aplicação do modelo ARIMA a aplicar posteriormente. Para tal, primeiramente será necessário agregar os valores da variável a uma frequência semanal, pois as nossas previsões serão dessa mesma natureza. Posteriormente, e para efetivamente verificar este pressuposto, recorre-se ao teste de Dickey-Fuller Aumentado (ADF), e tal como na verificação da sazonalidade, este também é conduzido seguindo um nível de significância de 5% ($\alpha = 0{,}05$). Como neste caso, a hipótese nula diz que a série temporal possui uma raiz unitária, ou seja, que a mesma não é estacionária, caso o valor de p seja inferior ao nosso nível de significância, rejeita-se essa mesma hipótese. Desta forma, a função de verificação da estacionariedade apresenta a seguinte estrutura:

- Agregação dos dados da variável “Quantidade” em intervalo semanal
- Aplicação do teste ADF à respetiva base de dados/variável (nível de significância 5%)
- Retorna um booleano de valor $True$ caso a série seja estacionária

Uma vez mais, esta função será particularmente útil para alimentar aquilo que será um dos parâmetros presentes no modelo ARIMA que indica qual o grau de diferenciação necessário para atingir a estacionariedade, caso esta não se verifique.

Bloco 9 - Verificação de estacionariedade

```
def check_stationarity(df):
    # Weekly aggregation
    df = df.groupby(pd.Grouper(key = 'Data', freq = 'W'))['Quantidade'].sum()
    df.reset_index()
    # Check p-value
    test_stat = adfuller(df['Quantidade'].dropna())
    p_value = test_stat[1]
    # Set significance level
    alpha = 0.05
    return pValue < alpha
```

3.4.1.4. Preparação da série temporal

Esta, é a função que antecede a aplicação do modelo e, de forma geral, visa ultimar a preparação da série temporal para que essa aplicação seja possível de acordo com os pressupostos exigidos para o cálculo dos níveis de procura. Numa primeira fase são eliminadas variáveis não necessárias à aplicação do respetivo modelo, antes de reindexar a base de dados através da coluna da data. Por fim, todas as datas (frequência diária) são reindexadas à variável "Data" para adicionar datas possivelmente inexistentes na base de dados original e atribuir a estas o valor zero para a variável "Quantidade", isto pois inicialmente não teriam valor. Para sintetizar a estrutura da função, apresentam-se os pontos abaixo:

- Eliminar variáveis desnecessárias
- Selecionar coluna "Data" como índice
- Reindexar todas as datas com frequência diária
- Corrigir possíveis datas com valores nulos (atribuição de valor zero)

3.4.2. Modelo ARIMA

Uma vez desenvolvidas as funções de apoio necessárias, o passo seguinte será aplicar os modelos selecionados, numa primeira fase, o modelo ARIMA. Este processo é então delineado através da execução sequencial das etapas abaixo (com o apoio das funções desenvolvidas e apresentadas nos pontos anteriores), repetidas em ciclo $(for\ loop)$ para os sete produtos selecionados previamente:

Adaptação da série: através da função de adaptação, para cada produto é extraído um subset da base de dados original contendo apenas observações do respetivo material.

Verificação de componentes: verificação da componente sazonal, conduzida pela aplicação do teste não paramétrico de Kruskal-Wallis, e da estacionariedade, conduzida pela aplicação do teste ADF. Ambas geram um booleano de acordo com o output do mesmo.

Preparação da série: a preparação da série garante a divisão entre o subset de treino e de teste para possibilitar mais tarde a comparação dos dados e consequente avaliação – utilizada a regra 80/20%, o equivalente aos últimos 154 dias da base de dados para o set de teste.

Modelação: é conduzida através da função $auto\_arima$ e do critério de informação de Akaike (AIC). O critério calcula automaticamente os melhores valores para os respetivos parâmetros e a função recebe como argumentos os outputs da verificação de sazonalidade e estacionariedade para indicar que valores utilizar nesses mesmos parâmetros da função.

- Sazonalidade: parâmetro assume valor 12 caso se verifique sazonalidade e 0 se o contrário.
- Estacionariedade: parâmetro assume valor 0 caso se verifique estacionariedade e $None$ se o contrário (o último valor garante cálculo automático para d – componente de diferenciação)

Previsão e avaliação: as previsões são geradas para os últimos 154 dias, agrupadas em formato semanal para finalmente serem comparadas contra as observações reais, tudo compilado num subset análogo que permite a posterior avaliação do modelo de acordo com as métricas escolhidas (MAE e MAPE).

Bloco 10 - Modelação ARIMA

```
def auto_arima_model(train_set, season, station):
    # Set seasonal according to seasonality test
    season_period = 12 if season else 0
    # Set d according to stationarity test
    d_param = 0 if station else None
    # Use auto_arima and AIC to find the best parameters
    model = auto_arima(train_set,
                       start_p = 0, start_q = 0, max_p = 5, max_q = 5,
                       d = d_param,
                       start_P = 0, start_Q = 0, max_P = 2, max_Q = 2,
                       D = 1 if season else 0,
                       Seasonal = season, m = season_period,
                       stepwise = True,
                       suppress_warnings = True,
                       error_action = 'ignore',
                       trace = True,
                       information_criterion = 'aic',
                       n_fits = 50)
    # Fit the model
    result = model.fit(train_set)
    # Forecast the next 154 days
    forecast = result.predict(n_periods = 154).astype(int)
    return forecast
```

3.4.3. Modelo Facebook Prophet

Uma vez aplicado o modelo ARIMA, o próximo na lista é o Facebook Prophet, um modelo também de carácter estatístico. Uma vez mais, aplicação deste modelo é também executada de forma individual e a lista de produtos é exatamente a mesma que a utilizada na aplicação do modelo anterior. A implementação deste modelo seguiu a sequência de etapas abaixo apresentadas:

Adaptação da série: através da função de adaptação, para cada produto é extraído um subset da base de dados original contendo apenas observações do respetivo material. Uma adaptação extra é necessária, as variáveis têm de ser renomeadas pada “ds” (data) e “y” (quantidade) respetivamente.

Verificação de sazonalidade: a verificação de sazonalidade, conduzida pela aplicação do teste de Kruskal-Wallis, gera um booleano de acordo com o output do mesmo.

Divisão de sets: da mesma forma que no modelo ARIMA, para garantir a comparabilidade dos modelos, o set de treino corresponde a todas as datas observadas até 154 dias ates da última, enquanto o set de teste corresponde a esses mesmos 154 dias finais.

Modelação: a instância do modelo é criada com dois argumentos, um ajustando o nível do intervalo de previsão e outro garantindo que a componente sazonal é ativada ou desativada de acordo com a verificação executada anteriormente.

Previsão e avaliação: uma vez mais, as previsões são geradas para os últimos 154 dias e posteriormente agrupadas em formato semanal para possibilitar comparação com as observações reais, tudo compilado num subset análogo que permite a posterior avaliação do modelo de acordo com as métricas escolhidas (MAE e MAPE).

Ao contrário do que acontece com o modelo ARIMA que requer que a série temporal apresente estacionariedade, este modelo não impõe esse pressuposto. O Facebook Prophet foi desenvolvido com a capacidade de lidar com séries que apresentem tendências não lineares e sazonalidades recorrentes, permitindo desta forma a modelação direta de séries não estacionárias.

Bloco 11 - Modelação Facebook Prophet

```
def prophet_model(train_set, test_set, seasonality):
    # Create model instance
    model = Prophet(interval_width = 0.25, yearly_seasonality = seasonality)
    # Fit model to train set
    model.fit(train_set)
    # Run predictions on test set
    forecast = model.predict(test_set).astype(int)
    return forecast
```

3.4.4. Modelo XGBoost

Para termo de comparação entre aquilo que são métodos estatísticos e aquilo que será a implementação de um método de aprendizagem automática, foi selecionado o modelo XGBoost. Tal como nos dois anteriores, a aplicação deste modelo foi executada de forma sequenciada para cada um dos artigos selecionados previamente. Desta forma, as etapas do desenvolvimento e implementação do respetivo modelo podem ser definidas desta forma:

Adaptação da série: tal como nos anteriores, a função de adaptação extrai um subset da base de dados original contendo apenas observações do material em iteração.

Criação de novas variáveis: através dos dados existentes, são criadas as variáveis abaixo:

- Variáveis de calendário: dia da semana, semana, mês, ano, dia do mês, se é fim-de-semana, se é início do mês, se é fim do mês.
- Variáveis cíclicas: deteção de padrões cíclicos em intervalos semanais ou mensais (7 e 30).
- Defasagens: introdução de variáveis refletoras dos valores desfasados de carácter diário, semanal, bissemanal e mensal (1, 7, 14 e 30).
- Médias móveis: criação de variáveis de médias móveis de carácter semanal e mensal (7 e 30).

Divisão de sets: da mesma forma que nos modelos anteriores, para garantir a comparabilidade entre os mesmos, o set de treino corresponde a todas as datas observadas até 154 dias ates da última e o set de teste corresponde aos últimos 154.

Modelação: a instância do modelo é criada com três argumentos, um ajustando o número de estimadores (100), outro ajustado o rácio de aprendizagem (0,1) e um último garantindo que o modelo é aplicado sempre aos mesmos dados.

Previsão e avaliação: as previsões são geradas para os últimos 154 dias e posteriormente agrupadas em formato semanal. Para a comparação com as observações reais, é tudo compilado num subset análogo que permite aplicar as métricas de avaliação.

Bloco 12 - Modelação XGBoost

```
def xgb_model(X_train, y_train, X_test):
    # Create model instance
    model = xgb.XGBRegressor(n_estimators = 100, learning_rate = 0.1, \
    random_state = 42)
    # Fit model to train set
    model.fit(X_train, y_train)
    # Forecast
    y_pred = model.predict(X_test).astype(int)
    return forecast
```

3.4.5. Modelo atual

Tal como abordado num dos pontos iniciais deste capítulo, o modelo atual assenta fundamentalmente em inputs manuais (informações provenientes da área comercial) sobre aquilo que é a criação de um orçamento global de periodicidade anual. Desta feita, também para termo de comparação, foram comparadas as séries das observações reais, com as previsões lançadas pelo método que a organização utiliza atualmente. Para tal, foram seguidos alguns dos passos dados nos modelos anteriores com a pequena particularidade de que neste caso, as previsões já estariam lançadas e foram importadas diretamente de um ficheiro Excel (.xlsx).

Adaptação das séries: filtragem de ambas as séries por código de produto e agregação semanal apenas da série de observações reais, uma vez que as previsões atuais já trazem de origem o mesmo formato.

Divisão de sets: à semelhança do utilizado nos modelos anteriores, utilizou-se como set de teste as últimas 22 semanas da base de dados, valor que corresponde aos 154 dias utilizados nos modelos anteriores.

Avaliação: comparando os valores reais observados aos valores lançados pelo modelo atual, construíram-se novamente ambas as métricas para cada um dos produtos – MAE e MAPE.

3.5. Avaliação e seleção

Tal como as funções de apoio à fase de modelação, a etapa de avaliação também tem por base a utilização de uma função capaz de medir e avaliar a *performance* de cada modelo. Esta tem como base os valores reais observados e previstos por cada modelo, compilados numa nova tabela comparativa, que posteriormente permitirá executar a análise de *performance* de acordo com as métricas utilizadas. Para tal, como já mencionado anteriormente, foram selecionadas duas métricas estatísticas:

Erro Absoluto Médio: medição do erro médio entre os valores reais e os previstos;

Erro Percentual Absoluto Médio: medição do erro médio, em pontos percentuais.

Bloco 13 - Métricas de avaliação

```
def eval_metrics(df):
    # Calculate metrics
    mae = mean_absolute_error(df['Actual'], df['Forecast']).astype(int)
    mape = round(mean_absolute_percentage_error(df['Actual'], df['Forecast']), 2)
    return mae, mape
```

Para ambas as métricas, a interpretação será inversamente proporcional ao seu valor, isto é, valores mais reduzidos, indicam maior precisão e, consequentemente, modelos mais eficientes.

3.5.1. Desempenho ARIMA

A sequência de gráficos abaixo (Gráfico 8) demonstra a comparação entre aquilo que são os valores reais observados (a azul) e os valores previstos para cada uma das séries (a vermelho). Apesar de não acompanhar de forma tão ajustada as oscilações presentes na linha das observações reais, as linhas de previsão resultantes da aplicação do modelo estarão relativamente perto daquilo que será uma linha de tendência média. Este mesmo facto é mais notório nos resultados apresentados pelos produtos 200618, 200622 e 301998 onde a linha de previsão traça virtualmente ao meio aquilo que é a linha das observações reais. O mesmo acontece nas séries dos artigos 200018 e 302007 ainda que neste caso, as linhas não apresentem irregularidades, mantendo um comportamento praticamente constante.

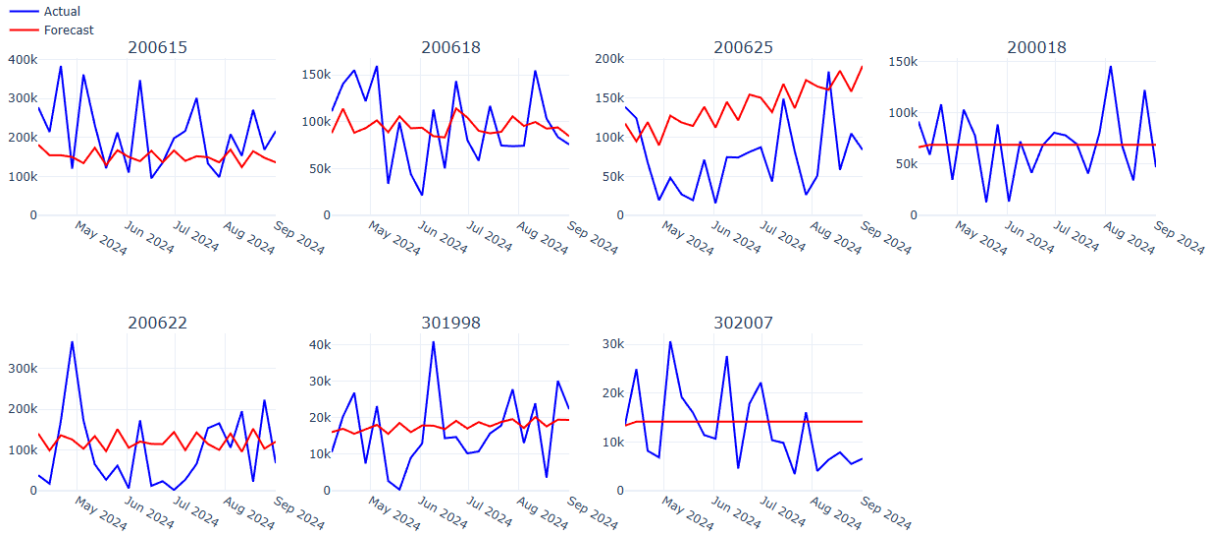

Gráfico 8 - Comparação de valores – modelo ARIMA

Os valores para as métricas de avaliação resultantes da aplicação do modelo ARIMA aos produtos selecionados traduz-se nos valores apresentados na Tabela 15, onde existe predominância de produtos que apresentam comportamento sazonal.

Tabela 15 - Métricas do modelo ARIMA

Material	Modelo	MAE	MAPE
200018	ARIMA	25 534	0,69
200615	SARIMA	75 531	0,32
200618	SARIMA	32 465	0,52
200622	SARIMA	88 119	6,05
200625	SARIMA	72 521	1,81
301998	SARIMA	7 714	3,38
302007	ARIMA	6 810	0,83

A comparação entre o erro absoluto e o erro percentual é também importante de destacar pois ao analisar apenas um destes de forma individual, poderiam se tirar conclusões precipitadas. Nem sempre um valor de MAE mais reduzido significa melhor *performance* do modelo no mesmo caso, e isso verifica-se principalmente nos últimos dois produtos da tabela, onde o valor do erro médio absoluto é inferior aos restantes, no entanto, o valor do erro médio percentual indica claramente que a *performance* não é a melhor (quando comparada aos restantes). Isto reforça que, é fundamental avaliar os modelos através de métricas distintas, para possibilitar tanto a captação da dimensão absoluta como da relativa do erro de previsão.

3.5.2. Desempenho Facebook Prophet

Ao analisar os gráficos de forma integrada (Gráfico 9), é possível identificar que em certa parte, o modelo consegue captar as tendências gerais de cada uma das séries, ainda que com maior sucesso numas em detrimento de outras. Ainda assim, os resultados gerados para os produtos 200018 e 302007 voltam a constituir uma preocupação ao nível da modulação executada já que os níveis de previsão seguem um valor praticamente constante ao longo de todo o intervalo de teste, com ligeiríssima tendência ascendente. Curiosamente, estes são os mesmos produtos que não apresentavam sazonalidade de acordo com o teste realizado anteriormente através da função desenvolvida para esse mesmo objetivo. Nestes, para uma melhor resposta, talvez se justificasse a modelação mais refinada dos parâmetros disponíveis (otimização de parâmetros), fator que poderia contribuir de forma positiva para a melhoria no desempenho destes dois artigos. Por outro lado, ajustar a componente sazonal poderia ser outro cenário a explorar para verificar se existiam diferenças nos resultados obtidos.

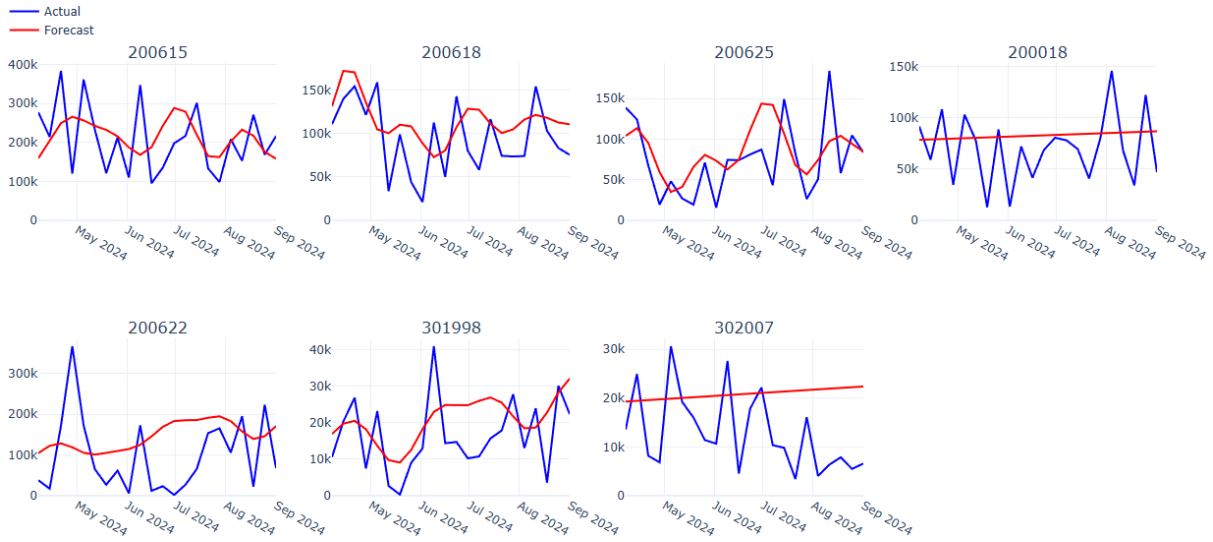

Gráfico 9 - Comparação de valores – modelo Facebook Prophet

Atendendo às métricas propriamente ditas, a Tabela 16 apresenta, para cada um dos produtos, os valores das respetivas métricas utilizadas para avaliar a aplicação do modelo.

Aqui podemos observar que os resultados são relativamente equilibrados entre si, com exceção do artigo 200622, que apresenta um grau de eficiência significativamente inferior aos restantes. Para além desse caso, este modelo volta a ter dificuldades em prever os níveis de procura para os artigos de azeite, onde habitualmente os volumes de vendas são mais reduzidos.

Tabela 16 - Métricas do modelo FB Prophet

Material	Modelo	MAE	MAPE
200018	FB Prophet	28 066	0,86
200615	FB Prophet	74 066	0,42
200618	FB Prophet	35 928	0,61
200622	FB Prophet	94 069	7,45
200625	FB Prophet	32 171	0,76
301998	FB Prophet	8 784	2,12
302007	FB Prophet	10 156	1,47

No sentido inverso, a referência 200615 apresenta o erro percentual mais baixo de todos os produtos, demonstrando assim um desempenho preditivo bastante satisfatório quando comparado aos restantes. Estes resultados estarão alinhados com a análise anterior quando se retiraram as primeiras ilações através da visualização dos gráficos presentes na Figura 12.

3.5.3. Desempenho XGBoost

Da mesma forma que nos dois modelos anteriores, a sequência de gráficos é analisada na perspetiva análoga de compreender a relação entre as linhas de valores reais e previstos.

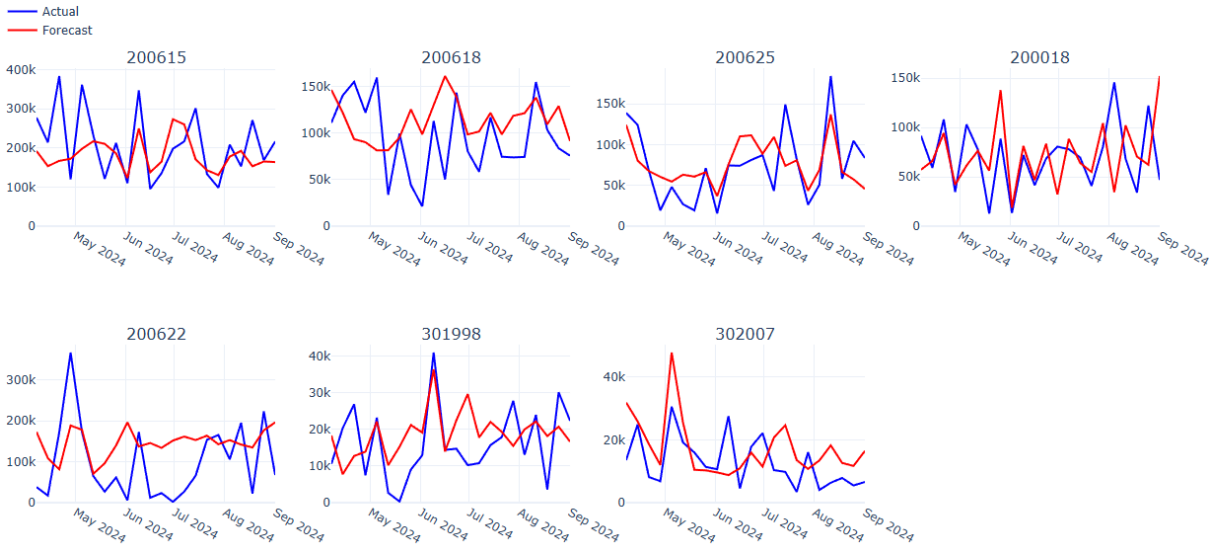

Gráfico 10 - Comparação de valores – modelo XGBoost

Desta feita, os valores obtidos através da aplicação do único modelo de aprendizagem automática utlizado no projeto demonstram uma capacidade superior de corresponder principalmente às oscilações dos valores reais observados, mas também à tendência geral de cada série – Gráfico 10. As únicas exceções estão situadas nos artigos 200618 e 200622 onde é percetível que as linhas não se acompanham tão aproximadamente como nos restantes, sobrestimando em certa parte os níveis de procura. Uma razão plausível para este acontecimento pode estar centrada na existência de valores de venda bastante baixos e até mesmo presença de observações muito próximas de zero, podendo dar indicação de se tratar de um produto com uma presença menos firme no mercado.

Tabela 17 - Métricas do modelo XGBoost

Material	Modelo	MAE	MAPE
200018	XGBoost	30 998	0,57
200615	XGBoost	64 930	0,30
200618	XGBoost	37 878	0,66
200622	XGBoost	87 151	7,18
200625	XGBoost	27 205	0,59
301998	XGBoost	8 222	2,92
302007	XGBoost	8 497	0,95

De forma geral, e analisando os valores da Tabela 17, a aplicação do modelo demonstrou algumas melhorias nos valores de ambas as métricas de avaliação de desempenho quando comparado com os dois anteriores, principalmente no artigo 200615 com uma redução do valor para o erro médio absoluto em cerca de dez mil litros.

3.5.4. Desempenho atual

A análise e exercício análogo dos valores obtidos através do modelo utilizado atualmente consta também no projeto pois, apesar da noção de que pode ser melhorada a sua eficiência, é sempre importante para termo de comparação. O processo de análise de desempenho deste método será assim realizado exatamente da mesma forma que nos modelos anteriores, garantindo uma análise da sequência de gráficos de linhas bem como da tabela demonstrativa dos valores para ambas as métricas de avaliação. Dito isto, a comparação de valores apresentada na Gráfico 11, apesar de demonstrar algumas debilidades em certos artigos, não estará tão fora da realidade quanto se poderia imaginar. Artigos como o 200615, tal como nos modelos anteriores, apresenta um comportamento bastante aceitável naquilo que é o lançamento de previsões e é possível observá-lo através da proximidade entre ambas as linhas durante o período de teste. Uma vez mais, é também notória a sobrestimação em

alguns casos, particularmente nos artigos 200618, 200625 e 200622, onde as linhas até parecem seguir a mesma tendência e acompanhar as irregularidades de forma consistente, mas com ligeiro exagero naquilo que será a quantidade de litros prevista.

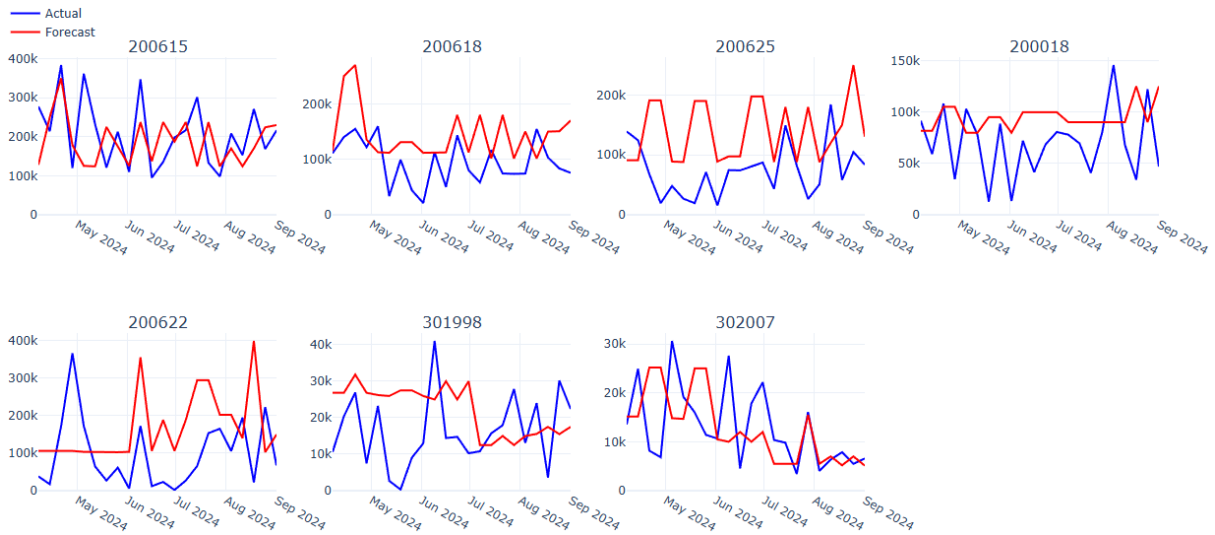

Gráfico 11 - Comparação de valores – modelo atual

Observando os valores apresentados na Tabela 18, é possível identificar uma clara melhoria no último artigo, uma vez que os valores de ambas as métricas são bem inferiores aos atingidos pelos restantes modelos. Ainda assim, de forma geral, é também possível afirmar que esses mesmos modelos se comportam de forma mais eficiente quando comparados ao modelo atual, isto pois, para todos os produtos, os valores das respetivas métricas ficam abaixo do obtido pelo método utilizado de momento.

Tabela 18 - Métricas do modelo atual

Material	Modelo	MAE	MAPE
200018	Atual	36 054	1,08
200615	Atual	73 111	0,36
200618	Atual	59 794	0,94
200622	Atual	120 700	6,01
200625	Atual	78 811	2,03
301998	Atual	11 843	5,14
302007	Atual	6 929	0,63

Tal como nos restantes modelos, a referência 200622 volta a ser o produto mais difícil de prever, registando novamente valores em ambas as métricas muito superiores aos dos restantes e confirmando, assim, a tendência constante na dificuldade de encontrar um modelo capaz de gerar níveis de procura suficientemente bons para este caso.

3.5.5. Seleção

Uma vez terminado o processo de avaliação de *performance*, o projeto atinge a fase final onde é selecionado o modelo que se adeque melhor a cada produto, respeitando uma métrica composta que é desenvolvida através de uma combinação ponderada de ambas as métricas utilizadas no estudo. Para tal, os valores das mesmas são previamente normalizados para permitir uma comparação equitativa entre elas garantindo desta feita que ambas as métricas se encontram na mesma escala de grandeza. Esta agregação ponderada de métricas de avaliação de *performance* dão origem a um rácio indicador de desempenho de acordo com o cálculo abaixo:

$$r = (0{,}2 \times MAE_{norm}) + (0{,}8 \times MAPE_{norm}) \quad (10)$$

A criação deste rácio segue uma ponderação relativamente maior para o valor do erro médio percentual (80%) pois entende-se que este irá refletir melhor o impacto em escala, de acordo com o volume de cada produto, sem descurar o erro médio absoluto, já que poderá indicar desvios na unidade de medida em estudo, neste caso litros. A Tabela 19 proporciona uma visão global de todos os valores das métricas em cada modelo e permite a comparação entre os mesmos.

Tabela 19 - Comparação de métricas entre modelos

	ARIMA		FB Prophet		XGBoost		Atual	
Material	**MAE**	**MAPE**	**MAE**	**MAPE**	**MAE**	**MAPE**	**MAE**	**MAPE**
200018	25 534	0,69	28 066	0,86	30 998	0,57	36 054	1,08
200615	75 531	0,32	74 066	0,42	64 930	0,30	73 111	0,36
200618	32 465	0,52	35 928	0,61	37 878	0,66	59 794	0,94
200622	88 119	6,05	94 069	7,45	87 151	7,18	120 700	6,01
200625	72 521	1,81	32 171	0,76	27 205	0,59	78 811	2,03
301998	7 714	3,38	8 784	2,12	8 222	2,92	11 843	5,14
302007	6 810	0,83	10 156	1,47	8 497	0,95	6 929	0,63

Em alguns casos a opção pode ser bastante direta, pois os valores entre os três modelos são bastante díspares, no entanto, em alguns casos essa escolha pode não ser assim tão fácil de tomar, como é o caso do produto 200615 que apresenta muito pouca diferença entre aquilo que é o modelo FB Prophet e o modelo ARIMA por exemplo. Em sentido contrário, é possível fazer a mesma análise, e verificar através dos valores apresentados na tabela acima que a seleção do melhor modelo para gerar os níveis de procura do artigo 200625 muito provavelmente será o modelo de aprendizagem automática – XGBoost.

Assim sendo, depois de realizado o cálculo do rácio acima apresentado a todos os produtos e respetivos modelos, a opção naturalmente cairá sobre o modelo que apresentar melhor eficiência, ou seja, neste caso, um menor valor para o rácio ponderado. A Tabela 20 apresenta o modelo escolhido, os valores de ambas as métricas e o respetivo rácio ponderado:

Tabela 20 - Seleção final de modelos

Material	Modelo	MAE	MAPE	Rácio
200018	XGBoost	25 534	0,69	0,07
200615	XGBoost	64 930	0,30	0,10
200618	SARIMA	32 465	0,52	0,07
200622	SARIMA	88 119	6,05	0,79
200625	XGBoost	27 205	0,59	0,07
301998	FB Prophet	8 784	2,12	0,21
302007	Atual	6 929	0,63	0,04

Em resumo, convém comparar aquilo que é o rácio médio calculado para ambos os cenários, ou seja, para o atual, onde todos os artigos têm as suas previsões baseadas num só método, e para um híbrido, onde cada produto vê os seus níveis de procura serem gerados pelo modelo que melhor se adapta ao comportamento da respetiva série – Tabela 21.

Tabela 21 - Comparação entre rácios

Material	Modelo	Rácio	Modelo	Rácio
200018	Atual	0,14	XGBoost	0,07
200615	Atual	0,12	XGBoost	0,10
200618	Atual	0,16	SARIMA	0,07
200622	Atual	0,84	SARIMA	0,79
200625	Atual	0,32	XGBoost	0,07
301998	Atual	0,55	FB Prophet	0,21
302007	Atual	0,04	Atual	0,04
	Média (a)	0,31	**Média** (h)	0,19

Calculando um rácio que compara o desempenho isolado através do modelo atual frente ao desempenho da adoção desta nova abordagem híbrida, obtém-se uma melhoria (m) de cerca de 38%, valor superior ao inicialmente proposto nos objetivos iniciais do projeto. Abaixo a exemplificação do cálculo deste rácio comparativo:

$$m = \frac{a - h}{a} \times 100\% \quad (11)$$

3.6. Operacionalização

Uma vez terminado o processo de avaliação e seleção, entra-se na fase que ultima o projeto, a aplicação do respetivo modelo a cada artigo para efetivamente realizar as primeiras previsões. Esta tarefa é suportada uma vez mais por uma função iterativa que atua sobre cada um dos produtos, onde será aplicada uma condição para a seleção e aplicação do modelo de acordo com a análise feita anteriormente, e onde serão previstos os 28 dias posteriores à data da observação final da base de dados, agregando posteriormente os respetivos valores por semana. Neste caso, são previstos os níveis de procura dos próximos 28 dias pois a organização atualmente gere o processo de planeamento de necessidades com um horizonte a quatro semanas, e o decidido foi manter esse mesmo princípio. Estas previsões são então guardadas num ficheiro .csv que mais tarde irá alimentar, junto com as observações reais, o painel interativo desenvolvido em Power BI para suportar a tomada de decisão nos processos de planeamento de necessidades e produção. Para tal, foram o passo-a-passo foi efetuado de acordo com os pontos abaixo:

1. **Importação de dados:** importação do ficheiro .csv com previsões e vendas reais.
2. **Divisão de séries:** divisão entre série de vendas reais e previsões.
3. **Adição de coluna:** adição de coluna com tipologia de dados – “Atual” e “Previsão”.
4. **Agregação:** combinação de ambas as séries numa final.
5. **Adição de coluna:** adição de coluna com tipologia de dados – “OL” e “AZ”.
6. **Criação de medidas:** criação de medidas – médias, mínimos e máximos.
7. **Criação de limites preditivos:** limite inferior e superior calculado através do valor da previsão subtraído ou somado ao valor do respetivo erro absoluto médio.

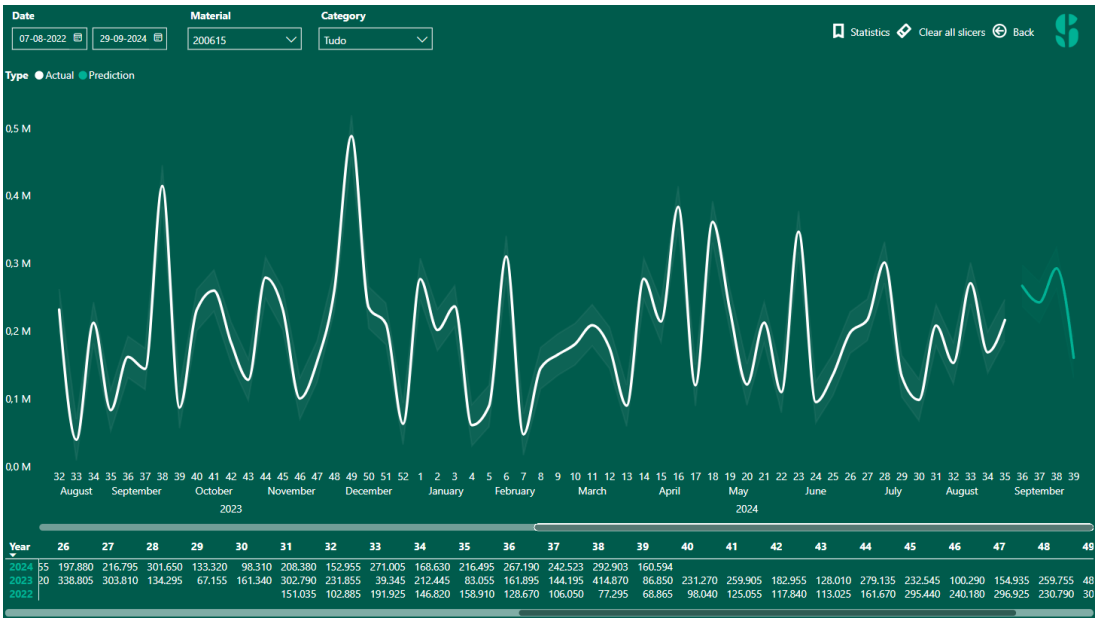

Figura 11 - Painel interativo em Power BI

Conclusão e investigação futura

Durante este percurso, o projeto foi fundamentalmente desenvolvido para promover uma melhoria na eficiência do lançamento dos níveis de procura que suportam o processo de planeamento de necessidades da organização. O trabalho realizado consiste num novo processo de avaliação comparativa entre modelos distintos, um modelo utilizado atualmente, dois modelos de carácter mais estatístico – ARIMA e Facebook Prophet – e um final relacionado com a área de aprendizagem automática – XGBoost. Para possibilitar uma analogia entre os diferentes modelos, as métricas utilizadas foram o Erro Absoluto Médio e o Erro Percentual Absoluto Médio, que mais tarde suportam a criação de um rácio ponderado que ditará o veredito final sobre qual o modelo a implementar em cada um dos produtos.

A comparação dos valores gerados por esta nova abordagem permitem criar um cenário híbrido, onde para cada produto se utiliza o modelo que melhor se adequa ao comportamento da sua série temporal e onde os resultados obtidos demonstram claramente, superando as expectativas iniciais, que esta nova metodologia será uma mais-valia naquilo que é o processo de previsão da procura e consequentemente os processos consequentes. Desta feita, é possível afirmar que tendo como pilar o objetivo geral de aplicar um método capaz de gerar previsões para os níveis de procura em contexto multiproduto, os objetivos específicos foram integralmente alcançados e isso ficou demonstrado através dos resultados obtidos. Revisitando os objetivos específicos por ordem:

- **Implementação de diferentes modelos preditivos:** foram selecionados três modelos, dois de natureza estatística e um de aprendizagem automática. Cada um deles foi implementado a cada uma das séries temporais – cenário multiproduto.
- **Comparação do desempenho entre modelos:** conforme referenciado na Tabela 19, foi possível comparar os valores das métricas de avaliação de desempenho entre os diferentes modelos. Os Gráficos 8 a 11 também suportam parte deste processo de comparação de forma mais visual e intuitiva numa perspetiva de análoga de verificar o comportamento de ambas as linhas de observações, as reais e as previstas.
- **Proposta de abordagem híbrida:** a criação de um rácio ponderado (Fórmula 10) permitiu atribuir pontuações aos diferentes modelos de forma mais objetiva. A Tabela 20 demonstra que o método híbrido seleciona o modelo de acordo com o valor mais baixo do respetivo rácio.
- **Avaliar impacto na eficiência preditiva:** a Tabela 21 vem evidenciar a melhoria aproximada na eficiência em cerca de 38%. Este resultado é fruto da aplicação da Fórmula 11, assumindo como argumentos o valor do rácio médio do método atual – 0,31 – contra o valor do rácio médio obtido através da abordagem híbrida – 0,19.

Impacto organizacional

Para além desta validação técnica, a solução proposta neste projeto pode representar algumas alterações a nível estratégico para a organização, não só através de uma vertente operacional que beneficia com melhores e mais fiáveis níveis de previsão para a procura dos seus produtos, mas também da sua maturidade analítica e capacidade melhorada no processo de tomada de decisão.

Desta feita, os impactos mais relevantes neste sentido podem destacar-se por:

- **Fiabilidade nas previsões:** a capacidade de prever cenários futuros com maior precisão permite um planeamento mais assertivo nas necessidades de produção, reduzindo consequentemente riscos relacionados com inventário, seja em caso de rutura ou excesso – efeito direto na redução de custos operacionais.
- **Agilidade organizacional:** maior capacidade de adaptação a novos padrões de procura que vão surgindo no mercado, permite à empresa responder com maior agilidade a essas mesmas variações fortalecendo a sua componente de vantagem competitiva.
- **Alavanca para a transformação digital:** a introdução deste tipo de métodos contribui para a evolução da maturidade analítica da organização e serve como ponte para reforçar a importância na criação de condições para futuros desenvolvimentos no âmbito digital.

Investigação futura

Ainda assim, apesar dos resultados obtidos através da abordagem desenvolvida durante o projeto, alguns desafios permanecem por desbravar e também esses merecem atenção para investigação futura. Uma das limitações encontrada prende-se fundamentalmente na dificuldade de prever os níveis de procura de artigos novos no mercado, ou que contenham poucos dados históricos para alimentar os respetivos modelos. Esta centra-se como um dos obstáculos mais predominante naquilo que são os projetos no âmbito da previsão de séries temporais, onde será importante também ganhar algum terreno e tentar explorar novas técnicas de aprendizagem que permitam por exemplo, gerar os níveis de previsão necessários através de atributos específicos para cada produto, como é o caso do formato, tipo de produto ou até mesmo marca. Precisamente por este motivo, a lista que compôs o leque selecionado de produtos para teste tem a presença dos produtos mais transacionados pela empresa ao longo do período em estudo, permitindo uma avaliação mais focada e representativa de todo o processo.

Mais relacionado com o desenvolvimento do projeto em si, uma opção válida a explorar em investigações futuras seria a aplicação de métodos de otimização de parâmetros nos diferentes modelos que permitam maximizar o seu desempenho, ajustando-os assim às particularidades de cada combinação modelo/série temporal.

Adicionalmente, e neste caso mais centrado naquilo que seria uma possível continuação deste projeto, propor-se-ia o desenvolvimento de um pipeline capaz de automatizar todo este processo, ou seja, algo escalável e integrado que possibilitasse não só o cálculo dos níveis de procura, mas também uma atualização periódica dos dados (com periodicidade a definir), e consequentemente os modelos. Isto, pois, nem todos os modelos são infalíveis, e o comportamento que determinado artigo está a apresentar neste momento, pode nem sempre se manter durante o período expectável. Esta adição final poderia significar um toque preponderante para garantir a eficiência operacional e a maturidade analítica da empresa.

Referências bibliográficas

Abd Elmonem, M. A., Nasr, E. S., & Geith, M. H. (2016). Benefits and challenges of cloud ERP systems – A systematic literature review. *Future Computing and Informatics Journal*, *1*(1–2), 1–9. https://doi.org/10.1016/j.fcij.2017.03.003

Agyemang, E. F., Mensah, J. A., Ocran, E., Opoku, E., & Nortey, E. N. N. (2023). Time series based road traffic accidents forecasting via SARIMA and Facebook Prophet model with potential changepoints. *Heliyon*, *9*(12). https://doi.org/10.1016/j.heliyon.2023.e22544

Aishah, S., Selamat, M., Prakoonwit, S., Sahandi, R., Khan, W., & Ramachandran, M. (2018). Big Data Analytics-A Review of Data Mining Models for SMEs in the Transportation Sector. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*.

Akkermans, H. A., Bogerd, P., Y€ U Ucesan, E., & Van Wassenhove, L. N. (2003). The impact of ERP on supply chain management: Exploratory findings from a European Delphi study. In *European Journal of Operational Research* (Vol. 146). www.elsevier.com/locate/dsw

An, H., Liu, Z., & Tian, F. (2025). A machine learning framework for coal gasification process simulation and operation optimization. *Fuel*, *401*, 135934. https://doi.org/10.1016/j.fuel.2025.135934

Bai, P., Chen, Y., Chen, L., Zhang, X., Wang, X., & Wang, X. (2024). Research on the influence of bad working state on air traffic control effect based on multi-independent sample Kruskal-Wallis test. *Journal of Air Transport Management*, *120*. https://doi.org/10.1016/j.jairtraman.2024.102653

Balan, G. S., Kumar, V. S., & Raj, S. A. (2025). Machine learning and artificial intelligence methods and applications for post-crisis supply chain resiliency and recovery. In *Supply Chain Analytics* (Vol. 10). Elsevier B.V. https://doi.org/10.1016/j.sca.2025.100121

Ben Hamou, K. A., Jarir, Z., & Elfirdoussi, S. (2025). Using machine learning for production scheduling problems in the supply chain: A review. *Computers and Industrial Engineering*, *206*. https://doi.org/10.1016/j.cie.2025.111243

Bhatsada, A., Payomthip, P., Itsarathorn, T., Lwin, Y. N. N., Wahyanti, E., Towprayoon, S., Patumsawad, S., Chiemchaisri, C., & Wangyao, K. (2025). AI-powered machine learning models for monitoring and optimization of biodrying process. *Results in Engineering*, *26*, 105584. https://doi.org/10.1016/j.rineng.2025.105584

Bi, Z., Xu, L. Da, & Wang, C. (2014). Internet of things for enterprise systems of modern manufacturing. *IEEE Transactions on Industrial Informatics*, *10*(2), 1537–1546. https://doi.org/10.1109/TII.2014.2300338

Büyüközkan, G., & Göçer, F. (2018). Digital Supply Chain: Literature review and a proposed framework for future research. *Computers in Industry*, *97*, 157–177. https://doi.org/10.1016/j.compind.2018.02.010

Chatfield, C. (2000). Time-series forecasting. In *Time-Series Forecasting*.

Cheng, J., Tiwari, S., Khaled, D., Mahendru, M., & Shahzad, U. (2024). Forecasting Bitcoin prices using artificial intelligence: Combination of ML, SARIMA, and Facebook Prophet models. *Technological Forecasting and Social Change*, *198*. https://doi.org/10.1016/j.techfore.2023.122938

Choi, T. M., Wallace, S. W., & Wang, Y. (2016). Risk management and coordination in service supply chains: Information, logistics and outsourcing. *Journal of the Operational Research Society*, *67*(2), 159–164. https://doi.org/10.1057/jors.2015.115

Chopra, R., Sawant, L., Kodi, D., & Terkar, R. (2022). Utilization of ERP systems in manufacturing industry for productivity improvement. *Materials Today: Proceedings*, *62*, 1238–1245. https://doi.org/10.1016/j.matpr.2022.04.529

Cooper, M. C., Lambert, D. M., & Pagh, J. D. (1997). Supply Chain Management: More Than a New Name for Logistics. *The International Journal of Logistics Management*, *8*(1), 1–14. https://doi.org/10.1108/09574099710805556

Croxton, K. L., Lambert, D. M., García-Dastugue, S. J., & Rogers, D. S. (2002). The Demand Management Process. *The International Journal of Logistics Management*, *13*(2), 51–66. https://doi.org/10.1108/09574090210806423

Culot, G., Podrecca, M., & Nassimbeni, G. (2024). Artificial intelligence in supply chain management: A systematic literature review of empirical studies and research directions. In *Computers in Industry* (Vol. 162). Elsevier B.V. https://doi.org/10.1016/j.compind.2024.104132

de Mattos, C. A., Correia, F. C., & Kissimoto, K. O. (2024). Artificial Intelligence Capabilities for Demand Planning Process. *Logistics*, *8*(2), 53. https://doi.org/10.3390/logistics8020053

de Oliveira, E. M., & Cyrino Oliveira, F. L. (2018). Forecasting mid-long term electric energy consumption through bagging ARIMA and exponential smoothing methods. *Energy*, *144*, 776–788. https://doi.org/10.1016/j.energy.2017.12.049

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.

Fattah, J., Ezzine, L., Aman, Z., El Moussami, H., & Lachhab, A. (2018). Forecasting of demand using ARIMA model. *International Journal of Engineering Business Management*, *10*. https://doi.org/10.1177/1847979018808673

Feizabadi, J. (2022). Machine learning demand forecasting and supply chain performance. *International Journal of Logistics Research and Applications*, *25*(2), 119–142. https://doi.org/10.1080/13675567.2020.1803246

Fleischmann, M., Bloemhof-Ruwaard, J. M., Dekker, R., Van Der Laan, E., Van Nunen, J. A. E. E., & Van Wassenhove, L. N. (1997). EUROPEAN JOURNAL OF OPERATIONAL RESEARCH Quantitative models for reverse logistics: A review. In *European Journal of Operational Rcsearch* (Vol. 103).

Garay-Rondero, C. L., Martinez-Flores, J. L., Smith, N. R., Caballero Morales, S. O., & Aldrette-Malacara, A. (2020). Digital supply chain model in Industry 4.0. *Journal of Manufacturing Technology Management*, *31*(5), 887–933. https://doi.org/10.1108/JMTM-08-2018-0280

George E. P. Box, & Gwilym M. Jenkins. (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day.

Germain, R., Claycomb, C., & Dröge, C. (2008). Supply chain variability, organizational structure, and performance: The moderating effect of demand unpredictability. *Journal of Operations Management*, *26*(5), 557–570. https://doi.org/10.1016/j.jom.2007.10.002

Getachew, M., Beshah, B., Mulugeta, A., & Kitaw, D. (2024). Application of artificial intelligence to enhance manufacturing quality and zero-defect using CRISP-DM framework. *International Journal of Production Research*. https://doi.org/10.1080/00207543.2024.2407919

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). *Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation*. https://scholarworks.utep.edu/cs_techrep/1209

Han, G., & Dong, M. (2015). Trust-embedded coordination in supply chain information sharing. *International Journal of Production Research*, *53*(18), 5624–5639. https://doi.org/10.1080/00207543.2015.1038367

Harvey, A. C., & Peters, S. (1990). Estimation Procedures for Structural Time Series Models. In *Journal of Forecasting* (Vol. 9).

Hofmann, E., & Rutschmann, E. (2018). Big data analytics and demand forecasting in supply chains: a conceptual analysis. *International Journal of Logistics Management*, *29*(2), 739–766. https://doi.org/10.1108/IJLM-04-2017-0088

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Jahani, H., Jain, R., & Ivanov, D. (2023). Data science and big data analytics: a systematic review of methodologies used in the supply chain and logistics research. *Annals of Operations Research*. https://doi.org/10.1007/s10479-023-05390-7

Kraus, S., Durst, S., Ferreira, J. J., Veiga, P., Kailer, N., & Weinmann, A. (2022). Digital transformation in business and management research: An overview of the current status quo. *International Journal of Information Management*, *63*. https://doi.org/10.1016/j.ijinfomgt.2021.102466

Kühl, N., Goutier, M., Baier, L., Wolff, C., & Martin, D. (2022). Human vs. supervised machine learning: Who learns patterns faster? *Cognitive Systems Research*, *76*, 78–92. https://doi.org/10.1016/j.cogsys.2022.09.002

Lambert, D. M., & Enz, M. G. (2017). Issues in Supply Chain Management: Progress and potential. *Industrial Marketing Management*, *62*, 1–16. https://doi.org/10.1016/j.indmarman.2016.12.002

Manavalan, E., & Jayakrishna, K. (2019). A review of Internet of Things (IoT) embedded sustainable supply chain for industry 4.0 requirements. *Computers and Industrial Engineering*, *127*, 925–953. https://doi.org/10.1016/j.cie.2018.11.030

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8), 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

Mentzer, J. T., DeWitt, W., Keebler, J. S., Min, S., Nix, N. W., Smith, C. D., & Zacharia, Z. G. (2001). DEFINING SUPPLY CHAIN MANAGEMENT. *Journal of Business Logistics*, *22*(2), 1–25. https://doi.org/10.1002/j.2158-1592.2001.tb00001.x

Negre, P., Alonso, R. S., Prieto, J., García, Ó., & de-la-Fuente-Valentín, L. (2024). Prediction of footwear demand using Prophet and SARIMA. *Expert Systems with Applications*, *255*. https://doi.org/10.1016/j.eswa.2024.124512

Nie, Y., Zhang, Q., Hou, L., Du, L., Zhao, X., Xue, Y., Lin, Z. C., Cao, X., & Wang, Z. (2025). TBM rock mass classification using XGBoost and Interpretable Machine learning. *Advanced Engineering Informatics*, *66*. https://doi.org/10.1016/j.aei.2025.103459

Org, E., Paparoditis, E., & Politis, D. N. (2018). The Asymptotic Size and Power of the Augmented Dickey-Fuller Test for a Unit Root. *Econometric Reviews*, *37*(9), 955–973.

Raman, S., Patwa, N., Niranjan, I., Ranjan, U., Moorthy, K., & Mehta, A. (2018). Impact of big data on supply chain management. *International Journal of Logistics Research and Applications*, *21*(6), 579–596. https://doi.org/10.1080/13675567.2018.1459523

Said, S. E., & Dickey, D. A. (1984). Testing for unit roots in autoregressive moving-average models with unknown order. *Biometrika*, *71*(3), 599–607.

Saltz, J. S. (2021). CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps. *2021 IEEE International Conference on Big Data (Big Data)*, 2337–2344. https://doi.org/10.1109/BigData52589.2021.9671634

Sovena Group. (2025, March 21). *Segmentos de atuação*. https://www.sovenagroup.com/pt/o-nosso-mundo/segmentos-de-atuacao/

Taylor, S. J., & Letham, B. (2018). Forecasting at Scale. *American Statistician*, *72*(1), 37–45. https://doi.org/10.1080/00031305.2017.1380080

Umble, E. J., Haft, R. R., & Michael Umble, M. (2003). Enterprise resource planning: Implementation procedures and critical success factors. *European Journal of Operational Research*, *146*(2), 241–257. https://doi.org/10.1016/S0377-2217(02)00547-7

Yadav, A., Garg, R. K., & Sachdeva, A. (2024). Artificial intelligence applications for information management in sustainable supply chain management: A systematic review and future research agenda. *International Journal of Information Management Data Insights*, *4*(2). https://doi.org/10.1016/j.jjimei.2024.100292

Yu, T. K., Chang, I. C., Chen, S. De, Chen, H. L., & Yu, T. Y. (2025). Predicting potential soil and groundwater contamination risks from gas stations using three machine learning models (XGBoost, LightGBM, and Random Forest). *Process Safety and Environmental Protection*, *199*. https://doi.org/10.1016/j.psep.2025.107249

Zolbanin, H., & Aubert, B. (2025). A process model for design-oriented machine learning research in information systems. In *Journal of Strategic Information Systems* (Vol. 34, Issue 1). Elsevier B.V. https://doi.org/10.1016/j.jsis.2024.101868

Zouari, D., Ruel, S., & Viale, L. (2021). Does digitalising the supply chain contribute to its resilience? *International Journal of Physical Distribution and Logistics Management*, *51*(2), 149–180. https://doi.org/10.1108/IJPDLM-01-2020-0038