INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Previsão de vendas para apoio em gestão de farmácias

Ricardo Manuel Sampaio Jorge Rolim

Mestrado em Ciência de Dados

Orientadora:
Prof. Doutora Patrícia Filipe, Professora Associada
ISCTE Business School

outubro, 2023

Departamento de Métodos Quantitativos para Gestão e Economia

Departamento de Ciências e Tecnologias da Informação

Previsão de vendas para apoio em gestão de farmácias

Ricardo Manuel Sampaio Jorge Rolim

Mestrado em Ciência de Dados

Orientadora:
Prof. Doutora Patrícia Filipe, Professora Associada
ISCTE Business School

outubro, 2023

Agradecimentos

Manifesto o meu entusiasmo pela oportunidade da realização da dissertação de mestrado nesta instituição, proporcionada pelos docentes do ISCTE. Este desafio foi possível pelo interesse e disponibilidade demonstrado pela empresa que facultou os dados.

O meu especial agradecimento vai para a minha família, amigos e para a minha orientadora Professora Doutora Patrícia Filipe.

Resumo

Este estudo tem como objetivo apresentar uma revisão de literatura e um projeto prático de ciência de dados sobre previsão de vendas farmacêuticas e suporte aos serviços e à gestão farmacêutica. A previsão de vendas é um fator crítico na logística e na gestão da cadeia de distribuições. Na componente prática são utilizadas ferramentas como *SQL* e *Python*, em etapas realizadas no contexto *CRISP-DM*. No âmbito técnico do presente projeto, tem maior destaque o uso de modelos de séries temporais para a realização de previsões de vendas de embalagens de medicamentos, a partir de dados de histórico reais, de diferentes produtos e características de venda. A presente dissertação oferece informações recentes sobre previsão de vendas e conhecimento sobre cadeias de distribuição, fornecendo indícios sobre tendências atuais e revelando lacunas interessantes. Explorando as oportunidades e objetivos desta área de negócio relativa à venda de medicamentos, foram nas diferentes fases do projeto, utilizados dados disponibilizados por uma empresa deste setor, existindo uma recolha, limpeza e preparação dos dados para a realização da modelação para as previsões de vendas em períodos futuros. As previsões finais foram obtidas utilizando o modelo *Prophet*, tendo sido efetuadas avaliações de desempenho para conclusão da afinação do modelo. Com este projeto é possível entregar à área de negócio dados quantitativos das vendas de medicamentos previstas e gráficos que revelam comportamentos de vendas e indicações das tendências e sazonalidades de vendas, para alcance de reduções em ruturas de medicamentos e em custos de gestão de *stock*.

Palavras-Chave: Modelos de Previsão de Vendas, Produtos Farmacêuticos, Ruturas de Medicamentos, Séries Temporais, Prophet

Abstract

This study aims to present a literature review and a practical data science project on pharmaceutical sales forecasting and support for pharmaceutical services and management. Sales forecasting is a critical factor in logistics and distribution chain management. The practical component uses tools such as SQL and Python, in stages carried out in the CRISP-DM context. Within the technical scope of this project, the use of time series models for forecasting sales of pharmaceutical packaging, based on real historical data for different products and sales characteristics, is of particular importance. This dissertation offers recent information on sales forecasting and knowledge about distribution chains, providing clues about current trends and revealing interesting gaps. Exploring the opportunities and objectives of this business area relating to the sale of medicines, the different phases of the project used data provided by a company in this sector, collecting, cleaning and preparing the data for modelling sales forecasts for future periods. The final forecasts were obtained using the Prophet model, and performance evaluations were carried out to finalize the model. With this project it is possible to provide the business area with quantitative data on forecast medicine sales and graphs that reveal sales behavior and indications of sales trends and seasonality, in order to achieve reductions in medicine out-of-stocks, shortages and stock management costs.

Keywords: Sales Forecasting Models, Pharmaceutical Products, Medicines Shortages, Time Series, Prophet

Índice Geral

Índice de Figuras

CAPÍTULO 1

Introdução

O presente projeto de investigação desenvolvido no mestrado de ciência de dados, possui no contexto do mercado da área de negócio da saúde o objetivo de direcionar farmácias e empresas especializadas em dados de produtos farmacêuticos e respetivas vendas, a alcançarem soluções de otimização de *stock* e diminuição de ruturas para períodos futuros, utilizando modelos de previsão de vendas de medicamentos e dispositivos médicos. Ao ter sido desenvolvido na língua portuguesa, este estudo pretende complementar a falta das referidas soluções em território português.

1.1. Panorama atual

Há cerca de duas décadas, as transações eletrónicas (transações com suporte de tecnologias da informação e comunicação) na indústria farmacêutica eram efetivamente mais simples e básicas, e a aplicação da tecnologia era limitada. Atualmente, existe cada vez mais tecnologia a ser usada para ajudar as empresas farmacêuticas a gerir o seu *stock* [1]. A previsão de vendas na indústria farmacêutica tem uma estrutura mais complexa do que em outros setores. Fatores humanos, doenças sazonais e epidémicas, participação de mercado dos produtos concorrentes e condições de comercialização são considerados os principais fatores externos para a previsão de produtos farmacêuticos. Para além destas variáveis, as substâncias ativas dos medicamentos também representam um fator importante para o processo de previsão [2].

As elevadas margens de lucro obtidas com a venda de medicamentos de referência (originais) permitiram à indústria farmacêutica ter elevados custos de processo logístico. Expiração de patentes e, como resultado, um aumento considerável no número de empresas de produção de genéricos, que se concentram no desenvolvimento de produtos eficientes e eficazes, obriga a indústria farmacêutica a focar-se nos desafios na previsão da futura oferta e gestão de *stocks*, confirmando assim a importância da eficácia do ciclo de compra e abastecimento das farmácias para um maior desenvolvimento da indústria. Sendo esta uma das maiores indústrias, esta carece do desenvolvimento da tecnologia da gestão da cadeia logística. Esta necessidade originou a criação de vários estudos que aplicam inteligência artificial (IA) para melhorar a previsão de vendas e *stock*. Tais estudos propõem soluções na gestão de *stock*, fornecem informações sobre as quantidades mínimas e máximas a serem mantidas e definem as quantidades que precisam ser reabastecidas com base em tendências de consumo registadas para diferentes produtos farmacêuticos. Atualmente existem esforços focados na revisão dos métodos de gestão e negociação das farmácias, para estas alcançarem

uma redução no nível médio de *stock* e consequentemente nos custos de armazenamento e espaço necessário para armazenamento [3].

Conforme o comunicado do Ministro da Saúde Manuel Pizarro (novembro de 2022) o problema da falta e medicamentos nas farmácias manifesta-se um pouco por toda a Europa. Tendo apontado para a capacidade das organizações farmacêuticas assegurarem que todos os medicamentos em falta nas farmácias têm substitutos. Estas entidades encontram-se envolvidas neste processo para criar alternativas que mitiguem o número e frequência destas faltas [4].

1.2. Objetivos

Este projeto de investigação aborda vários estudos e problemas de domínio semelhante, identificando e documentando diferentes abordagens, modelos e *workflows*, com a finalidade de avaliar e refletir sobre as soluções propostas. Em suma, o objetivo é a realização de uma investigação que aborde a solução, delineando, estabelecendo e validando uma metodologia de *forecast (*previsão*)* transversal para produtos e serviços que utilizando os dados de histórico e variáveis, projete a dimensão da procura e a distribuição dos fármacos num período pretendido determinado.

Ao avaliar o estado da arte de soluções de previsão de vendas de produtos farmacêuticos, foi realizada uma revisão sistemática da literatura. Para uma melhor definição dos objetivos e melhoria da qualidade dos resultados obtidos com este trabalho de investigação existiu o foco na utilização de trabalhos que utilizam modelos de previsão e que efetuam uma avaliação dos mesmos.

Este projeto tem como visão o domínio geral da ciência de dados para problemas de previsão, através da utilização de modelos de séries temporais, direcionados à previsão de produtos farmacêuticos. É pretendido inovar as perspetivas e resultados já alcançados internacionalmente, fazendo a investigação para auxílio de respostas, tais como:

- Quais os modelos (algoritmos) utilizados?
- Qual a variável target (variável a modelar e prever)?
- Quais as maiores limitações na previsão de vendas de produtos farmacêuticos?

CAPÍTULO 2

Revisão da Literatura

2.1. Âmbitos da problemática

A saúde é área que mais se destaca para resolução de problemas utilizando aplicações de *Machine Learning (ML)*. A indústria da saúde é constituída por diferentes serviços de saúde e entidades, sendo que estes diferentes domínios da saúde possuem em comum o problema da consistência dos dados e a genuinidade dos dados recolhidos. O método de como os dados são recolhidos, se são confiáveis e como são processados é a diretriz principal, necessária a ter em atenção para a implementação de qualquer tipo de modelo de previsão. Um cientista de dados tem tarefas desafiantes porque recolhe dados de diferentes fontes e efetua o pré-processamento da informação para a reutilização dessas informações para diferentes finalidades [5].

A rápida transformação do mercado obriga às empresas adaptarem-se insistentemente. Para tais ajustes, planear o que pode vir a acontecer é indispensável, sendo que a previsão é uma componente da ciência de dados que ajuda as empresas a prepararem-se. Uma empresa que se adapta rapidamente tem melhores oportunidades de prosperar. Ao utilizar previsões para antecipar mudanças nas suas ações, um negócio pode responder melhor às necessidades dos consumidores [6].

A precisão da previsão é um fator crítico de sucesso e continua a ser um grande desafio na gestão de processos de negócio. A IA pode ajudar as empresas e organizações a melhorar a sua previsão e reabastecimento, devido a esta capacitar diferentes a que algoritmos permitam que vários dispositivos prevejam ações, processos e tendências. Foram feitos estudos que defendem a necessidade que a tais soluções precisam de ser implementadas na cadeia logística para obter vantagens competitivas com o melhoramento de métricas. Com a previsão de compra de produtos pelos clientes, redução no custo da quantidade de *stock*, redução da superprodução, as empresas capacitam-se para possuírem uma rápida reação às mudanças no mercado e, por consequente, proporcionam aos clientes uma melhor experiência [7].

Nos tempos atuais da tecnologia moderna, o paradigma competitivo entre as empresas sofre alterações a um ritmo sem precedentes, sendo aplicadas novas medidas de sucesso para o desempenho das cadeias de distribuição, para superarem a concorrência. No entanto, tais métodos apenas podem ser obtidos e sustentados se a empresa possuir uma cadeia de distribuição eficaz e eficiente contida num ambiente adequado a técnicas de previsão. Na indústria farmacêutica, a previsão de procura é essencial para otimizar e gerir os complexos processos de negócios. Sendo que

fontes de dados sobre as vendas, operações e previsões de planeamentos colaborativos, reabastecimentos e gestão de *stock*, ajudam os fabricantes a uma previsão com precisão [18].

Previsões precisas das necessidades e exigências dos consumidores continua a ser um desafio no ambiente competitivo e dinâmico dos negócios atuais, sendo que mesmo pequenas melhorias na previsão de vendas ajudam diversos retalhistas a reduzir os custos operacionais ao mesmo tempo que melhoram as vendas e a satisfação dos clientes. Previsão de vendas em cada loja retalhista é crucial para o sucesso de todas as empresas que atuam como intermediários da cadeia de distribuição porque ajuda no controlo de inventário, resultando numa melhor distribuição de produtos em lojas. Este método também minimiza o excesso de *stock* e a rutura de *stock* em cada loja, minimizando assim as perdas, e o mais importante, maximizando as vendas e a satisfação do cliente. Devido ao alto risco e consequências envolvidas com a previsão de vendas, este torna-se um problema essencial a ser resolvido por todas as empresas de retalho. A venda de produtos pode depender de uma variedade de fatores externos, como competição, clima ou tendências sazonais e por ações internas como promoções, eventos de vendas, preços, planeamento no fornecimento de mercadorias, o que aumenta a complexidade do problema [13]. Para fornecerem aos seus clientes os cuidados de que necessitam o mais rapidamente possível, as farmácias são abastecidas por grossistas que lhes fornecem uma garantia de entrega de meio-dia para a maioria das referências de produtos. Para isso, eles definiram uma cadeia de abastecimento eficiente e complexa. Contudo, para melhorar ainda mais a eficiência dos seus serviços de entrega, algumas distribuidoras pretendem usar ferramentas de *ML* para prever pedidos futuros e antecipar as suas necessidades de *stock* [14].

Num ambiente dinâmico como o da compra e consumo de medicamentos, é considerado como essencial a necessidade das indústrias farmacêuticas de aumentar a vantagem na competitividade comercial. Existem muitos fatores ocultos que podem afetar o consumo de medicamentos. Uma maneira para prever o consumo de medicamento e da produção é através da identificação desses fatores latentes. Com o uso de técnicas de *data mining (DM)*, relações ocultas entre as variáveis podem ser identificadas. Enquanto isso, variáveis efetivas podem ser utilizadas para prever as variáveis dependentes. No entanto, o uso de *DM* requer acesso a informações de históricos sobre o uso de medicamentos, juntamente com características de compra e características dos compradores [15].

Os modelos de previsão permitem que as empresas farmacêuticas tenham sucesso no mercado global. Métodos de previsão da procura possuem como objetivo a recomendação de estratégias de vendas e marketing, apoiando-se nas tendências e efeitos sazonais e estes representam avaliações e procedimentos metódicos que são relativos à futura procura de produtos críticos [19].

Ciência de dados é uma atividade centrada no ser humano, dedicada à extração de conhecimento de dados complexos com o objetivo de criar mais *insights*. É um campo geral que abrange IA, recolha e manipulação de dados, avanços em bases de dados e infraestruturas de tecnologia de informação (TI). A IA aplica abordagens e tratamentos de matemática e algorítmicos bastante diferentes, desde a investigação operacional até à programação. Os programas de *ML*, campo da IA, são capazes de se ajustar quando expostos a novos conjuntos de dados, ou seja, "aprendem" sem serem explicitamente programados. Normalmente, estes são desenvolvidos para encontrar: padrões, tendências e associações, descobrir ineficiências, aprender e melhorar, executar planos, prever resultados futuros com base em tendências históricas e para informar decisões baseadas em factos. *ML* é geralmente categorizado em dois tipos: supervisionado e não supervisionado. A aprendizagem supervisionada adapta um modelo para reproduzir um conhecido output de um conjunto de treino enquanto a aprendizagem não supervisionada refere-se à inferência de padrões existentes de um conjunto de dados não rotulado, sem qualquer referência a variáveis de resultados ou previsões rotuladas. À medida que as infraestruturas de TI se desenvolvem em *Cloud*, várias estruturas de *ML* prontas para uso foram propostas para o número crescente de não cientistas de dados, embora o *ML* exija muitas parametrizações e experiência para fornecer informações relevantes e robustas [8].

Infelizmente, a cadeia de conhecimento e os conjuntos de dados em saúde têm falhas críticas em muitos pontos, portanto, os especialistas estão sempre a tomar decisões com base no conhecimento "menos prejudicial" e adicionam as suas próprias intuições. Nesses cenários, a inserção direta de IA, sem interoperabilidade com a intuição humana, terá consequências imprevisíveis. Face aos contextos em que a IA se encontra aplicada é necessário realizar alinhamentos, regular e determinar com sensibilidade os impactos na saúde que os parâmetros de IA provocam, sendo esta uma prática importante tendo em conta a alta velocidade a que os avanços tecnológicos acontecem [9].

Conforme a notícia publicada no jornal *on-line The Independent*, foi afirmado pelo *CEO* da *Google* que o alinhamento internacional será fundamental para fazer com que os padrões globais funcionem porque as grandes empresas não podem simplesmente construir uma nova tecnologia promissora e deixar que as forças do mercado decidam como ela será usada. Assim como defende que a IA tem o potencial de melhorar milhões de vidas, e o maior risco pode ser não o fazer. Sendo idealmente garantir o desenvolvimento de forma responsável de uma forma que beneficie a todos de modo que inspire as gerações futuras a acreditar no poder da tecnologia [10]. Por sua vez, consoante um comunicado de *Margrethe Vestager*, vice-presidente executiva da Comissão Europeia para a era digital, defende que a União Europeia (EU) lidera o desenvolvimento de novas normas globais para garantir que a IA seja confiável. Acrescenta também que a UE garante permanecer competitiva ao definir padrões que podem abrir caminhos para uma tecnologia ética em todo o mundo [11].

As autoridades de saúde e as farmácias estão cada vez mais interessadas em obter significativas reduções de custos através do planeamento e reorganização de toda a cadeia logística de medicamentos e dispositivos médicos, sem comprometer a segurança do processo e a prevenção de erros. Atualmente, a gestão de *stock* efetuada pelas farmácias locais segue, para a maioria dos medicamentos, uma lógica, invés de uma previsão. A lógica é um processo que é iniciado quando é alcançado um determinado nível de *stock*, conhecido como nível de reabastecimento, assim procede-se à emissão de um novo pedido de abastecimento. Como resultado, esta política de gestão, para além de não fornecer uma constante monitorização do processo de *stock*, implica grandes custos de armazenamento. A eficiência da distribuição também requer uma boa logística para todos os produtos farmacêuticos. A poupança que pode ser atingida através do uso de previsões, de facto, representa uma elevada contribuição para a estabilidade e sustentabilidade do sistema. Muitos benefícios podem ser obtidos a partir de uma correta gestão das atividades de distribuição, em dinheiro, tempo e pessoas [12].

O conceito de produção em massa não é apenas aplicado para empresas que fabricam produtos de consumo rápido, é também aplicado à indústria farmacêutica porque as estimativas de procura são recolhidas pelo consumo do paciente. A rede de produção da indústria farmacêutica oferece associações para a construção dos seus planos, para lidar com a concorrência e para os perigos na distribuição de *stock* de produtos farmacêuticos. A escassez de substâncias ativas dos produtos farmacêuticos pode resultar em perdas de vidas. Desta forma, os processos de previsões adequadas à procura de tais recursos e medicamentos que salvam vidas são obrigatórios conforme é indicado por autoridades reguladoras mundiais. A fidelidade do cliente é perdida quando a empresa perde a visão da procura do mercado, desta forma a previsão da procura leva à criação de um relacionamento entre os fornecedores e clientes. Para manter a oferta e a procura a nível mundial, é essencial ter um melhor modelo de previsão que aumentará a viabilidade dos produtos, portanto, quanto melhor a previsão, mais acessíveis serão os produtos. Assim sendo, a computação de previsões, como a previsão sobre procura de medicamentos por clientes, de maior precisão é uma componente fundamental para garantir um fornecimento adequado de medicamentos que salvam vidas. Como a indústria farmacêutica é volátil e a rivalidade entre as empresas aumenta, para ganhar competitividade é essencial ter mecanismo adequados para a previsão da procura de medicamentos com a máxima precisão e baixo erro de previsão. A partilha de informações entre os fornecedores e fabricantes aumenta a precisão da previsão relativa à procura pelos clientes. Na previsão de procura de medicamentos, não existem apenas fatores internos, sendo que fatores externos, ou seja, doenças sazonais e epidémicas, taxas de substâncias ativas, fatores humanos, quota do mercado de produtos competitivos e condições de *marketing*, também contribuem significativamente no enviesamento da precisão da previsão. A indústria farmacêutica tem erros de previsão mais elevados devido a fatores

externos como regulamentos, políticas, entrada de novos produtos com a mesma forma farmacêutica e mudança na tecnologia. Os farmacêuticos enfrentam situações com ruturas de *stock* e de *stock* excessivo devido a frequentes erros de previsão [18].

2.2. Investigação em ciência de dados

O processo de *DM* pode ser aplicado a conjuntos de dados de qualquer tamanho, e embora possa ser usado para descobrir padrões, no entanto, este processo não consegue descobrir padrões que ainda não estão presentes no conjunto de dados. A limpeza de dados e integração de dados para pré-processamento, ajuda a melhorar a precisão do treino e desempenho de modelos de previsão. A limpeza de dados serve para detetar e corrigir (ou remover) registos corrompidos ou imprecisos de um *dataset*, tabela ou base de dados. O processo também é aplicado à identificação de partes incompletas, incorretas, imprecisas ou irrelevantes dos dados, para de seguida, substituir, alterando ou removendo o ruído dos dados. Após a limpeza, um conjunto de dados será consistente com outros conjuntos de dados semelhantes no sistema. As inconsistências de dados detetadas ou em falta podem ter sido originalmente causadas por diferentes definições em bases de dados, efetuadas por entidades da mesma área de diferentes lojas, podem ter sido causadas por erros de input do utilizador ou os dados podem ter ficado corrompidos na transmissão, ou armazenamento. Limpeza de dados difere da validação de dados, porque neste segundo cenário de processo, existem dados que são rejeitados do sistema no momento da sua entrada. A validação de dados deve ser rigorosa, como, por exemplo, rejeitar qualquer endereço que não tenha um código postal válido, ou deverá corrigir registos que correspondem parcialmente a registos já existentes ou conhecidos. A integração de dados é o processo de combinação dos dados existentes em diferentes fontes e fornecer ao utilizador uma visão unificada desses dados. Este processo surge em diversas situações, tanto comerciais (quando duas empresas semelhantes precisam convergir as suas bases de dados) e científico (combinando resultados de pesquisas de diferentes repositórios). A integração de dados aparece com uma frequência crescente conforme o volume e a necessidade de partilhar dados existentes se multiplicam e esta referida integração representa um foco em extensivos trabalhos teóricos que apontam para numerosos problemas em aberto que ainda necessitam de resolução [16]. A falta de dados reduz a representatividade da amostra e pode distorcer as conclusões sobre a previsão. Com algumas aplicações práticas, é possível controlar o nível de dados perdidos e evitar valores ausentes. Muitas vezes não é possível passar para a deteção de *outlier* (*data point* ou registo que difere significativamente de outras observações) até que os dados ausentes não sejam tratados. A presença de *outliers* nos dados de vendas pode enviesar significativamente uma previsão. A deteção de *outliers*

é especialmente crítica para fazer previsões baseadas em dados históricos onde o tamanho das vendas é importante. Este é especialmente o caso da compra por pânico de produtos populares que estavam em falta por algum período [17].

Cross-industry standard process for data mining, conhecido como *CRISP-DM* é uma metodologia popular para aumentar o sucesso do projeto de *DM*. A metodologia define uma sequência flexível de seis fases, que permite a construção e implementação de um modelo de *DM* para ser utilizado em contexto real, auxiliando no suporte às decisões do negócio. O *CRISP-DM* define um projeto como um processo cíclico, onde diversos passos e ações podem ser utilizadas para permitir um resultado mais direcionado com os objetivos do negócio. Após a identificação do objetivo a ser atingido, realizada na primeira fase *CRISP-DM,* "Compreensão do Negócio", os dados necessitam ser analisados na seguinte fase de "Compreensão dos Dados" para de seguida serem processados durante a fase de "Preparação dos Dados". A fase de "Modelação" representa a construção e desenvolvimento do modelo ou algoritmo que, devido aos dados nesta fase já se encontrarem selecionados, limpos e preparados, permite a pesquisa e análise de úteis padrões nos dados que possuem informação e conhecimento que será aprendido pelo modelo (por exemplo, o modelo pode ser usado para prever o valor alvo que representa o objetivo definido). Depois o modelo é analisado e avaliado, em termos de desempenho e utilidade, na fase de "Avaliação". Se o modelo executado não for bom o suficiente para ser utilizado no suporte ao negócio, então uma nova iteração para o *CRISP-DM* é definida. Caso contrário, o modelo é implementado num ambiente em tempo real, concretizando a fase de disponibilização para uso pelos utilizadores da área de negócio (fase de "*Deployment"*). Um dos principais objetivos da metodologia *CRISP-DM* é a afinação e suporte a projetos de *DM*, sendo que cada iteração do *CRISP-DM* demonstra ser muito importante [20].

Modelos para previsão de vendas são ferramentas que são utilizadas pela maioria das corporações de laboratórios farmacêuticos internacionais com o duplo objetivo de igualar a oferta para quaisquer eventos de aumento ou queda na procura pelos produtos e para manter em *stock* o mínimo possível. Através de um sistema implementado eficazmente, é possível uma contribuição para uma gestão eficiente das cadeias de distribuição. Tal iria contribuir numa melhoria geral nos lucros da organização. É importante indicar que as cadeias de distribuição das empresas farmacêuticas consistem em variadas atividades complexas que por vezes estão interligadas. Sendo que às complexidades existentes somam-se distintos desafios que necessitam de ser enfrentados pelas empresas farmacêuticas de todo o mundo. A gestão eficiente da cadeia de distribuição é vital para qualquer empresa de serviços essenciais, e tal é uma realidade muito presente no caso da indústria farmacêutica. A pandemia de *Covid-19,* inicialmente considerada como uma epidemia restrita à China, tornou-se numa pandemia global. Esta pandemia é uma situação única dos últimos 100 anos e afetou a dimensão da oferta e da procura, inclusive na bem estabelecida indústria farmacêutica global. Isto devido a surtos repentinos

terem dado origem a vagas de aumento de procura, assim como os bloqueios, desenvolvimento de vacinas e medicamentos que contribuíram para a imprevisibilidade de abastecimento dos produtos farmacêuticos. Devido às dificuldades existentes nas empresas em identificar mudanças rápidas na quantidade de procura, origina a constrangimentos na entrega da quantidade requerida dos produtos dentro do prazo. A pandemia de *Covid-19* produziu quebras na cadeia de distribuição e criou ruturas de medicamentos ambulatórios. A pandemia aumentou a procura por produtos farmacêuticos e são relatas ruturas desde o início do surto [19].

Existe a necessidade, em vários estudos científicos, de fazer previsões de comportamentos futuros. A crescente disponibilidade de grandes quantidades de dados de histórico requer a definição de técnicas eficientes para deduzir a dependência estocástica entre o passado e o futuro a partir de observações. Os dados de histórico são frequentemente uma série de pontos ordenados no tempo. Em séries temporais o tempo será frequentemente uma variável independente, sendo que objetivo é realizar uma previsão para o futuro. Uma série temporal é uma sequência S de medições históricas yt de uma variável y em intervalos iguais de tempo t. Métodos clássicos como modelos *Autoregressive integrated moving average* (*ARIMA*) são muito utilizados na previsão de séries temporais. Tais abordagens, normalmente focadas em relações lineares, possuem complexidade, mas efetivamente funcionam para uma ampla diversidade de problemas, desde que os dados sejam devidamente preparados e os parâmetros bem configurados. As previsões de séries temporais podem ser consideradas problemas de aprendizagem supervisionada. Portanto, além da previsão baseada algoritmos de *ML*, a análise de séries temporais fornece uma alternativa na previsão de compras. As redes neurais de *deep learning* evoluíram nos últimos anos e atualmente são uma ferramenta valiosa em tarefas de previsão [6].

Os modelos *ARIMA* são modelos de previsão frequentemente utilizados devido à simplicidade e capacidade de generalização para séries temporais não estacionárias. Utilizado como métrica de desempenho o *Root mean squared error* (*RMSE*) é definido como a raiz quadrada do erro médio da soma das diferenças dos quadrados entre as previsões de valores de vendas e valores de vendas reais. O objetivo é minimizar o valor *RMSE* em modelos de previsão. Os dados de vendas semanais de produtos farmacêuticos possibilitam prever a procura por tipos de medicamentos específicos ou produtos específicos. As tendências das vendas, ao serem apresentadas através de gráficos fornecem uma imagem clara da natureza dos dados obtidos. Os modelos e previsão de procura possuem como objetivos principais, o alcance das metas de produção planeadas e estabilização da produção face à procura [19].

Artificial neural network (*ANN*) é um sistema de processamento de informações que é semelhante ao biológico do sistema nervoso ou ao funcionamento do cérebro humano. *ANN* são modelos para reconhecer padrões nos dados. Um algoritmo frequentemente usado é o *Backpropagation ANN* porque este algoritmo pode ser usado para obter um output mais preciso. *ANN*, como uma técnica que simula o processo de aprendizagem de redes neurais biológicas tornou-se uma ferramenta de previsão popular. *ANN* é uma estrutura de muitos neurónios conectados que são arranjados em camadas em maneiras sistemáticas. As conexões entre os neurónios têm pesos associados a eles, dependendo da quantidade de influência que um neurónio exerce sobre outro. Existem algumas vantagens em usar redes neurais em alguns problemas. Por exemplo, devido a conterem muitos neurónios e também pelo peso atribuído a cada conexão, redes neurais são bastante robustas em relação a conjuntos de dados com ruído e erros [6].

Modelos univariados de séries temporais, como *Seasonal auto-Regressive integrated moving average* (*SARIMA)* e *Prophet* fornecem previsões satisfatórias e acertadas, sendo que no contexto de previsão de vendas os modelos univariados apenas utilizam dados que apenas contêm as datas de venda e a variável com o número de quantidades vendidas. Em contraste, modelos como regressão linear, *Random forest* (floresta aleatória) e *XGBoost* são métodos de aprendizagem supervisionada que dependem fortemente em recursos mais ricos, em muitas características e variáveis, para obter previsões ideais [6]. O *XGBoost* é usado para problemas de aprendizagem supervisionada, onde são usados dados de treino x para prever uma variável de destino y. O termo de regularização controla a complexidade do modelo, o que ajuda a evitar erros de *overfitting* (este tipo de erro ocorre em modelos quando este ajusta-se muito bem ao conjunto de dados de histórico, mas demonstra ineficácia para prever novos resultados) [13].

O modelo *Prophet* devido a ser aplicado sem a definição de hiperparâmetros (atributos ou configurações definidas antes do treino do modelo, que controlam e determinam a aprendizagem efetuada por este), demonstra interesse no contexto de modelação para a previsão afinada e ajustada por produto. A aplicação deste modelo é mais simples, enquanto o desempenho do algortimo *SARIMA* depende principalmente de hiperparâmetros. Independentemente do método escolhido, a definição de hiperparâmetros é um processo que pode demorar muito tempo. As previsões diárias são sensíveis às mudanças de data. Uma pequena mudança pode aumentar significativamente o erro de uma previsão. Os totais semanais não possuem tanto flutuação como as vendas diárias. É possível existirem situações reais em que um cliente tenha registado erradamente a quantidade no pedido, sendo que irá ajustar as quantidades no próximo pedido no dia seguinte. Desta forma, as previsões com a soma em vários dias podem evitar erros aleatórios e disponibilizar uma melhor previsão [6].

CAPÍTULO 3

Metodologia

3.1. Contexto da metodologia

No capítulo da introdução encontra-se o contexto atual abordando problemas e constrangimentos reais e atuais, em Portugal e na Europa. O contexto científico foi demonstrado no segundo capítulo deste documento, retratando e analisando os factos comprovados cientificamente, sendo este uma base importante para a transformação social e tecnológica porque consolida o conhecimento e desafia as estruturas atuais que permanecem apoiadas em métodos que não respondem à altura das necessidades do panorama nacional e internacional. Ao ser possível delimitar minimamente o objeto de estudo e garantir a disponibilidade de informação e dos dados, encontram-se reunidas as condições para dar início ao projeto de investigação.

O domínio do problema integrado no capítulo de revisão de literatura desenvolvido foi se particularizando através dos estudos lidos e pesquisados durante as primeiras etapas da investigação. As etapas iniciais do presente estudo não continham ainda os problemas definidos e por esse motivo determinados avanços mais concisos foram apenas alcançados em etapas posteriores da investigação e após efetivado o começo do projeto de previsão em séries temporais. A definição concreta do problema de investigação ocorreu depois da análise dos dados existentes para a base das previsões e somente após foi delineada a solução para o problema exposto neste documento. Os passos realizados como a análise da qualidade, o período dos dados e verificação das variáveis a utilizar, ocorreram na fase de compreensão dos dados, sendo que este percurso passou também pela exploração e preparação dos dados e seleção dos modelos de previsão. Após os progressos concretizados e referidos anteriormente foi possível especificar as pesquisas para a exploração e problematização do projeto. A “Figura 1” sintetiza o processo inicial da investigação para a pesquisa de artigos científicos.

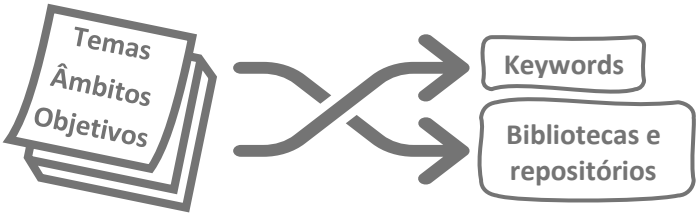

Figura 1 – Início do processo da pesquisa científica.

Para ser possível ultrapassar dificuldades foi essencial dividir o projeto de investigação a realizar, separando o percurso e as tarefas em duas vertentes. A primeira, a vertente do problema teórico e a segunda, a vertente mais técnica focada no problema dos dados e no caso prático. Este método permitiu que existisse diariamente avanços no projeto de investigação. Inicialmente na vertente do problema teórico foram realizados avanços através do desenvolvimento da revisão de literatura. Na primeira etapa da investigação teórica foram constantemente definidos quais os próximos passos a seguir. Como exemplo dos passos realizados que serviram de guia para uma melhor tomada de decisões na vertente teórica, destacam-se os seguintes:

- Pesquisa e leitura de artigos científicos e dissertações específicas e relacionadas com a investigação, pesquisando por *keywords* e termos definidos e associados ao tema;
- Seleção e arquivo de 38 referências pesquisadas, sendo que um dos pontos da seleção é a validação da relevância para este estudo, analisando a introdução dos artigos e pesquisa de subtemas e métodos contidos no mesmo (estes últimos são, por exemplo, nomes de fases do *CRISP-DM* e métodos como "*Model*" ou "*Evaluation*", mais abrangentes devido a incertezas existentes nas primeiras etapas relativas aos dados e modelos de previsão a utilizar);
- Realização de pontuação dos 38 artigos analisados em três categorias definidas: "Problemas, modelos, variáveis e target", "*CRISP-DM*, metodologias e técnicas" e "Contextualizações, introduções e estados de arte";
- Na sequência da pontuação dos 38 artigos, foi feita a soma e comparação das características dos mesmos e estes foram arquivados separadamente em 3 pastas, de forma a organizar e dividir estes pela avaliação dos mais interessantes e menos interessantes;
- Análise da informação da empresa como missão e valores e toda a proposta da temática indicada pelo ISCTE e objetivos indicados pela empresa;
- Leitura de livros sobre *Python* e Séries Temporais e de livro para auxílio e facilitação na elaboração de teses, dissertações e relatórios de investigação;
- Pesquisa em mais repositórios científicos, bibliotecas e jornais *on-line* e leitura de novas notícias, conferências e artigos;
- Interlocução entre o autor e a professora orientadora durante todo o período do estudo, foi importante para chegar às seguintes fases e conclusão do projeto e procurar otimizar a qualidade da investigação;
- Para familiarização do projeto prático em *Python* foram estudados projetos de trabalhos, aulas, exercícios, cursos e vídeos de origem nas Unidades Curriculares do Mestrado em Ciência de Dados e disponíveis através de comunidades online.

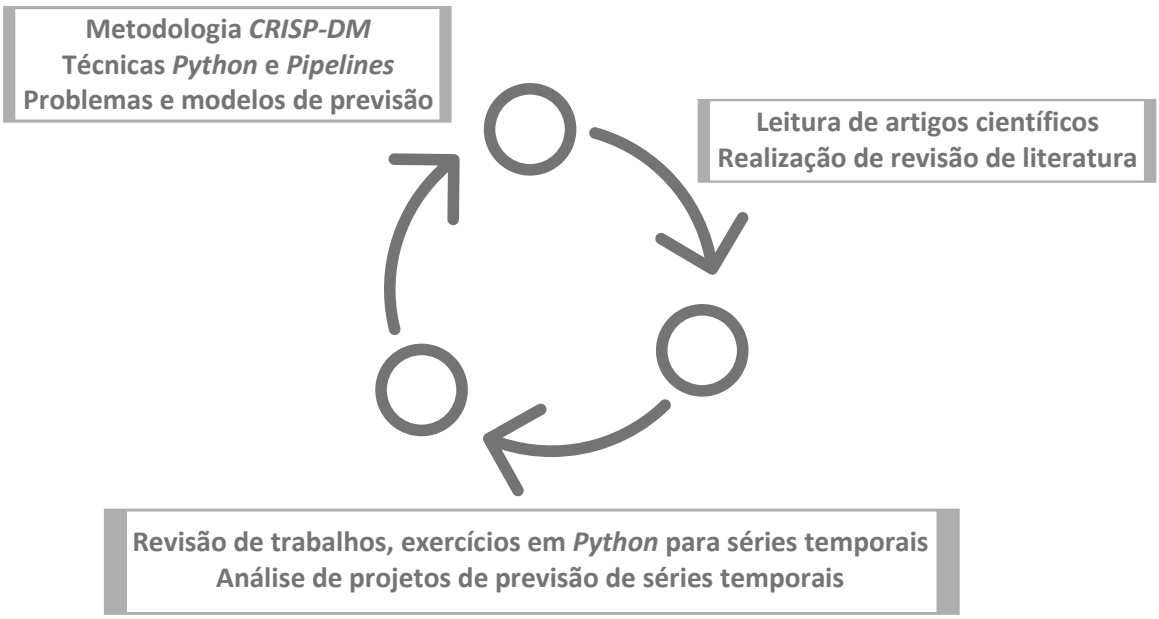

Figura 2 – Análises essências para progresso da investigação.

Conforme o esquema da "Figura 2" foram realizadas três atividades distintas para o avanço da investigação.

Após percorridas as primeiras etapas conceptuais, é feita a transição para a segunda etapa que se apoia na metodologia *CRISP-DM* e que se inicia através da familiarização com a base de dados no seu contexto para a sua caracterização, definição dos objetivos mais concretos e começo da preparação da base de dados para o avanço das análises e aplicação dos métodos de previsão. Na "Figura 3" encontram-se as fases constituintes do *CRISP-DM*.

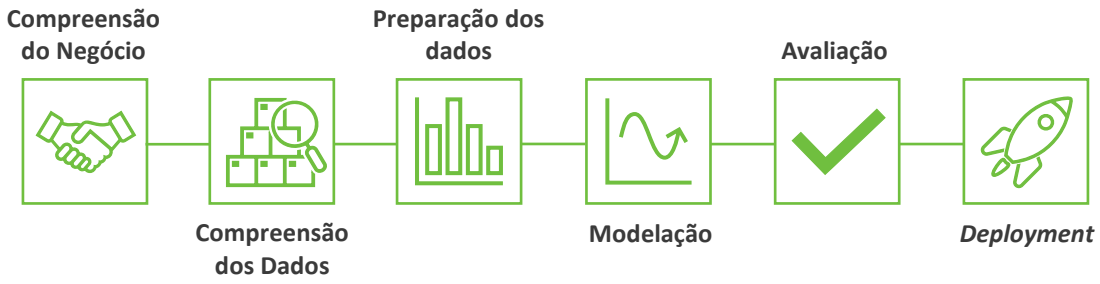

Figura 3 – As diferentes fases da metodologia CRISP-DM.

3.2. *CRISP-DM* Fase 1: Compreensão do Negócio

3.2.1. Contexto

Após a análise de propostas de temas de dissertação e apresentações de projetos de investigação com empresas foi demonstrado interessante, pelo autor deste documento, no desenvolvimento do presente tema na área de negócio da saúde. A proposta aborda problemáticas atuais existentes em Portugal relativamente à gestão de *stocks* de medicamentos e foi possibilitada para alunos do mestrado de Ciência de Dados, pelo Departamento de Métodos Quantitativos para a Gestão e Economia (DMQGE) e comunicada através da professora Diana Mendes. Após o acordo de colaboração entre a escola, a empresa e o discente foi iniciada a análise e definição dos objetivos e características dos dados. Este projeto foi desenvolvido em colaboração com a empresa que disponibilizou os dados. A missão e valores da empresa consistem em desenvolver soluções utilizando tecnologia de saúde, oferecendo resposta e apoio especializado para serem alcançadas melhorias através de análises, informações, vendas e *stocks*.

3.2.2. Objetivos

O objetivo geral é poder contribuir para a otimização do *stock* e das compras de produtos farmacêuticos para ajudar as farmácias e para este ser atingido será efetuada a criação de um modelo de previsão. Estando o objetivo definido à partida é ainda necessário explorar todos os dados para definir a seleção de quais serão utilizados nos modelos de previsão. Para as empresas terem sucesso prolongado têm de saber gerir o *stock* e a previsão de vendas, porque caso contrário, podem perder clientes e lucros obtidos. Na sequência deste pressuposto, é necessário aceder aos dados de histórico de vendas, realizando uma análise à base de dados, para definição do problema em concreto e para ser efetuada posteriormente a preparação do *dataset* a utilizar no projeto. O referido *dataset* tem como função ser importado para um *dataframe* em linguagem *Python*. O problema passa a ficar definido e intitulado de "Previsão de vendas para apoio em gestão de farmácias". Este passo tem uma grande importância e representa uma etapa crítica para a conclusão deste documento e foi alcançado pelo cruzamento dos objetivos do autor com os objetivos da empresa que disponibiliza os dados. Foram também assinalados os valores finais pretendidos de obtenção e o seu impacto de melhoria na diminuição de ruturas de *stock* de medicamentos, assim como foram analisados potenciais problemas e condicionamentos. Os objetivos do autor para o projeto de investigação e projeto técnico foram formados pela prática profissional em bases de dados de área de negócio da saúde e de medicamentos

e delineados durante o período de um ano da realização da dissertação de mestrado que engloba as várias vertentes da investigação e trabalho prático em previsões de séries temporais.

3.2.3. Intervenientes e tecnologias

Foi realizada pela empresa, uma contextualização teórica da área de negócio e uma introdução técnica ao acesso e visualização das tabelas da base de dados.

Os *softwares* presentes e utilizados na realização do projeto são os seguintes:

- Uma ligação *VPN* e uma aplicação universal de gestão de base de dados;
- Utilização das linguagens de programação *SQL* e *Python*;
- Criação de documentos *Notebook* interativos com linguagem de programação *Python* e organizadas, e explicadas em texto com formato *Markdown*;
- Ferramentas para a criação de documentos de texto para escrita da dissertação de mestrado.

3.3. *CRISP-DM* Fase 2: Compreensão dos Dados

Os principais objetivos da presente fase *CRISP-DM* é a rigorosa análise do *schema* da base de dados e definição do problema. É necessário definir o problema para definir o *dataset*. Este *dataset* será criado através da exportação dos resultados de uma *query*, tendo esta a funcionalidade de selecionar e filtrar os dados.

Esta fase do projeto foi efetuada inicialmente com a linguagem *SQL* para visualização, seleção e exportação dos dados para um ficheiro *CSV*. A etapa de acesso e a familiarização dos dados foi inicialmente realizada pela empresa através de uma passagem de conhecimento e contextualização dos mesmos. Nesta procedeu-se à análise das poucas dezenas de tabelas existentes no *schema*, estas, contudo são compostas por uma vasta quantidade de colunas e na sua maioria apresentam uma boa consistência relativamente à quantidade de linhas de registos com valores presentes para o tipo de negócio em causa. Na análise referida a ordenação e a organização da totalidade das tabelas demonstrou-se ser um método importante devido à seleção de grupos de dados, sendo definidos sete grupos em que cada grupo é uma categoria dos dados existentes nas tabelas. Foi também indicado o grau de importância, curiosidades, oportunidades e lógicas de ligações da maioria das tabelas e ainda explicada de forma ligeira o significado e contexto de algumas das suas colunas.

Devido ao tamanho e detalhes dos dados e por não existir possibilidade de observação dos modelos relacionais para as tabelas em causa, foi necessário iniciar novas análises individuais para as mesmas. Na “Figura 4” a seguir apresentada no documento, são demonstrados os diversos métodos utilizados para análise das tabelas da base de dados e seleção do conteúdo. No exemplo de uma tabela com 200,000 linhas em que são detetadas colunas que possuem um pequeno número de valores

distintos como, por exemplo, categorias de medicamentos, tipo de produto ou colunas binárias de Sim, ou Não, para esta foram também apontados os diferentes valores contidos para fácil visualização das possibilidades dos valores dessas colunas. Através de técnicas em *SQL* foi simplificada a visualização de todo o universo do *schema* da base de dados da empresa.

Figura 4 – As duas etapas de análises em SQL na base de dados.

Após as análises e apontamentos realizados nos métodos efetuados e apresentados na “Figura 4” conclui-se que as tabelas mais importantes para este estudo são as dos Produtos e das Vendas. Encontram-se também apresentadas as análises efetuadas posteriormente em *SQL* no *schema* da base de dados. Completando o resumo da segunda tabela infra, sobre os métodos aplicados em *SQL* para a definição do *dataset* final a ser preparado e modelado em *Python*, foram utilizadas diversas condições e seleções para estudar os dados. Relativamente à referida seleção e exclusão de produtos, esta tem impacto na apresentação dos dados realizada nesta fase que é constituída pelo universo total de produtos farmacêuticos. Este universo de produtos são embalagens com vendas porque foi efetuado o cruzamento entre produtos e vendas e criada a condição para produtos do tipo mercadoria. Foi finalizada a construção da *query* do universo total de embalagens com venda, sendo que para esta não se encontram incluídos produtos denominados "Produto desconhecido" ou os produtos sacos de papel e plástico, e excluídos os identificadores (ids) de embalagens que possuem várias linhas no catálogo de produtos com a mesma informação, mas com múltiplos ids.

O total de vendas não é o mesmo que o número de unidades de embalagens vendidas, porque uma venda pode incluir mais que uma unidade. O total de vendas é apresentada na respetiva coluna da *query SQL* utilizando a função de agregação *count* e o comando *groupy by* para agregar por embalagem ou por dia e embalagem. Tendo assim respetivamente para cada embalagem o total de vendas para todos os dias ou o total de vendas por dia. Com esta abordagem foi possível transformar a *query* do universo total para apresentar os seguintes conjuntos de dados:

- O número de dias em que ocorreu venda, para cada produto do universo total ou para um produto específico;
- O total de vendas realizadas no total de dias, para cada produto do universo total ou para um produto específico;
- O total de vendas diárias para todos os dias, para cada produto do universo total ou produto específico.

O número de unidades de embalagens vendidas é o número de todas as unidades de embalagens/caixas de medicamentos vendidas. A soma de unidades vendidas é apresentada na respetiva coluna da *query SQL* utilizando a função de agregação *sum* e o comando *group by* para agregar por embalagem ou por dia e embalagem. Tendo assim respetivamente para cada embalagem a soma das unidades vendidas para todos os dias ou a soma de unidades vendidas por dia. Com esta abordagem foi possível transformar a *query* do universo total para apresentar os seguintes conjuntos de dados:

- O número total de unidades vendidas no total dos dias, para cada produto do universo total ou para um produto específico;
- O número total de unidades vendidas por dia para todos os dias, para cada produto do universo total ou produto específico.

Para possibilitar o início da escolha final dos campos e dos conjuntos de dados a exportar para a continuação da análise dos dados em *Python* foram realizadas as seguintes ordenações na *query* universal:

- Apresentação dos produtos com vendas em mais dias;
- Apresentação dos produtos que tiveram o maior número de unidades vendidas no total dos dias.

Após o acesso e análise dos dados foi possível analisar o período de vendas para os medicamentos e embalagens, sendo que este passo foi essencial para definir concretamente os modelos de previsão a serem utilizados. A frequência das linhas transacionais das vendas é diária e o período pouco superior a dois anos e foi identificado na sequência dos dias das vendas a falta de seis dias. Quatro desses dias são feriados e os outros dois são dias com falha ou falta de registo dos dados. Para as muitas colunas existentes com dados relativos às vendas de produtos farmacêuticos foi realizada a seleção das mais importantes para o problema definido de previsão de unidades vendidas, assim como foram escolhidas as ligações entre tabelas e as funções para definir como são calculados e consequentemente apresentados os dados das colunas. No caso deste projeto em concreto e adequando à qualidade dos dados registados na base de dados para as vendas de produtos farmacêuticos, foi tomada a decisão de obter o total diário de unidades vendidas, através da soma de todas as vendas diárias para cada produto selecionado, com o intuito de melhorar a qualidade dos dados do *dataset* a ser exportado. Esta questão tornou-se evidente por existirem originalmente transações de venda de embalagens com unidades negativas (inferiores a zero), relativas a devoluções e correções para transações registadas (como uma ou mais unidades vendidas) no mesmo dia ou dias anteriores. Devido a ser realizada a soma dos valores de unidades para todas as transações de vendas diárias, é obtido o número positivo (que é um número inteiro) de embalagens efetivamente vendidas, sendo esta técnica uma solução adaptada para a existência do referido comportamento nos dados originais.

Na sequência das análises e *queries* construídas foi escolhida uma embalagem específica, denominada neste documento por "Produto Um". Conforme a “Figura 5” foi para este produto, realizada a exportação do *dataset* para o arranque da análise dos dados em *Python*, devido a este

possuir um elevado volume de vendas e estas possuírem uma equilibrada distribuição por todos os dias do período diário das vendas da base de dados.

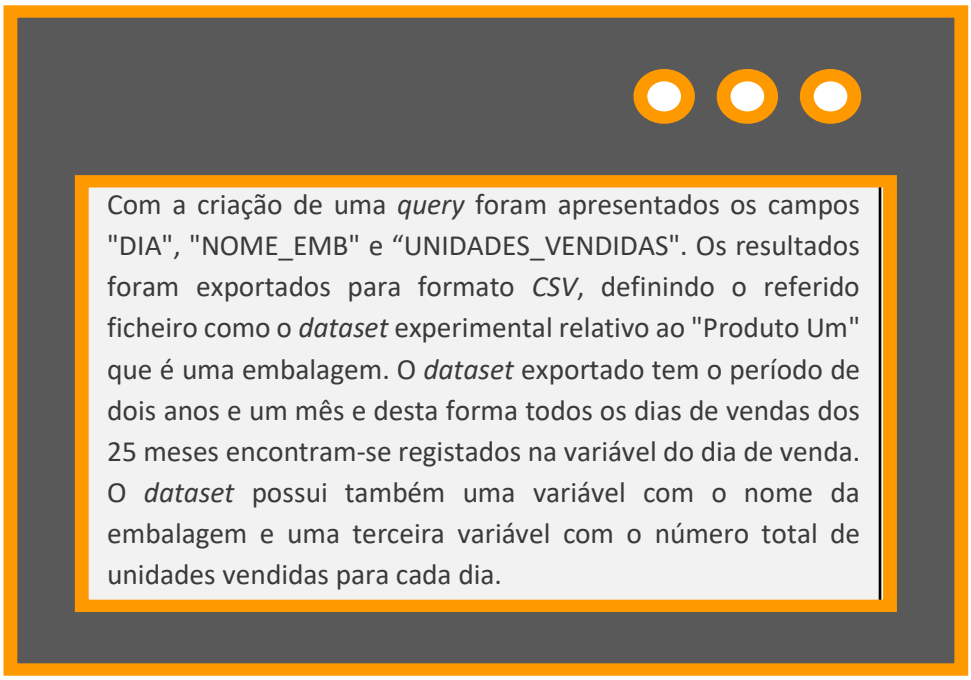

Figura 5 – Criação do dataset de teste para análise e manipulação de dados.

A exportação do conjunto de dados final foi efetuada após definidos pelo autor do projeto e pela empresa os medicamentos que necessitam de soluções com maior prioridade para mitigação de ruturas de stock, ou seja, *face* às necessidades reais e mais críticas das farmácias nacionais. Foi utilizado o *dataset* relativo ao "Produto Um" para importação em *Python*.

Entrando no ambiente de programação *Python* do *Jupyter Notebook*, foram realizadas as seguintes análises e validadas as seguintes questões:

- Importação do *CSV* para *dataframe Python*;
- Reconhecimento da variável "DIA" com o tipo de dados Data;
- Indexação da coluna "DIA" no *dataframe*;
- Visualização de informação, contagem, tipo de variáveis, primeiras e últimas linhas;
- Criação de *date_range* entre o primeiro e último dia do *dataset* para verificação de dias em falta no *dataframe*;
- Tratamento para preencher os registos dos dias em falta;
- Visualização do total de vendas diárias para todo o período;
- Visualização da distribuição e *outliers*;
- Teste de métodos de identificação de *outliers* e tratamento dos mesmos;

- Expansão do *dataframe* introduzindo variáveis com o dia da semana e o nome do mês;
- Visualização do total de vendas diárias para períodos mais curtos definidos;
- Visualização dos valores das variáveis "DIA" e "UNIDADES_VENDIDAS" com a reamostragem (utilizando o método de *Pandas resample*) com a média, valores máximos e soma nas várias frequências: Anual, Quadrimestral, Mensal e Semanal (gráfico presente na "Figura 20" do Anexo A);
- Visualização do *resample* com a média comparando entre as várias frequências;
- Decomposição da série temporal para visualização das componentes da série temporal Tendência, Sazonalidade e Ruído para melhor entendimento da estrutura básica e facilitação para a fase de modelação.

Na "Figura 19" do Anexo A, encontra-se o gráfico das vendas diárias do "Produto 1", sendo este o produto utilizado para as experiências de preparação de dados e previsões iniciais com vários modelos de previsão para séries temporais.

Em resumo, foram efetuadas manipulações de dados recorrendo maioritariamente às bibliotecas *Pandas* e *NumPy* e foram visualizados vários e diferentes gráficos tendo sido utilizadas as bibliotecas *Matplotlib* e *Seaborn*. Para a decomposição pelas componentes da série temporal foi utilizada a biblioteca *Statsmodels*. Em linguagem de programação *Python* um objeto do tipo "série temporal" possui um *índice* que são os registos da variável "DIA" e uma variável com o número total das embalagens vendidas. Os conceitos e técnicas descritos anteriormente sobre o âmbito de visualização e manipulação de dados em *Python* apresentam-se neste documento na fase seguinte do CRISP-DM (Fase 3), cuja é a fase da preparação dos dados.

Um dos objetivos nesta etapa é a validação da abordagem escolhida na exportação dos dados e atual *dataset*, verificando se se encontram reunidos os requisitos essenciais para o trabalho em *Python* através de métodos, processos e modelos de previsão para séries temporais. Paralelamente à atividade de exploração e compreensão da base de dados e tratamento dos dados exportados em *Python* foram estudados os modelos com capacidade de entregar as soluções de previsão de vendas de medicamentos. É uma mais-valia na presente fase de compreensão dos dados, saber os modelos de previsão a serem utilizados e nesta etapa foi validada a possibilidade de utilização de determinados modelos e definido o problema prático e técnico. Desta forma o problema ficou definido como a previsão de vendas de medicamentos fazendo o uso dos modelos de séries temporais *ARIMA*, *SARIMA* e *Prophet*. Foram escolhidos os modelos de *ARIMA* e modelo *Prophet*, sendo que o *ARIMA* e *SARIMA* são uma abordagem clássica destes problemas e o *Prophet* apresenta simplicidade e rapidez na preparação dos dados e execução da previsão. O *Prophet* é conhecido por ser resiliente e engenhoso

ao lidar com valores em falta e *outliers*, sendo que os resultados preditos e avaliações ao desempenho do modelo serão ainda melhores com uma rigorosa preparação de dados e parametrização do modelo.

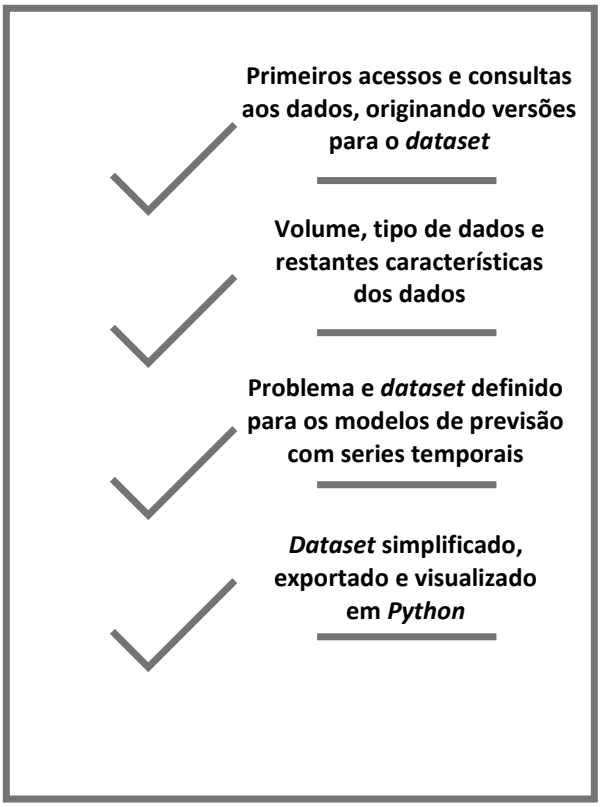

Figura 6 – Análises concluídas na finalização da fase 1 do CRISP-DM.

Como resultado da troca de ideias e escolhas entre a empresa e o autor para escolha dos produtos contidos no *dataset* final para a sua previsão foram escolhidos 13 medicamentos. O *dataset* exportado do *SQL* para estes 13 medicamentos possui a coluna "DIA", "NOME_MED" e "UNIDADES_VENDIDAS". Foram exportados os dados com a soma das vendas de todas as embalagens de cada medicamento. Foi feita a escolha de seleção conjunta dos 13 medicamentos englobando as vendas de qualquer uma das embalagens dos mesmos porque estes medicamentos têm menos vendas que a embalagem do Produto A e possuem vendas para diferentes embalagens. Desta forma o nosso *dataset* é mais rigoroso tendo menos dias em falta e responde diretamente ao problema existente de ruturas de várias ou todas as embalagens destes medicamentos. A *query* para seleção e junção dos 13 medicamentos em causa teve a utilização da *query* universal e nesta para a seleção dos medicamentos pretendidos foi adicionada uma condição referente ao preço de venda das suas embalagens e foram utilizados agregações e ordenações. Os objetivos concretizados na presente fase apresentam-se na “Figura 6”.

3.4. *CRISP-DM* Fase 3: Preparação dos Dados

A fase de preparação dos dados é muito importante para um cientista de dados porque serve de união dos dados originais às aplicações de modelos. É a fase de entrada para muitas análises e experiências feitas aos dados, nesta existe uma análise ampliada e pormenorizada do universo dos dados e exposição de inúmeros comportamentos e curiosidades e é ainda nesta fase que é aplicado a Análise Exploratório de Dados, conhecida como *EDA*. Esta aplicação possibilita a exploração e familiarização dos dados do *dataset* e como esta etapa é realizada em *Python* torna-se muito facilitada a tarefa de manipulação de dados, sendo que é muito mais rápido utilizar funções da biblioteca *Pandas* face aos diferentes métodos mais usados e necessários para realizar manipulações e consultas em *SQL*. A *EDA* para além de ser obrigatória para o funcionamento em modelos da fase *CRISP-DM* seguinte, é um guia porque as análises realizadas e curiosidades validadas feitas no percurso que antecede a modelação, impulsionam e capacitam a aplicação dos dados para diferentes soluções e oportunidades. Para a descrição da *EDA* o *dataset* utilizado é o segundo que engloba os dados diários das vendas para 13 medicamentos diferentes.

Os passos que de seguida se descrevem foram aplicados individualmente a todos os medicamentos existentes no segundo *dataset*. Para a partida do primeiro passo essencial a ser elaborado em *Python* para a análise do *dataset* que contém os 13 medicamentos, foi criado um ambiente Anaconda para instalação das bibliotecas necessárias para todo o projeto que inclui a análise dos dados, visualização, manipulação, modelação e avaliação. Estando o ambiente preparado é garantida a correta importação e execução das vastas funções a serem executados nos scripts dos *Notebooks* através do *Jupyter Notebook*. As primeiras bibliotecas *Python* a serem importadas são: *Pandas*, *Matplotlib*, *Itertools*, *Datetime*, *Numpy*, *Seaborn* e *Statsmodels* de seguida é realizada a importação do *CSV* para um *dataframe*, indicando a variável "DIA" como sendo do tipo data, para ser interpretada e visualizada com formato de data correto. De seguida foi atribuído ao *dataframe* a coluna "DIA" como índice, para, por exemplo, ser possível aceder aos dados a partir de intervalos de data indicados contidos no período total do *dataset* ou para permitir a filtragem por dia, mês ou ano. Para a segunda coluna do *dataset* "NOME_MED" foram apresentados os valores únicos que são os 13 nomes dos medicamentos e foi redefinido o atual *dataframe* para conter apenas as vendas diárias e respetivos dias para o primeiro dos 13 medicamentos denominado “Produto A”. São também apresentadas informações do *dataframe* como as linhas com os registos, o resumo do índice, comprimento e tipos das variáveis, e corrigidos nomes de variáveis para o seu uso mais simples na fase de preparação dos dados, sendo que neste documento vão ser utilizados os nomes anteriormente indicados.

Após a visualização em formato tabela dos dias e número de vendas presentes no *dataframe*, foi utilizado o método *plot*() de *Pandas* para apresentação de um gráfico simples, onde o eixo X é o período total de vendas do "Produto A" e o eixo Y é o número de vendas diárias, com a frequência original do *dataset* que é a diária. O gráfico referido encontra-se na “Figura 7”.

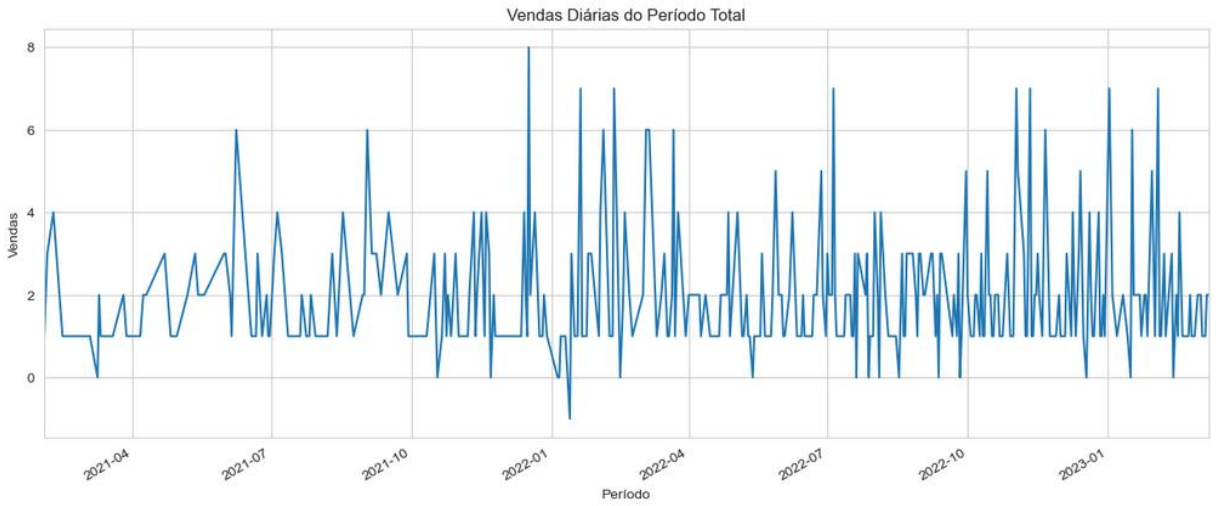

Figura 7 – Gráfico das vendas diárias do “Produto A”.

Para verificação dos dias em falta foi criada uma nova variável, externa ao *dataframe*, que através do método *date_range*() contém a sequência de todos os dias entre o primeiro e último existente no *dataframe*. Com o método *difference*() foram apresentados os dias e o respetivo número de total que se encontram em falta na sequência do *dataframe* (para o período entre o primeiro e último dia de venda para o "Produto A") e para o seu tratamento foi efetuado o preenchimento dos dias em falta através do método *asfreq*('D', *fill_value*=0) para que em todos os dias em falta sejam inseridos os respetivos dias com o número de vendas a zero para o medicamento em causa. Neste passo o *dataframe* já não apresenta dias em falta, contudo pelo analisado anteriormente diretamente nas vendas da base de dados, existem seis dias em falta que afetam todos os produtos. Estes seis dias vão ser tratados, atualizando o valor atual de zero para o número de vendas existente no dia anterior ao mesmo. Para tal foi apresentado para cada, o resultado da sua linha anterior, fazendo o uso da técnica *loc* e indicado o intervalo de dias pretendido. De seguida apenas é necessário atualizar o valor verificado, selecionando o dia e coluna pretendida para inserir número de vendas verificado. Para casos de produtos farmacêuticos que possuam vendas em todos os dias, exceto para os dias de feriado ou dias sem registo na base de dados, poderá ser apenas usada a técnica *asfreq*('D', *method='pad'*) para completar os dias em falta contendo o mesmo número de vendas diárias que o registo do dia

anterior. Por último para verificação do número de dias existentes no *dataframe* foi usado o método *count*().

Na sequência do tratamento de dias, é explicado que devido a cada um dos medicamentos do segundo *dataset*, não ter muitos picos de vendas e possuírem muitos dias sem vendas foi inserido o número de dias existente do dia anterior em cada um dos seis dias de vendas que existem em falta para todos os medicamentos, devido a serem constituídos por dias de feriado ou dias com falha de registo na base de dados. O tratamento dos dias em falta vai ter influência no passo posterior de tratamento de *outliers*, secção onde se encontra apresentado o motivo da escolha do método combinado descrito para o preenchimento dos dias em falta para cada série temporal de vendas dos medicamentos.

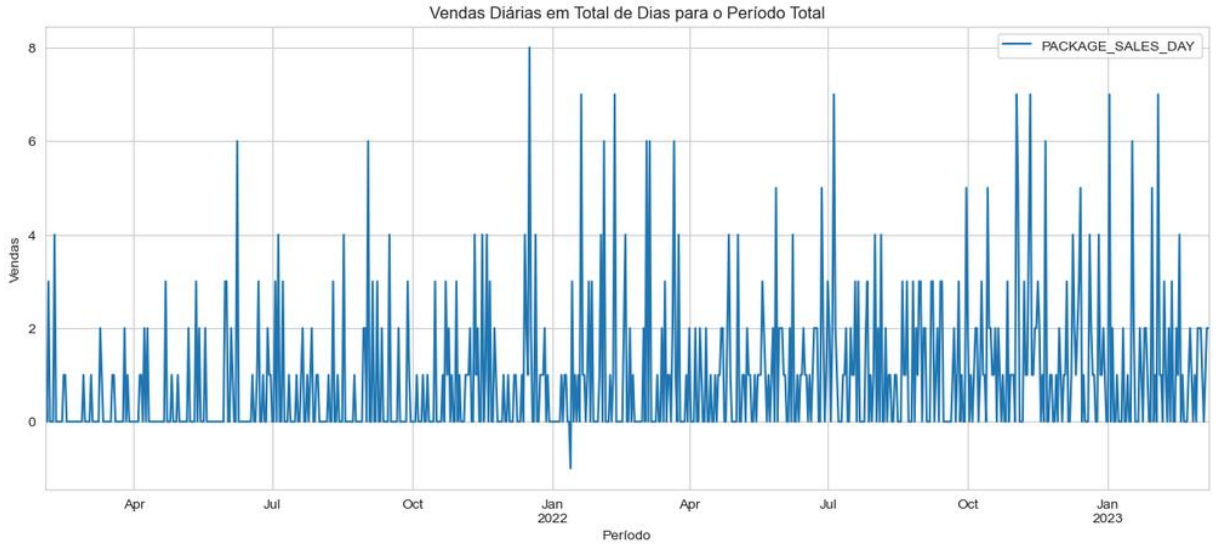

Figura 8 – Gráfico das vendas diárias do "Produto A" após tratamento de dias em falta.

A criação do gráfico das vendas de unidades para o período total de vendas, presente na "Figura 8", teve o propósito de verificar o comportamento e distribuição das vendas com maior simplicidade. Para uma análise com maior detalhe, foram acrescentadas duas linhas ao gráfico de vendas diárias que demonstram duas conversões para a série temporal em causa, uma com a média de vendas semanal e outra linha com a média de vendas mensal. Este gráfico é apresentado na "Figura 9". A técnica usada origina a visualização do *resample* com a média de vendas que permite a visualização em conjunto e comparação do total de vendas diárias, da média semanal de vendas e da média mensal de vendas. Foram também criados gráficos semelhantes, com as vendas diárias e inclusão das vendas semanais e mensais com o uso do método *resample*, para períodos mais curtos. Através da limitação do período do *dataframe* para apresentação de apenas um ano para os anos 2021, 2022 e 2023, assim como gráficos para visualização de cada semestre deste ano e ainda gráficos de apenas 3 meses para

visualizar os efeitos sazonais na venda de medicamentos que ocorrem durante o período do verão e durante o período de dezembro e transição de ano. Foram também realizados gráficos e verificados os dados através do *Python*, em formato de tabela, dos valores das variáveis "DIA" e "UNIDADES_VENDIDAS" com o *resample* pelos cálculos de média, de valores máximos e pela soma de valores. As frequências associadas e selecionadas como as melhores para análise da série temporal em causa são as semanais e mensais, por outro lado, as frequências quadrimestral e anual não acrescentaram curiosidades oportunas para aprofundamento da exploração dos dados.

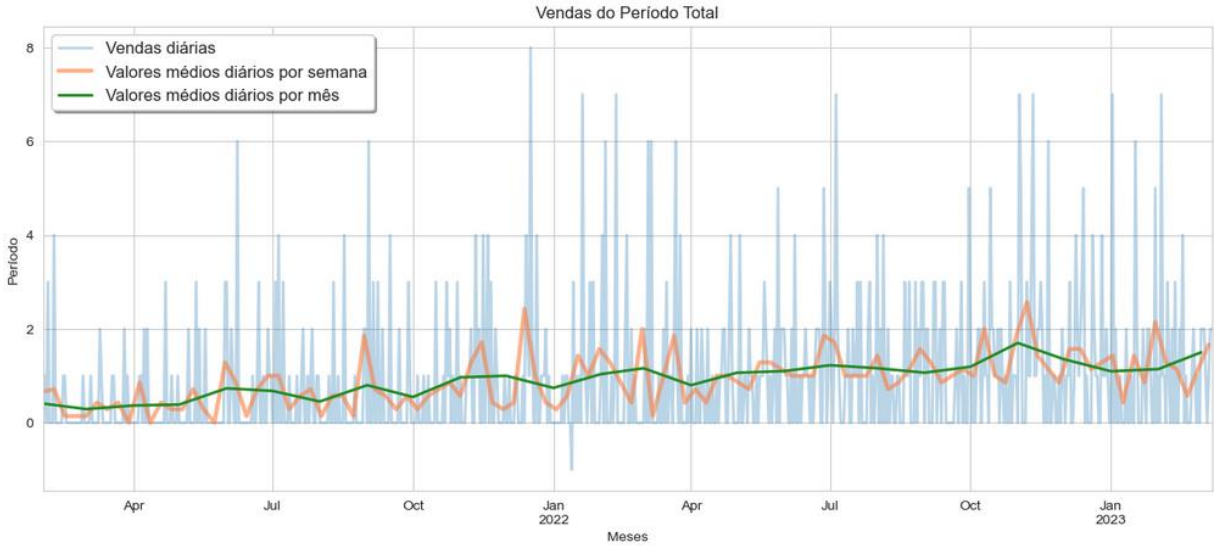

Figura 9 – Gráfico de vendas diárias e da média das vendas semanais e mensais.

Devido à necessidade de responder de forma simples e direta com valores de vendas totais para auxílio das farmácias na gestão de *stocks* e devido à frequência semanal explicar e retratar com maior versatilidade, mas também objetividade as transações reais de unidades vendidas, foi definido um gráfico com o *resample* apenas para apresentação do total de vendas por semana para todo o período de vendas. A construção e visualização do gráfico presente na “Figura 10” tem como objetivo a analisar a viabilidade da opção para a realização de previsão para vendas semanais.

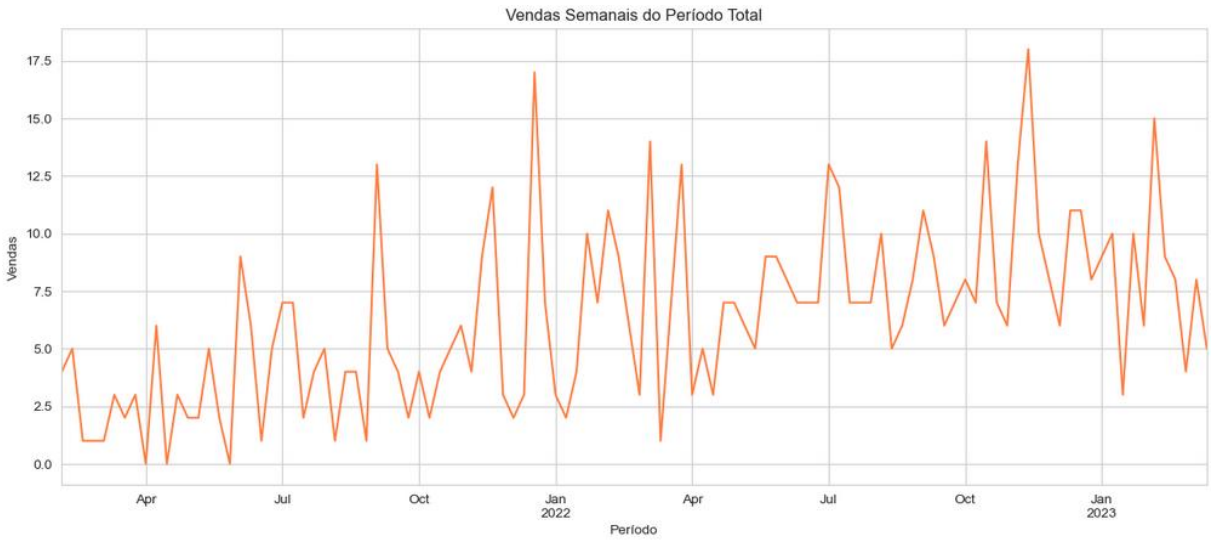

Figura 10 – Gráfico de vendas semanais do "Produto A".

Na continuação da preparação para o *dataframe* desta fase, cujo contém ainda a venda total de unidades do "Produto A" por dia, foi criada uma função em *Python* para serem apresentados todos os dias que são segunda-feira para os períodos definidos. Foram definidos dois períodos, nomeadamente o período para os primeiros 15 dias do *dataframe* e outro período para os últimos 15 dias do *dataframe*. Através dos dias apresentados é possível confirmar os dias concretos que são contemplados com o método *resample* semanal com a soma das vendas, devido a este método contar as semanas de segunda-feira a domingo. O objetivo deste passo consiste em efetuar um arranjo das semanas, para garantir que todas as semanas existentes no período total estão completas com os sete dias. O arranjo semanal é a definição do primeiro e último dia do *dataframe* para serem segunda-feira e domingo respetivamente, assim sendo, conforme o período de vendas próprio de cada produto do segundo *dataset*, foram descartados os dias que originalmente são os limites do período do *dataframe* que são de uma semana incompleta. Esta limitação no período do *dataframe* é adaptada para cada medicamento, sendo que para medicamentos com menos vendas que tenham um dia da semana em falta, foram mantidas as semanas iniciais ou finais do período para existirem mais dias contidos no *dataframe* para maior qualidade de previsão. Esta preparação é necessária para o método optado que calcula a soma das vendas de todos os dias de cada semana para obter o total semanal, porque uma semana que esteja incompleta irá apresentar menos vendas e tal não demonstra com rigor as vendas totais dessa semana, por não possuir todos os dias, portanto, com o arranjo das semanas é evitado o enviesamento dos dados. O tratamento dos dias em falta e das semanas inteiras, influência numa etapa posterior da *EDA*, o resultado do método *Interquartile range* (*IQR*) para o limite superior e tal tem impactos na correção aos *outliers* e consequentemente no desempenho do modelo. Sendo que uma análise rigorosa aos *outliers* será mais correta quando todos os dias existem na sequência do

período total da série temporal de vendas de cada medicamento. O limite superior pode então ser mais fidedigno com o tratamento de dias realizado, mas é também necessário a analisar na próxima etapa a existência de *outliers* que são naturais, mas que são identificados pela sobrevalorização dos valores positivos de vendas existentes, devido a terem sido inseridos dias sem unidades vendas. A função *Python* para verificação do início de cada semana encontra-se no Anexo B na "Figura 21".

A próxima etapa da preparação dos dados consiste na deteção e tratamento de *outliers*. Para a sua concretização vão ser primeiramente realizadas diversas análises e neste seguimento são introduzidas duas novas variáveis ao *dataframe*. Uma variável contém o nome do dia da semana e a segunda variável acrescentada possui o nome do mês. Para a referida integração das duas novas variáveis foi primeiro retirada como índice a coluna "DIA" para poder ser reconhecida e analisada apenas como variável do tipo data do *dataframe* e utilizadas as funções dt.*day_name*() e dt.*month_name*() da biblioteca *Pandas*. De seguida foi definido novamente o índice do *dataframe* para a variável "DIA". O *dataframe* passa a possuir cinco variáveis nesta etapa, sendo as duas adicionadas denominadas por "DIA_DA_SEMANA" e "MES". Com as novas variáveis inseridas foram criados dois gráficos circulares (gráfico do tipo *Pie Chart*) para ser apresentada a distribuição de unidades vendidas através da média pelos dias da semana e pelos meses do ano. Os gráficos *Pie Chart* realizados encontram-se nas figuras 22 e 23 do Anexo B. Conjuntamente foram criadas tabelas ordenadas para a visualização dos dias da semana e dos meses que possuem mais vendas. Esta análise no contexto da fase atual do *CRISP-DM* tem como objetivo o apoio para a posterior interpretação dos *outliers*.

Para identificação dos *outliers*, foi primeiramente realizado um gráfico do tipo caixa de bigodes (gráfico do tipo *Boxplot*) e um gráfico de dispersão (gráfico do tipo *Scatter Plot*). A identificação dos *outliers* foi iniciada pela apresentação de um gráfico *Boxplot*, apresentado na "Figura 26" do Anexo B, para análise da variação dos registos únicos de totais de vendas unitárias. O gráfico *Boxplot* tem como principal interesse a apresentação dos valores únicos dos potenciais *outliers* que em concreto para os casos de cada uma das séries temporais dos 13 medicamentos, são pontos fora dos limites relativos a *outliers* superiores. Para ser de fácil visualização todos os dados da coluna "UNIDADES_VENDIDAS" para cada medicamento foi realizado um gráfico *Scatter Plot*, apresentado na "Figura 11", que apresenta todos os registos de vendas totais diárias pelo que ajuda a identificação de picos e a sua regularidade.

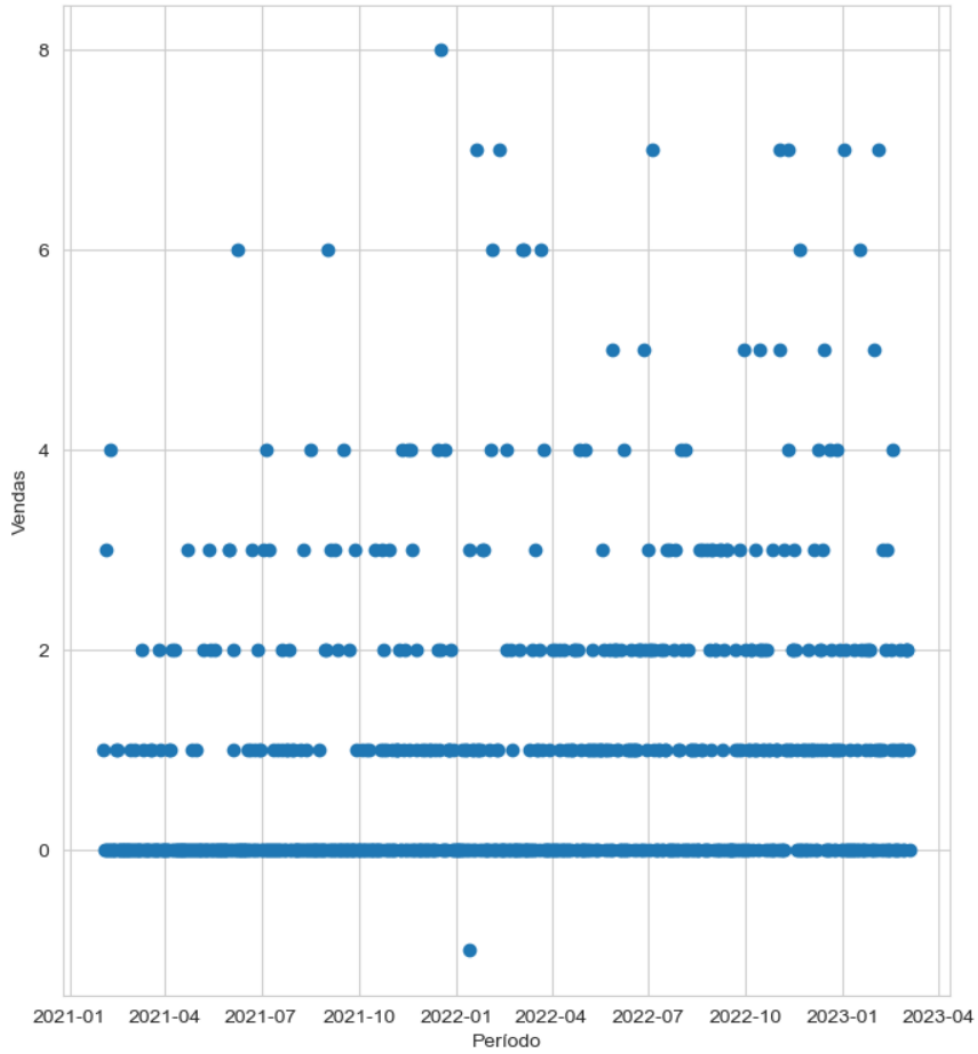

Figura 11 – Gráfico do tipo Scatter Plot sobre as vendas diárias do "Produto A".

Foram também agregados os diferentes sete dias da semana e agregados os doze meses, para as apresentações de dois gráficos individuais do tipo *Boxplot* e dois gráficos do tipo *Scatter Plot* para a observação do conjunto dos valores únicos e distribuição das vendas totais diárias existentes para cada dia da semana ou para cada mês do ano. Os quatro gráficos referidos encontram-se representados pelas figuras 24, 25, 27 e 28 do Anexo B.

Para a identificação específica dos valores de possíveis *outliers* e das respetivas linhas do *dataframe* para esses casos foi utilizado em *Python* através da biblioteca *NumPy* os cálculos necessários para verificar o número da quantidade total dos valores existentes acima do limite superior calculado, assim como a quantidade de valores ou linhas existentes inferiores ao limite inferior. Conforme visualizado no passo anterior pelo gráfico de *Boxplot* apenas existem casos superiores ao limite superior. De seguida é feito o *print* para a identificação da posição das linhas que contêm valores acima do limite superior e utilizando o iloc de *Pandas* é realizada, através do índice de números inteiros que são as posições recolhidas, a seleção para apresentação em tabela das linhas para análise. Devido

ao *dataframe* possuir as colunas "DIA_DA_SEMANA" e "MES", as linhas analisadas contêm cinco variáveis e não apenas as variáveis de data, nome do medicamento e total de vendas diárias. Na observação das linhas é possível verificar o dia da semana e mês em que estão inseridas e recorrendo à análise dos dois *Pie Charts* e das duas tabelas ordenadas anteriormente efetuadas, foi validado se os dias com vendas superiores são em meses ou dias da semana com mais vendas. Ao ser facilmente analisada esta informação é possível decidir com maior certeza se os *outliers* são dados extremos ou exagerados que não são reais ou se são reais, sem erro e naturais do seu âmbito. A decisão de tratamento dos *outliers* para cada previsão é apenas tomada após a visualização do gráfico do tipo histograma para observação da distribuição. Na “Figura 29” do Anexo B pode ser visualizado o gráfico do tipo histograma.

Após a análise da distribuição foi verificado o total de cada valor de vendas diárias e o total de registos existentes com zero vendas. Foi optado pelo não tratamento dos possíveis *outliers* identificados se existirem mais de metade de linhas com as vendas de unidades com zero. Pelas análises feitas, tratar os *outliers* tanto através da sua remoção ou limitação, para situações em que existe cerca de metade de linhas com zero vendas, irá enviesar a modelação e causar impactos negativos na previsão. Pelo contrário se os dias com zero vendas são menos que metade do total, são adaptados os valores superiores identificados como *outlier*, limitando os mesmos para conterem como valor, em todos os campos nesta situação, o número de vendas igual ao limite superior. O limite superior foi o calculado anteriormente pela regra de intervalo interquartis, conhecida como 1.5 x *IQR*. O valor numérico inserido para atualizar os campos com *outliers* é o arredondamento superior do limite superior, para definir que é utilizado um número inteiro no tratamento optado, sendo conhecido pelo método *capping*. Desta forma está finalizado o tratamento de *outliers* através do estabelecimento de limites (*capping*) e do uso do método *IQR* para as séries temporais de vendas dos medicamentos, mitigando impactos não pretendidos nas previsões a efetuar, para serem alcançados valores preditos mais rigorosos e melhores resultados nas métricas da avaliação ao ajuste do modelo. Para os casos de medicamentos com séries temporais que possuem casos excecionais em que o total de unidades vendidas para um dia é negativo, foi optada a não eliminação ou correção do referido valor, devido a este estar associado à correção de um valor alto de total de vendas para um dia anterior vizinho. Esta decisão é apenas para séries temporais de medicamentos em que após a análise dos possíveis *outliers* foi decido pelo não tratamento dos mesmos. Para séries temporais de medicamentos com mais vendas diárias foi realizado o *capping* dos *outliers* superiores e foram também atualizados para zero os valores a negativo relativos a totais de vendas diárias.

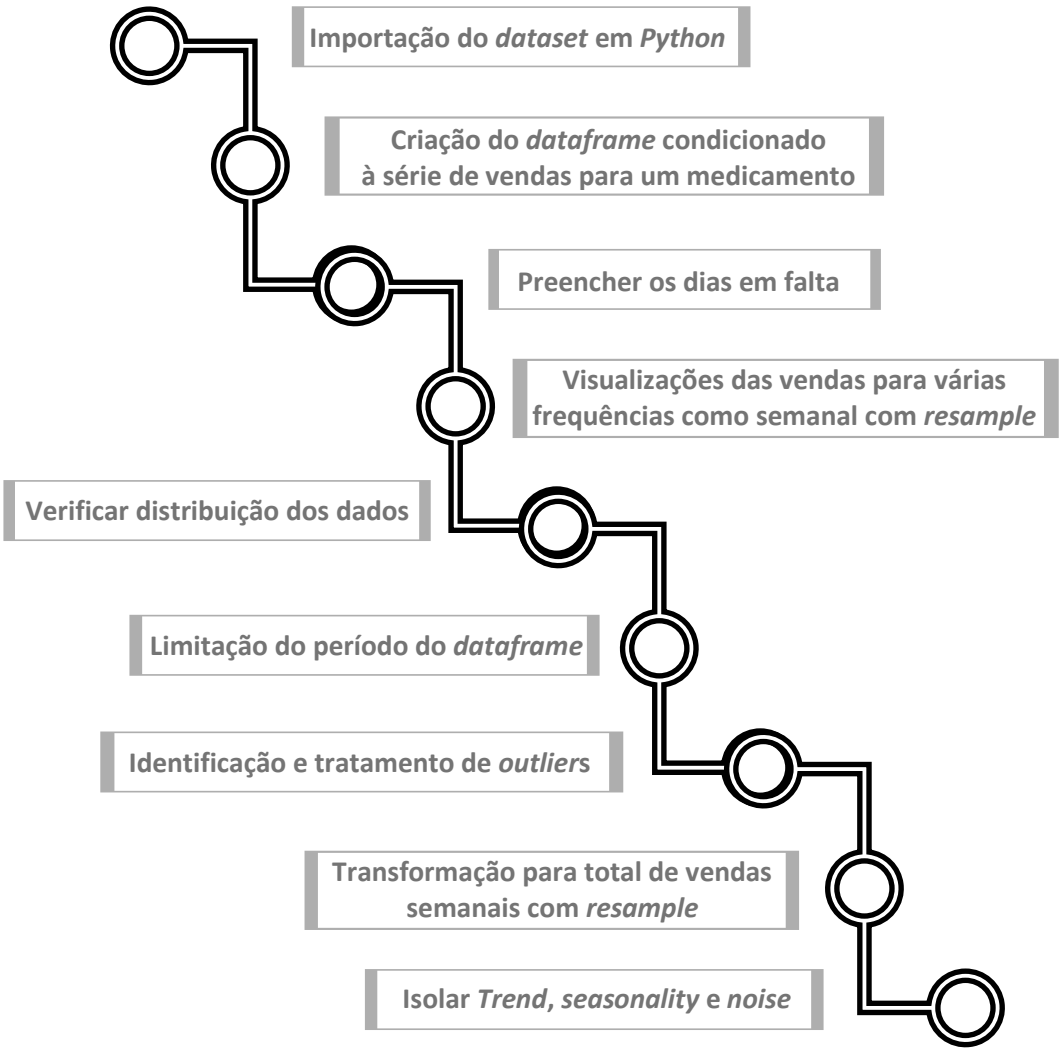

Figura 12 – Percurso de análises essenciais realizadas na segunda fase CRISP-DM.

O passo final da fase preparação de dados é a observação das componentes das séries temporais, sendo necessário analisar os movimentos de tendência e sazonalidade para entender os seus efeitos aplicados à previsão. Para estender a análise de apoio até este passo realizada, procurando uma compreensão aprofundada das séries temporais utilizadas neste projeto, é uma boa prática a decomposição da série temporal, existente no *dataframe* no estado atual. Na sequência dos requisitos de análise identificados é utilizada a técnica *seasonal_decompose* da biblioteca *Statsmodels* que divide a série temporal entre as três componentes: tendências, sazonalidade e resíduos. A importância da descomposição é devido à sua capacidade de estruturar o entendimento para o problema da previsão e tem como vantagens a obtenção de informação a ser parametrizada no modelo de previsão escolhido para o problema. Observando o gráfico de vendas é possível concluir antecipadamente que a série temporal em causa é aditiva porque o padrão de crescimento ou diminuição dos pontos

relativos às vendas diárias é semelhante durante período o total. No caso de não ser aditiva seria então multiplicativa, contudo para isso o crescimento ou diminuição iria ser exponencial com o tempo. Pela apresentação dos gráficos da decomposição é confirmado que a variação da sazonalidade é constante ao longo do período, se, por outro lado, esta aumentasse com o tempo seria identificado um modelo multiplicativo e não como este caso que é aditivo. Na “Figura 12” são descritos os passos essenciais no percurso de preparação dos dados para este projeto.

De forma a preparar o *dataframe* para a seguinte etapa do modelo é necessário efetuar a transformação optada para a alteração da frequência de vendas diárias para vendas semanais. Esta opção permite a elaboração de um processo mais adaptável para diferentes previsões de vendas de medicamentos, que possuem muitas ou poucas vendas, sendo facilitada a realização de um modelo ou solução padrão, como foi procurado alcançar no presente projeto. A realização de uma previsão diária das vendas de medicamentos apresenta riscos associados à imprecisão inerente aos muitos fatores que influenciam a procura pelos clientes de medicamentos. De forma sintética, a escolha da transformação para vendas semanais auxilia o presente projeto nomeadamente na fase atual e seguintes fases de modelação e avaliação, onde são efetuadas afinações ao modelo e comparações de resultados e das métricas de avaliação. Paralelamente às vantagens associadas às decisões de tornar mais fácil e eficaz a preparação e conclusão do presente projeto de previsão, existem também vantagens para os recetores das previsões realizadas, as farmácias. Se o período de vendas fosse igual ou superior a dez anos, a previsão de vendas mensais podia ser considerada, por ser possível a sua execução pelos modelos *Python* escolhidos, contudo a previsão mensal demonstra ter menos interesse para a área de negócio beneficiária. Desta forma a análise e definição para a previsão com dados semanais para semanas futuras, encontra-se concluída.

A versão final da série temporal, presente na “Figura 13”, fica concluída após a transformação do *dataframe* para a frequência semanal, somando as vendas diárias de cada semana. O resultado da transformação com o *resample* é a criação da série temporal a ser modelada para previsão quantitativa das vendas, sendo esta constituída apenas pelas variáveis "Semana" e "Vendas_semanais".

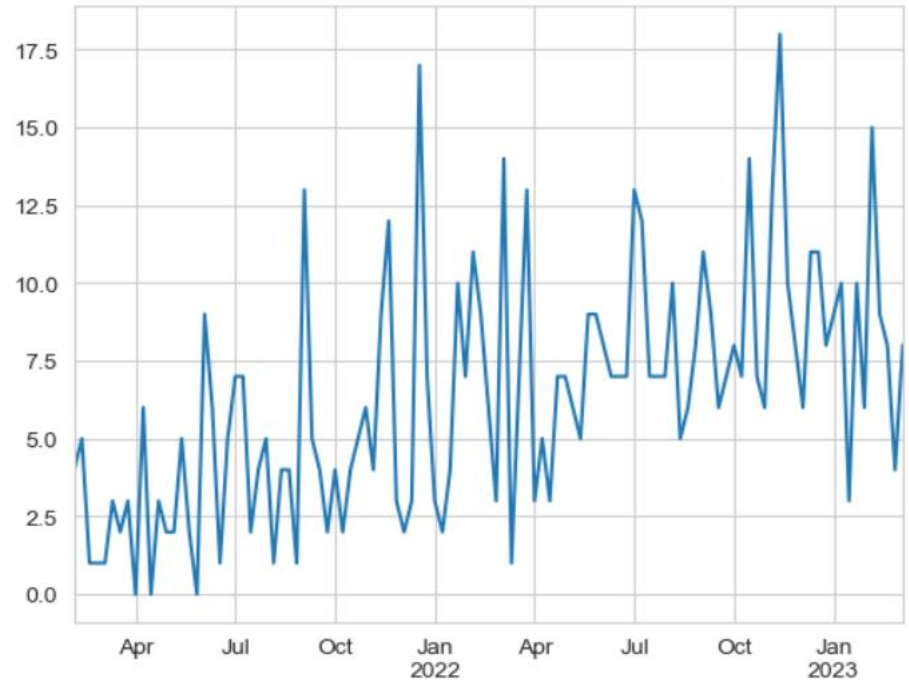

Figura 13 – Gráfico da série de vendas semanais".

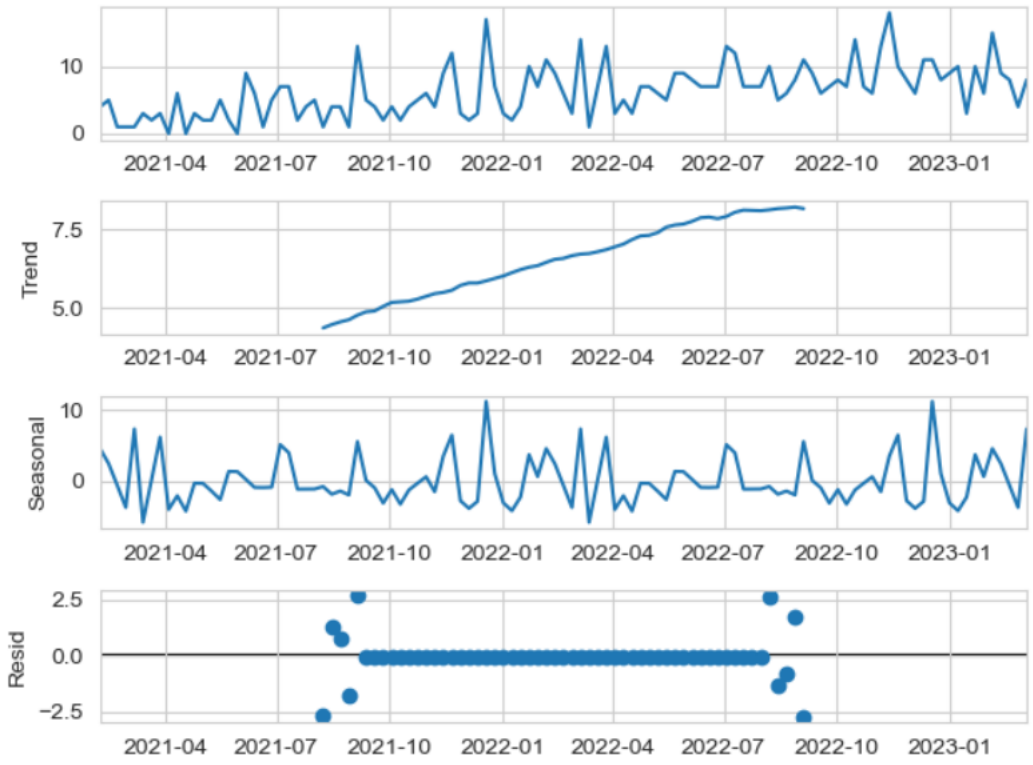

Figura 14 – Gráfico da decomposição da série de vendas semanais.

Foram novamente apresentadas as diferentes componentes para a última versão da série temporal e foi confirmado que a tendência é linear e que tanto a tendência e sazonalidade são constantes com o decorrer do período sendo que estas contribuem para a existência de um modelo

aditivo. O gráfico da decomposição da série temporal na frequência semanal apresenta-se pela “Figura 14”.

Após esta etapa, utilizando o primeiro *dataset* relativo ao “Produto 1” foram ainda feitas experiências realizando previsões com os Modelos *ARIMA* e *Prophet*.

Como escolha resultante da gestão do projeto de investigação, tendo em conta o fator do tempo e da quantidade de previsões finais a executar, foi decidido que o modelo a utilizar na próxima fase de modelação é o *Prophet*. Face às experiências efetuadas, este modelo transmite confiança e poderá ainda ser parametrizado com mais detalhe para atingir melhores resultados. Este modelo também simplifica a escolha de modelo final para o foco na sua afinação, assim como:

- Permite execução e resultados rápidos e acessíveis para comparação com os dados reais e a sua visualização em gráfico;
- Preparação e exportação dos dados preditos;
- Avaliação para o conjunto de treino e para o conjunto de teste para avaliar impactos negativos de *underfitting* e *overfitting*.

Em resumo, nesta fase de preparação dos dados foram desenvolvidas e asseguradas as abordagens mais benéficas para um projeto inovador. Foram percorridos vários passos através da definição e sequenciação de diferentes técnicas que foram ajustadas entre si e feita a remoção de outras, para serem mantidas as que contribuem para a solução identificada e ambicionada. Todos os restantes passos realizados para o "Produto um" e modelação com os modelos *ARIMA* não se encontram descritos devido ao modelo para solução do problema selecionado ser o *Prophet*.

3.5. *CRISP-DM* Fase 4: Modelação

Foi possível iniciar fase de modelação tendo os objetivos bem definidos, devido à conclusão das fases *CRISP-DM* anteriores e pelas análises realizadas através de experiências com os modelos *ARIMA* e *Prophet*, comparando entre estes os gráficos com os valores das previsões e valores das métricas de avaliação *MSE, RMSE* e *MAE*. No entanto, o contexto da referência ao modelo *ARIMA* é indicada neste projeto apenas como uma etapa no percurso. À semelhança dos testes realizados numa primeira etapa de experiências de modelação, vão ser utilizados os 13 medicamentos. É essencial possuir o entendimento que todas as séries temporais dos medicamentos são univariadas, desta forma é possível escolher o modelo de previsão em séries temporais de forma correta.

Para ser possível em contexto real e com dados reais, dar resposta às necessidades das farmácias é exigido que num curto espaço de tempo seja testado um modelo simples de previsão (sem muitas

pré-análises e preparações) com ciência de dados que facilite a execução e transmissão de resultados para várias previsões que providenciem confiança para escolhas de soluções. Face ao problema e às suas características a escolha optada é pela utilização do modelo *Prophet* (este modelo foi criado pela equipa *Facebook's Core Data Science*), que pode ser executado no código *Python* através de uma biblioteca open-source disponível em *Python* e *R*. O processo principal do *Prophet* é realizar previsões em séries temporais univariadas (com apenas uma variável) tendo como base o modelo aditivo conforme confirmado na anterior fase *EDA*. Uma série temporal univariada é constituído por observações únicas registadas sequencialmente em intervalos de tempo igualmente espaçados.

O *Prophet* tem notoriedade devido a disponibilizar uma parametrização intuitiva e simples, sendo que possibilita a técnicos e investigadores que não tenham muita experiência em modelos de previsão de séries temporais, a realização de previsões com consistência em diferentes problemas e áreas de negócio. Este modelo tem a capacidade de melhores execuções para séries temporais com várias sazonalidades e assim como para dados de histórico de vendas com grande vulnerabilidade devido a impactos de sazonalidade. Este modelo possui também a vantagem de obter bons desempenhos para séries temporais com dados em falta, mudança nas tendências e aplicação na existência de *outliers*.

Para início da modelação das séries temporais, é necessário nesta etapa a instalação e importação de mais uma biblioteca no *Notebook* em *Python*, biblioteca *Prophet*. Para executar as previsões com *Prophet* é sempre necessário renomear as colunas do *dataframe Python*, sendo constituído pela série temporal com as variáveis "Semana" (definida como índice do *dataframe*) e "Vendas_semanais". Para produzir este requisito foi removido o atual índice do *dataframe* para de seguida ser possível identificar as duas variáveis "Semana" e "Vendas_semanais" e renomear as mesmas para "ds" e "y" respetivamente. Estes nomes para as variáveis são os obrigatórios a serem executados pelo algoritmo do modelo *Prophet* e devido a este modelo necessitar que a coluna "ds", que no caso deste projeto contém as datas semanais, não seja do tipo índice, vai ser mantido o índice de inteiros padrão, criado automaticamente pelo *Python* para índice de *dataframes*.

Para a verificação da existência de *underfitting* e *overfitting*, e para posterior avaliação a ser realizada sobre a qualidade das previsões, foi efetuada a divisão do *dataframe* em dois conjuntos de dados. A divisão do *dataframe* para os conjuntos de treino e de teste é importante e deve ser adotada para ser possível avaliar de forma adequada o desempenho dos modelos de previsão. O conjunto de treino passa a conter 70% da informação e o conjunto de teste fica com 30%. Pelo motivo das previsões a modelar serem em séries temporais, não deve ser feita a divisão aleatória, mas sim a divisão ordenada no tempo. Acrescentando às análises e definições consideradas a adotar foi essencial também proceder à visualização do número de linhas do *dataframe* de vendas semanais, que em concreto é de 108 ou 109 para os medicamentos em causa. Na sequência do total de semanas verificado foi criado o *dataframe* "treino" que contém a seleção das primeiras 76 linhas, através do

método iloc associado ao *dataframe* do conjunto completo. Através do mesmo procedimento foi criado o *dataframe* "teste" para as linhas seguintes à linha presente na posição ordenada 76. Após a divisão através da razão 70:30, foi verificado que o conjunto de treino contém 76 linhas ou semanas e o conjunto de teste contém 32 semanas. Foi também realizada a apresentação das linhas dos limites dos períodos de cada conjunto para validar a conformidade das datas e valores de vendas.

Esta etapa consiste na parametrização do modelo de previsão *Prophet*, em que são escolhidas as afinações do modelo e consolidadas as parametrizações a serem integradas na execução do ajuste de modelo (*fit* do modelo). As análises e parâmetros do modelo *Prophet* utilizados para as previsões de vendas para cada um dos 13 medicamentos encontram-se infra-apresentadas:

- Parâmetro *growth='linear'*:
    - Tendo com opções os argumentos "*linear*" ou "*logistic*" relativos ao movimento da tendência. As séries temporais em causa não são de crescimento logístico porque não apresentam um aumento crescente com estabilização dos valores com o aproximar do seu pico e posterior diminuição de valores. No caso de ser escolhido o crescimento logístico, o modelo *Prophet* requer que seja especificado um ponto máximo alcançável.

- Parâmetro *seasonality_mode='additive'*:
    - Para modelos aditivos como no presente caso, contudo para parametrização deste ponto também existe a opção multiplicativo.

- Devido à série temporal encontrar-se para a frequência semanal não são usados os parâmetros *weekly seasonality* e *daily_seasonality*. Sendo que o modelo ao fazer o ajuste do conjunto de treino irá detetar a situação referida e desativar a sazonalidade semanal e sazonalidade diária.

- Existem dois tipos de parâmetros que podem ser utilizados individualmente ou em conjunto para a modelação das previsões das vendas de medicamentos, sendo os mais importantes e que permitiram o sucesso da utilização do modelo *Prophet* para a previsão das séries temporais dos 13 medicamentos, devido à sua afinação permitir os valores preditos a melhor ajuste dos picos de vendas existentes nos dados reais. Os parâmetros em causa são o *yearly_seasonality* e um parâmetro personalizado para adaptações de sazonalidades anuais ou mensais.

- Estes dois parâmetros efetuam várias combinações de sazonalidades que permite regular e melhorar os valores preditos, encontrando um equilibro entre a variância e a distorção dos mesmos. Contudo, existem combinações de uso de ambos os parâmetros que não podem ser utilizados em conjunto devido à redundância de impactos causada por ambos porque podem dar origem a efeitos muito negativos de *overfitting*, se a variância for muito alta que causa o problema da aprendizagem do modelo demasiado vincada e adaptada às observações históricas, incluindo ao ruído dos dados (*white noise*), o que resulta em projeções desses movimentos copiando-os para a previsão futura. O *white noise* foi muito reduzido nas séries temporais em estudo após ter sido utilizado o útil e poderoso método *resample* para uma conversão de termo *downsampling*, de dia para semana. O arranjo final constituído pela realização da preparação dos dados da fase anterior representa uma proposta de percurso para chegar à solução final, sendo que esta possui como uma prioridade a prevenção para alguns problemas que podem ocorrer na presente fase de modelação.

- Na continuação dos esforços aplicados para a concretização de passos no projeto, contemplando e garantido a análise de cuidados a serem considerados foram efetuadas as seleções dos valores parametrizados e dos números de séries de *Fourier* (*Fourier terms*) a definir, tendo sido essencial efetuar vários testes para ser possível proceder a comparações de gráficos e de métricas de avaliação, através de tentativa e erro pelas de experiências efetuadas, de modo a realizar a melhor afinação e ajuste do modelo.

- O primeiro parâmetro dos dois mais influentes na presente modelação realizada é o parâmetro *year_seasonality* e o argumento ou valor a aplicar ao mesmo pode ser um número específico. O *year_seasonality* demonstrou ser extremamente importante e para além de permitir a definição manual de um valor também possui como argumentos padrões os seguintes: *TRUE*, *FALSE* ou "*auto*". Se a escolha for *yearly_seasonality=True* o algoritmo do modelo associa o valor dez, que será interpretado como dez ciclos de mudança num ano. A opção "*auto*" apresentou impactos mais positivos ou mais negativos, consoante as diferentes séries de vendas e para os casos que melhorou os resultados preditos foi devido ao impacto concreto nas previsões realizadas ser semelhante a um número maior que dez, como, por exemplo 12, definindo mudanças de ciclos com mais rapidez e em maior número. Nos casos de outras previsões realizadas foi a escolha de um número específico e mais elevado, como 15, que melhorou a adaptação e ajuste dos valores preditos para com os movimentos que representam os valores reais.

- Um segundo parâmetro igualmente importante foi adicionado manualmente através do método *add_seasonality* para treino do modelo para ciclos de sazonalidade anual ou mensal, onde foram integrados os argumentos *name*, *period*, *fourier_order* e *prior_scale*. A definição do nome é livre e tem como objetivo designar a esta componente na apresentação da previsão. O período é relativo ao número de dados ou vendas existentes, sendo que este será definido consoante o *fourier_order* definido. Para as previsões realizadas em que o *fourier_order* foi definido com cinco (*fourier_order*=5), é pretendido que a sazonalidade retrate o período mensal e desta forma o argumento *period* contém um valor, entre 4 a 5, devido a ser este ser o número de semanas e vendas existentes para o período de um mês. O *prior_scale* é um input adicional para o parâmetro "*add_seasonality*", sendo utilizado para evitar o *overfitting* através da suavização da previsão. Por definição o valor deste parâmetro é de dez e tal possibilita pouca regulação. O valor utilizado é de dez (*prior_scale*=0.01) tendo sido escolhido pela capacidade de atenuar os efeitos de sazonalidade provenientes da definição especificada. A criação do novo parâmetro específico capacita o treino do modelo *Prophet* a uma previsão mais flexível, através da interpretação de uma nova sazonalidade anual ou mensal. Contudo, a análise e validação dos impactos na previsão apenas podem ser concluídas após a visualização dos dados, dos gráficos e das métricas de avaliação dos dados preditos face aos dados reais, para o período de teste.

- O último parâmetro utilizado na definição para o treino do modelo *Prophet* a ser executado é o "*interval_width*" que regula a largura dos intervalos de incerteza. Este último parâmetro não tem impacto nos valores preditos concretos, sendo que a sua utilidade é na posterior apresentação da previsão executada pelo modelo *Prophet*, devido a especificar através da marcação de uma sombra com a cor azul, os limites superiores e inferiores mais prováveis a conter os valores desconhecidos e intencionados para a previsão.

Os próximos passos para conclusão da definição das características do modelo *Prophet* e execução do mesmo são:

- Ajuste do modelo conforme o parametrizado para os dados do conjunto de treino;
- Uso do método *Prophet make_future_dataframe*() para definir os períodos pretendidos para previsão. Através dos argumentos "*periods*" e "*freq*" foi definido a utilização de 52 períodos na frequência de semanas, desta forma foi definido a previsão até um ano. O número de período definido é relativo a períodos futuros, contudo o comportamento padrão do *Prophet* inclui as

datas utilizadas na aprendizagem do modelo e realiza previsões também para este período, cujo é o período de treino, o que possibilita a posterior avaliação ajustada;

- Por último é utilizado o método *predict*() para atribuição de previsões para as linhas futuras, da série temporal das vendas semanais. Sendo o número de linhas especificado no passo anterior. Este método cria um *dataframe* que utiliza a coluna "ds" para as datas das semanas e atribui o valor previsto na coluna que denomina por "yhat". As outras colunas existentes no novo *datafame* criado pela previsão do modelo *Prophet* são referentes a componentes e intervalos de incerteza.

Após a rápida execução do modelo *Prophet* através de código *Python*, foi realizada a visualização das primeiras e últimas linhas do novo *dataframe* que possui os valores preditos relativos ao número de unidades vendidas semanalmente, para as semanas existentes nos dados reais e para as 52 semanas posteriores à última semana existente no conjunto de treino. Nas figuras 30 e 31 do Anexo B encontram-se os parâmetros definidos do modelo *Prophet* para a previsão do "Produto A", assim como as linhas com os resultados dos valores preditos.

De seguida foi criada a visualização da previsão através do gráfico do *Prophet*, onde foi colocado o argumento *uncertainty=True* para ser apresentado no gráfico a marcação do intervalo que correspondem à incerteza, cujo foi definido no argumento "*interval_width*" parametrizado anteriormente para o modelo. Para a análise da previsão é apresentada a linha azul para os valores resultantes da previsão e os pontos a negro são os valores reais. A sombra a azul-claro é o intervalo de confiança existente foi definida em 80% e engloba toda a previsão. Foi também apresentado o gráfico das várias componentes existentes no modelo aditivo aplicado pelo ajuste efetuado através do *Prophet*. Para uma melhor visualização e validação do desempenho e resultados alcançados pelo ajuste do modelo com os parâmetros definidos, foi criado um gráfico, presente na "Figura 15", para o período de teste onde são apresentadas as linhas relativas às vendas reais e às vendas preditas. Este gráfico é essencial para analisar o resultado do modelo para poderem ser realizadas melhorias e é utilizado para avaliar a previsão.

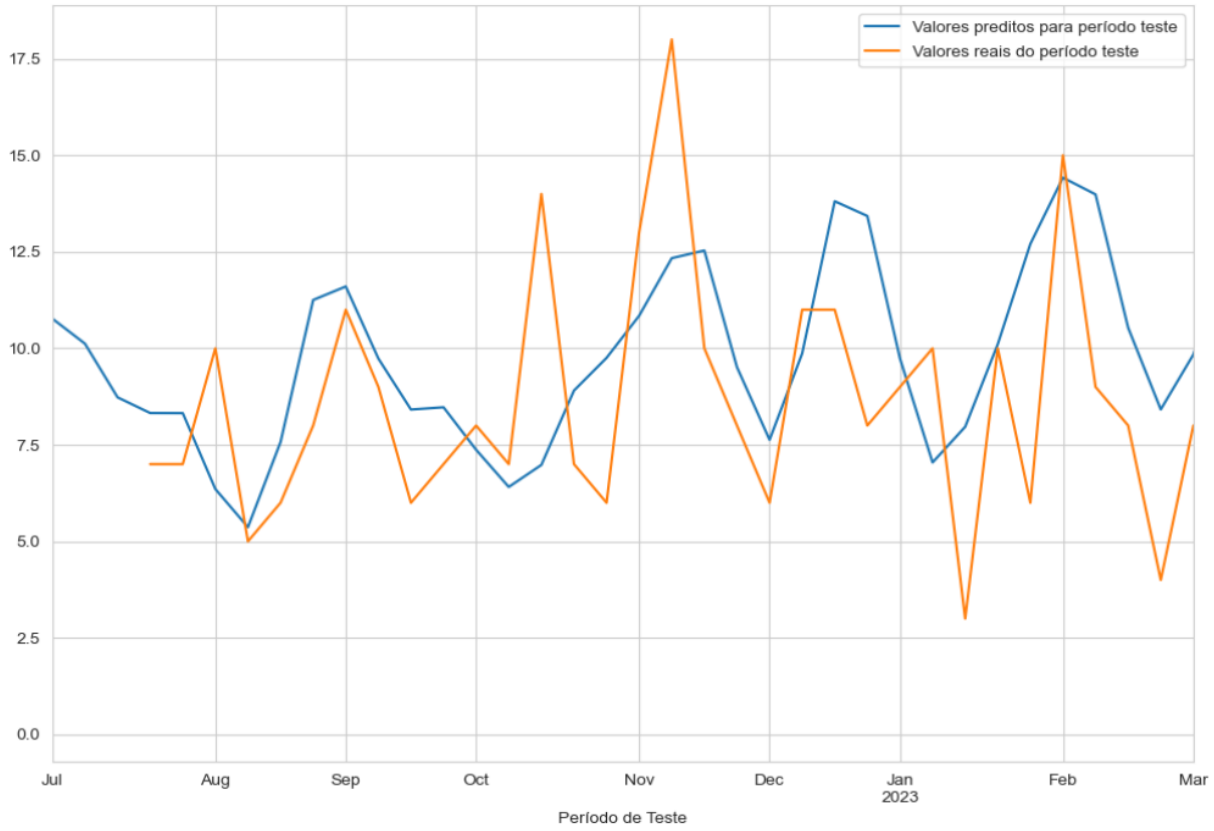

Figura 15 – Gráfico da comparação da previsão dos dados não treinados com os dados reais.

Antes de ser iniciada a fase da avaliação é realizado um novo *dataframe* que é constituído pela junção das colunas dos valores reais e dos valores preditos para as semanas que contemplam o período do conjunto de teste. Desta forma o novo *dataframe* possui colunas importantes, como a coluna "ds" para as datas de cada semana do período de teste, para os valores preditos é inserida a coluna "yhat" de origem no *dataframe* da previsão e é integrada a coluna "y" com os valores reais provenientes do conjunto de teste. Após criado o *dataframe* de junção dos valores para o período de teste, são visualizadas as suas linhas de forma validar a sua integridade para os posteriores cálculos a serem realizados para criação de métricas de avaliação.

Através das mesmas técnicas e lógicas referidas no anterior parágrafo foi criado um *dataframe* com a junção dos dados reais e dos preditos para o período de treino. Ao ter o novo *dataframe* para a previsão ajustada são verificadas as linhas contidas na mesma, sendo que também é criado um gráfico à semelhança do criado para o período de teste, sendo este último para o período de treino. Desta forma conseguimos facilmente identificar o desempenho do modelo para os dados de ambos os conjuntos e prosseguir com as indispensáveis avaliações na próxima fase.

Após a avaliação do modelo e a definição final de todas as preparações efetuadas inicialmente aos dados e dos parâmetros ajustados para aprendizagem do modelo, este encontra-se pronto a ser executado uma última vez. A previsão final tem como objetivo a transmissão e a disponibilização da previsão dos valores das vendas semanais para o período do próximo ano para a área de negócio afeta à atividade de gestão de *stock* de produtos farmacêuticos. A previsão final encontra-se na "Figura 16". De forma a concluir a previsão é feito o ajuste do modelo *Prophet* com todos os dados históricos existentes (período total das séries de vendas dos medicamentos), porque o conjunto total dos dados possui os mesmos movimentos de sazonalidade existentes e parametrizados no conjunto de treino. Nos últimos passos, são realizadas as apresentações da previsão e obtida a tabela com as linhas das 108 ou 109 semanas de dados reais incluindo também as 52 semanas futuras. Para as datas referidas são apresentados os valores inteiros das vendas resultantes da aplicação da previsão com o modelo *Prophet*, através do arredondamento superior afeto à coluna "yhat" (utilizando o método *ceil*() da biblioteca *NumPy*). Esta exportação da tabela com os resultados com o número de venda semanais preditas para o ano seguinte só pode ser concluída após terem sido alcançados os melhores parâmetros para o ajuste do modelo e avaliação do conjunto de teste para os dados do mesmo período, sendo que nesta fase e na fase seguinte foram feitas experiências com diferentes preparações, não alterando os parâmetros do modelo para validação de que os passos da *EDA* respondiam melhorando o resultado da previsão.

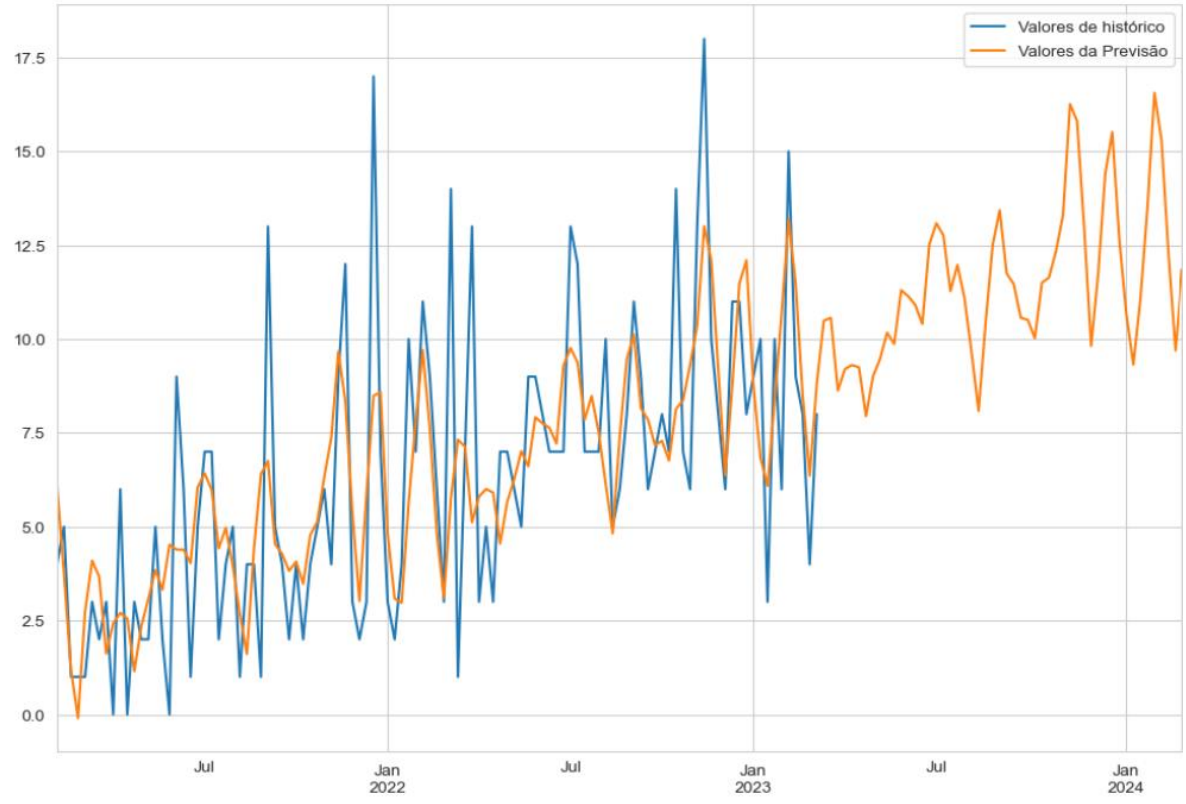

Figura 16 – Gráfico da previsão do "Produto A" pelo ajuste de todos os dados históricos

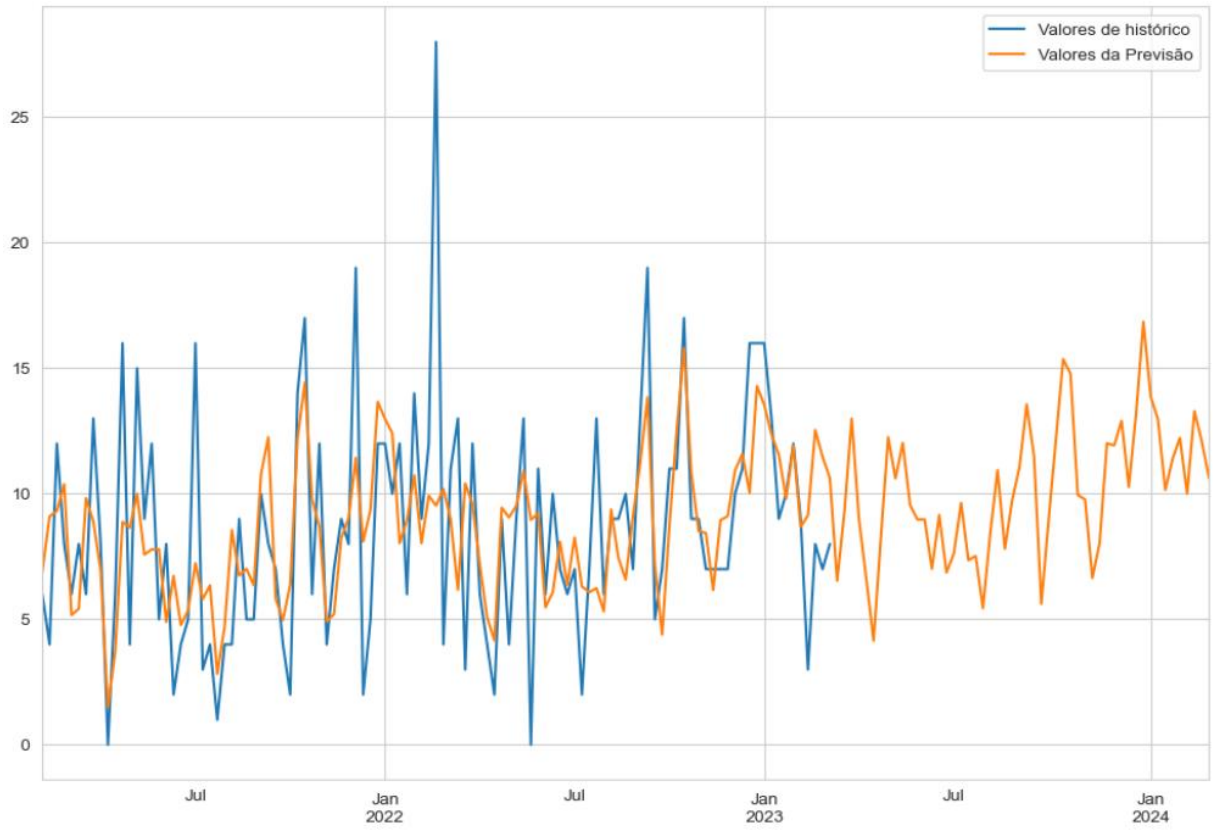

Figura 17 – Gráfico da previsão do "Produto B" pelo ajuste de todos os dados históricos.

No gráfico da "Figura 17" é possível verificar o ajuste de outra previsão realizada em *Prophet*, para o "Produto B".

3.6. *CRISP-DM* Fase 5: Avaliação

A avaliação do desempenho do modelo inicia-se depois da previsão realizada pelo *Prophet* e da criação de dois *dataframes* que possuem o Y real e Y predito. O objetivo desta fase é analisar os erros presentes entre os dados reais e os dados preditos, sendo que o tipo de métricas escolhidas é:

- Erro Quadrático Médio (*MSE*), calculado a partir da média dos erros absolutos;
- Raiz Quadrada do Erro Médio (*RMSE*), raiz quadrada do *MSE*, onde o valor do erro retorna à unidade de medida do modelo (no *MSE*, a unidade de medida é quadrática);
- Erro Médio Absoluto (*MAE*), calculado a partir da média dos erros absolutos.

Para em *Python* serem facilmente apresentados os cálculos das três métricas optadas para a avaliação, é necessário realizar a importação de cada uma destas através das bibliotecas *Scikit-Learn* e *Statsmodels*. De seguida foram criadas as duas novas varáveis "y_true" e "y_pred", em que a primeira integra os valores reais de vendas semanais e a segunda os valores preditos. Estes valores tinham sido

anteriormente reunidos num único *dataframe*, desta forma torna-se fácil realizar o presente passo. De seguida são criadas três novas variáveis para cada um dos resultados das três métricas e a cada uma é aplicado o método para a respetiva métrica, sendo que para cada método relativo a cada uma das três métricas é necessário apenas indicar as duas variáveis "y_true" e "y_pred". Após definido o código *Python* para cálculo das métricas de avaliação são realizados os *prints* para cada um dos três valores, sendo especificada a limitação em três casas décimas. As técnicas referidas para cálculo das três métricas são aplicadas duas vezes, a primeira para avaliação dos dados do período de teste e a segunda para a avaliação dos dados do período de treino.

As métricas associadas ao período de teste são as mais importantes para verificar a capacidade do modelo para realizar previsões em dados futuros não aprendidos pelo modelo. No entanto, ambas as métricas dos períodos necessitam de ser comparadas em cada nova experiência de previsão com o *Prophet*, para analisar o *overfitting* e *underfitting*. O desejável é observar o gráfico da previsão para o período de teste e verificar uma boa adaptação da previsão comparando-a no mesmo gráfico com as vendas reais. Se a referida primeira análise for satisfatória, são então verificados os resultados das métricas em ambos os períodos, sendo que os melhores resultados das métricas são:

- Resultados mais baixos possíveis (próximos de zero) nas métricas *MSE, RMSE* e *MAE* na avaliação da previsão (período de teste);
- Resultados com a menor diferença entre a avaliação ajustada (período de treino) e avaliação da previsão, sendo comparados os resultados *MSE, RMSE* e *MAE* das duas avaliações.

Na sequência das experiências efetuadas dos modelos e com a análise das respetivas avaliações, foi concluído que caso o resultado *RMSE* diminua para o conjunto de treino e aumente para o conjunto de teste, significa que as alterações realizadas aos parâmetros do *Prophet* produzem o efeito não desejado de *overfitting*. Por outro lado, se através da análise do gráfico com o resultado da previsão não for detetado o efeito de *underfitting* e se foi alcançada uma diminuição da diferença dos resultados das métricas entre a avaliação ajustada e a avaliação da previsão é validado que os valores nos argumentos de parametrização definidos para execução do modelo proporcionam um melhor desempenho e consequentemente previsões de maior qualidade. Completando este último cenário, cujo é o pretendido alcançar em cada uma das 13 previsões, é também validada uma melhoria no desempenho do modelo se após a alteração nos parâmetros do modelo for criado um pequeno aumento do *RMSE* na avaliação da previsão, mas verificada uma diminuição significativa na diferença dos resultados *RMSE* entre as avaliações dos dois conjuntos. Os resultados das previsões realizadas com o modelo *Prophet* para os 13 medicamentos, encontram-se na “Figura 18”, sendo o valor registado, relativo à métrica de desempenho *RMSE*.

Previsão para cada produto farmacêutico	Previsão para o conjunto de treino	Previsão para o conjunto de teste
Produto A	**2.72**	**3.03**
Produto B	**3.86**	**4.51**
Produto C	**2.14**	**3.20**
Produto D	**4.32**	**6.41**
Produto E	**2.74**	**3.55**
Produto F	**4.97**	**8.65**
Produto G	**3.80**	**7.06**
Produto H	**5.38**	**9.17**
Produto I	**4.23**	**7.60**
Produto J	**5.63**	**10.65**
Produto K	**3.63**	**4.76**
Produto L	**2.90**	**3.74**
Produto M	**2.80**	**4.85**

Figura 18 – Resultados da avaliação do modelo Prophet com a métrica RMSE.

CAPÍTULO 4

Resultados e Discussão

O presente projeto foi realizado com o intuito da sua aplicação em contexto real no mercado das empresas farmacêuticas.

O projeto realizado bem como o tema (previsão de vendas) podem ser empregues para previsão de produtos semelhantes, ou seja, aplicados e integrados noutras soluções de ciência de dados. No entanto, este centrou-se numa abordagem concreta sem grandes desvios, de modo a serem obtidos os resultados que permitam responder à problemática principal de previsões de vendas de medicamentos.

O problema habitual nestes contextos com dados reais foi ultrapassado através da delapidação e da preparação dos dados efetuada, o que permitiu evitar a existência de resíduos ou confusão nos dados. Dos métodos de limpeza de dados utilizados, destaca-se a reamostragem para transformação do conjunto de dados para a frequência semanal.

Foram realizadas experiências com os modelos *ARIMA* e *SARIMA*, contudo o objetivo definido foi direcionado para a utilização do modelo *Prophet*, para o aprofundamento das capacidades deste modelo de previsão de séries temporais. Neste sentido, foram determinados e afinados os parâmetros mais importantes para um bom desempenho das previsões. Nesta fase do projeto concluiu-se que os parâmetros automáticos do *Prophet* não cumprem os requisitos necessários, como, por exemplo, *yearly_seasonality*="*auto*" devido ao problema de *underfitting*. Este problema é relativo ao facto de o modelo não ter conseguido aprender o suficiente com os dados do conjunto de treino, o que causa impactos na previsão devido à baixa variação dos seus valores e origina um desempenho enviesado do modelo.

Para a previsão final a ser realizada em todos os produtos do segundo *dataset* do projeto, o percurso de afinação do modelo foi o seguinte:

- Validação que o desempenho com o parâmetro *yearly_seasonality*="*auto*" é insatisfatório;
- Devido ao parâmetro referido realizar um ajuste com comportamentos de *underfitting*, é definido o *yearly_seasonality*=14;
- Existindo melhorias na previsão, observando o gráfico do período de teste e as métricas de avaliação, são realizadas várias execuções com o argumento anterior, baixando o valor do mesmo;

- Se ao ir diminuindo o valor do *yearly_seasonality* o *RMSE* de teste desce e o *RMSE* de treino sobe, significa que está a ser corrigido um caso de *overfitting* e devem ser realizadas mais execuções com a mesma lógica;
- Pelo contrário, se consoante a diminuição do valor no parâmetro referido aumenta o *RMSE* de teste e de treino, é necessário ir incrementando o valor para 15, 16, 17 ou 18;
- Parar de diminuir o valor do parâmetro *yearly_seasonality* quando deixam de haver ganhos no resultado do *RMSE* teste, ou quando é verificado no gráfico que a previsão começa a apresentar *underfitting* (devido à pouca variância das vendas por isso deixa de prever os picos de vendas dos dados históricos);
- É importante analisar a decomposição da série antes de parametrizar o *Prophet*;
- O argumento do *Prophet prior_scale* apresenta melhorias no *RMSE* de teste, porque este impede que sejam aplicados a 100% os movimentos de sazonalidade definidos;
- O melhor cenário é baixar o *RMSE* de teste e diminuir a diferença para com o *RMSE* de treino.

O facto de o *Prophet* ser mais rápido e simples vai ao encontro da solução necessária para aplicação por equipas de TI que trabalham com dados de negócio da saúde. Sendo que métodos de previsão para ajuda na distribuição de medicamentos em farmácias são cada vez mais essenciais e muitas vezes o tempo disponível e alocado a equipa ou profissionais técnicos desta área de negócio é bastante curto. A divulgação do modelo *Prophet* como uma opção viável para métodos de previsão, como demonstrado neste estudo, afigura-se de extrema importância para que as farmácias e empresas do ramo tenham conhecimento deste modelo de previsão e o considerem como uma solução. Desta forma, o trabalho com dados da saúde pode ser complementado fazendo o uso de métodos de previsão inseridos no patamar de ciência de dados.

CAPÍTULO 5

Conclusões e Recomendações

A área da saúde é uma área crucial para a governação e para a população, o que explica o interesse pela realização de um projeto desta natureza, que implicou a colaboração com entidades de negócio.

Os dados disponibilizados pela empresa representaram uma das componentes mais importantes para o desenvolvimento do presente estudo. Os dados disponibilizados enquadram-se para o primeiro estado do *Business Analytics* que consiste na análise descritiva. Esta primeira fase é relativa às análises descritivas de organização dos dados, sendo que infelizmente muitas empresas não possuem ainda esta parte bem desenvolvida. A segunda fase que corresponde às análises preditivas só deve ser iniciada após a conclusão da fase anterior e é nesta fase que a ciência de dados tem um contributo decisivo. Comparativamente com outras áreas, a área da medicina possui avanços significativos nesta fase. Por fim, na terceira fase o objetivo é alcançar análises prescritivas, estas podem ser solucionadas através da automatização de vendas de produtos farmacêuticos e de previsão para apoio em gestão e serviços de farmácias.

Para responder a este desafio de realizar uma análise preditiva através de métodos de previsão em séries temporais, foi feita uma seleção rigorosa e consistente dos dados e efetuada com detalhe a preparação dos mesmos. Paralelamente a estas atividades foram analisados alguns dos modelos de previsão possíveis a serem utilizados com os dados de histórico existentes. Nas fases finais deste projeto foi selecionado o modelo de previsão em séries temporais *Prophet*, sendo este o modelo que foi melhorado e adaptado neste projeto para ser possível alcançar o objetivo de otimização de *stocks*. O modelo *Prophet* aplicado neste projeto de investigação e projeto de ciência de dados com dados reais permite prever as vendas semanais de embalagens de medicamentos para períodos futuros. Os dados resultantes das previsões efetuadas ajudam as farmácias a evitar ruturas de *stock* e mitigam gastos desnecessários de reabastecimento e armazenamento de medicamentos.

Com este projeto foi criado um processo para realização de previsões de vendas que pode ser utilizado para diferentes produtos, farmacêuticos ou não, como, por exemplo, entregas de refeições e produtos ao domicílio. É ainda possível efetuar previsões para grupos de produtos, invés de produtos específicos. Após um aprofundamento do comportamento do *Prophet* para as séries de vendas em causa, foi possível adaptar o modelo para este realizar o ajuste e previsão quantitativa para várias séries de vendas contidas no *dataframe Python* de diferentes produtos.

Bibliografia

1. Ranjan, Jayanthi. (2007). Application of data mining techniques in pharmaceutical industry. Journal of Theoretical and Applied Information Technology.

2. G. Candan, M. F. Taşkin, H. R. Yazgan, "Demand Forecasting in Pharmaceutical Industry Using Artificial Intelligence: Neuro-Fuzzy Approach", Year 2014, Volume 2, Issue 2, 41 - 49, 21.03.2014, https://doi.org/10.17858/jmisci.06816

3. G. Merkuryeva, A. Valberga, A. Smirnov, "Demand forecasting in pharmaceutical supply chains: A case study", Procedia Computer Science, Volume 149, 2019, Pages 3-10, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2019.01.100.

4. "Ministro da Saúde quer mitigar falta de medicamentos nas farmácias " (in PÚBLICO, https://www.publico.pt/2022/11/15/sociedade/noticia/ministro-saude-quer-mitigar-falta-medicamentos-farmacias-2027806 [consultado em 20-10-2023])

5. G. Sasubilli and A. Kumar, "Machine Learning and Big Data Implementation on Health Care data", 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 859-864, doi: 10.1109/ICICCS48265.2020.9120906.

6. B. K. Almentero, J. Li and C. Besse, "Forecasting pharmacy purchases orders", 2021 IEEE 24th International Conference on Information Fusion (FUSION), Sun City, South Africa, 2021, pp. 1-8, doi: 10.23919/FUSION49465.2021.9627017.

7. M. Shashi, "Artificial Intelligence Digital Enablers in Facilitating Demand Forecasting of Biopharmaceutical Supply Chains", Walden University, DOI:10.1729/Journal.30285

8. D. Gruson, T. Helleputte, P. Rousseau, D. Gruson, "Data science, artificial intelligence, and machine learning: Opportunities for laboratory medicine and the value of positive regulation", Clinical Biochemistry, Volume 69, 2019, Pages 1-7, ISSN 0009-9120, https://doi.org/10.1016/j.clinbiochem.2019.04.013.

9. E. Godber, "Uses of Artificial Intelligence in Health", 2018 International Conference on Artificial Intelligence Applications and Innovations (IC-AIAI), Nicosia, Cyprus, 2018, pp. 31-36, doi: 10.1109/IC-AIAI.2018.8674444.

10. "Artificial intelligence must be regulated to stop it damaging humanity, Google boss Sundar Pichai says" (in The Independent, https://www.independent.co.uk/tech/google aiartificial-intelligence-sundar-pichai-regulation-ft-a9292861.html [consultado em 10-01-2023])

11. "EU proposes rules for high-risk artificial intelligence uses" (in The Independent, https://www.independent.co.uk/news/eu-proposes-rules-for-highrisk-artificial-intelligence uses-london-margrethe-vestager-european-commission-b1835096.html [consultado em 10-01 2023])

12. F. Fruggiero, R. Iannone, G. Martino, S. Miranda and S. Riemma, "A forecast model for pharmaceutical requirements based on an artificial neural network", Proceedings of 2012 IEEE International Conference on Service Operations and Logistics, and Informatics, Suzhou, China, 2012, pp. 263-268, doi: 10.1109/SOLI.2012.6273543.

13. Jain, A., Menon, M.N., & Chandra, S. (2015). Sales Forecasting for Retail Chains.

14. J. Renaud, R. Couturier, C. Guyeux, B. COURJAL and C. GIOT, "A Comparative Stud of Predictive Models for Pharmaceutical Sales Data", 2022 2nd International Conference on Computer, Control and Robotics (ICCCR), Shanghai, China, 2022, pp. 191-195, doi: 10.1109/ICCCR54399.2022.9790232.

15. R. Ghousi, S. Mehrani and M. Momeni, "Application of Data Mining Techniques in Drug Consumption Forecasting to Help Pharmaceutical Industry Production Planning", Proceedings of the 2012 International Conference on Industrial Engineering and Operations Management, Istanbul, Turkey, July 3 – 6, 2012

16. C. Qingkui and R. Junhu, "Study on the Demand Forecasting of Hospital Stocks Based on Data Mining and BP Neural Networks", 2009 International Conference on Electronic Commerce

and Business Intelligence, Beijing, China, 2009, pp. 284-289, doi: 10.1109/ECBI.2009.81.

17. A. Burinskiene, “Forecasting Model: The Case of the Pharmaceutical Retail”, Front Med (Lausanne). 2022 Aug 3;9:582186. doi: 10.3389/fmed.2022.582186. PMID: 35991643; PMCID: PMC9381873.

18. Raheel Siddiqui, Muhammad Azmat, Shehzad Ahmed & Sebastian Kummer (2022) A Hybrid demand forecasting model for greater forecasting accuracy: the case of the Pharmaceutical industry, Supply Chain Forum: An International Journal, 23:2, 124-134, DOI: 10.1080/16258312.2021.1967081

19. Rathipriya, R., Abdul Rahman, A.A., Dhamodharavadhani, S. et al. Demand forecasting Model for time-series pharmaceutical data using shallow and deep neural network model. Neural Comput & Applic 35, 1945–1957 (2023). https://doi.org/10.1007/s00521-022-07889-9

20. Moro, Sérgio & Cortez, Paulo & Laureano, Raul. (2011). Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. Proceedings of the European Simulation and Modelling Conference.

Anexos

Anexo A – Visualização do produto utilizado em experiências de preparação de dados

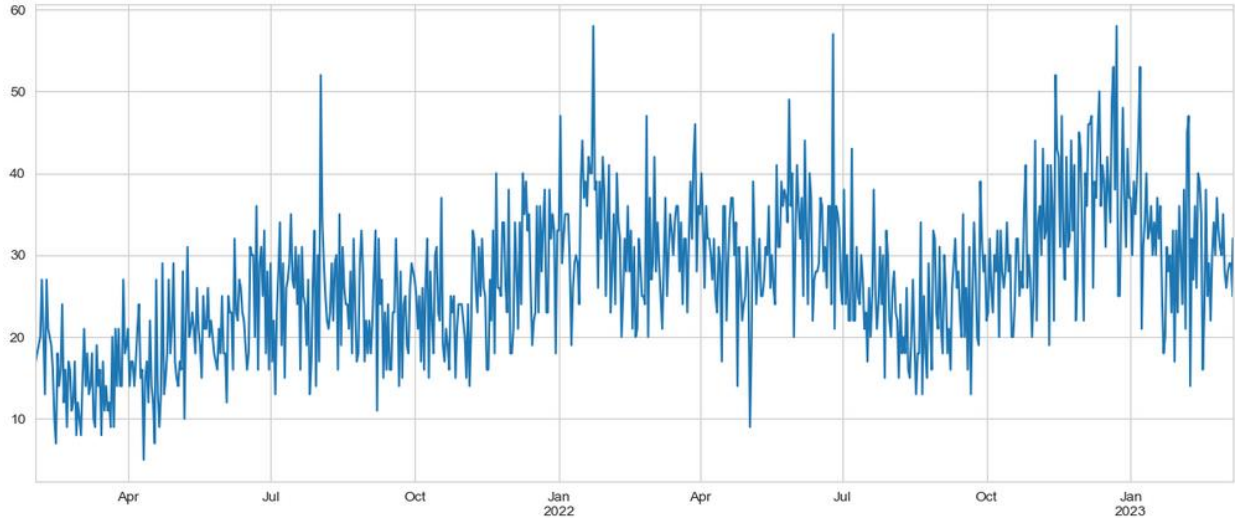

Figura 19 – Gráfico das vendas diárias do "Produto 1".

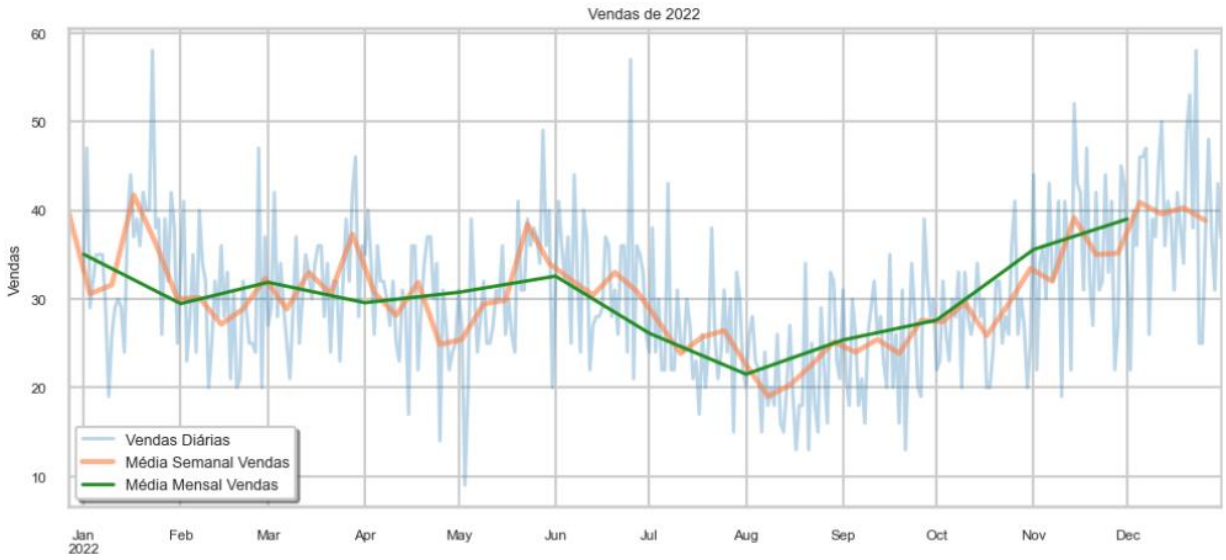

Figura 20 – Gráfico de vendas diárias e da média das vendas semanais e mensais do "Produto 1" para 2022.

Anexo B – Preparação dos dados e modelação para os produtos farmacêuticos finais

```python
import datetime

def get_mondays(date_start, date_end):
    date_start = datetime.datetime.strptime(date_start, "%Y-%m-%d")
    date_end = datetime.datetime.strptime(date_end, "%Y-%m-%d")

    result = []
    while date_start <= date_end:
        if date_start.weekday() == 0:   #0 == Monday
            result.append(date_start.strftime("%Y-%m-%d"))
        date_start += datetime.timedelta(days=1)

    return result

print(get_mondays('2021-01-25', '2021-02-15'))
print(get_mondays('2023-02-25', '2023-03-15'))
```

```
['2021-01-25', '2021-02-01', '2021-02-08', '2021-02-15']
['2023-02-27', '2023-03-06', '2023-03-13']
```

```python
df = df.loc["2021-02-01": "2023-03-05"]
```

Figura 21 – Função Python para verificação dos limites semanais.

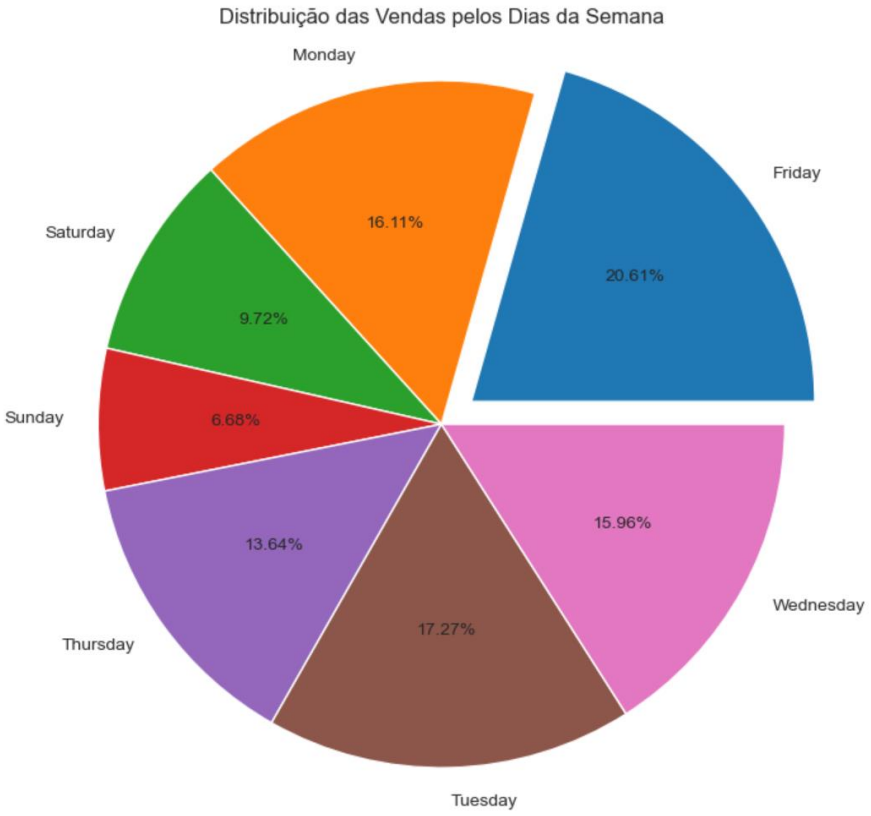

Figura 22 – Pie Chart da distribuição das vendas pelos dias da semana.

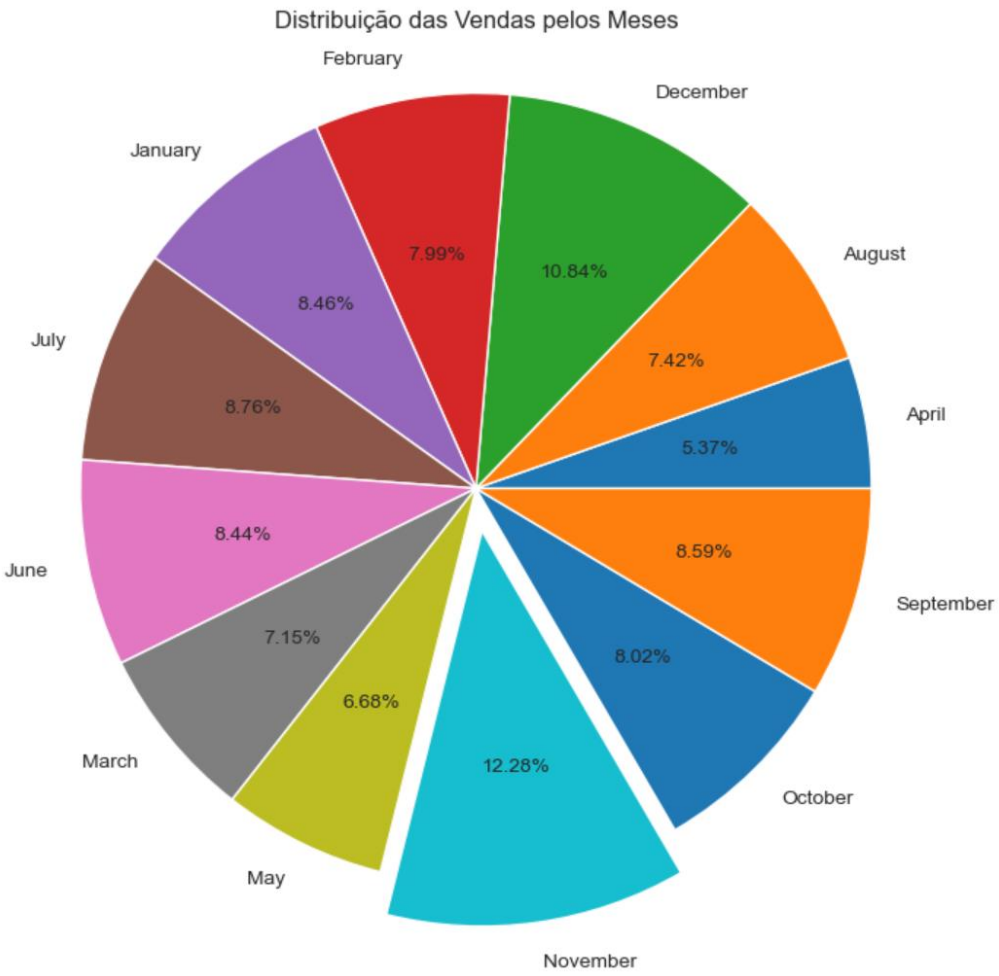

Figura 23 – Pie Chart da distribuição das vendas pelos meses.

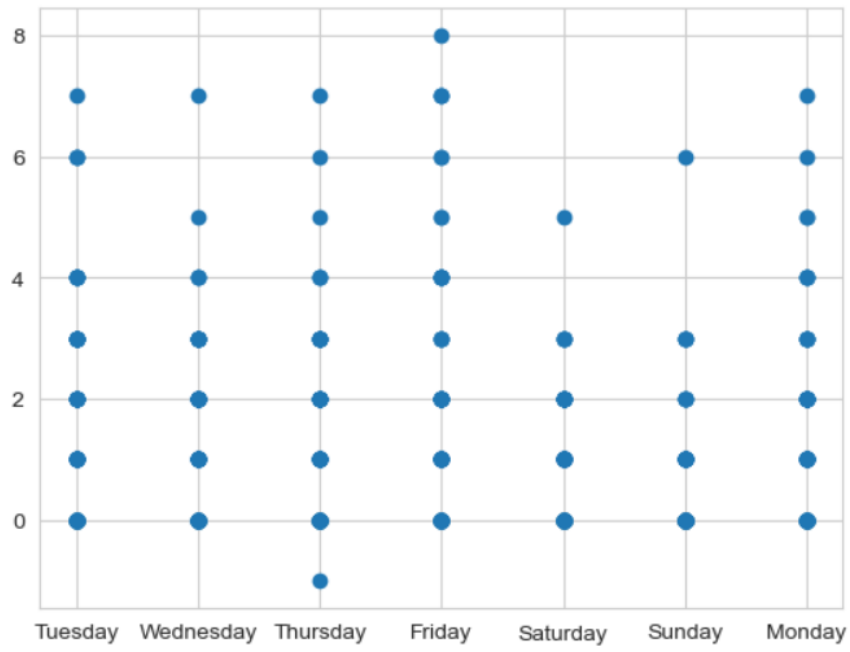

Figura 24 – Scatter Plot da distribuição das vendas pelos dias da semana.

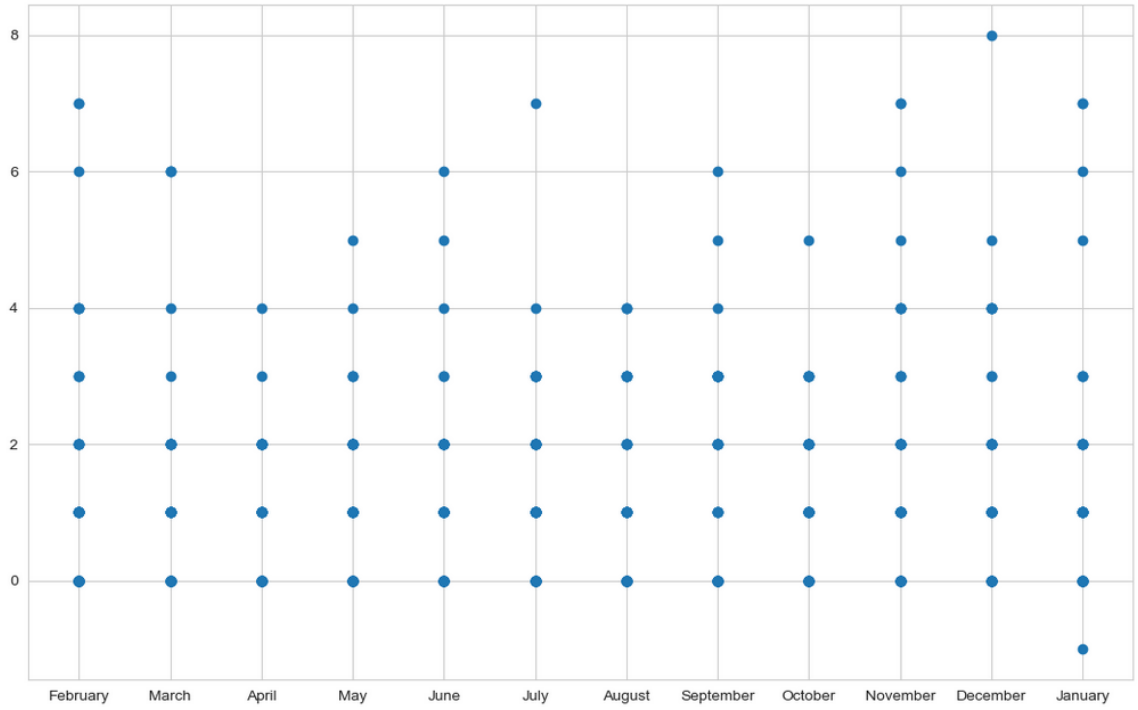

Figura 25 – Scatter Plot da distribuição das vendas pelos meses.

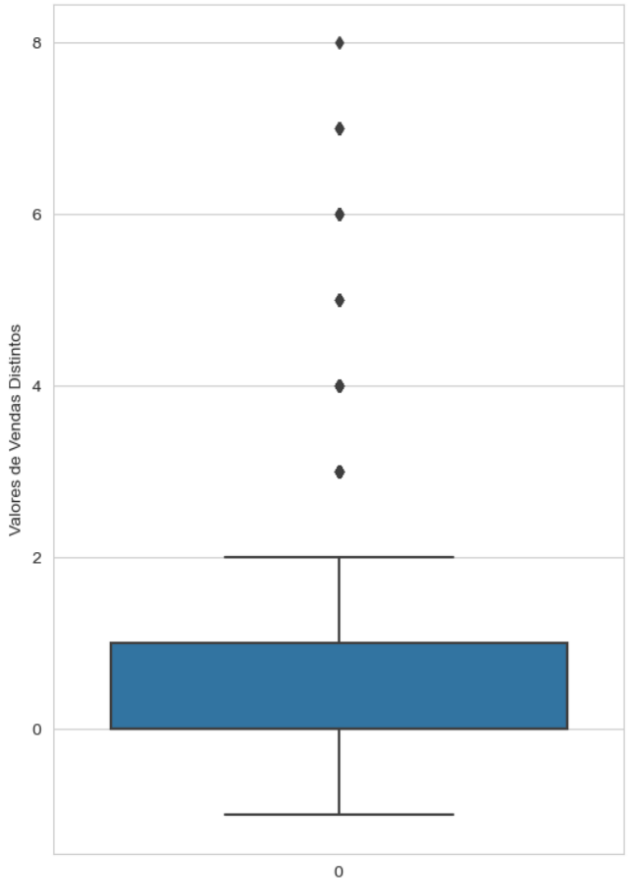

Figura 26 – Gráfico Boxplot.

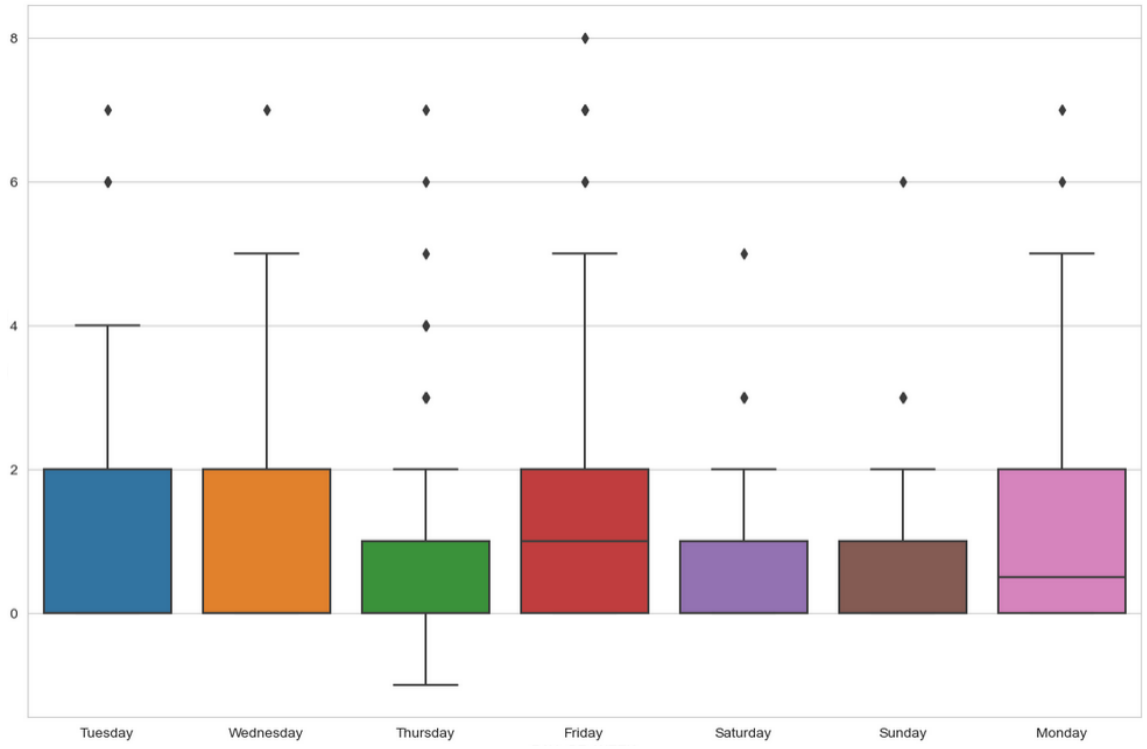

Figura 27 – Gráfico Boxplot para as vendas pelos dias de semana.

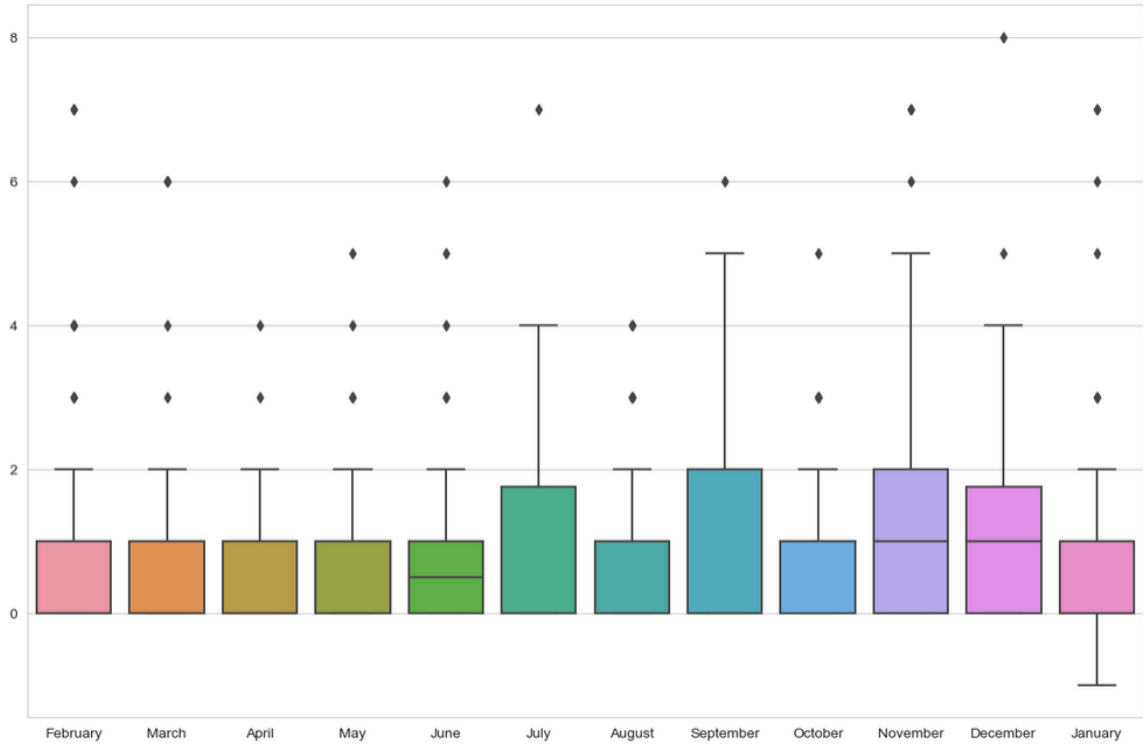

Figura 28 – Gráfico Boxplot para as vendas pelos meses.

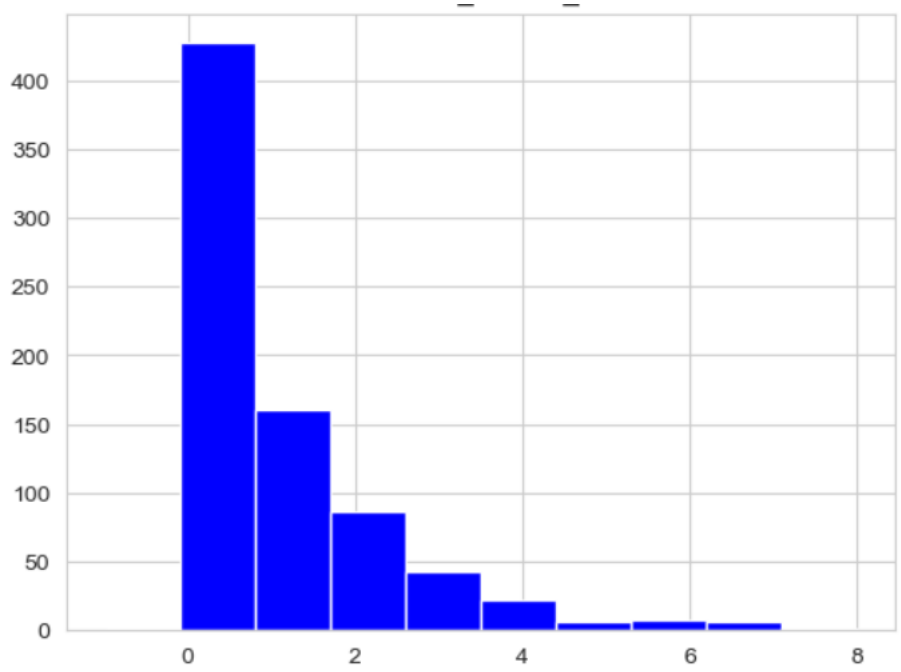

Figura 29 – Gráfico do tipo Histograma.

```
m = Prophet(growth="linear",yearly_seasonality=10, interval_width=0.80)
m.add_seasonality(name='weekly', period=4, fourier_order=5, prior_scale=0.5)
m.fit(train)
future=m.make_future_dataframe(periods=52,freq='W')
forecast=m.predict(future)

forecast.head()
```

	ds	trend	yhat_lower	yhat_upper	trend_lower	trend_upper	additive_terms	additive_terms_lower	additive_terms_
0	2021-02-07	0.170572	2.703663	9.644416	0.170572	0.170572	6.013183	6.013183	6.013183
1	2021-02-14	0.250826	0.328581	7.161988	0.250826	0.250826	3.436210	3.436210	3.436210
2	2021-02-21	0.331080	-2.070852	4.536460	0.331080	0.331080	0.948271	0.948271	0.948271
3	2021-02-28	0.411334	-2.561689	4.656028	0.411334	0.411334	0.560205	0.560205	0.560205
4	2021-03-07	0.491589	-1.038645	5.701309	0.491589	0.491589	1.802448	1.802448	1.802448

Figura 30 – Parâmetros para o modelo Prophet e linhas dos valores resultantes.

yearly	yearly_lower	yearly_upper	multiplicative_terms	multiplicative_terms_lower	multiplicative_terms_upper	yhat
3.913798	3.913798	3.913798	0.0	0.0	0.0	6.183755
1.336825	1.336825	1.336825	0.0	0.0	0.0	3.687036
-1.151114	-1.151114	-1.151114	0.0	0.0	0.0	1.279352
-1.539180	-1.539180	-1.539180	0.0	0.0	0.0	0.971539
-0.296937	-0.296937	-0.296937	0.0	0.0	0.0	2.294036

Figura 31 – Resultado do modelo com os valores preditos na coluna "yhat".

```
fit_models = {}
for stock_line in lines:
    frame = df[df['NOME_MED'] == stock_line].copy()
    frame.drop('NOME_MED', axis=1, inplace=True)
    frame = frame.groupby(pd.Grouper(freq='W')).sum()
    frame = frame.reset_index()
    frame.columns = ['ds', 'y']

    m = Prophet(growth="linear",yearly_seasonality=13, interval_width=0.95)
    m.add_seasonality(name='weekly', period=5, fourier_order=5, prior_scale=0.5)

    model = m.fit(frame)

    fit_models[stock_line] = m
```

Figura 32 – Função for loop para ajuste do Prophet a múltiplos produtos.