Instituto Superior
de Engenharia
Politécnico de Coimbra

DEPARTAMENTO DE ENGENHARIA QUÍMICA E BIOLÓGICA

# Modelo de Gestão e Previsão de Vendas Baseado em Aprendizagem Computacional

Estudo de Caso numa Empresa de Soluções Técnicas para a Indústria

Trabalho de Projeto para a obtenção do grau de Mestre em Engenharia e Gestão Industrial

Autor

**Jorge Filipe Costa Lopes**

Orientadores

**Prof. Doutor Mateus Daniel Almeida Mendes**

**Prof. Doutor José Manuel Torres Farinha**

INSTITUTO POLITÉCNICO
DE COIMBRA

INSTITUTO SUPERIOR
DE ENGENHARIA
DE COIMBRA

Coimbra, junho de 2025

# Resumo

Este trabalho apresenta um modelo de gestão e previsão de vendas aplicado ao contexto específico de uma empresa nacional de engenharia e soluções técnicas industriais, com o objetivo de apoiar as tomadas de decisão estratégicas através de previsões fiáveis e fundamentadas. A organização em estudo tem presença em diversas regiões de Portugal e opera fundamentalmente num contexto B2B, fornecendo uma vasta gama de componentes técnicos e serviços especializados para o sector industrial.

O projeto segue uma abordagem baseada em dados históricos reais de faturação, abrangendo mais de uma década de atividade da empresa, e recorre a técnicas de *Machine Learning* (ML) para a construção de modelos preditivos. A metodologia adotada é a CRISP-DM, que permite estruturar o processo desde a compreensão do negócio até à sua validação e avaliação dos resultados.

No desenvolvimento deste trabalho foram testados cinco modelos preditivos, LSTM, GRU, MLP Regressor, XGBoost e um modelo híbrido, que combinava o LSTM e o XGBoost. A análise comparativa e avaliação sobre os modelos foi realizada com base em métricas como MSE, RMSE, MAE e MAPE. O modelo XGBoost destacou-se como o mais eficaz, apresentando o menor MAPE (16,45%), confirmando o seu desempenho superior em ambientes com grande variabilidade de dados.

O documento discute ainda o impacto do pré-processamento de dados, engenharia de *features* e ajuste de hiperparâmetros, destacando a importância destas etapas para o sucesso dos modelos. A viabilidade da modelação híbrida foi também avaliada, embora para este caso em concreto não tenha superado o modelo base mais robusto.

Os resultados obtidos estão em consonância com evidências presentes na literatura científica, reforçando a aplicabilidade prática do ML em ambientes industriais. É esperado que o modelo proposto se revele uma ferramenta valiosa para apoiar a gestão comercial e estratégica da empresa, contribuindo para um planeamento mais informado e uma resposta mais eficaz às dinâmicas e necessidades de mercado.

**Palavras-chave**: Previsão de Vendas; *Machine Learning*; Modelos Preditivos; CRISP-DM; LSTM; GRU; MLP Regressor; XGBoost

## Abstract

This work presents a sales management and forecasting model applied to the specific context of a national engineering and industrial technical solutions company, with the aim of supporting strategic decision-making through reliable and founded forecasts. The organization under study has a presence in several regions of Portugal and operates mainly in a B2B context, providing a wide range of technical components and specialized services for the industrial sector.

The project follows an approach based on real historical billing data, covering more than a decade of the company's activity, and uses Machine Learning (ML) techniques to build predictive models. The methodology adopted is CRISP-DM, which allows the process to be structured from understanding the business to validating and evaluating the results.

In the development of this work, five predictive models were tested, LSTM, GRU, MLP Regressor, XGBoost and a hybrid model, which combined LSTM and XGBoost. The comparative analysis and evaluation of the models was carried out based on metrics such as MSE, MMR, MAE and MAPE. The XGBoost model stood out as the most effective, presenting the lowest MAPE (16.45%), confirming its superior performance in environments with high data variability.

The document also discusses the impact of data pre-processing, feature engineering and hyperparameter tuning, highlighting the importance of these steps for the success of the models. The feasibility of hybrid modelling was also evaluated, although for this specific case it did not outperform the more robust base model.

The results obtained are consistent with evidence found in scientific literature, reinforcing the practical applicability of ML in industrial environments. The proposed model is expected to be a valuable tool to support the commercial and strategic management of the company, contributing to more informed planning and a more effective response to market dynamics and needs.


**Keywords**: Sales Forecast; Machine Learning; Predictive Models; CRISP-DM; LSTM; GRU; MLP Regressor; XGBoost

# Epígrafe

*Não tenhamos pressa, mas não percamos tempo.*
José Saramago

# Dedicatória

Quero dedicar este documento e o culminar desta fase às pessoas mais importantes da minha vida, aos meus pais e à minha mulher, Jéssica Lopes, que sempre me apoiaram incondicionalmente para que este percurso fosse possível.

# Agradecimentos

Este projeto foi desenvolvido com o objetivo de aplicabilidade empresarial, pelo que o meu primeiro agradecimento não poderia deixar de ser para a empresa alvo de estudo e sua administração, na pessoa do Dr. Alfredo Lima, pelo apoio e incentivo prestados desde o início deste trajeto. Aos Engenheiros Ana Pais e Vasco Lima pelo suporte e colaboração sempre demonstrados.

Aos meus orientadores, Professor Doutor Mateus Mendes e Professor Doutor Torres Farinha pela sempre pronta disponibilidade, motivação e dedicação a este projeto. O seu apoio, orientação pedagógica e rigor científico contribuíram para o meu desenvolvimento pessoal e profissional, tornando este caminho mais enriquecedor e desafiante.

Aos meus pais, pelo seu amor incondicional e por serem os meus melhores amigos, conselheiros, fontes de uma confiança e força incomparáveis, o meu agradecimento especial, amo-os mais do que tudo na minha vida.

Aos meus amigos por toda a motivação e amizade porque, independentemente de por vezes estar ausente, sempre demonstraram o seu apoio para esta etapa da minha vida.

Por último, mas não menos importante, ao meu amor, Jéssica Lopes, a minha companheira de vida, o meu pilar, o meu porto seguro, muitas vezes a minha razão, aquela que incondicionalmente está lá para me incentivar e motivar diariamente para todo e qualquer desafio que enfrente. A ela, o meu desmedido e sentido obrigado, porque sem a sua persistência, motivação e resiliência, nada disto seria possível.

# Índice

# Índice de tabelas

# Índice de figuras

# Lista de siglas e acrónimos

| | |
|---|---|
| ANN | *Artificial Neural Network* |
| B2B | *Business-to-Business* |
| BI | *Business Intelligence* |
| CRISP-DM | *Cross-Industry Standard Process for Data Mining* |
| DM | *Data Mining* |
| EL | *Ensemble Learning* |
| GB | *Gradient Boosting* |
| GRU | *Gated Recurrent Unit* |
| IEEE | *Institute of Electrical and Electronics Engineers* |
| ISEC | Instituto Superior de Engenharia de Coimbra |
| KDD | *Knowledge Discovery in Databases* |
| LSTM | *Long Short-Term Memory* |
| MAE | *Mean Absolut Error* |
| MAPE | *Mean Absolut Percentage Error* |
| ML | *Machine Learning* |
| MLP | *Multi-Layer Perceptron* |
| MLPRegressor | *Multi-Layer Perceptron Regressor* |
| MSE | *Mean Squared Error* |
| OHE | *One-Hot Encoding* |
| RMSE | *Root Mean Squared Error* |
| RNN | *Recurrent Neural Network* |
| SAS Institute | *Statistical Analysis System Institute* |
| SEMMA | *Sample, Exploration, Modification/Manipulate, Model and Assessment* |
| XGBoost | *Extreme Gradiante Boosting* |

# 1 Introdução

## 1.1 Enquadramento do tema

O mundo de hoje está cada vez mais evoluído, mais industrializado e com um desenvolvimento tecnológico já algo distante daquilo que se pensaria na primeira revolução industrial.

Com a revolução industrial veio a revolução tecnológica, fruto da constante procura por soluções mais eficazes e eficientes nos diferentes setores. Com isso, foram surgindo ferramentas muito poderosas que têm vindo a dar um contributo muito forte para o crescimento das organizações e da indústria. Associado a este crescimento temos também o aumento da competitividade e da concorrência entre as empresas, levando aqui vantagem aquelas que maior capacidade de adaptação ao mercado apresentem e cujas soluções melhor colmatem as necessidades da procura. As organizações que se distinguirem nesta vertente poderão prosperar nas suas áreas de atuação e, assim, alcançar o sucesso.

A globalização é um fenómeno que tem vindo a ganhar cada vez maior destaque no que ao setor industrial diz respeito. Se, por um lado, é positiva porque abrange um largo leque de opções para que as organizações possam socorrer as suas necessidades, por outro tem vindo a afetar determinados sectores, nomeadamente os que estão relacionados com cadeias de abastecimento, fornecimento de material e serviços. A realidade de hoje em dia é que a competição entre empresas para a obtenção de maior quota de mercado e lucros transitou de um nível local para uma competição global. O mercado atual é altamente competitivo, tendo aqui um forte contributo a quantidade de informação disponível, o avanço da tecnologia industrial e a globalização da oferta e da procura [1].

No atual ambiente competitivo a que o mundo dos negócios está sujeito, os clientes têm a possibilidade de escolher sobre várias opções, ponderando diversos fatores quando se trata de decidir sobre os seus fornecedores, mediante as suas necessidades. Se o cliente estiver satisfeito com o serviço, a tendência normal passa por trabalhar com os fornecedores com os quais já tem uma relação criada, já há confiança. No entanto, quando os clientes estão insatisfeitos, há uma clara tendência para que mudem o serviço e, consequentemente, o seu fornecedor, causando assim algum impacto sobre aquele que é um dos principais objetivos de uma organização, a sua lucratividade [2].

É, por isso, fundamental que os decisores possam delinear as melhores estratégias e to-

mar as melhores decisões, que sendo devidamente fundamentadas podem contribuir e muito para o sucesso das suas empresas. Grandes organizações estão normalmente associadas a grandes quantidades de dados, sendo por isso preponderante uma reflexão ponderada sobre toda essa informação. Só assim se pode transformar essa informação numa vantagem competitiva, num conhecimento útil para o negócio, contribuindo para a definição de estratégias que permitam não só fidelizar os clientes já existentes como atrair novos [3].

## 1.2 Empresa

A Organização alvo de caso de estudo é uma empresa nacional quase centenária, especialista em soluções industriais. A sua estrutura é composta por sete filiais, dentro das quais integra duas oficinas técnicas e um gabinete de engenharia, trabalhando uma vasta gama de marcas *premium*, nomeadamente ao nível de rolamentos, movimento linear, transmissão de potência mecânica, lubrificação e sistemas de vedação, ferramentas e máquinas, manipulação de fluídos e serviços de reparação, entre outros produtos complementares. O grupo fornece serviços técnicos e logísticos ajustados à real necessidade dos seus clientes, orientando-os para a inovação, tendo por base um profundo conhecimento sobre as suas áreas de atuação.

## 1.3 Objetivos

O objetivo do projeto apresentado passa por desenvolver uma metodologia que proporcione aos gestores da empresa um forte auxílio na definição de estratégias que lhe sejam mais adequadas. Estas decisões terão por base um modelo de previsão de vendas indexado ao mercado nacional, especialmente sobre a sua área de atuação e do seu setor, que fará uma previsão sobre as suas vendas futuras considerando o histórico de faturação da empresa e determinados *insights* decorrentes de possíveis padrões temporais identificados.

Com o desenvolvimento deste projeto é esperado o avanço na compreensão sobre o impacto de determinados fatores no comportamento do mercado, das vendas e o desenvolvimento de um modelo de previsão robusto, capaz de identificar e antecipar determinadas tendências e outros fatores que poderão influenciar as vendas. A principal meta a atingir passa pela implementação do modelo na aplicação prática da gestão estratégica da empresa, nomeadamente ao nível de vendas. Assim, são proporcionadas tomadas de decisão estruturadas e devidamente fundamentadas sobre dados o que criará, à partida, uma vantagem competitiva à organização perante a sua principal concorrência.

## 1.4 Metodologia

O desenvolvimento do modelo baseou-se em técnicas avançadas de *Machine Learning* e análise estatística, seguindo a construção do modelo uma metodologia CRISP-DM. Os dados foram providenciados por uma fonte, neste caso específico a organização sobre a qual se realiza o estudo, incluindo históricos de vendas, indicadores económicos e tendências de vendas consoante as zonas de atuação. A validação do modelo ocorreu por intermédio de comparações com métodos convencionais, recorrendo para tal a determinadas métricas de avaliação específicas para modelos preditivos.

## 1.5 Questões de investigação

Para o estudo em causa, foram propostas as seguintes questões de investigação:

1. A partir da análise do histórico da empresa (dados de faturação) é possível compreender as suas dinâmicas e evolução ao longo do tempo? Conseguem identificar-se padrões claros na evolução das vendas?

2. Qual o melhor modelo preditivo de vendas que leve em consideração o histórico de vendas da empresa? Qual será o desempenho preditivo de diferentes modelos de *Machine Learning* na previsão de vendas da empresa, tendo por base o seu histórico de faturação?

3. É possível identificar e incorporar variáveis relevantes no modelo de previsão de vendas? Essas variáveis podem influenciar a performance dos modelos de previsão de vendas?

4. Qual o impacto que poderá ter o ajuste de hiperparâmetros e o pré-processamento de dados na precisão dos modelos de previsão?

5. A combinação de modelos pode melhorar os resultados das previsões face aos modelos individuais?

## 1.6 Estrutura do relatório

No primeiro capítulo deste documento foi realizado o enquadramento teórico sobre o projeto, explicada a sua finalidade e apresentada a empresa em estudo sobre a qual se irá desenvolver o modelo. Nesta etapa foram ainda explicadas as metodologias base usadas para o desenvolvimento do processo e as respetivas questões de investigação que serviram como pilares para a formulação do problema/solução.

No segundo capítulo do projeto é apresentado o estado da arte sobre o tema em estudo, recorrendo para tal a quatro dos principais motores de busca disponíveis para a comunidade científica, Web of Science, Scopus, IEEE e Google Scholar. Neste ponto é

também explicada a importância da previsão de vendas, alguns tipos e modelos de previsão, nomeadamente redes neuronais e *Gradient Boosting* (GB), realçando com maior detalhe aqueles que são utilizados no projeto. Neste capítulo são ainda apresentados os métodos de avaliação que serão usados mais à frente para avaliação e posterior validação dos modelos desenvolvidos.

No capítulo seguinte, o terceiro, é realizada uma breve abordagem sobre as metodologias mais comuns para análise de dados e explicada detalhadamente a metodologia usada para o desenvolvimento do modelo, CRISP-DM, e as diferentes etapas que a constituem. Nesta divisão são ainda referidas as principais ferramentas utilizadas para análise e o desenvolvimento do projeto na integra, desde a análise de dados até ao desenvolvimento dos modelos.

O capítulo quatro descreve de forma pormenorizada as etapas de desenvolvimento do projeto, desde a análise e compreensão dos dados até à implementação e evolução dos modelos preditivos. Este capítulo aborda ainda o processo iterativo que decorreu na procura de melhoria dos modelos, incluindo o incremento de variáveis temporais, o ajuste de hiperparâmetros e a necessidade de um pré-processamento e filtragem dos dados. Neste ponto estão apresentados os motivos que justificaram algumas decisões e ponderações ao longo projeto e contextualiza a sua trajetória de evolução técnica e metodológica.

No quinto capítulo são discutidos os resultados obtidos, analisadas e respondidas as questões de investigação que estiveram na base do desenvolvimento deste projeto e realizada uma comparação entre os resultados alcançados e a literatura. Aqui, é também feita uma reflexão sobre os valores alcançados para os testes de avaliação para que se possa perceber a veracidade dos valores obtidos e a sua taxa de precisão.

Por fim, são apresentadas as conclusões sobre o projeto e discutidos sinteticamente os principais resultados obtidos e conhecimento adquirido. Ainda nesta secção podem ser verificadas as implicações práticas que este projeto poderá ter para a organização em estudo. Foi também realizada uma reflexão sobre as limitações verificadas no decorrer do projeto e sobre trabalhos que poderão ser desenvolvidos no futuro para tentar melhorar os modelos.

# 2 Estado da arte

Para qualquer organização, o objetivo de atuação no mercado define o planeamento operacional futuro a ser feito [1].

A estratégia é um tema bastante utilizado e debatido há largos séculos, sendo parte integrante e fundamental para o sucesso de qualquer unidade, seja ela de que área for. Para a definição das melhores estratégias é indispensável uma tomada de decisão devidamente fundamentada e que possibilite a escolha das melhores alternativas para aquilo que se prevê ser o melhor para uma organização. Para tal, são necessárias ferramentas, métodos e modelos matemáticos que possibilitem suportar essas decisões e, consequentemente, auxiliar a gestão de topo a delinear os melhores caminhos para as suas empresas.

## 2.1 Enquadramento global

Para iniciar este projeto foi necessário realizar uma revisão da literatura sobre os trabalhos efetuados dentro desta área, que envolvessem as *keywords* métodos de previsão e previsão de vendas. Estas pesquisas foram efetuadas através de plataformas bastante reconhecidas na área da investigação, nomeadamente a Scopus, na Web of Science, Google Scholar e ainda IEEE. Neste contexto, numa primeira fase foram analisados os títulos dos resultados obtidos nas pesquisas efetuadas e numa segunda fase os seus resumos, tendo a escolha recaído sobre aqueles que se adaptavam mais ao principal objetivo deste projeto.

Silva [3] utilizou diferentes modelos de previsão de vendas numa empresa líder do setor de retalho não alimentar a nível nacional. O objetivo passava por apresentar um modelo que fosse capaz de prever as vendas líquidas mensais da empresa por loja, não uma previsão para o total da loja mas sim para um determinado conjunto de produtos, neste caso específico, unidades de negócio. A autora recorreu a modelos lineares e não lineares para realizar o seu estudo, confrontando depois os respetivos resultados para verificar quais os que apresentavam menor erro. Analisando os resultados obtidos verificou que os métodos não lineares, de entre os quais as redes neuronais artificiais, apresentam melhores resultados do que os métodos considerados mais tradicionais, os modelos lineares e, por isso, resultados mais fidedignos. No entanto, para a construção do seu modelo final, optou por usar um método que engloba os vários modelos que testou para o seu projeto, um modelo múltiplo. No final do seu estudo,

e por forma a construir um modelo mais consistente nas previsões, a autora combina o modelo até então apresentado com o método de consideração dos diferentes níveis hierárquicos dos dados, o que levou à obtenção de erros significativamente inferiores quando comparados com as tradicionais abordagens, obtendo desta forma as taxas de erro desejadas pela empresa [3].

Almeida e Passari [4] realizaram também um estudo sobre métodos de previsão de venda para o retalho, recorrendo igualmente a redes neuronais artificias, comparando depois os resultados obtidos com os de outros modelos de previsão. Os autores identificaram que existia uma lacuna relativamente aos estudos realizados até então, que se prendia com a previsão de vendas de um produto tendo em consideração o impacto da procura simultânea de outros produtos relacionados com a sua venda, optando por explorar essa área. Foi observado que o uso de modelos de redes neuronais artificiais, em particular redes multicamada e o método de retropropagação proposto por Rumelhart, Hinton e Williams em 1986 era mais eficaz, apresentando uma maior precisão para a previsão de vendas a curto prazo do que outros métodos frequentemente utilizados, nomeadamente técnicas de modelação Naïve e de regressão linear. Apesar disso, os autores ressalvam que os erros verificados ainda são elevados para a previsão a curto prazo, o que certamente poderia representar valores de ainda maior desvio para horizontes temporais mais amplos, destacando as melhorias significativas que ainda se poderão fazer nesta metodologia e técnica de previsão [4].

Ponte [5] desenvolveu modelos previsão de vendas de produtos de uma marca de seguros de saúde recorrendo a métodos de regressão linear múltiplos. O estudo tinha como principal objetivo a criação de modelos de previsão de vendas para seguros de saúde que pudessem prever futuras vendas de produtos já existentes e ao mesmo tempo ser adaptados ao lançamento de futuros, por forma a auxiliar nas estratégias de marketing e na elaboração de orçamentos. Os modelos desenvolvidos apresentaram resultados satisfatórios, uma vez que os valores previstos foram adequadamente próximos aos valores reais depois verificados, apresentado apenas uma diferença notória no ano da pandemia mundial que desfalcou os valores obtidos, visto que se verificou uma redução nas vendas dos seguros. No entanto, os modelos criados foram aceites e considerados fidedignos para utilização futura [5].

Rumbe *et al.* [1] apresentaram uma comparação entre dois métodos distintos de previsão de vendas. Para o seu projeto, os autores pretendiam obter o modelo de previsão de vendas ideal para tendas comerciais e para tal consideraram três modelos de tendas vendidos por uma empresa americana. De acordo com o histórico de vendas obtido na empresa, os autores verificaram que para além de uma queda nas vendas, havia ainda uma sazonalidade associada aos valores, o que os motivou para desenvolverem a sua pesquisa recorrendo ao modelo Holt- Winters e a uma Rede Neuronal Artificial, a de retropropagação. Depois de analisados os dados referentes à aplicação dos dois mode-

los, os autores concluíram que o modelo de redes neuronais artificias é mais vantajoso que o modelo de Holt-Winters uma vez que apresentou resultados mais fidedignos e de precisão superiores, como comprovado pelos valores de erro associados a cada um deles [1].

Na pesquisa desenvolvida por Eiglsperger *et al.* [6], os autores apresentaram uma ampla gama de abordagens para o desenvolvimento de um modelo de previsão de vendas para produtos hortícolas. No seu estudo foram consideradas técnicas de previsão clássicas e ainda técnicas de previsão mais recentes, baseadas em *Machine Learning*, recorrendo para tal a um conjunto de dados cedidos por seis empresas alemãs da área em estudo. Posteriormente, os autores procederam à análise de valores obtidos, tendo observado que os modelos baseados em ML apresentaram melhor desempenho do que os modelos clássicos, apesar do desempenho dos modelos depender bastante do conjunto de dados específicos em análise. Os mesmo autores consideraram o método XGB como o melhor dos utilizados no seu estudo e deram ainda especial destaque ao modelo SARIMAX, método que, segundo os mesmos, pode ser considerado como híbrido e que apresentou um desempenho notável nas previsões que fez [6].

No trabalho de Yuan e Lee [7] foram realizados testes entre modelos de *Machine Learning* para prever o volume de vendas do setor automobilístico em Taiwan. Os autores recorreram a modelos de redes neuronais artificiais e um modelo de regressão por vetores de suporte otimizado por um algoritmo genético usado para melhorar os parâmetros do modelo de regressão, sendo capaz de obter soluções ótimas em pouco tempo. O estudo tinha como principal objetivo saber de entre os modelos selecionados qual é que apresentava maior precisão, baseando-se para tal no valor do erro absoluto médio em percentagem, o MAPE. De acordo com os valores obtidos, foi concluído que o modelo de regressão combinado com o algoritmo genético apresentava melhores resultados do que os restantes modelos em estudo, sendo por isso considerado o melhor para este caso [7].

Osawa *et al.* [8] quiseram desenvolver um modelo que lhes permitisse realizar previsões de vendas tendo em consideração não apenas variações sazonais mas também flutuações cíclicas. Os autores optaram por recorrer ao modelo de média móvel de soma autoregressiva, ARIMA tradicional, e comparar os resultados obtidos com um modelo ARIMA ao qual foram adicionados fatores periódicos, mais concretamente flutuações circulares, procurando assim obter um modelo de previsão mais preciso e robusto do que os modelos convencionais. Para o seu estudo utilizaram dados de três anos recolhidos de um site de comércio de informática e tecnologia, incluindo 70 categorias de produtos com vendas regulares. Após a análise dos valores obtidos, verificou-se que o modelo ARIMA com incorporação de flutuações circulares (F-ARIMA) obteve melhores resultados que o ARIMA convencional, o que revela que os modelos de previsão devem considerar os fatores periódicos para obter resultados mais precisos, sobretudo

para dados que revelam padrões cíclicos fortes. Apesar disso, é importante salientar que o modelo apresenta algumas limitações sobretudo quando sujeito a alterações bruscas e inesperadas [8].

Wu *et al.* [9] apresentam uma análise sobre diferentes métodos de previsão de vendas para a indústria de consumíveis eletrónicos. Os autores debateram-se com algumas limitações base, desde logo porque realizaram um estudo sobre previsões de vendas para novos produtos, o que pode representar uma procura ascendente numa fase inicial após o seu lançamento; depois porque o tipo de produtos em estudo pode apresentar sazonalidade e, por último, porque tratando-se estes produtos de um tipo de consumíveis com um ciclo de vendas normalmente compreendido entre dois e três anos, os poucos dados de que se dispõe pode ser um grande desafio para previsão de séries temporais. Neste estudo são utilizados modelos de previsão de vendas de séries temporais tendo por base apenas dados históricos e outros que adicionam ainda o impacto que poderão causar determinados fatores de entrada, ou fatores causais, como cotações, fatores de sazonalidade e até preços dos produtos. Os dados históricos são referentes a dois anos de vendas de uma empresa do setor de consumíveis eletrónicos e os dados que servem para usar como fatores de entrada são referentes a valores de cotações da empresa em vinte e sete meses. Depois de realizadas as previsões com os diferentes modelos em estudo, os autores verificaram que o método mais preciso era o ARMAV com tendência linear, sendo por isso considerado o mais indicado para fazer a previsão do volume de vendas de produtos novos com poucos dados históricos. Este modelo interpreta não só a tendência de crescimento como também a influência que poderão causar os fatores de entrada e consegue modelar melhor as relações de entrada e saída do que os modelos de regressão múltipla [9].

Hewage e Perera [10] apresentam uma perspetiva sobre a capacidade dos modelos de previsão baseadas em *Machine Learning* para a previsão de vendas no retalho em épocas promocionais e pós promocionais. Os autores recorreram a dados facultados por um retalhista dos Estados Unidos da América, referentes às vendas realizadas sobre determinados produtos durante 156 semanas em 75 lojas. O objetivo passou por perceber de entre modelos univariados e de ML, quais aqueles que apresentavam melhores resultados para a previsão de vendas em períodos promocionais e quais os modelos que conseguiam prever o declínio pós promocional, facto que normalmente se verifica após uma atividade promocional sobre qualquer produto. Para além de realizarem o estudo com os métodos considerados normais, os autores optaram ainda por verificar se incluindo parâmetros e variáveis adicionais aos modelos, tanto para os univariados como os de ML, os resultados poderiam demonstrar maior precisão. De acordo com os resultados obtidos, pode verificar-se que os modelos baseados em ML e com a inclusão de características adicionais (nomeadamente os períodos promocionais) apresentaram maior precisão nas previsões apresentadas, já que conseguem identificar bem a dimen-

são e magnitude do declínio nos períodos pós promocionais, levando assim a concluir que estes modelos podem ser uma ferramenta eficaz para o planeamento de futuras promoções [10].

Na pesquisa de Guo *et al.* [11], há uma preocupação dos autores em fazer um trabalho diferente do que havia sido feito até então para a previsão de vendas de retalho. Até à data da realização do seu estudo, os trabalhos realizados nesta área de previsão tinham em consideração os dados históricos de vendas dos produtos a serem previstos ou, para o caso de previsão de vendas de novos produtos, eram considerados históricos de vendas de produtos semelhantes. Até ao momento, apenas um autor havia realizado um estudo diferente sobre as previsões de vendas no retalho, recorrendo para tal aos valores das vendas iniciais para realizar a previsão do total de vendas. No entanto, este autor não considerou múltiplos fatores que poderiam influenciar o total de vendas e por isso não conseguia prever as alterações que poderiam resultar desses fatores. Os autores colmataram essa lacuna e realizaram a previsão do volume de vendas de produtos tendo por base as suas vendas iniciais numa unidade de vendas, considerando fatores influenciadores como o tipo de produto, o preço de venda, índices económicos, climáticos, entre outros. Os dados do estudo foram recolhidos de uma grande empresa de retalho de moda da China continental durante aproximadamente dois anos, foram tratados e posteriormente aplicados sobre eles modelos multivariados e o modelo proposto pelos autores, primeiro utilizando todas as variáveis de entrada e depois utilizando apenas as variáveis selecionadas. Após a análise aos resultados obtidos, pôde verificar-se que todos os modelos apresentaram melhores índices de previsão com o incremento apenas das variáveis selecionadas, sendo o modelo proposto pelos autores aquele que se destacou e apresentou maior precisão [11].

No trabalho de Giri *et al.* [12] os autores procuraram desenvolver um método que lhes permitisse a realização de previsões sobre a procura de produtos para o retalho, mais especificamente do setor de moda feminina, utilizando para tal uma abordagem que combinava regressão não linear e *Deep Learning*. Neste tipo de indústria consegue-se obter uma grande quantidade de dados, não só sobre as características dos produtos mas também sobre o seu público-alvo, os seus clientes. No início do estudo observaram a existência de dois pontos que seriam basilares para o desenvolvimento do seu modelo de previsão, os dados de imagem que possuem características estéticas sobre o produto, e os dados de vendas que indicam *insights* fundamentais sobre a sua procura. Esta combinação de dados, aliada aos modelos que utilizaram no seu estudo, modelos de *Deep Learning* e redes neuronais de regressão não linear, permitiu-lhes a obtenção de valores que consideraram como aceitáveis para a previsão das vendas de produtos de moda, comprovados pelos métodos de avaliação utilizados, sobretudo pelo RMSE e o MAE. Apesar dos resultados obtidos, os autores realçaram ainda que uma melhoria futura para o seu estudo seria o incremento de dados para treino, imagens e dados

de vendas, para que desta forma pudessem apresentar melhores resultados e com isso melhorar a *performance* do seu modelo [12].

No trabalho de Hicham *et al.* [13] é desenvolvido um modelo híbrido de previsão de vendas tendo como base redes neuronais de retropropagação com taxa de aprendizagem adaptativa e agrupamento fuzzy com o objetivo de realizar a previsão do volume de vendas de placas de circuito impresso (PCB) de uma empresa. Até à data de realização deste estudo, uma boa parte dos modelos que usavam agrupamento recorriam a métodos de agrupamento clássicos, nomeadamente rígidos, em que cada dado pertence a um único *cluster*. No modelo fuzzy utilizado pelos autores, o fuzzy c-means, os dados podem pertencer a mais do que um *cluster*, verificando-se depois os níveis de pertença de cada dado a cada *cluster*. Tal facto permite que um dado possa pertencer a um determinado *cluster* até certo ponto, levando a que esses *clusters* sejam maiores e, consequentemente, se consiga obter maior precisão nos resultados obtidos. Para treinar o seu modelo, os autores utilizaram dados históricos das vendas de três anos de uma empresa eletrónica de Taiwan e, para o testar, recorreram a dados de um ano. De acordo com os resultados obtidos através do seu modelo, e comparando os mesmos com resultados obtidos recorrendo a outros modelos de previsão de vendas mais tradicionais, os autores puderem verificar que o seu modelo superou os restantes em teste demonstrando resultados mais precisos [13].

Tabela 2.1: Resumo sobre trabalhos realizados na área de modelos de previsão de vendas

| **Autor** | **Problema** | **Métodos** | **Resultados** |
|---|---|---|---|
| Eiglsperger *et al.* (2024) [6] | Previsão para vendas de produtos hortícolas | Métodos estatísticos; Métodos de ML Clássico; Métodos de ML Profundo | Os métodos de ML foram os mais eficientes - XGB foi o que apresentou melhores resultados |
| Rumbe *et al.* (2024) [1] | Previsão de vendas de tendas | Modelo de Holt-Winters; Redes Neuronais Artificiais (Rede neuronal de retropropagação) | Modelo de Redes Neuronais artificiais foi melhor que o de Holt-Winters – previsões muito mais precisas |
| Maria Ponte (2022) [5] | Previsão de vendas de seguros de saúde | Regressão Linear Múltipla (método dos mínimos quadrados e método de seleção de variáveis backward) | O modelo usado foi considerado satisfatório |

*Continua na próxima página*

Tabela 2.1 – *Continuação da tabela*

| **Autor** | **Problema** | **Métodos** | **Resultados** |
|---|---|---|---|
| Osawa *et al.* (2021) [8] | Previsão de vendas que considere variações sazonais e cíclicas | ARIMA convencional e ARIMA considerando as flutuações circulares (F-ARIMA) | O modelo F-ARIMA obteve resultados de previsão mais precisos |
| Hewage e Perera (2021) [10] | Previsão do volume de vendas no retalho em épocas de promoções | Métodos estatísticos; Métodos de ML Clássico | Os Modelos de ML Clássico apresentaram melhores resultados |
| Giri *et al.* (2019) [12] | Prever a quantidade de novos produtos de vestuário feminino | Inception V3 (Rede neuronal profunda) e MLP (Rede neuronal- Multilayer Perceptron) | O modelo apresentado obteve resultados aceitáveis |
| Silva (2015) [3] | Previsão de vendas para produtos ou unidade de negócio do retalho não alimentar | Métodos estatísticos; Métodos de ML clássico; Métodos de ML Profundo | Modelo múltiplo apresentou melhores resultados, sobretudo quando combina as previsões dos diferentes níveis hierárquicos |
| Wu *et al.* (2012) [9] | Previsão de vendas de consumíveis eletrónicos | Métodos estatísticos | O modelo ARMAV com tendência linear apresentou maior precisão |
| Guo *et al.* (2012) [11] | Previsão de vendas de produtos no retalho | IELM; GLM; MID | O modelo MID revelou melhor resultados |
| Hicham *et al.* (2012) [13] | Modelo híbrido para previsão de vendas de placas de circuito impresso (PCB) | Modelos estatísticos; Modelos de ML clássico; Modelos de ML profundo | O modelo desenvolvido (Fuzzy + FCBPN) apresentou melhores resultados |

*Continua na próxima página*

Tabela 2.1 – *Continuação da tabela*

| **Autor** | **Problema** | **Métodos** | **Resultados** |
|---|---|---|---|
| Yuan e Lee (2011) [7] | Previsão do volume de vendas para a indústria automóvel em Taiwan | Métodos estatísticos; Modelos de ML clássico | O Modelo de Regressão por vetores de suporte, otimizado pelo algoritmo genético (GA-SVR) foi o mais preciso |
| Almeida e Passari (2006) [4] | Previsão de vendas para o retalho não alimentar | Modelos de ML clássicos; Naïve | O modelo de redes neuronais teve um melhor desempenho |

*Fim da tabela*

## 2.2 Previsão de vendas

A história das previsões de vendas remonta há já mais de 50 anos e abrange uma enorme variedade de aplicações no mundo industrial [11].

Para qualquer organização, a previsão do volume de vendas é considerada como uma capacidade organizacional crítica, sobretudo no que respeita ao planeamento estratégico, tático e operacional de uma empresa [7].

Num mercado que se apresenta cada vez mais dinâmico e extremamente competitivo é fundamental que as empresas otimizem as suas estratégias e que se tomem decisões fundamentadas para impulsionar o seu crescimento [14].

Atualmente, a concorrência industrial intensifica-se a um ritmo avassalador, o que leva as empresas a uma incessante procura por coisas novas, que as destaquem e que lhes permitam obter vantagens competitivas [15].

O progresso de uma organização pode ser determinado por vários fatores, nomeadamente o valor das suas vendas anuais e o lucro obtido no ano fiscal operacional [1].

Para uma organização, a receita gerada pelas vendas pode ser uma importante fonte de lucro [16]. A previsão de vendas ganha por isso um destaque preponderante para aquela que deve ser a atenção dos gestores de topo, uma vez que, quanto mais precisa for, maior será a eficiência operacional, mais otimizada será a gestão de inventários e menor serão os custos associados à cadeia de abastecimento [16].

A previsão de vendas é um dos pilares que está na base da construção dos planos da maioria das organizações [11]. Realizar esta tarefa de forma fiável é crucial para a tomada de decisões empresariais, uma vez que, quão mais precisa for esta previsão,

maior será o beneficio que as organizações poderão alcançar [13].

As organizações tendem a adotar diferentes estratégias para sondar os níveis de procura no mercado [1]. No entanto, a previsão de vendas é um desafio para as empresas dos mais diversos setores económicos, já que existe uma enorme volatilidade na procura, que depende de múltiplos fatores [17]. "A variabilidade associada à procura é força motriz por trás da necessidade de previsão" [1].

O principal objetivo da previsão de vendas é estimar um volume de vendas futuras tendo em consideração determinados padrões e tendências, possibilitando às empresas uma melhor gestão sobre o seu inventário, níveis de produção, estratégias de marketing e vendas [14].

Todo o processo associado à previsão do volume de vendas de uma qualquer organização é bastante crítico e, portanto, uma área nada fácil de gerir [7].

A previsão de vendas está a tornar-se um processo cada vez mais difícil e complexo para as organizações por múltiplos fatores, nomeadamente a elevada concorrência de mercado, o marketing extremamente agressivo, a enorme quantidade de produtos e soluções de mercado, os ciclos de vida mais reduzidos dos produtos, entre outros [10].

Se, por um lado, se pode considerar o processo de previsão de vendas como essencial para a maioria das organizações, já que pode ser encarado como uma ferramenta de gestão muito importante que auxilia na definição de planos de marketing, orçamentação, planos de produção, estratégias de publicidade e promoção, entre outros; por outro lado, deve assumir-se como sendo um processo difícil, uma vez que, por maior que seja a qualidade dos métodos utilizados, é um processo com um determinado nível de incerteza associado [13].

### 2.2.1 Modelos de previsão

Um bom modelo de previsão pode ser fundamental para qualquer organização, podendo desempenhar um papel preponderante no processo de tomada de decisão não só sobre investimentos como também sobre a distribuição de recursos humanos e materiais [3].

O recurso a modelos de previsão é uma prática comum na indústria, para que se consigam obter dados sobre volumes de vendas futuros e, através deles, poder tomar determinadas decisões (figura 2.1). Estes modelos podem incluir vários métodos, desde o mais simples ao mais complexo, sendo que para diferentes realidades podem existir métodos mais adequados do que outros.

Uma boa parte das organizações recorre a análise de regressão, a simulações computacionais mais sofisticadas ou ainda a métodos estatísticos para realizarem as suas previsões sobre os respetivos volumes de vendas [7].

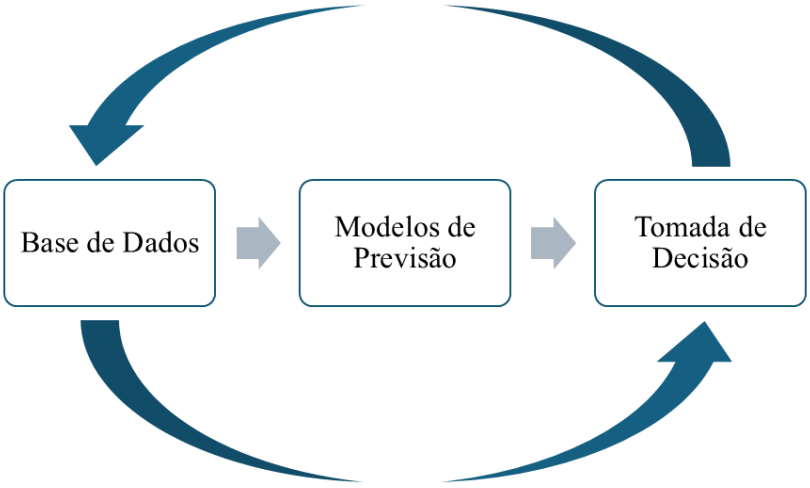


Figura 2.1: Arquitetura de um problema de previsão - adaptado de [18]

Os métodos de previsão podem ser aplicados em múltiplas tarefas, nas mais diversas áreas. Prever com precisão a procura de mercado futuro é de extrema importância para as organizações, e pode ser uma vantagem competitiva em muitos domínios já que auxiliam na tomada de decisões informadas sobre a gestão, mais concretamente ao nível de gestão de stocks, planeamento de produção ou até preços de mercado para os seus produtos [6].

A incapacidade de prever corretamente as quantidades de produto necessárias pode conduzir a um excesso de material, o que remete para excessivos custos de inventário ou, por outro lado, a uma escassez de material, o que pode levar à perda de fidelidade por parte dos clientes e, consequentemente, perda de lucros [12].

As empresas têm uma constante preocupação para evitar a falta de stock por forma a que disponham do necessário para satisfazer as suas necessidades e dos seus parceiros. No entanto, é também sua intenção a redução de custos de excesso de stock nas suas prateleiras. É, por isso, fundamental fornecer os produtos apropriados e nas quantidades certas [10].

Nas organizações, por forma a otimizar os seus planos de decisão é crucial prever futuras tendências de mercado com o objetivo de estabelecer as melhores estratégias de reabastecimento de stocks, permitindo assim dar uma resposta ajustada às futuras necessidades do mercado [12]. No entanto, convém salientar que por mais confiável que possa ser, o ato de prever tem sempre alguma incerteza associada [3]. A precisão é imperativa para tentar reduzir ao máximo os erros de previsão e, assim, alcançar empreendimentos comerciais bem sucedidos [1].

Numa primeira instância, os modelos de previsão eram vistos como um auxílio para a tomada de decisão sobre um qualquer negócio, normalmente realizados de forma qualitativa. Estes modelos apresentavam um elevado grau de incerteza uma vez que se baseavam na subjetividade dos especialistas das diferentes áreas de negócio e, para além disso, poderiam ainda ser facilmente afetados por alterações estruturais, gerando assim uma noção errada da procura. Em ambientes de incerteza, as pessoas tendem a adotar atitudes distintas, ou seja, um pessimista que tendencialmente prefere não correr grandes riscos, poderia sujeitar-se a ter uma procura bastante superior à oferta e com isso não ter a capacidade de satisfazer as necessidades do mercado, enquanto que um otimista, que tendencialmente espera obter o lucro máximo, pode conduzir a um desperdício de recursos humanos e materiais, sujeitando assim a empresa a ficar numa situação bastante desagradável [3]. A mesma autora refere ainda que, segundo Hooley & Hussey (1999), foi com o intuito de colmatar estas lacunas que foram surgindo os métodos quantitativos, menos falíveis e com a possibilidade de considerar alguns fatores que estão na base do processo da oferta e procura como a conjuntura sócio-económica atual, tendências de mercado, preços, sazonalidade de alguns produtos, campanhas publicitárias, entre outros.

Conseguir prever a quantidade ótima da procura dos produtos pode ser uma mais valia para qualquer organização e tem normalmente por base um padrão de vendas, seja ele histórico ou atual [12]. As empresas geram e armazenam diariamente grandes quantidades de dados associados às vendas, o que lhes permite a realização de previsões [15]. O uso de dados históricos possibilita um melhor desempenho do sistema, providenciando assim uma maior precisão na previsão de valores [1].

Os históricos de vendas das empresas são considerados como dados de séries temporais e podem ser facilmente influenciados por variações cíclicas. Para realizar previsões de séries temporais é necessário equacionar determinados fatores, nomeadamente a periodicidade. Existem estudos realizados onde a sazonalidade é equacionada, no entanto, nem todas as empresas são afetadas apenas pela sazonalidade, sendo necessário considerar as flutuações periódicas que, para além das flutuações sazonais, incluem ainda as flutuações circulares. Os fatores circulares podem influenciar determinados setores, como o económico e o financeiro, devendo por isso ser considerados para a realização de previsões, sobretudo no que ao campo industrial diz respeito [8].

As técnicas de previsão de séries temporais são muito utilizadas para a previsão de vendas e dividem-se sobretudo em dois grandes grupos, um deles referente a técnicas clássicas tendo como base modelos matemáticos e estatísticos e outro tendo por base o recurso a técnicas de inteligência artificial [11].

Muitos são os estudos que usam os modelos clássicos para a previsão de séries temporais, apoiando-se a sua investigação sobre a relação linear entre o objeto de previsão e as variáveis internas. Contudo, estes modelos apresentam as suas limitações, uma vez

que descuram do efeito de fatores não lineares internos, podendo por isso dar origem a previsões limitadas. As técnicas de *Machine Learning*, por sua vez, apresentam vantagens significativas para o processamento de grandes conjuntos de dados, e complexos, vindo por isso a ganhar destaque e a tornarem-se cada vez mais uma referência para o estudo e previsão de séries temporais [15].

Os modelos estatísticos têm por base uma combinação de funções matemáticas e princípios estatísticos que recorrem a amostras de dados para estimar e realizar projeções mais amplas. Existe uma grande diversidade de métodos estatísticos, destacando-se os modelos de média móvel integrada autoregressiva, ARIMA, e sua extensão com variáveis explicativas, SARIMA, como dos mais utilizados dentro deste tipo. Esta classe de modelos tem-se revelado eficaz perante determinados *datasets*, contudo, enfrentam algumas adversidades para processos de previsão que envolvam diversas variáveis, sobretudo pela maior complexidade de processamento, dificuldade na capacidade de generalização e o acréscimo nos tempos de computação [19].

O *Machine Learning* tem-se vindo a afirmar como uma tecnologia disruptiva com um impacto transversal em múltiplas áreas, nomeadamente a ciência e a indústria [20]. Nas últimas décadas têm emergido paulatinamente técnicas de previsão baseadas em ML, como máquina de vetores de suporte, árvores de regressão, redes neuronais artificiais, entre outras [10]. Estas técnicas apresentam diversas vantagens na área da previsão. A sua aptidão para lidar e gerir grandes quantidades de dados, aliada à capacidade que têm de considerar diversos fatores, tornam possível o estudo de processos de forma mais completa e detalhada [21]. Apesar de se tratar de métodos que exigem bastantes recursos computacionais, são soluções capazes de facultar uma enorme flexibilidade e uma elevada precisão no que às previsões diz respeito, sobretudo quando se está perante uma grande quantidade de observações [10].

A aplicação de metodologias preditivas baseadas em ML assenta sobretudo na construção e desenvolvimento de modelos mais robustos e precisos [22]. No entanto, estes modelos não apresentam apenas vantagens, tornando-se imperativo o reconhecimento de algumas das suas limitações. Uma delas está relacionada com a quantidade e qualidade de dados para treino que necessita, que têm que ser elevadas; outra prende-se com a engenharia de atributos e a seleção adequada de variáveis de entrada que são fundamentais para criar modelos relevantes; por último a possibilidade de ocorrer *overfitting* para situações em que a complexidade do modelo não seja bem gerida ou que o conjunto de treino seja demasiado pequeno [21].

Com o intuito de melhorar a precisão das previsões foram também desenvolvidos modelos de previsão híbridos. Estes integram dois ou mais modelos num só, combinados entre si, beneficiando das vantagens individuais de cada um [19]. Podem resultar da associação de vários modelos de ML ou da agregação de métodos matemáticos e estatísticos com modelos de ML [21].

Quando se têm disponíveis dados que apresentam uma determinada tendência, um certo padrão e até sazonalidade, é importante que sejam considerados, pois caso contrário podem surgir más previsões tendo por base estes valores. Nas últimas décadas, modelos mais tradicionais de previsão de séries temporais, como média móvel, regressões multivariadas, Box Jenkis ARIMA, de alisamento exponencial, entre outros, eram usados para processar este tipo de dados. No entanto, para casos em que se verificassem flutuações de mercado constantes e aleatórias, os resultados obtidos com estes modelos nem sempre eram os melhores, surgindo a necessidade de procurar modelos diferentes e com outro tipo de capacidades, como, por exemplo, as redes neuronais artificiais [7].

**Redes Neuronais Artificiais**

Os dados nem sempre se apresentam da mesma forma, possuindo cada conjunto as suas características. Uns tipos pela instabilidade, não linearidade, outros por múltiplo acoplamento ou multivariáveis, ou ainda outros que em algum momento possam apresentar simultaneamente uma índole sequencial, todos à sua maneira, têm o seu ADN, a sua natureza. Os métodos estatísticos não são, normalmente, capazes de lidar com dados multivariados e a sua não-linearidade, enquanto que os métodos de aprendizagem automática, por sua vez, não têm a capacidade de extrair por completo a informação temporal dos dados. Estes desafios podem ser ultrapassados pelas redes neuronais [23].

Dentro do *Deep Learning*, a classe de modelos que mais se tem destacado são as redes neuronais. Estas são amplamente utilizadas para extração de informação, reconhecimento de padrões e modelações de previsões dinâmicas, complexas e precisas a partir de grandes conjuntos de dados. Apesar das exigências computacionais, os recentes avanços a nível de hardware têm potenciado reduções substanciais nos tempos de processamento, causando grandes impactos em campos como o ML e a inteligência artificial [24].

Das técnicas de inteligência artificial, as redes neuronais artificiais são o modelo mais utilizado, já que se trata de um modelo com provas dadas no que respeita à modelação tanto de séries temporais como de não temporais, superando os modelos clássicos também quanto à sua capacidade de aproximação universal de funções, de não linearidade e de generalização [11].

As redes neuronais artificiais dizem respeito a uma tecnologia com enorme potencial utilizada em muitas áreas da ciência, nomeadamente a engenharia, computação, gestão, finanças, marketing, entre outras. Estas técnicas possibilitam a identificação e modelação de padrões não lineares dentro de um conjunto de dados, que não seriam de fácil interpretação recorrendo a métodos estatísticos ou outros modelos mais tradicionais [13] [7]. São modelos de inteligência artificial que têm por base a ciência biológica,

mais concretamente o funcionamento do cérebro e seus neurónios. Assim como no cérebro humano, também os componentes das redes estão ligados entre si, seguindo algum padrão de conectividade. Associados a esta conectividade estão diferentes pesos que são atualizados por intermédio da aprendizagem [13].

Se, por um lado, se tem que reconhecer que as redes neuronais artificiais se tratam de uma tecnologia com enorme potencial, com uma forte capacidade de processamento e resistência à falha, por outro, tem que se admitir e identificar algumas limitações, sobretudo no que diz respeito a uma velocidade lenta de convergência, ao *overfitting* e à facilidade para cair em extremos locais [7].

Dentro das redes neuronais artificiais existem alguns tipos que se destacam, nomeadamente as redes neuronais recorrentes, ou *Recurrent Neural Network* (RNN). As redes neuronais recorrentes são um tipo de rede neuronal artificial especial que, pelas ligações direcionadas entre os neurónios de uma mesma camada, consegue utilizar informação sequencial. Para cada elemento da sequência este tipo de rede executa a mesma tarefa, por isso o nome de recorrente. Cada saída destas redes depende de cálculos realizados anteriormente, o que desenvolveu a sua necessidade em armazenar memória. No entanto, tal como outros tipos de redes neuronais, também este apresenta as suas fraquezas, ligadas sobretudo ao desvanecimento do gradiente [25].

#### Gradient Boosting

O *Ensemble Learnig* (EL), ou aprendizagem por conjuntos, é um dos campos de investigação mais trabalhados dentro do *Machine Learning*, e tem como base a combinação de vários modelos mais fracos para formar um forte, devidamente estruturado e mais robusto, com a capacidade de realizar previsões com níveis de precisão bastante elevados [26].

O *boosting* diz respeito a uma abordagem da aprendizagem de EL fundamentalmente desenvolvida para incrementar valor às previsões através da utilização de vários modelos fracos, melhorando progressivamente a sua capacidade preditiva através do foco em exemplos de treino mais complexos, o que leva gradualmente ao aperfeiçoamento das suas previsões [27].

O Gradient Boosting (GB) é uma técnica de otimização bastante utilizada nos algoritmos de *boosting* dentro do EL, extremamente poderosa e especialmente eficiente para resolver problemas de classificação e regressão [26] [27]. Estes métodos caracterizam-se por atribuírem maior peso às previsões com erro mais elevado, sendo particularmente eficazes na modelação de relações mais complexas e na gestão de dados ausentes [21].

Esta técnica é mais do que uma simples sobreposição de diversos classificadores, uma vez que, mediante os parâmetros, consegue realizar uma combinação extremamente

otimizada de modelos mais fracos até alcançar um modelo mais completo e consolidado que apresente previsões com uma precisão significativamente superior [26]. As previsões corretas não sofrem penalização, contrariamente ao que acontece com as previsões consideradas como incorretas, remetendo desta forma para aquele que é o cerne do modelo, as partes mais difíceis de treino. O constante ajuste de hiperparâmetros é fundamental para os modelos de GB, uma vez que lhes possibilita o controlo total sobre o equilíbrio entre o *overfitting* e o *underfitting*, influenciando diretamente aquele que será o desempenho final do modelo [27].

**Long Short-Term Memory - LSTM**

As *Long Short-Term Memory* (*LSTM*), também conhecidas como redes de memória de longo e curto prazo, surgem em 1997, propostas por Jürgen Schmidhuber e Sepp Hochreiter, e dizem respeito a um tipo de redes neuronais recorrentes, ou *Recurrent Neural Network* (RNN) [28].

Este tipo de rede foi desenvolvida para enfrentar os desafios apresentados pelas RNN no que respeitava ao processamento de sequências de dados prolongadas [19]. As RNN convencionais apresentam determinados problemas relacionados com o gradiente, desvanecido ou explosivo, ultrapassados nas LSTM com a introdução de novas portas que possibilitam uma maior capacidade no controlo do gradiente e uma maior capacidade na preservação de dependências de longo alcance [25].

A LSTM é particularmente eficaz no que se refere à modelação e previsão de sequência de dados, superando algumas das limitações apresentadas por algumas RNN mais tradicionais que, pelo desvanecimento do gradiente, podem ter apenas memória de curto prazo [29][30]. Estas apresentam-se como uma boa solução para lidar com as dependências temporais, uma vez que combinam memória de curto e longo prazo, podendo atenuar até determinado ponto o problema da redução do gradiente [31] [30]. Este modelo introduz uma janela de memória capaz de reter informação por prolongados períodos, armazenando informação de longo prazo, e três tipos de porta que regulam o fluxo para o interior e exterior de cada célula, a porta de entrada, de esquecimento e de saída [19]. A porta de entrada define qual a informação, proveniente da nova entrada, que deve ser adicionada ao estado da célula ou atualizada; a porta de esquecimento controla a quantidade de informação que é mantida do estado anterior e a porta de saída indica qual a informação que, com base no estado da célula, será disponibilizada para saída. Estas portas regulam o fluxo de informação que entra, circula e sai da célula de memória, trabalhando conjuntamente para aprender e armazenar informação de sequências de curto, longo prazo e padrões complexos [28] [25]. Desta forma, as células de memória podem ser introduzidas nos neurónios da camada oculta da rede recorrente e registar o histórico de informação, que pode ser preservada [30]. Assim é possível armazenar e propagar retroativamente os erros através das camadas e no

decorrer do tempo, o que facilita o processo de aprendizagem [19].

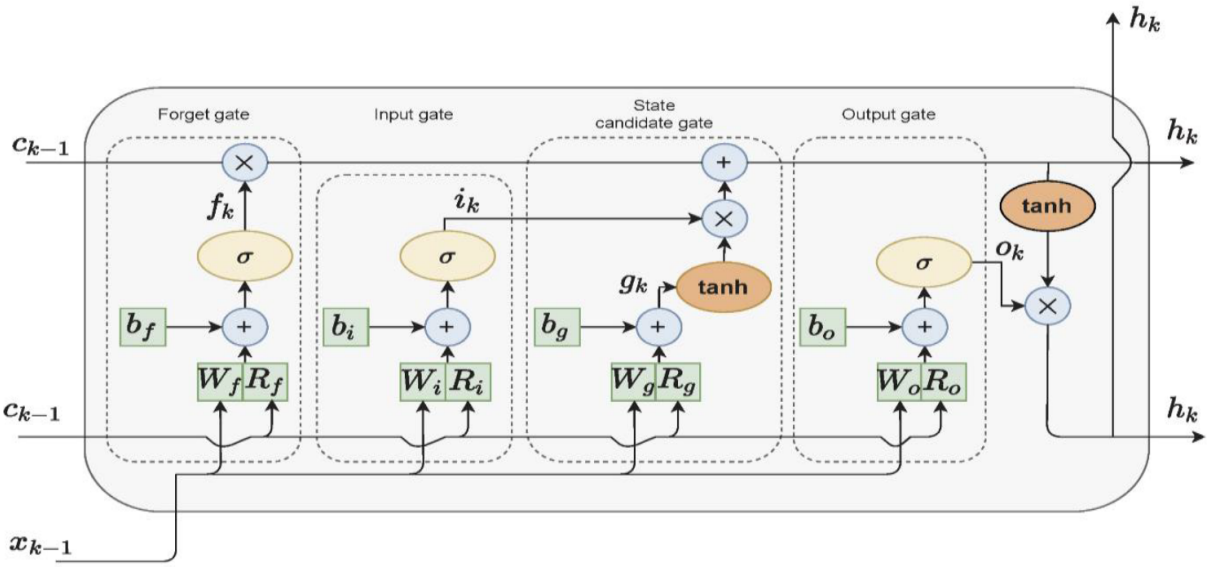


Figura 2.2: Arquitetura Long Short-Term Memory - LSTM [28]

A arquitetura do modelo LSTM pode ser apresentada segundo a figura 2.2 [28]. As operações de controlo das portas consistem basicamente em operações de produto escalar e numa função de ativação sigmoide [20]. Segundo os mesmos autores, a porta de esquecimento é responsável por determinar a informação que é para ser descartada e a útil que deve ser retida do estado anterior da célula e pode ser expressa por :

$$f_t = \sigma(w_f * [h_{t-1}, x_t] + b_f) \tag{2.1}$$

onde:

- $f_t$ - matriz de pesos da porta de esquecimento;
- $\sigma$ - função de ativação sigmoide;
- $w_f$ - peso de ligação da saída anterior;
- $h_{t-1}$ - saída anterior;
- $x_t$ - entrada atual;
- $b_f$ - vetor de polarização (constante).

A porta de entrada é a responsável por determinar qual a informação que deve ser atualizada [20]. O processo tem inicio a partir do momento em que é aplicada a função sigmoide à combinação entre a informação da entrada de unidade atual com a de saída da unidade anterior, o que cria um vetor que determina as partes que devem ser atualizadas. A essa informação combinada é também aplicada uma função de ativação tangente hiperbólica, "tanh", o que cria outro vetor com os valores candidatos ao novo

estado da célula. Estes dois vetores são posteriormente multiplicados e como resultado desta operação surge a nova informação que será adicionada à célula [19]. De acordo com Zeng *et al.* [20], estas operações podem ser expressas pelas seguintes equações:

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \tag{2.2}$$

$$\tilde{C_t} = tanh(W_c.[h_{t-1}, x_t] + b_c) \tag{2.3}$$

onde:

- $[h_{t-1}, x_t]$ - combinação dos vetores de saída anterior e de entrada atual num só vetor;
- $i_t$ - vetor criado pela função sigmoide;
- $W_i$ - matriz de pesos da porta de entrada;
- $b_i$ - vetor de bias de entrada;
- $\tilde{C_t}$ - vetor criado pela função tanh;
- $W_c$ - matrizes de pesos do estado da célula;
- $b_c$ - vetores de polarização do estado da célula.

A memória de longo prazo é atualizada através da combinação entre dados novos e anteriores, atualização esta que resulta da soma entre os produtos de cada elemento da porta de esquecimento com o estado anterior e do estado atual candidato com a porta de entrada [19]. De acordo com os mesmos autores, o estado da célula é dado por:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C_t} \tag{2.4}$$

onde:

- $C_t$ - novo estado da célula;
- $C_{t-1}$ - estado anterior da célula;

A porta de saída é a responsável por regular o fluxo de informação no estado atual [19]. A função de ativação sigmoide vai determinar em primeira instância qual a porta de saída, sendo o estado da célula posteriormente normalizado para o intervalo [-1;1] através da função tanh. Por último, é realizado o produto elemento a elemento, obtendo-se assim o valor final de saída. Segundo Zeng *et al.* [20], estas operações podem ser expressas pelas seguintes equações:

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \tag{2.5}$$

$$h_t = o_t * tanh(C_t) \tag{2.6}$$

onde:

- $o_t$ - taxa de esquecimento;
- $W_o$ - matriz de pesos da porta de saída;
- $b_o$ - vetor de polarização da porta de saída;
- $h_t$ - saída da camada oculta.

As LSTM destacam-se pela sua capacidade em capturar dependências de longo prazo e padrões sequenciais em dados históricos, incluindo sazonalidades [19]. O desempenho e a precisão destas redes são fortemente condicionados pela configuração dos seus hiperparâmetros. A profundidade da rede e o número de neurónios em cada uma das suas camadas assumem, por isso, uma posição preponderante. Contudo, o ajuste manual desses parâmetros é uma tarefa exigente e algo complexa, o que pode comprometer a possibilidade de alcançar um modelo eficiente e adequado a determinadas necessidades do mundo real [20].

**Gated Recurrent Unit - GRU**

A *Gated Recurrent Unit* (GRU) foi apresentado em 2014 pelos autores Cho, Merriënboer, Gulcehre, Bahdanau, Bougares, Schwenk e Bengio. Este modelo tinha como principal finalidade a captação de forma adaptativa de dependências em diferentes escalas temporais por parte de cada unidade recorrente [32].

A GRU é uma extensão de RNN bastante utilizada e capaz de controlar o fluxo de informação por intermédio da aprendizagem [23].

As RNN são um tipo de redes neuronais *feedforward* adequadas ao processamento de dados em sequência de comprimento variável mas que apresentam algumas limitações em aplicações práticas relativamente ao desvanecimento de gradiente, ultrapassados pela estrutura GRU, que utiliza mecanismos de portões inspirados nas LSTM, através dos quais consegue extrair informação temporal[15] [23]. Com a capacidade que têm de regular o fluxo dos grandientes, tanto as GRU como as LSTM são arquiteturas com capacidade de assegurar aprendizagens estáveis e previsões precisas [28].

Tal como a LSTM, também a GRU tem vindo a despertar interesse no que a modelos de previsão de séries temporais diz respeito [33].

As GRU e as LSTM são dois modelos que conseguem identificar e reter dependên-

cias de longo alcance em dados de vendas, o que as destaca na modelação de relações sequenciais [34]. As LSTM são recorrentemente utilizadas para modelação de dependência de longo prazo enquanto as GRU podem ser uma melhor opção para casos em que o *dataset* seja menos complexo ou que se procure uma maior eficiência computacional [28].

"As GRU reduzem o número de parâmetros das LSTM", sendo por isso menos complexas e apresentam como principal diferença o facto de com uma única unidade de GRU conseguir controlar em simultâneo a atualização do estado e a decisão sobre o fator de esquecimento [33] [15]. Assim, este modelo consegue manter a forte capacidade de captar eficientemente dependências de longo prazo e reduzir a complexidade computacional exigida ao mesmo tempo [22].

No entanto, apresentam também as suas limitações uma vez que não permitem a identificação da importância atribuída a cada variável [23].

As GRU são modelos de estrutura mais compacta que combinam as portas de esquecimento e de entrada, simplificando assim a sua arquitetura [28]. Desta forma, conseguem atingir tempos de treino mais rápidos uma vez que têm um menor número de parâmetros a otimizar [22].

A arquitetura das GRU pode ser apresentada consoante a figura 2.3 [28].

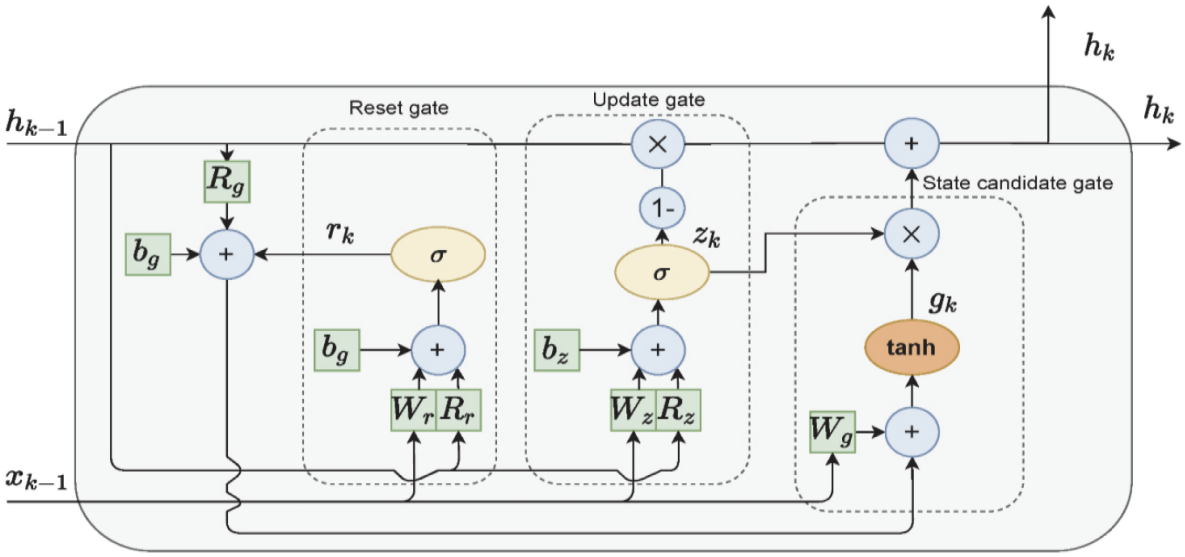


Figura 2.3: Arquitetura Gated Recurrent Unit - GRU [28]

Este modelo introduz duas portas, uma de *reset* e outra de atualização. Os valores destas portas são calculados, em cada instante temporal, recorrendo à entrada atual e ao estado oculto anterior, através de uma camada toda ela interligada e cuja função de ativação é a sigmoide [23]. Assim, as redes GRU conseguem atualizar as células de memória de forma adaptativa e controlar o fluxo de informação, o que lhes permite lidar de forma eficiente com padrões em evolução e sequências de diferentes dimensões

[28].

De acordo com Gasparin *et al.* [35], as portas de atualização e a porta de *reset* podem ser expressas, respetivamente, pelas seguintes equações:

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \tag{2.7}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1}) \tag{2.8}$$

onde:

- $z_t$ - vetor de atualização;
- $W_z$ - matriz de peso associada à entrada na porta de atualização;
- $x_t$ - entrada atual;
- $U_z$ - matriz de peso do estado oculto anterior na porta de atualização;
- $h_{t-1}$ - estado oculto anterior;
- $r_t$ - vetor de *reset*;
- $W_r$ - matriz de peso associada à entrada na porta *reset*;
- $x_t$ - entrada atual;
- $U_r$ - matriz de peso do estado oculto anterior na porta de *reset*;
- $h_{t-1}$ - estado oculto anterior.

A função de ativação sigmoide vai normalizar os elementos e transformar cada um para valores entre 0 e 1 [0,1], garantindo desta forma que nas portas de atualização e de *reset* todos os valores estão dentro deste intervalo [23]. A porta de atualização decide a medida em que a unidade irá atualizar o seu conteúdo enquanto que e a porta de reinicialização é a responsável por controlar como é que o estado oculto anterior influenciará o estado oculto candidato no instante atual [32] [23].

Para auxiliar o cálculo dos estados ocultos futuros, a GRU calcula o estado oculto candidato [23]. De acordo com Chung *et al.* [32], a ativação candidata é calculada por:

$$\tilde{h}_t = tanh(W x_t + U(r_t . h_{t-1})) \tag{2.9}$$

onde:

- $\tilde{h}_t$ - vetor do estado oculto candidato.

O estado oculto anterior pode conter informação referente à série temporal, podendo

aqui ser importante a porta de *reset* para descartar informação que não seja de todo relevante para o modelo [23]. A ativação da GRU no instante $t$, ou atualização do estado oculto, não é mais do que uma interpolação linear entre a ativação candidata e a ativação do estado oculto anterior [32]. Segundo Gasparin *et al.* [35], esta pode ser expressa pela equação:

$$h_t = z_t h_{t-1} + (1 - z_t)\tilde{h}_t \tag{2.10}$$

onde:

- $h_t$ - vetor do estado oculto.

**Multi-Layer Perceptron Regressor - MLP Regressor**

O Multi-Layer Perceptron (MLP) é um algoritmo bastante utilizado na área da previsão. Trata-se de um tipo de rede neuronal artificial *feedforward* com uma técnica de aprendizagem supervisionada, designada por retropropagação, e que utiliza uma função de ativação não linear para o treino. Para este modelo, o problema em estudo e o grau de aprendizagem do preditor são extremamente importantes, uma vez que o seu número de execuções e respetivos *timings* de previsão dependem significativamente destes dois fatores [36].

Entre os modelos de redes neuronais, o erro quadrático pode ser otimizado pelo gradiente estocástico ou pelo MLP Regressor através do método de Broyden-Fletcher-Goldfarb-Shanno (LBFGS) com memória limitada [37].

O MLP Regressor, também conhecido como *Multi-Layer Perceptron Regressor*, diz respeito a um percetrão multicamada treinado por retropropagação. A aprendizagem por retropropagação foi estabelecida para realizar as ligações entre entradas e saídas, desempenhando aqui o MLP Regressor uma função de extrema importância ao criar redes de forma aleatória, atualizando-as posteriormente pela comparação dos resultados desejados com os resultados iterativos da rede [36].

Este modelo tem capacidade para decifrar padrões e dependências não lineares complexas entre os dados [38], podendo ser um recurso bastante valioso para grande parte das organizações. O modelo consiste em múltiplas camadas de neurónios ligados individualmente à camada seguinte. Cada um destes neurónios aplica uma soma ponderada nas suas entradas, processando de seguida o resultado através de uma função de ativação para criar uma saída[37]. Os neurónios das camadas ocultas conferem ao modelo um comportamento não linear em virtude das suas funções de ativação, também elas de natureza não linear [38]. O MLPRegressor é treinado consecutivamente de forma iterativa, utilizando regularização para ajustar os parâmetros do modelo com o intuito de minimizar o *overfitting* [37].

De acordo com Rizk *et al.* [38], a equação geral para cada um dos nós que constitui as camadas da rede é:

$$z_j = \sum_{i=1}^{n} w_{ij}.x_i + b_j \tag{2.11}$$

em que:

- $z_j$ - soma ponderada das entradas numa determinada camada para o nó $j$;
- $w_{ij}$ - peso da ligação entre o nó $i$ da camada anterior e o nó $j$ da camada atual;
- $x_i$ - saída do nó $i$ da camada anterior;
- $b_j$ - termo de polarização para o nó $j$ da camada atual.

Para minimizar as diferenças entre os valores reais e os previstos, o modelo ajusta os pesos $wij$ e os termos de polarização $b_j$ por retropropagação durante o seu treino. A função de ativação $a_j = f(z_j)$ é aplicada depois de calculado o valor de $z_j$, sendo as mais comuns a tangente hiperbólica (tanh), a unidade linear retificada (Relu) e a função sigmoide [38].

Segundo Rizk *et al.* [38], a arquitetura geral do MLP Regressor é dada por:

$$a_{saida} = f_{saida}(W_{saida}.f_{oculto}(W_{oculto}.X + b_{oculto}) + b_{saida}) \tag{2.12}$$

onde:

- $a_{saida}$ - saída da última camada, ou seja, valor previsto para a regressão;
- $X$ - vetor de entrada;
- $f_{saida}$ - função de ativação para a camada de saída;
- $W_{saida}$ - matriz de pesos para camada de saída;
- $f_{oculto}$ - função de ativação para camada oculta;
- $W_{oculto}$ - matriz de pesos para camada oculta;
- $b_{oculto}$ - vetor de polarização para camada oculta;
- $b_{saida}$ - vetor de polarização para camada de saída.

Devido ao seu baixo grau de complexidade computacional, é um método fácil de implementar [36]. A grande força deste modelo provém sobretudo da sua arquitetura, perceptrão multicamada. Contudo, apresenta também algumas lacunas, sobretudo no que respeita ao número de parâmetros envolvidos, que pode ser ainda mais otimizado. Apesar disso, é um modelo que pode ser encarado como um instrumento com

capacidade para identificar relações complexas entre os dados e não apenas como uma simples ferramenta de previsão, demonstrando capacidade para captar e interpretar interações complexas entre variáveis [38].

**Extreme Gradient Boosting - XGBoost**

O *Extreme Gradient Boosting* (XGBoost) surgiu em 2016 por Tianqi Chen e Carlos Guestrin [39]. Trata-se de uma ferramenta de ML bastante poderosa baseada em *Gradient Boosting* (GB) [31].

Este modelo, altamente poderoso e eficiente, foi projetado para otimizar o algoritmo de GB, combinando múltiplos modelos preditivos mais fracos, árvores de decisão, para criar um modelo mais robusto, com o intuito de melhorar o seu desempenho [19][29]. Estes modelos preditivos são melhorados por intermédio da expansão de Taylor de segunda ordem, que aborda de forma eficaz a complexidade do cálculo das derivadas em determinadas funções de perda de primeira ordem, processando assim a sua otimização de forma rápida, contribuindo também para atenuar a possibilidade de *overfitting* [40] [39].

O XGBoost é um modelo extremamente utilizado em múltiplas aplicações, destacando-se sobretudo pela sua portabilidade e versatilidade [37]. É bastante reconhecido em modelagem preditiva pela sua eficácia e pelo seu sucesso na ciência dos dados [29]. Este método evidencia-se dos demais pela capacidade que apresenta para captar tendências dinâmicas de curto prazo com precisão, sendo especialmente eficaz para solucionar problemas de previsão de curto prazo [19]. Foi projetado para alcançar uma elevada flexibilidade, portabilidade e eficiência, apresentando uma enorme capacidade de generalização, sendo por isso capaz de resolver muitos problemas de dados de forma rápida e precisa [31]. Este algoritmo apresenta pouca sensibilidade a valores muito díspares, também designados por *outliers*, e uma ótima resistência ao *overfitting*, uma vez que introduz um termo de regularização na sua função de perda [37] [40]. É reconhecido pela sua velocidade de desempenho e capacidade de lidar com inúmeros conjuntos de dados, mesmo aqueles que apresentem grandes dimensões [29].

Este modelo tem sido alvo de grande destaque para múltiplas tarefas, nomeadamente para problemas de ordenação (*ranking*) e classificação [29]. As suas características tornam este método bastante eficiente e particularmente útil no que respeita a problemas com um *dataset* onde é aplicada classificação ou procedimentos de regressão [37]. O XGBoost é orientado para a otimização de resíduos e baseia-se essencialmente na adição de novos "fracos aprendizes", árvores de decisão, que vão corrigindo os resíduos deixados pelos anteriores [31][19]. Neste modelo, o conjunto de entrada da próxima árvore de decisão é influenciado pelos resultados de treino e de previsão da árvore anterior [26]. As imperfeições de cada aprendiz são analisadas individualmente, podendo apresentar resultados nada satisfatórios. A previsão final do XGBoost corres-

ponde à soma dos valores de todos os "fracos aprendizes" que, todos combinados, levam a uma previsão final mais precisa do que aquela que se obteria recorrendo a modelos de árvores de decisão individualmente [19].

Segundo Pandey *et al.* [37], a função objetivo do XGBoost na etapa $i$ é dada pela seguinte equação:

$$L^{(t)} = \sum_{i} l(P^{(t)}, R) + \sum_{k} \Omega(f_k) \tag{2.13}$$

onde:

- $P^{(t)}$ - valor previsto no instante $t$;
- $R$ - valor real;
- $l(P^{(t)}, R)$ - quantifica a diferença entre o valor real e o valor previsto, ou seja, representa a perda;
- $\Omega(f_k)$ - representa a complexidade da $k$-ésima árvore de decisão.

A tabela 2.2 resume as principais características dos modelos preditivos abordados nesta secção, permitindo uma visão comparativa entre as suas arquiteturas, funcionalidades, vantagens e limitações. Esta análise constitui a base para a seleção prática dos modelos no contexto específico deste projeto, sendo posteriormente complementada com a avaliação sobre o deu desempenho na previsão de vendas.

### 2.2.2 Avaliação sobre os modelos de previsão

A qualidade e o tipo de dados são fundamentais para que se consigam realizar boas previsões e, com isso, tomar decisões sobre os valores que apresentem maior precisão.

As previsões são estimativas que têm por base um conjunto de dados ou observações, normalmente sustentadas por cálculos com um determinado nível de precisão [41].

A avaliação sobre a eficiência dos modelos de previsão é bastante importante, permitindo a identificação e escolha sobre os modelos que, de acordo com o *dataset* disponível, consigam obter os resultados mais fiáveis e, por isso, mais precisos.

Dada a relevância e o impacto que as previsões podem ter para o sucesso de uma empresa, é preponderante perceber de que forma é que se podem aprimorar os métodos utilizados até então com o objetivo de melhorar a qualidade dessas previsões.

Para analisar a qualidade dos métodos de previsão existem alguns indicadores que, depois de calculados, permitem que o utilizador tenha uma perceção quantitativa do desempenho dos seus modelos e, consequentemente, da qualidade das previsões que estão a gerar.

Tabela 2.2: Comparação entre modelos de previsão - LSTM, GRU, MLP Regressor e XGBoost

| Características | Modelos | | | |
|---|---|---|---|---|
| | **LSTM** | **GRU** | **MLP Regressor** | **XGBoost** |
| **Tipo** | Rede neuronal Recorrente (RNN) avançada | Rede neuronal Recorrente (RNN) simples | Rede neuronal Feed-Forward | Gradient Boosting |
| **Arquitetura** | Células de memória + 3 portas (entrada, esquecimento, saída) | 2 portas (reset, atualização) | Densas camadas totalmente interligadas | Árvores de decisão em boosting |
| **Memória de longo prazo** | sim (captura de dependências longas) | Sim (captura de dependências mas menos profundas que LSTM) | Não | Não |
| **Resistência a ruído/outliers** | Alta | Alta | Média | Alta |
| **Exigência computacional** | Elevada | Média | Baixa | Média |
| **Aplicações** | Séries temporais longas | Séries temporais | Regressão tabular | Regressão/Classificação, Rankings |
| **Vantagens** | Memória longa, precisão alta | Memória, simplicidade e velocidade | Simplicidade e rapidez | Robusto e bom desempenho em tabular |
| **Desvantagens** | Treino pesado | Limitações em dependências muito longas | Engenharia de *features* obrigatória; Probabilidade de *overfitting* | Menor eficácia em dados muito sequenciais |

### Métodos de avaliação de precisão

A qualidade de uma previsão pode ser avaliada segundo vários testes estatísticos e métricas determinísticas [41].

Para a avaliação da performance dos modelos de previsão temos as métricas de avaliação, de entre as quais podemos destacar a MSE (*Mean Squared Error*) ou erro quadrático médio, a RMSE (*Root Mean Squared Error*) ou raíz do erro quadrático médio, a MAE (*Mean Absolut Error*) ou erro médio absoluto e ainda o MAPE (*Mean Absolut Percentage Error*) ou erro médio absoluto percentual.

Todos as métricas evidenciadas são consideradas como indicadores de baixo desempenho, uma vez que a exatidão dos modelos é tanto maior quanto menor for o valor calculado para cada uma delas, o que indica uma melhor e mais robusta capacidade de previsão [15]. Quanto menor for o valor obtido para qualquer um dos métodos, maior será o seu nível de precisão [8].

### MSE - Erro Quadrático Médio

O Erro Quadrático Médio (MSE) avalia a diferença quadrática média entre os valores reais e os valores previstos [41]. Esta métrica é fundamental para avaliar o desempenho dos modelos de previsão uma vez que indica o nível de generalização do modelo treinado relativamente aos dados não observados, conseguindo assim percecionar a

sua utilidade para aplicações práticas [28].

De acordo com Zermane *et al.* [28] este valor pode ser calculado matematicamente como a média das diferenças quadráticas para cada par correspondente de valores reais e previstos, segundo a equação:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(R_t - P_t)^2 \tag{2.14}$$

onde:

- $n$ - número de elementos em teste, quantidade de vezes que as iterações acontecem;
- $R_t$ - valor real no tempo $t$;
- $P_t$ - valor previsto no tempo $t$.

**RMSE - Raiz do Erro Quadrático Médio**

A Raiz do Erro quadrático Médio (RMSE) é um indicador recorrentemente utilizado para medir a magnitude média dos erros entre os valores reais e os valores previstos de um conjunto de dados, avaliando desta forma a eficiência de um modelo sobre a previsão de valores-alvo. [37].

A RMSE corresponde matematicamente ao desvio padrão dos resíduos, resíduos estes que representam a distância entre os pontos de dados e a reta de regressão [41].

Este indicador permite obter uma perceção sobre a dimensão das variações entre os valores reais e os previstos [34].

De acordo com Zhai *et al.* [23], a equação para o cálculo da raíz do erro quadrático médio é:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(R_t - P_t)^2} \tag{2.15}$$

onde:

- $n$ - número de elementos em teste, quantidade de vezes que as iterações acontecem;
- $R_t$ - valor real no tempo $t$;
- $P_t$ - valor previsto no tempo $t$.

O RMSE é um método sensível para grandes disparidades entre os valores reais e os valores previstos, atribuindo-lhes maior peso, enfatizando por isso os grandes erros

[8].

**MAE - Erro Médio Absoluto**

O Erro Médio Absoluto (MAE) é uma estatística que mede a diferença absoluta média entre os valores reais e os valores previstos, indicando o valor médio do erro absoluto para um determinado conjunto de dados [37] [41].

O MAE mede a magnitude média dos erros num conjunto de dados, não considerando para tal a sua direção [34].

Conforme o explicado por Rafi *et al.* [34], o cálculo do erro absoluto médio é realizado segundo a equação:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|R_t - P_t| \tag{2.16}$$

onde:

- $n$ - número de elementos em teste, quantidade de vezes que as iterações acontecem;
- $R_t$ - valor real no tempo $t$;
- $P_t$ - valor previsto no tempo $t$.

**MAPE - Erro Médio Absoluto Percentual**

O Erro Médio Absoluto Percentual (MAPE) mede a precisão de uma previsão em percentagem [34]. Esta métrica indica de forma clara a diferença entre os valores reais e os valores previstos, o que facilita a análise sobre o impacto das previsões [19]. Trata-se portanto de uma medida intuitiva uma vez que mostra a diferença que existe em termos percentuais entre os valores previstos e os valores reais [8].

De acordo com Guo *et al.* [19], a equação para calcular o erro médio absoluto em percentagem é:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{R_t - P_t}{R_t}\right| \tag{2.17}$$

onde:

- $n$ - número de elementos em teste, quantidade de vezes que as iterações acontecem;
- $R_t$ - valor real no tempo $t$;
- $P_t$ - valor previsto no tempo $t$.

Este método oferece benefícios extremamente importantes para os estudos efetuados graças à sua simplicidade, sobretudo no que diz respeito à facilidade de comunicação e interpretação [10]. No entanto, é um método que apresenta também algumas limitações, demonstrando-se errático sobretudo no que respeita ao zero e valores próximos dele.

# 3 Metodologia e ferramentas de trabalho

A ciência dos dados é uma área interdisciplinar que tem por base modelos analíticos e matemáticos que podem fornecer *insights* extremamente valiosos para as empresas [42].

No decorrer das últimas décadas é de realçar o crescente destaque que se tem vindo a dar a toda esta área, sobretudo pela emergência de grandes bases de dados nas mais diversas organizações [43].

O processamento de grandes quantidades de dados tem vindo a tornar-se um suporte imprescindível para grande parte das organizações[42]. Por isso, tornou-se fundamental o desenvolvimento e aplicação de técnicas próprias de extração de conhecimento útil para que se consiga retirar todo o valor da informação contida nesses dados [43].

Nesta secção é apresentada a metodologia utilizada para o desenvolvimento do modelo de previsão de vendas em estudo, bem como as diferentes etapas que a constituem.

Inicialmente é realizado um enquadramento teórico relativamente aos processos que fazem parte do método de desenvolvimento utilizado e, posteriormente, são explicadas as ferramentas usadas no decorrer do projeto, tanto para análise e tratamento de dados como para o desenvolvimento dos modelos.

## 3.1 Metodologia para o desenvolvimento do modelo

Nas últimas décadas a crescente digitalização de processos tem levado à proliferação de grandes quantidades de dados apoiadas em abordagens como a mineração de dados [44].

A mineração de dados, em inglês *Data Mining* (DM), é referente aos processos através dos quais são retirados e enumerados padrões a partir de um determinado conjunto de dados [43]. É definido como o processo de identificação de padrões novos, válidos e potencialmente úteis através de dados, vastamente utilizado em múltiplas aplicações [45].

A integração do *Machine Learning* na rota industrial veio revelar-se como uma grande mais valia para muitas organizações e indivíduos, sobretudo no que à tomada de decisões diz respeito. Ao permitir que as máquinas aprendam através de experiências e exemplos é possível descobrir informação oculta num determinado conjunto de dados, identificar padrões e, recorrendo à lógica preditiva, alcançar uma previsão de resulta-

dos que possam beneficiar as organizações das mais diversas maneiras [46].

### 3.1.1 Metodologias de apoio ao processo de modelação

Os métodos de *Data Mining* (DM) têm vindo a ganhar cada vez maior reconhecimento e, consequentemente, maior popularidade em determinadas áreas de conhecimento, nomeadamente a empresarial e industrial, sobretudo pela redução de custos e melhoria de desempenho que lhe estão inerentes [45].

As organizações mostram cada vez maior disponibilidade para explorar o potencial do DM integrando e aplicando técnicas nos seus projetos para análise e previsão de processos [44].

Esta área tem vindo a consolidar-se cada vez mais ao longo dos anos, surgindo assim a necessidade de estabelecer padrões, tanto a nível académico como industrial [43].

Quando se pretende aplicar uma técnica de DM, independente da que for, é importante que seja seguida uma determinada metodologia, isto é, uma sequência de processos que seja comum, independentemente do problema [3].

Dentro das metodologias de DM conhecidas, existem três que se podem destacar como das mais aplicadas pelas organizações para garantir melhores resultados nos seus projetos, nomeadamente a KDD (Knowledge Discovery in Databases), a SEMMA (Sample, Exploration, Modification/Manipulate, Model and Assessment) e a CRISP-DM (Cross-Industry Standard Process for Data Mining) [46].

O KDD, que em português significa descoberta de conhecimento em base de dados, é um processo de DM que extrai informação até então desconhecida, que pode ser considerada como conhecimento e, possivelmente útil, de um conjunto de dados, no qual o analista tem total controlo sobre a orientação e validação do processo de extração. Este modelo divide o processo analítico em cinco fases, nomeadamente a seleção, o pré-processamento, a transformação, a mineração de dados e a interpretação/avaliação, podendo algumas delas serem repetidas para a obtenção de resultados mais precisos [47][43]. Contudo, este modelo apresenta algumas limitações, sobretudo pelo facto de não apresentar nenhuma etapa de compreensão de dados, de não incluir a fase de implementação de processo e ainda a falta de ciclos de *feedback* [47].

A abordagem SEMMA, desenvolvida pelo SAS Institute, diz respeito a um processo de orientação para a um projeto de DM composto por uma sequência de etapas, nomeadamente a amostragem, exploração, modificação, modelação e a posterior avaliação[43]. Como todos os modelos, também este apresenta as suas vantagens e desvantagens. Quanto às vantagens, este método destaca-se pela fácil compreensão dos processos que estão na base para a solução de diversos problemas empresariais e para o desenvolvimento dos seus projetos, o que contribui para um desenvolvimento e manutenção organizados e adequados [47][43]. Este modelo foca-se muito mais na variante prá-

tica do problema do que na compreensão do negócio, que desmerece, sendo esta uma grande desvantagem que apresenta, para além de desprezar também processos como a avaliação e posterior implementação [47].

A metodologia utilizada para o desenvolvimento do modelo de previsão do volume de vendas neste projeto foi a CRISP-DM. Este modelo é bastante utilizado neste tipo de abordagens e uma referência no que respeita ao estudo e tratamento de grandes quantidades de dados, sendo bastante completo e, por isso, o escolhido.

### 3.1.2 CRISP-DM

O modelo CRISP-DM surgiu em 1996 e foi desenvolvido pela indústria no sentido de colmatar a necessidade que existia de um modelo padrão para orientação e processamento do DM [48] [42]. Este modelo foi criado por um consórcio que compreende empresas como Daimler-Chrysler, Teradata, NCR, SPSS (ISL), OHRA e significa "Cross Industry Standard Process for Data Mining", ou seja, processo padrão intersectorial para *data mining*, com o objetivo de ser independente em relação ao tipo de indústria, aplicação e ferramentas utilizadas [43] [45] [48].

Alguns autores descrevem o modelo como um processo fácil, confiável, bem estruturado e independente do tipo de indústria, permitindo assim a sua aplicação em diferentes cenários [42]. A grande vantagem deste modelo está relacionada com o facto de considerar, para além dos atributos técnicos, o principal objetivo do negócio em estudo [43].

O CRISP-DM é amplamente utilizado em diversos domínios, como a saúde, a educação, a engenharia, entre outros, demonstrando uma enorme versatilidade e eficácia, provando tratar-se de uma ferramenta fidedigna tanto na pesquisa como na prática industrial [42]. Esta metodologia apresenta ainda um ponto de destaque que pode ser benéfico para a maioria das organizações, que se relaciona com o facto de motivar os diversos departamentos envolvidos no projeto a um constante *brainstorming*, o que leva à intervenção de diferentes pessoas com pontos de vista distintos, acabando normalmente por contribuir para a construção de um modelo mais flexível e completo [3].

O modelo CRISP-DM é um dos principais modelos de DM e é composto por seis fases, iniciando-se na compreensão de negócio, passando depois pela compreensão dos dados, a sua preparação, modelação e posterior avaliação e implementação (figura 3.1). Ao longo de todas estas etapas é possível adaptar o modelo para que se consigam atingir os objetivos da empresa [42].

#### Compreensão do negócio (Problema)

A compreensão do negócio diz respeito ao passo de maior preponderância para que um projeto de *Data Mining* seja bem sucedido [44]. Nesta fase é efetuada uma avaliação

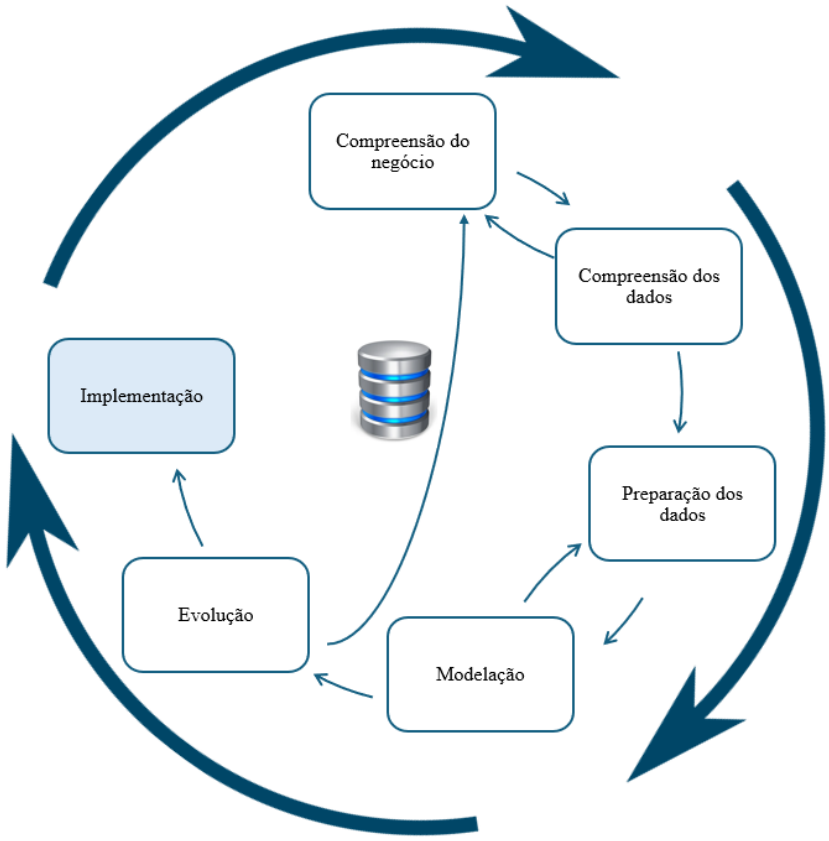


Figura 3.1: Arquitetura da metodologia CRISP-DM - adaptado de [46]

sobre o negócio em estudo com o objetivo de interpretar os recursos de que se dispõe e quais os necessários [42]. É aqui que o projeto é analisado, compreendido e definido [44]. Esta etapa foca-se na compreensão dos objetivos e dos requisitos do projeto de um ponto de vista empresarial, convertendo-os depois num problema de DM [43]. Nesta fase é preponderante que seja determinado o principal objetivo e projetado um plano preliminar realizado para atingir esse objetivo, onde deve constar informação sobre o tipo e os critérios de avaliação sobre o DM [42] [45].

### Compreensão dos dados

Para qualquer estudo é fundamental que se compreendam os dados de que se dispõe, a informação a que se tem acesso. Nesta etapa é normalmente indicada a origem dos dados e quais as suas fontes. Para esta parte do modelo são muito importantes fases como a recolha de dados, a respetiva análise, interpretação, descrição e validação da sua qualidade, sendo posteriormente determinados os seus atributos e correlações [42]. É aqui que ocorre a familiarização com os dados, onde são descobertos os primeiros *insights*, identificados os problemas relativamente à sua qualidade, detetando, por exemplo, valores ausentes ou discrepantes, ou até, de se identificarem subconjuntos para formar hipóteses sobre determinados padrões [43][3]. No entanto, apesar do principal objetivo do DM estar relacionado com a identificação de relações entre parâ-

metros ocultos e até então desconhecidos em grandes conjuntos de dados, o número de parâmetros é limitado e por isso é necessária a redução dos dados [44].

**Preparação dos dados**

Na preparação dos dados é muito importante que se realize a seleção dos dados a tratar, definindo para tal os critérios de inclusão e exclusão de informação [42]. Esta etapa envolve todos as atividades necessárias para a construção do conjunto de dados final tendo por base o conjunto de dados brutos obtidos inicialmente [43]. Os dados que sejam considerados de baixa qualidade devem ser excluídos, contrariamente aos dados que se considerem bons, que serão posteriormente tratados. As tarefas de preparação dos dados podem ser realizadas inúmeras vezes e incluem seleção de tabelas, atributos (tendo por base o plano definido na primeira etapa), registos, transformação e limpeza de dados para serem posteriormente utilizados nas ferramentas de modelação [45] [42]. Aqui é então realizada a integração, seleção, limpeza, construção e formatação dos dados como preparação para a próxima fase [44].

**Modelação**

De acordo com o tipo de negócio presente na base do problema e do tipo de dados disponíveis é realizada a modelação dos dados. Este processo consiste na definição da técnica de modelação a utilizar, construção do modelo e criação de casos de teste a serem usados [42]. Pela grande quantidade de técnicas e algoritmos de modelação existentes, a fase da modelação em projeto de DM é bastante complexa [44]. É aqui que são selecionadas as técnicas de modelação a aplicar e realizada a calibração dos seus parâmetros para valores ótimos [43]. Algumas técnicas têm requisitos específicos na forma de dados, o que significa que, mediante as técnicas selecionadas, pode surgir a necessidade de voltar à fase de preparação dos dados [45]. Nesta etapa torna-se benéfico treinar, testar e analisar mais do que um algoritmo ou modelo para que, na fase seguinte, se possa fazer a comparação entre eles através dos resultados obtidos [42]. No caso de serem utilizados vários modelos, estes devem ser testados individualmente. Nesta fase deve ainda ser feita uma reflexão e análise com os especialistas dos domínios em estudo para que se comecem a discutir os resultados obtidos até então no contexto do negócio [3].

**Avaliação**

Nesta fase são definidas as métricas a ser usadas para analisar e comparar os modelos utilizados [42]. Um ponto chave desta etapa passa pela verificação de algumas questões de negócio importantes que, por um qualquer motivo, possam não ter sido consideradas [45]. Tendo por base os objetivos definidos é realizada a avaliação e possível validação dos modelos, interpretados os resultados obtidos, e, caso se verifique a

necessidade, definidas novas ações a serem tomadas [42]. É aqui que os modelos obtidos são validados. No final desta fase é fundamental que seja alcançada uma decisão sobre o uso dos resultados obtidos, se avança para posterior implementação ou se volta para uma das fases iniciais do processo [45].

**Implementação**

Depois de todos os processos descritos anteriormente vem a implementação do modelo, fase que integra todo o planeamento que envolve a sua aplicação, bem como subsequente monitorização e manutenção [42]. A criação do modelo pode não ser o fim do projeto, ainda que normalmente permita aumentar o conhecimento sobre os dados, existe a necessidade de o organizar e apresentar de forma compreensível para auxiliar a suportar as tomadas de decisão nas organizações [43][45]. De acordo com os requisitos do projeto, a fase de implementação pode ser uma coisa tão simples como a realização de um relatório final ou tão complexo como a criação de um software ou implementação de um processo de DM para toda a empresa [45]. De destacar que, em muitos estudos esta etapa não chega a ser realizada, dependendo dos resultados e avaliação alcançados na fase anterior, que, não estando de acordo com o pretendido, podem levar a essa tomada de decisão [42] .

## 3.2 Ferramentas

Para o desenvolvimento do projeto foram utilizadas algumas ferramentas, selecionadas de forma ponderada e criteriosa com o objetivo de atingir os resultados pretendidos. Para esta seleção foi realizada uma pesquisa e analisados vários tipos de programas para que dentro das possibilidades existentes, de acordo com os recursos disponíveis e tendo em conta os objetivos pretendidos, fosse realizada a seleção daqueles que se consideraram ser os mais indicados.

### 3.2.1 Folhas de cálculo

Numa fase inicial, depois de recolhidos os dados da empresa, foi utilizado o Excel, uma vez que os dados facultados pela empresa também vinham nesse formato. O Excel é uma ferramenta da Microsoft bastante reconhecida e utilizada para estudo, análise de dados e realização de cálculos. Esta ferramenta possibilitou uma breve verificação da quantidade dos dados logo à partida e uma rápida compreensão sobre a qualidade dos mesmos para que depois se procedesse ao seu estudo.

Através deste programa foi possível realizar uma exploração preliminar dos valores disponíveis. Através da aplicação de filtros sobre os dados, da realização de tabelas dinâmicas para estudar os valores e de diversas representações gráficas foi possível

agrupar, segmentar e resumir a informação de forma flexível. Assim, conseguiu-se obter uma primeira perceção sobre o comportamento dos dados, a sua distribuição, tendência e estudar determinados pontos específicos que se viriam a revelar importantes para o projeto.

Esta abordagem exploratória teve como principal objetivo compreender a estrutura do *dataset*, identificar padrões ou anomalias e preparar a base para análises mais aprofundadas em ferramentas analíticas mais especializadas.

### 3.2.2 Ferramentas de *Business Intelligence*

Para esta etapa do problema optou-se por uma solução diferente, que fosse capaz de proporcionar uma análise de dados completa, fidedigna e ao mesmo tempo visual. Foi aí que se começou a entrar no mundo do *Business Intelligence* (BI), uma área que permite uma tomada de decisões e a definição de estratégias devidamente fundamentadas tendo por base uma análise de dados, sejam eles históricos ou atuais. O mundo do BI possibilita a aquisição de informação a partir de um conjunto de dados, informação de valor que poderá ser determinante para as organizações do ponto de vista estratégico, analítico e operacional[1].

Dentro da área do BI existem inúmeras soluções que, mediante o pretendido, podem ser equacionadas para a realização do estudo sobre um qualquer conjunto de dados. Se por um lado existem soluções que são de acesso gratuito, diga-se que aqui o número de oferta é extremamente limitado, por outro lado, existem outras que são pagas, e mediante os objetivos e o que se pretenda efetivamente do programa, existem até diferentes valores para as respetivas mensalidades. As que são gratuitas são também conhecidas como as de código aberto e, por isso mesmo, qualquer pessoa consegue ter acesso a elas, ou seja, qualquer pessoa pode ter acesso ao código fonte e com isso alterar, seja com o intuito de o melhorar, ou simplesmente para o atualizar ou aferir/inspecionar. As ferramentas que são pagas são desenvolvidas por um determinada empresa, também elas sujeitas a atualizações do *software* base para eliminação de pequenos *bugs* ou mesmo em formato de progressão para implementar novas funcionalidades do programa. No entanto, apresentam um nível de segurança diferente, muito superior, com acesso restrito ao utilizador em que é garantido desde logo que apenas ele tem acesso aos dados em estudo, ou alguém com quem ele os queira partilhar.

Inicialmente foi efetuado um estudo sobre os tipos de ferramentas existentes e quais as principais vantagens de cada uma delas, ficando evidentes determinadas lacunas que umas tinham em detrimento de outras. Contudo, grande parte das ferramentas poderia ser considerada como solução para o que era pretendido. Dentro da análise efetuada foram ponderados diversos fatores, nomeadamente a facilidade de aprendizagem,

[1] https://www.zendesk.com.br/blog/o-que-e-business-intelligence-para-que-serve/ (última consulta em 2025-05-18).

qual a dimensão dos dados a trabalhar, quais as plataformas de maior reconhecimento e que por isso fossem de mais fácil implementação futura na organização, quais as que dariam a possibilidade de produzir relatórios com notas sobre os dados em análise, quais as que teriam já uma ligação com a inteligência artificial, qual a limitação de utilizadores para a plataforma, qual o custo, a compatibilidade com a fonte dos dados, quais os recursos de colaboração entre os utilizadores e disponibilizados pela empresa que o desenvolveu, entre outras.

Dentro das ferramentas de BI que existem, que são já bastantes e com tendência a surgirem cada vez mais, houve três que desde o início da pesquisa se destacaram e por isso foram alvo de uma análise mais profunda, nomeadamente o Apache Superset, o Tableau e o Power BI.

Fundado em 2003, o Tableau surge como resultado de um projeto de ciência da computação em Stanford. O objetivo dos seus cofundadores, Chris Stolte, Christian Chabot e Pat Hanrahan, passava por tornar o estudo sobre dados mais acessível às pessoas. Com esse intuito, desenvolveram e patentearam a tecnologia que está na base deste programa, o VizQL, que tem a capacidade de transformar ações de arrastar e soltar em análise de dados por intermédio de uma "interface intuitiva". Posteriormente, esta ferramenta foi adquirida e atualmente faz parte da Salesforce, uma empresa líder no desenvolvimento de softwares de gestão de empresas, sobretudo no que ao CRM (Customer Relationship Management) integrado diz respeito. Esta plataforma foi desenvolvida para proporcionar às empresas um maior aproveitamento da potencialidade dos seus dados, transformando a forma como estes podem ser analisados, potencializando assim a sua capacidade para a resolução de problemas. Este programa possibilita melhorar as análises sobre os dados e torná-las mais acessíveis e visualmente mais atraentes a toda a gente. Por intermédio de uma *interface* intuitiva é possível que se consigam realizar análises bastante rápidas sobre um determinado conjunto de dados e com isso descobrir *insights* que se podem tornar bastante valiosos para as organizações[2].

O Power BI é uma ferramenta de BI desenvolvida pela Microsoft e que pode ser explorada de diversas formas, consoante o pretendido. Por isso, existem diferentes variantes, com diferentes valores associados, desde o gratuito até ao empresarial. O incremento de funcionalidades a explorar no programa é proporcional ao valor que o utilizador, ou a organização, terão que pagar para conseguir usufruir das respetivas potencialidades. Esta ferramenta é bastante completa e por isso uma referência no que toca à análise de dados num formato mais interativo e visual, com a capacidade de transformar dados não relacionados em informação extremamente útil e coerente, visualmente interativas e envolventes para os utilizadores[3]. Para além desta análise, o programa pode ainda auxiliar ao nível de relatórios, seja para realização ou apenas publicação, o que pode

[2]https://www.tableau.com/why-tableau/what-is-tableau (última consulta em 2025-05-18).

[3]https://learn.microsoft.com/pt-pt/power-bi/fundamentals/power-bi-overview (última consulta em 2025-05-18).

ser uma mais valia para determinadas posições de responsabilidade, tanto para orientação nos dados importantes a realçar ou simplesmente ao nível da organização. O Power Bi é uma ferramenta bastante recorrente no mundo dos negócios, utilizada por muita gente e muitas vezes de forma completamente diferente, tratando-se de um programa que se adapta facilmente às diferentes realidades, funções e necessidades do seu utilizador.

Outra ferramenta muito conhecida nesta área, e que por isso foi também equacionada, é o Apache Superset, uma alternativa *opensource* (código aberto) para análise e exploração visual de dados[4]. Esta plataforma foi desenvolvida inicialmente pela Airbnb, sendo posteriormente cedida à Apache Software Foundation, e é bastante utilizada nas mais diversas áreas de gestão. Gradualmente, o Apache tem vindo a ganhar destaque e a tornar-se um dos principais instrumentos de análise de dados em tempo real, com a capacidade de integrar dados de diversas fontes. Outra grande vantagem desta ferramenta é que é extremamente personalizável, proporcionando múltiplas opções de visualização e análise, além de outros recursos avançados de colaboração e segurança, o que possibilita o trabalho de diversas pessoas no mesmo projeto[5]. Esta plataforma esteve muito próxima de ser a escolhida para este projeto, uma vez que estamos a falar de um programa com bastante potencial e que é gratuito. Este critério, o custo da ferramenta, é bastante importante para o projeto uma vez que não existe um financiamento para o suportar, ou seja, quanto menores forem os custos para o seu desenvolvimento, melhor. No entanto, depois de alguma ponderação e analisadas as vantagens e desvantagens da utilização deste programa, outras questões acabaram por surgir e por se sobrepor, acabando a opção por recair sobre o Power BI.

Na tabela 3.1 são comparadas algumas das principais características que foram consideradas e analisadas na pesquisa realizada sobre as três ferramentas acima descritas, retiradas e adaptadas do site docskanaries[6].

De acordo com a pesquisa efetuada constatou-se que, a nível de aspeto visual, o Power Bi e o Tableau se tratam de ferramentas que possuem recursos visuais mais interessantes do que o Apache Superset, que dispõe também de visualizações bastante funcionais e de extrema eficácia, mas que podem não ser tão atraentes esteticamente. O Apache superset é a único das três ferramentas que é de código aberto, o que possibilita que vários utilizadores desenvolvam código para o programa e que o utilizador, percebendo ele de programação, consiga adaptar o programa à sua necessidade em específico. No que confere a custos, temos aqui uma grande diferença entre os programas, já que o Apache Superset é de acesso completamente gratuito, ou seja, qualquer pessoa é livre de utilizar. O Power Bi que tem duas variantes, uma que é também ela gratuita,

[4]https://superset.apache.org/docs/intro/ (última consulta em 2025-05-18).

[5]https://jlgjosue.medium.com/apache-superset-uma-plataforma-moderna-de-exploração-e-visualização-de-dados-em-tempo-real-c5b23e7ddccb (última consulta em 2025-05-18).

[6]https://docs.kanaries.net/pt/articles/apache-superset-vs-tableau (última consulta em 2025-05-18)

Tabela 3.1: Resumo de ferramentas de Business Intelligence - Fonte: docskanaries

| **Cacterísticas** | **Modelos BI** | | |
|---|---|---|---|
| | **Apache Superset** | **Power BI** | **Tableau** |
| Aspeto Visual | Moderado | Elevado | Elevado |
| Código Aberto | Sim | Não | Não |
| Custo | Sem custo | Sem custo e Pago | Pago |
| Personalização | Elevada | Moderada | Elevada |
| Quantidade de Base de Dados Suportadas | Elevada | Elevada | Elevada |
| Suporte técnico | Moderado | Elevado | Elevado |
| Usabilidade | Moderada | Elevada | Elevada |

mas com limitações no que toca a determinadas funcionalidades, e outras variantes pagas cujos valores diferem de acordo com o incremento de funcionalidades que se pretendam explorar. Por último o Tableau, que apenas tem a versão paga, e a mais dispendiosa de todas, sendo por isso de acesso mais reservado e limitado. No que toca à personalização existe uma forte capacidade de personalização de *dashboards*, painéis, gráficos e dos restantes componentes estruturais de análise que possibilitam aos utilizadores uma visualização muito mais caraterística, distinta no Apache Superset e no Tableau do que no Power BI que dispõe de algumas opções de personalização mas é mais limitada em determinados aspetos. Relativamente à quantidade de base de dados suportadas temos uma ampla gama para todos os programas já que todos possibilitam trabalhar com uma grande quantidade de base de dados, permitindo assim aos seus utilizadores a capacidade de analisar dados das mais diversas fontes. Quanto ao suporte técnico, existem algumas diferenças a considerar, isto porque, o Power BI e o Tableau, como aplicações pagas, conferem um suporte técnico diferenciado, neste caso superior, quando comparado com o suporte técnico de um programa que é totalmente gratuito, como o Apache Superset. Todos eles dispõem de suporte ao nível da comunidade, nomeadamente através de *blogs*, páginas de utilizadores, fóruns de debate, entre outros, no entanto, para os utilizadores do Apache, este é um auxílio preponderante já que estamos a falar de uma plataforma de código aberto e gratuita. No que à usabilidade diz respeito existem também aqui diferentes níveis a considerar, já que se está perante programas que conferem diferentes padrões de facilidade para a sua utilização, desde a instalação até ao manuseamento. Neste ponto, o Power BI e o Tableau são

considerados mais leves quanto à facilidade de uso e de instalação, com uma *interface* mais acessível relativamente ao Apache Superset que é considerado o mais complicado dos três.

Na realidade, todas estas ferramentas seriam bastante válidas para o projeto e ao nível de funcionalidades, qualquer uma delas podia ser a escolhida. Contudo, salientaram-se aqui algumas funcionalidades e determinados critérios que, para futura implementação na empresa, como é objetivo, pesaram bastante na decisão e por isso o escolhido foi o Power Bi em detrimento das restantes opções. O primeiro foi o facto de ser um programa intuitivo, de relativa facilidade de aprendizagem e de simples instalação; outro fator esteve relacionado com o sistema operativo utilizado em todos os computadores da empresa, o Windows, o que facilita a operacionalidade e integração do programa; a terceira razão foi o facto da empresa já utilizar o Office, o que facilita na introdução de um programa deste tipo e ao nível da aprendizagem auxilia bastante porque existe uma determinada correlação entre as diferentes aplicações que o constituem, logo, seria mais fácil a adaptação à ferramenta; por último, o facto da empresa ter de renovar com determinada periodicidade as licenças para o Office, sendo tudo da mesma empresa, a Microsoft, acaba por ser mais benéfico negociar tudo junto.

### 3.2.3 Linguagem de programação

Para se realizarem análises de dados de forma rápida, intuitiva, altamente visual, em formatos de rápida compreensão, analogia e com a possibilidade de obtenção de *insights* que se podem revelar como mais valias para as mais diversas organizações, temos então o BI, como explicado acima. No entanto, este projeto tem como principal objetivo não só estes pontos como também o desenvolvimento de um modelo de previsão de vendas tendo por base um conjunto de dados, sobretudo históricos. Nesta fase do projeto começaram então a equacionar-se as ferramentas que se poderiam utilizar para a execução do modelo, tendo surgido aqui o Python como forte possibilidade.

O Python é uma linguagem de programação extremamente poderosa e altamente reconhecida na ciência de dados, para o desenvolvimento de *software*, de ferramentas, aplicações, *Machine Learning*, entre outros. Esta ferramenta foi escolhida sobretudo pela sua enorme capacidade, pelo reconhecimento no aumento de produtividade, uma vez que se pode escrever um programa com menos linhas de código do que em muitas outras linguagens de programação, pela enorme quantidade de bibliotecas padrão de que dispõe e pelo grande apoio disponível na internet, já que existem milhares de programadores Python no mundo, múltiplas plataformas de suporte e conteúdo de apoio e aprendizagem disponíveis. Posteriormente, verificou-se a forma mais intuitiva para se utilizar esta linguagem e qual seria a melhor forma para a sua introdução, surgindo assim o Anaconda.

O Anaconda é muito popular no mundo da programação sobretudo porque contém

as principais ferramentas para inteligência artificial, ciência de dados e ML. Esta ferramenta é reconhecida como um ecossistema amplamente utilizado no que a análise de dados e Python diz respeito, permitindo a instalação de diferentes versões de linguagem para criar ambientes de desenvolvimento diferentes, consoante as necessidades[7]. Dentro desta plataforma, optou-se pelo Spyder, um IDE (Integrated Development Environment), ambiente de desenvolvimento integrado Python que possibilita ciclos curtos de interação, muitos testes e rápido *feedback* durante a sua programação, para além de compatibilidade e integração de várias bibliotecas[8]. Este ambiente apresenta muitos recursos de visualização e diversas funcionalidades de edição, análise e desenvolvimento de ferramentas para exploração, análise, execução e inspeção de dados[9], sendo por isso considerada a ferramenta ideal para o desenvolvimento do modelo preditivo.

[7]https://dojo.bylearn.com.br/python/o-que-e-anaconda-para-python/ (última consulta em 2025-05-18).

[8]https://www.spyder-ide.org/ (última consulta em 2025-05-18).

[9]https://sourceforge.net/projects/spyder.mirror/ (última consulta em 2025-05-18).

# 4 Análise exploratória dos dados

Para a realização deste projeto foi utilizado um conjunto de dados cedidos pela empresa em estudo, referentes à sua atividade durante o período compreendido entre 2010 e 2020.

O conjunto de dados explorados estava separado em dois ficheiros, um deles que continha informação apenas sobre a faturação da empresa e outro com informação referente apenas a notas de crédito emitidas durante o mesmo período. Cada um destes ficheiros continha diversos campos de informação, nomeadamente números das faturas, datas de faturação, a que filial da empresa pertenciam, o código de produto, a família de produto, a quantidade faturada e ainda o valor de faturação. Contudo, para o estudo realizado, não foram utilizados todos estes campos de informação, sendo selecionados apenas aqueles que foram considerados como fundamentais, ou seja, datas de faturação, a filial à qual pertenciam e valores de faturação. Inicialmente foi efetuada uma análise separada sobre cada um dos ficheiros, primeiro sobre a faturação e depois sobre as notas de crédito, o que permitiu tirar algumas conclusões importantes sobretudo sobre as filiais a título individual. Posteriormente foram juntos os dois ficheiros num só, e realizou-se uma análise semelhante à realizada anteriormente mas agora referente à faturação global efetiva, conseguindo-se aqui uma análise mais apurada sobre o comportamento global da empresa. No decorrer deste projeto, considerou-se como faturação efetiva o valor da faturação bruta sem o valor das notas de crédito associadas. Esta métrica representa o montante líquido efetivamente realizado pela empresa, refletindo de forma mais precisa o volume de receita após correções, devoluções ou ajustes comerciais realizados.

Numa primeira fase foi verificada a qualidade dos dados de que se dispunha, ou seja, começou por se verificar se todos os dados eram válidos, coerentes, se não existiam zeros, se havia alguns que tinham de ser eliminados ou desconsiderados. De seguida, começaram então a ser criados os filtros para fazer a seleção efetiva dos dados considerados como ideais para o desenvolvimento do estudo, com o objetivo de obter os melhores resultados possíveis.

Na organização em estudo existem sete filiais atualmente em atividade, cada uma delas com um número atribuído para identificação desde a sua criação. A sede da empresa, localizada na cidade do Porto, é representada pelo número 1, a filial de Lisboa pelo número 2, a de Coimbra pelo 3, Braga pelo 4, Olhão pelo 6, Setúbal pelo 7 e Aveiro pelo 8 (tabela 4.1). Ao longo dos anos outras filiais foram abertas mas entretanto já fecharam,

Tabela 4.1: Tabela de filiais e respetivas localizações

| Filial | Localização |
|---|---|
| 1 | Porto |
| 2 | Lisboa |
| 3 | Coimbra |
| 4 | Braga |
| 6 | Olhão |
| 7 | Setúbal |
| 8 | Aveiro |

ficando apenas em consideração aquelas que estão atualmente a laborar. Dentro dos dados disponibilizados pela empresa existia informação referente a uma filial representada pelo número 9, que foi desconsiderada precisamente por não estar a trabalhar desde o primeiro trimestre de 2011, sensivelmente.

## 4.1 Dataset

Selecionados os dados sobre os quais se iria trabalhar, começou então a parte da análise dos mesmos, o estudo mais detalhado já sobre a seleção que iria ser a base para o desenvolvimento de todo o projeto. Como já referido, começou por se analisar o documento referente apenas à faturação e verificar qual o nível de faturação global da empresa, para as sete filiais durante o período em estudo, obtendo-se um valor próximo dos 64,92 milhões de euros. De seguida começou a verificar-se qual o nível de faturação por filial para o total de anos em estudo (Figura 4.1) e posteriormente por ano. Relativamente à faturação global pôde verificar-se um destaque para duas filiais, a que corresponde à sede da empresa, a filial 1, e para a filial 2, sendo as filiais que apresentam maior volume de faturação todos os anos, a rondar os 20,5 milhões de euros, sendo a filial 2 a que se coloca no pódio. Logo depois vêm as filiais 3 e 4 respetivamente, com valores próximos dos 8,2 milhões de euros, seguidas da filial 8 com uma faturação global à volta dos 4 milhões de euros, a filial 6 com uma faturação de 2,2 milhões e, por último, a filial 7 com metade da faturação desta última, cerca de 1,1 milhões de euros.

Depois de uma primeira análise relativamente à faturação da empresa baseada apenas nas vendas realizadas nos dez anos em estudo, foi também realizada uma análise sobre as notas de crédito. O principal objetivo passava agora por compreender quais os valores creditados no geral, para cada uma das filiais individualmente, por ano, e ainda se era possível retirar alguma analogia relativamente aos dados para ambas as variantes estudadas até aqui, faturação e notas de crédito. Depois de analisados os valores, pôde então verificar-se que a filial 2 é aquela que apresenta o maior nível de créditos passados, com um valor de notas de crédito a rondar os 1,5 milhões de euros no total de anos em estudo. De seguida encontra-se a filial 4 com um valor na ordem dos 400

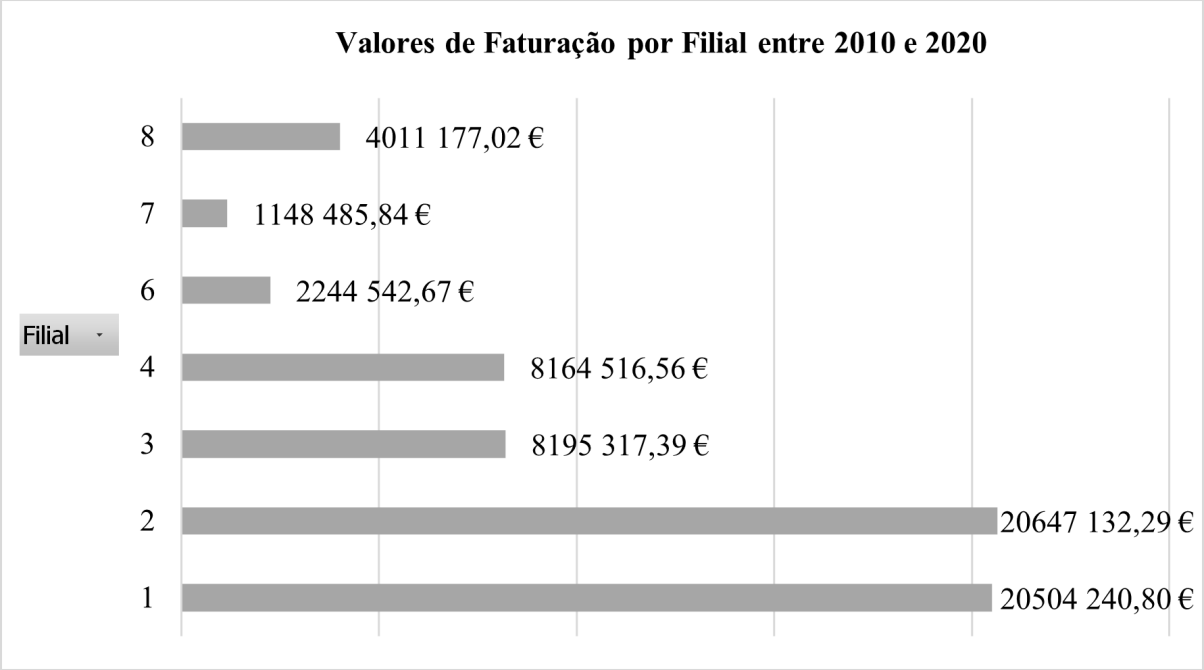


Figura 4.1: Valores de faturação por filial da empresa de 2010 a 2020

mil euros, seguindo-se a filial 1 com um valor creditado pouco acima dos 320 mil euros. Logo depois vem a filial 3, com 150 mil euros, a filial 8 com um valor creditado muito próximo dos 56 mil euros e, por último, as filiais 6 e 7 com valores inferiores a 30 mil euros (Figura 4.2). Após esta análise, pôde-se constatar e tirar algumas ilações interessantes, sobretudo quando são comparados os resultados de faturação com os resultados referentes às notas de crédito. A filial 2, que de acordo com a análise inicial, os dados de faturação baseado apenas em vendas, é aquela que mais fatura, é também a que maior valor em notas de crédito passa, enquanto que a filial 1, que apresenta um nível de faturação semelhante, tem um valor de notas de crédito passadas menor cerca de cinco vezes. A filial 4, por sua vez, apresenta um valor de notas de crédito passadas maior que a filial 1, no entanto, fazendo a mesma relação relativamente à faturação, a mesma filial 4 fatura cerca de metade do valor da filial 1. Posteriormente foram calculados os rácios entre os valores de faturação e de notas de crédito para que se percebesse quanto é que as notas de crédito passadas por cada filial representavam face ao respetivo nível de faturação. A filial que apresenta melhor valor é a filial 6, com um rácio de 1,17%, o que significa que as notas de crédito passadas por esta filial representam cerca de 1,17% do seu valor de faturação para o total de anos em estudo. De seguida vem a filial 8 com 1,39%, seguindo-se a filial 1 com 1,57% , depois a filial 7 com 1,71%, a seguir a filial 3 com 1,81%, a filial 4 com 4,85% e por último a filial 2 com 7,20%, porque como explicado acima, sendo das filiais que mais fatura, é também a que apresenta maiores valores creditados (Figura 4.3).

Depois de realizadas estas primeiras análises sobre os dois ficheiros iniciais de que se dispunha em separado, ou seja, faturação apenas com valor das vendas por um lado

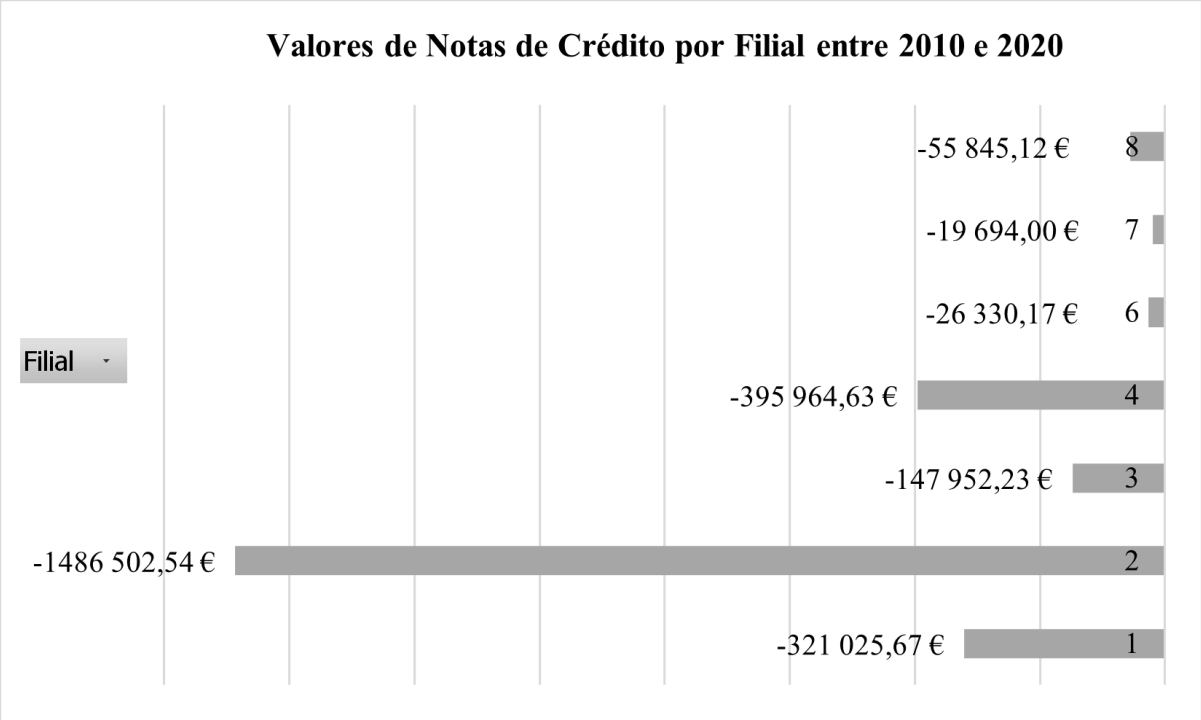


Figura 4.2: Valores de notas de crédito por filial de 2010 a 2020

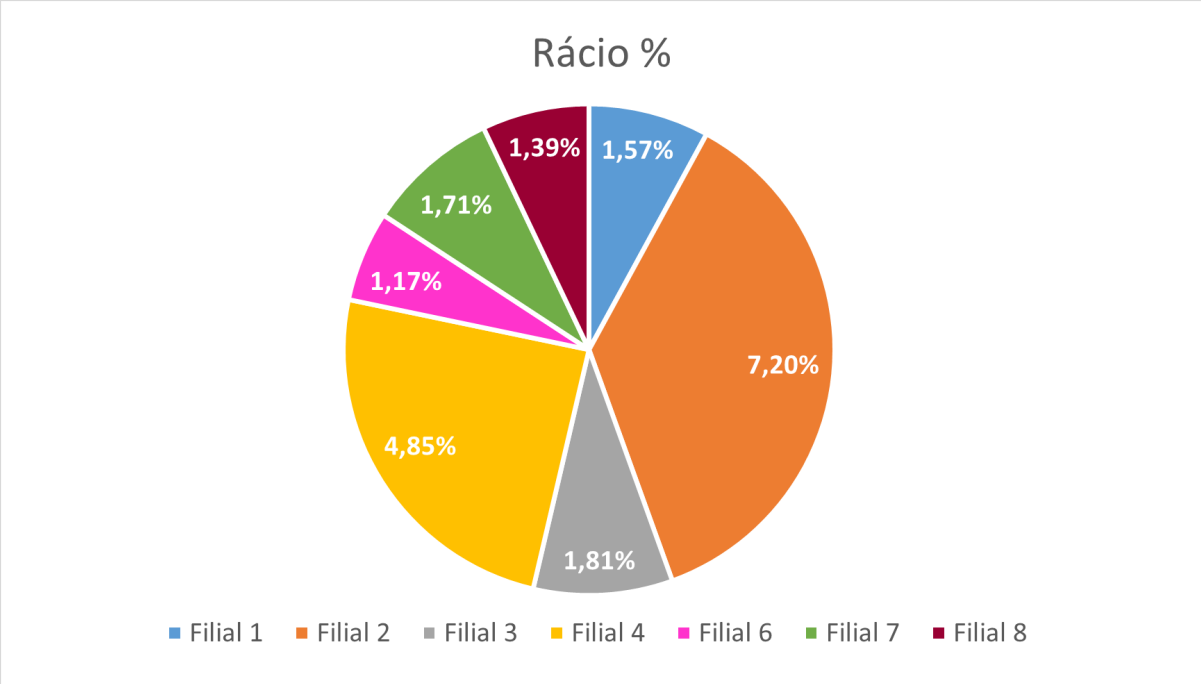


Figura 4.3: Rácio entre valores de notas de crédito e faturação por filial de 2010 a 2020

e notas de crédito por outro, procedeu-se à junção dos dados num mesmo ficheiro para que se procede-se à análise geral dos dados da empresa como um todo, ou seja, faturação efetiva, trabalhando-se apenas com este ficheiro a partir daí para a realização do projeto.

Já com os dados todos juntos, foi-se verificar qual o nível de faturação efetiva da em-

presa no geral, obtendo-se um valor de cerca de 62,46 milhões de euros para o período em estudo. De seguida foram-se analisar quais os níveis de faturação efetiva para cada uma das filiais da empresa individualmente e, face ao descrito anteriormente, a maior diferença a assinalar foi referente à filial 2. Na realidade, e feita a análise da faturação com notas de crédito, a filial 1 foi a que apresentou maior nível de faturação, um pouco acima dos 20 milhões de euros, seguida da filial 2, que faturou um valor ligeiramente superior a 19 milhões de euros, invertendo agora os lugares das duas primeiras posições no que toca à faturação, isto comparando os valores com a primeira análise realizada. De seguida vem a filial 3 com um valor de faturação a rondar os 8 milhões de euros, a filial 4 com aproximadamente 7,8 milhões de euros e a filial 8 com cerca de 4 milhões. Posicionadas nas duas últimas posições temos na mesma as filiais 6 e 7, com uma faturação muito semelhante à verificada na primeira análise, a rondar os 2,2 milhões de euros e os 1,1 milhões de euros respetivamente (Figura 4.4).

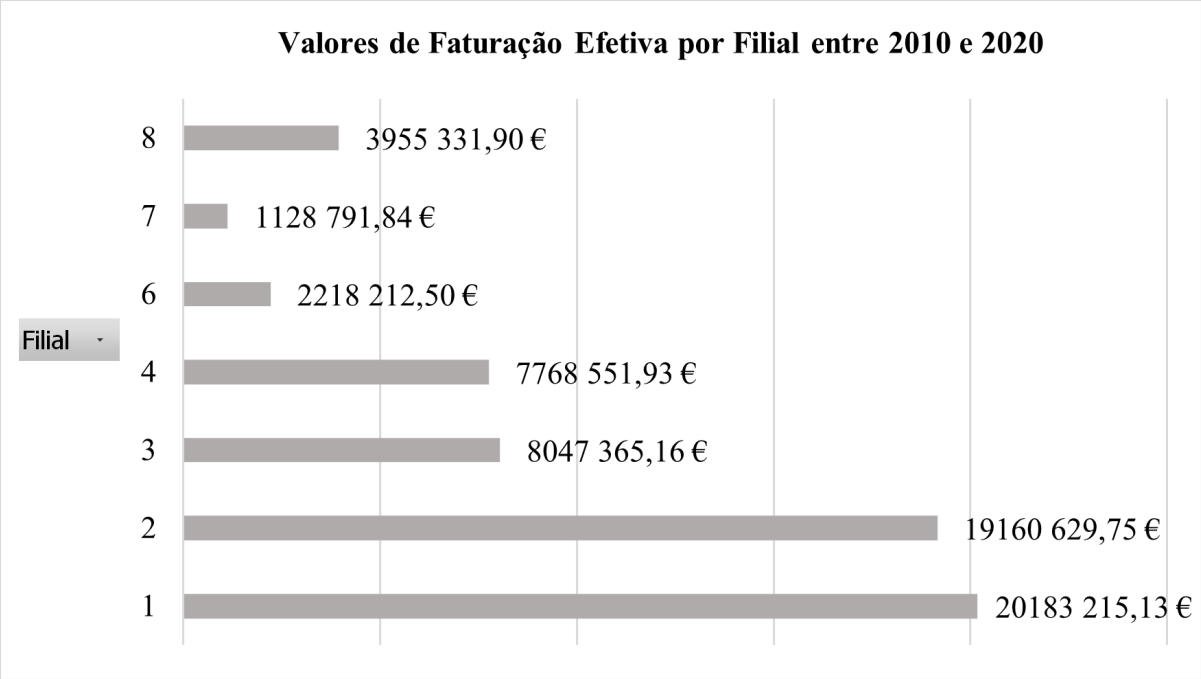


Figura 4.4: Valores de faturação efetiva por filial de 2010 a 2020

Outra análise realizada recaiu sobre o nível de faturação efetiva por ano da empresa, entre 2010 e 2020 (Figura 4.5). De acordo com esta análise e recorrendo ao gráfico da figura 4.5 para a fundamentar, pode-se verificar que os melhores anos de faturação da empresa foram os anos de 2019 e 2018, respetivamente, com um volume de faturação acima dos 6 milhões de euros e o pior foi o ano de 2013 com um valor faturado pouco acima dos 5 milhões. A média de faturação apresentada pela empresa para o total de filiais em estudo é de aproximadamente 5,5 milhões, um valor perfeitamente compreensível já que os níveis de faturação estão compreendidos entre os 5 e os 6,5 milhões para os anos em estudo. Um detalhe importante a realçar dentro deste módulo da análise é que no ano de 2020, ano em que surgiu a pandemia do COVID-19 a

nível mundial, a empresa apresentou um volume de faturação bastante positivo face aos anos anteriores, sendo inclusivamente o quarto melhor ano de todo o período em estudo, o que significa que, independentemente disso, a empresa não foi abalada por esse fator externo.

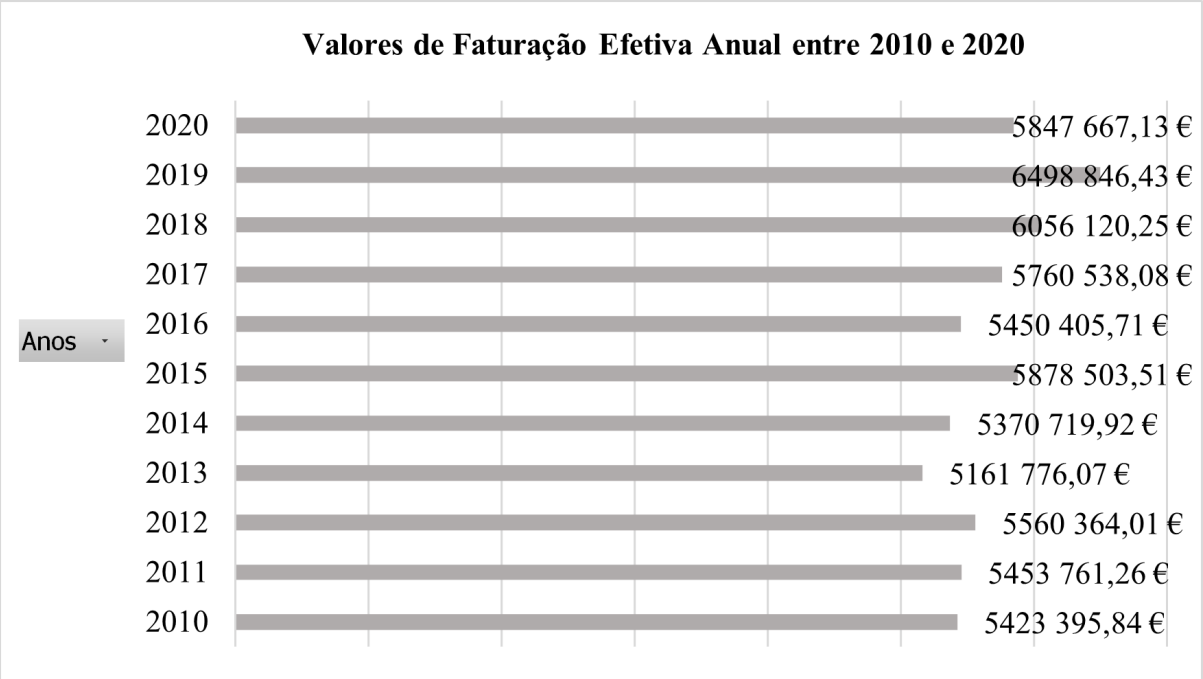


Figura 4.5: Valores de faturação efetiva anual da empresa de 2010 a 2020

Um ponto importante para o estudo dos dados e por isso também alvo de análise foi a verificação da evolução do volume de faturação ao longo dos meses para cada uma das filiais durante os anos em estudo (Figura 4.6). O objetivo desta parte da análise prendia-se com a possibilidade de existir algum tipo de sazonalidade no que à faturação das filiais diz respeito. Analisado o gráfico da figura 4.6 pode-se verificar que existem diferentes comportamentos para cada uma das filiais, contudo, consegue-se ter uma noção dos períodos em que grande parte das variações são semelhantes, como era a intenção. As três filiais com menor volume de faturação apresentam um nível de faturação mais homogéneo, ou seja, com muito poucas variações, contrariamente ao que se verifica nas quatro filiais com maior faturação, cujas variações são constantes. De acordo com os dados de que se dispõe, o meses de março, maio, julho e setembro são aqueles que apresentam maior crescimento, contrariamente aos meses de abril, agosto e dezembro que apresentam sempre reduções de faturação na linha cronológica, permitindo assim tirar algumas relações.

O incremento no mês de julho pode ser justificado pelas muitas intervenções de manutenção nas empresas das mais diversas indústrias que são realizadas no mês de agosto, sendo o material necessário para sua realização preparado antecipadamente, normalmente no mês de julho. Como a empresa comercializa produtos de desgaste, consu-

míveis relacionados com a manutenção industrial, é normal que tenha um aumento de faturação neste mês. Para o mês de agosto pode verificar-se uma redução na faturação precisamente pelo descrito anteriormente, já que é o mês em que as intervenções são realizadas e por isso o material já está adquirido antecipadamente pelas empresas, resta intervir e solucionar algo que seja para uma manutenção mais corretiva ou de ocorrência. Para além deste motivo, um outro que justifique estes valores em agosto pode ser o facto de muita gente tirar férias nesta altura, inclusivamente uma grande parte dos colaboradores da empresa, ficando apenas a laborar os serviços mínimos, ou ainda o facto de muitas fábricas de diversas marcas fecharem completamente durante todo este período. O mês de setembro é um mês pós férias, pós intervenções e de reabertura das fábricas das marcas, e, por isso, um mês de alavancagem para o funcionamento da indústria na sua normalidade, o que poderá ser um forte argumento para o aumento dos volumes de faturação neste período. No final do ano, no mês de dezembro, é tendencialmente verificada a redução da faturação, o que se pode justificar por ser um mês com dois feriados em duas semanas seguidas, que, sendo estes durante a semana são muito propícios a que se tirem férias, levando assim a um abrandamento do ritmo e do normal funcionamento do mercado. Para além disso é um período de festas, de fecho de ano, controlo de inventário e de pequenas intervenções ao nível da manutenção na maioria das empresas para que se inicie o novo ano com força, motivando tudo isto a que ocorra essa redução.

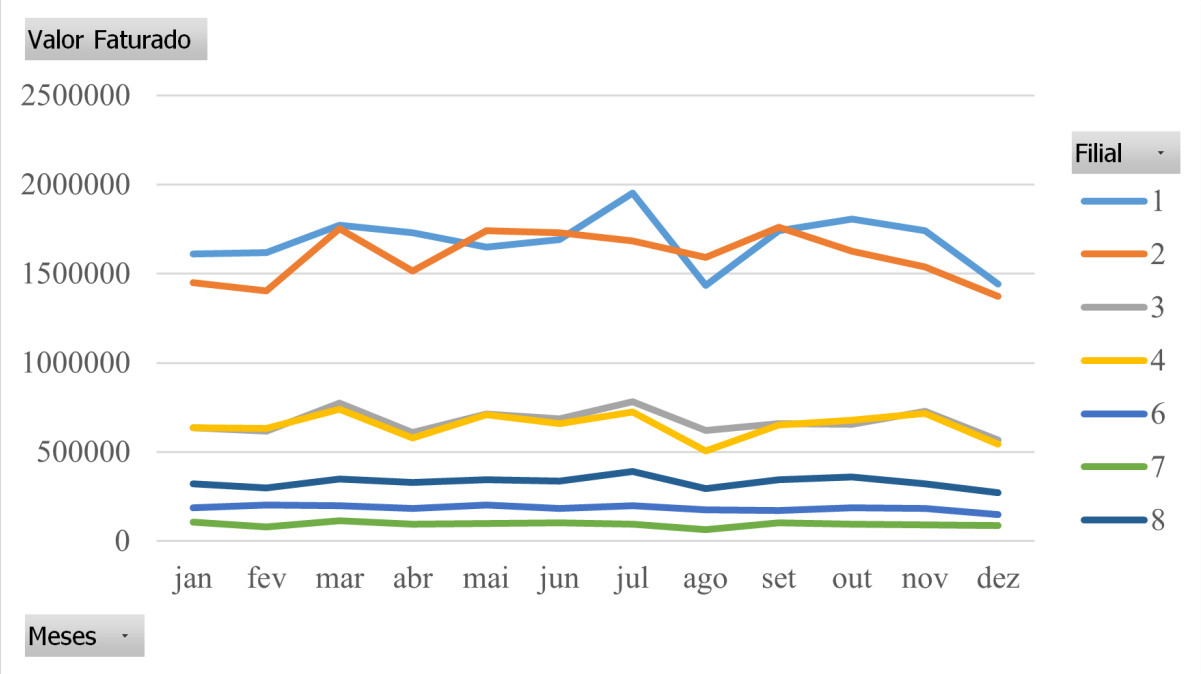


Figura 4.6: Evolução da faturação efetiva mensal por filial de 2010 a 2020

Para além desta análise, fez-se ainda uma outra onde foram comparados os volumes de faturação ao longo dos anos para a mesma filial, com o objetivo de verificar se os resultados financeiros alcançados em cada ano tinham sido melhores ou piores face ao

ano anterior. Para além deste objetivo pretendia-se ainda perceber qual o percentual que correspondia a essa diferença de faturação para que depois se conseguisse analisar se esse valor estaria acima ou abaixo do valor da taxa de inflação para o respetivo ano. Os valores das taxas de inflação para o período em estudo, entre 2010 e 2020 foram retiradas do site dadosmundiais[1], onde se verificou o histórico das taxas de inflação anuais em Portugal (Tabela 4.2).

Tabela 4.2: Histórico de taxas de inflação anual em Portugal - Fonte: dadosmundiais

| **Ano** | **Taxa de inflação em Portugal (%)** |
|---|---|
| 2010 | 1,40 |
| 2011 | 3,65 |
| 2012 | 2,77 |
| 2013 | 0,27 |
| 2014 | -0,28 |
| 2015 | 0,49 |
| 2016 | 0,61 |
| 2017 | 1,37 |
| 2018 | 0,99 |
| 2019 | 0,34 |
| 2020 | -0,01 |

No que respeita às diferenças de faturação pôde verificar-se que nenhuma das filiais da empresa apresentou resultados crescentes contínuos relativamente ao ano anterior, tendo-se observado variações constantes no decorrer do período em estudo. Em determinados anos verificaram-se aumentos de faturação face aos anos transatos enquanto que para outros pôde verificar-se uma diminuição. Por exemplo, no gráfico da figura 4.7 estão representadas as diferenças de faturação efetiva anual da filial 1 ao longo do tempo, onde se podem verificar se as variações são positivas, para o caso da filial ter faturado mais face ao ano anterior, ou negativas, caso a filial tenha faturado menos. Neste caso em específico, e de acordo com os dados apresentados na figura, pode verificar-se que nos anos de 2011, 2012, 2013, 2016, 2019 e 2020 a filial em causa apresentou resultados de faturação inferiores ao que tinha apresentado no ano imediatamente anterior, contrariamente ao verificado nos anos de 2014, 2015, 2017 e 2018, onde foram verificados aumentos nos valores de faturação.

Para além de analisadas em valor monetário, as mesmas variações foram também analisadas em valor percentual. Convertendo estas variações para percentagem obtiveram-se os dados apresentados na tabela 4.3. Como se pode constatar, tanto o gráfico da figura 4.7 como a tabela 4.3 apresentam as variações referentes à faturação, em valor monetário ou em percentagem, apenas a partir de 2011, o que se justifica pelo facto de ser a partir desse ano que se conseguem analisar as variações face aos anos imediatamente anteriores, uma vez que o primeiro ano de que se dispõe de dados é o

[1]https://www.dadosmundiais.com/europa/portugal/inflacao.php (última consulta em 2025-05-14).

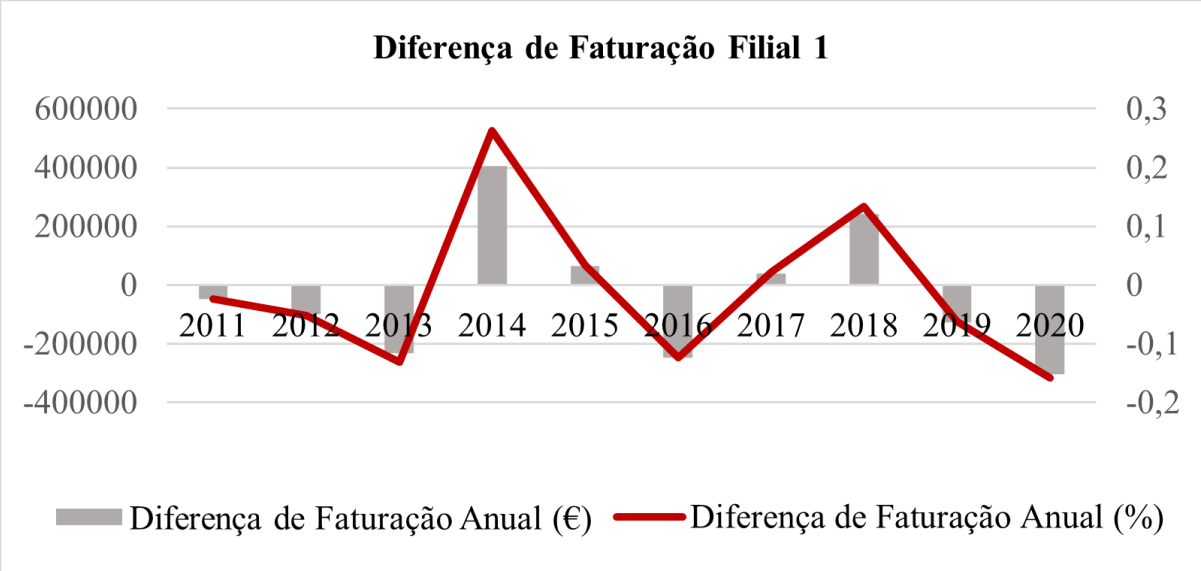


Figura 4.7: Valores da diferença de faturação anual efetiva da filial 1

ano de 2010. Depois da conversão foi então realizada a comparação com os valores da taxa de inflação de Portugal para perceber que tipo de crescimento ao nível da faturação se tinha verificado, se era ou não superior à taxa de inflação do respetivo ano. Para tal, apenas se verificaram os anos que apresentaram aumento de faturação face ao ano anterior, ou seja, neste caso de exemplo da filial 1, os anos 2014, 2015, 2017 e 2018. Assim, e de acordo com os valores apresentados na tabela 4.3, podemos verificar que para os quatro anos em que a filial demonstrou crescimento apresentou taxas de crescimento maiores que os valores das taxas de inflação verificadas no país para os mesmos períodos (sobretudo quando comparados os valores de 2014 em que a taxa de inflação registada foi negativa), sendo por isso os resultados superiores ao mínimo esperado. Nesta análise em específico destacar aqui os anos de 2014 e de 2018 porque foram aqueles que apresentaram valores consideravelmente superiores face às taxas de inflação verificadas nos respetivos anos, cerca de 95 e 13 vezes respetivamente.

Tabela 4.3: Histórico das taxas de inflação anual em Portugal e com as variações sobre as faturações anuais da filial 1 em percentagem

| Ano | Taxa de inflação em Portugal (%) | Diferenças de Faturação Anual Filial 1 (%) |
|---|---|---|
| 2011 | 3,65 | -2,44 |
| 2012 | 2,77 | -5,23 |
| 2013 | 0,27 | -13,10 |
| 2014 | -0,28 | 26,31 |
| 2015 | 0,49 | 3,33 |
| 2016 | 0,61 | -12,29 |
| 2017 | 1,37 | 2,18 |
| 2018 | 0,99 | 13,39 |
| 2019 | 0,34 | -6,27 |
| 2020 | -0,01 | -15,79 |

Após a realização desta análise para a filial 1, foram seguidos os procedimentos para as restantes filiais da empresa e retiradas as respetivas conclusões sobre os dados. Para a realização destas análises foram também elaborados gráficos para todas as filiais, traçando-se posteriormente as respetivas linhas de tendência para perceber as suas evoluções ao longo do tempo. Através dos dados obtidos pôde-se constatar que a filial que apresentou maior margem de crescimento foi a filial 3, com um total de sete anos a apresentar resultados superiores face aos anteriores, como se pode verificar pelo gráfico da figura 4.8 . As filiais 4 e 8 apresentaram ambas seis anos com crescimento de faturação face aos anos anteriores, as filiais 2, 6 e 7 com 5 anos de variações positivas e por último a filial 1 com apenas 4.

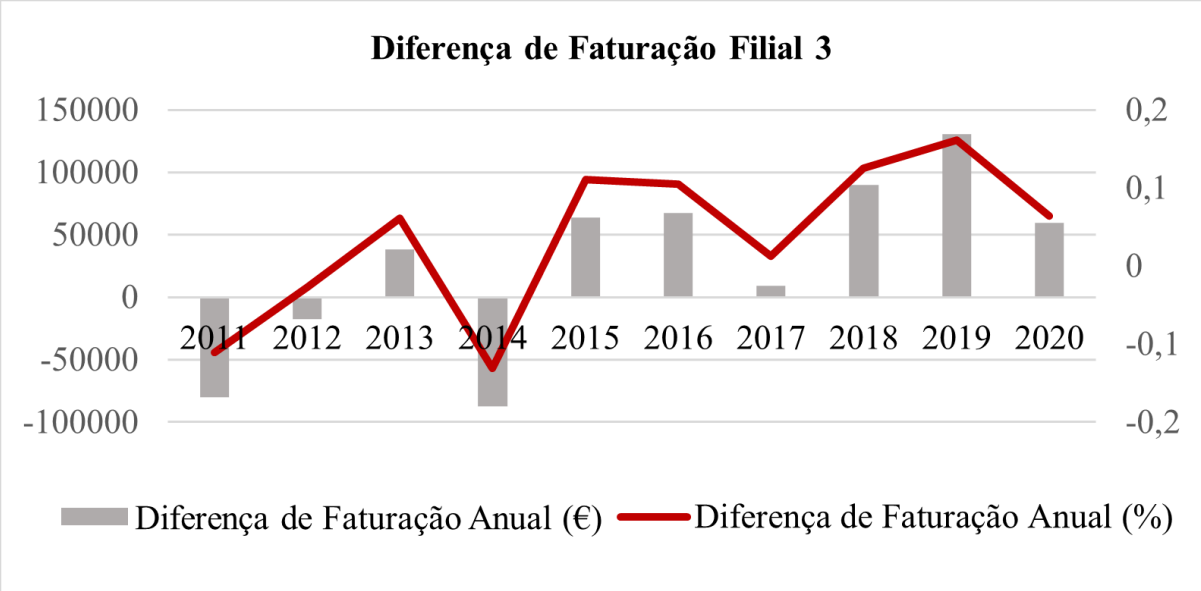


Figura 4.8: Valores da diferença de faturação efetiva anual da filial 3

Nesta fase, a análise gráfica sobre os dados de cada filial e das respetivas linhas de tendência foi realizada recorrendo ao Excel, uma ferramenta bastante conhecida e poderosa no que respeita a análise de dados. Através das linhas de tendência consegue-se verificar, tal como nome indica, qual é a tendência dos valores ao longo do tempo de acordo com os dados de que se dispõe e até fazer uma potencial previsão de valores futuros durante um determinado período pré definido pelo utilizador. De acordo com os dados disponíveis, o programa permite realizar diversos modelos de linhas de tendência, adotando por omissão a linha de tendência linear, a utilizada também no estudo descrito. Esta linha de tendência assenta na regressão linear simples, uma técnica estatística de análise de dados que, através dos dados de que se dispõe, modela matematicamente as variáveis dependentes e independentes como uma equação linear, possibilitando a previsão de valores futuros sobre uma determinada variável[2]. Analisadas as linhas de tendência referentes às variações de faturação de cada uma das

[2]https://aws.amazon.com/pt/what-is/linear-regression/ (última consulta em 2025-05-18).

filiais, tanto em valor monetário como percentual, pôde verificar-se que a filial 8, apesar dos seus 6 anos de resultados superiores face ao ano anterior, apresentou uma linha de tendência negativa, bem como as filiais 1 e 2, com 4 e 5 anos de variações positivas de faturação, respetivamente. As restantes filiais, isto é, a 3, 4, 6 e 7 apresentaram todas linhas de tendência positivas face às variações de faturação. De todas estas filiais com linhas de tendência positivas, a que apresentou uma linha de tendência com declive mais acentuado foi a filial 3, muito por causa dos seus 7 anos a apresentar resultados superiores face aos anteriores, 6 deles consecutivos (Figura 4.9).

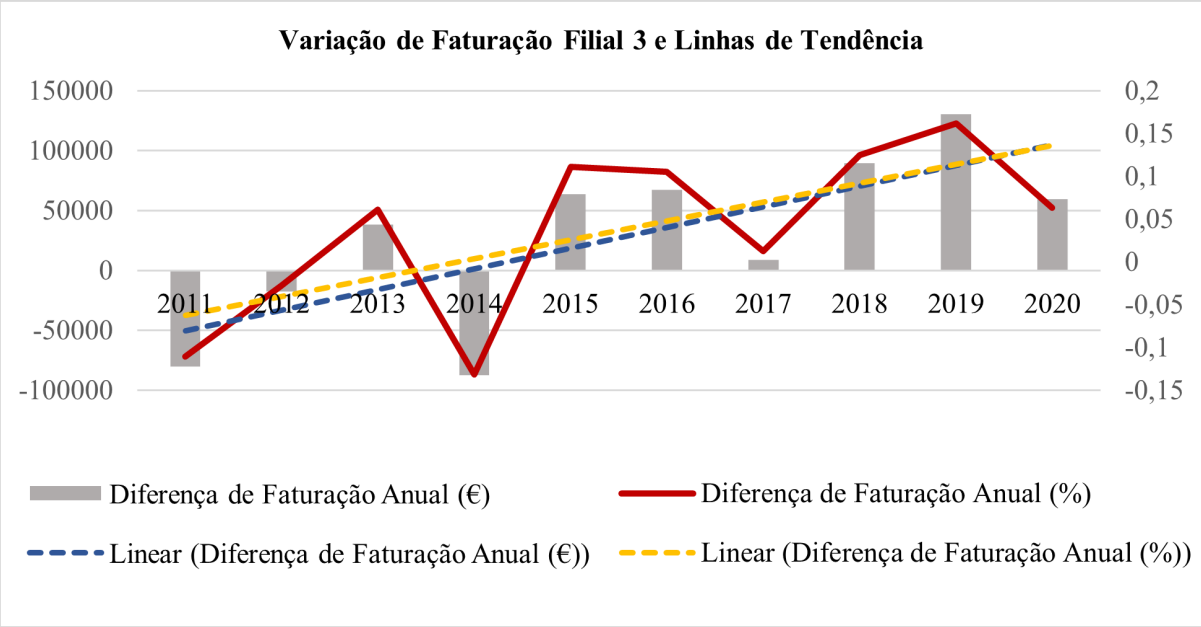


Figura 4.9: Valores da variação de faturação efetiva anual da filial 3 em valor monetário e percentual com respetivas linhas de tendência

Para analisar ao detalhe todas as variações e fazer a analogia para todas as filiais do respetivo aumento percentual perante a taxa de inflação anual para os mesmos anos, foi construida a tabela 4.4. Nesta tabela estão presentes os valores referentes ao histórico da taxa de inflação anual em Portugal para o período em estudo, entre 2010 e 2020, as variações percentuais das faturações de cada filial para cada um dos anos e para a empresa no geral, sejam elas positivas ou negativas. Como já explicado acima, apenas aparecem dados referentes às variações a partir de 2011 uma vez que para o período em estudo, apenas em 2011 se consegue verificar a primeira variação de faturação face ao ano anterior.

Analisada a tabela 4.4 ao detalhe, e individualizando aqui a mesma por filiais, pode verificar-se que de todas as filiais em estudo e para todos os anos, uma grande parte das variações positivas de faturação foi superior às taxas de inflação referentes ao mesmo ano, algumas até consideravelmente. No entanto, excetuando aqui o afirmado anteriormente, existem dois anos e duas filiais diferentes em que os crescimentos verificados face aos anos anteriores não ultrapassou o valor das taxas de inflação para os respeti-

Tabela 4.4: Histórico das taxas de inflação anual em Portugal e variações das faturações anuais das filiais

| Ano | Inflação (%) | Variação da Faturação Anual (%) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Filial 1 | Filial 2 | Filial 3 | Filial 4 | Filial 6 | Filial 7 | Filial 8 | Geral |
| 2011 | 3,65 | -2,44 | 11,60 | -11,04 | -6,12 | -5,15 | 38,17 | -6,04 | 0,56 |
| 2012 | 2,77 | -5,23 | 12,09 | -2,76 | 4,12 | -13,08 | -11,41 | 19,19 | 1,95 |
| 2013 | 0,27 | -13,10 | -8,03 | 6,08 | -2,72 | -2,42 | -20,25 | 0,53 | -7,17 |
| 2014 | -0,28 | 26,31 | -6,49 | -13,15 | 6,58 | 0,61 | -64,98 | 16,85 | 4,05 |
| 2015 | 0,49 | 3,33 | 14,00 | 11,07 | 4,44 | 23,12 | 27,91 | 20,64 | 9,45 |
| 2016 | 0,61 | -12,29 | -5,08 | 10,51 | -15,91 | 0,16 | -23,10 | -6,18 | -7,28 |
| 2017 | 1,37 | 2,18 | 9,27 | 1,27 | 13,71 | 9,59 | 5,05 | -0,31 | 5,69 |
| 2018 | 0,99 | 13,39 | -2,45 | 12,49 | 4,97 | -15,99 | 32,51 | -1,14 | 5,13 |
| 2019 | 0,34 | -6,27 | 13,29 | 16,15 | 20,94 | 11,41 | 6,44 | 5,10 | 7,31 |
| 2020 | -0,01 | -15,79 | -10,06 | 6,36 | -13,71 | -3,26 | -13,78 | -15,10 | -10,02 |
| **Desvio Padrão** | | 0,12 | 0,09 | 0,09 | 0,11 | 0,11 | 0,30 | 0,11 | 0,06 |

vos anos, sendo esses casos a filial 3 no ano de 2017 e a filial 6 no ano de 2016. Olhando depois para a coluna geral, onde estão expressas as variações de faturação por ano para a empresa no global, podem verificar-se variações positivas durante sete anos, o que significa que nesses mesmos sete anos a empresa apresentou melhores resultados do que no ano imediatamente anterior, sobrepondo-se estas variações na sua maioria aos valores da taxa de inflação para os mesmos anos. Contudo, tal como o verificado para a filial 3 no ano de 2017 e a filial 6 no ano de 2016 na análise anterior, também nos dois primeiros anos em que a empresa apresentou variações positivas, 2011 e 2012, os valores apresentados não superaram os valores das taxas de inflação referentes ao mesmo ano. De acordo com o Conselho de Finanças Públicas, a taxa de inflação diz respeito à "variação percentual do Índice de Preços no Consumidor de um determinado período relativamente ao valor registado num período anterio"[3], o que significa que, apesar do crescimento apresentado, uma vez que as filiais supramencionadas e a empresa no geral não ultrapassaram o valor referente à taxa de inflação para os referidos anos, esses aumentos acabam por ser um pouco fictícios já que, na realidade, o custo de vida e dos produtos para esses anos em Portugal acabou por ser superior face aos aumentos de faturação verificados. Na análise sobre esta tabela observou-se também que determinados valores se destacam dos restantes pela sua ordem de grandeza, como se pode constatar através dos valores nela sublinhados. Precisamente porque se tratavam de valores que sobressaíam perante os demais, pensou-se que talvez pudessem ter algum erro associado, ou que existisse algum problema com os dados, confirmando-se depois que não, que correspondem mesmo a valores reais de faturação da empresa e por isso que estão corretos para observação.

Uma outra análise realizada sobre estes dados assentou na discrepância de valores, isto é, se existia uma variação muito grande entre as diferenças de faturação de uns anos

[3]https://www.cfp.pt/pt/glossario/taxa-de-inflacao (última consulta em 2025-05-18)

para os outros, tanto ao nível de filiais por si só como da empresa no geral. Esta análise foi realizada tendo por base o valor do desvio padrão, que foi calculado para cada uma das filiais e para a empresa globalmente no sentido de tentar perceber o nível de homogeneidade das respetivas variações de faturação. Para esta medida, considerada uma medida de avaliação da dispersão dos dados, quanto menor o valor, maior será a homogeneidade dos dados e, por isso, menor as variações entre eles; quanto maior for o valor, maior será a sua distribuição, ou seja, mais heterogéneos serão. Depois de calculados os desvios padrão verificou-se que uma das filiais apresentou um valor muito diferente das restantes para esta variável, a filial 7, com um desvio padrão de 0,3, sendo a filial com maior dispersão entre os respetivos dados, o que já seria de esperar face ao nível de variações apresentadas (comprovada pelos valores sublinhados na tabela 4.4). As restantes filiais, por sua vez, apresentaram valores compreendidos entre os 0,09 e os 0,12, o que significa que as respetivas variações de faturação foram mais homogéneas ao longo do tempo. Analisando posteriormente os valor do desvio padrão da empresa no seu todo pode-se constatar que o valor mais baixo para esta medida se encontrava aí, com o valor de 0,06. Este facto pode ser justificado porque foi também aqui que se verificaram as menores variações de faturação ao longo dos anos, isto quando comparadas com as das filiais individualmente, não apresentando valores excecionalmente grandes, tanto para os resultados positivos como para os negativos. Assim, podemos concluir que mesmo quando determinadas filiais apresentavam um grande aumento ou redução no nível de faturação para um determinado ano, esses valores acabavam depois por ser compensados pelos valores de faturação de outras filiais, levando a empresa no seu global a apresentar estes resultados.

Toda a análise anterior foi realizada tendo por base o Excel como ferramenta de análise de dados. No entanto, e por forma a preparar os dados já para o desenvolvimento do modelo preditivo, procurou-se também fazer uma análise recorrendo a uma outra ferramenta tão bem conhecida na área da programação e uma das mais utilizadas no mundo da exploração e análise de dados, o Python.

Para trabalhar esta ferramenta foi utilizado o ambiente de desenvolvimento Spyder, já que, segundo uma pesquisa realizada sobre ambientes de desenvolvimento Python, este é um dos que melhor permite trabalhar dados graficamente, recaindo então a opção por este. Numa primeira fase começou-se por carregar os dados do projeto já minimamente filtrados, ou seja, o *dataset* explorado já continha os dados referentes à faturação efetiva da empresa para o período em estudo, entre 2010 e 2020, inclusive. Para além desta seleção, dentro destes dados foram selecionados apenas aqueles que eram referentes às sete filiais atualmente a laborar, excluindo uma outra que entretanto encerrou atividade e sobre a qual se dispunha de dados até ao primeiro trimestre do ano de 2011. Depois de carregar a base de dados, optou-se por realizar uma análise mais detalhada dentro do Python sobre uma das filiais apenas, designada a filial piloto, po-

dendo depois adaptar-se o código desenvolvido para as restantes filiais e no final para a empresa no seu todo. Neste caso, a filial piloto selecionada para o efeito foi a filial 3, uma vez que, depois de realizada a análise em Excel, foi aquela que apresentou maior margem de crescimento ao longo dos anos em estudo. Após a realização de algumas representações gráficas referentes aos dados desta filial puderam retirar-se mais alguns *insights* interessantes para esta fase do projeto, não apenas para esta filial em específico mas sobre a empresa no geral. Através do gráfico da figura 4.10 verificou-se que para o máximo de 31 dias de faturação possíveis num mês, havia uma clara evidência de maior volume de faturação nos dias 15 e 30 ou 31. Este facto é justificado porque alguma da faturação é realiza apenas nestes dias, muito por causa das guias de remessa. Para efeitos de faturação, as guias de remessa passadas em cada mês apenas são faturadas no meio ou no final desse mesmo mês, ou seja, as guias de remessa passadas entre os dias 1 e 15 são faturadas no último dia útil antes do dia 15, enquanto que as que são passadas entre os dias 15 e 30, ou 31 de cada mês apenas são faturadas no último dia útil, excetuando aqui o mês de Fevereiro, que pode ser nos dias 28 ou 29 consoante o ano. Outro aspeto importante que se retirou da análise aos gráficos obtidos nesta fase do projeto está relacionado com a faturação diária ao longo dos onze anos em estudo, entre 2010 e 2020. No gráfico da figura 4.11 pode verificar-se o volume de faturação diária da filial 3 para os onze anos em estudo. Observando o comportamento dos valores na figura, pôde-se constatar que desde o início de 2010 até julho de 2013 os valores mínimos de faturação eram mais baixos do que a partir dessa data até ao final de 2020. Através do gráfico pode verificar-se que há um padrão até junho de 2013 e outro a partir daí, que é manifestamente diferente. Este padrão deve-se sobretudo ao facto de até ao início de Julho de 2013 a empresa trabalhar bastante a sua faturação sob a forma de guias de remessa, o que significava que havia volumes de faturação muito fortes a meio e no final de cada mês, contrariamente ao verificado normalmente para os restantes dias dos meses. Assim, acumulava-se grande parte da faturação que deveria ser diária para esses dias e, por isso, existem valores tão díspares no meio e no final de cada mês quando comparados com os restantes. Outra análise que se tentou fazer foi baseada nas quantidades de transações realizadas por dia, sendo vendas ou notas de crédito, com o intuito de se perceber quais os dias com maior movimento, quais os dias mais calmos dos meses, e se eventualmente existiria algum padrão ou sazonalidade também aqui. Depois de verificar o gráfico da figura 4.12 comprovou-se o analisado anteriormente para os volumes de faturação diários, i.e., que até ao final de junho de 2013 a empresa tinha picos muito fortes de transações para os dias de meio e fecho dos meses, contrariamente ao verificado para os restantes dias, mantendo-se assim o padrão verificado no gráfico 4.11. A partir do mês de Julho de 2013 até ao último mês em estudo, dezembro de 2020, verificou-se uma ligeira homogeneidade dos dados, ou seja, passaram a ser realizadas mais faturações diárias, reduzindo assim o sistema de faturação que mais vigorava até então, as guias de remessa. Por este mo-

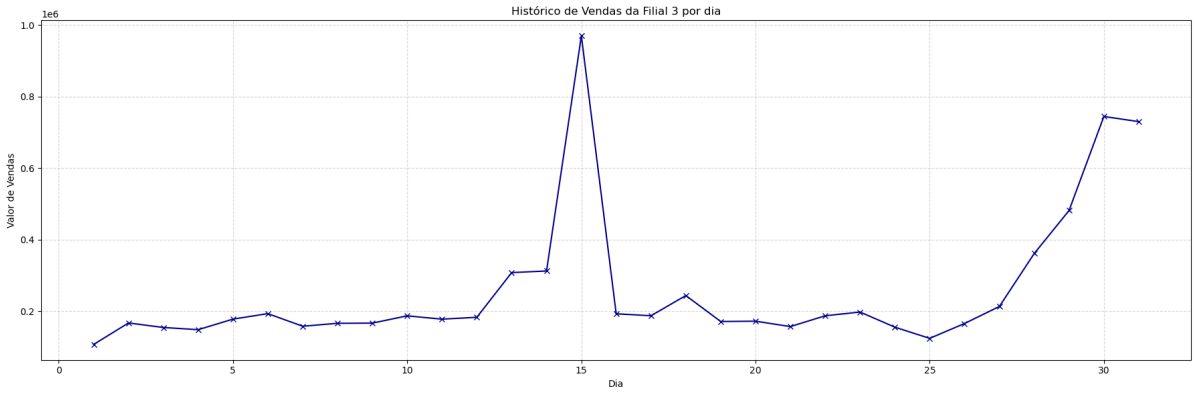


Figura 4.10: Valores do histórico de faturação por dia da filial 3

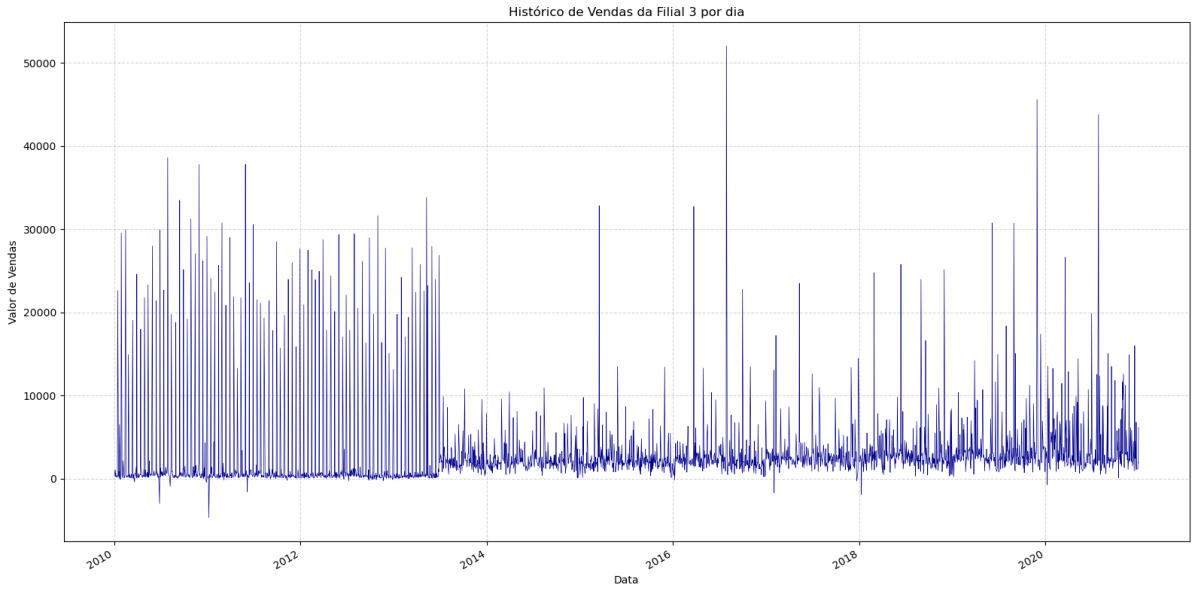


Figura 4.11: Valores do histórico de faturação por dia da filial 3 de 2010 a 2020

tivo, e para tentar construir um modelo preditivo o mais fidedigno possível, decidiu-se que para a sua construção apenas se iriam considerar os dados desde julho de 2013 até ao final de 2020, porque apesar de serem os mais recentes, eram também aqueles que apresentavam um padrão durante a maior parte dos anos.

## 4.2 Construção do modelo

Efetuada a análise sobre o *dataset* disponível, fundamental para interpretar o tipo de dados e a sua qualidade, deu-se início ao processo de definição do modelo de previsão de vendas que fosse adequado ao projeto.

A empresa em estudo realizava o processo de previsão de consumo recorrendo a métodos tradicionais, modelos matemáticos e estatísticos, sendo as previsões realizadas tendo por base o histórico de consumo dos últimos dois anos para prever um mês.

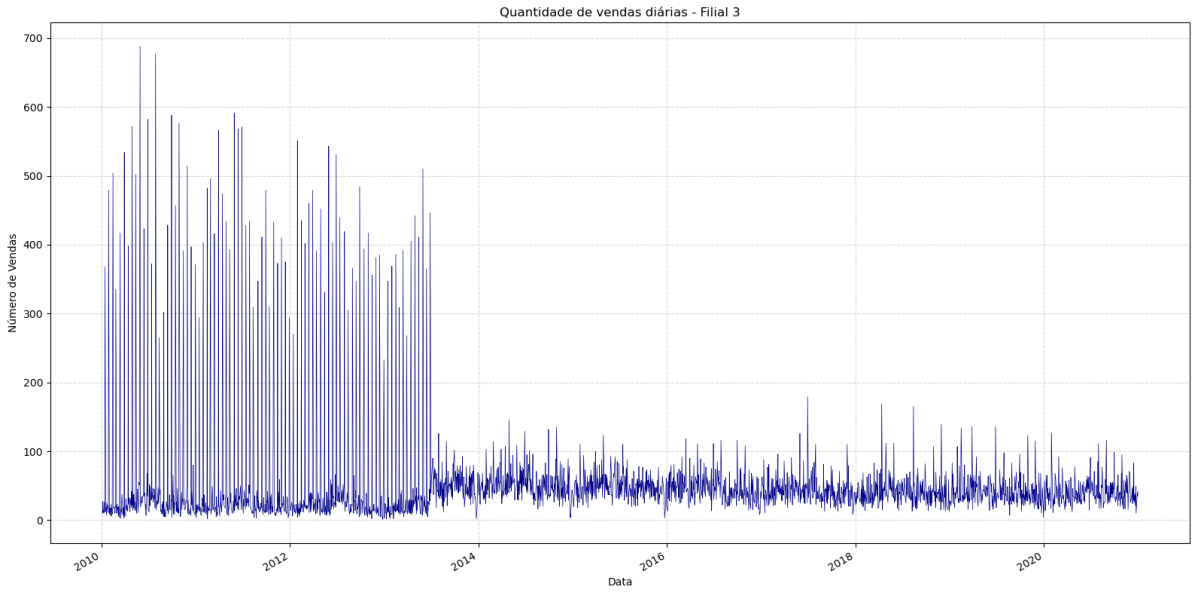


Figura 4.12: Quantidade de transações realizadas por dia da filial 3 de 2010 a 2020

O objetivo principal deste projeto passava por desenvolver uma ferramenta de previsão de vendas assente em métodos mais recentes, recorrendo ao *Machine Learning*, nomeadamente modelos de redes neuronais, que pudessem realizar previsões de forma mais dinâmica, interativa e com a maior precisão possível dentro da área de atuação da empresa e respetivo comportamento de vendas ao longo dos anos, complementando assim as previsões realizadas até então.

### 4.2.1 LSTM - Long Short-Term Memory

Para ir ao encontro do inicialmente proposto, e de acordo com os dados disponíveis, equacionou-se o modelo LSTM (*Long Short-Term Memory*), denominado em português de Memória Longa de Curto Prazo, para suportar o desenvolvimento do programa. A escolha recaiu sobre este modelo por ser adequado no que respeita à análise de séries temporais, com alta capacidade de auxiliar em dependências temporais e extremamente útil no que confere à identificação de padrões, nomeadamente fatores sazonais ou tendências.

O *dataset* disponível continha os dados de faturação da empresa em estudo entre 2010 e 2020, inclusive, e era pretendida a realização de previsões de vendas através desse histórico de valores. Recorrendo à linguagem de programação Python, começou a desenvolver-se o modelo, inicialmente apenas para uma filial da empresa, a três, considerada nesta fase a filial piloto, adaptando-se posteriormente para a realização de previsões referentes a cada uma das restantes filiais de forma individualizada e, finalmente, para a empresa no seu todo.

De acordo com a análise aos dados efetuada anteriormente e no seguimento do ex-

plicado, foi programada uma primeira versão do modelo de previsão de vendas recorrendo apenas ao histórico dos dados de faturação compreendidos entre julho de 2013 e dezembro de 2020. Para o desenvolvimento deste modelo projetou-se um programa onde se utilizavam os valores de vendas de cinco dias para prever um. Ou seja, o vetor de entrada para este primeiro programa era constituído por sequências de 5 valores normalizados consecutivos referentes apenas à série de vendas (equação 4.1). Relativamente aos dados foi definido que 80% do *dataset* selecionado seria para treino e 20% para teste. Ao nível da arquitetura, foram utilizadas duas camadas de rede, a primeira com 100 neurónios e a segunda com 50, recorrendo à função de ativação "relu"(unidade linear retificada) para ambas. Para treinar o modelo definiu-se inicialmente o otimizador "Adam" com a taxa de aprendizagem padrão utilizada por defeito pelo Python para este modelo, ou seja, 0.001. Ainda no que ao treino diz respeito, foram definidas cem épocas para treinar o modelo, com um ajuste dos pesos a cada 32 amostras. Dentro deste programa foram também definidas algumas métricas de avaliação para analisar os resultados obtidos e, consoante esses valores, proceder-se à validação do modelo ou a uma análise sobre futuras melhorias a implementar. As métricas inicialmente utilizadas foram a MSE (*Mean Squared Error*), a RMSE (*Root Mean Squared Error*), MAE (*Mean Absolut Error*) e MAPE (*Mean Absolute Percentage Error*). Compilada esta primeira versão, verificados os resultados destas métricas e os graficamente obtidos (figura 4.13) observou-se que o modelo desenvolvido estava muito aquém do pretendido, iniciando-se aí a procura por possíveis melhorias.

$$I_n = (V_{n-1}, V_{n-2}, V_{n-3}, V_{n-4}, V_{n-5}) \tag{4.1}$$

onde:

- $I_n$ - vetor de entrada para prever vendas do dia $n$;
- $V_{n-1}$ - valor de vendas do dia anterior a $n$;
- $V_{n-5}$ - valor de vendas referente ao quinto dia anterior a $n$.

Neste primeiro protótipo do modelo apenas foram considerados os valores disponíveis das vendas. Porém, talvez fosse possível melhorar o modelo através da engenharia de *features*, começando nesta fase a equacionar-se a consideração de fatores externos como as datas dos dados de vendas, por exemplo. Após essa avaliação foi ponderada a adição de *features* temporais, nomeadamente o "Dia da Semana", "Dia do Mês", "Mês"e "Dia do Ano"com o principal objetivo de alargar o vetor de entrada e assim tentar incrementar valor ao programa. Desta forma, poderiam ser reconhecidos determinados padrões e comportamentos nas vendas que até então pudessem não estar a ser identificados, determinadas sazonalidades ou até dependências e tendências que não estivessem a ser consideradas pelo programa no treino de dados e que, indubitavelmente, iriam ter

impacto nos valores previstos. Foram então adicionadas as *features* temporais ao programa, recorrendo à codificação de inteiros, com os dias da semana codificados de 0 a 4, onde os zero seriam referentes às segundas-feiras e os quatro às sextas-feiras (já que a empresa labora apenas em dias úteis), os dias do mês codificados de 1 a 31, sendo que cada um corresponde ao respetivo dia do mês, os meses codificados de 1 a 12 referentes ao respetivo mês do ano a que os dados dizem respeito e, ainda, o dia do ano codificado de 1 a 365 ou 366 para anos bissextos, correspondendo cada um ao número do dia do ano a que os dados em análise são referentes. Para este caso, o vetor de entrada iria ser maior do que o inicialmente equacionado, porque além dos valores normalizados das vendas, iria também conter os valores normalizados com informação sobre o dia de semana, o dia do mês, o mês e o dia do ano de cada uma das vendas realizadas (equação 4.2). Depois de efetuada esta alteração, fez-se a compilação do programa e foram novamente analisados os resultados, tanto gráficos como os obtidos para as métricas de avaliação, tendo-se constatado que estes se encontravam ainda consideravelmente distantes do pretendido.

$$I2_n = (V_{n-i}, DS_{n-i}, DM_{n-i}, M_{n-i}, DA_{n-i}) \tag{4.2}$$

para $i \in [1, ..5]$; $DS \in [0, ..4]$; $DM \in [1, ..31]$; $M \in [1, ..12]$; $DA \in [1, ..366]$

onde:

- $I2_n$ - vetor de entrada para prever vendas do dia $n$;
- $V_{n-i}$ - valor de vendas referente a um dos cinco dias anteriores a $n$;
- $DS_{n-i}$ - informação sobre o dia da semana em que se concretizou a venda $Vn - i$;
- $DM_{n-i}$ - informação sobre o dia do mês em que se concretizou a venda $Vn - i$;
- $M_{n-i}$ - informação sobre o mês em que se concretizou a venda $Vn - i$;
- $DA_{n-i}$ - informação sobre o dia do ano em que se concretizou a venda $Vn - i$.

Para além destas *features* inicialmente equacionadas, foram posteriormente acrescentados o "Trimestre", "Ano"e "Semana do ano", procurando com isto proporcionar ainda mais informação de entrada para o modelo e acrescentar ainda mais detalhe, contudo, depois de se voltar a compilar o programa, pôde verificar-se que este incremento de informação não se reverteu numa melhoria significativa nos resultados alcançados, voltando assim a utilizar-se as *features* temporais inicialmente definidas. Observados estes resultados, continuou a ponderação sobre a tentativa de melhorar o programa, surgindo aí uma alteração na codificação da engenharia de *features* que poderia auxiliar bastante nesse sentido.

A codificação de inteiros para as *features* temporais tem as suas limitações uma vez que

codifica de igual forma todas as segundas, terças, e por aí em diante até às sextas-feiras para a variável "Dia da semana", o que poderia dificultar a identificação de padrões referentes a determinados dias em específico, já que todas as segundas são 0 e todas as sextas-feiras são 4. Para além disso, este tipo de codificação podia levar a falsas relações ordinais, já que podia interpretar que as sextas-feiras, que têm o número 4 podiam ser mais importantes do que as segundas-feiras que têm o número 0, ou do que as terças-feiras que têm o número 1, o que não é verdade. Assim, o modelo podia gerar previsões erradas uma vez que podia não conseguir compreender que os valores presentes nessa variável diziam respeito a uma categoria, o dia da semana, e não propriamente a quantidade. O mesmo acontecia para as restantes categorias, "Dia do mês", "Mês" e "Dia do Ano", fazendo-se aqui a mesma analogia que a explicada para o dia de semana, mas agora referente a cada uma destas categorias em específico.

Uma outra tentativa de melhoria surgiu depois desta análise, ou seja, fazer a codificação das *features* temporais por *One-Hot Encoding* (OHE) e não através de codificação de inteiros. O OHE é um tipo de codificação mais rigoroso, que por um lado faz crescer consideravelmente a base de dados disponível, mas por outro é mais eficaz no que toca à codificação das *features* temporais. Este tipo de codificação passa cada categoria para um código binário (0 e 1), ou seja, cada *feature* temporal tem o seu código binário único associado, evitando assim a interpretação de falsas relações ordinais, como acontecia com a codificação de inteiros. Desta forma, era esperado que o modelo LSTM melhorasse a sua aprendizagem e, consequentemente, as suas previsões, já que interpretaria valores contínuos e não números inteiros como categorias. Nesta altura surgiram duas possibilidades, sendo ambas testadas. Para este tipo de codificação foram realizados dois códigos, um recorrendo à biblioteca Sklearn e outro recorrendo à biblioteca Pandas do Python, sendo posteriormente efetuadas as respetivas análises aos resultados obtidos. Com esta modificação, o vetor de entrada era muito semelhante ao da equação 4.2 mas iria ser maior porque iria transformar em código binário cada uma das *features* temporais, ou seja, o dia da semana (codificado em binário, com 7 colunas), o dia do mês(codificado em binário com 31 colunas), o mês (codificado em binário com 12 colunas) e dia do ano (codificado em binário com 366 colunas).

Alterada a codificação para OHE, compilaram-se novamente os programas e, se graficamente foram obtidos melhores resultados (figura 4.14), por outro lado, verificadas as métricas de avaliação, os resultados em nada foram ao encontro aos esperados. Como se pode verificar através da tabela 4.5, ocorreu uma ligeira melhoria recorrendo à biblioteca Pandas, o que não sucedeu quando analisados os valores referentes à biblioteca Sklearn, que pioraram.

Na tabela 4.5 estão representados os resultados obtidos nas métricas de avaliação para os programas realizados até então. A linha "Vendas" é referente ao modelo LSTM onde foram realizadas as previsões de vendas tendo em consideração apenas o histórico dos

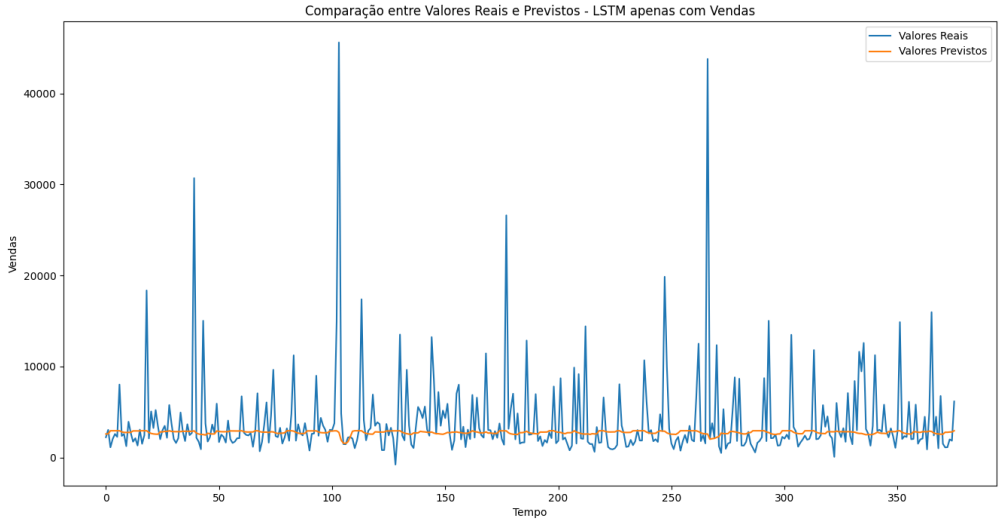


Figura 4.13: Comparação entre valores reais e previstos para filial 3 com modelo LSTM usando apenas dados de vendas

valores das vendas. A linha "V + T (Cod. Int)" diz respeito ao programa onde foram consideradas não só as vendas, mas também as *features* temporais, recorrendo aqui à codificação de inteiros para as categorias. As linhas "V + T(OHE Sklearn)" e "(V + T(OHE Pandas)", por sua vez, são referentes aos programas onde se consideraram os valores do histórico das vendas e as *features* temporais codificadas por OHE. No entanto, como explicado acima, cada uma delas recorreu a uma biblioteca do Python diferente.

Tabela 4.5: Métricas de avaliação do modelo LSTM - Filial 3

| **Modelos LSTM** | **Métricas de avaliação** | | | |
|---|---|---|---|---|
| | **MSE (€)** | **RMSE (€)** | **MAE (€)** | **MAPE (%)** |
| **Vendas** | 23997830,90 | 4898,76 | 2281,05 | 66,41 |
| **V + T (Cod. Int)** | 24663264,26 | 4966,21 | 2305,99 | 58,73 |
| **V + T (OHE Sklearn)** | 24392993,27 | 4938,93 | 2332,36 | 59,25 |
| **V + T (OHE Pandas)** | 24511796,99 | 4950,94 | 2355,98 | 58,35 |

Para além da engenharia de *features* foi ainda equacionada outra forma de adicionar valor ao modelo, uma vez que os resultados estavam longe de corresponder ao inicialmente pretendido.

Depois de realizadas as etapas supra mencionadas, foi efetuada uma série de revisões e diversos ajustes sobre os hiperparâmetros que compunham o modelo. Dentro destes foram inúmeras as alterações e testes realizados com o intuito de melhorar o modelo de previsão, nomeadamente o ajuste ao número da janela de previsão, ao número de camadas e de neurónios que compõem cada camada da rede neuronal, colocação ou não

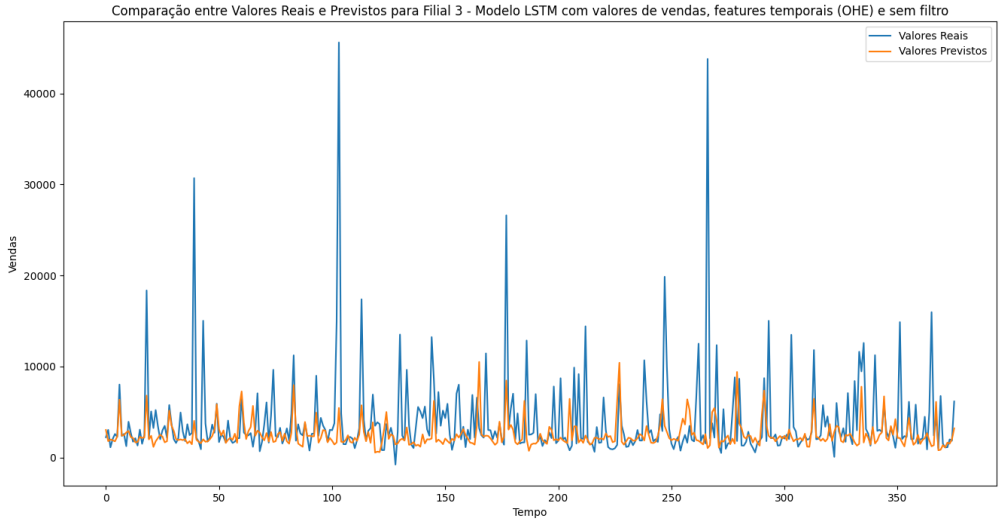


Figura 4.14: Comparação entre valores reais e previstos para filial 3 com modelo LSTM usando dados de vendas e *features* temporais com OHE

de "dropout" para tentar antecipar a possibilidade de *overfitting*, alteração da função de ativação, da taxa de aprendizagem para uma taxa mais lenta para que pudesse captar melhor os padrões e generalizar melhor, alteração das épocas de treino, do "batch size", entre outras. Contudo, mesmo com todas estas alterações e testes, a melhoria desejada não foi alcançada, refletindo-se em resultados para cada um dos programas desenvolvidos na tabela 4.5.

Analisados os valores das métricas, começou a equacionar-se se poderia ser vantajoso para o modelo aplicar um filtro nos dados que iriam ser utilizados como entrada no modelo para a realização da previsão, procurando desta forma minimizar o seu ruído, destacar tendências e, consequentemente, melhorar as previsões geradas. Inicialmente foi aplicado o filtro de média móvel simples aos dados de vendas, conseguindo desta forma suavizar as grandes flutuações já que se estava a substituir o valor de cada ponto de dados pela média do valores que lhes são mais próximos, dentro de um conjunto definido pelo programador. Depois de vários testes realizados, verificou-se que o melhor seria definir esta janela para um conjunto de cinco valores, com a ideia de corresponder aos valores de vendas de cinco dias úteis, uma semana. Programado esse filtro, voltou a compilar-se o modelo e, apesar de se conseguirem obter melhores resultados, visíveis através da figura 4.15, estes estavam ainda distantes de valores que pudessem ser considerados como satisfatórios para um modelo de previsão.

Depois de analisados os resultados obtidos para os diferentes programas do modelo de previsão de vendas LSTM realizados, pôde-se verificar que todos estão bastante distantes dos valores considerados aceitáveis para as métricas de avaliação, começando nesta fase a surgir bastantes dúvidas de que este fosse o modelo ideal para a realização

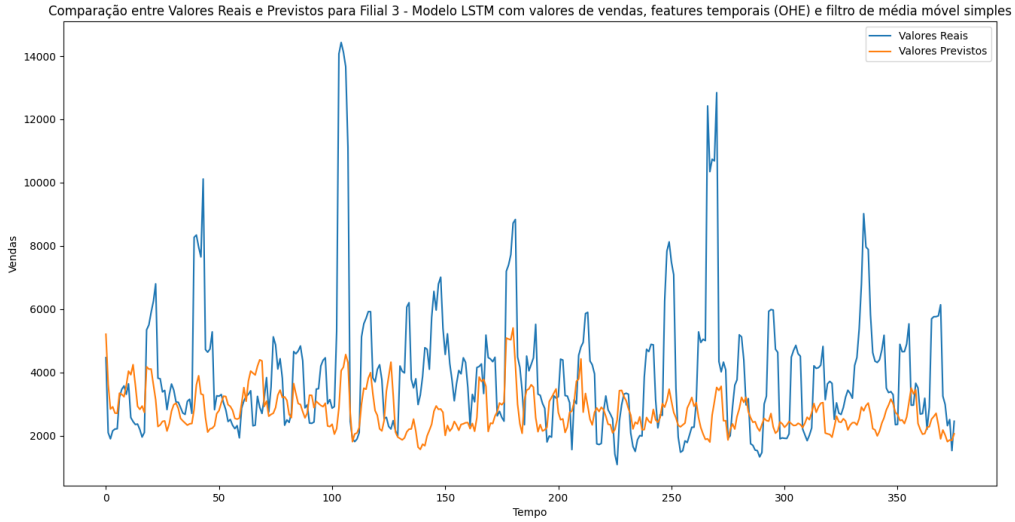


Figura 4.15: Comparação entre valores reais e previstos para filial 3 com modelo LSTM usando dados de vendas, *features* temporais com OHE e um filtro de média móvel simples

de previsão de vendas tendo em conta o *dataset* disponível. Um dos problemas poderia estar associado à dimensão da base de dados, que poderia ser demasiado pequena para este tipo de modelo. Outro problema identificado e que podia também estar na origem dos resultados insatisfatórios alcançados para as métricas de avaliação foi a forte oscilação de valores de faturação, uma vez que os dados de vendas diária variam muito de dia para dia e em determinados dias existem picos de faturação muito fortes.

### 4.2.2 GRU - Gated Recurrent Unit

Após alguma reflexão sobre os vários modelos existentes e alguma pesquisa sobre modelos que pudessem dar algumas garantias quanto ao tratamento de sequências temporais, optou-se por avançar para o modelo *Gated Recurrent Unit* (GRU). Tal como o LSTM, este modelo é também ele de redes neuronais recorrentes, adequado para séries temporais, no entanto pode ser considerado como uma versão mais simples do LSTM. Como o objetivo era identificar padrões, sazonalidades ou tendências dentro dos dados disponíveis, e constatado o facto de que o tamanho do *dataset* era uma lacuna, já que estamos a usar os dados de vendas referentes a sete anos e meio, que é pouca quantidade, tentou-se usar um modelo indicado para o estudo de séries temporais, menos complexo e que funcionasse melhor com uma quantidade de dados reduzida.

Aproveitando parte do que tinha sido a base de programação para o modelo LSTM, adaptou-se o código para trabalhar agora o modelo GRU para a previsão de vendas. No desenvolvimento do GRU foi igualmente definida uma janela de 5 dias para prever um e foram logo adicionadas as *features* temporais, codificadas com OHE com a biblioteca

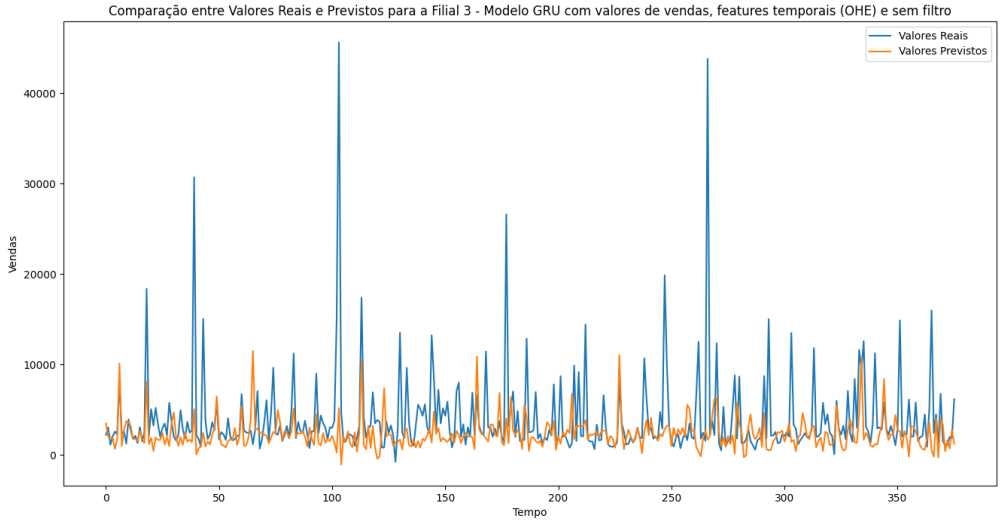


Figura 4.16: Comparação entre valores reais e previstos para filial 3 com modelo GRU usando dados de vendas e *features* temporais com OHE

Pandas, verificado no modelo anterior que tinha alcançado melhores resultados, para tentar incrementar mais valor ao programa desde o início. O vetor de entrada para este modelo era igual ao do LSTM que considerava o valor das vendas e as *features* temporais com codificação OHE. Na sua programação foi também estipulado que 80% dos dados seriam para treino e 20% para teste. Na arquitetura desta rede foram definidas duas camadas, cada uma composta por 100 e 50 neurónios respetivamente, com a função de ativação "tanh" (tangente hiperbólica) para ambas. Para aprendizagem do modelo recorreu-se ao otimizador "adam", com a sua taxa de aprendizagem padrão. Para o seu treino foram definidas 100 épocas, com um ajuste de parâmetros a cada 32 amostras (batch size=32). Depois de compilar o modelo, ao contrário do que era esperado, não houve melhorias significativas nos resultados obtidos para as métricas de avaliação da previsão - definidas as mesmas que no modelo LSTM - verificando-se precisamente o oposto. Procurou-se ajustar alguns parâmetros, realizara-se diferentes testes com diversos valores, alterou-se a função de ativação, o otimizador, tudo com o intuito de melhorar as previsões, contudo, os resultados alcançados estiveram sempre bastante distantes do pretendido, como se pode verificar graficamente através da figura 4.16.

Como no primeiro modelo tinha resultado numa melhoria considerável dos valores das métricas de avaliação, também para este modelo se tentou incrementar algum valor colocando um filtro nos seus dados de entrada, o filtro de média móvel simples. Nesta etapa foram igualmente realizados alguns testes e optou-se por escolher também o filtro de média móvel para um intervalo de cinco dias para que se pudesse enquadrar para cada valor a média de valores correspondentes aos dias úteis de uma semana de trabalho normal. Aplicado este filtro puderam-se verificar melhorias significativas no

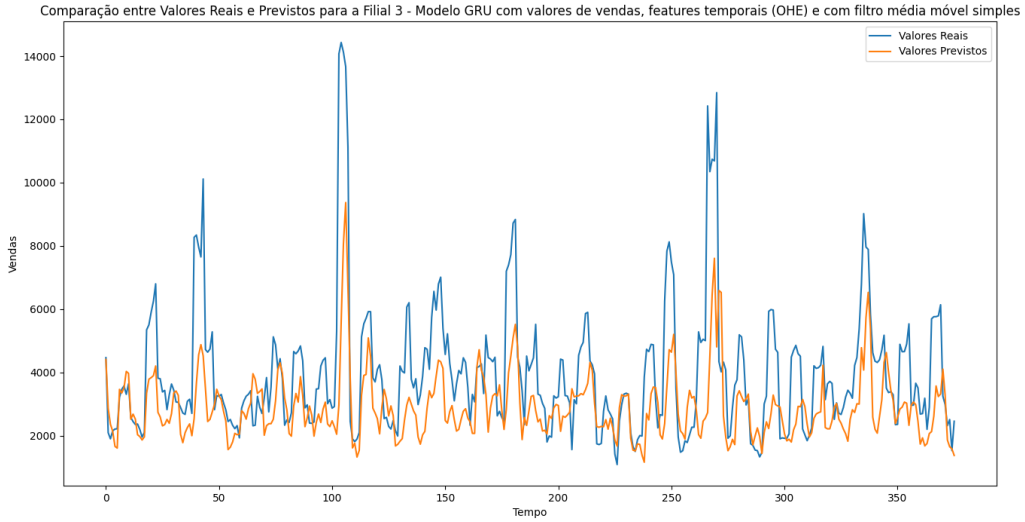


Figura 4.17: Comparação entre valores reais e previstos para filial 3 com modelo GRU usando dados de vendas, *features* temporais com OHE e um filtro de média móvel simples

que respeitava às métricas de avaliação dos modelos, bem como ao nível gráfico, já que os valores previstos conseguiam acompanhar melhor os valores reais, como se pode verificar na figura 4.17. No entanto, e apesar da melhoria verificada, os valores estavam ainda distantes dos que eram os pretendidos para um bom modelo de previsão de vendas.

Mais uma vez, começou a ponderar-se outros tipos de modelo que pudessem adequar-se ao projeto, já que os testados até agora não tinham conseguido proporcionar os resultados que eram esperados.

### 4.2.3 MLP Regressor - Multi-Layer Perceptron Regressor

Analisados novamente os vários modelos disponíveis, optou-se por um modelo diferente, um *Multi-Layer Perceptron*, neste caso de regressão, mais conhecido no mundo do *Machine Learning* como MLP Regressor. Este é um modelo de redes neuronais, mas de *feedfoward*, ainda mais simples que os usados anteriormente mas que exige *features* temporais para conseguir identificar padrões temporais e determinadas tendências. Continuando a usar a base de programação do LSTM e GRU, adaptou-se agora o código para programar o MLP Regressor na realização de previsões de vendas. De igual forma ao que tinha sido pensado para o desenvolvimento do GRU, também no MLP Regressor foram adicionadas as *features* temporais desde a base, sendo que para este caso em específico, estas eram fundamentais logo desde o início. Para este modelo foram também definidos cinco dias na janela de previsão para realizar a previsão de um, sendo o seu vetor de entrada igual ao definido para os modelos GRU e o LSTM que con-

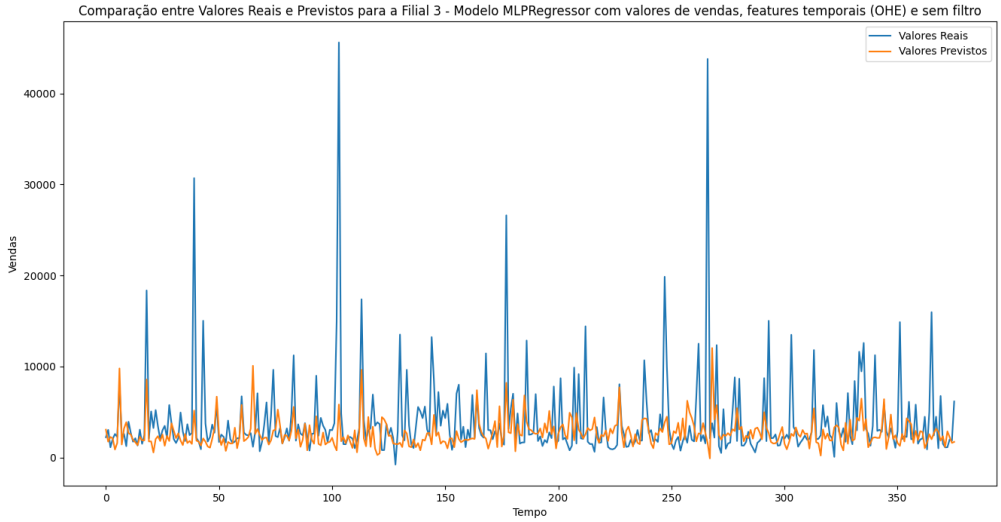


Figura 4.18: Comparação entre valores reais e previstos para filial 3 com modelo MLP Regressor usando dados de vendas e *features* temporais com OHE

siderava as vendas e as *features* temporais codificadas com OHE. Tal como nos modelos anteriores, o *dataset* foi também dividido em dois blocos, 80% para treino e 20% para teste. Ao nível da sua arquitetura, o modelo foi desenvolvido com duas camadas ocultas, a primeira composta por 800 neurónios e a segunda por 400. Aqui determinou-se também que o "randon state" iria ser 1[4], um número de 2000 interações para o treino, uma tolerância[5] de 0.01 e ainda o "verbose" como verdadeiro para mostrar as informações durante o treino. Depois de compilar o programa conseguiu-se verificar que os resultados obtidos eram os piores alcançados até então, tanto ao nível de métricas de avaliação, como se pode comprovar através da tabela 4.7, como a nível gráfico, verificado através da figura 4.18.

Para este modelo foi também aplicado o filtro de média móvel simples para o conjunto de cinco dias e mesmo depois de vários testes realizados alterando hiperparâmetros e aplicando filtro, os valores das métricas de avaliação, apesar de melhorarem consideravelmente, continuavam bastante insatisfatórios, como se pode verificar também graficamente através da figura 4.19, onde os valores previstos têm bastante dificuldade em acompanhar os valores reais.

Após a análise sobre os valores obtidos para as métricas de avaliação dos modelos LSTM, GRU e MLP Regressor, tornava-se cada vez mais preponderante encontrar um modelo com o qual se conseguissem alcançar os valores pretendidos, ou pelo menos, próximo disso.

[4]garante a reprodutibilidade dos resultados, os pesos dos neurónios iniciam da mesma forma e o comportamento do treino vai ser o mesmo

[5]tolerância mínima de melhoria - se a função de erro não melhorar entre iterações o número de vezes definida neste parâmetro, o treino pode parar antes do número máximo de interações

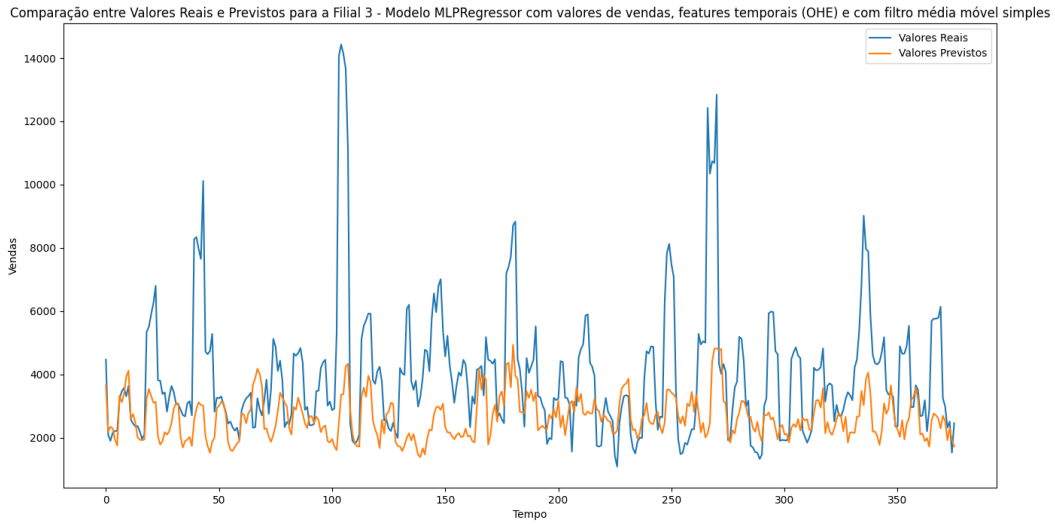


Figura 4.19: Comparação entre valores reais e previstos para filial 3 com modelo MLP Regressor usando dados de vendas, *features* temporais com OHE e um filtro de média móvel simples

### 4.2.4 XGBoost - Extreme Gradient Boosting

Depois de uma reflexão sobre os restantes modelos de previsão, o *dataset* disponível e o tipo de dados em estudo, colocou-se como possibilidade o desenvolvimento de um modelo *Extreme Gradiente Boosting* (XGBoost) para a previsão do volume de vendas. Para o efeito iria alterar-se o tipo de modelo quando comparado com os anteriores, que eram de redes neuronais. Face ao observado, ponderou-se o desenvolvimento um modelo que não fosse tão complexo, neste caso especifico, de *Machine Learning* baseado em árvores de decisão.

Para a realização deste modelo, à semelhança do que aconteceu com os restantes, conseguiu aproveitar-se, dentro do possível, alguma da programação que tinha sido desenvolvida para os modelos anteriores. Nesta fase foram igualmente criadas sequências de dados para a previsão, com uma janela de cinco dias para prever um, e colocadas logo no primeiro programa as *features* temporais codificadas com OHE. Ao contrário do que tinha sido feito inicialmente para o LSTM e depois com o GRU, que conseguem identificar sequências, reconhecer padrões e dependências temporais ao longo do tempo sem a necessidade de adicionar as *features* temporais, para o XGBoost, tal como aconteceu com o MLP Regressor, verificava-se a obrigatoriedade de as adicionar logo desde o princípio para conseguir melhorar o seu desempenho, uma vez que o modelo trata tudo como variáveis independentes e assim poderiam perder-se alguns padrões temporais para a realização das previsões. O vetor de entrada para o XGBoost era também ele igual aos anteriores, ou seja, composto por valores de vendas normalizados e por *features* temporais codificadas com OHE. No desenvolvimento deste programa determinou-

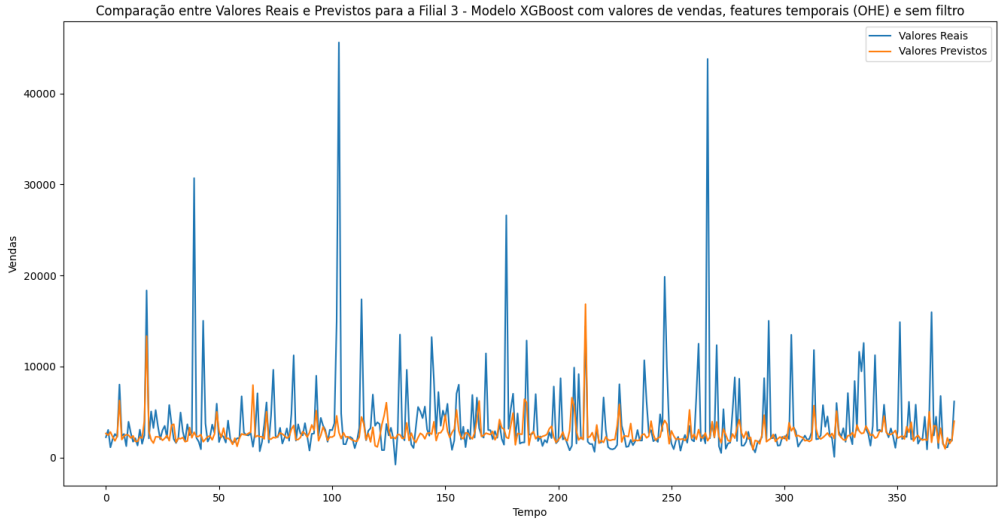


Figura 4.20: Comparação entre valores reais e previstos para filial 3 com modelo XGBoost usando dados de vendas e *features* temporais com OHE

se que 80% dos dados seriam para treino e que 20% seriam para teste. Para este modelo foi definido o problema como sendo de regressão, com minimização do erro quadrático médio entre os valores reais e os previstos e um "randon state" de 42, que garante que o modelo vai ter pesos inicias iguais, essencial para a reprodutibilidade. Os parâmetros usados foram os definidos como padrão para este modelo, i.e., 100 árvores, com uma profundidade máxima de 6, uma taxa de aprendizagem de 0.1 para cada árvore, usando a totalidade dos dados por árvore e todas as suas variáveis. Depois de compilado o modelo foram analisados os resultados obtidos para as mesmas métricas de avaliação estudadas para os modelos anteriores e pôde verificar-se que apesar de se terem obtido melhores valores, estes ainda estavam bastante distantes do pretendido, como se pode verificar através da tabela 4.7 e graficamente através da figura 4.20. Procedeu-se ao ajuste de alguns hiperparâmetros e foi aplicado o filtro de média móvel simples nos dados de entrada, tal como nos programas anteriormente desenvolvidos, mas aí, os valores obtidos para as métricas foram consideravelmente superiores, conseguindo-se assim atingir os melhores resultados alcançados até ao momento, como se pode verificar através da figura 4.21, onde os valores previstos conseguem acompanhar bastante bem os valores reais de vendas.

Aplicado o filtro de média móvel simples no XGBoost e atingindo nessa altura os melhores resultados até então, tentou-se analisar se havia outra forma de melhorar ainda mais estes resultados. Como o filtro aplicado nos dados melhorou significativamente a performance do modelo, começou a pesquisar-se se havia uma outra maneira de melhorar esse filtro, surgindo, depois de alguma pesquisa, a ideia do filtro de média móvel exponencial. Este filtro é relativamente diferente do anterior, uma vez que o filtro de

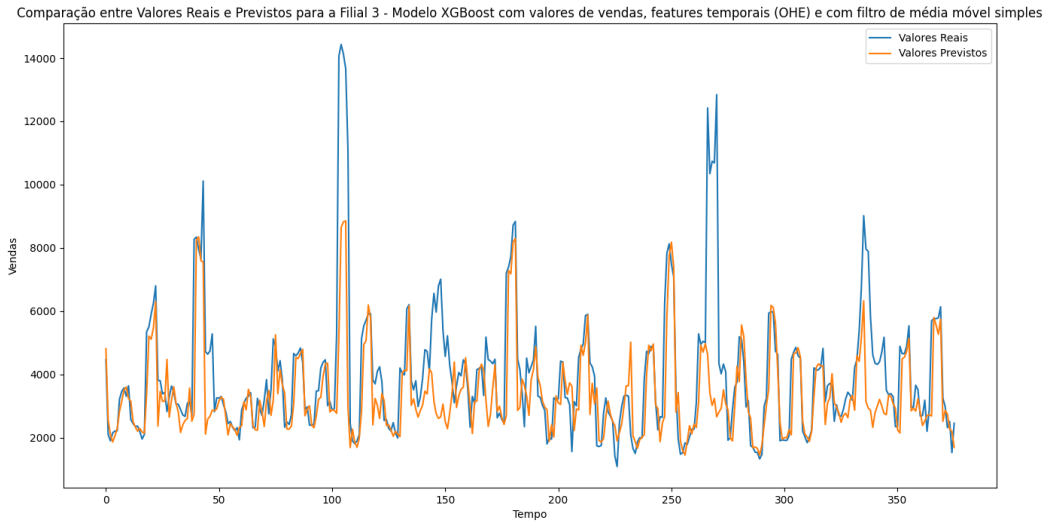


Figura 4.21: Comparação entre valores reais e previstos para filial 3 com modelo XGBoost usando dados de vendas, *features* temporais com OHE e um filtro de média móvel simples

média móvel simples atribui o mesmo peso à janela de dias definida pelo programador para calcular o valor da média que atribuirá a cada dia, enquanto que o filtro de média móvel exponencial atribui maior peso aos dias mais recentes e, por isso, consegue reagir mais rapidamente a novas tendências e à mudança de valores nos dados. Aplicou-se este filtro ao XGBoost inicialmente desenvolvido e os resultados obtidos foram ainda superiores aos que tinham sido atingidos com o filtro da média móvel simples, como se pode verificar através do gráfico da figura 4.22. Depois de se verificar este melhoria, não só através de gráfico como também nos valores das métricas de avaliação do XGBoost, comprovado pelos valores na tabela 4.7, optou-se por aplicar esse mesmo filtro nos restantes modelos desenvolvidos para verificar se, à semelhança do que aconteceu com o XGBoost, também com eles os resultados obtidos seriam superiores. Programando também o filtro da média móvel exponencial para os restantes modelos estudados pode-se constatar que, tanto ao nível das métricas como na variante gráfica, os resultados obtidos foram melhores do que os anteriormente alcançados, como se pode observar através dos gráficos das figuras 4.23, 4.24 e 4.25, contudo, ainda longe dos resultados que eram pretendidos e esperados para estes modelos.

### 4.2.5 Modelo híbrido - LSTM e XGBoost

Após a análise sobre os modelos já desenvolvidos e respetivas métricas de avaliação, começou-se a equacionar se eventualmente poderia ser benéfico combinar mais do que um modelo, criando assim um modelo híbrido. O objetivo deste modelo passaria por utilizar o que de melhor têm alguns dos modelos testados e colmatar as respetivas

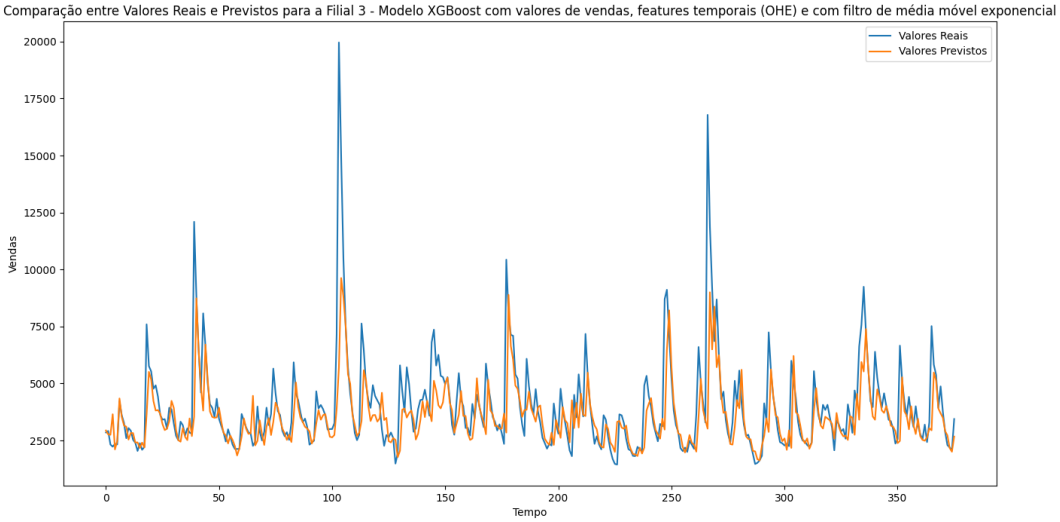


Figura 4.22: Comparação entre valores reais e previstos para filial 3 com modelo XGBoost usando dados de vendas, *features* temporais com OHE e um filtro de média móvel exponencial

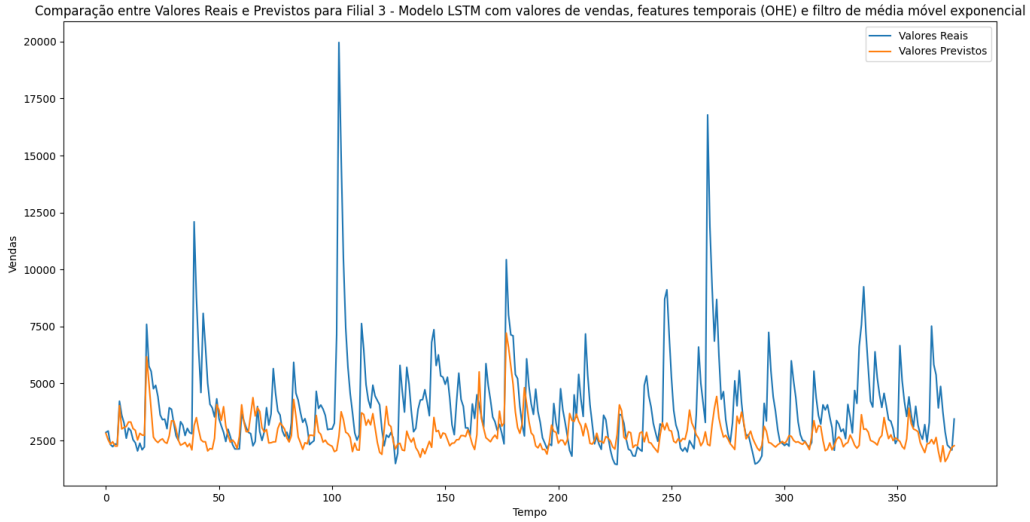


Figura 4.23: Comparação entre valores reais e previstos para filial 3 com modelo LSTM usando dados de vendas, *features* temporais com OHE e um filtro de média móvel exponencial

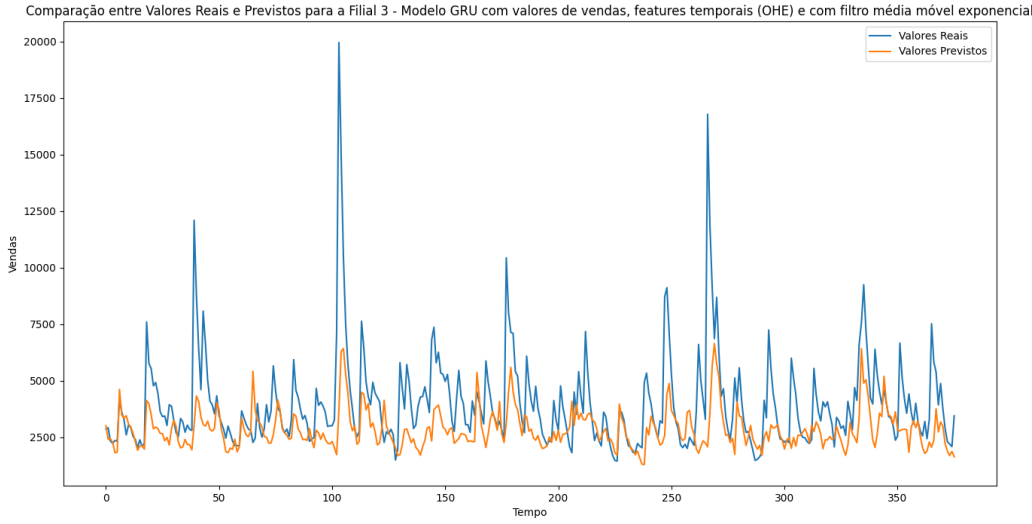


Figura 4.24: Comparação entre valores reais e previstos para filial 3 com modelo GRU usando dados de vendas, *features* temporais com OHE e um filtro de média móvel exponencial

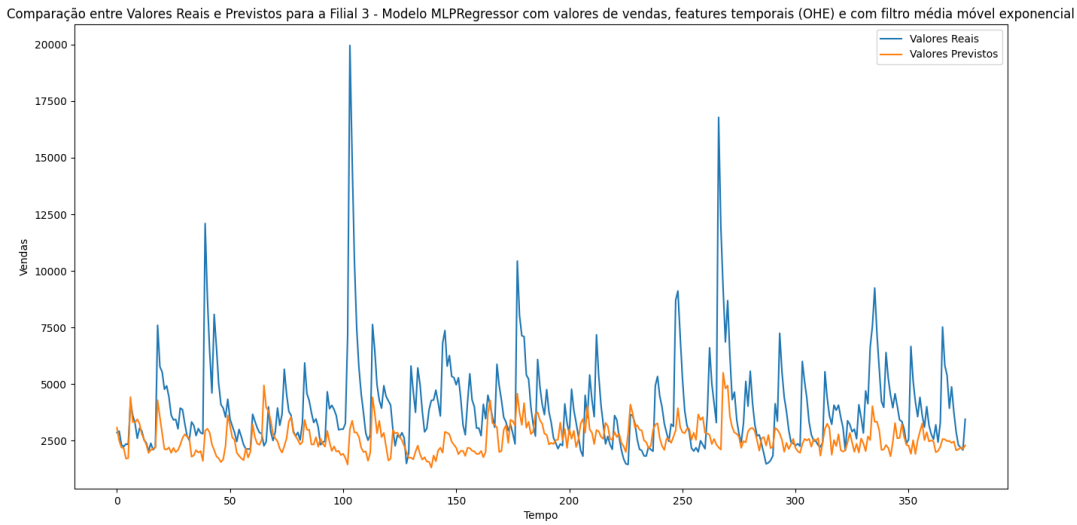


Figura 4.25: Comparação entre valores reais e previstos para filial 3 com modelo MLP Regressor usando dados de vendas, *features* temporais com OHE e um filtro de média móvel exponencial

limitações para tentar alavancar os resultados obtidos, melhorando a precisão das previsões e, consequentemente, as suas métricas de avaliação. Observando os modelos estudados pensou-se que a junção do LSTM, modelo inicialmente equacionado, com o XGBoost poderia possivelmente resultar na construção de um modelo que pudesse satisfazer isso mesmo, afinar as previsões de vendas e apresentar melhores resultados nas métricas de avaliação. Esta junção dos dois modelos num só poderia ser abordada de duas formas diferentes, uma delas em que se usa o modelo LSTM para identificar as *features* e se adicionam esses valores como entrada no modelo XGBoost para que este realize as previsões finais e outra em que se realizam as previsões de vendas individualmente e no final juntam-se esses valores para se obter o valor de previsão de vendas.

Recorrendo à primeira possibilidade, juntaram-se as forças de ambos os modelos para criar um. O LSTM foi usado para identificar padrões, tendências, sazonalidade e dependências de longo prazo, *features* estas que alimentaram depois o XGBoost como *inputs*, juntamente com as *features* temporais definidas na base do programa, codificadas com OHE, e o valor das vendas. Posteriormente, o XGBoost realizou as previsões das vendas com *features* enriquecidas pelo LSTM e mais rapidamente do que este o faria individualmente, dando assim origem aos resultados para as previsões de vendas. Nesta opção, para a parte do LSTM, foi definido, tal como nos modelos anteriores, um bloco de cinco dias para prever um e dividido o conjunto de dados em 80% para treino e 20% para teste. As *features* temporais foram adicionadas logo desde a sua génese, também codificadas por OHE, ou seja, o vetor de entrada para este bloco do LSTM seria igual ao aplicado na maioria dos modelos, constituído por dados de vendas e *features* temporais codificadas com OHE. A arquitetura da rede foi construida com duas camadas LSTM, uma com 100 neurónios e outra com 50. No entanto, ao contrário do que vinha acontecendo com os modelos anteriores, para este utilizou-se um camada densa intermédia de 25 neurónios para tentar refinar a aprendizagem. Quanto ao treino foram também definidas 100 épocas e um ajuste dos pesos a cada 32 elementos. Já para o bloco da previsão final, a fase em que entra o XGBoost, foi definido que este seria composto por 100 árvores sequenciais, com uma profundidade de 6 níveis e com uma taxa de aprendizagem baixa, ou seja, 0,1. Nesta parte do modelo, o vetor de entrada para XGBoost é maior do que o anterior porque além dos dados iniciais, adicionam-se a este vetor os valores da relações temporais retirados do modelo LSTM na primeira fase. Para esta possibilidade, cada modelo fica responsável por assumir uma parte do modelo híbrido. Compilou-se esta versão e analisaram-se os resultados obtidos, que, ao contrário da expetativa, não foram ao encontro do esperado.

Depois de analisados os valores para as métricas de avaliação e na procura de os melhorar, equacionou-se o desenvolvimento de uma otimização diferente, recorrendo para tal ao "grid search", ou seja, à técnica de validação cruzada em grelha. Com isto era obje-

Tabela 4.6: Parâmetros para otimização - *Grid Search*

| **Grid Search** | |
|---|---|
| **Conjunto de parâmetros** | **Valores para combinações** |
| Nº de árvores | [300; 500; 700; 900] |
| Taxa de aprendizagem | [0,01; 0,05; 0,1] |
| Profundidade das árvores | [3; 6; 9] |
| % de amostras por árvore | [0,8; 1] |
| % *features* por árvore | [0,8; 1] |

tivo testar o máximo de combinações possíveis de parâmetros para o XGBoost para que a partir daí se tentassem realizar as previsões o mais precisas possível. Este otimizador foi configurado para realizar combinações entre vários números de árvores, taxas de aprendizagem, profundidade das árvores, percentagens de amostras por árvore e de *features* utilizadas em cada uma (tabela 4.6). Aqui, foi usada uma validação cruzada de cinco dobras e o erro quadrático médio negativo para avaliar os modelos desenvolvidos com as diferentes combinações (quanto mais próximo de zero, melhor seria a combinação). Depois de programada e testada esta nova tentativa de melhoria, surgiram os resultados da combinação e os valores para as métricas de avaliação do modelo híbrido com o XGBoost a realizar as previsões com aqueles que foram considerados os melhores parâmetros. Surpreendentemente, os valores obtidos depois desta alteração foram praticamente os mesmos do que os verificados anteriormente, ou seja, algo distantes do que era pretendido.

A outra forma de desenvolver o modelo híbrido passava por cada modelo realizar as suas previsões individualmente, realizando-se no final o *ensemble* das mesmas. Ao contrário do que acontece com a possibilidade anterior, neste caso tanto o LSTM como o XGBoost fazem as suas previsões, realizando posteriormente a média ponderada entre ambos os resultados, obtendo-se assim os valores de previsão de vendas final. Para esta versão do modelo híbrido foram usados muitos parâmetros iguais aos definidos para os modelos anteriores. Também aqui foram logo adicionadas as *features* temporais ao programa, codificadas com OHE. A janela de previsão de cinco valores para prever um manteve-se, o vetor de entrada era também ele igual ao da maioria dos modelos testados, valor de vendas normalizado com a codificação das *features* temporais com OHE, bem como a divisão dos dados de 80% para treino e 20% para teste. Para o LSTM recorreu-se a uma arquitetura com duas camadas, uma com 100 e outra com 50 neurónios, respetivamente, utilizando a função "relu" para a sua ativação. O treino foi definido para 100 épocas com uma atualização de pesos a cada 32 amostras. Para o otimizar usou-se o algoritmo "adam" com uma taxa de aprendizagem de 0,0005. Quanto ao XGBoost, os parâmetros usados foram aqueles que tinham sido considerados como os melhores no "grid search" do modelo anterior, ou seja, 900 árvores, com uma taxa de aprendizagem de 0,05 em cada uma e uma profundidade máxima de 6. Para esta pos-

sibilidade, pôde ainda definir-se qual o modelo que preferencialmente iria apresentar maior peso dentro do modelo híbrido, sendo atribuída uma determinada percentagem a cada um relativamente ao nível de importância que representariam no valor de previsão final. Uma vez que o XGBoost foi o modelo que apresentou melhores valores nas métricas de avaliação quando trabalhados individualmente, optou-se por lhe atribuir maior peso nesta fase, obtendo-se valores muito melhores do que para o inverso, ou seja, atribuição de maior peso aos valores do LSTM para a obtenção das previsões finais. Para retirar estas conclusões foram testados várias combinações, verificando-se que os resultados iam melhorando consoante fosse aumentado o peso do modelo XGBoost.

Ambas as variantes do modelo explicadas anteriormente foram testadas, desde logo com a adição de *features* temporais e com filtro de média móvel exponencial nos dados de entrada, uma vez que se vinha demonstrando a melhor opção nos resultados dos modelos anteriores. Contudo, e apesar de serem combinadas as forças do LSTM e do XGBoost, os resultados alcançados não foram os esperados, continuando as melhores métricas de avaliação e os melhores gráficos a pertencer ao XGBoost individualmente para a previsão dos valores de vendas da empresa.

Na tabela 4.7 estão apresentados os valores das métricas de avaliação para cada modelo estudado, LSTM, GRU, MLP Regressor, XGBoost e ainda um modelo híbrido onde se juntaram os modelos LSTM e XGBoost. Nesta tabela podem ser verificados os resultados para as métricas dos modelos estudados com os valores de vendas e *features* temporais codificadas com o OHE, os valores das métricas acrescentando em cada modelo o filtro de média móvel simples (Filtro MMS) com intervalo de cinco dias e ainda os mesmos valores para os modelos mas com o filtro de média móvel exponencial (Filtro MME), igualmente de cinco dias. Os resultados apresentados na tabela para as métricas referentes ao modelo híbrido com *enssemble*, a segunda opção, tem em consideração o modelo que gera as previsões com peso de 40% do LSTM e 60% do XGBoost.

De acordo com os resultados obtidos para as métricas de avaliação dos diferentes modelos, que podem ser observados na tabela 4.7, pôde-se verificar que todos os resultados eram mais vantajosos depois de aplicados os filtros nos dados de entrada, verificando-se ainda melhores valores para o filtro da média móvel exponencial relativamente aos da média móvel simples. Contudo, mesmo aplicando esses filtros, os resultados das métricas estavam ainda distantes do pretendido. O único valor que se aproximou ligeiramente do que se poderia considerar como um resultado satisfatório na previsão de vendas foi o modelo XGBoost com um valor de MAPE a rondar os 16,45%.

Tabela 4.7: Métricas de avaliação dos modelos para a filial 3 sem filtro e com filtros de média móvel simples (MMS) e exponencial(MME)

| **Modelo** | **Métricas de avaliação** | | | |
|---|---|---|---|---|
| | **MSE (€)** | **RMSE (€)** | **MAE (€)** | **MAPE (%)** |
| LSTM | 24 511 796,99 | 4 950,94 | 2 355,98 | 58,35 |
| GRU | 24 929 343,67 | 4 992,93 | 2 497,31 | 71,63 |
| MLPRegressor | 22 674 147,37 | 4 761,74 | 2 398,68 | 75,39 |
| XGBoost | 22 792 838,22 | 4 774,18 | 2 110,69 | 56,47 |
| LSTM + Filtro MMS | 6 422 168,74 | 2 534,20 | 1 724,91 | 36,69 |
| GRU + Filtro MMS | 3 359 415,74 | 1 832,87 | 1 258,18 | 28,38 |
| MLPRegressor + Filtro MMS | 5 476 238,58 | 2 340,14 | 1 587,37 | 34,16 |
| XGBoost + Filtro MMS | 2 710 266,22 | 1 646,29 | 878,98 | 19,20 |
| LSTM + Filtro MME | 6 090 623,37 | 2 467,92 | 1 620,14 | 34,37 |
| GRU + Filtro MME | 4 650 941,79 | 2 156,60 | 1 350,59 | 28,52 |
| MLPRegressor + Filtro MME | 5 908 926,96 | 2 430,83 | 1 579,37 | 33,37 |
| XGBoost + Filtro MME | 2 739 846,14 | 1 655,25 | 791,77 | 16,45 |
| 1ª Opção - Hibrido + Filtro MME | 5 436 172,56 | 2 331,56 | 1 451,04 | 29,78 |
| 2ª Opção - Hibrido + Filtro MME | 3 564 107,61 | 1 887,88 | 1 029,23 | 20,50 |

# 5 Análise crítica e discussão dos resultados

Este capítulo apresenta uma visão crítica e objetiva sobre os resultados obtidos ao longo deste projeto, desde a análise dos dados até ao desenvolvimento dos diferentes modelos de previsão testados no seu âmbito.

O principal objetivo deste projeto passava por integrar uma ferramenta que auxiliasse na definição de estratégicas mais robustas e devidamente suportadas, baseadas num modelo de previsão de vendas assente no histórico de faturação da empresa.

A discussão apresentada é estruturada tendo por base as questões centrais de investigação formuladas inicialmente. Aqui é estabelecida uma conexão entre os resultados e as hipóteses inicialmente delineadas, o que possibilita a avaliação sobre o trabalho desenvolvido e respetiva apreciação face aos seus objetivos. É também estabelecida uma ligação entre os resultados e a literatura de suporte ao tema, procurando demonstrar de que forma é que o observado pode contribuir para a área da previsão de vendas, sobretudo modelos de previsão baseados em *Machine Learning*.

Os resultados são cuidadosamente discutidos e interpretados, procurando apresentar não apenas o que se obteve mas também o seu significado e relevância para o projeto.

**1. A partir da análise do histórico da empresa (dados de faturação) é possível compreender as suas dinâmicas e evolução ao longo do tempo? Conseguem identificar-se padrões claros na evolução das vendas?**

A análise exploratória dos dados permitiu verificar uma evolução dinâmica e heterogénea das vendas ao longo dos anos para as diferentes filiais que constituem a empresa.

Ao nível da faturação ficou claro que os comportamentos das filiais são bastante diferentes, com a filial 1 e a filial 2 a liderarem em volumes de vendas, mas com perfis distintos no que respeita à emissão de notas de crédito, um fator que tem impacto direto na análise sobre a *performance* real de faturação. As variações nos rácios de notas de crédito e faturação por filial revelaram alguns detalhes interessantes, sobretudo em relação aos perfis comerciais, que se manifestaram bastante diferentes. Neste ponto, a filial 2 esteve em destaque, porque apesar de liderar em faturação, também concentrava os maiores valores em notas de crédito. Estas observações permitiram perceber que a empresa não apresenta um comportamento uniforme em todas as unidades, variando de zona para zona, o que revela a importância que poderá ter um modelo que seja adaptável por filial.

De acordo com os dados disponíveis é possível concluir que o histórico de faturação

da empresa contém padrões estruturais bem definidos, como sazonalidade, tendências locais e oscilações cíclicas. Existem efetivamente flutuações muito grandes no que respeita à faturação diária, com uma incidência por demais evidente nos meios e finais de cada mês. Este destaque resulta do tipo de faturação que se verificava na empresa, que decorria sob a forma de guias de remessa. Neste tipo de faturação, acontece que as guias de remessa lançadas até ao meio de cada mês têm a respetiva fatura emitida apenas no dia útil que corresponde efetivamente ao meio do mês, normalmente a 15, enquanto que as guias de remessa lançadas entre o meio e o final de cada mês, têm as respetivas faturas emitidas apenas no último dia útil do mês. Esta forma de faturação leva a que se manifestem, inequivocamente, valores discrepantes no meio e finais de cada mês, o que contribui para uma forte variação das faturações diárias, com picos constantes nestes períodos mensais, sobretudo até julho de 2013.

Através do estudo detalhado sobre cada filial e da análise temporal foi possível identificar determinados padrões mensais. Meses como julho e setembro são consistentemente os meses de maior volume de faturação, o que pode estar relacionado com a sazonalidade industrial e as operações de manutenção planeada dos clientes da empresa, enquanto agosto e dezembro revelam quebras, que pode ser justificado por períodos de férias e encerramentos de fábricas. Estes elementos foram essenciais para justificar a introdução de variáveis temporais nos modelos e sustentam que a evolução das vendas é cíclica e poderá ser previsível em determinadas alturas.

Outro facto a destacar na análise dos dados está relacionado com a comparação realizada entre a evolução de faturação para cada filial e para a empresa no seu todo, fazendo depois uma analogia com as taxas de inflação verificadas para Portugal nos respetivos anos. O objetivo passava por perceber se os crescimentos de faturação anual, quando verificados, eram superiores ou não à taxa de inflação, ou se poderiam ser crescimentos "fictícios". Nos anos em que se verificaram crescimentos, quase todas as filiais apresentaram um crescimento superior face às taxas de inflação dos respetivos anos, destacando aqui dois casos em que isso não se verificou, para a filial 3 em 2017 e a filial 6 em 2016, que tiveram os seus crescimentos inferiores à taxa de inflação para os referidos anos. Apesar disso, a comparação entre a evolução da faturação e a taxa de inflação nacional permitiu verificar que, em vários anos, o crescimento da empresa foi real e não meramente nominal, destacando um crescimento económico sustentado em algumas das filiais, destacando a filial 3.

**2. Qual o melhor modelo preditivo de vendas que leve em consideração o histórico de vendas da empresa? Qual será o desempenho preditivo de diferentes modelos de *Machine Learning* na previsão de vendas da empresa, tendo por base o seu histórico de faturação?**

No decorrer do projeto foram testados cinco modelos de previsão, nomeadamente, LSTM, GRU, MLP Regressor, XGBoost e um modelo híbrido que combinava LSTM e

XGBoost.

As métricas definidas para avaliação dos modelos foram o MSE, RMSE, MAE e MAPE, tendo sido possível efetuar uma comparação rigorosa sobre a qualidade das suas previsões (tabela 4.7).

O modelo inicialmente considerado, LSTM, foi à partida o escolhido por se tratar de um modelo amplamente reconhecido no que a séries temporais, identificação de padrões e tendências diz respeito. Depois de alguns testes, verificou-se que os resultados obtidos estavam bastante distantes do pretendido, o que pode ser justificado pelo reduzido tamanho do *dataset* e as muitas flutuações nos dados disponíveis. Procurou-se então recorrer a um modelo que por um lado desse também as suas garantias no que respeita a séries temporais mas que fosse menos complexo, equacionando-se o modelo GRU. Depois de testado sob várias formas verificou-se que os resultados das previsões geradas por este algoritmo eram igualmente distantes do pretendido, começando a equacionar-se outro modelo de previsões. De acordo com Rafi *et al.* [34], embora os modelos LSTM e GRU se evidenciem na ciência de dados como bons modelos para a previsão de séries temporais, pode verificar-se que não são ideais para todo o tipo de dados.

Posteriormente considerou-se o MLP Regressor, que apresentou resultados também consideravelmente distantes do esperado, até que se equacionou um modelo de *Machine Learning*, desta vez não de redes neuronais mas de *Gradient Boosting*, o XGBoost. Por último, verificou-se a possibilidade de juntar dois modelos para criar um, o LSTM e o XGBoost, o designado modelo híbrido, desenvolvido de duas formas. Uma delas em que cada modelo era responsável por uma parte do modelo combinado, o LSTM gerava *features* aprimoradas que seriam depois usadas pelo XGBoost para a previsão final dos valores. A outra forma em que cada modelo realizava as previsões individualmente, fazendo-se depois a média ponderada entre as previsões de ambos para obter as previsões finais.

De todos os modelos, o XGBoost com filtro de média móvel exponencial foi o modelo que apresentou o melhor desempenho absoluto, com um MAPE de 16,45%. O modelo híbrido, na segunda versão desenvolvida, com uma média ponderada de previsões, 60% XGBoost e 40% LSTM, aproximou-se consideravelmente dos melhores resultados com MAPE de 20,5%, demonstrando potencial na combinação de modelos, mas sem superar o XGBoost isolado. Os modelos LSTM, GRU e MLPRegressor, mesmo como filtros e tentativas de otimizações testadas, ficaram aquém do desejado, com MAPE acima de 28%, revelando maior sensibilidade às oscilações e menor capacidade de generalização com o volume e o tipo de dados disponível.

Tendo por base o histórico de faturação da empresa, o XGBoost foi o modelo preditivo mais adequado, apresentado melhor combinação entre desempenho, velocidade e estabilidade. Esta comparação evidencia que, para o contexto específico da empresa

e com o volume dos dados com que se desenvolveram os modelos, sete anos e meio, modelos de *boosting* baseado em árvores de decisão como o XGBoost são mais eficazes do que os de redes neuronais.

**3. É possível identificar e incorporar variáveis relevantes no modelo de previsão de vendas? Essas variáveis podem influenciar a performance dos modelos de previsão de vendas?**

Na tentativa de melhorar os modelos de previsão que foram desenvolvidos ao longo do projeto, foram testadas diferentes alternativas de otimização. A adição de *features* temporais (dia da semana, dia do mês, mês e dia do ano) e a sua codificação por OHE permitiram alcançar uma melhoria nos resultados das métricas de avaliação e, consequentemente na *performance* dos modelos.

A introdução das *features* temporais nos modelos auxiliou no reconhecimento de determinados padrões, tendências e sazonalidades que sem elas poderiam não ser identificadas, sobretudo nos modelos não sequenciais (MLP Regressor e XGBoost) que requerem variáveis explícitas para esse efeito.

O modelo LSTM, embora mais apto a aprender relações sequenciais, também beneficiou ligeiramente da inclusão de variáveis temporais como entrada, no entanto o ganho não foi suficiente.

O uso da codificação OHE nas *features* permitiu evitar relações ordinais artificiais, reforçando a capacidade dos algoritmos em captar padrões cíclicos.

No entanto, convém salientar que, apesar de melhorar a *performance* dos modelos, os benefícios alcançados não foram tão significativos quanto era esperado.

**4. Qual o impacto que poderá ter o ajuste de hiperparâmetros e o pré-processamento de dados na precisão dos modelos de previsão?**

O processo de ajuste de hiperparâmetros, também designado por *tuning*, contribuiu para otimizações subtis dos modelos mas não com impacto preponderante e transformador.

Ao longo do projeto, e para cada modelo, foram testadas diversas combinações, alterando-se os parâmetros como a janela de previsão, a arquitetura (quantidade de camadas, de neurónios, funções de ativação, número de árvores, profundidade), funções de otimização, taxas de aprendizagem, alterações ao nível do treino, taxas de atualização de pesos, entre outras. No entanto, estes ajustes mostraram-se sempre insuficientes para ultrapassar as limitações impostas pela qualidade dos dados. Mesmo no modelo híbrido, em que numa das versões desenvolvidas se recorreu ao "grid search" para encontrar a combinação ótima de parâmetros, a melhoria que se verificou não foi significativa.

O pré-processamento de dados, por sua vez, acrescentou valor significativo a todos os

modelos, de forma transversal. A aplicação de filtros de suavização, sobretudo a média móvel exponencial, teve um impacto claro na melhoria das métricas de avaliação, reduzindo o ruído e as elevadas flutuações nos dados, tornando as tendências e padrões subjacentes mais visíveis. A precisão dos modelos aumentou consideravelmente com a aplicação dos filtros, destacando-se o XGBoost, que beneficiou particularmente do pré-processamento dos dados.

O pré-processamento mostrou ter mais impacto prático do que o ajuste de hiperparâmetros, o que reforça a importância de um *pipeline* de dados bem construido.

**5. A combinação de modelos pode melhorar os resultados das previsões face aos modelos individuais?**

Com o desenvolvimento do modelo híbrido procurou-se concatenar as mais-valias dos modelos que o constituíram, com o intuito de se obter um modelo de previsão que fosse mais robusto e preciso.

O modelo híbrido desenvolvido seguiu duas abordagens distintas, uma em que se recorreu ao modelo LSTM para gerar *features* mais refinadas, que juntamente com as definidas inicialmente no programa foram utilizadas como entrada para o modelo XGBoost realizar as previsões finais e outra em que se combinavam as previsões, realizadas individualmente, dos dois modelos através de um *ensemble* ponderado, originando assim as previsões finais.

A primeira abordagem não obteve resultados muito satisfatórios e, por isso, ficou um pouco aquém do que era esperado quando se pensou no desenvolvimento deste modelo. A segunda abordagem, por sua vez, apresentou resultados bastante superiores ao da primeira, sobretudo à medida que se atribuía maior peso ao XGBoost na previsão de valor final. Contudo, e apesar da segunda versão apresentar melhores resultados do que a primeira, nenhum deles foi superior ao melhor modelo individual, o XGBoost, que apresentou um MAPE de 16,45%.

A modelação híbrida, sobretudo a segunda versão, demonstrou potencial teórico, contudo, no que respeita à prática, e neste projeto especificamente, o XGBoost acabou por se revelar mais robusto do que qualquer combinação.

Os resultados obtidos neste projeto estão, em vários aspetos, alinhados com alguns estudos analisados na secção 2.1, onde se apresenta um enquadramento global de abordagens utilizadas na previsão de vendas. Em diferentes pontos, pode verificar-se uma convergência entre algumas das conclusões apresentadas pelos autores e os dados empíricos produzidos neste estudo, o que reforça a validade e a atualidade da abordagem adotada.

Tal como neste projeto, a literatura analisada confirma a crescente tendência na adoção de modelos baseados em ML para a previsão de vendas. Este facto pode ser confirmado pelos trabalhos de Giri *et al.* [12], de Rumbe *et al.* [1] e Hewage e Perera [10] pelos

resultados obtidos quando comparam modelos ML com métodos clássicos.

Silva [3] destaca que modelos não lineares, como as redes neuronais artificiais, superam os modelos lineares tradicionais em termos de precisão, apresentando previsões mais fidedignas. Embora essa comparação direta não tenha sido realizada neste estudo, uma vez que foram testados apenas modelos não lineares, os resultados obtidos ajudam a confirmar a eficácia dessa classe de modelos na previsão de vendas. Foram implementados os modelos LSTM, GRU, MLP Regressor e foi o XGBoost, um modelo não linear baseado em *ensemble learning*, que se destacou com o melhor desempenho global, com um MAPE de 16,45%, ultrapassando até os modelos neuronais em precisão, estabilidade e robustez (com um erro cima de 28%). No entanto a autora acabou por optar por um modelo múltiplo, combinando diferentes abordagens para melhorar o desempenho preditivo. A tentativa de construir um modelo híbrido para tentar melhorar os resultados, combinando LSTM e XGBoost, vai ao encontro da abordagem de Silva [3], embora neste caso os ganhos adicionais não tenham justificado a complexidade acrescida, ao contrário do observado pela autora.

Almeida e Passari [4] também sustentam a eficácia das redes neuronais multicamada e de retropropagação em vendas de curto prazo quando comparados com técnicas de regressão linear e modelação Naïve. No entanto, alertam para a sua fragilidade, reconhecendo que os erros são elevados em horizontes temporais curtos e que tenderiam a agravar-se a longo prazo. Esta observação está em consonância com o presente projeto, onde se observou que tanto o LSTM como o GRU, embora com maior capacidade teórica de modelação sequencial, foram penalizados pela elevada variação das vendas diárias e a falta de sequências contínuas de dados limpos, o que confirma as limitações apontadas pelos autores.

Já Ponte [5] recorreu à regressão linear múltipla com resultados satisfatórios para a precisão de vendas de seguros de saúde. Neste projeto, devido à complexidade do domínio industrial e considerando a natureza não linear e dinâmica dos dados analisados, os modelos lineares simples não foram considerados adequados e, por isso, não foram equacionados, uma vez que a variabilidade do mercado alvo exigia abordagens mais sofisticadas, robustas e adaptadas a relações não lineares.

Eiglsperger *et al.* [6] reforçam que, entre todos os modelos testados para previsão de vendas de produtos hortícolas recorrendo ao histórico de seis empresas, os modelos de ML foram os que apresentaram os melhores resultados. Dentro destes, os autores destacaram o XGBoost como o modelo com melhor desempenho. Esta evidência é inteiramente confirmada pelo presente estudo, no qual o XGBoost também se revelou o modelo mais eficaz, com um MAPE de 16,45%, o mais reduzido entre as abordagens testadas.

Por fim, Hicham *et al.* [13] desenvolveram um modelo híbrido, demonstrando que a complementaridade entre técnicas pode ser benéfica para reduzir o erro. Esta ideia

foi igualmente explorada neste projeto. Contudo, apesar de promissor, o modelo não ultrapassou o desempenho do XGBoost isolado, o que poderá indicar que a eficácia dos modelos híbridos depende da qualidade e complementaridade dos modelos base estudados.

Outro ponto comum entre os estudos analisados e este projeto está relacionado com a importância dos dados, da sua qualidade e da engenharia de *features* para o sucesso dos modelos. Eiglsperger *et al.* [6] salientam que a *performance* dos modelos preditivos é fortemente influenciada pelos conjuntos de dados utilizados, observação esta que se confirmou igualmente neste projeto, revelando-se um ponto determinante para os resultados alcançados. Silva [3], Guo *et al.* [11], Hewage e Perera [10] destacam a relevância de selecionar, combinar e adicionar as variáveis cuidadosamente para compreender o comportamento das vendas ao longo do tempo e para melhorar os modelos. Este princípio foi aplicado também neste projeto para tentar melhorar os modelos, neste caso específico com a inclusão das *features* temporais e a codificação de OHE, o que permitiu aos modelos desenvolvidos, sobretudo ao XGBoost, captar padrões sazonais e cíclicos que de outro modo seriam ignorados.

Depois de realizada a análise sobre os resultados obtidos, verificou-se que não existe um modelo universal que possa ser considerado o melhor. O melhor desempenho resulta de uma combinação equilibrada de técnica utilizada, qualidade dos dados, pré-processamento e ajuste ao contexto em que se insere.

# 6 Conclusão

Este projeto foi pensado e desenvolvido com o propósito de otimizar a previsão de vendas de uma empresa, tornando-a mais dinâmica, moderna, interativa e eficiente, com o objetivo de obter resultados mais robustos e precisos.

No atual contexto da indústria, a previsão de vendas é um processo cada vez mais complicado, pela variabilidade de procura, pela elevada concorrência, globalização de mercado, comunicação e marketing agressivos das empresas do setor e ciclos curtos de determinados produtos. No entanto, este desafio realça a relevância prática no trabalho desenvolvido, na medida em que poderá dotar a empresa de uma ferramenta preditiva concreta, mais dinâmica, com fundamentação técnica e potencial de aplicação operacional.

Através da análise de dados históricos de faturação da empresa e da aplicação de uma metodologia CRISP-DM, foram desenvolvidos e comparados vários modelos preditivos, desde o LSTM, ao GRU, ao MLP Regressor, o XGBoost e um modelo híbrido que combinava o LSTM com o XGBoost. O modelo XGBoost destacou-se com um MAPE de 16,45%, demonstrando maior robustez e capacidade de adaptação face aos restantes, especialmente em cenários com dados não lineares e descontinuidades temporais.

No decorrer do desenvolvimento deste projeto e consoante os resultados alcançados, que embora não sendo perfeitos revelaram-se encorajadores, foi possível identificar algumas limitações que devem ser reconhecidas. A primeira diz respeito à qualidade de dados, uma vez que o *dataset* disponível apresentava bastante ruído, manifestado sobretudo por grandes variações entre os valores de faturação e um número considerável de *outliers*, tornando mais difícil o tratamento de dados e a capacidade de generalização dos modelos para a realização das previsões. O volume de dados foi outro problema identificado. Apesar de serem disponibilizados dados de histórico referente a onze anos de faturação, apenas foram utilizados sete anos e meio. Esta decisão foi positiva porque auxiliou os modelos na sua capacidade de generalização, contudo, trouxe consigo algumas desvantagens sobretudo para modelos sequenciais, como é o caso do LSTM e o GRU. Por último, a adaptação do modelo que atingiu melhores resultados para outras realidades. O modelo foi treinado e testado no contexto específico da empresa em estudo, recorrendo ao histórico de faturação de um determinado período. Para sua generalização e futura adaptação a outras realidades, o modelo carecerá sempre de uns pequenos ajustes ou, eventualmente, de uma nova revisão para se verificar se é o modelo mais adequado ao pretendido.

Para a organização em estudo, este projeto poderá vir a ter um impacto bastante positivo. O mais relevante deverá estar relacionado com o apoio que poderá providenciar nas decisões financeiras e comerciais, auxiliando na definição de estratégias das respetivas áreas, podendo auxiliar na melhoria e controlo de investimento e tesouraria. No que respeita à antecipação de variações na procura, poderá auxiliar no ajuste de *stocks*, possibilitando a definição de uma estratégia de *stock* e compra mais ajustada à necessidade que se previrá vir a demonstrar. Outra vantagem que a implementação deste modelo poderá originar está relacionada com a otimização de recursos humanos e com o planeamento de futuras ações de comunicação, adaptando as equipas e a realização de determinadas campanhas publicitárias de acordo com os ciclos de vendas identificados.

Com o intuito de colmatar algumas das lacunas identificadas no desenvolvimento do projeto e para lhe tentar incrementar valor, são deixadas algumas recomendações para investigação futura. Para desenvolvimentos posteriores é sugerido que sejam realizadas previsões multivariadas, ou seja, tentar formatar o *dataset* para obter dados com maior qualidade e trabalhar a informação para que se consigam realizar previsões por produto, gama de produto, família, entre outras. Desta forma, será possível proporcionar informação mais detalhada e com bastante potencial para auxiliar nas definições de estratégias e na tomada de decisão por parte dos gestores de topo. Outra sugestão está relacionada com a adição de variáveis externas no modelo. Fatores como indicadores económicos, calendário laboral, calendário de determinadas campanhas poderiam ser um forte incremento para enriquecer o modelo e aumentar a sua capacidade preditiva. De futuro, poderão também ser explorados outros modelos de previsão por forma a verificar se eventualmente poderão ser mais indicados para o tipo e quantidade de dados disponível. Outra sugestão poderia ser a extensão e integração do modelo preditivo com a análise de dados, implementada através de um *dashboard*. Desta forma, as informações poderiam circular e ser apresentadas em tempo real, permitindo auxiliar a administração no seu processo de decisão sobre a empresa, nas suas análises e na definição de estratégias futuras.

A partir da análise aos resultados obtidos, pode-se então verificar que o histórico de faturação de uma empresa é preponderante para que se consigam desenvolver modelos de previsões de vendas que auxiliem a gestão de topo na definição estratégica da empresa e, com isso, contribuir para o sucesso de uma organização. A qualidade e o volume dos dados disponíveis são determinantes para estes modelos. A utilização de modelos mais complexos de redes neuronais, como LSTM, ou até da combinação de modelos, como os híbridos, só se justifica para casos em que o volume de dados é maior, existem mais variáveis externas ou para situações em que os objetivos preditivos possam ser mais complexos. Na procura da otimização dos modelos, a engenharia de *features* é bastante importante em determinados modelos, fundamental para outros.

Contudo, se os dados forem muito oscilantes e ocorrerem muitas e vincadas variações nos seus valores, torna-se difícil tirar o máximo proveito do benefício que podem conferir. O pré-processamento dos dados, com a aplicação de filtros nos dados de entrada dos modelos, revelou-se fundamental para a obtenção de melhores métricas e, consequentemente, de modelos mais precisos. O XGBoost, modelo que apresentou os melhores resultados no paradigma do projeto efetuado, demonstrou ser um modelo robusto, eficaz e confiável para a realização de previsões, revelando-se o mais adequado para implementar na empresa em estudo.

# Referências bibliográficas

[1] G. Rumbe, M. Hamasha, and S. Al Mashaqbeh, "A comparison of holts-winter and artificial neural network approach in forecasting: A case study for tent manufacturing industry," *Results in Engineering*, vol. 21, p. 101899, 2024.

[2] R. Thangeda, N. Kumar, and R. Majhi, "A neural network-based predictive decision model for customer retention in the telecommunication sector," *Technological Forecasting and Social Change*, vol. 202, p. 123250, 2024.

[3] A. M. M. da Silva, "Modelos preditivos aplicados ao retalho," 2015.

[4] F. C. de Almeida and A. F. L. Passari, "Previsão de vendas no varejo por meio de redes neurais," *Revista de Administração-RAUSP*, vol. 41, no. 3, pp. 257–272, 2006.

[5] M. B. V. Ponte, "Modelos preditivos de vendas de produtos da multicare," Ph.D. dissertation, 2022.

[6] J. Eiglsperger, F. Haselbeck, V. Stiele, C. G. Serrano, K. Lim-Trinh, K. Menrad, T. Hannus, and D. G. Grimm, "Forecasting seasonally fluctuating sales of perishable products in the horticultural industry," *Expert Systems with Applications*, p. 123438, 2024.

[7] F.-C. Yuan and C.-H. Lee, "Sales volume forecasting decision models," in *2011 International Conference on Technologies and Applications of Artificial Intelligence*. IEEE, 2011, pp. 239–244.

[8] K. Osawa, T. Tabata, and T. Hosoda, "Product sales forecast model considering circular fluctuations," in *2021 10th International Congress on Advanced Applied Informatics (IIAI-AAI)*. IEEE, 2021, pp. 897–902.

[9] L. Wu, J. Y. Yan, and Y. J. Fan, "Data mining algorithms and statistical analysis for sales data forecast," in *2012 Fifth International Joint Conference on Computational Sciences and Optimization*. IEEE, 2012, pp. 577–581.

[10] H. C. Hewage and H. N. Perera, "Comparing statistical and machine learning methods for sales forecasting during the post-promotional period," in *2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*. IEEE, 2021, pp. 462–466.

[11] Z. Guo, M. Li, and W. K. Wong, "Intelligent multivariate sales forecasting using wrapper approach and neural networks," in *IEEE 10th International Conference on Industrial Informatics*. IEEE, 2012, pp. 145–150.

[12] C. Giri, S. Thomassey, J. Balkow, and X. Zeng, "Forecasting new apparel sales using deep learning and nonlinear neural network regression," in *2019 International Conference on Engineering, Science, and Industrial Applications (ICESI)*. IEEE, 2019, pp. 1–6.

[13] A. Hicham, B. Mohamed, and E. F. Abdellah, "A model for sales forecasting based on fuzzy clustering and back-propagation neural networks with adaptive learning rate," in *2012 IEEE International Conference on Complex Systems (ICCS)*. IEEE, 2012, pp. 1–5.

[14] M. Niranjanamurthy, J. A. Babu *et al.*, "Enactment of sales forecasting application using artificial intelligence techniques," in *2024 Second International Conference on Data Science and Information System (ICDSIS)*. IEEE, 2024, pp. 1–7.

[15] C. Zhang and R. Shi, "Research on sales forecasting model based on gru neural network and machine learning model," in *2023 IEEE 3rd International Conference on Data Science and Computer Application (ICDSCA)*. IEEE, 2023, pp. 575–579.

[16] W. Xu, Y. Cao, and R. Chen, "A multimodal analytics framework for product sales prediction with the reputation of anchors in live streaming e-commerce," *Decision Support Systems*, vol. 177, p. 114104, 2024.

[17] Z.-L. Sun, T.-M. Choi, K.-F. Au, and Y. Yu, "Sales forecasting using extreme learning machine with applications in fashion retailing," *Decision support systems*, vol. 46, no. 1, pp. 411–419, 2008.

[18] Y. Ali and S. Nakti, "Sales forecasting: A comparison of traditional and modern times-series forecasting models on sales data with seasonality," in *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 2023, pp. 159–163.

[19] F. Guo, H. Mo, J. Wu, L. Pan, H. Zhou, Z. Zhang, L. Li, and F. Huang, "A hybrid stacking model for enhanced short-term load forecasting," *Electronics*, vol. 13, no. 14, p. 2719, 2024.

[20] X. Zeng, C. Liang, Q. Yang, F. Wang, and J. Cai, "Enhancing stock index prediction: A hybrid lstm-pso model for improved forecasting accuracy," *PloS one*, vol. 20, no. 1, p. e0310296, 2025.

[21] V. Kumar, N. Kedam, K. V. Sharma, D. J. Mehta, and T. Caloiero, "Advanced machine learning techniques to improve hydrological prediction: A comparative analysis of streamflow prediction models," *Water*, vol. 15, no. 14, p. 2572, 2023.

[22] E. Gothai, R. Rajalaxmi, R. Thamilselvan, and H. SM, "Forecasting price prediction for vegetables and fruits using recurrent neural network," in *2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE, 2024, pp. 1889–1896.

[23] N. Zhai, P. Yao, and X. Zhou, "Multivariate time series forecast in industrial process based on xgboost and gru," in *2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, vol. 9. IEEE, 2020, pp. 1397–1400.

[24] G. Cassales, S. Salekin, N. Lim, D. Meason, A. Bifet, B. Pfahringer, and E. Frank, "A comparative study of four deep learning algorithms for predicting tree stem radius measured by dendrometer: A case study," *Ecological Informatics*, vol. 86, p. 103014, 2025.

[25] S. Bouktif, A. Fiaz, A. Ouni, and M. A. Serhani, "Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches," *Energies*, vol. 11, no. 7, p. 1636, 2018.

[26] H. Wei and Q. Zeng, "Research on sales forecast based on xgboost-lstm algorithm model," in *Journal of Physics: Conference Series*, vol. 1754, no. 1. IOP Publishing, 2021, p. 012191.

[27] A. H. Alharbi, D. S. Khafaga, A. M. Zaki, E.-S. M. El-Kenawy, A. Ibrahim, A. A. Abdelhamid, M. M. Eid, M. El-Said, N. Khodadadi, L. Abualigah *et al.*, "Forecasting of energy efficiency in buildings using multilayer perceptron regressor with waterwheel plant algorithm hyperparameter," *Frontiers in Energy Research*, vol. 12, p. 1393794, 2024.

[28] H. Zermane, H. Madjour, A. Ziar, and A. Zermane, "Forecasting material quantity using machine learning and times series techniques," *Journal of Electrical Engineering*, vol. 75, no. 3, pp. 237–248, 2024.

[29] A. Gifty and Y. Li, "A comparative analysis of lstm, arima, xgboost algorithms in predicting stock price direction," *Engineering and Technology Journal*, vol. 9, no. 8, pp. 4978–4986, 2024.

[30] A. Tian, "Enhancing vegetable sales forecasting with a cnn-lstm-transformer hybrid model," *Highlights in Business, Economics and Management*, vol. 25, pp. 112–121, 01 2024.

[31] T. Liwei, F. Li, S. Yu, and G. Yuankai, "Forecast of lstm-xgboost in stock price based on bayesian optimization." *Intelligent Automation & Soft Computing*, vol. 29, no. 3, 2021.

[32] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[33] K. Honjo, X. Zhou, and S. Shimizu, "Cnn-gru based deep learning model for demand forecast in retail industry," in *2022 international joint conference on neural*

*networks (ijcnn)*. IEEE, 2022, pp. 1–8.

[34] M. A. Rafi, G. N. Rodrigues, M. N. H. Mir, M. S. M. Bhuiyan, M. Mridha, M. R. Islam, and Y. Watanobe, "A hybrid temporal convolutional network and transformer model for accurate and scalable sales forecasting," *IEEE Open Journal of the Computer Society*, 2025.

[35] A. Gasparin, S. Lukovic, and C. Alippi, "Deep learning for time series forecasting: The electric load case," *CAAI Transactions on Intelligence Technology*, vol. 7, no. 1, pp. 1–25, 2022.

[36] S. Bravo and Á. H. Moreno, "Prediction model based on neural networks for microwave drying process of amaranth seeds," in *Proceedings of the 2019 3rd International Conference on Compute and Data Analysis*, 2019, pp. 88–93.

[37] B. Pandey, S. Giri, R. D. Pant, M. Jalan, A. Chaudhary, and N. P. Adhikari, "Prediction of binding energy using machine learning approach," *AIP Advances*, vol. 14, no. 10, 2024.

[38] F. H. Rizk, M. E. Mohamed, B. Sameh, A. M. Zaki, M. M. Eid, and E.-S. M. Elkenawy, "Predictive modeling of portuguese student performance: comparative machine learning analysis," in *2024 International Telecommunications Conference (ITC-Egypt)*. IEEE, 2024, pp. 26–31.

[39] L. Semmelmann, S. Henni, and C. Weinhardt, "Load forecasting for energy communities: a novel lstm-xgboost hybrid model based on smart meter data," *Energy Informatics*, vol. 5, no. Suppl 1, p. 24, 2022.

[40] L. Lu, T. Yang, Z. Chen, Q. Ge, J. Yang, and G. Sen, "Prediction analysis of human brucellosis cases in ili kazakh autonomous prefecture xinjiang china based on time series," *Scientific Reports*, vol. 15, no. 1, p. 1232, 2025.

[41] M. Sumorek and A. Idzkowski, "Time series forecasting for energy production in stand-alone and tracking photovoltaic systems based on historical measurement data," *Energies*, vol. 16, no. 17, p. 6367, 2023.

[42] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying crisp-dm process model," *Procedia Computer Science*, vol. 181, pp. 526–534, 2021.

[43] A. Azevedo and M. F. Santos, "Kdd, semma and crisp-dm: a parallel overview," *IADS-DM*, 2008.

[44] F. Schäfer, C. Zeiselmair, J. Becker, and H. Otten, "Synthesizing crisp-dm and quality management: A data mining approach for production processes," in *2018 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)*. IEEE, 2018, pp. 190–195.

[45] A. Nadali, E. N. Kakhky, and H. E. Nosratabadi, "Evaluating the success level of data mining projects based on crisp-dm methodology by a fuzzy expert system,"

in *2011 3rd International Conference on Electronics Computer Technology*, vol. 6. IEEE, 2011, pp. 161–165.

[46] S. Maataoui, G. Bencheikh, and G. Bencheikh, "Predictive maintenance in the industrial sector: a crisp-dm approach for developing accurate machine failure prediction models," in *2023 Fifth International Conference on Advances in Computational Tools for Engineering Applications (ACTEA)*. IEEE, 2023, pp. 223–227.

[47] Z. Ahmad, S. Yaacob, R. Ibrahim, and W. F. W. Fakhruddin, "The review for visual analytics methodology," in *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*. IEEE, 2022, pp. 1–10.

[48] C. Catley, K. Smith, C. McGregor, and M. Tracy, "Extending crisp-dm to incorporate temporal data mining of multidimensional medical data streams: A neonatal intensive care unit case study," in *2009 22nd IEEE International Symposium on Computer-Based Medical Systems*. IEEE, 2009, pp. 1–5.

**Instituto Superior de Engenharia**
Politécnico de Coimbra