



Previsão de preços de eletricidade no mercado spot baseado em Machine Learning

EDUARDO SANTOS NOVO

Setembro de 2025

Previsão de preços de eletricidade no mercado spot baseado em Machine Learning

Eduardo Santos Novo

**Dissertação para obtenção do Grau de Mestre em
Engenharia Informática, Área de Especialização em
Engenharia de Dados**

Orientador: Ana Madureira

Co-orientador: Judite Ferreira

Declaração de Integridade

Declaro ter conduzido este trabalho académico com integridade.

Não plagiei ou apliquei qualquer forma de uso indevido de informações ou falsificação de resultados ao longo do processo que levou à sua elaboração.

Portanto, o trabalho apresentado neste documento é original e de minha autoria, não tendo sido utilizado anteriormente para nenhum outro fim.

Declaro ainda que tenho pleno conhecimento do Código de Conduta Ética do P.PORTO.

ISEP, Porto, 28 de setembro de 2025

Resumo

A transição energética na Europa e a crescente integração de fontes renováveis têm intensificado a volatilidade dos preços no mercado de eletricidade da Península Ibérica. Este cenário apresenta desafios para o equilíbrio entre a oferta e a procura, agravados pela intermitência associada à integração de energias renováveis. O problema central desta dissertação é desenvolver um sistema avançado de previsão de preços no mercado spot, que permita mitigar os impactos desta volatilidade, promovendo estabilidade e eficiência. Para atingir este objetivo, foi seguido um planeamento estruturado segundo a metodologia *Action Research*, incluindo uma revisão da literatura com PRISMA e a estruturação do processo de desenvolvimento com CRISP-DM. Esta abordagem permitiu identificar lacunas e avanços na aplicação de modelos de *Machine Learning* na transição energética e nos mercados de energia elétrica. Foram estudados e comparados modelos lineares, modelos baseados em árvores, modelos de *gradient boosting* e redes neuronais recorrentes. Os resultados evidenciaram a superioridade dos métodos de *boosting* na redução do erro absoluto, com o *XGBoost* a obter o melhor desempenho, seguido dos modelos de ensemble, superando as *RNNs*. As principais limitações encontradas relacionam-se com a janela temporal, a parametrização uniforme nas *RNNs* e a manutenção de *outliers*.

Palavras-chave: Mercado de Eletricidade, Transição energética, Volatilidade, Modelos de previsão, *Machine Learning*

Abstract

The energy transition in Europe and the growing integration of renewable sources have intensified price volatility in the Iberian Peninsula electricity market. This scenario presents challenges for balancing supply and demand, exacerbated by the intermittency associated with the integration of renewable energies. The central problem of this dissertation is to develop an advanced spot market price forecasting system that mitigates the impacts of this volatility, promoting stability and efficiency. To achieve this goal, a structured plan was followed according to the Action Research methodology, including a literature review with PRISMA and the structuring of the development process with CRISP-DM. This approach allowed us to identify gaps and advances in the application of Machine Learning models in the energy transition and electricity markets. Linear models, tree-based models, gradient boosting models, and recurrent neural networks were studied and compared. The results showed the superiority of boosting methods in reducing absolute error, with XGBoost achieving the best performance, followed by ensemble models, surpassing RNNs. The main limitations encountered relate to the time window, uniform parameterization in RNNs, and the maintenance of outliers.

Keywords: Electricity market, Energy transition, Volatility, Forecasting models, Machine Learning

Índice

1	Introdução	17
1.1	Contexto	17
1.2	Problema	19
1.3	Objetivos	19
1.4	Planeamento	20
1.5	Metodologia Seguida	21
1.6	Considerações Éticas	23
1.7	Estrutura do Documento	25
2	Estado da Arte	27
2.1	Machine Learning	27
2.1.1	Regressão Linear e Regularização Lasso e Ridge	27
2.1.2	Árvores de Decisão e Random Forest	30
2.1.3	Gradient Boosting Machines	32
2.1.4	Redes Neurais Recorrentes	37
2.1.5	Sistematização dos Modelos	39
2.2	Questões de Investigação	40
2.3	Critérios de elegibilidade	41
2.4	Fontes de informação	42
2.5	Seleção de estudos e processo de extração de dados	42
2.6	Discussão	43
2.6.1	Qual é o impacto direto da produção de energia renovável nas flutuações dos preços de eletricidade?	43
2.6.2	Como os dados históricos de produção de energia e procura têm sido usados para prever variações nos preços?	44
2.6.3	Quais são as estratégias atuais e potenciais para mitigar a volatilidade causada pela intermitência das energias renováveis?	45
2.6.4	Que papel a previsão meteorológica desempenha na gestão da volatilidade dos preços e nas operações do mercado elétrico?	47
2.6.5	Sistematização da Discussão	48
2.7	Sumário	49
3	Previsão de preços de eletricidade no mercado spot	51
3.1	Descrição do problema	51
3.2	Descrição da Metodologia a Seguir	52
3.3	Descrição do Dataset	53
3.4	Pré-processamento de Dados	55
3.4.1	Análise de dados	56

3.4.2	Tratamento de dados	63
4	Avaliação de Modelos	67
4.1	Afinação dos hiperparâmetros	67
4.2	Critérios de Avaliação	69
4.2.1	Erro Médio Absoluto (MAE)	69
4.2.2	Erro Quadrático Médio (MSE) e Raiz do Erro Quadrático Médio (RMSE).....	69
4.2.3	Coeficiente de determinação (R^2)	70
4.2.4	Erro Percentual Absoluto Médio (MAPE)	70
4.2.5	Resultados dos modelos	70
4.3	Discussão de resultados	73
5	Conclusões	77

Lista de Figuras

Figura 1 - Natureza periódica do uso energia renovável (Verdelho, 2024) . Error! Bookmark not defined.	
Figura 2 – Consumo primário de energia em Portugal (Verdelho, 2024)	18
Figura 3 - WBS por ciclo de vida.....	20
Figura 4 – Cronograma do projeto.....	20
Figura 5 - Estrutura <i>Action Research</i> (Western Sydney University, 2024).....	21
Figura 6 - Metodologia PRISMA (Moher et al., 2010).....	22
Figura 7 - Metodologia CRISP-DM (Chapman et al., 1999).....	23
Figura 8 – Exemplo Regressão Linear (Montgomery, D. C., Peck, E. A., & Vining, G. G., 2012). 28	
Figura 9 – Exemplo do comportamento de uma Árvore de Decisão (Data Courses, 2020)	30
Figura 10 - Exemplo do comportamento de uma Random Forest (Chaya, 2020)	31
Figura 11 – Representação do Gradient Boosting (<i>GeeksforGeeks</i> , 2025).....	33
Figura 12 – Representação do XGBoost (Öztornaci, s.d.)	34
Figura 13 – Algoritmo Leaf-wise do LightGBM (<i>LightGBM</i> , s.d.).....	35
Figura 14 – Aprendizagem do algoritmo CatBoost (CatBoost, 2018)	37
Figura 15 - Diagrama do PRISMA	43
Figura 16 – Ciclo diário médio de <i>RT_Demand</i>	58
Figura 17 - Ciclo semanal médio de <i>RT_LMP</i> e <i>System_Load</i> por dia da semana	59
Figura 18 - Médias mensais e móveis a 30 dias de <i>Dry_Bulb</i> e <i>RT_Demand</i>	60
Figura 19 – Matriz de Correlação.....	61
Figura 20 - Escala de correlação com <i>RT_LMP</i>	62
Figura 21 - Boxplots de Outliers.....	63
Figura 22 - RMSE por modelo	74
Figura 23 – R ² por modelo.....	74
Figura 24 – Previsão com os valores de Teste	74
Figura 25 – Curva de Aprendizagem do <i>XGBoost</i>	75
Figura 26 – Resíduos da previsão com os valores de Teste	75

Lista de Tabelas

Tabela 1 – Detalhes do Planeamento	21
Tabela 2 – Comparação ética da União Europeia com Estados Unidos	24
Tabela 3 – Comparação Árvore de Decisão com Random Forest	32
Tabela 4 – Comparação LSTM e GRU	38
Tabela 5 – Sistematização de Técnicas de ML	40
Tabela 6 – Critérios de Inclusão	41
Tabela 7 – Critérios de Exclusão.....	42
Tabela 8 – Termos de Pesquisa	42
Tabela 9 – Sistematização da discussão.....	48
Tabela 10 – Resultados Análise estatística	57
Tabela 11 - Médias semanais do <i>System_Load</i> e do <i>RT_LMP</i>	59
Tabela 12 – Número de <i>Outliers</i> por Variável.....	64
Tabela 13 – Parâmetros utilizados pelo <i>GridSearch</i>	68
Tabela 14 - Resultados das parametrizações base	71
Tabela 15 - Resultados das Redes Neurais	71
Tabela 16 – Resultados dos testes de hipótese	72

Acrónimos e Símbolos

Lista de Acrónimos

AI Act	Regulamento sobre Inteligência Artificial da União Europeia
AI RMF	AI Risk Management Framework
CRISP-DM	Cross-Industry Standard Process for Data Mining
EFB	Exclusive Feature Bundling
ERSE	Entidade Reguladora dos Serviços Energéticos
GOSS	Gradient-based One-Side Sampling
GRU	Gated Recurrent Units
IEEE	Institute of Electrical and Electronics Engineers
LSTM	Long Short-Term Memory
MAE	Erro Médio Absoluto
MAPE	Erro Percentual Absoluto Médio
ML	Machine Learning
MSE	Erro Quadrático Médio
NEMO	Nominated Electricity Market Operators
NIST	National Institute of Standards and Technology
OMIE	Operador do Mercado Ibérico de Eletricidade
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RGPD	Regulamento Geral sobre a Proteção de Dados
RMSE	Raiz do Erro Quadrático Médio
RNNs	Redes Neurais Recorrentes
SVMs	Máquinas de Vetores de Suporte

1 Introdução

Este capítulo descreve o contexto, a definição do problema a resolver, os principais objetivos e opções metodológicas a seguir no âmbito desta dissertação de mestrado. Apresenta ainda o planeamento do trabalho a realizar e as questões éticas. Por último, apresenta uma visão da estrutura global do documento.

1.1 Contexto

A transição energética na Europa tem sido acompanhada por uma importante reformulação de estruturas e políticas de sustentabilidade e eficiência no sector elétrico. Ao mesmo tempo, os mercados energéticos marginais transfronteiriços caracterizam a integração energética na progressão do objetivo da redução de emissões (European Commission, 2024b). Neste sentido, vale a pena sublinhar que a sigla OMIE corresponde à denominação do Operador do Mercado Ibérico de Eletricidade, que, com a nova designação de *Nominated Electricity Market Operators*, NEMO, continua a manter sob a sua responsabilidade os mercados diário e intradiário da Península Ibérica (OMIE, 2023).

Uma das características mais imprevisíveis do mercado da eletricidade é a volatilidade dos preços, especialmente com a crescente inserção de fontes de energia renováveis. Estas fontes, embora muito benéficas para o ambiente e para a sustentabilidade a longo prazo, criam instabilidade nos preços de eletricidade devido à sua natureza periódica (IEA, 2021). A Figura 1 apresenta o consumo de energias renováveis em 2021. O OMIE pretende enfrentar o desafio de gerir eficazmente estas flutuações de preços, com a garantia de que a estabilidade e a sustentabilidade energética são mantidas na região.



Figura 1 - Natureza periódica do uso energia renovável (Verdelho, 2024)

Em Portugal, a Entidade Reguladora dos Serviços Energéticos, ERSE, é responsável pela regulação e monitorização do setor elétrico. Assim, estabelece normas e diretrizes que incentivam a integração de energias renováveis, enquanto procura diminuir o impacto da volatilidade nos preços de eletricidade (ERSE, 2025). A utilização das energias renováveis é uma realidade cada vez mais presente em Portugal, como representado na Figura 2 (Verdelho, 2024).

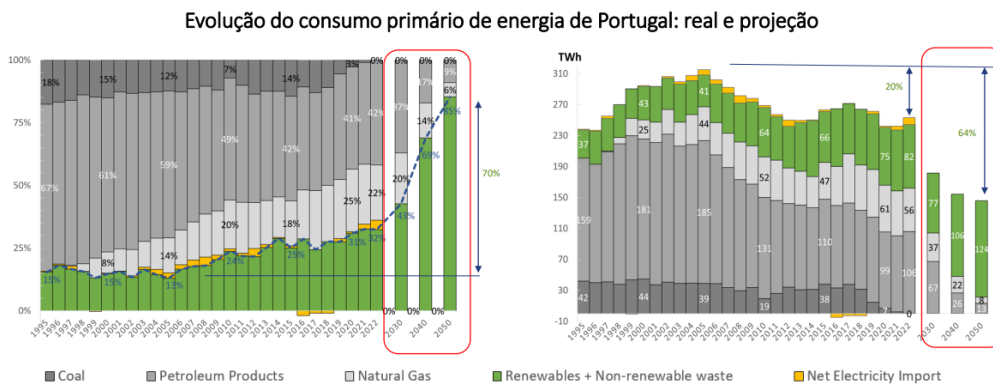


Figura 2 – Consumo primário de energia em Portugal (Verdelho, 2024)

Uma vez que as ações para mitigar os efeitos da volatilidade dos preços têm de ser imediatas para evitar impactos adversos tanto para os consumidores como para os produtores de energia, o desenvolvimento de ferramentas eficientes para a previsão e mitigação dos riscos relacionados com estas flutuações é de principal importância (ERSE, 2023).

Nesta perspetiva, o presente trabalho centrar-se-á no desenvolvimento de um sistema avançado de análise e previsão de preços de eletricidade. Pretende-se que o sistema preveja,

com maior precisão, desvios nos preços e ofereça soluções capazes de aumentar a eficiência e a resiliência do mercado de eletricidade.

1.2 Problema

Na gestão diária e intradiária dos mercados de eletricidade da Península Ibérica, o OMIE enfrenta um desafio crítico decorrente da volatilidade dos preços de eletricidade. Esta volatilidade é aumentada pela crescente inserção de fontes de energia renováveis, cuja produção é altamente dependente de condições meteorológicas variáveis. À medida que o mercado se torna mais dependente de energias renováveis, o OMIE acumula uma quantidade significativa de dados sobre a produção de energia, a procura, e flutuações de preços em tempo real. Estes dados são de grande valor para o OMIE, o que facilita a criação de uma linha temporal para a evolução histórica das condições de mercado e de preços (OMIE, 2023; ERSE, 2023). Assim, estes dados permitem gerar um historial detalhado das condições de mercado e dos padrões de preços. No entanto, o carácter intermitente das fontes renováveis introduz uma complexidade adicional na previsão de preços, o que dificulta a tarefa de equilibrar eficazmente a oferta e a procura. Neste contexto, o desenvolvimento de um sistema avançado de previsão de preços não é apenas uma necessidade operacional, mas uma exigência estratégica para manter a resiliência do mercado elétrico (ERSE, 2023).

1.3 Objetivos

Face ao problema identificado da volatilidade dos preços no mercado elétrico, e à necessidade urgente de uma melhor gestão dessa volatilidade, este trabalho tem como objetivo o desenvolvimento de um sistema de análise e previsão de preços no mercado spot de eletricidade. A implementação deste sistema visa não só reduzir o impacto das flutuações do mercado, mas também, alavancar a sustentabilidade e estabilidade energética na Península Ibérica alinhada com as diretrizes regulatórias europeias até 2030 (European Commission, 2024b; ERSE, 2024). Para atingir este objetivo principal, este trabalho de mestrado foi estruturado em sete fases, cada uma projetada para abordar diferentes aspetos para a criação de um sistema robusto e eficaz, sendo elas:

1. Revisão da Literatura sobre *Machine Learning* e suas Aplicações no Mercado de Eletricidade;
2. Realização de uma revisão sistemática para melhor compreensão do estudo do mercado elétrico;
3. Modelar o sistema a desenvolver para o problema de previsão de preços de eletricidade;
4. Desenvolvimento de modelos preditivos baseados em *Machine Learning*;
5. Análise, validação e otimização dos modelos;
6. Implementar o protótipo do sistema de análise e previsão de preços;

7. Estudo computacional do protótipo;
8. Escrita da dissertação a documentar todos os desenvolvimentos.

Esta estruturação garante uma abordagem metodológica e abrangente, assegurando que todos os objetivos sejam abordados e que os resultados sejam documentados de forma completa e acessível.

1.4 Planeamento

No planeamento do projeto, foi definido que este seria dividido em sete fases, sendo estas as seguintes: revisão da literatura, realização de uma revisão sistemática da literatura, modelação do sistema, desenvolvimento de modelos preditivos, análise, validação e otimização dos modelos, implementação do protótipo do sistema e o estudo computacional do protótipo. A escrita deste relatório foi um processo contínuo durante todas estas sete fases. Cada uma destas fases foi um marco para o sucesso do projeto, como está representado na Figura 3 e na Figura 4. Na Tabela 1 estão representados os detalhes de cada fase do planeamento.

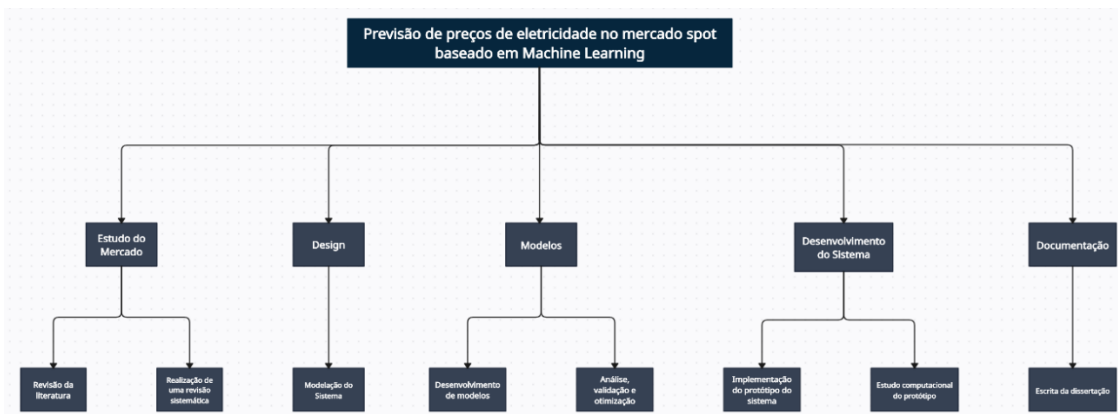


Figura 3 - WBS por ciclo de vida

CRONOGRAMA



Figura 4 – Cronograma do projeto

Tabela 1 – Detalhes do Planeamento

Planeamento	Data	Descrição
Revisão da literatura	01/nov - 04/jan	Estudo sobre a aplicação de <i>Machine Learning</i> no mercado da eletricidade para suportar a revisão sistemática
Realização de uma revisão sistemática	01/nov – 04/jan	Revisão sistemática com o objetivo de uma melhor compreensão do mercado elétrico
Modelação do sistema	08/jan – 10/fev	Desenho do sistema a desenvolver.
Desenvolvimento de modelos	06/fev – 15/mar	Construção os modelos de previsão para solucionar o problema
Análise, validação e otimização	05/mar – 15/abr	Avaliação dos modelos com objetivos de otimização e precisão
Implementação do protótipo do sistema	20/mar – 10 mai	Desenvolvimento do protótipo do sistema através dos modelos criados
Estudo computacional do protótipo	05/mai – 20/jun	Estudo computacional sobre o protótipo criado.
Escrita da dissertação	01/nov – 30/jun	Documentação detalhada do projeto

1.5 Metodologia Seguida

A metodologia utilizada neste trabalho combina as abordagens do *Action Research* como estrutura ampla de orientação, demonstrado na Figura 5, do *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) no suporte à revisão da literatura e do *Cross-Industry Standard Process for Data Mining* (CRISP-DM) no suporte ao processo de extração do conhecimento.



Figura 5 - Estrutura *Action Research* (Western Sydney University, 2024)

A aplicação da metodologia PRISMA (Page et al., 2021) está direcionada para a revisão sistemática da literatura, fundamental para estabelecer o contexto teórico e identificar lacunas de pesquisa na área de previsão de preços de eletricidade usando *Machine Learning*. Conforme as diretrizes do PRISMA, foi realizada uma seleção criteriosa de estudos, garantindo que a revisão fosse tanto abrangente quanto transparente, como demonstrado na Figura 6. Isto permite uma documentação clara de todos os estudos considerados, facilitando a sua replicação e validação. Assim, abordagem metodológica do PRISMA ajuda a alinhar este projeto com as exigências de transparência e reprodutibilidade, uma prática recomendada na compilação e análise de contribuições científicas (Sohrabi et al., 2021).

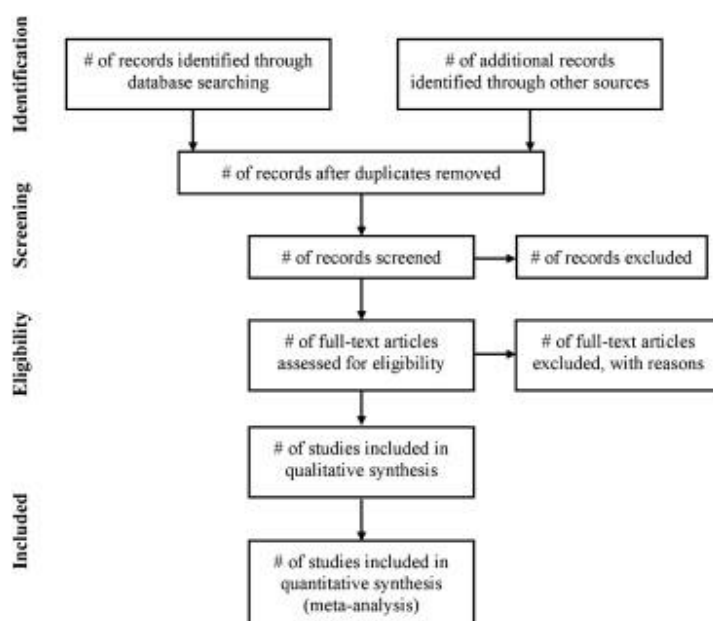


Figura 6 - Metodologia PRISMA (Moher et al., 2010)

Paralelamente, o CRISP-DM (Wirth & Hipp, 2000) será utilizado para estruturar o processo de mineração de dados. Este modelo é aplicado para documentar e orientar as fases do desenvolvimento do sistema de previsão de preços, desde a compreensão dos dados até a implementação do modelo final, como demonstrado na Figura 7. O uso do CRISP-DM permite uma gestão do ciclo de vida do desenvolvimento do sistema proposto, garantindo que cada fase seja claramente reportada e que as transições entre as fases do projeto sejam avaliadas e validadas (Chapman et al., 1999). A implementação deste modelo assegura que a análise de dados seja realizada de forma sistemática e que os resultados possam ser diretamente aplicados para melhorar a precisão das previsões de preços no mercado de eletricidade.

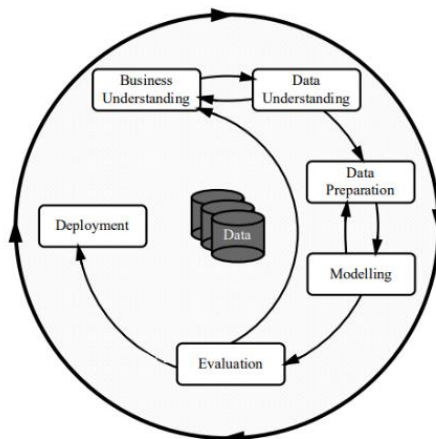


Figura 7 - Metodologia CRISP-DM (Chapman et al., 1999)

A metodologia *Action Research* com a combinação das metodologias PRISMA e CRISP-DM enriquece este estudo ao proporcionar uma estrutura de orientação, uma base teórica sólida e um método prático robusto para a análise de dados. Enquanto o PRISMA orienta a estruturação da revisão sistemática, garantindo a cobertura abrangente e imparcial do estado da arte, o CRISP-DM estrutura o processo de descoberta de conhecimento, desde a compreensão do negócio até a avaliação do modelo desenvolvido. Esta abordagem não só aprofunda a compreensão teórica do problema, como também melhora a aplicabilidade dos desenvolvimentos práticos.

1.6 Considerações Éticas

Este projeto aborda conscientemente os potenciais problemas éticos desde o início do seu desenvolvimento. No que diz respeito à privacidade, todos os dados utilizados são considerados dados indiretos, não podendo ser utilizados para identificar indivíduos, como definido pelo Regulamento Geral sobre a Proteção de Dados (RGPD), especialmente os Art.º 5 e 32, que exigem que os dados pessoais sejam processados de forma legal, justa e transparente, e que medidas técnicas e organizativas adequadas sejam implementadas para garantir a segurança dos dados (RGPD, 2016). A transparência é assegurada com a documentação altamente detalhada dos algoritmos utilizados, permitindo revisões e auditorias independentes. Além disso, a análise e a otimização de modelos preditivos são realizadas em conformidade com os princípios do Art.º 25, que exige a proteção de dados desde a concepção e por defeito, garantindo assim que o sistema desenvolvido não pratica enviesamento. A conformidade com o Regulamento sobre Inteligência Artificial da União Europeia (AI Act), publicado em 2024, é uma consideração crucial, especialmente porque este regulamento visa garantir que os sistemas de inteligência artificial sejam seguros e os direitos fundamentais sejam protegidos, fornecendo assim resultados imparciais (European Commission, 2024a). O AI Act defende a importância de utilizar dados de treino precisos, representativos e livres de viés, conforme o

Art.º 14, garantindo que as previsões sejam justas e não perpetuem desigualdades. O Art.º 13 defende que os sistemas de inteligência artificial sejam explicáveis para os *stakeholders*, para que possam compreender decisões tomadas.

No entanto, os Estados Unidos seguem uma abordagem mais liberal e orientada pelo mercado, onde a regulação é fragmentada, variando entre estados e setores, e onde a ênfase recai sobre a inovação e competitividade económica (National Institute of Standards and Technology, 2023). Em vez de um regulamento abrangente e vinculativo como o AI Act, as diretrizes federais, como o AI Risk Management Framework (AI RMF) do National Institute of Standards and Technology (NIST), são de carácter voluntário, funcionando como orientações de boas práticas e não como obrigações legais. Esta filosofia regulatória favorece ciclos de inovação mais rápidos, mas levanta questões quanto à consistência na proteção de dados e à prevenção de desvios e danos éticos. A Tabela 2 apresenta uma comparação entre as abordagens da União Europeia e dos Estados Unidos da América relativamente a esta regulação ética.

Tabela 2 – Comparação ética da União Europeia com Estados Unidos

Abordagens	União Europeia	Estados Unidos
Estratégia reguladora	Foco principal na proteção de direitos fundamentais e prevenção de riscos.	Liberal, com foco na inovação e autorregulação pelo mercado.
Base legal	RGPD (2016) e AI Act (2025), obrigatórios e uniformizados para todos os Estados-Membros.	Diretrizes voluntárias (como o NIST AI RMF) e legislação estadual/setorial.
Proteção de dados	Elevada, com segurança e privacidade por defeito, desde a sua criação (Art.º 5, 25 e 32 RGPD).	Fragmentada, sem equivalente federal ao RGPD.
Garantia de imparcialidade	Obrigatória (AI Act, Art.º 14).	Recomendação não vinculativa.
Explicabilidade	Obrigatória (AI Act, Art.º 13).	Incentivada, mas não obrigatória.
Penalizações	Multas até 6 % do volume de negócios global ou €30 milhões (European Commission, 2024a).	Variáveis por estado/setor; geralmente inferiores.
Impacto na inovação	Ciclos de adoção mais lentos devido a requisitos de conformidade rigorosos.	Ciclos de adoção mais rápidos devido a menor carga regulatória.

Observa-se que a estratégia reguladora da União Europeia segue um paradigma preventivo, centrado na proteção de direitos fundamentais e na prevenção de riscos, enquanto a abordagem dos Estados Unidos é mais liberal, privilegiando a inovação e a autorregulação pelo mercado. Como base legal, a União Europeia adota instrumentos vinculativos e uniformizados entre Estados-Membros, nomeadamente o RGPD (RGPD, 2016) e o AI ACT (AI ACT, 2025), que estabelecem requisitos obrigatórios de conformidade. Em contraste, os Estados Unidos dispõem de um enquadramento fragmentado, baseado em diretrizes voluntárias, como o NIST AI RMF (AI RMF, 2023), e legislação estadual ou setorial. A proteção de dados na União Europeia é elevada, sendo exigida segurança e privacidade por defeito e desde a conceção, conforme disposto nos Art.º 5, 25 e 32 do RGPD. Já nos Estados Unidos, a ausência de uma lei federal equivalente ao RGPD conduz a uma proteção menos uniforme e dependente de regulamentações locais. Relativamente a assegurar imparcialidade no modelo, esta é obrigatória na União Europeia (AI Act, Art.º 14), enquanto nos Estados Unidos permanece como uma recomendação não vinculativa. A explicabilidade, prevista no Art.º 13 do AI Act, é igualmente uma exigência obrigatória na União Europeia, garantindo que as decisões dos sistemas de AI sejam compreensíveis para os *stakeholders*. Nos Estados Unidos, esta prática é incentivada, mas não imposta por lei federal. Em termos de penalizações, a União Europeia estabelece multas significativas para incumprimentos, até 6 % do volume de negócios global ou €30 milhões (European Commission, 2024a). Nos Estados Unidos, as sanções variam de acordo com o estado ou setor, sendo geralmente inferiores. Com isto, estas diferenças impactam diretamente a inovação. Na União Europeia, os ciclos de adoção à AI tendem a ser mais lentos devido aos requisitos de conformidade rigorosos, no entanto, nos Estados Unidos, a menor carga regulatória permite ciclos de implementação mais rápidos, mesmo com potenciais riscos acrescidos no que diz respeito à proteção de dados e aos princípios éticos.

Além disto, o código de conduta do Institute of Electrical and Electronics Engineers, IEEE, é igualmente relevante. Princípios do IEEE (IEEE Code of Ethics, 2020), como honestidade, integridade, manutenção da transparência e veracidade ao longo deste trabalho, bem como competência profissional, asseguram que o trabalho apresentado (IEEE Code of Ethics, 2020). Estes critérios reforçam, para além da qualidade, a credibilidade e a validade científica das contribuições deste estudo.

1.7 Estrutura do Documento

Este documento está estruturado em cinco principais capítulos fundamentais que compõem esta dissertação, em que cada um, descreve um aspeto diferente do trabalho desenvolvido.

No Capítulo 1, a Introdução, é descrita a contextualização na área da transição energética e dos mercados de energia elétrica, assim como a definição do problema sobre a volatilidade dos preços de eletricidade. Apresenta também os objetivos desta dissertação de mestrado e a metodologia para o seu desenvolvimento, destacando os benefícios da utilização de uma abordagem em duas fases para a revisão da literatura e o desenvolvimento.

No Capítulo 2, o Estado da Arte apresenta a revisão sistemática realizada sobre o mercado de energias e os seus problemas de instabilidade de preços e explora como diferentes métodos de Machine Learning podem contribuir para a resolução do problema. A estrutura segue os passos pré-requeridos para a realização da pesquisa bibliográfica de acordo com a metodologia PRISMA, seguida de uma discussão dos resultados apresentados conforme as questões de investigação alinhadas com a definição do problema e dos objetivos.

No Capítulo 3, Previsão de preços de eletricidade no mercado spot, apresenta-se detalhadamente o problema em estudo, a metodologia aplicada, a descrição do dataset utilizado e o pré-processamento efetuado. Este pré-processamento inclui uma análise exploratória dos dados, com estatísticas, sazonalidade, tendências e correlação entre variáveis, bem como o tratamento de valores em falta, normalização, criação de variáveis derivadas e seleção de atributos relevantes.

O Capítulo 4, Avaliação de Modelos, descreve a afinação dos hiperparâmetros dos modelos, a escolha dos critérios de avaliação e a discussão dos resultados. São apresentadas métricas de erro, procedimentos de validação e discussão comparativa dos resultados obtidos, salientando pontos fortes e limitações no contexto de desempenho de modelos e do mercado elétrico.

Por último, no Capítulo 5, Conclusões, é realizado um resumo dos capítulos desenvolvidos. Este resumo, através da Introdução, apresenta uma visão clara da proposta e, do Estado da Arte, um estudo aprofundado sobre os mercados de energias e os preços de eletricidade. Adicionalmente, menciona os resultados obtidos, conclusões da discussão destes resultados, limitações encontradas e futuras melhorias.

2 Estado da Arte

Neste capítulo, é realizada a contextualização dos temas e áreas científicas desta proposta. Para o efeito, a seguinte secção expõe como e quais métodos de Machine Learning poderiam ajudar a encontrar uma solução para o problema. Para além disso, é utilizada a metodologia PRISMA, revisão sistemática esta que permite uma análise rigorosa e estruturada da literatura relevante, assegurando a inclusão de estudos pertinentes e a exclusão de viés. De seguida, é realizada uma descrição dos resultados obtidos e uma discussão detalhada dos tópicos comuns e distintos encontrados. Esta análise não só destaca as principais tendências e desafios enfrentados pelo setor, como também cria uma base de conhecimento para suportar futuras decisões no desenvolvimento de soluções inovadoras para a gestão e previsão dos preços de eletricidade.

2.1 Machine Learning

Nesta secção, procede-se a apresentação das principais abordagens de Machine Learning utilizadas na previsão de preços de eletricidade, com especial ênfase nos modelos supervisionados aplicáveis ao contexto do mercado spot. Através de uma análise na literatura, são descritas as características de cada modelo, os respetivos mecanismos de aprendizagem, assim como as suas vantagens e limitações, permitindo não só compreender o enquadramento tecnológico da previsão de preços, como também identificar as abordagens com maior robustez, escalabilidade e capacidade de generalização.

2.1.1 Regressão Linear e Regularização Lasso e Ridge

A *Regressão Linear* foi uma das primeiras abordagens utilizadas para a previsão de preços de eletricidade, devido à sua simplicidade e capacidade de interpretação (Weron, 2014).

Este modelo assume uma relação linear entre as variáveis independentes e a variável dependente, como demonstrado na equação seguinte,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (1)$$

onde Y representa a variável dependente e X a variável independente, o que nem sempre captura as complexidades do mercado elétrico. Para uma maior compreensão, o parâmetro β_0 representa o valor esperado de Y quando $X = 0$, sendo designado por intercepto, enquanto β_1 representa a variação esperada de Y por unidade de variação em X . O termo ε_i representa o erro, que tem como objetivo agregar todas as influências não modeladas pelo modelo linear, como variações aleatórias ou fatores não observados (Weron, 2014). Este modelo pode ser representado como consta na Figura 8, onde se pode encontrar um exemplo da relação entre o *Delivery Volume* e o *Delivery Time* presente no artigo “*Introduction to Linear Regression Analysis*” de Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012).

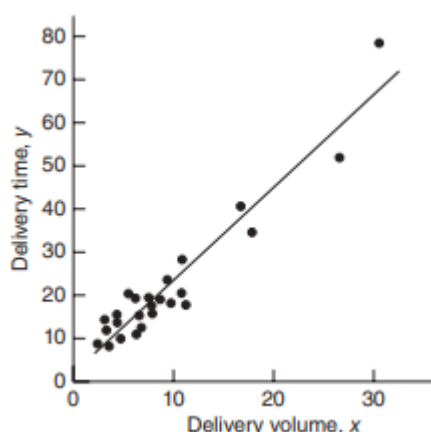


Figura 8 – Exemplo Regressão Linear (Montgomery, D. C., Peck, E. A., & Vining, G. G., 2012)

Para lidar com problemas de multicolinearidade e *overfitting*, foram desenvolvidas variantes como a *Regressão de Lasso* e *Ridge*, que introduzem termos de penalização no processo de ajuste do modelo (Tibshirani, 1996). Apesar das suas limitações na modelação de padrões não-lineares, estas técnicas continuam a ser uma referência inicial devido à sua robustez em quantidades de dados reduzidas e à elevada interpretabilidade dos resultados (Tibshirani, 1996).

A **Regressão de Lasso** (*Least Absolute Shrinkage and Selection Operator*), também conhecida como Penalização L1, proposta por Tibshirani (1996), é uma técnica de regularização que visa resolver algumas limitações da regressão linear tradicional, nomeadamente a multicolinearidade e o *overfitting*, sobretudo em contextos onde o número de variáveis explicativas é elevado ou há redundância entre elas. Numa *Regressão Linear* sem regularizações, o objetivo é encontrar os coeficientes β que minimizam o erro quadrático total, definido pela soma dos quadrados das diferenças entre os valores observados e os valores ajustados. Este erro é representado pela seguinte equação:

$$\beta = \arg \min_{\beta} \left\{ \sum_{i=1}^n (Y_i - X_i \beta)^2 \right\} \quad (2)$$

No entanto, para a *Regressão de Lasso* altera esta função de erro e adiciona-se um termo de penalização, com o objetivo dos coeficientes terem um valor menor e, eventualmente, tornarem-se zero, como demonstrado na equação (3).

$$\beta = \arg \min_{\beta} \left\{ \sum_{i=1}^n (Y_i - X_i \beta)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\} \quad (3)$$

Nesta equação, o *lambda*, λ , tem como função controlar o equilíbrio da regularização. Esta penalização afeta diretamente o viés e a variância, tendo como objetivo evitar o *overfitting* e melhorar a generalização do modelo.

A ***Regressão de Ridge***, também conhecida como Penalização L2, proposta por Hoerl e Kennard (1970) é uma técnica de regularização que tem como objetivo principal combater a multicolinearidade e o *overfitting*, assim como a regularização de *Lasso* (Hoerl, A. E., & Kennard, R. W., 1970).

A *Regressão de Ridge* introduz um termo de penalização quadrático sobre os coeficientes, com o objetivo de estabilizar o modelo e controlar a sua complexidade, onde a nova função passa a ser:

$$\beta = \arg \min_{\beta} \left\{ \sum_{i=1}^n (Y_i - X_i \beta)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\} \quad (4)$$

O termo adicional $\lambda \sum_{j=1}^p \beta_j^2$ representa penalização L2 e atua como uma força que reduz os coeficientes no sentido do zero, sem os anular completamente.

Com isto, conclui-se que enquanto a Regularização de *Ridge* apresenta vantagens quando todas as variáveis são potencialmente úteis, a regularização de *Lasso* destaca-se quando se acredita que apenas algumas variáveis são verdadeiramente relevantes. Contudo, em situações onde se pretende beneficiar simultaneamente da estabilidade da Penalização L2 e da exclusão da Penalização L1, uma abordagem híbrida torna-se desejável. Para tal, surge o *Elastic Net* (Zou, H., & Hastie, T., 2005).

O ***Elastic Net***, criado por Hui Zou e Trevor Hastie, é uma técnica que combina as vantagens dos métodos de regularização, proporcionando uma solução equilibrada (Zou, H., & Hastie, T., 2005). Esta abordagem visa acrescentar as duas penalizações à função dos coeficientes β , como demonstra a equação seguinte.

$$\beta = \operatorname{arg\,min}_{\beta} \left\{ \sum_{i=1}^n (Y_i - X_i\beta)^2 + \lambda \sum_{j=1}^p |\beta_j| + \lambda \sum_{j=1}^p \beta_j^2 \right\} \quad (5)$$

Assim, o Elastic Net surge como uma alternativa flexível, robusta e versátil para situações onde se pretende simultaneamente regularizar, selecionar variáveis e estabilizar o modelo com colinearidade.

2.1.2 Árvores de Decisão e Random Forest

As **Árvores de Decisão** são modelos de previsão que estruturam os dados de forma hierárquica, criando regras de decisão baseadas nos atributos dos dados. Embora simples, tendem a sofrer uma alta variância e *overfitting* (Breiman, 2001). No entanto, estas são amplamente valorizadas pela sua interpretação intuitiva e capacidade de modelar relações não lineares entre variáveis. Para tal, contêm uma estrutura hierárquica em forma de árvore, composta por diferentes tipos de nós e ligações. Esta estrutura permite representar, de forma sequencial e lógica, o processo de tomada de decisão com base nas variáveis explicativas. A árvore inicia-se no nó raiz, onde se aplica a primeira divisão com base numa condição sobre uma variável explicativa. A condição poderá ser, por exemplo “ $X > Y$ ”, onde X e Y são variáveis do problema em questão. O conjunto de dados é então dividido em dois subconjuntos conforme o resultado do teste, onde é avaliado conforme o menor erro quadrático médio (MSE), e este processo repete-se recursivamente em nós internos, originando ramos que conduzem a novas divisões, como representado na Figura 9. O objetivo em cada etapa da construção de uma árvore de decisão é escolher a variável explicativa e o respetivo ponto de corte que melhor dividem os dados. A qualidade da divisão é medida pela variância da variável dependente nos subconjuntos criados (Mienye & Jere, 2024). Este processo recursivo termina quando forem satisfeitas as condições de paragem definidas, como profundidade máxima ou número mínimo de observações por nó.

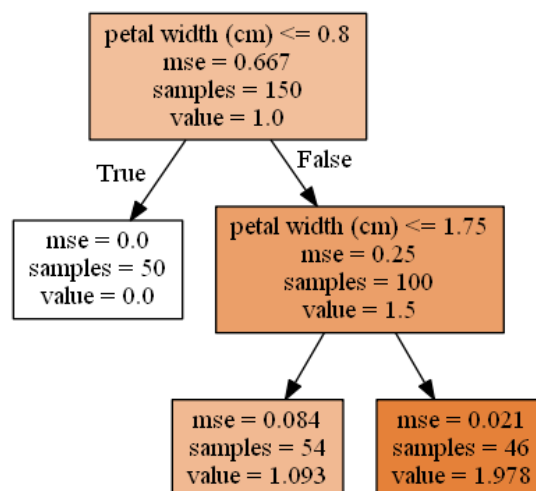


Figura 9 – Exemplo do comportamento de uma Árvore de Decisão (Data Courses, 2020)

O **Random Forest**, introduzido por Breiman, supera estas limitações ao combinar múltiplas árvores de decisão treinadas em subconjuntos aleatórios dos dados, aumentando assim a consistência e a capacidade de previsão. Esta técnica tem mostrado resultados na previsão de preços de eletricidade, especialmente em cenários com grande número de variáveis explicativas e interações complexas (Lago et al., 2021).

O processo de construção de um modelo *Random Forest* começa pela criação de várias amostras de treino por *bootstrap*, isto é, subconjuntos extraídos aleatoriamente com reposição a partir do conjunto de treino original. Cada árvore é construída de forma independente com base nestas amostras, como demonstrado na Figura 10. Em cada divisão dentro de cada árvore, o algoritmo seleciona aleatoriamente apenas um subconjunto de variáveis explicativas como candidatas para a divisão, promovendo a diversidade estrutural entre as árvores e diminuindo a correlação entre elas. A agregação das previsões individuais é feita através do cálculo da média das previsões numéricas geradas por todas as árvores da floresta (Breiman, 2001).

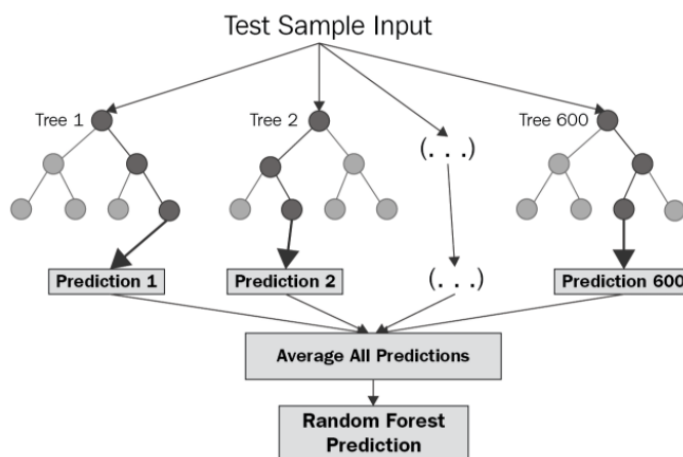


Figura 10 - Exemplo do comportamento de uma Random Forest (Chaya, 2020)

Devido a isto, apresenta uma elevada resistência ao *overfitting*, característica comum em árvores de decisão isoladas, sobretudo quando muito profundas. Por outro lado, devido à natureza aleatória e ao processo de agregação, o modelo tende a ter baixa variância, mantendo ao mesmo tempo um viés relativamente reduzido. Isto torna o *Random Forest* particularmente preferencial em contextos com relações não lineares e interações complexas entre variáveis, mesmo na presença de ruído ou **outliers** (Louppe, 2014). Na Tabela 3 está representada uma sistematização da comparação destes modelos.

Tabela 3 – Comparação Árvore de Decisão com Random Forest

Modelo	Árvore de Decisão	Random Forest
Tipo de modelo	Modelo único, determinístico	Conjunto de modelos com aleatoriedade
Interpretação	Alta (visualizável, explicável)	Baixa (difícil de interpretar individualmente)
Variância	Alta (muito sensível aos dados)	Baixa (reduzida pela média de várias árvores)
Overfitting	Elevado, especialmente em árvores profundas	Reduzido, devido à agregação e aleatoriedade
Robustez	Baixa (pode variar muito com pequenas alterações)	Alta (estável e consistente)
Velocidade de treino	Rápida	Mais lenta (devido à construção de múltiplas árvores)

2.1.3 Gradient Boosting Machines

O **Gradient Boosting** é uma técnica de *ensemble* que constrói modelos aditivos de forma sequencial, corrigindo os erros dos modelos anteriores. Bibliotecas como **XGBoost** (Chen & Guestrin, 2016), **LightGBM** (Ke et al., 2017) e **CatBoost** (Prokhorenkova et al., 2018) melhoraram significativamente a eficiência e o desempenho deste método. Em estudos de previsão de preços de eletricidade, estas técnicas têm-se destacado pela sua capacidade de lidar com dados tabulares heterogêneos e de capturar interações complexas (Singh et al., 2021).

Esta abordagem foi introduzida por Friedman (2001) como uma generalização dos métodos de *boosting*, propondo uma estrutura que utiliza a descida do gradiente como base para ajustar modelos sucessivos a partir dos erros residuais acumulados. Ao contrário de técnicas como o *bagging*, que combinam modelos de forma paralela e independente, o *Gradient Boosting* constrói os modelos de forma sequencial, sendo cada um treinado com o objetivo de corrigir os erros cometidos pelos anteriores. Assim, o *Gradient Boosting* visa encontrar uma função $F(x)$ que minimize uma função de perda $L(y, F(x))$, onde y é a variável independente e x o vetor de variáveis explicativas. O processo inicia-se com uma função inicial $F_0(x)$, frequentemente definida como a constante que minimiza L no conjunto de treino. Com isto, através da equação,

$$F_m(x) = F_{m-1}(x) + \beta_m h(x; a_m) \quad (6)$$

onde $F_{m-1}(x)$ representa o modelo anterior, β_m representa a taxa de aprendizagem, que é responsável por controlar a contribuição de cada modelo na construção do modelo final e $h(x; a_m)$ a nova função base ajustada aos resíduos através do a_m (Friedman, J. H., 2001).

A versatilidade do *Gradient Boosting* reside no facto de poder ser utilizado com qualquer tipo de modelo que este seja capaz de aproximar os gradientes da função de perda. Entre os exemplos possíveis encontram-se modelos lineares, *splines*, métodos baseados em *kernels*, ou até modelos probabilísticos simples. A escolha da função de perda é igualmente flexível, podendo ser adaptada ao problema em questão. Na Figura 11 está representado um exemplo do *Gradient Boosting* aplicado a Árvores de Decisão.

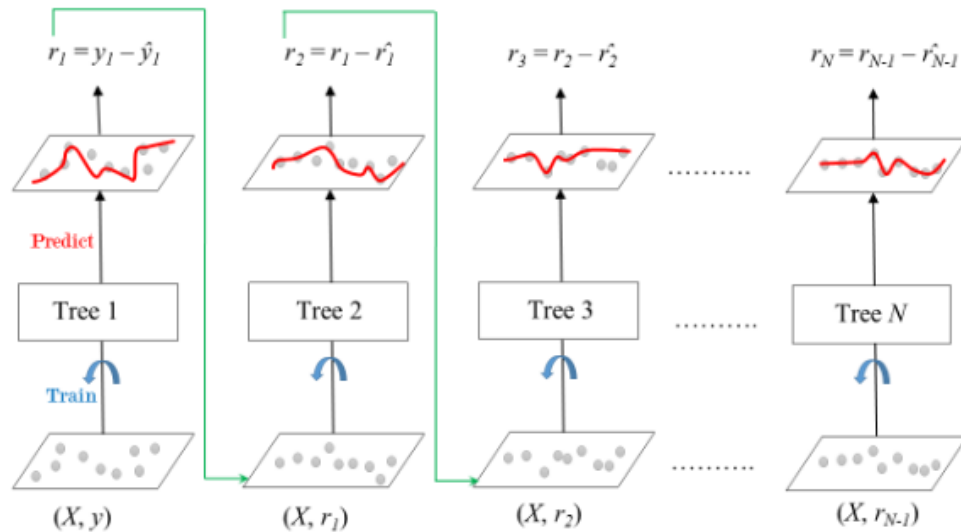


Figura 11 – Representação do Gradient Boosting (GeeksforGeeks, 2025)

2.1.3.1 XGBoost

O *Extreme Gradient Boosting*, conhecido por *XGBoost*, é uma implementação escalável do algoritmo de *Gradient Boosting* que utiliza exclusivamente árvores de decisão, proposto por Chen e Guestrin (2016). Desenvolvido com o intuito de aumentar o desempenho computacional e a consistência do *boosting* clássico, o *XGBoost* tornou-se um benefício para a área do *machine learning* e aplicações reais, particularmente quando se trata de problemas com dados estruturados e tabulares (Chen & Guestrin, 2016).

O *XGBoost* baseia-se na mesma lógica iterativa que o *Gradient Boosting*, através da construção sequencial de modelos, onde cada modelo novo ajusta-se aos resíduos do anterior, com vista à minimização de uma função de perda cumulativa. Contudo, a principal inovação introduzida pelo *XGBoost* consiste na inclusão de um termo de regularização explícito na função objetivo, que penaliza a complexidade do modelo, controlando tanto o número de folhas como os valores atribuídos a cada folha da árvore. Esta regularização é representada através do seguinte termo:

$$\Omega(f) = YT + \frac{1}{2}\lambda\|\omega\|^2 \quad (6)$$

onde T representa o número de folhas, ω representa os pesos atribuídos às folhas e, Y e λ são parâmetros de regularização ajustáveis. A função objetivo passa, assim, a ser composta por dois termos: o primeiro corresponde à função de perda cumulativa sobre o conjunto de treino, e o segundo é o termo de regularização $\Omega(f)$, que visa evitar o *overfitting* e promover a generalização. Outra inovação do *XGBoost* é a utilização de uma expansão de Taylor de segunda ordem para aproximar a função de perda, incorporando tanto o gradiente, primeira derivada, como a matriz de *Hessiano*, segunda derivada, em cada iteração de treino. Esta matriz de *Hessiano* é responsável por descrever a curvatura da função (Chen & Guestrin, 2016; Friedman, J.H., 2001).

Assim, através destas novas funções é permitida uma otimização mais rápida e estável, em comparação com métodos baseados apenas no gradiente. Adicionalmente, o algoritmo inclui técnicas específicas de poda das árvores com base na informação relevante, através da variável Y , permitindo interromper o crescimento de ramos com impacto marginal na melhoria do modelo. Na Figura 12 está representada um exemplo da arquitetura do modelo *XGBoost*.

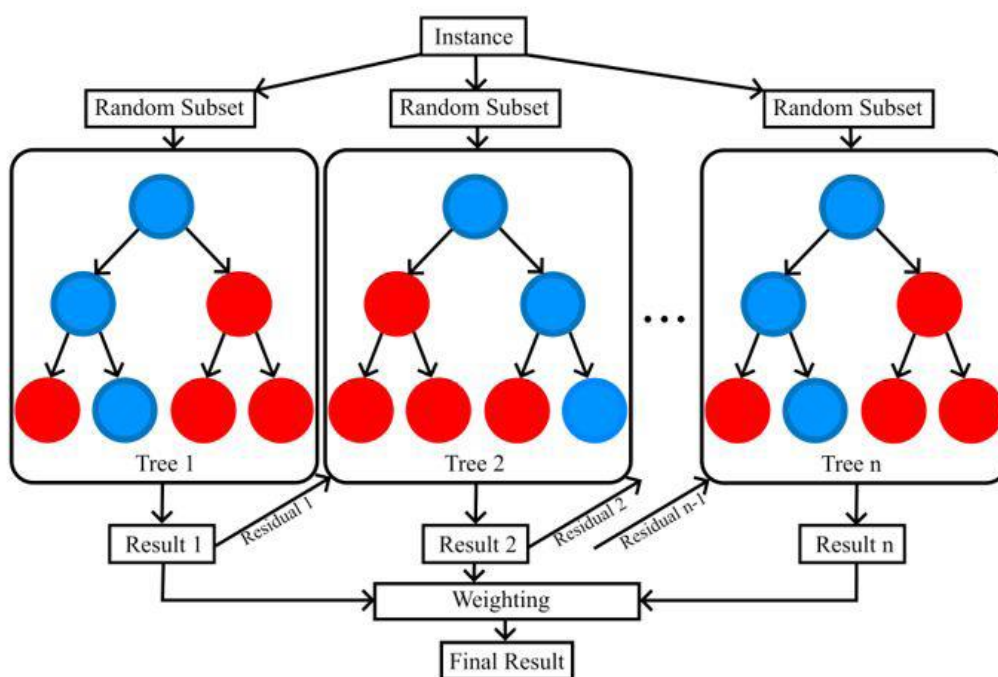


Figura 12 – Representação do XGBoost (Öztornaci, s.d.)

Do ponto de vista computacional, o *XGBoost* foi desenvolvido para suportar paralelização, valores em falta e dados esparsos, com otimizações que tornam o desempenho altamente competitivo em grandes volumes de dados. Além disso, permite avaliar a importância relativa das variáveis explicativas, tornando-o útil não só para previsão, mas também para interpretação de modelos.

2.1.3.2 LightGBM

O **Light Gradient Boosting Machine** é uma *framework* de *gradient boosting* desenvolvida pela equipa da Microsoft (Ke et al., 2017), desenhada com o objetivo de aumentar significativamente a eficiência do treino e reduzir o uso de memória em comparação com implementações tradicionais. Tal como o *XGBoost*, o *LightGBM* segue o princípio do *gradient boosting*, por isso, treina modelos de forma sequencial para minimizar a função de perda. Contudo, distingue-se pelo modo como constrói as árvores e controla os dados, oferecendo melhorias substanciais de escalabilidade, especialmente em conjuntos de dados de grande dimensão e esparsos.

O *LightGBM* é caracterizado pela sua utilização do algoritmo *Leaf-wise*, crescimento por folha, em vez do tradicional *Level-wise*, crescimento por nível. No método *Leaf-wise*, a árvore é expandida a partir da folha com maior valor de perda, maximizando a redução da função objetivo a cada iteração. Este crescimento seletivo leva geralmente a modelos mais profundos e com melhor capacidade de generalização, mas também aumenta o risco de *overfitting*, sendo assim comum a definição de um parâmetro de profundidade máxima. Este comportamento contrasta com o algoritmo *Level-wise*, onde todas as folhas de um nível são divididas em simultâneo, resultando em árvores mais equilibradas, como demonstrado na Figura 13.

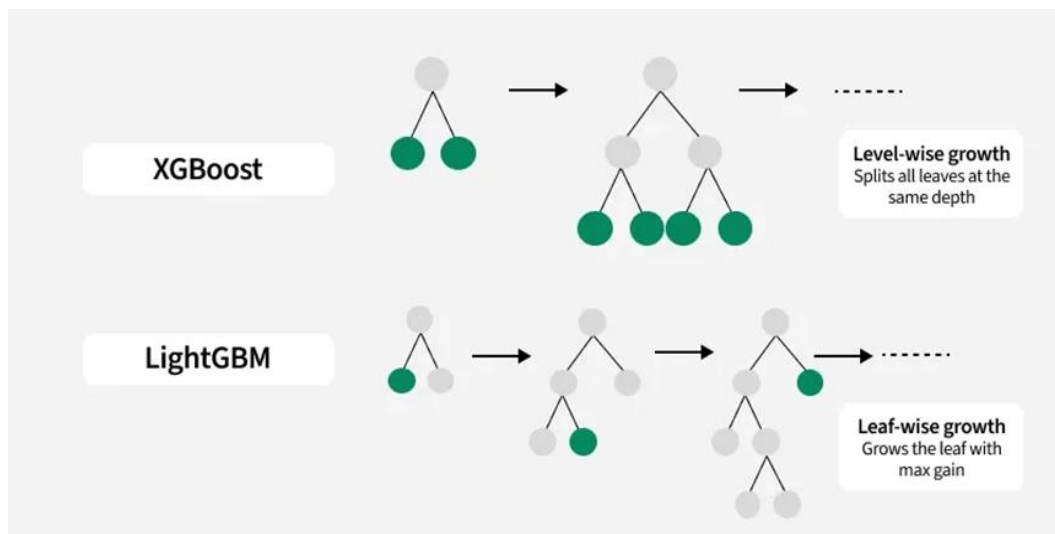


Figura 13 – Algoritmo Leaf-wise do LightGBM (*LightGBM*, s.d.)

Uma das principais razões para o desempenho elevado do *LightGBM* vem da introdução de técnicas específicas de otimização computacional que permitem acelerar o processo de treino e reduzir significativamente o consumo de memória. Entre estas, destaca-se o método *Histogram-based Decision Tree Learning*, no qual os valores contínuos das variáveis explicativas são previamente agrupados em intervalos discretos, construindo histogramas que permitem encontrar os pontos de divisão de forma muito mais eficiente do que a abordagem exaustiva tradicional. Este procedimento reduz o custo computacional da procura por *splits* e facilita a manipulação de grandes volumes de dados. Adicionalmente, o *LightGBM* recorre à técnica

Gradient-based One-Side Sampling (GOSS), que visa reduzir o número de amostras utilizadas no cálculo dos gradientes, mantendo apenas todas as instâncias com grandes valores de gradiente, onde o erro é elevado, e desvaloriza as restantes. Este método preserva a informação mais relevante para o ajuste do modelo, enquanto acelera o processo de otimização. Outra técnica implementada é a *Exclusive Feature Bundling (EFB)*, que consiste na combinação de variáveis mutuamente exclusivas, que raramente assumem valores diferentes de zero na mesma instância, num único atributo artificial, sem perda de informação. Esta fusão reduz eficazmente a dimensão do espaço de entrada, tornando o treino mais rápido e menos exigente em termos de memória, o que é particularmente vantajoso em contextos esparsos ou com um número elevado de variáveis.

2.1.3.3 CatBoost

O **CatBoost** é um algoritmo de *Gradient Boosting* desenvolvido pela Yandex Research, concebido especialmente para lidar com variáveis categóricas de forma eficaz sem necessidade de codificação manual, como a tradicional codificação *one-hot* (Samuels, 2024). Proposto por Prokhorenkova et al. (2018), este algoritmo implementa técnicas avançadas como a codificação baseada em alvos, que calcula estatísticas das variáveis categóricas condicionadas à variável alvo com regularização incorporada, reduzindo substancialmente o risco de *overfitting* associado a categorias raras ou com muitos níveis.

Sobre a arquitetura, o modelo é construído de forma sequencial, adicionando árvores de decisão que aprendem a corrigir erros cometidos pelas árvores anteriores. Em cada iteração, o algoritmo calcula os resíduos, que correspondem aos gradientes negativos da função de perda em relação às previsões. A nova árvore é então ajustada a esses resíduos e integrada no modelo com um fator de aprendizagem, permitindo uma melhoria progressiva das previsões (Prokhorenkova et al., 2018).

Uma inovação na arquitetura do *CatBoost* reside na utilização de árvores de decisão simétricas (*Symmetric Decision Trees*), uma abordagem em que as divisões realizadas em cada nível da árvore seguem critérios idênticos em todos os ramos, resultando numa estrutura equilibrada e homogénea, como demonstrado na Figura 14. Esta configuração estrutural contribui para a melhoria da generalização, aumenta a eficiência computacional e mitiga o risco de *overfitting*, especialmente em contextos de regressão com dados complexos ou ruidosos (Prokhorenkova et al., 2018).

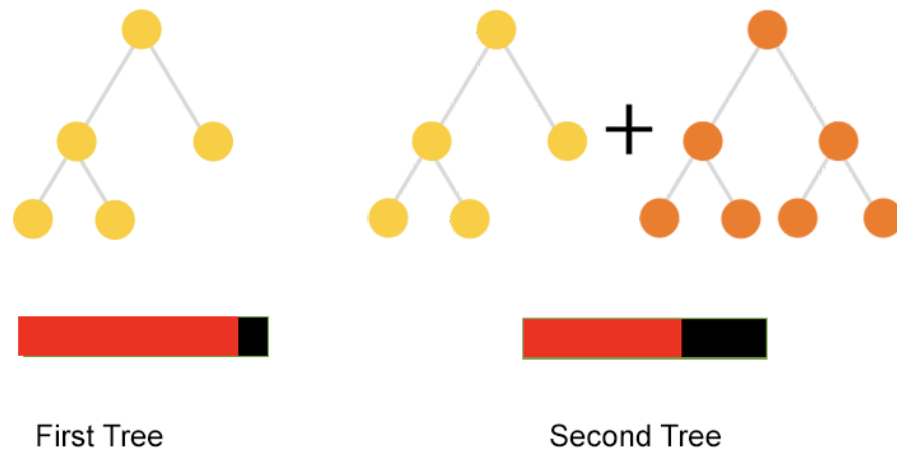


Figura 14 – Aprendizagem do algoritmo CatBoost (CatBoost, 2018)

Estas características tornam o *CatBoost* particularmente indicado para aplicações em contextos empresariais, financeiros e de *marketing*, onde é comum a presença de variáveis categóricas com alta cardinalidade. Este algoritmo apresenta geralmente desempenho superior em conjuntos de dados que possuem muitas categorias e é reconhecido por sua eficiência computacional e robustez em diversos cenários práticos (Hancock & Khoshgoftaar, 2020; Prokhorenkova et al., 2018).

2.1.4 Redes Neurais Recorrentes

As **Redes Neurais Recorrentes (RNNs)** são particularmente adequadas para séries temporais, pois conseguem capturar dependências dinâmicas, mantendo uma memória interna dos dados anteriormente processados (Rumelhart et al., 1986). Este estado interno possibilita modelar dependências temporais e captar padrões dinâmicos ao longo do tempo, tornando as particularmente eficazes com problemas como previsão de séries temporais, processamento de linguagem natural, reconhecimento de fala e análise de dados sequenciais (Lipton, 2015).

Ao contrário das redes neurais tradicionais, *feedforward*, em que a informação flui numa única direção, nas *RNNs*, a saída de um passo no tempo pode influenciar a entrada do passo seguinte. Matematicamente, esta estrutura é representada pelas equações (Lipton, 2015):

$$h^{(t)} = \sigma(W^{hx}x^{(t)} + W^{hh}h^{(t-1)} + b_h) \tag{7}$$

$$\hat{y}^{(t)} = \text{softmax}(W^{yh}h^{(t)} + b_y) \tag{8}$$

onde $x^{(t)}$ é o vetor de entrada no instante t , $h^{(t)}$ é o estado oculto no instante t , $\hat{y}^{(t)}$ representa a saída da rede, W^{hx} , W^{hh} e W^{yh} são as matrizes de pesos ajustáveis, b_h e b_y são os vetores de viés e σ é uma função de ativação não-linear. *Softmax* é uma função de ativação

usada habitualmente em problemas de classificação para gerar probabilidades, no entanto esta pode ser substituída por outra dependendo do problema.

Tabela 4 – Comparação LSTM e GRU

Característica	LSTM	GRU
Estrutura	3 portas (entrada, esquecimento, saída) (Hochreiter & Schmidhuber, 1997)	2 portas (atualização, reset) (Cho et al., 2014)
Complexidade	Mais complexa devido ao maior número de parâmetros (Hochreiter & Schmidhuber, 1997)	Menos complexa devido a conter menos parâmetros (Cho et al., 2014)
Capacidade de memória	Potencialmente maior devido à célula de memória separada (Hochreiter & Schmidhuber, 1997)	Menor, sem célula de memória dedicada (Cho et al., 2014)
Velocidade de treino	Relativamente mais lenta devido à maior complexidade (Lipton, 2015)	Mais rápida por ter menos parâmetros (Lipton, 2015)
Eficiência computacional	Menor eficiência devido à complexidade adicional (Lipton, 2015)	Maior eficiência por simplicidade estrutural (Lipton, 2015)
Aplicações típicas	Problemas com dependências temporais longas e complexas (Hochreiter & Schmidhuber, 1997)	Problemas que requerem maior velocidade e menos memória (Cho et al., 2014)

Contudo, as *RNNs* tradicionais enfrentam dificuldades com longas dependências temporais devido ao problema do *vanishing gradient*. Este problema ocorre especialmente durante o processo de retropropagação através do tempo (*Backpropagation Through Time*). Para contornar esta limitação, surgiram variantes avançadas como as *Long Short-Term Memory (LSTM)* e as *Gated Recurrent Units (GRU)*. Estas arquiteturas introduzem mecanismos de portas que controlam o fluxo de informação, facilitando a captura de dependências temporais de longo prazo (Hochreiter & Schmidhuber, 1997; Cho et al., 2014). Na Tabela 4 estão detalhadas as principais diferenças entre estas duas arquiteturas.

Em termos de estrutura interna, as *LSTM* (Kjærø et al., 2021) utilizam três portas principais, entrada, esquecimento e saída, enquanto as *GRU* recorrem apenas a duas portas, atualização e *reset*, conforme descrito por Hochreiter e Schmidhuber (1997) e Cho et al. (2014). Na complexidade do modelo, a *LSTM* é mais complexa devido ao maior número de parâmetros,

resultante da existência de múltiplas portas e da célula de memória separada, enquanto a *GRU* apresenta uma arquitetura mais simples e menos parametrizada (Hochreiter & Schmidhuber, 1997; Cho et al., 2014). Esta diferença tem impacto direto na capacidade de memória de cada arquitetura, onde as *LSTM* conseguem reter dependências temporais mais longas devido à presença de uma célula de memória dedicada, enquanto as *GRU* tendem a perder informação mais rapidamente (Hochreiter & Schmidhuber, 1997; Cho et al., 2014). Devido à sua maior complexidade, as *LSTM* são geralmente mais lentas no treino, em comparação com as *GRU*, que por terem menos parâmetros, permitem um treino mais eficiente (Lipton, 2015). Esta simplicidade estrutural das *GRU* traduz-se também numa maior eficiência computacional, sobretudo em aplicações onde os recursos computacionais são limitados (Lipton, 2015). No que diz respeito às aplicações típicas, as *LSTMs* são particularmente adequadas a tarefas que requerem a modelação de dependências temporais longas e complexas, como tradução automática (Sutskever et al., 2014), reconhecimento de voz (Graves et al., 2013) e séries temporais financeiras (Fischer & Krauss, 2018; Nelson et al., 2017).

Assim, as *LSTM* são particularmente indicadas para problemas complexos que requerem a retenção de informações por períodos temporais prolongados, enquanto as *GRU* são preferidas em cenários onde a eficiência computacional e a velocidade de treino são fatores prioritários. No entanto, estas podem ser menos eficazes em tarefas com dependências temporais muito longas e complexas (Hochreiter & Schmidhuber, 1997; Cho et al., 2014). A eficácia das *RNNs* em aplicações práticas depende significativamente da seleção apropriada das variáveis explicativas, da definição da janela temporal, da configuração dos hiperparâmetros do modelo, assim como, da dimensão da camada oculta e da função de ativação escolhida.

Na área da previsão de preços de eletricidade, estas arquiteturas têm mostrado resultados promissores. Lago et al. (2021) revela que *LSTM* e *GRU* superam tanto modelos estatísticos clássicos como técnicas baseadas em *Machine Learning* tradicional, sobretudo quando os dados apresentam elevada variabilidade temporal, sazonalidade e padrões não-lineares. Um dos grandes benefícios destas redes é a sua capacidade de captar ciclos de consumo, variações climáticas e efeitos semanais ou sazonais, características comuns no mercado spot de eletricidade.

2.1.5 Sistematização dos Modelos

Na Tabela 5 encontra-se uma sistematização dos métodos mencionados. Nesta tabela são apresentadas as vantagens e desvantagens da implementação e do comportamento face a estes modelos.

Tabela 5 – Sistematização de Técnicas de ML

Modelo	Vantagens	Desvantagens
<i>Regressão Linear / Ridge / Lasso</i>	Simples de implementar; Alta interpretabilidade; Baixo custo computacional.	Não captura relações não-lineares; Fraco desempenho com variáveis interdependentes.
<i>Árvores de Decisão</i>	Intuitiva e fácil de interpretar; Bom desempenho em problemas simples.	Propensa a overfitting com uma grande quantidade de dados; Frac generalização individualmente.
<i>Random Forest</i>	Reduz overfitting em relação a árvores de decisão; Lida bem com variáveis categóricas e contínuas.	Menos interpretável; Tempo de treino moderado.
<i>Gradient Boosting (XGBoost, LightGBM, CatBoost)</i>	Elevada precisão; Lida bem com dados tabulares complexos; Boa generalização com os hiperparâmetros adequados.	Mais difícil de ajustar; Propenso a overfitting sem regularização adequada.
<i>RNN / LSTM / GRU</i>	Possibilita sequências temporais; Interpreta séries históricas e padrões sazonais.	Treino mais complexo que modelos tradicionais; Sensível a parâmetros; Custo computacional.

2.2 Questões de Investigação

Nesta secção, define-se a principal questão de investigação que se pretende analisar no âmbito deste trabalho de mestrado: “Como a volatilidade do mercado energético afeta a previsão dos preços de eletricidade na Península Ibérica?”. De forma a responder a esta questão, uma análise detalhada à literatura deve ser feita para fornecer respostas às seguintes questões provenientes:

1. Qual é o impacto direto da produção de energia renovável nas flutuações dos preços de eletricidade?
2. Como os dados históricos de produção de energia e procura têm sido usados para prever variações nos preços?
3. Quais são as estratégias atuais e potenciais para mitigar a volatilidade causada pela intermitência das renováveis?
4. Que papel a previsão meteorológica desempenha na gestão da volatilidade dos preços e na operação do mercado elétrico?

Esta primeira questão visa identificar e investigar como a influência das flutuações da produção de energia renovável, tais como a solar e a eólica, impactam diretamente os preços de eletricidade. O foco reside em entender a correlação entre a produção intermitente de energia e as oscilações de preços no mercado. Serão discutidas metodologias para quantificar esse impacto, tendo em consideração fatores como a produção efetiva e as condições meteorológicas.

A segunda questão examina como os grandes volumes de dados sobre a produção e a procura de energia são utilizados para desenvolver modelos de previsão de preços de eletricidade. Serão estudadas técnicas de análise de dados e modelos estatísticos para compreender como estes dados são transformados em previsões, bem como de que modo poderão ser adaptados aos modelos de forma a melhor corresponder à realidade do mercado.

Na terceira questão, são identificadas as abordagens e tecnologias atuais e emergentes adotadas para diminuir o efeito da flutuação dos preços de eletricidade face à volatilidade introduzida pela produção de energia renovável. Também serão consideradas inovações tecnológicas e políticas regulatórias que possam facilitar a gestão dessa volatilidade.

Por último, a quarta questão menciona como as previsões meteorológicas são utilizadas para antecipar a produção de energia renovável e, conseqüentemente, influenciar a gestão e a previsão do mercado de eletricidade. Será analisado o impacto das previsões de curto e longo prazo na tomada de decisões operacionais e estratégicas, além de explorar as técnicas e ferramentas utilizadas para integrar estas previsões nos sistemas de planeamento energético.

2.3 Critérios de Elegibilidade

Para a realização do estudo, definiram-se critérios de inclusão e exclusão que podem ser consultados na Tabela 6 e na Tabela 7, respetivamente. Os estudos que satisfizeram pelo menos um ou mais critérios de inclusão foram considerados, enquanto os estudos que corresponderam a um ou mais critérios de exclusão foram descartados.

Tabela 6 – Critérios de Inclusão

<i>ID</i>	<i>Critérios de Inclusão</i>
<i>IC1</i>	A fonte centra-se na análise da volatilidade dos preços de eletricidade;
<i>IC2</i>	A fonte explora o impacto da integração de energias renováveis na volatilidade dos preços;
<i>IC3</i>	A fonte avalia métodos de previsão de preços de eletricidade em mercados com alta inserção de energias renováveis;
<i>IC4</i>	A fonte identifica os benefícios e desafios atuais na gestão da volatilidade dos preços.

Tabela 7 – Critérios de Exclusão

ID	Critérios de Exclusão
EC1	As fontes não estão escritas em português ou inglês;
EC2	As fontes são publicadas antes de 2010;
EC3	As fontes não apresentam estudos relacionados com a volatilidade dos preços de eletricidade ou gestão de energia renovável;
EC4	As fontes são duplicadas.

2.4 Fontes de Informação

As bases de dados utilizadas para encontrar as informações adquiridas foram o IEEE Xplore (IEEE Xplore, 2025), a ACM Library (ACM Digital Library, 2025) e o ScienceDirect (ScienceDirect, 2025). Para tal, os termos de pesquisa, identificados na Tabela 8, foram utilizados como uma combinação para encontrar os melhores resultados. As cadeias de pesquisa foram adaptadas a cada metodologia de pesquisa de bases de dados.

Tabela 8 – Termos de Pesquisa

Escopo	Termos de Pesquisa
Energia	(Energy OR Electricity) AND
Volatilidade	(Volatility OR Volatile) AND
Previsão	(Prediction OR Predictive) AND
Modelo	(Algorithm OR model)

2.5 Seleção de estudos e processo de extração de dados

Após a definição de cada condição necessária para incluir ou excluir um determinado estudo, o processo de seleção e extração de dados foi realizado utilizando as diretrizes PRISMA divididas em quatro fases diferentes: Identificação, Triagem, Elegibilidade e Inclusão (Liberati et al. 2009).

Como demonstrado na Figura 15, na fase de **Identificação** obteve-se um total de 1453 resultados recuperados das bases de dados selecionadas. Na fase de **Triagem**, excluíram-se 24 registos duplicados (com o critério de exclusão EC4), seguidos de 936 registos excluídos após a revisão dos títulos e resumos. Na fase de **Elegibilidade**, efetuou-se uma análise do texto integral dos restantes artigos de acordo com os critérios de Inclusão e Exclusão definidos. Apenas 493 estudos foram considerados elegíveis para a fase de Inclusão. Finalmente, na fase de **Inclusão**, recolheu-se um número total de 21 estudos elegíveis.

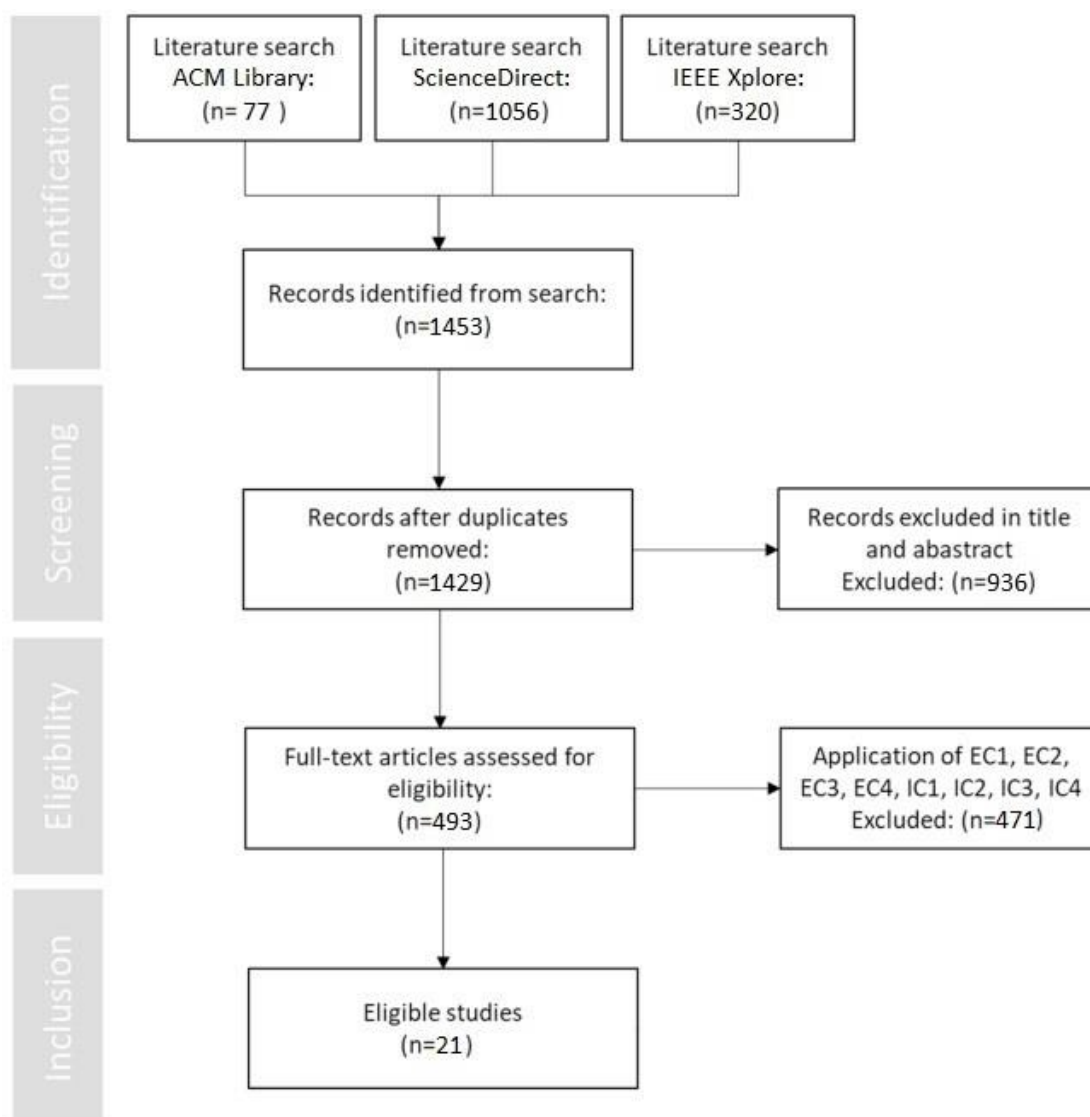


Figura 15 - Diagrama do PRISMA

2.6 Discussão

Nesta secção serão analisadas as questões de investigação propostas anteriormente através da análise dos artigos recolhidos.

2.6.1 Qual é o impacto direto da produção de energia renovável nas flutuações dos preços de eletricidade?

A crescente integração de fontes de energia renováveis nos sistemas energéticos internacionais tem implicações significativas na volatilidade dos preços de eletricidade. A natureza intermitente da produção de energia a partir de fontes, solar e eólica, introduz uma camada de incerteza, o que resulta em elevadas flutuações nos preços de eletricidade

(Nandakumar et al., 2017). Este fenômeno ocorre devido à limitada capacidade de armazenamento de energia, que, é crucial para equilibrar a oferta e a procura em tempo real (Zheng et al., 2014).

Estudos têm explorado como a exploração de energias renováveis afeta os mercados de energia, incluindo o mercado de gás natural, que está diretamente ligado à produção de eletricidade. A análise de Nandakumar e Annaswamy (2017) destaca como a dependência do gás natural, para complementar a produção de energia renovável, impacta não apenas os preços do gás, mas também os preços de eletricidade. Esta relação é crucial para entender as dinâmicas do mercado em regiões com alta inserção de energias renováveis, tal como na Península Ibérica. A previsão a curto prazo da velocidade do vento e a subsequente produção de energia eólica são componentes importantes para a gestão eficaz da volatilidade dos preços. Estudos como o de Bunn et al. (2021) demonstram que a previsibilidade dos preços do mercado de equilíbrio britânico, apesar de sua complexidade e da crescente influência de fontes renováveis, sugere que modelos avançados de previsão podem desempenhar um papel importante para mitigar os riscos associados às flutuações de preços. Para complementar, o estudo de Al-Hajj et al. (2023) demonstra que modelos de *Machine Learning*, como os *Support Vector Regressors* e as *Random Forests*, aplicados na previsão da velocidade do vento podem melhorar significativamente a qualidade das previsões de curto prazo, permitindo uma resposta mais ágil às variações na produção de energia eólica.

A fim de controlar adequadamente a volatilidade dos preços e integrar com eficiência as fontes de energias renováveis ao *mix* energético, será necessário desenvolver e implementar métodos avançados de previsão e estratégias robustas de armazenamento de energia. A combinação de previsões adequadas e a capacidade de resposta rápida ao armazenamento pode reduzir os riscos financeiros associados às flutuações de preço, e propiciar uma operação de mercado mais estável e previsível.

2.6.2 Como os dados históricos de produção de energia e procura têm sido usados para prever variações nos preços?

A análise histórica de dados de produção de energia e procura ajuda a prever flutuações de preços com maior precisão e a desenvolver estratégias eficazes para a gestão de riscos no mercado energético. O artigo "*Developing Two RNN-Based Algorithms for Electricity Price Forecasting in Markets with High Penetration of Renewable Energy Resources*" explora o uso de Redes Neurais Recorrentes (RNNs) para criar modelos de previsão baseados em vastos conjuntos de dados históricos (Kock, Robin et al. 2024). Estes modelos são treinados para identificar padrões e tendências nos dados que frequentemente precedem em variações significativas nos preços de eletricidade. Com isto, os modelos permitem previsões de preço para o dia seguinte que são críticas para operadores e investidores no mercado energético.

Além disso, no artigo "*Extreme Supervised Algorithm for Day Ahead Market Price Forecasting*" ilustra como métodos de *Machine Learning* supervisionados, como Máquinas de Vetores de Suporte (SVMs), são utilizadas para prever preços em mercados específicos,

utilizando dados históricos de preços e outras variáveis económicas (Loizidis et al., 2023). Este estudo destaca a eficácia de combinar modelos lineares com técnicas de *bootstrap* para melhorar a precisão das previsões de preços. O estudo demonstrou que, em anos como 2019, 2021 e 2022, o uso exclusivo de dados históricos proporcionou previsões precisas. Em 2020, devido à pandemia, que introduziu padrões anômalos nos mercados de eletricidade, os resultados apresentaram pior desempenho, concluindo que estas anomalias podem ser consideradas limitações dos modelos.

Hang Yang et al. (2014) em "*A Method to Predict the Intermittent Power by Classification Model.*" investigaram o potencial das técnicas de *data mining* para melhorar a previsão e utilização da energia solar, com o objetivo de mostrar como métodos avançados podem otimizar o uso de dados históricos para prever a produção de energia. Estas abordagens são fundamentais para antecipar as necessidades de capacidade de produção e ajustar os mecanismos de resposta da procura de forma mais eficiente.

Outro estudo relevante é o "*Day-ahead probabilistic load forecasting for individual electricity consumption*" que avalia métodos probabilísticos e baseados em intervalos para prever o consumo de energia. Este artigo demonstra como modelos baseados em dados históricos podem fornecer previsões robustas e adaptáveis para diversos cenários de mercado, favorecendo uma otimização para decisões de compra e venda de energia no mercado energético (Zhang et al., 2023). No artigo "*Optimal Allocation of Energy Storage System for Risk Mitigation of DISCOs With High Renewable Penetrations*" é descrito como ferramentas de previsão avançadas, como *OptiLoad* e *OptiSolar*, através de dados como a temperatura, condições climáticas sazonais, feriados e dados históricos podem ser integrados para prever a produção e procura de energia, permitindo aos operadores planejar melhor as transações no mercado e reduzir os riscos associados à volatilidade dos preços (Zheng et al., 2014).

Estes estudos demonstram que o uso avançado de técnicas de análise de dados e *Machine Learning* pode significativamente melhorar a capacidade de previsão dos preços de eletricidade. Assim, além de mitigar os riscos financeiros, também potencia a estabilidade do mercado ao permitir que os seus participantes tomem decisões informadas baseadas em previsões confiáveis.

2.6.3 Quais são as estratégias atuais e potenciais para mitigar a volatilidade causada pela intermitência das energias renováveis?

A intermitência natural das fontes de energia renováveis favorece o aumento da volatilidade nos preços de eletricidade, o que pode levar à desestabilização dos mercados energéticos e a um aumento dos riscos financeiros para os produtores e consumidores de eletricidade. Abordagens e estratégias inovadoras são essenciais para mitigar estes impactos e garantir uma transição energética sustentável.

Um dos métodos mais eficazes encontrados na literatura para mitigar esta instabilidade é a utilização de sistemas de armazenamento de energia, conforme discutido em "*Optimal*

Allocation of Energy Storage System for Risk Mitigation of DISCOs With High Renewable Penetrations" (Zheng et al., 2014). Este estudo explora como a alocação estratégica de sistemas de armazenamento em baterias pode ser utilizada para suavizar as flutuações de preço causadas pela produção de energia intermitente. Ao armazenar energia durante períodos de baixa procura e alta produção, e libertá-la durante picos de procura ou baixa produção, os sistemas de armazenamento podem ajudar a estabilizar os preços de eletricidade e reduzir a volatilidade do mercado. No artigo *"Pumped Storage Hydroelectricity for a Sustainable Electricity Transition"* é destacado o papel da energia hidrelétrica de bombeamento como uma solução de armazenamento em larga escala. Este sistema equilibra variações na oferta e na procura, atuando como um *buffer* eficiente para lidar com a intermitência das fontes renováveis e mantendo a estabilidade da rede elétrica (Johansson et al., 2023). Outros estudos como *"Research on Optimization Strategy of Energy Storage and Charging Based on Photovoltaic Power Prediction Model"* (Wang et al., 2023b) e o *"The Smoothing Strategy of Wind Power Combined with Storage Based on Least Square Fitting"* (Liu et al., 2016) também defendem sistemas de armazenamento de energias devido à baixa dependência de respostas reativas e a uma melhor estabilidade da flutuação de preços.

No artigo *"Impact of increased renewables on natural gas markets in eastern United States"* (Nandakumar et al., 2017) é analisado a forma como o aumento da produção de energia renovável afeta o mercado de gás natural, que serve como um *backup* flexível para as fontes de energia renováveis. Este estudo sugere que um entendimento profundo da interação entre os mercados de gás e eletricidade pode permitir uma melhor gestão dos recursos energéticos e uma resposta mais eficaz às flutuações dos preços.

O artigo *"Adjustable Load Resource Pool Operation Analysis"* (Wang et al., 2023a) destaca o papel da flexibilidade da procura para suavizar as flutuações das energias renováveis. Este trabalho descreve como a agregação de recursos ajustáveis pode melhorar a estabilidade da rede, utilizando respostas rápidas da procura para equilibrar a oferta e a procura em diferentes escalas de tempo.

Outra abordagem complementar é apresentada em *"A Theory for the Economic Operation of a Smart Grid with Stochastic Renewables"*. Este estudo propõe um sistema descentralizado de resposta à procura, que utiliza mecanismos de preços dinâmicos para otimizar a integração de energias renováveis e reduzir a volatilidade dos preços de forma eficiente, enquanto preserva a privacidade dos utilizadores e garante a flexibilidade operacional do sistema (Singh et al., 2015).

O uso de modelos avançados de previsão também desempenha um papel vital na mitigação da volatilidade. Métodos de *data mining* e classificação para prever a produção intermitente de energia renovável, como os descritos no artigo *"A Method to Predict the Intermittent Power by Classification Model"*, ajudam a antecipar flutuações de produção e a ajustar as operações de mercado de forma proativa (Yang et al., 2014).

Estas estratégias, juntamente com políticas regulatórias que incentivem a integração e o suporte de fontes de energia renováveis, são essenciais para desenvolver um sistema

energético mais resiliente e menos suscetível à volatilidade. A implementação de tais medidas não só ajudará a mitigar os riscos associados à intermitência da produção de energia a partir de fontes renováveis, mas também poderá suportar a transição para uma matriz energética mais verde e sustentável.

2.6.4 Que papel a previsão meteorológica desempenha na gestão da volatilidade dos preços e nas operações do mercado elétrico?

A previsão meteorológica é uma ferramenta essencial na gestão da volatilidade dos preços e nas operações dos mercados de eletricidade, especialmente em contextos com alto crescimento do mercado de energias renováveis. A precisão destas previsões é crucial para a tomada de decisões operacionais e estratégicas, pois influencia diretamente a estimativa de produção de energias renováveis, como a solar e a eólica.

No artigo "*Ultra-Short-Term Forecasting of Wind Speed Using Lightweight Features and Machine Learning Models*" de Al-Hajj et al. (2023) é destacada a importância da previsão da velocidade do vento a curto prazo. Este estudo demonstra como previsões meteorológicas precisas podem ajudar a prever a disponibilidade de energia eólica, o que é vital para ajustar as operações de mercado em tempo real e mitigar riscos associados à volatilidade dos preços.

Estas previsões são integradas em sistemas de gestão energética para melhorar a eficiência das operações de rede, tendo como objetivo que a oferta e a procura de energia sejam balanceadas de forma eficaz. A capacidade de antecipar a produção de energias renováveis permite que os operadores do mercado elétrico planeiem com mais eficácia, evitando desequilíbrios que podem levar a picos de preço ou à necessidade de intervenções de última hora.

Conforme discutido em "*Developing Two RNN-Based Algorithms for Electricity Price Forecasting in Markets with High Penetration of Renewable Energy Resources*" (Kock et al., 2024), a incorporação de dados de previsão meteorológica em modelos de previsão de preços de eletricidade também demonstra uma melhoria significativa na precisão das previsões. Estes modelos, ao utilizar dados meteorológicos, conseguem ajustar as previsões de preço para refletir de forma mais precisa a influência das condições climáticas sobre a produção de energia.

Adicionalmente, no estudo "*Day-Ahead Power Output Forecasting for Small-Scale Solar Photovoltaic Electricity Generators*" de Zhang et al. (2015) é explorado como os modelos de previsão baseados em dados meteorológicos podem aumentar significativamente a precisão na previsão de produção fotovoltaica. Para operações de mercado com alta exploração de energia solar é informação bastante valiosa, onde a variabilidade meteorológica pode causar flutuações imprevisíveis nos preços, considerando o efeito do incremento da oferta.

2.6.5 Sistematização da Discussão

Na Tabela 9 encontra-se uma sistematização das respostas às questões de investigação. Nesta tabela são mencionadas as perguntas de investigação, os artigos encontrados e conclusões sobre a informação retirada dos artigos.

Tabela 9 – Sistematização da discussão

Pergunta de Investigação	Estudo(s) Relacionado(s)	Conclusões
Qual é o impacto direto da produção de energia renovável nas flutuações dos preços de eletricidade?	(Nandakumar et al., 2017) (Zheng et al., 2014) (Bunn et al., 2021) (Al-Hajj et al., 2023)	A produção de energia renovável é variável, o que torna a flutuação dos preços de eletricidade maior. Esta variação pode derivar da meteorologia. A capacidade de armazenamento não é suficiente para solucionar o problema da volatilidade, logo uma solução seria aumentar o armazenamento.
Como os dados históricos de produção de energia e a procura têm sido usados para prever variações nos preços?	(Kock, Robin et al., 2024) (Loizidis et al., 2023) (Hang Yang et al., 2014) (Zhang et al., 2023) (Zheng et al., 2014)	Dados históricos beneficiam as previsões dos modelos devido às tendências registadas nos dados, no entanto existem limitações. Dados históricos sozinhos não são suficientes para capturar anomalias significativas (ex.: pandemia).
Quais estratégias atuais e potenciais podem mitigar a volatilidade causada pela intermitência das renováveis?	(Zheng et al., 2014) (Johansson et al., 2023) (Wang et al., 2023a) (Liu et al., 2016) (Nandakumar et al., 2017) (Wang et al., 2023b) (Singh et al., 2015) (Yang et al., 2014)	Sistemas de armazenamento de energia podem reduzir a flutuação dos preços devido a armazenarem a energia das fontes renováveis. No entanto, a implementação destes sistemas implica custos. Modelos de previsão avançados de <i>Machine Learning</i> podem prever com maior eficiência esta volatilidade, no entanto exigem uma infraestrutura computacional robusta e qualidade nos dados utilizados pelos modelos.
Que papel a previsão meteorológica desempenha na gestão da volatilidade dos preços e na operação do mercado elétrico?	(Al-Hajj et al., 2023) (Kock et al., 2024) (Zhang et al., 2015)	Previsões meteorológicas a longo prazo ainda apresentam alta incerteza. No entanto, dados meteorológicos combinados com modelos de previsão avançados podem ter maior sucesso nas previsões, devido à meteorologia influenciar as fontes de energias renováveis.

2.7 Sumário

No âmbito deste Capítulo foi realizada uma revisão do estado da arte, com o objetivo de compreender os fatores que influenciam a volatilidade dos preços de eletricidade e as abordagens utilizadas para a sua previsão. A produção de energia renovável tem um impacto direto nas flutuações dos preços de eletricidade. Estudos indicam que a variabilidade inerente às fontes renováveis, dependente das condições meteorológicas, aumenta a volatilidade do mercado elétrico (Nandakumar et al., 2017; Zheng et al., 2014; Bunn et al., 2021; Al-Hajj et al., 2023). A capacidade atual de armazenamento de energia não é suficiente para compensar estas flutuações. Uma solução para o problema seria o aumento de sistemas de armazenamento, conseguindo assim mitigar parcialmente este efeito da volatilidade. Dados históricos da oferta e da procura tem sido amplamente utilizados para prever variações nos preços. A informação relativamente ao passado permite aos modelos captar tendências e padrões, melhorando a previsão (Kock et al., 2024; Loizidis et al., 2023; Hang Yang et al., 2014; Zhang et al., 2023; Zheng et al., 2014). Contudo, os dados históricos apresentam limitações, eventos atípicos, como crises ou pandemias, e até podem não ser capturados adequadamente, comprometendo a precisão das previsões.

Diversos métodos e estratégias têm sido propostos para mitigar a volatilidade causada pela intermitência das fontes renováveis. O uso de sistemas de armazenamento de energia permite reduzir flutuações ao disponibilizar energia acumulada, embora a implementação envolva custos significativos (Zheng et al., 2014; Johansson et al., 2023; Wang et al., 2023a; Liu et al., 2016; Nandakumar et al., 2017; Wang et al., 2023b; Singh et al., 2015; Yang et al., 2014). Paralelamente, modelos avançados de previsão baseados em *Machine Learning* oferecem maior eficiência na antecipação das variações de preços, mas dependem de uma infraestrutura computacional robusta e da qualidade dos dados. Finalmente, o papel da previsão meteorológica na gestão da volatilidade dos preços e na operação do mercado elétrico é destacado em vários estudos (Al-Hajj et al., 2023; Kock et al., 2024; Zhang et al., 2015). Apesar das previsões a longo prazo ainda apresentarem elevada incerteza, a integração de dados meteorológicos com modelos avançados de previsão permite uma melhoria significativa na antecipação das flutuações, dado que a meteorologia influencia diretamente a produção das fontes renováveis.

No domínio de *Machine Learning*, diferentes abordagens apresentam vantagens e limitações específicas. Modelos lineares, como *Regressão Linear*, *Ridge* e *Lasso*, são simples de implementar, interpretáveis e computacionalmente económicos, mas não capturam relações não lineares e apresentam desempenho limitado em situações de variáveis interdependentes. Árvores de Decisão são intuitivas e apresentam bom desempenho em problemas simples, mas têm tendência ao *overfitting*. *Random Forest* melhora a robustez das árvores isoladas, enquanto modelos de *Gradient Boosting*, como *XGBoost*, *LightGBM* e *CatBoost*, proporcionam elevada precisão e boa generalização, embora exijam maior cuidado na parametrização. Redes neurais recorrentes (*RNN*, *LSTM*, *GRU*) permitem modelar sequências temporais e padrões sazonais, mas requerem treino mais complexo e maior recurso computacional.

Em síntese, o estado da arte demonstra que a previsão de preços de eletricidade exige uma abordagem integrada, combinando dados históricos, informação meteorológica e modelos de *Machine Learning* capazes de capturar relações não lineares e padrões temporais, considerando a parametrização e capacidade computacional dos modelos como fatores críticos para o desempenho.

3 Previsão de preços de eletricidade no mercado spot

Neste capítulo é detalhado o problema em estudo, a metodologia adotada, o *dataset* e o respetivo pré-processamento, incluindo a análise exploratória e o tratamento de dados necessário para a previsão nos modelos.

3.1 Descrição do Problema

A previsão de preços de eletricidade no mercado spot constitui atualmente uma das tarefas mais complexas e críticas para o setor energético, especialmente num contexto de transição para fontes renováveis. No caso da Península Ibérica, o operador OMIE (Operador do Mercado Ibérico de Eletricidade) é responsável pela organização dos mercados diário e intradiário, e enfrenta o desafio crescente de manter a estabilidade do sistema elétrico num cenário altamente dinâmico e incerto (OMIE, 2023). A principal dificuldade reside na volatilidade dos preços, fenómeno diretamente influenciado por fatores como a flutuação da procura, a variabilidade da produção renovável, o comportamento dos mercados internacionais de energia, e condições meteorológicas instáveis. A integração de energias renováveis, como a solar e a eólica, embora essencial para cumprir metas climáticas, introduz um elevado grau de intermitência e imprevisibilidade no sistema elétrico (IEA, 2021). Esta intermitência e imprevisibilidade ocorre porque a produção não pode ser controlada da mesma forma que as fontes convencionais, sendo dependente de variáveis exógenas como o vento ou a radiação solar.

Esta instabilidade desafia a tradicional gestão do equilíbrio entre oferta e procura, essencial para a definição de preços no mercado spot. Como resposta, os operadores do sistema e os agentes de mercado têm investido em estratégias de previsão mais sofisticadas, que permitam reduzir riscos e melhorar a eficiência operacional (Weron, 2014). Em particular,

a previsão de preços torna-se uma ferramenta indispensável para produtores, comercializadores, consumidores industriais e reguladores, permitindo otimizar decisões de compra e venda de energia, estratégias de armazenamento, e até mesmo políticas públicas de regulação. No entanto, apesar do avanço de técnicas estatísticas e econométricas, os métodos tradicionais demonstram limitações face à complexidade não-linear e multivariada do problema. A elevada granularidade temporal (previsões horárias ou mesmo em intervalos de 15 minutos), combinada com o volume e heterogeneidade dos dados disponíveis, exige abordagens mais robustas, capazes de modelar dependências temporais e incorporar múltiplas fontes de informação contextual. Neste sentido, a previsão de preços já não pode ser encarada como um problema puramente estatístico, mas sim como um problema multidimensional que requer ferramentas computacionais avançadas, capazes de integrar séries temporais, variáveis exógenas e até fontes semiestruturadas de informação. Os métodos tradicionais, embora úteis como base, têm dificuldade em captar interações não-lineares e dependências temporais de médio a longo prazo, especialmente em contextos de elevada variabilidade como o setor energético (Lago et al., 2021).

Neste enquadramento, surge a necessidade de aplicar modelos de *Machine Learning* e *Deep Learning* que não só lidem eficazmente com grandes volumes de dados históricos, como também aprendam padrões complexos e adaptem-se à dinâmica evolutiva do mercado. A aplicação de RNNs, particularmente das variantes LSTM, tem demonstrado resultados promissores na previsão de séries temporais energéticas, graças à sua capacidade de capturar relações de longo prazo e sequências dependentes no tempo (Bianchi et al., 2017; Abumohsen et al., 2023).

3.2 Descrição da Metodologia a Seguir

Para orientar o desenvolvimento, é seguida a metodologia CRISP-DM (Wirth & Hipp, 2000), amplamente adotada em projetos de estudo de modelos de *Machine Learning* e análise preditiva. Esta metodologia caracteriza-se pela sua estrutura iterativa e flexível, permitindo uma abordagem sistemática e adaptativa à resolução de problemas complexos através de dados. A sua utilização justifica-se pela capacidade de abranger todo o ciclo de vida analítico, desde a compreensão do problema até à avaliação e implementação do modelo.

Esta metodologia é composta por seis fases principais: compreensão do negócio, compreensão dos dados, preparação dos dados, modelação, avaliação e implementação. Estas fases serão adaptadas ao contexto da previsão de preços de eletricidade no mercado spot, conforme descrito posteriormente.

A fase de compreensão do negócio foi desenvolvida no subcapítulo anterior, Neste capítulo é detalhado o problema em estudo, a metodologia adotada, o *dataset* e o respetivo pré-processamento, incluindo a análise exploratória e o tratamento de dados necessário para a previsão nos modelos.

Descrição do Problema, onde se analisou a volatilidade dos preços de eletricidade no mercado spot e a sua importância estratégica para a estabilidade do sistema. Foram identificados os principais fatores de incerteza, como a crescente produção de fontes renováveis e a necessidade de mecanismos preditivos robustos para apoiar a tomada de decisão no OMIE.

A fase de compreensão dos dados consiste na análise exploratória do conjunto de dados, que compreende de registos históricos horários de variáveis como procura real e a prevista, preços do mercado diário e de tempo real, variáveis meteorológicas, e indicadores de serviços de regulação. Será realizada uma avaliação da qualidade dos dados, identificação de padrões sazonais, correlações entre variáveis e possíveis *outliers* ou valores em falta.

A etapa de preparação dos dados envolverá o tratamento das variáveis para uso em modelos de *Machine Learning*. Para tal, incluirá normalização, tratamento de valores em falta, criação de variáveis derivadas, como *lags* e médias móveis, e a transformação de variáveis temporais em representações cíclicas.

A fase de modelação contempla a seleção, parametrização e treino de diferentes algoritmos de previsão. Serão considerados modelos supervisionados, com especial foco em Redes Neurais Recorrentes, bem como modelos comparativos como Regressão Linear, Árvores de Decisão, *Random Forest* e algoritmos baseados em *Gradient Boosting*, como por exemplo, XGBoost, LightGBM e CatBoost. Os modelos serão avaliados quanto à sua capacidade de prever preços horários de eletricidade com base em dados históricos e exógenos.

Na fase de avaliação, será conduzida uma análise comparativa do desempenho dos modelos com recurso a métricas quantitativas como Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE) e o Mean Absolute Percentage Error (MAPE). Adicionalmente, será feita uma análise qualitativa dos erros, da robustez dos modelos e da interpretabilidade dos resultados.

Por fim, embora a fase de implementação não seja desenvolvida de forma operacional, serão discutidas as implicações práticas da adoção de modelos preditivos no contexto real do mercado elétrico, incluindo as suas aplicações para operadores, comercializadores e entidades reguladoras.

A adoção do CRISP-DM, aliada à experimentação e comparação de diferentes modelos de previsão, permitirá não só estruturar a abordagem metodológica, como também garantir que os resultados obtidos sejam sólidos, reproduzíveis e alinhados com os objetivos estratégicos do setor elétrico.

3.3 Descrição do Dataset

Como base de dados para o desenvolvimento, será utilizado um conjunto de dados históricos com registos horários do mercado de eletricidade, abrangendo o período de 2020 a

2025. Estes dados provêm de sistemas de gestão de mercado organizados por operadores regionais, sendo a fonte o **ISO New England** (ISO New England, 2025), uma organização de transmissão regional independente e sem fins lucrativos, com sede em *Holyoke, Massachusetts*. Esta entidade é responsável pela coordenação do sistema elétrico e dos mercados grossistas de eletricidade nos estados de Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island e Vermont, nos Estados Unidos.

A utilização de dados do *ISO New England* (ISO New England, 2025) justifica-se pela sua disponibilidade pública, qualidade e estruturação, fatores que contrastam com as restrições de acesso e privacidade associadas aos dados detalhados do mercado ibérico. Embora o foco do presente estudo seja a previsão de preços no contexto ibérico, os padrões operacionais e estruturais dos mercados elétricos organizados seguem princípios semelhantes em muitos sistemas liberalizados, como os mercados marginalistas, permitindo assim a aplicação dos métodos preditivos num cenário representativo e rigoroso.

Assim, este dataset inclui as seguintes variáveis:

- *DA_Demand* e *RT_Demand*: correspondem, respetivamente, à procura prevista no mercado diário (*Day-Ahead Demand*) e à procura observada em tempo real (*Real-Time Demand*), ambas expressas em megawatts. A *DA_Demand* resulta da soma de ofertas aceites no mercado diário, incluindo ofertas fixas e sensíveis ao preço, enquanto a *RT_Demand* traduz a procura efetiva do sistema no momento, incorporando a produção medida, consumos líquidos e a resposta da procura agregada;
- *DA_LMP* e *RT_LMP*: referem-se aos preços marginais locais de eletricidade, no mercado diário (*Day-Ahead Locational Marginal Price*) e no mercado em tempo real (*Real-Time Locational Marginal Price*), respetivamente, expressos em dólares por megawatt-hora. Estes preços refletem o custo marginal de fornecimento de energia em cada nó ou zona da rede elétrica, considerando restrições físicas e operacionais do sistema. O *Locational Marginal Price* é amplamente utilizado em mercados liberalizados para determinar preços e garantir sinais de investimento adequados na produção e na rede;
- *DA_EC*, *DA_CC*, *DA_MLC* e as respetivas métricas equivalentes em tempo real constituem a decomposição do preço marginal local de eletricidade nos seus principais componentes:
 - *DA_EC/RT_EC*: componente da energia, que representa o custo marginal de produção;
 - *DA_CC/RT_CC*: componente da congestão, refletindo custos adicionais associados a restrições de transmissão e redistribuição de fluxos na rede;
 - *DA_MLC/RT_MLC*: componente das perdas marginais, que traduz o custo incremental associado às perdas elétricas no transporte de energia.
- *System_Load*: indica a carga total do sistema em cada hora, medida em megawatts. Trata-se da soma da produção distribuída e dos consumos líquidos registados, representando a procura efetiva do sistema;
- *Dry_Bulb* e *Dew_Point*: representam variáveis meteorológicas sobre a temperatura do ar, sem considerar os efeitos da humidade, e o ponto de orvalho, temperatura à qual o

ar, com a sua atual pressão e conteúdo de vapor de água, se tornaria saturado, ou seja, com 100% de humidade relativa, respetivamente. Estas duas métricas são representadas em graus Fahrenheit;

- *Reg_Service_Price*: corresponde ao preço de liquidação do serviço de regulação de mercado, expresso em dólares por megawatt-hora. Este valor reflete o custo pago aos produtores ou consumidores que ajustam a sua inserção ou remoção de energia em tempo real, contribuindo para manter a frequência do sistema e corrigir desequilíbrios entre a procura e a oferta;
- *Reg_Capacity_Price*: representa o preço da capacidade de regulação disponibilizada ao operador do sistema, em dólares por megawatt-hora. Ao contrário do serviço efetivamente prestado, este valor indica a disponibilidade de recursos que podem ser ativados caso seja necessário regular a rede. É, portanto, um indicador da flexibilidade do sistema em termos de reservas de regulação;
- *Min/Max 5min RSP*: correspondem, respetivamente, ao menor e ao maior preço do serviço de regulação observados no mercado em intervalos de cinco minutos dentro de cada hora. Estes indicadores captam a variabilidade intra-horária do custo da regulação, refletindo a rapidez com que os preços podem oscilar em resposta a desajustes súbitos entre a produção e o consumo;
- *Min/Max 5min RCP*: referem-se ao preço mínimo e máximo da capacidade de regulação registados em intervalos de cinco minutos dentro da hora. Estes valores permitem observar a variação do indicador da disponibilidade de recursos para regulação em curtos períodos, sendo assim representado como um indicador da liquidez e flexibilidade da rede elétrica.

Estes dados apresentam uma granularidade horária, essencial para captar as flutuações de curto prazo que caracterizam o mercado spot. A presença simultânea de variáveis energéticas, meteorológicas e operacionais confere ao *dataset* um carácter multivariado, adequado à aplicação de técnicas avançadas de *Machine Learning* orientadas para séries temporais. Antes da sua utilização em modelos preditivos, o conjunto de dados será submetido a uma fase rigorosa de pré-processamento, onde se procederá o tratamento de valores em falta, normalização, criação de *lags* e tratamento de variáveis temporais. Adicionalmente, serão avaliadas correlações entre variáveis para suportar a seleção de atributos relevantes. É importante salientar que a dimensão do *dataset* permitirá testar a generalização dos modelos propostos a dados não vistos durante o treino, refletindo a sua robustez em contextos de elevada variabilidade. Esta abordagem temporal também será essencial para avaliar o impacto de fenómenos mais recentes no mercado elétrico, como a evolução da introdução das energias renovável ou alterações regulatórias.

3.4 Pré-processamento de Dados

Esta secção tem como objetivo descrever de forma sistemática os procedimentos realizados antes da aplicação de modelos de *Machine Learning*. A análise incide sobre a variável

RT_LMP, Real-Time Locational Marginal Price, selecionada como variável-alvo para o presente estudo.

A secção encontra-se dividida em duas subsecções principais. Na subsecção 3.4.1, designada Análise de Dados, procede-se à caracterização exploratória do conjunto de dados, identificando padrões temporais, correlações e potenciais problemas como valores em falta ou observações atípicas. Na subsecção 3.4.2, dedicada ao Tratamento de Dados, são descritas as técnicas de pré-processamento aplicadas, incluindo normalização, criação de variáveis derivadas, codificação de variáveis temporais e estratégias para lidar com valores em falta e *outliers*.

3.4.1 Análise de Dados

A análise exploratória de dados constitui a primeira etapa da pré-processamento do *dataset* e tem como finalidade caracterizar as variáveis disponíveis, detetar padrões estruturais e identificar potenciais limitações que possam influenciar o desempenho dos modelos preditivos.

3.4.1.1 Análise Estatística

A análise estatística das variáveis evidencia diferentes escalas e padrões de variação entre os três grupos de dados, energéticos, meteorológicos e operacionais. Os resultados estão representados na Tabela 10.

Tabela 10 – Resultados Análise estatística

Variável	Média	Mediana	Mínimo	Máximo	Q1	Q3	Desvio Padrão
<i>RT_LMP</i>	48.14	34.03	-150.91	2254.35	22.17	56.29	47.28
<i>DA_LMP</i>	49.12	35.77	6.33	374.98	32.35	58.03	39.86
<i>RT_Demand</i>	13147.95	12842.13	7557.68	25424.07	11682.51	14625.48	2602.34
<i>DA_Demand</i>	12882.73	12644.70	6417.30	25352.70	10966.65	14301.18	2557.08
<i>System_Load</i>	13375.60	13068.38	7711.00	25801.00	11484.21	14844.00	2632.80
<i>RT_MLC</i>	0.14	0.08	-0.26	15.03	0.00	0.21	0.29
<i>DA_MLC</i>	0.14	0.08	-2.01	5.98	0.00	0.21	0.29
<i>Dry_Bulb</i>	51.36	51.00	-10.0	95.0	37.0	66.0	21.68
<i>Dew_Point</i>	39.42	40.0	-29.0	76.0	25.0	55.0	18.81
<i>Max_5min_RCP</i>	31.47	14.69	0.0	3812.42	8.32	33.71	62.91
<i>Min_5min_RCP</i>	11.54	7.47	0.0	1169.73	1.43	12.24	20.93
<i>Max_5min_RSP</i>	0.28	0.01	0.0	17.78	0.0	0.01	0.83
<i>Min_5min_RSP</i>	0.08	0.01	0.0	7.60	0.0	0.01	0.43
<i>Reg_Capacity_Price</i>	20.43	10.71	0.0	2354.12	6.66	20.97	36.22
<i>Reg_Service_Price</i>	0.17	0.16	0.0	10.0	0.01	0.19	0.55

No grupo das variáveis energéticas, representado por *RT_LMP*, *DA_LMP*, *RT_Demand*, *DA_Demand*, *System_Load*, *RT_MLC* e *DA_MLC*, observa-se que os preços marginais horários apresentam forte variabilidade. O *RT_LMP* apresenta média de cerca de 48 dólares por megawatt-hora e mediana de 34, o que indica uma distribuição assimétrica à direita, marcada por valores extremos positivos, máximo de 2254, mas também negativos, como mínimo de -150, correspondentes a situações de oferta excessiva ou eventos excepcionais de mercado. O *DA_LMP* revela valores semelhantes, embora com menor dispersão. Já a procura, *RT_Demand* e *DA_Demand*, e a carga total, *System_Load*, registam valores médios próximos dos 13 mil megawatts, com amplitudes significativas entre mínimos de 7500–7700 MW e máximos acima de 25 mil MW, refletindo variações sazonais e diárias no consumo de eletricidade. As componentes de perdas marginais, *RT_MLC* e *DA_MLC*, apresentam valores médios muito reduzidos, 0,14 ambas, com alguns picos acima de 5 unidades, confirmando a sua contribuição marginal, mas não negligenciável, para a formação dos preços.

Sobre as variáveis meteorológicas, a temperatura do ar, *Dry_Bulb*, apresenta média de 51 °F, com variações entre -10 °F no inverno e 95 °F no verão, evidenciando a influência das estações do ano. O ponto de orvalho, *Dew_Point*, varia entre -29 °F e 76 °F.

No grupo das variáveis operacionais, *Max_5min_RCP*, *Min_5min_RCP*, *Max_5min_RSP*, *Min_5min_RSP*, *Reg_Capacity_Price* e *Reg_Service_Price*, os preços associados aos serviços de regulação e capacidade apresentam maior dispersão relativa. O *Max_5min_RCP* atinge valores muito elevados, com um máximo superior a 3800 dólares por megawatt-hora, revelando episódios de forte escassez de capacidade em curtos intervalos temporais, enquanto o valor mediano permanece relativamente baixo, 14,7. O mesmo ocorre com o *Reg_Capacity_Price*, cujo máximo ultrapassa 2300, mas a mediana é apenas de 10,7. Em contraste, os preços do serviço de regulação, *Reg_Service_Price*, mantêm-se bastante estáveis e baixos, com média de 0,17, refletindo uma menor volatilidade.

3.4.1.2 Sazonalidade e Tendências

Esta análise temporal permite identificar padrões recorrentes que caracterizam a procura, os preços e a carga do sistema. No ciclo diário representado na Figura 16, observa-se uma procura mínima de madrugada, entre as 3h e as 5h, período em que os consumos residenciais e industriais são reduzidos. A partir da manhã, há uma subida progressiva da procura, acentuando-se ao longo da tarde até atingir o pico entre as 18h e 19h, coincidindo com o retorno das atividades domésticas e comerciais. Após este máximo, a procura decresce gradualmente durante a noite.

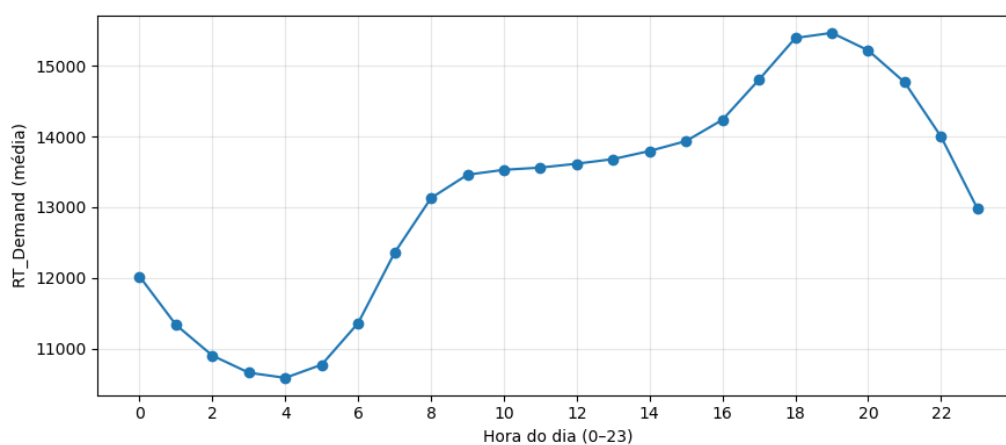


Figura 16 – Ciclo diário médio de *RT_Demand*

O ciclo semanal demonstrado na Figura 17 mostra diferenças estruturais entre dias úteis e fim de semana. De segunda a sexta-feira, a carga do sistema mantém-se relativamente estável, refletindo os perfis de consumo industrial e comercial. Aos sábados e domingos, observa-se uma redução significativa, sobretudo no *System_Load*, sinalizando menor atividade económica e industrial.

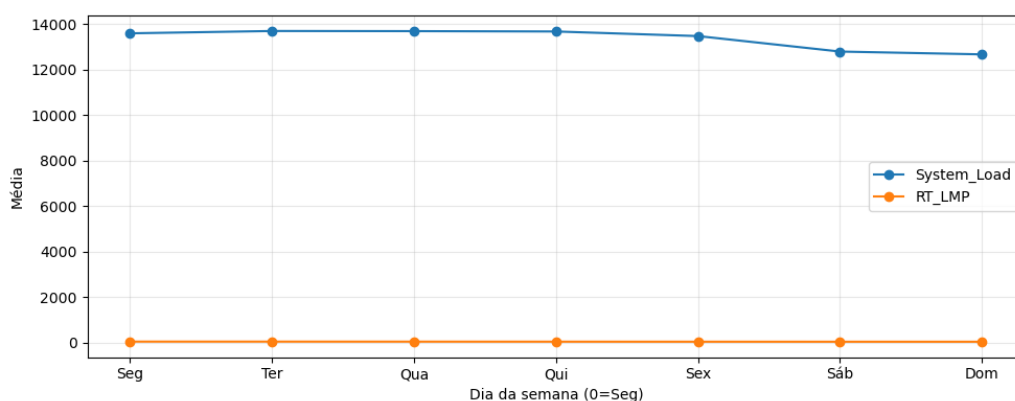


Figura 17 - Ciclo semanal médio de *RT_LMP* e *System_Load* por dia da semana

Este comportamento é confirmado pelos valores médios apresentados na Tabela 11, onde se verifica que, no início da semana, o *System_Load* se mantém em torno dos 13.600 MW, enquanto no final da semana, nomeadamente aos domingos, a média desce para valores próximos de 12.700 MW. Já o *RT_LMP* acompanha estas oscilações de forma menos acentuada, apresentando uma descida gradual ao longo da semana, de cerca de 51 \$/MWh para valores próximos de 45 \$/MWh.

Tabela 11 - Médias semanais do *System_Load* e do *RT_LMP*

Dia da semana	System_Load (MW)	RT_LMP (\$/MWh)
Segunda	13 599.58	51.06
Terça	13 701.65	50.66
Quarta	13 695.56	49.40
Quinta	13 680.27	48.08
Sexta	13 478.54	46.49
Sábado	12 798.13	45.90
Domingo	12 674.35	45.39

A análise das médias mensais e móveis a 30 dias evidencia a presença de ciclos anuais regulares tanto na procura real de eletricidade, *RT_Demand*, como na temperatura do ar, *Dry_Bulb*. Na Figura 18 observa-se que as duas séries apresentam oscilações de natureza semelhante, com variações cíclicas que acompanham as alterações climáticas das estações do ano. Durante os meses mais quentes registam-se valores mais elevados de temperatura e, em paralelo, uma tendência de aumento da procura. Nos meses mais frios, ambas as variáveis reduzem os seus valores médios, no entanto a *RT_Demand* superior à *Dry_Bulb*, devido à necessidade de energia para superar o inverno, verificando-se também maior dispersão nos consumos elétricos.

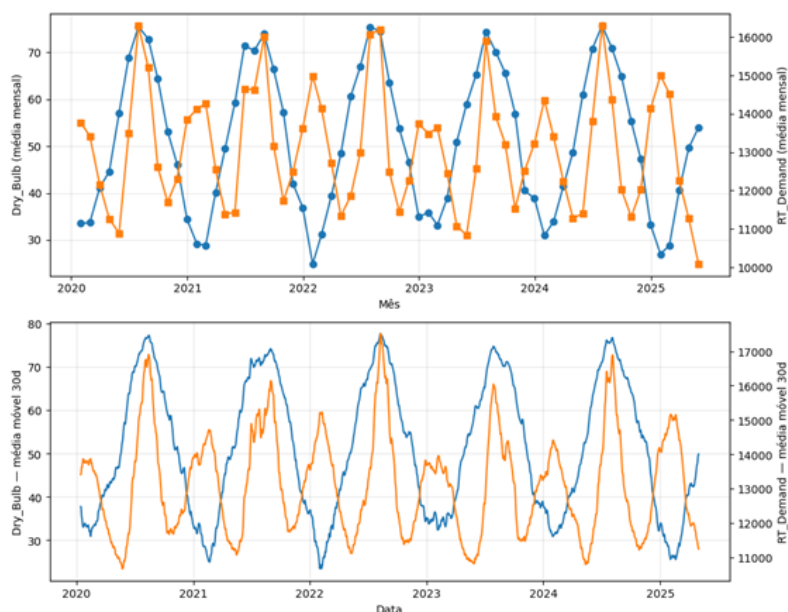


Figura 18 - Médias mensais e móveis a 30 dias de *Dry_Bulb* e *RT_Demand*

Assim, a realização prévia desta análise temporal contribui para uma seleção mais detalhada de atributos e facilita o processo de afinação dos hiperparâmetros dos modelos, ao fornecer indicações sobre as janelas temporais mais relevantes a considerar.

3.4.1.3 Correlação entre variáveis

A análise de correlação entre as variáveis energéticas, meteorológicas e operacionais permite identificar relações significativas com a variável de previsão *RT_LMP*. A correlação presente na análise é a de *Pearson* com o objetivo de identificar relações lineares entre as variáveis. Os resultados são apresentados através de duas visualizações: a *heatmap* da matriz de correlação e a escala de correlação com a variável *RT_LMP*.

O *heatmap* da matriz de correlação, representado na Figura 19, revela as correlações globais entre todas as variáveis do *dataset*. As correlações mais fortes estão presentes entre a *RT_Demand* e a *DA_Demand*, o *DA_LMP* e a *DA_EC*, o *RT_LMP* e a *RT_EC*, a *Dry_Bulb* e o *Dew_Point*, a *System_Load* com a *RT_Demand* e a *DA_Demand*. As correlações mais fracas estão representadas entre a *Dry_Bulb* e o *Dew_Point*, e a *RT_MLC*.

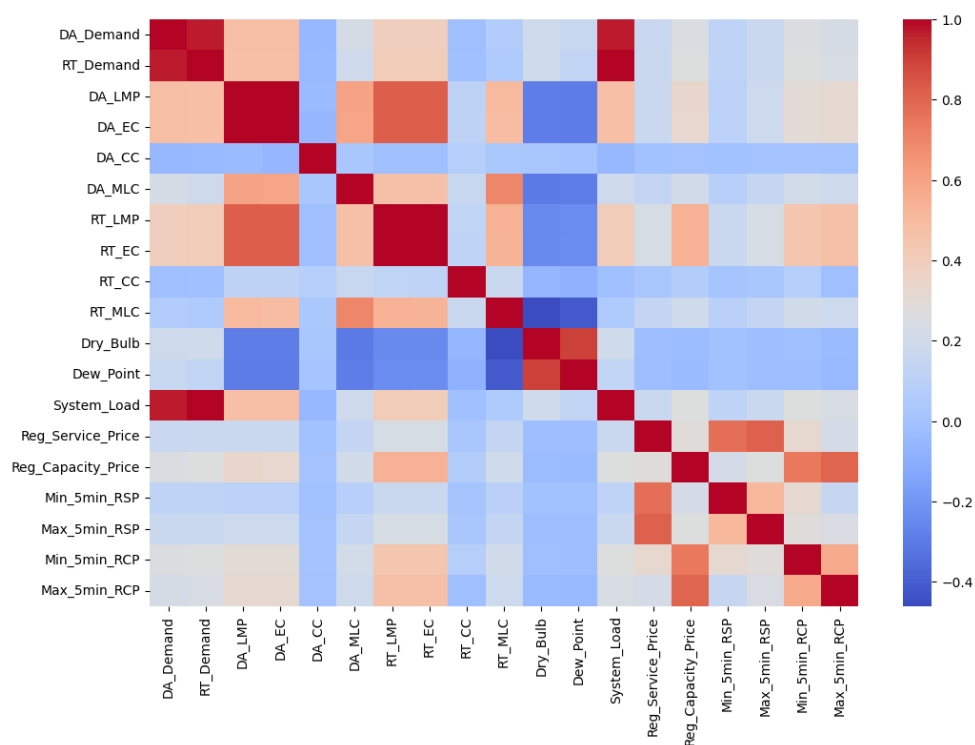


Figura 19 – Matriz de Correlação

Na Figura 20 está presente a escala de correlação com o *RT_LMP* que destaca as variáveis mais influentes sobre o preço. A *RT_EC* apresenta uma correlação forte com o *RT_LMP*, de 0.99996, refletindo uma dependência entre os preços e a componente de energia em tempo real. O *DA_LMP* e a *DA_EC* também mostram uma forte correlação com o *RT_LMP*, de 0.81886 e 0.81864, respectivamente, concluindo que os preços e a componente de energia, do mercado diário têm um impacto significativo nos preços em tempo real. Outras variáveis energéticas como a *RT_MLC*, com 0.54197, o *Reg_Capacity_Price*, com 0.54018, e a *DA_MLC* com 0.47492, revelam correlações moderadas. A variável *System_Load* (0.41011) também apresenta correlação moderada com o *RT_LMP*, destacando a importância da carga do sistema na determinação do preço.

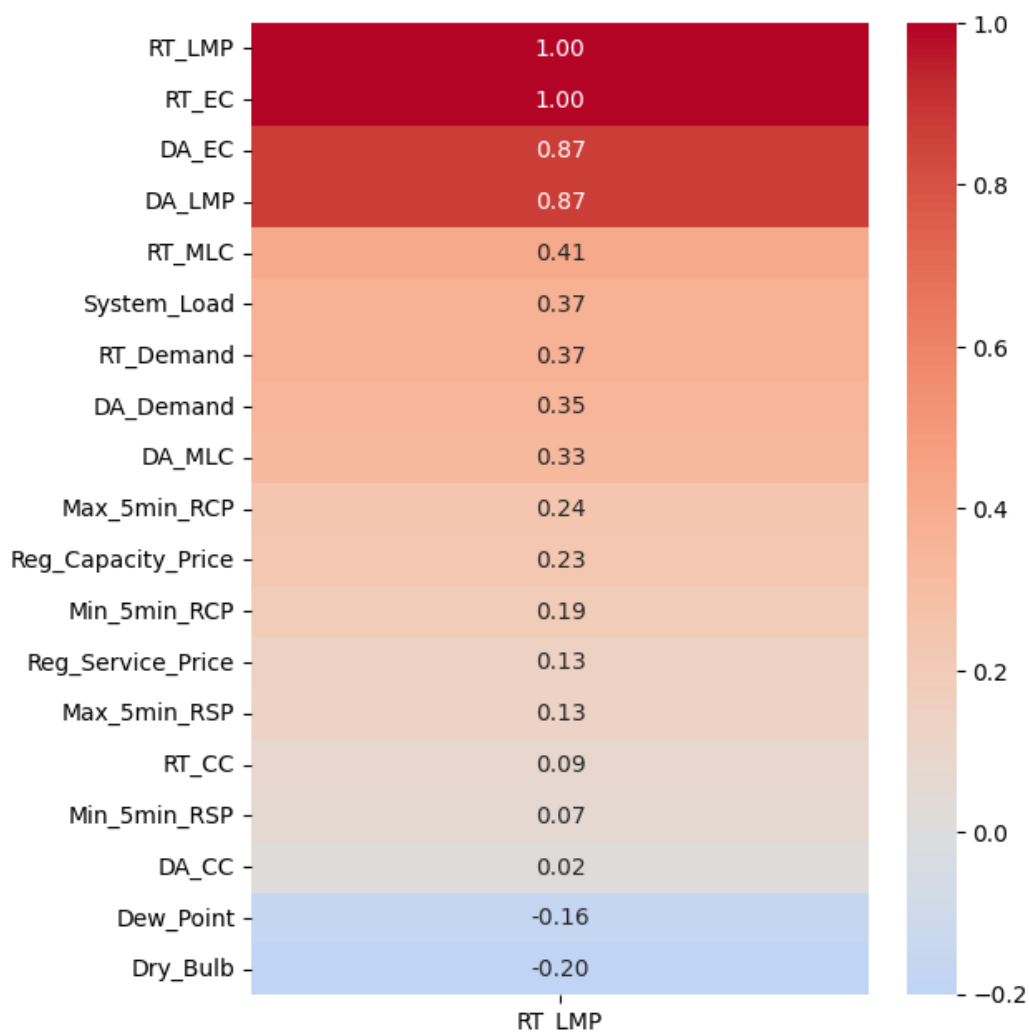


Figura 20 - Escala de correlação com RT_LMP

Por outro lado, as variáveis *Reg_Service_Price* (0.23512), *Max_5min_RSP* (0.23387), *Min_5min_RSP* (0.16785) e *RT_CC* (0.12274) mostram correlações mais fracas, mas ainda assim contribuem para a dinâmica do mercado. As variáveis *DA_CC* (-0.0195), *Dew_Point* (-0.24016) e *Dry_Bulb* (-0.24866) apresentam correlações negativas. No entanto, independentemente da correlação, as variáveis meteorológicas são úteis para os modelos identificarem padrões de sazonalidade no *dataset*.

3.4.1.4 Análise de valores em falta e duplicados

Na verificação da sequência temporal do *dataset* foram identificadas inconsistências que afetam a sua continuidade. Constatou-se a existência de um registo duplicado no instante de 3 de novembro de 2024 às 02:00 e a ausência de dois registos, correspondentes a 10 de março de 2024 às 02:00 e a 9 de março de 2025 às 02:00. Optou-se por manter a primeira ocorrência do registo duplicado e eliminar a repetição, assegurando a consistência da série. Relativamente às linhas em falta, foi aplicada o método de interpolação, de modo a reconstruir os valores

ausentes com base na continuidade da evolução das variáveis. Esta abordagem permitiu restabelecer a coerência da série temporal e preservar a integridade dos dados para utilização nos modelos de previsão.

3.4.2 Tratamento de Dados

Após a Análise de dados inicial, procede-se o tratamento de dados para garantir a consistência e a utilidade do *dataset* nos modelos de previsão. Este tratamento inclui a identificação e tratamento de *outliers*, a criação de variáveis derivadas através de histórico temporal, a codificação de variáveis temporais para representar a sua natureza cíclica, a seleção de variáveis relevantes e, por fim, a normalização e o escalonamento das variáveis.

3.4.2.1 Detecção e Tratamento de outliers

A identificação de *outliers* foi realizada com base no método do Intervalo Interquartil (IQR), que considera como valores anómalos todas as observações que se encontram abaixo de $Q1 - 1,5 \times IQR$ ou acima de $Q3 + 1,5 \times IQR$, onde $Q1$ e $Q3$ representam, respetivamente, o primeiro e terceiro quartis. Os resultados foram organizados em *boxplots*, representados para diferentes grupos de variáveis segundo a sua escala de variação, como apresentado na Figura 21.

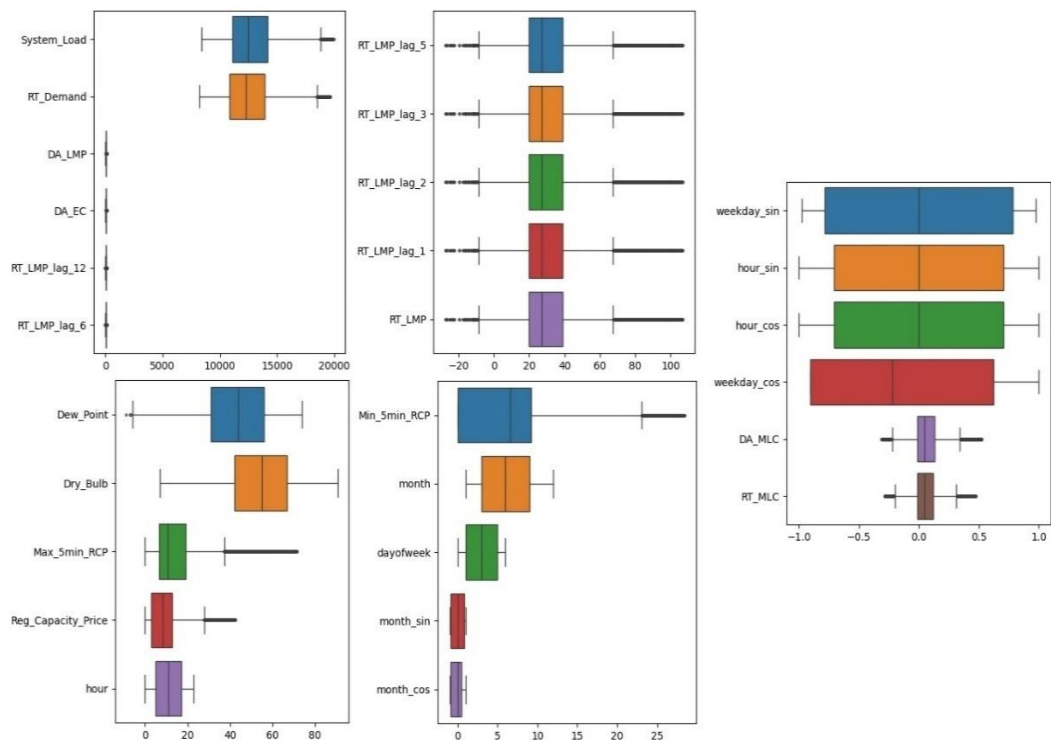


Figura 21 - Boxplots de Outliers

Através da análise da Figura 21 conclui-se que variáveis relacionadas com a procura e a carga do sistema (*RT_Demand*, *DA_Demand*, *System_Load*) apresentam um número elevado de *outliers*, refletindo picos de consumo ou situações de procura abruptas. As variáveis de preços e componentes elétricas (*RT_LMP*, *DA_LMP*, *RT_EC*, *DA_EC*) apresentam igualmente dispersões significativas, associadas a oscilações pontuais do mercado. Nas variáveis meteorológicas (*Dry_Bulb* e *Dew_Point*), os *outliers* são menos frequentes, mas relacionam-se a condições climáticas extremas.

Para complementar esta análise, na Tabela 12 está presente a contagem de *outliers* por variável, a qual permite quantificar objetivamente a incidência destes valores no *dataset*.

Tabela 12 – Número de *Outliers* por Variável

Variável	Nº de Outliers
<i>DA_CC</i>	10 824
<i>RT_CC</i>	9 679
<i>Min_5min_RSP</i>	7 480
<i>Max_5min_RSP</i>	5 857
<i>Reg_Capacity_Price</i>	4 974
<i>Max_5min_RCP</i>	4 740
<i>RT_MLC</i>	4 421
<i>Reg_Service_Price</i>	3 998
<i>RT_EC</i>	3 857
<i>RT_LMP</i>	3 855
<i>DA_EC</i>	3 715
<i>DA_LMP</i>	3 707
<i>Min_5min_RCP</i>	3 501
<i>DA_MLC</i>	3 414
<i>DA_Demand</i>	1 093
<i>RT_Demand</i>	1 085
<i>System_Load</i>	1 067
<i>Dew_Point</i>	19
<i>Dry_Bulb</i>	9

No entanto, os *outliers* encontrados não foram removidos, dado que estes podem refletir padrões relevantes do mercado elétrico, como situações de congestionamento, volatilidade extrema de preços ou variações súbitas na procura. Assim, optou-se por preservar estes registos do *dataset*, garantindo que os modelos preditivos captam adequadamente o comportamento real e os episódios de maior variabilidade do sistema.

3.4.2.2 Criação de variáveis derivadas e variáveis temporais

Com o objetivo de capturar melhor as dependências temporais presentes no mercado de eletricidade, foram criadas variáveis derivadas a partir do histórico temporal das séries originais. Este procedimento consistiu na introdução de deslizamentos temporais da variável alvo *RT_LMP*. A inclusão destas transformações permite que os modelos incorporem informação proveniente de valores passados, promovendo a identificação de padrões de curto e médio prazo. Assim, foram considerados os seguintes deslizamentos temporais, expressos em horas: 1, 2, 3, 5, 6 e 12. A inclusão destes intervalos foi necessária para capturar dependências de curta duração, tendo como objetivo o modelo capturar melhores padrões da variável alvo *RT_LMP*.

Para além da criação de variáveis derivadas, procedeu-se à criação explícita de atributos temporais com o objetivo de representar a estrutura cíclica do tempo no *dataset*. Para tal, foram extraídas das séries temporais as seguintes componentes: hora do dia, dia da semana e mês do ano. Estas variáveis foram posteriormente transformadas através das funções trigonométricas seno e cosseno, de forma a preservar a sua natureza circular.

3.4.2.3 Seleção de variáveis

A seleção de variáveis tem como finalidade reduzir a dimensão do *dataset*, eliminando redundâncias e resíduos, para preservar apenas as variáveis com maior contributo para a previsão do *RT_LMP*.

Através da análise de correlações entre as variáveis, foram identificados e removidos atributos altamente correlacionados que poderiam introduzir problemas de multicolinearidade. Adicionalmente, foram considerados padrões de sazonalidade e características teóricas com o objetivo de captar relações não lineares e dependências complexas. Com isto em consideração, foram descartadas as variáveis que tinham menor correlação com a variável alvo. Assim, optou-se pela exclusão das seguintes variáveis: *DA_LMP*, *RT_EC*, *DA_EC*, *Reg_Capacity_Price*, *Reg_Service_Price*, *Max_5min_RSP*, *Min_5min_RSP*, *RT_CC* e *DA_CC*.

Este procedimento tem como objetivo fazer com que os modelos de previsão consigam capturar os padrões necessários na sua aprendizagem e evitar o *overfitting*, conseguindo assim generalizar corretamente.

3.4.2.4 Normalização dos dados

Definido o conjunto final de variáveis, procedeu-se à normalização dos dados, com o objetivo de colocar todas as variáveis na mesma ordem de grandeza e evitar que atributos com magnitudes mais elevadas prevaleçam no processo de treino dos modelos.

Para este efeito, foi utilizado o método de normalização *StandardScaler*, da biblioteca *Scikit-learn* (Scikit-learn, 2025), aplicado tanto às variáveis explicativas como à variável alvo. Este procedimento transforma cada variável segundo a seguinte expressão:

$$z = \frac{x - \mu}{\sigma} \quad (9)$$

Onde x representa o valor original, μ a média da variável e σ o respetivo desvio padrão. Neste caso, a normalização foi ajustada com base nos dados de treino, assegurando que a informação estatística das fases de validação e teste não contaminava o processo. Os mesmos parâmetros de transformação foram posteriormente aplicados aos subconjuntos de validação e de teste, garantindo consistência entre todos os conjuntos de dados.

Adicionalmente, procedeu-se também à normalização da variável alvo. Embora não seja necessário em modelos lineares clássicos, revelou-se vantajoso para algoritmos mais sensíveis à escala dos dados, como redes neuronais e métodos baseados em gradiente. Ao reduzir a variância do alvo, melhora-se a estabilidade numérica durante o treino e obtém-se uma convergência mais eficiente, sem comprometer a interpretabilidade dos resultados após a inversão da transformação.

Este procedimento contribui para melhorar a estabilidade numérica dos algoritmos e acelerar a convergência nos métodos de otimização, além de garantir que os modelos não são influenciados por diferenças de escala entre variáveis heterogêneas, como preços, cargas e indicadores meteorológicos.

3.4.2.5 Divisão do *dataset* e Criação de janelas temporais

Para o treino e previsão nos modelos, o conjunto de dados foi dividido em 70% para treino, 15% para validação e 15% para teste, respeitando a ordem temporal das observações, de forma a evitar fuga de informação entre subconjuntos.

No contexto de séries temporais, a previsão de valores futuros requer a transformação do *dataset* em sequências de observações passadas que servem de entrada para os modelos de redes neuronais. Para este efeito, foi implementada a técnica de janelas temporais deslizantes, que consiste em organizar a série em blocos de tamanho fixo, cada um representando um histórico de observações consecutivas.

Este procedimento foi aplicado à variável alvo *RT_LMP* e às variáveis exógenas selecionadas, garantindo que o modelo dispõe de informação suficiente sobre a evolução recente do sistema para prever o valor futuro. Após a realização da análise temporal e de testes preliminares, foi selecionada uma dimensão de 48 horas para a janela, uma vez que permite captar dois ciclos diários completos de variação, os quais refletem padrões intradiários recorrentes no mercado de eletricidade. A criação de janelas temporais permite, assim, estruturar os dados de forma compatível com modelos sequenciais, redes neuronais recorrentes, assegurando que a dimensão temporal da informação é devidamente incorporada no processo de previsão.

4 Avaliação de Modelos

Neste capítulo está presente o processo de definição da arquitetura e de afinação dos hiperparâmetros dos modelos testados, de forma a maximizar o seu desempenho. Seguidamente, apresenta-se a escolha dos critérios de avaliação adotados, englobando as métricas de erro e os procedimentos de validação utilizados para medir a qualidade das previsões. Por fim, procede-se à discussão comparativa dos resultados obtidos, destacando os pontos fortes e limitações no enquadramento do mercado elétrico.

4.1 Afinação dos hiperparâmetros

A hiperparametrização realizou-se através da técnica *GridSearch*, que permite encontrar as melhores combinações de hiperparâmetros para os modelos testados. O objetivo foi otimizar os modelos para maximizar o seu desempenho no conjunto de validação. Para tal, na Tabela 13 estão representados os parâmetros utilizados pelo *GridSearch*:

Tabela 13 – Parâmetros utilizados pelo *GridSearch*

Modelo	Hiperparâmetros	Valores testados
<i>Árvore de Decisão</i>	max_depth	[5, 10, 15, None]
	min_samples_split	[2, 5, 10]
<i>Random Forest</i>	n_estimators	[50, 100, 200]
	max_depth	[5, 10, 15, None]
	min_samples_split	[2, 5, 10]
<i>XGBoost</i>	n_estimators	[100, 200]
	max_depth	[3, 5, 6, 7]
	learning_rate	[0.01, 0.1, 0.2, 0.5]
<i>LightGBM</i>	n_estimators	[100, 200]
	max_depth	[5, 10, -1]
	learning_rate	[0.01, 0.1, 0.2, 0.5]
<i>CatBoost</i>	iterations	[100, 200]
	depth	[4, 6, 8]
	learning_rate	[0.01, 0.03, 0.1, 0.2]

Os modelos de Regressão Linear e Redes Neurais não foram incluídos no processo de *GridSearch*. Como a Regressão Linear obtém melhor desempenho através da regularização, optou-se por manter o desempenho deste modelo com a parametrização *default*. As redes neurais exigem um processo de parametrização mais complexo, logo a otimização dos seus hiperparâmetros foi realizada separadamente.

No entanto, para avaliar o desempenho das RNNs, foi decidido manter a parametrização para todos os modelos, *SimpleRNN*, *LSTM* e *GRU*, com o objetivo de permitir uma comparação direta entre as arquiteturas. Para tal, todos os modelos foram configurados com três camadas recorrentes, contendo 64, 32 e 16 unidades, respectivamente. Cada camada foi seguida de uma camada de *dropout* de 40%, com o objetivo de aplicar regularização para evitar o *overfitting*. A camada de saída foi composta por uma *dense layer* com 32 unidades e uma função de ativação ReLU, para ajudar a capturar relações não lineares, sendo que a camada final consistiu em uma unidade linear para prever a variável alvo, RT_LMP. Além disto, foram utilizadas 100 epochs, um *batch_size* de 128 e um *early stop* de 15 de *patience*, de acordo com o valor de perda. A escolha destes parâmetros resultou de uma procura prática da melhor relação entre a qualidade dos resultados obtidos e os recursos computacionais disponíveis. A configuração de camadas e unidades permite capturar padrões temporais relevantes sem exigir capacidade computacional excessiva. O dropout de 40% foi selecionado para manter a regularização adequada, equilibrando aprendizagem e prevenção de *overfitting*. O número de epochs, *batch_size* e *early stopping* foram definidos de forma a treinar as redes, garantindo estabilidade na convergência da função de perda.

Esta abordagem foi adotada para garantir uma avaliação consistente entre as diferentes arquiteturas, permitindo determinar qual destas apresentaria melhor desempenho e comparar com modelos de *machine learning* mais simples e menos exigentes.

4.2 Critérios de Avaliação

A avaliação quantitativa do desempenho dos modelos foi realizada através de métricas de erro que medem a discrepância entre os valores previstos e os valores reais. No presente estudo, foram utilizadas as seguintes métricas: Erro Médio Absoluto (MAE), Erro Quadrático Médio (MSE), Raiz do Erro Quadrático Médio (RMSE), o Coeficiente de determinação (R^2) e Erro Percentual Absoluto Médio (MAPE).

4.2.1 Erro Médio Absoluto (MAE)

O MAE mede a magnitude média dos erros absolutos entre previsões e observações, sendo definido como:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (10)$$

Onde y_i corresponde ao valor real e \hat{y}_i ao valor previsto. Esta métrica expressa, em termos absolutos, o desvio médio das previsões face à realidade, sendo de mais fácil interpretação e implementação relativamente às outras métricas (Jadon, Nain & Choudhary, 2022; Hodson, T. O, 2022).

4.2.2 Erro Quadrático Médio (MSE) e Raiz do Erro Quadrático Médio (RMSE)

O MSE calcula a média dos quadrados dos erros, sendo penalizado de forma mais acentuada através de desvios de maior magnitude. O RMSE é a raiz quadrada do MSE, frequentemente utilizada por estar na mesma unidade que a variável em análise. Estas métricas são calculadas através das seguintes expressões:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (11)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (12)$$

Enquanto o MSE é útil para captar grandes erros, o RMSE facilita a interpretação prática ao expressar o erro médio em termos comparáveis com o valor real (Jadon, Nain & Choudhary, 2022; Hodson, T. O., 2022).

4.2.3 Coeficiente de determinação (R^2)

O R^2 mede a proporção da variabilidade da variável prevista pelo modelo de acordo com variáveis independentes. Assim, é calculado através da seguinte expressão:

$$R^2 = 1 - \frac{\sum_{i=1}^m (y_i - \hat{y}_i)^2}{\sum_{i=1}^m (y_i - \bar{y})^2} \quad (13)$$

Este coeficiente normalmente varia entre 0 e 1, onde 1 indica que um modelo prevê os dados perfeitamente, e 0 que o modelo não teria a capacidade para prever a variável alvo com base nas variáveis independentes. Caso esta métrica seja negativa, o modelo prevê pior do que uma média dos dados (Chicco, Warrens & Jurman, 2021).

4.2.4 Erro Percentual Absoluto Médio (MAPE)

O MAPE expressa o erro médio em termos relativos, ou seja, em percentagem do valor observado. É calculado através da seguinte expressão:

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i} \cdot 100\% \quad (14)$$

Esta métrica permite avaliar a precisão de previsão de forma normalizada. Como é uma métrica percentual é independente da escala das variáveis originais. Contudo, apresenta limitações ao criar distorções quando os valores reais são muito próximos de zero (Jadon, Nain & Choudhary, 2022; Chatterjee & Hadi, 2012).

4.2.5 Resultados dos modelos

Os modelos foram avaliados em duas etapas: execução das parametrizações base para obter referências e definir os intervalos de procura e a parametrização pelo *GridSearch* aplicado aos modelos baseados em árvores e *boosting*. As *RNNs* foram treinadas isoladamente, com a mesma parametrização, para permitir a comparação direta entre arquiteturas. A Tabela 14 e a Tabela 15 apresentam os melhores resultados dos modelos. As métricas de avaliação presentes são o RMSE, o MSE, o MAE, o R^2 e o MAPE.

Tabela 14 - Resultados das parametrizações base

Modelo	RMSE	MSE	MAE	R ²	MAPE
<i>Linear Regression</i>	14.6821	215.5647	7.2177	0.7249	25.4571
<i>Ridge Regression</i>	14.6822	215.5682	7.2174	0.7249	25.4553
<i>Lasso Regression</i>	14.7247	216.8173	7.1816	0.7233	25.0599
<i>Elastic Net</i>	14.9359	223.0820	7.0946	0.7153	24.5089
<i>Decision Tree</i>	17.8252	317.7395	8.3724	0.5946	27.1610
<i>Random Forest</i>	14.4250	208.0807	6.1534	0.7345	19.5857
<i>XGBoost</i>	13.3691	178.7333	6.1571	0.7716	22.3166
<i>LightGBM</i>	13.3787	178.9914	6.0084	0.7719	19.9256
<i>CatBoost</i>	13.7690	189.5867	6.3896	0.7581	23.0609

Nos modelos base, o desempenho organiza-se por modelos de *ensemble* com menores erros, seguido do *Random Forest*, modelos lineares e, por último, a árvore de decisão. Os modelos lineares, *Linear*, *Ridge*, *Lasso*, *Elastic Net*, concentram-se num valor de RMSE 14,68 e 14,93, MSE de 215.56 a 223.08, o R² entre 0,71e 0,72, o MAE com aproximadamente 7,1 e o MAPE entre 24 e 26%. A *Decision Tree* isolada apresenta valores mais elevados de erro, com aproximadamente RMSE de 17,82, MSE de 317,73, MAE de 8,37, MAPE de 27,16% e menor R² de 0,59, o que é consistente com a variância típica de uma única árvore.

Entre os métodos não lineares, destaca-se o *XGBoost*, com um RMSE de 13,36, MSE de 178,733, MAE de 6,15, R² de 0,77 e um MAPE de 22,31%, seguido do modelo *LightGBM* com resultados muito semelhantes, RMSE de 13,37, R² de 0,77 e MAPE de 19,58%, e *Catboost*, com RMSE de 13,76, R² de 0,75 e MAPE de 23,06%. O *Random Forest* situa-se entre os modelos lineares e os de *ensemble* (RMSE 14,42; MSE 208,08; MAE 6,15; R² 0,73; MAPE 19,92%), demonstrando ganhos relativamente à *Decision tree* e uma aproximação às soluções de *boosting* nas métricas de MAE e MAPE.

Assim, nas métricas RMSE e R², observa-se a seguinte ordem de desempenho, *XGBoost*, *LightGBM*, *Catboost* e *Random Forest*. No MAE, o *LightGBM* supera ligeiramente o *XGBoost*. No *Random Forest*, o MAPE apresenta o menor valor, o que indica melhor desempenho em termos percentuais nos níveis típicos do preço, sendo estes valores médios. Contudo, este modelo é mais penalizado nos extremos, refletindo erros absolutos maiores quando ocorrem picos ou quebras acentuadas.

Tabela 15 - Resultados das Redes Neurais

Modelo	RMSE	MSE	MAE	R ²	MAPE
<i>SimpleRNN</i>	19.9453	397.8179	10.2610	0.4924	43.3000
<i>LSTM</i>	18.3070	335.1463	11.4090	0.5724	42.2066
<i>GRU</i>	24.7970	614.8952	12.0917	0.2155	46.5777

Para as RNNs, com a mesma parametrização entre arquiteturas, a *LSTM* apresenta os melhores valores, um RMSE de 18,30, MSE de 335,14, MAE de 11,41, R^2 de 0,5724 e um MAPE de 42,20%. A *SimpleRNN* surge próxima, mas com desempenho inferior (RMSE de 19,94, MSE de 397,81, MAE de 12,09, R^2 de 0,49 e um MAPE de 43,30). A *GRU* evidencia resultados substancialmente mais fracos, com um RMSE de 24,79, MSE de 9,16, MAE de 4,71, R^2 de 0,21 e um MAPE de 46,57.

Para complementar as métricas de desempenho RMSE, MAE, MAPE e R^2 , avaliou-se a existência de um desvio sistemático nas previsões através de testes de hipótese aos resíduos. Aplicou-se um *t-test* onde a Hipótese nula indica que a média dos resíduos é igual a zero, e a Hipótese alternativa, que a média dos resíduos é diferente de zero, justificando a presença de diferenças significativas. Além deste, aplicou-se também o teste de *Wilcoxon signed-rank*, onde a Hipótese nula representa que a mediana dos resíduos é igual a zero e a Hipótese alternativa que a mediana dos resíduos é diferente de zero.

O *t-test* (Student, 1908) é indicado para verificar o desvio médio dos resíduos, uma vez que testa se a média dos resíduos é significativamente diferente de zero, enquanto o teste de *Wilcoxon* (Wilcoxon, 1945) serve como uma abordagem não paramétrica para avaliar o desvio da mediana. Este teste é útil quando se suspeita de desvios da normalidade ou da presença de *outliers* nos resíduos (Hollander & Wolfe, 1999). A combinação estes dois testes permite garantir a avaliação do desvio das previsões e reforçam a análise do desempenho dos modelos além das métricas utilizadas (García, Fernández, Luengo, & Herrera, 2009). Na Tabela 16 estão representados os resultados dos testes hipótese efetuados.

Tabela 16 – Resultados dos testes de hipótese

Modelo	<i>t-test p-value</i>	<i>Wilcoxon p-value</i>
<i>Linear Regression</i>	3.66E-14	8.03E-64
<i>Ridge Regression</i>	3.68E-14	7.84E-64
<i>Lasso Regression</i>	2.34E-17	3.00E-76
<i>Elastic Regression</i>	1.08E-11	3.28E-60
<i>Decision Tree</i>	3.64E-08	7.36E-18
<i>Random Forest</i>	4.50E-07	4.82E-57
<i>XGBoost</i>	1.38E-05	1.71E-44
<i>LightGBM</i>	0.002559	1.67E-24
<i>CatBoost</i>	4.23E-05	1.13E-38
<i>SimpleRNN</i>	2.53E-08	0.023882
<i>LSTM</i>	7.38E-37	6.53E-261
<i>GRU</i>	1.12E-118	2.41E-209

Os resultados demonstram que todos os modelos rejeitam as hipóteses nulas, ou seja, a média e a mediana dos resíduos são estatisticamente diferentes de zero. Relativamente ao *t-*

test, todos os *p-values* estão abaixo de 0,05, indicando que a média dos resíduos é significativamente diferente de zero. Da mesma forma, os *p-valores* obtidos com o teste de *Wilcoxon* são também extremamente baixos, reforçando a conclusão de que a mediana dos resíduos é também diferente de zero em todos os modelos analisados.

A rejeição das hipóteses nulas sugere que todos os modelos apresentam viés significativo. Assim, as previsões dos modelos não estão corretamente calibradas, apresentando um desvio sistemático nas estimativas. Embora os modelos mostrem bom desempenho geral, como os resíduos não estão centrados em zero indica que os modelos ainda podem ser melhorados em termos de calibração. Alterações, como a correção de viés, podem ser necessárias para melhorar a precisão das previsões, especialmente em contextos com alta volatilidade e variabilidade nos dados, como no caso da previsão de preços de eletricidade.

4.3 Discussão de resultados

Nos resultados obtidos, os modelos de *boosting* obtiveram os menores RMSE, MSE e MAE e os maiores R^2 . O *Random Forest* ficou abaixo destes em RMSE, MAE e R^2 , mas apresentou o menor MAPE. Os modelos lineares agruparam-se numa faixa de desempenho intermédia e a *Decision tree* foi a que teve o desempenho mais baixo. Com a mesma parametrização, as RNNs ficaram abaixo dos modelos de *ensemble*.

Como os modelos de *boosting* combinam múltiplas árvores pouco profundas com *shrinkage* e regularização, capturam não-linearidades e interações sem aumentar demasiado a variância, tendem a reduzir erros absolutos, o RMSE e MAE, e a captar mais variância, R^2 . O *Random Forest*, ao utilizar muitas árvores sem *boosting*, reduz variância, mas não corrige sistematicamente os resíduos. Assim, erra menos em termos percentuais nos níveis médios, tendo um melhor MAPE, mas acumula erros absolutos maiores nos extremos, piores RMSE e MAE, o que é prejudicial em séries de preços com flutuações maiores. Os modelos lineares captam tendências médias e relações aproximadamente lineares, mas não identificam bem padrões intrínsecos na serie temporal. As RNNs ficaram atrás dos modelos de *boosting* com a janela temporal de 48 horas e a parametrização comum entre estas. Redes neuronais recorrentes tendem a exigir mais recursos e maior sensibilidade na escolha de camadas e na sua parametrização. Além disto, exigem um treino mais exigente para captarem melhor padrões e tendências na serie temporal.

De acordo com a Figura 22 e a Figura 23, e definido como critério de seleção minimizar o RMSE e maximizar o R^2 , dada a relevância de limitar erros absolutos em séries de preços com elevada volatilidade, o *XGBoost* foi escolhido como modelo final. Este modelo foi reutilizado para avaliação com os dados de teste.

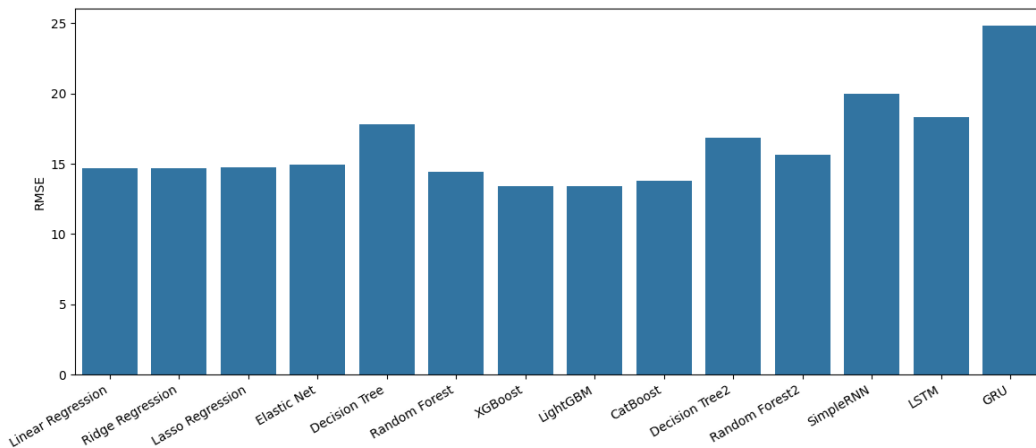


Figura 22 - RMSE por modelo

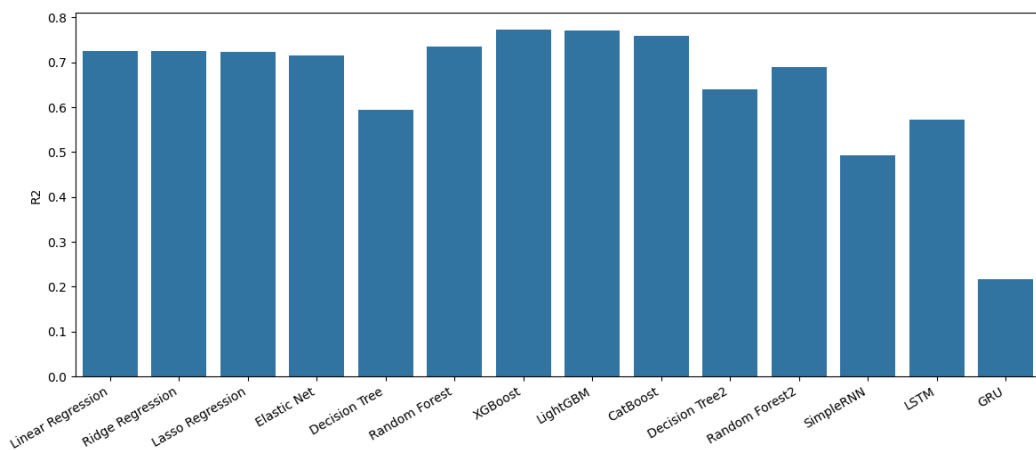


Figura 23 – R² por modelo

No teste, a série prevista acompanha a estrutura temporal do preço, com maiores desvios concentrados em episódios de volatilidade extrema, como representado nos picos da Figura 24. A curva de aprendizagem evidencia estabilização da perda de validação, como consta na Figura 25. A análise de resíduos, representada na Figura 26, demonstra resíduos centrados em zero e uma dispersão crescente para níveis elevados de preço, o que significa que apesar do modelo ter os RMSE e R² melhores, este continua a errar mais em valores atípicos.

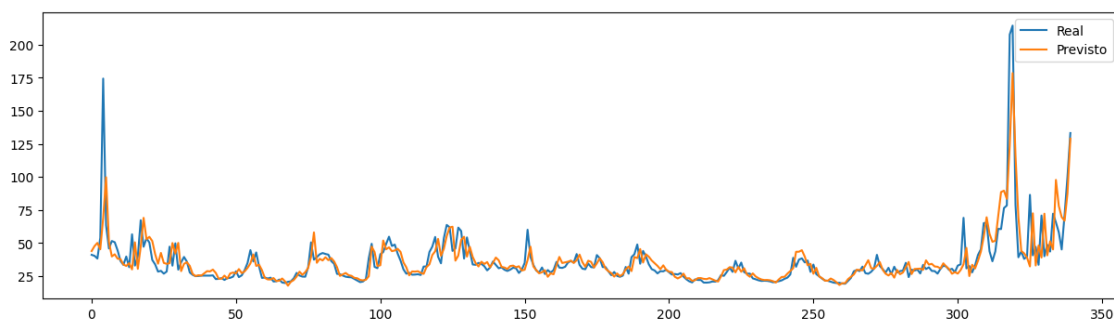


Figura 24 – Previsão com os valores de Teste

De acordo com a Figura 24, o modelo acompanha a estrutura temporal do preço. Este prevê bem as oscilações de pequena e média amplitude. Observa-se, contudo, uma ligeira compressão de amplitude e atraso nos episódios de volatilidade extrema. Conclui-se então que o modelo tende a subestimar os picos e as quebras. No regime central, o alinhamento é estável, tendo boa captação dos ciclos diários, contudo os desvios concentram-se nos extremos.

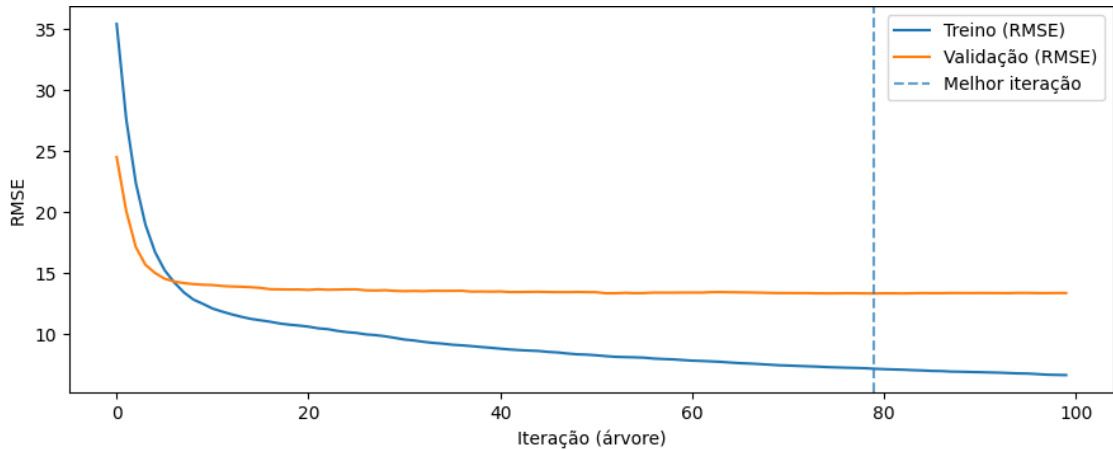


Figura 25 – Curva de Aprendizagem do XGBoost

A partir da Figura 25, conclui-se que o erro de treino do modelo decresce continuamente com o número de árvores, enquanto o erro de validação desce nas primeiras iterações e depois estabiliza. A linha vertical tracejada assinala a melhor iteração, onde ocorre a *early stopping*, a partir da qual ganhos adicionais em treino não se traduzem em melhoria fora da amostra e podem induzir em *overfitting*. Além disto, a Figura 25 demonstra que a maior parte do ganho é captado nas primeiras árvores, ficando o modelo saturado com as *features* atual.

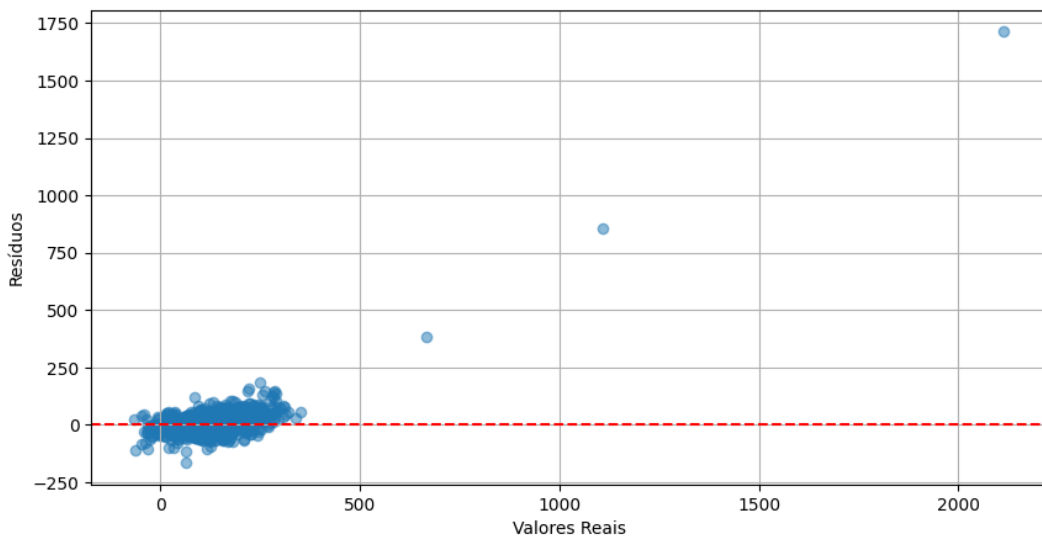


Figura 26 – Resíduos da previsão com os valores de Teste

Na Figura 26 está representada a distribuição dos resíduos, onde a maioria dos resíduos encontram-se próximos de zero para níveis baixos a médios, sem padrão sistemático visível. À medida que o preço aumenta, surge um padrão, com dispersão crescente. Nos níveis de preço mais altos, há maior incidência de resíduos positivos, indicando subestimação nos picos. O gráfico evidencia heterocedasticidade, sendo o erro, pequeno no regime central e maior nos extremos, o que ajuda a explicar o RMSE mais elevado em dias atípicos.

A avaliação foi efetuada com uma partição temporal única e uma janela fixa de 48 horas. Esta configuração capta dois ciclos diários, mas pode ser insuficiente. A opção da parametrização uniforme nas RNNs possibilitou a comparação entre os modelos, mas condiciona o seu desempenho relativo. O espaço de procura de hiperparâmetros aplicado nos modelos de árvore e *boosting* poderia ter sido maior, e a parametrização por *GridSearch* não demonstrou melhorias significativas. No pré-processamento, a decisão de manter *outliers* conserva acontecimentos informativos, como picos e quebras, ao custo de maior variância do erro, o que pode ser mais dispendioso para os modelos captarem estas irregularidades, causa de uma potencial influência desproporcionada em métricas baseadas no erro quadrático. Além disto, o conjunto de variáveis no *dataset* poderia incluir mais variáveis referentes aos picos ou quedas, com o objetivo dos modelos conseguirem identificar melhor estas instabilidades.

Para futuras melhorias, pretende-se aprofundar o uso de RNNs com maior recurso computacional e parametrizações específicas por arquitetura, explorando janelas temporais maiores e tamanhos de *batch* superiores, para assim promover melhor a otimização. Será também estudada a combinação de modelos que conjuguem as vantagens sequenciais das RNNs com a modelação não linear dos métodos de *boosting*, para reduzir o erro em picos, sem deteriorar o comportamento dos resultados em níveis médios, para atingir uma melhor capacidade de generalização.

5 Conclusões

Neste capítulo apresentam-se as conclusões do trabalho, organizadas em três subcapítulos: Sumário e Contribuições, Limitações e Trabalho Futuro. Assim, sintetizam-se os principais resultados e contributos, identificam-se as limitações do trabalho e delineiam-se as linhas de continuidade para uma futura melhoria.

5.1 Sumário e Contribuições

O desenvolvimento deste trabalho teve como objetivo abordar o problema da volatilidade dos preços de eletricidade no mercado da Península Ibérica, através do desenvolvimento de um sistema avançado de previsão de preços. No capítulo da Introdução são descritos o contexto e a definição do problema onde são apresentados os desafios impostos pela crescente inserção de fontes de energia renováveis e sua influência na instabilidade dos preços. Também são definidos os objetivos e o planeamento para uma melhor gestão do trabalho a desenvolver, assim como a metodologia a utilizar, sendo esta uma combinação do PRISMA (Page et al., 2021) e do CRISP-DM (Wirth & Hipp, 2000), e as considerações éticas a ter, respeitando os regulamentos do RGPD (RGPD, 2016), AI ACT (AI ACT, 2025) e do IEEE (IEEE Code of Ethics, 2020).

No capítulo do Estado da Arte, é utilizada a metodologia PRISMA para a realização da revisão sistemática onde são propostas questões de investigação que serão respondidas através dos artigos encontrados no subcapítulo da Discussão. Assim, é possível uma melhor compreensão do problema a resolver e da área da transição energética e dos mercados de energia elétrica. Além disto, foram explorados diversos modelos de *Machine Learning*, modelos lineares como Regressão Linear, *Ridge* e *Lasso*, e modelos de árvore como Árvores de Decisão e *Random Forest*. Também foram discutidos métodos baseados em *Gradient Boosting*, incluindo *XGBoost*, *LightGBM* e *CatBoost*, e RNNs, com ênfase no LSTM e no GRU, devido à sua

relevância na modelagem de dados temporais e na previsão de séries temporais, concluindo assim o **objetivo 1 e 2**.

A metodologia adotada, ajudou a concluir o **objetivo 3**, e, aliada ao estudo dos modelos de *Machine Learning*, permitiu identificar quais as técnicas mais adequadas para a previsão de preços de eletricidade. A escolha do modelo final foi baseada na avaliação das métricas de erro e na análise de consistência nos resultados obtidos, sendo o *XGBoost* o modelo escolhido para a previsão, concluindo o **objetivo 4 e 5**. Os resultados demonstraram que os métodos de *boosting* se destacaram na redução de erros absolutos, sendo o melhor o *XGBoost*, seguido do *LightGBM* e do *CatBoost* a apresentarem um desempenho superior relativamente às RNNs, que se mostraram mais exigentes na parametrização e no treino, resultando em um desempenho inferior. Adicionalmente, o **objetivo 8**, escrita da dissertação, também se encontra concluído, documentando o trabalho desenvolvido, os resultados e as implicações, de forma a transmitir reprodutividade.

Apesar dos resultados alcançados, dois objetivos ficaram por completar: (6) implementação de um protótipo do sistema de análise e previsão de preços e (7) estudo computacional do protótipo. Foi dada prioridade à realização da revisão sistemática e à comparação dos modelos, esgotando o tempo e os recursos disponíveis. A implementação do protótipo (6) implicava um pipeline completo e a integração de dados, que excedia o intervalo de tempo disponível, o que inviabilizou também o estudo computacional (7).

5.2 Limitações

Apesar de terem sido cumpridos a maior parte dos objetivos, foram encontradas limitações. Estas limitações incluem a escolha de uma janela temporal, que, embora adequada para capturar ciclos diários, pode não ser suficiente para refletir padrões sazonais ou eventos extremos. Além disso, a parametrização uniforme nas RNNs condicionou o seu desempenho, sendo sugerido que em futuras abordagens se explorem parametrizações mais específicas para cada arquitetura. Estas limitações exigem maior recurso computacional e parametrizações específicas para cada arquitetura das RNNs. Além disso, manutenção de *outliers* no conjunto de dados, embora tenha preservado eventos informativos como picos de preço, resultou em maior variância nos erros, o que pode afetar a estabilidade dos modelos.

5.3 Trabalho Futuro

Para futuras melhorias, seria vantajoso o uso de RNNs com maior poder computacional e parametrizações específicas, além da exploração de janelas temporais maiores e tamanhos de *batch* superiores. Além disso, a combinação de modelos de *boosting* e redes neurais pode ser uma abordagem para tirar proveito das vantagens sequenciais das RNNs e dos métodos de *boosting*, permitindo capturar melhor os padrões temporais e, ao mesmo tempo, melhorar a generalização e o desempenho durante episódios de elevada volatilidade. Adicionalmente,

propõe-se concluir os objetivos em falta: (6) implementação de um protótipo do sistema de análise e previsão de preços, incluindo o modelo final num pipeline reprodutível, e (7) estudo computacional do protótipo, com medições sistemáticas de desempenho e análise custo-benefício, de forma a sustentar decisões de funcionamento e escalabilidade.

Referências

- (Abumohsen et al., 2023)
(ACM Digital Library, 2025)
- Abumohsen, M. (2023). Electrical Load Forecasting Using LSTM, GRU, and RNN. *Energies*.
ACM Digital Library (2025). *Association for Computing Machinery Digital Library*. Disponível em: <https://dl.acm.org>. Acesso em: 02 de janeiro de 2025.
- (AI ACT, 2025)
- European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act). Official Journal of the European Union
- (AI RMF, 2023)
- National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). (NIST AI 100-1). U.S. Department of Commerce.
- (Al-Hajj et al., 2023)
- Al-Hajj, Rami et al. (2023). "Ultra-Short-Term Forecasting of Wind Speed Using Lightweight Features and Machine Learning Models." *Proceedings of the 12th IEEE International Conference on Renewable Energy Research and Applications*
- (Bianchi et al., 2017)
- Bianchi, F. M., Maiorino, E., Kampffmeyer, M. C., Rizzi, A., & Jenssen, R. (2017). An overview and comparative analysis of Recurrent Neural Networks for Short Term Load Forecasting.
- (Breiman, 2001)
- Breiman L. (2001). *Random Forests*. Robert E. Schapire. Kluwer Academic Publishers.
- (Bunn et al., 2021)
- Bunn, Derek W., John N. Inekwe, e David MacGeehan. (2021). "Analysis of the Fundamental Predictability of Prices in the British Balancing Market." *IEEE Transactions on Power Systems*, Vol. 36, No. 2
- (CatBoost, 2018)
- CatBoost. (2018, 18 de dezembro). CatBoost enables fast gradient boosting on decision trees using GPUs. Disponível em: <https://catboost.ai/news/catboost-enables-fast-gradient-boosting-on-decision-trees-using-gpus>
- (Chapman et al., 1999)
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., & Wirth, R. (1999). *The CRISP-DM Process Model: Discussion Paper*. CRISP-DM Consortium
- (Chatterjee & Hadi, 2012)
- Chatterjee, S., & Hadi, A. S. (2012). *Regression Analysis by Example (5ª ed.)*. Hoboken, NJ: John Wiley & Sons, Inc.
- (Chaya, 2020)
- Chaya. (2020). *Random Forest Regression*. Level Up Coding. Disponível em: <https://levelup.gitconnected.com/random-forest-regression-209c0f354c84>
- (Chen & Guestrin, 2016)
- Chen, T., Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD*.
- (Chicco, Warrens & Jurman, 2021)
- Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE, and RMSE in regression analysis evaluation.
- (Cho et al., 2014)
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation.
- (Data Courses, 2020)
- Data Courses. (2020). *Decision Trees in Scikit-Learn*. Disponível em: <https://www.datacourses.com/decision-trees-in-sci-kit-learn-1261>
- (ERSE, 2023)
- Entidade Reguladora dos Serviços Energéticos (2023). "C24 Portugal-NL." ERSAR, Relatório de Volatilidade nos Preços do Mercado de Energia

- (ERSE, 2024) Entidade Reguladora dos Serviços Energéticos (2024). Desafios Regulatórios da Integração de Energias Renováveis. Disponível em: https://www.erse.pt/media/faieivum/20241113_feup_pv_final.pdf. Acesso em: 30 de dezembro de 2024.
- (ERSE, 2025) Informações sobre o mercado de energia e renováveis. Entidade Reguladora dos Serviços Energéticos. Disponível em: <https://www.erse.pt>. Acesso em: 03 de janeiro de 2025.
- (European Commission, 2024a) European Commission (2024). "Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)." European Commission.
- (European Commission, 2024b) European Commission (2024). "Clean Energy for All Europeans Package." European Union
- (Friedman, J. H., 2001) Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.
- (García, Fernández, Luengo, & Herrera, 2009) García, S., Fernández, A., Luengo, J., & Herrera, F. (2009). A study of statistical techniques and performance measures for genetics-based machine learning: Accuracy and interpretability. *Soft Computing*, 13(10), 959–977.
- (GeeksforGeeks, 2025) GeeksforGeeks. (2025). ML: Gradient Boosting. Disponível em: <https://www.geeksforgeeks.org/machine-learning/ml-gradient-boosting>
- (Graves et al., 2013) Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 6645-6649.
- (Hancock & Khoshgoftaar, 2020) Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for Big Data: an interdisciplinary review. *Journal of Big Data*
- (Hang Yang et al., 2014) Hang Yang, Fuzheng Zhang, Aidong Xu (2014). "A Method to Predict the Intermittent Power by Classification Model." 9th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing
- (Hochreiter & Schmidhuber, 1997) Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*.
- (Hodson, T. O ,2022) Hodson, T. O (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not, *Geosci. Model Dev.*, 15, 5481–5487.
- (Hoerl, A. E., & Kennard, R. W., 1970) Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*.
- (Hollander & Wolfe, 1999) Hollander, M., & Wolfe, D. A. (1999). *Nonparametric statistical methods* (2nd ed.). Wiley.
- (IEA, 2021) IEA (2021). "Renewables 2021: Analysis and Forecast to 2026." International Energy Agency
- (IEEE Code of Ethics, 2020) IEEE Code of Ethics (2020). IEEE Code of Ethics. IEEE Board of Directors. Disponível em: <https://www.ieee.org/about/corporate/governance/p7-8.html>. Acesso em: 04 de janeiro de 2025.
- (IEEE Xplore, 2025) IEEE Xplore (2025). IEEE Xplore Digital Library. Disponível em: <https://ieeexplore.ieee.org>. Acesso em: 02 de janeiro de 2025.
- (ISO New England, 2025) ISO New England Inc. (n.d.). About us. Retrieved September 24, 2025, from <https://www.iso-ne.com>
- (Jadon, Nain & Choudhary, 2022) Jadon, A., Patil, A., & Jadon, S. (2022). A comprehensive survey of regression based loss functions for time series forecasting.
- (Johansson et al., 2023) Johansson, A. et al. (2023). "Pumped Storage Hydroelectricity for a Sustainable Electricity Transition - With a Case Study of Juktan Power Station." *Renewable Energy Journal*
- (Ke et al., 2017) Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*.
- (Kjærran et al., 2021) Kjærran, A., Bugge, E. S., & Vennerød, C. B. (2021, May 17). Time Series: Long short-term memory RNN

- (Kock, Robin et al., 2024) Kock, Robin et al. (2024). "Developing Two RNN-Based Algorithms for Electricity Price Forecasting in Markets with High Penetration of Renewable Energy Resources." Proceedings of the 12th IEEE International Conference on Renewable Energy Research and Applications
- (Lago et al., 2021) Lago, J., Marcjasz, G., & De Schutter, B. (2021). Forecasting spot electricity prices: Deep learning approaches and empirical comparison. Applied Energy.
- (Liberati et al. 2009) Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of Clinical Epidemiology*, 62(10), e1-e34. <https://doi.org/10.1016/j.jclinepi.2009.06.006>
- (LightGBM, s.d.) LightGBM. (s.d.). Leaf-wise (Best-first) Tree Growth (Features). Disponível em: <https://lightgbm.readthedocs.io/en/latest/Features.html>
- (Lipton, 2015) Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning.
- (Liu et al., 2016) Liu, Y., Chen, H., Li, N., Zhang, Y., Zong, J. (2016). The smoothing strategy of wind power combined with storage based on least square fitting. 2016 International Conference on Smart Grid and Clean Energy Technologies, 192-195.
- (Loizidis et al., 2023) Loizidis, Stylianos et al. (2023). "Extreme Supervised Algorithm for Day Ahead Market Price Forecasting." Proceedings of the IEEE
- (Louppe, 2014) Louppe, G. (2014). Understanding Random Forests: From Theory to Practice (Doctoral dissertation, University of Liège). arXiv:1407.7502
- (Mienye & Jere, 2024) Mienye, I. D., & Jere, N. (2024). A survey of decision trees: Concepts, algorithms, and applications. *IEEE Access*, 12, 9536–9549
- (Moher et al., 2010) Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2010). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *International Journal of Surgery*, 8(5), 336-341. <https://doi.org/10.1016/j.ijsu.2010.02.007>.
- (Montgomery, D. C., Peck, E. A., & Vining, G. G., 2012) Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis (5th ed.). Wiley.
- (Nandakumar et al., 2017) Nandakumar, Neha, e Annaswamy, Anuradha M. (2017). "Impact of increased renewables on natural gas markets in eastern United States." *Journal of Modern Power Systems and Clean Energy*
- (OMIE, 2023) *Relatório Anual de Atividades 2023*. Operador do Mercado Ibérico de Eletricidade. Disponível em: <https://www.omie.es>
- (Öztornaci, s.d.) Öztornaci, B. (s.d.). Simplified structure of XGBoost. Figure 5. Analysing Household Food Consumption in Turkey Using Machine Learning Techniques. Disponível em: https://www.researchgate.net/figure/Simplified-structure-of-XGBoost_fig5_381857634
- (Page et al., 2021) Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews.
- (Prokhorenkova et al., 2018) Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*.
- (RGPD, 2016) Regulamento Geral sobre a Proteção de Dados (RGPD). (2016). Regulamento (UE) 2016/679 do Parlamento Europeu e do Conselho de 27 de abril de 2016. Acesso em: 30 de dezembro de 2024

- (Rumelhart et al., 1986) Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors.
- (Samuels, 2024) Samuels, J. I. (2024). *One-Hot Encoding and Two-Hot Encoding: An Introduction*.
- (ScienceDirect, 2025) ScienceDirect (2025). ScienceDirect Digital Library. Disponível em: <https://www.sciencedirect.com>. Acesso em: 02 de janeiro de 2025.
- (Scikit-learn, 2025) scikit-learn developers. (n.d.). scikit-learn (stable documentation). Retrieved September 24, 2025, from <https://scikit-learn.org/stable>
- (Singh et al., 2015) Singh, Rahul et al. (2015). "A Theory for the Economic Operation of a Smart Grid with Stochastic Renewables, Demand Response and Storage." IEEE 54th Annual Conference on Decision and Control (CDC)
- (Singh et al., 2021) Singh, U.; Rizwan, M.; Alaraj, M.; Alsaidan, I. A Machine Learning-Based Gradient Boosting Regression Approach for Wind Power Production Forecasting: A Step towards Smart Grid Environments. *Energies* 2021, 14, 5196
- (Sohrabi et al., 2021) Sohrabi, C., Franchi, T., Mathew, G., Kerwan, A., Nicola, M., Griffin, M., & Agha, R. (2021). "PRISMA 2020 statement: What's new and the importance of reporting guidelines." *International Journal of Surgery*, 88, 105918. DOI: 10.1016/j.ijisu.2021.105918
- (Student, 1908) Student. (1908). The probable error of a mean. *Biometrika*, 6(1), 1-25.
- (Sutskever et al., 2014) Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 27, 3104-3112.
- (Tibshirani, 1996) Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*.
- (Verdelho, 2024) Verdelho, P. (2024). *Desafios Regulatórios da Integração de Energias Renováveis*. Apresentação na Faculdade de Engenharia da Universidade do Porto, 13 de novembro de 2024. Disponível em: <https://www.erse.pt>. Acesso em: 03 de janeiro de 2025.
- (Wang et al., 2023a) Wang, Guoji et al. (2023). "Adjustable Load Resource Pool Operation Analysis and Capacity Verification Software Development Service Project." 2023 International Conference on Networking, Informatics and Computing (ICNETIC)
- (Wang et al., 2023b) Wang, T., Huang, J., Zhou, L., Zhang, Y., Wang, S., & Tan, C. (2023). Research on Optimization Strategy of Energy Storage and Charging Based on Photovoltaic Power Prediction Model. 2023 3rd International Conference on New Energy and Power Engineering (ICNEPE), 144-148.
- (Weron, 2014) Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*.
- (Western Sydney University, 2024) Western Sydney University (2024). Action Research. Disponível em: https://www.westernsydney.edu.au/learning_futures/home/professional_learning/evidencing_your_practice/action_research. Acesso em: 04 de dezembro de 2025.
- (Wilcoxon, 1945) Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80-83.
- (Wirth & Hipp, 2000) Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining.
- (Yang et al., 2014) Yang, Hang et al. (2014). "A Method to Predict the Intermittent Power by Classification Model." 9th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing
- (Zhang et al., 2015) Zhang, Wei et al. (2015). "Day-Ahead Power Output Forecasting for Small-Scale Solar Photovoltaic Electricity Generators." *IEEE Transactions on Sustainable Energy*

- (Zhang et al., 2023) Zhang, Lei et al. (2023). "Day-ahead probabilistic load forecasting for individual electricity consumption - Assessment of point-and interval-based methods." IEEE Transactions on Power Systems
- (Zheng et al., 2014) Zheng, Yu et al. (2014). "Optimal Allocation of Energy Storage System for Risk Mitigation of DISCOs With High Renewable Penetrations." IEEE Transactions on Power Systems
- (Zou, H., & Hastie, T., 2005) Zou, H., & Hastie, T. (2005). Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society: Series B (Statistical Methodology)