Lisbon School of Economics & Management
Universidade de Lisboa

MESTRADO EM

MÉTODOS QUANTITATIVOS PARA A DECISÃO ECONÓMICA E EMPRESARIAL

TRABALHO FINAL DE MESTRADO

RELATÓRIO DE ESTÁGIO

TENDÊNCIAS DOS MERCADOS *DAY-AHEAD* E *INTRADAY* DE ELETRICIDADE NO MIBEL:

ANÁLISE DE DADOS MULTIVARIADA

RODRIGO JOSÉ PRATA DOS SANTOS

ORIENTAÇÃO:

PROF.ª DOUTORA ALEXANDRA BUGALHO DE MOURA

ENG.º JOÃO PARDELHAS

DOCUMENTO ESPECIALMENTE ELABORADO PARA OBTENÇÃO DO GRAU DE MESTRE

JANEIRO - 2025

Agradecimentos

Em primeiro lugar gostaria de agradecer aos meus familiares mais próximos com especial atenção para a minha mãe, o meu pai e irmã. Embora a distância geográfica tenha sido uma realidade, o apoio que sempre me proporcionaram foi fundamental para a conclusão desta etapa. Agradeço também a toda a equipa docente do mestrado em Métodos Quantitativos para a Decisão Económica e Empresarial nomeadamente à professora orientadora, Alexandra Bugalho de Moura. Não menos importante, agradeço ao João Pardelhas e à restante equipa de *Digital Systems & Data* da Galp que foram uma importante fonte de conhecimento e motivação para realização deste projeto. Finalmente, agradeço aos colegas do programa *ready.set.galp*, que tornaram os dias mais leves e repletos de momentos divertidos, contribuindo para uma experiência enriquecedora e memorável.

Resumo

A atual penetração de energias renováveis no portefólio de geração da *commodity* eletricidade tem vindo a crescer de forma galopante ano após ano. Estes contributos acarretam incertezas devido aos fenómenos externos como previsões meteorológicas, pouco previsíveis e não controláveis. Por outro lado, a geração de energia elétrica está também dependente do preço de outras *commodities,* como o gás, petróleo e carvão. Como resultado tem-se assistido a uma maior volatilidade do preço da eletricidade.

O presente trabalho final de mestrado foca-se em dar respostas no que toca à previsão do mercado à vista *(spot)*: *day-ahead* ou *intraday*, mais vantajoso na aquisição/venda da *commodity* eletricidade, a fim de maximizar as vantagens competitivas. O mercado ibérico de eletricidade (MIBEL), mais especificamente o espanhol foi o mercado alvo do estudo. Optou-se por usar bases de dados de acesso público nomeadamente o *esios*, o *mibgas* e o *entso-e*, tendo-se realizado transformações nas variáveis independentes e criado a da variável dependente "*legenda*", que constitui uma variável binária indicadora do mercado, *day-ahead* ou *intraday*, em que a energia elétrica tem o preço mais reduzido. No estudo são utilizadas abordagens multivariadas de dados tais como a Análise Discriminante, a Regressão Logística e Redes Neuronais Artificiais. De acordo com os resultados obtidos, esta última tipologia de modelos foi o que apresentou um valor mais elevado da taxa de acertos, em aproximadamente 70%.

Com a metodologia aqui aplicada é possível fornecer contributos para os *traders* que atuam no mercado *power*, suportando a tomada de decisão sobre o mercado mais vantajoso para aquisição ou venda de eletricidade no contexto atual.

Palavras-chave: Análise Multivariada; Mercado Eletricidade; Análise Discriminante; Regressão Logística; Redes Neuronais Artificiais

Abstract

The current penetration of renewables in the generation portfolio of the commodity electricity has been growing at a galloping rate year after year. These contributions entail uncertainties due to external phenomena such as meteorological forecasts, that are poorly predictable and uncontrollable. On the other hand, electrical energy generation also depends on the price of other commodities, such as gas, oil, and coal. As a result, there has been a higher volatility in the price of electricity.

The present master's final work focuses on providing answers regarding the forecast of the spot market: day-ahead or intraday, more advantageous in the acquisition/sale of the electricity commodity, in order to maximize competitive advantages. The Iberian electricity market (MIBEL), more specifically the Spanish one was the target of the study. It was opted for the use of publicly accessible databases, namely esios, mibgas, and entso-e, where transformations were made in the independent variables and it was created the dependent variable "legenda", which constitutes a market indicator binary variable, day-ahead or intraday, in which electricity has the lowest price. In the study, multivariate data approaches are used, such as Discriminant Analysis, Logistic Regression, and Artificial Neural Networks. According to the results obtained, this last typology of models was the one that presented the highest value of the success rate, at approximately 70%.

According to the methodology here applied it is possible to provide contributions to traders, who operate in the power market, supporting decision-making about the most advantageous market for the purchase or sale of electricity in the current context.

Keywords: Multivariate Analysis; Power Market; Discriminant Analysis; Logistic Regression; Artificial Neural Networks

Índice

Índice de Tabelas

Índice de Figuras

Lista de Abreviaturas

ACP - Análise dos Componentes Principais

AD - Análise Discriminante

B2B - Business-to-Business

B2C - Business-to-Consumer

CET - Central European Time

CMVM - Comissão do Mercado de Valores Mobiliários

CNE - Comisión Nacional de Energía

CP - Componente Principal

DA – Mercado do Dia Seguinte de Eletricidade *(day-ahead)*

DS&D - Digital Systems & Data

ERSE - Entidade Reguladora dos Serviços Energéticos

ID01 - Primeiro Mercado Intradiário de eletricidade (1º mercado do *intraday*)

MAPE - Mean Absolute Percentage Error

MIBEL - Mercado Ibérico de Eletricidade

ODS - Objetivos de Desenvolvimento Sustentável

OMIE - Operador de Mercado Ibérico, polo espanhol

OMIP - Operador de Mercado Ibérico, polo português

ONU - Organização das Nações Unidas

PBF - Programa Diário Base de Funcionamento

PVP - Programa Provisório Resultante da Aplicação de Restrições Técnicas

RL - Regressão Logística

RMSE- Root Mean Square Error

RNA - Redes Neuronais Artificiais

ROC - Receiver Operating Characteristic Curve

VIF - Variance Inflation Factor

Capítulo 1. Introdução

O estudo que se segue foi desenvolvido em colaboração com a Galp, mais concretamente no departamento de *Digital Systems & Data (DS&D)*. Este departamento está inserido na unidade de negócios, a nível operacional, do *Energy Management*, cujas principais atividades incluem a compra/venda *(trading)* e expedição de várias *commodities* como petróleo, biocombustíveis, gás natural e eletricidade. Especificamente o departamento *DS&D* integra dois objetivos principais: (i) automação de processos e (ii) tarefas de *data science*, como previsões de produção e consumo de energia. Desta forma, o departamento não só simplifica processos através da implementação de automações, como também faz a ligação entre o negócio/mercado e os modelos desenvolvidos, desde provas de conceito até à entrada em produção dos vários modelos.

Através do programa *ready.set.galp* foi possível integrar o departamento *DS&D*, com o intuito de desenvolver o trabalho final de mestrado. As tarefas, que serão detalhadas nos capítulos subsequentes, consistem num estudo no âmbito de *data science* aplicada ao mercado da eletricidade, mais especificamente o espanhol, que está inserido no mercado ibérico de eletricidade (MIBEL). Dessa forma, é necessário compreender o mercado, recolher, analisar e tratar os dados, por fim, aplicar os métodos mais apropriados de análise multivariada, com o intuito de responder ao objetivo do estudo, o qual também será explicado em seguida.

Atualmente, a Galp atua no mercado à vista *(spot)*, ou seja, contratações de compra/venda de energia elétrica para entrega no dia seguinte ao da negociação. Estas contratações podem ocorrer no mercado diário (DA) ou em contrapartida nos vários mercados intradiários. A questão que se pretende responder, é em que mercado, no DA ou no primeiro mercado intradiário (ID01), se deve realizar a aquisição/venda de energia elétrica, de forma a gerar vantagem competitiva. Esta predição é importante para os agentes que atuam como gestores de portefólio de energia, especialmente no mercado desta *commodity*, onde o preço assume dinâmicas muito voláteis podendo chegar a diferenças de duas ordens de grandeza [1]. Este fator pode ser explicado pelo preço das matérias-primas que contribuem para a geração da *commodity*, bem como pelo atual cenário da penetração das energias renováveis que trazem consigo incertezas externas que

são difíceis de prever e controlar. Além disso, como é insustentável armazenar a energia elétrica em grande escala, poderão ocorrer situações em que apesar da elevada contribuição de fontes renováveis, não exista capacidade suficiente para escoar essa energia para a rede, fenómeno conhecido como *curtailment*. Nos últimos anos não foi apenas a produção de eletricidade que sofreu transformações, o consumo de energia elétrica também, influenciado por eventos socioeconómicos, como crises, alteração dos padrões de trabalho ou de mobilidade, consciencialização ambiental, entre outras causas.

Para o efeito é conduzido um estudo através do qual são utilizadas bases de dados públicas com variáveis referentes à geração de energia elétrica sob as suas diversas formas de produção, a procura, entre outas variáveis que foram alvo de transformação. No que respeita à variável dependente, esta constituirá uma variável binária indicando qual mercado, DA ou ID01, apresenta o preço da *commodity* mais baixo.

O estudo é dividido em cinco partes principais: no Capítulo 2, é alvo de análise o contexto atual do mercado de eletricidade, mais especificamente o MIBEL; no Capítulo 3 são apresentados os contributos de vários autores no que toca às previsões no mercado da *commodity*; o Capítulo 4 é dedicado aos dados que foram empregues no estudo, incluindo técnicas de pré-processamento, criação de novas variáveis e da variável dependente, além de pressupostos que foram adotados; no Capítulo 5 são especificadas as metodologias de análise de dados utilizadas no estudo; por fim, o Capítulo 6 destina-se à exposição e discussão dos resultados.

Capítulo 2. O Mercado da Eletricidade

A partir da década de 80, muito devido à crise petrolífera de 1973, foram alterados os hábitos de consumo de eletricidade na Europa. A maior liberalização e reestruturação do mercado de eletricidade passou a ser uma realidade tendo-se assistido à entrada de novas empresas no mercado. Previamente a esse período as atividades de produção, transporte, distribuição e comercialização de eletricidade estavam fortemente integradas verticalmente, ou seja, cabendo a um conjunto limitado de empresas ou até aos governos, desenvolverem todas estas atividades. Na verdade, atualmente constata-se que atividades como o transporte e a distribuição continuam a ser consideradas como "monopólios naturais" regulados, uma vez que se torna mais eficiente economicamente que assim seja por acarretarem um elevado investimento e níveis de operação. Já atividades como a produção e a comercialização de energia elétrica atualmente estão abertas à concorrência, visando uma maior eficiência na gestão de recursos[1]. A produção de energia elétrica é caraterizada como um mercado grossista, enquanto a sua comercialização um mercado retalhista, onde agentes concorrem para fornecer clientes finais [2].

O mercado grossista, especialmente nos contextos português e espanhol, é composto por empresas que atuam como produtores de eletricidade. Este mercado é caracterizado por uma alta liquidez e concorrência, o que acarreta impactos ao longo da cadeia de eletricidade. Posteriormente, as empresas de comercialização no mercado retalhista têm a liberdade de escolher as suas fontes de abastecimento.

O Decreto-Lei n.º 29/2006 de 15 de fevereiro, artigo 16, estabelece a existência de dois tipos de produção de eletricidade: (i) Produção em regime ordinário[2], em que a eletricidade é gerada através de fontes de energia tradicionais não renováveis, como o carvão, o gás natural, o petróleo, assim como as grandes centrais hídricas; (ii) Produção em regime especial[3], em que se produz eletricidade através de recursos "endógenos renováveis ou de tecnologias de produção combinada de calor e eletricidade" [3][4]. No

[1] Art.º 4 do Decreto-Lei n.º 29/2006 de 15 de fevereiro
[2] Art.º 17 do Decreto-Lei n.º 29/2006 de 15 de fevereiro
[3] Art.º 18 do Decreto-Lei n.º 29/2006 de 15 de fevereiro

que se refere ao tipo de contratação de eletricidade, este mercado em Portugal e Espanha, divide-se em [2]:

- Mercado de contratação a prazo, onde se estabelecem contratos de compra ou venda de eletricidade em datas futuras a preços previamente acordados;
- Mercado de contratação à vista *(spot)*, caracterizado por transações imediatas com ajustes diários ou intradiários;
- Mercado contínuo, onde se realiza a equalização instantânea entre a produção e o consumo de energia elétrica;
- Mercado de contratação bilateral, caracterizado por acordos de compra e venda diretos entre as partes, para vários horizontes temporais.

O mercado retalhista é constituído por um conjunto de entidades que compra e vende eletricidade para fins comerciais a clientes finais ou a outros agentes, sob a celebração de acordos bilaterais ou participação noutros mercados como o MIBEL, no caso de Portugal e Espanha[4]. Existem duas principais formas de contratação para o fornecimento de energia elétrica: (i) o mercado regulado, onde os preços e tarifas são definidos anualmente pela ERSE, no caso de Portugal. Os comercializadores que atuam neste tipo de mercado designam-se por "comercializador de último recurso" e devem adquirir toda a energia proveniente da produção em regime especial. Já no (ii) mercado liberalizado há a possibilidade de acordar condições de negociação de energia em contratações bilaterais ou através dos mercados organizados onde os preços da energia elétrica são definidos em função da oferta-procura [2][5][6].

2.1 O MIBEL - Mercado Ibérico de Eletricidade

Na Europa não existe um mercado único de comercialização de eletricidade, devido essencialmente a "questões de acesso à rede, a questões de tarifação e à diversidade de graus de abertura dos mercados existentes nos Estados-Membros" ([7], pág.1). Desta forma, foram criados, a partir da década de 90, mercados regionais de eletricidade. O seu propósito é migrar de uma abordagem monopolista da comercialização de eletricidade, para uma dinâmica de concorrência, visando atrair novas empresas que

[4] Art.º 42 do Decreto-Lei n.º 29/2006 de 15 de fevereiro

possam competir nestes mercados. Desta forma, os consumidores finais ficam mais protegidos do poder dominante de grandes empresas, que até então detinham o controlo no mercado da eletricidade.

A primeira bolsa de energia multinacional foi a *NordPoll*, que produz e distribui energia nos países nórdicos e bálticos. O segundo mercado regional criado foi precisamente o que atua em Portugal e Espanha, o MIBEL. Posteriormente assistiu-se ao surgimento de vários outros mercados regionais na Europa, como o *N2EX*, que é uma extensão da *NordPoll* e cujo mercado alvo é o Reino Unido, o *GME* em Itália, a *APX-ENDEX* que regula os mercados de eletricidade e gás natural na Holanda e na Bélgica, a *Powernext* em França, entre outros [8]. Todos estes mercados regionais almejam a construção de um futuro mercado interno europeu de energia [2], que trará benefícios "em termos de aumento de eficiência, reduções de preços, padrões de serviço mais elevados e maior competitividade", ([7], pág.1).

O acordo entre Portugal e Espanha, que mais tarde constituirá o MIBEL, foi assinado em Santiago de Compostela a 1 de outubro de 2004. No entanto, só a 1 de julho de 2007 o MIBEL teve todas as suas atividades e funções em curso. Esta integração dos sistemas elétricos ibéricos apresenta benefícios para consumidores e produtores de eletricidade na região, permitindo a livre entrada no mercado a novos agentes, produtores ou comerciantes, portugueses e espanhóis. Desta forma, o MIBEL "deverá basear-se nos princípios de transparência, livre concorrência, objetividade e liquidez, autofinanciamento e auto-organização dos mercados" ([9], pág.2), onde todos os intervenientes estarão em igualdade de direitos, informações e obrigações [9].

O MIBEL, atualmente, é subdividido em dois polos, o polo português (OMIP) e o polo espanhol (OMIE). O OMIP é responsável pelos mercados a prazo, nos quais são contratadas transações de eletricidade para dias seguintes aos da negociação. Existem quatro tipos de contratos estabelecidos: futuros, *forwards, swaps* e de opção [10]. De referir que a Comissão do Mercado de Valores Mobiliários (CMVM) é a entidade supervisora do OMIP [2].

Por sua vez, o OMIE está encarregue das transações realizadas no mercado à vista *(spot)* onde se estabelecem negociações diárias, com o fornecimento de eletricidade no dia seguinte ao da contratação e intradiárias. Este mercado é também conhecido como de

"ajustes", uma vez que tenta oferecer uma adequação mais precisa entre a oferta e a procura de eletricidade. O OMIE é supervisionado pela *Comisión Nacional de Energía (CNE)* [2].

Em suma, o OMIP gere o mercado a prazo, "em que se estabelecem compromissos a futuro de produção e de compra de energia elétrica", ([2], pág.18). Já o OMIE gere o mercado à vista, "com uma componente de contratação diária e uma componente de ajustes intradiários em que se estabelecem programas de venda (produção) e de compra de eletricidade para o dia seguinte ao da negociação", ([2], pág. 18). As características destes dois tipos de mercados, diário e intradiário, estão detalhadas nos próximos pontos, bem como a caracterização da curva da oferta de eletricidade. Referir também que apesar do MIBEL atuar em Portugal e Espanha, o mesmo não está isolado, permitindo, por exemplo, fluxos de energia com França.

- **Mercado Diário de Eletricidade**

O mercado diário ou *day-ahead* é um dos mais utilizados para fazer transações de compra e venda de energia elétrica, permitindo aos agentes realizarem as suas ofertas para as 24 horas do dia seguinte. Além de Portugal e Espanha, este mercado funciona desde 2014 noutros mercados na Europa, e tem como finalidade alcançar um mercado interno europeu. Desta forma, no MIBEL, são admitidas propostas de compra e venda de energia elétrica do dia seguinte (D+1) até às 12h00 *CET* do dia anterior (D), com um total de 24 ofertas de compra ou venda, correspondendo às 24 horas do dia. Após esse período, é feito um leilão para cada uma das horas, de forma a cruzar a curva da oferta dos produtores de energia com a curva da procura. Este cruzamento é designado de preço de encontro e indica qual o preço e o volume de eletricidade a negociar para cada hora do dia seguinte [2][5][11]. Na *Figura 2.1* encontra-se um exemplo das curvas da oferta e da procura da eletricidade para um dia e hora específicos, onde é possível observar o preço e a quantidade de energia a negociar no ponto de equilíbrio.

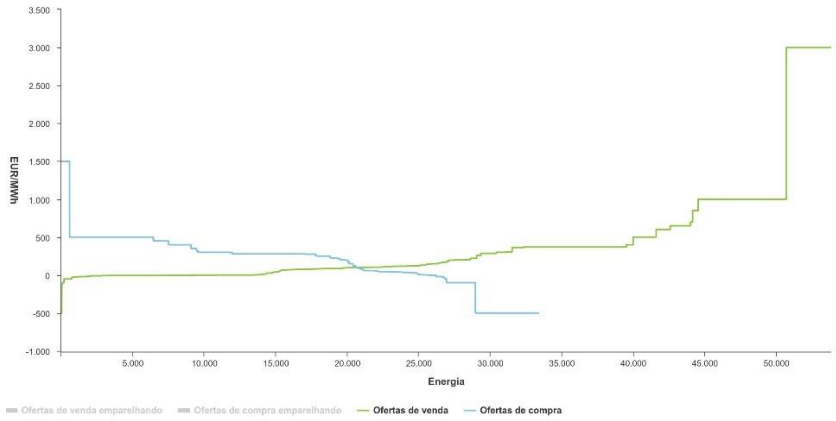

Figura 2.1 Curvas agregadas da oferta e da procura de energia para Portugal e Espanha, registadas a 08/07/2024 na 1ª hora do dia[5]

No mercado diário, que inclui Portugal e Espanha, no caso em que o preço de encontro não exceda a capacidade comercial disponível, o mercado permanece integrado, resultando num único preço da *commodity* para ambos os países. No entanto, se a capacidade de integração entre os dois países não suportar o fluxo de energia determinado pelo mercado, procede-se à separação dos mercados espanhol e português. Como resultado do encontro das curvas da oferta e procura de cada um dos países, são obtidos preços distintos para cada uma das regiões. Esta separação, conhecida como *market splitting*, visa otimizar o uso da capacidade disponível sem comprometer a segurança do mercado [2].

- **Mercado Intradiário de Eletricidade**

O mercado intradiário de eletricidade ou *intraday* é uma plataforma complementar ao mercado diário, uma vez que permite que sejam feitas várias negociações de eletricidade com o objetivo de aferir uma melhor adequação ao consumo e produção da *commodity*. Assim, existe uma melhor otimização do portefólio dos agentes durante horizontes temporais sucessivos de curto prazo. Atualmente, este mercado está organizado em seis sessões, permitindo aos intervenientes ajustarem as suas opções de compra/venda de eletricidade, conforme demonstrado na *Tabela A*. Este mercado de

[5] Gráfico retirado do site do operador OMIE. Acesso a 08 julho 2024, em: https://www.omie.es/pt/market-results/daily/daily-market/aggragate-suply-curves

ajustes é importante caso ocorram problemas de congestionamento na rede de distribuição, ou avarias em equipamentos na rede elétrica [5][12].

Tabela A - Horário (CET) das sessões do Mercado Intradiário, retirado de [12]

	SESSÃO 1ª	SESSÃO 2ª	SESSÃO 3ª	SESSÃO 4ª	SESSÃO 5ª	SESSÃO 6ª
Abertura de sessão	14:00	17:00	21:00	1:00	4:00	9:00
Encerramento de sessão	15:00	17:50	21:50	1:50	4:50	9:50
Emparelhamento	15:00	17:50	21:50	1:50	4:50	9:50
Publicação do programa acumulado (PIBCA)	15:07	17:57	21:57	1:57	4:57	9:57
Publicação PHF dos OSs	16:20	18:20	22:20	2:20	5:20	10:20
Horizonte da Programação (Períodos horários)	24 horas (1-24 D+1)	28 horas (21-24 y 1-24 D+1)	24 horas (1-24 D+1)	20 horas (5-24)	17 horas (8-24)	12 horas (13-24)

- **Caracterização da Curva da Oferta**

Todas as ofertas de energia propostas são submetidas pelos agentes do mercado até às 12h00 *CET* (para o mercado diário), onde são ordenadas de forma crescente no que respeita ao seu custo marginal, formando assim a curva de oferta para cada hora do dia seguinte. Tipicamente, a eletricidade proveniente de fontes renováveis apresenta custos marginais mais baixos, uma vez que não depende do preço das matérias-primas, em oposição à energia gerada a partir de fontes como o carvão ou o petróleo, onde o custo marginal é mais elevado, posicionando-se, portanto, em zonas mais elevadas da curva. Estes últimos tipos de produção de eletricidade são responsáveis pela cobertura da procura de energia durante períodos extremos do ano, ou quando as reservas hídricas são escassas. Como mencionado anteriormente, o cruzamento da curva da oferta com o da procura determina a quantidade e o preço da eletricidade a ser transacionada e, por conseguinte, quanto maior o contributo das energias renováveis no portefólio das ofertas, menor será o preço da eletricidade, tornando mais vulneráveis os produtores de energia cujos custos marginais de produção são mais elevados [2][13].

2.2 Galp Energia

A Galp Energia é uma empresa portuguesa, fundada em 1999, através da junção da Petrogal[6] e da GDP (Gás de Portugal). Porém, a sua origem remonta ao ano de 1846, quando foi atribuída a concessão da iluminação pública da capital portuguesa à

[6] A Petrogal foi constituída em 1976 com a fusão da SACOR, CIDLA, SONAP e PETROSUL.

Campanhia Lisbonense de Iluminação a Gás. Desde então, a empresa tem vindo a crescer e a adaptar-se aos desafios energéticos: atualmente está presente em 10 países distribuídos por 3 continentes, conta com um portefólio alargado de serviços e produtos energéticos e é listada na *Euronext* Lisboa desde 23 de outubro de 2006 [14].

Figura 2.2 Logotipo da Galp

A organização atua no setor energético sendo pioneira em comercializar as três principais formas de energia na península ibérica: produtos petrolíferos, gás natural e eletricidade [15]. A Galp está envolvida na exploração e produção de petróleo e gás natural, em países da América do Sul, como o Brasil, e da África, como a Namíbia, Moçambique, Angola e São Tomé e Príncipe. Relativamente à compra e venda de eletricidade, a Galp atua no MIBEL, tanto em contratos a prazo, onde estabelece compras de caráter futuro de energia, bem como contratos à vista, com uma componente de ajustes diários e intradiários e também acordos bilaterais [16].

No âmbito da sua atividade comercial, a Galp opera em três principais segmentos de mercado [14]:

- No *B2C* a empresa mantém uma extensa oferta de serviços, detendo uma forte presença no setor de produtos petrolíferos, com 1 273 estações de serviço na Península Ibérica, representando uma quota de mercado de cerca de 26,3%, em Portugal. Na mobilidade elétrica é líder em Portugal, detendo aproximadamente 26% do mercado, contando com mais de 2 400 postos de carregamento e com a meta de alcançar 10 000 até 2025. No fornecimento de gás natural e eletricidade, atende aproximadamente 365 mil clientes na Península Ibérica, com uma quota de mercado de cerca de 25,3% em gás natural e 4,9% em eletricidade em Portugal.
- No segmento *B2B*, possui quase 10 mil clientes de gás natural e eletricidade e cerca de 21 mil clientes em produtos petrolíferos, distribuídos em diversos setores de atividade.

- No mercado internacional, a Galp tem presença em cinco países africanos, sendo líder de mercado em Cabo Verde, Guiné-Bissau e Eswatini, com um total de 202 estações de serviço nesses três países.

De referir que a Galp está comprometida não só em promover um valor acrescentado para todos os seus *stakeholders*, como também em viabilizar um importante papel na transição energética, social e de *governance*. Assim, a empresa rege-se pelos ODS da ONU, os quais são categorizados com base no seu potencial impacto na organização bem como na relevância para os *stakeholders*. Os principais estão representados na *Figura 2.3* [14].

Figura 2.3 "ODS materiais" considerados pela Galp

Capítulo 3. Revisão da Literatura

As previsões do preço de eletricidade são relevantes para os participantes de mercado, cujo foco reside na comercialização dessa *commodity* de forma a otimizar a carteira de negócios e realizar escolhas economicamente rentáveis. A complexidade e o risco inerente às previsões são amplificadas no contexto em estudo, particularmente pela penetração das energias renováveis na geração de eletricidade. A produção de eletricidade sob formas renováveis está muito correlacionada com fenómenos externos, que são muito difíceis de prever ou controlar, como a inclinação do sol, a velocidade do vento e até a precipitação. Além de que, a energia renovável tem ocupado uma percentagem substancial no portefólio de eletricidade produzida na península ibérica. Segundo a REN[7] e a *Red Eléctrica*[8], em 2023, foram ultrapassados todos os *records* no que diz respeito à percentagem de geração das energias renováveis, 50,3% em Espanha e 61% em Portugal. Além das referidas, outras causas que tornam a previsão enviesada são os eventos quotidianos e socioeconómicos, que têm impacto no preço da eletricidade, como horários de ponta da utilização e dias do ano específicos (feriados e fins de semana) [17][18].

Apesar dos desafios mencionados, a comunidade científica está empenhada em dar respostas. Nesta secção, são discutidos os esforços e limitações conhecidos até à data sobre as principais abordagens adotadas nas previsões no mercado da eletricidade. *Rafał Weron*[9], professor e investigador, é uma das principais referencias no que respeita à previsão no mercado da energia. O objetivo deste capítulo não passa pelo escrutínio de cada uma das técnicas utilizadas na previsão do preço da eletricidade, mas sim em ressaltar os principais métodos e os que são amplamente conhecidos na literatura. Desta forma, utilizou-se o trabalho de revisão feito por de *Weron* [1] como fio condutor da

[7] REN - "Produção de energia renovável bate recorde em 2023". Acesso a 13 de maio 2024, em https://www.ren.pt/pt-pt/media/noticias/producao-de-energia-renovavel-bate-recorde-em-2023

[8] *Red Eléctrica* – "*España pone en servicio en 2023 la mayor cifra de potencia instalada solar fotovoltaica de su historia*". Acesso a 13 de maio 2024, em https://www.ree.es/es/sala-de-prensa/actualidad/nota-de-prensa/2024/03/espana-pone-en-servicio-en-2023-la-mayor-cifra-de-potencia-instalada-solar-fotovoltaica-de-su-historia

[9] As publicações de Rafał Weron podem ser consultadas em *(*Acesso a 08 julho 2024*):* https://p.wz.pwr.edu.pl/~weron.rafal/Publ

literatura aqui apresentada. O autor destaca, entre outras, duas tipologias de técnicas distintas: as Estatísticas/econométricas e as de Inteligência computacional.

As técnicas estatísticas preveem o preço da eletricidade com o uso de combinações matemáticas do histórico de preços registados e/ou de outras variáveis exógenas como consumo, produção e dados meteorológicos. Dentro desta categoria existe uma variedade de modelos específicos que são referidos em seguida.

O modelo dos dias semelhantes, também conhecido como *naïve model*, baseia-se na suposição que o valor futuro do preço da *commodity* é determinado pelo preço observado em iterações passadas, que exibem características muito semelhantes às que se procuram prever. Este modelo também pode considerar uma combinação linear de múltiplas observações que compartilham características semelhantes [19][20]. Assim, de acordo com [21], um dia semelhante pode ser caraterizado de acordo com o dia da semana: por exemplo o preço horário do sábado, domingo e segunda-feira é semelhante aos dos respetivos dias da semana anterior. A terça-feira é semelhante à segunda feira imediatamente anterior (na mesma semana), e da mesma forma para os restantes dias uteis da semana.

O modelo de suavização exponencial é constituído a partir da média exponencialmente ponderada de observações passadas. Ou seja, cada valor previsto é calculado como a média ponderada de observações anteriores, onde a importância diminui em relação às observações mais antigas. Em [22] é comparado o método de suavização exponencial com o dos dias semelhantes, para o mercado *Nord Pool*: por meio da suavização exponencial, os autores obtiveram uma performance ligeiramente superior, com um *RMSE* de 8,64, comparado com os 9,36 obtidos com o modelo *naïve*.

Os modelos de regressão com séries temporais são provavelmente a técnica mais comum no que se refere à previsão dos preços da eletricidade. Estes modelos têm a vantagem de poder expressar o valor atual do preço da eletricidade através dos seus valores passados e um determinado termo de erro, ou adicionalmente combinar variáveis exógenas que ajudem a determinar o valor do preço atual. De referir que estes modelos servem igualmente de base para a previsão do preço de outras *commodities*, como por exemplo do petróleo e gás [23][24]. Incluem-se nestes modelos os autorregressivos (AR), que apenas têm em conta os valores passados, e os ARMA, junção do modelo AR com o

de médias móveis (MA) onde o valor da *commodity* é relacionado com os termos de erro observados em momentos anteriores. As aplicações destes modelos pressupõem que a série temporal do preço da eletricidade seja fracamente estacionária. Caso contrário, esta pode ser alvo de diferenciações até o ser, constituindo assim o modelo ARIMA. Por fim, os modelos SARIMA são utilizados quando são requeridos desfasamentos mais longos, com um comportamento sazonal, como é o caso do preço da eletricidade.

Em [25], foi realizada uma análise dos padrões do preço da eletricidade no mercado alemão, especificamente na bolsa de energia de *Leipzig*, onde foram usados modelos univariados como AR's e ARMA's. Paralelamente no mesmo artigo os autores desenvolvem um modelo hora-a-hora de forma separada, para preços *spot* de eletricidade, demonstrando uma melhoria significativa na capacidade de previsão. Noutro estudo [26], modelos ARIMA foram aplicados para investigar os mercados da Califórnia e do MIBEL. O objetivo passou pela previsão a curto prazo das 24 horas do dia D+1.

Em [27] são comparadas previsões realizadas por métodos *naïve* e ARIMA no mercado espanhol, para o dia seguinte. A novidade que é introduzida neste artigo é o uso do *wavelet*, um método de transformação da série. Por meio desta técnica, a série temporal é dividida em múltiplas séries, tipicamente quatro, menos complexas que a original e com variâncias mais estáveis. Aqui, os autores aplicaram a inversa da função *wavelet* das séries constituintes de forma a prever o preço da eletricidade. Os resultados revelam que esta abordagem supera os métodos convencionais do modelo ARIMA e do método *naïve*.

O modelo de séries temporais tem a vantagem de agregar variáveis exógenas que podem conter informações relevantes para a previsão futura. Assim, os modelos não estão apenas relacionados com o seu passado, mas também com variáveis que têm impacto no preço da eletricidade. Em [26], foram também acrescentadas variáveis "exploratórias", a procura e disponibilidades hídricas para o modelo de previsão no MIBEL, e a procura para o mercado da Califórnia. Como resultado, os autores concluíram que as métricas de desempenho são genericamente melhores para o modelo com variáveis exógenas.

Estes modelos com séries temporais captam a atenção da comunidade científica pelo facto de conseguirem dar uma interpretação intuitiva dos resultados. A previsão depende não apenas dos algoritmos usados, mas também da qualidade dos dados e da capacidade de incorporarem fatores relevantes, no caso de dados externos. No entanto,

estes modelos apresentam um fraco desempenho quando existem comportamentos não lineares e valores "extremos". Deste modo, existem artigos que abordam a possibilidade de substituir os valores "extremos" por outros mais "razoáveis", embora haja falta de consenso sobre se se devem retirar ou manter estas observações [22][28][29].

Os modelos de inteligência computacional constituem uma outra família de modelos que conseguem dar uma resposta mais eficiente às limitações dos modelos estatísticos, como a habilidade de captar não linearidades e valores extremos, sendo por isso designados como "inteligentes". As redes neuronais artificiais (RNA), os sistemas *fuzzy*, os *support vector machines* (SVM) e *evolutionary computation* são as técnicas mais conhecidas dentro dos modelos de *machine learning*, possuindo características que os tornam modelos adequados para previsões a curto prazo do preço da eletricidade. As redes neuronais têm recebido especial atenção na comunidade científica, enquanto as demais são frequentemente empregues em contextos híbridos, por exemplo em [30] e [31], são combinadas técnicas de modelos estatísticos ARIMA com SVM.

As RNA, como já referido, têm sido amplamente utilizadas na previsão de preços da eletricidade, muito pelo aumento da capacidade computacional, que possibilitou a reestruturação das redes para uma estrutura mais profunda e complexa, ou com características *feedforward*. As RNA podem ser classificadas de acordo com os seus nós de saída. Quando apresentam apenas um, estão destinadas a prever o preço num determinado momento da *commodity*: hora imediatamente seguinte, *n* horas à frente, preço de pico no dia seguinte ou também preço médio do dia seguinte. Por outro lado, podem conter múltiplos nós de saída, tipicamente 24 ou 48, de forma a prever o perfil completo dos preços horários de um ou dois dias [1][28][29].

Em [32], são utilizadas técnicas de redes neuronais, que posteriormente são comparadas com modelos estatísticos, mais concretamente com os modelos *naïve* e ARIMA. Os mercados alvo desta pesquisa foram os de Espanha e da Califórnia. Em ambos o modelo RNA superou os modelos *naïve* e ARIMA no que diz respeito à métrica de desempenho *MAPE* (*Mean Absolute Percentage Error*). Em [19], e no mesmo sentido do estudo anterior, no mercado da *Nord Pool*, foram comparados modelos de regressão estimados pelos mínimos quadrados (*OLS*), modelos *Ridge* e *Lasso*, com RNA. Os resultados indicam *MAPE* de 7.09%, 7.11%, 7.07% e 6.53%, respetivamente.

Por estas razões, as técnicas de inteligência computacional têm-se tornado populares. No entanto, além da severa carga computacional que acarretam, existe um leque muito diverso de modelos e respetivos hiperpârametros que têm de ser fixados, a acrescer a perda de interpretabilidade que estes modelos introduzem. Desta forma, torna-se difícil comparar e encontrar a solução ideal. Por outro lado, a flexibilidade, bem como a resposta à não linearidade, pode não resultar necessariamente em melhores previsões pontuais [1][29].

Todas as técnicas referidas acima têm como objetivo a previsão direta do preço da eletricidade a curto e médio prazo. No entanto, o objetivo da presente tese passa pela comparação entre o preço da eletricidade no mercado diário e no intradiário. A literatura, até onde foi possível apurar, é bastante omissa quanto ao tipo de técnicas e tratamento de dados a empregar na comparação entre estes dois mercados. O trabalho conduzido por *Maciejowska, Nitka* e *Weron* em 2019 [33] é o que mais se aproxima do âmbito pretendido, no entanto usa uma metodologia distinta da que aqui se propõe. Os autores identificam a disparidade do preço da eletricidade entre os mercados diário e intradiário recorrendo a modelos AR e ao modelo *probit*, de forma a avaliar os mercados da Alemanha e da Polónia. Já no nosso trabalho são aplicadas outras técnicas de análise de dados multivariados, que serão detalhados mais adiante no capítulo da metodologia, e o mercado alvo, é o espanhol. Assim o objetivo em ambos os estudos é prever em qual mercado o preço da eletricidade é mais reduzido, a fim de tomar decisões economicamente vantajosas.

Os mercados da Alemanha e da Polónia, analisados em [33], exibem comportamentos distintos no que se refere ao preço da eletricidade. Uma das causas apontadas pelos autores refere-se à constituição do portefólio de geração da energia. Na Alemanha assistiu-se a um sucesso na penetração das energias renováveis, o que não é observado na Polónia, onde a geração de eletricidade é baseada essencialmente em carvão, resultando em preços mais elevados e voláteis. No que se refere aos modelos, foram considerados dois tipos, AR e *probit*, onde ambas as abordagens a ligar a variável dependente a preços passados e a um conjunto de variáveis exógenas como a procura, a geração de energia eólica e as reservas previstas de energia disponível. Desta forma, foram considerados dados horários e uma média para construir dados agregados diários. A variável dependente em ambos os modelos é binária e assume o valor um, caso o

produtor decida vender a eletricidade no mercado intradiário, e zero caso contrário. Na técnica linear AR com variáveis exógenas (ARX), foram considerados dois cenários: (i) Os preços dos mercados foram modelados separadamente e em seguida, a diferença foi calculada; (ii) O *spread* do preço é calculado diretamente. Em ambos os casos os parâmetros são estimados com recurso ao método dos mínimos quadrados. Relativamente ao modelo *probit*, este é utilizado de forma a descrever diretamente a distribuição de probabilidade da variável binária e os parâmetros são estimados com recurso ao método da máxima verossimilhança.

No que diz respeito aos resultados, a métrica base usada pelos autores indica a frequência com que as previsões coincidem com os valores verdadeiros. De uma forma prática essa métrica é idêntica à *accuracy*, que também será considerada nesta tese. No caso polaco os modelos que obtiveram uma maior precisão foram os ARX, especialmente os que descrevem de forma separada os preços dos mercados diário e intradiário, com uma taxa de sucesso de 57,3%. Contrariamente a estes resultados, no caso alemão o modelo *probit* apresenta um melhor poder de classificação, 54,2%.

Assim, de forma a responder ao desafio proposto pela equipa de *DS&D* da Galp, neste trabalho é conduzido um estudo de análise multivariada de dados para fornecer *insights* importantes sobre a seleção do mercado mais vantajoso, diário ou o primeiro mercado intradiário, com vista à aquisição da *commodity* eletricidade. Note-se que, apesar do objetivo do estudo ser igual ao de [33], a nossa tese utiliza outras técnicas de análise de dados, tais como a análise discriminante, regressão logística e redes neuronais artificiais, em vez dos modelos AR e *probit*. Por outro lado, o mercado alvo do nosso estudo é o MIBEL mais especificamente o espanhol. Lembre-se que se o preço de encontro não exceder a capacidade comercial disponível, o preço permanece o mesmo para ambas as regiões, Portugal e Espanha. Relativamente a trabalhos realizados por pares sobre o MIBEL, estes incidem sobre previsões do preço da eletricidade, e não na comparação entre mercados. Por exemplo em [34], são usados modelos ARIMA-GARCH, e em [35] redes neuronais artificiais de forma a prever o preço no mercado diário.

Capítulo 4. Extração e Tratamento dos Dados

Este capítulo tem como objetivo apresentar as fontes dos dados utilizados bem como o pré-processamento, transformação e criação de variáveis e a definição da variável dependente. No que diz respeito aos dados utilizados, optou-se por utilizar repositórios de acesso público, nomeadamente o *esios*[10], o *mibgas*[11] e o *entso-e*[12]. A descrição e o horário de publicação de cada variável considerada estão detalhados no Glossário das variáveis no Anexo A. O *esios* é o sistema de informação do operador do mercado gerido pela *Red Eléctrica de Espanha* (REE), cujas principais funções incluem executar processos que permitam a exploração segura e económica do sistema elétrico espanhol em tempo real. Como operador do sistema elétrico, a REE tem a obrigação de tornar públicos os dados do setor. Desta forma, são-lhe comunicados resultados de várias fontes, como é o caso do OMIE que fornece os dados referentes ao preço de eletricidade nos mercados diários e intradiários. O *mibgas* é o operador de mercado organizado do gás, na península ibérica, desde 2015. Um dos seus principais objetivos e funções passa pela transparência e igualdade de condições com todos os seus agentes participantes. Como resultado, o *mibgas* publica diariamente os preços e volumes negociados de cada produto do mercado. Por fim, o *entso-e* é uma associação para a cooperação dos operadores de sistemas de transmissão de eletricidade europeus. A sua missão é garantir a segurança do sistema elétrico em todos os horizontes temporais na Europa. Assim, de forma genérica, as variáveis utilizadas são referentes tanto à previsão da energia, que pode ser gerada pelas mais diversas formas de produção (eólica, solar, derivados de petróleo e carvão, cogeração e ciclo combinado), como à geração real referente a essas unidades de produção, para Espanha. Também foram consideradas para este trabalho as procuras previstas e programadas de energia elétrica em Espanha e trocas de energia entre Portugal e França (importações e exportações). De seguida, são detalhadas as técnicas de pré-processamento, criação de novas variáveis e pressupostos aplicados aos dados para o nosso estudo.

[10] *esios* – último acesso a 10 de abril 2024, disponível em https://www.esios.ree.es/es
[11] *mibgas* - último acesso a 10 de abril 2024, disponível em https://www.mibgas.es/pt
[12] *entso-e* - último acesso a 10 de abril 2024, disponível em https://www.entsoe.eu/

4.1 Pré-processamento dos Dados

A linha temporal de dados considerado para a análise encontra-se entre o dia 04/12/2020 à 01h00 *CET*, e o dia 06/12/2023 às 00h00 *CET*, com uma granularidade horaria. A razão pela qual foi considerado este período deve-se ao comportamento distinto do preço da eletricidade neste período, relativamente a datas anteriores, tanto no mercado diário (DA) como no primeiro mercado intradiário (ID01), conforme mostra a *Figura 4.1*.

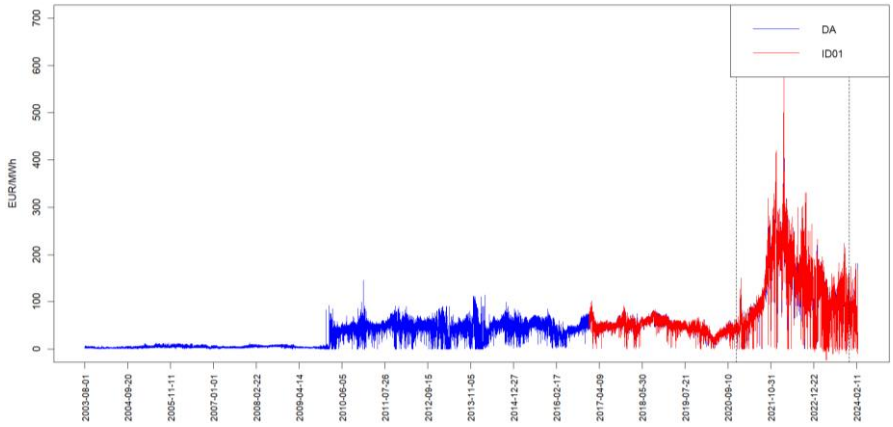

Figura 4.1 Valor em EUR/MWh do preço da eletricidade no mercado Diário e Intradiário

A variável referente ao preço do gás (ID número 27 no Anexo A), presente na base de dados do operador de mercado do gás, o *mibgas*, é referente ao preço diário da *commodity* gás. Uma vez que a granularidade considerada no estudo é horária e não diária, procedeu-se à transformação da variável diária em horas, sob a suposição de que o preço é uniforme ao longo do dia. Vale a pena ressaltar que os valores diários dos preços são referentes ao período das 06h00 *CET* do dia D+1 às 05h00 *CET* do dia D+2. Quanto à variável relacionada com as disponibilidades hídricas, (ID número 28 no Anexo A), a informação foi retirada da base de dados *entso-e*, tendo esta variável registos semanais. Assim, para a análise realizada foi necessária uma transformação semelhante à aplicada na variável do preço do gás, sob a mesma suposição de valor constante ao longo da semana, a fim de ajustar a granularidade. Para todas as restantes variáveis não foi necessário fazer qualquer alteração em relação a este aspeto, uma vez que cada entrada se apresentava de forma horária. No que diz respeito a valores ausentes, como os registados nas variáveis geração PBF ciclo combinado (ID número 25 no Anexo A), geração programada PBF solar térmico (ID número 23 no Anexo A), geração programada

PVP solar térmico (ID número 24 no Anexo A) e geração PBF carvão (ID número 26 no Anexo A), estes foram considerados com o valor de zero. Esta abordagem pressupõe que, quando uma observação está ausente, não houve produção de energia nesse período, ou seja, zero megawatt-hora (MWh) foram gerados nesse período.

No que se refere a *outliers*, optou-se por não se realizar nenhum tratamento específico. Tomou-se esta decisão porque, em princípio, os dados não contêm erros de natureza humana, quer estes sejam de transcrição ou outros, uma vez que provêm de fontes confiáveis. Além disso, no contexto do estudo em questão, torna-se importante avaliar fenómenos extremos para garantir que os modelos possam ser treinados adequadamente, facilitando assim a tomada de decisões. Por outro lado, embora o número de observações pareça substancial (26 304), os dados apresentam várias características que ocorrem de forma pouco frequente. A título de exemplo, o horário das 12h00 *CET* do dia 25 de dezembro está representado apenas três vezes no conjunto de dados considerado. Assim, ao manter os *outliers*, está a ser preservada a integridade dos dados originais e a importância de entender e incorporar eventos que são excecionais na análise.

Quando se lida com variáveis com diversas unidades de medida como MW (*megawatts*), MWh (*megawatt*-hora) e €/MWh (euros por *megawatt*-hora), é importante reconhecer que cada uma destas unidades representa interpretações e escalas distintas. Enquanto MW representa uma unidade de energia, €/MWh indica o preço associado a essa unidade de energia ao longo do tempo. Comparar diretamente estas variáveis pode ser enganoso e erróneo. De forma a mitigar este efeito, procedeu-se à estandardização das variáveis. Isto garante que todas tenham uma média de zero e uma variância de um, colocando-as numa escala comum e comparável, sem que percam as suas características iniciais.

4.2 Criação de Novas Variáveis e Variável Dependente

De forma a enriquecer a análise, foram concebidas variáveis adicionais, visando ampliar a compreensão do caso em estudo. Entre elas, destacam-se a variável Buraco Térmico e variáveis circulares, cuja descrição será detalhada nesta secção. Também se procedeu à criação da variável dependente, através da transformação de duas variáveis: preço da eletricidade para o mercado diário (DA) e intradiário (ID01).

O Buraco Térmico é uma variável criada com o objetivo de simplificar e reduzir o número total de variáveis, de forma a moderar o custo computacional e o tempo de treino de cada modelo. Esta é obtida através da transformação linear de outras variáveis, que posteriormente serão retiradas de forma a evitar problemas de multicolinearidade. As variáveis usadas são referentes a previsões e a fórmula aplicada foi a seguinte [36]:

$$BT = PP - PE - GS - GN - CTI + CTE$$

Com: BT - Buraco Térmico;

PP - Procura prevista (ID número 2 no Anexo A);

PE - Previsão da produção eólica peninsular (ID número 4 no Anexo A);

GS - Geração solar prevista (ID número 5 no Anexo A);

GN - Potência disponível de geração nuclear (ID número 3 no Anexo A);

CTI - Capacidades de troca previstas importações com Portugal e França (CTI = ID 8 + ID10 no Anexo A);

CTE - Capacidades de troca previstas exportações com Portugal e França (CTE = ID9 + ID11 no Anexo A).

Esta variável relaciona a procura prevista da eletricidade com as produções ou gerações de eletricidade proveniente de fontes de energia renováveis, considerando também a energia nuclear[13]. Estas acarretam custos marginais baixos, de tal forma que "empurram" a curva da oferta para a direita e por consequência fazem reduzir o preço da eletricidade. Desta forma, o Buraco Térmico representa a quantidade de energia prevista que terá de ser gerada por energias não renováveis e que está sujeita ao custo dos seus combustíveis e às emissões de dióxido de carbono [17][36].

Um aspeto que se torna particularmente relevante, no caso em estudo, é o tempo. Uma das formas para se captar esta propriedade é transformando-o numa variável circular ou angular [37]. Deste modo, o que distingue este tipo de variáveis das demais é a sua caraterística periódica e a não existência de um início ou fim estabelecido. De forma

[13] A energia nuclear, apesar de não renovável demonstra uma eficiência significativa em termos de emissões de gases de efeito estufa, quando comparada com os combustíveis fósseis, sendo que também não está tão dependente do preço da sua matéria-prima.

prática, se considerarmos um relógio analógico, a diferença de tempo entre a 01h00 e as 03h00 é a mesma que o das 23h00 e a 01h00, ou seja, estão ambas separadas por duas horas. Em [38], são enunciadas aplicações desta técnica com base em artigos de vários outros autores, em áreas como a meteorologia, com o objetivo de captar as direções do vento, ecologia, de forma a mapear o movimento de animais, e até na identificação de padrões temporais (diários e semanais) em incidentes criminais e policiais, de forma a contribuir para a gestão dos recursos policiais e detetar mudanças nos padrões dos incidentes.

No mercado da eletricidade o preço, a procura e a produção da energia elétrica assumem dinâmicas específicas de sazonalidade e periodicidade. Por exemplo, a procura e a produção de energia admitem picos em diferentes horas do dia, dias da semana e do mês, e por conseguinte, torna-se conveniente incorporar estas características nos modelos. A inclusão das variáveis circulares foi igualmente recomendada pela equipa *DS&D* da Galp. Assim, procedeu-se à criação das novas variáveis, através da data e hora que cada observação representa, e da sua localização no círculo trigonométrico. Desta forma, para conseguir mapear estes valores no círculo são necessárias duas componentes, o cosseno e o seno, como demostra a seguinte fórmula:

$$\left[cos\left(2\pi\frac{t}{T}\right);\ \ sin\left(2\pi\frac{t}{T}\right)\right]$$

Com: t - Tempo atual da variável temporal em questão, por exemplo hora do dia;

T - Período total do ciclo, por exemplo 24 horas.

A hora do dia, o dia da semana, o dia do ano e a estação do ano são as variáveis que foram consideradas neste contexto. As suas descrições e fórmulas de calculo estão detalhadas e apresentadas no glossário das variáveis circulares presente no Anexo A. De referir que as variáveis circulares não foram alvo de estandardização, uma vez que o seu contradomínio já está contido entre -1 e 1 e têm média em torno de zero.

Relativamente à variável dependente, procedeu-se à subtração entre o preço da eletricidade do *day-ahead* (ID número 29 no Anexo A) e o preço da *commodity* no primeiro mercado *intraday* (ID número 30 no Anexo A) de forma a perceber qual a discrepância de preço nos dois mercados. Posteriormente, o resultado da subtração foi convertido numa variável binária, do seguinte modo:

$$\begin{cases} 0, & se\ DA > ID01 \\ 1, & se\ DA \leq ID01 \end{cases}$$

Portanto, se a variável dependente, doravante designada como *legenda*, assumir o valor 0, isso indicará que o mercado mais vantajoso para investimento e, consequentemente, para adquirir a *commodity*, é o primeiro mercado intradiário, uma vez que seu o preço seria inferior. Por outro lado, se *legenda* for 1, isso revela que o preço no primeiro mercado intradiário ou é superior, e desta forma mais vantajoso adquirir a eletricidade no mercado diário, ou igual ao preço da *commodity* no mercado diário. Neste último caso, é indiferente em qual mercado adquirir energia elétrica. Assim, obtém-se um problema de classificação binário, onde a variável dependente é categórica, e as variáveis independentes constituem variáveis quantitativas.

4.3 Pressupostos Adotados

Tratando-se este estudo da aplicação de técnicas multivariadas em dados reais, foi necessário recorrer a algumas simplificações de forma a conseguir dar resposta ao problema, sem que houvesse uma alteração significativa do âmbito do estudo.

Um pressuposto adotado foi o de que todos os dias do ano são constituídos por 24 horas. Na verdade, devido a eventos como a mudança de horário de verão e de inverno, ou nas normas de tempo, alguns dias do ano podem apresentar uma hora a mais ou a menos. Nesse sentido, os dias 31/10/2021, 30/10/2022 e 29/10/2023 apresentam uma hora adicional, sendo necessário excluir uma hora a cada um deles. Para esse efeito, foi calculada a média aritmética da hora repetida na série temporal das variáveis (02h00 *CET*). Em contrapartida, nos dias 28/03/2021, 27/03/2022 e 26/03/2023 foi adicionada uma hora que estava ausente (02h00 *CET*), sendo atribuído o valor da variável correspondente à mesma hora do dia anterior, uma vez que as variáveis, na sua generalidade, apresentam uma maior correlação com o "*lag*" de 24 horas, como será discutido mais adiante. No que se refere à variável do preço da eletricidade no mercado intradiário, não existem valores referentes ao dia 13/10/2023 e por consequência, foi mimetizada a ideia proposta anteriormente, sendo utilizado o valor do preço da eletricidade do dia anterior para cada hora.

Adicionalmente, foi necessário realizar uma aproximação dos valores de variáveis que não estão disponíveis até à divulgação do preço da eletricidade nos mercados diário

(DA) e intradiário (IN01). Ou seja, no dia D, os operadores do mercado de eletricidade ibérico disponibilizam o preço e a quantidade de eletricidade transacionada a cada hora do dia D+1, às 12h20 *CET* para o mercado DA e da mesma forma, para cada hora do dia D+1 às 15h20 *CET* no caso do mercado ID01. No entanto, os dados só são publicados oficialmente no site do *esios* algumas horas depois, no caso do DA por volta das 14h00 *CET* e no ID01 às 16h00 *CET*. Relativamente às variáveis que representam previsões da geração de eletricidade, nas suas diversas formas de produção, podem ser obtidas antes das 12h00 *CET*. Já as variáveis com informações de caráter real só estão disponíveis após, ou muito próximo, do horário relativo ao conhecimento das componentes que constituem a variável *legenda*, como representado na *Figura 4.2*. Portanto, torna-se evidente a dificuldade de se realizarem previsões sobre qual mercado seria mais vantajoso quanto à aquisição da *commodity* eletricidade, uma vez que as variáveis podem não estar disponíveis para utilização imediata. Isso é particularmente relevante ao se considerar a implementação futura do modelo para produção. Desta forma, procedeu-se a um *"shift"* das variáveis que sofrem destas condições. Para este deslocamento foi considerado um período de 24 horas, por ser o *"lag"* que tem uma maior autocorrelação na generalidade das variáveis que apresentam esta característica, conforme se pode ver no conjunto de figuras presente no Anexo C. Assim, para prever qual o mercado mais vantajoso no dia D+1, são utilizados dados das variáveis, que não estão disponíveis até às 12h00 CET, referentes ao dia D, uma vez que representam a melhor aproximação disponível no momento. Este pressuposto faz com que o início do *dataset* utilizado avance um dia e passe para 05/12/2020 à 01h00 *CET*.

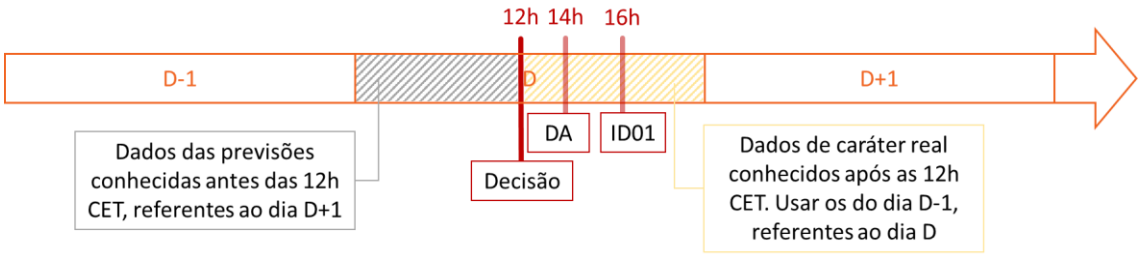

Figura 4.2 Diagrama do horário de publicação dos dados no esios

Capítulo 5. Metodologia

Este capítulo está dividido em quatro secções, onde são apresentados os métodos de análise de dados utilizados neste estudo. As três primeiras contêm uma breve descrição dos modelos empregues, análise discriminante (AD), regressão logística (RL) e redes neuronais artificiais (RNA). Na quarta e última, são discutidos os indicadores de desempenho utilizados no estudo, com o objetivo de comparar e avaliar os diferentes modelos.

5.1 Análise Discriminante

A análise discriminante (AD) é uma técnica clássica de análise multivariada, introduzida em 1936 por *Fisher*. Em [39] o autor introduz esta técnica de forma a distinguir espécies de flores através de características físicas que as plantas apresentam, com a finalidade de classificar novas plantas. Atualmente, a AD é amplamente utilizada em diversos domínios, incluindo medicina, por exemplo para classificar se um paciente está em risco de desenvolver uma determinada doença [40][41], a psicologia [42] ou a análise de risco de crédito [43], entre outras. A AD é aplicada em contextos muito variados devido à sua flexibilidade em classificar observações de diferentes grupos com base em múltiplas variáveis.

A AD é empregue quando as variáveis independentes são quantitativas e a variável dependente é categórica, tipicamente com dois grupos, resultando num caso binário [44][45]. No entanto, a análise pode ser expandida para mais de dois grupos, conhecida também como análise discriminante múltipla. A AD tem dois principais objetivos: (i) identificar quais são as variáveis que têm um maior poder para discriminar os grupos, (ii) criar uma função discriminante que tem como objetivo classificar novas observações. De forma semelhante à regressão linear múltipla, a AD busca encontrar uma combinação linear das variáveis independentes, mas que melhor distinga os grupos definidos, maximizando as diferenças entre eles, como sugerido pela *Figura 5.1*. Consequentemente, a função discriminante corresponde ao eixo para o qual a separação dos grupos é máxima.

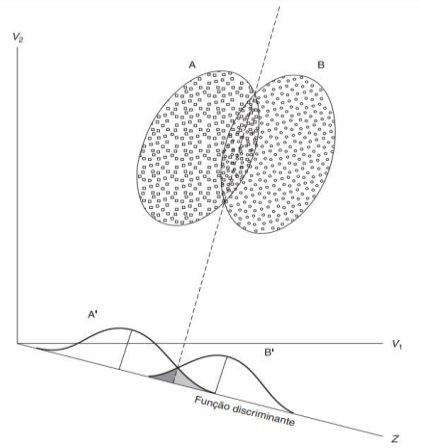

Figura 5.1 Ilustração gráfica da análise discriminante de dois grupos, retirada de [44]

A aplicação desta técnica requer que sejam cumpridos alguns pressupostos, de forma a não comprometerem a classificação, estimação e interpretação. Assume-se que as variáveis independentes seguem uma distribuição normal multivariada; para avaliar essas condições devem-se realizar testes de normalidade às variáveis (*Shapiro-Wilks*). Cumpridos os testes, é possível avaliar o poder que as variáveis independentes têm para discriminar a variável *legenda*. O *teste-t* de diferença de médias, que pressupõe uma população normal, é usado de forma a perceber se uma determinada variável tem poder individual de discriminar os grupos. Por sua vez, o teste de *Wilks* avalia de forma conjunta a igualdade de médias das variáveis, ou seja, o poder discriminante conjunto de todas as variáveis. Se a normalidade multivariada das variáveis não for respeitada, como se verifica neste estudo, a análise discriminante pode ser empregue numa ótica de comparação de resultados com outros métodos.

No caso em estudo foi conduzida uma AD aos dados, com o propósito de conseguir classificar a variável *legenda*. O **R** foi o software usado e, especificamente, a biblioteca *MASS* de forma a estimar os parâmetros do modelo de AD. A função discriminante é constituída por uma combinação linear das variáveis originais, para a qual a separação dos grupos é máxima, dada por:

$$Z = w_1X_1 + w_2X_2 + \cdots + w_nX_n$$

Com: $w_1, w_2, \ldots, w_n$ – Conjunto dos coeficientes a estimar para cada variável independente;

$X_1, X_2, \ldots, X_n$ – Conjunto de variáveis independentes, com poder discriminante.

Os pesos das variáveis são estimados de forma a maximizar a soma dos quadrados entre os grupos e minimizar a soma dos quadrados dentro de grupos, mais especificamente o objetivo é maximizar λ dado por:

$$\lambda = \frac{SS_B}{SS_W}$$

Com: SS_B – Soma dos quadrados das observações, entre as classes;

SS_W – Soma dos quadrados das observações, dentro das classes.

Em suma, através da métrica λ, é possível obter os valores dos coeficientes da função discriminante e assim proceder à obtenção dos *scores* de cada observação, tendo como fim a classificação das classes. Neste contexto, é necessário definir a partir de que valor é plausível classificar uma observação num ou noutro grupo, com base nos scores gerados, surgindo assim o conceito de *cutoff*. Este valor é utilizado para dividir o espaço de discriminação em duas regiões, de modo a permitir a classificação em dois grupos. A técnica usada neste trabalho assume que as dimensões dos grupos são aproximadamente iguais (12 993 observações com o valor da variável *legenda* igual a 0 e 13 311 observações com valor de 1), pelo que o *cutoff* é definido por:

$$cutoff = \frac{Z_1 + Z_2}{2}$$

Com: Z_1, Z_2 – O valor médio dos *scores* para o grupo 1 e 2, respetivamente.

5.2 Regressão Logística

Tal como a AD, a Regressão Logística (RL) é uma técnica multivariada supervisionada, ou seja, existe um conhecimento prévio da variável que se deseja prever, a variável *legenda*. No entanto, ao contrário da AD, a RL não impõe restrições de normalidade multivariada e é frequentemente recomendada para superar tais limitações. Assim a RL estima a probabilidade de uma variável categórica pertencer a um grupo, dado um conjunto de variáveis independentes [44]. Esta técnica tem aplicações no mesmo âmbito da AD, ou seja, problemas de classificação, por exemplo em áreas como a medicina, na deteção precoce de distúrbios como o autismo [46], ou na identificação de características relevantes para posterior previsão de casos de cancro da mama [47]. Pode também ser usada na análise de risco de crédito [48] e ainda pode ser estendida a contextos

sociais, como por exemplo na previsão do risco de quedas entre idosos, permitindo uma triagem mais precisa e promovendo intervenções de saúde direcionadas [49].

No caso em estudo, o objetivo da aplicação da RL é perceber qual a probabilidade da variável dependente, *legenda*, assumir o valor um, ou seja, investir no mercado DA em detrimento do ID01. No modelo RL a variável dependente é a função *logit* da probabilidade que se pretende determinar, dada pelo logaritmo natural entre a razão da probabilidade de que a variável dependente seja um e o seu complementar, ou seja, o logaritmo das chances da variável dependente tomar valor de 1, conforme a seguinte equação:

$$logit\ (p) = \ln\left(\frac{p}{1-p}\right)$$

Com: p – Probabilidade de investimento no mercado DA, por apresentar um preço inferior do que no mercado ID01, dadas as variáveis dependentes $X_1, X_2, \dots, X_n$.

A RL é representada da seguinte forma [44]:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Com: $\beta_0, \beta_1, \beta_2, \dots, \beta_n$ – Coeficientes da regressão que representam o efeito das variáveis independentes na variável *legenda*;

$X_1, X_2, \dots, X_n$ – Valores das variáveis independentes para cada observação.

Resolvendo em ordem a p, obtém-se diretamente os valores de probabilidade:

$$p = \frac{1}{1 + e^{-\beta_0 - \beta_1 X_1 - \beta_2 X_2 - \cdots - \beta_n X_n}}$$

Através do pacote em **R**, *stats*, mais especificamente o comando *glm (Generalized Linear Models)* é possível estimar os parâmetros dos modelos de RL, através do método da máxima verossimilhança. De forma a avaliar o ajuste do modelo aos dados, é importante ter em atenção algumas medidas, tais como o critério de informação de *Akaike* (AIC) [44] que pode ser útil de forma a comparar modelos e onde valores mais baixos representam um melhor ajuste; o *pseudo-*R^2 de *McFadden* [50] que avalia a adequação global do modelo, onde maiores valores representam uma melhor adequação; por fim, os coeficientes também podem ser alvo de testes à significância individual.

De forma a comparar diferentes modelos com as variáveis disponíveis, procedeu-se à criação de três cenários dentro da RL, descritos abaixo, dependendo da forma como as variáveis independentes são selecionadas:

A. Utilização da totalidade das 29 variáveis da base de dados inicial;

B. Redução do número de variáveis através da Análise de Componentes Principais;

O método multivariado da análise dos componentes principais (ACP) é uma técnica usada no pré-processamento dos dados. A ACP Baseia-se num problema de mudança de base e visa reduzir o número de variáveis originais, contendo o máximo de informação possível. As novas variáveis, designadas por componentes principais (CP), são uma combinação linear das variáveis originais, de tal forma que são ortogonais entre si e como resultado não são correlacionadas umas com as outras, mitigando problemas de multicolinearidade nos dados. Esta metodologia visa maximizar iterativamente a variabilidade de cada componente criada [44].

A percentagem de variabilidade explicada pelas CP vai reduzindo à medida que cada nova CP é criada. O número máximo de CP é igual ao número de variáveis originais consideradas, que por sua vez explicam a totalidade da variabilidade presente nos dados originais. A ideia é considerar apenas as CP que apresentam maior quantidade de informação. Para aplicação desta técnica, é importante e necessário estandardizar os dados, caso contrário, as variáveis que têm uma maior variação contribuirão de forma mais significativa para a ponderação das CP, em vez de captar as relações importantes dentro dos dados [51].

Uma questão implicitamente subjacente quando se pretende usar a ACP para a redução de variáveis é o número de CP que devem ser retidas numa ACP. Conforme delineado em [52], diversas abordagens podem ser adotadas para responder a este problema, algumas das quais são mais plausíveis e fáceis de entender e outras são baseadas em testes formais de hipóteses. O critério de *Kaiser*, desenvolvido em 1960 pelo autor que lhe dá o nome, é aplicado em matrizes de correlação, i.e., dados estandardizados, com a premissa de que apenas os CP que apresentam um valor próprio

superior a um devem ser mantidos. Este critério, que é o usado neste trabalho, sugere que tais componentes contêm mais informação (variância) do que uma única variável original dos dados. Existem outros métodos, como o proposto por *Horn* em 1965 ou *Cattell* em 1966 através da observação do "cotovelo" da curva no gráfico *scree plot*, no entanto envolve algum grau de subjetividade.

C. Redução do número de variáveis através de uma análise de multicolinearidade e correlação com a variável dependente;

Antes de proceder à especificação do modelo de RL, é importante realizar uma análise de multicolinearidade e correlação das variáveis com a variável *legenda*. Idealmente, as variáveis independentes devem, por um lado, estar muito correlacionadas com a variável dependente, por outro, apresentarem pouca correlação entre si. Identificar a multicolinearidade entre variáveis independentes não se resume apenas à análise da matriz de correlações, uma vez que a colinearidade pode ser causada pelo efeito combinado das variáveis [44]. Como resultado, surgem medidas para avaliar este efeito. A tolerância representa a quantidade de variabilidade da variável independente selecionada e que não é explicada pelas restantes variáveis independentes, enquanto o fator de inflação da variância *(VIF)* é calculado como o inverso da tolerância. Facilmente se conclui que existem sinais de multicolinearidade quando existem valores de tolerância baixos, ou seja, valores elevados de *VIF*. *Joseph F. Hair et al.* [44] indicam que um valor comum de corte para o *VIF* seria de 10, métrica usada neste estudo, que corresponde a uma tolerância de 0,1.

No que respeita à correlação entre as variáveis independentes e a variável *legenda*, foi usada a correlação *biserial*. Esta é uma estimativa do coeficiente de correlação linear de *Pearson* quando a variável dependente é dicotómica, como é o caso em estudo. De acordo com [53], as suposições básicas para a utilização da técnica são que ambas as variáveis correlacionadas são realmente mensuráveis de forma contínua, mas uma das duas é reduzida a categórica [54].

5.3 Redes Neuronais Artificiais

As Redes Neuronais Artificiais (RNA) têm uma forte inspiração na biologia, pois de certa forma mimetizam as sinapses entre neurónios do cérebro humano, com o objetivo de processar grandes quantidades de dados. O primeiro marco neste campo ocorreu em 1943, quando Warren McCulloch, neurofisiologista, e Walter Pitts, matemático, descreveram o neurónio como uma unidade que recebe inputs binários $X_1, X_2, \ldots X_p$, ponderados pelos respetivos pesos $W_1, W_2, \ldots W_p$ e geram um output também ele binário, como representado na *Figura 5.2* [55]. Posteriormente Donald Hebb em 1949, propôs uma importante tese de que os pesos poderiam ser alterados à medida que as redes aprendem diferentes tarefas funcionais. Embora estas ideias possam parecer triviais, estes modelos inspiraram pesquisas futuras na área das redes neurais, de tal forma que nos dias de hoje estão por detrás de vários modelos e interfaces que se usam de forma recorrente nas mais diversas áreas, como medicina, engenharia aeroespacial, análise financeira, telecomunicações, entre outras [56].

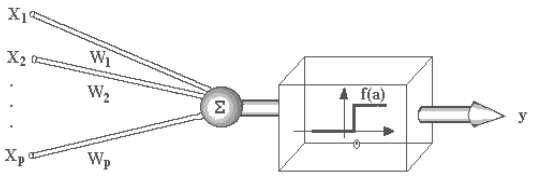

Figura 5.2 Esquema de unidade McCullock-Pitts, disponível em: https://sites.icmc.usp.br/andre/research/neural/

As RNA oferecem uma vasta gama de benefícios no contexto do nosso estudo. Uma das principais vantagens é a capacidade de generalização e a eficiência em lidar com não linearidades e valores extremos nos dados. Também são tolerantes a eventuais falhas, ou seja, mesmo que um neurónio tenha sido danificado, a informação está distribuída e armazenada nos restantes nós da rede, não sendo esperada uma alteração significativa no resultado.

Existem muitas variantes, técnicas e hiperparâmetros a serem ajustados nas RNA, no entanto, não é foco deste estudo enunciá-las. Como tal, serão apenas alvo de análise os métodos aplicados no presente estudo. Para mais informações é recomendada a leitura da obra [55].

No que se refere à topologia das redes aqui usadas, a arquitetura adotada é do tipo *feedforward* multicamadas. Esta estrutura é composta por um conjunto de nós iniciais (*inputs*), tantos quantos o número de variáveis consideradas, e dois nós finais, responsáveis pela classificação do mercado de eletricidade a investir (*output*). Entre essas, existem várias camadas ocultas (*hidden layers*) com os respetivos nós. São assim denominadas pois os seus valores não estão diretamente visíveis, funcionando como uma espécie de "caixa negra". No entanto, cada camada processa um grande número de funções e captura relações não lineares nos dados de entrada, permitindo a resolução de problemas complexos. Esta arquitetura tem também a característica de que o fluxo dos dados circula apenas numa direção, dos nós iniciais para os finais.

Com o objetivo de aplicar técnicas de RNA, recorreu-se ao uso da biblioteca *Keras* do **R**. Esta está integrada com o *TensorFlow*[14], uma API (*Application Programming Interface)*, que apesar de ser uma poderosa ferramenta de *machine learning*, exige conhecimentos prévios de sintaxe e de código específicos. Assim através do *Keras* é possível definir de forma mais simples a estrutura e hiperparâmetros das redes criadas.

As redes desenvolvidas no estudo apresentam uma arquitetura do tipo sequencial, ou seja, cada camada é empilhada uma após a outra, e cada nó é ligado aos nós da camada seguinte. A função de ativação dos nós que estão nas várias camadas ocultas é a *ReLu* de forma a introduzir não linearidades na rede. A camada final, tratando-se de um problema de classificação, é composta por apenas dois nós de saída, que correspondem ao número da classe que o modelo procura prever, e onde de acordo com [56], a função de ativação que deve estar associada aos nós de saída é a *Softmax*, uma vez que tal função normaliza o *output* para uma distribuição de probabilidade.

No que respeita à *loss function*, a *cross-entropy*[15] foi a que se considerou mais adequada para problemas de classificação, uma vez que mede a discrepância entre as probabilidades previstas pelo modelo e as verdadeiras classes das observações. Adicionalmente e de forma a melhorar a eficiência no treino, foi usado o optimizador

[14] Mais informações sobre a API *TensorFlow* em: https://www.tensorflow.org/
[15] Mais especificações da *loss function Cross-entropy* em: https://keras.io/api/losses/probabilistic_losses/

RMSprop[16] que permite uma convergência mais rápida e eficiente do modelo através do histórico da média móvel dos gradientes anteriores e usa essa média para estimar o gradiente. Além do referido, comparativamente a outros optimizadores o *RMSprop*, parece revelar-se superior, *Figura 5.3*.

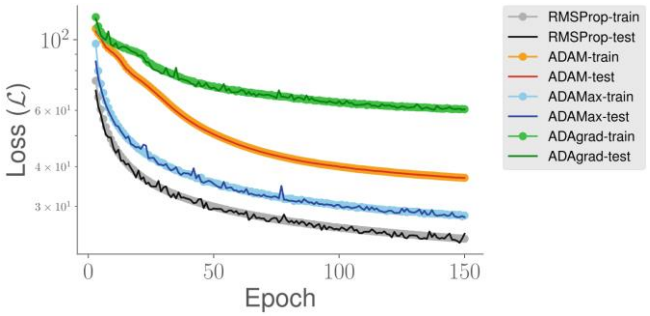

Figura 5.3 Desempenho de optimizadores, no decorrer de 150 épocas, retirado de [57]

O número máximo de épocas considerado foi de 50. Em cada época todas as observações na amostra de treino são usadas para atualizar os pesos da rede e o termo de erro de cada nó (*bias*), permitindo um ajuste gradual ao longo das épocas. Relativamente ao hiperparâmetro de *batch size* foi considerado um valor de 50, ou seja, os parâmetros são atualizados a cada 50 observações na amostra de treino.

As RNA, contudo, enfrentam um problema bastante comum: o *overfitting*. Este fenómeno traduz-se num excessivo ajuste aos dados de treino, impedindo que a rede generalize adequadamente os resultados para a amostra de teste ou outros dados. Com o objetivo de mitigar este problema, foram consideradas três técnicas que permitem identificar, ao longo do treino em várias épocas, sinais de *overfitting*:

- Amostra de Validação

A amostra de treino é posteriormente dividida de forma a gerar a amostra de validação, correspondendo a 20% dos dados de treino. Os restantes 80% dos dados de treino foram utilizados para calcular os gradientes e as atualizações dos pesos a cada iteração, *Figura 5.4*. A amostra de validação é usada como indicador de desempenho da rede. Quando se observarem variações muito grandes entre os resultados, quer da *loss function*, quer da *accuracy*, entre a amostra de treino e a de validação, então existem indícios de *overfitting*.

[16] Mais especificações do optimizador *RMSprop* em: https://keras.io/api/optimizers/rmsprop/

Figura 5.4 Divisão da base de dados para os modelos de RL, AD e RNA

- *Dropout*

A ideia por detrás do *dropout* é a possibilidade de se desativarem temporariamente nós que se encontram nas camadas ocultas da rede. Isto evita que as unidades da rede se adaptem excessivamente aos dados de treino, como referido em [58]. Os nós são desativados de forma aleatória, o que faz com que estes se tornem unidades mais independentes promovendo uma aprendizagem mais robusta e generalizável para receber a amostra de teste.

- *Early-Stopping*

O método da interrupção antecipada[17] é uma técnica usada para controlar e evitar o *overfitting*, como demonstrado na *Figura 5.5*. Durante o treino, é possível verificar na amostra de treino que o erro da rede, em princípio, vai diminuindo à medida que as épocas passam. No entanto, durante esse processo, pode ocorrer *overfitting*. Idealmente, a amostra de validação deverá acompanhar esta diminuição do erro. Assim, no caso em estudo, se a função perda aumentar durante três épocas consecutivas, na amostra de validação, o treino é interrompido e os pesos que produziram um menor valor da *loss function* no conjunto de validação são usados como parâmetros do modelo. Deve-se ter em consideração que o erro de validação pode não diminuir de forma contínua, e apresentar mínimos locais, tornando a análise mais desafiante.

[17]Mais especificações da técnica *early-stopping* em: https://keras.io/api/callbacks/early_stopping/

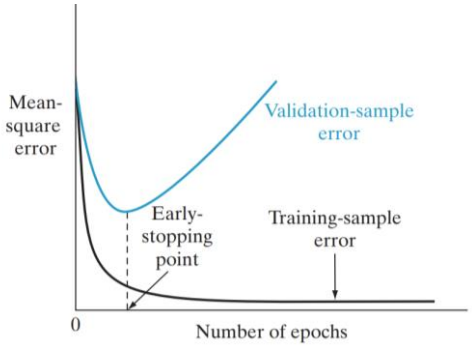

Figura 5.5 Ilustração do early-stopping, retirado de [55]

5.4 Indicadores de Desempenho

Após a escolha dos modelos, é de extrema importância compará-los e usar métricas para definir quais são os que geram uma melhor performance. Aqui são apresentadas as técnicas de medição de desempenho mais frequentes no que toca a problemas de classificação binária. As métricas utilizadas são, *accuracy*, *precision*, *recall* e *F1 score* [59], todas elas obtidas a partir da matriz de confusão. Esta descreve o desempenho do modelo de classificação no conjunto de teste dos dados. A matriz de confusão, *Tabela B*, é organizada em linhas e colunas, sendo que as linhas representam as observações reais e as colunas as classes previstas pelos modelos. Desta forma, se o problema for binário, como é o caso, podem ser distinguidos quatro quadrantes:

- *True positive* (*TP*), onde se prevê corretamente que o preço no mercado ID01 é inferior ao do DA;
- *True negatives* (*TN*), onde se prevê corretamente que o preço no mercado DA é inferior ou igual ao do ID01;
- *False positive* (*FP*), onde se prevê de forma errada que o preço da *commodity* no mercado ID01 é inferior ao do DA, quando na realidade ele é superior ou igual;
- *False negatives* (*FN*), onde se prevê de forma errada que o preço da eletricidade no mercado DA é inferior ou igual ao do ID01, quando na realidade é superior.

Tabela B - Matriz de Confusão

		Previsão	
		DA > ID01	DA ≤ IN01
Real	DA > ID01	*True positive*	*False negatives*
	DA ≤ IN01	*False positive*	*True negatives*

A partir desta matriz é possível construir as métricas pretendidas:

- A *accuracy* avalia a proporção de classificações corretas obtidas e é dada por:

$$accuracy = \frac{TN + TP}{TP + FN + FP + TN}$$

- A *precision* avalia quantas das observações o modelo previu como positivas e quantas na realidade são verdadeiramente positivas. Assim, esta métrica é obtida através da seguinte formula:

$$precision = \frac{TP}{TP + FP}$$

- O *recall*, ou sensibilidade, mede a fração de instâncias positivas que são corretamente classificadas e é obtido como:

$$recall = \frac{TP}{TP + FN}$$

- Por fim, o e *F1 score* corresponde à média harmónica da *precision* e do *recall*, quando o parâmetro que determina os pesos relativos das duas medidas, na métrica *F score*, é igual a 1. O *F1 score* é usado quando os cenários estão desequilibrados, como é o caso do trabalho em análise. É calculado através da seguinte fórmula:

$$F1\ score = 2 \times \frac{precision\ \times recall}{precision + recall}$$

Para efeitos de visualização e de forma a comparar a taxa de verdadeiros positivos, com a taxa de falsos positivos, utilizou-se também a curva *ROC (Receiver Operating Characteristic Curve)*, em que no eixo das ordenadas se representa a sensibilidade e no das abcissas, $1 - especificidade$. A especificidade corresponde à proporção de verdadeiros negativos em relação ao total de casos negativos, dada por:

$$especificidade = \frac{TN}{TN + FP}$$

Através das métricas apresentadas é possível comparar e avaliar os modelos desenvolvidos, de forma a eleger aqueles a serem utilizados por equipas de *trading* de eletricidade. Para o caso em estudo, a métrica mais adequada como critério de decisão é a *accuracy*, uma vez que mede, efetivamente, a taxa de classificações corretas do modelo, para ambos os mercados, DA e ID01.

Capítulo 6. Discussão dos Resultados

Neste capítulo serão discutidos os resultados que resultaram da aplicação da metodologia proposta aos dados descritos no Capítulo 4, para o mercado de eletricidade espanhol, que como já foi referido, refletem o preço da energia elétrica também para Portugal, quando não ocorre a separação dos mercados. O *software* que serviu de base aos cálculos foi o ***R***, com o objetivo de estimar os parâmetros de cada modelo. De forma semelhante ao Capítulo 5, este também será dividido em quatro secções: as primeiras três correspondem à exposição dos resultados para cada modelo utilizado, e por fim, na última, é feita uma comparação dos resultados reunidos.

Um procedimento transversal a todos os métodos foi a divisão da amostra num conjunto de treino e de teste. Desta forma a amostra de treino contém 21 043 observações, correspondendo a 80% da base de dados inicial, e o conjunto de teste conta com 5 261 observações. De referir que a divisão das amostras foi feita de forma aleatória e mantida para todos os modelos, de forma a se realizar uma comparação justa. O nível de significância adotado para os vários testes estatísticos realizados foi de 5%.

6.1 Análise Discriminante

No que se refere à AD, o primeiro passo consistiu em verificar se o pressuposto da normalidade multivariada das variáveis era cumprido. Para tal, foi empregue o teste *Shapiro-Wilks,* no qual a hipótese nula sugere que os dados seguem uma distribuição normal. A *Tabela N*, presente no Anexo B contém os valores da estatística do teste, bem como os respetivos *valores-p*, para cada variável. De acordo com o *output* fornecido, é possível concluir que nenhuma variável segue uma distribuição normal, uma vez que existem fortes evidências para se rejeitar *H0*. Apesar de não se ter cumprido o pressuposto, optou-se por prosseguir com a análise de forma a explorar e comparar os resultados obtidos. Posteriormente, foram realizados testes à igualdade de médias, com o intuito de perceber se cada variável possui poder discriminante individual para distinguir os dois grupos na variável *legenda*, mesmo sabendo que estes testes requerem que as variáveis assumam uma distribuição normal. Neste teste, com a rejeição da hipótese nula conclui-se que a variável tem poder para distinguir os grupos. Conforme está representado na *Tabela O* no Anexo B, apenas algumas das variáveis originalmente consideradas

apresentam de acordo com o teste, poder discriminante individual, estas são precisamente as que têm um *valor-p* inferior a 0,05. Foram mantidas as 18 variáveis que discriminam os grupos e que correspondem à procura, à geração de eletricidade através de energia solar, petróleo, carvão e ciclo combinado e ainda as variáveis circulares. A variável com o ID 37, estação do ano seno, foi a única variável circular que não conseguiu superar o teste. No entanto, também foi considerada para esta análise. Em seguida, realizou-se o teste *Wilks*, para avaliar se as variáveis consideradas têm poder discriminante de forma conjunta. A hipótese nula do teste suporta a ideia de que, de forma conjunta, a média de cada variável independente é igual para os dois grupos. Como resultado o *valor-p* assume o valor de 2,2e-16, logo para o nível de significância considerado, rejeitamos *H0* e concluímos que as variáveis selecionadas têm poder de discriminar a variável *legenda* de forma conjunta, com a ressalva de que a validade destes testes é discutível, uma vez que não é satisfeita a normalidade multivariada.

Identificadas as variáveis discriminantes, procedeu-se à estimação dos coeficientes da função discriminante, representados na *Tabela P* no Anexo B, através do comando *lda* (*linear discriminant analysis*) do **R**. Assim, é possível classificar a amostra de teste com base nos *scores* gerados e no valor de corte definido. Este último apresenta um valor de -0,00359, no entanto foi considerado o valor arredondado de 0. Nas *Figuras 6.1* e *6.2*, estão representados, para a amostra de teste *(out-of-sample)*, os valores dos *scores* para cada observação e o gráfico de densidade dos dois grupos. Por fim, os resultados da classificação podem ser sumarizados na matriz apresentada na *Tabela C*.

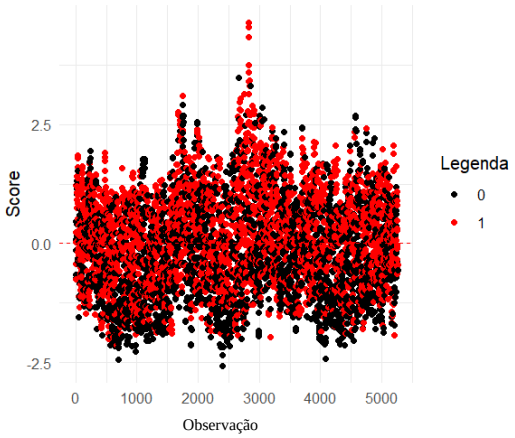

Figura 6.1 Scores na amostra de teste, para cada interação

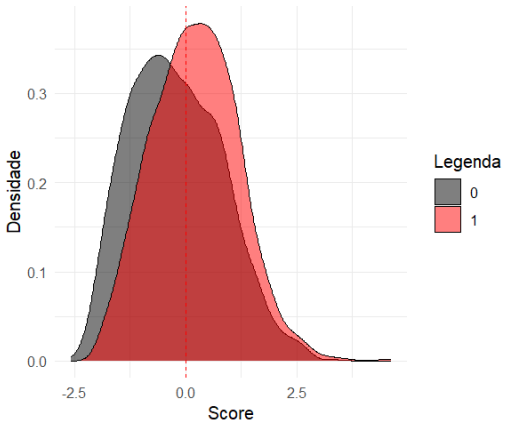

Figura 6.2 Gráfico de densidade para os dois grupos na amostra de teste

Tabela C - Matriz de Confusão da AD

		Previsão	
		DA > ID01	DA ≤ IN01
Real	DA > ID01	1554	1085
	DA ≤ IN01	1105	1517

Através da *Tabela C*, é possível apurar que das 2 639 observações em que o preço da eletricidade no mercado ID01 é inferior ao DA, 1 554 foram classificadas corretamente e 1 085 foram classificadas de forma errada. Quando o preço no mercado DA é inferior ou igual ao do ID01, 1 517 observações foram classificadas corretamente e as restantes 1 105 de forma errada. Foram também calculadas as métricas de desempenho, obtendo-se um valor de *accuracy* de 0,584. No que respeita à métrica de *precision*, foi obtido um valor de 0,584 e, no *recall*, de 0,589. Por fim, na métrica de *F1* score, o valor obtido foi de 0,587.

6.2 Regressão Logística

A RL surge no presente estudo como forma de resolver o problema de classificação pretendido, ultrapassando as limitações da AD no que diz respeito aos pressupostos de normalidade das variáveis independentes e determinação do seu poder discriminante. Conforme delineado no Capítulo 5, foram utilizadas diversas técnicas de seleção de variáveis para este modelo. A seguir, são detalhados os métodos usados, bem como os resultados obtidos:

A. Utilização de todas as 29 variáveis da base de dados inicial;

Neste ponto, foram usadas todas as 29 variáveis consideradas inicialmente, não sendo previamente realizada nenhuma pré-seleção das mesmas. Através do comando *glm (generalized linear model)* do **R**, é possível estimar os coeficientes $\beta_0, \beta_1, \beta_2, \dots, \beta_{29}$ da regressão logística, através do método de máxima verosimilhança. O valor de cada um está apresentado na *Tabela Q* no Anexo B. Com o *output* da estimação é possível perceber que existem algumas variáveis no modelo que são estatisticamente não significativas: através do *t-test*, todas as variáveis que apresentarem um *valor-p* superior a 0,05 são consideradas não significativas. Os coeficientes estimados indicam o impacto que cada

variável tem na variável *legenda*. Por exemplo, para a variável com ID 18, verifica-se que, para cada aumento unitário na previsão de energia fotovoltaica a ser produzida no dia seguinte, e mantendo-se tudo o resto constante, as chances de se adquirir a *commodity* no DA em detrimento do ID01 (*legenda* = 1), aumentam aproximadamente em 8,47%. Esta interpretação é mimetizada para todos os coeficientes da função logística.

Adicionalmente, através da estimação dos coeficientes é possível formular e determinar a probabilidade de cada observação pertencer a uma das classes de *legenda*. Resta apenas definir o valor de *cutoff*. Foram realizados testes adicionais para determinar o valor de corte que maximiza o valor da *accuracy*. Na *Figura 6.3* está representado o valor de *cutoff* que maximiza o valor da métrica na amostra de treino *(in-sample).*

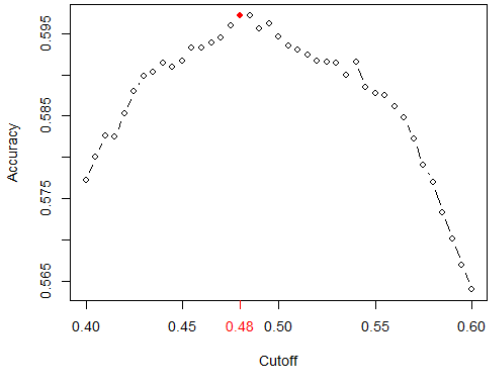

Figura 6.3 Accuracy VS Cutoff, RL in-sample

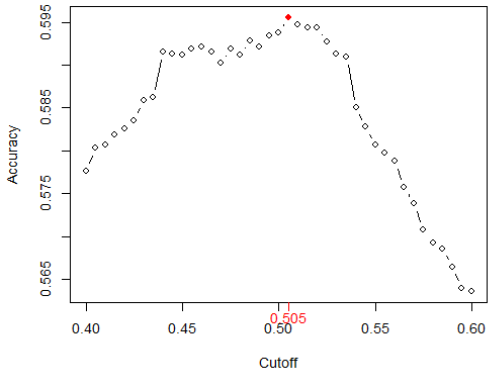

Figura 6.4 Accuracy VS Cutoff, RL out-of-sample

Apesar do valor de corte *in-sample* ser de 0,48, foi considerado um valor de 0,5, pela proximidade destes valores e pelo facto de não ser esse o valor que maximiza a *accuracy out-of-sample*, como demonstrado na *Figura 6.4*. Para efeitos de comparação o *cutoff* será mantido para os restantes modelos de RL. Assim, foi aplicada a função logística à amostra de teste procedendo-se à classificação da mesma, como se pode ver na *Figura 6.5*.

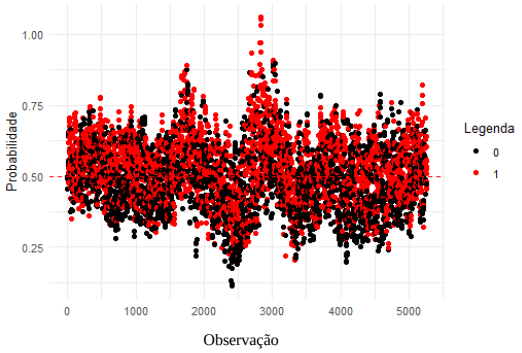

Figura 6.5 Probabilidades previstas pela RL, na amostra de teste, com todas as variáveis

Tabela D - Matriz de Confusão da RL, com todas as variáveis

		Previsão	
		DA > ID01	DA ≤ IN01
Real	DA > ID01	1507	1132
	DA ≤ IN01	1005	1617

A partir da *Tabela D* é possível concluir que das 2 639 observações em que o preço da *commodity* é inferior no mercado ID01, o modelo previu corretamente 1 507 e 1 132 foram classificadas erradamente. Quando o preço é inferior ou igual no DA, 1 617 observações foram classificadas corretamente e 1 005 de forma errada. Quanto às métricas de desempenho a *accuracy* fixou-se nos 0,594, a *precision* em 0,600 e o *recall* em 0,571. Finalmente, na métrica de *F1* score, o valor obtido foi de 0,585.

Uma vez que nem todas as variáveis passaram no *t-test* de significância individual, foi realizado um outro teste de forma que todas as variáveis inseridas no modelo fossem significativas individualmente. Assim e de forma iterativa, foram retiradas as variáveis que apresentavam um maior valor de não rejeição do *t-teste*, i.e., um maior *valor-p*. As variáveis que foram consideradas após este procedimento, bem como os respetivos valores de teste, estão representadas na *Tabela R* no Anexo B.

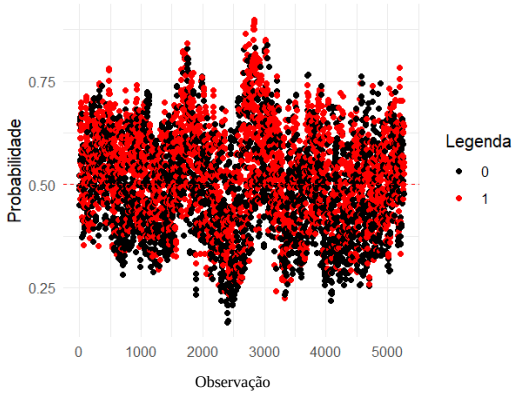

Figura 6.6 Probabilidades previstas pela RL, na amostra de teste, com as variáveis estatisticamente significativas

Tabela E - Matriz de Confusão da RL, com as variáveis estatisticamente significativas

		Previsão	
		DA > ID01	DA ≤ IN01
Real	DA > ID01	1497	1142
	DA ≤ IN01	1002	1620

Na *Figura 6.6* está representada a probabilidade de cada observação da amostra de teste pertencer a uma determinada classe da variável *legenda*. Já através da *Tabela E* é possível calcular as métricas de desempenho: a *accuracy* fixou-se nos 0,592, a *precision* em 0,599, o *recall* em 0,567 e na métrica de *F1* score o valor obtido foi de 0,582.

B. Redução do número de variáveis através da Análise de Componentes Principais;

A utilização da técnica de ACP centra-se na eliminação da multicolinearidade entre as variáveis independentes, bem como a redução do número de variáveis a usar na RL. Por conseguinte, utilizou-se o comando *prvomp*, em **R**, para transformar as variáveis originais num novo conjunto de variáveis, designadas componentes principais (CP). Na *Tabela S* no Anexo B, estão representadas as importâncias que cada CP tem, relativamente à quantidade de informação dos dados (variância). Também é possível retirar a informação de quantas CP devem ser selecionadas. Usando o critério de *Kaiser*, são retidos 7 componentes, como demonstrado na *Figura 6.7*, que contêm 76.18% da variabilidade total das 29 variáveis originais.

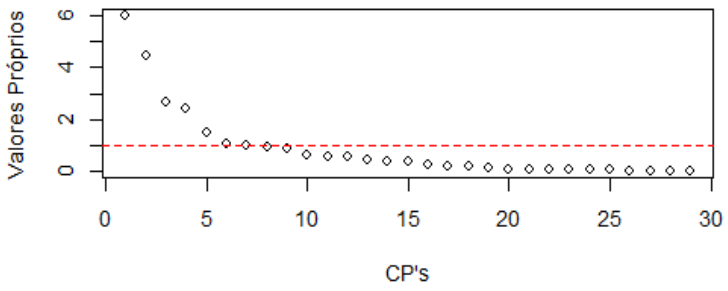

Figura 6.7 Scree plot, critério de Kaiser

De forma complementar, é possível realizar-se uma análise sobre que variáveis originais estão na base de cada CP, os *loadings* representam a correlação entre as variáveis originais e as CP, onde quanto maior esse valor, maior é a influencia dessa variável na formação da CP. Os cinco valores mais elevados estão representados na *Tabela T* no Anexo B. Esta informação é relevante, uma vez que permite dar interpretabilidade às componentes, assim:

- 1ª CP - As variáveis que mais contribuem para a formação desta componente são as ID {7,16,17,23,24}. Estas representam variáveis referentes à geração de energia solar;
- 2ª CP - As variáveis que mais contribuem para a formação da componente são as ID {25,20,27,26,19}. São variáveis relativas à geração de energia por ciclo combinado, carvão e referentes ao preço do gás e a desvios de preço da eletricidade.
- 3ª CP - As variáveis que contribuem para a formação desta componente são as ID {6, 15, 13, 18}. Genericamente representam variáveis referentes à geração de energia eólica;
- 4ª CP - As variáveis que mais contribuem para a formação desta componente são as ID {14, 28, 12}. Estas representam variáveis referentes à geração de energia por cogeração, disponibilidades hídricas e geração através do petróleo ou carvão;
- 5ª CP - As variáveis que mais contribuem para a formação desta componente são as ID {1, 13}. Genericamente representam variáveis referentes à procura programada de energia elétrica e ao preço de banda secundária;
- 6ª CP - As variáveis que mais contribuem para a formação desta componente são as ID {21, 22}, que são indicadores referentes à comparação dos preços dos desvios com o preço no mercado diário;
- 7ª CP - As variáveis que mais contribuem para a formação desta componente são as ID {36, 37}, relacionadas com as variáveis circulares.

Foi conduzida uma RL com as 7 componentes acima referidas. Novamente através do comando *glm* em **R**, foram estimados os parâmetros para a função da RL, estando estes representados na *Tabela U* no Anexo B. É possível observar que as últimas duas CP não são estatisticamente significativas. No entanto foram integradas na análise. No que se refere ao *cutoff*, foi usado o valor de 0,5 de forma a colocar todas as análises em igualdade de circunstâncias. Na *Figura 6.8* está representada a probabilidade de cada observação pertencer a uma determinada classe da variável *legenda*.

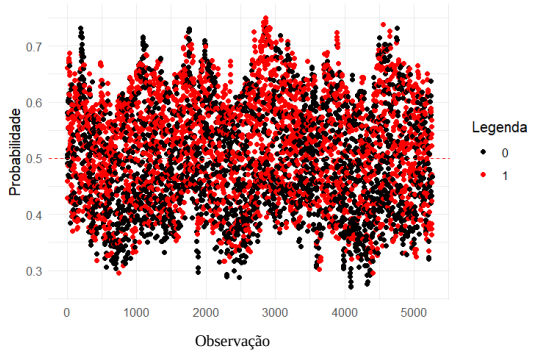

Figura 6.8 Probabilidades previstas pela RL, na amostra de teste, com uso da ACP

Tabela F - Matriz de Confusão da RL, com uso da ACP

		Previsão	
		DA > ID01	DA ≤ IN01
Real	DA > ID01	1480	1159
	DA ≤ IN01	1020	1602

A partir da *Tabela F* conclui-se que das 2 639 observações em que o preço da *commodity* é inferior no mercado ID01, o modelo previu corretamente 1 480 e 1 159 foram classificadas de forma errada. Quando o preço da eletricidade é mais reduzido no DA em detrimento do ID01, 1 602 observações foram classificadas corretamente e 1 020 de forma errada. Quanto às métricas de desempenho a *accuracy* fixou-se nos 0,586, a *precision* em 0,592 e o *recall* em 0,561. Finalmente, na métrica de *F1* score, o valor obtido foi de 0,576.

C. Redução do número de variáveis através de uma análise de multicolinearidade e correlação com a variável dependente;

Dentro da RL, esta é a última técnica de seleção das variáveis utilizada. Inicialmente foi realizada uma análise com o valor de *VIF*, onde iterativamente foi retirada a variável que apresentava um maior valor da métrica, até que todas tivessem um valor inferior ou igual a 10. Desta forma, foram retiradas as variáveis com ID {17, 23, 7, 15}. De seguida, avaliou-se quais as variáveis que apresentavam um maior valor de correlação *biserial* com a variável dependente *legenda*. Foi considerado um valor de corte de 0,05 e, conforme a *Figura 6.9*, as variáveis que prosseguiram para a construção do modelo foram as ID {18, 1, 16, 24, 25, 34, 31, 32}. Também se acrescentou a variável circular com ID 33, uma vez que só a combinação das duas, ID 32 e ID 33, faz sentido na análise. Estas variáveis estão associadas à procura de eletricidade, à geração da *commodity* através de energia solar térmica, fotovoltaica e ciclo combinado, a acrescer as variáveis circulares.

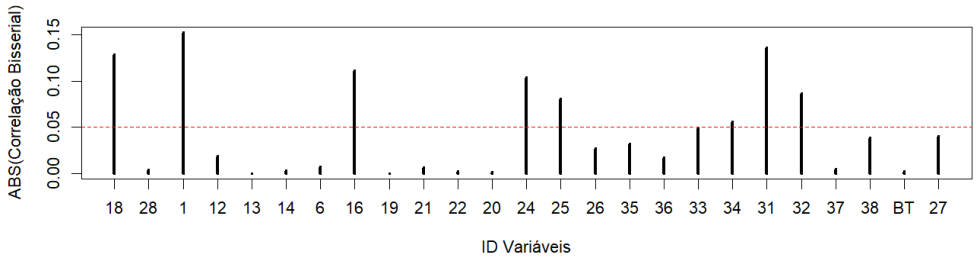

Figura 6.9 Valor absoluto da correlação biserial para cada variável

Constituída a seleção das variáveis, passou-se à estimação dos parâmetros do modelo, apresentadas na *Tabela V*, no Anexo B. É possível perceber que as variáveis com ID {16, 24, 32, 33}, não são estatisticamente significativas, no entanto permaneceram na análise. Na *Figura 6.10* estão apresentados, para a amostra de teste, o valor das probabilidades de cada observação pertencer a um determinado grupo.

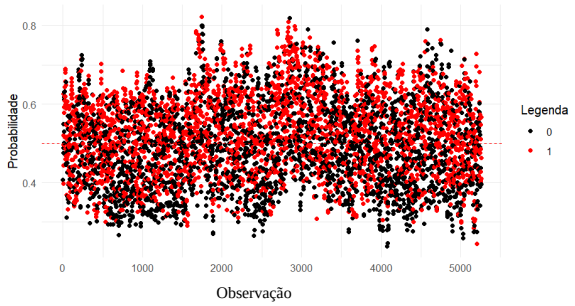

Figura 6.10 Probabilidades previstas pela RL, na amostra de teste, com estudo da multicolinearidade e correlação

Tabela G - Matriz de Confusão da RL, com estudo da multicolinearidade e correlação

		Previsão	
		DA > ID01	DA ≤ IN01
Real	DA > ID01	1488	1151
	DA ≤ IN01	1006	1616

A partir da *Tabela G* é possível concluir que o modelo previu corretamente 1 488 observações onde o preço no mercado ID01 é inferior e 1 616 quando o preço no mercado DA é inferior ou igual ao mercado ID01. No que ser refere às métricas de desempenho, a *accuracy* fixou-se nos 0,590, a *precision* em 0,597 e o *recall* em 0,564. Finalmente, na métrica de *F1* score, o valor obtido foi de 0,580.

6.3 Redes Neuronais Artificiais

A última técnica que surge no estudo são as RNA. Conforme discutido no capítulo 5, esta técnica apresenta vantagens em relação aos modelos descritos anteriormente. Desta forma, procedeu-se à criação dos modelos de RNA, através da biblioteca *Keras*[18], uma

[18] Mais sobre o *Keras* em: https://keras.io/about/, último acesso a 03 junho 2024

API escrita na linguagem de programação *Python*, mas que também pode ser executada em **R**.

Para a formulação dos modelos de RNA, foram aplicadas as técnicas de estandardização e a utilização de todas as 29 variáveis disponíveis. A arquitetura de cada rede foi definida da seguinte forma:

$$ARQ\ (i, h_1, h_2, \dots, h_n, o)(d);$$

Onde i representa o número de entradas na rede (*inputs*), n representa o número de camadas ocultas (*hidden layer*), h é o número de nós que cada uma contém, o representa as camadas de saída (*output*) (no caso serão sempre 2) e por fim d representa a percentagem de nós que são desativados temporariamente na rede (*dropout*). Além da divisão tradicional da amostra em amostra de teste e de treino, subdividiu-se a amostra de treino em amostra de treino e amostra de validação, com 20% das observações destinadas à validação. Assim a amostra de treino passa a conter 16 834 observações e a amostra de validação 4 209 observações. Como referido na metodologia, os pesos são ajustados de 50 em 50 observações (*batch size*), portanto, no total os pesos são atualizados 337 vezes para cada época. Em seguida, serão dispostas as arquiteturas que foram testadas, bem como os respetivos resultados e valores das métricas geradas:

A. $ARQ\ (29, 128, 64, 64, 2)(20)$

Nesta arquitetura o treino foi interrompido na época 28 de forma a prevenir o *overfitting*, relativamente ao número de parâmetros que a rede tem de estimar, estes são de 16 386, incluindo os pesos entre cada nó e o respetivo *bias*. É importante referir que a *loss function* atingiu na época 28, 0,5704 para a amostra de treino e 0,5891 na amostra de validação, como demonstra a *Figura 6.11*. Assim, em *out-of-sample* a RNA obteve uma *accuracy* de 0,693, uma *precision* de 0,703, um *recall* de 0,671 e *F1 score* de 0,687, métricas que são possíveis calcular através da matriz de confusão representada na *Tabela H*.

Figura 6.11 ARQ (29, 128, 64, 64, 2) (20), Loss Function & Accuracy

Tabela H - ARQ (29, 128, 64, 64, 2) (20), Matriz de Confusão

		Previsão	
		DA > ID01	DA ≤ IN01
Real	DA > ID01	1772	867
	DA ≤ IN01	749	1873

B. $ARQ\ (29, 64, 32, 32, 2)(20)$

O número de parâmetros a estimar nesta RNA é de 5 122. Através da *Figura 6.12* é de notar que o treino foi interrompido na época 26. Relativamente à *loss function*, na 26ª época, tem um valor de 0,6094 para a amostra de treino e 0,6129 para a amostra de validação. No que se refere às métricas de desempenho *out-of-sample*, a *accuracy* é de 0,653, a *precision* de 0,665, o *recall* de 0,622 e o *F1 score* de 0,643, *Tabela I*.

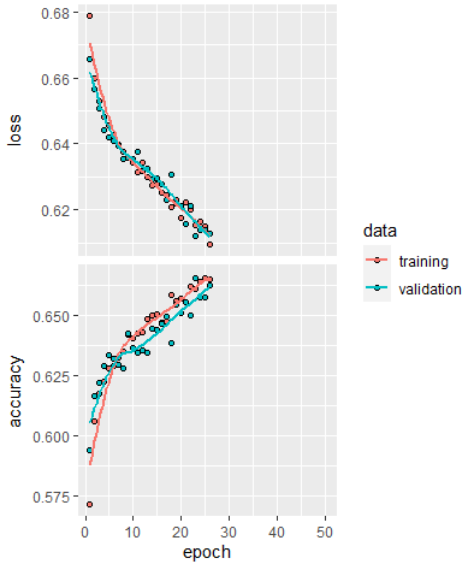

Figura 6.12 ARQ (29, 64, 32, 32, 2) (20), Loss Function & Accuracy

Tabela I - ARQ (29, 64, 32, 32, 2) (20), Matriz de Confusão

		Previsão	
		DA > ID01	DA ≤ IN01
Real	DA > ID01	1642	997
	DA ≤ IN01	826	1796

C. $ARQ\ (29, 64, 64, 2)(0)$

Com a arquitetura definida, o número de parâmetros que o modelo estimou foi de 6 210. Através da *Figura 6.13*, é possível observar que a série teve uma paragem precoce na época 17, onde o valor da *loss function*, atingiu 0,5736 na amostra de treino e 0,6130 na de validação. Por fim, pela *Tabela J* a métrica *out-of-sample* para a *accuracy* foi de 0,667, a *precision* de 0,668, um *recall* de 0,616 e *F1 score* de 0,650.

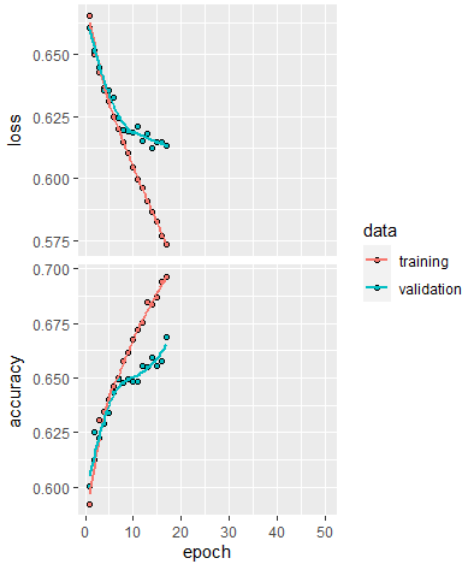

Figura 6.13 ARQ (29, 64, 64, 2) (0), Loss Function & Accuracy

Tabela J - ARQ (29, 64, 64, 2) (0), Matriz de Confusão

		Previsão	
		DA > ID01	DA ≤ IN01
Real	DA > ID01	1626	1013
	DA ≤ IN01	739	1883

D. $ARQ\ (29, 64, 64, 2)(20)$

Nesta arquitetura de RNA, o treino foi interrompido na época 44, *Figura 6.14*, tendo sido alvo de estimações os mesmos 6 210 parâmetros, uma vez que a arquitetura é igual à da alínea anterior. A *loss function* apresentou na última época um valor de 0,5834 para o treino e 0,5958 para a validação. Através da comparação com a arquitetura $ARQ\ (29, 64, 64, 2)(0)$, é possível observar a importância da técnica de *dropout* no que toca à discrepância no valor da *loss funcion* nas amostras de treino e validação. Conclui-se pela *Tabela K* que o valor das métricas relativas à amostra de teste, a *accuracy* fixou-se nos 0.668, a *precision* foi de 0,706, um *recall* de 0,579 e *F1 score* de 0,637.

Figura 6.14 ARQ (29, 64, 64, 2) (20), Loss Function & Accuracy

Tabela K - ARQ (29, 64, 64, 2) (20), Matriz de Confusão

		Previsão	
		DA > ID01	DA ≤ IN01
Real	DA > ID01	1528	1111
	DA ≤ IN01	634	1988

E. $ARQ\ (29, 128, 128, 2)(20)$

O número de parâmetros a estimar pela arquitetura $ARQ\ (29, 128, 128, 2)(20)$ é de 20 610, um valor superior em comparação às outras arquiteturas. O treino desta RNA cessou na época 32 e obteve um valor de 0,5532 na *loss function* para a amostra de treino e 0,5790 na amostra de validação. Relativamente às métricas de desempenho, a *accuracy* foi de 0,696, a *precision* foi de 0,686, o *recall* de 0,726 e *F1 score* de 0,706, *Tabela L*.

Figura 6.15 ARQ (29, 128, 128, 2) (20), Loss Function & Accuracy

Tabela L - ARQ (29, 128, 128, 2) (20), Matriz de Confusão

		Previsão	
		DA > ID01	DA ≤ IN01
Real	DA > ID01	1917	772
	DA ≤ IN01	878	1744

6.4 Comparação dos Resultados Reunidos

Esta secção destina-se à comparação dos resultados obtidos pelas três famílias de modelos. No que se refere à AD, apesar de não cumprir os requisitos de normalidade multivariada das variáveis independentes, resultou em valores muito próximos aos gerados pela RL. Na *Figura 6.16* está representada a curva ROC para a AD, onde idealmente a taxa de verdadeiros positivos é elevada (*sensitivity*) e a taxa de falsos positivos ($1 - especificidade$) é reduzida. Relativamente à RL, são propostos quatro modelos diferentes no que se refere à escolha das variáveis independentes, que necessitam ser comparados. No que concerne à medida AIC, o modelo que inclui todas as variáveis apresenta um valor de 29 321, o modelo que usa apenas as variáveis significativas individualmente apresenta um valor de 27 935, o modelo que utiliza uma ACP tem um valor de 29 781 e por fim o que foi alvo de uma análise da multicolinearidade e correlação tem um valor de 29 571. Assim, o que apresenta um melhor ajuste aos dados de acordo com esta métrica é o que usa apenas as variáveis significativas. No que diz respeito ao *pseudo-*R^2 de *McFadden*, mais uma vez, o modelo que apresenta um melhor ajuste relativo é o que usa todas e apenas, as variáveis individualmente significativas, com um valor de 0,044. Na *Figura 6.17* é possível observar as curvas *ROC* do modelo de RL, para cada uma das técnicas de seleção das variáveis, com os respetivos valores da métrica *accuracy*. A técnica que tem um valor mais elevado da *accuracy* é o que usa todas as 29 variáveis.

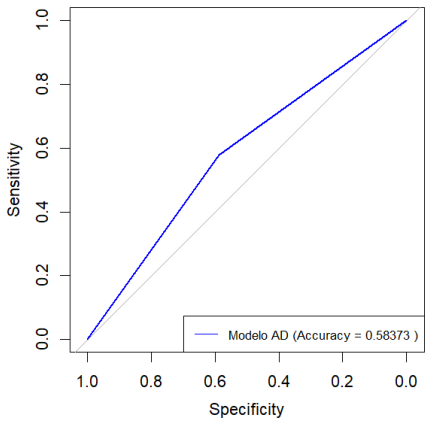

Figura 6.16 Curva ROC, para a AD

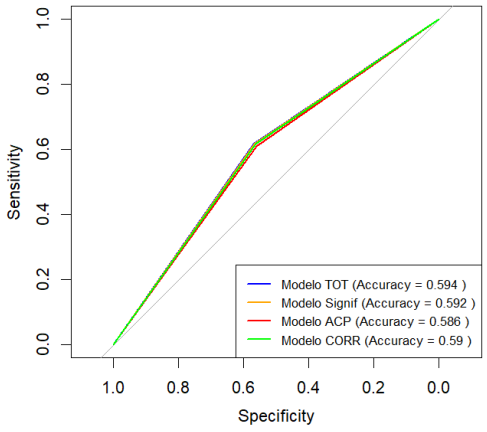

Figura 6.17 Curvas ROC, para a RL

No que concerne aos modelos desenvolvidos com recurso às redes neuronais, estes apresentaram, na globalidade, um valor de *accuracy* superior à AD e à RL. A arquitetura, de entre as consideradas, que maximizou a taxa de acertos, em aproximadamente 70%, foi a $ARQ\ (29, 128, 128, 2)(20)$. Esta apresenta apenas duas camadas ocultas, cada uma com 128 nós. Na *Figura 6.18*, é possível observar uma comparação entre as cinco arquiteturas de RNA consideradas para análise.

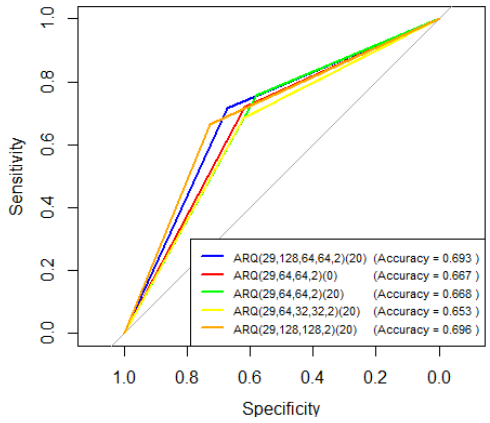

Figura 6.18 Curvas ROC, para a RNA

Ao se analisar os resultados da *Tabela M*, fica evidente que a técnica que apresenta um melhor desempenho com o conjunto específico dos dados apresentados, são as RNA. O valor da métrica *accuracy* para este tipo de modelo é o mais elevado entre todos os modelos considerados. Contudo, vale a pena ressalvar que este método está muito dependente da seleção adequada dos hiperparâmetros, o que pode tornar o processo de ajuste uma tarefa subjetiva e dependente de tentativa e erro. Além disso, é importante destacar a perda de interpretabilidade, que é transversal a este tipo de modelos e que dificulta a compreensão dos padrões e relações subjacentes entre dados.

Tabela M - Quadro resumo dos resultados obtidos para cada modelo

Modelo	**Técnica/Arquitetura**	***accuracy***	***precision***	***recall***	***F1 score***
AD	Variáveis com Poder Discriminante	0.584	0.584	0.589	0.587
RL	Todas as Variáveis	0.594	0.600	0.571	0.585
	Variáveis Significantes	0.592	0.599	0.567	0.582
	Análise Componentes Principais	0.586	0.592	0.561	0.576
	Análise de Multicolinearidade e Correlação	0.590	0.597	0.564	0.580
RNA	ARQ (29,128,64,64,2) (20)	0.693	0.703	0.671	0.687
	ARQ (29,64,32,32,2) (20)	0.653	0.665	0.622	0.643
	ARQ (29,64,64,2) (0)	0.667	0.688	0.616	0.650
	ARQ (29,64,64,2) (20)	0.668	0.706	0.579	0.637
	ARQ (29,128,128,2) (20)	0.696	0.686	0.726	0.706

Capítulo 7. Conclusão e Trabalhos Futuros

O estudo levado a cabo comporta técnicas multivariadas de dados, tais como a AD, RL e RNA, aplicadas a um conjunto de dados específicos do mercado *spot* de eletricidade espanhol, parte integrante do MIBEL. Através das metodologias utilizadas é possível fundamentar a escolha, no dia D, de qual mercado de eletricidade, DA ou IN01, tem um valor do preço mais baixo para o perfil de cada uma das 24 horas do dia D+1.

No que diz respeito aos modelos usados, a violação da hipótese de normalidade das variáveis independentes comprometeu a aplicabilidade da AD, no entanto a mesma foi usada para fins de comparação com as restantes técnicas. Relativamente à RL, foram desenvolvidas várias técnicas de seleção das variáveis independentes, no entanto o resultado da taxa de acertos *out-of-sample* foi bastante similar em todas elas, com um valor médio da métrica *accuracy* de 0,591, nos quatro casos analisados. A técnica que demonstrou um melhor desempenho foi a RNA, que, apesar de apresentar uma metodologia muito especifica e subjetiva na escolha dos hiperparâmetros a fixar, revelou-se ser a mais eficaz de forma a prever o mercado mais vantajoso para se efetuarem negociações de eletricidade, com um valor da métrica *accuracy* de aproximadamente 70%.

É importante referir que, tratando-se o estudo de um mercado onde o preço da *commodity* assume dinâmicas voláteis é necessário estar ciente das relaxações e pressupostos que foram tomados. Destaco como principal limitação a necessidade do "*shift"* realizado às variáveis às quais não dispomos de acesso aos dados antes do mercado DA encerrar (12h00 *CET*). Outra limitação significativa refere-se ao uso de séries temporais em análises multivariadas, o que pode introduzir desafios, como a possível não estacionaridade das variáveis, responsável por alterações na média e variância ao longo do tempo. Para mitigar esses impactos, é essencial avaliar a necessidade de diferenciações nas variáveis e realizar testes de estacionaridade como o de *Dickey-Fuller* aumentado. A acrescer, e uma vez que a crescente geração de eletricidade por fontes renováveis é uma realidade, seria importante, para trabalhos futuros, incorporar variáveis meteorológicas como: velocidade do vento, precipitação, radiação etc., para que os modelos capturarem flutuações na geração proveniente de fontes verdes. Outro conjunto de variáveis que são relevantes para o caso são variáveis macroeconómicas globais. Crescimento económico,

inflação, taxa de câmbio, índice global de paz (GPI), são fatores que direta ou indiretamente têm impacto no preço das *commodities* e que são usadas para a geração da energia elétrica.

Referências Bibliográficas

[1] R. Weron, 'Electricity price forecasting: A review of the state-of-the-art with a look into the future', *International Journal of Forecasting*, vol. 30, no. 4, pp. 1030–1081, Oct. 2014, doi: 10.1016/j.ijforecast.2014.08.008.

[2] Conselho de Reguladores do MIBEL, 'Descrição do Funcionamento do MIBEL', 2009. Accessed: Jan. 27, 2024. [Online]. Available: https://www.mibel.com/wp-content/uploads/2018/08/Descricao_Funcionamento_MIBEL_Marco_2009.pdf

[3] M. Neves, 'A Produção de Eletricidade em Portugal em Regime Ordinário: Evolução e Perspectivas', 2017. Accessed: Feb. 05, 2024. [Online]. Available: https://icjp.pt/sites/default/files/papers/producao_de_electricidade_em_portugal.pdf

[4] T. Baptista, 'A «Liberalização do Mercado Energético em Portugal» - Verdadeira Concorrência?', 2014.

[5] D. B. Pereira, 'Análise do Comportamento dos Preços do Mercado Ibérico no ano de 2018', 2018.

[6] Entidade Reguladora dos Serviços Energéticos, 'Estrutura Tarifária do Setor Elétrico em 2024', 2023. Accessed: Feb. 06, 2024. [Online]. Available: https://www.erse.pt/media/fsgnhkmw/estrutura-tarif%C3%A1ria-se-2024-dez2023.pdf

[7] Parlamento Europeu e do Conselho, 'Directiva 2003/54/CE', 2003, Accessed: Jan. 31, 2024. [Online]. Available: https://eur-lex.europa.eu/resource.html?uri=cellar:caeb5f68-61fd-4ea8-b3b5-00e692b1013c.0010.02/DOC_1&format=PDF

[8] K. Imran and I. Kockar, 'A technical comparison of wholesale electricity markets in North America and Europe', *Electric Power Systems Research*, vol. 108, pp. 59–67, 2014, doi: 10.1016/j.epsr.2013.10.016.

[9] Entidade Reguladora dos Serviços Energéticos, 'Acordo entre a República Portuguesa e o Reino de Espanha relativo à constituição de um mercado ibérico

da energia elétrica', 2006, Accessed: Jan. 31, 2024. [Online]. Available: https://www.mibel.com/wp-content/uploads/2018/07/documentos_SITE_MIBEL_RESOLUCAO_ASS._REPUBLICA_23_2006_67196764.pdf

[10] OMIP, 'Regulamento da Negociação', 2018. Accessed: Feb. 01, 2024. [Online]. Available: https://www.omip.pt/pt/system/files?file=2020-01/omip_regulamento_da_negociacao_29.junho.2018_pt_0.pdf

[11] OMIE, 'Funcionamento do Mercado Diário'. Accessed: Feb. 09, 2024. [Online]. Available: https://www.omie.es/sites/default/files/inline-files/mercado_diario_p_1.pdf

[12] OMIE, 'Detalhes do funcionamento do Mercado Intradiário'. Accessed: Jun. 07, 2024. [Online]. Available: https://www.omie.es/sites/default/files/inline-files/mercados_intradiario_y_continuo_p.pdf

[13] S. Almeida, 'Redesign the European Energy Markets to better allocate the increment of renewable energy suppliers', 2023.

[14] Galp, 'Regenerating the Future, Relatório Integrado de Gestão 2022', 2023, Accessed: Jan. 27, 2024. [Online]. Available: https://www.galp.com/corp/Portals/0/Recursos/Investidores/2023_IR/1Q_RESULTS_2023/GALP_RC22_PT_ESEF.pdf

[15] A. M. Jorge, 'Projeto SmartGalp: Um estudo sobre a inovação na Galp Energia, SA', 2016.

[16] Galp, 'Jornada de Sustentabilidade, Relatório Integrado de Gestão 2022', 2023. Accessed: Jan. 30, 2024. [Online]. Available: https://www.galp.com/corp/Portals/0/Recursos/Investidores/SharedResources/Relatorios/pt/2022/AIRGalp2022PT2Book2SustainabilityJourney.pdf

[17] L. Gelabert, X. Labandeira, and P. Linares, 'An ex-post analysis of the effect of renewables and cogeneration on Spanish electricity prices', *Energy Economics*, vol. 33, pp. S59–S65, Dec. 2011, doi: 10.1016/j.eneco.2011.07.027.

[18] S. G. Jensen and K. Skytte, 'Interactions between the power and green certificate markets', *Energy Policy*, vol. 30, no. 5, pp. 425–435, Apr. 2002, doi: 10.1016/S0301-4215(01)00111-2.

[19] T. Kristiansen, 'Forecasting Nord Pool day-ahead prices with Python', *SINTEF Energy Research*, 2018. Accessed: Jun. 07, 2024. [Online]. Available: https://core.ac.uk/download/pdf/230921024.pdf

[20] B. Uniejewski, J. Nowotarski, and R. Weron, 'Automated variable selection and shrinkage for day-ahead electricity price forecasting', *Energies*, vol. 9, no. 8, 2016, doi: 10.3390/en9080621.

[21] A. J. Conejo, J. Contreras, R. Espínola, and M. A. Plazas, 'Forecasting electricity prices for a day-ahead pool-based electric energy market', *International Journal of Forecasting*, vol. 21, no. 3, pp. 435–462, Jul. 2005, doi: 10.1016/j.ijforecast.2004.12.005.

[22] R. Beigaitė and T. Krilavičius, 'Electricity price forecasting for Nord Pool data', 2017. Accessed: Jun. 07, 2024. [Online]. Available: https://ceur-ws.org/Vol-1856/p07.pdf

[23] M. D. Chinn, M. Leblanc, and O. Coibion, 'The predictive content of energy futures: an update on petroleum, natural gas, heating oil and gasoline', *National Bureau of Economic Research*, Jan. 2005, doi: 10.3386/w11033.

[24] C. Morana, 'A semiparametric approach to short-term oil price forecasting', *Energy Economics*, vol. 23, no. 3, pp. 325–338, May. 2001, doi: 10.1016/S0140-9883(00)00075-X.

[25] J. Cuaresma, J. Hlouskova, S. Kossmeier, and M. Obersteiner, 'Forecasting electricity spot-prices using linear univariate time-series models', *Applied Energy*, vol. 77, no. 1, pp. 87–106, Jan. 2004, doi: 10.1016/S0306-2619(03)00096-5.

[26] J. Contreras, R. Espínola, F. J. Nogales, and A. J. Conejo, 'ARIMA models to predict next-day electricity prices', *IEEE Transactions on Power Systems*, vol. 18, no. 3, pp. 1014–1020, Aug. 2003, doi: 10.1109/TPWRS.2002.804943.

[27] A. J. Conejo, M. A. Plazas, R. Espínola, and A. B. Molina, 'Day-ahead electricity price forecasting using the wavelet transform and ARIMA models', *IEEE Transactions on Power Systems*, vol. 20, no. 2, pp. 1035–1042, May. 2005, doi: 10.1109/TPWRS.2005.846054.

[28] M. Pinhão, 'Iberian energy market: spot price forecast by modelling market offers', May. 2022. Accessed: Jun. 07, 2024. [Online]. Available: https://run.unl.pt/bitstream/10362/144914/1/Pinhao_2022.pdf

[29] M. Pinhão, M. Fonseca, and R. Covas, 'Electricity Spot Price Forecast by Modelling Supply and Demand Curve', *Mathematics*, vol. 10, no. 12, Jun. 2022, doi: 10.3390/math10122012.

[30] J. Che and J. Wang, 'Short-term electricity prices forecasting based on support vector regression and Auto-regressive integrated moving average modeling', *Energy Conversion and Management*, vol. 51, no. 10, pp. 1911–1917, Oct. 2010, doi: 10.1016/j.enconman.2010.02.023.

[31] X. Yan and N. A. Chowdhury, 'Mid-term electricity market clearing price forecasting: A hybrid LSSVM and ARMAX approach', *International Journal of Electrical Power & Energy Systems*, vol. 53, pp. 20–26, Dec. 2013, doi: 10.1016/j.ijepes.2013.04.006.

[32] J. P. S. Catalão, S. J. P. S. Mariano, V. M. F. Mendes, and L. A. F. M. Ferreira, 'Short-term electricity prices forecasting in a competitive market: A neural network approach', *Electric Power Systems Research*, vol. 77, no. 10, pp. 1297–1304, Aug. 2007, doi: 10.1016/j.epsr.2006.09.022.

[33] K. Maciejowska, W. Nitka, and T. Weron, 'Day-Ahead vs. Intraday—Forecasting the Price Spread to Maximize Economic Benefits', *Energies*, vol. 12, no. 4, p. 631, Feb. 2019, doi: 10.3390/en12040631.

[34] L. P. Wilson, 'Desenvolvimento de um modelo de previsão de preços diários MIBEL', 2023. Accessed: Aug. 13, 2024. [Online]. Available: https://www.repository.utl.pt/handle/10400.5/30253

[35] I. Teixeira, 'Um modelo de previsão de preços spot de eletricidade através de redes neuronais artificiais', 2021. Accessed: Aug. 13, 2024. [Online]. Available: https://run.unl.pt/bitstream/10362/120676/1/Teixeira_2021.pdf

[36] AleaSoft, 'Electricity demand, fundamental factor in the electricity market price', 2018, Accessed: Jun. 07, 2024. [Online]. Available: https://aleasoft.com/wp-content/uploads/2018/12/20181210-aleasoft-electricity-demand-fundamental-factor-market-price.pdf

[37] R. P. Mahan, 'Circular Statistical Methods: Applications in Spatial and Temporal Performance Analysis', *United States Army Research Institute*, 1991. Accessed: Jun. 07, 2024. [Online]. Available: https://apps.dtic.mil/sti/tr/pdf/ADA240751.pdf

[38] A. Lee, 'Circular data', *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 477-486, Jul. 2010. doi: 10.1002/wics.98.

[39] R. A. Fisher, 'The Use of Multiple Measurements in Taxonomic Problems', *Annals of Eugenics*, pp. 179–188, 1936, Accessed: Jun. 07, 2024. [Online]. Available: https://digital.library.adelaide.edu.au/dspace/handle/2440/15227

[40] F. Temurtas, K. Gorur, O. Cetin, and I. Ozer, 'Machine learning for thyroid cancer diagnosis', in *Computational Intelligence in Cancer Diagnosis*, *Elsevier*, pp. 117–145, 2023. doi: 10.1016/B978-0-323-85240-1.00011-0.

[41] M. Toğaçar, B. Ergen, and Z. Cömert, 'Application of breast cancer diagnosis based on a combination of convolutional neural networks, ridge regression and linear discriminant analysis using invasive breast cancer images processed with autoencoders', *Med Hypotheses*, vol. 135, pp. 109-503, Feb. 2020, doi: 10.1016/j.mehy.2019.109503.

[42] N. E. Betz, 'Use of discriminant analysis in counseling psychology research.', *Journal of Counseling Psychology*, vol. 34, no. 4, pp. 393-403, Oct. 1987, doi: 10.1037/0022-0167.34.4.393.

[43] G. C. Stefania *et al.*, 'Credit Risk Scoring Model Based on The Discriminant Analysis Technique', *Procedia Computer Science*, vol. 220, pp. 928–933, 2023, doi: 10.1016/j.procs.2023.03.127.

[44] J. F. Hair, W. C. Black, B. J. Babin, R. E. Anderson, and R. L. Tatham, *Multivariate Data Analysis*, 6th ed. Bookman, 2006.

[45] S. Sharma, *Applied Multivariate Techniques*. John Wiley & Sons, 1996.

[46] D. Jayaprakash and C. S. Kanimozhiselvi, 'Multinomial logistic regression method for early detection of autism spectrum disorders', *Measurement: Sensors*, vol. 33, p. 101-125, Jun. 2024, doi: 10.1016/j.measen.2024.101125.

[47] Z. Khandezamin, M. Naderan, and M. J. Rashti, 'Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier', *Journal of Biomedical Informatics*, vol. 111, p. 103-591, Nov. 2020, doi: 10.1016/j.jbi.2020.103591.

[48] M. Lin and J. Chen, 'Research on Credit Big Data Algorithm Based on Logistic Regression', *Procedia Computer Science*, vol. 228, pp. 511–518, Feb. 2024, doi: 10.1016/j.procs.2023.11.058.

[49] P. J. Pan, C. H. Lee, N. W. Hsu, and T. L. Sun, 'Combining principal component analysis and logistic regression for multifactorial fall risk prediction among community-dwelling older adults', *Geriatric Nursing*, vol. 57, pp. 208–216, May. 2024, doi: 10.1016/j.gerinurse.2024.04.021.

[50] D. McFadden, 'Conditional Logit Analysis of Qualitative Choice Behavior', in *Frontiers in Econometrics*, Academic Press, ch. 4, pp. 105–142, 1974.

[51] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, NY: Springer New York, 2013. doi: 10.1007/978-1-4614-6849-3.

[52] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York: Springer Science & Business Media, 2002.

[53] J. P. Guilford, *Fundamental Statistics in Psychology and Education*, 1st ed. McGraw-Hill, 1942.

[54] S. Lira, 'Análise de Correlação: Abordagem Teórica e Construção dos Coeficientes com Aplicações', Universidade Federal do Paraná, Curitiba, 2004. Accessed: Jun. 07, 2024. [Online]. Available:

https://www.ipardes.pr.gov.br/sites/ipardes/arquivos_restritos/files/documento/2019-09/sachiko_dissertacao_2004.pdf

[55] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed. Pearson Education, 2009.

[56] M. T. Hagan, H. B. Demuth, M. H. Beale, and O. Jesús, *Neural Network Design*, 2nd ed. Martin Hagan, 2014.

[57] D. Bhowmik, S. Gao, M. T. Young, and A. Ramanathan, 'Deep clustering of protein folding simulations', *BMC Bioinformatics*, vol. 19, no. S18, pp. 484, Dec. 2018, doi: 10.1186/s12859-018-2507-5.

[58] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov, 'Dropout: A Simple Way to Prevent Neural Networks from Overfitting', *The Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[59] M. Vakili, M. Ghamsari, and M. Rezaei, 'Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification'.

Anexo A – Glossário de variáveis

A. Glossário das variáveis utilizadas.

As informações, para cada variável, foram retiradas das bases de dados do *esios*[19], *mibgas*[20] e do *entso-e*[21]. Todas as horas apresentadas estão no fuso *CET*.

Id	Nome	Descrição	Publicação	Fonte
1	Procura programada (MW)	A procura programada de energia é determinada através da combinação das bolsas internacionais e dos programas de geração dos grupos responsáveis pela oferta de energia nos mercados diário e intradiário.	Diariamente às 00h00 com a informação do dia D e às 13h30 com as três primeiras horas do dia D+1.	*esios*
2	Procura Prevista (MW)	Previsão de consumo considerando os dados registados em períodos anteriores semelhantes, além de outros fatores como horário de trabalho, condições meteorológicas e atividade económica. Os dados apresentados nesse indicador são referentes à Península Ibérica.		*esios*
3	Potencia Disponível de Geração Nuclear (MW)	Calculado, para cada tipo de produção de geração convencional, como a diferença entre a soma da potência líquida instalada registada no sistema da *esios* e o total de potência indisponível declarada pelos sujeitos do mercado. Especificamente, este indicador refere-se a grupos com tipo de produção nuclear.	Diariamente a partir das 4h30, com dados por dia e hora do trimestre atual e do seguinte.	*esios*
4	Previsão Da Produção Eólica Peninsular (MW)	Previsão horária de energia eólica de acordo com o modelo de previsão de vento do operador do sistema.	A cada hora com a previsão desde a hora atual do dia D até a última hora do dia D+1.	*esios*
5	Geração Prevista Solar (MW)	Previsão horária da geração de energia solar através dos modelos de predição. A decomposição deste indicador é composta por previsões para a energia solar térmica e fotovoltaica.	A cada hora com informações do dia D da hora H+1 até o final do dia e todo o dia D+1.	*esios*
6	Geração Real Eólica (MW)	Geração medida em tempo real de energia produzida por meio de fontes eólicas na península ibérica.	A cada 5 minutos com informações das últimas três horas do dia D-1 até o horário atual do dia D.	*esios*
7	Geração Real Solar (MW)	Geração medida em tempo real de energia produzida por meio de fontes solares. Este indicador é composto em energia gerada	A cada 10 minutos com informações das últimas três horas do dia D-1 até	*esios*

[19] Último acesso a 10 de abril 2024, disponível em https://www.esios.ree.es/es
[20] Último acesso a 10 de abril 2024, disponível em https://www.mibgas.es/pt
[21] Último acesso a 10 de abril 2024, disponível em https://www.entsoe.eu/

		como solar térmica e como solar fotovoltaica.	o horário atual do dia D.	
8	Capacidade de Troca Prevista com Portugal Importação (MW)	Este indicador diz respeito à previsão de capacidade na interconexão em MW, com Portugal no sentido PT, ES para cada hora do dia seguinte.	No dia imediatamente anterior ao dia de programação, antes das 11h00.	*esios*
9	Capacidade de Troca Prevista com Portugal Exportação (MW)	Este indicador diz respeito à previsão de capacidade na interconexão em MW, com Portugal no sentido ES, PT para cada hora do dia seguinte.	No dia imediatamente anterior ao dia de programação, antes das 11h00.	*esios*
10	Capacidade de Troca Prevista com França Importação (MW)	Este indicador diz respeito à previsão de capacidade na interconexão em MW, com França no sentido FR, ES para cada hora do dia seguinte.	No dia imediatamente anterior ao dia de programação, antes das 11h00.	*esios*
11	Capacidade de Troca Prevista com França Exportação (MW)	Este indicador diz respeito à previsão de capacidade na interconexão em MW, com França no sentido ES, FR para cada hora do dia seguinte.	No dia imediatamente anterior ao dia de programação, antes das 11h00.	*esios*
12	Geração Programada PBF Derivados de Petróleo ou Carvão (MWh)	Este indicador aborda a programação das unidades de produção que utilizam petróleo ou carvão como fonte de energia.	Diariamente a partir das 13h45, com informações do dia D+1.	*esios*
13	Preço de Banda Secundária (€/MWh)	Este indicador refere-se ao preço marginal resultante das ofertas nas várias unidades de programação de manter a estabilidade e o equilíbrio das flutuações da procura.	Diariamente a partir das 15h30, com informação do dia D+1.	*esios*
14	Geração Programada PBF Cogeração (MWh)	Este indicador refere-se à quantidade de energia que está programada ser gerada pelos diferentes tipos de unidades de produção de energia por cogeração, incluindo contratos bilaterais com entrega física confirmada.	Diariamente a partir das 13h45 com informações do dia D+1.	*esios*
15	Geração Programada PBF Eólica (MWh)	Este indicador é referente à quantidade de energia que está programada ser gerada pelos diferentes tipos de unidades de produção de energia eólica, incluindo contratos bilaterais com entrega física confirmada.	Diariamente a partir das 13h45, com informações do dia D+1.	*esios*
16	Geração Programada PBF Fotovoltaica (MWh)	Este indicador refere-se à quantidade de energia que está programada ser gerada pelos diferentes tipos de unidades de produção de energia solar fotovoltaica, incluindo contratos bilaterais com entrega física confirmada.	Diariamente a partir das 13h45, com informações do dia D+1.	*esios*

17	Geração Programada PVP Fotovoltaica (MWh)	Geração de energia que incorpora as alterações introduzidas no PBF para resolução de restrições técnicas, identificadas por critérios de segurança e posterior reequilíbrio da produção-procura. Este indicador refere-se às unidades de produção do tipo solar fotovoltaica.	Diariamente a partir das 16h00, com informação do dia D+1.	*esios*
18	Previsão Diária D+1 Eólica e Fotovoltaica (MWh)	Este indicador refere-se à previsão diária da geração de energia eólica e energia fotovoltaica para a península ibérica, no dia seguinte.	Diariamente às 11h00 para D+1.	*esios*
19	Desvios De Preço De Pagamento Aumentar (€/MWh)	Os desvios são a diferença entre a energia programada e a energia medida.	Cada hora com a informação disponível até aquele momento do dia D tem um preço provisório. No dia seguinte as informações de todo o dia anterior são atualizadas.	*esios*
20	Desvios de Preço de Pagamento Baixo (€/MWh)	Os desvios são a diferença entre a energia programada e a energia medida.	Cada hora com a informação disponível até aquele momento do dia D tem um preço provisório. No dia seguinte as informações de todo o dia anterior são atualizadas.	*esios*
21	Preço de Desvio Medido Entre o Preço Marginal Diário a Baixar (€/MWh)	Este indicador refere-se ao quociente entre o preço do desvio medido a descer e o preço marginal do mercado diário.	Diariamente às 7h45 com as informações do dia anterior e a partir das 19h30 para dias passados.	*esios*
22	Preço de Desvio Medido Entre o Preço Marginal Diário a Subir (€/MWh)	Este indicador refere-se ao quociente entre o preço do desvio medido para subir e o preço marginal do mercado diário.	Diariamente às 7h45 com as informações do dia anterior e a partir das 19h30 para dias passados.	*esios*
23	Geração Programada PBF Solar Térmico (MWh)	Este indicador refere-se à quantidade de energia que está programada ser gerada pelos diferentes tipos de unidades de produção de energia solar térmica, incluindo contratos bilaterais com entrega física confirmada.	Diariamente a partir das 13h45 com informações do dia D+1	*esios*
24	Geração Programada PVP Solar Térmico (MWh)	Geração de energia que incorpora as alterações introduzidas no PBF para resolução de restrições técnicas, identificadas por critérios de segurança e posterior reequilíbrio da produção-procura.	Diariamente a partir das 16h00, com informação do dia D+1.	*esios*

		Este indicador refere-se às unidades de produção do tipo solar térmica.		
25	Geração PBF Ciclo Combinado (MWh)	Este indicador refere-se à quantidade de energia que está programada ser gerada pelos diferentes tipos de unidades de produção de energia de ciclo combinado, incluindo contratos bilaterais com entrega física confirmada.	Diariamente a partir das 13h45, com informações do dia D+1.	*esios*
26	Geração PBF Carvão (MWh)	Este indicador refere-se à quantidade de energia que está programada ser gerada pelos diferentes tipos de unidades de produção de energia de carvão, incluindo contratos bilaterais com entrega física confirmada.	Diariamente a partir das 13h45 com informações do dia D+1	*esios*
27	Preço do Gás Day-Ahead (€/MWh)	Índice diário de gás a partir do preço de referência do período de negociação, do produto diário (*Day-Ahead*) com entrega no PVB (Ponto de Balanço Virtual) - Espanhol, segundo o dia de entrega.	Diariamente a partir das 21h30, valores diários correspondentes ao período das 06h00 do dia D+1 às 05h00 do dia D+2	*mibgas*
28	Reservatórios de Água e Usinas Hidrelétricas (MW)	Taxa de enchimento média semanal agregada de todos os reservatórios de água e centrais hidroelétricas (MWh) em Espanha, incluindo o valor da mesma semana do ano anterior.	A informação é publicada no terceiro dia útil seguinte à semana a que a informação diz respeito.	*entso-e*
29	Preço do Mercado Spot Diário (DA) (€/MWh)	O Mercado Diário é um mercado no qual se estabelecem transações de energia elétrica para o dia D+1, por meio da apresentação de ofertas de venda e aquisição de energia elétrica por parte dos participantes do mercado. No caso de Portugal e Espanha é o MIBEL que fornece estes dados através do cruzamento da oferta-procura de eletricidade na região.	Diariamente, por volta das 14h00, com informação referente ao dia D+1.	*esios*
30	Preço do Mercado Spot Intradiário (ID01) (€/MWh)	Este mercado tem por objeto atender, por meio da apresentação de ofertas de venda e aquisição de energia elétrica por parte dos sujeitos do mercado diário, os ajustes que são necessários sobre o programa diário referido no ID 29. Atualmente existem seis sessões diárias de mercado intradiário. Esta concretamente diz respeito ao preço da primeira sessão deste mercado. Tal como no DA, a fonte da informação é o MIBEL do polo Espanhol (OMIE).	Diariamente, por volta das 16h00, com informação referente ao dia D+1.	*esios*

B. Glossário das variáveis circulares:

ID	Nome	Descrição	Fórmula	Contradomínio
31	Hora do dia Seno	Através da hora de cada iteração é possível determinar a posição	$\sin\left(2\pi\frac{hora}{24}\right)$	$[-1,1]$

32	Hora do dia Cosseno	temporal relativa de cada uma dentro do dia. Posteriormente, para cada observação foram criadas duas variáveis que representam a posição da hora do dia que ocupam no círculo trigonométrico: seno e cosseno.	$\cos\left(2\pi\frac{hora}{24}\right)$	
33	Dia da semana Seno	Através da data de cada iteração é possível determinar a posição temporal relativa de cada uma dentro semana. Posteriormente, para cada observação foram criadas duas variáveis que representam a posição do dia da semana que ocupam no círculo trigonométrico: seno e cosseno.	$\sin\left(2\pi\frac{dia\ semana}{7}\right)$	
34	Dia da semana Cosseno		$\cos\left(2\pi\frac{dia\ semana}{7}\right)$	
35	Dia do ano Seno	Através da data de cada iteração é possível determinar a posição temporal relativa de cada dia dentro do ano. Posteriormente, para cada observação foram criadas duas variáveis que representam a posição do dia do ano que ocupam no círculo trigonométrico: seno e cosseno. *Tratando-se o ano de 2022 um ano bissexto foi considerado no denominador da fórmula um valor de 366 dias, para o respetivo ano.	$\cos\left(2\pi\frac{dia\ ano}{(365^{*})}\right)$	
36	Dia do ano Cosseno		$\cos\left(2\pi\frac{dia\ ano}{(365^{*})}\right)$	
37	Estação do ano Seno	Através da data de cada iteração é possível atribuir a que estação do ano pertence. Posteriormente, para cada observação foram criadas duas variáveis que representam a posição da estação do ano que ocupam no círculo trigonométrico: seno e cosseno.	$\cos\left(2\pi\frac{estação\ ano}{4}\right)$	
38	Estação do ano Cosseno		$\cos\left(2\pi\frac{estação\ ano}{4}\right)$	

Anexo B – *Outputs*

Tabela N - Teste Shapiro-Wilks

ID da Variável	Valor do Teste	*Valor-p*
18	0.97187	2.2e-16
28	0.96732	2.2e-16
1	0.98933	2.2e-16
12	0.94536	2.2e-16
13	0.64851	2.2e-16
14	0.92949	2.2e-16
15	0.95508	2.2e-16
6	0.95251	2.2e-16
7	0.76862	2.2e-16
16	0.76577	2.2e-16
17	0.77905	2.2e-16
19	0.94677	2.2e-16
21	0.00907	2.2e-16
22	0.01148	2.2e-16
20	0.93505	2.2e-16
23	0.81150	2.2e-16
24	0.80577	2.2e-16
25	0.83461	2.2e-16
26	0.79074	2.2e-16
35	0.90131	2.2e-16
36	0.90441	2.2e-16
33	0.88462	2.2e-16
34	0.84002	2.2e-16
31	0.89285	2.2e-16
32	0.89520	2.2e-16
37	0.80792	2.2e-16
38	0.80749	2.2e-16
Buraco Térmico	0.99714	3.934e-08
27	0.87928	2.2e-16

Tabela O - Test-t à igualdade de médias, Análise Discriminante

ID da Variável	Valor do Teste-*t*	*Valor-p*
18	-19.749	0
28	0.27108	0.78634
1	-22.81061	0
12	-2.84205	0.00449
13	-0.47083	0.63777
14	0.32287	0.7468
15	-0.72202	0.47029
6	-1.37444	0.16932
7	-16.73032	0
16	-16.71679	0
17	-16.62352	0
19	0.22664	0.82071
21	1.0291	0.30345
22	0.43737	0.66185
20	-0.14026	0.88846
23	-15.74512	0
24	-15.03648	0
25	-12.05813	0
26	-3.80104	0.00014
35	4.25539	2e-05

36	-2.4938	0.01265
33	7.49047	0
34	8.32312	0
31	19.61607	0
32	13.37169	0
37	0.76646	0.44341
38	-5.19679	0
Buraco Térmico	0.1387	0.88969
27	-5.74095	0

Tabela P - Coeficientes da Função Discriminante

ID da Variável	Coeficientes da Função Discriminante
18	0.54910400
1	0.25455748
12	0.07310560
7	0.15304900
16	1.25023680
17	-1.52485997
23	1.29946599
24	-1.09326290
25	0.46282025
26	-0.15103329
35	-0.12450122
36	0.23648543
33	-0.06985469
34	-0.36152487
31	-0.37356927
32	-0.06022715
37	0.07236321
38	-0.05225251
27	0.03971337

Tabela Q - Coeficientes estimados para a RL, com todas as variáveis consideradas

ID da Variável	Estimativa	*Std. Error*	*t-test*	*Valor-p*
(intercept)	0.5079758	0.0033518	151.552	2.2e-16
18	0.0846746	0.0088518	9.566	2.2e-16
28	0.0934999	0.0092813	10.074	2.2e-16
1	-0.0128773	0.0070119	-1.836	0.06630
12	-0.0010376	0.0076291	-0.136	0.89182
13	0.0138538	0.0042059	3.294	0.00099
14	0.0360785	0.0084028	4.294	1.77e-05
15	0.0419461	0.0178635	2.348	0.01888
6	-0.0263896	0.0171764	-1.536	0.12446
7	0.0060826	0.0228719	0.266	0.79028
16	0.1996788	0.1345010	1.485	0.13767
17	-0.1909385	0.1364589	-1.399	0.16176
19	-0.0025389	0.0089114	-0.285	0.77572
21	-0.0040339	0.0030194	-1.336	0.18156
22	-0.0007520	0.0030090	-0.250	0.80266
20	-0.0130146	0.0085426	-1.523	0.12765
23	0.1376968	0.0330749	4.163	3.15e-05
24	-0.1314320	0.0326274	-4.028	5.64e-05
25	0.0840586	0.0062114	13.533	2e-16
26	-0.0040906	0.0058486	-0.699	0.48430
35	-0.1118166	0.0139280	-8.028	1.04e-15

36	0.0316711	0.0121983	2.596	0.00943
33	-0.0109983	0.0054429	-2.021	0.04333
34	-0.0313640	0.0051848	-6.049	1.48e-09
31	-0.0483281	0.0076701	-6.301	3.02e-10
32	0.0006255	0.0097641	0.064	0.94892
37	0.0227018	0.0111681	2.033	0.04209
38	-0.0176183	0.0112478	-1.566	0.11728
Buraco Térmico	0.0266363	0.0079911	3.333	0.00086
27	0.0524692	0.0075103	6.986	2.91e-12

Tabela R - Coeficientes estimados para a RL, com todas as variáveis significantes de forma individual

ID da Variável	Estimativa	*Std. Error*	*t-test*	*Valor-p*
(intercept)	0.03626	0.01424	2.546	0.01091
18	0.38358	0.03321	11.550	2e-16
28	0.39121	0.03514	11.132	2e-16
1	-0.05386	0.02729	-1.973	0.04845
13	0.05748	0.01753	3.278	0.00105
14	0.13415	0.02219	6.044	1.50e-09
15	0.05141	0.02130	2.414	0.01580
20	-0.06138	0.02565	-2.393	0.01669
23	0.77163	0.12888	5.987	2.13e-09
24	-0.70732	0.12680	-5.578	2.43e-08
25	0.34454	0.02586	13.324	2e-16
35	-0.40725	0.04405	-9.245	2e-16
36	0.09423	0.04457	2.114	0.03450
33	-0.04805	0.02245	-2.140	0.03232
34	-0.13270	0.02185	-6.072	1.26e-09
31	-0.19729	0.02677	-7.371	1.70e-13
37	0.09282	0.04720	1.966	0.04925
Buraco Térmico	0.11777	0.03308	3.560	0.00037
27	0.20253	0.02617	7.738	1.01e-14

Tabela S - Importância das CP

Componente Principal	*Standard Deviation*	*Proportion of Variance*	*Cumulative Proportion*
PC1	2.4487	0.2398	0.2398
PC2	2.1133	0.1786	0.4185
PC3	1.6260	0.1058	0.5242
PC4	1.5570	0.0970	0.6212
PC5	1.2126	0.0588	0.6800
PC6	1.0215	0.0417	0.7218
PC7	1.0000	0.0400	0.7618
PC8	0.9776	0.0382	0.8000
PC9	0.9363	0.0351	0.8351
PC10	0.7885	0.0249	0.8599
PC11	0.7558	0.0229	0.8828
PC12	0.7359	0.0217	0.9044
PC13	0.6615	0.0175	0.9220
PC14	0.6275	0.0158	0.9377
PC15	0.5862	0.0138	0.9514
PC16	0.5047	0.0102	0.9616
PC17	0.4495	0.0081	0.9697
PC18	0.4330	0.0075	0.9772
PC19	0.3627	0.0053	0.9825
PC20	0.2983	0.0036	0.9860
PC21	0.2957	0.0035	0.9895

PC22	0.2639	0.0028	0.9923
PC23	0.2630	0.0028	0.9951
PC24	0.2118	0.0018	0.9969
PC25	0.1909	0.0015	0.9983
PC26	0.1370	0.0008	0.9991
PC27	0.1250	0.0006	0.9997
PC28	0.0835	0.0003	0.9999
PC29	0.0174	0.00001	1.0000

Tabela T – Valores dos 5 Loading's mais elevados das 7 CP

Componente Principal	**ID das 5 variáveis com um maior valor de *Loading***	***Loading***
PC1	7	0.37713
	16	0.37412
	17	0.37405
	23	0.35012
	24	0.34616
PC2	25	0.34677
	20	0.33999
	27	0.33442
	26	0.32269
	19	0.31691
PC3	6	0.41412
	15	0.40671
	13	0.34712
	18	0.32656
	12	0.28886
PC4	14	0.55085
	28	0.37457
	12	0.33849
	1	0.26060
	35	0.25954
PC5	1	0.57362
	13	-0.44178
	Buraco Térmico	0.27105
	25	0.24577
	33	-0.22529
PC6	21	-0.71383
	22	-0.69653
	19	-0.02634
	27	0.02378
	31	-0.02114
PC7	36	0.37067
	37	0.32556
	24	-0.29546
	23	-0.29173
	13	-0.27621

Tabela U - Coeficientes estimados para a RL, com uso da ACP

Componente Principal	**Estimativa**	***Std. Error***	***t-test***	***Valor-p***
(intercept)	0.507979	0.003384	150.114	2e-16
PC1	0.020973	0.001383	15.170	2e-16
PC2	0.018331	0.001601	11.450	2e-16
PC3	0.021393	0.002083	10.271	2e-16

PC4	0.007066	0.002176	3.247	0.00117
PC5	0.049379	0.002789	17.702	2e-16
PC6	0.004178	0.002977	1.404	0.16046
PC7	-0.001790	0.003379	-0.530	0.59630

Tabela V - Coeficientes estimados para a RL, com estudo da multicolinearidade e correlação

ID da Variável	**Estimativa**	***Std. Error***	***t-test***	***Valor-p***
(intercept)	0.507968	0.003367	150.864	2e-16
18	0.061319	0.004936	12.423	2e-16
1	0.028939	0.005046	5.735	9.87e-09
16	-0.008825	0.008593	-1.027	0.304
24	0.003199	0.006182	0.517	0.605
25	0.043794	0.003852	11.370	2e-16
34	-0.039294	0.004766	-8.245	2e-16
31	-0.048410	0.006927	-6.989	2.85e-12
32	-0.012340	0.008908	-1.385	0.166
33	-0.008712	0.005354	-1.627	0.104

Anexo C – Gráficos Auxiliares

Todos os *lags* estão em representação temporal horaria.

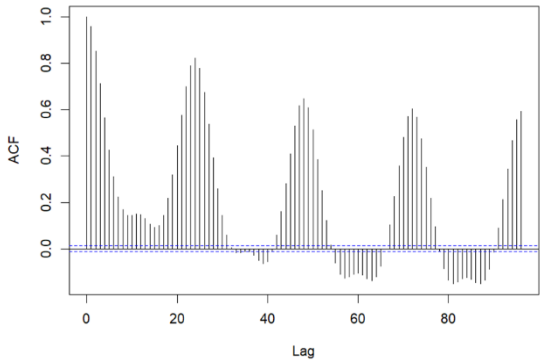

Figura Anexo C.1 Autocorrelação da procura programada

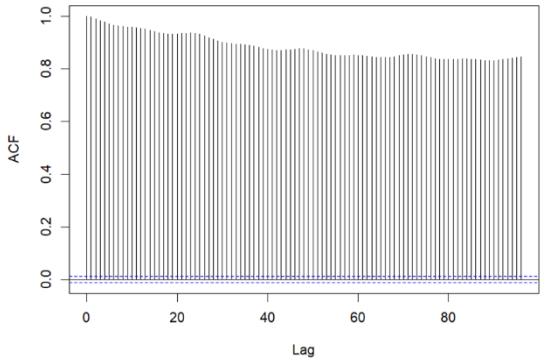

Figura Anexo C.2 Autocorrelação da geração programada PBF derivados de petróleo ou carvão

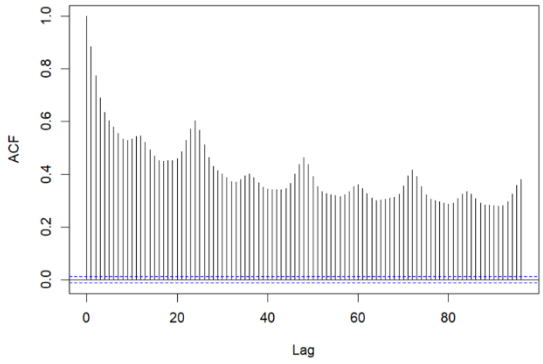

Figura Anexo C.3 Autocorrelação do preço de banda secundária

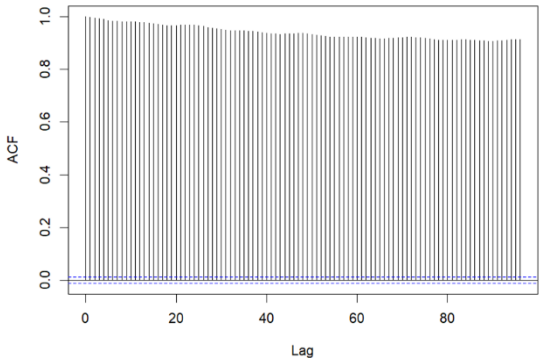

Figura Anexo C.4 Autocorrelação da geração programada PBF cogeração

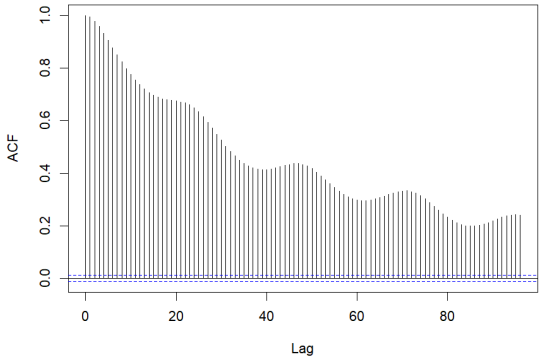

Figura Anexo C.5 Autocorrelação da geração programada PBF eólica

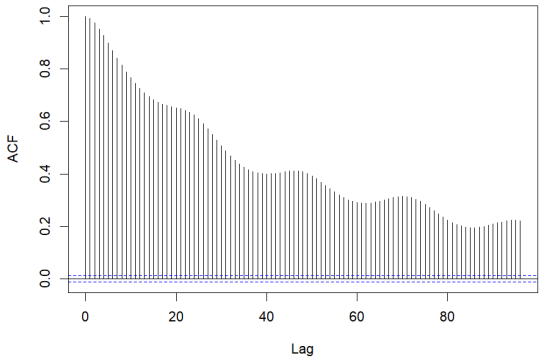

Figura Anexo C.6 Autocorrelação da geração real eólica

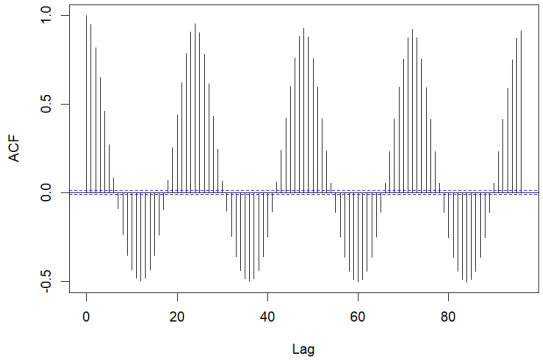

Figura Anexo C.7 Autocorrelação da geração real solar

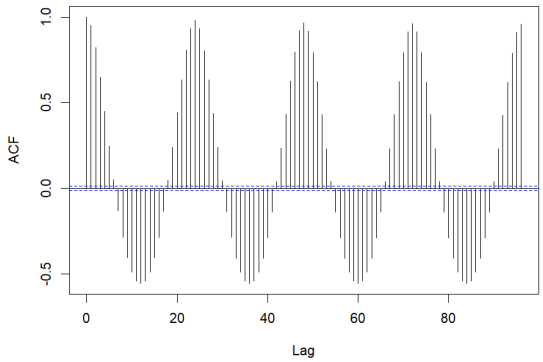

Figura Anexo C.8 Autocorrelação da geração programada PBF fotovoltaica

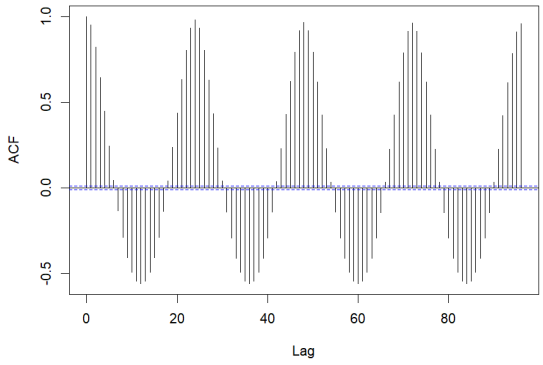

Figura Anexo C.9 Autocorrelação da geração programada PVP fotovoltaica

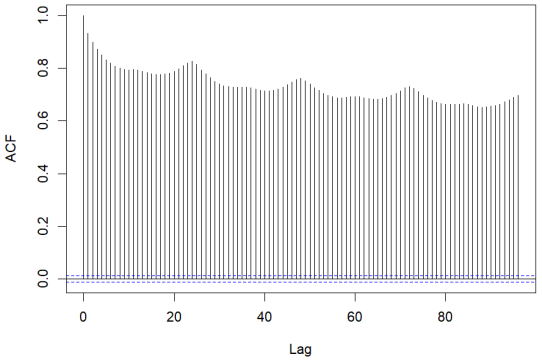

Figura Anexo C.10 Autocorrelação dos desvios de preço de pagamento aumentar

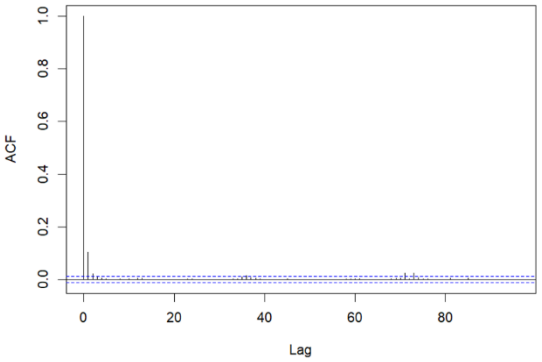

Figura Anexo C.11 Autocorrelação do preço de desvio medido entre o preço marginal diário a baixar

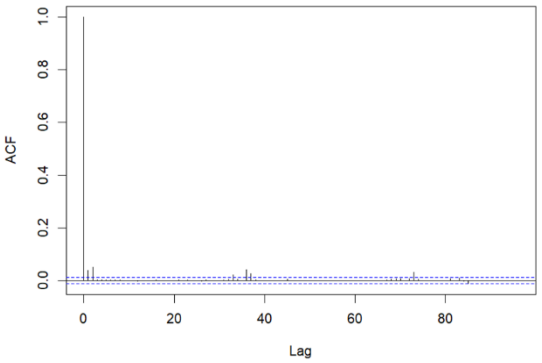

Figura Anexo C.12 Autocorrelação do preço de desvio medido entre o preço marginal diário a aumentar

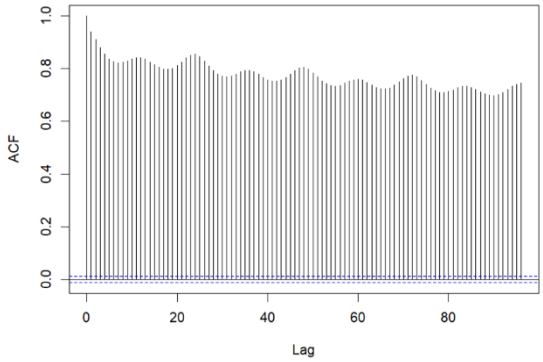

Figura Anexo C.13 Autocorrelação dos desvios de preço de pagamento baixo

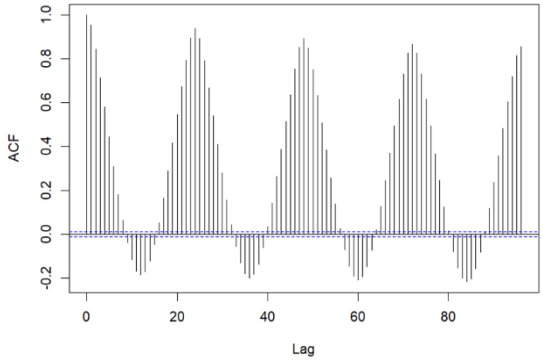

Figura Anexo C.14 Autocorrelação da geração programada PBF solar térmico

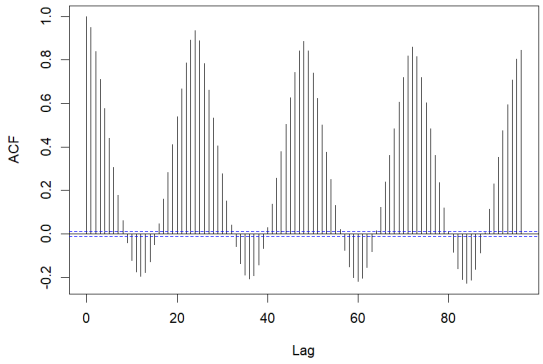

Figura Anexo C.15 Autocorrelação da geração programada PVP solar térmica

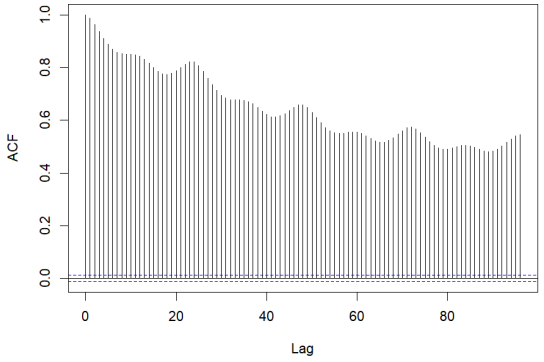

Figura Anexo C.16 Autocorrelação da geração de ciclo combinado

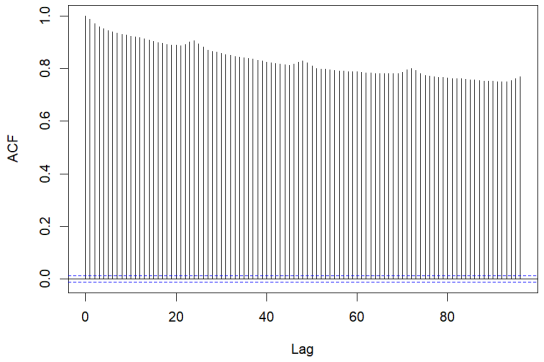

Figura Anexo C.17 Autocorrelação da geração PBF carvão

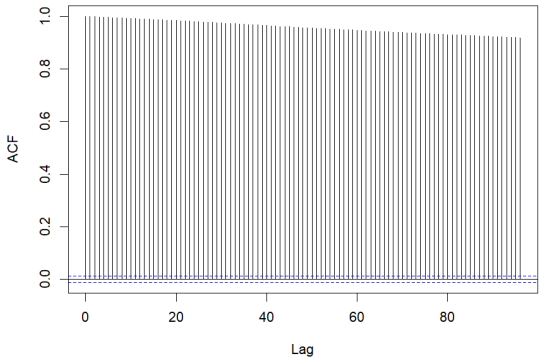

Figura Anexo C.18 Autocorrelação do preço do gás Day-Ahead