

Análise e previsão de quebras energéticas: E-Redes

Realizado no âmbito da Unidade Curricular de Projeto Final
Aplicado a Ciência de Dados da Licenciatura em Ciência de
Dados no ISCTE-IUL

Allan Kardec Rodrigues, 103380, CDC1
aksrs@iscte-iul.pt

André Plancha, 105289, CDC2
Andre_Plancha@iscte-iul.pt

Diogo Freitas, 104841, CDC1
daafs@iscte-iul.pt

João Francisco Botas, 104782, CDC1
Joao_Botas@iscte-iul.pt

Marco Esperança, 110451, CDC1
mdeao@iscte-iul.pt

Índice

1. Introdução	3 / 30
2. Business Understanding	4 / 30
3. Data Understanding	5 / 30
3.1. Indicadores gerais de continuidade de serviço	5 / 30
3.2. Zonas de Qualidade de Serviço	5 / 30
3.3. Mini-Revisão de literatura	6 / 30
3.4. Problema → Limitações iniciais	6 / 30
3.4.1. A Descoberta dos Eventos Excepcionais	6 / 30
3.4.2. Recolha dos dados	6 / 30
3.5. Compreensão dos dados principais	7 / 30
3.6. Dados de Extras/Terceiros	7 / 30
3.7. Visualização do número das quebras energéticas	7 / 30
4. Data Preparation	8 / 30
4.1. Tratamento dos dados principais	8 / 30
4.1.1. Evolução de tempo dos relatórios	8 / 30
4.1.2. Datas inseridos de forma incorreta	9 / 30
4.1.3. Códigos do relatório inconsistentes	9 / 30
4.1.4. Tratamento do número de quebras energéticas	9 / 30
4.2. Criação de variáveis	10 / 30
4.2.1. Percentagem das zonas de qualidade de serviço	10 / 30
4.2.2. Trigonometria Data	11 / 30
4.2.3. Classe das causas	12 / 30
4.2.4. Percentagem de ruralidade de cada Concelho	13 / 30
4.3. Datasets Finais	13 / 30
5. Modelling	14 / 30
5.1. Não Supervisionada	14 / 30
5.1.1. Pré-Clustering	14 / 30
5.1.1.1. Visualização e interpretação dos clusters dos concelhos	15 / 30
5.1.2. Clustering - Quebras energéticas	18 / 30
5.1.2.1. Visualização e interpretação dos clusters das quebras energéticas	19 / 30
5.2. Supervisionada	22 / 30
5.2.1. Time Series	22 / 30
5.2.1.1. Previsão do valor absoluto	22 / 30
5.2.1.2. Previsão binária de ocorrência de eventos	23 / 30
5.2.2. Modelo supervisionada de classificação	24 / 30
5.2.2.1. Oversampling	25 / 30
5.2.2.2. Undersampling	25 / 30
6. Tentativa de 2º iteração do CRISP-DM	26 / 30
6.1. Distância entre eventos previstos	26 / 30
6.2. Correspondência de eventos	27 / 30
6.3. Métricas de Avaliação	27 / 30

6.4. Modelação da segunda interação	27 / 30
6.4.1. AR(1,1)	28 / 30
6.4.2. ARIMA(2,1,0)	29 / 30
7. Conclusão e deployment	30 / 30
8. Anexos	31 / 30
Anexo A - IPMA	31 / 30
Anexo B - GPP - Gabinete de Planeamento, Políticas e Administração Geral	32 / 30
Anexo C - Dados Demográficos - Pordata	32 / 30
Anexo D - Desequilíbrio das Classes	33 / 30
Anexo E - Taxa de ruralidade	34 / 30
Anexo F - Eventos excepcionais de grande impacto	35 / 30
Anexo G - Exemplo de cluster “errado”	36 / 30
Anexo H - Exemplo de previsão “errada”	37 / 30
Anexo I - Gradient Boosting com overfitting	38 / 30

1. Introdução

O presente trabalho foi realizado no âmbito da unidade curricular de Projeto final aplicado a Ciência de Dados, onde era pedido que escolhêssemos um tema e uma entidade para trabalhar com, ao mesmo tempo que éramos orientados por um docente da UC. Dentro das várias opções trabalhámos os dados da E-Redes, fornecidos pela OpenDataSoft¹, com especial enfoque para as falhas na rede elétrica, onde fomos orientados pelo professor Luís Nunes.

O consumo de energia é um tema cada vez mais relevante e desafiador, impulsionado pela evolução da ciência, tecnologia e pela integração da energia na vida humana, surgindo, assim, novos problemas como a sustentabilidade, a segurança energética e o consumo consciente de energia.

Neste projeto, propomos utilizar o CRISP-DM (*Cross-Industry Standard Process for Data Mining*) para realizar uma análise aprofundada dos dados disponíveis no portal de dados abertos da E-Redes. O objetivo principal é prever o número de falhas energéticas, identificar os seus locais específicos, otimizar o processo de reconhecimento de incidentes e oferecer *insights* para a implementação de medidas de segurança proativas, visando a prevenção de falhas no futuro.

O CRISP-DM é um modelo de processo de *data mining* que descreve abordagens usualmente utilizadas por especialistas para “atacar” e resolver problemas. Ele é composto por seis fases:

1. **Business Understanding:** Definir os objetivos do projeto e entender as necessidades da E-Redes;
2. **Data Understanding:** Explorar e analisar os dados disponíveis no portal de dados abertos da E-Redes e de outras entidades;
3. **Data Preparation:** Limpar, transformar e preparar os dados para a análise;
4. **Modelling:** Desenvolver modelos preditivos para prever o número e a localização de falhas energéticas;
5. **Evaluation:** Avaliar o desempenho dos modelos preditivos e selecionar o melhor modelo;
6. **Deployment:** Implementar o modelo selecionado na E-Redes e monitorizar seu desempenho.

Acreditamos que este projeto pode contribuir para uma gestão mais eficiente e resiliente da rede de distribuição elétrica da E-Redes, de forma a garantir um fornecimento de energia mais seguro e confiável para todos os consumidores.

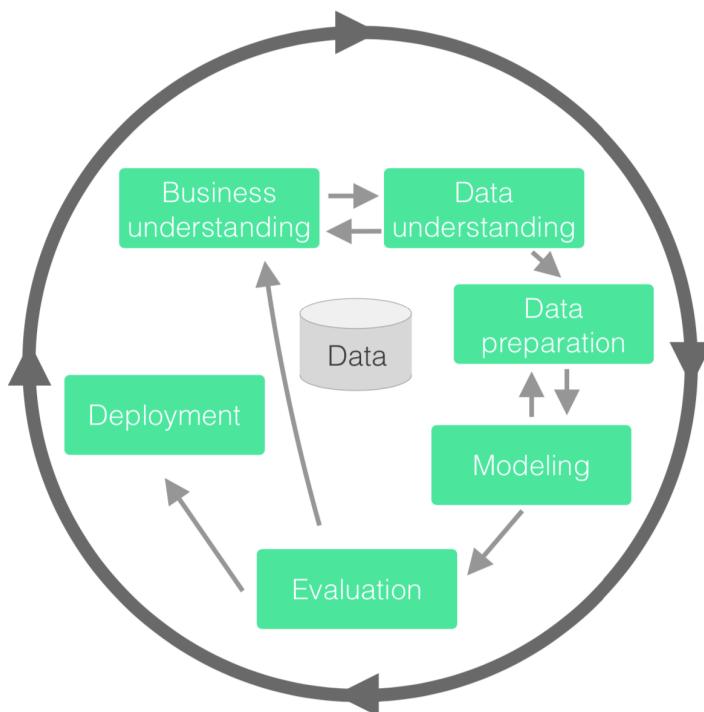


Figura 1: CRISP-DM

¹<https://e-redes.opendatasoft.com/explore/?exclude.keyword=internal&sort=modified>

2. Business Understanding

O consumo de energia tem vindo a tornar-se progressivamente um tema cada vez mais relevante e desafiador. Esta crescente relevância é impulsionada pela constante evolução da ciência e das tecnologias, que integram a energia na vida humana. Consequentemente, surgem novos problemas a serem combatidos, tais como a **sustentabilidade**, a **segurança energética** e o **consumo consciente de energia**.

A E-Redes é uma distribuidora de energia em Portugal e líder no setor, com mais de 40 anos de história. Em 2023, contava com 6.4 milhões de clientes e uma rede de distribuição de 234.669 km, que permite satisfazer as necessidades de todos os clientes de forma prática e eficaz. A E-Redes adota um conjunto de máximas que asseguram um serviço de excelência de norte a sul do país, sendo fundamental a sua análise e estudo²:

- Garantir o fornecimento de eletricidade a todos os consumidores com qualidade, segurança e eficiência, através de linhas aéreas e cabos subterrâneos de baixa, média e alta tensão;
- Promover o desenvolvimento da rede de distribuição que suporte a transição energética;
- Assegurar, de forma isenta, a disponibilidade de serviços aos agentes de mercado.

Com este propósito em mente, a E-Redes dedica-se há mais de 40 anos a evoluir e a superar-se no setor energético. A sua missão é garantir que este fluxo vital, invisível, mas crucial, funcione de forma impecável, levando energia a todas as casas dos consumidores. Dia após dia, a empresa trabalha incansavelmente para que a luz nunca se apague. **Mas, e se falhar?**

Uma falha no fornecimento de energia teria consequências imediatas e devastadoras para a sociedade portuguesa. Serviços essenciais, como hospitais, sistemas de comunicação e transportes públicos, seriam paralisados, colocando em risco vidas e causando enormes transtornos na vida das pessoas. As atividades económicas e sociais também sofreriam um impacto imediato, com fábricas a parar, lojas a fechar e escolas a cancelar aulas. Para combater este problema, a E-Redes conta com trabalhadores preparados para lidar com qualquer tipo de acidente que possa ocorrer. No entanto, esta solução não é perfeita, pois a resolução dos problemas pode levar algumas horas, prejudicando todos os tipos de consumidores. **Mas e se fosse possível prever e identificar padrões nestes acidentes?**

Propomo-nos a realizar uma análise aprofundada dos dados disponíveis no portal de dados abertos da E-Redes, acessível em e-redes.opendatasoft.com. O nosso objetivo principal consiste na previsão não apenas do número de falhas energéticas, mas também na identificação dos seus locais específicos, visando contribuir para a implementação de inspeções mais controladas e reguladas. Além disso, almejamos identificar padrões recorrentes nas falhas energéticas, otimizando o processo de reconhecimento desses incidentes. Por exemplo, ao analisar se determinadas falhas estão associadas a condições climáticas específicas numa determinada localidade, pretendemos oferecer *insights* que possam informar a implementação de medidas de segurança proativas, visando a prevenção de tais falhas no futuro. Este enfoque rigoroso e analítico visa fortalecer a eficácia das operações da E-Redes, contribuindo para uma gestão mais eficiente e resiliente da rede de distribuição elétrica.

Assim, o nosso foco incide, sobretudo, na seguinte questão, que será respondida no desenrolar do estudo:
“Estudar as causas das quebras energéticas, prever e classificar padrões de tais eventos.”



²<https://www.e-redes.pt/pt-pt/noticias/2021/01/29/edp-distribuicao-agora-e-e-redes>

3. Data Understanding

Uma quebra energética, segundo a E-Redes, ocorre quando a tensão de alimentação no ponto de entrega é inferior a 5% da tensão declarada. Se durar mais de 3 minutos, é considerada longa; se for inferior a um segundo é momentânea; as restantes são breves. Estas podem ser previstas (programadas) ou acidentais (inesperadas). Somado a isto, é importante ter em conta que a rede de distribuição pode ser:

- AT → Alta Tensão / MT → Média Tensão / BT → Baixa Tensão

3.1. Indicadores gerais de continuidade de serviço

Estas quebras afetam a continuidade do serviço e, para avaliar esta continuidade, a E-Redes utiliza os seguintes indicadores gerais relativos aos pontos de entrega às instalações de consumo:

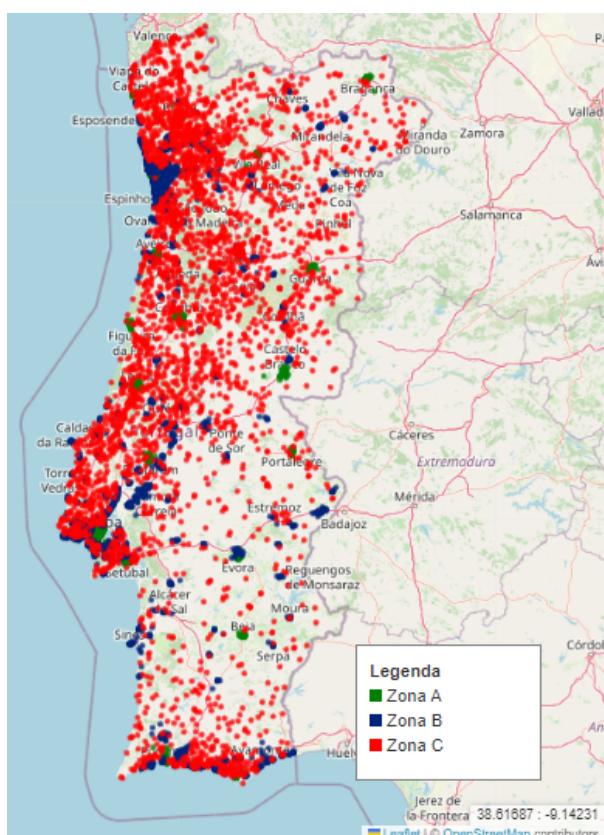
Tabela 1: Principais indicadores gerais; período de cálculo: 1 ano (indicadores anuais)

Indicadores Gerais	Quantidade	Duração	Rede de Distribuição
SAIFI	F - Frequência	-	(AT, MT, BT)
SAIDI	-	D - Duração	(AT, MT, BT)
MAIFI	F - Frequência	-	(AT, MT)
TIEPI	-	-	(MT)
END	-	-	(MT)

Relativamente aos indicadores Gerais, temos os seguintes:

- Frequência Média de Interrupções Longas do Sistema (**SAIFI**)
- Duração Média das Interrupções Longas do Sistema (**SAIDI**)
- Frequência Média das Interrupções Breve do Sistema (**MAIFI**)
- Tempo de Interrupção Equivalente da Potência Instalada (**TIEPI**)
- Energia Não Distribuída (**END**)

3.2. Zonas de Qualidade de Serviço



Zona A:

- Capitais de distrito e lugares com mais de 25000 clientes.
- Infraestrutura elétrica mais desenvolvida e manutenção mais frequente.

Zona B:

- Lugares com um número de clientes entre 2 500 e 25 000.
- Qualidade do serviço média, com uma probabilidade de interrupções maior do que na Zona A.

Zona C:

- Lugares com um número de clientes inferior a 2 500.
- Infraestrutura elétrica menos desenvolvida e a manutenção menos frequente.

Figura 2: Zonas de qualidade de serviço

3.3. Mini-Revisão de literatura

A previsão de quebras energéticas é extensivamente explorada em literatura académica em vários contextos, objetivos e locais, especialmente no contexto de catástrofes naturais. Por exemplo, [Cerrai et al. \(2019\)](#) treinaram vários modelos para prever o número de quebras energéticas durante tempestades no nordeste dos Estados Unidos, utilizando variáveis de meteorologia, dados do bioma do local (o tipo de floresta que era, variáveis de vegetação da área), e variáveis da infraestrutura elétrica do local (número de transformadores, número de interruptores... por cada célula); os resultados depois são enviados para equipas de resposta para que elas se preparem adequadamente.

Outros artigos, como o de [Wang et al. \(2024\)](#), não têm este foco de desastre naturais. Eles usaram dois *perceptron* multi-camadas (redes neurais) para prever o número de utilizadores que perdem energia hora à hora numa área do estado de Michigan, concluindo que a inclusão de variáveis socio-económicas e demográficas sobre o local melhoraram os resultados.

Alguns artigos defletem de métodos de aprendizagem tradicionais, como o de [Zhang et al. \(2019\)](#). Eles preveem o número de quebras energéticas anuais na área de Turim usando um Modelo Cinzento GM(1, 1), um método “clássico para estudar a tendência a partir de séries de dados discretos com amostras limitadas e informação inadequada”, e otimização por enxame de partículas (PSO).

Finalmente, [Mahmoud et al. \(2021\)](#), na sua revisão sistemática em Manutenção preditiva em *Smart Grids*, dividem os métodos de previsão em métodos convencionais - método que olham para a eletricidade e infraestrutura diretamente para a previsão; e métodos de aprendizagem automática, como SVMs, ANNs, *Random Forests*, e RNNs. Eles associam análise de causas e técnicas adequadas, com base nos artigos que reviram.

A nosso conhecimento, não existe nenhum artigo académico que faça previsão de quebras energéticas de nenhum local de Portugal.

3.4. Problema → Limitações iniciais

O primeiro grande desafio surgiu com a obtenção dos dados da E-Redes através do portal [Open Data](#). Os indicadores de continuidade de serviço estavam disponíveis apenas numa base [anual](#), o que inicialmente impossibilitava qualquer análise detalhada das quebras energéticas ao longo do tempo. Essa limitação prejudicou gravemente o início do projeto, pois, sem dados temporais detalhados, tornava-se inviável realizar uma análise precisa e significativa desde o começo. Consequentemente, isso afetou a nossa capacidade de conduzir estudos aprofundados e detalhados sobre a frequência e o impacto das quebras energéticas. Era necessário encontrar uma solução!

3.4.1. A Descoberta dos Eventos Excepcionais

Após uma extensa pesquisa e análise, identificámos os [eventos excepcionais](#) como uma oportunidade única para estudar as quebras energéticas. Os eventos excepcionais são ocorrências raras e significativas de interrupções no fornecimento de energia elétrica, documentadas detalhadamente no portal da E-Redes. Focar nesses eventos permitiu-nos não só compreender melhor as causas e os efeitos das quebras energéticas, mas também explorar as respostas e medidas adotadas pela operadora de rede para mitigar tais incidentes. Apesar de lidarmos com um número limitado de quebras energéticas (apenas aquelas avaliadas quanto à sua excepcionalidade), esta foi a melhor solução encontrada. Além disso, analisamos os [documentos de qualidade de serviço](#) para aprofundar o conhecimento sobre o tema.

3.4.2. Recolha dos dados

Inicialmente, enfrentámos o desafio de coletar os dados [manualmente](#), uma vez que os relatórios disponíveis consistiam, principalmente, em imagens de tabelas, impossibilitando a automação do processo ([Link com um exemplo de relatório](#)). No entanto, a nossa persistência levou-nos à descoberta de um recurso valioso: o [Power BI](#) da ERSE, que continha relatórios PDF bem estruturados e formatados. Esta ferramenta não só simplificou a coleta de dados, mas também proporcionou informações adicionais e contextuais que enriqueceram a nossa análise.

3.5. Compreensão dos dados principais

De forma geral, cada relatório seguia um esquema de tabelas semelhante ao representado na [Tabela 2](#). Essas tabelas incluíam não apenas códigos, datas e clientes afetados, mas também os valores dos indicadores gerais, conforme descrito na [Tabela 1](#).

Tabela 2: Esquema da tabela dos respetivos eventos excepcionais dos relatórios de qualidade e serviço

Código do Relatório	Concelho Origem	Data	Nível De Tensão	Causa	Duração incidente	Nº Cliente afetados	Indicadores (...)	Decisão
...
...

- **Código do Relatório:** identificador único que contém a data e número do evento;
- **Nível de tensão:** qual foi o tipo de rede de distribuição afetada (AT, MT, BT);
- **Causa:** o que originou o evento (ex: Veículos, Aves, Furtos, Escavações, Descarga Atmosférica, entre outras);
- **Duração incidente:** quanto tempo, em minutos, durou o incidente;
- **Nº clientes afetados:** qual o número de clientes que foram afetados pelo evento;
- **Decisão:** evento considerado excepcional → Sim / evento não considerado excepcional → Não

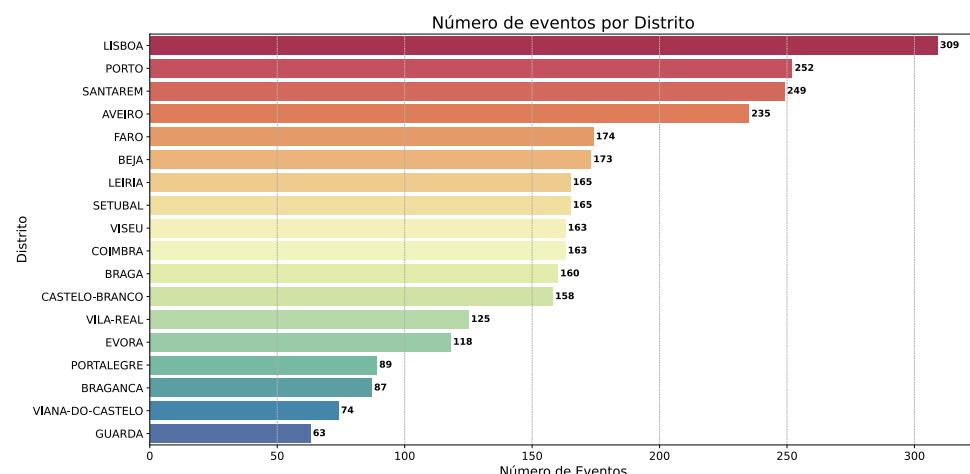
No entanto, foi necessário realizar uma limpeza desses dados, como será discutido mais adiante ([Secção 4.1](#)), pois o esquema das tabelas mudou ao longo dos anos e ainda existem alguns erros aparentes.

3.6. Dados de Extras/Terceiros

Para enriquecer as conclusões do trabalho, foi necessário recolher dados de outras fontes, pois fatores relativos ao território e ao clima têm uma grande influência nas quebras energéticas. Abaixo, estão listados os respetivos dados que auxiliaram as conclusões ao longo do trabalho, com uma breve explicação dos mesmos nos anexos:

- Dados Meteorológicos → IPMA: [Anexo A](#)
- Dados Territoriais → GPP: [Anexo B](#)
- Dados Socioeconómicos → Pordata: [Anexo C](#)
- Localização das subestações → E-REDES: [RND](#) & [PTD](#)
- Shapefile dos concelhos → [dados.gov](#): [shp](#)
- Zonas de Qualidade de Serviços → [ERSE Power BI](#) (Secção “Postos de Transformação”)

3.7. Visualização do número das quebras energéticas



À esquerda, podemos visualizar o número de interrupções energéticas que ocorreram em cada distrito. Lisboa destaca-se em primeiro lugar, de forma significativa, enquanto a Guarda é o distrito com o menor número de interrupções energéticas.

Figura 3: Barplot com o número de quebras energéticas por distrito

O relatório dos eventos excepcionais foca-se nos concelhos como ponto geográfico de análise. Assim, optou-se por usar os concelhos na representação visual para facilitar a compreensão dos dados e respeitar as informações fornecidas. Embora esta abordagem assuma que os concelhos são independentes entre si, consideramos que é a melhor forma de visualizar as informações. Entre todos os concelhos, Lisboa destaca-se de forma significativa, apresentando o maior número de quebras energéticas. Este concelho registou um número de eventos excepcionalmente elevado, o que o coloca numa posição de destaque, comparado com os restantes concelhos.

Por outro lado, alguns concelhos encontram-se representados a preto no mapa. Esta escolha visual deve-se à inexistência de quebras energéticas nestas localidades. Além disso, esta decisão também foi influenciada por uma deliberação do grupo, que será explicada mais à frente no relatório ([Secção 4.1.4](#)). Desta forma, procuramos proporcionar uma visualização clara e intuitiva dos dados relativos às interrupções energéticas em cada concelho.

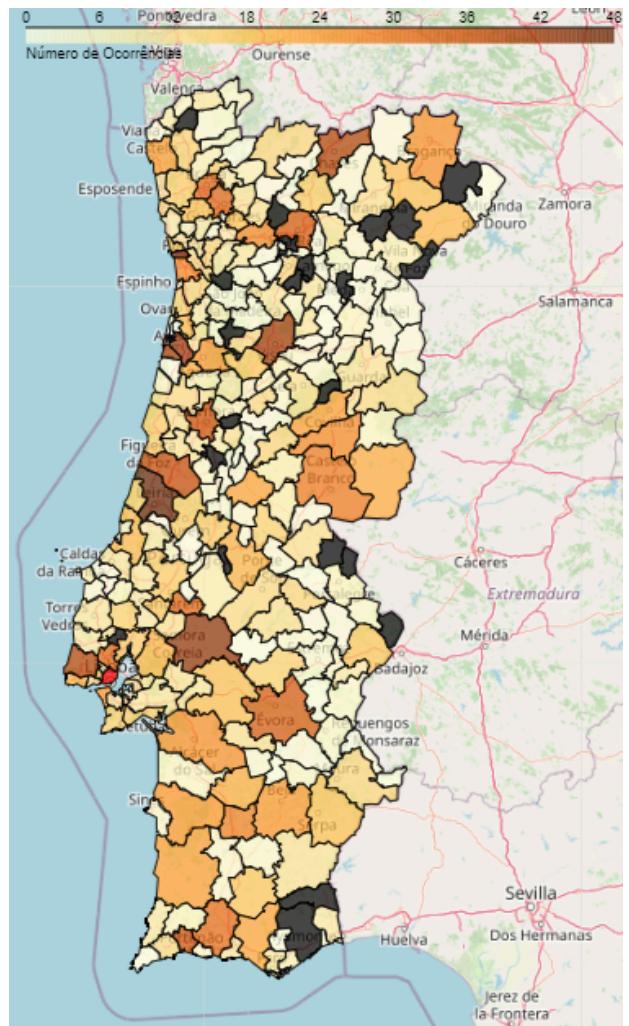


Figura 4: Barplot com o número de quebras energéticas por distrito

4. Data Preparation

4.1. Tratamento dos dados principais

Nesta secção, serão abordados todos os problemas que encontramos nos dados recolhidos nos diferentes relatórios da E-Redes, tanto na recolha, quanto no tratamento dos mesmos.

4.1.1. Evolução de tempo dos relatórios

No processo de recolha, deparamo-nos com um pequeno pormenor: com o passar do tempo, a E-Redes foi evoluindo a forma como guardava os dados relativos às quebras energéticas, resultando no surgimento de novas variáveis e no desaparecimento de outras. Um exemplo simples pode ser encontrado no seguinte [link](#), onde podemos visualizar um relatório onde as tabelas possuem uma coluna intitulada de Código do Relatório, coluna esta que só existe nos relatórios de 2016. Outro exemplo também pode ser o aparecimento da coluna Qualidade de Energia Elétrica em 2017 ([Antes](#) → [Depois](#)).

Para superar este problema, foi necessário realizar vários testes e atualizar o código até conseguirmos ter sucesso. Este processo envolveu a adaptação contínua às mudanças nos relatórios, garantindo que todos os dados relevantes fossem corretamente identificados e integrados na nossa análise.

4.1.2. Datas inseridos de forma incorreta

Outro problema encontrado foram os valores incorretos relativos às datas dos incidentes. Em vez de apresentarem datas comuns como, por exemplo, 2024-06-17, apareciam números que, à primeira vista, pareciam incorretos e sem significado. Contudo, após uma breve pesquisa, descobrimos que esses números escondiam um segredo.

O Excel tem a capacidade de armazenar datas como números sequenciais para que possam ser usados em cálculos. Por padrão, o dia 1 de janeiro de 1900 é representado pelo número serial 1, e o dia 1 de janeiro de 2008 é o número serial 39448 porque corresponde a 39448 dias após 31 de dezembro de 1899.

Além disso, apenas as quebras energéticas ocorridas após 2020 (exceto em 2021) incluíam a hora exata dos incidentes. Infelizmente, foi necessário abdicar da precisão horária devido à falta de informação sobre o horário das restantes quebras energéticas.

DataOld	DataNew
27/10/2017	27/10/2017
39448	01/01/2008
03/12/2019 11:09	03/12/2019

Figura 5: Tabelas com datas incorretas e solução encontrada para corrigir

4.1.3. Códigos do relatório inconsistentes

Código do Relatório (adaptado)	Concelho Origem	Data do incidente	Nº Clientes Afetados
2016_FEV_085	ALMADA	2016-02-24	10
2016_FEV_085	ALMADA	2016-02-27	3
2016_FEV_085	ALMADA	2016-02-28	5

Relativamente ao Código do Relatório, existiam linhas que possuíam o mesmo código, mas apresentavam diferentes características das quebras (Clientes afetados, indicadores gerais, Data do incidente, etc...). Como esses casos não eram numerosos, optou-se por eliminar essas entradas para evitar discordâncias e erros na lógica dos eventos que pudessem comprometer o modelo no futuro.

Figura 6: Tabelas com datas incorretas

4.1.4. Tratamento do número de quebras energéticas

Realizámos um tratamento minucioso dos dados provenientes de entidades externas, corrigindo valores omissos e erros de digitação. Restringimos o conjunto de dados ao período posterior a 2017, pois a partir de 2018 o número de eventos tornou-se mais uniforme, com cerca de 500 eventos por ano. Incluir dados anteriores a 2017 poderia prejudicar a consistência e precisão do modelo devido à disparidade no número de eventos e à maior incidência de erros nesses dados mais antigos.



Figura 7: Jitterplot com as quebras energéticas disponíveis por mês e ano

4.2. Criação de variáveis

4.2.1. Percentagem das zonas de qualidade de serviço

As Zonas de Qualidade de Serviço (ZQSs) recolhidas, ao contrário da nossa expectativa, estão representadas por coordenadas geográficas, em vés de um conjunto de áreas ou freguesias/concelhos associados. Devido à sua definição, nós suspeitamos e assumimos que os pontos sejam apenas como a ERSE decidiu representar as áreas no documento fonte.

De forma a poder representar o concelho em termos de número de clientes no tal, usando as ZQSs, decidimos usar o rácio de zona X que o concelho tem comparados com as outras zonas. Esta solução podia assim cobrir casos onde um concelho tem apenas pontos de zona A (1:0:0) como capitais de distrito; concelhos que tem maioria zona A, mas alguns pontos de zona B (e.g. 9:1:0), potencialmente vindo de algumas localidades dentro do concelho que não passam os 25000 clientes; concelhos apenas com pontos de zona C (0:0:1), como alguns concelhos do interior; entre outros.

Veremos num exemplo prático o concelho de Santarém. Este concelho tem ao todo 100 coordenadas de ZQSSs, sendo 67 delas classificadas como *zona A* e 33 delas *zona C* (ver [Figura 2](#)). Visto que nenhuma é considerada *zona B*, neste concelho, a `percentagem_zonaB` = $\frac{0}{100} = 0$. Da mesma forma, a `percentagem_zonaA` = $\frac{67}{100} = 0.67$ e `percentagem_zonaC` = $\frac{33}{100} = 0.33$.

Observando a [Figura 8](#) verificamos que ter a percentagem é uma mais valia, dado que há um grande número de *zonas B* nos concelhos vizinhos ao de Santarém. Há também uma forte incidência de *zonas A* na cidade de Santarém dentro do concelho e alguma dispersão de *zonas C* por todo o concelho. Desta maneira conseguimos diferenciar bem dois concelhos vizinhos que tenham diferente número de zonas totais e A, B e C separadamente.

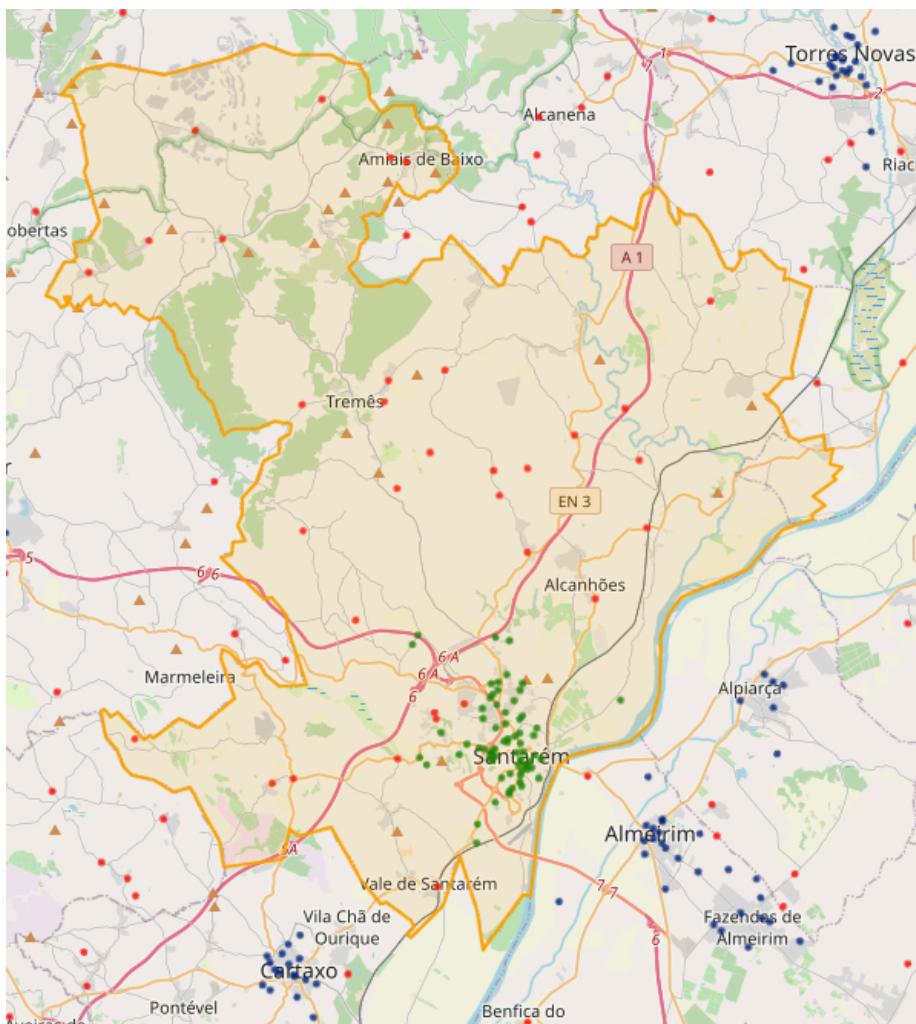


Figura 8: Coordenadas de ZOSs de Santarém (área amarela) e arredores.

A verde estão as zonas A, a azul as zonas B, e a vermelho as zonas C³

³Os pontos em forma de triângulo não representam zonas, apenas são resíduos do mapa, semelhante às estradas.

4.2.2. Trigonometria Data

Na elaboração das nossas transformações e seleção de variáveis, deparamos-nos com o desafio de codificar a altura do ano e a altura do mês. Em outras palavras, poderíamos simplesmente ter registado os dias e meses decorridos, mas isso não informaria ao nosso modelo que a data é cíclica. Assim, acidentes que ocorrem no final de um ano teriam uma grande distância daqueles que ocorrem no início do ano seguinte. Por exemplo:

$$\begin{aligned} \text{Tempo1} : 30/12/2022 &\Rightarrow 364 \text{ dias ocorridos até ao final do ano} \\ \text{Tempo2} : 02/01/2023 &\Rightarrow 2 \text{ dias ocorridos até ao final do ano} \\ 3 &\Rightarrow \text{dias de diferença entre as duas datas} \end{aligned} \quad (1)$$

$$\text{Fórmula} \Rightarrow \text{dias ocorridos até o final do ano} * \frac{2\pi}{\text{Nº de dias do ano}} \quad (2)$$

$|\text{Tempo } 1|$

$$30/12/2022 \Rightarrow 364 * \frac{2\pi}{365} \quad (3)$$

$$\sin\left(364 * \frac{2\pi}{365}\right) \approx -0.01721 \mid \cos\left(364 * \frac{2\pi}{365}\right) \approx 0.99985 \quad (4)$$

$|\text{Tempo } 2|$

$$02/01/2023 \Rightarrow 2 * \frac{2\pi}{365} \quad (5)$$

$$\sin\left(2 * \frac{2\pi}{365}\right) \approx 0.03442 \mid \cos\left(2 * \frac{2\pi}{365}\right) \approx 0.99940 \quad (6)$$

Na imagem ao lado, podemos visualizar as duas datas convertidas acima. É notório que elas agora se encontram próximas uma da outra, alcançando assim o objetivo de representar a ciclicidade das datas. Além disso, podemos utilizar o gráfico para analisar a distribuição dos eventos ao longo das estações do ano. Por exemplo, quando o valor de $\cos(\alpha)$ é próximo de 1 e o de $\sin(\alpha)$ é próximo de 0, estamos no inverno; já os valores opostos representam o verão.

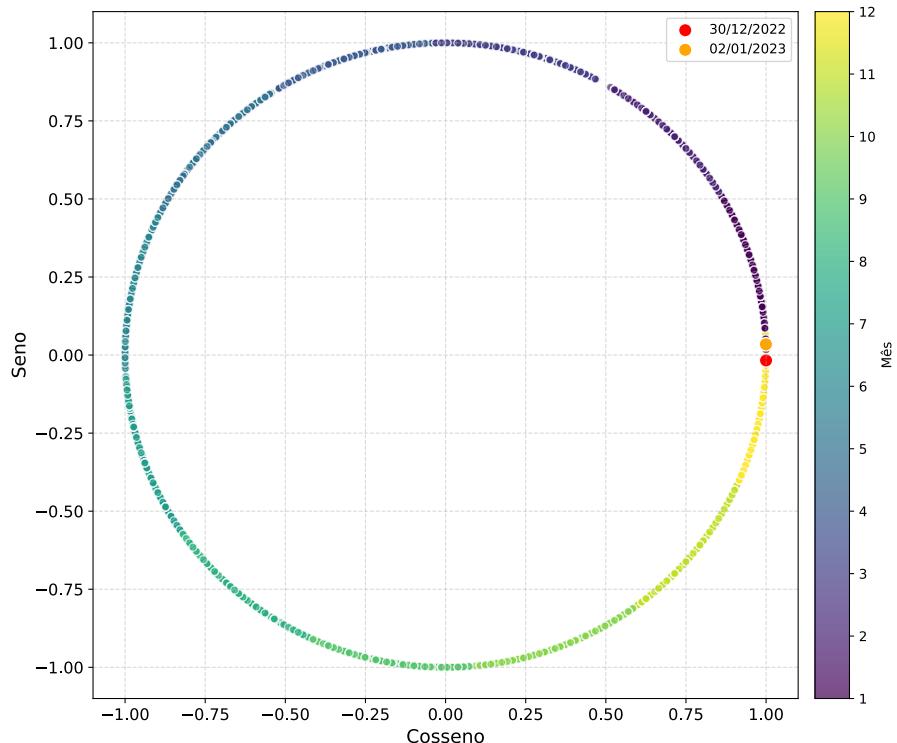


Figura 9: Trigonometria do Ano

Também é importante referir que o mesmo procedimento foi aplicado para os dias de cada um dos meses, ou seja, utilizamos a informação cíclica dos dias dentro de um mês. Desta forma, o modelo pode compreender padrões relacio-

nadas com a altura do mês: o modelo pode, por exemplo detetar que na primeira quinzena de cada mês há uma maior incidência de quebras energéticas, ou aglomerar quebras energéticas das últimas e primeiras semanas do mês.

4.2.3. Classe das causas

É fundamental ter em conta uma característica importantíssima das quebras energéticas: **a sua causa**; ou seja, a ação que originou a quebra energética. Esta característica é essencial para agrupar as diferentes quebras energéticas. No entanto, estas classes apresentam um grande desequilíbrio, com algumas causas semelhantes descritas de maneira diferente. No [Anexo D](#), podemos visualizar um gráfico circular que ilustra a situação inicial, demonstrando um número muito elevado de classes, algumas das quais com um número baixo de ocorrências. Abaixo, podemos ver a solução encontrada para equilibrar as diferentes classes:

Para resolver o desequilíbrio das classes, começamos por identificar causas de incidentes semelhantes, chegando a 4 *classes-chave*:

- **Clima** → Quebras energéticas causadas por condições climatéricas adversas
- **Animais** → Quebras energéticas causadas por Animais
- **Humanos_Intenção** → Quebras energéticas causadas por humanos, mas de forma intencional, como vandalismo e furto
- **Humanos - Acidentes** → Quebras energéticas causadas por humanos, mas de forma não intencional, como escavações e veículos.

Consequentemente, essas classes foram utilizadas como variáveis dummy.

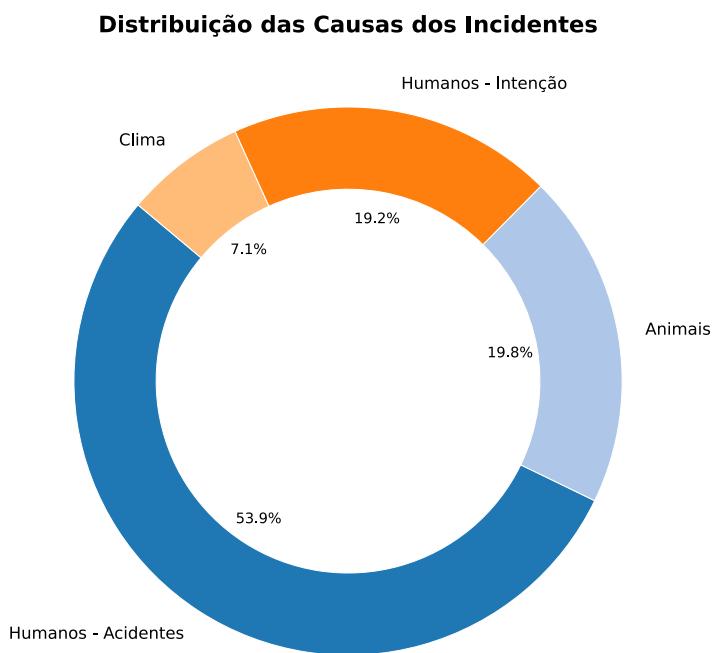


Figura 10: Distribuição das causas dos Incidentes

4.2.4. Percentagem de ruralidade de cada Concelho

Inicialmente, esta variável representava uma taxa de ruralidade, ou seja, em cada concelho eram contadas as freguesias rurais e dividia-se esse número pelo total de freguesias, obtendo-se assim a taxa de ruralidade (mais detalhes disponíveis no [Anexo E](#)). No entanto, durante a apresentação, um dos docentes sugeriu uma alteração. Em vez de se utilizar o número bruto de freguesias, foi proposto usar a área de cada freguesia. Como utilizámos um [shapefile](#), bastou simplesmente realizar:

• `Freguesia['area_m2'] = Freguesia.geometry.area`

De seguida, bastou utilizar a mesma formula, com uma breve alteração:

• **Percentagem de Ruralidade** = $\frac{\sum(\text{área das freguesias rurais})}{\text{área total do concelho}}$

Se analisarmos, a figura assemelha-se bastante à da [Figura 37](#), mas possui algumas diferenças que enriquecem significativamente o modelo, como a noção da dimensão das freguesias rurais e urbanas.

Percentagem de ruralidade de cada concelho

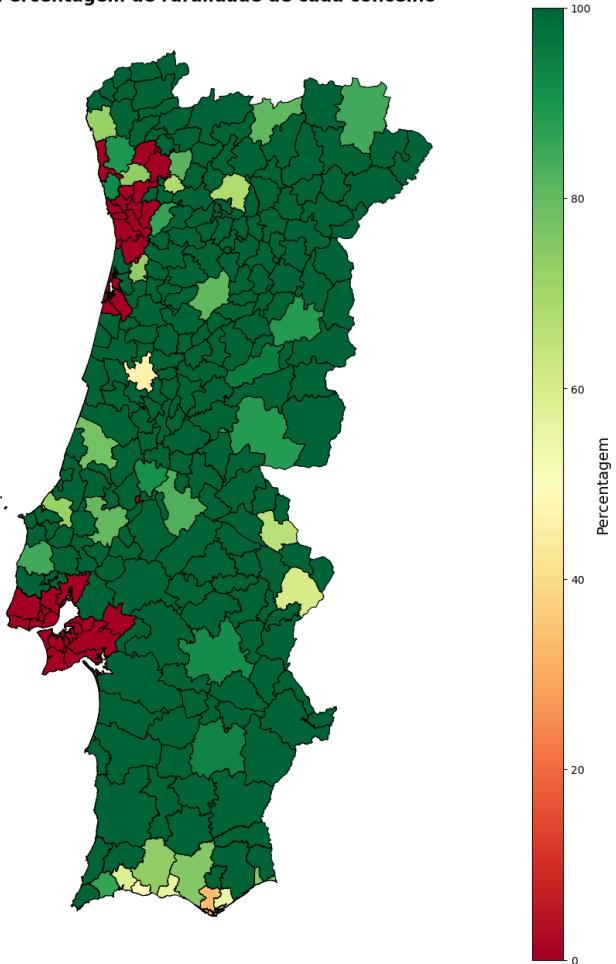


Figura 11: Taxa de ruralidade de cada concelho

4.3. Datasets Finais

Após realizado todo o processo de limpeza e tratamento de dados, ficamos com dois *datasets* finais:

Tabela 3: Conjuntos de dados final

Dataset de eventos	Dataset com todos os dados diários
Cada observação representa um evento excepcional ou não previsto (tratados), e dados associados a tal (como a causa ou a rede de distribuição alvo); tirado das tabelas de decisão da ERSE, entre 2018 e 2023. Ao termos todos os eventos que ocorreram conseguimos identificar padrões nestes.	Cada observação representa um dia por concelho, e dados associados a tais, como dados metereológicos do dia, e dados demográficos do concelho, entre 2018 a 2023. O objetivo principal de ter este <i>dataset</i> com todas as combinações possíveis é de poder prever casos em que estes eventos não existem.

Nós dividimos em dois *datasets* para podermos ter os dados diários dos concelhos e dados sobre os eventos, para poder ter flexibilidade na modelação e não ter perda de informação caso fosse necessário. Esta divisão facilitou também a análise exploratória feita ao longo do projeto.

5. Modelling

Nesta fase do CRISP-DM, avançamos para a criação dos modelos. Serão explorados modelos de aprendizagem não supervisionada ([Secção 5.1](#)) e modelos de aprendizagem supervisionada ([Secção 5.2](#)).

5.1. Não Supervisionada

Na implementação do modelo de não supervisionada, o grupo realizou vários testes e chegou à conclusão que existiam 2 grandes problemas que prejudicavam o modelo, sendo eles:

- **Eventos excepcionais de grande impacto:**

- Devido à ocorrência de eventos excepcionais de grande impacto (EEGIs), qualquer modelo desenvolvido acabava sempre por criar dois grupos distintos: um para os eventos excepcionais de grande impacto e outro para aqueles que não o são, resultando num desequilíbrio que comprometia a interpretação e a precisão do modelo. Por isso, o grupo viu-se obrigado a retirar esses eventos (além de outras limitações, como a abrangência geográfica ou o tipo de tensão afetado). No [Anexo F](#) encontram-se alguns exemplos destes EEGIs.

- **Características dos concelhos mais fortes que as características das quebras energéticas:**

- O modelo considerava que as características dos concelhos tinham uma importância maior do que as características das quebras energéticas (Exemplo: [Anexo G](#)). Para tentar resolver este problema, inicialmente tentamos fazer Análise de Componentes Principais (PCA), mas os resultados eram difíceis de interpretar. Assim, optamos por realizar o que nós chamámos de “pré-clustering”.

5.1.1. Pré-Clustering

Para iniciar o processo de pré-clustering, foram consideradas todas as características relacionadas com os concelhos discutidas ao longo do trabalho, abrangendo desde indicadores socioeconómicos até o número de subestações em cada concelho. Foi utilizado um modelo de *clustering* hierárquico aglomerativo com o método de ligação de [Ward](#), para variância mínima. Abaixo, apresenta-se o dendrograma com os resultados obtidos:

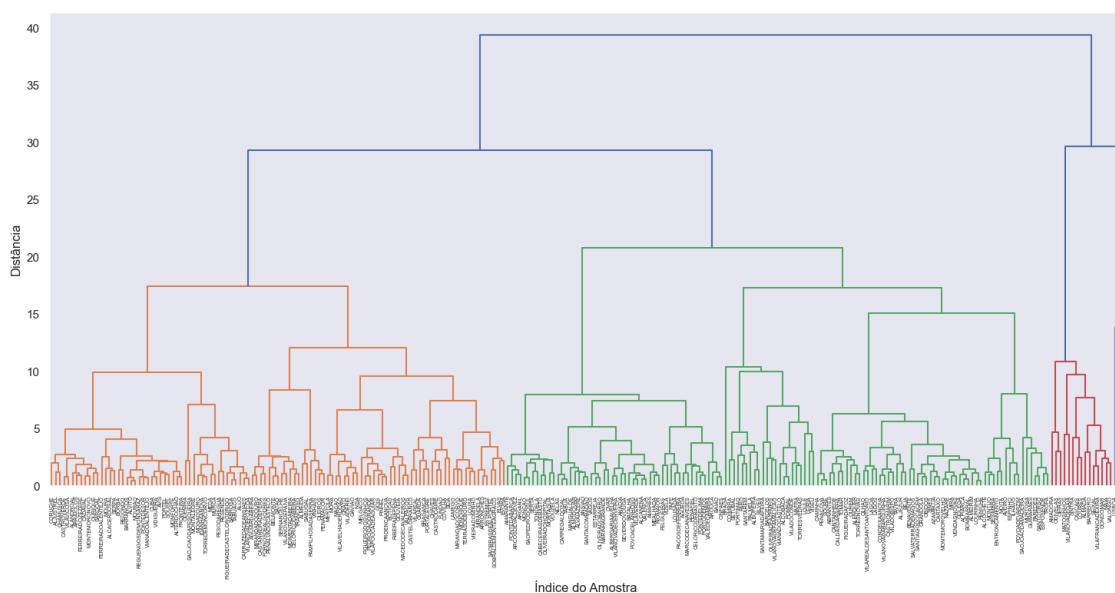


Figura 12: Dendrogramas Hierárquico dos Concelhos

É de notar que os concelhos de Lisboa e Porto ficaram separados em um *cluster* apenas, daí a escolha dos 4 *clusters* para não descharacterizar estes concelhos que são verdadeiros *outliers* em termos energéticos. A Silhouette média para o modelo Agglomerative Clustering foi 0.1924.

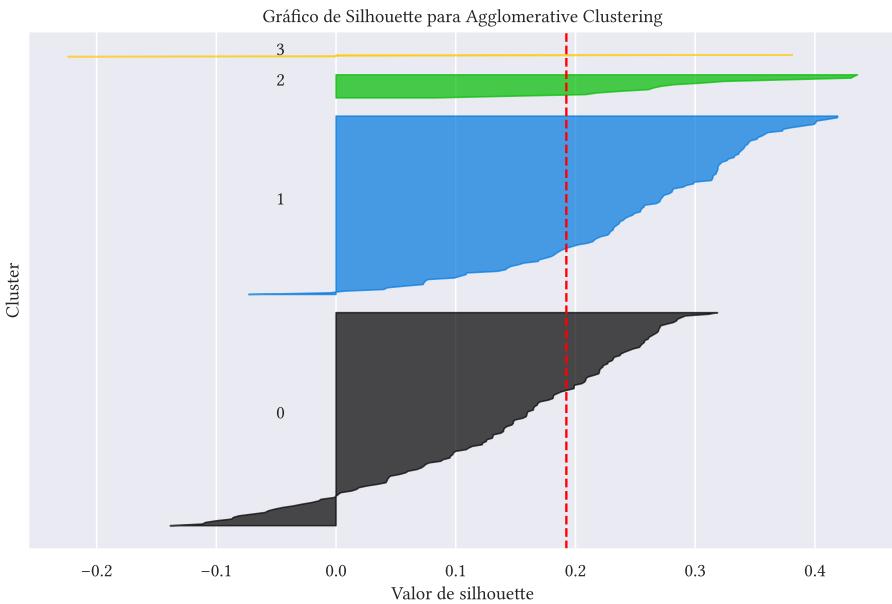


Figura 13: Silhueta dos clusters dos Concelhos

5.1.1.1. Visualização e interpretação dos clusters dos concelhos

CLUSTER 0 → LITORAL E ZONAS INDUSTRIAIS

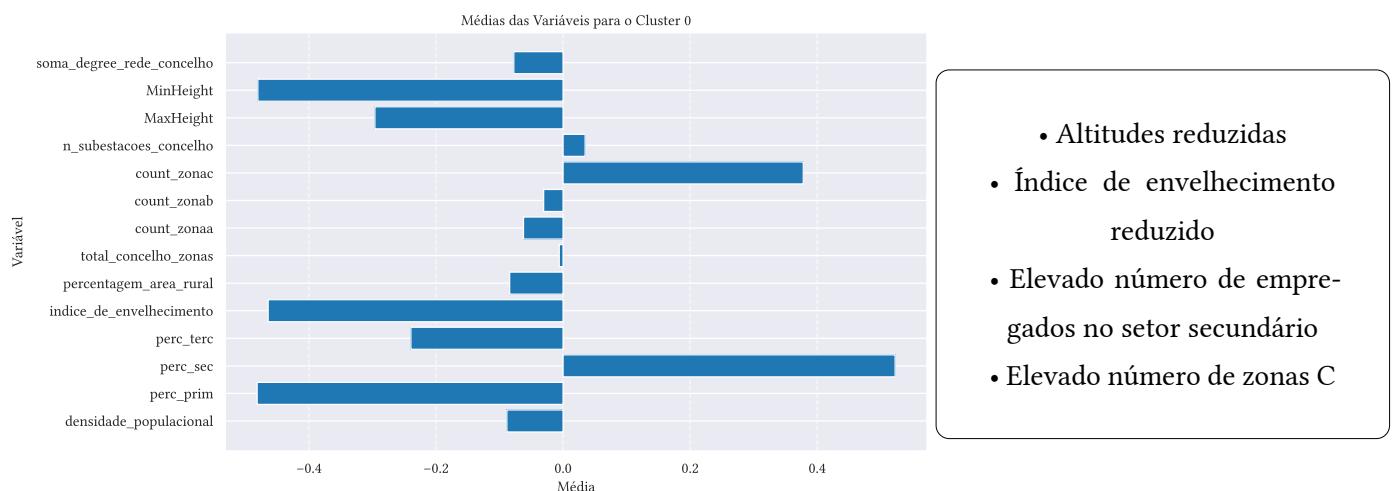


Figura 14: Cluster 0 → Litoral e Zonas Industriais

CLUSTER 1 → INTERIOR

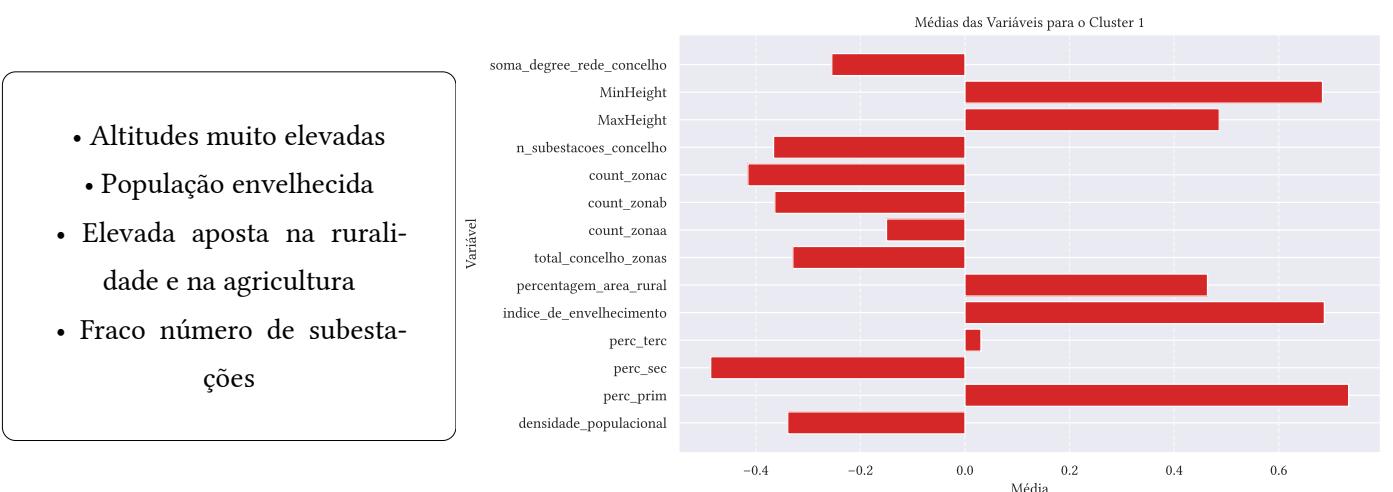
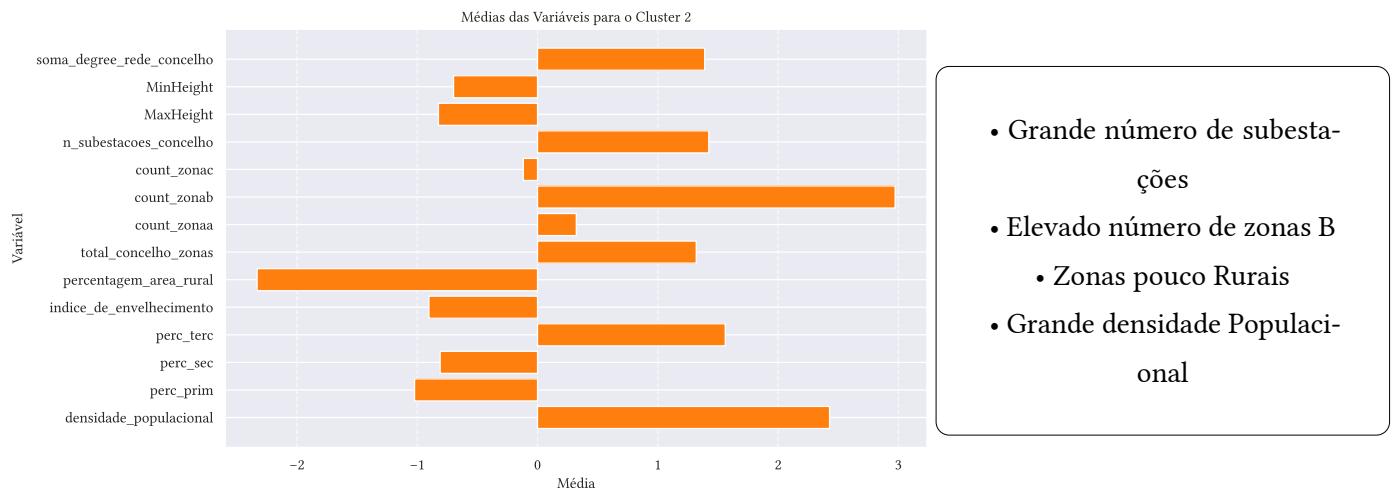


Figura 15: Cluster 1 → Interior

CLUSTER 2 → ARREDORES DAS GRANDES CIDADES



- Grande número de subestações
- Elevado número de zonas B
- Zonas pouco Rurais
- Grande densidade Populacional

Figura 16: Cluster 2 → Arredores das grandes cidades

CLUSTER 3 → GRANDES CIDADES

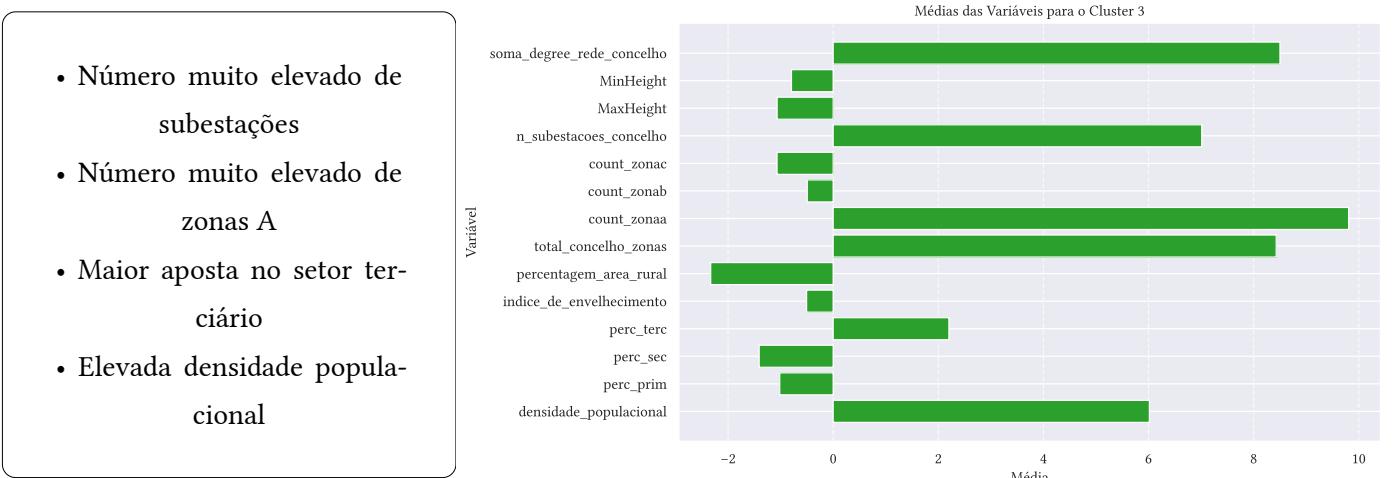


Figura 17: Cluster 3 → Grandes Cidades

Agora que os clusters foram criados e interpretados, o próximo passo é substituir as numerosas variáveis que descrevem as características de cada concelho pelos seus valores de cluster, de forma semelhante a uma tabela dinâmica com variáveis dummy. Cada linha representará um concelho (onde ocorreu a quebra energética) e cada coluna corresponderá a um cluster (com o seu devido nome), onde o valor será 1 se o concelho pertencer ao cluster correspondente e 0 caso contrário.

Além disso, abaixo é possível visualizar um mapa onde a cor de cada concelho representa o seu cluster. Concelhos sem quebra energética são destacados a preto:

Mapa dos Clusters dos concelhos

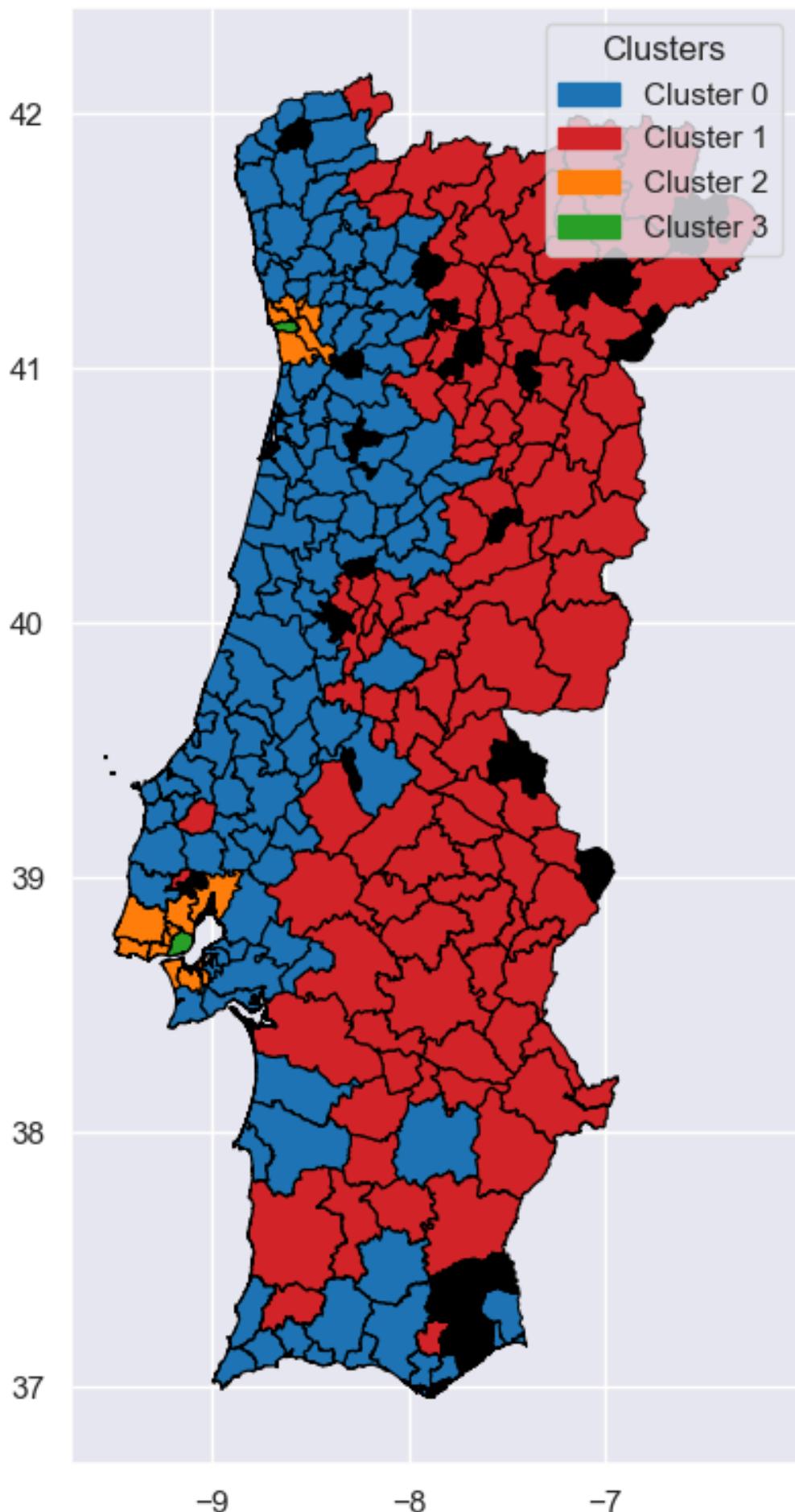


Figura 18: Mapa dos clusters dos concelhos

5.1.2. Clustering - Quebras energéticas

Agora que temos as características das quebras energéticas e os clusters atribuídos a cada concelho no conjunto de dados de **eventos**, podemos avançar com a criação do modelo.

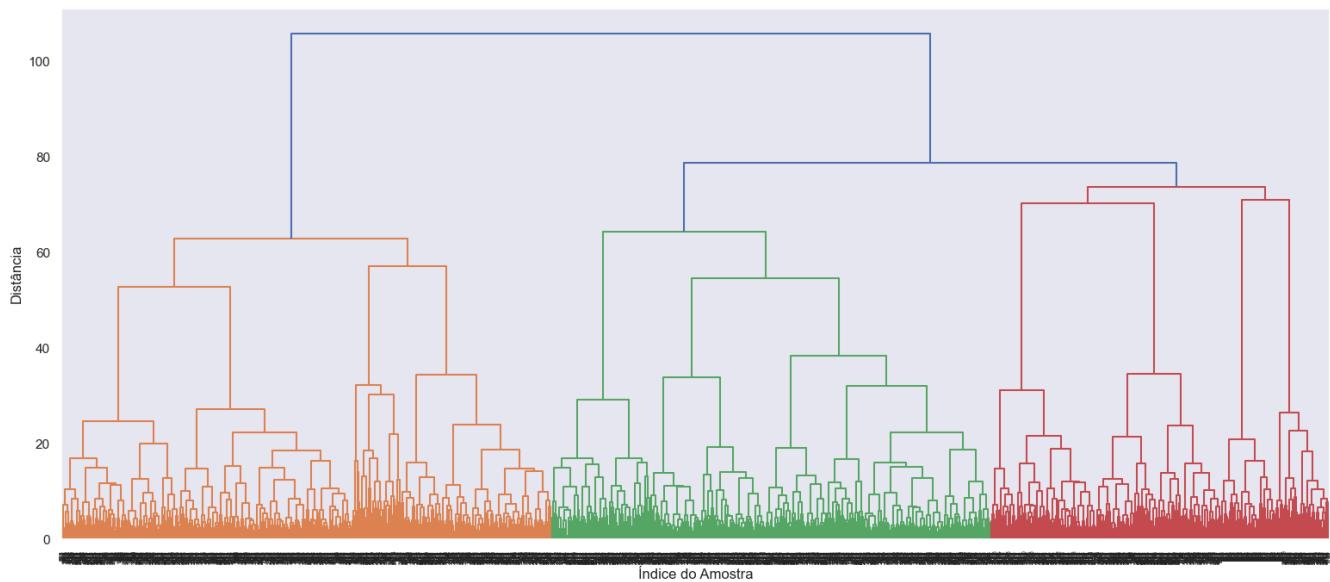


Figura 19: dendrogramas Hierárquico das quebras energéticas

Aqui também foi utilizado um modelo de *clustering* hierárquico aglomerativo com o método de ligação de [Ward](#), para variância mínima. Foi desafiante determinar o número ideal de clusters. Um número muito baixo resultaria em clusters demasiado amplos e pouco distintos, enquanto um número excessivo tornaria os clusters não só difíceis de interpretar, mas também propensos a capturar variações irrelevantes entre os concelhos. Por isso, foi chegado a um consenso definir 3 clusters como a melhor opção. A seguir, é possível visualizar a silhueta correspondente a esses clusters.

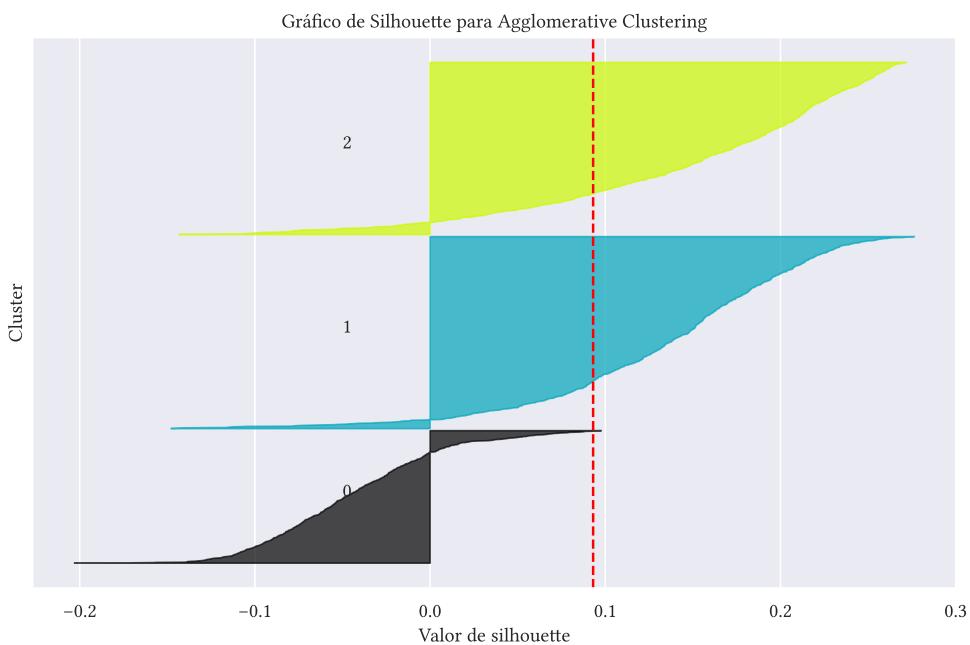


Figura 20: Silhueta dos clusters dos Concelhos

A silhueta média do modelo *Agglomerative Clustering* é 0.0930, o que indica uma estrutura e qualidade relativamente baixa. Este valor sugere que os clusters não estão claramente separados e há uma sobreposição considerável entre eles. Embora os clusters possam fornecer alguma informação útil, a sua definição não é robusta e as diferenças entre os grupos não são muito acentuadas.

5.1.2.1. Visualização e interpretação dos clusters das quebras energéticas

Primeiramente, é importante destacar que serão apresentadas as médias das variáveis de cada um dos clusters, acompanhadas por um mapa de incidência desses clusters em cada concelho. No mapa, quanto mais escuro estiver o concelho, maior será o número de clusters presentes nele.

CLUSTER 0 → QUEBRAS DE BAIXO IMPACTO EM ÁREAS DESENVOLVIDAS POR MÃO HUMANA

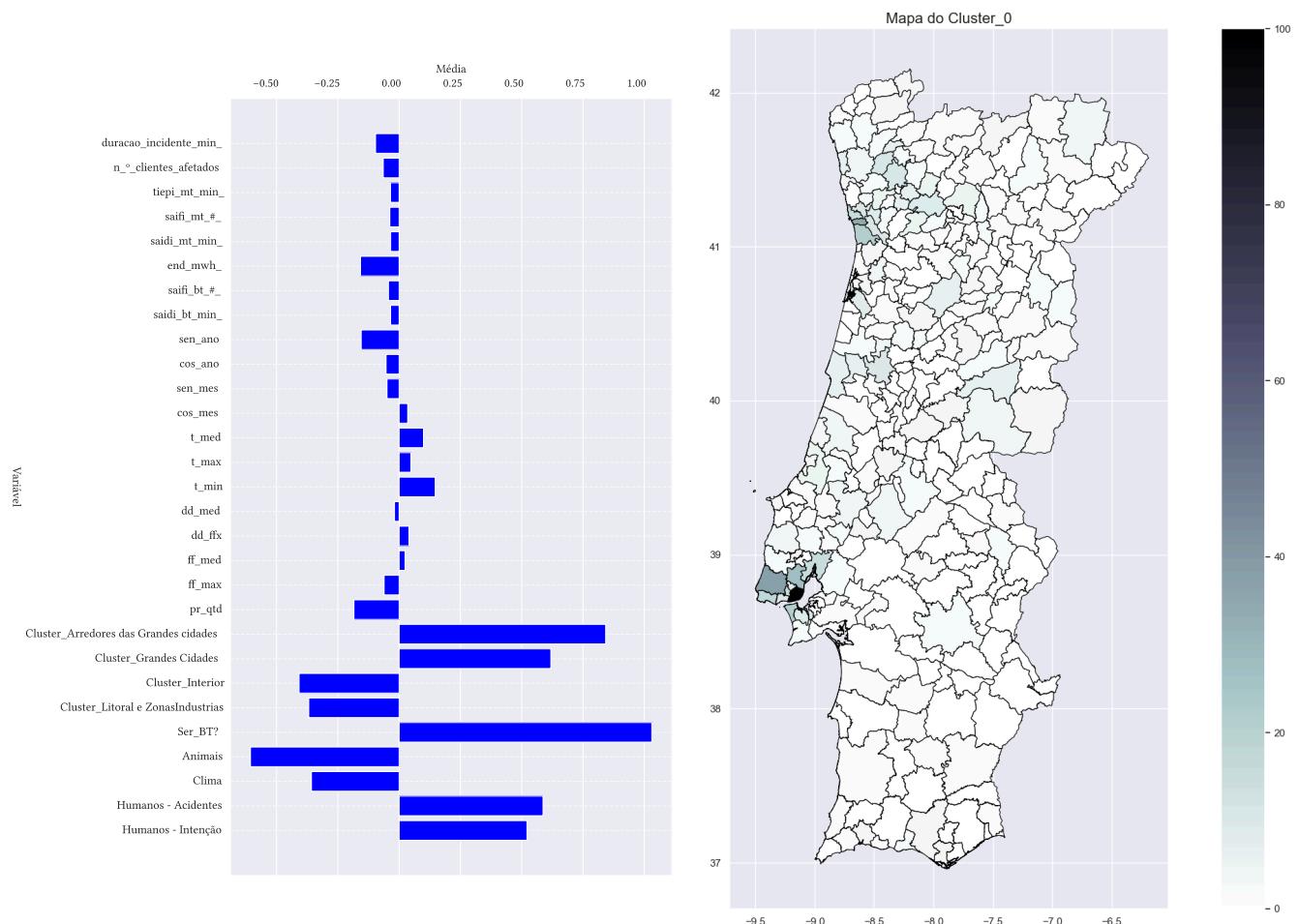


Figura 21: Cluster 0 → Quebras de Baixo Impacto em Áreas Desenvolvidas por mão humana

As quebras energéticas deste cluster ocorrem predominantemente nas grandes cidades e nos seus arredores, onde a densidade populacional e a atividade urbana são mais intensas. Este cluster caracteriza-se por incidentes em sistemas de Baixa Tensão, que são comuns em áreas residenciais e comerciais onde a demanda por energia elétrica é elevada, mas não requer infraestruturas de Alta Tensão.

A maioria destas quebras energéticas são atribuídas a ações humanas, quer sejam resultantes de acidentes ou de atividades intencionais. Acidentes podem incluir falhas causadas por obras de construção, manutenção inadequada ou erros operacionais. As interrupções intencionais, por outro lado, podem envolver vandalismo, roubo de cabos ou outros atos maliciosos que visam a infraestrutura elétrica.

A confluência destes fatores – a ocorrência em zonas urbanas e periurbanas, a predominância de sistemas de Baixa Tensão e a origem humana das falhas – proporciona uma compreensão nítida dos desafios enfrentados na gestão da rede elétrica em áreas de alta densidade populacional.

CLUSTER 1 → ACIDENTES COM ANIMAIS EM PERÍODOS FRIOS NO LITORAL E ZONAS INDUSTRIAS

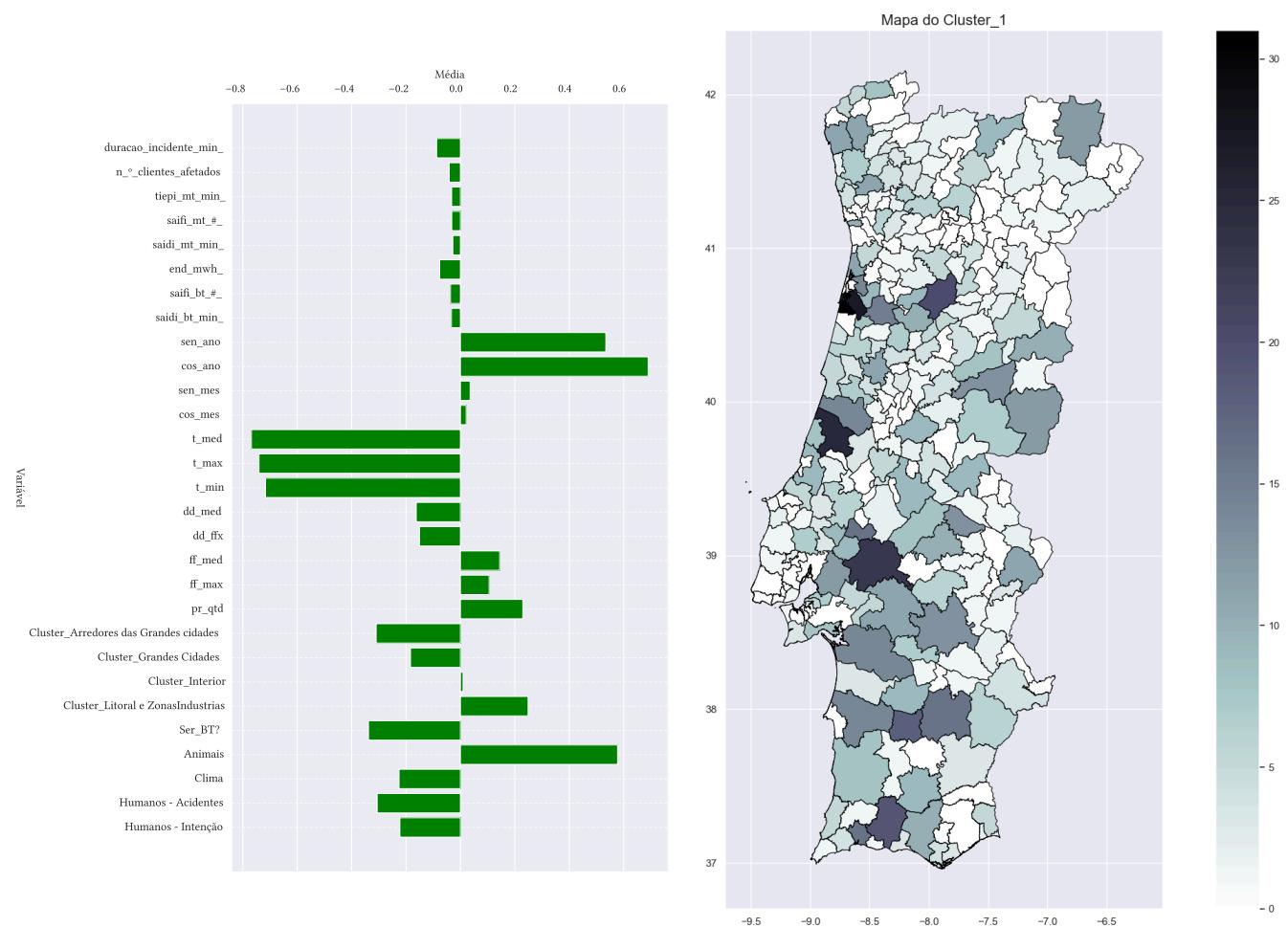


Figura 22: Cluster 1 → Acidentes com Animais em Períodos Frios no Litoral e Zonas Industriais

As quebras energéticas neste cluster são frequentemente atribuídas a interferências causadas por animais, afetando principalmente áreas não classificadas como Baixa Tensão no litoral e zonas industriais de Portugal, em épocas frias e chuvosas.

Este fenómeno é mais prevalente durante o início do ano, especialmente durante o Inverno ($\sin(\alpha)$ e $\cos(\alpha)$ são positivos, logo estamos no primeiro quadrante, e $\cos(\alpha) > \sin(\alpha)$); para uma melhor visualização, poderá rever a [Figura 9](#)), quando as temperaturas frias tendem a aumentar a procura de energia e ainda, devido às condições meteorológicas adversas, o comportamento dos animais tende a ser alterado, muitas vezes causado pela procura de calor ou alimento, que podem levar a incidentes que comprometem a rede elétrica.

Estas interrupções sublinham a necessidade urgente de implementar estratégias de gestão da fauna e de infraestrutura elétrica específicas para estas regiões. Medidas como a proteção de equipamentos elétricos contra danos causados por animais e a adoção de tecnologias que minimizem o impacto dessas interferências são essenciais para assegurar a fiabilidade contínua do fornecimento de energia. Além disso, investimentos em monitorização avançada e resposta rápida a incidentes podem mitigar os efeitos adversos dessas quebras energéticas, garantindo assim um serviço elétrico mais estável e seguro para os residentes e indústrias locais.

CLUSTER 2 → ACIDENTES EM ONDAS DE CALOR NO INTERIOR

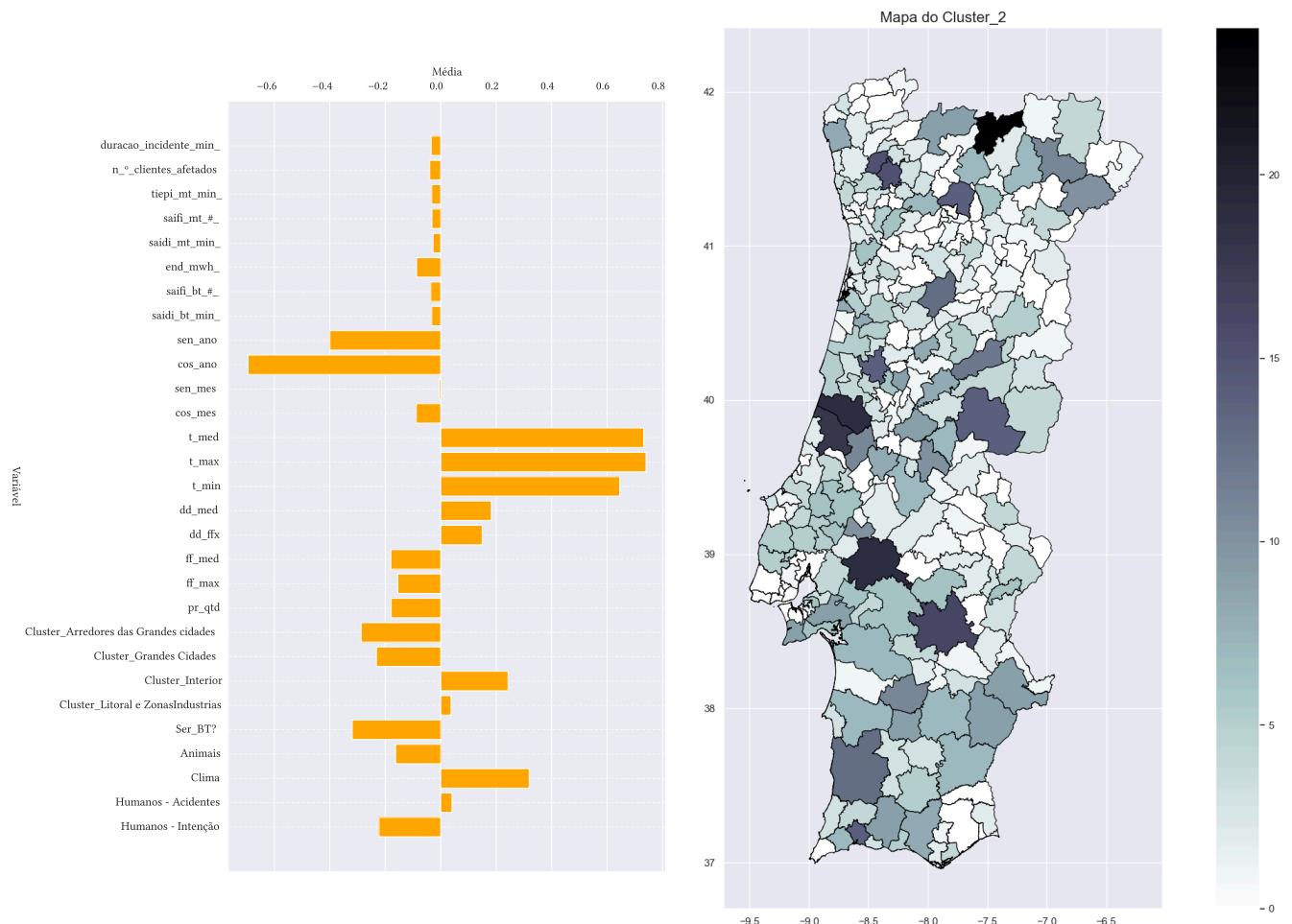


Figura 23: Cluster 2 → Acidentes em Ondas de Calor no Interior

As quebras energéticas neste cluster são frequentemente observadas no interior de Portugal durante o final do Verão (o $\sin(\alpha)$ e o $\cos(\alpha)$ são negativos, logo estamos no terceiro quadrante, e o $\cos(\alpha) < \sin(\alpha)$; para uma melhor visualização, poderá rever a [Figura 9](#)), caracterizado por temperaturas extremamente elevadas e condições meteorológicas com pouco vento e chuva. Nesta época do ano, a demanda por energia atinge picos devido ao uso intensivo de ar condicionado e outros sistemas de refrigeração. As altas temperaturas também podem afetar diretamente a infraestrutura elétrica, aumentando o risco de falhas em equipamentos sensíveis ao calor.

A falta de vento e chuva contribui para um ambiente seco que pode ser propenso a incêndios florestais, representando um desafio adicional para a manutenção da rede elétrica. A combinação desses fatores sublinha a importância de medidas preventivas robustas, como a inspeção regular da infraestrutura e a implementação de tecnologias de monitorização climática avançada.

Para mitigar o impacto das quebras energéticas, é crucial desenvolver estratégias adaptadas às condições climáticas específicas do interior de Portugal. Isso inclui o reforço da resiliência da rede elétrica e a promoção de práticas sustentáveis para garantir um fornecimento contínuo e confiável de energia durante períodos críticos de demanda e condições climáticas desfavoráveis.

5.2. Supervisionada

5.2.1. Time Series

5.2.1.1. Previsão do valor absoluto

Esta foi, de longe, a fase mais complicada do trabalho. Como seria possível prever um evento raro? E de que maneira? Para tal, testámos vários métodos.

Primeiramente, tentamos prever os valores dos indicadores que tinham uma relação direta com as quebras energéticas. No entanto, após realizar um pequeno teste ([Anexo H](#)), rapidamente percebemos que esta abordagem não seria a mais adequada. Isso ocorreu porque os valores dos indicadores gerais nos dias sem quebras energéticas não estavam disponíveis. Além disso, não faria sentido atribuir a esses dias o valor zero, uma vez que isso implicaria, incorretamente, que não houve pequenas quebras energéticas ou interrupções para manutenção, algo que não podíamos assegurar devido ao uso exclusivo de relatórios de eventos excepcionais.

Em resposta à necessidade de prever o número de quebras energéticas diárias, optamos por utilizar o modelo *Long-Short Term Memory* (LSTM). Essa escolha se baseou na robustez e nas características vantajosas do LSTM, que o tornam ideal para lidar com sequências temporais e dados complexos.

A seguir, apresentamos os resultados obtidos com a implementação do modelo LSTM (é importante ter em conta que todos os modelos de LSTM apresentados são referentes ao concelho de Lisboa, para simplificar a visualização):

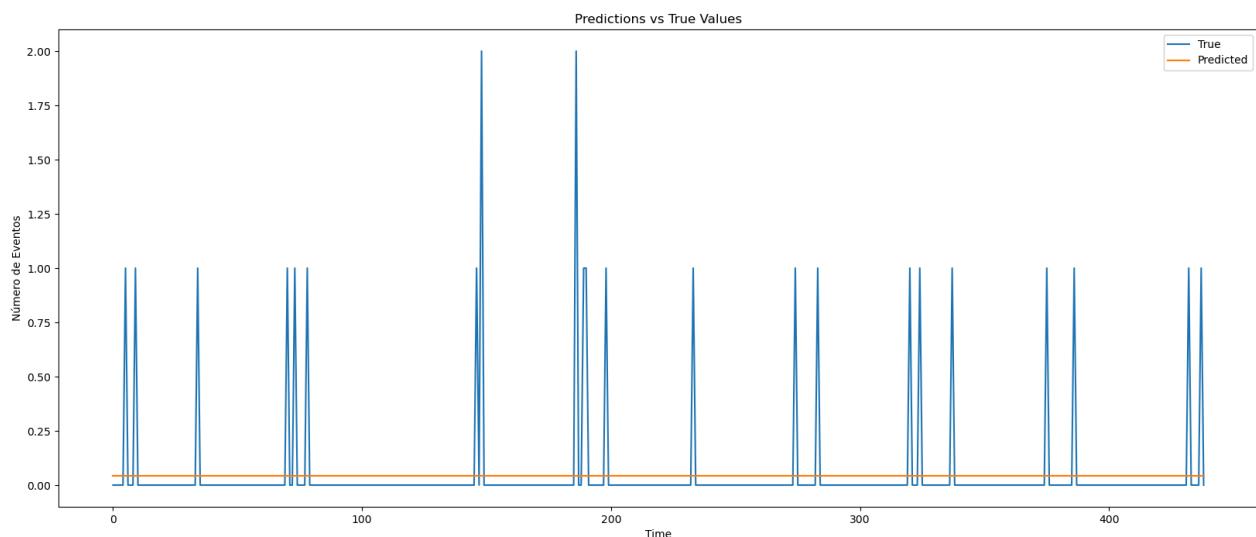


Figura 24: LSTM do Concelho de Lisboa

Se analisarmos corretamente, é fácil concluir que este modelo não é bom, pois prevê quase sempre zero quebras energéticas. Este problema deve-se ao desequilíbrio existente entre as classes, havendo muito mais dias sem quebras energéticas do que com. Abaixo, podemos ver o quanto desequilibradas são as classes:

Tabela 4: Ocorrência de eventos ao longo de 6 anos (2018 - 2023)

Número de Eventos Diários	Ocorrências
0	558072
1	2737
2	77
3	9
4	1

Com base nessa abordagem, para resolver o problema, começamos por atribuir pesos às classes. Isto significa que informamos o modelo para atribuir maior importância aos casos de quebras energéticas do que aos casos em que não há quebras. Os resultados obtidos foram os seguintes:

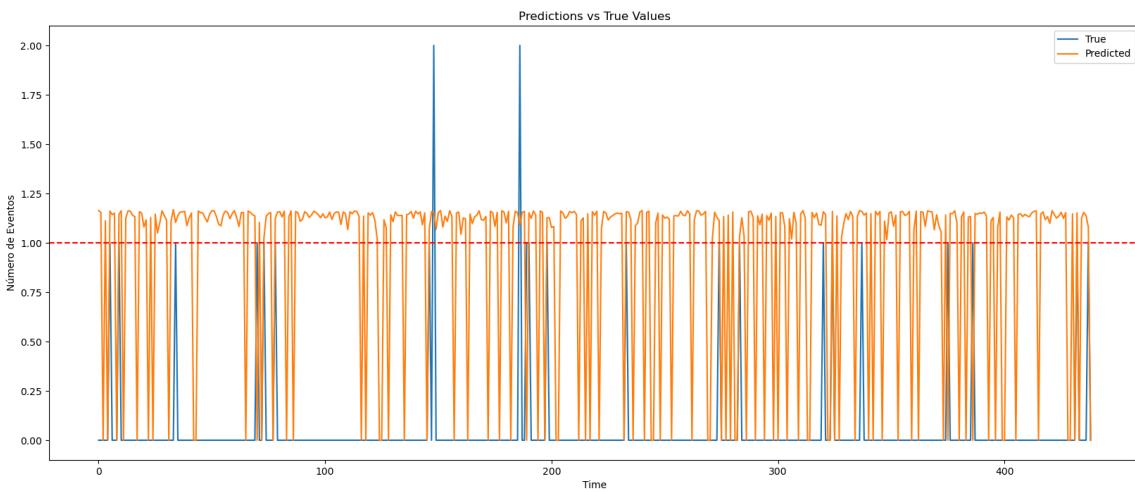


Figura 25: LSTM com pesos do Concelho de Lisboa

Embora os resultados apresentem uma melhoria em comparação ao modelo original, ainda não alcançam um nível satisfatório de precisão. Diante dessa constatação, optamos por simplificar o problema, focando na previsão binária de ocorrência ou não de quebras energéticas, em vez de tentar prever o número absoluto de eventos.

Essa mudança estratégica visa direcionar os esforços para um objetivo mais realista e alcançável, considerando as limitações do modelo e a natureza complexa do problema. Para uma análise mais crítica do modelo, abaixo podemos visualizar as métricas deste modelo:

```

1 14/14 ━━━━━━ 0s 3ms/step
2 Mean Squared Error (MSE): 1.0896107969400848
3 Root Mean Squared Error (RMSE): 1.043844239788717
4 Mean Absolute Error (MAE): 1.0114270582506624
5 Mean Absolute Percentage Error (MAPE): 4506763245347768.0
6 Mean Bias Deviation (MBD): 0.9956297822771827
7 Explained Variance Score: -0.6175031679929905

```

Listing 1: Métricas do modelo LSTM com pesos

5.2.1.2. Previsão binária de ocorrência de eventos

A conversão do problema para binário foi uma etapa simples, bastando definir qualquer valor de quebra energética superior a zero como 1, exemplo:

```
QuebrasFinal["HouveEvento?"] = np.where(QuebrasFinal["n_incidentes"] > 0, 1, 0).
```

Apesar da expectativa de que a conversão em binário equilibraria as classes, a análise revela que o reequilíbrio não foi significativo, como evidenciado pela [Tabela 5](#).

Tabela 5: Ocorrência, binário, de quebras energéticas ao longo de 6 anos (2018 - 2023)

Ocorreu alguma quebra energética?	Ocorrências
0	558072
1	2824

Mesmo assim, decidimos aplicar o mesmo modelo LSTM com os pesos, e o resultado foi o seguinte:

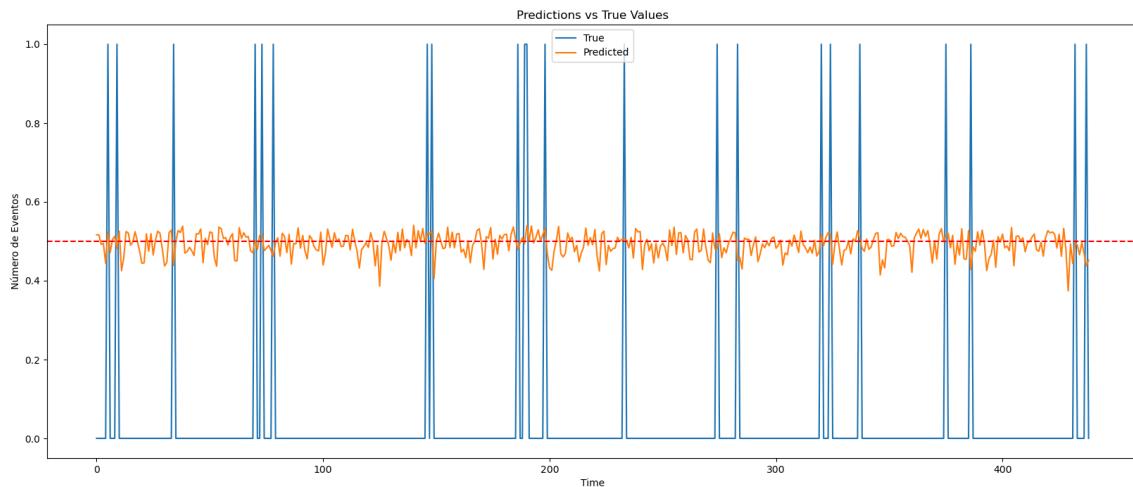


Figura 26: Previsão binária com o LSTM com pesos do Concelho de Lisboa não tratado

Embora o modelo apresente uma tendência em concentrar-se em valores entre 0,4 e 0,6, podemos ajustá-lo para melhorar o seu desempenho na classificação binária.

Ao definir uma reta de 0,5, podemos converter os valores acima desse limite em 1 (predizendo uma quebra energética) e os valores abaixo em 0 (predizendo a ausência de quebra). Essa estratégia visa forçar o modelo a tomar decisões mais definitivas, categorizando cada evento como 0 ou 1.

Os resultados obtidos com essa abordagem são apresentados a seguir:

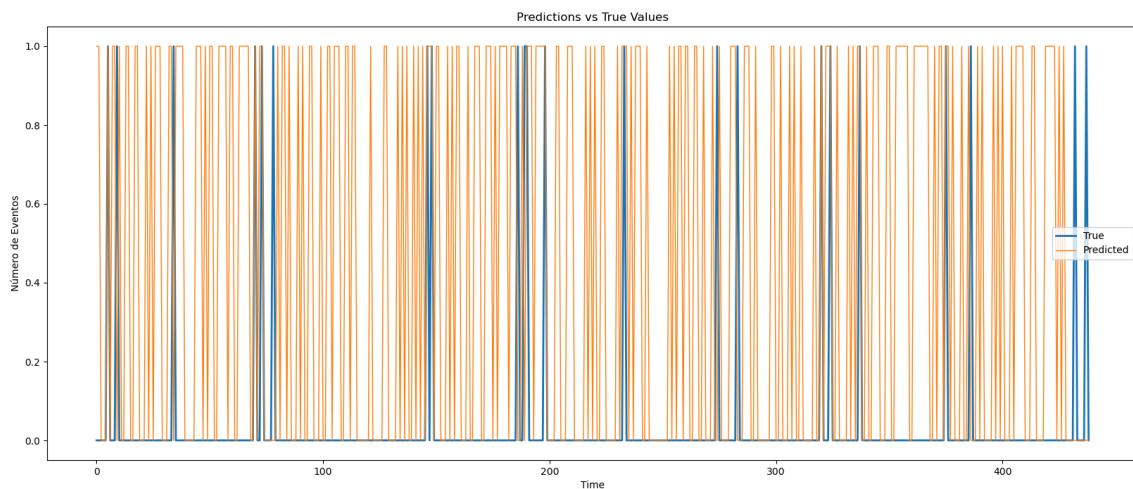


Figura 27: Previsão binária com o LSTM com pesos do Concelho de Lisboa tratado

```

1 14/14 ━━━━━━ 0s 3ms/step
2 Mean Squared Error (MSE): 0.24366735223373623
3 Root Mean Squared Error (RMSE): 0.49362673371053983
4 Mean Absolute Error (MAE): 0.49276345316532805
5 Mean Absolute Percentage Error (MAPE): 2105084156437207.8
6 Mean Bias Deviation (MBD): 0.44208170653475715
7 Explained Variance Score: -0.01320570042571756

```

Listing 2: Métricas do modelo LSTM com pesos

5.2.2. Modelo supervisionada de classificação

Diante dos resultados insatisfatórios obtidos com o modelo LSTM na [Secção 5.2.1](#), o grupo optou por explorar uma abordagem alternativa para o problema em questão. Reconhecendo a necessidade de uma solução mais robusta e confiável, a equipe decidiu direcionar seus esforços para modelos de classificação mais tradicionais.

Com base em conhecimento prévio e *expertise* na área, o modelo Gradient Boosting foi selecionado como a ferramenta mais adequada para o desafio. A robustez e a confiabilidade deste modelo, reconhecido em diversos trabalhos realizados pelo grupo, faz-nos acreditar que este trará resultados mais positivos e consistentes.

O grupo teve logo a noção que, caso aplicássemos os nossos dados binários no modelo, ele sofreria de [Overfitting](#), graças ao grande desequilíbrio de classes (como é possível visualizar na [Tabela 5](#)). Em [Anexo I](#) é possível visualizar a respetiva matriz de confusão do modelo base com o [Overfitting](#).

Para esta fase serão utilizadas duas abordagens: **Oversampling** e **Undersampling**

5.2.2.1. Oversampling

Nesta secção, aplicamos um oversampling clássico. A seguir, será demonstrada a *feature importance* do modelo, bem como a matriz de confusão deste modelo quando aplicado na base de dados original. Isto porque faz mais sentido avaliar o nosso modelo com as métricas de um caso real em vez de um caso “sintético”.

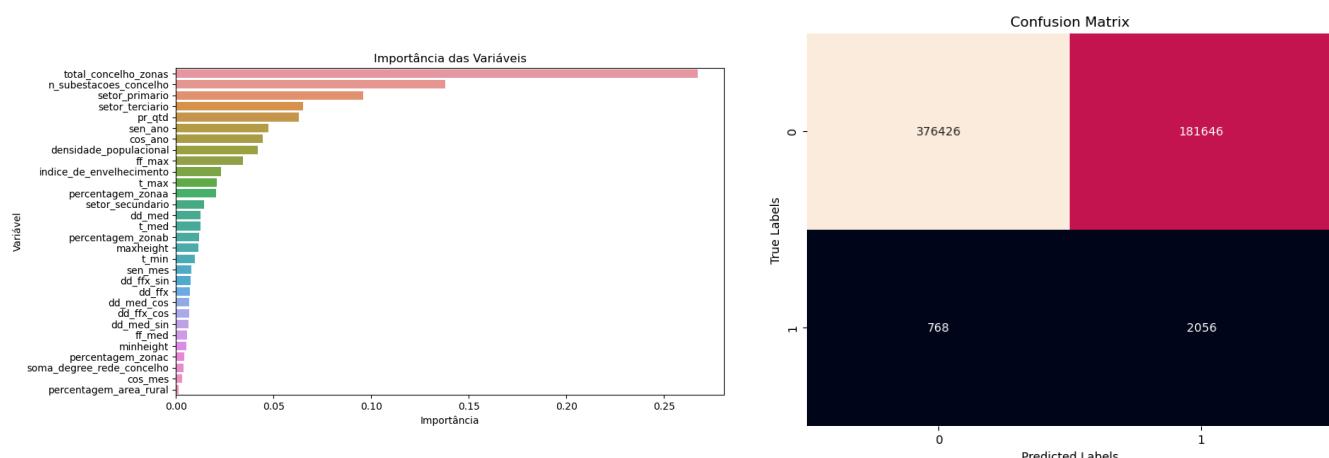


Figura 28: Gradient Boosting - Oversampling

- 1 Acurácia: **0.675**
- 2 Precisão: **0.011**
- 3 Recall: **0.728**
- 4 F1-score: **0.022**

Listing 3: Métricas do Gradient Boosting com oversampling

É necessário analisar e avaliar as métricas com atenção! Apesar da acurácia ser elevada, a precisão e o F1-Score apresentam valores muito baixos, o que demonstra que o modelo não é totalmente fidedigno na previsão de ocorrências de quebras. Mesmo assim, ele aparenta conseguir identificar algumas quebras energéticas.

5.2.2.2. Undersampling

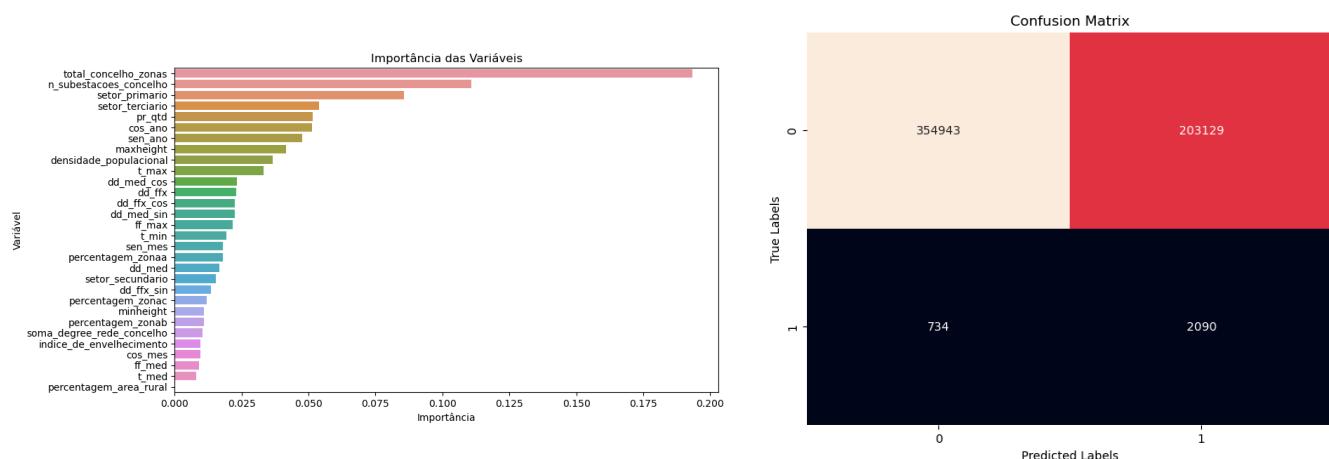


Figura 29: Gradient Boosting - Undersampling

- 1 Acurácia: **0.637**
- 2 Precisão: **0.010**
- 3 Recall: **0.740**
- 4 F1-score: **0.020**

Listing 4: Métricas do Gradient Boosting com undersampling

Se comparados com atenção, tanto o *oversampling* como o *undersampling* apresentam resultados muito parecidos, quer seja pelas variáveis mais importantes, como pelas métricas. Mesmo assim, se tivéssemos que optar por uma delas, optaríamos pelo *oversampling* devido ao maior valor no F1-Score e na precisão. Infelizmente, os resultados não foram muito enriquecedores.

6. Tentativa de 2º iteração do CRISP-DM

Foi realizada uma segunda iteração do CRISP-DM, onde se apostou mais na fase de *modelling*, devido aos seus resultados não satisfatórios. Para tal, foi realizada uma elevada pesquisa, que consumiu uma maior parte do tempo.

Para esta segunda iteração, uma das nossas preocupações seria arranjar uma solução para deixar de assumir a independência entre concelhos. Algumas das formas que consideramos, em vés de treinar um concelho de cada vez e ter um modelo para cada concelho, seria treinar sobre os concelhos todos e o modelo iria prever para todos os dias e para todos os concelhos; ou usar um codificador do concelho (usando a informação demográfica e climática de tal), treinar e prever sobre o conjunto de teste, e de seguida usar um descodificador, semelhante ao apresentado na Figura 7 no artigo de [Zhao \(2021\)](#). Considerando a natureza do estudo de previsão de eventos, utilizar métricas de avaliação tradicionais pode conduzir a resultados inconclusivos e/ou conclusões erradas. Em vez disso, focamos na previsão de eventos em concelhos adjacentes no dia em que eventos reais ocorreram em um determinado concelho. Essa abordagem alinha-se melhor com os nossos objetivos, pois não leva em consideração a distância entre os eventos previstos e os eventos reais.

6.1. Distância entre eventos previstos

Para desenvolver métricas neste ciclo de avaliação, precisamos determinar a distância entre eventos previstos \hat{y} e eventos reais y . Essa distância é calculada considerando o número de dias entre as datas dos eventos e a distância geográfica entre os concelhos envolvidos. Optamos por calcular a distância usando o caminho mais curto entre os concelhos, em vez da distância euclidiana entre os seus centros, para refletir melhor o comportamento real das redes, mesmo que ambos os métodos frequentemente produzem resultados semelhantes. Implementamos um grafo de concelhos vizinhos e um algoritmo de caminho mais curto para facilitar este cálculo. Isso é particularmente útil quando concelhos próximos geograficamente estão separados por corpos de água, como Lisboa e Almada.

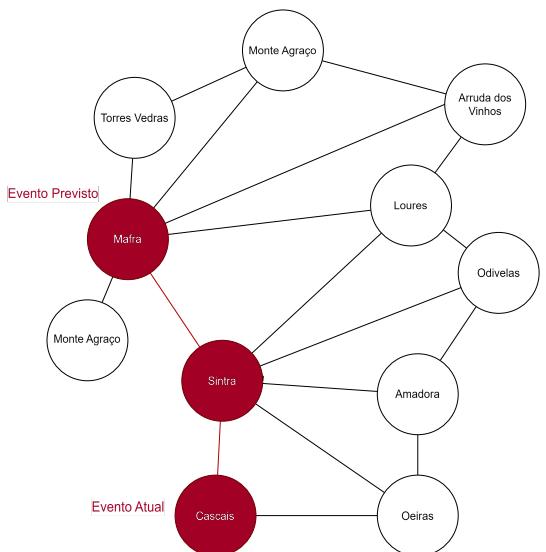


Figura 30: Distância geográfica entre concelhos implementada

Sendo $\Lambda(y, \hat{y})$ a distância entre o evento real y e o evento previsto \hat{y} , $P(y, \hat{y})$ o tamanho do caminho mais pequeno entre os concelhos dos eventos, e $\Delta(y, \hat{y})$ a diferença de dias entre os eventos, assentámos na seguinte função:

$$\begin{aligned}\Lambda(y, \hat{y}) &= \beta_1 P(y, \hat{y}) + \beta_2 |\Delta(y, \hat{y})| \\ , \beta_1 &= 7, \beta_2 = 1 \text{ (rácio de 7:1)}\end{aligned}\tag{7}$$

Este rácio significa que acertar ao lado do concelho terá a mesma distância do que acertar no concelho com uma semana de diferença. Achámos o rácio mostrado apropriado, sendo que achamos mais importante para o modelo acertar no concelho ou perto do concelho, do que acertar exatamente no dia.

6.2. Correspondência de eventos

Para elaborar as métricas, seria necessário primeiro determinar como devíamos ligar eventos reais e eventos previstos. Como achámos mais apropriado fazer correspondência bipartida ([Zhao, 2021](#)), isto consegue-se transformar em um [problema de atribuição de tarefas](#). Nós no início elaboramos uma heurística para ligar os eventos reais com eventos previstos, mas ao pesquisar melhor decidimos usar o algoritmo de Kuhn-Munkres ([Hungarian method](#)). Como este funciona não faz parte do âmbito deste projeto, mas de forma reduzida, este vai encontrar os melhores pares de eventos previstos e eventos reais, minimizando $\sum_{l \in L} \Lambda(l_y, l_{\hat{y}})$, sendo l o par de evento real e evento previsto, e L os pares resultados do algoritmo.

6.3. Métricas de Avaliação

As métricas que desenvolvemos para esta segunda iteração são de dois tipos, baseados no estudo de [Zhao \(2021\)](#):

- Qualidade de previsões, que avalia o quanto perto o modelo prevê os eventos;
- Adequação de previsões, que avalia o desempenho do modelo a prever eventos.

Nesta primeira, desenvolvemos as seguintes métricas:

$$\begin{aligned}\delta_{\text{normal}} &= \sum_{l \in L} \Lambda(l_y, l_{\hat{y}}) + 10 \cdot |\#Y - \#\hat{Y}| \\ \delta_{\text{punishing}} &= \sum_{l \in L} \Lambda(l_y, l_{\hat{y}}) + 100 \cdot |\#Y - \#\hat{Y}|\end{aligned}\tag{8}$$

Estas métricas são simplesmente a soma das distâncias dos pares determinados, mas como pode haver eventos não previstos, ou eventos previstos a mais, é necessário adicionar o argumento da direita, para poder ter isso em conta. $\delta_{\text{punishing}}$, comparado com δ_{normal} , vai punir 10 vezes mais quando o modelo prevê eventos a mais ou a menos, podendo assim verificar de forma mais precisa a qualidade de previsões.

No segundo tipo, vamos adicionar um limiar t , e designar todos os pares $\Lambda(l_y, l_{\hat{y}}) \leq t$ como pares aceitáveis, classificando-os como *True Positives* (TP). Com esta noção podemos usar a métrica de classificação F_1 , levando à seguinte métrica:

$$F_1 = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}, \begin{cases} \text{TP} = \#\{l \in L : \Lambda(l_y, l_{\hat{y}}) \leq t\} \\ \text{FN} = \#\{y : (\exists l_{y,\hat{y}} : \Lambda(l_y, l_{\hat{y}}) > t \vee \nexists y)\} \\ \text{FP} = \#\{\hat{y} : (\exists l_{y,\hat{y}} : \Lambda(l_y, l_{\hat{y}}) > t \vee \nexists \hat{y})\} \end{cases}\tag{9}$$

Esta métrica vai medir, baseado num limiar adequado, o quanto bom o modelo é a prever eventos. Nós recomendamos um $t = 25$, sendo que isto dá alguma liberdade do modelo prever em concelhos ao lado com alguma distância temporal também, mas este valor depende também do rácio usado para Λ . Para prever exatamente para o dia, semelhante a um modelo de classificação de *time series*, $t = 0$ seria adequado.

6.4. Modelação da segunda interação

Devido à falta de tempo, não foi possível implementar alguns dos modelos que prevíamos. No entanto, só como prova de conceito, treinámos dois modelos auto-regressivos: um AR e um ARIMA. Semelhante ao primeiro ciclo, assumimos independência entre concelhos e fizemos um modelo para cada concelho. Devido à natureza dos modelos, em vés de

termos os eventos diários, acumulámos os eventos para ter uma tendência estritamente positiva. Analogamente também ao primeiro ciclo, treinámos 2018 a 2022, deixando 2023 como teste.

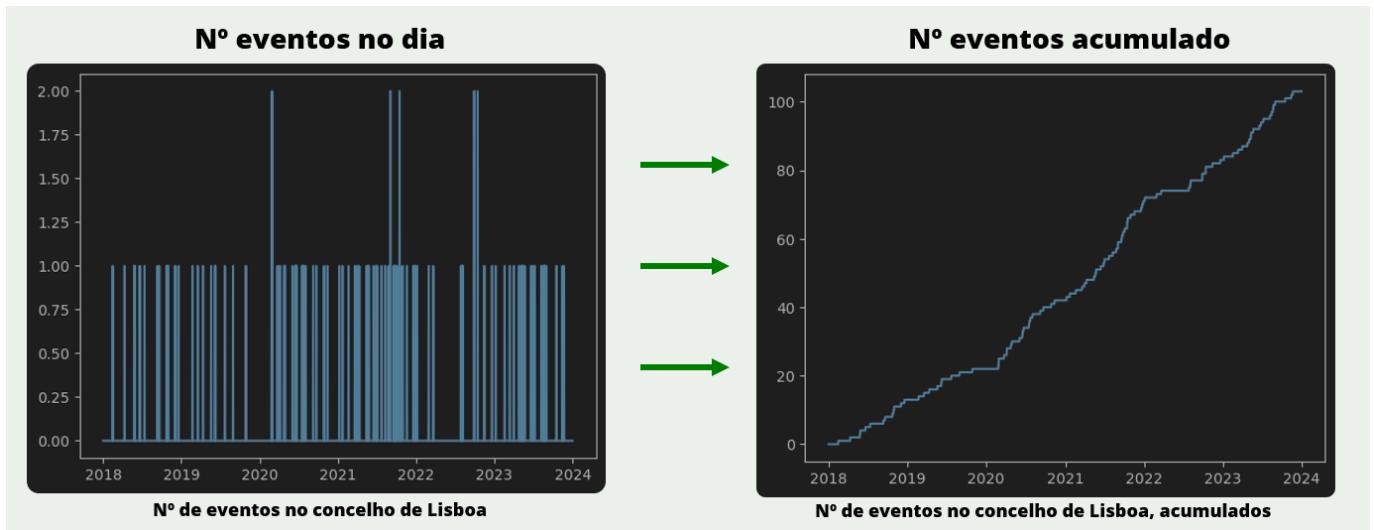


Figura 31: Representação de eventos acumulados

Em conformidade com a primeira iteração também, na modelação, arredondámos para o inteiro mais perto depois da modelação. Desta forma, conseguimos extrair quando é que o modelo considerou evento e quando não.

6.4.1. AR(1,1)

O modelo teve os seguintes resultados:

- $\delta_{\text{normal}} = 79982$
- $\delta_{\text{punishing}} = 101672$
- $F_1 \approx 0.0038$
- $\#\hat{Y} - \#Y = 241$

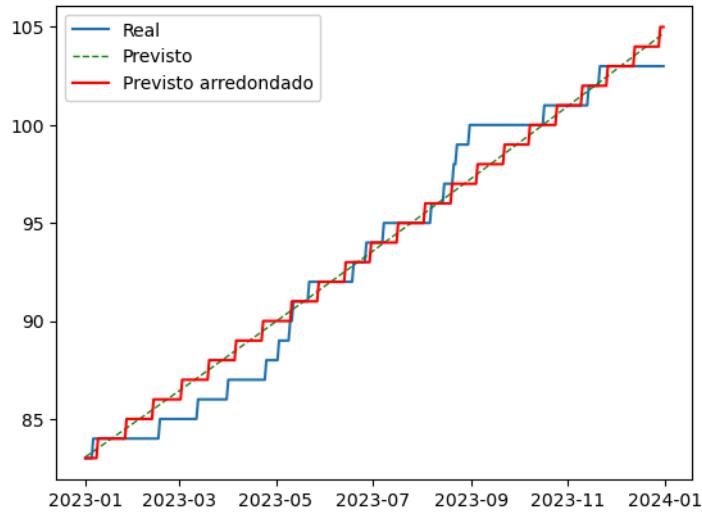


Figura 32: Resultados do modelo AR(1,1) em Lisboa

Os resultados indicam que há um número gigante de eventos previstos a mais, o modelo não tem boa qualidade de previsões devido aos altos δ , e não é um modelo adequado para previsão dos eventos, baseado no pequeno valor de F_1 . Embora a Figura 32 pareça demonstrar um bom resultado em Lisboa, temos que ter em noção que isto não é um modelo de regressão, e que há muitos outros concelhos que não mostramos.

6.4.2. ARIMA(2,1,0)

Para determinar parâmetros ARIMA adequados, usámos o `auto_arima` da [package `pmdarima`](#) no concelho de Lisboa. Embora este possa não ser os melhores parâmetros para todos os concelhos, foi bom o suficiente para o que queremos demonstrar. Os resultados foram os seguintes:

- $\delta_{\text{normal}} = 69677$ (melhorou 12.88%)
- $\delta_{\text{punishing}} = 81017$ (melhorou 20.32%)
- $F_1 \approx 0.0043$ (melhorou 11.63%)
- $\#\hat{Y} - \#Y = 126$ (melhorou 47.72%)

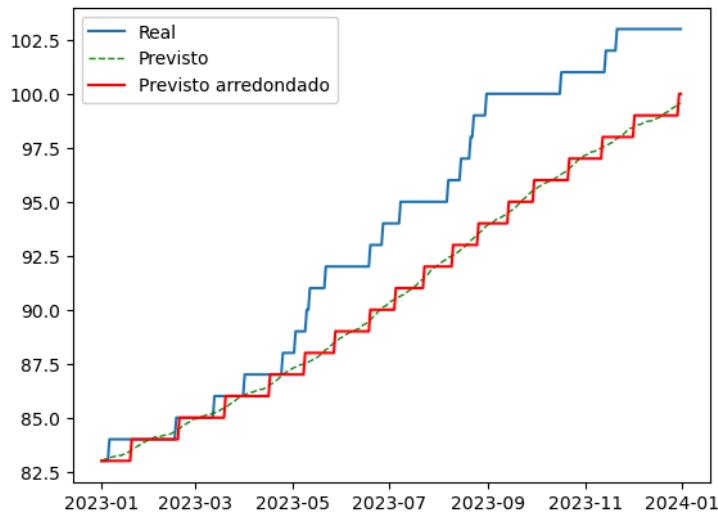


Figura 33: Resultados do modelo ARIMA(1,1) em Lisboa

Este modelo, independentemente do outro, continua a ter má qualidade, má adequabilidade, e continua a prever eventos a mais. Comparado com o modelo anterior, o modelo melhorou todas as métricas, mostrando que este modelo é melhor que o outro a prever os eventos. Olhando para a , parece que este modelo não se adequa tanto como no modelo anterior em Lisboa; no entanto, lembrando que este não é um modelo de regressão, olhando para quando o modelo considera os eventos, até parece que o modelo está a prever relativamente perto dos dias reais, apenas não correspondendo à frequência de eventos que Lisboa apresenta.

7. Conclusão e deployment

A deteção precisa de quebras de energia não previstas e excepcionais continua a ser um desafio significativo hoje em dia. A raridade e a natureza complexa desses eventos dificultam o desenvolvimento de modelos eficazes.

A falta de especificidade geotemporal impede análises precisas e previsões confiáveis. Além disso, os indicadores gerais de quebras de energia não possuem poder discriminatório suficiente para distinguir entre eventos distintos, exceto em casos excepcionais de grande impacto. Apesar dos avanços alcançados no decorrer do projeto, os resultados atuais dos modelos demonstram limitações em termos de fidedignidade.

Embora os desafios persistam, o potencial para aprimorar a deteção de quebras de energia é considerável e traduz para um investimento em pesquisas futuras, como:

- Usar métodos de aprendizagem mais robustos para captar padrões complexos nos dados e lidar com maior volume de informações;
- Usar modelos que usam dados sobre o evento, como Cadeias de Markov Escondidas, ou tratar os eventos como censurados;
- Recolha de dados mais integrada com a E-Redes, e implementação de sensores para uma previsão melhor;
- Integração de dados das ilhas para uma análise mais aprofundada.

Este trabalho foi de extrema importância para percebermos como é o funcionamento de toda a rede elétrica e das quebras energéticas.

Por fim, gostaríamos de agradecer à E-Redes pela disponibilidade de dados e o conhecimento especializado e ao nosso professor orientador Luís Nunes pela sua orientação inestimável e apoio contínuo.

Trabalhos futuros:

- <https://dl.acm.org/doi/pdf/10.1145/3450287>
- <https://ragulpr.github.io/2016/12/22/WTTE-RNN-Hackless-churn-modeling/>

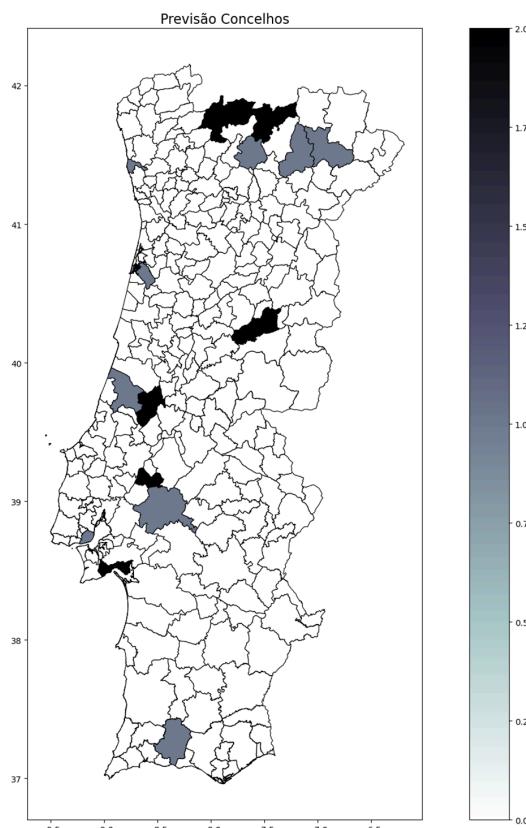


Figura 34: Previsão de quebras energéticas para o dia da apresentação utilizando o modelo LSTM

8. Anexos

Anexo A - IPMA

Nos dados fornecidos pela entidade tínhamos acesso a algumas estações espalhadas pelos distritos de Portugal Continental que continham a informação das variáveis mostradas na [Tabela 6](#) e [Tabela 7](#).

Tabela 6: As diferentes estações do IPMA

Nº da Estação	Nome da Estação
1200551	Viana do Castelo / Chafé
1210622	Braga / Merelim
1200567	Vila Real / Aeródromo
1200575	Bragança
1200545	Porto / Pedras Rubras
1210702	Aveiro / Universidade
1200560	Viseu / Centro Coordenador
1210683	Guarda
1200548	Coimbra / Cernache / Aeródromo
1200570	Castelo Branco
1210718	Leiria
1210734	Santarém / Fonte Boa Est. Zootécnica
1200571	Portalegre
1200579	Lisboa / Gago Coutinho
1210770	Setúbal / Estação de Fruticultura
1200558	Évora / Aeródromo
1200562	Beja
1200554	Faro / Aeroporto
1210615	PONTE DE LIMA / Escola Agrícola
1210716	Ansião

Tabela 7: As diferentes variáveis fornecidas pelo IPMA

Variável	Descrição	
T_MED	°C	Temperatura média do ar a 1,5 m
T_MAX	°C	Temperatura máxima do ar a 1,5 m
T_MIN	°C	Temperatura mínima do ar a 1,5 m
DD_MED	°	Rumo médio do vento
DD_FFX	°	Rumo do vento máximo
FF_MED	m/s	Intensidade média do vento 10 m
DD_MAX	m/s	Intensidade máxima instantânea do vento
PR_QTD	mm	Quantidade de precipitação

Anexo B - GPP - Gabinete de Planeamento, Políticas e Administração Geral

[Programa de Desenvolvimento Rural](#) de 2020, que nos demonstra todas as freguesias que são consideradas *zonas rurais*, em Portugal Continental.

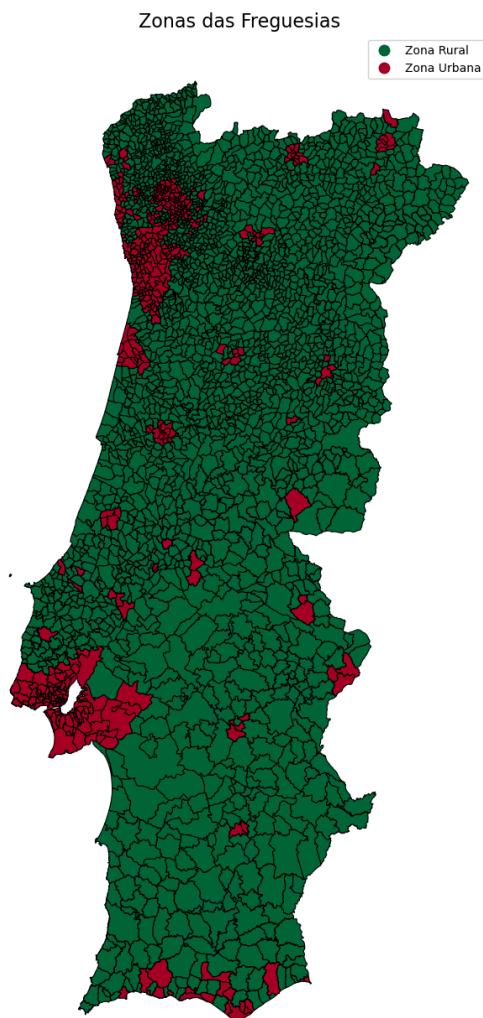


Figura 35: Zonas Rurais em Portugal Continental

Anexo C - Dados Demográficos - Pordata

- **Densidade Populacional:** Ao entendermos a distribuição da população no concelho, podemos identificar áreas com elevada densidade populacional, que podem enfrentar uma maior procura de energia, assim como áreas mais dispersas que podem ter necessidades diferentes em termos de infraestrutura energética.
- **Índice de Envelhecimento:** O conhecimento da proporção de idosos em relação à população mais jovem pode ajudar-nos a antecipar demandas específicas de energia, como cuidados de saúde, sistemas de aquecimento e adaptações para acomodar necessidades especiais.
- **Setor de Atividade:** Ao analisarmos os setores económicos predominantes no concelho, podemos identificar padrões de consumo de energia em diferentes indústrias e setores, permitindo uma melhor alocação de recursos e planeamento energético.

Anexo D - Desequilíbrio das Classes

Distribuição das Causas dos Incidentes

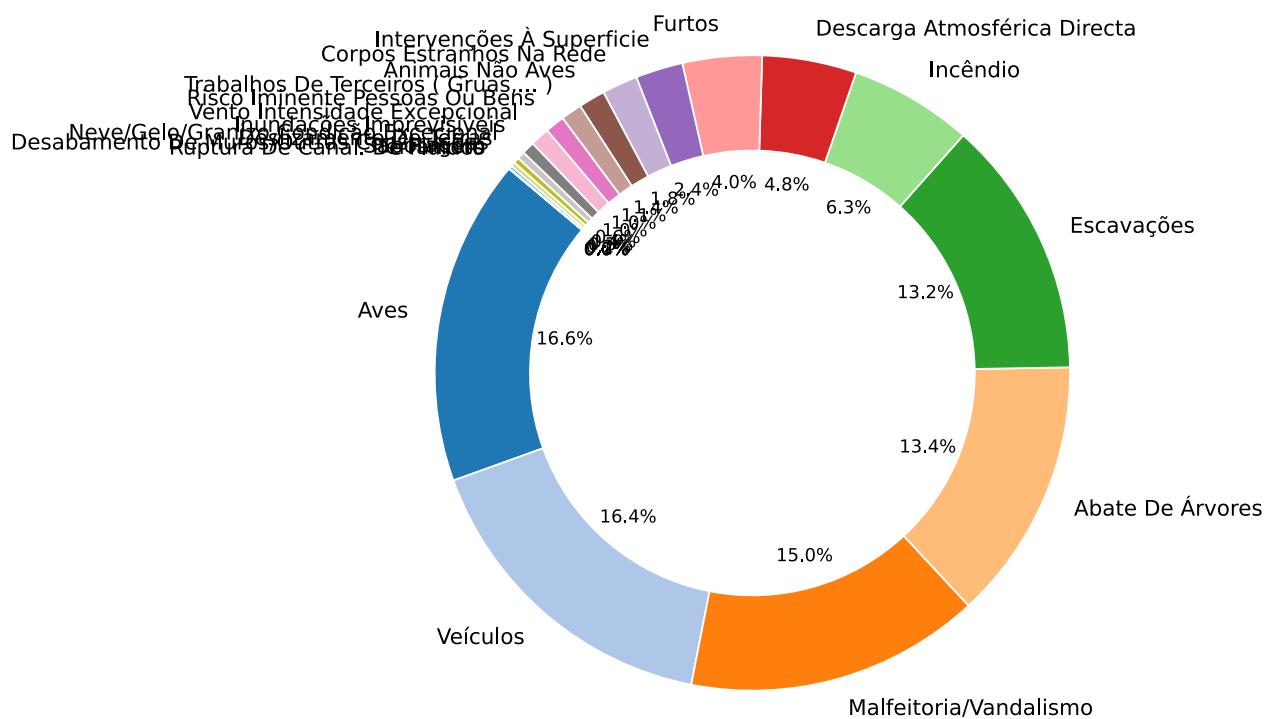


Figura 36: Distribuição desequilibrada das causas dos incidentes

“A imagem acima é meramente ilustrativa e não foi utilizada ou considerada pelo grupo ao longo do trabalho final, pelo que não deve ser interpretada ou tomada como referência para qualquer conclusão.“ (Grupo E-REDES 3)

Anexo E - Taxa de ruralidade

Para esquematizarmos alguma informação relevante e não termos dados a mais, decidimos agregar as variáveis da Zona Rural e Zona Urbana que retirámos da [Secção Anexo B](#). Estas duas variáveis representam o número de freguesias existentes em cada um dos concelhos, indicando quantas são consideradas rurais e quantas são consideradas urbanas, respetivamente.

Por isso, através do código desenvolvido na [Listing 5](#), foi produzido o coeficiente que designámos de “taxa de ruralidade”, dado pela expressão: Tx Ruralidade =

$$\frac{\text{Zona Rural}}{\text{Zona Urbana} + \text{Zona Rural}}.$$

Com isto conseguimos perceber se um concelho tem um rácio maior ou menor de freguesias rurais. Abaixo, podemos visualizar o respetivo código, como também três exemplos distintos no distrito de Braga, com valores diferentes de “taxa de ruralidade”:

Taxa de Ruralidade de cada concelho

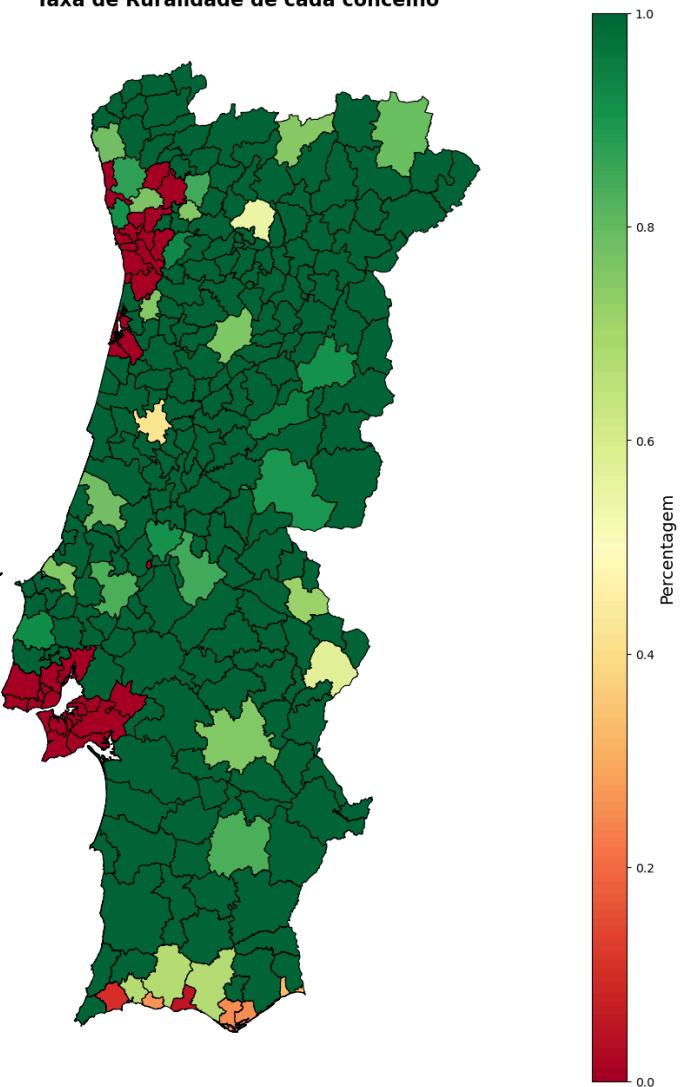


Figura 37: Taxa de ruralidade de cada concelho

```
1 file_path = data_folder / "ZonasCount.csv"
2 ZonasRurais = pd.read_csv(file_path)
3 ZonasRurais["Rural"] = ZonasRurais["Zona Rural"] / (ZonasRurais["Zona Urbana"] +
ZonasRurais["Zona Rural"])
```

Listing 5: Cálculo da taxa de ruralidade *python*

Vila Nova de Famalicão:

$$\text{tx. ruralidade} = \frac{27}{27+7} \equiv \\ \text{tx. ruralidade} \approx 0.79$$

Vizela:

$$\text{tx. ruralidade} = \frac{5}{5+0} \equiv \\ \text{tx. ruralidade} \approx 1.00$$

Espordeste:

$$\text{tx. ruralidade} = \frac{0}{0+7} \equiv \\ \text{tx. ruralidade} \approx 0.00$$

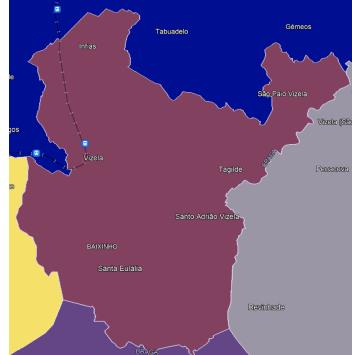
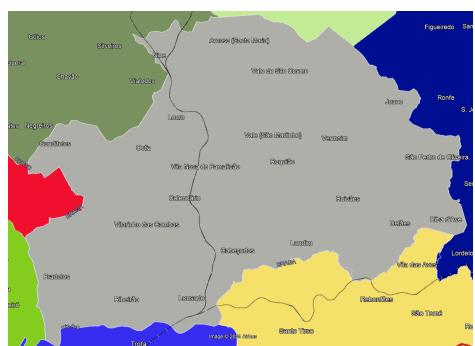


Figura 38: Taxa de ruralidade para vários concelhos do distrito de Braga e contorno dos mesmos

Anexo F - Eventos excepcionais de grande impacto

Tabela 8: Eventos excepcionais descritos entre o período de 2019 e 2022

Nome	Tipo de Evento	Período	Média de duração da interrupção	Impacto QEE	Nº pedidos atraso	Nº clientes afetados	Boletim Informativo
Temporal Região Sul	Trovoadas e inundações	7 e 8/12/2022	< 1 min	significativo	13	32 809	Notícia IPMA
Depressão Efrain	Condições meteorológicas adversas	12 a 13/12/2022	6 min	significativo	43	387 223	Notícia IPMA
Depressão Hortense	Condições meteorológicas adversas	21 e 22/01/2021	3 min	significativo	15	297 636	Notícia Expresso 21/1/2021
Deslastre de Carga Automático	Sobrecarga e disparo de ligações	24/07/2021	7 min, em média	—	exclusão do serviço	977 394	ERSE
Rio Atmosférico	Condições metereológicas adversas	29/10/2021	3 min, em média	significativo	33	237 413	ERSE
Depressão Glória	Condições metereológicas adversas	19/01/2020	7 min, em média	significativo	36	181 984	ERSE
Depressão 1 e 2/03/2020	Condições metereológicas adversas	1 e 2/03/2020	3 min, em média	relevante	17	266 104	ERSE
Tempestade Helena	Condições metereológicas adversas	1/02/2019	3 min, em média	significativo	32	240 299	ERSE
Depressões Elsa e Fabien	Condições metereológicas adversas	18 a 23/12/2019	1h 20min, em média, visto nas três tensões	relevante	800	1 699 906	ERSE

Anexo G - Exemplo de cluster “errado”

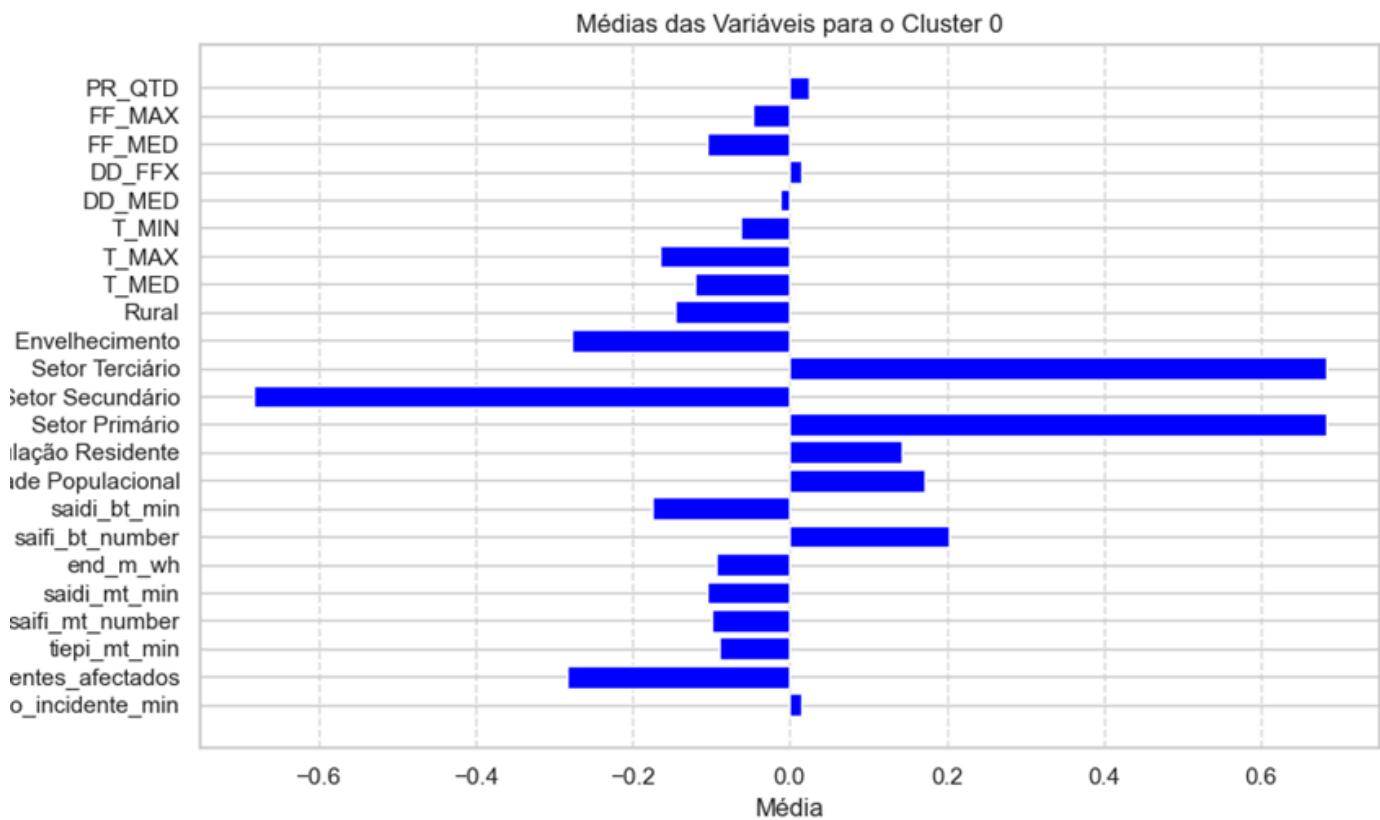


Figura 39: Distribuição desequilibrada das causas dos incidentes

“A imagem acima é meramente ilustrativa e não foi utilizada ou considerada pelo grupo ao longo do trabalho final, pelo que não deve ser interpretada ou tomada como referência para qualquer conclusão.” (Grupo E-REDES 3)

Anexo H - Exemplo de previsão “errada”

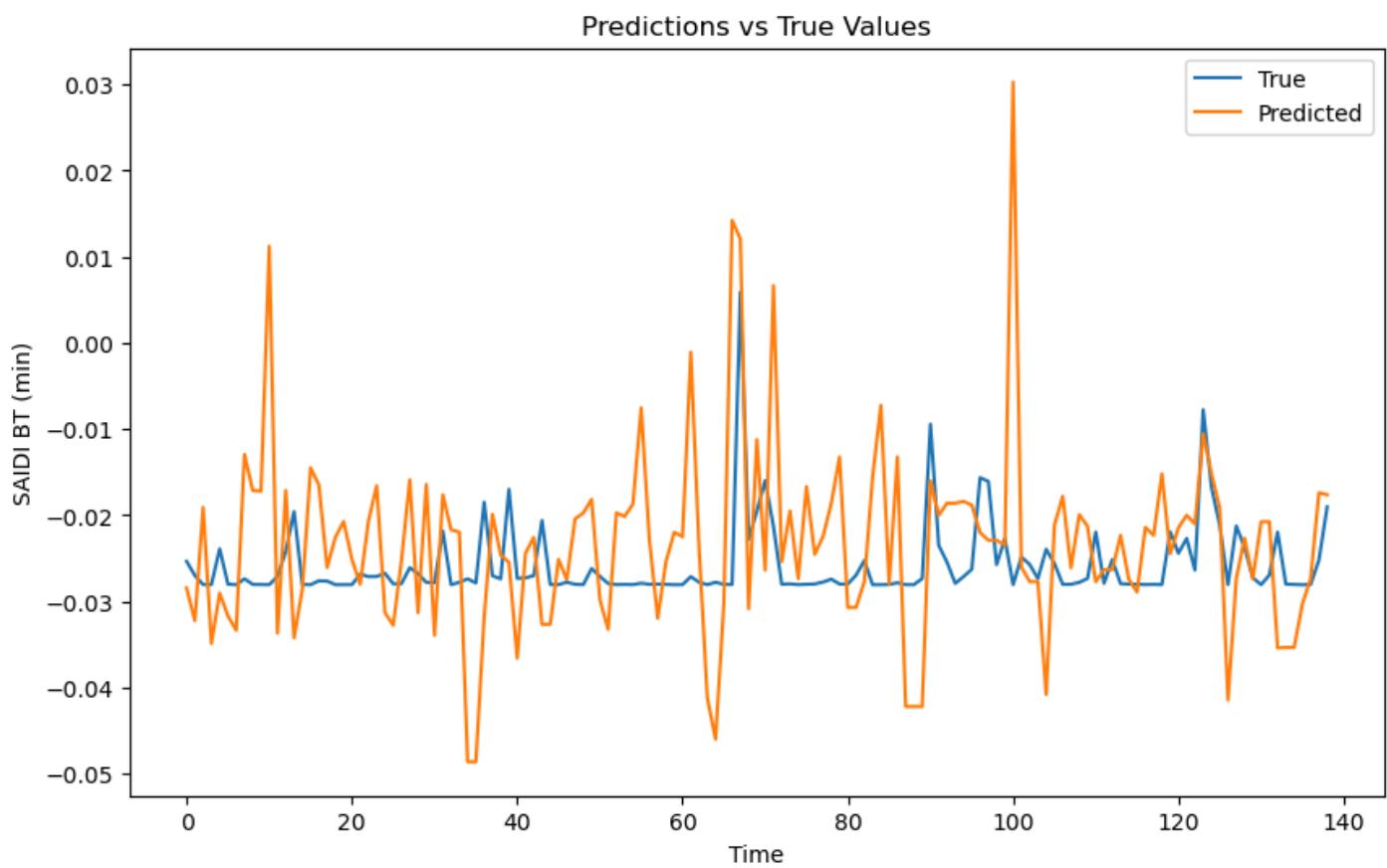


Figura 40: Previsão do indicador SAIDI BT (min) em Lisboa

“A imagem acima é meramente ilustrativa e não foi utilizada ou considerada pelo grupo ao longo do trabalho final, pelo que não deve ser interpretada ou tomada como referência para qualquer conclusão.” (Grupo E-REDES 3)

Anexo I - Gradient Boosting com overfitting

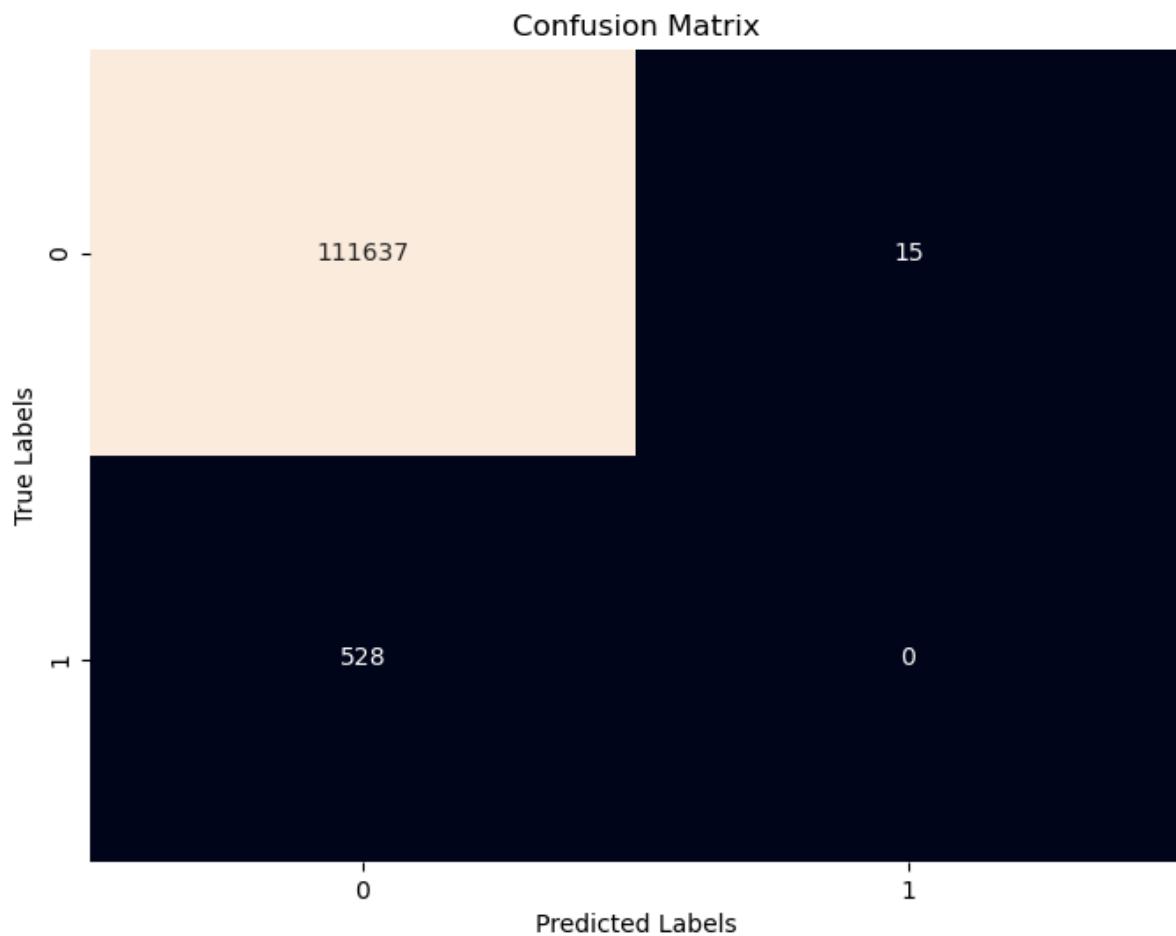


Figura 41: Matriz de confusão do Gradient Boosting com overfitting