POLITÉCNICO DE LEIRIA
ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO

Data Mining

Aplicada à Gestão de Incêndios

No âmbito de Data Mining

Mestrado em Ciência de Dados

Eliana Oliveira
2240276

Jéssica Grácio
2240549

Maria Fialho
2240286

27 de Junho de 2025

Índice

Índice de Figuras

Índice de Tabelas

Introdução

A problemática dos incêndios florestais em Portugal tem assumido, nas últimas décadas, um impacto social, económico e ambiental significativo.

Neste contexto, a análise de dados históricos sobre incêndios surge como uma ferramenta fundamental para compreender padrões de ocorrência e apoiar decisões estratégicas.

O caso de estudo selecionado baseia-se numa base de dados disponibilizada pelo Instituto da Conservação da Natureza e das Florestas (ICNF). Esta base de dados contém informação, desde 2001 até à atualidade, sobre os incêndios detetados pelo Sistema de Gestão de Informação de Incêndios Florestais (SGIF) - sistema que cataloga os incêndios desde o momento que existe despacho de meios de combate a incêndios. A base de dados tem informação detalhada sobre ocorrências de incêndios, incluindo localizações, datas, causas prováveis, áreas ardidas, tempos de resposta, entre outros atributos relevantes.

Torna-se, assim, pertinente utilizar abordagens analíticas avançadas que permitam extrair conhecimento útil a partir dos dados disponíveis, contribuindo para a prevenção, resposta e mitigação de futuros incidentes.

A problemática acima descrita motivou a realização deste trabalho, no âmbito da unidade curricular de Data Mining, integrada no Mestrado em Ciência de Dados do Politécnico de Leiria. O seu principal objetivo é aplicar a metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*), tendo por base a análise de dados reais relacionados com incêndios florestais em Portugal.

A metodologia CRISP-DM foi adotada como estrutura base para a execução do projeto, sendo reconhecida como uma das abordagens mais completas e flexíveis para a condução de projetos de data mining. Esta metodologia divide-se em seis fases principais: compreensão do negócio, compreensão dos dados, preparação dos dados, modelação, avaliação e colocação em produção (implementação).

Com este projeto, espera-se não só aplicar e aprofundar os conhecimentos adquiridos na unidade curricular de Data Mining, mas também contribuir para uma melhor compreensão da problemática dos incêndios florestais em Portugal, fornecendo informação útil que possa auxiliar na prevenção, gestão e mitigação destes eventos devastadores.

Fase 1: Compreensão do Negócio

Nesta fase inicial do projeto, procura-se adquirir uma compreensão aprofundada do problema de negócio em análise, dos objetivos estratégicos que se pretendem alcançar e dos fatores que podem influenciar a execução e o sucesso do projeto. A metodologia CRISP-DM recomenda que esta etapa seja desenvolvida antes de qualquer exploração técnica dos dados, garantindo que todas as decisões técnicas estejam alinhadas com as necessidades reais da organização ou domínio em estudo.

Neste capítulo, serão abordados os seguintes temas: descrição do negócio, definição dos objetivos de negócio e respetivos critérios de sucesso; avaliação do cenário atual, incluindo inventário de recursos, requisitos, assunções, restrições, riscos e terminologia; análise dos custos e benefícios; identificação dos objetivos de data mining e critérios de sucesso associados; e, por fim, a definição do plano de projeto e das ferramentas/metodologias inicialmente consideradas.

1.1 Identificação dos objetivos do negócio

1.1.1 Descrição do Negócio

A ocorrência de incêndios florestais em Portugal constitui um problema recorrente e complexo, com impactos profundos na economia, no ambiente e na segurança das populações. O ICNF é a entidade responsável pela gestão e monitorização destes eventos, mantendo um registo sistemático das ocorrências, com dados que cobrem múltiplas dimensões dos incidentes.

O negócio, neste contexto, consiste na gestão do território florestal e dos recursos afetos à prevenção e combate aos incêndios, envolvendo entidades como a Proteção Civil, Bombeiros, Municípios e forças de segurança. A análise detalhada e sistemática dos dados históricos de incêndios permite apoiar a tomada de decisão, melhorar os tempos de resposta, prevenir reincidências e otimizar a alocação de meios.

No âmbito deste projeto, foi possível estabelecer contacto com o responsável pela gestão dos dados no ICNF, Comandante Rui Almeida, o que possibilitou um melhor entendimento do funcionamento do SGIF (Sistema de Gestão de Informação de Incêndios Florestais), nomeadamente os fluxos de informação, critérios de catalogação e estrutura das variáveis. Esta contextualização foi essencial para assegurar a correta interpretação dos dados e a definição de abordagens analíticas adequadas.

Em anexo (Anexo I), encontra-se um documento técnico fornecido pelo ICNF que apresenta a evolução histórica do SGIF, destacando marcos relevantes nas suas transformações ao longo do tempo, impulsionadas por mudanças tecnológicas e institucionais.

1.1.2 Objetivos de Negócio e Critérios de Sucesso

No âmbito deste projeto foram definidos quatro objetivos de negócio, alinhados com os desafios enfrentados pelas entidades responsáveis pela prevenção e combate

aos incêndios florestais em Portugal. Para cada objetivo, foi também estabelecido um critério de sucesso claro, que permitirá avaliar a eficácia das ações e análises desenvolvidas.

O primeiro objetivo de negócio é a redução de reacendimentos, que visa diminuir em 5% a percentagem de ocorrências que originam reacendimentos após terem sido considerados extintos. Este critério de sucesso tem como propósito aumentar a fiabilidade das ações de rescaldo e evitar que incêndios reapareçam, exigindo novos recursos de combate.

O segundo objetivo é a otimização do tempo de primeira Intervenção, com o critério de sucesso definido como a redução em 10% do tempo médio entre o despacho dos meios e a sua chegada ao local do incêndio. Uma resposta mais célere é fundamental para conter o fogo numa fase inicial e evitar a sua propagação.

Segue-se a classificação de reacendimentos, que pretende reduzir em 10% o número de reacendimentos. Este critério de sucesso procura prever se um incêndio pode gerar reacendimentos com base no tipo, duração, local, condições atmosféricas e área ardida.

Por fim, destaca-se a criação de perfis territoriais de risco, com o critério de sucesso centrado na identificação dos 20 concelhos com maior risco, tendo por base dados históricos e variáveis ambientais. Esta classificação permitirá direcionar medidas preventivas para os territórios mais vulneráveis.

De forma a tornar estes objetivos mais claros e tangíveis, apresenta-se de seguida uma tabela de síntese que relaciona cada objetivo com o seu critério de sucesso e impacto esperado:

Tabela 1 - Objetivos de negócio e critérios de sucesso

Objetivo de Negócio	Critério de Sucesso	Impacto Esperado
1. Redução de Incêndios	Diminuir em 5% a percentagem de ocorrências que originam reacendimentos.	Redução do número de incêndios que voltam a deflagrar após terem sido considerados extintos, de forma a aumentar a fiabilidade das ações de resolução.
2. Otimização do tempo da primeira intervenção	Reduzir em 10% o tempo médio entre o início do despacho dos meios de combate e a primeira intervenção.	Maior eficácia operacional, resposta mais rápida.
3. Classificação de reacendimentos	Reduzir em 10% o número de reacendimentos.	Prever o número de reacendimentos e melhor gestão dos recursos naturais.

4. Criação de perfis territoriais de risco	Desenvolver um sistema de classificação territorial que identifique zonas de alto risco com base no histórico e fatores ambientais, identificando os 20 concelhos com maior risco para orientar ações preventivas.	Permitir intervenções preventivas localizadas, como limpezas, restrições ao uso do fogo e reforço da vigilância.

1.2. Descrição do contexto

1.2.1 Inventário de Recursos, Requisitos, Assunções e Restrições

Para a concretização deste projeto, é essencial ter uma visão clara dos recursos disponíveis e dos requisitos necessários à sua execução. A nível de dados, o recurso central é a base de dados disponibilizado pelo ICNF, o qual fornece informação detalhada sobre milhares de ocorrências de incêndios florestais em Portugal. Esta base de dados inclui atributos como data e hora de despacho de meios de combate, localização (freguesia, concelho, distrito), área ardida, causa provável, tipo de vegetação afetada, entre outros. A qualidade e riqueza destes dados permite desenvolver análises com elevado potencial informativo.

Do ponto de vista humano, o projeto é conduzido por três estudantes do Mestrado em Ciência de Dados, com conhecimentos em estatística, programação e visualização de dados. Poderá haver, sempre que necessário, colaboração informal com o ICNF e com o responsável pelo desenvolvimento da estrutura de dados, Rui Almeida, sobretudo para clarificação de conceitos técnicos ou validação de interpretações dos dados.

Relativamente aos recursos tecnológicos, serão utilizados computadores pessoais com capacidade suficiente para processamento e análise dos dados da base de dados fornecida. Ferramentas *open source* como Google Colab, utilizando a linguagem de programação Python (bibliotecas pandas, matplotlib, numpy, entre outras), também serão utilizados, assim como será utilizado o Python para visualização e criação de gráficos. Estas ferramentas são adequadas para análise exploratória, tratamento de dados e comunicação visual dos resultados.

Quanto aos requisitos, destaca-se a necessidade de concluir o projeto dentro do prazo previsto pelo plano curricular da disciplina Data Mining, do Mestrado em Ciência de Dados. Os resultados devem ser claros, fiáveis e compreensíveis, mesmo por pessoas sem formação técnica em ciência de dados. Além disso, é fundamental garantir o cumprimento das normas éticas e legais relativas à utilização dos dados, assegurando que são utilizados apenas para fins académicos e de investigação.

Foram também assumidos alguns pressupostos. Um dos principais é a suposição de que os dados são suficientemente completos e representativos para as análises previstas. Parte-se ainda do princípio de que os registos disponíveis seguem um padrão consistente ao longo do tempo e que as variáveis estão bem definidas e corretamente codificadas.

Contudo, existem algumas restrições que importa reconhecer. O projeto está limitado à informação constante na base de dados fornecida, o que significa que variáveis externas (por exemplo, dados socioeconómicos) não serão consideradas nesta fase. Além disso, o tempo disponível para a realização do projeto é restrito, o que impõe uma seleção criteriosa das análises a realizar, dando prioridade às mais relevantes e com maior potencial de impacto.

1.2.2 Riscos e Contingências

Como em qualquer projeto de análise de dados, existem riscos que podem comprometer a qualidade dos resultados ou a própria execução do trabalho. Um dos principais riscos está relacionado com a qualidade dos dados. Pode haver campos com valores em falta, inconsistências nos dados, ou registos incompletos que dificultam a análise ou limitam a comparabilidade entre anos ou regiões.

Outro risco prende-se com a possível interpretação incorreta de variáveis ou categorias específicas.

Existe ainda o risco de sobrecarga de complexidade – ou seja, tentar explorar demasiados caminhos analíticos no tempo limitado disponível para o projeto.

Para mitigar estes riscos, foram delineadas algumas estratégias de contingência. No caso de dados em falta ou inconsistentes, serão aplicadas técnicas de limpeza, ou, em último caso, poderão ser excluídas determinadas variáveis da análise. Para clarificar conceitos ambíguos, será feita investigação bibliográfica e, se possível, consulta a documentos técnicos ou relatórios do ICNF. Por fim, será mantido um foco constante na relevância prática das análises, selecionando apenas aquelas que ofereçam valor real para os objetivos de negócio definidos.

1.2.3 Terminologia

Dado o caráter multidisciplinar do projeto, é importante clarificar desde o início a terminologia utilizada, tanto no domínio dos incêndios florestais como na área da análise de dados.

Termos do domínio florestal:

- Área ardida: Superfície total afetada por um incêndio, geralmente medida em hectares.
- Fogacho: Incêndio de pequena dimensão, de curta duração e impacto reduzido.
- Rescaldo: O intervalo de tempo entre o momento em que o incêndio é dado como controlado e o momento em que a operação é oficialmente encerrada.
- FWI - *Fire Weather Index*: índice que avalia o perigo de incêndio com base em condições meteorológicas.
- Causa provável: Origem presumível do incêndio (ex: negligência, intencional, natural, desconhecida).

- Reacendimento: Reinício de um foco de incêndio anteriormente extinto, geralmente no mesmo local.

Termos da área de Data Mining:

- *Outlier*: Valor atípico que se desvia significativamente da tendência geral dos dados.
- Clusterização: Técnica que permite agrupar dados com características semelhantes, útil na identificação de padrões.
- Classificação: Técnica preditiva que atribui uma categoria ou classe a uma observação, com base em variáveis preditivas. Ex: prever se um incêndio terá ou não reacendimento.
- Regressão: Técnica que estima um valor numérico contínuo com base em variáveis independentes. Ex: prever o tempo de chegada ao local de um incêndio.
- *Data cleaning*: Etapa em que os dados são preparados, corrigindo erros, valores em falta ou inconsistências.
- Visualização de dados: Representação gráfica de informação, essencial para comunicar resultados de forma intuitiva.

A construção e uso consistente desta terminologia ao longo do trabalho é fundamental para assegurar clareza e coerência, tanto na análise como na interpretação dos resultados.

1.2.4 Custos e Benefícios

Este projeto apresenta um perfil de custos reduzido, sobretudo porque recorre a ferramentas gratuitas e dados públicos. Os custos principais estão associados ao tempo dedicado à sua execução, estimando-se cerca de 200 a 250 horas de trabalho, ECTS definidos pela unidade curricular, que incluem a preparação dos dados, análise, interpretação, redação e apresentação final. Poderá haver também necessidade de consulta de especialistas, mas pontuais.

Em contrapartida, os benefícios são significativos. A nível académico, o projeto permite o desenvolvimento de competências em análise de dados aplicada a problemas reais, com forte componente de impacto social. Para as entidades que gerem ou intervêm no contexto dos incêndios florestais, os resultados poderão oferecer *insights* úteis para melhorar a eficiência das operações, reforçar a prevenção e otimizar a alocação de recursos. A médio prazo, este tipo de análises pode contribuir para decisões mais fundamentadas e eficazes na gestão do risco de incêndio, com potenciais ganhos em termos de segurança, proteção ambiental e poupança económica.

1.3. Identificação dos objetivos de Data Mining

1.3.1 Objetivos de Data Mining

Com base nos objetivos de negócio previamente definidos, esta secção apresenta a identificação e formulação dos objetivos de Data Mining (DM), em conformidade com a metodologia CRISP-DM. Cada objetivo de DM foi desenvolvido com o propósito de transformar os problemas e necessidades operacionais em tarefas analíticas concretas, passíveis de serem tratadas com técnicas de Data mining.

Para cada um dos objetivos de Data Mining, são indicados o tipo de técnica a aplicar (ex: classificação, regressão, clusterização ou associação), bem como os critérios de sucesso mensuráveis que permitirão avaliar a eficácia das soluções propostas.

Tabela 2 - Objetivos de negócio e data mining

Objetivos de Negócio	Objetivos de Data Mining
1. Redução de Incêndios	Identificar as características mais associadas aos incêndios, com base em variáveis como tipo de causa, condições meteorológicas após extinção, tempo de rescaldo, localização, entre outras.
2. Otimização do tempo da primeira Intervenção	Analisar fatores que influenciam o tempo de resposta, como a distância à unidade de bombeiros, hora e data de início do despacho, ou número de ocorrências simultâneas, e prever situações de risco de atrasos operacionais.
3. Classificação de reacendimentos	Prever se um incêndio pode gerar reacendimentos com base no tipo, duração, local, condições atmosféricas e área ardida.
4. Criação de perfis territoriais de risco	Criação de clusters com base no número médio de incêndios, área ardida, perigosidade, tipo de causa e meteorologia média que representem o risco nas diferentes localidades.

1.3.2 Descrição dos objetivos de Data Mining

1. **Identificar as características mais associadas aos incêndios, com base em variáveis como tipo de causa, condições meteorológicas após extinção, tempo de rescaldo, localização, entre outras.**

Descrição: Desenvolver um modelo que permita identificar as variáveis mais associadas aos incêndios, ou seja, as características que podem estar na origem destes acontecimentos. Esta situação representa um risco operacional relevante e um desafio para as equipas de combate, especialmente quando ocorre fora do período ativo da operação. Para tal, serão analisadas variáveis como o tipo e a causa do incêndio, o

tempo de rescaldo, localização geográfica e outras condições meteorológicas. O modelo de associação procurará identificar padrões mais associados aos incêndios que reacendem dos que permanecem controlados, contribuindo assim para decisões mais eficazes de vigilância e prevenção após o término das operações.

Critério de sucesso: Pelo menos 3 regras com Suporte > 3%

Técnica: Associação

2. **Analisar fatores que influenciam o tempo de resposta, como a distância à unidade de bombeiros, hora e data de ínicio do despacho, ou número de ocorrências simultâneas, e prever situações de risco de atrasos operacionais.**

Descrição: Este objetivo visa desenvolver um modelo preditivo capaz de estimar o tempo de resposta operacional dos bombeiros, ou seja, o intervalo entre o despacho de meios de combate e a chegada da primeira equipa ao local do incidente. A análise focar-se-á em identificar os principais fatores que influenciam este tempo, permitindo antecipar situações de risco de atrasos e otimizar a alocação de recursos.

Devem ser consideradas variáveis como a distância às unidades de bombeiros, hora e data do despacho, número de ocorrências simultâneas.

Este modelo permitirá às entidades responsáveis antecipar e mitigar potenciais atrasos na resposta a emergências, com o intuito de melhorar a eficácia e eficiência dos serviços de combate a incêndios.

Critério de sucesso: R2 (coeficiente de determinação) ≥ 70%

Técnica: Regressão

3. **Prever se um incêndio pode gerar reacendimentos com base no tipo, duração, local, condições atmosféricas e área ardida.**

Descrição: Este objetivo foca-se na construção de modelos preditivos para identificar se um incêndio tem potencial de gerar reacendimentos, considerando variáveis como o tipo do incêndio, a sua duração, localização, condições atmosféricas e a área ardida. O modelo permitirá antecipar eventos de reacendimentos, facilitando a alocação eficiente de recursos para evitar que o fogo volte a crescer após o controle inicial.

Critério de sucesso: F1-score ≥ 70%

Técnica: Classificação

4. **Criação de clusters com base no número médio de incêndios, área ardida, perigosidade, tipo de causa e meteorologia média que representem o perfil de risco nas diferentes localidades.**

Descrição: Neste objetivo pretende-se aplicar técnicas de clusterização para agrupar concelhos com perfis de risco semelhantes em relação à ocorrência e impacto de incêndios florestais. A análise será baseada em variáveis como o número médio de incêndios por zona, área ardida, perigosidade e condições meteorológicas.

Através da definição de clusters é possível realizar uma abordagem eficiente na prevenção, no planeamento e na distribuição de recursos.

Critério de sucesso: Índice de Silhouette ≥ 0.6

Técnica: Clusterização

1.4. Produção do plano do projeto

O presente capítulo apresenta o plano que orienta a execução do projeto, desde a definição dos objetivos até à entrega final. São descritas as atividades principais, os recursos técnicos e humanos envolvidos, os riscos identificados e as ações de mitigação previstas. Inclui-se também o cronograma detalhado com entregas faseadas, dependências entre tarefas e os mecanismos de revisão ao longo do projeto.

1.4.1. Objetivos do projeto

O presente projeto tem como principal objetivo aplicar a metodologia CRISP-DM na análise de dados históricos de incêndios florestais em Portugal, com o intuito de identificar padrões relevantes, relações significativas e fatores que influenciam a área ardida. Através de abordagens de Data Mining, pretende-se extrair conhecimento útil para apoiar estratégias de prevenção, resposta e mitigação de incêndios. Os objetivos específicos incluem a análise de variáveis como o tipo e causa do incêndio, reacendimentos, tempo de rescaldo, condições meteorológicas e distribuição geográfica e temporal, contribuindo para uma compreensão mais profunda dos fatores que afetam a propagação do fogo.

1.4.2. Planeamento das atividades

Com base na Tabela 2 do Capítulo 1.3.1, foi possível estruturar um plano de trabalho que abrange todas as atividades a serem realizadas ao longo do projeto. Este plano contém todo o detalhe desde a identificação e exploração da base de dados até à entrega final dos resultados. Trata-se, no entanto, de uma estimativa inicial, sujeita a ajustes ao longo do tempo, de acordo com a evolução do projeto e os fatores de risco identificados no Capítulo 1.4.4. O seu desenvolvimento teve em consideração as duas fases estipuladas para entrega, assegurando uma organização coerente e alinhada com os objetivos definidos.

Tabela 3 - Planeamento das Atividades com base na Metodologia CRISP-SM

Nº	Fase da Metodologia CRISP-DM	Atividades e Entregas	Dependências	Datas
Fase I				
	Identificação da Base de dados	Pesquisa, seleção da base de dados a utilizar no projeto.	Nenhuma (atividade inicial do projeto).	21/03 a 08/04
1	**Compreensão do Negócio**	Determinar objetivos de negócio; Avaliar o cenário;	Após a identificação da base de dados.	09/04 a 14/04

		Objetivos de Data Mining; Plano de Projeto.		
2	**Compreensão dos Dados**	Recolha dos Dados Iniciais; Descrição dos Dados; Exploração dos Dados; Qualidade dos Dados.	Após a compreensão do negócio.	15/04 a 05/05
Entrega da Parte I		Submissão do relatório com fases 1 e 2 (relatório + scripts).	**Conclusão das fases 1 e 2.**	**05/05**
Fase II				
3	**Preparação dos Dados**	Seleção dos Dados; Limpeza dos Dados; Derivar Novos Dados; Integrar Dados; Formatar Dados; Criação do Dataset.	Após compreensão dos dados (Fase 2).	06/05 a 13/05
4	**Modelação**	Escolha de Técnicas de Modelização; Planificação de Testes; Construção do Modelo; Avaliar o Modelo.	Após preparação dos dados (Fase 3).	14/05 a 14/06
5	**Avaliação**	Avaliar os Resultados; Revisão do Processo; Determinar Ações Futuras.	Inicia após primeiros modelos gerados (decorrer paralelamente à Fase 4).	07/06 a 20/06

6	Colocação em Produção	Planeamento da colocação em produção; Planeamento da monitorização e manutenção; Elaboração do relatório final; Revisão do projeto.	Inicia com os modelos avaliados (em paralelo com finalização da Fase 5).	17/06 a 27/06
Entrega Final + Prova Oral		Submissão do projeto completo e apresentação oral.	**Conclusão das fases de 3 a 6.**	**27/06**

O diagrama de Gantt apresenta uma representação visual do planeamento já detalhado na tabela anterior. Este recurso permite visualizar a distribuição temporal das atividades, bem como a sua sequência, duração e sobreposição ao longo do semestre.

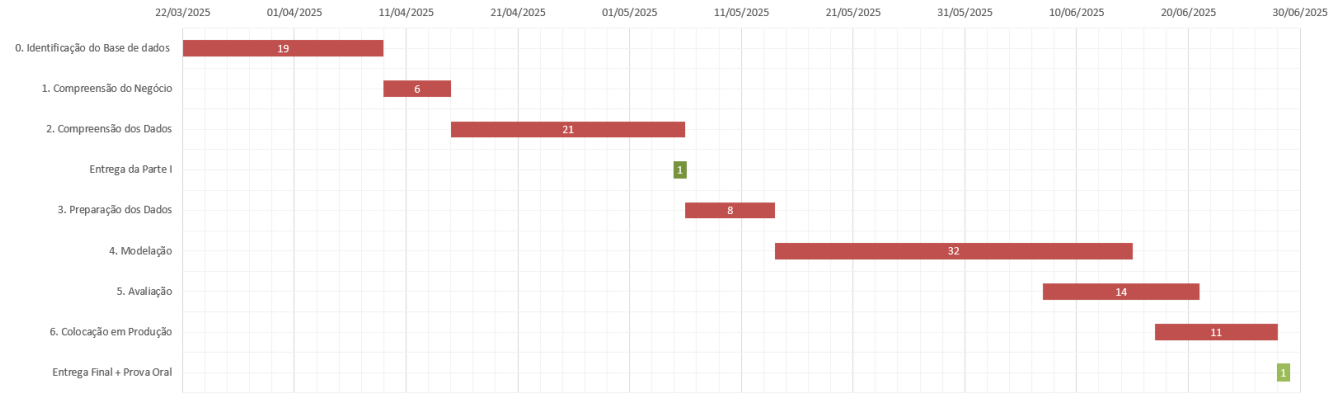


Figura 1 - Diagrama de Gantt com o Planeamento das Atividades do Projeto

1.4.3 Recursos do Projeto

1.4.3.1 *Stakeholders*

No contexto do presente projeto de Data Mining, é essencial reconhecer os principais *stakeholders* envolvidos, pois cada um desempenha um papel específico e relevante no desenvolvimento e validação do projeto. A tabela seguinte resume os principais grupos intervenientes, os seus papéis e os respetivos objetivos no âmbito do projeto:

Tabela 4 - Tabela dos Stakeholders

Stakeholder	**Papel**	**Objetivo no Projeto**

Grupo de estudantes: Eliana Oliveira, Jéssica Grácio e Maria Fialho	Realizam todas as etapas do projeto.	Aplicar conhecimentos da unidade curricular e desenvolver uma solução relevante e prática.
Professores Responsáveis pela UC de Data Mining - Prof. Maria Piedade e Prof. Ricardo Malheiro	Acompanham, orientam e validam as decisões e a execução do projeto.	Garantir rigor metodológico, qualidade académica e adequação às exigências do plano curricular.
Instituto da Conservação da Natureza e das Florestas (ICNF)	Entidade pública responsável pela disponibilização da base de dados utilizada no projeto.	Fornecer dados relevantes e oficiais sobre incêndios florestais, fundamentais para a realização das análises e extração de conhecimento.
Comando Nacional da Força de Sapadores Bombeiros Florestais (CNFSBF) – Comandante Rui Almeida	Colaborador institucional com experiência no terreno e envolvimento direto na temática dos incêndios florestais.	Potencial utilizador dos *insights* produzidos, com interesse na aplicação de resultados para apoiar a gestão de incêndios.

1.4.3.2 Recursos Técnicos e Humanos

A tabela seguinte apresenta uma visão clara dos recursos humanos e técnicos que serão utilizados ao longo das diversas fases do projeto. Cada recurso é descrito pela sua atividade alocada e duração prevista.

Tabela 5 - Tabela de Recursos Técnicos e Humanos

Recurso (Humano/Técnico)	Atividade alocada	Duração (dias)
Grupo de estudantes: **Eliana Oliveira, Jéssica Grácio e Maria Fialho**	Todas as fases do projeto	98

Python (Pandas, seaborn, matplotlib, numpy)	Desenvolvimento e execução do código	66
Google Colab	Desenvolvimento e execução do código em ambiente de Notebook	66
Google Drive	Para armazenamento e backup dos dados e implementações dos Notebooks.	66
Microsoft Teams	Comunicação diária e reuniões de progresso ao longo do projeto e também repositório de documentos.	98

A duração atribuída a cada recurso corresponde ao período em que este é utilizado de forma efetiva no projeto. O grupo de estudantes e a plataforma Microsoft Teams permanecem ativos ao longo de praticamente todo o cronograma (98 dias), dado o seu envolvimento direto nas tarefas e na comunicação regular. Já ferramentas como Python, Google Colab e Google Drive são utilizadas nas fases técnicas do projeto, como a exploração dos dados, preparação dos mesmos, a modelação e a avaliação dos modelos, o que justifica a duração aproximada de 66 dias. Estes valores mantêm coerência com o planeamento das atividades definido anteriormente.

1.4.4 Identificação e Mitigação de Riscos

Durante a definição do plano de projeto, foram identificados alguns riscos que podem impactar o cumprimento dos prazos ou a qualidade das entregas. Os principais riscos incluem a sobrecarga de trabalho académico, devido à coincidência de entregas de outras unidades curriculares do mestrado; as limitações técnicas do grupo de estudantes, tendo em conta a complexidade de certas abordagens analíticas e a aprendizagem em curso; e as dificuldades de conciliação de horários entre os elementos do grupo, dada a existência de outras responsabilidades profissionais.

Para mitigar estes riscos, foram delineadas diversas ações. Serão realizadas reuniões de esclarecimento com os docentes e promovido o contacto com o Rui Almeida, com o objetivo de esclarecer dúvidas sobre os dados e garantir maior precisão na análise. Pretende-se também utilizar ferramentas já conhecidas pelo grupo de estudantes, como Python, Google Colab, entre outras, para reduzir a curva de aprendizagem. O planeamento antecipado das tarefas e entregas parciais será adotado como estratégia para distribuir melhor o esforço ao longo do tempo, evitando a acumulação de tarefas em períodos críticos. Além disso, será feita uma coordenação periódica através do Microsoft Teams, garantindo comunicação constante e visibilidade sobre o progresso das tarefas.

1.4.5 Plano de Revisão do Projeto

Foram definidos dois momentos-chave de revisão: o primeiro após a entrega da Parte I (05/05/2025), com o objetivo de refletir sobre os métodos aplicados e ajustar o planeamento e o segundo (Parte II) antes da entrega final (28/06/2025), para garantir a coerência do projeto e preparar a apresentação oral. Estes pontos estão integrados no cronograma para facilitar a sua articulação com o restante plano de atividades.

Fase 2: Compreensão dos Dados

Após a definição dos objetivos de negócio e de Data Mining, esta fase visa garantir uma compreensão dos dados disponíveis. As etapas que incluem a recolha, descrição detalhada, análise exploratória e avaliação da qualidade dos dados. Desta forma, foi possível identificar padrões relevantes, detetar inconsistências ou anomalias, compreender a distribuição das variáveis e avaliar o grau de completude e fiabilidade da base de dados.

2.1. Recolha dos dados

Na presente etapa, foi realizada a recolha da base de dados necessária para dar resposta aos objetivos definidos na fase anterior. Os dados utilizados foram obtidos a partir do portal oficial dos incêndios florestais do ICNF (Instituto da Conservação da Natureza e das Florestas). Abaixo, encontra-se uma tabela com mais informações sobre a base de dados em questão.

Nome do Ficheiro	"Data2001_now.csv"
Fonte dos Dados	https://www.icnf.pt/florestas/gfr/gfrgestaoinformacao/estatisticas
Formato	CSV
Período Abrangido	2001 até 2025 (momento atual)
Número total de Registos	486715
Número de Atributos	99
Acesso	Público

Tabela 6 - Informações sobre a base de dados utilizada

A escolha recaiu sobre o ficheiro "Data2001_now.csv" por ser o mais atualizado e abrangente entre os disponibilizados, incluindo mais de 486 mil registos de incidentes associados a incêndios em território nacional, ao longo de um período superior a duas décadas.

Contudo, a base de dados inicialmente recolhida no início do projeto encontra-se, atualmente, em fase de manutenção pela entidade responsável (ICNF). Após contacto com o ICNF, foi-nos comunicado que poderíamos continuar a utilizar a versão previamente obtida para efeitos de tratamento de dados e desenvolvimento do projeto, desde que fosse feita referência oficial à fonte de dados indicada na Tabela 6. Desta forma, apesar da base de dados se encontrar em manutenção, mantém-se a utilização da versão original para assegurar a continuidade do trabalho e a consistência das análises já realizadas, sem comprometer a validade dos resultados.

Em termos de integridade e estrutura, verificou-se que:

- ✓ O ficheiro encontrava-se em bom estado e não apresentava erros de leitura.
- ✓ Os dados estavam devidamente estruturados, em formato CSV, com cabeçalhos e valores separados por vírgulas.

- ✓ A cobertura temporal é extensa e adequada para os objetivos propostos.

No entanto, foi identificada uma limitação relacionada com a documentação da base de dados, pois o portal ICNF não fornece um dicionário oficial de dados com a descrição dos atributos disponíveis. Dado isto, inicialmente, a compreensão das variáveis foi feita com base em análises dos próprios valores, nos nomes das colunas e em pesquisas externas.

Após contacto direto com o ICNF, foi possível solicitar a descrição formal dos atributos da base de dados. Este pedido foi aceite e pelo que, foi-nos fornecida alguma documentação relativa às descrições dos atributos, e que permitiram validar e complementar a interpretação já efetuada durante a análise preliminar.

Além disso, durante a preparação dos dados para análise e modelação, foi realizada uma avaliação às 99 variáveis presentes no ficheiro. A partir dessa análise, e tendo como base os objetivos de negócio e os objetivos de Data Mining definidos na Fase 1, foi feita uma pré-seleção dessas variáveis, com base nos seguintes critérios:

- Relevância direta para os objetivos do projeto (ex.: previsão de tempos, causas de incêndios, impacto).
- Qualidade e completude dos dados.
- Eliminação de redundâncias (ex.: colunas duplicadas ou derivadas).
- Clareza na interpretação do conteúdo.
- Exclusão de variáveis cuja utilização está dependente de softwares específicos não disponíveis no âmbito do projeto (ex.: variáveis relacionadas com o FARSITE, cuja análise e modelação requerem esse sistema).

Embora nem todas as variáveis avancem para a Análise Exploratória de Dados (AED) e fases seguintes, todos os atributos foram mantidos na base de dados e devido também à dimensão da base de dados e do número de atributos da mesma, apenas as variáveis que obtiveram resultados mais relevantes foram mencionadas no relatório, sendo que as restantes ficaram documentadas nos *notebooks*.

2.2. Descrição dos dados

Nesta secção apresenta-se a descrição detalhada das variáveis constantes na base de dados utilizada no projeto. A informação relativa aos 99 atributos da mesma encontra-se organizada em formato tabular e inclui os seguintes elementos para cada variável:

- Nome do atributo;
- Tipo da variável;
- Descrição;
- Relevância para o projeto, com base nos critérios definidos no capítulo 2.1;
- Motivo para exclusão, nos casos em que a variável não foi selecionada para análise posterior.

A análise destas variáveis teve como base os critérios mencionados anteriormente, como a clareza da descrição, a relevância analítica, a ausência de redundância, e a dependência de ferramentas ou códigos externos (ex: FARSITE, códigos técnicos sem documentação pública). Esta análise permitiu realizar uma seleção fundamentada das variáveis mais adequadas para a fase de análise exploratória e modelação.

Devido ao enorme volume de atributos da base de dados, a tabela em questão encontra-se no Anexo II. Abaixo, na tabela 7, pode ser visualizado o número de variáveis por tipo presentes na base de dados original.

Tabela 7 - Tipo de Variável da Base de Dados

Tipo de Variável	Número de Variáveis
Numérica	68
Nominal (Categórica)	23
Binária	3
Data/Hora	6
Identificador	1
Total	**99**

2.3 Análise Exploratória dos dados

A análise exploratória de dados corresponde à etapa 3 da fase de Preparação dos Dados no ciclo de vida de um projeto de Data Mining, conforme a metodologia CRISP-DM. Esta fase visa obter uma compreensão inicial e um pouco mais aprofundada dos dados, permitindo identificar padrões, características relevantes, inconsistências e possíveis relações entre as diferentes variáveis presentes.

O principal objetivo desta etapa é garantir que a informação contida nos dados seja bem compreendida, de forma a fundamentar adequadamente as decisões a tomar nas fases seguintes de modelação.

2.3.1 Análise Exploratória das Variáveis Quantitativas

Começaram por serem analisadas as variáveis numéricas contínuas onde foram aplicadas as seguintes análises.

1. Estatísticas descritivas básicas: Cálculo da média, mediana, desvio padrão, mínimos e máximos, complementados por histogramas e boxplots;
2. Correlação entre variáveis: Avaliação das correlações entre as variáveis numéricas através da matriz de correlação de *Pearson* e *heatmap*.
3. Deteção de *outliers*: Identificação de valores extremos (outliers).

2.3.1.1 Estatísticas descritivas das variáveis numéricas contínuas

Para esta análise recorreu-se ao tratamento estatístico das medidas de tendência central (média e mediana), bem como de dispersão (desvio padrão, valores mínimos e máximos). O objetivo foi avaliar a amplitude dos dados, verificar a existência de valores distantes entre si e identificar possíveis *outliers*. Abaixo apresenta-se a tabela com as estatísticas calculadas.

Tabela 8 - Estatísticas descritivas das variáveis numéricas contínuas

Variável	Média	Mediana	Desvio padrão	Mínimo	Máximo
DuraçãoHoras (horas)	1.91	1.02	8.14	0.00	2474.32
Tempo1Intervenção (min)	12.63	12.00	7.83	0.00	120.00
Tempo Resolução (min)	74.94	90.00	71.88	0.00	9829.00
Tempo Rescaldo (min)	71.33	90.00	74.40	0.00	17121.00
Área Povoamento (ha)	3.33	0.00	148.51	0.00	47215.43
Área Mato (ha)	2.57	0.01	67.98	0.00	15647.00

Área Agrícola (ha)	0.47	0.00	23.10	0.00	5675.65
Área Total (ha)	6.37	0.05	208.11	0.00	53618.81
Altitude Média (m)	256.41	189.55	227.80	-64.21	1848.80
Declive Médio (%)	13.74	11.86	9.14	0.00	82.34
Horas Exposição (h)	12.19	12.55	1.70	0.00	15.00
Rugosidade	9.36	8.10	6.38	0.00	78.28
Temperatura (°C)	21.29	21.28	6.23	-4.67	46.37
Humidade Relativa (%)	56.12	55.16	18.84	0.00	100.14
Vento Intensidade (km/h)	11.59	10.54	6.41	0.00	53.76
Precipitação (mm)	0.01	0.00	0.12	-0.26	8.06
Perigosidade (de -1 a 5)	2.44	2.28	1.27	-1.00	5.00

Como se pode verificar na tabela acima, as variáveis associadas aos tempos ("DuracaoHoras", "Tempo1Intervencao", "TempoResolucao" e "TempoRescaldo") mostram que, de forma geral, os incêndios tendem a ser resolvidos rapidamente, com valores medianos não muito altos. No entanto, registam-se alguns valores máximos bastante elevados que indicam a ocorrência de casos excecionais com durações muito prolongadas. Estes valores extremos sugerem uma elevada dispersão e a possibilidade de *outliers*.

Por outro lado, relativamente às áreas ardidas, na maioria dos casos, as áreas afetadas são reduzidas, com medianas próximas de zero ou muito baixas. Contudo, tal como nos tempos, também aqui surgem alguns incêndios de grande dimensão, evidenciados pelos valores máximos muito elevados, o que contribui para um desvio padrão significativo e possivelmente uma distribuição assimétrica.

As variáveis "AltitudeMedia", "DecliveMedio", "Rugosidade" e "HorasExposicaoMedia" indicam que os incêndios ocorreram em zonas com relevo e exposição solar relativamente variados, mas sem extremos marcados. A "Rugosidade" evidencia alguma diversidade na complexidade do terreno, enquanto as "HorasExposicaoMedia" mostram uma exposição solar relativamente homogénea.

Finalmente, as variáveis meteorológicas "Temperatura", "HumidadeRelativa", "VentoIntensidade", e "Precepitacao" e a variável de risco "Perigosidade", mostram variações que são coerentes com o contexto de incêndios florestais. As temperaturas médias são elevadas, a "HumidadeRelativa" tende a ser reduzida, e a "VentoIntensidade" apresenta valores moderados a altos em alguns casos, todos fatores que favorecem a propagação dos fogos. A "Precepitacao" é muito reduzida, como

esperado e a "Perigosidade" apresenta alguma variabilidade, mas com valores medianos em torno de níveis moderados.

Para complementar esta análise, apresenta-se de seguida os histogramas, que permitem observar a distribuição dos dados e perceber se os valores estão concentrados em torno de certos intervalos ou se existe uma maior dispersão (tabela 8).

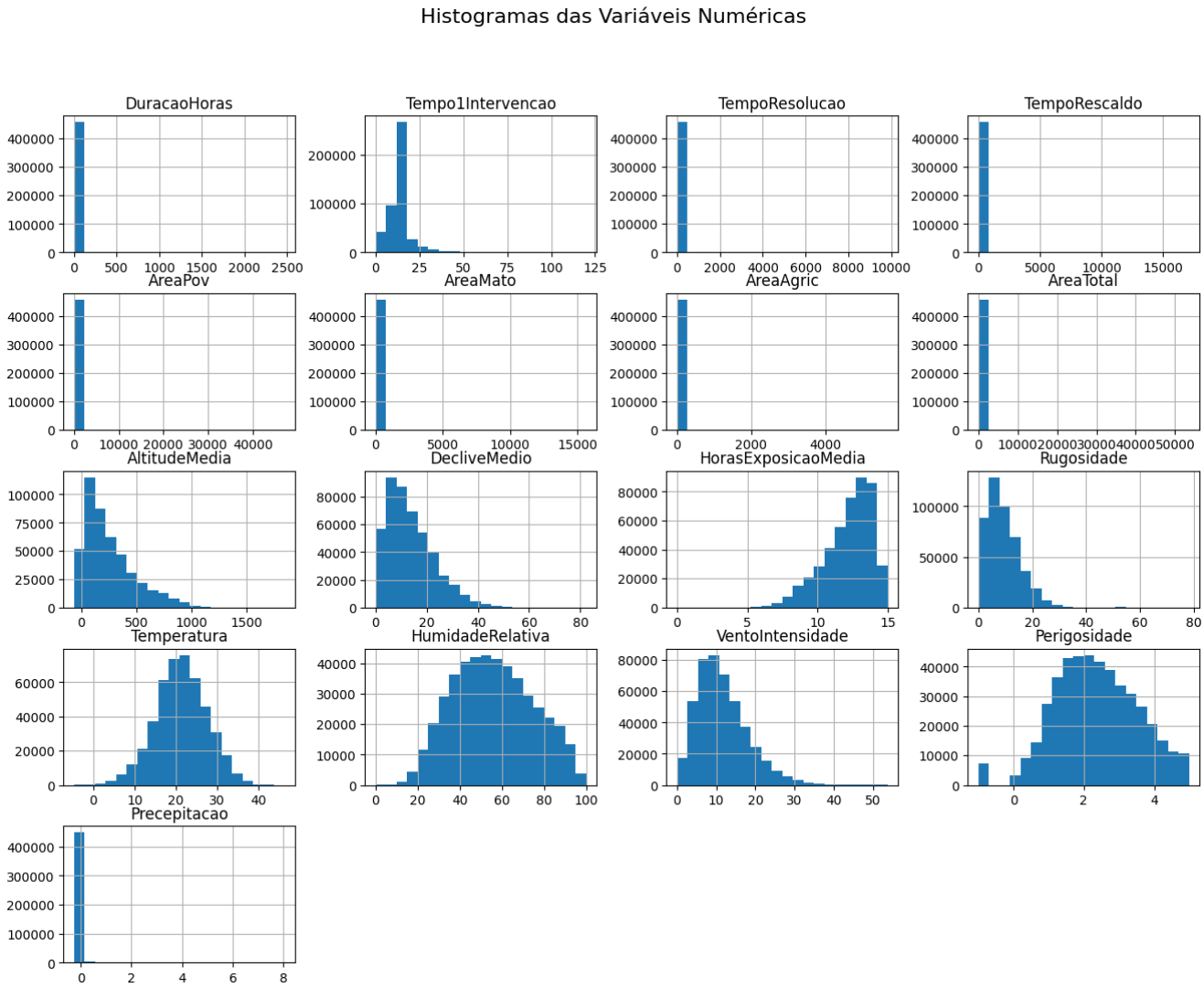


Figura 2 - Histogramas das Variáveis Numéricas

Tal como indicado pelos valores das estatísticas descritivas, os histogramas vieram provar que as variáveis relativas aos tempos (“Tempo1Intervencao”, “TempoResolucao”, “TempoRescaldo”), às áreas (“AreaPov”, “AreaMato”, “AreaAgric”, “AreaTotal”) e também as variáveis “DuracaoHoras” e “Precipitacao”, apresentam distribuições bastante assimétricas, indicando que a maior parte dos registos está concentrada em valores baixos, com poucos eventos que assumem valores muito elevados.

Por outro lado, variáveis como “AltitudeMedia”, “DecliveMedio”, “Rugosidade”, “HorasExposicaoMedia” e “VentoIntensidade” apresentaram distribuições mais amplas, com dispersão considerável dos dados, embora ainda com alguma assimetria.

As variáveis com distribuição mais próxima da simetria foram “Temperatura”, “HumidadeRelativa” e “Perigosidade”. Estas sugerem uma dispersão mais equilibrada em torno da média, sem caudas muito alongadas e com uma variação mais estável entre os eventos analisados.

Adicionalmente, também o boxplot das variáveis em causa abaixo (Figura 3) permite confirmar a presença de valores extremos (*outliers*) e validar as tendências observadas na tabela de estatísticas e nos histogramas das variáveis.

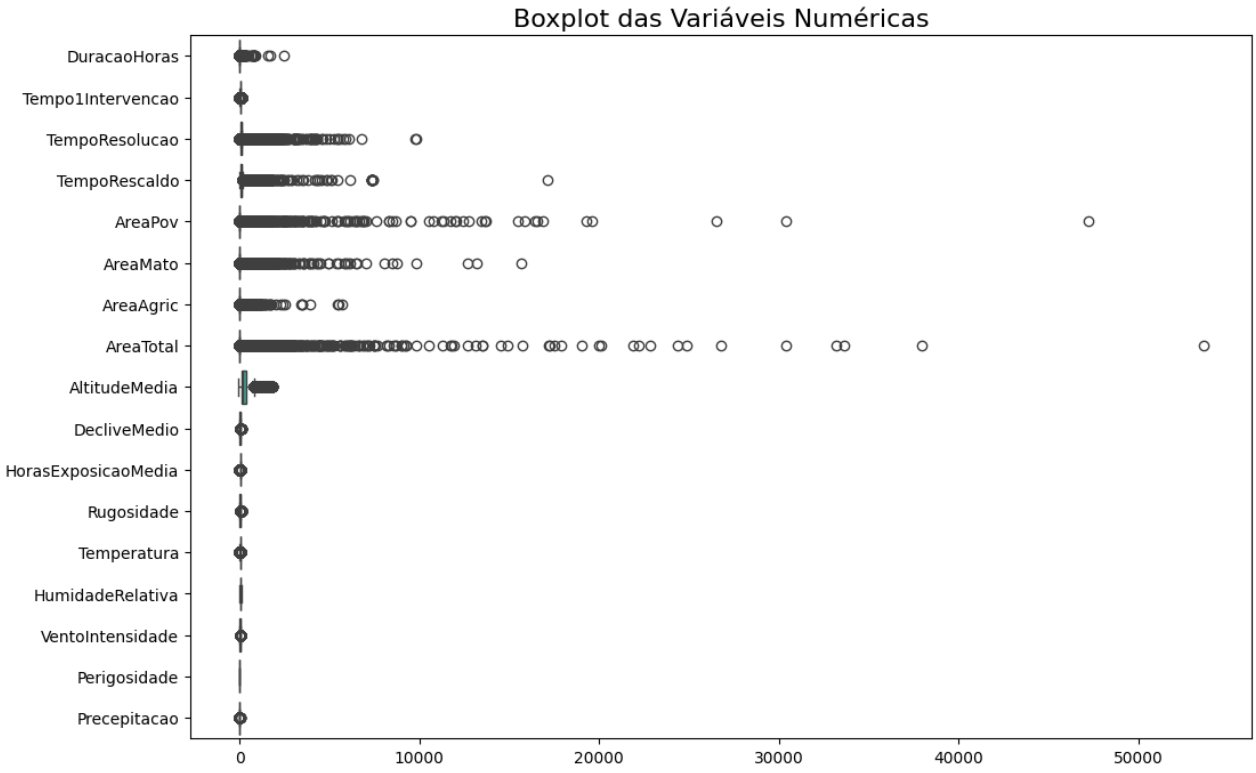


Figura 3 - Boxplot das Variáveis Numéricas

A análise do boxplot revela uma grande quantidade de valores *outliers* nas variáveis relativas às áreas e aos tempos. Esses *outliers* indicam ocorrências com características extremas, que podem distorcer medidas como a média e afetar modelos sensíveis a essas variações. Já variáveis como a Temperatura, HumidadeRelativa e Perigosidade demonstraram menor presença de *outliers*.

Contudo, a diferença de escalas entre variáveis também é evidente no boxplot e destaca a importância de aplicar normalização nas variáveis antes de as usar para modelação, principalmente em algoritmos que dependem de distâncias ou magnitudes comparáveis.

2.3.1.2 Correlação entre variáveis.

Nesta etapa da análise exploratória, foi avaliada a correlação entre as variáveis numéricas contínuas, com o objetivo de identificar relações lineares significativas que possam ser relevantes para a modelação e interpretação futura.

Para isso, utilizou-se a matriz de correlação de *Pearson*, representada visualmente através de um *heatmap,* como se mostra na figura abaixo.

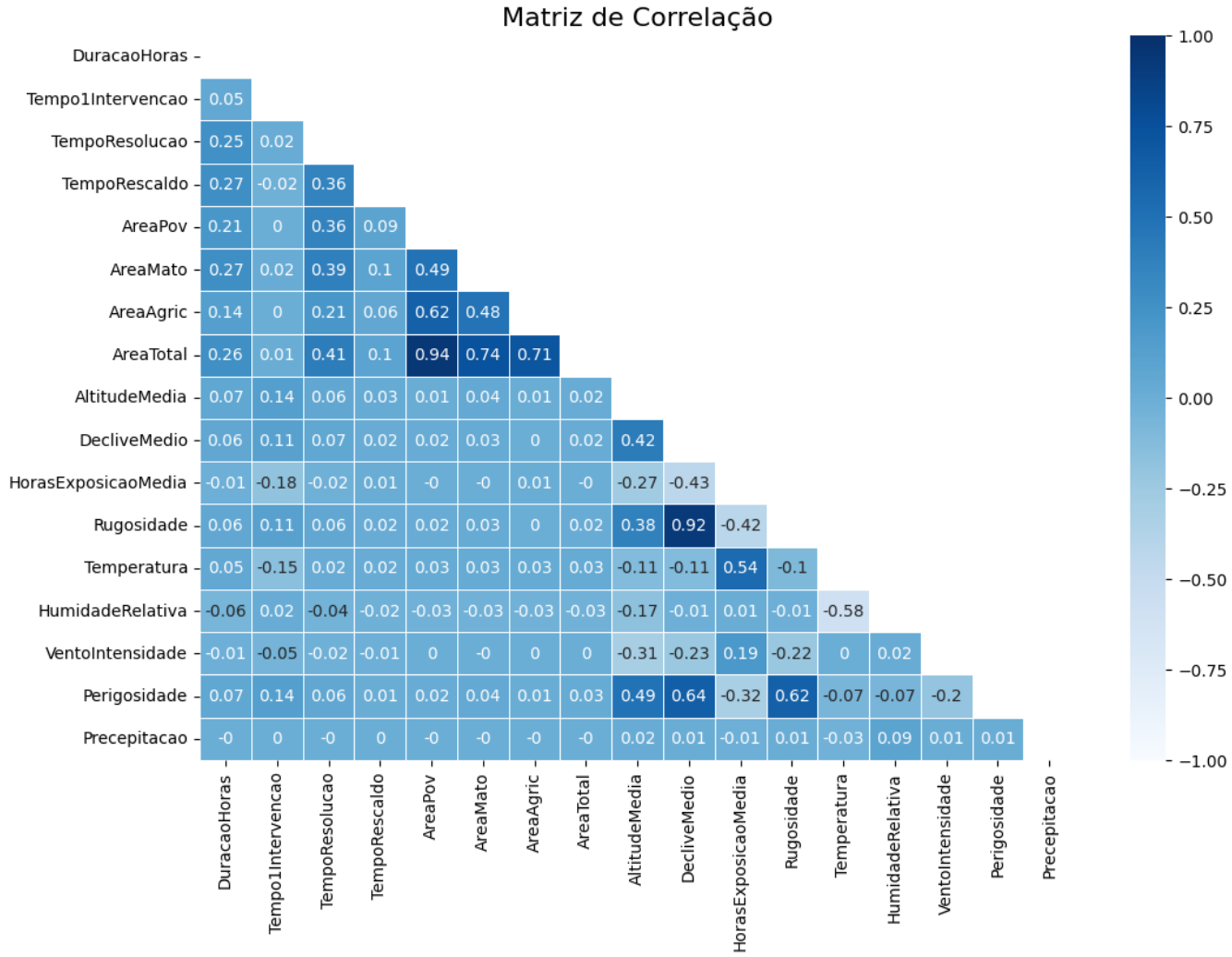


Figura 4 - Matriz de Correlação

Ao analisar a matriz de correlação das variáveis numéricas, verifica-se o seguinte:

Correlações inexistente (r = 0) **ou fraca** (|r ≤ 0.3|)

- A maioria das correlações entre variáveis situa-se entre -0.3 e 0.3, o que indica uma correlação inexistente ou fraca.

Correlação moderada (|0.3 < r ≤ 0.75|)

- A "AreaAgric" com "AreaMato" (r = 0.48) e "AreaPov" com "AreaMato" (r = 0.36) indicam que diferentes tipos de ocupação do solo podem ser afetados em simultâneo.

- A “Perigosidade” com “DecliveMedio” (r = 0.64) e a “Perigosidade” com “Rugosidade (r = 0.62) indicam que zonas com maior de declive e zonas com mais rugosidade tendem a apresentar maior risco de perigo.
- “AltitudeMedia” com “DecliveMedio” (r = 0.42) indicam que altitudes mais elevadas se associam a maiores declives.
- A “Temperatura” com “HumidadeRelativa” têm uma correlação de r = -0.58, o que indica que se trata de uma relação inversa (quando uma aumenta a outra tende a diminuir). Como esperado, temperaturas elevadas tendem a coincidir com menor humidade, um fator crítico no risco de incêndio.
- As “HorasExposicaoMedia” com a “Rugosidade” também apresentam uma relação inversa (r = -0.42) o que indica que terrenos mais irregulares tendem a receber menos exposição solar média.

Correlação forte (|r > 0.75|)

- A “AreaTotal” com a “AreaMato” (r = 0.94): demonstra que a componente de mato tem grande peso na área total queimada, o que sugere que a maioria dos incêndios ocorra nestas áreas.
- A “Rugosidade” com o “DecliveMedio” (r = 0.92) mostra também que terrenos mais inclinados são também mais irregulares.
- A “AreaTotal” com “AreaAgric” (r = 0.74): reforça a ideia de que áreas agrícolas também contribuem significativamente para a área total afetada pela ocorrência de incêndios.

2.3.2 Análise Exploratória das Variáveis Qualitativas

Para sistematizar o processo e garantir a uniformidade da análise, foram desenvolvidas várias funções em Python, aplicada às seguintes variáveis: Tipo, ClasseArea, TipoCausa, FonteAlerta, GrupoCausa, Distrito, Conselho, Freguesia, Nut2, Fogacho, Agricola e Reacendimentos.

Para as variáveis qualitativas mais relevantes, elaboraram-se tabelas de frequências absolutas e relativas, bem como gráficos de barras ou gráficos de pizza consoante o número de variáveis, com o objetivo de identificar categorias dominantes e categorias raras. A análise respeita os princípios de visualização de dados abordados nas sessões práticas.

2.3.2.1 Tipo

Tipo	Frequência Absoluta	Frequência Relativa (%)
Agrícola	85676	17.6 %
Florestal	401026	82.39 %
Queima	13	0.0 %

Figura 5 - Tabela de frequência absoluta e frequência relativa para a coluna Tipo

A variável *Tipo* apresenta três categorias distintas: "Agrícola", "Florestal" e "Queima". A análise das frequências absoluta e relativa mostra que a maioria dos registos corresponde ao tipo "Florestal", com 401.026 ocorrências (82,39%), seguido do tipo "Agrícola", com 85.676 ocorrências (17,60%). A categoria "Queima" é residual, com apenas 13 ocorrências, o que representa uma percentagem praticamente nula (0,00%). Estes valores mostram uma forte predominância de incêndios florestais no conjunto de dados analisado.

Figura 6 - Gráfico de Pizza para a coluna Tipo

O gráfico de pizza complementa esta análise ao mostrar visualmente a elevada concentração de ocorrências no tipo "Florestal", com a categoria "Agrícola" com menor expressão e a "Queima" praticamente inexistente. A escolha pelo gráfico de pizza deve-se à existência de poucas categorias e permite observar de forma clara e intuitiva a distribuição proporcional entre elas.

2.3.2.2 ClasseArea

ClasseArea	Frequência Absoluta	Frequência Relativa (%)
[Area+1000]	473	0.1 %
[Area-1]	415310	85.33 %
[Area-50-100]	2181	0.45 %
[Area1-10]	56796	11.67 %
[Area10-50]	9156	1.88 %
[Area100-500]	2341	0.48 %
[Area500-1000]	458	0.09 %

Figura 7 - Tabela de frequência absoluta e frequência relativa para a coluna ClasseArea

A variável *ClasseArea* revela a distribuição das áreas afetadas pelos incêndios em diferentes intervalos de dimensão. A análise das frequências absoluta e relativa mostra que a grande maioria dos registos corresponde a áreas inferiores a 1 hectare, com 415.310 ocorrências (85,33%). Em contraste, apenas 473 registos (0,10%) dizem

respeito a áreas superiores a 1000 hectares. Estes valores destacam uma predominância clara de incêndios de pequena dimensão no conjunto de dados analisado.

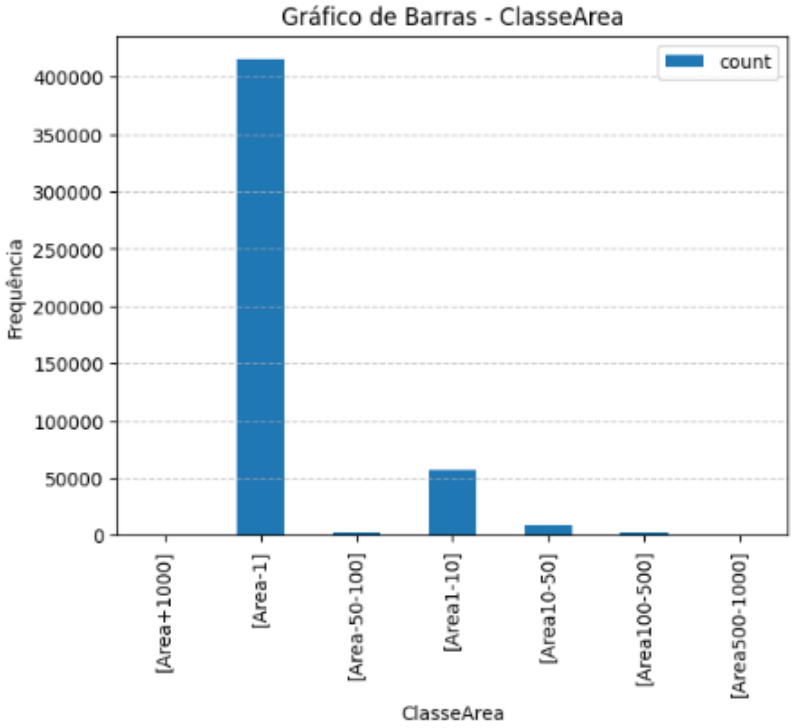


Figura 8 - Gráfico de Barras para a coluna ClasseArea

O gráfico de barras ilustra visualmente esta concentração de incêndios em áreas muito pequenas, demonstrando que a esmagadora maioria dos eventos analisados foi de reduzida extensão territorial. A diferença de proporções entre as classes é muito acentuada, tornando claro o predomínio de incêndios de pequena dimensão no conjunto de dados.

2.3.2.3 TipoCausa

TipoCausa	Frequência Absoluta	Frequência Relativa (%)
Desconhecida	82019	34.26 %
Intencional	46069	19.24 %
Natural	1970	0.82 %
Negligente	80657	33.69 %
Reacendimento	28714	11.99 %

Figura 9 - Tabela de frequência absoluta e frequência relativa para a coluna TipoCausa

A variável *TipoCausa* mostra que a maioria dos incêndios tem causas Desconhecidas (82.019 ocorrências, 34,26%) e Negligentes (80.657 ocorrências, 33,69%). As causas Intencionais aparecem em menor escala (46.069 ocorrências, 19,24%), seguidas pelos Reacendimentos (28.714 ocorrências, 11,99%). As causas Naturais são as menos frequentes, com apenas 1.970 ocorrências (0,82%). Estes dados destacam a dificuldade em identificar a origem dos incêndios e evidenciam a relevância das ações negligentes como um fator de risco significativo.

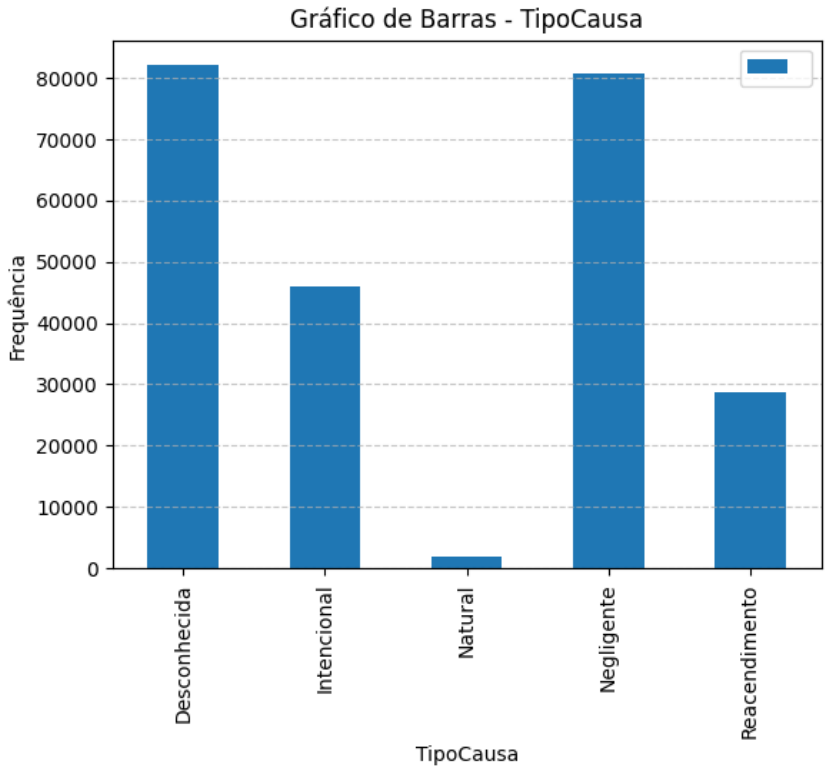


Figura 10 - Gráfico de Barras para a coluna TipoCausa

O Gráfico de Barras (Figura 10) complementa a análise anterior, reforçando a perceção da predominância das causas Desconhecida e Negligente no conjunto de dados, e mostrando a pouca representatividade das causas Naturais.

2.3.2.4 Fonte Alerta

FonteAlerta	Frequência Absoluta	Frequência Relativa (%)
112	32184	6.63 %
117	51802	10.67 %
Autarquias	4	0.0 %
BAV	3	0.0 %
BRISA	1	0.0 %
BT	4	0.0 %
Bombeiros	210	0.04 %
CB	1018	0.21 %
CCO	40701	8.38 %
CDOS	581	0.12 %
CM	24	0.0 %
CNGF	241	0.05 %
CODU	4	0.0 %
Concessionários rodoviários	1	0.0 %
DRA	505	0.1 %
GNR	263	0.05 %
Outros	69617	14.34 %
P.Mun.	88	0.02 %
PJ	1	0.0 %
PNM	1	0.0 %
PNPG	17	0.0 %
PSP	50	0.01 %
PV	36475	7.51 %
Part	1	0.0 %
Particular	12901	2.66 %
Populares	235420	48.49 %
RNPV	92	0.02 %
S.M.Prot. Civil	36	0.01 %
Sapadores	3056	0.63 %
Sapadores Flo	17	0.0 %
Sapadores florestais	9	0.0 %
VMT/Agris	17	0.0 %
VT	84	0.02 %
Vig. Aérea	17	0.0 %
Vig. Movel Terr	39	0.01 %
pv	1	0.0 %

Figura 11 - Tabela de frequência absoluta e frequência relativa para a coluna Fonte Alerta

A variável *FonteAlerta* apresenta uma elevada diversidade de categorias, com destaque para “Populares”, responsável por 235.420 ocorrências (48,43% do total), seguida pela linha de emergência “117”, com 51.802 ocorrências (10,67%) e “Outros”, com 69.617 (14,34%). Outras fontes relevantes incluem “112” (6,63%), “CCO” (8,38%) e “PV” (7,51%). As restantes categorias apresentam percentagens residuais. Estes dados sugerem que a maioria dos alertas de incêndios provém de cidadãos e linhas telefónicas de emergência, evidenciando a importância da participação popular e da acessibilidade dos meios de reporte na deteção precoce de incêndios.

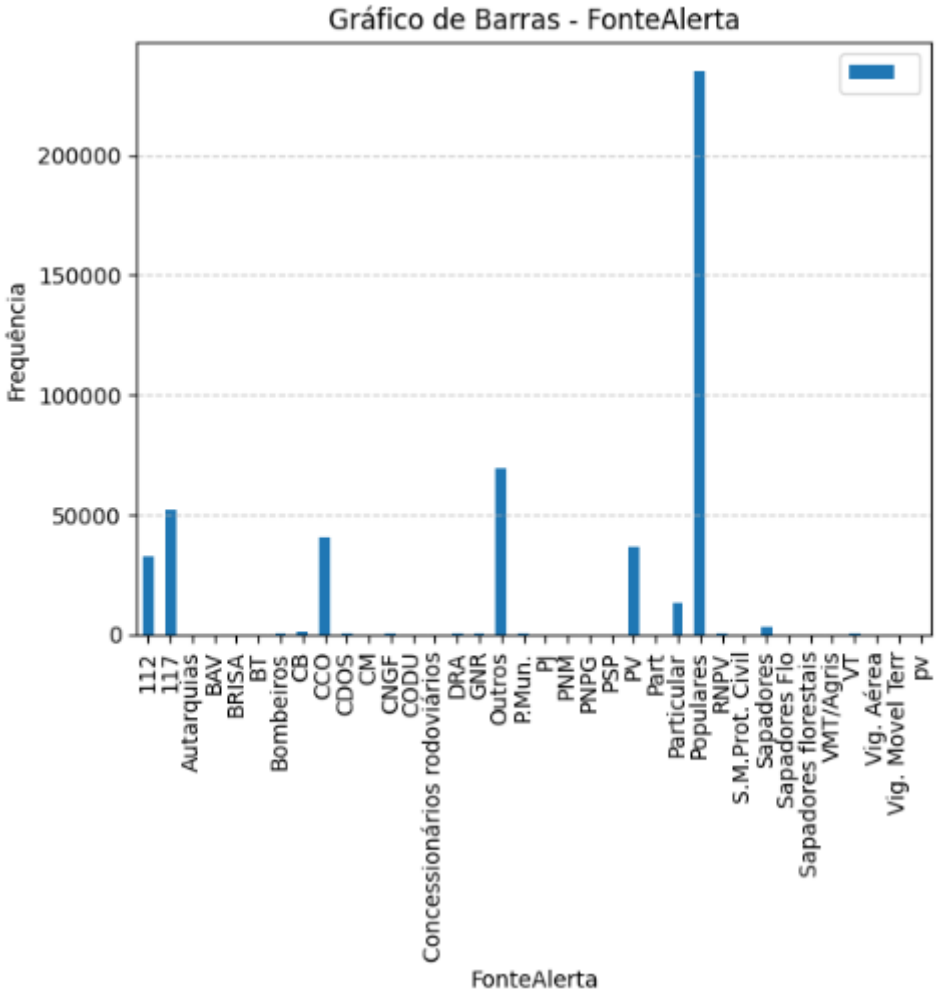


Figura 12- Gráfico de Barras para a coluna FonteAlerta

O gráfico de barras mostra que a maioria dos alertas de incêndio foi gerada por "Populares", seguida pelo número "117" e pela categoria "Outros", evidenciando o papel crucial da população na deteção e comunicação dos incêndios.

2.3.2.5 Distrito

Distrito	Frequência Absoluta	Frequência Relativa (%)
Aveiro	40338	8.29 %
Beja	6937	1.43 %
Braga	55697	11.45 %
Bragança	15947	3.28 %
Castelo Branco	12637	2.6 %
Coimbra	14869	3.06 %
Faro	9149	1.88 %
Guarda	17548	3.61 %
Leiria	17244	3.54 %
Lisboa	40176	8.26 %
Portalegre	6103	1.25 %
Porto	101649	20.89 %
Santarém	21900	4.5 %
Setúbal	18736	3.85 %
Viana Do Castelo	30877	6.35 %
Vila Real	31908	6.56 %
Viseu	38776	7.97 %
Évora	6003	1.23 %

Figura 13 - Tabela de frequência absoluta e frequência relativa para a coluna Distrito

A variável *Distrito* evidencia uma distribuição desigual do número de ocorrências por distrito. O Porto apresenta o maior número absoluto (101.649 ocorrências), seguido por Braga (55.697) e Viseu (37.876). Em contraste, Évora (6.003) e Portalegre (6.103) registam os menores valores. A análise da frequência relativa confirma este padrão: o Porto representa aproximadamente 20,89% do total de ocorrências registadas, enquanto distritos como Beja e Évora têm uma participação inferior a 2%. Estes dados indicam uma concentração significativa de incêndios em regiões específicas, com especial destaque para o litoral norte.

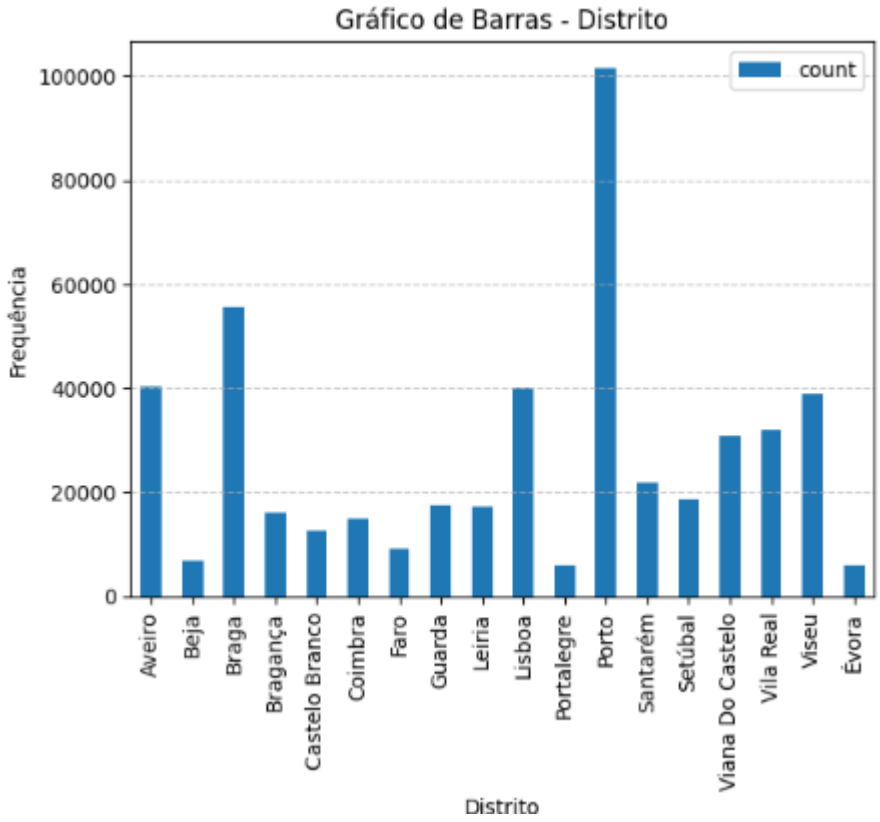


Figura 14 - Gráfico de Barras para a coluna Distrito

O gráfico de barras reforça esta interpretação, evidenciando visualmente a predominância do distrito do Porto no conjunto de dados, seguido de Braga, Viseu e Lisboa. Esta distribuição poderá indicar fatores como maior densidade populacional, características específicas da vegetação ou outras variáveis regionais associadas à incidência de incêndios.

2.3.2.5 Fogacho

Fogacho	Frequência Absoluta	Frequência Relativa (%)
Não	161031	33.09 %
Sim	325684	66.91 %

Figura 15 - Tabela de frequência absoluta e frequência relativa para a coluna Fogacho

A variável *Fogacho* apresenta duas categorias: “Sim” (presença de fogacho) e “Não” (ausência). A análise das frequências absoluta e relativa mostra que a maioria dos registos refere a ocorrência de fogachos, com 325.684 casos, o que corresponde a 66,91% do total. Em contraste, 161.031 registos (33,09%) não registam a presença deste fenómeno. Este padrão indica que a ocorrência de fogachos constitui um fator relevante no conjunto de dados analisado.

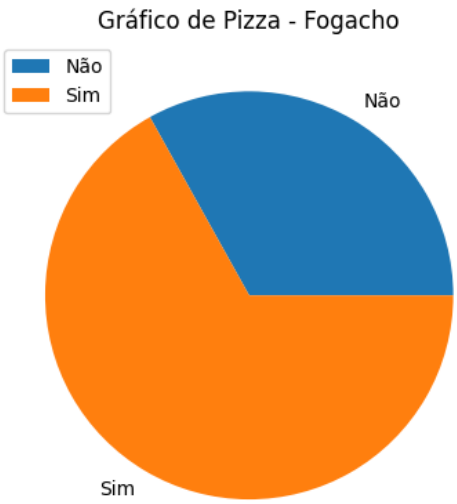


Figura 16 - Gráfico de Pizza para a coluna Fogacho

O Gráfico de Pizza permite visualizar de forma intuitiva esta distribuição, onde a maior fatia representa os registos em que o fogacho esteve presente. A representação gráfica reforça a conclusão de que a maioria dos casos da base de dados analisada está associada à ocorrência de fogachos.

2.3.2.6 Reacendimentos

Reacendimentos	Frequência Absoluta	Frequência Relativa (%)
Não	458303	94.16 %
Sim	28412	5.84 %

Figura 17 - Tabela de frequência absoluta e frequência relativa para a coluna Reacendimentos

A variável *Reacendimentos* indica se um incêndio sofreu ou não reacendimento. A análise combinada das frequências absoluta e relativa mostra que 458.303 ocorrências (94,16%) não apresentaram reacendimentos, enquanto 28.412 ocorrências (5,84%) envolveram reacendimentos. Estes resultados revelam que os incêndios sem reacendimentos são largamente predominantes no conjunto de dados analisado.

Figura 18 - Gráfico de Pizza para a coluna Reacendimentos

O gráfico de pizza reforça esta análise, ilustrando visualmente que a vasta maioria dos incêndios não se reacendeu, com apenas uma pequena fração do total a sofrer este fenómeno.

2.3.3 Análise Exploratória das Variáveis Temporais

2.3.3.1 Análise Exploratória por Mês

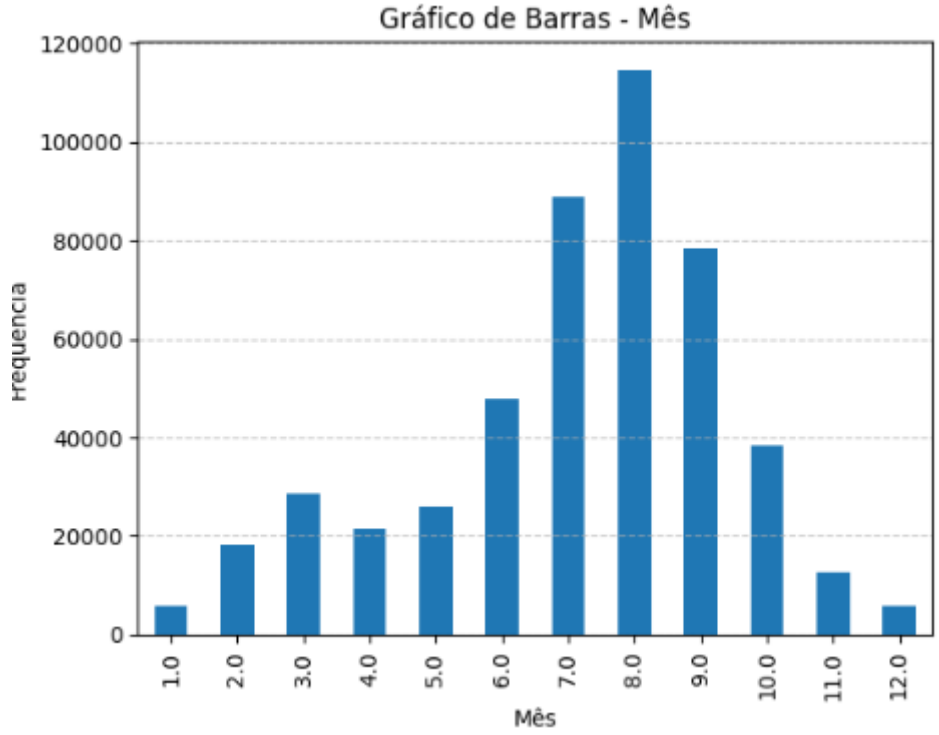


Figura 19 - Gráfico de Barras para a coluna Mês

O gráfico de barras por mês mostra uma clara sazonalidade nos incêndios, com maior número de ocorrências durante os meses de julho (7) e agosto (8), correspondendo ao verão, período de maior calor e seca. Estes dois meses representam o pico absoluto de frequência, seguidos por junho (6) e setembro (9). Já os meses de inverno (de novembro a fevereiro) apresentam os menores números de ocorrências, como esperado.

2.3.3.2 Análise Exploratória por Ano

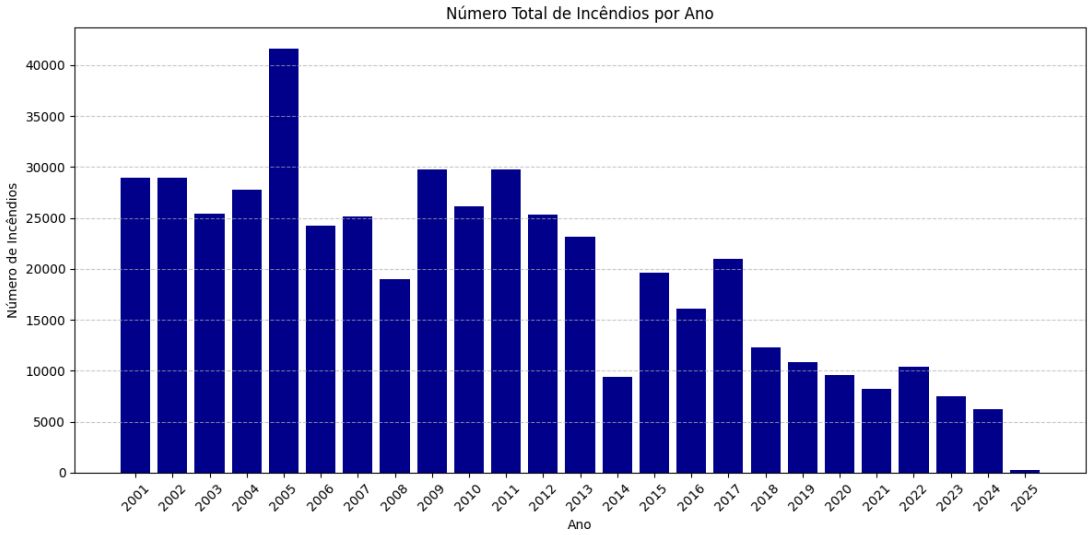


Figura 20 - Número total de incêndios por Ano

O gráfico da Figura 20 mostra a evolução do número total de incêndios por ano em Portugal, no período de 2001 a 2025. Observa-se um pico acentuado em 2005, seguido de uma oscilação até cerca de 2013, após a qual se verifica uma tendência clara de redução no número de incêndios anuais. A partir de 2017, a diminuição torna-se mais evidente, com os valores mais baixos registados entre 2020 e 2025. O ano de 2025 apresenta um valor residual, possivelmente devido à incompletude dos dados até à data de recolha.

2.3.3.3 Análise Exploratória por Hora do Dia

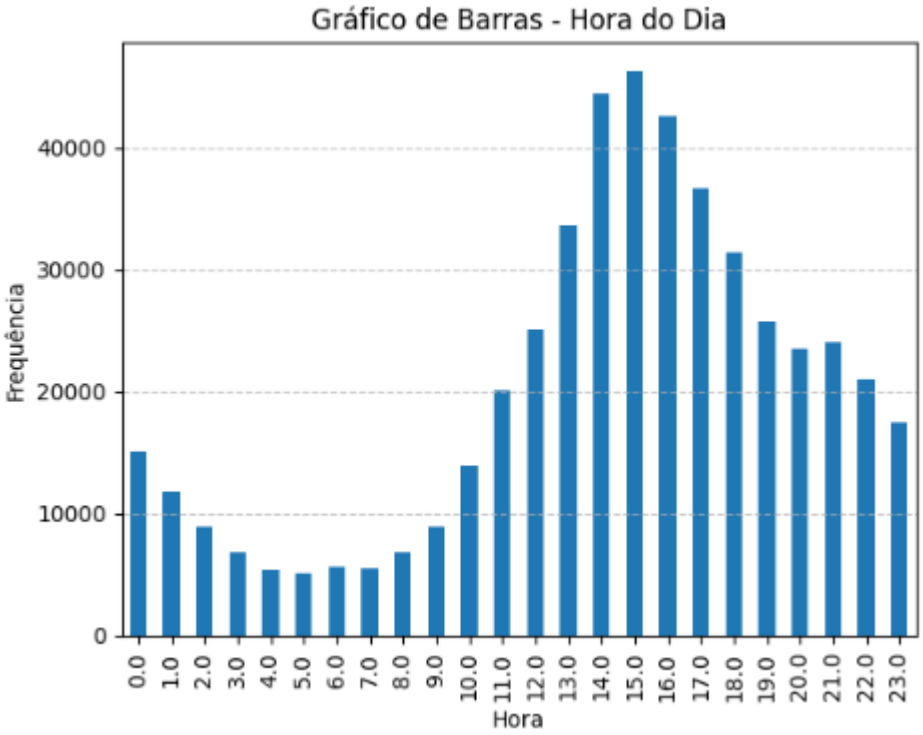


Figura 21 - Gráfico de Barras para a coluna Hora

Em relação ao horário do dia, os incêndios ocorrem principalmente entre as 12h e as 18h, com um pico claro por volta das 15h. Esta concentração no período da tarde coincide com as horas de maior temperatura e radiação solar, fatores que favorecem o início e propagação de incêndios. Durante a noite e madrugada (entre as 0h e as 8h), a frequência é muito menor. Estes padrões temporais reforçam a importância de reforçar a vigilância e os meios de prevenção especialmente nos meses de verão e durante o período da tarde.

2.3.4 Investigação de subgrupos

Para uma compreensão mais profunda dos dados e uma melhor fundamentação dos modelos de Data Mining, foi realizada uma análise exploratória com foco na identificação de subgrupos relevantes. A análise de subgrupos permite destacar padrões específicos que podem passar despercebidos numa abordagem mais global, fornecendo *insights* úteis tanto para os objetivos de negócio como para o desenvolvimento dos modelos preditivos.

A tabela seguinte apresenta os principais subgrupos considerados, acompanhados da respetiva justificação e ligação aos objetivos do projeto. Para além disso, foram criado gráficos que ilustram a distribuição de cada subgrupo e os mesmo podem ser encontrados desde o Anexo III até ao Anexo XIV.

Tabela 9 - Subgrupos de variáveis analisados

Nº	Subgrupo	Justificação
1	Tipo de incêndio × Classe de área ardida	Para identificar quais tipos causam mais área ardida.
2	Tipo de causa × Área ardida	Para verificar se causas negligentes ou intencionais estão associadas a maior impacto.
3	Distrito × Número de incêndios / Área ardida	Avaliar onde ocorrem mais incêndios e com que intensidade.
4	Ano × Tipo de incêndio ou Área ardida	Estudar evolução temporal da gravidade e frequência.
5	Reacendimentos × Tipo de incêndio ou Causa	Verificar se reacendimentos estão ligados a causas específicas.
6	Número de reacendimentos × Classe de área	Avaliar se reacendimentos contribuem para maior área ardida.
7	Tempo de rescaldo × Reacendimentos	Para verificar se tempos de rescaldo mais curtos estão associados a uma maior probabilidade de reacendimento.
8	Índices meteorológicos × Área ardida	Para avaliar se condições meteorológicas extremas influenciam a propagação dos incêndios e aumentam a área ardida.

2.3.4.1 Tipo de Incêndio x Classe de Área Ardida

Esta análise visa compreender a relação entre o tipo de incêndio e a classe de área ardida, em que permite identificar padrões relevantes sobre a frequência e a gravidade dos incêndios.

A tabela e os gráficos seguintes mostram o número total de incêndios e a área ardida agregados por tipo de incêndio (Agrícola, Florestal e Queima) e por classe de área ardida (de 0-1 ha até >500 ha).

	Tipo	ClasseArea	Nº Incêndios	Área Ardida Total (ha)
0	Agrícola	0-1 ha	79718	8.461566e+03
1	Agrícola	1-10 ha	5378	1.513821e+04
2	Agrícola	10-50 ha	486	9.690355e+03
3	Agrícola	50-100 ha	61	4.207600e+03
4	Agrícola	100-500 ha	31	6.015520e+03
5	Agrícola	>500 ha	2	3.024000e+03
6	Florestal	0-1 ha	335587	5.581215e+04
7	Florestal	1-10 ha	51410	1.675198e+05
8	Florestal	10-50 ha	8670	2.009207e+05
9	Florestal	50-100 ha	2120	1.523233e+05
10	Florestal	100-500 ha	2310	5.050600e+05
11	Florestal	>500 ha	929	2.079391e+06
12	Queima	0-1 ha	6	6.000000e+00
13	Queima	1-10 ha	7	1.800000e+01
14	Queima	10-50 ha	0	0.000000e+00
15	Queima	50-100 ha	0	0.000000e+00
16	Queima	100-500 ha	0	0.000000e+00
17	Queima	>500 ha	0	0.000000e+00

Figura 22 - Tabela do subgrupo Tipo de Incêndio por Classe de Área Ardida

Relativamente à Figura anterior, pode concluir-se o seguinte:

1. Incêndios “Florestais” com 0-1 ha registaram mais de 335 mil ocorrências, constituindo a maioria dos casos presentes no conjunto de dados. Embora tenham pequena dimensão, representam um esforço operacional elevado devido à sua quantidade.
2. Incêndios “Florestais” com área superior a 500 ha ocorreram em menor número (929 ocorrências), mas causaram a maior área ardida total, com mais de 2 milhões de hectares queimados. Este subgrupo destaca-se pelo elevado impacto ambiental.

3. Os incêndios “Agrícolas” surgem como o segundo tipo mais comum, sobretudo na classe de 0-1 ha, com cerca de 79 mil ocorrências. No entanto, a área ardida é consideravelmente inferior à dos incêndios florestais.
4. Os incêndios do tipo “Queima” foram residuais e apresentaram impacto praticamente nulo.

2.3.4.2 Tipo de Causa x Classe de Área Ardida

Esta análise visa compreender a relação entre o tipo de causa dos incêndios (Desconhecida, Intencional, Natural, Negligente ou Reacendimento) e a respetiva área ardida, agrupada por classes de dimensão. O objetivo passa por identificar se determinados tipos de causa estão mais frequentemente associados a incêndios de grande dimensão, o que pode orientar ações de prevenção e fiscalização.

	TipoCausa	ClasseArea	Nº Incêndios	Área Ardida Total (ha)
0	Desconhecida	[Area+1000]	76	254864.293067
1	Desconhecida	[Area-1]	70354	9646.719543
2	Desconhecida	[Area-50-100]	429	30851.526382
3	Desconhecida	[Area1-10]	8887	28563.714019
4	Desconhecida	[Area10-50]	1742	39915.531026
5	Desconhecida	[Area100-500]	455	98457.055164
6	Desconhecida	[Area500-1000]	76	53234.973485
7	Intencional	[Area+1000]	177	672777.770371
8	Intencional	[Area-1]	34834	6888.502220
9	Intencional	[Area-50-100]	532	38125.637077
10	Intencional	[Area1-10]	7767	25291.765312
11	Intencional	[Area10-50]	1945	45683.328871
12	Intencional	[Area100-500]	672	149039.171649
13	Intencional	[Area500-1000]	142	99899.347457
14	Natural	[Area+1000]	20	94206.222753
15	Natural	[Area-1]	1583	207.256902
16	Natural	[Area-50-100]	15	1112.877000
17	Natural	[Area1-10]	235	766.700587
18	Natural	[Area10-50]	68	1520.667017
19	Natural	[Area100-500]	41	9508.388204
20	Natural	[Area500-1000]	8	5728.992529
21	Negligente	[Area+1000]	92	313213.837568
22	Negligente	[Area-1]	62620	11508.737904
23	Negligente	[Area-50-100]	686	49001.084637
24	Negligente	[Area1-10]	13371	44115.470057
25	Negligente	[Area10-50]	3072	70082.708337
26	Negligente	[Area100-500]	691	146360.125678
27	Negligente	[Area500-1000]	125	87880.698483
28	Reacendimento	[Area+1000]	36	150091.340076
29	Reacendimento	[Area-1]	25553	2792.413174
30	Reacendimento	[Area-50-100]	115	8403.636373
31	Reacendimento	[Area1-10]	2286	7758.002229
32	Reacendimento	[Area10-50]	529	12341.418855
33	Reacendimento	[Area100-500]	160	37110.434465
34	Reacendimento	[Area500-1000]	35	24849.737058

Figura 23 - Tabela do subgrupo Tabela do subgrupo Tipo de Causa por Classe de Área Ardida

Relativamente à Figura anterior, pode concluir-se o seguinte:

1. As causas desconhecidas e negligentes estão associadas à maior frequência de incêndios, especialmente na classe [Area1-] (0-1 ha), com mais de 70 mil e 62 mil ocorrências, respetivamente. No entanto, estas causas apresentam valores médios de área ardida relativamente baixos, o que sugere menor impacto por ocorrência.
2. Os incêndios de causa intencional ocorrem com menor frequência, mas estão ligados a valores elevados de área ardida total, principalmente nas classes entre 50 e 1000 ha, onde destacam-se várias subcategorias com mais de 40.000 ha queimados. O destaque vai para os casos superiores a 1000 ha, que registam mais de 670 mil hectares queimados, um dos valores mais elevados da tabela.
3. Os incêndios causados por reacendimentos também revelam áreas ardidas significativas em classes superiores, como 100-500 ha e 500-1000 ha, o que sugere que reacendimentos podem contribuir para a propagação de incêndios já iniciados, agravando o impacto total.
4. As causas naturais representam o menor número de ocorrências e, no geral, causam áreas ardidas muito reduzidas em todas as classes.

2.3.4.3 Distrito x Número de incêndios / Área ardida

Esta análise tem como objetivo identificar os distritos com maior número de incêndios e os que registam maior área ardida, permitindo reconhecer regiões prioritárias para ações de prevenção e combate.

Distrito	Nº Incêndios	Área Ardida Total (ha)
Aveiro	40338	142612.398243
Beja	6937	78332.917117
Braga	55697	174911.714299
Bragança	15947	201985.714961
Castelo Branco	12637	356743.776069
Coimbra	14869	232137.860454
Faro	9149	188048.512717
Guarda	17548	359783.351528
Leiria	17244	150987.015841
Lisboa	40176	35041.165375
Portalegre	6103	114145.515498
Porto	101649	156179.049386
Santarém	21900	171282.081982
Setúbal	18736	30416.069007
Viana Do Castelo	30877	186909.375051
Vila Real	31908	270512.506865
Viseu	38776	311065.924055
Évora	6003	46140.282017

Figura 24 - Tabela do subgrupo Distrito por Número de incêndios e Área ardida

Relativamente à Figura anterior, pode concluir-se o seguinte:

1. O distrito do Porto apresenta o maior número de incêndios (mais de 101 mil ocorrências), seguido por Braga (55 mil), Aveiro (40 mil) e Lisboa (40 mil). Estes valores sugerem elevada densidade de ignições, possivelmente associada à atividade humana e à ocupação do solo.
2. Apesar de não liderarem em número de ocorrências, os distritos do interior centro e norte destacam-se pela área ardida total. Guarda (359 mil ha), Castelo Branco (356 mil ha), Viseu (311 mil ha) e Vila Real (270 mil ha) encontram-se entre os mais afetados, refletindo incêndios de maior dimensão e complexidade.
3. Alguns distritos, como Lisboa e Setúbal, apresentam número elevado de incêndios, mas áreas ardidas relativamente baixas, o que pode indicar eficácia no combate inicial ou predominância de fogos pequenos e rapidamente controlados.

4. Por outro lado, Coimbra e Faro apresentam valores significativos tanto em número de incêndios como em área ardida, o que releva uma combinação de frequência e gravidade que exige atenção especial.

2.3.4.4 Ano × Tipo de Incêndio ou Área Ardida

Esta análise permite observar a evolução temporal do número de incêndios e da área ardida, segmentada por tipo de incêndio (Agrícola, Florestal e Queima), e também o total anual de ocorrências e área afetada. Esta visão global ajuda a identificar anos críticos e tendências gerais, orientando estratégias de prevenção e resposta.

Tipo Ano	Agrícola	Florestal	Queima	TotalIncendios
2001	1973	26940	0	28913
2002	2495	26482	0	28977
2003	1698	23748	0	25446
2004	5816	21954	0	27770
2005	5979	35659	0	41638
2006	4250	19993	0	24243
2007	6117	19014	0	25131
2008	5254	13705	0	18959
2009	4784	24999	0	29783
2010	4870	21244	0	26114
2011	5593	24189	0	29782
2012	4795	20557	0	25352
2013	4427	18701	0	23128
2014	2545	6843	0	9388
2015	4163	15480	0	19643
2016	3070	13034	0	16104
2017	3610	17396	0	21006
2018	2356	9918	0	12274
2019	2761	8070	0	10831
2020	2115	7504	0	9619
2021	2036	6150	0	8186
2022	1927	8453	10	10390
2023	1626	5898	0	7524
2024	1389	4864	2	6255
2025	27	231	1	259

Figura 25 - Tabela do subgrupo Ano por Tipo de Incêndio

Tipo Ano	Agrícola	Florestal	Queima	AreaTotal
2001	714.641210	116704.597447	0.0	117419.238657
2002	1118.302200	129720.344150	0.0	130838.646350
2003	1121.006420	468566.300511	0.0	469687.306931
2004	3171.798710	148171.436100	0.0	151343.234810
2005	2159.342700	344526.411532	0.0	346685.754232
2006	3619.119000	80086.744650	0.0	83705.863650
2007	2375.397510	34036.139850	0.0	36411.537360
2008	2504.151800	17393.067499	0.0	19897.219299
2009	1584.698290	90540.964141	0.0	92125.662431
2010	2156.258080	138796.927880	0.0	140953.185960
2011	2417.449849	74686.075951	0.0	77103.525800
2012	1780.469360	116204.313016	0.0	117984.782376
2013	3769.823354	156617.404510	0.0	160387.227864
2014	1705.707057	21114.115621	0.0	22819.822678
2015	2256.182753	64944.042778	0.0	67200.225531
2016	1382.219760	166425.166851	0.0	167807.386611
2017	3005.464033	536915.531076	0.0	539920.995109
2018	1200.645093	43376.854962	0.0	44577.500055
2019	1966.669268	40117.249512	0.0	42083.918780
2020	3288.541337	63881.392447	0.0	67169.933783
2021	1395.335612	26964.672720	0.0	28360.008331
2022	687.168412	109391.395885	18.5	110097.064297
2023	496.711317	34012.498354	0.0	34509.209671
2024	655.531797	136992.893010	3.0	137651.424807
2025	4.611500	840.226706	2.5	847.338206

Figura 26 - Tabela do subgrupo Ano por Área Ardida

Relativamente às Figuras anteriores, pode concluir-se o seguinte:

1. O ano 2005 destacou-se com o maior número total de incêndios (41.638), impulsionado sobretudo por 35.659 incêndios florestais. Este valor reflete um período de elevada atividade, com numerosos focos de ignição, muitos deles de causa negligente ou intencional. Ocorreram incêndios relevantes nos distritos de Coimbra, Ourém e Castelo Branco, embora com uma área média por evento inferior à de anos mais extremos.
2. A maior área ardida total da série ocorreu em 2017, com mais de 539.000 hectares queimados. Este ano foi marcado por dois eventos catastróficos:
    1. Em junho, o incêndio de Pedrógão Grande provocou 66 mortes e centenas de feridos.
    2. Em outubro, incêndios simultâneos atingiram os distritos de Coimbra, Leiria e Viseu, causando mais de 40 vítimas mortais e destruição em larga escala. Estes eventos mostram como condições meteorológicas extremas e falhas na gestão do território podem amplificar tragédias.
3. Em 2003, registou-se a segunda maior área ardida da série (469.687 ha), com destaque para os incêndios que afetaram o Algarve (Monchique e Silves) e vastas zonas do centro do país. Foi um ano de verão severo, com elevadas temperaturas e ausência de chuva.
4. O ano 2020, com apenas 9.619 incêndios, demonstrou que a gravidade de um ano não depende apenas da frequência. Incêndios em Oleiros, Proença-a-Nova e Monchique causaram uma área ardida superior a 67 mil hectares, mostrando que menos ignições podem originar grandes danos se as condições forem adversas.
5. A partir de 2011, observou-se uma tendência geral de redução no número de incêndios, com os valores mais baixos registados em 2023 (7.524) e 2024. Apesar disso, alguns anos recentes, como 2016 e 2017, mostraram que a severidade dos eventos aumentou, exigindo maior preparação e capacidade de resposta.

2.3.4.5 Reacendimentos × Tipo de Incêndio

Esta análise tem como objetivo compreender a relação entre os reacendimentos e o tipo de incêndio (Agrícola, Florestal ou Queima), de modo a identificar que categorias apresentam maior propensão para reacender após terem sido consideradas extintas.

Tipo	Agrícola	Florestal	Queima
Reacendimentos			
Não	84526	373764	13
Sim	1150	27262	0

Figura 27 - Tabela do subgrupo Reacendimentos por Tipo de Incêndio

Relativamente à Figura anterior, pode concluir-se o seguinte:

1. Os dados mostram que os reacendimentos ocorrem quase exclusivamente em incêndios do tipo Florestal, totalizando 27.262 casos. Isto representa a esmagadora maioria dos reacendimentos registados na base de dados.
2. Em contraste, os incêndios Agrícolas registaram apenas 1.150 reacendimentos, enquanto os de Queima não apresentaram qualquer ocorrência de reacendimento.
3. Estes resultados sugerem que os incêndios florestais, devido à sua extensão, complexidade do terreno, e tipo de combustível (vegetação densa e combustível residual), apresentam maior dificuldade em serem totalmente extintos, o que aumenta o risco de reacendimento.
4. O tempo de rescaldo, as condições meteorológicas após a extinção (como vento e temperatura) e a estrutura do solo podem também desempenhar um papel importante na probabilidade de um incêndio florestal reacender, e deverão ser considerados no modelo preditivo.

2.3.4.6 Número de Reacendimentos × Classe de Área

Esta análise visa compreender como os reacendimentos se distribuem em função da classe de área ardida, procurando identificar se existe maior propensão para reacendimentos em incêndios pequenos ou de maior dimensão.

ClasseArea	Reacendimentos
[Area-1]	25251
[Area1-10]	2286
[Area10-50]	529
[Area100-500]	160
[Area-50-100]	115
[Area+1000]	36
[Area500-1000]	35

Figura 28 - Tabela do subgrupo Número de Reacendimentos por Classe de Área

Relativamente à Figura anterior, pode concluir-se o seguinte:

1. A grande maioria dos reacendimentos ocorreu em incêndios com área ardida muito pequena, destacando-se:
    - [Area-1] (até 1 hectare): 25.251 reacendimentos
    - [Area1-10]: 2.286 casos
    - [Area10-50]: 529 casos

2. Classes com áreas superiores a 100 hectares apresentam números muito reduzidos de reacendimentos, com apenas:
    - 160 casos em [Area100-500]
    - 36 em [Area+1000]
    - 35 em [Area500-1000]
3. Este padrão sugere que os reacendimentos estão muito mais associados a incêndios de pequena dimensão, o que pode ocorrer por vários motivos:
    - Rescaldo menos rigoroso por se tratar de ocorrências aparentemente "simples" ou "resolvidas rapidamente".
    - Ignorações iniciais localizadas e pouco profundas, que se reativam com variações meteorológicas.
4. Já nos grandes incêndios, a dimensão e complexidade operacional provavelmente conduzem a maior vigilância e ações de rescaldo mais robustas, o que pode explicar a baixa taxa de reacendimentos nas classes com maior área ardida.

2.3.4.7 Tempo de Rescaldo × Reacendimentos

Esta análise pretende explorar a relação entre o tempo de rescaldo (tempo decorrido entre a extinção do incêndio e o momento de reacendimento) e a ocorrência de reacendimentos, de forma a identificar padrões de risco que possam auxiliar na gestão pós-combate.

Reacendimentos	Não	Sim
FaixaRescaldo		
0-10min	75197	3770
10-30min	31229	2305
30-60min	11959	1017
1-2h	337458	21038
2-4h	1664	189
4-8h	462	61
8-24h	249	28
+24h	85	4

Figura 29 - Tabela do subgrupo Tempo de Rescaldo por Reacendimentos

Relativamente à Figura anterior, pode concluir-se o seguinte:

1. A maioria dos reacendimentos ocorreu na faixa de 1 a 2 horas após o término do incêndio, totalizando 21.038 casos. Esta faixa destaca-se como o intervalo mais crítico, concentrando o maior número absoluto de reacendimentos.

2. Ainda assim, observa-se que faixas mais curtas de tempo, como 0-10 minutos (3.770 reacendimentos) e 10-30 minutos (2.305 reacendimentos), também apresentam valores expressivos, sugerindo que reacendimentos podem ocorrer quase imediatamente após o fim oficial da operação de combate.
3. À medida que o tempo de rescaldo aumenta, o número de reacendimentos diminui drasticamente, com apenas:
    - 189 reacendimentos entre 2-4h
    - 61 entre 4-8h
    - 28 entre 8-24h
    - 4 reacendimentos após 24h
4. Este padrão indica que o período mais vulnerável ao reacendimento é até cerca de 2 horas após a extinção, e que a probabilidade de reacender reduz drasticamente com o tempo.

2.3.4.8 Índices Meteorológicos × Área Ardida

Esta análise tem como objetivo verificar a relação entre o valor do índice meteorológico FWI *(Fire Weather Index*) — que reflete o potencial de propagação de incêndios com base nas condições meteorológicas — e a área total ardida.

ClasseFWI	AreaTotal
Muito Baixo	4.903450e+04
Baixo	1.735706e+05
Moderado	5.565146e+05
Elevado	1.668071e+06
Muito Elevado	5.760490e+05
Extremo	1.469131e+05
Severo	3.739942e+04
Crítico	0.000000e+00

Figura 30 - Tabela do subgrupo Índices meteorológicos por Área Ardida

Relativamente à Figura anterior, pode concluir-se o seguinte:

1. A maior área ardida total foi registada na classe “Elevado”, com mais de 1,66 milhões de hectares queimados. Este resultado indica que condições de risco elevado são as mais críticas em termos de extensão da área queimada.
2. As classes “Muito Elevado” (576 mil ha) e “Moderado” (556 mil ha) também apresentam valores expressivos de área ardida, sugerindo que não são apenas os dias extremos que resultam em grandes perdas, mas também os de risco

moderado e elevado, provavelmente devido à frequência mais comum dessas condições.

3. Curiosamente, a classe "Crítico" apresenta zero hectares ardidos, o que pode dever-se a:
    - baixa frequência de ocorrência dessa classe no período analisado;
    - medidas preventivas mais fortes adotadas em dias de risco extremo;
    - ou possível ausência de registos em que o índice tenha atingido esses valores.
4. As classes "Muito Baixo" e "Baixo" correspondem a valores significativamente menores de área ardida, o que é coerente com o comportamento esperado: condições meteorológicas menos favoráveis à propagação resultam em menor impacto.

2.3.4.9 Hipóteses Derivadas da Análise Exploratória

Com base na análise exploratória realizada, foram identificados alguns padrões que permitem levantar hipóteses para investigações futuras. Estas hipóteses podem orientar análises mais aprofundadas ou testes estatísticos, de forma a contribuir para a construção de modelos preditivos mais completos e para a compreensão de fatores críticos associados à ocorrência e impacto dos incêndios.

Nº	Análise	Hipótese
1	Tipo de Incêndio × Classe de Área Ardida	Incêndios florestais tendem a ocorrer em classes de maior área ardida, enquanto os agrícolas concentram-se nas classes menores.
2	Tipo de Causa × Área Ardida	Causas intencionais podem originar incêndios mais severos; causas negligentes tendem a estar associadas a ignições mais frequentes, mas com menor área.
3	Distrito × Número de Incêndios / Área Ardida	Distritos do interior centro e norte apresentam maiores áreas ardidas médias; os distritos mais urbanos registam mais ignições, mas de menor gravidade.
4	Ano × Tipo de Incêndio ou Área Ardida	Existem anos críticos (ex. 2003, 2005, 2017) com picos de área ardida e/ou número de incêndios.

5	Reacendimentos × Tipo de Incêndio	A maioria dos reacendimentos ocorre em incêndios florestais, devido ao tipo de vegetação, terreno e maior dificuldade de extinção completa.
6	Classe de Área Ardida × Número de Reacendimentos	Incêndios de pequena área (< 1 ha) são os mais suscetíveis a reacender, possivelmente por menor rigor nas ações de rescaldo.
7	Tempo de Rescaldo × Reacendimentos	Tempos curtos de rescaldo podem aumentar a probabilidade de reacendimentos.
8	Índices meteorológicos (FWI) × Área Ardida	Valores elevados dos índices meteorológicos favorecem maior área ardida, por influenciarem a propagação do fogo.

Figura 31 - Hipóteses sugeridas com base na análise exploratória dos subgrupos

2.3.5 Verificar qualidade dos dados

No âmbito da análise da qualidade dos dados, foi realizada uma breve reunião com o Rui Almeida, responsável do ICNF, que nos deu informações relevantes sobre a fiabilidade da base de dados utilizada. Segundo o Rui Almeida, os dados registados nos primeiros anos do sistema (especialmente entre 2001 e 2005) apresentam algumas limitações de qualidade, uma vez que o Sistema de Gestão de Incêndios Florestais (SGIF) ainda se encontrava numa fase inicial de implementação e estabilização.

Foi referido que, a partir de 2012, a qualidade dos registos melhora substancialmente, refletindo uma maior maturidade do sistema e melhores práticas de recolha. Ainda assim, existem algumas inconsistências conhecidas, como a ausência de datas de fim em algumas ocorrências ou casos em que a data de fim é registada como anterior à data de início - o que reforça a recomendação de considerar a data de início como o registo mais confiável para efeitos de análise. Esta informação é essencial para orientar decisões sobre a preparação e limpeza dos dados, garantindo maior rigor nas análises posteriores.

Com base nas observações recolhidas, e tendo em conta as limitações identificadas na qualidade dos dados, procede-se agora a uma análise exploratória inicial focada na avaliação da completude, consistência e formato dos dados disponíveis. Esta análise foi realizada num ambiente Python (notebook), onde foram aplicados procedimentos para:

- Identificar e quantificar valores em falta (*nulls*);
- Verificar a coerência temporal entre datas de início e fim;
- Avaliar o formato e tipo das variáveis;
- Detetar possíveis valores anómalos ou fora do intervalo esperado.

1. Identificação de Valores Omissos

Na análise inicial, foi verificado que a base de dados contém valores omissos em várias colunas, sendo alguns dos mais significativas os RCM, EstadoRegisto e Observacoes.

Estes valores ausentes indicam que nem todos os registos estão completos, o que sugere que estas colunas não serão relevantes para os modelos de análise.

2. Verificação de Duplicados

Durante a análise de duplicidade, foi confirmado que a base de dados não contém registos duplicados, o que é positivo para a integridade da mesma.

3. Verificação de Valores Negativos

Detetaram-se alguns valores negativos em variáveis que não são esperados, como:

- Precipitação: valores negativos indicam possíveis erros de entrada ou falhas nos registos.
- A coluna Perigosidade também apresentou 7.856 registos negativos.

Estes dados precisam de ser analisados, uma vez que valores negativos nestas variáveis podem prejudicar as análises e modelos preditivos.

4. Validação de Colunas Binárias

As colunas esperadas como binárias, tais como OriginouReacendimento, Fogacho e Agricola foram verificadas. As análises confirmaram que estas colunas estão corretamente configuradas com valores 0 e 1, sem irregularidades nos valores apresentados.

5. Coerência Temporal Entre Datas

Na reunião com o Rui Almeida, responsável do ICNF, foi mencionado que poderiam existir algumas incoerências temporais nos dados, especialmente no que se refere à sequência e relações entre as datas de início, intervenção, resolução, conclusão e fim. No entanto, ao realizar a análise detalhada dos dados, não foram detetadas nenhuma inconsistência nas relações temporais:

- Datas superiores a DHFim: Não foram identificados casos em que qualquer uma das datas (de início, intervenção, resolução ou conclusão) fosse posterior à data de fim (DHFim), indicando que a sequência temporal entre os eventos está corretamente ordenada.
- Datas inferiores a DHInicio: Da mesma forma, não foram detetados registos em que as datas de intervenção, resolução, conclusão ou fim fossem anteriores à data de início (DHInicio). Desta forma, é garantida que as sequências de eventos respeitam a ordem lógica e cronológica esperada.
- Duração do Incêndio: Também foi confirmada a consistência na duração do incêndio, com os valores calculados a partir das datas de início e fim coincidindo com os valores registados na coluna DuracaoHoras, sem diferenças significativas.

Portanto, apesar da reunião com o ICNF ter sugerido possíveis incoerências temporais, a análise realizada na base de dados não revelou problemas nas relações de tempo entre os eventos. Isto garante que, no conjunto de dados analisado, as sequências temporais estão corretas e as durações registadas estão em conformidade.

6. Verificação de Causas de Incêndio

Foi também avaliada a consistência entre o campo CodCausa e a variável associada TipoCausa, com o objetivo de garantir que cada código de causa está associado apenas a um tipo de causa.

A análise indicou que não existem inconsistências: cada CodCausa está sempre relacionado com um único valor de TipoCausa. Esta coerência reforça a integridade do modelo de codificação utilizado nos registos e dá confiança para a utilização destes campos em análises futuras sobre a origem dos incêndios.

De forma geral, a análise realizada permite concluir que a base de dados apresenta um nível de qualidade adequado para prosseguir com as etapas seguintes do projeto. Embora existam limitações conhecidas, especialmente em registos mais antigos, as

verificações efetuadas demonstraram uma estrutura consistente, sem incoerências temporais ou formais significativas. A limpeza e validação dos dados asseguram uma base fiável para a realização de análises robustas e fundamentadas.

Fase 3: Preparação dos Dados

3.1 Seleção dos Dados

Nesta etapa, foi realizada uma primeira seleção das variáveis disponíveis no conjunto de dados, com o objetivo de identificar quais seriam relevantes para os objetivos da análise. A escolha levou em consideração a utilidade das variáveis, a sua completude (presença de dados omissos) e a ausência de redundâncias ou dependências externas.

A tabela do Anexo II resume as variáveis inicialmente consideradas, indicando quais foram mantidas e quais foram excluídas, junto com os motivos da exclusão (como ausência de dados, redundância de informação ou irrelevância para os objetivos definidos).

É importante destacar que esta seleção não foi definitiva: na etapa seguinte de limpeza dos dados, alguns atributos foram novamente avaliados. Essa reavaliação permitiu ajustes e, em alguns casos, levou à exclusão de variáveis que inicialmente haviam sido mantidas, ou à reintrodução de variáveis que passaram a fazer sentido após o pré-processamento.

3.2 Limpeza dos Dados

3.2.1 Valores omissos

Muitas variáveis no *dataset* apresentavam valores em falta, desde 5 até 475606 (quase a totalidade do dataset). Pelo que, diferentes métodos foram aplicados a fim de tratar dos mesmos.

Para a maioria das variáveis numéricas com poucos *missing values* (até cerca de 10%) a mesma foi substituída pela média.

Tabela 10 - Variáveis numéricas com missing values

Atributo	Nº de valores em falta	Atributo	Nº de valores em falta
DensidadeRV	5	dc	28
Perigosidade	11	dmc	28
Dist_CBS_m	11	ffmc	28
NIncSimul5000	11	bui	28
NIncSimulDistrito	11	DHFim	38
NIncSimulConcelho	11	DuracaoHoras	38
fwi	28	Temperatura	455

dsr	28	HumidadeRelativa	455
isi	28	VentoIntensidade	484
bui	28	DistIncSimul5000	619
ffmc	28	AreaTotalIncSimul	619
DHFim	38	Precipitação	569
DuracaoHoras	38	Rugosidade:	4671
HorasExposicaoMedia	4671	VentoDirecao:	6515
hFWI	6515	MaxFWIh_48h_PosExtincao	7083
hFFMC	6515	MaxFFMCh_48h_PosExtincao	7083
hISI	6515	MaxDMC_48h_PosExtincao	7083
MaxDC_48h_DiaPosExtincao	7083	MaxISIh_48h_PosExtincao	7083
MaxBUI_48h_PosExtincao	7083	DensidadeResidentes	7083
DensidadeEdificios	10818	DensidadeResidentes	12744

Relativamente aos valores omissos das variáveis categóricas, às mesmas foi atribuida uma nova classe chamada "desconhecido". Na tabela abaixo encontram-se as variáveis categóricas e os respetivos missing values.

Tabela 11 - Variáveis categóricas com missing values

Variáveis Categóricas	Nº Missing Values
FonteAlerta	1230
Distrito	221
Concelho	221
Nut2	221

Quanto aos atributos que apresentavam uma quantidade muito grande de *missing values* (mais de 70% do total) os mesmos foram analisados um por um a fim de tentar perceber o motivo da ausência dos mesmos e se as poucas ocorrências que apresentavam podiam transmitir ainda assim valor nas análises. Abaixo encontram-se as variáveis em causa, com respetivos *missing values* e uma pequena justificação.

- DH1Intervencao: 211798 *missing values*
- DHConclusao: 223161 *missing values*
- DHResolucao: 345673 *missing values*

Como estas três variáveis apresentaram muitos *missing values* optou-se por eliminar estas colunas e optar por outras colunas temporais sem valores em falta, como é o caso da variável "DuracaoHoras" (tempo do incêndio).

- TipoCausa: 247286 *missing values*
- GrupoCausa: 247286 *missing values*

As variáveis referentes às causas apresentavam muitos valores omissos, pelo que, foram excluídas. No entanto, a fim de tentar tirar partido máximo dos poucos dados que tinham disponíveis, foi feita uma análise desses mesmos valores no modelo de associação. Para ver os fatores que estavam mais relacionados a estas causas (implementação e resultados na Fase 4).

- RCM: 276984 *missing values*

Este campo foi eliminado devido à ocorrência de *missing values*, porém este atributo foi calculado novamente com ajuda de outros campos (Fase 3.3 Derivar Novos Dados).

- Reacendimento_IncendioPai: 458469 *missing values*

O campo "Reacendimento_IncendioPai" para além de ter muitos *missing values*, os valores que apresentavam representavam apenas códigos únicos, pelo que essa informação não foi considerada.

- Perimetro: 475606 *missing values*

Por último, a variável Perímetro tinha quase todos os valores em falta, e ao consultar os valores que apresentava, constatou-se que os mesmos eram localidades, pelo que a variável não foi considerada.

3.2.1 Tratamento de *Outliers*

Durante a fase exploratória dos dados, foi possível identificar que o *dataset* apresentava uma quantidade significativa de *outliers* em várias variáveis numéricas (Fase 2.3.1.1).

Alguns destes valores extremos foram considerados pouco plausíveis ou possivelmente introduzidos de forma incorreta, como no caso da variável "DuracaoHoras", que apresentava valores equivalentes a mais de 100 dias de incêndio contínuo. Após pesquisa e consulta a fontes externas, constatou-se que o maior incêndio registado em Portugal teve uma duração de 36 dias, o que sugere que alguns dos valores mais extremos podem dever-se a erros de introdução no registo dos dados.

Apesar disso, a aplicação direta do método tradicional de remoção de outliers com IQR (Interquartile Range) que considera o intervalo entre o 1.º quartil (Q1 = 0.25) e o 3.º quartil (Q3 = 0.75) resultaria numa perda significativa de observações válidas, prejudicando a representatividade do dataset. Por esse motivo, optou-se por ajustar o critério de remoção de *outliers*, utilizando uma abordagem mais conservadora com limiares baseados nos percentis 10 e 90 (Q0.10 e Q0.90). Esta abordagem permitiu reduzir o impacto dos valores extremos mais improváveis, preservando ao mesmo tempo um volume adequado de dados para análise.

A imagem abaixo representa os boxplots das variáveis numéricas padronizadas após a aplicação deste critério ajustado, permitindo visualizar a redução de outliers em várias variáveis sem comprometer a variabilidade natural dos dados.

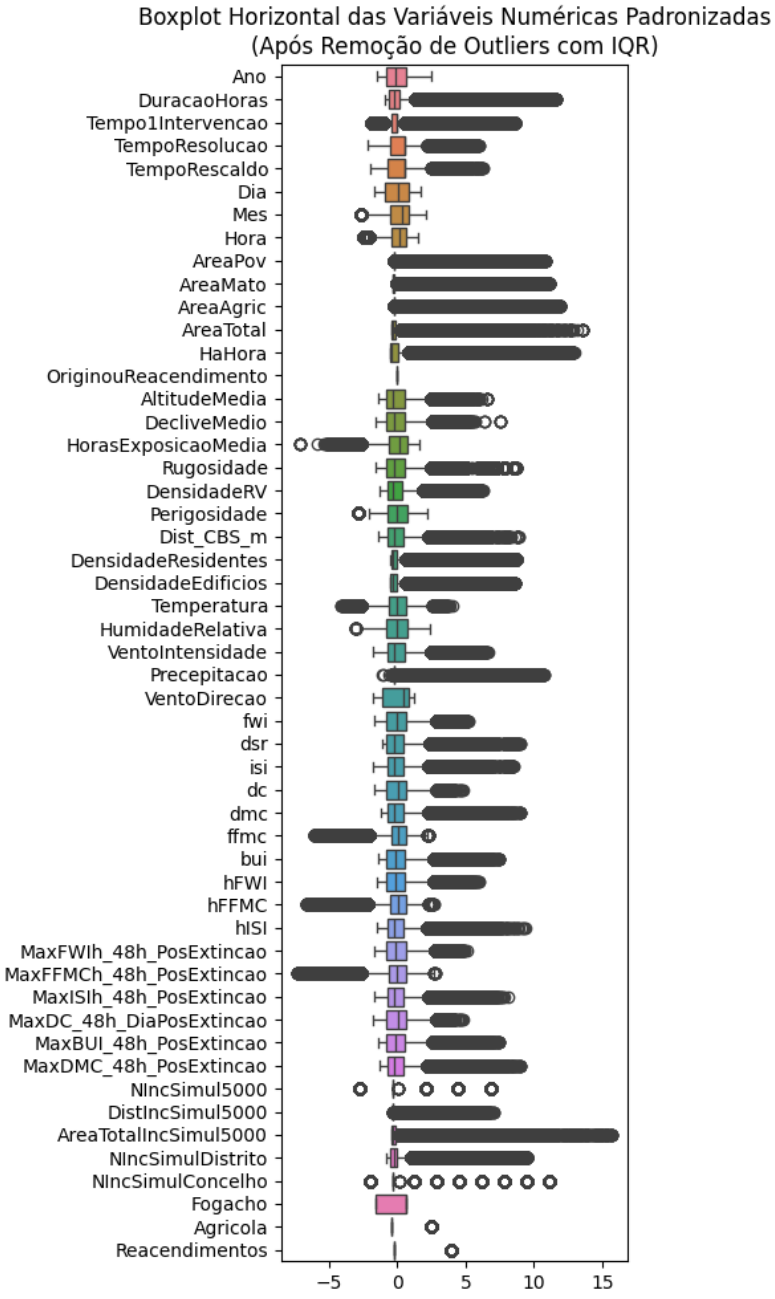


Figura 32 - Boxplot Horizontal das Variáveis Numéricas

A figura acima apresenta os *boxplots* horizontais das variáveis numéricas *standardizadas* após a aplicação do método IQR ajustado (percentis 10 e 90) para remoção de *outliers*. É possível observar alguma redução de valores extremos em várias variáveis, mantendo-se a distribuição geral dos dados e facilitando a comparação entre variáveis em escalas normalizadas.

3.2.2 Valores negativos

Durante a análise exploratória da base de dados, foram identificadas que diversas variáveis apresentavam valores negativos, o que não faz sentido para o seu contexto. Por esse motivo, foi realizado um processo de verificação, correção e validação desses valores.

Inicialmente, foi elaborada uma lista de variáveis que, por definição, não deveriam conter valores negativos. Esta lista incluía variáveis binárias, variáveis de contagem, indicadores meteorológicos e variáveis que estavam relacionadas com áreas. Em seguida, percorreu-se cada uma dessas variáveis e contabilizou-se os valores negativos encontrados. Verificou-se que algumas colunas, como Precipitacao, RCM e Perigosidade, apresentavam milhares de valores negativos, enquanto outras como fwi, isi, dc, dmc, ffmc e bui apresentavam exatamente 22 valores negativos, todos com o valor -80.

Para tratar esses casos, definiram-se abordagens específicas, tais como, no caso da variável Precipitação, concluiu-se que não faz sentido existirem valores negativos, uma vez que esta representa a quantidade de chuva observada (em milímetros) no dia do incêndio. Assim, todos os valores negativos foram substituídos por zero, o que indica que não choveu nesse dia.

As variáveis fwi, isi, dc, dmc, ffmc e bui, que são índices meteorológicos padronizados, apresentavam os mesmos 22 valores negativos com o valor -80. Este valor, claramente fora do intervalo aceitável, indica provavelmente a ausência ou erro na recolha dos dados. Segundo a documentação do ECMWF[1] (*European Centre for Medium-Range Weather Forecasts*), os valores esperados para estas variáveis são:

- FWI: 0 a 100
- ISI: 0 a 100
- DC, DMC, BUI: 0 a 1000
- FFMC: 0 a 101

Com base nisso e considerando que estas ocorrências estavam limitadas a apenas 22 linhas, optou-se por eliminá-las do *dataset*.

No caso da variável Perigosidade, consultou-se o documento técnico do IPMA e ICNF sobre a metodologia de cálculo do RCM, que indica que a classificação de Perigosidade deve estar entre 1 (menos perigoso) e 5 (mais perigoso), dividida por quintis nacionais. Os valores negativos e inferiores a 1 foram, portanto, substituídos por 1, de forma a manter a coerência com a escala e a evitar que afetassem o cálculo do RCM.

Em relação à variável RCM, decidiu-se recalculá-la com base nos valores de Perigosidade e FWI. Para isso, criou-se uma função que classificou o FWI em 6 classes, conforme os intervalos definidos pelo IPMA[2], e utilizou-se uma matriz de correspondência para determinar o valor de RCM a partir da classe de FWI e da

Perigosidade. O novo valor calculado foi então atribuído ao *dataset* original, substituindo os anteriores. Por fim, a variável auxiliar usada no processo (fwi_classe) foi eliminada.

Após todas as alterações, realizou-se uma validação final para garantir que nenhuma das variáveis da lista apresentava valores negativos. A verificação confirmou que todas as correções foram aplicadas corretamente e que os dados estavam consistentes com os valores esperados.

Este processo de limpeza e validação foi essencial para garantir a qualidade dos dados e a fiabilidade das análises subsequentes.

3.3 Derivar Novos Dados

Conforme mencionado na secção anterior (3.2 Limpeza de Dados), durante o processo de preparação dos dados, foi necessário reconstruir a variável RCM, com base nos atributos existentes Perigosidade e FWI.

Para isso, classificou-se o FWI em seis categorias, de acordo com os intervalos definidos pelo IPMA. Em seguida, uma matriz de correspondência foi aplicada entre a classe de FWI e a Perigosidade, de modo a atribuir um novo valor de RCM para cada registo. Esta operação resultou na criação de um novo atributo derivado, que substituiu o valor original da variável RCM no conjunto de dados.

Durante esse processo, foi utilizada uma variável auxiliar (fwi_classe), que serviu de apoio para a lógica de classificação. Após a geração do novo valor de RCM, essa variável auxiliar foi eliminada.

Por fim, foi realizada uma verificação final para garantir que nenhuma das variáveis apresentava valores inconsistentes (como negativos indevidos), assegurando que os dados estavam preparados de forma coerente e fiável para as fases seguintes.

3.4 Integrar Dados

A etapa de integração dos dados tem como objetivo juntar informações de diferentes origens, de forma a criar uma base de dados mais completa para análise. No entanto, neste caso, como só foi utilizada uma única fonte de dados, não houve necessidade de integrar tabelas ou combinar diferentes conjuntos de informações.

3.5 Formatar Dados

A etapa de formatação dos dados consiste em realizar transformações principalmente sintáticas, ou seja, mudanças na forma dos dados que não alteram o seu significado, mas que são necessárias para que o conjunto de dados esteja adequado às exigências da ferramenta de modelagem utilizada.

Inicialmente, foram verificados os tipos das variáveis na base de dados para entender que tipo de dados existem (numéricos, categóricos, datas, etc.). Por exemplo, foi identificado que há variáveis numéricas (como *float64* e *int64*) e variáveis nominais (do tipo *object*), que geralmente são textos ou categorias.

Nas variáveis nominais, foram listados os valores distintos para identificar inconsistências ou diferentes formas de representar o mesmo conceito. Por exemplo, em variáveis como *ClasseArea* e *FonteAlerta*, foram encontradas diferentes formas de escrever a mesma categoria (como “Sapadores”, “Sapadores florestais” e “Sapadores Flo”), o que pode gerar interpretações ambíguas durante a análise. Para resolver isso, foi criado um mapa de substituição para unificar esses valores, garantindo que cada categoria tenha uma única representação padronizada.

Também foi verificado se alguns dos valores numéricos estão dentro dos limites esperados, conforme definido na secção 3.2.2. Os valores fora desses intervalos (por exemplo, um índice *fwi* maior que 100, quando o máximo aceitável é 100), foram corrigidos para os limites máximos permitidos e, desta forma, evitou-se dados que possam prejudicar o modelo.

Outro ponto importante foi a formatação das variáveis de data e hora (DHInicio e DHFim). Foram detetadas que algumas datas estavam com formatos inválidos, o que poderia causar erros em fases posteriores. Para corrigir, foi adicionada uma hora padrão "00:00:00" quando a data estava incompleta. Desta forma garantiu-se que todas as datas estivessem no formato correto (YYYY-MM-DD HH:MM:SS).

Por fim, foi verificada a quantidade de categorias distintas em cada variável nominal para confirmar se a padronização foi eficaz e que não existiam valores duplicados ou inconsistentes que possam afetar a análise.

Essas transformações garantiram que os dados fossem consistentes, padronizados e no formato adequado para serem utilizados pelos modelos, facilitando assim a interpretação dos resultados e a qualidade do modelo final.

3.6 Criação do Dataset

Após as etapas de seleção, limpeza, tratamento e formatação dos dados, foi gerado o dataset final que será utilizado nas fases subsequentes da análise. Este novo conjunto de dados reflete a aplicação de todas as transformações, remoção de variáveis com muitos valores omissos, tratamento de *outliers*, correção de inconsistências e padronização das categorias.

O *dataset* final contém 415.071 observações e 63 variáveis, reduzido em relação ao conjunto inicial que possuía 486.715 observações e 100 variáveis. Essa redução deve-se principalmente à exclusão de colunas com alta proporção de dados omissos e à eliminação de registos com valores inconsistentes ou irreparáveis.

As variáveis remanescentes foram cuidadosamente escolhidas para garantir a qualidade e relevância dos dados para os objetivos da análise. Além disso, o *dataset* final está devidamente formatado para facilitar a aplicação dos modelos preditivos e análises estatísticas posteriores.

Esta etapa foi fundamental para garantir que as análises realizadas a seguir sejam baseadas em dados confiáveis, consistentes e representativos.

Fase 4: Modelação

Nesta fase do processo CRISP-DM, inicia-se a construção e avaliação de modelos que respondem aos objetivos de negócio definidos, como a identificação de fatores associados aos reacendimentos, a otimização do tempo de primeira intervenção, a classificação de reacendimentos e a criação de perfis territoriais de risco. Com os dados devidamente preparados, no passo seguinte foram aplicadas técnicas de modelização como algoritmos de associação, regressão, classificação e *clustering*.

4.1 Escolha das Técnicas de Modelação

Dado os objetivos de *Data Mining* definidos, as técnicas de modelação foram escolhidas com base na natureza do problema e nos dados disponíveis, considerando diferentes variáveis desde o tipo de causa, condições meteorológicas, área ardida e outras características.

Abaixo encontram-se as informações relativas aos objetivos, tipos de modelos utilizados assim como algoritmos escolhidos e considerações importantes.

4.1.1 Modelo de Associação (Objetivo 1)

Tabela 12 - Modelo de Associação (Objetivo 1)

Objetivo de Data Mining:	Identificar fatores associados aos Incêndios
Tipo de Modelo utilizado:	Modelo de Associação
Algoritmo utilizado:	Apriori

Como o objetivo não era prever diretamente uma variável alvo, mas sim descobrir padrões frequentes entre atributos relacionados aos reacendimentos, optou-se pelo uso do algoritmo Apriori. Este método permite identificar combinações de variáveis que ocorrem com frequência em casos de Incêndios. As regras geradas ajudam a compreender melhor os fatores de risco e podem ser utilizadas para orientar estratégias preventivas.

O modelo de associação Apriori exige que todas as variáveis utilizadas sejam categóricas. Uma vez que o conjunto de dados continha diversas variáveis numéricas, estas foram transformadas em categorias através da discretização em três intervalos. Como por exemplo a variável tempo ficou com as categorias:

- temperatura_baixa;
- temperatua_média;
- temperatura_alta;

Esta abordagem manteve-se para as restantes variáveis numéricas e, desta forma, as várias variáveis contínuas puderam ser utilizadas no modelo.

Foram consideradas todas as variáveis previamente selecionadas durante a fase de preparação dos dados.

4.1.2 Modelo de Regressão (Objetivo 2)

Tabela 13 - Modelo de Regressao (Objetivo 2)

Objetivo de Data Mining:	Analisar fatores que influenciam o tempo de resposta aos incêndios (Tempo1Intervencao) e prever situações de risco de atrasos operacionais
Tipo de Modelo utilizado:	Modelo de Regressão
Algoritmos utilizados:	Regressão Linear Múltipla (RLM) Regressão Linear Polinomial (RLP) Regressão por Árvores de Decisão (DTR) Regressão Random Forests (RFR)

Para atingir o objetivo de prever o tempo de resposta às ocorrências de incêndio - particularmente o tempo da primeira intervenção - foram selecionadas técnicas de regressão supervisionada. A escolha por técnicas de regressão justifica-se pela natureza contínua da variável alvo, "Tempo1Intervencao".

Foram aplicados quatro algoritmos de modelação distintos, de forma a permitir uma análise comparativa entre modelos simples e complexos, lineares e não lineares:

- **Regressão Linear Múltipla (RLM)**: modelo estatístico tradicional que assume uma relação linear entre a variável dependente e múltiplas variáveis independentes.
- **Regressão Linear Polinomial (RLP)**: extensão da RLM que incorpora termos polinomiais, permitindo capturar curvaturas nos dados.
- **Árvores de Decisão para Regressão (DTR)**: modelo não paramétrico que segmenta iterativamente os dados com base nos valores das variáveis preditoras.
- **Random Forest Regressor (RFR)**: conjunto de múltiplas árvores de decisão, que melhora a capacidade preditiva e reduz a variância do modelo.

Cada técnica possui diferentes pressupostos e implicações:

- A RLM assume linearidade entre variáveis, ausência de multicolinearidade, homocedasticidade dos resíduos e normalidade dos erros.
- A RLP flexibiliza a linearidade, mas mantém os restantes pressupostos estatísticos.
- As DTR e o RFR são métodos mais robustos e menos exigentes em termos de pressupostos, sendo capazes de lidar com não linearidades, variáveis categóricas e interações complexas sem necessidade de normalização dos dados.

A utilização de múltiplas técnicas permite não apenas verificar a qualidade de ajuste, mas também a interpretabilidade dos modelos e a sua capacidade de generalização.

4.1.3 Modelo de Classificação (Objetivo 3)

Tabela 14 - Modelo de Classificação (Objetivo 3)

Objetivo de Data Mining:	Prever se um incêndio pode gerar reacendimentos com base no tipo, duração, local, condições atmosféricas e área ardida.
Tipo de Modelo utilizado:	Modelo de Classificação
Algoritmos utilizados:	Regressão Logística Random Forest

Como o objetivo consistia em prever se um incêndio poderia gerar reacendimentos com base em variáveis como tipo, duração, localização, condições atmosféricas e área ardida, optou-se por utilizar um modelo de Classificação. Foram aplicados dois algoritmos: Regressão Logística e *Random Forest*.

A Regressão Logística permitiu avaliar o impacto individual de cada variável na probabilidade de reacendimento, o que ofereceu uma interpretação estatística dos fatores mais relevantes. Já o algoritmo *Random Forest*, por ser um modelo baseado em múltiplas árvores de decisão, apresentou-se como uma opção segura e eficaz para lidar com interações complexas entre variáveis e com possíveis padrões não lineares nos dados.

Para aplicar a Regressão Logística, foi necessário normalizar os dados com recurso ao *StandardScaler*, o que garantiu assim, que todas as variáveis tivessem a mesma escala. Dado o desequilíbrio da variável alvo, foi utilizada a técnica SMOTE para gerar exemplos da classe minoritária e equilibrar o conjunto de treino. Para melhorar o desempenho do modelo, foi também feita uma seleção de variáveis através do método de seleção sequencial (SFS), com a escolha apenas dos atributos mais relevantes para a predição do reacendimento.

No caso do modelo *Random Forest*, foi seguido o mesmo processo de normalização, balanceamento com SMOTE e seleção de variáveis. Embora este algoritmo não exija pressupostos estatísticos específicos, beneficiou igualmente do tratamento prévio dos dados. Ambos os modelos foram avaliados com base na matriz de confusão, precisão, recall e F1-score, o que garantiu uma análise da capacidade de previsão.

4.1.4 Modelo de Clustering (Objetivo 4)

Tabela 15 - Modelo de Clustering (Objetivo 4)

Objetivo de Data Mining:	Criação de clusters com base no número médio de incêndios, área ardida, perigosidade, tipo de causa

	e meteorologia média que representem o risco nas diferentes localidades.
Tipo de Modelo utilizado:	Modelo de Clustering
Algoritmos utilizados:	K-MEANS Clustering Hierárquico

Com o objetivo de identificar grupos de regiões com características semelhantes em termos de risco de incêndio, foi utilizado um modelo de *Clustering* não supervisionado. Foram aplicados dois algoritmos distintos: *K-Means* e *Clustering* Hierárquico.

Para ambos os métodos, foram selecionadas variáveis com relevância na caracterização do risco, como área ardida, duração dos incêndios, perigo, variáveis meteorológicas, índices de propagação do fogo e dados simulados sobre o histórico de reacendimentos. Estas variáveis, todas numéricas, foram previamente normalizadas com recurso ao *StandardScaler*, de forma a garantir que estivessem na mesma escala e não influenciassem desproporcionalmente o agrupamento.

Além disso, foi aplicada uma Análise de Componentes Principais (PCA) nos dois modelos, com o objetivo de reduzir a dimensionalidade e facilitar a interpretação dos clusters formados.

No caso do K-Means, essa transformação permitiu uma melhor visualização da distribuição dos grupos e auxiliou na definição do número de clusters. Já no *Clustering* Hierárquico, a redução dimensional foi complementada com a construção do dendrograma a partir do método *linkage*. A definição dos agrupamentos foi feita com base nos cortes visuais deste dendrograma.

Esta abordagem permitiu segmentar as diferentes localidades em perfis de risco, contribuindo para uma melhor compreensão dos padrões associados à ocorrência e reincidência de incêndios.

4.2 Definição de Planificação de Testes

Nesta etapa, são definidos os critérios e métodos de avaliação utilizados no processo de teste dos modelos. Esta fase é essencial para garantir que os modelos desenvolvidos são avaliados de forma consistente e comparável.

4.2.1 Modelo de Associação (Objetivo 1)

Para o modelo de associação Apriori, a avaliação das regras não segue os métodos clássicos de validação (como *cross-validation*), uma vez que não há uma variável alvo a prever, mas sim a identificação de padrões frequentes entre os atributos. Assim, a qualidade das regras geradas é analisada com base em três métricas principais: suporte *(support*), confiança (*confidence*) e *lift*.

- Suporte indica a frequência com que uma determinada combinação de itens ocorre no conjunto total de transações. Neste caso, foi definido um suporte mínimo de 0.002, o que significa que uma regra de associação só será considerada se os seus itens estiverem presentes em pelo menos 0.2% das 415.000 ocorrências.
- Confiança mede a probabilidade de ocorrência do item consequente (por exemplo, reacendimento) dado que o item antecedente ocorreu. Foram testados vários valores de confiança mínima, tendo em conta que valores mais elevados indicam maior fiabilidade da regra.
- *Lift* avalia a força da associação entre os itens, indicando se a ocorrência conjunta é maior do que o esperado pelo acaso. Valores de *lift* superiores a 1 revelam associação positiva, sendo estas regras as mais relevantes do ponto de vista interpretativo.

Para restringir a complexidade das regras e garantir maior interpretabilidade, definiram-se ainda os parâmetros min_length = 2 e max_length = 2, limitando a geração de regras a pares de itens.

4.2.2 Modelo de Regressão (Objetivo 2)

Para assegurar a fiabilidade e validade dos resultados obtidos pelos modelos de regressão, foi implementada uma estratégia de avaliação baseada na separação dos dados em conjunto de treino e conjunto de teste.

Foi utilizada uma divisão estratificada de 80/20, onde 80% dos dados foram utilizados para treinar os modelos e os restantes 20% para os testar. Esta divisão foi aplicada de forma aleatória, garantindo a representatividade da variável dependente em ambos os conjuntos.

As métricas de desempenho previstas para avaliação incluem:

- **R^2 (Coeficiente de Determinação)**
- **MSE (Mean Squared Error)**

Estas métricas fornecem uma análise abrangente do erro de previsão, sendo o R^2 útil para compreender a variabilidade explicada pelo modelo e o MSE indicador direto do erro de previsão.

4.2.3 Modelo de Classificação (Objetivo 3)

Para o modelo de classificação, foi realizada a divisão do conjunto de dados em conjunto de treino (80%) e conjunto de teste (20%), o que garantiu assim, que a avaliação dos modelos fosse feita com dados nunca vistos durante o processo de treino. Além disso, foi aplicado o método SMOTE apenas sobre o conjunto de treino, de forma a evitar vazamento de informação e assegurar uma avaliação mais realista da capacidade de generalização dos modelos.

O desempenho foi avaliado no conjunto de teste com base em métricas como a acurácia, precisão, *recall*, F1-score e matriz de confusão, permitindo uma comparação consistente entre modelos.

4.2.4 Modelo de *Clustering* (Objetivo 4)

Os modelos de clustering são não supervisionados (K-means e *Clustering* Hierárquico), pelo que não foi feita uma separação formal dos dados em treino e teste. A totalidade do conjunto de dados foi utilizada para identificar grupos com características semelhantes.

Para garantir a robustez da segmentação, os dados foram normalizados previamente e foi aplicada a técnica de PCA, o que permitiu reduzir a dimensionalidade e evitar o impacto de variáveis com escalas distintas.
No caso do K-Means, a definição do número ideal de clusters foi feita com base na análise da variabilidade intra-cluster, utilizando o método do cotovelo. No *clustering* hierárquico, a formação dos grupos baseou-se na análise visual do dendrograma.

4.3 Construção do Modelo

4.3.1 Modelo de Associação (Objetivo 1)

Na fase de construção deste modelo, foi utilizada a técnica Apriori, com o objetivo de identificar padrões frequentes e relações significativas entre atributos do dataset.

Inicialmente, foram consideradas todas as variáveis disponíveis no conjunto de dados. No entanto, para reduzir a complexidade do modelo e evitar redundância nas regras, optou-se por manter apenas uma variável representativa do local - o "Distrito" - em vez de múltiplas variáveis geográficas como "Concelho", "NUTS II" e "Local".

As variáveis numéricas do dataset foram categorizadas em três categorias com base nos seus intervalos de valores: "Baixo", "Médio" e "Alto". Esta transformação foi essencial para converter os dados em formato transacional, compatível com a aplicação do algoritmo Apriori.

Durante a definição dos parâmetros do modelo, foram definidos valores mínimos para suporte e confiança, de forma a garantir que apenas regras relevantes e com grau suficiente de ocorrência fossem geradas.

Suporte mínimo (min_support):

Define o limiar mínimo de frequência com que uma regra (ou item) deve aparecer nas transações para ser considerada relevante.

Com base em 415.000 transações e a natureza crítica dos incendios, um *min_support* de 0.002 é um ponto de partida equilibrado, permitindo descobrir padrões relevantes, sem ser demasiado permissivo. Valores mais baixos podem ser explorados, mas exigem controlo rigoroso do lift/confidence.

Número mínimo e máximo (Min_length e Max_length):

Ambos definidos como 2, para restringir a análise a regras com exatamente dois elementos, facilitando a interpretação e aplicação prática dos resultados.

Confiança mínima (min_confidence): Escolhido o valor de 0.2 para garantir que as regras apresentem um nível mínimo de fiabilidade (pelo menos 20% das vezes, o consequente ocorre quando o antecedente está presente), permitindo captar padrões relevantes mesmo que não muito fortes.

Força mínima (min_lift = 3): Define-se um limiar de 0.003, para garantir que as associações descobertas são três vezes mais prováveis do que ao acaso, focando apenas em relações com forte associação positiva.

Desta forma, o algoritmo Apriori foi aplicado, resultando na identificação de um conjunto de regras de associação com os respetivos valores de suporte, confiança e lift. Essas regras foram posteriormente analisadas com o intuito de extrair padrões úteis e interpretáveis no contexto do domínio dos dados.

4.3.2 Modelo de Regressão (Objetivo 2)

4.3.2.1 Regressão Linear Múltipla (RLM)

Após a seleção das variáveis explicativas mais relevantes e a remoção de colinearidades por meio da análise de VIF e do método de Backward Elimination, foi construído um modelo de Regressão Linear Múltipla (RLM) para prever o tempo de resposta da primeira intervenção (Tempo1Intervencao).

Pré-processamento e seleção de variáveis:

- Foram removidas variáveis com p-value elevado (> 0.05) por meio de eliminações iterativas, conforme a abordagem Backward Elimination.
- Foi realizada análise de multicolinearidade com o cálculo do VIF (Variance Inflation Factor), removendo variáveis como dsr, bui e MaxDMC_48h_PosExtincao por apresentarem VIF > 10.
- A versão final incluiu 24 variáveis explicativas, entre elas Hora, Dia, DuracaoHoras, fwi, Temperatura, VentoIntensidade, NIncSimul5000, Dist_CBS_m, entre outras.

Construção do modelo:

O modelo foi treinado com um conjunto de treino de 80% dos dados, utilizando a biblioteca sklearn.linear_model.LinearRegression.

4.3.2.2 Regressão Linear Polinomial (RLP)

O modelo de Regressão Linear Polinomial (RLP) foi construído a partir das variáveis numéricas selecionadas com base na Regressão Linear Múltipla (RLM), utilizando critérios como significância estatística (*Backward Elimination*) e multicolinearidade (VIF).

A transformação polinomial foi realizada com *PolynomialFeatures*, da biblioteca *sklearn.preprocessing*, e aplicada sobre as variáveis preditoras para considerar interações e curvaturas até o grau 3. Após a transformação, foi aplicado o modelo *LinearRegression* da *sklearn*.

4.3.2.3 Regressão por Árvores de Decisão (DTR)

O modelo de Regressão por Árvores de Decisão (DTR) foi construído com base numa ampla seleção de variáveis categóricas e numéricas, incluindo dados meteorológicos, topográficos, e características do incêndio. As variáveis categóricas foram transformadas via *One-Hot Encoding* com *handle_unknown='ignore',* e as numéricas foram normalizadas com *StandardScaler*. O alvo (Tempo1Intervencao) também foi normalizado.

As seguintes alterações foram efetuadas:

- Variáveis com baixa variância (*VarianceThreshold*).
- Variáveis altamente correlacionadas (*corr > 0.9*), com base na matriz de correlação entre as variáveis numéricas.

O modelo foi treinado com *DecisionTreeRegressor*, utilizando uma profundidade máxima (*max_depth=3*) para evitar sobreajuste e manter a interpretabilidade.

As variáveis mais importantes segundo o *atributo .feature_importances_* foram:

- Ano (30.74%)
- DuracaoHoras (30.73%)
- Dist_CBS_m (29.68%)
- FonteAlerta_112 (8.84%)

Estas variáveis foram utilizadas para reconstruir um segundo modelo mais simples e provou-se que apenas as variáveis acima são importantes para este modelo.

4.3.2.4 Random Forest Regressor (RFR)

O modelo de *Random Forest Regressor* (RFR) foi desenvolvido utilizando uma combinação de variáveis categóricas e numéricas representando características do incêndio, condições meteorológicas, localização, e contexto temporal.

À semelhança do modelo de regressão por árvores de decisão, as variáveis categóricas foram tratadas com *One-Hot Encoding*, e as numéricas foram normalizadas com *StandardScaler*. O alvo (Tempo1Intervencao) também foi normalizado.

Além disso, foram removidas variáveis altamente correlacionadas (correlação > 0.9) para evitar redundância e melhorar a generalização do modelo. A preparação de dados resultou num conjunto robusto, com variáveis diversificadas e escaladas adequadamente.

O modelo foi treinado com *RandomForestRegressor* usando 10 estimadores (*n_estimators=10*) e uma *seed* fixa para reprodutibilidade (*random_state=0*).

4.3.3 Modelo de Classificação (Objetivo 3)

4.3.3.1 Regressão Logística

Antes de treinar o modelo, as variáveis explicativas foram normalizadas com a função *StandardScaler*, assegurando que todas apresentavam a mesma escala. Como a variável dependente se encontrava desbalanceada, foi aplicada a técnica SMOTE (*Synthetic Minority Over-sampling Technique*) ao conjunto de treino, o que permitiu equilibrar a representação das classes.

Para identificar as variáveis mais relevantes, utilizou-se o método de seleção sequencial de atributos (*SFS - Sequential Forward Selection*), em validação cruzada com 5 folds. O critério de avaliação adotado foi o F1-score ponderado, dada a assimetria na distribuição das classes. O modelo utilizou random_state=0, garantindo a reprodutibilidade dos resultados.

Após a seleção dos atributos, procedeu-se à otimização dos hiperparâmetros através de um *Grid Search* com validação cruzada de 10 folds (*GridSearchCV*). Foram avaliadas várias combinações de penalizações e *solvers*, selecionando-se os melhores parâmetros com base no F1-score.

O modelo final foi treinado sobre os dados balanceados e testado com o conjunto de teste original. A avaliação de desempenho incluiu as métricas acurácia, precisão, recall, F1-score e matriz de confusão, permitindo aferir a capacidade de generalização do modelo.

4.3.4 Modelo de Clustering (Objetivo 4)

4.3.4.1 K-Means

A construção do modelo de K-Means iniciou-se com a seleção de variáveis contínuas consideradas relevantes para o risco de incêndio, como características do terreno, meteorologia, índice de perigosidade, simulações de reacendimentos, entre outras.
De forma a evitar distorções causadas por escalas diferentes, todas as variáveis foram normalizadas com o *StandardScaler*.

Aplicou-se a Análise de Componentes Principais (PCA) para reduzir a dimensionalidade do conjunto de dados. Esta transformação permitiu preservar a maior parte da variância com um menor número de componentes, facilitando a visualização e o agrupamento dos dados.

Para determinar o número ideal de grupos, recorreu-se ao método do cotovelo, com base no cálculo do WCSS (*Within-Cluster Sum of Squares*). A análise do gráfico indicou a presença de quatro clusters bem definidos (n_clusters=4).

O modelo foi treinado com os parâmetros n_clusters=4, random_state=42 e n_init=10. O valor elevado de n_init permitiu efetuar várias inicializações, reduzindo o risco de convergência para mínimos locais.
Após o treino, os rótulos dos clusters foram adicionados ao dataset original, o que possibilitou a caracterização dos diferentes grupos com base nas variáveis de entrada.

4.3.4.2 Clustering Hierárquico

No modelo de Clustering Hierárquico foi utilizado o método aglomerativo, mais concretamente o linkage 'ward', que tende a formar grupos compactos e com variância interna reduzida.

A preparação dos dados incluiu normalização com o *StandardScaler* e a aplicação de PCA, à semelhança do modelo de K-Means.

Inicialmente, foi retirada uma amostra de 500 observações para criar o dendrograma. Este foi construído com a função *linkage*, utilizando a distância euclidiana como medida base.

A análise visual do dendrograma revelou a existência de três agrupamentos bem diferenciados (n_clusters=4).

Com base nesse número, aplicou-se o modelo *AgglomerativeClustering* à totalidade dos dados, com uma amostra maior de 10.000 observações, e com os parâmetros n_clusters=4 e linkage='ward'.

Os rótulos dos clusters foram também integrados no dataset original. Esta informação permitiu identificar perfis de risco distintos em termos de condições geográficas, meteorológicas e de propagação do fogo.

4.4 Avaliar os Modelos

Nesta fase, os modelos obtidos são analisados com base nos critérios definidos, tendo em conta o contexto do problema. A análise permite avaliar a qualidade das regras geradas e identificar aquelas com maior relevância para os objetivos do projeto.

4.4.1 Modelo de Associação (Objetivo 1)

A avaliação dos modelos gerados com o algoritmo Apriori foi realizada com base em três métricas principais: suporte, confiança e lift. Estas métricas permitem medir, respetivamente, a frequência com que a regra ocorre no conjunto de dados, a fiabilidade da implicação entre antecedente e consequente, e a força da associação, comparando-a com a expectativa de ocorrência aleatória.

Foram analisadas com maior detalhe as regras que apresentaram valores de lift superiores a 3, confiança elevada e suporte significativo (superior a 3%), dado o seu potencial valor interpretativo e operacional. De seguida apresentam-se algumas das regras mais relevantes:

Regra 1: Agrícola=Sim ⇒ Fogacho=Não

- Suporte: 0,1415 — Esta regra ocorre em 14,15% de todas as observações, o que representa uma presença significativa no conjunto de dados.
- Confiança: 1,00 — Em todos os casos em que o incêndio ocorreu em terreno agrícola, não se registou fogacho.
- Lift: 3,61 — A ausência de fogacho é 3,6 vezes mais provável em incêndios agrícolas do que por ocorrência aleatória.

Esta regra revela um padrão forte e claro: incêndios em terrenos agrícolas raramente são fogachos (incêndios pequenos). Ou seja, tendem a ter maior dimensão e não se limitam a ocorrências pequenas. Este padrão pode ser explicado pela combinação de diferentes fatores como:

- potencial para rápida propagação;
- possíveis atrasos na deteção;
- condições climáticas críticas no período em que normalmente ocorrem;

Esta informação pode ser útil para orientar a priorização de monitorização nestas zonas e na alocação de recursos.

Regra 2: Declive_media ⇒ Perigosidade_alta

- Suporte: 0,0686 — Presente em cerca de 6,9% das ocorrências.
- Confiança: 0,5112 — Em mais de metade dos casos com declive médio, o local foi classificado com perigosidade alta.
- Lift: 3,37 — Esta associação é 3,3 vezes mais forte do que o esperado por acaso.

Esta regra sugere que são as áreas que apresentam declives nem muito altos nem muito baixos que estão frequentemente associadas a maior perigosidade. A regra pode servir como critério adicional na classificação de zonas críticas e no planeamento preventivo.

Regra 3: Perigosidade_alta ⇒ Rugosidade_media

- Suporte: 0,0406 — Esta regra ocorre em aproximadamente 4,06% das observações, o que representa uma presença moderada, mas estatisticamente relevante, no conjunto de dados.
- Confiança: 0,2681 — Em 26,8% dos casos classificados como de perigosidade alta, a rugosidade do terreno era média.
- Lift: 4,06 — Esta associação é mais de quatro vezes mais forte do que seria de esperar por acaso, indicando uma relação significativa entre perigosidade elevada e rugosidade média.

Esta regra indica que, quando zonas de perigosidade alta, estão muitas vezes associadas a terrenos que apresentam alguma rugosidade.

Regra 4: 112 ⇒ TempoResolucao_baixa

- Suporte: 0,0571 — Esta regra ocorre em 5,71% das observações, o que representa uma presença estatisticamente significativa no conjunto de dados.
- Confiança: 0,8644 — Em 86,4% dos casos em que a origem do alerta foi o 112, o tempo de resolução da ocorrência foi considerado baixo.
- Lift: 3,13 — A probabilidade de um tempo de resolução baixo, dado que o alerta veio do 112, é mais de 3 vezes superior ao que seria esperado por acaso.

Esta regra sugere uma associação forte e positiva entre alertas provenientes do 112 e um tempo de resolução mais rápido das ocorrências de incêndio. Este padrão indica que o comportamento do sistema de resposta funciona bastante bem.

Regra 5: dmc_media ⇒ dc_alta

- Suporte: 0,0417 — Esta regra aparece em 4,17% das observações, o que representa uma ocorrência estável e útil para análise.
- Confiança: 0,8611 — Sempre que o índice DMC (Duff Moisture Code) está em nível médio, em 86,1% dos casos o índice DC (Drought Code) está elevado.
- Lift: 3,62 — A associação entre estas duas variáveis é 3,6 vezes mais forte do que o esperado por acaso.

Esta regra relaciona dois índices meteorológicos usados no sistema canadiano de previsão de incêndios florestais:

- O DMC mede a humidade da camada intermédia da vegetação (ramos finos, folhas secas), sendo sensível à precipitação recente.
- O DC mede a humidade de camadas mais profundas e densas do solo, sendo muito mais lento a reagir à chuva ou à humidade.

A regra mostra que mesmo quando a humidade nos materiais combustíveis mais leves é moderada, o solo mais profundo e os combustíveis pesados ainda se mantêm muito secos. Isto indica um cenário de risco acumulado e persistente, onde o fogo pode não começar facilmente, mas, uma vez iniciado, tem maior probabilidade de se tornar intenso e prolongado.

Dado isto, as regras apresentadas demonstram que o modelo Apriori foi eficaz na identificação de padrões relevantes. Várias associações encontradas confirmam relações já conhecidas no domínio dos incêndios florestais, o que valida o processo. Outras regras revelam correlações que, embora menos óbvias, podem representar novas oportunidades de análise e apoio à decisão.

4.4.2 Modelo de Regressão (Objetivo 2)

4.4.2.1 Regressão Linear Múltipla (RLM)

A performance do modelo foi avaliada no conjunto de teste (20% dos dados). Os principais resultados foram:

- **Erro Quadrático Médio (MSE)**: 38.90
- **Coeficiente de Determinação (R^2)**: 0.10

Interpretação:

- O valor de **R^2 = 0.10** indica que o modelo consegue explicar apenas 10% da variabilidade da variável Tempo1Intervencao. Este valor sugere que, embora algumas variáveis tenham influência, existe ainda grande variabilidade não capturada pelo modelo linear.
- O **MSE relativamente elevado** também indica que o modelo apresenta erros consideráveis nas previsões individuais.
- A análise dos resíduos e o summary() do modelo indicam que várias variáveis têm baixa significância estatística, e que o modelo linear pode ser insuficiente para capturar a complexidade dos dados.

Comparação entre modelos (parcial):

Outros modelos como Regressão Polinomial, Árvore de Decisão Regressora e Random Forest Regressor foram também considerados e comparados (a serem detalhados nas próximas secções). A baixa performance da RLM levanta a hipótese de que modelos não lineares possam ser mais adequados.

Revisão de parâmetros:

Dado o fraco desempenho do modelo linear simples, foi recomendada:

- a remoção de variáveis redundantes com baixa significância estatística;
- a experimentação de transformações não lineares;
- e a priorização de modelos de árvore para iterações seguintes.

4.4.2.2 Regressão Linear Polinomial (RLP)

O modelo polinomial de grau 3 foi avaliado com as seguintes métricas no conjunto de teste:

- **Mean Squared Error (MSE)**: 32.13
- **Coeficiente de Determinação (R^2)**: 0.26

Estes valores representam uma melhoria face à regressão linear múltipla, indicando que a inclusão de termos polinomiais permitiu capturar padrões mais complexos entre as variáveis e o tempo de intervenção.

A performance do modelo é considerada moderada, e embora o R^2 esteja abaixo de 0.30, é superior aos outros modelos testados até este ponto (como a RLM, que obteve R^2 = 0.10). O modelo é capaz de prever o tempo da primeira intervenção com maior precisão, especialmente em cenários com múltiplos fatores combinados.

Nas próximas subsecções vamos verificar os modelos baseados em árvores.

4.4.2.3 Regressão por Árvores de Decisão (DTR)

A avaliação do modelo no conjunto de teste revelou os seguintes resultados:

- **Mean Squared Error (MSE)**: 0.89
- **Coeficiente de Determinação (R^2)**: 0.11

Estes valores demonstram uma capacidade preditiva modesta, com um R^2 ainda baixo, indicando que o modelo explica apenas cerca de 11% da variabilidade do tempo até a primeira intervenção.

Apesar da limitação na performance, o modelo tem a vantagem de ser interpretável, evidenciando quais variáveis mais contribuem para o tempo de resposta. Essa característica pode ser particularmente útil para tomada de decisão operacional.

O modelo será mantido como referência comparativa para o outro modelo baseado em árvore, *Random Forest*.

4.4.2.4 Random Forest Regressor (RFR)

A avaliação do modelo no conjunto de teste produziu os seguintes resultados:

- **Mean Squared Error (MSE):** 0.61
- **Coeficiente de Determinação (R^2)**: 0.39

O modelo de *Random Forest* demonstrou melhor desempenho em relação à árvore de decisão simples (DTR), com um ganho significativo em R^2 (+0.28). O valor do coeficiente R^2 indica que o modelo consegue explicar aproximadamente 39% da variabilidade da variável alvo, um desempenho considerado razoável dada a complexidade do problema e as variáveis disponíveis.

A natureza do *Random Forest* permite reduzir *overfitting* ao combinar várias árvores com diferentes subconjuntos de dados e variáveis, mantendo um bom equilíbrio entre viés e variância.

Este modelo representa uma solução mais robusta para o problema, sendo particularmente útil para explorar a importância relativa das variáveis e oferecer previsões mais estáveis.

4.4.3 Modelo de Classificação (Objetivo 3)

4.4.3.1 Regressão Logística

Ao longo do desenvolvimento do modelo de Regressão Logística foram testadas três abordagens distintas, com o objetivo de comparar o desempenho obtido em cada uma e selecionar a versão mais adequada de acordo com os critérios definidos.

Na primeira abordagem, utilizou-se o conjunto de variáveis originais, com normalização e aplicação de SMOTE para corrigir o desbalanceamento entre classes. O modelo alcançou uma acurácia de 50% e um F1-score de 0,16 para a classe minoritária (reacendimentos). O *recall* para essa classe foi de 74%, o que indica uma elevada taxa de deteção de reacendimentos, mas com uma precisão reduzida (9%), revelando um número significativo de falsos positivos.

Na segunda abordagem, procedeu-se à seleção sequencial de atributos com o método SFS (*Sequential Forward Selection*), com o intuito de reduzir a complexidade do modelo e melhorar a sua capacidade preditiva. O modelo resultante manteve um *recall* elevado (66%), mas apresentou novamente precisão baixa (7%) e um F1-score de 0,12, confirmando a dificuldade do modelo em gerar previsões equilibradas. A acurácia global registou um ligeiro decréscimo para cerca de 41%, o que é coerente com o reforço da sensibilidade em detrimento da especificidade.

Por fim, aplicou-se uma afinação de hiperparâmetros com *Grid Search* e validação cruzada, testando várias combinações de C, penalty e solver. Durante a validação cruzada, o modelo obteve um F1-score médio de 93,87%, sugerindo um bom desempenho com os dados de treino balanceados. No entanto, ao aplicar o modelo otimizado ao conjunto de teste, verificou-se que os resultados se mantiveram consistentes com as versões anteriores: acurácia próxima dos 50%, *recall* elevado (74%) e precisão baixa (9%). Isto indica que a complexidade do problema e o desequilíbrio de classes continuam a limitar o desempenho prático do modelo, mesmo após afinação.

A comparação entre os três modelos demonstra que a Regressão Logística é eficaz na deteção de reacendimentos (devido ao elevado *recall*), mas mantém uma precisão insuficiente, o que compromete a sua aplicabilidade em contextos operacionais onde falsos positivos representam custos elevados. Apesar disso, o modelo cumpre o seu objetivo de fornecer interpretação estatística dos fatores de risco e poderá ser utilizado como ferramenta complementar de apoio à decisão.

4.4.4 Modelo de Clustering (Objetivo 4)

4.4.4.1 K-Means

Para o modelo de K-means foi utilizado o método do cotovelo como critério principal para determinar o número ideal de clusters. A análise do gráfico revelou uma inflexão evidente no ponto correspondente a quatro clusters, sendo este o valor escolhido para o agrupamento final.

Após a segmentação dos dados em quatro grupos, foi realizada uma análise descritiva das médias das variáveis por cluster. Os resultados indicaram diferenças claras entre os grupos, com destaque para variáveis como área ardida (ÁreaTotalIncSimul5000), precipitação, intensidade do vento, perigosidade e número de reacendimentos. Estas variáveis mostraram-se relevantes na distinção dos perfis de risco de incêndio entre as diferentes regiões. A presença de valores contrastantes em variáveis meteorológicas e de severidade operacional sugere que os clusters captam padrões diferenciados de risco, alinhados com o objetivo definido para esta tarefa de *clustering*.

Para avaliar a qualidade da segmentação, foram calculados dois índices internos de validação:

O índice de Silhueta apresentou um valor reduzido de 0,082, o que indica uma separação pouco definida entre os clusters.

Por outro lado, o índice de *Calinski-Harabasz* obteve um valor elevado de 47053,89, o que sugere que os grupos gerados apresentam boa coesão interna e separação externa.

Embora o índice de Silhueta revele uma sobreposição entre alguns clusters, o elevado valor de Calinski-Harabasz, aliado à análise interpretativa das variáveis por grupo, reforça a validade prática da segmentação. Os clusters gerados permitem identificar regiões com diferentes perfis de risco, o que pode ser útil para apoiar decisões estratégicas de prevenção, planeamento e gestão de incêndios.

De forma geral, o modelo de K-means cumpriu o seu propósito de estruturar os dados em grupos distintos com base em características ambientais, sociais e operacionais, fornecendo uma visão exploratória complementar à modelação preditiva, com potencial aplicabilidade em contextos de gestão territorial.

4.4.4.2 Clustering Hierárquico

Para o modelo de Clustering Hierárquico, aplicou-se o método aglomerativo com ligação do tipo Ward, definindo-se a criação de quatro clusters com base na análise prévia do dendrograma. Após a segmentação dos dados, recorreu-se a dois indicadores internos para avaliar a qualidade dos agrupamentos.

O índice de Silhueta registou um valor de 0,080, o que indica uma separação modesta entre os grupos gerados. Este valor sugere que a maioria das observações se encontra próxima das fronteiras dos clusters ou que os grupos partilham características comuns, comprometendo a nitidez da segmentação.

O índice de Calinski-Harabasz apresentou o valor de 977,93, o que, embora superior a alguns limiares mínimos, permanece significativamente abaixo do valor observado no modelo K-means. Este resultado revela uma coerência interna e separação entre grupos limitadas, o que enfraquece a robustez do modelo hierárquico nesta aplicação específica.

Embora o Clustering Hierárquico permita uma estrutura de agrupamento interpretável e útil para análises exploratórias, os resultados obtidos indicam baixa qualidade na segmentação dos dados. Em termos comparativos, o modelo de K-means demonstrou desempenho superior, tanto nos valores dos índices de validação como na interpretação prática dos grupos gerados.

Fase 5: Avaliação

5.1 Avaliar os Resultados

5.1.1 Modelo de Associação (Objetivo 1)

O objetivo principal do negócio definido neste projeto foi a redução do número de incêndios. A técnica de associação (Apriori) foi usada para descobrir padrões entre variáveis que se repetem com frequência com base num *dataset* de mais de 400 mil ocorrências ao longo de 25 anos, contendo 53 variáveis.

Tabela 16 - Critério de Sucesso e Resultado (objetivo 1)

Critério de Sucesso	Pelo menos 3 regras com suporte > 3%, relevância interpretável e ligação com reacendimentos.
Resultado	Este critério foi cumprido com sucesso. Foram identificadas mais de 10 regras com suporte superior a 3% e lift elevado (>3), sendo várias diretamente relacionadas com o risco de reacendimento ou com condições favoráveis à extinção eficaz

Com isto, conclui-se que o modelo de associação identificou regras valiosas com a realidade operacional. As relações descobertas contribuem para a redução de incêndios, pois permitem antecipar situações de risco e atuar preventivamente e os critérios de sucesso estabelecidos foram cumpridos com sucesso.

5.1.2 Modelos de Regressão (Objetivo 2)

Após testar diferentes modelos para prever o tempo até à primeira intervenção, concluiu-se que o modelo de *Random Forest* foi o que apresentou o melhor desempenho. Ele conseguiu explicar cerca de 39% da variação do tempo, o que indica que, apesar de não ser perfeito, tem uma capacidade razoável de previsão. Os modelos mais simples, como a regressão linear e polinomial, tiveram resultados inferiores, enquanto a árvore de decisão, apesar de ser fácil de interpretar, mostrou menor precisão.

É importante destacar que nenhum dos modelos capturou toda a complexidade do problema, o que pode indicar que existem fatores importantes não contemplados nos dados disponíveis. Se possível, seria ideal testar o modelo diretamente em situações reais para avaliar a sua eficácia prática.

Tabela 17 - Critério de Sucesso e Resultado (objetivo 2)

Critério de Sucesso	Coeficiente de determinação (R2) ≥ 70%
Resultado	Este critério não foi cumprido com sucesso. O melhor coeficiente de determinação (R2) foi através do Random Forest e foi de 39%.

Em resumo, o modelo de *Random Forest* é o recomendado para avançar no projeto.

Modelo	MSE (aprox.)	R^2	Observações
Regressão Linear Múltipla (RLM)	36.5	0.17	Modelo base, captura apenas relações lineares.
Regressão Polinomial (RLP, grau 3)	32.1	0.26	Melhoria face à RLM, mas maior risco de sobreajuste.
Decision Tree Regressor (DTR)	0.89*	0.11	Baixa generalização, útil para interpretação.
Random Forest Regressor (RFR)	0.63*	0.37	Melhor desempenho geral, mais robusto.

5.1.3 Modelo de Classificação (Objetivo 3)

5.1.3.1 Regressão Logística

O modelo de Regressão Logística teve como principal objetivo prever a ocorrência de reacendimentos com base em variáveis como a localização, condições atmosféricas, duração e área ardida do incêndio. A aplicação prática deste modelo insere-se num contexto operacional real, sendo fundamental avaliar a sua eficácia no apoio à tomada de decisão preventiva.

A primeira versão do modelo apresentou um desempenho modesto, com valores de *recall* na classe minoritária (reacendimentos) a atingir 74%, mas com uma *precision* bastante reduzida (9%), o que indicou uma elevada taxa de falsos positivos. Apesar disso, a capacidade do modelo para detetar a maioria dos reacendimentos reais

mostrou-se relevante, considerando o impacto que essas ocorrências podem ter sobre os recursos de combate e a segurança pública.

Após a aplicação da técnica de seleção sequencial de variáveis (*Sequential Forward Selection*), a performance do modelo manteve-se dentro do mesmo intervalo, com ligeiras variações. O *recall* continuou a apresentar valores acima dos 65%, enquanto a *precision* permaneceu baixa. Esta estabilidade sugeriu que o modelo não perdeu capacidade preditiva ao utilizar um subconjunto mais reduzido de variáveis, o que representa um ganho em interpretabilidade.

Por fim, procedeu-se à afinação dos hiperparâmetros através de uma pesquisa em grelha (*Grid Search*), o que permitiu melhorar ligeiramente o desempenho do modelo, atingindo um valor médio de *F1-score* de aproximadamente 93.87% nos dados de validação cruzada. Esta afinação otimizou o equilíbrio entre sensibilidade e precisão, mantendo a robustez do modelo mesmo em cenários com desbalanceamento de classes.

A avaliação global indica que, embora o modelo tenda a gerar falsos positivos, a sua sensibilidade elevada torna-o adequado para um cenário onde o custo de não prever um reacendimento pode ser significativamente superior ao custo de um alarme falso. Assim, considera-se que o modelo cumpre, em grande parte, os objetivos do projeto, apresentando-se como uma ferramenta viável para integrar em sistemas de apoio à decisão em operações de combate a incêndios.

Tabela 18 - Critério de Sucesso e Resultado (objetivo 3)

Critério de Sucesso	F1-score ≥ 70%
Resultado	Este critério foi cumprido com sucesso, uma vez que utilizando o modelo de Regressão Logística com Grid Search conseguiu-se um F1-score de 93.87%

5.1.4 Modelo de Clustering (Objetivo 4)

A presente fase visa avaliar se os modelos desenvolvidos respondem de forma adequada aos objetivos definidos, tendo em conta os critérios de sucesso do projeto e a sua aplicabilidade prática. Para além da análise técnica dos resultados, importa considerar o alinhamento com os objetivos de negócio e a utilidade da informação gerada. A avaliação contempla também a identificação de possíveis limitações ou descobertas adicionais que possam orientar futuras análises ou apoiar a tomada de decisão.

5.1.4.1 K-Means

Para avaliar a qualidade da segmentação, foram utilizados dois índices quantitativos:

- O índice de silhueta, com valor médio de 0.082, indica que a separação entre os clusters é baixa, o que pode refletir sobreposição entre grupos ou estruturas complexas nos dados.
- O índice de Calinski-Harabasz apresentou um valor elevado (47,053.89), o que sugere que os grupos possuem boa separação intercluster e baixa dispersão interna, o que reforça a estrutura estatística da segmentação.

A análise visual com PCA mostrou uma separação razoável entre os grupos, confirmando parcialmente os resultados dos indicadores quantitativos.

Em termos qualitativos, a interpretação das médias das variáveis por cluster permitiu identificar quatro perfis distintos:

Cluster 0 - Risco Muito Elevado: zonas florestais de alto risco, com elevada perigosidade (3.40), fogacho (0.84) e condições críticas (altitude, declive, vento, fwi).

Cluster 1 - Risco Elevado: regiões com condições meteorológicas severas, destacando-se pela temperatura (25.16 °C), fwi (36.58) e vento intenso (12.70).

Cluster 2 - Baixo Risco: áreas de baixo impacto, com os menores valores nas variáveis analisadas, indicando incêndios moderados ou controlados.

Cluster 3 - Risco Urbano de longa duração: zonas urbanas com elevada densidade populacional (2.82), grande área ardida (3.63) e longa duração dos incêndios (3.76 h).

Apesar do índice de silhueta indicar uma separação modesta, os demais critérios — perfil distinto dos clusters, boa separação visual em PCA e índice de Calinski-Harabasz elevado — demonstram que o modelo foi capaz de gerar uma segmentação estatisticamente consistente e interpretável, respondendo adequadamente aos objetivos propostos.

5.1.4.2 Clustering Hirárquico

A qualidade da segmentação foi avaliada através de métricas quantitativas:

- O índice de silhueta apresentou um valor de 0.080, o que indica uma separação limitada entre os clusters, com alguma sobreposição entre os grupos.
- O índice de Calinski-Harabasz registou um valor de 977.93, o que sugere uma separação aceitável entre clusters e boa coesão interna, embora inferior ao valor obtido com o modelo K-Means.

A análise visual dos dados em PCA reforçou a consistência estrutural dos clusters, mesmo com margens de interseção.

Em termos qualitativos, a interpretação das médias das variáveis por cluster permitiu identificar quatro perfis distintos:

Cluster 0 – Risco Urbano Moderado: zonas urbanas com elevada densidade de edifícios (1.83) e população residente (1.46), mas com baixos níveis de perigosidade (-0.26) e reacendimentos (-0.18), sugerindo impacto urbano mais estrutural do que climático.

Cluster 1 – Risco Médio: regiões com incêndios de intensidade média, caracterizadas por duração mais elevada (0.50), temperatura ligeiramente abaixo da média (-0.24) e fogacho reduzido (-0.18), sem extremos meteorológicos ou topográficos relevantes.

Cluster 2 – Baixo Risco: áreas com valores baixos em quase todas as variáveis, incluindo perigosidade (-0.34), fogacho (-0.19), reacendimentos (-0.25) e área ardida (-0.34), o que reflete incêndios menos severos e bem controlados.

Cluster 3 – Risco Elevado: zonas com área total ardida extremamente elevada (3.94), acompanhada por valores elevados de fogacho (0.45), reacendimentos (0.39) e perigosidade (0.18), o que indica ocorrências críticas e de grande extensão.

5.1.4.3 Conclusão da Avaliação

Com base nas métricas de avaliação, na análise visual e na interpretação qualitativa dos clusters, conclui-se que o modelo K-Means apresenta melhor desempenho no contexto deste estudo.

A segmentação gerada por este modelo demonstrou maior coesão interna e separação entre os grupos, conforme evidenciado pelo índice de Calinski-Harabasz significativamente superior.

Além disso, os perfis identificados foram mais distintos e representativos, o que facilitou na interpretação e validação da segmentação com base nos critérios de sucesso estabelecidos.

Assim, o modelo K-Means é o mais adequado para responder ao objetivo de clustering neste caso, sendo o modelo escolhido para interpretações e ações futuras.

Tabela 19 - Critério de sucesso e Resultado (objetivo 4)

Critério de Sucesso	Índice de Silhouette ≥ 0.6
Resultado	Este critério não foi cumprido com sucesso. O melhor índice de Silhouette foi através do K-means e foi de 0.082.

5.2 Revisão do Processo

5.2.1 Modelo de Associação (Objetivo 1)

O processo de desde modelo foi realizado de forma sistemática e a qualidade do trabalho foi garantida através de:

- Discretização cuidadosa das variáveis contínuas para se adaptarem ao algoritmo Apriori.
- Remoção de outliers com base em percentis ajustados (10 e 90) para preservar observações críticas.
- Validação interpretativa das regras, confrontando os resultados com o conhecimento do domínio (combate a incêndios).

Contudo, foram identificadas algumas limitações e pontos de melhoria:

- Nem todas as regras têm utilidade operacional direta é necessário um "filtro humano" para interpretação e seleção das associações mais relevantes.
- A discretização de variáveis pode reduzir a sensibilidade do modelo em contextos mais especificos (ex: dias com condições extremas).

5.2.2 Modelos de Regressão (Objetivo 2)

Ao rever todo o processo, notou-se que o tratamento dos dados foi feito de maneira cuidadosa: as variáveis categóricas foram codificadas corretamente e as variáveis numéricas foram normalizadas. A divisão dos dados em treino e teste foi feita de forma adequada, garantindo que os resultados são confiáveis e que evitamos o *overfitting*. Além disso, foram eliminadas variáveis com alta correlação para evitar redundância.

Apesar disso, ainda há espaço para melhorias, como testar mais modelos, ajustar melhor os parâmetros e explorar outras técnicas de validação. Também seria interessante verificar com mais profundidade a qualidade dos dados, procurando possíveis erros ou valores fora do esperado. Outro ponto importante é garantir que os dados usados hoje estarão disponíveis no futuro para novas previsões.

De modo geral, o processo foi bem conduzido, mas existem sempre oportunidades para ajustes e refinamentos.

5.2.3 Modelo de Classificação (Objetivo 3)

Durante o desenvolvimento do objetivo de clustering, foram aplicadas duas técnicas distintas — K-Means e Clustering Hierárquico — com o intuito de segmentar os territórios com base em características de risco de incêndio.

A aplicação dos métodos permitiu atingir os objetivos inicialmente propostos, com destaque para o modelo K-Means, que demonstrou resultados mais robustos e interpretáveis segundo os critérios definidos. A escolha deste modelo foi fundamentada

na sua melhor coesão interna, separação entre clusters e clareza na interpretação dos perfis.

Apesar disso, o processo evidenciou alguns desafios, nomeadamente a baixa silhueta geral dos modelos, sugerindo a existência de sobreposição ou complexidade estrutural nos dados. Ainda assim, a utilização de métricas complementares, como o índice de Calinski-Harabasz e a análise qualitativa, permitiu validar os resultados obtidos.

De forma geral, o processo de modelação foi bem-sucedido, com as técnicas de clustering a revelarem-se adequadas para o tipo de problema. Contudo, uma análise futura poderia considerar o uso de outras técnicas ou incorporar variáveis adicionais para melhorar a separabilidade dos grupos. Esta experiência reforça a importância de combinar métricas quantitativas com avaliação interpretativa para garantir a utilidade prática dos resultados obtidos.

5.2.4 Modelo de Clustering (Objetivo 4)

Durante a aplicação dos modelos de clustering, foram identificados alguns desafios e decisões metodológicas relevantes que merecem ser destacadas.

Inicialmente, foi necessário realizar normalização das variáveis para garantir que todas as dimensões tivessem o mesmo peso, evitando que variáveis com maior amplitude dominassem a segmentação. Além disso, a aplicação de PCA (Análise de Componentes Principais) foi uma etapa fundamental para reduzir a dimensionalidade dos dados e facilitar a visualização dos grupos.

A definição do número ideal de clusters no K-Means foi apoiada pelo método do cotovelo, mas esta escolha pode ser subjetiva, dependendo da interpretação da curva. No modelo hierárquico, a escolha do número de grupos baseou-se na inspeção visual do dendrograma, o que também introduz um elemento de julgamento humano no processo.

Foi também observado que o índice de silhueta apresentou valores baixos em ambos os modelos, o que sugere que a estrutura dos dados pode não formar agrupamentos perfeitamente bem definidos. Ainda assim, a validação por métricas complementares (como o índice de Calinski-Harabasz) e a interpretação qualitativa reforçaram a robustez da segmentação, especialmente no K-Means.

Por fim, optou-se por testar os modelos numa amostra reduzida antes de aplicar à totalidade dos dados (10.000 observações), garantindo maior eficiência computacional e estabilidade na estrutura dos clusters.

5.3 Determinar Ações Futuras

5.3.1 Modelo de Associação (Objetivo 1)

Com base nos resultados obtidos, recomenda-se a continuação do projeto com foco na aplicação prática e no aprofundamento analítico. Em primeiro lugar, as regras mais relevantes extraídas através do modelo Apriori deverão ser integradas em sistemas de apoio à decisão, contribuindo para a vigilância ativa e para a definição de prioridades no terreno. A sua aplicação poderá ser particularmente útil em contextos de monitorização pós-extinção, ajudando a reduzir a probabilidade de reacendimentos.

Além disso, é aconselhável a criação de mapas de risco territoriais baseados nas associações encontradas, permitindo identificar zonas críticas com maior propensão para reacendimentos em função de variáveis meteorológicas, topográficas e operacionais. Paralelamente, o modelo poderá ser complementado com abordagens supervisionadas, como algoritmos de classificação, com o objetivo de prever de forma mais direta a probabilidade de reacendimento com base nas características do evento.

Outra ação importante será a atualização regular do modelo com dados mais recentes, garantindo a sua validade temporal e ajustamento à realidade operacional atual.

5.3.2 Modelos de Regressão (Objetivo 2)

Diante dos resultados e da revisão do processo, sugere-se seguir em frente com a implementação de um piloto usando o modelo de Random Forest. Este teste prático permitirá avaliar se o modelo funciona bem no contexto real e identificar ajustes necessários. Paralelamente, vale a pena continuar a ajustar os parâmetros do modelo para tentar melhorar seu desempenho.

Também seria interessante explorar a inclusão de novas variáveis e dados que possam ajudar a compreender melhor o problema, além de utilizar técnicas como validação cruzada para reforçar a confiabilidade das previsões.

Caso o modelo não atenda às expectativas na aplicação prática, talvez seja necessário repensar o problema ou incluir dados adicionais. Por fim, é importante avaliar o custo-benefício desta implementação para garantir que o esforço faz sentido para a organização.

Assim sendo, a recomendação é avançar com o piloto do *Random Forest*, enquanto se continua a melhorar a análise para garantir resultados cada vez melhores.

5.3.3 Modelo de Classificação (Objetivo 3)

O objetivo deste processo consistiu em prever a ocorrência de reacendimentos com base em variáveis como tipo de ignição, duração, localização, condições atmosféricas e área ardida. Para isso, foi aplicado o algoritmo de classificação Regressão Logística.

Durante a construção do modelo, efetuaram-se várias iterações com técnicas de seleção de variáveis (Sequential Forward Selection) e afinação de hiperparâmetros (Grid Search), de forma a melhorar o desempenho preditivo. A métrica de recall foi priorizada nas decisões, dado o interesse prático em identificar a maioria dos casos positivos de reacendimento, mesmo à custa de um menor valor de precisão.

Apesar dos esforços aplicados, o modelo apresentou limitações no desempenho, com valores reduzidos de precisão e F1-score. Estes resultados podem estar associados ao forte desbalanceamento entre as classes, à possível ausência de variáveis com maior capacidade explicativa, ou a padrões não lineares difíceis de captar com o modelo utilizado.

Embora se tenha considerado a aplicação de outros algoritmos mais robustos, como o Random Forest, o tempo disponível para a realização do projeto não permitiu a sua implementação e validação.

Ainda assim, o processo permitiu uma melhor compreensão dos fatores associados ao reacendimento e trouxe indicadores úteis sobre as limitações e potencialidades dos dados disponíveis.

Ações futuras recomendadas:

- Explorar modelos mais robustos para dados desbalanceados, como XGBoost ou LightGBM, com técnicas integradas de penalização.
- Avaliar a viabilidade de abordagens de oversampling alternativas ao SMOTE, como ADASYN ou SMOTE-ENN.
- Analisar em maior detalhe os falsos negativos, de modo a identificar padrões que possam ser explorados por regras específicas ou modelos híbridos.

5.3.4 Modelo de Clustering (Objetivo 4)

Os resultados obtidos com o modelo de clustering, especialmente com o K-Means, permitem identificar perfis distintos de risco e impacto associados a incêndios florestais, o que representa uma base sólida para diversas aplicações futuras.

Com base nessa segmentação, recomenda-se a implementação prática dos clusters identificados como ferramenta de apoio à tomada de decisão, nomeadamente em:

1. **Planeamento de prevenção e resposta**: priorizando recursos e ações de vigilância nas regiões classificadas como “Risco Muito Elevado” ou “Risco Elevado”.
2. **Comunicação com populações e autoridades locais**: adaptar estratégias de sensibilização consoante os perfis de risco dos clusters.
3. **Integração em sistemas de alerta ou dashboards operacionais**,que permite uma visão geográfica e estratégica dos perfis de incêndio.

Para investigações futuras, sugere-se:

1. A **exploração de novas abordagens de clustering**, como métodos baseados em densidade (ex: DBSCAN), que podem identificar estruturas menos lineares ou agrupamentos mais subtis nos dados.
2. A **avaliação da estabilidade dos clusters ao longo do tempo**, dividir os dados por períodos (ex: por década ou por anos críticos) para verificar se os perfis se mantêm ou evoluem.

A redefinição do problema não é necessária, uma vez que os objetivos propostos foram atingidos com sucesso. Contudo, há espaço para refinar e expandir a aplicação do modelo, mantendo o equilíbrio entre complexidade computacional e utilidade prática.

Fase 6: Colocação em Produção

6.1 Planeamento da colocação em produção

Com os modelos treinados, testados e avaliados, o passo seguinte foi torná-los realmente úteis - ou seja, garantir que podem ser usados de forma simples por qualquer pessoa envolvida na gestão ou análise de incêndios.

Para isso, foi desenvolvida uma aplicação interativa em *Streamlit*, que reúne todos os modelos criados ao longo do projeto, desde regressão até *clustering*. Esta aplicação permite inserir alguns dados manualmente (por exemplo, temperatura, duração do incêndio, FWI), e com base nisso, o sistema devolve previsões ou padrões relevantes.

A aplicação foi pensada para ser leve, acessível e adaptável:

- Os modelos já estão guardados e prontos a usar, sem necessidade de reprocessar tudo.
- A interface é clara e intuitiva, mesmo para quem não tem *background* técnico.
- Está preparada para ser usada localmente, em contexto académico, operacional ou numa eventual versão online segura.

Este planeamento assegura que o trabalho desenvolvido tem continuidade e utilidade prática, tornando-se numa ferramenta real de apoio à decisão, e não apenas num exercício teórico.

6.2 Planeamento da monitorização e manutenção

Colocar os modelos em produção é apenas o início. Para garantir que eles continuam a funcionar bem ao longo do tempo, é essencial definir um plano claro de monitorização e manutenção.

Com o passar dos meses, os dados podem mudar, os padrões de incêndios podem evoluir e, com isso, o desempenho do modelo pode diminuir. Por isso, será importante acompanhar periodicamente os resultados, avaliando se as previsões continuam fiáveis.

A manutenção pode incluir atualizações dos dados, reajustes nos modelos ou até o seu re-treinamento com novos registos. Tudo isso deve ser feito com cuidado e com registos adequados, para garantir confiança nas previsões e consistência no apoio à tomada de decisão.

Conclusão

Pode concluir-se que a implementação da metodologia CRISP-DM permitiu uma abordagem sistemática e bem fundamentada à análise dos incêndios florestais em Portugal, com base em dados extensos e detalhados disponibilizados pelo ICNF. A fase inicial do projecto possibilitou uma compreensão aprofundada da problemática, identificando os principais desafios enfrentados pelas entidades responsáveis: a redução de reacendimentos, a optimização do tempo de resposta, a mitigação de incêndios de grande escala e a identificação de territórios de maior risco.

A análise exploratória dos dados revelou a predominância de incêndios de pequena dimensão, mas também evidenciou episódios extremos com impacto significativo. Variáveis como o tipo e a causa do incêndio, as condições meteorológicas e a localização geográfica mostraram-se fundamentais para compreender a dinâmica dos eventos e orientar as fases subsequentes do projecto.

Foram desenvolvidos quatro modelos principais. O modelo de associação identificou padrões frequentes na combinação de variáveis, como determinadas causas associadas a áreas mais extensas, permitindo gerar regras úteis para a prevenção. No entanto alguns objetivos não conseguiram ser objetivos, devido à complexidade dos dataset e à qualidade de alguns atributos.

Por fim, pode-se concluir que alguns modelos desenvolvidos demonstraram utilidade prática contribuindo com *insights* relevantes, no entanto havia espaço para melhorias.

Este projecto demonstra que a aplicação de técnicas de *Data Mining* a dados históricos de qualidade tem um enorme potencial para melhorar a gestão e a prevenção dos incêndios florestais, com impactos positivos na segurança das populações, na protecção ambiental e na eficácia operacional das entidades envolvidas.

Bibliografia

Instituto da Conservação da Natureza e das Florestas (ICNF). (2025). *Sistema de Gestão de Informação de Incêndios Florestais*. Disponível em: https://www.icnf.pt

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide* [PDF]. CRISP-DM Consortium. Disponível em https://ead.ipleiria.pt/2024-25/pluginfile.php/172551/mod_resource/content/1/CRISP-DM_1.0.pdf

Instituto Politécnico de Leiria. (2025). *Plataforma Moodle – Unidade Curricular de Data Mining*. Disponível em https://ead.ipleiria.pt/2024-25/course/view.php?id=2796

Alves, G. S. (2023). *Utilização do CRISP-DM para criação de um modelo de previsão de NPS da central de atendimento de um grande banco brasileiro* (Trabalho de Conclusão de Curso de Graduação). Escola Politécnica, Universidade de São Paulo, São Paulo. Recuperado de https://bdta.abcd.usp.br/directbitstream/601ed9be-269a-4339-8b04-4d484c986053/GuilhermeStortoAlvesPRO23%20%281%29.pdf:contentReference[oaicite:2]{index=2}

Referências

[1] ECMWF, “User Guide,” Copernicus Emergency Management Service, 2024. [Online]. Available: https://confluence.ecmwf.int/display/CEMS/User+Guide

[2] Instituto Português do Mar e da Atmosfera (IPMA), “Boletim Técnico – Meteorologia Operacional,” IPMA, 2023. [Online]. Available: https://www.ipma.pt/export/sites/ipma/bin/docs/relatorios/meteorologia/nt-rcm.pdf

Anexos

Anexo I - Sistema de Gestão de Informação de Incêndios Florestais (SGIF)

Em 2001, foi estabelecido um protocolo entre o Centro Nacional de Informação Geográfica (CNIG) e a Direção-Geral das Florestas (DGF) com vista à criação de uma aplicação informática designada por Sistema de Gestão de Informação de Incêndios Florestais (SGIF), com o objetivo de gerir a base de dados nacional sobre incêndios rurais.

No mesmo ano, foi instalada em todos os Centros de Prevenção e Detecção (CPDs) uma aplicação que permitiu informatizar o processo de registo de dados de incêndios rurais de forma sistemática e uniforme e deu os primeiros passos para a transmissão digital dos dados para uma base de dados central única.

A estrutura de dados e o desenvolvimento da aplicação foram estabelecidos no início de 2001, tendo a instalação e formação do operador ocorrido no início da época de incêndios, em junho.

Desde então, o sistema tem sofrido várias evoluções à medida que as necessidades foram surgindo, impulsionadas pelos avanços tecnológicos e pelas reorganizações institucionais nos setores florestal e de proteção civil.

O resumo que se segue apresenta datas significativas na história do SGIF (Ano e Ações).

- 2001: Protocolo estabelecido entre o Centro Nacional de Informação Geográfica (CNIG) e a Direção-Geral das Florestas (DGF) para a criação do SGIF. Desenvolvimento da aplicação e base de dados central. Aquisição de computadores e modems. Instalação da aplicação SGIF em CPDs.
- 2002: Conversão da base de dados central para SQL Server. Abertura para importar dados diretamente da aplicação do Serviço Nacional de Bombeiros (SNB).
- 2003: Instalação de internet permanente nos CPD. Melhoria no processo de importação de dados para a base de dados central.
- 2006: Conversão da aplicação para uma aplicação web. A operação de registo passa a ser assegurada pela Guarda Nacional Republicana (GNR), sendo extintos os CPD e criadas as Equipas de Manutenção e Exploração de Informação Florestal (EMEIF). Os registos são agora criados pelo Comando Distrital de Operações de Socorro (CDOS) e importados diretamente para a base de dados central. Estudo realizado pela COTEC Portugal sobre vigilância com recurso a dados SGIF.
- 2009: Desenvolvimento de uma nova versão. O SGIF passa a ter dados meteorológicos associados a incêndios. O SGIF passa a controlar o acesso ao registo e a entrada no sistema. Módulos desenvolvidos para os Gabinetes

Técnicos Florestais (GTF). Desenvolvimento de um módulo estatístico para acesso público.

- 2011: Início dos estudos de tendências.
- 2012: Melhoria dos módulos de verificação da consistência dos dados. A localização das ocorrências passou a ser feita pela Autoridade Nacional de Emergência e Proteção Civil (ANEPC). Ajuste do conceito de reacendimento.
- 2013: Atualização da nova lista de paróquias. Protótipo criado para autorização de queimadas e queimadas controladas para acesso público.
- 2014: Módulos desenvolvidos para a gestão dos planos de Defesa Florestal Municipal e sensibilização. É desenvolvido um novo módulo para a gestão da atividade de fogo controlado.
- 2015: Versão atual do SGIF produzida. Iniciado o desenvolvimento de um portal de análise estatística.
- 2016: Revisão do módulo de avaliação de inconsistência de dados. É instituída a análise de locais críticos, com recurso a Mapas de Calor para distribuição de ocorrências.
- 2018: As áreas ardidas superiores a 10 hectares são digitalizadas por uma equipa do ICNF, sendo o preenchimento das áreas das parcelas feito automaticamente pelo perímetro da área. É instituída uma nova verificação para verificar se o ponto de partida se encontra dentro ou fora do perímetro da área ardida. Uma nova versão do sistema de registo de incêndios e incêndios controlados é disponibilizada ao público. O componente geográfico dos perímetros das áreas ardidas é armazenado numa base de dados do servidor SQL.
- 2021: A série de dados meteorológicos tem agora apenas uma fonte de dados, utilizando a série de reanálise global ERA5 do Centro Europeu de Previsões Meteorológicas a Médio Prazo (ECMWF).
- 2022: A codificação das ocorrências passa a ser a mesma no sistema SADO da ANEPC e do SGIF, adotando a codificação SADO. As listas de codificação para as causas e as listas de alerta para as ocorrências são revistas. São disponibilizados vários serviços de dados geográficos com informação SGIF.
- 2024: A série de dados meteorológicos foi alargada até 1980, utilizando a série de reanálise global ERA5 do Centro Europeu de Previsões Meteorológicas a Médio Prazo (ECMWF).

Marcos importantes:

No final de 2001, iniciou-se o processo que levou à conversão da base de dados central para uma base de dados SQL Server, que ainda hoje se encontra em utilização. Em 2006, foi necessário desenvolver uma nova versão devido às alterações de responsabilidades decorrentes da aplicação dos conceitos estabelecidos no Plano Nacional de Defesa Florestal contra Incêndios. Em 2009, foi lançado um módulo

estatístico de acesso público. Em 2018, houve uma integração muito maior entre os dados alfanuméricos e os dados geográficos, e foi lançada a primeira versão pública do sistema de registo de permissões de queima.

Onde se Encaixa o Projeto SGIF:

O SGIF fazia parte de um projeto mais vasto, iniciado em 1999 no National Geographic Information Center, denominado Emergency Situations Information Network (RISE). Os principais objectivos da Rede de Informação de Situações de Emergência eram dotar o país de uma rede electrónica de informação georreferenciada, proporcionando um fluxo automático e actualizado de dados relativos a situações de emergência. Um subconjunto destes dados foi disponibilizado ao público através do site do National Geographic Information System. O RISE era um serviço desenvolvido pelo Grupo de Cartografia de Risco de Incêndio Florestal (CRIF) do CNIG, em estreita colaboração com o Serviço Nacional de Bombeiros, o Serviço Nacional de Proteção Civil, a Direção-Geral de Florestas, o Instituto Geográfico do Exército e, aquando da dissolução do CNIG, estavam a ser estabelecidos acordos com a Direção-Geral de Trânsito e o Instituto Nacional de Emergência Médica. Através de um protocolo, foi formalizada a adesão das entidades geradoras de registos referentes a situações de emergência ao RISE, garantindo:

- Envio remoto de informação para a base de dados dessa instituição ou para a base de dados de outra instituição aderente ao presente protocolo.
- Consulta de informação cartográfica produzida pelo Instituto Geográfico do Exército, disponibilizada pelas aplicações RISE.
- A utilização da informação disponibilizada pelo RISE exige às instituições o cumprimento de um conjunto de condições, entre as quais a utilização exclusiva dos dados cartográficos a título estritamente de emergência ou para fins internos, sendo proibida a sua reprodução ou transferência para terceiros.
- Esta cooperação com outras instituições da rede resultou em benefícios em termos de partilha de experiências, redução de custos e otimização de benefícios.
- Houve uma troca de informação científica e técnica relacionada com a gestão de informação de emergência georreferenciada.
- Verificou-se um reforço da capacidade de intervenção nacional no desenvolvimento de projetos de cooperação internacional, o que foi reconhecido, por exemplo, pela Agência Espacial Europeia. Em 2001, esta agência abriu um processo de candidatura específico para apoiar o desenvolvimento desta rede, do qual resultou o projecto “PremFIRE” (CNIG, 2001).

O que é hoje o SGIF e como se mantém?

Hoje, o SGIF é um sistema que recolhe informação sobre incêndios rurais e gera dados estatísticos sobre o mesmo tema. É a base de dados de referência nacional para incêndios rurais. Após a criação do SGIF em 2001, a primeira grande mudança organizacional ocorreu em 2006. O SGIF tornou-se a base de dados de referência nacional para incêndios rurais, o que exigiu uma adaptação, evoluindo de uma aplicação de gestão de informação da DGF para um sistema que integra informação das três instituições-chave da Defesa Florestal Contra Incêndios (DFCI) (Autoridade Nacional de Proteção Civil - ANPC, Guarda Nacional Republicana - GNR e Autoridade Nacional Florestal (AFN). A segunda grande mudança organizacional ocorreu após 2017, com a introdução do Sistema Integrado de Gestão de Incêndios Rurais. Além disso, vários estudos científicos utilizaram dados do SGIF. A responsabilidade pela manutenção do SGIF cabe ao ICNF, mas desde a dissolução do CNIG em 2002, todos os desenvolvimentos foram realizados por Rui Almeida, João Moreira, Raquel Onofre e Pedro Venâncio.

O SGIF sofreu diversas melhorias ao longo do tempo. Em 2010, foi realizada uma análise completa dos dados. Este processo exigiu uma revisão exaustiva de toda a base de dados de incêndios rurais, durante a qual foram detectadas algumas incongruências, sendo necessária a republicação de séries estatísticas e de dados base.

Desde 2010, o sistema evoluiu para incluir procedimentos internos mais rigorosos para a verificação da consistência dos dados. Foram também implementados procedimentos para a associação de outras informações, como dados meteorológicos (valores observados e previstos).

Os dados foram novamente reavaliados em 2015 e 2022, com a decisão de utilizar uma nova série de dados meteorológicos para garantir a uniformidade na fonte de dados. Sempre que foi possível associar novos dados à informação de base sobre os incêndios rurais, tal permitiu o desenvolvimento de novos estudos, como o desta tese.

Um dos objetivos atuais do SGIF é expandir o conjunto de utilizadores que acedem à sua informação. Para isso, estão a ser desenvolvidos novos módulos para utilizadores não registados. Exemplos incluem parcerias que estão a ser desenvolvidas com a Pordata e o Instituto Nacional de Estatística, onde estão a ser criadas novas abordagens para a divulgação de informação.

A base de dados SQL do SGIF contém dados sistematicamente organizados sobre incêndios rurais desde 2001. Inclui também dados sobre outros temas relacionados com incêndios rurais, como a meteorologia, a orografia, as infraestruturas e o uso do solo. A tabela seguinte apresenta a lista e a descrição das variáveis associadas a cada incêndio rural.

Anexo II – Descrição dos dados

Nº	Nome da Variável	Tipo da variável	Descrição	Relevante	Motivo para exclusão
1	Codigo	Identificador	Código SGIF para o incêndio – código do Serviço Florestal.	Sim	____
2	CodigoSado	Numérico	Código SADO para o incêndio – código de proteção civil	Não	Apresenta valores omissos; Duplicação / redundância.
3	Ano	Numérico	Ano da data do alerta	Sim	____
4	Tipo	Nominal	Tipo de incêndio (Florestal ou Agrícola).	Sim	____
5	DHInicio	Data/Hora	Data e hora do alerta	Sim	____
6	DH1Intervencao	Data/Hora	Data e hora da primeira intervenção	Sim	____
7	DHResolucao	Data/Hora	Data e hora da resolução.	Sim	____
8	DHConclusao	Data/Hora	Data e hora de conclusão do evento (após extinção).	Sim	____
9	DHFim	Data/Hora	Data e hora de término final do incêndio.	Sim	____
10	DuracaoHoras	Numérico	Duração total do incêndio em horas.	Sim	____
11	Tempo1Intervencao	Numérico	Tempo (em horas) até a primeira intervenção.	Sim	____
12	TempoResolucao	Numérico	Tempo até a resolução do incêndio.	Sim	____
13	TempoRescaldo	Numérico	Tempo, em minutos, dedicado ao rescaldo após a resolução do incêndio.	Sim	___
14	Dia	Numérico	Dia em que o incêndio ocorreu.	Sim	____
15	Mes	Numérico	Mês em que o incêndio ocorreu.	Sim	____
16	Hora	Numérico	Hora do dia em que o incêndio começou.	Sim	____

17	AreaPov	Numérico	Área ardidade de povoamentos florestais, em hectares (ha).	Sim	____
18	AreaMato	Numérico	Área de mato ardida, em hectares (ha).	Sim	____
19	AreaAgric	Numérico	Área de terrenos agrícolas ardida, em hectares (ha).	Sim	____
20	AreaTotal	Numérico	Área total ardida, em hectares (ha).	Sim	____
21	HaHora	Numérico	Área ardida (em hectares) por hora de duração do incêndio.	Sim	___
22	ClasseArea	Nominal	Categoria da área total ardida (agrupada por classes predefinidas)	Sim	____
23	FonteAlerta	Nominal	Identifica a fonte do alerta de comunicação do incêndio às autoridades.	Sim	____
24	CodCausa	Numérico	Código numérico da causa do incêndio.	Não	Apresenta valores omissos; Redundância.
25	TipoCausa	Nominal	Tipo de causa do incêndio (ex.: humana, natural, etc.).	Sim	____
26	GrupoCausa	Nominal	Grupo específico da causa (ex.: incêndio por descuido, incêndio criminoso, etc.).	Sim	____
27	DescricaoCausa	Nominal	Descrição detalhada da causa específica do incêndio.	Não	Apresenta valores omissos; Redundância.
28	Reacendimento_IncendioPai	Nominal	Se houve reacendimento, código do incêndio pai.	Sim	____
29	OriginouReacendimento	Binário	Indica se o incêndio originou reacendimento (0 ou 1).	Sim	____
30	Distrito	Nominal	Distrito do ponto de origem do incêndio.	Sim	____
31	Concelho	Nominal	Concelho do ponto de origem do incêndio.	Sim	____
32	Freguesia	Nominal	Freguesia do ponto de origem do incêndio.	Não	Redundância.
33	Local	Nominal	Localização mais próxima ao ponto de origem do incêndio.	Sim	____
34	Nut2	Nominal	Código ou nome da unidade territorial de nível II, conforme classificação da União Europeia.	Sim	____

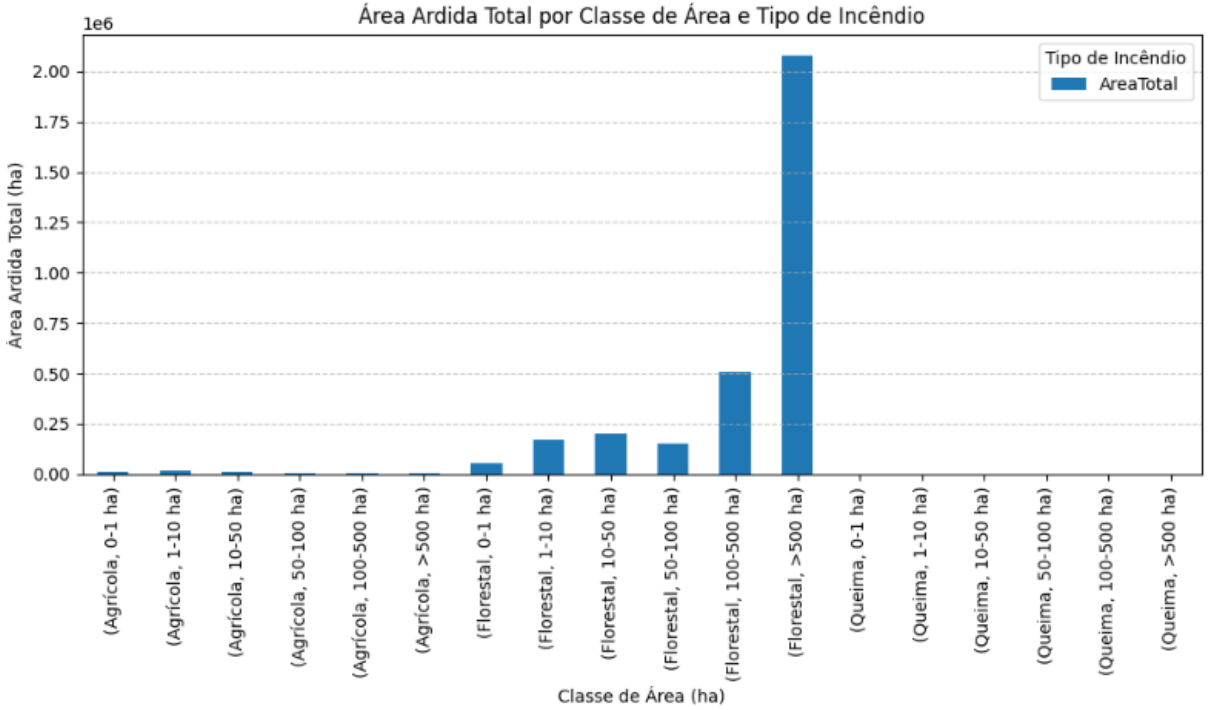
35	Nut3	Nominal	Nome da unidade territorial de nível III (NUTS III), que representa subdivisões regionais dentro das NUTS II	Não	Redundância.
36	INE	Nominal	Código do INE da freguesia no ponto de origem do incêndio.	Não	Redundância.
37	Perimetro	Categórica	Origem do Incêndio	Não	Apresenta valores omissos;
38	APS	Numérico	Área protegida (RNAP) no ponto de origem do incêndio.	Não	Irrelevante para os objetivos definidos;
39	x_20790	Numérico	Coordenada X em EPSG:20790 Lisboa.	Não	Irrelevante para os objetivos definidos;
40	y_20790	Numérico	Coordenada Y em EPSG:20790 Lisboa.	Não	Irrelevante para os objetivos definidos;
41	Lat_4326	Numérico	Latitude em EPSG:4326 WGS 84 -- Sistema Geodésico Mundial 1984.	Não	Irrelevante para os objetivos definidos;
42	Lon_4326	Numérico	Longitude em EPSG:4326 WGS 84 -- Sistema Geodésico Mundial 1984.	Não	Irrelevante para os objetivos definidos;
43	x_3763	Numérico	Coordenada X em EPSG:3763 ETRS89 / Portugal TM06.	Não	Irrelevante para os objetivos definidos;
44	y_3763	Numérico	Coordenada Y em ETRS89 (PT-TM06).	Não	Irrelevante para os objetivos definidos;
45	QO	Numérico	Código de qualidade operacional associado ao incêndio.	Não	Irrelevante para os objetivos definidos;
46	ModFarsite	Nominal	Modelo Farsite da área ardida no ponto de origem.	Não	Dependência de programas externos;
47	AreaManchaModFarsite	Numérico	Área ardida do modelo Farsite no ponto de origem (ha).	Não	Dependência de programas externos;
48	AltitudeMedia	Numérico	Altitude média da grelha 1x1 km no ponto de origem.	Sim	____
49	DecliveMedio	Numérico	Declive médio da grelha 1x1 km no ponto de origem (%)	Sim	____
50	HorasExposicaoMedia	Numérico	Num. médio de horas de exposição ao fogo para as áreas afetadas.	Sim	____
51	Rugosidade	Numérico	Medição da rugosidade do terreno.	Sim	____
52	DensidadeRV	Numérico	Densidade de vias na grelha 1x1 km (m/ha)	Sim	____

53	CosN5Variedade	Numérico	Número de classes de COS N5 diferentes na grelha 1x1 km.	Não	Irrelevante para os objetivos definidos;
54	Perigosidade	Numérico	Perigosidade estrutural.	Sim	____
55	Dist_CBS_m	Numérico	Distância em linha reta até a estação de bombeiros mais próxima (m).	Sim	___
56	CBS	Numérico	Estação de bombeiros mais próxima em linha reta.	Não	Redundância.
57	DensidadeResidentes	Numérico	Densidade de residentes na área afetada pelo incêndio.	Sim	___
58	DensidadeEdificios	Numérico	Densidade de edifícios na área afetada pelo incêndio.	Sim	___
59	Temperatura	Numérico	Temperatura do ambiente no momento do incêndio (em °C).	Sim	____
60	HumidadeRelativa	Numérico	Nível de umidade relativa do ar durante o incêndio.	Sim	____
61	VentoIntensidade	Numérico	Intensidade do vento durante o incêndio (em km/h).	Sim	____
62	VentoIntensidade_vetor	Numérico	Intensidade do vento associada ao vetor direcional.	Não	Redundância.
63	VentoDirecao_vetor	Nominal	Direção do vento associada ao vetor direcional.	Não	Redundância.
64	Precepitacao	Numérico	Quantidade de precipitação durante o incêndio em mm.	Sim	____
65	VentoDirecao	Numérico	Direção do vento durante o incêndio (em graus)	Sim	____
66	fwi	Numérico	Índice meteorológico.	Sim	___
67	dsr	Numérico	Índice meteorológico.	Sim	___
68	isi	Numérico	Índice meteorológico.	Sim	___
69	dc	Numérico	Índice meteorológico.	Sim	___
70	dmc	Numérico	Índice meteorológico.	Sim	___
71	ffmc	Numérico	Índice meteorológico.	Sim	___

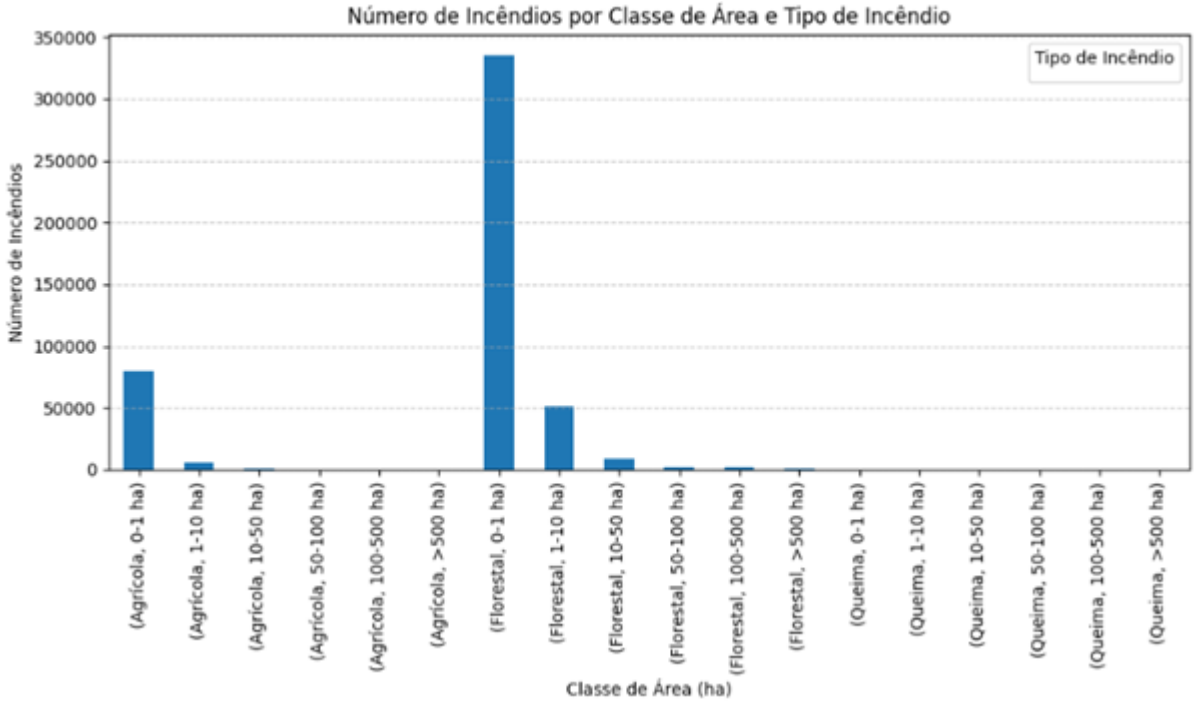
72	bui	Numérico	Índice meteorológico.	Sim	___
73	hFWI	Numérico	É o valor do FWI (Fire Weather Index) calculado para a hora exata em que foi emitido um alerta de incêndio.	Sim	___
74	hFFMC	Numérico	Valor na hora do alerta (HFFMC) – Fine Fuel Moisture Code.	Sim	___
75	hISI	Numérico	Valor na hora do alerta (HISI) – Initial Spread Index.	Sim	___
76	RCM	Numérico	RCM (Risco Conjuntural e Meteorológico)	Sim	___
77	MaxFWIh_48h_PosExtincao	Numérico	Máximo do Fire Weather Index nas últimas 48h após extinção.	Sim	___
78	MaxFFMCh_48h_PosExtincao	Numérico	Máximo do Fine Fuel Moisture Code nas últimas 48h após extinção.	Sim	___
79	MaxISIh_48h_PosExtincao	Numérico	Máximo do Initial Spread Index nas últimas 48h após extinção.	Sim	___
80	MaxDC_48h_DiaPosExtincao	Numérico	Máximo do Drought Code nas últimas 48h após extinção.	Sim	___
81	MaxBUI_48h_PosExtincao	Numérico	Máximo do Burning Index nas últimas 48h após extinção.	Sim	___
82	MaxDMC_48h_PosExtincao	Numérico	Máximo do Drought Moisture Code nas últimas 48h após extinção.	Sim	___
83	NIncSimul5000	Numérico	Número de incêndios simultâneos na área de 5000 metros.	Sim	___
84	DistIncSimul5000	Numérico	Distância de incêndios simultâneos a partir de 5000 metros.	Sim	___
85	AreaTotalIncSimul5000	Numérico	Área total de incêndios simultâneos na área de 5000 metros.	Sim	___
86	NIncSimulDistrito	Numérico	Número de incêndios simultâneos (total ou parcialmente) no mesmo distrito do incêndio.	Sim	___
87	NIncSimulConcelho	Numérico	Número de incêndios simultâneos (total ou parcialmente) no mesmo concelho do incêndio.	Sim	___
88	NIncSimulDistrito90	Numérico	Número de incêndios simultâneos (total ou parcialmente) nos primeiros 90 minutos do incêndio, no mesmo distrito.	Não	Redundância.

89	NIncSimulConcelho90	Numérico	Número de incêndios simultâneos (total ou parcialmente) nos primeiros 90 minutos do incêndio, no mesmo concelho.	Não	Redundância.
90	NIncSimul500090	Numérico	Número de incêndios simultâneos (total ou parcialmente) nos primeiros 90 minutos do incêndio, num raio de 5000 metros do ponto de origem.	Não	Redundância.
91	DistIncSimul500090	Numérico	Número de incêndios simultâneos (total ou parcialmente) nos primeiros 90 minutos do incêndio, num raio de 5000 metros no mesmo distrito.	Não	Redundância.
92	AreaTotalIncSimul500090	Numérico	Soma das áreas ardidas de todos os incêndios simultâneos nos primeiros 90 minutos do incêndio, num raio de 5000 metros no mesmo distrito.	Não	Redundância.
93	Fogacho	Binário	Indica a ocorrência de fogachos.	Sim	____
94	Agrícola	Binário	Indica se o incêndio foi classificado como.	Sim	____
95	Reacendimentos	Numérico	Número de reacendimentos no incêndio (se houver).	Sim	____
96	ClassificacaoRegisto	Nominal	Classificação do registo do incêndio.	Não	Irrelevante para os objetivos definidos;
97	EstadoRegisto	Nominal	Estado do registo do incêndio (ex.: ativo, finalizado).	Não	Irrelevante para os objetivos definidos;
98	DHFimEstimado	Data/Hora	Estimativa da data e hora de término do incêndio.	Não	Irrelevante para os objetivos definidos;
99	Observacoes	Nominal	Observações adicionais sobre o incêndio.	Não	Irrelevante para os objetivos definidos;

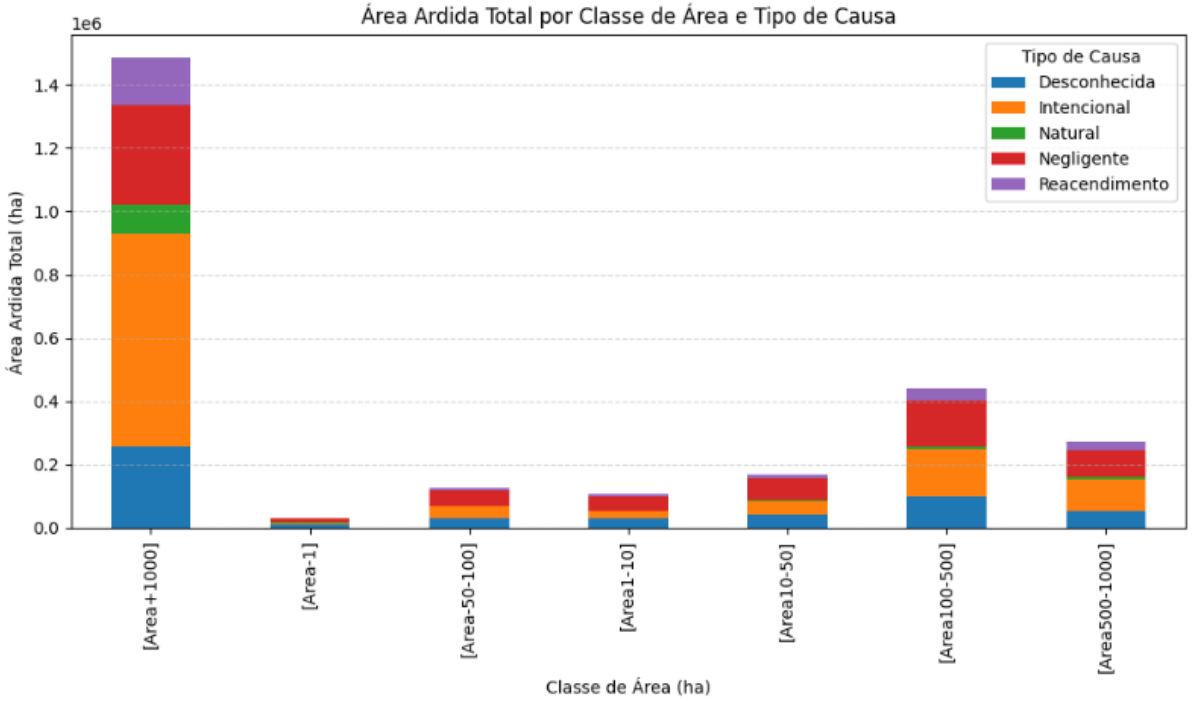
Anexo III - Gráfico Área Ardida Total por Classe de Área e Tipo de Incêndio



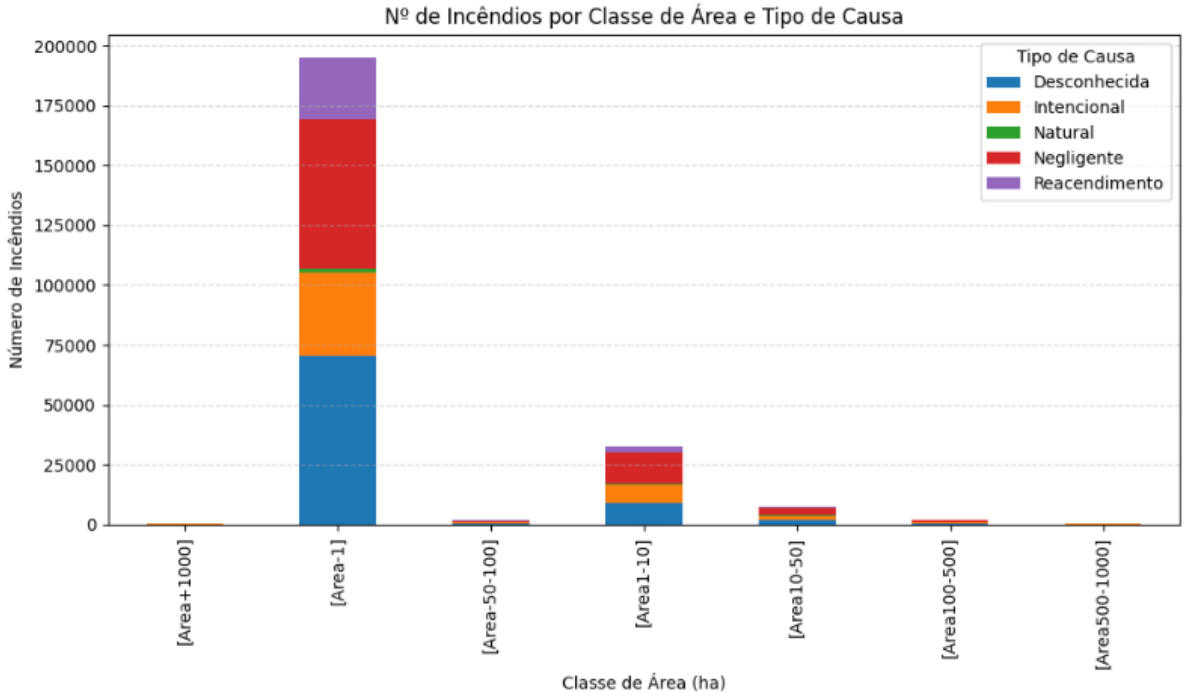
Anexo IV - Gráfico Número de Incêndios por Classe de Área e Tipo de Incêndio



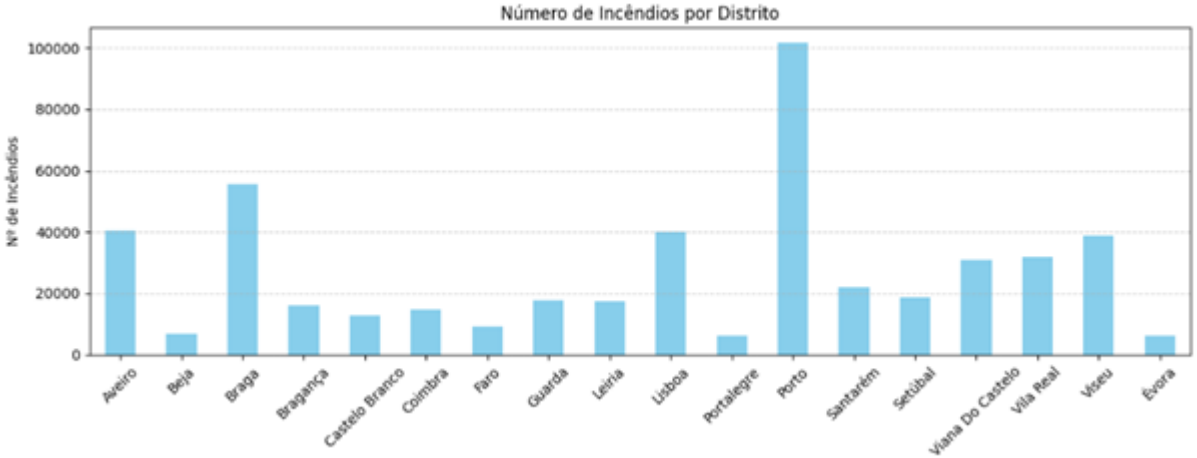
Anexo V - Gráfico Área Ardida Total por Classe de Área e Tipo de Causa



Anexo VI - Gráfico Número de Incêndios por Classe de Área e Tipo de Causa

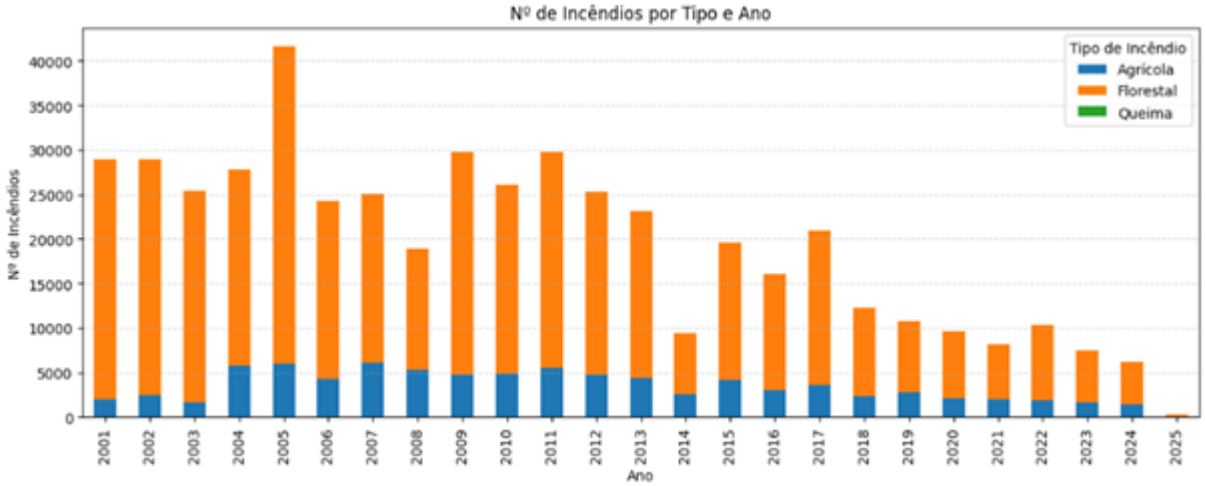


Anexo VII - Gráfico Número de Incêndios por Distrito

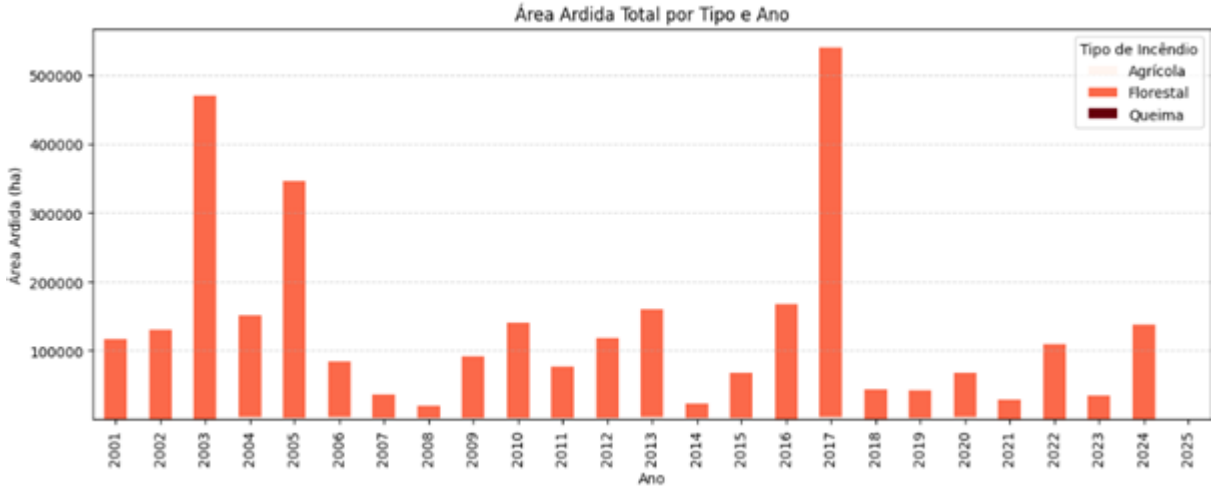


Anexo VIII - Gráfico Área Ardida Total por Distrito

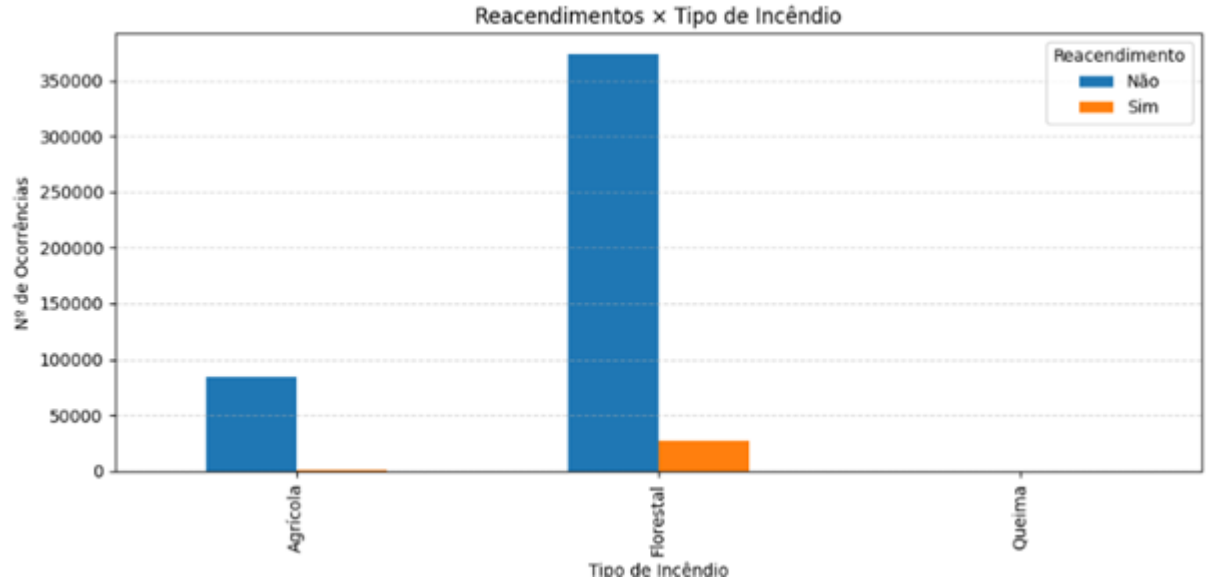
Anexo IX - Gráfico Número de Incêndios por Tipo e Ano



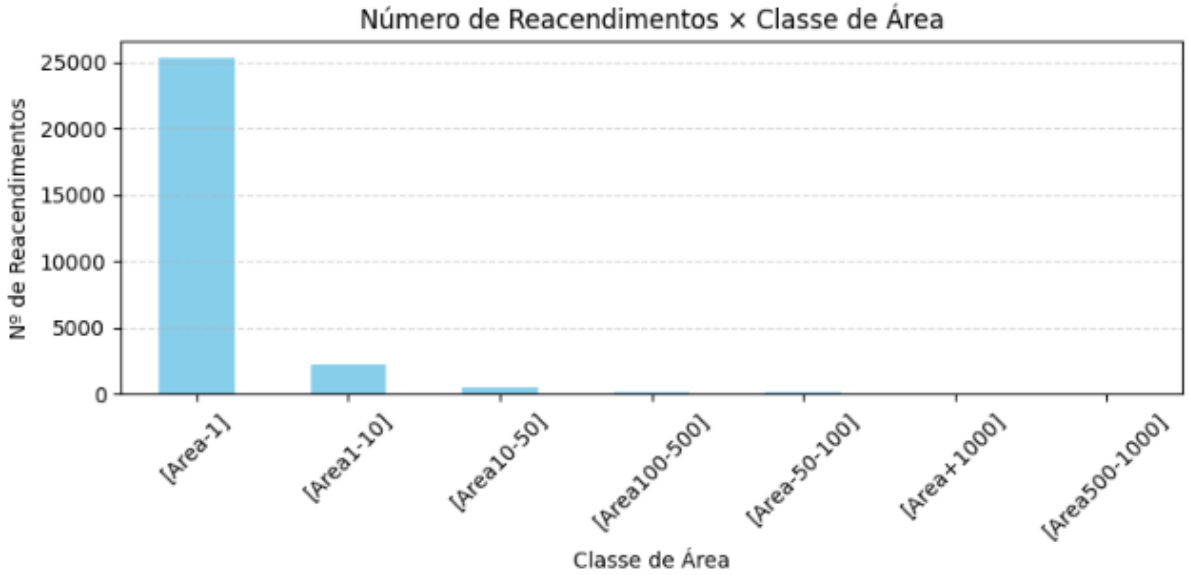
Anexo X - Gráfico Área Ardida Total por Tipo e Ano



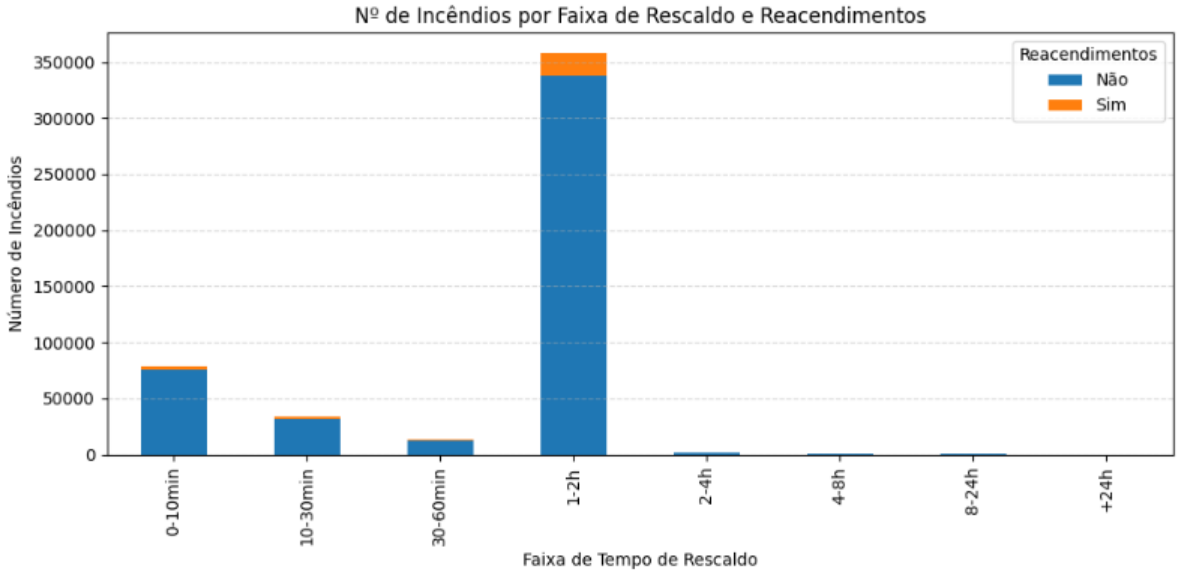
Anexo XI - Gráfico Reacendimentos x Tipos de Incêndio



Anexo XII - Gráfico Número de Reacendimentos x Classe de Área



Anexo XIII - Gráfico Número de Incêndios por Faixa de Rescaldo e Reacendimentos



Anexo XIV - Gráfico Área Ardida Total por Classe de FWI