Uma proposta de *Data Mining* para análise de dados referentes aos incêndios florestais ocorridos em Portugal

Universidade Fernando Pessoa

Paulo César de Almeida Júnior

Faculdade de Ciência e Tecnologia

Universidade Fernando Pessoa

Uma tese submetida para obtenção do grau de

MSc em Engenharia Informática

2024

Resumo

Incêndios florestais representam um desafio mundial expressivo, exigindo uma compreensão aprofundada de seus fatores desencadeantes para uma gestão eficaz. Este estudo aborda a necessidade de prevenção, detecção e supressão de incêndios, bem como a consideração das interações ecológicas envolvidas. A mineração de dados de incidentes históricos de incêndios florestais revela-se crucial para a previsão e a compreensão desses eventos.

Sendo assim a pesquisa se concentra na construção de modelos preditivos, baseados em algoritmos de aprendizado de máquina, um modelo de aprendizagem supervisionada, que relaciona variáveis independentes (como datas de ocorrências, localidades, duração, índices de severidade meteorológica e de perigo de incêndios e causas) com uma variável dependente (a classe de área ardida). Dois pontos fundamentais são abordados: uma análise exploratória de dados de incêndios ocorridos em Portugal entre 2011 e 2022 e a criação de um modelo preditivo para classificar a faixa de área ardida em registros históricos do conjunto de dados.

Os resultados revelaram *insights* significativos. Visto que a análise exploratória dos dados forneceu uma visão abrangente dos incêndios, identificando áreas suscetíveis e destacando o impacto da ação humana na ampliação desses incidentes. Os fatores meteorológicos, representados pelos índices de severidade meteorológica e risco de incêndio, demonstram uma associação direta com o aumento das ocorrências.

A pesquisa superou desafios iniciais, como o desbalanceamento de classes, por meio do método *Synthetic Minority Oversampling Technique (SMOTE)*, resultando em modelos de alta qualidade. O *Random Forest*, após o balanceamento das classes, emergiu como uma abordagem promissora, obtendo métricas de desempenho notáveis, incluindo uma accuracy de 96% e valores de F1-score consistentemente acima de 87%.

As várias análises e dados estatísticos gerados por esta pesquisa contribuem para a compreensão e a prevenção de incêndios florestais, com implicações práticas na gestão desses eventos. A capacidade de predição aprimorada e a identificação de fatores-chave oferecem uma base sólida para estratégias de prevenção e resposta mais eficazes.

Abstract

Forest fires represent a significant global challenge, demanding an in-depth understanding of their triggering factors for effective management. This study addresses the need for fire prevention, detection, and suppression, taking into consideration the involved ecological interactions. Data mining of historical forest fire incidents proves to be crucial for predicting and comprehending these events.

Therefore, the research focuses on building predictive models, based on machine learning algorithms, a supervised learning model, that relate independent variables (such as occurrence dates, locations, duration, meteorological severity indices, fire danger indices, and causes) to a dependent variable (the burned area class). Two key points are addressed: an exploratory data analysis of fire incidents that occurred in Portugal between 2011 and 2022 and the creation of a predictive model to classify the burned area range in historical records from the dataset.

The results have revealed significant insights. As the exploratory data analysis provided a comprehensive view of fires, identifying susceptible areas and highlighting the impact of human actions in amplifying these incidents. Meteorological factors, represented by meteorological severity and fire risk indices, demonstrate a direct association with the increase in occurrences.

The research has overcome initial challenges, such as class imbalance, through the Synthetic Minority Oversampling Technique (SMOTE) method, resulting in high-quality models. Random Forest, after class balancing, emerged as a promising approach, achieving notable performance metrics, including an accuracy of 96% and consistently F1-scores above 87%.

The various analyses and statistical data generated by this research contribute to the understanding and prevention of forest fires, with practical implications in the management of these events. Enhanced prediction capability and the identification of key factors provide a solid foundation for more effective prevention and response strategies.

Em primeiro lugar dedico a Deus a realização deste trabalho!

E a toda minha família pelo amor, apoio e carinho!

Agradecimentos

Em primeiro lugar a Deus por seu infinito amor e permitir que chegasse até aqui com saúde e coragem para concluir a realização deste trabalho!

A minha família por acreditarem, me incentivarem e me apoiarem durante todo percurso desta caminhada.

Aos meus orientadores, os professores Doutores Christophe Soares e José Manuel Torres, pela atenção, orientação e direcionamento necessários para realização deste trabalho.

Índice

Índice vi

Lista de Figuras ix

Lista de Tabelas xi

Acrónimos xii

1 Introdução 1

1.1 Motivação . . . 2
1.2 Objetivos . . . 2
1.3 Contribuições . . . 3
1.4 Estrutura do Documento . . . 4

2 Revisão da literatura 5

2.1 Incêndios Florestais . . . 5
2.2 Índices Meteorológicos de Risco de Incêndio . . . 6
2.2.1 Fire Weather Index (FWI) - índice canadiano de risco de incêndio florestal . . . 6
2.2.2 Estrutura do Índice Meteorológico de Incêndio FWI . . . 7
2.3 Ciência dos Dados e *Inteligência Artificial* (IA) . . . 9
2.4 *Inteligência Artificial* e *Aprendizagem de Máquina* . . . 10
2.4.1 *Inteligência Artificial* Clássica . . . 10
2.4.2 Inteligência Artificial Computacional . . . 11
2.4.3 *Aprendizagem de Máquina* . . . 11
2.4.4 Paradigmas de Aprendizagem . . . 11
2.5 Modelos Analíticos Descritivos e Preditivos . . . 12
2.5.1 *Análise Descritiva de Dados* (ADD) . . . 12
2.5.2 Mineração de Dados . . . 12
2.6 Trabalhos Relacionados e Discussão . . . 14
2.6.1 Trabalhos Relacionados . . . 14
2.6.2 Discussão . . . 15

3 Metodologia **19**
3.1 Introdução . . . 19
3.2 Fontes dos Dados Utilizadas . . . 20
3.3 Descrição Sumária dos Dados . . . 20
3.4 Pré-processamento dos Dados . . . 22
3.4.1 Seleção dos Dados . . . 23
3.4.2 Limpeza dos Dados . . . 24
3.4.3 Descritização e Normalização dos Dados . . . 25
3.5 Exploração dos Dados . . . 25

4 Análise dos Dados **26**
4.1 Descrição dos Dados . . . 26
4.2 Estatísticas Anuais . . . 27
4.2.1 Quantidade de Incêndios por Tipo de Área em Portugal . . . 27
4.2.2 Áreas de Incêndios por Ano . . . 29
4.3 Dimensão dos Incêndios . . . 31
4.4 Análise das Causas . . . 34
4.4.1 Incêndios Rurais por TipoCausa . . . 34
4.4.2 Incêndios Rurais por GrupoCausa . . . 36
4.5 Análises Regionais . . . 37
4.5.1 Quantidade de Incêndios Rurais por Distritos e Respectivas Extensões de Áreas Ardidas (Ordem Alfabética) . . . 37
4.5.2 Área Ardida por Distritos . . . 39
4.5.3 Quantidade de Incêndios por Concelhos . . . 40
4.5.4 Área Ardida por Concelhos . . . 42
4.6 Análises Mensais . . . 43
4.6.1 Quantidade de Incêndios por Tipo de Área em Cada Mês . . . 44
4.6.2 Distribuição de Áreas Ardidas por Mês . . . 45
4.7 Análise da Severidade Meteorológica . . . 46
4.7.1 Número de Incêndios Rurais Anuais por Classe de Severidade Meteorológica 2011-2022 (DSR) . . . 46
4.7.2 Número de Incêndios Rurais Mensais por Classe de Severidade Meteorológica Janeiro - Dezembro . . . 48
4.8 Análise do Perigo de Incêndio Rural FWI . . . 49
4.8.1 Número de Incêndios Rurais Anuais por Classe de Perigo de Incêndio Rural . . . 49
4.8.2 Número de Incêndios Rurais Mensais por Classe de Perigo de Incêndio Rural . . . 50
4.9 Correlação entre Variáveis . . . 51

4.9.1 Verificando a Correlação com ClasseArea . . . 51
4.9.2 Verificando a Correlação com a Área Total Ardida . . . 52
4.10 Discussão . . . 52
4.10.1 Contextualização dos Resultados . . . 53
4.10.2 Relação com a Literatura . . . 54
4.10.3 Fatores Influenciadores . . . 54
4.10.4 Implicações para a Gestão e Prevenção: . . . 55
4.10.5 Limitações e Futuras Pesquisas . . . 55

5 *Machine Learning*: Modelagem Preditiva dos Incêndios Florestais 56
5.1 Tratamento Inicial do *Dataframe* . . . 56
5.1.1 Tratamento dos Dados Faltantes . . . 56
5.1.2 Tratamento dos Dados Inconsistentes . . . 57
5.1.3 Tratamento das Datas . . . 58
5.1.4 Codificação dos Dados . . . 58
5.1.5 Conferida na Duração dos Incêndios . . . 58
5.1.6 Verificando a Distribuição dos Registros da Classe . . . 58
5.2 Prever a Classe Área . . . 59
5.2.1 Conjunto de Treinamento e Teste . . . 60
5.3 Aprendizagem de Máquina . . . 60
5.3.1 Algoritmos de *Regressão Logística* e *Redes Neurais* . . . 60
5.3.2 *Aprendizagem Bayesiana* . . . 61
5.3.3 *Árvore de Decisão* . . . 62
5.3.4 *Random Forest* . . . 63
5.3.5 Conclusão do Tópico . . . 64
5.4 Abordagens para Tratar o Problema do Desbalanceamento de Dados . . . 64
5.4.1 *Regressão Logística* para Dados Balanceados com *Smote* . . . 65
5.4.2 *Aprendizagem Bayesiana* para Dados Balanceados com *Smote* . . . 66
5.4.3 *Árvore de Decisão* para Dados Balanceados com *Smote* . . . 67
5.4.4 *Random Forest* para Dados Balanceados com *Smote* . . . 69
5.5 Discussão . . . 70
5.5.1 Tratamento do Desbalanceamento de Dados: . . . 70
5.5.2 Impacto das Técnicas de Amostragem: . . . 70
5.5.3 Eficácia dos Algoritmos: . . . 70
5.5.4 Limitações e Sugestões para Trabalhos Futuros: . . . 70

6 Conclusão 72

Referências 74

Lista de Figuras

2.1 Estrutura do Índice Meteorológico de Incêndio FWI (*Fire Weather Index*). Fonte (IPMA, 2023b) . . . 7
2.2 Processo de *Knowledge Discovery in Databases*. Adaptado de (Castro and Ferrari, 2016) . . . 13

3.1 Geração do *dataframe* a partir da biblioteca pd do pandas . . . 21
3.2 Código para Inspecionar Dados Faltantes. . . . 24
3.3 Pré processamento para tratar dados faltantes . . . 24

4.1 Incêndios por tipo de área afetada 2011 - 2022 . . . 28
4.2 Incêndios rurais por ano . . . 29
4.3 Comportamento dos incêndios ao longo dos anos . . . 29
4.4 Áreas afetadas pelos incêndios por ano (tipos de área) . . . 31
4.5 Evolução da área total afetada pelos incêndios ao longo dos anos . . . 31
4.6 Incêndios por Classe . . . 32
4.7 Quantidade de incêndios por TipoCausa . . . 35
4.8 Distribuição dos tipos de causas de incêndios por Distrito . . . 36
4.9 Quantidade de incêndios por GrupoCausa . . . 36
4.10 Quantidade de incêndios por Distrito em ordem decrescente . . . 38
4.11 Lista Top10 da Evolução dos Incêndios quanto a quantidade por Distrito . 38
4.12 Concentração dos focos de incêndios nos Distritos . . . 39
4.13 Área ardida por Distrito . . . 39
4.14 Evolução das Áreas ardidas por Distritos . . . 40
4.15 Quantidade de incêndios por Concelhos da lista Top20 . . . 41
4.16 Área ardida nos concelhos Concelhos da lista Top20 . . . 41
4.17 Os vinte Concelhos que mais arderam . . . 42
4.18 Quantidade de incêndios nos Concelhos que mais arderam . . . 43
4.19 Incêndios por mês . . . 43
4.20 Gráfico de Incêndios por Mês . . . 44
4.21 Gráfico da quantidade de tipos de incêndios por mês . . . 45
4.22 Gráfico dos tipos de áreas dos incêndios por mês . . . 46

4.23 Quantidade de Incêndios por Classe de Severidade Durante os Anos . . . 47
4.24 Quantidade de Incêndios por Classe de Severidade Durante os Meses . . . 48
4.25 Quantidade de Incêndios por Classe de Perigo de Incêndio Rural Durante Anos . . . 49
4.26 Quantidade de Incêndios por Classe de Perigo de Incêndio Rural Durante Meses . . . 50
4.27 Correlação com a variável ClasseArea . . . 51
4.28 Correlação com a variável área total ardida . . . 52

5.1 Quantidade de registros por classe de área categórico . . . 59
5.2 Base de dados de treinamento e teste . . . 60
5.3 Matriz de Confusão: *Aprendizagem Bayesiana* . . . 61
5.4 Matriz de Confusão: *Árvore de Decisão* . . . 62
5.5 Matriz de Confusão: *Random Forest* . . . 63
5.6 Base de dados de treinamento e teste após smote . . . 65
5.7 Matriz de Confusão: *Regressão Logística Oversampling* . . . 66
5.8 Matriz de Confusão: *Aprendizagem Bayesiana Oversampling* . . . 67
5.9 Matriz de Confusão: *Árvore de Decisão* Oversampling . . . 68
5.10 Matriz de Confusão: *Random Forest Oversampling* . . . 69

Lista de Tabelas

4.1 Dataset Processado 27
4.2 Tabela: Quantidade de incêndios por tipo de área 28
4.3 Área ardida pelos incêndios por tipo de área 30
4.4 Tabela de Classes de Área por Ano 32
4.5 Tabela Contendo os Vinte Incêndios Rurais de Maior Dimensão 33
4.6 Quantidade de incêndios por tipo de causa 34
4.7 Tabela de Incêndios por Distrito e Causa 35
4.8 Tabela de Incêndios Rurais e Áreas Correspondentes 37
4.9 Incêndios Rurais e Áreas Correspondentes 40
4.10 Área ardida nos Concelhos 42
4.11 Quantidade de incêndios rurais por tipo de área em cada mês 44
4.12 Tabela de áreas ardidas por mês 45
4.13 Escala de *Daily Severity Rating* (DSR) por ano 47
4.14 Escala de DSR por mês 48
4.15 Escala de FWI por ano 49
4.16 Escala de FWI por Mês 50
4.17 Resumo de padrões identificados 53

5.1 Quantidade de incêndios por Distritos: registros padronizados 57
5.2 Quantidade de Incêndios por Faixas de Área 59
5.3 Resultados das Métricas de Classificação para *Aprendizagem Bayesiana* . 62
5.4 Resultados das Métricas de Classificação para *Árvore de Decisão* 63
5.5 Resultados das Métricas de Classificação para *Random Forest* 64
5.6 Resultados das Métricas de Classificação *Regressão Logística* com *Oversampling* 66
5.7 Resultados das Métricas de Classificação para *Aprendizagem Bayesiana* com *Oversampling* 67
5.8 Resultados das Métricas de Classificação *Árvore de Decisão* com Oversampling 68
5.9 Resultados das Métricas de Classificação para *Random Forest Oversampling* 69

Acrónimos

AB *Aprendizagem Bayesiana*

AD *Árvore de Decisão*

ADD *Análise Descritiva de Dados*

AE *Algoritmos Evolutivos*

AM *Aprendizagem de Máquina*

AUC *Area Under Receiver Operating Characteristic*

AUROC *Area Under Receiver Operating Characteristic*

BD *Base de Dados*

BUI *Buildup Index*

CFFWIS *Canadian Forest Fire Weather Index System (mesmo que FWI)*

CRISP-DM *Cross-Industry Standard Process for Data Mining*

DC *Drought Code*

DM *Data Mining*

DMC *Duff Moisture Code*

DSR *Daily Severity Rating*

DT *Decision Trees*

DW *Data Warehouse*

FFMC *Fine Fuel Moisture Content*

FR *Floresta Randômica*

FS *Fuzzy Systems*

FWI *Fire Weather Index*

ha *hectares*

IA *Inteligência Artificial*

IC *Inteligência Computacional*

ICNF *Instituto da Conservação da Natureza e das Florestas*

ID *Identificador Único*

IMISC *Índice Meteorológico de Incêndio do Sistema Canadiano*

IPMA *Instituto Português do Mar e da Atmosfera*

ISI *Initial Spread Index*

KDD *Knowledge Discovery in Databases*

KNN *k-Nearest Neighbor*

MD *Mineração de Dados*

ML *Machine Learning*

RF *Random Forest*

RL *Regressão Logística*

RNAP *Rede Nacional de Áreas Protegidas*

RNMNPF *Rede Nacional de Matas Nacionais e Perímetros Florestais*

RN *Redes Neurais*

RNA *Redes Neurais Artificiais*

SGIF *Sistema de Gestão de Informação de Incêndios Florestais*

SMOTE *Synthetic Minority Over-sampling Technique*

SN *Sistemas Nebulosos*

SVM *Support Vector Machines*

Capítulo 1

Introdução

Os incêndios florestais podem ter ignição naturalmente a partir de fenômenos naturais ou combustão provocada por ação humana, por descuido ou por maneira intencional (Nunes, 2009). Esse fenômeno tem se tornado constante e intenso, levando à destruição comunidades e ecossistemas por onde se espalham (Programme, 2022) e afetando milhões de hectares todos os anos (Bem, 2017). De acordo com Teodoro esse fenômeno é reconhecido como um dos eventos mais críticos na mudança global (Teodoro and Duarte, 2013).

De Rigo identifica que Espanha, Portugal e Turquia são os países mais propensos aos incêndios florestais (De Rigo et al., 2017). Portugal Continental é um país mediterrânico, singularmente conhecido pela reincidência de incêndios florestais em termos de áreas ardidas, perdas e danos (Farinha et al., 2022) com sérias implicações em nível social, económico e ecológico (Ramos et al., 2023).

Freitas ressalta que a compreensão do comportamento do fogo e a antecipação dos fatores que influenciam a sua ocorrência são aspectos muito importantes para o seu controle (Freitas, 2021). A gestão bem-sucedida depende da prevenção, detecção e pré-supressão eficazes, não obstante uma capacidade adequada de supressão e da consideração das relações ecológicas do fogo também se fazem necessárias (Teodoro and Duarte, 2013). Dessa forma Wood defende a prática da mineração de dados eficiente e penetrante de conjuntos de dados de incidentes históricos de incêndios florestais (Wood, 2021), corroborando com Finney que reafirma ser essencial para melhorar nossa capacidade de predizer e compreender melhor os fatores que influenciam a quantidade total de áreas que podem sofrer queimadas sob condições específicas (Finney et al., 2013) assim como modelagem em tempo corrente da propagação de incêndios florestais (de Gennaro et al., 2017).

Na previsão de incêndios, modelos estatísticos e algoritmos de aprendizado de máquina fundamentam-se na relação entre as variáveis independentes (clima, topografia, tipo de vegetação, etc.) e uma variável de resposta (quantidade de área queimada, número de focos, etc.), ou dependente.

Para auxiliar na prevenção de ocorrências e consequentemente minimização dos impactos ocasionados pelos incêndios, este trabalho é uma proposta de *Data Mining* para

análise de dados referentes aos incêndios ocorridos em Portugal. Esses dados são referentes a registros reais de incêndios oriundos do *Instituto da Conservação da Natureza e das Florestas* (ICNF) que aconteceram entre 2011 e 2022. O objetivo é a realização de uma análise exploratória sobre esses dados para obtenção de *insights* importantes para compreender os fatores relacionados ao evento contribuindo assim para a tomada de decisão e a geração de um modelo preditivo que seja capaz de classificar a qual faixa de área ardida pertence um registro submetido ao classificador com um certo grau de precisão a partir de registros históricos da base de dados.

1.1 Motivação

A realização deste trabalho foi motivada por um desejo de poder agregar e contribuir na minimização dos impactos ocasionados pelos incêndios florestais em Portugal alicerçado pela *Mineração de Dados*. Empregando uma abordagem *Data Mining* associada à *Inteligência Artificial* para análise, exploração de dados para compreensão do fenômeno é possível explorar grande quantidade de dados para extrair informações valiosas e relevantes para a tomada de decisões. Essas técnicas permitem que padrões e tendências sejam identificados em dados de maneira mais rápida e precisa do que seria possível com métodos tradicionais de análise de dados. Dessa forma este estudo pode acrescentar a outros estudos relacionados a incêndios florestais em Portugal ao apresentar estatísticas e *insights* valiosos para compreensão dos incêndios em todo o território Português sem se limitar a apenas regiões específicas do País, e ao adotar uma abordagem de aprendizagem de máquina para prever a faixa de área total ardida por um incêndio a partir de várias técnicas de aprendizagem de máquina. O fato de poder somar forças nas estratégias de combate aos incêndios apoiado pelo uso de ferramentas da tecnologia nessa pesquisa proporciona grande contentamento pela sua realização.

1.2 Objetivos

O objetivo deste estudo é utilizar a mineração e a modelagem de dados para melhorar nossa capacidade de predizer e de compreender os fatores que influenciam os incêndios florestais em Portugal, permitindo assim uma abordagem mais eficiente na prevenção e no controle desses eventos. Dessa forma, a estratégia é realizar o tratamento adequado dos dados do ICNF, utilizando técnicas de mineração de dados, obtenção de *insights* úteis e geração de um modelo preditivo capaz de prever a faixa de área total ardida por um registro com métricas confiáveis de acerto.

1.3 Contribuições

A pesquisa pode contribuir para:

- Este trabalho contribuiu para área de *Data Mining* numa prespectiva de Análise Descritiva: a *Análise Descritiva de Dados* ressaltou a importância das ferramentas de exploração e visualização de dados para a obtenção de estatísticas e *insights* úteis presentes nos dados, podendo servir também para outras áreas que explorem o comportamento dos dados a partir de conjuntos massivos de dados. A amostragem dos dados com o *oversampling* (*Synthetic Minority Over-sampling Technique* - SMOTE), reverteu o desbalanceamento da base e melhorou o resultado dos algoritmos de *Machine Learning*, o que pode ser útil em estudos de *Mineração de Dados* que possuem dados desbalanceados.

- Este trabalho contribuiu para a área de IA numa prespectiva de Análise Predictiva: ao aplicar técnicas de aprendizagem supervisionada aos dados históricos, a pesquisa fornece um modelo capaz de prever a faixa de área total ardida de um registro submetido ao classificador, o que pode ser útil em outras áreas que também envolvam previsões baseadas em dados.

- Para a prevenção e o controle de incêndios em Portugal: ao obter *insights* valiosos com base nos dados analisados e ajudando a compreender os fatores que influenciam os incêndios, pode ajudar na adoção de medidas mais eficientes para prevenir e controlar os incêndios em Portugal.

- Para a sociedade: a pesquisa pode trazer benefícios significativos para a sociedade, ao ajudar a minimizar os impactos ocasionados pelos incêndios em Portugal, contribuir para a redução de perdas materiais e de vidas humanas, bem como a preservação do meio ambiente.

Em resumo, o trabalho pode auxiliar na prevenção de incêndios e consequentemente na minimização dos impactos ocasionados por esse fenômeno, tendo em vista os diversos prejuízos já causados nos locais onde tiveram combustão. Ao prever em qual faixa de área total ardida um registro se enquadrada por meio da modelagem de dados históricos, os combatentes podem ter uma ideia da severidade e até mesmo se antecipar em ações estratégicas mais eficazes para prevenir ou minimizar os danos causados pelo incêndio. Portanto montar uma análise exploratória de dados e um modelo que possa nos dizer a faixa de área total ardida por um registro poderia somar forças às diversas estratégias de prevenção e combate auxiliando os combatentes na tarefa de reduzir novas ocorrências minimizando assim os impactos ocasionados.

1.4 Estrutura do Documento

Este documento está subdividido em seis capítulos diferentes: Introdução, Revisão da literatura, Metodologia, Análise de dados, *Machine Learning*: Modelagem Preditiva dos Incêndios Florestais e Conclusão. O primeiro Capítulo apresenta o tema da dissertação, juntamente com as principais motivações para o desenvolvimento deste trabalho, contribuições, objetivos e a descrição do problema a ser resolvido. O Capítulo seguinte realiza uma análise do trabalho relacionada aos incêndios florestais e o uso de *Data Mining* neste tipo de cenário. O terceiro Capítulo foca na metodologia utilizada, com uma pequena introdução, apresentação da metodologia, fonte de dados utilizados e uma descrição sumária dos dados utilizados. Além disso, no quarto é apresentada a análise dos dados, tratando diretamente da análise exploratória sobre os dados utilizados. No quinto Capítulo *Machine Learning* é tratado da modelagem preditiva dos incêndios florestais com aplicação dos algoritmos de *Machine Learning* para geração do modelo preditivo. Já o sexto e último Capítulo conclui esta dissertação, deixando abertas as possibilidades para desenvolvimentos futuros no tema da aplicação de *Data Mining* a incêndios florestais.

Capítulo 2

Revisão da literatura

2.1 Incêndios Florestais

Conforme Carvalho, a relação que o homem tem tido com a floresta ao longo da história não tem sido de convivência harmoniosa, pois, de acordo com ele, as pessoas enxergavam os espaços florestais como um entrave ao desenvolvimento econômico e, por isto, acreditavam que a sua destruição seria a alternativa para obtenção de mais área para a pastorícia e agricultura (Carvalho, 2006). Esse pensamento negativo também esteve presente na Europa, resultando em uma destruição significativa de florestas durante os períodos de frequentes revoltas e de turbulência social como nos acontecimentos após a Revolução de Abril. Mesmo sabendo da sua importância em escala global, ainda assim a floresta continua a ser delapidada pelo homem que utiliza continuamente a "ferramenta" mais poderosa e de fácil utilização que tem à sua disposição: o fogo (Carvalho, 2006) .

Incêndios florestais são um problema mundial, queimando áreas na ordem de milhões de *hectares* todos os anos (Bem, 2017), sendo reconhecido como um dos eventos mais críticos na mudança global (Teodoro and Duarte, 2013).

Incêndios quando ocorrem de forma natural, como acontece em florestas subtropicais estacionais, e sazonalmente secas, tem significativa importância, pois contribuem para a manutenção da biodiversidade (Pausas and Keeley, 2019) e também para evolução de alguns ecossistemas (Carvalho, 2006). Todavia, quando a ignição é provocada pela ação humana, ocorre um drástico aumento na sua incidência, prejudicando a capacidade de recuperação natural dos ecossistemas e ampliando irreversivelmente os efeitos danosos sobre o meio ambiente (Carvalho, 2006). Aliás este tipo de ocorrência também pode provocar uma série de outros prejuízos como: alteração da paisagem afetando a floração e a frutificação, provocando desastres ecológicos, como atenuação (degradação) da qualidade da água, do solo e da vegetação, ocasionando ainda perda de vidas humanas e de animais, problemas respiratórios associados às fumaças emitidas, entre outros (Pourtaghi et al., 2016). De acordo com De Rigo, na Europa os países mais suscetíveis à ocorrência de

incêndios são Espanha, Portugal e Turquia (De Rigo et al., 2017). Enquanto que Grécia, a região costeira dos Balcãs, regiões centrais e sul da Itália e sul da França apresentam um risco menor. Portugal é atingindo todos os anos por um número elevado de ocorrências que, conforme Pinto, a destruição causada é equiparada a uma catástrofe, haja vista a devastação das florestas, das habitações, ao ponto de colocar em risco a vida de muitos seres, dentre os quais o ser humano (Pinto, 2020). Em 2017, em Portugal, os incêndios resultaram em seu pior cenário registrado um valor recorde em relação à área total queimada cerca de (540.000 ha) desde 1980, contabilizando um trágico número de 114 óbitos provenientes dos eventos de junho e outubro desse ano, sendo considerado pelo setor de seguros local o desastre natural mais caro com prejuízos pagos em mais de 295 milhões de dólares pelo setor de seguros (Ramos et al., 2023). Para entender melhor os incêndios, a análise partiu de uma perspectiva composta na qual se considerou a influência de fatores meteorológicos nos quais uma seca prolongada ocasionou um estresse hídrico cumulativo pré-condicionado da vegetação de outubro de 2017, a passagem do furacão Ophelia e o agente humano causador de uma elevada quantidade de ignições negligentes possibilitaram entender melhor o cenário de 2017 (Ramos et al., 2023).

Vale ressaltar que, de acordo com Castro o excessivo acumulo de combustível ocasionado pela concentração de biomassa lenhosa nos povoamentos florestais e na paisagem é outro fator que formou as condições para ocorrência de grandes incêndios, intensos, e potencialmente catastróficos (Castro Rego et al., 2020a).

Dessa forma, a compreensão do comportamento do fogo, dos fatores que originam sua ocorrência e a antecipação na prevenção dos fatores que influenciam a sua ocorrência são aspectos muito importantes para o seu manejo (Freitas, 2021). Hoje com disponibilidade de conjuntos massivos de dados, maior poder de processamento dos computadores propiciado pela evolução contínua da tecnologia tem tornado a mineração de dados uma realidade cada vez mais essencial para auxiliar nas medidas de prevenção e combate a incêndios.

2.2 Índices Meteorológicos de Risco de Incêndio

A influência dos elementos climáticos no surgimento de eventos críticos de fogo é devidamente avaliada através de índices meteorológicos de risco de incêndio que estimam a possibilidade de um incidente de incêndio (Finney, 2005).

2.2.1 Fire Weather Index (FWI) - índice canadiano de risco de incêndio florestal

O Serviço Canadiano de Florestas elaborou o *Fire Weather Index* (FWI), também designado como CFFWIS, com o objetivo de avaliar o risco de incêndio a partir das condições

dos diversos materiais inflamáveis existentes no solo e também por elementos meteorológicos obtidos indiretamente (Pinto, 2020). O FWI corresponde ao índice de perigo de incêndio rural o qual integra seis índices que levam em conta os efeitos da humidade do combustível e a influência do vento no comportamento do fogo (IPMA, 2023a). De acordo com Ramos, essa é uma das metodologias de avaliação de incêndios mais confiáveis e aplicadas globalmente ao FWI, dada a importância à Europa Mediterrânica que, desde o século passado, utiliza-o para avaliar o perigo de incêndios meteorológicos em seus ecossistemas (Ramos et al., 2023). O cálculo desses índices é realizado com base nos seguintes parâmetros: valores de temperatura e humidade relativa do ar a 2 metros acima do solo, intensidade do vento a 10 metros acima do solo e precipitação acumulada em 24 horas (IPMA, 2023b). Dos seis componentes que integram o sistema os três primeiros aferem a humidade do combustível, representados através de códigos que são classificações numéricas do teor de humidade do solo da floresta e outras matérias orgânicas mortas. Já os três restantes são índices de comportamento do fogo que correspondem à taxa de propagação do fogo, a quantidade de combustível disponível para queima e a intensidade do fogo frontal, conforme esses valores aumentam, aumenta também o perigo de incêndio (Natural Resources Canada, 2023).

2.2.2 Estrutura do Índice Meteorológico de Incêndio FWI

No diagrama esquemático 2.1, estão ilustrados os componentes do Índice Meteorológico de Incêndio FWI. A partir de observações diárias e consecutivas de temperatura, humidade relativa, velocidade do vento e precipitação de 24 horas é que são calculados os valores desses componentes (Natural Resources Canada, 2023).

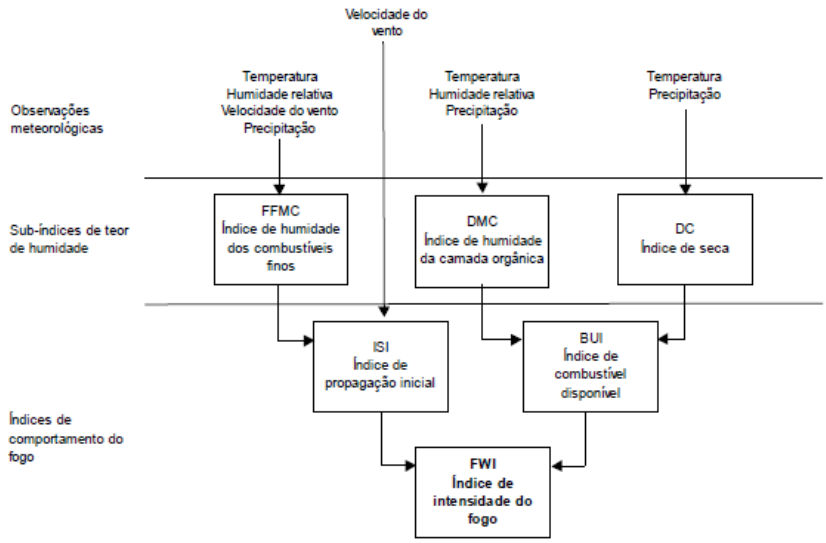


Figura 2.1: Estrutura do Índice Meteorológico de Incêndio FWI (*Fire Weather Index*). Fonte (IPMA, 2023b)

Índice de Humidade dos Combustíveis Finos (FFMC)

O Índice de Humidade dos Combustíveis Finos (*Fine Fuel Moisture Content* - FFMC) corresponde à quantidade de humidade presente nos combustíveis finos (0.25kg/m2) mortos que se encontram na camada superficial do solo, atingindo até 2 cm de profundidade e representam as condições meteorológicas dos últimos 2 a 3 dias (IPMA, 2023b). Este índice é um indicador da facilidade de ignição e da inflamabilidade de combustíveis finos (Natural Resources Canada, 2023).

Índice de Húmus (DMC)

O Índice de Húmus (*Duff Moisture Code* - DMC) – corresponde ao teor de humidade dos combustíveis médios mortos (5kg/m2) abaixo da camada superficial entre 5 a 10 cm de profundidade (IPMA, 2023b). É um indicador do consumo de combustível em camadas moderadas de duff e material lenhoso de tamanho médio (Natural Resources Canada, 2023).

Índice de Seca (DC)

O Índice de Seca (*Drought Code* - DC) - corresponde ao teor de humidade dos combustíveis grossos (25kg/m2) mortos entre 10 e 20 cm de profundidade (IPMA, 2023b). É um indicador dos impactos sazonais da seca nos combustíveis florestais e da quantidade de combustão lenta em camadas mais profundas (Natural Resources Canada, 2023).

Índice de Combustível Disponível (BUI)

O Índice de Combustível Disponível (*Buildup Index* - BUI) - corresponde à quantidade de combustível disponível para a combustão (IPMA, 2023b). É baseado no DMC e no DC, e o seu valor é geralmente menor que o dobro do valor DMC (Natural Resources Canada, 2023).

Índice de Propagação Inicial (ISI)

O Índice de Propagação Inicial (*Initial Spread Index* - ISI) - corresponde à velocidade inicial de progressão do fogo (IPMA, 2023b). É baseado na velocidade do vento e FFMC (Natural Resources Canada, 2023).

Índice Meteorológico de Incêndio (FWI)

O Índice Meteorológico de Incêndio (*Fire Weather Index* - FWI) – corresponde à intensidade de fogo, estabelecida como a libertação de energia por unidade de comprimento da

frente de chamas (IPMA, 2023b). É utilizado como um índice geral de perigo de incêndio baseado no ISI e no BUI (Natural Resources Canada, 2023).

Índice de Severidade Meteorológica DSR (*Daily Severity Rating*)

O índice DSR interpreta, indiretamente, a severidade meteorológica diária local. Valores excessivos de DSR equivalem a níveis de severidade meteorológica excessivos (tendencialmente, temperaturas elevadas, vento forte, ausência de precipitação e humidade relativa baixa) (ICNF, 2022). O DSR é um índice aferido diretamente por meio do valor do FWI que representa a intensidade de um potencial incêndio e a sua dificuldade de contensão (de Meteorologia, 2008). A dificuldade de contensão de um incêndio pode ser estimada a partir do cálculo do valor médio diário do DSR que reflete a severidade média das condições climáticas diárias no continente (Castro Rego et al., 2020b). Portanto, de acordo com o (Natural Resources Canada, 2023), DSR é uma classificação numérica da dificuldade de controlar incêndios que utiliza como referência o índice de clima de incêndio reproduzindo com maior precisão o empenho esperado imprescindível para a eliminação do incêndio.

2.3 Ciência dos Dados e *Inteligência Artificial* (IA)

Os dados são considerados como um recurso valioso de informação e conhecimento. No entanto, para que sejam úteis, é necessário que sejam adequadamente interpretados. Apesar dos humanos terem habilidades para processar e descobrir conhecimento nos dados, essas habilidades não são suficientes atualmente, tendo em vista a enorme quantidade de dados gerados, necessitando de ferramentas mais apropriadas para exploração. Com o avanço da computação e dos sistemas de informação, tem sido possível aumentar a capacidade humana no armazenamento, gestão e tratamento desses dados (Carvalho, 2006). No estudo dos incêndios, a análise exploratória de incêndios florestais pode envolver a utilização de técnicas estatísticas e de mineração de dados. De acordo com Cortez e Morais, o uso de técnicas estatísticas e de mineração de dados pode ajudar na identificação de padrões e fatores de risco associados aos incêndios florestais, permitindo a elaboração de modelos preditivos e a tomada de decisão mais informada pelos órgãos responsáveis pela prevenção e combate aos incêndios (Cortez and Morais, 2007). A análise exploratória de dados pode ser utilizada para geração de relatórios de incêndios florestais como faz o ICNF, para contabilizar quantidade de ocorrências, dimensão dos incêndios, análises das causas, da severidade meteorológica, obtidos através de relatórios individuais anuais de incêndios que proporcionam uma visão e compreensão mais detalhada do comportamento do fenômeno por meio dos *insights* obtidos (ICNF, 2022). Os algoritmos de aprendizagem de máquina podem ser utilizados para a classificação de áreas afetadas por incêndios

como foi realizado no estudo proposto por Pacheco, no qual dados de imagens dos satélites Landsat-8, Sentinel-2 e Terra e as peculiaridades de cada uma dessas plataformas com o apoio das estatísticas de separabilidade Jeffries–Matusita foram utilizados para verificar o desempenho dos classificadores *k-Nearest Neighbor* (KNN) e *Random Forest* (RF) para a classificação de uma área afetada por incêndios no centro de Portugal (Pacheco et al., 2021).

Neste trabalho pretende-se utilizar *Data Mining* para análise e exploração de dados com intuito de obter *insights* úteis que possam ajudar na compreensão dos fenômenos relacionados e na identificação de padrões e elementos de perigo relacionados aos incêndios florestais, permitindo a construção de um modelo que classifique a área total afetada pelo incêndio auxiliando o processo de tomada de decisão na prevenção e contenção de incêndios.

2.4 *Inteligência Artificial* e *Aprendizagem de Máquina*

Muitas técnicas para solução de problemas e algoritmos computacionais que surgiram nas últimas décadas vem sendo utilizadas por vários segmentos - desde pesquisadores, grupos de pesquisa, empresas, consultores e até mesmo indivíduos comuns têm aproveitado esses recursos computacionais, teóricos, práticos e estatísticos, a fenômenos só observados na natureza para resolver os mais variados problemas. Essa variedade de envolvidos e técnicas fez com que a literatura apresentasse diferentes nomenclaturas. Dentre essas diversas nomenclaturas disponíveis, algumas muito utilizadas são: *Inteligência Artificial*, *Inteligência Computacional*, *Aprendizagem de Máquina* (Castro and Ferrari, 2016).

2.4.1 *Inteligência Artificial* Clássica

A *Inteligência Artificial* (IA) é uma área de conhecimento que busca compreender e desenvolver sistemas inteligentes, e uma das motivações para esse estudo é aprofundar o conhecimento sobre a natureza do ser humano (Russell and Norvig, 2022). Ela está associada ao processo de utilizar computadores para compreender a inteligência humana, mas se limitando essencialmente aos métodos influenciados pela biologia. Essas definições inspiravam-se na inteligência humana e a forma utilizada para construir o sistema inteligente baseava-se em uma visão procedural recomendando que sistemas inteligentes fossem planejados codificando-se conhecimentos especialistas em algoritmos específicos (Castro and Ferrari, 2016).

2.4.2 Inteligência Artificial Computacional

A IA clássica encontrou muitas dificuldades na proposta de projetar máquinas e organismos inteligentes que fossem capazes de realizar as mais diversas tarefas como robôs inteligentes, veículos auto-guiados etc. Isso fez com que surgissem diversas discordâncias entre ela e as abordagens que tinham essencialmente outras formas de atuar, como as *Redes Neurais Artificiais*, os *Sistemas Nebulosos* (*Fuzzy Systems*) e os *Algoritmos Evolutivos* principalmente por causa da disputa por financiamentos, Dessa forma, houve então a necessidade de separar essas áreas das técnicas que compunham a IA clássica e, para isso, desenvolveu-se uma nova linha de pesquisa intitulada *Inteligência Computacional* (Castro and Ferrari, 2016).

2.4.3 *Aprendizagem de Máquina*

A *Aprendizagem de Máquina* (AM) é a área de conhecimento que tem o intuito de construir programas computacionais que sejam capazes de melhorar automaticamente seu funcionamento através da experiência (Mitchell, 1997). Alpaydin define AM como a programação de computadores para aprimorar um critério de atuação a partir de experiências transcorridas, intituladas de exemplos ou simplesmente dados de entrada (Alpaydin, 2016). O intuito é que, de alguma maneira, essas técnicas possam ter a capacidade de aprender a resolver problemas. Sistemas com habilidade de aprendizagem são aptos a se adaptarem ou mudarem seu comportamento a partir de exemplos, de modo que manipule informações (Castro and Ferrari, 2016). O enfoque da *Aprendizagem de Máquina* é extrair informação a partir de dados de modo automático. Logo ela está essencialmente correlacionada à mineração de dados, à estatística, à *Inteligência Artificial* e à teoria da computação, além de outras áreas como computação natural, sistemas complexos adaptativos e computação flexível (Castro and Ferrari, 2016). Algumas das técnicas desenvolvidas em AM são: Indução de Regras e de Árvores de Decisão, Modelos Conexionistas e o Aprendizado Baseado em Instâncias (Goldschmidt et al., 2015). A mineração de dados emprega os métodos de *Aprendizagem de Máquina* para descobrir regularidades, padrões ou conceitos em bases de dados (Goldschmidt et al., 2015).

2.4.4 Paradigmas de Aprendizagem

Na área da mineração de dados, o processo de aprendizagem, ou treinamento, refere-se ao ajuste e/ou construção de um modelo a partir da apresentação dos objetos do conjunto de dados. Dessa forma, um algoritmo de aprendizagem ou treinamento define explicitamente como ensinar uma técnica de aprendizado de máquina. Por outro lado, o paradigma de aprendizagem é influenciado pelo ambiente em que a técnica é aplicada durante o processo de aprendizagem (Castro and Ferrari, 2016). Os dois paradigmas de aprendizagem

mais predominantes são:

- Aprendizado supervisionado: é fundamentado em um conjunto de objetos para os quais as saídas desejadas são conhecidas, ou outras informações que descrevam o comportamento esperado do sistema (Castro and Ferrari, 2016);
- Aprendizado não supervisionado: leva em consideração apenas os objetos da base, cujo rótulos não são conhecidos. O próprio algoritmo deve aprender a categorizar ou rotular os objetos (Castro and Ferrari, 2016).

2.5 Modelos Analíticos Descritivos e Preditivos

2.5.1 *Análise Descritiva de Dados* (ADD)

A *Análise Descritiva de Dados* (ADD) é uma técnica utilizada que proporciona uma visão geral das principais características de um conjunto de dados. Ela permite uma sumarização e compreensão dos objetos da base e seus atributos, sendo essencial para qualquer análise quantitativa de dados, pois descreve, simplifica ou sumariza as características mais importantes de um conjunto de dados. A diferença entre ADD e Mineração de dados está no fato de que a ADD objetiva descrever e encontrar o que está presente nos dados realizando investigação na distribuição das frequências, procurando medidas de centro e variação, medidas de posição relativa, associação dos dados aplicando técnicas elementares de visualização, enquanto que os algoritmos de mineração procuram conclusões que vão além dos dados mais que permitem inferir algo a partir deles (Castro and Ferrari, 2016).

2.5.2 Mineração de Dados

De acordo com Castro e Ferrari, a mineração de dados surgiu na década de 90 como uma nova área de pesquisa e aplicação independente (Castro and Ferrari, 2016). Entretanto, as suas origens vieram de períodos bem mais anteriores, a partir de disciplinas como a matemática, a estatística e a computação.

Segundo Goldschmidt, a mineração de dados é um dos passos ou etapas do processo de descoberta de conhecimento KDD, e emprega os métodos de AM para obter regularidades, padrões ou conceitos em bases de dados (Goldschmidt et al., 2015). Embora o termo seja utilizado por muitos como sinônimo de KDD, ficou definido na primeira conferência Internacional de KDD em 1995 no Canadá que a terminologia *Knowledge Discovery in Databases* é referente a todo o processo de extração de conhecimento a partir dos dados; e que mineração de dados fosse exclusivamente utilizada para etapa de descoberta do processo de KDD (Adriaans and Zantinge, 1996).

De acordo com Fayyad, a expressão KDD é definida como um processo não trivial, interativo e iterativo, para encontrar padrões inteligíveis, válidos, novos e úteis obtidos de grandes bases de dados (Fayyad et al., 1996).

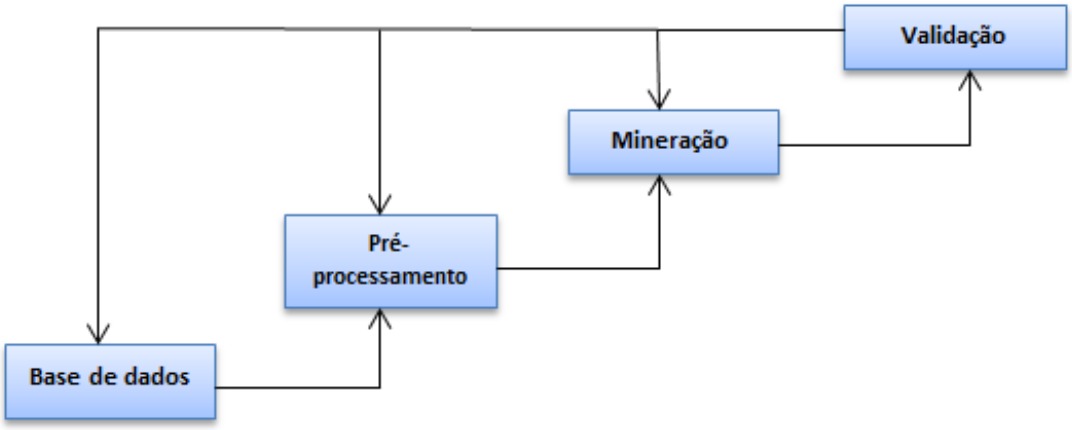


Figura 2.2: Processo de *Knowledge Discovery in Databases*. Adaptado de (Castro and Ferrari, 2016)

Resumo do Processo de KDD em Quatro Etapas:

- Base de dados: conjunto organizado de dados que possibilita uma recuperação satisfatória dos dados que podem ser valores quantitativos ou qualitativos pertencentes ao acervo (Castro and Ferrari, 2016). O processo tem início com a organização da base de dados composta pelos dados de interesse. Uma série de procedimentos de pré-processamento são aplicados aos dados que incluem, mas não se restringem a, organizar os dados em um único repositório, como um *Data Warehouse* (DW) (Silva et al., 2016).

- Pré-processamento: de acordo com Goldschmidt, esta etapa engloba todas as funções relacionadas à captação, à organização e ao tratamento dos dados objetivando preparar os dados para etapa de mineração de dados (Goldschmidt et al., 2015). O tratamento abrange a eliminação de instâncias repetidas e valores discrepantes; a seleção dos dados que realmente tenham relevância para a mineração de dados; a normalização para transformar os dados para a mesma escala a fim de evitar que os algoritmos considerem alguns dados mais relevantes que outros por estarem em escalas diferentes (Silva et al., 2016); transformação para converter os dados para os formatos adequados para mineração; e a integração a partir da combinação de dados de várias fontes (Castro and Ferrari, 2016).

- Mineração: a etapa correspondente à busca efetiva por conhecimento pelos algoritmos de mineração a partir dos dados pré-processados (Castro and Ferrari, 2016).

De acordo com Goldschmidt, é a etapa mais importante do processo de KDD, tanto que alguns autores utilizam os termos mineração de dados e descoberta de conhecimento em base de dados como sinônimos (Goldschmidt et al., 2015).

- Validação: avaliar o quão úteis são os conhecimentos obtidos pelo resultado da mineração de dados (Castro and Ferrari, 2016).

  O processo de KDD como um todo pode ser iterativo e interativo, no qual cada fase pode ser executada mais de uma vez, na sequência normal ou fora dela, a depender do conjunto de dados originais e/ou de decisões tomadas pelo analista (especialista no domínio dos dados) (Silva et al., 2016).

2.6 Trabalhos Relacionados e Discussão

2.6.1 Trabalhos Relacionados

Conforme Oliveira afirma, a existência de uma fonte de combustão e das condições de espalhamento do fogo são o suficiente para ativar a ocorrência de um incêndio. Esses componentes são de grande relevância para a avaliação do risco de incêndios (Oliveira et al., 2012).

Oliveira, propôs modelar os padrões espaciais de ocorrência de incêndios na Europa Mediterrânea, fazendo uso de duas técnicas: Regressão Linear Múltipla tradicional e Floresta Aleatória (Oliveira et al., 2012). Através dessas técnicas, os autores procuraram identificar os principais fatores que influenciam a ocorrência de incêndios na região. Ambos os modelos mostram que a distribuição espacial da probabilidade de ocorrência de incêndios é altamente variável naquela região: maior probabilidade de incêndio na região noroeste da Península Ibérica e sul da Itália, enquanto é baixa no norte da França, nordeste da Itália e norte da Grécia.

No estudo proposto por Bem foram utilizados modelos que relacionem a ocorrência do fogo às variáveis que o influenciam (Bem, 2017). Mais precisamente, dois modelos distintos de previsão: *Regressão Logística* (RL) e uma *Redes Neurais Artificiais* (RNA) foram aplicados à região do Distrito Federal brasileiro, que se encontra inserida dentro do bioma Cerrado. Os modelos tiveram performances similares e foram otimizados em função da melhor medida de *Area Under Receiver Operating Characteristic* (AUROC, ou simplesmente AUC) a partir da seleção de atributos, e posteriormente validados utilizando dados reais de áreas queimadas. Concluiu-se que os métodos foram adequados para a elaboração de mapas de risco de incêndio no Cerrado. Na área onde o estudo foi realizado é apenas uma pequena porção do bioma, e, portanto, os modelos elaborados podem ter dificuldades ao serem aplicados à totalidade da região, não só devido à baixa representatividade dos dados em relação à escala do Cerrado, mas também devido às maiores

amplitudes das variáveis em maiores escalas.

Safi e Bouroumi apresentaram e discutiram uma abordagem baseada em *Redes Neurais* para o problema de prever incêndios florestais (Safi and Bouroumi, 2013). Os dados utilizados para previsão foram dados referentes ao parque natural de Montesinho em Portugal em que o algoritmo de aprendizado *backpropagation* foi utilizado para treinar a rede neural. Os resultados obtidos foram satisfatórios e a área de aplicação foi o parque natural de Montesinho. Como continuação dos resultados, o estudo incentiva reduzir a sensibilidade do método a parâmetros arquitetônicos e algorítmicos.

Cortez e Morais propuseram em seu estudo, uma abordagem de *Data Mining* (DM) para prever a área queimada de incêndios florestais empregando cinco técnicas diferentes de DM (Cortez and Morais, 2007). *Support Vector Machines* (SVM) e *Random Forest*, e quatro configurações distintas de seleção de recursos (usando componentes espaciais, temporais, FWI e atributos climáticos), foram testados em dados recolhidos referentes à região nordeste de Portugal. A melhor configuração utilizou um SVM e quatro entradas meteorológicas (ou seja, temperatura, humidade relativa, chuva e vento) e é capaz de prever a área queimada de pequenos incêndios, que são mais frequentes. A abordagem é particularmente útil para melhorar o gerenciamento de recursos de combate a incêndios.

O ICNF tem gerado diversos relatórios provisórios de incêndios rurais que variam de 1º de janeiro a quinze de outubro em Portugal (da Conservação da Natureza e das Florestas , ICNF). Esses relatórios são emitidos anualmente e incluem estatísticas e dados relacionados aos incêndios florestais, como o número total de ocorrências, a área ardida, as causas dos incêndios e a distribuição geográfica dos mesmos. Esse é um trabalho que ajuda a compreender a evolução dos incêndios ao longo do tempo possibilitando identificar tendências e padrões a partir dos dados.

2.6.2 Discussão

Em relação aos estudos e trabalhos relacionados, este trabalho apresenta um diferencial nos seguintes aspectos: 1) quanto aos dados utilizados - o conjunto de dados é mais abrangente, considerando registros ocorridos em todo continente Português e não apenas em localidades específicas como observou-se em alguns trabalhos; 2) quanto ao intervalo das ocorrências - os registros compreendem a um período superior a uma década, mais precisamente a onze anos de ocorrências registradas pelo ICNF, abrangendo o período compreendido entre janeiro de 2011 e dezembro de 2022, com os dados mais atuais disponíveis até o momento do estudo; 3) quanto à quantidade de registros do conjunto - como os dados são referentes a Portugal Continental e não a uma região específica do País, os registros utilizados nesse estudo formam uma extensa base de dados com 195.705 registros de ocorrências registradas pelo ICNF contendo diversos atributos. Apesar de não possuir variáveis meteorológicas diretas como temperatura, humidade relativa do ar, ve-

locidade do vento e precipitação, a base possui para cada registro de ocorrência um valor de DSR que traduz, ainda que de forma indireta, a severidade meteorológica diária local (ICNF, 2022) e de FWI índice de perigo de incêndio rural integra seis índices: (DC, DMC, FFMC, ISI, BUI e FWI) que quantificam os efeitos da humidade do combustível e do vento no comportamento do fogo (IPMA, 2023a).

O enfoque da abordagem foi pautado em duas estratégias. No primeiro momento foi realizada uma extensa análise exploratória da massa de dados, não se limitando apenas à produção de relatório para geração de estatísticas unicamente. Nesse trabalho optou-se por uma estratégia mais *Data Mining* do que simplesmente estatística descritiva convencional. Foi utilizada a linguagem de programação Python para enriquecimento do processo e permitir uma maior compressão e entendimento do fenômeno a partir da geração de diversos tipos de gráficos como barras, boxplot mapas de calor, geração de tabelas para visualização e compreensão dos dados, realização de análises para cada resultado, aprofundando e direcionando a análise para Mineração de dados para obtenção de *insights* úteis presentes nos dados. Vale a pena ressaltar que, segundo (Goldschmidt et al., 2015), uma maneira de tornar inteligível o conhecimento ou os padrões é através da representação gráfica. Sendo assim, há alguns pontos importantes a destacar sobre a análise exploratória realizada:

- Quanto a dimensão dos incêndios: este trabalho não se limitou às quantidades de incêndios por ano apenas de forma geral. A análise foi mais profunda, detalhou a quantidade por tipo de área afetada: Povoamento, Matos, Agrícolas e Total, informações valiosas para o estudo e análises. Priorizando melhorar a compreensão e enriquecer as análises, foram gerados gráficos, relacionando à quantidade de incêndios por tipo de área afetada durante os anos de 2011 a 2022, gráfico para apresentar a quantidade das ocorrências por ano em ordem decrescente e uma plotagem para que se pudesse perceber a evolução da quantidade de incêndios no decorrer dos anos. Em relação à área total afetada, os gráficos foram utilizados para representar a dimensão dos tipos de áreas atingidas durante os anos, a plotagem auxiliou a perceber o comportamento dos incêndios ao apresentar sua evolução quanto à área total atingida no período de 2011 a 2022. Em relação à dimensão dos incêndios por classes de áreas, utilizou-se um gráfico para representar essas dimensões durante os anos 2011 a 2022.

- Para análises das causas: este trabalho não se restringiu a apresentar somente a distribuição percentual dos incêndios rurais por tipos de causas mais frequentes, mas considerou analisar o número de incêndios rurais por Tipo de Causa e Grupo de Causa em cada ano por meio de gráficos representativos.

- Quanto às análises regionais: este trabalho não se contentou em apresentar o número de incêndios rurais e a extensão de área ardida por distrito e concelhos apenas

em uma tabela. O estudo considerou a geração de gráficos mais um recurso válido para visualizar o número geral e de tipos de incêndios por distritos e por concelhos, para apresentar a extensão dos tipos de área ardida nos distritos e concelhos, assim como para que se pudesse verificar e perceber o comportamento e evolução da quantidade de ocorrências e da evolução das áreas atingidas nos distritos a partir de uma plotagem dessa evolução no decorrer dos anos.

- Análises mensais, aqui se enriqueceu o trabalho com a apresentação de gráficos como boxplot, permitindo visualizar os dados de uma outra perspetiva e plotagens com gráficos de barras para perceber a evolução entre os meses de janeiro a dezembro do período de 2011 a 2022 . Quanto à quantidade de incêndios, preocupou-se também em apresentar a quantidade por tipos de área, objetivando enriquecer ainda mais as análises. A quantidade de incêndios mensais e as áreas atingidas por tipos de área também foram apresentados em gráficos sem detrimento ao uso de tabelas.

- Quanto à severidade meteorológica DSR, este trabalho acrescentou a quantidade de ocorrências das classes DSR de 2011 – 2022 por meses de janeiro a dezembro, uma estatística importante para perceber a influência das condições climáticas mais severas no que diz respeito à quantidade e à dimensão de áreas afetadas pelos incêndios. Gráficos e tabelas foram gerados para apresentar a quantidade de ocorrências anuais e mensais.

- FWI, tendo em vista que o aumento de cada um destes componentes corresponde a um aumento de perigo de incêndio (IPMA, 2023a). Este trabalho realizou análises para obter o número de ocorrências por classe FWI durante os anos e também durante meses. A quantidade de ocorrências foi representada por meio de tabelas e gráficos.

- Por fim, foi acrescentada uma plotagem representando um mapa de calor geográfico com o objetivo de se perceber a concentração de incêndios nos Distritos de Portugal por meio de uma visão diferenciada, panorâmica, a qual pode contribuir como fonte de alerta para os combatentes dos incêndios e para a população desses distritos.

Por fim, vale ressaltar que essas análises não se restringiram a um ano específico ou apenas sobre registros compreendidos entre 1º de janeiro e 15 de outubro de determinado ano, as análises deste estudo levaram em consideração toda uma base de dados composta por incidentes de registros históricos reais compreendidos no período de 2011 a 2022.

Já no segundo momento, a abordagem foi um pouco mais ousada, propôs prever incêndios florestais em Portugal Continental, mais precisamente qual é a classe de área total ardida por um incêndio a partir de um registro submetido ao classificador. Ao invés de se empregar duas técnicas como alguns dos trabalhos anteriores, para a modelagem nesse

estudo foram utilizadas cinco técnicas de *Machine Learning*: *Regressão Logística*, *Aprendizagem Bayesiana*, *Decision Trees*, *Random Forest* e *Redes Neurais*, utilizando diversos atributos importantes para predição pertencentes ao conjunto. Apesar da base não possuir componentes espaciais e climáticos diretos, de forma indireta, a severidade meteorológica diária local foi representada pelo DSR (ICNF, 2022) e o perigo de incêndio rural pelo FWI e seus componentes (ISI, DC, DMC, FFMC e BUI) (IPMA, 2023b). Atributos como: Mês da ocorrência, Tipo de área afetada, área afetada em ha, local da ocorrência, além dos tipos de causas foram empregados para geração do classificador. O estudo comparou a acurácia de cada modelo e mostrou aquele que melhor se adequou ao conjunto de dados reais de Portugal Continental. No geral tanto análise exploratória quanto o modelo de previsão gerado no estudo visou auxiliar na tomada de decisões relacionadas à prevenção e ao combate a incêndios florestais, permitindo que as autoridades competentes ajustem as estratégias e aloquem recursos de forma mais eficiente e, consequentemente, minimização dos impactos ocasionados pelos incêndios.

Capítulo 3

Metodologia

3.1 Introdução

Neste trabalho utilizou-se a metodologia CRISP-DM (*Cross-Industry Standard Process for Data Mining*) para guiar o processo de mineração de dados, reconhecido como um método comprovado para orientar trabalhos de mineração de dados (IBM, 2021). A metodologia CRISP-DM é composta por seis fases principais (Pellegrino, 2020):

- Entendimento do Negócio: definiu-se claramente os objetivos do projeto e os requisitos específicos do gerenciamento de incêndios florestais. O objetivo principal era desenvolver um modelo preditivo para prever a faixa de área total ardida por um incêndio florestal;

- Entendimento dos Dados: foram coletados dados relevantes de fontes específicas, como o ICNF. Realizou-se uma análise exploratória detalhada e abrangente com intuito de compreender e identificar características importantes dos incêndios florestais em Portugal.

- Preparação dos Dados: foram utilizados métodos para seleção de dados, limpeza dos dados e codificação dos dados para assegurar a qualidade dos fatos representados.

- Modelagem: utilizou-se técnicas de *Aprendizagem de Máquina*, como *Regressão Logística*, *Aprendizagem Bayesiana*, *Árvore de Decisão*, *Random Forest* e *Redes Neurais*. Os modelos foram treinados utilizando a função *train_test_split* do *scikit-learn* para garantir sua robustez e precisão.

- Avaliação: os modelos foram avaliados com base em métricas de desempenho, accuracy, precision, recall e f1-score, para garantir que atendiam aos objetivos do estudo e eram capazes de fornecer previsões úteis e precisas.

- Implantação: embora a fase de implantação tipicamente envolva a implementação dos modelos no ambiente de produção, neste estudo, não realizou-se a implantação. Em vez disso, focou-se na geração e teste dos modelos para verificar suas métricas de desempenho. A decisão de não prosseguir com a implantação foi tomada devido à natureza exploratória deste trabalho, cujo objetivo principal era avaliar a viabilidade e a precisão dos modelos preditivos para suporte à tomada de decisões no gerenciamento de incêndios florestais.

A metodologia CRISP-DM proporcionou uma estrutura robusta e adaptável para o nosso projeto, garantindo que cada etapa do processo fosse conduzida de maneira organizada e eficiente, resultando em modelos preditivos eficazes para a gestão de incêndios florestais.

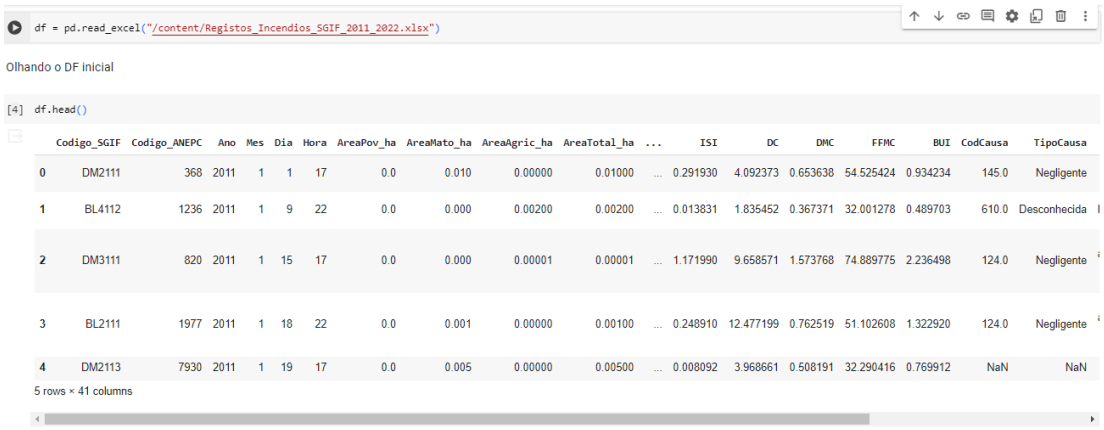
3.2 Fontes dos Dados Utilizadas

O *Sistema de Gestão de Informação de Incêndios Florestais* (SGIF) é uma ferramenta importante para coleta, organização, análise e apresentação de dados estatísticos sobre diferentes aspectos florestais, como área florestal, espécies florestais, incêndios florestais, entre outros. Através dele, o ICNF pode acompanhar e avaliar o estado e a evolução das florestas em Portugal, dando suporte à tomada de decisões relacionadas à gestão e conservação desses recursos naturais (da Conservação da Natureza e das Florestas , ICNF). A partir do SGIF, é possível acessar dados sobre incêndios florestais, relatórios, mapas, gráficos e outras informações estatísticas sobre as florestas portuguesas, contribuindo para uma melhor compreensão do seu estado atual e auxiliando na implementação de políticas de proteção e gestão sustentável (Instituto da Conservação da Natureza e das Florestas, 2010). Com intuito de obter e explorar dados reais referentes aos incêndios florestais em Portugal, dados históricos de incêndios foram acessados e colhidos para este estudo a partir dos registos individuais de incêndios 2011 a 2022, informações oficiais do SGIF disponíveis no ICNF, uma fonte de dados contendo registros organizados em planilhas por intervalos de anos, atualmente contendo registros até o ano de 2022.

3.3 Descrição Sumária dos Dados

Os dados do ICNF estão estruturados em planilhas no formato xls em seu site de Informação oficial do SGIF. Dessa forma, durante o processo de coleta dos dados desse trabalho, foi realizada a integração dos conjuntos 2011 a 2020 com os registros de 2021 a 2022, para formar uma planilha única, uma base de dados composta por registros de 2011 a 2022. Esses dados foram utilizados para o processamento no Pandas a partir da geração

de um *dataframe* para inicializar o processo de análise e exploração dos dados utilizando os recursos e bibliotecas disponíveis pelo *Phyton* Figura 3.1.



```
df = pd.read_excel("/content/Registos_Incendios_SGIF_2011_2022.xlsx")
```

Olhando o DF inicial

```
[4] df.head()
```

	Codigo_SGIF	Codigo_ANEPC	Ano	Mes	Dia	Hora	AreaPov_ha	AreaMato_ha	AreaAgric_ha	AreaTotal_ha	...	ISI	DC	DMC	FFMC	BUI	CodCausa	TipoCausa
0	DM2111	368	2011	1	1	17	0.0	0.010	0.00000	0.01000	...	0.291930	4.092373	0.653638	54.525424	0.934234	145.0	Negligente
1	BL4112	1236	2011	1	9	22	0.0	0.000	0.00200	0.00200	...	0.013831	1.835452	0.367371	32.001278	0.489703	610.0	Desconhecida
2	DM3111	820	2011	1	15	17	0.0	0.000	0.00001	0.00001	...	1.171990	9.658571	1.573768	74.889775	2.236498	124.0	Negligente
3	BL2111	1977	2011	1	18	22	0.0	0.001	0.00000	0.00100	...	0.248910	12.477199	0.762519	51.102608	1.322920	124.0	Negligente
4	DM2113	7930	2011	1	19	17	0.0	0.005	0.00000	0.00500	...	0.008092	3.968661	0.508191	32.290416	0.769912	NaN	NaN

5 rows × 41 columns

Figura 3.1: Geração do *dataframe* a partir da biblioteca pd do pandas

A base de dados é composta por milhares de registros históricos de incêndios reais que ocorreram em Portugal, com diversos atributos relevantes e importantes para análise e compreensão do fenômeno resumidos a seguir:

- Data de ocorrência dos incêndios: essas informações estão expressas também nas colunas hora, dia, mês e ano, nas quais há informações do momento de ocorrência de um incêndio;
- Data de Alerta: informam a data e hora de alerta do incêndio;
- Data de intervenção: informam a data e hora de uma intervenção;
- Data de extinção: informam a data e hora de uma extinção;
- Duração do incêndio: informam a duração em horas e intervalos superior ou não a 24 horas de um incêndio;
- Áreas ardidas: também são apresentadas por tipo área ardida, se a área foi uma área de povoamento, área de mato, área agrícola e área total;
- Classe de área ardida por um incêndio: informa em que classe de área ardida se enquadra um registro, por exemplo: 0 e 1 ha, 1 a 10 ha etc;
- Localidades afetadas: informa em que Distritos, Concelhos, Freguesia e locais os incêndios ocorreram;
- Coordenadas geográficas: dispõe das coordenadas de um registro como: X_Militar, Y_Militar, Latitude, longitude, X_ETRS89 e Y_ETRS89.

- Índice de severidade meteorológica: representado pelo DSR que, conforme ICNF, traduz, ainda que de forma indireta, a severidade meteorológica diária local (ICNF, 2022);
- O *Índice Meteorológico de Incêndio do Sistema Canadiano* CFFWIS ou genericamente designado FWI composto pelos índices: FFMC, DMC, DC, ISI, BUI e FWI (IPMA, 2023b);
- Causas dos incêndios: estas informações estão presentes nas colunas Tipo de Causa, Grupo da Causa e Descrição da Causa.

É importante salientar que muitos conjuntos de dados do mundo real não oferecem entrada de treinamento suficiente para classificadores regulares fazendo com que algumas classes fiquem mais representadas do que outras. Dados desbalanceados levantam problemas na classificação do *Machine Learning* e prever um resultado torna-se difícil quando não há dados suficientes para aprender (Padurariu and Breaban, 2019). Não obstante do cenário real, esta base também apresenta-se desbalancaeada, na qual o desequilíbrio presente nos dados prejudica muito a classificação dos registros. Como alternativa, a estratégia foi recorrer aos métodos de reamostragem que visam modificar o conjunto de dados para reduzir a discrepância entre os tamanhos das classes. Um dos métodos empregados que não gerou resultados satisfatórios foi a subamostragem que elimina instâncias da classe majoritária (Padurariu and Breaban, 2019). Já com a sobreamostragem baseada na criação de instâncias sintéticas para as classes minoritárias, uma técnica na qual o algoritmo utiliza cada amostra da classe minoritária e introduz amostras sintéticas ao longo da linha que conecta a instância atual a alguns de seus k vizinhos mais próximos da mesma classe (Padurariu and Breaban, 2019), essa técnica mostrou-se eficiente para obter bons resultados, melhorando assim a classificação.

3.4 Pré-processamento dos Dados

Neste tópico será apresentada a metodologia de pré-processamento dos dados com a descrição dos passos e técnicas específicas que foram utilizadas na pesquisa.

Pyle, estima que a preparação de dados sozinha representa 60% de todo o tempo e esforço despendido no processo de mineração de dados (Pyle, 1999); para Mckinney, essas tarefas gastam de 80% a mais do tempo de um analista (McKinney, 2018); já o Projectpro concluiu que a maior parte do tempo de um cientista de dados é gasta na preparação de dados (coleta, limpeza e organização), antes que eles possam começar a fazer a análise de dados (ProjectPro, 2023). Neste capítulo, será discutido como os dados foram coletados, preparados para análise, o que inclui a limpeza dos dados, tratamento de valores ausentes, seleção de atributos relevantes, codificação, entre outros.

3.4.1 Seleção dos Dados

Para a mineração de dados, o subconjunto dos dados que foi considerado para a realização do processo teve enfoque na seleção de atributos e seleção de registros. Conforme Goldschmidt, a seleção de atributos despreza aqueles atributos que são totalmente irrelevantes para a classificação, por exemplo, o nome de um cliente é um atributo sem qualquer relevância em uma aplicação com objetivo de prever o comportamento de um cliente quanto ao pagamento de futuros créditos a eles concedidos (Goldschmidt et al., 2015). Desta maneira, Codigo_SGIF (Identificador único SGIF do incêndio rural) e Codigo_ANEPC (Identificador da ocorrência na base-de-dados SADO da ANEPC igualmente designado por código NCCO), foram desprezados tendo em vista que não agregam relevância alguma para a classificação pois são apenas identificadores únicos dos registros da base de dados.

Já na seleção de registros, pode haver situações nas quais não é possível utilizar todo o conjunto de dados por algum motivo, como ausência de valores de atributos em alguns registros, ou o conjunto de dados completo pode ser muito grande para ser manipulado etc. (Goldschmidt et al., 2015). Conforme Aggarwal, quando os dados têm um número muito alto de dimensões, muitos algoritmos de mineração de dados não funcionam de forma eficaz (Aggarwal, 2015). Sendo assim, um subconjunto dos registros deve ser selecionado para compor o conjunto de dados a ser selecionado. Na base de dados foi observado que já existia uma coluna com ano, uma com o mês, uma com o dia e uma informando duração do incêndio. Foi possível, então, reduzir ainda mais a base de dados descartando as colunas 'DataHora_PrimeiraIntervencao', 'DataHora_Extincao' sem trazer prejuízos para o processo.

Por fim, com intuito de obter uma base de dados com os registros mais interessantes para a classificação foram utilizadas técnicas estatísticas de correlação de dados como correlação de Pearson e Spearman para averiguar os atributos mais promissores. O coeficiente de correlação de Pearson (nomeado em homenagem a Karl Pearson) é utilizado para resumir a força da relação linear entre duas amostras de dados (Brownlee, 2019), ele é usado para mensurar a dependência linear entre as variáveis, ou seja, determina se existe uma relação linear entre as variáveis (Castro and Ferrari, 2016). Já a correlação de Spearman é um procedimento estatístico projetado para medir a relação entre duas variáveis em uma escala ordinal de medição. A intuição por trás da correlação de Spearman é que ela calcula uma correlação de Pearson usando os valores de postos em vez dos valores reais (Brownlee, 2019). Ambas ajudam a quantificar a relação entre as variáveis, variam de -1 a 1, onde -1 indica uma correlação negativa perfeita, 1 indica uma correlação positiva perfeita e 0 indica ausência de correlação. Neste trabalho a correlação de Pearson foi utilizada para selecionar os atributos mais relevantes para a fase de seleção dos atributos.

3.4.2 Limpeza dos Dados

De acordo com Aggarwal os dados podem possuir entradas errôneas ou ausentes (Aggarwal, 2015). Assim, alguns registros poderão vir a ser serem descartados, as entradas ausentes serem estimadas, ou as inconsistências poderão ser removidas.

Como forma de assegurar a qualidade dos fatos representados, considera-se a limpeza dos dados qualquer forma de tratamento realizado sobre os dados selecionados a fim de garantir essa qualidade (Goldschmidt et al., 2015).

Sendo assim, foram realizadas investigações na base de dados para verificar a existência de registros inconsistentes. Por exemplo, na duração de incêndios e áreas ardidas foi verificado se havia a existência de registros com valores negativos. Uma vez que essas inconsistências não foram encontradas, os dados foram preservados.

Uma investigação foi feita na procura de ocorrências de registros ausentes, e outra na busca por registros com valores nulos ou faltantes. Verificou-se que colunas como *Rede Nacional de Matas Nacionais e Perímetros Florestais* (RNMNPF) e *Rede Nacional de Áreas Protegidas* (RNAP) possuíam dados faltando em mais de 70% dos casos, conforme a Figura 3.2.

```
[ ]  pNaN = df.isnull().sum()/df.shape[0]*100
     pNaN[pNaN > 20].sort_values(ascending=False)

     RNMPF    96.824302
     RNAP     96.422677
     dtype: float64
```

Figura 3.2: Código para Inspecionar Dados Faltantes.

Com o propósito de assegurar a qualidade dos fatos representados e a irrelevância desses dados para a classificação, essas colunas foram removidas da base de dados, conforme a Figura 3.3.

```
[ ]  dfEnxuto = df.drop(columns=['RNMPF', 'RNAP'])
```

Figura 3.3: Pré processamento para tratar dados faltantes

Já as linhas que apresentavam valores nulos também foram removidas para garantir o funcionamento dos algoritmos de aprendizagem. Dessa forma, a fim de realizar o tratamento na base, passou-se a considerar apenas colunas e linhas sempre com valores. A fim de evitar que os dados das colunas Distrito, Concelho, Freguesia e Local fossem diferenciados por letras maiúscula e minúscula como observou-se durante a construção de gráficos para análises, o tratamento foi realizado sobre os mesmos com o objetivo de deixá-los padronizados e assegurar a qualidade dos fatos representados.

3.4.3 Descritização e Normalização dos Dados

Goldschmidt advoga que os dados devem ser codificados para que possam ser usados como entrada dos algoritmos de mineração de dados (Goldschmidt et al., 2015). Desta forma, variáveis categóricas, como ClasseArea, Distritos, Concelho, Local, TipoCausa, precisaram ser codificadas em valores numéricos para que pudessem servir de entrada para o processamento pelos algoritmos de *Machine Learning*.

3.5 Exploração dos Dados

Nesta etapa muitos gráficos e tabelas foram gerados para auxiliar no processo de análise e exploração dos dados, fazendo uso de importantes ferramentas que apoiaram e permitiram realização do processo.

De acordo com Morgan, *Python* é uma linguagem de programação simples, clara e intuitiva, sendo escolhida por muitos engenheiros e cientistas para diversas aplicações científicas e numéricas (Morgan, 2016). Talvez essa escolha se deva ao fato de que muitos prefiram entrar rapidamente na tarefa principal, como descobrir o efeito ou a correlação de uma variável com uma saída, em vez de passar inúmeras horas aprendendo os detalhes de uma linguagem de programação "complexa". Além do mais, a existência de muitos pacotes e ferramentas que tornam o uso do *Python* na análise de dados e aprendizado de máquina muito mais fácil. Levando-se em consideração todas essas vantagens, somadas à intenção de iniciar o projeto mais rapidamente, obtendo *insights* valiosos em menos tempo e recursos, optou-se pela linguagem *Python* para a exploração dos dados desse trabalho. Com o intuito de facilitar a construção de gráficos para visualização dos dados, e permitir a identificação de padrões, análises descritivas foram realizadas com o uso de recursos como as bibliotecas Pandas, uma ferramenta de análise e manipulação de dados de código aberto rápida, poderosa, flexível e fácil de usar, construída sobre a linguagem de programação *Python* (Team, 2023); *Matplotlib*, uma biblioteca abrangente para criar visualizações estáticas, animadas e interativas em *Python* (Team, 2012 – 2023) e *Seaborn* (Team, 2012-2023) uma biblioteca de visualização de dados *Python* baseada em *matplotlib* que fornece uma interface de alto nível para desenhar gráficos estatísticos atraentes e informativos (Team, 2012-2023).

Capítulo 4

Análise dos Dados

Introdução

Incêndios florestais são uma grave adversidade ambiental, gerando danos econômicos e ecológicos, além de colocar vidas humanas em risco (Cortez and Morais, 2007). Objetivando melhor compreender os fatores que desencadeiam este fenômeno, neste trabalho exploramos uma abordagem em *Data Mining* rica em gráficos e tabelas para visualização dos dados, obtenção de estatísticas e *insights* que são de suma importância para compreensão do fenômeno estudado.

4.1 Descrição dos Dados

A base de dados utilizada neste trabalho é resultado da junção de duas planilhas referentes aos registros de incêndios reais ocorridos em Portugal oriundos da base nacional de dados de incêndios rurais do ICNF. Uma das planilhas refere-se aos registros de 2011 a 2020, e a outra aos registros de 2021 a 2022. Com as planilhas integradas formando uma única planilha, a base de dados ficou composta por registros que ocorreram em Portugal desde 2011 até os mais atuais de 2022.

Após o tratamento dos dados realizado na etapa de pré-processamento de dados, o *dataset* resultante que foi submetido aos algoritmos *Machine Learning* na etapa de modelagem dos dados ficou representado da seguinte maneira, Tabela 4.1. Nela são apresentados os campos para cada registro seguidos de sua descrição. Os últimos campos da tabela seguidos _cod são os campos que foram codificados de valores categóricos para valores numéricos para que pudessem ser submetidos aos algoritmos de *Machine Learning*.

Tabela 4.1: Dataset Processado

Campo	Descrição
Mes	Mês da data de alerta do incêndio
Hora	Hora do alerta do incêndio
Duracao_Horas	Duração do incêndio (período entre a data/hora de alerta e a data/hora de extinção)
IncSup24horas	Incêndio rural com duração superior a 24 horas (período entre a data/hora de alerta e a data/hora de extinção)
X_Militar	Coordenada X do ponto de início do incêndio em Datum Lisboa Hayford Gauss (EPSG: 20790)
Y_Militar	Coordenada Y do ponto de início do incêndio em Datum Lisboa Hayford Gauss (EPSG: 20790)
Latitude	Latitude do ponto de início do incêndio (unidade: graus decimal)
Longitude	Longitude do ponto de início do incêndio (unidade: graus decimal)
X_ETRS89	Coordenada X do ponto de início do incêndio no sistema de coordenadas PT-TM06/ETRS89 (EPSG: 3763)
Y_ETRS89	Coordenada Y do ponto de início do incêndio no sistema de coordenadas PT-TM06/ETRS89 (EPSG: 3763)
DSR	Índice de perigo meteorológico de incêndio rural (calculado com base no FWI)
FWI	Índice meteorológico de perigo de incêndio rural
ISI	Índice meteorológico de propagação inicial do fogo
DC	Índice meteorológico de seca
DMC	Índice meteorológico de humidade da manta-morta
FFMC	Índice meteorológico de humidade do combustível fino
BUI	Índice meteorológico de combustível disponível
Distrito_cod	Distrito do ponto de início do incêndio após codificação de categórico para numérico
Concelho_cod	Concelho do ponto de início do incêndio após codificação de categórico para numérico
CodCausa	Código da causa do incêndio | ver correspondência na folha "TipoCausa"
TipoCausa_cod	Tipificação da causa após codificação de categórico para numérico
GrupoCausa_cod	Grupo da causa após codificação de categórico para numérico
DescricaoCausa_cod	Descrição da causa após codificação de categórico para numérico
ClasseArea_cod	Classe de área ardida total (ha) após codificação de categórico para numérico

4.2 Estatísticas Anuais

Neste tópico serão apresentadas as estatísticas anuais referentes aos registros de incêndios que ocorreram em Portugal durante o período de 2011 a 2022.

4.2.1 Quantidade de Incêndios por Tipo de Área em Portugal

O número de ocorrências de incêndios anuais por tipo de área ardida está apresentada na Tabela 4.2 com as informações dispostas nas colunas: Povoamento, Mato, Agrícola e Povoamento + Mato + Agrícola, referentes ao intervalo de 2011 a 2022.

Para realização da contagem das ocorrências para cada um dos tipos de área, considerou-se na base de dados, aqueles registros cuja a área ardida apresentada fosse superior zero *hectares*. Desta forma, observou-se que a maior quantidade de ocorrências de incêndios registrada aconteceu em 2011 com o tipo de área Mato chegando a registrar 18619 ocorrências. Por outro lado, o tipo Povoamento foi o que registrou o menor valor, obtendo 1691 registros no ano de 2021.

Tabela 4.2: Tabela: Quantidade de incêndios por tipo de área

Ano	Povoamento	Mato	Agrícola	Pov + Mato + Agr
2011	6901	18619	4756	30276
2012	7026	15704	4453	27183
2013	6413	15088	4083	25584
2014	2107	5665	2434	10206
2015	4979	12853	3992	21824
2016	3715	10986	3043	17744
2017	5518	14076	3902	23496
2018	2976	8213	2167	13356
2019	2222	6795	2830	11847
2020	1954	6433	2194	10581
2021	1691	5223	2065	8979
2022	2677	6983	2397	12057
Total	48179	126638	38316	213133

De acordo com a Figura 4.1, é possível perceber que, durante todos os anos, o tipo de área que registrou a maior quantidade de incêndios foi a área de Mato.

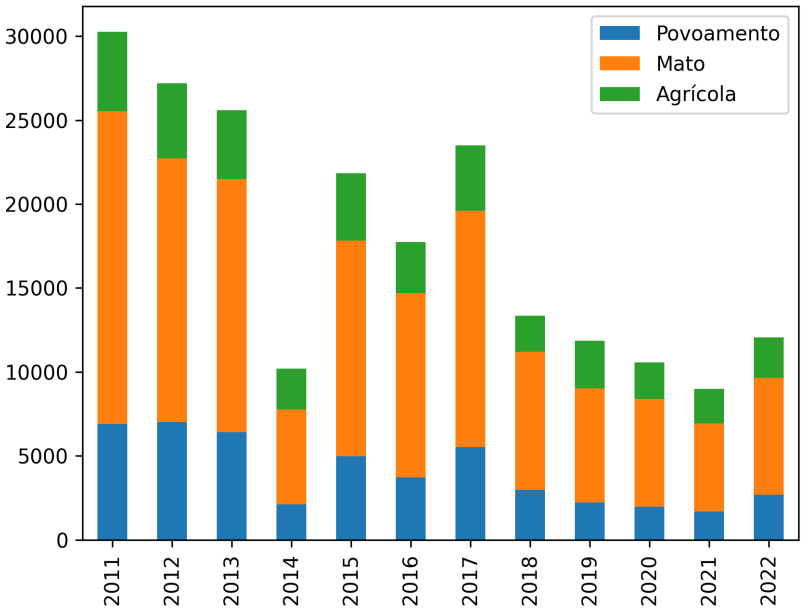


Figura 4.1: Incêndios por tipo de área afetada 2011 - 2022

Quanto ao total de incêndios rurais registrados por ano, conforme a Figura 4.2, 2011 foi o ano com mais registros de incêndios, seguido por 2012, 2013 e 2017, quando esse número permaneceu muito elevado.

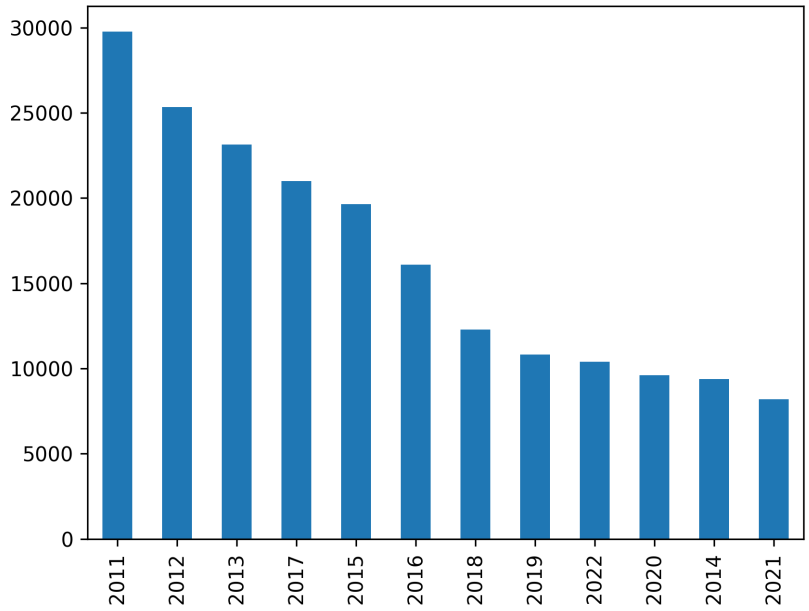


Figura 4.2: Incêndios rurais por ano

Já a evolução da quantidade de incêndios ao longo dos anos é apresentada na Figura 4.3, indicando que o número de incêndios começou muito alto em 2011, diminuindo até 2014. A partir daí voltou a crescer novamente, oscilando até 2018 quando novamente teve uma queda, e depois percebeu-se que, mesmo que pequeno, houve um aumento em 2022.

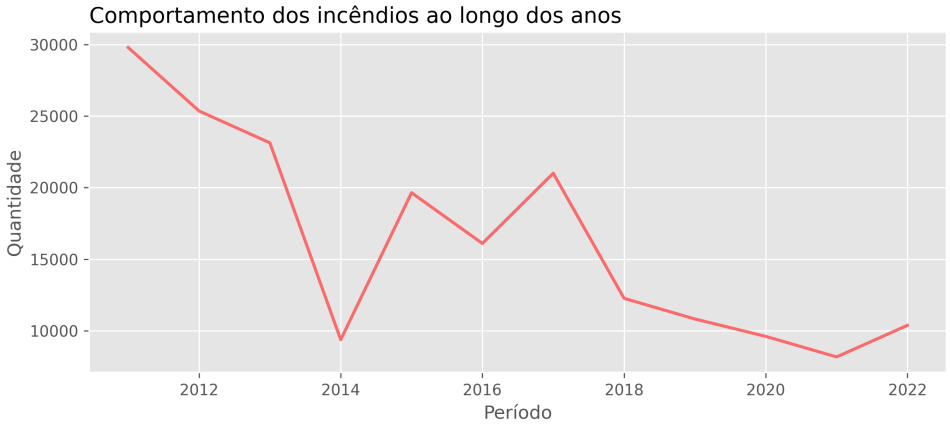


Figura 4.3: Comportamento dos incêndios ao longo dos anos

4.2.2 Áreas de Incêndios por Ano

Nesta análise, estão representadas na Tabela 4.3 as áreas ardidas em ha de cada tipo de incêndio por ano. Pela tabela é possível perceber que as áreas de povoamento, sem levar

em consideração a área total que inclui os três tipos de área, foram as áreas que mais arderam quanto a soma dos valores de cada ano para cada tipo, chegando a um total de 701552.09 ha de área afetada; logo em seguida, as áreas de mato, registrando um total de 640971.02 ha; e, por último, as áreas agrícolas com 102989.78 ha de área ardida. A área total ardida durante todos esses anos correspondeu a 1445512.89 ha referentes ao intervalo de 2011 a 2022.

Tabela 4.3: Área ardida pelos incêndios por tipo de área

Ano	**AreaPov_ha**	**AreaMato_ha**	**AreaAgric_ha**	**AreaTotal_ha**
2011	20036.87	52476.65	4590.01	77103.53
2012	48022.79	61312.19	8649.79	117984.78
2013	55660.14	96657.17	8070.42	160387.73
2014	8723.60	11129.07	2967.15	22819.82
2015	23539.94	39828.40	3831.89	67200.23
2016	77490.83	83696.51	6620.05	167807.39
2017	329513.86	170585.04	39822.10	539920.99
2018	21940.95	19485.81	3150.74	44577.50
2019	21431.59	15912.71	4739.62	42083.92
2020	31725.00	28953.92	6491.02	67169.93
2021	8157.99	17172.24	3029.77	28360.01
2022	55308.54	43761.31	11027.21	110097.06
Soma	**701552.09**	**640971.02**	**102989.78**	**1445512.89**

De acordo com a Figura 4.4, nesse intervalo de 2011 a 2016 as maiores concentrações de áreas ardidas foram em áreas de Mato. A partir de 2017, com exceção de 2021, as áreas de Povoamento ultrapassaram as áreas de Mato em ha de área ardida, sendo 2017 o pior ano, atingindo as maiores áreas em Portugal.

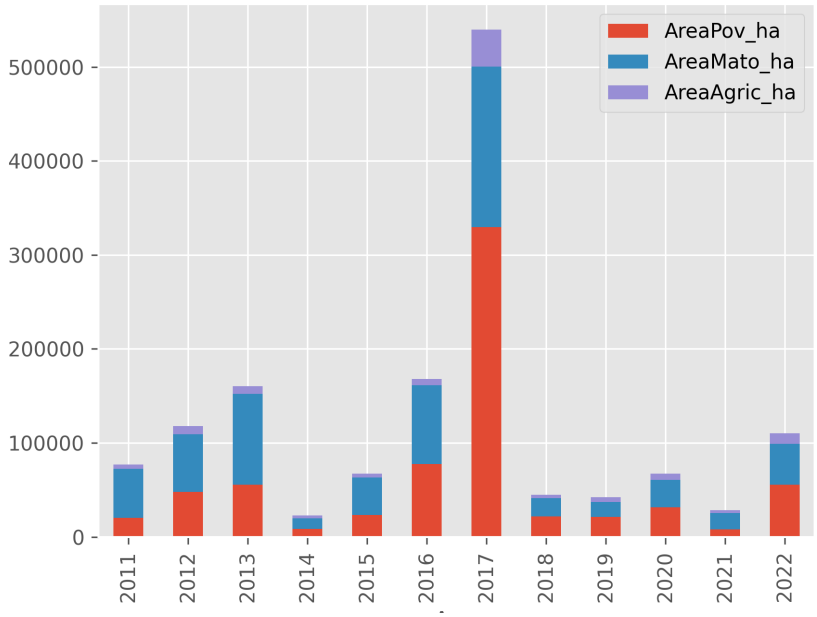


Figura 4.4: Áreas afetadas pelos incêndios por ano (tipos de área)

Quanto à evolução dos incêndios ao longo dos anos por área total ardida, conforme a Figura 4.5, houve uma oscilação nos anos iniciais, e em 2017 ocorreu o pior momento, com um pico representativo em relação aos demais anos, após este pico uma queda em 2018, acompanhado de outras oscilações, com um certo crescimento em 2022.

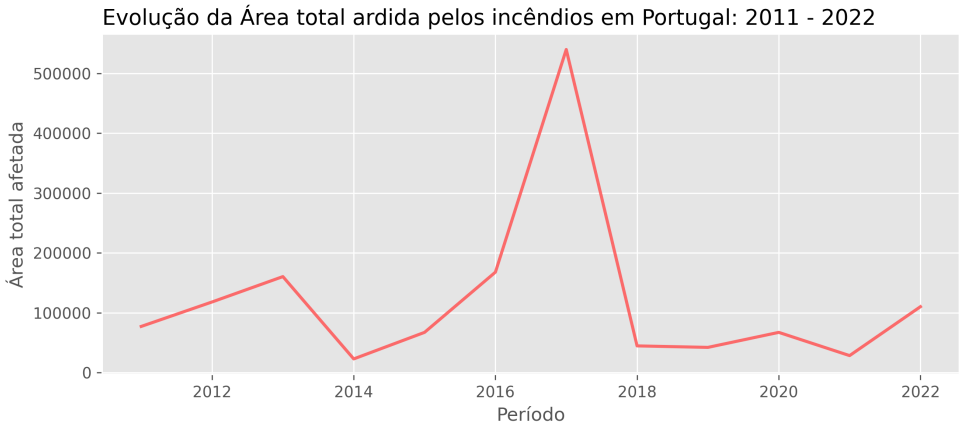


Figura 4.5: Evolução da área total afetada pelos incêndios ao longo dos anos

4.3 Dimensão dos Incêndios

A distribuição do número de incêndios rurais por classe de área ardida Tabela 4.4 aponta que os incêndios com área ardida inferior a 1 ha são os mais frequentes seguidos dos de

área entre 1 a 10 ha. No que se refere a incêndios de maior dimensão, com área ardida superior a 1000 ha, só em 2022, foram registrados 17 incêndios.

De acordo com ICNF, incêndios que possuem a área ardida total igual ou superior a 100 ha, são considerados como grandes incêndios (ICNF, 2022). Entre 2011 e 2022 foram registrados 1386 incêndios enquadrados nesta categoria; já em 2022, o número chegou a 100 ocorrências.

Tabela 4.4: Tabela de Classes de Área por Ano

Ano	]0 a 1 ha[	[1 a 10 ha]	[10 a 20 ha]	[20 a 50 ha]	[50 a 100 ha]	[100 a 500 ha]	[500 a 1000 ha]	[superior a 1000 ha]
2011	24341	4602	319	267	132	97	18	6
2012	20626	3893	300	281	108	117	16	11
2013	18916	3348	259	267	114	166	30	29
2014	8069	1100	75	79	35	26	1	3
2015	16000	3008	253	204	77	77	16	8
2016	13045	2319	240	181	112	151	31	25
2017	16936	3082	311	269	135	178	32	63
2018	10736	1302	110	78	23	21	3	1
2019	9179	1254	144	138	55	47	13	2
2020	8258	1034	115	103	42	45	11	11
2021	6772	1078	136	117	52	28	1	2
2022	8582	1312	176	147	73	71	12	17
Soma	161460	27332	2438	2131	958	1024	184	178

Analisando a Figura 4.6, é possível perceber que as classes com os maiores números de ocorrências foram, respectivamente, ClasseArea (0 a 1 ha) e ClasseArea (1 a 10 ha). Isso quer dizer que a maior parte dos incêndios foram inferiores ou iguais a 10 ha, com pico do número de incêndios em 2011. Em 2017 foi registrado um aumento das ocorrências em áreas maiores.

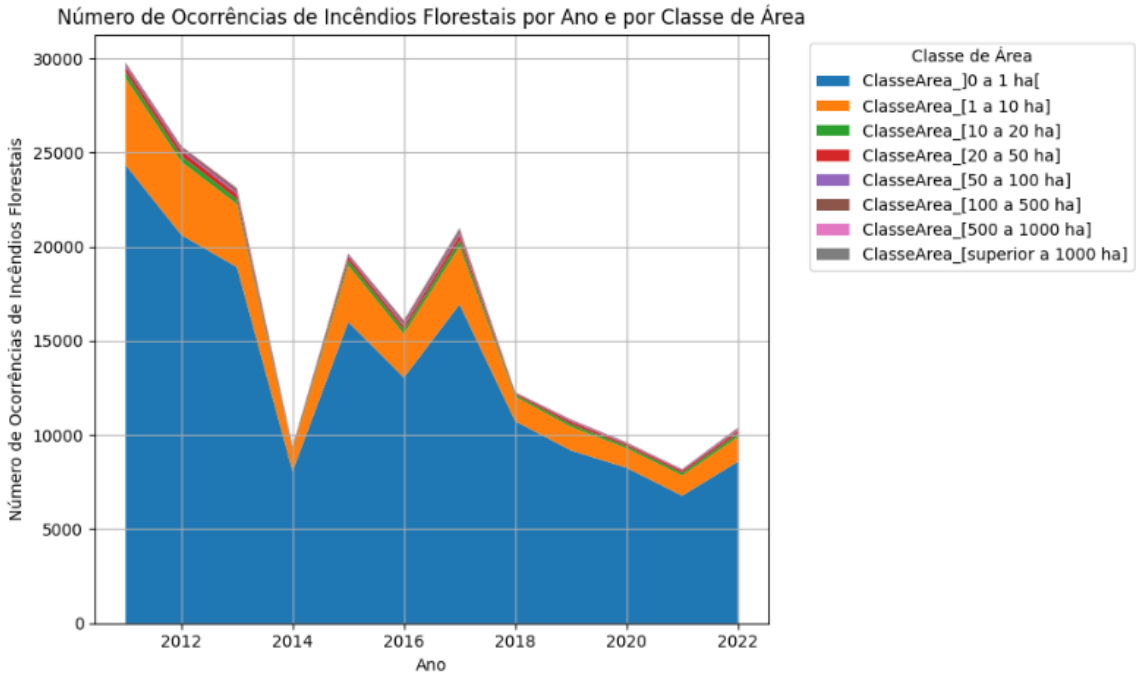


Figura 4.6: Incêndios por Classe

Na Tabela 4.5 estão listados os vinte incêndios de maior dimensão e os respectivos distritos, concelhos e freguesias a que pertencem juntamente com suas respectivas datas de alerta das ocorrências.

Tabela 4.5: Tabela Contendo os Vinte Incêndios Rurais de Maior Dimensão

Distrito	**Concelho**	**Freguesia**	**DataHoraAlerta**	**Pov**	**Mato**	**Agrícola**	**Total**
Coimbra	Lousã	Vilarinho	2017-10-15 08:41:00	47215.43	909.95	5493.42	53618.81
Coimbra	Arganil	Coja	2017-10-15 12:28:00	26532.87	5989.70	5435.94	37958.50
Castelo Branco	Sertã	Várzea Dos Cavaleiros	2017-07-23 13:47:00	16579.06	13178.86	3881.76	33639.68
Castelo Branco	Sertã	Figueiredo	2017-10-15 12:02:00	19279.20	12703.93	1209.48	33192.61
Leiria	Pedrógão Grande	Pedrógão Grande	2017-06-17 14:43:00	30358.84	0.00	0.00	30358.84
Faro	Monchique	Monchique	2018-08-03 13:32:00	15835.82	9801.76	1126.25	26763.82
Faro	Tavira	Cachopo	2012-07-18 14:10:00	5790.00	15647.00	3406.00	24843.00
Castelo Branco	Covilhã	Vila Do Carvalho	2022-08-06 03:18:00	13603.71	8722.87	2006.65	24333.24
Leiria	Alvaiázere	Pussos	11/08/2017 19:40:00	19610.57	1550.19	1663.73	22824.49
Aveiro	Arouca	Janarde	2016-08-08 14:35:00	16431.00	5478.00	0.00	21909.00
Coimbra	Figueira Da Foz	Quiaios	2017-10-15 14:36:00	15487.68	3537.85	0.00	19025.53
Coimbra	Góis	Alvares	2017-06-17 14:52:00	9483.80	8036.85	0.00	17520.65
Leiria	Alcobaça	Pataias	2017-10-15 14:33:00	16864.05	84.12	316.42	17264.59
Viseu	Vouzela	Campia	2017-10-15 17:21:00	11213.96	3110.46	1420.97	15745.38
Castelo Branco	Proença -A-Nova	Sobreira Formosa	2020-09-13 13:43:00	12039.88	2301.75	536.05	14877.68
Bragança	Alfândega Da Fé	Ferradosa	2013-07-09 13:47:00	1982.77	11723.37	429.95	14136.09
Guarda	Seia	Sabugueiro	2017-10-15 06:03:00	5389.85	6469.50	65.28	11924.63
Guarda	Seia	Sandomil	2017-10-15 10:26:00	11328.05	0.00	479.89	11807.94
Castelo Branco	Vila De Rei	Fundada	2019-07-20 14:50:00	6880.95	1861.86	506.14	9248.95
Aveiro	Vale De Cambra	Macieira De Cambra	2017-10-15 07:15:00	6470.61	1746.48	756.54	8973.63

4.4 Análise das Causas

Para análise das causas, as estatísticas foram concentradas no TipoCausa e GrupoCausa, colunas do conjunto de dados que apontam a tipificação da causa e em qual grupo da causa se enquadram um registro de incêndio da base de dados.

4.4.1 Incêndios Rurais por TipoCausa

Começando as análises por TipoCausa, a Tabela 4.6 apresenta a distribuição da quantidade incêndios em cada ano por TipoCausa, essa coluna é correspondente à tipificação da causa de um incêndio (da Conservação da Natureza e das Florestas , ICNF).

Tabela 4.6: Quantidade de incêndios por tipo de causa

Anos	**Desconhecida**	**Intencional**	**Natural**	**Negligente**	**Reacendimento**
2011	6351	3545	89	7076	3736
2012	5826	3624	58	7259	2292
2013	5396	3986	86	5947	2416
2014	2890	1615	49	3253	319
2015	5392	3315	156	5853	1535
2016	4722	2569	75	4042	1384
2017	6661	3439	134	5279	1758
2018	3813	1400	135	4207	723
2019	3578	1977	146	3619	595
2020	3293	2039	117	2738	525
2021	2809	1306	116	3504	205
2022	3765	2011	134	3660	482
Soma	54496	30826	1295	56437	15970

A Figura 4.7 apresenta a quantidade de ocorrências de TipoCausa por ano. Conforme o gráfico, foi possível observar que, na maioria dos anos, o TipoCausa que mais ocasionou incêndios foi o TipoCausa_Negligente, seguido de TipoCausa_Desconhecida e TipoCausa_Intencional. Sendo que, em 2016, 2017, 2020 e 2022, o TipoCausa_Desconhecida superou em número de ocorrências o TipoCausa_Negligente.

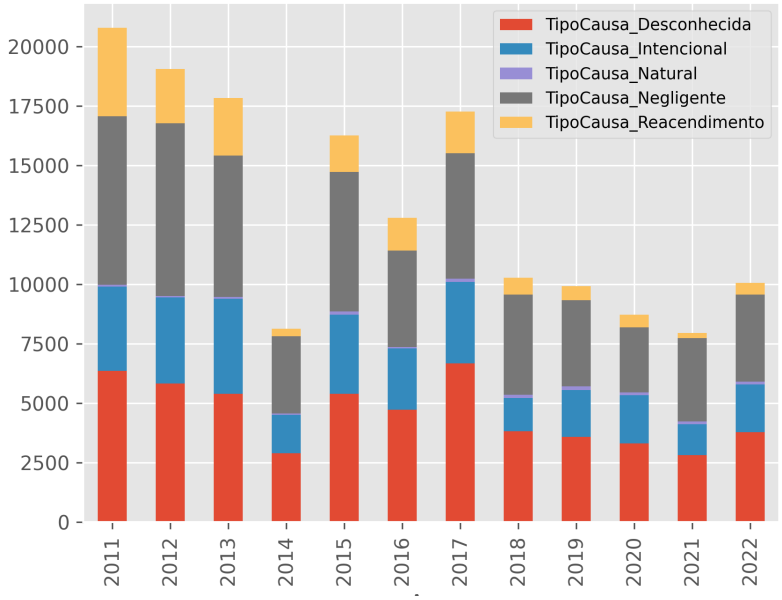


Figura 4.7: Quantidade de incêndios por TipoCausa

Já a quantidade de ocorrências de tipos de causas de incêndios por Distrito está disposta na Tabela 4.7.

Tabela 4.7: Tabela de Incêndios por Distrito e Causa

Distrito	**Desconhecida**	**Intencional**	**Natural**	**Negligente**	**Reacendimento**
Aveiro	7565	1429	14	3433	2694
Beja	1405	619	44	1433	15
Braga	4317	3753	22	3925	2553
Bragança	1093	816	184	3350	315
Castelo Branco	1486	1465	172	1911	20
Coimbra	1009	2137	95	2784	303
Faro	1498	367	7	1930	26
Guarda	916	1055	196	3179	265
Leiria	3028	1313	23	2308	376
Lisboa	9500	116	6	987	65
Portalegre	756	287	75	1823	21
Porto	10793	2267	18	6609	3934
Santarém	1909	2898	49	3792	169
Setúbal	3733	231	10	1493	69
Viana Do Castelo	366	4594	18	4654	2369
Vila Real	569	3199	138	6106	973
Viseu	2921	4239	198	5748	1775
Évora	1632	41	26	972	28

A representação gráfica da distribuição dos tipos de causas de incêndios por Distritos pode ser observada na Figura 4.8. A partir dela verificou-se que no Distrito que registrou a maior quantidade de ocorrências, o Porto, as causas mais predominantes de incêndios foram: desconhecidas, seguidas de negligentes e reacendimentos.

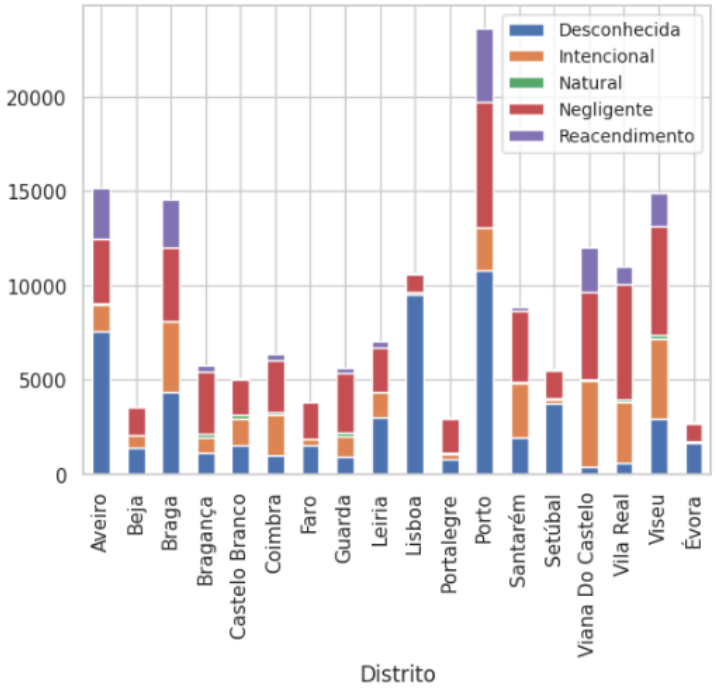


Figura 4.8: Distribuição dos tipos de causas de incêndios por Distrito

4.4.2 Incêndios Rurais por GrupoCausa

De acordo com a Figura 4.9, os grupos de causas que tiveram mais ocorrência foram GrupoCausa_Indeterminadas seguido de GrupoCausa_Incendiarismo - Imputáveis e GrupoCausa_Queimadas de sobrantes florestais ou agrícolas.

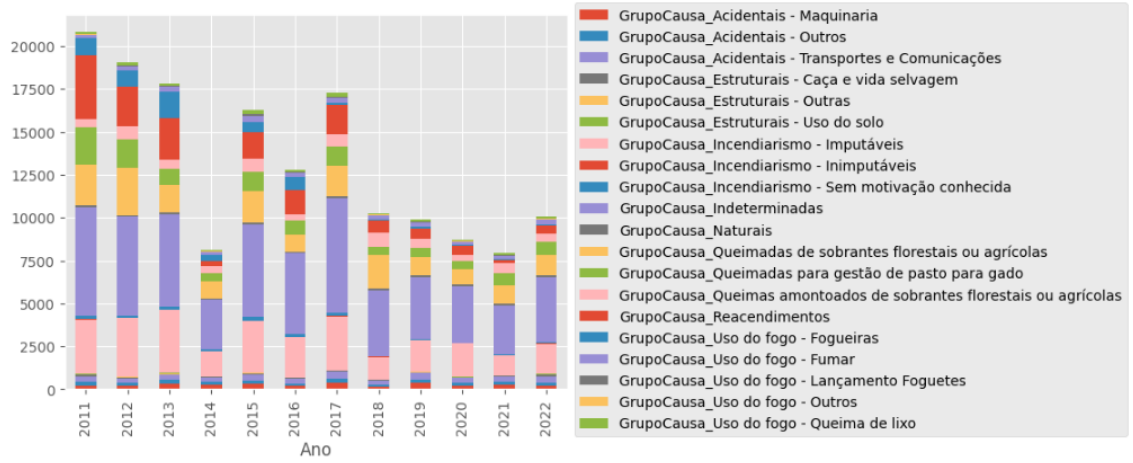


Figura 4.9: Quantidade de incêndios por GrupoCausa

4.5 Análises Regionais

Nesta seção serão apresentadas as análises regionais nas quais o enfoque foi direcionado para a quantidade de incêndios por distrito e concelhos, e nas análises referentes à área ardida por distritos e concelhos. Assim como nos tópicos anteriores, tabelas e figuras foram utilizadas para visualização e auxiliar na compreensão do fenómeno.

4.5.1 Quantidade de Incêndios Rurais por Distritos e Respectivas Extensões de Áreas Ardidas (Ordem Alfabética)

A Tabela 4.8 apresenta a quantidade de incêndios ocorridos em cada Distrito de Portugal de 2011 a 2022, juntamente com as respectivas áreas ardidas por tipo de área ardida e área total. O Porto registrou o maior número de ocorrências em comparação aos outros distritos, chegando a 43200 ocorrências, um total de 60451.27 ha de áreas ardidas. Dentre os tipos de áreas ardidas, AreaMato_ha foi a mais afetada com 34066.08 ha. Por outro lado, Évora registrou apenas 2731 ocorrências de queimadas, sendo a área total afetada de 9728.65 ha e o tipo de área mais afetada foi AreaAgric_ha com 4816.05 ha.

Tabela 4.8: Tabela de Incêndios Rurais e Áreas Correspondentes

Distrito	**Nº Incêndios**	**AreaPov_ha**	**AreaMato_ha**	**AreaAgric_ha**	**AreaTotal_ha**
Aveiro	15250	60725.40	14671.37	1257.07	76653.84
Beja	3722	7440.37	1116.05	11707.90	20264.32
Braga	19657	32033.59	47254.20	443.65	79731.44
Bragança	6758	19201.18	84527.69	5957.71	109686.58
Castelo Branco	5101	104878.68	61035.33	14735.21	180649.21
Coimbra	6528	124582.98	25922.96	11329.58	161835.52
Faro	4376	32229.85	37023.25	6472.94	75726.04
Guarda	5637	51045.03	84611.86	20544.20	156201.10
Leiria	7123	82775.90	10168.26	4027.40	96971.56
Lisboa	14642	2301.74	7424.78	1251.23	10977.75
Portalegre	3045	7876.46	3111.27	4702.87	15690.60
Porto	43200	25768.86	34066.08	616.32	60451.27
Santarém	9283	25195.73	9032.15	6931.49	41159.37
Setúbal	7827	4239.71	1669.52	1357.46	7266.68
Viana Do Castelo	13013	27431.73	56549.25	305.18	84286.17
Vila Real	12844	40859.43	77404.41	4733.79	122997.64
Viseu	14968	48688.53	84746.91	1799.73	135235.16
Évora	2731	4276.92	635.68	4816.05	9728.65

Já a Figura 4.10 exibe, em ordem decrescente, a quantidade de incêndios por Distrito que ocorreram durante o período de 2011 a 2022.

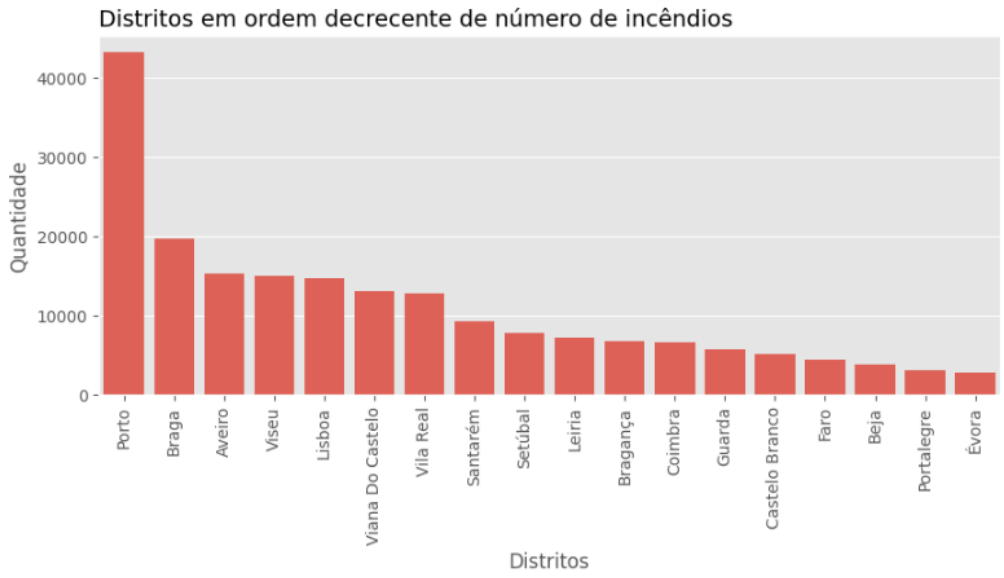


Figura 4.10: Quantidade de incêndios por Distrito em ordem decrescente

Quanto à evolução dos incêndios, foi criada uma lista com os 10 Distritos que registraram as maiores ocorrências ao longo dos anos. Pela Figura 4.11, pode-se acompanhar a evolução da quantidade de incêndios em cada Distrito. O Porto, que obteve as maiores quantidades, iniciou com valores altos, oscilando até 2013. Em 2014 despencou, mas mesmo assim ainda com valores maiores que os demais distritos dessa lista. De 2014 a 2017, esses valores voltaram a crescer. Já de 2017 a 2019 houve diminuição, ocorrendo uma oscilação de 2020 a 2021, tendo sido registrado crescimento em 2022. Outra importante observação é que, em 2014, foi quase que unânime a diminuição da quantidade de incêndios nos Distritos.

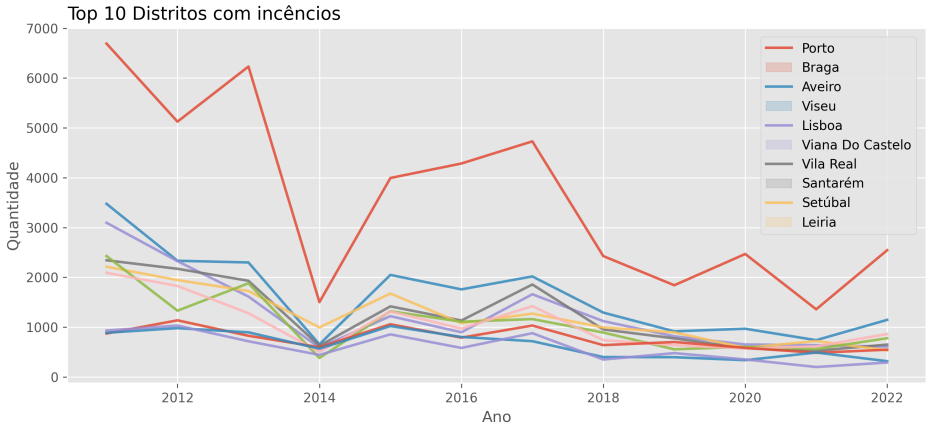


Figura 4.11: Lista Top10 da Evolução dos Incêndios quanto a quantidade por Distrito

Como pode-se observar no mapa de calor gerado 4.12, o Porto é que apresenta a maior concentração de focos de incêndios rurais dentre os Distritos de Portugal.

Figura 4.12: Concentração dos focos de incêndios nos Distritos

4.5.2 Área Ardida por Distritos

Os três distritos com maiores áreas ardidas, por ordem decrescente foram Castelo Branco, Coimbra e Guarda; enquanto que os tipos de áreas mais afetadas foram de Povoamento e Mato, Figura 4.13.

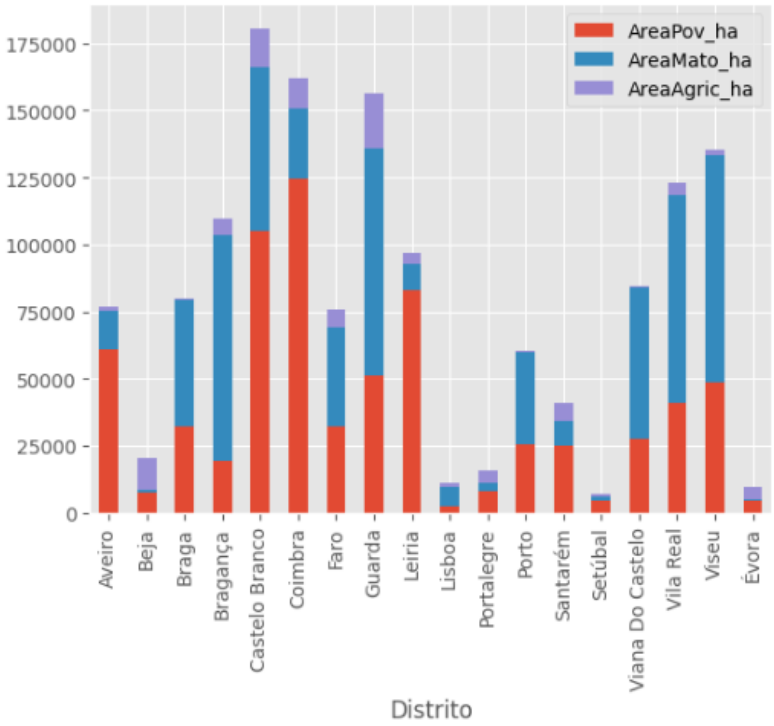


Figura 4.13: Área ardida por Distrito

Também é possível acompanhar a evolução das áreas ardidas nos 10 Distritos que possuem as maiores áreas atingidas por incêndios e percebe-se que, em 2017, houve um

pico na área ardida, um salto muito grande em relação aos demais anos no território Português, Figura 4.14.

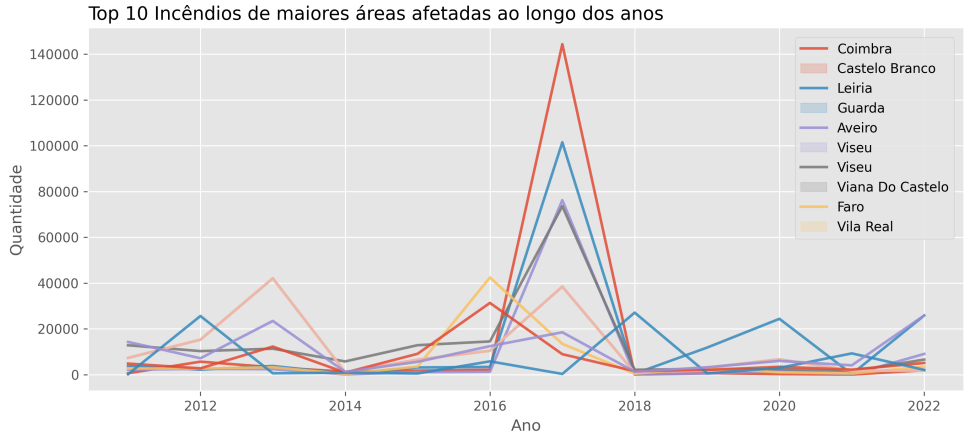


Figura 4.14: Evolução das Áreas ardidas por Distritos

4.5.3 Quantidade de Incêndios por Concelhos

A Tabela 4.9 apresenta a quantidade de incêndios por concelhos que pertencem a lista dos 20 concelhos com as maiores quantidades de ocorrências de incêndios, tendo Paredes como aquele no qual houve o registro de maior número de incêndios 5065 ocorrências, sendo o seu tipo de área ardida mais afetada a AreaMato_ha com 2248.88 ha e sua área total ardida de 4441.69 ha.

Tabela 4.9: Incêndios Rurais e Áreas Correspondentes

Concelhos	Nº Incêndios	AreaPov_ha	AreaMato_ha	AreaAgric_ha	AreaTotal_ha
Paredes	5065	2152.99	2248.88	39.81	4441.69
Penafiel	4809	2770.25	4193.42	39.57	7003.23
Vila Nova De Gaia	3687	531.58	475.27	5.95	1012.80
Gondomar	3463	2318.51	1345.13	19.23	3682.87
Amarante	3232	3687.46	4914.07	70.55	8672.09
Santa Maria Da Feira	3159	1051.89	1354.68	5.28	2411.85
Felgueiras	3151	1154.76	881.93	25.29	2061.99
Santo Tirso	2871	2941.63	850.94	38.75	3831.32
Marco De Canaveses	2802	2170.25	7570.29	101.27	9841.81
Montalegre	2801	2977.08	14849.95	162.78	17989.81
Guimarães	2618	2853.85	1606.06	58.59	4518.51
Ponte De Lima	2544	3270.50	3695.09	18.85	6984.44
Arcos De Valdevez	2519	5038.68	14774.29	18.73	19831.71
Paços De Ferreira	2394	931.60	833.14	12.52	1777.26
Lousada	2372	1006.60	866.93	27.77	1901.30
Sintra	2372	141.78	877.29	229.75	1248.82
Cinfães	2217	1396.68	10502.22	5.92	11904.81
Vila Verde	2139	2518.47	5893.54	43.07	8455.08
Viana Do Castelo	2098	4896.56	6297.81	73.13	11267.50
Fafe	1974	3989.62	6980.37	40.27	11010.26

A Figura 4.15 apresenta, em ordem decrescente, os concelhos com as maiores ocorrências de registros, com Paredes e Penafiel apresentando as maiores quantidades de ocorrências e Fafe e Viana Do Castelo registrando as menores quantidades.

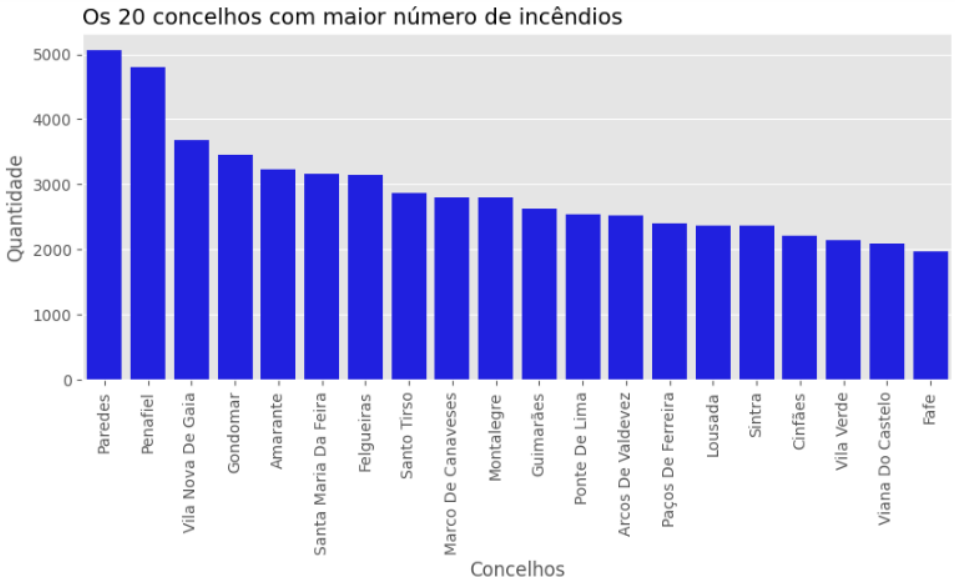


Figura 4.15: Quantidade de incêndios por Concelhos da lista Top20

Apesar de Paredes e Penafiel serem os concelhos onde mais ocorreram incêndios dessa lista, vê-se que eles não tiveram as áreas mais afetadas. Conforme a Figura 4.16, Arcos de Valdevez e Montalegre arderam extensões maiores dentre os vinte concelhos de maiores ocorrências, sendo que áreas de Mato e de Povoamento são as áreas que mais sofreram.

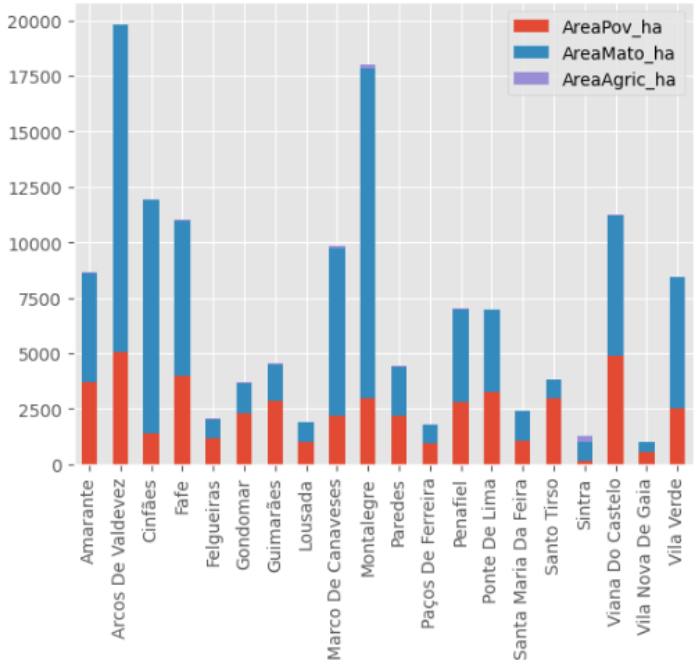


Figura 4.16: Área ardida nos concelhos Concelhos da lista Top20

4.5.4 Área Ardida por Concelhos

Em relação à extensão de área ardida, apesar de não possuir o maior número de ocorrências dentre os Concelhos, Sertã teve 502 ocorrências registrou a maior área ardida, chegando a um total de 73702.46 ha, sendo povoamento o tipo de área mais atingida, com 40892.16 ha, Tabela 4.10.

Tabela 4.10: Área ardida nos Concelhos

Concelhos	Nº Incêndios	A_Pov ha	A_Mato ha	A_Agric ha	A_Total ha
Sertã	502	40892.16	27248.49	5561.81	73702.46
Lousã	247	47650.22	977.19	5496.14	54123.55
Arganil	350	27645.93	6417.75	5442.40	39506.08
Covilhã	956	20685.03	14790.72	3176.99	38652.74
Monchique	172	18452.02	13035.19	1267.06	32754.27
Seia	537	23839.44	7871.12	754.90	32465.46
Pedrógão Grande	245	30924.80	69.48	4.18	30998.46
Arouca	1117	20938.45	8449.49	19.96	29407.91
Guarda	556	7106.51	13371.32	5208.95	25686.78
Tavira	387	5846.07	15771.00	3464.65	25081.71
Alvaiázere	250	20142.60	1904.99	1680.08	23727.67
Alcobaça	713	19464.88	477.19	362.48	20304.55
Figueira Da Foz	724	16165.25	3731.58	23.69	19920.52
Arcos De Valdevez	2519	5038.68	14774.29	18.73	19831.71
Torre De Moncorvo	721	1075.75	18128.20	102.56	19306.51
Góis	168	10882.43	8351.06	6.58	19240.06
Chaves	1638	6938.71	10035.83	2160.35	19134.89
Vouzela	290	12853.97	4364.49	1422.13	18640.59
Montalegre	2801	2977.08	14849.95	162.78	17989.81
Freixo De Espada À Cinta	475	2090.91	14080.80	1764.47	17936.19

A Figura 4.17 apresenta os Concelhos por ordem decrescente de área ardida, com Sertã e Lousã sendo os mais atingidos. Por outro lado, Freixo De Espada À Cinta e Montalegre obtiveram as menores áreas ardidas.

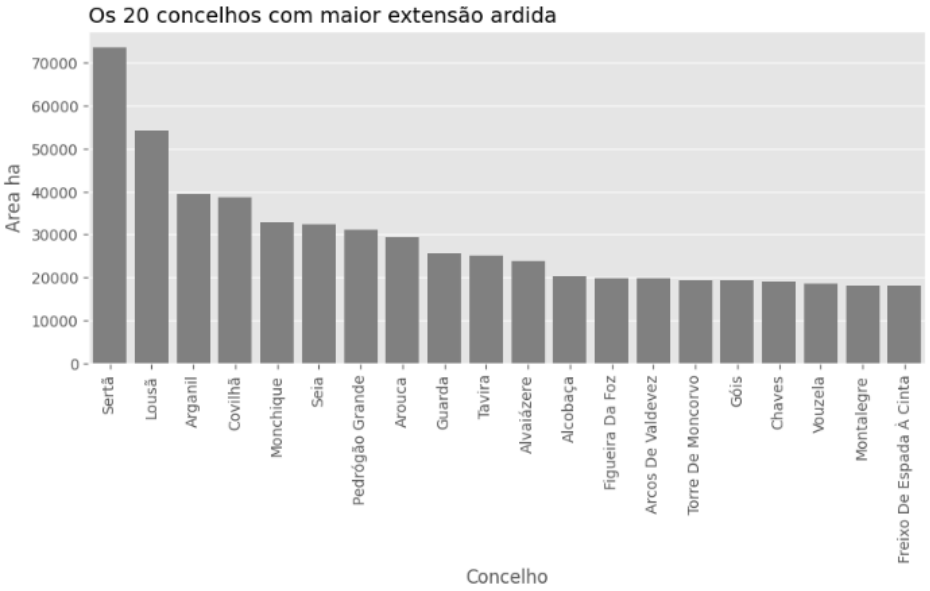


Figura 4.17: Os vinte Concelhos que mais arderam

Conforme a Figura 4.18, Montalegre e Arcos De Valdevez possuem os maiores números de ocorrências entre os concelhos desta lista, enquanto que Góis e Monchique as menores quantidades.

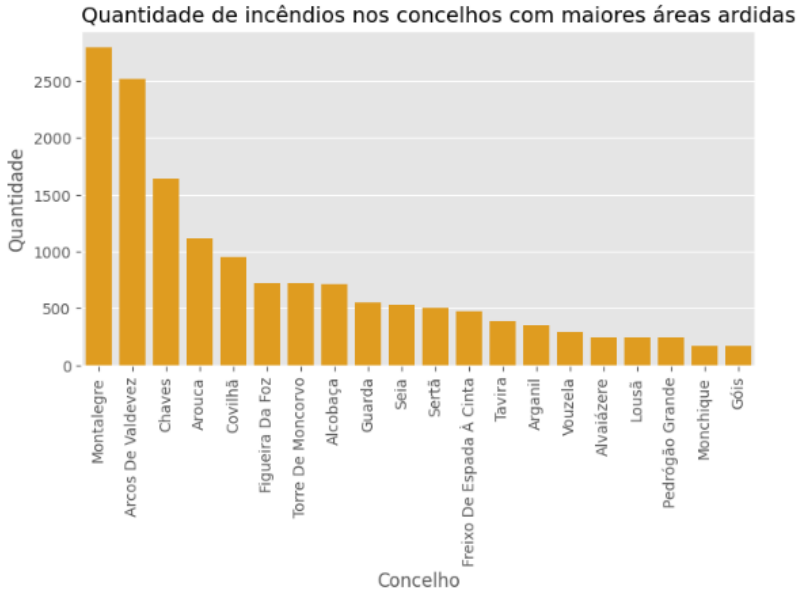


Figura 4.18: Quantidade de incêndios nos Concelhos que mais arderam

4.6 Análises Mensais

Nesta seção as análises são referentes às ocorrências mensais dos incêndios florestais em Portugal. Conforme a Figura 4.19, os incêndios estão concentrados entre os meses de julho a outubro, potencializando o aumento do número de ocorrências e a devastação de maiores áreas por *hectares* conforme análises anteriores. Um outro aspecto perceptível neste gráfico é a existência de *outliers* presentes nos meses de fevereiro a abril e setembro a dezembro, o que demonstra uma quantidade de ocorrências atípicas dos incêndios durante estes períodos.

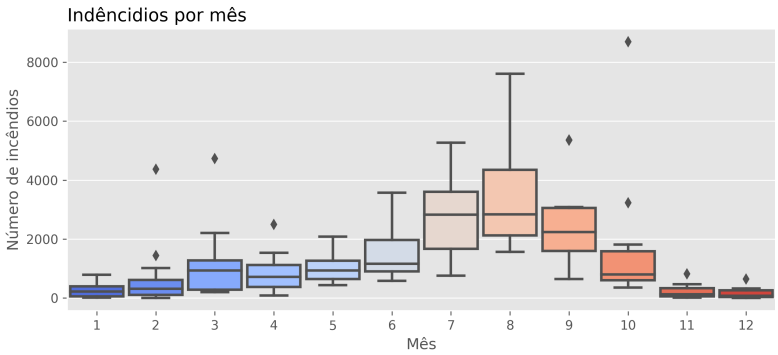


Figura 4.19: Incêndios por mês

O momento mais agitado é justamente entre os meses de julho a outubro Figura 4.19, com um pico representativo formado no mês de agosto Figura 4.20, indicando que as maiores ocorrências do ano acontecem justamente neste mês.

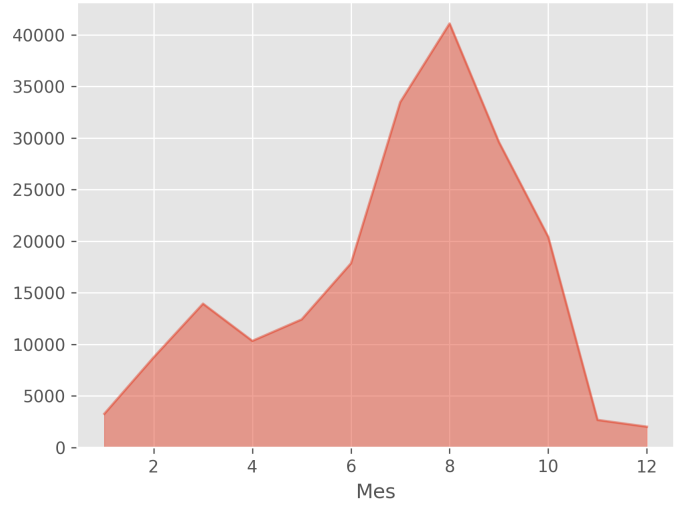


Figura 4.20: Gráfico de Incêndios por Mês

4.6.1 Quantidade de Incêndios por Tipo de Área em Cada Mês

Durante todos os meses de 2011 a 2022, o tipo de área ardida que registrou a maior quantidade de incêndios foi a área de Mato, com 126638 registros, Tabela 4.11. Por outro lado, a menor quantidade de incêndios foi registrada em área Agrícola com 38316 ocorrências. O mês mais intenso foi agosto, que registrou a maior quantidade de ocorrências 27651, em áreas de Mato. Já as menores quantidades foram registradas em dezembro, com 376 ocorrências em áreas de Povoamento.

Tabela 4.11: Quantidade de incêndios rurais por tipo de área em cada mês

Mês	Povoamento	Mato	Agrícola	Pov + Mato + Agr
1	803	2403	536	3742
2	2239	6234	1328	9801
3	4374	9602	1759	15735
4	3576	6942	1166	11684
5	3272	7445	2738	13455
6	3998	9820	4745	18563
7	7799	20368	7786	35953
8	10180	27651	7606	45437
9	7006	19859	5361	32226
10	4122	13355	4245	21722
11	434	1703	661	2798
12	376	1256	385	2017
Soma	48179	126638	38316	213133

Confirmando as estatísticas anteriores, a Figura 4.21, apresenta a concentração dos incêndios entre os meses de julho a outubro, com destaque para agosto, tendo como tipo de área ardida predominante durante todos os meses a área de Mato. Quanto a área de Povoamento, percebeu-se que quantidade de incêndios começou a aumentar entre os meses de junho a agosto, e diminuir nos meses seguintes, setembro a dezembro.

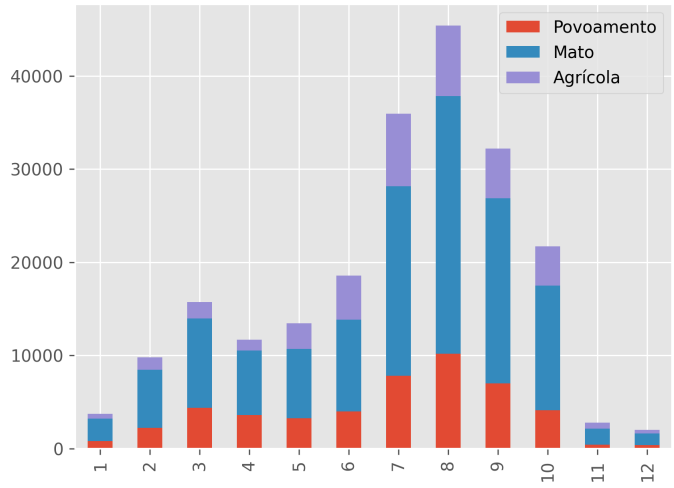


Figura 4.21: Gráfico da quantidade de tipos de incêndios por mês

4.6.2 Distribuição de Áreas Ardidas por Mês

Conforme a Tabela 4.12, o mês de agosto foi responsável pelas maiores áreas ardidas no território Português, registrando o seu maior valor de área total ardida 482638.89 ha. Desse total, 239649.98 ha foram de área de mato, o que a colocou em primeiro lugar em área ardida por tipo de área, seguida por área de povoamento com 218081.20 ha.

Tabela 4.12: Tabela de áreas ardidas por mês

Mês	ÁreaPov_ha	ÁreaMato_ha	ÁreaAgric_ha	ÁreaTotal_ha
1	2059.52	9541.55	231.57	11832.64
2	4409.08	14504.04	429.96	19343.08
3	18000.02	27475.21	567.01	46042.24
4	10507.97	12096.94	262.11	22867.03
5	5953.78	5058.74	1917.18	12929.71
6	52780.16	21279.97	7566.37	81626.50
7	109074.43	114591.34	32874.15	256539.92
8	218081.20	239649.98	24907.70	482638.89
9	75952.54	86875.91	9259.81	172088.26
10	203501.94	104899.57	24545.96	332947.48
11	647.37	2950.56	321.55	3919.49
12	584.06	2047.20	106.40	2737.67

Conforme pode-se observar na Figura 4.22, as maiores áreas ardidas concentraram-se entre os meses de julho a outubro assim como aconteceu com o número de ocorrências. Em agosto as áreas ardidas atingiram os patamares mais elevados, sendo que em outubro e julho estes valores também continuaram elevados. Em relação aos tipos de áreas ardidas as áreas de matos superaram os demais tipos na maior parte dos meses, seguidas das áreas de povoamento, que nos meses de maio, junho e outubro ultrapassaram as áreas de matos em *hectares* ardidos.

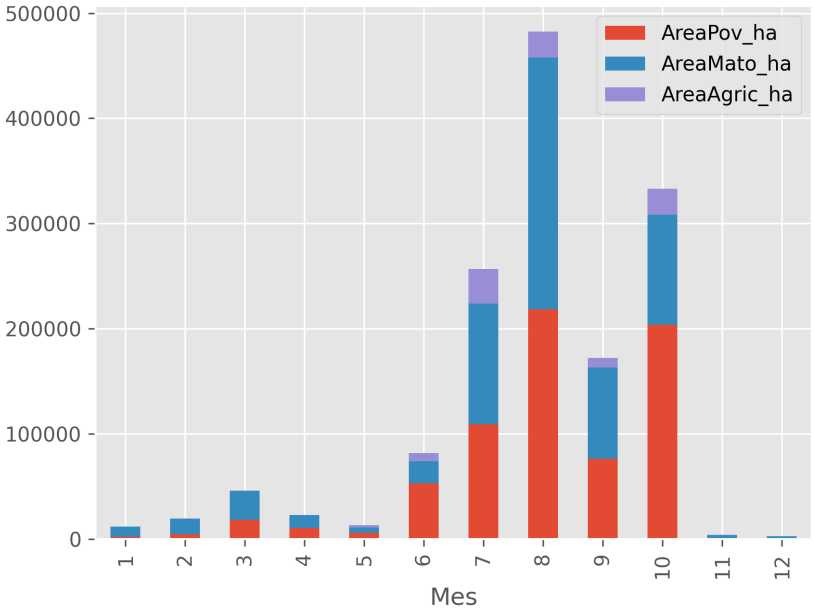


Figura 4.22: Gráfico dos tipos de áreas dos incêndios por mês

4.7 Análise da Severidade Meteorológica

Nesta seção as estatísticas estão concentradas na quantidade de incêndios rurais por classe de severidade meteorológica durante os anos e ao longo dos meses!

4.7.1 Número de Incêndios Rurais Anuais por Classe de Severidade Meteorológica 2011-2022 (DSR)

Conforme a Tabela 4.13, as classes estão representadas por escalas que variam de 0 a 6. A escala 0 registrou a maior quantidade de incêndios rurais, sendo que, em 2011, a quantidade chegou a 25394 registros, a maior que nos outros anos.

Tabela 4.13: Escala de DSR por ano

Ano	DSR_Esc 0	DSR_Esc 1	DSR_Esc 2	DSR_Esc 3	DSR_Esc 4	DSR_Esc 5	DSR_Esc 6
2011	25394	4244	144	0	0	0	0
2012	20030	4356	746	79	0	0	0
2013	15235	7460	331	19	6	0	0
2014	8270	1002	83	0	0	0	0
2015	16728	2700	201	0	3	0	0
2016	11409	3642	982	50	7	0	0
2017	15756	4410	674	126	15	0	0
2018	10477	1597	182	11	2	0	0
2019	9120	1343	211	25	0	0	0
2020	7837	1566	170	14	3	0	0
2021	7141	932	100	11	1	0	0
2022	7437	2178	532	135	55	24	5

A severidade meteorológica diária local é representada indiretamente pelo índice DSR, onde valores elevados de DSR correspondem a níveis de severidade meteorológica elevada (tendencialmente, temperaturas elevadas, vento forte, ausência de precipitação e humidade relativa baixa)(ICNF, 2022).

A Figura 4.23 mostra que as maiores ocorrências de incêndios aconteceram em 2011 e foram ocasionadas por DSR_Escala 0. Entretanto, em 2012, 2016, 2017 e 2022, percebeu-se que houve uma elevação da Escala do DSR em relação aos demais anos. Vale ressaltar que valores elevados do índice DSR estão intrinsecamente relacionados a temperaturas elevadas, vento forte, ausência de precipitação e humidade relativa baixa (ICNF, 2022), o que pode ajudar a compreender um pouco do cenário de 2017 em Portugal.

No ano de 2017, essa elevação do índice DSR pode ter causado impactos em Portugal pois, de acordo com Ramos, a influência de fatores meteorológicos como uma seca prolongada que ocasionou um stresse hídrico cumulativo pré-condicionado da vegetação de outubro de 2017 somados à passagem do furacão Ophelia e o agente humano causador de uma elevada quantidade de ignições negligentes levaram Portugal ao seu pior cenário (Ramos et al., 2023).

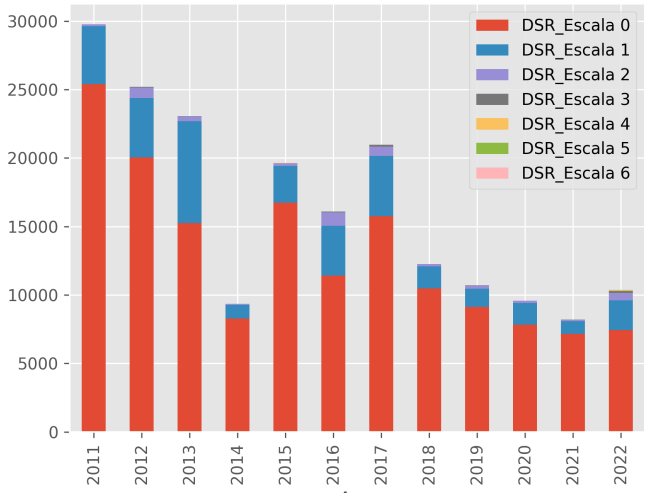


Figura 4.23: Quantidade de Incêndios por Classe de Severidade Durante os Anos

4.7.2 Número de Incêndios Rurais Mensais por Classe de Severidade Meteorológica Janeiro - Dezembro

Conforme a Tabela 4.14, o mês de agosto foi o mês que registrou as maiores quantidades de ocorrências, com a classe DSR_Esc 0 registrando o maior número de incêndios - 26701 ocorrências.

Tabela 4.14: Escala de DSR por mês

Mês	DSR_Esc 0	DSR_Esc 1	DSR_Esc 2	DSR_Esc 3	DSR_Esc 4	DSR_Esc 5	DSR_Esc 6
1	3248	0	0	0	0	0	0
2	8663	8	0	0	0	0	0
3	13641	191	43	3	0	0	0
4	10247	45	0	0	0	0	0
5	11408	853	109	6	0	0	0
6	14828	2734	226	13	2	0	0
7	23888	8123	1105	189	57	24	5
8	26701	12325	1850	106	19	0	0
9	21346	7415	665	70	3	0	0
10	16204	3734	358	83	11	0	0
11	2656	2	0	0	0	0	0
12	2004	0	0	0	0	0	0

A quantidade mensal de incêndios por escala DSR está apresentada na Figura 4.24. As escalas com maiores concentrações de incêndios foram as DSR 0 e DSR 1, sendo de julho a outubro como os meses de maiores incidências. Nesse período, percebeu-se que escalas mais elevadas do índice DSR passaram a registrar uma quantidade bem maior de incêndios quando comparadas aos demais meses. Conforme exposto anteriormente, essa elevação é um reflexo das condições climáticas e meteorológicas que impactam diretamente no número de ocorrências.

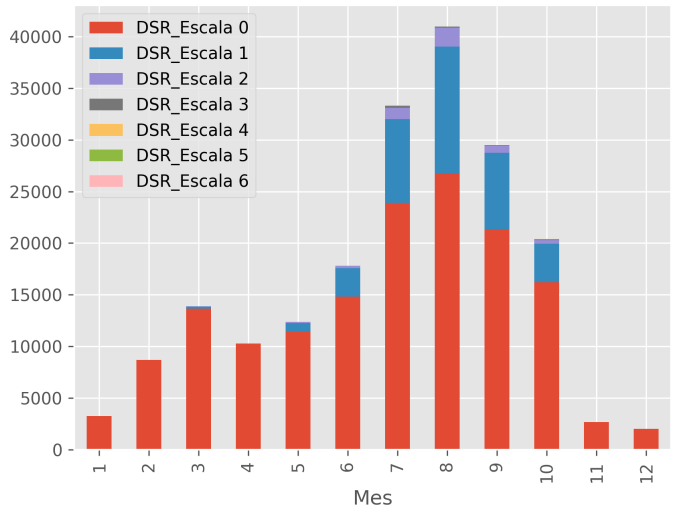


Figura 4.24: Quantidade de Incêndios por Classe de Severidade Durante os Meses

4.8 Análise do Perigo de Incêndio Rural FWI

Nesta seção são apresentadas as estatísticas relacionadas à quantidade de incêndios rurais por classe de perigo de incêndio rural que ocorreram ao longo dos anos e dos meses.

4.8.1 Número de Incêndios Rurais Anuais por Classe de Perigo de Incêndio Rural

De acordo com a Tabela 4.15, o ano de 2011 registrou a maior quantidade de incêndios por classe de perigo de incêndio rural, e a classe FWI_escala 1 responsável pelo maior número, 13636 ocorrências.

Tabela 4.15: Escala de FWI por ano

Ano	FWI_Esc 0	FWI_Esc 1	FWI_Esc 2	FWI_Esc 3	FWI_Esc 4	FWI_Esc 5	FWI_Esc 6
2011	7422	13636	7952	743	29	0	0
2012	10110	7977	5217	1651	256	0	0
2013	3981	7126	10472	1399	67	6	0
2014	3923	3507	1677	246	2	0	0
2015	6486	8301	4086	723	33	3	0
2016	3334	5823	4918	1809	199	7	0
2017	6495	7205	5232	1816	215	18	0
2018	6328	3263	2151	477	48	2	0
2019	4942	3357	1810	521	67	2	0
2020	3120	3566	2402	461	37	4	0
2021	3947	2553	1362	299	23	1	0
2022	3834	2482	2698	993	269	72	18

A Figura 4.25 apresenta a distribuição anual dos incêndios por classes FWI. As classes FWI_Escala 0, FWI_Escala 1 e FWI_Escala 2 concentraram as maiores quantidades de registros de incêndios. Conforme análises anteriores 2011, 2012, 2013 e 2017 registraram os maiores números de ocorrências. Em 2022, o número de incêndios em classes de escalas maiores aumentou.

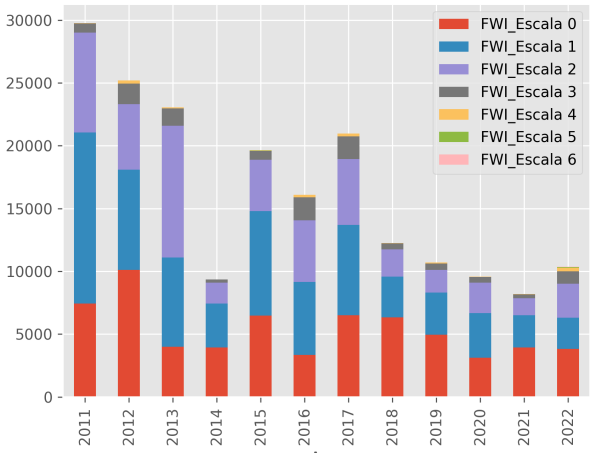


Figura 4.25: Quantidade de Incêndios por Classe de Perigo de Incêndio Rural Durante Anos

4.8.2 Número de Incêndios Rurais Mensais por Classe de Perigo de Incêndio Rural

Conforme a Tabela 4.16, agosto foi o mês que registrou a maior quantidade de incêndios, sendo a classe FWI_escala 2 a que registrou a maior quantidade de ocorrências com 16754 focos de incêndios.

Tabela 4.16: Escala de FWI por Mês

Mês	FWI_Esc 0	FWI_Esc 1	FWI_Esc 2	FWI_Esc 3	FWI_Esc 4	FWI_Esc 5	FWI_Esc 6
1	3226	22	0	0	0	0	0
2	7903	752	16	0	0	0	0
3	10912	2585	279	93	9	0	0
4	7373	2789	127	3	0	0	0
5	7416	3413	1230	292	25	0	0
6	5619	7491	3822	821	48	2	0
7	3629	15534	10838	2885	414	73	18
8	3379	16272	16754	4207	369	20	0
9	4739	11943	10791	1806	214	6	0
10	5476	7596	6107	1031	166	14	0
11	2324	321	13	0	0	0	0
12	1926	78	0	0	0	0	0

A Figura 4.26 apresenta a concentração dos incêndios nas classes FWI_Escala 0, FWI_Escala 1 e FWI_Escala 2. Entretanto entre os meses de maior incidência concentrados entre julho a outubro, percebeu-se um aumento do número de incêndios nas classes de escalas maiores. Tal fato leva a crer que isso tenha relação direta com o verão e a seca durante essa época do ano, tendo em vista que quantificam os efeitos da humidade do combustível e do vento no comportamento do fogo, evidenciando uma relação entre a quantidade de incêndios e as escalas mais elevadas de perigo de incêndio.

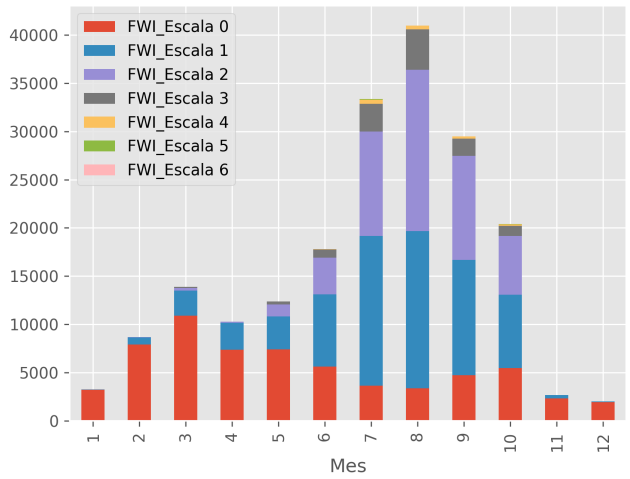


Figura 4.26: Quantidade de Incêndios por Classe de Perigo de Incêndio Rural Durante Meses

4.9 Correlação entre Variáveis

Durante a etapa de mineração, o enfoque central foi encontrar um modelo preditivo que permitisse classificar em que classe de área ardida se enquadraria um registro histórico da base de dados. Em busca das variáveis mais promissoras para atingir o objetivo, foi utilizada a técnica correlação com uma característica, um recurso proveniente da biblioteca *yellowbrick*.

Dessa forma, foram realizados testes com a variável ClasseArea para identificar dentre as características da base de dados quais que melhor se relacionavam com ela. Também foram realizados testes com a variável AreaTotal_ha com intuito de encontrar aquelas que possuíam uma relação mais forte consequentemente gerando uma maior influência quanto a Área total ardida por um registro de incêndio.

4.9.1 Verificando a Correlação com ClasseArea

A Figura 4.27 apresenta a correlação com uma característica, no caso ClasseArea, uma variável inicialmente categórica, codificada em numérica para aplicação das técnicas de correlação, que indica em qual faixa de área ardida um registro se enquadra. Como pode-se perceber, as correlações apresentaram-se muito fracas, existindo tanto correlações negativas quanto positivas. Conforme Castro e Ferrari, a relação entre as variáveis, variam de -1 a 1, onde o valor 0 indica ausência de correlação; o valor igual a -1 indica uma correlação negativa perfeita em que o valor de uma variável é inversamente proporcional ao valor da outra variável e, quando o valor é igual a 1, indica uma correlação positiva perfeita em que o valor de uma variável é diretamente proporcional ao valor dá outra variável (Castro and Ferrari, 2016).

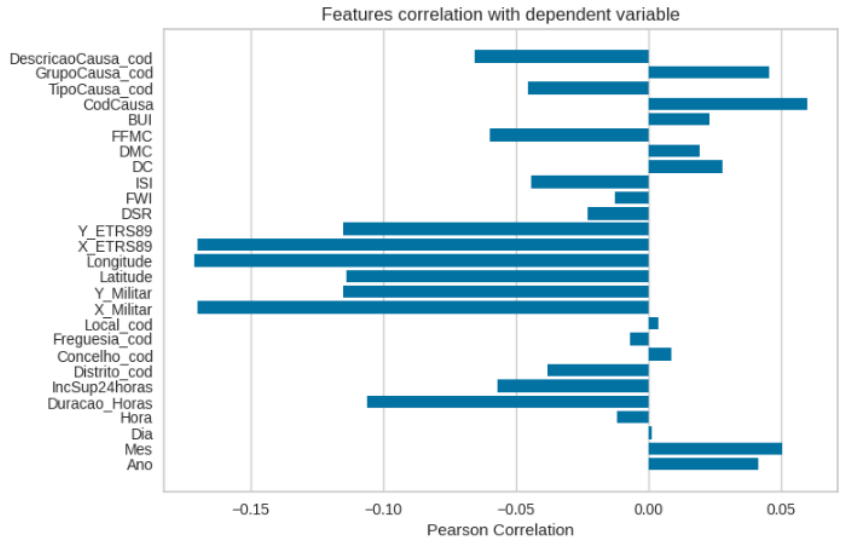


Figura 4.27: Correlação com a variável ClasseArea

Conforme a figura, as correlações negativas representadas pelas variáveis de coordenadas geográficas destacaram-se apresentando os maiores valores negativos. Outras variáveis relevantes foram respectivamente de características referentes à duração dos incêndios: "Duracao_Horas", "IncSup24horas", as de causa: DescricaoCausa_cod, TipoCausa_cod, as do componente do índice genérico de perigo de incêndio rural FWI: FFMC e ISI, as de local: Distrito_cod e a de severidade meteorológica representada pelo DSR. Já em relação às correlações positivas, o destaque foi para as variáveis de causa: CodCausa e GrupoCausa_cod, nas de período de ocorrência: Mes e Ano, nas do componente do índice genérico de perigo de incêndio rural FWI: DC, BUI e DMC e na de localidade: Concelho_Cod.

4.9.2 Verificando a Correlação com a Área Total Ardida

Ao se analisar a Figura 4.28, percebe-se que a área total ardida por um incêndio, uma variável numérica mensurada em ha obteve correlações negativas e positivas, sendo as positivas mais fortes de acordo com a representação gráfica. Uma relação maior foi verificada com a duração de um incêndio: IncSup24horas e Duracao_Horas. Também se observou que a severidade meteorológica e os componentes do índice genérico de perigo de incêndio rural FWI também influenciam na área afetada por um incêndio rural, conforme apresenta a Figura.

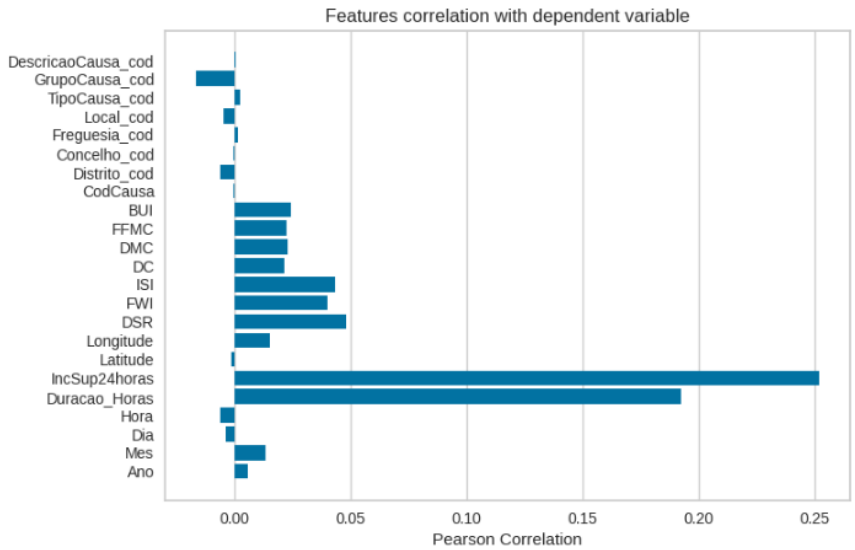


Figura 4.28: Correlação com a variável área total ardida

4.10 Discussão

Em meio a um processo de elevado crescimento e volume de dados, surgiu a *Mineração de Dados*, como um processo sistemático, interativo e iterativo de preparação e extração de

conhecimento oriundo de grandes conjuntos de dados (Castro and Ferrari, 2016). Morgan ressalta o uso de programação associada à ciência de dados para lidar com grandes bases de dados facilitando a extração de *insights* para usá-los para tomar decisões melhores (Morgan, 2016). Dessa forma, ao se optar por uma estratégia mais *Data Mining* para exploração dos dados do que apenas estatística convencional, levou-se em conta que esse tipo de abordagem:

- é mais eficaz para lidar com grandes volumes de informações, podendo lidar com dados em grande escala e extrair conhecimento útil deles;
- permite descobrir padrões complexos e não triviais nos dados que podem não ser detectados por análises estatísticas tradicionais;
- as descobertas do *Data Mining* podem fornecer informações valiosas para apoiar a tomada de decisões; e
- permite a construção de modelos preditivos e modelos descritivos mais avançados, o que pode ser útil para prever tendências futuras e tomar decisões baseadas em análises mais sofisticadas.

Partindo do exposto, neste tópico é apresentado alguns pontos importantes da discussão sobre a exploração dos dados realizada neste capítulo.

4.10.1 Contextualização dos Resultados

A análise abrangente dos registros de incêndios florestais em Portugal, em consonância com o que foi apresentado nas seções anteriores, revela uma imagem complexa e multifacetada desse fenômeno recorrente e impactante. Ao mergulharmos nas estatísticas anuais, áreas afetadas, dimensão dos incêndios, causas e análises regionais, pudemos discernir padrões notáveis e tendências que nos fornecem *insights* valiosos sobre os fatores subjacentes e as dinâmicas que moldam o cenário dos incêndios florestais no país. A Tabela 4.17 apresenta um resumo desses padrões identificados nos dados.

Tabela 4.17: Resumo de padrões identificados

Fatores	**Associados**
Meses: De julho a Outubro	Ampliação dos incêndios florestais
Valores elevados de DSR	Ampliação dos incêndios florestais
Valores elevados de FWI	Ampliação dos incêndios florestais
Ação Humana: negligente, desconhecida ou intencional	Ampliação dos incêndios florestais
Incêndio superior a 24 horas	Ampliação da área ardida
Longa duração do incêndio	Ampliação da área ardida

As descobertas refletem uma realidade preocupante, onde incêndios florestais são uma preocupação constante e desafiadora para as comunidades, gestores de recursos naturais e autoridades governamentais em Portugal. Os números impressionantes de ocorrências, áreas ardidas e tipos de áreas afetadas indicam a magnitude do impacto desses eventos, bem como a necessidade urgente de estratégias de gestão e prevenção mais eficazes.

4.10.2 Relação com a Literatura

Comparando os resultados obtidos neste trabalho com estudos anteriores sobre incêndios florestais a exemplo dos relatórios gerados pelo ICNF, podemos constatar tanto consonâncias quanto discrepâncias significativas. Os achados corroboram com a compreensão amplamente estabelecida de que os meses de verão, especialmente entre julho e outubro, são caracterizados por um aumento substancial na quantidade de incêndios, corroborando a influência sazonal de condições meteorológicas adversas. Essa consistência com a literatura existente confirma a confiabilidade dos dados utilizados em nossa análise.

No entanto, destaca-se a importância de considerar fatores contextuais específicos de Portugal. O aumento notável na severidade dos incêndios em certos anos, como em 2017, revela a interação complexa entre fatores meteorológicos (por exemplo, a passagem de furacões), causas humanas e condições do terreno. Nesse sentido, nossos resultados acrescentam um entendimento mais detalhado das implicações de eventos climáticos extremos ao incluir variáveis como componentes do índice genérico de perigo de incêndio rural FWI em nossas análises, explorar mais a severidade meteorológica representada pelo DSR e detalhar mais as causas dos incêndios.

4.10.3 Fatores Influenciadores

Uma análise mais aprofundada dos fatores influenciadores revela que a duração do incêndio, as condições meteorológicas, o perigo de incêndio e as causas desempenham papéis interconectados na determinação da quantidade e gravidade dos incêndios. A correlação entre a área total ardida e a duração do incêndio, assim como a relação entre o índice FWI e a área afetada, destacam a importância das condições climáticas e do comportamento do fogo no desenvolvimento e propagação dos incêndios.

As causas humanas, especialmente a negligência e ações intencionais, emergem como um fator crucial que requer atenção urgente. O aumento das ocorrências de causas desconhecidas em anos recentes levanta questões sobre a necessidade de um maior esforço na identificação e compreensão das origens desses eventos.

4.10.4 Implicações para a Gestão e Prevenção:

Os resultados da análise têm implicações significativas para a gestão e a prevenção de incêndios florestais em Portugal. A concentração de incêndios em determinados meses e regiões sugere que estratégias de prevenção e preparação devem ser intensificadas durante os períodos críticos de maior risco. A implementação de medidas de monitoramento, alerta precoce e educação pública pode desempenhar um papel fundamental na redução do impacto desses eventos.

A análise também aponta para a necessidade de abordagens integradas que considerem tanto fatores climáticos quanto comportamentais. A promoção de práticas sustentáveis de uso da terra, a conscientização da população sobre o risco de incêndios e a adoção de medidas preventivas podem contribuir para minimizar a incidência e a severidade dos incêndios florestais.

4.10.5 Limitações e Futuras Pesquisas

Apesar do avanço proporcionado por esta análise, ela possui limitações e algumas questões permanecem em aberto e sugerem direções para pesquisas futuras:

- Padrões Temporais e Climáticos: embora tenha sido explorada a relação entre a época do ano e a incidência de incêndios, uma análise mais detalhada dos padrões temporais e climáticos pode proporcionar *insights* adicionais. Como as mudanças climáticas estão afetando os padrões de incêndios? Existem outras variáveis climáticas que podem estar contribuindo para o aumento dos incêndios florestais?

- Causas Específicas: por mais que tenha sido identificado algumas causas predominantes dos incêndios, como negligência e intencionalidade, uma investigação mais profunda poderia esclarecer os fatores subjacentes a essas causas. Quais são os principais motivos por trás das ações negligentes ou intencionais? Como os fatores socioeconômicos e culturais influenciam esses comportamentos. Essas são limitações decorrentes da própria base de dados que não possui esses dados adicionais.

Capítulo 5

Machine Learning: Modelagem Preditiva dos Incêndios Florestais

Neste Capítulo, as atividades foram concentradas na geração de um modelo preditivo que fosse capaz de prever a qual faixa de área ardida um registro histórico da base de dados pertenceria com métricas confiáveis de acerto.

A base de dados, conforme exposto em seções anteriores, é composta por registros reais de incêndios que ocorreram em Portugal de 2011 a 2022.

Também serão discutidas e apresentadas as estratégias empregadas para alcançar os resultados desejadas.

5.1 Tratamento Inicial do *Dataframe*

Tratamentos realizados para assegurar a qualidade dos fatos representados.

5.1.1 Tratamento dos Dados Faltantes

Ao realizar uma averiguação quanto aos dados faltantes, percebeu-se que colunas como: RNMNPF e RNAP possuíam dados faltando em mais de 90% dos casos. Dada a ausência dos dados e a irrelevância dessas colunas para geração do modelo preditivo, pois são ID, (*Identificador Único*), e ID não traz significância para modelagem, essas colunas foram descartadas. Também foi realizada a remoção das linhas que apresentavam algum nulo para não prejudicar o funcionamento dos algoritmos de *Machine Learning*. Em valores numéricos (como idade, número de ocorrências etc...) a substituição do valor ausente pela valor da média dos valores é muito comum e satisfatório, mais existem situações em que inviável a correção de todos os registros, como é caso dos registros deste conjunto de dados, sendo mais interessante a remoção. Apesar do impacto que a remoção de um registro pode causar, essa estratégia também é utilizada na impossibilidade de estimar de

forma precisa ou na impossibilidade de obter o valor do dado faltante junto a fonte. Desta forma, passou-se a considerar as colunas e as linhas sempre com valores.

5.1.2 Tratamento dos Dados Inconsistentes

As investigações mostraram outro problema no conjunto de dados. Observou-se que haviam registros como Viana Do Castelo que estava sendo diferenciado por letras maiúsculas e minúsculas (Viana **do** Castelo e Viana **Do** Castelo), o que gerava inconsistência ao apresentar estatísticas. Ao aplicar a padronização, Tabela 5.1, o número de ocorrências de incêndios por Distritos torna-se coerente pois Viana Do Castelo passa a ser representado de um única forma.

Tabela 5.1: Quantidade de incêndios por Distritos: registros padronizados

Índice	Distritos	Número de Incêndios
0	Porto	43200
1	Braga	19657
2	Aveiro	15250
3	Viseu	14968
4	Lisboa	14642
5	**Viana Do Castelo**	**13013**
6	Vila Real	12844
7	Santarém	9283
8	Setúbal	7827
9	Leiria	7123
10	Bragança	6758
11	Coimbra	6528
12	Guarda	5637
13	Castelo Branco	5101
14	Faro	4376
15	Beja	3722
16	Portalegre	3045
17	Évora	2731

Dessa forma, para solucionar problemas deste tipo na base de dados, a estratégia foi padronizar os dados em caixa alta para que não houvesse esta distinção para Distritos, Concelhos, Freguesias e Localidades, evitando assim estatísticas com resultados inconsistentes; a Tabela 5.1 apresenta o resultado da aplicação da correção para os Distritos.

5.1.3 Tratamento das Datas

Como a base já possuía uma coluna com ano, uma com o mês, uma com o dia e uma informando duração do incêndio, pode-se descartar as colunas de data: 'DataHora_PrimeiraIntervencao', 'DataHora_Extincao' para enxugar mais a base de dados, eliminando assim registros repetidos ou desnecessários.

5.1.4 Codificação dos Dados

Para dar prosseguimento a aplicação dos algoritmos de *Machine Learning*, as variáveis categóricas presentes na base de dados, foram codificadas em variáveis numéricas.

Após a codificação de categórico para numérico a classe passou a ser reproduzida pelos valores de 0 a 7, onde:

- O valor **0**: corresponde a ClasseArea]0 a 1 ha[.
- O valor **1**: corresponde a ClasseArea [1 a 10 ha].
- O valor **2**: corresponde a ClasseArea [10 a 20 ha].
- O valor **3**: corresponde a ClasseArea [20 a 50 ha].
- O valor **4**: corresponde a ClasseArea [50 a 100 ha].
- O valor **5**: corresponde a ClasseArea [100 a 500 ha].
- O valor **6**: corresponde a ClasseArea [500 a 1000 ha].
- O valor **7**: corresponde a ClasseArea [superior a 1000 ha].

5.1.5 Conferida na Duração dos Incêndios

Foi inspecionado a duração dos incêndios para ver se havia registros de duração negativos, mas não foram encontradas inconsistências.

5.1.6 Verificando a Distribuição dos Registros da Classe

O atributo alvo da base de dados é a ClasseArea, que corresponde à área ardida representada em faixas conforme a Tabela 5.2. Nesta tabela a quantidade incêndios por cada classe de área apresenta-se atualizada. Os registros nulos e incompletos que haviam na base de dados foram removidos durante o processo de tratamento da mesma. A quantidade incêndios por cada classe de área após o pré-processamento está representado a seguir.

Tabela 5.2: Quantidade de Incêndios por Faixas de Área

ClasseArea	Quantidade
]0 a 1 ha[	127348
[1 a 10 ha]	24319
[10 a 20 ha]	2375
[20 a 50 ha]	2089
[50 a 100 ha]	937
[100 a 500 ha]	1015
[500 a 1000 ha]	183
superior a 1000 ha	175

A quantidade de registros também foi representada graficamente na Figura 5.1.

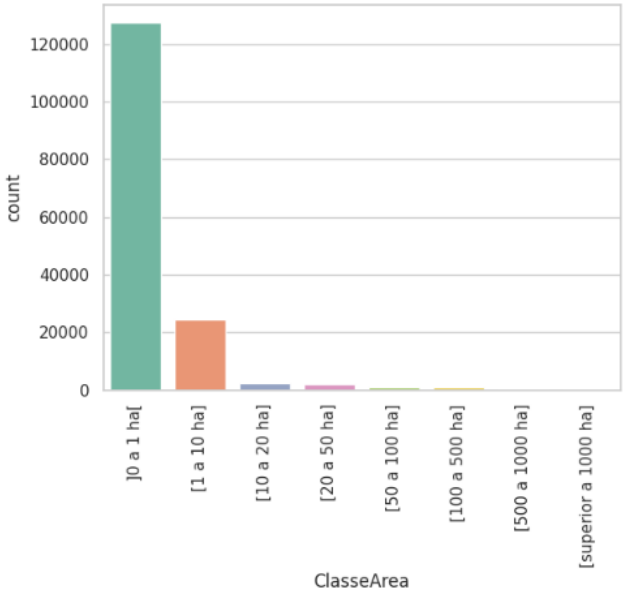


Figura 5.1: Quantidade de registros por classe de área categórico

5.2 Prever a Classe Área

Conforme apresentado na Tabela 5.2, o atributo ClasseArea corresponde a diferentes valores categóricos, de 0 a 1 ha, a superiores a 1000 ha, que são as possíveis faixas de área que um registro pode arder. Esse atributo representa a classe do conjunto de dados deste estudo cuja o objetivo é prever a qual faixa de área ardida um registro histórico da base de dados pertence.

Para conferir a distribuição dos registros por classe área, foi gerada a Figura 5.1, a partir da qual se verificou o desbalanceamento da classe, tendo como classe majoritária a ClasseArea ’]0 a 1 ha[’, que corresponde ao valor 0.

Após realizar a codificação, o conjunto de dados foi dividido em base de dados de treinamento, 75% dos dados e base de dados de teste, 25% dos dados. A base de dados de

treinamento foi submetida aos algoritmos de *Regressão Logística*, *Aprendizagem Bayesiana*, *Árvore de Decisão*, *Random Forest* e *Redes Neurais* para geração do modelo. Cada classificador gerado foi testado submetendo-lhe os registros da base de dados de teste para verificar o desempenho e a precisão entregues por cada modelo. Conforme mencionado, a variável alvo, a classe, mostrou-se totalmente desbalanceada com quase que a totalidade dos registros da base pertencentes à classe 0, o que inicialmente levou os classificadores a gerarem resultados inconsistentes e não satisfatórios.

5.2.1 Conjunto de Treinamento e Teste

De acordo com scikit-learn, aprender os parâmetros de uma função de predição e testá-la nos mesmos dados é um erro metodológico: um modelo que apenas repetisse os rótulos das amostras que acabou de ver teria uma pontuação perfeita, mas não conseguiria prever nada útil para dados diferentes. Esta situação é chamada de *overfitting* (scikit-learn, 2023a). Para que os resultados aqui apresentados pelos modelos de aprendizagem de máquina pudessem apresenta-se mais próximos do mudo real, a base de dados foi dividida em um conjunto para treinamento e outro para teste.

Para uma divisão aleatória em conjuntos de treinamento e teste, existe uma função auxiliar do *scikit-learn* - a função *train_test_split* - que, por padrão, divide 25% dos dados para teste e 75% restantes para o treinamento (scikit-learn, 2023b). Neste trabalho, a função *train_test_split* do *scikit-learn* foi utilizada para dividir a base de dados em base de dados de treino e base de dados de teste, mantendo-se as proporções padrões de divisão, conforme a Figura 5.2.

```
X_train.shape, y_train.shape
((118830, 23), (118830,))

X_test.shape, y_test.shape
((39611, 23), (39611,))
```

Figura 5.2: Base de dados de treinamento e teste

5.3 Aprendizagem de Máquina

5.3.1 Algoritmos de *Regressão Logística* e *Redes Neurais*

Os modelos obtidos pelos algoritmos de *Machine Learning Regressão Logística* e *Multilayer Perceptron classifier* de *Redes Neurais* com uma arquitetura com uma ou mais ca-

madas ocultas (que realizam cálculos intermediários e extraem características dos dados), cada uma contendo um número variável de neurônios, possui pelo menos uma camada de entrada (para receber os dados) e uma camada de saída (para produzir as previsões); não foram capazes de prever classes diferentes de 0. Devido ao desbalanceamento do conjunto de dados, ambos tenderam a classificar todos os registros da base de teste conforme o valor da classe majoritária, sendo tais modelos ineficazes para prever classes diferentes de 0.

5.3.2 *Aprendizagem Bayesiana*

O modelo gerado pelo algoritmo de *Aprendizagem Bayesiana* realizou previsões em todas as classes. Entretanto, a sua matriz de confusão 5.3 mostrou a presença de muitos erros de classificação, tendendo o algoritmo a classificar a maioria dos registros da base como classe 0. Novamente, o desbalanceamento da classe influenciou no resultado da classificação.

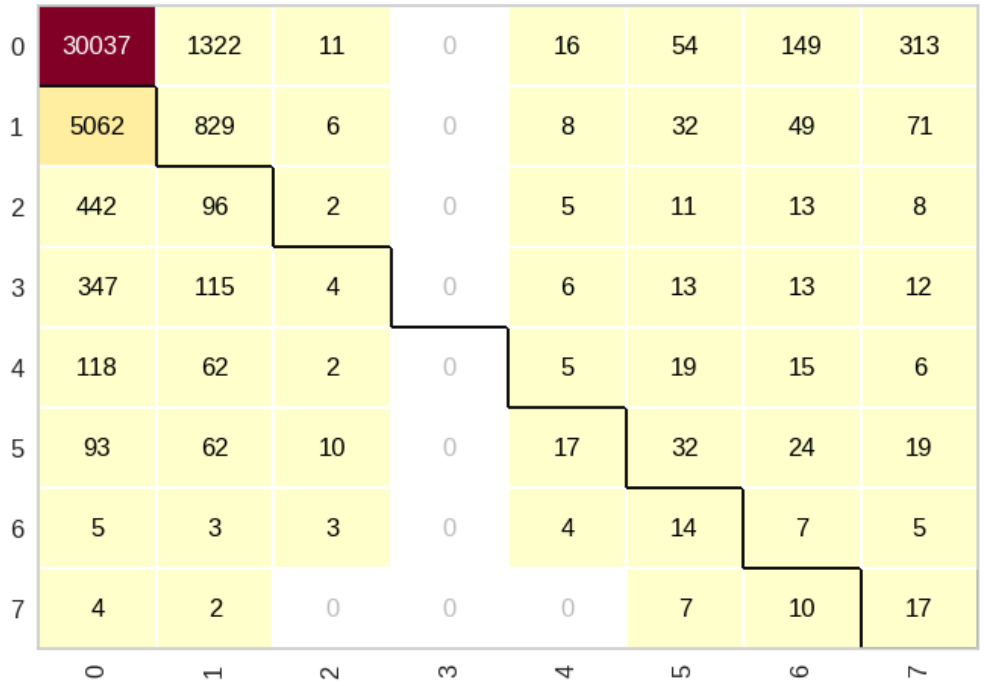


Figura 5.3: Matriz de Confusão: *Aprendizagem Bayesiana*

A tabela de métricas 5.3 do algoritmo apresentou o baixo desempenho do classificador para classes diferentes de 0, não sendo interessante para previsão um modelo que não classifique corretamente registros de outras classes.

Tabela 5.3: Resultados das Métricas de Classificação para *Aprendizagem Bayesiana*

Classe	Precisão	Revocação	F1-Score	Suporte
0	0.83	0.94	0.88	31902
1	0.33	0.14	0.19	6057
2	0.05	0.00	0.01	577
3	0.00	0.00	0.00	510
4	0.08	0.02	0.03	227
5	0.18	0.12	0.15	257
6	0.03	0.17	0.04	41
7	0.04	0.42	0.07	40
Acurácia			0.78	39611
Média Macro	0.19	0.23	0.17	39611
Média Ponderada	0.72	0.78	0.74	39611

5.3.3 *Árvore de Decisão*

O algoritmo de *Árvore de Decisão* obteve uma acurácia menor em relação aos algoritmos anteriores, mas isso não o impediu que obtivesse resultados melhores para acertar as classes diferentes de 0. A diagonal principal de sua matriz de confusão 5.4 apresentou uma quantidade maior de acertos para as classes diferentes de 0.

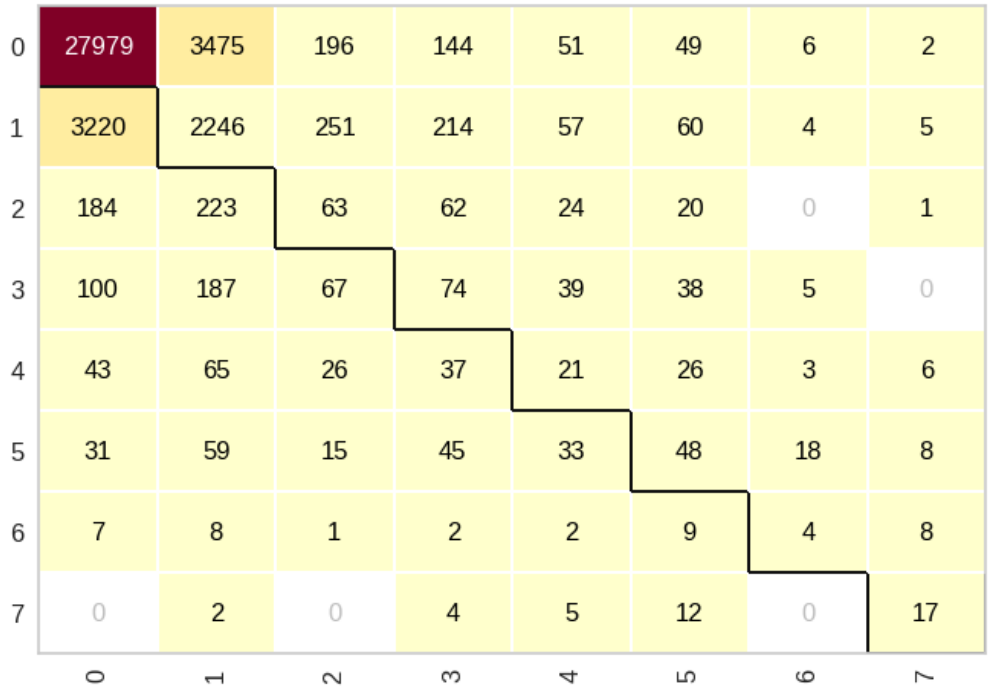


Figura 5.4: Matriz de Confusão: *Árvore de Decisão*

A tabela de métricas do algoritmo 5.4 também apresentou uma leve melhoria nos resultados para classes diferentes de 0, mostrando uma melhor adaptação do algoritmo para prever estas classes. Mesmo assim, os valores do precision e recall estavam baixos,

levando a uma média harmônica igualmente baixa.

Tabela 5.4: Resultados das Métricas de Classificação para *Árvore de Decisão*

Classe	Precisão	Revocação	F1-Score	Suporte
0	0.89	0.88	0.88	31902
1	0.36	0.37	0.36	6057
2	0.10	0.11	0.11	577
3	0.13	0.15	0.14	510
4	0.09	0.09	0.09	227
5	0.18	0.19	0.18	257
6	0.10	0.10	0.10	41
7	0.36	0.42	0.39	40
Acurácia			0.77	39611
Média Macro	0.28	0.29	0.28	39611
Média Ponderada	0.77	0.77	0.77	39611

5.3.4 *Random Forest*

Os testes realizados no classificador gerado pelo algoritmo *Random Forest* mostraram através de sua matriz de confusão 5.5, que o modelo prever registros para todas as classes; assim como o modelo de *Árvore de Decisão*, *Random Forest* obteve acertos ao classificar registros em todas as classes, mas, mesmo acertando e com uma acurácia melhor que a de *Árvore de Decisão*, ainda não foi possível obter métricas interessantes para um classificador confiável.

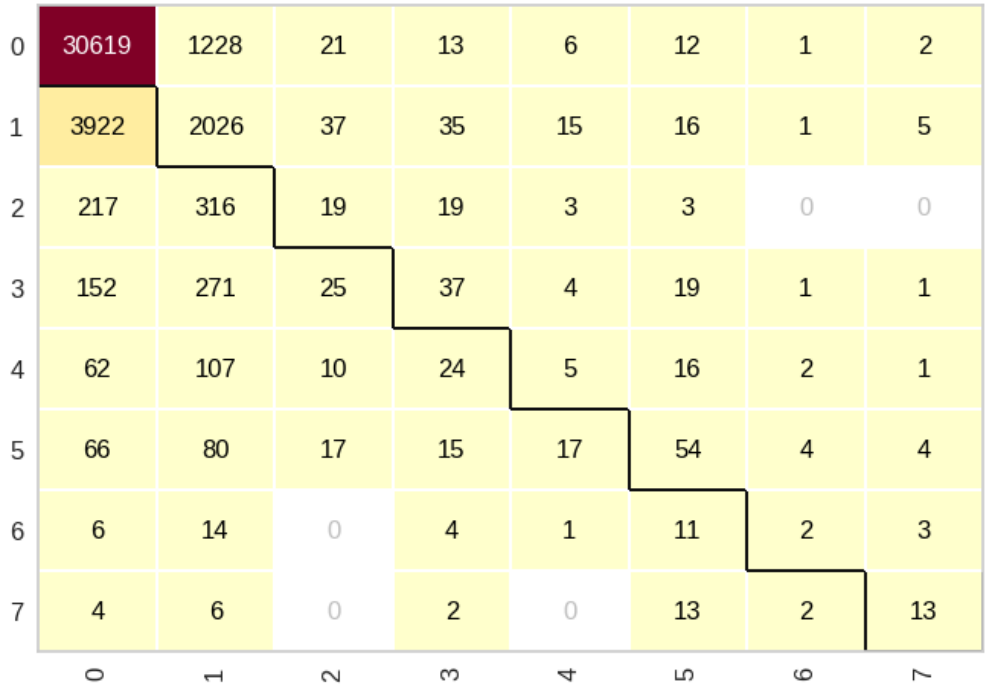


Figura 5.5: Matriz de Confusão: *Random Forest*

Conforme sua tabela de métricas 5.5, os resultados de seus acertos para classificar registros de classes diferentes de 0 mostraram-se muito baixos.

Tabela 5.5: Resultados das Métricas de Classificação para *Random Forest*

Classe	Precisão	Revocação	F1-Score	Suporte
0	0.87	0.96	0.91	31902
1	0.50	0.33	0.40	6057
2	0.15	0.03	0.05	577
3	0.25	0.07	0.11	510
4	0.10	0.02	0.04	227
5	0.38	0.21	0.27	257
6	0.15	0.05	0.07	41
7	0.45	0.33	0.38	40
Acurácia			0.83	39611
Média Macro	0.36	0.25	0.28	39611
Média Ponderada	0.79	0.83	0.80	39611

5.3.5 Conclusão do Tópico

Pode-se perceber o quanto o desbalanceamento da classe impactou negativamente nos resultados, influenciando e prejudicando os modelos gerados pelos algoritmos a acertarem previsões para as classes minoritárias durante as previsões com o conjunto de testes. Os classificadores classificavam todos os registros como classe 0 ou obtinham poucos acertos ao tentarem classificar registros que pertenciam a outras classes. Como esses resultados não refletiam a realidade esperada, as abordagens a seguir foram no intuito de contornar o problema e conseguir alcançar resultados mais apurados.

5.4 Abordagens para Tratar o Problema do Desbalanceamento de Dados

De acordo com Barbosa, as técnicas de amostragem realizam um papel importante na ajuda aos classificadores que aprendem com base de dados desbalanceadas, uma vez que essas técnicas retornam uma versão mais balanceada do conjunto de dados desbalanceados (Barbosa et al., 2019).

O *Under sampling* é uma técnica que reduz o número de exemplos da classe majoritária para equilibrar o número de exemplos de cada classe mais isso pode fazer com que exemplos importantes para generalização sejam excluídos. O *Over sampling* geralmente usa todos os exemplos disponíveis na classe minoritária para sintetizar novas instâncias, que podem incluir dados ruidosos ou outliers (Barbosa et al., 2019).

Chawla apresentou um método de *oversampling* conhecido como *Synthetic Minority Oversampling Technique (SMOTE)* - um método que adiciona novos exemplos minoritários artificiais, extrapolando entre instâncias minoritárias pré-existentes em vez de simplesmente duplicar exemplos originais (Chawla et al., 2002). A técnica primeiro encontra os k vizinhos mais próximos da classe minoritária para cada exemplo minoritário (Chawla et al., 2002) recomenda k = 5. Os exemplos artificiais são então gerados na direção de alguns ou todos os vizinhos mais próximos, dependendo da quantidade de *oversampling* desejada (Van Hulse et al., 2007).

Das técnicas de amostragem utilizadas na base de dados, a que realmente proporcionou melhorias ao desempenho dos algoritmos utilizados foi o método de *oversampling* conhecido como *SMOTE*. Por este motivo, apenas os resultados com *oversampling* serão apresentados.

Solucionado o problema do desbalanceamento dos dados com *SMOTE*, a base de dados passou de 158441 registros para 1018784 registros após a geração dos registros sintéticos. Figura 5.6, para uma divisão aleatória em conjuntos de treinamento e teste, a função auxiliar do scikit-learn - a função train_test_split - que, por padrão, divide 25% dos dados para teste e 75% restantes para o treinamento (scikit-learn, 2023b). A partir daí os algoritmos empregados anteriormente foram utilizados para geração de novos modelos a partir dos registros tratados e dessa forma testar e analisar os resultados obtidos.

```
[33] X_train_o.shape, y_train_o.shape

     ((764088, 23), (764088,))

[32] X_test_o.shape, y_test_o.shape

     ((254696, 23), (254696,))
```

Figura 5.6: Base de dados de treinamento e teste após smote

5.4.1 *Regressão Logística* para Dados Balanceados com *Smote*

Para o algoritmo de *Regressão Logística*, observou-se, a partir da sua matriz de confusão 5.7, que, após a amostragem da classe com o *SMOTE*, o modelo conseguiu classificar registros que pertenciam a todas as classes, o que não tinha sido possível quando a classe estava desbalanceada, a qual as previsões anteriores do conjunto de testes obtiveram como resultado classe 0.

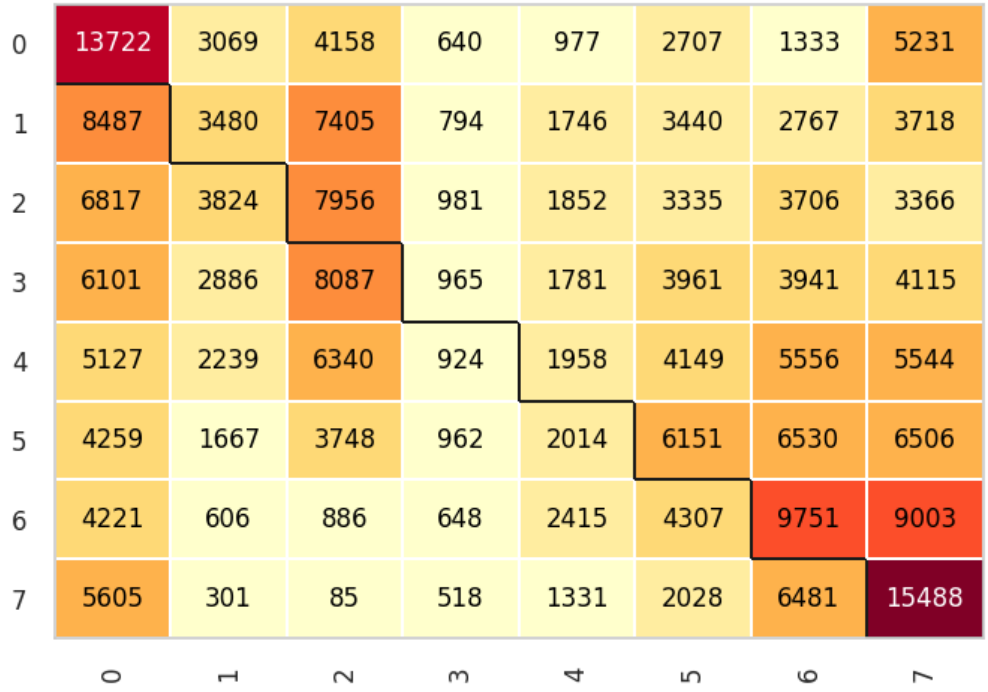


Figura 5.7: Matriz de Confusão: *Regressão Logística Oversampling*

Todavia, apesar de conseguir classificar registros de todas as classes, a tabela 5.6 mostrou que o algoritmo ainda possuía métricas muito baixas para classificação, o que, de certa forma, faz com que o modelo não seja considerado um bom classificador para o conjunto de dados.

Tabela 5.6: Resultados das Métricas de Classificação *Regressão Logística* com *Oversampling*

Classe	Precisão	Revocação	F1-Score	Suporte
0	0.25	0.43	0.32	31837
1	0.19	0.11	0.14	31837
2	0.21	0.25	0.23	31837
3	0.15	0.03	0.05	31837
4	0.14	0.06	0.09	31837
5	0.20	0.19	0.20	31837
6	0.24	0.31	0.27	31837
7	0.29	0.49	0.37	31837
Acurácia			0.23	254696
Média Macro	0.21	0.23	0.21	254696
Média Ponderada	0.21	0.23	0.21	254696

5.4.2 *Aprendizagem Bayesiana* para Dados Balanceados com *Smote*

Para *Aprendizagem Bayesiana*, a amostragem não foi interessante, pois, conforme a matriz de confusão 5.8 e a tabela de métricas 5.7, o F1-Score alcançado não chegou a 50%

para nenhuma das classe.

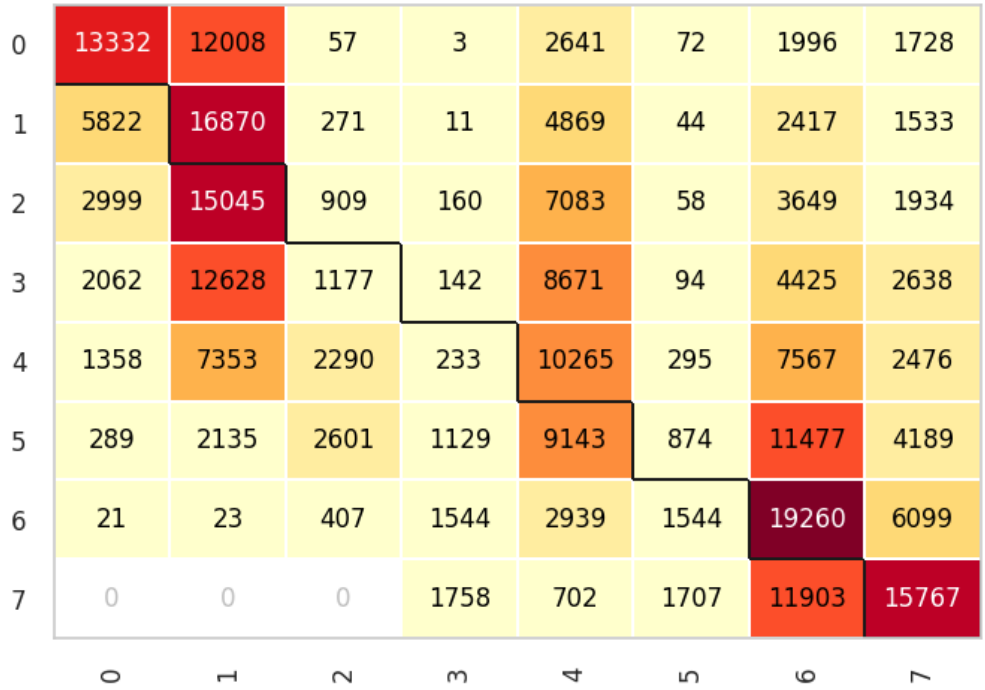


Figura 5.8: Matriz de Confusão: *Aprendizagem Bayesiana Oversampling*

A acurácia do modelo também foi inferior, visto que antes era 78% e agora diminuiu para 30% com *oversampling*. Os baixos resultados obtidos pelo modelo o tornam inapto para previsão.

Tabela 5.7: Resultados das Métricas de Classificação para *Aprendizagem Bayesiana* com *Oversampling*

Classe	Precisão	Revocação	F1-Score	Suporte
0	0.52	0.42	0.46	31837
1	0.26	0.53	0.34	31837
2	0.12	0.03	0.05	31837
3	0.03	0.00	0.01	31837
4	0.22	0.32	0.26	31837
5	0.19	0.03	0.05	31837
6	0.31	0.60	0.41	31837
7	0.43	0.50	0.46	31837
Acurácia			0.30	254696
Média Macro	0.26	0.30	0.26	254696
Média Ponderada	0.26	0.30	0.26	254696

5.4.3 *Árvore de Decisão* para Dados Balanceados com *Smote*

A sobreamostragem da base de dados com *smote* proporcionou melhores resultados ao algoritmo de *Árvore de Decisão*. Sua matriz de confusão 5.9 apresentou-se mais limpa,

observada pela menor quantidade de erros do modelo.

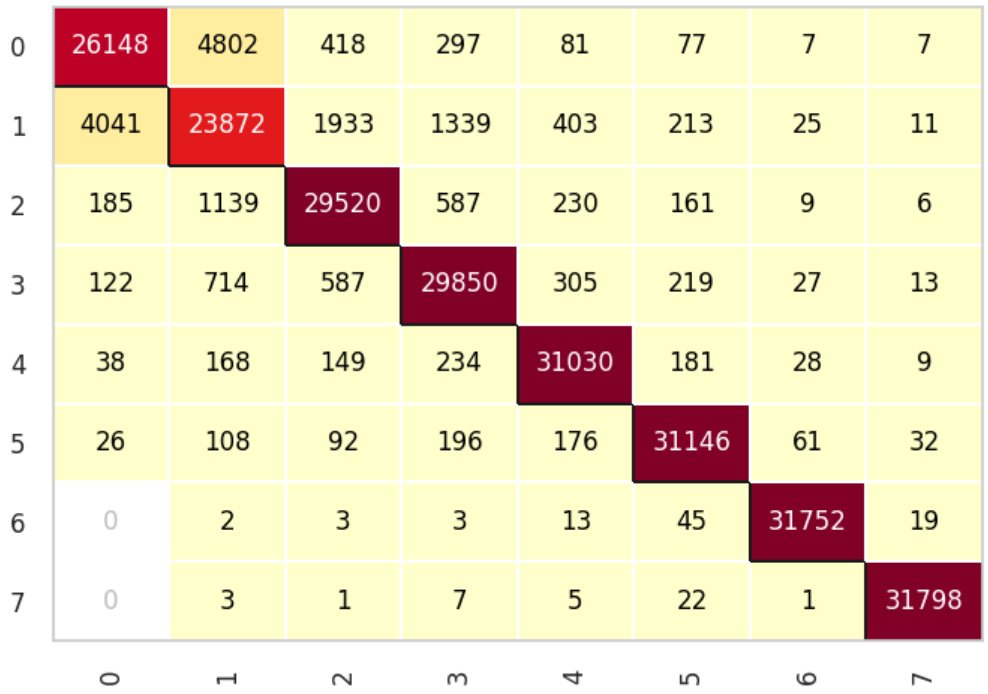


Figura 5.9: Matriz de Confusão: *Árvore de Decisão* Oversampling

A tabela de métricas do algoritmo 5.8 mostrou que o classificador obteve um bom desempenho, apresentando valores altos de acertos em quase todas as classes. Sua acurácia chegou a 92% e o valor mais baixo de seu F1-score foi de 76% para classe 1 e 84% para classe 0. Os demais valores superaram 90%, para algumas classes 100%, uma média harmônica interessante entre *precision* e *recall*, quando comparada com a de modelos anteriores.

Tabela 5.8: Resultados das Métricas de Classificação *Árvore de Decisão* com Oversampling

Classe	Precisão	Revocação	F1-Score	Suporte
0	0.86	0.82	0.84	31837
1	0.77	0.75	0.76	31837
2	0.90	0.93	0.91	31837
3	0.92	0.94	0.93	31837
4	0.96	0.97	0.97	31837
5	0.97	0.98	0.97	31837
6	1.00	1.00	1.00	31837
7	1.00	1.00	1.00	31837
Acurácia			0.92	254696
Média Macro	0.92	0.92	0.92	254696
Média Ponderada	0.92	0.92	0.92	254696

5.4.4 *Random Forest* para Dados Balanceados com *Smote*

O balanceamento da classe permitiu que modelo gerado pelo algoritmo *Random Forest* obtivesse melhorias significativas em seus resultados, tornando-o, até aqui, o melhor dos já testados. A matriz de confusão do algoritmo 5.10 foi a que apresentou menos erros de classificação.

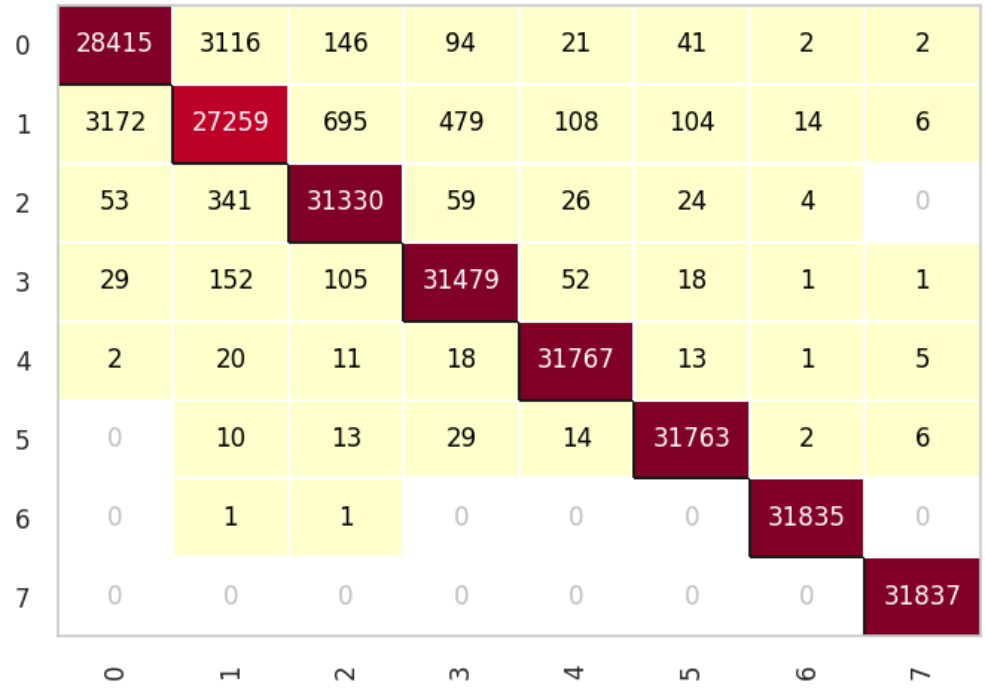


Figura 5.10: Matriz de Confusão: *Random Forest Oversampling*

O modelo gerado pelo algoritmo foi muito preciso ao realizar as previsões do conjunto de testes. Sua tabela de métricas 5.9 apresentou uma excelente acurácia chegando a 96% de acertos. Apenas duas classes apresentaram valores de F1-Score menores que 98%, as classes 1 e 0, com 87% e 89% respectivamente. As demais classes apresentaram valores entre 98% e 100% o que tornou o classificador excelente para prever os registros em todas as classes.

Tabela 5.9: Resultados das Métricas de Classificação para *Random Forest Oversampling*

Classe	Precisão	Revocação	F1-Score	Suporte
0	0.90	0.89	0.89	31837
1	0.88	0.86	0.87	31837
2	0.97	0.98	0.98	31837
3	0.98	0.99	0.98	31837
4	0.99	1.00	1.00	31837
5	0.99	1.00	1.00	31837
6	1.00	1.00	1.00	31837
7	1.00	1.00	1.00	31837
Acurácia			0.96	254696
Média Macro	0.96	0.96	0.96	254696
Média Ponderada	0.96	0.96	0.96	254696

5.5 Discussão

5.5.1 Tratamento do Desbalanceamento de Dados:

Um grande desafio encarado neste estudo foi o desbalanceamento dos dados. Conforme verificado, a maioria dos algoritmos de *Machine Learning* tiveram dificuldades para prever classes diferentes de 0 quando os dados estavam desbalanceados. A técnica de *oversampling*, especialmente o método *SMOTE*, foi utilizada para lidar com esse problema. Todavia, apesar do balanceamento dos dados, alguns algoritmos ainda assim apresentaram dificuldades para classificar corretamente as classes minoritárias.

5.5.2 Impacto das Técnicas de Amostragem:

A partir da análise dos resultados ficou evidente que a utilização da técnica *oversampling*, com *SMOTE*, teve impacto significativo no desempenho dos algoritmos. Mais precisamente, os modelos gerados após a aplicação do *SMOTE* apresentaram melhorias significativas em relação aos modelos obtidos antes do balanceamento dos dados. No entanto, ainda houve variação no desempenho entre os diferentes algoritmos após o balanceamento.

5.5.3 Eficácia dos Algoritmos:

Quanto à eficácia, os resultados obtidos revelaram que alguns algoritmos se sobressaíram em relação a outros na tarefa de classificação da dimensão de um incêndio, ou seja, prever a qual classe de área que um registro de incêndio florestal poderia pertencer. Por exemplo, a *Árvore de Decisão* e o *Random Forest* se destacaram como os algoritmos mais promissores, alcançando altas taxas de acurácia e F1-Score após o balanceamento dos dados. Em contrapartida, a *Regressão Logística* e as *Redes Neurais* apresentaram os desempenhos mais baixos, mesmo após o tratamento do desbalanceamento.

5.5.4 Limitações e Sugestões para Trabalhos Futuros:

É necessário reconhecer as limitações deste estudo. Por exemplo, a qualidade dos dados pode influenciar significativamente o desempenho dos modelos. Apesar de termos utilizado técnicas de balanceamento de dados e algoritmos de aprendizado de máquina adequados e eficientes, a inserção de atributos adicionais, como relevo, temperatura e dados culturais por exemplo, poderia refinar ainda mais a capacidade preditiva dos modelos. Ademais, a seleção de características pode desempenhar um papel crucial no desenvolvimento de modelos mais apurados. Apesar de termos nos concentrado em algumas técnicas de seleção de características, explorar outras abordagens poderia revelar *insights*

valiosos e melhorar a performance dos modelos. Dessa forma, como sugestão para trabalhos futuros, a adição de novas características e a exploração de diferentes abordagens de modelagem e técnicas de *Machine Learning*. Ao considerar esses pontos de vista e explorar continuamente novas abordagens, é possível avançar no desenvolvimento de sistemas de prevenção e combate a incêndios mais eficientes. Esta discussão fornece uma visão abrangente dos resultados obtidos neste estudo e destaca áreas para futuras pesquisas no contexto de modelagem preditiva de incêndios florestais.

Capítulo 6

Conclusão

Neste trabalho intitulado "Uma proposta de *Data Mining* para análise de dados referentes aos incêndios florestais ocorridos em Portugal", procurou-se compreender e abordar de maneira sistemática e abrangente o complexo desafio dos incêndios florestais em território português. Dois pontos fundamentais direcionaram essa investigação: a análise exploratória dos dados e a modelagem preditiva.

A análise exploratória dos dados possibilitou uma visão minuciosa, mais abrangente e holística desse fenômeno recorrente e impactante. A partir de uma variedade de gráficos e tabelas, foi possível mapear as áreas afetadas em diferentes níveis geográficos, identificando distritos, concelhos e localidades mais suscetíveis. Essa abordagem apresentou uma imagem complexa, porém indispensável, dos incêndios em Portugal.

As estatísticas resultantes da análise de dados apresentaram *insights* significativos. Em particular, destaca-se a ação do homem como um fator significante na ampliação dos incêndios florestais. Enquanto os incêndios naturais tendem a apresentar menores escalas, aqueles desencadeados pela intervenção humana, seja por negligência, causas desconhecidas ou ações intencionais, dominam as estatísticas, responsáveis pela grande maioria dos incêndios computados. Entretanto, é essencial não negligenciar o papel dos fatores meteorológicos nas quantidades de ocorrências. Durante os períodos críticos do ano, de julho a outubro, observou-se que a quantidade de incêndios aumentava, aumentando também o número de ocorrências que pertenciam a escalas mais elevadas de índices de severidade meteorológica e do risco de incêndio, evidenciando a associação direta entre condições meteorológicas adversas e o aumento das ocorrências. Considerando a gravidade e o impacto dos incêndios para a sociedade e meio ambiente, é incontestável e urgente a adoção de estratégias mais eficazes em sua gestão e prevenção.

Todavia, uma das lições mais importantes obtidas a partir deste estudo foi a necessidade de abordagens flexíveis e adaptáveis na pesquisa científica. O desbalanceamento das classes apresentou-se como um empecilho expressivo nos empenhos iniciais de aprendizado de máquina. No entanto, a aplicação do método *Synthetic Minority Oversampling Technique (SMOTE)* provou ser o caminho para superar esse desafio, permitindo que nos-

sos modelos alcançassem resultados de alta qualidade.

Na avaliação dos algoritmos de Aprendizado de Máquina, o *Random Forest*, após o balanceamento das classes com *SMOTE*, despontou como uma abordagem promissora. Com métricas de desempenho significativamente altas, incluindo uma acurácia de 96% e valores de F1-score consistentemente acima de 87%, essa abordagem destacou-se por sua capacidade de lidar eficazmente com os dados desbalanceados e produzir previsões precisas.

Todavia, é importante ressaltar que este estudo não elimina as possibilidades de pesquisa no campo da análise e modelagem de incêndios. A inclusão de fontes adicionais de dados, como dados culturais, climáticos e geoespaciais, oferecem uma oportunidade promitente para aprimorar ainda mais a precisão dos modelos e a percepção das causas subjacentes dos incêndios.

À medida que a pesquisa avança, deseja-se que os resultados aqui apresentados estimulem novas investigações e abordagens na área da prevenção e diminuição de incêndios, não apenas em Portugal, mas de uma forma global. Combinando análise exploratória de dados sofisticada e técnicas avançadas de *Machine Learning*, pode-se continuar a melhorar nossa capacidade de aprender com os dados e assim antecipar e responder a esse desafio crítico, garantindo a segurança das comunidades e a preservação do meio ambiente. Vale ressaltar que esta é uma caminhada que apenas iniciou, ficando assim aberta a futuras contribuições e sugestões para aprimorar o combate aos danos às nossas florestas e comunidades.

Referências

Pieter Adriaans and Dolf Zantinge. *Data Mining*. Addison-Wesley, 1996. ISBN 978-0201403800. 12

Charu C. Aggarwal. *Data Mining: The Textbook*. Springer, New York, NY, 2015. ISBN 978-3-319-14142-8. 23, 24

E. Alpaydin. *Introduction to Machine Learning*. MIT Press, 3 edition, 2016. ISBN 9780262028189. 11

Gian Barbosa, Péricles Miranda, Rafael Mello, and Ricardo Silva. Sequenciamento de algoritmos de amostragem para aumentar o desempenho de classificadores em conjuntos de dados desequilibrados. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 413–423, Porto Alegre, RS, Brasil, 2019. SBC. doi: 10.5753/eniac.2019.9302. URL `https://sol.sbc.org.br/index.php/eniac/article/view/9302`. 64

Pablo Pozzobon de Bem. Previsão de vulnerabilidade a incêndios florestais utilizando regressão logística e redes neurais artificiais : um estudo de caso no distrito federal brasileiro. Master's thesis, Universidade de Brasília, 2017. URL `https://repositorio.unb.br/handle/10482/31544`. 1, 5, 14

Jason Brownlee. *Statistical Methods for Machine Learning: Discover how to Transform Data into Knowledge with Python*. Machine Learning Mastery, 2019. 23

Paulo Alexandre Cabral Freire de Carvalho. Modelação do risco de incêndio florestal com redes neuronais artificiais: aplicação ao parque natural de montesinho. Master's thesis, Universidade Nova de Lisboa, 2006. URL `http://hdl.handle.net/10362/3632`. 5, 9

Leandro Nunes de Castro and Daniel Gomes Ferrari. *Introdução à Mineração de Dados Conceitos Básicos, Algoritmos e Aplicações*. Saraiva, 2016. ISBN 978-85-472-0098-5. ix, 10, 11, 12, 13, 14, 23, 51, 53

Francisco Castro Rego, Paulo Fernandes, João Sande Silva, João Azevedo, José Manuel Moura, Eurico Oliveira, Rui Cortes, Domingos Xavier Viegas, Duarte Caldeira, and Filipe Duarte Santos. Redução do risco de incêndio através da utilização de biomassa lenhosa para energia (2020). Technical report, Observatório Técnico Independente, Lisboa, 2020a. URL https://www.esquerda.net/sites/default/files/otiabril2020.pdf. Acesso em: 02 de maio de 2023. 6

Francisco Castro Rego, Paulo Fernandes, João Sande Silva, João Azevedo, José Manuel Moura, Eurico Oliveira, Rui Cortes, Domingos Xavier Viegas, Duarte Caldeira, and Filipe Duarte Santos. Análise de indicadores de desempenho do sistema de defesa da floresta contra incêndios na transição (2018-2020) para o sistema de gestão integrada de fogos rurais. Technical report, Observatório Técnico Independente, Lisboa, 2020b. URL https://www.parlamento.pt/Parlamento/Documents/oti/Estudotecnico-dez2020.pdf. Acesso em: 02 de maio de 2023. 9

N. Chawla, K. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *ArXiv*, abs/1106.1813, 2002. URL https://api.semanticscholar.org/CorpusID:1554582. 65

P. Cortez and A. Morais. A data mining approach to predict forest fires using meteorological data. In M. F. Santos J. Neves and J. Machado, editors, *New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence*, pages 512–523, Guimarães, Portugal, December 2007. APPIA. ISBN 978-989-95618-0-9. 9, 15, 26

Instituto da Conservação da Natureza e das Florestas (ICNF). Gfr | estatísticas, 2017 - 2023. URL https://www.icnf.pt/florestas/gfr/gfrgestaoinformacao/estatisticas. 20 de junho de 2023. 34

Instituto da Conservação da Natureza e das Florestas (ICNF). Relatórios provisórios de incêndios rurais: Áreas ardidas e ocorrências, 2017-2023. URL https://www.icnf.pt/florestas/gfr/gfrgestaoinformacao/grfrelatorios/areasardidaseocorrencias. Acesso em: 20 de junho de 2023. 15, 20

Matthieu de Gennaro, Yann Billaud, Yannick Pizzo, Savitri Garivait, Jean-Claude Loraud, Mahmoud El Hajj, and Bernard Porterie. Real-time wildland fire spread modeling using tabulated flame properties. *Fire Safety Journal*, 91:872–881, 2017. ISSN 0379-7112. doi: https://doi.org/10.1016/j.firesaf.2017.03.006. URL https://www.sciencedirect.com/science/article/pii/S0379711217302187. Fire Safety Science: Proceedings of the 12th International Symposium. 1

Instituto de Meteorologia. Apoio meteorológico na prevenção e combate aos incêndios florestais - relatório anual 2008. Technical report, Instituto de Meteorologia, 2008.

URL https://www.ipma.pt/resources.www/docs/im.publicacoes/edicoes.online/20090310/IFjkcXWwGQeGojfivCUc/met_20080501_20081015_fog_ex_co_pt.pdf. Acesso em: 02 de maio de 2023. 9

D. De Rigo, G. Liberta', T. Durrant, T. Artes Vivancos, and J. San-Miguel-Ayanz. Forest fire danger extremes in europe under climate change: variability and uncertainty. Technical Report EUR 28926 EN, Publications Office of the European Union, Luxembourg, 2017. URL https://publications.jrc.ec.europa.eu/repository/handle/JRC108974. 1, 6

Joana Farinha, Lúcio Cunha, and Luca Antonio Dimuccio. Exploratory spatial analysis of social vulnerability and forest fire risk in the pinhal interior sul (central portugal). *Sustainability*, 14(5), 2022. ISSN 2071-1050. doi: 10.3390/su14053010. URL https://www.mdpi.com/2071-1050/14/5/3010. 1

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37, Mar. 1996. doi: 10.1609/aimag.v17i3.1230. URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230. 13

Mark A. Finney. The challenge of quantitative risk analysis for wildland fire. *Forest Ecology and Management*, 211(1):97–108, 2005. ISSN 0378-1127. doi: https://doi.org/10.1016/j.foreco.2005.02.010. URL https://www.sciencedirect.com/science/article/pii/S0378112705000563. Relative Risk Assessments for Decision –Making Related To Uncharacteristic Wildfire. 6

Mark A Finney, Jack D Cohen, Sara S McAllister, and W Matt Jolly. On the need for a theory of wildland fire spread. *International Journal of Wildland Fire*, 22:25–36, 2013. ISSN 1049-8001. doi: 10.1071/WF11117. URL https://doi.org/10.1071/WF11117. 1

Flávia Silva de Freitas. Comparação de metodologias para o mapeamento do risco de incêndio florestal no município de braga, portugal. Master's thesis, Universidade Federal de Viçosa, 2021. URL https://locus.ufv.br//handle/123456789/29692. 1, 6

R. Goldschmidt, E. Passos, and E. Bezerra. *Data Mining Conceitos, técnicas, algoritmos, orientações e aplicações*. Elsevier, 2015. ISBN 978-85-352-7822-4. 11, 12, 13, 14, 16, 23, 24, 25

IBM. Crisp-dm: Cross-industry standard process for data mining. https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview, 2021. Acesso em 05 de junho de 2024. 19

ICNF. 8.º relatório provisório de incêndios rurais – 2022 – informação estatística sobre incêndios rurais 1 de janeiro a 15 de outubro de 2022. Relatório técnico, Divisão de Gestão do Programa de Fogos Rurais, 2022. URL https://www.icnf.pt/api/file/doc/4e8a66514175d0f7. Acesso em: 20 de junho de 2023. 9, 16, 18, 22, 32, 47

Instituto da Conservação da Natureza e das Florestas. Sistema de gestão integrada de fogos rurais (sgif), 2010. URL https://fogos.icnf.pt/sgif2010/InformacaoPublicalist.asp. Acesso em: 22 de junho de 2023. 20

IPMA. Índice meteorológico de incêndio [fwi] e probabilidade de extremos – mapa dinâmico. https://www.ipma.pt/pt/riscoincendio/fwi/, 2023a. Acesso em 16 de abril de 2023. 7, 16, 17

IPMA. Perigo de incêndio rural - Índice meteorológico de incêndio [fwi]. https://www.ipma.pt/pt/enciclopedia/otempo/risco.incendio/index.jsp?page=pirfwi.xml, 2023b. Acessado em 16 de abril de 2023. ix, 7, 8, 9, 18, 22

Wes McKinney. *Python para Análise de Dados: Tratamento de dados com Pandas, NumPy e IPython*. Editora Novatec, São Paulo, SP, Brasil, 2018. ISBN 9788575227510. 22

Tom M Mitchell. *Machine learning*. McGraw-Hill, 1997. ISBN 0070428077. 11

Peters Morgan. *Data Analysis from Scratch with Python: Step By Step Guide*. AI Sciences LLC, 2016. ISBN 978-1721942817. Edited by Davies Company, Ebook Converted and Cover by Pixels Studio. 25, 53

Natural Resources Canada. Canadian forest fire weather index (fwi) system, 2023. URL https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi. Accessed: 2023-04-16. 7, 8, 9

Diogo Pereira Fiadeiro Mesquita Nunes. Estimativa das emissões de incêndios florestais na europa. Master's thesis, Universidade de Aveiro, 2009. URL http://hdl.handle.net/10773/656. 1

Sandra Oliveira, Friderike Oehler, Jesús San-Miguel-Ayanz, Andrea Camia, and José M.C. Pereira. Modeling spatial patterns of fire occurrence in mediterranean europe using multiple regression and random forest. *Forest Ecology and Management*, 275:117–129, 2012. ISSN 0378-1127. doi: https://doi.org/10.1016/j.foreco.2012.03.003. URL https://www.sciencedirect.com/science/article/pii/S0378112712001272. 14

Admilson da Penha Pacheco, Juarez Antonio da Silva Junior, Antonio Miguel Ruiz-Armenteros, and Renato Filipe Faria Henriques. Assessment of k-nearest neighbor and

random forest classifiers for mapping forest fire areas in central portugal using landsat-8, sentinel-2, and terra imagery. *Remote Sensing*, 13(7), 2021. ISSN 2072-4292. doi: 10.3390/rs13071345. URL https://www.mdpi.com/2072-4292/13/7/1345. 10

Cristian Padurariu and Mihaela Elena Breaban. Dealing with data imbalance in text classification. *Procedia Computer Science*, 159:736–745, 2019. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2019.09.229. URL https://www.sciencedirect.com/science/article/pii/S1877050919314152. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 23rd International Conference KES2019. 22

Juli G. Pausas and Jon E. Keeley. Wildfires as an ecosystem service. *Frontiers in Ecology and the Environment*, 2019. URL https://api.semanticscholar.org/CorpusID:164879541. 5

Mateus Pellegrino. Crisp-dm: metodologia ideal para ciência de dados. *Blog Escola DNC*, 2020. URL https://www.escoladnc.com.br/blog/metodologia-crisp-dm/. 19

Gonçalo Filipe Lucas Menino Rodrigues Pinto. Sistema inteligente de previsão e prevenção de fogos florestais. Master's thesis, UNIVERSIDADE D COIMBRA, 2020. URL http://hdl.handle.net/10316/93937. 6, 7

Zohre Sadat Pourtaghi, Hamid Reza Pourghasemi, Roberta Aretano, and Teodoro Semeraro. Investigation of general indicators influencing on forest fire and its susceptibility modeling using different data mining techniques. *Ecological Indicators*, 64:72–84, 2016. ISSN 1470-160X. doi: https://doi.org/10.1016/j.ecolind.2015.12.030. URL https://www.sciencedirect.com/science/article/pii/S1470160X15007359. 5

United Nations Environment Programme. Spreading like wildfire – the rising threat of extraordinary landscape fires. Technical report, United Nations Environment Programme - UNEP, Nairobi, 2022. URL https://www.unep.org/resources/report/spreading-wildfire-rising-threat-extraordinary-landscape-fires. 1

ProjectPro. Why data preparation is an important part of data science? https://www.projectpro.io/article/why-data-preparation-is-an-important-part-of-data-science/242, 2023. Acessado em 20/06/2023. 22

Dorian Pyle. *Data Preparation for Data Mining*. Morgan Kaufmann, San Francisco, CA, 1999. ISBN 1558605290. 22

Alexandre M. Ramos, Ana Russo, Carlos C. DaCamara, Silvia Nunes, Pedro Sousa, P.M.M. Soares, Miguel M. Lima, Alexandra Hurduc, and Ricardo M. Trigo. The compound event that triggered the destructive fires of october 2017 in portugal. *iScience*, 26(3):106141, 2023. ISSN 2589-0042. doi: https://doi.org/10.1016/j.isci.2023.106141. URL https://www.sciencedirect.com/science/article/pii/S2589004223002183. 1, 6, 7, 47

Stuart Jonathan Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson India Education Services Private Limited, Harlow, England, 4 edition, 2022. ISBN 9356063575. 10

Youssef Safi and Abdelaziz Bouroumi. Prediction of forest fires using artificial neural networks. *Applied mathematical sciences*, 7:271–286, 2013. URL https://api.semanticscholar.org/CorpusID:53321936. 15

scikit-learn. Cross-validation: evaluating estimator performance. https://scikit-learn.org/stable/modules/cross_validation.html, 2023a. Accessedo em 18 julho de 2023. 60

scikit-learn. sklearn model selection train test split. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html, 2023b. Accessedo em 18 julho de 2023. 60, 65

Leandro Augusto da Silva, Sarajane Marques Peres, and Clodis Boscarioli. *Introdução à mineração de dados: com aplicações em R*. Elsevier, 2016. ISBN 9788535284461. 13, 14

Matplotlib Development Team. Matplotlib: Visualization with python. https://matplotlib.org/, 2012 – 2023. Acessado em 16 de junho de 2023. 25

Pandas Development Team. Pandas documentation. https://pandas.pydata.org/docs/index.html, 2023. Acessado em 16 de junho de 2023. 25

Seaborn Development Team. seaborn: statistical data visualization. https://seaborn.pydata.org/, 2012-2023. Acessado em 16 de junho de 2023. 25

A.C. Teodoro and L. Duarte. Forest fire risk maps: a gis open source application – a case study in norwest of portugal. *International Journal of Geographical Information Science*, 27(4):699–720, 2013. doi: 10.1080/13658816.2012.721554. URL https://doi.org/10.1080/13658816.2012.721554. 1, 5

Jason Van Hulse, Taghi M. Khoshgoftaar, and Amri Napolitano. Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th International Con-*

ference on Machine Learning, pages 935–942. ACM, 2007. doi: 10.1145/1273496.1273614. URL https://doi.org/10.1145/1273496.1273614. 65

David A. Wood. Prediction and data mining of burned areas of forest fires: Optimized data matching and mining algorithm provides valuable insight. *Artificial Intelligence in Agriculture*, 5:24–42, 2021. ISSN 2589-7217. doi: https://doi.org/10.1016/j.aiia.2021.01.004. URL https://www.sciencedirect.com/science/article/pii/S2589721721000118. 1