Sistema Inteligente De Previsão E Prevenção De Fogos Florestais Gonçalo Filipe Lucas Menino Rodrigues Pinto

Gonçalo Filipe Lucas Menino Rodrigues Pinto

Sistema Inteligente De Previsão E Prevenção De Fogos Florestais

VOLUME 1

Relatório de Estágio no âmbito do Mestrado em Engenharia Informática, especialização em Sistemas Inteligentes orientada pelo Prof. Dr. António Dourado e Eng. Rafael Maia, e apresentada à Faculdade de Ciências e Tecnologias / Departamento de Engenharia Informática.

Outubro de 2020

Agradecimentos

Quero deixar o meu agradecimento em primeiro lugar aos meus pais por me terem proporcionado as devidas condições para que pudesse seguir os meus sonhos e efectuar todo o percurso académico na área que tanto me fascina.

Agradecer também de uma forma especial à Grama e a todos os meus colegas de trabalho, que me ajudaram a crescer não só a nível profissional como a nível pessoal, bem como me ajudaram dia após dia para que o presente projecto ganhasse vida. Sem eles nada disto seria possível.

Agradecer ainda ao Professor António Dourado pelo acompanhamento prestado ao longo do período de estágio, e por ter partilhado comigo o seu conhecimento ao nível daquela que é a Área de Aprendizagem Computacional.

Por último mas não menos importante quero deixar o meu agradecimento a todo o resto da minha família e a todos os meus amigos que viveram esta experiência ao meu lado.

Abstract

Forest fires are the biggest natural disasters that have troubled Portugal in the last few years. Forest fires result in great economic and environmental damages but they also put the life of many life forms at risk, humans included. Being able to predict a forest fire helps reduce or even prevent the damage caused.

This paper documents the development of a prototype of a web application whose main purpose is help predict and prevent forest fires using machine learning.

This application uses data pertaining to the number of habitants in each Portuguese municipality on top the weather data, terrain data and data collected from previous forest fires which is already used in other similar wild fires predicting applications.

Several machine learning methods were used to treat the different data sets to create a single data set to use in the prediction models.

To create a model capable of predicting forest fires two machine learning algorithms were considered, Support Vector Machine and Random Forest.

The different machine learning methods and machine learning algorithms that were given due consideration are described and analyzed at another point in this document

In the end, the best model was obtained by Random Forest with 61.29% following the Area Under the Receiver Operating Characteristics Curve (AUC) metric.

The prototype developed during my internship is totally functional when it comes to predicting forest fires and all its functionalities are described at another point in this document.

Keywords

Forest fires, wildfires, web application, prediction, prevention, Machine Learning, Random Forest, Support Vector Machine

Resumo

Os incêndios florestais são uma das catástrofes naturais que mais têm afectado Portugal nos últimos anos. A ocorrência de incêndios florestais provoca grandes danos económicos, ambientais e coloca a vida de diversos seres vivos como o ser humano em risco. A capacidade de prever a ocorrência de um incêndio florestal é um factor determinante para amenizar ou até mesmo evitar os danos causados pelos mesmos.

O foco deste relatório incide no desenvolvimento de um protótipo de uma aplicação web para ajudar na previsão e prevenção de incêndios florestais.

Para além dos dados habitualmente considerados em sistemas de previsão e prevenção de incêndios florestais já existentes, como dados meteorológicos, das condições do solo terrestre e de ocorrências de incêndios florestais, o presente protótipo considera também dados relativos ao número de habitantes nos diferentes concelhos de Portugal.

Foram aplicadas diferentes técnicas de Aprendizagem Computacional para efectuar o tratamento dos diferentes conjuntos de dados considerados e proceder à criação de um único conjunto de dados a ser utilizado por parte dos modelos na previsão de ocorrências de incêndios florestais.

Para efectuar a previsão da ocorrência de incêndios foram considerados dois algoritmos, respectivamente o algoritmo Support Vector Machine e o algoritmo Random Forest.

As diferentes técnicas de Aprendizagem Computacional consideradas bem como os diferentes modelos construídos encontram-se descritos e analisados adiante no presente relatório.

O melhor modelo obtido foi um modelo do algoritmo Random Forest que apresentava um desempenho de 61.29% relativamente à métrica Area Under the Receiver Operating Characteristics Curve (AUC).

O protótipo desenvolvido durante o período de estágio encontra-se totalmente funcional no que diz respeito à previsão da ocorrência de incêndios florestais e as suas diferentes funcionalidades encontram-se descritas no documento em questão.

Palavras-Chave

Fogos florestais, incêndios florestais, aplicação web, previsão, prevenção, Aprendizagem Computacional, Random Forest, Support Vector Machine

Conteúdo

1 Introdução **1**
1.1 Ambiente de trabalho . . . 1
1.1.1 Contexto . . . 1
1.1.2 Motivação . . . 2
1.1.3 Objectivos . . . 3
1.1.4 Estrutura do documento . . . 4

2 Estado da Arte **7**
2.1 Sistemas base de recolha de dados . . . 7
2.1.1 FWI . . . 7
2.1.2 Moderate Resolution Imaging Spectroradiometer (MODIS) . . . 9
2.1.3 Geographic information system (GIS) . . . 10
2.1.4 Dados meteorológicos . . . 10
2.2 Conceitos de Aprendizagem Computacional . . . 11
2.2.1 Aprendizagem supervisionada . . . 11
2.2.2 Aprendizagem não supervisionada . . . 12
2.2.3 Aprendizagem em conjunto . . . 12
2.2.4 Overffiting . . . 12
2.3 Algoritmos de Aprendizagem Computacional . . . 12
2.3.1 Rede Neuronal Artificial (RNA) . . . 13
2.3.2 Support Vector Machine (SVM) . . . 14
2.3.3 Decision Trees (DT) . . . 16
2.3.4 Random Forest . . . 16
2.3.5 Gradient Boosting . . . 17
2.3.6 Logistic Regression . . . 18
2.3.7 K-means clustering . . . 18
2.3.8 Bootstrap aggregating . . . 19
2.4 Revisão de literatura . . . 20
2.4.1 Learning to Predict Forest Fires with Different Data Mining Techniques 20
2.4.2 A Data Mining Approach to Predict Forest Fires using Meteorological Data . . . 21
2.4.3 Artificial Intelligence for Forest Fire Prediction . . . 21
2.4.4 Pattern Clustering of Forest Fires based on Meteorological Variables and its Classification using Hybrid Data Mining Methods . . . 22
2.4.5 Applying Decision Tree Algorithm and Neural Networks to Predict Forest Fires in Lebanon . . . 22
2.4.6 Predicting Size of Forest Fire using Hybrid Model . . . 23
2.4.7 Computational Intelligence and Data Mining Techniques using the Fire Data Set . . . 24
2.4.8 Predicting the Burned Area in Forest using Machine Learning Techniques . . . 25

2.4.9 Assessing the Suitability of Soft Computing Approaches for Forest Fires Prediction . . . 26
2.4.10 FireCast: Leveraging Deep Learning to Predict Wildfire Spread . . . 27
2.5 Plataformas Web Existentes . . . 28
2.5.1 IPMA . . . 28
2.5.2 Fogos.pt . . . 32
2.5.3 EFFIS . . . 34
2.5.4 GWIS . . . 36

3 Especificações **39**
3.1 Metodologia de desenvolvimento de software . . . 39
3.1.1 Cargos . . . 39
3.1.2 Artefactos . . . 40
3.1.3 Eventos . . . 40
3.1.4 Vantagens . . . 42
3.2 Requisitos não funcionais . . . 42
3.3 Requisitos funcionais . . . 43
3.3.1 Diagramas de Casos de Uso . . . 43
3.3.2 Casos de Uso . . . 51
3.4 Análise de Riscos . . . 66
3.4.1 R01 - Desconhecimento das novas tecnologias . . . 67
3.4.2 R02 - Obtenção do conjunto de dados de operadores móveis . . . 67
3.4.3 R03 - Qualidade do conjunto de dados meteorológicos . . . 67
3.5 Arquitectura do sistema . . . 67
3.5.1 Amazon Web Services . . . 68
3.6 Tecnologias e ferramentas utilizadas . . . 70
3.6.1 Angular . . . 70
3.6.2 Python . . . 70
3.6.3 scikit-learn . . . 70
3.6.4 Pandas . . . 71
3.6.5 JSON . . . 71
3.6.6 REST . . . 71
3.7 Plano de trabalho . . . 71
3.7.1 Primeiro Semestre . . . 72
3.7.2 Segundo Semestre . . . 72
3.8 Planeamento de tarefas . . . 73

4 Conjunto de dados **75**
4.1 Dados considerados . . . 75
4.1.1 Dados históricos de ocorrências de incêndios florestais . . . 75
4.1.2 Dados históricos de condições meteorológicas . . . 77
4.1.3 Dados populacionais . . . 78
4.2 Processamento de dados . . . 78
4.2.1 Dados históricos de ocorrências de incêndios florestais . . . 78
4.2.2 Dados históricos de condições meteorológicas . . . 80
4.2.3 Dados populacionais . . . 81
4.3 Criação do conjunto de dados . . . 81
4.3.1 Adição do conjunto de dados históricos de condições meteorológicas . . . 82
4.3.2 Criação do target do conjunto de dados de ocorrências de incêndio . . . 83
4.3.3 Conjunto de dados de não ocorrências de incêndio . . . 84
4.3.4 União dos conjuntos de dados de ocorrências e não ocorrências de incêndio . . . 86

4.3.5 Adição do conjunto de dados populacionais . . . 87
4.3.6 Criação de novas features . . . 87
4.3.7 Adição de novas features . . . 88
4.4 Conjunto de dados final . . . 92
4.5 Remoção de outliers . . . 92

5 Desenvolvimento **95**
5.1 Aprendizagem Computacional . . . 95
5.1.1 Algoritmos de Aprendizagem Computacional . . . 95
5.1.2 Divisão do conjunto de dados em conjuntos de treino e teste . . . 98
5.1.3 Cross-validation . . . 98
5.1.4 Ajuste de hiperparâmetros . . . 100
5.1.5 Undersampling . . . 101
5.1.6 Métrica utilizada . . . 102
5.1.7 Coeficiente da correlação de Pearson . . . 103
5.1.8 Condições meteorológicas mais influentes nas ocorrências de incêndio 104
5.2 Web Application . . . 106
5.2.1 Autenticação . . . 106
5.2.2 Carregamento de dados . . . 107
5.2.3 Integração do modelo de Aprendizagem Computacional na web application . . . 111
5.2.4 Consulta de ocorrências históricas de incêndios florestais . . . 115
5.2.5 Visualização dos dados carregados . . . 117

6 Testes e resultados obtidos **121**
6.1 Período de treino . . . 122
6.1.1 Resultados obtidos durante o período de treino . . . 122
6.1.2 Melhores combinações de hiperparâmetros . . . 123
6.2 Período de teste . . . 126
6.2.1 Support Vector Machine . . . 126
6.2.2 Random Forest . . . 129
6.3 Comparação face a outros modelos . . . 132
6.4 Análise dos resultados obtidos . . . 133

7 Conclusão **137**
7.1 Trabalho futuro . . . 140

Anexos **147**

A Resultados obtidos durante o período de treino **149**
A.1 Support Vector Machine . . . 149
A.2 Random Forest . . . 152

Acrónimos

AWS Amazon Web Services. 138, 139

CNN Convolutional Neural Network. 27

DT Decision Trees. ix, 16, 20–24

ETL Extract, Transform and Load. 138

FCM Fuzzy C-Means. 23, 24

FWI Canadian Forest Fire Weather Index System. ix, xv, 7, 8, 21, 31, 32

GIS Geographic Information System. xv, 10, 20, 27

IG Information Gain. 27

JSON JavaScript Object Notation. 70, 71

MAE Mean absolute error. 27

MODIS Moderate Resolution Imaging Spectroradiometer. 9

MSE Mean squared error. 27

NASA National Aeronautics and Space Administration. 9

RAE Relative absolute error. 27

REST Representational State Transfer. 70, 71

RMSE Root mean squared error. 24, 25, 27

RNA Rede Neuronal Artificial. xv, 13, 14, 25, 27

SVM Support Vector Machine. ix, 14, 15, 21, 22, 26, 27

URI Uniform Resource Identifier. 71

Lista de Figuras

2.1 Estrutura do sistema Canadian Forest Fire Weather Index System (FWI) [38] 8
2.2 Mapa com recurso ao sistema Geographic Information System (GIS) [25] . . 10
2.3 Técnicas de Aprendizagem Computacional [35] 12
2.4 Modelo base de uma Rede Neuronal Artificial (RNA) (Perceptron) [33] . . . 13
2.5 Modelo de uma RNA organizada em camadas (Multilayer Perceptron) [69] . 14
2.6 Construção de um hiperplano com base nos vetores de suporte [50] 15
2.7 Aplicação de uma função de kernel [37] 15
2.8 Decision Tree [11] 16
2.9 Random Forest [61] 17
2.10 Gradient Boosting [13] 17
2.11 Função Sigmoid [63] 18
2.12 K-means clustering [31] 19
2.13 Bootstrap aggregating (Bagging) [40] 20
2.14 Plataforma IPMA - Risco de Incêndio Rural 29
2.15 Plataforma IPMA - Condições meteorológicas previstas para o concelho de Coimbra 30
2.16 Plataforma IPMA - Risco de Incêndio Rural do Concelho de Coimbra . . . 31
2.17 Plataforma IPMA - FWI, índice de perigo de incêndio 32
2.18 Plataforma Fogos 33
2.19 Detalhes de um fogo activo 33
2.20 Estatísticas dia anterior 34
2.21 Dashboard situação atual plataforma EFFIS 35
2.22 Estatísticas EFFIS relativas ao número de incêndios por país 35
2.23 Dashboard notícias sobre incêndio plataforma EFFIS 36
2.24 Dashboard plataforma GWIS 36
2.25 Estatísticas GWIS relativas ao número de incêndios por país na Europa . . 37

3.1 Desenvolvimento de um produto com a metodologia Scrum [43] 42
3.2 Legenda dos componentes dos diagramas 43
3.3 Utilizadores da aplicação 44
3.4 Diagrama de alto nível do utilizador 45
3.5 Diagrama de funcionalidades do utilizador 46
3.6 Diagrama de funcionalidades do utilizador 2 47
3.7 Diagrama de alto nível do fornecedor de dados 48
3.8 Diagrama de funcionalidades do fornecedor de dados 49
3.9 Diagrama de funcionalidades do fornecedor de dados 2 50
3.10 Diagrama de funcionalidades do sistema inteligente 51
3.11 Arquitetura Amazon Web Services utilizada 69
3.12 Diagrama de Gantt relativo ao primeiro semestre 73
3.13 Diagrama de Gantt relativo ao segundo semestre 74

4.1 Z-Score Normal Distribution [70] 93

5.1 Time series com 5 iterações e 1000 amostras [36] 99
5.2 K-fold cross-validation [32] 100
5.3 Separação das classes e respectiva ROC Curve para uma AUC de valor 1 . . 103
5.4 Coeficiente da correlação de Pearson entre as variáveis X e Y [66] 103
5.5 Coeficiente da correlação de Pearson -1, 0 e 1 [66] 104
5.6 Ecrã de login 107
5.7 Diagrama de fluxo de carregamento de dados 108
5.8 Ecrã relativo ao carregamento de dados 110
5.9 Condições oferecidas por diferentes aplicações de previsão meteorológica [9] 112
5.10 Ecrã relativo à previsão da ocorrência de incêndios florestais 115
5.11 Ecrã relativo à consulta de ocorrências históricas de incêndios florestais . . . 116
5.12 Ecrã relativo à consulta dos detalhes de uma ocorrência histórica de incêndio 117
5.13 Ecrã relativo à visualização de dados 118
5.14 Ecrã relativo à edição de dados 118
5.15 Ecrã relativo à eliminação de dados 119

6.1 Matriz de confusão do modelo do algoritmo Support Vector Machine com os hiperparâmetros escolhidos 127
6.2 ROC curve do modelo do algoritmo Support Vector Machine com os hiperparâmetros escolhidos 128
6.3 Matriz de confusão do modelo do algoritmo Random Forest com os hiperparâmetros escolhidos 130
6.4 ROC curve do modelo do algoritmo Random Forest com os hiperparâmetros escolhidos 131

Lista de Tabelas

3.1 Actores da aplicação 44
3.2 CU01 - Autenticação 52
3.3 EX01 - Conta inválida 52
3.4 CU02 - Terminar sessão 52
3.5 CU03 - Repor palavra-passe 53
3.6 EX02 - Nova palavra-passe diferente da confirmação 53
3.7 CU04 - Alterar dados da conta 53
3.8 CU05 - Alterar palavra-passe 54
3.9 EX03 - Palavra-passe incorrecta 54
3.10 CU06 - Alterar nome 54
3.11 EX04 - Nome inválido 54
3.12 CU07 - Alterar Email 55
3.13 EX05 - Email inválido 55
3.14 CU08 - Alterar número de telemóvel 55
3.15 EX06 - Número de telemóvel inválido 55
3.16 CU09 - Visualizar dados históricos 56
3.17 CU010 - Visualizar dados históricos de incêndios 56
3.18 CU11 - Visualizar condições meteorológicas mais relevantes na ocorrência de um incêndio 56
3.19 CU12 - Visualizar propagação de um incêndio 57
3.20 CU13 - Visualizar dados históricos de condições meteorológicas 57
3.21 CU14 - Visualizar dados populacionais 57
3.22 CU15 - Visualizar a previsão da ocorrência de incêndios 58
3.23 CU16 - Carregar dados 58
3.24 CU17 - Carregar dados históricos de incêndios 59
3.25 CU18 - Carregar dados históricos de condições meteorológicas 59
3.26 CU19 - Carregar dados populacionais 59
3.27 CU20 - Consultar dados 60
3.28 CU21 - Consultar dados históricos de incêndios 60
3.29 CU22 - Consultar dados históricos de condições meteorológicas 60
3.30 CU23 - Consultar dados populacionais 61
3.31 CU24 - Editar dados 61
3.32 CU25 - Editar dados históricos de incêndios 61
3.33 CU26 - Editar dados históricos de condições meteorológicas 62
3.34 CU27 - Editar dados populacionais 62
3.35 CU28 - Eliminar dados 63
3.36 CU29 - Eliminar dados históricos de incêndios 63
3.37 CU30 - Eliminar dados históricos de condições meteorológicas 63
3.38 CU31 - Eliminar dados populacionais 64
3.39 CU32 - Informar resultado do carregamento de um ficheiro de dados 64
3.40 CU33 - Gerir conflitos de dados 64

3.41 CU34 - Remover ruído dos dados de entrada 65
3.42 CU35 - Normalizar os dados de entrada 65
3.43 CU36 - Inserir dados 65
3.44 CU37 - Prever a ocorrência de incêndios 66

5.1 Análise dos algoritmos utilizados por autores no estudo do Estado da Arte . 96
5.2 Hiperparâmetros do algoritmo Support Vector Machine considerados 97
5.3 Hiperparâmetros do algoritmo Random Forest considerados 98

6.1 Melhores resultados do algoritmo Support Vector Machine 123
6.2 Melhores resultados do algoritmo Random Forest 125
6.3 Comparação do desempenho obtido pelos modelos considerados face a outros modelos de classificação 133

A.1 Resultados obtidos durante o período de treino por parte dos modelos do algoritmo Support Vector Machine 151
A.2 Resultados obtidos durante o período de treino por parte dos modelos do algoritmo Random Forest 154

Capítulo 1

Introdução

O presente relatório tem o objectivo de descrever todo o trabalho desenvolvido na empresa Grama [27] durante o período de estágio.

Este capítulo encontra-se dividido em 5 sub-capítulos. O primeiro sub-capítulo, denominado de ambiente de trabalho expõe o ambiente de trabalho onde o presente estágio decorre. O segundo apresenta um enquadramento do projecto desenvolvido neste estágio. O terceiro incide nos factores que motivaram o seu desenvolvimento. O quarto, onde se encontram apresentados os objectivos a alcançar com o desenvolvimento do projecto, e por último, o quinto onde se encontra apresentada a estrutura que compõe o presente relatório.

1.1 Ambiente de trabalho

O estágio reportado no presente relatório decorre na empresa Grama [27].

A Grama [27] é uma empresa de desenvolvimento de *software* fundada em Janeiro de 2017 e sediada em Coimbra que desenvolve software à medida essencialmente para a área das telecomunicações.

O local de trabalho definido para o período de estágio são as instalações da empresa, mais precisamente no escritório sediado na Avenida Sá da Bandeira em Coimbra. Durante todo o período de estágio o aluno tem ao seu dispor os equipamentos necessários para o desempenho das suas tarefas bem como a ajuda de Engenheiros experientes para qualquer dificuldade vivida no desenvolvimento do projecto.

1.1.1 Contexto

Os incêndios florestais são nos dias de hoje uma das catástrofes naturais que provoca mais danos em Portugal. A ocorrência deste tipo de incêndios provoca a destruição de florestas, habitações e chega mesmo a colocar em risco a vida de uma grande quantidade de seres vivos, de entre os quais o ser humano.

Todos os anos em Portugal se verifica um elevado número de ocorrências de incêndios florestais que fazem arder hectares de floresta, destroem habitações e por vezes acabam mesmo por tirar a vida a pessoas. No ano de 2017 presenciou-se em Portugal o pior cenário desde 1980, verificaram-se mais de 21000 ocorrências de incêndios florestais que resultaram numa área ardida de 539.921 hectares [19].

A aprendizagem Computacional é uma área que tem vindo a evoluir cada vez mais nos dias de hoje e a permitir que o ser humano tire partido desta mesma evolução. Há alguns anos não se pensava que fosse possível prever a ocorrência de um ataque cardíaco num ser humano, existirem carros com diferentes níveis de autonomia ou prever a ocorrência de catástrofes naturais. Nos dias de hoje estas e outras vertentes da Aprendizagem Computacional são possíveis e encontram-se presentes no nosso dia a dia, com o intuito de oferecer ao ser humano um estilo de vida melhor.

O projecto desenvolvido neste estágio consiste no desenvolvimento de um protótipo de uma aplicação web para ajudar na previsão e prevenção de fogos florestais com base em técnicas de Aprendizagem Computacional.

Inicialmente fora pretendido que tal projecto incorporasse a iniciativa *Big Data for Social Good* da GSMA [28]. Esta iniciativa visava identificar usos que beneficiassem as comunidades para a grande quantidade de dados que os operadores móveis recolhem durante a sua actividade normal. Pretendia-se que fossem considerados dados provenientes de operadoras telefónicas portuguesas com o intuito de através destes extrair informação da movimentação da população de Portugal.

A não disponibilização dos dados por parte das operadoras telefónicas portuguesas, um dos riscos identificados e apresentados no capítulo 3 na secção 3.4, originou com que o âmbito do presente projecto fosse alterado.

Apesar da clara necessidade da alteração do âmbito do projecto em questão, era pretendido que tal alteração não resultasse em algo totalmente diferente daquilo que era o âmbito inicial do projecto.

Desta forma, apesar de não ser possível a consideração de dados provenientes de operadores móveis pretendia-se que fossem considerados outros tipos de dados que permitissem obter conhecimento da população de Portugal

Assim sendo passaram a ser considerados dados relativos ao número de habitantes dos diferentes concelhos de Portugal, provenientes dos resultados dos Censos [20] em contrapartida aos dados provenientes das operadoras telefónicas portuguesas inicialmente idealizados para o desenvolvimento deste projecto.

A aplicação web desenvolvida permite efectuar a previsão da ocorrência de incêndios florestais com base em dados de ocorrências de incêndios florestais históricas, de condições meteorológicas históricas e de dados populacionais.

A previsão efectuada é resultante da aplicação de diversas técnicas e da construção de diferentes modelos de Aprendizagem Computacional que serão abordados no presente relatório no capítulo 5 na secção 5.1.

1.1.2 Motivação

Nos dias de hoje existem já soluções que permitem efectuar a previsão e prevenção de incêndios florestais. Estas soluções consideram dados de ocorrências de fogos florestais, meteorológicos, das condições da superfície terrestre, mas não consideram dados relacionados com a população.

Numa actualidade em que incêndios florestais causados pelo ser humano são uma realidade cada vez mais comum faz sentido existir uma solução que tenha em consideração o factor humano.

O projecto em questão pretende incorporar dados relativos ao número de habitantes nos diferentes concelhos de Portugal com o intuito de investigar até que ponto o número de habitantes é influente na ocorrência de incêndios florestais e extrair padrões relevantes que permitam melhorar a previsão e prevenção de incêndios florestais.

1.1.3 Objectivos

Nesta secção encontram-se apresentados os objectivos definidos para o período deste estágio.

Uma vez que os objectivos foram modificados devido à alteração do âmbito do presente estágio, serão apresentados em primeiro lugar os objectivos inicialmente idealizados e posteriormente os objectivos definidos aquando da alteração do âmbito do estágio em questão.

Os objectivos inicialmente definidos e os objectivos definidos após a alteração do âmbito do estágio foram definidos pela empresa Grama [27] e encontram-se de seguida retratados.

Objectivos iniciais

1. Desenvolvimento de um protótipo de uma aplicação web para ajudar entidades governamentais na previsão e prevenção de fogos florestais
2. Investigar soluções para carregamento e processamento de grandes quantidades de dados anonimizados de operadores móveis
3. Aplicação de técnicas de Aprendizagem Computacional sobre o banco de dados de operadores móveis de modo a retirar padrões que possam ser utilizados na previsão e prevenção de fogos florestais

Objectivos após a alteração do âmbito do estágio

1. Desenvolvimento de um protótipo de uma aplicação web para ajudar entidades governamentais na previsão e prevenção de fogos florestais
2. Investigar soluções alternativas à utilização de dados de operadores móveis que considerem o factor humano
3. Aplicação de técnicas de Aprendizagem Computacional sobre o novo conjunto de dados a considerar de forma a extrair padrões que possam ser utilizados na previsão e prevenção de fogos florestais

Para além dos objectivos acima mencionados, espera-se que o aluno tenha a autonomia e capacidade crítica necessárias para recolher, analisar e estruturar toda a informação acerca de soluções existentes no mercado, informação esta que deverá ser posteriormente aplicada no desenvolvimento do protótipo.

O aluno deve também efectuar a análise de requisitos do projecto, realizar um plano de desenvolvimento e definir um *projet backlog* com as funcionalidades apresentadas na Secção 3.3 - Requisitos funcionais do Capítulo 3 - Especificações.

O aluno deve ainda possuir a capacidade de aplicar conceitos e conhecimento da área de Aprendizagem Computacional adquiridos durante o mestrado no desenvolvimento do protótipo.

No final do estágio o aluno deve ter desenvolvido um protótipo funcional que seja facilmente demonstrável e que demonstre bem os conceitos explorados durante o período de estágio.

1.1.4 Estrutura do documento

Nesta secção encontram-se apresentados os capítulos que constituem o presente relatório bem como o conteúdo presente em cada um dos diferentes capítulos.

- Capítulo 1 - Introdução
  - Este primeiro capítulo do documento é tido como um capítulo explicativo. É o capítulo onde é apresentado o ambiente de trabalho do estágio, o contexto do projecto em questão, a motivação que levou ao desenvolvimento do projecto, os objectivos determinados para a realização do estágio e do projecto, e a forma de como o relatório se encontra estruturado.
- Capítulo 2 - Estado da arte
  - É o capítulo onde é feito o estudo da arte relativa à previsão e prevenção de fogos florestais. Neste capítulo é feita inicialmente uma revisão de literatura de artigos notáveis relativos à previsão e prevenção de fogos florestais. Posteriormente são analisados os sistemas base de recolha de dados bem como algoritmos de Aprendizagem Computacional capazes de apresentar solução para o problema a retratar neste projecto. Por fim são analisadas plataformas já existentes para a previsão e prevenção de incêndios florestais.
- Capítulo 3 - Especificações
  - É o capítulo onde se encontram apresentadas todas as especificações do projecto em questão, nomeadamente o modelo de desenvolvimento escolhido, os requisitos funcionais e não funcionais do projecto, a análise de riscos associados, a arquitectura adoptada e as tecnologias e ferramentas utilizadas no desenvolvimento do projecto em questão.
- Capítulo 4 - Conjunto de dados
  - É o capítulo onde são apresentadas todas as etapas realizadas na construção do conjunto de dados considerado no presente projecto
- Capítulo 5 - Desenvolvimento
  - É o capítulo onde se encontram detalhadas ao nível da implementação as diferentes fases do desenvolvimento do protótipo desenvolvido neste estágio. Encontra-se descrito neste capítulo todo o processo de desenvolvimento, nomeadamente à aplicação de técnicas de Aprendizagem Computacional, à construção dos modelos de Aprendizagem Computacional considerados bem como às funcionalidades implementadas ao nível do back-end e front-end da aplicação web.

- Capítulo 6 - Testes e resultados obtidos
  - É o capítulo onde se encontram apresentados os diferentes testes e resultados obtidos relativamente ao desempenho dos modelos desenvolvidos no presente estágio. É também efectuada neste capítulo a análise crítica dos respectivos resultados obtidos.
- Capítulo 7 - Conclusão
  - Este último capítulo do documento é tido como um capítulo de sintetização. É o capítulo onde se sintetiza todo o trabalho desenvolvido e se destacam os aspectos cruciais e sucessos alcançados.

Capítulo 2

Estado da Arte

O presente capítulo tem por objectivo fazer uma análise do estado da arte no que diz respeito à previsão e prevenção de fogos florestais.

Neste capítulo começará por ser feita a análise dos sistemas base de recolha de dados para a previsão e prevenção de fogos florestais, bem como dos algoritmos de **Aprendizagem Computacional** capazes de apresentar solução para o problema a retratar neste projecto.

De seguida será efectuada a revisão de literatura de artigos notáveis sobre o tema abordado, nomeadamente no que diz respeito às metodologias e dados utilizados, bem como aos feitos alcançados em cada um destes mesmos artigos.

Por último mas não menos importante serão analisadas as plataformas já existentes para o tema em questão, nomeadamente no que diz respeito às funcionalidades oferecidas por cada uma das mesmas.

2.1 Sistemas base de recolha de dados

A presente secção tem por objectivo apresentar os sistemas de recolha de dados utilizados por parte de sistemas de previsão de fogos florestais, bem como utilizados em artigos notáveis que abordam o tema em questão.

2.1.1 Canadian Forest Fire Weather Index System (FWI)

O FWI [38] foi desenvolvido pelo Serviço Canadiano de Florestas com o intuito de permitir estimar o risco de incêndio com base no estado dos diferentes combustíveis presentes no solo florestal, bem como em elementos meteorológicos de forma indirecta.

Encontra-se dividido em 6 componentes distintas calculadas com base nos valores dos elementos meteorológicos que realizam a avaliação dos diferentes estados do solo possíveis para apresentar o índice final, **Fire Weather Index**, índice este que representa uma classificação numérica da intensidade do fogo.

O diagrama que se segue ilustra as diferentes componentes do FWI interligadas.

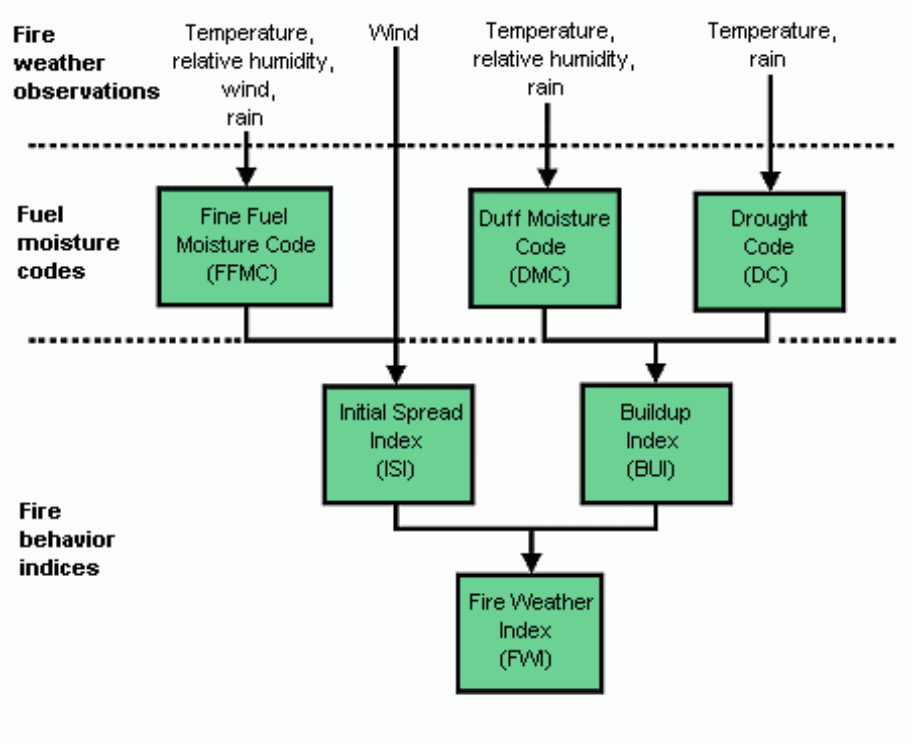


Figura 2.1: Estrutura do sistema FWI [38]

O cálculo das diferentes componentes é baseado em observações diárias constantes de temperatura, humidade relativa, velocidade do vento e precipitação durante as 24 horas.

O sistema encontra-se repartido em diferentes componentes interligadas. De seguida são contextualizadas cada uma das diferentes componentes.

Índice de humidade dos combustíveis finos

O índice de humidade dos combustíveis finos (FFMC) representa uma classificação numérica do teor de humidade dos combustíveis finos mortos, de secagem rápida. Representa o grau de inflamabilidade deste tipo de combustíveis presentes à superfície do solo. [38]

Índice de humidade

O índice de humidade (DMC) representa uma classificação numérica do nível médio de humidade das camadas orgânicas pouco compactadas e de profundidade significativa, até cerca de 8cm de profundidade. Este índice fornece informação ao nível do consumo de combustível em camadas de profundidade significativa. [38]

Índice de seca

O código de seca (DC) representa uma classificação numérica do nível médio de humidade das camadas orgânicas compactas e profundas, entre 8cm e 20 cm de profundidade. Este índice serve de indicador dos efeitos sazonais de seca nos combustíveis florestais. [38]

Índice de dispersão inicial

O índice de dispersão inicial (ISI) representa uma classificação numérica da taxa prevista de propagação de incêndio. Faz recurso dos efeitos do vento e do índice de humidade dos combustíveis finos explicado anteriormente para o cálculo da taxa de propagação. [38]

Índice de combustível disponível

O índice de acumulação (BUI) é um factor de avaliação dos combustíveis presentes no solo que podem alimentar um fogo, representa uma classificação numérica da quantidade de combustível existente para combustão. Para tal faz recurso do índice de humidade e do índice de seca. [38]

Índice de risco de incêndio

O índice de risco de incêndio (FWI) representa uma classificação numérica da intensidade do fogo. Faz recurso do índice de dispersão inicial e do índice de acumulação para tal. [38]

Classificação diária de gravidade

A classificação diária de gravidade (DSR) representa uma classificação numérica da dificuldade de controlar incêndios. Baseia-se no índice de risco de incêndio, porém oferece mais precisão no que diz respeito às necessidades para o controlo e término do incêndio. [38]

Este sistema é utilizado por vários países do mundo para estimar o risco de incêndio com base no estado dos combustíveis presentes no solo, em particular na Europa.

Estudos como [48], [29], [44], [60], [46] e [71] tem por base este sistema para realizar a previsão de fogos florestais.

2.1.2 Moderate Resolution Imaging Spectroradiometer (MODIS)

Moderate Resolution Imaging Spectroradiometer (MODIS) [4] da National Aeronautics and Space Administration (NASA) é o sensor de satélite mais utilizado para detectar incêndios em grandes regiões.

MODIS é um sensor que se encontra a bordo do satélites Terra (EOS AM-1) e Aqua (EOS PM-1). Estes dois satélites encontram-se a visualizar a superfície da Terra a cada 1 a 2 dias adquirindo dados em 36 bandas espectrais ou grupos de comprimentos de onda. Juntamente com todos os dados de outros instrumentos a bordo destes satélites os dados do MODIS são transferidos para estações terrestres.

Os diferentes dados extraídos das observações do MODIS descrevem características da superfície terrestre, oceanos e atmosfera como a temperatura da superfície, mapas da vegetação global ou nível de concentração de aerossóis e que podem ser considerados em diferentes estudos a nível global.

Em [12] presenciamos a utilização de dados provenientes do sensor MODIS.

2.1.3 Geographic information system (GIS)

Geographic Information System (GIS) é um sistema concebido para capturar, armazenar, verificar e exibir dados referentes a posições da superfície terrestre. GIS é uma parte relevante da infraestrutura de dados espaciais. Este sistema pode utilizar qualquer informação que disponha de localização, sendo que esta mesma localização pode ser expressa por diversas formas como a correlação latitude/longitude.

O GIS pode incluir dados de diferentes áreas interligados, como dados relativos às características da paisagem existente, diferentes tipos de vegetação, diferentes tipos de solo, postos de combustível e acessos rodoviários bem como dados relativos à população existente.

Através deste sistema podemos proceder à comparação de dados e encontrar relações entre os mesmos.

No caso do tema em questão, ao utilizarmos o GIS num mapa, com o recurso a dados históricos de fogos florestais, podemos verificar que características possuem as zonas mais propícias de incêndio bem como da população existente em redor dessas mesmas zonas, o que fornece ajuda à previsão e prevenção de fogos florestais.

Estudos ao nível da previsão de fogos florestais como [12] e [14] baseam-se no sistema GIS para obter dados relativos às características da superfície da Terra.

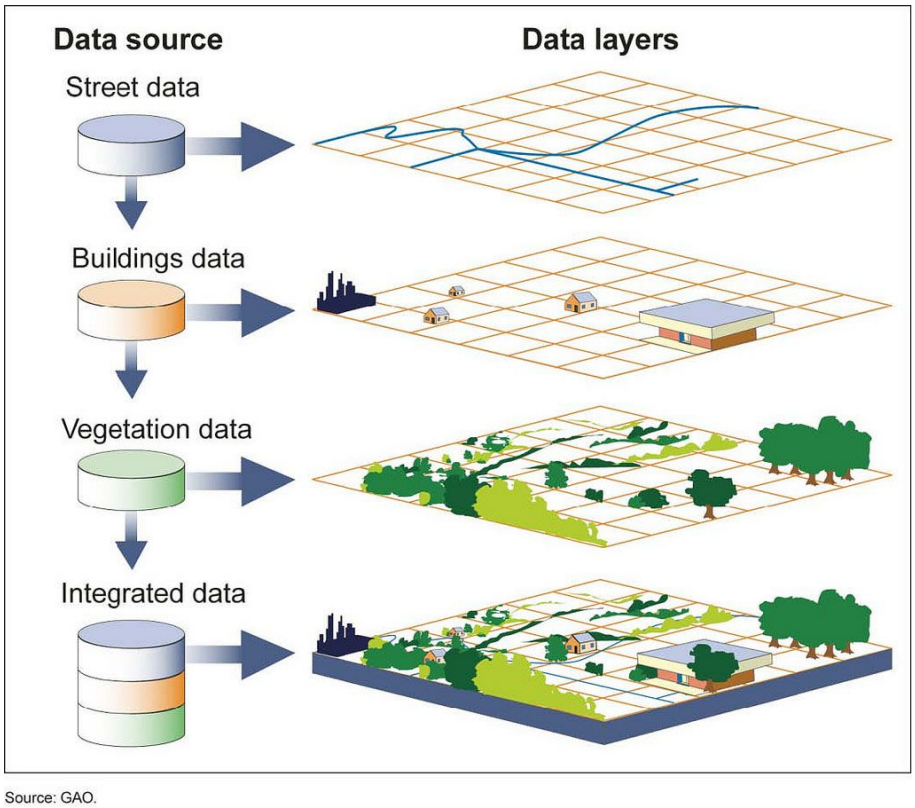


Figura 2.2: Mapa com recurso ao sistema GIS [25]

2.1.4 Dados meteorológicos

As condições meteorológicas são um factor bastante condicionante na ocorrência de fogos

florestais.

A recolha de dados meteorológicos com recurso a sistemas de previsão meteorológica bem como a dados meteorológicos históricos são bastante relevantes e utilizados em praticamente todos os sistemas e estudos existentes no que diz respeito à previsão de fogos florestais.

São dados que podem ser recolhidos em tempo real e com custos muito baixos, um dos motivos pelo qual são bastante utilizados.

Estudos como [12],[29],[44],[26],[71],[60] e [46] fazem recurso a dados meteorológicos com o intuito de prever a ocorrência de fogos florestais.

2.2 Conceitos de Aprendizagem Computacional

Aprendizagem Computacional é a área que se dedica a ensinar os computadores a fazer o que os humanos e animais fazem intuitivamente, aprender com a experiência. Os algoritmos de **Aprendizagem Computacional** utilizam métodos computacionais para "aprender"informação directamente dos dados, sem a necessidade de uma equação predeterminada como modelo. Estes algoritmos melhoram progressivamente o seu desempenho à medida que conjunto de dados para aprendizagem aumenta. [35]

Nesta secção serão abordados alguns conceitos de **Aprendizagem Computacional** pertinentes para a compreensão da secção que se segue.

2.2.1 Aprendizagem supervisionada

Este tipo de aprendizagem tem por objectivo a construção de um modelo capaz de realizar previsões baseadas em evidências na presença de incerteza. [35]

Para tal são apresentados ao sistema exemplos de dados de entrada e saída desejados, com o intuito de treinar um modelo capaz de prever a resposta de novos dados de entrada.

A **Aprendizagem supervisionada** é feita com base em técnicas de **Classificação** e **Regressão**.

Classificação

Técnica de **Aprendizagem supervisionada** que prevê uma resposta discreta de um conjunto de dados, como por exemplo a ocorrência/não ocorrência de um fogo florestal.

Os modelos de **Classificação** classificam os dados de entrada em diferentes categorias (classes).

Regressão

Técnica de **Aprendizagem supervisionada** que prevê uma resposta contínua de um conjunto de dados, como por exemplo a área queimada por um fogo florestal.

2.2.2 Aprendizagem não supervisionada

Neste tipo de aprendizagem, ao contrário da aprendizagem supervisionada não são apresentados ao sistema exemplos de dados de entrada e saída desejados.

Assim sendo, este tipo de aprendizagem é baseada na observação e descoberta. O sistema procura encontrar padrões úteis para o problema de uma forma autónoma.

Clustering

É a técnica de **Aprendizagem não supervisionada** mais utilizada. É utilizada com o intuito de encontrar padrões relevantes nos dados em questão.

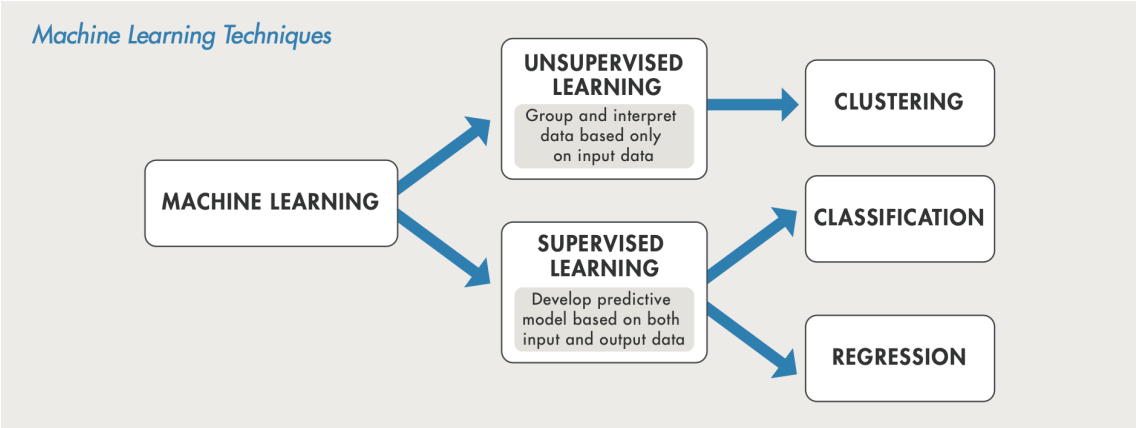


Figura 2.3: Técnicas de Aprendizagem Computacional [35]

2.2.3 Aprendizagem em conjunto

Este tipo de aprendizagem tem por base o agrupamento de decisões de vários modelos para melhorar o desempenho geral. Ajuda a minimizar factores de erro nos modelos de aprendizagem como o ruído e a variância, foca-se em melhorar a estabilidade e a precisão dos algoritmos de **Aprendizagem Computacional**.

2.2.4 Overffiting

O fenómeno de overfitting ocorre quando no período de treino são estabelecidas fronteiras de decisão demasiado ajustadas ao conjunto de treino. Isto resulta num bom desempenho do modelo durante a fase de treino, no entanto o modelo perde capacidade de generalização tornando-se mais propício a erros de classificação para dados que lhe são desconhecidos.

2.3 Algoritmos de Aprendizagem Computacional

Na presente secção será feita a análise de diferentes algoritmos de Aprendizagem Computacional utilizados nos artigos que serão referidos adiante na secção 2.4 para fazer frente

ao problema que se pretende combater com o desenvolvimento deste projecto.

2.3.1 Rede Neuronal Artificial (RNA)

Uma Rede Neuronal Artificial (RNA) é um modelo computacional na sua génese baseado no conhecimento das redes neuronais biológicas, que tem a característica de conseguir adquirir conhecimento através da experiência.

É representada como um sistema de "neurónios" ligados entre si capazes de processar valores de entrada em valores de saída, permitindo, através da variação dos seus pesos, obter relações funcionais entre as entradas e as saídas muito variadas.

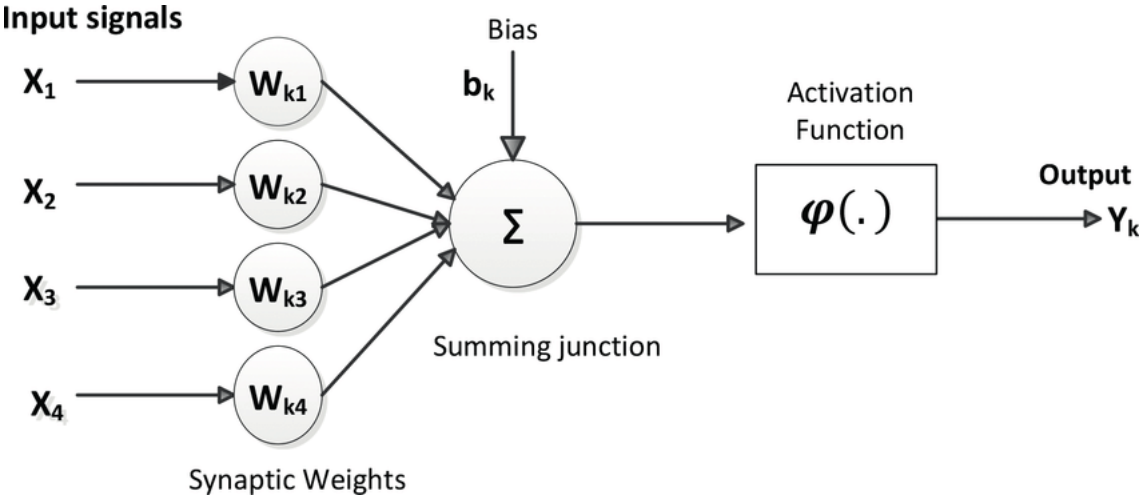


Figura 2.4: Modelo base de uma Rede Neuronal Artificial (RNA) (Perceptron) [33]

A RNA recebe como entrada um conjunto de sinais (X1,...,Xn). De seguida, cada sinal é multiplicado por um peso (Wk1,...,Wkn) que representa a importância final do sinal de entrada na saída da unidade de processamento. Posto isto é feita a soma ponderada dos diferentes sinais que produz um dado valor. Por fim, consoante o valor obtido e a função de activação em causa é devolvido um valor de saída (Yk).

As RNA apresentam uma regra de treino com o intuito de ajustar os pesos entre as ligações para alcançar os resultados desejados, para que o resultado obtido por parte da RNA para um dado conjunto de dados e o resultado esperado para esse mesmo conjunto sejam o mais próximo possíveis. As RNA aprendem através do treino com base em exemplos e estão geralmente organizadas em camadas.

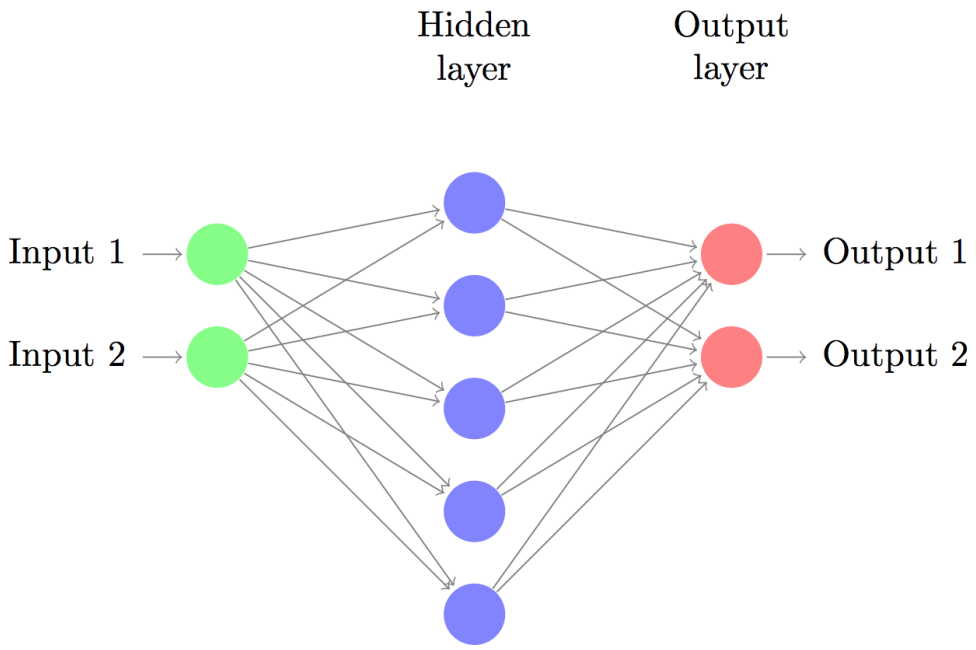


Figura 2.5: Modelo de uma RNA organizada em camadas (Multilayer Perceptron) [69]

As RNA são muito utilizadas para resolver problemas de classificação e regressão. Têm a capacidade de encontrar uma solução ideal para problemas não lineares complexos, como o foco deste trabalho, os incêndios florestais, descobrindo padrões entre os factores e respostas relativamente úteis para o tema em questão. Em [48],[29],[44],[71],[5] e [60] deparamo-nos com a utilização de redes neuronais artificiais com o intuito de solucionar o problema do estudo em questão.

2.3.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) são um algoritmo de aprendizagem supervisionada que analisam dados e reconhecem padrões. Este algoritmo é utilizado para resolver problemas de classificação e regressão.

No que diz respeito à classificação, este algoritmo procede à separação dos dados em duas classes, classes estas onde os elementos englobados apresentam um conjunto de características semelhantes, e posteriormente, procede à construção do hiperplano que apresenta a maior separação possível para estas mesmas classes.

Para calcular a maior separação possível entre as classes no momento de construção do hiperplano o algoritmo faz recurso aos **vetores de suporte**, mais concretamente à distância que os separa. Os **vetores de suporte** representam os elementos da classe que fixam o limite de separação das classes, designado de margem.

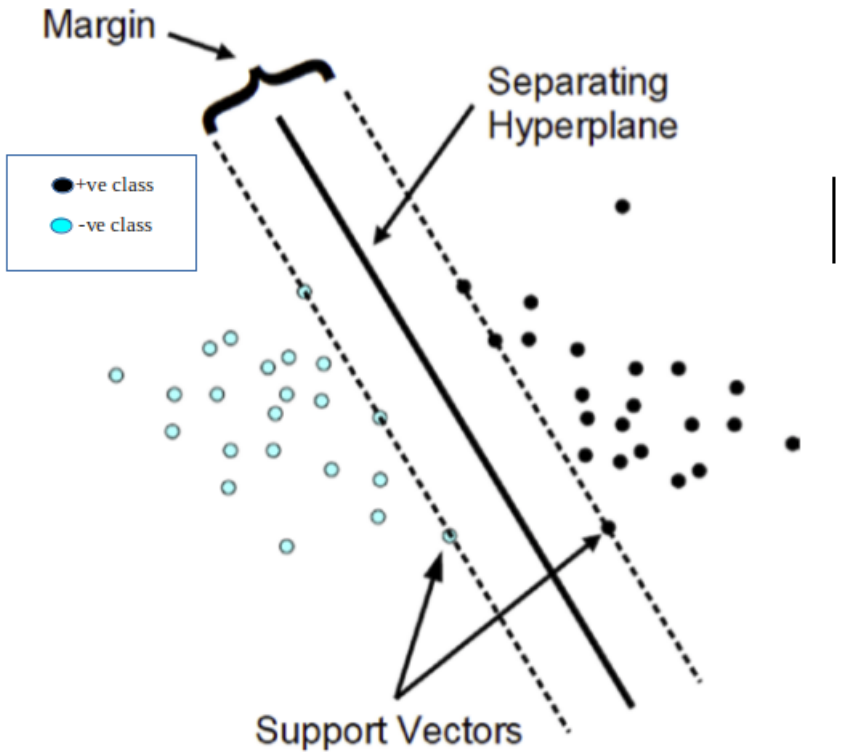


Figura 2.6: Construção de um hiperplano com base nos vetores de suporte [50]

Certos problemas não permitem que seja feita uma separação linear das classes. As Support Vector Machine (SVM) têm a capacidade de resolver problemas não-linearmente separáveis através da utilização de funções de kernel que aumentam a dimensão do problema, de tal modo que no espaço aumentado o problemas são linearmente separáveis. As funções de kernel permitem que seja possível mapear os dados numa dimensão diferente da original e ao mesmo tempo seja possível a construção de um hiperplano linear que assegure a divisão das classes. Posto isto os dados voltam a ser mapeados na sua dimensão original.

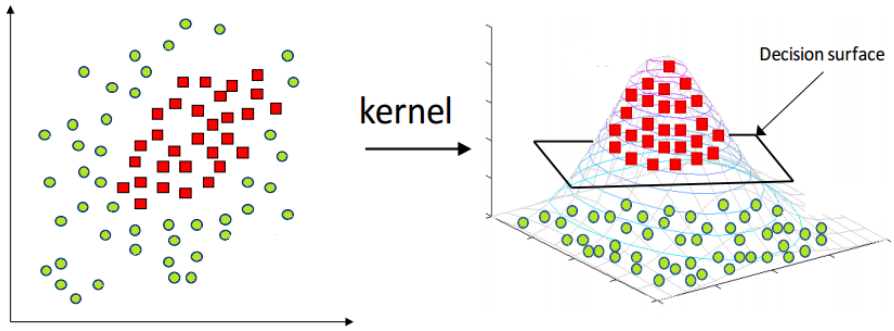


Figura 2.7: Aplicação de uma função de kernel [37]

Na figura **2.6** observamos a utilização de uma função de kernel a um conjunto de dados. Através de um hiperplano linear de dimensão 2 não seria possível separar os dados em duas classes, verde e vermelha. Assim, para que a separação em duas classes seja possível fazemos recurso a uma função de kernel que transforma os pontos originais para pontos de dimensão

3 onde já é possível construir o hiperplano. No fim de o hiperplano estar construído necessitamos apenas de projetar os dados juntamente com o hiperplano calculado na sua dimensão original.

Estudos como [48], [44], [26] e [46] consideram máquinas de vetor de suporte como tentativa de resposta ao problema em causa.

2.3.3 Decision Trees (DT)

Decision Trees (DT) são um algoritmo de aprendizagem supervisionada muito utilizados em problemas de classificação e regressão e permitem a análise de várias variáveis. As DT permitem prever, explicar e classificar um resultado. Estas fazem a classificação da raiz para as folhas onde os nós representam as características das instâncias, os ramos os diferentes possíveis cenários que podem ser tomados e as folhas que correspondem ao resultado final, à decisão tomada pela execução do algoritmo.

Em [5], [44] e [12] deparamo-nos com a utilização de DT como tentativa de solução ao problema retratado nestes estudos.

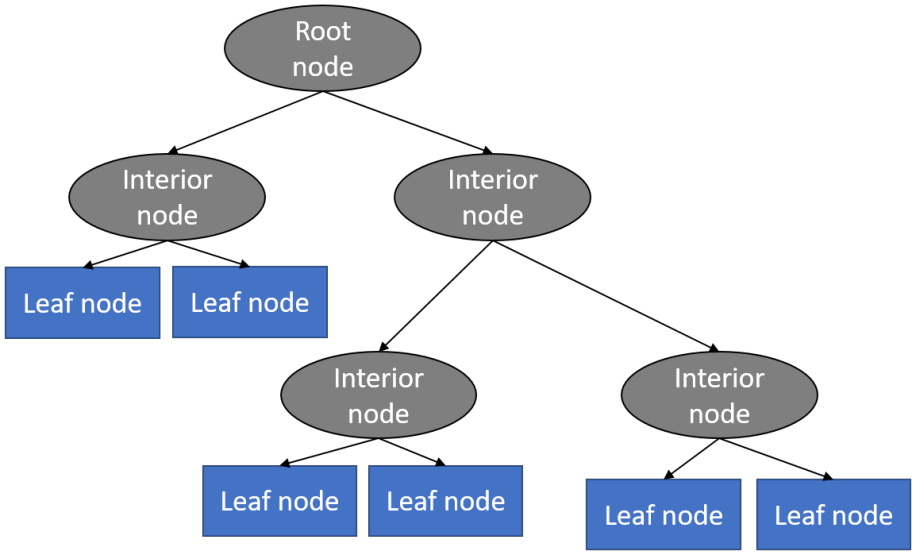


Figura 2.8: Decision Tree [11]

2.3.4 Random Forest

Random Forest é um algoritmo de aprendizagem supervisionada utilizado em problemas de classificação e regressão. Este algoritmo recorre à criação de um dado número de **Decision Trees (DT)** (2.3.3) com base nas características de um determinado conjunto de dados de entrada e posteriormente faz recurso ao algoritmo **Majority Vote** para atribuir a classe correspondente ao conjunto de dados de entrada, onde a classe predominante nas diferentes DT é a classe final atribuída.

Estudos como [44], [12] e [46] consideram Random Forest na procura de uma solução ao problema retratado no estudo em causa.

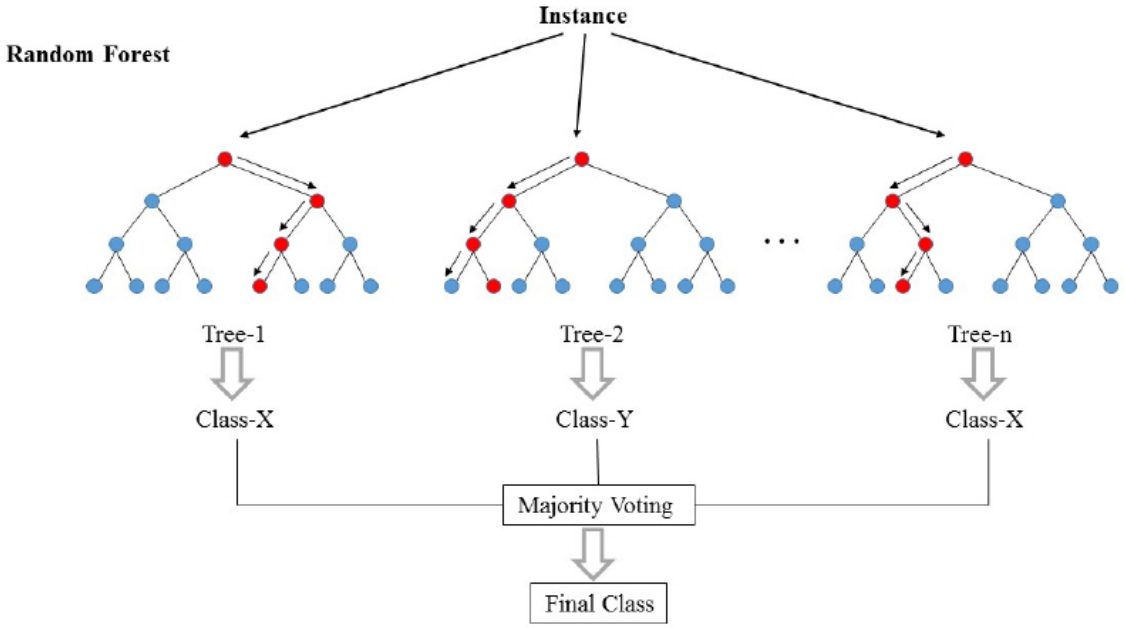


Figura 2.9: Random Forest [61]

2.3.5 Gradient Boosting

Gradient Boosting é um algoritmo de aprendizagem supervisionada utilizado em problemas de regressão e classificação que produz um modelo de previsão baseado num conjunto de modelos de previsão, geralmente **Árvores de previsão**. Trata-se de um algoritmo iterativo, as diferentes árvores criadas na sua execução são criadas de forma sequencial, utilizando os resultados das árvores anteriores para criar as árvores seguintes e minimizar possíveis erros.

Em [46] uma das abordagens utilizadas na procura da melhor solução é Gradient Boosting.

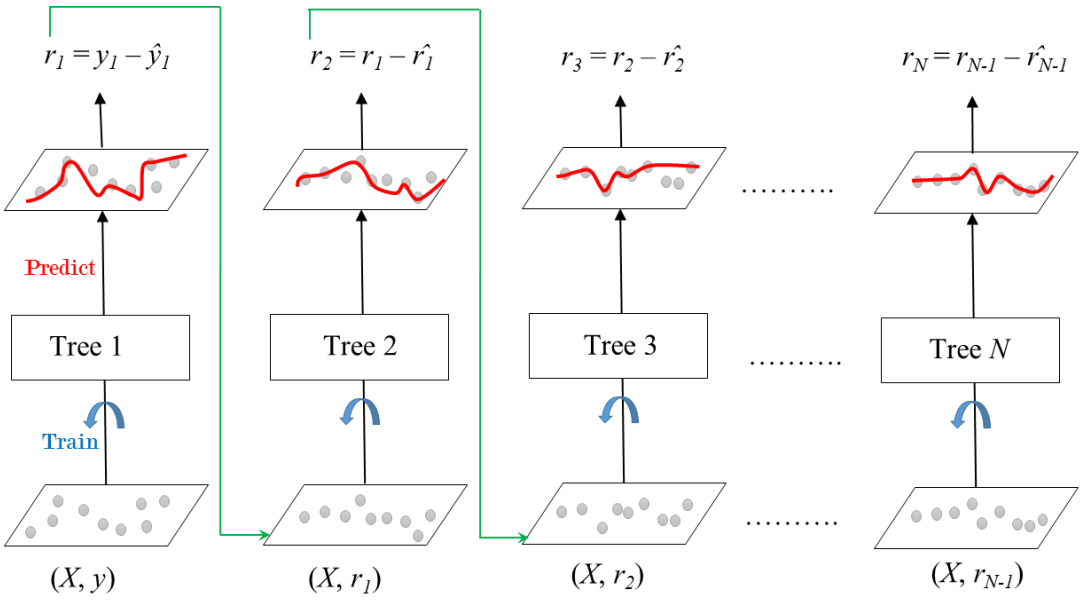


Figura 2.10: Gradient Boosting [13]

2.3.6 Logistic Regression

Logistic Regression é um algoritmo de aprendizagem utilizado em problemas de classificação que tem por base o conceito de probabilidade.

Este algoritmo permite a previsão de um resultado discreto a partir de um conjunto de variáveis. A variável dependente ou de resposta é dicotómica, como sucesso/falha. No caso do tema em questão podemos abordar como a presença/ausência de um fogo florestal.

Logistic Regression utiliza a função de custo **Sigmoid**, também conhecida como **Logistic Function** para mapear valores previstos para probabilidades. A função **Sigmoid** tem a capacidade de mapear qualquer valor real presente no intervalo [0,1]. [63]

Por forma a classificar as saídas de acordo com os diferentes valores de probabilidades, é depois definido um **threshold** como medida de separação, onde no caso de existirem duas classes um valor menor que o **threshold** respresenta um elemento de uma classe e um valor maior um elemento de outra classe. [63]

Estudos como [12] e [46] utilizam Logistic Regression na procura da melhor solução ao problema em questão.

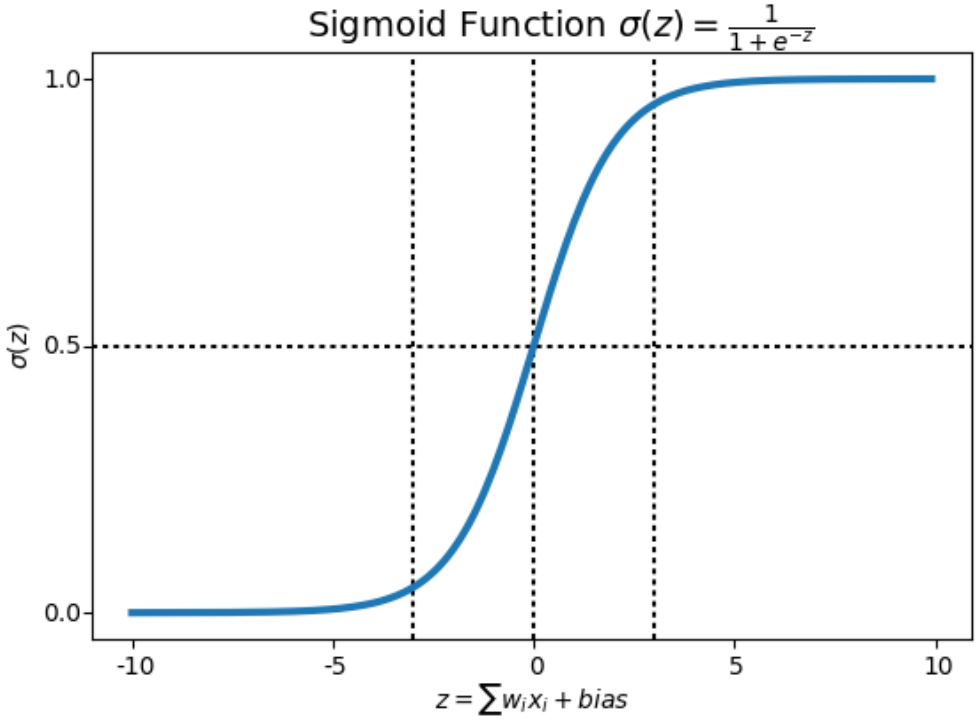


Figura 2.11: Função Sigmoid [63]

2.3.7 K-means clustering

K-means clustering é um algoritmo de aprendizagem não supervisionado. O objectivo deste algoritmo é agrupar conjuntos de dados semelhantes e descobrir padrões existentes

nos dados em causa. Para atingir este objectivo este algoritmo procura um dado número de **clusters**(n) no conjunto de dados, sendo que um cluster se refere a uma conjunto de dados que se encontram agregados devido às semelhanças existentes. O número de **clusters**(n) define também o número de **centroids**(n) existentes. Entende-se por **centróide** o local que representa o centro do **cluster**. Na execução deste algoritmo é identificado previamente o número de **centróides**(n) e depois os dados existentes são alocados para o **cluster** mais próximo. Com a execução iterativa do algoritmo os **centróides** podem ser alterados se tal permitir a optimização dos **centróides**.

Em **K-means clustering**, **means** refere-se à média dos dados, ou seja, a encontrar o **centroid** em questão.

Em [60] deparamo-nos com a utilização de K-means na procura da melhor solução para o problema em causa.

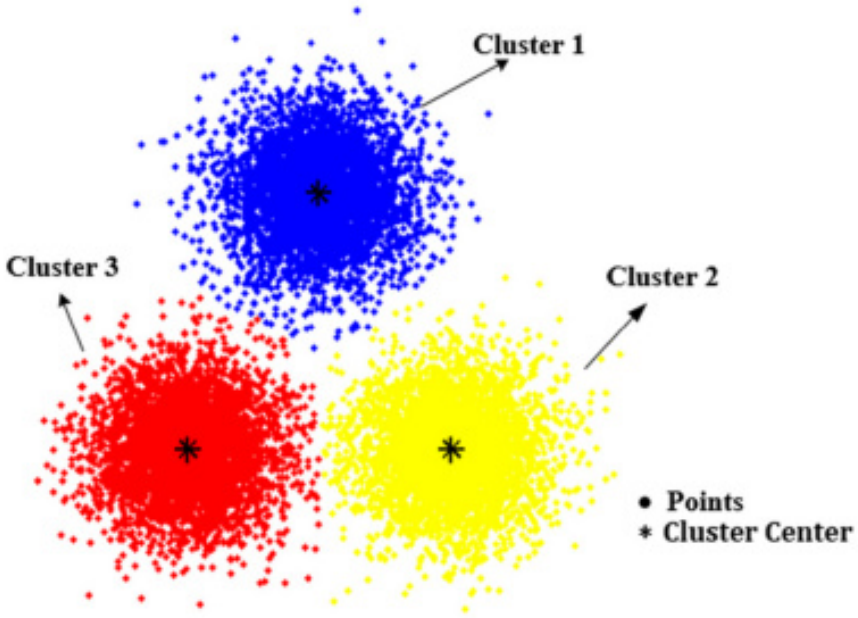


Figura 2.12: K-means clustering [31]

2.3.8 Bootstrap aggregating

Bootstrap aggregating também conhecido como **Bagging** é um algoritmo de aprendizagem em conjunto concebido para melhorar a estabilidade e a precisão de algoritmos de aprendizagem utilizados em problemas de classificação e regressão. Este algoritmo ajuda também no que toca a reduzir a variância e a evitar overffiting (2.2.4).

Neste algoritmo os modelos são criados independentemente uns dos outros e treinados com base numa amostra de dados de treino. Esta amostra é alterada de modelo para modelo através da substituição dos dados contidos na mesma, no momento de treino de cada modelo. No final os resultados são combinados através da média, no caso de se tratar de um problema regressão, ou da votação, no caso de um problema de classificação.

Em [12] observamos que a utilização de Bootstrap aggregating ao problema em questão garantiu os melhores resultados no que diz respeito ao nível de **precisão**.

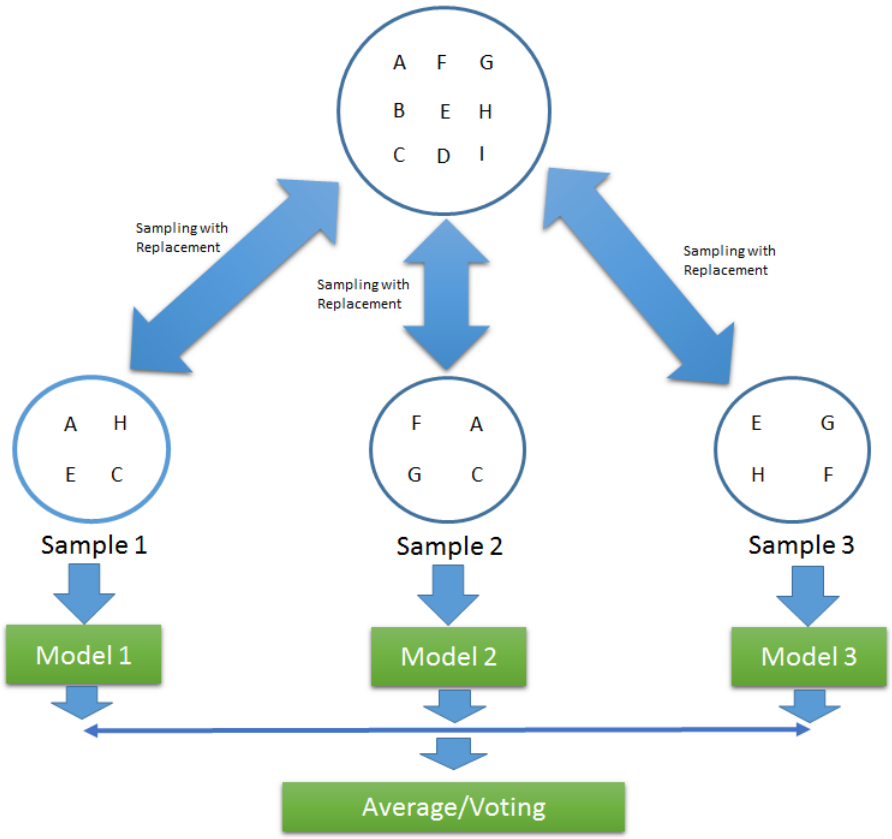


Figura 2.13: Bootstrap aggregating (Bagging) [40]

2.4 Revisão de literatura

A presente secção tem por objectivo efectuar a revisão de literatura de artigos notáveis referentes à previsão e prevenção de fogos florestais.

Em cada artigo serão analisadas as metodologias e dados utilizados bem como os sucessos alcançados em cada um dos mesmos.

Os artigos considerados encontram-se de seguida apresentados por ordem cronológica.

2.4.1 Learning to Predict Forest Fires with Different Data Mining Techniques

Neste artigo Stojanova et al (2006) [12] realizaram um estudo ao nível da previsão de fogos florestais que incide na Eslovénia.

Para tal os autores utilizaram dados provenientes de GIS para recolher dados relativos às condições florestais, sistema de previsão meteorológico Aladin para dados relativos às condições meteorológicas e dados de satélite provenientes do sensor **MODIS**.

No desenvolvimento do estudo foram aplicados diferentes algoritmos de **Aprendizagem Computacional** por forma de criar modelos de previsão de fogos florestais. **Logistic regression**, **Decision Trees (DT)**, **Random Forest**, **Bootstrap aggregating** e **AdaBoost** foram os algoritmos escolhidos para o desenvolvimento deste estudo.

Os melhores resultados de **precisão**, de entre todos os algorimtos utilizados, foram obtidos com recurso ao algoritmo **Bootstrap aggregating**.

2.4.2 A Data Mining Approach to Predict Forest Fires using Meteorological Data

Neste artigo Cortez et al (2007) [44] fizeram recurso a diferentes algoritmos de **Aprendizagem Computacional**, a dados meteorológicos, e componentes do Canadian Forest Fire Weather Index System que tinham em conta factores espaciais e temporais.

Os algoritmos de **Aprendizagem Computacional** utilizados no desenvolvimento deste estudo foram **Support Vector Machine**, **Random Forest**, **Multiple Regression**, **Decision Trees** e **Redes neuronais artificiais**.

Este estudo incide sobre o parque natural de Montesinho localizado em Trás os Montes. Os dados utilizados ao nível das condições meteorológicas e ocorrências de fogos foram recolhidos entre Janeiro de 2000 e Dezembro de 2003 [47].

Foram utilizados quatro conjuntos de dados no período experimental deste estudo. Um primeiro englobando dados espaciais, temporais e quatro componentes do Canadian Forest Fire Weather Index System. Um segundo englobando dados espaciais, temporais e quatro variáveis meteorológicas. Um terceiro considerando quatro componentes do Canadian Forest Fire Weather Index System e por fim, um quarto considerando quatro variáveis meteorológicas.

De entre os diferentes algoritmos e conjuntos de dados utilizados na fase experimental, a melhor solução foi alcançada com recurso à utilização de uma **Support Vector Machine** e um conjunto de dados que considerava unicamente quatro variáveis meteorológicas, solução esta capaz de prever fogos de menor dimensão.

Variável	Intervalo de valores
Coordenadas X e Y	1 a 9
Mês	Janeiro a Dezembro
Dia	Segunda-feira a Domingo
FFMC	18.7 a 96.2
DMC	1.1 a 291.3
DC	7.9 a 860.6
ISI	0 a 56.1
Temperatura (°C)	2.2 a 33.3
Humidade Relativa (%)	15 a 100
Velocidade do vento (km/h)	0.4 a 9.4
Chuva (mm/m2)	0 a 6.4
Área queimada (ha)	0 a 1090.84

Conjunto de dados utilizados no estudo de Cortez et al (2007) [44]

2.4.3 Artificial Intelligence for Forest Fire Prediction

Neste artigo Sakr et al (2010) [26] os autores fazem recurso ao algoritmo **SVM** e a dados de entrada meteorológicos por forma a alcançar os melhores resultados para o estudo em questão.

Os dados utilizados neste estudo são provenientes do **Instituto de Pesquisa Agrícola do Líbano** e disponibilizam dados relativos ao território libanês entre 2000 e 20008. Para além dos dados meteorológicos foram também utilizados dados relativos ao número diário de fogos florestais, cedidos pelo **Ministério do Meio Ambiente do Líbano** e utilizados para criar quatro escalas de perigo de ocorrência de fogo florestal (1 a 4).

Um dos objectivos dos autores na realização deste estudo foi eliminar a necessidade de mecanismos de previsão de meteorológica por forma a reduzir possíveis fontes de erro. Assim, a estratégia utilizada pelos autores para este fim foi a de associar as variáveis meteorológicas de um dado dia à escala de de perigo de incêndio florestal do dia seguinte na fase de aprendizagem da **SVM**.

O desempenho da arquitectura implementada foi testada com recurso ao cálculo do erro médio do número de fogos florestais previstos.

Os resultados obtidos foram bastante favoráveis. Com duas classes de risco de ocorrência de fogo florestal a **SVM** conseguiu alcançar uma **precisão** de 96%, bem como a previsão de quatro classes com um erro baixo ao nível do número de fogos florestais previstos e ao nível da escala de perigo de fogo florestal prevista.

2.4.4 Pattern Clustering of Forest Fires based on Meteorological Variables and its Classification using Hybrid Data Mining Methods

Neste artigo Yu et al (2011) [71] realizaram um estudo ao nível da previsão do tamanho de um fogo florestal tendo em conta variáveis meteorológicas.

Para a realização do estudo foram utilizados os dados provenientes do estudo de Cortez et al. (2007) [44] já citado anteriormente. Para agrupar as variáveis meteorológicas presentes neste estudo os autores utilizaram um **Self organizing map** para tal efeito.

Após se encontrarem agrupados, estes dados foram utilizados como dados de entrada para os dois algoritmos de **Aprendizagem Computacional** considerados neste estudo, **Rede neuronal de retropropagação** e **Sistema baseado em regras**.

A **Rede neuronal de retropropagação** recebeu para treino quatro variáveis meteorológicas de treino de cada **cluster**. O resultado do treino realizado pela **Rede neuronal de retropropagação** foi a área queimada sob a forma de uma variável categórica, cujas classes definidas foram pequeno, médio, grande e extremamente grande.

O **Sistema baseado em regras** foi também desenvolvido com base nas variáveis meteorológicas, dados de entrada, e área queimada, dados de saída.

Ao nível da fase experimental os resultados obtidos comprovaram que os dois algoritmos utilizados obtiveram uma precisão considerável, embora a **Rede neuronal de retropropagação** tenha conseguido atingir um nível de **precisão** mais elevado que o **Sistema baseado em regras**, visto que o **Sistema baseado em regras** não tem a capacidade de prever nenhum critério que não se encontre no conjunto de regras.

2.4.5 Applying Decision Tree Algorithm and Neural Networks to Predict Forest Fires in Lebanon

Neste artigo Karouni et al (2014) [5] os autores consideraram os algoritmos **Decision Trees** e **Redes neuronais de retropropagação** e dados meteorológicos com o intuito

de apresentar uma solução o problema em questão.

Os dados meteorológicos utilizados dizem respeito à região norte do Líbano em 2012 e foram disponibilizados pelo **Instituto de Pesquisa Agrícola do Líbano**. Foram também utilizados dados relativos à ocorrência de fogos florestais na região norte do Líbano.

Um dos objectivos dos autores na execução deste estudo foi a utilização de técnicas de **Aprendizagem Computacional** fáceis de implementar para prever, tomar acções rapidamente no que diz respeito a evitar fogos florestais, e, que oferecem-se um custo baixo.

Assim, os autores rejeitaram na realização deste estudo a utilização de índices climáticos devido a serem dispendiosos e complicados de considerar.

Ao nível da parte experimental foi utilizado o algoritmo **ID3** nas **Decision Trees**. Na construção da árvore as variáveis meteorológicas encontram-se agrupadas em 3 categorias (baixa, moderada e alta) e a decisão é dada por F (risco de ocorrência de fogo florestal elevado) e NF (risco de ocorrência de fogo florestal imprevisível).

Nas **Redes neuronais de retropropagação** foram aplicadas duas redes, **Rede neuronal Feedforward** e **Rede neuronal Cascade-forward** com 3 algoritmos de treino diferentes, **trainlm**, **traingdx** e **trainrp**. A entrada das **Redes neuronais de retropropagação** é um vector com 4 ou 2 variáveis meteorológicas e a saída é respectivamente 1 (risco de fogo florestal) ou 0 (ausência de risco de fogo florestal).

A **Árvore de decisão** implementada possuía a capacidade de classificar as variáveis meteorológicas, mais significativas e menos significativas, ignorar as menos significativas, e prever melhor a ocorrência de fogos florestais.

As **Redes neuronais de retropropagação** com algoritmos de treino diferentes com 2 dados meteorológicos de entrada comprovaram obter melhores resultados do que **Redes neuronais de retropropagação** com 4 dados meteorológicos de entrada com menos erro quadrado médio.

2.4.6 Predicting Size of Forest Fire using Hybrid Model

Neste artigo Shidik et Mustofa (2016) [29] seguiram uma abordagem com base em dois algoritmos de **Aprendizagem Computacional**, **Rede neuronal de retropropagação** e **Fuzzy C-Means (FCM)**.

Para a realização do estudo foram utilizados os dados provenientes do estudo de Cortez et al (2007) [44] já citado anteriormente.

Os dados são divididos em duas categorias de acordo com o campo respectivo à área queimada presente no conjunto de dados utilizados. Os dados cujo valor referente à área queimada é 0 são considerados como área não queimada. Os dados cujo valor é diferente de zero são agrupados de acordo com o algoritmo FCM para serem divididos em duas categorias, área pouco queimada e área muito queimada.

Posto isto, os dados de entrada considerados neste estudo sofreram um processo de normalização **min max** para que ficassem compreendidos no intervalo [0,1] e fossem utilizados para classificação por parte de uma **Rede neuronal de retropropagação**.

Após o processo de classificação por parte da **Rede neuronal de retropropagação** os resultados do desempenho obtido foram avaliados com recurso a uma matriz de confusão para medir a **precisão** do classificador auxiliada da medição estatística **Cohen's Kappa**

para avaliar a confiabilidade da observação de variáveis categóricas.

Os resultados obtidos pelos autores neste estudo apresentaram um desempenho da **Rede neuronal de retropropagação** de 97.5% no que diz respeito à **precisão** e 0.961 no que diz respeito ao índice de **Cohens Kappa**.

O melhor desempenho do modelo apresentado neste estudo é obtido com FCM utilizando **Cosine Similarity**.

Os autores compararam também o desempenho da **Rede neuronal de retropropagação** com outros algoritmos de classificação como **Máquinas de vetor de suporte**, **K-nearest neighbors**, **Decision Trees** e **Naive Bayes** que utilizaram também FCM para categorizar os dados.

A análise dos resultados obtidos comprovou que o desempenho da **Rede neuronal de retropropagação** apresentava melhores resultados quando comparado com o desempenho dos outros algoritmos de classificação.

2.4.7 Computational Intelligence and Data Mining Techniques using the Fire Data Set

Neste artigo Storer (2016) [60] propôs uma nova abordagem para a regressão dos dados utilizados anteriormente por Cortez et al (2007) [44].

Na análise do conjunto de dados a utilizar o autor apercebeu-se de dois problemas presentes no conjunto de dados.

O primeiro problema retratava que no conjunto de dados utilizados, dados que apresentassem um valor inferior a 100 m^2 de área queimada eram considerados insignificantes a nível estatístico e registados com um valor de área queimada de 0,0.

Tal facto dificultava a capacidade de prever fogos florestais de dimensões maiores com precisão devido a serem **outliers** e consequentemente sub-representados. Por forma a reduzir este problema o autor aplicou uma transformação logarítmica ao conjunto de dados.

O segundo problema constatava que as variáveis numéricas do conjunto de dados a utilizar não se encontravam normalizados pelo que certas variáveis do conjunto de dados podiam ser favorecidas em relação a outras devido à diferença de escalas utilizadas.

Para solucionar este problema o autor realizou a normalização de todas as variáveis numéricas no intervalo [0,1].

As variáveis categóricas foram codificadas de acordo com a codificação **1-of-C**. Este processo fez com que um total de 29 variáveis de entrada fossem consideradas em vez das 12 variáveis presentes no conjunto de dados.

Foram utilizadas dois tipos de **Redes neuronais artificias** como medida de solução para o problema. Uma treinada com **Backpropagation** e outra com **Particle Swarm Optimization**.

Um dos objetivos do autor foi verificar se variações na arquitetura da **Rede neuronal artificial** afectavam a taxa **RMSE** do problema de regressão. Foram assim considerados diferentes números de nós na camada oculta da **Rede neuronal artificial** em questão bem como consideradas cinco funções de activação diferentes, **Rectifier**, **Hyperbolic Tangent**, **Hard Sigmoid**, **Sigmoid** e **Linear**.

A parte experimental realizada por parte do autor demonstrou que ambas as **Redes neuronais artificiais** apresentavam um bom desempenho na tarefa de regressão de prever a área queimada de um fogo florestal.

A RNA treinada com **Particle Swarm Optimization** apresentou no entanto melhores resultados a nível de **RMSE** face à RNA treinada com **Backpropagation**, respectivamente 15,65 (12 variáveis de entrada) e 42,48 (29 variáveis de entrada). Tais resultados mostraram uma grande melhoria em relação a estudos anteriores.

Apesar dos melhores valores a nível de **RMSE** terem sido alcançados com o recurso a 12 variáveis de entrada o autor defende a utilização de 29 variáveis de entrada na utilização da **Rede neuronal artificial** treinada com **Particle Swarm Optimization** por forma a percorrer o espaço de pesquisa com mais confiabilidade.

O autor propõe também uma abordagem ao problema considerando as RNA já faladas anteriormente auxiliadas de algoritmos de **clustering**.

Os algoritmos de **clustering** utilizados pelo autor neste estudo foram respectivamente **K-means** e **Spectral**. Ambos foram utilizados para agrupar os dados em 2, 3 e 4 **clusters** e efectuados nos dados de entrada e de saída.

Cada número de **clusters** foi avaliado em consideração das 12 ou 29 variáveis de entrada já faladas anteriormente. Estes **clusters** foram depois utilizados por parte das RNA para executar a tarefa de classificação.

Os resultados da **precisão** são derivados de **cross validation**, nomeadamente **stratified k-Folds**.

No que diz respeito à parte experimental foram obtidos por parte dos dois algoritmos de **clustering** resultados muito próximos.

Os melhores resultados foram obtidos por parte das RNA quando os algoritmos de **clustering** foram aplicados aos dados de entrada, tendo a RNA treinada com **Backpropagation** e com 29 variáveis de entrada alcançado uma taxa de 100% de sucesso na tarefa de classificação.

2.4.8 Predicting the Burned Area in Forest using Machine Learning Techniques

Neste artigo Ramasubramanian (2017) [46] considerou os dados utilizados no estudo de Cortez et al (2007) [44].

Na análise do conjunto de dados o autor decidiu realizar algumas alterações ao conjunto de dados para que fosse posteriormente mais fácil a implementação dos algoritmos de **Aprendizagem Computacional** considerados no estudo.

O autor começou assim por transformar as duas variáveis categóricas (dia e mês) em variáveis numéricas.

Posto isto o autor procedeu à criação de um mapa de calor para recolher informação relativa à área queimada de acordo com os dados considerados, onde o esquema de cores diferencia a intensidade da área queimada.

Devido à grande existência de dados com o valor 0 no que diz respeito à variável de saída, área queimada, o autor decidiu aplicar uma transformação logarítmica para reduzir a assimetria e fazer com que a área queimada fosse normalmente distribuída, para posteriormente

ser utilizada pelos algoritmos de **Aprendizagem Computacional**.

Antes de proceder à implementação dos diferentes algoritmos o autor fez uma análise para verificar quais as variáveis que possuíam maior impacto na variável de saída e reduzir assim o tempo de treino dos algoritmos ao eliminar variáveis que possuem menos impacto na variável de saída. Este processo foi feito com recurso do algoritmo **Random Forest**.

Relativamente à parte experimental o autor considerou 70% dos dados para treino e os restantes 30% para teste. Tanto os dados de treino como os dados de teste sofreram um processo de padronização por forma a que todas as variáveis consideradas tivessem a mesma ordem de grandeza para serem utilizadas pelos diferentes modelos de **Aprendizagem Computacional**.

Como forma de solução para o problema no estudo em questão o autor implementou algoritmos de dois tipos, algoritmos de regressão e algoritmos de classificação.

Os algoritmos considerados para regressão foram respectivamente **Regressão linear**, **Regressão logística** e **SVM** com funções de **Kernel RBF** e **Poly**.

Para classificação os algoritmos escolhidos foram **Bagging**, **Random Forest** e **Gradient Boosting**.

Estes algoritmos foram auxiliados por uma matriz de confusão por forma a avaliar a **precisão** de cada um dos mesmos.

De entre os algoritmos considerados os resultados experimentais demonstraram que **Random Forest** apresentou os melhores resultados com uma **precisão** de 95% relativamente a treino e 56% relativamente a teste. No entanto **Gradient Boosting** apresentou resultados também muito próximos, com **precisão** de 96% relativamente a treino e 54% relativamente a teste.

2.4.9 Assessing the Suitability of Soft Computing Approaches for Forest Fires Prediction

Neste artigo _Janabi et al (2017) [48] consideraram os dados utilizados no estudo de Cortez et al (2007) [44].

Antes de procederem à implementação dos algoritmos de **Aprendizagem Computacional** considerados os autores realizaram previamente um processamento do conjunto de dados.

Os autores começaram por converter as variáveis categóricas dia e mês em variáveis numéricas e normalizar estas mesmas variáveis.

Posto isto os autores procederam à verificação da correlação das diferentes variáveis presentes no conjunto de dados e verificaram que todas as variáveis apresentavam uma forte correlação entre si excepto a **Humidade relativa**. Assim, os autores eliminaram esta variável no momento de agrupar os dados através de **Particle swarm optimization**.

Tal facto levou a que o custo computacional e o erro gerado aumentasse ao nível dos algoritmos de **Aprendizagem Computacional** e a que os autores aplicassem o método **Principal component analysis** para a selecção das variáveis com maior correlação do conjunto de dados.

Com os resultados obtidos por parte deste método e com o intuito de melhorar a **precisão** todas as variáveis foram consideradas para utilização devido à forte correlação das mesmas

no conjunto de dados.

Após as variáveis a serem utilizadas estarem definidas, os autores utilizaram o método **Particle swarm optimization** para dividir os dados em diferentes **clusters** de acordo com o grau de semelhança, para que dados presentes em cada um destes fossem o mais semelhante possível.

Para obter as distâncias entre os **clusters** foram atribuídos pesos a cada um dos mesmos e calculada a distância euclidiana.

No que diz respeito à parte experimental o desempenho de cada algoritmo foi medido de acordo com cinco medidas de qualidade, **RMSE**, **MSE**, **RAE**, **MAE** e **IG**.

O conjunto de dados foi dividido em 50% para treino e os restantes 50% para teste.

Os algoritmos escolhidos pelos autores neste estudo foram **RNA** e **SVM**.

Dentro das RNA as escolhidas foram **Radial basis function**, **Multilayer perceptron**, **Polynomial** e **Cascade correlation**.

Os resultados obtidos pelos diferentes algoritmos escolhidos pelos autores demonstraram que a **SVM** apresenta melhores resultados no que diz respeito à previsão de fogos florestais quando comparada com os outros algoritmos.

A **SVM** implementada apresentou o menor **RMSE** (54,0), **MSE** (2926,4), **RAE** (10,5) e **MAE** (2,656) bem como o **IG** mais alto na fase de teste (2,656).

2.4.10 FireCast: Leveraging Deep Learning to Predict Wildfire Spread

Neste artigo Radke et al (2019) [14] desenvolveram um estudo ao nível da previsão de fogos florestais, nomeadamente no que diz respeito à propagação de fogos florestais.

Para combater o problema apresentado os autores propuseram um novo sistema denominado **Firecast** que segue uma abordagem baseada em técnicas de **Aprendizagem Computacional** e em dados recolhidos de Geographic Information System (GIS).

Os autores propuseram uma abordagem diferente das já existentes, uma abordagem baseada em **Deep learning**.

Para tal implementaram uma **Convolutional Neural Network (CNN) 2D** composta por duas camadas convolucionais de 32 e 64 nós ocultos que utilizam as funções de activação **Sigmoid** e **ReLU** respectivamente.

A janela deslizante implementada explora uma área quadrada de 30 pixels à volta de cada pixel de interesse de uma camada visual, com um kernel de tamanho 3x3.

O tensor de saída da CNN é conectado com os dados atmosféricos do local e utilizado como entrada para uma camada densa com função de activação **Sigmoid** que atribui a cada pixel um único valor de saída.

Os autores desenvolveram FireCast de forma a que este fosse treinado para prever áreas circundantes de um perímetro de incêndio inicial no período de 24 horas e tendo em consideração o perímetro inicial, características de localização e dados meteorológicos de entrada.

O modelo desenvolvido recolhe aleatoriamente pixeis de interesse à volta do perímetro de incêndio e atribui um valor de previsão a cada pixel compreendido no intervalo [0,1], onde o valor atribuído denota a probabilidade de o modelo prever que o pixel em questão se

encontra no perímetro de incêndio.

Para testar o modelo desenvolvido os autores decidiram considerar um incêndio histórico com perímetros mapeados consecutivamente para teste, incêndio este que não foi incluído na parte de treino do modelo.

Por forma a avaliar os resultados de teste os autores decidiram considerar a **precisão** total da previsão, **precisão** da área queimada e **F-score**.

Para além destas 3 medidas de avaliação os autores compararam também os resultados de teste obtidos visualmente com os perímetros reais do incêndio de teste.

Os resultados experimentais obtidos foram comparados com um modelo de previsão aleatória e com o modelo de propagação de fogos florestais Farsite.

A comparação dos resultados demonstrou que o modelo desenvolvido pelos autores apresentava um desempenho superior no que diz respeito à **precisão** total da previsão, **precisão** da área queimada e **F-score**. O modelo desenvolvido pelos autores apresentou resultados de 87,7% no que diz respeito à **precisão** total de previsão, 91,1% no que diz respeito à **precisão** da área queimada e 6,4% no que diz respeito ao **F-score**.

Além da obtenção de melhores resultados, o modelo implementado fornece também informações visuais acerca da expansão do fogo florestal em vez de apresentar apenas dados de fogos florestais passados e fornecer previsões meteorológicas.

A avaliação das informações visuais demonstrou a capacidade do modelo implementado realizar previsões gerais de áreas de alto risco de propagação de um fogo florestal com duas semanas de antecedência.

2.5 Plataformas Web Existentes

A presente secção visa apresentar e analisar as diferentes plataformas web existentes para a prevenção e previsão de fogos florestais.

2.5.1 IPMA

Risco de Incêndio Rural

Na plataforma do IPMA podemos encontrar na secção Fogos Rurais uma sub-secção destinada ao risco de incêndio rural.

Nesta sub-secção é possível ao utilizador, sem a necessidade de qualquer registo e de uma forma totalmente gratuita, verificar a previsão do risco de ocorrência de fogos florestais no território de Portugal Continental para um intervalo de até 5 dias.

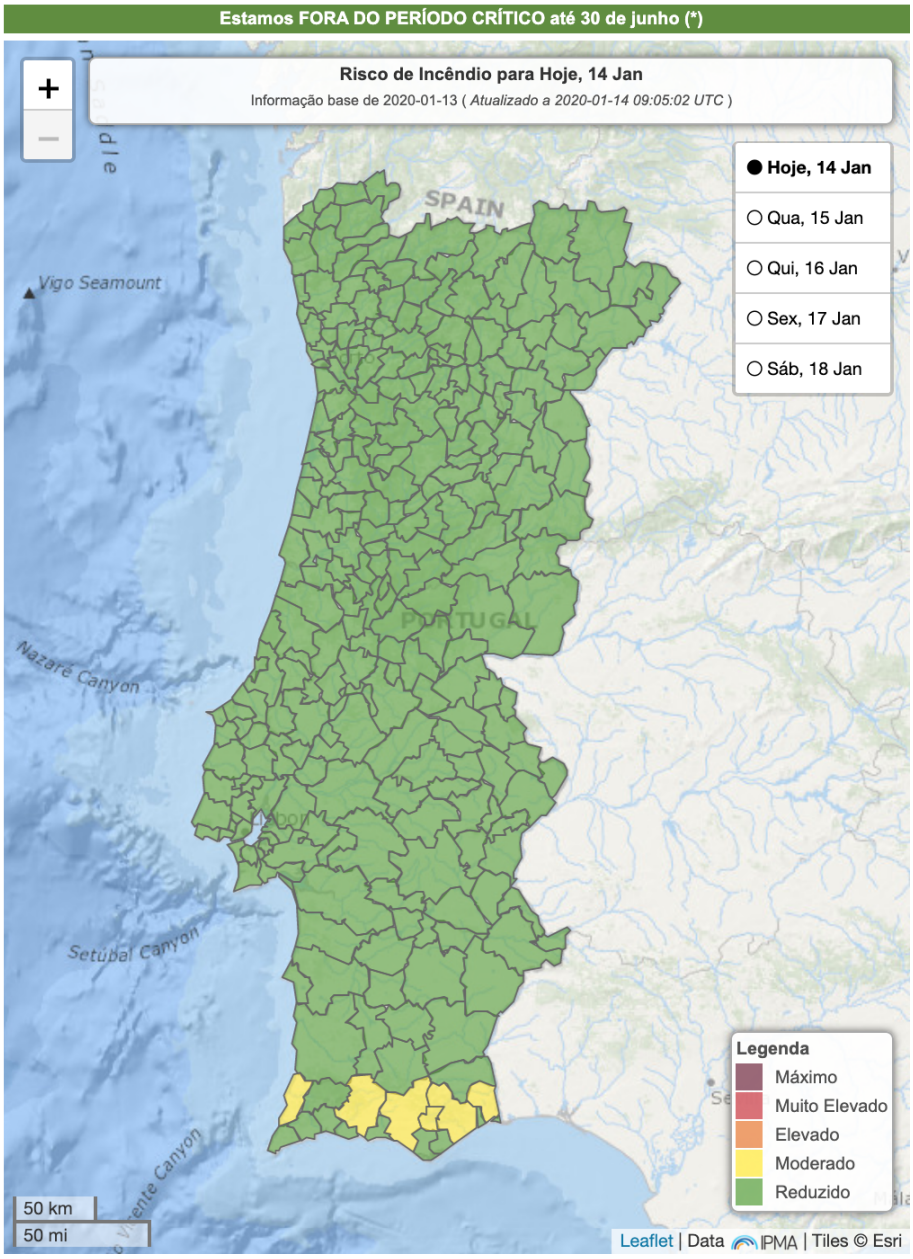


Figura 2.14: Plataforma IPMA - Risco de Incêndio Rural

Ainda nesta secção da plataforma é possível ao utilizador consultar com mais detalhe as condições meteorológicas previstas a nível da temperatura (mínima e máxima), vento, humidade relativa (mínima e máxima) e precipitação para um determinado Distrito e Concelho num dado dia.

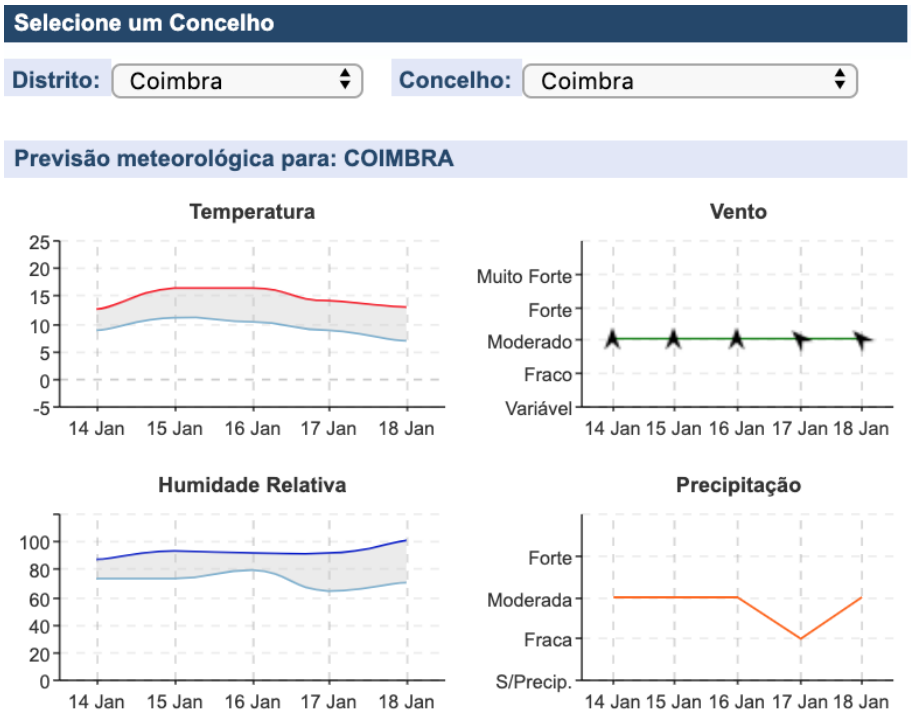


Figura 2.15: Plataforma IPMA - Condições meteorológicas previstas para o concelho de Coimbra

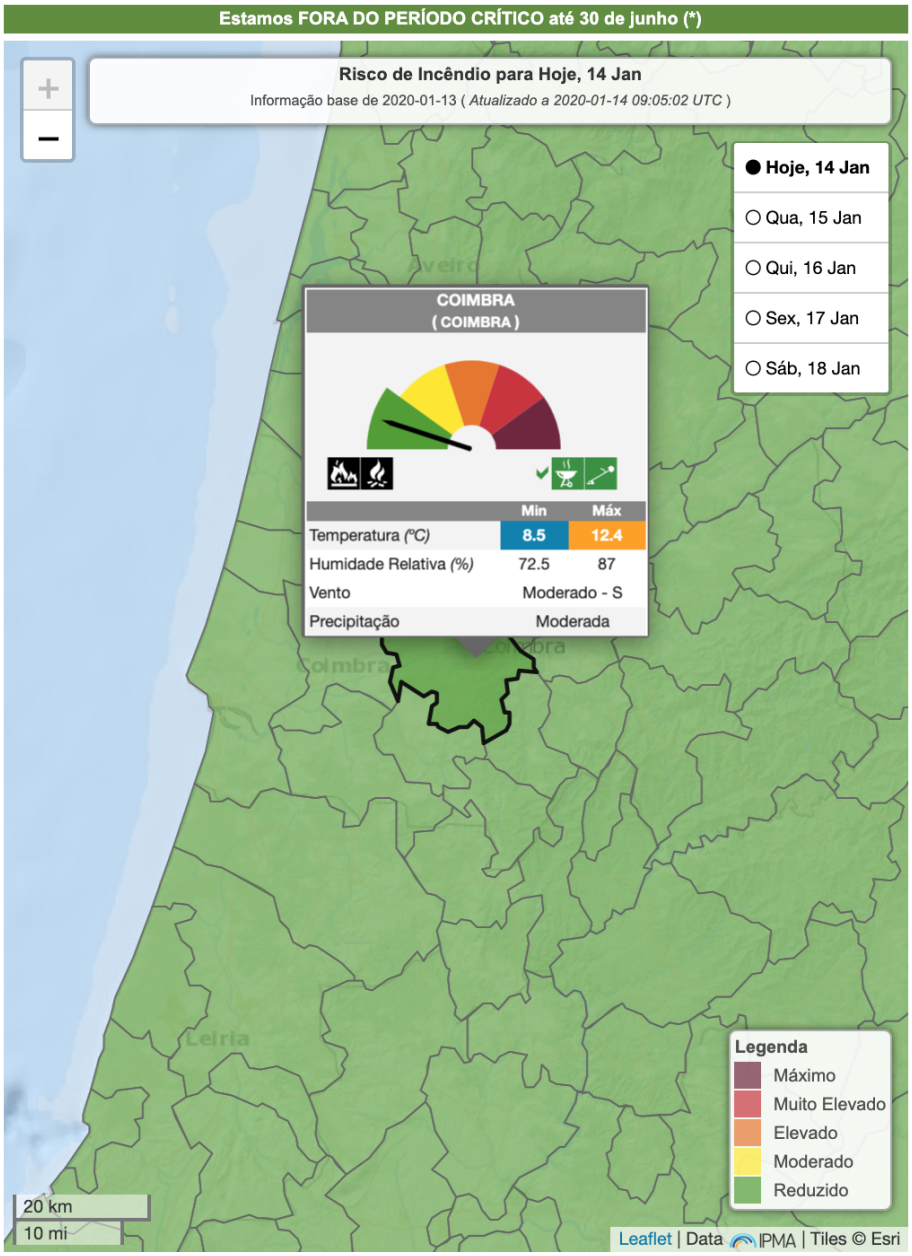


Figura 2.16: Plataforma IPMA - Risco de Incêndio Rural do Concelho de Coimbra

FWI, índice de perigo de incêndio

A plataforma do IPMA permite disponibiliza também na secção Fogos Rurais uma sub-secção destinada ao índice de perigo de incêndio FWI. Nesta sub-secção é possível ver ao longo de Portugal Continental o índice de perigo de incêndio rural de acordo com o sistema Canadian Forest Fire Weather Index System (FWI).

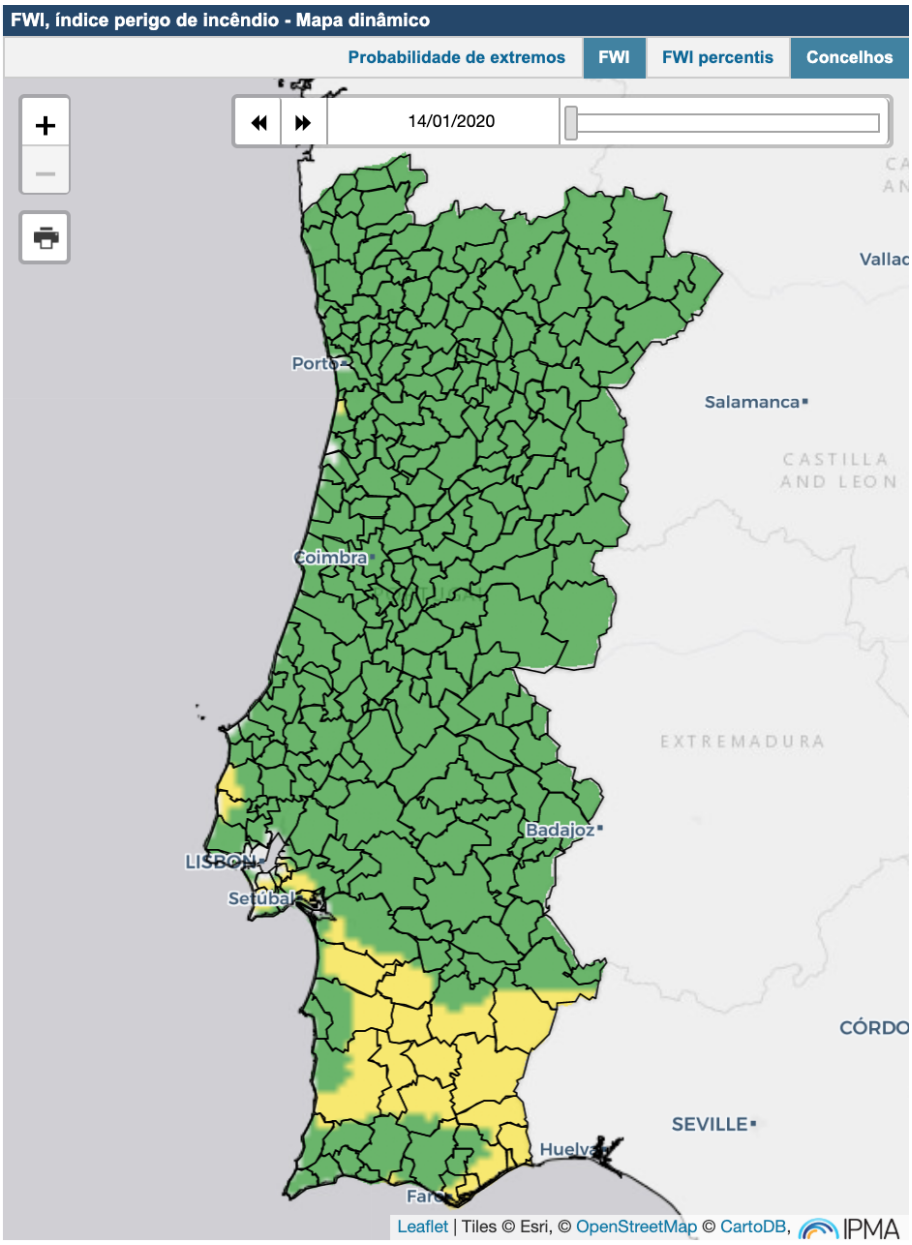


Figura 2.17: Plataforma IPMA - FWI, índice de perigo de incêndio

2.5.2 Fogos.pt

A plataforma **Fogos.pt** [2] é uma plataforma gratuita sem a necessidade de qualquer registo que oferece a possibilidade de consultar os fogos que se encontram activos bem como o seu estado de actividade. É também possível consultar o registo histórico de fogos já finalizados embora apresente um registo histórico relativamente curto.

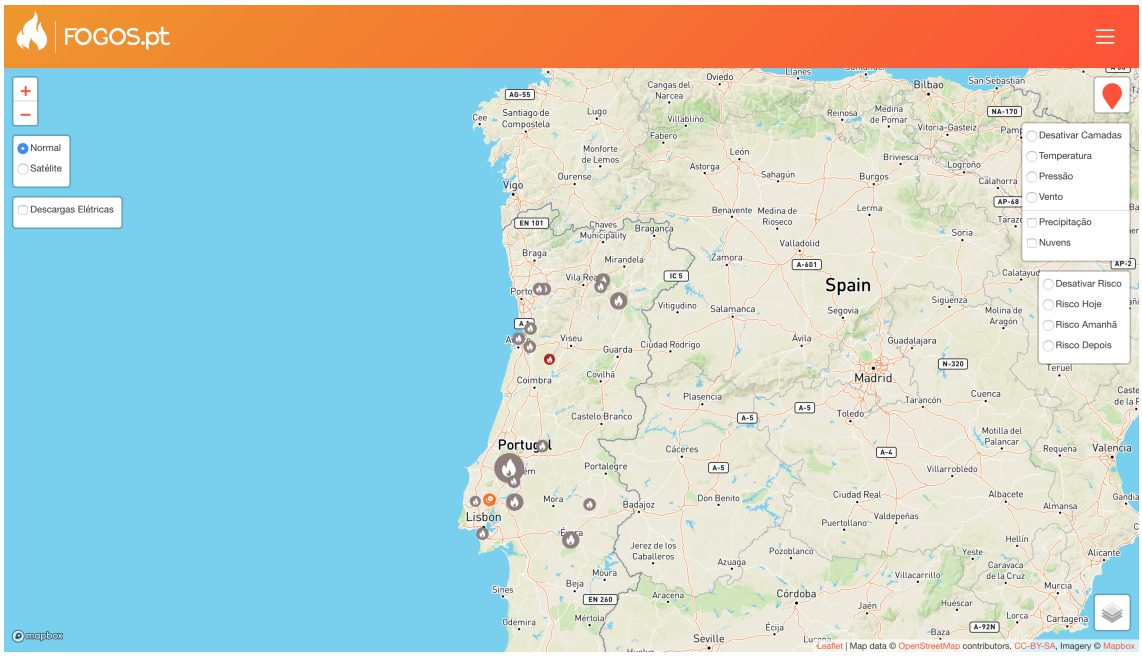


Figura 2.18: Plataforma Fogos

Para consultarmos os detalhes relativos a um fogo activo nesta plataforma devemos clicar sobre ele no mapa. De seguida são assim apresentadas as informações relativas ao fogo em questão.

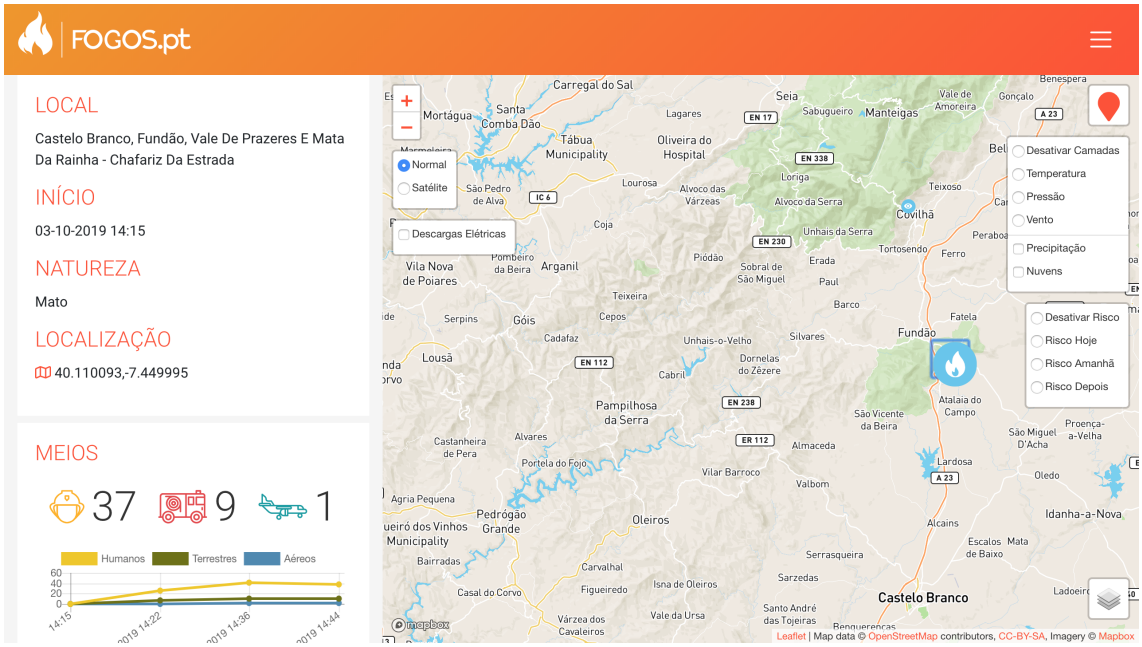


Figura 2.19: Detalhes de um fogo activo

Ao pressionarmos o botão que se encontra no canto superior direito e de seguida pressionarmos o botão **Estatísticas** podemos consultar dados estatísticos passados relativos ao dia anterior apenas.

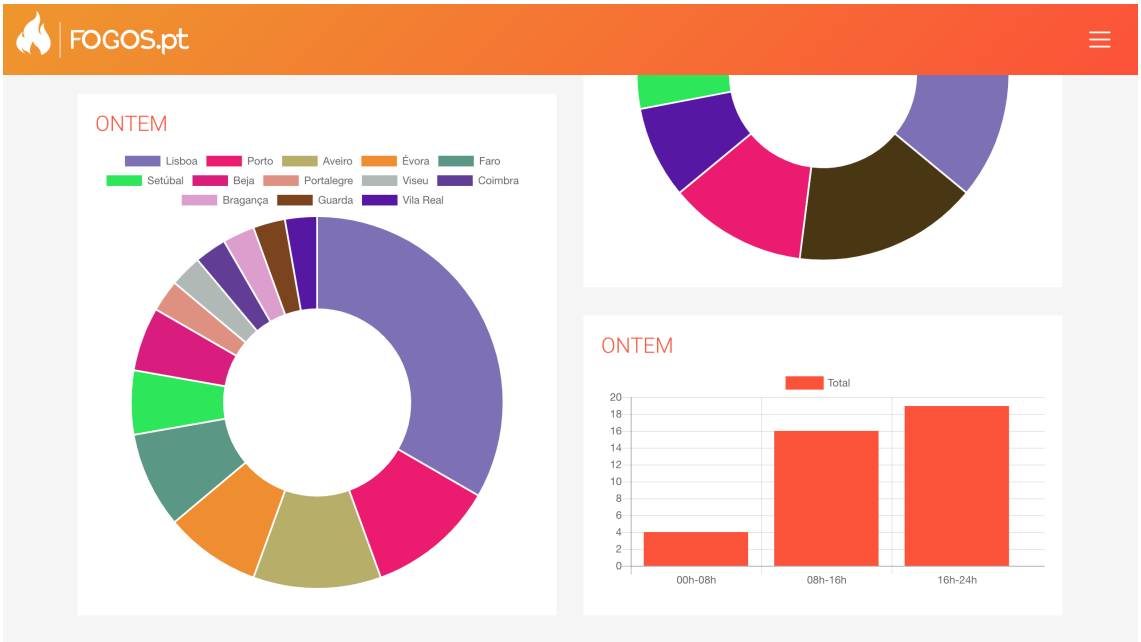


Figura 2.20: Estatísticas dia anterior

2.5.3 EFFIS

A plataforma **EFFIS** [1] é também uma plataforma gratuita isenta de qualquer registo.

Esta plataforma oferece já mais funcionalidades do que **Fogos.pt** [2], nomeadamente no que diz respeito à previsão de fogos. É possível realizar a previsão da ocorrência de fogos florestais com uma margem de 6 dias. Ao nível de registo histórico de fogos já finalizados permite a consulta de fogos que já se encontram finalizados há um grande período de tempo.

Quanto à sua organização, a plataforma encontra-se dividida em duas partes.

Na primeira parte intitulada de **Situação actual** podemos acompanhar a actividade de incêndios florestais em tempo quase real bem como a previsão da ocorrência dos mesmos.

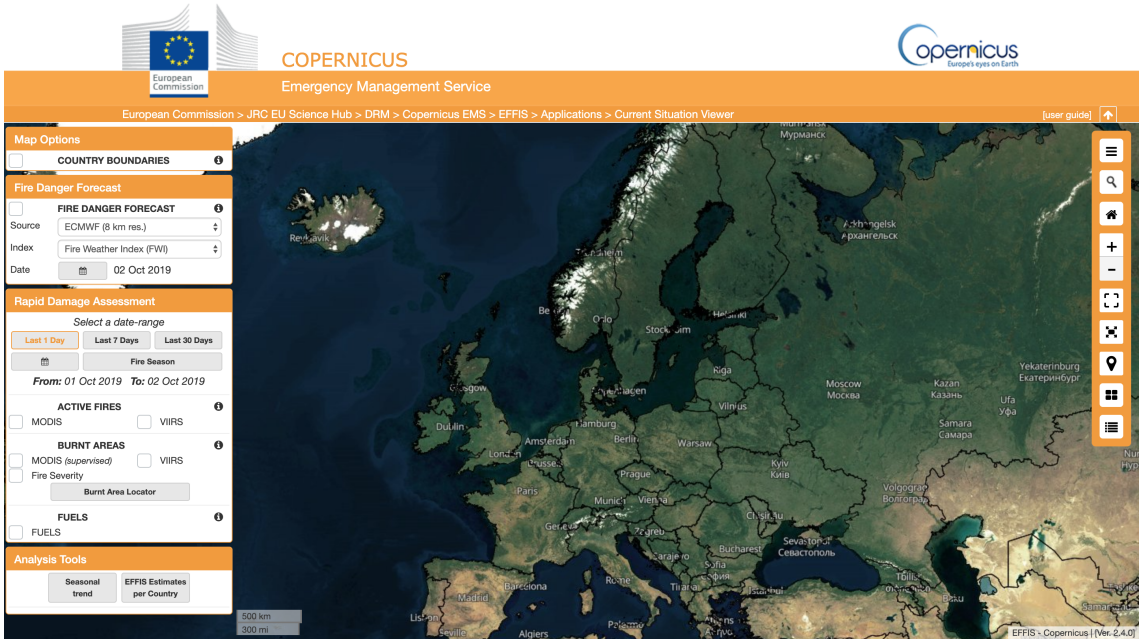


Figura 2.21: Dashboard situação atual plataforma EFFIS

Ainda na primeira parte existe uma funcionalidade que nos permite consultar dados estatísticos relativos ao número de fogos e à dimensão de área queimada dos países em questão desde 2008. Para visualizarmos tal funcionalidade devemos selecionar o botão **EFFIS Estimates per Country** que se encontra no canto inferior esquerdo dentro da componente **Analysis Tools**.

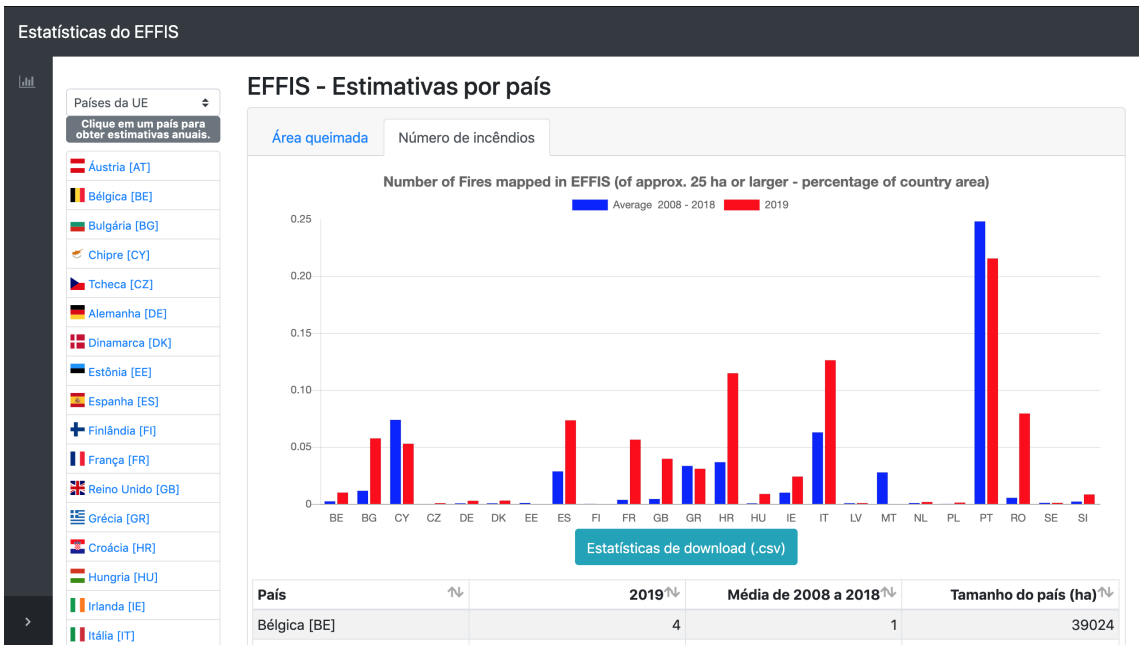


Figura 2.22: Estatísticas EFFIS relativas ao número de incêndios por país

A segunda parte da plataforma, intitulada de **Notícias sobre incêndio**, disponibiliza um vasto conjunto de informações históricas relativas a fogos florestais que podem ser pesquisadas para países específicos ou até mesmo cidades.

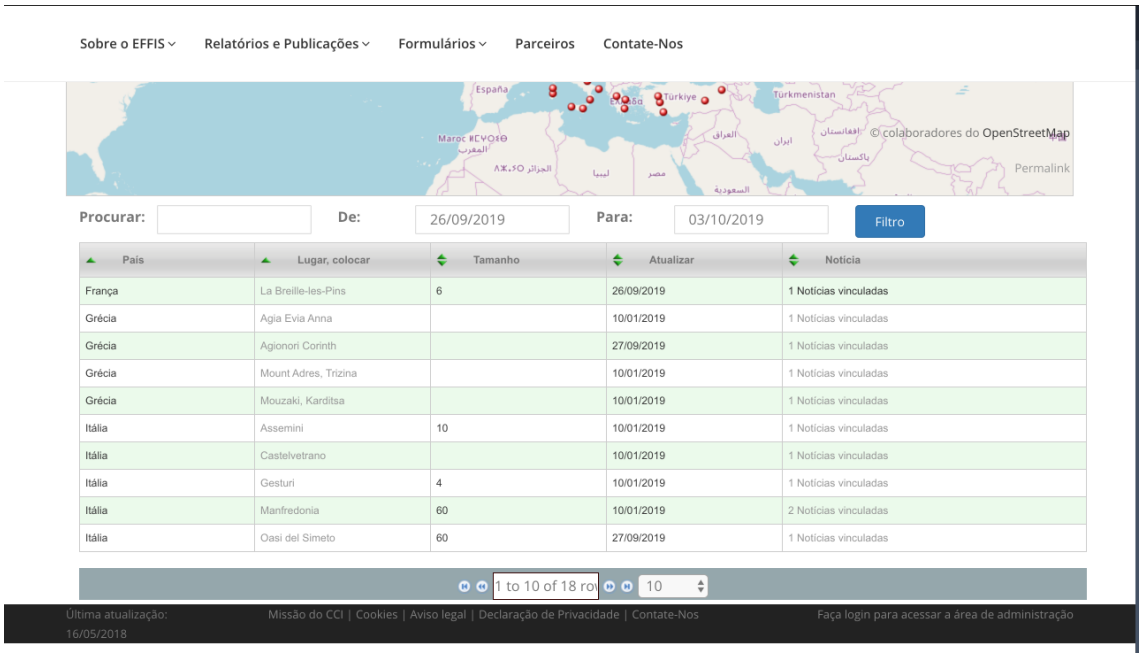


Figura 2.23: Dashboard notícias sobre incêndio plataforma EFFIS

2.5.4 GWIS

A plataforma **GWIS** [3] apresenta na generalidade grandes semelhanças à plataforma **EFFIS** [1] (2.4.2), destacando-se no aspecto em que cobre todos os continentes e não exclusivamente a Europa, Oriente Médio e Norte de África. Tal como a plataforma **EFFIS** também a plataforma **GWIS** é totalmente gratuita e acessível sem a necessidade de qualquer tipo de registo. Oferece também a possibilidade de previsão de fogos florestais com uma margem de 6 dias bem como de um vasto registo histórico de fogos florestais já finalizados.

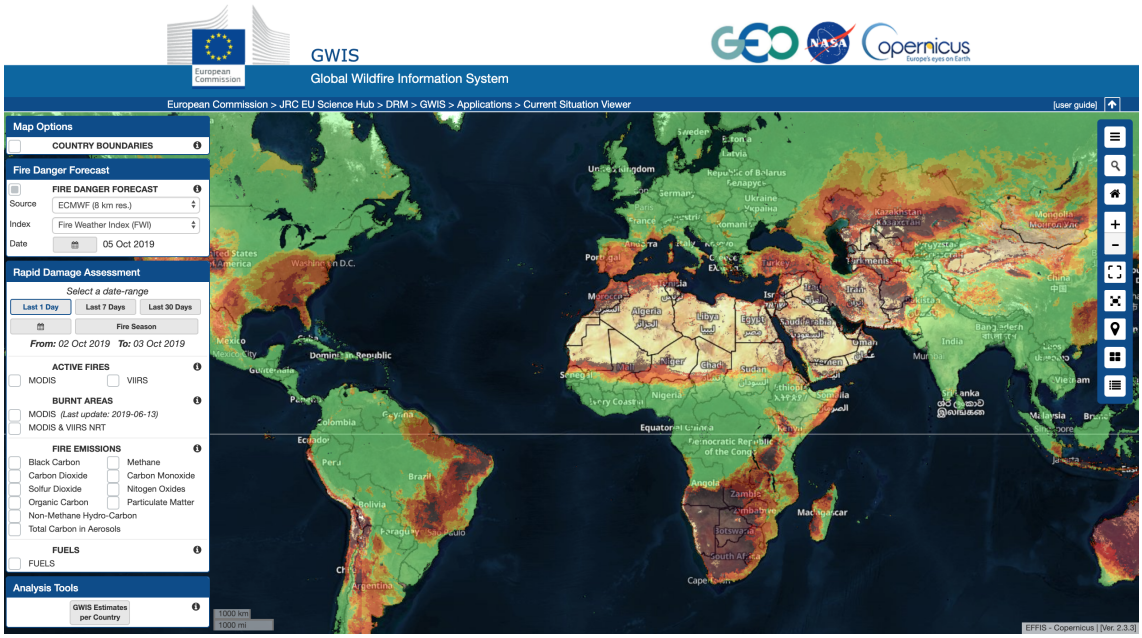


Figura 2.24: Dashboard plataforma GWIS

Para consultarmos dados estatísticos relativos ao número de fogos e à dimensão de área queimada dos países em questão desde 2001 até 2017 devemos selecionar o botão **GWIS Estimates per Country** que se encontra no canto inferior esquerdo dentro da componente **Analysis Tools**.

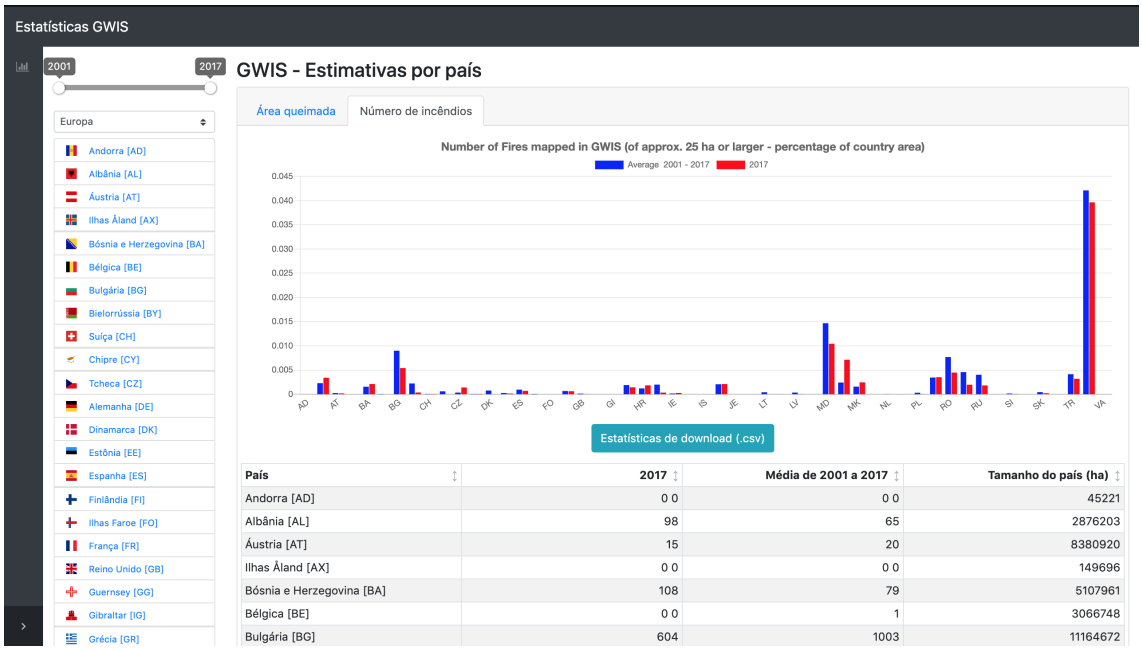


Figura 2.25: Estatísticas GWIS relativas ao número de incêndios por país na Europa

Capítulo 3

Especificações

O presente capítulo tem por objectivo apresentar as especificações do projecto em questão, nomeadamente no que diz respeito à metodologia de desenvolvimento de software utilizada, aos requisitos funcionais e não funcionais definidos, à análise de riscos associados, à arquitectura do sistema adoptada, às tecnologias e ferramentas utilizadas e ao plano de trabalho definido para o desenvolvimento do projecto em questão.

3.1 Metodologia de desenvolvimento de software

A realização deste projecto irá ter por base uma metodologia ágil fortemente baseada em **Scrum**. Por forma a perceber melhor o funcionamento desta metodologia de seguida encontram-se presentes os diferentes conceitos de **Cargos**, **Artefactos** e **Eventos** que caracterizam a metodologia **Scrum**.

3.1.1 Cargos

Na metodologia **Scrum** estão definidos 3 cargos diferentes, **Product Owner**, **Scrum Master** e **Scrum Team**. De modo a compreender melhor o papel e a importância de cada um destes cargos de seguida encontram-se explicados cada um dos mesmos.

Product Owner

Representa o cliente que define o que pretende para o produto através de **User Stories**.

É o responsável por transmitir a sua visão do produto por forma a que este alcance o maior valor possível.

É também responsável por atribuir os diferentes níveis de prioridade no **Product Backlog** e definir o conteúdo e a data de conclusão dos diferentes **Sprint Backlog**.

No projecto em questão o **Product Owner** é o Engenheiro Rafael Maia da empresa Grama [27].

Scrum Master

Representa o elemento da equipa que detém um grande conhecimento da metodologia **Scrum** e é responsável por ajudar os restantes elementos da equipa no desenvolvimento do produto de acordo com esta mesma metodologia. É também o responsável por acompanhar todo o trabalho da equipa e assegurar os recursos necessários para a implementação do produto.

No projecto em questão o **Scrum Master** é o Engenheiro Miguel Oliveira.

Development Team

Representa a equipa responsável por implementar o produto em questão.

No projecto em questão a equipa é unicamente composta por mim, Gonçalo Pinto.

3.1.2 Artefactos

Na metodologia **Scrum** de entre os artefactos existentes estão definidos 3 para o desenvolvimento deste projecto, **Product Backlog**, **Sprint Backlog** e **Increment**. De modo a compreender melhor o papel e a importância de cada um destes artefactos de seguida encontram-se explicados cada um dos mesmos.

Poduct Backlog

Representa a lista de todas as tarefas necessárias no desenvolvimento do produto, uma lista de requisitos do produto. Esta lista pode ser alterada durante o desenvolvimento do produto e, como já referido anteriormente, é da responsabilidade do **Product Owner**.

Sprint Backlog

Representa um sub-conjunto de tarefas do **Product Backlog** seleccionados para uma **Sprint**.

Increment

Representa o estado atual do produto, todas as tarefas do **Product Backlog** concluídas até ao **Sprint** que se encontra em desenvolvimento. O **Increment** é actualizado sempre que um **Sprint** termina.

3.1.3 Eventos

Na metodologia **Scrum** estão definidos 5 eventos, **Sprint**, **Sprint Planning**, **Sprint Review**, **Sprint Retrospective** e **Daily Scrum**. De modo a compreender melhor o

papel e a importância de cada um destes eventos de seguida encontram-se explicados cada um dos mesmos.

Sprint

Representa o intervalo de tempo para a implementação de um **Sprint Backlog**. No fim de um **Sprint** todas as tarefas presentes no **Sprint Backlog** correspondente devem estar concluídas.

A duração de um **Sprint** varia por norma entre 1 a 3/4 semanas. No desenvolvimento de um produto com base na metodologia **Scrum** é aconselhável que todos os **Sprints** tenham a mesma duração.

Sprint Planning

Ocorre no início de cada **Sprint** e tem por objetivo definir o **Sprint Backlog** e o período de duração para o **Sprint** em questão.

Sprint Review

Ocorre no final de cada **Sprint** e tem por objetivo verificar o **Increment** bem como adaptar o **Product Backlog** em caso de necessidade.

No caso de existirem tarefas incompletas é também planeada a conclusão das mesmas.

Sprint Retrospective

Ocorre no final de cada **Sprint** e tem por objetivo que a equipa realize uma reflexão a respeito do trabalho realizado durante o **Sprint**, o que correu melhor e o que correu pior, e idealizar melhorias a implementar no **Sprint** seguinte para que se alcance o maior sucesso possível.

Daily Scrum

Ocorre no final de cada dia e tem por objectivo cada elemento da equipa expor o que foi fez durante o dia de trabalho bem como apresentar o seu plano de trabalho para o dia seguinte. Esta reunião é planeada por forma a não exceder um tempo de 15 minutos.

Qualquer problema verificado por um elemento da equipa no desenvolvimento do produto devem ser apresentados na **Daily Scrum**.

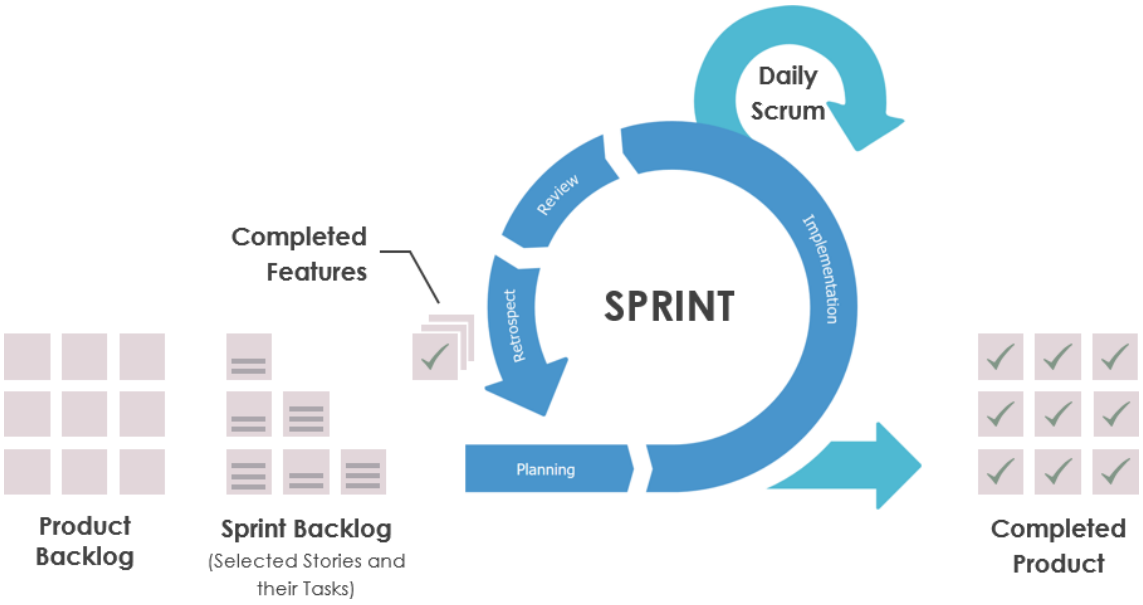


Figura 3.1: Desenvolvimento de um produto com a metodologia Scrum [43]

3.1.4 Vantagens

A utilização da metodologia **Scrum** no desenvolvimento de um produto pode trazer vantagens ao mesmo. Algumas destas vantagens são por exemplo:

- É possível proceder à realização de mudanças a meio do projecto.
- Os intervalos de tempo bem definidos nos diferentes eventos e a gestão de trabalho através destes mesmos eventos permitem que se atinjam níveis de eficiência mais altos.
- Redução de riscos na execução do projecto na medida em que as tarefas são divididas em pequenos períodos de trabalho, sendo possível detectar problemas e proceder à correcção dos mesmos mais facilmente.
- Permite a visualização do progresso do projecto de forma clara na medida em que todas as etapas do produto se encontram definidas e divididas em **Sprints**.
- Prioriza os pontos mais importantes do produto, ou seja as características que oferecem maior valor ao **Product Owner**.

3.2 Requisitos não funcionais

Nesta secção encontram-se apresentados os diferentes requisitos não funcionais considerados no desenvolvimento da aplicação em questão.

- Escalabilidade - o sistema deve conseguir controlar e assegurar a conexão de vários utilizadores em simultâneo
- Usabilidade - o sistema deve ser fácil e intuitivo de utilizar por parte de qualquer tipo de utilizador

- Integridade de dados - o sistema deve garantir a precisão, integridade e consistência dos dados em utilização
- Desempenho - o sistema deve assegurar que um pedido não deve demorar mais que 3 segundos a ser satisfeito

3.3 Requisitos funcionais

Nesta secção serão especificados os diferentes requisitos funcionais deste projecto. Foram elaborados diferentes casos de uso de forma a retratar os diferentes requisitos funcionais.

A alteração do âmbito do projecto devido à não disponibilização dos dados de operadores móveis parte das operadoras telefónicas portuguesas, um dos riscos associados ao desenvolvimento do presente projecto e mencionado na secção 3.4 levou a alteração dos requisitos inicialmente definidos.

Uma vez quer foi apenas alterado o âmbito do presente projecto, que os objectivos definidos permaneceram inalterados, foi apenas necessário alterar o âmbito dos requisitos inicialmente definidos. Deste modo os requisitos que anteriormente se referiam a dados de operadores móveis referem-se agora a dados populacionais.

Denota-se caso de uso por **CU** e excepção por **Ex** respectivamente.

3.3.1 Diagramas de Casos de Uso

Por forma a retratar as funcionalidades que os diferentes tipos de actores possuem na aplicação foram desenhados vários diagramas de casos de uso.

Os diferentes diagramas seguem a notação UML e encontram-se de seguida apresentados.

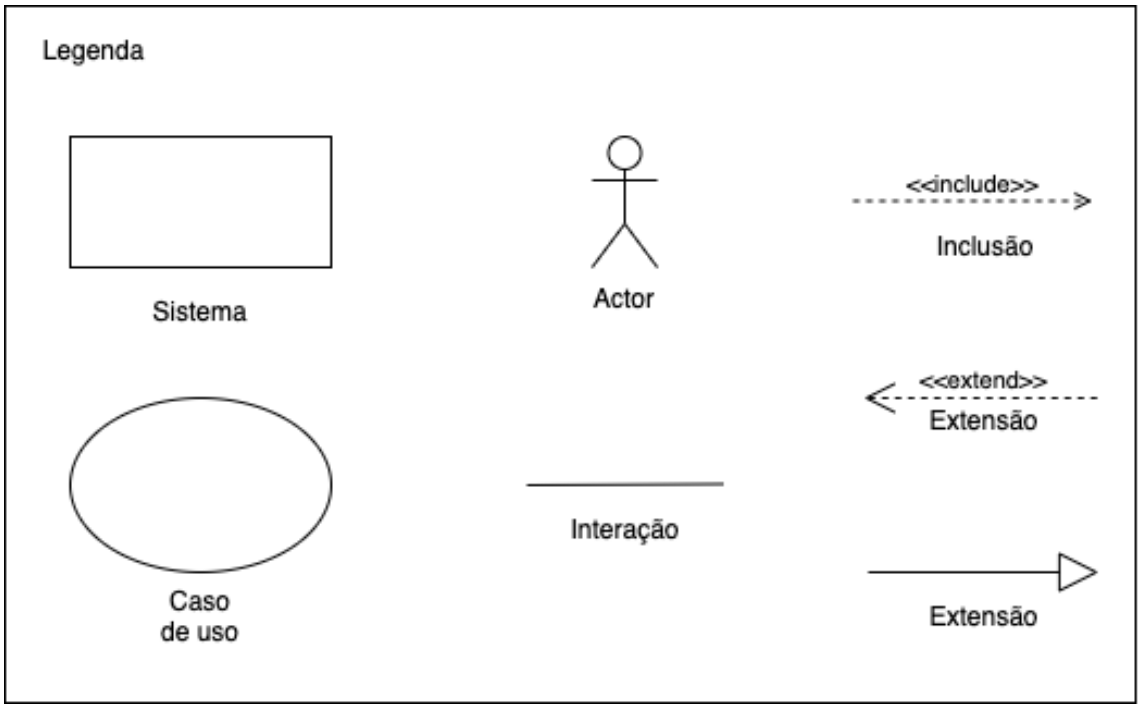


Figura 3.2: Legenda dos componentes dos diagramas

Na figura 3.2 observamos a descrição dos diferentes componentes que constituem os diferentes diagramas de casos de uso apresentados neste documento.

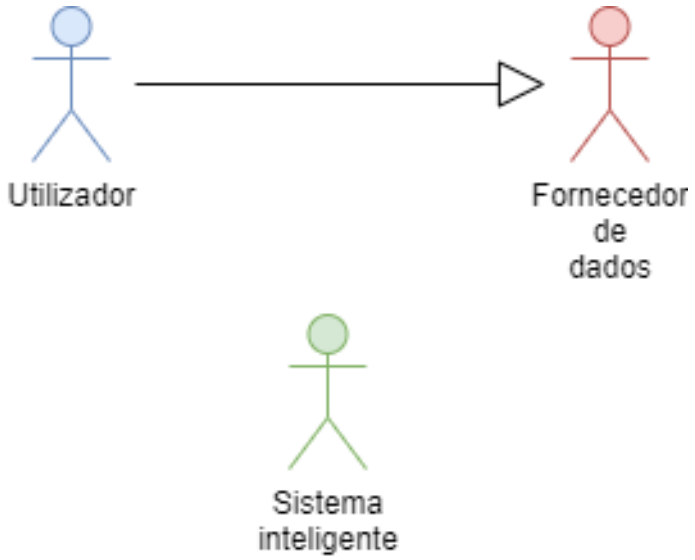


Figura 3.3: Utilizadores da aplicação

Na figura 3.3 encontram-se apresentados os diferentes actores da aplicação em questão.

Por forma a compreender o público alvo representado por estes actores, encontra-se apresentada de seguida uma tabela que descreve cada um dos mesmos.

Utilizador	Pessoa que deseja obter uma previsão para a ocorrência de incêndios florestais, ou consultar dados passados de incêndios florestais
Fornecedor de dados	Pessoa que recolhe dados em diferentes fontes e os insere na aplicação e responsável por garantir a integridade dos dados e correcto funcionamento da aplicação
Sistema inteligente	Sistema responsável por efectuar a previsão da ocorrência de incêndios bem como todo o tratamento e processamento de dados

Tabela 3.1: Actores da aplicação

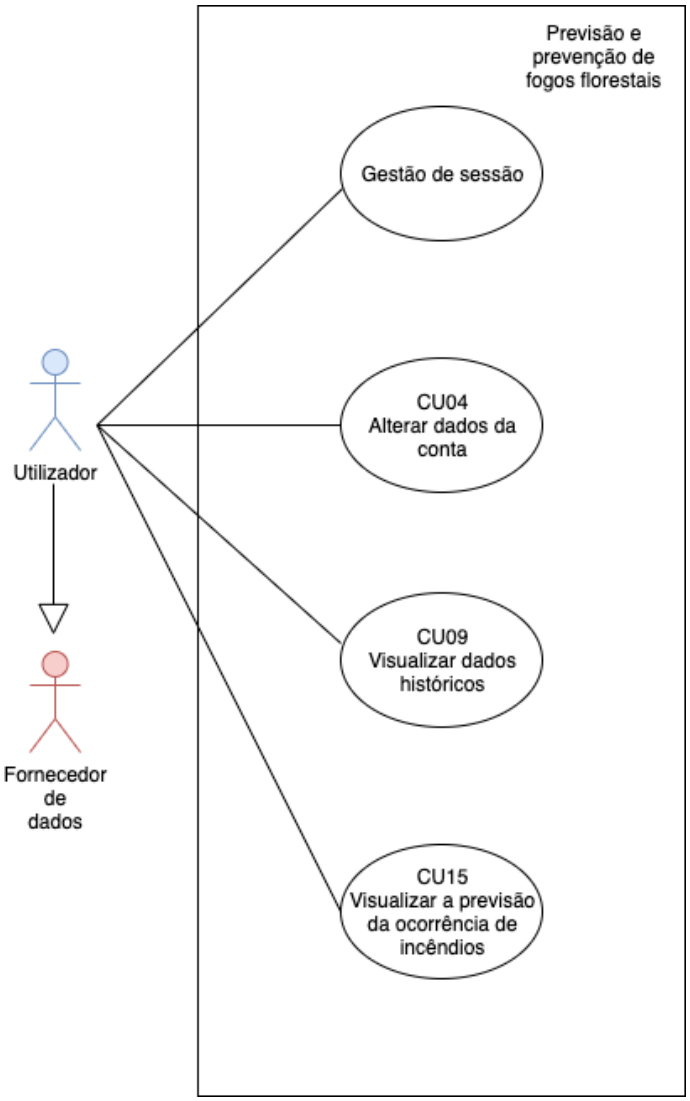


Figura 3.4: Diagrama de alto nível do utilizador

Na figura 3.4 visualizamos o diagrama de alto nível do utilizador.

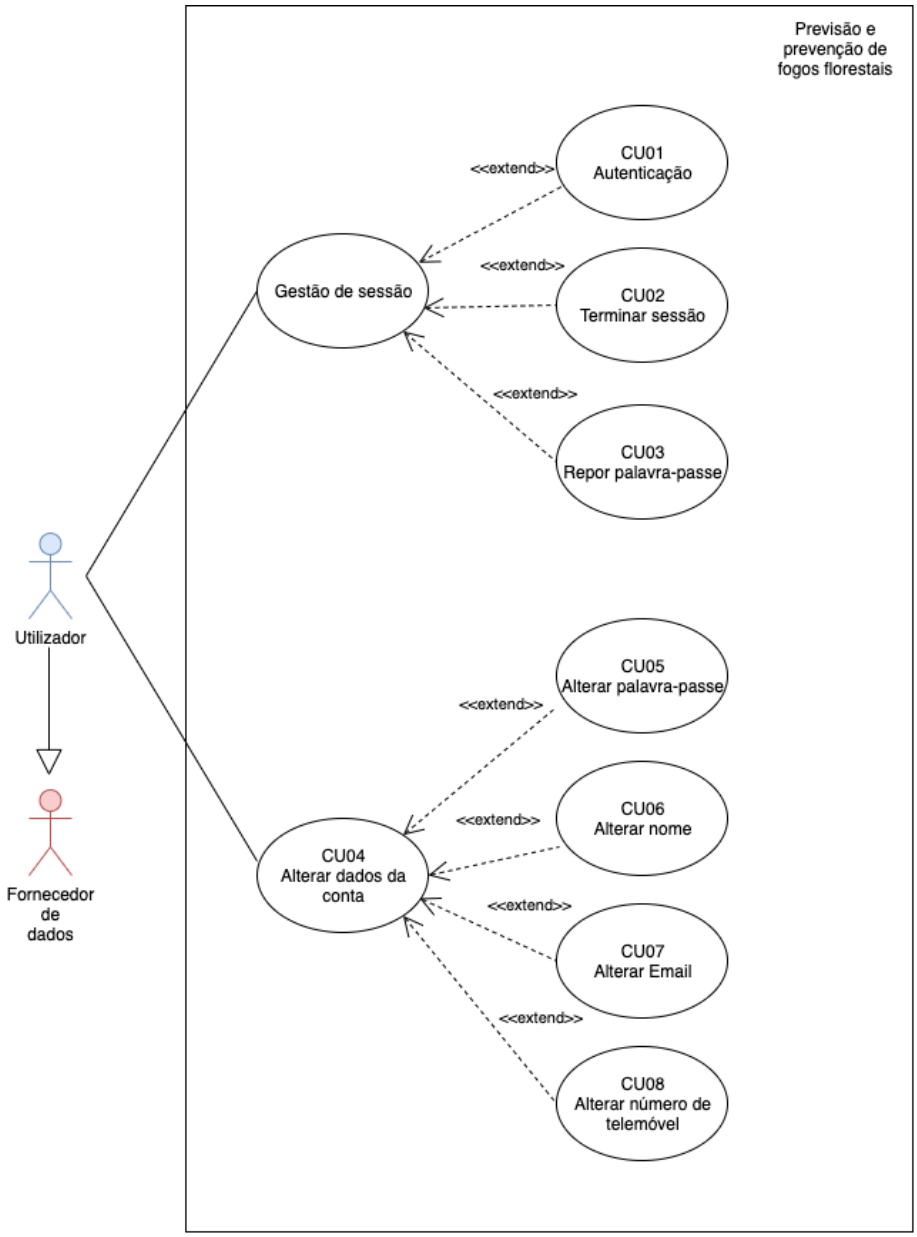


Figura 3.5: Diagrama de funcionalidades do utilizador

Na figura 3.5 são apresentadas as funcionalidades relativas à gestão de sessão e à alteração dos dados da conta. A alteração dos dados da conta só é possível a um utilizador que se encontre autenticado.

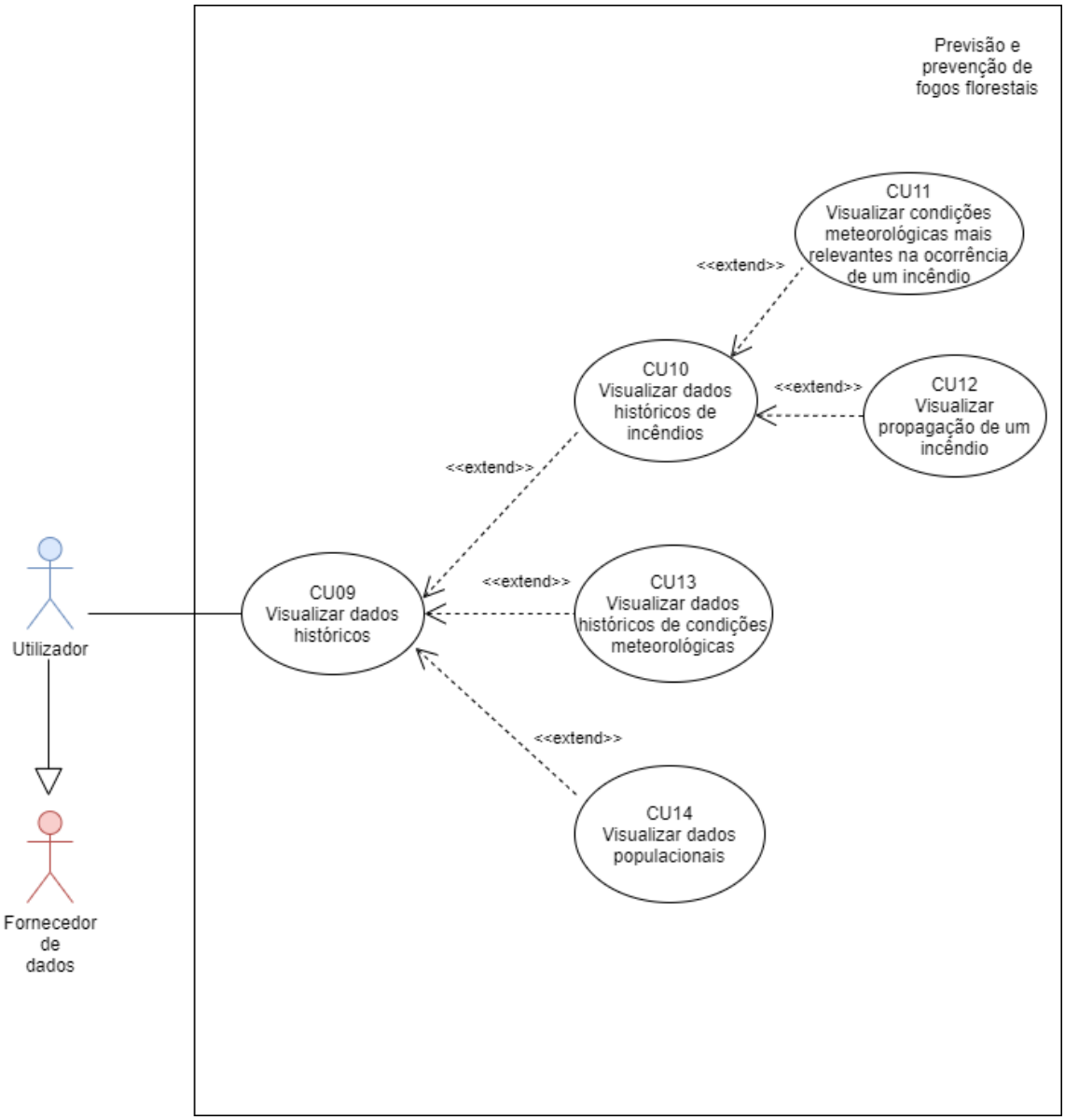


Figura 3.6: Diagrama de funcionalidades do utilizador 2

Na figura 3.6 encontra-se apresentada a funcionalidade de visualização de dados históricos, possível a um utilizador que se encontre autenticado.

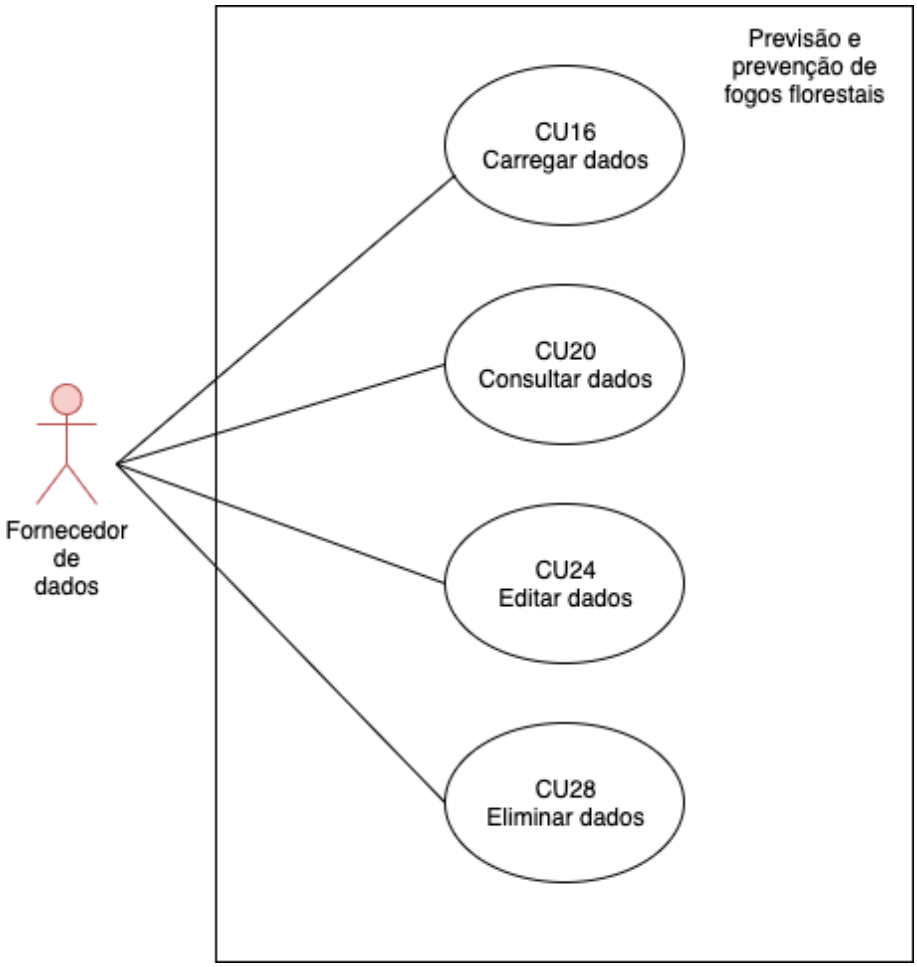


Figura 3.7: Diagrama de alto nível do fornecedor de dados

Na figura 3.7 observamos o diagrama de alto nível do fornecedor de dados.

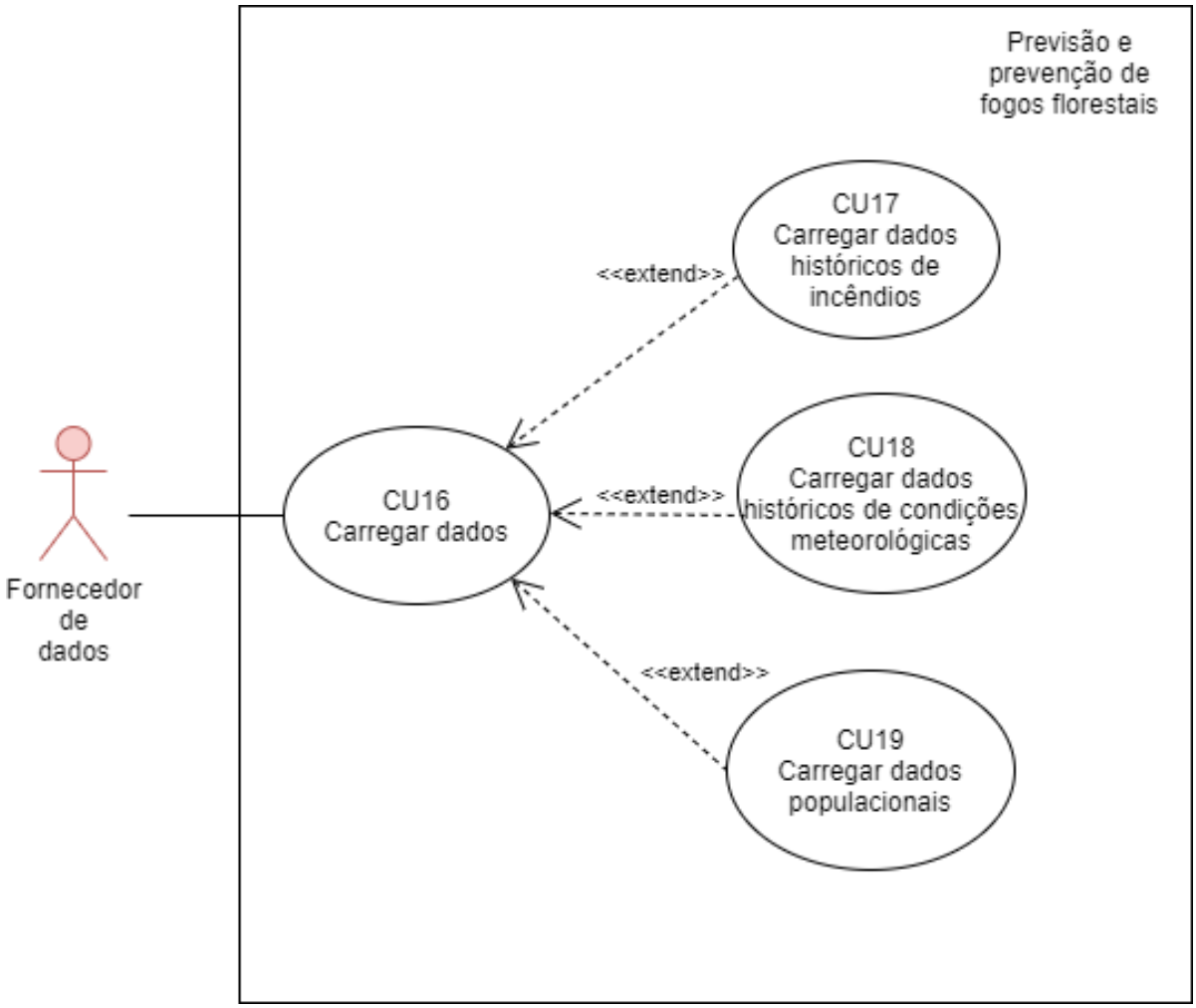


Figura 3.8: Diagrama de funcionalidades do fornecedor de dados

Na figura 3.8 observamos a funcionalidade de carregar dados que um fornecedor de dados pode utilizar quando autenticado na aplicação.

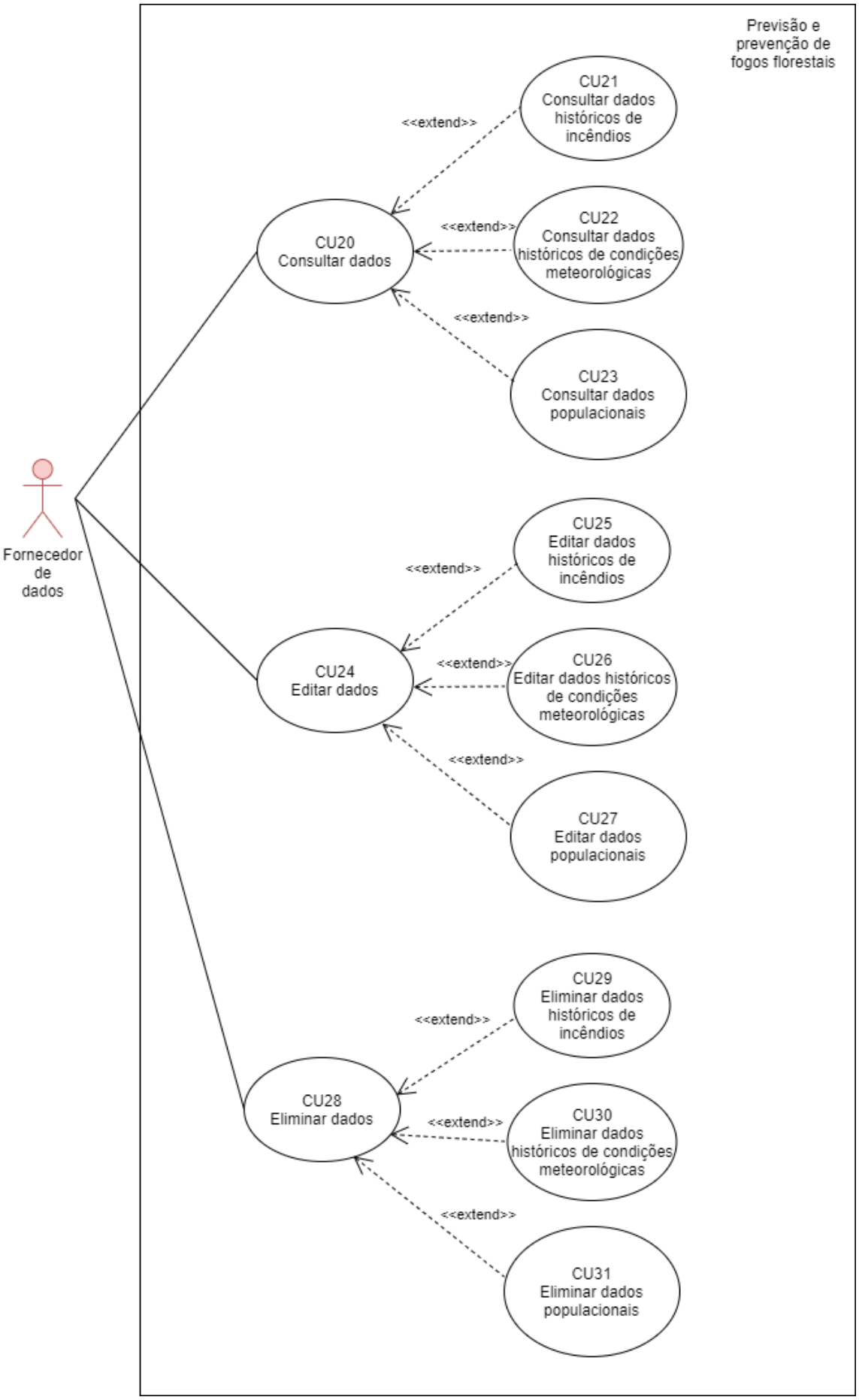


Figura 3.9: Diagrama de funcionalidades do fornecedor de dados 2

Na figura 3.9 encontram-se apresentadas as funcionalidades de consultar, editar e eliminar dados que um fornecedor de dados pode utilizar quando autenticado na aplicação.

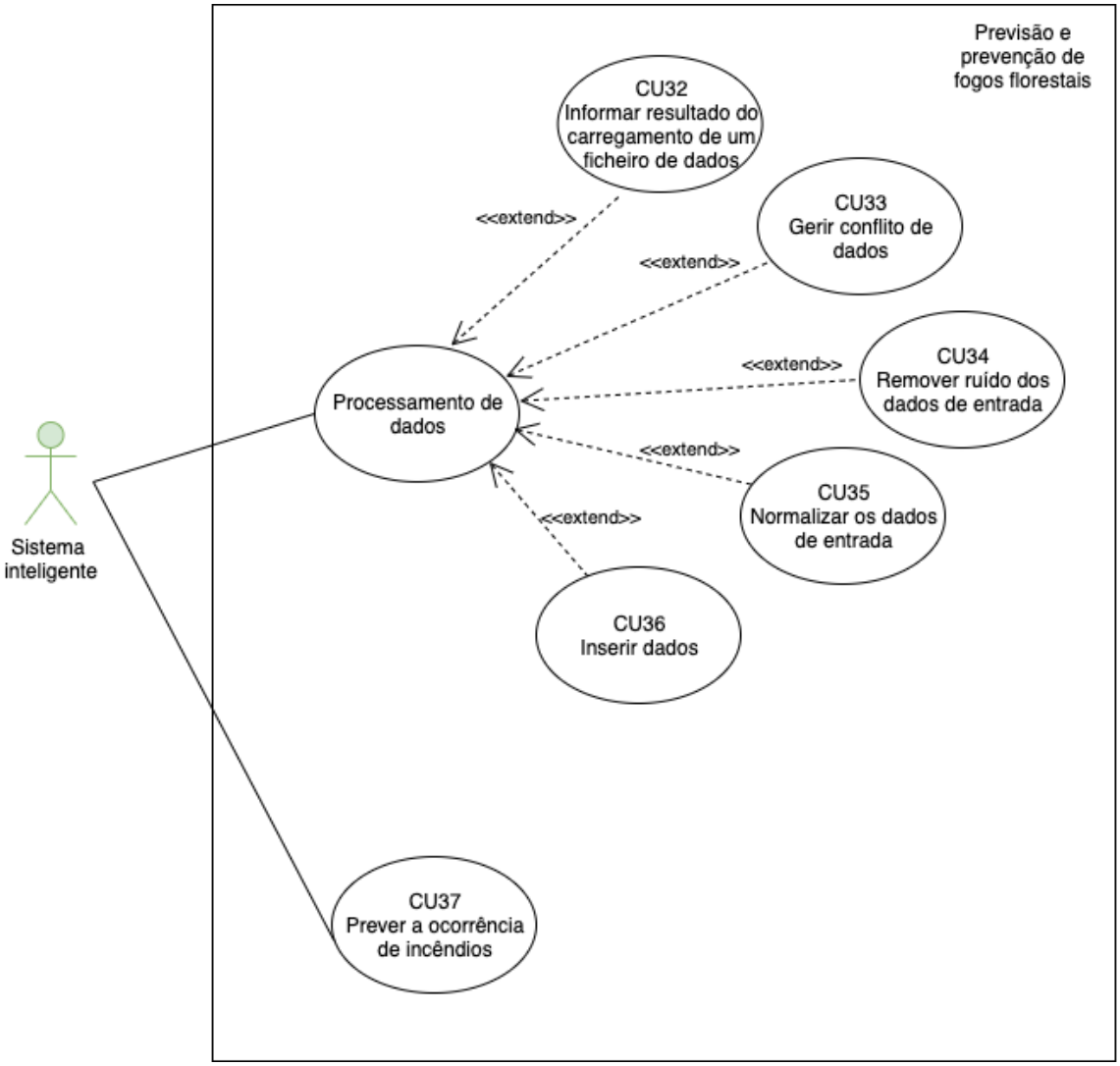


Figura 3.10: Diagrama de funcionalidades do sistema inteligente

Na figura 3.10 observamos as diferentes funcionalidades do sistema inteligente.

3.3.2 Casos de Uso

Por forma a retratar os diferentes requisitos funcionais da aplicação em questão de uma forma mais clara e de uma interpretação mais acessível, foram desenvolvidos casos de uso para cada um dos diferentes requisitos.

A prioridade de cada caso de uso foi definida pelo cliente, a empresa Grama [27]. A sigla NA denota não aplicável.

Os diferentes casos de uso encontram-se de seguida apresentados.

CU01 - Autenticação

Prioridade	Média
Descrição	O utilizador autentica-se na aplicação com os dados da sua conta
Actores	Utilizador
Pré-Condições	O utilizador tem que ter uma conta válida
Fluxo Básico	O utilizador insere o nome de utilizador O utilizador insere a palavra-passe O utilizador clica em "Login"
Pós-Condições	O utilizador é autenticado com sucesso
Dependências	NA

Tabela 3.2: CU01 - Autenticação

EX01 - Conta inválida

Fluxo de eventos	No passo 2 do FB o utilizador carrega em "Login" O sistema não reconhece os dados do utilizador
Pós-Condições	O sistema não consegue autenticar o utilizador

Tabela 3.3: EX01 - Conta inválida

CU02 - Terminar sessão

Prioridade	Baixa
Descrição	O utilizador termina a sessão da sua conta
Actores	Utilizador
Pré-Condições	O utilizador tem que estar autenticado
Fluxo Básico	O utilizador clica em "Options" O utilizador clica em "Logout"
Pós-Condições	O utilizador termina a sessão com sucesso
Dependências	NA

Tabela 3.4: CU02 - Terminar sessão

CU03 - Repor palavra-passe

Prioridade	Baixa
Descrição	O utilizador repõe a palavra-passe esquecida
Actores	Utilizador
Pré-Condições	O utilizador tem que ter uma conta válida
Fluxo Básico	O utilizador clica em "Forgot password?" O utilizador insere o seu Email O utilizador recebe um link no seu Email para recuperar a palavra-passe O utilizador abre o link O utilizador é direccionado para uma nova página O utilizador insere a nova palavra-passe O utilizador confirma a nova palavra-passe O utilizador carrega em "Finish"
Pós-Condições	O utilizador repõe a palavra-passe da sua conta com sucesso
Dependências	NA

Tabela 3.5: CU03 - Repor palavra-passe

EX02 - Nova palavra-passe diferente da confirmação

Fluxo de eventos	O utilizador introduz a nova palavra-passe O utilizador confirma a nova palavra-passe O sistema notifica que a nova palavra-passe e a confirmação são diferentes
Pós-Condições	O sistema não consegue repor a palavra-passe

Tabela 3.6: EX02 - Nova palavra-passe diferente da confirmação

CU04 - Alterar dados da conta

Prioridade	Baixa
Descrição	O utilizador pode alterar os diferentes dados da sua conta
Actores	Utilizador
Pré-Condições	O utilizador tem que estar autenticado
Fluxo Básico	O utilizador clica em "Options" O utilizador clica em "Change personal settings" O utilizador pode alterar os diferentes dados da sua conta: - alterar nome (CU05) - alterar Email (CU06) - alterar número de telemóvel (CU07) O utilizador clica em "Finish"
Pós-Condições	NA
Dependências	NA

Tabela 3.7: CU04 - Alterar dados da conta

CU05 - Alterar palavra-passe

Prioridade	Baixa
Descrição	O utilizador altera a palavra-passe da sua conta
Actores	Utilizador
Pré-Condições	O utilizador tem que estar autenticado
Fluxo Básico	O utilizador clica em "Options" O utilizador clica em "Change password" O utilizador insere a palavra-passe actual O utilizador insere a nova palavra-passe O utilizador clica em "Finish"
Pós-Condições	O utilizador altera a sua palavra-passe com sucesso
Dependências	CU04

Tabela 3.8: CU05 - Alterar palavra-passe

EX03 - Palavra-passe incorrecta

Fluxo de eventos	No passo 3 do FB o utilizador insere a palavra-passe atual incorreta O utilizador introduz a nova palavra-passe O utilizador carrega em "Finish" O sistema notifica o utilizador que a palavra-passe actual está incorrecta
Pós-Condições	O sistema não consegue alterar a palavra-passe

Tabela 3.9: EX03 - Palavra-passe incorrecta

CU06 - Alterar nome

Prioridade	Baixa
Descrição	O utilizador altera o nome associado à sua conta
Actores	Utilizador
Pré-Condições	O utilizador tem que estar autenticado
Fluxo Básico	O utilizador clica em "Change name" O utilizador insere o novo nome desejado O utilizador clica em "Finish"
Pós-Condições	O utilizador altera o nome com sucesso
Dependências	CU04

Tabela 3.10: CU06 - Alterar nome

EX04 - Nome inválido

Fluxo de eventos	No passo 2 do FB o utilizador introduz um nome inválido O utilizador carrega em "Finish" O sistema notifica o utilizador de que o nome introduzido não é válido
Pós-Condições	O sistema não consegue alterar o nome

Tabela 3.11: EX04 - Nome inválido

CU07 - Alterar Email

Prioridade	Baixa
Descrição	O utilizador altera o Email associado à sua conta
Actores	Utilizador
Pré-Condições	O utilizador tem que estar autenticado
Fluxo Básico	O utilizador clica em "Change Email" O utilizador insere o novo Email desejado O utilizador clica em "Finish"
Pós-Condições	O utilizador altera o Email com sucesso
Dependências	CU04

Tabela 3.12: CU07 - Alterar Email

EX05 - Email inválido

Fluxo de eventos	No passo 2 do FB o utilizador introduz um Email inválido O utilizador carrega em "Finish" O sistema notifica o utilizador de que o Email introduzido não é válido
Pós-Condições	O sistema não consegue alterar o Email

Tabela 3.13: EX05 - Email inválido

CU08 - Alterar número de telemóvel

Prioridade	Baixa
Descrição	O utilizador altera o número de telemóvel associado à sua conta
Actores	Utilizador
Pré-Condições	O utilizador tem que estar autenticado
Fluxo Básico	O utilizador clica em "Change phone number" O utilizador insere o novo número de telemóvel desejado O utilizador clica em "Finish"
Pós-Condições	O utilizador altera o número de telemóvel com sucesso
Dependências	CU04

Tabela 3.14: CU08 - Alterar número de telemóvel

EX06 - Número de telemóvel inválido

Fluxo de eventos	No passo 2 do FB o utilizador introduz um número de telemóvel inválido O utilizador carrega em "Finish" O sistema notifica o utilizador de que o número de telemóvel introduzido não é válido
Pós-Condições	O sistema não consegue alterar o número de telemóvel

Tabela 3.15: EX06 - Número de telemóvel inválido

CU09 - Visualizar dados históricos

Prioridade	Muito alta
Descrição	Um utilizador pode visualizar dados
Actores	Utilizador
Pré-Condições	O utilizador tem que estar autenticado
Fluxo Básico	O utilizador clica em "Data" O utilizador tem a opção de: Visualizar dados históricos de incêndios (CU09) Visualizar dados históricos de condições meteorológicas (CU10) Visualizar dados populacionais (CU11)
Pós-Condições	Os dados históricos de incêndios existentes são apresentados no ecrã
Dependências	NA

Tabela 3.16: CU09 - Visualizar dados históricos

CU010 - Visualizar dados históricos de incêndios

Prioridade	Muito alta
Descrição	Um utilizador visualiza dados históricos de incêndios
Actores	Utilizador
Pré-Condições	O utilizador tem que estar autenticado Têm que já ter sido carregados na aplicação dados históricos de incêndios
Fluxo Básico	O utilizador clica em "Forest fires storical data"
Pós-Condições	Os dados históricos de incêndios existentes são apresentados no ecrã
Dependências	CU09

Tabela 3.17: CU010 - Visualizar dados históricos de incêndios

CU11 - Visualizar condições meteorológicas mais relevantes na ocorrência de um incêndio

Prioridade	Média
Descrição	Um utilizador visualiza as meteorológicas que tiveram mais impacto na ocorrência de um fogo florestal
Actores	Utilizador
Pré-Condições	Têm que já ter sido carregados dados históricos de incêndios O utilizador tem de estar autenticado O utilizador tem de estar na página destinada a visualizar dados históricos de incêndios
Fluxo Básico	O utilizador selecciona um dos incêndios existentes O utilizador clica em "Conditions that most likely cause the fire event"
Pós-Condições	São apresentadas no ecrã as condições meteorológicas que tiveram mais impacto no incêndio em questão
Dependências	CU10

Tabela 3.18: CU11 - Visualizar condições meteorológicas mais relevantes na ocorrência de um incêndio

CU12 - Visualizar propagação de um incêndio

Prioridade	Média
Descrição	Um utilizador visualiza a propagação que um incêndio teve ao longo do tempo em que esteve activo
Actores	Utilizador
Pré-Condições	Têm que já ter sido carregados dados históricos de incêndios O utilizador tem de estar autenticado O utilizador tem de estar na página destinada a visualizar dados históricos de incêndios
Fluxo Básico	O utilizador selecciona um dos incêndios existentes O utilizador clica em "Display trend"
Pós-Condições	É apresentado ao utilizador a propagação que o incêndio em questão teve durante o período em que esteve activo
Dependências	CU10

Tabela 3.19: CU12 - Visualizar propagação de um incêndio

CU13 - Visualizar dados históricos de condições meteorológicas

Prioridade	Muito alta
Descrição	Um utilizador visualiza dados históricos de condições meteorológicas
Actores	Utilizador
Pré-Condições	O utilizador tem que estar autenticado Têm que já ter sido carregados na aplicação dados históricos de condições meteorológicas
Fluxo Básico	O utilizador clica em "Weather data"
Pós-Condições	Os dados históricos de condições meteorológicas existentes são apresentados no ecrã
Dependências	CU09

Tabela 3.20: CU13 - Visualizar dados históricos de condições meteorológicas

CU14 - Visualizar dados populacionais

Prioridade	Alta
Descrição	Um utilizador visualiza dados populacionais
Actores	Utilizador
Pré-Condições	O utilizador tem que estar autenticado Têm que já ter sido carregados na aplicação dados populacionais
Fluxo Básico	O utilizador clica em "Population data"
Pós-Condições	Os dados populacionais existentes são apresentados no ecrã
Dependências	CU09

Tabela 3.21: CU14 - Visualizar dados populacionais

CU15 - Visualizar a previsão da ocorrência de incêndios

Prioridade	Muito Alta
Descrição	Através de algoritmos de aprendizagem computacional o sistema apresenta ao utilizador uma previsão para a ocorrência de incêndios num certo dia
Actores	Utilizador
Pré-Condições	Têm que já ter sido carregados dados históricos de incêndios, dados históricos de condições meteorológicas e dados populacionais O utilizador tem de estar autenticado
Fluxo Básico	O utilizador clica em "Predict forest fires" O utilizador escolhe o dia que pretende receber uma previsão É apresentada no ecrã a previsão da ocorrência de incêndios nesse dia
Pós-Condições	O sistema efectua a previsão para a ocorrência de incêndios num certo dia com sucesso
Dependências	NA

Tabela 3.22: CU15 - Visualizar a previsão da ocorrência de incêndios

CU16 - Carregar dados

Prioridade	Muito Alta
Descrição	Um fornecedor de dados pode carregar um novo ficheiro de dados
Actores	Fornecedor de dados
Pré-Condições	O fornecedor de dados tem que estar autenticado
Fluxo Básico	O fornecedor clica em "Load data file" O fornecedor tem a opção de: Carregar dados históricos de incêndios (CU17) Carregar dados históricos de condições meteorológicas (CU18) Carregar dados populacionais (CU19)
Pós-Condições	NA
Dependências	NA

Tabela 3.23: CU16 - Carregar dados

CU17 - Carregar dados históricos de incêndios

Prioridade	Muito Alta
Descrição	Um fornecedor de dados carrega um novo ficheiro com dados históricos de incêndios
Actores	Fornecedor de dados
Pré-Condições	O fornecedor de dados tem que estar autenticado O ficheiro a carregar tem que se encontrar no formato CSV, XLS, XSLX, JSON ou XML
Fluxo Básico	O fornecedor clica em "Choose forest fires data file" O fornecedor de dados escolhe um ficheiro do seu computador O fornecedor de dados clica em "Upload file"
Pós-Condições	O ficheiro seleccionado é carregado com sucesso na aplicação
Dependências	CU16

Tabela 3.24: CU17 - Carregar dados históricos de incêndios

CU18 - Carregar dados históricos de condições meteorológicas

Prioridade	Muito Alta
Descrição	Um fornecedor de dados carrega um novo ficheiro com dados históricos de condições meteorológicas
Actores	Fornecedor de dados
Pré-Condições	O fornecedor de dados tem que estar autenticado O ficheiro a carregar tem que se encontrar no formato CSV, XLS, XSLX, JSON ou XML
Fluxo Básico	O fornecedor clica em "Choose weather data file" O fornecedor de dados escolhe um ficheiro do seu computador O fornecedor de dados clica em "Upload file"
Pós-Condições	O ficheiro seleccionado é carregado com sucesso na aplicação
Dependências	CU16

Tabela 3.25: CU18 - Carregar dados históricos de condições meteorológicas

CU19 - Carregar dados populacionais

Prioridade	Alta
Descrição	Um fornecedor de dados carrega um novo ficheiro com dados populacionais
Actores	Fornecedor de dados
Pré-Condições	O fornecedor de dados tem que estar autenticado O ficheiro a carregar tem que se encontrar no formato CSV, XLS, XSLX, JSON ou XML
Fluxo Básico	O fornecedor clica em "Choose population data file" O fornecedor de dados escolhe um ficheiro do seu computador O fornecedor de dados clica em "Upload file"
Pós-Condições	O ficheiro seleccionado é carregado com sucesso na aplicação
Dependências	CU16

Tabela 3.26: CU19 - Carregar dados populacionais

CU20 - Consultar dados

Prioridade	Alta
Descrição	Um fornecedor de dados pode consultar dados
Actores	Fornecedor de dados
Pré-Condições	O fornecedor de dados tem que estar autenticado
Fluxo Básico	O fornecedor de dados clica em "View data" O fornecedor de dados tem a opção de: Consultar dados históricos de incêndios (CU21) Consultar dados históricos de condições meteorológicas (CU22) Consultar dados populacionais (CU23)
Pós-Condições	NA
Dependências	NA

Tabela 3.27: CU20 - Consultar dados

CU21 - Consultar dados históricos de incêndios

Prioridade	Alta
Descrição	Um fornecedor de dados consulta os dados históricos de incêndios existentes e quando e por quem foram inseridos
Actores	Fornecedor de dados
Pré-Condições	O fornecedor de dados tem que estar autenticado Têm que já ter sido carregados na aplicação dados históricos de incêndios
Fluxo Básico	O fornecedor de dados clica em "Forest fires data"
Pós-Condições	Os dados históricos de incêndios são apresentados no ecrã
Dependências	CU20

Tabela 3.28: CU21 - Consultar dados históricos de incêndios

CU22 - Consultar dados históricos de condições meteorológicas

Prioridade	Alta
Descrição	Um fornecedor de dados consulta os dados históricos de condições meteorológicas existentes e quando e por quem foram inseridos
Actores	Fornecedor de dados
Pré-Condições	O fornecedor de dados tem que estar autenticado Têm que já ter sido carregados na aplicação dados históricos de condições meteorológicas
Fluxo Básico	O fornecedor de dados clica em "Weather data"
Pós-Condições	Os dados históricos de condições meteorológicas existentes são apresentados no ecrã
Dependências	CU20

Tabela 3.29: CU22 - Consultar dados históricos de condições meteorológicas

CU23 - Consultar dados populacionais

Prioridade	Alta
Descrição	Um fornecedor de dados consulta os dados populacionais existentes e quando e por quem foram inseridos
Actores	Fornecedor de dados
Pré-Condições	O fornecedor de dados tem que estar autenticado Têm que já ter sido carregados na aplicação dados populacionais
Fluxo Básico	O fornecedor de dados clica em "Population data"
Pós-Condições	Os dados populacionais são apresentados no ecrã
Dependências	CU20

Tabela 3.30: CU23 - Consultar dados populacionais

CU24 - Editar dados

Prioridade	Média
Descrição	Um fornecedor de dados pode editar dados
Actores	Fornecedor de dados
Pré-Condições	O fornecedor de dados tem que estar autenticado
Fluxo Básico	O fornecedor de dados clica em "Edit data" O fornecedor de dados tem a opção de: Editar dados históricos de incêndios (CU25) Editar dados históricos de condições meteorológicas (CU26) Editar dados populacionais (CU27)
Pós-Condições	NA
Dependências	NA

Tabela 3.31: CU24 - Editar dados

CU25 - Editar dados históricos de incêndios

Prioridade	Média
Descrição	Um fornecedor de dados edita os dados históricos de incêndios existentes
Actores	Fornecedor de dados
Pré-Condições	O fornecedor de dados tem que estar autenticado Têm que já ter sido carregados na aplicação dados históricos de incêndios
Fluxo Básico	O fornecedor de dados clica em "Forest fires data" O fornecedor de dados selecciona o incêndio em que pretende editar dados O fornecedor de dados edita os dados pretendidos O fornecedor de dados clica em "Save"
Pós-Condições	Os dados históricos de incêndios são editados com sucesso
Dependências	CU24

Tabela 3.32: CU25 - Editar dados históricos de incêndios

CU26 - Editar dados históricos de condições meteorológicas

Prioridade	Média
Descrição	Um fornecedor de dados edita os dados históricos de condições meteorológicas existentes
Actores	Fornecedor de dados
Pré-Condições	O fornecedor de dados tem que estar autenticado Têm que já ter sido carregados na aplicação dados históricos de condições meteorológicas
Fluxo Básico	O fornecedor de dados clica em "Weather data" O fornecedor de dados selecciona a entrada das condições meteorológicas em que pretende editar dados O fornecedor de dados edita os dados pretendidos O fornecedor de dados clica em "Save"
Pós-Condições	Os dados históricos de condições meteorológicas são editados com sucesso
Dependências	CU24

Tabela 3.33: CU26 - Editar dados históricos de condições meteorológicas

CU27 - Editar dados populacionais

Prioridade	Média
Descrição	Um fornecedor de dados edita os populacionais existentes
Actores	Fornecedor de dados
Pré-Condições	O fornecedor de dados tem que estar autenticado Têm que já ter sido carregados na aplicação dados populacionais
Fluxo Básico	O fornecedor de dados clica em "Population data" O fornecedor de dados selecciona a entrada dos dados populacionais em que pretende editar dados O fornecedor de dados edita os dados pretendidos O fornecedor de dados clica em "Save"
Pós-Condições	Os dados populacionais são editados com sucesso
Dependências	CU24

Tabela 3.34: CU27 - Editar dados populacionais

CU28 - Eliminar dados

Prioridade	Média
Descrição	Um fornecedor de dados pode eliminar dados
Actores	Fornecedor de dados
Pré-Condições	O fornecedor de dados tem que estar autenticado
Fluxo Básico	O fornecedor de dados clica em "Delete data" O fornecedor de dados tem a opção de: Eliminar dados históricos de incêndios (CU29) Eliminar dados históricos de condições meteorológicas (CU30) Eliminar dados populacionais (CU31)
Pós-Condições	NA
Dependências	NA

Tabela 3.35: CU28 - Eliminar dados

CU29 - Eliminar dados históricos de incêndios

Prioridade	Média
Descrição	Um fornecedor de dados elimina dados históricos de incêndios existentes
Actores	Fornecedor de dados
Pré-Condições	O fornecedor de dados tem que estar autenticado Têm que já ter sido carregados na aplicação dados históricos de incêndios
Fluxo Básico	O fornecedor de dados clica em "Forest fires data" O fornecedor de dados selecciona o(s) incêndio(s) que pretende eliminar O fornecedor de dados clica em "Delete"
Pós-Condições	Os dados históricos de incêndios seleccionados são eliminados com sucesso
Dependências	CU28

Tabela 3.36: CU29 - Eliminar dados históricos de incêndios

CU30 - Eliminar dados históricos de condições meteorológicas

Prioridade	Média
Descrição	Um fornecedor de dados elimina dados históricos de condições meteorológicas existentes
Actores	Fornecedor de dados
Pré-Condições	O fornecedor de dados tem que estar autenticado Têm que já ter sido carregados na aplicação dados históricos de condições meteorológicas
Fluxo Básico	O fornecedor de dados clica em "Weather data" O fornecedor de dados selecciona a(s) entrada(s) que pretende eliminar O fornecedor de dados clica em "Delete"
Pós-Condições	Os dados históricos de condições meteorológicas seleccionados são eliminados com sucesso
Dependências	CU28

Tabela 3.37: CU30 - Eliminar dados históricos de condições meteorológicas

CU31 - Eliminar dados populacionais

Prioridade	Média
Descrição	Um fornecedor de dados elimina dados populacionais existentes
Actores	Fornecedor de dados
Pré-Condições	O fornecedor de dados tem que estar autenticado Têm que já ter sido carregados na aplicação dados populacionais
Fluxo Básico	O fornecedor de dados clica em "Population data" O fornecedor de dados selecciona a(s) entrada(s) que pretende eliminar O fornecedor de dados clica em "Delete"
Pós-Condições	Os dados populacionais seleccionados são eliminados com sucesso
Dependências	CU28

Tabela 3.38: CU31 - Eliminar dados populacionais

CU32 - Informar resultado do carregamento de um ficheiro de dados

Prioridade	Baixa
Descrição	O sistema inteligente informa ao fornecedor de dados se o ficheiro em causa foi carregado com sucesso ou existiram problemas no seu carregamento
Actores	Sistema inteligente
Pré-Condições	Tem de ser carregado um ficheiro de dados
Fluxo Básico	NA
Pós-Condições	O sistema inteligente apresenta uma mensagem com a informação relativa ao carregamento do ficheiro
Dependências	NA

Tabela 3.39: CU32 - Informar resultado do carregamento de um ficheiro de dados

CU33 - Gerir conflitos de dados

Prioridade	Baixa
Descrição	O sistema inteligente verifica se existem dados anteriormente carregados iguais aos novos de entrada e, caso existam, os dados são alterados pelos novos dados de entrada
Actores	Sistema inteligente
Pré-Condições	Tem de ser carregado um ficheiro de dados
Fluxo Básico	NA
Pós-Condições	Os dados existentes são actualizados com sucesso
Dependências	NA

Tabela 3.40: CU33 - Gerir conflitos de dados

CU34 - Remover ruído dos dados de entrada

Prioridade	Muito alta
Descrição	O sistema inteligente remove os dados de entrada não relevantes
Actores	Sistema
Pré-Condições	Tem de ser carregado um ficheiro de dados
Fluxo Básico	NA
Pós-Condições	Os dados de entrada não relevantes são removidos com sucesso
Dependências	NA

Tabela 3.41: CU34 - Remover ruído dos dados de entrada

CU35 - Normalizar os dados de entrada

Prioridade	Muito alta
Descrição	O sistema inteligente coloca todos os dados de entrada nas unidades de medida correctas
Actores	Sistema inteligente
Pré-Condições	Os dados de entrada não podem conter ruído
Fluxo Básico	NA
Pós-Condições	Os dados de entrada são normalizados com sucesso
Dependências	NA

Tabela 3.42: CU35 - Normalizar os dados de entrada

CU36 - Inserir dados

Prioridade	Muito alta
Descrição	O sistema inteligente insere os dados de entrada na base de dados
Actores	Sistema inteligente
Pré-Condições	Os dados de entrada têm de se encontrar normalizados
Fluxo Básico	NA
Pós-Condições	Os dados de entrada são inseridos na base de dados com sucesso
Dependências	NA

Tabela 3.43: CU36 - Inserir dados

CU37 - Prever a ocorrência de incêndios

Prioridade	Muito alta
Descrição	O sistema inteligente prevê a ocorrência de incêndios para os próximos dias
Actores	Sistema inteligente
Pré-Condições	Têm de já ter sido carregados dados históricos de incêndios, condições meteorológicas e populacionais
Fluxo Básico	NA
Pós-Condições	O sistema realiza a previsão da ocorrência de incêndios dos próximos dias com sucesso
Dependências	NA

Tabela 3.44: CU37 - Prever a ocorrência de incêndios

3.4 Análise de Riscos

Encontrando-se o presente projecto numa fase completamente inicial, foi necessário efectuar uma análise de quais os possíveis riscos associados ao desenvolvimento do mesmo de modo a precaver a ocorrência de possíveis problemas durante o período de desenvolvimento.

A presente secção tem por objectivo efectuar o levantamento dos possíveis riscos associados ao desenvolvimento projecto em questão e definir as suas respectivas estratégias de mitigação.

Os riscos descritos nesta secção são caracterizados pela probabilidade da sua ocorrência e pelo seu impacto no desenvolvimento do presente projecto.

De modo a caracterizar quer a probabilidade de ocorrência, quer o impacto de um risco no desenvolvimento do projecto, foi considerada uma escala qualitativa cuja sua representação se encontra de seguida apresentada.

- Probabilidade
  - **Baixa**: inferior a 50%
  - **Média**: entre 50% a 80%
  - **Alta**: superior a 80%

- Impacto
  - **Baixo**: Origina dificuldades à realização do projecto sem no entanto afectar o produto final
  - **Médio**: Provoca dificuldades ao desenvolvimento do projecto e promove a probabilidade de o produto final não ficar de acordo com o que era esperado
  - **Alto**: Potencializa a incapacidade de obtenção do produto final

De seguida encontram-se apresentados os riscos que foram levantados para o desenvolvimento do presente projecto.

3.4.1 R01 - Desconhecimento das novas tecnologias

- **ID**: R01
- **Descrição**: A utilização de tecnologias desconhecidas por parte do aluno no desenvolvimento do projecto pode implicar a ocorrência de atrasos que comprometam a implementação de determinadas funcionalidades.
- **Estratégia de mitigação**: O aluno deverá utilizar horas extras para compensar atrasos que possam vir a ser verificados de modo a não comprometer a implementação de nenhuma tarefa planeada.
- **Probabilidade**: Média
- **Impacto**: Médio

3.4.2 R02 - Obtenção do conjunto de dados de operadores móveis

- **ID**: R02
- **Descrição**: A não disponibilização ou a disponibilização tardia do conjunto de dados de operadores móveis pode originar o atraso de tarefas que sejam dependentes deste tipo de dados e comprometer o desenvolvimento do projecto.
- **Estratégia de mitigação**: 0 aluno deverá investigar soluções alternativas à consideração de dados de operadores móveis de modo a que no início de Março de 2020 se não possuir tal conjunto de dados tenha uma solução que não comprometa a execução do desenvolvimento do projecto.
- **Probabilidade**: Alta
- **Impacto**: Alto

3.4.3 R03 - Qualidade do conjunto de dados meteorológicos

- **ID**: R03
- **Descrição**: A não obtenção de um conjunto de dados meteorológicos com a devida qualidade pode comprometer o desempenho dos modelos a serem desenvolvidos.
- **Estratégia de mitigação**: O aluno deverá efectuar uma análise de técnicas de Aprendizagem Computacional que permitam atenuar a falta de qualidade do conjunto de dados e aplicar tais técnicas estudadas no respectivo conjunto de dados
- **Probabilidade**: Média
- **Impacto**: Médio

3.5 Arquitectura do sistema

No desenvolvimento deste projecto iremos utilizar uma arquitectura do tipo **Serverless**.

Designamos por **Serverless** uma arquitectura em que a equipa responsável pelo desenvolvimento do software não necessita de gerir a infraestrutura dos servidores no que diz

respeito à capacidade de processamento, sistemas de armazenamento, etc. Todas estas funções são asseguradas pela **cloud** a utilizar.

A utilização deste tipo de arquitectura apresenta diferentes vantagens ao desenvolvimento do projecto. As principais vantagens são:

- Os desenvolvedores dispõem de mais tempo para se dedicarem às suas funções principais e consequentemente conseguirem desenvolver mais funcionalidades num dado período de tempo.

- A escalabilidade é realizada de forma automática, o sistema oferece maior ou menor disponibilidade consoante os acessos que recebe, tem a capacidade de se ajustar à situação actual.

- O tempo pago é apenas o tempo utilizado, o tempo em que existe acesso a funções do servidor. Isto permite uma redução de custos na medida em que apenas é paga a utilização do servidor, ao contrário dos servidores "tradicionais"em que é paga uma mensalidade, quer o servidor esteja a ser utilizado ou não o valor é sempre o mesmo.

Uma outra vantagem da utilização deste tipo de arquitectura é o facto de ser uma arquitectura tipicamente utilizada por parte da empresa no desenvolvimento dos seus projectos. Tal facto permite um maior apoio por parte da empresa no desenvolvimento deste projecto.

3.5.1 Amazon Web Services

Dentro das arquitecturas **Serverless** a solução de infraestrutura escolhida foi a **Amazon Web Services**.

Dentro da arquitectura existem dois módulos para servir conteúdo:

- Conteúdo estático: o conteúdo estático utilizado na construcção do Front-end (ex. ficheiros HTML, CSS e JavaScript) e é servido com recurso aos produtos **Amazon CloudFront** e **AWS S3**.

- Conteúdo dinâmico: o conteúdo dinâmico presente na aplicação (ex. ficheiros Python). Tal conteúdo é servido com recurso aos produtos **Amazon API Gateway**, **AWS Lambda**, **Amazon DynamoDB** e **Amazon Cognito**.

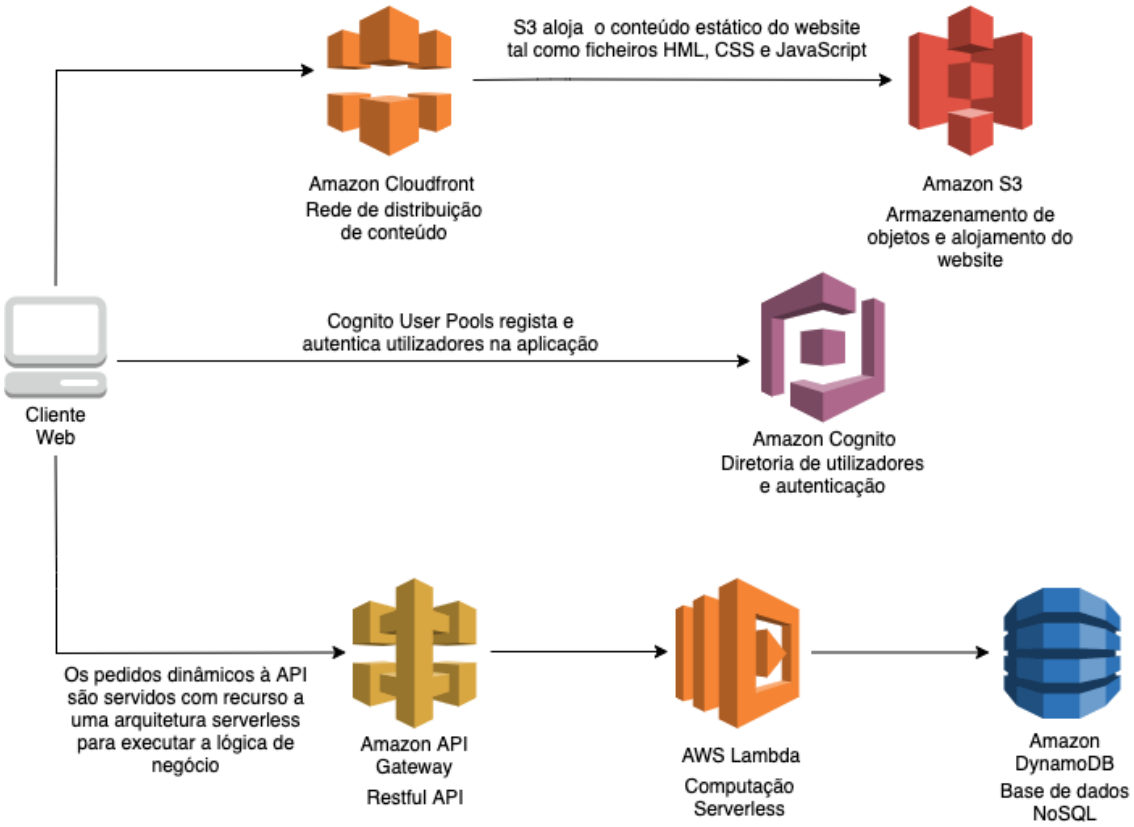


Figura 3.11: Arquitetura Amazon Web Services utilizada

Módulo de conteúdo estático

A **Amazon CloudFront** é uma rede de distribuição de conteúdo, do inglês Content Delivery Network (CDN) que consiste numa rede de servidores que armazenam "cópias"de conteúdo de sites para depois distribuir este mesmo conteúdo aos utilizadores com base na localização geográfica. Os utilizadores ligam-se assim ao servidor que se encontra mais próximo e a transferência de dados é efectuada de forma mais rápida.

A **Amazon S3** é o serviço de armazenamento da **Amazon** onde o site será alojado, é a origem e arquivo do conteúdo.

Módulo de conteúdo dinâmico

A **Amazon API Gateway** é o serviço da **Amazon** responsável por administrar as tarefas envolvidas na recepção e processamento de pedidos à **API**.

A **AWS Lambda** é onde se encontram as diferentes funções da **API** que representam a lógica de negócio. As funções **AWS Lambda** fornecem operações de criação, leitura, atualização e eliminação na base de dados **Amazon DynamoDB**. A execução e escalabilidade das funções é assegurada pela **AWS Lambda** por forma a garantir o melhor funcionamento possível do sistema.

A **Amazon DynamoDB** é a base de dados NoSQL utilizada pela **AWS Lambda** para a execução de funções. Permite a consulta e escrita de dados com uma latência bastante baixa

A **Amazon Cognito** é a directoria onde se encontram guardados os diferentes utilizadores da aplicação. É responsável pela gestão dos diferentes utilizadores bem como da autenticação dos mesmos na aplicação.

3.6 Tecnologias e ferramentas utilizadas

Para a realização deste projecto irão ser utilizadas diferentes tipos de tecnologias e ferramentas.

No que diz respeito ao **Front-end** será utilizado **Angular**.

No que diz respeito ao **Back-end** bem como às tarefas de **Aprendizagem Computacional** será utilizado **Python** auxiliado das bibliotecas **sciki-learn** e **Pandas**.

De forma a servir a comunicação entre o cliente e o servidor será utilizado REST.

Será utilizado JSON como formato do conteúdo das mensagens HTTP trocadas pelo cliente e servidor através de REST e como formato do conteúdo da base de dados **Amazon DynamoDB**.

São utilizados neste projecto também o **Slack** como ferramenta de comunicação com os restantes membros da empresa e o **Git** como ferramenta de controlo de versões do projecto.

3.6.1 Angular

Angular é uma plataforma de código aberto utilizada na construção do frontend de aplicações que tem por base ficheiros HTML, Sass e TypeScript.

O desenvolvimento é feito por módulos e componentes web, o que permite a manutenção do código fonte de uma forma mais fácil e clara.

Trata-se da plataforma utilizada pela empresa na construção do frontend dos seus projectos.

3.6.2 Python

Python é uma linguagem de programação de alto nível orientada a objectos e interpretada.

Trata-se de uma linguagem muito utilizada na área de **Aprendizagem Computacional** devido a possuir uma grande quantidade de bibliotecas direccionadas para esta área, o que torna a sua implementação bastante mais acessível quando comparado com outras linguagens existentes.

3.6.3 scikit-learn

scikit-learn é uma biblioteca de código aberto para a linguagem de programação de programação **Python** que disponibiliza diversos algoritmos e técnicas de Aprendizagem Computacional.

3.6.4 Pandas

Pandas é uma biblioteca de código aberto para a linguagem de programação **Python** que disponibiliza estruturas de dados de desempenho elevado bem como ferramentas para análise de dados.

3.6.5 JSON

JavaScript Object Notation (JSON) é um formato utilizado para guardar e efectuar trocas de dados. É frequentemente utilizado no envio de dados por parte de um servidor devido a ser um formato leve para enviar dados, o que torna o envio mais rápido.

Este formato de dados será assim considerado na utilização do Representational State Transfer (REST) como conteúdo das mensagens HTTP trocadas entre o cliente e o servidor.

3.6.6 REST

REST é um estilo de arquitectura de software para comunicação de aplicações através do protocolo **HTPP** muito utilizado no desenvolvimento de serviços web.

Utiliza um protocolo cliente/servidor sem estado, onde cada mensagem (pedido/resposta) **HTTP** contém toda a informação necessária para compreender o pedido. Nem o cliente nem o servidor necessitam de guardar o estado da comunicação entre mensagens.

REST faz uso de uma sintaxe universal para identificar os recursos, cada recurso é identificado através do respectivo Uniform Resource Identifier (URI).

São utilizados quatro métodos distintos para executar operações. Estes métodos são:

- GET - solicitar recursos pedidos pelo cliente através do uso do URI do recurso. Não provoca a modificação dos recursos pedidos.
- POST - criar um novo recurso com a informação fornecida pelo cliente caso este recurso ainda não exista. Caso exista, nada é feito.
- PUT - actualizar um dado recurso com a informação fornecida pelo cliente ou, caso o recurso não exista, criar o mesmo. O método PATCH permite também alterar recursos, no entanto através deste método é possível alterar apenas parâmetros do recurso. Não é necessário actualizar o recurso em questão na totalidade como no método PUT. A utilização do método PATCH num recurso inexistente, ao contrário de PUT, não cria esse mesmo recurso.
- DELETE - remover um recurso. Esta operação é irreversível, o recurso é eliminado para sempre.

3.7 Plano de trabalho

Nesta secção será detalhado o plano estruturado para o desenvolvimento deste projecto. O plano apresentado foi definido pelo Engenheiro Rafael Maia, responsável pela orientação do estágio por parte da Grama.

3.7.1 Primeiro Semestre

1. Levantamento do estado da arte;
2. Estudo da infraestrutura da Amazon AWS para hosting de plataformas web;
3. Análise de requisitos;
4. Elaboração do plano de desenvolvimento;
5. Prototipagem de uma aplicação básica de visualização de dados históricos de fogos florestais;
6. Desenvolvimento de um protótipo com as seguintes funcionalidades:
   (a) Carregamento de um banco de dados móveis;
   (b) Carregamento de um banco de dados históricos de fogos florestais;
   (c) Carregamento de um banco de dados históricos com a descrição do estado meteorológico;
   (d) Plataforma web para visualização da informação recolhida no espaço (mapa de Portugal) e tempo;
7. Preparação dos protótipos para demonstrações internas;
8. Documentação intermédia do estágio.

3.7.2 Segundo Semestre

1. Processamento de dados aplicando técnicas como filtragem, redução de ruído, enriquecimento, regularização, normalização, segmentação e agregação;
2. Aplicação de técnicas avançadas de **Aprendizagem Computacional** para categorizar, agrupar e reconhecer padrões no banco de dados;
3. Implementação de mecanismos para prever ocorrências futuras de fogos florestais, e onde;
4. Implementação de técnicas de visualização de dados passados e dos padrões encontrados;
5. Implementação de técnicas de visualização de ocorrências de fogos florestais previstos pela aplicação;
6. Preparação do protótipo para demonstrações;
7. Testes funcionais;
8. Testes de usabilidade;
9. Avaliação de requisitos não funcionais;
10. Documentação final de estágio.

3.8 Planeamento de tarefas

De acordo com o plano de trabalho definido por parte da empresa Grama e com os requisitos funcionais levantados já anteriormente neste capítulo, foram criados dois diagramas de Gantt para efectuar a gestão da duração das diferentes tarefas, um relativo ao primeiro semestre e outro ao segundo semestre.

O diagrama de Gantt referente ao primeiro semestre apresenta a calendarização das tarefas que foram realizadas ao longo do primeiro semestre.

Por outro lado o diagrama de Gantt referente ao segundo semestre apresenta o planeamento cronológico inicial para o desenvolvimento das tarefas referentes ao segundo semestre.

Tais diagramas encontram-se de seguida apresentados.

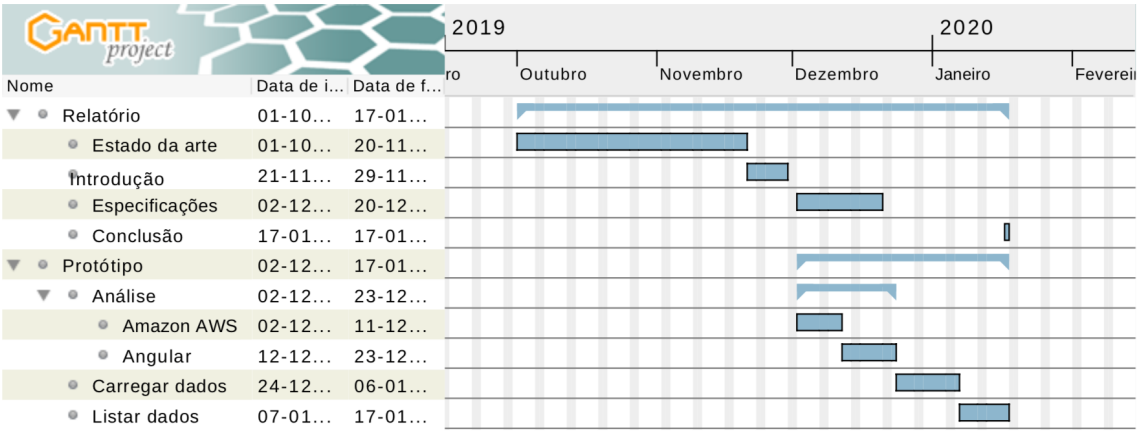


Figura 3.12: Diagrama de Gantt relativo ao primeiro semestre

Na figura 3.12 observamos o diagrama de Gantt que retrata as tarefas realizadas no primeiro semestre bem como a duração aproximada que cada uma das mesmas.

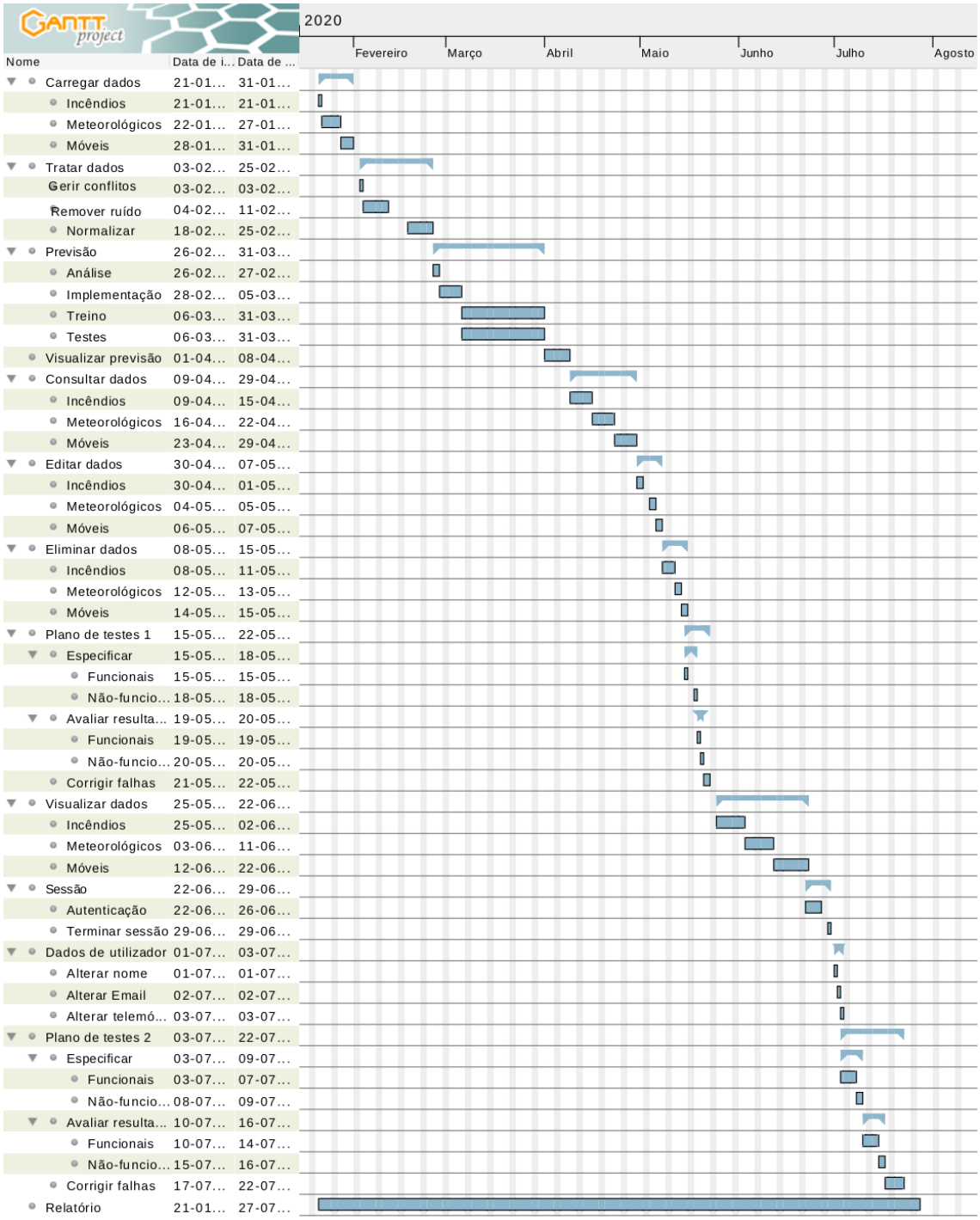


Figura 3.13: Diagrama de Gantt relativo ao segundo semestre

Na figura (3.12) observamos o diagrama de Gantt que apresenta o planeamento proposto para a execução das diferentes tarefas que compõe o segundo semestre.

Capítulo 4

Conjunto de dados

O presente capítulo tem por objectivo detalhar todo o processo de criação do conjunto de dados utilizado no projecto em questão e encontra-se dividido em cinco secções, nomeadamente Dados considerados, Processamento de dados, Criação do conjunto de dados, Conjunto de dados final e Remoção de outliers.

Na secção denominada de Dados considerados são apresentados os dados considerados no desenvolvimento deste projecto bem como a fonte de onde são provenientes.

Adiante na secção denominada de Processamento de dados encontram-se descritas as diferentes etapas envolvidas no processamento dos conjuntos de dados considerados.

Na secção denominada de Criação do conjunto de dados são apresentadas as diferentes etapas envolvidas na criação do conjunto de dados a ser utilizado no desenvolvimento do presente projecto. Posteriormente na secção denominada de Conjunto de dados final são apresentadas as features que se encontram presentes neste mesmo conjunto de dados.

Por último mas não menos importante, é descrito na secção denominada de Remoção de outliers o processo de remoção de outliers do conjunto de dados a ser utilizado.

4.1 Dados considerados

No desenvolvimento deste projecto foram utilizados dados de três diferentes tipos. Dados históricos de ocorrências de incêndios florestais, dados históricos de condições meteorológicas e dados populacionais referentes ao número de habitantes por concelho em Portugal.

Por forma a compreender os dados considerados na realização do projecto em questão, encontram-se de seguida apresentada a proveniência dos diferentes conjuntos de dados e as features presentes em cada um dos diferentes conjuntos.

4.1.1 Dados históricos de ocorrências de incêndios florestais

Os dados relativos a ocorrências de incêndios florestais históricos são provenientes do portal da Central de Dados [16]. Os dados presentes neste portal englobam ocorrências de incêndios florestais entre 1980 e 2015, no entanto foram apenas considerados os dados compreendidos entre 2001 e 2015.

De seguida encontram-se apresentadas as features que compõe este conjunto de dados.

- ano - ano de ocorrência do incêndio
- codigo_sgif - identificador da ocorrência de incêndio do Sistema de Gestão de Informação de Incêndios Florestais
- codigo_anpc - identificador da ocorrência de incêndio da Autoridade Nacional de Emergência e Protecção Civil
- tipo - tipo de ocorrência de incêndio (florestal, agrícola, queimada, falso alarme)
- distrito - distrito em que se verificou a ocorrência de incêndio
- concelho - concelho em que se verificou a ocorrência de incêndio
- local - local em que se verificou a ocorrência de incêndio
- ine - identificador da ocorrência de incêndio do Resultados da procura Instituto Nacional de Estatística
- x - coordenada X da localização da ocorrência de incêndio (EPSG:20790 - Hayford-Gauss Datum Lisboa (Militar))
- y - coordenada Y da localização da ocorrência de incêndio ((EPSG:20790 - Hayford-Gauss Datum Lisboa (Militar))
- lat - latitude relativa à localização da ocorrência de incêndio (EPSG:4326 - WGS84)
- lon - longitude relativa à localização da ocorrência de incêndio (EPSG:4326 - WGS84)
- data_alerta - data relativa ao alerta da ocorrência de incêndio
- hora_alerta - hora relativa ao alerta da ocorrência de incêndio
- data_extincao - data relativa à conclusão da ocorrência de incêndio
- hora_extincao - hora relativa à conclusão da ocorrência de incêndio
- data_primeira_intervencao - data relativa à primeira intervenção sobre a ocorrência de incêndio
- hora_primeira_intervencao - hora relativa à primeira intervenção sobre a ocorrência de incêndio
- fonte_alerta - informação da origem do alerta relativa à ocorrência de incêndio
- nut - geocódigo que referencia as subdivisões de Portugal
- area_povoamento - área povoada que foi carbonizada por parte do incêndio
- area_mato - área de mato que foi carbonizada por parte do incêndio
- area_agricola - área agrícola que foi carbonizada por parte do incêndio
- area_pov_mato - área povoada e de mato carbonizada por parte do incêndio
- area_total - área povoada, de mato e agrícola que foi carbonizada por parte do incêndio
- reacendimento - informação relativa a reacendimento (feature binária)

- reacendimento - informação relativa a ocorrência de incêndio resultar de um reacendimento (feature binária)
- queimada - informação relativa à ocorrência de incêndio ser uma queimada (feature binária)
- falso_alarme - informação relativa à ocorrência de incêndio ser um falso alarme (feature binária)
- fogacho - informação relativa à ocorrência de incêndio ser um fogacho (feature binária)
- incendio - informação relativa à ocorrência de incêndio ser um incêndio (feature binária)
- agricola - informação relativa à ocorrência de incêndio ser agrícola (feature binária)
- perimetro - perímetro da ocorrência de incêndio
- aps - cobertura de danos por parte da Associação Portuguesa de Seguradores
- causa- feature númerica referente à causa da ocorrência de incêndio
- tipo_causa - feature categórica referente à causa da ocorrência de incêndio (negligente, desconhecida, reacendimento, intencional)
- regiao_prof - identificador dos Programas Regionais de Ordenamento Florestal
- ugf - identificador da Unidade de Gestão Florestal

4.1.2 Dados históricos de condições meteorológicas

Os dados relativos às condições meteorológicas históricas são provenientes do portal do Sistema Nacional de Informação de Recursos Hídricos [22].

Os dados considerados para o desenvolvimento do projecto em questão foram obtidos através de Web Scraping em Python com o recurso às bibliotecas Requests e Beautiful Soup.

A utilização de Web Scraping deve-se ao facto de no portal do Sistema Nacional de Informação de Recursos Hídricos [22] os dados se encontrarem organizados por estações meteorológicas. Entre 2001 e 2015 existem 586 estações que dispõe de dados relativos às condições meteorológicas pretendidas e originam um total de 8790 ficheiros CSV, pelo que a utilização de Web Scraping automatizou e diminuiu o tempo despendido na recolha destes mesmos dados.

De seguida encontram-se apresentadas as features presentes para cada estação entre 2001 e 2015.

- Data - data de medição das condições meteorológicas
- Direcção do vento horária - direcção do vento para uma determinada hora medida em graus
- Humidade relativa horária - humidade relativa para uma determinada hora medida em %
- Precipitação horária - precipitação para uma determinada hora medida em mm

- Temperatura do ar horária - temperatura do ar para uma determinada hora medida em graus Celsius
- Velocidade do vento horária - velocidade do vento para uma determinada hora medida em m/s

4.1.3 Dados populacionais

Os dados relativos ao número de habitantes por concelho em Portugal são provenientes do portal Pordata - Base de Dados Portugal Contemporâneo [20].

Os dados são resultantes dos Censos da população portuguesa e encontram-se agrupados em três grandes grupos etários, 0-14 anos, 15-64 anos e 65 anos ou mais. O total de habitantes em todos os grupos é também contabilizado num grupo.

De seguida encontram-se apresentadas as features presentes para cada um dos diferentes grupos acima mencionados.

- Território - território português considerado
- Âmbito Geográfico - designação do tipo de território considerado (Continente, Região Autónoma dos Açores, Região Autónoma da Madeira, região ou município)
- 1960 - valor do número de habitantes no ano de 1960
- 1981 - valor do número de habitantes no ano de 1981
- 2001 - valor do número de habitantes no ano de 2001
- 2011 - valor do número de habitantes no ano de 2011

4.2 Processamento de dados

Os dados considerados para a realização deste projecto, descritos na secção 4.1, não se encontravam no formato desejado e prontos a serem utilizados.

Para que os dados se encontrassem nas condições desejadas para a realização do projecto em questão, foi necessário que os diferentes conjuntos de dados fossem previamente processados. Esta secção visa abordar o processamento de dados efectuado para cada um dos diferentes conjuntos de dados considerados.

4.2.1 Dados históricos de ocorrências de incêndios florestais

A prévia análise do conjunto de dados históricos de ocorrências de incêndios florestais verificou que existiam features irrelevantes para o problema que se pretendia solucionar com o desenvolvimento do presente projecto.

Assim sendo, o primero passo no que diz respeito ao processamento de dados foi analisar o conjunto de dados e compreender quais as features que face ao problema seriam irrelevantes. Posta esta análise certas features foram então removidas do conjunto de dados. Tais features encontram-se de seguida apresentadas.

- codigo_sgif
- codigo_anpc
- ine
- x
- y
- fonte_alerta
- nut
- area_pov_mato
- area_total
- reacendimento
- queimada
- falso_alarme
- reacendimento
- queimada
- falso_alarme
- agricola
- perimetro
- aps
- causa
- regiao_prof
- ugf

Após as features irrelevantes para o problema em questão terem sido removidas do conjunto de dados foi necessário processar as features mantidas por forma a que estas ficassem nas condições pretendidas.

Assim, nesta etapa o primeiro passo foi identificar no conjunto de dados a presença de valores a **NaN** e substituir estes mesmos valores. Valores que se encontravam a **NaN** passaram assim a tomar o valor "Desconhecida".

Após os valores inicialmente com o valor a **NaN** terem sido identificados e substituídos, o passo seguinte foi juntar cada feature data e hora existentes numa só feature. Este processo fez com que passássemos de 6 features (data_alerta, hora_alerta, data_extincao, hora_extincao, data_primeira_intervencao e hora_primeira_intervencao) para 3 features (data_alerta, data_extincao e data_primeira_intervencao).

Por forma a garantir que todas as variáveis categóricas presentes no conjunto de dados seguiam a mesma notação, e para efeitos de compatibilidade futura com features de outros conjunto de dados, as variáveis categóricas presentes neste conjunto de dados foram transformadas de modo a que a primeira letra de cada palavra ficasse capitalizada.

No que diz respeito às coordenadas geográficas latitude e longitude, após uma prévia análise foi verificado que conjuntos de dados inferiores ao ano de 2014 não apresentavam esta informação no seu conjunto de features. Foi também verificado que para os anos de 2014 e 2015 estas mesmas coordenadas apresentavam diferentes formatos de representação de coordenadas geográficas.

Assim foi necessário aplicar uma transformação sobre as coordenadas latitude e longitude presentes nos conjuntos de dados de 2014 e 2015 por forma a que todas estas coordenadas seguissem a notação de graus decimais.

No que diz respeito aos conjuntos de dados inferiores ao ano de 2014 foi desenvolvido um script com o intuito de obter as coordenadas geográficas de cada uma das ocorrências de incêndios florestais. Este script foi desenvolvido em Python com o recurso à biblioteca geopy. As coordenadas geográficas obtidas foram posteriormente concatenadas aos conjunto de dados em questão.

4.2.2 Dados históricos de condições meteorológicas

Cada ficheiro possui a informação relativa às medições efectuadas por parte da estação em questão para o respectivo ano entre 2001 e 2015. A análise dos diferentes conjuntos de dados históricos de condições meteorológicos verificou que a qualidade dos dados difere entre os diversos conjuntos existentes no que diz respeito à quantidade de medições anuais por parte da estação meteorológica em questão bem como à ausência de registo de valores.

Derivado desta análise foi necessário verificar quais os conjuntos de dados que teriam qualidade para serem considerados no desenvolvimento deste projecto. Por forma a verificar a qualidade dos dados, cada um dos 8790 conjuntos de dados existentes foi decomposto em sub conjuntos de dados de acordo com os dias em que se verificava a presença de registos, para cada dia que apresentasse registos foi assim criado um sub conjunto de dados.

Cada um dos sub conjuntos referentes aos dias com presença de registos foi assim analisado de forma individual no que diz respeito à qualidade dos dados. O indicador de qualidade estabelecido para a consideração ou não de cada um dos sub conjuntos de dados foi a existência de registo de pelo menos 4 features das 5 que nos encontrávamos a considerar.

Caso o sub conjunto de dados apresentasse assim registos de 4 das 5 features consideradas, passaríamos a lidar com a ausência de valores presentes no sub conjunto de dados, caso contrário os dados presentes no sub conjunto de dados em questão não seriam considerados no desenvolvimento deste projecto.

Por forma a que valores que se encontravam a **NaN** possuíssem um valor concreto, foi calculada a média de cada uma das features em questão e os valores que se encontravam a **NaN** foram substituídos pelo valor médio da respectiva feature.

Este processamento sobre os dados históricos de condições meteorológicas fez com que fosse possível reduzir os 8790 ficheiros CSV iniciais em apenas 15 ficheiros CSV, um para cada ano entre 2001 e 2015.

Cada um dos novos 15 ficheiros CSV continha assim as seguintes features:

- Estação - estção em que foi efectuada a medição dos valores das condições meteorológicas
- Data - data de medição das condições meteorológicas

- Direcção do vento horária - direcção do vento para uma determinada hora medida em graus
- Humidade relativa horária - humidade relativa para uma determinada hora medida em %
- Precipitação horária - precipitação para uma determinada hora medida em mm
- Temperatura do ar horária - temperatura do ar para uma determinada hora medida em gaus Celsius
- Velocidade do vento horária - velocidade do vento para uma determinada hora medida em m/s

4.2.3 Dados populacionais

O conjunto de dados a considerar encontrava-se originalmente no formato XLSX e não se apresentava pronto a ser lido. Por forma a que o ficheiro se encontrasse nas devidas condições a fim de ser correctamente lido e processado, a informação desejada presente no ficheiro original foi guardada num ficheiro do tipo CSV com recurso à biblioteca Pandas do Python.

Após os dados se encontrarem nas devidas condições para serem lidos foi então feita uma análise ao nível do conteúdo presente no conjunto de dados.

A análise do respectivo conjunto de dados verificou que determinados grupos e features presentes neste conjunto de dados demonstravam-se irrelevantes para o problema que se pretende combater com o desenvolvimento do presente projecto e foram removidas do conjunto de dados a considerar.

As features que não foram removidas encontram-se de seguida apresentadas:

- Território
- Grupo total
  - 1960
  - 1981

4.3 Criação do conjunto de dados

Após os 3 diferentes conjuntos de dados considerados no desenvolvimento deste projecto referidos na secção 4.1 terem sido devidamente processados como descrito na secção 4.2, os diferentes conjuntos de dados foram agrupados num único que englobasse todas as features presentes nos 3 conjuntos de dados.

A junção dos diferentes conjuntos de dados seguiu diferentes etapas que serão descritas nesta secção.

4.3.1 Adição do conjunto de dados históricos de condições meteorológicas

De forma a que fossem conhecidas as condições meteorológicas registadas para a data e local em que se verificaram as diferentes ocorrências de incêndio foi necessário efectuar a correspondência entre os dois conjuntos de dados.

Uma vez que o conjunto de dados históricos de condições meteorológicas apresentava apenas a informação relativa ao nome da estação onde foram medidas as diferentes condições meteorológicas, foi necessário em primeiro lugar conhecer a localização de cada uma das estações meteorológicas em questão.

Localização das estações meteorológicas

A localização de cada uma das respectivas estações meteorológicas foi efectuada de forma manual através da consulta das coordenadas geográficas latitude e longitude que se encontram na informação relativa a cada estação presente no portal do Sistema Nacional de Informação de Recursos Hídricos [23].

À medida que foram consultadas, as coordenadas geográficas referentes a cada estação foram guardadas juntamente com o nome da respectiva estação meteorológica num ficheiro CSV que apresentava a seguinte estrutura:

- Nome da estação meteorológica
- Latitude
- Longitude

Estações meteorológicas das diferentes ocorrências de incêndio

Após ser conhecida a localização das diferentes estações meteorológicas em termos de coordenadas geográficas latitude e longitude, foi necessário averiguar qual a estação meteorológica que se encontrava mais próxima da localização de cada uma das diferentes ocorrências de incêndio.

De forma a atribuir a cada ocorrência de incêndio a estação meteorológica mais próxima, foi desenvolvido um script em Python que calcula a distância geodésica entre dois pontos com recurso à biblioteca geopy.

Para cada ocorrência de incêndio foi verificado qual de entre as estações meteorológicas que apresentavam registos para o ano da ocorrência em questão se encontrava mais próxima. A distância máxima estabelecida entre a localização da ocorrência de incêndio e a localização da estação meteorológica foi 30 km.

Ocorrências de incêndio cuja estação meteorológica mais próxima se encontrava a uma distância igual ou inferior a 30 km da localização da ocorrência de incêndio receberam como nova feature o nome da estação meteorológica que apresentava a menor distância. Por sua vez, ocorrências de incêndio cuja estação meteorológica mais próxima se encontrava a uma distância superior a 30 km foram devidamente assinaladas e excluídas do conjunto de dados final.

Condições meteorológicas das diferentes ocorrências de incêndio

Após se encontrarem atribuídas as estações meteorológicas referentes às diferentes ocorrências de incêndio, foi então necessário adicionar o valor das condições meteorológicas registadas por parte da respectiva estação meteorológica à ocorrência de incêndio.

Deste modo foi desenvolvido um script em Python com o intuito de verificar se para a data da ocorrência de incêndio a estação meteorológica apresentava registos. Para cada ocorrência de incêndio foi verificado se no respectivo ano a estação meteorológica teria registado o valor das condições meteorológicas para a data da ocorrência de incêndio.

Uma vez que o registo das condições meteorológicas por parte das estações meteorológicas fora efectuado por hora e as ocorrências de incêndios florestais contabilizavam a hora e minutos, foi necessário considerar apenas a hora das respectivas ocorrências de incêndio no momento de verificar a existência de registos de condições meteorológicas por parte da respectiva estação meteorológica.

Assim sendo, para cada ocorrência de incêndio foi verificado se a estação meteorológica que fora associada à respectiva ocorrência de incêndio disponibilizava o registo das condições meteorológicas desejadas para a hora da respectiva data da ocorrência.

Ocorrências de incêndio em que se verificava a existência de registos por parte da respectiva estação meteorológica receberam assim os valores da direcção do vento, humidade relativa, precipitação, temperatura e velocidade do vento que foram registados para a respectiva hora da data em questão. Por outro lado, ocorrências de incêndio em que não se verificava a existência de registos por parte da respectiva estação meteorológica foram devidamente assinaladas e excluídas do conjunto de dados final.

4.3.2 Criação do target do conjunto de dados de ocorrências de incêndio

Como já referido anteriormente, o presente projecto tinha como foco a previsão da ocorrência de incêndios florestais nos diferentes concelhos de Portugal.

Para tal efeito foi necessário proceder à criação do target que representasse as ocorrências de incêndios que haviam sido verificadas no conjunto de dados.

O conjunto de dados contemplava a nomenclatura para a ocorrência de dois tipos de ocorrências de incêndio, fogacho e incêndio.

Denominava-se por fogacho uma ocorrência de incêndio cuja área ardida era inferior a 1 hectare e por incêndio uma ocorrência de incêndio cuja área ardida era igual ou superior a 1 hectare.

Deste modo foi desenvolvido um script de Python com recurso à biblioteca Pandas que tinha por objectivo verificar se uma dada entrada do conjunto de dados relativo às ocorrências de incêndio era um fogacho ou um incêndio.

Sendo as features fogacho e incêndio duas features binárias, a feature que apresentasse o valor 1 representava o tipo de ocorrência de incêndio. Caso o valor de ambas fosse 0 significava que a entrada em questão não representava uma ocorrência de incêndio.

Após este processo os dois tipos de ocorrências de incêndio possíveis encontravam-se unidos e o target desejado para o desenvolvimento deste projecto havia sido criado.

4.3.3 Conjunto de dados de não ocorrências de incêndio

Após o target desejado ter sido criado, a sua análise verificou que devido às diferentes etapas realizadas no de processamento de dados e na criação do conjunto de dados anteriormente descritas nas secções 4.2 e 4.3, que o conjunto de dados se encontrava notoriamente desfasado no que diz respeito à proporção de entradas de ocorrências de incêndio face à proporção de ocorrências de não incêndio.

Desta forma foi necessário efectuar a inserção de entradas de não ocorrências de incêndio a fim de equilibrar o número de ocorrências de incêndio face ao número de não ocorrências de incêndio.

Nesta secção serão descritas todas as etapas que foram realizadas na adição de entradas de não ocorrência de incêndio ao conjunto de dados.

União dos dados das estações meteorológicas registados anualmente

Os dados relativos às condições meteorológicas registadas pelas diferentes estações meteorológicas encontravam-se agrupados por ano, onde cada ano contemplava a informação relativa às medições efectuadas por parte das diferentes estações meteorológicas no respectivo ano num ficheiro CSV.

De modo a que os registos das condições meteorológicas existentes ao longo dos diferentes anos fossem dispostos num único ficheiro foi desenvolvido um script em Python com recurso à biblioteca Pandas para tal efeito.

O script desenvolvido efectuou a leitura da informação presente em cada um dos diferentes ficheiros e transcreveu esta mesma informação para um novo ficheiro.

Desta forma a informação relativa às condições meteorológicas que inicialmente se encontrava disposta por 15 ficheiros passou agora a ser apresentada num único ficheiro, o que ofereceu uma maior praticabilidade do acesso aos registos de condições meteorológicas.

Selecção das entradas de não ocorrências de incêndio

Posteriormente aos registos das condições meteorológicas terem sido agrupados num único ficheiro como descrito na sub secção 4.3.3, foi então feita a selecção das entradas de ocorrências de incêndio a partir das quais se extrairia a informação das condições meteorológicas relativas às não ocorrências de incêndio.

Uma vez que o objectivo da adição de entradas de não ocorrências de incêndio era balancear o conjunto de dados no que diz respeito à proporção de ocorrências de incêndio face à proporção de não ocorrências de incêndio, e que em cada ano o número de ocorrências de incêndio existentes diferia, a selecção das entradas de ocorrências de incêndio foi feita anualmente.

Para tal efeito foi desenvolvido um script em Python com recurso à biblioteca Pandas.

O script desenvolvido começava por efectuar a leitura do conjunto de dados relativo às ocorrências de incêndio e restringir as ocorrências de incêndio às respectivas ocorrências do ano em questão.

De seguida, as ocorrências de incêndio referentes ao ano em questão foram baralhadas por forma a garantir a disposição aleatória das ocorrências desse mesmo ano e, depois de

baralhadas, foram seleccionadas 40% das ocorrências do respectivo ano.

Após se encontrarem seleccionadas 40% das ocorrências de incêndio do ano em questão foi então efectuada a leitura do ficheiro relativo às condições meteorológicas anteriormente abordado na sub secção 4.3.3.

Para cada uma das ocorrências de incêndio existentes foi verificado se existiam registos de condições meteorológicas anteriores à data de cada ocorrência registada. Foi definido um intervalo de 5 dias anteriores à data da ocorrência de incêndio na procura de registos meteorológicos existentes.

Sendo que os registos de condições meteorológicas foram efectuados de hora em hora e as ocorrências de incêndio podiam ser verificadas a qualquer minuto, foi garantido que para a hora da ocorrência de um incêndio não existia também informação relativa a uma não ocorrência de incêndio.

As condições meteorológicas presentes no conjunto de dados que verificavam as condições acima mencionadas eram assim gravadas num ficheiro CSV que contemplava a informação relativa às não ocorrências de incêndio do ano em questão.

Este processo foi efectuado para os anos compreendidos entre 2001 e 2015.

Validação do conjunto de dados de não ocorrências de incêndio

A análise dos ficheiros que contemplavam a informação relativa às não ocorrências de incêndio de cada um dos anos considerados, demonstrou que estes mesmos ficheiro apresentava um elevado número de entradas, pelo que não poderiam ser todas consideradas.

Assim sendo, cada um destes ficheiros foi baralhado por forma a garantir a disposição aleatória das entradas do conjunto de dados. Depois de baralhado, foi verificado o número de ocorrências de incêndio que foram registadas para cada um dos anos compreendidos entre 2001 e 2015, e foram seleccionadas tantas entradas para cada ano quanto o número de ocorrências de incêndio registado para esse mesmo ano.

Uma vez que o intervalo estabelecido na procura de condições meteorológicas fora 5 dias anteriores à data de uma ocorrência de incêndio, a existência de ocorrências de incêndio neste intervalo de tempo poderia resultar em registos de não ocorrências de incêndio repetidos, e/ou registos de ocorrências e não ocorrências de incêndio referentes à mesma data e hora.

Por forma a garantir a não existência de registos de não ocorrências de incêndio repetidos, e/ou registos de ocorrências e não ocorrências de incêndio referentes à mesma data e hora foram desenvolvidos 2 scripts em Python com recurso à biblioteca Pandas.

O primeiro script desenvolvido tinha por objectivo verificar o conjunto de dados referente às não ocorrências de incêndios quanto à presença de registos repetidos e remover estes mesmos registos que se encontravam repetidos.

O segundo script visava efectuar a leitura dos conjuntos de dados relativos às ocorrências e não ocorrências de incêndio, e, verificar para cada uma das não ocorrências de incêndio registadas a existência de ocorrências de incêndio para a mesma data e hora.

Caso se verificasse a existência de registos de ocorrências e não ocorrências de incêndio referentes à mesma data e hora, tais não ocorrências de incêndio eram assim removidas do respectivo conjunto.

Em ambos os scripts a verificação foi efectuada com base nos valores do concelho e do Timestamp presentes em cada registo.

No final de todo este processo, o conjunto de dados relativo às não ocorrências de incêndios garantia assim a qualidade e confiabilidade necessárias e apresentava as seguintes features:

- Timestamp
- Data
- Estação
- Concelho
- Direcção do vento horária
- Humidade relativa horária
- Precipitação horária
- Temperatura do ar horária
- Velocidade do vento horária
- Temperatura mínima do ar horária
- Temperatura média do ar horária
- Temperatura máxima do ar horária
- Área ardida (0)

4.3.4 União dos conjuntos de dados de ocorrências e não ocorrências de incêndio

Posteriormente ao conjunto de dados de não ocorrências de incêndio ter sido validado como anteriormente referido na sub secção 4.3.3, os conjuntos de dados de ocorrências e não ocorrências de incêndio foram então unidos num único conjunto.

Para tal efeito foi desenvolvido um script em Python com recurso à biblioteca Pandas que tinha por objectivo efectuar a leitura dos ficheiros de dados relativos às ocorrências e não ocorrências de incêndio e transcrever a informação presente em cada um dos mesmos para um novo ficheiro CSV.

Após a informação relativa às ocorrências e não ocorrências de incêndio se encontrar presente em apenas num único ficheiro, foi efectuado ainda neste mesmo script a ordenação cronológica das entradas de dados de acordo com o respectivo valor do timestamp.

O conjunto de dados contemplava agora as features que se encontram de seguida apresentadas.

- Timestamp
- Concelho
- Direcção do vento horária

- Velocidade do vento horária
- Humidade relativa horária
- Precipitação horária
- Temperatura do ar horária

4.3.5 Adição do conjunto de dados populacionais

De forma a que o número de habitantes dos diferentes concelhos de Portugal fosse adicionado ao conjunto de dados que se encontra a ser criado foi necessário efectuar a correspondência entre o conjunto de dados populacionais e o conjunto de dados em criação.

A correspondência efectuada entre os dois conjuntos de dados foi feita com base no respectivo concelho, uma vez que era uma feature que se encontrava presente nos dois conjuntos de dados e permitiria assim que a ligação entre ambos fosse possível.

Uma vez que os dados populacionais considerados são resultantes dos Censos da população portuguesa, e de entre 2001 e 2015 os Censos terem sido calculados em 2001 e 2011 foi necessário também ter em consideração o ano do conjunto de dados das ocorrências de incêndios para adicionar o respectivo valor populacional. Assim, entre 2001 e 2010 foram considerados os valores obtidos por parte dos Censos realizados em 2001 e entre 2011 e 2015 foram considerados os valores obtidos por por dos Censos realizados em 2011.

Assim foi desenvolvido um script em Python com o intuito de que todas as ocorrências de incêndio possuíssem também a informação relativa ao número de habitantes existentes de acordo com o concelho e o ano em questão.

4.3.6 Criação de novas features

A análise das diferentes features presentes no conjunto de dados demonstrou que seria possível, através da aplicação de técnicas de feature engineering, originar novas features que poderiam acrescentar informação relevante ao problema a ser retratado.

Uma vez que o conjunto de dados considerado apresentava informação relativa ao concelho em que tinha sido verificada cada ocorrência de incêndio e que o número de habitantes presente no conjunto de dados era também referente ao número de habitantes por concelho em Portugal, poderia ser útil considerar novas features tendo em conta o concelho.

Deste modo foram assim originadas 3 novas features tendo em o concelho em consideração, número de ocorrências de incêndios por concelho do ano anterior, ranking por concelho do ano anterior e área ardida por concelho do ano anterior.

De seguida encontram-se apresentadas cada uma das novas features originadas bem como todo o processo envolvido na criação de cada uma das mesas.

Número de ocorrências de incêndios por concelho do ano anterior e Ranking por concelho do ano anterior

Por forma a tentar compreender se o número de ocorrências de incêndio num determinado ano e concelho influenciaria o número de ocorrências do mesmo concelho no ano seguinte foi desenvolvida uma Slide Window em Python.

A Slide Window desenvolvida tinha por objectivo calcular o número de ocorrências de incêndio para um dado ano, e, atribuir um ranking para os diferentes concelhos de Portugal em função do número de ocorrências de incêndios verificadas em cada um dos mesmos.

O concelho que apresentava mais ocorrências de incêndio para o ano em questão era o detentor da primeira posição do ranking, e o concelho que apresentava menos ocorrências assumia o último lugar do ranking.

Os valores obtidos quer para o número de ocorrências quer para o ranking eram tidos em consideração para o ano seguinte, de forma a que a cada ano tivesse em consideração a informação relativa aos concelhos onde existiram mais ocorrências de incêndios e qual o respectivo número de ocorrências.

Os dados iniciais utilizados na execução da presente Slide Window foram os dados referentes ao ano de 2001.

Área ardida por concelho do ano anterior

Os solos ardidos necessitam de tempo de regeneração até que se verifique uma nova ocorrência de incêndio. Um solo que tenha ardido há relativamente pouco tempo apresenta uma menor probabilidade de ocorrência de incêndio face a um solo em que já não se verifique uma ocorrência de incêndio há um tempo considerável.

Desta forma foi desenvolvida uma Slide Window em Python com o intuito de verificar se a área ardida anual nos diferentes concelhos tem influência no número de ocorrências de incêndio no ano seguinte.

A Slide Window desenvolvida pretendia efectuar a soma do valor da área ardida resultante das diferentes ocorrências de incêndio anuais para cada um dos diferentes concelhos de Portugal de forma a ter em consideração estes mesmos valores para o ano seguinte, de forma a que em cada ano fosse tido em consideração o valor da área ardida do último ano.

Os dados iniciais utilizados na execução desta mesma Slide Window foram também os dados referentes ao ano de 2001.

4.3.7 Adição de novas features

Com o intuito de adicionar novas features ao conjunto de dados foi efectuada uma pesquisa de features que poderiam ser relevantes de ser adicionadas face às features já existentes e ao problema que a combater com o desenvolvimento deste projecto.

Uma vez que a feature concelho efectuava a "interligação"de dados, como referido anteriormente na secção 4.3.6, foram assim tidas em consideração 3 novas features que tinham em consideração esta mesma feature.

As novas features adicionadas foram respectivamente a dimensão dos concelhos, a altitude mínima dos concelhos face ao nível médio das águas do mar e a altitude máxima dos concelhos face ao nível médio das águas do mar.

De seguida encontram-se assim apresentadas cada uma destas mesmas features bem como a origem e o respectivo processamento realizado na adição das mesmas.

Dimensão dos concelhos

Na secção 4.3.6 anteriormente apresentada verificámos que fora criada uma nova feature que contemplava a área ardida por concelho relativa ao ano anterior.

Deste modo, considerar a dimensão dos concelhos onde tiveram origem as respectivas ocorrências de incêndio poderia ser útil ao problema a solucionar com o desenvolvimento do presente projecto.

Através da adição da feature referente à dimensão do concelho conseguiríamos verificar a relação entre a área ardida de um dado concelho no ano anterior e a área total desse mesmo concelho, verificar se a quantidade de área não ardida de um concelho (diferença entre a área total e a área ardida no ano anterior do concelho) teria influência no número de ocorrências de incêndio.

O conjunto de dados relativo à dimensão dos diferentes concelhos de Portugal são provenientes do portal Pordata - Base de Dados Portugal Contemporâneo [21].

A análise do conjunto de dados em questão demonstrou que existiam diversas features desnecessárias para a adição da informação relativa à dimensão dos diferentes concelhos. A informação presente no respectivo conjunto de dados encontra-se de seguida apresentada.

- Território - território português considerado
- Âmbito Geográfico - designação do tipo de território considerado (Continente, Região Autónoma dos Açores, Região Autónoma da Madeira, região ou município)
- 2001 - dimensão de um dado concelho em Km^2 no ano de 2001
- 2009 - dimensão de um dado concelho em Km^2 no ano de 2009
- 2010 - dimensão de um dado concelho em Km^2 no ano de 2010
- 2011 - dimensão de um dado concelho em Km^2 no ano de 2011
- 2012 - dimensão de um dado concelho em Km^2 no ano de 2012
- 2013 - dimensão de um dado concelho em Km^2 no ano de 2013
- 2014 - dimensão de um dado concelho em Km^2 no ano de 2014
- 2015 - dimensão de um dado concelho em Km^2 no ano de 2005
- 2016 - dimensão de um dado concelho em Km^2 no ano de 2006
- 2017 - dimensão de um dado concelho em Km^2 no ano de 2017
- 2018 - dimensão de um dado concelho em Km^2 no ano de 2018
- 2019 - dimensão de um dado concelho em Km^2 no ano de 2019

A diferença verificada entre o valor da dimensão dos diferentes concelhos entre os anos de 2001 e 2015 era relativamente pequena de ano para ano face à dimensão dos concelhos em questão, pelo que foi considerada como não significativa.

Desta forma foram assim considerados os valores da dimensão dos concelhos relativos ao ano de 2015 como valores da nova feature a ser adicionada.

A adição dos valores relativos à dimensão dos concelhos foi feita através de um script em Python, onde para cada Concelho, foram averiguadas as ocorrências que haviam sido detectadas nesse mesmo concelho e adicionada assim a cada uma destas ocorrências o valor da dimensão do respectivo concelho.

Altitude mínima e máxima dos concelhos face ao nível médio das águas do mar

O nível médio das águas do mar é um indicador que permite representar a altitude de uma região acima do nível médio do mar através da especificação de curvas de nível e das suas respectivas cotas.

A altitude em relação ao nível médio das águas do mar permite que seja possível compreender a que altura se encontra uma dada região.

Com o intuito de compreender se a altitude seria um factor que apresentava impacto na ocorrência de incêndios foram consideradas 2 novas features para tal efeito, a altitude mínima e máxima em relação dos concelhos de Portugal face ao nível médio das águas do mar.

Através da informação relativa à altitude mínima e máxima dos diferentes concelhos face ao nível médio das águas do mar conseguimos ainda ter a noção da declividade de cada um destes mesmos concelhos, e se esta declividade apresenta apresenta impacto no que diz respeito à ocorrência de incêndios.

Os dados referentes à altitude mínima e máxima dos concelhos face ao nível médio das águas do mar são provenientes do portal Pordata - Base de Dados Portugal Contemporâneo [18][17].

A análise dos 2 conjuntos de dados demonstrou que os 2 conjuntos tinham em consideração as mesmas features. Tais features encontram-se de seguida apresentadas.

- Território - território português considerado
- Âmbito Geográfico - designação do tipo de território considerado (Continente, Região Autónoma dos Açores, Região Autónoma da Madeira, região ou município)
- 2009 - altitude mínima/máxima de um dado concelho em m no ano de 2009
- 2010 - altitude mínima/máxima de um dado concelho em m no ano de 2010
- 2011 - altitude mínima/máxima de um dado concelho em m no ano de 2011
- 2012 - altitude mínima/máxima de um dado concelho em m no ano de 2012
- 2013 - altitude mínima/máxima de um dado concelho em m no ano de 2013
- 2014 - altitude mínima/máxima de um dado concelho em m no ano de 2014
- 2015 - altitude mínima/máxima de um dado concelho em m no ano de 2005
- 2016 - altitude mínima/máxima de um dado concelho em m no ano de 2006
- 2017 - altitude mínima/máxima de um dado concelho em m no ano de 2017
- 2018 - altitude mínima/máxima de um dado concelho em m no ano de 2018
- 2019 - altitude mínima/máxima de um dado concelho em m no ano de 2019

Uma vez que os 2 conjuntos de dados apresentavam apenas informação a partir do ano de 2009 e que não se apresentam significativas alterações em relação à altitude mínima e máxima dos diferentes concelhos de Portugal face ao nível médio das águas do mar, foram assim considerados os valores relativos ao ano de 2015 como valores da nova feature a ser adicionada.

A adição dos valores relativos à altitude mínima e máxima dos concelhos face ao nível médio das águas do mar foi feita através de um script em Python onde, para cada concelho, foram averiguadas as ocorrências que haviam sido detectadas nesse mesmo concelho, e adicionada assim a cada uma destas ocorrências os respectivos valores da altitude mínima e máxima do concelho em questão face ao nível médio das águas do mar.

Latitude e longitude dos concelhos

O presente projecto, como já mencionado anteriormente tem por objectivo principal a previsão da ocorrência de incêndios dos concelhos de Portugal.

Uma vez que os modelos de Aprendizagem Computacional não possuem a capacidade de suporte de features categóricas, e que o concelho é uma feature categórica, foi necessário fazer com que os diferentes concelhos fossem considerados de forma numérica.

Com o intuito de obter uma representação numérica para a feature concelho foram analisados dois tipos de codificação, **Ordinal Encoding** [55] e **One-Hot Encoding** [54].

Sendo que a feature concelho contempla os diferentes concelhos existentes em Portugal, não existe uma relação ordinal valores assumidos pela mesma.

Desta forma utilização de **Ordinal Encoding** não se adequava ao cenário em questão, "Forcing an ordinal relationship via an ordinal encoding and allowing the model to assume a natural ordering between categories may result in poor performance or unexpected results (predictions halfway between categories)."[34].

Foi então utilizado **One-Hot Encoding** para representar a feature concelho, onde cada valor tomado por parte desta feature é representado de forma binária, "Each bit represents a possible category. If the variable cannot belong to multiple categories at once, then only one bit in the group can be "on."This is called one-hot encoding ..."[6].

A utilização de One-Hot Encoding originou com que fossem adicionadas 278 novas features para representar os 278 concelhos diferentes existentes em Portugal Continental.

Devido a esta estratégia não ter alcançado os melhores resultados foi desenvolvida uma nova estratégia com base nas coordenadas geográficas latitude e longitude centrais de cada Concelho.

Para tal foi desenvolvido um script em Python com recurso à biblioteca geopy que visava descobrir e guardar as coordenadas geográficas latitude e longitude centrais de cada um dos diferentes concelhos.

Após as coordenadas geográficas centrais de todos os concelhos se encontrarem guardadas, foi desenvolvido um script em Python com recurso à biblioteca Pandas que tinha por objectivo atribuir a cada ocorrência de incêndio as coordenadas geográficas centrais do concelho em que foi verificada a mesma ocorrência.

4.4 Conjunto de dados final

Após todas as etapas descritas na secção 4.3 terem sido realizadas o conjunto de dados englobava agora todas as features pretendidas.

As features presentes no conjunto de dados final encontram-se de seguida apresentadas:

- Timestamp
- Latitude do concelho
- Longitude do concelho
- Número de habitantes do concelho
- Direcção do vento horária
- Velocidade do vento horária
- Humidade relativa horária
- Precipitação horária
- Temperatura do ar horária
- Altitude mínima do concelho face ao nível médio das águas do mar
- Altitude máxima do concelho face ao nível médio das águas do mar
- Dimensão do concelho
- Área ardida do concelho no ano anterior
- Número de ocorrências de incêndio no concelho no ano anterior
- Ranking de ocorrências de incêndio no concelho no ano anterior

4.5 Remoção de outliers

Após o conjunto de dados final, apresentado anteriormente na secção 4.4 se encontrar criado foi ainda necessário efectuar a remoção de outliers do conjunto de dados antes do mesmo ser utilizado por parte dos modelos de Aprendizagem Computacional.

A presença de outliers foi verificada com base no cálculo do Z-Score, que pode ser expresso através da fórmula que se segue:

$$z = \frac{x - \mu}{\sigma}$$

onde:

μ = média

σ = desvio padrão

O Z-Score pode ser definido como a medida de desvios padrão abaixo ou acima da população e expresso através de uma curva de distribuição normal.

O Z-Score varia entre -3 desvios padrões e 3 desvios padrões.

De seguida encontra-se apresentada uma imagem que ilustra uma Z-Score Normal Distribution.

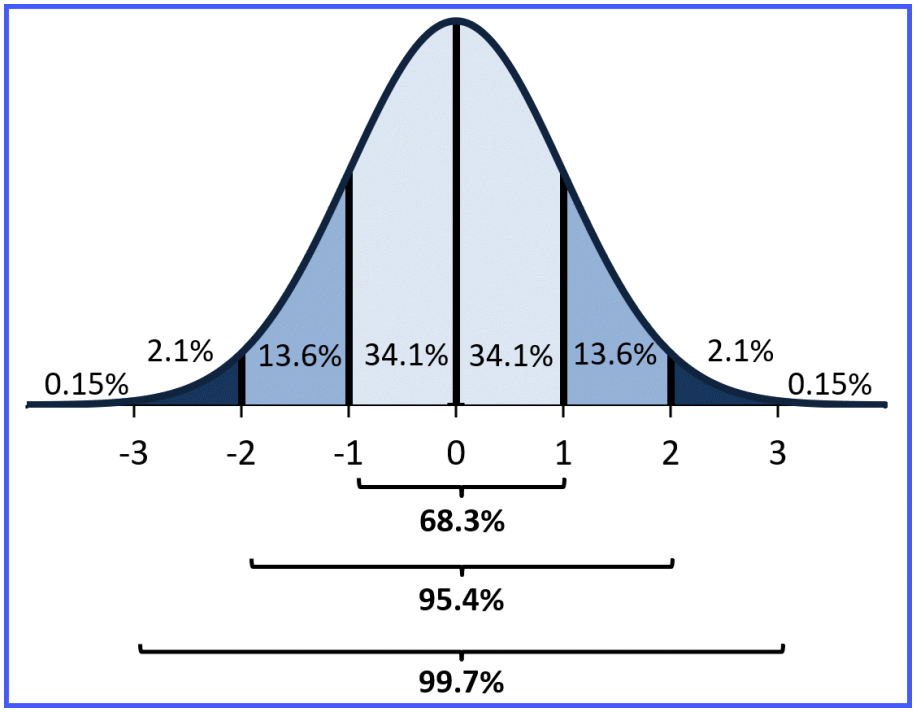


Figura 4.1: Z-Score Normal Distribution [70]

A imagem anteriormente apresentada permite-nos visualizar uma aproximação da distribuição dos dados em conformidade com a variação do desvio padrão.

Foram considerados como outliers dados cujo seu Z-Score fosse em módulo igual ou superior a 3 desvios padrões.

Podemos verificar através da análise da distribuição dos dados que variações em módulo iguais ou superiores a 3 desvios padrões representam uma percentagem de aproximadamente 3% dos dados.

Desta forma a remoção dos outliers foi efectuada em registos de dados onde pelo menos uma das diferentes features consideradas apresentasse um Z-Score em módulo igual ou superior a 3 desvios padrões.

Capítulo 5

Desenvolvimento

O presente capítulo tem por objectivo apresentar as diferentes fases associadas ao desenvolvimento do projecto em questão e encontra-se dividido em três secções, nomeadamente Aprendizagem Computacional, Back-end e Front-end.

Na secção denominada de Aprendizagem Computacional serão abordadas todas as técnicas utilizadas e os diferentes modelos construídos.

Adiante na secção denominada de Back-end serão apresentadas as diferentes funcionalidades desenvolvidas ao nível do servidor.

Por último mas não menos importante serão apresentadas na secção Front-end todas as funcionalidades implementadas ao nível da web.

5.1 Aprendizagem Computacional

A presente secção tem por objectivo apresentar os diferentes algoritmos e técnicas de Aprendizagem Computacional que foram desenvolvidas com o intuito de apresentar solução para o problema que se pretende combater com o desenvolvimento deste projecto.

Nesta secção serão apresentados em primeiro lugar os algoritmos de Aprendizagem Computacional que foram desenvolvidos.

Posteriormente serão apresentadas as diferentes técnicas de Aprendizagem Computacional e especificações que foram utilizadas na construção dos modelos adoptados.

5.1.1 Algoritmos de Aprendizagem Computacional

No capítulo relativo ao estudo do estado da arte, capítulo 2, foi possível verificar a utilização de algoritmos de Aprendizagem Computacional distintos por parte de diferentes autores para apresentar solução ao problema dos incêndios florestais que pretendiam combater.

De seguida encontra-se apresentada a tabela com a informação relativa aos algoritmos considerados por cada um dos diferentes autores estudados no estudo do Estado da Arte, 2, bem como o/os algoritmo/algoritmos que levaram estes mesmos autores a alcançar os melhores resultados.

Autor	Algoritmos considerados	Melhores algorimtos
Stojanova et al (2006)	Logistic Regression, Decision Tree, Random Forest, Bootstap Aggregating, AdaBoost	Bootsrap aggregating
Cortez et al (2007)	Support Vector Machine, Random Forest, Multiple Regression, Decision Tree, Rede Neuronal Artificial	Support Vector Machine
Sakr et al (2010)	Support Vector Machine	Support Vector Machine
Yu et al (2011)	Rede Neuronal Artificial, Sistema Baseado em Regras	Rede Neuronal Artificial
Karouni et al (2014)	Decision Tree, Rede Neuronal Artificial	Rede Neuronal Artificial
Shidik et Mustofa (2016)	Rede Neuronal Artificial, Fuzzy C-Means	Rede neuronal, Fuzzy C-Means
Storer (2016)	Rede Neuronal Backpropagation, Rede Neuronal Particle Swarm Optimization, K- means Optimization, K- means Clustering, Spectral Clustering	Rede neuronal Backpropagation, K-means Clustering, Spectral Clustering
Ramasubramanian (2017)	Linear Regression, Logistic Regression, Support Vector Machine, Bagging, Random Forest e Gradient Boosting	Random Forest
_Janabi et al (2017)	Rede Neuronal Artificial Radial Basis Function, Multilayer Perceptron, Polynomial e Cascade Correlation, Support Vector Machine	Support Vector Machine

Tabela 5.1: Análise dos algoritmos utilizados por autores no estudo do Estado da Arte

Ao analisarmos a tabela anteriormente apresentada podemos verificar que certos algoritmos se destacam na coluna referente aos melhores algoritmos, que se encontra assinalada a verde.

Podemos considerar, com base na análise dos diferentes resultados por parte dos diferentes autores, a existência de algoritmos mais adequados à previsão de incêndios florestais.

A análise da tabela anteriormente apresentada permite-nos observar que os algoritmos que demonstraram apresentar um melhor desempenho no que diz respeito à previsão de ocorrências de inêndios florestais foram respectivamente os algoritmos **Support Vector Machine** 2.3.2 e **Random Forest** 2.3.4. A respectiva análise foi efectuada com base na comparação dos algoritmos que apresentaram um melhor desempenho face aos algoritmos considerados.

Feita a análise dos diferentes algoritmos considerados por parte dos diferentes autores, e tendo o estudo do Estado da Arte como suporte, os algoritmos **Support Vector Ma-**

chine e **Random Forest** foram assim os algoritmos de Aprendizagem Computacional considerados no desenvolvimento do presente projecto.

Cada um dos algoritmos considerados será de seguida apresentados de forma a conhecermos a fonte da sua proveniência, bem como os hiperparâmetros e os respectivos valores que foram considerados por parte de cada algoritmo.

Para denotar um conjunto foi considerado []. Para denotar um intervalo foi considerado (x, y, z), onde x representa o valor inicial, y o valor final e z o valor de incremento do intervalo.

De salientar que de seguida se encontram apresentados os hiperparâmetros considerados, o seu respectivo intervalo de valores e a descrição de cada hiperparâmetro. No capítulo 6 será posteriormente abordado de que forma estes hiperparâmetros possuem influência na obtenção de resultados por parte de cada um dos algoritmos em questão.

Support Vector Machine

O algoritmo Support Vector Machine considerada no presente projecto foi desenvolvida com recurso à biblioteca scikit-learn do Python [51].

Encontram-se de seguida apresentados os hiperparâmetros e respectivo intervalo de valores considerado para o presente algoritmo.

Hiperparâmetro	Descrição	Valores Considerados
penalty	Norma utilizada na penalização	[l1, l2]
loss	Função de perda	squared_hinge
tol	Tolerância para stopping criteria	(0.0001, 0.01, 0.001))
max_iter	Número máximo de iterações	(1000, 10000, 500)
C	Parâmetro de regularização	[0.001, 0.01, 0.1, 1, 10, 100, 1000]
dual	Selecciona o algoritmo para resolver o problema de otimização dual ou primal	False

Tabela 5.2: Hiperparâmetros do algoritmo Support Vector Machine considerados

Random Forest

O algoritmo Random Forest considerado no presente projecto foi desenvolvida com recurso à biblioteca scikit-learn do Python [56].

Encontram-se de seguida apresentados os hiperparâmetros e respectivo intervalo de valores considerado para o presente algoritmo.

Hiperparâmetro	Descrição	Valores Considerados
max_depth	Profundidade máxima das árvores	[None, (1,28,1)]
n_estimators	Número de árvores criadas	(100, 1000, 50)
min_samples_split	Número mínimo de amostras necessárias para efectuar um split de um nó interno	(0.01, 0.25, 0.05)
max_features	Número de features a considerar na procura do melhor split	[sqrt, log2, (1, 14, 1)]
criterion	Função que mede a qualidade de um split	[gini, entropy]

Tabela 5.3: Hiperparâmetros do algoritmo Random Forest considerados

5.1.2 Divisão do conjunto de dados em conjuntos de treino e teste

A divisão do conjunto de dados em conjuntos de treino e teste é um ponto fulcral na obtenção de resultados por parte de modelos de Aprendizagem Computacional.

A divisão do conjunto de dados pode originar com que os resultados obtidos sejam maus, bons ou até mesmo enganosamente bons.

Certas divisões do conjunto de dados podem apresentar bons resultados que na verdade não correspondem a bons resultados mas sim a uma boa adaptação do modelo ao conjunto de dados de treino, overfitting.

A presença de overfitting é um factor bastante negativo na construção de um bom modelo de Aprendizagem Computacional. Um modelo que contenha overfitting não possui capacidade de generalização, de conseguir prever correctamente novos conjuntos de dados.

No presente projecto a divisão conjunto de dados em conjuntos de treino e teste foi feita com base na consideração dos 70% iniciais dos dados como conjunto de treino e os restantes 30% como conjunto de teste.

Por forma a descobrir a melhor combinação de hiperparâmetros e a prevenir a ocorrência de overffiting nos modelos desenvolvidos foi efectuada cross-validation no conjunto de treino criado, que será abordada adiante na sub secção 5.1.3.

No conjunto de teste foi avaliada a capacidade de generalização do modelo treinado na classificação de novos dados.

5.1.3 Cross-validation

A cross-validation é uma técnica de Aprendizagem Computacional utilizada para garantir o ajuste dos hiperparâmetros, que será abordado adiante na sub secção 5.1.4, e a robustez do desempenho de um modelo.

K-fold cross-validation e hold-out cross-validation são os tipos de cross-validation mais utilizados.

O problema a combater com o desenvolvimento deste projecto lida com eventos em que o factor temporal é um factor importante a ter em consideração no momento da divisão do conjunto de dados para treino e teste.

Ao lidarmos com eventos temporais não podemos assumir uma distribuição aleatória para a construção do conjunto de treino e teste, pois não faz sentido utilizar valores do futuro para prever valores do passado. Existe uma relação temporal entre as diferentes entradas

de dados e essa relação tem de ser preservada.

Desta forma não poderíamos utilizar a habitual abordagem de cross-validation, seria necessário ter em conta uma abordagem de time series cross-validation.

A time series cross-validation teve por base a utilização de forward chaining, conhecido também como rolling-origin, onde os conjuntos de treino e teste vão avançando no tempo [64] [36].

Por forma a implementar a time series cross-validation foi utiizado a time series split cross-validation presente na biblioteca scikit-learn do Python [57].

A time series split cross-validation tem por objectivo efectuar a divisão do conjunto de dados para treino e validação a cada iteração com a garantia que o conjunto de validação será sempre posterior ao conjunto de treino.

De seguida encontra-se apresentado um gráfico que pretende ilustrar a separação dos conjunto de dados para treino e validação ao longo das diferentes iterações.

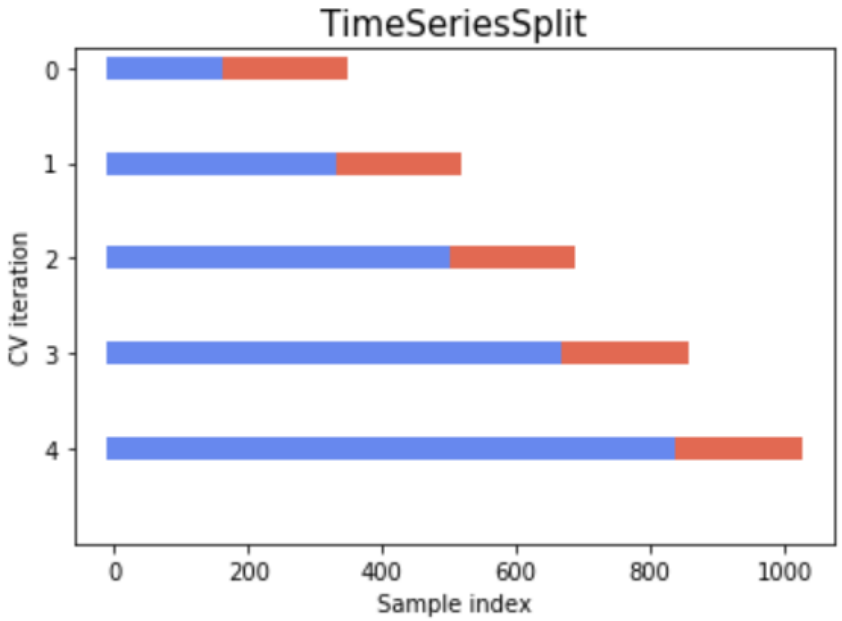


Figura 5.1: Time series com 5 iterações e 1000 amostras [36]

O gráfico apresentado ilustra a presença de 5 iterações (folds) e 1000 amostras, onde a cada iteração o conjunto de treino é assinalado a azul e o conjunto de validação assinalado a vermelho.

A cada iteração os resultados relativos ao desempenho do modelo, medidos de acordo com a métrica definida que será abordada adiante na sub secção 5.1.6 são guardados.

No fim é efectuado o cálculo do valor médio do desempenho do modelo obtido nas diferentes iterações, valor este que corresponde ao desempenho do modelo com os hiperparâmetros em questão.

A combinação de hiperparâmetros que permita ao modelo obter o desempenho mais alto é seleccionada para voltar a treinar o modelo em questão.

Todos os dados presentes no conjunto de treino são agora considerados unicamente para o treino do modelo.

Após o treino do modelo ser concluído é efectuado o cálculo do desempenho do mesmo no conjunto de teste.

Com o intuito de ilustrar as diferentes etapas seguidas ao longo da time-series cross-validation considerada no desenvolvimento do projecto em questão, encontram-se de seguida apresentadas as diferentes etapas da K-fold cross-validation como meio de comparação.

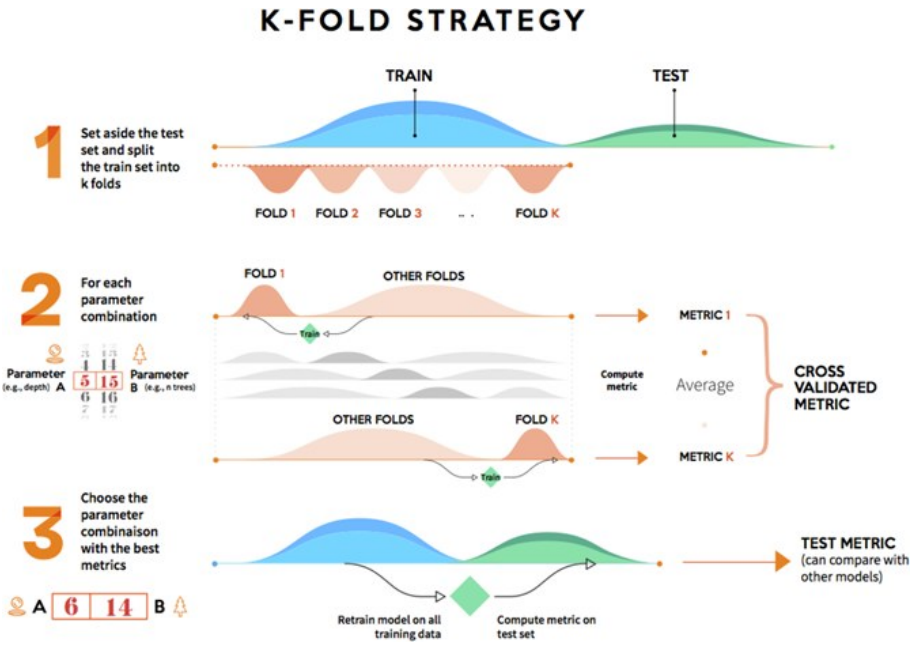


Figura 5.2: K-fold cross-validation [32]

A time-series cross-validation considerada no desenvolvimento do presente projecto segue as mesmas etapas que a K-fold cross-validation, com a diferença que na **etapa 2** as folds de treino e validação seguem a ordem cronológica dos acontecimentos, de modo a que a fold relativa à validação seja sempre posterior às folds de treino, tal como observado na figura 5.1.

5.1.4 Ajuste de hiperparâmetros

A definição dos hiperparâmetros de um modelo de Aprendizagem Computacional é um factor crucial na construção do mesmo.

A incorrecta definição dos hiperparâmetros de um modelo podem prejudicar de forma abrupta o desempenho de um modelo.

Existem diferentes técnicas para o ajuste dos hiperparâmetros de um modelo de Aprendizagem Computacional como Grid Search, Random Search e Bayesian Optimization.

No presente projecto foi considerada uma Bayesian Optimization para descobrir a melhor combinação de hiperparâmetros de cada um dos modelos considerados.

O primeiro passo a ser tomado no desenvolvimento da Bayseian Optimization foi a definição dos hiperparâmetros e dos diferentes valores referentes a cada um dos hiperparâmetros a serem considerados para análise.

Após os hiperparâmetros e os seus respectivos valores para análise se encontrarem devi-

damente definidos foi definida a função objectivo, função esta através da qual a Baysean Optimization avalia o desempenho de um modelo com determinados hiperparâmetros.

A função objectivo definida contempla uma time series cross-validation anteriormente abordada na sub secção 5.1.3.

Para cada fold da time series cross-validation é efectuada a avaliação do desempenho da combinação dos hiperparâmetros considerados para o modelo em questão, na classificação do conjunto de validação de acordo com a métrica adoptada no presente projecto 5.1.6, e o respectivo resultado é guardado.

Após o desempenho do modelo ser devidamente avaliado na classificação do conjunto de validação de cada uma das diferentes folds é efectuado o cálculo da média do desempenho do modelo, que é tido como o desempenho geral da combinação dos hiperparâmetros considerados para o modelo em questão.

Durante a execução da Baysean Optimization diferentes combinações de hiperparâmetros são avaliadas por forma a que no fim da mesma a melhor combinação de hiperparâmetros seja devolvida, que seja devolvida a combinação de hiperparâmetros que permita ao o modelo em causa obter os melhores resultados.

5.1.5 Undersampling

Uma vez que foi utilizado time series split cross-validation na divisão do conjunto de dados como referido anteriormente na secção 5.1.2 originou com que nas diferentes iterações o número de instâncias de ocorrências de incêndio e não ocorrências de incêndio fosse discrepante, com que o conjunto de dados em cada iteração não se encontrasse balanceado.

Tal facto levou a que no período de treino os modelos aprendessem a classificar a classe predominante com maior facilidade face à classe que apresentava menos instâncias e, consequentemente, o desempenho dos modelos no período de teste fosse pior como mais tarde iremos observar no capítulo 6.

Desta forma foi necessário efectuar um balanceamento dos dados de modo a garantir que o número de instâncias de ocorrências de incêndio e não ocorrências de incêndio se encontrava equilibrado e os modelos conseguiriam o melhor período de aprendizagem possível.

Para satisfazer tal propósito foram consideradas técnicas de under-sampling através da utilização da biblioteca imbalanced-learn do Python [30].

A biblioteca imbalanced-learn apresenta dois tipos de algoritmos diferentes para under-sampling, **prototype generation** e **prototype selection**.

Os diferentes tipos de algoritmos servem o mesmo propósito, reduzir o número de instâncias da classe predominante de forma a equilibrar o número de instâncias entre as duas classes, "... undersampling, that consists of reducing the data by eliminating examples belonging to the majority class with the objective of equalizing the number of examples of each class ..."[24].

A classe prototype generation efectua a redução do número de amostras através da geração de novas amostras de alguma forma representativos das existentes no conjunto de dados. Por outro lado, a classe prototype selection efectua a redução do número de amostras através da selecção de amostras do conjunto de dados.

Os algoritmos do tipo prototype generation podem ser ainda divididos em dois grupos,

controlled under-sampling techniques e **cleaning under-sampling techniques**.

O primeiro grupo permite a especificação por parte de utilizador do número de amostras a considerar. Em contrapartida o segundo grupo não permite tal especificação, é destinado a efectuar a limpeza do feature space, "Cleaning under-sampling techniques do not allow to specify the number of samples to have in each class. In fact, each algorithm implement an heuristic which will clean the dataset."[30].

No desenvolvimento do presente projecto foi considerada uma controlled under-sampling techniques, mais concretamente **RandomUnderSampler**, onde o número de instâncias de ocorrências de incêndio e não ocorrências de incêndio foi estabelecido como sendo igual ao número de instâncias com menor predominância.

5.1.6 Métrica utilizada

Como já referido anteriormente o presente projecto tem como principal objectivo a previsão da ocorrência de incêndios nos diferentes Concelhos de Portugal.

Por forma a verificar o desempenho dos diferentes modelos considerados no desenvolvimento do projecto em questão foi adoptada a métrica de **Area Under the Receiver Operating Characteristics Curve (AUC)**.

A Receiver Operating Characteristic (ROC) Curve é o gráfico que representa o desempenho de um modelo de classificação em diferentes thresholds de classificação.

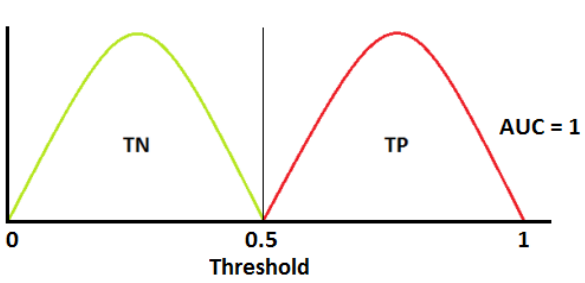
A ROC Curve é gerada com base no True Positive Rate (TPR), presente no eixo dos x, e com base no False Positive Rate (FPR), presente no eixo dos y, e representa o TPR vs FPR em diferentes thresholds de classificação.

A Area Under the ROC Curve representa a área que se encontra abaixo da ROC Curve e fornece uma medida de classificação em vários thresholds de classificação. Representa o grau de separação das classes, a capacidade que um modelo possui de distinguir entre as classes positiva (TP) e negativa (TN), que representam a ocorrência e não ocorrência de incêndio para o problema em questão.

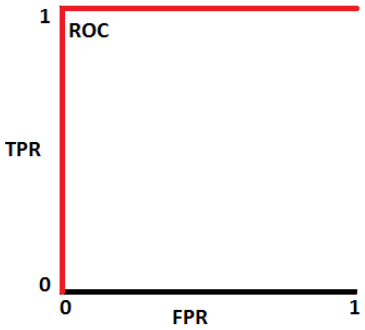
Quanto maior for o valor da AUC, maior capacidade terá o modelo de prever correctamente as classes.

Os valores relativos à AUC variam entre 0 e 1, onde um modelo do qual as suas previsões estão 100% erradas apresenta um valor de AUC de 0 e um modelo do qual as suas previsões estejam 100% correctas apresenta um valor de AUC de 1.

Modelos que apresentem um valor da AUC próximo de um apresentam uma excelente capacidade de separação das classes e, por outro lado, modelos que apresentem um valor da AUC igual a 0.5 denotam que o modelo não possui a capacidade de separação de classes.



(a) Separação das classes [65]



(b) ROC Curve [65]

Figura 5.3: Separação das classes e respectiva ROC Curve para uma AUC de valor 1

A escolha da métrica em questão teve por base a verificação da veracidade das previsões de ocorrência de incêndio efectuadas.

É desejável que o modelo desenvolvido tenha a capacidade de prever correctamente o maior número de ocorrências de incêndio mas também seja capaz de prever correctamente o maior número de não ocorrências de incêndio, que o modelo consiga apresentar uma boa capacidade de previsão de ocorrências de incêndio sem para isso apresentar demasiadas classificações falsas positivas (não ocorrências de incêndio classificadas como ocorrências de incêndio).

A métrica AUC permite que tal propósito seja satisfeito e permite também considerar o pior cenário existente para o problema em questão, a presença de classificações falsas negativas (classificar uma ocorrência de incêndio como não ocorrência de incêndio) na medida em que o número de classificações falsas negativas é considerado no denominador do quociente do True Positive Rate (TPR).

5.1.7 Coeficiente da correlação de Pearson

O coeficiente da correlação de Pearson, representado por **r**, é um indicador da relação linear entre duas variáveis.

A correlação de Pearson tenta traçar uma recta que melhor se ajusta aos dados das duas respectivas variáveis e o coeficiente da correlação de Pearson indica a distância a que se encontram os diferentes pontos de dados da recta traçada, o quão bem os pontos de dados se ajustam à reta traçada.

O cálculo do coeficiente da correlação de Pearson entre duas variáveis pode ser efectuado através da seguinte fórmula:

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Figura 5.4: Coeficiente da correlação de Pearson entre as variáveis X e Y [66]

Ao analisarmos a fórmula acima apresentada verificamos que o cálculo do coeficiente da correlação de Pearson é feito através da divisão da covariância pelo produto dos desvios padrão.

O coeficiente da correlação de Pearson pode variar entre -1 e 1 e assume 3 valores de referência:

- -1 - indica que estamos perante uma relação perfeitamente negativa e linear por parte das duas variáveis
- 0 - indica que as duas varáveis não apresentam qualquer relação linear
- 1 - indica que estamos perante uma relação perfeitamente positiva e linear por parte das duas variáveis

De seguida encontram-se apresentados graficamente os coeficientes da correlação de Pearson entre duas variáveis para os 3 valores de referências acima referidos:

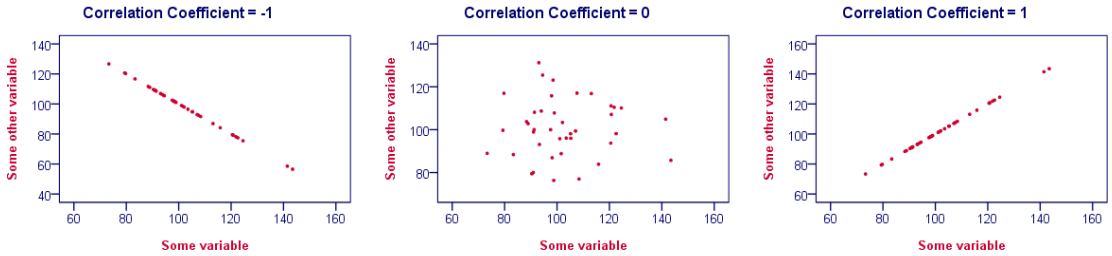


Figura 5.5: Coeficiente da correlação de Pearson -1, 0 e 1 [66]

5.1.8 Condições meteorológicas mais influentes nas ocorrências de incêndio

Para além da previsão da probabilidade de ocorrência de incêndios florestais nos diferentes concelhos de Portugal era também desejado que através de técnicas de Aprendizagem Computacional fosse possível saber as condições meteorológicas que haviam sido mais influentes em cada uma das ocorrências de incêndio históricas existentes.

A presente sub secção tem por objectivo descrever as diferentes etapas que foram seguidas de forma a que tal propósito fosse satisfeito.

Correlação das diferentes condições meteorológicas com as ocorrências de incêndio

Uma vez que pretendemos saber quais as condições meteorológicas mais influentes em cada uma das ocorrências de incêndio existentes, queremos saber em que medida as features meteorológicas se correlacionam com o target.

Desta forma foi desenvolvido um script em Python com o recurso à biblioteca Pandas que tinha por objectivo criar a matriz de correlação das diferentes features meteorológicas com o respectivo target.

A matriz de correlação foi criada de acordo com o coeficiente da correlação de Pearson, anteriormente abordado na sub secção 5.1.7, uma vez que estamos na presença de features numéricas onde não existe uma relação ordinal.

Normalização dos dados

Após conhecermos a correlação de cada uma das condições meteorológicas consideradas com o target em questão foi necessário proceder à normalização dos dados.

As condições meteorológicas consideradas apresentavam intervalos de valores distintos uma vez que representavam condições meteorológicas diferentes e medidas em unidades diferentes.

Por forma a que o intervalo de valores das condições meteorológicas consideradas não fosse um factor dominante no momento da verificação das condições meteorológicas mais influentes numa ocorrência de incêndio foi efectuada uma normalização nos dados de modo a que o valores das diferentes condições meteorológicas se encontrassem todos no mesmo intervalo.

Para tal propósito foi escolhida uma normalização min-max na qual o novo intervalo de cada condição meteorológica passaria a estar no intervalo [0.1 , 1].

O valor de cada condição meteorológica no novo intervalo podia ser obtido com recurso à seguinte fórmula:

$$novo\_valor = \frac{valor - minimo}{maximo - minimo} * (novo\_maximo - novo\_minimo) + novo\_minimo$$

Na fórmula acima apresentada **novo_minimo** e **novo_maximo** referem-se ao novo valor mínimo e máximo para o intervalo de valores da condição meteorológica em questão, respectivamente 0.1 e 1.

Por outro lado **minimo** e **maximo** representam o valor mínimo e máximo da condição meteorológica em causa e **valor** o valor da condição meteorológica que se pretende converter para o intervalo [novo_minimo,novo_maximo].

O processo de normalização em questão é aplicado a todas as condições meteorológicas de modo a que no fim deste processo todas as condições meteorológicas se encontrem no mesmo intervalo de valores, no intervalo [0.1 , 1].

Influência das diferentes condições meteorológicas

Após a correlação das diferentes features meteorológicas com o target ser conhecida e de as mesmas se encontrarem todas no mesmo intervalo de valores é então avaliada a influência que cada uma tem numa determinada ocorrência de incêndio.

De modo a que seja possível medir a influência de cada uma das diferentes featueres é multiplicado o valor correspondente à correlação da feature em questão pelo valor no novo intervalo da mesma feature. O valor resultante da multiplicação é então guardado para ser avaliado.

A avaliação é feita com base num valor de comparação estabelecido em 50% do valor da correlação da feature meteorológica em questão.

Uma vez que existem features que apresentam uma correlação positiva e features que apresentam uma correlação negativa, a avaliação efectuada tem por objectivo verificar se o valor resultante da multiplicação da correlação da feature em questão pelo novo valor da mesma feature é superior ou inferior a 50% do valor da correlação da respectiva feature.

Caso seja superior ou inferior a feature em questão é considerada como influente na respectiva ocorrência de incêndio.

A avaliação em questão é feita para todas as features meteorológicas consideradas e no final é obtido um array que contem as features meteorológicas mais influentes na ocorrência de incêndio em causa.

5.2 Web Application

A presente secção tem por objectivo apresentar as diferentes funcionalidades que foram desenvolvidas no presente projecto durante o período de estágio.

Como referido anteriormente no capítulo 3 na secção 3.6 as funcionalidades relativas ao back-end foram desenvolvidas em Python e as funcionalidades relativas ao front-end foram desenvolvidas em Angular.

5.2.1 Autenticação

O presente projecto visava assegurar que o conteúdo do back office fosse exclusivamente destinado a utilizadores devidamente autorizados.

Para tal efeito foi integrado um sistema de autenticação no presente projecto.

O sistema de autenticação escolhido foi a Amazon Cognito [58].

Através da utilização da Amazon Cognito foram asseguradas as seguintes funcionalidades:

- Login - Um utilizador devidamente registado insere o seu endereço de email e palavra-passe para aceder ao back office
- Logout - Um utilizador devidamente autenticado termina a sua sessão no back office
- Registar utilizador - Um utilizador é registado de modo a que o mesmo tenha acesso ao conteúdo exclusivo do back office
- Editar dados pessoais de um utilizador - São editados os dados pessoais pretendidos de um utilizador devidamente registado
- Redefinição da palavra-passe - É enviado um código para o endereço de email indicado para que posteriormente a palavra-passe possa ser reposta

Embora todas as funcionalidades acima apresentadas se encontrem implementadas e funcionais ao nível do servidor, foi apenas possível colocar funcional o ecrã de login na aplicação web.

De seguida encontra-se então apresentado o ecrã relativo ao login de um utilizador na web application.

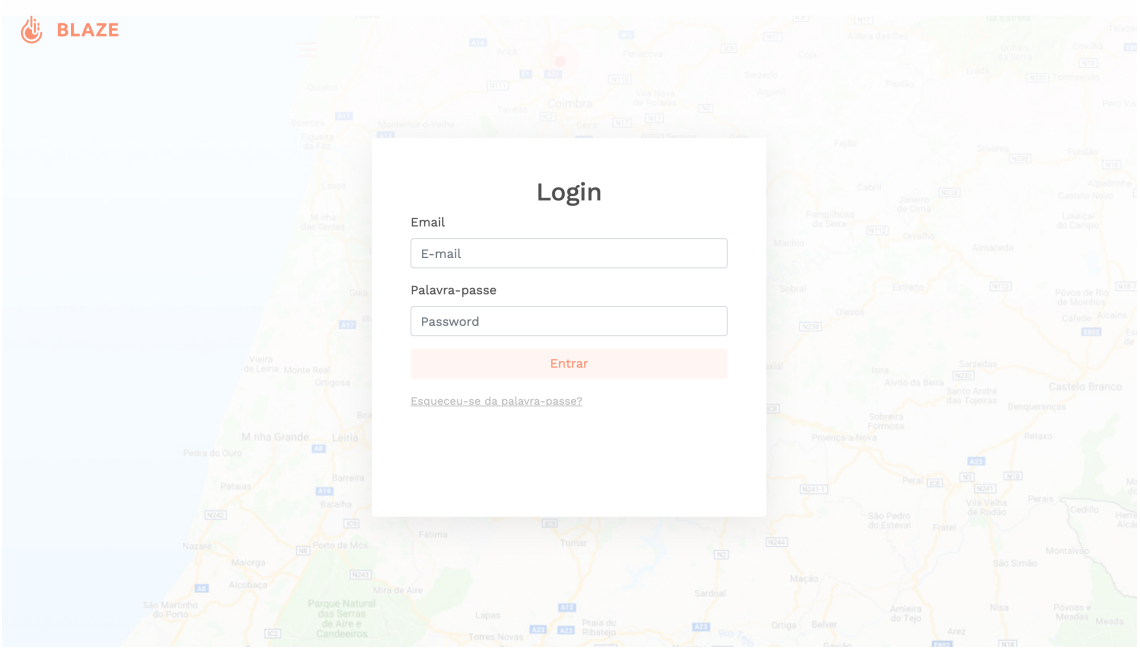


Figura 5.6: Ecrã de login

5.2.2 Carregamento de dados

Um dos requisitos funcionais levantados para a realização do presente projecto foi que fosse possível efectuar o carregamento de ficheiros com novos registos históricos por parte de um utilizador devidamente autorizado para tal através da web application.

Por forma a compreender o processo de carregamento de dados efectuado encontra-se de seguida apresentado um diagrama que pretende ilustrar as diferentes etapas deste mesmo processo.

De salientar que o diagrama de seguida apresentado se refere a um utilizador com permissões de acesso ao conteúdo do back office devidamente autenticado.

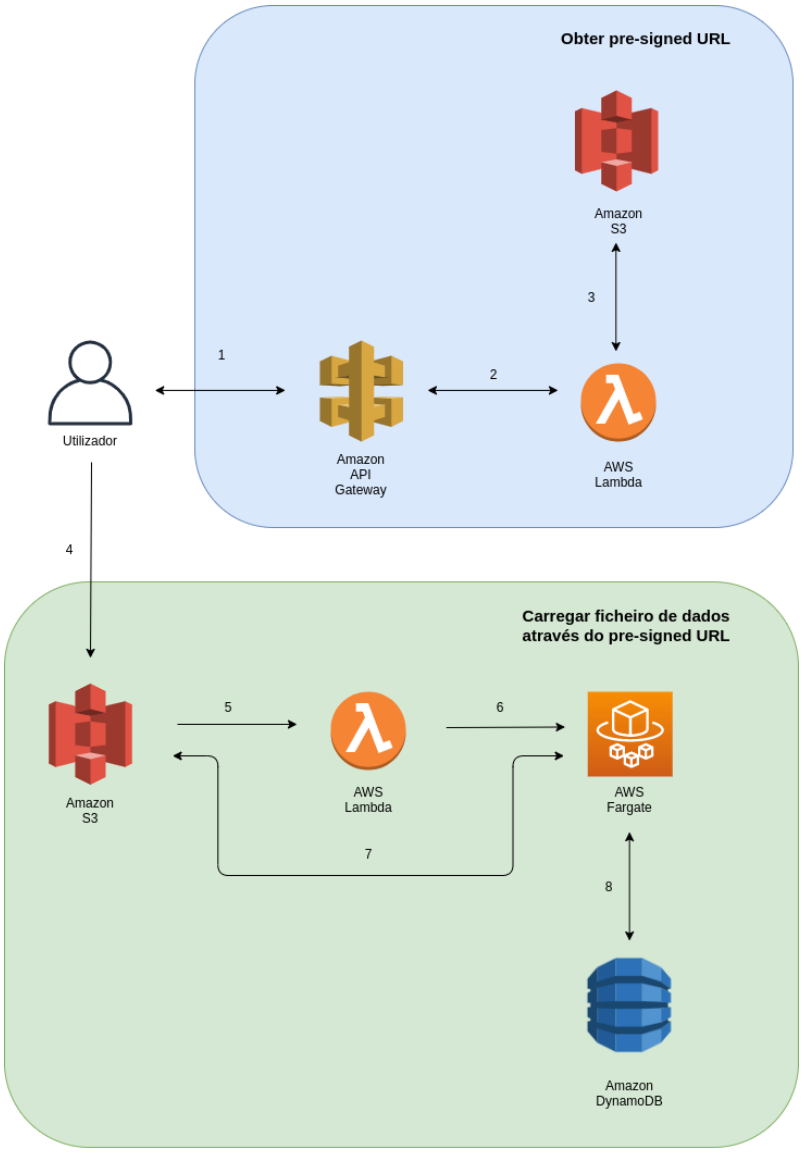


Figura 5.7: Diagrama de fluxo de carregamento de dados

Ao observarmos o diagrama anteriormente apresentado na figura 5.7 podemos identificar a presença de dois blocos principais no processo de carregamento de dados.

De modo a compreender cada uma das diferentes etapas do processo de carregamento de dados com uma maior clareza, encontra-se de seguida apresentada a descrição de cada uma destas etapas para cada um dos blocos anteriormente mencionados.

Obter pre-signed URL

O presente bloco, assinalado a azul no diagrama apresentado na figura 5.7, é o responsável por garantir ao utilizador um URL de acesso ao Amazon S3 Bucket com as devidas permissões.

De seguida encontram-se descritas cada uma das etapas da obtenção do pre-signed URL assinaladas no diagrama 5.7.

- 1 - É efectuado um pedido HTTP para o endpoint referente à obtenção de um pre-

signed URL;

- 2 - A AWS Lamda associada ao endpoint em questão é invocada;
- 3 - A AWS Lambda comunica com a Amazon S3 para solicitar um pre-signed URL para o Amazon S3 Bucket desejado.

No fim das 3 etapas presentes neste bloco se encontrarem concluídas o utilizador dispõe de um URL que lhe permite efectuar o carregamento ficheiros de dados no Amazon S3 Bucket desejado por um determinado período de tempo.

Carregar ficheiro de dados através do pre-signed URL

O bloco em questão, assinalado a verde no diagrama apresentado na figura 5.7, é o responsável for efectuar o carregamento do conteúdo de um ficheiro de dados na respectiva base de dados da Amazon DynamoDB.

Encontram-se de seguida descritas cada uma das diferentes etapas do carregamento de um ficheiro de dados assinaladas no diagrama 5.7.

- 4 - Através do pre-signed URL gerado anteriormente no bloco denominado de "Obter pre-signed URL"o utilizador efectua o carregamento do ficheiro de dados desejado directamente no Amazon S3 Bucket;
- 5 - Ao ser carregado um ficheiro de dados no Amazon S3 Bucket é accionada uma AWS Lambda que irá verificar qual o tipo do respectivo ficheiro de dados (dados de ocorrências de incêndio; históricas, dados de condições meteorológicas históricas ou dados populacionais)
- 6 - Consoante o tipo do ficheiro de dados carregado no Amazon S3 Bucket é lançado um AWS Fargate job que será responsável por efectuar ou não o carregamento do conteúdo do ficheiro de dados em causa;
- 7 - É efectuado o pedido ao Amazon S3 do ficheiro de dados previamente carregado e o mesmo é aberto para leitura no AWS Fargate job;

- 8
  - São solicitadas à Amazon DynamoDB a tabela onde se encontram guardados os registos já carregados e a tabela de metadados referentes ao tipo de dados em questão;
  - É efectuada a validação do conteúdo do ficheiro de dados a ser carregado por forma a verificar se existem conflitos de dados, se existem entradas de dados que já constam na base de dados;
  - Caso não seja verificado nenhum conflito de dados o conteúdo do ficheiro de dados a ser carregado é inserido na tabela da base de dados e o ficheiro é marcado como não tendo conflitos na tabela de metadados;
  - Caso sejam verificados conflitos de dados o ficheiro é marcado na tabela de metadados como tendo conflitos e nenhum do conteúdo do ficheiro de dados a ser carregado é inserido na tabela de dados.

De modo a que um utilizador devidamente autenticado pudesse efectuar o carregamento de dados através da web application foi desenvolvido o ecrã que se encontra de seguido apresentado.

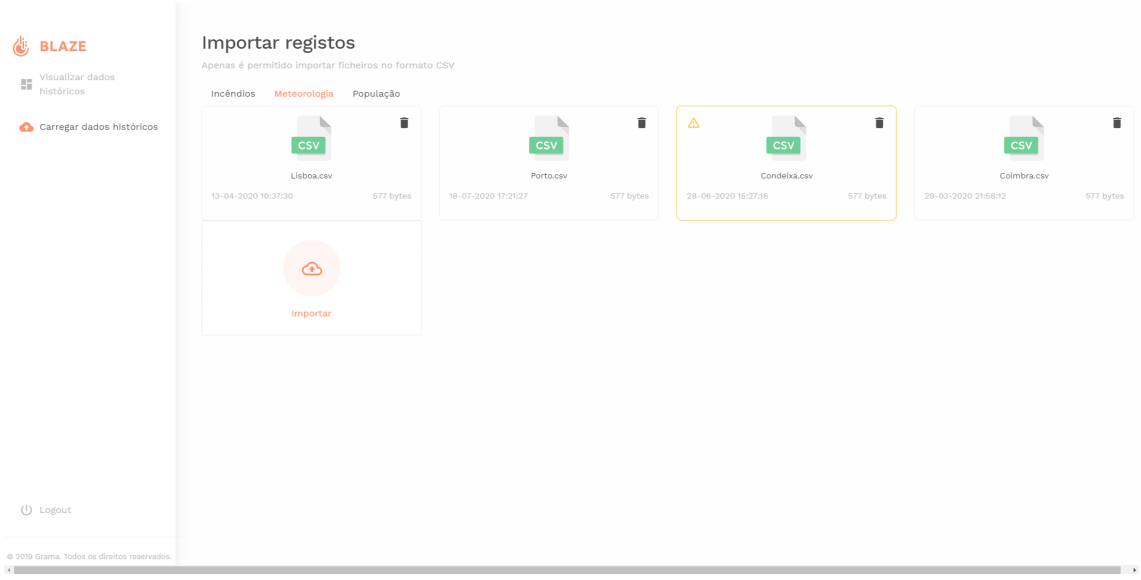


Figura 5.8: Ecrã relativo ao carregamento de dados

Na parte superior da página anteriormente apresentada verificamos a presença de três divisórias, **Incêndios**, **Meteorologia** e **População**, que correspondem aos três diferentes tipos de dados considerados no desenvolvimento do projecto em causa.

Assim, um utilizador devidamente autorizado para efectuar o carregamento de novos registos históricos selecciona em primeiro lugar a divisória referente ao tipo de dados que pretende carregar.

Após se encontrar na divisória desejada o utilizador deve assim clicar em **Importar** e será aberta uma nova janela para o utilizador seleccionar qual o ficheiro presente no seu computador que deseja carregar.

Após ser carregado, o ficheiro em questão aparecerá na divisória correspondente com o nome com que foi carregado. Se existirem conflitos com o novo ficheiro carregado, será

apresentado um sinal e uma borda amarela como apresentado na figura 5.8 para o ficheiro **Condeixa.csv**.

5.2.3 Integração do modelo de Aprendizagem Computacional na web application

Um dos requisitos estabelecidos para o desenvolvimento do projecto em questão era garantir a previsão de ocorrências para os diferentes Concelhos de Portugal através de um ambiente de web application.

Esta sub secção tem por objectivo retratar as diferentes etapas que foram seguidas de modo a satisfazer tal propósito.

Carregamento do modelo de Aprendizagem Computacional

Após ter sido identificado no capítulo 6 na sub secção 6.1 o modelo que apresentava o melhor desempenho foi necessário guardar tal modelo para utilização do mesmo na web application.

Para guardar o modelo em questão foi utilizado o módulo pickle do Python [45], onde o respectivo modelo foi assim serializado e guardado num ficheiro.

O ficheiro criado com recurso ao módulo pickle encontrava-se no formato SAV e foi guardado no bucket S3 para que posteriormente pudesse ser utilizado na web application ao nível da previsão de ocorrências de incêndios dos diferentes Concelhos de Portugal.

A utilização do modelo na web application será explicada mais adiante na sub secção 5.2.3.

Aplicação de previsão meteorológica

Uma vez que um dos pontos fulcrais no desenvolvimento do presente projecto era apresentar uma previsão da ocorrência de incêndios nos diferentes Concelhos de Portugal, era necessário que as condições meteorológicas previstas para um certo dia e um certo Concelho fossem conhecidas.

Deste modo foi necessário proceder à integração de uma aplicação de previsão meteorológica que permitisse satisfazer tal propósito.

Foram analisadas diferentes aplicações de previsão meteorológica por forma a verificar qual a mais adequada às necessidades do presente projecto. As aplicações que foram alvo de análise foram respectivamente OpenWeatherMap [39], Weatherbit [68], AccuWeather [8], Dark Sky [59], Weather2020 [67] e ClimaCell [10].

A escolha da aplicação em questão teve em conta diferentes factores como a capacidade de efectuar previsões até 5 dias para os diferentes Concelhos, integração com a linguagem de programação adoptada, e em especial o preço de utilização da aplicação.

Um factor bastante importante na escolha da aplicação foram os custos de utilização da mesma. Pretendia-se que a aplicação escolhida fosse totalmente gratuita e não apresentasse qualquer tipo de restrições no que diz respeito ao número de pedidos.

De seguida encontra-se apresentada uma tabela que ilustra as condições oferecidas por parte das aplicações meteorológicas referidas anteriormente.

API	Forecast Data	Historical Data	Current/Real-Time Weather	Pricing	Number of Free Requests
OpenWeatherMap	Yes	No	Yes	Free	Unlimited
Weatherbit	Yes	No	Yes	Freemium	150 per day
AccuWeather	Yes	No	Yes	Free Trial	50 per day
Dark Sky	Yes	Yes	No	Freemium	100 per month
Weather2020	Yes	No	No	Freemium	1000/day
ClimaCell	Yes	No	Yes	Freemium	100/day

Figura 5.9: Condições oferecidas por diferentes aplicações de previsão meteorológica [9]

A análise da tabela acima apresentada permite-nos observar que de entre as diferentes aplicações que foram comparadas apenas a aplicação OpenWeatherMap [39] oferece as condições desejadas, visto que é a única aplicação totalmente gratuita e sem restrições quanto ao número de pedidos.

Perante tal facto, a aplicação OpenWeatherMap [39] foi assim a aplicação escolhida para ser integrada na web application.

Features necessárias ao modelo

De modo a que o modelo considerado conseguisse efectuar as respectivas previsões da probabilidade de ocorrência de incêndio nos diferentes Concelhos de Portugal num ambiente de web application, era necessário fazer chegar ao modelo a informação das diferentes features consideradas.

A informação relativa às condições meteorológicas consideradas era garantida por parte da aplicação meteorológica anteriormente referida na sub secção 5.2.3.

Para que o modelo conseguisse obter a informação relativa às restantes features, mencionadas anteriormente na secção 4.4, foi necessário guardar tal informação num ficheiro CSV.

Desta forma foi desenvolvido um script em Python com recurso à biblioteca Pandas que tinha por objectivo atribuir a cada Concelho o valor correspondente já conhecido das diferentes features consideradas.

Uma vez que na web application a previsão estava dividida por distritos, onde o utilizador selecciona o distrito pretendido e recebe a previsão da probabilidade de ocorrência de cada um dos diferentes Concelhos do respectivo distrito, foi necessário atribuir também a cada Concelho o seu respectivo distrito.

Para tal efeito foi desenvolvido um script em Python com recurso às bibliotecas Pandas e geopy que tinha como intuito descobrir o distrito dos diferentes Concelhos com base nas suas coordenadas geográficas centrais latitude e longitude.

Após os distritos dos diferentes Concelhos estarem identificados, estes foram adicionados ao ficheiro CSV e o mesmo passou assim a contemplar as features de seguida apresentadas:

- Distrito
- Concelho
- Latitude do Concelho
- Longitude do Concelho
- Dimensão do Concelho
- Número de ocorrências de incêndio do Concelho relativas ao ano anterior
- Ranking de incêndio do Concelho relativo ao ano anterior
- Área ardida do Concelho no ano anterior
- Altitude mínima do Concelho face ao nível médio das águas do mar
- Altitude máxima do Concelho face ao nível médio das águas do mar

O respectivo ficheiro foi assim armazenado no bucket S3 para posterior utilização do modelo. Tal utilização será descrita na sub secção 5.2.3.

Previsão da probabilidade de ocorrência de incêndios florestais

De modo a que fosse possível utilizar o modelo implementado para efectuar a previsão da probabilidade de ocorrência de incêndios num ambiente de web application foi então desenvolvida uma lambda function.

A lambda function desenvolvida começa por receber o distrito e a data que haviam sido seleccionados por parte do utilizador na web application.

Sendo o distrito pretendido conhecido, a lambda function passa então a abrir para leitura o ficheiro CSV que contemplava as features necessárias ao modelo para cada Concelho, abordado anteriormente na sub secção 5.2.3, e a seleccionar as entradas cujo distrito era igual ao distrito que fora recebido por parte do utilizador, a seleccionar os Concelhos do respectivo distrito.

Após os Concelhos pertencentes ao distrito em questão se encontrarem seleccionados é feito recurso da API meteorológica, referida anteriormente na sub secção 5.2.3, para conhecer os valores referentes às condições meteorológicas de cada Concelho para a data indicada.

Na sua utilização a aplicação meteorológica começa por receber o nome do Concelho e a respectiva data para os quais se pretende conhecer as condições meteorológicas previstas.

Posto isto é efectuado um pedido http à API desejada, onde no url são enviadas a chave da API, o nome do Concelho desejado e as unidades em que se deseja receber os valores que resultam do pedido.

A API considerada oferece uma previsão das condições meteorológicas num intervalo de 5 dias, onde neste intervalo as medições das condições meteorológicas são efectuadas de 3 em 3 horas.

Tendo em consideração as 24 horas existentes num dia, a API em questão são apresenta assim 8 previsões diárias, às 00:00, 03:00, 06:00, 09:00, 12:00, 15:00, 18:00 e 21:00 horas.

De entre as previsões efectuadas maioritariamente verificam-se condições meteorológicas mais quentes e secas às 15:00 horas, pelo que as condições meteorológicas obtidas por parte da API e utilizadas para efectuar a previsão da probabilidade de ocorrência de incêndios de um dado Concelho tem em consideração as condições meteorológicas previstas para as 15:00 horas do dia pretendido.

Por forma a obter o valor das condições meteorológicas previstas para o dia desejado, a resposta proveniente do pedido http é convertida para um objecto JSON onde se encontram presentes as previsões efectuadas de 3 em 3 horas para cada um dos 5 dias abrangidos pela API, e são procurados no mesmo objecto JSON os valores referentes à data desejada, que são depois guardados.

Posto isto, foi efectuada a junção dos valores provenientes da resposta da API meteorológica com os valores das restantes features para cada Concelho. Cada Concelho dispunha agora de todas as features necessárias para que o modelo pudesse efectuar a previsão da sua respectiva probabilidade de ocorrência de incêndio.

Após os diferentes Concelhos disporem de todas as features necessária à previsão da probabilidade de ocorrência de incêndio por parte do modelo de Aprendizagem Computacional, tal modelo foi assim carregado.

Anteriormente na sub secção 5.2.3 observamos que o modelo em questão se encontrava serializado no bucket S3.

O modelo necessitaria agora de ser desserializado para a sua utilização fosse possível. Para tal efeito foi feito recurso do módulo pickle [45] do Python.

Após este procedimento o modelo encontrava-se apto a efectuar as previsões de ocorrência de incêndios, no entanto como se tratava de um classificador binário o mesmo apresentava a classificação sob a forma de classe positiva (ocorrência de incêndio) e classe negativa (não ocorrência de incêndio).

Um dos requisitos do presente projecto era que a respectiva previsão fosse apresentada sobre uma probabilidade de ocorrência de incêndio foi utilizada a função predic_proba da biblioteca scikit-learn [52] do Python, cujo resultado da sua utilização é expresso sob a forma de um array com dois valores, que representam respectivamente a probabilidade da entrada em questão pertencer à classe positiva (ocorrência de incêndio) ou negativa (não ocorrência de incêndio).

Dado que era pretendido efectuar uma previsão da probabilidade de ocorrência de incêndio, a previsão de cada Concelho efectuada por parte do modelo resultava do valor da probabilidade referente a esse mesmo Concelho pertencer à classe positiva, à probabilidade de nesse mesmo Concelho se verificar uma ocorrência de incêndio.

Posto todo este processo, são então fornecidas ao modelo as features referentes a cada um dos Concelhos do distrito em causa e são devolvidas as respectivas previsões da probabilidade de ocorrência de incêndios e a lamda function é terminada.

Previsão na web application

De forma a conhecer a probabilidade de ocorrência de incêndios florestais de uma determinada região através da aplicação web, o utilizador deve seleccionar o distrito e a data

pretendidos.

Para seleccionar o distrito o utilizador deve clicar na barra presente no canto superior esquerdo e escolher o distrito desejado.

Para seleccionar a data o utilizador deve clicar sobre a barra que se encontra na parte inferior do mapa presente no ecrã e escolher de entre os dias apresentados o dia desejado.

Quando o distrito e a data desejados se encontrarem seleccionados é assim efectuado o cálculo da probabilidade de ocorrência de incêndios florestais para a data em questão dos diferentes concelhos do distrito escolhido.

Após a probabilidade de ocorrência de incêndio relativa a cada um dos diferentes concelhos se encontrar calculada, tais probabilidades são então apresentadas ao utilizador.

De seguida encontra-se apresentado o resultado da previsão de ocorrência de incêndios florestais no distrito de Coimbra no dia 27 de Outubro de 2020.

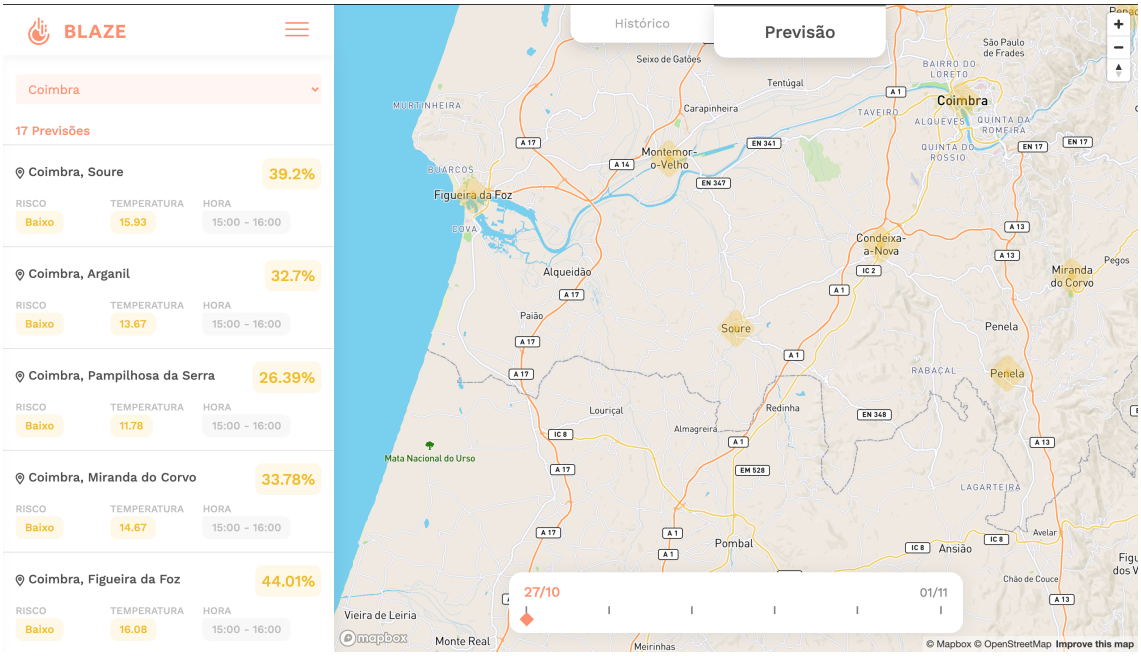


Figura 5.10: Ecrã relativo à previsão da ocorrência de incêndios florestais

5.2.4 Consulta de ocorrências históricas de incêndios florestais

Por forma a que fosse possível consultar as ocorrências históricas de incêndios florestais verificadas foi desenvolvida uma página web para satisfazer tal propósito.

As ocorrências históricas de incêndios florestais foram representadas no mapa de acordo com a origem de cada ocorrência.

De seguida encontra-se apresentado o ecrã que ilustra a representação das ocorrências históricas no distrito de Coimbra para o ano de 2008.

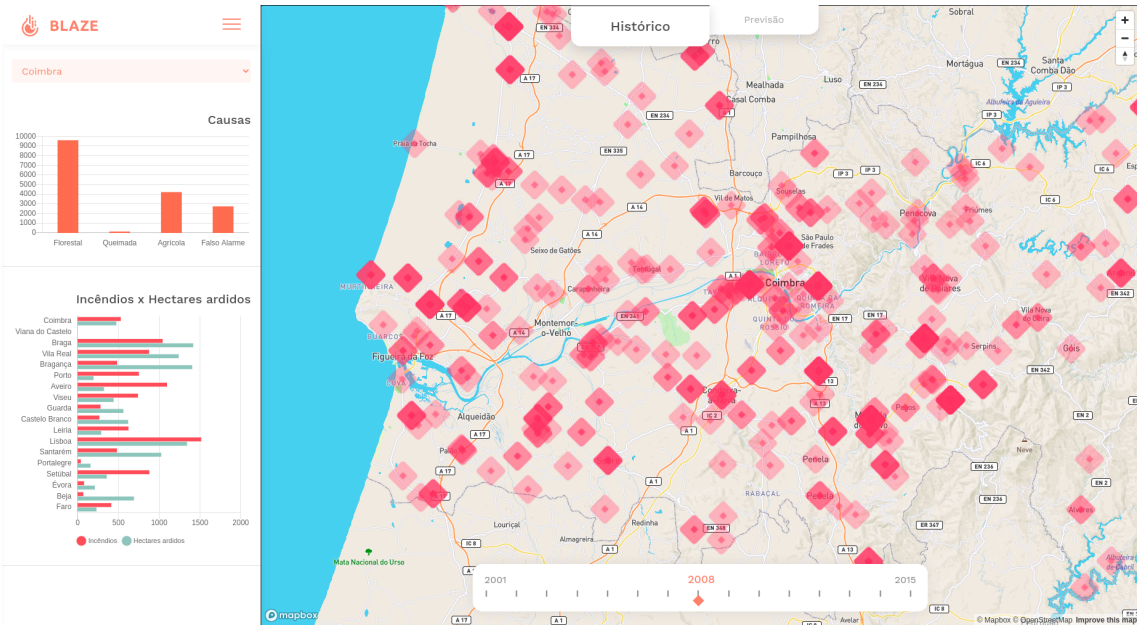


Figura 5.11: Ecrã relativo à consulta de ocorrências históricas de incêndios florestais

Para consultar outras ocorrências históricas de incêndios o utilizador deve seleccionar o distrito pretendido no canto superior esquerdo e/ou o ano pretendido na barra que se encontra na parte inferior do mapa do ecrã anteriormente apresentada.

Detalhes de uma ocorrência histórica de incêndio

Era pretendido no desenvolvimento do presente projecto que fosse possível consultar os detalhes de uma ocorrência de incêndio que havia sido verificada no passado como a data de ocorrência, o tipo de ocorrência ou até mesmo o número de hectares ardidos.

Para que tal fosse possível foi implementada uma interacção do cursor do rato sobre as ocorrências históricas assinaladas no mapa.

A passagem do cursor do rato sobre uma das ocorrências de incêndio históricas assinaladas no mapa faz com que seja apresentado um pop-up com os detalhes relativos à ocorrência em questão.

Condições mais influentes de uma ocorrência histórica de incêndio

Anteriormente na sub secção 5.1.8 foi descrita a forma de como através de técnicas de Aprendizagem Computacional foi possível verificar as condições meteorológicas que haviam sido mais influentes numa determinada ocorrência de incêndio histórica.

De forma a que fosse possível visualizar tais condições através da web application foi implementada uma interacção do cursor do rato sobre as ocorrências históricas assinaladas no mapa.

Através da passagem do cursor do rato sobre uma ocorrência de incêndio histórica assinalada no mapa são calculadas e apresentadas as condições meteorológicas que foram mais influentes na ocorrência do incêndio em questão.

Gráficos estatísticos

É ainda possível consultar na página web em questão alguma informação estatística calculada anualmente sobre a forma de gráficos.

Para cada ano seleccionado são calculados dois tipos de gráficos, um primeiro gráfico com informação relativa ao tipo de ocorrências históricas verificadas para o ano e distrito em questão, e um segundo gráfico com informação relativa ao número de ocorrências verificadas Vs. quantidade de área ardida para os diferentes distritos de Portugal no ano em questão.

Podemos observar no ecrã de seguida apresentado a página web implementada que integra o conteúdo abordado na sub secções 5.2.4, 5.2.4 e 5.2.4 de uma ocorrência de incêndio histórica registada no ano de 2008 no concelho de Penela, distrito de Coimbra.

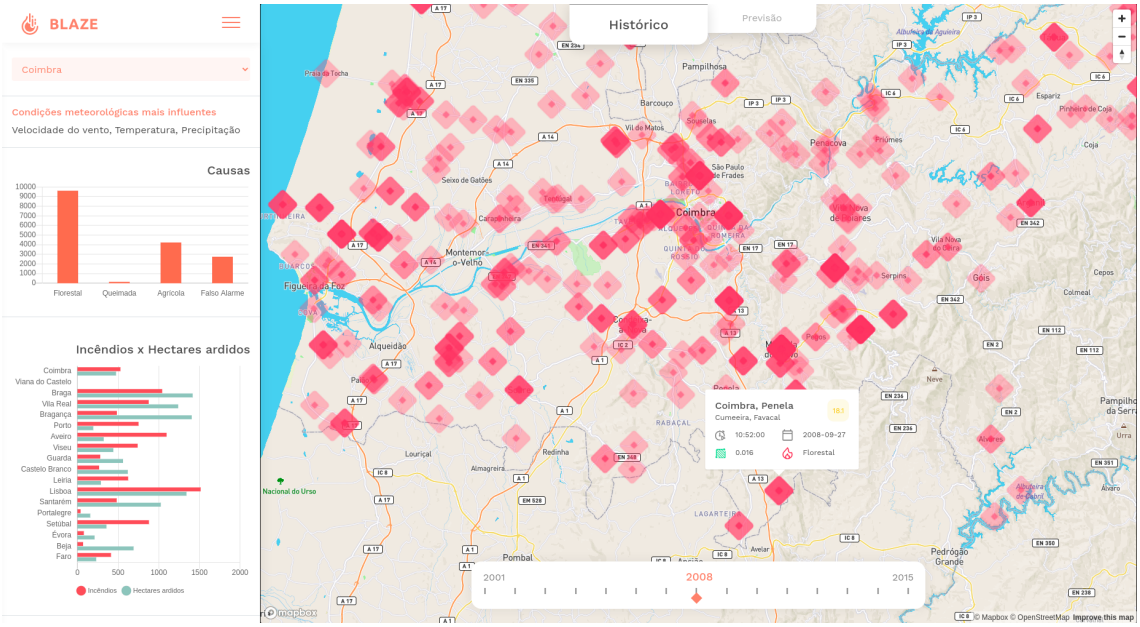


Figura 5.12: Ecrã relativo à consulta dos detalhes de uma ocorrência histórica de incêndio

5.2.5 Visualização dos dados carregados

Era pretendido que os dados carregados através do back office pudessem ser posteriormente consultados de modo a que fosse possível visualizar se toda a informação carregada se encontrava correcta e, no caso em que tal não se verificasse, fosse possível editar ou até mesmo apagar a respectiva informação.

De tal forma foi desenvolvida uma página web para satisfazer tal propósito, onde os dados são apresentados sob a forma de tabelas.

De seguida podemos observar a tabela que contempla os dados relativos às ocorrências de incêndio históricas.

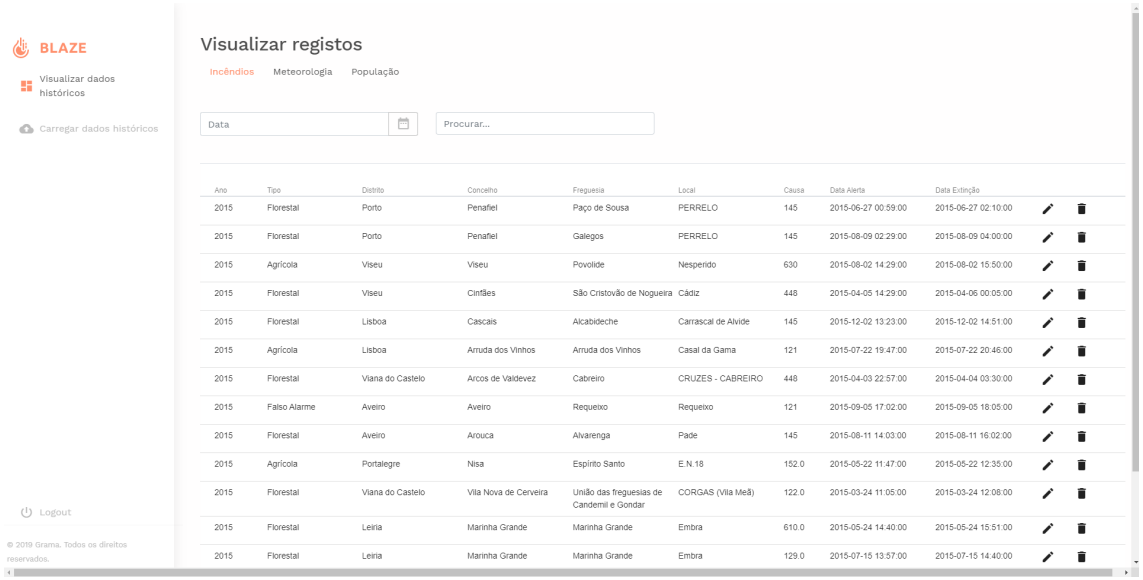


Figura 5.13: Ecrã relativo à visualização de dados

Para editar uma dada entrada de dados o utilizador deve clicar no botão com o ícone de um lápis ilustrado na figura 5.13 e o seguinte ecrã será apresentado.

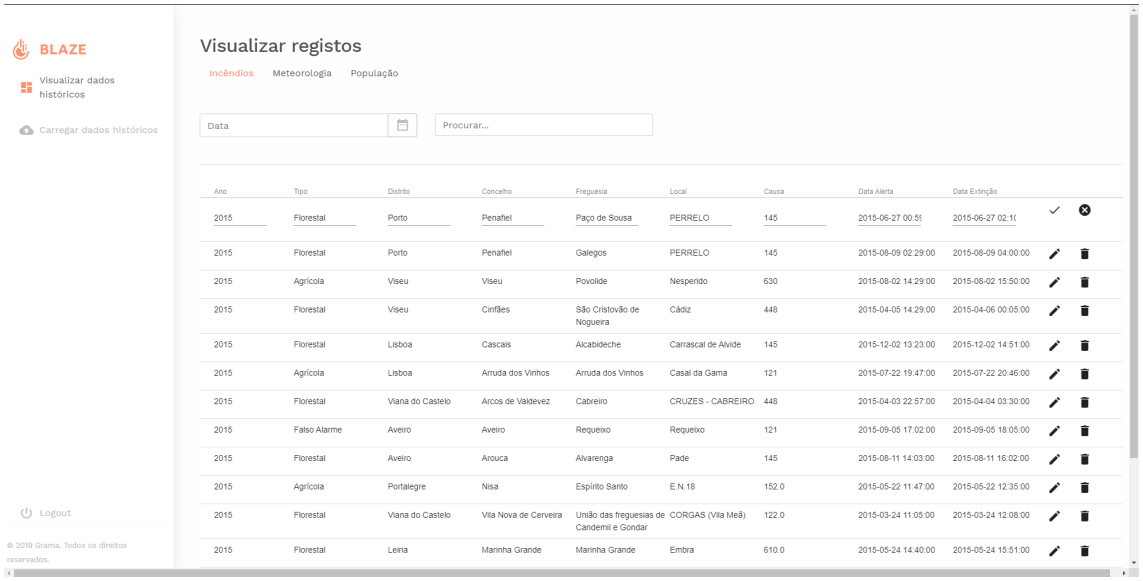


Figura 5.14: Ecrã relativo à edição de dados

Podemos verificar na imagem anteriormente apresentada que ao clicarmos no botão anteriormente mencionado, a entrada de dados seleccionada fica editável, permitindo ao utilizador alterar os campos desejados da respectiva entrada.

Após ter editado os campos pretendidos o utilizador deve clicar no botão com o ícone de um certo para guardar as alterações em causa na base de dados Amazon DynamoDb, ou clicar no botão com o ícone de uma cruz caso se tenha enganado e não pretenda editar a entrada de dados seleccionada.

De seguida podemos visualizar o ecrã destinado à edição de entradas de dados.

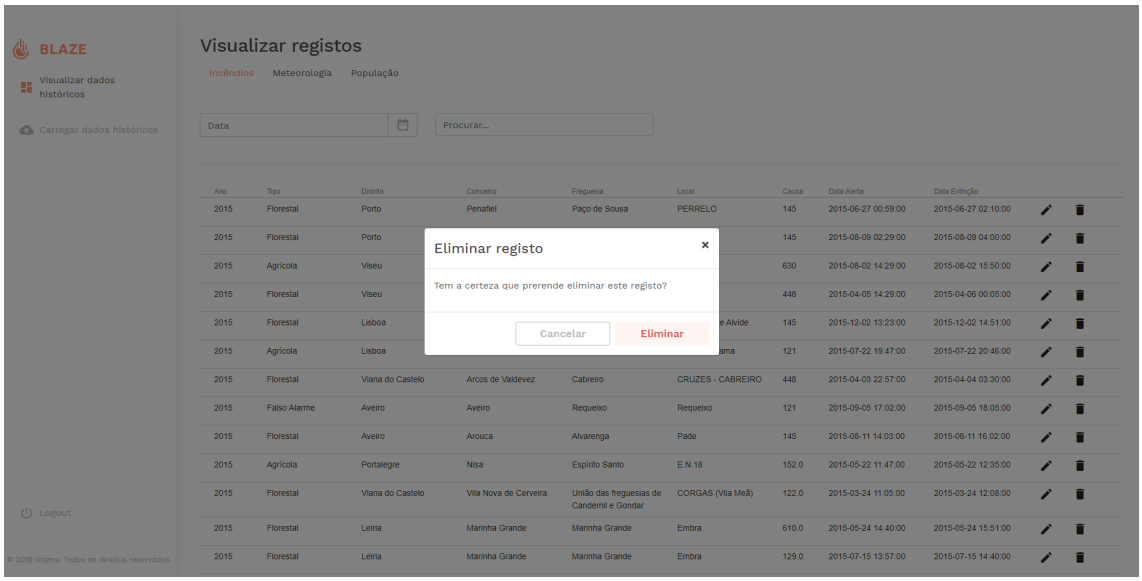


Figura 5.15: Ecrã relativo à eliminação de dados

Para eliminar uma determinada entrada de dados o utilizador deve clicar no botão com o ícone de um caixote do lixo ilustrado na figura 5.15 e o seguinte ecrã será apresentado.

Podemos verificar na imagem anteriormente apresentada que ao clicarmos no botão anteriormente mencionado surgirá uma janela de confirmação de modo a validar se o utilizador deseja efectivamente eliminar a entrada de dados em questão.

Caso o utilizador deseje efectivamente apagar a respectiva entrada de dados deve clicar no botão "Sim"e a entrada de dados será removida da base de dados.

Caso o utilizador se tenha enganado e pretenda reverter a opção de eliminar a respectiva entrada de dados deve clicar no botão "Não"e a entrada de dados permanecerá na base de dados.

Capítulo 6

Testes e resultados obtidos

O presente capítulo tem por objectivo apresentar os resultados obtidos por parte dos modelos de Aprendizagem Computacional desenvolvidos anteriormente apresentados no capítulo 5 na secção 5.1.1 bem como efectuar uma análise crítica destes mesmos resultados obtidos.

Um dos objectivos associados ao desenvolvimento do presente projecto era, para além de considerar o número de habitantes ao nível dos diferentes concelhos de Portugal na previsão da ocorrência de incêndios florestais, compreender de que forma este factor influência a ocorrência destes mesmos incêndios.

Uma vez que os modelos desenvolvidos não apresentavam um desempenho considerável na previsão da ocorrência de incêndios florestais, sem a consideração do número de habitantes ao nível dos diferentes concelhos de Portugal face à comparação com os resultados obtidos por parte dos autores considerados no estudo do estado da arte no capítulo 2. Assim sendo não foi possível retirar conclusões ao nível da influência do número de habitantes dos diferentes concelhos de Portugal na ocorrência de incêndios florestais.

O facto de os modelos desenvolvidos apresentarem dificuldades em efectuar a previsão da ocorrência de incêndios sem a consideração deste factor relativo ao número de habitantes, implica que através da consideração do mesmo não seja possível retirar conclusões a respeito da sua influência. Deste modo apesar de o número de habitantes por concelho de Portugal ser considerado como uma feature por parte dos modelos abordados neste capítulo, conforme anteriormente explicado não será apresentada a análise da influência desta mesma feature.

Assim sendo, serão em primeiro lugar apresentados e analisados neste capítulo os resultados relativos ao desempenho dos diferentes modelos obtidos no período de treino.

Posteriormente serão apresentados e analisados os resultados obtidos relativamente ao desempenho dos modelos obtidos no período de teste.

De seguida é efectuada a comparação dos resultados obtidos por parte dos modelos considerados face a outros modelos de Aprendizagem Computacional.

Por último mas não menos importante é efectuada a análise crítica a respeito de todos os resultados obtidos no presente capítulo.

6.1 Período de treino

Na presente secção será efectuada a análise dos resultados obtidos relativos ao desempenho dos modelos considerados durante o período de treino.

Os resultados obtidos encontram-se apresentados em A em duas secções, uma secção relativa ao modelo Support Vecto Machine e uma secção relativa ao modelo Random Forest. Em cada uma destas secções encontra-se apresentada a informação relativa ao desempenho do respectivo modelo de acordo com os diferentes hiperparâmetros no período de treino.

Será efectuada em primeiro lugar uma análise global sobre os resultados obtidos por parte dos diferentes modelos e posteriormente a análise das combinações de hiperparâmetros que permitiram a obtenção dos melhores resultados relativos ao desempenho de cada um dos modelos desenvolvidos.

Os resultados obtidos resultam do ajuste de hiperparâmetros abordado anteriormente na secção 5.1.4 do capítulo 5, mais concretamente da utilização de optimização Bayesiana, integrada na time-series cross-validation abordada também anteriormente na secção 5.1.3 do capítulo 5.

Os hiperparâmetros considerados bem como o seu intervalo de valores encontra-se para cada algoritmo apresentado na secção 5.1.1 do capítulo 5, e foram consideradas 100 avaliações no que diz respeito ao processo de ajuste de hiperparâmetros efectuado para cada um dos algoritmos considerados.

Os resultados que foram obtidos por parte deste processo podem ser consultados em A.

6.1.1 Resultados obtidos durante o período de treino

Ao consultarmos as tabelas relativas ao desempenho dos diferentes modelos desenvolvidos durante o período de treino em A verificamos que em cada tabela se encontra apresentada uma linha a vermelho e uma linha a verde.

A linha assinalada a verde representa a combinação de hiperparâmetros que permitiu a obtenção do melhor desempenho registado. Em oposição a linha assinalada a vermelho representa a combinação de hiperparâmetros que obteve o pior desempenho registado.

A análise das diferentes tabelas mostra-nos que entre as diferentes combinações de hiperparâmetros de cada algoritmo os resultados relativos ao desempenho das diferentes combinações são muito próximos, não se verifica uma diferença considerável ao nível do desempenho alcançado por parte das diferentes combinações de hiperparâmetros.

Relativamente aos resultados obtidos por parte do algoritmo Support Vector Machine verificamos que o intervalo de valores se encontra compreendido entre os 58.83% e os 59.34%.

Por outro lado, relativamente aos resultados obtidos por parte do algoritmo Random Forest verificamos que o intervalo de valores se encontra compreendido entre os 57.66% e os 60.39%.

Tais resultados obtidos ficam um pouco aquém das expectativas inicialmente previstas. Uma vez que os resultados obtidos resultam de uma time-series cross validation com um número de folds igual a **5** tais resultados demonstram que os modelos construídos não conseguiram obter uma capacidade de generalização notável face a novos conjuntos de dados.

A dificuldade de generalização dos modelos durante o período de treino demonstra que tais modelos apresentam dificuldades de classificação face a novos conjuntos de dados, factor este que compromete de certo modo a qualidade dos modelos construído na previsão de futuros dados de ocorrências de incêndios florestais.

6.1.2 Melhores combinações de hiperparâmetros

A presente sub secção tem por objectivo verificar qual a melhor combinação de hiperparâmetros de cada um dos algoritmos considerados, de modo a que tal combinação seja considerada na construção do modelo final de cada um dos respectivos algoritmos.

De modo a descobrir a melhor combinação de hiperparâmetros para cada um dos diferentes algoritmos considerados foi efectuada a análise das combinações de hiperparâmetros que permitiram obter os melhores desempenhos durante o período de treino.

Esta análise foi efectuada para cada um dos algoritmos considerados e encontra-se de seguida apresentada.

Support Vector Machine

De modo a verificar quais as combinações de hiperparâmetros que apresentavam resultados de desempenho próximos dos resultados obtidos por parte do melhor modelo (59.34%), foram verificadas as combinações que apresentavam valores superiores a 59.33%.

As combinações de hiperparâmetros que conseguiram obter um desempenho superior a 59.33 % encontram-se de seguida apresentadas.

C	dual	loss	max_iter	penalty	tol	Performance
100.0	False	squared_hinge	5000	l1	0.0011	0.5933253995102716
0.1	False	squared_hinge	4000	l1	0.0081	0.59336605843485
100.0	False	squared_hinge	7500	l1	0.0051	0.5933217501370198
1000.0	False	squared_hinge	4000	l1	0.0021	0.5934416457438745
1000.0	False	squared_hinge	4000	l2	0.0021	0.5933431047302111
10.0	False	squared_hinge	1500	l1	0.0021	0.5933901899658606
10.0	False	squared_hinge	1000	l1	0.0021	0.5933731613125957
1.0	False	squared_hinge	1500	l1	0.0021	0.5933953783710617

Tabela 6.1: Melhores resultados do algoritmo Support Vector Machine

A análise da tabela anteriormente apresentada mostra-nos que existem soluções que apresentam valores diferentes relativos aos hiperparâmetros considerados mas resultados muito semelhantes entre si.

Dado que não se verificou uma diferença consideravelmente significativa por parte das diferentes combinações de hiperparâmetros que permitiram alcançar os melhores desempenhos, tais combinações foram então avaliadas por forma a compreender qual a combinação de hiperparâmetros a ser considerada.

Ao analisarmos o hiperparâmetro **max_iter** verificamos que foi possível obter resultados muito semelhantes com um número baixo de iterações e com um número elevado de iterações. Tal facto remete que o número de iterações não é um factor condicionante ao

desempenho dos modelos e que a utilização de um valor baixo relativo a este hiperparâmetro nos permite obter níveis de desempenho praticamente iguais. Assim sendo podemos descartar os valores de **max_iter** superiores a **4000** dos candidatos à melhor combinação de hiperparâmetros.

De entre as combinações com um valor do hiperparâmetro **max_iter** inferior a **4000** podemos observar que entre as mesmas o hiperparâmetro **tol** assume sempre o valor de **0.0021**, valor este que é considerado em **5** das **8** melhores combinações aqui apresentadas e nos pode indicar que a utilização de uma tolerância para o critério de paragem igual a 0.0021 origina bons resultados por parte dos modelos.

Relativamente à função de penalização podemos verificar também que de entre as combinações com um hiperparâmetro **max_iter** inferior a **4000** todas utilizam a função **l1**, assim como **7** das **8** melhores combinações apresentadas, o que nos demonstra com clareza que a consideração da função de penalização **l1** permite a obtenção de melhores níveis de desempenho.

No que diz respeito ao parâmetro de regularização, hiperparâmetro **C** verificamos a presença de dois valores para as combinações com um hiperparâmetro **max_iter** inferior a **4000**, sendo eles **1** e **10**.

O hiperparâmetro **C** é hiperparâmetro responsável por evitar classificações incorrectas durante o período de treino através da escolha das margens associadas ao hiperplano. A atribuição de um valor de **C** alto resulta em que seja considerada uma margem menor no período de treino e respectivamente a consideração de um valor de **C** baixo resulta em que seja atribuída uma margem maior, mesmo que tal hiperplano classifique mais pontos erradamente.

Sendo o objectivo da Support Vector Machine apresentar a melhor classificação possível sobre o conjunto de dados, um dos paradigmas presentes neste algoritmo é a procura de uma margem que potencialize a separação máxima entre as classes mas que também garanta a maior classificação correcta dos diferentes pontos através desta mesma margem.

O facto de para um valor de **C** igual a **1** e um valor de **C** igual a **10** o desempenho ser praticamente igual leva-me assim a escolher o valor de **C** igual **1** uma vez que me permite uma maior separação entre as classes e um nível de desempenho idêntico.

Na tabela 6.1 encontra-se assinalado a vermelho a melhor combinação de hiperparâmetros obtida em termos de valor de desempenho e a verde encontra-se aquela que após fruto de análise foi definida como a melhor combinação para o modelo Random Forest.

Random Forest

De modo a verificar quais as combinações de hiperparâmetros que apresentavam resultados de desempenho próximos dos resultados obtidos por parte do melhor modelo (60.39%), foram verificadas as combinações que apresentavam valores superiores a 60.33%.

As combinações de hiperparâmetros que conseguiram obter um desempenho superior a 60.33 % encontram-se de seguida apresentadas.

criterion	max_depth	max_features	min_samples_split	n_estimators	Performance
gini	24	8	0.01	850	0.6038598499171769
gini	24	5	0.01	100	0.6035185947325737
gini	24	5	0.01	100	0.6033069174815815
gini	22	5	0.01	850	0.6035202969313402
gini	21	log2	0.01	850	0.6034900689495607
entropy	23	5	0.01	950	0.6033659140369159
entropy	24	7	0.01	900	0.6036668776252145
entropy	24	7	0.01	150	0.6037048296883133
entropy	23	7	0.01	450	0.6039353765077446

Tabela 6.2: Melhores resultados do algoritmo Random Forest

A análise da tabela anteriormente apresentada permite-nos observar que existem soluções muito próximas à melhor solução obtida.

Uma vez que a diferença de desempenho em função dos diferentes hiperparâmetros não pode ser considerada significativa devido ao seu baixo valor, foram então analisadas as combinações que obtiveram os melhores resultados com o intuito de avaliar qual o melhor modelo de Random Forest a ser considerado.

Pela análise da tabela anterior podemos verificar que em todas as combinações seleccionadas o hiperparâmetro **min_samples_split** toma sempre o valor de 0.001, pelo que podemos concluir para obter os melhores resultados deverá ser feito o split dos nós quando existirem 1% dos dados em cada lado (aproximadamente 3200 dos dados).

Podemos verificar que relativamente ao hiperparâmetro **n_estimators** os valores se encontram compreendidos entre 150 e 950. Tal facto indica-nos que modelos com um baixo número de árvores de decisão conseguem alcançar resultados idênticos a modelos com um número de árvores de decisão elevados e que o número de árvores de decisão não representa um factor influente na obtenção dos melhores resultados.

Deste modo dentro daquele que é o conjunto das melhores combinações de hiperparâmetros podemos ainda excluir as combinações que apresentam um valor alto par o hiperparâmetro **n_estimators**. Podemos estabelecer este valor em 500 visto que é o valor intermédio do conjunto de valores deste hiperparâmetro.

Das combinações que não são excluídas verificamos que apresentam valores de 5 e 7 relativos ao hiperparâmetro **max_depth**. Apesar de estarmos perante um hiperparâmetro que representa o número de features consideradas para verificar o melhor split, a documentação da biblioteca scikit-learn [52] indica que a procura por um split não termina até que seja encontrada uma partição válida das amostras.

Desta forma não é possível garantir que o número de features para verificar o melhor split seja respeitado e por essa mesma razão não o considerarei como factor de decisão na escolha da melhor combinação de hiperparâmetros.

Restando os hiperparâmetros **criterion** e **max_depth** verificamos que relativamente ao hiperparâmetro **criterion** tanto a função **gini** como a função **entropy** para efectuar a medição da qualidade dos splits apresentam um bom desempenho. Verificamos também que o valor do hiperparâmetro **max_depth** é igual, pelo que não pode ser considerado também para desempate.

Assim sendo o hiperparâmetro **n_estimators** foi considerado como factor de desempate

entre as duas combinações em causa.

Na tabela 6.2 encontra-se assinalado a vermelho a melhor combinação de hiperparâmetros obtida em termos de valor de desempenho e a verde encontra-se aquela que após fruto de análise foi definida como a melhor combinação para o modelo Random Forest.

6.2 Período de teste

Anteriormente na sub secção 6.1 foram analisados os resultados obtidos por parte das diferentes combinações de hiperparâmetros por forma a descobrir a melhor combinação para cada um dos algoritmos considerados.

A presente secção tem por objectivo considerar o conjunto de hiperparâmetros que permitiu a cada um dos diferentes algoritmos obter os melhores resultados durante o período de treino, e considerar esses mesmos hiperparâmetros na construção dos modelos finais de cada um dos algoritmos considerados.

Tal como abordado anteriormente no capítulo 5 na secção 5.1.3, após a combinação de hiperparâmetros que permitiu a cada algoritmo obter o melhor desempenho no período de treino ser descoberta, é construído um novo modelo com os respectivos hiperparâmetros do algoritmo em causa que é treinado considerando todos os dados presentes no conjunto de treino exclusivamente para treino.

Por forma a ilustrar os resultados obtidos por parte dos melhores modelos de cada algoritmo anteriormente apresentados na secção 6.1 foram construídos os gráfico relativos à ROC curve e à matriz de confusão com recurso às bibliotecas matplotlib e seaborn do Python.

De seguida encontram-se apresentados os respectivos gráficos construídos para os melhores modelos de cada um dos algoritmos considerados.

6.2.1 Support Vector Machine

De acordo com a análise dos hiperparâmetros que permitiram a obtenção dos melhores resultados por parte do algoritmo Support Vector Machine, efectuada na sub secção 6.1.2 do presente capítulo verificámos que a combinação de hiperparâmetros escolhida foi respectivamente:

- **C**: 1.0
- **dual**: False
- **loss**: squared_hinge
- **max_iter**: 1500
- **penalty**: l1
- **tol**: 0.0021

De seguida encontram-se apresentados em primeiro lugar o gráfico relativo à matriz de confusão da classificação do modelo do algoritmo Support Vector Machine com os hiperparâmetros anteriormente apresentados e posteriormente o gráfico da sua ROC curve.

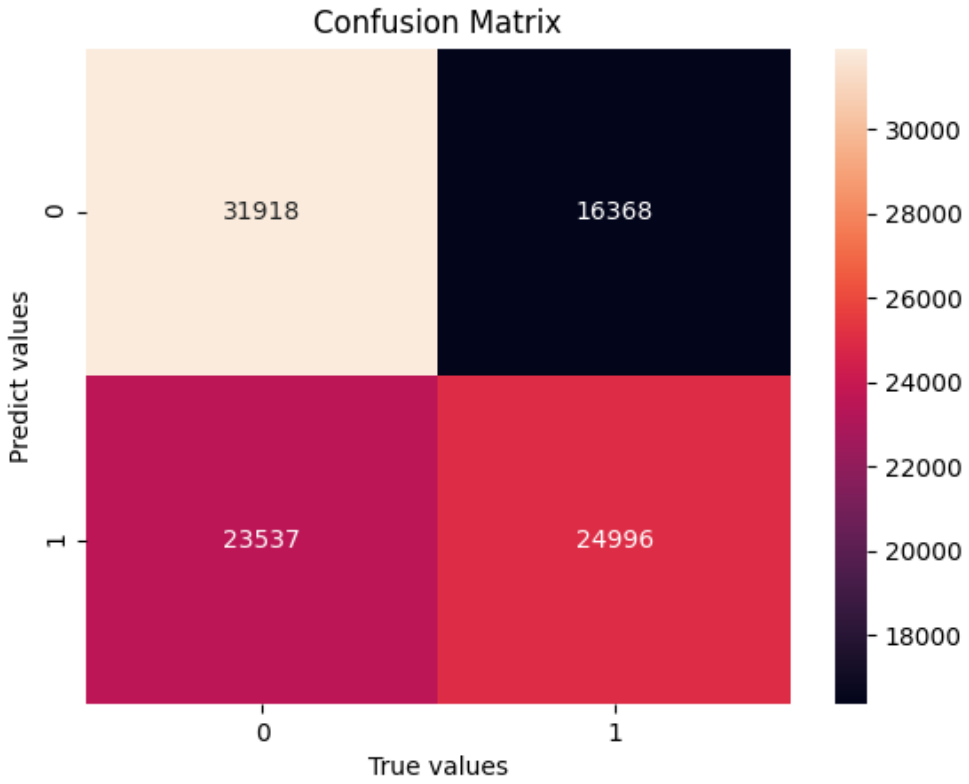


Figura 6.1: Matriz de confusão do modelo do algoritmo Support Vector Machine com os hiperparâmetros escolhidos

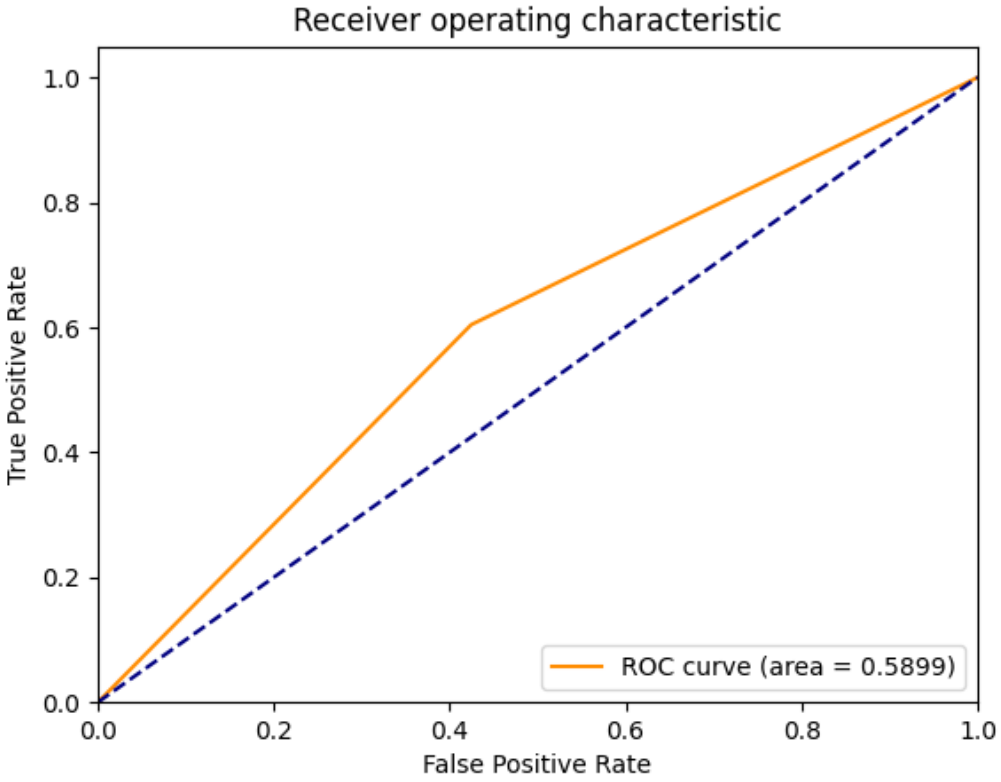


Figura 6.2: ROC curve do modelo do algoritmo Support Vector Machine com os hiperparâmetros escolhidos

Ao observarmos os gráficos anteriormente apresentados podemos observar na figura 6.1 a respectiva matriz de confusão de classificação do modelo face ao conjunto de teste, e na figura 6.2 observar o desempenho do modelo medido com a métrica Area Under the Receiver Operating Characteristics Curve (AUC).

Ao compararmos o desempenho do modelo no período de treino e de teste verificamos que o modelo piorou o seu desempenho, passando de um desempenho de 59.33% obtido durante o período de treino para um desempenho de 58.99% obtido no período de teste.

Uma vez que o desempenho obtido durante o período de teste face ao período de treino representa em módulo uma diferença inferior a 1% tal facto não nos permite afirmar que o modelo construído não apresentou a devida capacidade de classificação face ao conjunto de dados de teste, o modelo pode simplesmente ter tido mais dificuldades em prever determinadas ocorrências de incêndio que constavam no conjunto de teste e eram desconhecidas por parte do modelo durante o seu treino.

A análise da matriz de confusão apresentada na figura 6.1 demonstra-nos que um dos objectivos da utilização da métrica considerada Area Under the Receiver Operating Characteristics Curve (AUC) e abordada em 5.1.6 foi cumprido.

Ao analisarmos a respectiva matriz de confusão verificamos que o modelo apresenta mais instâncias classificadas de forma correcta. Apresenta respectivamente 31918 **Verdadeiros Negativos**, 24996 **Verdadeiros Positivos**, 16368 **Falsos Negativos** e 23537 **Falsos Positivos**.

Para além de o modelo apresentar mais instâncias classificadas de forma correcta verifi-

camos que o número de **Falsos Negativos** é o mais baixo aquando comparado com os outros. Tal facto demonstra que o modelo, de acordo com a métrica considerada, e tal como desejado, promove a não ocorrência de **Falsos Negativos**, que no problema em questão representam a situação de o modelo dizer que não irá ocorrer um incêndio e o respectivo incêndio se verificar.

No que diz respeito ao número de **Falsos Positivos** verificamos que o número apresentado é elevado quando comparado com o número de **Verdadeiros Positivos**. Observamos que o modelo apresenta um número de **Falsos Positivos** igual a 23537 e um número de **Verdadeiros Positivos** igual a 24996.

A pequena diferença entre estes valores remete-nos para que numa previsão efectuada por parte do modelo em que irá ocorrer um incêndio haja quase uma igual probabilidade de o incêndio se verificar ou ser um falso alarme.

Tal facto denota alguma incapacidade de classificação do modelo do que são ocorrências ou não ocorrências de incêndio. No entanto o modelo prioriza a ocorrência de **Falsos Positivos** face a **Falsos Negativos**, o que no problema em questão é mais aceitável visto que é preferível prever a ocorrência de um incêndio que depois não se verifique, do que a não ocorrência de um incêndio que posteriormente se venha a verificar.

6.2.2 Random Forest

De acordo com a análise dos hiperparâmetros que permitiram a obtenção dos melhores resultados por parte do algoritmo Random Forest, efectuada na sub secção 6.1.2 do presente capítulo verificámos que a combinação de hiperparâmetros escolhida foi respectivamente:

- **criterion**: gini
- **max_depth**: 24
- **max_features**: 5
- **min_samples_split**: 0.01
- **n1_estimators**: 100

De seguida encontram-se apresentados em primeiro lugar o gráfico relativo à matriz de confusão da classificação do modelo do algoritmo Random Forest com os hiperparâmetros anteriormente apresentados e posteriormente o gráfico da sua ROC curve.

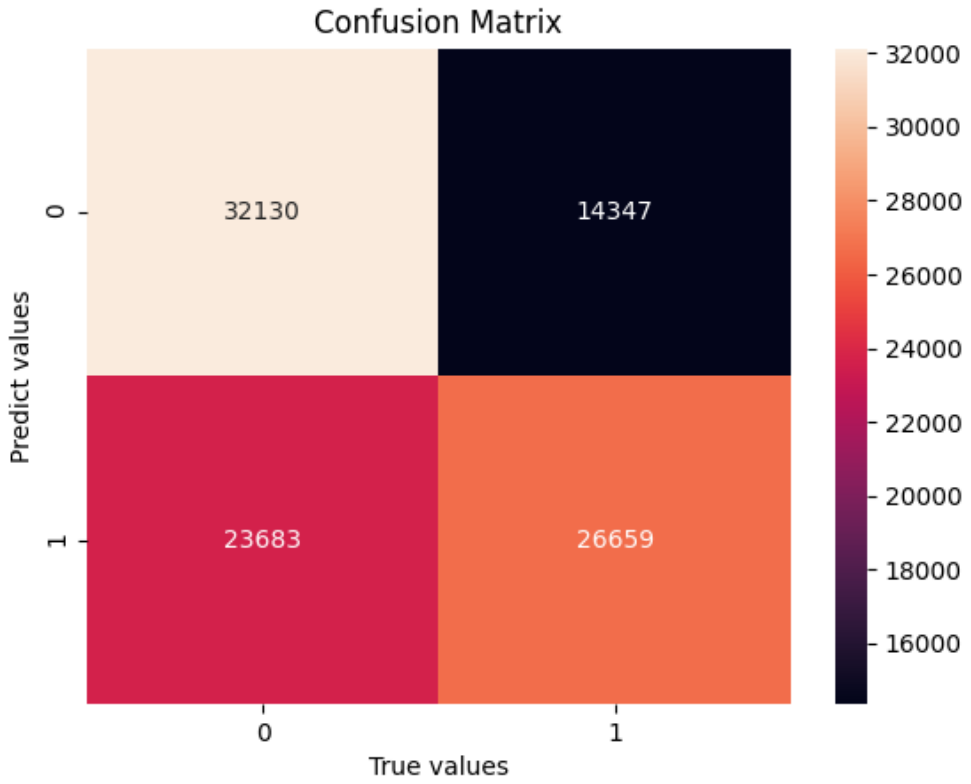


Figura 6.3: Matriz de confusão do modelo do algoritmo Random Forest com os hiperparâmetros escolhidos

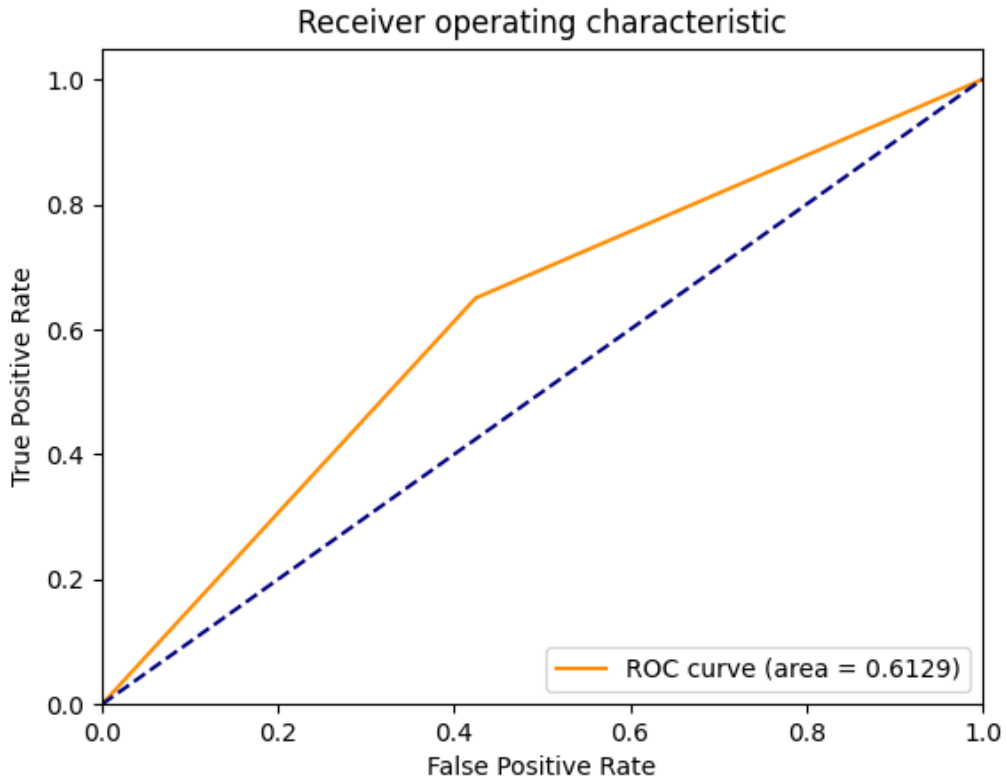


Figura 6.4: ROC curve do modelo do algoritmo Random Forest com os hiperparâmetros escolhidos

Ao observarmos os gráficos anteriormente apresentados podemos observar na figura 6.3 a respectiva matriz de confusão de classificação do modelo face ao conjunto de teste, e na figura 6.4 observar o desempenho do modelo medido com a métrica Area Under the Receiver Operating Characteristics Curve (AUC).

Ao compararmos o desempenho do modelo no período de treino e de teste verificamos que o modelo conseguiu melhorar o seu desempenho, passando de um desempenho de 60.35% obtido durante o período de treino para um desempenho de 61.29% obtido no período de teste.

Tal facto demonstra que o modelo construído conseguiu classificar um novo conjunto de dados de forma superior à classificação que apresentou durante o período de treino, que de acordo com os dados considerados para o seu período de treino o respectivo modelo apto se encontrou apto a classificar o conjunto de dados de teste visto que apresentou um desempenho superior ao obtido anteriormente durante o período de treino.

A análise da matriz de confusão apresentada na figura 6.3 demonstra-nos que um dos objectivos da utilização da métrica considerada Area Under the Receiver Operating Characteristics Curve (AUC) e abordada em 5.1.6 foi cumprido.

Ao analisarmos a respectiva matriz de confusão verificamos que o modelo apresenta mais instâncias classificadas de forma correcta. Apresenta respectivamente 32130 **Verdadeiros Negativos**, 26659 **Verdadeiros Positivos**, 14347 **Falsos Negativos** e 23683 **Falsos Positivos**.

Para além de o modelo apresentar mais instâncias classificadas de forma correcta verifi-

camos que o número de **Falsos Negativos** é o mais baixo aquando comparado com os outros. Tal facto demonstra que o modelo, de acordo com a métrica considerada, e tal como desejado, promove a não ocorrência de **Falsos Negativos**, que no problema em questão representam a situação de o modelo dizer que não irá ocorrer um incêndio e o respectivo incêndio se verificar.

No que diz respeito ao número de **Falsos Positivos** verificamos que o número apresentado é elevado quando comparado com o número de **Verdadeiros Positivos**. Observamos que o modelo apresenta um número de **Falsos Positivos** igual a 23683 e um número de **Verdadeiros Positivos** igual a 26659.

A pequena diferença entre estes valores remete-nos para que numa previsão efectuada por parte do modelo em que irá ocorrer um incêndio haja quase uma igual probabilidade de o incêndio se verificar ou ser um falso alarme.

Tal facto denota alguma incapacidade de classificação do modelo do que são ocorrências ou não ocorrências de incêndio. No entanto o modelo prioriza a ocorrência de **Falsos Positivos** face a **Falsos Negativos**, o que no problema em questão é mais aceitável visto que é preferível prever a ocorrência de um incêndio que depois não se verifique, do que a não ocorrência de um incêndio que posteriormente se venha a verificar.

6.3 Comparação face a outros modelos

A presente sub secção tem como intuito comparar o desempenho dos melhores modelos construídos face a outros modelos de classificação.

A comparação efectuada face a outros modelos tem como intuito verificar o desempenho de classificação de outros modelos do conjunto de dados considerados no estudo do estado da arte efectuado no capítulo 2.

Os modelos considerados como fonte de comparação apresentam a combinação de hiperparâmetros padrão e são originários da biblioteca scikit-learn [52].

Tais modelos são respectivamente Gradient Boosting, Decision Tree, AdaBoost e Logistic Regression.

De seguida encontra-se a tabela que apresenta a comparação dos modelos considerados no desenvolvimento do presente projecto face aos modelos de classificação anteriormente mencionados.

Algoritmo	Desempenho(%)
Support Vector Machine	58.99
Random Forest	61.29
Gradient Boosting	60.20
Decision Tree	54.53
AdaBoost	59.67
Logistic Regression	58.10

Tabela 6.3: Comparação do desempenho obtido pelos modelos considerados face a outros modelos de classificação

Perante a análise da tabela anteriormente apresentada observamos que o modelo do algoritmo Random Forest apresenta o melhor desempenho ao nível do período de teste quer em relação ao modelo do algoritmo Support Vector Machine quer em relação aos outros modelos de classificação acima abordados.

Apesar de os modelos dos algoritmos escolhidos como comparação serem modelos padrão e não se encontrarem optimizados tal como os modelos dos algoritmos considerados no presente projecto, a presente comparação permite-nos ter uma ideia da qualidade da classificação por parte dos modelos considerados, partindo do pressuposto que a optimização dos modelos comparativos não irá apresentar um grande aumento no que diz respeito ao desempenho destes mesmos modelos.

Fruto desta comparação podemos concluir que os modelos considerados apresentam a devida qualidade de classificação e que modelo do algoritmo Random Forest é, de entre os diferentes modelos analisados, aquele que se encontra mais apto a combater o problema de previsão da ocorrência de incêndios florestais.

6.4 Análise dos resultados obtidos

Durante período de treino presenciamos a obtenção de um desempenho de 59.33% por parte do algoritmo Support Vector Machine e de um desempenho de 60.33% por parte do algoritmo Random Forest. Posteriormente durante o período de teste foi obtido um nível de desempenho de 58.99% por parte do algoritmo Support Vector Machine e 61.29% por parte do algoritmo Random Forest.

Os resultados obtidos por parte dos diferentes modelos não podem ser considerados satisfatórios aquando comparados com os resultados obtidos por parte de outros autores no estudo do estado da arte.

Os baixos resultados obtidos por parte dos diferentes algoritmos considerados podem ser fruto de diferentes factores.

Pode num primeiro cenário demonstrar a dificuldade de classificação por parte dos algoritmos considerados para o conjunto de dados utilizado.

Num segundo cenário pode denotar que o conjunto de hiperparâmetros considerado para serem testados não são os que apresentam uma maior influência no problema em questão, ou ainda que o intervalo de valores escolhido para estes mesmos hiperparâmetros não foi o mais acertado para o problema a combater.

Dado que os resultados obtidos por parte dos diferentes algoritmos considerados foi semelhante aos resultados obtidos por parte de outros algoritmos como analisado anteriormente na secção 6.3, remetendo para resultados baixos e ainda assim o algoritmo Random Forest apresentar os melhores resultados, podemos rejeitar o segundo cenário acima apresentado.

Rejeitando este segundo cenário, e por sua consequência validando o primeiro cenário, cenário de dificuldade de classificação por parte dos algoritmos considerados sobre o conjunto de dados utilizado podemos apontar alguns aspectos notáveis a este nível.

No estudo do estado da arte assistimos em regra a estudos que apresentavam a previsão da ocorrência de incêndios para uma zona relativamente pequena, como é o exemplo do conjunto de dados do parque natural Montesinho considerado por Cortez et al. (2007) [44].

O presente projecto desenvolvido visa efectuar a previsão da ocorrência de incêndios florestais ao nível de Portugal Continental, pelo que assistimos a um grande aumento do território abrangido pela previsão de ocorrências aquando comparado por exemplo com Cortez et al. (2007) [44].

O aumento de território implica o aumento das características presentes no conjunto de dados, uma vez que ao longo do território de Portugal Continental os diferentes concelhos apresentam diferentes características quer ao nível das condições meteorológicas quer ao nível da ocorrência de incêndios florestais, pelo que se não existir uma quantidade de dados significativa para cada concelho os modelos podem sentir dificuldades ao nível da previsão da ocorrência de incêndios florestais.

A baixa qualidade do conjunto de dados meteorológicos obrigou a que diversas ocorrências históricas de incêndios fossem descartadas devido a não existirem registos meteorológicos para a data da ocorrência como abordado anteriormente no capítulo 5.

A existência de anos com baixo número de registos de ocorrências de incêndio devido à consequente inexistência de correspondência entre dados, pode implicar a que não haja um número suficiente de instâncias de um concelho nos conjunto de treino, validação e teste, e que os modelos ao utilizarem estes mesmos dados não consigam classificar correctamente as ocorrências de incêndio, uma vez que não foi definida a obrigatoriedade de uma dada percentagem de instâncias de cada concelho nos diferentes conjuntos de treino, validação e teste.

Dado que estamos na presença de um único modelo de previsão, um modelo capaz de prever as ocorrências de incêndios florestais para todo o território de Portugal Continental, a existência de anos em que se verifica um baixo número de instâncias de um determinado concelho poderá prejudicar o treino do modelo e consequentemente o desempenho do modelo, e ser uma das causas de maior impacto nos resultados obtidos por parte dos diferentes modelos considerados e apresentados em A.

Um outro factor que pode também ter apresentado impacto na obtenção dos respectivos resultados foi o processamento de dados efectuado na construção do conjunto de dados considerado por parte dos modelos de Aprendizagem Computacional.

O facto de os dados meteorológicos não apresentarem uma informação devidamente abrangente ao longo do intervalo de tempo de ocorrências de incêndio históricas considerado levou a que bastantes ocorrências fossem descartadas devido a não existirem registos me-

teorológicos correspondentes à data da ocorrência, e que informação que informação que acrescentaria valor aos modelos fosse à priori descartada.

A elevada quantidade de dados em falta em determinados ficheiros de dados meteorológicos obrigou também a que fossem aplicadas estratégias de imputação de dados na construção do conjunto de dados.

Através da imputação de dados foi possível não descartar uma quantidade de entradas de dados tão grande na construção do conjunto de dados, o que representa um factor positivo na medida em que os modelos possuirão mais entradas de dados para realizar o seu período de treino, adquirirão um maior leque de conhecimento e uma maior capacidade de generalização.

No entanto a imputação de dados não apresenta só pontos positivos. A consideração de dados imputados implica a consideração de dados não reais, pelo que ao serem criadas entradas de dados com base nos registos de dados existentes podemos obter valores aproximados aos que realmente se verificaram na altura em questão e promoverão uma aprendizagem dos modelos com base no que realmente aconteceu, ou, devido à incerteza da natureza e neste preciso caso devido à incerteza das condições meteorológicas, obter dados registos discrepantes do que realmente aconteceu e fazer com que os modelos efectuem uma aprendizagem com base em condições erradas que poderá originar implicar uma dificuldade por parte dos modelos no período de treino.

A imputação de dados meteorológicos apesar de ter contribuído por um lado para um maior número de registos de dados a serem considerados pode também ter causado dificuldades no treino dos modelos. No entanto, devido aos dados meteorológicos obtidos para o desenvolvimento do presente projecto foi uma decisão necessária de forma a não comprometer o desenvolvimento do mesmo. A respectiva imputação de dados permitiu a criação de modelos capazes de efectuar a previsão de incêndios florestais embora tais modelos apresentem um desempenho inferior ao desejado tal como foi verificado no presente capítulo.

Capítulo 7

Conclusão

Um dos objectivos associados ao desenvolvimento do presente projecto era, para além de considerar o número de habitantes ao nível dos diferentes concelhos de Portugal na previsão da ocorrência de incêndios florestais, compreender de que forma este factor influência a ocorrência destes mesmos incêndios.

Uma vez que os modelos desenvolvidos não apresentavam um desempenho considerável na previsão da ocorrência de incêndios florestais, sem a consideração do número de habitantes ao nível dos diferentes concelhos de Portugal face à comparação com os resultados obtidos por parte dos autores considerados no estudo do estado da arte no capítulo 2. Assim sendo não foi possível retirar conclusões ao nível da influência do número de habitantes dos diferentes concelhos de Portugal na ocorrência de incêndios florestais.

O facto de os modelos desenvolvidos apresentarem dificuldades em efectuar a previsão da ocorrência de incêndios sem a consideração deste factor relativo ao número de habitantes, implica que através da consideração do mesmo não seja possível retirar conclusões a respeito da sua influência.

Em primeiro lugar é preciso obter um modelo capaz de apresentar um bom desempenho relativo à previsão de ocorrências de incêndio e conseguir competir contra outros modelos já desenvolvidos até ao momento.

Quando estivermos na presença de um modelo que verifique tal cenário podemos então considerar o número de habitantes e retirar conclusões acerca da sua influência na ocorrência de incêndios. Até esse momento tais conclusões não são possíveis de serem retiradas.

Os resultados obtidos relativamente ao desempenho dos modelos considerados que se encontram apresentados e analisados no capítulo 6 demonstram que o modelo do Algoritmo Random Forest conseguiu alcançar o melhor desempenho durante o período de teste com um valor de 61.29%, no entanto abaixo do idealizado tal como é referido no respectivo capítulo.

Perante toda a análise dos resultados obtidos, efectuada no capítulo 6 somos remetidos para a ideia de que o desempenho inferior ao esperado por parte dos modelos considerados se deve ao conjunto de dados utilizado, nomeadamente à qualidade do conjunto de dados meteorológicos considerado.

No entanto, embora os dados meteorológicos considerados não tenham permitido aos modelos alcançar um desempenho de acordo com o idealizado no início do presente projecto todo o processo de tratamento de dados se encontra delineado através da aplicação de diferentes técnicas de Aprendizagem Computacional, a métrica de avaliação de resultados

encontra-se bem definida, o processo de construção de modelos de Aprendizagem Computacional encontra-se definido e o treino, validação e teste dos diferentes modelos encontram-se implementados.

Toda a estrutura necessária para efectuar a previsão de incêndios florestais se encontra construída, pelo que é possível com recurso a pequenas alterações a consideração de um novo conjunto de dados meteorológicos que apresente uma grau de qualidade superior.

A consideração de um novo conjunto de dados meteorológicos de qualidade superior pode resultar na obtenção de resultados iguais ou superiores aos idealizados inicialmente no desenvolvimento do presente projecto, em oposição aos resultados obtidos até ao momento.

O pedido de dados meteorológicos a fontes de informação fidedigna foi efectuado aquando do início do projecto, no entanto a disponibilização de tal conjunto de dados estava implicada do pagamento de quantias monetárias elevadas e, sendo o estado actual do projecto uma prova de conceito não foi possível efectuar tal investimento sem a existência de uma versão que comprovasse a funcionalidade e utilidade do respectivo projecto.

O desenvolvimento de um protótipo totalmente funcional de uma aplicação web que integra a previsão da ocorrência de incêndios florestais, a consulta de ocorrências históricas de anos anteriores e seus respectivos detalhes, e o carregamento e gestão de dados históricos através de uma forma acessível e intuitiva comprovaram a utilidade do presente projecto na ajuda da previsão da ocorrência de incêndios florestais em Portugal Continental.

De tal forma a adição de melhorias ao trabalho efectuado por mim enquanto aluno e estagiário, como a consideração de um conjunto de dados meteorológicos de maior qualidade, poderá tornar o protótipo actualmente existente num produto inovador e competitivo com soluções actualmente existentes ao nível da previsão de incêndios florestais em Portugal.

O desenvolvimento do presente projecto permitiu também verificar que a estrutura inicialmente considerada para o processo de Extract, Transform and Load (ETL) de dados não conseguia satisfazer as necessidades do respectivo projecto.

Inicialmente fora planeado efectuar o processo de ETL de dados através de uma lambda function, no entanto devido às lamda functions serem funções que têm o objectivo de serem executadas num espaço de tempo muito curto, o tempo máximo definido pela Amazon Web Services (AWS) para a execução de uma lamda function é de 15 minutos.

A capacidade de leitura e de escrita de cada tabela é um factor cobrado por parte da Amazon Web Services a partir de um limite estabelecido em 25 unidades para todas as tabelas.

A existência de ficheiros com um elevado número de dados implicava a consideração de uma capacidade de leitura e escrita pagas para tentar, sem garantia, que o processo de ETL fosse efectuado durante o tempo máximo permitido por uma lamda function.

Tal facto teria custos demasiado elevados para o estado actual do projecto e iria também contra o conceito de uma lamda function pelo que o processo de ETL passou assim a ser feito através da AWS Fargate onde o processo de ETL é efectuado pelo tempo necessário.

Os riscos verificados no presente projecto carecem também da devida análise uma vez que apresentaram um impacto significativo no seu desenvolvimento, nomeadamente no que diz respeito ao atraso das tarefas.

Devido aos atrasos verificados no desenvolvimento do presente projecto não foi possível que todos os ecrãs da aplicação web fossem implementados.

Apesar de todas as funcionalidades previstas para o protótipo em questão se encontrarem desenvolvidas no que diz respeito à parte do servidor, não foram no entanto desenvolvidos os ecrãs relativos à reposição da palavra-passe, alteração de definições pessoais e criação de utilizador.

Uma vez que não são funcionalidades que comprometem a utilização do protótipo desenvolvido e que tais funcionalidades necessitam apenas de serem integradas num ecrã da aplicação, podemos considerar que tal não compromete o grau de sucesso do desenvolvimento do presente projecto e que tais ecrãs poderão facilmente ser implementados num futuro próximo de modo a poder tirar partido das referidas funcionalidades através da aplicação web.

Outro ponto que merece a devida análise neste mesmo relatório é o desempenho ao nível da previsão da ocorrência de incêndios florestais através da aplicação web.

Apesar da previsão da ocorrência de incêndios florestais se encontrar totalmente funcional a mesma não é instantânea, factor que nos sugere que possa não ter sido considerada a melhor estrutura para integrar o modelo de Aprendizagem Computacional escolhido na aplicação web.

Visto que estamos perante uma prova de conceito tal não representa um aspecto negativo. O objectivo principal no desenvolvimento do presente projecto era a garantia de que todas as funcionalidades ao nível da previsão da ocorrência de incêndios florestais fossem satisfeitas com sucesso. Dado que todas as funcionalidades se encontram satisfeitas tal objectivo foi cumprido.

Para além de todo o trabalho ao nível de Aprendizagem Computacional que representa o foco do presente projecto, foi também efectuado todo o trabalho ao nível de back-end e front-end necessários à construção da aplicação web bem como da integração de todas estas componentes na Amazon Web Services (AWS) por forma a ter um protótipo totalmente funcional e apto a ser utilizado na ajuda da previsão de ocorrências de incêndios florestais.

Sendo a área da Aprendizagem Computacional uma área que visa poder ajudar o ser humano a ter uma melhor qualidade de vida, a obtenção de um protótipo totalmente funcional no fim do presente estágio vem cumprir este mesmo propósito, ajudar o ser humano.

Deste modo ao olhar não só para os pontos positivos, como também para os pontos negativos do desenvolvimento do presente projecto verificamos que todos os objectivos idealizados foram cumpridos.

Durante o decorrer do presente estágio existiram momentos bons e momentos menos bons durante o período de desenvolvimento, como em qualquer projecto relacionado com a área de Aprendizagem Computacional que, enquanto aluno e futuro engenheiro informático, me fizeram crescer não só a nível profissional como a nível pessoal.

Por forma a finalizar o presente relatório gostava de expressar uma frase da minha autoria, "Na área de Aprendizagem Computacional nem sempre alcançamos os resultados idealizados. Mais importante que alcançar os resultados idealizados é compreender a causa de situações adversas verificadas, de modo a que com empenho e determinação consigamos chegar a estes mesmos resultados".

Bem-haja!

7.1 Trabalho futuro

Embora os objectivos principais previstos para o desenvolvimento do presente projecto terem sido cumpridos, faltaram implementar algumas funcionalidades que haviam sido pedidas.

De tal forma de seguida encontram-se apresentadas as tarefas futuras que pretendem dar continuidade ao protótipo actualmente implementado por forma a acrescentar um maior valor ao mesmo.

Tais tarefas encontram-se divididas em **Curto prazo**, **Médio prazo** e **Longo prazo**. Quando nos referimos a **Curto prazo** falamos de tarefas que podem ser inicializadas neste preciso mês. Já quando nos referimos a **Médio prazo** falamos em tarefas possíveis de se inicializarem entre Dezembro de 2020 e Janeiro de 2021. Por último quando nos referimos a **Longo prazo** falamos de tarefas possíveis de serem inicializadas entre Maio e Junho de 2021.

- **Curto prazo**
  - Implementação do ecrã de reposição da palavra-passe
  - Implementação do ecrã de criação de utilizador
  - Implementação do ecrã de edição dos dados de um utilizador
- **Médio prazo**
  - Aquisição de um conjunto de dados meteorológicos com um grau de qualidade superior
  - Construção de um modelo de Aprendizagem Computacional capaz de apresentar um nível de desempenho competitivo face a outros modelos
  - Consideração do número de habitantes dos diferentes concelhos de Portugal na previsão da ocorrência de incêndios florestais
  - Analisar a influência do número de habitantes dos diferentes concelhos de Portugal na ocorrência de incêndios florestais
- **Longo prazo**
  - Aquisição de um conjunto de dados de operadores móveis
  - Processamento do conjunto de dados de operadores móveis
  - Analisar a informação presente no conjunto de dados de operadores móveis, e verificar de que forma tal informação pode apresentar influência na ocorrência de incêndios florestais

Referências

[1] Effis. https://effis.jrc.ec.europa.eu/. Accessed: 2019-10-17.

[2] Fogos.pt. https://fogos.pt/. Accessed: 2019-10-17.

[3] Gwis. https://gwis.jrc.ec.europa.eu/static/gwis_current_situation/public/index.html. Accessed: 2019-10-17.

[4] National Aeronautics and Space Administration. Moderate resolution imaging spectroradiometer. https://modis.gsfc.nasa.gov/about/. Accessed: 2019-11-26.

[5] PIERRE CHAUVET ALI KAROUNI, BASSAM DAYA. Applying decision tree algorithm and neural networks to predict forest fires in lebanon. Lebanese University, Ecole Doctorale des Sciences et de Technologie, Lebanon, Umm Al-Qura University, Faculty of Social Sciences, Information Science Department, Makkah – KSA, LARIS EA, L'UNAM Universite, Universite Catholique de l'Ouest, 3 place Andre-Leroy,Angers-France,LARIS EA, L'UNAM Universite, Universite d'Angers, Angers - France, 2014.

[6] Amanda Casari Alice X. Zheng. Feature engineering for machine learning: Principles and techniques for data scientists. 2018.

[7] Inc. or its affiliates Amazon Web Services. Serverless applications lens, aws well-architected framework, 2018.

[8] AccuWeather APIs. Forecast api. https://developer.accuweather.com/accuweather-forecast-api/apis. Accessed: 2020-09-08.

[9] The RapidAPI Blog. Top 6 best free weather apis to access global weather data (updated for 2020). https://rapidapi.com/blog/access-global-weather-data-with-these-weather-apis/?utm_source=google&utm_medium=cpc&utm_campaign=DSA&gclid=EAIaIQobChMI3cvjg6jZ6wIVx_hRCh1KsgHFEAAYAiAAEgKip_D_BwE. Accessed: 2020-09-08.

[10] ClimaCell. Global weather api platform. https://www.climacell.co/weather-api/. Accessed: 2020-09-08.

[11] Python Course. What are decision trees? https://www.python-course.eu/Decision_Trees.php. Accessed: 2019-11-22.

[12] Andrej Kobler Saso Dzeoski Katerina Taskova Daniela Stojanova, Pance Panov. Learning to predict forest fires with different data mining techniques. Slovenian Forestry Institute, Ljubljana, Slovenia, Department of Knowledge Technologies, Jozef Stefan Institute, Ljubljana, Slovenia and Faculty of Information Technology, European University, Skopje, Macedonia, 2006.

[13] DataCamp. Machine learning with tree-based models in python.

[14] Anna Hessler David Radke1 and Dan Ellsworth. Firecast: Leveraging deep learning to predict wildfire spread. David R. Cheriton School of Computer Science, University of Waterloo and Department of Mathematics and Computer Science, Colorado College, 2019.

[15] Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo. Redes neuronais artificiais. `http://conteudo.icmc.usp.br/pessoas/andre/research/neural/`. Accessed: 2019-11-25.

[16] Central de Dados. Incêndios. `http://centraldedados.pt/incendios/`. Accessed: 2020-08-17.

[17] Pordata Base de Dados Portugal Contemporâneo. Altitude máxima. `https://www.pordata.pt/Municipios/Altitude+m%C3%A1xima-50`. Accessed: 2020-08-25.

[18] Pordata Base de Dados Portugal Contemporâneo. Altitude mínima. `https://www.pordata.pt/Municipios/Altitude+m%C3%ADnima-49`. Accessed: 2020-08-25.

[19] Pordata Base de Dados Portugal Contemporâneo. Incêndios rurais e área ardida – continente. `https://www.pordata.pt/Portugal/Inc%C3%AAndios+rurais+e+%C3%A1rea+ardida+%E2%80%93+Continente-1192`. Accessed: 2020-01-09.

[20] Pordata Base de Dados Portugal Contemporâneo. População residente segundo os censos: total e por grandes grupos etários. `https://www.pordata.pt/Municipios/Popula%C3%A7%C3%A3o+residente+segundo+os+Censos+total+e+por+grandes+grupos+et%C3%A1rios-22`. Accessed: 2020-08-17.

[21] Pordata Base de Dados Portugal Contemporâneo. Superfície. `https://www.pordata.pt/DB/Municipios/Ambiente+de+Consulta/Tabela`. Accessed: 2020-08-24.

[22] Sistema Nacional de Informação de Recursos Hídricos. Dados de base. `https://snirh.apambiente.pt/index.php?idMain=2&idItem=1`. Accessed: 2020-08-17.

[23] Sistema Nacional de Informação de Recursos Hídricos. Pesquisa rápida de estações. `https://snirh.apambiente.pt/index.php?idMain=2&idItem=3`. Accessed: 2020-08-20.

[24] Galar M. Prati R.C. Krawczyk B. Herrera F. Fernández A., García S. *Learning from Imbalanced Data Sets*. Springer, Cham, 2018.

[25] National Geographic. Gis (geographic information system). `https://www.nationalgeographic.org/encyclopedia/geographic-information-system-gis/`. Accessed: 2019-11-22.

[26] George Mitri George E. Sakr, Imad H. Elhajj and Uchechukwu C. Wejinya. Artificial intelligence for forest fire prediction. IEEE/ASME International Conference on Advanced Intelligent Mechatronics Montréal, Canada, 2010.

[27] grama. About. `https://www.grama.io/about.html`. Accessed: 2020-01-09.

[28] GSMA. About us. `https://www.gsma.com/aboutus/`. Accessed: 2020-01-09.

[29] Khabib Mustofa Guruh Shidik. Predicting size of forest fire using hybrid model. Universitas Dian Nuswantoro (Indonesia) and Universitas Gadjah Mada (Indonesia), 2016.

[30] imbalanced learn. Under-sampling. `https://imbalanced-learn.readthedocs.io/en/stable/under_sampling.html`. Accessed: 2020-09-02.

[31] MINGYI GAO GANGXIANG SHEN JUNFENG ZHANG, WEI CHEN. K-means-clustering-based fiber nonlinearity equalization techniques for 64-qam coherent optical communication system. School of Electronic and Information Engineering, Soochow University, No. 1 Shizi Street, Suzhou, Jiangsu Province, 215006, China, 2017.

[32] KDnuggets. Making predictive models robust: Holdout vs cross-validation. `https://www.kdnuggets.com/2017/08/dataiku-predictive-model-holdout-cross-validation.html`. Accessed: 2020-10-11.

[33] Habib Fathallah Latifa Guesmi and Mourad Menif. Modulation format recognition using artificial neural networks for the next generation optical networks. 2018.

[34] Machine Learning Mastery. Ordinal and one-hot encodings for categorical data. `https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/`. Accessed: 2020-09-01.

[35] MathWorks. Introducing machine learning. 2016.

[36] Medium. Cross validation in time series. `https://medium.com/@soumyachess1496/cross-validation-in-time-series-566ae4981ce4`. Accessed: 2020-09-02.

[37] Medium. What is the kernel trick? why is it important? `https://medium.com/@zxr.nju/what-is-the-kernel-trick-why-is-it-important-98a98db0961d`. Accessed: 2019-11-22.

[38] Government of Canada. Canadian forest fire weather index (fwi) system. `https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi`. Accessed: 2019-11-22.

[39] OpenWeatherMap. 5 day weather forecast. `https://openweathermap.org/`. Accessed: 2020-09-08.

[40] Packt. Bagging. `https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781788830577/3/ch03lvl1sec34/bagging`. Accessed: 2019-11-22.

[41] Visual Paradigm. What are the scrum events? `https://www.visual-paradigm.com/scrum/what-are-scrum-events/`. Accessed: 2019-11-22.

[42] Visual Paradigm. What are the three scrum roles? `https://www.visual-paradigm.com/scrum/what-are-the-three-scrum-roles/`. Accessed: 2019-11-22.

[43] Visual Paradigm. What is a sprint in scrum? `https://www.visual-paradigm.com/scrum/what-is-sprint-in-scrum/`. Accessed: 2019-11-22.

[44] Aníbal Morais Paulo Cortez. A data mining approach to predict forest fires using meteorological data. Department of Information Systems/RD Algoritmi Centre, University of Minho, Guimarães, Portugal, 2007.

[45] Python. pickle — python object serialization. `https://docs.python.org/3/library/pickle.html`. Accessed: 2020-09-08.

[46] SRINIVAS RAMASUBRAMANIAN. Predicting the burned area in forest using machine learning techniques. National College of Ireland, 2017.

[47] UCI Machine Learning Repository. Forest fires data set. `https://archive.ics.uci.edu/ml/datasets/forest+fires`. Accessed: 2019-11-18.

[48] Mahdi A. Salman Samaher Al_Janabi, Ibrahim Al_Shourbaji. Assessing the suitability of soft computing approaches for forest fires prediction. Department of Computer Science, Faculty of Science for Women (SCIW), University of Babylon, Iraq and Department of Computer Network, Faculty of Computer Science and Information System, University of Jazan, Jazan, Saudi Arabia, 2017.

[49] Saed Sayad. Logistic regression. `https://www.saedsayad.com/logistic_regression.htm`. Accessed: 2019-11-22.

[50] Towards Data Science. Demystifying maths of svm — part 1. `https://towardsdatascience.com/demystifying-maths-of-svm-13ccfe00091e`. Accessed: 2019-11-22.

[51] scikit learn. Linear support vector classification. `https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html#sklearn.svm.LinearSVC`. Accessed: 2020-10-18.

[52] scikit learn. Machine learning in python. `https://scikit-learn.org/stable/index.html`. Accessed: 2020-09-09.

[53] scikit learn. Multi-layer perceptron classifier. `https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html`. Accessed: 2020-10-18.

[54] scikit learn. One hot encoder. `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html#sklearn.preprocessing.OneHotEncoder`. Accessed: 2020-09-01.

[55] scikit learn. Ordinal encoder. `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OrdinalEncoder.html`. Accessed: 2020-09-01.

[56] scikit learn. Random forest classifier. `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`. Accessed: 2020-10-18.

[57] scikit learn. Time series split. `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html`. Accessed: 2020-09-02.

[58] Amazon Web Services. Amazon cognito. `https://aws.amazon.com/pt/cognito/`. Accessed: 2020-09-10.

[59] Dark Sky. Dark sky api. `https://darksky.net/dev`. Accessed: 2020-09-08.

[60] Jeremy Storer. Computational intelligence and data mining techniques using the fire data set. Graduate College of Bowling Green State University, 2016.

[61] TEX. Illustrating the random forest algorithm in tikz. `https://tex.stackexchange.com/questions/503883/illustrating-the-random-forest-algorithm-in-tikz`. Accessed: 2019-11-22.

[62] towards data science. Classification: Roc curve and auc. `https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc`. Accessed: 2020-09-01.

[63] towards data science. Logistic regression. `https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148`. Accessed: 2020-01-14.

[64] towards data science. Time series nested cross-validation. https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9. Accessed: 2020-09-02.

[65] towards data science. Understanding auc - roc curve. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5. Accessed: 2020-09-01.

[66] SPSS tutorials. Pearson correlations – quick introduction. https://www.spss-tutorials.com/pearson-correlation-coefficient/. Accessed: 2020-09-09.

[67] Weather2020. Weather2020 data platform. https://www.weather2020.com/products. Accessed: 2020-09-08.

[68] Weatherbit.io. Weather api documentation. https://www.weatherbit.io/api. Accessed: 2020-09-08.

[69] WILDML. Implementing a neural network from scratch in python - an introduction. http://www.wildml.com/2015/09/implementing-a-neural-network-from-scratch/. Accessed: 2020-01-14.

[70] Dawn Wright. Empirical rule and z-score probability, Nov 2019.

[71] Rhett D. Harrison Mohan Kumar Sammathuria Yong Poh Yu, Rosli Omar and Abdul Rahim Nik. Pattern clustering of forest fires based on meteorological variables and its classification using hybrid data mining methods. Department of Electrical Engineering, University of Malaya, 50603 Lembah Pantai, Kuala Lumpur, Malaysia, Xishuangbanna Tropical Botanical Garden, Menglun, Mengla, 666303, Yunnan, China, Malaysia Meteorological Department, Jalan Sultan, 46667 Petaling Jaya, Malaysia and Forest Research Institute, 52110 Kepong, Selangor, Malaysia, 2011.

Anexos

Anexo A

Resultados obtidos durante o período de treino

A.1 Support Vector Machine

De seguida encontram-se apresentados os resultados obtidos durante o período de treino por parte dos modelos do algoritmo Support Vector Machine.

C	dual	loss	max_iter	penalty	tol	Performance
100.0	False	squared_hinge	9500	l1	0.0081	0.5929035943974041
0.01	False	squared_hinge	8500	l1	0.0081	0.5928295036695573
0.001	False	squared_hinge	1500	l1	0.0041	0.5884944643766612
1.0	False	squared_hinge	8500	l1	0.0031	0.5931631384177161
0.01	False	squared_hinge	9500	l2	0.0021	0.5930752933416595
1000.0	False	squared_hinge	2000	l2	0.0051	0.5931867843920879
100.0	False	squared_hinge	1500	l2	0.0031	0.5931839354523406
100.0	False	squared_hinge	8000	l1	0.0031	0.5930552447823435
1.0	False	squared_hinge	7500	l1	0.0011	0.5932197546949689
1000.0	False	squared_hinge	8500	l1	0.0051	0.5931840695809458
100.0	False	squared_hinge	4000	l1	0.0001	0.5932044593699731
100.0	False	squared_hinge	8000	l2	0.0021	0.5931353027712453
0.001	False	squared_hinge	7000	l1	0.0091	0.5884569870513443
100.0	False	squared_hinge	3000	l1	0.0071	0.5928262427232089
1.0	False	squared_hinge	3000	l1	0.0061	0.5928082891232169
0.01	False	squared_hinge	6500	l1	0.0071	0.5929142336288897
1.0	False	squared_hinge	1500	l1	0.0091	0.5931030801938574
100.0	False	squared_hinge	5000	l1	0.0011	0.5933253995102716
0.01	False	squared_hinge	3500	l1	0.0021	0.5929217994248269
10.0	False	squared_hinge	1500	l1	0.0041	0.5931214698655174
0.1	False	squared_hinge	5000	l1	0.0011	0.5932762878108206
0.1	False	squared_hinge	5000	l2	0.0011	0.5929618531463656
0.1	False	squared_hinge	5000	l1	0.0011	0.5932072438719741
0.1	False	squared_hinge	5000	l1	0.0011	0.5929576871892048
10.0	False	squared_hinge	9000	l1	0.0011	0.5930264080801889
0.1	False	squared_hinge	4500	l2	0.0011	0.5932656862669763
0.1	False	squared_hinge	1000	l1	0.0001	0.5931945466970052

C	dual	loss	max_iter	penalty	tol	Performance
0.1	False	squared_hinge	5500	l1	0.0061	0.5930467831140274
10.0	False	squared_hinge	6000	l1	0.0081	0.5931920783344711
1000.0	False	squared_hinge	2500	l2	0.0011	0.5931043030473067
0.001	False	squared_hinge	5000	l1	0.0011	0.5884714865122838
100.0	False	squared_hinge	5000	l1	0.0041	0.5932649483664587
0.1	False	squared_hinge	4000	l1	0.0081	0.59336605843485
100.0	False	squared_hinge	4000	l1	0.0081	0.5928581328629721
0.001	False	squared_hinge	4000	l2	0.0081	0.5927249828905308
1000.0	False	squared_hinge	9500	l1	0.0081	0.5930349533860234
0.01	False	squared_hinge	2000	l2	0.0081	0.5930821280518659
100.0	False	squared_hinge	7500	l1	0.0051	0.5933217501370198
0.1	False	squared_hinge	7000	l1	0.0031	0.5930807374672796
10.0	False	squared_hinge	4000	l1	0.0091	0.5931632360311003
1.0	False	squared_hinge	5500	l2	0.0001	0.5929651239057756
100.0	False	squared_hinge	1000	l1	0.0071	0.5928865276136431
1000.0	False	squared_hinge	8500	l1	0.0051	0.5932492267782721
100.0	False	squared_hinge	3500	l2	0.0061	0.5932504406935637
0.001	False	squared_hinge	6500	l1	0.0021	0.5883981715278963
0.01	False	squared_hinge	2500	l1	0.0031	0.5929573657905329
1.0	False	squared_hinge	4500	l1	0.0041	0.5932227856837623
100.0	False	squared_hinge	6000	l1	0.0081	0.5930270901224144
0.1	False	squared_hinge	8000	l2	0.0091	0.5930359166544928
100.0	False	squared_hinge	9000	l1	0.0071	0.5929458148558405
0.001	False	squared_hinge	9500	l1	0.0001	0.5885928194516793
1000.0	False	squared_hinge	4000	l1	0.0021	0.5934416457438745
1000.0	False	squared_hinge	4000	l1	0.0021	0.5932467046190262
1000.0	False	squared_hinge	4000	l2	0.0021	0.5933431047302111
1000.0	False	squared_hinge	3000	l1	0.0021	0.5931339554369202
1000.0	False	squared_hinge	4000	l1	0.0021	0.5928924655544119
1000.0	False	squared_hinge	2000	l1	0.0061	0.5930382602524311
0.1	False	squared_hinge	7500	l2	0.0051	0.5932619242860709
0.01	False	squared_hinge	7000	l1	0.0081	0.5928389046655053
10.0	False	squared_hinge	1500	l1	0.0021	0.5933901899658606
10.0	False	squared_hinge	6500	l1	0.0021	0.5930541361714505
10.0	False	squared_hinge	1500	l1	0.0021	0.5931904880034076
10.0	False	squared_hinge	1500	l2	0.0021	0.5929283490611246
10.0	False	squared_hinge	1500	l1	0.0021	0.5931241023304474
10.0	False	squared_hinge	8500	l1	0.0031	0.5931034957195994
1.0	False	squared_hinge	4000	l1	0.0041	0.5931905764169235
0.1	False	squared_hinge	3500	l1	0.0021	0.5932753101434238
10.0	False	squared_hinge	8000	l1	0.0081	0.593088141642151
1000.0	False	squared_hinge	1500	l1	0.0091	0.593082050848953
0.1	False	squared_hinge	9000	l1	0.0071	0.5931554473933544
0.01	False	squared_hinge	4000	l1	0.0001	0.5929061262822476
10.0	False	squared_hinge	1000	l1	0.0021	0.5933731613125957
10.0	False	squared_hinge	1000	l1	0.0021	0.5932210460350685
10.0	False	squared_hinge	1000	l1	0.0021	0.5930562989888133
10.0	False	squared_hinge	1000	l1	0.0021	0.5927568672655383
10.0	False	squared_hinge	6000	l1	0.0021	0.5930198033197982

C	dual	loss	max_iter	penalty	tol	Performance
10.0	False	squared_hinge	5500	l1	0.0061	0.593122346059309
0.001	False	squared_hinge	4500	l1	0.0021	0.588592077558253
1.0	False	squared_hinge	1000	l2	0.0051	0.5930812436087802
1000.0	False	squared_hinge	9500	l1	0.0041	0.5932662178197562
10.0	False	squared_hinge	3000	l1	0.0031	0.5930942116885469
10.0	False	squared_hinge	2500	l1	0.0021	0.5931297928397047
1000.0	False	squared_hinge	2000	l2	0.0091	0.5929804105299599
0.001	False	squared_hinge	7500	l1	0.0071	0.5883071422925941
0.01	False	squared_hinge	6500	l1	0.0001	0.5927384994001262
1.0	False	squared_hinge	1500	l1	0.0021	0.5933953783710617
1.0	False	squared_hinge	1500	l1	0.0061	0.5932384687039708
1.0	False	squared_hinge	1500	l2	0.0021	0.5931043939673517
1.0	False	squared_hinge	1500	l1	0.0051	0.593114135552238
1.0	False	squared_hinge	7000	l1	0.0041	0.5929821205711779
1.0	False	squared_hinge	1500	l1	0.0021	0.5929468390176487
1.0	False	squared_hinge	8500	l2	0.0031	0.5930042808628404
1000.0	False	squared_hinge	3500	l1	0.0011	0.5930550852410397
1.0	False	squared_hinge	8000	l1	0.0021	0.5930802951820818
1000.0	False	squared_hinge	5500	l1	0.0091	0.5932397234903664
0.001	False	squared_hinge	4500	l1	0.0071	0.5883715685956112
0.01	False	squared_hinge	9000	l2	0.0021	0.5931516480221116
1000.0	False	squared_hinge	1500	l1	0.0001	0.5932000227917886
1.0	False	squared_hinge	6000	l1	0.0061	0.5930396733193011
100.0	False	squared_hinge	2500	l1	0.0021	0.5930183295361527

Tabela A.1: Resultados obtidos durante o período de treino por parte dos modelos do algoritmo Support Vector Machine

A.2 Random Forest

De seguida encontram-se apresentados os resultados obtidos durante o período de treino por parte dos modelos do algoritmo Support Vector Machine.

As abreviaturas do cabeçalho da tabela que se encontra de seguida apresentada representam respectivamente:

- **c**: criterion
- **m_d**: max_depth
- **m_f**: max_features
- **m_s_s**: min_samples_split
- **n_e**: number_estimators
- **P**: Performance

c	m_d	m_f	m_s_s	n_e	P
gini	24	8	0.01	850	0.6038598499171769
entropy	7	log2	0.060000000000000005	250	0.5947788357483658
gini	27	11	0.01	400	0.6030081626501478
gini	14	13	0.11	750	0.5901932679548605
gini	10	2	0.21000000000000002	600	0.5877753229909551
entropy	5	6	0.01	850	0.5961953330826699
entropy	10	10	0.21000000000000002	550	0.5864710320594806
entropy	18	13	0.060000000000000005	450	0.5931436916228675
entropy	8	3	0.11	250	0.5925118350123222
gini	15	14	0.21000000000000002	150	0.5804758446810776
entropy	3	log2	0.01	950	0.5906200243583062
entropy	21	12	0.01	150	0.6028234062119184
entropy	14	1	0.060000000000000005	450	0.5937656609542108
entropy	1	1	0.16000000000000003	300	0.5796433089284541
entropy	9	3	0.060000000000000005	300	0.5963341804268765
gini	23	9	0.060000000000000005	700	0.5943356963800464
entropy	10	2	0.21000000000000002	950	0.5879562657030528
entropy	24	9	0.060000000000000005	250	0.5945995416913774
gini	12	sqrt	0.11	250	0.5929445904795003
gini	14	4	0.21000000000000002	900	0.5895022788307451
gini	27	8	0.01	400	0.6030322621647084
gini	13	8	0.01	800	0.6032479508704831
gini	13	8	0.01	800	0.6028326792196868
gini	25	8	0.16000000000000003	800	0.5889004148043145
gini	24	5	0.01	100	0.6035185947325737
gini	24	5	0.01	100	0.6033069174815815
gini	4	5	0.01	500	0.5940471189530954
gini	24	7	0.16000000000000003	100	0.5886376516514688
gini	22	5	0.01	850	0.6035202969313402
gini	22	7	0.01	850	0.6028428131472328
gini	20	11	0.01	350	0.6025878416429122

c	m_d	m_f	m_s_s	n_e	P
gini	22	14	0.11	650	0.5897231456087326
gini	11	5	0.16000000000000003	850	0.5899209048283505
gini	17	sqrt	0.01	200	0.6031422338223924
gini	26	6	0.01	850	0.6025806783471853
gini	2	12	0.11	750	0.5831688900257006
gini	None	10	0.01	600	0.6028767912960566
gini	7	13	0.21000000000000002	850	0.5826822224967161
gini	6	log2	0.01	550	0.5985417643321838
gini	28	4	0.11	850	0.5926883522829264
gini	19	11	0.16000000000000003	200	0.5872547819805014
gini	16	6	0.01	500	0.6026276612153862
entropy	22	10	0.21000000000000002	700	0.5862103810822006
gini	18	2	0.060000000000000005	350	0.5957494986514243
entropy	5	3	0.01	650	0.5960835265884161
gini	15	1	0.01	900	0.6010665617402486
entropy	8	14	0.060000000000000005	450	0.5930398846384997
gini	3	13	0.11	150	0.5872921478530199
gini	1	8	0.21000000000000002	400	0.5793007995472357
entropy	12	9	0.16000000000000003	750	0.588897447034569
gini	21	log2	0.01	850	0.6034900689495607
gini	9	12	0.060000000000000005	300	0.5939797034206136
entropy	23	5	0.01	950	0.6033659140369159
gini	25	8	0.21000000000000002	600	0.5872753575283628
gini	27	4	0.01	550	0.6029947904974229
entropy	10	sqrt	0.11	850	0.5927503294994692
gini	4	2	0.16000000000000003	250	0.5885631701732826
gini	22	3	0.01	700	0.6032343473412324
gini	20	5	0.060000000000000005	450	0.5959442464463407
entropy	24	7	0.01	900	0.6036668776252145
entropy	24	7	0.01	900	0.6032361935949527
entropy	11	8	0.21000000000000002	900	0.5873215423656596
entropy	24	7	0.11	900	0.5916760976425112
entropy	24	7	0.01	150	0.6037048296883133
entropy	14	1	0.16000000000000003	150	0.5871624038191852
entropy	24	7	0.01	150	0.6031031861331002
entropy	26	7	0.01	150	0.6026944395605989
entropy	24	7	0.01	400	0.6030697525619456
entropy	17	7	0.01	950	0.6030102396241461
entropy	24	7	0.01	800	0.602992589376228
entropy	2	9	0.01	500	0.5837117354158838
entropy	None	11	0.01	350	0.6025297584318031
entropy	6	6	0.01	300	0.5985308535184546
entropy	19	14	0.060000000000000005	200	0.5930656254139206
entropy	7	10	0.01	650	0.5993097151448583
entropy	13	8	0.01	900	0.6031952581806959
entropy	28	13	0.01	150	0.6026012230707705
entropy	24	12	0.21000000000000002	750	0.5852249263027026
entropy	16	sqrt	0.11	250	0.5922282113796854
entropy	18	log2	0.16000000000000003	100	0.590201103440496

c	m_d	m_f	m_s_s	n_e	P
entropy	5	7	0.01	600	0.5959977959800548
entropy	8	2	0.060000000000000005	900	0.595940697043405
entropy	15	3	0.01	550	0.6030618900233538
entropy	3	4	0.01	700	0.5904505409731723
entropy	21	8	0.21000000000000002	150	0.5876805578259146
entropy	9	1	0.11	800	0.5901656873123545
entropy	1	11	0.16000000000000003	850	0.5766449171422741
entropy	23	7	0.01	450	0.6039353765077446
gini	12	6	0.01	450	0.603026993411287
entropy	23	9	0.060000000000000005	450	0.5953345357277445
gini	23	10	0.01	450	0.6028537944701655
entropy	23	14	0.01	450	0.6017910629455671
gini	25	8	0.21000000000000002	500	0.5876018666643665
entropy	27	13	0.11	400	0.589138082862804
gini	4	12	0.16000000000000003	200	0.5871260304154838
entropy	10	4	0.01	350	0.6023540086466093
gini	11	7	0.01	650	0.6025708364953501
entropy	20	log2	0.060000000000000005	100	0.596313032574083
gini	23	sqrt	0.01	950	0.6028435574687407
entropy	14	7	0.21000000000000002	300	0.5876065170595263

Tabela A.2: Resultados obtidos durante o período de treino por parte dos modelos do algoritmo Random Forest