NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENTO DE INFORMÁTICA

GONÇALO BERNARDO SEVERINO

Licenciado em Engenharia Informática

INDUÇÃO DE REGRAS PARA A CLASSIFICAÇÃO AUTOMÁTICA DA SEVERIDADE DE RISCO DE INCÊNDIOS FLORESTAIS

MESTRADO EM ENGENHARIA INFORMÁTICA

Universidade NOVA de Lisboa
Janeiro, 2023

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENTO DE INFORMÁTICA

INDUÇÃO DE REGRAS PARA A CLASSIFICAÇÃO AUTOMÁTICA DA SEVERIDADE DE RISCO DE INCÊNDIOS FLORESTAIS

GONÇALO BERNARDO SEVERINO

Licenciado em Engenharia Informática

Orientadora: Susana Nascimento
Professora Auxiliar, Universidade Nova de Lisboa

Coorientadores: Carlos Viegas Damásio
Professor Associado, Universidade Nova de Lisboa

Maria Lourdes Bugalho
Meteorologista, Instituto Português do Mar e da Atmosfera

MESTRADO EM ENGENHARIA INFORMÁTICA

Universidade NOVA de Lisboa
Janeiro, 2023

Indução de Regras para a Classificação Automática da Severidade de Risco de Incêndios Florestais

Resumo

O objetivo desta dissertação é aplicar métodos de mineração de dados a dados de incêndios florestais de forma a extrair regras interpretáveis na classificação dos graus de severidade dos incêndios. Foram construídos 12 conjuntos de dados de estudo referentes a Portugal Continental e a cinco regiões geográficas e caracterizados por duas séries de índices de risco. Os dados referem-se aos incêndios com uma área ardida maior que 100 hectares ocorridos no território de Portugal Continental entre os anos de 2001 e 2020.

Os dados foram classificados de forma automática não supervisionada com o algoritmo *fuzzy c-means* com suporte às escalas existentes dos índices de risco. Os resultados obtidos com este algoritmo foram validados com o índice de validação interno Xie-Beni e a com o método de visualização *fuzzy sammon mapping*.

Com os dados etiquetados foi realizado um estudo experimental de algoritmos de classificação com extração de regras: árvores de decisão, RIPPER e SIRUS; e subsequente avaliação comparativa dos conjuntos de regras com métricas de interpretabilidade e com análise por especialista na área.

Conclui-se que o FCM e os algoritmos de extração de regras foram eficazes na abordagem, o classificador RIPPER obteve melhores resultados e a análise por região foi satisfatória.

Palavras-chave: *fuzzy c-means*, árvores de decisão, florestas aleatórias, RIPPER, SIRUS, extração de regras, medidas de interpretabilidade, canadian forest fire risk index (FWI), continuous Haines index (CHI)

Abstract

The objective of this dissertation is to apply data mining methods to forest fire data in order to extract interpretable rules for classifying fire severity degrees. Twelve sets of study data referring to Mainland Portugal and five geographic regions were constructed and characterized by two different series risk indices. The data refer to fires with a burned area greater than 100 hectares that occurred in Mainland Portugal between 2001 and 2020.

The data were classified in an unsupervised automatic way with the *fuzzy c-means* algorithm and the existing risk index scales. The results obtained with this algorithm were validated with the Xie-Beni internal validation index and with the *fuzzy sammon mapping* visualization method.

With the tagged data, a comparative experimental study was applied with the classification algorithms with rule extraction: decision trees, RIPPER and SIRUS; and subsequent comparative evaluation of the sets of rules with interpretability metrics and analysis by an expert in the field.

It is concluded that the FCM and the rule extraction algorithms were effective in the approach, the RIPPER classifier obtained better results and the analysis by region was positive.

Keywords: fuzzy c-means, decision tree, random forest, RIPPER, SIRUS, rule extraction, interpretability scores, canadian forest fire risk index (FWI), continuous Haines index (CHI)

Índice

1

Introdução

1.1 Motivação

Os incêndios são um desastre natural que afeta várias regiões do globo. Estes fenómenos têm um impacto tanto ambiental, com a destruição de vastas áreas de vegetação e população animal, erosão do solo, como também um impacto social e económico com perdas de pessoas e bens, de infraestruturas, danos físicos ou prejuízos à saúde das populações.

As escalas de previsão de risco de incêndio florestal permitem, como o nome indica, atribuir a uma região o risco de ocorrência de incêndio, com base em índices de risco. Os índices de risco de incêndio florestal permitem traduzir as condições meteorológicas propícias a fogos florestais. Um índice amplamente utilizado é o índice canadiano de risco de incêndio florestal (FWI), sendo este índice calculado com base nos valores de temperatura do ar, humidade relativa do ar, velocidade do vento e precipitação acumulada de 24 horas anterior. Para efeitos de comparação de resultados, e porque se pretende que os resultados deste trabalho sejam úteis na operacionalização da severidade de fogo florestal foram utilizadas as escalas de risco deste índice adotadas pelas instituições oficiais da área a nível nacional e europeu, nomeadamente a escala do Instituto Português do Mar e da Atmosfera (IPMA) [25] e a escala do Sistema de Informação Europeu sobre Incêndios Florestais (EEFIS - *European Forest Fire Information System*) [14].

As escalas de severidade baseadas neste índice consistem em atribuir a intervalos de FWI uma escala ordinal de risco. As situações meteorológicas, traduzidas por diferentes índices de risco, sintetizam condições necessárias para a existência de incêndios, mas não suficientes, devido a várias outras causas. A análise de diferentes índices e escalas conduzem a uma melhor apreciação do risco

O trabalho desta dissertação tem como principal objetivo contribuir para o desenvolvimento de uma escala de risco de incêndio, baseada em vários índices ajustada às cinco regiões geográficas de Portugal continental (Norte Litoral, Norte Interior, Centro Litoral, Centro Interior e Sul). Esta dissertação dá continuidade ao estudo feito em uma dissertação intitulada "Identificação automática de valores de índices de risco de incêndios

florestais aos níveis de risco de incêndio extremo florestal em Portugal Continental" [8]. O objetivo dessa dissertação passou por identificar as condições meteorológicas associadas a diferentes níveis de risco de incêndio aplicando métodos não supervisionados de agrupamento difuso e métodos preditivos a dados de incêndios florestais extremos e aos índices de risco de incêndio associados. O autor concluiu que o algoritmo *fuzzy c-means* e as árvores de decisão foram eficazes como abordagem ao problema. Verificou-se também que os intervalos dos índices nem sempre entram em concordância com as escalas de risco de referência.

Um dos aspetos indicados como possível aprofundamento do estudo é o uso de algoritmos de florestas aleatórias como modelo preditivo com métodos de interpretação para extrair as condições meteorológicas, traduzidas por índices de severidade de incêndios florestais. As florestas aleatórias, pela sua caracterização, possibilitam obter precisão mais elevada do que árvores de decisão, mas não são passíveis de interpretação por especialistas. Os modelos de caixa fechada (*black box*), aos quais as florestas aleatórias pertencem, são modelos nos quais é difícil entender os mecanismos que levam a um determinado resultado.

Existem algumas diferenças nos dados que vão ser usados nesta dissertação, em comparação com [8], especificamente o uso de incêndios de menor área, aumentando o número total de incêndios analisados, a inserção no estudo de regiões mais pequenas para os valores dos índices meteorológicos e o não uso da média em alguns valores por estes já serem cumulativos.

Esta dissertação recorre ao algoritmo de *fuzzy c-means* para a atribuição de uma escala ordinal aos dados, utilizar árvores de decisão, o algoritmo RIPPER de poda de árvores de decisão e o algoritmo SIRUS de extração de regras de florestas aleatórias para caracterizar e obter as condições dessa escala. Os conjuntos de regras obtidos com os diferentes métodos foram alvo de um estudo comparativo com o uso das medidas de interpretabilidade: previsibilidade, estabilidade e simplicidade.

1.2 Principais objetivos e contribuições

Esta dissertação segue na continuação do estudo realizado em [8], e o grande objetivo continua a ser mostrar que o agrupamento com o algoritmo *fuzzy C-means* com a abordagem de analisar os protótipos é uma abordagem viável e robusta para a classificação não supervisionada dos incêndios a partir de escala(s) de risco de incêndio do domínio (escala IPMA e escala EEFIS). Os dados classificados são subsequentemente alvo de um processo de extração de regras onde se vão comparar três tipos de classificadores: árvores de decisão, RIPPER e SIRUS.

Para tal construiu-se um conjunto alargado de dados dos incêndios em Portugal nos últimos 20 anos. Estes dados encontram-se divididos pelas cinco regiões de Portugal continental e classificados com o auxílio do algoritmo de agrupamento difuso *fuzzy c-means* e de 2 escalas de risco de incêndio existentes para o índice FWI. Este algoritmo

atribui a uma observação um nível de pertença a cada partição, permitindo modelar a incerteza encontrada nas fronteiras entre classes de risco.

Com os dados classificados são extraídas e comparadas regras associadas aos níveis de risco com o auxílio de três métodos: algoritmo CART de árvores de decisão [27], algoritmo RIPPER [9] de poda de árvores de decisão e algoritmo SIRUS [2] de extração de regras de florestas aleatórias. As regras obtidas com os diferentes métodos são alvo da aplicação de um procedimento de avaliação baseado em medidas de interpretabilidade recentemente propostas na literatura ([30]).

As principais contribuições deste estudo foram:

- Desenvolvimento de um protocolo experimental para suportar os objetivos do trabalho proposto.
- Construção de conjuntos de dados de estudo referentes a Portugal Continental e a 5 regiões geográficas (Norte Litoral, Norte Interior, Centro Litoral, Centro Interior e Sul). Dados estes que dizem respeito a 1946 incêndios, com mais de 100 hectares de área ardida, ocorridos entre os anos 2001 a 2020. Os dados encontram-se divididos em 12 conjuntos caracterizados por a combinação das regiões com 2 séries de índices de risco de incêndio, uma série composta de 5 índices e a outra de 7.
- Classificação automática não supervisionada da severidade de incêndio com o uso do algoritmo *fuzzy c-means* e das escalas existentes para o índice FWI (EEFIS e IPMA), seguido da validação e visualização das mesmas com o algoritmo *fuzzy sammon mapping*.
- Estudo experimental de algoritmos de classificação com extração de regras: árvores de decisão, RIPPER e SIRUS; e subsequente avaliação comparativa dos conjuntos de regras dos modelos de classificação de severidade com métricas de interpretabilidade (previsibilidade, estabilidade e simplicidade). Análise por especialista na área das regras de severidade de risco de incêndio obtidas com a comparação com as escalas de risco existentes.

O trabalho desenvolvido propõe-se contribuir para a definição de escalas de severidade de risco de incêndios mais informativas e que, consequentemente, possam melhorar os sistemas de previsão de incêndios florestais tanto em Portugal Continental como a nível regional.

1.3 Organização do documento

Após este capítulo de introdução, com a apresentação do problema de incêndios florestais, das limitações das escalas de risco existentes e do consequente objetivo deste estudo segue-se um capítulo com a apresentação dos índices de risco com o qual se pretende trabalhar. No capítulo 3 e 4 são apresentados os métodos e algoritmos de agrupamento e

de extração de regras empregues nesta dissertação. No capítulo seguinte são descritos os conjuntos de dados estudados e o procedimento envolvido na sua construção. Nos capítulos 6 e 7 é apresentada a abordagem experimental desenvolvida seguida dos resultados obtidos com a mesma. Por fim, são apresentadas as principais conclusões do trabalho e as vantagens e limitações da abordagem utilizada.

2

Grandes fogos e índices de risco de incêndio florestal

2.1 Grandes fogos

Não existe consenso na classificação da severidade dos fogos florestais [43], o critério mais comum para esta classificação é a área ardida, mas aspetos como imprevisibilidade e capacidade de ser suprimido também são importantes para a definição de severidade de um fogo.

A velocidade de propagação do fogo (*rate of spread* - ROS) mede a velocidade de propagação do perímetro do incêndio [42]. Esta medida depende do tipo de combustível disponível, da topografia do local e das condições meteorológicas, especialmente o vento. Um aumento rápido do perímetro do fogo dificulta as ações de combate e dificulta o processo de evacuação dos habitantes locais [43]. O ROS pode atingir valores de 450 m/min e é considerado extremo quando ≥50 m/min.

Uma grandeza útil para medir a imprevisibilidade do fogo é a intensidade da linha de incêndio (*fire line intensity* - FLI), que mede a quantidade de energia, em Joule, emitida por unidade de tempo e de comprimento da linha de incêndio. Este valor pode ser calculado com o valor do ROS e de combustível consumido, pode ser estimado através do comprimento da linha de incêndio (*flame length* - FL) ou a partir do FRP (*fire radiative power*), taxa de calor radiante estimada por radiómetros instalados em satélites.

Embora estes parâmetros sejam importantes para caracterizar um incêndio florestal, devido à falta dos dados, e por a maior parte das definições terem como base este valor, será usado apenas a área ardida. As classificações baseadas em intervalos de área ardida são subjetivas e variam por região. Uma solução é usar percentis da distribuição estatística da dimensão dos incêndios na região. Será utilizado o percentil 90% dos incêndios em Portugal Continental o que corresponde a uma área ardida de cerca de 100 hectares, sendo estes considerados como grandes fogos [17].

2.2 Índices meteorológicos de risco de incêndio

2.2.1 Índice de Haines contínuo

O índice de estabilidade da baixa atmosfera proposto em [21] (*LASI - lower atmospheric stability index*), mais conhecido por índice de Haines (*HI - Haines index*) descreve a estabilidade da baixa atmosfera através de valores de temperatura em diferentes níveis de pressão. O índice é composto por dois componentes. O termo A, a diferença de temperatura do ar entre dois níveis de pressão atmosférica $P1$ e $P2$.

$$A = (TP1 - TP2)$$

Este termo mede a estabilidade da baixa atmosfera, onde quanto maior a diferença, maior a instabilidade da atmosfera.

O componente B,

$$B = (TP3 - TdP3)$$

mede a diferença entre a temperatura do ar num nível de pressão atmosférica $P3$ e a temperatura do ponto de orvalho nesse mesmo nível. Quanto maior a diferença mais "seca" a atmosfera e mais favorável à propagação do fogo.

Os valores de $Pi(i = 1,2,3)$ variam consoante a elevação do local de interesse seja baixa, média ou alta.

Após o cálculo das diferenças A e B os valores são transformados num número inteiro, de acordo com regras tabeladas [21, 38].

Os dados necessários para calcular o índice de Haines podem ser obtidos por rádio sondagem ou a partir de um modelo atmosférico. Em Portugal, os valores são obtidos pelo modelo AROME ou ECMWF [6, 7].

Pelo facto dos valores tabelados pelo índice de Haines, que apresenta valores de 2 a 6, terem sido obtidos empiricamente com base em dados do Estados Unidos existe um potencial problema de usar este índice em outras regiões, problema que foi verificado em [48, 29, 32]. Estes estudos mostram que o *HI* satura facilmente, não se observando discriminação suficiente dos dias com instabilidade e secura fora do comum.

O índice de Haines contínuo (*CHI - continous Haines index*) [34] tem uma composição semelhante ao *HI*, sendo composto por um componente CA, que mede a instabilidade da atmosfera e um componente CB responsável pela humidade do ar.

$$CA = 0.5(T850 - T700) - 2$$

$$CB = 0.3333(T850 - DP850) - 1$$

$$CHI = CA + CB$$

Onde $T850$ e $T700$ são as temperaturas nos níveis com pressão atmosférica de 850 hPa e 700 hPa; e $DP850$ é a temperatura do ponto de orvalho a 850 hPa. Foi verificado que o

valor de CB pode atingir valores desproporcionalmente grandes, sendo por isso aplicadas as seguintes restrições:

$$se\ (T850 - DP850)\ >\ 30°C,\ \text{então}\ (T850 - DP850)\ =\ 30°C$$

e

$$se\ CB > 5.0,\ \text{então}\ CB = 5.0 + (CB - 5.0)/2$$

O índice de Haines contínuo toma valores entre 0 e 14 onde quanto mais alto o índice mais erraticamente é provável que o incêndio se comporte, como pode ser observado na tabela da Figura 2.2.

Significado dos valores do Índice de Haines Contínuo (CHI) no comportamento do fogo.

CHI	**Comportamento provável do fogo e confiança da predição do fogo**
<4	Fogo facilmente controlado. Os modelos predizem facilmente o trajecto do fogo
4-8	Fogos podem ser difíceis de controlar e o comportamento do fogo pode ser errático. É provável a modelação do comportamento do fogo ser próxima da realidade
8-10	Fogos serão de difícil controlo e o comportamento do fogo será errático. É provável a modelação do comportamento do fogo subestimar a realidade
>10	Fogos não controláveis e extremamente difíceis de extinguir. É provável a modelação do comportamento do fogo subestimar dramaticamente a realidade

Figura 2.1: Categorias de fogos atendendo ao índice de Haines contínuo (CHI). Reproduzido de Bugalho [2018a].

Por ser um valor contínuo, o *CHI* permite uma maior discriminação entre os valores de elevado risco, sendo mais apropriado para climas mediterrânicos [7].

2.3 Índice canadiano de risco de incêndio florestal

O FWI permite avaliar o risco de incêndio com base nos valores meteorológicos: temperatura do ar, humidade relativa do ar, velocidade do vento e precipitação acumulada de 24 horas anteriores, para as 12 horas no fuso horário UTC. [45].

Como pode ser observado na Figura 2.3 este índice é composto por seis elementos. Os três índices primários, denominados códigos de humidade (*moisture codes*) avaliam o teor de humidade do solo, o que se traduz para combustível disponível para o fogo.

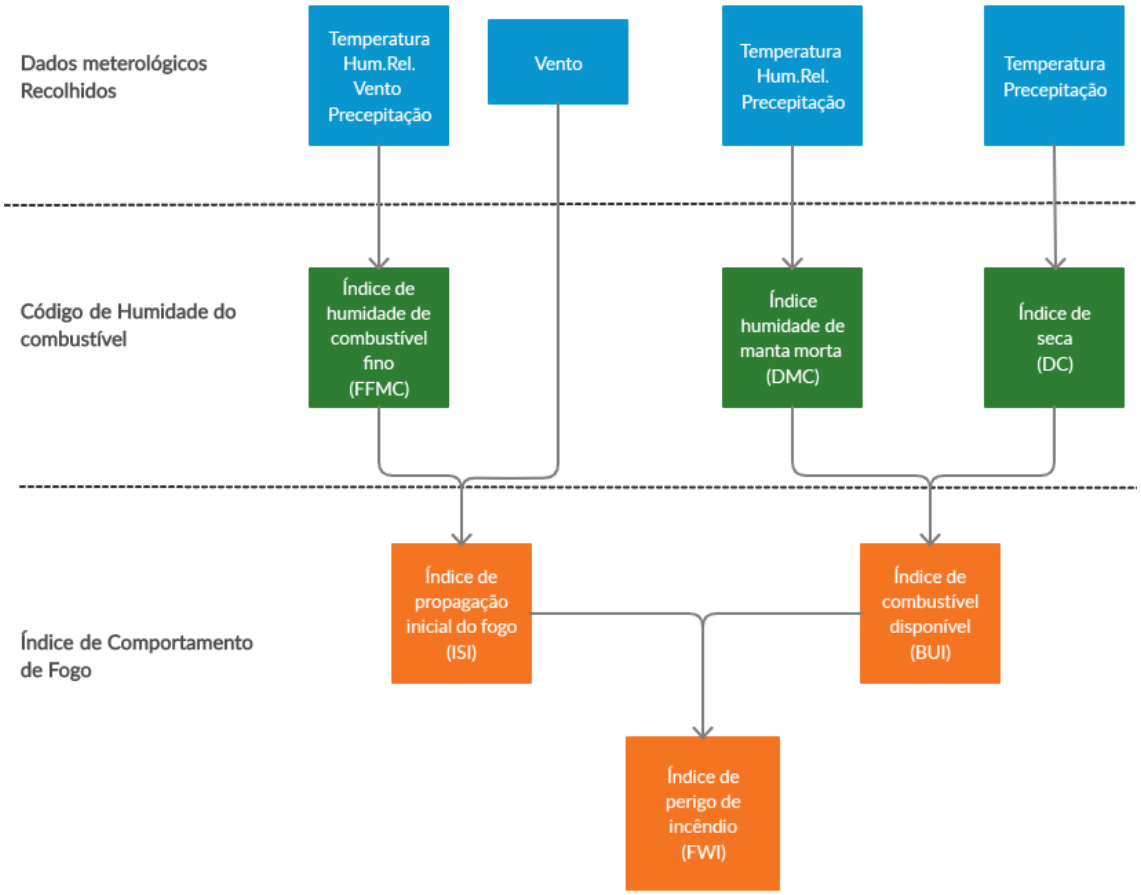


Figura 2.2: Estrutura do índice canadiano de risco de incêndio florestal (FWI, Forest Fire Weather Index). Reproduzido de [8]

- Índice de humidade de combustível fino *(FFMC - fine fuel moisture content)*

  Este índice avalia a humidade na camada mais fina e superficial do solo, 1-2 cm de profundidade [12]. Nesta camada encontra-se o combustível inicial para o possível fogo. A inflamabilidade deste combustível está relacionada com a humidade. Por ser referente à camada mais superficial, este subíndice é o que mais é afetado por fatores meteorológicos como a temperatura, humidade relativa, vento e precipitação. O FFMC toma valores entre 0 e 99, sendo o único subíndice com um limite final na escala. De acordo com [12] os fogos começam com valores próximos a 70.

- Índice de humidade de manta morta *(DMC - duff moisture code)*

  Este subíndice considera a humidade das camadas de material orgânico pouco compactado, a profundidade moderada, 5-10 cm. Por não estar em contacto com a superfície este índice não têm em conta a velocidade do vento. Este índice pode ser usado para prever a probabilidade de fogos iniciados por relâmpagos. O DMC é calculado tendo em conta os seus valores em dias anteriores, tendo assim um efeito cumulativo das condições anteriores.

Tabela 2.1: Classificação do índice FWI de acordo com a escala IPMA e dos restantes índices de acordo com a escala EEFIS [15]

Risco	**FWI**	**FFMC**	**DMC**	**DC**	**ISI**	**BUI**
Muito baixo	8.5	82.7	15.7	256.1	3.2	24.2
Baixo	8.5 - 17.2	82.7 - 86.1	15.7 - 27.9	256.1 - 334.1	3.2 - 5.0	24.2 - 40.7
Moderado	17.2 - 24.6	86.1 - 89.2	27.9 - 53.1	334.1 - 450.6	5.0 - 7.5	40.7 - 73.3
Alto	24.6 - 38.3	89.2 - 93.0	53.1 - 140.7	450.6 - 749.4	7.5 - 13.4	73.3 - 178.1
Muito alto	38.3 - 50.0	≥ 93.0	≥ 140.7	≥ 749.4	≥ 13.4	≥ 178.1
Extremo	50.0 - 64.0					
Máximo	≥ 64.0					

- Índice de seca *(DC - drought code)*

  Este subíndice estima a humidade das camadas de material orgânico compactado, a 10-20 cm de profundidade. Devido à profundidade a humidade relativa e o vento não afetam este índice, sendo este apenas afetado pela temperatura e pela precipitação. Este índice contém também um efeito cumulativo das condições anteriores por ser calculado tendo em conta os valores de dias anteriores. O índice de seca é um indicador do potencial de reacendimento e da possibilidade de ocorrência de fogo subterrâneo [16, 47].

Os três índices primários de humidade do combustível são usados para calcular os dois índices intermédios de comportamento de fogo.

- Índice de propagação inicial do fogo *(ISI - initial spread index)*

  Este índice combina o combustível disponível à superfície (FFMC) com a velocidade do vento para avaliar a propagação inicial do fogo.

- Índice de combustível disponível *(BUI - buildup index)*

  Combina o DMC e o DC para representar o total de combustível disponível

Com os valores do ISI e BUI, representando a velocidade de propagação inicial com o combustível a ser consumido é calculado o FWI que representa a intensidade potencial da frente do fogo. Este valor é útil para determinar os requisitos necessários para suprimir o fogo. A Tabela 2.1 apresenta as escalas de classificação dos índices utilizada neste trabalho.

Para além do FWI e dos subíndices conhecidos é muitas vezes utilizado o DSR (*Daily Severity Index*), uma função do FWI

$$DSR = 0.0272(FWI)^{1.77}$$

que pode ser acumulado, ao contrário do FWI. O DSR, por poder ser acumulado indicando o risco acumulado num período de tempo, permite comparar os anos ou períodos de meses em termos de risco acumulado.

2.4 Índice de perigosidade

- **Índices de risco estrutural** Este índice classifica uma área em cinco classes de perigosidade tendo em conta o seu declive, altitude e vegetação. Este índice pode ser encontrado em cartas publicadas pelo Instituto de Conservação da Natureza e das Florestas (ICNF) com periodicidade decadal.

- **Índices de risco conjuntural** Este índice, também publicado pelo ICNF, classifica uma área em seis classes tendo em conta o risco estrutural e o efeito dos incêndios ocorridos nos últimos três anos.

A metodologia e o modo de apresentação dos índices foram alterados em 2020 de acordo com as definições anteriores. Para os anos de 2012 a 2019, com exceção de 2018 em que não foi produzido, existe apenas a carta de perigosidade estrutural com periodicidade anual. De acordo com a nova metodologia, onde a produção da carta de risco estrutural foi alterada para um período decenal, existe, para 2020 e 2021 (Figura 2.3) cartas de risco conjuntural e uma carta de risco estrutural para a década 2020-2030 [23].

2.5 Índices de vegetação

Índices de vegetação são um conjunto de índices que medem propriedades da vegetação numa determinada área utilizando espectro da radiação solar refletidas pela mesma. Em concreto estes índices são estimados com base na diferença de reflexão para 2 ou 3 comprimentos de onda. Um dos mais usados é o índice de vegetação normalizado (NDVI - *Normalized Difference Vegetation Index*) que avalia a saúde das vegetações a partir da diferença da reflexão nos comprimentos de onda vermelho e infravermelho próximo do espectro de radiação solar. O índice de vegetação saudável (VHI - *Vegetation Health Index*) avalia o nível de secura; este índice é útil devido à relação entre períodos de seca e de fogos. Estes índices, e outros, podem ser calculados através de valores obtidos por imagens de satélites ou de drones. A estes dados alguns índices acrescentam informação sobre temperatura.

2.6 Fontes de Extração dos Dados

Os dados dos índices FWI (e subíndices FFMC, DMC,DC,ISI e BUI), e CHI para os anos de 2001 a 2020, calculados com a temperatura a 2m, humidade relativa a 2m, componente horizontal e meridional do vento a 10m de análises do ECMWF (*European Centre for Medium-Range Weather Forecasts*) e das componentes convectiva e de larga escala da precipitação obtidas da previsão do modelo ECMWF para H+24, foram fornecidos pelo IPMA. Para cada dia do ano existe uma medição para cada ponto de uma malha geográfica

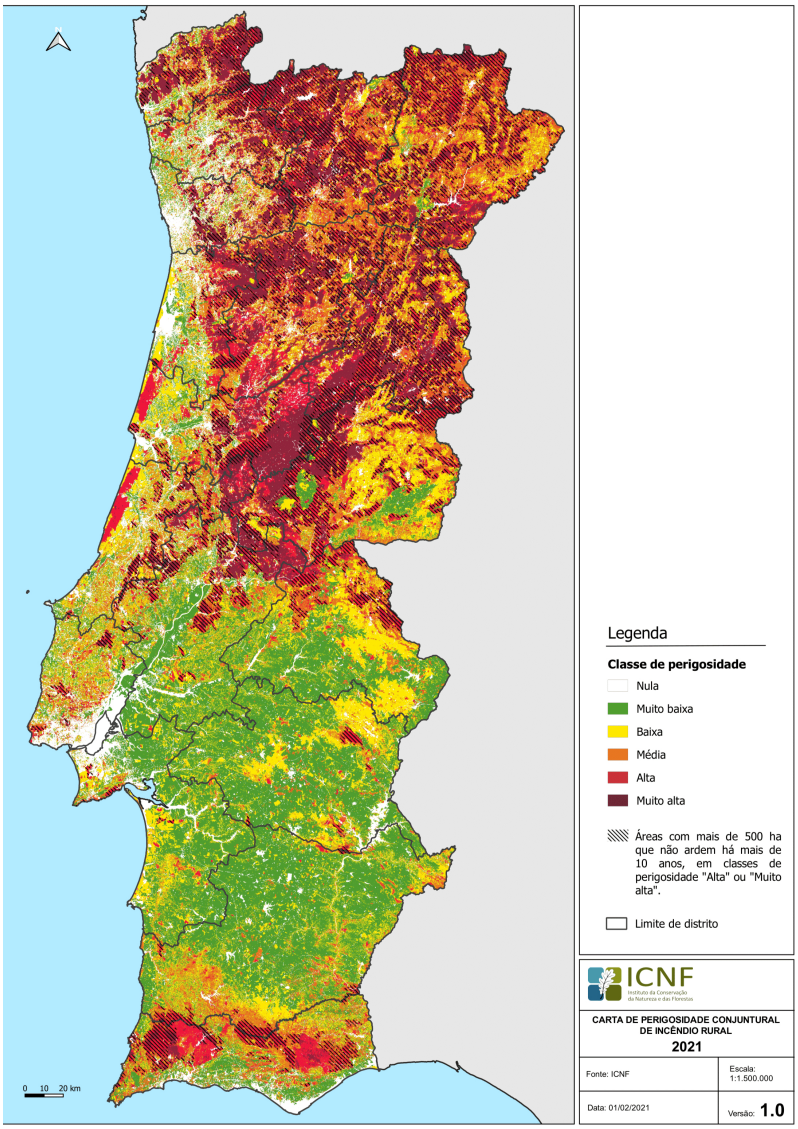


Figura 2.3: Carta de perigosidade conjuntural de incêndio rural 2021 [23]

regular de 0,125º de latitude e de 0,125º de longitude, cobrindo o território de Portugal Continental.

Para obter a área ardida foram usados os *shapefiles* disponibilizados pelo Instituto da Conservação da Natureza e das Florestas (ICNF) referentes aos anos entre 1990 a 2020. Os *shapefiles* contem informação, como as coordenadas da localização do incêndio, a área ardida, as datas em que começou e terminou, entre outras.

Por não terem um formato uniforme e, nalguns casos, carecerem de informação essencial, adicionou-se informação das áreas ardidas de outras fontes. Existem diversas fontes que mantêm repositórios com os dados das áreas ardidas obtidos através do satélite MODIS. Um desses repositórios é disponibilizado pelo Copernicus, um "programa europeu para a criação de uma capacidade europeia de observação da Terra"[10]. Outro repositório que pode ser usado é o disponibilizado pela Universidade de Maryland [20].

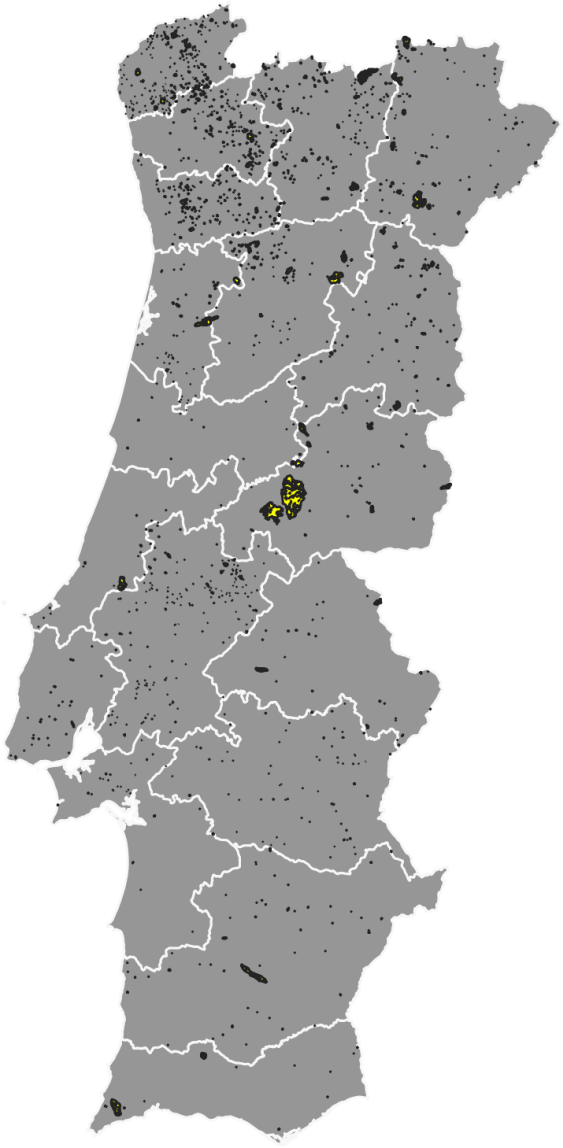


Figura 2.4: *Shapefile* das áreas ardidas de 2020, reproduzido a partir do QGis.

Os vários índices de vegetação podem ser obtidos com dados do satélite MODIS, disponíveis em vários repositórios.

Referente ao terreno existem vários dados que podem ser usados:

1. Elevação do terreno

   Pode ser obtido através de modelos digitais de elevação (DEM - *Digital Elevation Models*). A *Shuttle Radar Topography Mission* (SRTM), disponibiliza modelos com 30 metros de resolução para 80% do globo [40, 41].

2. Cobertura dos solos

   O *CORINE Land Cover* (CLC) é um inventário da cobertura dos solos europeus em 44 classes, com resolução de 5 hectares.

3. Inventario florestal

   Inventário florestal nacional produzido pelo ICNF inclui uma distribuição geográfica das espécies.

3

Agrupamento difuso: fuzzy c-means

O algoritmo *fuzzy c-means* é um algoritmo de partição difusa amplamente utilizado em áreas como a medicina, bioinformática, economia e psicologia [36]. Um estudo, relevante para este trabalho, onde o algoritmo apresentou resultados positivos foi na classificação de risco para departamentos florestais na Grécia de acordo com o número de incêndios e a área ardida [24].

Os algoritmos de agrupamento têm como objetivo dividir os dados em grupos, denominados *clusters*, a partir das características intrínsecas dos pontos. Um *cluster* é então um grupo de pontos com maior semelhança entre si do que com pontos de outros *clusters* [1]. Uma interpretação matemática da semelhança entre dois pontos amplamente utilizada é a distância.

Os algoritmos de agrupamento podem ser divididos em dois grupos de acordo com a forma que os agrupamentos são representados. Os agrupamentos rígidos (*hard clustering*) ou não-difusos, classificam uma observação como pertencente a apenas um grupo e os agrupamentos difusos (*soft clustering*) que atribuem a uma observação um nível de pertença a cada partição.

O agrupamento difuso permite aproximar os agrupamentos frequentemente observados em situações reais onde as fronteiras entre grupos apresentam alguma incerteza e não são divididos com fronteiras fixas.

3.1 O algoritmo fuzzy c-means (FCM)

O objetivo do algoritmo *fuzzy c-means* é particionar os dados em um número c de *clusters*. O resultado deste algoritmo é a partição difusa dos dados, esta partição pode ser representada por uma matriz $c \times N, \mathbf{U} = [\mu_{i,k}]$ onde $\mu_{i,k}$ indica o grau de pertença da observação k para cada um dos *clusters* i.

Partindo da matriz U é possível obter os centroides dos *clusters*. Os centroides são os pontos representativos de cada grupo, estes pontos são obtidos calculando a média dos pontos pertencentes ao *cluster* ponderada com o valor de pertença ao mesmo. Quanto maior o nível de pertença de um ponto a um *cluster*, maior o seu efeito na determinação

do centroide.

Este algoritmo tem como base a minimização da função objetivo:

$$J(\mathbf{X}; U, V) = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{i,k})^m \|\mathbf{x}_k - \mathbf{v}_i\|_{\mathbf{A}}^2$$

onde U é a matriz de pertença, V a matriz dos centroides e m um número real maior ou igual a 1 que define o grau de *fuzzyness* da partição.

O FCM pode ser descrito pelos seguintes passos:

- Dado um conjunto de dados X, é escolhido um número de clusters $1 < c < N$, o expoente de fuzzyness $m > 1.0$, o erro de paragem $\epsilon > 0$ e a matriz para o cálculo de distâncias A.
- Inicializar a matriz de partição, $\mathbf{U}^{(0)}$, aleatoriamente
- Repetir para cada iteração l
  1. Calcular os centroides

$$\mathbf{v}_i^{(l)} = \frac{\sum_{k=1}^{N} \left(\mu_{i,k}^{(l-1)}\right)^m \mathbf{x}_k}{\sum_{k=1}^{N} \left(\mu_{i,k}^{(l-1)}\right)^m}, 1 \leq i \leq c$$

  2. Calcular as distâncias dos pontos aos centroides

$$D_{i,k\mathbf{A}}^2 = (\mathbf{x}_k - \mathbf{v}_i)^T \mathbf{A} (\mathbf{x}_k - \mathbf{v}_i), \quad 1 \leq i \leq c, \quad 1 \leq k \leq N$$

  3. Atualizar a matriz de partição

$$\mu_{i,k}^{(l)} = \frac{1}{\sum_{j=1}^{c} \left(D_{i,k\mathbf{A}} / D_{j,k\mathbf{A}}\right)^{2/(m-1)}}$$

- Repetir até $\left\|\mathbf{U}^{(l)} - \mathbf{U}^{(l-1)}\right\| < \epsilon$

Uma métrica comum para o valor da matriz A é $A = I$, induzindo a norma Euclidiana. Neste caso, o cálculo das distâncias é simplificado para:

$$D_{i,k\mathbf{A}}^2 = (\mathbf{x}_k - \mathbf{v}_i)^T (\mathbf{x}_k - \mathbf{v}_i)$$

Este algoritmo é sensível à inicialização, diferentes matrizes iniciais $\mathbf{U}^{(0)}$ podem originar diferentes matrizes de partição finais. Um processo de validação usado para amenizar este problema é correr o algoritmo múltiplas vezes e selecionar a partição que obteve o melhor valor da função objetivo. Outra desvantagem deste algoritmo é a necessidade de especificar o número de grupos, esta dificuldade é resolvida, dependendo do problema, ao estudar o conjunto de dados ou, como é mais frequente, a comparar partições com diferentes números de grupos.

3.2 Validação de partições FCM

O objetivo dos índices de validação é associar à qualidade de uma partição um valor numérico permitindo a comparação entre múltiplos agrupamentos. Os índices de validação podem-se dividir entre índices de validação internos e externos.

Nos índices de validação externos é efetuada uma comparação com informação exterior aos dados, como a classificação "real" dos mesmos. Pelo contrário, os índices de validação internos utilizam apenas as propriedades intrínsecas das partições como a compactação, nível de separação e diferença dos grupos formados.

Um índice de validação interna utilizado para partições difusas é o índice de Xie-Beni [49]. Este índice tem como objetivo quantificar o rácio total da variação de pontos pertencentes a um *cluster* e a separação dos vários *clusters*, através da seguinte formula:

$$XB(c) = \frac{\sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{i,k})^m \|\mathbf{x}_k - \mathbf{v}_i\|^2}{N \min_{i,k}\|\mathbf{x}_k - \mathbf{v}_i\|^2}$$

3.3 Visualização de partições FCM

Os algoritmos de agrupamento, particionam os dados mesmo quando este procedimento não se adequa aos dados. Os índices de validação reduzem a qualidade do agrupamento a apenas um valor o que pode não permitir detetar alguns problemas devido à perda de informação. Existe então uma utilidade na informação obtida com a análise gráfica das partições, mas, caso os dados tenham uma dimensão superior a três é necessário realizar uma redução de dimensionalidade.

O método de *Sammon mapping* é um método de redução de dimensionalidade que tenta preservar a distância pontos. Este algoritmo reduz a dimensionalidade dos dados de n para q dimensões ao encontrar N pontos na dimensão q que melhor aproximam a distância entre pontos no espaço original de n dimensões. Isto é obtido com a minimização da função de erro denominada *Sammon's stress*:

$$E = \frac{1}{\lambda}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{\left(d_{i,j} - d^*_{i,j}\right)^2}{d_{i,j}},$$

onde i e j são pontos distintos, $d_{i,j}$ é a distância dos pontos no espaço de dimensões n e $d^*_{i,j}$ é a distância no espaço de dimensões q.

O *fuzzy Sammon mapping*, adapta o método de *Sammon mapping* para partições difusas. Para tal utiliza a característica das partições difusas onde apenas são consideradas as distâncias entre os pontos e os centroides, ignorando a distância entre pontos, sendo utilizado apenas $N \times c$ distâncias. Tendo em conta esta simplificação a função de erro deste algoritmo, baseado no *Sammon's stress* é dada por:

$$E_{fuzz} = \sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{i,k})^m \left(d(\mathbf{x}_k, \eta_i) - d^*(\mathbf{y}_k, \mathbf{z}_i)\right)^2,$$

onde $d(\mathrm{x}_k, \eta_i)$ é a distância entre a observação x_k e o centroide v_i medida no espaço original e $d^*(\mathrm{x}_k, \eta_i)$ corresponde à distância na projeção. Desta forma, no espaço projetado cada *cluster* é representado por um ponto e resultara em *clusters* com um formato, aproximadamente, esférico. Esta projeção é facilmente interpretada devido ao uso da distância Euclidiana.

No algoritmo de *fuzzy c-means*, o grau de pertença de um ponto está associado a uma distância, por esse motivo, estes graus também podem ser ilustrados utilizando a equação:

$$\mu^*_{i,k} = 1/\sum_{j=1}^{c}\left(\frac{d^*(\mathrm{x}_k, \eta_i)}{d^*\left(\mathrm{x}_k, \eta_j\right)}\right)^{\frac{2}{m-1}}$$

4

Classificadores baseados em árvores e extração de Regras

O objetivo da extração de regras é retirar de um modelo uma descrição passível de ser interpretada [22]. A descrição extraída tem de ser compreensível, mas também manter a capacidade preditiva do modelo. Nos métodos aqui utilizados esta descrição é feita com o uso de regras condicionais, mas também pode ser feita com árvores de decisão ou outros modelos gráficos. As regras condicionais são um conjunto de regras na forma:

IF c_1 e c_2 e, ..., c_k
THEN Previsão = p

A parte "IF" da regra é uma conjunção de várias condições (c) que testam se a observação tem ou não uma propriedade especifica. A parcela "THEN" contem a previsão da regra caso todas as condições sejam sucedidas.

4.1 Árvores de decisão

Árvores de decisão [33] são um modelo que permite dividir dados numa estrutura semelhante à de uma árvore, composto por uma coleção de nós, conectados por ramos. Em cada nó interno da árvore um parâmetro é testado, e cada resultado possível resulta num ramo. Cada ramo liga um nó a outro nó interno, ou a um nó-folha. Nós-folha são o resultado do conjunto de regras que levou a esse nó, são o "*output*" da árvore de decisão e representam um ou um grupo de dados.

O objetivo das árvores de decisão é dividir o conjunto de dados em grupos da melhor forma possível. Para construir a árvore são realizadas sucessivas divisões dos dados em subconjuntos disjuntos até que a divisão não acrescente nenhum valor às previsões. As divisões, realizadas nos nós internos são efetuadas no parâmetro que minimiza uma função de custo. Esta função varia consoante a implementação, o algoritmo e o tipo de árvore. Existem dois tipos de árvores de decisão, de acordo com o que os nós-folha representam, árvores de classificação e árvores de regressão [27].

4.1.1 Árvores de classificação

As árvores de classificação dividem os dados em grupos classificados. O resultado da árvore de classificação para um elemento é uma etiqueta que descreve, em algum grau, esse elemento.

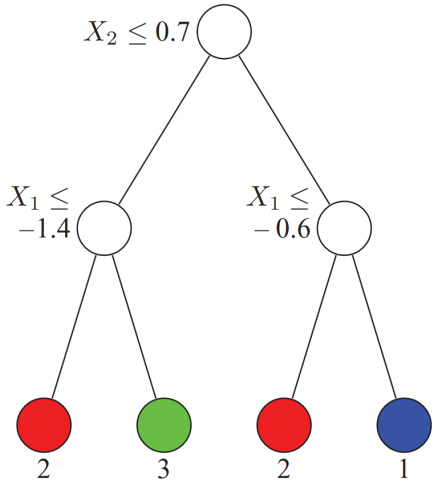


Figura 4.1: Exemplo de árvore de classificação. Adaptado de [28]

Para árvores de classificação as duas principais fórmulas de avaliação, usadas para decidir qual a divisão ótima, são a impureza de Gini e a entropia de informação.

A **impureza de Gini** mede a probabilidade de um grupo de dados ser classificado de forma errada se a classificação for feita de forma aleatória de acordo com a distribuição do subconjunto. A impureza de Gini para um elemento é dada por:

$$G = \sum_k P(k) \times (1 - P(k))$$

onde $P(k)$ é a probabilidade da classe k no conjunto. A melhor divisão possível seria uma com um valor de $G = 0$, onde as classes seriam perfeitamente divididas.

Para um conjunto de elementos, com J classes a impureza de Gini é dada por:

$$\mathrm{I}_G(p) = \sum_{i=1}^{J}\left(p_i \sum_{k\neq i} p_k\right) = \sum_{i=1}^{J} p_i\,(1-p_i) = \sum_{i=1}^{J}\left(p_i - p_i^2\right) = \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} p_i^2 = 1 - \sum_{i=1}^{J} p_i^2$$

sendo p_i a fração de elementos com classe i.

A **entropia de informação** baseia-se no princípio da entropia e ganho de informação existentes na teoria de informação. Para um conjunto com J classes onde p_i com $i \in \{1, 2, \ldots, J\}$ é a fração de elementos com classe i a entropia é definida como:

$$\mathrm{H}(T) = \mathrm{I}_E\left(p_1, p_2, \ldots, p_J\right) = -\sum_{i=1}^{J} p_i \log_2 p_i$$

O ganho de informação ao dividir sobre um atributo a será a diferença entre a entropia do conjunto e a soma das entropias dos subconjuntos resultantes da divisão.

$$\overbrace{IG(T,a)}^{\text{Ganho de informação}} = \overbrace{\mathrm{H}(T)}^{\text{Entropia (nó ascendente)}} - \overbrace{\mathrm{H}(T \mid a)}^{\text{Soma das entropias (nós descendentes)}}$$

$$= -\sum_{i=1}^{J} p_i \log_2 p_i - \sum_{i=1}^{J} -p_i(i \mid a) \log_2 p_i(i \mid a)$$

A melhor divisão é a que apresenta uma maior redução da entropia e, consequentemente, um maior ganho de informação.

4.1.2 Árvores de regressão

Nas árvores de regressão os nós-folha apresentam o melhor valor numérico associado com o nó.

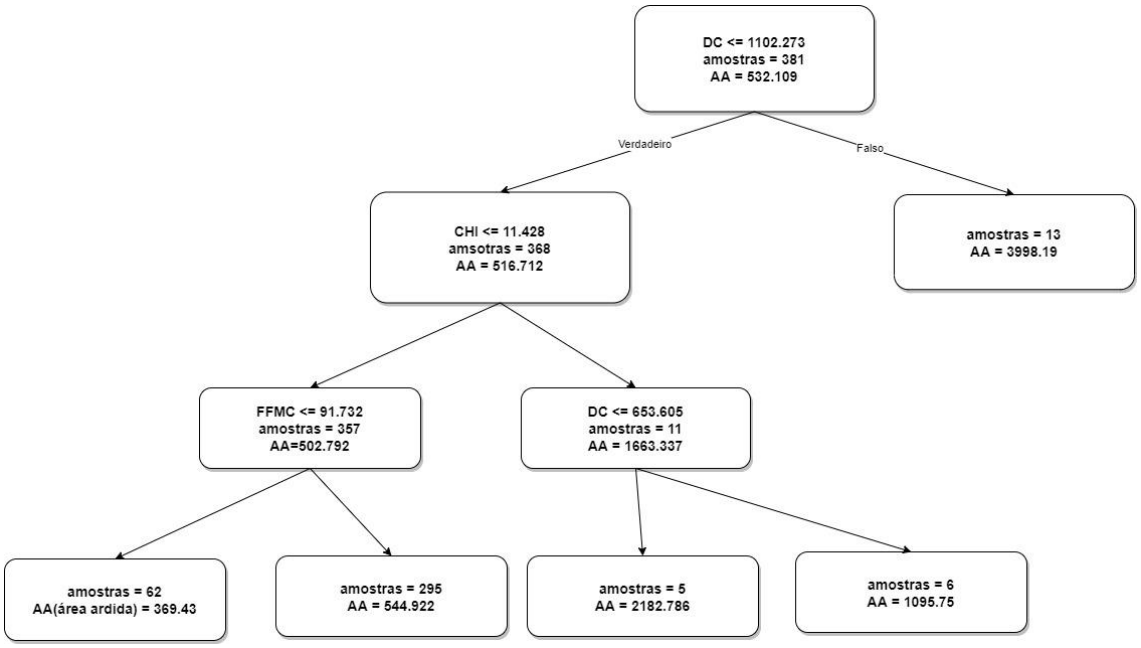


Figura 4.2: Exemplo de árvore de regressão para estimar área ardida. Reproduzido de [8]

Uma das funções de custo das mais usada para árvores de regressão é o erro quadrático médio (*MSE - Mean square error*):

$$\mathrm{MSE}_{\text{node}} = \sum_{i \in \text{node}} \left(\hat{y}_{\text{node}} - y^{(i)}\right)^2$$

$$\hat{y}_{\text{node}} = \frac{1}{m_{\text{node}}} \sum_{i \in \text{node}} y^{(i)}$$

este valor calcula a média dos quadrados dos erros entre valores espectáveis e valores reais.

4.1.3 Variantes de extração de regras para árvores de decisão

Os algoritmos de criação de árvores de decisão tentam encontrar a árvore que melhor se ajusta aos dados de treino, um problema comum é o de sobreajuste, onde o modelo se ajusta demasiado ao conjunto de treino diminuindo a sua capacidade de generalização.

A poda de uma árvore de decisão é um método que tem como objetivo reduzir o sobreajuste da mesma. Os métodos de poda podem-se dividir em duas categorias, pré-poda, onde durante o processo de treino alguns exemplos são ignorados e consequentemente não explicados pela árvore e a pós-poda onde uma árvore é treinada para explicar todos as instâncias de treino e em seguida alguns ramos são retirados.

De acordo com [30] o RIPPER resulta em regras com maior interpretabilidade e estabilidade quando comparado com regras obtidas por árvores puras, sem poda, obtidas com o algoritmo CART (*Classification And Regression Trees*).

4.1.3.1 REP

REP (*Reduced Error Pruning*) [5] é uma técnica de pós-poda de árvores de decisão. Neste método os dados de treino são inicialmente divididos em dois conjuntos, um *growing set* e um *pruning set*. De seguida é gerado um modelo que explique todas as instâncias do *growing set*. O modelo gerado é então generalizado com a remoção de condições que maximizem a redução de erro no *pruning set*. A poda termina quando não existir nenhuma remoção possível que diminua o erro.

4.1.3.2 IREP

O IREP (*Incremental Reduced Error Pruning*) [19] é um método que tenta resolver alguns problemas do REP descritos pelos autores. Este método integra ambos pré-poda e pós-poda. A principal diferença é que cada regra é podada logo após ser gerada.

Tal como no REP inicialmente os dados de treino são divididos em dois conjuntos, o *growing set* e um *pruning set*. De seguida é gerada uma regra com o *growing set* e são removidas condições de forma a maximizar a precisão no *pruning set*. Após a poda todos exemplos cobertos pela regra, positivos ou negativos são removidos dos dados de treino. Os dados de treino são outra vez divididos e usados para gerar uma nova regra. Este processo é repetido até que a precisão da regra após a poda seja inferior à regra vazia.

4.1.3.3 RIPPER

O algoritmo RIPPER (*Repeated Incremental Pruning to Produce Error Reduction*) [9] é baseado no algoritmo IREP com alterações de forma a aumentar a precisão e diminuir o tempo computacional.

Este método implementa três modificações em relação ao IREP:

- **Métrica de valor de regra**
  A métrica utilizada para podar uma regra no IREP dado uma regra *Rule*, e os exemplos cobertos pela regra, positivos e negativos *PrunePos* e *PruneNeg* respetivamente é a seguinte

$$v(\text{ Rule, PrunePos, PruneNeg }) \equiv \frac{p + (N - n)}{P + N},$$

  onde P e N são o número total de exemplos em *PrunePos* e *PruneNeg* respetivamente e p e n são o número de exemplos em *PrunePos* e *PruneNeg* respetivamente cobertos pela regra. Esta métrica dá preferência a uma regra que cobre 2000 exemplos positivos e 1000 negativos a uma que cobre 1000 positivos e um negativo, o que os autores consideram não intuitivo. O RIPPER substitui esta métrica pela seguinte:

$$v^*(\text{ Rule, PrunePos, PruneNeg }) \equiv \frac{p - n}{p + n}$$

- **Condição de paragem**
  No RIPPER após uma regra ser gerada a sua descrição de comprimento e a descrição de comprimento dos exemplos são calculadas. O algoritmo termina quando a descrição da nova regra for d bits maior que a descrição mais pequena obtida até agora. O algoritmo também termina quando não houver mais exemplos positivos.

- **Otimização de regras**
  O processo do IREP de sucessivas seleções de regras individuais seguidas de poda da mesma, pode produzir resultados que diferem de métodos de poda convencionais que analisam a árvore como um todo (ex. REP, *Reduced Error Pruning*). O RIPPER utiliza um método de otimização diferente. Para otimizar um conjunto de regras $R_1, \ldots, R_k$, cada regra é considerada individualmente na ordem em que foi gerada. Para cada regra R_i duas alternativas são construídas. A regra para substituir R_i é obtida por gerar e podar uma regra R_i' onde a poda é feita de forma a minimizar o erro do conjunto de regras $R_1, \ldots, R_i', \ldots, R_k$. É construída outra regra, chamada de revisão de R_i onde se adicionam condições à regra R_i. De seguida é decidido se é melhor manter R_i ou alterar por R_i' ou pela revisão de R_i.

4.2 Florestas aleatórias

Outra possível solução para o problema de sobreajuste das árvores de decisão é o uso de métodos de reunião (*ensemble methods*) como as florestas aleatórias.

Floresta aleatória (FA) [4] é um método de reunião que combina várias árvores independentes, treinadas com um subconjunto aleatório do conjunto de treino. Os subconjuntos são obtidos por meio de um método de nome *bootstrap aggregating* ou *bagging*. Cada subconjunto é obtido selecionando, de forma aleatória, elementos do conjunto de treino com reposição, existindo a possibilidade de selecionar um elemento mais do que uma vez.

Outra forma que este método utiliza para colocar aleatoriedade no modelo, de forma a diminuir o *overfitting*, é o uso de parâmetros aleatórios. Em cada nó a divisão apenas pode ser feita sobre um parâmetro de um subconjunto aleatório de todos os parâmetros. Para problemas de classificação o resultado da floresta é a classe escolhida por o maior número de árvores, para problemas de regressão, é a média de todos os resultados. Este modelo já foi aplicado, com dados de satélite, para classificação de terreno [44, 11, 46].

O uso de FA permite obter modelos com maior precisão, mas retira a facilidade da interpretação das árvores de decisão [3]. Uma informação sobre os dados que se pode retirar das FA é uma classificação da importância dos parâmetros para a decisão final. Uma medida possível de ser usada para classificar a importância dos parâmetros é a diminuição média da precisão (MDA - *Mean Decrease in Accuracy*), este valor mede o decréscimo de precisão do modelo ao ser excluído um parâmetro. Esta métrica permite classificar os parâmetros de acordo com a sua importância nos resultados da floresta, mas esta informação não apresenta uma explicabilidade do conjunto de dados muito significativa [37]. Uma forma de extrair mais informação explicativa é o uso de métodos de extração de regras para florestas aleatórias.

4.3 Métodos de extração de regras para florestas aleatórias

4.3.1 SIRUS

O algoritmo *Stable and Interpretable RUle Set* - SIRUS proposto em [2] é um algoritmo de classificação baseado em florestas aleatórias e que resulta numa lista de regras. O algoritmo é composto pelos quatro passos seguintes.

4.3.1.1 Geração de regras

O SIRUS utiliza, para gerar as regras, uma floresta aleatória com uma alteração ao método original. No método original cada árvore da floresta é gerada de uma forma *greedy* no valor onde é feita a divisão de cada nó. No caso do SIRUS, inicialmente são calculados os q-quantis da distribuição do conjunto de dados, normalmente q=10. Durante a criação das árvores a divisão em um nó só pode ser feito com um dos valores dos quantis. Esta alteração permite estabilizar a floresta aleatória sem observar alterações significativas na precisão.

Cada nó de cada árvore pode ser representado por um caminho que vamos denotar de $\mathscr{P}$. Um caminho descreve a sequência de divisões da raiz até ao nó. Sendo Π a lista de todos os caminhos possíveis, todo $\mathscr{P} \in \Pi$ pode ser interpretado como uma regra. São então extraídas todas as regras de cada árvore da floresta, este conjunto terá, usualmente, 10^4 regras, [2].

4.3.1.2 Seleção de regras

De seguida são selecionadas as regras mais relevantes tendo em conta a sua frequência, denotada por $\hat{p}_{M,n}(\mathscr{P})$ e um valor mínimo, o hiperparâmetro $p_0 \in (0,1)$

$$\hat{\mathscr{P}}_{M,n,p_0} = \{\mathscr{P} \in \Pi : \hat{p}_{M,n}(\mathscr{P}) > p_0\}$$

4.3.1.3 Pós-tratamento das regras

Devido à forma como as regras são extraídas da árvore, tendo em conta tanto nós internos como folhas, existe sobreposição entre regras na lista $\hat{\mathscr{P}}_{M,n,p_0}$. Neste passo são removidos todos os caminhos $\mathscr{P} \in \hat{\mathscr{P}}_{M,n,p_0}$ onde a regra induzida por este for uma combinação linear de regras associadas com outros caminhos de maior frequência.

4.3.1.4 Agregação de regras

O conjunto de regras extraído no passo anterior é composto por um número pequeno de regras, passível de interpretação. Para obter um modelo de classificação é feita a média do valor obtido pelas regras de forma a obter a estimativa da função objetivo: $\hat{\eta}_{M,n,p_0}(\mathbf{x}) = \frac{1}{|\hat{\mathscr{P}}_{M,n,p_0}|} \sum_{\mathscr{P} \in \hat{\mathscr{P}}_{M,n,p_0}} \hat{g}_{n,\mathscr{P}}(\mathbf{x})$, onde $\hat{g}_{n,\mathscr{P}}$ é uma regra associada a um caminho $\mathscr{P}$.

4.4 Medidas de avaliação de regras

O seguinte método, proposto em [30], tem como objetivo criar uma medida que permite comparar diferentes modelos preditivos. A pontuação de interpretabilidade (*interpretability score*) é uma média ponderada de três outras pontuações: previsibilidade, estabilidade e simplicidade.

4.4.1 Pontuação de previsibilidade

A pontuação de previsibilidade mede a precisão de um modelo independentemente do intervalo de valores possíveis da variável dependente.

Sendo $\mathscr{L}_n$ o risco empírico de um dado modelo preditivo, a pontuação de previsibilidade é definida como

$$\mathscr{P}_n(g_n, h_n) := 1 - \frac{\mathscr{L}_n(g_n)}{\mathscr{L}_n(h_n)}$$

onde g_n é o modelo que queremos analisar e h_n é um modelo *naive* com o qual o modelo g_n é comparado. Por exemplo, se $Y \in \mathbb{R}$ onde Y é a variável dependente, podemos usar a seguinte função

$$h_n = \frac{1}{n} \sum_{i=1}^{n} Y_i,$$

para um modelo de classificação pode ser utilizado a moda do conjunto de dados.

No caso de $\mathscr{P}_n(g_n, h_n) < 0$ o modelo *naive* teria melhor precisão que o modelo g_n, logo seria melhor utilizar o modelo *naive* como modelo preditivo. Para esse caso o novo modelo

g_n seria igual ao h_n, logo teria uma pontuação de previsibilidade igual a 0. Por esse motivo, podemos assumir então que a pontuação de previsibilidade é um valor entre 0 e 1.

4.4.2 Pontuação de estabilidade

A pontuação de estabilidade (*q-stability score*) quantifica a sensibilidade do modelo a ruído nos dados e permite avaliar a robustez do modelo.

Um modelo é estável se a partir de dois conjuntos de dados diferentes da mesma distribuição obter regras similares. Caso os parâmetros sejam contínuos obtiver dois conjuntos de regras semelhantes é praticamente impossível. Por esse motivo é necessário discretizar os parâmetros envolvidos nas regras.

Tendo X, um parâmetro contínuo, começamos por atribuir a cada inteiro $p \in \{1,\dots,q\}$, um intervalo $\left[x_{(p-1)/q}, x_{p/q}\right]$. De seguida é construído $Q_q(X)$, uma versão discreta do parâmetro X obtida pela substituição cada valor de X pelo valor p correspondente. Finalmente os intervalos das regras são também substituídos pelos valores discretos.

Sendo $\mathcal{A}$ um algoritmo e R_n and R'_n dois conjuntos de regras obtidos com esse algoritmo a partir de duas amostras independentes de uma distribuição $\mathbb{Q}$, a *q-stability score* é dado por

$$\mathcal{S}_n^q(\mathcal{A}) := \frac{2\left|Q_q(R_n) \cap Q_q(R'_n)\right|}{\left|Q_q(R_n)\right| + \left|Q_q(R'_n)\right|},$$

onde $Q_q(R)$ é a versão discreta do conjunto de regras R, e com a convenção que $0/0 = 0$.

A pontuação de estabilidade é um número positivo entre 0 e 1. Se os 2 conjuntos de regras não tiverem nenhuma regra em comum a pontuação de estabilidade será zero. Se tiverem as mesmas regras será um.

4.4.3 Pontuação de simplicidade

A pontuação de simplicidade (*simplicity score*) mede a facilidade como uma previsão é verificada.

Sendo o índice de interpretabilidade (*interpretability index*) definido em [31] como

$$\mathrm{Int}(g_n) := \sum_{r \in R_n} \mathrm{length}(r)$$

para um estimador g_n gerado por um conjunto de regras R_n, onde length(r) é o número de condições de uma regra.

A pontuação de simplicidade consiste em comparar o índice de interpretabilidade do algoritmo para o qual queremos calcular a pontuação com os índices de um conjunto de algoritmos $\mathcal{A}_1^m = \{\mathcal{A}_1,\dots,\mathcal{A}_m\}$ e é definida por

$$\mathbb{S}_n(\mathcal{A}_i, \mathcal{A}_1^m) = \frac{\min\left\{\mathrm{Int}\left(g_n^A : A \in \mathcal{A}_1^m\right)\right\}}{\mathrm{Int}\left(g_n^{\mathcal{A}_i}\right)}$$

Para calcular a pontuação de simplicidade de apenas um algoritmo é necessário obter um valor de *threshold*, um valor máximo para o tamanho das regras.

5

Extração e tratamento dos dados

5.1 Índices de risco de incêndio

Os dados referentes aos índices de risco de incêndio para os anos de 2001 a 2020 foram fornecidos pelo IPMA, calculados com base nas análises do ECMWF (*European Centre for Medium-Range Weather Forecasts*). Para cada dia do ano existe uma medição para cada ponto de uma malha geográfica regular de 0,125º de latitude e de 0,125º de longitude (≈ 193 km^2), cobrindo o território de Portugal Continental (Figura 5.1).

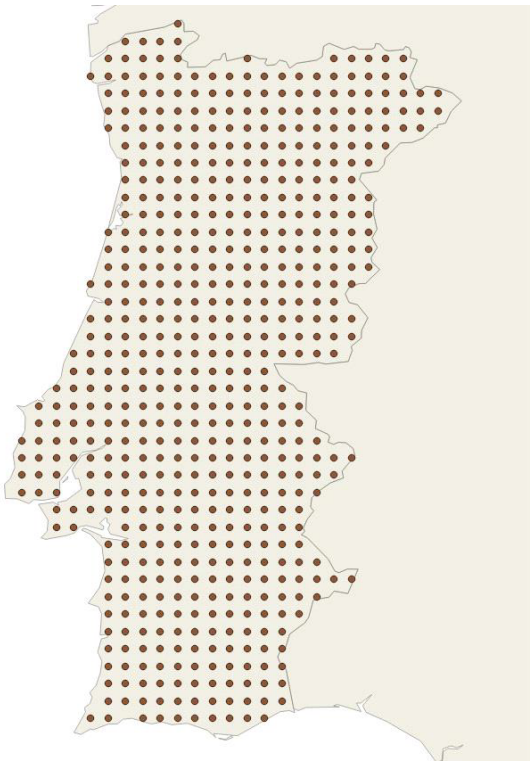


Figura 5.1: Malha geográfica dos pontos dos índices de risco de incêndio ao longo do território de Portugal Continental. Reproduzido de [8]

Os dados consistem num conjunto de ficheiros em formato *ASCII* para cada um dos índices, onde para cada dia existe um ficheiro com uma lista de medições em cada região

da malha identificada com um id. Os índices disponíveis são: o FWI e os subíndices (FFMC, DC, ISI, BUI e DMC), o CHI e o DSR.

5.2 Área ardida

Os dados referentes aos incêndios (data de início, coordenadas e área ardida) foram obtidos de duas fontes: os *shapefiles* disponibilizados pelo Instituto da Conservação da Natureza e das Florestas (ICNF) e os dados disponibilizados pela Universidade de Maryland com dados obtidos do satélite MODIS.

5.2.1 Incêndios ICNF

Apesar de estarem disponíveis dados desta fonte entre os anos 1990 a 2019, os incêndios de 1990 a 2011 apenas têm a informação do ano no qual o incêndio ocorreu, não podendo ser usados por esse motivo neste estudo. Foram utilizados os *shapefiles* referentes aos anos de 2012 a 2019 com exceção do ano 2017, por não estar disponível. Estes ficheiros contem as demarcações da área ardida no mapa, a data (dia, mês e ano) de início e de conclusão. Os dados foram extraídos dos *shapefiles* com o auxílio de algumas bibliotecas de *python*. A biblioteca *geopandas*, para leitura dos ficheiros .shp e as bibliotecas *pyproj* e *pycrs* para a leitura dos ficheiros de projeção .prj e transformações entre sistemas de coordenadas.

De cada um dos ficheiros anuais foram extraídos os incêndios com área ardida maior ou igual a 100 hectares. Nos ficheiros relativos a alguns anos está disponível a informação do tipo de incêndio (florestal, agrícola, queimada, reacendimento), nestes ficheiros apenas foram considerados os incêndios classificados como "florestal".

5.2.2 Incêndios Maryland

Este repositório, disponibilizado pela Universidade de Maryland [20] é composto por imagens recolhidas diariamente pelo satélite MODIS, com uma resolução de 500 m^2. Para cada ano o repositório contém 24 ficheiros .tiff para cada uma das 24 janelas espaciais apresentadas na Figura 5.2. Para este estudo foram usados os ficheiros .tiff relativos à zona da Europa, a janela espacial oito, aos quais se retirou uma área retangular contendo Portugal Continental.

Aos ficheiros .tiff foi aplicado o processo descrito em [20] com o auxílio das bibliotecas de *python geopandas*, *pyproj*, *pycrs* e *gdal*. Este processo consiste em unir as camadas referentes a cada mês em apenas uma camada e associar a cada incêndio o dia no qual ocorreu. Com a camada obtida a data foi convertida para o formato dia/mês/ano e a partir das delimitações do incêndio foi calculada a área ardida em hectares.

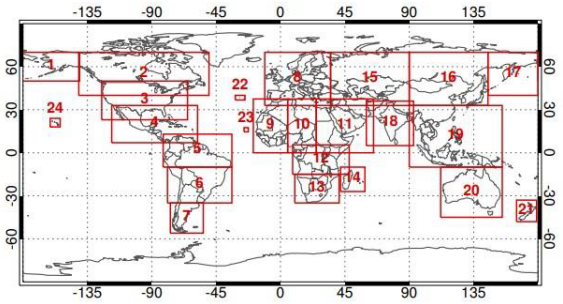


Figura 5.2: Esquema das 24 janelas espaciais [20]

5.2.3 Interceção com malha geográfica e união dos dados

De forma a obter os valores dos índices de risco para cada incêndio foi necessário fazer a sua sobreposição com a malha geográfica da Figura 5.1. Para cada incêndio foi calculada a percentagem de área contida em cada quadrícula e atribuído a cada incêndio um parâmetro com o identificador da quadrícula com maior sobreposição. Destes dados foram excluídos os incêndios sem sobreposição a nenhuma quadrícula por se encontrarem fora da área de Portugal Continental.

Com a informação de data de início do incêndio e o identificador da malha geográfica foram iterados os ficheiros com os índices de risco fornecidos pelo IPMA e extraídos os índices referentes a cada incêndio. Finalmente os dados originados das duas fontes foram unidos. Existindo anos com dados das duas fontes foi dado prioridade aos dados obtidos dos *shapefiles* disponibilizados pelo ICNF. O conjunto de dados final integra então os incêndios de 2012, 2013, 2014, 2015, 2016, 2018 e 2019 com origem nos dados do ICNF e os incêndios de 2001 a 2011, 2017 e 2020 com origem nos dados de Maryland. Este conjunto é composto por informação referente a 2949 incêndios com os parâmetros que podem ser vistos no excerto da Tabela 5.1

5.3 Índice de perigosidade

Para obter os índices de perigosidade foi utilizada a carta de perigosidade estrutural 2020-2030 disponibilizada em um ficheiro .tif pelo ICNF. Esta carta atribui uma classe de perigosidade de "Nula", "Muito baixa", "Baixa", "Média", "Alta" ou "Muito alta" a cada área de $25m^2$. No ficheiro .tif esta escala de risco é representada por um inteiro entre 0 e 5. De forma a obter um valor de perigosidade para incêndio foi calculada a sua sobreposição com a carta de perigosidade e atribuído um valor de perigosidade igual à média dos valores de risco da área.

5.4 Índices de vegetação

Os índices de vegetação (SMN - *Smoothed NDVI*, SMT - *Smoothed Brightness Temperature*, TCI - *Temperature Condition Index*, VCI - *Vegetation Condition Index* e VHI - *Vegetation*

Tabela 5.1: Excerto da tabela (2949 x 14) da área ardida e dos valores de risco de incêndio. Parâmetro **geometry** contém um objeto *python* com a informação da delimitação do incêndio, simplificado na tabela para leitura. O parâmetro **id:area** contem uma lista da razão de área ardida por quadrícula. O parâmetro **max_id** contém o identificador da quadrícula com maior área ardida.

date	**area_ha**	**geometry**	**source**	**id:area**
03/09/2005	2677.142	POLYGON	MARYLAND	[[3320, 0.239], [3321, 0.749], [3421, 0.012]]
01/09/2002	53.46362	POLYGON	MARYLAND	[[2512, 1.0]]
27/08/2013	181.8424	POLYGON	ICNF	[[3518, 0.269], [3519, 0.731]]
31/07/2001	127.6214	POLYGON	MARYLAND	[[3819, 1.0]]
03/09/2009	89.64348	POLYGON	MARYLAND	[[2918, 1.0]]

max_id	**CHI**	**FWI**	**ISI**	**DC**	**FFMC**	**BUI**
3321	8.069381	44.38953	11.926211	1192.52771	93.29565	253.2061
2512	7.58462	24.06036	5.735005	608.167358	90.67004	129.7835
3519	6.348304	42.95666	11.35731	832.096558	91.35573	245.3542
3819	7.367481	30.82406	7.195733	619.454041	89.11017	181.4567
2918	4.003566	22.17302	4.716711	590.361816	87.21174	153.9188

DMC	**DSR**
172.34491	22.40038
88.498955	7.576586
194.28709	21.13649
143.1414	11.74637
114.16608	6.556601

Health Index) foram obtidos com dados do satélite MODIS, disponíveis num repositório de ficheiros .tif mantido pela NOAA (*National Oceanic and Atmospheric Administration*). O repositório contém um ficheiro .tif para cada índice e para cada semana (algumas semanas não estão disponíveis) com dados do planeta inteiro.

Devido ao tamanho do repositório (177 GB) foi necessário ter em conta a forma como as informações necessárias foram obtidas. Os incêndios foram agrupados e iterados por semana, para cada semana foram colocados todos os incêndios num *shapefile*. Seguidamente, foram baixados os .tif dos índices referentes a essa semana. Os ficheiros .tif dos índices foram sobrepostos com o *shapefile* dos incêndios e a cada incêndio foi dado um valor de índice igual à média simples da sobreposição. Os ficheiros foram em seguida apagados e repetido o processo para todas as semanas.

Os 4 índices de vegetação foram então concatenados com os parâmetros anteriores. Dos 2949 incêndios, 70 pertencem semanas sem dados no repositório. Estes incêndios foram removidos, sendo obtido um conjunto de dados com 2879 incêndios. Finalmente foram removidos os parâmetros referentes a delimitação do incêndio e a sua sobreposição com a malha geográfica mantendo-se os seguintes 18 parâmetros:

- **date**, data de inicio do incêndio
- **area_ha**, área ardida em hectares

- **source**, fonte dos dados (MARYLAND ou ICNF)
- **CHI**, índice de Haines contínuo
- **FWI**, índice canadiano de risco de incêndio florestal
- **ISI**, índice de propagação inicial do fogo
- **DC**, índice de seca
- **FFMC**, índice de humidade de combustível fino
- **BUI**, índice de combustível disponível
- **DMC**, índice de humidade de manta morta
- **DSR**, índice de severidade diário
- **PERIG**, índice de perigosidade
- **SMN**, índice de vegetação de diferença normalizada, suavizado
- **SMT**, *Smoothed Brightness Temperature*
- **TCI**, *Temperature Condition Index*
- **VCI**, *Vegetation Condition Index*
- **VHI**, índice de sáude de vegetação
- **region**, região

5.5 Agrupamento dos incêndios por região

Com os parâmetros obtidos, foram finalmente selecionados os incêndios com mais de 100 hectares e divididos pelas regiões estabelecidas. Para cada incêndio foi atribuído um parâmetro "região" com base no mapeamento do id da malha geográfica com as regiões "Norte Litoral", "Norte Interior", "Centro Interior", "Centro Litoral" e "Sul", ilustrada na Figura 5.3. Um excerto da tabela é apresentado na Tabela 5.2.

Figura 5.3: Esquema das regiões estabelecidas

Tabela 5.2: Excerto da tabela (1946 x 18) dos dados extraídos para incêndios com mais de 100 hectares de área ardida.

date	**area_ha**	**source**	**CHI**	**FWI**	**ISI**
04/08/2003	161.313	MARYLAND	5.23046	38.0751	11.0316
13/09/2019	199.336	ICNF	5.5405	33.971	10.8997
27/08/2001	545.362	MARYLAND	3.24283	15.4542	2.80621
09/11/2008	262.063	MARYLAND	7.07845	4.85101	1.93486
13/09/2019	116.97	ICNF	6.17178	35.1694	11.901

DC	**FFMC**	**BUI**	**DMC**	**DSR**	**PERIG**
568.59412	91.5696	136.347	97.3553	17.0728	4.67537
568.42798	92.0386	104.016	67.4321	13.9518	4.9746
816.75464	83.3996	166.331	111.566	3.46083	4.99406
783.27429	79.712	38.3772	20.4405	0.44513	1.07255
468.49756	91.9837	98.8205	67.1028	14.8348	4.98636

SMN	**SMT**	**TCI**	**VCI**	**VHI**	**region**
0.317738	305.389	36.7952	75.7136	56.2544	norte_interior
0.453399	295.94	70.5995	98.4722	84.536	norte_interior
0.372076	299.193	61.0607	61.1168	61.0886	centro_interior
0.211403	293.836	15.3067	37.9396	26.6185	sul
0.447	297.7	43.9	92.86	68.38	norte_litoral

Tabela 5.3: Número de incêndios com mais de 100 hectares por região.

	2001-2018	2019-2020	
Norte Interior	260	19	279
Norte Litoral	565	28	593
Centro Interior	619	36	655
Centro Litoral	183	11	194
Sul	209	16	225
Portugal Continental	1836	110	1946

5.6 Conjuntos de dados construídos e pré-processamento

Numa fase inicial do projeto foi testada a abordagem com todos os parâmetros extraídos. Os resultados obtidos na fase de agrupamento não supervisionado não foram satisfatórios tendo sido decidido não usar os índices de vegetação nem o índice de perigosidade. O índice DSR também não será utilizado por ser um valor obtido em função do FWI. Foram então tido em conta duas combinações de índices, **some** com 5 parâmetros (FWI, CHI, ISI, DC, FFMC) e **all**, com 7 (FWI, CHI, ISI, DC, FFMC, BUI e DMC). Os dados foram agrupados em 6 regiões: Portugal Continental (PT), Norte Litoral (NL), Norte Interior (NI), Centro Litoral (CL), Centro Interior (CI) e Sul (S). Os dados foram ainda divididos de acordo com o ano de ocorrência do incêndio, os incêndios ocorridos nos anos

2001 a 2018 foram usados no treino dos modelos e os incêndios ocorridos nos anos 2019 a 2020 foram utilizados para teste. Foram obtidos 24 conjuntos de dados com a combinação das regiões, dos conjuntos de índices e dos dois intervalos de tempo, descritos na Tabela 5.4. A Tabela 5.5 apresenta o número de exemplos por região utilizados para treino e para teste.

Tabela 5.4: Tabela com os conjuntos de dados estudados. Séries de parâmetros: **some** com 5 parâmetros (FWI, CHI, ISI, DC, FFMC) e **all**, com 7 (FWI, CHI, ISI, DC, FFMC, BUI e DMC).

Região	**Atributos**	**Anos**	**Conjunto de dados**	**Dimensões**
Portugal Continental (PT)	some	01-18	PT some	1836 x 5
		19-20	PT some 1920	110 x 5
	all	01-18	PT all	1836 x 7
		19-20	PT all 1920	110 x 7
Norte Interior (NI)	some	01-18	NI some	260 x 5
		19-20	NI some 1920	19 x 5
	all	01-18	NI all	260 x 7
		19-20	NI all 1920	19 x 7
Norte Litoral (NL)	some	01-18	NL some	565 x 5
		19-20	NL some 1920	28 x 5
	all	01-18	NL all	565 x 7
		19-20	NL all 1920	28 x 7
Centro Interior (CI)	some	01-18	CI some	619 x 5
		19-20	CI some 1920	36 x 5
	all	01-18	CI all	619 x 7
		19-20	CI all 1920	36 x 7
Centro Litoral (CL)	some	01-18	CL some	183 x 5
		19-20	CL some 1920	11 x 5
	all	01-18	CL all	183 x 7
		19-20	CL all 1920	11 x 7
Sul (S)	some	01-18	S some	209 x 5
		19-20	S some 1920	16 x 5
	all	01-18	S all	209 x 7
		19-20	S all 1920	16 x 5

5.7 Sumário

Os dados usados neste estudo experimental são constituídos pelos incêndios com área ardida superior a 100 hectares e com data de inicio no período de 2001 a 2020. Para cada um destes incêndios são utilizados os seguintes atributos: índice canadiano de risco de incêndio florestal (FWI), índice de Haines contínuo (CHI), índice de propagação inicial do fogo (ISI), índice de seca (DC), índice de humidade de combustível fino (FFMC), índice de combustível disponível (BUI) e índice de humidade de manta morta (DMC). Estes índices

Tabela 5.5: Número de incêndios no conjunto de dados de treino e no conjunto de dados de teste por região.

Região	**nº de exemplos**	
	treino	**teste**
Portugal Continental (PT)	1836	110
Norte Interior (NI)	260	19
Norte Litoral (NL)	565	28
Centro Interior (CI)	619	36
Centro Litoral (CL)	183	11
Sul (S)	209	16

serão estudados em duas séries distintas, **some** com 5 parâmetros (FWI, CHI, ISI, DC, FFMC) e **all**, com 7 (FWI, CHI, ISI, DC, FFMC, BUI e DMC).

Durante uma fase inicial do estudo, foram obtidos e testados índices de vegetação e de perigosidade. Os resultados obtidos com estes índices não foram satisfatórios não comparecendo por isso no conjunto de dados final.

Devido a Portugal Continental possuir regiões com diferentes tipos de climas, topografias, tipos de vegetação e condições meteorológicas os dados foram divididos em 5 regiões: Norte Litoral (NL), Norte Interior (NI), Centro Litoral (CL), Centro Interior (CI) e Sul (S). Os dados serão estudados por região e, com a sua concatenação, para Portugal Continental (PT).

Os dados foram divididos em dados de treino, compostos pelos incêndios de 2001 a 2018 e dados de teste com os incêndios de 2019 a 2020.

6

Protocolo experimental desenvolvido

6.1 Plano geral

A abordagem escolhida é constituída por três estágios, apresentados na Figura 6.1 e descritos nesta secção. O protocolo será aplicado individualmente aos dados das seis regiões definidas (PT, NL, NI, CL, CI e S). Os conjuntos de dados são compostos por 5 (**some**) ou 7 (**all**) índices de risco de incêndio.

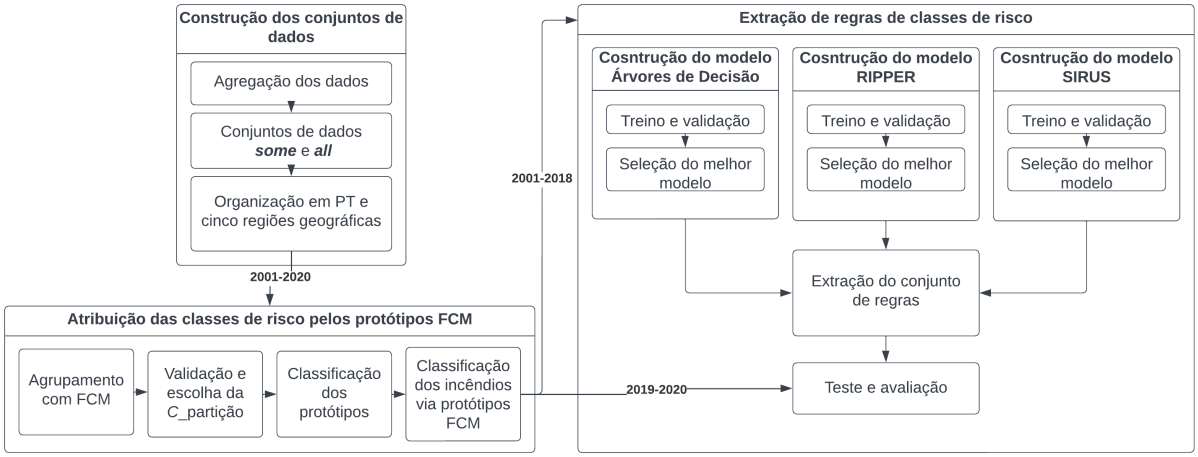


Figura 6.1: Esquema da abordagem desenvolvida.

- Estágio 1 - organização de dados de incêndios na forma tabular

  1. Construção dos conjuntos de dados como descritos no ponto 5.6.
  2. Os conjuntos de dados foram construidos e organizados na forma tabular com dois propósitos:
     a) construção dos modelos de *clustering* e classificação
     b) avaliação de teste dos classificadores

3. Sumarizar os dados e designá-los com "2001-2018" e "2019-2020" de acordo com os aspetos a) e b). Tendo como objetivo a construção com os dados "2001-2018" e o teste dos mesmos "2019-2020'.
4. Construir duas coleções, Tabela 5.4, some/all com o objetivo de escolher e avaliar a abordagem com o CHI, FWI e alguns sub-índices (some) ou CHI, FWI e todos os sub-índices (all).
5. Normalização por amplitude dos conjuntos de dados.

- Estágio 2 - agrupamento dos incêndios e atribuição das classes de risco

1. Agrupamento dos incêndios via FCM
   - Este algoritmo foi escolhido devido à incerteza nas fronteiras das escalas de severidade de risco de incêndio.
   - Hiperparametrização do FCM, é necessário escolher o número de grupos
2. Validação e seleção da partição difusa do FCM
   - Correr o algortimo com nº de grupos de c_min a c_max.
   - Utilizar o índice de Xie e Beni para escolher o melhor número de grupos pelo *elbow method*.
   - Obter a C-partição difusa
3. Atribuição de classes de perigosidade aos incêndios
   - Classificação dos protótipos da fuzzy C-partição através da escala de risco
   - Utilizar os protótipos etiquetados para a classificação dos incêndios, adicionando aos dados um novo atributo com a severidade
4. Sumarização de erro de classificação de risco de incêndio
   - O erro da classificação não supervisionada utilizado é definido como o número de incêndios classificados com um nível de risco inferior com a escala obtida com o FCM ao nível de risco obtido com a escala de referência.
   - O resultado de 3. e 4. é a C-partição com os parâmetros some e a C-partição com os parâmetros all.
   - Escolher a melhor C-partição com o apoio do erro e dos protótipos.
5. Visualização da C-partição com *fuzzy sammon map*.

- Estágio 3 - definição e avaliação de regras de regras de classes de risco de incêndio

1. São comparados três classificadores *rule based* com extração de regras: árvores de decisão, RIPPER e SIRUS.
2. Construção, validação e seleção do melhor modelo para cada um dos classificadores

3. Seleção do melhor modelo de entre os diferentes classificadores com o auxilio das métricas de pontuação de conjuntos de regras
4. Análise do conjunto de regras e comparação com à escala de referência
5. Avaliação do modelo com dados de teste

6.2 Organização dos dados de incêndios na forma tabular

Como resultado da extração e tratamento dos dados descrita na secção 5 obtiveram-se 24 conjuntos de dados derivados da combinação das seis regiões, dois conjuntos de parâmetros (some, all) e dois intervalos de tempo ("2001-2018", "2019-2020"'). Para cada um dos 12 conjuntos de dados referentes ao incêndios de 2001-2018 foi efetuada a normalização por amplitude, de acordo com a fórmula:

$$y_i = \frac{x_i - \mu(X_i)}{\max(X_i) - \min(X_i)}$$

,onde X_i é o conjunto de valores de um atributo i, x_i um elemento deste conjunto e $\mu(X_i)$ é média do conjunto. Os restantes 12 conjuntos de dados "2019-2020", foram normalizados com a mesma fórmula mas com os valores de $max(X_i)$, $min(X_i)$ e $\mu(X_i)$ do conjunto de dados "2001-2018" da região e parâmetros correspondente.

6.3 Agrupamento dos incêndios e atribuição das classes de risco

6.3.1 Agrupamento dos incêndios via FCM

Pelo facto de os dados não serem etiquetados foi necessário utilizar um algoritmo de agrupamento não supervisionado. Esta classe de algoritmos permite agrupar os dados em partições apenas tendo em conta a "similaridade" entre os dados.

Tendo em conta as escalas existentes de risco de incêndio, observa-se que os valores limiares entre níveis comportam alguma incerteza por serem obtidos através de um estudo empírico. Por esse motivo optou-se por utilizar um algoritmo de partição difusa (*soft* ou *fuzzy clustering*). Estes algoritmos atribuem a cada ponto um nível de pertença a cada partição, em comparação, os algoritmos de *hard clustering*, como o *k-means*, classificam cada ponto em apenas uma partição, criando fronteiras entre partições definidas com valores fixos. A partição difusa permite uma análise preservando a incerteza dos valores limiares.

O algoritmo *fuzzy c-means* (FCM) foi escolhido por ser um dos algoritmos de partição difusa não supervisionada mais amplamente utilizado e por se terem obtidos resultados favoráveis com este algoritmo num estudo realizado em dados deste mesmo domínio.

6.3.2 Validação e seleção da partição difusa do FCM

Os algoritmos de aprendizagem automática, no geral, contêm parâmetros que controlam o comportamento do mesmo, denominados de hiperparâmetros. O algoritmo *fuzzy*

c-means (FCM), tem alguns hiperparâmetros relacionados com a sua natureza iterativa: erro de paragem, número máximo de iterações e matriz de pertenças inicial. Pelo facto de o algoritmo ser corrido várias vezes, é esperado que estes valores, se definidos de forma razoável, não tenham impacto no resultado final.

O principal hiperparâmetro tido em conta, por ter maior impacto no resultado, foi o número de grupos. Durante o estudo experimental foram tidos em conta 3 escalas de previsão distintas de risco de incêndio, com 5,6 e 7 níveis de risco. Pelo facto de a escala de risco utilizada com menor número de níveis conter 5 níveis foi definido um número mínimo de partições, c_min de 4 e, de forma arbitrária, foi definido um máximo, cmax, de 10. Os restantes hiperparâmetros foram definidos de forma padrão: erro mínimo de critério de paragem de 0.001, número máximo de iterações igual a 1000 e grau de *fuzzyness*, m, igual a 2.0.

Tendo como objetivo selecionar a melhor C-partição é necessário comparar os agrupamentos derivados de diferentes números de grupos. De forma a comparar a qualidade de classificação de diferentes agrupamentos, utilizam-se índices de validação. Um índice associa a qualidade do agrupamento a um valor numérico, permitindo a comparar múltiplas classificações.

Por se tratar de um problema de classificação não supervisionada e não existirem dados etiquetados não é possível calcular índices de validação externos. Os índices de validação internos utilizam apenas as propriedades intrínsecas aos dados, como a o nível de separação entre grupos, o nível de semelhança entre pontos do mesmo grupo e entre pontos de grupos distintos. Uma partição é melhor quanto mais coesos os elementos de um mesmo grupo e mais distintos elementos de grupos diferentes.

O índice de validação utilizado neste estudo experimental foi o índice de Xie e Beni [1, 49], índice este extensivamente usado para validação interna do FCM. Este índice quantifica a razão entre a variação total nos *clusters* com a separação dos mesmos. O valor do índice de Xie e Beni tende a decrescer com o aumento do número de *clusters*, por esse motivo, este índice não deve ser utilizado isoladamente como indicador do número de *clusters* ótimo, sendo necessário contrapor com outros índices. No entanto neste estudo apenas é utilizado um único índice de validação interna (Xie e Beni), esta incorreção é amenizada pelo uso de métodos de validação externa (comparação com escala de risco existentes).

Pelo facto de este índice não ter limite superior definido, e pela necessidade de comparar índices obtidos com diferentes conjuntos de dados será aplicado o procedimento proposto em [35] que consiste em normalizar os valores dos índices.

De fora a obter a melhor C-partição será corrido o FCM e calculado o índice de Xie-Beni com de c_min a c_max grupos. Os valores deste índice para as serão então normalizados por amplitude e, com a representação dos valores em um gráfico de linhas, utilizado o *elbow method* para escolher a melhor C-partição.

6.3.3 Atribuição de classes de perigosidade aos incêndios

Para permitir a classificação dos incêndios é necessário atribuir aos protótipos obtidos com o FCM uma classe de risco. Esta atribuição foi feita classificando cada um dos índices dos protótipos de acordo com a escala de severidade de risco da Tabela 2.1, escala IPMA para o índice FWI e escala EEFIS para os restantes índices, o índice CHI foi considerado mas não classificado. A cada protótipo foi atribuída uma classificação em concordância com a classificação dos seus índices. Um exemplo da aplicação desta metodologia pode ser observada na Tabela 6.1 com us protótipos de uma partição com dados de Portugal Continental. Esta metodologia permite comparar os protótipos com a escala existente e aferir se são ou não fiáveis. Os incêndios serão então etiquetados de acordo com a classificação do protótipo do *cluster* ao qual têm maior pertença.

Finalizado este estágio, os dados dos incêndios terão um parâmetro com a severidade, designado de "atributo classe", acrescido aos parâmetros **some** ou **all**. Este parâmetro de severidade será utilizado, no 3º estágio, como a variável dependente nos modelos de extração de regras.

Tabela 6.1: Classificação dos protótipos do FCM com c=7 e parâmetros **all** com dados de Portugal Continental. Inicialmente os índices de risco dos protótipos são classificados com a escala existente (IPMA e EEFIS). Seguidamente com a classificação dos vários índices, e tendo em conta o valor do CHI, é atribuído a cada protótipo uma classificação.

Classificação	FWI	FFMC	DMC	DC	ISI	BUI	CHI
???	6.69	78.46	26.49	183.38	3.46	34.58	4.34
???	19.89	85.69	81.98	554.42	5.24	114.81	3.16
???	25.06	89.59	79.49	537.86	7.24	111.58	7.10
???	31.96	89.93	137.87	681.98	8.23	177.38	5.09
???	34.14	91.69	120.03	604.32	9.38	155.82	8.35
???	39.28	92.19	225.56	801.52	10.29	259.54	7.34
???	48.54	93.45	154.78	642.89	15.01	188.43	8.85

Risco
Muito baixo
Baixo
Moderado
Alto
Muito alto
Extremo

Classificação	FWI	FFMC	DMC	DC	ISI	BUI	CHI
Baixo	6.69	78.46	26.49	183.38	3.46	34.58	4.34
Moderado	19.89	85.69	81.98	554.42	5.24	114.81	3.16
Alto	25.06	89.59	79.49	537.86	7.24	111.58	7.10
Alto	31.96	89.93	137.87	681.98	8.23	177.38	5.09
Alto	34.14	91.69	120.03	604.32	9.38	155.82	8.35
Muito alto	39.28	92.19	225.56	801.52	10.29	259.54	7.34
Extremo	48.54	93.45	154.78	642.89	15.01	188.43	8.85

6.3.4 Sumarização de erro de classificação de risco de incêndio

As partições selecionadas foram em seguida comparadas com o auxílio de uma matriz de confusão entre a classificação com a C-partição e a classificação com a escala de risco existente. Sendo este um problema associado a um risco, uma classificação pessimista

é valorizada face a uma classificação otimista. Por esse motivo, na análise da matriz de confusão foi analisada principalmente o número de incêndios classificados com risco menor com a partição do que com a escala original, sendo dado preferência à partição com menor número de casos. Para tal, foram construídas matrizes de confusão com ambas as classificações, Tabela 6.2

Tabela 6.2: Matriz de confusão do erro de classificação para os protótipos da partição com c=7 e parâmetros **all** com dados de Portugal Continental.

	Baixo	Moderado	Alto	VeryAlto	Extremo	
VeryBaixo	148	43	3	0	0	194 (10.0%)
Baixo	50	58	28	2	0	138 (7.1%)
Moderado	**3**	108	157	10	0	278 (14.3%)
Alto	0	**54**	656	105	0	815 (41.9%)
VeryAlto	0	0	**115**	129	135	379 (19.5%)
Extremo	0	0	**2**	**23**	97	122 (6.3%)
Maximum	0	0	0	0	**20**	20 (1.0%)
	201 (10.3%)	263 (13.5%)	961 (49.4%)	269 (13.8%)	252 (12.9%)	1946
Erro	3	54	117	23	20	**217 (11.2%)**

Com um intuito de apresentar e comparar várias partições serão utilizadas tabelas apenas com o valor do erro (última linha da Tabela 6.2). No entanto as matrizes de todas as partições estudadas foram analisadas, sendo apenas referidas quando se observou resultados atípicos.

6.3.5 Visualização das partições

Outra método de validação utilizado foi a visualização das partições. Pelo conjunto de dados ter um número de atributos superior a três, foi necessário converter os dados do agrupamento difuso para 2 dimensões. Foi, para isso, utilizado o método de *fuzzy sammon mapping*, descrito na secção 4.3.

6.3.6 Agrupamento e atribuição das classes de risco para os incêndios de Portugal Continetal

Com o intuito de exemplificar o procedimento descrito, nesta secção serão apresentados os resultados da aplicação da metodologia para os dados de Portugal Continental.

	Clusters	4	5	6	7	8	9	10
XB	some	241.82	266.54	140.51	118.06	121.80	129.49	91.45
	all	327.22	189.63	243.24	152.09	159.02	139.02	153.11

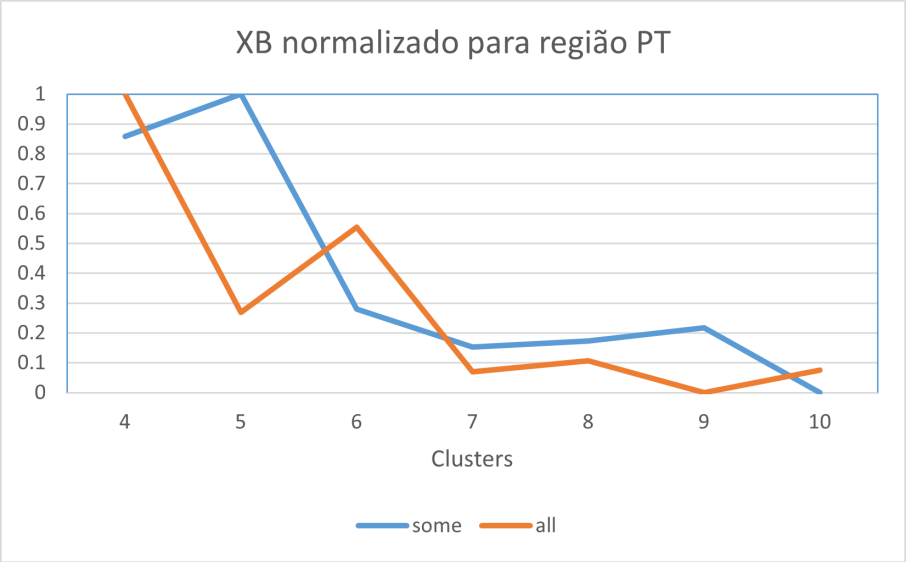


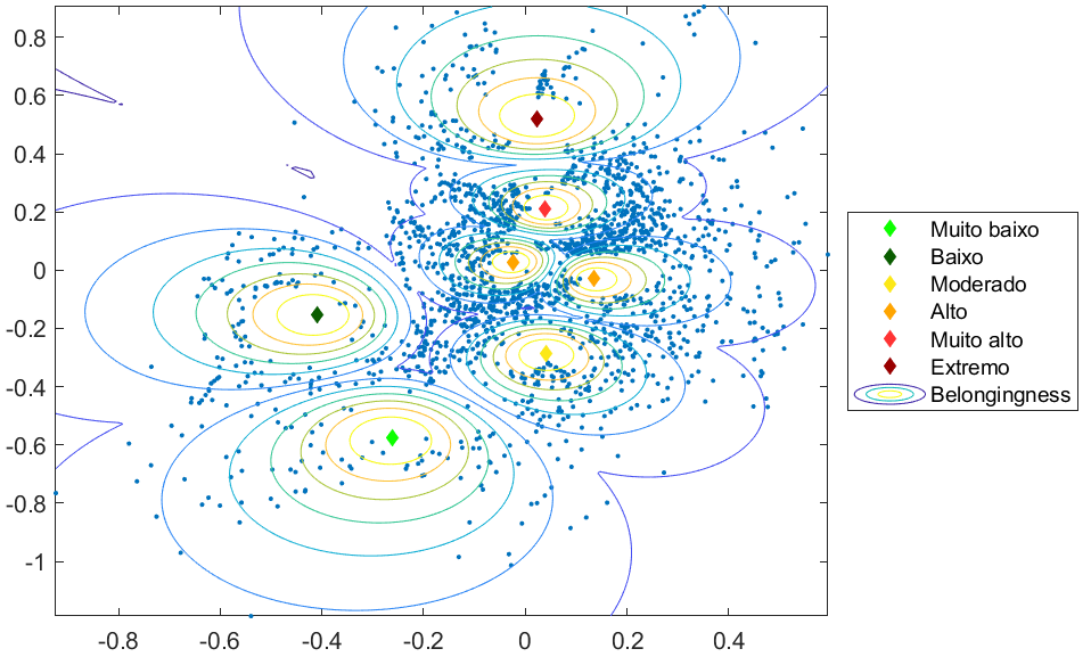
Figura 6.2: Índice de Xie-Beni para Portugal Continental com 5 parâmetros (some) e 7 parâmetros (all).

Tabela 6.3: Portugal Continental. Erro por partições, obtido com a comparação da classificação obtida com o FCM com a classificação com a escala IPMA do índice FWI.
Erro definido como o número de incêndios classificados com um nível de risco inferior ao nível de risco obtido com a escala de referência.

	Muito baixo	Baixo	Moderado	Alto	Muito alto	Extremo	Erro
someC6	—	1	112	261	—	20	394 (20.2%)
allC6	—	4	156	64	—	142	366 (18.8%)
someC7	22	23	136	106	0	20	307 (15.8%)
allC7	—	3	54	117	23	20	217 (11.2%)
someC8	9	16	98	107	6	20	256 (13.2%)
allC8	—	3	29	56	19	20	127 (6.5%)

Classificação	FWI	FFMC	DC	ISI	CHI
Muito baixo	5.12	63.08	567.10	1.43	2.59
Baixo	9.35	84.93	134.03	4.81	5.90
Moderado	24.12	88.36	585.44	6.14	3.19
Alto	27.68	90.63	578.87	7.46	7.68
Alto	34.55	91.04	717.09	9.10	5.45
Muito alto	37.81	92.64	662.37	10.35	8.84
Extremo	52.43	93.93	654.46	16.81	9.08

(a) Protótipos da partição e comparação com escala de referência, escala IPMA para o índice FWI e escala EEFIS para os restantes índices (exceto CHI). Índice a verde indica uma classificação semelhante à classificação do protótipo (coluna "Classificação"). Amarelo, uma classificação a um nível de risco de diferença. Vermelho, mais de um nível de risco de diferença.

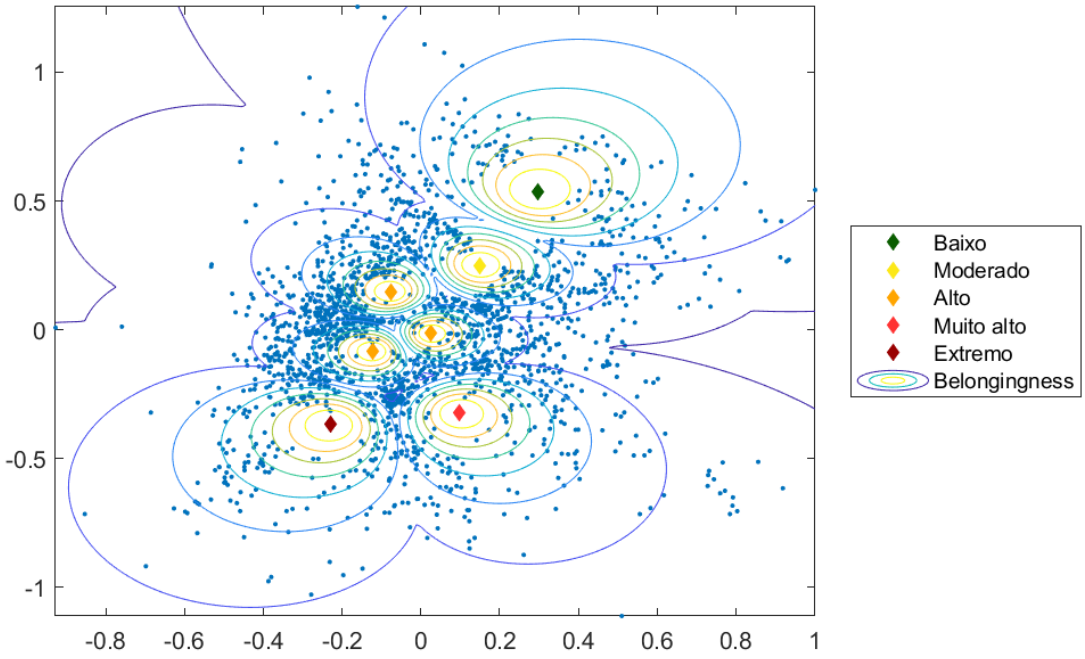


(b) Visualização das partições com o algoritmo de FUZZSAM e com redução para duas dimensões com PCA. O centroide de cada partição é representado com um losango com a cor correspondente à classe de risco. Os pontos a azul representam os incêndios e as isolinhas indicam o grau de pertença (belongingness) a cada partição.

Figura 6.3: FCM com C = 7 com os parâmetros **some** para a região Portugal Continental.

Classificação	FWI	FFMC	DMC	DC	ISI	BUI	CHI
Muito baixo / Baixo	6.69	78.46	26.49	183.38	3.46	34.58	4.34
Moderado	19.89	85.69	81.98	554.42	5.24	114.81	3.16
Alto	25.06	89.59	79.49	537.86	7.24	111.58	7.10
Alto	31.96	89.93	137.87	681.98	8.23	177.38	5.09
Alto	34.14	91.69	120.03	604.32	9.38	155.82	8.35
Muito alto	39.28	92.19	225.56	801.52	10.29	259.54	7.34
Extremo	48.54	93.45	154.78	642.89	15.01	188.43	8.85

(a) Protótipos da partição e comparação com escala de referência, escala IPMA para o índice FWI e escala EEFIS para os restantes índices (exceto CHI). Índice a verde indica uma classificação semelhante à classificação do protótipo (coluna "Classificação"). Amarelo, uma classificação a um nível de risco de diferença. Vermelho, mais de um nível de risco de diferença.



(b) Visualização das partições com o algoritmo de FUZZSAM e com redução para duas dimensões com PCA. O centroide de cada partição é representado com um losango com a cor correspondente à classe de risco. Os pontos a azul representam os incêndios e as isolinhas indicam o grau de pertença (belongingness) a cada partição.

Figura 6.4: FCM com C = 7 com os parâmetros **all** para a região Portugal Contiental

Observando o gráfico do índice de Xie e Beni, Figura 6.2, o melhor número de centroides, de acordo com este índice, é 6 ou 7, para o conjunto de parâmetros **some** e 5 ou 7 para o conjunto de parâmetros **all**. Pelo facto de com 5 centroides com o conjunto de parâmetros **all** o algoritmo FCM não resultar numa classificação com o nível de risco "Muito alto" este valor será ignorado. Pelo mesmo motivo será ignorada a partição com 6 centroides para o conjunto de parâmetros **some**. As partições escolhidas com o índice de Xie e Beni para Portugal Continental são então **someC7** e **allC7**.

Na tabela com o erro das partições, Tabela 6.3, é possível observar que a partição

someC7 resulta em um erro de 15.8% com 136 incêndios com classificação "Moderado", subestimados em relação à escala de referência. A partição **allC7** resulta num erro inferior, 11.2% e com uma classificação mais correta para o nível de risco "Moderado", apenas com 54 incêndios. Ambas as partições apresentam valores não satisfatórios na classificação de risco "Alto", de acordo com esta métrica de erro.

Observando os protótipos das partições **someC7** e **allC7**, Figura 6.3 e Figura 6.4 respetivamente, é possível observar que o centroide com classificação "Muito baixo" não parece correto, tendo um valor de DC no intervalo de "Alto" na escala de referência. A partição **allC7** não resulta em nenhum protótipo com classificação "Muito baixo", mas pelo estudo se focar em incêndios com mais de 100 hectares de área ardida, isso não é necessariamente um ponto negativo. Os protótipos da partição **allC7** aparentam ser mais corretos e não aparenta existir nenhum problema observável na visualização das partições. Com base nos critérios apresentados para a região Portugal Continental foi escolhido a partição **allC7**.

6.4 Metodologia de classificação

6.4.1 Árvores de decisão

A principal restrição aos modelos neste estudo experimental foi a necessidade de extrair regras para classificação. A escolha pelas árvores de decisão foi feita por este modelo resultar num conjunto de regras condicionais. As árvores de decisão são também um método de classificação amplamente utilizado e que obteve resultados favoráveis num estudo realizado em dados deste mesmo domínio, [8]. Os resultados das árvores de decisão são comparados com os resultados com o modelo floresta aleatória de forma a, comparativamente, verificar a sua robustez.

Os hiperparâmetros da árvore de decisão tidos em consideração foram os seguintes:

- **Critério de divisão**. Este parâmetro refere-se à função que mede a qualidade da divisão de cada nó. Durante a criação da árvore, em cada nó, é escolhida uma divisão, tendo em conta todos os parâmetros e todas divisões possíveis, que maximiza esta função. As funções testadas foram impureza de Gini e entropia.

- **Profundidade máxima**. Este parâmetro restringe o tamanho máximo de um ramo de uma árvore. Uma árvore com pouca profundidade é fácil de ser lida, mas não têm capacidade de classificar os dados corretamente. Uma árvore com maior profundidade tende a conseguir classificar melhor, mas dando origem a uma árvore mais complexa. Uma árvore muito profunda pode também se sobreajustar aos dados.

- **Número mínimo de dados por divisão**. Durante a criação da árvore se o número de dados em um nó for igual ou inferior a este valor, não é feita a divisão e o nó finaliza como um nó folha. Este valor limita a complexidade da árvore.

- **Número mínimo de dados por folha**. Este valor restringe o número de dados em um nó folha. Durante o processo de divisão de um nó, apenas são consideradas divisões que resultem em dois nós com um número de dados em cada nó maior que este valor. Tal como o valor anterior, este valor limita a complexidade da árvore.

De forma a obter a melhor combinação de hiperparâmetros foi realizado um processo de validação. Durante a validação os dados são divididos em dois conjuntos, um conjunto de treino e um de validação. Como os nomes indicam, o conjunto de treino será usado para treinar os modelos e o conjunto de validação será um conjunto "desconhecido" pelo modelo usado para o avaliar. Esta divisão é necessária para detetar situações onde o modelo se sobreajusta aos dados de treino. Este procedimento não pode ser efetuado apenas uma vez, escolher um modelo que melhor classifica o conjunto de validação poderia levar a um modelo que se sobreajuste a este conjunto. O método de validação usado é, portanto, denominado de validação cruzada, onde o procedimento descrito é efetuado para diferentes divisões dos dados. Podem-se especificar dois tipos de validação cruzada, a exaustiva e a não exaustiva.

Na validação cruzada exaustiva são tidas em consideração todas as possíveis divisões entre conjunto de treino e de validação. Este procedimento exige bastante tempo computacional e, com um conjunto de dados na casa das centenas, como é o caso deste estudo, não têm vantagem em relação à versão não exaustiva.

Será por isso aplicada a técnica de validação cruzada não exaustiva *stratified shufflesplit*. Neste método o processo de validação é efetuado um número pré-definido de vezes para conjuntos de treino e validação obtidos através de uma seleção aleatória dos dados originais. O resultado de um modelo será então a média do resultado obtido para cada divisão. O facto de a seleção ser aleatória leva a que exista sobreposição e omissão de dados usados nas diferentes validações, mas, pelo tamanho do conjunto de dados, não é esperado que isto afeta os resultados obtidos. Uma técnica que não omite ou sobrepõem dados que poderia ser usado é o *stratified k-fold*.

O *stratified shufflesplit*, como o nome indica, faz uma divisão estratificada dos dados. A divisão estratificada resulta em conjuntos de treino e de validação com a mesma distribuição de dados por classe entre si e igual ao conjunto original.

Tabela 6.4: Valores testados para os hiperparâmetros do algoritmo de árvores de decisão

Hiperparâmetros	Possibilidades
Critério de ramificação	Gini ou entropia
Quantidade mínima de dados por nó-folha	1 a 20
Quantidade mínima de dados por ramificação	2 a 40

Realizando o processo acima descrito, com os hiperparâmetros apresentados na Tabela 6.4 foram geradas árvores com profundidade máxima de 2 a 10 com os melhores hiperparâmetros encontrados. A árvore escolhida foi a que maximiza o F1 e que tenha um número de nós e de nós-folha que permita uma leitura clara da árvore.

Todo o processo descrito, referente às árvores de decisão foi efetuado com o auxílio da biblioteca *Scikit-learn* em *python*. Esta biblioteca implementa as árvores de decisão com o algoritmo CART e implementa, de entre outros métodos de validação, o *stratified shufflesplit*. Para obter a representação gráfica da árvore foi utilizada a biblioteca *graphviz*.

6.4.2 RIPPER

Outro método estudado, que permite a extração de regras, foi o algoritmo RIPPER. Uma explicação deste algoritmo encontra-se na secção 3.1.3. De forma simplificada, este algoritmo consiste no seguinte processo iterativo:

1. Divisão dos dados em dois conjuntos, o *growing set* e o *pruning set*.
2. Treino de uma árvore com o *growing set*
3. Poda da árvore com o *pruning set*, e obtenção de um ramo, correspondente a uma regra condicional
4. Remoção dos dados cobertos pela regra

O resultado final é um conjunto de regras encadeadas no formato:

IF {c_1 e c_2 e, ..., c_k
ou
c'_1 e c'_2 e, ..., c'_k
ou
...
} **THEN** Previsão = p_1
ELSE
IF{ ... } **THEN** Previsão = p_2
ELSE
...

Estas regras têm a dificuldade, em relação às regras condicionais das árvores de decisão, de ao ler uma regra é necessário ter em conta a negação de todas as regras anteriores. No exemplo acima, a previsão p_2 seria feita se a conjunção de condições c'' for verdadeira e as conjunções das condições c e c' forem falsas.

Para este método foi utilizada a biblioteca *RWeka* em *R*, com os hiperparâmetros padrão do algoritmo RIPPER. Para integrar o código *R* em *python* foi utilizada a biblioteca *rpy2*.

6.4.3 Resultados obtidos com Árvores de decisão e RIPPER com incêndios de Portugal Continetal

Tabela 6.5: Portugal Continental (PT allC7). Melhores valores encontrados dos hiperparâmetros para árvores de decisão, métricas de validação e número de nós e de nós-folha. DT X, refere-se à árvore de profundidade máxima X. RF aos resultados da floresta aleatória (random forest). Os valores de min_leaf e min_split referem-se ao número mínimo de dados por folha e por ramificação, respetivamente.

Modelo	Critério	min_leaf	min_split	Nº nodes	Nº folhas	Precisão	Recall	F1
DT 2	Gini	1	2	7	4	0.58	0.68	0.60
DT 3	Gini	1	2	15	8	0.78	0.77	0.75
DT 4	Gini	1	19	29	15	0.80	0.80	0.77
DT 5	Gini	1	40	45	23	0.80	0.80	0.80
DT 6	Gini	3	18	71	36	0.85	0.85	0.84
DT 7	Entropia	3	4	133	67	0.81	0.82	0.81
DT 8	Gini	2	21	129	65	0.86	0.86	0.86
DT 9	Gini	1	24	145	73	0.87	0.87	0.87
RIPPER	—	—	—	—	—	0.84	0.84	0.84
RF	—	—	—	—	—	0.89	0.89	0.88

(a) Parâmetros **some**

Modelo	Critério	min_leaf	min_split	Nº nodes	Nº folhas	Precisão	Recall	F1
DT 2	Gini	1	2	7	4	0.61	0.70	0.62
DT 3	Gini	1	3	15	8	0.79	0.80	0.76
DT 4	Gini	11	2	29	15	0.87	0.86	0.86
DT 5	Gini	1	3	53	27	0.87	0.86	0.86
DT 6	Gini	3	2	73	37	0.89	0.89	0.89
DT 7	Gini	1	9	95	48	0.88	0.88	0.87
DT 8	Gini	3	9	113	57	0.90	0.90	0.90
DT 9	Gini	2	2	149	75	0.90	0.90	0.90
RIPPER	—	—	—	—	—	0.90	0.89	0.89
RF	—	—	—	—	—	0.92	0.91	0.91

(b) Parâmetros **all**

Como é possível observar na Tabela 6.5 para ambos os conjuntos de parâmetros a árvore com melhor F1 e com número de nós reduzido é a **DT 4**. Com os modelos **RIPPER**, obtêm-se, com ambos os conjuntos, melhores resultados que com as árvores de decisão. No geral, o conjunto de parâmetros **all** aparenta ser melhor nesta situação pelo facto dos modelos obterem melhores métricas e por essas métricas se aproximarem mais dos valores da **RF**.

6.4.4 SIRUS

Um método bastante usado em problemas de classificação é o algoritmo de florestas aleatórias. Neste método várias árvores são treinadas independentemente e, para problemas de classificação, o resultado é a classe escolhida por o maior número de árvores. O problema deste método neste estudo é não permitir a extração de regras. O algoritmo SIRUS é um algoritmo que permite gerar uma floresta aleatória e retirar um conjunto de regras da mesma.

6.4.4.1 Problemas encontrados com SIRUS

Numa fase inicial do estudo experimental observou-se um problema com os resultados do algoritmo para o conjunto de dados em estudo. Cada árvore da floresta é criada com o objetivo de melhor separar os dados e as regras extraídas vão ser, de forma simplificada, as regras com maior frequência. O problema identificado foi que, pelo conjunto de dados conter um número não balanceado de elementos de cada classe, as regras extraídas referiam-se quase na sua totalidade às classes com mais elementos.

Com o intuito de resolver este problema o algoritmo foi corrido para que no treino das árvores de decisão fosse utilizado uma versão balanceada dos dados. Para tal para cada processo de treino foi selecionado para cada classe um número N de elementos com N igual ao número de elementos da classe com menos elementos. Após o balanceamento dos dados observou-se outro problema das regras obtidas. Estas, na sua maioria, referiam-se aos níveis extremos da escala ("Muito baixo" e "Extremo"). Uma explicação possível para este comportamento é o facto de as classes formarem uma escala ordinal, o que faz com que a separação de um extremo dos restantes dados ser mais "fácil" sendo por isso o identificado em todas as árvores.

Para solucionar este problema e pelo facto de o algoritmo SIRUS apresentar resultados muito promissores para classificação binária [2], foram desenvolvidos protocolos baseados em classificação *one vs all*.

6.4.4.2 Classificação *one vs all*

A classificação *one vs all*, [39] ,é um método que permite realizar uma classificação multiclasse com base em um classificador binário. Este método de classificação consiste em gerar um classificador para cada classe que separe esta dos restantes e em seguida agrupar todos os classificadores para obter uma classificação final.

Concretamente, foram desenvolvidos dois protocolos experimentais para o algoritmo SIRUS baseados na classificação *one vs all*, denominados de **SIRUS else** e **SIRUS boost**.

6.4.4.3 SIRUS else

Este protocolo foi concebido com o objetivo de gerar um conjunto de regras encadeadas semelhantes às obtidas pelo algoritmo RIPPER. O método consiste nos seguintes

passos:

1. Para cada classe presente nos dados.
    a) Executar o algoritmo SIRUS
    b) Calcular, com o modelo SIRUS, para cada ponto a probabilidade de pertencer à classe
    c) Realizar uma regressão logística entre as probabilidades e o classe original
    d) Calcular o valor necessário da probabilidade do SIRUS para obter uma probabilidade de 0.5 no modelo de regressão logística
    e) Classificar, com a classe, os pontos com uma probabilidade SIRUS maior ou igual à encontrada no passo anterior
    f) Calcular a função *log-loss* para os dados classificados
2. Selecionar a classe, e o respetivo modelo SIRUS, que obtém melhor valor de *log-loss*
3. Remover os pontos classificados com a classe do passo anterior
4. Repetir até existir um modelo por classe

Com o processo descrito é obtido um conjunto de regras com o formato das regras na Tabela 6.6a. Por exemplo, com este conjunto de regras, dado um incêndio seriam testadas as primeiras quatro regras e obtido quatro valores de probabilidade. Esta probabilidade é obtida empiricamente, nos dados do Centro Litoral 5.8 % dos incêndios com FFMC 94.49 têm uma classificação de "Extremo". Em seguida é calculada a média das probabilidades obtidas, caso este valor seja maior ou igual a 0.334 (o valor que obtém uma probabilidade de 0.5 no modelo de regressão logística), o incêndio é classificado como "Extremo", caso contrário são lidas as regras para a classe "Muito alto".

Pelo facto de haver um número fixo de probabilidades possíveis de serem obtidas pelas regras é possível simplificar o conjunto de regras. Por exemplo para a média das probabilidades das regras associadas a classe "Extremo"na Tabela 6.6a ser maior que 0.334 é único e necessário que duas das regras sejam verdadeiras. Podendo agrupar-se as regras numa conjunção, como pode ser observado na Tabela 6.6b

Outra forma de simplificação é a efetuada com as regras associadas à classe "Moderado", na Tabela 6.6a. Para tal é utilizada a função **COUNT()**, função esta que interpreta um conjunto de proposições e devolve o número de verdadeiros.

Este formato permite não só a leitura das regras sem a necessidade de efetuar cálculos, mas também um entendimento imediato das consequências das mesmas. Neste caso, apenas são necessárias duas das três condições para ser classificado como "Moderado".

Tabela 6.6: Regras extraídas com o método Sirus else para região Centro Litoral

Risco	**Regras**
Extremo	Se probabilidade >= 0.334 FFMC < 94.494 then 0.057692 (n=104) else 0.66667 (n=12) FWI < 51.542 then 0.028846 (n=104) else 0.91667 (n=12) ISI < 15.956 then 0.028846 (n=104) else 0.91667 (n=12) Se não:
Muito alto	Se probabilidade >= 0.443 FFMC < 92.717 then 0.25352 (n=71) else 0.90323 (n=31) FWI < 37.099 then 0.13115 (n=61) else 0.92683 (n=41) ISI < 10.772 then 0.22535 (n=71) else 0.96774 (n=31) Se não:
Moderado	Se probabilidade >= 0.438 CHI < 3.3931 then 0.81818 (n=11) else 0.088889 (n=45) FFMC < 86.823 then 0.70588 (n=17) else 0.025641 (n=39) ISI < 3.1245 then 1 (n=6) else 0.14 (n=50) Se não:
Alto	

(a) Regras extraídas

Risco	**Regras**
Extremo	FWI >= 51.54 and ISI >= 16.96 Se não:
Muito alto	FWI >= 37.1 e ISI >= 10.77 Se não:
Moderado	**COUNT**(CHI < 3.39; FFMC < 86.82; ISI < 3.12) >= 2 Se não:
Alto	

(b) Regras simplificadas

COUNT(): interpreta um conjunto de proposições e devolve o número de verdadeiros.

6.4.4.4 SIRUS boost

Este protocolo consiste em gerar um modelo SIRUS para cada classe e juntar todos os modelos através do método de *boosting* [13]. Em aprendizagem automática *boosting* é um método de reunião (*ensemble method*) que permite combinar vários modelos "fracos"para obter um modelo mais eficaz. Começou-se por gerar um modelo SIRUS para cada classe

no formato *one vs all*. Para um determinado ponto, é obtido uma probabilidade de pertencer a cada classe, obtido por cada classificador. Uma opção possível de juntar os modelos seria escolher a classificação com maior probabilidade. Pelo facto de os valores de probabilidade derivados a partir do algoritmo SIRUS serem calculadas com base nos dados de treino, uma classe com maior número de casos teria uma maior probabilidade de ocorrer. Para o caso específico dos dados estudados, para quase todos os incêndios a probabilidade maior estava associada ao nível "Alto".

A decisão de adotar versões do método de *boosting* ao invés de usar métodos amplamente usados como *AdaBoost* [18], foi o facto de estes métodos não permitirem manter a interpretação das regras dos modelos individuais. Foram desenvolvidos três métodos diferentes para combinar as regras dos vários modelos SIRUS:

- **Sirus boost plus**. Neste método, aos valores de probabilidade obtidos com o SIRUS, são somados valores constantes consoante a classe do modelo. A classificação é feita escolhendo o maior valor de probabilidade somado com a constante. As constantes são obtidas num processo de validação que testa diferentes combinações de valores e opta pela combinação que minimiza o *log-loss*.

- **Sirus boost stand**. Partindo dos valores de probabilidade obtidos com o SIRUS, os valores de cada classe são normalizados por desvio padrão. Na classificação, é escolhida a classe cujo modelo obtém maior probabilidade após a normalização.

- **Sirus boost multinom**. Neste método as probabilidades obtidas com os modelos SIRUS são colocadas como *input* em um modelo de regressão logística multinomial [26]. Este modelo é uma variante da regressão logística para problemas multi-classe e o seu output é uma lista com a probabilidade de pertencer a cada classe. A classificação é feita com os valores de probabilidade da regressão logística multinomial.

As regras obtidas com o **Sirus boost plus** e com o **Sirus boost stand** podem ser simplificadas de forma semelhante, aplicando as operações às probabilidades das regras individuais, no caso **Sirus boost plus** a soma da constante e no caso do **Sirus boost stand** a normalização. A aplicação destas operações nos valores individuais de probabilidade das regras é semelhante a aplicação da operação no valor final obtido com a média das várias regras.

A classificação com os valores obtidos com o **Sirus boost multinom** é realizada multiplicando os valores obtidos com o SIRUS pelos valores da matriz obtida pela regressão logística multinomial, Tabela 6.7d. Este método, ainda que mantendo as regras obtidas com o SIRUS e permitindo realizar a análise da matriz da regressão logística, não permite uma leitura clara do classificador final, afastando-se do propósito deste estudo. Este método não será, por isso, tido em conta na escolha dos conjuntos de regras, servindo apenas como referência para os restantes métodos **Sirus boost**.

Tabela 6.7: Trecho das regras extraídas com o método Sirus boost para região Centro Litoral

Risco	**Regras**
Baixo	FWI < 18.779 then 0.529 else 0
	...
Moderado	CHI < 3.370 then 0.647 else 0.088
	...
Alto	FWI < 39.116 then 0.643 else 0.015
	...
Muito alto	FWI < 39.116 then 0.02 else 0.667
	...
Extremo	FFMC < 94.638 then 0.075 else 0.588
	...

(a) Regras por classe de risco

	Baixo	Moderado	Alto	Muito alto	Extremo
P + ->	0.2	0.3	0	0.1	0.3

(b) Valores das constantes obtidas pelo método Sirus boost plus.

	Baixo	Moderado	Alto	Muito alto	Extremo
Desvio padrão	0.133	0.146	0.186	0.201	0.126
Média	0.055	0.146	0.39	0.28	0.128

(c) Valores de desvio padrão e média obtidos pelo método Sirus boost stand

	X	Baixo	Moderado	Alto	Muito alto	Extremo
Baixo	-9.97476	124.303	-20.9779	-2.08265	-93.3823	-2.1506
Moderado	-9.39557	-24.1587	59.5004	29.4634	3.51143	-61.0135
Alto	-6.92145	-29.1502	42.8965	29.5818	-3.07695	-7.07945
Muito alto	-9.42927	-54.5865	46.8062	23.3489	14.8911	-16.1792

(d) Matriz da regressão logística multinomial obtida com o método Sirus boost multinom

6.4.5 Resultados obtidos com SIRUS com incêndios de Portugal Continetal

Tabela 6.8: Portugal Continental (PT allC7). Métricas de validação dos modelos baseados no algoritmo SIRUS.

Model	Precisão	Recall	F1
else 3	0.71	0.72	0.69
else 4	0.71	0.72	0.70
else 5	0.69	0.66	0.63
else 6	0.68	0.70	0.67
boost 3 plus	0.73	0.73	0.72
boost 3 stand	0.71	0.71	0.69
boost 4 plus	0.70	0.72	0.70
boost 4 stand	0.71	0.70	0.69
boost 5 plus	0.65	0.68	0.65
boost 5 stand	0.70	0.66	0.66
boost 6 plus	0.70	0.70	0.70
boost 6 stand	0.69	0.68	0.67
boost 3 multi	0.74	0.74	0.72
boost 4 multi	0.76	0.75	0.72
boost 5 multi	0.76	0.75	0.73
boost 6 multi	0.75	0.74	0.73
RF	0.89	0.89	0.88

(a) Parâmetros **some**

Model	Precisão	Recall	F1
else 3	0.80	0.80	0.80
else 4	0.83	0.83	0.82
else 5	0.83	0.82	0.81
else 6	0.79	0.79	0.79
boost 3 plus	0.79	0.79	0.78
boost 3 stand	0.80	0.78	0.78
boost 4 plus	0.81	0.80	0.80
boost 4 stand	0.80	0.78	0.78
boost 5 plus	0.80	0.79	0.79
boost 5 stand	0.81	0.79	0.79
boost 6 plus	0.82	0.82	0.81
boost 6 stand	0.81	0.80	0.80
boost 3 multi	0.86	0.85	0.85
boost 4 multi	0.85	0.84	0.83
boost 5 multi	0.84	0.83	0.82
boost 6 multi	0.85	0.84	0.84
RF	0.92	0.91	0.91

(b) Parâmetros **all**

Observando a Tabela 7.12 para ambos os conjuntos de parâmetros valores obtidos pelos modelos **SIRUS** são inferiores aos obtidos com **RF**. Os valores obtidos com os modelos **boost plus** e **boost stand** são inferiores mas próximos aos obtidos com **boost multi**, como seria esperado. Os modelos obtidos com os parâmetros **all** resultam em melhores métricas e métricas mais próximas dos valores obtidos com **RF**.

6.5 Índices de validação das regras

De forma a validar e comparar as regras obtidas com os vários modelos serão utilizadas três métricas de pontuação de conjuntos de regras:

- Pontuação de previsibilidade. Mede a precisão do conjunto de regras comparando o erro empírico obtido com o modelo em causa e um modelo *naive*.
- Pontuação de estabilidade. Quantifica a sensibilidade do modelo a ruído nos dados e avalia a robustez do modelo. Um modelo é estável se partindo de dois conjuntos de dados da mesma distribuição obter regras similares

- Pontuação de simplicidade. Mede a facilidade como uma previsão é verificada. Este índice é menor quanto maior o número de desigualdades nas regras, tomando o valor de 1 para o modelo com menor número.

Estas pontuações foram implementadas em *python* de acordo com os índices descritos na secção 4.4.

6.6 Sumário

A metodologia pode ser dividida em três fases: agrupamento dos dados para a classificação não supervisionada dos mesmos, extração de regras e validação das regras extraídas.

Na fase de agrupamento será usado o algoritmo não supervisionado de agrupamento difuso fuzzy c-means. Optou-se por este algoritmo pelo facto de as fronteiras entre as escalas de risco existentes conterem alguma incerteza. Este algoritmo será executado para 4 a 10 *clusters*, e escolhido a c-partição com melhor valor do índice de validação interno Xie e Beni. Para a c-partição encontrada os centroides foram classificados com a escala de risco de referência e os incêndios etiquetados de acordo com a classificação do centroide ao qual têm maior pertença. Finalmente, para cada região, a melhor c-partição com a série de atributos **some** e a melhor c-partição com a série de atributos **all** são comparadas por meio da visualização com *fuzzy Sammon maps* e da comparação com a escala de referência. Após esta fase cada região será representada por apenas uma C-partição.

Na fase de extração de regras serão extraídos conjuntos de regras de três métodos diferentes: árvores de decisão, RIPPER e SIRUS. Para as árvores de decisão, serão comparadas árvores com diferentes profundidades máximas de forma a escolher a árvore que obtém melhor valor de F1 com um número de nós reduzido. Quanto ao algoritmo SIRUS, por este ter apresentado problemas na classificação multiclasse foram desenvolvidas duas metodologias baseadas na classificação *one vs all*. O método ao qual foi dado o nome de **SIRUS else** onde é gerado um conjunto de regras encadeadas e o método **SIRUS boost** onde é obtido um conjunto de regras para cada classe de risco existente.

Finalmente, na fase de validação das regras, os melhores conjuntos de regras obtidos com os diferentes métodos são comparados com o auxílio das métricas de precisão, *recall* e F1, e com as métricas de interpretabilidade de regras (previsibilidade, estabilidade e simplicidade). Com a métricas será encontrado o melhor conjunto de regras para cada uma das 6 regiões estudas, os quais serão alvo de uma análise por especialista na área baseada na comparação dos índices obtidos com os limites desses índices nas escalas existentes.

7

Estudo experimental

7.1 Principais objetivos do estudo experimental

Os objetivos do trabalho experimental são os seguintes:

- Construção e pré-processamento de doze coleções de conjuntos de dados de incêndios (2001 - 2018) para análise comparativa da severidade dos incêndios a nível regional, considerando-se cinco regiões (Norte Litoral e Norte Interior, Centro Litoral e Centro Interior e Sul), versus em Portugal Continental bem como análise ortogonal da caracterização dos incêndios com duas séries de índices de severidade de risco de incêndio, **some** com 5 parâmetros (CHI, FWI e principais subíndices) e **all**, com 7 (CHI, FWI e todos os subíndices);
- Aplicação do *fuzzy c-means* (FCM) para o agrupamento não supervisionado de incêndios (i.e. sem recurso a classificação de incêndios) com estudo do número de *clusters* das partições avaliados por índices internos e visualização por *Fuzzy Sammon Mappings*. Classificação dos centroides dos agrupamentos obtidos com o FWI de acordo com escala existente (IPMA e EEFIS). Classificação dos pontos com a classificação do centroide do agrupamento com maior pertença.
- Mostrar que o FCM é um método não supervisionado com resultados positivos para a classificação não supervisionada neste domínio.
- Análise das árvores de classificação geradas para diferentes valores de profundidade e comparação com valores obtidos com floresta aleatória. Extração das regras e análise quantitativa e por especialista.
- Treino, validação e testes de classificadores com indução de regras:
  - árvores de decisão (DT) com comparação de desempenho de classificação com referência a florestas aleatórias;
  - algoritmo RIPPER
  - algoritmo de florestas aleatórias SIRUS

- Comparação e avaliação de regras de indução, derivadas dos melhores modelos de classificação construídos, sobre escalas de severidade de incêndios por
  - medidas de avaliação de regras de indução (preditividade, estabilidade e simplicidade)
  - análise por especialista do domínio
- Aferir a separação por regiões geográficas na extração de regras de risco de incêndio
- Aferir o uso de diferentes sub-índices neste domínio

7.2 Classificação não supervisionada do nível de risco dos incêndios extremos

Nesta secção é descrito a aplicação do algoritmo de agrupamento difuso não supervisionado *fuzzy c-means (FCM)* e a classificação dos incêndios com o seu resultado.

A análise foi feita por região, onde para cada região foram analisadas e comparadas as melhores partições para os dois conjuntos de parâmetros **some** e **all**. Para cada conjunto de parâmetros foram selecionadas partições "candidatas" de acordo com o índice de validação interna de Xie e Beni. As partições selecionadas foram em seguida comparadas com o auxílio do erro entre a classificação com a partição e a classificação com a escala de risco existente.

Finalmente aferiu-se a correção e robustez das partição analisando os protótipos das mesmas face a escala de risco existente, e com o uso de mapas de agrupamento difuso. Com a conciliação destes vários pontos escolheu-se a melhor partição para cada uma das regiões estudadas.

7.2.1 Norte Litoral

	Clusters	4	5	6	7	8	9	10
XB	some	47.59	76.65	52.78	38.47	49.80	42.31	35.22
	all	86.48	51.41	49.25	54.85	32.25	32.03	28.56

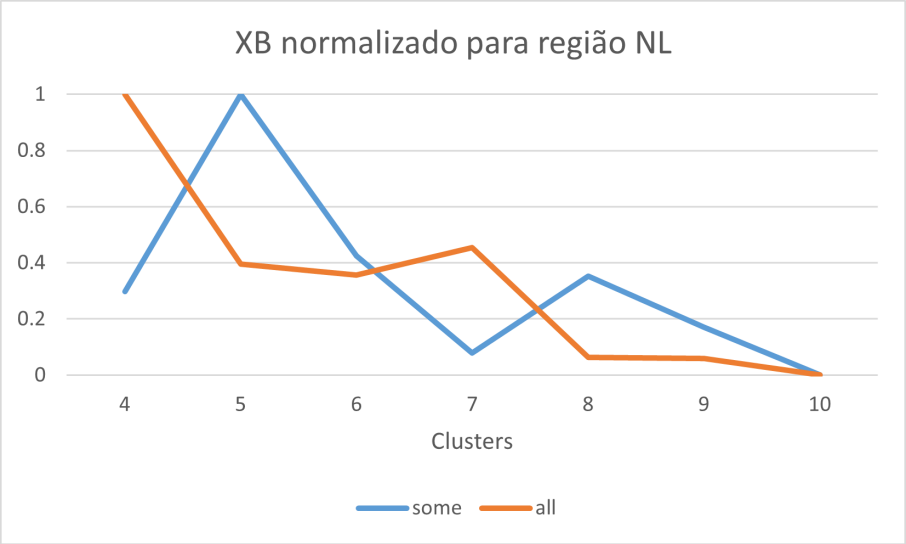


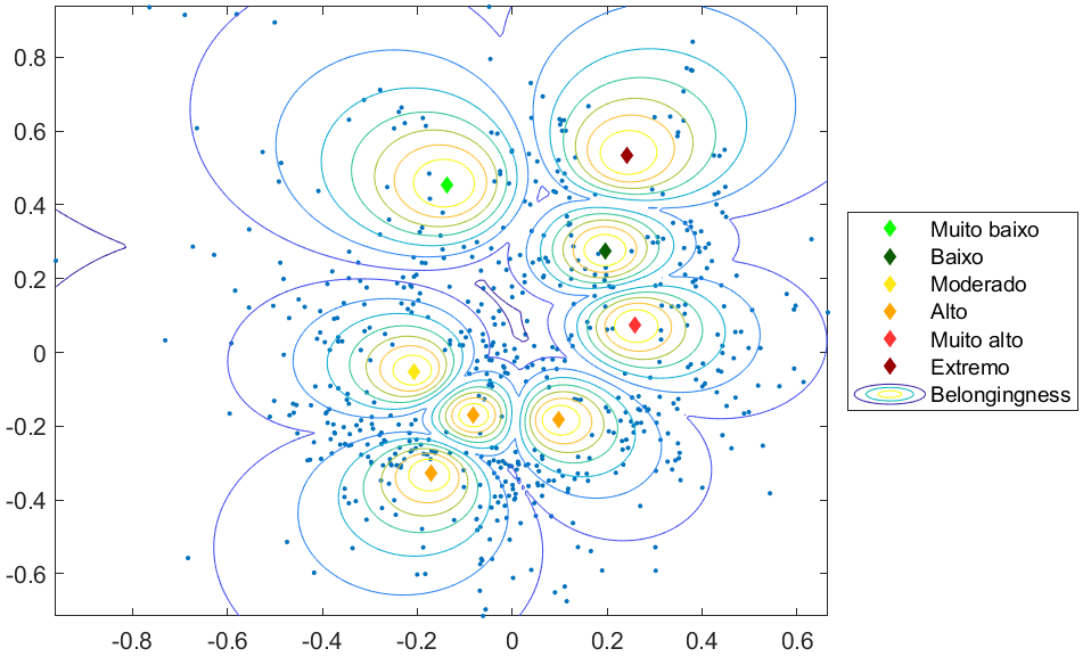
Figura 7.1: Índice de Xie-Beni para Norte Litoral com 5 parâmetros (some) e 7 parâmetros (all).

Tabela 7.1: Norte Litoral. Erro por partições, obtido com a comparação da classificação obtida com o FCM com a classificação com a escala IPMA do índice FWI.
Erro definido como o número de incêndios classificados com um nível de risco inferior ao nível de risco obtido com a escala de referência.

	Muito baixo	Baixo	Moderado	Alto	Muito alto	Extremo	Erro
someC6	7	12	38	40	—	5	102 (17.2%)
allC6	—	3	12	40	—	5	60 (10.1%)
someC7	6	9	21	49	—	5	90 (15.2%)
allC7	8	11	33	8	0	5	65 (11.0%)
someC8	4	9	33	5	0	5	56 (9.4%)
allC8	7	7	12	8	0	5	39 (6.6%)

Classificação	FWI	FFMC	DMC	DC	ISI	BUI	CHI
Muito baixo	4.44	76.29	16.75	89.02	2.82	20.1	2.5
Baixo	7.09	85.4	17.69	57.34	4.72	19.35	7.1
Moderado	20.76	88.8	54.54	459.89	6.69	81.53	6.6
Alto	20.93	87.37	75.13	488.96	5.61	106.23	3
Alto	28.71	90.98	76.13	467.77	8.75	105.73	8.5
Alto	29.48	90.4	99.57	551.05	7.88	134.8	6.3
Muito alto	38.1	92.26	136.27	610.98	10.32	172.4	8.3
Extremo	56.49	93.89	156.05	517.72	18.9	175.16	9.8

(a) Protótipos da partição e comparação com escala de referência, escala IPMA para o índice FWI e escala EEFIS para os restantes índices (exceto CHI). Índice a verde indica uma classificação semelhante à classificação do protótipo (coluna "Classificação"). Amarelo, uma classificação a um nível de risco de diferença. Vermelho, mais de um nível de risco de diferença.



(b) Visualização das partições com o algoritmo de FUZZSAM e com redução para duas dimensões com PCA. O centroide de cada partição é representado com um losango com a cor correspondente à classe de risco. Os pontos a azul representam os incêndios e as isolinhas indicam o grau de pertença (belongingness) a cada partição.

Figura 7.2: FCM com C = 8 com os parâmetros **all** para a região Norte Litoral

Observando o gráfico do índice de Xie e Beni, Figura 7.1, o melhor número de centroides, de acordo com este índice, é 7, para o conjunto de parâmetros **some** e 5 ou 8 para o conjunto de parâmetros **all**. Pelo facto de com 5 centroides com o conjunto de parâmetros **all** o algoritmo FCM não resultar numa classificação com o nível de risco "Muito alto" este valor será ignorado. As partições escolhidas com o índice de Xie e Beni para Norte Litoral são então **someC7** e **allC8**.

Na tabela com o erro das partições, Tabela 7.1, é possível observar que a partição

someC7 resulta em um erro de 15.2% com 49 incêndios com classificação "Alto", subestimados em relação à escala de referência e sem nenhum centroide com classificação de nível "Muito alto". A partição **allC7** resulta num erro inferior, 6.6%, uma classificação mais correta para o nível de risco "Alto", apenas com 8 incêndios e com classificação de nível "Muito alto". A omissão do nível de risco "Muito alto" apesar de acordo com a escala de referência 21.2% dos incêndios terem esta classificação, com a existência dos níveis adjacentes "Alto" e "Extremo" é fundamento para a exclusão da partição **someC7**.

Observando os protótipos da partição **allC8**, Figura 7.2, no mapa de agrupamento difuso existe alguns pontos com maior pertença a *clusters* de nível alto ("Muito alto" e "Extreno") e ao *cluster* de nível "Baixo", o que não seria esperado. No entanto na comparação dos protótipos com a escala de referência não aparenta existir nenhum problema, exceto os valor do índice de FFMC e DC serem inferiores no protótipo "Extremo" quando comparados com o protótipo "Muito alto". Com base nos critérios apresentados para a região Norte Litoral foi escolhido a partição **allC8**.

7.2.2 Norte Interior

	Clusters	4	5	6	7	8	9	10
XB	some	36.60	20.24	21.30	16.13	12.59	12.10	12.34
	all	28.94	28.38	17.98	21.05	27.90	31.97	40.62

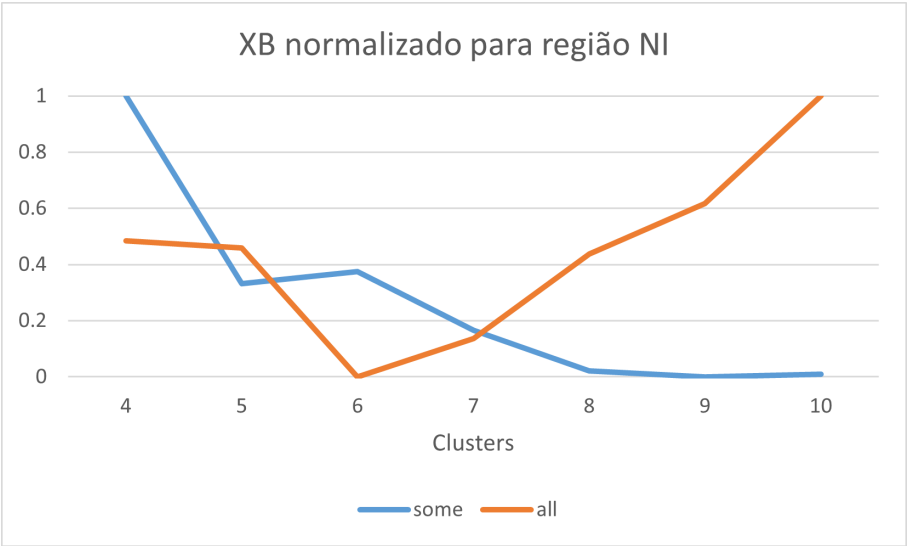


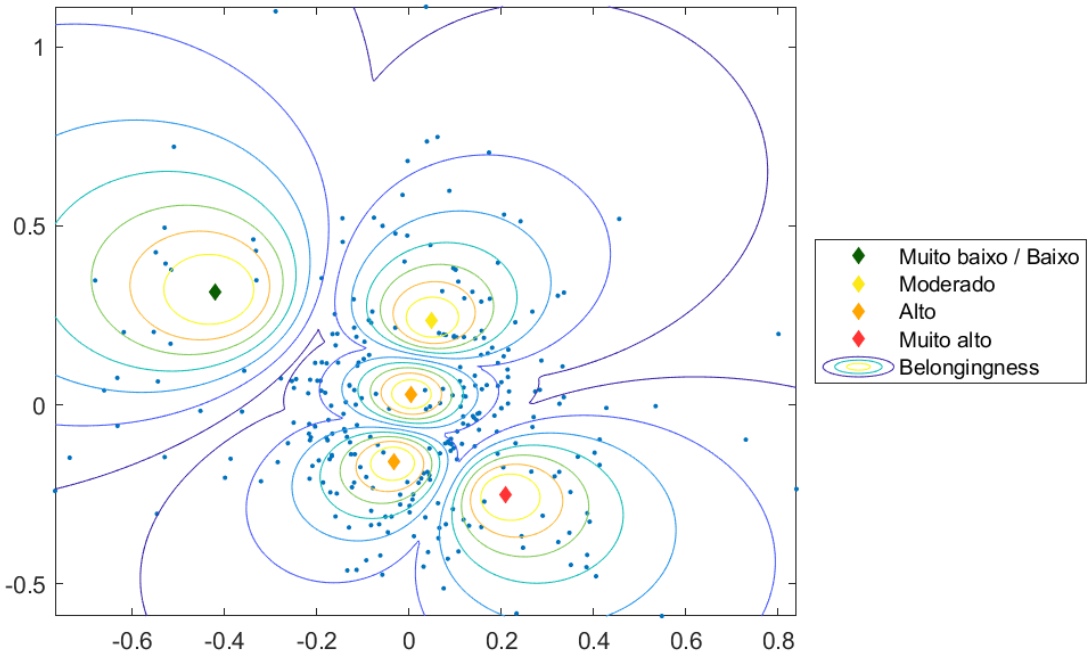
Figura 7.3: Índice de Xie-Beni para Norte Interior com 5 parâmetros (some) e 7 parâmetros (all).

Tabela 7.2: Norte Interior. Erro por partições, obtido com a comparação da classificação obtida com o FCM com a classificação com a escala IPMA do índice FWI.
Erro definido como o número de incêndios classificados com um nível de risco inferior ao nível de risco obtido com a escala de referência.

	Muito baixo	Baixo	Moderado	Alto	Muito alto	Erro
someC5	—	7	14	10	4	35 (12.5%)
someC6	—	7	13	10	4	34 (12.2%)
allC6	—	8	11	4	4	27 (9.7%)
someC7	2	8	17	10	4	41 (14.7%)
allC7	—	8	10	3	4	25 (9.0%)
someC8	1	7	12	17	4	41 (14.7%)
allC8	—	7	5	5	4	21 (7.5%)

Classificação	FWI	FFMC	DC	ISI	CHI
Baixo	8.46	82.86	109.29	4.96	5.49
Moderado	21.61	87.61	577.54	5.46	3.04
Alto	27.4	90.58	662.49	6.87	7.72
Alto	33.11	90.67	723.57	8.41	4.18
Muito alto	40.02	92.7	678.79	11.35	8.41

(a) Protótipos da partição e comparação com escala de referência, escala IPMA para o índice FWI e escala EEFIS para os restantes índices (exceto CHI). Índice a verde indica uma classificação semelhante à classificação do protótipo (coluna "Classificação"). Amarelo, uma classificação a um nível de risco de diferença. Vermelho, mais de um nível de risco de diferença.

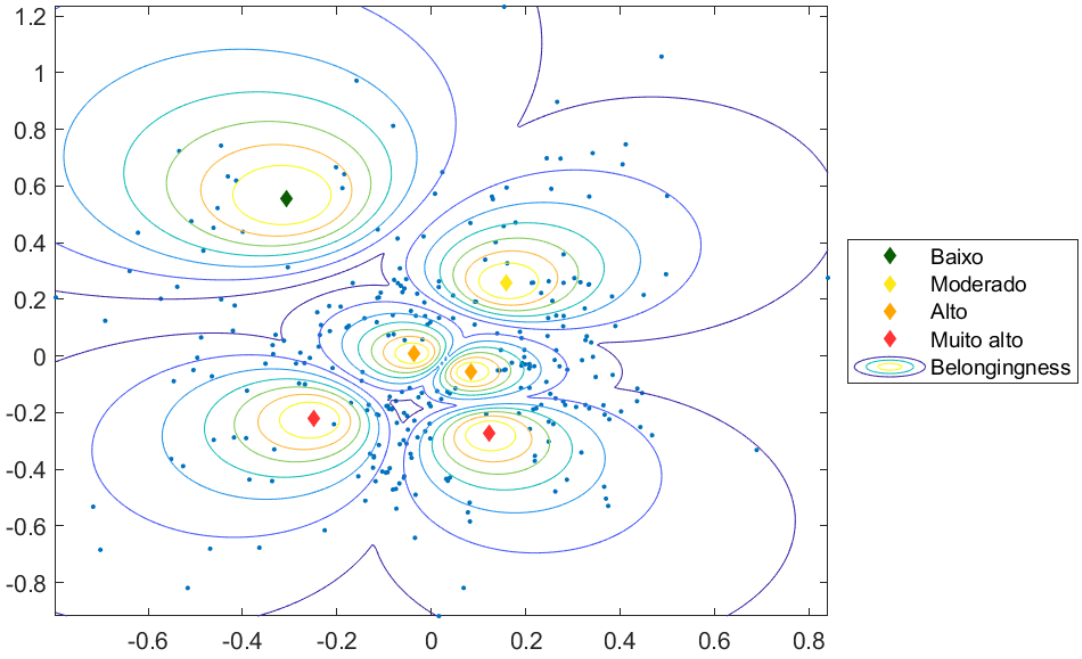


(b) Visualização das partições com o algoritmo de FUZZSAM e com redução para duas dimensões com PCA. O centroide de cada partição é representado com um losango com a cor correspondente à classe de risco. Os pontos a azul representam os incêndios e as isolinhas indicam o grau de pertença (belongingness) a cada partição.

Figura 7.4: FCM com C = 5 com os parâmetros **some** para a região Norte Interior

Classificação	FWI	FFMC	DMC	DC	ISI	BUI	CHI
Baixo	8.19	83.55	21.16	90.39	5.2	24.69	5.71
Moderado	21.46	87.51	86.57	567.76	5.46	122.27	3.04
Alto	25.67	90.07	87.33	632.8	6.78	126.85	7.19
Alto	31.04	90.15	130.93	695.99	7.82	174.1	4.51
Muito alto	36.31	91.72	209.45	745.8	9.1	242.61	5.76
Muito alto	36.66	92.26	128.1	668.23	10.05	169.82	8.51

(a) Protótipos da partição e comparação com escala de referência, escala IPMA para o índice FWI e escala EEFIS para os restantes índices (exceto CHI). Índice a verde indica uma classificação semelhante à classificação do protótipo (coluna "Classificação"). Amarelo, uma classificação a um nível de risco de diferença. Vermelho, mais de um nível de risco de diferença.



(b) Visualização das partições com o algoritmo de FUZZSAM e com redução para duas dimensões com PCA. O centroide de cada partição é representado com um losango com a cor correspondente à classe de risco. Os pontos a azul representam os incêndios e as isolinhas indicam o grau de pertença (belongingness) a cada partição.

Figura 7.5: FCM com C = 6 com os parâmetros **all** para a região Norte Interior

Observando o gráfico do índice de Xie e Beni, Figura 7.3, o melhor número de centroides, de acordo com este índice, é 5, para o conjunto de parâmetros **some** e 6 para o conjunto de parâmetros **all**. As partições escolhidas com o índice de Xie e Beni para Norte Interior são então **someC5** e **allC6**.

Na tabela com o erro das partições 7.2, é possível observar que ambas as partições selecionadas **someC5** e **allC6** têm erro baixo, 12.5% e 9.7% respetivamente.

Observando os protótipos das partições **someC5** e **allC6**, Figura 7.4 e Figura 7.5 respetivamente, é possível observar que os protótipos da partição **someC5** aparentam ser

melhores que os protótipos da partição **allC6** na classificação "Moderado" e mais significativamente na classificação "Muito alto". Não aparenta existir nenhum problema observável nos mapas de agrupamento difuso. Com base nos critérios apresentados para a região Norte Interior foi escolhido a partição **someC5**.

7.2.3 Centro Litoral

	Clusters	4	5	6	7	8	9	10
XB	some	31.55	28.98	18.97	19.90	17.36	12.09	12.07
	all	22.93	23.70	29.98	34.11	18.97	17.92	13.07

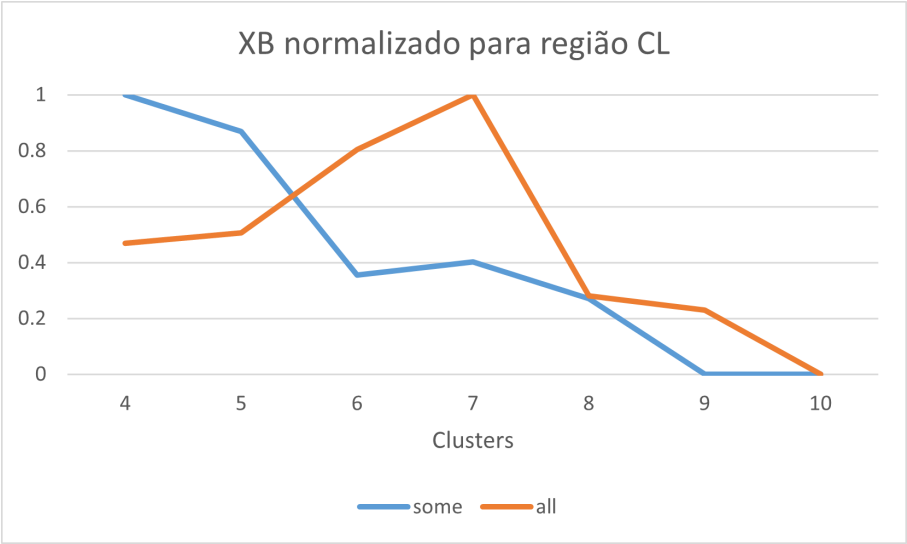


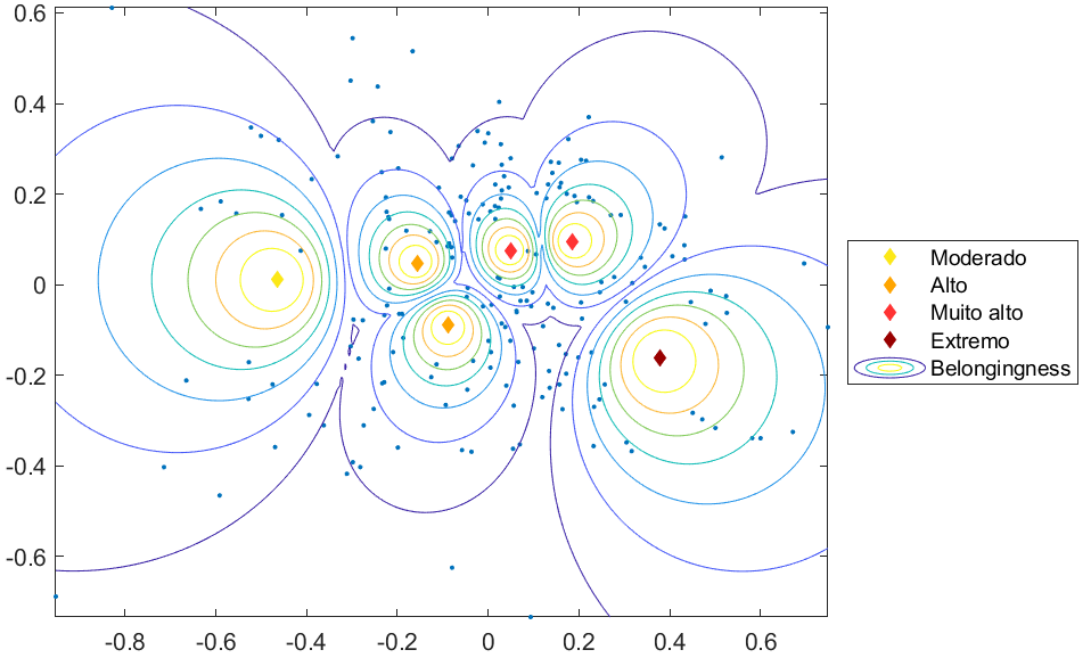
Figura 7.6: Índice de Xie-Beni para Centro Litoral com 5 parâmetros (some) e 7 parâmetros (all).

Tabela 7.3: Centro Litoral. Erro por partições, obtido com a comparação da classificação obtida com o FCM com a classificação com a escala IPMA do índice FWI.
Erro definido como o número de incêndios classificados com um nível de risco inferior ao nível de risco obtido com a escala de referência.

	Moderado	Alto	Muito alto	Extremo	Erro
someC6	5	1	5	2	13 (6.7%)
allC6	5	1	1	2	9 (4.6%)
someC7	4	7	4	2	17 (8.8%)
allC7	4	10	2	2	18 (9.3%)
someC8	4	9	5	2	20 (10.3%)
allC8	4	0	4	2	10 (5.2%)

Classificação	FWI	FFMC	DC	ISI	CHI
Moderado	20.01	84.7	643.76	5.25	2.7
Alto	26.53	88.83	725.83	6.86	7.4
Alto	28.97	90.16	538.41	8.35	7.5
Muito alto	37.46	92.39	699.64	10.04	9.2
Muito alto	43.47	92.31	699.9	12.67	6.8
Extremo	52.82	93.98	661.59	17.56	9

(a) Protótipos da partição e comparação com escala de referência, escala IPMA para o índice FWI e escala EEFIS para os restantes índices (exceto CHI). Índice a verde indica uma classificação semelhante à classificação do protótipo (coluna "Classificação"). Amarelo, uma classificação a um nível de risco de diferença. Vermelho, mais de um nível de risco de diferença.

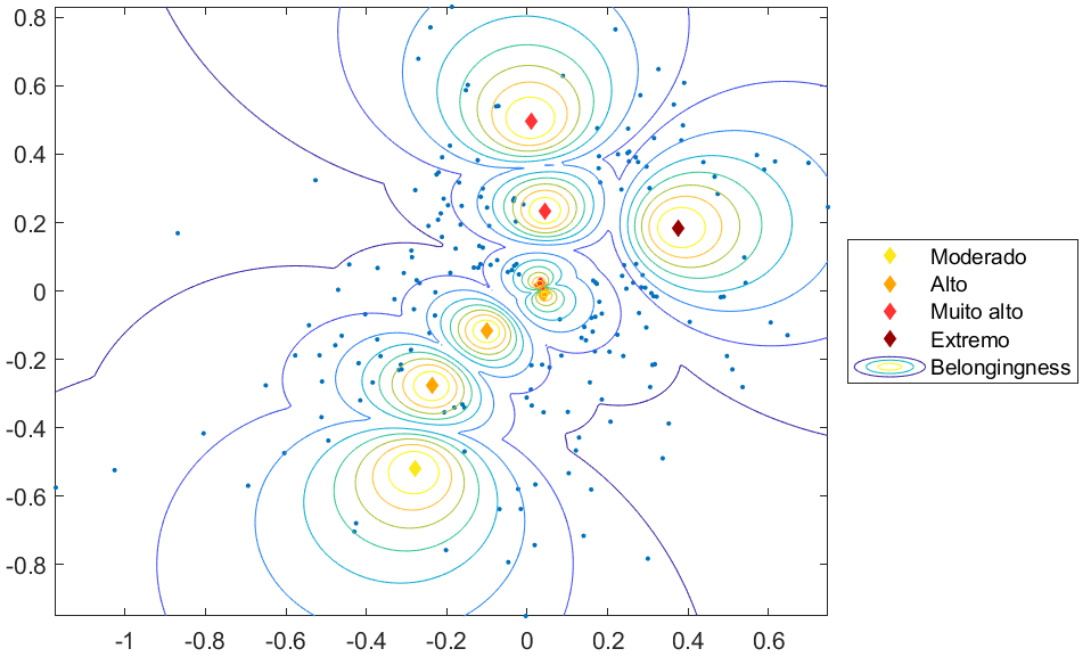


(b) Visualização das partições com o algoritmo de FUZZSAM e com redução para duas dimensões com PCA. O centroide de cada partição é representado com um losango com a cor correspondente à classe de risco. Os pontos a azul representam os incêndios e as isolinhas indicam o grau de pertença (belongingness) a cada partição.

Figura 7.7: FCM com C = 6 com os parâmetros **some** para a região Centro Litoral

Classificação	FWI	FFMC	DMC	DC	ISI	BUI	CHI
Moderado	17.98	85.1	60.1	517.17	5.52	87.69	4.3
Alto	25.68	89.06	61.28	615.36	8.09	95.28	7.3
Alto	29.86	89.48	101.41	704.83	8.06	144.09	7.2
Alto	32.95	90.78	135.01	675.11	8.41	176.4	8.1
Muito alto	40.05	92.63	227.56	794.1	10.37	261.79	8.1
Muito alto	40.78	91.87	113.63	654.22	12.1	154.65	7.4
Muito alto	42.33	93.32	167.13	688.32	11.68	205.17	8.9
Extremo	53.25	93.91	143.89	652.05	17.3	181.17	8.6

(a) Protótipos da partição e comparação com escala de referência, escala IPMA para o índice FWI e escala EEFIS para os restantes índices (exceto CHI). Índice a verde indica uma classificação semelhante à classificação do protótipo (coluna "Classificação"). Amarelo, uma classificação a um nível de risco de diferença. Vermelho, mais de um nível de risco de diferença.



(b) Visualização das partições com o algoritmo de FUZZSAM e com redução para duas dimensões com PCA. O centroide de cada partição é representado com um losango com a cor correspondente à classe de risco. Os pontos a azul representam os incêndios e as isolinhas indicam o grau de pertença (belongingness) a cada partição.

Figura 7.8: FCM com C = 8 com os parâmetros **all** para a região Centro Litoral

Observando o gráfico do índice de Xie e Beni, Figura 7.6, o melhor número de centroides, de acordo com este índice, é 6, para o conjunto de parâmetros **some** e 8 para o conjunto de parâmetros **all**. As partições escolhidas com o índice de Xie e Beni para Norte Interior são então **someC6** e **allC8**.

Na tabela com o erro das partições 7.3, é possível observar que ambas as partições selecionadas **someC6** e **allC8** têm erro baixo, 6.7% e 5.2%, respetivamente.

Observando os protótipos das partições **someC6** e **allC8**, Figura 7.7 e Figura 7.8 respetivamente, a partição **allC8** os protótipos referentes ao níveis "Alto" e "Muito alto" são mais concordantes com a escala de referência. Não aparenta existir nenhum problema observável nos mapas de agrupamento difuso. Com base nos critérios apresentados para a região Centro Litoral foi escolhido a partição **allC8**.

7.2.4 Centro Interior

	Clusters	4	5	6	7	8	9	10
XB	some	89.99	66.57	59.91	45.14	35.42	32.60	26.26
	all	95.95	93.08	95.03	66.04	69.47	56.83	47.25

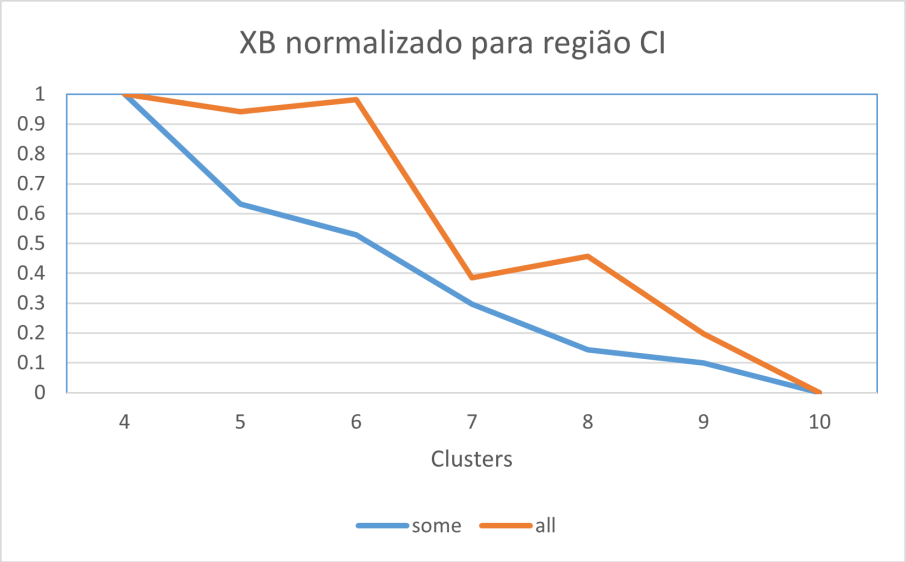


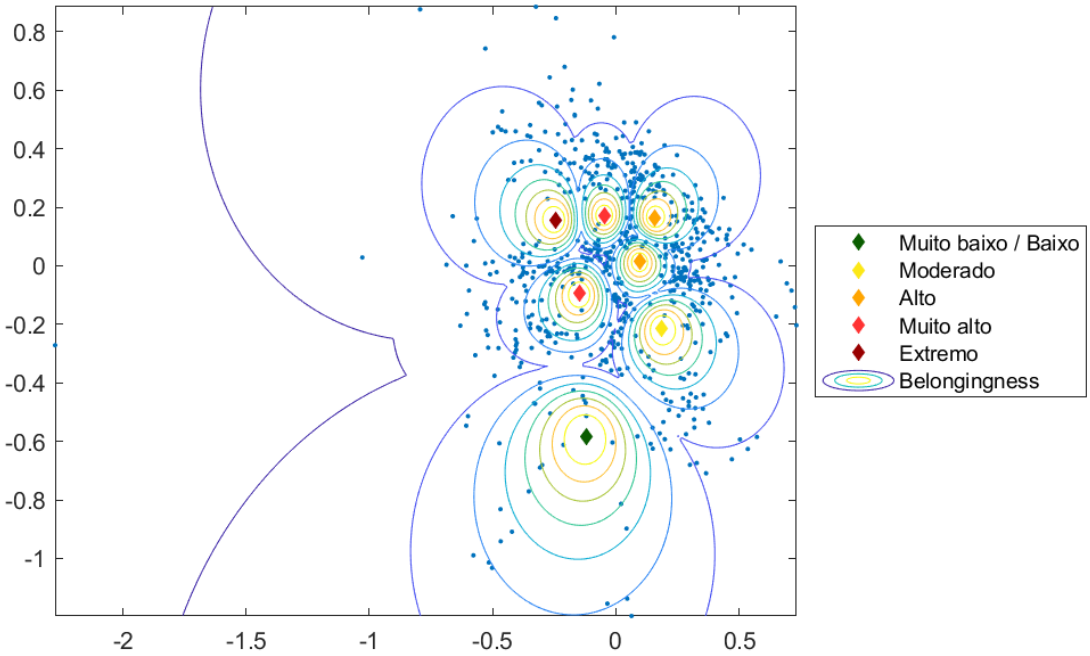
Figura 7.9: Índice de Xie-Beni para Centro Interior com 5 parâmetros (some) e 7 parâmetros (all).

Tabela 7.4: Centro Interior. Erro por partições, obtido com a comparação da classificação obtida com o FCM com a classificação com a escala IPMA do índice FWI.
Erro definido como o número de incêndios classificados com um nível de risco inferior ao nível de risco obtido com a escala de referência.

	Muito baixo/Baixo	Moderado	Alto	Muito alto	Extremo	Erro
someC7	0	38	45	1	4	88 (13.4%)
allC7	0	47	7	10	4	68 (10.4%)
someC8	0	30	36	2	4	72 (11.0%)
allC8	0	22	22	7	4	55 (8.4%)

Classificação	FWI	FFMC	DMC	DC	ISI	BUI	CHI
Baixo	5.90	64.45	64.19	604.30	1.49	98.41	2.46
Moderado	24.24	88.58	93.71	545.99	6.43	126.39	3.92
Alto	30.81	90.67	115.30	678.85	8.08	156.60	6.24
Alto	31.84	91.70	107.82	550.10	8.83	140.04	8.20
Muito alto	35.94	90.94	196.60	747.62	9.11	231.39	4.60
Muito alto	39.57	92.85	144.21	650.84	11.07	180.67	8.39
Extremo	43.19	92.96	203.61	784.18	11.93	241.14	7.53

(a) Protótipos da partição e comparação com escala de referência, escala IPMA para o índice FWI e escala EEFIS para os restantes índices (exceto CHI). Índice a verde indica uma classificação semelhante à classificação do protótipo (coluna "Classificação"). Amarelo, uma classificação a um nível de risco de diferença. Vermelho, mais de um nível de risco de diferença.



(b) Visualização das partições com o algoritmo de FUZZSAM e com redução para duas dimensões com PCA. O centroide de cada partição é representado com um losango com a cor correspondente à classe de risco. Os pontos a azul representam os incêndios e as isolinhas indicam o grau de pertença (belongingness) a cada partição.

Figura 7.10: FCM com C = 7 com os parâmetros **all** para a região Centro Interior

Observando o gráfico do índice de Xie e Beni, Figura 7.9, o melhor número de centroides, de acordo com este índice, é 5, para o conjunto de parâmetros **some** e 7 para o conjunto de parâmetros **all**. Pelo facto de com 5 centroides com o conjunto de parâmetros **some** o algoritmo FCM não resultar numa classificação com o nível de risco "Extremo" este valor será ignorado. Como índice de Xie e Beni, para Centro Interior selecionou-se a partição **allC7**.

Na tabela com o erro das partições 7.4 é possível verificar que a partição **allC7** apresenta um bom valor de erro, 10.4%. Com o conjunto de parâmetros some, a partição que

resulta em um menor erro é a partição **someC8** com 11.0%. Apesar de um erro total semelhante e de um erro nas classes "Moderado" e "Muito alto" ligeiramente menor esta partição apresenta um erro no nível de risco "Alto" significativamente maior.

Observando os protótipos da partição **allC7**, Figura 7.10, os protótipos com exceção do de nível "Baixo" são concordantes com a escala de referência. O protótipo com classificação "Baixo" é caracterizado por índices DMC, DC e BUI referentes a uma classificação de nível "Alto" na escala de referência. No entanto, tanto para o índice DMC e ISI este valor é o menor valor de todos os protótipos o que pode indicar que isto deve-se ao conjunto de dados utilizados ou às características da região e não a um incorreto comportamento do procedimento utilizado. Com base nos critérios apresentados para a região Centro Interior foi escolhido a partição **allC7**.

7.2.5 Sul

	Clusters	4	5	6	7	8	9	10
XB	some	16.84	12.73	14.65	13.68	7.93	6.16	6.60
	all	21.16	19.25	15.47	16.98	18.85	18.23	10.98

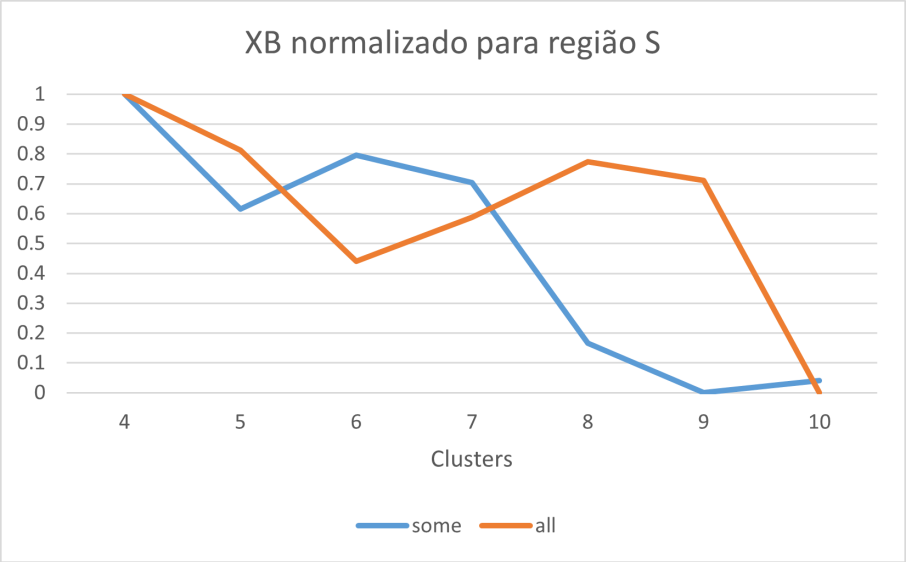


Figura 7.11: Índice de Xie-Beni para Sul com 5 parâmetros (some) e 7 parâmetros (all).

Tabela 7.5: Sul. Erro por partições, obtido com a comparação da classificação obtida com o FCM com a classificação com a escala IPMA do índice FWI.
Erro definido como o número de incêndios classificados com um nível de risco inferior ao nível de risco obtido com a escala de referência.

	Muito baixo	Baixo	Moderado	Alto	Muito alto	Extremo	Erro
someC5		0	17	7	2	8	34 (15.1%)
someC6	0	3	18	14	2	8	45 (20.0%)
allC6	0	4	—	12	4	8	28 (12.4%)
someC7	0	1	17	27	20	8	73 (32.4%)
allC7	0	2	—	4	13	8	27 (12.0%)
someC8	0	1	17	43	11	8	80 (35.6%)
allC8	0	1	—	19	6	8	34 (15.1%)

Observando o gráfico do índice de Xie e Beni, Figura 7.11, o melhor número de centroides, de acordo com este índice, é 5 ou 8, para o conjunto de parâmetros **some** e 6 para o conjunto de parâmetros **all**. Como índice de Xie e Beni, para a região Sul selecionou-se as partições **someC5**, **someC8** e **allC6**.

Na tabela com o erro das partições 7.5 é possível observar que a partição **someC8** têm um erro bastante alto (35.6%), justificando a exclusão desta partição. Das restantes duas, **someC5** e **allC6** ambas apresentam valores de erro semelhantes, 15.1% e 12.4%, respetivamente. A classificação com a partição **allC6** têm a desvantagem de não conter o

nível "Moderado", mas pelo conjunto de dados conter apenas 5.3% de incêndios classificados com este nível de risco com a escala de referência este não é motivo suficiente para desconsiderar esta partição.

A apresentação apenas do erro por classe de risco e não da matriz de confusão completa vêm por este valor ser suficiente para avaliar as classificações e para simplificar a apresentação dos resultados. No entanto as matrizes completas para cada partição foram calculadas e analisadas para averiguar a correção das classificações. Apenas na região Sul se observou um problema com estas matrizes, justificando a ligeira alteração na metodologia.

Observando as matrizes completas da comparação da classificação das partições com a classificação de referência, Tabelas 7.6, é possível observar que em ambas as partições, **someC5** e **allC6**, existem um elevado número de incêndios com classificação "Alto" a ser classificados com "Muito alto", a classificação obtida com a partição **allC7** resolve este problema ao mesmo tempo que mantém um erro total reduzido, semelhante às outras duas partições (12.0%).

Tabela 7.6: Matrizes completas do erro da classificação obtida com a partição quando comparada com a escala do índice FWI do IPMA, para a região Sul.

	Baixo	Moderado	Alto	Muito alto	Extremo	
Muito baixo	39	4	1	0	0	44 (19.6%)
Baixo	2	6	3	1	0	12 (5.3%)
Moderado	0	8	4	0	0	12 (5.3%)
Alto	0	17	21	27	0	65 (28.9%)
Muito alto	0	0	7	31	12	50 (22.2%)
Extremo	0	0	0	2	32	34 (15.1%)
Máximo	0	0	0	0	8	8 (3.6%)
	41 (16.1%)	35 (13.7%)	36 (14.1%)	61 (23.9%)	52 (20.4%)	225
Erro	0	17	7	2	8	34 (15.1%)

(a) someC5

	Muito baixo	Baixo	Alto	Muito alto	Extremo	
Muito baixo	27	17	0	0	0	44 (19.6%)
Baixo	0	9	1	2	0	12 (5.3%)
Moderado	0	3	2	7	0	12 (5.3%)
Alto	0	1	31	33	0	65 (28.9%)
Muito alto	0	0	12	27	11	50 (22.2%)
Extremo	0	0	0	4	30	34 (15.1%)
Máximo	0	0	0	0	8	8 (3.6%)
	27 (12.0%)	30 (13.3%)	46 (20.4%)	73 (32.4%)	49 (21.8%)	225
Erro	0	4	12	4	8	28 (12.4%)

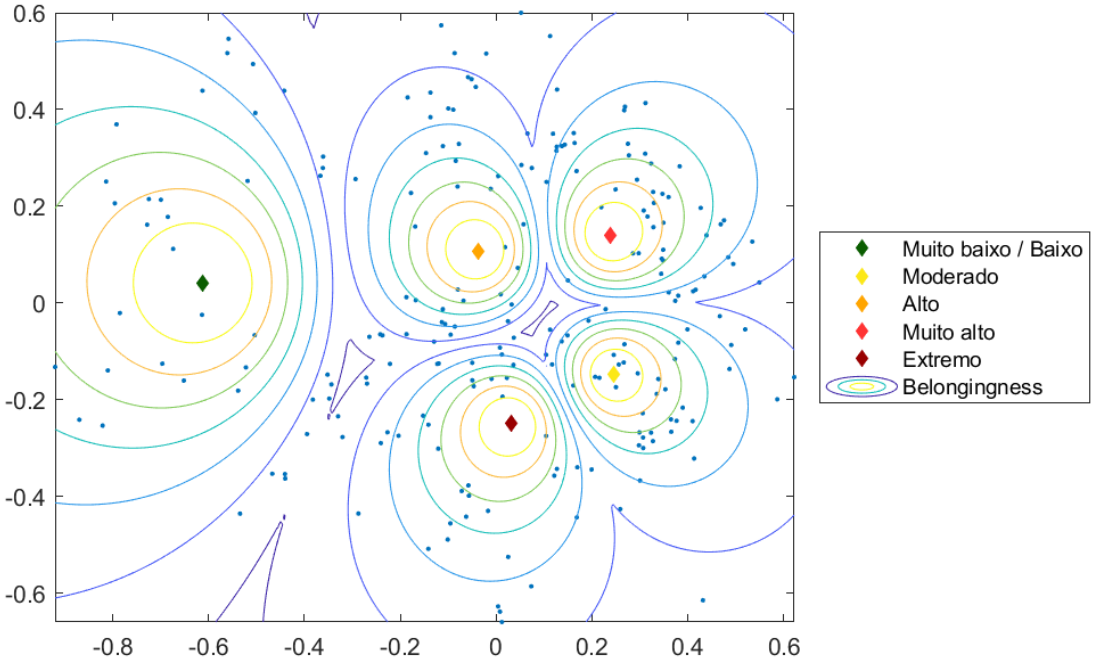
(b) allC6

	Muito baixo	Baixo	Alto	Muito alto	Extremo	
Muito baixo	27	17	0	0	0	44 (19.6%)
Baixo	0	7	5	0	0	12 (5.3%)
Moderado	0	2	10	0	0	12 (5.3%)
Alto	0	0	37	28	0	65 (28.9%)
Muito alto	0	0	4	45	1	50 (22.2%)
Extremo	0	0	0	13	21	34 (15.1%)
Máximo	0	0	0	0	8	8 (3.6%)
	27 (12.0%)	26 (11.6%)	56 (24.9%)	86 (38.2%)	30 (13.3%)	225 (100.0%)
Erro	0	2	4	13	8	27 (12.0%)

(c) allC7

Classificação	FWI	FFMC	DC	ISI	CHI
Baixo	2.54	56.8	582.74	0.88	3.11
Moderado	24.63	86.98	1111.55	5.39	4.18
Alto	29.44	88.46	658.73	7.49	4.84
Muito alto	39.09	92.17	784.75	10.27	9.03
Extremo	56.63	95.44	734.73	18.1	10.27

(a) Protótipos da partição e comparação com escala de referência, escala IPMA para o índice FWI e escala EEFIS para os restantes índices (exceto CHI). Índice a verde indica uma classificação semelhante à classificação do protótipo (coluna "Classificação"). Amarelo, uma classificação a um nível de risco de diferença. Vermelho, mais de um nível de risco de diferença.

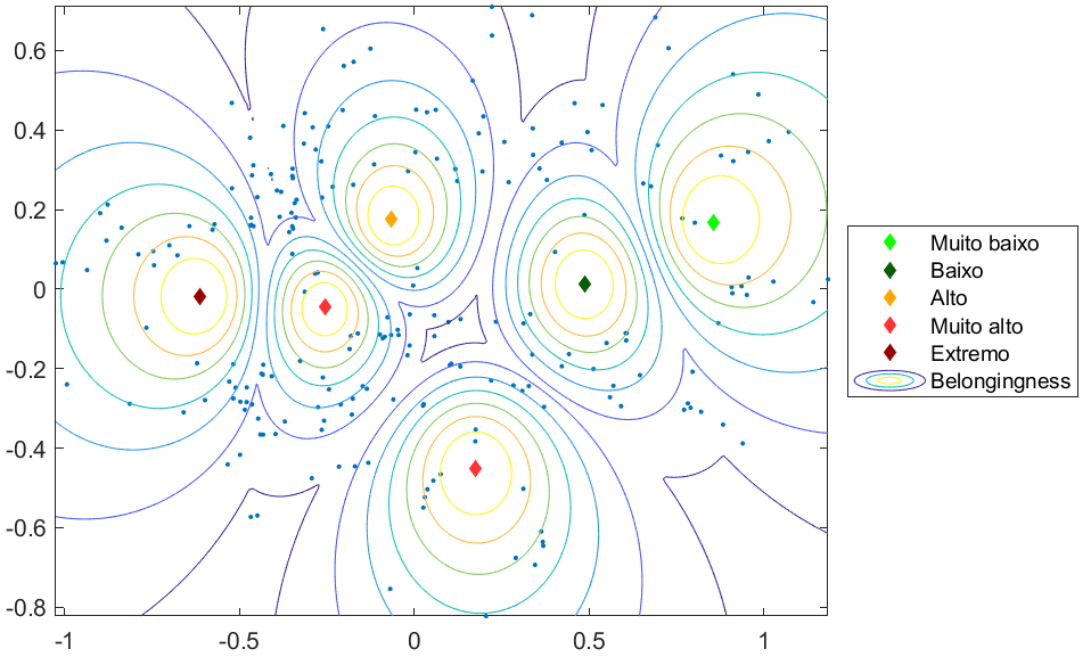


(b) Visualização das partições com o algoritmo de FUZZSAM e com redução para duas dimensões com PCA. O centroide de cada partição é representado com um losango com a cor correspondente à classe de risco. Os pontos a azul representam os incêndios e as isolinhas indicam o grau de pertença (belongingness) a cada partição.

Figura 7.12: FCM com C = 5 com os parâmetros some para a região Sul

Classificação	FWI	FFMC	DMC	DC	ISI	BUI	CHI
Muito baixo	2.09	54.43	28.01	538.21	0.82	46.05	3.24
Baixo	10.2	77.58	58.67	726.86	2.75	90.11	3.1
Alto	35.08	90.49	156.5	686.6	9.31	190.27	6.41
Muito alto	26.78	88.95	427.6	1211.78	5.66	453.86	4.2
Muito alto	40.02	92.46	245.68	820.21	10.45	276.69	8.86
Extremo	56.07	95.31	216.27	729.07	17.86	245.6	10.3

(a) Protótipos da partição e comparação com escala de referência, escala IPMA para o índice FWI e escala EEFIS para os restantes índices (exceto CHI). Índice a verde indica uma classificação semelhante à classificação do protótipo (coluna "Classificação"). Amarelo, uma classificação a um nível de risco de diferença. Vermelho, mais de um nível de risco de diferença.

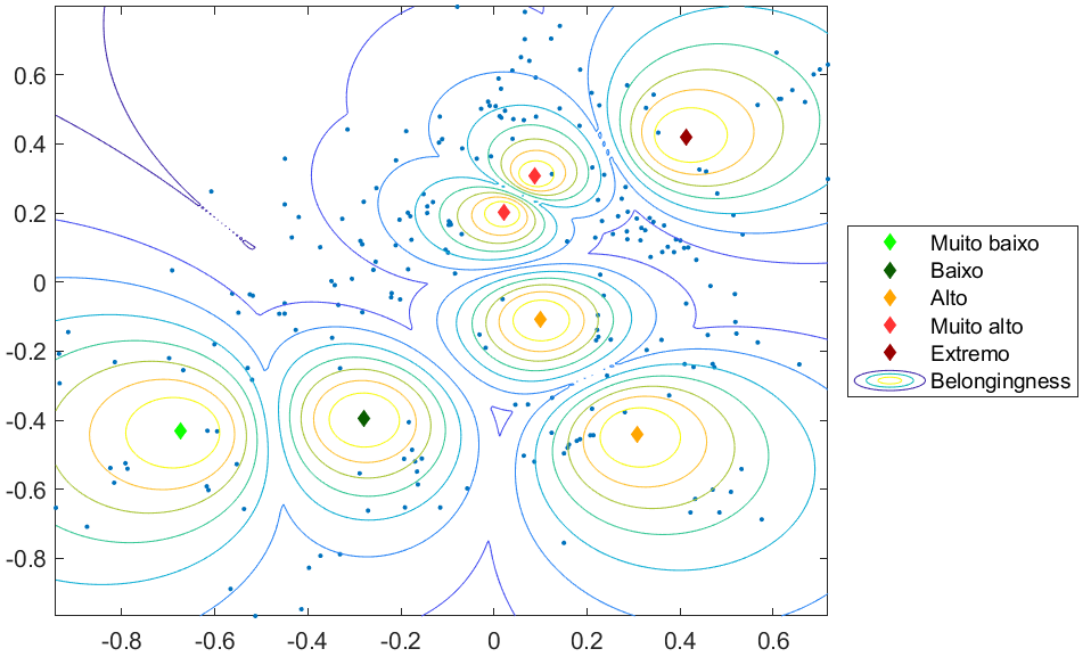


(b) Visualização das partições com o algoritmo de FUZZSAM e com redução para duas dimensões com PCA. O centroide de cada partição é representado com um losango com a cor correspondente à classe de risco. Os pontos a azul representam os incêndios e as isolinhas indicam o grau de pertença (belongingness) a cada partição.

Figura 7.13: FCM com C = 6 com os parâmetros all para a região Sul

Classificação	FWI	FFMC	DMC	DC	ISI	BUI	CHI
Muito baixo	1.77	53.76	25.9	521.53	0.74	43.05	3.21
Baixo	8.31	75.96	52.58	736.1	2.27	83.73	2.92
Alto	26.3	88.92	435.01	1225.25	5.48	460.68	4.04
Alto	31.12	89.24	146.05	669.34	7.96	179.78	5.23
Muito alto	39.1	92.24	264.22	869.86	10.04	296.37	8.55
Muito alto	42.43	92.94	186.52	699.29	11.77	218.11	8.99
Extremo	59.21	95.59	223.72	744.23	19.36	253.14	10.5

(a) Protótipos da partição e comparação com escala de referência, escala IPMA para o índice FWI e escala EEFIS para os restantes índices (exceto CHI). Índice a verde indica uma classificação semelhante à classificação do protótipo (coluna "Classificação"). Amarelo, uma classificação a um nível de risco de diferença. Vermelho, mais de um nível de risco de diferença.



(b) Visualização das partições com o algoritmo de FUZZSAM e com redução para duas dimensões com PCA. O centroide de cada partição é representado com um losango com a cor correspondente à classe de risco. Os pontos a azul representam os incêndios e as isolinhas indicam o grau de pertença (belongingness) a cada partição.

Figura 7.14: FCM com C = 7 com os parâmetros **all** para a região Sul

Observando os protótipos das partições **someC5** (Figura 7.12), **allC6** (Figura 7.13) e **allC7** (Figura 7.14), os protótipos dos níveis "Muito baixo" e "Baixo" têm índices bastante mais altos do que com a escala de referência. O protótipo com classificação "Moderado" do **someC5** apresenta um valor de DC bastante elevado, classificado como "Muito alto" na escala de referência. Um dos protótipos do **allC6** com classificação "Muito alto", apresenta valores baixos, o que pode explicar o problema do elevado número incêndios "Alto" estarem a ser classificados neste nível. Comparando os protótipos da partição **allC6** e

allC7 é possivél observar que o protótipo adicional do **allC7** resolve alguma da ambiguidade entre o nível "Alto" e "Muito alto". Vale referir, que apesar de a partição **allC7** não ter sido selecionada pelo resultado do índice de Xie e Beni está não é necessariamente pior visto que nenhum índice de validação é universal.

Ambas as partições **someC5** e **allC7** apresentam resultados satisfatórios e não aparenta existir nenhum problema observável nos mapas de agrupamento difuso. Apesar de a partição **someC5** resultar numa classificação com nível "Moderado", pela comparação com a escala de referência e pelo facto de apenas 5.3% dos incêndios terem esta classificação (com a escala de referência) esta não aparenta ser uma classificação robusta. Com base nos critérios apresentados para a região Sul foi escolhido a partição **allC7**.

7.3 Extração de regras de risco

Concluído o estágio anterior, os dados dos incêndios foram classificados com as classes de risco de acordo com a classificação dos protótipos da C-partição difusa selecionada. Este parâmetro de severidade será utilizado, em sequência, como a variável dependente dos modelos de extração de regras.

7.3.1 Extração de regras de risco com árvores de decisão e RIPPER e comparação de desempenho com referência a florestas aleatórias

Nesta secção é conduzido o estudo comparativo com os índices de validação: precisão, *recall* e F1, entre os resultados das árvores de decisão, da floresta aleatória e do RIPPER.

7.3.1.1 Norte Litoral

Tabela 7.7: Norte Litoral (NL allC8). Melhores valores encontrados dos hiperparâmetros para árvores de decisão, métricas de validação e número de nós e de nós-folha. DT X, refere-se à árvore de profundidade máxima X. RF aos resultados da floresta aleatória (random forest). Os valores de min_leaf e min_split referem-se ao número mínimo de dados por folha e por ramificação, respetivamente.

Modelo	Critério	min_leaf	min_split	Nº nodes	Nº folhas	Precisão	Recall	F1
DT 2	Gini	1	2	7	4	0.64	0.74	0.67
DT 3	Gini	2	22	13	7	0.76	0.65	0.68
DT 4	Gini	20	2	19	10	0.76	0.72	0.73
DT 5	Gini	4	4	31	16	0.76	0.74	0.74
DT 6	Gini	3	10	43	22	0.78	0.75	0.76
DT 7	Gini	2	17	55	28	0.79	0.77	0.78
DT 8	Gini	2	14	65	33	0.79	0.77	0.78
DT 9	Gini	1	4	105	53	0.83	0.79	0.79
RIPPER	—	—	—	—	—	0.76	0.75	0.75
RF	—	—	—	—	—	0.87	0.86	0.86

(a) Parâmetros **some**

Modelo	Critério	min_leaf	min_split	Nº nodes	Nº folhas	Precisão	Recall	F1
DT 2	Gini	1	2	7	4	0.61	0.72	0.66
DT 3	Gini	9	2	15	8	0.80	0.77	0.75
DT 4	Entropia	7	2	27	14	0.93	0.89	0.90
DT 5	Entropia	4	16	41	21	0.92	0.88	0.88
DT 6	Entropia	4	5	49	25	0.94	0.91	0.92
DT 7	Entropia	1	2	69	35	0.89	0.86	0.86
DT 8	Entropia	1	3	71	36	0.88	0.84	0.84
DT 9	Entropia	1	4	67	34	0.89	0.86	0.86
RIPPER	—	—	—	—	—	0.88	0.88	0.87
RF	—	—	—	—	—	0.92	0.91	0.91

(b) Parâmetros **all**

Como é possível observar na Tabela 7.7 para ambos os conjuntos de parâmetros a árvore com melhor F1 e com número de nós reduzido é a **DT 4**. Com os modelos **RIPPER**, obtêm-se valores próximos aos com as árvores de decisão. No geral, o conjunto de parâmetros **all** aparenta ser melhor nesta situação pelo facto dos modelos obterem melhores métricas e por essas métricas se aproximarem mais dos valores da **RF**.

7.3.1.2 Norte Interior

Tabela 7.8: Norte Interior (NI someC5). Melhores valores encontrados dos hiperparâmetros para árvores de decisão, métricas de validação e número de nós e de nós-folha. DT X, refere-se à árvore de profundidade máxima X. RF aos resultados da floresta aleatória (random forest). Os valores de min_leaf e min_split referem-se ao número mínimo de dados por folha e por ramificação, respetivamente.

Modelo	Critério	min_leaf	min_split	Nº nodes	Nº folhas	Precisão	Recall	F1
DT 2	Entropia	20	2	7	4	0.60	0.69	0.63
DT 3	Entropia	6	2	15	8	0.88	0.85	0.84
DT 4	Entropia	4	20	21	11	0.90	0.88	0.88
DT 5	Gini	3	2	33	17	0.84	0.85	0.84
DT 6	Entropia	6	29	21	11	0.88	0.85	0.84
DT 7	Gini	2	3	43	22	0.90	0.88	0.88
DT 8	Gini	2	7	37	19	0.90	0.88	0.88
DT 9	Gini	3	2	39	20	0.84	0.85	0.84
RIPPER	—	—	—	—	—	0.91	0.88	0.88
RF	—	—	—	—	—	0.93	0.92	0.91

(a) Parâmetros **some**

Modelo	Critério	min_leaf	min_split	Nº nodes	Nº folhas	Precisão	Recall	F1
DT 2	Entropia	19	2	7	4	0.60	0.69	0.63
DT 3	Entropia	1	9	15	8	0.88	0.85	0.84
DT 4	Entropia	3	4	25	13	0.93	0.92	0.92
DT 5	Gini	2	7	33	17	0.91	0.88	0.88
DT 6	Entropia	3	4	35	18	0.90	0.88	0.88
DT 7	Gini	4	5	35	18	0.80	0.81	0.80
DT 8	Gini	2	4	43	22	0.90	0.88	0.88
DT 9	Gini	3	5	39	20	0.76	0.77	0.76
RIPPER	—	—	—	—	—	0.88	0.85	0.84
RF	—	—	—	—	—	0.93	0.92	0.91

(b) Parâmetros **all**

Como é possível observar na Tabela 7.8, para ambos os conjuntos de parâmetros a árvore com melhor F1 e com número de nós reduzido é a **DT 4**. Com os modelos **RIPPER**, obtêm-se valores próximos aos com as árvores de decisão, mas ao contrário do que acontece com as árvores de decisão obtêm-se melhores valores com o conjunto de parâmetros **some**.

7.3.1.3 Centro Litoral

Tabela 7.9: Centro Litoral (CL allC8). Melhores valores encontrados dos hiperparâmetros para árvores de decisão, métricas de validação e número de nós e de nós-folha. DT X, refere-se à árvore de profundidade máxima X. RF aos resultados da floresta aleatória (random forest). Os valores de min_leaf e min_split referem-se ao número mínimo de dados por folha e por ramificação, respetivamente.

Modelo	Critério	min_leaf	min_split	Nº nodes	Nº folhas	Precisão	Recall	F1
DT 2	Entropia	1	2	7	4	0.79	0.79	0.79
DT 3	Entropia	5	19	13	7	0.72	0.79	0.75
DT 4	Gini	20	2	13	7	0.79	0.79	0.79
DT 5	Gini	3	18	21	11	0.72	0.79	0.75
DT 6	Gini	3	18	23	12	0.72	0.79	0.75
DT 7	Entropia	2	3	37	19	0.51	0.68	0.58
DT 8	Entropia	1	15	23	12	0.54	0.74	0.63
DT 9	Gini	1	2	51	26	0.80	0.74	0.72
RIPPER	—	—	—	—	—	0.72	0.79	0.75
RF	—	—	—	—	—	0.92	0.89	0.87

(a) Parâmetros **some**

Modelo	Critério	min_leaf	min_split	Nº nodes	Nº folhas	Precisão	Recall	F1
DT 2	Entropia	20	2	7	4	0.79	0.79	0.79
DT 3	Entropia	1	23	11	6	0.79	0.79	0.79
DT 4	Entropia	1	2	17	9	0.72	0.79	0.75
DT 5	Entropia	1	2	21	11	0.73	0.84	0.78
DT 6	Entropia	1	2	23	12	0.72	0.79	0.75
DT 7	Gini	3	19	21	11	0.83	0.84	0.83
DT 8	Gini	1	2	31	16	0.70	0.79	0.74
DT 9	Entropia	1	4	23	12	0.72	0.79	0.75
RIPPER	—	—	—	—	—	0.66	0.74	0.69
RF	—	—	—	—	—	0.92	0.89	0.87

(b) Parâmetros **all**

Observando a tabela 7.9a, com os parâmetros some, o valor de F1 obtido com a melhor árvore de decisão e o valor obtido com o **RIPPER** são bastante inferiores ao valor obtido com a **RF**. No caso modelos obtidos com os parâmetros **all**, 7.9b o valor obtido com a **DT 7** (0.83), é o único que se aproxima do valor obtido com a **RF** (0.87). As **RF** obtidas com ambos os conjuntos de parâmetros resultam em métricas iguais o que pode indicar que o conjunto de parâmetros some é suficiente para classificar os dados. Isso não se traduz, no entanto, para os modelos de árvore de decisão e RIPPER.

7.3.1.4 Centro Interior

Tabela 7.10: Centro Interior (CI allC7). Melhores valores encontrados dos hiperparâmetros para árvores de decisão, métricas de validação e número de nós e de nós-folha. DT X, refere-se à árvore de profundidade máxima X. RF aos resultados da floresta aleatória (random forest). Os valores de min_leaf e min_split referem-se ao número mínimo de dados por folha e por ramificação, respetivamente.

Modelo	Critério	min_leaf	min_split	Nº nodes	Nº folhas	Precisão	Recall	F1
DT 2	Gini	1	2	7	4	0.66	0.63	0.61
DT 3	Gini	1	3	15	8	0.80	0.68	0.68
DT 4	Gini	1	7	29	15	0.81	0.79	0.79
DT 5	Entropia	1	2	57	29	0.80	0.77	0.78
DT 6	Gini	1	33	45	23	0.77	0.73	0.73
DT 7	Entropia	1	14	87	44	0.78	0.76	0.76
DT 8	Entropia	3	3	111	56	0.83	0.82	0.82
DT 9	Entropia	2	14	101	51	0.83	0.82	0.82
RIPPER	—	—	—	—	—	0.83	0.81	0.81
RF	—	—	—	—	—	0.88	0.85	0.86

(a) Parâmetros **some**

Modelo	Critério	min_leaf	min_split	Nº nodes	Nº folhas	Precisão	Recall	F1
DT 2	Gini	1	2	7	4	0.66	0.66	0.64
DT 3	Gini	1	2	15	8	0.75	0.71	0.72
DT 4	Gini	15	32	25	13	0.79	0.76	0.76
DT 5	Gini	1	19	37	19	0.75	0.73	0.73
DT 6	Entropia	1	8	63	32	0.83	0.81	0.81
DT 7	Entropia	1	3	87	44	0.86	0.85	0.86
DT 8	Gini	1	3	107	54	0.82	0.79	0.80
DT 9	Gini	1	2	115	58	0.82	0.81	0.81
RIPPER	—	—	—	—	—	0.83	0.82	0.82
RF	—	—	—	—	—	0.88	0.87	0.87

(b) Parâmetros **all**

Como é possível observar na Tabela 7.10 para ambos os conjuntos de parâmetros a árvore com melhor F1 e com número de nós reduzido é a **DT 4**. Com os modelos **RIPPER**, obtêm-se, com ambos os conjuntos, melhores resultados que com as árvores de decisão. Os modelos obtêm valores próximos com os diferentes conjuntos de parâmetros.

7.3.1.5 Sul

Tabela 7.11: Sul (S allC7). Melhores valores encontrados dos hiperparâmetros para árvores de decisão, métricas de validação e número de nós e de nós-folha. DT X, refere-se à árvore de profundidade máxima X. RF aos resultados da floresta aleatória (random forest). Os valores de min_leaf e min_split referem-se ao número mínimo de dados por folha e por ramificação, respetivamente.

Modelo	Critério	min_leaf	min_split	Nº nodes	Nº folhas	Precisão	Recall	F1
DT 2	Entropia	20	2	7	4	0.46	0.62	0.52
DT 3	Gini	18	3	11	6	0.87	0.86	0.86
DT 4	Entropia	20	2	11	6	0.88	0.86	0.86
DT 5	Gini	2	27	15	8	0.87	0.81	0.80
DT 6	Entropia	1	4	35	18	0.87	0.86	0.86
DT 7	Gini	2	34	17	9	0.87	0.81	0.80
DT 8	Gini	1	2	49	25	0.87	0.86	0.86
DT 9	Gini	2	6	39	20	0.84	0.81	0.81
RIPPER	—	—	—	—	—	0.81	0.86	0.82
RF	—	—	—	—	—	0.93	0.90	0.90

(a) Parâmetros **some**

Modelo	Critério	min_leaf	min_split	Nº nodes	Nº folhas	Precisão	Recall	F1
DT 2	Entropia	20	2	7	4	0.46	0.62	0.52
DT 3	Gini	10	9	13	7	0.87	0.81	0.80
DT 4	Entropia	3	5	19	10	0.90	0.86	0.86
DT 5	Gini	1	27	15	8	0.87	0.81	0.80
DT 6	Entropia	1	4	33	17	0.84	0.81	0.81
DT 7	Entropia	3	5	37	19	0.87	0.86	0.86
DT 8	Gini	2	2	41	21	0.93	0.90	0.90
DT 9	Gini	7	4	29	15	0.84	0.81	0.81
RIPPER	—	—	—	—	—	0.73	0.71	0.70
RF	—	—	—	—	—	0.96	0.95	0.95

(b) Parâmetros **all**

Observando a Tabela 7.11 para ambos os conjuntos de parâmetros a árvore escolhida obtém um F1 de 0.86, mas, com os parâmetros **some** esta árvore têm uma profundidade inferior, sendo por isso escolhida a árvore **DT 3** com os parâmetros **some**. Em ambos os casos os modelos **RIPPER** resultam em métricas piores que as árvores de decisão.

7.3.2 Extração de regras de risco com algoritmo SIRUS

Nesta secção é descrito o estudo comparativo dos diferentes modelos baseados no algoritmo de extração de regras de florestas aleatórias SIRUS e a comparação com floresta aleatória.

Os modelos podem ser divididos em duas categorias, os modelos **else** e os modelos **boost**. Nos modelos **else** as regras são encadeadas, se tiver determinadas características é classificado com uma classe de risco, se não, é testado para outro conjunto de características. Nos modelos **boost**, é obtido um valor de pertença a cada nível, sendo a classificação feita com base nesses valores. Os modelos **boost**, dividem-se ainda em **boost plus**, **boost stand** e **boost multinom**, diferenciando-se apenas na forma a classificação é feita dado os níveis de pertença.

Tanto no **boost plus** como no **boost stand** as regras podem ser simplificadas para que a classificação possa ser feita à partir da leitura das regras e sem a necessidade de efetuar nenhum cálculo. No **boost multinom** é necessário multiplicar as pertenças por uma matriz. Como é necessário obter um conjunto de regras fácil de ler, este método não será tido em conta na seleção do modelo mas será usado para testar a robustez dos restantes métodos.

Nesta secção é escolhido o modelo **else** e modelo **boost** (**plus** ou **stand**) com maior F1 e com menor número de regras. Foi dada preferência a modelos com menor número de regras com valores de F1 muito próximos.

7.3.2.1 Portugal Continental

Tabela 7.12: Portugal Continental (PT allC7). Métricas de validação dos modelos baseados no algoritmo SIRUS.

Model	Precisão	Recall	F1
else 3	0.71	0.72	0.69
else 4	0.71	0.72	0.70
else 5	0.69	0.66	0.63
else 6	0.68	0.70	0.67
boost 3 plus	0.73	0.73	0.72
boost 3 stand	0.71	0.71	0.69
boost 4 plus	0.70	0.72	0.70
boost 4 stand	0.71	0.70	0.69
boost 5 plus	0.65	0.68	0.65
boost 5 stand	0.70	0.66	0.66
boost 6 plus	0.70	0.70	0.70
boost 6 stand	0.69	0.68	0.67
boost 3 multi	0.74	0.74	0.72
boost 4 multi	0.76	0.75	0.72
boost 5 multi	0.76	0.75	0.73
boost 6 multi	0.75	0.74	0.73
RF	0.89	0.89	0.88

(a) Parâmetros **some**

Model	Precisão	Recall	F1
else 3	0.80	0.80	0.80
else 4	0.83	0.83	0.82
else 5	0.83	0.82	0.81
else 6	0.79	0.79	0.79
boost 3 plus	0.79	0.79	0.78
boost 3 stand	0.80	0.78	0.78
boost 4 plus	0.81	0.80	0.80
boost 4 stand	0.80	0.78	0.78
boost 5 plus	0.80	0.79	0.79
boost 5 stand	0.81	0.79	0.79
boost 6 plus	0.82	0.82	0.81
boost 6 stand	0.81	0.80	0.80
boost 3 multi	0.86	0.85	0.85
boost 4 multi	0.85	0.84	0.83
boost 5 multi	0.84	0.83	0.82
boost 6 multi	0.85	0.84	0.84
RF	0.92	0.91	0.91

(b) Parâmetros **all**

Observando a Tabela 7.12 para ambos os conjuntos de parâmetros valores obtidos pelos modelos **SIRUS** são inferiores aos obtidos com **RF**. Os valores obtidos com os modelos **boost plus** e **boost stand** são inferiores mas próximos aos obtidos com **boost multi**, como seria esperado. Os modelos obtidos com os parâmetros **all** resultam em melhores métricas e métricas mais próximas dos valores obtidos com **RF**.

7.3.2.2 Norte Litoral

Tabela 7.13: Norte Litoral (NL allC8). Métricas de validação dos modelos baseados no algoritmo SIRUS.

Model	Precisão	Recall	F1
else 3	0.74	0.70	0.68
else 4	0.63	0.67	0.63
else 5	0.74	0.70	0.71
else 6	0.79	0.74	0.76
boost 3 plus	0.68	0.65	0.62
boost 3 stand	0.74	0.67	0.66
boost 4 plus	0.64	0.61	0.59
boost 4 stand	0.72	0.63	0.63
boost 5 plus	0.67	0.63	0.62
boost 5 stand	0.71	0.63	0.64
boost 6 plus	0.66	0.63	0.60
boost 6 stand	0.78	0.67	0.67
boost 3 multi	0.62	0.61	0.61
boost 4 multi	0.71	0.68	0.69
boost 5 multi	0.69	0.68	0.68
boost 6 multi	0.68	0.67	0.66
RF	0.87	0.86	0.86

(a) Parâmetros **some**

Model	Precisão	Recall	F1
else 3	0.75	0.74	0.74
else 4	0.75	0.74	0.73
else 5	0.75	0.74	0.74
else 6	0.83	0.82	0.82
boost 3 plus	0.80	0.77	0.78
boost 3 stand	0.73	0.70	0.71
boost 4 plus	0.66	0.63	0.64
boost 4 stand	0.67	0.65	0.65
boost 5 plus	0.66	0.65	0.65
boost 5 stand	0.69	0.65	0.65
boost 6 plus	0.74	0.74	0.72
boost 6 stand	0.68	0.65	0.66
boost 3 multi	0.81	0.77	0.78
boost 4 multi	0.76	0.74	0.74
boost 5 multi	0.75	0.72	0.73
boost 6 multi	0.80	0.75	0.77
RF	0.92	0.91	0.91

(b) Parâmetros **all**

Observando a Tabela 7.13 para ambos os conjuntos de parâmetros os modelos **SIRUS else** resultam em melhores métricas. Os modelos obtidos com os parâmetros **all** resultam em melhores métricas e métricas mais próximas dos valores obtidos com **RF**.

7.3.2.3 Norte Interior

Tabela 7.14: Norte Interior (NI someC5). Métricas de validação dos modelos baseados no algoritmo SIRUS.

Model	Precisão	Recall	F1
else 3	0.74	0.73	0.72
else 4	0.72	0.65	0.64
else 5	0.82	0.77	0.76
else 6	0.83	0.81	0.81
boost 3 plus	0.69	0.65	0.66
boost 3 stand	0.70	0.69	0.69
boost 4 plus	0.73	0.69	0.70
boost 4 stand	0.70	0.69	0.69
boost 5 plus	0.81	0.77	0.77
boost 5 stand	0.77	0.77	0.77
boost 6 plus	0.84	0.85	0.84
boost 6 stand	0.82	0.81	0.81
boost 3 multi	0.71	0.69	0.70
boost 4 multi	0.80	0.81	0.80
boost 5 multi	0.86	0.85	0.84
boost 6 multi	0.89	0.88	0.88
RF	0.93	0.92	0.91

(a) Parâmetros **some**

Model	Precisão	Recall	F1
else 3	0.83	0.81	0.80
else 4	0.82	0.77	0.76
else 5	0.77	0.77	0.77
else 6	0.76	0.73	0.73
boost 3 plus	0.84	0.85	0.84
boost 3 stand	0.84	0.85	0.84
boost 4 plus	0.89	0.85	0.84
boost 4 stand	0.84	0.85	0.84
boost 5 plus	0.85	0.85	0.85
boost 5 stand	0.81	0.81	0.81
boost 6 plus	0.86	0.85	0.84
boost 6 stand	0.77	0.77	0.76
boost 3 multi	0.89	0.85	0.84
boost 4 multi	0.86	0.85	0.84
boost 5 multi	0.89	0.88	0.88
boost 6 multi	0.86	0.85	0.84
RF	0.93	0.92	0.91

(b) Parâmetros **all**

Observando a Tabela 7.14 para ambos os conjuntos de parâmetros os modelos **SIRUS boost** resultam em melhores métricas. Os modelos obtêm métricas semelhantes independentemente do conjunto de parâmetros, mas os modelos treinados com os parâmetros **all** têm um menor número de regras.

7.3.2.4 Centro Litoral

Tabela 7.15: Centro Litoral (CL allC8). Métricas de validação dos modelos baseados no algoritmo SIRUS.

Model	Precisão	Recall	F1
else 3	0.89	0.89	0.89
else 4	0.89	0.89	0.89
else 5	0.83	0.84	0.83
else 6	0.79	0.79	0.78
boost 3 plus	0.84	0.84	0.84
boost 3 stand	0.84	0.84	0.84
boost 4 plus	0.77	0.79	0.78
boost 4 stand	0.77	0.79	0.78
boost 5 plus	0.73	0.74	0.73
boost 5 stand	0.84	0.84	0.84
boost 6 plus	0.83	0.84	0.83
boost 6 stand	0.77	0.79	0.78
boost 3 multi	0.86	0.84	0.82
boost 4 multi	0.92	0.89	0.87
boost 5 multi	0.67	0.79	0.72
boost 6 multi	0.77	0.79	0.78
RF	0.92	0.89	0.87

(a) Parâmetros **some**

Model	Precisão	Recall	F1
else 3	0.86	0.84	0.82
else 4	0.89	0.89	0.89
else 5	0.83	0.84	0.83
else 6	0.89	0.89	0.89
boost 3 plus	0.79	0.79	0.79
boost 3 stand	0.79	0.79	0.79
boost 4 plus	0.83	0.84	0.83
boost 4 stand	0.79	0.79	0.79
boost 5 plus	0.83	0.84	0.83
boost 5 stand	0.79	0.79	0.79
boost 6 plus	0.79	0.79	0.79
boost 6 stand	0.79	0.79	0.79
boost 3 multi	0.79	0.79	0.79
boost 4 multi	0.83	0.84	0.83
boost 5 multi	0.83	0.84	0.83
boost 6 multi	0.79	0.79	0.79
RF	0.92	0.89	0.87

(b) Parâmetros **all**

Observando a Tabela 7.15 é possível verificar que os modelos treinados com os parâmetros **some** obtêm valores idênticos aos treinados com parâmetros **all** mas com um número menor de regras. Em ambos os conjuntos de parâmetros com o modelo **SIRUS else** obtêm-se métricas melhores que com a **RF**.

7.3.2.5 Centro Interior

Tabela 7.16: Centro Interior (CI allC7). Métricas de validação dos modelos baseados no algoritmo SIRUS.

Model	Precisão	Recall	F1
else 3	0.54	0.48	0.46
else 4	0.68	0.61	0.61
else 5	0.57	0.52	0.53
else 6	0.52	0.47	0.48
boost 3 plus	0.65	0.60	0.58
boost 3 stand	0.73	0.65	0.65
boost 4 plus	0.68	0.63	0.63
boost 4 stand	0.71	0.66	0.66
boost 5 plus	0.68	0.56	0.58
boost 5 stand	0.64	0.60	0.60
boost 6 plus	0.64	0.61	0.61
boost 6 stand	0.65	0.60	0.60
boost 3 multi	0.70	0.63	0.61
boost 4 multi	0.74	0.66	0.67
boost 5 multi	0.73	0.68	0.68
boost 6 multi	0.70	0.65	0.65
RF	0.88	0.85	0.86

(a) Parâmetros **some**

Model	Precisão	Recall	F1
else 3	0.75	0.73	0.72
else 4	0.73	0.71	0.71
else 5	0.75	0.74	0.74
else 6	0.79	0.76	0.76
boost 3 plus	0.67	0.63	0.63
boost 3 stand	0.68	0.65	0.65
boost 4 plus	0.67	0.61	0.62
boost 4 stand	0.70	0.66	0.67
boost 5 plus	0.70	0.68	0.68
boost 5 stand	0.76	0.73	0.73
boost 6 plus	0.72	0.71	0.71
boost 6 stand	0.70	0.65	0.65
boost 3 multi	0.78	0.74	0.75
boost 4 multi	0.80	0.76	0.77
boost 5 multi	0.77	0.74	0.75
boost 6 multi	0.80	0.77	0.78
RF	0.88	0.87	0.87

(b) Parâmetros **all**

Observando a Tabela 7.16 o modelo **else** obtêm melhores resultados com os parâmetros **all** com qualquer número de regras. Nos modelos **boost** os modelos com menor número de regras obtém valores semelhantes, mas apenas os modelos treinados com os parâmetros **all** obtêm valores acima de 0.70. São obtidos melhores resultados com o conjunto de parâmetros **all** mas este são bastante inferiores aos resultados da **RF**.

7.3.2.6 Sul

Tabela 7.17: Sul (S allC7). Métricas de validação dos modelos baseados no algoritmo SIRUS.

Model	Precisão	Recall	F1
else 3	0.90	0.86	0.86
else 4	0.89	0.81	0.81
else 5	0.87	0.76	0.77
else 6	0.87	0.81	0.81
boost 3 plus	0.93	0.90	0.90
boost 3 stand	0.87	0.86	0.86
boost 4 plus	0.91	0.90	0.90
boost 4 stand	0.93	0.90	0.90
boost 5 plus	0.90	0.90	0.90
boost 5 stand	0.93	0.90	0.90
boost 6 plus	0.87	0.86	0.86
boost 6 stand	0.87	0.81	0.81
boost 3 multi	0.84	0.81	0.81
boost 4 multi	0.92	0.90	0.90
boost 5 multi	0.96	0.95	0.95
boost 6 multi	0.93	0.90	0.90
RF	0.93	0.90	0.90

(a) Parâmetros **some**

Model	Precisão	Recall	F1
else 3	0.93	0.90	0.90
else 4	0.79	0.86	0.81
else 5	0.82	0.76	0.76
else 6	0.72	0.76	0.72
boost 3 plus	0.71	0.76	0.72
boost 3 stand	0.87	0.81	0.80
boost 4 plus	0.88	0.86	0.85
boost 4 stand	0.90	0.86	0.86
boost 5 plus	0.75	0.81	0.77
boost 5 stand	0.90	0.86	0.86
boost 6 plus	0.75	0.81	0.77
boost 6 stand	0.85	0.81	0.81
boost 3 multi	0.75	0.81	0.77
boost 4 multi	0.75	0.81	0.77
boost 5 multi	0.75	0.81	0.77
boost 6 multi	0.75	0.81	0.77
RF	0.96	0.95	0.95

(b) Parâmetros **all**

Observando a Tabela 7.17 o melhor modelo else em ambos os casos é o com 3 regras por nível de risco o que justificaria um estudo com modelos com menor número de regras. O F1 do melhor modelo para os dois conjuntos de parâmetros é 0.90 em ambos.

7.4 Comparação e avaliação de regras de indução, derivadas dos melhores modelos

7.4.1 Melhor modelo por região

Com o objetivo de escolher o melhor conjunto de regras para cada uma das regiões estudadas serão comparados os melhores modelos obtidos com os diferentes algoritmos e com os parâmetros **some** e **all** com o auxilio da métrica F1 da métrica de interpretabilidade. Inicialmente serão selecionados os conjuntos de regras que obtêm melhores valores de F1. Em seguida, dos modelos selecionados, é escolhido apenas um levando em consideração os valores da métrica de interpretabilidade e o valor de F1.

7.4.1.1 Portugal Continental

Tabela 7.18: Portugal Continental. Métricas de validação dos conjuntos de regras obtidos pelos diferentes métodos para o dataset PTallC7.
score: métrica de interpretabilidade, obtida com o cálculo da média das métricas de preditividade (pred), estabilidade(stab) e simplicidade(simp).

Parâmetros	Modelo	Precisão	Recall	F1	pred	stab	simp	score
some	DT 4	0.80	0.80	0.77	0.62	0.00	0.24	0.29
	RIPPER	0.84	0.84	0.84	0.69	0.06	0.17	0.30
	SIRUS else 3	0.71	0.72	0.69	0.46	0.24	1.00	0.57
	SIRUS boost 3 plus	0.73	0.73	0.72	0.48	0.24	0.67	0.46
	RF	0.89	0.89	0.88	—	—	—	—
all	DT 4	0.87	0.86	0.86	0.73	0.00	0.36	0.36
	RIPPER	0.90	0.89	0.89	0.79	0.06	0.22	0.36
	SIRUS else 4	0.83	0.83	0.82	0.67	0.29	0.63	0.53
	SIRUS boost 4 plus	0.81	0.80	0.80	0.62	0.25	0.48	0.45
	RF	0.92	0.91	0.91	—	—	—	—

Com os dados da Tabela 7.18, selecionou-se com base no valor de F1 os modelos **RIPPER** e **DT 4** com os parâmetros **all**. De entre estes escolheu-se o **RIPPER**, por ter um F1 maior com o mesmo valor de score.

7.4.1.2 Norte Litoral

Tabela 7.19: Norte Litoral. Métricas de validação dos conjuntos de regras obtidos pelos diferentes métodos para o dataset NLallC8.
score: métrica de interpretabilidade, obtida com o cálculo da média das métricas de preditividade (pred), estabilidade(stab) e simplicidade(simp).

Parâmetros	Modelo	Precisão	Recall	F1	pred	stab	simp	score
some	DT4	0.76	0.72	0.73	0.43	0.00	0.94	0.46
	RIPPER	0.76	0.75	0.75	0.50	0.24	1.00	0.58
	SIRUS else 6	0.79	0.74	0.76	0.46	0.15	0.53	0.38
	SIRUS boost 3 stand	0.74	0.67	0.66	0.32	0.23	0.94	0.50
	RF	0.87	0.86	0.86	—	—	—	—
all	DT 4	0.93	0.89	0.90	0.79	0.00	0.77	0.52
	RIPPER	0.88	0.88	0.87	0.75	0.19	0.94	0.63
	SIRUS else 6	0.83	0.82	0.82	0.64	0.08	0.49	0.40
	SIRUS boost 3 plus	0.80	0.77	0.78	0.54	0.30	0.94	0.59
	RF	0.92	0.91	0.91	—	—	—	—

Com os dados da Tabela 7.19, selecionou-se com base no valor de F1 os modelos **RIPPER** e **DT 4** com os parâmetros **all**. De entre estes escolheu-se o **RIPPER**, por ter um valor de score 0.11 valores maior com apenas 0.03 de F1 menor.

7.4.1.3 Norte Interior

Tabela 7.20: Norte Interior. Métricas de validação dos conjuntos de regras obtidos pelos diferentes métodos para o dataset NIsomeC5.
score: métrica de interpretabilidade, obtida com o cálculo da média das métricas de preditividade (pred), estabilidade(stab) e simplicidade(simp).

Parâmetros score	Modelo	Precisão	Recall	F1	pred	stab	inter	simp
some	DT 4	0.90	0.88	0.88	0.80	0.00	0.24	0.35
	RIPPER	0.91	0.88	0.88	0.80	0.17	1.00	0.66
	else 6	0.83	0.81	0.81	0.67	0.18	0.42	0.42
	boost 6 plus	0.84	0.85	0.84	0.73	0.23	0.33	0.43
	RF	0.93	0.92	0.91	—	—	—	—
all	DT 4	0.93	0.92	0.92	0.87	0.00	0.21	0.36
	RIPPER	0.88	0.85	0.84	0.73	0.20	1.00	0.64
	else 3	0.83	0.81	0.80	0.67	0.08	0.89	0.54
	boost 3 plus	0.84	0.85	0.84	0.73	0.18	0.67	0.53
	RF	0.93	0.92	0.91	—	—	—	—

Com os dados da Tabela 7.20, selecionou-se com base no valor de F1 os modelos **DT 4** e **RIPPER** com os parâmetros **some** e o modelo **DT 4** com os parâmetros **all**. De entre estes escolheu-se o **RIPPER** com parâmetros **some** devido ao maior valor de score.

7.4.1.4 Centro Litoral

Tabela 7.21: Centro Litoral. Métricas de validação dos conjuntos de regras obtidos pelos diferentes métodos para o dataset CLallC8.
score: métrica de interpretabilidade, obtida com o cálculo da média das métricas de preditividade (pred), estabilidade(stab) e simplicidade(simp).

Parâmetros	Modelo	Precisão	Recall	F1	pred	stab	simp	score
some	DT 2	0.79	0.79	0.79	0.69	0.00	1.00	0.56
	RIPPER	0.72	0.79	0.75	0.69	0.25	0.67	0.54
	else 3	0.89	0.89	0.89	0.85	0.08	0.86	0.59
	boost 3 plus	0.84	0.84	0.84	0.77	0.14	0.50	0.47
	RF	0.92	0.89	0.87	—	—	—	—
all	DT 7	0.83	0.84	0.83	0.77	0.00	0.33	0.37
	RIPPER	0.66	0.74	0.69	0.62	0.17	0.60	0.46
	else 4	0.89	0.89	0.89	0.85	0.03	0.50	0.46
	boost 4 plus	0.83	0.84	0.83	0.77	0.16	0.38	0.43
	RF	0.92	0.89	0.87	—	—	—	—

Com os dados da Tabela 7.21, selecionou-se com base no valor de F1 o modelo **else 3** com parâmetros some e o modelo **else 4** com parâmetros **all**. De entre estes escolheu-se o **else 3** com parâmetros **some** devido ao maior valor de score.

7.4.1.5 Centro Interior

Tabela 7.22: Centro Interior. Métricas de validação dos conjuntos de regras obtidos pelos diferentes métodos para o dataset CIallC8.
score: métrica de interpretabilidade, obtida com o cálculo da média das métricas de preditividade (pred), estabilidade(stab) e simplicidade(simp).

Parâmetros	Modelo	Precisão	Recall	F1	pred	stab	simp	score
some	DT 4	0.81	0.79	0.79	0.73	0.00	0.30	0.34
	RIPPER	0.83	0.81	0.81	0.75	0.00	0.44	0.40
	SIRUS else 4	0.68	0.61	0.61	0.50	0.14	1.00	0.55
	SIRUS boost 3 stand	0.73	0.65	0.65	0.54	0.22	0.93	0.57
	RF	0.88	0.85	0.86	—	—	—	—
all	DT 4	0.79	0.76	0.76	0.69	0.00	0.44	0.38
	RIPPER	0.83	0.82	0.82	0.77	0.00	0.58	0.45
	SIRUS else 6	0.79	0.76	0.76	0.69	0.09	0.67	0.48
	SIRUS boost 5 stand	0.76	0.73	0.73	0.65	0.22	0.50	0.46
	RF	0.88	0.87	0.87	—	—	—	—

Com os dados da Tabela 7.22, selecionou-se com base no valor de F1 os modelos **DT 4** e **RIPPER** com os parâmetros **some** e o modelo **RIPPER** com os parâmetros **all**. De entre estes escolheu-se o **RIPPER** com parâmetros **all** devido ao maior valor de score.

7.4.1.6 Sul

Tabela 7.23: Sul. Métricas de validação dos conjuntos de regras obtidos pelos diferentes métodos para o dataset SallC7.
score: métrica de interpretabilidade, obtida com o cálculo da média das métricas de preditividade (pred), estabilidade(stab) e simplicidade(simp).

Parâmetros	Modelo	Precisão	Recall	F1	pred	stab	simp	score
some	DT 3	0.87	0.86	0.86	0.80	0.00	0.82	0.54
	RIPPER	0.81	0.86	0.82	0.80	0.15	0.90	0.62
	SIRUS else 3	0.90	0.86	0.86	0.80	0.13	0.90	0.61
	SIRUS boost 3 plus	0.93	0.90	0.90	0.87	0.22	0.60	0.56
	RF	0.93	0.90	0.90	—	—	—	—
all	DT 4	0.90	0.86	0.86	0.80	0.00	0.50	0.43
	RIPPER	0.73	0.71	0.70	0.60	0.13	1.00	0.58
	SIRUS else 3	0.93	0.90	0.90	0.87	0.19	0.75	0.60
	SIRUS boost 4 stand	0.90	0.86	0.86	0.80	0.20	0.45	0.48
	RF	0.96	0.95	0.95	—	—	—	—

Com os dados da Tabela 7.23, selecionou-se com base no valor de F1 o modelo **boost 3 plus** com parâmetros some e o modelo **else 3** com parâmetros **all**. De entre estes escolheu-se o **else 3** com parâmetros **all** devido ao maior valor de score.

7.4.2 Análise das regras

7.4.2.1 Portugal Continental

Tabela 7.24: Portugal Continental. Regras obtidas com o algoritmo RIPPER.
Suporte: (nº casos que tornam a condição verdadeira) / (nº de casos que tornam a condição verdadeira com classificação diferente)

Risco	**Regras**	**Suporte**
Baixo	FWI ≤ 14.86 *e* DC ≤ 375.63	131/1
	FWI ≤ 3.38 *e* BUI ≤ 77.75	30/1
	DC ≤ 137.41 *e* FWI ≤ 20.28	7/1
	FWI ≤ 11.76 *e* DC ≤ 613.56 *e* 2.35 ≤ CHI ≤ 6.73	9/1
	Se não:	
Extremo	FWI ≥ 42.91 *e* BUI ≤ 236.28 *e* CHI ≥ 5.3	172/0
	FWI ≥ 41.16 *e* CHI ≥ 9.44 *e* DC ≤ 802.01	31/5
	ISI ≥ 14.45 *e* FWI ≥ 54.24 *e* BUI ≤ 327.63	10/2
	Se não:	
Muito alto	BUI ≥ 214.88 *e* FWI ≥ 32.63 *e* DMC ≥ 216.12	115/1
	CHI ≥ 6.84 *e* BUI ≥ 226.02	46/0
	BUI ≥ 202.58 *e* ISI ≥ 10.51	29/6
	BUI ≥ 295.15	17/2
	BUI ≥ 208.18 *e* FWI ≥ 32.98 *e* CHI ≥ 5.31	19/6
	Se não:	
Moderado	CHI ≤ 4.52 *e* FWI ≤ 28.22 *e* BUI ≤ 171.54	178/1
	CHI ≤ 5.06 *e* FWI ≤ 22.63 *e* DMC ≤ 163.55 *e* DC ≥ 457.66	17/0
	CHI ≤ 2.07 *e* FWI ≤ 32.49 *e* BUI ≤ 178.28	9/0
	ISI ≤ 3.54 *e* CHI ≤ 6.25	5/0
	CHI ≤ 4.96 *e* 379.35 ≤ DC ≤ 480.11 *e* FWI ≤ 29.11	6/0
	CHI ≤ 0.0	3/0
	Se não:	
Alto		818/16

Comparação das regras de maior suporte com a escala de referência. Numeração da regra em concordância com a ordem por nível de risco apresentada na Tabela 7.24.

- Baixo

  1. FWI com valor Baixo. DC no limite entre Baixo e Moderado.
  2. FWI abaixo do Muito baixo. BUI no limite Moderado / Alto.

- Extremo

1. FWI com valor Muito alto (extremo ≥ 50). BUI com valor bastante acima do Muito alto (> 178.1).
2. FWI e DC com valor Muito alto. CHI bastante elevado.

- Muito alto

1. BUI e DMC com valor bastante acima do Muito alto. FWI com valor Alto.
2. BUI bastante acima do Muito alto com valor alto de CHI.

- Moderado

1. FWI e BUI perto do limite Moderado / Alto

No geral as regras estão em concordância com a escala de referência. Algumas regras tem índices com valores abaixo ou acima do esperado mas normalmente a regra contém ambos os casos "balanceando-se", como por exemplo a regra de risco Baixo: "FWI ≤ 3.38 *e* BUI ≤ 77.75 ", é composta por um valor de BUI classificado como Alto mas um valor de FWI muito abaixo do limite do Muito baixo (< 5.2). Existe um elevado número de regras com suporte baixo, cinco regras em dezanove com suporte menor a dez.

7.4.2.2 Norte Litoral

Tabela 7.25: Norte Litoral. Regras obtidas com o algoritmo RIPPER.
Suporte: (nº casos que tornam a condição verdadeira) / (nº de casos que tornam a condição verdadeira com classificação diferente)

Risco	**Regras**	**Suporte**
Extremo	FWI ≥ 47.45	42/3
	Se não:	
Muito baixo	FFMC ≤ 84.92 *e* DC ≤ 298.43 *e* CHI ≤ 5.26 FWI ≤ 10.79 *e* CHI ≤ 3.65 *e* DC ≤ 514.39	33/1 8/1
	Se não:	
Baixo	DC ≤ 152.03	50/2
	Se não:	
Muito alto	BUI ≥ 145.39 *e* FWI ≥ 32.57 *e* CHI ≥ 5.18 BUI ≥ 178.91 CHI ≥ 9.26 *e* BUI ≥ 135.41	62/1 9/1 9/2
	Se não:	
Moderado	DMC ≤ 58.08 *e* 4.56 ≤ CHI ≤ 8.14 FWI ≤ 21.54 *e* CHI ≥ 5.42	51/3 29/6
	Se não:	
Alto		215/12

Comparação das regras de maior suporte com a escala de referência. Numeração da regra em concordância com a ordem por nível de risco apresentada na Tabela 7.25.

- Extremo
    1. FWI ≥ 47.45, próximo do valor de referência (≥ 50)
- Muito baixo
    1. FFMC e DC ligeiramente acima dos limites Muito baixo / Baixo
- Baixo
    1. DC ≤ 152.03, muito inferior ao limite do Muito baixo (< 256.1). Valor inesperado, adicionado ao facto de estar a ser utilizado para separar incêndios Baixo de incêndios de risco superior (visto que as regras anteriores classificaram os incêndios Muito baixo).
- Muito alto

1. BUI e FWI com valor Alto mas mas o valor do CHI pode justificar a classificação

- Moderado

1. DMC perto do limite Moderado / Alto
2. FWI perto do limite Moderado / Alto

Com exceção das classes Baixo e Muito baixo os valores encontrados são próximos dos valores limiares correspondentes na escala de referência.

7.4.2.3 Norte Interior

Tabela 7.26: Norte Interior. Regras obtidas com o algoritmo RIPPER.
Suporte: (nº casos que tornam a condição verdadeira) / (nº de casos que tornam a condição verdadeira com classificação diferente)

Risco	**Regras**	**Suporte**
Baixo	DC $\leq$ 390.45 *e* FWI $\leq$ 23.09	18/0
	Se não:	
Muito alto	ISI $\geq$ 10.09 *e* CHI $\geq$ 6.56	30/0
	FWI $\geq$ 35.41 *e* CHI $\geq$ 6.3	8/0
	Se não:	
Moderado	CHI $\leq$ 5.22 *e* FWI $\leq$ 27.42	54/4
	Se não:	
Alto		124/7

Comparação das regras de maior suporte com a escala de referência. Numeração da regra em concordância com a ordem por nível de risco apresentada na Tabela 7.26.

- Baixo

1. DC com valor Moderado. FWI perto do limite Moderado / Alto.

- Muito alto

1. ISI com valor Alto, abaixo do limite Alto / Muito alto (13.4), mas CHI pode justificar a classificação
2. FWI perto do limite Alto / Muito alto.

- Moderado

1. FWI perto do limite Moderado / Alto

Com exceção da Classe Baixo, os valores encontrados estão em concordância com a escala de referência.

7.4.2.4 Centro Litoral

Tabela 7.27: Centro Litoral. Regras obtidas com o modelo SIRUS else.

Risco	**Regras**
Extremo	Se probabilidade ≥ 0.334 FFMC < 94.494 then 0.057692 (n=104) else 0.66667 (n=12) FWI < 51.542 then 0.028846 (n=104) else 0.91667 (n=12) ISI < 15.956 then 0.028846 (n=104) else 0.91667 (n=12) Se não:
Muito alto	Se probabilidade ≥ 0.443 FFMC < 92.717 then 0.25352 (n=71) else 0.90323 (n=31) FWI < 37.099 then 0.13115 (n=61) else 0.92683 (n=41) ISI < 10.772 then 0.22535 (n=71) else 0.96774 (n=31) Se não:
Moderado	Se probabilidade ≥ 0.438 CHI < 3.3931 then 0.81818 (n=11) else 0.088889 (n=45) FFMC < 86.823 then 0.70588 (n=17) else 0.025641 (n=39) ISI < 3.1245 then 1 (n=6) else 0.14 (n=50) Se não:
Alto	

(a) Original

Risco	**Regras**
Extremo	FWI ≥ 51.54 *e* ISI ≥ 16.96 Se não:
Muito alto	FWI ≥ 37.1 e ISI ≥ 10.77 Se não:
Moderado	**COUNT**(CHI < 3.39; FFMC < 86.82; ISI < 3.12) ≥ 2 Se não:
Alto	

(b) Simplificado.

COUNT( ): interpreta um conjunto de proposições e devolve o número de verdadeiros.

Comparação das regras de maior suporte com a escala de referência. Numeração da regra em concordância com a ordem por nível de risco apresentada na Tabela 7.27.

- Extremo

1. FWI perto do limite do Extremo. ISI bastante acima do limite do Muito alto.

- Muito alto

1. FWI perto do limite Alto / Muito Alto. ISI com valor Alto

- Moderado

1. FFMC perto do limite Baixo / Moderado. ISI perto do limite Muito baixo / Baixo.

A classificação com Extremo e Muito alto é concordante com a escala de referência. As regras associadas ao risco Moderado são abaixo do que seria esperado. A classe Moderada é a classe de risco mais baixo nesta classificação, na prática estas regras estão apenas a dividir os incêndios de risco Alto dos de risco inferior.

7.4.2.5 Centro Interior

Tabela 7.28: Centro Interior. Regras obtidas com o algoritmo RIPPER.
Suporte: (nº casos que tornam a condição verdadeira) / (nº de casos que tornam a condição verdadeira com classificação diferente)

Risco	**Regras**	**Suporte**
Baixo	FFMC $\leq$ 84.81	46/7
	Se não:	
Extremo	BUI $\geq$ 219.72 *e* FFMC $\geq$ 92.84	58/6
	CHI $\geq$ 6.84 *e* BUI $\geq$ 220.81	17/2
	FWI $\geq$ 55.59 *e* CHI $\leq$ 8.73	7/0
	Se não:	
Moderado	FFMC $\leq$ 90.27 *e* CHI $\leq$ 4.69 *e* BUI $\leq$ 185.49	50/3
	FWI $\leq$ 26.28 *e* CHI $\leq$ 6.75 *e* BUI $\leq$ 121.48	21/2
	CHI $\leq$ 6.14 *e* DC $\leq$ 578.22 *e* BUI $\leq$ 173.36	12/2
	Se não:	
Alto	FWI $\leq$ 35.58 *e* BUI $\leq$ 186.32	143/3
	FWI $\leq$ 31.69 *e* CHI $\geq$ 5.74	15/3
	BUI $\leq$ 146.82 *e* FWI $\leq$ 41.03 *e* DC $\leq$ 763.30 *e* FFMC $\leq$ 93.94	10/1
	Se não:	
Muito alto		178/15

Comparação das regras de maior suporte com a escala de referência. Numeração da regra em concordância com a ordem por nível de risco apresentada na Tabela 7.28.

- Baixo

1. FFMC Baixo, abaixo do limite Baixo / Moderado

- Extremo

  1. BUI bastante acima do limite do Muito alto. FFMC perto do limite do Muito alto.
  2. Bui bastante acima do limite do Muito alto e CHI alto.

- Moderado

  1. FFMC perto do limite Moderado / Alto. BUI acima do esperado, acima do limite Alto / Muito alto.
  2. FWI perto do limite Moderado / Alto. BUI com valor Alto.

- Alto

  1. FWI e BUI perto do limite Alto / Muito alto.

A classificação parece estar concordante com a escala de referência. No geral os valores do índice BUI são acima do esperado.

7.4.2.6 Sul

Tabela 7.29: Sul. Regras obtidas com o modelo SIRUS else.

Risco	**Regras**
Muito baixo	Se probabilidade ≥ 0.622 FFMC < 60.575 then 1 (n=14) else 0.024793 (n=121) ISI < 0.95086 then 0.92857 (n=14) else 0.033058 (n=121) FWI < 2.9755 then 0.92857 (n=14) else 0.033058 (n=121) Se não:
Baixo	Se probabilidade ≥ 0.357 BUI < 132.97 then 0.70833 (n=24) else 0 (n=94) ISI < 4.5573 then 0.70833 (n=24) else 0 (n=94) FFMC < 87.247 then 0.70833 (n=24) else 0 (n=94) Se não:
Extremo	Se probabilidade ≥ 0.517 ISI < 15.599 then 0.025 (n=80) else 0.7619 (n=21) FWI < 52.901 then 0.025 (n=80) else 0.7619 (n=21) CHI < 9.8718 then 0.028571 (n=70) else 0.51613 (n=31) Se não:
Alto	Se probabilidade ≥ 0.367 CHI < 5.4084 then 0.88 (n=25) else 0.17241 (n=58) ISI < 7.8859 then 0.78788 (n=33) else 0.12 (n=50) FWI < 33.448 then 0.78788 (n=33) else 0.12 (n=50) Se não:
Muito alto	

(a) Original

Risco	**Regras**
Muito baixo	**COUNT**(FFMC < 60.58 ; ISI < 0.95 ; FWI < 2.98) ≥ 2 Se não:
Baixo	**COUNT**(BUI < 132.97 ; ISI < 4.56 ; FFMC < 87.25) ≥ 2 Se não:
Extremo	ISI ≥ 15.6 *e* FWI ≥ 52.9 *e* CHI ≥ 9.87 Se não:
Alto	CHI < 5.41 *ou* (ISI < 7.89 *e* FWI < 33.45) Se não:
Muito alto	

(b) Simplificado.

COUNT( ): interpreta um conjunto de proposições e devolve o número de verdadeiros.

Comparação das regras de maior suporte com a escala de referência. Numeração da regra em concordância com a ordem por nível de risco apresentada na Tabela 7.29.

- Muito baixo
    1. FFMC, ISI e FWI bastante abaixo do limite Muito baixo / Baixo.
- Baixo
    1. BUI com valor Alto. ISI com valor Baixo. FFCM com valor Moderado
- Extremo
    1. ISI acima do limite do Muito alto. FWI acima do limite do Extremo. CHI muito alto.
- Alto
    1. ISI perto do limite Moderado / Alto. FWI perto do limite Alto / Muito alto

As regras das classes de risco Baixo e Muito baixo não são muito concordantes com a escala de referência. Os índices da regra com risco Extremo são próximos dos limites na escala de referência. Na divisão dos incêndios entre Alto e Muito alto o FWI é concordante mas os restantes índices não. No entanto os equívocos nas regras de risco Alto e Muito alto podem ser explicados pela falta de classe Moderado no conjunto de regras (devido ao reduzido número de casos no conjunto de dados).

7.5 Discussão dos resultados

Os modelos selecionados foram treinados com os dados de 2001 a 2018. Os dados com os incêndios ocorridos em 2019 e 2020 foram separados para testar a capacidade preditiva das regras para dados "novos". As métricas de classificação dos modelos para estes dados podem ser encontrados na Tabela 7.30.

Tabela 7.30: Características e métricas de validação dos melhores modelos de cada região, com os incêndios de 2001 a 2018 e com os incêndios de 2019 a 2020. Comparação do F1 do modelo selecionado com o F1 obtido com a floresta aleatória para as mesmas condições.

	FCM		Regras		2001-2018		2019 - 2020	
	Parâmetros	Nº clusters	Parâmetros	Modelo	F1	F1 com RF	F1	F1 com RF
PT	all	7	all	RIPPER	0.89	0.91	0.85	0.87
NL	all	8	all	RIPPER	0.87	0.91	0.78	0.91
NI	some	5	some	RIPPER	0.88	0.91	0.85	0.83
CL	all	8	some	SIRUS else 3	0.89	0.87	0.48	0.72
CI	all	7	all	RIPPER	0.82	0.87	0.85	0.83
S	all	7	all	SIRUS else 3	0.90	0.95	0.75	0.84

Começando pelos resultados do agrupamento difuso, verificou-se que para 5 das 6 regiões a melhor partição foi a obtida com o conjuntos e parâmetros all. As partições selecionadas têm para todos os casos 7 ou 8 nº de *clusters*, sendo consistente com a escala de referência com 7 classes de risco.

Em relação à extração de regras, pelos dados da Tabela 7.30, é possível observar que para quatro das seis regiões o melhor conjunto de regras é obtido com o modelo RIPPER, três destas com o conjunto de parâmetros all. Nas quatro regiões com o modelo RIPPER as métricas com os incêndios de 2019 a 2020 são ligeiramente piores, com exceção do Centro Interior. Este não é um resultado necessariamente negativo sendo esperado visto que o treino e a seleção dos modelos foi feito com dados apenas de 2001 a 2019.

Em relação as restantes duas regiões, ambas com o melhor conjunto de regras obtido com o algoritmo SIRUS os valores obtidos com o conjunto de dados de 2019 e 2020 é significativamente inferior, No caso do modelo do Centro Litoral, esta diferença é de 0.41 valores. Uma explicação plausível para este comportamento é o modelo SIRUS else estar a sobreajustar-se ao conjunto de treino.

O modelo RIPPER treinado com os parâmetros all, aparenta ser o melhor modelo para este conjunto de dados.

O F1 médio das cinco regiões geográficas com os dados de 2019-2020 é de 0. 742. Por termos concluído que o conjunto de regras associado à região Centro Litoral está a sobreajustar-se, e por ser inutilizável, calculando o F1 médio das restantes quatro regiões geográficas obtemos um valor de 0.808, um valor bom para dados reais. Com as regras sem separação geográfica (Portugal Continental) obtém-se um F1 de 0.85, bastante próximo do valor de 0.87 obtido com floresta aleatória mas com a possível interpretação do modelo.

8

Conclusão e trabalho Futuro

Este trabalho consistiu num estudo comparativo de diferentes algoritmos de classificação baseados em árvores de decisão com extração de regras para a definição das classes de risco de incêndio florestal com os principais índices de risco de incêndio e com a análise conjunto do FWI e sub-índices com o CHI.

Para tal, foi necessário criar os conjuntos de dados de estudo concatenando as informações de data e localização dos incêndios dos anos de 2001 a 2020 com as medições diárias dos índices.

As ocorrências de incêndios foram classificadas de forma não supervisionada recorrendo ao algoritmo de agrupamento difuso *fuzzy c-means*. Os resultados obtidos com este algoritmo foram validados com o índice de validação interno Xie e Beni e com a visualização gráfica com o método de visualização *fuzzy sammon mapping*. As segmentações obtidas com este algoritmo foram positivas, obtendo-se partições com um número de *clusters* igual a 7 ou 8, igual ou próximo à escala de referência com 7 classes de risco. Os protótipos dos clusters, enquanto "pontos representativos" dos clusters são nos valores dos seus índices concordantes com a escala de risco de referência.

Com os dados etiquetados foram, em seguida treinados diferentes modelos de classificação e extraídas as suas regras. Das seis regiões geográficas (Portugal Continental, Norte Litoral, Norte Interior, Centro Litoral, Centro Interior e Sul) o melhor conjunto de regras encontrado está associado a diferentes modelos, quatro com o modelo RIPPER e duas com o modelo SIRUS. Os modelos SIRUS parecem estar a sobreajustar-se devido a grande diferença entre métricas (precisão, *recall* e F1) obtida com o conjunto de dados de treino e com o conjunto de dados de teste. O modelo RIPPER aparenta ser o melhor de entre os três modelos comparados, obtendo para os dados de Portugal Continental um valor de F1 de 0.85 quando para os mesmos dados uma floresta aleatória obtém 0.86. De entre os conjuntos de regras das seis regiões estudadas três obtêm um valor de F1 igual a 0.85, um valor cosiderado muito bom para dados reais. Dos restantes conjuntos de regras dois obtêm um valor maior ou igual a 0.75, valores estes satisfatórios. Apenas para uma das regiões os resultados são negativos, aparentando haver sobreajuste do modelo selecionado.

Os índices das regras associadas aos níveis de risco estão, no geral em concordância com o conhecimento do comportamento dos mesmos. Observando-se uma maior concordância e uma melhor classificação nas classes de risco de maior severidade. Uma justificação possível para a ligeira incerteza nas classes de menor risco é o limite de 100 hectares de área ardida mínimo nos dados utilizados.

Quanto às medidas de avaliação das regras, estas podem não ter sido a métrica mais indicada para comparar os diferentes tipos de modelos. Na medida de preditividade o modelo é comparado com um modelo *naïve* (foi utilizada a moda dos dados). Esta comparação é feita a partir do número de classificações incorretas. Esta métrica vai, como consequência, favorecer modelos que classificam corretamente um maior número de casos com o *label* mais comum, em prol de um modelo que faça uma correta divisão de classes com um número mais reduzido de elementos. Para além disto, o facto de a escala de risco se tratar de uma escala ordinal, faz com que um incêndio, por exemplo, Baixo a ser classificado com Extremo esteja "mais errado" do que se esse incêndio classificado como Moderado. Outro problema com as medidas é em relação a medida de simplicidade, por esta medida utilizar a soma do comprimento de todas as regras. O problema surge de os quatro modelos comparados terem conjuntos de regras com diferentes regras de leitura. As regras do **SIRUS else** e do **RIPPER** são encadeadas, parando a execução quando uma regra é verdadeira, mas sendo necessário ter em conta a negação de todas as condições falsas anteriores. No **SIRUS boost** é necessário aplicar todas as regras, obtendo um valor para cada nível de risco, escolhendo finalmente o nível com maior valor. Com as árvores de decisão são obtidas regras disjuntas, onde um dado é apenas verdadeiro para uma regra e toma o nível de risco associado a esta. As árvores de decisão possuem ainda a vantagem de poderem ser representadas num modelo gráfico.

Quanto ao trabalho futuro na progressão deste estudo, existem dois aspetos passíveis de aprofundar:

- Estudo experimental mais aprofundado com diferentes combinações de índices. Numa fase inicial do projeto foi descartada a ideia de usar os índices de vegetação e o índice de perigosidade devido aos resultados das partições não serem satisfatórios. Há a possibilidade de existir uma série de índices capaz de obter melhores resultados. Este estudo poderia ser feito com a comparação das partições obtidas com diferentes combinações de índices ou com a análise dos dados com o intuito de encontrar os índices mais informativos.

- Estudo com múltiplos índices de validação interna

- Estudo comparativo do fuzzy c-means com outros métodos de agrupamento (e.x.: *archetypal analysis*)

- Estudo mais aprofundado com o algoritmo RIPPER. O algoritmo RIPPER apresentou melhor resultados, no geral, mas em alguns modelos gerou um número elevado de regras com suporte baixo.

Bibliografia

[1] J. Abonyi e B. Feil. *Cluster analysis for data mining and system identification*. Springer Science & Business Media, 2007 (ver pp. 14, 37).

[2] C. Bénard et al. "Sirus: Stable and interpretable rule set for classification". Em: *Electronic Journal of Statistics* 15.1 (2021), pp. 427–505 (ver pp. 3, 23, 47).

[3] L. Breiman. "Bagging predictors". Em: *Machine learning* 24.2 (1996), pp. 123–140 (ver p. 23).

[4] L. Breiman. "Random forests". Em: *Machine learning* 45.1 (2001), pp. 5–32 (ver p. 22).

[5] C. A. Brunk e M. J. Pazzani. "An investigation of noise-tolerant relational concept learning algorithms". Em: *Machine Learning Proceedings 1991*. Elsevier, 1991, pp. 389–393 (ver p. 21).

[6] L. Bugalho. *Caracterização Climática do Índice de Haines e do Índice de Haines Contínuo em Portugal no período 2011 a 2017. Nota Técnica DivMV, 3/2018*. 2018 (ver p. 6).

[7] L. Bugalho. *Temporal variability of the Haines index and its relationship with forest fire in Portugal. In Advances in Forest Fire Research 2018, Viegas, D. X. (Ed.)* 2018 (ver pp. 6, 7).

[8] H. Coelho. "Identificação automática de valores de índices de risco de incêndios florestais aos níveis de risco de incêndio extremo florestal em Portugal Continental". Tese de mestrado. 2021 (ver pp. 2, 8, 20, 26, 43).

[9] W. W. Cohen. "Fast effective rule induction". Em: *Machine learning proceedings 1995*. Elsevier, 1995, pp. 115–123 (ver pp. 3, 21).

[10] Copernicus Climate Change Service. *Fire burned area from 2001 to present derived from satellite observations*. 2019. DOI: 10.24381/CDS.F333CF85. URL: https://cds.climate.copernicus.eu/doi/10.24381/cds.f333cf85 (ver p. 11).

[11] J. M. Corcoran, J. F. Knight e A. L. Gallant. "Influence of multi-source and multi-temporal remotely sensed and ancillary data on the accuracy of random forest classification of wetlands in Northern Minnesota". Em: *Remote Sensing* 5.7 (2013), pp. 3212–3238 (ver p. 23).

[12] W. J. De Groot et al. "Interpreting the Canadian forest fire weather index (FWI) system". Em: *Proc. of the Fourth Central Region Fire Weather Committee Scientific and Technical Seminar*. 1998 (ver p. 8).

[13] H. Drucker et al. "Boosting and other ensemble methods". Em: *Neural Computation* 6.6 (1994), pp. 1289–1301 (ver p. 49).

[14] EFFIS. *Fire danger forecast*. 2022. URL: https://effis.jrc.ec.europa.eu/about-effis/technical-background/fire-danger-forecast (acedido em 26/09/2022) (ver p. 1).

[15] EFFIS. *Fire danger forecast*. 2022. URL: https://effis.jrc.ec.europa.eu/about-effis/technical-background/fire-danger-forecast (acedido em 19/12/2022) (ver p. 9).

[16] P. Fernandes. "Estudo de adaptação para Portugal do Sistema Canadiano de Indexação do Perigo de Incêndio". Em: *Universidade de Trás-os-Montes e Alto Douro* (2005) (ver p. 9).

[17] P. M. Fernandes et al. "Bottom-up variables govern large-fire size in Portugal". Em: *Ecosystems* 19.8 (2016), pp. 1362–1375 (ver p. 5).

[18] Y. Freund e R. E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". Em: *Journal of computer and system sciences* 55.1 (1997), pp. 119–139 (ver p. 50).

[19] J. Fürnkranz e G. Widmer. "Incremental reduced error pruning". Em: *Machine Learning Proceedings 1994*. Elsevier, 1994, pp. 70–77 (ver p. 21).

[20] L. Giglio et al. "Collection 6 modis burned area product user's guide version 1.0". Em: *NASA EOSDIS Land Processes DAAC: Sioux Falls, SD, USA* (2016) (ver pp. 11, 27, 28).

[21] D. A. Haines. "A lower atmosphere severity index for wildlife fires". Em: *National Weather Digest* 13 (1988), pp. 23–27 (ver p. 6).

[22] J. Huysmans, B. Baesens e J. Vanthienen. "Using rule extraction to improve the comprehensibility of predictive models". Em: (2006) (ver p. 18).

[23] ICNF. *Perigosidade de Incêndio Rural*. 2021. URL: http://www2.icnf.pt/portal/florestas/dfci/inc/cartografia/perigosidade (ver pp. 10, 11).

[24] L. Iliadis, M. Vangeloudh e S. Spartalis. "An intelligent system employing an enhanced fuzzy c-means clustering model: Application in the case of forest fires". Em: *Computers and Electronics in Agriculture* 70.2 (2010), pp. 276–284 (ver p. 14).

[25] IPMA. *Dataservices: Perigosidade meteorológica de fogos florestais*. 2022. URL: http://mf2.ipma.pt/about (acedido em 26/09/2022) (ver p. 1).

[26] B. Krishnapuram et al. "Sparse multinomial logistic regression: Fast algorithms and generalization bounds". Em: *IEEE transactions on pattern analysis and machine intelligence* 27.6 (2005), pp. 957–968 (ver p. 50).

[27] B. Li et al. "Classification and regression trees (CART)". Em: *Biometrics* 40.3 (1984), pp. 358–361 (ver pp. 3, 18).

[28] W.-Y. Loh. "Classification and regression trees". Em: *Wiley interdisciplinary reviews: data mining and knowledge discovery* 1.1 (2011), pp. 14–23 (ver p. 19).

[29] M. Long. "A climatology of extreme fire weather days in Victoria". Em: *Australian Meteorological Magazine* 55.1 (2006), pp. 3–18 (ver p. 6).

[30] V. Margot e G. Luta. "A new method to compare the interpretability of rule-based algorithms". Em: *AI* 2.4 (2021), pp. 621–635 (ver pp. 3, 21, 24).

[31] V. Margot et al. "Consistent regression using data-dependent coverings". Em: *Electronic Journal of Statistics* 15.1 (2021), pp. 1743–1782 (ver p. 25).

[32] L. McCaw et al. "Bushfire weather climatology of the Haines Index in southwestern Australia". Em: *Australian Meteorological Magazine* 56.2 (2007) (ver p. 6).

[33] R. Messenger e L. Mandell. "A modal search technique for predictive nominal scale multivariate analysis". Em: *Journal of the American statistical association* 67.340 (1972), pp. 768–772 (ver p. 18).

[34] G. A. Mills e W. L. McCaw. "Atmospheric stability environments and fire weather in Australia: extending the Haines Index". Em: (2010) (ver p. 6).

[35] S. Nascimento e N. Madaleno. "Unsupervised Initialization of Archetypal Analysis and Proportional Membership Fuzzy Clustering". Em: *Intelligent Data Engineering and Automated Learning – IDEAL 2019*. Ed. por H. Yin et al. Cham: Springer International Publishing, 2019, pp. 12–20. ISBN: 978-3-030-33617-2 (ver p. 37).

[36] J. Nayak, B. Naik e H. Behera. "Fuzzy C-means (FCM) clustering algorithm: a decade review from 2000 to 2014". Em: *Computational intelligence in data mining-volume 2* (2015), pp. 133–149 (ver p. 14).

[37] D. Petkovic et al. "Improving the explainability of Random Forest classifier–user centered approach". Em: *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium*. World Scientific. 2018, pp. 204–215 (ver p. 23).

[38] B. E. Potter et al. "Computing the low-elevation variant of the Haines index for fire weather forecasts". Em: *Weather and Forecasting* 23.1 (2008), pp. 159–167 (ver p. 6).

[39] R. Rifkin e A. Klautau. "In defense of one-vs-all classification". Em: *The Journal of Machine Learning Research* 5 (2004), pp. 101–141 (ver p. 47).

[40] SRTM. *SRTM 90m DEM Digital Elevation Database*. 2018. URL: https://srtm.csi.cgiar.org/ (ver p. 12).

[41] SRTM. *U.S. Releases Enhanced Shuttle Land Elevation Data*. URL: https://www2.jpl.nasa.gov/srtm/ (ver p. 12).

[42] A. L. Sullivan e J. S. Gould. "Wildland Fire Rate of Spread". Em: *Encyclopedia of Wildfires and Wildland-Urban Interface (WUI) Fires*. Ed. por S. L. Manzello. Cham: Springer International Publishing, 2018, pp. 1–4. ISBN: 978-3-319-51727-8. DOI: 10.1007/978-3-319-51727-8_55-1. URL: https://doi.org/10.1007/978-3-319-51727-8_55-1 (ver p. 5).

[43] F. Tedim et al. "Defining extreme wildfire events: difficulties, challenges, and impacts". Em: *Fire* 1.1 (2018), p. 9 (ver p. 5).

[44] S. Tian et al. "Random forest classification of wetland landcovers from multi-sensor data in the arid region of Xinjiang, China". Em: *Remote Sensing* 8.11 (2016), p. 954 (ver p. 23).

[45] J. A. Turner e B. D. Lawson. "Weather in the canadian forest fire danger rating system. a user guide to national standards and practices". Em: (1978) (ver p. 7).

[46] S. Van Beijma, A. Comber e A. Lamb. "Random forest classification of salt marsh vegetation habitats using quad-polarimetric airborne SAR, elevation and optical RS data". Em: *Remote Sensing of Environment* 149 (2014), pp. 118–129 (ver p. 23).

[47] D. X. Viegas et al. "Calibração do sistema canadiano de perigo de incêndio para aplicação em Portugal". Em: *Silva Lusitana* 12.1 (2004), pp. 77–93 (ver p. 9).

[48] J. Werth e P. Werth. "Haines Index climatology for the western United States". Em: *Fire management notes (USA)* (1998) (ver p. 6).

[49] X. L. Xie e G. Beni. "A validity measure for fuzzy clustering". Em: *IEEE Transactions on pattern analysis and machine intelligence* 13.8 (1991), pp. 841–847 (ver pp. 16, 37).