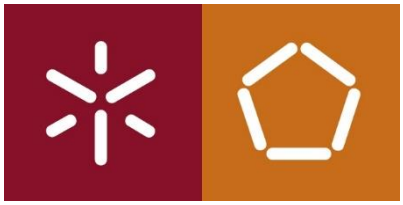

Universidade do Minho
Escola de Engenharia
Mestrado em Engenharia Informática

Aprendizagem Automática 1

Ano Letivo de 2020/2021

Introdução à Aprendizagem Estatística

Parque Natural de Montesinho – Incêndios Florestais

Hugo Miguel Ramos Alves de Faria – pg44415
João Pedro Matos Ribeiro Soares – pg42836
Simão Pedro Martins Gonçalves – pg42850

Dezembro, 2020

Índice

1. Resumo Executivo

O nosso trabalho incidiu na análise de um *dataset* sobre incêndios do Parque Natural de Montesinhos. A escolha deste tema deveu-se, principalmente, por se tratar de um assunto preocupante e que todos os anos causa tantos desastres na sociedade. Assim sendo, o nosso objetivo foi utilizar os dados que o dataset continha, para tentarmos prever a área queimada de incêndios florestais do Parque Natural de Montesinho.

A priori, colocamos algumas questões que, *a posteriori*, conseguimos responder com a análise exploratória que realizamos ao dataset. Essas questões e consequentes respostas foram:

- ✓ **Meses e dias da semana em que ocorrem incêndios com maior frequência?**
  Os meses que ocorrem mais incêndios são os meses de agosto e setembro, com uma larga diferença. Relativamente aos dias da semana, percebemos que à sexta, sábado e domingo são os dias em que ocorrem mais incêndios.
- ✓ **Quais os atributos climatéricos que mais influenciam a ocorrência de um incêndio de maior escala?**
  Não conseguimos afirmar que um determinado atributo interfere de forma muito significativa na dimensão de um incendio. Percebemos que os dados que nos são fornecidos no nosso *dataset* são mais importantes para prever a ocorrência de um incêndio do que o tamanho que ele pode vir a ter.
- ✓ **Qual a área/extensão total da floresta queimada de 2000 a 2003?**
  Neste trabalho observamos que a área total ardida do Parque neste intervalo de 3 anos foi de um total de 66642.05ha/100m2, sendo importante de referir que neste total não estão adicionados os incêndios que tiveram menos de 1ha/100m2 já que esses incêndios têm valor de 0 no nosso *dataset*.
- ✓ **Qual a zona geográfica mais afetada por incêndios?**
  Conseguimos perceber que grande percentagem dos incêndios ocorreram na zona em que o nosso Y (que corresponde a um atributo espacial do nosso dataset) possuía o valor de 4, sendo que podemos considerar essa zona como a que mais foi afetada por os incêndios.

Depois de realizarmos a análise exploratória e de conseguir responder às questões que propusemos, começamos a desenvolver vários modelos para tentarmos prever a dimensão da área ardida, sendo que recorremos a inúmeras técnicas de avaliação de performance de modelo, bem como á técnica de partição de dados cross-validation. Os vários resultados que obtivemos foram os que podemos ver na Tabela 1.

Cross Validation	Técnica Avaliação	K-NN	LDA	QDA	MLR
Não	Nenhuma	63.08%	60.00%	60.00%	57.03%
Não	AIC	---------	---------	---------	60.10%
Não	Adj_r^2	62.30%	64.61%	60.00%	---------
Não	CP	63.10%	64.61%	63.08%	---------
Não	BIC	63.10%	64.61%	63.08%	---------
Sim	Nenhuma	57.82%	---------	---------	59.9%
Sim	AIC	58.49%	---------	---------	61.24%
Sim	CP	57.82%	---------	---------	61.12%

Tabela 1 – Comparação dos vários métodos utilizados

Através da tabela, conseguimos perceber que em nenhum dos modelos utilizados conseguimos obter uma Accuracy e percentagem de acerto muito elevada. Os vários modelos que desenvolvemos classificam com alta percentagem de acerto os incêndios de pequena escala, contudo os incêndios de maior escala não consegue classifica-los corretamente. Assim, ao longo do trabalho, conseguimos perceber que os atributos contidos no nosso *dataset* são atributos mais importantes para a previsão de incêndios do que para previsão da dimensão dos mesmos.

2. Introdução

Este projeto insere-se na unidade curricular de Aprendizagem Automática I, onde se pretende aperfeiçoar os conhecimentos obtidos, *a priori*, de forma a aplicá-los num contexto prático. Assim sendo, foi escolhido um problema que consideramos conter questões interessantes que necessitam de ser analisadas e resolvidas.

Após uma procura minuciosa, optamos por um *dataset* sobre incêndios do Parque Natural de Montesinho, isto porque, consideramos tratar-se de um tema interessante e bastante preocupante. Posto isto, será explicado, no decorrer do relatório, todo o processo de elaboração e obtenção de resultados.

3. Sistema *FWI - Forest Fire Weather Index System*

O Sistema *FWI* trata-se de um sistema canadiense de índice climático de incêndio florestal que classifica o perigo de um incêndio. Este sistema subdivide-se em seis componentes que são responsáveis pelos efeitos da humidade do combustível e das condições climáticas no comportamento do fogo.

Os primeiros três componentes presentes no sistema *FWI* são os componentes que estão relacionados com o combustível do fogo, sendo eles:

- ***FFMC* (*Fine Fuel Moisture Code*) –** Denota o teor de humidade contida nos produtos vegetais de superfície e outros combustíveis secos que influenciam a ignição e a propagação do fogo. Mostra a relativa facilidade de ignição e a combustibilidade dos combustíveis finos.
- ***DMC* (Duff Moisture Code) –** Denota o teor de humidade de camadas orgânicas de profundidade moderada, ou seja, em níveis orgânicos compactos, mas soltos. O código indica a profundidade a que o fogo se produzirá em materiais vegetais de médio volume e na manta morta.
- ***DC* (*Drought Code*) –** Denota o teor de humidade de camadas orgânicas de profundidade significativa, ou seja, em materiais compactos e de grande volume. É um indicador útil de seca que mostra a probabilidade de o fogo atingir enorme profundidade nos materiais.

Os restantes componentes que completam o sistema canadiense *FWI* são:

- ***ISI* (*Initial Spread Index*) –** Denota uma pontuação que correlaciona a velocidade de propagação do fogo.
- ***BUI* (*Build-Up Index*)** – Denota a quantidade de combustível disponível.
- ***FWI* (*Forest Fire Weather*) –** Denota a intensidade do fogo através da combinação dos componentes *ISI* e *BUI*.

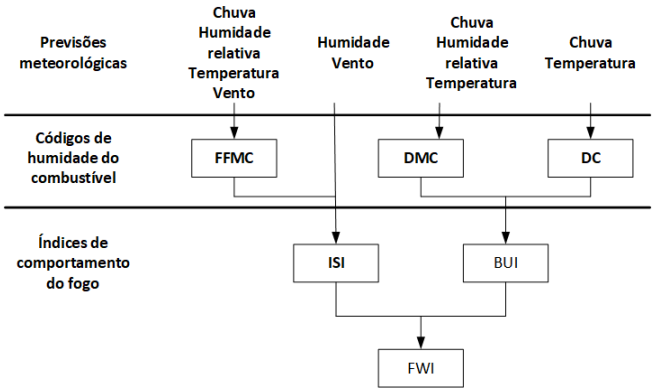

Figura 1 : Estrutura do sistema FWI

4. Descrição dos dados

O nosso *dataset* corresponde a um conjunto de dados sobre incêndios ocorridos no Parque de Montesinho, sendo que este conjunto de dados contém treze atributos divididos em cinco categorias. A primeira categoria inclui os atributos espaciais, como X e Y, que possuem as características geográficas da ocorrência dos incêndios. De seguida, nos atributos temporais, temos os atributos mês e dia da semana em que ocorreram os incêndios. A terceira categoria

inclui os atributos do sistema *FWI* sendo que os atributos *BUI* e *FWI* foram descartados porque, dependem dos restantes atributos. A quarta categoria contém os atributos climáticos usados pelo sistema *FWI*. Por fim, a quinta categoria, é o nosso *target*, ou seja, a nossa variável de interesse, que é a área queimada pelos incêndios.

Nome	Tipo	Atributos	Descrição
X	integer	Quantitativa	Coordenada espacial do eixo x dentro do mapa do parque de Montesinho: 1 a 9
Y	integer	Quantitativa	Coordenada espacial do eixo y dentro do mapa do parque de Montesinho: 2 a 9
month	string	Qualitativa	Mês do ano: 'jan' a 'dec'
day	string	Qualitativa	Dia da Semana: 'mon' a 'sun'
FFMC	float	Quantitativa	Índice FFMC do sistema FWI: 18,7 a 96,20
DMC	float	Quantitativa	Índice DMC do sistema FWI: 1,1 a 291,3
DC	float	Quantitativa	Índice DC do sistema FWI: 7,9 a 860,6
ISI	float	Quantitativa	Índice ISI do sistema FWI: 0,0 a 56,10
temp	float	Quantitativa	temperatura em graus Celsius: 2,2 a 33,30
RH	float	Quantitativa	Humidade Relativa em %: 15,0 a 100
wind	float	Quantitativa	Velocidade do vento em km/h: 0,40 a 9,40
rain	float	Quantitativa	Chuva externa em mm/m2: 0,0 a 6,4
area	float	Quantitativa	A área queimada da floresta (em ha): 0,00 a 1090,84

Tabela 2- Descrição dos dados do dataset

Assim, depois de analisado o *dataset*, torna-se agora importante referirmos o nosso propósito.

- ✓ **Objetivo:**
  O nosso objetivo será utilizar os dados do mundo real, recolhidos na região nordeste de Portugal, para prever se a área queimada de incêndios florestais do Parque Natural de Montesinho será uma área pequena, média ou elevada, para permitir aos bombeiros dinamizar a melhor estratégia de combate, tal como, a seleção dos meios de combate a incêndios.

- ✓ **Perguntas:**
  No final do nosso trabalho, tencionamos conseguir responder às seguintes questões:
  1) Meses e dias da semana em que ocorrem incêndios com maior frequência?
  2) Quais os atributos climatéricos que mais influenciam a ocorrência de um incêndio de maior escala?
  3) Qual a área/extensão total da floresta queimada de 2000 a 2003?
  4) Qual a zona geográfica mais afetada por incêndios?

Posto isto, passaremos agora para as abordagens que utilizaremos no decorrer do nosso trabalho.

5. Resumo das Abordagens

Inicialmente, tivemos que entender que tipo de modelo de aprendizagem de máquina iriamos utilizar no nosso trabalho, ou seja, se iriamos utilizar um modelo supervisionado ou não supervisionado. Um modelo supervisionado é um modelo de predição que utiliza tanto preditores como atributos objetivos para treinar o modelo enquanto que o modelo não supervisionado apenas utiliza os preditores para treinar o modelo. Assim, podemos entender facilmente, que iremos utilizar um método de aprendizagem supervisionado já que o nosso atributo objetivo se encontra presente no nosso *dataset*. Este tipo de aprendizagem pode ser utilizado em problemas de regressão ou classificação. Primeiramente, pensamos em utilizar técnicas de problemas de regressão e tentamos prever o valor contínuo da área ardida, contudo, visto que a nossa variável de interesse possuía um comportamento enviesado, com elevada tendência para valores próximos de zero, optamos por utilizar os problemas de classificação e categorizamos o atributo área em “Incêndio Pequeno”, “Incêndio Médio” e “Incêndio Grande”

consoante o valor do mesmo. Antes de categorizarmos a área, realizamos uma análise exploratória do nosso *dataset* para tentar perceber a relação da nossa variável de interesse (area) com as restantes, sendo que durante esta análise, decidimos transformar os atributos "month" e "day" em *integer*. Depois de terminada a nossa análise exploratória, começamos a desenvolver os nossos modelos de previsão. Os modelos que decidimos utilizar foram o *K-NN Classification, LDA, QDA e Multinomial Logistic Regression*, sendo que particionamos os dados em 75% para treino e 25% para teste e aplicamos a mesma partição em todos os modelos. Inicialmente, como não conseguíamos obter resultados concretos na análise exploratória, decidimos testar todos os modelos com todos os preditores, isto porque, não havia um determinado atributo que possuísse elevada relevância significativa. Posto isto, analisamos a performance dos modelos, de forma mais detalhadamente, através de diversas técnicas, tais como o AIC, R^2 *adjusted*, BIC, CP e mudamos as variáveis. Além disso, também realizamos a técnica de partição de dados "Cross Validation".

6. Abordagens Utilizadas e Suas Conclusões

6.1. Análise Exploratória

Inicialmente, começamos por realizar uma análise exploratória dos dados. Começamos por explorar o nosso atributo objetivo, ou seja, a área, para perceber como o seu valor varia. Assim, em relação à média é importante afirmar que o *dataset* contém muitos valores próximos de 0, ou seja, incêndios que tiveram uma área ardida inferior a 1ha/100m2. Isso ajuda-nos a entender o porquê de a média da área ardida ser baixa. Além disso, podemos observar através da *box plot* que a variável possui alguns *outliers* e que a maior parte dos valores da mesma correspondem a valores inferiores a 200ha/100m2. Além disso, também é importante vermos que a área total ardida foi de 66642.05 ha/100m2.
Depois de entendermos como varia a nossa variável objetivo iremos agora tentar perceber como ela se relaciona com as restantes variáveis.

	Média	Máximo	Soma
area	12.85	1090.84	6642.05

Tabela 3 – Dados do atributo "area"

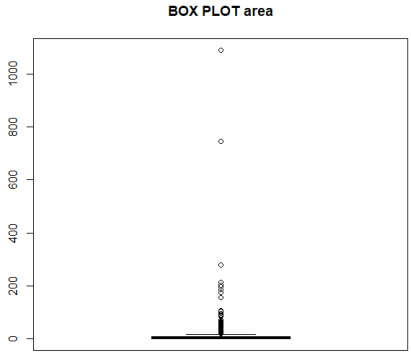

Figura 2 – Box Plot "área"

✓ **Atributos Temporais**

Uma ideia generalizada é o facto de haver maior ocorrência de incêndios de elevada dimensão ao longo do verão e por isso fomos verificar se essa ideia corresponde à realidade.

month	Frequência	Frequência Relativa
1	2	0.39%
2	20	3.87%
3	54	10.45%
4	9	1.74%
5	2	0.39%
6	17	3.28%
7	32	6.19%
8	184	35.59%
9	172	33.27%
10	15	2.90%
11	1	0.19%
12	9	1.74%

Tabela 4 – Ocorrência de incêndios por "month"

day	Frequência	Frequência Relativa
1	95	18.38%
2	74	14.31%
3	64	12.38%
4	54	10.45%
5	61	11.79%
6	85	16.44%
7	84	16.25%

Tabela 5 – Ocorrência de incêndios por "day"

Através destas tabelas, podemos perceber claramente que os meses de agosto e setembro, apresentam uma percentagem muito superior aos restantes meses, sendo que dos 517 incêndios registados, estes dois meses correspondem a 356 incêndios.
Quanto aos dias da semana, podemos perceber que a sexta, sábado e domingo apresentam quantidades superiores de ocorrência de incêndios.

Podemos também perceber a relação entre a área ardida, os meses e os dias na "Figura 3". Através da figura, vemos claramente que os meses em que os incêndios de maior escala ocorrem é sobretudo entre julho e setembro. Já os dias, concluímos que não interferem tanto com a seriedade dos incêndios, contudo, percebemos claramente que os dois incêndios de maior seriedade foram durante uma quinta-feira e um sábado.

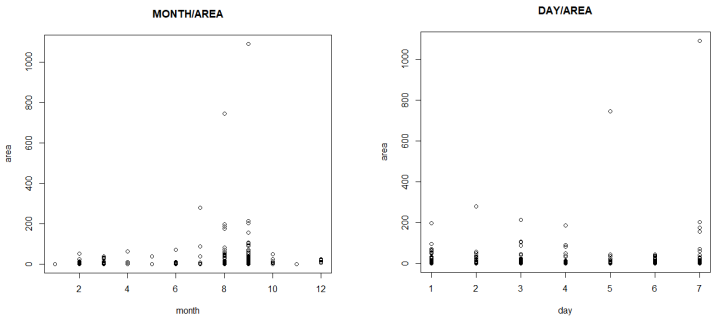

Figura 3 – Scatter Plot month/area e day/area

- ✓ **Atributos do sistema FWI**

  Após termos analisado o FWI torna-se fundamental na nossa análise perceber como se relacionam os preditores com a área ardida, e, também, como os valores desses preditores variam. Para isso, recorremos a diversos "Scatter Plots" e "Box Plots", sendo que estes podem ser consultados nos ***anexos***.

  - FFMC

  Podemos reparar que existe a ocorrência de mais incêndios e consequentemente incêndios de maior escala, quando o FFMC possui valores elevados (superiores a 80).

  - DC

  Através dos gráficos relativos ao DC, podemos concluir que valores mais elevados possuem maior probabilidade de haver mais incêndios e destes serem de maior escala. Os valores neste atributo variam especialmente ente 400 e 700.

  - DMC

  O atributo DMC é um atributo em que podemos constatar que existe muitos incêndios quando este atributo possui valores entre 60 e 150.

  - ISI

  Em relação a este atributo, este possui especialmente valores muito baixos que variam entre 0 e 20. Podemos ver que valores intermédios neste intervalo tem maior probabilidade de o incêndio ser de maior escala.

- ✓ **Atributos climáticos usados pelo FWI**

  Para terminarmos a análise da relação dos atributos com a nossa variável objetivo só nos falta avaliar os atributos climáticos. Os "Scatter Plots" e "Box Plots" que criamos podem ser consultados nos ***anexos***.

  - temp

  Podemos ver que a maioria dos valores deste atributo variam entre 15 e 25. Quanto à dimensão do incêndio, podemos ver que valores superiores a 15 e inferiores a 30 têm maior probabilidade de serem incêndios em maior escala.

  - rain

  A maior parte dos incêndios e os incêndios em maior escala, ocorrem quando a precipitação é nula.

  - wind

  Em relação a este atributo podemos constatar que tem sobretudo valores entre 2 e 8 e quanto ao tamanho do incêndio, este tem maiores chances de ser maior quando possui valores intermédios entre o intervalo que referimos.

  - RH

  Em relação a este atributo podemos reparar que a maior parte dos valores desta variável estão compreendidos entre 35 e 55, e que parece existir a tendência de ocorrer mais incêndios quando a RH possui valores entre 30 e 50.

Também decidimos observar a relação das nossas variáveis geográficas, isto para, conseguirmos verificar se existe alguma zona geográfica do parque mais afetada por incêndios. (Imagem do Parque em ***anexo***).

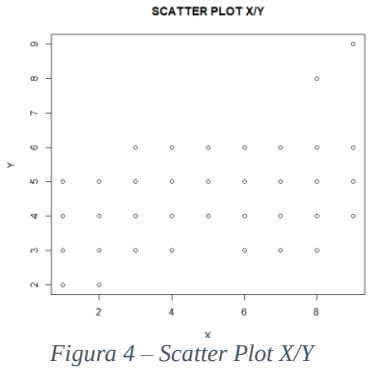

Figura 4 – Scatter Plot X/Y

Y	X	Número de Casos	%
6	8	52	10,06%
5	6	49	9,48%
4	7	45	8,70%
4	3	43	8,32%
4	4	36	6,96%

Tabela 6 – Top 5 de Casos de incêndios

Podemos denotar que, a zona geográfica em que houve mais casos foram as que se situam geograficamente em Y=6 e X=8.

De reparar que, 10,06% dos incêndios ocorreram nesta zona do parque. É de realçar que, a zona que possui Y=4 foi massacrada com pelo menos 23% dos incêndios. Através da "Figura 12", conseguimos perceber claramente que as zonas que possuem Y=7, Y=8 e Y=9 são claramente zonas em que raramente ocorrem incêndios.

Até agora, podemos ver que o nosso atributo objetivo, possui muitos valores a tender para 0 e, além disso, um comportamento muito enviesado. Assim, iria ser muito complicado treinar um modelo suficientemente bom para prever o valor contínuo da área e por isso decidimos categorizar a "área" em "Incêndio Pequeno" (entre 0 e 2), "Incêndio Medio" (entre 2 e 20) e "Incêndio Grande" (maior que 20).

Incêndio Pequeno	310	59,96%
Incêndio Médio	148	28,63%
Incêndio Grande	59	11,41%

Tabela 7 – Ocorrência de incêndios por classificação

Analisando a ocorrência de cada classe que criamos, podemos observar que existiu a ocorrência de mais incêndios com pouca seriedade (Pequenos) do que os restantes somados.

Para terminar a nossa análise exploratória decidimos avaliar a correlação entre os vários atributos (***anexos***). Através disso, percebemos que o atributo que possui maior correlação com os restantes trata-se do atributo DC, sendo que este, possui correlação positiva com o atributo DMC e month.

Em conclusão, com a nossa análise exploratória entendemos como variam os valores da nossa variável de interesse ("area") e a relação da mesma com todas as restantes variáveis, sendo que, conseguimos perceber como a variação das mesmas influência a ocorrência de incêndios. Assim, graças a esta análise exploratória, conseguimos responder a todas as questões que colocamos anteriormente (Página 5).

6.2. Modelos de classificação desenvolvidos

Tendo em conta a categorização já referida anteriormente iremos descrever agora, de forma detalhada, as técnicas que desenvolvemos para prevermos o tipo de incêndio através dos vários preditores. Assim, inicialmente desenvolvemos um código no R para normalizarmos os dados, excepto os atributos exceto o "day" e o "month", visto que, estes devem ser considerados como classes. Fizemos isto, pois ao normalizarmos os dados contínuos, estes ficarão todos na mesma escala o que aumentará a performance dos nossos modelos. Para todos os modelos, depois de normalizar os dados, dividimos os mesmos treino e teste, sendo que, selecionamos 75% dos dados de forma aleatória para treino e os restantes 25% para teste. Após a partição dos dados para treino e para teste, veremos agora cada modelo desenvolvido de forma individual.

✓ **KNN Classification**

O KNN trata-se de um modelo em que o atributo objetivo é classificado por uma pluralidade de votos dos seus vizinhos, com o objeto sendo atribuído à classe mais comum entre os seus k vizinhos mais próximos.

Assim, inicialmente neste modelo, desenvolvemos o "método de força bruta" para conseguirmos determinar o melhor "K" para o nosso modelo. Na figura 5 podemos observar os resultados que obtivemos com a utilização deste método, sendo que realizamos um gráfico para visualizarmos como os valores variaram (Figura 6).

```
k   Accuracy   Kappa
 5  0.5160698   0.0207637959
 7  0.5378851   0.0197407923
 9  0.5300519  -0.0237203749
11  0.5559284   0.0028941378
13  0.5704333   0.0112285069
15  0.5729998   0.0037581435
17  0.5814815   0.0168788181
19  0.5824972   0.0092967812
21  0.5833317   0.0045638881
23  0.5858058   0.0048939034
25  0.5841189  -0.0015405354
27  0.5831528  -0.0015935743
29  0.5849534   0.0027200174
31  0.5850648   0.0012813258
33  0.5832844  -0.0052411566
35  0.5857619  -0.0005585366
37  0.5874950   0.0012362322
39  0.5883936   0.0011143318
41  0.5858070  -0.0051456335
43  0.5866865  -0.0018647486

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 39.
```

Figura 5- Resultados para o método de força bruta

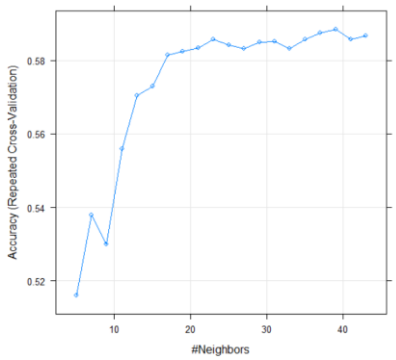

Figura 6- Gráfico dos resultados do método de força bruta

Podemos observar através das imagens, que o K que obteve melhor Accuracy foi quando o K era igual a 39, sendo que por este motivo, utilizamos o k=39 no nosso modelo KNN. Depois de percebido como escolhemos o K está na altura de ver os resultados que obtivemos.

	Incêndio Pequeno	Incêndio Médio	Incêndio Grande	TOTAL
Incêndio Pequeno	82	0	0	82
Incêndio Médio	35	0	0	35
Incêndio Grande	13	0	0	13

Tabela 8 – Resultados do K-NN

Accuracy: 63.08%

Através da tabela podemos entender que o nosso modelo a nível de Accuracy atingiu um valor aproximado de 63.1%, contudo, se analisarmos com atenção, conseguimos perceber que o modelo só tende a prever corretamente os incêndios da classe "Incêndios Pequenos". Os "Incêndios Médios" e os "Incêndios Grandes" tiveram 0 acertos.

Assim, percebemos que o modelo KNN só está a ter uma boa performance para a classe "Incêndios Pequenos".

✓ **Linear Discriminant Analysis (LDA)**

De uma forma mais geral, o LDA tem como objetivo maximizar a variância entre classes e minimizar a variância dentro das classes, por meio de uma função discriminante linear, partindo do pressuposto de que os dados em todas as classes são descritos por uma função densidade de probabilidade gaussiana com a mesma covariância.

Percebido o algoritmo LDA veremos agora os resultados que obtivemos através da sua execução.

	Incêndio Pequeno	Incêndio Médio	Incêndio Grande	TOTAL
Incêndio Pequeno	72	29	11	112
Incêndio Médio	9	6	2	17
Incêndio Grande	1	0	0	1

Tabela 9 – Resultados do LDA

Accuracy: 60.00%

Podemos ver que neste modelo o nível de Accuracy foi um pouco inferior ao que foi conseguido no KNN, sendo que obteve 60% de acerto, contudo, também podemos constatar

que aqui a percentagem de acerto nos Incêndios Médios foi superior á que tínhamos obtido no KNN. O melhor nível de acerto que obtivemos foi na classe "Incêndios Pequenos", mas com um resultado bastante pior do que obtidos no KNN, enquanto que, nos "Incêndios Grandes" continuamos sem conseguir classificar algum corretamente.

✓ **Quadratic Discriminant Analysis (QDA)**

O QDA está relacionado com o LDA, onde se assume que as medições são normalmente distribuídas. Contudo, ao contrário do LDA, no QDA não há suposição de que a covariância de cada uma das classes seja idêntica.

Percebido o QDA passaremos a ver os resultados obtidos no mesmo.

	Incêndio Pequeno	Incêndio Médio	Incêndio Grande	TOTAL
Incêndio Pequeno	72	29	11	112
Incêndio Médio	9	6	2	17
Incêndio Grande	1	0	0	1

Tabela 10 – Resultados do QDA

Accuracy: 60.00%

Se atentarmo-nos bem dos resultados que obtivemos, podemos ver que obtivemos exatamente os mesmos resultados obtidos que no modelo do LDA, quer a nível de Accuracy quer a nível de quantidades de acerto nas nossas classes.

✓ **Multinomial Logistic Regression (MLN)**

Como o nosso modelo tinha 3 classes não podemos recorrer à Logistic Regression porque esta só funciona para modelos binários. Assim, tivemos que recorrer á Multinomial Logistic Regression que se trata de um método de classificação que generaliza a regressão logística para problemas multiclasse. Primeiramente, os resultados que obtivemos neste modelo foram os seguintes:

	Incêndio Pequeno	Incêndio Médio	Incêndio Grande	Total
Incêndio Pequeno	71	5	1	77
Incêndio Médio	35	2	0	37
Incêndio Grande	14	0	0	14

Tabela 11 – Resultados do MLN

AIC: 768.71 Accuracy: 57.03%

Podemos observar mais uma vez que o nosso modelo não conseguiu prever corretamente a ocorrência de "Incêndios Grandes". Além disso, este modelo foi aquele que obteve a pior Accuracy com apenas 57.03%, sendo que, através da percentagem de acertos conseguimos perceber porque obtivemos uma Accuracy tão baixa. Os únicos incêndios que foram quase totalmente previstos foram os "Incêndios Pequenos". Quanto aos "Incêndios Médios" obtivemos uma percentagem de acerto não significativa, e, relativamente aos "Incêndios Grandes" obtivemos uma percentagem de acerto nula.

Posto isto, passamos a realizar uma avaliação aos atributos neste modelo, a ver se conseguíamos arranjar atributos para eliminar comparando os coeficientes e os *standards errors*. Assim veremos agora na "Tabela 11" os resultados que obtivemos, sendo que nela, colocamos apenas um dos doze months e um dos sete days porque os resultados eram similares em todos os meses e dias e assim a tabela fica mais pequena e mais percetível.

	Incêndio Pequeno (Coefficients)	Incêndio Pequeno (Std. Errors)	Incêndio Médio (Coefficient)	Incêndio Médio (Std. Errors)
X	1.298	0.745	1.144	0.801
Y	1.104	1.245	2.071	1.336
month(8)	-137.923	1.499	34.495	1.555
day(7)	0.042	0.580	-0.498	0.640
FFMC	-2.931	6.342	-3.025	6.562

DMC	-2.595	1.579	-2.763	1.659
DC	4.627	3.776	8.576	4.072
ISI	-0.129	3.652	0.326	3.843
temp	-1.533	2.149	-1.637	2.300
RH	-0.545	1.667	-0.412	1.758
wind	-1.661	0.992	-0.895	1.053
rain	372.285	2.084	375.464	2.084

Tabela 12 – Avaliação dos features

Ao observarmos os valores da tabela atentamente e como o nosso objetivo agora será melhorar a percentagem de acerto nos restantes incêndios, percebemos que quase todas as variáveis são pouco significativas para aumentar a percentagem de acerto dos "Incêndios Médios". Assim, decidimos experimentar o modelo apenas com os atributos "month", "DC" e "rain" já que estas possuem um Std. Errors de pelo menos metade do seu coeficiente. Obtivemos os resultados que podemos ver na Tabela 12.

	Incêndio Pequeno	**Incêndio Médio**	**Incêndio Grande**	**TOTAL**
Incêndio Pequeno	75	2	0	77
Incêndio Médio	35	0	2	37
Incêndio Grande	14	0	0	14

Tabela 13 – Resultados do MLR

AIC: 742.42 Accuracy: 60.16%

Podemos observar que o nosso AIC baixou ao utilizarmos somente as variáveis referidas anteriormente, contudo, continuamos só a prever corretamente os "Incêndios Pequenos" e não prevemos agora nenhuns "Incêndios Médios" nem "Incêndios Grandes". A nossa Accuracy teve uma subida, pouco significativa, para 60.16%.

6.3. Técnicas de Avaliação

Depois de vermos os resultados que obtivemos ao colocar todas as variáveis no nosso modelo vamos agora utilizar técnicas de avaliação para vermos como podemos melhorar a performance dos nossos modelos. Inicialmente, utilizamos uma função no R para fazermos a seleção dos melhores *features*, sendo que podemos ver na Tabela 13 os resultados que obtivemos dessa seleção de forma generalizada. Para a tabela ficar menos confusa e também menos extensa, colocamos apenas um "month" e um "day". (Podemos ver a tabela completa que nos foi dada nos ***anexos***)

	X	**Y**	**month**	**day**	**FFMC**	**DMC**	**DC**	**ISI**	**temp**	**RH**	**wind**	**rain**
1 -2			X									
3			X								X	
4-6			X						X		X	
7			X	X					X		X	
8			X	X		X			X		X	
9-10	X		X	X		X			X		X	
11-14	X		X	X		X	X		X		X	
15-16	X		X	X		X	X		X	X	X	
17-20	X		X	X	X	X	X	X	X	X	X	
21-24	X	X	X	X	X	X	X	X	X	X	X	
25-27	X	X	X	X	X	X	X	X	X	X	X	X

Tabela 14 – Avaliação dos features

- ✓ **Adj_r^2**

Começamos então por avaliar o modelo utilizando o R^2 *adjusted*. Nesta técnica, quanto maior fosse o resultado que obtivéssemos melhor é a avaliação do mesmo. Podemos observar na tabela 14 que o melhor valor que obtivemos foi

	Adj_r^2
9	0.03197
10	0.03179
11	0.0312

Tabela 15 – Resultados obtidos para o Adj_r^2

para quando estávamos na presença de 9 *features*, sendo que podemos observar os *features* que são na tabela 13. Na Figura 7 podemos perceber que até aos 10 *features* houve um aumento relativamente gradual da performance do modelo, contudo a partir daí foi descendo a performance do mesmo também de forma relativamente gradual.

Veremos agora os resultados obtidos para cada modelo consoante as conclusões tiradas com a utilização desta técnica.

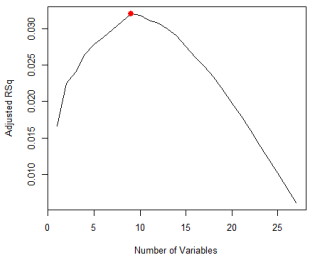

Figura 7 – Gráfico da variação do ADJ_r^2

	K-NN			
	Incêndio Pequeno	**Incêndio Médio**	**Incêndio Grande**	**TOTAL**
Incêndio Pequeno	81	1	0	82
Incêndio Médio	35	0	0	35
Incêndio Grande	13	0	0	13

Tabela 16 – Resultados do K-NN

Accuracy: 62.30%

Podemos ver que o desempenho do algoritmo K-NN piorou relativamente aos resultados que tínhamos obtido utilizando todas as variáveis, contudo não foi uma mudança significativa. Também é importante referir que continuou somente a acertar os "Incêndios Pequenos".

	LDA			
	Incêndio Pequeno	**Incêndio Médio**	**Incêndio Grande**	**TOTAL**
Incêndio Pequeno	82	33	13	128
Incêndio Médio	0	2	0	2
Incêndio Grande	0	0	0	0

Tabela 17 – Resultados do LDA

Accuracy: 64.61%

O desempenho do LDA melhorou, embora não tenha sido uma mudança significativa. Contudo, há que salvaguardar que na nossa amostra existe uma maioria de elementos pertencentes à classe "Incêndios Pequenos", sendo que, como vimos até agora, é o que melhor desempenho possui comparativamente com os restantes modelos.

	QDA			
	Incêndio Pequeno	**Incêndio Médio**	**Incêndio Grande**	**TOTAL**
Incêndio Pequeno	72	29	11	77
Incêndio Médio	9	6	2	37
Incêndio Grande	1	0	0	14

Tabela 18 – Resultados do QDA

Accuracy: 60.00%

Contrariamente ao que aconteceu anteriormente quando usamos todas as variáveis, o QDA agora não obteve a mesma Accuracy que o LDA. Contudo, a percentagem de acerto é a mesma que tínhamos obtido anteriormente.

✓ **CP**

Na técnica do CP quanto menor for o resultado que obtemos melhor é a avaliação do mesmo. Assim, ao analisar os valores, conseguimos entender facilmente que o menor valor que obtemos é quando temos 4 *features*, sendo que podemos ver os *features* na Tabela 13. Quanto à variação da performance consoante os *features*, podemos perceber através da Figura 8 que no início vai melhorando e a partir do 4 vai piorando até ao final.

Veremos agora os resultados obtidos com as conclusões tiradas aqui.

	CP
4	-5.479
2	-5.415
3	-5.244

Tabela 19 – Resultados obtidos para o CP

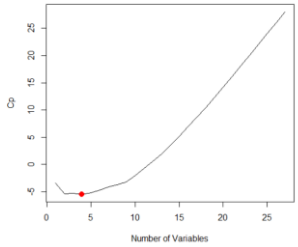

Figura 8 – Gráfico da variação do CP

	K-NN			
	Incêndio Pequeno	Incêndio Médio	Incêndio Grande	TOTAL
Incêndio Pequeno	82	0	0	82
Incêndio Médio	35	0	0	35
Incêndio Grande	13	0	0	13

Tabela 20 – Resultados do K-NN

Accuracy: 63.10%

Com este modelo, conseguimos aumentar a nossa Accuracy, contudo, continua a não acertar nenhum incêndio exceto os "Incêndios Pequenos".

	LDA			
	Incêndio Pequeno	Incêndio Médio	Incêndio Grande	TOTAL
Incêndio Pequeno	82	33	13	128
Incêndio Médio	0	2	0	2
Incêndio Grande	0	0	0	0

Tabela 21 – Resultados do LDA

Accuracy: 64.61%

Aqui, obtivemos resultados melhores que no K-NN, nas mesmas condições. Também é importante referir que correspondem a resultados exatamente iguais ao método LDA com a técnica de avaliação de performance Adj_r^2.

	QDA			
	Incêndio Pequeno	Incêndio Médio	Incêndio Grande	TOTAL
Incêndio Pequeno	82	35	13	130
Incêndio Médio	0	0	0	0
Incêndio Grande	0	0	0	0

Tabela 22 – Resultados do QDA

Accuracy: 63.08%

Podemos observar que o QDA foi aquele que obteve pior Accuracy dos métodos utilizados com a avaliação de performance CP.

✓ **BIC**

A última técnica que utilizamos para avaliar a performance do nosso modelo foi o BIC. Nesta técnica, quanto menor fosse o resultado que obtivéssemos melhor é a avaliação do mesmo. Assim, podemos ver na Tabela 22 que para este modelo o melhor resultado que obtemos é só na utilização de 1 *feature* sendo se trata do "month". Quanto á variação dos valores podemos ver na Figura 9 que á medida que as *features* aumentam a avaliação piora gradualmente.

Passaremos agora a ver os resultados que obtivemos com este modelo.

	BIC
1	2.809
2	5.014
3	9.388

Tabela 23 – Resultados obtidos para o BIC

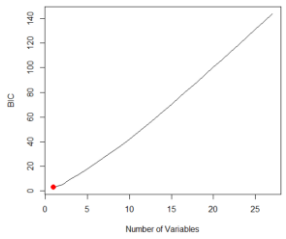

Figura 9 – Gráfico da variação do BIC

	K-NN			
	Incêndio Pequeno	Incêndio Médio	Incêndio Grande	TOTAL
Incêndio Pequeno	82	0	0	82
Incêndio Médio	35	0	0	35
Incêndio Grande	13	0	0	13

Tabela 24 – Resultados do K-NN

Accuracy: 63.10%

Os resultados que obtivemos aqui foram exatamente iguais aos resultados que obtivemos para o KNN com o método de avaliação CP.

	LDA			
	Incêndio Pequeno	Incêndio Médio	Incêndio Grande	TOTAL
Incêndio Pequeno	82	35	13	130
Incêndio Médio	0	0	0	0
Incêndio Grande	0	0	0	0

Accuracy: 63.08%

Tabela 25 – Resultados do LDA

Os resultados que obtivemos aqui foram piores aos resultados que obtemos para o LDA com o método de avaliação de Performance CP. Também é importante referir que o LDA teve piores resultados que o K-NN.

	QDA			
	Incêndio Pequeno	Incêndio Médio	Incêndio Grande	TOTAL
Incêndio Pequeno	82	35	13	130
Incêndio Médio	0	0	0	0
Incêndio Grande	0	0	0	0

Accuracy: 63.08%

Tabela 26 – Resultados do QDA

Podemos ver que o QDA obteve uma Accuracy e resultados exatamente iguais que ao LDA.

6.4. Cross Validation

Para terminar, decidimos realizar a partição dos dados através da técnica do Cross Validation, sendo que, utilizamos esta partição para os modelos de MLR e para o K-NN. Para utilizar esta técnica recorremos a uma função do R que nos permite realizar isto e ao utilizá-la obtivemos os valores de Accuracy que podemos observar na tabela 26, sendo que, testamos também com os resultados que obtivemos do AIC e CP. A nossa função utilizou o Repeated K-fold e colocamos "K" igual a 10 a repetir 3 vezes.

Técnica Avaliação	Nenhuma	AIC	CP
K-NN	57.82%	58.49%	57.82%
MLR	59.9%	61.24%	61.12%

Tabela 27 – Resultados da aplicação do Cross Validation

Os melhores resultados que obtivemos ao realizar esta técnica de partição foi para o MLR com os resultados que tínhamos obtidos anteriormente para a técnica de avaliação AIC. Contudo, a diferença de Accuracy que nos foi dada em todos os modelos não é muito significativa.

7. Conclusão

Ao longo do trabalho, desenvolvemos inúmeros algoritmos de aprendizagem de máquina para prevermos de forma classificativa que tipo de incêndio se tratava. Se repararmos na Tabela 28 podemos ver os vários resultados que obtivemos com o desenvolvimento dos nossos modelos.

Cross Validation	Técnica Avaliação	K-NN	LDA	QDA	MLR
Não	Nenhuma	63.08%	60.00%	60.00%	57.03%
Não	AIC	---------	---------	---------	60.10%
Não	Adj_r^2	62.30%	64.61%	60.00%	---------
Não	CP	63.10%	64.61%	63.08%	---------
Não	BIC	63.10%	63.08%	63.08%	---------
Sim	Nenhuma	57.82%	---------	---------	59.90%
Sim	AIC	58.49%	---------	---------	61.24%
Sim	CP	57.82%	---------	---------	61.12%

Tabela 28 – Comparação dos vários métodos utilizados

Conseguimos perceber que, o resultado que nos deu melhor performance foi o LDA sem a utilização da partição dos dados através do Cross Validation e com a utilização das conclusões que tiramos para as Técnicas de Avaliação CP e Adj_r^2. Estes modelos, atingiram os 64.61% de Accuracy, sendo que, podemos considerar que se trata de um valor mediano. Isto porque, ao longo do trabalho percebemos que os atributos contidos no nosso *dataset* não são muito bons para prever a extensão dos incêndios, mas sim para prever a ocorrência de incêndios, sendo que, além disso, também se trata de um *dataset* pouco extenso (517 linhas). Assim, todos os modelos tinham grande percentagem de acerto na ocorrência de incêndios de pequenas dimensões, mas para os restantes, a percentagem de acerto era muito baixa ou mesmo nula.