Incêndios Florestais em Portugal - Relatório

Grupo 7

Cristiana Silva up201505454, Nuno Tomás up201503467, Rui Santos up201805317

14/01/2022

Definição do problema

Os incêndios florestais são uma questão muito importante, que afeta negativamente as mudanças climáticas, cujas causas normalmente são os descuidos, acidentes e negligências cometidos por indivíduos, atos intencionais e causas naturais. Estes podem ter impactos e efeitos nocivos sobre os ecossistemas, levando ao desaparecimento de espécies e até ao aumento dos níveis de dióxido de carbono.

Assim sendo, e com um intuito de analisar com melhor detalhe e tentar encontrar formas que possam ajudar a evitar estas tragédias, desenvolvemos este projeto com o intuito de encontrar um modelo que nos permitisse determinar se a causa de um incêndio foi intencional ou não.

Preparação dos dados

Após fazer importação dos dados, o nosso ponto de partida foi remover as variáveis desnecessárias bem como fazer um pré-processamento dos dados.

Assim sendo, das 21 colunas iniciais tomamos as seguintes decisões:

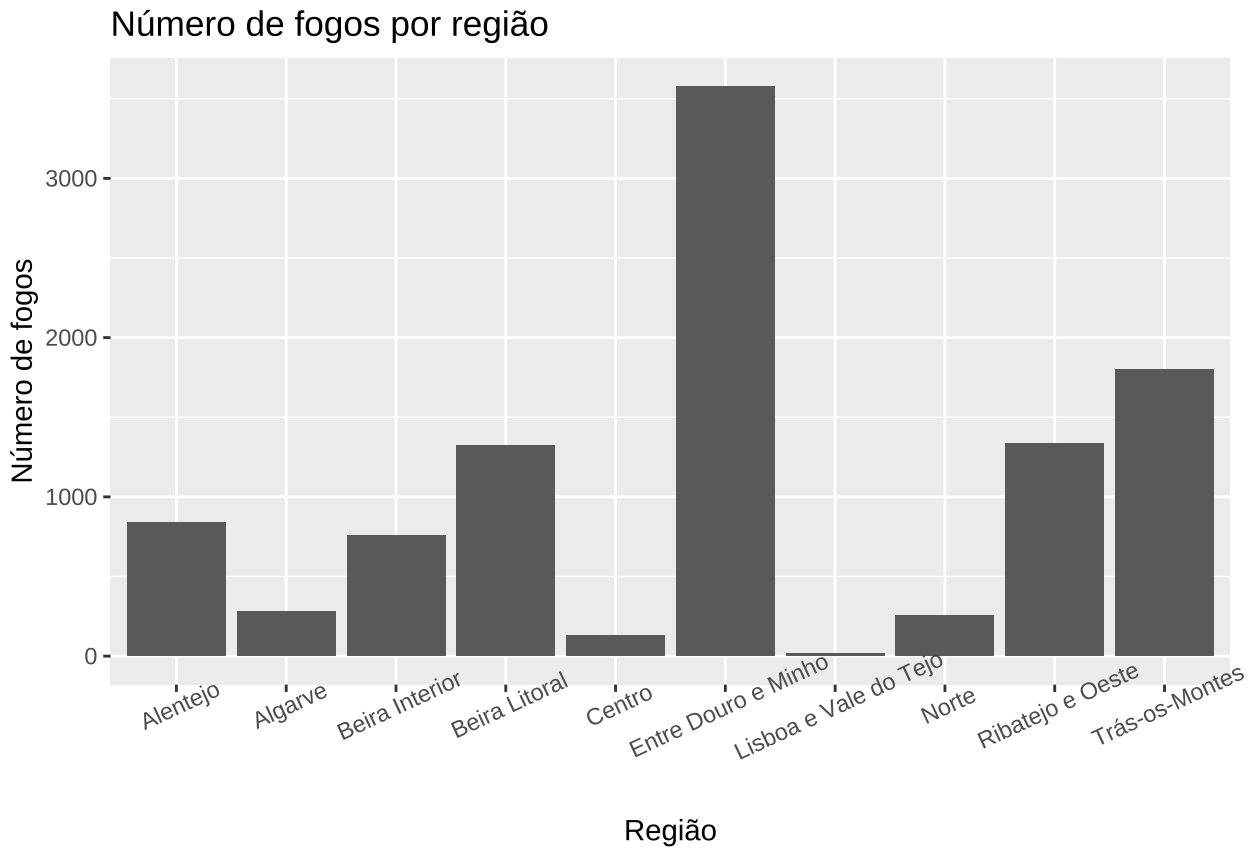
- O **id** optamos por manter uma vez que este atributo é único e necessário para identificar os diferentes dados.
- A **region**, **district**, **municipality** e **parish** são fatores importantes uma vez que ajudam na identificação da localização do incêndio.
- A **lat** e **lon** tal como os anteriores têm influência na localização do fogo e permitem localizá-lo com maior precisão numa dada zona.
- A **origin** é o atributo que indica a origem do incêndio, o que por si só já o torna um fator importante
- A **alert_date** e **alert_hour**, a data e hora do incêndio que nos podem ajudar a determinar o período o dia bem como a época do ano em que ocorreu o mesmo.
- A **extinction_date**, **extinction_hour**, **firstInterv_date**, **firstInterv_hour** são atributos são relativos à extinção do incêndio, ou seja, não nos dizem nada sobre a origem ou causa do mesmo.
- A **alert_source** como possui todos os valores como **NA** não tem importância nenhuma pra o resultado final.
- A **village_area**, **vegetation_area**, **farming_area**, **village_veget_area** e **total_area** permitem-nos identificar se a zona em que ocorreu o incêndio é habitável ou uma zona vegetacional, o que pode ter influência na causa do incêndio então decidimos manter inicialmente todos.
- Por fim a **intentional_cause** é o atributo que nos permite saber se a causa do incêndio e será importante para o nosso modelo de marchine learning.

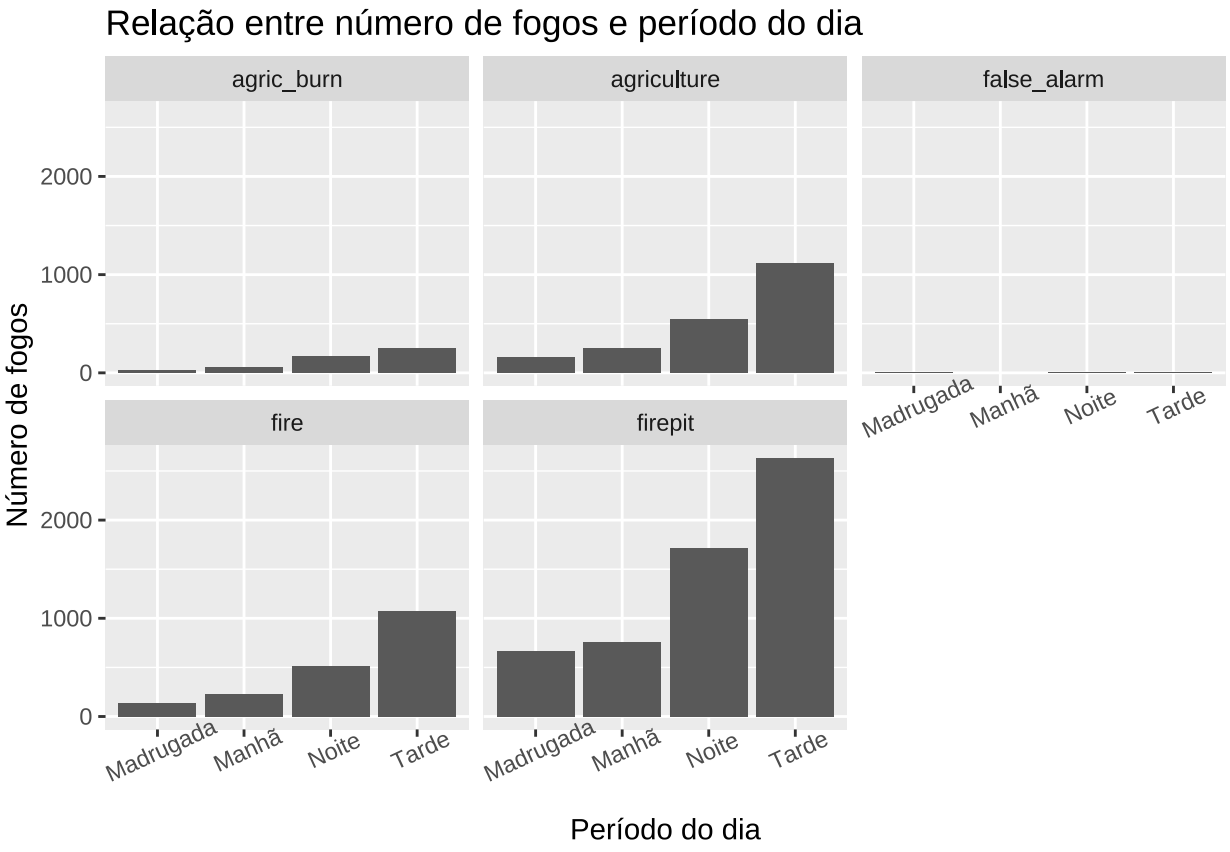
Após isto, verificamos que ainda existiam atributos em falta na **region** e assim sendo optamos por ordená-los por região e em seguida agrupa por distrito de forma a eliminar esses missing values. Depois, notamos um

incorreta formatação dos valores da latitude e longitude e assim sendo no caso do primeiro remove-mos o elemento com o formato de data **1900-01-01** e em ambos, trocamos a **,** por um **.** e como algumas deste valores possuíam um número elevado de casas decimais, limitamos o tamanho de cada um a 9 carateres. Por fim a nível de formatação apenas tivemos de corrigir o formato da data para **YY-MM-DD**.

Tendo todos os dados nos formatos corretos, adicionamos duas novas colunas com dados, a **timePeriod** e **tmax**. Na primeira fica registado a altura do dia em que ocorreu o incendio a partir da hora do mesmo e na segunda a temperatura máxima no dia do incêndio naquela zona, usando como auxiliar o **getTemperatureNOAA.R**.

Exploração dos dados e análise





Configuração experimental

Para este ponto começamos inicialmente por ver que tipo de predictive modelling melhor se enquadrava neste problema e que neste caso, como a variável é nominal uma vez que o objetivo é prever se a causa do incêndio foi intencional ou não, escolhemos os algoritmos Partindo desta doutrina, escolhemos três modelos mais intuitivos e robustos: o **Random Forests**, **Naive Bayes** e o **k-Nearest Neighbors**.

O **Random Forests** é um ensemble learning, o que que permite que seja usado tanto para classificação como para regressão, para além de que é um dos mais fáceis de usar, fornecendo um nível mais alto de precisão na previsão de resultados para além de que o parâmetro mais importante a ser ajustado é o número de árvores, normalmente quanto maior melhor, não sendo assim necessário ajustes muito elaborados. Em relação ao **Naive Bayes** escolhemos porque funcionava rapidamente e permitia economizar muito tempo, contudo como os predictors são considerados independentes, o que não é o caso neste cenário, não é possível ter a certeza da precisão dos resultados da probabilidade porque suas estimativas podem estar erradas em alguns casos. Por fim o **k-Nearest Neighbors** que é dos melhores algoritmos de classificação e uma vez que não requer nenhuma etapa de treino explícita, ou seja, aprende por analogia, tendo como base a noção de semelhança entre casos o que nesta situação seria vantajoso, só que que tal como o anterior considera os predictors independentes. Com estas informações em mente e após analisarmos as vantagens de desvantagens de cada um decidimos optar pelo **Random Forest** uma vez que este pode ser usado tanto para classificação como para regressão, para além de que é dos mais fáceis de usar.

Resultados

Ao aplicar os modelos, fomos fazendo submissões no **kaggle** e podemos chegar a alguns resultados. ### Naive Bayes Tentamos implementar o **Naive Bayes** mas foi preciso categorizar a maior parte das variáveis

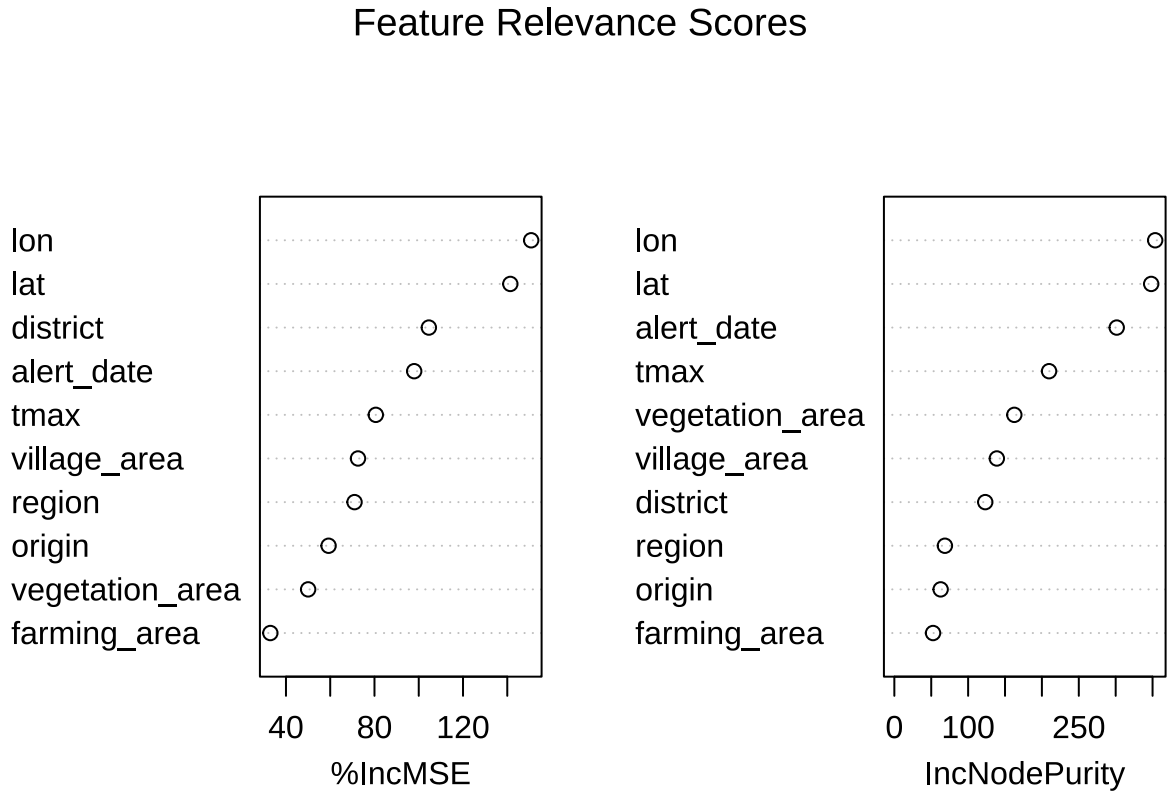
sendo que algumas delas ficaram com muitas categorias. Ainda assim obtivemos 0.52615.

randomForest

Finalmente, e como referimos anteriormente, o **randomForest** foi o que inicialmente nos levou a melhor resultados mesmo antes de aplicarmos a temperatura. Assim que esta foi usada notamos que houve um melhoramento o que nos levou a concluir que poderia ser um bom fator de previsão.

Inicialmente usamos um número de arvores igual a 1000. De seguida experimentamos aumentar o número de árvores para 2000, mas os resultados pioraram. Tentamos por último remover colunas (latitude e longitude) o que piorou os resultados (0.77255) sendo assim o nosso melhor resultado de 0.83223 após inserir o distrito e a temperatura máxima.

```
##
## Call:
##  randomForest(formula = intentional_cause ~ ., data = aux, ntree = 1000,      importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 1000
## No. of variables tried at each split: 3
##
##           Mean of squared residuals: 0.1418395
##                     % Var explained: 30.13
```



O gráfico da importância da **randomForest** mostra-nos a relevância das variáveis latitude e longitude e a razão pela qual se as tirarmos o nosso score diminui.

Conclusões, limitações e trabalhos futuros

As limitações que encontramos foram que se passássemos mais tempo com o **KNN** e o **Naive Bayes** talvez conseguíssemos obter melhores resultados. Também poderíamos possivelmente obter um melhor score se tivéssemos mais dados extra, por exemplo informações sobre o vento ou até mesmo sobre precipitação. Por fim uma limitação que encontramos poderá ser o facto de o **kaggle** ter um número de submissões limitado a duas por dia. Em suma, este trabalho permitiu-nos conhecer os diferentes modelos de previsão existentes bem como aprofundar os nossos conhecimentos da linguagem **R**.