Incêndios Florestais em Portugal - Apresentação

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

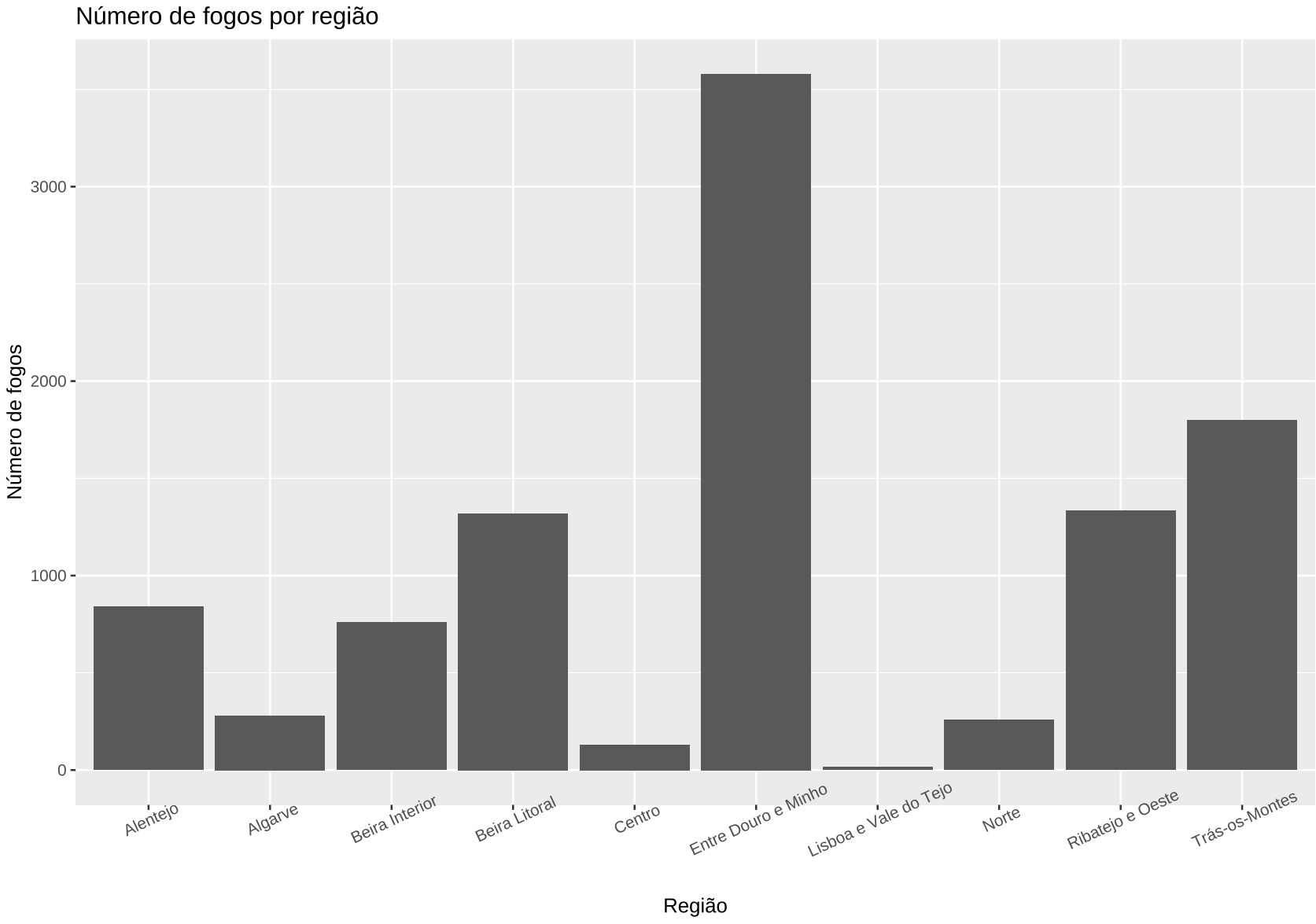
Após fazer importação dos dados, o nosso ponto de partido foi remover as variáveis desnecessárias bem como fazer um pré-processamento dos dados. Assim sendo removemos: - A **extinction_date**, **extinction_hour**, **firstInterv_date**, **firstInterv_hour** - A **alert_source** como possui todos os valores como **NA** não tem importância nenhuma para o resultado final. - Já a **village_veget_area** e **total_area** como eram a soma dos anteriores decidimos remover e ficar com os só com os atributos da área referidos anteriormente. Os restantes atributos mantivemos pois achamos que seriam importantes para a previsão.

Preparação dos dados

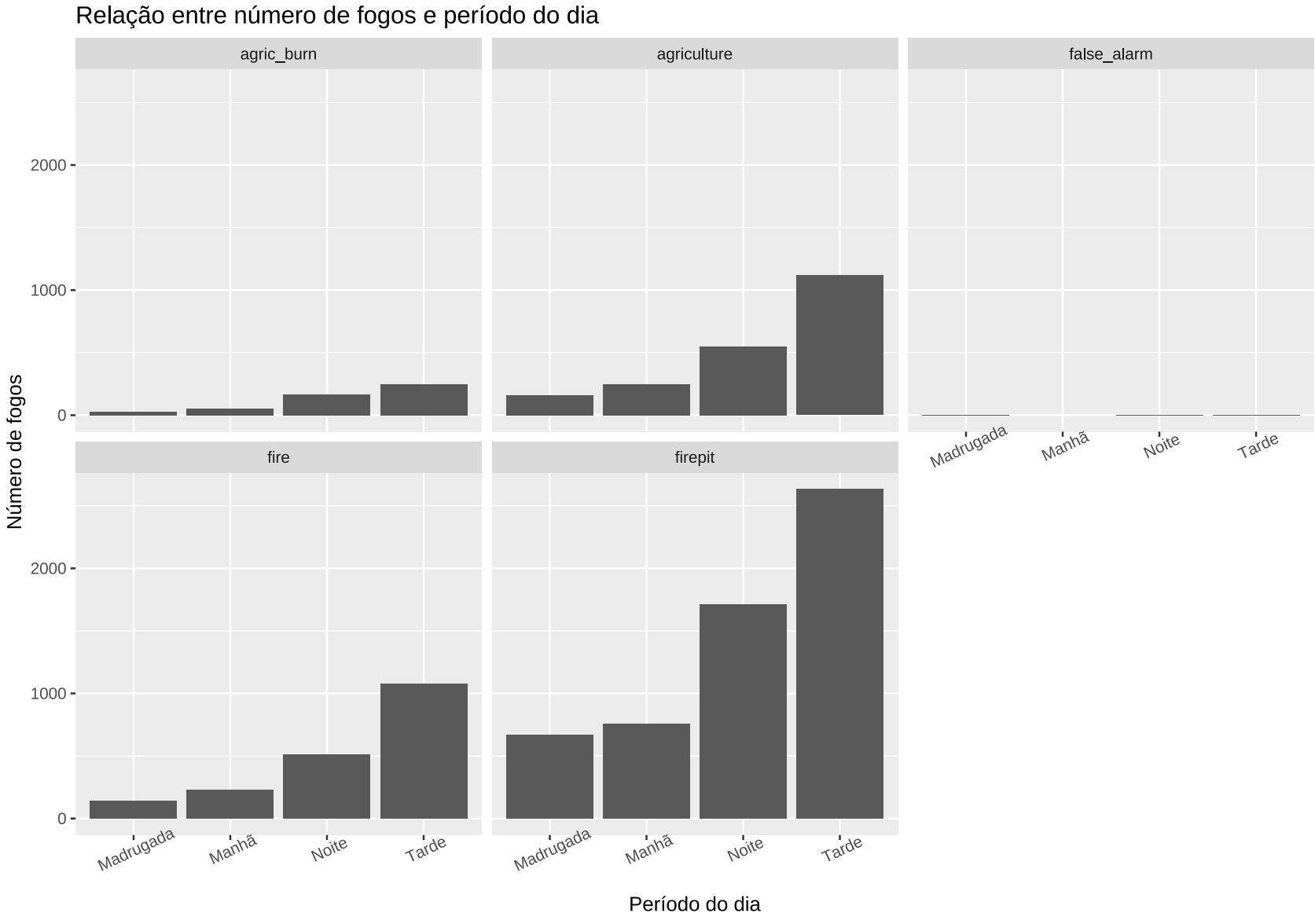
Após isto, verificamos que ainda existiam atributos em falta na **region** e assim sendo optamos por ordená-los por região e em seguida agrupa por distrito de forma a eliminar esses missing values. Depois, notamos um incorreta formatação dos valores da latitude e longitude e assim sendo no caso do primeiro 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



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**.

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

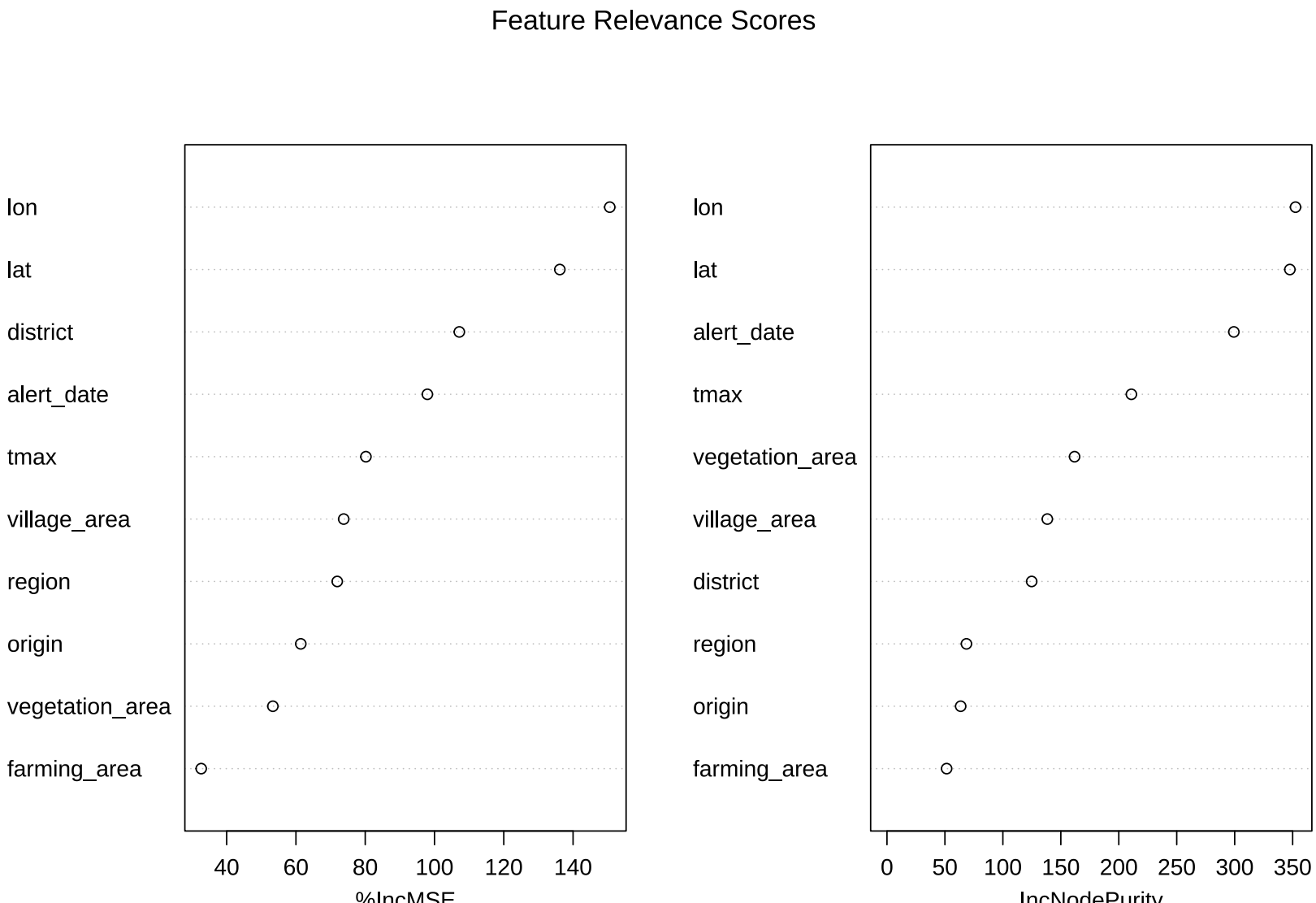
Finalmente, e como referimos anteriormente, o **randomForest** foi o que inicialmente nos levou a melhor resultados mesmo antes de aplicarmos a temperatura. Assim que esta foi usada notamos que houve um melhoramento o que nos levou a concluir que poderia ser um bom fator de previsão.

Resultados

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.

Resultados

```
## Warning in randomForest.default(m, y, ...): The response
## unique values. Are you sure you want to do regression?
```



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. **R**.