

PREVISÃO DE FOGO POSTO EM PORTUGAL 2014-2015

Data Mining I - Trabalho Prático

01 de janeiro de 2023

Trabalho Realizado por:

Joana Pereira (201805191)

Pedro Azevedo (201905966)

Pedro Santos(201904529)

Docente:

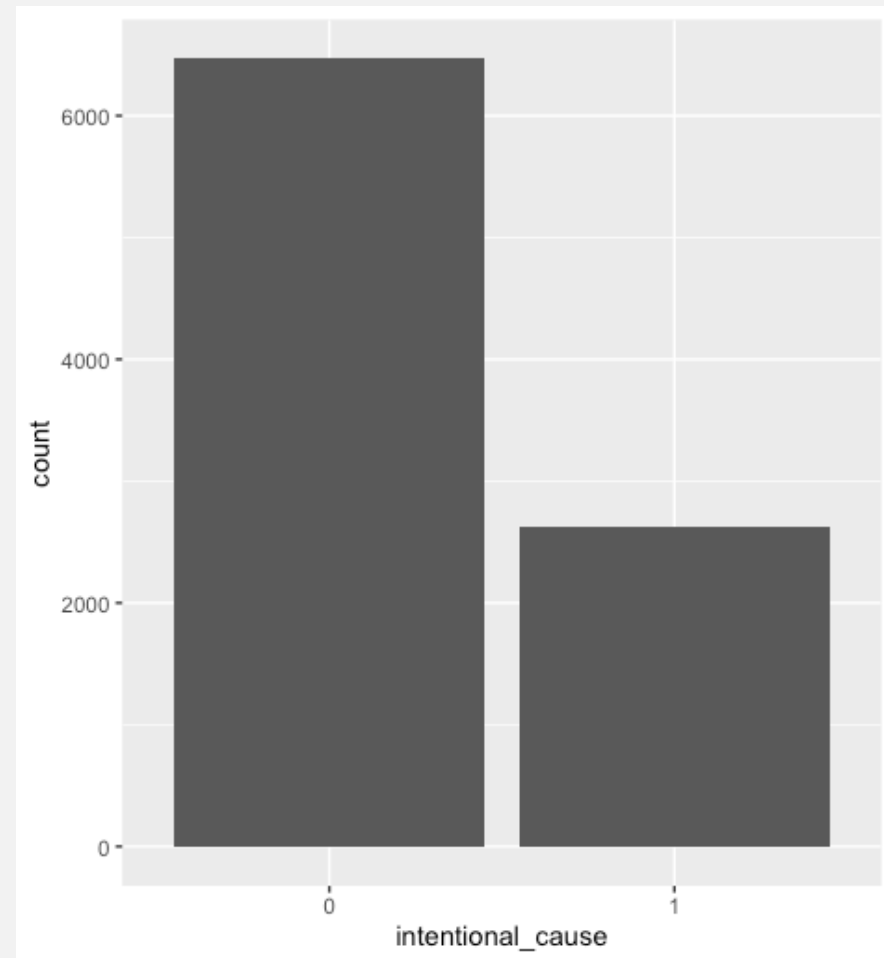
Rita Ribeiro

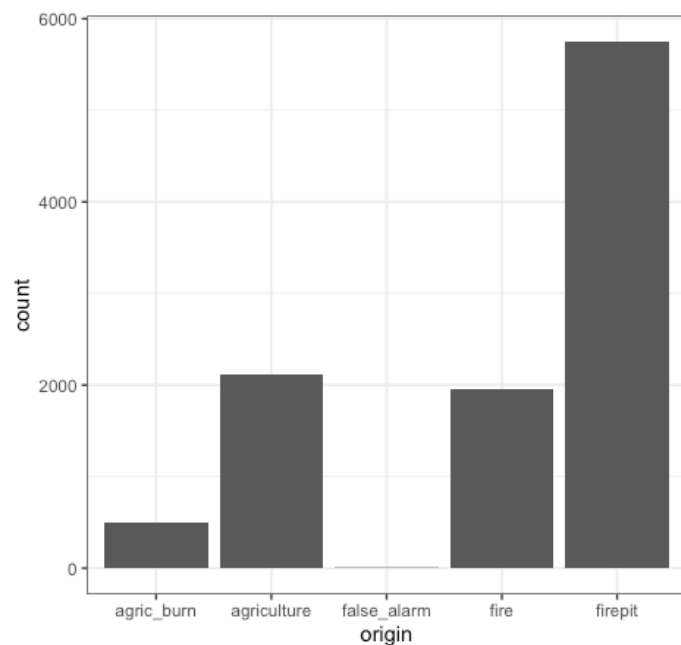
ÍNDICE

- Definição do problema
- Data Understanding
- Data Preparation
- Melhorar o conjunto de dados
- Avaliar features
- Predictive Modelling
- Conclusão

DATA UNDERSTANDING

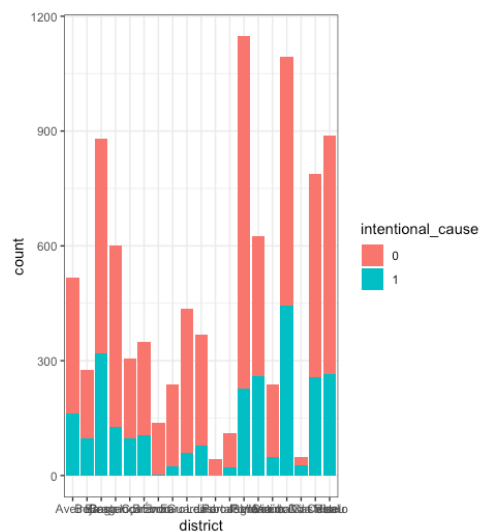
- Características:
 - Id, region, district, municipality, parish, lat, lon, origin, alert_date, alert_hour, extinction_date, extinction_hour, firstInterv_date, firstInterv_hour, alert_source, village_área, vegetation_área, farming_área, village_veget_área, total_area
- Output
 - intentional_cause.
 - 0: 72%
 - 1: 28%





Pode, ainda, concluir que a grande parte dos fogos são iniciados por fogueira.

Viana do Castelo apresenta a maior taxa de fogo posto.



DATA
UNDERSTANDING

DATA PREPARATION

- Apagar colunas que não tem valores como alert_source e uma coluna que é muito específica e que não traria valor.

```
fire_Train_Data <- fire_Train_Data %>% select(-c(alert_source, parish))
```

- Descobrir onde existem valores nulos:

```
apply(X = is.na(fire_Train_Data), MARGIN = 2, FUN = sum)
```

- Coluna region tem poucos e por isso foram preenchidos:

```
fire_Train_Data <- fire_Train_Data %>% mutate(region = ifelse(is.na(region),  
"Ribatejo e Oeste", region))
```

DATA PREPARATION

- Por fim, apagar linhas com algum valor nulo.

```
y = c("extinction_hour", "firstInterv_date", "firstInterv_hour")
```

```
vars <- "y"
```

```
fire_Train_Data <- drop_na(fire_Train_Data, any_of(y))
```

- No caso de teste, os valores foram preenchidos:

```
fire_Test_Data <- fire_Test_Data %>% fill(extinction_date)
```

```
fire_Test_Data <- fire_Test_Data %>% fill(extinction_hour)
```

```
fire_Test_Data <- fire_Test_Data %>% fill(firstInterv_date)
```

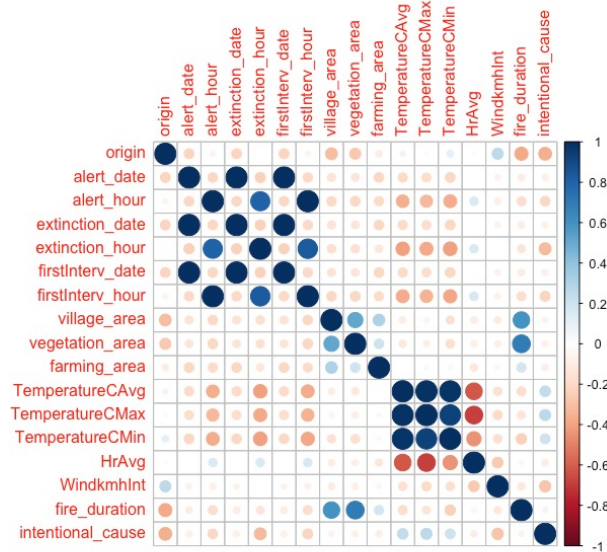
```
fire_Test_Data <- fire_Test_Data %>% fill(firstInterv_hour)
```

MELHORAR O CONJUNTO DE DADOS

- Usando uma biblioteca externa obteve-se mais características, usando a lat, lon e date do conjunto original de dados adicionou-se mais características, tanto ao conjunto de treino como de teste.

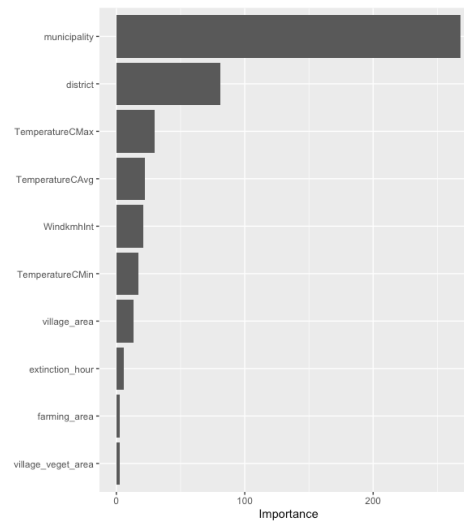
```
install_github("bczernecki/climate", force = TRUE)
```

- Assim, o conjunto final de dados ficou com as colunas:
 - District, municipality, origin, alert_date, alert_hour, extinction_date, extinction_hour, firstInterv_date, firstInterv_hour, village_area, vegetation_area, farming_area, village_veget_area, total_area, TemperatureCAvg, TemperatureCMax, TemperatureCMin, HrAvg, WindkmhInt, fire_duration e intentional_cause
- Também houve tratamento de dados.



Foi possível observar as correlações entre as features.

As caraterísticas com mais importância.



AVALIAR
FEATURES

PREDICTIVE MODELLING

- Modelos testados:
 - Multiple Linear Regression
 - Ridge Regression
 - Lasso Regression
 - Cart Trees
 - KNN

PREDICTIVE MODELLING (LASSO REGRESSION)

- Usando as colunas que foram avaliadas como mais importantes:

```
model_glm_lasso <- linear_reg(engine="glmnet",penalty = 10^-2,mixture=1)
```

```
glm_lasso_fit <- model_glm_lasso %>%  
  fit(intentional_cause ~ district + TemperatureCMax + WindkmhInt +  
  TemperatureCAvg + TemperatureCMin + village_area + extinction_hour +  
  farming_area + village_veget_area, data = fire_train)
```

```
glm_lasso_preds <-  
  fire_test %>% dplyr::select(intentional_cause) %>%  
  bind_cols(predict(glm_lasso_fit,fire_test))  
glm_lasso_preds %>% metrics(truth=intentional_cause,estimate=.pred)
```

- Avaliar Root Mean Squared Error, R-square e Mean absolute Error.

```
# A tibble: 3 × 3  
  .metric .estimator .estimate  
  <chr>    <chr>         <dbl>  
1 rmse     standard      0.433  
2 rsq      standard      0.0892  
3 mae      standard      0.377
```

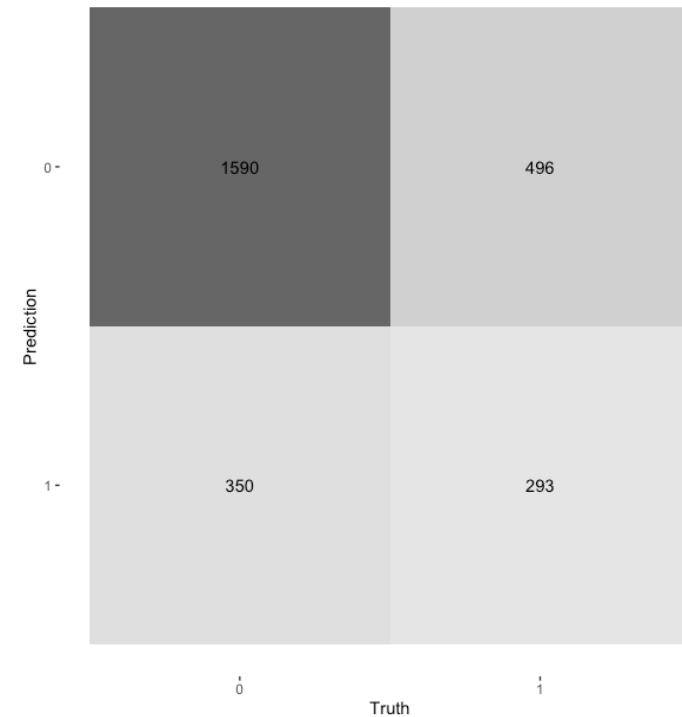
PREDICTIVE MODELLING

K Nearest-Neighbor Model

- usou-se `recipe()` como pré-processamento e `sem recipe()`, ambos com a mesma matriz de confusão, e, 69% de acurácia.

```
model_knn<-  
nearest_neighbor(mode="classification")
```

```
knn_fit <- model_knn %>%  
  fit(intentional_cause ~ district +  
    TemperatureCMax + WindkmhInt +  
    TemperatureCAvg + TemperatureCMin +  
    village_area + extinction_hour + farming_area +  
    village_veget_area, data = f_train, na.action =  
    na.exclude)
```



PREDICTIVE MODELLING

```
path <- paste(getwd(),
"/Rdata/Test_Data_noNa.rds",sep = "")
```

```
fire_Test_Data <- readRDS(path)
```

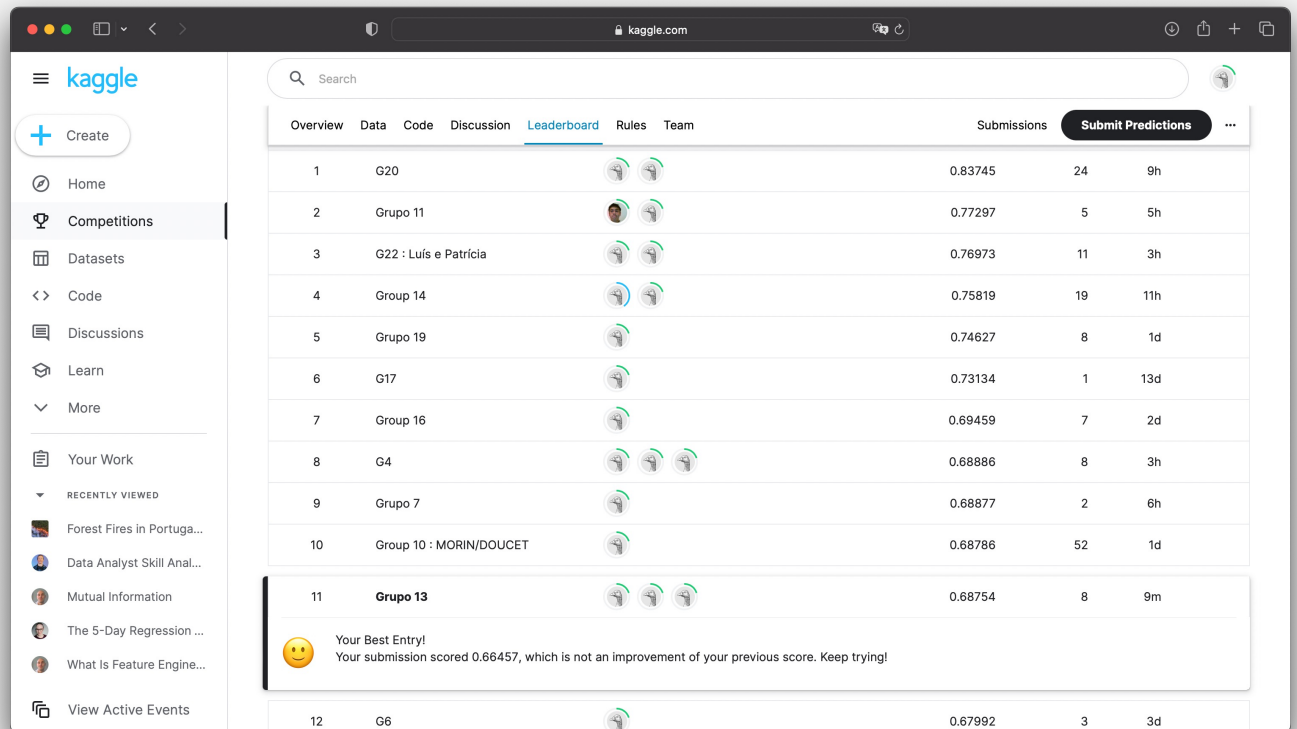
```
prevLG <- predict(glm_lasso_fit, fire_Test_Data)
```


```
names(prevLG)[length(names(prevLG))<-
"intentional_cause"]
```

```
prevLG$id <- fire_Test_Data$id
```

```
prevLG <- prevLG[c("id", "intentional_cause")]
```

```
write.csv(prevLG, "grupo13_DMI_13LG.csv",
row.names=FALSE)
```



Overview	Data	Code	Discussion	Leaderboard	Rules	Team	Submissions	Submit Predictions	...
1	G20						0.83745	24	9h
2	Grupo 11						0.77297	5	5h
3	G22 : Luis e Patricia						0.76973	11	3h
4	Group 14						0.75819	19	11h
5	Grupo 19						0.74627	8	1d
6	G17						0.73134	1	13d
7	Group 16						0.69459	7	2d
8	G4						0.68886	8	3h
9	Grupo 7						0.68877	2	6h
10	Group 10 : MORIN/DOUCET						0.68786	52	1d
11	Grupo 13						0.68754	8	9m
 Your Best Entry! Your submission scored 0.66457, which is not an improvement of your previous score. Keep trying!									
12	G6						0.67992	3	3d

CONCLUSÃO

- Para ter um melhor resultado, seria melhor ter uma outra abordagem aos dados, e, ao seu processamento.
- Ainda que com algumas dificuldades, foi possível prever resultados e publicá-los.
- De salientar, que foi possível aplicar alguns modelos de previsão e, contudo, é de salientar a dificuldade em obter bons modelos de previsão.