

Forest Fires in Portugal - What Are The Causes?

Practical Assignment of Data Mining I

By Robson Teixeira, Eduardo Rodrigues and Claudio Rocha

M:CC – FCUP, 10/01/2021

Contents

1	Introduction	2
2	Problem Definition	3
3	Forest Fire Dataset	4
3.1	Dataset Variables	6
4	Exploratory data analysis and engineering	8
4.1	Data Cleaning and Transforms:	8
4.2	Statistical exploration of the variables	10
4.3	Feature Engineering	10
4.4	Data exploration	10
4.5	Feature Selection	17
5	Models	18
5.1	Evaluation of the models used	18
6	Conclusions and Prospects	19
7	References	20

Chapter 1

Introduction

In this project, we try to find the best machine learning model that more accurately predicts whether a forest fire occurs negligently, intentionally, naturally or recurrently. From a database that was given to us, we divided the work into several parts.

The remainder of this report is organized as follows: in chapter 2, we describe the importance of predicting forest fires that are a big problem actually; in chapter 3 is described the causes of the occurrences that is the variable that the model will be predict and a table with all the variables of the original dataset; chapter 4 is dedicated to the exploration, cleaning and engineering of the data. Some graphics are plotted in this chapter and they help to visualize the origins and locations of the forest fires; the models used to test the dataset are described and compared in chapter 5; in chapter 6, We finalize with the main conclusions and the last chapter includes references.

Chapter 2

Problem Definition

Forest fires are a very important issue that negatively affects climate change. Typically, the causes of forest fires are those oversights, accidents and negligence committed by individuals, intentional acts and natural causes. The latter is the root cause for only a minority of the fires.

Their harmful impacts and effects on ecosystems can be major ones. Among them, we can mention the disappearance of native species, the increase in levels of carbon dioxide in the atmosphere, earth's nutrients destroyed by the ashes, and the massive loss of wildlife.

Data mining techniques can help in the prediction of the cause of the fire and, thus, better support the decision of taking preventive measures in order to avoid tragedy. In effect, this can play a major role in resource allocation, mitigation and recovery efforts.

Chapter 3

Forest Fire Dataset

The ICFN - Nature and Forest Conservation Institute has the record of the list of forest fires occurred in Portugal for several years. For each fire, there is information such as the site, the alert date/hour, the extinction date/hour, the affected area and the cause type. A classifications for causes types are presented in table @ref(tab:cause_type).

Table 1: (#tab:cause_type) Classifications of causes of forest fires.

Cause	Description
Unknown	absence of sufficient objective evidence to determine the cause of the ignition of fire
Natural	lightning generated in thunderstorms
Negligence	the misguided use of fire in activities such as burning trash, mass burning of agricultural and forest fuels, fun and leisure activities; failure to properly extinguish cigarettes by smokers; the dispersal and transport of incandescent particles from chimneys; etc.
Intentional	incendiarism and arson, mostly resulting from behaviors and attitudes reacting to the constraints of agroforestry management systems and to conflicts related to land use
Rekindling	reburning of an area over which a fire has previously passed, but where fuel has been left that is later ignited by latent heat, sparks, or embers

The dataset used in this study was provided by ICFN, and it contains the data on reported forest fires during 2015 and its respective causes. The data are distributed in files:

- **fires2015train.csv** — the file contain the data of 7511 reported forest fires during 2015

A summary of the structure of the dataset is provided below.

```
## Rows: 7,511
## Columns: 21
```

```
## $ id <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1...
## $ region <chr> "Entre Douro e Minho", "Entre Douro e Minho", "T...
## $ district <chr> "Viana do Castelo", "Porto", "Vila Real", "Vila ...
## $ municipality <chr> "Ponte de Lima", "Marco de Canaveses", "Boticas"...
## $ parish <chr> "Serdedelo", "Vila Boa de Quires", "Cerdedo", "G...
## $ lat <chr> "41:44:48.5663999999878'", "41:12:58.4280000000...
## $ lon <chr> "8:31:12.32760000000027'", "8:12:28.3788000000002...
## $ origin <chr> "fire", "fire", "fire", "firepit", "firepit", "f...
## $ alert_date <chr> "2015-03-24", "2015-03-24", "2015-03-24", "2015-...
## $ alert_hour <chr> "17:01:00", "17:10:00", "21:40:00", "16:00:00", ...
## $ extinction_date <chr> "2015-03-24", "2015-03-24", "2015-03-25", "2015-...
## $ extinction_hour <chr> "18:09:00", "18:47:00", "05:45:00", "17:00:00", ...
## $ firstInterv_date <chr> "2015-03-24", "2015-03-24", "2015-03-24", "2015-...
## $ firstInterv_hour <chr> "17:10:00", "17:16:00", "22:00:00", "16:14:00", ...
## $ alert_source <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ village_area <dbl> 2.50, 0.00, 0.50, 0.00, 0.10, 0.00, 0.35, 0.50, ...
## $ vegetation_area <dbl> 0.000, 1.350, 38.000, 0.010, 0.000, 0.100, 14.82...
## $ farming_area <dbl> 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, ...
## $ village_veget_area <dbl> 2.500, 1.350, 38.500, 0.010, 0.100, 0.100, 15.17...
## $ total_area <dbl> 2.5000, 1.3500, 38.5000, 0.0100, 0.1000, 0.1000,...
## $ cause_type <chr> "negligent", "negligent", "negligent", "negligen...
```

```
## # A tibble: 6 x 21
```

```
##       id region district municipality parish lat lon origin alert_date
##   <int> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
## 1     1 Entre~ Viana d~ Ponte de Li~ Serde~ 41:4~ 8:31~ fire 2015-03-24
## 2     2 Entre~ Porto Marco de Ca~ Vila ~ 41:1~ 8:12~ fire 2015-03-24
## 3     3 Trás~ Vila Re~ Boticas Cerde~ 41:3~ 07:5~ fire 2015-03-24
## 4     4 Trás~ Vila Re~ Montalegre Gralh~ 41:5~ 7:42~ firep~ 2015-03-25
## 5     5 Trás~ Vila Re~ Valpaços Alger~ 41:3~ 07:2~ firep~ 2015-03-12
## 6     6 Entre~ Vila Re~ Mondim de B~ Ermelo 41:2~ 07:5~ firep~ 2015-03-13
## # ... with 12 more variables: alert_hour <chr>, extinction_date <chr>,
## # extinction_hour <chr>, firstInterv_date <chr>, firstInterv_hour <chr>,
## # alert_source <lgl>, village_area <dbl>, vegetation_area <dbl>,
## # farming_area <dbl>, village_veget_area <dbl>, total_area <dbl>,
## # cause_type <chr>
```

```
##       variable q_zeros p_zeros q_na p_na q_inf p_inf type unique
## 1          id         0    0.00    0  0.00    0    0 integer    7511
## 2         region         0    0.00  501  6.67    0    0 character     10
## 3       district         0    0.00    0  0.00    0    0 character     19
## 4   municipality         0    0.00    0  0.00    0    0 character    297
## 5          parish         0    0.00    0  0.00    0    0 character   2270
## 6           lat         0    0.00    0  0.00    0    0 character   5858
```

```
## 7          lon      0    0.00    0    0.00    0    0 character 5867
## 8         origin    0    0.00    0    0.00    0    0 character    5
## 9        alert_date  0    0.00    0    0.00    0    0 character  317
## 10       alert_hour  0    0.00    0    0.00    0    0 character 1312
## 11    extinction_date  0    0.00    9    0.12    0    0 character  319
## 12    extinction_hour  0    0.00    9    0.12    0    0 character 1201
## 13 firstInterv_date    0    0.00  214    2.85    0    0 character  318
## 14 firstInterv_hour    0    0.00  215    2.86    0    0 character 1202
## 15      alert_source    0    0.00 7511 100.00    0    0   logical    0
## 16      village_area 5349 71.22    0    0.00    0    0   numeric   591
## 17    vegetation_area 2648 35.25    0    0.00    0    0   numeric 1052
## 18      farming_area 5976 79.56    0    0.00    0    0   numeric   650
## 19 village_veget_area 1413 18.81    0    0.00    0    0   numeric 1377
## 20         total_area    8    0.11    0    0.00    0    0   numeric 1781
## 21         cause_type    0    0.00    0    0.00    0    0 character    4
```

```
## # A tibble: 6 x 21
##       id region district municipality parish lat   lon   origin alert_date
##   <int> <chr>  <chr>      <chr>          <chr> <chr> <chr> <chr>  <chr>
## 1  7506 Trás~ Bragança Mirandela   Carva~ 41:3~ 7:10~ firep~ 2015-07-18
## 2  7507 Beira~ Castelo~ Idanha-a-No~ Oledo 39:5~ 7:20~ firep~ 2015-08-07
## 3  7508 Entre~ Porto    Penafiel    São M~ 41:1~ 8:12~ firep~ 2015-08-08
## 4  7509 Entre~ Porto    Amarante    Telões 41:1~ 8:6::~ firep~ 2015-08-08
## 5  7510 Entre~ Braga    Celorico de Gêmeos 41:2~ 8:0::~ firep~ 2015-08-08
## 6  7511 Ribat~ Santarém Ourém      Nossa~ 39:3~ 8:34~ firep~ 2015-08-08
## # ... with 12 more variables: alert_hour <chr>, extinction_date <chr>,
## #   extinction_hour <chr>, firstInterv_date <chr>, firstInterv_hour <chr>,
## #   alert_source <lgl>, village_area <dbl>, vegetation_area <dbl>,
## #   farming_area <dbl>, village_veget_area <dbl>, total_area <dbl>,
## #   cause_type <chr>
```

3.1 Dataset Variables

Table 3.2 describes all variables contained in the `fires.raw` of the data set. Clearly, the type of some of variables is incorrect and inconvenient for analysis and that was taken care of in section ??.

Table 3.2: List of variables present in the original file from the `fires2015train.csv` data set.

Variable	Type	Description
id	integer	id number
region	character	region name
district	character	district name

Variable	Type	Description
municipality	character	municipality name
parish	character	parish name
lat	character	latitude value
lon	character	longitude value
origin	character	how the fire started
alert_date	character	date when fire started
alert_hour	character	alert hour
extinction_date	character	date of the end of fire
extinction_hour	character	hour of the end of fire
firstInterv_date	character	date of intervention
firstInterv_hour	character	hour of intervention
alert_source	logical	alert source
village_area	numeric	village area affected
alert_source	logical	alert source
village_area	numeric	village area affected
vegetation_area	numeric	vegetation area affected
farming_area	numeric	farming area affected
village_veget_area	numeric	total village+veget affected
total_area	numeric	total area affected
cause_type	character	cause of the fire

Chapter 4

Exploratory data analysis and engineering

4.1 Data Cleaning and Transforms:

We started to verify the dataset to identify and correct the mistakes and errors in the data. First we tested if there were duplicated observations.

```
## [1] 0
```

The result was negative.

The lat variable had to be corrected because we observed that incorrect values like a date were mixed in it.

The number of observations with wrong value in lat variable were detected and the total number of them is showed below.

```
## [1] "There are 38 observations with wrong value '1900-01-01'"
```

On the wrong values, an imputation was made, based on another observation that has the same region, district, municipality and parish than these.

A cleaning was made on lat and lon variables to remove some characters and change to ". This process was necessary to convert their contents from GPS coordinate to decimals.

After the imputation, 8 NA's values were assigned to observations in latitude and longitude variables

For parish with missing values we insert latitude and longitude

Alentejo - Évora - Mora - Cabeção

Alentejo - Évora - Montemor-o-Novo - Cortiçadas de Lavre

Alentejo - Évora - Montemor-o-Novo - Cíborro

Alentejo - Évora - Mourão - Granja

Alentejo - Évora - Évora - Horta das Figueiras

Alentejo - Évora - Montemor-o-Novo - Cortiçadas de Lavre

Alentejo - Évora - Estremoz - São Lourenço de Mamporcão

Alentejo - Évora - Mora - Brotas

Data imputation: firstInterv_date and firstInterv_hour

Changing type of some variables to factor

We create new features using the alerts that can be used on the analysis.

##	variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
## 1	id	0	0.00	0	0.00	0	0	integer	7511
## 2	region	0	0.00	501	6.67	0	0	character	10
## 3	district	0	0.00	0	0.00	0	0	character	19
## 4	municipality	0	0.00	0	0.00	0	0	character	297
## 5	parish	0	0.00	0	0.00	0	0	character	2270
## 6	lat	0	0.00	0	0.00	0	0	character	5858
## 7	lon	0	0.00	0	0.00	0	0	character	5867
## 8	origin	0	0.00	0	0.00	0	0	character	5
## 9	alert_date	0	0.00	0	0.00	0	0	character	317
## 10	alert_hour	0	0.00	0	0.00	0	0	character	1312
## 11	extinction_date	0	0.00	9	0.12	0	0	character	319
## 12	extinction_hour	0	0.00	9	0.12	0	0	character	1201
## 13	firstInterv_date	0	0.00	214	2.85	0	0	character	318
## 14	firstInterv_hour	0	0.00	215	2.86	0	0	character	1202
## 15	alert_source	0	0.00	7511	100.00	0	0	logical	0
## 16	village_area	5349	71.22	0	0.00	0	0	numeric	591
## 17	vegetation_area	2648	35.25	0	0.00	0	0	numeric	1052
## 18	farming_area	5976	79.56	0	0.00	0	0	numeric	650
## 19	village_veget_area	1413	18.81	0	0.00	0	0	numeric	1377
## 20	total_area	8	0.11	0	0.00	0	0	numeric	1781
## 21	cause_type	0	0.00	0	0.00	0	0	character	4

-----XXXXXXXXXXXXXXXXXXXX-----

4.2 Statistical exploration of the variables

4.3 Feature Engineering

Deriving new variables from available data and merging datasets WEATHER DATA AND FOREST FIRES getting new variables that can help on the predictions. We insert some new variables like TAVG, TMIN, TMAX and PRCP and did an imputation of value in tavg variable based on tmax and tmin, in tavg variable based on tavg15d (15 days before each occurrence), in tavg15d variable if NaN, in tmax variable based on tavg and tmin, in tmin variable based on tavg and tmax. We eliminated the maximum of observations with zeros and the dataset stayed like below:

##	variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
## 1	district	0	0.00	0	0	0	0	factor	19
## 2	origin	0	0.00	0	0	0	0	factor	5
## 3	alert_month	0	0.00	0	0	0	0	factor	12
## 4	alert_period	0	0.00	0	0	0	0	factor	4
## 5	duration	4	0.05	0	0	0	0	numeric	598
## 6	village_area	5342	71.21	0	0	0	0	numeric	590
## 7	vegetation_area	2646	35.27	0	0	0	0	numeric	1051
## 8	farming_area	5968	79.55	0	0	0	0	numeric	650
## 9	tavg	8	0.11	0	0	0	0	numeric	335
## 10	tavg15d	56	0.75	0	0	0	0	numeric	1549
## 11	prcp	4694	62.57	0	0	0	0	numeric	138
## 12	cause_type	0	0.00	0	0	0	0	factor	4

4.4 Data exploration

Based on the dataset we plotted some graphics that helped us to get some conclusions and showed a general notion about the problem of the forests fires.

Figure 4.1 depicts the bar graphic of the distribution of forests fires during 2015. The x-axis represents the months along the year 0f 2015 and the y-axis represents the total of fires that occurred by month.

This graphic showed us that july and august are the months with the largest occurrences and the period between march and september needs more attention. Probably we will consider the variable month as important to the analisys.

In another graphic, figure 4.2 we plotted the bar graphic of the distribution of forests fires by region. The x-axis represents the regions and the y-axis represents the total of fires that occurred by region.

Observing this grafic we saw that Entre Douro e Minho was the region with more forests fires and other regions like Centro, Lisboa and Norte were with minimum occurrences.

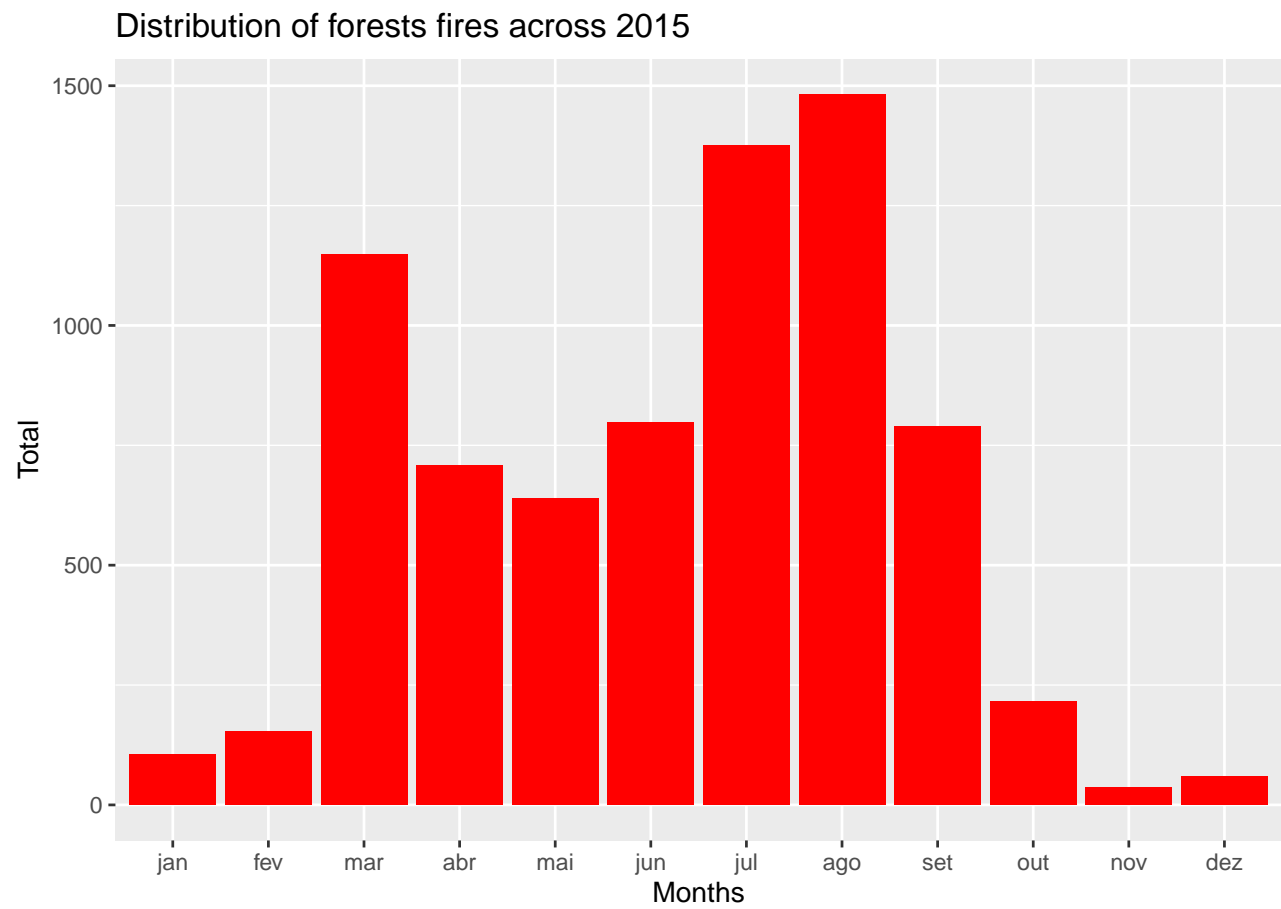


Figure 4.1: Barplot of the distribution of forests fires during 2015.

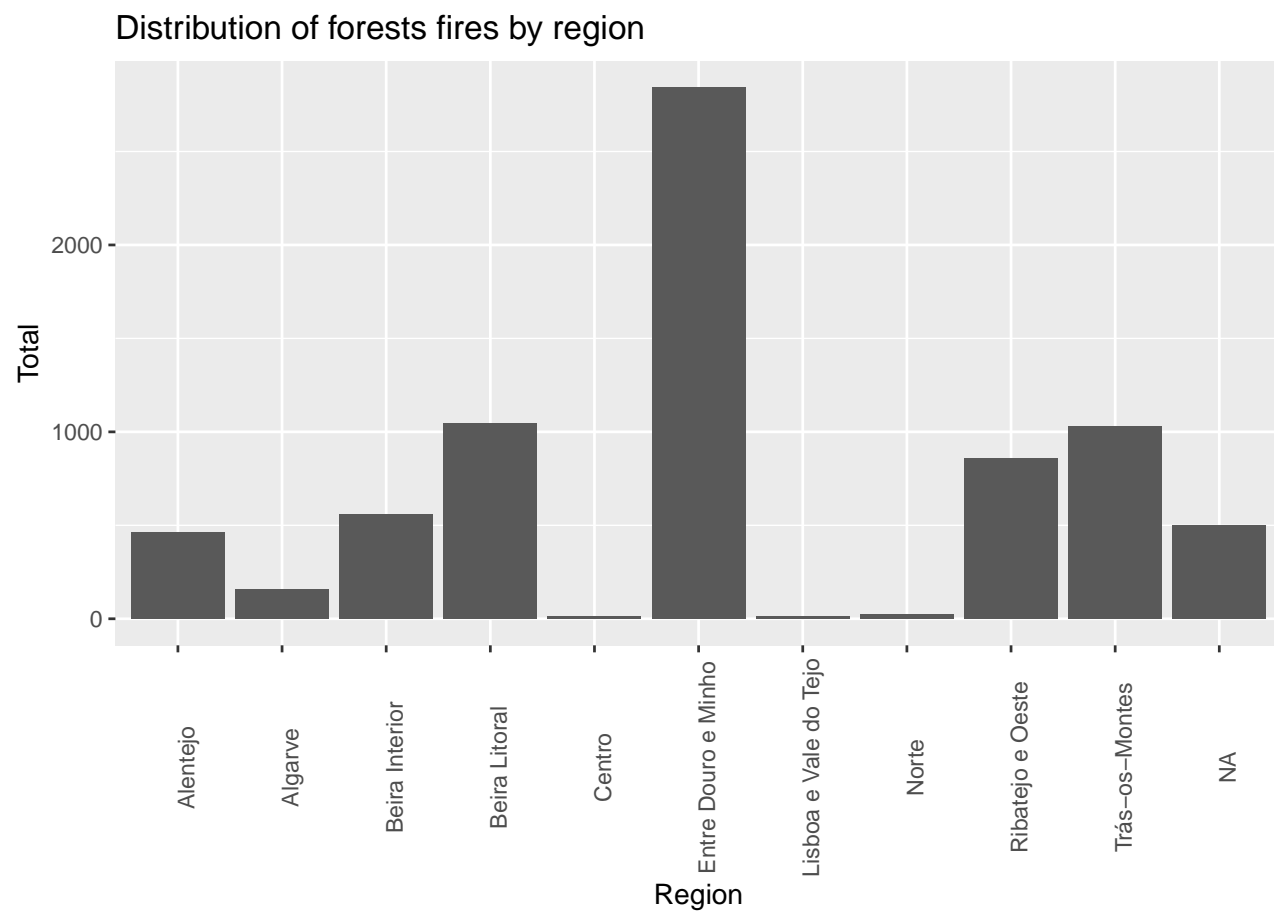


Figure 4.2: Barplot of the distribution of forests fires by regions.

Another important graphic is the figure 4.3 that corresponds to the distribution of forests fires by district. The x-axis represents the districts and the y-axis represents the total of fires that occurred by region.

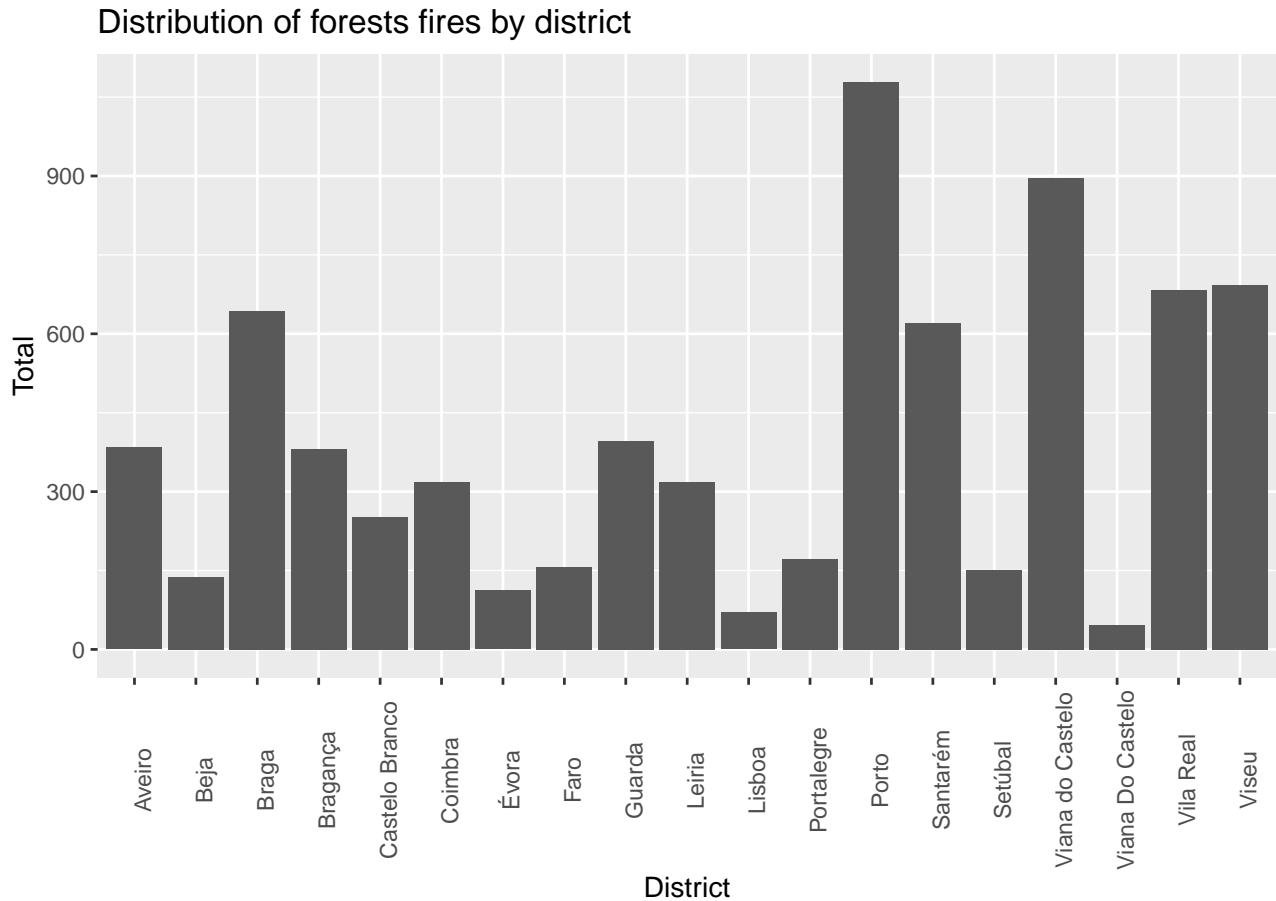


Figure 4.3: Barplot of the distribution of forests fires by districts.

This graphic indicates that Porto and Viana do Castelo were the districts with more forests fires.

A bar plot of forests fires related with the causes of them, is being reported on figure 4.4. On the x-axis are listed the different causes and on y-axis represents the total of fires that occurred.

A thing that calls our attention is the difference between the number of fires that were caused by negligence and by natural causes. The number of forest fires caused intentionally were almost the half of the negligent causes what is an alarmant number.

We plotted too a bar plot of forests fires during 2015 by origin, figure 4.5. On the x-axis are listed the different origins and on y-axis represents the total of fires that occurred.

The firepit was the origin of the most forests fires comparing it with the other origins.

The relationship between district, month and causes is represented on figure 4.6. The x-axis includes the different districts, y-axis represents the months of occurrences and the variable cause is showed by colours listed on the labels.

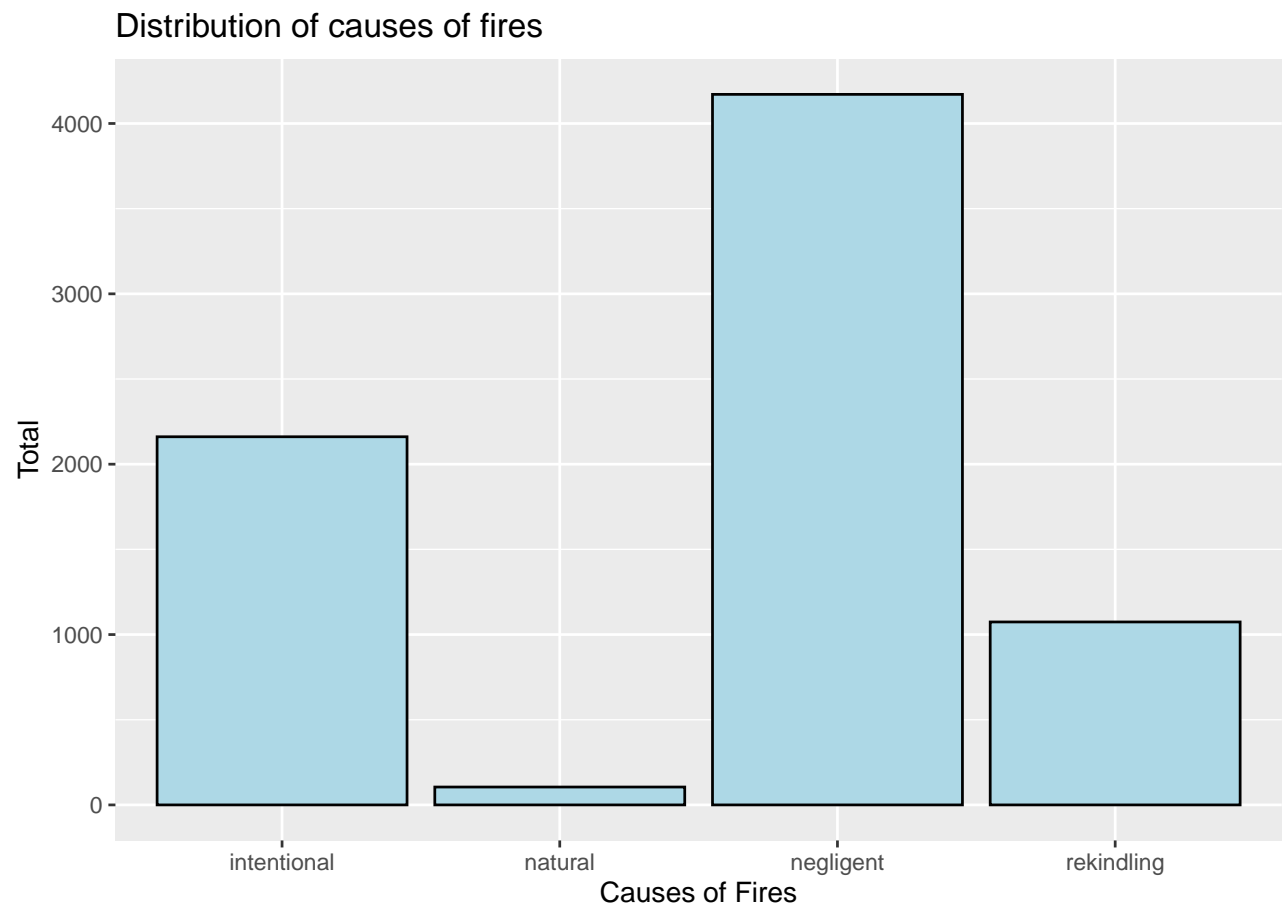


Figure 4.4: Barplot of the distribution of forests fires by causes.

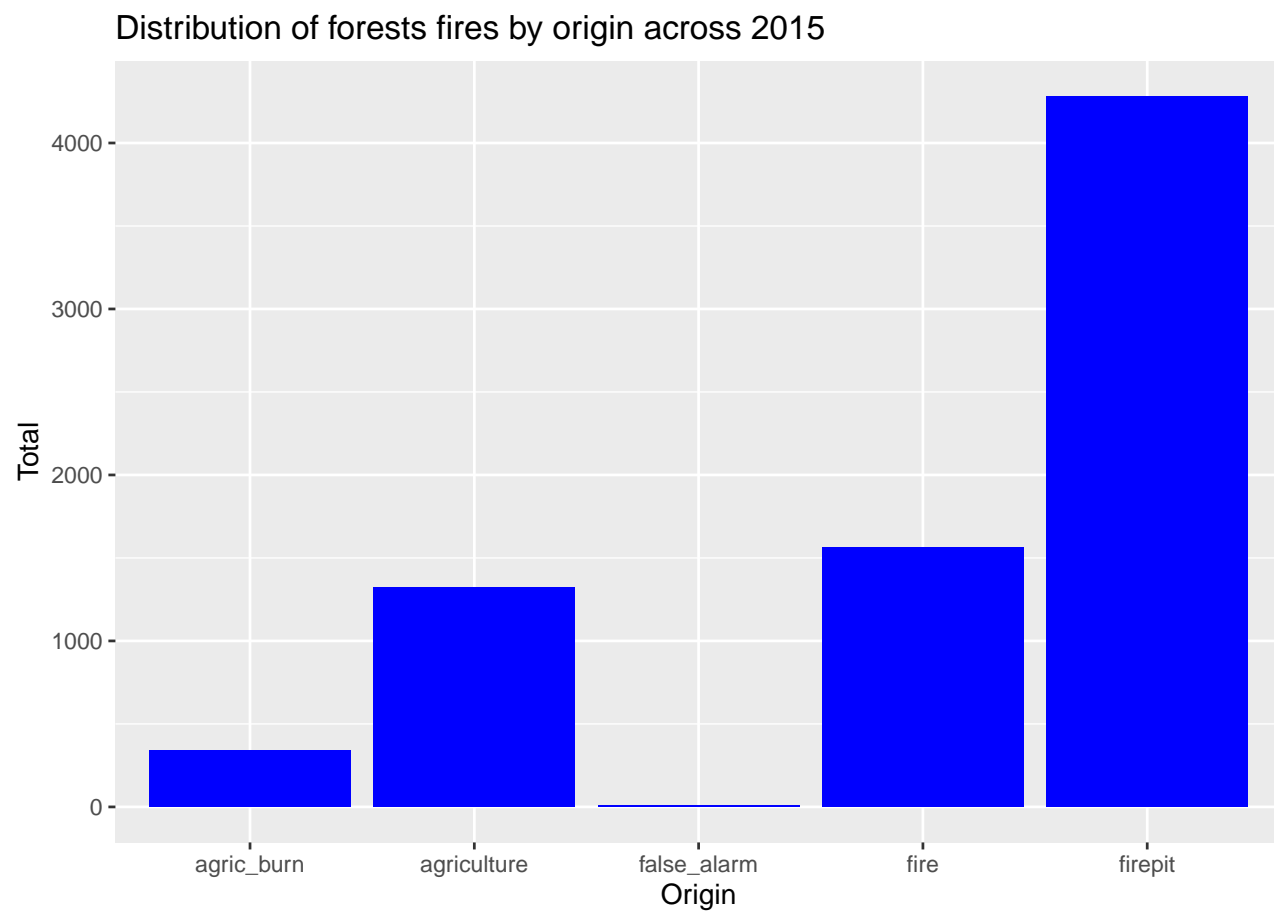


Figure 4.5: Barplot of the distribution of forests fires by origins.

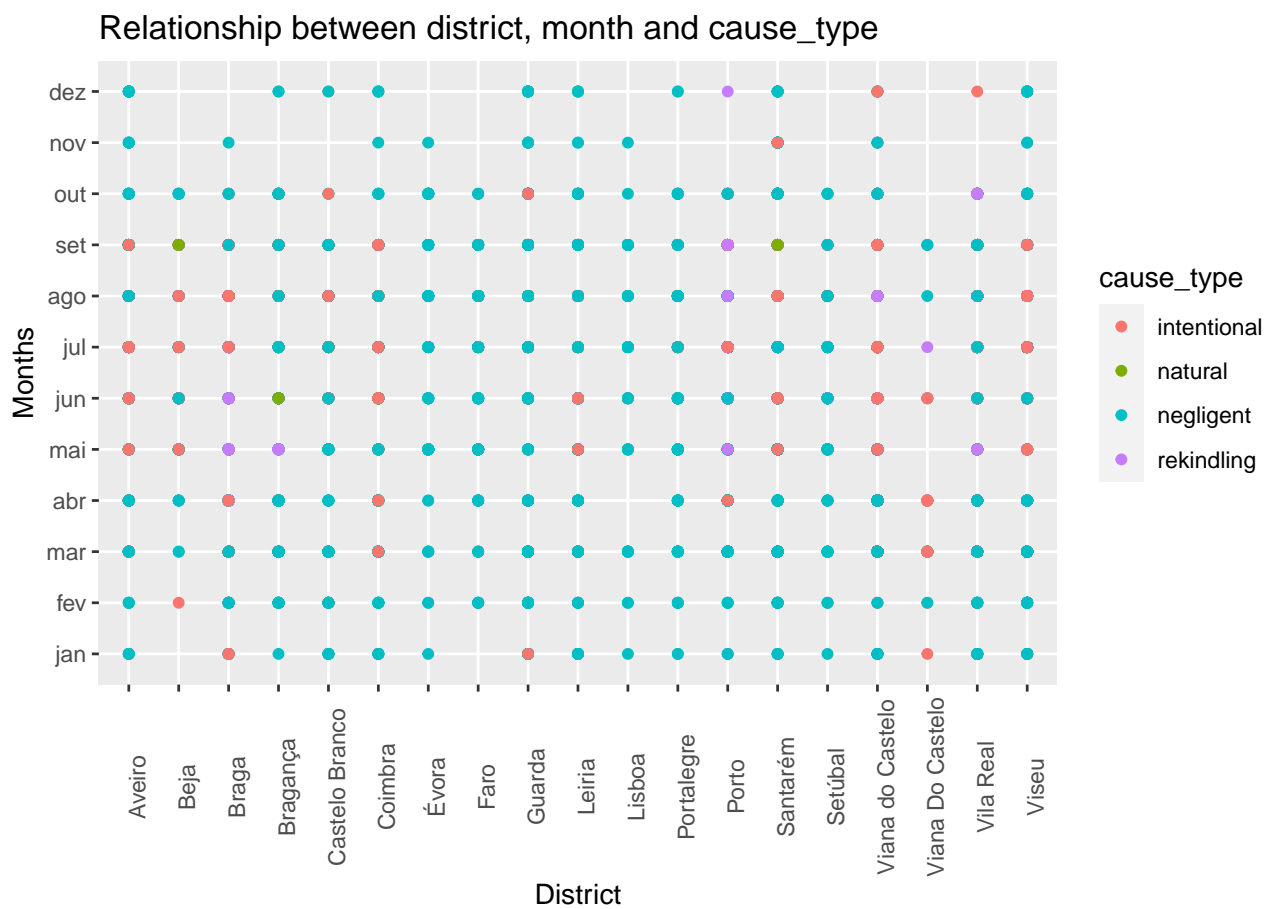


Figure 4.6: Distribution of forests fires relating district, month and causes.

The relationship between region, month and causes is represented on figure 4.7. The x-axis includes the different regions, y-axis represents the months of occurrences and the variable cause is showed by colours listed on the labels.

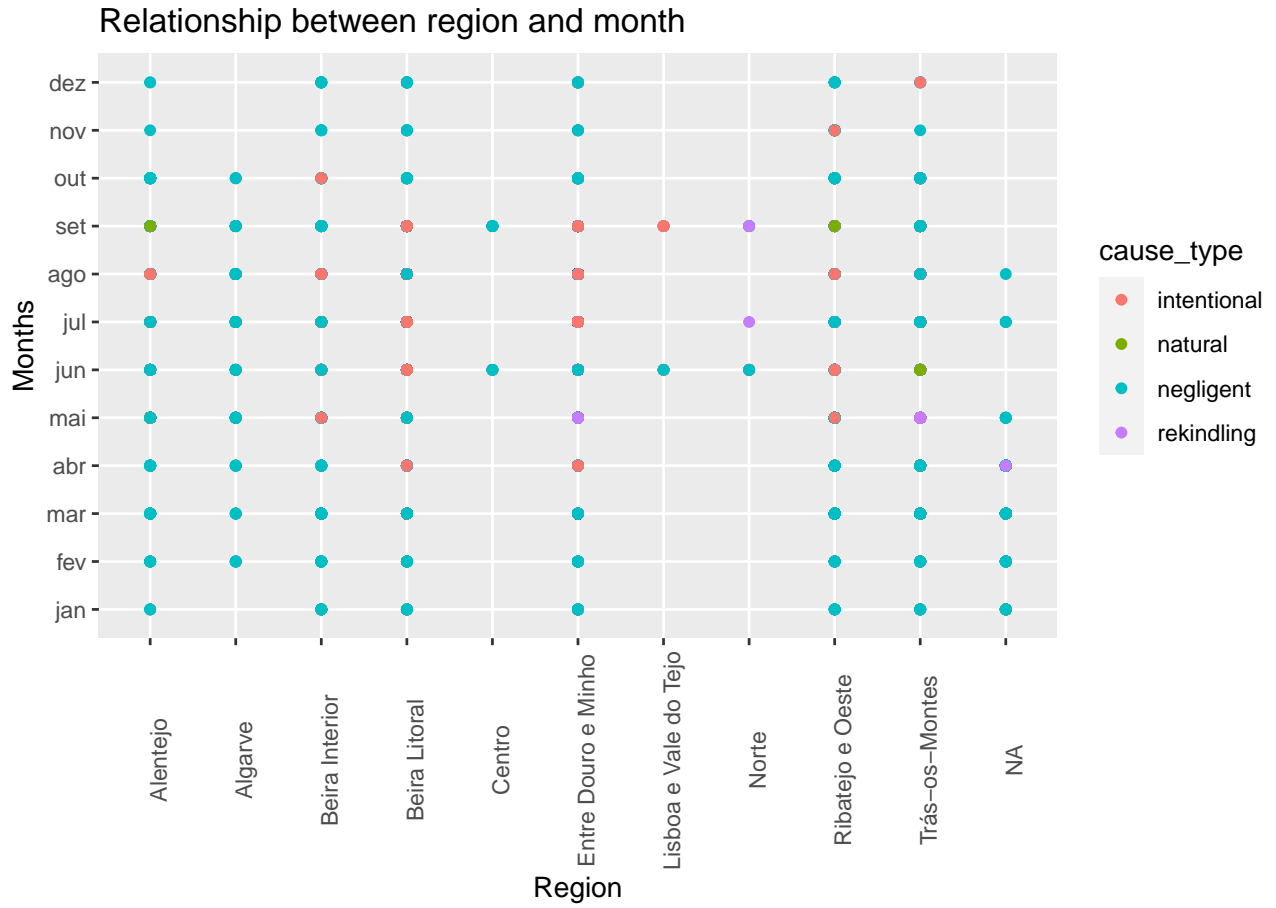


Figure 4.7: Distribution of forests fires relating region, month and causes.

4.5 Feature Selection

Chapter 5

Models

5.1 Evaluation of the models used

Chapter 6

Conclusions and Prospects

Chapter 7

References