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

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 ofnthe ocurrences 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 Institute for Nature Conservation and Forests (ICNF) is the governmental body responsible for the nature and forest policies, including the management of protected areas and state managed national, municipal, and communal forests of mainland Portugal. The ICNF has been maintained a database with data of all forest fires that occurred in Portugal over several years. The data set used in this study is a subset extracted from this database regarding the fires that occurred over 2015. It consist of **7511** records of fires and for each one, there is relevant information such as the GPS coordinates (latitude and longitude) where occur the fire, the date and time of fire alert, the date and time of the first intervention, and the date and time of fire extinction, besides the origin of the ignition, the affected area, and the cause type. The table 3 describes all variables contained in `Forest Fires` data set:

Table List of variables in `Forest Fires` data set.

Variable	Type	Description
id	integer	id number
region	character	region name
district	character	district name
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
vegetation_area	numeric	vegetation area affected

Variable	Type	Description
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

A classification for causes types are presented in table 3.2.

Table 3.2: Classifications of causes of forest fires.

Cause	Description
Unknown	absence of suficient 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 theconstraints 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

A glimpse of the structure of the `Forest Fires` data set is provided below:

Table: A glimpse of the structure of the data set.

```
## 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> "8:31:12.3276000000027’’", "8:12:28.378800000002...
## $ 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, ...
## $ 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 summary for each variable present in dataset is provided below. The metrics displayed are: quantity and percentage of zeros, quantity and quantity and percentage of NA's, data type and quantity of unique values.

Table 4: A summary of variables of the dataset.

```
##              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 sample the first observations is provided below:

```
## # 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>
```

Chapter 4

Data Preparation

Data preparation consists of the process of cleaning and transforming raw data in a form that can be used by machine learning algorithms. Next sections, we exploit the `Forest Fires` dataset in order to perform the steps of cleaning a transforming, when need.

4.1 Data Cleaning

4.1.1 Latitude and Longitude

The `Forest Fires` dataset store the latitude and longitude of the place where occurred the fire into variables `lat` e `lon` respectively. These values are in format of *Degreesº Minutes'Seconds"* and for the reason contain special characters `º`, `'`, `:` and `"`. Besides, there are wrong values into variables as dates between the coordinates and values with scientific notation `E-12`, `E-11` and `E-02`. A sample of these inconsistencies is provided in the tables below:

```
## # A tibble: 1 x 2
##   lat                       lon
##   <chr>                     <chr>
## 1 41º41'25.821599999997'' 8º20'37.446000000002''
```

```
## # A tibble: 1 x 2
##   lat                 lon
##   <chr>               <chr>
## 1 1900-01-01 14:19:38 07:30:27
```

```
## # A tibble: 1 x 2
##   lat                          lon
##   <chr>                        <chr>
## 1 38:36:5.11590769747272E-12 8:35:49.9999999999972
```

A cleaning and transformation steps were perfomed on `lat` and `lon` variables to remove the special caracters and scientific notation. For the values wrongs where there is a date among the coordinates, it was performed an data imputation based on another observations that has the same `region`, `district`, `municipality` and `parish`. After the cleaning steps, the values were transformated from GPS coordinates to decimals coordinates in order to be able retrieve historical data from nearest weather stations using the RNOAA package

The data imputation and transformation generated 8 NA's in `lat` and `lon` variables for parishes listed below:

```
## # A tibble: 8 x 6
##   region   district municipality    parish                    lat   lon
##   <chr>    <chr>    <chr>           <chr>                     <chr> <chr>
## 1 Alentejo Évora    Mora            Cabeção                   <NA>  <NA>
## 2 Alentejo Évora    Montemor-o-Novo Cortiçadas de Lavre       <NA>  <NA>
## 3 Alentejo Évora    Montemor-o-Novo Ciborro                   <NA>  <NA>
## 4 Alentejo Évora    Mourão          Granja                    <NA>  <NA>
## 5 Alentejo Évora    Évora           Horta das Figueiras       <NA>  <NA>
## 6 Alentejo Évora    Montemor-o-Novo Cortiçadas de Lavre       <NA>  <NA>
## 7 Alentejo Évora    Estremoz        São Lourenço de Mamporcão <NA>  <NA>
## 8 Alentejo Évora    Mora            Brotas                    <NA>  <NA>
```

In order to fixing this, the latitude and longitude values for these parishes were imputed directly from the localization data retrieved from the internet.

4.1.2 District

Mainland Portugalis is divided into 18 districts and the variable `district` from `Forest Fires` dataset refer the place where occurred the fires. As seen in table @ref(tab:summary_data), this variable has 19 unique values, so there are some inconsistent data. The table below display the unique values for this variable:

```
##  [1] "Viana do Castelo" "Porto"            "Vila Real"        "Bragança"
##  [5] "Braga"            "Portalegre"       "Santarém"         "Viseu"
##  [9] "Guarda"           "Leiria"           "Castelo Branco"   "Aveiro"
## [13] "Évora"            "Faro"             "Coimbra"          "Viana Do Castelo"
## [17] "Lisboa"           "Beja"             "Setúbal"
```

As seen in table above, there are two references for the same district: *Viana do Castelo* and *Viana Do Castelo*. So a step of cleaning was performed into this variable values.

4.1.3 First Intervention and Extinction

The variables `firstInterv_date` and `firstInterv_hour` store the date and time that occured the the first intervention by autorities after the fire alert. As seem in table 4, these variables have a total of NA's values equals 214 and 215, respectively. In order to reduce these quantity, a data imputation were performed based on values of `extinction_date` and `extinction_hour` assumption that if there are values for extinction date and time it because some intervention was realized. After data imputatio the quantity of NA's was reduced to 7 in both variables as can be seen below:

```
##           variable q_zeros p_zeros q_na p_na q_inf p_inf      type unique
## 1 firstInterv_date       0       0    7 0.09     0     0 character    318
## 2 firstInterv_hour       0       0    7 0.09     0     0 character   1209
## 3  extinction_date       0       0    9 0.12     0     0 character    319
## 4  extinction_hour       0       0    9 0.12     0     0 character   1201
```

The remaining quantity of NA's values in `firstInterv_date`, `firstInterv_hour`, `extinction_date`, and `extinction_hour` represent 0.9% and 1.2% respectively of the total of observations. As these values are relatively low, these observations were removed from dataset.

4.1.4 Variable: Alert Source

As can be seen in table below, the variable `alert_source` has 100% of values with NA's, so this variable were removed from dataset.

```
##   variable q_zeros p_zeros q_na p_na q_inf p_inf    type unique
## 1      var       0       0 7511  100     0     0 logical      0
```

4.2 Data Transformation

In order to

Changing type of some variables to factor

Creating new features

Variable alert

Marking a check point

———————————-XXXXXXXXXXXXXXXXXXX————————-

It was necessary to do a cleaning on lat and lon variables and convert their contents from GPS coordinate to decimals.Before the transformation they were like below:

```
## [1] "41.746824"        "41.21623"         "41.6352777777778" "41.851153"
## [5] "41.5897222222222" "41.3505555555556"
```

After the transformation, the values were corrected.

Here are the first lines of lat variable

and here are the first lines of lon variable

———————??????????????????????????????——————— Data imputation: firstInterv_date and firstInterv_hour

Fix data type as factor on the variables below.

Creating new features Variable alert

——————-save(fires.raw, file = “fires.raw.RData”)——

4.3 Data Exploration and Analysis

Understanding the structure of the data, the distribution of the variables, and the relationships between them is fundamental to build a solid model.

Based on the dataset we ploted 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 ocurrences and the period between march and september needs more atention. Probably we will consider the variable month as important to the analisys.

4.3.1 region and district

These two variables represents the areas of the occurrences and we can observe by summary that the variable region has a lot of NA´s (501) that corresponds to 6,67% of the total of lines. The distribution of the occurrences by region can be observed in the grafic 4.2. The y-axis represents the regions and the x-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.

The relationship between region, month and causes is represented on figure 4.3. The x-axis includes the diferent regions, y-axis represents the months of ocurrences and the variable cause is showed by colours listed on the labels.

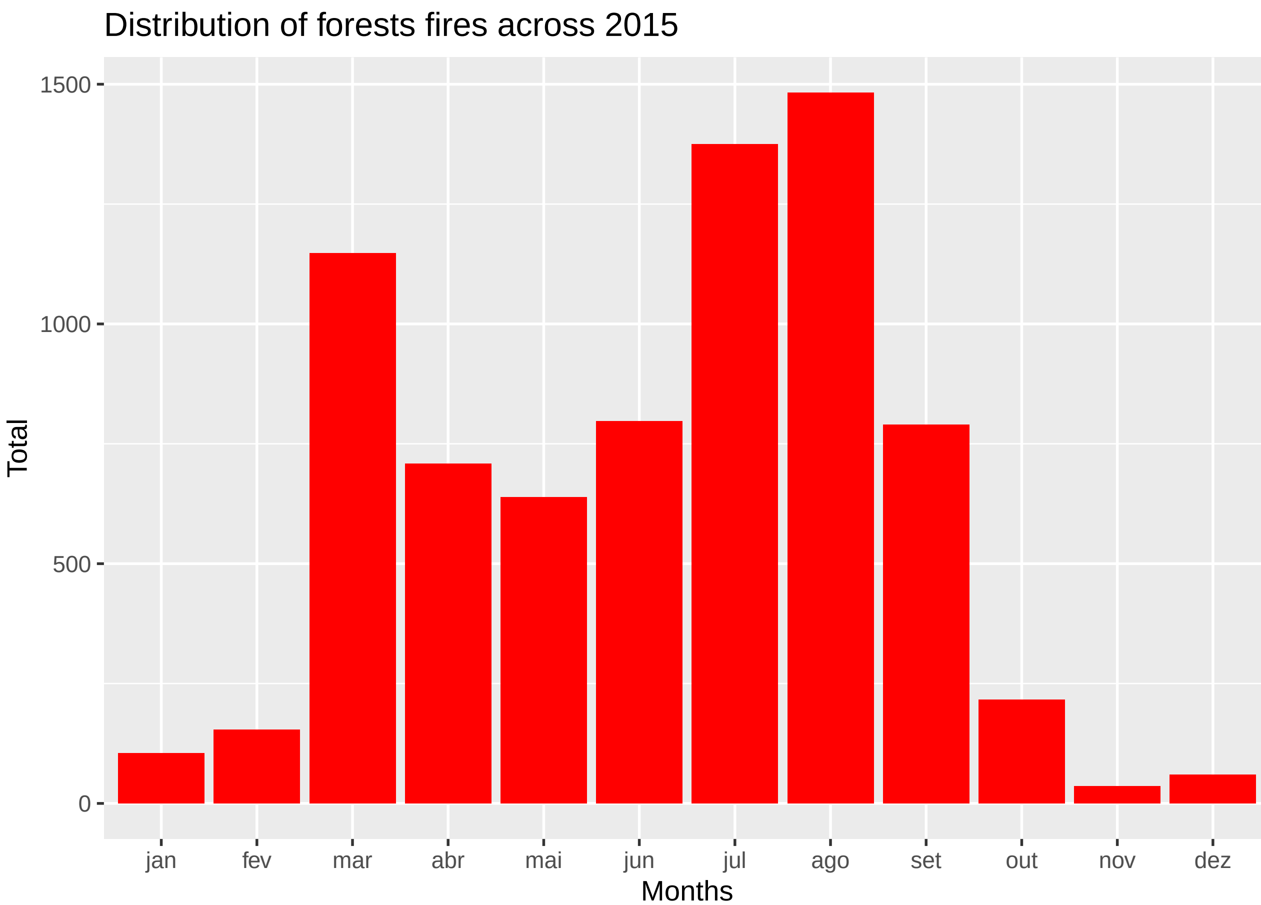

Figure 4.1: Barplot of the distribution of forests fires during 2015.

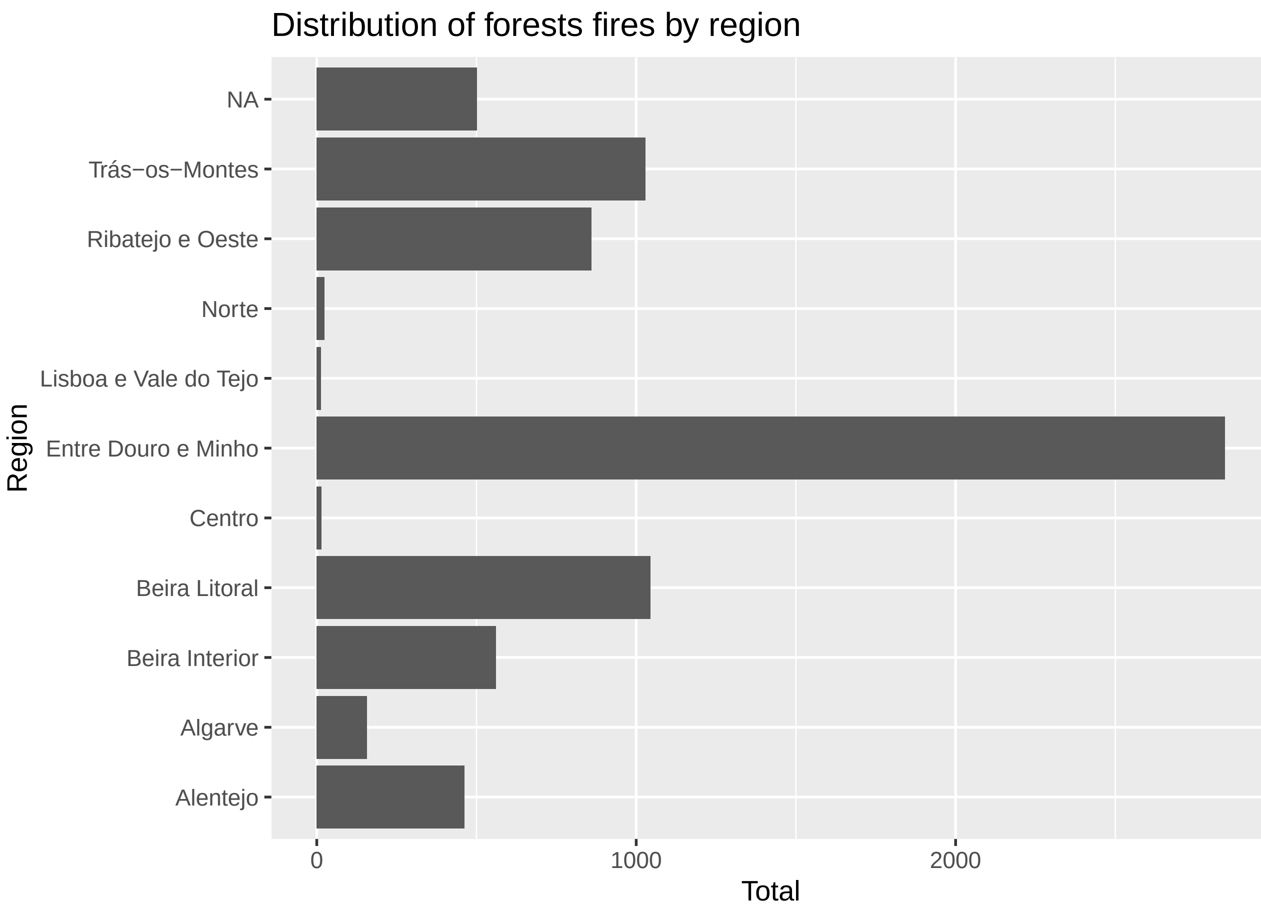

Figure 4.2: Barplot of the distribution of forests fires by regions.

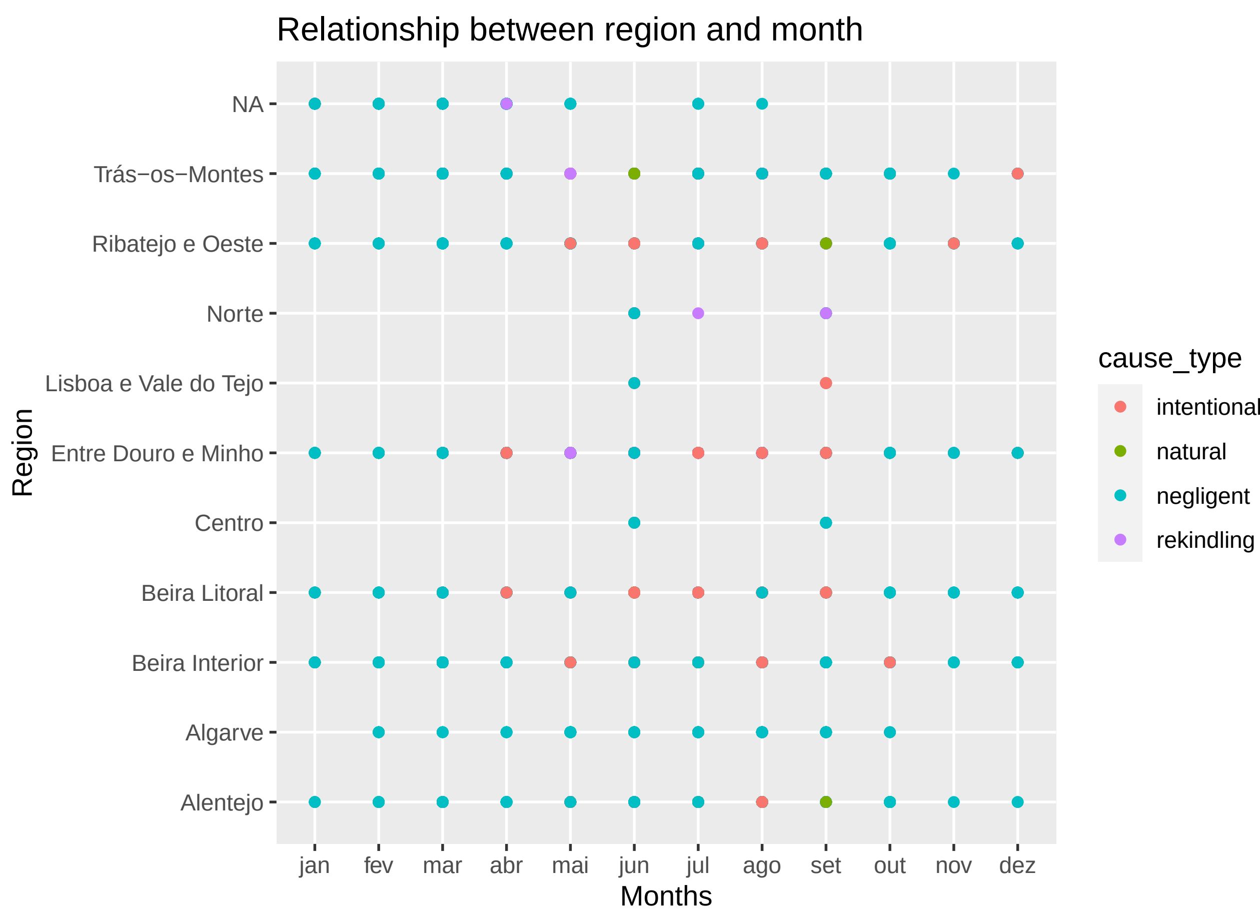

Figure 4.3: Distribution of forests fires relating region, month and causes.

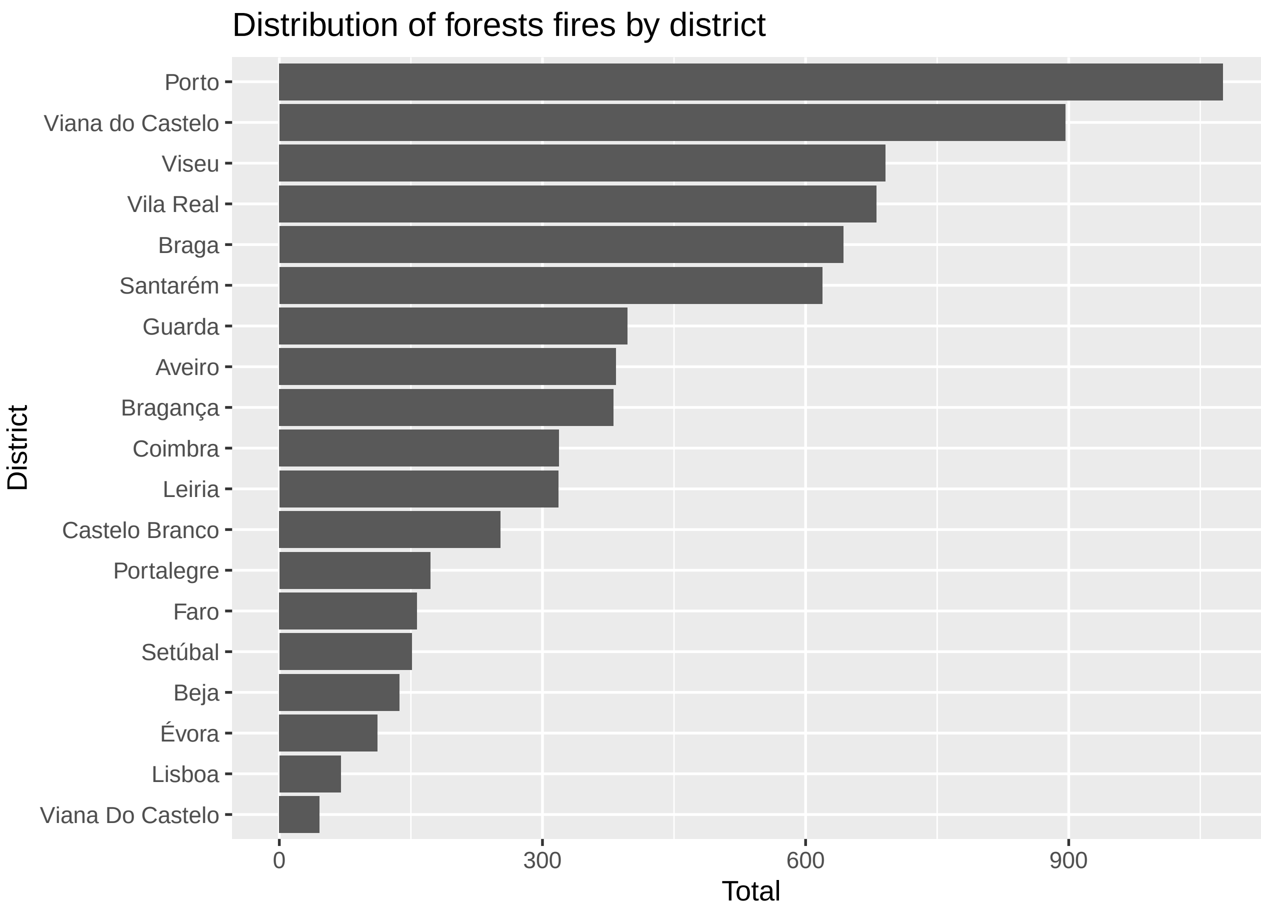

Figure 4.4: Barplot of the distribution of forests fires by districts.

Another important graphic is the figure 4.4 that corresponds to the distribution of forests fires by district. The y-axis represents the districts and the x-axis represents the total of fires that occurred by region.

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

The relationship between district, month and causes is represented on figure 4.5. The x-axis includes the diferent districts, y-axis represents the months of ocurrences and the variable cause is showed by colours listed on the labels.

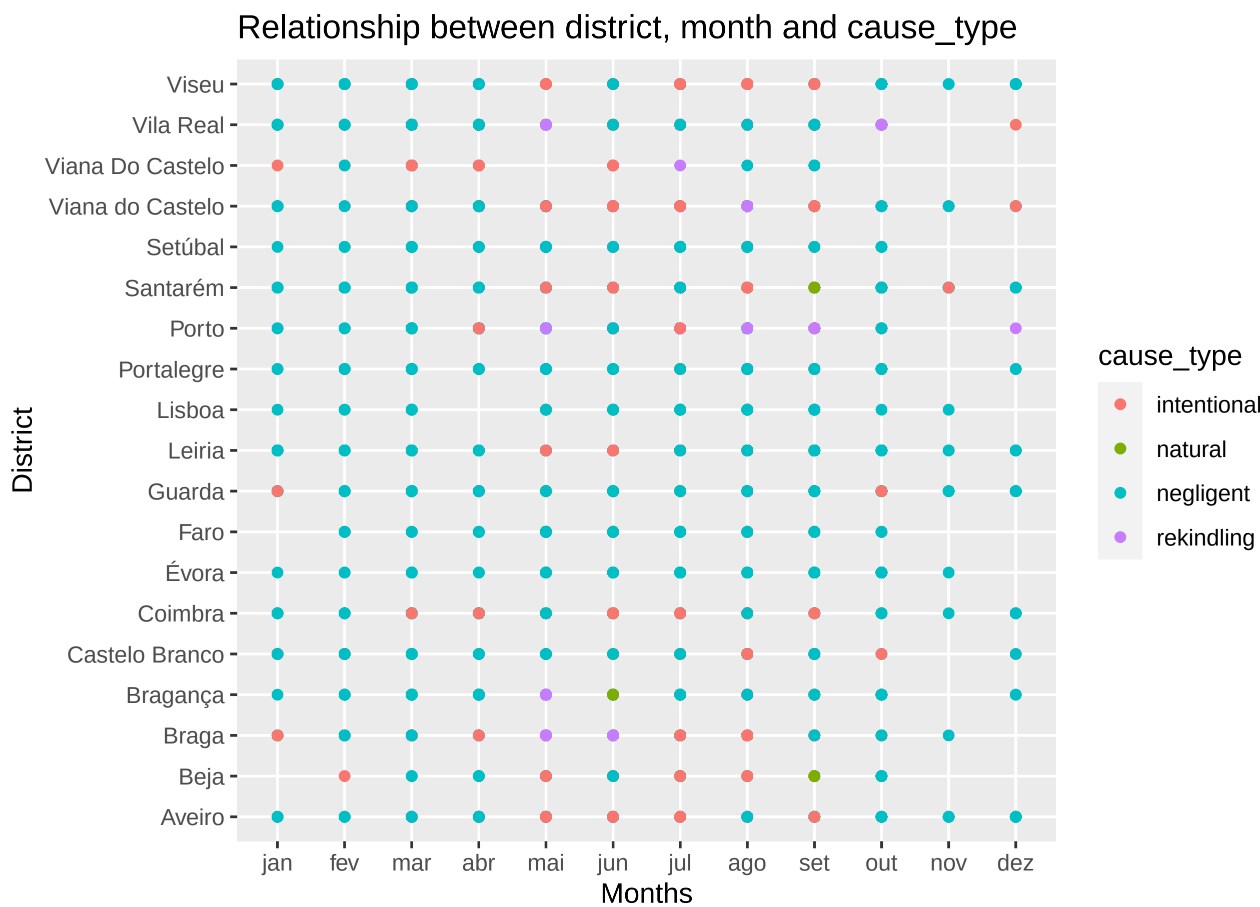

Figure 4.5: Distribution of forests fires relating district, month and causes.

4.3.2 origin

Origin informs the reason that the fire started and apparently appears to be an important observation for evaluation.It can be observed in figure 4.6.

On the x-axis are listed the diferent 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.

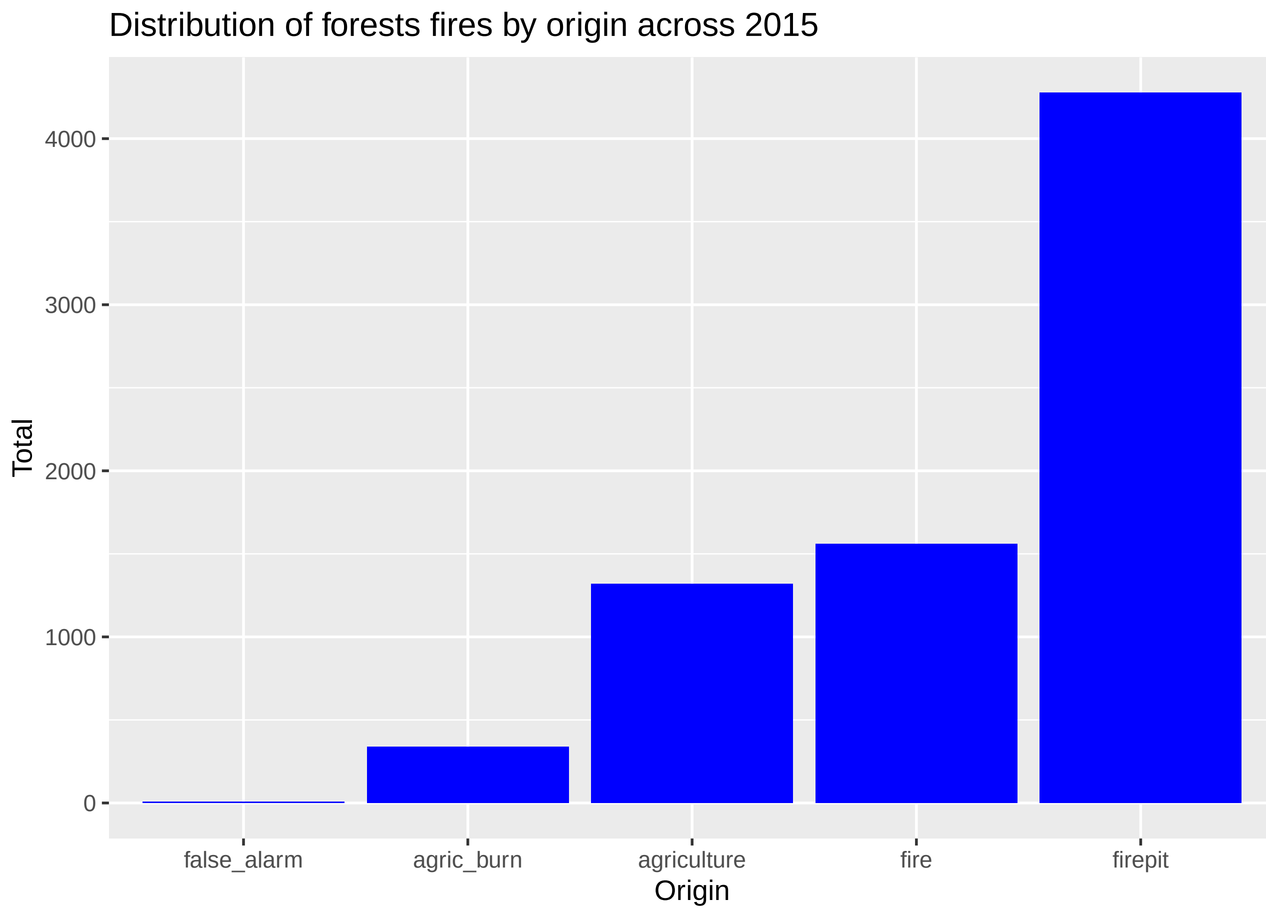

Figure 4.6: Barplot of the distribution of forests fires by origins.

4.3.3 cause_type

This is the variable to be predicted by the model that will be chosen.It shows the four causes of the occurrences: intentional, natural, negligent and rekindling. They can be observed in the graphic 4.7 below. On the x-axis are listed the diferent causes and on y-axis represents the total of fires that occurred.

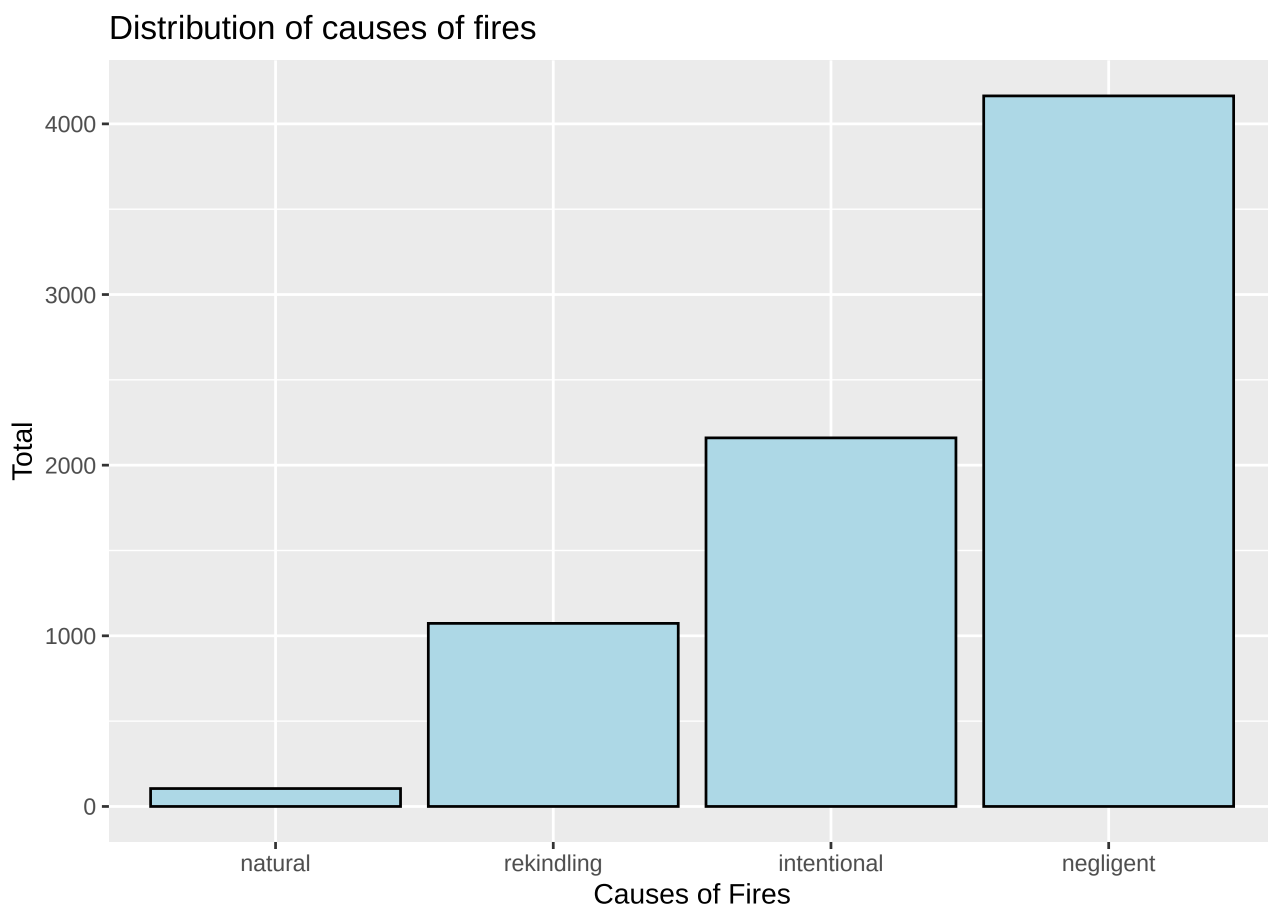

Figure 4.7: Barplot of the distribution of forests fires by causes.

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.

4.4 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 did an imputation of value in tavg variable based on tmax and tmin, in tavg variable based on tavg15d, in tavg15d variable if NaN, in tmax variable based on tavg and tmin, in tmin variable based on tavg and tmax

4.5 Feature Selection

Identifying those input variables that are most relevant to the task. First, we rename the dataset and extracted some variables to minimize it. The variables "parish" and "municipality" have many diferent observations, "id" can be retired because it is only the number of the lines of dataset and we don´t need this column, "region" has 501 NA´s and we don´t have how to substitute them, "alert_source" only has NA´s and it can be removed.

changing to numeric

##Creating train and test datasets

4.6 Dimensionality Reduction

Creating compact projections of the data.

Chapter 5

Prediction Models

In this section was created

5.1 Distance-based Approach

5.1.1 K-Nearest Neighbor

It is an instance-based learning algorithm: it learns by analogy that is, they are based on the notion of similarity between cases.

5.2 Probabilistic Approach

5.2.1 Naive Bayes

According to Bayes' theorem, it is possible to find the probability that a certain event will occur, given the probability of another event that has already occurred. Naive Bayes is a particular class of Bayesian classifiers that predicts the probability that a case belongs to a certain class. Due to its simplicity and high predictive power, it is one of the most used algorithms. This algorithm assumes that there is no dependency relationship between the attributes. However, this is not always possible. The algorithm reads the database and builds a probability table.In Bayesian classification, the main interest is to find the posterior probabilities, the probability of a label given some observed features.

5.3 Mathematical Formulas

5.3.1 Linear Discriminants

5.4 Logical Approaches

5.4.1 Decision Trees

The Decision Tree classification method works as a tree-shaped flowchart, where each node indicates a test done on a value. The connections between the nodes represent the possible values of the upper node test, and the leaves indicate the class to which the record belongs. After the decision tree is assembled, to classify a new record, just follow the flow in the tree starting at the root node until reaching a leaf. Due to the structure they form, decision trees can be converted into Classification Rules.

5.5 Optimization Approaches

5.5.1 Neural Networks

It basically consists of simulating the behavior of neurons. A neural network can be seen as a set of input and output units connected by intermediate layers and each connection has an associated weight. During the learning process, the network adjusts these weights to be able to correctly classify an object.Neural networks can work in ways that do not suffer from wrong values and can also identify patterns for which they have never been trained.

5.5.2 SVM

It is used for both classification and prediction tasks. It consists of separating classes that can be separated by a straight line, called linearly separated classes. The model tries to trace the separation based on the best distance between the closest points. There are variations of SVM, such as the Kernell trick, which allows applying SVM to a set of nonlinearly separable data.

5.6 Ensemble Approaches

5.6.1 Random Forests

Random forest is a supervised learning algorithm which is used mainly used for classification problems. This algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting.

Chapter 6

Conclusions, Shortcomings and Future Work

Chapter 7

Appendix