Forest Fires in Portugal

Inês Aguiar, João Aguiar, José Almeida

15/12/2021

The following report was made for the subject of Data Mining in the year 2021/2022.

Problem definition

The dataset, **fires_train.csv**, provided by ICFN, list all forest fires that occurred in Portugal during 2014 and 2015 along with attributes that have information such as the site, alert date/hour, extinction date/hour, affected area and the cause type which is classified as intentional, natural, negligent or unknown.

Given the provided information, the goal of this practical assignment is to build a machine learning model to predict the cause type of a forest fire: intentional or non-intentional.

In order to provide further information and achieve our goal we also added the following dataset **station_data.RData**.

Data preparation

In order to facilitate our posterior analysis, the imported data present in **fires_train.csv** and **fires_test.csv**, needed to be in an appropriate R format.

After importing the data set to a tibble, we removed attributes and NA's values (NA's only from the train data set) as their presence was deemed, by us, as redundant or irrelevant in order to simplify our data set using available tools provided by libraries such as **na.tools**, **dyplr** and **tidyverse**.

For instance, **alert_source** had all entries as NA's so its column was removed,**extinction_hour** and **firstInter_hour** was irrelevant givin our set goal, and **parish** we considered to be redundant as we are given also **municipality**, **region** and **district** to work with which are in a simpler format that we observed as easier to comprehend.
ID was also irrelevant to our findings and was removed.

Due to some conflict and UTF-8, **district**, **municipality** and **region** we're distorted and had to be cleaned up.
To do this, we manually iterated through the possible names and mutated it with the correct format.
Latitude and **Logitude** we're also in a incorrect format and we used a similar process in order to present the correct format.
This part of the code that corrects the format is extensive and as we progressed more with our work in R, we realized that there are libraries that are specialized in sorting these conflicts. Dates which are in a string type are converted to a numeric date format.

We created columns for **days** and **months**, and in the **months** column, in particular, was converted to their nominal equivalent instead of their numeric format to facilitate the comprehension of data.

In order to further integrate data, we create columns for **max_temperature**, **min_temperature**, **avg_temperature** and **precipitation** that are obtained from additional weather data. We also proceed to remove any lines from these columns with NA's values from train.
After these tweaks and to optimize the run times of our code, we saved our alterations to different csv files, **fires_train_alt.csv** and **fires_test_alt.csv**.

Exploratory data analysis

Given our previous data clean-up and pre-processing, we visualized the data and provide the following insights.

1)Number of Fires by Region

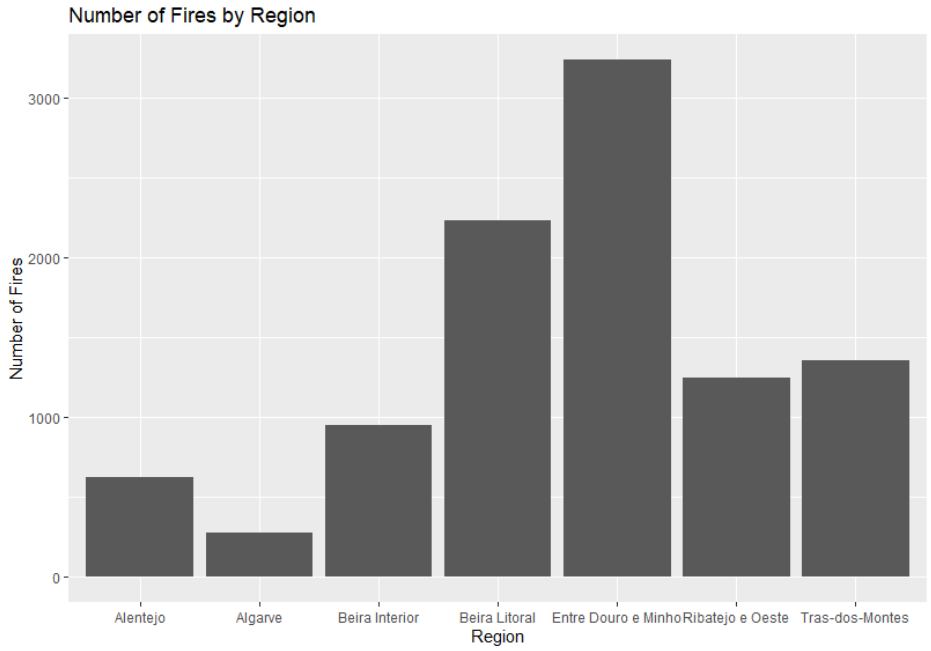


Figure 1: Number of fires per Region

As a group we questioned which presented a higher number of fires. Which left us wondering what districts had the most.

2)Number of Fires by District

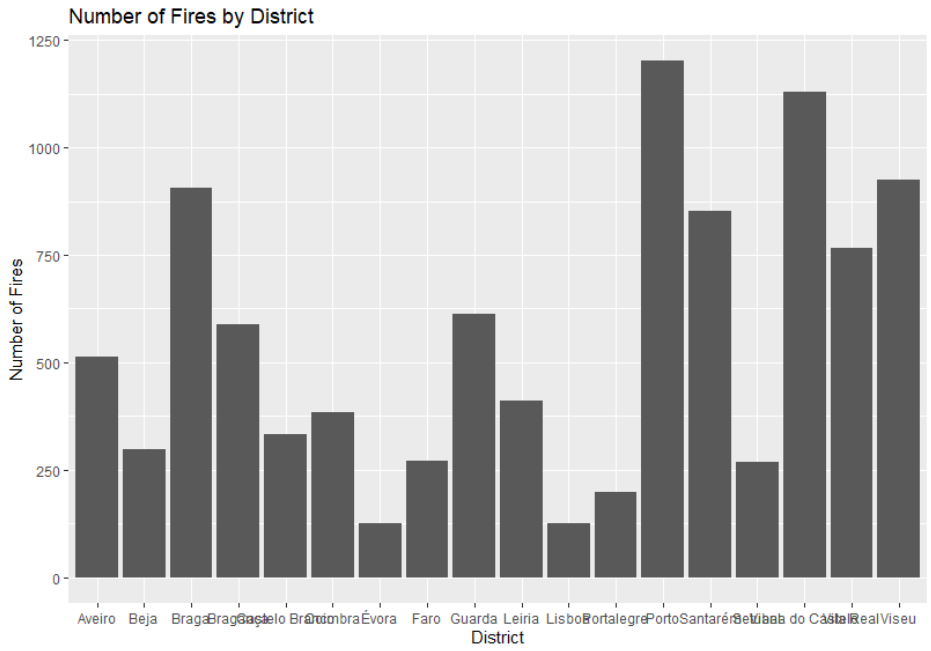
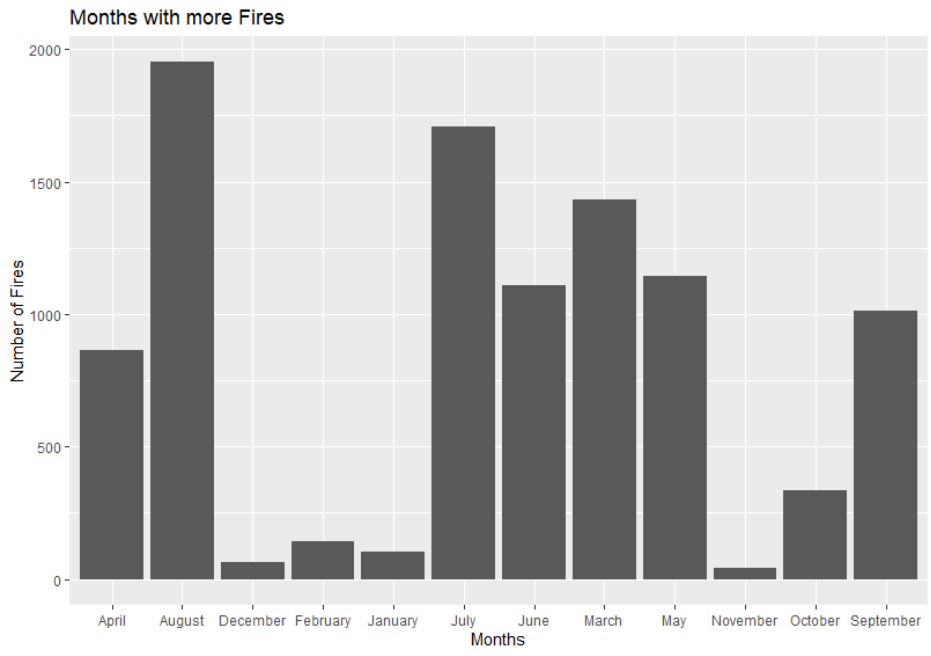


Figure 2: Number of fires per District

Porto, Viana do Castelo and Viseu appear to have the most fires. While Lisboa, Évora and Portalegre have the least.

3)Months with more fires In discussion, we assumed that the months of Summer would have have the most fires and we generated the following graphic to corroborate our theory.



.

4)Type of Origin of Fires

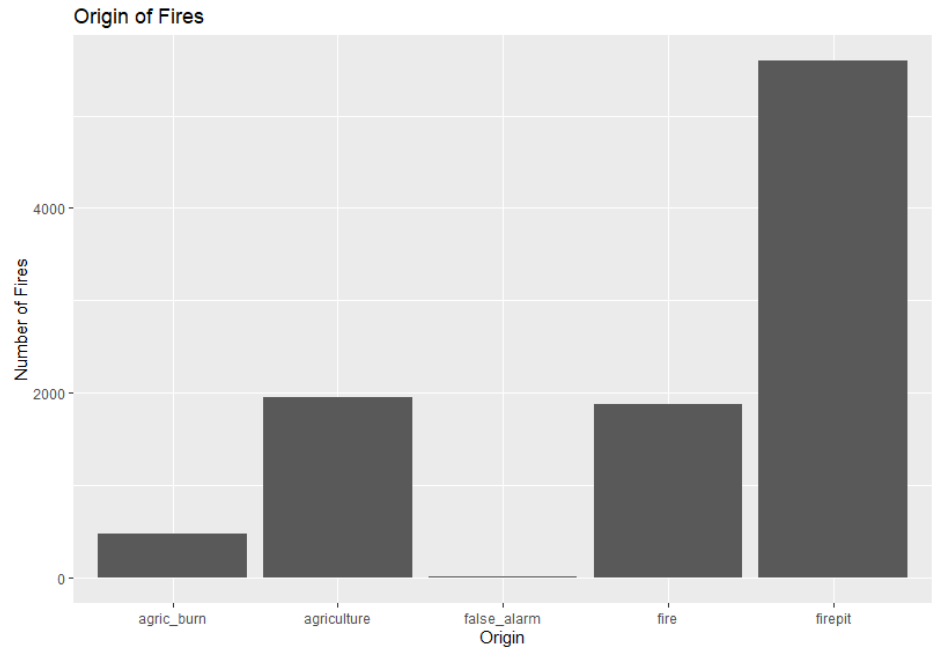
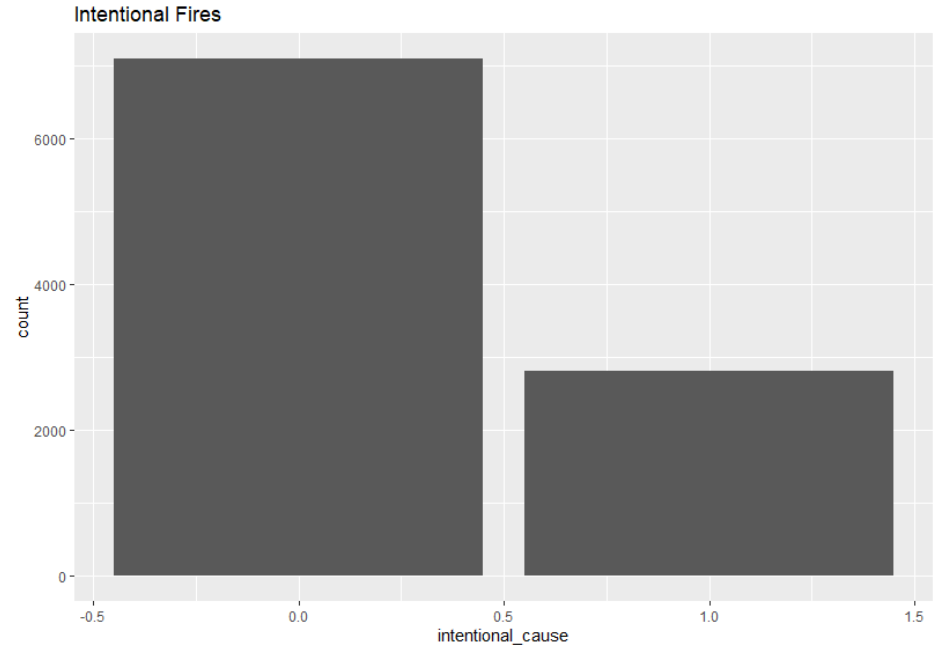


Figure 3: Type Of Origin Of Fires

The type of Origin of Fires allows to observe that the most fires, by a vast majority, start from fire pits.

5)Intentional Fires



The number of fires given their cause, we observe the number of intentional represented by 1 as opposed to unintentional ones that are represent by 0.

The following section analyzes relations between different attributes.

6)Relation between District and Origin

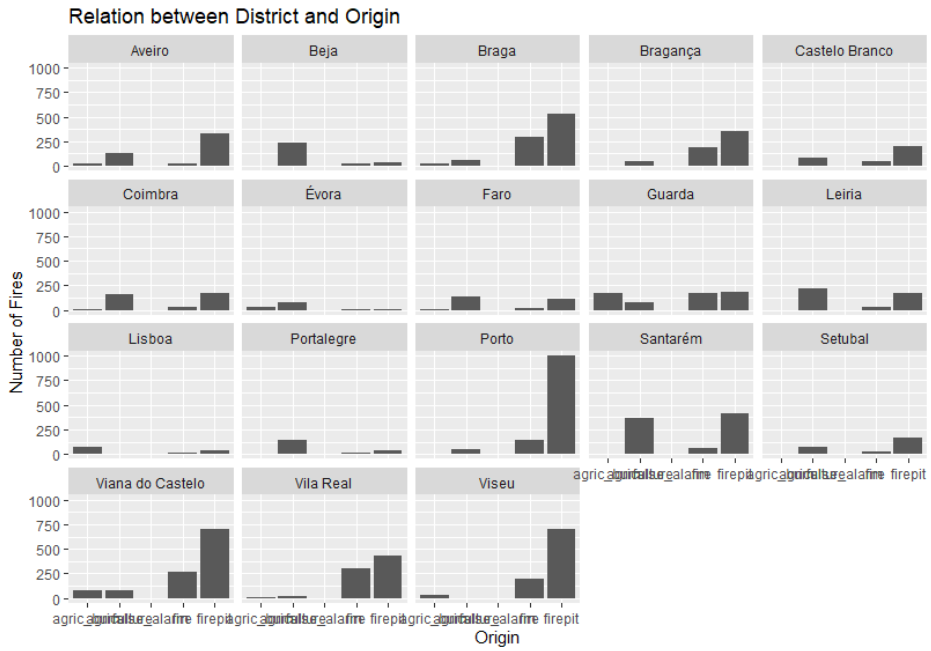
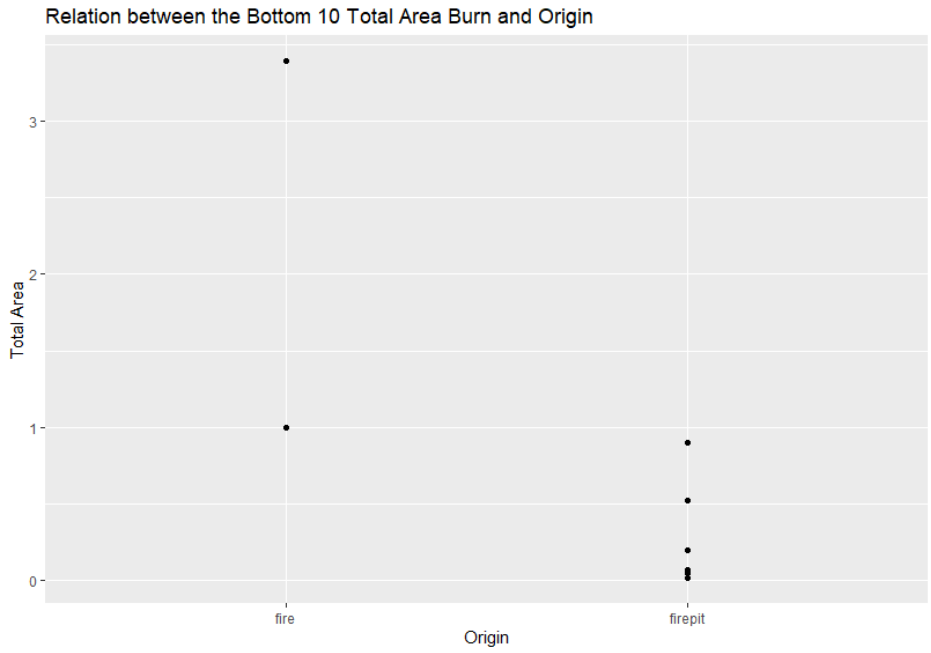


Figure 4: Relation between District and Origin

The previous graphic shows a perceptive on what originates the most fires given a district. For instance, it shows at first glance how in Braga, Porto, Viseu and Viana do Castelo and most fires are due to firepits.

7)Relation between the Bottom 10 Total Area Burn and Origin



To oppose this view, by observing the fires with the lowest area burn, we can see how they mostly stem from **firepit** and a few from **fire**.

9)Relation between District and the Maximum Temperature

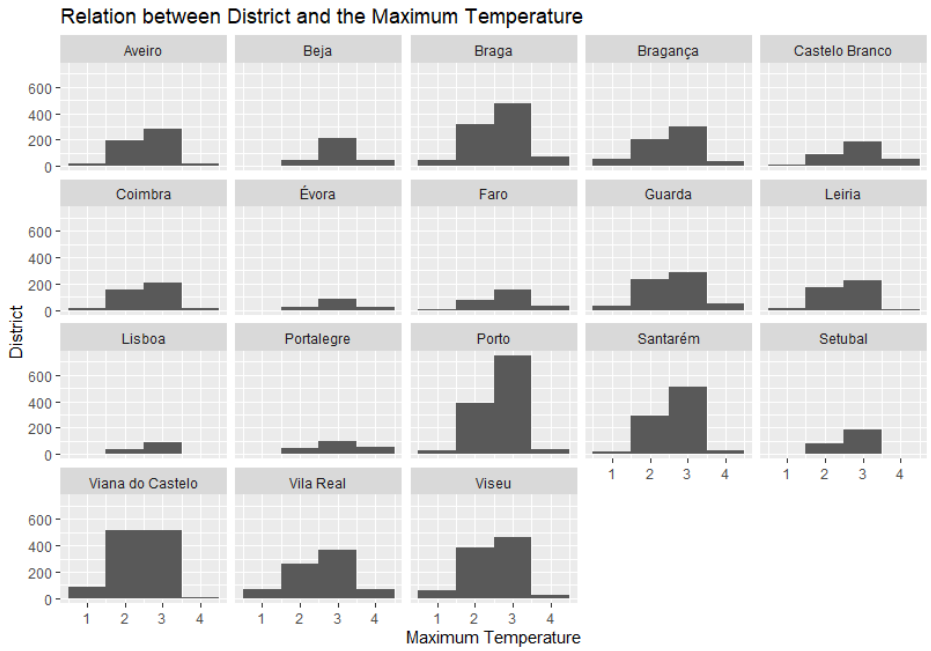
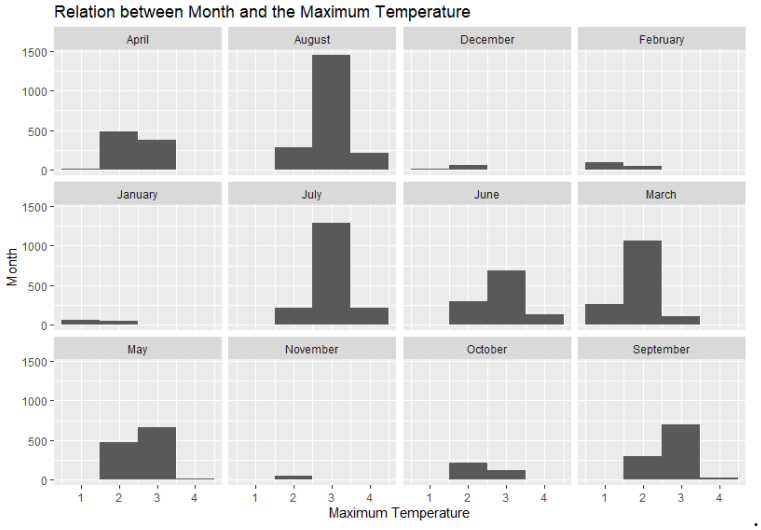


Figure 5: Relation between District and the Maximum Temperature

The maximum temperature achieved in a day when a fire occurred in relation to districts.
From cultural knowledge we see that districts in northern and inner Portugal coincide with districts where fires with the highest registered temperature occurred.
Viseu, being a district in inner Portugal, is notable for having high temperatures and as such, more prone to fires.
The same can be stated for **Porto** and **Braga**.
All three of this cities that serve as an example, are located in northern Portugal.
While, for instance, **Faro**, **Évora** and **Beja**, present fires with lowest temperatures and are southern districts of Portugal.

10)Relation between Month and the Maximum Temperature



.

This graphic justifies even more our previous observation that in summer months (June, July, August and September), the hotter months have the highest number of fires.

11)Relation between the Top 10 Total Area Burn and the Maximum Temperature

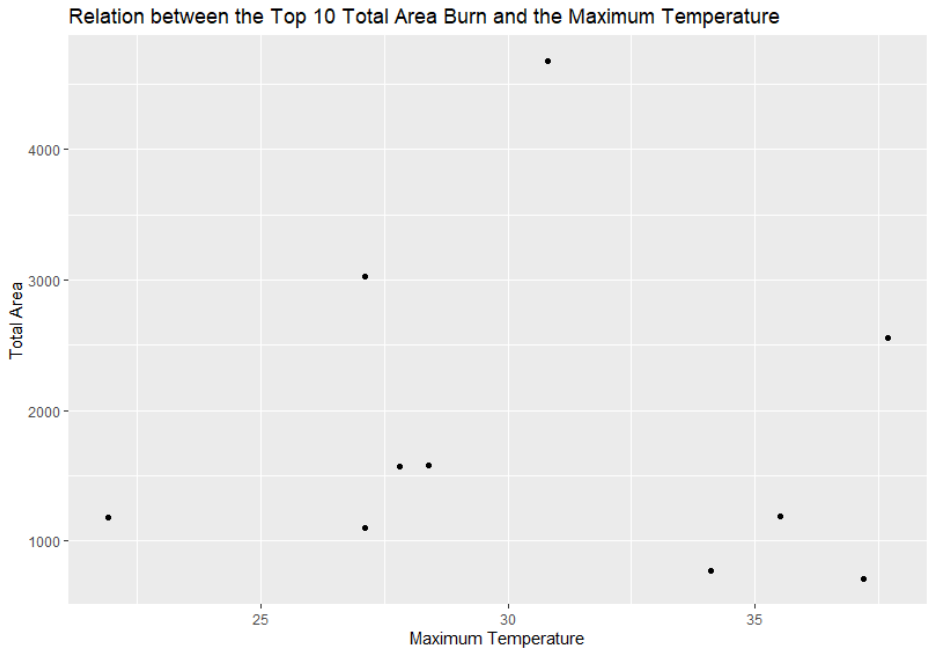


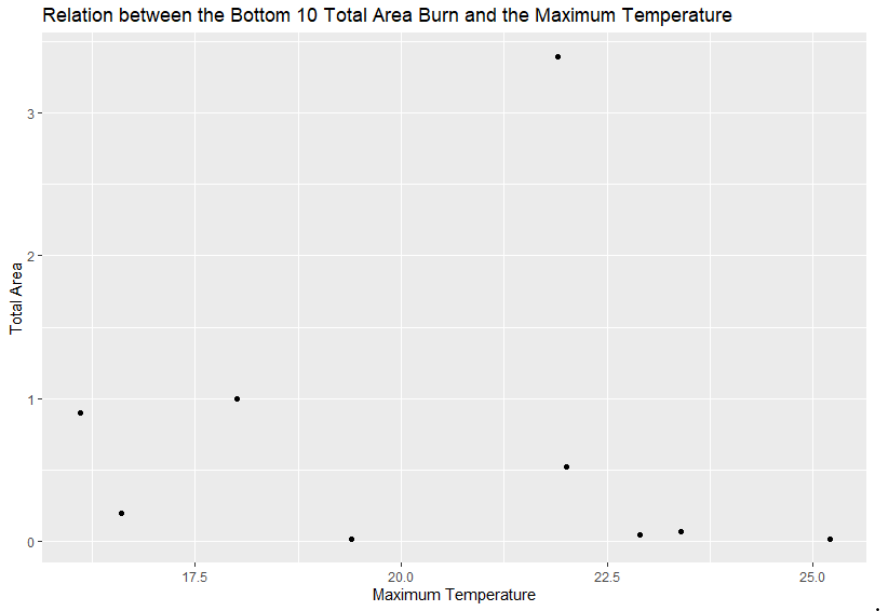
Figure 6: Relation between the Top 10 Total Area Burn and the Maximum Temperature

Here we observed that, between area and maximum temperature, the fires with the most area burn have temperatures that vary between approximately

$$6, 42$$

.

12)Relation between the Bottom 10 Total Area Burn and the Maximum Temperature



.

Coincidentally, the fires with the lowest maximum Temperature, that varies between approximately

$$5, 26$$

, have lower total areas burned except for an outlier value.

13)Map of Portugal with origin of fires

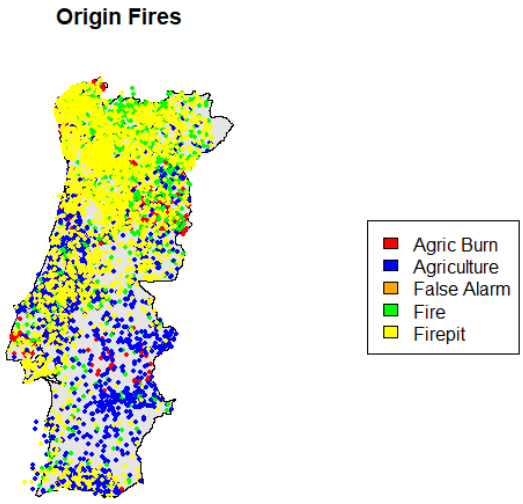


Figure 7: Map of Portugal with origin of fires

To obtain a birds eye view, we generated the following Portugal map with the type of origin of fires.
It shows a clear distinction between the North and South of Portugal where in the first one prevail Firepit induced fires and in the South, **Agriculture**.
Focus of agglomeration of **Agric Burns** are also notable in Portugal.

Experimental setup

Before experimenting with different predictive models, the training data set was divided, 30% to tests and 70% into training.

The following predictive models we're considered and applied to our data set. It follows a brief description of each one.

- Random Forest

Random Forest, is a machine learning ensemble method that starts with a selection of random samples from the given data set that will be latter used in forming decision trees.
After creating all possible decision trees, the obtained results are evaluated and the best outcome is chosen.
It is possible to limit the number of intended decision trees and the number of variables that are part of the division of the same decision trees.
In particular, the value of possible decision trees should not be to low as to guarantee that each observation is tested various times.

- Adaptive Boosting or Adaboost

It consists of an iterative classification process where new models are added form an ensemble in order to improve the performance of a base algorithm.
At each new iteration of the algorithm, a new model is built that tries to overcome the errors made in the previous iterations, the weights of each iteration are adjusted according to how wrongly they we're predicted.
The higher the weight the less accurate was the prediction.

- Naive Bayes

Naive Bayes classifiers, based on the Bayes statistic method, takes in account the probability of each observation belonging to a class.
Given a data set and the attribute we indeed to predict, it estimates how likely it should appears in the data set, being that all attributes are independent and equal among themselves.

- Bagging or Bootstrap Aggregating

Another ensemble model that obtains itself a set of models using different bootstrap samples of the given training data.
For each model a sample with replacements of the same size as the available data is gathered, which results in within each model there is a small proportion of examples that differs.

Results

In order to guide us, we used **Kaggle** to evaluate the performance of our models, but during a preliminary testing phase, we observed how we obtained higher values than the ones presented on **Kaggle**.
We understood it was due to the type tests made available on the website.
In order to summarize the models' performance we list the highest registered score obtained by each model.

- Random Forest

kaggle_fires_rf.csv 0.59876

- AdaBoost

kaggle_fires_b.csv 0.64310

- Naive Bayes

kaggle_fires_nb.csv 0.68976

- Bagging

kaggle_fires_bg.csv 0.52584

As we can attest, the model that scored the highest was the Naive Bayes and the lowest was Bagging.

Conclusions

In this project we had the opportunity to discover a little more about the area of data science, we realized with this work that one of the most important parts of data analysis is the pre-processing of information. It was also important to understand that any task related to data in the real world requires a somewhat deep knowledge about the subject that the data portrays, in this case, it will be important to know what kind of attributes would be the most suitable to be able to generate a prediction, this knowledge proved essential to save processing time in our program and work that otherwise would have overloaded us even more.
Another positive point of this work was the way it was proposed, in the form of an open competition. The Kaggle platform allowed each group to have the opportunity to see the results of other groups which generates a friendly competition and boosts everyone's results.
We also noticed that for each dataset there is an algorithm that best applies and for this algorithm there are also configurations and small details that best take advantage of this algorithm.
Finally, we consider that the task was successfully completed even with barriers, such as the lack of positive inputs or with a non-uniform dataset, we achieved a result of 69% which we consider to be a good result.

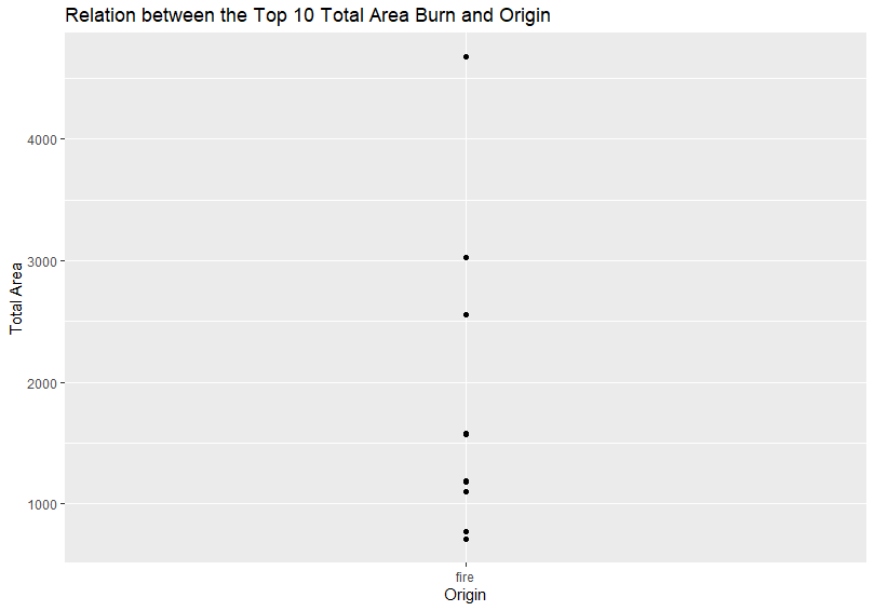
Future Work

In order to further improve our work, we would like to do more experiments, try more algorithms, more variables and more configurations, in particular we would like to explore with Artificial Neural Networks, since we only have to predict a theoretical approach and a practical introduction.
It seems a growing scientific field with evidence that it is the way to go in the predictive field, for the established success and because it captured our interest during the classes, we thought it would be interesting to reproduce our work with a neural networks algorithm in order to obtain better results.

Annexes

In this section, we present some other insights we obtained.

1)Relation between the Top 10 Total Area Burn and Origin



The following graphics analyze the relation between the highest or lowest areas burned and their origin. Here we state that the top 10 fires with the most area burn in our data, all originate from **fire**.

Referências

[1] Tutorialspoint, https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_random_forest.htm , consultado em 09/01/2021

[2] Rita Ribeiro, https://www.dcc.fc.up.pt/~rpribeiro/aulas/DMI2021/#Slides, consultado em 09/01/2021