U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

DatAttack 2023
Faculdade de Engenharia da Universidade do Porto

Project

Predicting the quantity of resources needed for every Civil Protection emergency

Filipa Costa (filipabcosta@tecnico.ulisboa.pt), Miguel Neves (neves.miguel8@gmail.com), Vasco Reid (vascoreid@gmail.com)

ABSTRACT — *The present work aims to predict the quantity of resources needed for every Civil Protection emergency based on the location, the nature of emergency and time, in Portugal. To fulfill this purpose, data from the year of 2016 was retrieved from the publicly available data collected by The National Emergency and Civil Protection Authority. Data was pre-processed and an exploratory analysis was conducted. A strong temporal dependency was discovered. Finally, regression and classification models were created to predict the number of people and vehicles needed to be allocated for every emergency, with an accuracy of 60% and 78%, respectively.*

Keywords: National Emergency and Civil Protection Authority, Linear Regression, Naive Bayes, Decision Trees

1 Introduction

Civil protection emergencies like fires, car accidents, and natural catastrophes can happen at any time, causing extensive damage and endangering lives. To save lives and reduce damage in such circumstances, timely and efficient disaster management is crucial. Predicting the quantity of resources needed to properly address the emergency situation is an important component of disaster management. Emergency responders can more efficiently distribute resources, such as personnel and operational equipment, if the necessary resources are known in advance and can be made available where and when they are needed. This may be crucial in preserving lives and lessening the disaster's effects. Although civil protection emergencies happen across the entire Portugal, some regions are more susceptible to some disasters. For example, the northern and central regions of Portugal, such as Minho and Douro, are known for their dense forested areas and have historically been prone to fires. In contrast, the southern regions, such as the Algarve, have a more arid climate and are less prone to fires.

With this in mind, the goal of this project is to study how the time, location and Nature of civil protection emergency influence both the number of involved vehicles (in land) and the number of involved people (in land), in the year of 2016.

Firstly, in section 2, data was pre-processed and submitted to a preliminary analysis. In section 3, regression and classification models were set into place. Finally, the conclusions of this study are given in section 4.

2 Data Exploration & Preparation

To start the assessment, the data from 2016 provided by The National Emergency and Civil Protection Authority was analysed. Note that only Portuguese related data was retrieved, implying that the work in hand does not establish an assessment for other countries, aside from Portugal.

At first sight, given that the data was provided with 88 different natures of emergency under the variable *Natureza*, the granularity of this variable is very big, thus needs to be decreased. For this, GPT-4 was applied to condense the already-existing categories into only 10. Among the resulting categories, the ones with fewer than 100 records were eliminated because they account for a relatively small portion of all observations.

The analysis of missing values followed next. Since it was determined that records with missing values only made up 0.003% of all observations, they were eliminated.

Further, it was noticed that the quantity of resources needed to be allocated was not balanced among the records, for all Meios Terrestres (number of vehicles needed to be allocated in land), Operacionais Terrestres (number of people needed to be

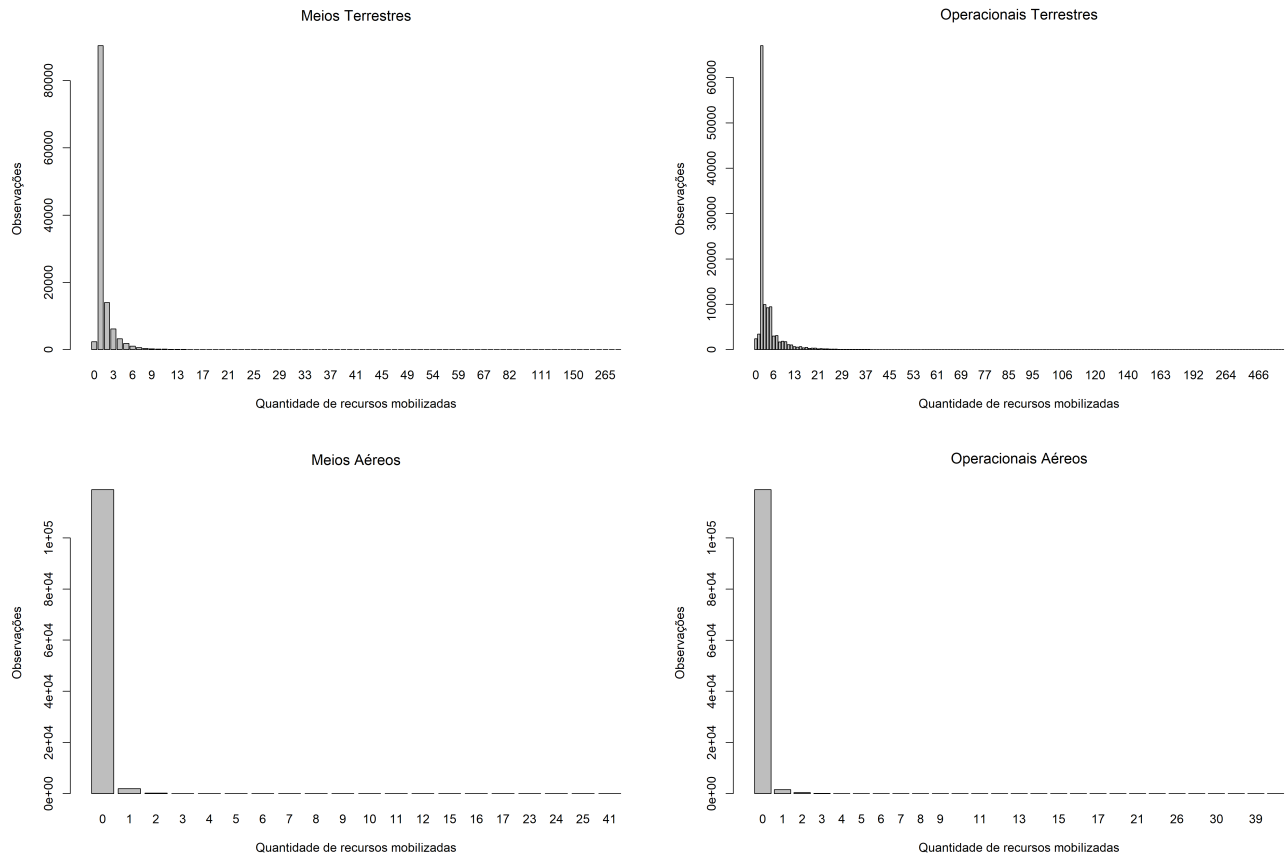

Fig. 1: Number of observations by quantity of resources needed to be allocated, in each type of resource.

allocated in land), Meios Aéreos (number of vehicles needed to be allocated in sky), Operacionais Aéreos (number of people needed to be allocated in sky) as it can be seen in Figure 1.

The enormous imbalance between the class=0 and the others, as seen in Figure 1, limits the Aerial resources from having enough data to estimate how many of these resources will need to be allocated. Thus, these Aerial types of resources won't be taken in consideration.

The vehicles required to assign land resources only seek a significant number of records when there are fewer than six vehicles required. In order to reduce the granularity of this variable, the records are all aggregated into one bin when the number of vehicles is equal to or more than 6, which lowers the maximum number of resources that could be allocated to 6. In a similar reasoning, when the quantity of people needed to be allocated is equal to 13 or more, the records are all grouped into the same bin, decreasing the possible quantity of resources to be allocated to 13.

For a brief analysis of the dataset, the temporal dependencies of the quantity of resources that needed to be allocated were then examined in Figures 8 3. For this, the variable *DataOcorrencia* was divided in *Month*, *Day*, *Weekday* and *Hour*.

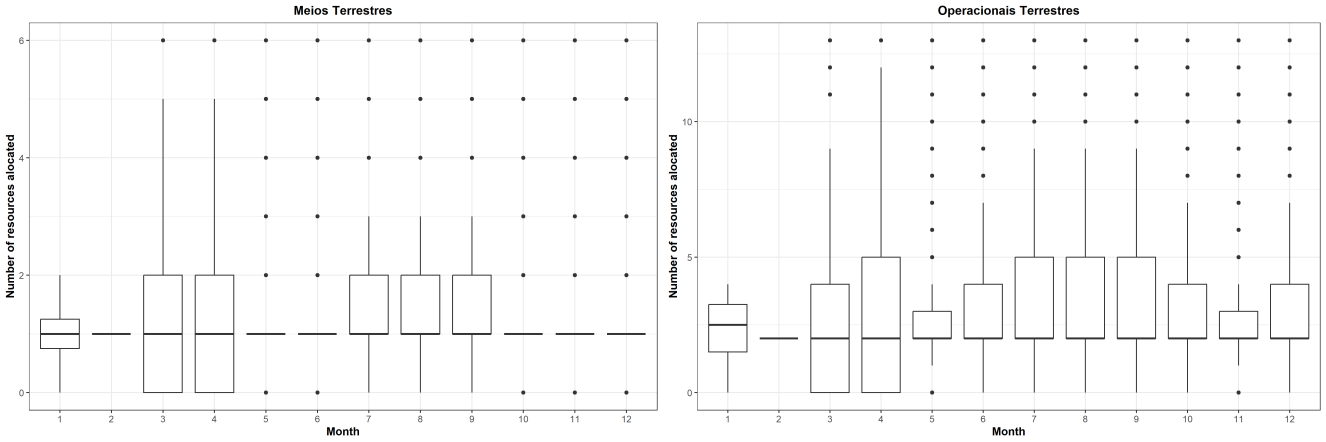

Fig. 2: Number of resources alocated by Month in each type of resource - number of vehicles in the left and number of people in the right.

As observed in Figure 8, some values go outside of the interquartile range, yet each ”point” actually represents a large number of observations. In this manner, it was decided to leave all of the values in to avoid losing information because they make up a significant portion of the entire data. Additionally, the amount of resources that must be allocated depends not only on the month variable but also on the nature of emergency and the district. Records that are considered outliers when analyzing data by month are not always considered outliers when analyzing data by another variable.

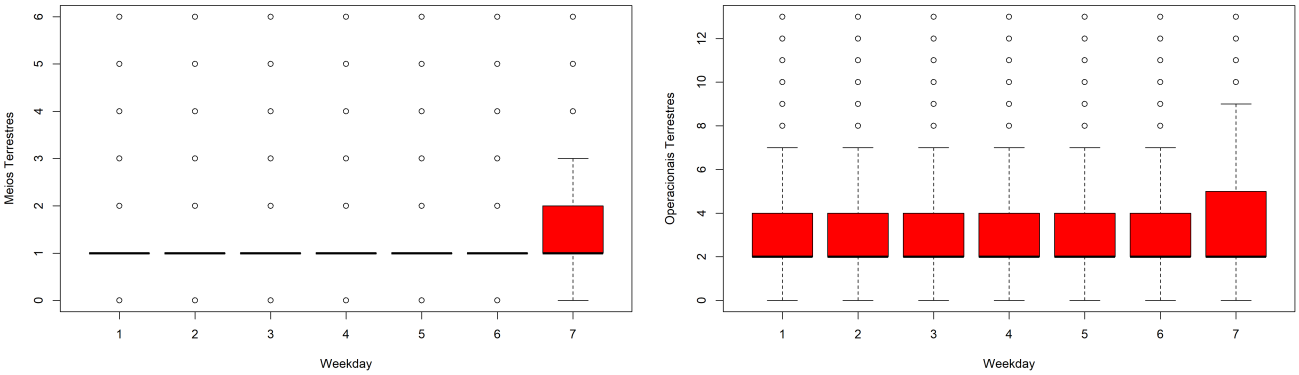

Fig. 3: Number of resources allocated by Day of the Week in each type of resource - number of vehicles in the left and number of people in the right.

When analysing the distribution of allocated resources by Day of the Week in Figure 3, it was found that on weekdays with a value of 7, that is, on a Saturday is when more vehicles and people need to be allocated.

As a result, it was chosen to keep these variables when building the model because the quantity of resources depends on both the Weekday and the Month.

Next, Figure 4 examines how resources were allocated based on the emergency nature.

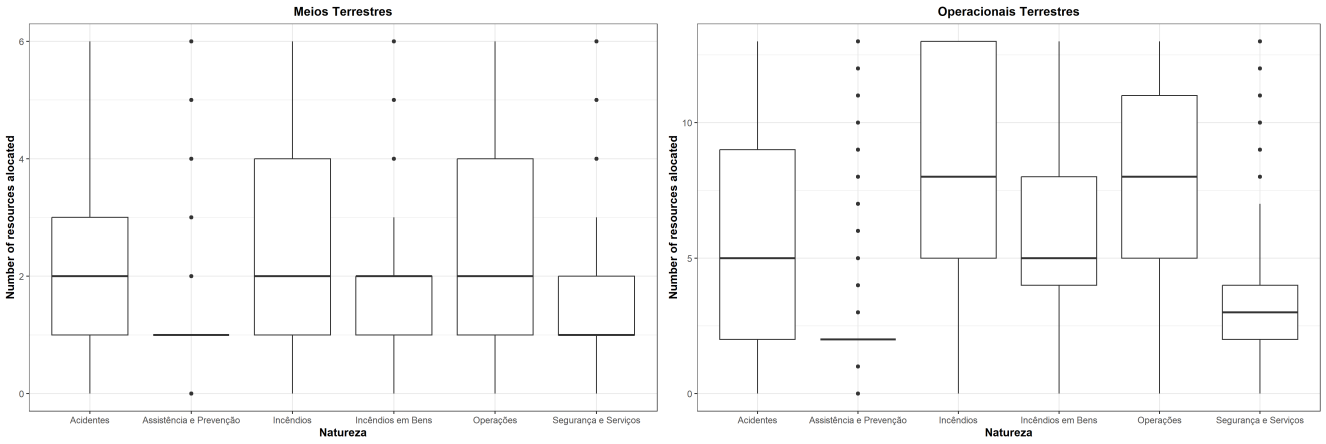

Fig. 4: Number of resources allocated by Emergency Nature in each type of resource - number of vehicles in the left and number of people in the right.

As it stands, *Incêndios* and *Operações* are the natural emergencies that require a greater quantity of both types of resources. It is expected that there would be a high demand for resources to combat fires in Portugal given the country's Mediterranean climate, which is characterized by hot, dry summers that are favorable to forest fires. In this manner, a greater number of firemen may need to be dispatched due to the recurrence of deadly wildfires in Portugal, which flames can develop quickly and are challenging to control.

The distribution of resources allocated among the districts was also examined in the code, but as it typically displayed the same behavior for every district, it was decided not to include it in the report.

In the variable *EstadoOcorrencia* it was also noticed that some events were still not "Encerrado", meaning that the event was still occuring. It was determined to only include records that fell under the category "Encerrado" because it contained 98 % of the observations.

Finally, there were several location features available such as *Distrito*, *Concelho*, *Freguesia* and *Localidade*, however the granularity of *Concelho*, *Freguesia* and *Localidade* was too big, meaning that the model would not have sufficient data of each category to learn a pattern.

For instance, "Vila Nova de Gaia" accounts for 3% of the total records and is the *Concelho* class that presents the most observations. Since these variables are all significantly correlated with one another, only one can be selected to be part of the model. District is the one that will be included because it has less granularity.

Further analysis was conducted to get a grasp of the distribution of incidents per district, per group (Natureza) of the incident, and per Month. As shown bellow.

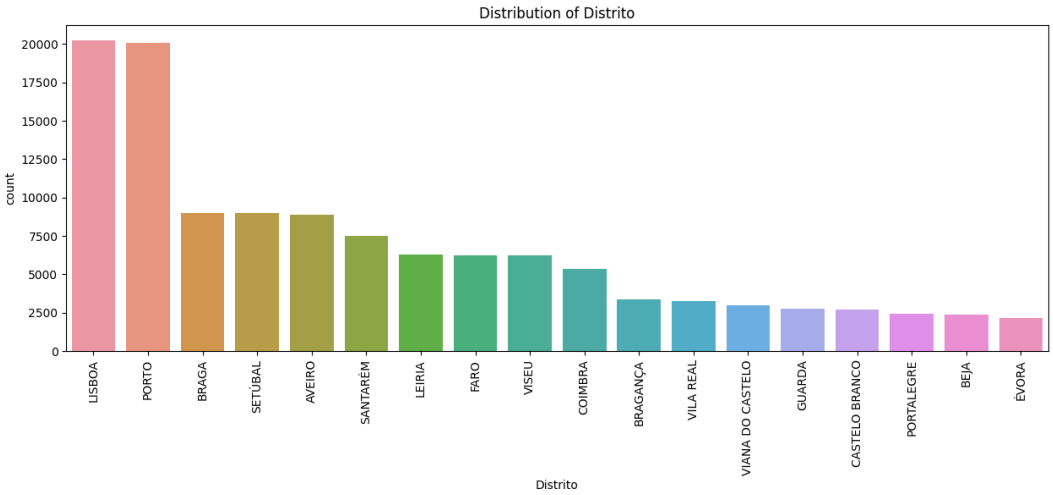

Fig. 5: Number of Incidents per District

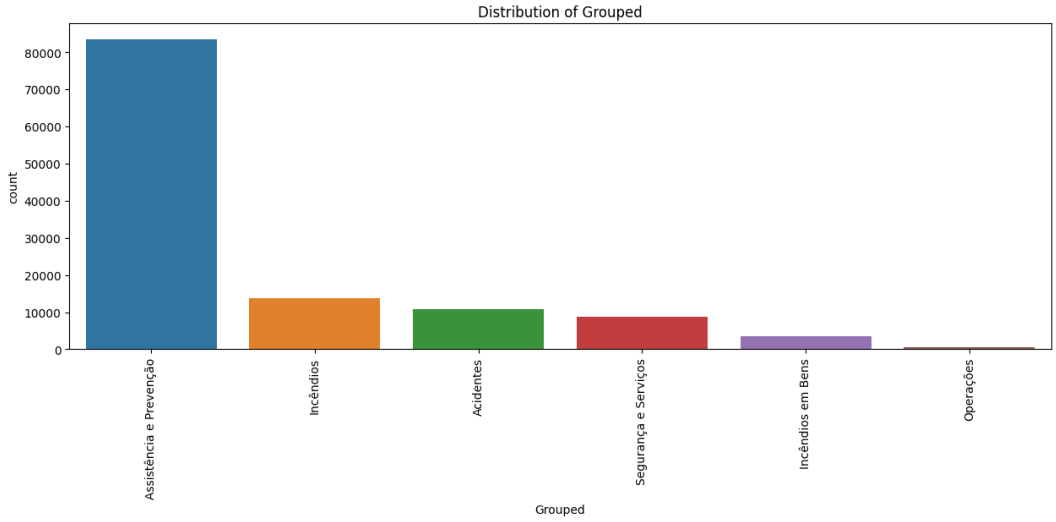

Fig. 6: Number of Incidents per Group of incident

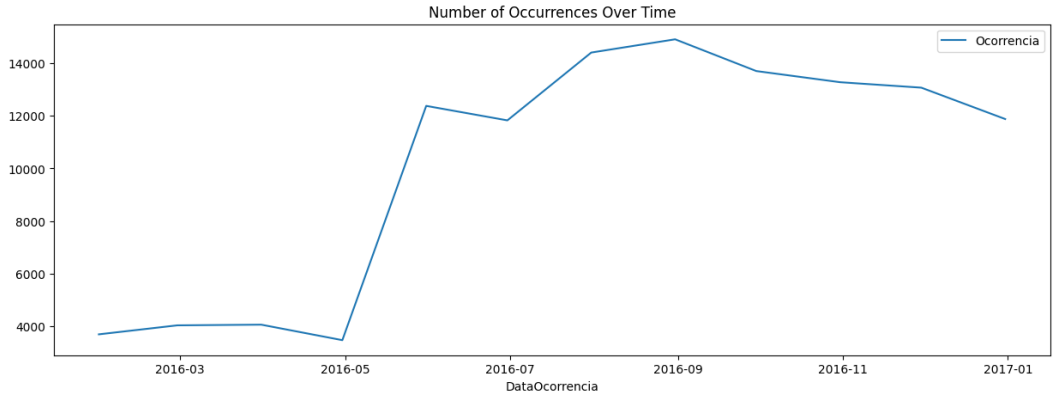

Fig. 7: Number of Incidents per Month

2.1 Encoding

For the categorical values of the *Natureza* and *Distrito*, a one hot encoding was nada in order for these values to be able to be sent as input to a model such as a neural network.

2.2 Geographical Heatmap of Portugal

In this analysis, we visualized the number of terrestrial operational personnel (”NumeroOperacionaisTerrestresEnvolvidos”) and the number of terrestrial resources involved (”NumeroMeiosTerrestresEnvolvidos”) based on their maximum, mean, and minimum values. The data was represented on a heatmap of Portugal’s districts for different months and groups.

We used GeoPandas to handle the geospatial data and Plotly to create interactive plots. The visualization process involved the following steps:

- Load the data into a Pandas DataFrame and preprocess it.
- Load the geospatial data (shapefile) of Portugal’s districts from a official website.
- Merge the geospatial data with the DataFrame.
- Create a heatmap function (*create_heatmap*) that takes the month, group, and the metric to visualize
- Then it generates a heatmap using GeoPandas.
- Create interactive dropdown menus to allow users to select the month, group, and variable for visualization.

- Update the heatmap based on the user's selection using the *update_chart* function, which calls the *create_heatmap* function with the chosen inputs.

The resulting interactive heatmap, as seen in figure , allows users to explore the distribution of terrestrial operational personnel and resources involved across different months, groups, and metrics (maximum, mean, and minimum values) in Portugal's districts. This visualization can be useful for understanding the spatial distribution of these variables and identifying areas with higher or lower numbers of personnel and resources.

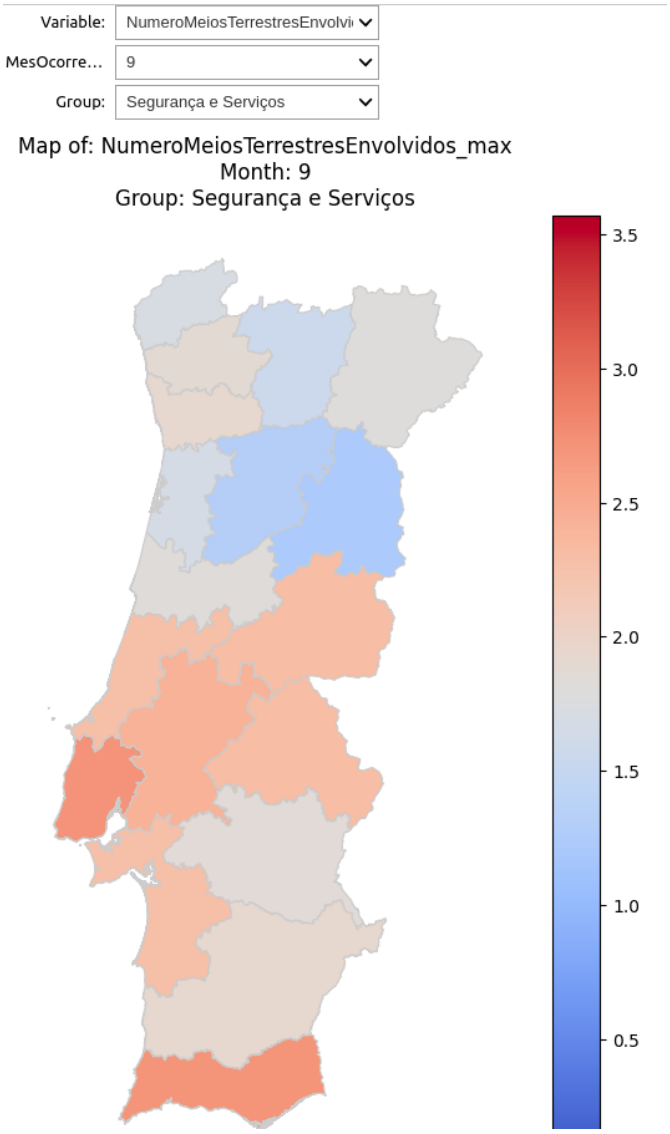

Fig. 8: Number of resources alocated by Month and District based on the selected Group of incidents - variable to be shown is also selected

3 Results and Discussion

3.1 Test and Training Data

The training set is composed by 80% of the original data and the test set is composed of the remaining 20%. The splitting of the data was performed randomly.

3.2 Methods

The problem we have in hand can be seen as a regression problem - where we want to predict a number between 0 and 6, and then round up, since we want to predict the number of resources needed to be allocated, and this number must be an integer - or a multiclass classification problem - where we want to classify to wich class between 0,1,2,3,4,5,6 a new observation belongs.

Since the competition only has 24 hours, the measures used to evaluate the quality of the resulting models will be simply the R^2 for the regression models and Accuracy for the classification models.

- R^2: Amount of dependent variable variability explained by the independent ones;
- Accuracy: Relative frequency of correctly classified observations in the sample.

In the end, in this project we ended up implementing 3 different modeling techniques, namely Linear Regression, Naive Bayes and Decision Trees.

In the classifiers, we decided to develop them through application of the function train of the package caret with cross-validation (number of folds equals to ten). Cross-validation allows to estimate the competence of a model on unseen data (data not used for training the model). The metric implemented to select the optimal model is the Accuracy.

All the model techniques will be constructed with the same independent and dependent variables, that is:

- Independent Variables: *Month*, *Distrito*, *Weekday*, *Natureza*;
- Dependent Variables: *NumeroMeiosTerrestresEnvolvidos*, *NumeroOperacionaisTerrestresEnvolvidos*.

The results are shown in the Table 1.

3.2.1 Linear Regression

Linear regression was used to seek for the relationship between the independent variables and the number of people and vehicles needed to be allocated. In other words, it seeks to predict the number of people and vehicles needed to be allocated based on the values of the independent variables, through estimating the values of the coefficients that best fit the data.

3.2.2 Naive Bayes

Naïve Bayes (NB) Classifier is a probabilistic algorithm that is usually used for classification problems. In this algoithm, the distribution of samples in each class is modelled using a probabilistic model which assumes that all the variables (given the class) are independent from each other. The Naive Bayes classifier combines Naive Bayes model with the maximum a posteriori decision rule, which instructs the model to choose the most probable hypothesis in the end. It finds the probability of a given set of inputs for all possible values of the possible outcomes and pick up the output with maximum probability. However, one disadvantage of this method is if some class is missing when the frequency-based probability estimate will be zero, so we will get a zero when all the probabilities are multiplied. To overcome this issue, we chose to apply Laplace Smoothing. Laplace Smoothing is a technique that ensures that each feature has a nonzero probability of occurring for each class.

3.2.3 Decision Trees

Decision Trees are a non-parametric supervised learning method used for classification. They learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules and the fitter the model. The Random Forest (RF) method is a combination of individual decision trees, in which each tree depends on the values of a random vector sampled independently, and with equal distribution for all trees involved in the forest. In this case, the square root of the number of variables is used in the input.

Models	R^2/Accuracy			
	Meios Terrestres		Meios Op. Terrestres	
	Train Set	Test Set	Train Set	Test Set
Linear Regression	0.37	0.36	0.48	0.47
Naive Bayes	0.74	0.74	0.56	0.56
Decision Trees	0.78	0.78	0.61	0.60

Tab. 1: Performance Measures

4 Conclusions

The aim of this report was to investigate the relationship between the quantity of resources needed for every Civil Protection emergency and the location, the nature of emergency and time, at a national level.

Unfortunately, the official data available posed many challenges and so the results given by all methods applied are to be taken cautiously. Regardless, it was still very much possible to predict the number of people and vehicles needed to be allocated with accuracies of 60% and 78%, respectively.

To best model the effect of the independent variables *Month*, *Distrito*, *Weekday* and *Natureza*, multiple distinct models were performed. The performance of these models was then measured by running it against test data.

Other interesting observations allow some conclusions, such that Incêndios and Operações are the natural emergencies that need a higher quantity of resources; the number of resources needed to be allocated depends both on the day of the week (Sunday, Monday...) and the Month - both variables were significative when building the regression model; In particular, it was verified that the quantity of resources needed tends to be higher on Saturdays.

Future works could and should include more variables, such as population density and geographical area, in order to find more patterns in the data, and consequently improve the prediction models performance.

Since Civil Protection emergencies is a worldwide problem that occurs every year a more broad and yearly approach considering worldwide data would also be valuable.

It is also important to have in mind that this project was done under 24 hours, thus there was not too much time dedicated to analyse different performance metrics. A more broad range of different methods would also be appreciated, for example, applying the method XGBoost.

Finally, a more narrower study on the different *Concelhos*, *Freguesias* e *Localidades* should also be done, but this was not possible due to data constraints.