U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

University of Porto - Faculty of Engineering

Hackathon Report

Danger Cast

Deep Learning application on the registered incidents by the Portuguese Civil Guard

Edgar Torre
up201906573@edu.fe.up.pt
João Andrade
up201905589@edu.fe.up.pt
Rodrigo Tuna
up201904967@edu.fe.up.pt
Sérgio Estêvão
up201905680@edu.fe.up.pt

April 2023

Contents

List of Figures

1 Introduction

The National Emergency and Civil Protection Authority plays a crucial role in ensuring the safety of citizens during emergency situations. However, the organization faces numerous challenges in optimizing its work processes to provide the most efficient response during crises. In the context of the "Data Attack" hackathon, we were challenged to develop a data-driven framework that would assist the National Emergency and Civil Protection Authority in optimizing their work processes. The framework would enable the organization to make data-driven decisions in real-time during emergency situations, thus improving their response time and efficiency.

Furthermore, using Convolutional Long Short-Term Memory Neural Networks (CNN-LSTM) to anticipate future patterns in emergency circumstances based on prior data might be beneficial. By training the model on previous emergency data, it can detect patterns and connections that human analysts might miss. These findings may be utilized to guide decision-making and assist the National Emergency and Civil Protection Authority in better preparing for future emergencies. Furthermore, the Convolutional LSTM may be utilized for anomaly detection, alerting authorities to any strange occurrences or departures from regular patterns, and allowing them to intervene quickly before the situation worsens. Overall, incorporating Convolutional LSTM into the data-driven framework can considerably improve the capacities of the National Emergency and Civil Protection Authority, allowing them to better respond to emergencies.

Considering this, in the first stage, we will explain the methodology we used to approach this challenge and the results we obtained in each step. After that, we will demonstrate the CLI prototype developed to utilize our findings and the predictive model developed in an accessible way.

2 Methodology

In this machine learning problem, we followed the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, which is a well-established framework for data mining and machine learning projects. To accomplish the objective of this study, CRISP-DM methodology was followed.

3 Development Roadmap

3.1 Business Understanding

It is a challenge for the National Emergency and Civil Protection Authority (Proteção Civil) to make their work in reacting to emergencies and natural disasters as efficient as possible. To do this, they are looking for a data-driven framework that can assist them in making predictions about the frequency of occurrences in the future, allocating resources effectively, and identifying patterns that can enhance their services.

The Authority has made a dataset with the history of incidents, including dates, places, and the number of experts involved, available to address this issue. Participants are asked to create a solution that can aid the Authority in optimizing its operations using this dataset as well as data from other sources.

The goal of the solution should be to forecast the volume of upcoming incidents and the volume of resources required to address them, per location. Additionally, it should seek to locate pertinent patterns in the data, such as relationships between the type of occurrence and the degree of development, regional topography, and population density.

The ultimate objective of this challenge is to assist Proteção Civil in bettering its services and allocating its resources, potentially resulting in cost and resource savings.

3.2 Data Understanding

The efficient and effective management of emergency situations presents difficulties for Proteção Civil. A data-driven framework can be created to forecast the number of upcoming events and the number of resources needed for each location in order to optimize the Authority's activity. To do this, we must first comprehend the information that is currently available regarding emergency situations, their locations, and the resources involved. In order to analyze data quality, identify any restrictions or biases, and define the breadth and constraints of the data, we will explore and summarize the data in this part.

3.2.1 General Analysis

Firstly we began by analyzing the overall general statistics of the dataset in order to confirm their types, some general information, and whether or not nulls were present.

Then we proceeded to plot some graphs in order to better understand some of the relations between the attributes.

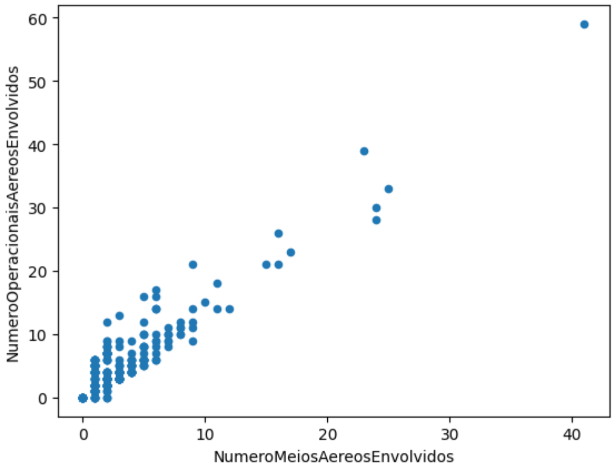

Figure 1: Relation between the NumeroMeiosAereosEnvolvidos and NumeroOperacionaisAereosEnvolvidos attributes

This one represents the relation between the 'NumeroMeiosAereosEnvolvidos' and 'NumeroOperacionaisAereosEnvolvidos' attributes

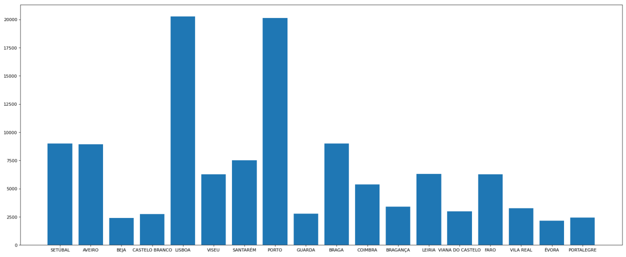

Figure 2: Number of occurrences per district

In this one we decided to check the number of occurrences in every district in Portugal, to better see if there were any hot spots or something that would go out of normal, but everything appeared to be as we expected.

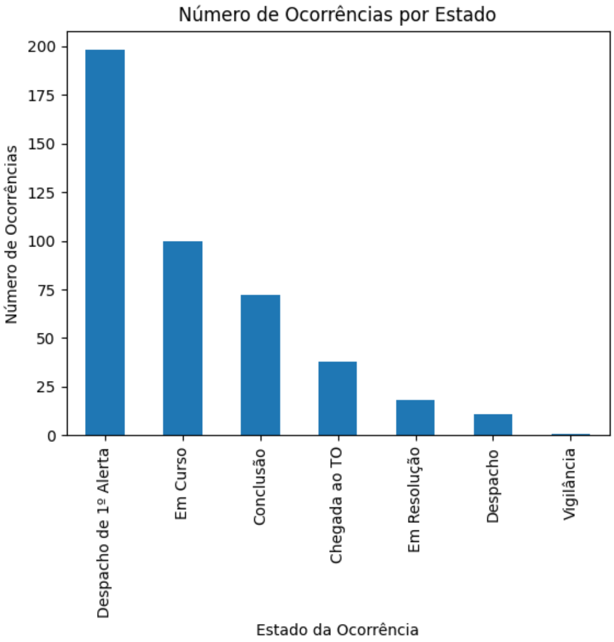

Figure 3: Number of occurrences which did not have a termination date and their correspondent type

We checked this because we wanted to better comprehend the amount of entries in the dataset that did not have a termination date even though the dataset goes back to the year of 2016, so in theory all of the occurrences or at least most of them would be over.

Carrying on our exploratory analysis we then checked for the top 3 occurrences in each of the districts where we concluded that in almost every one the 'Trauma' occurrence seemed to be the highest one.

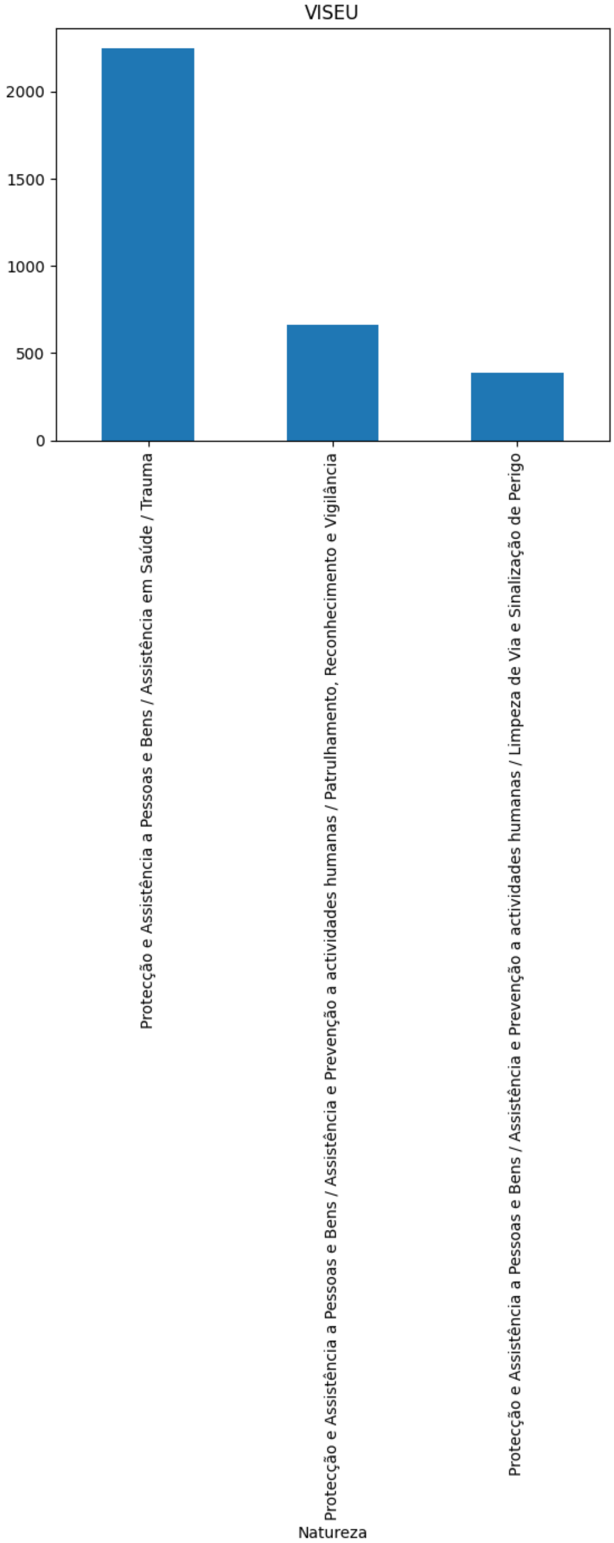

Figure 4: Top 3 occurrences in Viseu

When we analysed this data we also tried to find what would be the reason for the bottleneck of Proteção Civil and we concluded that in order to calculate it what would make sense was to compare the types of most frequent occurrences with the number of available resources in each of the districts that would help. This way it would be possible to mobilize resources from one district to another to better suit the needs of each one.

However we do not have access to the number of resources of each district (we only have the amount of resources utilized in previous occurrences). After realising this we thought of a few solutions such as calculating the average amount of necessary resources to allocate for each of the occurences, then calculate what occurrences would be most resource-intensive and detect any possible bottlenecks.

Now for an analysis more related to the number of occurrences over the time we decided to calculate the number of occurrences over each hour in a day, each day in a week and each month of the year.

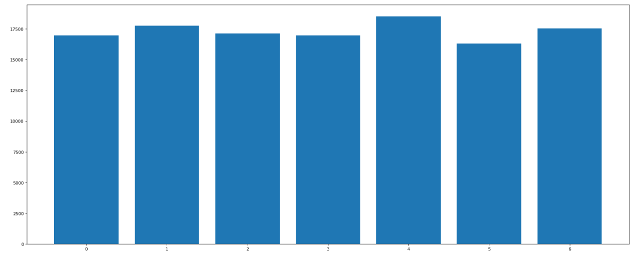

Figure 5: Number of occurrences for all districts for each day of the week

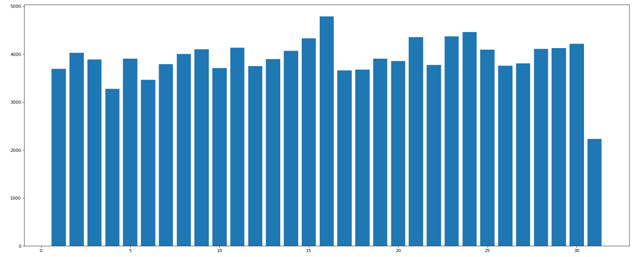

Figure 6: Number of occurrences for all districts for each day of the month

In these graphs we can see that the number of occurrences is evenly distributed along each day of the week and each day in a month

Now to conclude our data analysis we decided to see how the ratio of true/false alarms would vary with the hours of the day

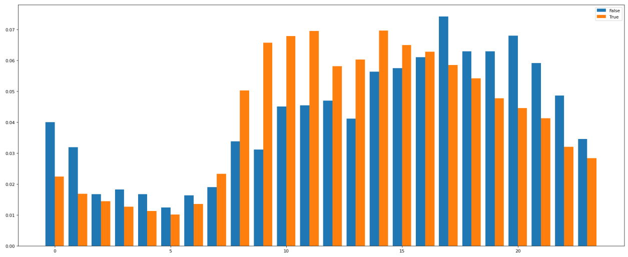

Figure 7: Percentage of false and true occurrences for all districts for each hour of the day

We can easily conclude that as we expected the number of false alarms increases a lot over the night. One possible reason is that it is often darker at night, which can cause shadows, reflections, and other visual effects that may be mistaken for signs of an emergency. Additionally, people may be more on edge and easily alarmed at night, especially if they are tired or anxious, which could lead to more false alarms.

3.3 Data Preparation

3.3.1 Base Dataset

We considered the datasets of the years 2016, 2017, and 2018 and concatenated them, since:

- We considered there wasn't enough data to use in our model for depth in every municipality
- Events with yearly seasonality such as forest fires would not be captured by our model if only one year was considered.

3.3.2 Data Prunning

Firstly, after transforming *DataOcorrencia* and *DataFechoOperacional* from a string to a DateTime, we removed the time part of the DateTime and kept only the date part.

Then we verified many rows had nulls in *DataOcorrencia* and *Latitude*, so those were removed. Also, about these coordinates, occurrences with invalid dates and longitude/latitude values were removed. Some longitude/latitude errors were detected, as Portugal's rectangle shape enables us to establish a lower and upper bound for both features. Some of the errors followed a clear logical path, for example, positive values that were supposed to be negative were positive, or values 1000 times bigger than the supposed value, so those entries were simply fixed. However, if there was not a clear explanation for further inconsistencies in the latitude/longitude, then these rows were removed.

Another outlier removed was an occurrence that called for 2000 operational and no vehicles for an occurrence, seen in the top left of Figure 8.

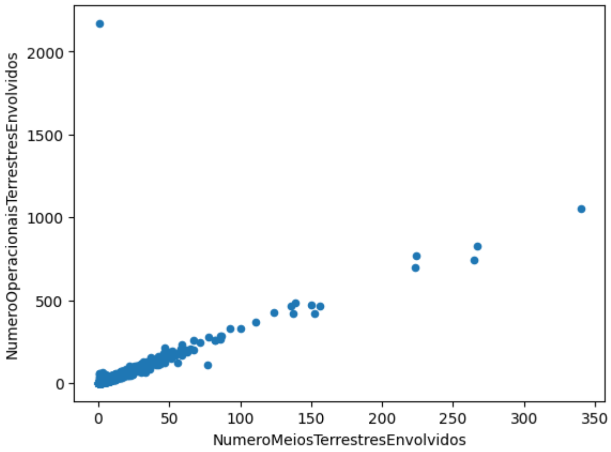

Figure 8: Relation between the NumeroMeiosTerrestresEnvolvidos and NumeroOperacionaisTerrestresEnvolvidos attributes

As there was a clear lack of occurrences in the first 5 months of 2016, seen in Figure 12 contrasting with the rest of the dataset, the occurrences until 11th May 2016 were discarded.

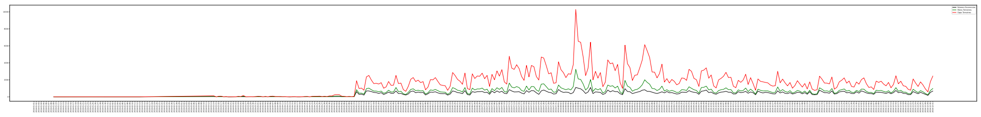

Figure 9: Time series of the number of occurrences, number of terrestrial means and vehicles used per day in 2016

3.3.3 Additional Data Gathered

- **Density**: By using a dataset containing the population density of every municipality [2], we added this data to our dataset by matching it with the *Concelho* column. This helped in labeling these municipalities' districts as rural or urban.
- **Weekday**: Every row was given another column called *Weekday* which contained a number, Monday being 0 up to Sunday being 6. This was used for analysis based on a weekly timeframe.
- **Holiday**: We used the date and municipality to account for national and municipal holidays, so every row got a new column that was *True* if the occurrence was during a national holiday or a municipal one. This calculation also accounts for all holidays whose date varies by year, such as Carnival and Easter. This accounted for 4,3% of all occurrences, which may have been influenced by the special date.
- **Atmospheric measurements** Initially we had planned on including a column for *DailyTemperature* and *PrecipitationLevels* based on date and location, but no free dataset was found after a lot of search. However, more work done on this subject by a bigger entity with more monetary resources can work on this aspect, which we expected to yield a lot of good data and better conclusions.

3.3.4 Data Aggregation

Occurrences were aggregated in blocks of 0.25 degrees of latitude * 0.25 degrees of longitude * 1 day, these blocks had 4 features. Addition of nonprediction features in this part of the process which are related to the geography or the specific day or the block, such as population density or if the day was a holiday or not, can be added as future work. This results in data of $T * N_f * H * W$ dimensions

3.4 Modeling

3.4.1 Convolutional LSTM

Convolutional LSTM [3] is a special type of recurrent neural network that is designed to predict spatiotemporal data, which is data that has both spatial and temporal dimensions. This type of network uses convolutional structures in both the input-to-state and state-to-state transitions to determine the future state of a specific cell in a grid. By analyzing the inputs and past states of the local neighbors, a convolution operator is applied to the transitions to make predictions.

If we consider the states to be hidden representations of moving objects, a Convolutional LSTM with a larger transitional kernel is better equipped to capture faster motions, while a smaller kernel is more suitable for slower motions.

To ensure that the states have the same dimensions as the inputs, padding is needed before applying the convolution operation. The padding of the hidden states on the boundary points allows the model to take into account the state of the outside world when making predictions. Additionally, before the first input is processed, all the states of the LSTM are initialized to zero, which helps the model start with no prior knowledge of the future.

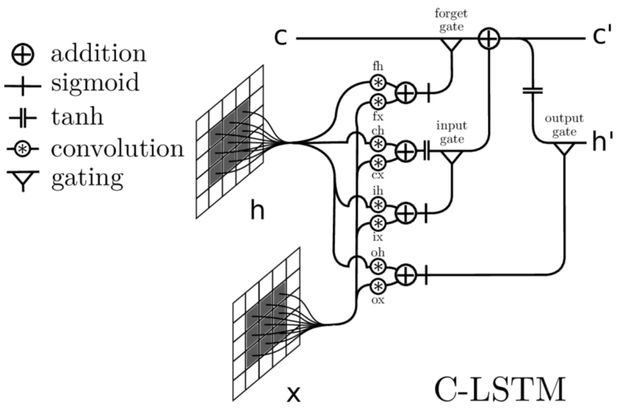

Figure 10: The inner workings of a Convolutional LSTM [1]

3.5 Implementation

We used the Pytorch implementation of the ConvLSTM available in https://github.com/ndrplz/ConvLSTM_pytorch, any quantity of layers is supported by the ConvLSTM class. In this scenario, both the kernel size of each layer and the hidden dimension (i.e., the quantity of channels) can be chosen. If there are other levels present but only one value is given, all of the layers receive the same value.

Train and test sets were split with a ratio of 0.80, and windowing was performed with size 12.

The parameters for the ConvLSTM were:

- $hidden\_dim = [64, 64, 4]$
- $kernel\_size = (3, 3)$
- $num\_layers = 3$

The model was trained during 30 epochs with a $learning\_rate = 0.04$.

3.6 Evaluation

Two evaluation metrics were chosen *Rooted Mean Squared Error* and *Mean Scaled Average Error*, the best-performing model obtained achieved 13.398 and 1.409, respectively. These values, according to the industry standard, are in agreement with the industry standard considering the dataset we had. In the future, these values could be lower with further tuning of the training parameters and longer training times.

The results obtained were heatmaps representing the density of the number of land units, land vehicles, air units, and air vehicles that should be present in a given time to help the population in possible incidents.

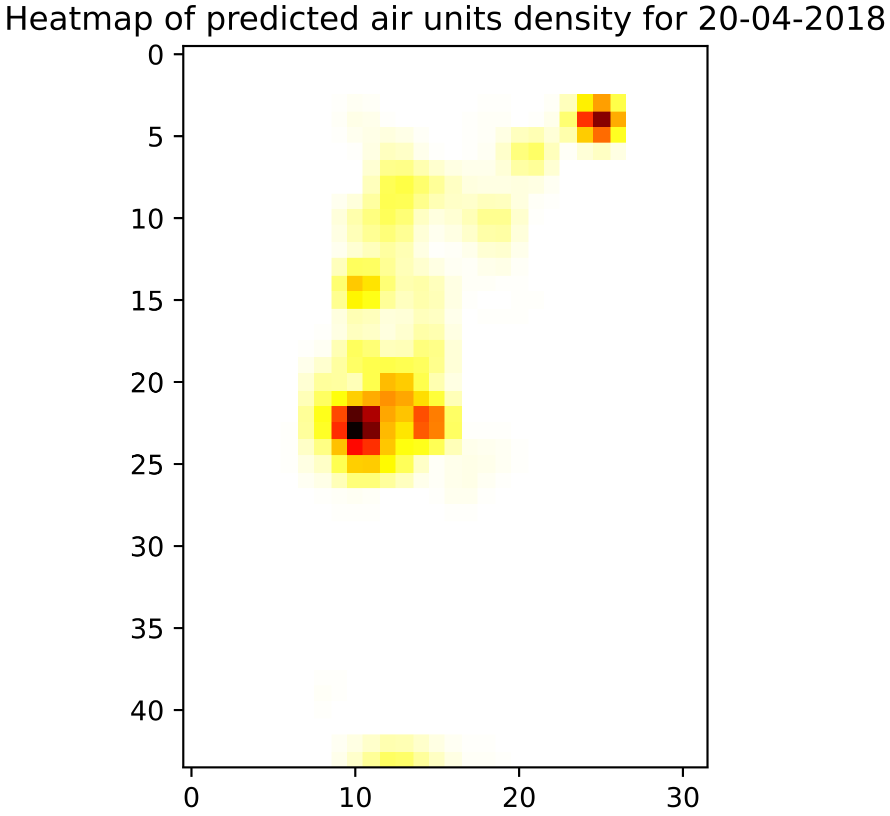

Figure 11: Heatmap prediction for the air units

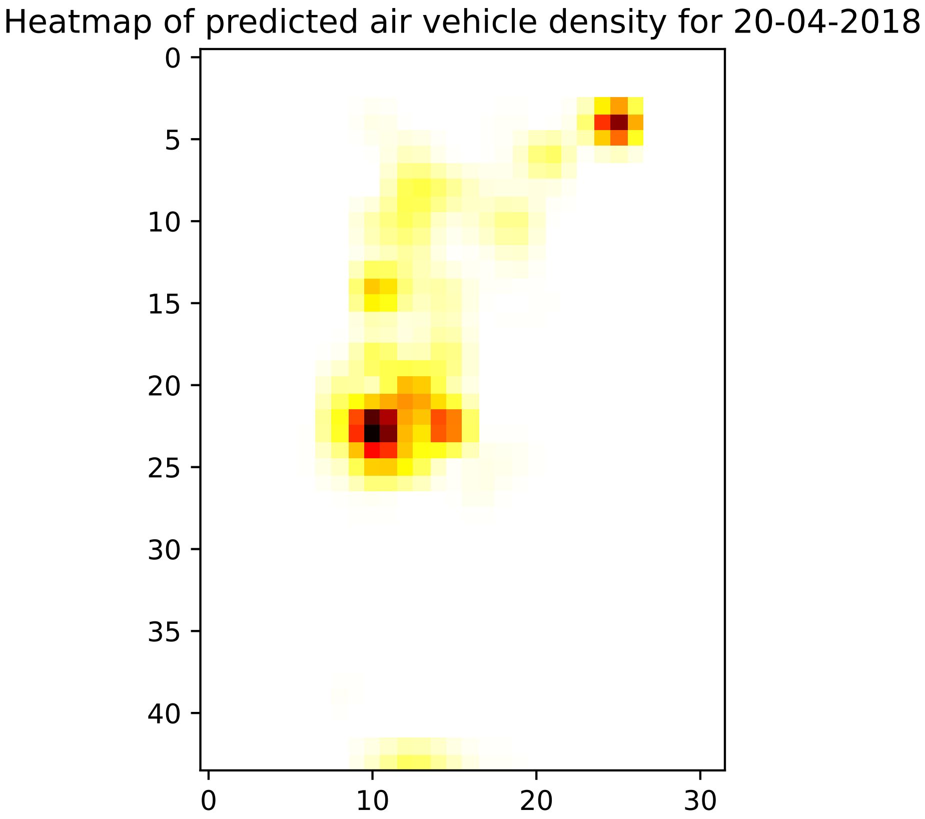

Figure 12: Heatmap prediction for the air vehicles

4 Prototype

While developing and analyzing the data from this project we thought it would be cool to somehow implement a frontend that would not only help us in viewing the data but also allow us to dynamically see the changes in occurrences over time and also the predicted occurrences from our predictive system.

However, we did not have time to fully implement this.

5 Conclusions

First of all, our work was more focused on helping the Portuguese authorities to better visualize the data and letting them elaborate on the measures to be taken according to the available data. This is due to their better understanding of their internal logistics and management of resources. Therefore, the conclusions drawn are more about the technical side of the work done, and less about what this data leads us to believe.

References

[1] M. Stollenga, *Advances in Humanoid Control and Perception.* PhD thesis, 05 2016.

[2] Pordata, "Densidade populacional." https://www.pordata.pt/Municipios/Densidade+populacional-452, accessed 2022. Accessed on: April 23, 2023.

[3] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W. kin Wong, and W. chun Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," 2015.