INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Data-driven disaster management in a smart city

Sandra de Jesus Pereira Gonçalves

Master's in Integrated Business Intelligence Systems

Supervisor:
PhD João Carlos Amaro Ferreira, Assistant Professor with habilitation,
ISCTE - University Institute of Lisbon

Co-Supervisor:
PhD Ana Maria Dias Madureira Pereira, Assistant Professor,
ISEP

October, 2021

TECNOLOGIAS
E ARQUITETURA

LISBOA
CÂMARA MUNICIPAL

Department of Information Science and Technology

Data-driven disaster management in a smart city

Sandra de Jesus Pereira Gonçalves

Master's in Integrated Business Intelligence Systems

Supervisor:
PhD João Carlos Amaro Ferreira, Assistant Professor with habilitation,
ISCTE - University Institute of Lisbon

Co-Supervisor:
PhD Ana Maria Dias Madureira Pereira, Assistant Professor,
ISEP

October, 2021

Aknowledgements

I thank and dedicate this dissertation to my family, my mother Iria, my father Víctor, my sister Dulceneia and my brothers Adilson, Jeremias and Bruno for their love, encouragement and unconditional support.

To my closest friends, thank you for always believing in me and always motivating me to keep going. A debt of gratitude is also owed to Marisa Faria for all her support during this journey.

I would also like to show my gratitude to my supervisors Dr. João Carlos Ferreira and Dr. Ana Maria Madureira for their teachings, support and incentive that have helped me to a very great extent to accomplish this task.

Sandra de Jesus Pereira Gonçalves

Resumo

Os desastres, tanto naturais quanto as provocadas pelo homem, são eventos complexos que se traduzem em perdas de vidas e/ou destruição de propriedades. Os avanços na área de Tecnologias de Informação e *Big Data Analysis* representam uma oportunidade para o desenvolvimento de ambientes resilientes dado que, a partir da aplicação das tecnologias de *Big Data* (BD), é possível não só extrair padrões de ocorrências dos eventos, mas também fazer a previsão dos mesmos.

O trabalho realizado nesta dissertação visa aplicar a metodologia CRISP-DM de forma a conduzir análises descritivas e preditivas sobre os eventos que ocorreram na cidade de Lisboa, com ênfase nos eventos que afetaram os edifícios.

A investigação permitiu verificar a existência de padrões temporais e espaciais eventos a ocorrer em certos períodos do ano, como é o caso das cheias e inundações que são registados com maior frequência nos períodos de alta precipitação. A análise espacial permitiu verificar que a área do centro da cidade é a área mais afetada pelas ocorrências sendo nestas áreas onde se concentram a maior proporção de edifícios com grandes necessidades de reparação.

Por fim, modelos de aprendizagem automática foram aplicados aos dados tendo o modelo *Random Forest* obtido o melhor resultado com accuracy de 58%.

Esta pesquisa contribui para melhorar o aumento da resiliência da cidade pois, a análise desenvolvida permitiu extrair insights sobre os eventos e os seus padrões de ocorrência que irá ajudar os processos de tomada de decisão.

Palavras-Chave: Gestão de desastres, Data Mining, Aprendizagem Automática, Cidades Inteligentes

Abstract

Disasters, both natural and man-made, are complex events that result in the loss of human life and/or the destruction of properties. The advances in Information Technology (IT) and Big Data Analysis represent an opportunity for the development of resilient environments, since from the application of Big Data (BD) technologies it is possible not only to extract patterns of occurrences of events, but also to predict them.

The work carried out in this dissertation aims to apply the CRISP-DM methodology to conduct a descriptive and predictive analysis of the events that occurred in the city of Lisbon, with emphasis on the events that affected buildings.

Through this research it was verified the existence of temporal and spatial patterns of occurrences with some events occurring in certain periods of the year, such as floods and collapses that are recorded more frequently in periods of high precipitation. The spatial analysis showed that the city center is the area most affected by the occurrences, and it is in these areas where the largest proportion of buildings with major repair needs are concentrated.

Finally, machine learning models were applied to the data, and the Random Forest model obtained the best result with an accuracy of 58%.

This research contributes to improve the resilience of the city since the analysis developed allowed to extract insights regarding the events and their occurrence patterns that will help the decision-making process.

Keywords: Disaster Management, Data mining, Machine Learning, Smart City.

Index

Tables Index

Figures Index

List of abbreviations

BD - **B**ig **D**ata

CP - Civil **P**rotection

CRISP-DM - Cross **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining

DM – **D**ata **M**ining

IoT - Internet **o**f **T**hings

IT - Information **T**echnology

LFBR -Lisbon **F**ire **B**rigade **R**egiment

LMA - Lisbon **M**etropolitan **A**rea

LMCPEP - Lisbon **M**unicipal **C**ivil **P**rotection **E**mergency **P**lan

MCP -Municipal **C**ivil **P**rotection

MP – **M**unicipal **P**olice

NACP – **N**ational **A**uthority for **C**ivil **P**rotection

NSPR - National **S**ystem of **P**rotection and **R**escue

SC – **S**mart **C**ity

TIU - Territorial **I**ntervention **U**nit

Chapter 1 – Introduction

1.1 Topic context

Disaster is defined, according to Kobiyama et al. [1] as the result of adverse events, natural or man-made events, on a (vulnerable) ecosystem causing damage to human life, to properties and/or to the environment, with consequent economic and social losses.

Disasters, both natural and man-made, have been occurring more frequently around the world. In this way, it becomes crucial to implement disaster management techniques to minimize the risks associated [2].

Disaster management is a set of organized processes that includes the organization and management of activities related to the different phases of a disaster - mitigation, rescue, response, and recovery. These disaster response activities are executed through collaboration between multiple entities, where the main objective is to integrate a set of interrelated tasks to provide adequate resources to analyze, control, and predict disasters [3].

As mentioned above, disasters can be categorized into two types - natural and man-made. According to Wellington and Ramesh [2], natural disasters include phenomena such as floods, tsunamis, cyclones, earthquakes, droughts, volcanic eruptions, landslides, and forest fires. On the other hand, phenomena such as terrorist attacks, structural damage, nuclear accidents, chemical accidents, and road accidents can be assigned to the category of man-made disasters. All these accidents have been occurring with more frequency and severity, owing their causes to urbanization and globalization.

In fact, in the last ten years, 3 751 natural disasters such as earthquakes, tsunamis, and floods were recorded worldwide, representing total damages of $1 658 billion and impacting more than two billion people [4]. In this way, need to analyze the spatio-temporal patterns and general dynamics underlying the occurrence of such calamities arises. These analyses are becoming essential for the development of efficient strategies to mitigate the negative effects of disasters and thus protect people and properties [5].

According to Shah et al. [3], the increase in population density in cities and the increase in the frequency of disasters in recent years arises the need for cities to provide better services and proper infrastructures to their population. In this context, the concept of

Smart City (SC) emerges, considered the ideal solution to overcome the challenges brought by globalization and urbanization.

In a SC, electronic devices and network infrastructures are incorporated to obtain high-quality services and as the cities get the latest network infrastructure, smart devices, and sensors, a substantial amount of data is generated, known as BD. This data can contain large amounts of information that can be contextual, spatial, or temporal [6].

The data generated from the infrastructures that integrate an SC enables the improvement of the way in which disaster situations are managed, since the BD technologies, by enabling data collection, data storage, data fusion, data analysis, data representation, and data visualization allow the application of Data Mining (DM) techniques to find patterns and thus increase the capacities of the authorities to cope with emergency situations.

In conclusion, the application of BD technologies assists agents in the decision-making process, since they enable identifying potential risks and, consequently, the development of appropriate strategies to deal with disasters, building an important tool to increase the resilience of the SC [4].

The concept of resilience in urban contexts is linked to the idea of resistance, an effort to fortify the city in order to anticipate the problems that communities are subject to and proactively seek solutions that can improve the quality of both public and private living spaces [7], therefore, city management allied with BD technologies represents a way of an informed and efficient management based on analytical results.

1.2 Motivation and topic relevance

Disaster management can be characterized as a multifaceted process where the primary goals is to avoid, reduce, respond, and recover from disaster impact on the system. Due to the complexity of these events, disaster response involves different organizations such as governmental, public, and private organizations as well as different layers of authority [8]. In this way, the involvement of different entities in the disaster management processes highlights the need for collaboration and coordination mechanisms since these agencies, to be effective in a disaster situation, need to communicate, coordinate, and collaborate with each other. Some factors may difficult the communication between

stakeholders, such as lack of situational awareness or difficulty in adopting technological systems for disaster response since they represent high costs [9].

The interdependence between different institutions emphasizes the need to implement decision support systems to assist the stakeholders in the decision-making processes. In this way, problems such as situational awareness and the elaboration of a shared operational framework among all stakeholders (who often have a partial view of the situation), become the most important needs of the disaster management area [8].

On the other hand, disaster events are complex and dynamic situations that put at risk a significant amount of assets, infrastructures, and communities. The combination of complexity and dynamics makes disasters difficult to understand and complex to predict [8]. As follows, unpredictability implies that the impacts of disasters on people and material assets cannot be predicted. It is in this context that the issue of resource allocation emerges since unpredictability prevents resources from being allocated in a timely manner. Therefore, the adoption of disaster management policies in combination with the application of adequate levels of IT allows the enhancement of the authorities capabilities when facing disaster scenarios [10].

Advances in the IT area have positively impacted the field of disaster management as the various stakeholders involved have begun to have access to knowledge extracted from unprecedented volumes of data. According to authors Gupta, Nair, and Röder [11], disaster data can be used in the development of a risk management system as it allows a spatio-temporal analysis of disaster occurrence patterns, building an important tool for effective disaster management since it enables prioritization of mitigation measures as well as analysis of how development practices have increased or reduced disaster impact in a given community.

In a short, data concerning disasters have significant value for the disaster awareness phase. However, it is still necessary to apply analytical methods in order to integrate and analyze in a credible way data that may come from various sources. Only in this way can disasters be characterized from the standpoint of situational awareness and in terms of their underlying spatio-temporal patterns. Therefore, decision-making processes can be based on the analytical results improving the authority's response capabilities [8].

1.3 Questions and research goals

Considering the problems mentioned above, the research carried out in this dissertation aims to apply a data-driven approach to extract information about disasters in the context of a SC in order to contribute to improving the way the city is managed. Therefore, DM techniques are going to be applied to extract knowledge from data concerning events in the city of Lisbon to improve disaster management and, consequently, increase the city's resilience.

Through the application of the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, the objective is to perform a descriptive and predictive analysis of the data provided by the Lisbon City Hall that contains information regarding incidents that occurred in the city. This analysis is going to be performed using two different data sources, where the first dataset corresponds to data regarding occurrences registered by firefighters between the years 2011 and 2018. The second dataset comes from the application *Na Minha Rua Lx*, which is an application for intervention request management in the city of Lisbon and this dataset contains data from 2017 to 2020.

In general terms, this analysis is going to be carried out in two level of analysis wherein in the first moment, the objective is common to both datasets. With this is mind, in this phase, a general analysis of the types of occurrences recorded and their frequency in both dataset is going to be carried out.

In a second moment, the analysis will focus only on events that affected buildings, and in this case, specific objectives were defined for each dataset. For the firefighters' dataset the following objectives were defined:

1. Characterize the types of events reported.
2. Identify the regions with the most occurrences.
3. Verify if there is an association between the incidents and the period of the year in which they occur.
4. Understand how human and material resources are allocated depending on the type of event.
5. Apply predictive algorithms to predict the occurrences.

Regarding the dataset coming from the application, due to the lack of richness of this dataset, only descriptive analysis is going to be carried, out and the following objectives were defined:

1. Analyze the type of occurrences that affected buildings.
2. Analyze the occurrences from the temporal perspective.
3. Analyze the occurrences according to the region.

To achieve the proposed objectives, two data analysis tools are going to be used, namely, the Python programming language [12] and a Microsoft data analysis tool named Power Bi [13].

1.4 Methodologic approach

This study is based on the CRISP-DM methodology, a standard methodology in DM processes [14]. As shown in Figure 1, this methodology is characterized as an iterative process that provides a framework for extracting knowledge from data.

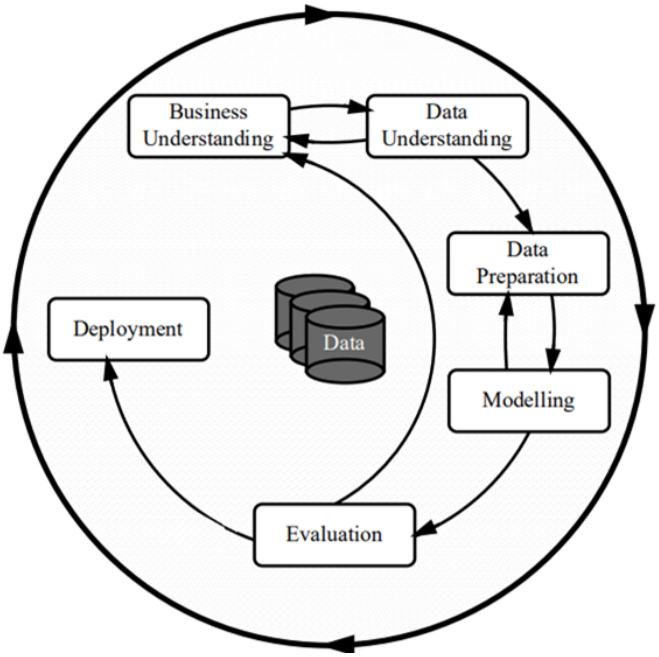


Figure 1 - CRISP-DM phases. Source: Chapman et al. [15]

The CRISP-DM model assumes six stages, which includes business understanding, data understanding, data preparation, modeling, evaluation, and deployment. However, in this dissertation, the steps of the methodology are not going to be presented in full as

only the first 5 phases are going to be presented, that is, business understanding, data understanding, data preparation, modeling, and evaluation.

The reason why the deployment phase will not appear in this dissertation is due to the fact that this phase corresponds to the organization and delivery of the results obtained from this research. In this sense, the knowledge acquired is organized in a report and presented to the Lisbon City Hall stakeholders in a suitable way so they can use it to improve the disaster management.

Applying this methodology, this process starts with the business understand where the focus is on understanding the project's main objectives from a business perspective.

The second step concerns the data understanding where the familiarization with the available data is done and the exploration to discover the problems related to data quality and draw the first insights from the variables.

The third phase is dedicated to data preparation. This phase covered the activities designed to build a final dataset that is going to be used in the following phases. In this step, null, duplicate, and outliers are going to be checked, and data cleansing techniques are going to be applied. Furthermore, at this phase, attribute selection is going to be conducted, new variables are going to be constructed from the original variables, and finally, new variables are going to be added from external data.

The next phase, corresponding to modeling, is focused on applying DM techniques according to the objectives initially defined. In this phase, the activities are oriented towards data visualization, in order to find patterns and consequently extract information that would allow the achievement of the proposed objectives. The application of DM techniques is also associated with predictive analyses to predict the events in the city of Lisbon.

The last phase is the evaluation, where the project is evaluated by those responsible for the TIU department at the Lisbon City Hall in order to verify if the defined objectives were achieved the project.

1.5 Structure and organization of dissertation

Considering the objectives outlined, this dissertation is organized into five chapters, including the Introduction (chapter 1), where an overview of the problem studied is

presented, and the general objectives of the research. Also included in the introduction is the presentation of the CRISP-DM methodology as the methodology adopted to conduct this research.

Chapter 2 is focused on the theoretical framework, or in other words, literature review. This chapter presents state of the art on data-driven disaster management, based on a survey conducted to identify relevant literature on the topic under study.

Chapter 3 presents the first three stages of the implementation of the CRISP-DM methodology through the presentation of business understanding, data understanding, data preparation.

Chapter 4 presents the implementation of the remaining phases of the methodology, namely data modeling and evaluation.

To conclude, chapter 5 presents the conclusions of the present study as well as the identification of the challenges to be considered in future work.

It is important to note that these steps of the CRISP-DM methodology, except for the business understanding, were applied to both datasets, separately, since it was not possible to merge the two datasets. The reason for the impossibility of merging the two datasets is due to the fact that there was no point of interest between the datasets, that is, a column that was common to both datasets and that had the exact same values. The join between the datasets could be made through the parish column, which is common to both, however, the datasets did not have a complete correspondence between the parishes and data could be lost when merging.

Another factor that made it impossible to merge the datasets is due to the fact that the datasets do not have the same dimension in terms of rows. If the join is made taking into account the *left* and *right join* criteria, data could be lost if the join was made by the firefighters' dataset, since it is smaller. Also, if the join is made using the dataset with the application data, there would be many fields with null values since the dataset is larger in terms of rows.

Chapter 2 – Literature review

2.1 Introduction

Cities that aim to become a smart city use digital and networked technologies to address different types of problems, such as improving the quality of services, becoming more sustainable, growing the local economy, improving the quality of life, and increasing the safety and security of their inhabitants [16]. In this way, emergency response systems are among the main dimensions of a smart city due to the increase in disruptions caused by natural or man-made disasters [17], which have severe and negative effects on society.

In this context, to provide protection and safety to the smart city, it is essential to develop appropriate strategies to protect the city from such hazards. To this purpose, it is possible to apply DM and analysis techniques to analyze patterns and predict disasters, allowing the development of appropriate disaster management strategies from the data collected from disasters that have occurred in the past [16].

Data-driven disaster management is a recent area that has been undergoing an evolution due to the number of works that have been developed [18]. In this sense, this literature review aims to present an overview of the state of art on data-driven disaster management, where it intends to survey the contributions on the topic and thus outline the evolution of knowledge in this area.

This state of the art is composed of two sections: section 2.1 includes the description of the research methodology adopted, the identification of the keywords used in the research, and the characterization of the analysis tools used in the research. Section 2.2 is dedicated to the presentation of the research results.

2.2 Review Methodology

The survey and critical appreciation of the literature related to the proposed theme were performed by applying the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology [19] in accordance with the Systematic Literature Review steps proposed by Okoli and Schabram [20]. Accordingly, a systematic search on the topic was conducted in two electronic databases: Scopus [21] and Google Scholar [22], and the main objective was to identify and select research papers related to data-driven disaster management in the context of a smart city.

The research process was conducted to identify papers whose title, abstract, and keywords included terms such as "Disaster management", "Data mining", "Smart City" and "Machine Learning". With this in mind, a query was formulated to make the selection of the works carried out in this area. The query is the following:

```
((("Disaster Management" OR "Incident Management") AND ("smart city" OR "data analysis" OR "data mining" OR “big data” OR "machine learning"))
```

Additionally, a ten-year time window was defined (2010-2020), and the search covered areas such as Decision Science, Computer Science, Environmental Science, and Engineering.

In terms of document typology, only journal articles, articles, and book chapters were considered. In a first analysis, it was found that the electronic databases shared most of the documents, i.e., documents found in Scopus were also found in the Google Scholar database.

The documents were selected through the abstract and in cases where the information contained in the abstract was not sufficiently complete, the document was consulted in its entirety and all the selected documents were stored in the reference management software, Zotero [23]. This tool besides allowing the management of bibliographic information also has a feature that allows the identification of duplicate documents (documents linked to several databases) [24].

The PRISMA methodology as a tool to systematically conduct systematic review studies and meta-analyses provides a flowchart to illustrate the literature review process. In other words, the flowchart illustrates the process of identifying and selecting articles for the systematic review by mapping the number of identified studies, the number of included and excluded studies, and the reason for the exclusions [25].

Figure 2 presents the practical application of the PRISMA methodology flowchart as well as the information related to each step of the process.

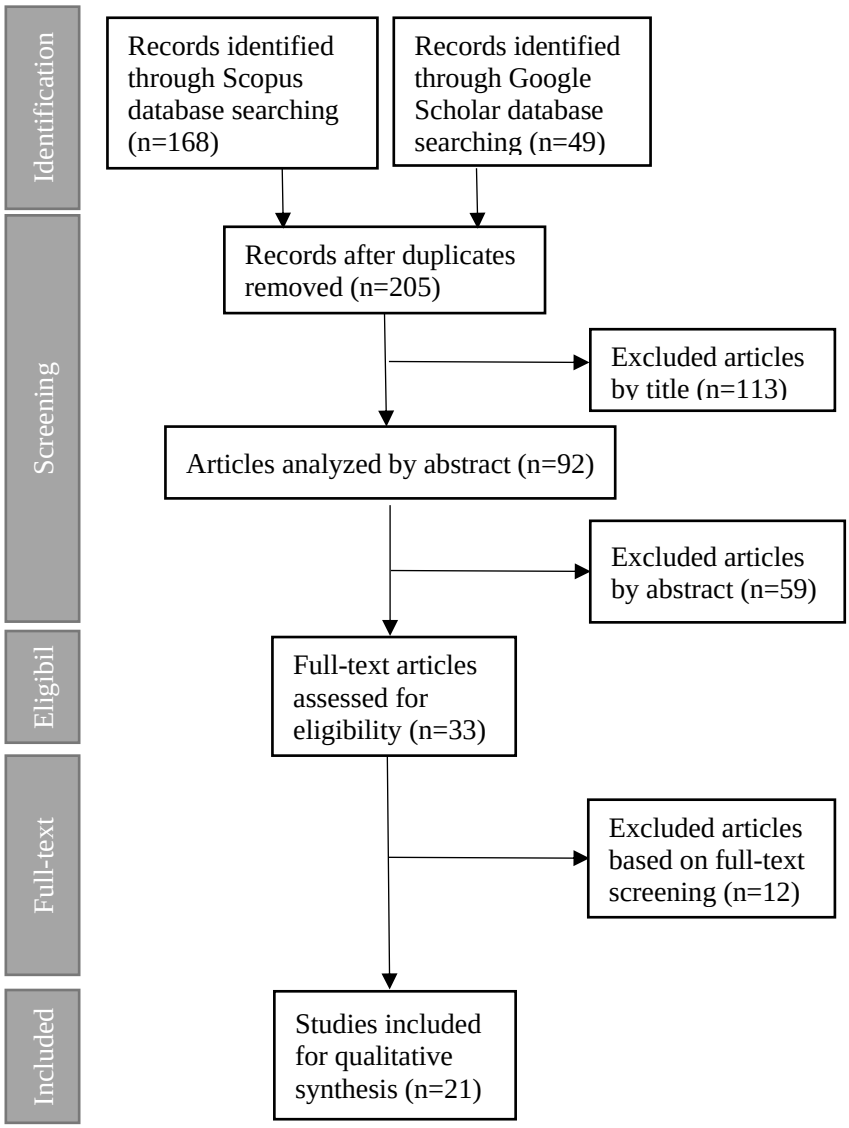


Figure 2 - PRISMA flow diagram

This process began with the identification of the publications in the two electronic databases where the searches were conducted, resulting in a total of 217 publications (Scopus:168 and Google Scholar:49). Before proceeding to the exclusion of articles taking into account the exclusion criteria, duplicates (12 articles) were removed.

The next step consisted in the exclusion of publications considering the title, and publications whose title was outside the scope of the topic of this dissertation were excluded. The application of this criterion resulted in the exclusion of 113 articles.

Next, abstract screening was conducted where publications were excluded through the information provided in the abstracts. At this stage, 59 articles were excluded.

At this stage the systematic review was reduced to 33 articles and these articles were targeted for full-text screening which resulted in the exclusion of 12 articles. After the

full-text screening stage was complete, 21 articles were considered eligible for the systematic review.

2.3 Results

The first analysis carried out on this set of articles was to get an overview of the countries where these studies were conducted and the year of their publication. The bar chart containing this information are presented in Figure 3 and Figure 4.

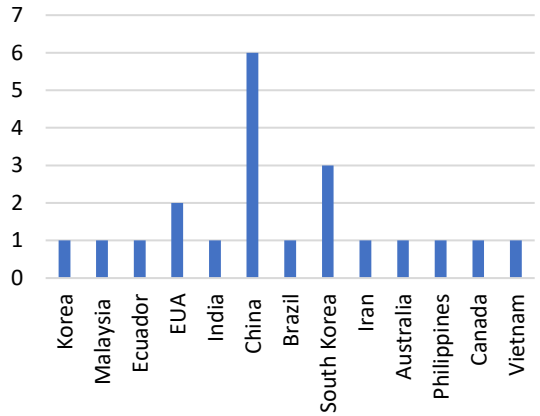


Figure 3 - Number of surveys per country

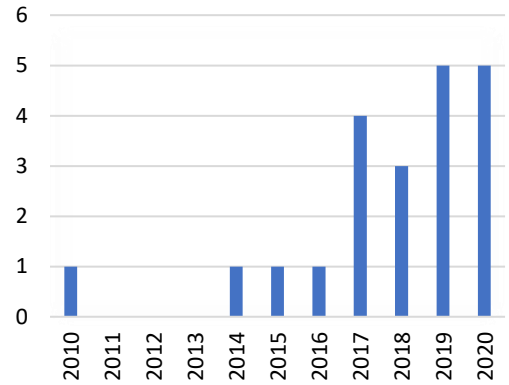


Figure 4 - Number of surveys per year

From the bar chart containing the number of surveys per country it, was possible to verify that most researches were developed in China (6 articles), followed by South Korea (3 articles) the USA (2 articles). The remaining countries present only one article on research in the area of data-driven disaster management.

Regarding the number of published articles per year, it was verified that in recent years there has been an increase in the number of studies developed in this area. During the first 3 years, i.e., between 2010 and 2013, only one work was published. From the year 2014 on, more papers started to be published and there was an increase in the number of studies carried out in this area, with the main highlight being the years of 2019 and 2020, with five publications each.

2.3.1 Natural disasters

Natural disasters are events that are characterized by the substantial impact they cause on society, interrupting its normal functioning. To mitigate the negative effects of these types

of events, it is necessary to analyse their patterns of occurrence, both spatial and temporal, and to apply techniques that enable disaster prediction and thus handle these emergencies situations in a timely manner.

Work has been done in this area of data-driven disaster management to provide decision support systems that assist decision makers in making decisions in a faster and more informed way, that is, based on analytical results. It was with this purpose that Jeong and Kim [26] developed a research where they conducted a statistical analysis of electrical incidents such as fires or failures occurring in Korea caused by climate changes that manifest themselves through unusual weather conditions such as high temperatures or sudden temperature drops. Weather changes such as heavy rains, floods, snow, typhoons, or lightning cause damage to electrical equipment, which can lead to malfunctions or fires. This study thus established a relationship between climate change and accidents involving electrical equipment.

Another study [27] conducted in 2017, reflected the link between BD systems and disaster management. Big Data Analytics technologies was implemented on a dataset from the National Hydraulic Research Institute of Malaysia in order to analyze the hydroclimate data. The goal was to extract insights on climate change and thus provide information to prepare, mitigate, respond and recover from natural disasters. The application of BD technologies allowed the detection of periods of extreme precipitation and runoff as well as tracing drought episodes.

Through the application of DM techniques, also in 2017, another research was developed by the authors Briones-Estébanez e Ebecken [28] to identify and analyse the patterns in the occurrence of extensive and intensive events, including floods, river overflows and landslides, related to precipitation intensity in five cities in Ecuador. The patterns were identified by applying the K-means technique and association rules between a given event and the amount of precipitation. In this way, patterns were identified in the relationship between the type of event and the precipitation level.

In addition to works developed to analyze disasters from a spatial and temporal perspective, other works have been developed to conduct a quantitative analysis of the damage caused by natural disasters. This is the case of the analysis carried out by the authors Alipour et al. [29]. They present a systematic framework that takes into account the different aspects that explain different types of risk (such as vulnerability and

exposure) and apply Machine Learning models to predict the damage caused by flash floods in the Southeast, US.

In this study the Random Forest model was applied in two distinct moments - the first moment as a binary classifier to predict if a region of interest suffered damage during a flash flood and the second moment the model was used to predict the amount of damage associated with each type of event.

With a similar approach, Park et al. [30] conducted a study aiming to quantify the possible effects or effectively the damage caused by three types of disasters namely typhoons, heavy rain, and earthquakes on water supply systems in Korea. For this, the damage prediction was predicted by applying the Random Forest and XGBoost models.

The work done in the area of data-driven disaster management is diverse as various techniques are adopted to make information available to decision makers. In the case of the study carried out by Saha et al. [31], they presented the analytical results in more iterative way by developing a dashboard to predict and identify areas vulnerable to flooding in West Bengal, India, using geographic map visualization.

Other research has also adopted the geographic map visualization method to identify and classify areas vulnerable to natural disasters as in the case of the study conducted by Zhou et al. [32] that assessed the integrated and relative risks of five major natural disasters (drought, earthquake, flood, low temperature/snow and windstorm) at the provincial scale in China, to identify regions with relatively higher multi-risk risks.

Another study [33] using geographic map visualization was conducted in Brazil to estimate the population living in a landslide and/or flood risk zone. To achieve this goal, the databases of the demographic census and the mapping of risk areas were crossed. The knowledge of the population in risk areas contributes to the identification of the most critical areas that require priority response actions in disaster situations, such as areas with a higher presence of elderly people, children and a higher concentration of residents in households without sanitation.

Lee et al.[34] also made their contribution in this area of data-driven disaster management with the development of a study that combined DM and Geographic Information System (GIS) methods [35] where they established the relationship between flood areas with hydrological factors with the main objective of mapping the flood susceptibility of the Seoul metropolitan area in South Korea. To this end, two DM models

Frequency Ratio and Logistic Regression were applied to create a flood risk map, and thus provide a tool to help decision makers implementing mitigation strategies in priority areas contributing to sustainable urban development.

In another research, the authors Liu et al. [36] used DM and GIS techniques to assess seismic vulnerability at the urban scale in the city of Urumqi, China. The combination of these methods allowed extracting relevant information through a dataset containing characteristics of the infrastructures and the population of the city allowing to represent and estimate the systemic risk of this city. Therefore, two DM models were applied, namely, Support Vector Machine and Association Rule Learning to establish the relationship between the characteristics of buildings and their vulnerability classes, allowing the construction of risk maps.

Other studies have also used a combination of DM and GIS techniques to construct disaster susceptibility maps. The central objective of these studies focuses on the identification and classification of vulnerable areas to natural disasters with the difference in the DM models used in the different research works. For example, Random forest, and Naïve Bayes were applied to extract insights and construct the landslide susceptibility map in the area of Longhai, China [37]. Evidential Belief Function Model, Random Forest and Boosted Regression Tree were applied individually to compare the effectiveness of these methods in flood susceptibility mapping in Iran [38]. Support Vector Machine and artificial neural network models were applied to predict landslide susceptibility in Seoul, South Korea [39]. Random forest and Radial Basis Function Neural Network were used for identification and risk assessment of regional flood disasters related to Yangtze River Delta, China [40]. Lastly, Lu et al. [41] took a new approach and applied multiple linear regression-TOPSIS, to analyse flood incident data having analysed and ranked flood risk for the 63 provinces and eight regions of Vietnam, which was used to construct a nationwide flood risk map.

2.3.2 Man-made Disasters

Regarding man-made disasters, Smith et al. [42] developed a research that consisted in the implementation of Big Data technologies for disaster management. They used the statistical tool R, as well as its visualization capabilities, to analyse a dataset regarding

fires that occurred in Australia. The goal was to determine the optimal response time for firefighters, thus minimizing losses.

Still in the context of fire data analysis Balahadia et al. [43] applied the K-means clustering algorithm to generate patterns and create clusters of fire events based on the recorded data of fires that occurred in the city of Manila, Philippines. In summary, the goal was to obtain characteristics of fire events that can be used for risk assessment and risk management concerning these types of disasters as well as to assist in the development of prevention measures.

In the study of Asgary et al. [44] an attempt was made to use spatiotemporal methods to analyse the spatial and temporal patterns of fire-related incidents in Toronto, Canada. Insights were extracted by analysing the relationship between the economic, physical, and environmental aspects of various neighbourhoods and the total number of fires that occurred in those neighbourhoods. In this way, they analysed how fire patterns vary depending on the time of day, the day of the week, and the month of the year.

Liu et al. [45] proposed a DM method based on using Bayesian Network to model building fires in urban area. From the historical records of fires in a city in China between 2014 and 2016, they analysed the potential fire risk according to building construction characteristics and external influences. Another study [46] aiming to analyse fire patterns was conducted by applying the Support Vector Machine model to analyse the correlation between building characteristics, occupants and fire incidents in Sydney.

Finally, in a study developed by Wan et al. [47] BD technologies were applied to analyse the distribution and influence factors of harmful gases in the urban underground sewage pipe network of Chongqing city, and explore the impact of smart city developments on harmful gases in the urban underground sewage pipe network.

In short, the literature review allowed to verify that most of the surveys developed in this area were in China and that this area is becoming more relevant since in the last 6 years the number of published surveys have been increasing. It was also found that the research in this field covers natural disasters events as well as man-made disasters and that in the case of natural disasters there is a predominance of analysis of flood incidents and in the case of man-made disasters there is a predominance of analysis of incidents related to fires. It was also possible to verify that from the selected publications eligible for the literature review not one was published in Portugal, which means that this area is

not as developed in Portugal when compared to other countries. However, in general it is possible to verify that data-driven disaster management is an emerging topic and that in the context of smart cities, large amounts of data are generated which enhances conscious management based on analytical results.

Table 1 presents a summary of the papers published in the last ten years as well as their approximation to the work developed in this dissertation

Table 1 - Summary of the publications found in the literature search. (*) Approach to the subject under study

Paper	Year	Disaster	Methodology Applied		(*)
			Descriptive Analysis	Predictive Analysis	
[26]	2019	Fires and equipment failures	x		
[27]	2017	Natural disaster	x	x	x
[28]	2017	Floods, river overflows, and landslides	x	x	x
[29]	2020	Floods		x	
[30]	2020	Typhoons, heavy rain, and earthquakes		x	
[31]	2018	Floods	x	x	x
[32]	2015	Drought, earthquake, flood, low temperature/snow and windstorm	x		
[33]	2019	Landslide and flood	x		
[34]	2018	Floods	x	x	x
[36]	2019	Earthquake	x	x	x
[37]	2018	Landslide	x	x	x
[38]	2017	Floods	x	x	x
[39]	2017	Landslide	x	x	x
[40]	2020	Flood		x	x
[41]	2019	Flood		x	x
[42]	2016	Fire	x		
[43]	2019	Fire	x	x	x
[44]	2010	Fire	x		
[45]	2017	Fire	x	x	x
[46]	2014	Fire	x	x	x
[47]	2020	Toxic gas	x		

Chapter 3 – Data-driven approach for disaster management

The proper identification, characterization, and identification of the patterns of disaster occurrence, both natural and man-made, that affect the safety and consequently the normal operation of communities is one of the main aspects to be considered in a process of the development of disaster response strategies. Although there are losses of human life and destruction of properties, these must be minimized as much as possible.

Cities have infrastructures with their specificities that are different from city to city. For this reason, data analysis must be conducted taking into account the specificities of each city in order to understand the circumstances in which they occur.

This dissertation aims to perform a descriptive and predictive analysis of events that occurred in a specific city, the city of Lisbon, the capital of Portugal. This is a real case where the Lisbon City Hall data is going to be treated to extract knowledge that can assist agents in the decision-making processes. The Lisbon City Hall has data about events that occurred in the city and a reporting system that can be used by citizens to make requests for intervention and in this way have an active participation in the city management. However, this data requires analysis, and this dissertation will contribute information that will support the decision making contributing to improve the way the city is managed.

This chapter is dedicated to the implementation of the first three stages of the CRISP-DM methodology in the case study of this dissertation. The chapter is organized into three sections, where in the first section business understanding is conducted, with the purpose of understanding and identifying the business problems, defining the objectives, and identifying of the data sources needed to achieve the proposed objectives.

In the second section, the data understanding is conducted. This process includes a set of tasks such as the identification of the size of the datasets, the comprehension of the type and meaning of each variable as well as the basic statistics of each variable. Also, in this phase, familiarization tasks are performed to identify problems related to data quality, such as null or duplicate values. Finally, exploratory analyses are going to be carried out to extract the first insights from the data.

In the third section, the data preparation process is conducted. In this phase, the activities are carried out to build the final dataset that is going to be used in the following stages. These tasks include the dataset cleansing that covers the treatment of null values,

duplicates, and outliers. Additionally, in this stage, the relevant variables for the analysis are selected, variables are created from others, and external variables are added to complete the dataset.

3.1 Business Understanding

The business understanding process is important since it allows contextualizing and understanding the problem that is intended to be solved. This chapter aims to understand the context of the business problem through an analysis of the aspects that characterize the city of Lisbon from different perspectives.

Therefore, this chapter is divided into three sections, where the first section characterizes the city geographically, in terms of population, and in terms of the buildings it has. Next, a characterization of the entity responsible for providing assistance in case of disaster situations is made, and lastly, the sources of data necessary to achieve the defined objectives are identified.

3.1.1 Characterization of the city of Lisbon

Lisbon is the capital of Portugal and the largest city in the country. With a population exceeding 500 000 citizens, the city is located on the southeast coast and is the country's main port as well as its commercial and political center [48] where there is a strong concentration of people and activities.

As its entire territory is considered as an urban area, Lisbon corresponds to the central consolidated and densely built-up area of a metropolitan area composed of 18 municipalities where, until 2011, 2 821 876 citizens lived around the Tagus Estuary. The city of Lisbon represents 3% of the territory corresponding to the Metropolitan Area, but the city concentrates 12% of the total buildings, 22% of the dwellings, 21% of the families and, 20% of the inhabitants. [49]. The size of this city in comparison to the Metropolitan area in which it is located as well as the number of buildings and inhabitants make it a densely populated territory.

In demographic terms, the city was densely populated at the end of the 19th century and a large part of the 20th century. However, in the second half of the 20th century, the city began to register a significant demographic decrease as well as an aging population

with difficulties in fixing the young population [50]. As shown in Table 2, between 2001 and 2011 the trend in demographic decline continued, with a loss of 3%. In addition, Lisbon is the oldest city in the Lisbon Metropolitan Area (LMA), with a total of 130 960 elderly people with 65 or more years of age, which represents 25.48% of the total number of elderly people in the LMA.

On the other hand, this demographic decrease is mitigated by the positive evolution of the 0-14 age group, with 7.55% concerning the 2001 census.

Table 2- Resident population in 2001 and 2011, according to the age groups. Source INE[1]

Geographic Area	Resident Population						Resident population - Variation between 2001 and 2011 (%)		
	In 2001			In 2011					
	Total	Age groups		Total	Age groups		Total Var.	Age groups	
	WM	0-14	65 or more	WM	0-14	65 or more		0-14	65 or more
Lisbon Metropolitan Area	2661850	396221	410046	2821876	437881	513842	6,01	10,51	25,31
Cascais	170683	25801	25757	206479	32655	36714	20,97	26,56	42,54
Lisboa	564657	65548	133304	547733	70494	130960	-3,00	7,55	-1,76
Loures	199059	31510	24394	205054	32056	35277	3,01	1,73	44,61
Mafra	54358	8746	8468	76685	14365	11344	41,07	64,25	33,96
Oeiras	162128	22685	24153	172120	26559	32969	6,16	17,08	36,50
Sintra	363749	65987	37311	377835	66633	51657	3,87	0,98	38,45
Vila Franca de Xira	18442	2644	3126	18197	2716	3699	-1,33	2,72	18,33
Amadora	175872	26230	24611	175136	25903	32742	-0,42	-1,25	33,04
Odivelas	133847	19771	16034	144549	21912	23501	8,00	10,83	46,57
Alcochete	13010	2115	2000	17569	3332	2538	35,04	57,54	26,90
Almada	160825	22662	26945	174030	25583	35725	8,21	12,89	32,58
Barreiro	79012	10184	12484	78764	11221	17011	-0,31	10,18	36,26
Moita	67449	11231	8691	66029	10549	11281	-2,11	-6,07	29,80
Montijo	39168	5879	6792	51222	8506	8569	30,78	44,68	26,16
Palmela	53353	8567	8051	62831	10680	10971	17,76	24,66	36,27
Seixal	150271	25092	15127	158269	25747	24433	5,32	2,61	61,52
Sesimbra	37567	6229	5513	49500	8615	7751	31,76	38,30	40,59
Setúbal	113934	17686	16825	121185	19557	21906	6,36	10,58	30,20

In addition to being one of the most aged municipalities, Lisbon is also characterized for being a city with an active growth in terms of buildings when compared with the census of previous years since, at the time of the 2011 census, Lisbon contained 52 496 buildings and 32 398 dwellings [51]. These are buildings constructed before the middle of the 20th century, and a large part of the buildings are in need of repair and others are very degraded [52]. Among the factors that have contributed to the degradation of buildings in Lisbon are the freezing of rents, the decapitalization of property owners, easy

[1] Data extracted and adapted from the INE website on April 7, 2021

access to credit for homeownership, and multi-property (associated with the process of partition, with implications for the decision process to decide the fate of the property). These factors have penalized the city center, which has led to the degradation of a high quantity of buildings, which in many cases has caused the abandonment of the building resulting in an increase in vacant buildings [53].

In terms of altimetry (Figure 5), the city is characterized by areas with altitudes lower than 100m, except the central zone of the *Serra do Monsanto* where the altitude values are higher than 150m, and in the remaining areas the values vary between 70 and 100m. The areas that correspond to the riverside zone and the valleys have altitudes less than 30m and the heights of the slopes are around 70m. In the northern part of the city, *Telheiras* and *Carnide* have altitudes varying between 100 and 150m, and these values are again verified in the *Campolide* area. These altimetric changes are associated with specific circumstances such as the affluent valleys of the Tagus River, the riparian zone of the Tagus River, and the Monsanto mountain range. The circumstances related to the affluent valleys of the Tagus River and the Tagus river margins present lower altitude values and are associated with the circulation and accumulation of water [54].

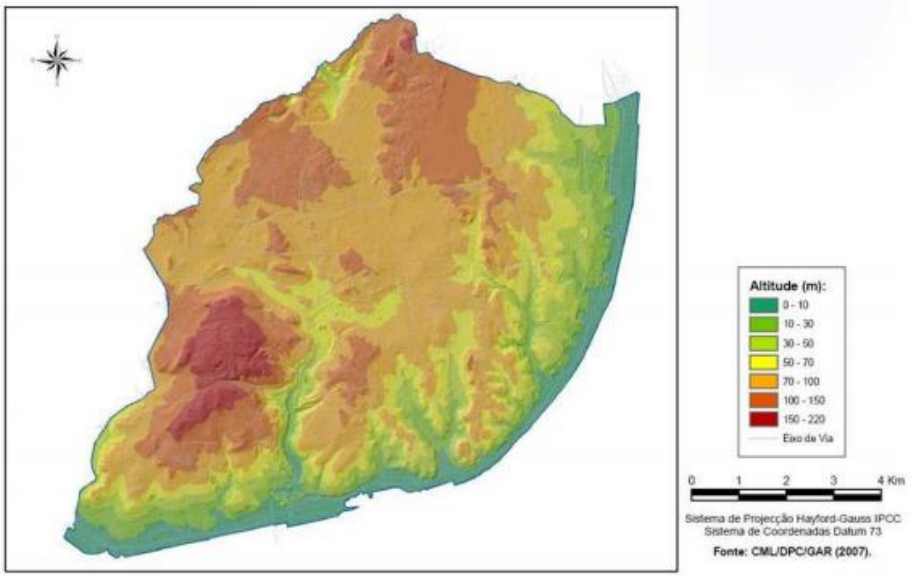


Figure 5 - Lisbon city altimetry. Source: Synthesis Report of Biophysical Characterization of Lisbon [54]

As mentioned above, the city has a riparian zone that is illustrated in Figure 6. It is an area formed by a border adjacent to the coastline of the Tagus River estuary [55] with altitudes lower than 10m that enter the inner city, which makes it vulnerable to flooding situations [56].

Figure 6 - Riparian zone of Lisbon city. Source: Lx_Risk - Geo-environmental characterization[56]

The water lines that flow into the riparian front are of quick drainage due to the accentuated relief of these areas. However, at the meeting point between the various water lines and the riparian front, the surface drainage is decelerated due to the flat relief that characterizes this area. This behavior is a result of the fact that the riparian front is an artificial area where construction has been made over the years by human action through the construction of embankments on the estuary to install port facilities, industrial or transport. This artificial character of the riparian front has the effect of favoring water accumulation, making the point of intersection where other water lines flow into a vulnerable area to the occurrence of urban flooding. [55]. In general, the city's vulnerability to flooding is associated with heavy rainfall situations, and its effects are aggravated when they occur during high tide [57].

In addition to the characteristics mentioned above, Lisbon is also a city vulnerable to situations of slope mass movements. These movements are associated with factors such as the geological morphology of the land and the circulation and accumulation of water [54].

In terms of climatic conditions, according to the Synthesis Report of Biophysical Characterization of Lisbon [54], in general terms, the city has a Mediterranean type of climate, where the summer is characterized by being hot and dry, and the concentration of most precipitation occurs between the months of October and April. The average annual temperature is around 16ºC, with the minimum values registered during the

months of December, January, and February (10ºC) and the maximum values between the months of July and September (20-25ºC). Regarding precipitation values, the average annual values vary between 650mm and 760mm with the minimum values recorded in the months of July and August and the maximum values recorded during the months of November to February.

The city of Lisbon is occasionally affected by unpredictable weather conditions that lead to atypical situations. Among these adverse weather conditions, the following can be highlighted:

- Extreme temperature values, both negative and positive.
- Strong wind or gusts of wind with very high speeds.
- Thunderstorm situations.
- High precipitation values in a short period.

These atmospheric conditions occur depending on the time of year, which allows the identification of two distinct climatic periods. This polarity is explained by the migration in latitude of the band of subtropical high pressures and the low pressures of the middle latitudes.

Therefore, Lisbon is a city vulnerable to various types of risks due to its geological characteristics and the way urban spaces have been built and occupied. To ensure the correct development of response strategies to disaster situations to minimize their effects, the Lisbon City Hall has developed a Lisbon Municipal Civil Protection Emergency Plan (LMCPEP) [58]. It is a document that contains a set of procedures and norms to minimize the damage caused by the occurrence of a disaster situation. This document identifies the risks to which the city is exposed, and these risks can be seen in Table 3.

As shown in Table 3, these risks are categorized according to the type of disaster, namely natural disasters such as diverse meteorological conditions, floods, earthquakes, tsunamis, and mass movements. Included in the category of man-made disasters are traffic accidents, accidents involving the transportation or storage of hazardous materials, collapses of structures, and urban fires. Lastly, mixed disasters include forest fires.

Table 3- Risk Identification. Source: adapted from the Lisbon Municipal Civil Protection Emergency Plan [58]

	Nº	Risk Type
Natural	1	**Diverse weather conditions**
		1.1 Extremes of maximum temperature
		1.2 Extremes of minimum temperature
		1.3 Heavy precipitation
		1.4 Strong wind and gusts
		1.5 strong sea or river agitation and high tide
	2	**Floods**
	3	**Earthquake**
		3.1 Nearby seismic source, intensity VIII
		3.2 Distant seismic source, intensity IX
	4	**Tsunamis**
	5	**Slope mass movements**
Technological	6	**Severe traffic accidents**
		6.1 Aerial
		6.2 Maritime/fluvial
		6.3 Road
		6.4 Railway
	7	**Accidents in the transport of dangerous goods**
	8	**Accidents in the storage of dangerous goods**
	9	**Accidents in pyrotechnic or explosive industries**
	10	**Accidents with radiological establishments**
	11	**Collapses in tunnels, bridges, infrastructures and other structures**
	12	**Urban fires**
Mixed	13	**Forest fires**

Some of these risk situations occur in all regions of the city, others occur in specific areas, some occur cyclically, while others occur suddenly with devastating consequences.

3.1.2 Lisbon Fire Brigade Regiment (LFBR)

With the elaboration of LMCPEP, it was necessary to adopt a system for emergency management that could ensure a concerted and effective response capable of minimizing the losses resulting from an accident or disaster situation. In this way, the responsibility of coordinating and providing assistance during the occurrence of an accident or disaster situation in all areas of the city of Lisbon was assigned to the LFBR [59].

With more than 600 years of existence, the LFBR is the oldest fire department in Portugal. The mission of this entity includes a set of activities in areas such as prevention, protection, and rescue to reduce not only risks, but also minimize the damage caused by the occurrence of major accidents or disasters [60].

According to the yearbook of the LFBR for the year 2012 [60], this entity provides a public security service to people and properties in the entire territory of Lisbon, covering an area of 84 km², with a permanent operational component working 24 hours a day and 365 days per year in a hierarchical structure consisting of seven positions, totaling a staff of 854 elements.

The daily tasks of the LFBR include responding to various types of occurrences - accidents, fire, pre-hospital, infrastructure, industrial technology, legal conflicts, among other activities. In addition to these activities, the LFBR has special units and nucleus (presented in Appendix B- List of special departments and nucleus) to respond to special situations that occur daily.

According to the 2016 Activities Report [61], in operational terms, the LFBR has 11 fire stations, strategically located in order to provide rapid and concerted support to any occurrence in the scope of protection and rescue that occurs in the Lisbon area. Each of these fire stations is responsible for the rescue and protection of citizens in the areas assigned to them, allowing the effectiveness in aiding citizens. All fire stations are equipped with materials and equipment, with the main emphasis on the accident and fire areas where they have vehicles for specific functions such as extrication vehicles or ladder vehicles [62].

The LFBR is an integral part of the National System of Protection and Rescue (NSPR) and has been directing efforts to ensure an effective articulation with the National Authority for Civil Protection (NACP), Lisbon's Voluntary Fire Brigades and other agents of CP. This collaboration between these entities has as its main objective the improvement of the effectiveness in assisting citizens [60].

In this context, one of the main purposes of this dissertation is to conduct an analysis with the objective of characterize, from a temporal and spatial point of view, the occurrences recorded in the city of Lisbon by the LFBR between 2011 and 2018, with emphasis on occurrences that occurred in buildings. To achieve this goal datasets provided by the Lisbon City Hall are going to be used, as well as external data provided by other entities such as IPMA[2] and INE[3] to complete the analysis.

[2] Portuguese Sea and Atmosphere Institute
[3] National Institute of Statistics

3.1.3 Characterization of data sources

The analysis carried out in this dissertation has two distinct focuses that serve the same purpose, i.e., spatial-temporal analysis of occurrences recorded in Lisbon to extract knowledge about their occurrence patterns. This analysis focuses on two datasets, namely the firefighters' dataset and data extracted from the *Na Minha Rua Lx* application. For the firefighter's dataset, external data is going to be used to complement the dataset. Therefore, this analysis is based on 4 data sources - LFBR occurrence management system [62], *Na Minha Rua Lx* application [63], INE Census [64] and IPMA [65].

Regarding the LFBR occurrence management system, it is a platform created in 2009 whose name is "Occurrence Management" and according to Martins [62], this is a pioneer system in Portugal and was developed to improve the response time to occurrences, optimizing resources and also simplifying information. Considering the objectives set for the design of this platform, a database was created to include in the same system as much information as possible on each occurrence recorded.

From the operational point of view, it is a web system where information regarding occurrences is registered and stored in an Oracle database, integrated in one of the Lisbon City Hall servers. This system allows the simultaneous use of several users without compromising its good performance. The registration of information about occurrences is done by the elements of the LFBR operations center, and the registration includes a set of procedures that allow a better coordination between the occurrence response teams. Thus, each occurrence goes through five main stages and they are as follows: Occurrence Creation, Activation, Operational Closure and Administrative Closure, illustrated in the Figure 7.

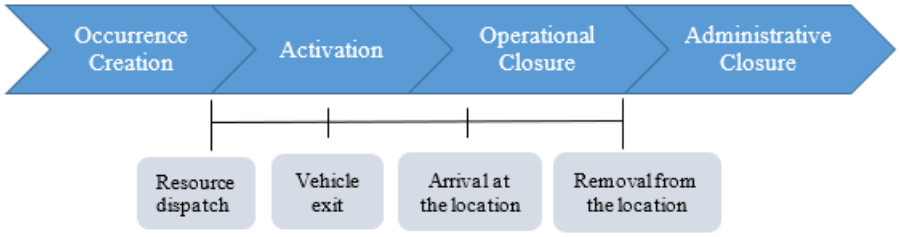


Figure 7 - States of an occurrence. Source: adapted from Martins [62]

The first step is the registration of the occurrence and between the activation and the operational closure there are four operational stages, namely the Resource dispatch, Vehicles exit, Arrival at the location and Removal from the location.

In this way, once the information is collected through the call, the next step is to dispatch the resources, which corresponds to the identification of the available and necessary vehicles to respond to the occurrence. Next, the temporal information is registered, namely the time the vehicle leaves the fire station, its arrival at the location of the occurrence and also its departure. This information allows the accounting of the response time. The operational closure relates to the departure of the elements of the corporation from the location of the occurrence and finally, there is the administrative closure that can only be completed when all data and reports are entered into the system and properly validated.

The use of this application is not exclusive to the LFBR, as it is also used by the Municipal Police (MP), the Municipal Civil Protection (MCP), the NACP and EMEL[4]. Besides the mentioned identities, the Mayor of Lisbon, heads of offices and councilmen can view the active occurrences in the city of Lisbon and the means used in each occurrence.

Therefore, the first descriptive and predictive analysis is going to be performed based on the data recorded by the LFBR occurrence management system and provided by the department of Territorial Intervention Unit (TIU) of the Lisbon City Hall. To complete the dataset provided, external data from two different sources is going to be used, namely INE and IPMA.

To enrich the firefighters' dataset to extract as much knowledge as possible, external data concerning the geographic characterization of the city of Lisbon is going to be added and these data come from the INE website. INE is a public institute that integrates the indirect administration of the state and has the administrative autonomy to produce and disseminate effectively and efficiently statistical information about the society. Statistical information is extracted respecting the techno-scientific methodologies and internationally established standards that ensure data quality. The statistical results must be made available simultaneously for the entire population [64]. In this context, the following data was extracted from INE website and added to the firefighters' dataset:

[4] Responsible entity for parking management and the improvement of mobility in Lisbon. Available in: https://www.emel.pt/pt/

average, age of buildings per parish, proportion of buildings in need of major repairs or very degraded per parish, and resident population per parish.

On the other hand, with the aim of correlating the occurrences recorded with the meteorological data, it was requested to IPMA meteorological data referring to a period from 2011 to 2018. IPMA holds the largest national network and meteorological observation infrastructures, and the largest meteorological data archive in the country [65]. The following data is going to be added to the firefighters' dataset: average air temperature, average wind intensity, humidity, and precipitation.

The external data extracted from the INE website and the data provided by the meteorological institute IPMA went through the process of cleansing, transformation, and integration to the firefighters' data dataset.

Finally, as previously mentioned, a second analysis is going to be performed on a second dataset corresponding to the data extracted from the application *Na Minha Rua Lx* (Figure 8).

This is a web portal and an application that simplifies the reporting of occurrences in the public spaces of the city of Lisbon, thereby encouraging citizens to have an active participation in the management of the municipality. This application can be installed on any smartphone and allows a quick report of situations that are grouped into the following categories: walkways and accessibility, urban Hygiene, public Lighting, roads and bike paths, trees and green spaces, municipal equipment (sports), municipal equipment (Culture), municipal equipment (education), public safety and noise, sanitation, municipal housing, and animals in urban environment.

It is important to understand how the application works in practice. The citizens make the registration of occurrences in the application and in the desktop application the registration is made by the operational services of the Lisbon City Hall. To register occurrences on the application, before accessing the application it is necessary to take a picture to record the situation. Afterwards, the registration is done on the application by creating a new occurrence where the first step is to upload the picture taken. After uploading the picture, the next step is to select the type of occurrence according to the options presented in the application and presented above. The next step is to wait for the location, which is extracted by the georeferencing of the uploaded picture. Before continuing, it is necessary to confirm if the occurrence has not yet been reported and if

not, the record will proceed to the next phase. In the next phase, is necessary to make a brief description of the occurrence, and the last phase corresponds to the submission of the occurrence. It is possible to follow the status of the occurrence in the main menu of the application.

Based on the data extracted from the application, exploratory analysis is going to be conducted to extract insights regarding the characterization of the occurrences reported in spatial and temporal terms, and the frequency of use of the application by citizens.

3.2 Data Understanding

In order to achieve the objectives outlined for this dissertation, it is essential to deepen the knowledge about the data that composes the firefighters' dataset and the dataset regarding the data from the application *Na Minha Rua Lx*. The process of familiarization with the data involves exploratory tasks that allow acquiring general knowledge, identifying flaws that affect the data quality, and identifying variables of possible interest for this analysis. In this sense, this step is important since it allows determining whether the available data contain the necessary information to achieve the proposed goals.

3.2.1 Firefighters' dataset

To work on the firefighters' dataset provided by the Lisbon City Hall, the dataset was first loaded. It is a CSV file, i.e., the values of each column are separated by commas. This file contains information regarding the occurrences registered by the firefighters and this information covers aspects such as the description of the occurrence, date of the occurrence, location of the occurrence, i.e., latitude, longitude and address, and the human and material resources (number of vehicles) allocated to each occurrence.

Initially, the number of rows and columns in the dataset was verified and it was concluded that the dataset contains data from 2011 to 2018 consisting of 135 200 records (rows in the CSV file) and 22 variables (columns in the CSV file).

Furthermore, it was applied the command that allows visualizing the first rows of the dataset with the purpose of having an overview of the variables that compose the dataset. The information regarding the type of data of each column was extracted, organized and presented in Table 4.

From the data summarization, it is possible to verify that the database has 13 columns with null values, and there are columns with high amounts of null values such as “OCO_FALSE_ALARM”, “OCO_RISCO”, “OCO_FRACCAO”, and “NUM_ELEMENTS” which have null values exceeding 100 000. The remaining columns with null values present less expressive values when compared to the columns mentioned previously. All columns are of type *object* and it is also possible to conclude that some of the data are categorical, i.e., they have a maximum limit of options as a possible value, such as the “OCO_STATE_DESIG” variable that can only assume the values "FO" for Operational Closure, "FA" for Administrative Closure and "A" for Acted, and the case of the “OCO_RISK” column that can only assume two possible values - 1 or 2.

Table 4 - Dataset description

Variable name	Variable description	First line data	Variable type	Null values
OCO_ID	Occurrence ID	1 160 721 082 011	Object	0
OCO_DT	Date of occurrence entry ("YYYY-MM-DD HH:MM:SS")	08/21/2011 18:02:00	Object	0
OCO_X	Longitude	-9,180194917	Object	1 600
OCO_Y	Latitude	38,75897253	Object	1 600
OCO_MORADA_ID	Cod_Via	34 780	Object	102
OCO_NUM_POL	Código_SIG	67100	Object	14 805
OCO_MORADA_DESIG	Identification of the occurrence location	Avenida de Ceuta	Object	0
OCO_FREGUESIA_ID	SIG code of the parish	110	Object	9 096
OCO_FREGUESIA_DESIG	Parish description	Campo de Ourique	Object	245
OCO_ESTADO_ID	Occurrence state ID	4	Object	0
OCO_DATA_ALTERACAO_ESTADO	Date of occurrence status change ("YYYY- MM-DD HH:MM:SS")	08/22/2011 17:40:00	Object	0
OCO_ESTADO_DESIG	Description of the occurrence status (FO - Operational Closure; FA - Administrative Closure; A - Acted)	FA	Object	0
OCO_FRACCAO	SIG code of the fraction	25348	Object	109 852
OCO_NAT_ID	ID Typology	11	Object	0
OCO_NAT_DESC	Typology	1300 - Incêndio - Inculto	Object	14 155
OCO_ENT_ID	ID of the entity that reported the occurrence	15	Object	4 708
OCO_ENT_DESC	Entity that reported the	CBV	Object	4 708
OCO_RISCO	Risk level indicator (1 - high; 2 - moderate)	2	Object	132 091
OCO_YN_SUSPENSO	Indicator of occurrence suspension before activation of	N	Object	0
OCO_FALSO_ALARME	False alarm identifier after media activation	1	Object	135 109
OCO_NUM_VIATURAS	Number of vehicles dispatched	2	Object	0
NUM_ELEMENTOS	Estimated number of people	6	Object	22 607

It is also important at this stage to check the presence of outliers. This analysis consists of a graphic tool, called Boxplot, that allows observing the data variation of a numerical variable through quartiles. From a boxplot it is possible to extract the data's asymmetry, dispersion, discrepant measures and identify the quartiles, where the first represents 25% of the data, the second represents 50% of the data (corresponding to the median), and the third is representative of 75% of the data [66].

Since the boxplots are made from the numeric variables and this dataset contains numeric variables, nevertheless, they are not relevant for the analysis since they are IDs. The numeric variables considered relevant for the analysis were selected to build the boxplots and they are the following: “OCO_RISCO”, “OCO_NUM_VIATURES”, and “NUM_ELEMENTS”. The respective boxplots are illustrated in Figure 8.

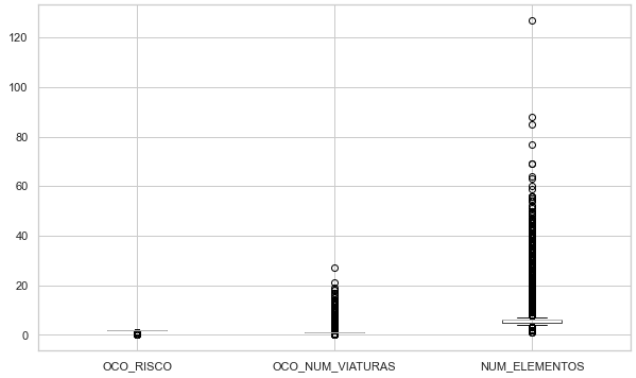


Figure 8- Outliers

From the direct analysis of the boxplot, it is possible to verify that there are no outliers that need to be treated. Lastly, the duplicate values were checked, and it was concluded that there is one duplicate value that is going to be treated in the next step.

3.2.2 *Na Minha Rua Lx* Dataset

Regarding the dataset containing the data related to the occurrences recorded on the *Na Minha Rua Lx* application, the same commands applied to the firefighters' dataset were applied to extract the first insights on the data. To work on this dataset that was provided in an *xlsx* format file (indicating that it is a Microsoft Excel spreadsheet), the first operation performed was its import.

A first analysis shows that the dataset is composed of 12 866 rows and eight columns. The summary of the information presented in Table 5, shows that the dataset contains

columns that aggregate information about the occurrences reported in the application such as the date on which the occurrence was reported, the type of occurrence, and the location of the occurrence.

Table 5 - Dataset Description

Variable names	Variable description	First line data	Variable type	Null values
numero	Occurrence ID	OCO/97587/2017	Object	0
data_criacao	Date of occurrence entry	31/08/2017 16:20	Object	0
area	Occurrence type	Segurança Pública e Ruído	Object	0
tipologia	Occurrence description	Edifício, muro, escarpa ou talude degradado	Object	0
freguesia	Parish	Carnide	Object	0
local	Address	Rua do Rio Zêzere	Object	0
latitude	Latitude	38.771585	Float64	0
longitude	Longitude	-9.185676	Float64	0

As shown in Table 5, there is no null value in the entire dataset. Furthermore, this summary shows that six of the eight columns that compose the dataset are of type *object*, except the columns latitude and longitude that are of type float. Regarding duplicate values, this dataset has one, which is going to be treated in the next phase.

3.3 Data Preparation

During the data description activities to extract the first insights, problems and inconsistencies that affect the data quality were identified. The most common problems are related to the existence of null values, inconsistency in the date formats, and the presence of duplicates values. The tasks in this phase include activities to solve the problems identified as well as activities to build the final dataset such as building new variables from existing variables and adding variables from external sources.

3.3.1 Firefighters' dataset

Regarding the firefighters' dataset, all problems and inconsistencies are summarized and presented in Table 6.

Table 6 - Summary of data quality

Attribute name	Format	Null values	Data quality
OCO_ID	Object	0	No problems detected
OCO_DT	Object	0	Date and time in the same column
OCO_X	Object	1 600	Presence of missing values
OCO_Y	Object	1 600	Presence of missing values
OCO_MORADA_ID	Object	102	Presence of missing values
OCO_NUM_POL	Object	14 805	Presence of missing values
OCO_MORADA_DESIG	Object	0	No problems detected
OCO_FREGUESIA_ID	Object	9 096	Presence of missing values
OCO_FREGUESIA_DESIG	Object	245	Presence of missing values
OCO_ESTADO_ID	Object	0	No problems detected
OCO_DATA_ALTERACAO_ESTADO	Object	0	No problems detected
OCO_ESTADO_DESIG	Object	0	No problems detected
OCO_FRACCAO	Object	109 852	Presence of missing values
OCO_NAT_ID	Object	0	No problems detected
OCO_NAT_DESC	Object	14 155	Presence of missing values
OCO_ENT_ID	Object	4 708	Presence of missing values
OCO_ENT_DESC	Object	4 708	Presence of missing values
OCO_RISCO	Object	132 091	Presence of missing values
OCO_YN_SUSPENSO	Object	0	No problems detected
OCO_FALSO_ALARME	Object	135 109	Presence of missing values
OCO_NUM_VIATURAS	Object	0	No problems detected
NUM_ELEMENTOS	Object	22 607	Presence of missing values

Considering the problems identified, the process of dataset preparation begins with the selection of the relevant variables for the analysis. This dataset has a large number of variables however, some are irrelevant since they do not add important information to the analysis that is intended to be developed. In this way, the selection of variables becomes essential since this technique allows reducing the size of the dataset in terms of variables, selecting only those considered important to achieve the proposed goals. The results of the variable selection process are presented in Table 7.

Table 7 - Result of the variable selection process

Attribute name	Included	Reason for inclusion/exclusion
OCO_ID	No	Identifies the occurrence, not adding relevant information according to the defined objectives
OCO_DT	Yes	Identifies the location of an occurrence in temporal terms
OCO_X	Yes	Identifies the occurrence location
OCO_Y	Yes	Identifies the occurrence location
OCO_MORADA_ID	No	Does not add relevant information to the analysis as it is an identifier code
OCO_NUM_POL	No	Does not add relevant information to the analysis as it is an identifier code
OCO_MORADA_DESIG	No	Does not add information to the analysis since the analysis will not be conducted at street level
OCO_FREGUESIA_ID	No	Does not add relevant information to the analysis as it is an identifier code
OCO_FREGUESIA_DESIG	Yes	Identifies the parish where an occurrence took place
OCO_ESTADO_ID	No	Does not add relevant information to the analysis as it is an identifier code
OCO_DATA_ALTERACAO_ESTADO	No	Does not add relevant information to the analysis
OCO_ESTADO_DESIG	No	Does not add relevant information to the analysis
OCO_FRACCAO	No	Does not add relevant information to the analysis as it is an identifier code
OCO_NAT_ID	No	Does not add relevant information to the analysis as it is an identifier code
OCO_NAT_DESC	Yes	An important attribute to identify the type of occurrence
OCO_ENT_ID	No	Does not add relevant information to the analysis as it is an identifier code
OCO_ENT_DESC	No	Does not add relevant information to the analysis
OCO_RISCO	No	Does not add relevant information to the analysis
OCO_YN_SUSPENSO	No	Does not add relevant information to the analysis
OCO_FALSO_ALARME	No	Does not add relevant information to the analysis
OCO_NUM_VIATURAS	Yes	Identifies the material resources allocated to each occurrence
NUM_ELEMENTOS	Yes	Identifies the human resources allocated to each occurrence

With the variable selection process concluded, the dataset is now reduced to seven variables and they are as follows: “OCO_DT”, “OCO_X”, “OCO_Y”,

"OCO_FREGUESIA_DESIG", "OCO_NAT_DESC", "OCO_NUM_VIATURES", and "NUM_ELEMENTS". With the identification of the relevant attributes, it becomes necessary to apply data cleansing techniques to improve data quality.

Table 8 presents the variables and their data cleansing processes to which they were subjected.

Table 8 - Description of data cleansing techniques applied

Attribute name	First line data	Format	Null values	Null values %	Cleansing treatment applied
OCO_DT	08/21/2011 18:02:00	Object	0	0%	Split date-time column into separate date and time columns
OCO_X	-9.180194917	Object	1 600	1.1%	Delete null values
OCO_Y	38.7589725275051	Object	1 600	1.1%	Delete null values
OCO_FREGUESIA_DESIG	Campo de Ourique	Object	245	0.1%	Delete null values
OCO_NAT_DESC	1300 - Incêndio - Inculto	Object	14 155	10.4%	Delete null values
OCO_NUM_VIATURAS	2	Object	0	0%	Convert to int32
NUM_ELEMENTOS	6	Object	22 607	16.7%	Delete null values and convert to int32

The cleansing techniques include variable format conversion such as the "OCO_NUM_VIATURES" and "NUM_ELEMENTS" variables that were converted to integers, column separation, namely, the "OCO_DT" column that refers to the temporal and spatial location of the occurrences was separated into two distinct columns, one with data related to the date and the other with data regarding the time of the registration of each occurrence. Finally, the cleansing treatments also included the elimination of null values since one possible way to deal with null values would be to replace them with the average value of each column, however, these variables such as geographic coordinates, parish, and description of the occurrences do not allow such a replacement. The only variable where it would be possible to apply this technique would be the variable "NUM_ELEMENTS", but it has a significant number of null values, so replacing the null values with the average could bias the results. In this way, all null values were eliminated.

After completing the dataset cleaning tasks, the data transformation process started, which was focused on activities related to creating new variables from variables already in the dataset and adding variables from external sources. This phase is essential as it allows transforming and adapting the information present in the dataset to the objectives that were previously defined for this dissertation. Before proceeding to the creation of

new variables, and in order to adapt the data to this analysis, the variables were renamed, and the result is presented in Table 9.

Table 9- Variables name transformation

Attribute name	New name
OCO_DT	Date-Time
OCO_X	Longitude
OCO_Y	Latitude
OCO_FREGUESIA_DESIG	Parish
OCO_NAT_DESC	Occurrence Description
OCO_NUM_VIATURAS	Num of vehicles
NUM_ELEMENTOS	Num of people

In terms of the construction of new variables, it was taken into account information that could allow a temporal analysis of the occurrences and, in this sense, three new variables were created from the "Date-Time" and they are the following: "Year", "Month", and "Hour".

The variable "Hour" underwent a discretization process since the objective of the analysis is to analyze the occurrences throughout the periods of the day and with this in mind, this variable was discretized and from this discretization, a new variable ("Period of the day") was created containing the four periods of the day: dawn, morning, afternoon, and evening.

As one of the main goals of this dissertation is to perform a statistical and descriptive analysis of the occurrences that took place in the buildings of Lisbon city, it became necessary to add external information that complements the dataset and therefore allowing a detailed analysis. This external data contains information that characterizes the city of Lisbon in terms of population, building characteristics, and meteorological conditions. The information about the buildings and population of the city of Lisbon was extracted from the INE site which contains information about statistical data of Portugal. On the other hand, the meteorological information was made available by the meteorological institute IPMA.

The first variable created from external data was "Avg Building Age". For this purpose, a dataset that contains the average age of buildings per parish was loaded. This information allows identifying the parishes where the buildings are older. It is a dataset that has neither null nor duplicate data with values that vary between 27 and 107 years.

To better understand the state of conservation of the buildings according to parishes, a variable called “Prop of degraded buildings (%)” was created, and this variable comes from a dataset that contains the proportion of buildings in need of major repairs or very degraded per parish, with values varying between 0.67% (*São Francisco Xavier*) and 67.36% (*Santa Justa*).

The number of resident inhabitants was also considered, so a dataset without null and duplicate values was loaded containing the number of inhabitants per parish. From this dataset resulted a new variable named “Resident Pop”. Lastly, to correlate the occurrences with the meteorological data, a dataset containing meteorological data per day from the period between 2011 and 2018 was loaded. This dataset has 5 variables, namely:

- “Avg air temp” - Average air temperature.
- “RH” - Relative Humidity.
- “Avg WS” - Average wind speed.
- “PRCP” - Precipitation.

All variables originating from external sources were properly integrated with the firefighters' dataset and all the transformations and constructions performed resulted in a final dataset composed of 17 columns that are going to be used in the remaining phases of the methodology. Table 10 shows the constructions and transformations performed on the firefighters' dataset.

Table 10 - Final dataset

Attribute name	Format	Resulted from a transformation	Resulted from external data
Longitude	Object		
Latitude	Object		
Parish	Object		
Types of occurrences	Object		
Num of vehicles	Int32		
Num of people	Int32		
Year	Int32	x	
Month	Int32	x	
Period of the day	Object	x	
Avg Building Age	Object		x
Prop of degraded buildings (%)	Float64		x
Avg Num of floors	Object		x
Resident Pop	Int64		x
Avg air temp	Float64		x
RH	Int64		x
Avg WS	Float64		x
PRCP	Object		x

After aggregating all the above-mentioned external variables to the firefighters' dataset, it was necessary to categorize the types of occurrences that took place in buildings, since they are a significant amount and a categorization helps facilitate the visual analysis. The information regarding the types of occurrences is found in the "Occurrence Description" variable and this variable has 25 types of occurrences involving buildings. These types were defined by the firefighters' occurrence management system and are as follows:

- Fire - Building (Infrastructures/Installation) – School.
- Fire - Building (Infrastructures/Installation) - Empty/Degraded Building.
- Fire - Building (Infrastructures/Installation) - Commercial/Shops/Fairs/Transport Station.
- Fire - Building (Infrastructures/Installation) – Parking.
- Fire - Building (Infrastructures/Installation) – Culture/Museum/Art/Library.
- Fire - Building (Infrastructures/Installation) – Military/Security forces.
- Fire - Building (Infrastructures/Installation) – High-rise building (>29 m).
- Fire - Building (Infrastructures/Installation) – Hotel and similar
- Fire - Building (Infrastructures/Installation) - Performance/Recreation Religious Worship.
- Fire - Building (Infrastructures/Installation) - Services.
- Fire - Building (Infrastructures/Installation) - Hospital/Home.
- Fire - Building (Infrastructures/Installation) - Industry/Workshop/Warehouse.
- Fire - Building (Infrastructures/Installation) - Housing.
- Accidents - Equipment - Lifts.
- Accidents - Equipment.
- Infrastructures and Communication Routes – Collapse.
- Infrastructures and Communication Routes - Collapse (Coating fall).
- Infrastructures and Communication Routes – Landslide.
- Infrastructures and Communication Routes – Floods.
- Infrastructures and Communication Routes - Private space flood.
- Industrial–technological - Suspicious Situations - Check Smoke.
- Industrial–technological - Suspicious Situations - Check Smells.
- Industrial–technological - Gas Leak - Plumbing/Conduct.

- Industrial–technological - Gas Leak - Bottle.
- Industrial–technological - Gas Leak - Deposit/Reservoir.

These 25 types of occurrence were grouped into the following seven categories:

- Infrastructures – Collapse.
- Infrastructures – Floods.
- Infrastructures – Landslide.
- Fire.
- Accidents (with equipment or with elevators)
- Ind. technol - Gas leak
- Ind. technol - Suspicious situations (check smoke or check smells).

After categorizing the occurrences, it was made the counting of each type of category created in order to verify its representation in the dataset, it was found that the category Infrastructures - Landslide is underrepresented, with only 17 registrations in a total of 15 766 registrations. For this reason, this category was eliminated.

In terms of the distribution of occurrences over the years, findings showed that the in the first two years, between 2011 and 2012 few occurrences were recorded when compared to the other years, and this is explained by the fact that these years contain only seven months of the year. The distribution of occurrences by month during these two years can be consulted in the table in Annex A - Distribution of occurrences by month in 2011 and 2012.

Since these first two years contain limited data, these data are not representative in the dataset once the information available for these years does not allow a leveled analysis when these data are analyzed in relation to the following years that have more information. In this way, only the years between 2013 and 2018 were considered for this analysis.

3.3.2 *Na Minha Rua LX* Dataset

Regarding the dataset corresponding to the data reported in the *Na Minha Rua Lx* application, the data preparation process began, as with the firefighters' dataset, by selecting the relevant attributes. The dataset, which initially consisted of eight variables, is now reduced to six after the exclusion of the variables considered irrelevant for the

analysis since they do not add relevant information. Table 11 presents the summary of the variable selection process that was performed.

Table 11 - Result of the variable selection process

Attribute name	Included	Reason for inclusion/exclusion
numero	No	Identifies the occurrence, not adding relevant information according to the defined objectives
data_criacao	Yes	Attribute useful to identify the location of an occurrence in temporal terms
area	Yes	An important attribute to identify the category of occurrence reported
tipologia	Yes	An important attribute to identify the type of occurrence
freguesia	Yes	An important attribute that allows identifying the parish where an occurrence took place
local	No	Does not add information to the analysis since the analysis will not be conducted at street level
latitude	Yes	An important attribute to identify the occurrence location
longitude	Yes	An important attribute to identify the occurrence location

After identifying the necessary variables to conduct this analysis, the next step consisted of applying data processing techniques to adequate the dataset to the analysis intended to be developed. However, no significant problems were identified, only one variable ("data-criacao") that is not in date format. In addition to the date format problem, this dataset has a duplicate value that needs to be handled. The description of the data cleansing treatment applied is presented in Table 12.

Table 12 - Description of data cleansing techniques applied

Attribute name	Format	Null values	Detected problem	Cleansing treatment applied
data_criacao	Object	0	Wrong format	Convert to format D/MM/YYY
area	Object	0	No problems detected	-
tipologia	Object	0	No problems detected	-
freguesia	Object	0	No problems detected	-
latitude	Float64	0	No problems detected	-
longitude	Float64	0	No problems detected	-
-	-	-	-	Duplicate Removal

With the data cleansing process complete, the next step focuses on the data transformation. This step includes tasks such as renaming variables where the result of

this transformation is present in Table 13, and building new attributes from existing attributes in the dataset.

Table 13 – Variables name transformation

Attribute name	New name
data_criacao	Date-Time
area	Occurence categoory
tipologia	Ocurrence type
freguesia	Parish
latitude	Latitude
longitude	Longitude

In terms of constructing new variables to enable a detailed analysis, information that would allow the temporal location of occurrences was considered and, three new variables were created from the Date-Time variable: “Year”, “Month”, and “Hour”.

From the variable “Hour”, a new variable (“Period of the day”) was created for the purpose of analyzing the frequency of use of the application according to the period of the day. The discretization of the “Hour” variable resulted in the creation of a new variable that contains the four periods of the day: dawn, morning, afternoon, and evening.

Still in terms of attribute transformation, the variable “Incident type” which is an important variable since it describes the type of occurrences reported underwent a transformation process concerning the translation of the type of occurrences reported in order to facilitate the analyzing process. Initially, this variable had 5 types of reported occurrences and the translation of these categories can be seen in Table 14.

Table 14 - Result of the translation of the "Occurrence type" variable

Original occurrence type	New occurrence type
Despejos/desocupações.	Evictions/Displacements
difício, muro, escarpa ou talude degradado.	Degraded building, wall, scarp or slope
Fiscalização de insalubridade em propriedades/terrenos/via pública.	Inspection of insalubrity in properties/lands/public roads
Obras ilegais - Edificado, via pública e ruído	Illegal constructions - Building, public road and noise
upação ilegal de edificado	Illegal occupation of buildings

With all the transformations on the dataset completed, the final dataset has 8 columns and 12 865 rows, presented in Table 15.

Table 15 - Final dataset

Attribute name	Format	Resulted from a transformation
Occurence categoory	Object	
Ocurrence type	Object	x
Parish	Object	
Latitude	Float64	
Longitude	Float65	
Year	Int64	x
Month	Int64	x
Period of the day	Object	x

After the data preparation is complete, the distribution of occurrences over the years was analyzed and the results showed that for the year 2020 there are few records, and these records refer to the first six months of the year. Given that there are no data for the remaining months of the year, this issue was addressed with those responsible for the by the department of TIU of the Lisbon City Hall in order to verify if it is possible to provide data for the remaining months of the year. However, it was not possible to gather data for the remaining months since it was concluded that the 2020 data may be biased due to the pandemic situation that started in that year, and in this way the 2020 data may not reflect the normal use of the application as it happens in the other years where no constraints of this nature were experienced. For this reason, it was decided to exclude the data for the year 2020 and the analysis was focused on the period from 2017 to 2019.

In summary, this chapter allowed to understand that the city of Lisbon, given its geological and socio-urbanistic characteristics, is one of the most elderly cities in the LMA, being vulnerable to several events such as diverse meteorological conditions, floods, earthquakes, tsunamis, mass movements, traffic accidents, accidents involving the transportation or storage of hazardous materials, collapses of structures, urban fires, and forest fires.

In order to develop a system that supports the decision-making process, two main data sources were identified, namely LFBR occurrence management system and the application *Na Minha Rua LX*. The datasets used in this research contain information about the occurrences registered in the city of Lisbon and, in order to complete this analysis, data from external sources such as IPMA and INE were added to the firefighters' dataset and after identifying the data sources the tasks were directed to the cleansing and preparation process that included feature selection, format conversion, creating new variables from existing variables and from data that has been aggregated.

Chapter 4 – Presentation of results and evaluation

This chapter is dedicated to the last two phases of the adapted CRISP-DM methodology and is divided into two sections. The first section (modeling) includes tasks related to data visualization and application of predictive models. Data visualization is an important step since it allows the data to be visually represented, making it possible to find patterns and therefore extract information to support the decision-making process.

In addition to the visual component, i.e., data visualization, this phase also includes the application of predictive models to predict disasters. For the model application process, the first step refers to selecting predictive models according to the project's objective, and in this case, classification algorithms are going to be used to predict the occurrences. After choosing the models, the test design is generated where the strategies to test the quality of the models are defined. Additionally, the dataset is split into test and training datasets where the models are going to be built using the training dataset and the quality is going to be evaluated using the test dataset.

Afterward, the models are going to be built based on the parameters defined for each one, and their evaluation is going to be conducted by listing the qualities of the generated models. These qualities, such as accuracy, are going to be analyzed comparatively among the generated models considering the defined quality criteria.

The second section is dedicated to the evaluation. The main objective of this section is to evaluate the results of this research with the stakeholders of the Lisbon City Hall in order to verify to what extent the results obtained are in accordance with the defined objectives.

4.1 Modeling

With the data preparation phase complete in the previous chapter, the modeling phase begins. This phase is focused on extracting knowledge that can help decision-makers to manage the city in an efficient way when it comes to disaster situations. This phase includes data mining tasks to extract patterns from the data, and tasks related to building predictive models to predict the occurrences. In this regard, the success of this phase is intrinsically related to the quality of the data prepared in the previous phase, the DM techniques used to graphically visualize the data, and the models selected to predict the occurrences.

Before building the predictive models, a data visualization was conducted on both datasets with the primary goal of verifying the relationship between the different variables and deepen the knowledge about the occurrences and the circumstances in which they affect the city.

For the firefighters' dataset, the visual component of this chapter focuses on a spatial-temporal analysis since it is intended to analyze the occurrences from a spatial point of view, i.e., the places with a higher number of occurrences, and correlate the occurrences with the characteristics of the city's buildings. On the other hand, on a temporal level, it is intended to analyze the impact of the periods of the year (months) on occurrences by correlating the occurrences with the meteorological conditions recorded in that period.

Regarding the dataset that contains the data extracted from the application *Na Minha Rua Lx*, the spatial-temporal analysis, similarly to the firefighters' dataset, is also focused on two aspects - spatial and temporal perspectives. The analysis in spatial terms intends to verify the parishes that concentrate a higher number of occurrences as well as an analysis of the types of occurrences reported. At the temporal level, the analysis is focused on extracting insights regarding the period where occurrences are recorded.

Once the visualization phase is over, the next step refers to the prediction process to predict the occurrences registered in the LFBR occurrence management system.

4.1.1 Data Visualization – Firefighters' dataset

The first visual analysis is focused on understanding the distribution of the data to get an overview of the occurrences registered and for this purpose, a histogram was generated to visualize the distribution of occurrences over the period 2013 to 2018 and the result is presented in the Figure 9.

As shown in Figure 9, between the years 2013 and 2014, there was an increase in the number of occurrences registered, reaching 17 607 occurrences in 2014. After 2014, the number of occurrences decreased until 2016, when 15 089 occurrences were recorded. In 2017, this trend was reversed, with 16 582 occurrences recorded. Between the years 2017 and 2018, there was a new decrease in the number of occurrences, with 13 368 registrations in 2018. In general, it is possible to conclude that between the period of 2013 and 2018 there was a downward trend in the number of occurrences recorded in the

LFBR occurrence management system, but this decrease was not linear as there were oscillations over the years.

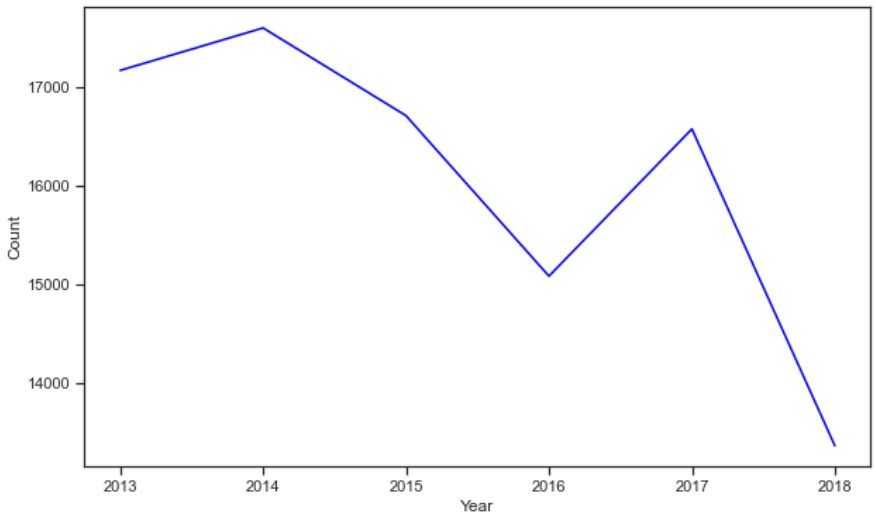


Figure 9- Occurrences distribution over the years

Firefighters respond to many different types of occurrences comprising several areas of action. For a better understanding of the activities performed by firefighters, Table 16 presents the types of occurrences (previously categorized) that firefighters respond to in their daily work.

Table 16 - Distribution of occurrences by category

Occurrence type	Frequency	%
Fire	8 460	8.5%
Accidents	10 058	10.1%
Infrastructures and Communication Routes	14 671	14.7%
Pre-hospital	9 166	9.2%
Legal conflicts	541	0.5%
Industrial Technological	5 170	5.1%
Services	45 442	45.7%
Activities	5 912	5.9%
Civil Protection Events - Technical Visit	4	0.004%

As shown in Table 16, the distribution is not balanced among the nine categories of occurrences recorded in the dataset, since there is an over-position of one category, namely the Services category, which represents 45.7% of occurrences recorded in the dataset. This category includes services such as road cleaning services, opening and

closing doors, hospital transport, water supply, and prevention services at shows, sports, and patrolling.

The occurrences related to Infrastructures and communication routes that include occurrences such as collapses, floods, falling trees and structures, and falling electric cables represent 14.7% of the occurrences recorded in the dataset. While the occurrences related to Accidents that include railroad accidents, road accidents, and accidents with equipment (elevators, escalators) present a proportion of 10.1%.

The categories that have the smallest representation in the dataset are Activities with 5.9%, Industrial-technological with a proportion of 5.1%, Legal conflicts with 0.5%, and Civil Potection events that represent 0.004% of the occurrences registered. The detailed distribution of each type of occurrence is presented in the table in Annex B - Occurrence Distribution.

Finally, still in this line of general analysis of the occurrences registered in the LFBR management system, the frequency of occurrences throughout the four periods of the day (dawn, morning, afternoon, and evening) was analyzed with the main purpose of verifying what time of the day the most occurrences are registered. This information is useful for managing and allocating human resources throughout the period of the day. The result is presented in figure 10.

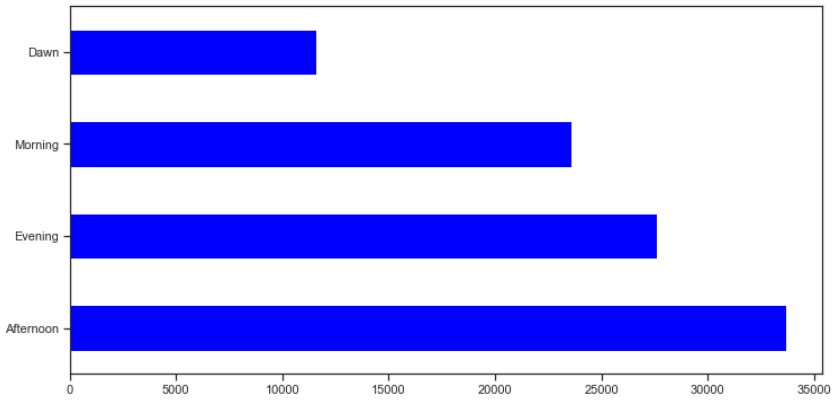


Figure 10 - Distribution of occurrences by period of the day

The results showed that the afternoon is the period of the day with the highest number of occurrences, as opposed to the dawns where a lower number of occurrences are registered.

After a general analysis of the type of occurrences, the study is focused on one of the main objectives of this dissertation - analysis of the occurrences that took place in the

buildings of the city of Lisbon to characterize them spatially and temporally. With this in mind, all occurrences that did not affect the buildings were excluded.

The first analysis consisted of verifying how the variables are correlated with the aim of obtaining a better understanding of the factors that influence the incidence of these occurrences. In this way, a correlation matrix between the different variables was created to identify which variables influence the occurrences. The correlation matrix is presented in Figure 11, and it was considered the variables that have a correlation with the types of occurrence with values greater than -1 and greater than 1.

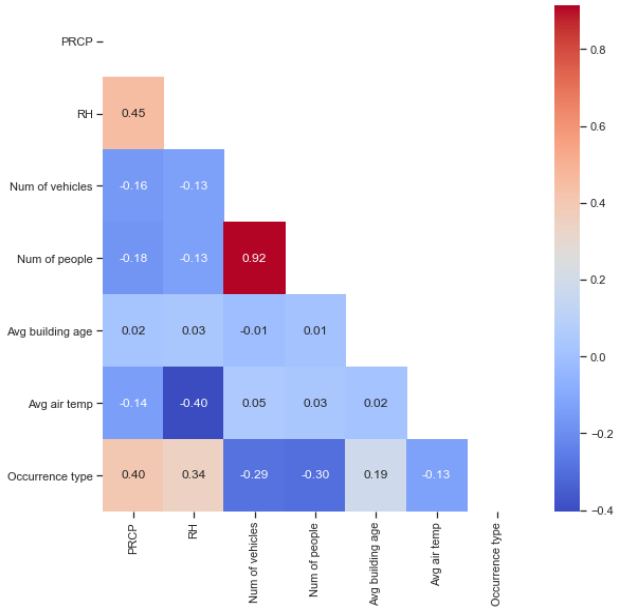


Figure 11 - Correlation matrix

The variables that show a higher correlation with the types of occurrences recorded are the following: precipitation (PRCP), relative humidity (RH), number of people allocated to occurrences (Num of people), number of vehicles involved in the response to occurrences (Num of vehicles), average age of buildings (Avg Building Age), average air temperature (Avg air temp) and resident population per parish (Resident Pop).

To begin the analysis of the events in buildings, a bar chart was created to extract the first insights into the occurrences that affect the buildings of the city. In this way, the categorization of the types of occurrences involving buildings is visually represented in Figure 12, and the figure shows that collapse with 3 742 records and floods with 3 356 records are the types of occurrences that most affect the buildings in the city of Lisbon, followed by occurrences related to suspicious situations that include verification of smells or smoke that count with 3 105 records. Also, with a significant proportion of incidences,

but less expressive when compared with the previously mentioned categories, are accidents involving equipment or elevators with a total of 2 399 records, fires with 1 892 records, and gas leaks with 1 259 records.

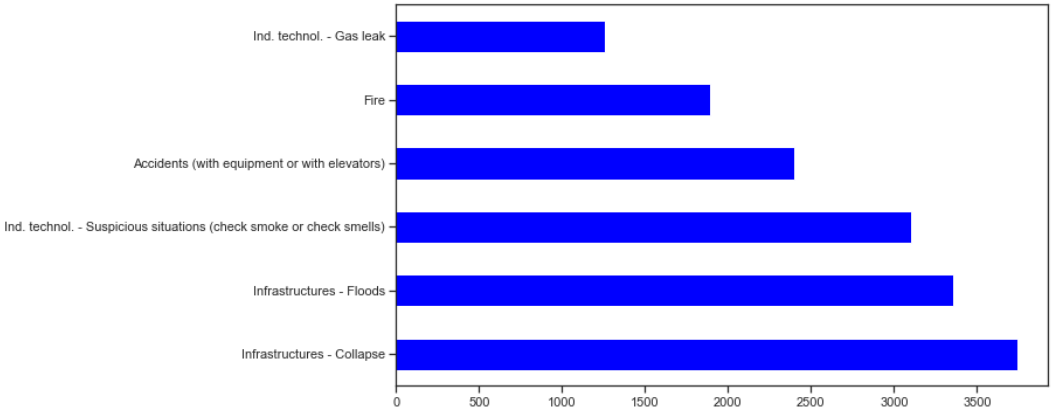


Figure 12 - Distribution of occurrences according to categories

When these occurrences are analyzed over time, i.e., their distribution by year (Figure 13), it is verified that there are occurrences that over the years occur in greater proportion, such as collapses, suspicious situations (checking smoke or smells), and accidents with equipment and elevators. The occurrences related to floods had a higher incidence in 2013 and 2014, with a decrease in the following years.

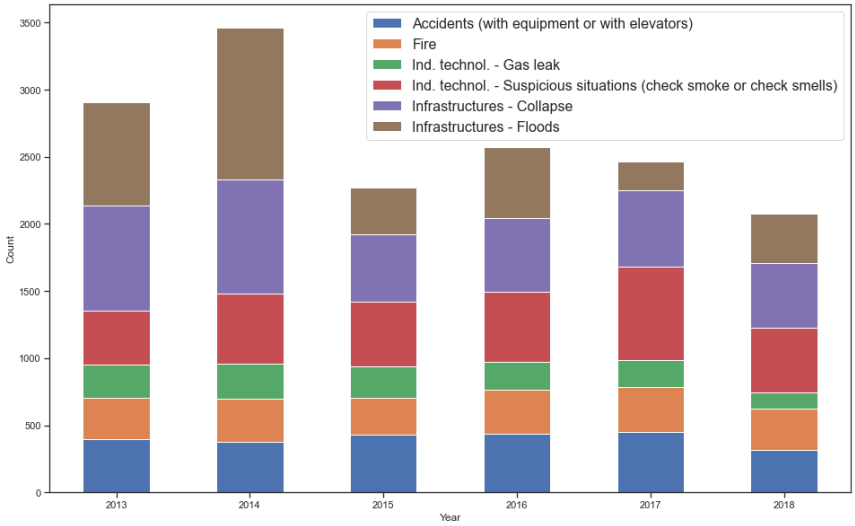


Figure 13 - Distribution of occurrences per year

Focusing the analysis on each occurrence to extract insights about its pattern of occurrence over the 12 months of the year where 1 corresponds to the month of January and 12 corresponds to the month of December, it is possible to verify that in the case of the occurrences referring to the infrastructure categories, i.e., collapses and floods represented in Figure 14, that in the case of collapses (A), these type of event occurred

more frequently in the autumn and winter months, reaching maximum values (over 400 records) in the months of October and January. As the spring and summer months approach, the number of records of this type of occurrence decreases, reaching lower values in the summer peak.

Regarding floods events (B), there is a higher incidence in the winter and autumn months with the highest values being recorded between the months of October to December and in the months of January and March. On the other hand, in the summer months these values are much lower when compared to the winter months.

Cases of suspicious situations (checking for smoke or smells) (C) occur, similar to the types described above, more frequently in the autumn and winter months. On the other hand, the occurrences related to gas leaks (D) show an oscillation during the months of the year, except for the month of January where values are higher than the other months, exceeding the 140 registered in this month.

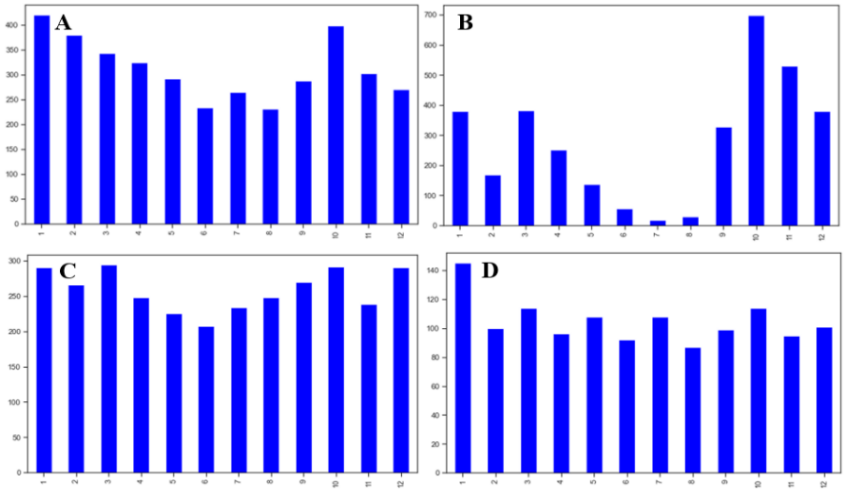


*Figure 14 - Temporal distribution of the occurrences. The bar chart from figure **A** shows the temporal distribution of Collapses, the bar chart from figure **B** shows the temporal distribution of Floods, the bar chart from figure **C** shows the temporal distribution of Suspicious situations (check smoke or check smells), and the bar chart from figure **D** shows the temporal distribution of Gas leaks.*

Lastly, the distribution of accidents involving equipment or elevators and fires is shown in Figure 15. Regarding accidents with equipment or elevators (A), this type of occurrence presents an incidence with similar values throughout the months except for the month of July where there is an increase and the month of November where there is a decrease.

In the case of fires (B), these events occur with a similar proportion throughout the months of the year, however, the month of December is highlighted since there is an

increase in the recording of this type of occurrence. These observations may be due to the fireplaces and candles that are used in greater density at this time of year.

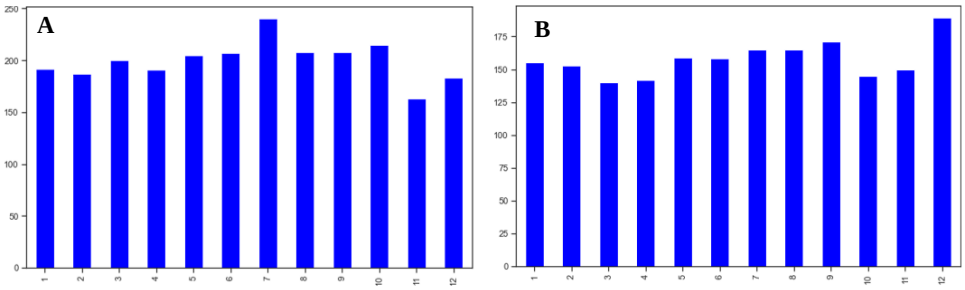


*Figure 15 - Temporal distribution of the occurrences. The bar chart from figure **A** shows the temporal distribution of accidents with equipment or elevators and the bar chart from figure **B** shows the temporal distribution of Fires*

Since a first analysis showed that there are types of occurrences that have a higher incidence in certain times of the year, such as collapses, floods, suspicious situations (checking for smoke or smells), occurring with a higher incidence in the winter/autumn months, the influence of weather conditions on the incidence of different types of events affecting the city of Lisbon has been verified. From the correlation matrix analysis (Figure 11), it was verified that the precipitation (PRCP) and relative humidity (RH) variables are highly correlated with the types of occurrences.

With this in mind, the next step consisted of analyzing the influence of precipitation on the different types of occurrences through segmented analysis of the data where the data was analyzed through four distinct periods, namely when it does not rain, when the rain is low, when the rain is moderate, and when the rain is heavy. The creation of these four levels allows the precipitation to be classified in qualitative terms.

For this purpose, an interquartile approach was adopted from the elaboration of a boxplot (Figure 16) to define the intervals of precipitation levels.

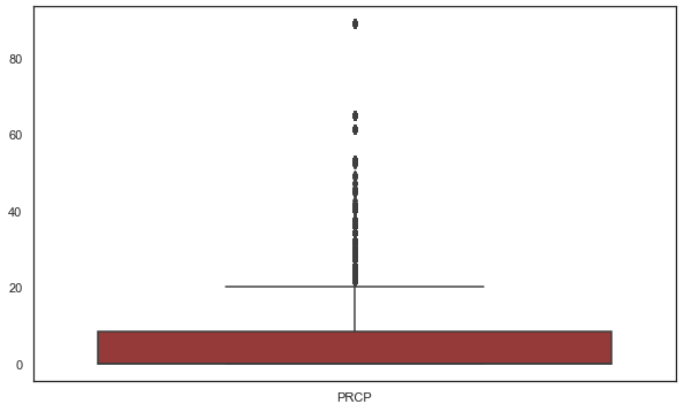


Figure 16 - Boxplot (PRCP variable)

The boxplot elaborated allows verifying the precipitation data distribution in the dataset and, as shown in the figure, the boxplot only presents data from the second quartile that corresponds to the median, therefore equivalent to 50% of the data. The fact that there is data from the second quartile signifies that it did not rain in 50% of the data. From the second quartile on, the first precipitation values are observed.

In order to complement the information extracted from the boxplot and thereby have a better understanding of the data's interquartile distribution, the basic statistics of the rainfall data were analyzed through the *describe()* command, and the results are presented in Table 17.

Table 17 - Statistical results from PRCP variable

Statistical results	PRCP
Count	16 252
Mean	7
std	13.9
Min	0
25%	0
50%	0
75%	8.6
Max	89.3

Considering the data for the minimum values, the three quartile values, and the maximum value, the following four datasets were built from the original dataset with the main objective of analyzing the distribution of occurrences when four rainfall intensities are recorded: no rain, low rain, moderate rain, and heavy rain. The datasets were created considering the quartile values, the minimum and maximum values registered. The following datasets were created:

- no_rain dataset - When the precipitation is zero. In this dataset, all precipitation data that are less than or equal to the median were considered.
- low_rain dataset - The data in this dataset comprise values that are higher than the median and lower than the third quartile values. This range is due to the fact that the third quartile is where the first precipitation values are observed.
- moderate_rain dataset – This dataset contains all precipitation data that are greater than or equal to the third quartile and less than the maximum interquartile.

- heavy_rain dataset – This dataset contains the data considered as outliers and for this reason, it was selected the values that are greater than or equal to the interquartile maximum up to the absolute maximum value of precipitation recorded in the dataset.

This approach based on quartile values is only applied in the context of this dissertation as it is not applicable to data from other countries since precipitation values are different. For instance, the values considered as moderate rainfall in Portugal are different from the values considered as moderate in a country characterized by high levels of precipitation as is the case of England [67].

With the datasets created, the *describe()* command was again applied to extract the statistical information from each dataset to verify that the data was correctly selected according to the different ranges. The statistical results are shown in Table 18.

Table 18 - Statistical results from the four new datasets created

Statistical results	no_rain dataset	low_rain dataset	moderate_rain dataset	heavy_rain dataset
Count	8 561	3 608	2 086	1 997
Mean	0	2.6	13.1	39
std	0	2.3	3	16
Min	0	0.1	8.6	21.7
25%	0	0.6	10.6	28.5
50%	0	2	12.9	32.3
75%	0	4.4	15.5	41.8
Max	0	8.5	20.4	89.3

From the statistical data on each dataset, it is possible to verify that the data are correctly selected since the dataset where it did not rain (precipitation 0) presents all values equal to zero as expected. Regarding the dataset with moderate rainfall, the values are also properly selected since its minimum value is 0.1, being this the value immediately greater than the median, and the maximum value 8.5 (value below the median). The same is verified in the moderate rainfall dataset since its minimum value (8.69) corresponds to the median and its maximum value (89.3) corresponds to the interquartile maximum. Lastly, the heavy rain dataset has the values selected correctly since its minimum and maximum values comprise values between the interquartile maximum and the absolute maximum.

Regarding the data distribution, it is possible to verify that the distribution is not balanced since the datasets have different sizes; the dataset with the highest amount of

data is the dataset where the precipitation is zero and the dataset with the lowest number of records is the dataset corresponding to heavy rain. Since the data are not balanced, the visual analysis of the distribution of occurrences according to precipitation levels has been performed in relative terms, that is, through their percentage in the dataset. The graphs containing the relative distribution of occurrences according to precipitation levels are presented in the following Figures.

Figure 17 shows that with precipitation values equal to zero, the occurrences recorded in greater proportion are suspicious situations (checking smoke or smells), collapses, and accidents (with equipment and elevators), with 25.7%, 23.7%, and 20%, respectively. With this level of precipitation, fires represent 15.9%, gas leakage with 10.5%, and finally floods with 4% indicating the low incidence of this last type of occurrence when precipitation values are low.

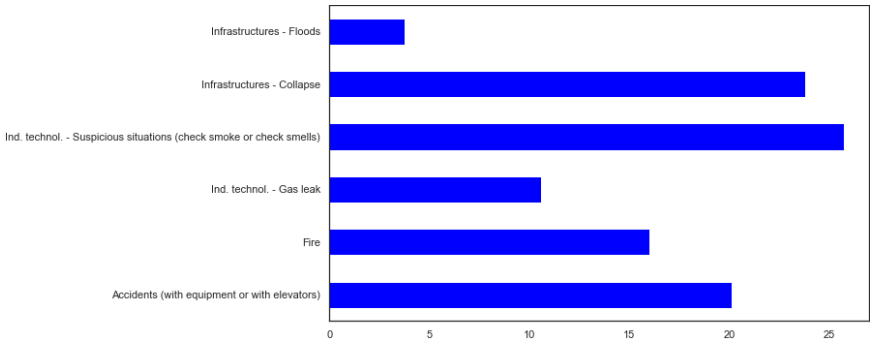


Figure 17 - Distribution of occurrences when precipitation is zero

Figure 18 shows that with the recording of the first precipitation values, although low precipitation, there is a change in the pattern of incidence of occurrences, with collapses appearing in greater proportion (28.3%) followed by floods with 19,23%, and suspicious situations 18,87%. Accidents involving equipment and elevators have a proportion of 14.6%, and fires 10.97%. Finally, for this level of precipitation, gas leaks are recorded with a residual proportion of 0.11%.

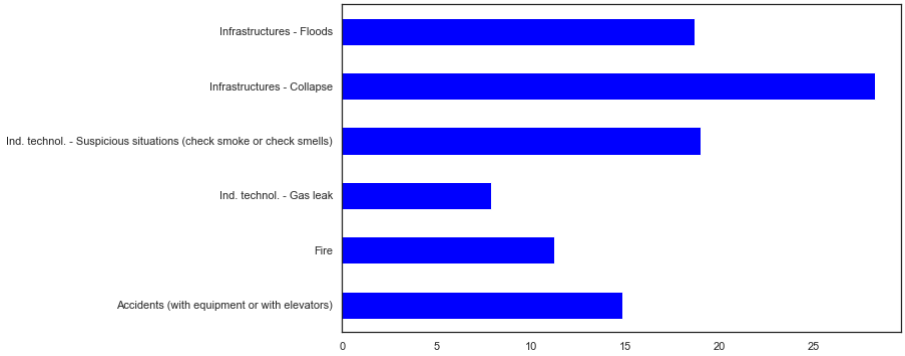


Figure 18 - Distribution of occurrences with low precipitation

The distribution of occurrences when precipitation is moderate is shown in Figure 19. With moderate precipitation levels there is an increase in flood-related occurrences with a proportion of 47.98%, followed by collapse with 22.57%. The remaining occurrences occur in smaller proportion and for suspicious situations presents a proportion of 11.64%, accidents with equipment and elevators 7.71%, fires with 5.99% and suspicious situations with 3.83%.

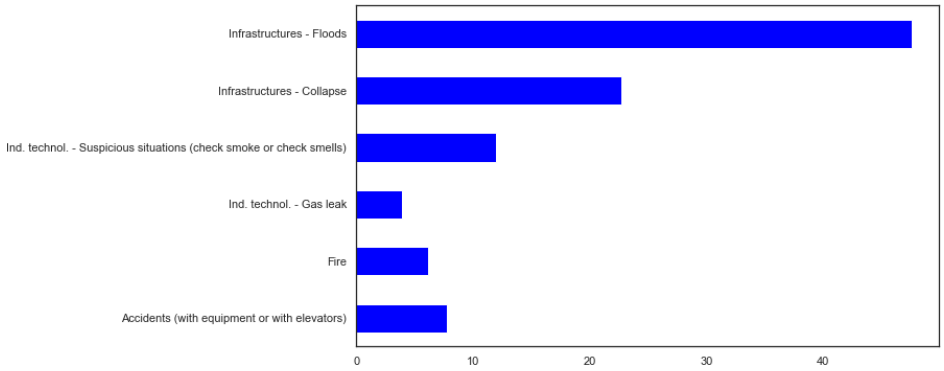


Figure 19 - Distribution of occurrences with moderate precipitation

Figure 20 shows the distribution of occurrences when precipitation levels are heavy. In this case, it is verified that the predominance of occurrences referring to floods with a proportion of 71.81% followed by collapses although in a lower proportion (16.37%). With this level of precipitation, there is a predominance of occurrences referring to floods, with the rest presenting less significant proportions with 2.45% for suspect situations, 2.15% for accidents with equipment and elevators, and 0.9% for gas leaks.

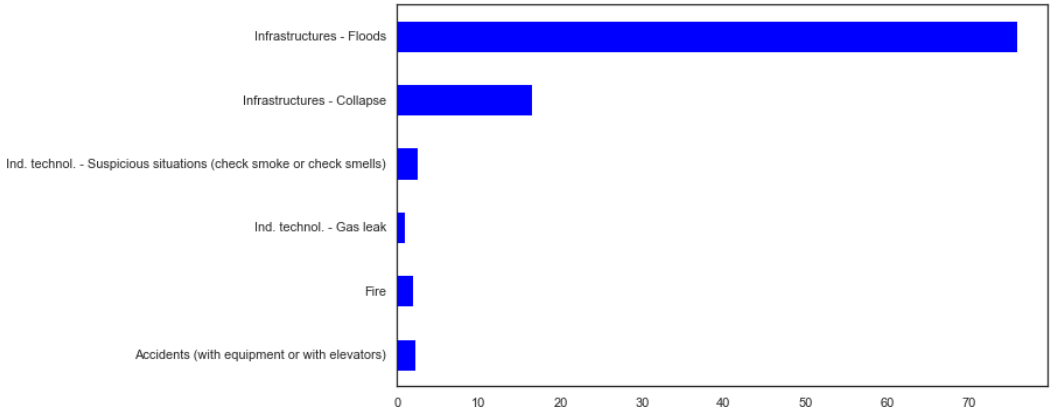


Figure 20 - Distribution of occurrences with heavy precipitation

From the analysis of occurrences according to the four precipitation levels, it is possible to conclude that there are two types of occurrences, namely floods and collapses that increase when precipitation levels increase. In the case of floods, the increase in

incidence depending on precipitation levels is outstanding, since in cases where the precipitation was zero its incidence was 4.02%, in situations of low precipitation it was 19.23%, in situations of moderate precipitation it was 47.98%, and finally in situations of heavy precipitation it was 75.81%.

Shifting the focus to an analysis of occurrences from a spatial perspective, to verify how occurrences are distributed throughout the city of Lisbon. The visualization of the spatial distribution of occurrences was carried out using Power Bi, a Microsoft analysis tool. This tool allows creating heatmaps that consist of the geospatial representation of data in an intuitive way where colors are used to represent areas with different concentrations of points, and thereby showing the patterns of occurrence through colors [68]. The color gradient represents the intensity of the phenomenon that is being analyzed, and for this case the stronger colors represent a higher incidence of occurrences and the less dense colors represent a lower incidence. To conclude, heatmaps constitutes an easy tool for analysis where the visual component is easy to understand.

In this sense, heatmaps were created for the six types of occurrences that most affect buildings in the city of Lisbon and analyzed according to their categories. The first category analyzed was the Infrastructures, which includes collapses and floods. Figure 21 show the spatial distribution of collapses (A) and floods (B).

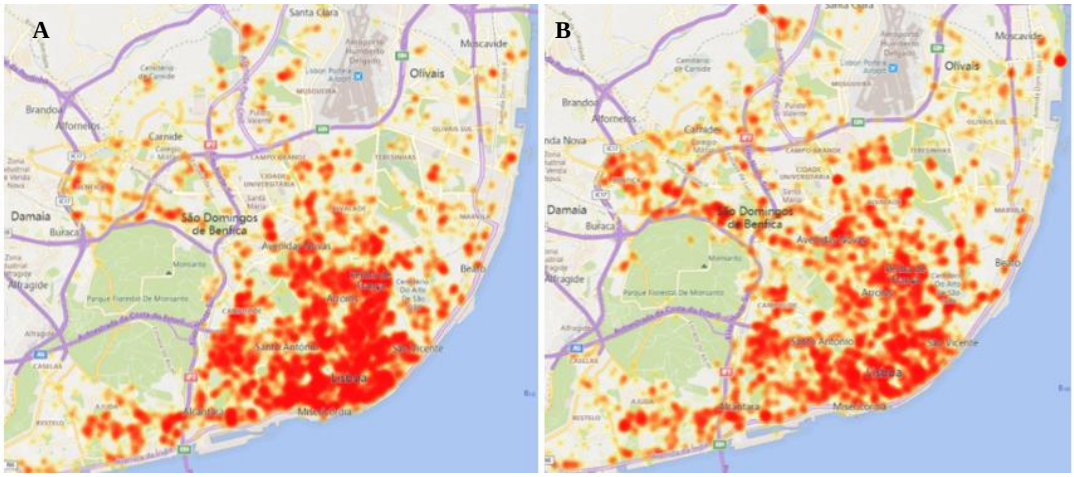


*Figure 21 - Spatial distribution of the occurrences. Figure **A** shows the spatial distribution of Collapses and figure **B** shows spatial distribution of Floods*

From the heatmaps presented above it is possible to infer that the occurrences related to collapses (A), which is the type of events that most affects the city of Lisbon, have a higher concentration of points in the central zone of the city, which means that collapses affect mainly parishes such *as Arroios, Santo António, São Vicente, Misericórdia,*

Campolide, Avenidas Novas, Penha de França, and areas of the Historical Center of Lisbon.

The occurrences referring to floods (B), similarly to collapses, have a higher concentration in the downtown area of the city with the difference that this type of occurrence also happens with high incidence in the northwestern part of the city, namely the parishes of *Benfica* and *São Domingos de Benfica*.

Figures 22 refer to the geospatial distribution of occurrences regarding the Technological-Industrial category. This type of occurrence includes Suspicious situations (check smoke or check smells) and Gas Leaks. The suspicious situations (A) present a higher concentration in the central zone of the city of Lisbon, namely the parishes of *Arroios, Santo António, São Vicente, Misericórdia, Campolide, Avenidas Novas, Penha de França* and areas in the Historical Center of Lisbon. On the other hand, situations concerning gas leaks (B) are more concentrated in the Lisbon Historical Center area and the districts of *Penha de França, Arroios*, and *Benfica*.

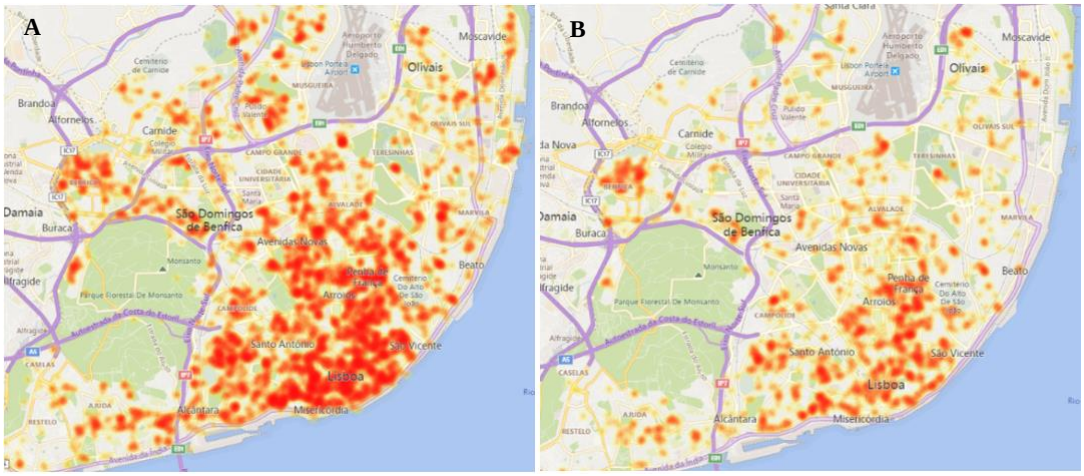


*Figure 22 - Spatial distribution of the occurrences. Figure **A** shows the spatial distribution of Suspicious situations (check smoke or check smells) and figure **B** shows spatial distribution of Gas leaks*

Lastly, Figures 23 show the geospatial distribution of those related to accidents (with equipment and elevators) and fires.

In terms of accidents involving equipment or elevators (A), this type of occurrence, unlike the types of occurrences already analyzed, does not present a higher concentration in a single Lisbon area, but instead affects the entire Lisbon city area with a similar proportion. On the other hand, although fires (B) are a type of occurrence that in general is registered in the entire Lisbon area, their concentration is slightly higher in the Lisbon historic center area.

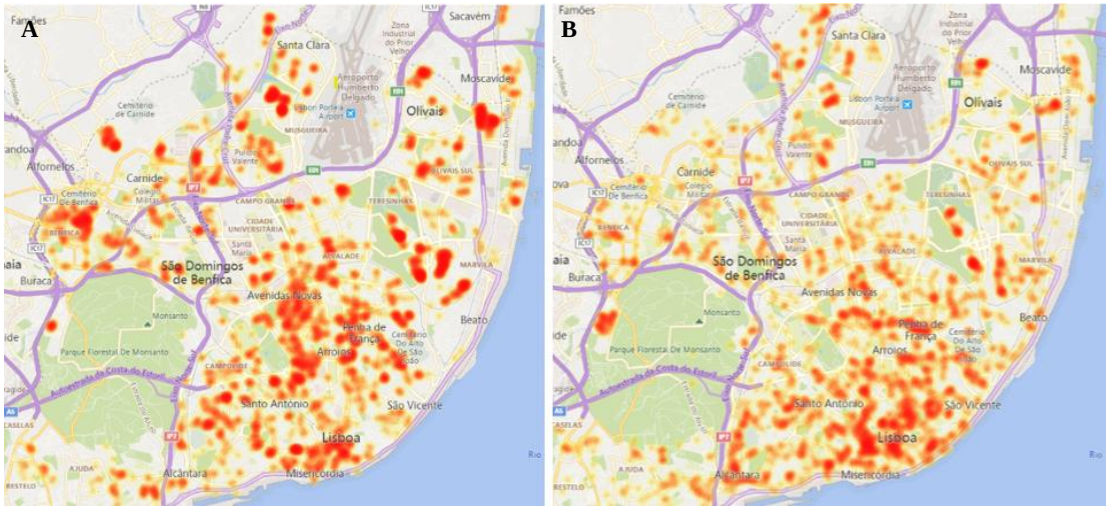


*Figure 23 - Spatial distribution of the occurrences. Figure **A** shows the spatial distribution of accidents (with equipment or elevators) and figure **B** shows spatial distribution of fires*

In summary, from the heatmaps, it is possible to verify that certain types of occurrences such as collapses, floods, suspicious situations (check smoke or check smells), gas leaks, and fires occur mainly in the central area of the city of Lisbon, except for accidents with equipment and elevators where their geospatial distribution is similar throughout the city.

Since there is a concentration of occurrences in a specific area of the city of Lisbon, it was sought to deepen the knowledge about the city by analyzing aspects such as the state of conservation of buildings and the average age of buildings in the different parishes. With this in mind, an analysis was conducted through the geospatial visualization of two aspects of the buildings in the city of Lisbon, namely the proportion of the buildings in the city that are degraded and in need of major repair and the average age of the buildings per parish.

Through the spatial visualization of the buildings that are degraded or in need of repair and through the visualization of the parishes where the oldest buildings are located, it is possible to establish the association between the spatial concentration of occurrences and the condition of the buildings. For this purpose, two heatmaps were prepared and are shown in Figure 24 and Figure 25.

Figure 24 shows that the downtown area of the city of Lisbon is where the most degraded buildings are located, with main emphasis on the historical central zone of the city including the following parishes: *São Vicente, Misericórdia, Arroios, Santo António,* and the Historical Center of Lisbon.

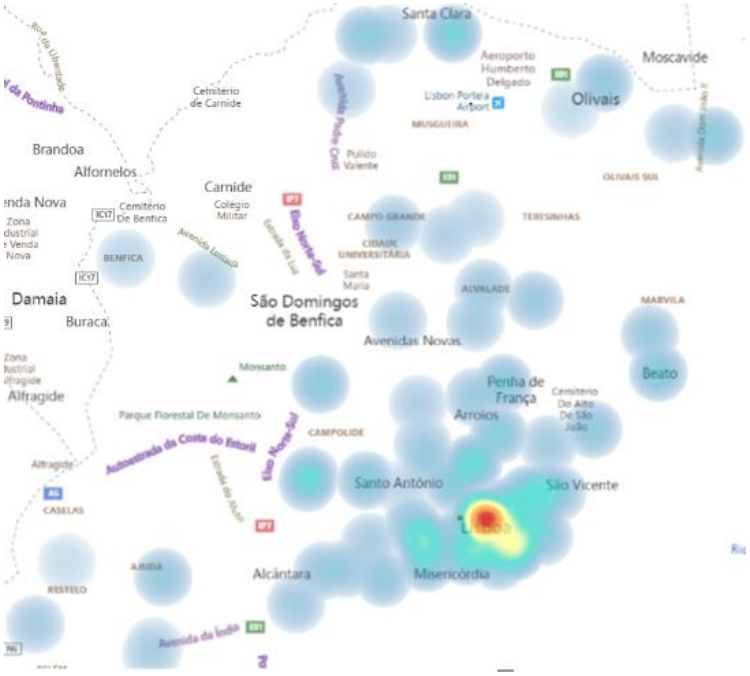


Figure 24 - Spatial representation of the proportion of buildings that are degraded or in need of major repairs

Regarding the spatial representation of the average age of the buildings by parish (Figure 25), it is possible to verify that the oldest buildings are mainly located in the downtown area of the city whereas the historical center area of the city concentrates the oldest buildings. Other parishes with old buildings are *São Vicente*, *Misericórdia*, *Arroios, Santo António, Alcântara, Olivais*, and *Beato*

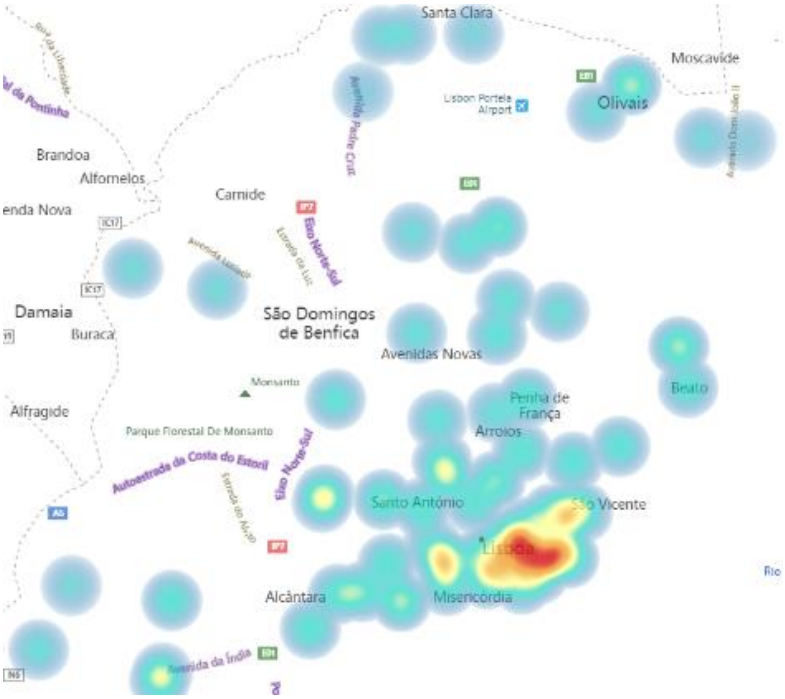


Figure 25 - Spatial representation of the average age of the buildings per parish

From the conclusions reiterated from the two heatmaps created, it is possible to infer that the areas where the older buildings are concentrated and where there is a greater proportion of degraded buildings or with major needs of repair are more affected by the types of occurrences such as collapses, floods, suspicious situations (check smoke or check smells), and gas leaks.

Lastly, one of the objectives of this dissertation consists of analyzing the human and material resources allocated to each type of occurrence. For this purpose, the bar charts in Figure 26 and Figure 27 were created, which show the average number of material resources (vehicles) and human resources spent in the response of each type of occurrence.

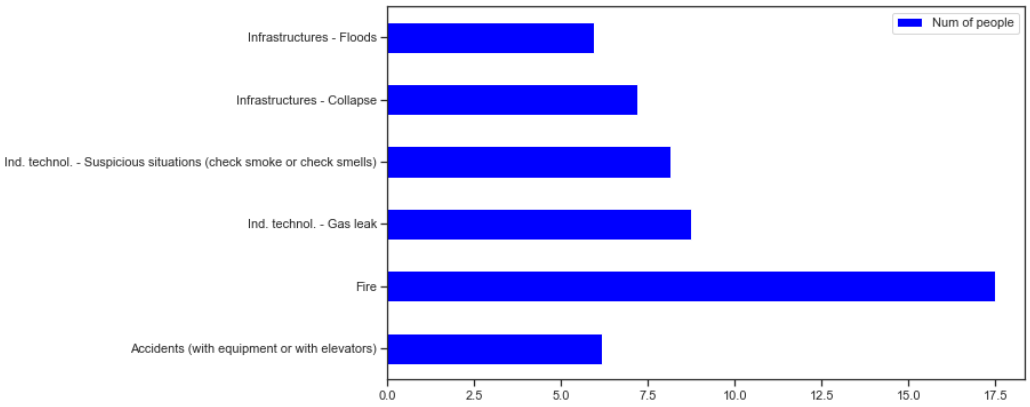


Figure 26 - Human resources allocated to each occurrence

Regarding human resources (Figure 26), it can be seen that fires, despite not being the type of occurrence that affects the buildings in the city of Lisbon, are the type of occurrence where a greater number of firefighters are allocated for the response. The remaining types of occurrences require fewer elements, with the average number of elements allocated varying between 6 to 10 elements.

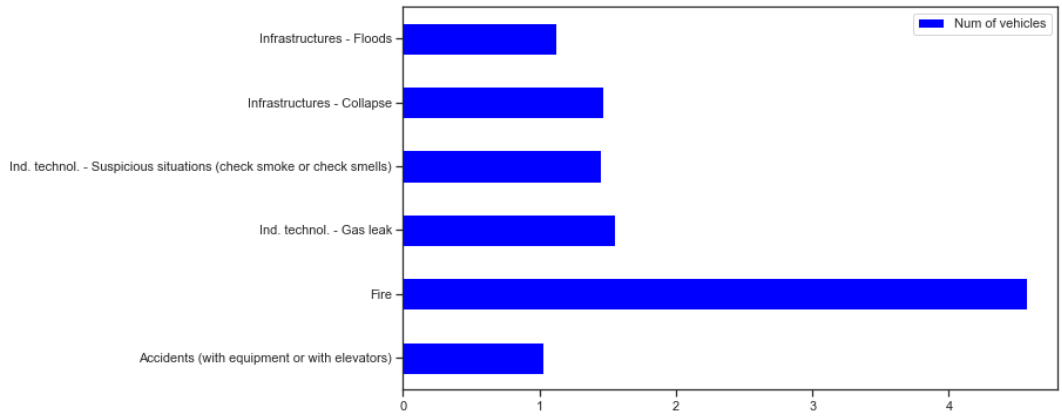


Figure 27 - Material resources allocated to each occurrence

In terms of the allocation of material resources (Figure 27), it is also in cases of fire that a greater number of vehicles are allocated, with an average of approximately 5 vehicles allocated. For the remaining types of occurrences, the number of vehicles allocated is lower.

4.1.2 Data Visualization – *Na Minha Rua Lx* dataset

The analysis of the data from the *Na Minha Rua Lx* application aims to deepen the knowledge about the types of occurrences reported in the application, constituting a tool to help decision-makers to manage the city in an informed and efficient way.

The first analysis performed had the purpose of analyzing how the variables are correlated. For this purpose, a correlation matrix was created, and as can be seen in Figure 28, no variable presents a strong correlation with the variable that we intend to analyze in this dissertation, which are the types of occurrences. Note that strong correlations are verified between variables that were created from the same variable, in this case, the variable "Month" and "Year" that resulted from the variable "Date-Time".

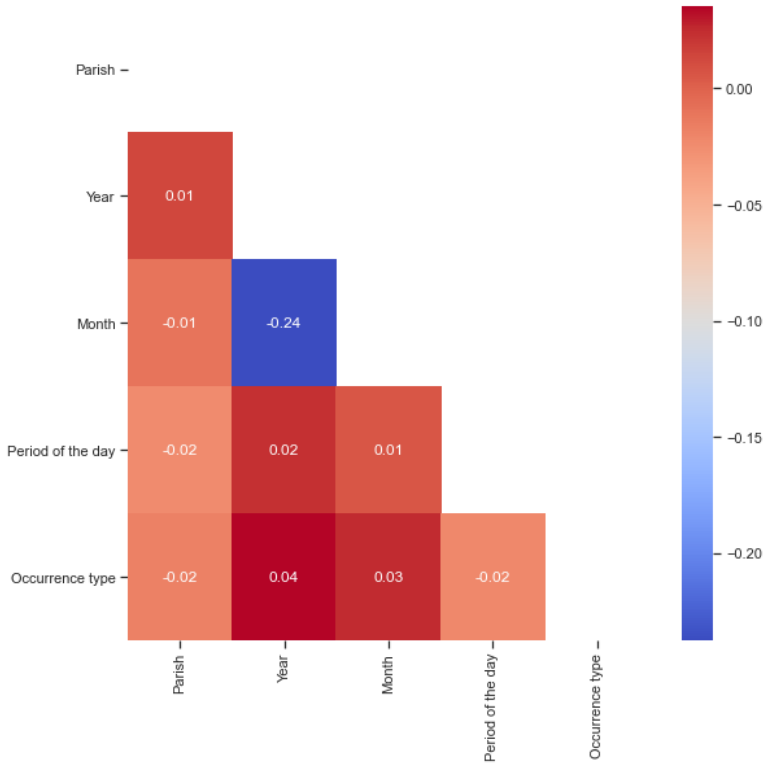


Figure 28 - Correlation matrix

Next, the analysis was focused on understanding the distribution of the data in the dataset with the purpose of obtaining an overview of the frequency of occurrences

registered in the application throughout the years. In this way, to understand how the data is distributed over the years, a histogram was elaborated, and the distribution of the occurrences reported between 2017 and 2019 is shown in Figure 29.

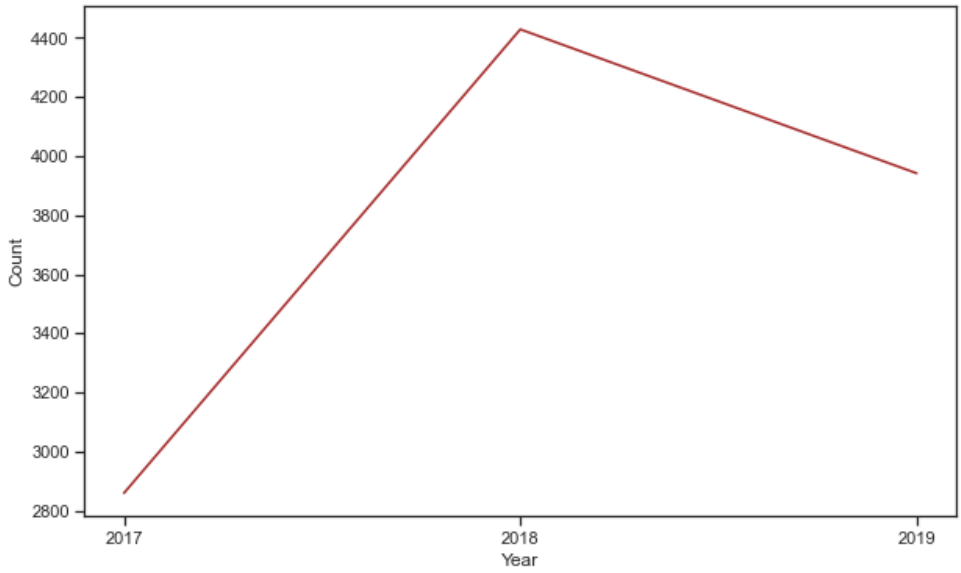


Figure 29 - Distribution of the reported occurrences over the year

From the direct analysis of the graph, it can be verified that there was an increasing trend in the recording of occurrences in the application between the year 2017 and 2018, wherein in 2017, 2 859 occurrences were reported and in the year 2018, the number of occurrences was 2 859. The year 2018 was when there was a greater number of registrations, from this year on there was a downward trend with 3 931 occurrences reported in 2019.

In order to deepen the knowledge about the types of the occurrences reported in the application, the distribution of each type of occurrence was analyzed and the results can be seen in Figure 30.

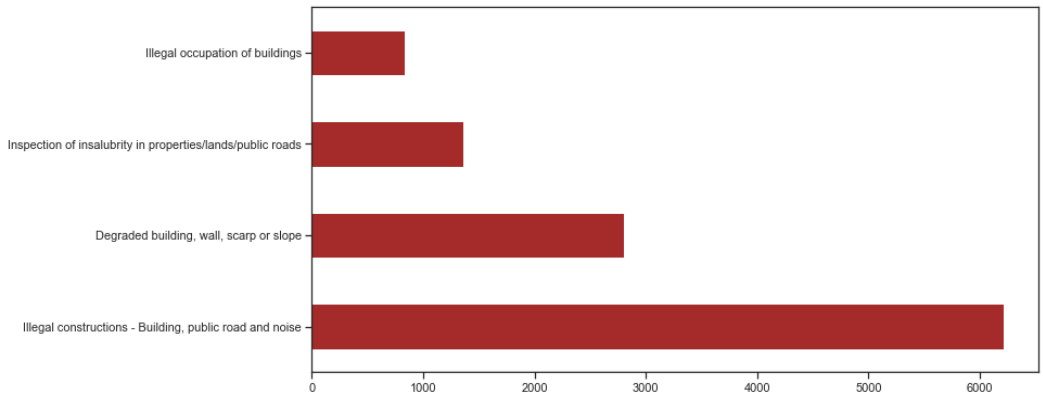


Figure 30 - Distribution of the types of occurrences

From Figure 30, it is possible to verify that there are four main types of occurrences reported in the application, namely Illegal constructions - building, public roads and noise, Degraded building, wall, scarp or slope, Inspection of insalubrity in properties/lands/public roads, and Illegal occupation of buildings.

In terms of their distribution, it is noted that the occurrences are not equally distributed in the dataset since there is an over-position of the type of occurrence referring to Illegal constructions in relation to the other types of occurrences once, during the period under analysis, 6 217cases of Illegal constructions - Building, public roads, and noise were reported. The other typologies are in lesser proportion with 2 799 cases of Degraded building, wall, scarp or slope, 1 365 cases of Inspection of insalubrity in properties/lands/public roads, and 834 Illegal occupations of buildings

Analyzing the distribution of these types and occurrences per year, it is clear from the graph shown in Figure 31 that every year there is a large number of reports corresponding to cases of illegal construction, while the other types of occurrences are reported less frequently.

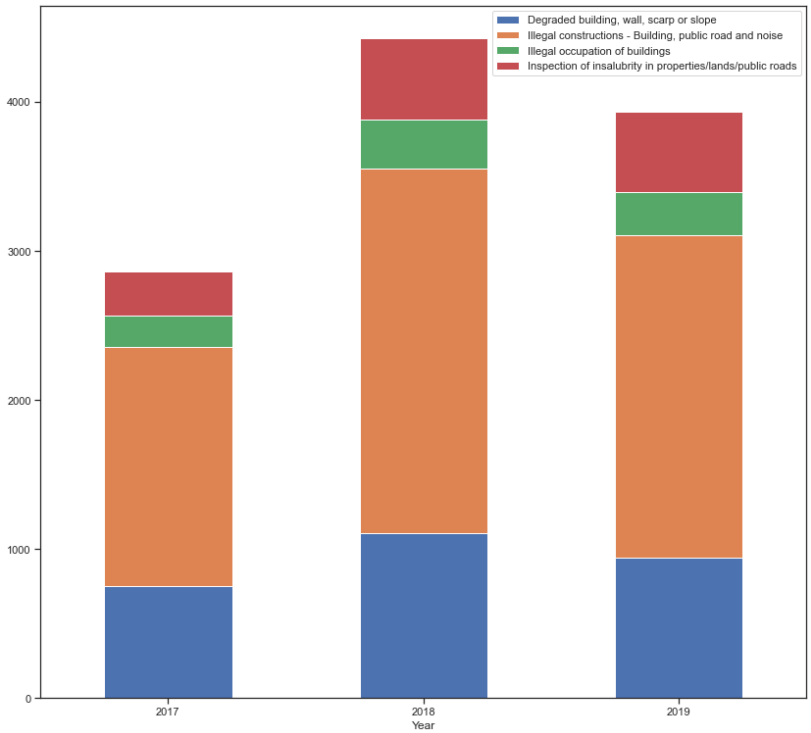


Figure 31 - Distribution occurrences per year

Finally, to complete the global analysis of the occurrences registered in the application, the frequency of use of the application throughout the day was analyzed (Figure32) and

it was found that, similar to the case of the firefighters' occurrences, the records in the application *Na Minha Rua Lx* are made mainly in the afternoon period and the dawn period is where a smaller number of registrations are made.

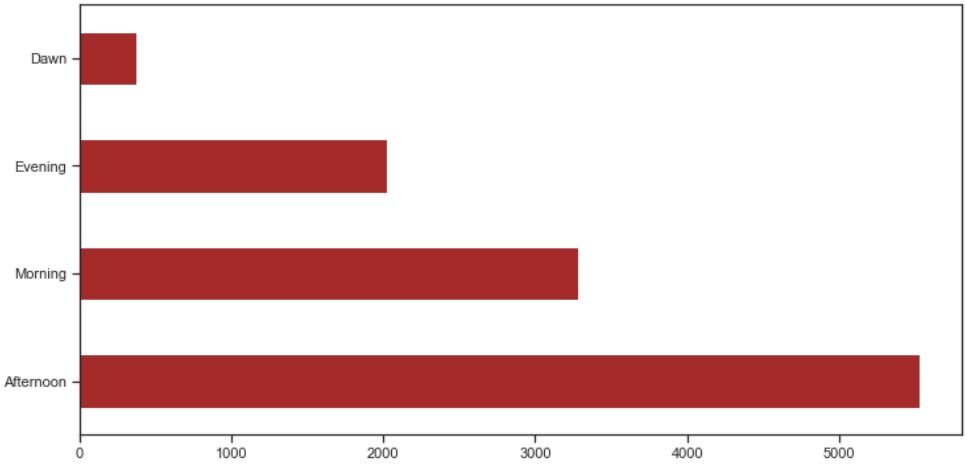


Figure 32 -Distribution of occurrences per period of the day

After a general description of the distribution of occurrences over the years, an analysis of the types of occurrences reported, and the analysis of the registrations of the occurrences by period of the day, as defined in the objectives of this dissertation, the focus changed to occurrences that affected buildings in the city of Lisbon. In this way the analysis is now focused on only two types of occurrences reported: Illegal occupation of buildings and Degraded building, wall, scarp or slope.

With this in mind, similarly to the temporal analysis carried out for the firefighters' dataset, the temporal distribution of occurrences regarding Illegal occupation of buildings and Degraded building, wall, scarp or slope was analyzed. The results are presented in Figure 33.

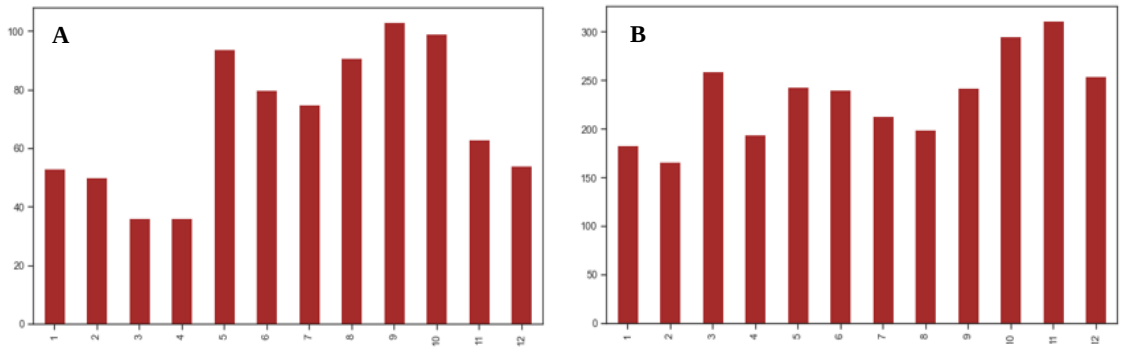


*Figure 33 - Temporal distribution of the occurrences. The bar chart from figure **A** shows the temporal distribution of illegal occupation of buildings and the bar chart from figure **B** shows the temporal distribution of degraded buildings, wall, scarp, or slope*

Cases regarding Illegal occupation of buildings (A) are reported in greater expression between the months of May to October. On the other hand, cases of Degraded buildings, wall, scarp, or slope (B), are reported in greater expression the last four months of the year and the months of March, May.

Shifting the focus to an analysis of occurrences from a spatial perspective, heatmaps were created that present the geospatial distribution of the two types of occurrences that affect the city's buildings with the main goal of verifying how occurrences are distributed throughout the city. From the heatmaps, it is possible to verify that both in cases of Degraded building, wall, scarp or slope and Illegal occupation of buildings these are registered with a higher incidence of the downtown area of the city.

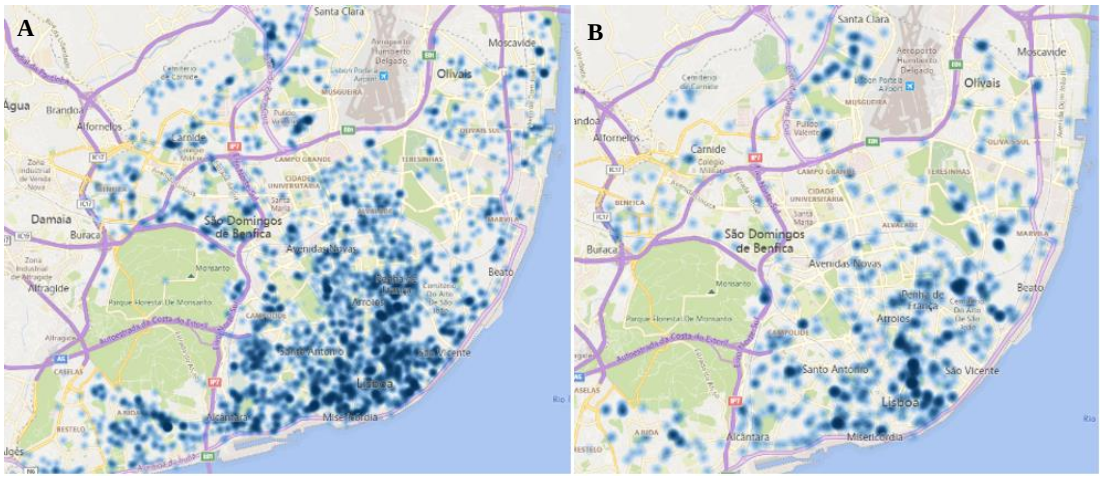


*Figure 34 - Spatial distribution of the occurrences. Figure **A** shows the spatial distribution of degraded buildings, wall, scarp, or slope and figure **B** shows the spatial distribution of illegal occupation of buildings*

As shown in Figure 34, the cases referring to Degraded building, wall, scarp or slope (A) have higher incidences in the following parishes: *Penha de França, Arrois, Avenidas Novas, Misericórdia, Santo António, São Vicente, Ajuda, São Domindos de Benfica* and *Campolide*.

The occurrences related to Illegal occupation of buildings (B) have a higher incidence in the parishes of *Arroios, São Vicente, Santo António, Penha de França, Misericórdia*, and the historical downtown area.

In summary, results from the modeling phase showed that the application of DM techniques to the firefighters' dataset allowed to verify the types of events that affect the city of Lisbon as well as the analysis of their temporal and spatial pattern.

On the other hand, it was verified that in the dataset with data from the application *Na Minha Rua Lx*, the data reported in the application are not of the same type as the data

registered in the firefighters' occurrence management system, since the occurrences involving buildings are Illegal occupation of buildings, Degraded building, wall, scarp, or slope. Furthermore, it was verified that the occurrences of both the firefighters and the application cover the same areas, and that in both cases there is a predominance of occurrences in the historic center area of the city.

4.2 Prediction process

This chapter is dedicated to one of the main objectives of this dissertation, which consists of applying predictive models to predict the occurrences. Since this is a classification problem and the variable that is intended to be predicted is a categorical variable ("Occurrence Type"), supervised classification algorithms [69] were applied and then compared to determine the most efficient classification algorithm for this case. The following predictive models were applied: Random Forest, Decision Tree Classifier, Support Vector Machine, Gaussian Naive Bayes, and Logistic Regression.

Before proceeding to the application of the predictive models it was necessary to make a feature selection, i.e., the selection of relevant variables for the construction of the predictive models, and the feature selection was conducted using the correlation matrix presented in Figure 11.

Considering the data presented in the correlation matrix, only the variables that have greater correlation with the variable Occurrence Type were selected, since it is the target variable (independent variable). Thus, the following variables were selected as dependent or explanatory variables: "PRCP", "RH" "Num of people", "Num of vehicles", "Avg building age", Avg air temp", "Resident pop", and Avg WS".

To train the predictive models, it was necessary to split the data to allow not only the training but also the testing of the models. The training set is used to find the relationship between independent and dependent variables, while the test set evaluates the model's performance. In numerical terms, the division was made so that 70% of the data was for training and the remaining 30% of the data was for model testing. In this way, it was generated the training and testing dataset with the following dimensions, which can be seen in Table 19.

Table 19 - Training and testing dataset dimension

Training dataset		Testing dataset	
Rows	Columns	Rows	Columns
11 027	8	4 726	8

After defining the X and Y variables and dividing the data for testing and training, the algorithms were applied, and the predictive results were analyzed. To this end, for each algorithm, the respective accuracy was calculated, which measures the effectiveness of a model as a proportion of actual results for total cases [70].

The first model applied was the Decision Tree model [71]., which is a non-parametric supervised learning method used for classification problems. It is a tree-like diagram used to define the course of an action, with each tree representing a possible decision, occurrence, or reaction. In other words, decision trees are based on algorithms that divide the original dataset into several homogeneous subsets, and these subsets can be further divided, resulting in subsets with a higher level of homogeneity. A decision tree consists of several nodes and arcs where each node represents a test that is applied to a dataset that serves to divide the data into smaller and more homogeneous subsets, and each arc connects a node to the next node or to a leaf. The leaves represent final nodes, i.e. the nodes with the most homogeneous data set that the tree can produce, and no further splits can be made.

In this way, the problem presented in the first node is decomposed into simpler problems until the final nodes (the leaves) are reached, which assign a certain class to the data. One of the advantages of this model is that this model allows different types of variables (nominal, ordinal, and categorical), knows how to deal with omitted values, and has a sensitivity to the scale factor.

The implementation of this model was conducted using the classifier *DecisionTreeClassifier* wherein the first moment the model was created based on the training data with known class labels. After the model creation, the prediction was made on the testing data, and, in this phase, the classifier was used to predict class labels of unknown data and it was obtained an accuracy of 50%.

The second step consisted of improving the performance of this model, and in this case, the *Grid Search* function [72] was used to find the best hyperparameters combination that provides the best predictive results. To use *Grid Search* function, the first step consists of creating a list of hyperparameters that is intended to test, saving them

as parameters. As the second step, the *Grid Search* function is provided with the classifier (model), and the parameters we want to see tested. The Grid Search function builds models for each set of combinations of hyperparameters and then evaluates the performance of each one. The combination that provides the best result is selected and the model is reapplied to the data, this time with the new hyperparameters. In this case, applying the model using the best hyperparameters, increased the accuracy value to 57%.

The second predictive model applied was the Random forest, which is a classification algorithm that can be used in both classification and regression problems. This algorithm is based on the ensemble learning principle that consists in combining different classifiers in order to find the best result for a given problem. In general, the Random Forest algorithm consists of a collection of decision trees (known as a forest) and the forest generated by the algorithm is trained through bagging or bootstrap aggregating (an ensembling method). The bagging technique is related to the use of different data samples (training data) rather than a single sample. In this way, using multiple sets of training data, the decision trees produce different outputs and these outputs are ranked, and the one with the highest score is selected as the final output [73].

In short, Random Forest is a classifier that contains a certain number of decision trees in the different subsets of a dataset and uses the average to improve the predictive accuracy of that dataset. In other words, instead of relying on results from a single tree, this algorithm considers the predictive results from each tree and the selection of the final output is made based on the majority of the prediction votes from each tree [74].

In this case, this model was implemented using the classifier *RandomForestClassifier* and, at first, the model was created, and afterward, the fit was made on the training data. With the fit process completed, the prediction was performed on the testing data, where it was obtained an accuracy of 56%.

In order to increase the accuracy, the performance of this predictive model was tuned using *RandomizedSearchCV* [75]. This function, similar to *Grid Search*, randomly passes a set of hyperparameters and calculates the scores, and then provides the best set of hyperparameters that provides the best predictive results. This model has several parameters that allow its performance to be improved, such as estimator, *param_distributions*, *cv*, n_*iter*, n_*jobs*, *verbose*, and *random_state* [76].

With the application of this function, the best parameters in terms of *n_estimators*, *max_depth*, and *max_features* were obtained and, using these parameters, there was no increase in accuracy as the model remained at 58%.

The third model applied was Support Vector Machine, which is a non-parametric classification algorithm that analyzes data, creates patterns, and classifies the data into classes, and this data is represented by points in N-dimensional space (N - number of features). The main goal of this algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the points that represent the data. In order to separate the data, there are several possible hyperplanes that can be chosen. The goal is to find the hyperplane that has the maximum margin, that is, the maximum distance between the points between classes. This maximization of the margin provides reinforcement for future data points to be classified with greater confidence [77].

For this model the SVC classifier was used and, similarly to the steps applied in the models described above, in a first moment the model was defined where the fit was made on the training data and later the prediction on the test data, obtaining an accuracy of 23%.

To improve the performance of the Support Vector Machine model, there are several parameters that can be analyzed to find the best predictive results, such *C* (regularization) which is a penalty parameter that represents the incorrect classification or, in other words, the error term. The higher the value of correct C is, the better the data is classified. The other hyperparameter is *gamma* whose function is to influence the calculation of the plausible separation line, i.e., the *gamma* value determines the fit of the model and the higher it is, the better fitted the model is.

To calculate the best combination of parameters mentioned above the *Grid Search* function was used to calculate the best value of C and gamma fitted to the training data. After finding the best estimator, the prediction was made using the test data, where there was an increase in accuracy from 23% to 57% after tuning the model.

The fourth model implemented was K-Nearest Neighbors. It is a supervised learning algorithm, and this algorithm is based on the idea that similar things are located close to each other. Thus, it is a model where the classification process is based on the nearest neighbor rule, that is, the classification process is based on a comparison made between the object that is intended to be classified and the objects stored in a training set. The objects, both those intended to be predicted and those that compose the training dataset,

are represented by a set of characteristics and, in the case of the objects in the training set, in addition to having characteristics that differentiate them from each other, they also have a classifying attribute that allows them to be assigned to a given class [78].

After the implementation of this model using the classifier *KNeighborsClassifier* and without any hyperparameter, an accuracy value of 44% was obtained. However, the performance of this model can be improved by changing the number of neighbors selecting the optimal value of *k* (neighbors) that has the lowest error rate value. The error rate corresponds to the correctness of the predictions, so the lower the error rate value the better the model performance and vice-versa. With this in mind, an interaction cycle was created with the number of neighbors fixed at 40, where the classifier will fit the training data, obtain the values of the prediction and append the average value of the error rate for k=1 to k=40 in order to find the value of *k* that has the lowest error rate value. The results of this iteration can be seen in the graph in Annex D - K versus Error Rate. After the implementation of the model with the number of neighbors that has the lowest error rate value, an accuracy of 44% was obtained.

The fifth model implemented Naive Bayes, a probabilistic model that is based on Bayes' theorem [79]. This theorem describes the probability of occurrence of a certain event based on knowledge that can be related to this event [80]. In other words, it is a probabilistic predictive model that calculates conditional probabilities of classes, given the attributes, in order to predict the most likely class according to the highest inferred probabilistic value [81].

This model was implemented using the classifier *GaussianNB* where first the model was created, then fitted to the training data, and when predicted on the testing dataset it was obtained an accuracy of 50%.

The performance of this model can be improved by using a hyperparameter named *var_smoothing*. This parameter calculates the stability to extend or smooth the curve of the normal distribution [82].

Therefore, it computes more samples that are far away from the mean of the distribution. Then, the *Search Grid* function was used to find the best value corresponding to the *var_smoothing* parameter and the model was reapplied, but the accuracy remained at 50%.

The last predictive model implemented was Logistic Regression, which is a machine learning algorithm that can be used in classification problems. This is a predictive algorithm based on the concept of probability that allows the calculation of the probability associated with a given event according to a set of explanatory variables. This predictive algorithm is considered appropriate for modeling cases where the variable to be predicted is binary or categorical. The prediction process of this algorithm is done by weighting the explanatory variables, dominated X, based on the influence these variables have on the occurrence of the variable intended to be predicted, the independent variable (Y), and in this way it is possible to estimate the probability of occurrence of the event under analysis [83] .

After implementation of this model through the *LogisticRegression* classifier, an accuracy value of 39% was obtained. This model was tuned through the following hyperparameters: *penalty* that concerns the penalty form, *C* that represents inverse of regularization strength, *solver* that concerns the algorithm to use in the optimization problem and *max_iter* that corresponds to the maximum number of iterations taken for the solvers to converge [84].

The *Grid Search* function was used to find the combination of the best parameters that provide the best predictive values for this data, and after implementing the model considering the parameters provided by the *Grid Search* function an accuracy of 55 % was obtained.

After applying the five predictive models, it was found that the Random Forest model performed best when applied to the firefighters' dataset, and using the best hyperparameters, an accuracy of 58% was obtained. These results can be seen in Figure 35 which contains the summary of the predictive results.

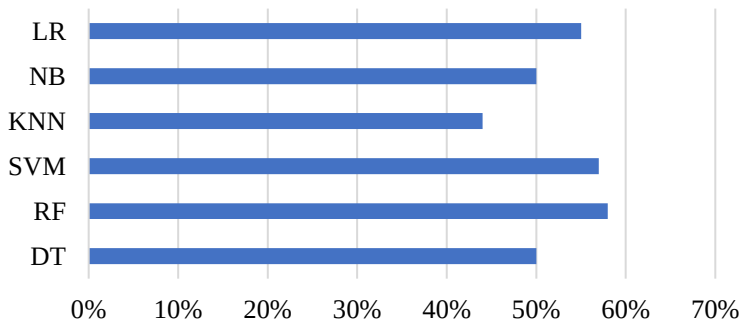


Figure 35 - Summary of the predictive results

In short, the prediction results were not satisfactory (best result 58%), and these results can be explained by the fact that there was no strong correlation between the variables and also by the fact that the independent variable (the one predicted) had six possible values: Infrastructures - collapses, Infrastructures - floods, Ind. Technol. -suspicious situations (check smoke or check smells), Ind. - Gas leak, Fires, and Accidents (with equipment or with elevators).

4.3 Evaluation

As previously mentioned, the methodology adopted in this research presupposes the evaluation of the project with the stakeholders. For this purpose, after the presentation of the results to those responsible for the TIU department of the Lisbon City Hall, a satisfaction questionnaire was applied with the main objective of evaluating the quality of the project and to what extent the results obtained meet the objectives and needs identified by the Lisbon City Hall. The satisfaction questionnaire was elaborated considering the main goal of this research which consists of analyzing the data made available by the Lisbon City Hall in order to extract insights that can help improving the way the city is managed.

The evaluations were carried out by 3 experts and, regarding the characterization of the panel of evaluators, they hold degrees in territorial engineering, geology and architecture and the average experience of the evaluators is 26 years. The questionnaire aims to evaluate five criteria, namely effectiveness, consistency with the organization's objectives, utility, level of detail, clarity, applicability, transferability, and impact. These criteria have four evaluation possibilities: NA (Not Achieved), PA (Partially Achieved), LA (Largely Achieved), and TA (Totally Achieved). The result is shown in Table 20.

The evaluation assigned to the criteria were PA and LA with the exception for the Clarity criterion which was assigned a TA classification by Evaluator 1. The reason why the criteria were classified as PA and LA is due to the fact that the stockholders responsible for this project initially expected this analysis to be conducted at the level of the buildings with a higher level of detail relating the state of conservation of each building with the events that occurred, but it was later concluded that the data did not allow such analysis and therefore the possible analysis would be from a high-level perspective.

Furthermore, it was expected that the events reported in the *Na Minha Lx* application could complement the firefighters' data, but results showed that the events reported in the application, despite covering the same areas, do not have events of the same extent and weight as the events recorded in the LFBR management system.

Table 20 - Results of the evaluation

Project evaluation				
Criterion	**Objective statement**	**#Eva1**	**#Eva2**	**#Eva3**
Efficacy	The research effectively informs about incidents that affect buildings in the city of Lisbon	LA	LA	LA
Consistency with the organization's objectives	The results achieved are aligned with the objectives set and correspond to the needs identified, providing relevant insights to the decision-makers	PA	PA	PA
Utility	Through the research, useful insights were extracted for the decision support process	LA	LA	LA
Detail level	The proposed solution provides the necessary level of detail to assist in the decision support process	LA	PA	PA
Clarity	The research provides clear and easy-to-understand information from the elaborate graphics	TA	LA	LA
Applicability	The solution has practical applicability in the field of disaster management and civil protection	PA	PA	PA
Transferability	The results gathered in this research can be applied to other contexts or areas	PA	LA	LA
Impact	The results achieved positively impact the way disaster situations are managed, thereby increasing the city's resilience	PA	PA	PA

In summary, through the data visualization techniques it was possible to extract patterns of occurrences of the events that affect the city of Lisbon, and it was verified that at the temporal level there is a predominance of certain events such as collapses, floods and suspicious situations in the autumn/winter months. In terms of spatial patterns, it was verified that the central area of the city of Lisbon is generally the area with the highest incidence of occurrences. In turn, the central area of the city is where are concentrated the oldest buildings and the buildings in need of repair.

With the data visualization phase completed, predictive algorithms were applied, and the results obtained were not satisfactory since the data was not rich and there were no

strong correlations between the variables. On the other hand, another fact that may have difficulted the achievement of better predictive results may be the fact that the variable that is intended predict have six possibilities of values.

Finally, the project was evaluated with the stakeholders responsible for the project and the evaluation obtained for each criterion varied between LA and PA.

Chapter 5 - Conclusion and future work

5.1 Conclusion

The research carried out in this dissertation shows that the evolution in the IT area has positively impacted the disaster management area since when city management policies are grounded on analytical results resulting from the application of DB technologies to the large amounts of data that have begun to be stored, it allows increasing the authorities' capabilities to cope with disaster situations.

The spatial-temporal analysis conducted in this research is important to understand the types of occurrences to which the city of Lisbon is vulnerable. The extraction of knowledge regarding the patterns of occurrence of these events is useful in the various stages of disaster management as it allows generating an overview of easy understanding among the various stakeholders.

In this way, it is safe to say that the objectives proposed in chapter 1.2 were reached since, in the case of firefighters ' dataset, it was verified that Lisbon's firefighters respond to several types of occurrences, being the incidents included in the category of Services which includes activities such as road cleaning services, opening and closing doors, hospital transport, water supply, and prevention services at shows, sports, and patrolling the events of greatest expression in the dataset since they represent 45.7% of all the data recorded. The first analyses also showed that the afternoon is the period of the day when the greatest number of occurrences are registered, followed by the evening, morning, and lastly dawn, when a reduced number of occurrences are registered. This information becomes useful when allocating material and human resources during the period of the day.

When this analysis was focused on the occurrences that affected the buildings and therefore discarded the remaining occurrences, results showed that collapses, floods, suspicious situations (check smoke or check smells), gas leak, fires, accidents (with equipment or with elevators) are events that most affect the buildings in the city of Lisbon, with main emphasis on the cases of collapses where 3 742 occurrences of this nature were registered between the period from 2013 to 2018.

The temporal analysis was performed using bar charts and was conducted at the month level with the goal of understanding how the incidences of occurrences vary throughout the 12 months of the year. The findings showed that there are events that occur more

frequently in certain periods of the year, such as floods, collapses, and suspicious situations which occur more frequently in the autumn and winter months. In the case of floods, the contrast in the frequency of occurrences between the winter/autumn months and the summer months is outstanding, since in the summer months this number is very low, reaching minimum record values of 9 occurrences. The remaining occurrences have similar incidence levels throughout the months of the year, and in the case of gas leaks the month of January stands out with an increase in records of this type of occurrence. In the case of fires, the month of December is highlighted for registering an increase, and finally, the cases of accidents with equipment or elevators, this type of occurrence presents similar values throughout the months except for the month of July where there is an increase.

The influence of weather conditions on the incidence of occurrences was also verified. This analysis was possible since external data (meteorological data) was added to the dataset which allowed correlating the types of occurrence with climatic factors. The results showed that there are two types of occurrences, namely floods and collapses that increase when precipitation levels increase. In the case of floods, the increase in incidence as a function of precipitation levels is remarkable, since in cases where precipitation was zero its incidence was 4.02%, in situations of low precipitation it was 19.23%, in situations of moderate precipitation it was 47.98%, and finally in situations of heavy precipitation it was 75.81%.

The spatial analysis was conducted using the Power BI tool, as well as its data visualization capabilities to create heatmaps and in this way analyze the spatial distribution of occurrences to extract knowledge about the areas with the highest incidence of certain types of occurrences, and results showed that events such as collapses, floods, suspicious situations (check smoke or check smells), gas leaks, and fires occur mainly in the central area of the city of Lisbon. On the other hand, accidents with equipment and elevators have a similar geospatial distribution throughout the city.

In order to deepen the knowledge about the characteristics of the buildings in the city of Lisbon, data from external sources were added that contained information about the area of the city where the oldest buildings and buildings in great need of repair are concentrated. The results obtained showed that the areas where the oldest buildings are concentrated and the areas with the highest number of buildings with great need for repair

are the areas in the Historical Center of Lisbon, which in turn are the areas with a higher incidence of occurrences.

Lastly, the analysis of the material and human resources allocated to the assistance of occurrences showed that fire is the type of occurrence where a greater number of elements and vehicles are allocated.

In the case of data extracted from the *Na Minha Rua Lx* application, it was verified that the data reported in the application are of four main types: Illegal constructions - Building, public roads and noise, Degraded building, wall, scarp or slope, Inspection of insalubrity in properties/lands/public roads, and Illegal occupation of buildings, being the afternoon the period of the day with the most records of occurrences.

The occurrences involving buildings are Illegal occupation of buildings, Degraded building, wall, scarp, or slope. The types of occurrences recorded in this application differ from the data recorded by firefighters as the data reported in the application is of lesser extension and severity.

Similar to the analysis made for the firefighters' dataset, the temporal analysis showed that cases regarding Illegal occupation of buildings are reported in greater expression between the months of August, September and October, with emphasis on the month of May where there is an increase in this type of occurrence. Cases of Degraded buildings, wall, scarp, or slope, the last four months of the year and the months of March, May, and June stand out, being the months where the highest number of this type of occurrence is recorded.

Spatial analysis showed that, similarly to the occurrences registered by the firefighters, the occurrences reported in the application have a higher incidence in the areas of the Historical Center of Lisbon

Finally, predictive algorithms were applied to the firefighters' data, however the results obtained were not satisfactory and the explanation for these results is due to the poor richness of the data since there were not very strong correlations between the dependent variables and the independent variable had six possible values. Thus, the fact that it was not possible to join the datasets due to their characteristics, the fact that the data were not rich and consequently did not allow a more detailed analysis (as intended by the stakeholders responsible for this project) and achieve better predictive results, were the main limitations identified during this research.

The main contribution of this dissertation consists in the fact that since it is a real case, where real data about occurrences were treated and analyzed, the information extracted from the data analysis techniques applied effectively allows helping in the management of the city since the stakeholders have access to information about events that affect the city and consequently have situational awareness that will allow them to manage the city in an efficient and informed way.

On the other hand, this research, besides the firefighters' dataset, also analyzed the dataset containing data extracted from the application *Na Minha Rua Lx*. In this last case, the analysis of the application data allows not only to verify the registered occurrences but also allows the Lisbon City Hall to be aware of the usefulness and acceptance of the application by the inhabitants, i.e., if the application is or is not widely used by the population.

Lastly, based on the work developed in this dissertation, a paper was accepted in the INTSYS 2021 (5th EAI International Conference on Intelligent Transport Systems).

5.2 Future work

As mentioned in chapter 4.3, the Lisbon City Hall stakeholders expected this analysis to be conducted with a greater level of detail, that is, at the level of buildings, relating occurrences to the state of conservation of each building. However, this was not possible once the available data did not allow such analysis to be conducted since the data provided by the Lisbon City Hall did not contain the buildings reported in the LFBR management system. In this sense, it was identified as future work to carry out this analysis considering the state of conservation of each building reported in the datasets.

It was also identified as future work the aggregation of more information that characterize the buildings in the city of Lisbon, such as the type of construction material of the buildings, year of construction of the buildings and quantity of vacant buildings in the city. This information would increase the richness of the data, enabling a more complete analysis. Additionally, the aggregation of more information would allow feeding the algorithms with relevant information, which would improve their performance.

Lastly, another aspect that was identified as future work was the application of other predictive models in order to verify if they present better results.

References

[1] M. Kobiyama, M. Mendonça, and D. A. Moreno, "Prevenção de desastres naturais," p. 124.

[2] J. J. Wellington and P. Ramesh, "Role of Internet of Things in disaster management," in *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Mar. 2017, pp. 1–4. doi: 10.1109/ICIIECS.2017.8275928.

[3] S. A. Shah, D. Z. Seker, M. M. Rathore, S. Hameed, S. B. Yahia, and D. Draheim, "Towards Disaster Resilient Smart Cities: Can Internet of Things and Big Data Analytics Be the Game Changers?," *IEEE Access*, vol. 7, pp. 91885–91903, 2019, doi: 10.1109/ACCESS.2019.2928233.

[4] C. Yang, G. Su, and J. Chen, "Using big data to enhance crisis response and disaster resilience for a smart city," in *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, Mar. 2017, pp. 504–507. doi: 10.1109/ICBDA.2017.8078684.

[5] Q. Huang, G. Cervone, and G. Zhang, "A cloud-enabled automatic disaster analysis system of multi-sourced data streams: An example synthesizing social media, remote sensing and Wikipedia data," *Comput. Environ. Urban Syst.*, vol. 66, pp. 23–37, 2017, doi: 10.1016/j.compenvurbsys.2017.06.004.

[6] S. A. Shah, D. Z. Seker, S. Hameed, and D. Draheim, "The Rising Role of Big Data Analytics and IoT in Disaster Management: Recent Advances, Taxonomy and Prospects," *IEEE Access*, vol. 7, pp. 54595–54614, 2019, doi: 10.1109/ACCESS.2019.2913340.

[7] L. J. Vale, "The politics of resilient cities: whose resilience and whose city?," *Build. Res. Inf.*, vol. 42, no. 2, pp. 191–201, Mar. 2014, doi: 10.1080/09613218.2014.850602.

[8] A. Zagorecki, D. Johnson, and J. Ristvej, "Data Mining and Machine Learning in the Context of Disaster and Crisis Management," *Int. J. Emerg. Manag.*, vol. 9, pp. 351–365, Jan. 2013, doi: 10.1504/IJEM.2013.059879.

[9] "(PDF) Crowdsourcing Disaster Response." https://www.researchgate.net/publication/268448750_Crowdsourcing_Disaster_Response (accessed Aug. 26, 2021).

[10] M. Yu, C. Yang, and Y. Li, "Big Data in Natural Disaster Management: A Review," *Geosciences*, vol. 8, p. 165, May 2018, doi: 10.3390/geosciences8050165.

[11] A. Gupta, S. Nair, and K. Röder, *Databases and Statistics for Disaster Risk Management*. 2013.

[12] "Welcome to Python.org." https://www.python.org/ (accessed Jul. 03, 2021).

[13] "Visualização de Dados | Microsoft Power BI." https://powerbi.microsoft.com/pt-pt/ (accessed Jun. 20, 2021).

[14] scar Marbn, G. Mariscal, and J. Segovi, "A Data Mining & Knowledge Discovery Process Model," in *Data Mining and Knowledge Discovery in Real Life Applications*, J. Ponce and A. Karahoc, Eds. I-Tech Education and Publishing, 2009. doi: 10.5772/6438.

[15] P. Chapman, R. Kerber, J. Clinton, T. Khabaza, T. Reinartz, and R. Wirth, "The CRISP-DM Process Model," *Discuss. Pap.*, p. 99.

[16] S. Chaudhari, A. Bhagat, N. Tarbani, and M. Pund, "Dynamic Notifications in Smart Cities for Disaster Management," in *Computational Intelligence in Data Mining*, Singapore, 2019, pp. 177–190. doi: 10.1007/978-981-10-8055-5_17.

[17] Z. Alazawi, O. Alani, M. B. Abdljabar, S. Altowaijri, and R. Mehmood, "A smart disaster management system for future cities," in *Proceedings of the 2014 ACM*

international workshop on Wireless and mobile technologies for smart cities, New York, NY, USA, Aug. 2014, pp. 1–10. doi: 10.1145/2633661.2633670.
[18] T. Li, N. Xie, C. Zeng, W. Zhou, L. Zheng, Y. Jiang, Y. Yang, H.-Y. Ha, W. Xue, Y. Huang, S.-C. Chen, J. Navlakha, and S. S. Iyengar, "Data-Driven Techniques in Disaster Information Management," *ACM Comput. Surv.*, vol. 50, no. 1, p. 1:1-1:45, Mar. 2017, doi: 10.1145/3017678.
[19] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *Int. J. Surg.*, vol. 8, no. 5, pp. 336–341, 2010, doi: 10.1016/j.ijsu.2010.02.007.
[20] C. Okoli and K. Schabram, "A Guide to Conducting a Systematic Literature Review of Information Systems Research," *SSRN Electron. J.*, 2010, doi: 10.2139/ssrn.1954824.
[21] "Scopus - Document search." https://www.scopus.com/search/form.uri?display=basic&edit.scft=1#basic (accessed Aug. 09, 2021).
[22] "About Google Scholar." https://scholar.google.com/intl/en/scholar/about.html (accessed Aug. 09, 2021).
[23] "Zotero | Your personal research assistant." https://www.zotero.org/ (accessed May 07, 2021).
[24] E. K. Yamakawa, F. I. Kubota, F. H. Beuren, L. Scalvenzi, P. A. C. Miguel, E. K. Yamakawa, F. I. Kubota, F. H. Beuren, L. Scalvenzi, and P. A. C. Miguel, "Comparing the bibliographic management softwares: Mendeley, EndNote and Zotero," *Transinformação*, vol. 26, no. 2, pp. 167–176, Aug. 2014, doi: 10.1590/0103-37862014000200006.
[25] "PRISMA." http://prisma-statement.org/prismastatement/flowdiagram.aspx (accessed Jul. 21, 2021).
[26] M.-C. Jeong and J. Kim, "Prediction and analysis of electrical accidents and risk due to climate change," *Int. J. Environ. Res. Public. Health*, vol. 16, no. 16, 2019, doi: 10.3390/ijerph16162984.
[27] M. F. Abdullah, M. Ibrahim, and H. Zulkifli, "Big Data Analytics Framework for Natural Disaster Management in Malaysia:," in *Proceedings of the 2nd International Conference on Internet of Things, Big Data and Security*, Porto, Portugal, 2017, pp. 406–411. doi: 10.5220/0006367204060411.
[28] K. M. Briones-Estébanez and N. F. F. Ebecken, "Occurrence of emergencies and disaster analysis according to precipitation amount," *Nat. Hazards*, vol. 85, no. 3, pp. 1437–1459, 2017, doi: 10.1007/s11069-016-2635-z.
[29] A. Alipour, A. Ahmadalipour, P. Abbaszadeh, and H. Moradkhani, "Leveraging machine learning for predicting flash flood damage in the Southeast US," *Environ. Res. Lett.*, vol. 15, no. 2, 2020, doi: 10.1088/1748-9326/ab6edd.
[30] J. Park, J.-H. Park, J.-S. Choi, J. C. Joo, K. Park, H. C. Yoon, C. Y. Park, W. H. Lee, and T.-Y. Heo, "Ensemble model development for the prediction of a disaster index in water treatment systems," *Water Switz.*, vol. 12, no. 11, pp. 1–19, 2020, doi: 10.3390/w12113195.
[31] S. Saha, S. Shekhar, S. Sadhukhan, and P. Das, "An analytics dashboard visualization for flood decision support system," *J. Vis.*, vol. 21, no. 2, pp. 295–307, Apr. 2018, doi: 10.1007/s12650-017-0453-3.
[32] Y. Zhou, Y. Liu, W. Wu, and N. Li, "Integrated risk assessment of multi-hazards in China," *Nat. Hazards*, vol. 78, no. 1, pp. 257–280, Aug. 2015, doi: 10.1007/s11069-015-1713-y.

[33] R. C. dos Santos Alvalá, M. C. de Assis Dias, S. M. Saito, C. Stenner, C. Franco, P. Amadeu, J. Ribeiro, R. A. Souza de Moraes Santana, and C. A. Nobre, "Mapping characteristics of at-risk population to disasters in the context of Brazilian early warning system," *Int. J. Disaster Risk Reduct.*, vol. 41, p. 101326, Dec. 2019, doi: 10.1016/j.ijdrr.2019.101326.

[34] S. Lee, S. Lee, M.-J. Lee, and H.-S. Jung, "Spatial Assessment of Urban Flood Susceptibility Using Data Mining and Geographic Information System (GIS) Tools," *Sustainability*, vol. 10, no. 3, Art. no. 3, Mar. 2018, doi: 10.3390/su10030648.

[35] H. Id, "Data analysis using GIS and data mining.," p. 10.

[36] Y. Liu, Z. Li, B. Wei, X. Li, and B. Fu, "Seismic vulnerability assessment at urban scale using data mining and GIScience technology: application to Urumqi (China)," *Geomat. Nat. Hazards Risk*, vol. 10, no. 1, pp. 958–985, Jan. 2019, doi: 10.1080/19475705.2018.1524400.

[37] W. Chen, S. Zhang, R. Li, and H. Shahabi, "Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naïve Bayes tree for landslide susceptibility modeling," *Sci. Total Environ.*, vol. 644, pp. 1006–1018, Dec. 2018, doi: 10.1016/j.scitotenv.2018.06.389.

[38] O. Rahmati and H. R. Pourghasemi, "Identification of Critical Flood Prone Areas in Data-Scarce and Ungauged Regions: A Comparison of Three Data Mining Models," *Water Resour. Manag.*, vol. 31, no. 5, pp. 1473–1487, 2017, doi: 10.1007/s11269-017-1589-6.

[39] S. Lee, M.-J. Lee, and H.-S. Jung, "Data Mining Approaches for Landslide Susceptibility Mapping in Umyeonsan, Seoul, South Korea," *Appl. Sci.*, vol. 7, no. 7, Art. no. 7, Jul. 2017, doi: 10.3390/app7070683.

[40] J. Chen, Q. Li, H. Wang, and M. Deng, "A Machine Learning Ensemble Approach Based on Random Forest and Radial Basis Function Neural Network for Risk Evaluation of Regional Flood Disaster: A Case Study of the Yangtze River Delta, China," *Int. J. Environ. Res. Public. Health*, vol. 17, no. 1, Art. no. 1, Jan. 2020, doi: 10.3390/ijerph17010049.

[41] C. Luu, J. von Meding, and M. Mojtahedi, "Analyzing Vietnam's national disaster loss database for flood risk assessment using multiple linear regression-TOPSIS," *Int. J. Disaster Risk Reduct.*, vol. 40, p. 101153, Nov. 2019, doi: 10.1016/j.ijdrr.2019.101153.

[42] S. Smith, V. Pang, K. Liu, M. Kavakli-Thorne, A. Edwards, M. Orgun, and R. Host, "Adoption of data-driven decision making in fire emergency management," presented at the 24th European Conference on Information Systems, ECIS 2016, 2016.

[43] F. F. Balahadia, B. G. Dadiz, R. R. Ramirez, M. Luvett, J. P. Lalata, and A. C. Lagman, "Application of Data Mining Approach for Profiling Fire Incidents Reports of Bureau of Fire and Protection," in *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, Dec. 2019, pp. 713–717. doi: 10.1109/ICCIKE47802.2019.9004420.

[44] A. Asgary, A. Ghaffari, and J. Levy, "Spatial and temporal analyses of structural fire incidents and their causes: A case of Toronto, Canada," *Fire Saf. J.*, vol. 45, no. 1, pp. 44–57, Jan. 2010, doi: 10.1016/j.firesaf.2009.10.002.

[45] X. Liu, Y. Lu, Z. Xia, F. Li, and T. Zhang, "A Data Mining Method for Potential Fire Hazard Analysis of Urban Buildings based on Bayesian Network," in *Proceedings of the 2nd International Conference on Intelligent Information*

Processing, New York, NY, USA, Jul. 2017, pp. 1–6. doi: 10.1145/3144789.3144811.
[46] E. W. Lee, G. Yeoh, M. Cook, and C. Lewis, "Data Mining on Fire Records of New South Wales, Sydney," *Procedia Eng.*, vol. 71, pp. 328–332, 2014, doi: 10.1016/j.proeng.2014.04.047.
[47] Z. Wang, J. Xu, X. He, and Y. Wang, "Analysis of spatiotemporal influence patterns of toxic gas monitoring concentrations in an urban drainage network based on IoT and GIS," *Pattern Recognit. Lett.*, vol. 138, pp. 237–246, 2020, doi: 10.1016/j.patrec.2020.07.022.
[48] "Lisboa-Capital Verde Europeia 2020 | Eurocid." https://eurocid.mne.gov.pt/artigos/lisboa-capital-verde-europeia-2020 (accessed Apr. 21, 2021).
[49] "Atlas Social de Lisboa." https://www.arcgis.com/apps/Cascade/index.html?appid=e63936cfadce405b805d7beded9543f0 (accessed Apr. 21, 2021).
[50] M. do R. G. Jorge, L. Baptista, and J. P. Nunes, "A Dança das Densidades no Contexto do Crescimento Urbano," *Os Espaç. Morfol. Urbana*, pp. 417–426, 2016.
[51] P. M. P. Nunes, "Incêndios em edifícios na cidade de Lisboa," p. 173.
[52] "Lx_Risk - Caracterização Sócio-Urbanística." http://webcache.googleusercontent.com/search?q=cache:9sRRU69yITkJ:lxrisk.cm-lisboa.pt/caract_socio_urb.html&hl=pt-PT&gl=pt&strip=1&vwsrc=0 (accessed Sep. 11, 2021).
[53] T. Santos, M. A. P. P. Esteves, S. Velez, and C. Álvaro, "Reabilitação de Edifícios Devolutos na Cidade de Lisboa (2009-2018)," *Proc. 25th APDR Congr.*, pp. 499–505, 2018.
[54] "Relatório síntese de Caracterização Biofísica de Lisboa." Accessed: May 09, 2021. [Online]. Available: https://www.lisboa.pt/fileadmin/cidade_temas/urbanismo/pdm/EstCarat_RSinteseCaracterizacaoBiofisica.pdf
[55] D. F. R. de Oliveira, "O risco de inundação urbana nas frentes de água de deltas e estuários em cenários de alterações climáticas. A frente ribeirinha de Lisboa," PhD Thesis, ISA, 2013.
[56] "Lx_Risk - Caracterização Geo-Ambiental." http://lxrisk.cm-lisboa.pt/caract_geo_amb.html (accessed Apr. 23, 2021).
[57] "Riscos da Cidade," *MUNICÍPIO de LISBOA*. https://www.lisboa.pt/cidade/seguranca-e-prevencao/protecao-civil/riscos-da-cidade (accessed Apr. 24, 2021).
[58] "Plano Municipal de Emergência de Proteção Civil de Lisboa." Accessed: Apr. 24, 2021. [Online]. Available: https://www.lisboa.pt/fileadmin/cidade_temas/seguranca/documentos/BM_1290_5Suplemento_proposta_330_2018.pdf
[59] "Reorganização do dispositivo de socorro da cidade de Lisboa." Accessed: Apr. 25, 2021. [Online]. Available: https://www.am-lisboa.pt/documentos/1406281283U9kWG7kg4Qo08FT4.PDF
[60] "Anuário do Regimento de Sapadores Bombeiros de Lisboa 2012 by Câmara Municipal de Lisboa - issuu." https://issuu.com/camara_municipal_lisboa/docs/anu__rio_rsb_2012/6 (accessed Apr. 26, 2021).
[61] P. Patrício, T. Lopes, I. Pinheiro, C. Pereira, and C. Bispo, "Regimento de sapadores bombeiros gabinete do comando | Planeamento e gestão," p. 70, 2016.

[62] A. L. P. Martins, “Georreferenciação e Análise das Ocorrências do Regimento de Sapadores Bombeiros na Cidade de Lisboa,” p. 121.
[63] “Na Minha Rua LX,” *Lisboa Inteligente*. https://lisboainteligente.cm-lisboa.pt/lxi-iniciativas/na-minha-rua-lx/ (accessed Aug. 10, 2021).
[64] “Portal do INE.” https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_inst_legislacao&xlang=pt (accessed May 04, 2021).
[65] “IPMA - Serviços.” https://www.ipma.pt/pt/produtoseservicos/index.jsp?page=dados.xml (accessed May 04, 2021).
[66] J. V. Neto, C. B. dos Santos, É. M. Torres, and C. Estrela, “Boxplot: um recurso gráfico para a análise e interpretação de dados quantitativos,” *Rev. Odontológica Bras. Cent.*, vol. 26, no. 76, Art. no. 76, Apr. 2017, doi: 10.36065/robrac.v26i76.1132.
[67] “The climate of the UK and recent trends.” Accessed: Jun. 14, 2021. [Online]. Available: https://ukcip.ouce.ox.ac.uk/wp-content/PDFs/UKCP09_Trends.pdf
[68] R. Nétek, T. Pour, and R. Slezakova, “Implementation of Heat Maps in Geographical Information System – Exploratory Study on Traffic Accident Data,” *Open Geosci.*, vol. 10, pp. 367–384, Aug. 2018, doi: 10.1515/geo-2018-0029.
[69] “Introdução aos Algoritmos de Aprendizagem Supervisionada.” Accessed: Jul. 20, 2021. [Online]. Available: https://fontana.paginas.ufsc.br/files/2018/03/apostila_ML_pt2.pdf
[70] “sklearn.metrics.accuracy_score — scikit-learn 1.0 documentation.” https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html (accessed Sep. 30, 2021).
[71] F. A. J. Serras, “Métodos de aprendizagem automática: um estudo baseado na avaliação e previsão de clientes bancários,” Mar. 2016, Accessed: Jul. 20, 2021. [Online]. Available: https://run.unl.pt/handle/10362/17371
[72] “3.2. Tuning the hyper-parameters of an estimator — scikit-learn 0.24.2 documentation.” https://scikit-learn.org/stable/modules/grid_search.html (accessed Jul. 27, 2021).
[73] “Introduction to Random Forest in Machine Learning | Engineering Education (EngEd) Program | Section.” https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/ (accessed Jul. 20, 2021).
[74] “Machine Learning Random Forest Algorithm - Javatpoint,” *www.javatpoint.com*. https://www.javatpoint.com/machine-learning-random-forest-algorithm (accessed Jul. 20, 2021).
[75] “sklearn.grid_search.RandomizedSearchCV — scikit-learn 0.16.1 documentation.” https://scikit-learn.org/0.16/modules/generated/sklearn.grid_search.RandomizedSearchCV.html (accessed Sep. 23, 2021).
[76] “sklearn.model_selection.RandomizedSearchCV — scikit-learn 0.24.2 documentation.” https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html (accessed Jul. 23, 2021).
[77] R. Gandhi, “Support Vector Machine — Introduction to Machine Learning Algorithms,” *Medium*, Jul. 05, 2018. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 (accessed Jul. 20, 2021).

[78] L. J. Moreira, “Classificação de dados combinando mapas auto-organizáveis com vizinho informativo mais próximo,” Dec. 2016, Accessed: Jul. 20, 2021. [Online]. Available: http://dspace.mackenzie.br/handle/10899/24442
[79] J. L. Puga, M. Krzywinski, and N. Altman, “Bayes’ theorem,” *Nat. Methods*, vol. 12, no. 4, pp. 277–278, Apr. 2015, doi: 10.1038/nmeth.3335.
[80] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, “Comparing different supervised machine learning algorithms for disease prediction,” *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, p. 281, Dec. 2019, doi: 10.1186/s12911-019-1004-8.
[81] L. C. D. Mello, “Um Assistente de Feedback para o Serviço de Filtragem do Software Direto,” p. 115.
[82] K. Jain, “How to Improve Naive Bayes?,” *Analytics Vidhya*, Apr. 03, 2021. https://medium.com/analytics-vidhya/how-to-improve-naive-bayes-9fa698e14cba (accessed Jul. 27, 2021).
[83] D. Dantas and E. A. Donadia, “Comparação entre as técnicas de regressão logística, árvore de decisão, bagging e random forest aplicadas a um estudo de concessão de crédito,” p. 67.
[84] “sklearn.linear_model.LogisticRegression — scikit-learn 0.24.2 documentation.” https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (accessed Jul. 27, 2021).

Annexes and appendix

Annex A – Distribution of occurrences by month in 2011 and 2012

2011		2012	
Month	**occurrence distribution**	**Month**	**occurrence distribution**
1	1	1	191
5	1	2	314
8	123	4	1
9	271	8	2
10	386	10	3
11	299	11	2
12	202	12	1 106
Total	1 282	**Total**	1 619

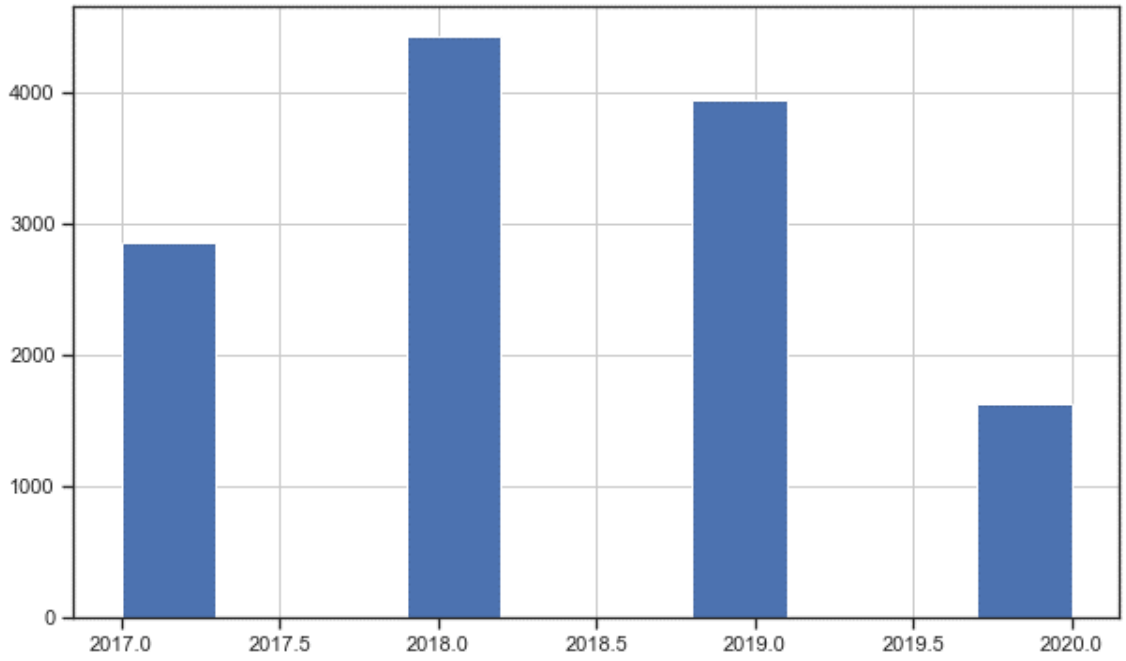
Annex B – Occurrence distribution

Occurrence	Count	Category	%
Fire - Forest stand	47	Fire	8,50%
Fire - Agricultural	3		
Fire - Uncultivated	1486		
Fire - Building (Infrastructures/Installation) - Housing	1468		
Fire - Building (Infrastructures/Installation) – Parking.	32		
Fire - Building (Infrastructures/Installation) - Services	37		
Fire - Building (Infrastructures/Installation) – School	16		
Fire - Building (Infrastructures/Installation) - Hospital/Home	14		
Fire - Building (Infrastructures/Installation) - Performance/Recreation Religious Worship.	9		
Fire - Building (Infrastructures/Installation) – Hotel and similar	126		
Fire - Building (Infrastructures/Installation) - Commercial/Shops/Fairs/Transport Station.	30		
Fire - Building (Infrastructures/Installation) – Culture/Museum/Art/Library	3		
Fire - Building (Infrastructures/Installation) – Military/Security forces.	3		
Fire - Building (Infrastructures/Installation) - Industry/Workshop/Warehouse	45		
Fire - Building (Infrastructures/Installation) - Empty/Degraded Building	141		
Fire - Building (Infrastructures/Installation) – High-rise building (>29 m).	1		
Fire - Equipment (without affecting the environment)	21		
Fire - Equipment (without affecting the environment) - Garbage containers	1379		
Fire - Transport - Road	928		
Fire - Transport - Air	2		
Fire - Transport - Railroad	1		
Fire - Transport - Aquatic	5		
Fire - Debris	2663		
Accidents - Road - Trampling	771	Accidents	10,10%
Accidents - Road - With vehicles	5759		
Accidents - Road - With incarcerated	963		
Accidents - Railroad - Trampling	46		
Accidents - Railroad - Collision	3		
Accidents - Railroad - Shock	2		
Accidents - Railroad - Derailment	2		
Accidents - Railroad - With incarcerated	6		
Accidents - Aquatic	2		

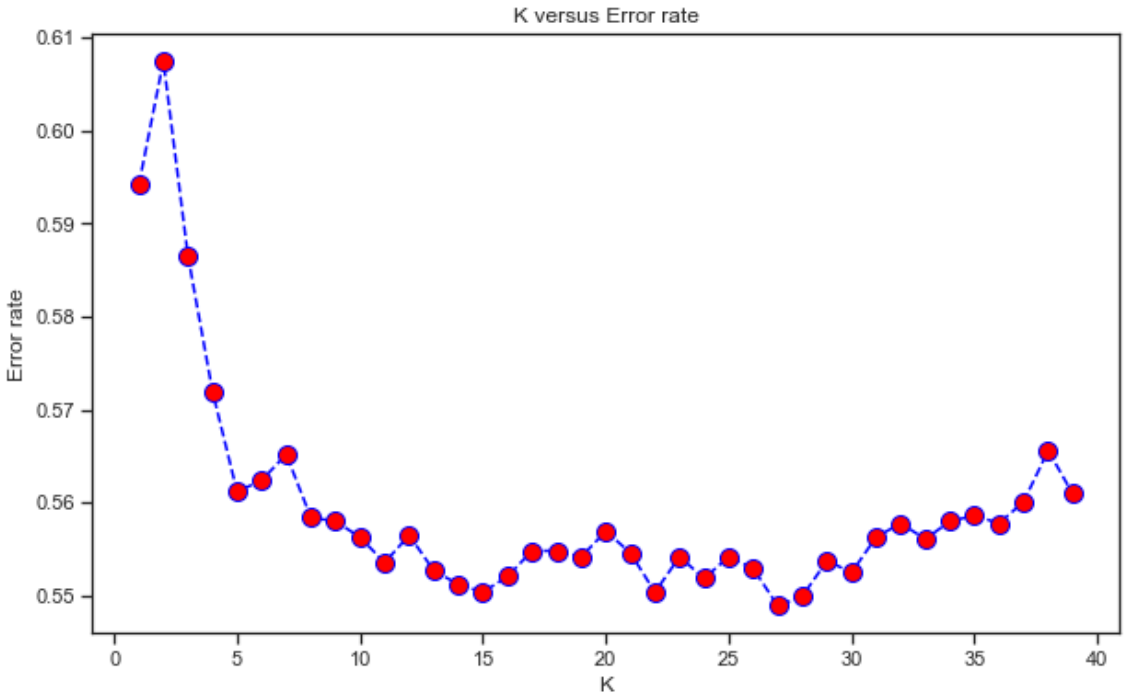
Accidents - Aquatic - Falls into a river	41		
Accidents - Equipment	18		
Accidents - Equipment - Lifts	2444		
Accidents - Equipment - Escalators	1		
Infrastructures and Communication Routes - Fall of Tree	2470	Infrastructures and Communication Routes	14,70%
Infrastructures and Communication Routes - Cutting off the power supply - Electricity	3		
Infrastructures and Communication Routes – Collapse	86		
Infrastructures and Communication Routes - Collapse (Coating fall)	3766		
Infrastructures and Communication Routes – Landslide	17		
Infrastructures and Communication Routes – Floods	3008		
Infrastructures and Communication Routes - Private space flood	546		
Infrastructures and Communication Routes - Public space flood	233		
Infrastructures and Communication Routes - Unclogging	86		
Infrastructure and Communication Routes - Damage/Fall of electric cables	347		
Infrastructures and Communication Routes - Damage/Fall electric cables - Short-circuit	1426		
Infrastructures and Communication Routes - Fall of structures	2700		
Pre-Hospital - Intoxication	41	Pre-hospital	9,20%
Pre-Hospital - Sudden Illness	7388		
Pre-Hospital - Trauma/Fall	1690		
Pre-Hospital - Burn	6		
Pre-Hospital - Childbirth	41		
Legal Conflicts - Explosives - Threat	4	Legal conflicts	0,50%
Legal Conflicts - Explosives - Explosion	7		
Legal Conflicts - Assault/Violation	56		
Legal Conflicts - Suicide/Homicide - Attempted	106		
Legal Conflicts - Suicide/Homicide - Consummated	5		
Legal Conflicts - Authority support	363		
Industrial-technological - Accidents with hazardous materials - Chemical	24	Industrial-technological	5,10%
Industrial-technological - Hazardous Materials in Transit - Chemical	1		
Industrial–technological - Gas Leak - Plumbing/Conduct	1142		
Industrial–technological - Gas Leak - Bottle	137		
Industrial–technological - Gas Leak - Deposit/Reservoir	7		
Industrial–technological - Suspicious Situations - Check smoke	1631		
Industrial–technological - Suspicious Situations - Check smells	1545		

Industrial-technology - Suspicious Situations - Check alarms	683		
Services - Prevention - Patrol/Surveillance	553	Services	45,60%
Services - Preventions - Entertainment	215		
Services - Preventions - Sports	173		
Services - Preventions - Bushfires	9		
Services - Preventions - Transportation	3		
Services - Prevention - Pre-Positioning means	362		
Services - Track Clearance/Conservation	2560		
Services - Track Clearance/Conservation - Pothole marking	1204		
Services - Track Clearance/Conservation - Oil on the road	3933		
Services - Water Supply - Population	8		
Services - Water Supply - Public entity	160		
Services - Water Supply - Private entity	58		
Services - Door Openning - With Assistance	9956		
Services - Door Openning - Without Assistance	7783		
Services - Water closure	15352		
Services - Towing	58		
Services - Patient Transportation - General	17		
Services - Patient Transportation - Inter-Hospital	2		
Services - Patient Transportation - Aid for patient transportation	1519		
Services - Animal Rescue	1517		
Activities - Search/Rescue (people and animals) - Land	118	Activities	5,90%
Activities - Search/Rescue (people and animals) - Aquatic	32		
Activities - Exercise/Simulacrum	232		
Activities - Displacement - General service	176		
Activities - Assistance to the population/Social support	4741		
Activities - Hospital Accompaniment	222		
Activities - Follow-up consults	85		
Activities - Follow-up returns	50		
Activities - Social support monitoring	10		
Activities - Social support evaluation	122		
Activities - Social support - Signaling	10		
Activities - Teleassistance - Membership	39		
Activities - Teleassistance for breakdown maintenance	46		
Activities - Teleassistance for equipment collection	10		
Activities - Teleassistance for documentationeExpedient	13		
Activities - Meetings - Events	6		
Civil Protection Events - Technical visit	4	Civil Protection Events	0,004%

Annex C – Occurrence distribution per year



Annex D – K versus Error Rate



Appendix A - Proportion of buildings in need of major repairs or very dilapidated (%). Source INE[5]

Geographic location	Proportion of buildings in need of major repairs or very dilapidated (%)
	2011
Ajuda	7,65
Alcântara	5,12
Alto do Pina	6,63
Alvalade	2,84
Ameixoeira	3,91
Anjos	12,09
Beato	13,10
Benfica	1,64
Campo Grande	5,16
Campolide	11,08
Carnide	12,97
Castelo	10,78
Charneca	15,81
Coração de Jesus	4,46
Encarnação	9,26
Graça	16,82
Lapa	3,95
Lumiar	2,20
Madalena	8,86
Mártires	17,19
Marvila	7,28
Mercês	9,29
Nossa Senhora de Fátima	4,35
Pena	19,73
Penha de França	4,80
Prazeres	6,51
Sacramento	9,70
Santa Catarina	13,82
Santa Engrácia	6,95
Santa Isabel	11,91
Santa Justa	67,36
Santa Maria de Belém	2,40
Santa Maria dos Olivais	3,17
Santiago	12,39
Santo Condestável	9,53
Santo Estêvão	8,07
Santos-o-Velho	8,27
São Cristóvão e São Lourenço	15,67
São Domingos de Benfica	3,14
São Francisco Xavier	0,67
São João	9,13
São João de Brito	1,47
São João de Deus	4,93
São Jorge de Arroios	6,43
São José	10,34
São Mamede	1,35
São Miguel	8,71
São Nicolau	17,95
São Paulo	13,89
São Sebastião da Pedreira	3,23
São Vicente de Fora	10,59
Sé	27,78
Socorro	17,12

[5] Data extracted and adapted from the INE website on April 27, 2021

Appendix B - List of special departments and nucleus. Source: yearbook of the LFBR for the year 2012[60]

Team	Mission
Disaster Intervention Detachment	Operate in disaster missions or severe accidents. The elements of this unit must be able to reach the scene of the occurrence in 32 hours and operate autonomously for 7 days.
Cynotechnical Rescue Unit	Search and rescue of people in collapsed structures due to human action or natural causes and search in large areas (forest or rural areas)
Divers Unit	As the city of Lisbon has an extended coastline and a busy riverfront, the team of divers performs search and rescue work for people, animals and property in the Tagus River estuary
Environmental Control Unit	Control and develop actions with the purpose of minimizing the impact of accidents involving dangerous materials. The elements of this unit have specific training and act in areas such as Recognition and intervention; Material; Decontamination; Protection.
Pre-hospital Emergency Unit	Ensure the management of the medical emergency activity through the regularization of the ambulance services , stock management of medical supplies, and distribution and replenishment of the first aid bags in the first line vehicles and in the Regiment's headquarters.
Social Intervention Center for Citizen Support	Provide support to the vulnerable population, through personal assistance, screening, analysis, and referral to the competent authorities, institutions, or organizations.