Data-Driven Disaster Management in a Smart City

Sandra P. Gonçalves[1], João C. Ferreira[1], and Ana Madureira[2]

[1] ISTAR, ISCTE-Instituto Universitário de Lisboa, 1649-026 Lisboa, Portugal
[2] 2ISRC, ISEP/P.PORTO, Porto, Portugal

Abstract. Disasters, both natural and man-made, are extreme and complex events with consequences that translate into a loss of life and/or destruction of properties. The advances in IT and Big Data analysis represent an opportunity for the development of resilient environments once the application of analytical methods allows extracting information from a significant amount of data, optimizing the decision-making processes. This research aims to apply the CRISP-DM methodology to extract information about incidents that occurred in the city of Lisbon with emphasis on occurrences that affected buildings, constituting a tool to assist in the management of the city. Through this research, it was verified that there are temporal and spatial patterns of occurrences that affected the city of Lisbon, with some types of occurrences having a higher incidence in certain periods of the year, such as floods and collapses that occur when there are high levels of precipitation. On the other hand, it was verified that the downtown area of the city is the area most affected by occurrences. Finally, machine learning models were applied to the data and the predictive model Random Forest obtained the best result with an accuracy of 58%.

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

1 Introductions

Disasters, both natural and man-made, have been occurring more frequently around the world with damaging consequences that are reflected in the loss of human life and material/facilities damage [1]. In fact, in the last ten years, 3,751 natural disasters such as earthquakes, tsunamis, and floods were detected worldwide, representing total damages of $1,658 billion and impacting more than 2 billion people [2]. In this way, it becomes crucial to implement disaster management techniques to minimize the risks associated.

Disaster management can be characterized as a multifaceted process where the primary goals is to avoid, reduce, respond, and recover from disaster impact in 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 [3]. 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 [4].

The increase in population density in cities and the increase in the frequency of disasters in recent years arise 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 [5]. Cities that aim to become a SC 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 [6].

In a SC, electronic devices and network infrastructures are incorporated to obtain high-quality services and as cities get the latest network infrastructure, smart devices, and sensors, a substantial amount of data is generated, known as Big Data (BD). This data can contain large amounts of information that can be contextual, spatial, or temporal [7].

In the case of disaster situations, BD plays an important role in disaster management processes since it is possible to apply data mining (DM) and analysis techniques to analyze patterns and predict disasters, allowing the development of appropriate disaster management strategies from the data collected that have occurred in the past [6].

In this way, 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 cope with disaster situations, thus increase the resilience of the SC [2].

This research aims to apply a data-driven approach to extract information about disasters in the context of a SC to contribute to improving the way the city is managed. 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: data regarding occurrences registered by firefighters between the years 2011 and 2018 both descriptive and predictive analysis are going to be carried out. The second dataset, where only a descriptive analysis of the data is going to be conducted, comes from the application "*Na Minha Rua Lx*" [8], which is an application for intervention request management in the city of Lisbon. In both cases, the analysis was conducted in two phases, where in a first moment a general analysis of the reported occurrences was carried out and in a second moment the analysis focused on occurrences that affected buildings in the city

2 State of the art

Data-driven disaster management is a recent area that has been undergoing an evolution due to the number of works that have been developed [9]. In this sense, a 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 [10] in accordance with the Systematic Literature Review steps proposed by Okoli and Schabram [11].

Accordingly, a systematic search on the topic was conducted in two electronic databases: Scopus [12] and Google Scholar [13], and the main objective was to identify and select research papers related to data-driven disaster management research area. 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 research 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. 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. The work done in this area covers both natural and man-made disasters.

2.1 Natural disasters

Natural disasters are events that are characterized by the substantial impact they cause on society, interrupting its normal functioning. 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 [14] developed a research where they conducted a statistical analysis of electrical incidents such as fires or failures occurring in Korea caused by climate changes. This study established a relationship between climate change and accidents involving electrical equipment.

Another study [15] conducted in 2017, reflected the link between BD systems and disaster management where Big Data Analytics technologies was implemented on a dataset from the National Hydraulic Research Institute of Malaysia in order to analyze the hydroclimate data to extract insights on climate change and thus provide information to prepare, mitigate, respond and recover from natural disasters. The applications of BD technologies allowed detecting periods of extreme precipitation and runoff projection 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 [16] to identify and analyze 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.

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, as the case of the analysis carried out by the authors Alipour Ahmadalipour, Abbaszadeh, and Moradkhani [17] where 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.

With a similar approach, Park et al. [18] 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.

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, Shekhar, and Sadhukhan [19], 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 studies [20], [21], [22], [23] 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 that different DM models are used in the different research works.

2.2 Man-made disasters

Regarding man-made disasters, Smith et al. [24] developed a research that consisted in the implementation of Big Data technologies for disaster management where they used the statistical tool R, as well as its visualization capabilities, to analyze 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. [25] 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 [26] an attempt was made to use spatiotemporal methods to analyze the spatial and temporal patterns of fire-related incidents in Toronto, Canada. Insights were extracted by analyzing the relationship between the economic, physical, and environmental aspects of various neighborhoods and the total number of fires that occurred in those neighborhoods.

In the study [27] was proposed a DM method based on using Bayesian Network to model building fires in urban areas. From the historical records of fires in a city in china between 2014 and 2016, they analyzed the potential fire risk according to building construction characteristics and external influences. Another study aiming to analyze fire patterns was conducted by Lee et al. [28] by applying the Support Vector Machine model to analyze the correlation between building characteristics, occupants, and fire incidents in Sydney.

Finally, in a study developed by Wan, Xu, He, and Wang [29] BD technologies were applied to analyze 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 researches developed in this area were in China and 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 the analysis of fire-related incidents.

3 Methodology

The analysis carried out in this research has two distinct focuses that serve the same purpose, i.e., spatial-temporal analysis of occurrences recorded in Lisbon to extract knowledge about the circumstances in which they occur. The Cross-Industry Standard Process for Data Mining (CRISP-DM) [30] methodology was applied separately on both datasets namely, the firefighters' dataset and data extracted from the application "*Na Minha Rua Lx*" to extract insights about disasters that affect the city of Lisbon with emphasis on buildings.

The analysis process based on the CRISP-DM methodology began with the business understanding that allows contextualizing and understanding the scope of the project. In this sense, an assessment of the business problem was made through the analysis of the aspects that characterize the city of Lisbon from different perspectives such as demographic, climatic, and edification aspects.

After the business understanding step was completed the next phases consisted of data understanding, data preparation, modeling, and evaluation that were individually applied on the two datasets.

3.1 Firefighters dataset

To work on the firefighter's dataset provided the Lisbon City Hall, the dataset was first loaded. It is a CSV file that 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 (number of people) and material resources (number of vehicles) allocated to each occurrence. The dataset contains data from 2011 to 2018 and contains 135 200 rows and 22 columns. All columns are of type "object", and 13 columns have null values.

During the data preparation phase, it was found that the years 2011 and 2012 have significantly less data when compared to the other years and, in order to perform an analysis where all years have representative data, it was necessary to eliminate these years. Also in this phase, cleansing techniques were applied that included column format conversion, the selection of relevant variables for the analysis where variables that did not add value to the scope of this research were eliminated. The null values were eliminated since those columns did not allow the replacement of the null values by the mean or median as they are geographic coordinates, parishes, and descriptions of the occurrences. This phase also included tasks such as building new variables from existing variables and adding variables from external sources. This external data contains information that characterizes the city of Lisbon in terms of population, building characteristics such as average age of buildings per parish, the proportion of buildings in need of major repairs or very degraded per parish, and meteorological conditions of the city such as average air temperature, relative humidity, average wind speed, and precipitation.

Lastly, 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" column and this variable has 25 types of occurrences that were defined by the firefighters´ occurrence management system. 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, and Ind. technol. - Suspicious situations (check smoke or check smells).

With the data preparation phase complete, 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. The first analysis is focused on understanding the distribution of the data over the years it was possible to conclude that between the period of 2013 and 2018 there was a downward trend in the number of occurrences recorded in the firefighters occurrence management system, however, this decrease was not linear as there were oscillations over the years as there were 17 176 occurrences registered in the year 2013, 17 607 occurrences in 2014, 16 717 occurrences in 2015, 15 089 occurrences in 2016, 17 582 occurrences in 2017, and 13 368 occurrences in 2018.

Firefighters respond to many different types of occurrences comprising several areas of action. For a better understanding of the activities performed by firefighters, the types of occurrences that firefighters respond to in their daily work were analyzed and it was verified that the distribution is not balanced among the nine categories of occurrences recorded in the dataset, as there is an over-position of one category, namely the Services category, which represents 45.6% 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, landslides, 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 protection events that represent 0.004% of the occurrences registered.

After a general analysis of the type of occurrences, the analysis is focused on the analysis of the occurrences that took place in the buildings of the city of Lisbon to characterize them spatially and temporally.

As shown in Fig. 1, 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 and 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.

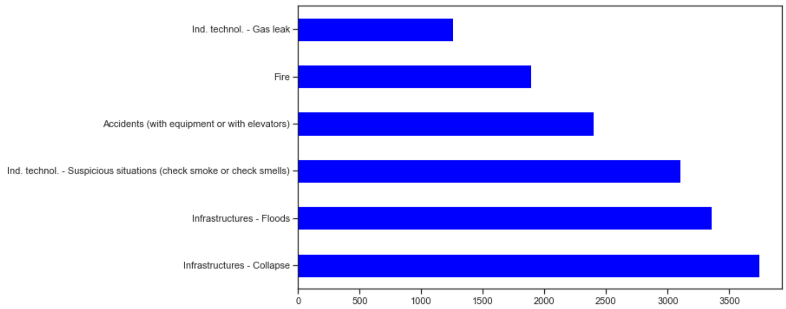

Fig. 1. Distribution of occurrences according to categories

When these occurrences are analyzed over time, i.e., their distribution over the years (Fig. 2), 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.

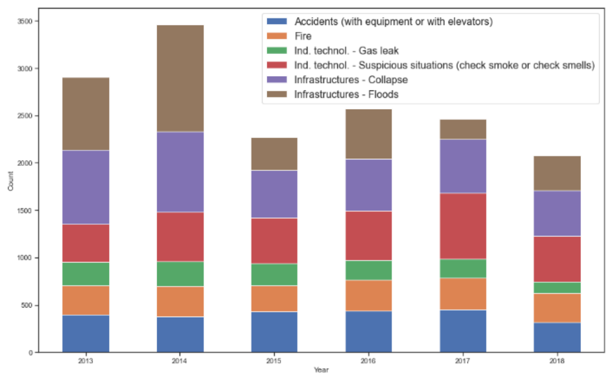

Fig. 2. Distribution of occurrences per year

Focusing the analysis on each occurrence to extract insights about its pattern of occurrence over the 12 months of the year, it is possible to verify that in the case of the occurrences referring to the infrastructure categories, i.e., collapses and floods represented in Fig.3, that in the case of collapses (A), these 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 (B), there is a higher incidence in the winter months with the highest values in the months of October to December, while during the summer months these values are much lower when compared to the winter months. Cases of suspicious situations (C) occur, similar to the types described above, more frequently in the winter months, especially in December. 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.

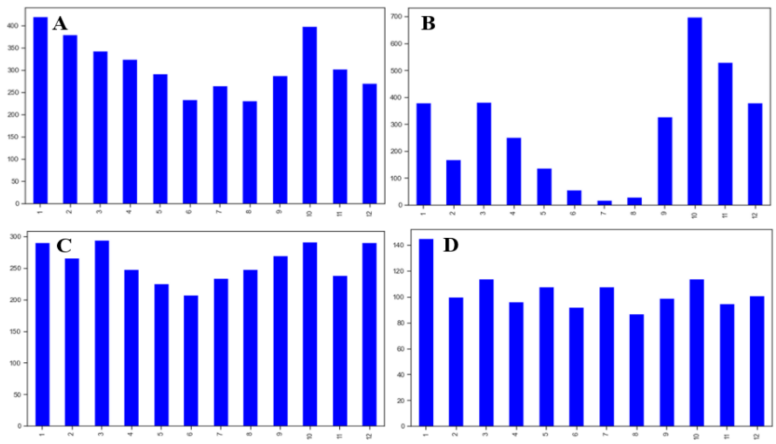

Fig. 3. Temporal distribution of the occurrences. The bar chart from figure **A** shows the temporal distribution of Collapses, the bar chart from figure **B** shows the temporal distribution of Floods, the bar chart from figure **C** shows the temporal distribution of Suspicious situations (check smoke 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 Fig. 4. 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 occur mainly in the last month of the year, and these observations may be due to the fireplaces and candles that are used in greater density at this time of year.

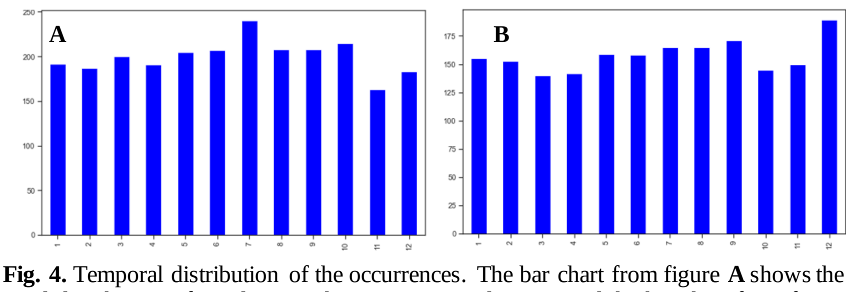

Fig. 4. Temporal distribution of the occurrences. The bar chart from figure **A** shows the temporal distribution of accidents with equipment or elevators and the bar chart from figure **B** shows the temporal distribution of Fires.

The first analysis showed that there are types of occurrences that have a higher incidence in certain seasons of the year, such as collapses, floods, suspicious situations (checking for smoke or smells) occurring with a higher incidence in the winter/spring months, the influence of weather conditions on the incidence of different types of events affecting the city of Lisbon has been verified. With this in mind, the influence of precipitation on the different types of occurrences data was analyzed through 4 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 interquartile ranges it was possible to build 4 datasets with the four precipitation levels previously mentioned.

From the analysis of occurrences according to the four precipitation levels, it was 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 cities of Lisbon, heatmaps were created for the six types of occurrences that most affect buildings in the city of Lisbon. Fig. 5 shows the spatial distribution of collapses and floods.

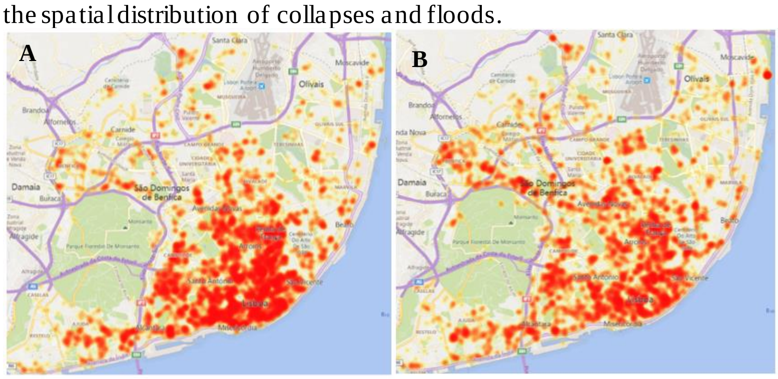

Fig. 5. Spatial distribution of the occurrences. Figure **A** shows the temporal distribution of Collapses and figure **B** shows spatial distribution of Floods,

From the heatmaps presented above it is possible to infer that the occurrences related to collapses, 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 in the central area of the city, 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, 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*.

Fig. 6 refer to the geospatial distribution of occurrence regarding Suspicious situations (check smoke or check smells) and Gas Leaks. The suspicious situations 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 are more concentrated in the Lisbon Historical Center area and the districts of *Penha de França*, *Arroios,* and *Benfica*.

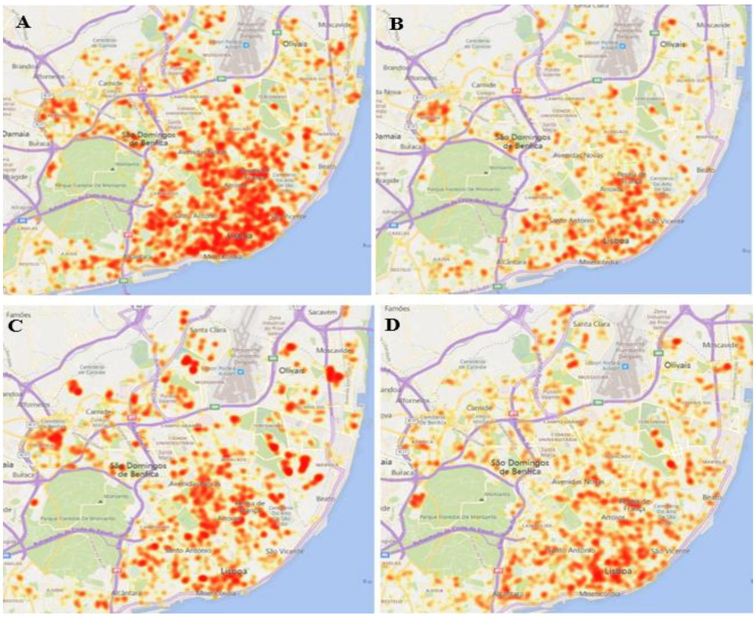

Fig. 6. Spatial distribution of the occurrences. Figure **A** shows the spatial distribution of Suspicious situations (check smoke or check smells), figure **B** shows the spatial distribution of Gas leaks, figure **C** shows the spatial distribution of accidents with equipment or elevators, and figure **D** shows the spatial distribution of Fires.

In terms of accidents involving equipment or elevators (C), 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 (D) 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.

Since there is a concentration of occurrences in a specific area of the city of Lisbon, it was sought to deepen the knowledge about Lisbon by analyzing aspects such as the state of conservation of buildings and the average age of buildings in the different parishes. 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.

Fig. 7. Figure **A** shows the spatial representation of the proportion of buildings that are degraded or in need of major repairs and figure **B** shows the spatial representation of the average age of the buildings per parish

From the conclusions reiterated from the two heatmaps presented in Fig. 7, it is possible to infer that it is in the locations where the older buildings are located and where there is a greater proportion of degraded buildings with major needs of repair that they are more affected by the types of occurrences such as collapses, floods, suspicious situations (check smoke or check smells), and gas leaks, occurring in greater proportion in these areas of the city.

3.2 *Na Minha Rua Lx* dataset

Regarding the dataset containing the data related to the occurrences recorded on *the Na Minha Rua Lx* application, the first analysis showed that the dataset covers the period between 2017 and 2020 and it is composed of 12 866 rows and eight 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. It was verified that there was one duplicate value and zero null values in the entire dataset. Furthermore, six of the eight columns that compose the dataset are of type *object*, except the columns "latitude" and "longitude" that are of type *float*.

The data preparation began with the selection of the relevant variables and after identifying the necessary variables to conduct this analysis, data processing techniques were applied to adequate the dataset to the analysis intended to be developed. However, no significant problems were identified, besides one variable that was not in date format and a duplicate value. Also, in the data preparation phase, new columns were created from an existing column since the information that allows locating the occurrences on a temporal level was extracted from the column "Date-Time". In this way, three new attributes were created from the "Date-Time" variable: "Year", "Month", and "Hour". Lastly, it was found that the year 2020 has significantly less data when compared to the other years and, to conduct an analysis where all years have representative data, it was necessary to eliminate the year 2020. With all the transformations on the dataset completed, the final dataset has seven columns and 12 865 rows.

With the data preparation phase completed, the modeling phase begins. 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.

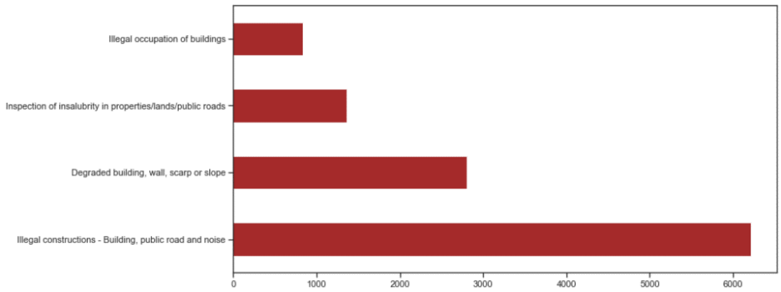

Fig. 8. Types of occurrences reported in the application

From Fig. 8 it is possible to verify that there are four main types of occurrences reported in the application and in terms of their distributions, it is noted that the occurrences are not equally distributed in the dataset since there is an over-position of the type of occurrence Illegal constructions in relation to the other types since during the period under analysis 6 217cases of Illegal constructions - Building, public roads, and noise were reported, while 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 was concluded 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.

After a general description of the distribution of reports over the years and an analysis of the types of occurrences reported, the analysis focused only on events that took place in the buildings of the city of Lisbon i.e, degraded building, wall, scarp or slope and illegal occupation of buildings.

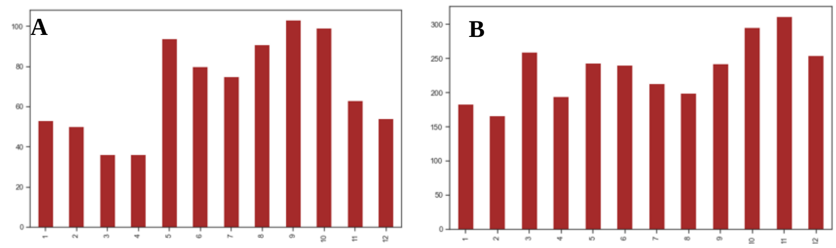

Fig. 9. Temporal distribution of the occurrences. The bar chart from figure **A** shows the temporal distribution of illegal occupation of buildings and the bar chart from figure **B** shows the temporal distribution of degraded buildings, wall, scarp, or slope.

With the purpose of deepening the knowledge about these two types of occurrence, their temporal distribution was analyzed and for cases of illegal occupation of buildings are reported in greater expression between the months of September and October, with emphasis on the month of May where there is an increase in this type of occurrence. On the other hands, cases of degraded buildings, wall, scarp, or slope, there is an increase in reports in the last four months of the year, i.e., from September to October, and then an increase again in March and May.

Shifting the focus to an analysis of occurrences from a spatial perspective, heatmaps were created that present the geospatial distribution of the above-mentioned types of occurrences with the goal of verifying how these occurrences are distributed throughout the city. From the heatmaps presented in Fig. 10, it is possible to verify that in both cases these events are registered with a higher incidence of the downtown area of the city.

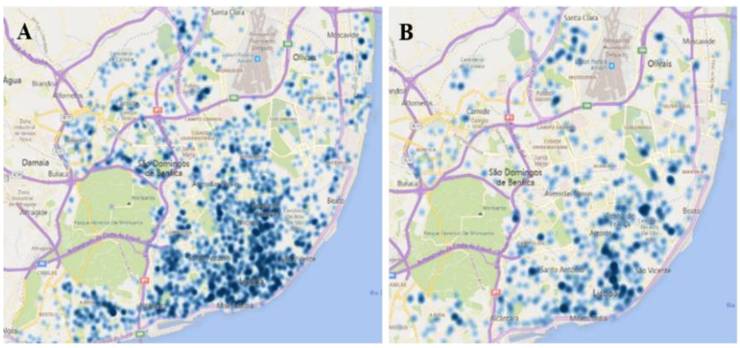

Fig. 10. Spatial distribution of the occurrences. Figure **A** shows the spatial distribution of degraded buildings, wall, scarp, or slope and figure **B** shows the spatial distribution of illegal occupation of buildings.

Cases regarding Degraded building, wall, scarp or slope 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*. On the other hand, the occurrences related to Illegal occupation of buildings have a higher incidence in the parishes of *Arroios*, *São Vicente*, *Santo António*, *Penha de França*, *Misericórdia*, and the historical downtown area.

3.3 Prediction process

In this phase, predictive models were applied to predict disasters. Since this is a classification problem and the variable that is intended to be predicted is a categorical variable ("Occurrence Type"), supervised classification algorithms [68] 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 Fig. 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 (dependent variables): "PRCP", "RH" "Num of people", "Num of vehicles", "Avg building age", Avg air temp", "Resident pop", and Avg WS".

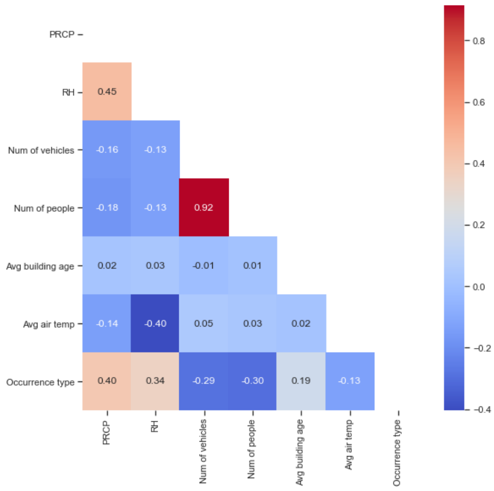

Fig. 11. Correlation matrix

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 dependent and independent 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.

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. It is important to emphasize that all algorithms were applied twice, wherein a first moment they were applied without hyperparameters and in a second moment, with the purpose of improving the performance of these algorithms, functions that present the best set of hyperparameters were applied and the algorithms were reapplied considering the hyperparameters. The graph in Fig. 12 shows the summary of the predictive results for each model after tuning.

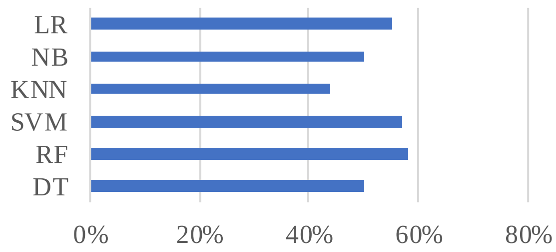

Fig. 12. Summary of the predictive results for each model

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 Conclusion

In short, through the application of DM techniques to the firefighters´ dataset, it was possible to conclude that the buildings in the city of Lisbon are affected by six types of events namely collapses, floods, suspicious situations (check smoke or check smells), gas leak, fires, accidents (with equipment or with elevators). It was also verified that there is a temporal pattern with regard to these occurrences since in some cases there is a greater predominance of certain occurrences at certain times of the year. In terms of the distribution of the occurrences, it was concluded that the historic center of the city is, in general, the area most affected by these types of occurrences and it is in this area where are concentrated the degraded buildings or with a great need for repair and also the older buildings.

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.

References

[1] 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.

[2] 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.

[3] 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.

[4] «(PDF) Crowdsourcing Disaster Response». https://www.researchgate.net/publication/268448750_Crowdsourcing_Disaster_Response (last accessed Aug. 26, 2021).

[5] S. A. Shah *et al.*, «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.

[6] 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.

[7] 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.

[8] «Na Minha Rua LX - Lisboa Inteligente». https://lisboainteligente.cm-lisboa.pt/lxi-iniciativas/na-minha-rua-lx/ (last accessed Aug. 10, 2021).

[9] T. Li *et al.*, «Data-Driven Techniques in Disaster Information Management», *ACM Comput. Surv.*, vol. 50, n. 1, p. 1:1-1:45, Mar. 2017, doi: 10.1145/3017678.

[10] 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, n. 5, pp. 336–341, 2010, doi: 10.1016/j.ijsu.2010.02.007.

[11] 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.

[12] «Scopus - Document search». https://www.scopus.com/search/form.uri?display=basic&edit.scft=1#basic (last accessed Aug. 09, 2021).

[13] «About Google Scholar». https://scholar.google.com/intl/en/scholar/about.html (last accessed Aug. 09, 2021).

[14] 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, n. 16, 2019, doi: 10.3390/ijerph16162984.

[15] 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.

[16] K. M. Briones-Estébanez and N. F. F. Ebecken, «Occurrence of emergencies and disaster analysis according to precipitation amount», *Nat. Hazards*, vol. 85, n. 3, pp. 1437–1459, 2017, doi: 10.1007/s11069-016-2635-z.

[17] 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, n. 2, 2020, doi: 10.1088/1748-9326/ab6edd.

[18] J. Park *et al.*, «Ensemble model development for the prediction of a disaster index in water treatment systems», *Water Switz.*, vol. 12, n. 11, pp. 1–19, 2020, doi: 10.3390/w12113195.

[19] S. Saha, S. Shekhar, S. Sadhukhan, and P. Das, «An analytics dashboard visualization for flood decision support system», *J. Vis.*, vol. 21, n. 2, pp. 295–307, Abr. 2018, doi: 10.1007/s12650-017-0453-3.

[20] R. C. dos Santos Alvalá *et al.*, «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, Dez. 2019, doi: 10.1016/j.ijdrr.2019.101326.

[21] 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, n. 3, Art. n. 3, Mar. 2018, doi: 10.3390/su10030648.

[22] 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, n. 1, pp. 958–985, Jan. 2019, doi: 10.1080/19475705.2018.1524400.

[23] 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, Dez. 2018, doi: 10.1016/j.scitotenv.2018.06.389.

[24] S. Smith *et al.*, «Adoption of data-driven decision making in fire emergency management», presented at the 24th European Conference on Information Systems, ECIS 2016, 2016.

[25] F. F. Balahadia *et al.*, «Application of Data Mining Approach for Profiling Fire Incidents Reports of Bureau of Fire and Protection», em *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, Dez. 2019, pp. 713–717. doi: 10.1109/ICCIKE47802.2019.9004420.

[26] A. Asgary, A. Ghaffari, e J. Levy, «Spatial and temporal analyses of structural fire incidents and their causes: A case of Toronto, Canada», *Fire Saf. J.*, vol. 45, n. 1, pp. 44–57, Jan. 2010, doi: 10.1016/j.firesaf.2009.10.002.

[27] 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.

[28] 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.

[29] 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.

[30] «[PDF] Crisp-dm: towards a standard process modell for data mining | Semantic Scholar». https://www.semanticscholar.org/paper/Crisp-dm%3A-towards-a-standard-process-modell-for-Wirth-Hipp/48b9293cfd4297f855867ca278f7069abc6a9c24 (last accessed Aug. 27, 2021).