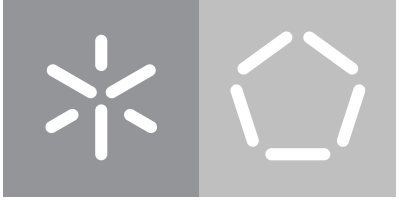
Universidade do Minho
Escola de Engenharia

Rui Nuno Vilaça Ribeiro

A Data Science Approach to Portuguese Road Accidents' Data

January, 2022



Universidade do Minho
Escola de Engenharia

Rui Nuno Vilaça Ribeiro

A Data Science Approach to Portuguese Road Accidents' Data

Master's Dissertation
Integrated Master's in Informatics Engineering

Work supervised by
Cesar Analide
Bruno Fernandes

January, 2022

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the Universidade do Minho.

_____, _____

(Location) (Date)

Assinado por : **Rui Nuno Vilaça Ribeiro**
Num. de Identificação: 14077496
Data: 2022.01.10 15:27:54+00'00'

_____ _____

(Rui Nuno Vilaça Ribeiro)

Acknowledgements

First and foremost, I'd like to thank professor Cesar Analide for this opportunity and professor Bruno Fernandes who helped me remotely throughout this tough pandemic period with suggestions, corrections and reviews for this work.

I'd also like to thank my all my friends that supported me, encouraged me, listened to my concerns and helped in times of need.

Finally, I want to thank my family specially my mother, father and brother who put up with me during the pandemic where we had to stay home most of the time.

A special mention to my father who is dealing and fighting a very complicated condition and never left my mind while working on this dissertation.

Abstract

We frequently hear about accidents and traffic news on television, radio and even social networks. Even though we have witnessed a decrease in mortality rate in Portuguese roads, the number of road victims have been increasing recently so we should be more aware of this problem, study it and come up with solutions to decrease the mortality rate and the number of victims in Portuguese roads. One possible solution to this problem is the identification of blackspots (areas with a high number of accidents or an abnormal number of fatalities) associated with temporal and spatial analysis, and relations between them. By doing this, we will be closer to decreasing accidents as well as the mortality rate on Portuguese roads. This dissertation is going to focus on these concerns using the information present on ANSR (*Autoridade Nacional de Segurança Rodoviária*) reports as well as other data gathered by the research team regarding road traffic incidents in Portuguese cities. After researching about the state of the art, we realize that, on one hand, there's a big problem which is traffic accidents and resultant victims that are still to this day very concerning to society, on the other hand, many techniques and methods have been developed and improved to help mitigate this problem. The data have shown that Portugal still has work to do on decreasing the number of accidents and victims according to those evolution curves, data collected in ANSR reports and the comparison between traffic numbers in EU countries. This dissertation focused on understanding, processing and exploring data in-depth, developing models to analyze data, preventing accidents and enhancing road safety and coming up with useful insights about the road network and publishing them in a dashboard platform open to the community.

Keywords: Accident Reports, Data Science, Road Accidents, Road Safety, Visual Analytics

Resumo

Frequentemente, ouvimos falar de acidentes e notícias sobre trânsito na televisão, rádio e redes sociais. Apesar de estarmos a testemunhar um decréscimo da taxa de mortalidade em estradas portuguesas, o número de vítimas resultantes de acidentes têm vindo a aumentar recentemente, por isso, devemos estar mais atentos a este problema, estudá-lo e arranjar soluções para diminuir a taxa de mortalidade e o número de pessoas vítimas de acidentes em estradas portuguesas. Uma possível solução para este problema é a identificação de zonas negras (zonas com um número elevado de acidentes ou um número anormal de óbitos) associado a uma análise temporal e espacial, juntamente com as relações entre eles. Ao fazer isto, estaremos mais perto de diminuir o número de acidentes, bem como a taxa de mortalidade nas estradas portuguesas. Esta dissertação irá focar-se nestes aspetos, utilizando a informação presente no relatórios da ANSR (Autoridade Nacional de Segurança Rodoviária) e também outros dados recolhidos pela equipa de investigação relativamente a incidentes rodoviários em estradas portuguesas. Depois de recolher dados sobre o estado de arte, percebemos que, por um lado, existe um grande problema com os acidentes rodoviários e vítimas dos mesmos que são até ao dia de hoje muito preocupantes para a sociedade, por outro lado, muitas técnicas e métodos que têm vindo a ser desenvolvidos e melhorados para ajudar a mitigar este problema. Os dados mostram que Portugal ainda tem trabalho a fazer para diminuir os números de acidentes e de vítimas tendo em consideração as curvas de evolução destes indicadores, dados recolhidos em relatórios da ANSR e a comparação entre dados rodoviários entre países da UE. Esta dissertação focou-se em perceber, processar e explorar os dados a fundo, desenvolver modelos para analisar os dados, prevenir acidentes e aumentar a segurança rodoviária e encontrar perceções sobre a rede rodoviária e publicá-las numa plataforma com painéis de informação disponíveis para a comunidade.

Palavras-chave: Acidentes Rodoviários, Análise Visual, Ciência de Dados, Relatórios de acidentes, Segurança Rodoviária

Contents

List of Figures

List of Tables

Acronyms

A3C Asynchronous Advantage Actor-Critic Algorithm

AI Artificial Intelligence

ANN Artificial Neural Network

BI Business Intelligence

CRISP-DM Cross-Industry Standard Process for Data Mining

DIC Deviance Information Criteria

DNN Deep Neural Network

DQN Deep Q Network

EB Empirical Bayesian

ENSR *Estratégia Nacional de Segurança Rodoviária*

ERSO European Road Safety Observatory

EU European Union

GB Great Britain

GIS Geographic Information System

HSM Highway Safety Manual

IQR Inter Quartile Range Method

KDE Kernel Density Estimation

KNN K-Nearest Neighbours

MAE Mean Average Error

ML Machine Learning

MSE Mean Squared Error

ONS Office for National Statistics

PCA Principal Component Analysis

RAM Random-Access Memory

RMSE Root Mean Squared Error

SI Severity Index

SPF Safety Performance Function

SVM Support Vector Machine

SVR Support Vector Regression

VRU Vulnerable Road User

WAN Weighted Accident Number

WHO World Health Organization

Chapter 1

Introduction

1.1 Context and Motivation

As most of us know, data science is growing in popularity in the last few years and for a good reason, it helps to better understand the problems around us through processed data and draw factual conclusions based on that data [1]. Also, many entities, from companies to even sports teams, have been using data science to justify their decisions and choices which means data science is very popular in many different areas [2][3]. On the other hand, road accidents are a serious problem that society has to face nowadays since it regularly involves fatalities or significant injuries so, consequently, peoples lives are at stake. Also, the road transport system is very complex and dangerous which does not make solving this problem any easier [4]. A better way to see how serious this problem is is to observe the numbers of WHO road safety reports. According to the 2018 report [5], the number of traffic deaths reached 1.35 million in 2016 and the rate of death per 100,000 people has remained equal. People are now more likely to die as a result of a road traffic injury than diseases like HIV/AIDS, tuberculosis or diarrhoeal diseases and is the main cause of death for children and young adults aged 5 to 29 years old. The main risk factors that influence the number of road traffic accidents are excessive speed and driving under the influence [6]. Other significant risk factors also influence the number of road crashes includes the non-use of helmets and/or seat belts, utilization of distracting devices such as cell phones, unsafe road structure and vehicles, inadequate law enforcement of traffic laws, etc [6].

One way to solve this problem is to use the algorithms, data science and AI since there are many ways to identify dangerous zones such as analyzing historical data, calculating accidents rate, analyzing roads' condition, signalization and safety measures, among others. For national authorities, the detection of blackspots is important since they identify locations that need attention by the responsible authorities.

There have been several studies that mention or perform analysis on Portuguese Roads [4] [7] [8] [9]

[10] [11] which can give some early insights of the things that were done regarding the state of Portuguese roads, their risks, what can be done to improve them, other approaches to identify dangerous scenarios, among others.

1.2 Objectives and Research Hypothesis

Considering the intent of this dissertation and the problem previously stated, the objectives and research hypothesis of this work are as follows:

1. Extract relevant information from the reports available in the ANSR which contain data about accidents, namely, time, location, climate conditions, people involved, etc;
2. Join the information mentioned above with other relevant data collected from Portuguese's road accidents which contain crash locations, timestamps, cause of the accident, affected roads, description of traffic, etc;
3. Find useful information about blackspots (areas with a high number of accidents or an abnormal number of fatalities), spatio-temporal insights, incident correlation and incident causes/consequences;
4. Develop models to enhance road safety, prevent road accidents and find causes for incident recurrences;
5. Present all this information in a dashboarding platform open to the community.

This dissertation intends to, ultimately, make the community more aware of this problem to, consequently, reduce the number of accidents and the mortality rate that occur on Portuguese roads.

1.3 Problems and challenges

As it was mentioned in a previous section, Portugal's road mortality rate has been decreasing in the past few years but road accidents and victims involved in accidents have been increasing. Comparing the numbers gathered in this Figure 1.1 developed by the EU [12], we can see that the number of road fatalities per million inhabitants in Portugal is more or less equal to the EU average. Also, the evolution curve of this same number, shown in Figure 1.2 shows an evolution similar to the EU average but always above it. All this goes to show that something about policies, advertisement or road analysis has not been effective so far.

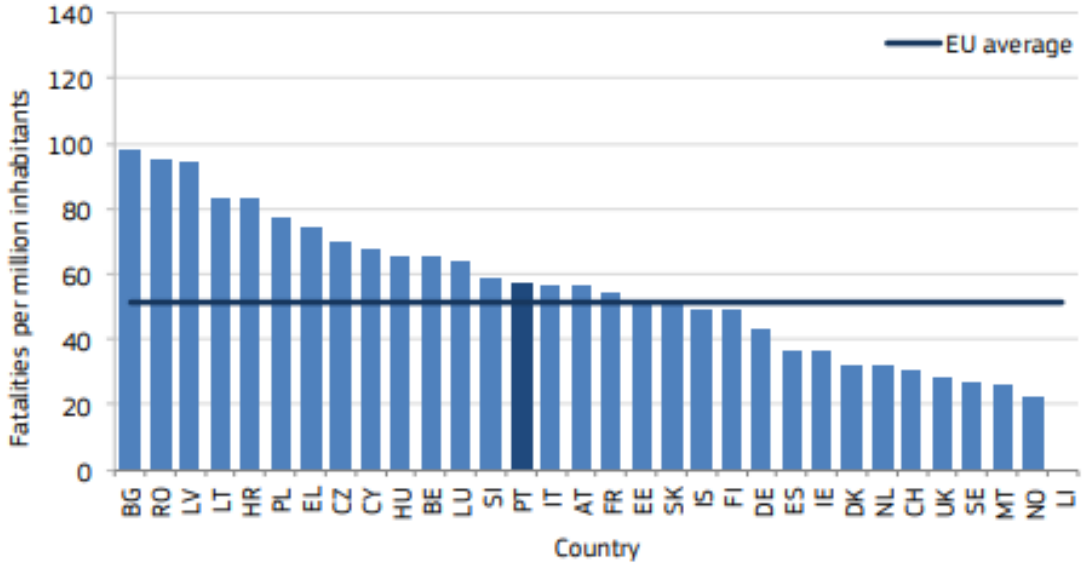


Figure 1.1: Fatalities per million inhabitants in 2015 with EU average. Adapted from [12]

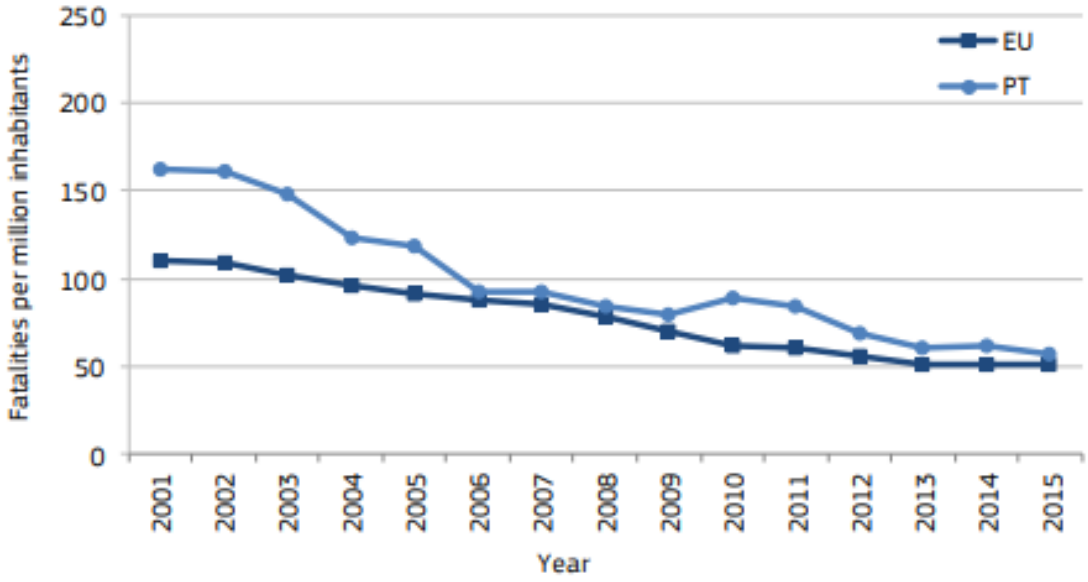


Figure 1.2: Development of fatalities per million inhabitants between 2001 and 2015 for Portugal and the EU average. Adapted from [12]

Another problem Portugal faces is that they're not using the most effective technologies to collect and analyze data gathered from their road network. As it was mentioned before, Portugal is still transitioning to more automatic tools, namely the Geographic Information System (GIS), and methods used to analyze the situation of road accidents around the country need to be reviewed since data science is growing and coming up with better solutions and could be used to solve these issues. Using the frequency of accidents or deaths to identify dangerous zones or periods instead of more complex models can jeopardise the veracity of the conclusions because they don't take into consideration things like randomness, a characteristic of

road accidents, and traffic volume. This dissertation intends to methodically analyze issues like blackspot identification, spatial and temporal relations with accidents and identification and interpretation of regular events or patterns in the road network and show the obtained insights to the community.

Some challenges are expected to occur. The first obstacle that may occur is to interpret data correctly. For the results to be accurate, it's necessary to know Portuguese road numbers well since they can give meaningful information when understanding data and consequently give the information to look at the problem and draw the correct conclusions.

The second challenge that may occur is when relating data. This means that it's going to be challenging to find which variables have more influence on the outcome and what can explain accidents in a certain zone or time. What makes this analysis difficult is that some accidents may have causes that are not noticeable through data, for example, poor road conditions or signalization at the time of the accident and driver's condition.

The third challenge that is expected is the withdrawal of useful and correct insights and information from data. The second and third challenges are somewhat correlated since to draw conclusions that are adequate to the problem in question, it's necessary to understand data, make relations between them and associate causes to the outcome.

The fourth and final challenge is presenting the conclusions of this study to the community. Even if all the work previously done is performed correctly, if that information is not presented in a way everyone can understand, internalize and act accordingly then all the work was done for nothing since people are not correctly informed.

1.4 Methodology

In the world of Data Science and Machine Learning, data mining is not an objective process, because it requires several different skills and there is no standard framework to carry out data mining projects. Also, the success of one project doesn't mean the next one can have the same success. This means data mining needs a standard methodology to eliminate the problems and subjectivity mentioned early so business problems can be translated into data mining tasks, data transformations and techniques are used appropriately and the results are effective.

This is where CRISP-DM comes since it offers a methodology for data mining projects to be independent of both the industry sector and technology. It also makes these projects less costly, more reliable, more repeatable, more manageable, faster and encourages best practices. In the same way that there are methods like Agile, PMI and others for developing software or SQL as a standard for relational databases there are also methods such as CRISP-DM for data mining projects [13].

This methodology's development was initially started in 1996 and was then refined through a series of workshops from 1997 to 1999 and published in 1999 with many organizations contributing to this process

model [14].

For the last 20 years, this process grew in popularity to the point of becoming the main methodology when carrying out data mining projects. In many pieces of research done to companies and users, according to Martínez-Plumed et al. (2019), CRISP-DM has been the standard for data mining projects and knowledge discovery. This domain has advanced significantly in 20 years with data science, being right now the preferred data mining method.

The CRISP-DM model provides an overview of the life cycle of a data mining project. It contains the phases of a project, their respective tasks and outputs. This life cycle is composed of six phases which can or not be sequential because the outcome of each phase dictates what has to be done next. Figure 1.3 represents the phases of the CRISP-DM process model for data mining, as follows:

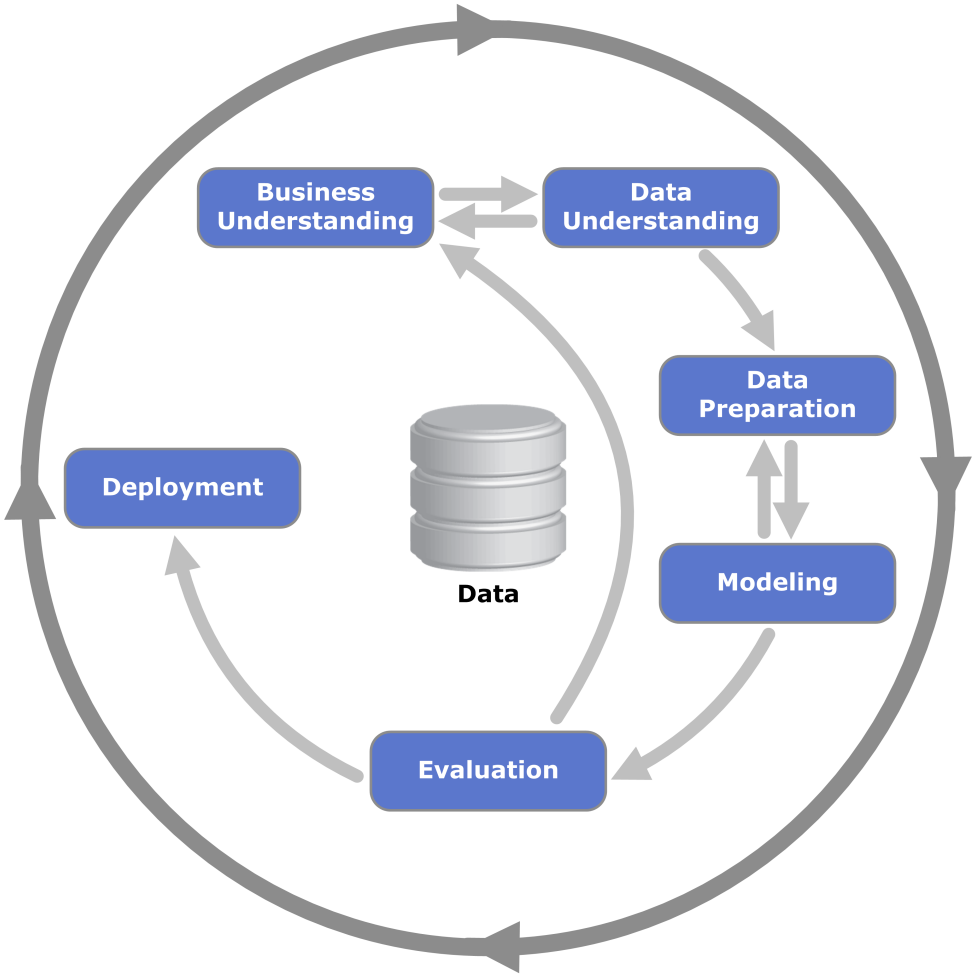


Figure 1.3: Phases of the Current CRISP-DM Process Model for Data Mining

- Business Understanding

  This is the starting point, it consists in knowing the business well and understanding the objectives of the data mining project from the business perspective. From there, it will transform into a data mining problem. It's considered one of the most if not the most important step in the whole process

since, without it, it's almost impossible to understand the problem and to give accurate results or to solve the problem.

- Data Understanding

  The first thing to do after understanding the business well is to understand the data. To understand it, we must collect it first to become more used to it, identify possible missing values, problems with data size, get some early insights that can be useful later one, etc. This is also a very important step since it can prevent future and unexpected problems in the next stage of the methodology.

- Data Preparation

  Normally, this is the longest stage of a project since it's where initial data is analyzed in-depth and processed to then result in the final dataset. The process of data preparation doesn't have a well-defined order nor mandatory tasks because, for example, it may be required to delete some data and then process and transform it or it may be better to process data and then delete data or instead attributes and then process it again. The final data will be used in the next step, the modelling stage.

- Modeling

  Many modelling techniques can be used depending on the problem at hand. This stage is where analysts execute modelling through various iterations while using different models with standard configurations and depending on the results, will adjust those configurations until, hopefully, optimized to deliver the best result possible. It's usual to go back to the data preparation stage since different models may require different ways to handle data.

- Evaluation

  Before the final stage, it's important to evaluate the model in-depth and make sure that it's done correctly while achieving the objectives previously laid out. One important aspect of this stage is to identify relevant problems that haven't been addressed yet and if there are it may be recommended to come back to the first stage and follow the previous flow again because results can be unsatisfying.

- Deployment

  The final stage of a data mining project is where the analyst has discovered new information about the problem he faced and has to present it in a way that the client can understand and use that knowledge to the benefit of its business. Deployment can be as easy as writing a report or as difficult as a data mining process applicable to an entire enterprise.

1.5 Document Structure

This dissertation is going to include state of the art, data understanding, processing and analysis as well as the model setup and finally results, discussion and conclusions. Firstly, in the state of the art, the goal is to discuss some of the topics that revolve around data science methods to enhance road safety along with non-data science methods for some contextualization, examples of some countries' situations on the road network and their data and the best dashboarding platforms available. The next two sections will be talking about the data exploration, how it was processed and analyzed as well as useful information for modelling. After that, the setup, experiments and decisions that led to optimal hyperparameter tuning for models to follow. Next, the results of the models will be discussed as well as the insights stemming from the dashboard platform that contains visualizations regarding the data mentioned. Lastly, the conclusion of this dissertation with what was achieved and what could have been better as well as subsequent work.

Chapter 2

State of the Art

This section will start by highlighting road data, prevention and analysis in some EU countries, followed by a description of methods applied to road safety based on data science as well as other algorithms and an overview of the best dashboard platforms discussing their characteristics as well as their pros and cons.

2.1 Road Data, Prevention and Analysis

To start, it makes sense to talk about road accidents' data, in this country, before going in-depth about details of some practices and solutions developed. After going through ANSR reports, which is the National Authority for Road Safety, we get to know more about what is the past and current situation. Table 2.1 shows that the number of accidents with victims suffered a decrease between 2009 and 2013, however, these numbers increased significantly until 2018 (close to 2009 numbers) [15]. Besides the number of lightly injured people, the numbers show a decrease in all other serious categories even though the numbers are still significant. One thing worth mentioning is the year 2017 because of the huge increase in all categories as shown in Table 2.1

Year	Accidents with victims	Accidents with fatalities and/or serious injuries	Accidents with fatalities	Fatalities	Serious injured	Light injured	Total injured	Severity index
2009	35484	2777	673	737	2624	43790	46414	2.1
2010	35426	2802	674	741	2637	43924	46561	2.1
2011	32541	2641	636	689	2436	39726	42162	2.1
2012	29867	2264	525	573	2060	36190	38250	1.9
2013	30339	2191	469	518	2054	36818	38872	1.7
2014	30604	2317	454	482	2152	37019	39171	1.6
2015	31953	2358	438	473	2250	38826	41076	1.5
2016	32299	2201	416	445	2102	39121	41223	1.4
2017	34416	2397	488	510	2198	41787	43985	1.5
2018	34235	2337	468	508	2141	41356	43497	1.5

Table 2.1: Historic information about Portuguese Road accidents [15]

To summarize, there has been a dangerous tendency, where we see an increase in road accidents and lightly injured in the last few years, even though the mortality rate and seriously injured have decreased in the same amount of time. This appears to be a growing problem since it does not show signs of deceleration any time soon, as long as there are no big changes in advertising and policies.

Some solutions that have been developed and are still operational include promoting education and training for a culture of traffic safety, improving protection on vulnerable road users, improving national and municipal road networks, optimizing help mechanisms and rehabilitation of road victims, improving legislation, inspection and sanctioning, etc. These are some of the measures included in the National Strategy of Road Security (ENSR) [16].

One of the main problems identifying the issues in the road network is the way they collect information since the police forces in Portugal have been slowly transitioning to a more automatic method [7]. There have also been efforts to include a GIS in the act of collecting data so, as long as there is education on this subject, national police forces can use this technology efficiently. It can be a great addition to help solve these problems. These are some of the advantages of applying GIS models to data:

- If the location is correct, it's easier to cross data with other equipment types like schools, hospitals, crosswalks, etc;
- Identifying possible problems the roads have regarding the accumulation of rain and drainage problems that might exist;
- Time of the day when accidents happen;
- Allows creation of accident prospecting models, in other words, predicting where new accidents might occur.

An analysis of other European countries, such as Croatia, shows that from 2001 to 2014, according to the ERSO in Croatia's Road Safety Country Overview [17], the fatality rate has been a bit higher than the EU average but has been steadily decreasing and it started to close the gap. Some other aspects to take into account are information about the accidents themselves from the last registered year of 2015. Pedestrians, car occupants and motorcyclists represent most of the fatalities registered, males are the group that was most reported on those fatalities but a big decrease happened in the age gaps 18 to 24 and 25 to 49, build-up areas and rural areas are the places where most fatalities happened (around 80-85%), fatalities are more likely to occur during daylight than night-time, those percentages are higher than EU average and fatalities while raining make up for 12% of those, 2% higher than EU average.

In Croatia, there is no systematic model to identify dangerous areas, unlike other European countries. The criteria available for determining blackspots is used for state roads that are under the authority of the Croatian roads, which means some blackspots are not being recognized, since some roads are not state roads which can lead to incomplete data and defective analysis. These state road's blackspots can be characterized as an intersection or road segment in the length of 300m or part of the length of 300 to 1000m, with the condition that satisfies one of the next three criteria:

- If on the critical location has occurred 12 or more road traffic accidents with injured people in the past three years;
- If in the monitored location, 15 or more accidents are recorded regardless of the consequences, in the past three years;
- If in the critical location three or more identical accidents have occurred with the same group of participants, with the same moving direction and with the same conflict area etc.

According to [17] and [18], road safety has been improving with the number of fatalities on the road decreasing whether it's per 100,000 inhabitants, vehicles or drivers. These stats are followed by an increase in the number of registered motor vehicle drivers and registered motor vehicles which can be a good sign. Even though these numbers have been dropping it seems that Croatia is one of the worst European countries, in 2011, in the number of deaths in traffic accidents per 100,000 inhabitants.

When it comes to road types, the reports say that a significant part of human lives is lost on road in the settlements (43%), followed by state roads (22%), highways (12%) and other county, state or local roads with less than 10%.

In Great Britain (GB), there is a database called STATS19 where all the information about road traffic accidents that resulted in a personal injury and were reported to the police within 30 days of the accident is stored (that occurred in GB). By reading reports that analyse the information inside the database, we can collect the tendencies with road accidents, injuries, mortality rate and also information about vehicles,

drivers, comparison with EU averages and the conditions in which accidents happened (p.e. the weather, kind of people involved, etc).

We can understand all this by reading and studying the report [19], namely the graphics, the numbers and descriptions that were created using the database, referenced earlier, from where was collected data until 2019. What data shows is that fatalities have been decreasing for a long time but have somewhat stagnated since 2010, even though the expectation was to continue the previous tendency. Regarding accidents involving seriously injured people, there seems to be a misinterpretation of these numbers due to the reporting systems since, with prior systems, some serious injuries were treated as light injuries, therefore, changing the real information. The data gathered from police shows a somewhat constant decrease of serious injuries from 2004 until 2015 and then an increase until 2019 but Office for National Statistics Methodology Advisory Service made a study to analyse the real numbers using the latest methodologies to compare the numbers between the data previously mentioned and the adjusted numbers which are a steady decrease from 2004 through 2019, as seen in Figure 2.1.

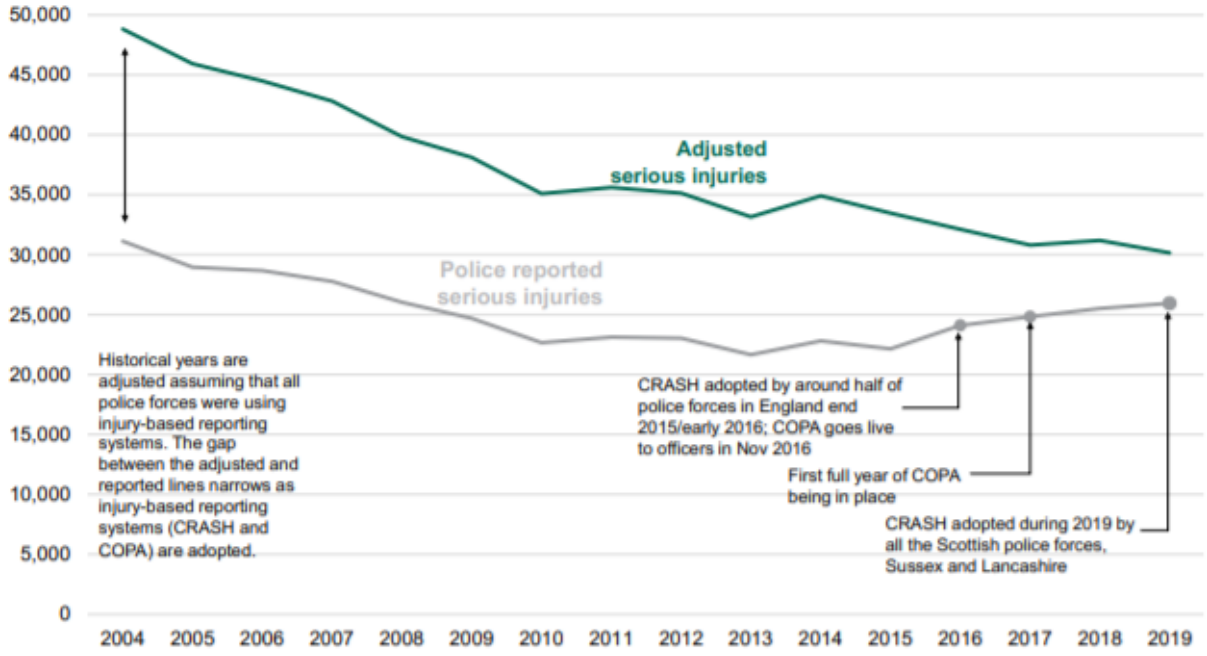


Figure 2.1: Serious injuries reported road accidents (adjusted and reported). Adapted from [19]

Since the previous point was influenced by the light injury numbers (the model used made some serious injuries get classified as light injuries) it's natural to assume that, here, instead of the adjusted curve being over the reported by the police, the adjusted curve is below of the reported by the police. What the report shows is that this assumption is correct, as shown in Figure 2.2, and even though it looks as both curves have a low amplitude between them, it's worth mentioning that the scale on the Y-axis of both graphs is very different so it can give a wrong perception of this data. What we can see from the 2 curves is that they both follow the same tendency decreasing from 2004 until 2013, with a slight increase

in 2014, followed by a steady decrease until 2019 but with the adjusted curve being a bit shifted down, as we explained earlier.

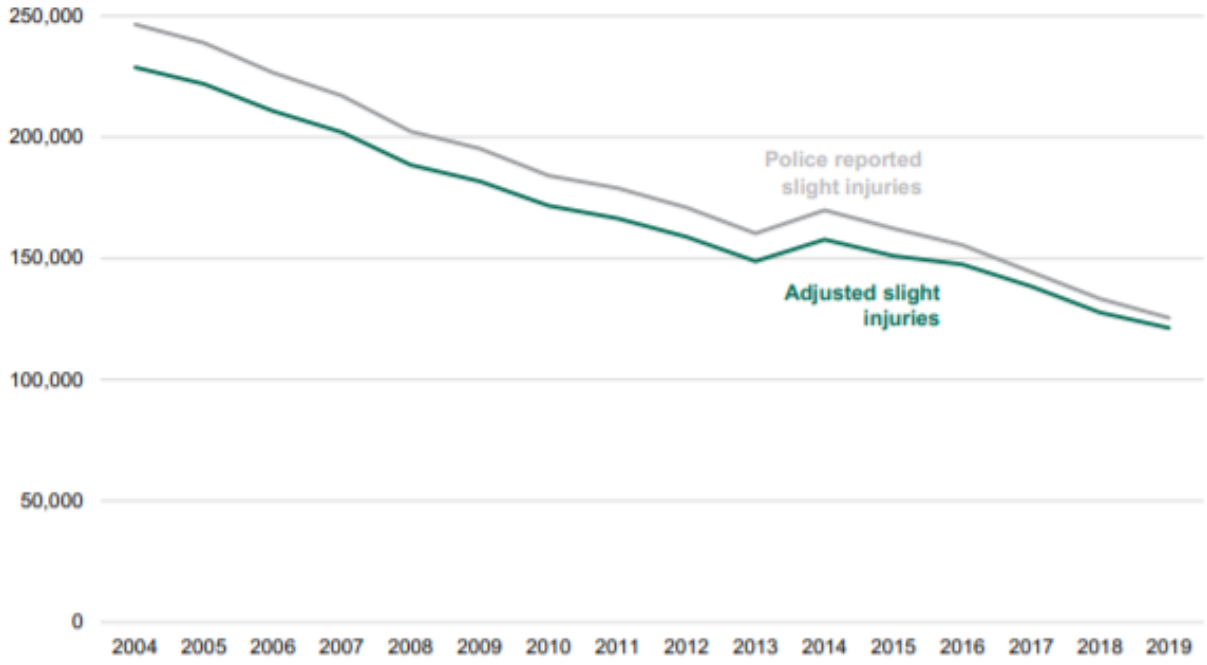


Figure 2.2: Light injuries reported road accidents (adjusted and reported). Adapted from [19]

In terms of casualties, from 1979 until around 2000, there have been some ups and downs, but, since then, they have been decreasing at a rapid pace.

2.2 Methods Applied to Road Safety

There are several methods known and used to enhance road safety which have their pros and cons. In the following section, we'll describe and explain those methods, along with the appropriate situations to use them due to their nature.

- **Accident Frequency method** - This method takes into consideration, as the name suggests, the number of accidents as an identification criterion, in other words, if that number is higher than the criterion, then the section is classified as a dangerous zone. This is pretty straightforward which means it's easy to calculate and have an idea of that section's accident risk, but the downside is that it's too simple. It doesn't consider many other variables such as the road, traffic or weather conditions which means we may be thinking one section is a blackspot just because it's a blackspot without attributing any other insight into it. In conclusion, it is suitable for sections and intersections where conditions are similar, and traffic is not heavy [20] [21] [22];

- **Accident rate method** - This method is slightly better than the previous one since it nullifies some of the flaws of the previous method. The defining criterion here is traffic volume (accident number of million motor-kilometres of one year), which means when the accident rate exceeds the criterion it's regarded as a dangerous zone. Generally speaking, it's better to use this metric to enhance road safety than the previous one, but there are some drawbacks, namely that the accident rate is high when traffic and accidents number values are low and, on the other side, the accident rate is low when traffic and accidents number values are high. This method requires lots of data, and it may still give inaccurate results. It's suitable to use this method when describing regional accident conditions [20] [21] [22];

- **Matrix analysis method** - This method utilizes both previous criteria to identify dangerous zones, accident rate and accident number. Each road section is represented by a cell that has a value which, in turn, make up a matrix, and those values show the risk of a road section. The higher the accident rate and accident number are, the higher will be the number inside the cell representing a section of the road. Even though this may be more accurate and flexible than other methods, it has some shortcomings, for example, it can show the risk of a road section but cannot distinguish road sections that have low accident numbers and high accident rate or high accident numbers and low accident rate without considering the criterion and severity of the accident. This method is suitable for road sections or intersections where conditions are similar, and traffic is not heavy [20] [21];

- **Equivalent Accident Number Method** - This method calculates the degree of severity of an accident through various aspects, such as death victims or fatalities, severe injuries, slight or minor injuries and property damages. To identify the ranking of each cell it's required to determine the weighted accident number (WAN) since there are several types of degrees of an accident based on the number of victims and seriousness of the accident. The formula to calculate WAN is by adding the aspects referenced earlier with certain coefficients. Indonesia has a particular situation when it comes to calculating this method since different coefficients are used within the country. Next, are going to be presented some of them [23] [24].

  - The Engineering Committee for Standardization of Transportation Infrastructure decided to assign the values 12, 3, 3, 1 to estimated economic losses caused by death (M), serious injury (B), minor injury (R) and property damage (K), respectively:

$$WAN = 12 * M + 3 * B + 3 * R + K \tag{2.1}$$

  - The Accident Point Average has the values 6, 3, 0.8, 0.2 assigned to those same categories, respectively:

$$WAN = 6 * M + 3 * B + 0.8 * R + 0.2 * K \tag{2.2}$$

- Indonesian Police utilize the values 10, 5, 1, 1 assigned to those same categories, respectively:

$$WAN = 10 * M + 5 * B + R + K \tag{2.3}$$

- The Directorate General of Land Transportation use the values 12, 6, 3, 1 assigned to those same categories, respectively:

$$WAN = 12 * M + 6 * B + 3 * R + K \tag{2.4}$$

- **Quality Control Method** - This method is quite different from the previous ones since it's based on some assumptions [25]. Practical application shows this method is often better than other traditional statistical methods. The steps to perform the calculation of this method are, firstly, to assume the number of accidents follows a Poisson distribution, which means the probability of *n* traffic accidents that happen within time *t* in one of the road sections, is as shown in the following formula. *u* is the number of road segment accidents, *n* is the number of traffic accidents and *t* is the time when those *n* accidents happen:

$$P(n|u,t) = \frac{\epsilon^{-ut}}{n!}(ut)^n \tag{2.5}$$

  Secondly, compare the road segment accident rate to the average rate of similar road segments, instead of comparing with all sections of the average accident rate. Lastly, and according to a significance level, determine the accident rate upper and lower limit. After determining the value range, if the accident rate is greater than the upper limit of the inspected section, it's considered a blackspot. To sum it up, by comparing to the other road section we are considering the road conditions, but it requires a lot of traffic data and classification work so it can be very accurate but with rather effort. It's suitable for road sections with low traffic flows [20] [21];

- **Critical Rate Method** - This method utilizes a measure called critical rate calculated by a function of the average crash rate of a reference group associated with the site, the traffic volume of the site, and the desired level of confidence [26]. Zones where the crash rate exceeds the critical rate value, are classified as blackspots. Since different sections of the same road can have different critical rates, it offers a more individualized analysis of those zones, however, it requires data to be updated often since the critical rate is calculated using any other information, other than the site itself. This is better than using the accident rate or frequency method because it's more robust as it

provides a means to statistically test how different the accident rate is at a site when compared to a referenced group. Still, it doesn't consider the severity of the accidents and assumes that traffic volume and accidents have a linear relationship [20];

- **Spacial Auto-Correlation** - This method utilizes the concept of auto-correlation (objects close to each other are more likely to have similar values) to identify relations between zones based on the spatial aggregation of contiguous spacial units (crashes) that are geographically close. To do this it's required the degree of co-variation between the different spatial points close to each other utilizing a global index I. If that value is positive, then there's a positive association between those spots and higher co-variation. If that value is negative, then there's a negative association between spots. If that value is 0 or near 0, then there's no association between those spots. It takes into account simultaneously discrete events' locations and values, in other words, defines similarities between them and put them into that index I. Nevertheless, this global index can sometimes fail to identify relations between zones when, for example, there are equal amounts of positive and negative clustering, so it's necessary to use a local index to detect these locations separately. This local auto-correlation is used to discover spatial variation and association between approximate spatial units. The procedure to identify blackspots using this method is the following:

  - Divide the road into small sections (i), for example, each one has 250 meters, and count the number of accidents (x_i) that occurred in each section.
  - Calculate the local index I_i for each i location with j values for all other locations.

$$I_i = z_i \sum w_{ij} * z_j \tag{2.6}$$

    Where: $z_i = x_i - \overline{x}, z_j = x_{ij} - \overline{x}$ ($\overline{x}$ here represents the critical number of accidents)
    And, $\sum_j w_{ij} = 1$ (the weight here is row standardization)

  The danger level depends on the value of the local index I, while the zone length determination depends on the weight matrix w_i. Therefore, the hazardous zone can be determined for various lengths, while it also depends on the critical number of accidents on the related zones [27];

- **Empirical Bayesian Method** - This method combines both observed and predicted accidents' frequencies, for a specific roadway network, in one statistical model, using the following equation:

$$N_E = w * N_p + (1 - w) * N_o \tag{2.7}$$

  The expected number of crashes (N_E) can be used to estimate the expected average crash frequency for both future and past periods, if only observed (N_o) and predicted (N_p) number of accidents are available. The weight factor (w) in the equation represents the degree of reliability

in obtaining N_p, and it's inversely proportional to its over-dispersion parameter that measures the degree of dispersing in N_p for the different study time included. If the dispersion of the predicted number of accidents (N_p) is more dispersed, then the weight in the Empirical Bayesian (EB) equation will be lower, and vice-versa. The predicted number of accidents can be anticipated using Safety Performance Function (SPF) which is a regression function that estimates (N_p) for the study period under given conditions. When calculating SPF, the segmentation is also another important issue which means how road sections are divided since it's recommended to use a homogeneous segmentation with a length of about 500 meters as an average. The procedure to calculate blackspots using this method is as follows:

- Calculate SPF for the selected road using the following function, after determining each road segment length (L), average annual daily traffic (AADT).

$$N_p = \exp(a + b * \ln(AADT) + \ln(L)) \tag{2.8}$$

  Where: a and b are regression parameters. Their values depend on the type of road (number of lane and median type) and type of collision. This equation is used for urban and suburban arterial roads according to Highway Safety Manual (HSM).

- The differences between N_E and N_p of the various segments are calculated, and the positive values refer to blackspots road segments.

- **Temporal and Spacial Analysis of Road Accidents** - It's common when talking about accident data, and performing analysis on that data, to use variables such as time and space (location) since they're of great importance when describing, for example, the most dangerous spots, hours of traffic jam, relations between them, etc. They are very important pieces of information when trying to understand road dynamics and to elaborate road safety plans according to that information. Normally the space analysis is done using the latitude and longitude coordinates, and then methods are used such as the ones explained in previous sections. Fortunately, a specific system that has been gaining popularity is the GIS [28]. It's a framework that provides the ability to gather, manage and analyze data. Using this data to feed a data model, we can then observe through visualizations, namely maps and 3D scenes. Consequently, GIS can give deeper insights into data, such as patterns, relationships and situations making this analysis and drawing conclusions easier. This system can be used for multiple purposes but is mainly used for these six topics:

  - Identify problems (that are driven by geography);
  - Monitor transformations/modifications (p.e. glacier retreat);
  - Manage and respond to events (p.e. weather events);

- Perform forecasting (p.e. traffic);
- Set priorities (p.e. assign officers to more problematic areas);
- Understand trends (p.e. employability and competitiveness);

For this problem, GIS is used to present hot spots of road accidents through GIS maps and other forms of visualizations and perform Spatio-temporal analysis through various methods/techniques so measures can be taken towards preventing road accidents. Other areas where GIS can be used besides road traffic-related issues are topography, biology, geology and remote sensors.

2.3 Machine Learning Approaches

Besides the methods mentioned previously, engineers and scientists use more advanced techniques that require programming and/or knowledge in computer science to perform them. The most notable ones are inserted in the machine learning category since this area has been growing in popularity in the past few years, and the number of libraries linked with machine learning has been increasing and improving at a rapid pace.

2.3.1 Supervised Learning

The defining characteristic of supervised learning is that these algorithms learn a mapping between a set of input and output variables. These algorithms induce models from this data and can then be used to classify other unlabelled data (that doesn't have output values/classification). Typically, data are divided into training data and testing data so the first one can be fed into the model for learning and the second one used for testing if the model is accurate enough to be used on unlabelled data. Supervised learning is the most common methodology in machine learning. The most popular supervised learning algorithms include support vector machines (SVM), linear/logistic regression, neural networks and k-nearest neighbour [29] [30].

A brief definition of deep neural networks (DNNs) is that they are artificial neural networks with multiple layers between the input and output layers. On the other hand, artificial neural networks (ANNs) are a collection of connected units or nodes called artificial neurons, which model the neurons in a biological brain. Like a human brain, these networks can also transmit signals from neurons to other neurons through connections between them. Each neuron that receives a signal, processes it and sends a signal to neurons connected to it until they reach the last layers and give an output for the problem. The first thing to do is create a DNN that adds information about the relation between data and their categories. The target data enters the input layer of a DNN generated based on the data-category information mentioned previously. Finally, the category information of the target data can be obtained from the output layer of the DNN.

The category information of the target data is obtained by establishing a DNN based on the category information of the data itself. Consequently, the category function used in the DNN is recognized, which is convenient to mine the deep regular pattern of the target data [31].

To use DNN, it's required to follow some steps that are going to be enumerated next [31].

1. Establish an initial DNN;
2. Generate a linear category analysis function after adding the data-category information in the locally saved initial linear analysis function according to the input training sample vector set;
3. Obtain an optimization function of the initial depth neural network according to the locally stored unsupervised coding model optimization function and the linear class analysis function;
4. Acquire parameters of the initial deep neural network according to the optimization function of the initial depth neural network;
5. Establish a deep neural network according to a locally stored classification neural network, an initial deep neural network, and parameters of the initial deep neural network;
6. Receive the input data to be identified;
7. First, the data to be identified is input to the input layer of the deep neural network. Then we obtain the category information of the data to be identified from the deep neural network's output layer.

2.3.2 Unsupervised Learning

In contrast with the previous paradigm, unsupervised learning algorithms learn from untagged data, meaning that it doesn't have output variables, only input variables. These algorithms build representations of the inputs that can be used for decision making, predicting future inputs, among others. Instead of finding patterns between input and output variables, they find patterns in data (input variables). However, it's still required to divide inputs into train and test data for the model to learn and verify its accuracy. The most popular unsupervised algorithms include k-means clustering, principal component analysis (PCA) and hierarchical clustering [32].

K-means clustering, as said above, is an unsupervised learning algorithm that classifies the input set of data into multiple clusters based on the distance of the representation of each input variable. For calculating the distance between them, different metrics can be specified on the algorithm to try and group the input data into various clusters. The points are clustered around center points called centroids. The steps to take to implement K-means clustering are as follows:

1. Compute the intensity distribution of the intensities;

2. Choose *k* centroids randomly;

3. Repeat the following steps until the cluster does not change anymore;

4. Cluster the points based on the distance of their intensities from the centroid intensities;

5. Compute the new centroid or mean point for each cluster.

2.3.3 Reinforcement Learning

Reinforcement learning algorithms interact with the environment by producing actions related to situations. These actions affect the environment that can return a response, which is called rewards, that, in turn, can be positive or negative. The algorithms' objective is to maximize the number of positive rewards received, consequently minimizing the negative ones. The learner doesn't know what actions to use so it must learn by discovering which actions bring more positive rewards. In some cases, actions may not only affect the rewards but also the next situations and all subsequent rewards, so trial-and-error and delayed rewards are the two most important distinguishing traits of reinforcement learning. Some reinforcement learning algorithms include Monte Carlo, Q-learning, SARSA, A3C, DQN, among others [32].

SARSA learning is known as an on-policy method, meaning that when updating the current state-action value, the next action will be taken into account. This means that for a state-action function $Q(s, a)$, where *s* represent the state and *a* represents the action, the reward, the next state and next action also belong in the training data (a quintuple (*s,a,r,s',a'*) where *r* represents the reward, *s'* the next situation and *a'* the next action). Another popular reinforcement learning algorithm is Q-learning which has many things in common with SARSA learning, but with a significant distinction that the next action is not considered to update the function. This means that the training data is composed of a quadruple (*s,a,r,s'*). The next action *a'* only serves for estimation because it's unknown, so the Q-learning algorithm tends to opt for more greedy actions since it has less information to work with [33].

2.4 Gathering and Analysis of Dashboarding platforms

Ultimately, this dissertation intends to warn people about dangerous zones, make them more careful when driving around those same zones and when driving in general. To make that information available to people it's necessary to present it in a way that everyone can check and understand. To do that, it's required to study some dashboard platforms that can be used for that purpose. So next, we're going to present and give a brief explanation about some of the best free dashboard platforms available that can be used for presenting road data information such as hazardous zones, traffic analysis, traffic patterns, spatial and temporal analysis, among others.

2.4.1 Google Data Studio

This was released in May 2016 as part of a paid package (Analytics360) and was made free a few months later [34]. Its main function is to present information about social media and web analytics such as YouTube analytics, but it could be used by researchers to interpret and visualize their data in the same way as its main intent. It strives in creating attractive and understandable data visualizations even though the process of creating them may not be very intuitive. It contains the same types of charts and graphs as other dashboard platforms but makes them better through features. It also impresses when allowing viewers to filter and adjust in real-time to enhance the dashboard's relevance to the viewer. Sharing reports publicly or privately is also possible with Google Data Studio with collaborators being able to modify reports without being able to change original data unless other access permissions are granted. One of this platform's drawbacks is that it usually doesn't protect personally identifiable data, not complying with IRB requirements, so it's recommended the usage of anonymous or aggregated data. Another thing is that it cannot modify underlying data and offer calculation and visualization options when compared to other platforms, even though it offers a simple user experience and basic reporting.

2.4.2 QlikView

It was released in 1993 [35] with the functionality of extracting data from database systems, then summarizing it and graphically presenting it through views. Now it's an advanced tool with major improvements such as being very fast at performing aggregations and calculations since it uses RAM for physically storing data. It also offers an interactive user experience because it presents all the information and data from the start, unlike other platforms, while delivering an attractive and interactive interface. Another great built-in feature is its patented query language *associative query logic*. It's easy to use and can perform many actions that are very important for users exploring and trying to take information from data. QlikView also maintains relationships among all data points in memory which means that extensive SQL queries only have to be written once and, from there, the front end has all the data and its associations intact only requiring adding the dimension and a measure to the configuration. Finally, this platform can be used for all solutions, it's an all-around free tool that can get the job done regardless of the "theme". However, all this comes with some drawbacks such as not being very efficient when analysing data in real-time and having a data load limit since it uses RAM for loading it so depending on the computer, it will be able to load more or fewer data, always with a cap associated to it. It can also be a bit weird when interacting with other software because they can have a very distinct appearance. For more complex projects or analysis, it may be required to purchase extra features which can make some small and medium users and enterprises opt for other affordable and convenient BI tools. When it comes to customer support, it doesn't deliver as it should, leading to occasional low-rated reviews damaging the tools' reputation.

2.4.3 Cluvio

Cluvio is cloud-based analytics aimed at data-driven teams, startups and companies. It makes use of SQL and R to analyse data, and, with it, create interactive dashboards in just a bit of time. All kinds of enterprises can use this platform whether they are startups or big companies since there are many packages, from a free one to a subscription around €2000/month. Like many other dashboard platforms, this provides an interactive analytics dashboard where a user can run queries, filter results and choose how the data is present overall, along with many chart types to choose from. It has a great collaborative characteristic since dashboards can be shared using an only-viewer option or editing option and can be shared through links, email and even schedule and automate the sending of dashboards to specific people with specific options. It has a SQL editor built-in, so it's not necessary to repeat code. SQL alerts are included, so users can be informed about specific conditions, receive an email when this happens and suggestions on how to best visualize and present their data. Besides this, it's pretty easy to use and learn about the platform and its features. Users that don't know SQL may be at a disadvantage when using this platform since it's required to know this language and have some experience. Even though, generally speaking, it's easy to learn and use Cluvio, some features could use some documentation and/or examples to show how to use them. This software's free subscription is mainly aimed at users that only want to create simple dashboards since some advanced chart options are not present in the free subscription.

2.4.4 Power BI

Power BI is a Microsoft tool intended for business intelligence problems by converting data into meaningful information by using premade tables and visualizations so people can make important business decisions based on the information gathered. It's an affordable tool since the desktop version of Power BI is free and if a user wants some more advanced features or wants to share reports on the cloud there's always a $10 plan per month which is much cheaper when compared to other BI tools. It offers a big range of general and custom visualizations, including several ones made by developers which are available in the Microsoft marketplace, including KPIs, maps, charts, graphs, etc. It has great data connectivity since the user can import data from a wide range of data sources, from data files such as XML, JSON or CSV, to databases such as SQL or Azure, even online services such as Google Analytics, Facebook or Twitter and being able to connect to Big Data sources directly. The visualizations on Power BI have an edge relative to other platforms since they are very appealing and interactive with a drag-and-drop system of adding views to a report where we can apply filters, select and highlight data, among others. It also gets updates and upgrades very often since Microsoft created communities where users can make suggestions to improve the software. Finally, it has the ability, through Power BI Embedded, to integrate visuals in web apps or other apps, which is great for programmers. Having said all this there are also some downsides of using Power BI, such as not being great at handling complex relationships between tables, so special

care is required when creating the data model. There are also not many configuration options regarding visualizations which can be a pain to more meticulous developers. The user interface is not very friendly to beginner users since there are many options to click that can be confusing and block the view of the dashboard or report. Power BI has an expression language called DAX that can perform lots of actions on data, but it's not very easy and requires study and experience to use since some formulas may not work on users' data. Finally, even though Power BI is an easy tool for importing data and developing reports and dashboards, it's another conversation when talking about all the other interrelated tools. If the intent is to do more than that, developers will need to put in hours to learn and master other tools they will need.

2.5 Literature Review

This section describes scientific articles that have done work similar to this work, particularly for Portuguese roads. One, in particular, is *Acumulação de Acidentes Rodoviários em Portugal Continental: Contributo dos Sistemas de Informação Geográfica* [7] which addresses the general and Portuguese situation on blackspots, collection of data, modelling spatial and temporal analysis, the impact that GISs can have on this process, and how can it help improve the current methodology. First, it mentions how dangerous zones are identified, why is it flawed, and the reasons why GISs can help improve the identification of dangerous zones. Next, urban and non-urban areas are differentiated so a different study can be performed on each one of them regarding the identification of dangerous zones. After that, a spatial analysis was made using GIS methods that identified density in the distribution of phenomenons, namely kernel density, Moran index and spatial accumulation techniques (clusters). It was said that models utilizing cluster analysis were better for analyzing intersections instead of roads but were tested anyway and, in the end, couldn't be used because of the lack of quality in the data and not enough geographic variables required for these models. Consequently, the two other methods were tested and compared. The Moran index method uses the normalized values to analyze the behaviour of spatial variability of traffic accidents. The kernel density estimation (KDE) estimated the intensity of a set of points inside an area and appeared to have significant advantages when performing spatial analysis. It used more information than the majority of methods, showed dangerous zones through spatial clusterization, comparing them in space, and provided more stable results but had a big disadvantage which was when it encountered a massive concentration of points (in this case accidents) because the virtual analysis became twisted and it couldn't be performed correctly. Through this method, a spatial analysis was done and several accidents accumulation zones, in Lisbon, were located.

Another article addresses the topic of predicting vulnerable road users (VRUs) risk injuries with multinomial logistic regression based on spatial and temporal assessments [8]. It starts by mentioning what methods are used to identify blackspots (KDE) and continues describing how the crash database was developed and then a spatial and temporal analysis of crashes involving VRUs. KDE was used to carry out the

spatial analysis, and a thorough annual, monthly, weekly and hourly analysis was done while relating them with many variables such as age group, injury severity, gender, road type, location, environment, etc. The results were then presented maps, tables and graphs. After this, a multinomial logistic regression model was used involving vehicle-VRUs crashes for 3 distinct cities, Lisbon, Porto and Aveiro with the intent of predicting the probability of cyclists getting involved in road crashes and to see which variables were statistically significant for that prediction. It's worth mentioning that each variable had a noticeable result, for gender, age group, severity, weekday, period and weather conditions were, greater than 65, fatal, Sunday, 8 to 11 pm and bad weather conditions, respectively. Overall, gender, age group and weather conditions were statistically significant in all three models where gender had a negative effect while the other two had a positive effect. It was concluded that high attraction places and intersections were the zones where most accidents and most severe happened, respectively, with more than 40% of pedestrian crashes occurring in the vicinity of sidewalks. Aveiro, which is a medium-sized city, has more to worry about the active age female group since they are the most vulnerable, Lisbon and Porto, which are big-sized cities, have to worry more about older adults both in injuries and severity. Concerning cyclists, as was mentioned earlier, gender, age group and weather were the most important variables to predict this type of crash.

A third one addresses the investigation of risk factors that impact road safety to better analyze and prevent road incidents so we can know which ones can cause more harm and try to control them or make changes, if possible, on those risk factors [36]. It uses a Bayesian hierarchical model to deduce the expected number of fatal and severe injury crashes using the observed number of fatal and severe injury crashes and the number of vehicles insured. This study applies to all the counties in Portugal and is in a period of 8 years so the values mentioned earlier vary between each county and each year. There's also a relative risk associated with the observed number of fatal and severe injury crashes which can be represented by a spatio-temporal model containing factors, such as, the covariates under study, spatial heterogeneity, unstructured heterogeneity, a linear time trend and interaction between time and space. These models were implemented using WinBUGS which is statistical software for Bayesian analysis, along with an add-on called GeoBUGS which fits spatial models and produces maps as outputs and another add-on called R-INLA which can be used for Bayesian model inference. All the software mentioned earlier is available when using R (software environment). The covariates under study are geographical area, in km^2, population size, in number of inhabitants, road length, in meters, differing for each county and year. Then, eight models were created, the first containing no covariates and the eighth containing all of them. The results were compared using deviance information criteria (DIC) which is effective in Bayesian model selection, and all eight models had close DIC, though the one that had the best score was the model that only had road length as a covariate. From there, maps were produced that held information about expected relative risks of fatal and severe injury crashes in 2000 and 2007. The results weren't very conclusive since the inclusion of more variables was important to gather more decisive conclusions, and some of them might have been, for example, average precipitation, average traffic density, county wealth,

etc. Nevertheless, the results showed that road length can be a factor associated with fatal and severe injury crashes, that these kinds of crashes decreased during the study period (from 2000 to 2007) and that spatial-structured effects explained a greater part of the variability in relative risk when comparing to the other variables.

The last one talks about modelling crash frequencies for different temporal and spatial aggregation of crash data in Portuguese two-lane highways [37]. It starts by referring that, even though crash prediction models offer good insights about crash frequencies, they are not usually satisfactory to provide safety performance estimates, instead, historical data, statistical models based on regression analysis, other studies and expert judgements are used to provide those safety performance estimates. After that, other useful information about other articles, study objectives and defining aspects of the study, data about the roads studied and their segments are described as well as their geometric characteristics such as lane width, shoulder width, lateral offset, among others, as well as some definitions and calculations. Right after, the traffic data and crash data is presented, which is ranging from 1999 to 2010, and some analysis is made to their records to then justify the use of an adequate crash predicting model which, in this case, uses generalized estimating equations with the negative binomial link function. The general expression representing these models calculate the expected number of crashes, at each segment, over a period through annual average daily traffic at each segment in a period, finding the variables that might explain the number of crashes and a few model parameters to be estimated. A model is created with the combination of all explanatory variables and are then eliminated by a backward elimination procedure (the one that explains least variance is removed) where only remain variables statistically significant above 5% error probability. Each model is evaluated using the cumulative residuals method and marginal R^2 for fitting the model using the difference between the number of observed and predicted crashes and the Quasi-likelihood Information Criterion for correlation structure evaluation. The results showed that the best models are better evaluated if using 400-meter road segments on a period of 1 and 2 years and have four major contributing factors to crash frequency which are traffic volume (average annual daily traffic), lane width, vertical sinuosity (ratio between curvilinear length and straight-line between endpoints of the curve) and density of access point, all influencing the model positively. Even though these factors were identified as possible reasons for higher crash frequencies, the data supporting these models didn't have many data records which can compromise the results gathered, although the process to reach them was well established and explained, and, with much more and better data, could reach better and more conclusive results.

2.6 Summary

In this chapter, we mentioned the state of road incidents in Portugal and other European countries as well as some current problems and solutions. We also mentioned some methods applied to road safety

used nowadays, most being outdated with the tools we have today, their pros and cons and where they are usually applied. Next, we mentioned the most known machine learning approaches where two of them are going to be used in this research, some of the most used methods for each approach and a brief explanation about them such as when are they used, what are each one's preconditions and what they achieve overall. Then, a description about some free and very popular dashboarding platforms, what they have to offer and their pros and cons. For future reference, PowerBI is going to be used as the chosen dashboard platform. Finally, we did a literature review with a few articles that focus on similar problems (road safety/incidents) and described their specified problems, the analysis and the results of each study.

Chapter 3

Data Understanding and Exploration

The first phase of the CRISP-DM methodology is to understand the problem well by studying it and the objectives of the project. This has already been done in the state of the art chapter, where many problems, researches, data and ways to solve challenges have been addressed. So, logically, the next step to take is to collect, visualize and understand data that is gathered from ANSR reports and from TOMTOM API that has been collected continuously since 2018 from the research team at the Synthetic Intelligence Laboratory [38]. These data are related to the year 2018 and the district of Braga. This district is going to be used as a case study, and its process will be analogous to other districts. Next, the data from those reports are going to be presented and interpreted to deduce some initial information about them and later confirm or change those conclusions according to the work developed later.

In the ANSR report of 2018, in Braga district, there are several tables and graphs exposing information about road accidents and many aspects regarding those accidents such as time, weather, localization, county, type of road, etc, for that year and also how have those numbers (accidents, victims and fatalities) been evolving for the past few years. This data is available at the ANSR website (`http://www.ansr.pt/Estatisticas/RelatoriosDeSinistralidade/Pages/default.aspx`) in pdf format. To extract them, an excel tool was used to identify tables in pdf files and import them to excel to save in CSV files. Next, are going to be presented all the information present in the ANSR report of 2018 in Braga.

3.1 Historic Data

Year	Accidents w/ victims	Accidents w/ fatalities	Accidents w/ fatalities and/or serious injuries	Severity index
2009	2854	49	207	1.9
2010	2893	54	220	1.9
2011	2753	47	216	1.7
2012	2669	41	190	1.6
2013	2706	35	179	1.3
2014	2721	28	214	1.0
2015	2881	31	202	1.1
2016	2807	28	166	1.0
2017	3062	29	155	1.0
2018	3139	29	152	0.9

Table 3.1: Number of accidents with victims from 2009 to 2018 from [15]

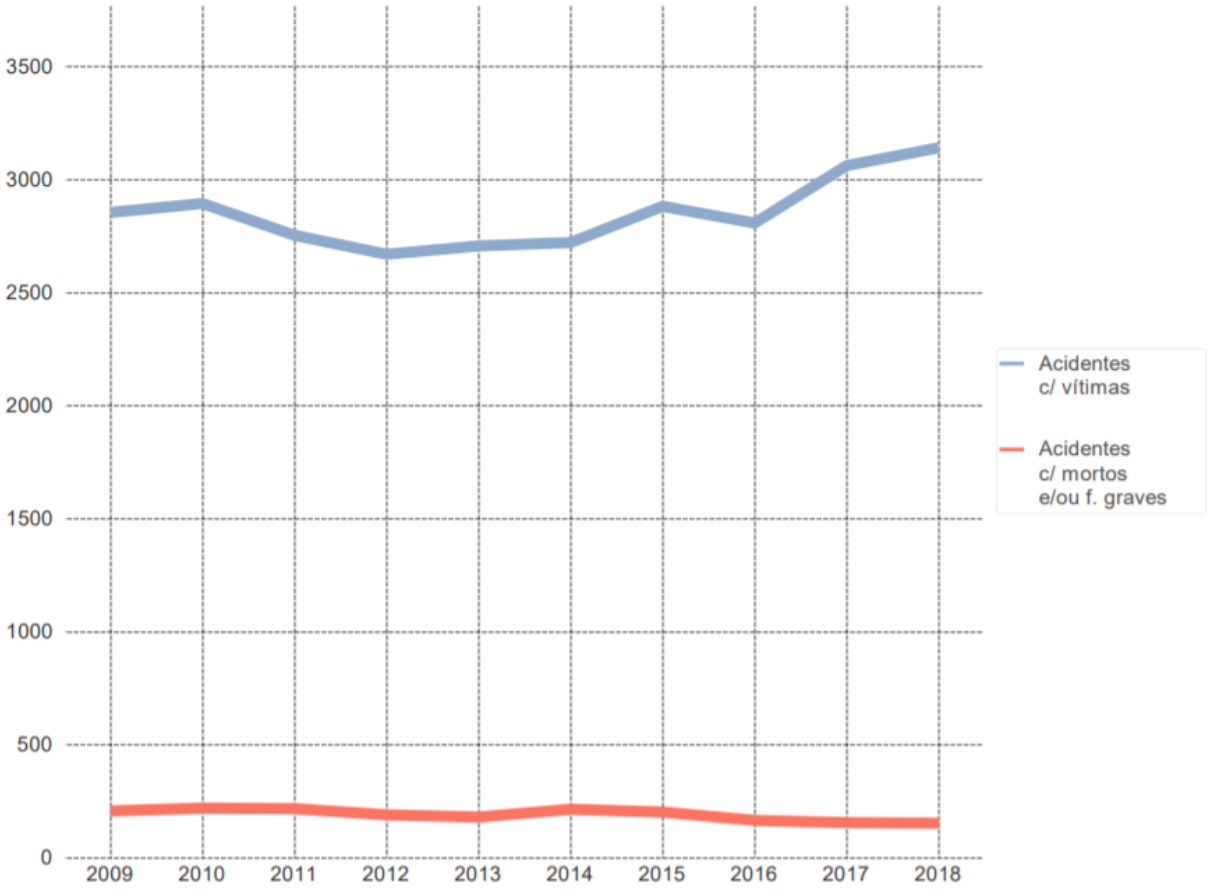


Figure 3.1: Evolution curve of accidents with victims, fatalities and/or serious injuries from 2009 to 2018. Adapted from [15]

Here we can see how the number of victims evolved from 2009 up to 2018. From the first Table 3.1, we see a decrease in the number of victims until 2012 and then a gradual increase until 2018 (first column). Concerning accidents with fatalities (second column), these decreased significantly until 2014 and somewhat stagnated until 2018. When talking about accidents with serious injuries adding the fatalities when applicable (third column), the conversation is a bit different. We can see a gradual decrease in this parameter until 2013 and increased significantly in 2014 while decreasing in further years. Finally, we observe a decrease in the severity index (fourth column) until 2014 and a stagnation once again until 2018, much like the evolution from accidents with fatalities. The severity index is calculated through the following equation:

$$Severityindex = 100 * Fatalities + 10 * SeriousInjuries + 3 * LightInjuries \tag{3.1}$$

From Figure 3.1 we can see a comparison between accidents with victims and accidents with fatalities and/or serious injuries, blue and red lines, respectively. The figure shows a red line relatively stable (relatively because the scale is somewhat large) in the time considered, and a pretty unstable blue line with an increase in 2010, a decrease until 2012, an increase again until 2015, a decrease in 2016 and finally a big increase until 2018, reaching the highest value registered in this period.

Year	Fatal victims	Serious injuries	Light injuries	Total victims
2009	55	197	3735	3987
2010	56	201	3755	4012
2011	47	201	3421	3669
2012	44	170	3424	3638
2013	35	163	3364	3562
2014	28	204	3370	3602
2015	31	192	3577	3800
2016	28	158	3486	3672
2017	30	143	3822	3995
2018	29	137	3924	4090

Table 3.2: Number of victims from 2009 to 2018 from [15]

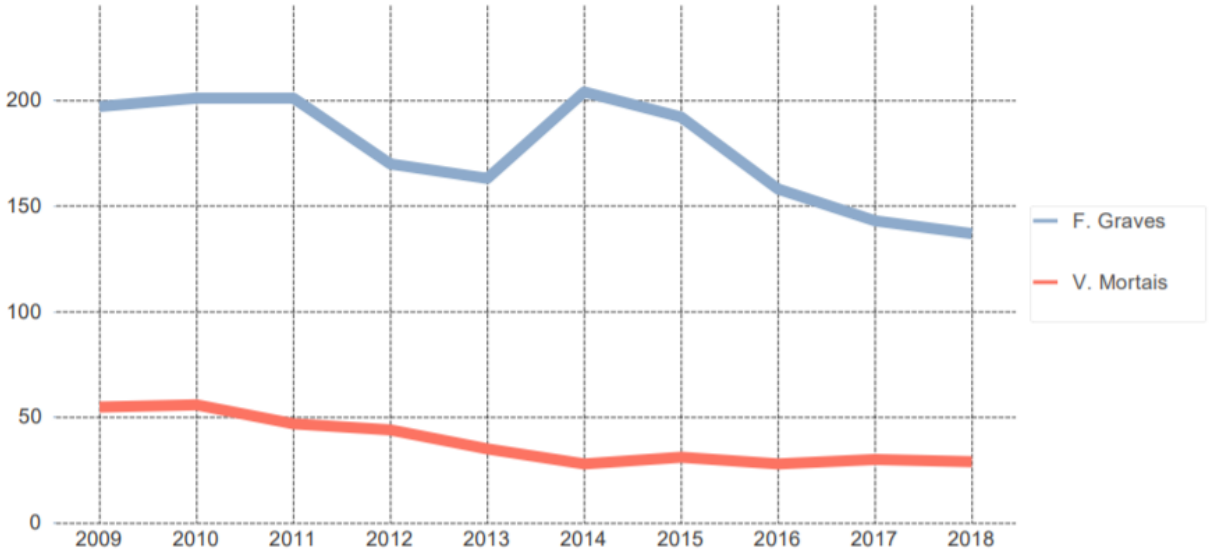


Figure 3.2: Evolution curve of the number of victims from 2009 to 2018. Adapted from [15]

Once again, Table 3.2 and Figure 3.2 are presented, showing the number of different types of victims from 2009 to 2018 and the evolution of seriously injured victims and fatalities in this same period, respectively. In Table 3.2 we see that fatalities (first column) and serious injuries (second column) have a similar evolution to their respective field in the previous Table 3.1. As for the light injured victims (third column), we can see a decrease until 2014, an up and down tendency in 2015 and 2016 and then a massive increase in 2017 and 2018, reaching the highest values registered in this period. As for the total victims (fourth column), the numbers fluctuate considerably, having many ups and downs but reaching, in 2018, the highest value registered. From Figure 3.2 we can see a comparison between the evolution of seriously injured victims and fatalities, in blue and red lines, respectively. The blue line appears to be well above the red line with a stagnate value until 2011 and decreasing ever since except for the year 2014 where a big increase happened, cancelling the decline up until that year. Such an event didn't happen in the red curve, where we see a somewhat steady decrease in the fatality numbers and stagnating from 2015 until 2018.

Generally speaking, by observing these tables and graphs, what jumps out the most is that, for the past few years, some numbers stopped dropping and remained constant, which should not happen since they're still substantial and should be dropping with the technological advance. Another aspect worth mentioning is that some values had a big spike in 2014, 2017 and 2018. The latter is concerning since it's a more recent date and integrates the focus of this study.

3.2 Temporal Analysis

	Accidents w/ victims	%	Fatal victims	%	Serious injuries	%	Light injuries	%	Total victims	%	Severity index
Jan	234	7.5	6	20.7	4	2.9	284	7.2	294	7.2	2.6
Feb	231	7.4	2	6.9	6	4.4	291	7.4	299	7.3	0.9
Mar	229	7.3	1	3.4	9	6.6	281	7.2	291	7.1	0.4
Apr	255	8.1	0	0.0	14	10.2	307	7.8	321	7.8	0.0
May	253	8.1	1	3.4	9	6.6	343	8.7	353	8.6	0.4
Jun	255	8.1	2	6.9	6	4.4	312	8.0	320	7.8	0.8
Jul	246	7.8	2	6.9	18	13.1	307	7.8	327	8.0	0.8
Aug	291	9.3	3	10.3	14	10.2	389	9.9	406	9.9	1.0
Sep	263	8.4	2	6.9	18	13.1	335	8.5	355	8.7	0.8
Oct	291	9.3	4	13.8	14	10.2	357	9.1	375	9.2	1.4
Nov	276	8.8	5	17.2	11	8.0	319	8.1	335	8.2	1.8
Dec	315	10.0	1	3.4	14	10.2	399	10.2	414	10.1	0.3
	3139	100	29	100	137	100	3924	100	4090	100	0.9

Table 3.3: Accidents and victims by month from [15]

Table 3.3 shows the number of accidents and victims that occurred in each month of 2018 together with their percentage from the total of that year. From the number of accidents with victims (first column), we can see higher values in August, October and December and a bit lower in November and September. The high value in August and December could be explained by the fact that many people enjoy vacations in August so it's expected a higher traffic volume in that month and December includes Christmas where many people visit their family, increasing once again the traffic volume combined with the winter season where the weather tends to worsen the safety conditions on the road. Concerning fatalities (third column), we note that January has the most number of fatalities racking up a staggering 20.7%, followed up by November and October, second and third with 17.2 and 13.8%, respectively. In the serious injuries (fifth column), we see that the second semester has the highest value of this category, except for November, summing up to 65% of the total serious injuries which is very close to 2/3 of the total value. The light injuries (seventh column) have similar percentages with accidents with victims since August, October and December have the highest values with the major differences being November and May, where the percentages seem to be somewhat higher than the respective percentages in accidents with victims. Summing the number of all types of victims (ninth column) we see that the percentages are very much alike the percentages from the light injury victims since the other numbers are so few that they don't affect the value and the percentage. Finally, we observe the severity index (eleventh column), calculated from equation 3.1, where the month of January has a value of 2.6 way above every other value and the months

of October and November with severity indexes above 1.

	Accidents w/ victims	%	Fatal victims	%	Serious injuries	%	Light injuries	%	Total victims	%	Severity index
Mon	464	14.8	1	3.4	15	10.9	574	14.6	590	14.4	0.2
Tue	433	14.1	6	20.7	12	8.8	549	14.0	567	13.9	1.4
Wed	455	14.5	1	3.4	23	16.8	564	14.4	588	14.4	0.2
Thu	410	13.1	2	6.9	13	9.5	503	12.8	518	12.7	0.5
Fri	539	17.2	7	24.1	22	16.1	645	16.4	674	16.5	1.3
Sat	462	14.7	7	24.1	38	27.7	577	14.7	622	15.2	1.5
Sun	366	11.7	5	17.2	14	10.2	512	13.0	531	13.0	1.4
	3139	100	29	100	137	100	3924	100	4090	100	0.9

Table 3.4: Accidents and victims by day of the week from [15]

In Table 3.4 we can see that the number of accidents with victims (first column) is similar except for Friday and Sunday, where the first has a higher value, probably due to the fact of being the end of the week where many people finish their job for the week and get to come home, sometimes tired and eager to rest and the second has a lower value than the other since Sunday is traditionally the resting day in Portugal where fewer people come out of their houses and so less traffic happens on the road. When it comes to fatalities (third column) we see that Monday, Wednesday and Thursday have a low value making only around 14% of the total while the other four days make up to around 86% of the total value. It's a major difference where the most surprising value is on Tuesday since the other three days include Friday and the weekend, so it appears to be somewhat random in comparison with the other three. In the serious injuries (fifth column), we see that Saturday has the highest number of them, followed by Wednesday and Friday, while the others tend to be around 9 and 10%. The light injuries (seventh column) has a similar distribution through all days of the week with some small differences, namely Thursday and Friday, that divert a bit from the norm that is 14%. Analogously with the previous table, the number of total victims follows a very similar distribution to the light injuries since they represent the biggest part of the total victims. The severity index (eleventh column) shows that four days have an index higher than 1 being Tuesday, Friday, Saturday and Sunday, mainly because they're the days of the week where most deaths happen, so they have a higher impact on this index than other days.

	Accidents w/ victims	%	Fatal victims	%	Serious injuries	%	Light injuries	%	Total victims	%	Severity index
00-03	98	3.1	0	0.0	12	8.8	143	3.6	155	3.8	0.0
03-06	59	1.9	1	3.4	6	4.4	82	2.1	89	2.2	1.7
06-09	365	11.6	4	13.8	12	8.8	443	11.3	459	11.2	1.1
09-12	492	15.7	3	10.3	17	12.4	572	14.6	592	14.5	0.6
12-15	607	19.3	5	17.2	13	9.5	742	18.9	760	18.6	0.8
15-18	597	19.0	5	17.2	27	19.7	779	19.9	811	19.8	0.8
18-21	639	20.4	10	34.5	34	24.8	801	20.4	845	20.7	1.6
21-24	282	9.0	1	3.4	16	11.7	362	9.2	379	9.3	0.4
	3139	100	29	100	137	100	3924	100	4090	100	0.9

Table 3.5: Accidents and victims by hour span from [15]

In Table 3.5 we can see how the accidents and victims are distributed among time spans. From 0 to 6 AM and from 9 to 12 PM are the time spans where fewer accidents and victims happen in comparison with the rest of the day. We can note that mainly in accidents, fatalities, light injuries and total victims. In the serious injury category (fifth column), we can also see that from 12 AM to 3 PM has similar values with the time spans mentioned before and from 9 to 12 PM has a bit higher value than the general tendency. In the severity index (last column), the highest values are actually between 3 and 6 AM and 6 and 9 PM, followed by the 6 to 9 AM period, which doesn't follow the general tendency of this table. The higher values registered between 6 and 9 PM can be explained by the fact that it includes the rush hour where almost all workers get out of their job and massively increase traffic volume. The lower values registered from 9 PM to 6 AM can be explained by the fact that people are usually at home during that time, so fewer accidents happen and fewer victims are registered since traffic volume is much lower than at other times.

3.3 Accident Location and Context

	Accidents w/ victims	%	Fatal victims	%	Serious injuries	%	Light injuries	%	Total victims	%	Severity index
Day	2285	72.8	20	69.0	79	57.7	2824	72.0	2923	71.5	0.9
Night	776	24.7	7	24.1	54	39.4	1009	25.7	1070	26.2	0.9
Dawn/Dusk	78	2.5	2	6.9	4	2.9	91	2.3	97	2.4	2.6
	3139	100	29	100	137	100	3924	100	4090	100	0.9

Table 3.6: Accidents and victims by brightness from [15]

Table 3.6 shows a higher number of every aspect when they happen during the day than during the night or the dusk/dawn combined, which is expected since the majority of people move around in the daytime, even though the visibility conditions are worse at night or dusk/dawn. The only thing that stands out here besides the point that has already been mentioned, is the fact that the severity index at dusk/dawn is way above the other two times of the day, possibly because people that drive during this time are probably tired or sleepy which can lead to accidents being more serious.

	Accidents w/ victims	%	Fatal victims	%	Serious injuries	%	Light injuries	%	Total victims	%	Severity index
Clear	2286	72.8	22	75.9	107	78.1	2847	72.6	2976	72.8	1.0
Rainy	831	26.4	7	24.1	29	21.2	1048	26.7	1084	26.5	0.8
Other	17	0.5	0	0.0	1	0.7	23	0.6	24	0.6	0.0
N.D.	5	0.2	0	0.0	0	0.0	6	0.2	6	0.1	0.0
	3139	100	29	100	137	100	3924	100	4090	100	0.9

Table 3.7: Accidents and victims by weather from [15]

As we can see from Table 3.7, all the categories in good weather represent around 75% of the total values while rainy represents around 25% of the total values which, can be a bit contradictory since it's expected to have higher values when weather conditions are worse. This may mean that people are more careful when driving around with unfavourable climates or that adverse weather may happen fewer times than clear weather which, may explain higher values in clear weather conditions. Since these are the most common weather conditions in this country and district, it explains why other types of weather have so lower numbers in comparison.

		Accidents w/ victims	%	Fatal victims	%	Serious injuries	%	Light injuries	%	Total victims	%	Severity index
Run over	Hit and run	30	1.0	1	3.4	2	1.5	29	0.7	32	0.8	3.3
	Run over animals	11	0.4	0	0.0	1	0.7	11	0.3	12	0.3	0.0
	Run over pedestrians	425	13.5	5	17.2	33	24.1	415	10.6	453	11.1	1.2
Total		466	14.8	6	20.7	36	26.3	455	11.6	497	12.2	1.3
Collision	Chain collision	54	1.7	1	3.4	0	0.0	92	2.3	93	2.3	1.9
	Hit and run	42	1.3	0	0.0	0	0.0	49	1.2	49	1.2	0.0
	Other collisions	136	4.3	0	0.0	4	2.9	179	4.6	183	4.5	0.0
	Vehicle or obstacle collision	72	2.3	0	0.0	6	4.4	96	2.4	102	2.5	0.0
	Frontal collision	492	15.7	6	20.7	23	16.8	771	19.6	800	19.6	1.2
	Lateral collision w/vehicle	530	16.9	4	13.8	16	11.7	680	17.3	700	17.1	0.8
	Rear collision w/vehicle	382	12.2	2	6.9	3	2.2	508	12.9	513	12.5	0.5
Total		1708	54.4	13	44.8	52	38.0	2375	60.5	2440	59.7	0.8
Skid	Skid w/ rollover	125	4.0	4	13.8	3	2.2	161	4.1	168	4.1	3.2
	Skid w/ collision	66	2.1	2	6.9	4	2.9	72	1.8	78	1.9	3.0
	Skid w/ retention mechanisms	44	1.4	0	0.0	0	0.0	50	1.3	50	1.2	0.0
	Skid and getaway	3	0.1	0	0.0	0	0.0	3	0.1	3	0.1	0.0
	Skid w/ lateral road transposal	19	0.6	0	0.0	4	2.9	27	0.7	31	0.8	0.0
	Skid w/o retention mechanisms	69	2.2	0	0.0	6	4.4	70	1.8	76	1.9	0.0
	Simple skid	639	20.4	4	13.8	32	23.4	711	18.1	747	18.3	0.6
Total		965	30.7	10	34.5	49	35.8	1094	27.9	1153	28.2	1.0
Aggregate total		3139	100	29	100	137	100	3924	100	4090	100	0.9

Table 3.8: Accidents and victims by nature of the accident from [15]

Table 3.8 is a bit more complex than the others since it's grouped by three labels, namely run over,

collision and skid. From the run over group, we see that the highest values belong to running over pedestrians representing close to the total value while hit and runs and running over animals have much lower values of accidents and victims. One thing to notice is that hit and runs have a severity index much higher than the other two, with an astonishing 3.3.

From the collision group, when talking about accidents with victims, we can see that front collisions, lateral collisions with a moving vehicle and rear collision with a moving vehicle have the highest values, while chain collisions, hit and run, other collisions and collisions with vehicle or obstacle have rather low values and only makeup to around 18% of the situations. In the fatalities category, the distribution of values and percentages is similar. In the serious injuries category, we can see a slight difference, namely in the rear collision with a moving vehicle, where the number and percentage are rather low, similar to the other types of collisions with low values. Besides that point, the distribution of values seems to be similar. The light injuries and total victims follow the same output as the accidents with victims and fatalities columns, high values and percentages in front collisions, lateral collisions with a moving vehicle and rear collision with a moving vehicle and low values and percentages in chain collisions, collision and run, collision with other objects and collision with vehicle or obstacle in the traffic lane. The severity index doesn't follow this tendency since the only ones above 1 are chain collisions and front collisions. This is fitting for the descriptions of these collisions since chain collisions assume a multiple car crash, so when they happen, they tend to bring a worse scenario than other types of collisions and front collisions where it's widely known that these collisions generate more force, resulting in more severe crashes than other types of collisions whether they are to the rear, lateral, or immobile object or vehicle.

In the skid group, we can see that a simple skid is the main cause of accidents with victims among skids, while the other six combined represent 1/3 of the skids. In the fatalities category, simple skids and skids with rollovers represent the most with four each, followed by skids with collision with 2, while the others didn't have any fatalities associated. In the serious injuries, light injuries and total victims columns, we can see similar percentages with the accidents with victims column where simple skids represent the biggest piece, again, with 2/3 of the total. Finally, the severity index appears to be high in skids with rollovers and skids with a collision with 3.2 and 3, respectively, which means these are very susceptible to the occurrence of serious events. Also, among the three groups, even though run over accidents happens less frequently than collisions or skids, they appear to be a little more severe due to the severity index values of the three groups.

	Accidents w/ victims	%	Fatal victims	%	Serious injuries	%	Light injuries	%	Total victims	%	Severity index
Inside localities	2930	93.3	24	82.8	127	92.7	3631	92.5	3782	92.5	0.8
Outside localities	209	6.7	5	17.2	10	7.3	293	7.5	308	7.5	2.4
	3139	100	29	100	137	100	3924	100	4090	100	0.9

Table 3.9: Accidents and victims inside and outside of localities from [15]

In Table 3.9, we see a tendency where the majority of accidents and victims happen inside localities (82.8% in fatalities and around 93% in the other categories). Even though there's a similar distribution of numbers and percentages, the severity index shows that, outside localities, this index is much higher than inside localities, which means that, when accidents happen outside localities, they tend to be more serious than when they happen inside localities even though the difference in numbers is considerably big.

	Accidents w/ victims	%	Fatal victims	%	Serious injuries	%	Light injuries	%	Total victims	%	Severity index
Street	1963	62.5	10	34.5	88	64.2	2340	59.6	2438	59.6	0.5
Highway	86	2.7	2	6.9	5	3.6	130	3.3	137	3.3	2.3
Municipal road	91	2.9	2	6.9	4	2.9	107	2.7	113	2.8	2.2
National road	856	27.3	15	51.7	37	27.0	1167	29.7	1219	29.8	1.8
Other road	143	4.6	0	0.0	3	2.2	180	4.6	183	4.5	0.0
	3139	100	29	100	137	100	3924	100	4090	100	0.9

Table 3.10: Accidents and victims by road type from [15]

Table 3.10 shows how accidents and victims are distributed among road types. The majority of accidents with victims happen in streets (62.5%) and national roads (27.3%) while the other three only represent around 10% of the total.

In the fatalities category, the data shows a bit different values since more people die on national roads than streets, while the highway and municipal roads have 6,9% each and none are registered in other routes (regional roads, forest roads, bridges, variant roads, etc). In the three remaining victim categories, we can see that the percentages are very much alike, even though the number is different, they are proportional,

with streets representing more than half and national roads representing around 30% of the total. In the severity index, we notice that highways and municipal roads have the highest indexes, followed closely by national roads. Freeways have the highest SI probably because highways have the highest limit speed of all the road types, which can lead to more serious accidents, even though they happen less frequently. Municipal roads also have a very high SI, probably due to these roads usually having poor conditions since they're considered secondary roads that connect other more important roads.

	Accidents w/ victims	%	Fatal victims	%	Serious injuries	%	Light injuries	%	Total victims	%	Severity index
Amares	58	1.8	0	0.0	4	2.9	76	1.9	80	2.0	0.0
Barcelos	441	14.0	6	20.7	13	9.5	562	14.3	581	14.2	1.4
Braga	685	21.8	7	24.1	32	23.4	821	20.9	860	21.0	1.0
Cabeceiras de Basto	44	1.4	0	0.0	0	0.0	58	1.5	58	1.4	0.0
Celorico de Basto	60	1.9	1	3.4	5	3.6	73	1.9	79	1.9	1.7
Esposende	139	4.4	1	3.4	6	4.4	157	4.0	164	4.0	0.7
Fafe	154	4.9	1	3.4	3	2.2	197	5.0	201	4.9	0.6
Guimarães	574	18.3	4	13.8	29	21.2	719	18.3	752	18.4	0.7
Póvoa de Lanhoso	112	3.6	1	3.4	6	4.4	145	3.7	152	3.7	0.9
Terras de Bouro	27	0.9	0	0.0	1	0.7	33	0.8	34	0.8	0.0
Vieira do Minho	34	1.1	1	3.4	4	2.9	35	0.9	40	1.0	2.9
V.N. de Famalicão	529	16.9	2	6.9	16	11.7	680	17.3	698	17.1	0.4
Vila Verde	197	6.3	5	17.2	16	11.7	259	6.6	280	6.8	2.5
Vizela	85	2.7	0	0.0	2	1.5	109	2.8	111	2.7	0.0
	3139	100	29	100	137	100	3924	100	4090	100	0.9

Table 3.11: Accidents and victims by each county from [15]

In Table 3.11, we can see that the number of accidents that happen in each county is generally proportional to its population, according to the numbers in PORDATA (Portuguese contemporary database) [39].

In the fatalities category, we notice two significant differences which are *Vila Nova de Famalicão* and *Vila Verde*. The first has a low number of fatalities when compared to its and other counties accidents

value, the second has a surprising increase in the percentage of fatalities when compared to its and other counties accidents percentage. In the serious injuries category, we notice a significant decrease in *Barcelos* and *Vila Nova de Famalicão* while *Braga* and *Guimarães* still have the highest percentages, in part due to the number of citizens, and *Vila Verde* having a significant increase once again in that percentage. The total victims follow the accidents tendency where we see that counties with more population have a higher percentage of victims. In the severity index, we have a whole different conversation. *Vieira do Minho* has the highest SI out of all the counties with 2.9, followed by *Vila Verde* with 2.5, then *Celorico de Basto* with 1.7 and after that *Barcelos* with 1.4 while the others have a SI of one or below.

3.4 People and Vehicles Involved

	Fatal victims	%	Serious injuries	%	Light injuries	%	Total victims	%
Drivers	22	75.9	83	60.6	2519	64.2	2624	64.2
Passengers	1	3.4	17	12.4	948	24.2	966	23.6
Pedestrians	6	20.7	37	27.0	457	11.6	500	12.2
	29	100	137	100	3924	100	4090	100

Table 3.12: Victims by user category from [15]

Table 3.12 shows that drivers represent the major part of fatalities with 75.9%, followed by pedestrians with 20.7% and then passengers with 3.4%. The seriously injured also show the same order but with a decline in drivers to 60.6% and an increase to the other two user categories, passengers and pedestrians. In the lightly injured and total victims categories, we see that this order is a bit flipped since drivers still represent the majority with 64.2%, but here passengers represent more than pedestrians with 24.2% and 11.6% in light injuries and 23.6% and 12.2% in total victims, each respectively.

	Fatal victims	%	Serious injuries	%	Light injuries	%	Total victims	%
Pedestrians	6	20.7	37	27.0	457	11.6	500	12.2
Light vehicle	10	34.5	47	34.3	2627	66.9	2684	65.6
Heavy vehicle	0	0.0	2	1.5	31	0.8	33	0.8
Velocipedes	0	0.0	6	4.4	144	3.7	150	3.7
Moped	2	6.9	9	6.6	238	6.1	249	6.1
Motorcycle	6	20.7	34	24.8	389	9.9	429	10.5
Others	5	17.2	2	1.5	37	0.9	44	1.1
N.D.	0	0.0	0	0.0	1	0.0	1	0.0
	29	100	137	100	3924	100	4090	100

Table 3.13: Victims by vehicle category from [15]

One thing to have into account is that all types of vehicles present in Table 3.13 include drivers and passengers. Light cars represent the most number of fatalities, followed by passengers and motorcycles with the same value, then “others” and finally mopeds. In the serious injuries, we notice that, surprisingly, “others” has a lower value than in fatalities with a huge percentage decrease while all remaining vehicle types raising their percentage or maintaining it. Light cars represent the majority of light injuries with 66.9% and pedestrians 11.6% while others are all under 10%. We notice that the major changes about serious injuries are light cars (big percentage increase), pedestrians (big percentage decrease) and “others” (big percentage decrease). The percentage of total victims also follow the light injuries values and tendencies since, as we already explained in some of the tables, they represent almost the totality of victims (in this case, around 96%). Light cars have the highest percentage in all categories, above motorcycles and mopeds, which are unquestionably more dangerous to ride, because these kinds of vehicles are way more abundant in Portuguese roads, according to data available in PORDATA [40].

	Fatal victims	%	Serious injuries	%	Light injuries	%	Total victims	%
<=14	0	0.0	5	3.6	237	6.0	242	5.9
15-19	1	3.4	9	6.6	284	7.2	294	7.2
20-24	5	17.2	14	10.2	467	11.9	486	11.9
25-29	1	3.4	13	9.5	342	8.7	356	8.7
30-34	2	6.9	9	6.6	293	7.5	304	7.4
35-39	0	0.0	6	4.4	276	7.0	282	6.9
40-44	2	6.9	20	14.6	318	8.1	340	8.3
45-49	5	17.2	7	5.1	349	8.9	361	8.8
50-54	4	13.8	13	9.5	305	7.8	322	7.9
55-59	0	0.0	7	5.1	275	7.0	282	6.9
60-64	2	6.9	10	7.3	227	5.8	239	5.8
65-69	3	10.3	8	5.8	182	4.6	193	4.7
70-74	0	0.0	9	6.6	161	4.1	170	4.2
>=75	4	13.8	7	5.1	208	5.3	219	5.4
	29	100	137	100	3924	100	4090	100

Table 3.14: Victims by age group from [15]

In Table 3.14, we notice that some age groups are more susceptible to die on the road, namely 20 to 24, 45 to 54, 65 to 69 and over 74, since they have higher mortality rates. The serious injuries have a somewhat even distribution with some age groups having a slight difference which is 20 to 29, 40 to 44 and 50 to 54 because their percentage is a bit higher than 9% while the rest fluctuate around 4 and 7%. Again, in the light injuries, we notice that only the 20 to 24 age group has a percentage higher than 10% while the other age groups have a more or less similar distribution. The total victims' category follows the same pattern as the light injuries category. One thing to take note of is that the numbers in this table 3.14 show that the age group of 20 to 24 has worrying high numbers, having the three highest percentages in those four categories so efforts should be made to lower these values, especially, for this young age range.

3.5 Incidents Overview

The ANSR report that we have been discussing also includes a list of accidents that have recorded fatalities and/or serious injuries that include county, date, time, number of deaths, serious injuries, road, at what kilometre of that road the accident happened and a description of the nature of the accident. Since this list is extensive, it's not going to be present here, but is available at the report itself [15].

The data collected from TomTom is composed of 34736 entries of incidents that occurred in 2018 in the Braga district. It contains many features where the most relevant are location, description of traffic and accidents, cause of accidents, amount and magnitude of the delay caused, affected roads, latitude, longitude and date of the incident. It's worth mentioning that this data contains incidents recorded from 24/7/2018 to 31/12/2018. The dataset contains 22 features, 8 of which are integer, 2 of them are float and the remaining 12 are strings.

By analysing the missing values in each feature, we notice that almost all the values in the cause of accident feature are missing (34715 in 34736 entries), and nearly half of the affected roads feature are missing as well, while all the others have no missing values. Regarding existing values, there are some features worth mentioning like the description of traffic, which can be queuing traffic, stationary traffic, slow traffic, closed, bridge closed, roadworks, heavy rain and traffic flow freely. Others include incident category description, which can be jam, road closed, road works, rain and dangerous condition, magnitude delay that can go from 0 to 4 and its description as major, moderate, minor, undefined or unknown delay.

Some features are worth taking a look at their distributions as well as some statistics, namely magnitude of delay, length of delay and time of delay. The first has a minimum and maximum value of 0 and 4, respectively, a mean around 3.24, which is quite close to the maximum value, a median of 3, and follows a direct proportional distribution (growing number of events when growing magnitude). The second one has a minimum and maximum value of 30 and 7315, respectively, a mean of 337.02, a median of 210, and follows an inverse proportional distribution (the lower the length, the higher number of events). The third one has a minimum and maximum value of 0 and 3605, respectively, a mean of 123.15, a median of 110, and follows an inverse proportional distribution. Other features have a descriptive function like ids, road names, latitude, longitude and dates, so it doesn't make much sense to analyse their values or see their statistics because they won't have a particular meaning.

Chapter 4

Data Pre-Processing and Analysis

4.1 Data Pre-Processing

For the more practical part of this thesis, Python is going to be used as the programming language because it's the most used language in data science projects and continues to grow in popularity in this and other areas such as machine learning and scientific computing in general [41]. It offers a wide array of useful libraries, simplifying data reading, interpretation, processing, modelling, among others.

As it was said before, the data were stored in CSV files so they could easily be imported to a python notebook. After the analysis done in the previous chapter, the first thing to do was to remove features that contained a high number of missing values as well as rows. In the tomtom data, since the cause of accident feature had nearly 99% of missing values, it was removed from the dataset. Furthermore, another problematic feature was the affected roads since it had around 50% missing values, but here we can assume that missing values mean that no other roads were affected, so it's going to stay in the dataset. No features were removed from the ANSR dataset. Besides features, rows were analysed after removing the previous feature, and none were discarded from either dataset since the affected roads feature was the only one causing missing values, which are going to be addressed later. One very important thing that was missing was detailed timestamps containing information about weekends, special dates and weekday together with date and time. The solution was to first separate the incident date feature into features, such as month, day, hour, minute and second, and then add the remaining features according to each specific date. Weekdays and weekends were fairly easy to add due to the DateTime library, which enabled the inclusion of weekdays given year, month, day and consequently weekend. As for the special dates, they had to be researched and added manually with conditional statements since they are different every year and differ from country to country. Next are going to be presented all the special dates that were considered.

- 1st January - New Year's Day
- 6th January - King's Day
- 13th February - Carnival
- 14th February - Valentine's Day
- 30th, 31st and 1st March and April, respectively - Easter
- 25th April - Freedom Day
- 1st May - Worker's Day
- 13th May - Holy Mary apparitions to 3 young children
- 31st May - Feast of *Corpus Christi*
- 10th June - Portugal Day
- 13th June - St. Anthony's Day
- 24th June - St. John's Day
- 15th August - Assumption of Mary
- 5th October - Republic Day
- 1st November - All Saints Day
- 23rd November - Black Friday
- 1st December - Restoration of Independence
- 8th December - Immaculate Conception
- 24th December - Christmas Eve
- 25th December - Christmas Day
- 31st December - New Year's Eve

Besides detailing time and data, information about roads was also specified, namely the road type where the accident happened in both datasets, dividing the instances of affected roads into different features and specifying the number of affected roads in the tomtom dataset. For the first one, regular expressions were used since road abbreviations in Portugal give information about its type, for example,

N14 stands for National Road 14 or A3 stands for Highway 3. For the second one, a string splitter was used since the affected roads caused by an accident were combined in a single feature separated by slashes or hyphens. For the third and last one, all different roads were counted in each instance and then inserted in the corresponding instance in a distinct feature. Table 4.1 shows all features existing in the tomtom dataset after processing, while Table 4.2 shows all features existing in the ANSR dataset after processing.

#	Features	Description
1	description	traffic description
2	from_road	where the incident starts
3	to_road	where the incident ends
4	incident_category	incident numerical category
5	incident_category_desc	description of the incident category
6	magnitude_of_delay	magnitude of the delay
7	magnitude_of_delay_desc	description of the magnitude of the delay
8	length	traffic length
9	delay_in_seconds	traffic delay (in seconds)
10	latitude	incident latitude
11	longitude	incident longitude
12	weekday	day of the week
13	month	month of the year
14	day	day of the month
15	hour	hour of the day
16	minute	minutes
17	second	seconds
18	holidays	if it's a special date or not
19	weekend	if it's weekend or not
20	from_road_type	type of road where the incident starts
21	to_road_type	type of road where the incident ends
22	number_affected_roads	number of affected roads
23	affected_roads1	affected roads first split
24	affected_roads2	affected roads second split
25	affected_roads3	affected roads third split

Table 4.1: Tomtom dataset features after processing

#	Features	Description
1	Concelho	locality name
2	M	number of deaths
3	FG	number of serious injuries
4	Via	road name
5	Km	kilometer where the incident happened
6	Natureza	Nature of the incident
7	dia_semana	day of the week
8	mes	month of the year
9	dia	day of the month
10	hora	hour of the day
11	minuto	minutes
12	datas_especiais	if it's a special date or not
13	fim_de_semana	if it's weekend or not
14	Tipo_estrada	type of road

Table 4.2: ANSR dataset features after processing

4.2 Data Visualization

4.2.1 Tomtom dataset

After performing some data processing, it was necessary to visualize some of the features' characteristics, such as their distribution and how they are related to others. To do that, some python graphs were coded to reveal the information previously mentioned. When it comes to the tomtom dataset, two features were studied, namely traffic's delay in seconds and its description and were compared with some other relevant features. To start, the traffic delay was analyzed initially, with a box plot containing its median, 1st and 3rd quartiles and possible outliers.

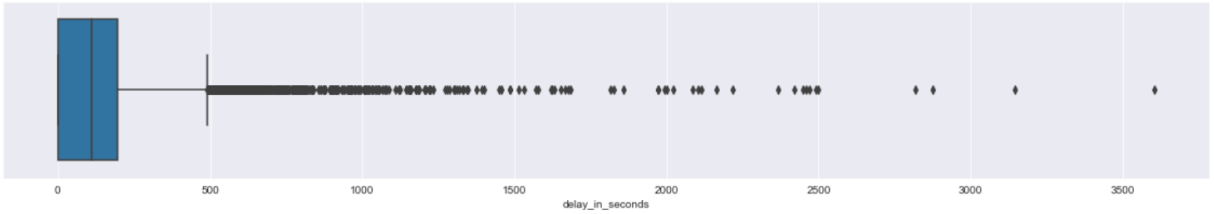


Figure 4.1: Traffic delay (seconds) feature

By looking at Figure 4.1, we can identify some outliers when the traffic delay is greater than 500. We can also observe that the box's minimum value and the 1st quartile have the same value, which is

zero, meaning the dataset contains a reasonable amount of events (accidents) with zero traffic delay. The removal of these outliers will be mentioned ahead in this chapter and the next figures are relative to data with these outliers. After this box plot, other graphs were made to show connections between the traffic delay and other important features that are going to be displayed next.

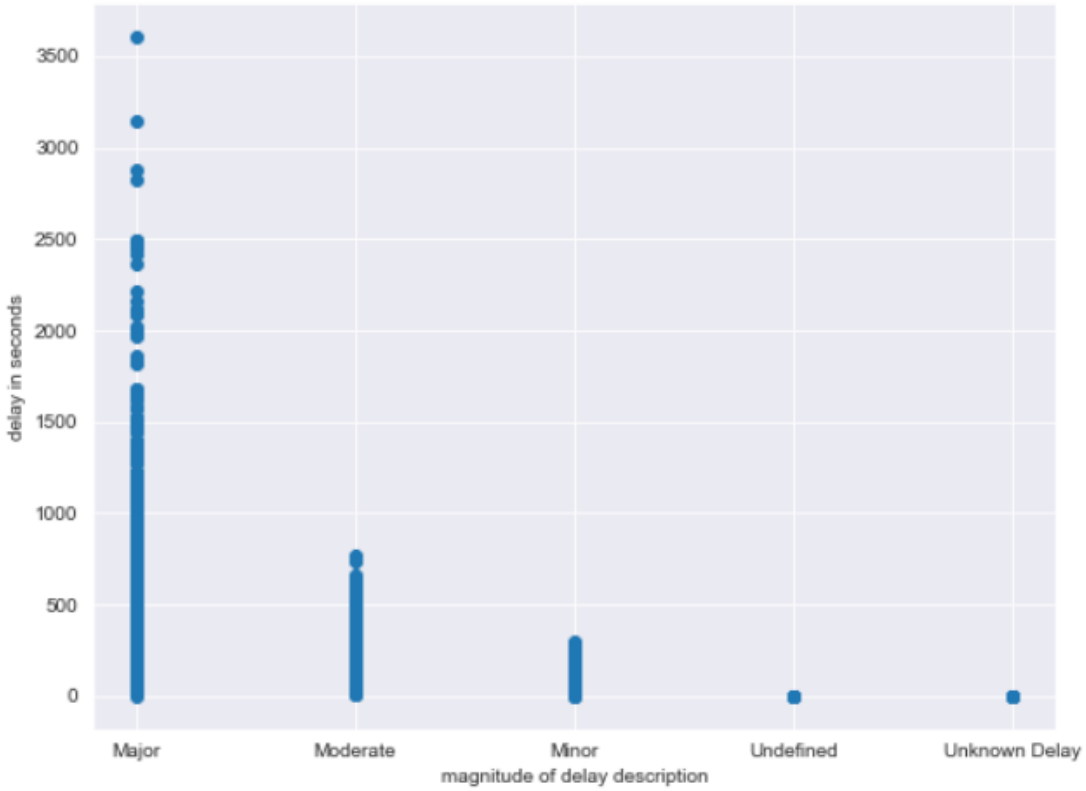


Figure 4.2: Relation between traffic delay (seconds) and magnitude of delay

In Figure 4.2, we can see that the major magnitude of delay contains more accidents with high traffics delays logically followed by moderate and then minor. We can also see two more categories available, namely undefined and unknown delay, having occurrences with zero traffic delay. The first one is due to exclusively including incident descriptions of roads closed which explains the value 0 in the respective figure. The second one is due to including only a few road works incident descriptions which their traffic delays couldn't be deduced or calculated successfully, so the value 0 isn't matching the reality of the situation but since having empty values in a dataset isn't a good practice it's left as it is.

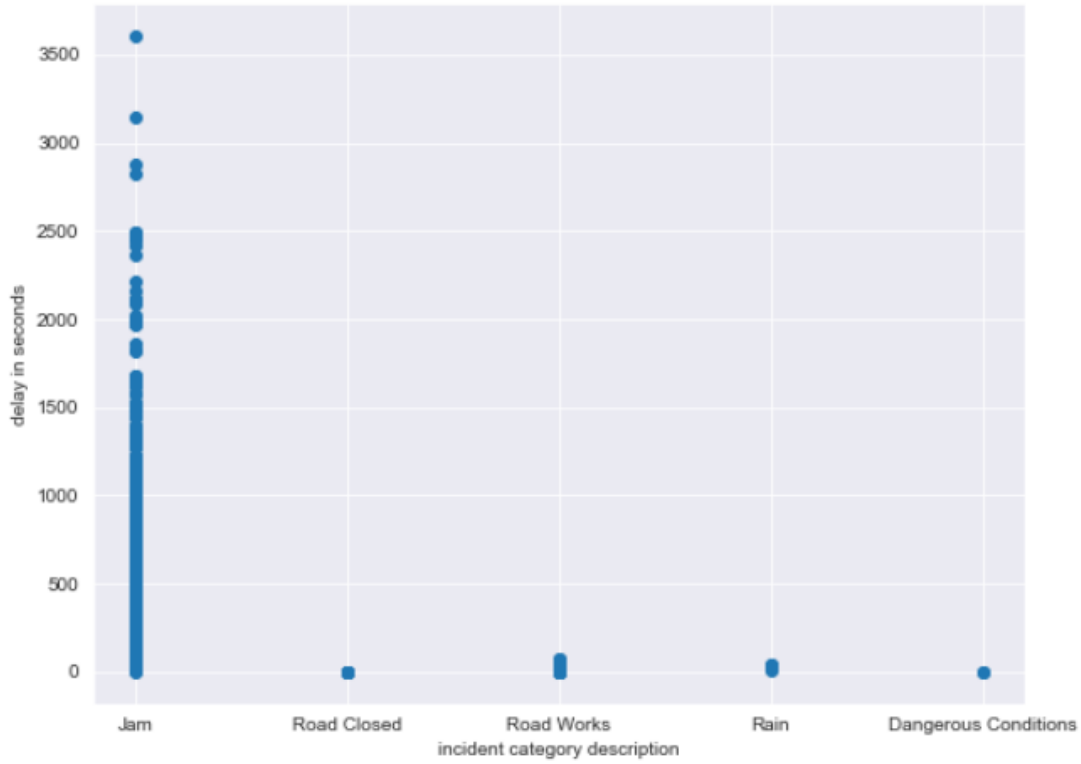


Figure 4.3: Relation between traffic delay (seconds) and incident category

In Figure 4.3, we can see that only the jam category in the incident category description has a diversity of traffic delay values ranging from 0 to up to 3500. Besides that, we can see the roads closed category that contains only traffic delays of zero seconds, as it is related to the explanation previously mentioned. Next, we have the road works category, which contains low values (from 0 to around 100 seconds) along with the rain category with a little shorter range and the dangerous conditions category with an even shorter range of traffic delay.

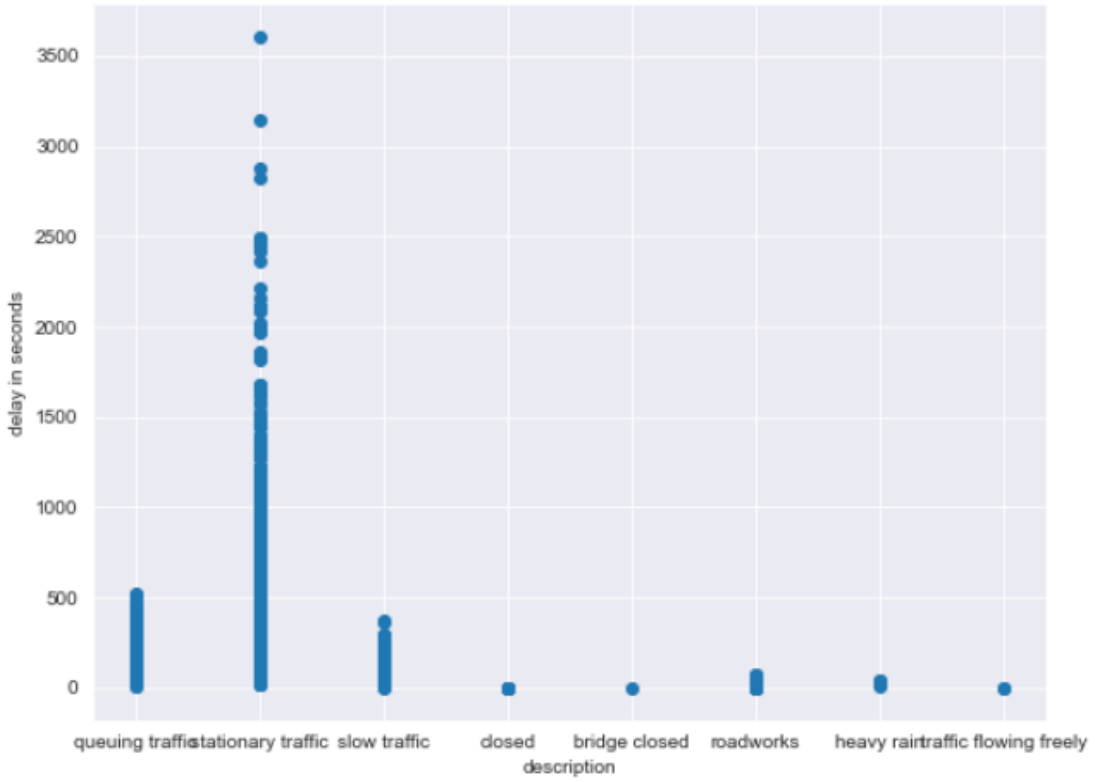


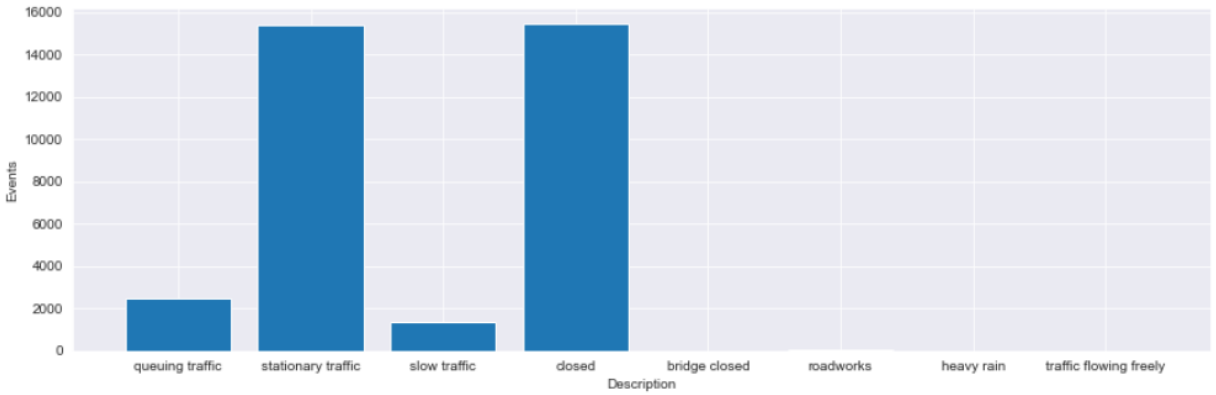
Figure 4.4: Relation between traffic delay (seconds) and traffic description

In Figure 4.4, stationary traffic has a wide range of time delays (from 0 to 3500 seconds) while all the others have much smaller ranges, ones bigger than others. Queuing traffic goes from 0 to 500 seconds, similar to slow traffic, while the other five have values close to zero except for bridge closed and closed (road closed), which have only zero seconds of traffic delay since there is no traffic at all. Then there're roadworks and heavy rain with short time delays and, logically, traffic flowing freely with almost no delay since, as the name says, it's flowing freely, so there are no special restrictions or restraints to traffic.

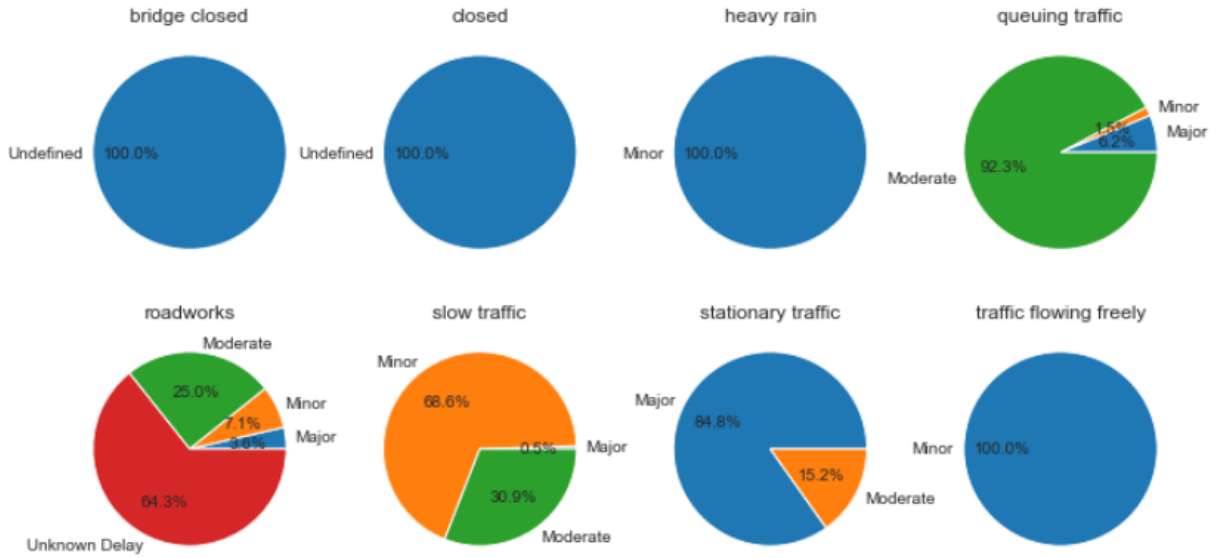
Figure 4.5: Relation between traffic delay (seconds) and incident length

In Figure 4.5, we can see the dispersion and correlation between the time delay and incident length and a concentration of incidents between 0 and 3000 meters (incident length) and 0 and 1000 seconds in traffic delay. There are also some other incidents besides the area mentioned, which are a minority and one lonely incident with around 7000 meters and 3500 seconds traffic delay, which doesn't seem appropriate to remain in the dataset given that it's way apart from the other incidents.

Besides all the information about the traffic delay, there are also relevant visualizations about the traffic description. This feature was analyzed, relating it with the number of incidents and the frequency of the magnitude of delay. First, a bar plot was used to represent the number of incidents that occurred for each traffic description, and several pie charts were utilized to show, for each traffic description, how frequent is the magnitude of the delay with percentages.



(a) Number of incidents for each traffic description



(b) Magnitude of delay distribution for each traffic description

Figure 4.6: Incident number and magnitude of delay for each traffic description

In Figure 4.6a, we quickly identify two traffic descriptions that are very common (around 15000 events) since a high number of incidents contain that descriptions which are stationary traffic and closed (closed road), followed by two other traffic descriptions that, although have a low number of events (around 2000 events), are still noticeable in Figure 4.6a. Besides these descriptions, four other traffic descriptions exist in the dataset but have a residual number of events.

In Figure 4.6b, we can see the different traffic descriptions with their respective pie chart containing the division of the delay magnitudes for each type of description. Reading the pie charts left to right and top to bottom, we can see in the first three figures that all of them have only one magnitude associated with the traffic description which is undefined for the first two and minor for the third. Since the first and second pie charts relate to closed paths, it's understandable that it only contains undefined magnitudes while the third one relates to rainy conditions, which for security reasons, results in vehicles slowing down, therefore resulting in 100% minor delay. In the fourth pie chart, we can see a distribution of 92.3% for moderate delay, 6.2% major delay and 1.5% minor delay. Since the description is queuing traffic, it implies accumulating

and waiting traffic while occasionally moving forward, so that's why the majority of magnitude is moderate, and there's a residual percentage of minor and major since there are always some exceptions. In the fifth pie chart, we can see that roadworks description has more than half unknown delay magnitude with 64.3% since it's related to changing roads' flow, closing lanes, introducing temporary traffic lights, among others, and due to roadworks it's many times hard to know the delay time and deduce the magnitude since those delays can be pretty inconsistent. Besides that, we have moderate delay with 25%, 7.1% for minor delay and 3.6% for major. In the sixth pie chart, we can see that slow traffic has more than half minor delay magnitude with 68.6%, moderate with 30.9% and major with 0.5%. Since slow traffic translates to a decrease of vehicles speed with little to no stops, which means that high delays are very uncommon, meaning that minor to moderate delays are logically more frequent. In the seventh pie chart, we can see that stationary traffic contains a majority of major delay magnitudes with 84.8% followed by moderate with 15.2%. This large percentage is easily explained by the fact that stationary traffic implies stopped vehicles in traffic, resulting in longer time delays supporting the pie chart in question. In the last pie chart, the only type of delay in freely flowing traffic description is minor since, as the name implies, vehicles are moving freely or at speeds close to or equal to regular traffic.

Several data records are considered outliers, and that can be seen clearly in Figure 4.1 where lots of dots are outside of what's considered normal by containing a time delay above the maximum displayed in the box plot. The method used to remove outliers was the Interquartile range (IQR) method and what it does is by using the first and third quartile and then calculating a lower and upper bound using those same quartiles to add them to IQR, which is the difference between Q3 and Q1 multiplied by 1.5. Then, after setting the lower and upper bound regarding the time delay feature, we remove the records that contain a value lower than the lower bound and higher than the upper bound. The resulting box plot is presented in Figure 4.7.

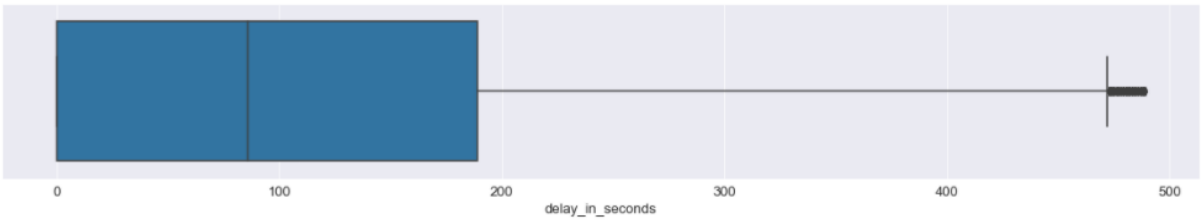


Figure 4.7: Box plot without outliers according to IQR method

4.2.2 ANSR dataset

Even though the ANSR dataset contained much fewer records than the tomtom dataset, there are still some relevant plots worth mentioning. First, there's the information about how many deaths and serious injuries occur on weekdays and the weekend.

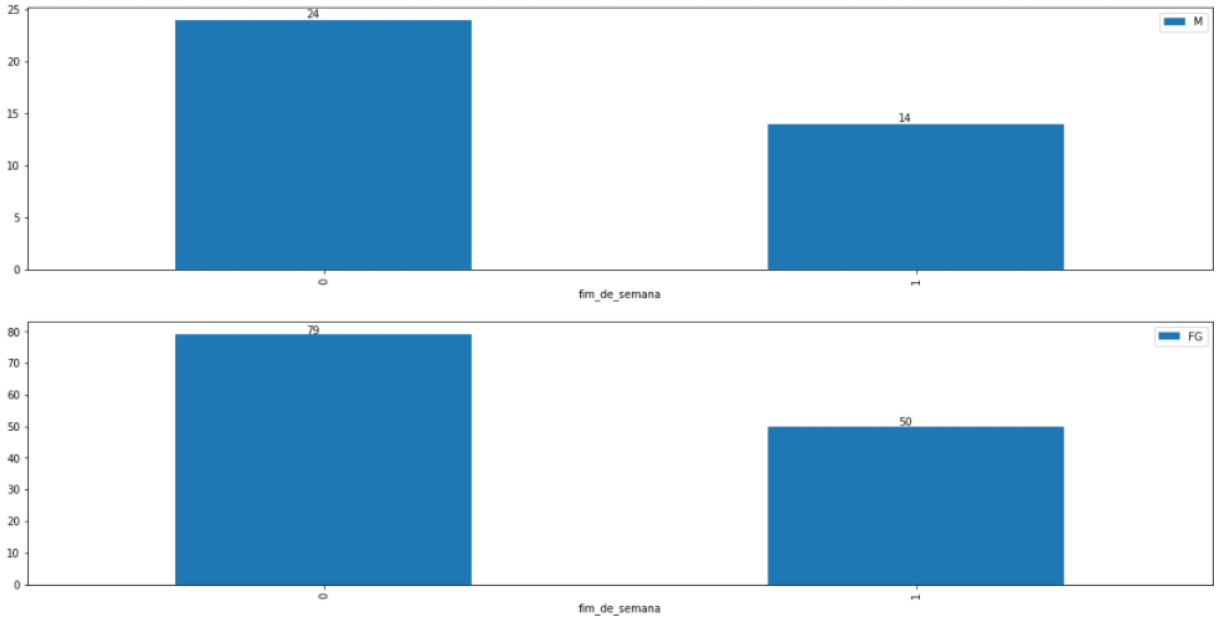


Figure 4.8: Number of deaths and serious injuries, respectively, in weekdays and weekend

Figure 4.8 shows that there is a correlation of around 2:1 for weekdays and weekend, respectively, whether it's fatalities or serious injuries. Since weekdays are constituted by five days while the weekend is constituted by two and the correlation between the two is the one mentioned previously, it reveals that weekend is more dangerous than weekdays, even though, in total, they have fewer records than weekdays because a 5:2 correlation (five weekdays to two weekend days) is greater than a 2:1 correlation.

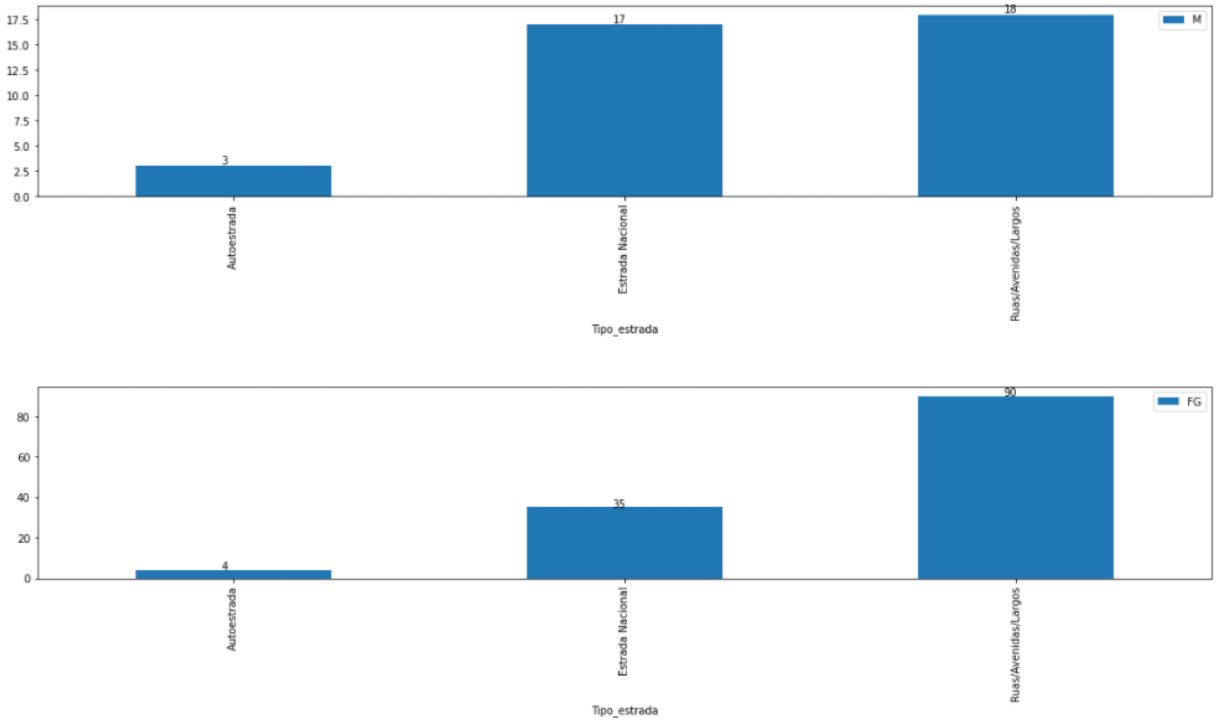


Figure 4.9: Number of deaths and serious injuries, respectively, for each road type

Figure 4.9 shows how fatalities and serious injuries are distributed between the road types available, which are highway, national roads and streets, respectively. In the fatalities category, we can see that national roads and streets have a high number of deaths when compared to highways with 17 and 18, respectively. Highways registered three deaths implying that these are safe and that the others need to be examined in more detail. When it comes to serious injuries, highways still have a low amount of it with four. However, the similarity that existed between national roads and streets no longer exists because they registered 35 and 90 serious injuries, respectively, implying that streets are much more dangerous when talking about serious injuries.

Having said this, because this dataset contains so few records, many of the visualizations that would be included here would also be included in the report that was built and is displayed in section 6.3. Further analysis will be displayed in that section to complement what was said in this one.

4.3 Data Correlation and Feature Analysis

One of the easiest ways to visualize data correlation is through heatmaps. These types of graphics allow seeing how features are related to each other (how much can a feature explain another) through somewhat of a matrix where lines and columns are defined by each feature and each cell has the amount of correlation (values between -1 and 1) between features. Normally, in many datasets, there is usually a feature that we try to know or get an answer through other features called output value. To see which features are more relevant when studying and modelling that output value, it's good practice to filter only the ones that correlate with an absolute value greater than a threshold relative to the output value since it helps the model give more accurate answers and not include "junk"in the model.

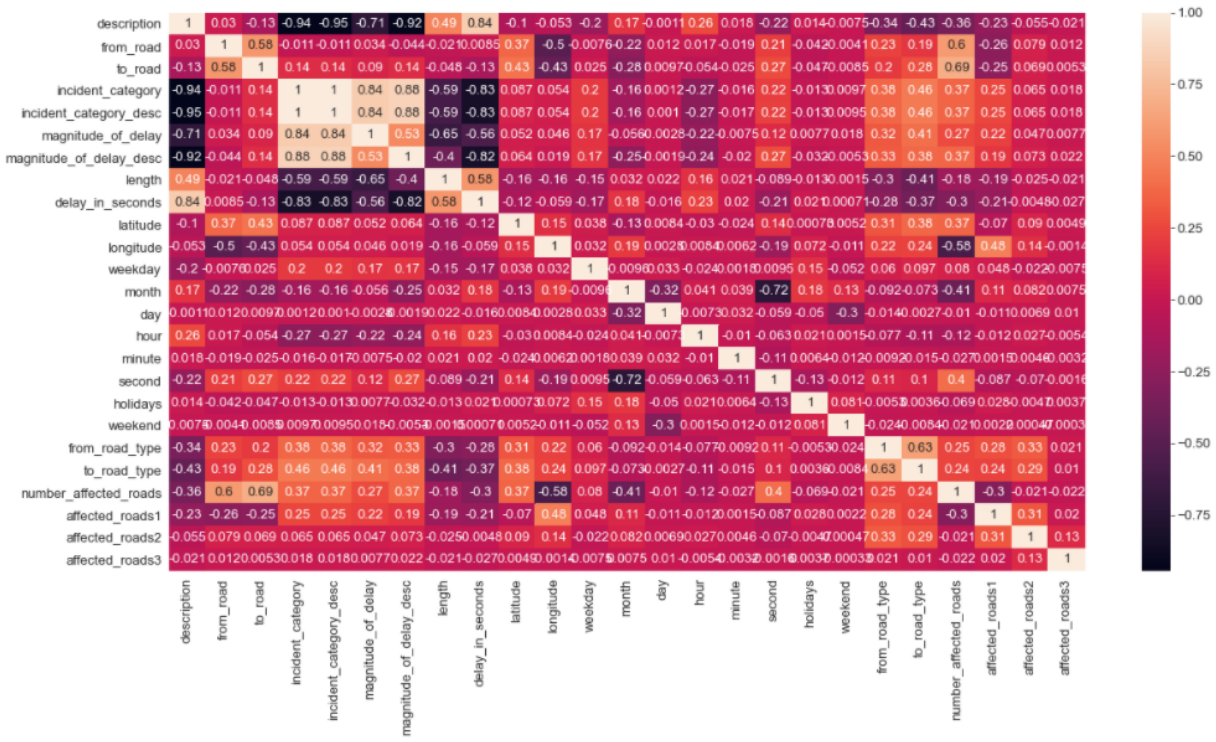


Figure 4.10: Tomtom heatmap

This is the heatmap collected from the tomtom dataset which, as we can see, shows correlation between all features present in the data where we can quickly identify strong connections such as “magnitude_of_delay_desc” and “description” or “incident_category” and “description” and also weak connections such as “weekday” and “month” or “hour” and “affected_roads1”. This was not the only heatmap required since, in the modelling section ahead, a study between models is going to be made using data with different sets of features. It’s also needed to point out a feature on which the heatmaps will focus and calculate the other features’ correlation to it, and this feature is the time delay (“delay_in_seconds”). The other heatmaps required for this work were for features that had an absolute correlation greater than 0.1 and 0.4 to the time delay and will be presented next.

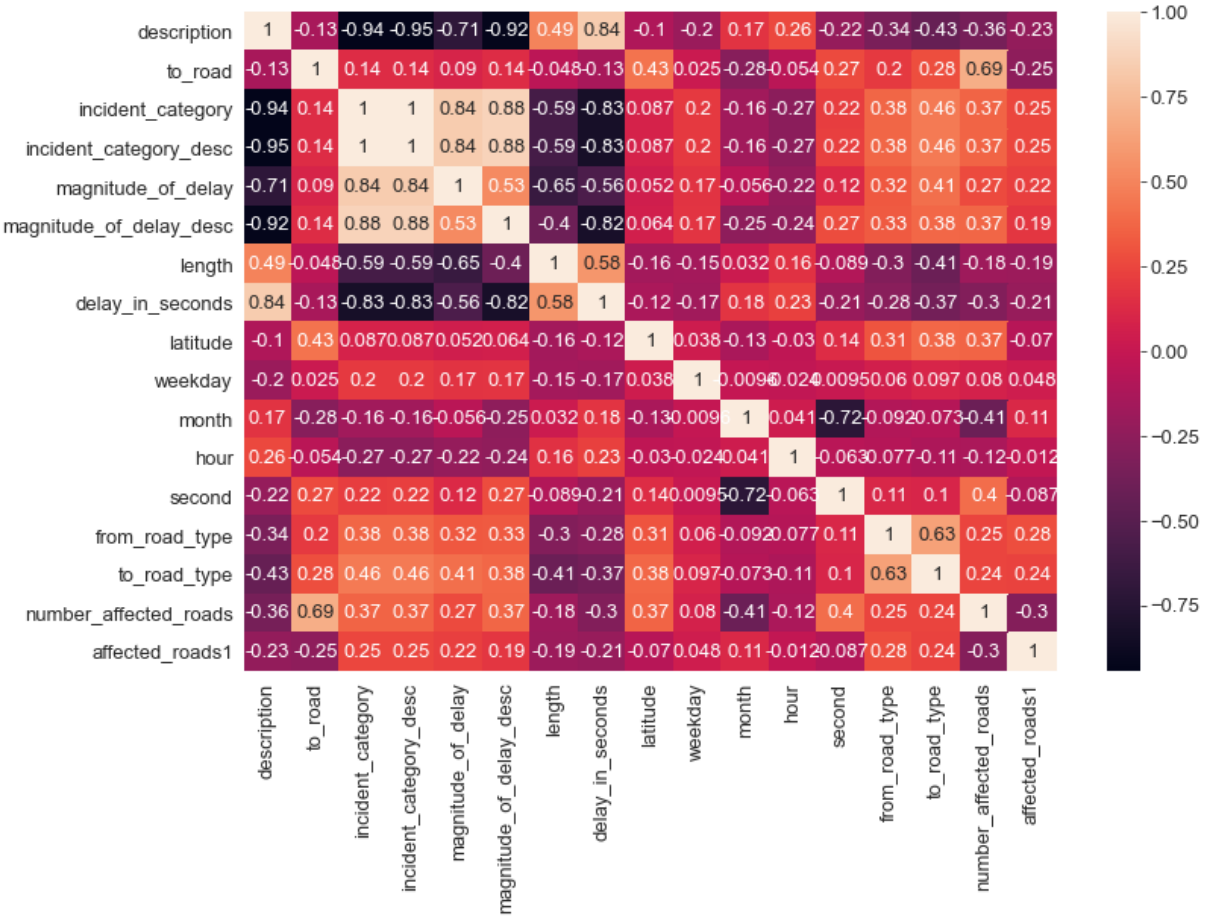


Figure 4.11: Tomtom heatmap with correlation greater than 0.1

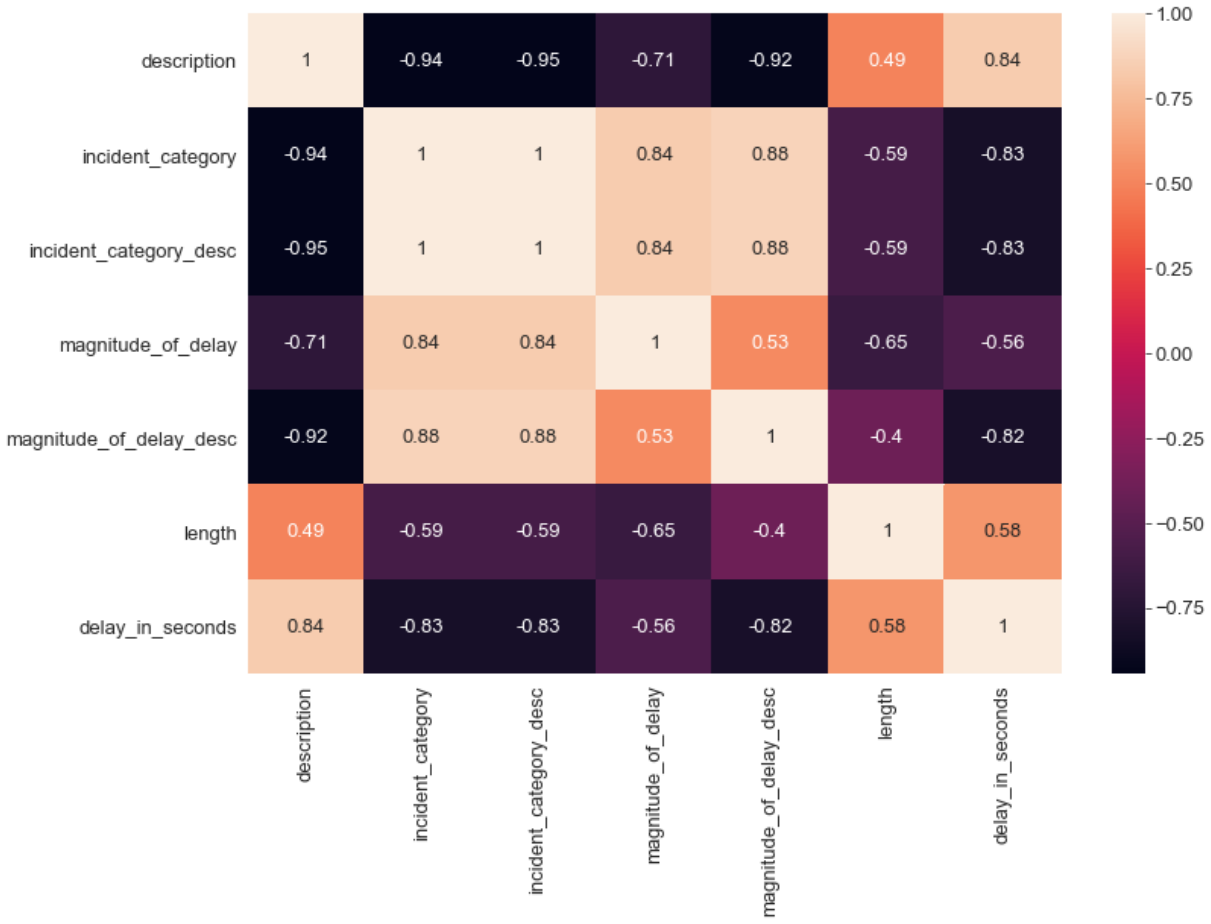


Figure 4.12: Tomtom heatmap with correlation greater than 0.4

As we can observe from Figures 4.10 and 4.11, eight features dropped meaning they have correlation below 0.1. The two most noticeable details are that the feature related to the starting location of the incident dropped while the finishing location stayed, and in the time hierarchy between month and second, only the day and minute features dropped, while month, hour and second remained which is strange since it gives us the impression that the second when an incident occurs can explain the traffic time delay better than the minute and day when an incident occurred. From Figures 4.11 and 4.12, ten features dropped, in other words, in total 17 features dropped from the dataset, and only six features had a correlation greater than 0.4. Only description about traffic, incident, magnitude of delay and length were correlated enough to be on this heatmap.

4.4 Technologies

This research was done on jupyter notebooks environment using the python programming language as well as PowerBI. These technologies were key throughout all the process where python was essential from inserting raw data to getting insights and results about data and models, respectively, whereas PowerBI

was decisive in presenting useful information in dynamic and practical visualizations. Many python libraries were used in this research, such as Pandas, NumPy, scikit-learn, Keras, matplotlib, SciPy and seaborn. The computer's main specs used in this research were an Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz processor, 16.0 GB RAM and an NVIDIA GeForce GTX 1050 Ti graphic card.

Chapter 5

Experiments

To reach insights and results, several scenarios were used along with various algorithms/models to reach definitive conclusions such as the best algorithm for these data or the impact of features on the results. These scenarios are only applicable to the tomtom dataset because it had data with a sufficient amount of records to model.

5.1 Unsupervised learning scenarios

There were two unsupervised learning models used, which were both clustering algorithms, namely, k-means clustering and hierarchical clustering (agglomerative). The intent of using clustering was to understand the distribution of data points in 2d and 3d scenarios and consequently deduce characteristics that distinguish concentrations of points. Before performing these clustering algorithms, it's necessary to turn all features into numeric values and then transform them in a way that we can concentrate information stemming from the features into a few dimensions. This transformation is known as dimensionality reduction and can be done using different techniques, but the most popular one is principal component analysis (PCA). This technique reduces the dimensionality of the data while still maintaining the data's variation. It identifies principal components that explain most of the variation in data so each data point can be represented with just a couple of numbers and consequently plot it to easily visualize similarities and differences between records [42]. In total, were used 33857 records for the unsupervised learning scenarios. The target was to reach around 95% of explained variance, and the minimum number of features required to reach close to it was four, so the whole dataset containing every feature was compacted into only four. The last observation is shown in Figure 5.1.

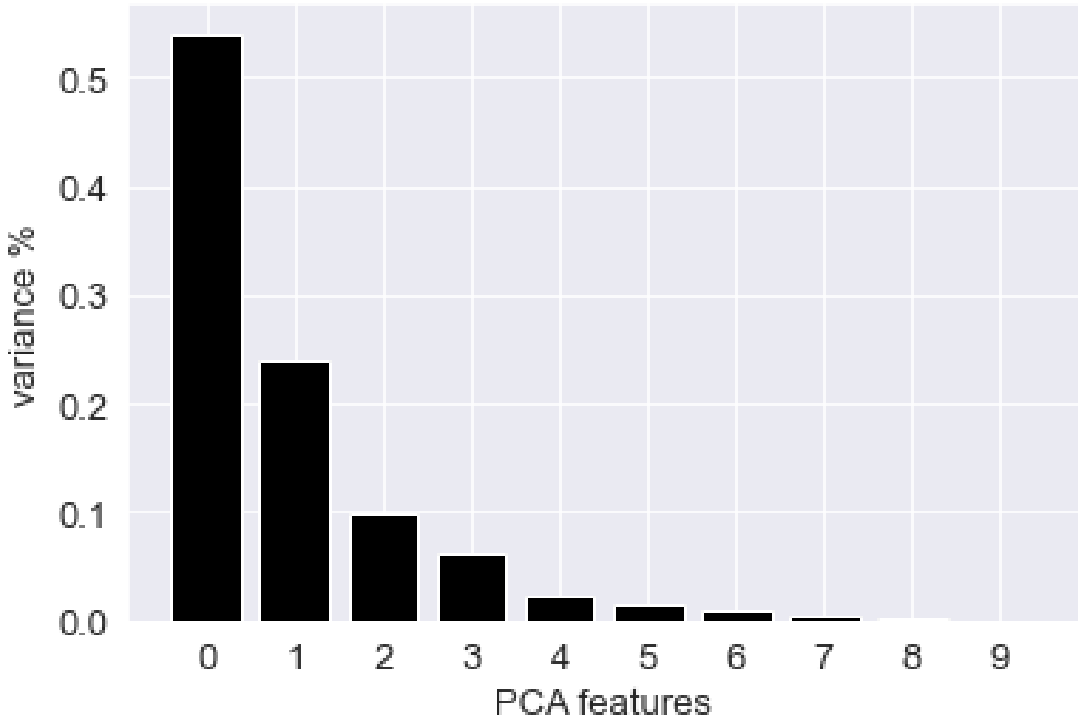


Figure 5.1: Variance evolution with number of features

5.1.1 K-Means Clustering

To get the ideal number of clusters for the algorithm a cycle was made in combination with the K-means scikit-learn function, making the "n_clusters" argument vary from one to eleven and using all the other arguments' standard values since it's quickly noticeable that slightly changing some arguments doesn't affect the result. One of the outputs of this function is the "inertia_" which is the sum of squared distances of samples to their closest center, and by gathering this value for all the number of clusters tested, we can achieve the optimal value. There were two methods used to obtain the ideal value, through a plot and a specific function used to locate the knee/elbow point of a line which is KneeLocator from the "kneed" library. This function required two special arguments indicating the direction and curve of the line, which was decreasing and convex, respectively. Even though the function gave the elbow point at four, by looking at Figure 5.2, three appeared to be a better choice since by then the benefit gained from raising from three clusters to four was significantly smaller, and so three was the number of clusters used in the K-means function.

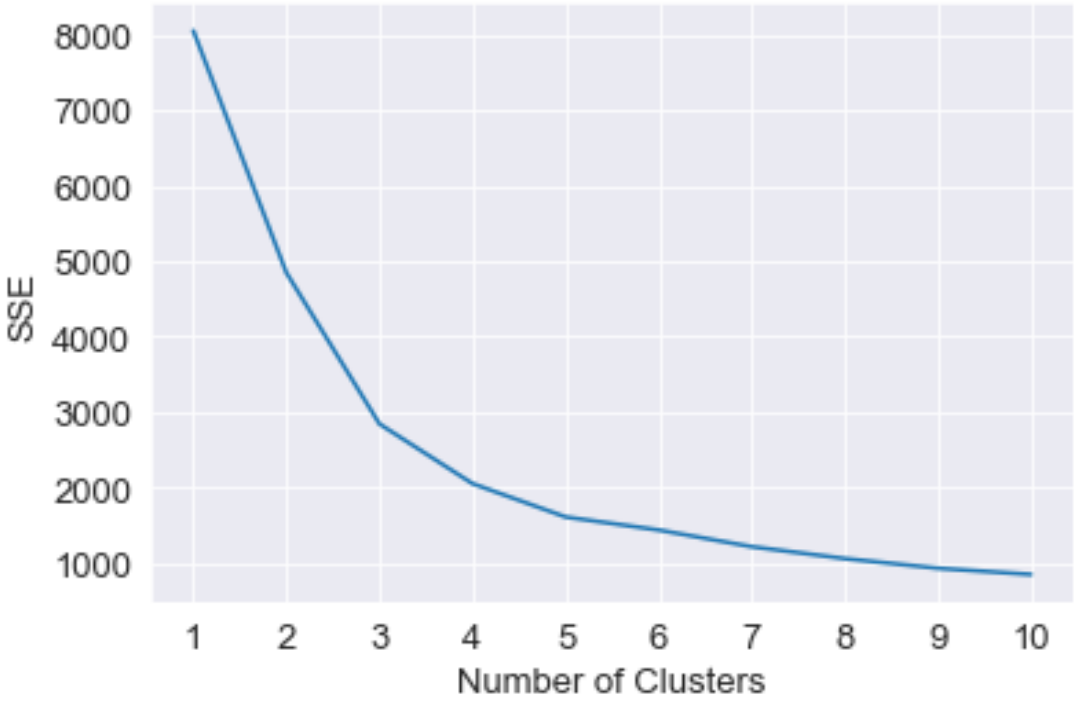


Figure 5.2: Elbow Method Curve

5.1.2 Hierarchical Clustering

For the hierarchical clustering, the "AgglomerativeClustering" function was used from the scikit-learn library, where the number of clusters was also an important argument to optimize. Besides the number of clusters, there weren't any more relevant arguments to optimize, so a cycle was made varying this value from two (minimum argument) to eleven utilizing various performance metrics to evaluate the optimal value. The metrics used were the silhouette score, Davies-Bouldin score and Calinski-Harabasz, and they evaluate, respectively:

- Difference between the mean nearest-cluster distance (distance between a sample and the nearest cluster that the sample is not a part of) and the mean intra-cluster distance (average dissimilarity between a sample and other objects that belong to the same cluster) and then dividing by which one is higher, for each sample. The result can be between 1 and -1, and the closer it is to 1, the better is the score [43];
- Ratio of the total within-cluster dispersion and between-cluster separation. This means that the closer the value is to zero, the better is the score [44];
- Ratio between the between-cluster dispersion and the within-cluster dispersion, in other words, the ratio between the sum of squares of the distances between the center of each cluster and the

centroid of the dataset and the sum of the squares of the distances between the center of each cluster and every point in the cluster. The higher the score, the better is the score [45].

After analyzing all the scores for all the possible number of clusters, the best values were 5, according to the silhouette and Calinski-Harabasz scores, and 2, according to the Davies-Bouldin score.

5.2 Supervised Learning scenarios

There were four supervised learning models used, which were support vector regression, linear regression, k-nearest neighbours and neural network. The objective for these models was to produce the lower traffic time delay loss (error). This feature indicates how much more time is needed to go through the same road when compared to a regular day. The data containing 33857 records had to be split into training and testing since these models require them. The "train_test_split" function from scikit-learn was used, and it required the predicting features as well as the response feature, together with a fraction to be assigned as testing data which, in this case, was 0.2 or 20% while the remaining 80% were assigned as training data. It's worth mentioning that 5-fold cross-validation was used in all supervised learning models, as well as a grid search technique to carry out hyperparameter tuning that's going to be explored later. There were also four data scenarios where these models were carried out.

- **Whole Dataset** - The first data scenario was the one with the whole dataset. The intent was to see the results using all data and then compare with the ones that had slight adjustments aiming to optimize the results;
- **Correlation >0.1** - The second data scenario was the one where all features had a correlation with the time delay greater than 0.1. The intent was to remove features that had little to no relevance to the feature in question. The features that remained in this scenario are "description", "to_road", "incident_category", "incident_category_desc", "magnitude_of_delay", "magnitude_of_delay_desc", "length", "delay_in_seconds", "latitude", "weekday", "month", "hour", "second", "from_road_type", "to_road_type", "number_affected_roads", "affected_roads1";
- **Correlation >0.4** - The third data scenario was the one where all features had a correlation with the time delay greater than 0.4. The intent was to only include features that had a great impact on the stated feature. The features that remained in this scenario are "description", "incident_category", "incident_category_desc", "magnitude_of_delay", "magnitude_of_delay_desc", "length", "delay_in_seconds";
- **Correlation >0.4 normalized** - The fourth and last data scenario used the previous scenario but added normalization. By adding normalization, all the records will be scaled to much lower values

in a specific range (p.e. between 0 and 1), giving the possibility to revert that scaling to the original values. This can be notably important since algorithms are frequently sensitive to high values, so it's common to use normalization before modelling.

5.2.1 Support Vector Regression

Support vector regression (SVR) is a popular model because it has good generalization performance and is not conducive to overfitting since it consists in finding a function that outlines the narrowest space (tube shape) while minimizing the distance between predicted and desired output [46][47]. All this process was developed with the "LinearSVR" function from the scikit-learn library. As mentioned earlier, these models try to predict the output feature "delay_in_seconds" (traffic time delay) so we can associate a score to the model as well as metrics such as MAE, MSE and RMSE, which evaluate the effectiveness of the candidate models to then perform comparisons and later evaluate them [48]. In this function, four arguments were tuned using grid search ("GridSearchCV" on scikit-learn), which is an exhaustive search over the parameters given. Furthermore, tuning was made using 5-fold cross-validation and the arguments tuned were:

- **epsilon**: Insensitive loss function or margin of tolerance where no penalty is attributed to errors;
- **C**: Regularization parameter or strictness about misclassification;
- **loss**: Loss function to be used (between epsilon insensitive or L1 loss and squared epsilon insensitive or L2 loss);
- **max_iter**: Number of iterations to be run.

These four hyperparameters were tuned using the values in Table 5.1:

Nominal Parameters			
Loss	["epsilon_insensitive", "squared_epsilon_insensitive"]		
Numerical Parameters			
	Start Value	Stop Value	Step
Epsilon	0	0.5	0.1
C	0.25	1.75	0.25
Maximum Iterations	1000	2500	250

Table 5.1: Set of hyperparameters used to tune SVR

The SVR models were tuned using all possible combinations shown above.

5.2.2 Linear Regression

Linear regression models interpret how the response variable y changes in regards to a certain number of predictors x and create a linear function that can predict the response variable utilizing the predictors with minimal error [49]. To develop this model, the "LinearRegression" function was used from the scikit-learn library, and the intent was the same as SVR, analyze, through metrics, the performance of these models and compare them afterwards. Linear regression is a simple algorithm when compared to other algorithms, so the function arguments are also pretty simple and don't need tuning since they are aimed at occasional situations. Furthermore, since this function didn't include cross-validation, another function had to be used, which was "cross_val_score" also from the scikit-learn library, which required arguments such as the model being used (linear regression function), training predictors and response as well as scoring method (r2) and cross-validation value which was five.

5.2.3 K-Nearest Neighbours

This algorithm classifies data points relative to their neighbours, meaning that, firstly, the neighbours have to be identified and then assign a classification or value using the identified neighbours as reference [29]. For this model, the "KNeighborsRegressor" function from scikit-learn was used, which had some arguments that needed tuning, unlike linear regression. Once again, grid search was used, just like SVR, for hyperparameter tuning to develop the best results. Finally, 5-fold cross-validation was used in model training. The arguments that were included in the grid search were the following:

- **n_neighbors**: Number of neighbours to be used for regression;
- **algorithm**: Algorithm used to compute the nearest neighbours;
- **leaf_size**: Number of leaves used in tree algorithms.

There's another parameter worth mentioning which is "p" that determines the metric used for calculating the distance between points. The value for "p" was 2 since it's the value for the euclidean distance that is the length of a line segment between two points. Next, Table 5.2 displays the parameters that were tuned and how they were tuned.

Nominal Parameters			
Algorithm	["auto", "ball_tree", "kd_tree"]		
Numerical Parameters			
	Start Value	Stop Value	Step
Leaf Size	10	30	5
Number of neighbors	4	11	1

Table 5.2: Set of parameters used to tune KNN

The KNN model was tuned using all the possible combinations displayed in table 5.2

5.2.4 Neural Network

Neural networks can learn complex input-output relationships and adjust internal parameters by themselves to better learn data [50]. There are some types of neural networks but, in this case, these neural networks are constituted by layers (at least two, input and output and possibly hidden layers) where each has neurons that send vectors of data to the neurons in the next layer and use backpropagation to adjust their weights to better fit data that's input [51]. They are more complex and require a significant number of computing resources. Some things worth mentioning is that 5-fold cross-validation was used in these neural networks, and the activation function "relu" was used as well since it's a linear function, thus being adequate with the values in question. Furthermore, the Adam algorithm was used for optimization, which is a stochastic gradient descent that changes attributes such as weights and learning rates to minimize loss method and is considered, arguably, the best overall optimizer existing [52]. Finally, MSE was used in the neural networks as the loss metric. Some parameters had to be tested and changed to optimize the neural network's performance. These parameters are going to be presented in Table 5.3.

Hidden Layers	[0,1,2,3]
Neurons	[8,16,32,64,128,256]
Dropout	[0.1,0.2,0.3,0.4]
Epochs	[100,200,500,1000,2000]
Batch size	[32,64,128]

Table 5.3: Set of parameters used to tune Neural network

Even though this tuning was more extensive than others, all these parameters were tested with the respective values until the network architecture and parameters were ideal.

Chapter 6

Results and discussion

The results in this chapter are all derived from models using the best parameters possible that were mentioned previously. Next, we are going to analyze the results, check their plausibility and model reliability, to then deploy the results in the dashboard platform.

6.1 Incident Correlation

To create a relation between incidents, unsupervised learning algorithms, namely clustering algorithms, were used to analyze connections between incidents and infer their patterns and tendencies. It's worth mentioning that a modified dataset was used when performing these models, which were explained in the previous section.

6.1.1 K-Means Clustering

The first unsupervised learning algorithm used is k-means clustering which places an arbitrary number of centroids (a number that is chosen and deliberated previously) in somewhat random positions (according to data points) or by using initializing techniques and associates each data point with its least distant centroid. After all, records have been matched, the centroids' positions are recalculated by summing all the points belonging to the respective cluster and then dividing by the number of points of the cluster. These were the data points' distribution using three clusters in two dimensions.

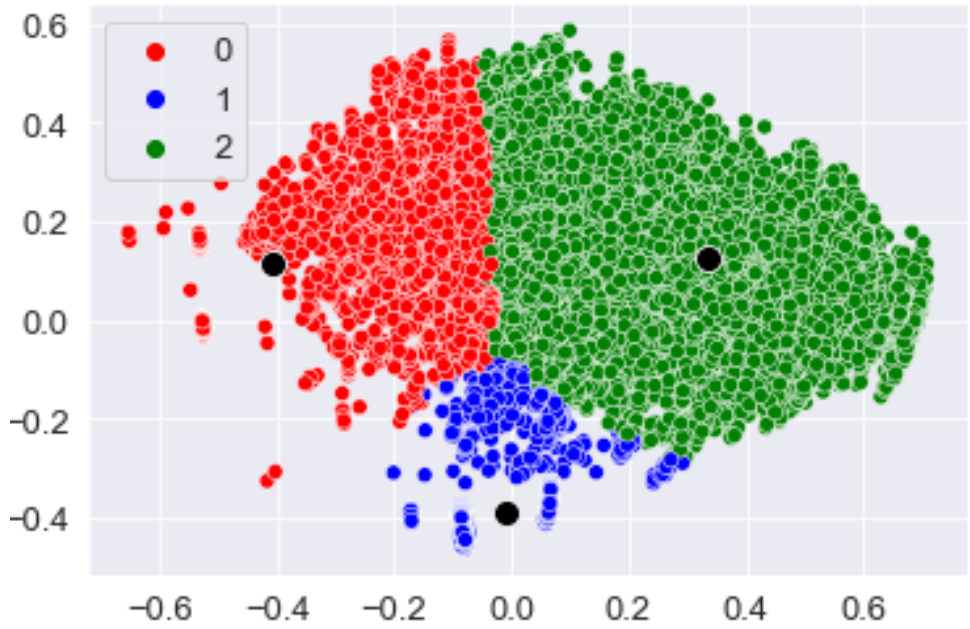


Figure 6.1: Representation of clusters and data points

We can see by Figure 6.1 the separation between the three groups of points, one of them gathering around 42% of the data records (green), the second one gathering around 34% (red) and the last one with around 24% (blue). Besides this two-dimension representation, it's also possible to represent these clusters in three dimensions which is shown in Figure 6.2.

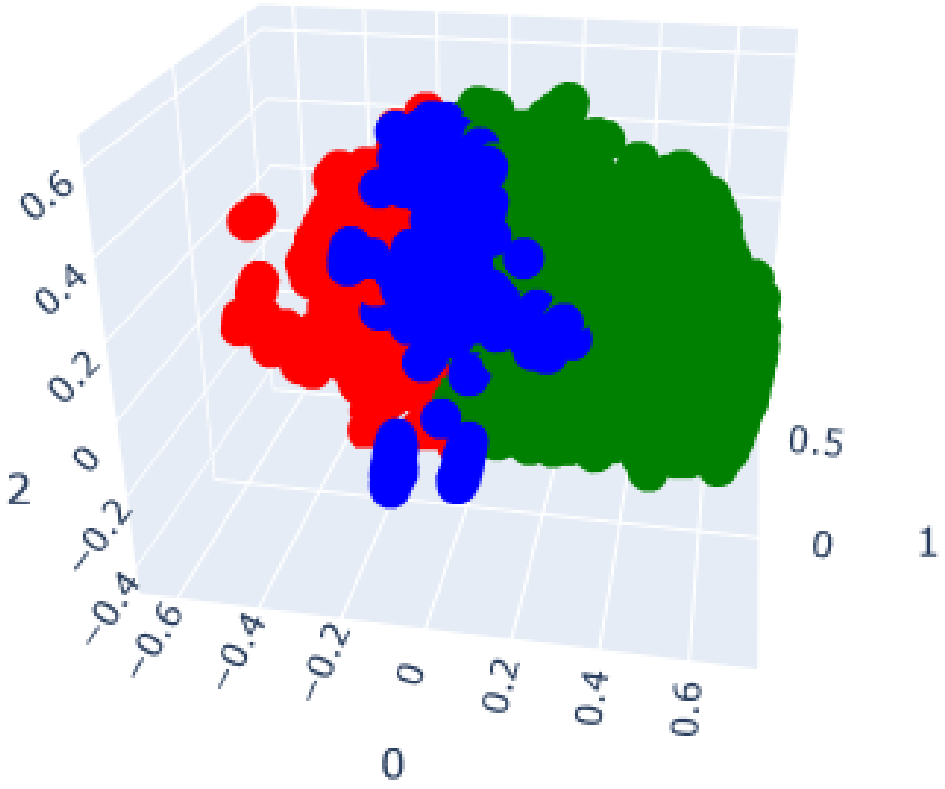


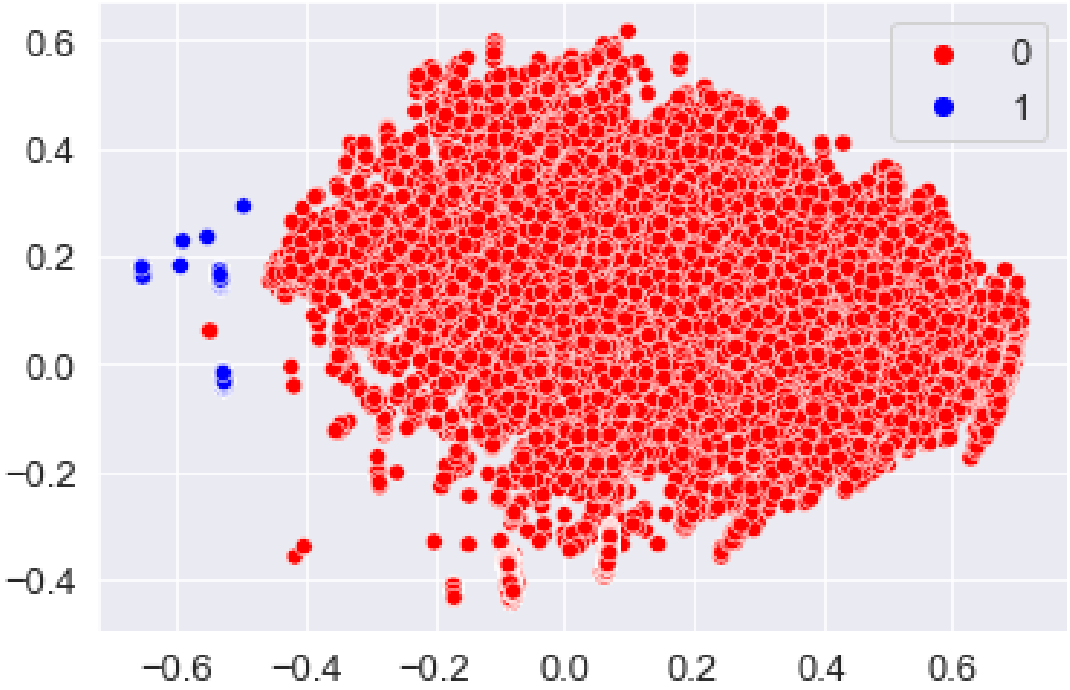
Figure 6.2: Three-dimensional representation of clusters and data points

To understand better why they were split like this, it's essential to look into them and gather the information that's exclusive to each cluster so we can explain the classification above. The logical way to assemble this information is to calculate, for each feature, some statistical values such as mean, median, mode or quartiles. One differentiating factor between the clusters is the delay magnitude of their respective data points. The smallest cluster appears to have a higher delay magnitude than the others since this feature's mean is approximately 3.94 for this cluster, more than 75% of the data records contain a magnitude of 4, which directly corresponds to undefined as its description, which also corresponds to closed roads. The middle cluster contains a delay magnitude of around 3.57, which is also higher compared to the dataset's mean having also more than 50% of records with a delay magnitude of 4. The biggest cluster contains a slightly lower mean of around 2.60 with a more diverse collection of data points when it comes to the delay magnitude. Another differentiating factor is the feature representing the length of an incident which has a significantly higher mean of around 553.17 and a standard deviation of around 350.83, meaning that their values are very dispersed and also have minimum, maximum and quartile values much higher than the respective metrics on the other two clusters. The middle cluster contains a mean of around 173.73 and the smaller cluster with a mean of around 83.47. An interesting fact is that the middle and smaller cluster contain a maximum incident length of 580 and 470, respectively, while the bigger cluster contains a maximum incident length of 4980 which is way above the ones mentioned previously. The time delay follows the same order as the incident length, with the biggest cluster having a mean of around 205.96,

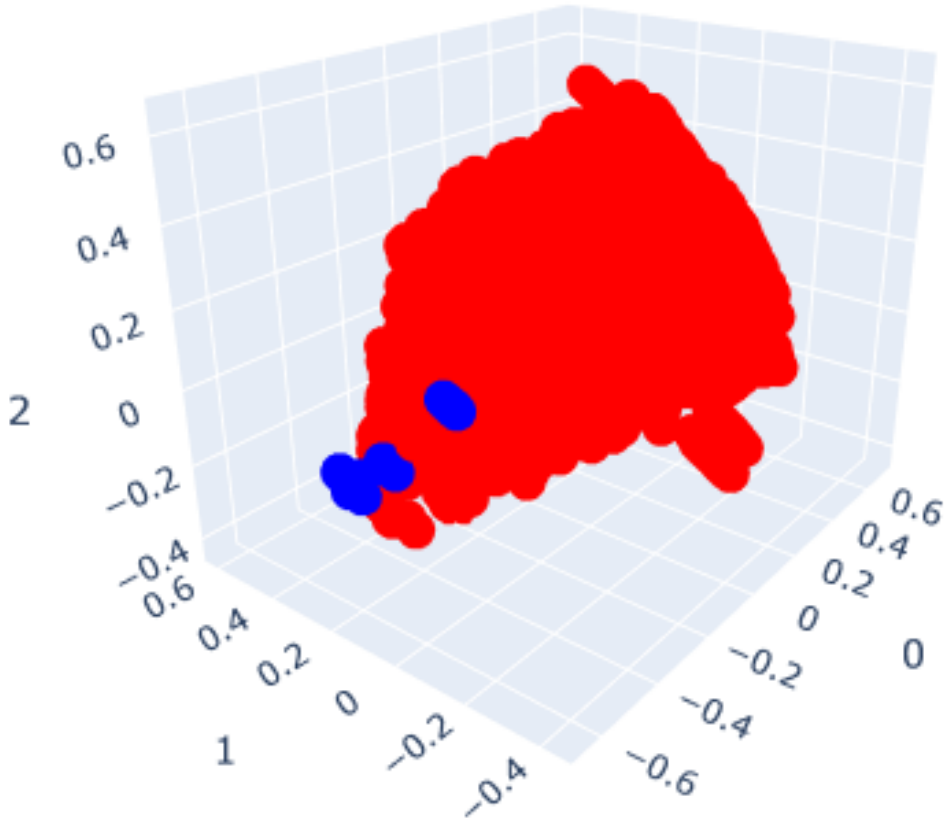
the middle cluster with a mean of around 56.74 and the smaller cluster with a mean of around 4.06, which is quite low but since the data records with undefined delay magnitude have a time delay of 0, these values are to be expected. Another somewhat distinguishing factor is the month, with the middle cluster having a bigger concentration (more than 50%) of incidents occurring earlier (between July and September) than the other two, which have a similar distribution when it comes to month (both appear to have similar distribution and variety). Even though special dates and weekends have few differences, they are worth mentioning since their occurrence is low, thus explaining those small differences that could have meaning when differentiating these clusters. The smaller cluster has a higher occurrence of special dates and weekends since their mean is higher than other clusters (around 0.046 and 0.076, respectively). The next in line is the bigger cluster having a special dates mean of around 0.038 and weekends of around 0.073, which is very close to the smaller cluster. Finally, we have the middle cluster with a special dates mean of around 0.023 and weekends of around 0.069. The number of roads affected in each incident also has significant differences mainly in the middle cluster, because, generally speaking, the incidents belonging to that cluster affect more other roads than the other 2 clusters with a mean of around 4.19 while the bigger and smaller clusters have a mean of around 2.57 and 2.06, respectively (minimum of affected roads is one while the maximum is five). The last thing to mention is that, when it comes to affected roads, there's a recurring road identified as the mode, which is the N201, as, for the others, there's no specific road identified.

6.1.2 Agglomerative Clustering

The second unsupervised learning algorithm used was agglomerative clustering, which is a type of hierarchical clustering that considers a different approach to finding clusters than the previous algorithm. In the beginning, each data point is a separate cluster which means that there are as many data points as there are clusters. At each iteration of the algorithm, similar clusters merge until k clusters are formed (k is inferred beforehand). To illustrate the results from the different number of clusters gotten in the previous chapter, next, will be presented plots in two and three dimensions regarding these two groupings.



(a) Two-dimensional representation of 2 clusters and data points



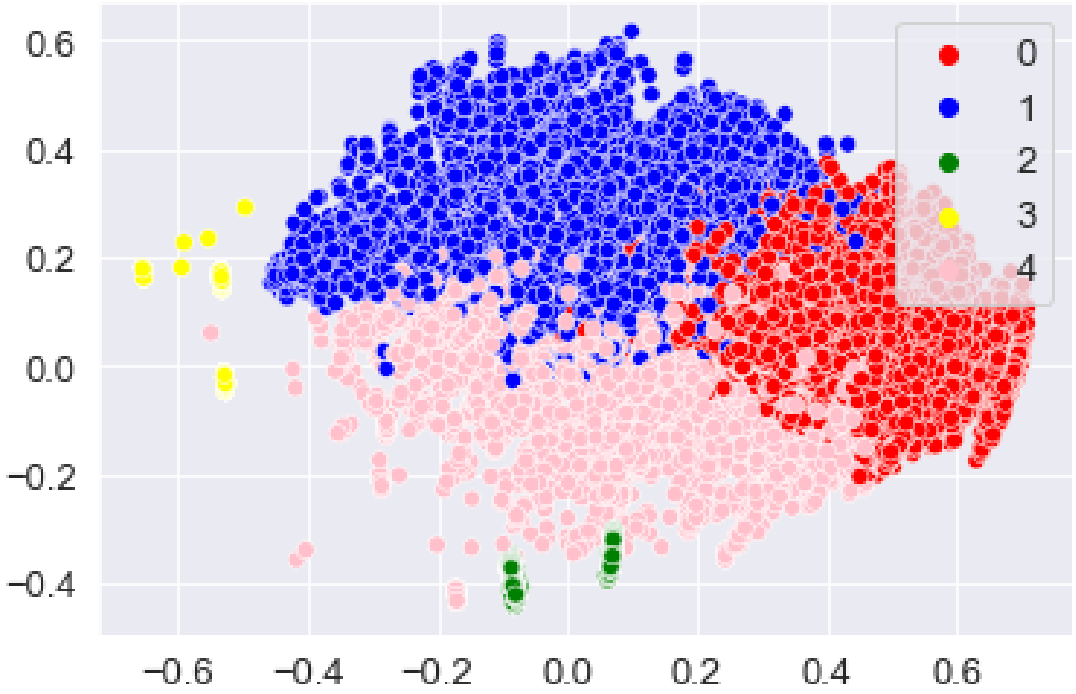
(b) Three-dimensional representation of 2 clusters and data points

Figure 6.3: Representation of agglomerative clustering with 2 clusters

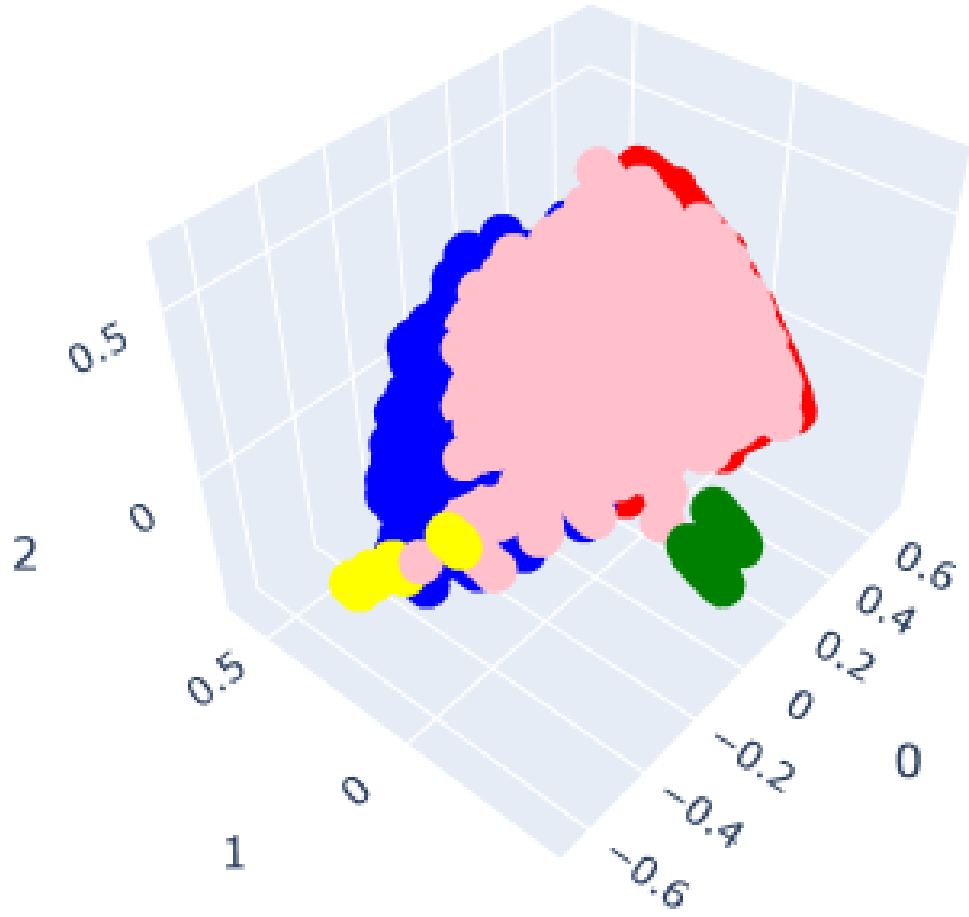
In Figures 6.3a and 6.3b, it appears that there's a big disproportion where one cluster contains almost

all data records and the other contains just a few ones but, as we are going to see ahead, the disproportion is not as big as it shows in these figures and some data records are just very concentrated in specific zones.

One of the clusters has 26431 data records while the second one contains 7426 data records, and even though there's a significant difference in these numbers, it doesn't reflect the difference shown in Figures 6.3a and 6.3b. One big difference between these two groups is that the smaller cluster contains a considerably bigger delay magnitude than the bigger cluster with a mean of around 4 and 3.03, respectively, which means that the smaller cluster contains almost exclusively delay magnitudes of 4 (since they're the maximum value) and its standard deviation is only 0.05 while the other one's standard deviation is around 0.81. This also means that the smaller one is mostly formed by undefined descriptions of the delay magnitude, which corresponds to incidents with roads closed. When it comes to incident length, there's also a similar difference such as the delay magnitude, where the bigger cluster contains a bigger mean of around 365.08 as well as a much bigger standard deviation of around 336.41, meaning that there's a big variety of incident length, and the smaller cluster contains a smaller mean of around 126.87 as well as a much smaller standard deviation of around 17.37 meaning that these values are much more concentrated around the mean value. The time delay follows the same direction, where the smaller cluster contains a much smaller mean of around 0.01 and a standard deviation of around 0.32 and the bigger cluster has a much bigger mean of around 137.56 and a standard deviation of around 118.19. These very small values are explained by the fact that, since the respective cluster contains almost exclusively roads closed, such as the description of delay magnitude, the time delay associated with that description is zero, making a definite impact on the time delay. There's also a significant difference in the average month between these two clusters, where the smaller one has a mean of around 9.15 while the bigger one has a mean of around 10.04. Besides, the smaller cluster doesn't contain incidents that occurred in November or December since the maximum month registered is October, while the other seems to contain incidents that happened in all months existing in the dataset. Both clusters show differences regarding incidents happening on special dates where the bigger cluster's mean is around 0.039 while the smaller one is around 0.019 whilst weekends seem to have a similar incidence. Another big difference is the number of affected roads, where the smaller cluster has the edge with a mean very close to the maximum value, which is five, and a very small standard deviation, while the bigger cluster has a mean of around 2.44 and a standard deviation of around 0.72. Finally, and similarly to what happened in the k-means model, the smaller cluster has a predominant road affected by its respective incident records, which is the N201. To sum it up a bit, it appears that the smaller cluster turned out to be a concentration of incidents that resulted in roads closed (which stand for very small time delay) with a considerably low length.



(a) Two-dimensional representation of 5 clusters and data points



(b) Three-dimensional representation of 5 clusters and data points

Figure 6.4: Representation of agglomerative clustering with 5 clusters

In Figures 6.4a and 6.4b, it appears once again that three clusters contain practically all the data records while the other two contain almost no points but, similarly to the previous situation, the disproportion is even smaller because both smaller (apparently) clusters even have roughly the same data records as the bigger ones meaning that some points are just really concentrated in specific zones (one of the apparent bigger clusters even has fewer data records than both apparent smaller clusters). Once again, it's important to dig deeper to find each cluster's characteristics by analyzing statistical values for each feature that was mentioned before.

It's necessary to make a different distinction between clusters since their size cannot be used as a factor for that purpose, so numbers are going to be used. Cluster numbers 0, 1 and 4 belong to that big agglomeration of points where cluster 0 is the most right one, number 1 is the top one, and number 4 is the bottom one. Cluster number 2 is the small concentration of data records below cluster 4, and cluster number 3 is the other small concentration of points left to clusters 1 and 4. Besides this, their main characteristics are going to be displayed through an ordered list so it's easier to express and understand information about them:

0. Contains 7935 data records, a relatively smaller delay magnitude (when compared to the whole dataset) with a mean of around 2.50, a very high incident length with a mean of around 693.61, followed by a very high standard deviation of around 402.40. It also contains a high time delay mean of around 207.35, month's mean is around 10.09, a similar special dates mean, when compared to the whole dataset, of around 0.036, similarly to the weekend mean of around 0.074 (close to the mean of the dataset) and a number of affected roads of around 2.49. There's also some interesting information given by the mode such as traffic description being stationary traffic or jam as traffic category description and having a predominance of national roads as what type of roads do these incidents happen since almost all other clusters, mentioned until now, have it as streets, avenues or squares.

1. Contains 7522 data records, a relatively smaller delay magnitude when compared to the whole dataset but higher than cluster 0 with a mean of around 2.83, a close to the average incident length of around 322.94 and standard deviation significantly lower of around 111.86 and time delay higher than average but smaller than the previous cluster of around 189.27. Besides that, the month when incidents happen has a mean of around 9.96, a similar special date's mean to cluster 0 of around 0.039, a weekend mean of around 0.067, which is one of the lowest among all five clusters and relatively smaller than the average and a number of affected roads of around 2.65 which is higher than the previous cluster but smaller than the dataset's average. The only relevant thing when it comes to the mode is that the predominant traffic description and traffic category description are stationary traffic and jam, respectively.

2. Contains 7529 data records, only contains delay magnitudes of 4 which means that its description is undefined, the incident categories are roads closed (since these are all connected between each other), and time delays are 0, has the lowest incident length out of all five clusters of around 74.34 and also the smallest standard deviation of around 14. The mean of months where incidents happen is the highest out of all five clusters, at around 10.17, has the highest mean, when it comes to special dates at around 0.048, the second-highest in weekends at around 0.076 and the number of affected roads is always 2 for every incident belonging to this cluster. Aside from this information, the mode doesn't add anything relevant that can be used to characterize this cluster.

3. Contains 7426 data records, has a delay magnitude of around 4 (rounded), so the same information can be derived as it was done previously, much like the former one with a very small standard deviation, a low incident length of around 126.87 and similar standard deviation to cluster 2. In contrast with cluster 2, the mean of the incidents' months and holidays are the lowest, at around 9.15 and 0.019, respectively, while the weekends is the highest at around 0.077 and the number of affected roads is the highest among the five clusters at around 5 (rounded) with minimal standard deviation. The only additional thing worth mentioning is that there is one road that appears predominantly, which is, once again, N201.

4. This cluster contains the least data records at 3445, a delay magnitude of around 2.61, an incident length of around 335.76, a time delay of around 164.51, a month mean of around 9.80, a special date mean of around 0.032, a weekend mean of around 0.063 and number of affected roads around 2.84. All the values mentioned are somewhat similar values to the ones in cluster 1, which doesn't come as much of a surprise since they appear to be fairly close to each other. Besides this, the predominant description is the same as clusters 0 and 1 with the same traffic description and traffic category description (stationary traffic and jam, respectively).

6.2 Preventive models

To prevent dangerous situations such as road incidents, we can use supervised learning algorithms, which are used mainly to predict a feature inside a dataset using all the other information available in it. Consequently, these algorithms can be used to create predictive models to prevent road incidents. In this case, the feature chosen to become label was the "delay_in_seconds". The four specified datasets were used in each model, and were evaluated using three different metrics, namely mean absolute error, mean square error and root mean square error. They have different purposes of usage, which are going to be listed next.

- Mean average error is the simplest and easiest to understand since it's just the average error;

- Mean squared error is more popular than mean average error because bigger errors have a greater weight when calculating this metric which is generally useful in real-life situations;
- Root mean squared error is a robust metric less sensitive to bigger discrepancies than mean square error.

Besides the loss metrics, a scoring method was used to grade the models, which was the r^2 score (r-squared score). This rating technique measures the proportion of variance explained by the model, so the higher this scoring is, the better [53]. All figures that are going to appear in this section are going to be from the best scoring or with the least error models.

6.2.1 Support Vector Regression

To run the model, some additional methods are necessary besides the "LinearSVR" and "GridSearchCV" such as the fit and predict methods, which fits training data using the parameters provided and predict testing dataset's outputs, respectively. These are complemented with some attributes such as "best_params" and "best_score" which give the combination of parameters that resulted in the model with less error and the mean cross-validated score of the best estimator, respectively [43].

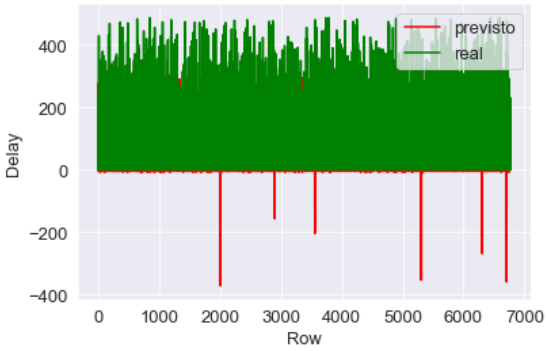
After performing this model in the four environments, the one that gave better results was the one using features with a correlation greater than 0.4 with normalization.

Dataset	MAE	MSE	RMSE	Score
Whole dataset	32.26	3157.26	56.19	0.769
Correlation >0.1	30.97	3100.90	55.69	0.783
Correlation >0.4	32.45	3570.66	59.76	0.778
Correlation >0.4 normalized	32.20	2956.39	54.37	0.795

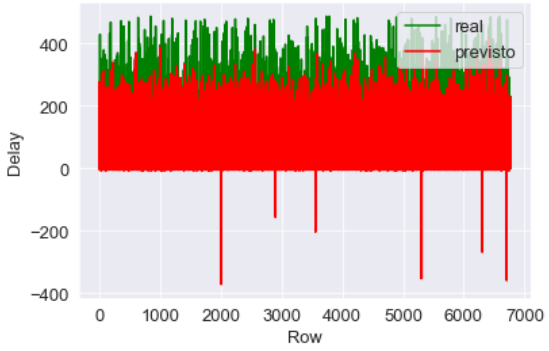
Table 6.1: Results from SVR model

Table 6.1 showed that the model using the whole dataset had the worst performance having a score of around 0.769, MAE, MSE and RMSE of around 32.26, 3157.26 and 56.19, respectively. When comparing the real values to the predicted ones, it showed that the model was consistently predicting values lower than real. Next was the one using the dataset containing features with correlation above 0.4 with a score around 0.778, MAE, MSE and RMSE of around 32.45, 3570.66 and 59.76, respectively. Even though the score on this model was higher, the metrics involving squared values were a bit worse, being more consistent but predicting even lower values than the previous model. Next was the model using the dataset with features that had a correlation greater than 0.1 with a score of around 0.783, MAE, MSE and RMSE of around 30.97, 3100.90 and 55.69, respectively. This model showed better metrics than the other two

but still showed some difficulty predicting higher values even though there was a noticeable improvement. The best model, mentioned earlier, had a score of around 0.795, MAE, MSE and RMSE of around 32.20, 2956.39 and 54.37, respectively. Even though the MAE was worse than the second, MSE and RMSE were better than all other models while having fewer error discrepancies (according to MSE and RMSE). In Figures 6.5a and 6.5b are shown the differences between the values SVR predicted and the ones present in data. Figure 6.6 shows the regression line stemming from joining the predicted and real values.



(a) SVR plot regrading differences between real and predicted values



(b) SVR plot regarding differences between real and predicted values (reversed)

Figure 6.5: SVR predicted and real values

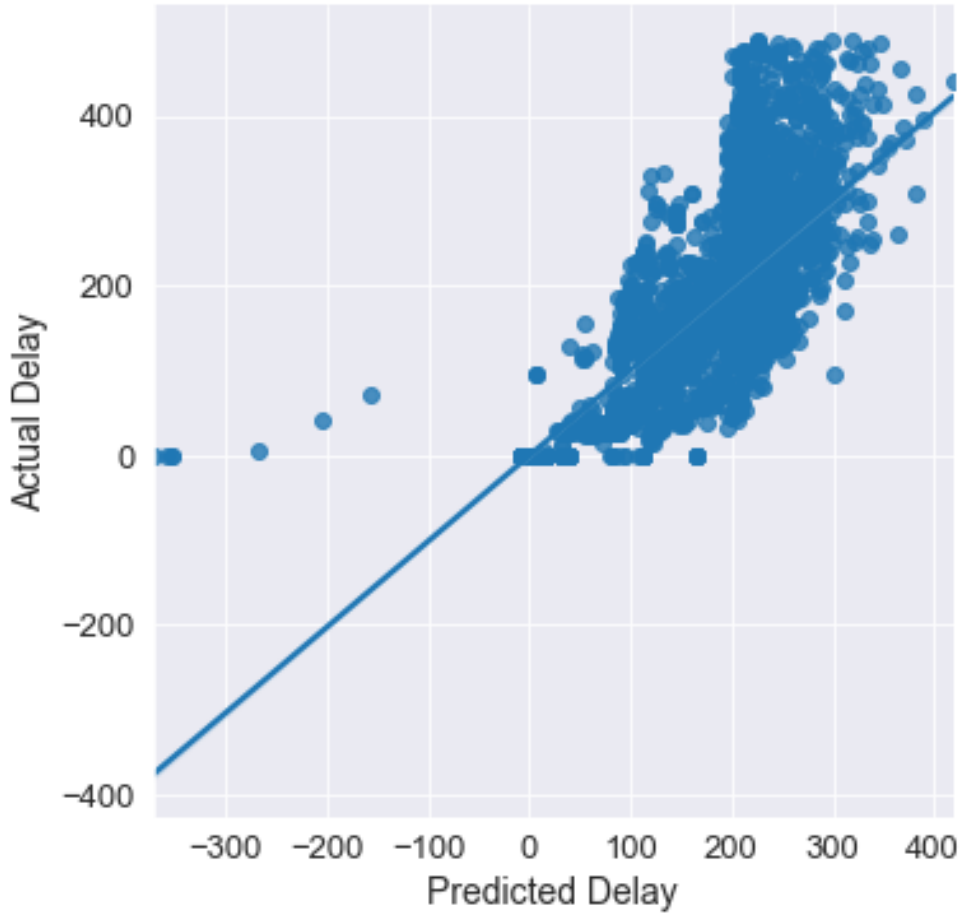


Figure 6.6: SVR regression between real and predicted values

6.2.2 Linear Regression

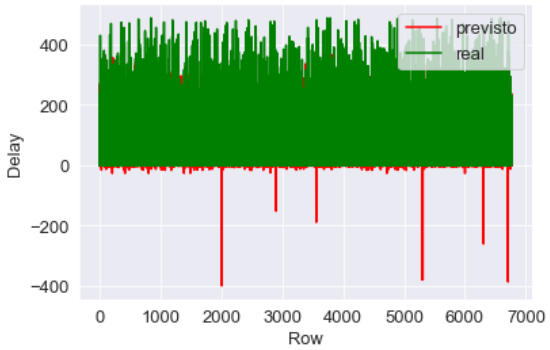
The scenario that gave better results with the linear regression model was the one using the whole dataset. Even though the differences between them were slim, there were still slight differences between scoring and error metrics. Table 6.2 shows the results of this model.

Dataset	MAE	MSE	RMSE	Score
Whole dataset	31.4	2820.25	53.11	0.804
Correlation >0.1	32.09	2867.08	53.55	0.801
Correlation >0.4	32.20	2956.46	54.37	0.795
Correlation >0.4 normalized	32.20	2956.46	54.37	0.795

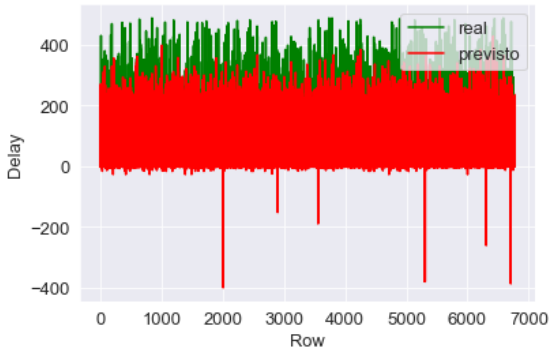
Table 6.2: Results from linear regression model

The better model had a score of around 0.804 and an MAE, MSE and RMSE of around 31.4, 2820.25 and 53.11, respectively. The graphs had data points where the delay was below and far from zero, which

was strange, and the predicted values, in general, were lower than real ones. This phenomenon happened in all models using linear regression. The next model used data with correlation greater than 0.1 with a score of around 0.801. It displayed higher errors having an MAE, MSE and RMSE of around 32.09, 2867.08 and 53.55, respectively. The two last were pretty similar, displaying a score of around 0.795 and almost identical error metrics with an MAE, MSE and RMSE of around 32.20, 2956.46 and 54.37, respectively, for both models. In Figures 6.7a and 6.7b are shown the differences between the values linear regression predicted and the ones present in data. Figure 6.8 shows the regression line stemming from joining the predicted and real values.



(a) Linear regression plot regrading differences between real and predicted values



(b) Linear regression plot regarding differences between real and predicted values (reversed)

Figure 6.7: Linear regression predicted and real values

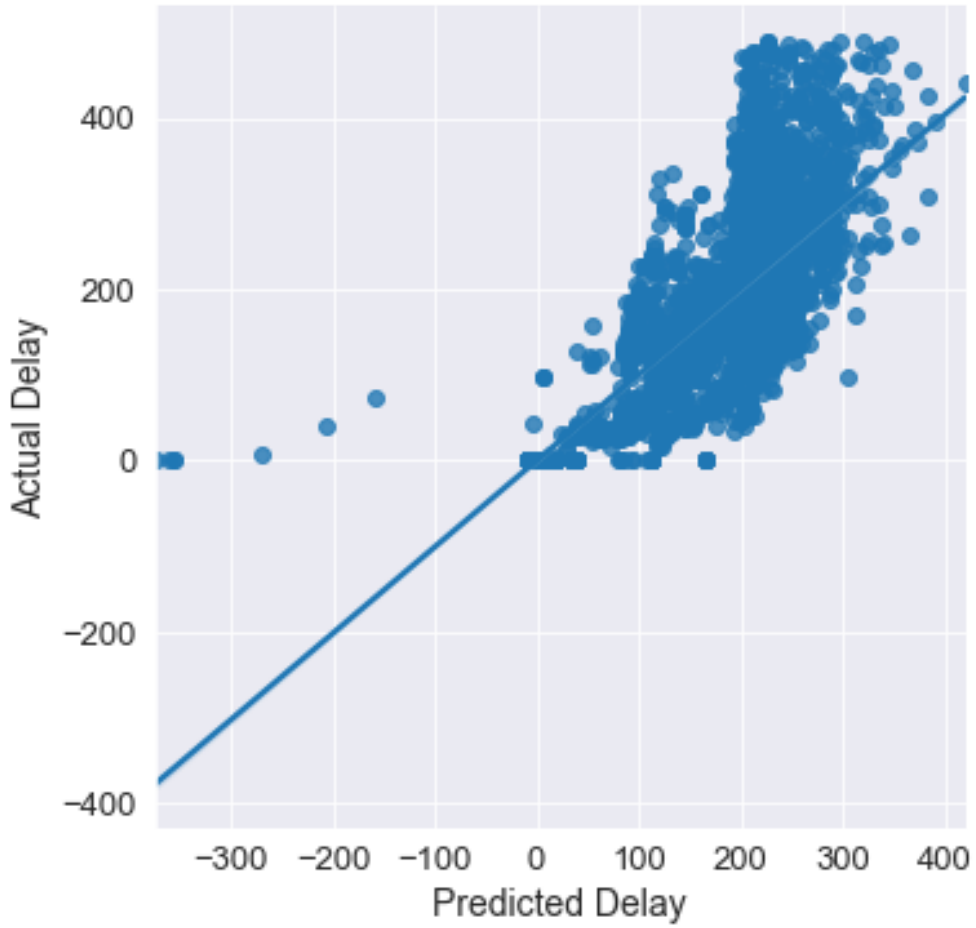


Figure 6.8: Linear regression between real and predicted values

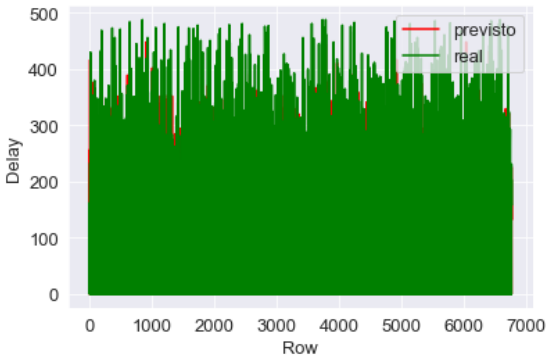
6.2.3 K-Nearest Neighbours

After performing this model in the four different environments, the one that gave better results was the one using features with a correlation greater than 0.4 with normalization, according to the attributes. The metrics showed that this model was, by a reasonable margin, better than all the other ones, and it's going to be displayed next. In contrast to what the graphs from support vector regression showed, these appeared to be more accurate and closer to higher values. Table 6.3 will show how results came out for this model.

Dataset	MAE	MSE	RMSE	Score
Whole dataset	31.83	3303.58	57.48	0.765
Correlation >0.1	30.75	3174.55	56.34	0.772
Correlation >0.4	30.60	3025.59	55.01	0.783
Correlation >0.4 normalized	27.81	2675.85	51.73	0.825

Table 6.3: Results from KNN model

The one that had the worst score was the model using the whole dataset with a score of 0.765 with an MAE, MSE and RMSE of around 31.83, 3303.58 and 57.48, respectively. The next best score belongs to the model that used features with a correlation greater than 0.1 of around 0.772. Moreover, it had an MAE, MSE and RMSE of 30.75, 3174.55 and 56.34, respectively. The second best was the model using features with a correlation greater than 0.4 but without normalization with a score of around 0.783. Furthermore, the metrics showed an MAE, MSE and RMSE of 30.60, 3025.59 and 55.01, respectively. Finally, the best model had a rather surprising score of around 0.825, which is a significant decrease relative to the other models. Additionally, it had an MAE, MSE and RMSE of around 27.81, 2675.85 and 51.73, respectively, which are also considerably lower than the other four models. Generally, except for the best model, all had some inability to predict higher values while the exception was noticeable better. In Figures 6.9a and 6.9b are shown the differences between the values K-nearest neighbours predicted and the ones present in data. Figure 6.10 shows the regression line stemming from joining the predicted and real values.



(a) KNN plot regrading differences between real and predicted values

(b) KNN plot regarding differences between real and predicted values (reversed)

Figure 6.9: KNN predicted and real values

Figure 6.10: KNN regression between real and predicted values

As we can see from Figures 6.9a and 6.9b, there are many instances where the real values surpass the predicted ones, but the difference is fairly small, probably the smallest out of all the models shown until here. From Figure 6.10 we can see a regression line with a slope very close to one, meaning that, generally speaking, this model is doing a good job, however, this may happen due to the unevenness on both sides of the line ending up balancing out each other.

6.2.4 Neural Networks

The network using the whole dataset had an MSE of around 2109.97 and an MAE between 25 and 28, which is surprising due to the results ahead. It showed that the network using the dataset whose features had a correlation greater than 0.1 had an MSE of around 2157.19 and an MAE between 26 and 30, which indicates a substantial decrease when compared to other algorithms. The network using the dataset whose features had a correlation greater than 0.4 had an MSE of around 2672.92 and an MAE between 29 and 31, which is unexpected and substantially worse than the previous one. Finally, the network using the dataset whose features had a correlation greater than 0.4 and normalized indicated an MSE of around

4301.71 and an MAE between 32 and 36, which is even more surprising due to the disappointing results when this dataset has been showing great results for other models. Besides, the network using the whole dataset had the best results, which were not expected at all, since all other sets of data were more carefully handled. Figure 6.11 shows the loss of each fold and consequent calculation of the 5-fold cross-validation average loss.

```
Score per fold
------------------------------------------------------------------------
> Fold 1 - Loss: 2187.515869140625
------------------------------------------------------------------------
> Fold 2 - Loss: 2078.595458984375
------------------------------------------------------------------------
> Fold 3 - Loss: 2098.338134765625
------------------------------------------------------------------------
> Fold 4 - Loss: 2010.8572998046875
------------------------------------------------------------------------
> Fold 5 - Loss: 2174.53466796875
------------------------------------------------------------------------
Average scores for all folds:
> Loss: 2109.9682861328124
```

Figure 6.11: Neural network average loss

6.2.5 Summary

When it comes to the r^2 score and by analyzing the results that each model had, we can see that the KNN model using data with correlation greater than 0.4 normalized had the biggest score of 0.825. However, the other results from KNN were worse or equal than results from other models, implying that the normalization had a great impact on that result. When it comes to MSE, the neural network using the whole dataset had the lowest value at 2109.97. It's important to mention both r^2 score and MSE since the neural network was only measured by its loss which was MSE, while the other models had other metrics and the score to complement the result analysis. In sum, the best model to be used for the improvement of road safety and prevention of road accidents would be neural networks, although KNN may also be viable since it got the best score out of the first three models.

6.3 Dashboard Platform

After modelling and analyzing all the results resulting from the previous algorithms, it's now important to create something that can give key information to complement the results drawn previously. As said before, the best way to convey information, given the data available, is to build a report containing different data visualization techniques that can give accurate information in a way that people can understand thoroughly. Besides, it's convenient that this report presents information that is not easy to obtain just by looking at

data since it can become redundant and uninteresting. With this being said, next is going to be listed the different pages of the report, what information they give and what people can withdraw from it. Before showing the visualizations and explaining the pages of the report, Figure 6.12 shows a general view of the final result.

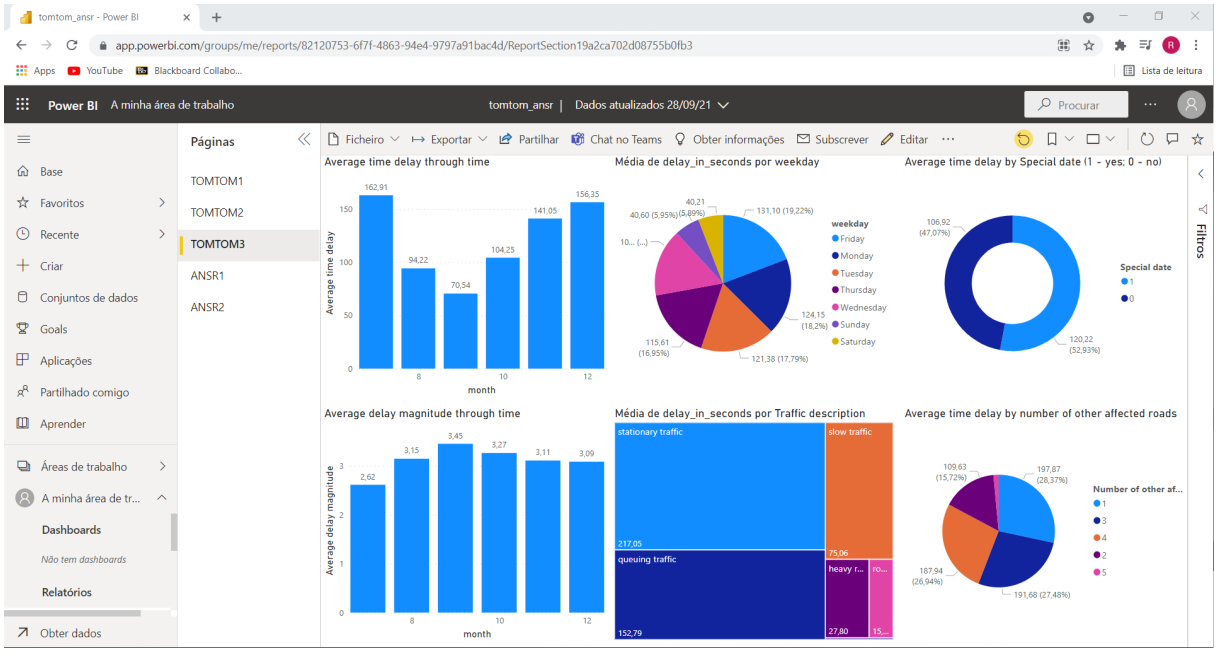


Figure 6.12: General report view

These next data visualizations mainly focus on the relationship between deaths/serious injuries and other important factors mentioned previously since it's what we want to prevent and analyze more thoroughly. Inside each page of the report, visualizations will be mentioned left to right and top to bottom.

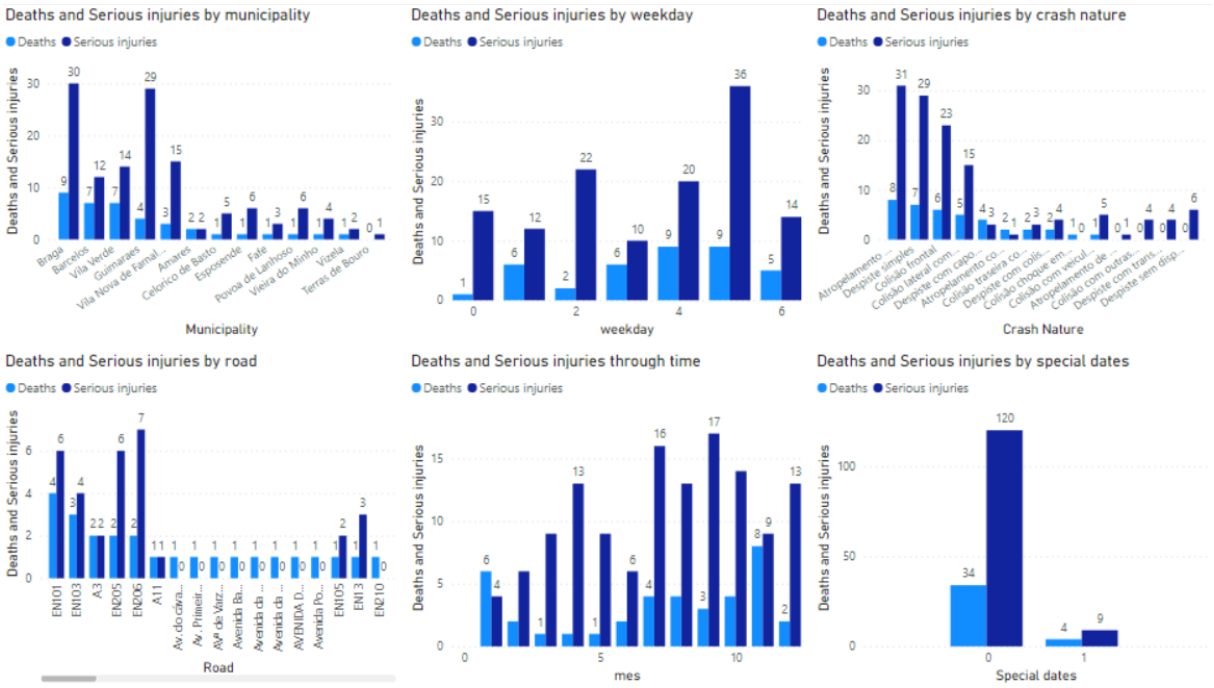


Figure 6.13: ANSR report - page 1

In Figure 6.13, the first chart show which municipalities contain more deaths and serious injuries. *Braga*, *Guimarães*, *Vila Nova de Famalicão*, *Vila Verde* and *Barcelos* have most incidents with serious injuries while *Braga*, *Barcelos* and *Vila Verde* have most incidents involving deaths. Excluding *Vila Verde*, all the municipalities mentioned have a higher number of inhabitants than the others, which can also lead to a higher probability of occurring serious incidents. Nonetheless, *Braga* and *Guimarães* show excessive records of serious injuries compared to the rest. The next chart shows the impact of the weekday (0 designates Monday and 6 Sunday) in deaths and serious injuries where Saturday has a noticeably higher number of serious injuries followed by Wednesday and Friday. When it comes to deaths, Friday and Saturday are the worst days with 9, followed by Tuesday and Thursday with 6. It's expected to see worse values on Friday and Saturday since it's the end of the working week, people may be tired or more relaxed which can lead to these poor results. Even though Wednesday has an unfortunate amount of serious injuries it has a very low number of deaths which is a positive aspect. The following chart shows what types of crashes lead to more serious incident outcomes, which are running over pedestrians, simple slipping, frontal collisions and lateral collisions with other moving vehicles. All remaining crash types have somewhat similar values when it comes to deaths and serious injuries. Talking now about the charts below, the first one shows a particular set of roads that appear to be more dangerous since more serious incidents happen and those are EN101, EN103, A3, EN205 and EN206. There's a predominance of national roads in the previously mentioned list, with four out of five being national roads while the only left is a highway and is also the least worrying out of the five roads when it comes to their negative results. The next chart shows how these incidents are distributed through time, where we can see at first sight a few highlighting months such as

January, July, September and November. July and September contain the highest values of serious injuries recorded at 16 and 17, respectively, while the other two contain an excessive amount of deaths recorded at 6 and 8, respectively, which are close if not above, in January's case, to the number of serious injuries which is worrying. Some other things to notice from this chart is that on the 29th of July were registered way more serious injuries than normal, however, this date doesn't seem to be any different from others since it is not a public holiday nor a special date in the district of *Braga*. In the final chart of this page, we can see the distribution of deaths and serious injuries on days where there are festivities, holidays or some other special date and regular days. It's expected to see higher values in regular days since special days are sporadic and only happen a few times a year, however, percentages can be used by dividing the number of deaths or serious injuries by the number of incidents that happened on special dates or in regular days to understand the ratio between the different kinds of dates. In terms of deaths, special dates show a higher ratio of 36.36% while the other shows a ratio of 23.94% even though the flat value mentioned in the report is way higher. When it comes to serious injuries, both types of dates show similar ratios with the first being 81.82% and the second 84.51%.

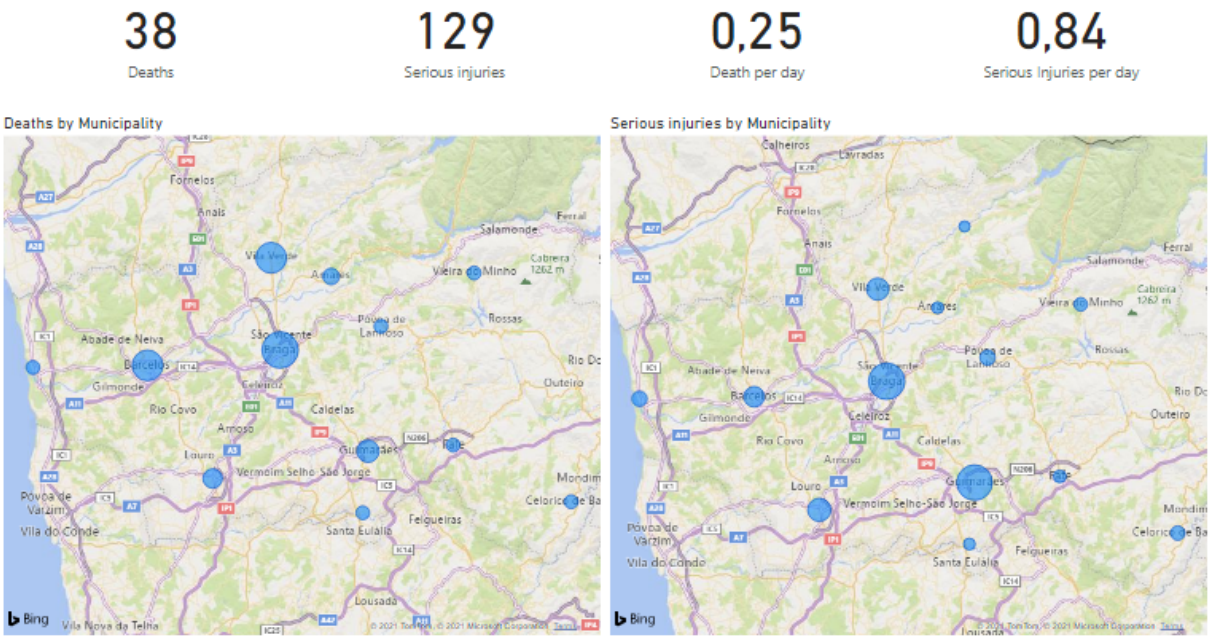


Figure 6.14: ANSR report - page 2

In Figure 6.14, the second page of the ANSR report, we can see some cards indicating the total amount of deaths and serious injuries registered as well as their daily average which is 0.25 or 1 death per 4 days and 0.84, respectively. There are also two maps indicating which municipalities contain more serious incidents, similar to the first chart but with a different visual approach. Their size is influenced by the number of deaths and serious injuries that occurred in each municipality (the bigger the size, the higher the value is).

Next, several pages from the report will be presented, studying two main properties, which are stemming traffic and the impact incidents have in surrounding areas, as well as their correlation between other features available.

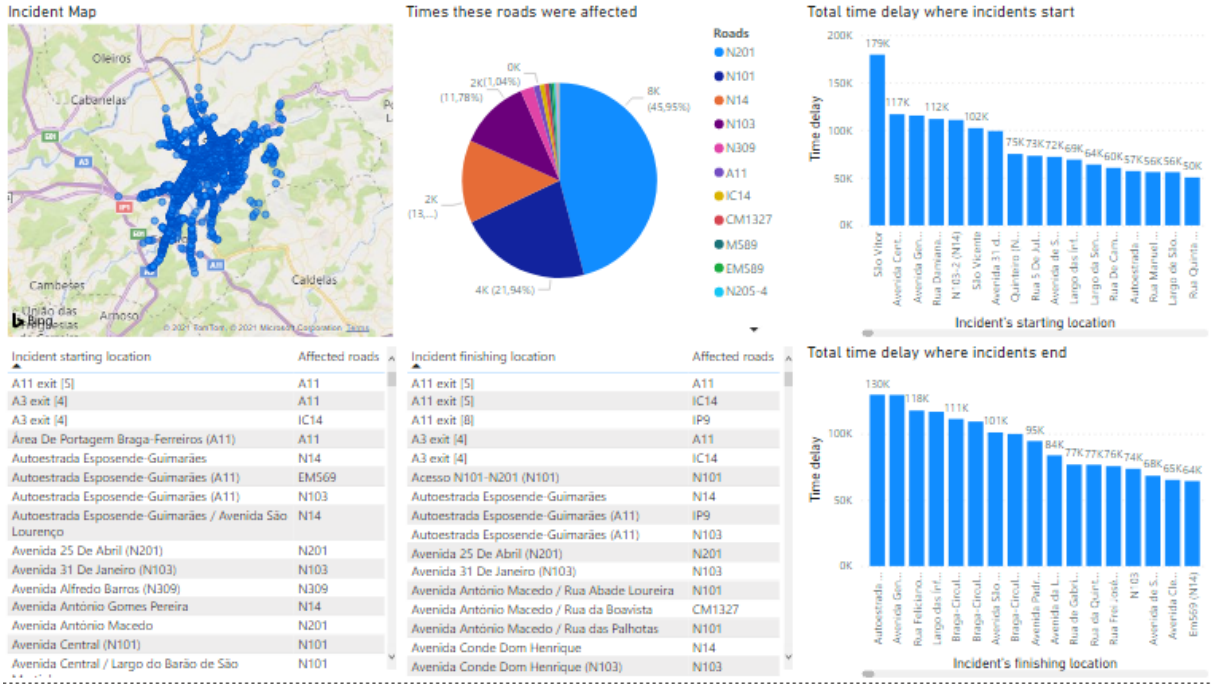


Figure 6.15: Tomtom report - page 1

Once again, we will be reading the visuals from left to right and then top to bottom. In Figure 6.15, the first visualization is a simple map where we can explore the exact location of all incidents since all records contain latitude and longitude, making it possible to have these precise locations. The second one is a pie chart that contains which roads were more affected due to an occurrence of an incident. We can see that N201 makes up almost 50% of these roads which is rather problematic, while the other three also have significant slices, them being N101, N14 and N103, from bigger to smaller proportions. These four roads make up to around 93/94% of total roads affected and they all are national roads which is a signal that these types of roadways need planning to become safer so these percentages can be flattened and total values lowered. Next to the pie chart, we have a bar chart with the total time delay with the roads where incidents start, and we can see one of them standing out, which is *São Vitor*, a street located in the center of *Braga*. One thing worth mentioning is that 5 of the top 7 roads in this chart are close to the *Braga* center, which can be a sign for authorities to take action and improve circulation in these streets/avenues near the city center. The first two tables below mention in the first column the starting and finishing incident location, respectively, while the second column shows which roads these tend to affect when incidents happen. The next one shows the same information as the one above, with the only difference being that these roads are when incidents end while the other was when they started. The two biggest total time

delays are highway *Esposende-Guimarães* and *Avenida General Norton de Matos* and the next ones are a mixture of surrounding and close to center roads.

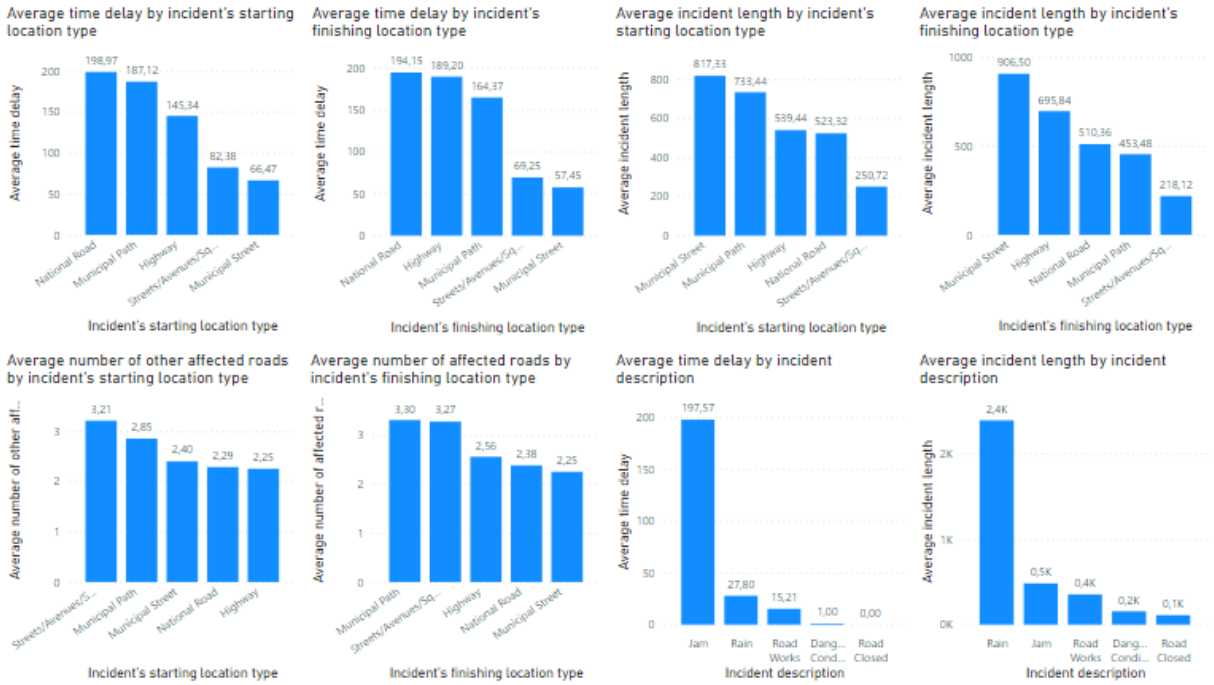


Figure 6.16: Tomtom report - page 2

In Figure 6.16 we have a collection of bar charts containing information about time delay, incident length and number of affected roads in regards to the type of road where it occurs and incident description. In the first two, we can see the connection between time delay and road type where the incident starts, and we can see, as we have been observing throughout the previous report, the average time delay is higher on national roads followed by municipal paths and then highways while streets/avenues/squares and municipal streets have a significantly lower time delay. When it comes to the finishing incident location, highways climb one position, municipal paths decrease one, and every other position stays the same, with values staying relatively equal. The third and fourth charts, instead of time delay, mentions incident length through the incident start and finish location type. On the starting location, we can see municipal streets as the highest average length value, followed by municipal paths, highways, national roads and streets/avenues/squares. On the finishing location, the highest and lowest averages belong to the same location types but the intermediate order is now highway, national roads and municipal paths. From these four sets of charts, we can draw that high time delays don't imply a high incident length as is the case with national roads where they have high average time delay but medium to low incident length or municipal streets which have very low time delays but very high incident lengths. On the bottom row, in the first two charts, we have the number of affected roads taking into account where the incidents started and ended. We can notice that, regarding the starting location, streets/avenues/squares have the highest value and the only one higher than three, followed by municipal paths with an average of 2.85 and the other three

close together right after. The differences we note when talking about the incident's finishing location are that municipal paths' average rose a bit and became the highest close to streets/avenues/squares, and highways' average also rose a bit and moved up two places but is still close to the last two types which are national roads and municipal streets. From the two charts, we see minor but not significant differences since they both convey agreement results. The last two charts show the average time delay and average incident length for each incident description, and we contemplate, on them both, one description that stands out. The first is jam, and the second is rain meaning that traffic jams generated by incidents cause a very high time delay, as a result, but raining causes a very high incident traffic length while all the others have a low average time delay or low average incident length.

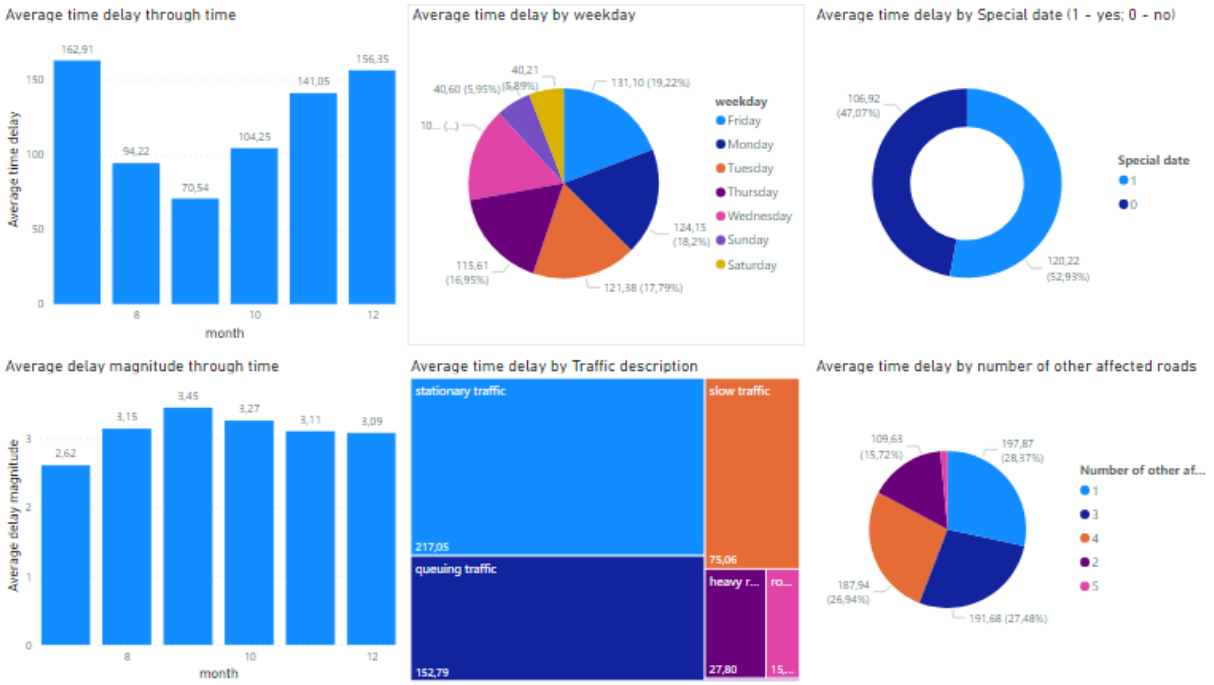


Figure 6.17: Tomtom report - page 3

On the last page, which is Figure 6.17, we have visualizations that exhibit information about some important dataset features over time, specifically, instances of time. The first one shows the average time delay over the time available in the dataset, and it's noticeable that July, November and December have a pretty high value while August, September and October have a lower average, with September standing out as the lowest. Besides July and December, which have fewer registered days than the remaining months, all of them have periods or specific days where a high average drop happens, meaning that it is not fairly consistent throughout the month. Furthermore, it seems that usual vacation months (August, September) turn out to have the lowest time delay, which is somewhat understandable taking into consideration that *Braga* isn't a particular holiday destination in contrast to, for example, *Algarve* so less traffic tends to happen during these months. In the second visualization, we have a pie chart with the distribution of time delays among weekdays where we can see that there is little difference among working days (Mondays

through Fridays) while weekends have much lower averages because they're considered resting days. On the next visualization, we have a donut chart, this time splitting the difference between average time delays in days considered special, according to the list of days already specified in data preprocessing. There's a higher time delay on special dates than on regular days, although the difference is slim and doesn't appear to have a significant impact. If a bigger amount of data was available with more days matching these special dates, we could see a greater difference. Moving now to the bottom row, the first chart shows the average delay magnitude over the time available, meaning the seriousness of traffic conditions after the incident. We note that every month has similar values except for July being a little bit lower. Furthermore, these values are quite high since the highest value possible is 4, and five out of the possible six months have an average above 3. The only extra matter worth mentioning is that, in August, there's a substantial increase of delay magnitude from 24 to the end of the month. Besides this chart, there's a treemap including the average time delay for each traffic description available, with stationary traffic having the highest by a substantial margin followed by queuing traffic and slow traffic while road works and heavy rain contain a residual time delay. Finally, the last pie chart conveys the correlation between the number of affected roads and time delay with three different values being very close, which are 1,3 and 4 around 190 and 27% each followed by the number 2 with around 109 and 16% and lastly, the number 5 with residual time delay.

6.4 Summary

In this chapter, we described the results of both supervised and unsupervised approaches, which were meant to fulfil correlation and preventive purposes, respectively, by presenting 2d and 3d representations of clusters and what feature characteristics distinguish them and bring them together as well as graphs with predictions and regression of data, comparing the results between models to settle which algorithms should be used with the data available, to prevent incidents in these areas. Furthermore, dashboards were built containing useful insights and information about blackspots, spatiotemporal analysis, traffic consequences and causes, among other things, and then explained what can be deducted from the visualizations to draw the bigger picture.

Chapter 7

Conclusions

7.1 Conclusion

This work and investigation aimed at getting to know more about the state of incidents, in this case, in Braga and what could be the defining factors as well as study models able to predict certain features and group data records with similarities so more insight can be gathered, regarding the data available. After carrying out all data processing tasks, it was settled that there were data clusters, and there were clear differences between each other and internal commonalities as well, which were identified and explained. Furthermore, the results about the other models showed that, in general, the more targeted dataset gave better results where the k-nearest neighbours algorithm seemed to be the more appropriate since it gave better results when trying to predict time delays than all other models. This was all taken from the tomtom dataset since the data collected from ANSR only contained a few dozen records which, at the time, were not sufficient to perform modelling. Moreover, it was possible to gather some conclusions regarding both datasets that can be useful and bring some insights about incident location, timing, consequences and outcomes. In hindsight, the intention, in the beginning, was to use data provided by the ANSR themselves with every single occurrence in the year of 2018 in the city of Braga, which would've brought useful insights with the upside of being complete (data throughout a whole year) being able to make a more broad study about road accidents in Portugal, namely in Braga. Unfortunately, that wasn't possible due to bureaucracy about data privacy which has to be respected and, nonetheless, I have to thank their availability to get in contact and discuss these issues.

7.2 Future Work

After concluding, there are tasks and conditions that, in the future, can be rectified to further analyze the topic of road accidents in Portugal. One of which is collecting more complete data with more interesting features such as characteristics about the driver(s), their driving qualification, vehicles involved, road conditions, signalling, etc, that are all included in data owned by the ANSR and that couldn't be obtained to assist in studying road accidents in Portugal. Another task that arises after having studied road accidents in the city of Braga is to do this process for other cities/districts so a big picture can be drawn about the state of road accidents in Portugal. The process to get the information and insights is already set, so now it's just a matter of repeating the process but with different data (from different locations and more complete, as said earlier). Besides that, more extensive and deep work can be done on modelling, since it was just defined which ones gave better results, but other stuff could be done to create an accurate predictive model such as training them with data from one year and trying to predict values from other years and do it well, develop other ways to predict data such as LSTM networks that are used in time series data which is the case for these data. Finally, as time goes by, more data are available from different years such as 2020, 2021, 2022 and so on, which allows repeating these same processes but with a whole new data enabling an improvement on the Portuguese road accidents situation so new problems can be discovered and people's safety on the road is improved.

Bibliography

[1] K. Srinath. "Python–the fastest growing programming language." In: *International Research Journal of Engineering and Technology (IRJET)* 4.12 (2017), pp. 354–357.

[2] E. Morgulev, O. H. Azar, and R. Lidor. "Sports analytics and the big-data era." In: *International Journal of Data Science and Analytics* 5.4 (2018), pp. 213–222. doi: 10.1007/s41060-017-0093-7.

[3] F. Provost and T. Fawcett. "Data science and its relationship to big data and data-driven decision making." In: *Big data* 1.1 (2013), pp. 51–59. doi: 10.1089/big.2013.1508.

[4] L. Ramos, L. Silva, M. Y. Santos, and J. M. Pires. "Detection of road accident accumulation zones with a visual analytics approach." In: *Procedia Computer Science* 64 (2015), pp. 969–976. doi: 10.1016/j.procs.2015.08.615.

[5] W. H. Organization et al. *Global status report on road safety 2018: Summary*. Tech. rep. World Health Organization, 2018.

[6] *WHO: Road Traffic Injuries*. https://www.who.int/health-topics/road-safety#tab=tab_1.

[7] D. P. Braceiro. "Acumulação de acidentes rodoviários em Portugal Continental: contributos dos Sistemas de Informação Geográfica." Master's thesis. Universidade Nova de Lisboa, 2016.

[8] M. Vilaça, E. Macedo, P. Tafidis, and M. C. Coelho. "*Multinomial logistic regression for prediction of vulnerable road users risk injuries based on spatial and temporal assessment.*" In: 26.4 (2019), pp. 379–390.

[9] C. Matos, N. Sillero, and E. Argaña. "Spatial analysis of amphibian road mortality levels in northern Portugal country roads." In: *Amphibia-Reptilia* 33.3-4 (2012), pp. 469–483. doi: 10.1163/15685381-00002850.

[10] S. M. Antunes, C. Cordeiro, and H. M. Teixeira. "Analysis of fatal accidents with tractors in the Centre of Portugal: Ten years analysis." In: *Forensic science international* 287 (2018), pp. 74–80. doi: 10.1016/j.forsciint.2018.03.048.

[11] J. O. d. Costa, E. F. Freitas, M. A. P. Jacques, and P. A. A. Pereira. "Collision prediction models with longitudinal data: an analysis of contributing factors in collision frequency in road segments in Portugal." In: *17th International Conference Road Safety On Five Continents (RS5C 2016), Rio de Janeiro, Brazil, 17-19 May 2016*. Statens väg-och transportforskningsinstitut. 2016, pp. 1–12.

[12] E. Comission. *Road Safety Country Overview - Portugal*. Tech. rep. European Road Safety Observatory, 2017.

[13] J. L. C. Ramos, R. L. Rodrigues, J. C. S. Silva, and P. L. S. de Oliveira. "CRISP-EDM: uma proposta de adaptação do Modelo CRISP-DM para mineração de dados educacionais." In: *Anais do XXXI Simpósio Brasileiro de Informática na Educação*. SBC. 2020, pp. 1092–1101.

[14] R. Wirth and J. Hipp. "CRISP-DM: Towards a standard process model for data mining." In: *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Springer-Verlag London, UK. 2000, pp. 29–39.

[15] ANSR. *Relatório Anual Sinistralidade Rodoviária 2018 -24 horas*. Tech. rep. ANSR, 2018.

[16] ANSR. *Plano Estratégico Nacional de Segurança Rodoviária — PENSE 2020 (Resolução do Conselho de Ministros n.º 85/2017)*. Tech. rep. ANSR, 2017.

[17] E. Comission. *Road Safety Country Overview - CROATIA*. Tech. rep. European Road Safety Observatory, 2016.

[18] G. Zovak, D. Brcic, and Z. Saric. "Analysis of road black spots identification method in republic of Croatia." In: *IX International Conference Road Safety in Local Communities*. 2014.

[19] A. Murphy. *Reported road casualties in Great Britain: 2019 annual report*. Tech. rep. Department for Transport, 2020.

[20] J Pei, J. Ding, and H. District. "Improvement in the quality control method to distinguish the black spots of the road." In: *Proceedings of the Eastern Asia Society for Transportation Studies*. Vol. 5. Citeseer. 2005, pp. 2106–2113.

[21] T. Yuan, X. Zeng, and T. Shi. "Identifying urban road black spots with a novel method based on the firefly clustering algorithm and a geographic information system." In: *Sustainability* 12.5 (2020), p. 2091. doi: 10.3390/su12052091.

[22] M. Ghadi and Á. Török. "A comparative analysis of black spot identification methods and road accident segmentation methods." In: *Accident Analysis & Prevention* 128 (2019), pp. 1–7. doi: 10.1016/j.aap.2019.03.002.

[23] G. Gito Sugiyanto, A. Ari Fadli, and Y. Mina Yumei Santi. "Identification of black spot and equivalent accident number using upper control limit method." In: *ARPN Journal of Engineering and Applied Sciences* 12.2 (2017), pp. 528–535.

[24] L. Leuhery and H. Hamkah. "Determination of Black Site Area Based on Equivalent Accident Number Analysis: Case Study National Roads in Ambon City." In: *Journal of Civil Engineering and Architecture* 8.5 (2020), pp. 1063–1073. doi: 10.13189/cea.2020.080533.

[25] X. Dawei and L. Xiansheng. "Identification of Speedway Accident Black Spots Based on the Quality Control Method." In: *2015 Seventh International Conference on Measuring Technology and Mechatronics Automation*. IEEE. 2015, pp. 541–544. doi: 10.1109/ICMTMA.2015.137.

[26] B. E. Chandler, M. Myers, J. E. Atkinson, T. Bryer, R. Retting, J. Smithline, J. Trim, P. Wojtkiewicz, G. B. Thomas, S. P. Venglar, et al. *Signalized intersections informational guide*. Tech. rep. United States. Federal Highway Administration. Office of Safety, 2013.

[27] M. Ghadi and Á. Török. "Comparison different black spot identification methods." In: *Transportation research procedia* 27 (2017), pp. 1105–1112. doi: 10.1016/j.trpro.2017.12.104.

[28] *ERSI - Environmental System Research Institute | What is GIS?* https://www.esri.com/en-us/what-is-gis/overview.

[29] P. Cunningham, M. Cord, and S. J. Delany. "Supervised learning." In: *Machine learning techniques for multimedia*. Springer, 2008, pp. 21–49. doi: 10.1007/978-3-540-75171-7_2.

[30] T. Hastie, R. Tibshirani, and J. Friedman. "Overview of supervised learning." In: *The elements of statistical learning*. Springer, 2009, pp. 9–41. doi: 10.1007/978-0-387-84858-7_2.

[31] Z. Fan, C. Liu, D. Cai, and S. Yue. "Research on black spot identification of safety in urban traffic accidents based on machine learning method." In: *Safety science* 118 (2019), pp. 607–616. doi: 10.1016/j.ssci.2019.05.039.

[32] Z. Ghahramani. "Unsupervised learning." In: *Summer School on Machine Learning*. Springer. 2003, pp. 72–112. doi: 10.1007/978-3-540-28650-9_5.

[33] D. Zhao, H. Wang, K. Shao, and Y. Zhu. "Deep reinforcement learning with experience replay based on SARSA." In: *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE. 2016, pp. 1–6. doi: 10.1109/SSCI.2016.7849837.

[34] G. Snipes. "Google data studio." In: *Journal of Librarianship and Scholarly Communication* 6.1 (2018). doi: 10.7710/2162-3309.2214.

[35] O. Troyansky, T. Gibson, and C. Leichtweis. "QlikView your business: an expert guide to business discovery with QlikView and Qlik Sense." In: John Wiley & Sons, 2015, pp. 17–38.

[36] C. Ribeiro, A. A. Turkman, and J. L. Cardoso. "Bayesian hierarchical models: An analysis of Portugal road accident data." In: *Spatial2 Conference: Spatial Data Methods for Environmental and Ecological Processes, Foggia (IT), 1-2 September 2011*. Università degli studi di Bergamo. 2011, pp. 1–4.

[37] J. O. Costa, A. P. Maria, P. A. Pereira, E. F. Freitas, and F. E. Soares. "Portuguese two-lane highways: modelling crash frequencies for different temporal and spatial aggregation of crash data." In: *Transport* 33.1 (2018), pp. 92–103. doi: 10.3846/16484142.2015.1073619.

[38] B. Fernandes, F. Silva, H. Alaiz-Moretón, P. Novais, C. Analide, and J. Neves. "Traffic flow forecasting on data-scarce environments using ARIMA and LSTM networks." In: *World Conference on Information Systems and Technologies*. Springer. 2019, pp. 273–282. doi: 10.1007/978-3-030-16181-1_26.

[39] *População residente, estimativas a 31 de Dezembro*. https://www.pordata.pt/Municipios/Popula%C3%A7%C3%A3o+residente++estimativas+a+31+de+Dezembro-120.

[40] *Veículos matriculados: total e por tipo de veículo*. https://www.pordata.pt/Portugal/Ve%c3%adculos+matriculados+total+e+por+tipo+de+ve%c3%adculo-3103-262502.

[41] S. Raschka, J. Patterson, and C. Nolet. "Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence." In: *Information* 11.4 (2020), p. 193. doi: 10.3390/info11040193.

[42] M. Ringnér. "What is principal component analysis?" In: *Nature biotechnology* 26.3 (2008), pp. 303–304. doi: 10.1038/nbt0308-303.

[43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python." In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[44] T. Sato, T. Suzuki, and K. Mabuchi. "Fast automatic template matching for spike sorting based on Davies-Bouldin validation indices." In: *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2007, pp. 3200–3203. doi: 10.1109/IEMBS.2007.4353010.

[45] J. Baarsch and M. E. Celebi. "Investigation of internal validity measures for K-means clustering." In: *Proceedings of the international multiconference of engineers and computer scientists*. Vol. 1. sn. 2012, pp. 14–16.

[46] C.-F. Juang and C.-D. Hsieh. "A fuzzy system constructed by rule generation and iterative linear SVR for antecedent and consequent parameter optimization." In: *IEEE Transactions on Fuzzy Systems* 20.2 (2011), pp. 372–384. doi: 10.1109/TFUZZ.2011.2174997.

[47] M. Awad and R. Khanna. “Support vector regression.” In: *Efficient learning machines*. Springer, 2015, pp. 67–80. doi: 10.1007/978-1-4302-5990-9_4.

[48] B. Fernandes, F. Silva, H. Alaiz-Moreton, P. Novais, J. Neves, and C. Analide. “Long Short-Term Memory Networks for Traffic Flow Forecasting: Exploring Input Variables, Time Frames and Multi-Step Approaches.” In: *Informatica* 31.4 (2020), pp. 723–749. doi: 10.15388/20-INFOR431.

[49] M. Tranmer and M. Elliot. “Multiple linear regression.” In: *The Cathie Marsh Centre for Census and Survey Research (CCSR)* 5.5 (2008), pp. 1–5.

[50] J. K. Basu, D. Bhattacharyya, and T.-h. Kim. “Use of artificial neural network in pattern recognition.” In: *International journal of software engineering and its applications* 4.2 (2010).

[51] R. Hecht-Nielsen. “Theory of the backpropagation neural network.” In: *Neural networks for perception*. Elsevier, 1992, pp. 65–93. doi: 10.1016/B978-0-12-741252-8.50010-8.

[52] S. Bock, J. Goppold, and M. Weiß. “An improvement of the convergence proof of the ADAM-Optimizer.” In: *arXiv preprint arXiv:1804.10587* (2018).

[53] J. Miles. “R-squared, adjusted R-squared.” In: *Encyclopedia of statistics in behavioral science* (2005). doi: 10.1002/0470013192.bsa526.