NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Predicting if accidents have incarcerated people with Machine Learning: Insights from Lisbon

Gonçalo Filipe Lacão Brancas

Dissertation

presented as partial requirement for obtaining the Master Degree Program in Data Science and Advanced Analytics

NOVA Information Management School

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Predicting if accidents have incarcerated people with Machine Learning: Insights from Lisbon

by

Gonçalo Brancas

Master Thesis presented as partial requirement for obtaining the Master's degree in Advanced Analytics, with a Specialization in Business Analytics

Supervisor: Professor Miguel de Castro Ferreira Neto

Co Supervisor*:* Professor Bruno Morais de Sousa Jardim

November 2024

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 acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Gonçalo Filipe Lacão Brancas

Lisboa, 30 de Novembro 2024.

ACKNOWLEDGEMENTS

I would like to give a special thanks to my family, they have been supporting me from the beginning specially my parents who never gave up on me even on tough times and always believed in me giving all the support that I needed and pushed me to be better.

To my friends, a special mention is also necessary since they provided me with several tips for the master itself and delivered the best advice to help me finish the master thesis.

Lastly a huge thanks to both my professors, Miguel Neto and Bruno Jardim who guided me and helped me improve this work with several tips over time along side the NOVA IMS Cidade with the Lisbon municipality that shared the data and made this work possible.

To all, this wouldn't be possible without your help.

RESEARCH MOTIVATION

These days road accidents are a major topic in our everyday life where we are most exposed to this topic on Tv(television) news, newspaper and in case we go to work by public transportation or our own vehicle certainly almost all of us have already experienced some type of road traffic that may or may not be connected to a road accident that happened on the path we do between our homes and work.

To give us an idea, before the pandemic in Portugal, there were a total of 34.235 road traffic accidents with injuries in the year of 2018 resulting in 508 fatalities ('Road Traffic Accidents with Victims, Injuries and Deaths - Mainland Portugal', 2023).

With this information, and the desire to find patterns that can contribute to help decrease this number or at least alert the people and competent authorities to act will serve and is the motivation behind this study to provide a better future for Portugal and the Portuguese people.

ABSTRACT

Accidents happen on a day-to-day basis all around the world, whether in major cities or around rural areas and they affect all of us directly or indirectly in various ways like traffic jams, on the way to work or on the news, and over the years there is a feeling that this phenomenon has been increasing with major studies focusing on understanding external factors that lead to road accidents and not the evolution itself. Here based on a real dataset shared by the NOVA Cidade lab we will dive further in this world with real data provided by the Lisbon Municipality making it unique to the Lisbon city and with an important gap that can be filled with the work around this subject. In this work a preliminary analysis of car accidents will take place where we will see the numbers over the years and the more favorable places where these events occur with heatmaps and charts that best provide these information's with geographic precision for the areas we need to highlight. Alongside that, this work will be guided through a CRISP-DM methodology and one of the outputs resulting from here will be a prediction model capable of indicating the code associated with the accident and has a result, what type of accident is most likely to happen on the next period of time so local authorities can allocate the resources faster or even the Lisbon municipality can see if they have an increase in road accidents and if measures need to be taken in places where they frequently occur to prevent and make Lisbon a safer city for its citizens.

KEYWORDS

Road accidents; Lisbon; Traffic analysis; Predictive model; Smart Cities

Sustainable Development Goals (SGD):

INDEX

1. Introduction 1
2. Literature review 3
   2.1. Approach for the literature review 3
   2.2. Literature Review Results 3
   2.3. Research studies context 4
3. Methodology 7
   3.1. Choosing and adaptation 7
   3.2. Business understanding 8
   3.3. Data Understanding 8
   3.4. Data preparation 9
   3.5. Modeling 10
      3.5.1. Prediction model 10
   3.6. Measures 12
4. Results and discussion 15
   4.1. Preliminary Analysis 15
   4.2. Modeling results 17
   4.3. Discussion 21
5. Conclusion 23
6. Limitations and recommendations for future works 24
7. References 25

Appendix 29
   7.1. Dashboard-All accidents & model 29
   7.2. Dashboard-Accidents in 2018 30
   7.3. Dashboard-Accidents in 2019 31
   7.4. Dashboard- Accidents in 2020 32
   7.5. Dashboard-Heatmap over the years 33
   7.6. Dashboard-Accidents with incarcerated 34
   7.7. Python descriptive measures 35
   7.8. Feature extraction 36
   7.9. Train test split and Regressors models 37

LIST OF FIGURES

Figure 1-PRISMA flow diagram 4
Figure 2-CRISP-DM phases. 7
Figure 3-Traffic accidents descriptive features 8
Figure 4-IPMA descriptive features 9
Figure 5- Target and Input variables 10
Figure 6-Logistic Regression tested hyperparameters 11
Figure 7- Random Forest tested hyperparameters 11
Figure 8- Gradient Boosting tested hyperparameters 12
Figure 9- Support Vector Machine tested hyperparameters 12
Figure 10- Confusion Matrix 14
Figure 11- Dashboard all years 15
Figure 12- Dashboard 2018 16
Figure 13- Best Hyperparameters 17
Figure 14- Accuracy and F1-Score for models 18
Figure 15- Confusion Matrix 18
Figure 16- ROC curve 19
Figure 17- Seaborn bar plot of variable importance 20
Figure 18- Error analysis heatmap 21

LIST OF ABBREVIATIONS AND ACRONYMS

ANSR - Autoridade Nacional Segurança Rodoviária

APP - Application

CRISP-DM – Cross Industry Standard Process for Data Mining

DBS – Databases

DIM - Dimension

GIS - Geographic Information System

IMT - Instituto da Mobilidade e Transportes

IPMA - Instituto Português do Mar e da Atmosfera

PRISMA - Preferred Reporting Items for Systematic reviews and Meta-Analyses

RMSE – Root Mean Squared Error

ROC – Receiver Operating Characteristics

RUN - Repositório Universidade Nova

TV - Television

1. INTRODUCTION

Every day, millions of people use some type of transport to travel from their current location to another one whether by necessity, job or need millions of people worldwide dislocate and with so many movements generated, accidents may result as a consequence of so many interactions.

Road accidents in particular are a big consequence from the millions of vehicles that travel the roads each day and interact with each other whether provoked by external factors such as the state of the road or even the human factor, the driver might be drunk or with excess of speed on an infrastructure that is designed for slow traffic.

Worldwide, road accidents remain a significant public health concern and Portugal is no exception. Each day when people turn on the news there is some type of heavy road accident involving several vehicles that perturbated for long hours several thousands of people or more serious ones where the involved passengers don't resist the accident and are fatalities from the incident.

According with report from PORDATA (2019) there were registered 35704 fatalities resulting from road accidents and even though the numbers over the past few decades have been decreasing, it still represents a very high number and no country wants so many fatalities on its roads with the consequence that Portugal has a country has been seeing its population decrease in the past decade.

The capital of Portugal, Lisbon is no exception to these events where the frequency and severity of these accidents have profound implications on its safety and population being very fundamental aspect these two implications on any developed society. Other aspects such as legal systems and healthcare are direct consequences that get impacted by this event.

Being the capital the most populous city in the country, it is with no surprise that a high volume of traffic affects the city on an everyday basic and that contributes to a prevalence of road accidents. The city´s dense urban structure, combined with a high number of vehicles on the road, creates several complex and challenging environments for traffic safety. In this study, there will be a focus on the city of Lisbon with the development of a prediction model and a small preliminary analysis regarding recent years and hotspots.

There will be a few challenges regarding this topic, starting with the number of actual reports and notices. According to a study by Yannis et al. (2014) there is a significant level of underreporting of road accidents injuries in several European countries and Portugal is not an exemption. Sometimes small accidents are not properly noticed and even the ones that are reported may lack information.

This under-reporting can have several factors at its origin such as the severity of the injuries, the type of accident and the users involved in the accidents. This represents a challenge that will be noticed in this work and is felt on all works related to this topic, representing a challenge in accurately understanding the full extent of road safety problems in Portugal and Lisbon. We can Observe that even tough with all these works related with car accidents there are some gaps that can be done and researched to fill and increase the knowledge around this topic.

For starter there isn't a work fully focused in the Lisbon municipality with direct data provided by the city council. This will allow more in-depth insights into the city. Another aspect is a prediction model that can help anticipate accidents or predict the type of accident. A great majority of the works consist

more in detail analysis of the data provided and don't have predictions models to complement and that's where we will differentiate.

The first steps will be to build prediction model for accidents depending on the data at hand and its possibilities with the compliment of descriptive preliminary analysis using PowerBI to allow the local Lisbon authorities or readers interested on the topic to have an in-depth view of the situation and for that reason a spatial view with access to maps and zones where these road accidents happened will be necessary to perform a good evaluation and a good analysis. With that in mind, there are three questions this work needs to answer at least. Based on the input we will put, what type of accident is most likely to happen(predictive model), where are the main areas where it its more likely to have or witness a car accident and have they been increasing over the years(descriptive preliminary analysis using Power BI).

In conclusion, road accidents in Portugal, particularly in Lisbon, are a significant issue that requires further investigation and intervention. Understanding the factors contributing to these accidents, the areas most affected, and the true extent of injuries and fatalities is crucial in developing effective strategies to improve road safety in the country.

This study is structured as follows. The literature review where we analyze past works, missing gap and how can we fill it. After that, the methodology followed by data description, preprocessing, modeling and lastly the obtained results and findings. We then conclude our work, providing significant contributions and insights of the research, identifying critical questions to be explored by future researchers, and acknowledges encountered limitations.

2. LITERATURE REVIEW

2.1. APPROACH FOR THE LITERATURE REVIEW

The research goal was to find/explore results that were connected with the topic and provided some enlightenment as well as different views for different problems in different places or different times but with an identical scenario.

To start, it was necessary an effective method that would provide help and guidance with the literature review. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology was the most adequate one. There are several PRISMA forms to perform literature reviews split into several phases. Each phase denotes a process and every time each process is performed it will ultimately guide to more clean and suitable solutions or in this case research. After all the steps or phases are complete, we get the final articles for the literature review.

The articles will be split by importance, being the first ones the more suitable and crucial for this work since they are similar in terms of final goals and outputs. For this research, several databases were consulted where it was seen and extracted master thesis, articles, papers and other works from Google Scholar, Web of Science and Repositório Universidade Nova (RUN). The keywords for identification and search in the repositories or databases were performed with the following query: road accident AND Lisbon and also in Portuguese acidentes rodoviários AND Lisboa. Additionally, only related works with no more than 10 years were considered for this purpose since the older the projects done, the more detached they could be from today's reality.

2.2. LITERATURE REVIEW RESULTS

The PRISMA flow diagram was performed to achieve the best sources and research associated with the topic being handled. The first step is the identification phase where all the search in a way related to road accidents on each repository was identified. The result was forty-five records and one additional record identified through other sources such as literature of reference in the field resulting in a total of forty-six records identified.

There were no duplicate records but after a look at the abstract, keywords and resume, twenty-four were excluded as they were off-topic or presented other reasons to be discarded such as outdated research. Also, for comparison reasons, some articles have a maximum of fifteen years in order to compare the results regarding different decades and see, if possible, tendencies that may exist. With all the steps done and full-text screening, there was a total of fourteen records eligible.

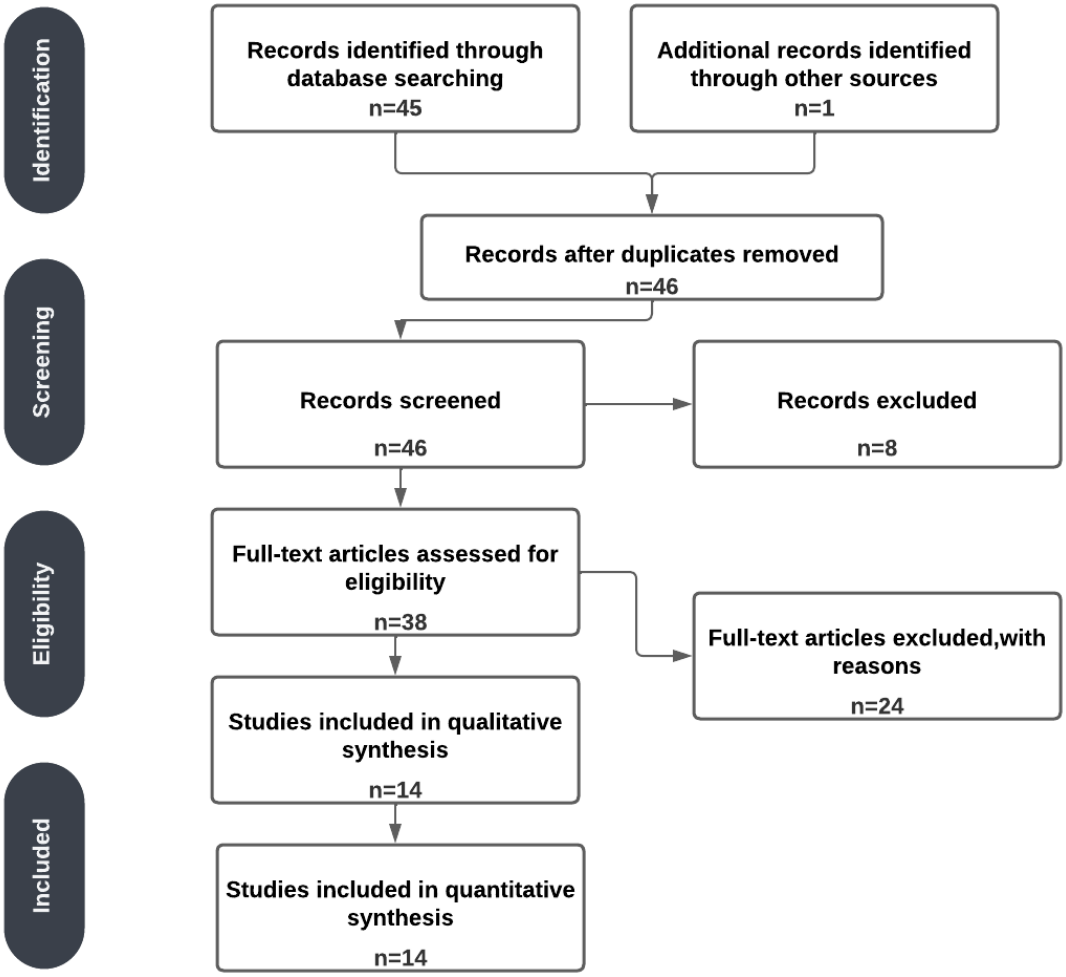


Figure 1-PRISMA flow diagram

2.3. RESEARCH STUDIES CONTEXT

Even though the research found and analyzed had different years and some different objectives they are all central to road accidents and we can extract some conclusions as well as information's that will help later as the thesis is developed. The studies are structured by importance and the closest we have to being ideal will serve as a base of comparison to extract ideas and if possible, compare choices made there with the choices done in this work. One work, conducted by Gomes studied the influence infrastructure characteristics had in urban road accidents, used a geocoded database of road accidents to visually/see a map of the occurrences and after, develop a model that estimated the frequency of the accidents on that area. Due to the small sample size, the results would not clarify the causes except in some situations where the explanatory variable could justify part of the problem like the intersections studied. Aside from that, legged intersections with only one direction proved to be associated with a decreased expected number of accidents when studying intersections (Gomes, 2013).

Subsequently, Mesquitela (2021) on its data driven approach to road accidents in Lisbon identified road accidents prone zones or hotspots and tried to correlate them with their influencing factors. Accordingly, to the same author, there were a few barriers regarding the accidents namely the causes of the accident, weather conditions and the respective time it happened. Once this information was obtained and the necessary steps performed the results indicated that the major contribution factor to traffic accidents was the human factor where there was a disrespect for traffic signs, excessive speed and the weather factor had little impact on these events. Important to stand out that the ArcGIS Pro tool was very important to the development of this work.

Similarly, Gomes et al. (2008) developed a model to predict the accidents on Lisbon urban areas and were faced with several barriers starting with the reports of each accident. Accidents that weren´t registered by the local police lacked crucial information or the vast majority wasn't even reported for quantification to the respective authorities and even when the police filled the forms, more information's were needed to perform a better access of the situation. Going back to the model, the most common explanatory variables used are the traffic volume suggesting that with high density traffic there is a bigger chance of an accident happening.

One other work from Mesquitela et al. (2022) using the Morans global autocorrelation and GIS-assisted technique (ArcGIS Pro) performed several density spatial view maps analysis that enabled them to identify and visualize the main road traffic zones or hotspots where these events were more common and came to some very interesting conclusions identifying twelve very problematic zones with the leading one being segunda circular. They also managed to combine Instituto Portugues do Mar e da Atmosfera (IPMA) data allowing them do conclude that the main factors for road accidents was the human factor and in some roads, the weather factor could aggravate and incitive road accidents to happened in case of rain for example.

There is a pattern forming where several works turn around visual maps pointing out the best way to deal with this subject. Along these lines, Braceiro (2016) trough ArcGis and Georeferenciação mapped the places where a road traffic accident occurred in urban areas and non-urban areas specifying the big differences and impact that urban and non-urban areas could have on accidents with urban areas having more roads and less accidents do to its number and non-urban areas having more accidents per road. We could also see the importance of ArGis on mapping all these events and having the right information/complete information helps to perform a good analysis.

Likewise Ferreira (2011) used similar techniques to study road accidents as we seen on recent researches like the ArcGis but with an additional peace that many of the other works didn't include which was a real map extracted from google earth or even pictures of the zones were these accidents happen more allowing for the reader to see in a more global and real life or in depth view the places.

These are the main works in which comparisons can be done for the future work developed here. Almost all turn around visualizations to better understand road accidents and try to make connections with external factors and only one offers a model for prediction besides the analysis. The remaining studies presented are also relevant but have a different approach to the topic as we will see next.

The first study we analysed, associated the cost of road accidents by Silva et al. (2021) found road accidents had a total cost of 5362.7 million euros, representing 2.53% of the Portuguese gross domestic product in 2019 and the male gender was the most predominant on the fatalities occurred

by a big margin. There is also a lot of high detailed information with all the cost associated with each accident, split between types of accidents, fatalities on each districts, months and others that give a very in depth view and enlightenment of the costs associated with road accidents.

Beginning with Martins (2010), she had research different from the studies previously talked, with an approach more targeted to the people involved on road accidents and not the road accidents in itself was chosen more for the human factor associated in these events and see possible correlations found in this research master thesis. With little data and a very sensitive topic the conclusions extracted in a subject more guided to our research was that many individuals recognized that factors such as speed and alcohol were the main road violations that could lead to a road accident or cause one being these two the most common ones.

Another interesting work, conducted by Ramos et al. (2015) through the identification of black spots, mapped the areas where the accidents occurred using a dynamic approach on each parameter for the variables. The results were similar to Mesquitela work but with more areas outside Lisbon such as Setubal peninsula and Portugal mainland.

The final one worth mentioning, showed Nunes (2011) performed one of the most complete researches regarding road accidents in general in Lisbon with all the side of effects, conditions and proper methodology to cover the hole topic and everything involved. Many statistical formulas and pros and cons tables highlight external factors which could increase or decrease the chances of the road accident to happened.

In all these previous works, even though the authors worked with accidents, the main topic spins around other categories like mental health that won't be treated in this project but might help certain behavior and patters latter down the line when the descriptive analysis is performed.

With the literature review performed to the major researches on the topic we start to see some similarities and aspects that most if not all have in common.

There are a lot of factors to take in consideration when a road accident happens, most of this information will not be available and in case of a more deeper research such as the external factors research, the data collected will only be possible to collect via other sources such as ASNR(Autoridade Segurança Nacional Rodoviária), insurance companys, IPMA(Institudo Português do Mar e da Atmosfera) or Lisbon municipality hall and even with that there is still a chance that most of the information required to answer the questions will not be present.

It highlights the importance prediction models have since they are very scarce with few to none focused on finding if the numbers have been increasing regarding road accidents or even predicting types of accidents or related factors. Several descriptive analysis are done but from external factors to the accident and not to the areas or zones they happen with details about those areas.

3. METHODOLOGY

3.1. CHOOSING AND ADAPTATION

On a work like this, it is important to define a methodology capable of satisfying all the steps and processes required to deliver a good result in the end. Has previously mentioned, the final work will deliver a prediction model capable of predicting car accidents(code for the type of accident) for the Lisbon municipality as well as a preliminary descriptive/analytic view performed around the main areas where these events occur in the city so local authorities can plan and act accordingly with the information and conclusions provided by this report.

In order to do so, the methodology associated with this work will be the CRISP-DM(Cross Industry Standard Process for Data Mining). It was created and is used in Data mining works to support and guide trough a framework the steps necessary to take in order to make projects more reliable, more repeatable, more manageable and faster regardless of the technology used or industry sector where its used (Wirth & Hipp, n.d.).

Its is designed and split in to six continuous and interactive, non-sequential phases and although this is a data mining cycle, it can be used on almost anything related to modeling and adapted accordingly with the topic.

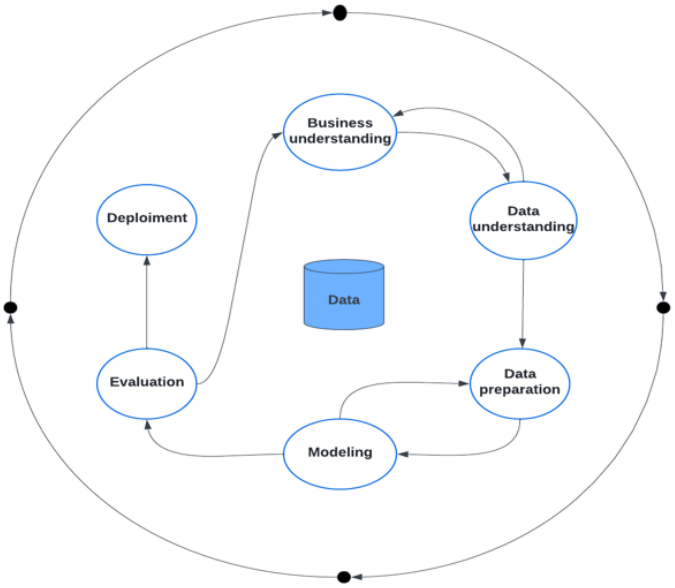


Figure 2-CRISP-DM phases.

Each phase when created was “built” with the intuition to guide the user, step by step till the final goal was achieved. For the prediction model, the CRISP-DM methodology will suit perfectly to deliver the intended output.

3.2. BUSINESS UNDERSTANDING

As previously mentioned on the introduction, road accidents happen on a daily basis and is important to understand how we can improve/understand and predict them in order to avoid the occurrence of this event. For an accident to happen, one vehicle needs to be involved on a crash due to many factors where some of them will be used to try and develop our model like weather factors. This study will be conducted on the city of Lisbon, the capital, where the data provided will undergo several changes and modifications so we can analyze and build a model that can predict what type of accident will happen(predict the code corresponding to the accident first and later see which type of accident that code corresponds to).With the model developed and a preliminary analysis taken, this work will help manage better the resources allocated to road accidents and if possible in less time saving more lives has consequence.

3.3. DATA UNDERSTANDING

Starting with the data understanding, since the NOVA Cidade Lab (NOVA Cidade Lab, 2024) has a partnership with the Lisbon Municipality (Lisboa, 2024) for projects related with smart cities, we were given access to a few sources of data collected in Lisbon regarding this subject and three main datasets were chosen to develop this work being traffic accidents from 2013 to 2020(the fundamental pillar of this project), IPMA dataset and traffic light areas.

Variable	Type	Description	Data example
OID_	Number	Unique identifier for each record	1,2,3,4,5...
OCO_ID	Number	Accident identification number	1,2,3,4,5...
OCO_DT	Long Date	Date and time of the accident	13/01/2013 13:12:00
LNG	Number	Longitude of the accident location	-9,10960988299999
LAT	Number	Latitude of the accident location	38,73503661
OCO_NAT_DESC	String	Description of the accident type	2103-Road accidents with Incarcerated; 2102-Road accident with vehicles
OCO_CODE	Number	Corresponding code to each accident type	2102;2103

Figure 3-Traffic accidents descriptive features

When understanding our data, it is not enough its description, further details need to be brought up to light. Being the main database we will use for the prediction model, several python tools are used to better understand it, starting with “data.info”, similar to the figure above but with datatype description, after “data.describe” where summary statistics show how the data is organized with

mean, rows count, quartile, minimum and maximum value. Another useful method also used was the "data.isnull" with a small formula that provides an output showing all missing values on all the dataset and since no missing values were found we could proceed to the next phase. All shown on Appendix 7.7.

Variable	Type	Description	Data example
datetime	Long Date	Date and time the record was taken	1/1/2020 12:00:00
id_station	Number	Weather Station unique identification	1200535
temp	Number	Temperature at the recording time in Celsius	13.5ºC; 20ºC
hum	Number	Humidity level, given as percentage at the recording	0...48,49...78,79...100
precip	Number	Precipitation at that exact moment in millimeters	0;0,1;0.3...
sun	Number	Sunshine at that exact moment	0...84...3200
wind_speed	Number	Wind speed recorded at that exact moment in meters per second	0...0,8...5...70...

Figure 4-IPMA descriptive features

The IPMA dataset was much cleaner and didn't contain any errors or missing values after the same analysis performed on the road accident dataset were done here implying the IPMA dataset was good and might only need some data engineering features before both datasets merge to simplify all the data preparation process.

3.4. DATA PREPARATION

With the understanding done, we proceed to the data preparation phase so we can build and perform the model without causing unnecessary issues or performance problems down the line.

On a python level, before we proceed to altering and moving data, the train/test split is performed using an 80-20 ratio before any normalization methods giving the models more reliable performances. After that, it is necessary to make sure the data is uniform so it doesn't cause constraints to the prediction or deviate us from the correct answer thus, it was required to perform some feature extractions meaning some variables would suffer "engineering" and transformation to facilitate the process. The first step would be the conversion of the date variable (OCO_DT) to datetime being a general manner of simplifying all the rows in to one type in case there was an error or wrong date along the dataset and the second step is the dropping of rows with error in case they persisted after the first step. Following this logic, some features were extracted from the date variable (OCO_DT) like the Year, Month, Day, Hour and Day of the week.

Lastly the OCO_NAT_DESC(Accident Description) and OCO_CODE(Accident Type) were transformed into a numerical type to have a more coercive dataset with one representing road accidents with incarcerated(code 2103) and zero road accidents with a vehicle(code 2102). All these changes are important steps to make the data smooth and normalized as much as possible to our model with the type of accident being the target variable since we want to predict what type of accident will happen on a determined period and features such as the Year, Month, Day, Latitude, Longitude, Humidity, Precipitation and Wind speed will be used to help predict our target variable. Now that we have the data arranged and ready, we can proceed to the modeling phase.

Target variable	Model input variables
Accident Type	Latitude; Longitude; Day; Hour ; Month; Year; Dayofweek; Humidity; Precipitation; Wind Speed

Figure 5- Target and Input variables

3.5. MODELING

Since the beginning of this project, when the gap and the objectives were established, two questions came up. Where are accidents most likely to happened and can we help local authorities by predicting what type of accident will happen so they can act quicker and better accordingly with the situation. To answer this, after the train-test split performed (80-20), all the necessary feature extractions done like we have seen before and other data operations performed we enter the modeling phase where we need to select the best and more accurate one.

3.5.1. Prediction model

To forecast the type of the next accident, several regression models were put in to test and try to determinate the outcome. Each model will grab the data (train data) use it to practice/train and after make the predictions comparing with the test data. For this, four models were put to the test like the Logistic Regression, Random Forest Classifier, Gradient Boosting Classifier and Support Vector Machine tested on other works with topics similar and related with road accidents.

In order to get the best model, each one has certain hyperparameters inside it that help the model to train and learn. For that reason, good hyperparameters are important to get the best models capable of delivering the best predictions and in that order, the GridSearch Algorithm, widely used in the data community to test and determine the best hyperparameters, will be put to the test in order to find the best ones.

In first model, Logistic Regression, the hyperparameters tested were the “penalty” since we wanted to see if regularization had any effect on the model, “C” to test the regularization, ”solver” an algorithm to use on optimization and finally “max_inter” or maximum number of interactions for the solvers to converge.

Model	Parameters	Testes Values
Logistic Regression	penalty	["l1", "l2", "elasticnet", "none"]
	C	[0.01,0.1,10,100]
	solver	["netwon-cg", "lbfgs", "liblinear", "sag", "saga"]
	max_inter	[100, 200, 500, 1000]

Figure 6-Logistic Regression tested hyperparameters

Regarding the second model, Random Forest, “n_estimators” represent the number of trees on the model, “max_depth” to see how far the model should go since sometimes the deeper it goes, the more information it can get from the data, “min_samples_split” representing the minimum number it takes to split each internal node, “min_samples_leaf” which is similar but for the minimum samples required to be a leaf node and finally “bootstrap”, when false the entire dataset is used for each tree and when true several samples are used to build trees.

Model	Parameters	Tested Values
Random Forest	n_estimators	[100, 200, 300]
	max_depth	[None, 10, 20, 30]
	min_samples_split	[2, 5, 10]
	min_samples_leaf	[1, 2, 4]
	max_features	["auto", "sqrt", "log2"]
	bootstrap	[True, False]

Figure 7- Random Forest tested hyperparameters

For the third model, Gradient Boosting, “n_estimators” represent the number of boosting stages the model should perform with higher numbers normally resulting in better performances,” learning_rate” which minimizes the contribution of each tree depending on the learning rate it is,”max_depth” parameter that limits the number of nodes each tree has, “min_samples_split” like we seen before are the minimum number it takes to split each internal node and “min_samples_leaf” also the minimum number of samples required to be a leaf node. Lastly, “subsample” represents the fraction of samples that should be used for fitting each individual base learners.

Model	Parameters	Tested Values
Gradient Boosting	n_estimators	[100, 200, 250]
	learning_rate	[0.01, 0.1, 0.2]
	max_depth	[3, 5, 7]
	min_samples_split	[2, 5, 10]
	min_samples_leaf	[1, 2, 4]
	subsample	[0.8, 0.9, 1]

Figure 8- Gradient Boosting tested hyperparameters

For the final model, Support Vector Machine, "C" where we test the regularization and with only positive values, "kernel" where a variety of types of kernels can be selected to our algorithm, "gamma" representing the kernel coefficient and can be selected between scale and auto, "degree" used when we have the poly function on the kernel only and "coef0" which is significant only when we previously selected poly and sigmoid on the kernel parameter.

Model	Parameters	Tested Values
Support Vector Machine	C	[0.1, 1, 10, 100]
	kernel	["linear, "poly", "rbf", "sigmoid"]
	gamma	["scale", "auto"]
	degree	[2, 3, 4]
	coef0	[0.0, 0.1, 0.5, 1]

Figure 9- Support Vector Machine tested hyperparameters

After the hyperparameters, the models need to be tested using the combined accuracy and f1-score of each one to determine the best one fit to help us predict what type of accident is going to happen using our target variable (Accident Type).

3.6. MEASURES

F1-Score: It's a widely used measure that combines precision and recall to calculate its score reaching a maximum value of 1 when there is perfect harmony and a minimum value of 0 when there is no harmony.

$$\frac{2 \times Precision \times Recall}{P + R}$$

(Eq.1)

Precision: The number of True Positives results divided by the number of all positive results even if those positive results are falsely identified (False Positives).

$$Precision = \frac{TP}{TP + FP}$$

(Eq.2)

Recall: The number of True Positives results divided by the number of examples that should have been identified has positive results(True Positives and False Negatives).

$$Recall = \frac{TP}{TP + FN}$$

(Eq.3)

Accuracy: Measures the number of correct predicted instances(True Positives plus True Negatives) divided by all the instances on the dataset. This ration is also one of the most used measures to evaluate models performance.

$$\frac{True\ Positives + True\ Negatives}{All\ Samples}$$

(Eq.4)

Confusion Matrix (misclassification evaluation): Widely used tool that helps evaluate where the model is making correct/incorrect predictions by presenting the count of True Positives, False Positives, False Negatives and True Negatives. As an example, when the model predicts well using the train data and matches and then tries to predict the same with the result being equal on the test data, it is called True Positive since the evaluation was correct. On the other hand, False Positive are predictions the model thought was correct but are incorrect.

		Actual Values	
		Positive	**Negative**
Predicted Values	**Positive**	True Positive	False Positive
	Negative	False Negative	True Negative

Figure 10- Confusion Matrix

4. RESULTS AND DISCUSSION

One of the last phases before the conclusions, the results represent all the outputs obtained from the python forecasting models and also the Power BI visualizations so the preliminary descriptive analysis could be performed. As previously mentioned, all steps would lead to this point where we can finaly see if the results are good where in the preliminary descriptive analysis, we will see the evolution and areas where accidents frequently happen and on the prediction model the results from the evaluation and the predictions performed.

4.1. PRELIMINARY ANALYSIS

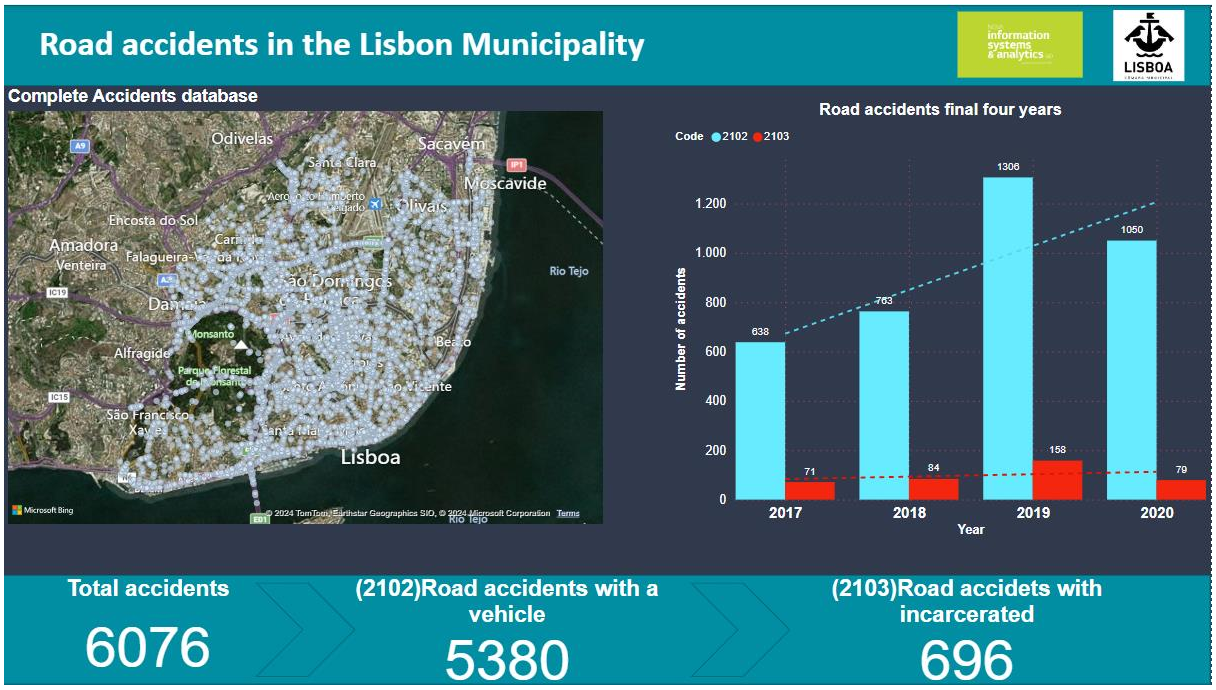


Figure 11- Dashboard all years

For a better understanding, this analysis will focus more between the years 2018 and 2020(3-year spam) since they are the closest to our current scenario. For this, I decided to show a picture of all the combined records on an street view map for viewing purposes and for the reader to have an idea of the data/information provided by accidents file. This image represents all road accidents available that we had access whether they are severe road accidents with incarcerated people or normal road accidents and not surprisingly, all events are spread out on the map since this image compiles all two types of accidents with no discrimination.

There is a total of 6076 accidents that occurred between 2013 and 2020 where 5.380 involved a vehicle and 696 had incarcerated people and through the years the numbers have been increasing following the trend represent on the top right chart (Road accidents final four years).

Now for the results that we are interested in so conclusions can be extracted.

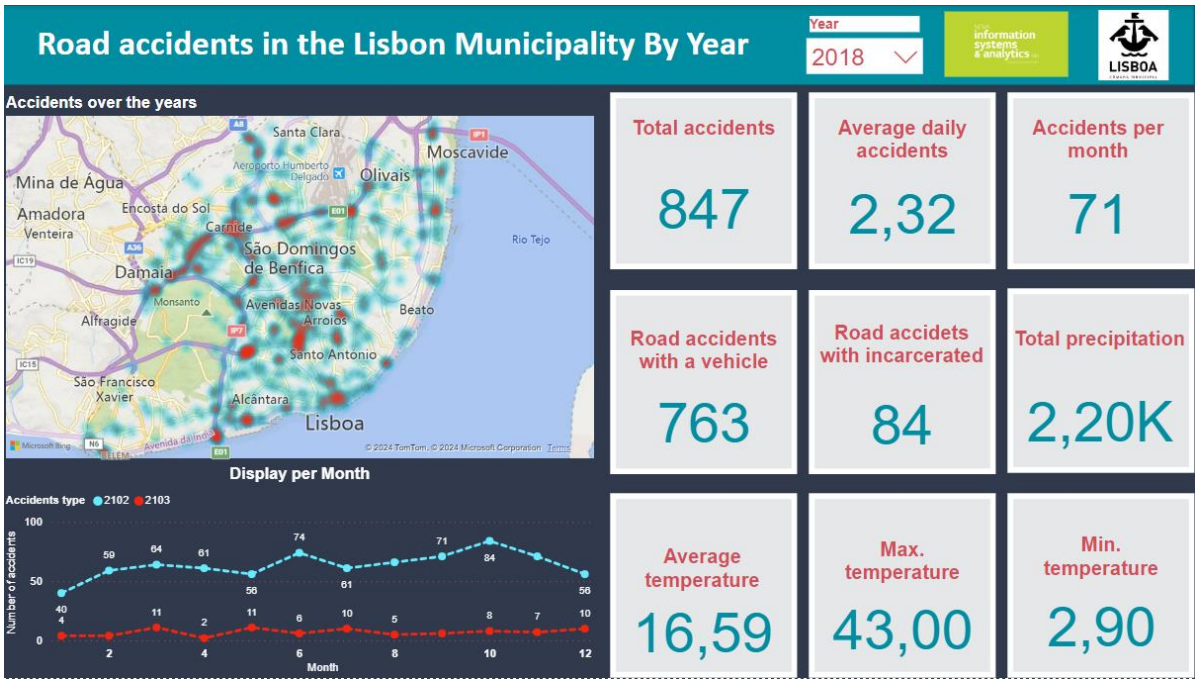


Figure 12- Dashboard 2018

Between the 3 years (2018;2019;2020), this year was the one with fewer accidents registered in the Lisbon municipality. Starting of with this year (2018), a total of 847 accidents were reported or at least registered since we had no indication, but this database seems to represent only severe accidents. Nevertheless, from these 847, 763 were accidents with a vehicle and the remaining 84 had incarcerated people.

We can see on the bottom left chart there is a steady number of incarcerated records not swaying too much but the records with vehicle accidents have large oscillations with October registering the highest number and January the lowest resulting from 2017 low accidents record that is not demonstrated here since it is outside of the time window we are studying but serves as an information year for this year.

A heatmap that concentrates the dots becoming more intense as they are together reveals the areas more affected or where this event occurs the most in the city. In the end, I will talk about the main areas where these hotspots concentrate since between the 3 years the map is very similar and all the main areas are represented in all years.

For 2019,the most pessimistic year, accidents in total almost doubled on both categories and a greater attention will be needed for this year when drawing conclusions.A total of 1464 records were registered compared with the previous 843(an increase of approximately 74%). From these records, 1306 were accidents involving another vehicle and another 158 had incarcerated.

The average number of accidents registered per month went from 71 to 122 and some interesting facts that is important to stand out starting off with the huge gap and dropdown between the months of

May and July making June the lowest month with registered accidents and October beating all records achieving almost 164 accidents with another vehicle.

Accidents with incarcerated follow the same distribution as the previous year but with three months reaching 19 accidents with incarcerated. These are alarming numbers since in the previous year we had a total of 84 incarcerated and it increased to 158(an increase of 88%)..

Regarding the heatmap, since we will discuss the hotspots in the end, the major difference we can see is even more aggravation in the areas where these accidents occur the most as we will see at the end of this chapter. Appendix 7.3 for the full picture of the dashboard.

Lastly the year 2020, more moderate year compared with the previous one and a little more severe than 2018. This particular year has an explanation for all number ups and downs on the Display per Month chart since it was the start of the COVID-19 pandemic in Portugal.

There were a total of 1129 registered accidents but here the situation has its particularity. Accidents involving another vehicle had 1050 records and accidents with incarcerated had a total of 79 records representing the lowest number between the three years. The average number of accidents per month was 94 like we can see on Appendix 7.4.

4.2. MODELING RESULTS

When developing each model, we saw that each one had several hyperparameters that needed tuning to perform well and for that reason, GridSearch was performed in order to get the best parameters for each one. The following figure represents the best hyperparameters found and used for the models.

Model	Hyperparameters from GridSearch
Logistic Regression	C:100 ; max_iter:100 ; penalty:l2 ; solver:netwon-cg
Random Forest	bootstrap:False ; max_depth:None ; max_features:auto ; min_samples_leaf:2 ; min_samples_split:2 ; n_estimators:100
Gradient Boosting	learning_rate:0.01 ; max_depth:7 ; min_samples_leaf:2 ; min_samples_split:5 ; n_estimators:250 ; subsample:1
Support Vector Machine	Infinite Kernel run, parameters will be done and adjusted by hand

Figure 13- Best Hyperparameters

For the type of accident prediction(OCO_CODE), after performing each model we need to evaluate all four models one by one using the accuracy and f1-score as previously mentioned and explained. For accuracy the closer to 100% or 1 the better the model and regarding the f1-score there are several classifications: under 0.5 means the model is not performing well, from 0.5 to 0.8 means the model is ok, 0.8 to 0.9 the model has a good prediction capability and anything above 0.9 is perfect or suits the data very well.

Model	Accuracy	F1-Score	Recall	Precision
Logistic Regression	0,8873	0,94	1	0,89
Random Forest	0,8873	0,94	1	0,89
Gradient Boosting	0,8848	0,94	0,99	0,89
Support Vector Machine	0,8873	0,94	1	0,89

Figure 14- Accuracy and F1-Score for models

From the results we can see that all models performed well, forecasting the type of accident and code variables but the Logistic Regression and Random Forest distinguished from the others slightly better. The difference is so small that any of these models would be good to forecast them and the hyperparameters had close to none impact on the final predictions with hand made adjustments representing of differences under 0.005 decimal numbers or less. With that in mind, the choice to keep going was by using the Random Forest and check for an outcome.

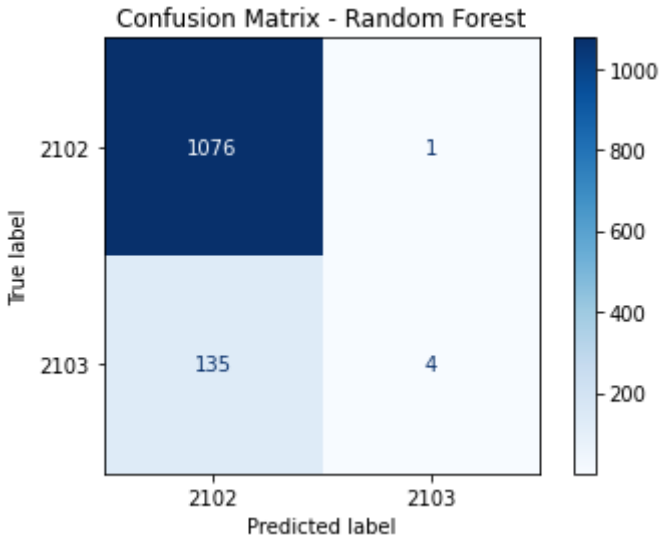


Figure 15- Confusion Matrix

Observing the Confusion Matrix, we can check that the model performance is good but with room for improvement since on making the predictions on our testing data, our Random Forest managed to predict 1076 accidents correctly (True Positives) but miss classified 135 accidents (False Negative) meaning there is room for improvement. One reason that could explain this behavior is the fact that our targer variable is bias, making it harder to know if the model is classifying well.

Besides the confusion matrix, The Receiver Operating Characteristic(ROC) was also performed to have a better evaluation of the model where the output consists in a visual with an line representing our model performance and the closer that line is to the top left corner the better the model performance with results above 0.90 meaning its perfect, between 0.7 and 0.8 good and above 0.6 satisfactory. The ROC visual comparing with the Confusion Matrix tells a different story, with the curve showing us an area of 0.67 which indicates the model is satisfactory but with room for improvement and adjustments to increase to a level of 0.70 or above. Even tough the ROC curve is slightly bellow the desired value, the 0,67 were only possible to obtain after several attempts with removing and introducing variables, feature scaling and also the application of the SMOTE method wich is a technique used on unbalanced datasets. The reason for this stands behind the fact the target variable, besides being biased, has a huge amount of records(close to 90%) belonging to one class and the remaining data representing the other class. Even with these techniques applied, the accuracy remained high but our ROC curve didn't go more higher indicating some miss classification cases were happening with the smallest class.

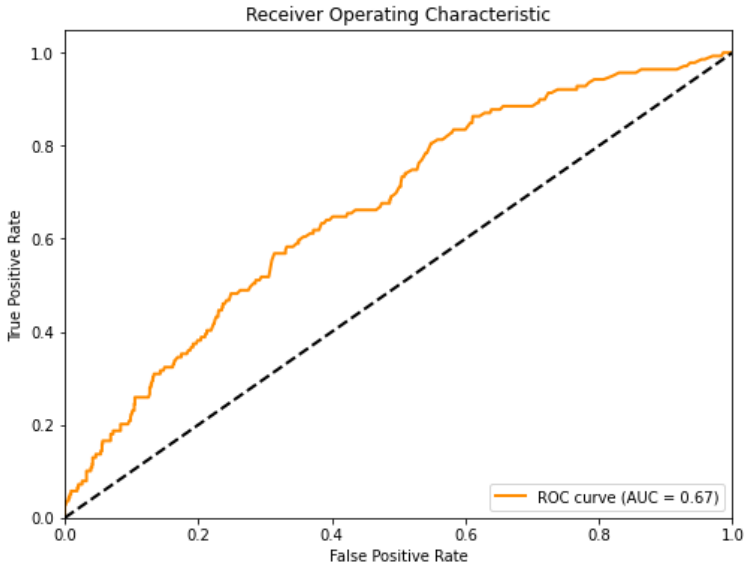


Figure 16- ROC curve

Before we dive into the final predicament whit an example, it is important to see the impact each variable has on the model using the Seaborn Bar Plot where a few conclusions and notes can be already taken. For starter, the Seaborn Bar Plot indicates the variables that have impact on our model with the greater ones on top and the less relevant on the bottom. The sum of the combined score from the variables will be 100.

Regarding the analysis, it comes with no surprise that "latitude" and "longitude" have great weights on our model with depending on the area as we previously saw on the preliminary explanatory analysis, some may be suitable or inclined to have one type of accident and as an example of that we have areas with roundabouts like Marques de Pombal and highways where the first is usually more suitable to accidents involving another vehicle and the other to accidents with incarcerated due to the nature of the infrastructure and speed practiced inside it .

One surprise that came with the variable analysis on our model is the fact that the day has more importance on predicting the type comparing with the hour or day of the week since there is this perception that at night accidents due to their nature or weather/other variables are more severe and

the same can be applied to the day of the week with the perception that at mondays, traffic is more intense and suitable to having more accidents with vehicles due to traffic jam comparing with an Friday. The last surprise was the fact that precipitation had close to zero impact since one would think with heavy precipitation, more accidents would follow but turned out that wind speed and humidity had greater impact on predicting compared with precipitation with almost no influence.

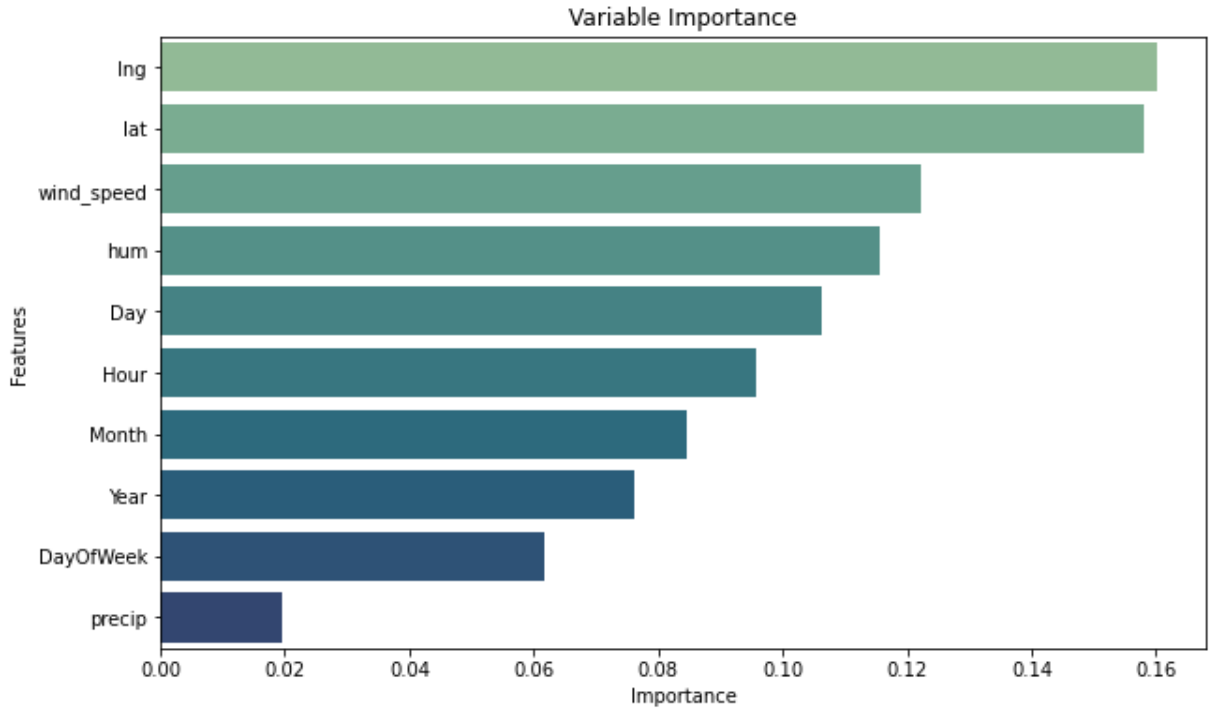


Figure 17- Seaborn bar plot of variable importance

To finish before the model deployment, an error analysis heatmap was performed in order to check if our theory was correct. On the Figure bellow, the X-axis represents the hours of the day going from 0 to 23 with each column corresponding to an hour and the Y-axis represent the day of the week with number one representing sunday, number two Monday all the way to number six which is Saturday and inside each square we had the color intensity with red or 1 meaning the model performed badly on that specific time of the day(combination of hour and day for that specific time) and zero or blue meaning the model performed well or as expected.

When performing an observation on the obtained figure we can take some key insights starting with Monday at five am where the model predictions were incorrect followed by Sunday at twenty-three, Monday at one am and Saturday at seven am where a high error rate was recorded. Aside from these, the great majority of the heatmap was blue (zero or very low error rate) meaning the model performed well but had very specific/sporadically days and hours were those periods were harder for it to classify or they represented some kind of outlier with unusual data that might be the meaning for the high isolated errors.

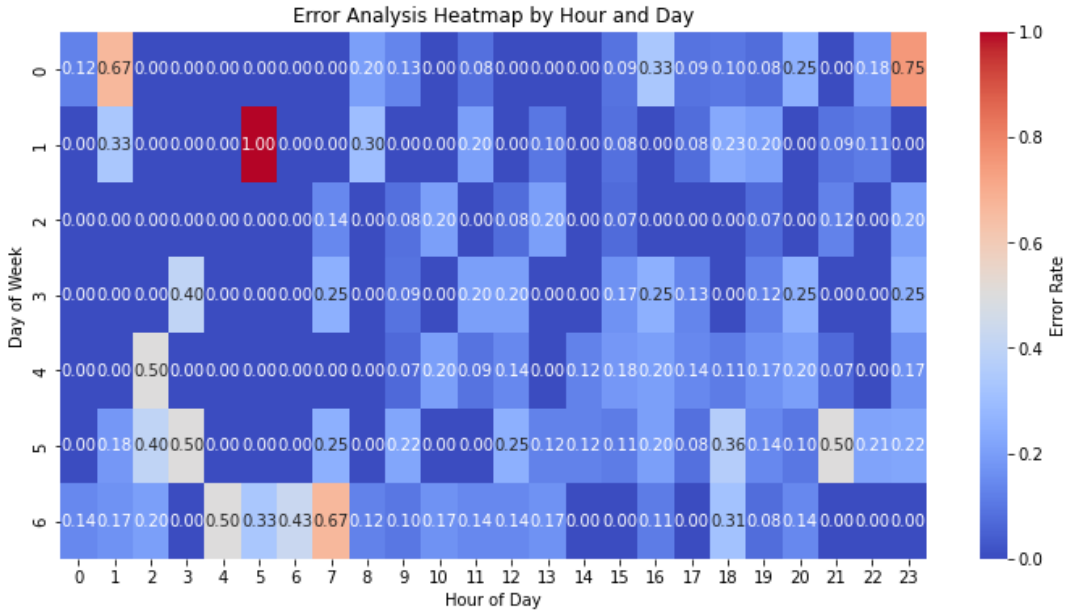


Figure 18- Error analysis heatmap

With the Random Forest being the best model picked and after its evaluation, it was time to return the results based on this model. Several parameters need to be filled with inputs we can choose to predict the best answer and in order to put the necessary inputs to help predict, one function was build that accepted the year, month, day, hour, latitude and longitude of the designated site and time zone we were trying to predict.

We can see on Appendix 7.9, based on the model predictions and according to the figure above, when we want to predict the type of accident that will happen on 26th June 2021 at Marques de Pombal, the most likely type of accident to happen is the 2102 corresponding to an accident with several vehicles. This will help authorities calculate and predict the type of assistance they will have to dispatch to that zone for example. This can be used alongside the descriptive analysis where we insert the major hotspots coordinates in here and see what is more likely to happen.

In the future, the database needs to be updated to the most recent accident and some variables need to be added since the pandemic is over and the trend is getting back to normal making it more reliable to predict and better to forecast next events with the possibility of a few adjustments in the code to forecast variables like the location using latitude and longitude without great efforts.

4.3. Discussion

Over this project, when performing the prediction model, several techniques and data transformation methods were used to better simplify and clean the performance of the prediction model and even though the model performance showed great accuracy, some punctual/sporadically miss classifications were happening on our target variable between both classes. For starter with the target variable (Accident type) being bias and with one class having a weight of almost 90% of the total representations, it influenced the model to accurately predict the first class very good but, the second class with a weight close to 10%,was harder to predict to the model with a few miss classifications

cases. The initial analysis, even though showed high accuracy, had very low ROC curves and more errors on the heatmap error analysis. After noticing this, some steps were taken to attenuate the situation with feature scaling, introduction and removal of variables with a little improvement but not enough to satisfy so techniques designed for datasets with unbalanced classes were used like the SMOTE and ADASYN (Adaptive Synthetic Sampling). Only SMOTE prove to be the most useful one where miss classifications decreased and our ROC curve increase but with room for improvement at the same. Clearly the class unbalance being so sizeable creates this problem that even several techniques designed for these specific situations can't completely attenuate this affect so for future work one suggestion would be introduction of more information inside the target variable with the need to split the same with more classes (previously its split between two). The introduction of more classes inside the dominant class would balance more the data with an example for this would be the creation of accidents with one vehicle, another with several vehicles and the final one with light injured has an example(only possible by applying the changes at the source when data is collected). This split and more data would solve the misclassification cases greatly. Another aspect of great importance would be the introduction of new variables, with more factors where we could see the impact they have on the model and better adjust the ones we need by removing the other that are only just making "noise" on our data.

The prediction model will give valuable insights and predictions so that public forces can be ready to allocate the means for that specific area where the accident will occur given the type provided meaning better resources will be available to help and faster response time to where it is needed. To complement this, has recently pointed out, with more data from other sources and more in-depth variables, a highly detailed descriptive analysis would perfectly alight with our model and help even more local authorities on evaluating the situation. With that said, a few insights were extracted from the preliminary analysis that are important to point out here for future works to develop. Analyzing the heatmap over the three-year spam, we can see that some areas are relevant and present in all years like "Segunda Circular" with IC19, Eixo Norte-Sul, downtown Lisbon,Marquês de Pombal and Avenida da Liberdade. These areas are major hot zones for accidents to happen and deserve special work fully focused on them and by incorporating the model developed here, autorithies might even begin to avoid the accidents from ever happening by taking preventing measures and allocating the necessary means previously to the event from ever happening.

Also from the preliminary analysis, summarizing the data obtained over the years and its ups and downs, we can conclude the following with the year 2018 being the closest to a "normal" year inside the standards. In 2019 we saw an increase in all aspects of accidents that could be connected with the amount of works and construction done in the city of Lisbon since in that year main areas went through a rehabilitation, areas like 2ª circular and downtown Lisbon suffered repairs and new renewal causing constraints and traffic in the city. The year 2020 was a special year, it had data implying that it would be equal to or even surpass 2019 but with the arrival of COVID-19 and the lockdowns prevented, it helped decline the accidents numbers in certain months lowering to some extent this phenomenon.

5. CONCLUSION

Overall, accidents in general with the data we had access to had been increasing with 2018 representing a standard year has it could be witnessed via the data provided and knowledge from previous years. The same cant be said about 2019, being a very tragic year with numbers way above average braking all records and if the situation didn't take a huge turning point, the following year could have been a disaster. When the COVID-19 pandemic hit, by forcing people to stay at home with lockdowns and remote work from homes, a major turning point happened that prevented the numbers from hitting records once again but the months that followed after the lockdown was over, represented also high numbers of accidents that could be explained by several reasons such as people being on their homes for so long that even a few months made them less aware and prepared for the day-to-day traffic. Besides that, some areas need more attention than others representing a risk for drivers due to the amount of traffic, location and even the speed that vehicles can circulate in those spots.

With the help of the prediction model, we already know what type of accident is more likely to happen on specific coordinates at a given time but nevertheless updated data, more variables and an increase on the prediction number should be done to better evaluate its performance for it to become a vital peace in forecasting and consequently saving lives with emergency services reaching the scene quicker. The target variable also needs more classes in order to attenuate its unbalanced with more classes meaning better chances of the correct means arriving at the accident with better preparation and tools.

The same model should be applied to the hot zones or hotspots that we saw on the preliminary descriptive analysis but with few adjustments through a more in-depth search in order to evaluate more factors that made those areas so alluring for crashes with local authorities need to be close by with good access points to arrive as soon as possible to the designated areas.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

This work throughout its development suffered from lots of limitations that were presented right in the beginning when reading other related projects with accidents in general.

When trying to develop a system or events that may affect accidents, in general/the reasons for them happening, a lot of factors need to be taken into account like speed, road condition, weather, traffic, condition of the tyers of the vehicle among so many other factors and the great majority of the available databases don't have a fraction of that. Even the official databases have only parts of those information's. This represents a challenge in finding ways to predict and even prevent these events since part of the problem also starts after the accident has happened with local authorities not pointing out every detail from the accident and the bureaucracy associated with it. On top of that, the data that is registered in the majority of the cases is exclusive to the insurance companies. These limitations are not only local but also at a European level at least.

For future works and to develop more in depth research, a partnership with "Autoridade Nacional Segurança Rodoviária" (ANSR) would be an added plus both from an information perspective but also for a better understanding of the subject and if possible data from the IMT(Instituto da Mobilidade e dos Transportes) since they have registered traffic flow and possibly other factors that Lisbon Municipality and even ANSR don't have.

Usually, the vast majority of studies is about Lisbon and it should be important to decentralize and focus on other cities like Porto, Évora or even on the entire country. With the data provided by these authorities plus the Lisbon municipality, I believe the necessary conditions would be reunited to provide a more in-depth perspective and possibly relate not all but most accidents to factors such as speed or the human factor.

7. REFERENCES

Bandeira, M. L. (2010). *Mortalidade rodoviária em Portugal: Uma abordagem sócio-demográfica*. [Master dissertation, Instituto Universitário de Lisboa]. ISCTE Campus Repository. https://repositorio.iscteiul.pt/bitstream/10071/2091/1/Disserta%c3%a7%c3%a3o_Mestrado_J.Faria.pdf

Braceiro, D. P. (2016). *Acumulação de acidentes rodoviários em Portugal Continental: Contributos dos Sistemas de Informação Geográfica* [Master dissertation, Faculdade de Ciências Sociais e Humanas]. http://hdl.handle.net/10362/18415

Câmara Municipal de Lisboa. (2022). Economia de Lisboa em números 2022. https://www.lisboa.pt/municipio/publicacoes/Economia_de_Lisboa_em_Numeros_2022.pdf

Câmara Municipal de Lisboa. (2024). Missão e composição. https://www.lisboa.pt/municipio/camara-municipal/missao-e-composicao

Datacamp (2023, December). *All-about-power-bi*. https://www.datacamp.com/blog/all-about-power-bi

esri. (2024). What is GIS. *The Science of Where*. https://www.esri.com/en-us/what-is-gis/overview

Ferreira, J. (2011). *Georeferencing Road Accidents with Google Earth: Transforming Information into Knowledge for Decision Support*. https://run.unl.pt/bitstream/10362/5339/1/ejise-volume14-issue1-article697.pdf

Garrido, R., Bastos, A., Almeida, A., & Elvas, J. P. (2014). Prediction of Road Accident Severity Using the Ordered Probit Model. In *Transportation Research Procedia* (Vol. 3, pp. 214–223). https://www.sciencedirect.com/science/article/pii/S2352146514002701

GeoJSON. (2023). *ArcGIS Online*. https://doc.arcgis.com/en/arcgis-online/reference/geojson.htm

GEOJSON. (2023). *GeoJSON*. https://geojson.org/

Gomes, S. V. (2013). The influence of the infrastructure characteristics in urban road accidents occurrence. *Accident Analysis & Prevention*, *60*, 289-297.

Gomes, S. V. (n.d.). The Influence of the Infrastructure Characteristics in Urban road Accidents Occurrence. In *Procedia—Social and Behavioral Sciences* (Vol. 48, pp. 1611–1621). https://www.sciencedirect.com/science/article/pii/S1877042812028728

Gomes, S. V., Carvalheira, C., Cardoso, J., & Santos, L. P. (2008). Accident Prediction Models In Urban Areas: Lisbon Case Study. In *Urban Transport XIV* (Vol. 101, pp. 619–627). C.A. Brebbia, Wessex Institute of Technology, UK. https://www.witpress.com/elibrary/wit-transactions-on-the-built-environment/101/19442

Majumder, A. (2019, March 19). *PRISMA for systematic reviews and meta-analysis*. Project Guru. https://www.projectguru.in/prisma-systematic-reviews-meta-analyses/

Martins, M. J. G. (2010). *Acidentes rodoviários: Culpa e comportamento preventivo*. [Master dissertation, Instituto Universitário de Lisboa]. ISCTE Campus Repository. https://repositorio.iscte-iul.pt/handle/10071/1641

Melanda Oliveira, A. C. (2021). *Road Accident Analysis in Lisbon* [Master dissertation, Instituto Universitário de Lisboa]. ISCTE Campus Repository https://repositorio.iscte-iul.pt/bitstream/10071/24020/1/master_ana_melanda_oliveira.pdf

Mesquitela, J. C. P. S. M. (2021). *A data-driven approach to road accidents in the municipality of Lisbon* [Master dissertation, Instituto Universitário de Lisboa]. ISCTE Campus Repository. http://hdl.handle.net/10071/23232

Mesquitela, J., Elvas, L. B., Ferreira, J. C., & Nunes, L. (2022, February 16). *Data Analytics Process over Road Accidents Data—A Case Study of Lisbon City*. 18. https://doi.org/10.3390/ijgi11020143

Microsoft. (2023). Power BI. *Create a Data-Driven Culture with BI for All*. https://www.microsoft.com/en-us/power-platform/products/power-bi

NOVA IMS LAB. (2024). Obtido de NOVA cidade urban analytics lab: https://www.novaims.unl.pt/en/nova-ims/labs/nova-cidade-urban-analytics-lab/

Nunes, A. R. D. T. (2011). *Modelação espacial de acidentes rodoviários na cidade de Lisboa* [Master dissertation, Faculdade de Ciências e Tecnologia]. UNL RUN. https://run.unl.pt/bitstream/10362/7733/1/Nunes_2011.pdf

PORDATA. (2023, November 13). Motor vehicles in circulation: Total and by type of vehicle. *PORDATA*. https://www.pordata.pt/en/portugal/motor+vehicles+in+circulation+total+and+by+type+of+vehicle-3100

Ramos, L., Silva, L., Santos, M. Y., & Pires, J. M. (2015). *Detection of road accident accumulation zones with a visual analytics approach*. 8. https://run.unl.pt/handle/10362/21164

Relatório de Tráfego na Rede Nacional de Autoestradas. (2022). Instituto da Mobilidade e dos Transportes, I.P. (IMT, I.P.). https://www.imt-ip.pt/sites/IMTT/Portugues/InfraestruturasRodoviarias/RedeRodoviaria/Relatrios/Relat%C3%B3rio%20de%20Tr%C3%A1fego%20-%202%C2%BA%20Trimestre%20de%202022.pdf

Road traffic accidents with victims, injuries and deaths—Mainland Portugal. (2023, June 28). *PORDATA*. https://www.pordata.pt/en/portugal/road+traffic+accidents+with+victims++injuries+and+deaths+++mainland+portugal-326-3585

Silva, C. M. P., Bravo, J. M. V., & Gonçalves, J. M. (2021). *Impacto Económico e Social da Sinistralidade Rodoviária em Portugal* [CEGE - Centro de Estudos de Gestão do ISEG]. https://run.unl.pt/bitstream/10362/134716/1/Impacto_Economico_Social_Sinistralidade_Rodoviaria_Portugal.pdf

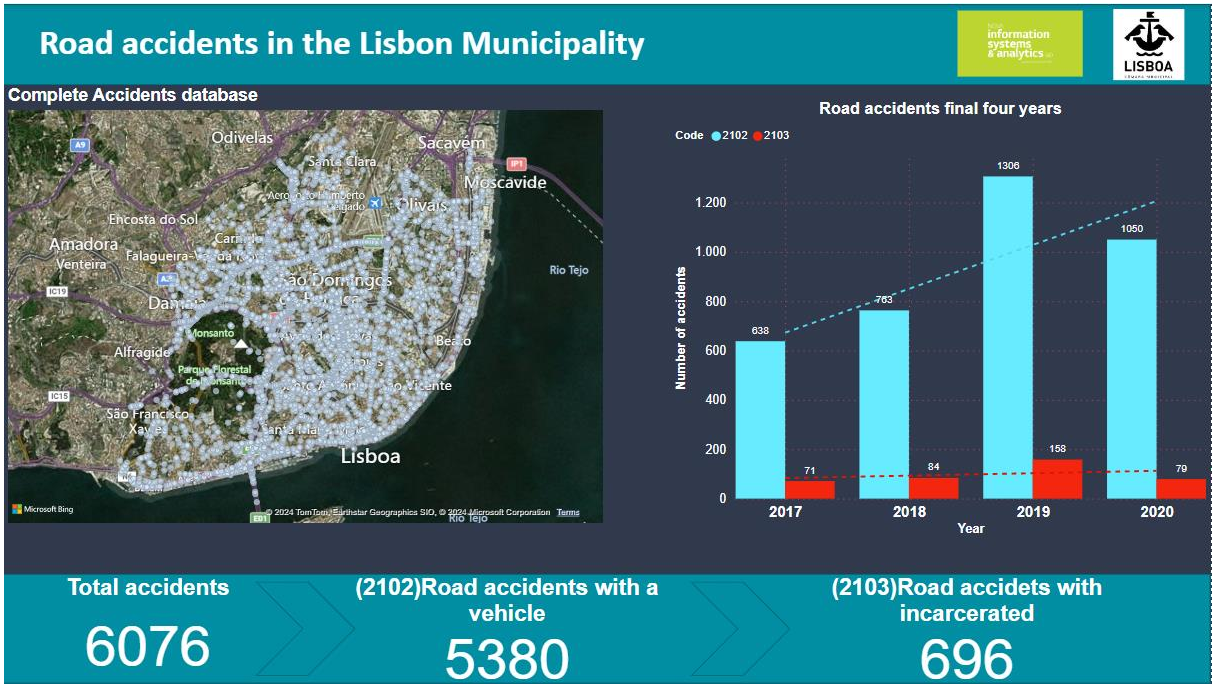
Who should use PRISMA? (n.d.). *Welcome to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)*. http://www.prisma-statement.org/

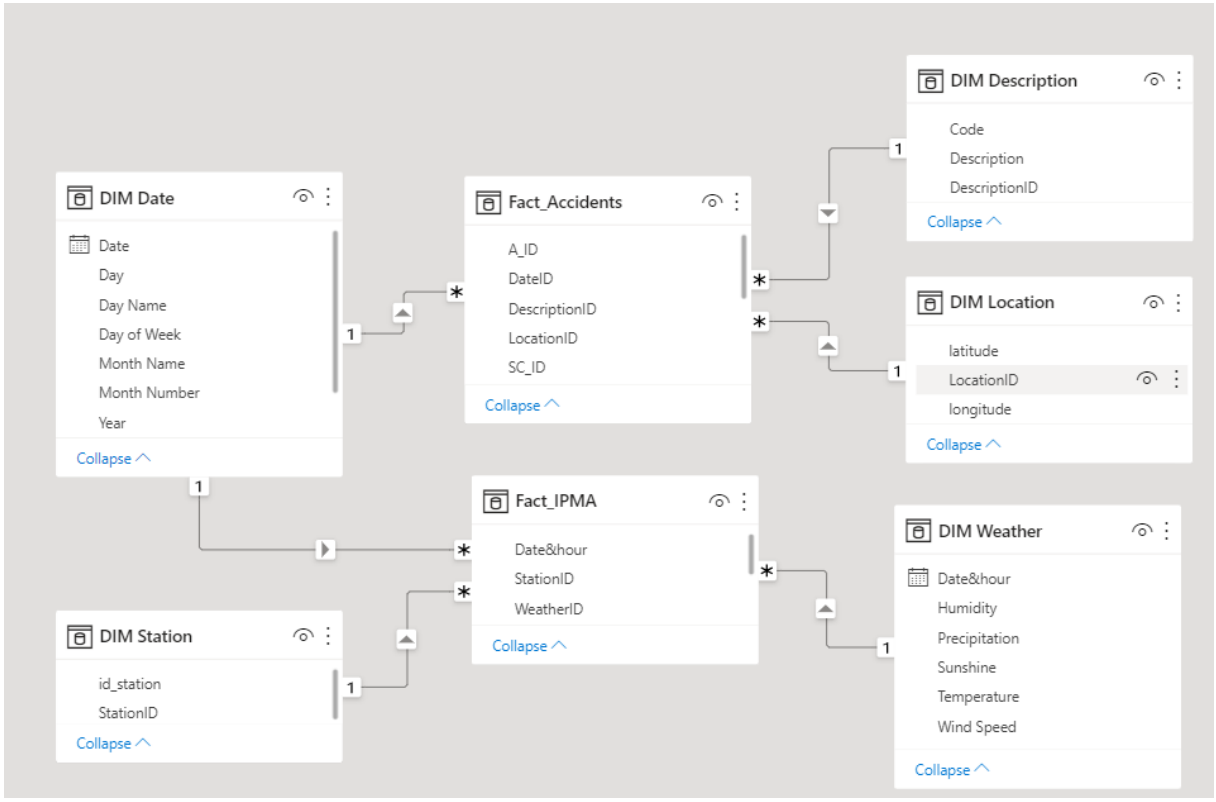
Wirth, R., & Hipp, J. (n.d.). *CRISP-DM: Towards a Standard Process Model for Data Mining*. https://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf

Yannis, G., Papadimitriou, E., Chaziris, A., & Broughton, J. (2014). Modeling road accident injury under-reporting in Europe. *European Transport Research Review*, *6*(4), 425–438. https://doi.org/10.1007/s12544-014-0142-4

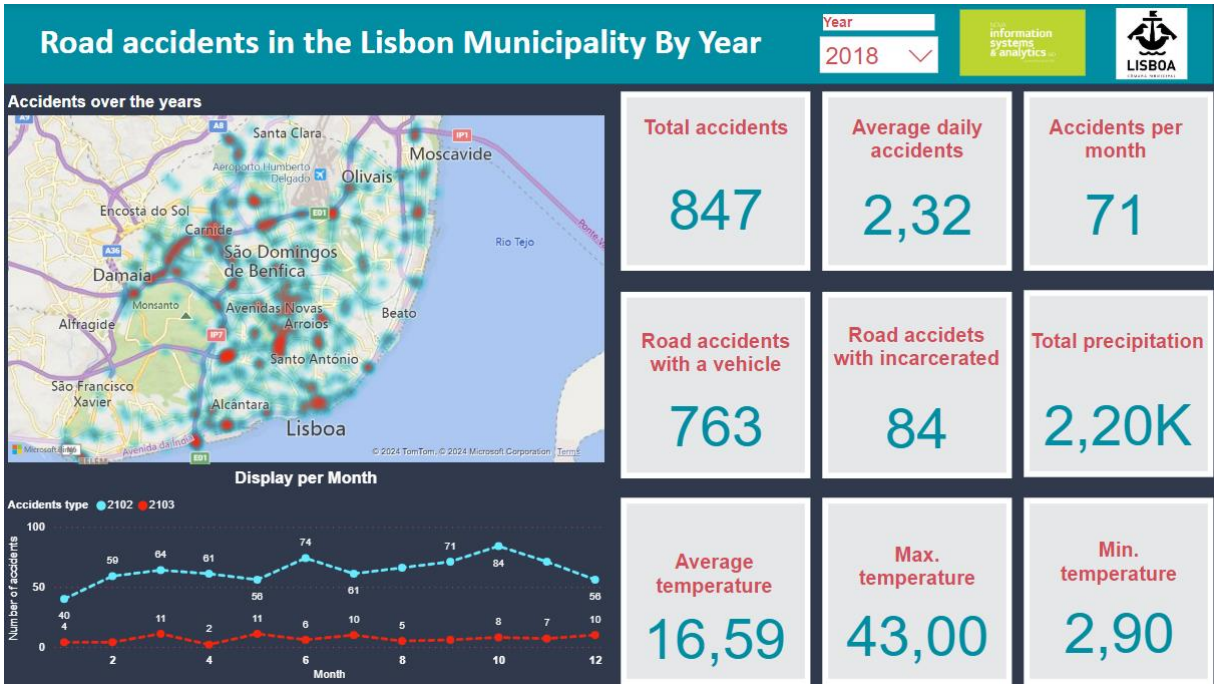
APPENDIX

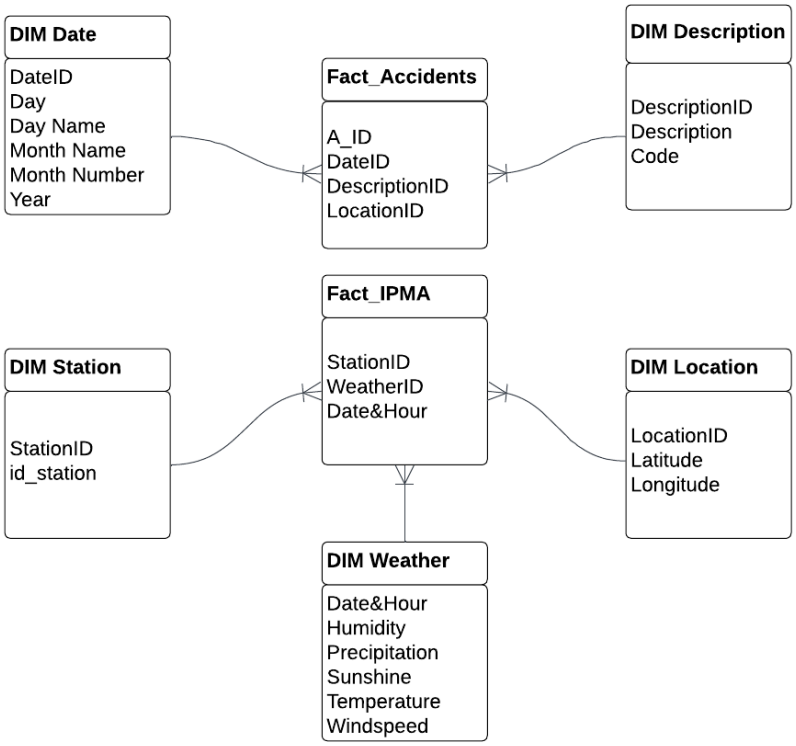
7.1. DASHBOARD-ALL ACCIDENTS & MODEL



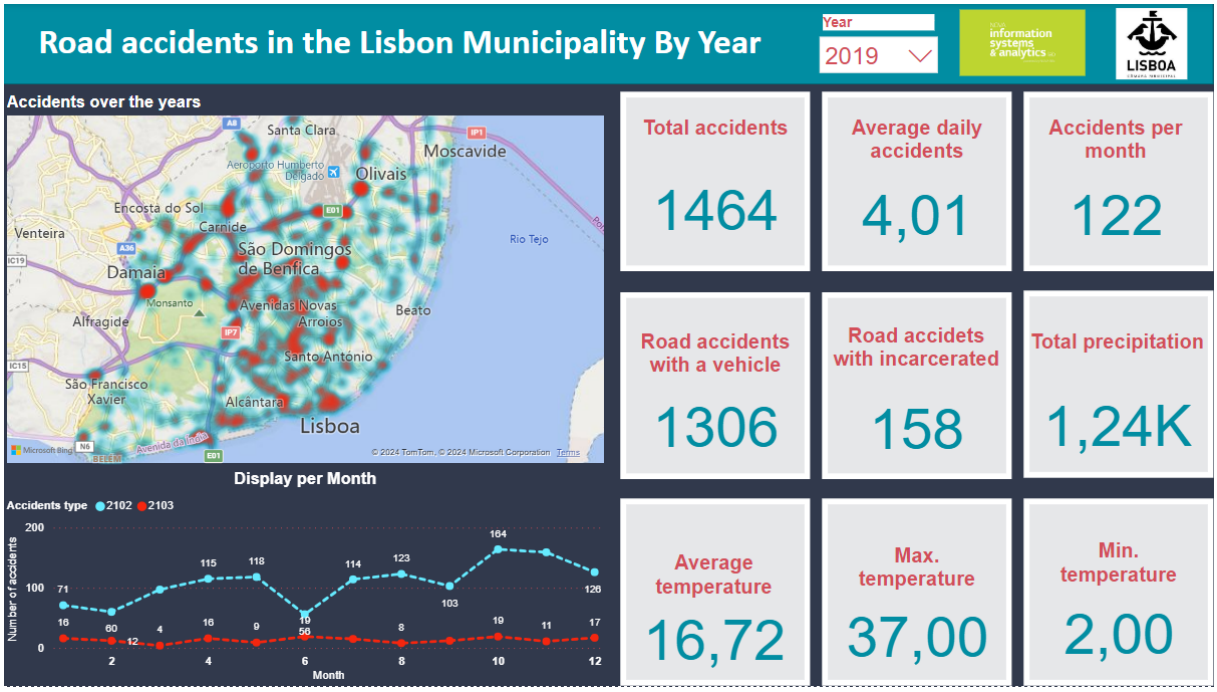


7.2. DASHBOARD-ACCIDENTS IN 2018

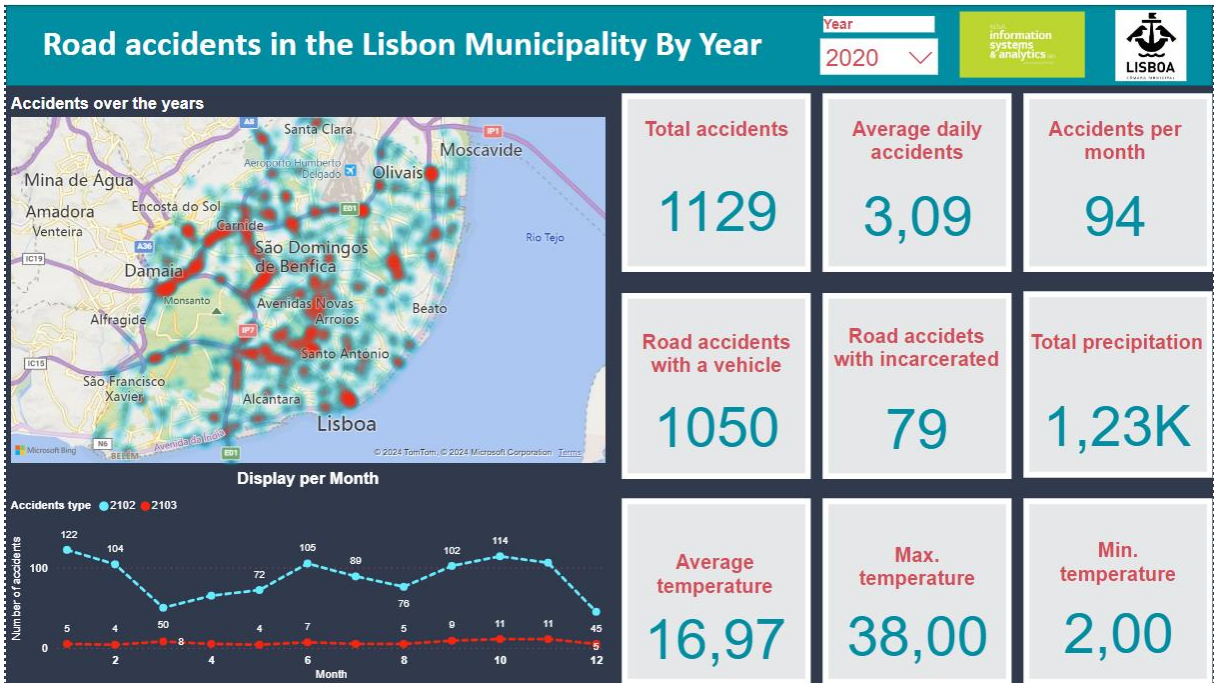




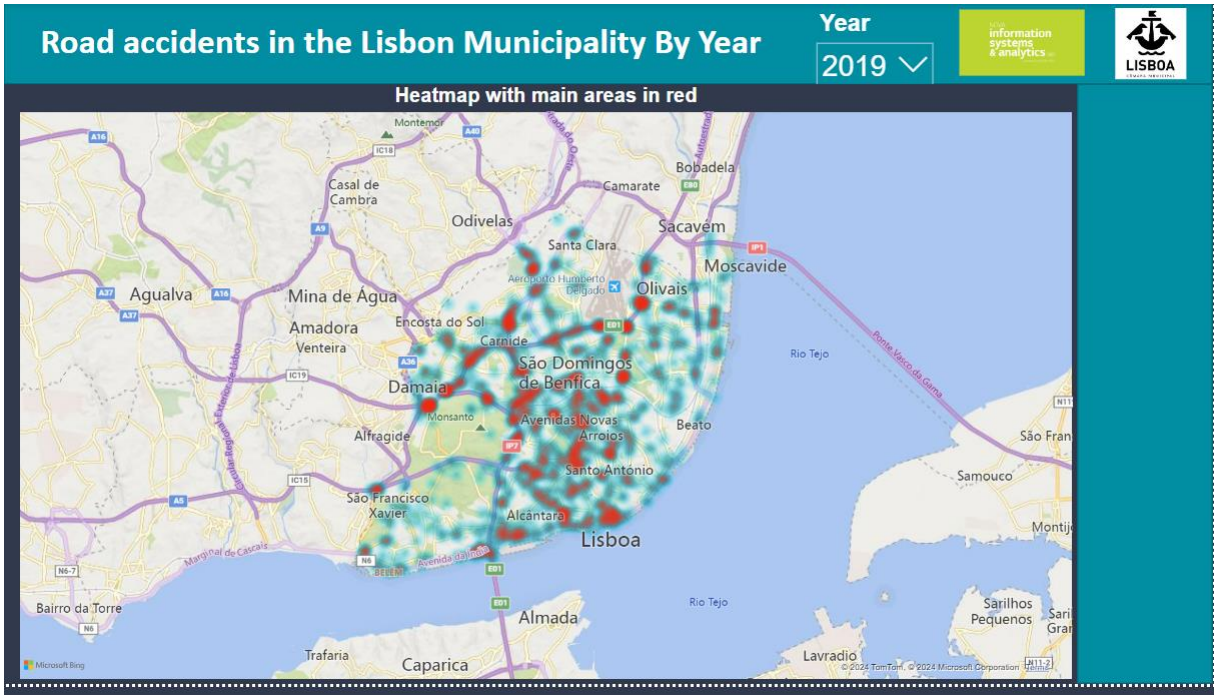
7.3. DASHBOARD-ACCIDENTS IN 2019



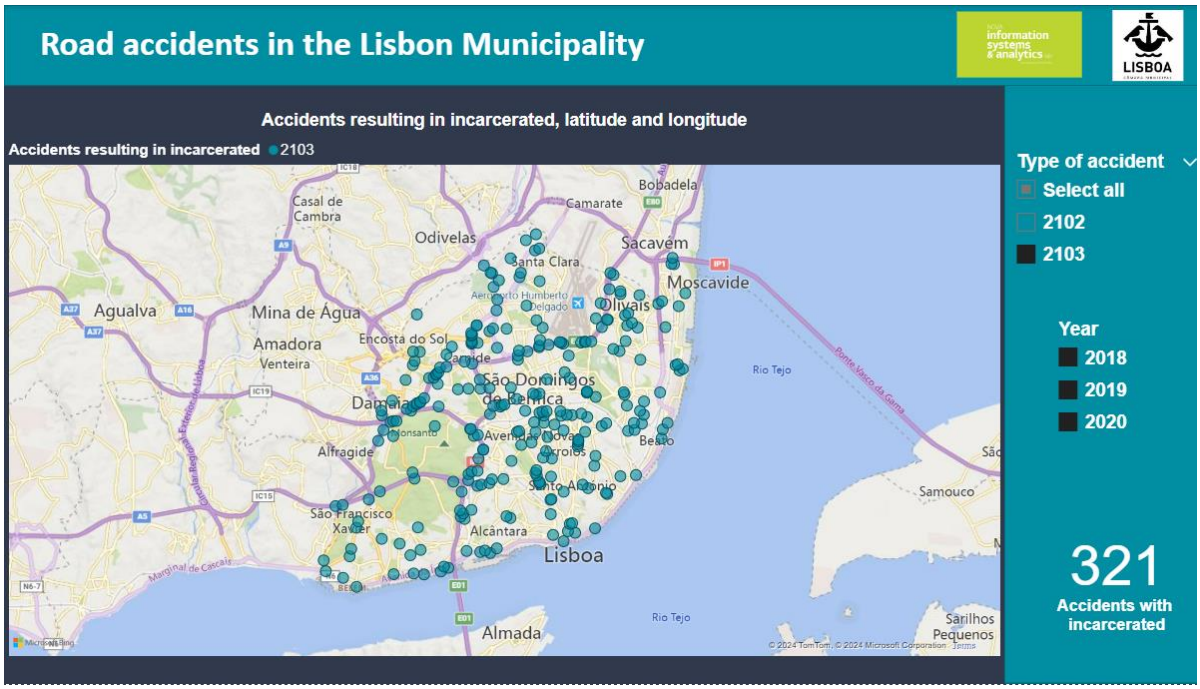
7.4. DASHBOARD- ACCIDENTS IN 2020



7.5. DASHBOARD-HEATMAP OVER THE YEARS



7.6. DASHBOARD-ACCIDENTS WITH INCARCERATED



7.7. PYTHON DESCRIPTIVE MEASURES

In [32]:
```
#Information regarding datatype of each variable of data_accidents
print(data_accidents.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6076 entries, 0 to 6075
Data columns (total 7 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   OID_          6076 non-null   int64
 1   OCO_ID        6076 non-null   int64
 2   OCO_DT        6076 non-null   object
 3   lng           6076 non-null   float64
 4   lat           6076 non-null   float64
 5   OCO_NAT_DESC  6076 non-null   object
 6   OCO_CODE      6076 non-null   int64
dtypes: float64(2), int64(3), object(2)
memory usage: 332.4+ KB
None
```

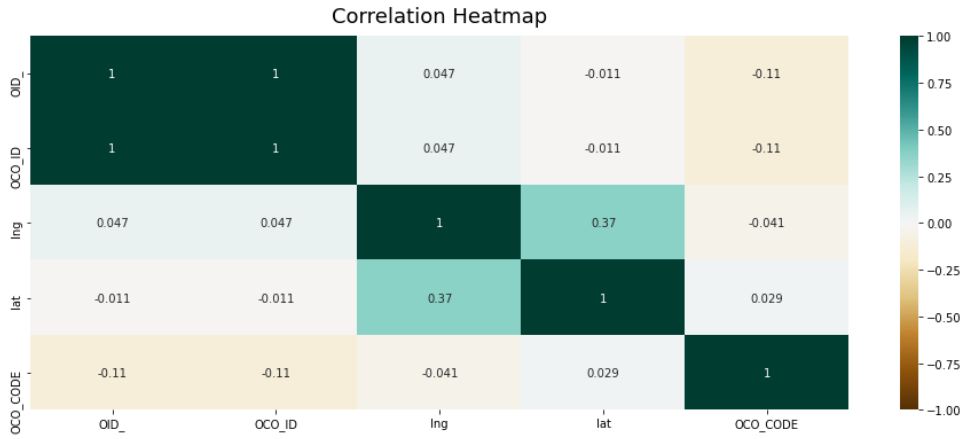
In [33]:
```
# Summary statistics
#Statistics regarding the dataset of road accidents
print(data_accidents.describe())
```

```
              OID_       OCO_ID          lng          lat     OCO_CODE
count  6076.000000  6076.000000  6076.000000  6076.000000  6076.000000
mean   3038.500000  3038.500000    -9.157340    38.739975  2102.114549
std    1754.134449  1754.134449     0.029577     0.022715     0.318503
min       1.000000     1.000000    -9.226871    38.689401  2102.000000
25%    1519.750000  1519.750000    -9.176121    38.722281  2102.000000
50%    3038.500000  3038.500000    -9.157991    38.741318  2102.000000
75%    4557.250000  4557.250000    -9.137489    38.757525  2102.000000
max    6076.000000  6076.000000    -9.091884    38.794842  2103.000000
```

In [34]:
```
# Checking for missing values
#No missing values found
print(data_accidents.isnull().sum())
```

```
OID_           0
OCO_ID         0
OCO_DT         0
lng            0
lat            0
OCO_NAT_DESC   0
OCO_CODE       0
```

In [37]:
```
plt.figure(figsize=(16, 6))
heatmap = sns.heatmap(data_accidents.corr(), vmin=-1, vmax=1, annot=True, cmap='BrBG')
heatmap.set_title('Correlation Heatmap', fontdict={'fontsize':18}, pad=12);
# save heatmap as .png file
# dpi - sets the resolution of the saved image in dots/inches
# bbox_inches - when set to 'tight' - does not allow the labels to be cropped
plt.savefig('heatmap.png', dpi=300, bbox_inches='tight')
```



7.8. FEATURE EXTRACTION

In [9]:
```
# Convert the OCO_DT column to datetime
data_accidents['OCO_DT'] = pd.to_datetime(data_accidents['OCO_DT'], errors='coerce')
```

In [10]:
```
# Handle any parsing errors by dropping rows with invalid dates
data_accidents = data_accidents.dropna(subset=['OCO_DT'])
```

In [11]:
```
# Split the data into training and testing sets (before feature extraction to prevent data leakege)
train_data, test_data = train_test_split(data_accidents, test_size=0.2, random_state=42)
```

In [12]:
```
#Feature extraction
```

In [13]:
```
# Extract additional datetime features for training data
train_data['Year'] = train_data['OCO_DT'].dt.year
train_data['Month'] = train_data['OCO_DT'].dt.month
train_data['Day'] = train_data['OCO_DT'].dt.day
train_data['Hour'] = train_data['OCO_DT'].dt.hour
train_data['DayOfWeek'] = train_data['OCO_DT'].dt.dayofweek
```

```
<ipython-input-13-7fb84efa7d35>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

[14]:
```
# Extract additional datetime features for testing data
test_data['Year'] = test_data['OCO_DT'].dt.year
test_data['Month'] = test_data['OCO_DT'].dt.month
test_data['Day'] = test_data['OCO_DT'].dt.day
test_data['Hour'] = test_data['OCO_DT'].dt.hour
test_data['DayOfWeek'] = test_data['OCO_DT'].dt.dayofweek
```

```
<ipython-input-14-1ee3d2e95766>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
```

[15]:
```
# Encode the target variable (OCO_NAT_DESC) as numerical values
label_encoder = LabelEncoder()
train_data['OCO_NAT_DESC'] = label_encoder.fit_transform(train_data['OCO_NAT_DESC'])
test_data['OCO_NAT_DESC'] = label_encoder.transform(test_data['OCO_NAT_DESC'])
```

```
<ipython-input-15-d915a3c2a697>:3: SettingWithCopyWarning:
```

7.9. TRAIN TEST SPLIT AND REGRESSORS MODELS

```
# Define features and target variables for classification
X_train = train_data[['Year', 'Month', 'Day', 'Hour','DayOfWeek', 'lng', 'lat','hum','precip','wind_speed']]
y_train = train_data['OCO_CODE']
X_test = test_data[['Year', 'Month', 'Day', 'Hour','DayOfWeek', 'lng', 'lat','hum','precip','wind_speed']]
y_test = test_data['OCO_CODE']
```

```
from sklearn.preprocessing import StandardScaler
# Feature scaling
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

```
# Apply SMOTE to balance the training data
from imblearn.over_sampling import SMOTE

smote = SMOTE(random_state=42)
X_train_smote, y_train_smote = smote.fit_resample(X_train_scaled, y_train)

print("Original class distribution:", np.bincount(y_train))
print("Class distribution after SMOTE:", np.bincount(y_train_smote))
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, accuracy_score
```

```
Original class distribution: [   0    0    0 ...    0 4303  557]
Class distribution after SMOTE: [   0    0    0 ...    0 4303 4303]
```

In [19]:
```
#MODEL TRAINING-testing 4 models to see the best one
```

In [20]:
```
#Logistic Regression
```

In [44]:
```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, accuracy_score

# Initialize and train the Logistic Regression model
logistic_model = LogisticRegression(C=100, max_iter=100, penalty='l2', solver='newton-cg')
logistic_model.fit(X_train, y_train)

# Make predictions
y_pred_logistic = logistic_model.predict(X_test)

# Evaluate the model
print("Logistic Regression")
print(classification_report(y_test, y_pred_logistic))
print("Accuracy:", accuracy_score(y_test, y_pred_logistic))
```

```
C:\Users\gonca\anaconda3\lib\site-packages\scipy\optimize\linesearch.py:478: LineSearchWarning: The line search algorithm did n
ot converge
  warn('The line search algorithm did not converge', LineSearchWarning)
C:\Users\gonca\anaconda3\lib\site-packages\scipy\optimize\linesearch.py:327: LineSearchWarning: The line search algorithm did n
ot converge
  warn('The line search algorithm did not converge', LineSearchWarning)
```

```
Logistic Regression
              precision    recall  f1-score   support

           0       0.89      1.00      0.94      1079
           1       0.00      0.00      0.00        94
           2       0.00      0.00      0.00        34
           3       0.00      0.00      0.00         9

    accuracy                           0.89      1216
   macro avg       0.22      0.25      0.24      1216
weighted avg       0.79      0.89      0.83      1216

Accuracy: 0.8873355263157895
```

```
Best Hyperparameters for Logistic Regression: {'C': 100, 'max_iter': 100, 'penalty': 'l2', 'solver': 'newton-cg'}
```

```
C:\Users\gonca\anaconda3\lib\site-packages\sklearn\utils\optimize.py:202: ConvergenceWarning: newton-cg failed to converge. Inc
rease the number of iterations.
  warnings.warn("newton-cg failed to converge. Increase the "
```

```
#randomforest
```

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, accuracy_score

# Initialize and train the Random Forest Classifier model
random_forest_model = RandomForestClassifier(bootstrap=False, max_depth=None, min_samples_leaf=2, min_samples_split=2, n_estimato
random_forest_model.fit(X_train, y_train)

# Make predictions
y_pred_rf = random_forest_model.predict(X_test)

# Evaluate the model
print("Random Forest Classifier")
print(classification_report(y_test, y_pred_rf))
print("Accuracy:", accuracy_score(y_test, y_pred_rf))
```

```
Random Forest Classifier
              precision    recall  f1-score   support

           0       0.89      0.99      0.94      1079
           1       0.41      0.07      0.11       137
```

```
grid_search_rf = GridSearchCV(estimator=random_forest_model, param_grid=param_grid_rf,
                              scoring='f1_weighted', cv=3, n_jobs=-1, verbose=2)
grid_search_rf.fit(X_train, y_train)

print("Best Hyperparameters for Random Forest:", grid_search_rf.best_params_)
```

```
Fitting 3 folds for each of 648 candidates, totalling 1944 fits
Best Hyperparameters for Random Forest: {'bootstrap': False, 'max_depth': None, 'max_features': 'auto', 'min_samples_leaf': 2,
'min_samples_split': 2, 'n_estimators': 100}
```

```
#gradientboosting
```

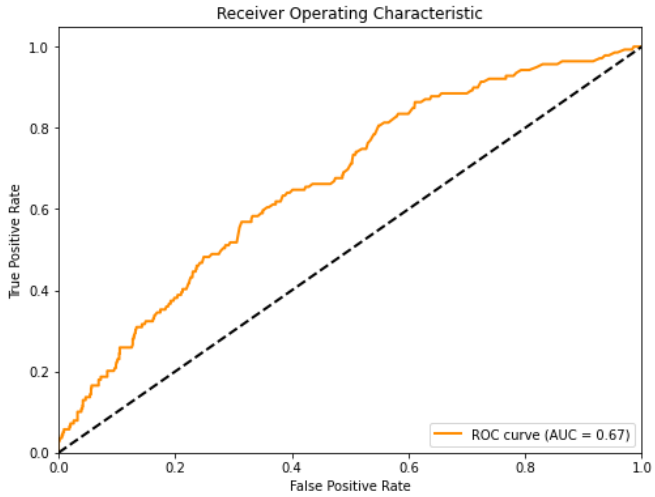
```
from sklearn.ensemble import GradientBoostingClassifier

# Initialize and train the Gradient Boosting Classifier model
gradient_boosting_model = GradientBoostingClassifier(n_estimators=250, learning_rate=0.1, random_state=42, subsample=1.0, max_dep
gradient_boosting_model.fit(X_train, y_train)

# Make predictions
y_pred_gb = gradient_boosting_model.predict(X_test)

# Evaluate the model
print("Gradient Boosting Classifier")
print(classification_report(y_test, y_pred_gb))
print("Accuracy:", accuracy_score(y_test, y_pred_gb))
```

```
Gradient Boosting Classifier
              precision    recall  f1-score   support

           0       0.89      0.98      0.93      1079
           1       0.41      0.13      0.20        94
           2       0.00      0.00      0.00        34
           3       0.00      0.00      0.00         9

    accuracy                           0.88      1216
   macro avg       0.33      0.28      0.28      1216
weighted avg       0.83      0.88      0.84      1216

Accuracy: 0.8791118421052632
```

In [50]:
```
from sklearn.svm import SVC

# Initialize and train the Support Vector Machine model
svm_model = SVC(kernel='linear', C=1.0, random_state=42, gamma='scale')
svm_model.fit(X_train, y_train)

# Make predictions
y_pred_svm = svm_model.predict(X_test)

# Evaluate the model
print("Support Vector Machine (SVM)")
print(classification_report(y_test, y_pred_svm))
print("Accuracy:", accuracy_score(y_test, y_pred_svm))
```

```
Support Vector Machine (SVM)
              precision    recall  f1-score   support

           0       0.89      1.00      0.94      1079
           1       0.00      0.00      0.00        94
           2       0.00      0.00      0.00        34
           3       0.00      0.00      0.00         9

    accuracy                           0.89      1216
   macro avg       0.22      0.25      0.24      1216
weighted avg       0.79      0.89      0.83      1216

Accuracy: 0.8873355263157895
```

```
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic')
plt.legend(loc="lower right")
plt.show()

# Calculate and display the overall AUC for multiclass classification
if not is_binary:
    overall_auc = roc_auc_score(y_test_binarized, y_pred_proba, multi_class='ovr')
    print(f'Overall AUC (One-vs-Rest): {overall_auc:.2f}')
```




```
#next step-choose the best model using the f1-score
```

```
#after that we can start to predict the accident type based on the variables we select
```

```
#Example with real latitude and longitude
import pandas as pd
from datetime import datetime

# Function to prepare a new data point
def predict_new_data_point(year, month, day, hour, lng, lat):
    date = datetime(year, month, day, hour, 0, 0)  # Minute, second, and microsecond are zero since we dont have that
    day_of_week = date.weekday()                   # information on the data received
    return pd.DataFrame({
        'Year': [year],
        'Month': [month],
        'Day': [day],
        'Hour': [hour],
        'DayOfWeek': [day_of_week],
        'lng': [lng],
        'lat': [lat]
    })
```

```
# Example: Predicting the next accident for a specific date and location #Marques de Pombal
new_data_point = predict_new_data_point(2021, 6, 24, 9, -9.1526354, 38.7252702)
```

```
# Make a prediction
predicted_accident_type_code = random_forest_model.predict(new_data_point)[0]
```

```
# Decode the predicted accident type
predicted_accident_type = label_encoder.inverse_transform([predicted_accident_type_code])[0]

print(f"Predicted Accident Type: {predicted_accident_type}")
```

```
Predicted Accident Type: 2102 - Acidentes - Rodoviários - Com viaturas
```

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa