Universidade de Évora - Escola de Ciências e Tecnologia

Mestrado em Engenharia Informática

Dissertação

Plataforma integrada de dados de acidentes de viação para suporte a processos de aprendizagem automática

Daniel Filipe Pé-leve dos Santos

Orientador(es) | José Saias
Paulo Miguel Quaresma
Vitor Beires Nogueira

Évora 2022

Universidade de Évora - Escola de Ciências e Tecnologia

Mestrado em Engenharia Informática

Dissertação

Plataforma integrada de dados de acidentes de viação para suporte a processos de aprendizagem automática

Daniel Filipe Pé-leve dos Santos

Orientador(es) | José Saias

Paulo Miguel Quaresma

Vitor Beires Nogueira

Évora 2022

A dissertação foi objeto de apreciação e discussão pública pelo seguinte júri nomeado pelo Diretor da Escola de Ciências e Tecnologia:

Presidente	|	Lígia Maria Ferreira (Universidade de Évora)
Vogais	|	José Saias (Universidade de Évora) (Orientador) Luís Rato (Universidade de Évora) (Arguente)

Évora 2022

I dedicate this to my family

Acknowledgements

I want to start by thanking my dissertation advisors, Prof. José Saias, Prof. Paulo Quaresma and Prof. Vítor Beires Nogueira, for their patience and availability during the course of the development of this work.

I also want to thank all the MOPREVIS team, for their work and dedication in the project, as well as granting me the opportunity to participate and contribute on this noble cause.

I also want to thank my family for their encouragement and words of wisdom.

Finally, I am truly grateful for the support of every person who supported me in some manner.

Contents

List of Figures

List of Tables

Acronyms

DT *Decision Tree*

RF *Random Forest*

LR *Logistic Regression*

NB *Naive Bayes*

AI *Artificial Intelligence*

MOPREVIS *Modeling and Prediction of Road Accidents in the District of Setúbal*

GNR *National Republican Guard*

FCT *Foundation for Science and Technology*

ANSR *Portuguese National Road Safety Authority*

PSP *Public Security Police*

ML *Machine Learning*

FNN *Feed-Forward Neural Networks*

SVM *Support Vector Machines*

FCM *Fuzzy C-means*

MNL *Multinomial Logit*

NNC *Nearest Neighbour Classification*

PTW *Powered Two-Wheeler*

CART *Classification And Regression Trees*

LSTM *Long-Short Term Memory*

BRF *Balanced Random Forest*

XGB *XG Boos*

kNN *k-Nearest Neighbors*

DOT *Department of Transportation*

AADT *annual average daily traffic*

CNN *Convolutional Neural Network*

NN *Neural Network*

DNN *Deep Neural Network*

HSID *HotSpot IDentification*

EB *Empirical Bayes*

DBSCAN *Density-Based Spatial Clustering of Applications with Noise*

KDE *Kernel Density Estimation*

BEAV *Statistical Bulletin of Traffic Accidents*

IPMA *Portuguese Institute for Sea and Atmosphere*

MDA *Mean Decrease Accuracy*

AUC *Area Under the ROC Curve*

ROC *Receiver Operating Characteristic*

FPR *False Positive Rate*

TPR *True Positive Rate*

TNR *True Negative Rate*

FNR *False Negative Rate*

ETL *Extract Transform Load*

Abstract

Integrated road accident data platform to support machine learning techniques

Traffic accidents are one of the most important concerns of the world, since they result in numerous casualties, injuries, and fatalities each year, as well as significant economic losses. There are many factors that are responsible for causing road accidents. If these factors can be better understood and predicted, it might be possible to take measures to mitigate the damages and its severity. The purpose of this dissertation is to identify these factors using accident data from 2016 to 2019 from the district of Setúbal, Portugal. This work aims at developing models that can select a set of influential factors that may be used to classify the severity of an accident, supporting an analysis on the accident data. In addition, this study also proposes a predictive model for future road accidents based on past data. Various machine learning approaches are used to create these models. Supervised machine learning methods such as decision trees (DT), random forests (RF), logistic regression (LR) and naive bayes (NB) are used, as well as unsupervised machine learning techniques including DBSCAN and hierarchical clustering. Results show that a rule-based model using C5.0 algorithm is capable of accurately detecting the most relevant factors describing a road accident severity. Furthermore, the results of the predictive model suggests the RF model could be a useful tool for forecasting accident hotspots.

Keywords: Machine Learning, Data Analysis, Road Accident Data, Clustering, Decision Trees, Random Forest

Sumário

Plataforma integrada de dados de acidentes de viação para suporte a processos de aprendizagem automática

Os acidentes de trânsito são uma grande preocupação a nível mundial, uma vez que resultam em grandes números de vítimas, feridos e mortes por ano, como também perdas económicas significativas. Existem vários fatores responsáveis por causar acidentes rodoviários. Se pudermos compreender e prever melhor estes fatores, talvez seja possível tomar medidas para mitigar os danos e a sua gravidade. O objetivo desta dissertação é identificar estes fatores utilizando dados de acidentes de 2016 a 2019 do distrito de Setúbal, Portugal. Este trabalho tem como objetivo desenvolver modelos capazes de selecionar um conjunto de fatores influentes e que possam vir a ser utilizados para classificar a gravidade de um acidente, suportando uma análise aos dados de acidentes. Além disso, este estudo também propõe um modelo de previsão para futuros acidentes rodoviários com base em dados do passado. Várias abordagens de aprendizagem automática são usadas para criar esses modelos. Métodos de aprendizagem supervisionada, como árvores de decisão (DT), random forest (RF), regressão logística (LR) e naive bayes (NB), são usados, bem como técnicas de aprendizagem automática não supervisionada, incluindo DBSCAN e clustering hierárquico. Os resultados mostram que um modelo baseado em regras usando o algoritmo C5.0 é capaz de detetar com precisão os fatores mais relevantes que descrevem a gravidade de um acidente de viação. Além disso, os resultados do modelo preditivo sugerem que o modelo RF pode ser uma ferramenta útil para a previsão de acidentes.

Palavras chave: Aprendizagem automática, Análise de dados, Road Accident Data, Clustering, Decision Trees, Random Forest

1

Introduction

In this chapter, we introduce the context of the work presented in this dissertation. Namely, we present topic of road traffic accidents and its importance. We provide the motivation and purpose of the work described as well as a brief description of the structure of this dissertation.

Road Traffic accidents are a global epidemic that is recognized as a major public health concern. According Association for Safe International Road Travel, approximately 1.35 million people lose their lives in road accidents each year, an average of 3700 each day. [7] A traffic accident includes all instances of collision between vehicle and pedestrians, animals, trees, other vehicles or any other type of obstacles on the road. Traffic collisions can result in great costs of propriety, serious injuries or loss of life.

Predicting traffic accidents can potentially improve road safety, decrease damage from road traffic accidents, give drivers alerts to potential dangers, or improve the emergency management system. A reduction in reaction time may be attained if authorities in an area receive advance notice or warning as to which portions of the district's roads are more likely to have an accident at various times of the day.

The work and approach described in this dissertation is based on the extraction of data from various sources and creating an integrated database, using AI methodologies to create new models, integrating

and evaluating different AI approaches (machine learning), assessment of the predictive power of the models and its validation. The ultimate goal is to create an approach that provides real-time assistance to drivers, pedestrians, and authorities.

1.1 Motivation

This work was conducted as part of the MOPREVIS (Modeling and Prediction of Road Accidents in the District of Setúbal) project. MOPREVIS [6] is a project of the University of Évora in partnership with the Territorial Command of the GNR (National Republican Guard) of Setúbal, Portugal, financed by the FCT (Foundation for Science and Technology). The project's primary goal is to reduce serious accidents in the Setúbal district, which, in 2017, despite not having the highest number of accidents, has the highest number of fatalities. The aim is to figure out what factors increase the likelihood of accidents and the severity of those accidents, develop predictive models for both the number and severity of accidents, and test a predictive model to predict the likelihood of accidents on specific road segments. Initially the project is to be carried out only in the district of Setúbal in Portugal, but the idea is to later extend it to the whole country [6].

1.2 Objectives and Methodology

The main objectives of the work described in this dissertation is to present a rule generation model to highlight factors responsible for severe accidents, as well as a predictive model. We would like to mention our contribution as follows:

- For both approaches, we collect and fuse datasets such as weather, time, traffic, and road information.
- The rule generation model supports an analysis on the accident dataset, addressing the responsible factors of severe traffic accidents.
- The predictive model aims at mapping accident hotspots, highlighting areas where, in given circumstances, accidents are likely to happen.

That said, the aim of this work will be a contribution to the MOPREVIS project. To achieve this goal, the following tasks were formulated:

- Creation of a data repository and data processing.
  This task includes the design and implementation of the information system. A large-scale data storage, processing and data management solution will be designed. This task will be changed as necessary throughout the project.
- Initial Experiments.
  After preparing the data, it will be important to develop simple experiments using accident data in order to assess the potential of research approaches, and in order to obtain a better vision of the future of this project, detection of problems and contribution for this work plan.
- Analysis and comparison of machine learning algorithms.
  In a more sophisticated way than the previous task, it will be necessary to create experiments with the chosen methods in order to be able to carry out an evaluation and comparison of the models. This task includes an analysis of the various algorithms using appropriate metrics.

- Apply additional variables to the models.
  Various combinations of variables and subsets of the data will be applied to the model in order to obtain more insights into the occurrence of accidents.
- Result Analysis.
  Finally, With the models defined, it will be necessary to interpret and analyze the results and, if necessary, alter or refine the models.

To achieve these objectives, various machine learning techniques will be used, such as Clustering, random forest and decision trees. Ideally, with the end result some new insights on the data and analysis of the spatial occurrences of the accidents might help in the prevention of future accidents.

1.3 Structure

This dissertation is structured in 5 chapters according to the following format:

In Chapter 2 we present various related works and approaches to road safety studies. We conclude the Chapter with a summary of the insights the state-of-the art has given us.

Chapter 3 is the proposal of the rule generation model that supports an analysis on our accident dataset. We introduce an in-depth description of the data used as well as the techniques used in this approach. We finalise the Chapter with the discussion of the results.

Chapter 4 describes the proposal of the predictive model. We describe the data processing applied, as well as the methods and algorithms implemented. A evaluation and discussion of the trained models is also included.

In Chapter 5, we present a conclusion to the work realized in this dissertation as well as describing the plans of the future work.

2

State of the Art

In this chapter, we present the State of the Art in the relevant fields for the work described, which includes: similar works, data mining techniques and machine learning algorithms applied to road accident data.

Vehicular accident data is usually used to model both the accident frequency as well as the degree of crash severity. Crash frequency models predict the number of accidents that will occur on a defined road segment or intersection over a specific time period [25], whereas crash severity models investigate the relationship between crash severity (injuries, fatal and non-fatal) and contributing factors such as driver behaviour, vehicle characteristics, roadway geometry, and weather conditions.

Similar to the first approach, the data can also be used to predict real-time crash occurrence.

In the following sections, we describe these approaches considering the state of the art, as well as more technical aspects.

2.1 General Traffic Accident Terms

In the topic of road safety studies, it's important to have the general terms well-defined.

In 2016, which marks the start of the data collection for this project, the Portuguese National Road Safety Authority (ANSR) defined road accident as "An occurrence on, or from, a public road where at least one running vehicle, known to supervising entities (GNR and PSP), resulted in victims and/or material damage" [3]. This definition remains unchanged to this day.

Depending on the consequences of the accident, the accident may be defined as an "accident with victims", containing at least one injured person; "fatal accident", resulting in at least one loss of life, "serious injuries accident", containing at least one seriously injured victim, and "light injuries accidents" containing at least one light injured victim.

2.2 Traffic Accident Severity Analysis

Analysis of crash severity outcomes is an established topic in road safety research. Crash data is frequently classified based on the severity of the injuries or the type of impact. For example, a crash could be classified as fatal, severe injury, non-severe injury or no injury. Alternatively, crashes might be classified as rear-end collisions, single-vehicle or multi-vehicle collisions, and so on.

In existing literature, the application of statistical techniques has remained the standard in finding the essential elements leading to crash severity outcomes.

In the existing road safety literature, researchers conducted descriptive statistical and temporal analyses on road crashes depending on a few risk factors, specific road users, or specific types of accidents. Some studies have applied advanced statistical approaches such as random parameter logit model, parameter ordered probit analysis, etc. [12] [14]. A full literature review of these statistical models is beyond the scope of this work. For a thorough literature review of these research, please see Savolainen et al. [32] and a more recent literature review by Slikboer et al. [37].

More recently, machine learning modeling approaches have emerged as a potential modeling tool for traffic accident severity classification and study of the association of road crash variables with respect to severity levels.

Siam et al. [34] used a number of machine learning methods to understand and predict the severity of the accidents in Bangladesh. The data used included traffic accidents from 2015, road information, weather conditions, and accident severity. The authors used the agglomerative hierarchical clustering method to extract homogeneous clusters, then used random forest to select the predictor variables for each cluster. From there, prediction rules were generated from the decision tree (C5.0) models for each cluster. Many rules were generated such as a national/regional/rural roads with no divider having more chances of fatal accidents which the authors claim to be true.

Assi et al.[9] carried out a study to predict accident severity that combined clustering techniques with Machine Learning (ML) algorithms. The authors used historical crash data of Great Britain and employed machine learning models, such as Feed-forward Neural Networks (FNN) and Support Vector Machines (SVM). Combining the ML models with Fuzzy C-Means (FCM) clustering the SVM-FCM model had obtained a good performance in terms of accuracy and F1 score. The clustering algorithm had significantly improved prediction accuracy.

To forecast the severity of traffic accidents, Iranitalab and Khattak [21] provided an extensive comparison of various ML methods. Polynomial logic (MNL), nearest neighbor classification (NNC), SVM and RF analysis are among the studied methods. The results show that NNC has the best overall prediction performance for more severe accidents, followed by RF, SVM, and MNL.

To investigate the likelihood and severity of road accidents, Theofilatos (2017) [40] used Random Forest to rank the most important variables of real-time traffic data of urban arterial roadways in Athens, Greece. The author then used the most important features as the input to logit models to look deeper into the factors that influence accident likelihood and severity. According to the findings, traffic variances had a considerable effect on accident occurrence but had a mixed effect on accident severity.

Other factors, pointed out by researchers that influence the occurrence of accidents or accident severity include: low visibility and unfavorable weather [44, 45]; Traffic flow and speed variations were found to influence Powered Two-Wheeler (PTW) crashes [43]; Theofilatos et al. (2012) [41] compared factors within and outside urban areas. Inside urban areas, factors such as young driver age, bicycles, intersections, and collisions with objects were found to affect accident severity, but outside urban areas, weather, and head-on and side collisions affected accident severity.

It is also worth noting that the majority of road accident data analysis employs data mining techniques, with the goal of identifying factors that influence the severity of an accident. According to Kumar and Toshniwal [23], to analyze the various circumstances of accident occurrences, data mining methods such as clustering algorithms, classification, and association rule mining, as well as defining the various accident-prone geographical locations, are very helpful in evaluating the various relevant factors of road accidents.

2.3 Accident Frequency Models

In literature, traffic safety studies are often implemented on road segments and intersections. The majority of these research focus on crash frequency as their primary subject.

Crash frequency prediction models have been explored by researchers in an attempt to find a relationship between the number of crashes and risk factors [35]. In other words, the goal is often to determine which factors contributes the most to traffic accidents events. These models' dependent variables are the number of crashes per segment or the number of crashes per segment per year.

Again, traditional statistical techniques have been widely used to model crash frequency. Caliendo et al. [11] used the poisson, negative binomial, and negative multinomial regression models to predict the frequency of accidents occurrence on multi-lane highways.

Lord and Mannering [25] presents a review of methodological approaches of researchers and provides a summary of usual employed statistical models such as Poisson regression, negative binomial regression, Poisson-lognormal regression, gama regression, zero-inflated Poisson regression, generalized estimating equation, negative multinomial model, random effects model and random parameters model.

Researchers have also conducted studies to investigate the applicability of machine learning approaches to road safety modeling.

Chang and Chen [13] implemented a CART (Classification And Regression Trees) model to analyse freeway crash frequency. For the test data, the CART model achieved a 52.6% accuracy.

Ren et al. [28] investigated the spatial and temporal patterns of traffic accident frequency and employed a Long-Short Term Memory (LSTM) model to predict the risk of citywide traffic accidents.

Zeng et al. [48] used a neural network model to investigate the association between crash frequency by severity and risk factors. The neural network was compared to the Poisson-lognormal multivariate model which showed that when neural networks are trained and optimized, they have a better fit as well as greater prediction performance. They also extracted two rule-sets from the neural networks to show the precise effect of each significant explanatory variable on crash frequency by severity under different conditions. The

authors claim that neural network models have great potential for modeling crash frequency by severity, and should be considered a good alternative for road safety analysis. According to the authors, neural network models have a high potential for modeling crash frequency by severity and should be considered a viable option for road safety studies.

According to Silva et al. [35], nearest neighbor classification, decision trees, evolutionary algorithms, support vector machines, and artificial neural networks are the usual techniques utilized for these purposes. Because of its capacity to deal with both regression and classification problems, as well as multivariate response models, the latter is employed in a variety of ways.

2.4 Prediction of Real-time Accident Occurrences

Prediction of accident occurrences and risk is typically a binary classification problem. Given the particular nature of traffic crash occurrences, where positive events (crashes) are rare and, in most cases, no data set for negative events (non-crashes) exists, it is necessary to create this data based on the absence of a crash event. Following that, predictive models can be applied to the data.

A Concordia University team [20] used a balanced random forest algorithm to study the accidents that occurred in Montreal. Accident data was obtained from three open datasets: Montreal Vehicle Collisions (2012 to 2018), the Historical Climate Dataset for meteorological information, and the National Road Network database, which contained information on roadway segments. BRF (Balanced Random Forest), RF (Random Forest), XGB (XG Boost), and a baseline model were among the models studied. A total of two billion negative samples were created, with the researchers choosing to use only 0.1% of them. Predictions were made for every hour in each segment. Overall, the algorithms predicted 85 percent of Montreal incidents, with a false positive rate (FPR) of 13%.

Lin et al. [24] investigated various machine learning algorithms, such as random forest, k-nearest neighbor, and bayesian network, to predict road accidents.The best model could predict 61% of accidents while having a false alarm rate of 38%.

Theofilatos et al. [42], compared several machine learning and deep learning techniques, including kNN, naive bayes, classification tree, random forest, SVM, shallow neural network, and deep neural network, finding that the deep learning approach produced the best results, while other, less complex methods, such as Naive Bayes, performed only slightly worse.

Gutierrez-Osorio and Pedraza [18] reviewed recent literature in the prediction of road accidents. The authors found that neural networks and deep learning methods have showed high accuracy and precision while integrating a wide range of data sources.

In [46], the authors used a ConvLSTM configuration that was applied to a research about vehicular accidents in Iowa, between 2006 and 2013. Data included crash reports from Iowa Department of Transportation (DOT), rainfall data, Roadway Weather Information System (RWIS) reports, and other data from the Iowa DOT such as speed limits, AADT (annual average daily traffic), and traffic camera counts. Reports from 2006 to 2012 were used for training, with 2013 being held for testing. The tests involved predicting locations for the next seven days based on data from the prior seven days. In terms of prediction accuracy, ConvLSTM outperformed all baselines. In addition, the system properly predicted accidents resulting from the case study of 8 December 2013, when a significant snowstorm occurred.

There have also been recent research that used neural networks to analyze visual imagery and predict traffic accidents.

Najjar et al. [27] used historical accident data and satellite pictures to train a CNN (Convolutional Neural Network) to estimate the likelihood of accidents at an intersection, achieving an accuracy of 73%.

In another work, Shah et al. [33] used CNN models to examine and investigate accidents using data from closed-circuit television traffic cameras.

For the prediction of accident occurrence, Table 2.1 describes the most relevant works addressed in this section. In addition to the authors who conducted the studies in the column "Reference", it is also mentioned the algorithms used, the Imbalance Ratio of the majority and minority classes and the performance of the best model of each study.

In order to evaluate the model's performance, authors use various performance metrics. Common used metrics are accuracy, precision, sensitivity, specificity and False Positive Rate (FPR). However, Roshandel et al. [29] have found that not many studies use all of these metrics to comprehensively evaluate their models. The author claims that using a wide range of metrics is important to validate any prediction model.

Reference	Algorithms used	Imbalance Ratio	Performance (best model)
Hébert et al.(2019) [20]	BRF, RF, XGB	42:1	85% Acc., 13% FPR
Lin and Wang (2017) [24]	RF, kNN, Bayes net	3.3:1	61% Acc., 38% FPR
Theofilatos et al.(2019) [42]	kNN, NB, DT, RF, SVM, LR , NN, DNN	2:1	68.95% Acc., 52% TPR, 77%TNR

Table 2.1: Description of Systematized Papers (Accident Occurrence)

The binary classification problems in the mentioned in Table 2.1 are imbalanced. Class imbalance refers to datasets with significantly more examples of one class than others. In other words, the class distribution is biased or skewed. In this cases, classes are "accident" and "non-accident", where "accident" is the minority class and the class of greater importance.

It is important to find a balance when creating negative samples. To handle this problem, there are procedures involving the re-sampling of the dataset to balance it, either by oversampling the minority class, undersampling the minority class, or a combination of both.

This problem is also common on accident severity studies where fatal crashes are more rare than no injuries. Fiorentini and Losa [17] mention various examples of literature where a number of papers utilize a highly imbalanced dataset and provide a solution on how to handle imbalanced accident datasets. The authors employ the undersampling technique to improve reliability of the classifiers.

2.5 Accident Hotspots Definitions and Identification

Another important aspect in road safety studies is the concept of hotspots (also called blackspots by other authors).

There is no commonly agreed definition of a 'hotspot' in the road accident literature [8]. Elvik, R. [16] conducted a survey on various European countries to describe the various hotspots definitions. The author

included Portugal in the survey, where one of the definitions used was road segments with a maximum length of 200 meters, with 5 or more accidents and a severity index greater than 20, in one year. Severity index weights fatal accidents with a greater value than accidents with only slight injuries. The detection is performed with the sliding window technique that moves along the road, which is a very commonly used method for these tasks. This and other definitions are usually applied depending on the road characteristics, typically highways, and the techniques are not optimal to deal with multiple road types or road junctions.

Table 2.2 includes a summary of various hotspot definitions across eight European countries and shows that there are no agreed methodology for determining hotspots.

Country	**Hotspot Definition**
Austria	3 or more similar injury accidents within 3 years and a calculated risk coefficient of at least 0.8 (based on the annual average daily traffic and the number of injury accidents) The sliding window method is used with a length of 250 m is applied.
Denmark	Based on statistical test. 4 or more accidents within 5 years identified by means of the sliding window approach with variable length
Belgium	3 or more accidents have occurred whithin 3 years Sliding window of 100m in length
Germany	5 or more accidents within 100 meters over the period of 1 year, or, 5 or more serious accidents within 3 year period. "Similar" sliding window method is used.
Hungary	Outside built-up areas: 4 or more accidents during 3 years on 1km road segments Inside built-up areas: 4 or more accidents during 3 years on 100m road segments Sliding window approach is used.
Norway	hazardous spots: at least 4 injury accidents in the last 5 years over 100m hazardous sections: at least 10 injury accidents during 5 years over 1km Sliding window is used.
Portugal	5 or more accidents in one year over 200m and a severity index greater than 20 (Equation 2.1) ANSR definition [3]
Switzerland	Various definitions depending on the type of section and intersections. Sections are fixed (no sliding window method is used)

Table 2.2: Summary of hotspots definitions across various European countries, adapted from [16] [3].

The severity index used by ANSR [3] is calculated by the following weighted sum:

$$SeverityIndex = 100 * D + 10 * SI + 3 * LI \tag{2.1}$$

Where D is the number of deaths, SI is severe injuries and LI light injuries.

Usually, the first step of road safety approaches, is the identification of accident hotspots. Errors in hotspot identification may lead to worse end results.

Montella [26] has compared various common HotSpot IDentification (HSID) methods. One of the methods is the empirical Bayes method (EB) which was proven to be the most consistent and reliable method, performing better than the other HSID methods.

The paper by Szénási and Csiba [38] present an alternative to the traditional HSID methods by applying a clustering method (DBSCAN) in order to search accident hotspots using the accident's GPS coordinates. The DBSCAN algorithm allows the identification of hot spots(or clusters) with shorter lengths and high density of accidents. The algorithm will also eliminate low density areas.

Thakali et al. [39] used Kernel Density Estimation (KDE), a popular spatial analysis technique, and kriging, an interpolation method, to identify accident hotspots applied to historical data in the Hennepin County of Minnesota, U.S.. Kriging method outperformed KDE and both methods resulted in somewhat different hotspots, highlighting the importance of selecting the appropriate method for hotspot identification.

2.6 Summary

The works previously described have provided valuable insights to support our proposed work, which are summarized below:

- Several authors combine various data sources such as weather information, road information and condition as well as the accident information as the main sources of data. Some works use limited features and small-scale traffic accident data.
- The work by Siam et al.[34] analyzed accident data by finding patterns for the severity of the accidents, thus resulting in a better understanding of the data.
- Most works fall under classification of the severity of accidents or the number of crashes per segment [35]. In the latter case, the study area typically refers to a specific highway which severely reduces the number accidents included in the dataset. By covering all of the district's hotspots more accidents are included, leading to a more general approach. The work in [38] is compatible with this approach.
- Decision trees, random forests, k-nearest neighbor, naive bayes and neural networks are some of the most common algorithms used in accident prediction. In some cases, more complex methods such as deep learning perform similarly to simpler probabilistic classifiers such as naive bayes.
- Using a wide set of evaluation metrics can be beneficial to present and compare performance of classification algorithms.

In summary of this chapter, we have presented two of the main approaches to road safety; Severity analysis and crash frequency.

We start by defining the basic terms in road safety studies relevant to the following approaches. The first approach can provide valuable information on the features that influence the injury severity of accidents or the type of accident. Statistical techniques are often applied but recently machine learning models have emerged as a potential modeling tool for crash severity. RF, DT, NN, SVM are among the techniques used in the mentioned papers. Data mining techniques are an important aspect of these studies.

Accident Frequency Prediction modelling is also an popular approach in road safety studies. Besides traditional statistical models, NNC, DT, SVM, NN are some of the techniques used to tackle these studies.

Then, we discussed some studies on real-time accident occurrences that are very related to this dissertation. We have pointed out a few key aspects of these types of studies such as class imbalance, algorithms commonly used and their performance. RF, kNN, DT, NN, SVM are among the utilized ML algorithms.

Finally we discussed Accident hotspot definitions and HSID methods, which are key aspects of road safety studies, mentioning very distinct approaches.

3

Rule-Based Model

In this chapter, we propose a rule-based model approach to support an analysis on our accident dataset. We introduce an in-depth description of the data used as well as the techniques used in this approach. We finalise the chapter with the discussion of the results.

3.1 Overview

The proposed approach aims at finding the most influential factors for accident severity and representing those factors in rule sets. Figure 3.1 illustrates the four stages of the proposed work, namely, Data-Processing, Clustering, Feature Selection and rule generation model.

Data processing is a common step in any study of this kind. For this stage, the data is cleaned and prepared so that it can be properly applied without problems to a machine learning model. This treatment consists of handling null values, encoding values, assigning types to variables, among other changes that are necessary for the correct reading of data by the algorithms.

Clustering or grouping of data is the creation of groups of data defined by their degree of similarity. The

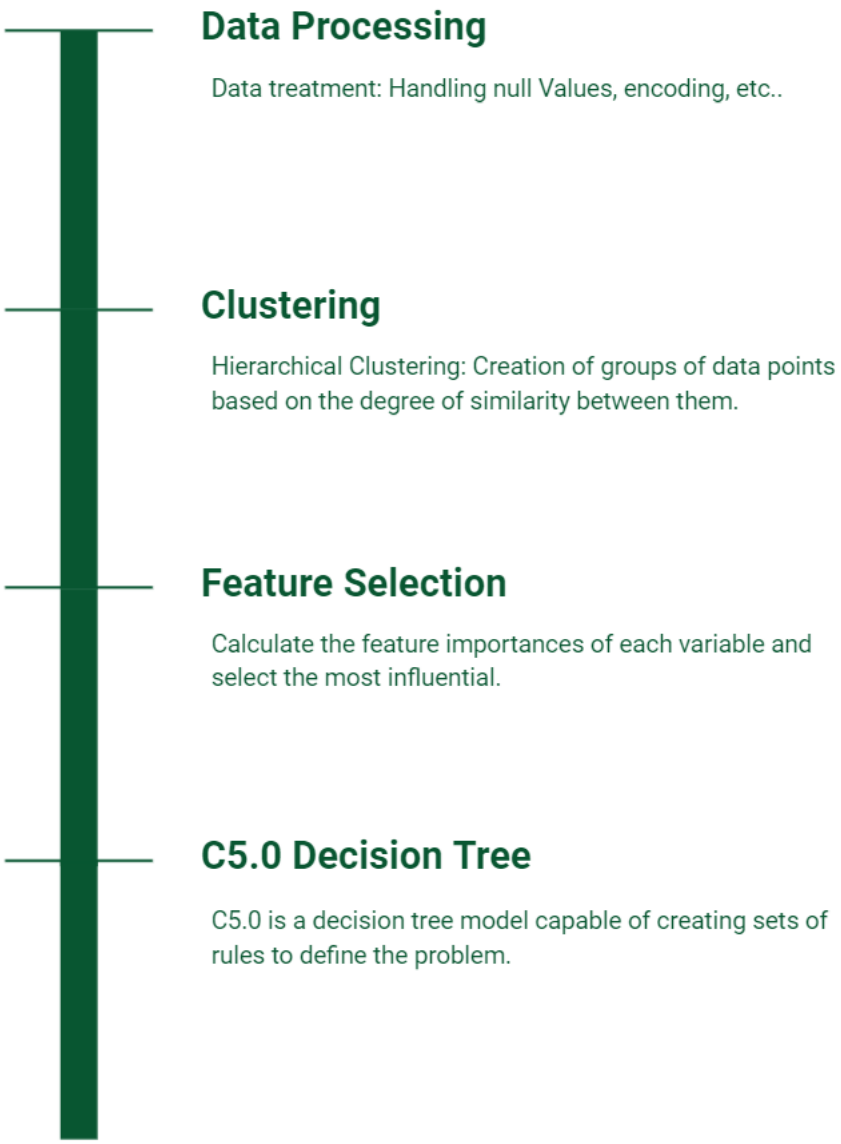

Figure 3.1: Description of the Rule Generation approach

main objective of this step is to facilitate the creation of rules for each cluster by the machine learning algorithms.

Feature selection or variable selection consists in identifying the most important/discriminatory variables in order to simplify the models and eliminate non-impact variables.

Finally, the last step applies the C5.0 algorithm to generate rules. These rules are formed by conditions of several variables to obtain a given class. This way, allowing new ideas and information about the data.

R Language was used to implement this approach, including the libraries "cluster", "randomForest", "C50", among others.

3.2 Data Integration

Within the MOPREVIS project, it was necessary to create a server where project members can perform operations on data. Due to the volume of the data, it is necessary to store them, justifying this solution. These data of a confidential nature must be treated confidentially, which leads to the need to apply security measures when accessing the data.

On the server, the users may install tools they deem necessary to their work. In addition, tools that offer features that facilitate cooperation between project members will be installed.

Data integration software allows the unified visualization of data accessible from multiple sources, transformable into relevant information. Data integration methods can include accessing databases, analyzing files, extracting files, transferring files and web services. Pentaho Data Integration CE (called PDI or Kettle) [5] was one of the solutions found that is capable of supporting access to different types of databases, with ETL (Extract Transform Load) strategies, ease of use and security. In addition to Pentaho Data Integration, there is also the Pentaho Business Analytics Platform (BA server), which is a data visualization and analysis platform with secure access from a web browser.

To complete the data flow, a MySQL database was created which consists of loading files or the database and executing transformations from the PDI where it can connect to the Pentaho BA server for data analysis and reporting. There is also an integrated repository with version control for the files that define the transforms.

An advantage of this software is that members can create transformations in a graphical interface and then, other members will be able to access these transformations and easily understand the operations performed on the data, facilitating cooperation.

Additionally, access to Zeppelin [4] notebooks was made available, which is a web tool with various user profiles where users can develop, organize and perform data operations and share reports with the other team members.

Zeppelin notebooks are documents that contain code from various programming languages, show their output and share conclusions and analyses. Each notebook is structured by blocks that can contain several programming languages, text, graphics, maps, etc..

3.3 Data processing

The data used, consists of 28102 observations of traffic accidents from 2016 to 2019 containing various data sources such as, weather, road, driver, victim, and vehicle information, along with many other variables.

The main data was retrieved from the Statistical Bulletin of Traffic Accidents (BEAV) validated by the National Republican Guard (GNR) of Setúbal and updated by the National Road Safety Authority (ANSR). Atmospheric data from 7 different weather stations across the district of Setúbal was provided by the Portuguese Institute for Sea and Atmosphere (IPMA).

3.3.1 BEAV Data

The BEAV is a statistical notation tool that is filled by the supervising authorities (GNR and PSP) for each case of a road accident events [2]. The BEAV is a main source of information about traffic accidents, supporting studies and assessments to the topic of road safety. Hence, all issues related to the quality and

reliability of the BEAV are of particular importance - missing, inconsistencies or errors in filling out the BEAV have repercussions on the results of the statistics carried out by studies based on this information, and may jeopardize the credibility of the information system of road accidents.

Appendix A [1] presents the BEAV and gives an idea of the data we started with. Section A of the Annexed document is filled for any type of accident, while all the other sections are only filled for accidents with victims. This isn't ideal when the study considers all types of accidents as we have to deal with many null values by finding alternative sources or by outright dismissing the variable in case of no better alternative.

3.3.2 IPMA Data

From the atmospheric data provided by IPMA, we have considered 6 columns of weather information plus the timestamp of the observations. The Variables are:

- Air Temperature in (ºC)
- Average relative humidity (%)
- Average wind direction (degrees)
- Average wind intensity (m/s)
- Rainfall duration (minutes)
- Precipitation amount (mm)
- Station ID
- Date, Hour and minute (in 10 minutes intervals)

To assign the atmospheric data to the accident occurrences, first the location of each accident is assigned to the closer weather station. Then the weather observation is matched to the accident occurrence by the hour. If a variable contains an error value, the variable from the second closest station is used and so on until all accidents have all valid weather variables assigned.

Also various new variables were constructed based on this data:

- Boolean variable for Portuguese holidays
- Boolean variable for Holiday OR Weekend OR Vacation time (between May and September)
- Boolean variable for Working hours (Weekday between 7h and 20h, no holiday)
- Boolean variable for Traffic Peak hours (Weekday between 6h and 10h OR between 17h and 21h, no holiday)
- Integer and String variable defining 4 possible time periods: "Going to Work" between 6h and 10h; "Morning and afternoon" between 10h and 17h; "Going out of work" between 17h and 21h; "Night" between 21h and 6h.
- Boolean variable for whether the date/time matches school hours

3.3.3 Dataset

Other data sources were processed and appended to the centralized dataset but of lesser importance to these these particular approaches. After a comprehensive analysis and error fixing of the initial data, various variables were chosen from this data.

In summary the following variables were used:

Variable	Description	Type
SeasonMov	Weekends and holidays between May and September	Binary
WorkHours	Between 7h and 20h, excluding weekends and holidays	Binary
School	If accident occurred during school hours	Binary
WindSpeed	Wind Speed in m/s on the nearest hour	Numerical Continuous
AirTemp	Air Temperature in ºC, nearest hour	Numerical Continuous
Parking	Accident occurred in a parking lot	Binary
TypeAcid	Type of accident (collision, crash, or run over)	Nominal (3 levels)
TotalDrivers	Total number of drivers involved	Numerical discrete
TypePlace	Urban or rural	Binary
RainQuant	Precipitation amount in mm on previous hour	Numerical Continuous
HitAndRun	A driver escaped the scene after the hit	Binary
WeekDay	Week Day (1 to 7)	Numerical discrete
County	County where accident occurred	Nominal (15 levels)
Month	Month when accident occurred (1 to 12)	Numerical discrete
HourNear	Nearest Hour	Numerical discrete
Motorcycle	Accident involved motorcycle or similar	Binary
LightVehicle	Accident involved light vehicles	Binary
HeavyVehicle	Accident involved heavy vehicles	Binary
RoadType	Type of road	Nominal (11 Levels)
Severity	Accident Severity	Binary, Dependent Variable

Table 3.1: Variables Description.

The Variable "Severity" is the dependent variable that will classify the accidents with or without victims.

3.4 Clustering

In the following paragraphs, we will explain some methods for clustering and respective algorithms.

Clustering is a common task of unsupervised learning, and unlike supervised learning, it does not require labelled dataset to work, instead, it discovers patterns on its own. Clustering is the process of grouping data points based on the similarity between them. The result of clustering will be groups of similar data points called clusters.

There are various types of Clustering algorithms, the most common being:

- Partitioning methods
- Hierarchical clustering
- Density-based clustering

Partitioning clustering groups a user specified number of clusters k based on a criterion function. Hierarchical clustering builds a hierarchy of clusters and can be represented in dendrograms. Density-based clustering groups together data points that are in a dense region of a data space. Low density regions separate the clusters and are classified as noise. For a more detailed overview consider for instance [36].

For this approach we will compare the algorithms k-means clustering and Agglomerative hierarchical clustering.

3.4.1 K-means

K-means is one of the most frequently used partitioning algorithms. This algorithm aims at creating groups of data points based on the K value that defines the number of clusters. Initially, K clusters are selected randomly within the data. The algorithm then reassigns the data points nearest to the centroid of the cluster and recalculates it. This process repeats until the criterion function converges [36].

Summarizing, the algorithms steps are:

- Define a value for K, the number of clusters
- Choose K random points as the cluster centers
- Assign points that are nearest to a cluster center to that cluster.
- Recompute the new clusters centers
- Repeat the steps until the criterion function is satisfied.

3.4.2 Agglomerative Hierarchical Clustering

Hierarchical Clustering aims to build a hierarchy of clusters. Specifically, the Agglomerative Hierarchical Clustering, starts with each data point as a singleton cluster (in the bottom) and iteratively merges the clusters until the final clusters includes all the data points (top cluster). The result of this method is a dendrogram of clusters as shown in Figure 3.2.

Summarizing, the algorithms steps are:

- Take each data point as a singleton cluster and specify the distance function measuring the proximity between clusters.

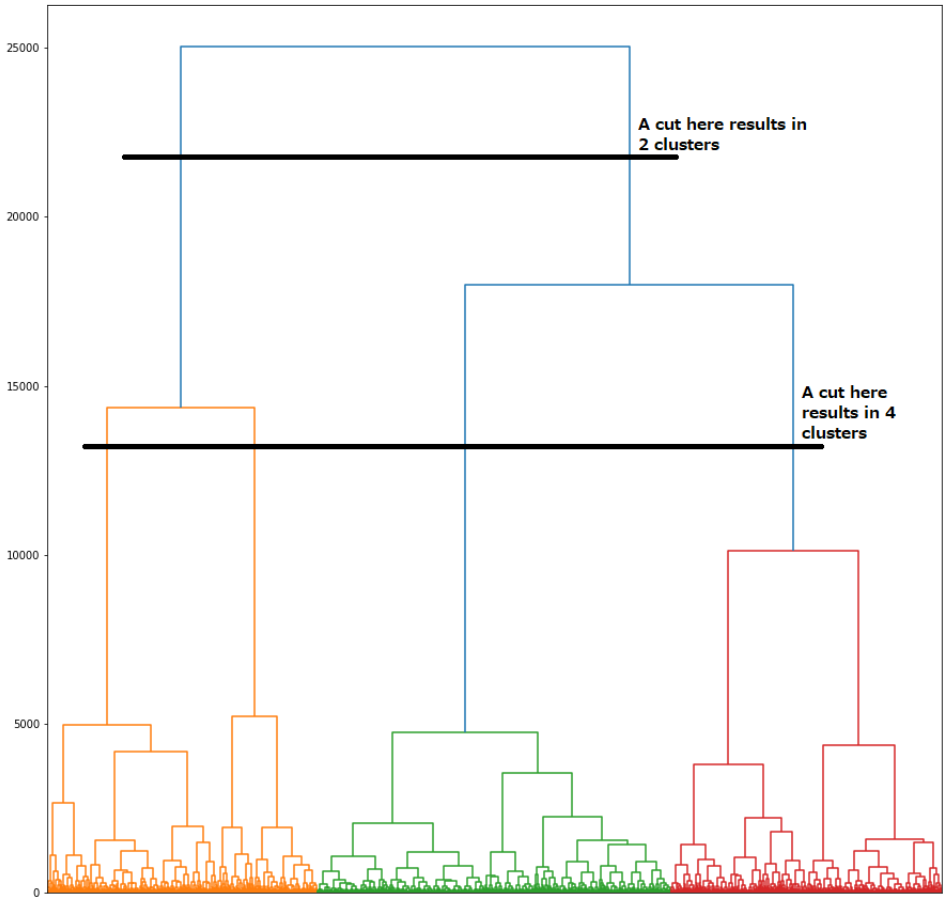

Figure 3.2: Dendrogram Example on a subset of our data

- Find the similar (closest) pair of clusters according to the distance function
- Merge the similar clusters
- Repeat steps until the last cluster containing all data points is formed

3.4.3 Evaluation

For this specific approach, the Silhouette Index [30] was used as an evaluation measure to compare the performance of the agglomerative hierarchical clustering and the k-means algorithms. The silhouette Index ranges from [-1,1], with -1 indicating poor consistency within clusters and 1 indicating excellent consistency within clusters. Values near 0 suggest overlapping clusters.

When both algorithms were applied to the data set and the silhouette index of the resulting groups was calculated, both approaches had a similar maximum value. Because none of the methods produced better indexes than the other in this circumstance, hierarchical clustering was selected.

When deciding on the number of groups, it was discovered that the two-cluster models produced the best silhouette index results for both algorithms. The fluctuation of the silhouette index and the number of clusters for hierarchical clustering is shown in Figure 3.3.

The data was then separated into two clusters, each with 20732 and 7371 observations. With a 0.18

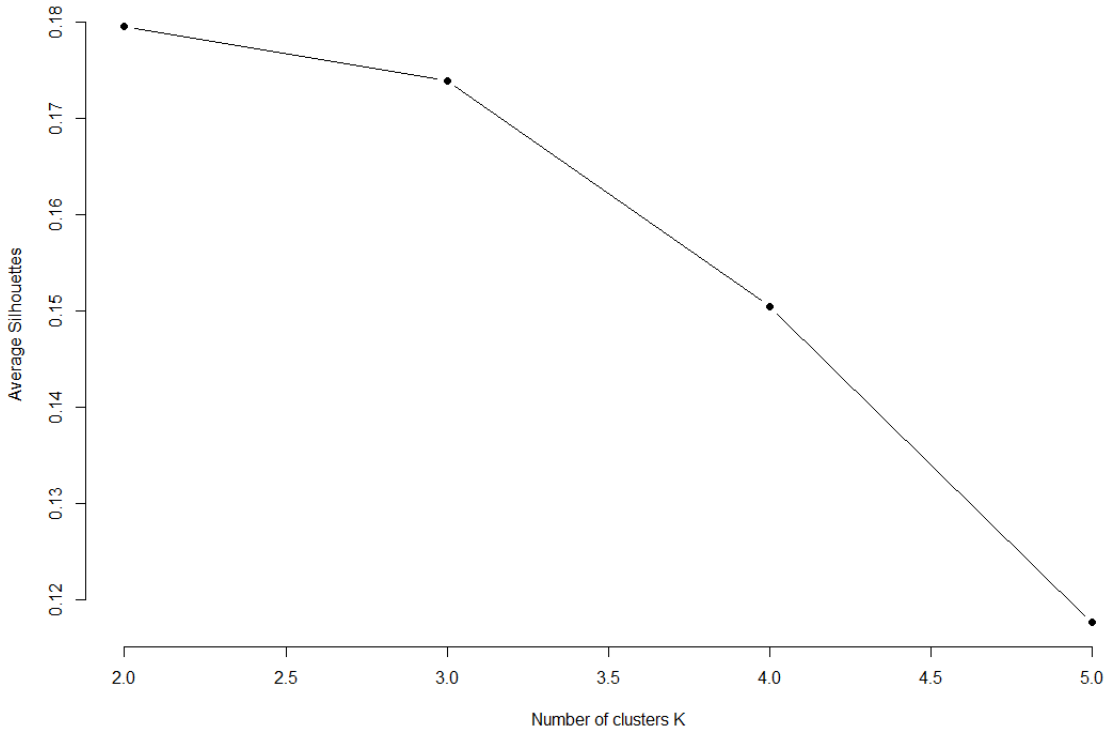

Figure 3.3: Description of the Rule Generation approach

Silhouette Index.

3.5 Feature Selection

The most influential variables of each cluster are selected in the next step. The "feature importances", which reflects the relevance of each variable, was calculated using random forest.

Random forests or random decision forests [19] are an example of ensemble learning technique for classification, regression and other tasks that operates by combining a collection of random decision trees at training time to achieve high classification accuracy. For classification tasks, the random forest's output is the class chosen by the majority of trees.

In a regression or classification problem, random forests may also be used to rank the importance of predictors/variables [10]. Based on the Mean Decrease Accuracy (MDA) values, we select the most influential variables in each cluster using the R package "randomForest." The MDA values of a variable tells us how much that particular variable reduces/decreases the accuracy of the model if removed. The higher the value of mean decrease accuracy, the higher the importance of the variable. The variables were ordered by descending order of the MDA values, and a cut was made on the top 6 variables or when a considerable drop in MDA value was present.

The resulting variables for the first cluster are specified in Figure 3.4 which lists the most important variables in descending order by MDA value.

The resulting variables for the second smaller cluster are specified in Figure 3.5.

It was also calculated the feature importances for the whole data set. This second experiment basically

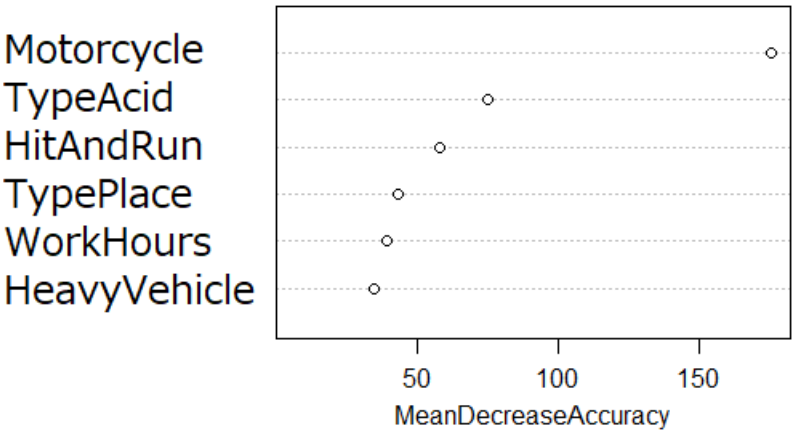

Figure 3.4: Most important variables for Cluster 1

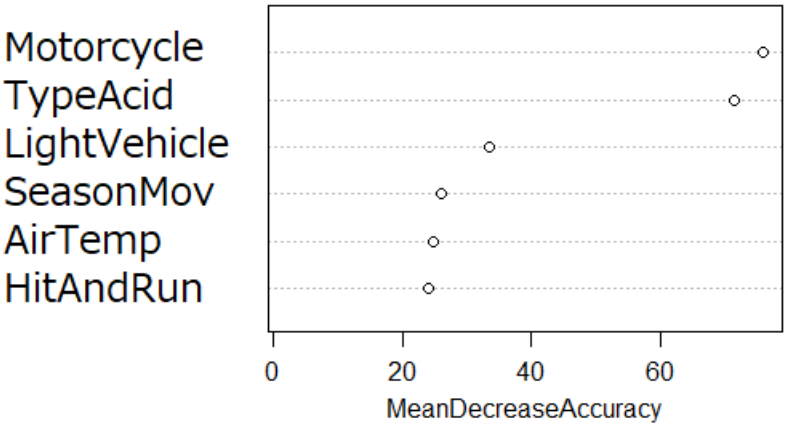

Figure 3.5: Most important variables for Cluster 2

ignores the clustering from the previous stage. The resulting main features are presented in Figure 3.6:

By applying the random forest to rank the variable importances we can already observe that accidents involving motorcycles are an important factor when classifying the severity of the accidents on our data. "TypeAcid" (whether the accident is a collision, single vehicle crash or trampling) also presents a considerable high MDA value. "HitAndRun" (whether a driver escaped after the accident) also appears in all the presented Figures.

3.6 Rule Generation Models

Machine learning algorithms usually fall into two categories, supervised and unsupervised learning. As previously mentioned, supervised learning uses labelled data, or more specifically, data points with correct outputs as opposed to unsupervised learning which is not a requirement [22]. Supervised learning algorithms will attempt to classify and predict the target output values based on the relationship between the outputs and inputs, which is learned from previous data sets.

Classification is a common supervised learning task that separates the data using a discrete target variable, more specifically, binary classification which has two possible output values, for example, "yes or no", "0 or

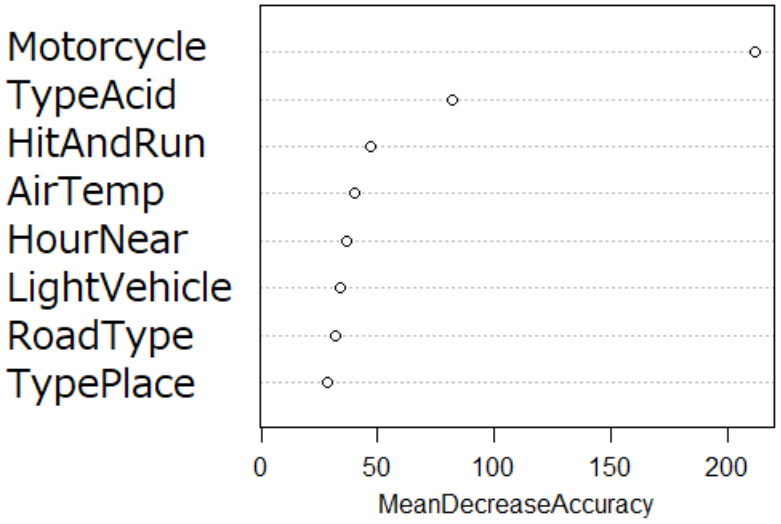

Figure 3.6: Most important variables for the whole data set

1". As an illustration of such algorithms, consider for instance logistic regression, random forest, support vector machines, decision trees, naive bayes, etc. [31]. In this particular work the output values are "No Victims" or "With Victims" i.e. 0 or 1, for the "Severity" variable.

The algorithm used in this approach was the C5.0 algorithm, which creates decision trees and can also generate rules. The rule sets represent a simplified version of the important information found the in the decision trees.

Models for each cluster and the entire data set were built. However, it is vital to note that the created rules in the models applied to each cluster should not be interpreted as a general rule because they only apply to a part of the data (cluster).

For cluster 1, four rules were created, and Table 3.2 depicts its result. The rules sets were adapted to be easier to understand and read. The following Tables show the Rule nº, the adapted rule, the number of observations that the rule can be applied to, the percentage of observations where the rule fails (incorrect class), and the class of the applied rule. For example, rule º1 says that "Trampling in urban areas,no motorcycles involved" result in accidents with victims. The data shows that 460 observations are "Trampling in urban areas,no motorcycles involved", of which 17% of these, are accidents with no victims (17 % fail the rule).

Rule nº	Rule	Obs. nº	Error %	Class
1	Trampling in urban areas, no motorcycles involved	460	15%	With Victims
2	Motorcycles involved	1575	37%	With Victims
3	Trampling in rural areas	263	16%	No Victims
4	Collisions and Crashes, no motorcycles involved	18434	16%	No Victims

Table 3.2: Rule results for Cluster 1

Following a thorough examination of some of these rules, the following findings were reached: When we apply the first rule to the complete dataset, we find that this holds true for 83 percent of the accidents. When the requirement "Motorcycle $= 1$" is removed from the rule, the result is substantially similar. In general, 63% of trampling have victims, therefore this rule indicates that trampling that occur within the localities are more severe. This observation is backed up by Rule 3. The possibility of being run over by animals in rural areas that do not produce "victims" could explain this observation.

A total of 21% of all accidents result in fatalities. When it comes to accidents involving motorcycles or similar vehicles, the number jumps to 69%. As a result, when motorcycles are involved, accidents are more serious. This is shown by Rules 2 and 4.

For cluster 2, three rules were created, and Table 3.3 shows the result.

Rule nº	Rule	Obs. nº	Error %	Class
1	Motorcycles involved	1133	24%	With Victims
2	Trampling	247	26%	With Victims
3	Collisions and Crashes, no motorcycles involved	6030	9%	No Victims

Table 3.3: Rule results for Cluster 2

Cluster 2's results are comparable to Cluster 1's results in certain ways. Again, we see how serious accidents are when motorcycles or similar vehicles are involved (rules 1 and 3). As previously said, 63 percent of pedestrians run overs become victims, which is a far larger percentage than crashes and collisions, as this rule demonstrates.

Finally, Table 3.4 shows results of the model applied to the whole dataset.

Rule nº	Rule	Obs. nº	Error %	Class
1	Trampling in urban areas	681	17%	With Victims
2	Accidents With Motorcycles, no Ligh Vehicles involved	915	19%	With Victims
3	Accidents With Motorcycles, no Hit and Run	2525	30%	With Victims
4	Trampling	977	37%	With Victims
5	Accidents with Light vehicles, with a Hit And Run	3803	4%	No victims
6	Accidents without motorcycles	25395	16%	No Victims

Table 3.4: Rule results for the whole dataset

The severity of pedestrian accidents is addressed under Rules 1 and 4. The severity of incidents involving motorbikes and similar vehicles is addressed under Rules 2, 3 and 6.

Rule 5 is particularly intriguing: there have been 3 990 hit and run incidents, yet only 233 (about 6%)

have resulted in victims. As previously stated, casualties are present in 21% of all accidents, a much higher percentage.

3.7 Summary

Although the model that skipped the clustering stage also presented useful information (Table 3.4), the clustering provides more homogeneity to the cluster data, which could allow the model to give more concise rules.

The feature selection stages gives us an initial idea of what factors are most important, and allows for the less relevant variables to be absent in the rule sets, otherwise the rules would include too many factors, decreasing its readability.

Finally, the generated rule sets, although requiring a deeper examination of each rule, it provides a good pointer as to what factors influence accident severity, thus finding hidden patterns for fatal and non-fatal traffic accidents.

Overall the models point out that motorcycles or similar vehicles and pedestrian accidents are more likely to result in accidents with victims, while also reinforcing these findings by pointing out the opposite, that other vehicle types (no motorcycles involved) and other accident types (collision or single vehicle crash) do not usually result in victims. Whether the pedestrian accidents happened inside or outside urban areas also affect accident severity. The models also found out that hit and runs occurrences have a lower chance of resulting in accidents with victims.

4

Prediction Model

In this chapter, we propose a real-time hotspot predictive model. It's introduced an in-depth description of the methods and algorithms used in this approach, including the hotspot generation and model training and evaluation process.

4.1 Overview

After the generation of rules and better understanding of the dataset, another approach was created that aims to build a system capable of predicting traffic accident hotspots. Figure 4.1 gives an outlook of the overall workings of the system.

The Figure shows the clustering algorithm taking as input the geographical coordinates of the accidents and adding its output as input of the predictive model. Moreover, the data inputs of the predictive model also contain weather, road, and time information. Finally, given a date and time, the system will then predict and map the traffic accidents hotspots.

Python Language was used to implement this approach and the libraries "pandas", "NumPy" and "Scikit-learn" were the main libraries used.

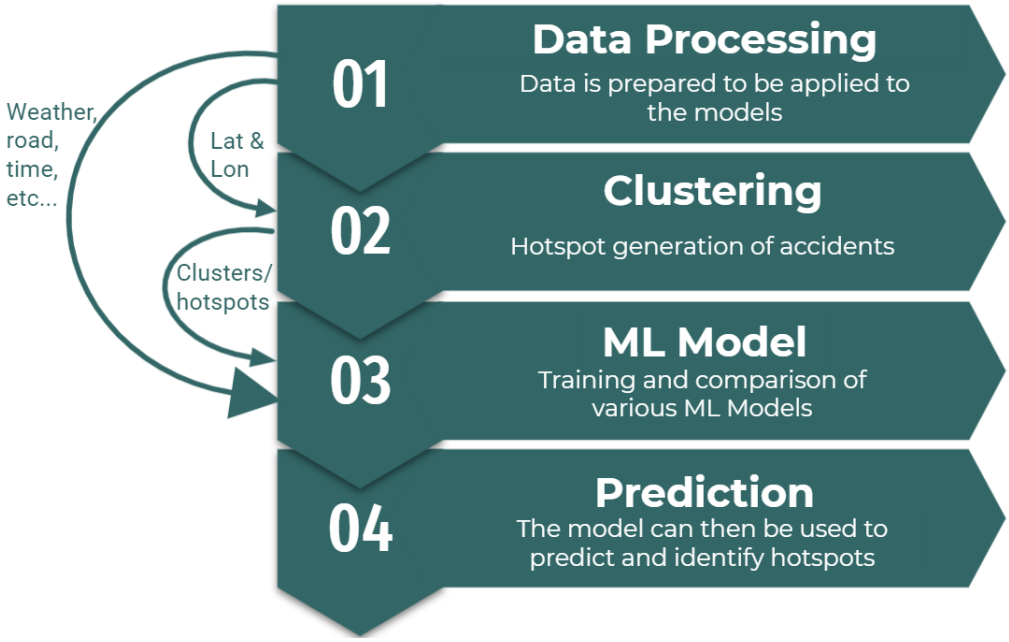

Figure 4.1: Description of the hotspot prediction approach

This work can be divided into data processing, clustering, predictive model training, and prediction. In the following sections, we briefly explain these stages.

4.2 Data Processing

Again, the same data containing the 28102 observations of traffic accidents were prepared for this new approach. The main criteria for variable selection in this approach is including variables that can be predicted for a set location and future date/time such as weather; thus, in this case, information regarding the date, time, weather condition, and location of the accidents are the main features. In summary, the following variables were used:

- "Latitude/Longitude"—The Latitude and Longitude coordinates.
- "WindDirection"—Average Wind Direction (Degrees).
- "DayOfYear"—Day of year (1 to 365/366).
- "Hour"—Nearest hour.
- "Day"—Day of month.
- "RegularSpeed"—Historical regular speed in segment in km/h.
- "DayShift"—Periods of the day.
- "Year"—Year when accident occurred.
- "HasDivider"—Road has a divider that separates the traffic flow in opposite direction.
- "TrafficPeak"—Accident occurred in a rush hour.

- "PathType"—Whether it occurred in a turn or straight.
- "Holidays"—The day is an holiday.
- "DamagedRoad"—Whether the road as significant damage.

Furthermore, variables described previously were used: SeasonalMov, WorkHours, School, AirTemp, TypePlace, RainQuant, DayOfWeek, Month, Hour, and RoadType. See Section 3.3.1 and Section 3.3.2 for more details on the datasources.

4.3 Clustering

One of the first steps in road safety improvement is the identification of hotspots or hazardous road locations, also known as black spots. As mentioned in Section 2.5, the paper by Szénási and Csiba [38] uses a clustering method (DBSCAN) in order to find accident hotspots using the accident's GPS coordinates. This gives us the possibility to use the whole dataset regardless of road characteristics and group accidents that are in proximity to one another. This hotspot concept and this method of identifying hotspots are also used in our work.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) works with two important parameters: epsilon and minimum points. Epsilon or eps, determines the maximum distance between two points to be considered neighbors (belonging to the same cluster). Minimum points or MinPts determine the minimum data points that are necessary to form a cluster. Otherwise, the data points are declared as noise (do not form a cluster).

There are no general methods for determining the ideal Eps and MinPts in this situation, as we want a certain area size for the clusters. Using such methods such as silhouette score, elbow curve, etc., would result in a few very large clusters of a few kilometers wide. The idea is to have a cluster size no larger than a few road segments or intersections, although it might still occur.

After a few experiments with changing Eps and MinPts, we found that assigning 150 m to Epsilon and 10 accidents as minimum points provided an acceptable size for the clusters, similar to the area size of a few intersections or road segments. Basically, a cluster will have at least 10 accidents, and in each cluster, the accidents will have in a 150 m radius, at least one other accident of the same cluster.

Figure 4.2 shows the overview of the resulting clustering while Figure 4.3 shows an acceptable size for the clusters.

Overall, the DBSCAN algorithm generated 298 clusters, for a total of 18457 observations assigned to some cluster. Other 9645 observations were not assigned to any cluster.

4.4 The Model

This work falls under the category of a classification problem, as we want to classify which hotspots are "activated" in given circumstances.

Figure 4.4 describes the general topic of the classification problem specific to this work.

Another important task is the generation of the negative samples. Negative samples are necessary for the binary classification models, as the original dataset contains only positive samples (actual accidents). For the negative sampling generation, we followed an approach used in [47] that basically generates three

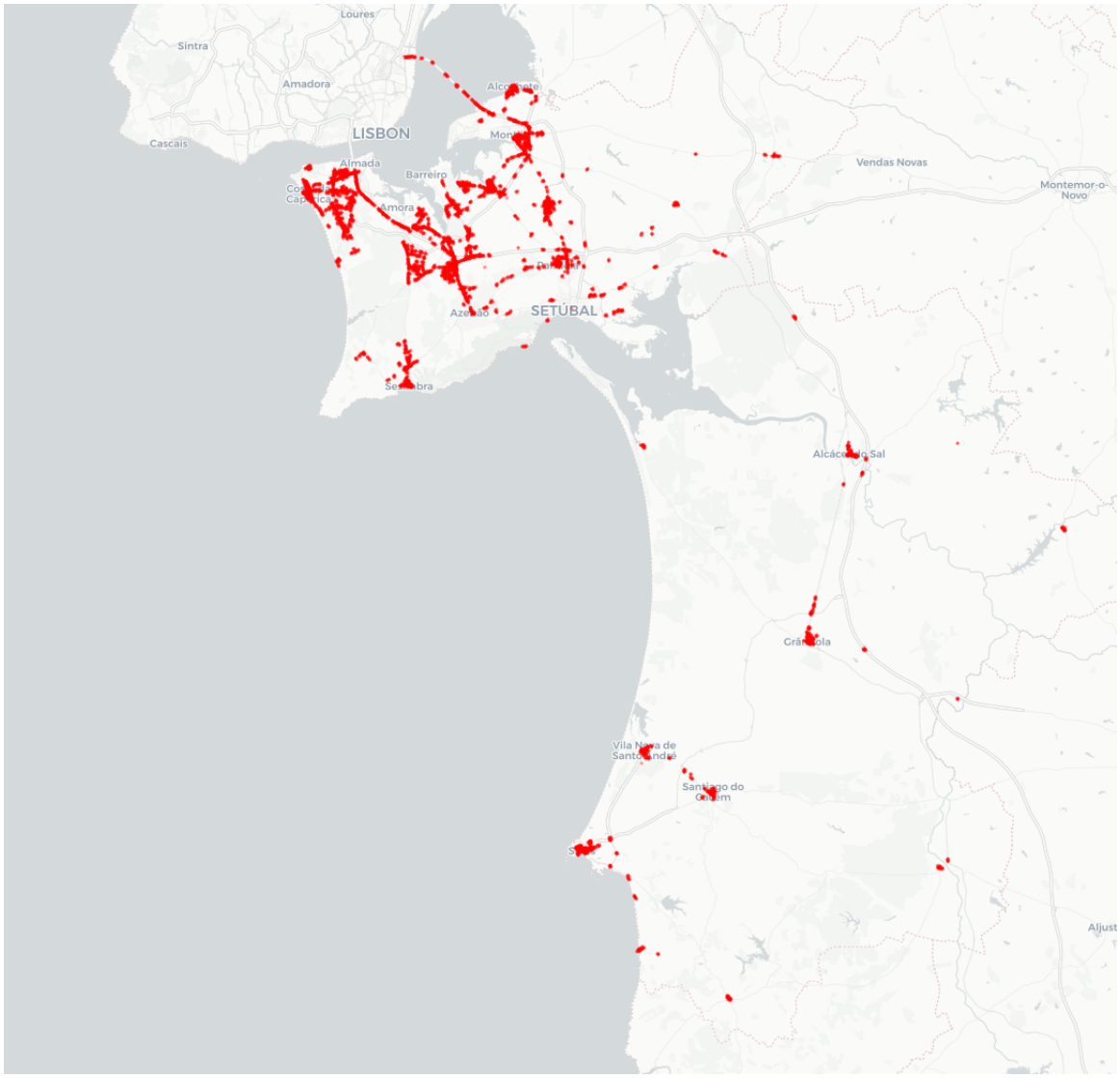

Figure 4.2: Overview of the clusters in the whole district

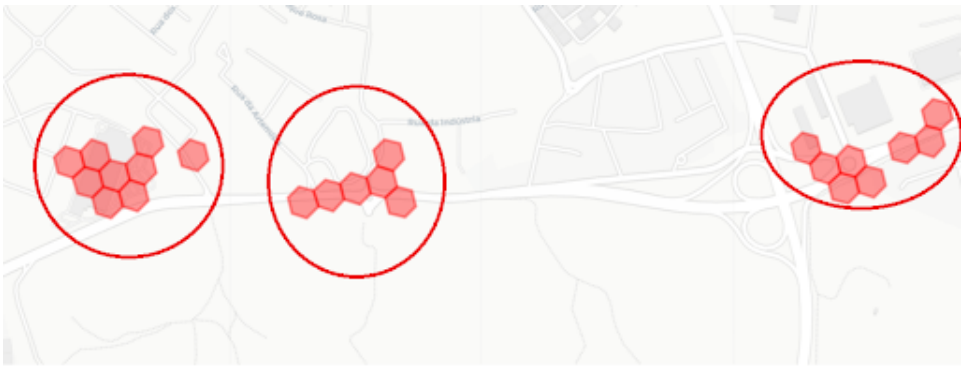

Figure 4.3: Zoom of the map that shows an acceptable size for the clusters (3 separate clusters)

negative samples for each accident randomly changing the date and time and consequently obtaining updated weather conditions for said date/time; making sure there are no negative samples equal to existing

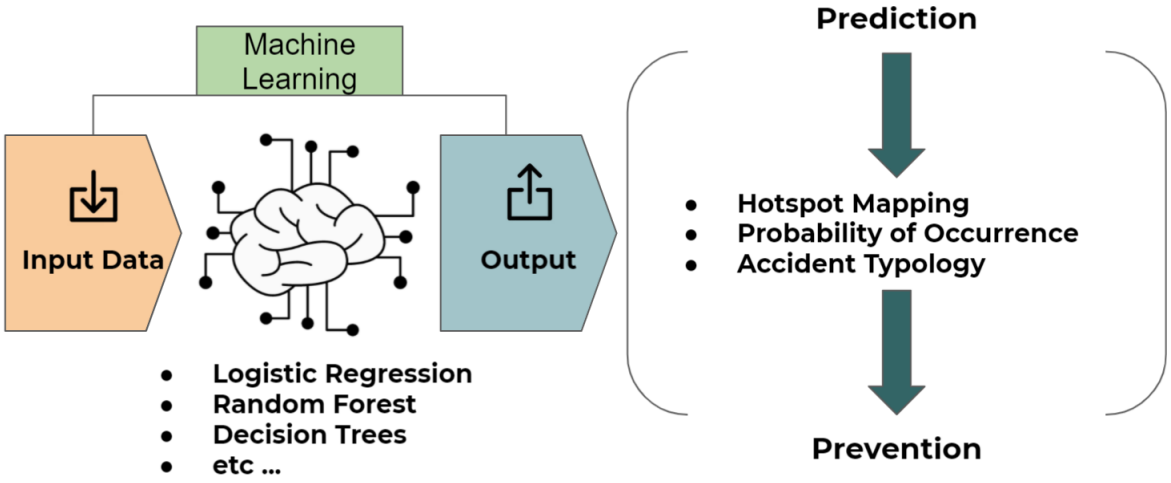

Figure 4.4: Overview of the classification problem in this work

positive samples.

Various tests were conducted with a different number of negative samples including full negative sampling (for every single hour when no accidents occurred). Results showed that having three times negative samples than positive samples had the best balance of sensitivity, specificity, and accuracy metrics in the model evaluation stage.

As we want to comprehensively validate and compare our models, we consider various performance metrics such as accuracy, sensitivity, specificity, precision, false positive rate, and AUC Score. To calculate these metric's values the following measures are needed:

- True positives (TP)—The model correctly predicts the positive class.
- False positives (FP)—The model of incorrectly predicts the positive class.
- True negatives (TN)—The model of correctly predicts the negative class.
- False negatives (FN)—The model of incorrectly predicts the negative class.

After counting the number of the different outcomes the following performance metrics can be calculated:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} = \frac{TP + TN}{All\ Samples} \tag{4.1}$$

$$Sensitivity = True\ Positive\ Rate(TPR) = \frac{TP}{TP + FN} \tag{4.2}$$

$$Specificity = True\ Negative\ Rate(TNR) = \frac{TN}{TN + FP} \tag{4.3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.4}$$

$$False\ Positive\ Rate(FPR) = \frac{FP}{FP + TN} \tag{4.5}$$

AUC score (area under the curve) measures the area underneath the ROC curve (receiver operating characteristic curve), which is a graph that plots the TPR and FPR. The higher the AUC value, the better the

model is at distinguishing crashes and non-crashes.

4.5 Model Results

The initial tests were made using logistic regression, decision trees, and random forests, the latter having the better results. The data was split into 70% training data and 30% test data. The evaluation can be seen in Table 4.1.

Model	Accuracy	AUC Score	Precision	Sensitivity	Specificity
Random Forest	0.73	0.68	0.44	0.08	0.97
Logistic Regression	0.73	0.66	0.27	0.00	1.00
Decision Trees	0.65	0.55	0.35	0.34	0.76
Naive Bayes	0.68	0.67	0.39	0.38	0.79

Table 4.1: Model Evaluation metrics

Results in Table 4.1 show that random forests have the best results. Its sensitivity and specificity tells us that the model is quite conservative, having many false negatives, but is quite good at preventing "false alarms" with a false positive rate (FPR) of 3%. Figure 4.5 shows the resulting ROC Curve for the random forest classifier, the AUC score indicating a fair value of 0.68.

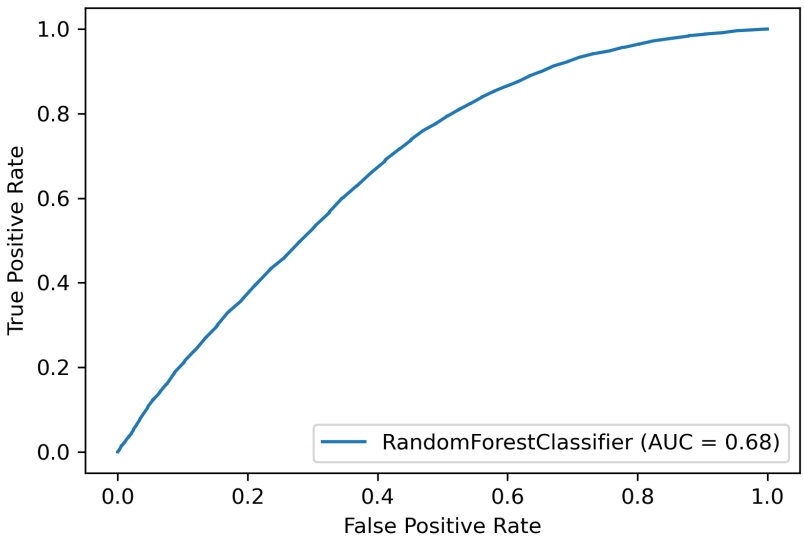

Figure 4.5: Random Forest AUC (Area under the ROC Curve)

An advantage of random forests is the visualization of the feature importance, as shown in Figure 4.6. Observing this figure, we can make changes to the dataset eliminating features that do not influence the results while having a better understanding of which features are the most influential.

The figure shows that the random forest model places a higher importance on the location of the accidents (cluster/hotspot) when compared to other variables.

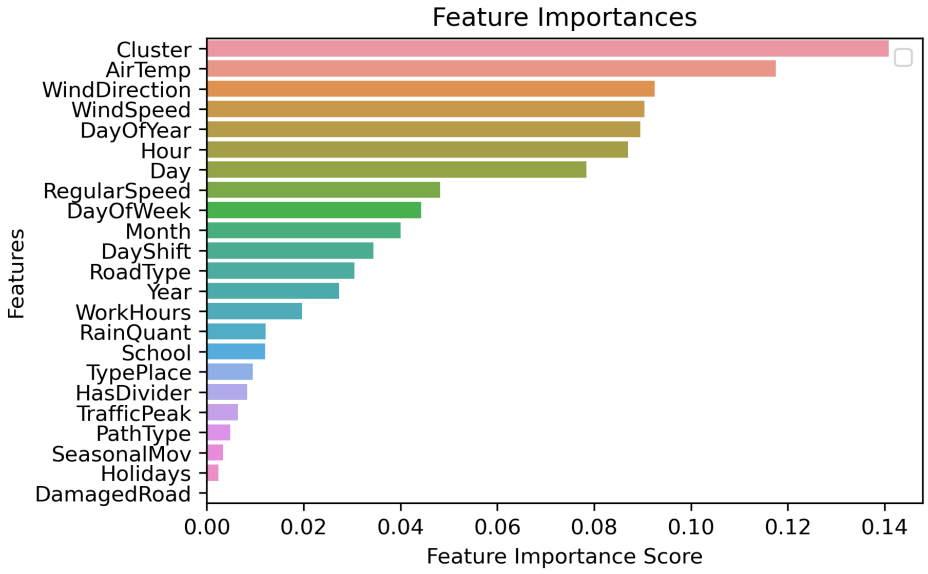

Figure 4.6: Random Forest Feature Importances

It could also be argued that Naive Bayes had the best results, it had the highest sensitivity of the trained models and an acceptable specificity. However, although it has a balance of the evaluation metrics, it does not excel at particularly anything unlike the random forest model. Nonetheless, it should be taken into account in future tests.

Decision Trees also has a balance of specificity and sensitivity values but it performed worse in all aspects when compared to NB.

We have also applied this approach and created models for specific accident types, single vehicle crashes, multi-vehicle collisions or pedestrian accident. Table 4.2 presents the results of the random forest models for each type.

Accident Type (model)	Accuracy	AUC Score	Precision	Sensitivity	Specificity
Multi-vehicle Collisions (RF)	0.73	0.71	0.47	0.11	0.95
Single-vehicle Crash (RF)	0.70	0.51	0.28	0.05	0.95
Pedestrian accident (RF)	0.77	0.68	0.39	0.26	0.90

Table 4.2: Random Forest models for each accident type

Overall, the DBSCAN clustering algorithm generated 253 clusters for collision type accidents, for a total of 14293 observations assigned to the clusters. The model's performance is similar to results shown in Table 4.1 (random forest) which is consistent with the fact that collisions constitutes the majority of the observations.

For single-vehicle crashes, the DBSCAN generated 21 clusters, for a total of 513 crashes assigned to the clusters. Of all the three models, this model had the worst performance.

For accidents involving pedestrians, using the same parameters of the DBSCAN as used before resulted in only 3 hotspots. This is because this type of accident is far less frequent than multi-vehicle collisions. Adjustments were made to the parameters, eps=200 m and minPts=5, which resulted in 16 clusters, and a

total of 176 accidents assigned to the clusters. Of all the random forest model this model had the highest sensitivity. Given the low number of data samples it's difficult to arrive to conclusions.

4.6 Summary

In this chapter we developed a hotspot predictive model. This approach starts by generating clusters/hotspots using the DBSCAN algorithm. Its parameters were chosen arbitrarily in a manner that it usually results in clusters positioned in road intersections or a few road segments. Otherwise, if these parameters where chosen based on existing evaluation metrics it would result in too big and few clusters.

Afterwards, various machine learning algorithms were tested and compared. Random forest and baive bayes models were among the best results, the first one having a very good FPR percentage but a low Sensitivity, while the naive bayes was more balanced.

Additional experiments were made with subsets of the data, namely multi-vehicle collisions, single-vehicle crashes and pedestrian accidents which showed interesting results. Random forest models were used for these experiments which slightly improved model performance except in single vehicle crashes, which worsened. Additional variables should be added to the dataset and these experiments repeated for a better interpretation of the results.

5

Conclusion

This chapter presents a summary of the the work described in this dissertation and some remarks about the conclusions observed. Some future work plans are also presented based on the research shown on this document.

5.1 Global Overview

This work analyzes road accidents that occurred in the district of Setúbal, joining data from various data sources.

The results from the rule generation model were successful in finding patterns for fatal and non-fatal traffic accidents. According to the model, pedestrian accidents and accidents involving motorcycles are the main factors that have a higher chance of resulting in victims, whereas most collisions and crashes that do not involve motorcycles do not result in injuries. Intriguingly, hit-and-run incidents are less likely to result in a victim. These results were discussed and validated by the MOPREVIS project team, which include road safety experts from GNR.

By clustering the data prior to the generation of the rules will facilitate the rule generation by the model but it should be taken into account that such rules should not be interpreted as a general rule of the entire data and should be tested for its veracity afterwards.

Comparing this results to the similar work by Siam et al. [34], both theirs and our work reached different conclusions and highlighted factors, which is expected as there are differences in the data itself and how it was processed and divided. For example they found that national/regional/rural roads with no divider have more chances of fatal accidents. The same can be said about other factors highlighted by researchers in Section 2.2. Overall, these results can help us understand hidden aspects of our data, that are not easily obtained in statistical data distributions or common univariate/bivariate analysis.

In the accident prediction study, we have used a dataset containing vehicle accidents in the road network of the district of Setúbal, as well as historical weather information for such accidents. Using this dataset, we extracted the relevant features for accident prediction and created positive examples, corresponding to the occurrence of a collision, single vehicle crash or pedestrian accidents, and negative examples corresponding to non-occurrences of accidents. We then generated the spatial clusters of high accident density using the DBSCAN algorithm. With the data now complete, we compared some ML algorithms such as random forest, decision trees, logistic regression, and naive bayes. Then we focused on random forest algorithm as it proved to have one of the best performances. Even so, the model proved to be quite conservative with a false positive rate (FPR) of 3%, specificity of 0.97, and sensitivity of 0.08, reaching an accuracy of 73%; further development of this approach is required to improve these results.

When comparing these results to literature, it is important to note that, as most examples belong to the negative class, the model that contains the higher negative/positive samples ratio is usually the one with the highest accuracy. Considering this, our work achieved an excellent FPR when compared to other works [20, 24, 42] mentioned in Section 2.4, Table 2.1, while sensitivity is still not in an acceptable range.

By utilizing only multi-vehicle collisions, which is the most common accident type, there is a slight improvement in performance of the model. Although the improvement not significant, this could indicate that raising the homogeneity of the dataset while retaining the high number of observations might improve performance.

In this project we successfully developed a rule generation model capable of highlighting factors responsible for severe accidents, as well as the development of a hotspot predictive approach.

We have created a data integration solution available to the MOPREVIS team, as well as contributing to the data repository that fuses various data sources relating to traffic accidents, enabling the development of various machine learning experiments on this data, its evaluation and comparison. Finally, we have presented our interpretation of the results.

5.2 Future Work

Although most of the objectives were achieved, the work must still be improved. Future work aims at further developing predictive models.

Initial tests are quite inconsistent and more work and data are required in this task before obtaining conclusive results. For future work, more recent data (2020/2021) will be provided to us that will allow us to improve the proposed work.

Since we have highlighted motorcycles accidents as the main factor influencing accident severity it would be interesting to include traffic parameters and intensity to our approaches and compare the results to the

results of Theofilatos et al. (2016) [43], which has identified traffic flow and speed variations to influence powered two-wheeler (PTW) crashes. This new variables should also potentially improve the predictive model performance.

Additional machine learning algorithms and especially neural networks and deep learning approaches will also be applied as it has proven to be successful and sometimes outperforming simpler algorithms [42, 28, 46].

The tweaking of the parameters, such as DBSCAN parameters, data sampling (undersampling and oversampling), and the model parameters should also be comprehensively experimented on.

Furthermore, other paths may be taken, such as using crash frequency as dependent variable, which is also a popular approach in literature, and/or focusing on particular arterial roads or highways, raising the homogeneity of the data.

Finally, with our contribution, a digital decision support tool will be developed to support GNR to make more informed decisions regarding road accidents prevention.

A

Statistical Bulletin of Traffic Accidents (BEAV)

Instrumento de notação registado no I.N.E., sob o n.º 10111, válido até 31/12/2014

Nº Boletim

Entidade Fiscalizadora

AUTORIDADE NACIONAL SEGURANÇA RODOVIÁRIA
Ministério da Administração Interna

BOLETIM ESTATÍSTICO DE ACIDENTES DE VIAÇÃO

(Reservado ao carimbo da unidade)

A - a preencher em todos os acidentes **B e seguintes - a preencher apenas em acidentes com vítimas**

A - IDENTIFICAÇÃO DO ACIDENTE

A1 DATA/HORA

Ano Mês Dia Hora Min.

A2 LOCALIZAÇÃO

1. ☐ Fora das localidades
   ☐ Dentro das localidades
2. Distrito
   Concelho
   Freguesia
   Povoação (ou a mais próxima)
   Coordenadas GPS
   Latitude + /
   Longitude - /
3. Designação de via
   Km /
   Arruamento
   n.º
4. Se houver separador central indique em que sentido
   1 ☐ Crescente
   2 ☐ Decrescente

A3 TIPO DE ACIDENTE

1 ☐ Acidente só com danos materiais
2 ☐ Acidente com vítimas
   Mortos
   Feridos graves
   Feridos leves

A4 NATUREZA DO ACIDENTE

1 ☐ Despiste
2 ☐ Colisão
3 ☐ Atropelamento

A5 NÚMERO DE VEÍCULOS INTERVENIENTES

Ciclomotor e motociclo
Veículo ligeiro
Veículo pesado
Outros

A6 CONDUTORES INTERVENIENTES

1. SEXO

	A	B	C
1 Masculino	☐	☐	☐
2 Feminino	☐	☐	☐

2. DATA DE NASCIMENTO

Ano Mês Dia
A
Ano Mês Dia
B
Ano Mês Dia
C

B - CIRCUNSTÂNCIAS EXTERNAS

B1 CARACTERÍSTICAS TÉCNICAS DA VIA

1. ESTRADA COM SEPARADOR
   1 Autoestrada – nº de vias de trânsito no sentido
   2 Outra via – nº de vias de trânsito no sentido
2. ESTRADA SEM SEPARADOR – nº de vias no sentido
3. VIA DE TRÂNSITO
   1 ☐ Esquerda
   2 ☐ Direita
   3 ☐ Central

B2 TRAÇADO DA VIA

1. EM PLANTA
   1 ☐ Reta
   2 ☐ Curva
2. EM PERFIL
   1 ☐ Em patamar
   2 ☐ Com inclinação
   3 ☐ Em lomba
3. 1 ☐ Sem berma ou impraticável
   2 ☐ Berma não pavimentada
   3 ☐ Berma pavimentada
4. SITUAÇÃO DO ACIDENTE
   1 ☐ Em plena via
   2 ☐ Na berma
   3 ☐ No passeio
   4 ☐ Em via ou pista reservada
   5 ☐ Em parque de estacionamento
5. INTERSEÇÃO DE VIAS
   1 ☐ **Fora da interseção**
   Em interseção de nível
   2 ☐ Em cruzamento
   3 ☐ Em entroncamento
   4 ☐ Em rotunda
   5 ☐ Em passagem de nível
   Em interseção desnivelada
   6 ☐ Em via de aceleração
   7 ☐ Em via de desaceleração
   8 ☐ Em ramo de ligação – entrada
   9 ☐ Em ramo de ligação – saída
6. ACIDENTE EM OBRAS DE ARTE
   1 ☐ Túnel
   2 ☐ Viaduto/Ponte
   3 ☐ Passagem estreita

B3 REGIME DE CIRCULAÇÃO

1. FAIXA DE RODAGEM COM
   1 ☐ Sentido único
   2 ☐ Dois sentidos
   3 ☐ Reversível
2. VELOCIDADE PERMITIDA NO LANÇO
   Limite geral _____ Km/h
   Limite local _____ Km/h

B4 PAVIMENTO

1. TIPO DE PISO
   1 ☐ Terra batida
   2 ☐ Betuminoso
   3 ☐ Betão de cimento
   4 ☐ Calçada
2. ESTADO DE CONSERVAÇÃO
   1 ☐ Em bom estado
   2 ☐ Em estado regular
   3 ☐ Em mau estado
3. OBSTÁCULOS OU OBRAS
   1 ☐ Inexistentes
   2 ☐ Não sinalizados
   3 ☐ Insuficientemente sinalizados
   4 ☐ Corretamente sinalizados
4. CONDIÇÕES DE ADERÊNCIA
   1 ☐ Seco e limpo
   2 ☐ Húmido
   3 ☐ Molhado
   4 ☐ Com água acumulada na faixa de rodagem
   5 ☐ Com gelo, geada ou neve
   6 ☐ Com lama
   7 ☐ Com gravilha ou areia
   8 ☐ Com óleo

B5 SINALIZAÇÃO

1. MARCAS NO PAVIMENTO
   1 ☐ Sem marcas rodoviárias ou pouco visíveis
   2 ☐ Com marcas – separadoras de sentido de trânsito
   3 ☐ Com marcas – separadoras de sentido e de vias de trânsito
2. SINALIZAÇÃO LUMINOSA
   1 ☐ Inexistente
   2 ☐ A funcionar normalmente
   3 ☐ Intermitente
   4 ☐ Desligada
3. SINAIS
   1 ☐ Stop
   2 ☐ Cedência de passagem
   3 ☐ Proibição de ultrapassagem
   4 ☐ Passagem de peões
   5 ☐ Outros

B6 LUMINOSIDADE

1 ☐ Em pleno dia
2 ☐ Sol encandeante
3 ☐ Aurora ou crepúsculo
4 ☐ Noite, sem iluminação
5 ☐ Noite, com iluminação

B7 FATORES ATMOSFÉRICOS

1 ☐ Bom tempo
2 ☐ Chuva
3 ☐ Vento forte
4 ☐ Nevoeiro
5 ☐ Neve
6 ☐ Nuvem de fumo
7 ☐ Granizo

C - NATUREZA DO ACIDENTE

DESPISTE
1 ☐ Despiste simples
Com transposição do separador central
2 ☐ Com dispositivo de retenção
3 ☐ Sem dispositivo de retenção
4 ☐ Com transposição do dispositivo de retenção lateral
5 ☐ Com capotamento
6 ☐ Com colisão com veículo imobilizado ou obstáculo
7 ☐ Com fuga

COLISÃO
8 ☐ Frontal
9 ☐ Traseira com outro veículo em movimento
10 ☐ Lateral com outro veículo em movimento
11 ☐ Com veículo ou obstáculo na faixa de rodagem
12 ☐ Choque em cadeia
13 ☐ Com fuga
14 ☐ Outras situações

ATROPELAMENTO
15 ☐ De peões
16 ☐ De animais
17 ☐ Com fuga

Incêndio posterior. A B C ☐ ☐ ☐ A preencher no caso de se verificar

D - VEÍCULOS INTERVENIENTES

D1 CATEGORIA/CLASSE

1. VEÍCULOS A, B e C

	A	B	C
1 Velocípede	☐	☐	☐
2 Velocípede c/motor	☐	☐	☐
3 Ciclomotor	☐	☐	☐
4 Triciclo	☐	☐	☐
5 Motociclo cilindrada ≤ 125cc	☐	☐	☐
6 Motociclo cilindrada > 125cc	☐	☐	☐
7 Automóvel ligeiro	☐	☐	☐
8 Automóvel pesado	☐	☐	☐
9 Veículo agrícola	☐	☐	☐
10 Máquina industrial	☐	☐	☐
11 Veículo sobre carris	☐	☐	☐
12 Veículo de tração animal	☐	☐	☐
13 Quadriciclo	☐	☐	☐
14 Desconhecido	☐	☐	☐

2. Se for automóvel ligeiro ou pesado, indicar o tipo:

	A	B	C
1 Passageiros	☐	☐	☐
2 Mercadorias	☐	☐	☐
3 Misto	☐	☐	☐
4 Trator	☐	☐	☐
5 Veículo especial. Qual?	☐	☐	☐

3. A B C
1 ☐ ☐ ☐ Sem semireboque/reboque
2 ☐ ☐ ☐ Com semireboque/reboque

D2 TIPO DE SERVIÇO
A B C
1 ☐ ☐ ☐ Particular
2 ☐ ☐ ☐ Público

D3 ANO DE MATRÍCULA
A ____ B ____ C ____

D4 INSPEÇÃO PERIÓDICA
A B C
1 ☐ ☐ ☐ Não obrigatória
2 ☐ ☐ ☐ Válida
3 ☐ ☐ ☐ Sem validade

D5 CERTIFICADO ADR
1. Preencher apenas no caso de transporte de mercadorias perigosas
A B C
1 ☐ ☐ ☐ Válido
2 ☐ ☐ ☐ Sem validade
3 ☐ ☐ ☐ Inexistente
2. MATÉRIA/OBJETO PERIGOSO TRANSPORTADO ____

D6 CARGA/LOTAÇÃO/PNEUS
1. CARGA/LOTAÇÃO
A B C
1 ☐ ☐ ☐ Sem carga
2 ☐ ☐ ☐ Com excesso de carga
3 ☐ ☐ ☐ Carga bem acondicionada
4 ☐ ☐ ☐ Carga mal acondicionada
5 ☐ ☐ ☐ Com lotação excedida

2. PNEUS
A B C
1 ☐ ☐ ☐ Sem deficiência
2 ☐ ☐ ☐ Com deficiência

3. TACÓGRAFO
A B C
1 ☐ ☐ ☐ Sem tacógrafo ou desativado
2 ☐ ☐ ☐ Com tacógrafo

D7 SEGURO
A B C
1 ☐ ☐ ☐ Com seguro
2 ☐ ☐ ☐ Sem seguro
3 ☐ ☐ ☐ Isento

E - CONDUTORES INTERVENIENTES

E1 CARACTERÍSTICAS DA HABILITAÇÃO DE CONDUÇÃO
1. LICENÇA/CARTA DE CONDUÇÃO
A B C
1 ☐ ☐ ☐ Com licença/carta adequada ao veículo
2 ☐ ☐ ☐ Com licença/carta não adequada ao veículo
3 ☐ ☐ ☐ Em situação de instrução/exame
4 ☐ ☐ ☐ Caducada/suspensa
5 ☐ ☐ ☐ Sem licença/carta
6 ☐ ☐ ☐ Não necessária ao veículo que conduz

2. PAÍS DE EMISSÃO
A B C
1 ☐ ☐ ☐ Portugal
2 ☐ ☐ ☐ Outro(s) A ____ B ____ C ____

3. ANO DA HABILITAÇÃO
Relativamente ao veículo que conduzia
A ____ B ____ C ____

4. CERTIFICADO ADR
A B C
1 ☐ ☐ ☐ Válido
2 ☐ ☐ ☐ Sem validade
3 ☐ ☐ ☐ Inexistente

E2 CONDIÇÕES PSÍCO/FÍSICAS
1. CONTRÓLO DO NÍVEL DE ALCOOLEMIA
A B C
1 ☐ ☐ ☐ **Submetido ao teste de alcoolemia**
Não submetido por
2 ☐ ☐ ☐ Doença
3 ☐ ☐ ☐ Lesão ou morte decorrente do acidente
4 ☐ ☐ ☐ Condutor não contactado na altura do acidente
5 ☐ ☐ ☐ Fuga
6 ☐ ☐ ☐ Recusa
7 ☐ ☐ ☐ Outra

2. TAXA DE ALCOOLEMIA
A ____,__ B ____,__ C ____,__

3. OUTROS FATORES
A B C
1 ☐ ☐ ☐ Normal
2 ☐ ☐ ☐ Droga por despistagem
3 ☐ ☐ ☐ Sono/sonolência
4 ☐ ☐ ☐ Distração
5 ☐ ☐ ☐ Doença súbita
6 ☐ ☐ ☐ Fadiga

4. TEMPO DE CONDUÇÃO CONTINUADA
A B C
1 ☐ ☐ ☐ Menos de 1 hora
2 ☐ ☐ ☐ De 1 a 3 horas
3 ☐ ☐ ☐ De 3 a 5 horas
4 ☐ ☐ ☐ Mais de 5 horas
5 ☐ ☐ ☐ Ignorada

E3 AÇÕES E MANOBRAS ANTES DO ACIDENTE
1. A B C
1 ☐ ☐ ☐ Início de marcha
2 ☐ ☐ ☐ Saída de parqueamento ou rua particular
3 ☐ ☐ ☐ Em marcha normal
4 ☐ ☐ ☐ Ultrapassagem pela esquerda
5 ☐ ☐ ☐ Ultrapassagem pela direita
6 ☐ ☐ ☐ Mudança de direção para a esquerda
7 ☐ ☐ ☐ Mudança de direção para a direita
8 ☐ ☐ ☐ Marcha atrás
9 ☐ ☐ ☐ Circulação em sentido oposto ao estabelecido
10 ☐ ☐ ☐ Travagem brusca
11 ☐ ☐ ☐ Parado ou estacionado
12 ☐ ☐ ☐ Inversão do sentido de marcha
13 ☐ ☐ ☐ Trânsito em filas paralelas
14 ☐ ☐ ☐ Mudança de via de trânsito para a esquerda
15 ☐ ☐ ☐ Mudança de via de trânsito para a direita
16 ☐ ☐ ☐ Desvio brusco/saída de fila de trânsito
17 ☐ ☐ ☐ Atravessando a via

2. ESQUEMA ____ (Ver esquema em anexo)

E4 INFORMAÇÃO COMPLEMENTAR A AÇÕES E MANOBRAS
A B C
1 ☐ ☐ ☐ Desrespeito da sinalização vertical
2 ☐ ☐ ☐ Desrespeito das marcas rodoviárias
3 ☐ ☐ ☐ Desrespeito da sinalização semafórica
4 ☐ ☐ ☐ Manobra irregular
5 ☐ ☐ ☐ Velocidade excessiva para as condições existentes
6 ☐ ☐ ☐ Não sinalização da manobra
7 ☐ ☐ ☐ Desrespeito das distâncias de segurança
8 ☐ ☐ ☐ Circulação afastada da berma ou passeio
9 ☐ ☐ ☐ Rebentamento pneumático
10 ☐ ☐ ☐ Queda de carga ou objeto
11 ☐ ☐ ☐ Falha mecânica do veículo
12 ☐ ☐ ☐ Ausência de luzes quando obrigatórias
13 ☐ ☐ ☐ Obstáculo imprevisto na faixa de rodagem
14 ☐ ☐ ☐ Abertura de porta
15 ☐ ☐ ☐ Encandeamento
16 ☐ ☐ ☐ Não identificada

E5 ACESSÓRIOS DE SEGURANÇA
A B C
1 ☐ ☐ ☐ Capacete
2 ☐ ☐ ☐ Cinto de segurança
3 ☐ ☐ ☐ Sem uso de cinto/capacete
4 ☐ ☐ ☐ Isento

F - CONSEQUÊNCIAS DO ACIDENTE

F1 CONDUTORES VÍTIMAS
1. GRAU DE GRAVIDADE DAS LESÕES
A B C
1 ☐ ☐ ☐ Morto
2 ☐ ☐ ☐ Ferido grave
3 ☐ ☐ ☐ Ferido leve

F2 PASSAGEIROS VÍTIMAS
Veículo A Veículo B Veículo C

1. SEXO

	a b c d	i j l m	r s t u	
1	☐☐☐☐	☐☐☐☐	☐☐☐☐	Masculino
2	☐☐☐☐	☐☐☐☐	☐☐☐☐	Feminino

2. IDADE

a b	i j	r s
c d	l m	t u

3. POSIÇÃO NO VEÍCULO

	a b c d	i j l m	r s t u	
1	☐☐☐☐	☐☐☐☐	☐☐☐☐	À frente
2	☐☐☐☐	☐☐☐☐	☐☐☐☐	À retaguarda
3	☐☐☐☐	☐☐☐☐	☐☐☐☐	Desconhecido

4. USO DE ACESSÓRIOS DE SEGURANÇA

	a b c d	i j l m	r s t u	
1	☐☐☐☐	☐☐☐☐	☐☐☐☐	C/ capacete/cinto segurança
2	☐☐☐☐	☐☐☐☐	☐☐☐☐	C/ sistema retenção de crianças
3	☐☐☐☐	☐☐☐☐	☐☐☐☐	S/ uso capacete/cinto segurança
4	☐☐☐☐	☐☐☐☐	☐☐☐☐	S/ sistema retenção de crianças

5. GRAU DE GRAVIDADE DAS LESÕES

	a b c d	i j l m	r s t u	
1	☐☐☐☐	☐☐☐☐	☐☐☐☐	Morto
2	☐☐☐☐	☐☐☐☐	☐☐☐☐	Ferido grave
3	☐☐☐☐	☐☐☐☐	☐☐☐☐	Ferido leve
4	☐☐☐☐	☐☐☐☐	☐☐☐☐	Ileso

F3 PEÕES VÍTIMAS
1. SEXO
a b c d
1 ☐☐☐☐ Masculino
2 ☐☐☐☐ Feminino

2.
a b c d
1 ☐☐☐☐ Peão isolado
2 ☐☐☐☐ Peões em grupo
3 ☐☐☐☐ Conduzindo à mão velocípedes, carros de crianças ou de deficientes físicos
4 ☐☐☐☐ Deslocando-se sobre patins, trotinetes ou outros

3. IDADE
a ____ b ____ c ____ d ____

4. CONDIÇÕES PSÍCO-FÍSICAS
a b c d
1 ☐☐☐☐ Sem restrições
2 ☐☐☐☐ Com visão deficiente
3 ☐☐☐☐ Com audição deficiente
4 ☐☐☐☐ Com deficiência motora
Influenciada pelo álcool
a ____,__ b ____,__ c ____,__ d ____,__
5

5. AÇÕES
a b c d
1 ☐☐☐☐ A sair ou entrar num veículo
2 ☐☐☐☐ Surgindo inesperadamente na faixa de rodagem de trás de um obstáculo
3 ☐☐☐☐ Em plena faixa de rodagem
4 ☐☐☐☐ Em trabalhos na via
5 ☐☐☐☐ Atravessando fora da passagem de peões, a menos de 50 m de uma passagem
6 ☐☐☐☐ Atravessando fora da passagem de peões a mais de 50 m de uma passagem ou quando não exista passagem
7 ☐☐☐☐ Atravessando em passagem sinalizada
8 ☐☐☐☐ Atravessando em passagem sinalizada com desrespeito da sinalização semafórica
9 ☐☐☐☐ Em ilhéu ou refúgio na via
10 ☐☐☐☐ Transitando pela direita da faixa de rodagem
11 ☐☐☐☐ Transitando pela esquerda da faixa de rodagem
12 ☐☐☐☐ Transitando pela berma ou passeio

6. UTILIZAÇÃO DE MATERIAL REFLETOR
a b c d
1 ☐☐☐☐ Sim
2 ☐☐☐☐ Não

7. GRAVIDADE DAS LESÕES
a b c d
1 ☐☐☐☐ Morto
2 ☐☐☐☐ Ferido grave
3 ☐☐☐☐ Ferido leve

DATA ___/___/___

Número de boletins utilizados neste acidente ____

Nome _____

(Posto) _____

Bibliography

[1] ANSR, "BEAV". http://www.ansr.pt/Estatisticas/BEAV/Documents/BEAV.pdf. Accessed: 2021-11-15.

[2] ANSR, "Manual de Preenchimento". http://www.ansr.pt/Estatisticas/BEAV/Documents/MANUALPREENCHIMENTOBEAV.pdf. Accessed: 2021-11-15.

[3] ANSR, "Relatório anual" 2016, Setúbal. http://www.ansr.pt/Estatisticas/RelatoriosDeSinistralidade/Pages/default.aspx. Accessed: 2021-11-15.

[4] Apache Zeppelin. https://zeppelin.apache.org. Accessed: 2021-11-29.

[5] Hitachi Vantara Lumada and Pentaho Documentation. https://help.hitachivantara.com/Documentation/Pentaho/8.2/Products/Data_Integration. Accessed: 2021-11-29.

[6] Moprevis. https://moprevis.uevora.pt/en/. Accessed: 2021-08-02.

[7] Road safety facts. =https://www.asirt.org/safe-travel/road-safety-facts/, Apr 2020.

[8] T. Anderson. Kernel density estimation and k-means clustering to profile road accident hotspots. *Accident; analysis and prevention*, 41:359–64, 06 2009.

[9] K. Assi, S. M. Rahman, U. Mansoor, and N. Ratrout. Predicting crash injury severity with machine learning algorithm synergized with clustering technique: A promising protocol. *International Journal of Environmental Research and Public Health*, 17:5497, 07 2020.

[10] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 10 2001.

[11] C. Caliendo, M. Guida, and A. Parisi. A crash-prediction model for multilane roads. *Accident; analysis and prevention*, 39:657–70, 08 2007.

[12] F. Chang, P. Xu, H. Zhou, A. H. Chan, and H. Huang. Investigating injury severities of motorcycle riders: A two-step method integrating latent class cluster analysis and random parameters logit model. *Accident Analysis & Prevention*, 131:316–326, 2019.

[13] L.-Y. Chang and W.-C. Chen. Data mining of tree-based models to analyze freeway accident frequency. *Journal of safety research*, 36:365–75, 02 2005.

[14] F. Chen, M. Song, and X. Ma. Investigation on the injury severity of drivers in rear-end collisions between cars using a random parameters bivariate ordered probit model. *International Journal of Environmental Research and Public Health*, 16(14), 2019.

[15] R. Elvik. State-of-the-art approaches to road accident black spot management and safety analysis of road networks. *Transportøkonomisk institutt: Oslo, Norway*, 2007.

[16] R. Elvik. A survey of operational definitions of hazardous road locations in some european countries. *Accident Analysis & Prevention*, 40(6):1830–1835, 2008.

[17] N. Fiorentini and M. Losa. Handling imbalanced data in road crash severity prediction by machine learning algorithms. *Infrastructures*, 5(7), 2020.

[18] C. Gutierrez-Osorio and C. Pedraza. Modern data sources and techniques for analysis and forecast of road accidents: A review. *Journal of Traffic and Transportation Engineering (English Edition)*, 7(4):432–446, 2020.

[19] T. K. Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1, 1995.

[20] A. Hébert, T. Guédon, T. Glatard, and B. Jaumard. High-resolution road vehicle collision prediction for the city of montreal. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 1804–1813, 2019.

[21] A. Iranitalab and A. Khattak. Comparison of four statistical and machine learning methods for crash severity prediction. *Accident; analysis and prevention*, 108:27–36, 08 2017.

[22] S. Kotsiantis, I. Zaharakis, and P. Pintelas. Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review*, 26:159–190, 11 2006.

[23] S. Kumar and D. Toshniwal. A data mining approach to characterize road accident locations. *Journal of Modern Transportation*, 24, 02 2016.

[24] L. Lin, Q. Wang, and A. W. Sadek. A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. *Transportation Research Part C: Emerging Technologies*, 55:444–459, 2015. Engineering and Applied Sciences Optimization (OPT-i) - Professor Matthew G. Karlaftis Memorial Issue.

[25] D. Lord and F. Mannering. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, 44(5):291–305, 2010.

[26] A. Montella. A comparative analysis of hotspot identification methods. *Accident; analysis and prevention*, 42:571–81, 03 2010.

[27] A. Najjar, S. Kaneko, and Y. Miyanaga. Combining satellite imagery and open data to map road safety. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb. 2017.

[28] H. Ren, Y. Song, J. Wang, Y. Hu, and J. Lei. A deep learning approach to the citywide traffic accident risk prediction. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3346–3351, 2018.

[29] S. Roshandel, Z. Zheng, and S. Washington. Impact of real-time traffic characteristics on freeway crash occurrence: Systematic review and meta-analysis. *Accident; analysis and prevention*, 79, 03 2015.

[30] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[31] I. Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2, 05 2021.

[32] P. Savolainen, F. Mannering, D. Lord, and M. Quddus. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident; analysis and prevention*, 43:1666–76, 09 2011.

[33] A. P. Shah, J.-B. Lamare, T. Nguyen-Anh, and A. Hauptmann. Cadp: A novel dataset for cctv traffic camera based accident analysis. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–9, 2018.

[34] Z. S. Siam, R. T. Hasan, S. S. Anik, A. Dev, S. I. Alita, M. Rahaman, and R. M. Rahman. Study of machine learning techniques on accident data. In M. Hernes, K. Wojtkiewicz, and E. Szczerbicki, editors, *Advances in Computational Collective Intelligence*, pages 25–37, Cham, 2020. Springer International Publishing.

[35] P. Silva, M. Andrade, and S. Ferreira. Machine learning applied to road safety modeling: A systematic literature review. *Journal of Traffic and Transportation Engineering (English Edition)*, 7:775–790, 12 2020.

[36] D. Sisodia, L. Singh, and S. Sisodia. Clustering techniques: A brief survey of different clustering algorithms. 2012.

[37] R. Slikboer, S. Muir, S. Silva, and D. Meyer. A systematic review of statistical models and outcomes of predicting fatal and serious injury crashes from driver crash and offense history data. *Systematic Reviews*, 9, 09 2020.

[38] S. Szenasi and P. Csiba. Clustering algorithm in order to find accident black spots identified by gps coordinates. In *In Proceedings of the 14th GeoConference on Informatics, Geoinformatics, and Remote Sensing, Ilza, Poland, 19–25*, volume 1, 06 2014.

[39] L. Thakali, T. J. Kwon, and L. Fu. Identification of crash hotspots using kernel density estimation and kriging methods – a comparison. *Journal of Modern Transportation*, 23, 03 2015.

[40] A. Theofilatos. Incorporating real-time traffic and weather data to explore road accident likelihood and severity in urban arterials. *Journal of Safety Research*, 61, 03 2017.

[41] A. Theofilatos, D. Graham, and G. Yannis. Factors affecting accident severity inside and outside urban areas in greece. *Traffic Injury Prevention*, 13(5):458–467, 2012.

[42] A. A. Theofilatos, C. Chen, and C. Antoniou. Comparing machine learning and deep learning methods for real-time crash prediction. *Transportation Research Record: Journal of the Transportation Research Board*, 2673:036119811984157, 04 2019.

[43] A. A. Theofilatos and G. Yannis. Investigation of powered-two-wheeler accident involvement in urban arterials by considering real-time traffic and weather data. *Traffic injury prevention*, 18, 06 2016.

[44] C. Xu, W. Wang, and P. Liu. Identifying crash-prone traffic conditions under different weather on freeways. *Journal of safety research*, 46:135–44, 09 2013.

[45] R. Yu, Y. Xiong, and M. Abdel-Aty. A correlated random parameter approach to investigate the effects of weather conditions on crash risk for a mountainous freeway. *Transportation Research Part C: Emerging Technologies*, 50, 11 2014.

[46] Z. Yuan, X. Zhou, and T. Yang. Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 984–992, New York, NY, USA, 2018. Association for Computing Machinery.

[47] Z. Yuan, X. Zhou, T. Yang, and J. Tamerius. Predicting traffic accidents through heterogeneous urban data : A case study. In *In Proceedings of the 6th international workshop on urban computing (UrbComp 2017), Halifax, NS, Canada, 13–17 August*, 2017.

[48] Q. Zeng, H. Huang, X. Pei, and S. Wong. Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks. *Analytic Methods in Accident Research*, 10:12–25, 2016.

Contacts:
Universidade de Évora
Escola de Ciências e Tecnologia
Colégio Luís António Verney, Rua Romão Ramalho, nº59
7000 - 671 Évora | Portugal
Tel: (+351) 266 745 371
email: geral@ect.uevora.pt