LUISS

Degree Program in Management And Computer Science

Course of Data Analysis for Business

Road Crashes Severity Prediction: an Analysis of Risk Factors in the Case Study of Rome

Prof. Iafrate Francesco

SUPERVISOR

Cervellini Mattia ID 255221

CANDIDATE

Academic Year 2022/2023

ABSTRACT

Road crashes are one of the deadliest plagues of contemporary times as they cause approximately 1.3 million deaths each year worldwide with enormous associated costs. Both the World Health Organization and the United Nations have often demonstrated their profound concern on the matter.

This paper presents a complete machine learning study of road accidents that took place on urban roads belonging to the province of Rome (Italy) during 2021. The analysis is structured as a binary classification task trying to differentiate between accidents presenting casualties (either injuries or deaths) and crashes with no casualties. The data for the study was retrieved directly from the official website of the municipality of Rome and included more than 20,000 distinct crashes. An extensive preprocessing was applied to the data to make them usable for classification purposes in order to be able to point out the most important risk factors in crashes.

The most relevant characteristics when predicting the severity of an accident were highlighted as being: the presence of two-wheeled vehicles in the crash, the involvement of pedestrians and the specific type of accident. The best model for predicting the severity of a crash according to the *AUC* statistic was an eXtreme Gradient Boosting Tree model with one hot encoding for categorical variables. The results of this study were insightful and could be the base for further in depth multinomial analysis in order to provide benefits to all road users and companies heavily involved in the matter (e.g., insurance companies, car manufacturers).

TABLE OF CONTENTS

1 INTRODUCTION

1.1 Road Accidents Worldwide

Road accidents cause approximately 1.3 million deaths each year and are the most common cause of death among people aged between 5 and 29 years, with a cost of 3% of the Gross Domestic Product of a country in most cases. The World Health Organization (WHO) states that the countries with a low to middle income are the stage for more than 90% of the deaths in traffic accidents and, in general, people from a lower economic status are more likely to be involved in a road crash. Besides that, it is also shown that the majority of traffic deaths are among vulnerable road users like pedestrians, cyclists and motorcyclists (WHO, 2022). For all these reasons combined, the United Nations General Assembly has set the target of halving the number of road casualties by 2030 requesting the full collaboration of the World Health Organization and Regional Commission of the United Nations (UnitedNations, 2020).

1.2 Road Accidents in Italy

The situation in Italy is in line with the global scenario regarding road crashes with ISTAT (National Institute for Statistics) publishing yearly reports on the topic. In 2021 ISTAT results reported 2,875 road fatalities (of which 2,397 died within 24 hours from the accident and 478 in the following 2 to 30 days), 204,728 injured people and 151,875 total road accidents, with a percentage increase compared to 2020 of 20.0%, 28.6% and 28.4% respectively. The numbers still show a strong decrease when compared with the year 2019 with -9.4% victims, -15.2% injured and -11.8% accidents overall. One of the few aspects that goes against this decreasing trend is the case of accidents involving electric scooters which have seen a rapid ascent in popularity (ISTAT, 2022). These results highlight the big role of the pandemic lockdowns in artificially changing the data associated with road traffic, making them look far better than they really are. In 2022, the first truly post-pandemic year, on the other hand, ISTAT presents an almost complete return to pre-pandemic levels of deaths, injuries and accidents, with a respective decrease compared to 2019 of -0.4%, -7.4% and -3.7% (ISTAT,

2023). These findings confirm a very weak positive trend over time which was inflated by the influence of COVID-19 restrictive measures on the amount of traffic.

1.3 Predicting Crash Severity

Reducing the amount of road accidents appears to be a titanic undertaking, for this reason in later years a parallel approach has been explored. Instead of trying to limit the number of crashes with its immense complexity, the new procedure aims at analysing all the available data from road accidents to predict their severity. For instance, it is possible to identify which factor is the most influential in predicting whether someone got injured or died in a car accident. Various machine and statistical learning approaches have been applied to data gathered all over the world with results varying slightly depending on the geographical and cultural area analysed. A common issue for this type of studies is the lack of a proper data management system or the scarcity of adequate data in the first place, which could hinder the validity of the researches: this was highlighted for some south-eastern European countries (Laiou et al., 2017).

1.4 Benefits for Society

A definitive breakthrough in accident severity prediction could benefit all the stakeholders involved. First, it would benefit hospitals and emergency services helping them in optimizing their resources to provide adequate medical treatments to the most urgent patients in a more efficient and rapid way. Secondly, it would benefit transportation and road planners in optimizing their efforts to build an infrastructure that is conductive to a reduction of high-risk accidents performing a cost benefit analysis on the expected severity of future accidents. Lastly, it would benefit insurance companies to better predict customer associated risks and fine tune premiums thanks to a more comprehensive economic analysis of accidents which is strongly dependent on the severity of the accident (Iranitalab & Khattak, 2017). All these consequences combined would have a great impact for road users which would be part of safer infrastructure with better access to emergency and insurance services in the case of any

unforeseen motor accident.

1.5 Why Rome

Rome is the capital city of Italy and its most populous city with over 2.7 million inhabitants, with a comprehensive area of 1,208 square kilometres. The city is also widely known for its impressive road system which includes around 8,000 kilometres of roads that develop around the historical city centre in almost every possible direction. Rome is also famous around the world for its dramatic traffic congestions that can cause road users delays of hours when trying to get from a place to another in the city.

According to the 2022 INRIX Global Traffic Scorecard (Figure 1) Rome is the 13th city worldwide when considering the number of hours lost in traffic congestions during peak hours with 107 hours lost for each citizen and an average city centre traffic speed of 20.92 kilometres per hour (INRIX, 2022). Given all these, Rome seemed the perfect subject for this study also thanks to its very diverse environment and heterogeneity of road types.

2022 Impact Rank (2021 Rank)	Urban Area	Country	2022 Delay per Driver (hours)	Change from 2021	Change from Pre-COVID	Downtown Speed (mph)	Change in Downtown Speed
1 (1)	London	UK	156	5%	5%	10	-9%
2 (6)	Chicago IL	USA	155	49%	7%	11	-27%
3 (2)	Paris	FRA	138	-1%	-16%	11	0%
4 (18)	Boston MA	USA	134	72%	-10%	11	-27%
5 (5)	New York City NY	USA	117	15%	-16%	11	-15%
6 (8)	Bogota	COL	122	30%	-36%	11	-15%
7 (22)	Toronto ON	CAN	118	59%	-13%	10	-29%
8 (13)	Philadelphia PA	USA	114	27%	-20%	11	-15%
9 (32)	Miami FL	USA	105	59%	30%	15	-21%
10 (9)	Palermo	ITA	121	11%	-12%	9	0%
11 (36)	Monterrey	MEX	116	66%	108%	19	-17%
12 (16)	Dublin	IRL	114	28%	-26%	12	-8%
13 (7)	Rome	ITA	107	0%	-36%	13	-7%

Figure 1: Global Traffic Ranking (2022). Source: INRIX, 2022.

1.6 Purpose of This Paper

This paper sets out to perform a statistics and machine learning based analysis on car accidents located in the municipality of Rome thanks to the data published on the site: https://dati.comune.roma.it/. Figure 2 shows a depiction obtained from the data of all the traffic accidents included in the paper, represented by red dots, superimposed with a map of the main roads of Rome.

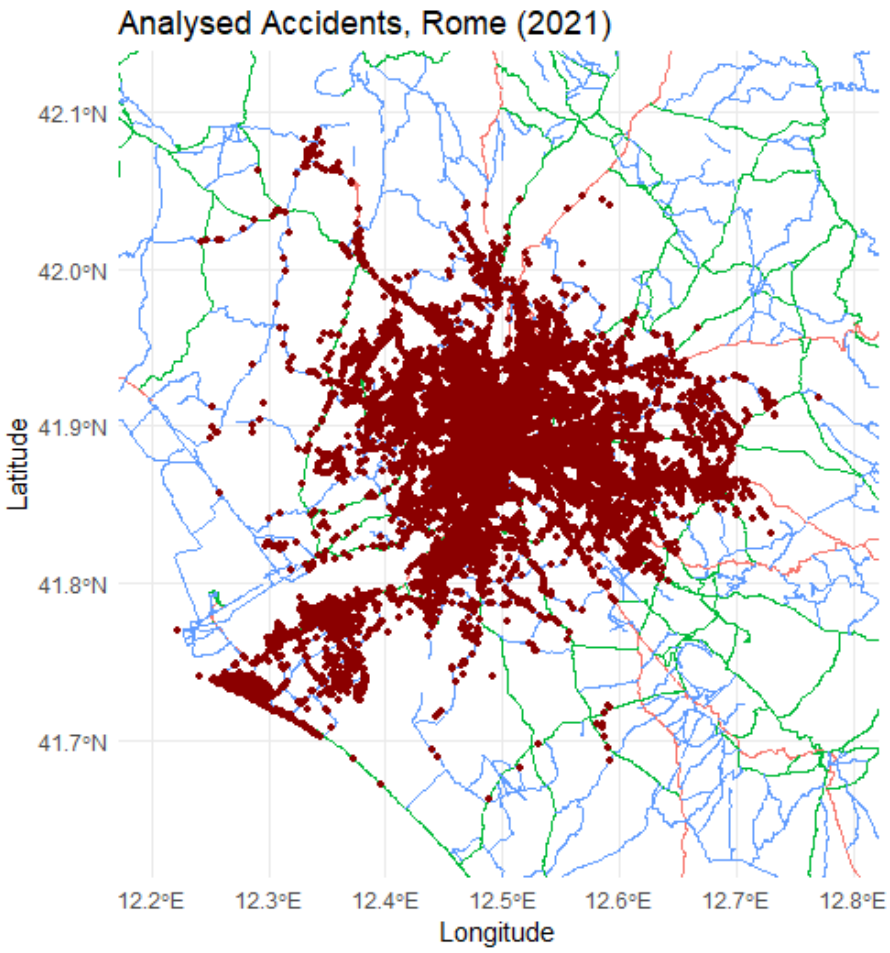

Figure 2: Traffic accidents in Rome (2021). Source: Self-elaboration.

These datasets include all the accidents in which any kind of law enforcement authority was involved except for the crashes that happened on the highway A90, also known as *Grande Raccordo Anulare*.

The final aim of this research is to give a first insight in understanding which are the most important factors in predicting the advent of any kind of casualty in crashes that happen in the city of Rome to gain a better understanding of the dynamics involved and possibly set out future research with more comprehensive data interlacing also with databases from hospitals and clinics. The data used for this modelling experiment regards accidents happened through-

out 2021, since the data for 2022 is to this day incomplete because of the maintenance of the digital infrastructure of the local police. The proposed models will use as predictors features gathered by the local authority on its arrival on the accident site to classify the accidents based on the occurrence or the lack of any kind of casualty (either deaths or injuries).

2 LITERATURE REVIEW

2.1 Variety of Studies

The field of severity prediction in car accidents has become more and more popular in recent years with studies popping up from all over the world analysing almost every possible different aspect of the matter. For this reason, the studies are very different from one another. Some of them explore one specific type of road accidents while others analyse a predetermined geographical area. The various studies also differ in the way the analysis is carried out: in some cases, a regression like approach is applied, while some other researchers have preferred a classification task depending on the specific aim of the paper and on the amount and type of data that were analysed.

This chapter will be focussed on a systematic review of the scientific literature concerning the prediction and evaluation of the severity of car crashes. The goal is to get some insights that could prove valuable for the analysis in this study.

2.2 Studies by Data Subjects

This section will focus on the review of various studies concerning specific characteristics of accidents or precise geographical areas, analysing in depth the kind of data that was subjected to scrutiny and the sources of such data. A concise synthesis of the results will be given alongside an overview of the learning methods used. A more in-depth analysis of common methodologies used for crash severity prediction will be given in section 2.3.

2.2.1 Crash-specific Analysis

Al Mamlook et al., 2020, explores in depth accidents in which elderly people were involved to grasp which are the most influential factors in predicting a severe injury for people over the age of 60. They used a database retrieved from the Office of Highway Safety Planning in Michigan containing over 100,000 crashes involving elderly drivers. They used a Synthetic Minority Oversampling Technique to balance the classes in their dataset and compared vari-

ous classification models: Logistic Regression, Decision Trees, Light-GBM, Random Forest and Naive Bayes Algorithm. The best model (Light-GBM) achieved outstanding results having an accuracy of around 87%. Figure 3 highlights the most relevant predictors according to the Light-GBM model: the age of the driver, the volume of traffic and the age of the car.

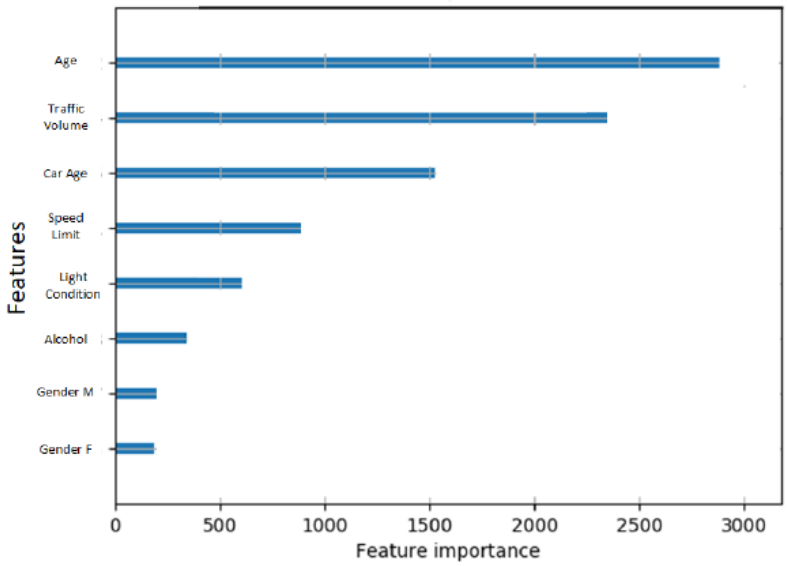

Figure 3: Feature importance according to Light-GBM. Source: Al Mamlook et al., 2020.

C. Lin et al., 2020, applies Random Forests and XGBoost models using both label encoding and one hot encoding to explore which factors contribute the most to the severity of an accident in which a driver aged between 15 and 20 years old is involved. They used a dataset of approximately 9,000 crashes with young drivers involved which took place in Texas. The results highlighted how the most important predictors of severity for young drivers are: unsafe speed, failure to control speed and failure to yield right of way. Another interesting finding pointed out in Figure 4 is that algorithms trained using one hot encoding perform better than the ones that use a label encoding.

Encoding Method	Predictive Algorithm	Mean Absolute Error (MAE)
One hot encoder	Random Forest	0.7271
One hot encoder	XGBoost	0.7140
Label encoder	Random Forest	1.0634
Label encoder	XGBoost	0.7804

Figure 4: Mean Absolutes Errors of the models. Source: C. Lin et al., 2020

Another interesting study for a particular category of accidents is Vajari et al., 2020, which analysed over 7,500 road crashes involving motorcycles at intersections in the state of Victoria, in Australia. A Multinomial Logit Model was used to perform a multilevel classification task dividing accidents in minor injuries, major injuries and fatal injuries. The most relevant predictors of serious injuries were the age of the motorcyclist, the characteristics of the intersection and the timing of the accident which reveals a particular relevance of crashes taking place during morning rush hours and weekends. Another interesting finding of the paper is that "it was observed that motorcycle crashes are about 3 times (RRR = 1/0.37) more likely to result in fatal injuries at mid-night/early morning as compared to serious injuries. Similarly, the results showed that the crashes occurring in morning rush hours are 1.37 times (RRR = 1/0.73) and 3.3 (RRR = 1/0.30) more likely to result in fatal injuries as compared to minor injuries and serious injuries respectively" (Vajari et al., 2020), where RRR stands for Relative Risk Ratio which measures the influence of all independent predictors on the response variable. M.-R. Lin and Kraus, 2008, also states how motorcycle users have a fatality rate 30 times higher than car users.

The severity for accidents involving heavy trucks was analysed in Chang and Chien, 2013, using a Classification and Regression Tree (CART) to build empirical connections between crash factors and accident severity. A database of 1,620 accidents which have taken place in the freeway system of Taiwan was analysed pointing out how drinking and driving, lack of seatbelt use, the number of vehicles involved were the most important contributors to severe or deadly injuries. Another study on the same matter is Hosseinzadeh et al., 2021,

which analysed 8,390 crashes differentiating them into three types of accidents depending on the number of vehicles involved and whether the truck driver was at fault or not. Using both Support Vector Machine and Random Parameter Binary Logit models trained on the three different categories of accidents (namely Multi-vehicle truck-involved crashes where the truck driver is at fault, Multi-vehicle truck-involved crashes where the truck driver is not at fault and Single-vehicle truck crashes), crashes were classified between fatal and non-fatal. According to their conclusions, "Results reveal the differences between the models and highlight the necessity of at-fault party classification" which poses the significant question of how different types of accidents may very well have different risk factors and may require a specialized approach.

Light Trucks rollovers were studied in Pillajo-Quijia et al., 2020, analysing data coming from the combination of a database of crashes collected by the "Dirección General de Tráfico" and a database of vehicle registrations to identify light trucks and access broader information about them. The vehicles were subsequently classified in four categories depending on their weight and a final dataset of approximately 9,000 accidents involving light trucks implicated in rollover (RO) or run-of-runway (ROR) types of accidents. The chosen methodology was to apply both RF and CART algorithms subsequently, exploiting the great predicting performance of the RF to highlight the significant variables and the interpretability of CART in order to clearly present results. Figure 5 presents the CART resulting from the study for rollover crashes and highlights the use of seatbelt as the most important predictor alongside physical and psychological conditions of the driver.

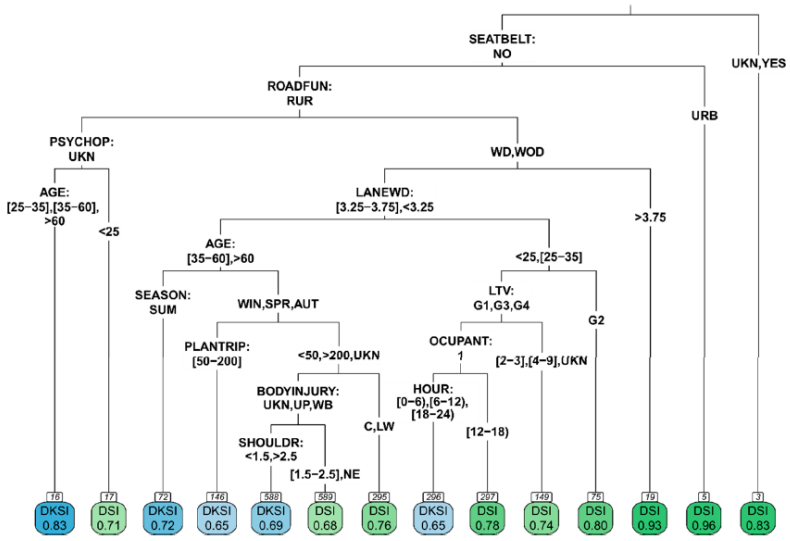

Figure 5: Classification tree for RO collisions. Source: Pillajo-Quijia et al., 2020

Pedestrians may also be part of road accidents, and the consequences for them are often very damaging since they are directly exposed to the forces involved in the impact (Mokhtarimousavi et al., 2020). Zajac and Ivan, 2003, analysed the cases of crashes involving motor vehicles and crossing pedestrians in rural Connecticut since a greater percentage of countryside accidents involving pedestrians results in fatalities. The study showed how the age of the pedestrian as well as the type of vehicle involved and the general presence of alcohol in the accident significantly impacted the severity of injuries suffered by its unwilling participants.

2.2.2 Geographical Analysis

Other studies on the matter are focussed on the analysis of the accidents in a particular geographical location without concentrating on any specific dynamic involved.

Chang and Wang, 2006, analysed a corpus of 12,604 accidents located in Taipei in 2001, provided by the Taiwan Ministry of Transportation and Communications. The data available was very diverse and concerned all aspects of the crash, from the age of the driver to the type of collision and the circumstances of the event. A preliminary data selection was applied to remove infrequent observations (e.g., snowy weather). A CART was used to avoid any pre-

defined relationships between predictors and the target variable which could lead to biased results. The results of the study validate many of the results achieved in crash-specific studies presenting the type of vehicles involved as the most important and relevant predictor.

Lee et al., 2019, conducted a similar study regarding the city of Seoul with the particularity of merging three different datasets to get more comprehensive results: one dataset on road geometry, one dataset on accidents and the last one on weather. Decision Trees (DT), Artificial Neural Networks (ANN) and Random Forests (RF) were applied to the data to validate the reliability of machine learning based models for this kind of tasks. Figure 6 shows how the results of the various models were in line with one another producing similar predictions. The RF also highlighted the weather conditions and the road geometry as the two most influential predictors.

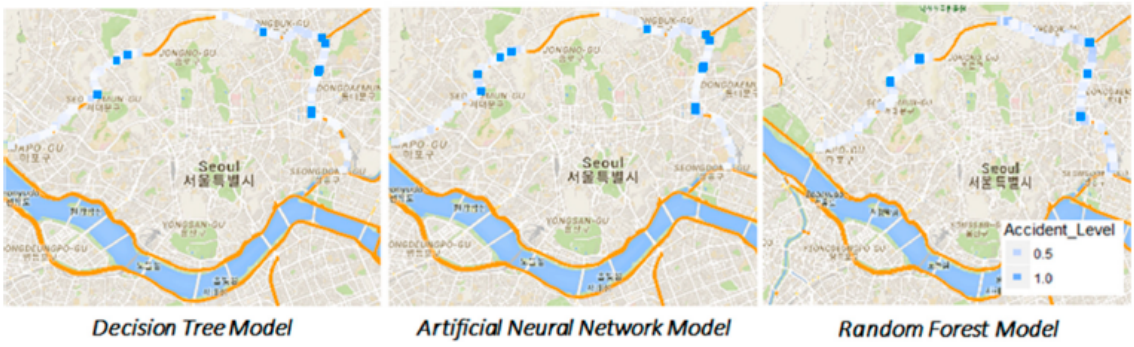

Figure 6: Predicted Accident Levels for Each Model. Source: Lee et al., 2019

AlMamlook et al., 2019, based their research on a database of 271,563 traffic accidents which took place on freeways in Michigan using data provided by the Office of Highway Safety Planning of Michigan. The study was conducted under the KABCO framework using SMOTE to create synthetic minority class observations. The classification methods used were Logistic Regression (LR), RF, Naïve Bayes (NB) and AdaBoost, with RF achieving the best performance showing an Area Under the Curve (AUC) of 0.755 demonstrating a satisfactory pattern recognition and prediction ability.

Another study was conducted by Evwiekpaefe and Umar, 2021, using the Waikato En-

vironment for Knowledge Analysis (WEKA) on the accidents in Kaduna metropolis. Applying DT, K-Nearest Neighbours (KNN), J-Repeated Incremental Pruning (JRIP), NB and Multi-layer Perceptron (MLP), KNN classifier demonstrates an outstanding superiority in performance in this task.

2.3 Methodologies Applied in Previous Studies

This section will review the various techniques utilized for severity prediction. There are various approaches to the task and the choice is influenced not only by the specific aim of the researchers but also by type of data available for each single study. A common ground for all studies is represented by categorical data representing most of the predictor variables with very few datasets presenting a relevant number of continuous variables. As it has been shown before, this outstanding prevalence has pushed many researchers to apply models that are known to work best with categorical data (e.g., RF and NB).

2.3.1 The KBACO Framework

This kind of task possesses an implicit categorical independent variable since it is impossible to evaluate the severity of an accident on a true continuous and objective scale because of the immense variety of possible injuries and future complications. A common practice among law enforcement users to evaluate the severity of an accident and partially circumvent this problem is to use the KABCO scale. This scale has been developed by the National Safety Council (NSC) in 1966 (Burdett et al., 2015) and exploits the intuitive ordinal properties of an accident to create a framework in which each crash can be properly classified. The possible levels on the KABCO scale are:

- K: Fatal Injury
- A: Incapacitating Injury
- B: Non-incapacitating Injury
- C: Possible Injury

- O: Property-damage-only

Recent research in US pointed out how the KABCO is not ideal in determining injury severity (Tarko et al., 2010). Tsui et al., 2009, found out also how only a very small percentage of the serious injuries reported by the Honk Hong police were confirmed as serious in the hospital. This inefficiency could be due to overestimation or underestimation of injuries by officers because of the lack of proper medical training. Burdett et al., 2015, states that "Overestimations of KABCO assessments of injury severity were found when injuries with a significant amount of bleeding were present. Conversely, underestimations of KABCO injury severity ratings were found when occult (not initially apparent) injuries were incurred by the crash victim", which is a very plausible explanation of the phenomena. Despite this flaws KABCO is still regarded as a satisfactory tool for differentiating severe crashes from modest crashes.

2.3.2 The AIS-90 Framework

The Abbreviated Injury Scale (AIS) describes over 2000 injuries in 9 different body areas, and it is to this day the most used medical injury scale in the world, widely adopted in Europe, North America, Japan, Australia, and New Zeeland. The scale works by assigning each injury a numerical score from 1, indicating a minor injury, to 6, indicating a potentially fatal injury. When multiple lesions are present the Maximum AIS (MAIS) approach is applied, synthetizing the overall condition of a patient with his most deadly trauma. For example, if a patient has three injuries with an AIS score 1, 4 and 5 respectively, his comprehensive MAIS score would be 5 (Stevenson et al., 2001). From past research it was established that MAIS classification of injuries is unquestionably better than KABCO classification (Tarko et al., 2010). Despite this, the KABCO classification is more prominent since MAIS data is often available only if there is some kind of intertwining between crash datasets and medical datasets which is usually not the case.

2.3.3 Binomial Classification

Binomial classification is a classic data analysis task based on supervised learning. An algorithm gets fed with a large number of observations which possess some kind of identifier that associates them with a predefined class. In a more formal fashion, there is a series of data points with the form (x, y) where $x = x_1, x_2, \ldots, x_k$ represents all of the dependent variables, also known as predictors, and y is equal to one among two possible classes. The aim of the classification algorithm is to approximate a function $f(x) = y$ which, given x without any information on y, can make predictions on previously unseen observations MIT, 2018. In practice, the algorithm starts with the training phase in which it analyses the characteristics of the various pre-labelled datapoints with the aim of discovering trends and relationships among the data. The algorithm is then used to make predictions on some new and unlabelled data: these predictions are then compared with the actual labels of the new data and the performance of the algorithm is extrapolated via some specialized metrics like the Area Under the Curve (AUC).

The choice of this binary approach in crash severity prediction is often determined by the lack of specific enough data available regarding the consequences of the accident; it may also be chosen because of its simplicity and intuitiveness allowing also for a wider variety of models to be applied. An example of this method is Hosseinzadeh et al., 2021, in which the response variable was considered as a simple indicator of fatality of the accident, where Y=1 corresponded to a fatal accident and Y=0 to a non-fatal accident. Also Beshah et al., 2013, proposed a similar approach using the presence of injuries as response variable, training its algorithms using Y=1 in the presence of post-crash injuries and Y=0 if they were absent.

2.3.4 Multinomial Classification

Multinomial classification, also known as multiclass classification, is a variant of classification that allows for the presence of more than two different classes in the response variable. This kind of tasks can be very easily achievable with some models that can naturally be extended to a number of response classes higher than two. The easily extendable models in-

clude Decision Trees, Neural Networks, K-Nearest Neighbour, Naive Bayes classifiers, and Support Vector Machines. Other models, on the other hand, may require some specific formulation to be adapted to the new kind of task. For example, it may be necessary to divide the multiclassification problem into various binary classification problems or to put a total ordinal scale on the various classes before predicting them (Mehra & Gupta, 2013). In order to decompose a binary classification problem into many binary classifications the most widely used approaches are the following.

The most intuitive method is One-Versus-All (OVA) in which each single binary classification task tries to differentiate between one singular class label and all the others. If there are K classes, there will also be K classifiers each trained having one specific class as positive outcome and all the others as negative cases (Rifkin & Klautau, 2004).

The other more complex approach is comparing all classes in a pairwise fashion, commonly called All-Versus-All (AVA) method. Having K classes this procedure requires $\frac{K(K-1)}{2}$ different binary models to be trained. When predicting the class of a new observation this latter will be fed to all learners and it will be assigned the class that receives the most votes among the models (Hastie & Tibshirani, 1997).

Multinomial classification seems to fit the purposes of crash severity analysis perfectly, given the fact that accidents can be intuitively and practically classified in more than two classes according to the amount of harm they cause. This property is also apparent from the intrinsic structure of the KABCO and AIS scales described above. Almost all of the studies on the matter who had access to a data corpus that was detailed enough to differentiate between more than two classes of accidents have chosen to do so.

Wang and Kim, 2019, is an example of the most basic way to apply multinomial classification, dividing their observations into three classes of accidents: property damage only crashes, crashes with injuries and fatal crashes, representing 60.87%, 38.66% and 0.47% of the observations respectively. A multinomial logit model and a RF were used in the study to differentiate between the three classes with the latter achieving better overall results. Despite the multinomial logit model being specifically designed to tackle multiclassification problems, it may be strongly affected by multicollinearity among the variables, and it may be less

effective in understanding nonlinear trends in the data.

Tang et al., 2019, applied the same concepts but with a more comprehensive framework, starting with a dataset including details on the injuries based on KABCO scale. The two most serious degrees of the scales were merged due to their scarcity in the dataset: 285 A-level injuries and 51 K-level injuries. The final response variable was divided into 4 levels: no-injury, invisible injury, non-incapacitating injury, highest level injury. Figure 4 illustrates the two-layer stacking framework applied by the study in which the algorithms of the first layer, namely RF, AdaBoost and Gradient Boosting Decision Trees (GBDT), are trained and then used to make predictions on the test set. Said predictions are subsequently fed into a meta-classifier in the second layer; in this case a "Logistic Regression model is designed to fuse the classifying results from the first layer through establishing a sigmoid function to minimize the loss function by using the gradient decent algorithm" (Tang et al., 2019).

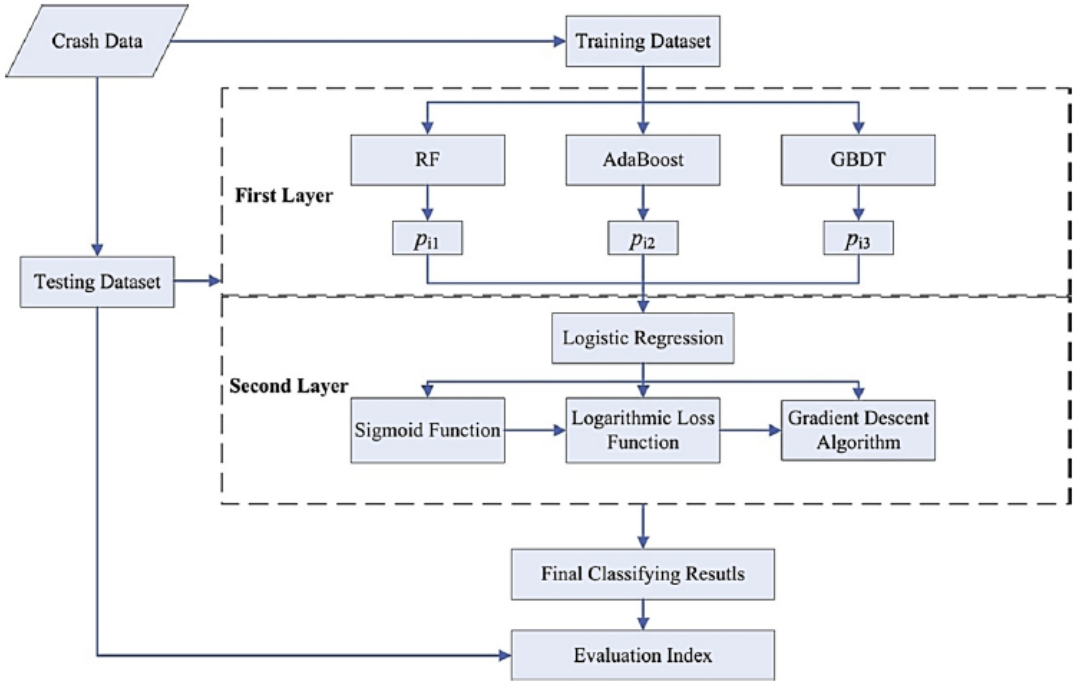

Figure 7: Two-layer framework used in Tang et al., 2019 . Source: Tang et al., 2019

2.4 Synopsis

As anticipated, the topic of crash severity prediction has been widely explored by researchers especially in recent years after the aforementioned declaration of intent by the United Nations General Assembly. The topic was subject to several different approaches regarding both different learning models and different data subjects. The most common learning techniques utilized were models which do not need particular assumptions on the starting dataset and are simultaneously able to detect nonlinear patterns and relationships in the data. The most prominent examples of these kind of models are RF, Boosting Models, KNN and SVM. The factors that presented themselves as most influential in classifying accidents were factors regarding the vehicles involved with particular attention to vulnerable road users (e.g., motorcycles and pedestrians), factors regarding driver's psychophysical state (e.g., age, alcohol consumption, fatigue) and some factors related with the road type (e.g., speed limit). All of these findings will be taken into account when defining the methodology for this study.

3 METHODOLOGY

This chapter will be focussed on the theoretical description of the methodologies applied in this study exploring all parts of the process.

3.1 Data Description

The data analysed in this study was collected from the official site of the municipality of Rome: https://dati.comune.roma.it/.

The datasets are in Italian with many of the categorical variables having italian descriptions. During the study, each time an analysed categorical predictor will present itself in Italian a brief English explanation of the latter will be also included.

The original dataset was divided in 12 *.csv* files, each corresponding to a month of the year, in which each row represented a particular person or vehicle involved in a crash. Table 1 presents the number of observations in each individual dataset.

Dataset	# Observations
January	4,610
February	5,073
March	4,538
April	5,274
May	6,467
June	6,514
July	6,392
August	4,468
September	6,691
October	7,770
November	7,848
December	7,443

Table 1: Number of observations in each dataset. Source: Self-Elaboration

All of the datasets combined generated a corpus of 73,088 observations spanning throughout all of 2021. In order to understand how data processing will be applied in future sections to reach a situation in which each entry of the dataset is a different crash, it is important to understand how the original data was composed. For this reason, a brief description of each of the variables contained in the initial dataset will be presented here.

- **Protocollo**: A numeric variable which represents the protocol number assigned by law enforcement authorities to each single road crash to distinguish them from each other. It will be fundamental in order to complete the transformation of the dataset into a one-crash-per-row style dataset.

- **Gruppo**: A grouping of accidents which was later discarded because of its inconsistency.

- **DataOraIncidente**: A timestamp of the accident precise to the minutes for most accidents. The cases in which the timing was unclear retained only the date.

- **Localizzazione1**: A categorical variable indicating the type of road in which the crash took place.

- **STRADA1**: A variable presenting the names of each street in which accidents took place. There are 4,844 different streets in the comprehensive dataset.

- **Localizzazione2, STRADA2, Strada02, Chilometrica, DaSpecificare**: All of them give additional details on the precise location of the accident in order to create an almost complete address. Most of these auxiliary variables are left empty or present very confused entries and were therefore left outside of the analysis.

- **NaturaIncidente**: A categorical variable indicating the type of accident occurred (e.g., Lateral Crash Between Moving Vehicles). It has 22 distinct values, some of which with very few occurrences.

- **particolaritastrade**: A categorical variable representing the particular road type in which the accident took place (e.g., Roundabout). It possesses 21 different levels with the same degree of sparsity as the previous variable.
- **TipoStrada**: A categorical variable indicating the way in which the road lanes are organized (e.g., double carriageway road), with 5 different categories.
- **FondoStradale**: Represents the state of the road surface (e.g., wet), with 10 unique levels.
- **Pavimentazione**: Represents the material composing the road surface (e.g, asphalt), with 11 distinct values.
- **Segnaletica**: Represents the type of signage present at the location of the accident (e.g., vertical signage), with 5 different levels.
- **CondizioneAtmosferica**: Indicates the weather conditions at the time of the crash (e.g., raining), with 8 unique levels.
- **Traffico**: Gives an overview of the traffic situation dividing accidents into low, average and intense traffic scenarios.
- **Visibilita**: Indicates the visibility on the site of the crash differentiating between insufficient, sufficient and good visibility.
- **Illuminazione**: A variable which was completely composed of not Assigned (*NA*) values but should have probably indicated the amount of lighting on the road.
- **NUM_FERITI, NUM_RISERVATA, NUM_MORTI, NUM_ILLESI**: They are four numerical variables indicating respectively: the number of injured people, the number of people with private medical records, the number of deceased people and the number of unharmed people. These four variables will be essential in constituting the response variable for this study.
- **Longitude** & **Latitude**: indicating precisely the location of the accidents.

- **Confermato**: A binary variable indicating whether the crash has been confirmed by law enforcers or not, taking values of -1 for a confirmed accident and 0 for a not confirmed one.

- **Progressivo**: A numerical variable which resets for each different protocol and indicates as a progressively increasing number every different vehicle involved in the accident. For example, a vehicle having four people inside will have 4 rows in the dataset with all of them having the same protocol and same progressive number, while a vehicle with no people inside will have only one row with one unique progressive number.

- **TipoVeicolo**: A categorical variable representing the type of vehicles (e.g., car, truck), with 46 distinct levels.

- **StatoVeicolo**: Representing the state of the vehicles in the accident, differentiating between moving, parked and fled vehicles.

- **TipoPersona**: Representing the role of each person in the accident differentiating among drivers, passengers and pedestrians.

- **Sesso**: A binary variable indicating the sex of the individuals involved in the accident.

- **TipoLesione**: A categorical variable presenting a very vague classification of the injuries resulting from the accident. It is incompatible with both the KABCO and the AIS-90 scales therefore it was not considered in the final study.

- **Deceduto** & **DecedutoDopo**: The first one is a binary variable indicating whether a person is deceased or not, while the second is a categorical variable indicating the time elapsed between the crash and the death.

- **CinturaCascoUtilizzato**: A categorical variable indicating whether the seat belt or the helmet were worn during the accident. It could have been a very interesting variable but it takes values of "Not Verified" in 53.37% of the cases and *NA* in 42.51% of the cases, limiting useful values at 4.12% of the comprehensive number of observations.

- **Airbag**: Reflecting whether the Airbag went off during the crash, however presenting too many *NA* values just like in the case of the previous variable.

The dataset presents overall a non-trivial amount of *NA* entries, representing 15.06% of the total number of cells in the data-frame. This factor may be due to various reasons spanning from poor dataset maintenance by the data controllers to the very peculiar structure of the dataset itself. Each row of the dataset, in fact, may represent a different type of participant in the accident, for example a single row may resemble an occupant of a moving car or a parked car that was involved in the accident or maybe even a pedestrian. Each of these different categories has diverse characteristics and therefore generates *NA* entries in different columns (e.g., a pedestrian would generate a *NA* in the *StatoVeicolo* column).

3.2 Data Pre-Processing

Given the findings of Section 3.1, it is evident how a correct handling of the data pre-processing is of paramount importance for the success of the subsequent modelling phase. The key idea for this task was that the final dataset needed each single row to present comprehensive information on a specific crash in order to be able to correctly perform the analysis. To pursue this objective, an in depth exploration of the possible cases in the dataset was necessary and subsequently required a vast amount of *Feature Engineering* in order to generate variables capable of synthesizing the characteristics of each accident. Due to the strongly categorical nature of the data, an intensive *Binning* procedure has been applied on some variables in order to produce more relevant and better represented categories that will be much more useful in the later analysis.

3.2.1 Feature Engineering (FE)

Duboue, 2020, defines Feature Engineering as *"the process of representing a problem domain to make it amenable for learning techniques. This process involves the initial discovery of features and their stepwise improvement based on domain knowledge and the observed performance of a given ML algorithm over specific training data"*.

In a more practical way, FE requires a deep observation of the raw data to then extract the most relevant information choosing a proper way to represent them for the expected recipients, which in the case of this study are classification algorithms. A very typical example of FE in data analysis is creating ratios or interaction factors between some of the pre-existent variables in order to capture the available information to a greater extent.

A widely known case of feature engineering is the Body Mass Index (BMI), whose formula is shown in equation 1 with w indicating the weight in kilograms and h the height in meters. The index takes as input the raw data available from the body (i.e., weight and height) and creates a metric that is more relevant and intuitive when analysing risk factors of some pathologies.

$$BMI = \frac{w}{h^2} \tag{1}$$

Another important aspect of FE is variable encoding which is especially relevant in the case of categorical predictors which, as previously mentioned, correspond to the vast majority of the available data for this study. The chosen approach followed the results coming from C. Lin et al., 2020, which indicated One-Hot encoding as the best approach for crash severity prediction resulting in a smaller Mean Absolute Error in their study.

One-Hot Encoding is an encoding process based on the subdivision of each discrete variable in a number n of different binary predictors where n is equal to the number of possible different realizations of the categorical predictor. Each of the created columns is a sparse vector and indicates whether the initial variable presented a particular level. This process leads to high dimensionality but it is widely used because of its simplicity and effectiveness (Bagui et al., 2021). Figure 8 shows a very basic example of One-Hot encoding.

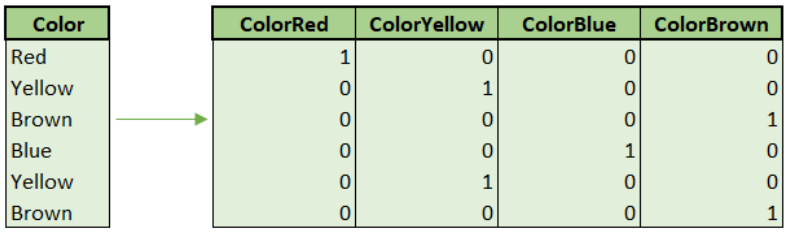

Figure 8: Visualization of One-Hot Encoding. Source: Self-Elaboration

3.2.2 Binning

The term Binning commonly refers to the data pre-processing operation of converting a continuos or semi-continuos variable into a categorical variable (Zeng, 2014). This operation is normally carried out by establishing a series of bins in which each observation can fit in, the number of available bins can vary depending on the granularity needed for each specific variable. This operation makes it so that values that are poorly represented in the data (e.g., extreme values) can be better included in the analysis without the need to remove them creating more significant and numerous classes. This is done also in the case of variables for which a very subtle change in value may not have a significant effect (e.g, the age of customers). Figure 9 shows a visualization of a classic binning example in which a normally distributed variable ranging from 1 to 100 is divided into five equal-size bins.

In this paper the binning approach has been applied in a slightly different fashion, since the dataset is composed of mostly categorical variables and some of them (e.g., *NaturaIncidente*, *particolaritastrade*) have many levels, some of which are very poorly represented in the data having just a couple of occurrences. The binning technique has therefore been applied to this type of variables in order to reduce the number of different categories, grouping together the ones that were either intuitively and logically similar or had similar effects on the response variable.

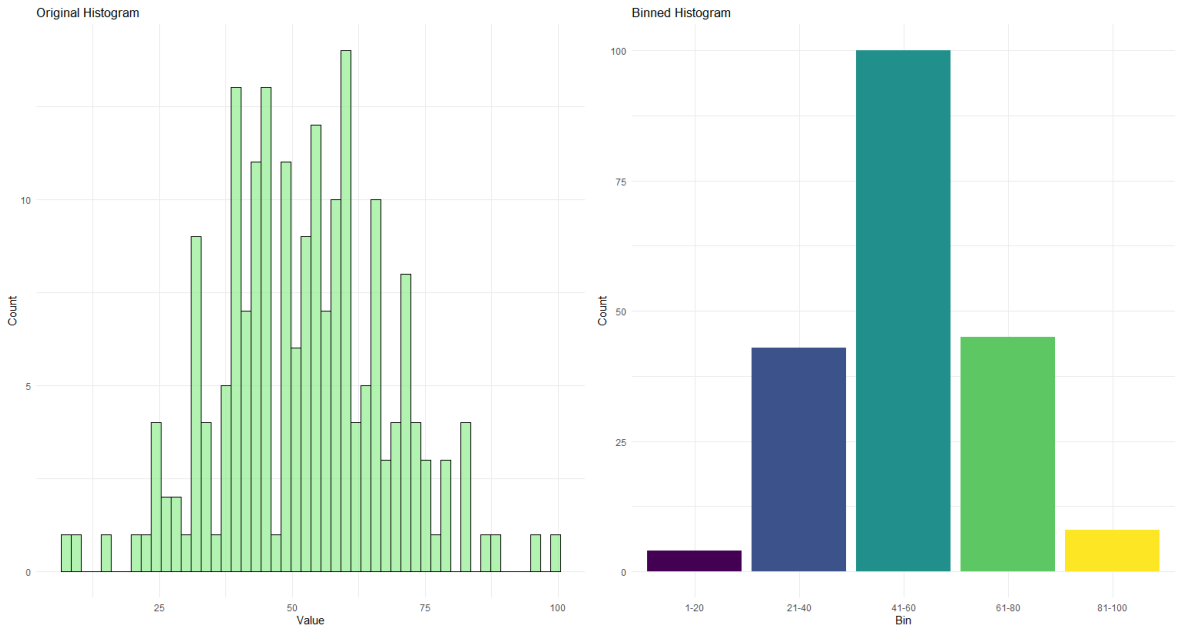

Figure 9: Visualization of Binning for a Continuos Variable. Source: Self-Elaboration

3.3 Applied Models

This section presents a brief theoretical explanation of all of the models employed in this study. All of the learners were trained and validated using the *caret* framework for ***R***. The analysis was structured as a binomial classification task in which the models had to classify accidents as *Casualty* in the case of injured or dead people and *No_casualty* otherwise.

3.3.1 K Nearest Neighbours (KNN)

As stated in Guo et al., 2003, KNN is a very intuitive yet often effective model for classification, based on the assumption that similar observations will reside close to one another according to whichever distance metric is chosen. In order to classify a new data-point, its k nearest neighbours are taken into consideration and the majority vote is considered as the outcome; it is also possible to produce probability predictions based on the ratios of votes among neighbours. As an obvious consequence of the process of the KNN algorithm, which

is intrinsically distance based, the model has a strong sensitivity to the magnitude of involved variables, making scaling numerical predictors crucial for valid results. KNN is a lazy learner, meaning that it defers its learning process to the prediction phase, resulting in very long classification times for new observations. The one and only hyper-parameter for KNN is k, which indicates how many neighbours to take into account, influencing heavily the performances of the model.

3.3.2 Classification and Regression Trees (CART)

Classification and Regression Trees are the base for all tree based models and work by recursively splitting the data in the attempt of minimizing some defined error metric (Loh, 2011). CART models stop creating splits when a certain stopping criteria is met: the most common method used for this purpose is to set a minimum number of observations in each leaf node. The splits applied by a CART are not random and depend normally on an impurity measure based on some variation of misclassification cost; their quality is measured by the decrease of said impurity metric that occurs between parent and child node: the greater the decrease the better the split. In particular, for CART models trained within the *caret* framework, the impurity of the splits is measured according to the *Gini Index* developed by the Italian Corrado Gini. Equation 2 shows the typical formulation of the *Gini Index* for decision trees, which represents basically the total variance between the various available classes. The lower the index the purer the split is, its lower-bound is 0 which means that all observations in a certain region (R_m) belong to the same class. The highest value possible, on the other hand, is $1 - \frac{1}{k}$, where k is the number of possible classes, occurring when classes are uniformly distributed in the analysed region.

$$G(R_m) = \sum_k \hat{p}_{mk}(1 - \hat{p}_{mk}) = 1 - \sum_k \hat{p}_{mk}^2 \tag{2}$$

After the CART tree has been built with the proper splits, a pruning protocol is normally applied in order to avoid over-fitting to the training data. A very deep tree normally presents an high degree of variance because of the few observations in each leaf while shallow trees

present higher bias since their structure is very rigid.

One of the greatest feats of CART and all tree-based models in general is that they handle well categorical variables even without encoding and can easily detect non-linear patterns. CART however fail in detecting simple linear patterns.

3.3.3 Random Forest (RF)

Howard and Bowles, 2012, stated *"ensembles of decision trees—often known as "random forests"—have been the most successful general-purpose algorithm in modern times"*.

RF were ideated by Breiman in the early years of the twenty-first century (Breiman, 2001), but in order to understand their functionality it is necessary to start from their closely related ancestor: Bagging Trees (BT).

Bagging trees were ideated in order to fight the high variance that is normally attributed to CART, approaching this task in the most intuitive way: by taking the average of the predictions of various CART. In order to do this, a large number of trees needed to be trained on the data belonging to the same population, for this purpose bootstrap resampling was applied. Equation 3 shows the formula for the prediction of a BT system, basically taking the average of the prediction of various learners $\hat{f}_n$ that are trained on a number N of different bootstrapped samples. The biggest weakness of this approach resides in the apparent correlation between the various trees which are trained on semi-different data and use all the same predictors.

$$\hat{f}_{bag}(X) = \frac{1}{N}\sum_{n=1}^{N}\hat{f}_n(X) \tag{3}$$

RF mitigate this correlation problem by considering only a subsample of the total number of available predictors in each different tree (Biau & Scornet, 2016). In the *caret* framework this parameter is called *mtry* and normally considering m total independent variables it is necessary to include either $\sqrt{m}$ or $\frac{m}{3}$ in each tree. The biggest concern with RF is normally their block-box nature which makes rigorous analysis of the model very intricate. At the same time, despite their notable computational complexity, RF are quite simple to optimize.

Another interesting feature is the built-in variable importance ranking which is obtained by calculating the average decrease in impurity for trees that include a specific predictor.

3.3.4 eXtreme Gradient Boosting Trees (XGBtree)

XGBtree is a tree boosting model, meaning that the base concept behind it is to ensemble numerous trees in order to get more stable and accurate predictions capturing even complex non-linear relations. It is in principle very similar to the RF approach (also based on ensembling) with a crucial difference in the training method. While RF trains the trees independently in a parallel manner, boosting algorithms work by training learners in an additive fashion, one after the other. Each learner is individually considered weak but their cumulative effort should create a comprehensively strong learner represented by the entirety of the boosted model. Each tree is in fact trained on a manipulation of the residuals of the previous one, basically trying to correct the errors of its predecessor. In simple terms, boosting models try to tackle a machine learning problem by learning its intricacies step by step.

XGBtree is a very popular model and has been widely known for performing very well in both regression and classification tasks. It is based on Gradient boosting which attempts at optimizing an objective function through a *Gradient descent* approach. Equation 4 shows the typical formulation of an objective function, composed by $L(\theta)$ representing the training loss, which is an indicator of goodness of fit, and $\Omega(\theta)$ representing the regularization term which prevents over-fitting.

$$obj(\theta) = L(\theta) + \Omega(\theta) \tag{4}$$

XGB in particular has various characteristics that make it so effective: it enables users to define custom loss functions for specific tasks while automatically handling the regularization of the model facilitating generalization to new data. It also comes with the capability of estimating feature importance and it is very well optimized for parallel processing, speeding up the training process for very large datasets.

When training an XGBtree model with the *caret* framework, the user has access to mul-

tiple hyper-parameters which can be cross-validated and tailored to each specific task. Available parameters are:

- **nrounds**: number of boosting iterations
- **max_depth**: maximum depth of each weak learner
- **eta**: shrinkage factor determining also the learning rate
- **gamma**: minimum loss reduction between learners
- **colsample_bytree**: ratio of columns used in each tree
- **min_child_weight**: minimum weight of a leaf node (if said node has a inferior weight no further splitting will be applied)
- **subsample**: ratio of the available data used for each tree

Each of these parameters is important in some capacity when trying to get the best available model, for this reason all of their combinations should be cross-validated at the same time. This validation process of course elongates running times in a tremendous manner, even considering the general efficiency of the XGBtree algorithm. As a consequence, in the context of this study the hyper-parameters for this model have been cross-validated in pairs, which is not as optimal but enormously less computationally expensive and time demanding.

3.4 Validation

K-fold Cross Validation (CV) was used as validation mechanism for every analysis in this paper. The CV approach requires the training dataset to be divided into k independent folds in which the ratios of classes remain similar to the original dataset (stratified sampling with no replacement). Each algorithm is then trained on $k-1$ folds and attempts at predicting on the last available fold. This process is then repeated until the same algorithm has been trained on all the k possible combinations of folds. Performance measures are then calculated each time thanks to the usage of the unused fold as pseudo test set and lastly a final CV score

is calculated taking the average performance over all combinations. This method allows the user to better approximate what will be the performance of each algorithm on new unseen data since it reduces dramatically the bias in the validation set. Figure 10 shows a visualization of a 5-fold Cross Validation. In this paper a 10-fold approach was chosen because of it's increased precision.

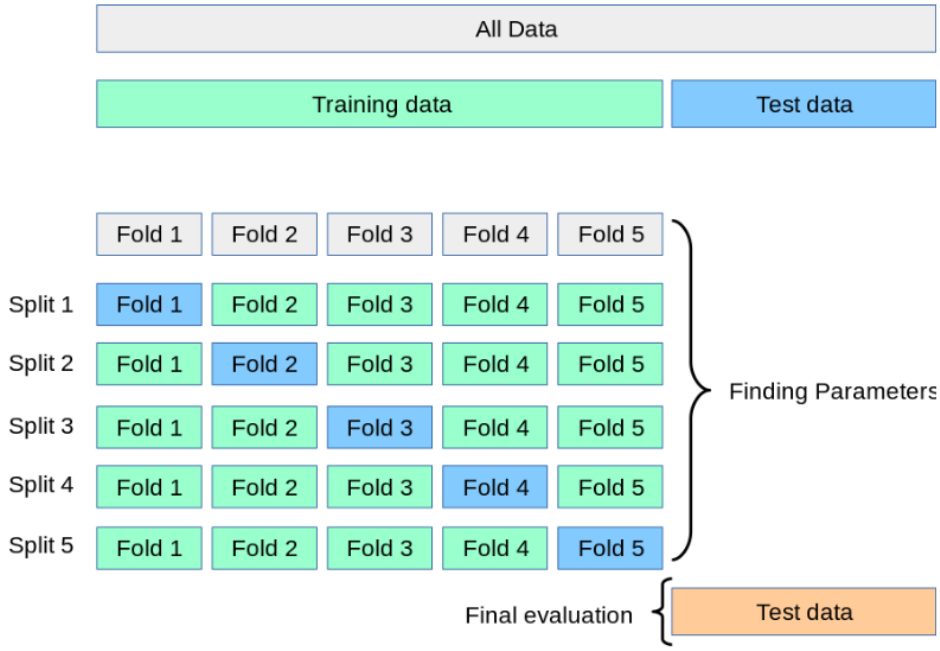

Figure 10: Visualization of 5-fold CV. Source: scikit-learn.org, n.d.

3.5 Performance Metrics

In order to measure the performances of the various models, the chosen performance metric was the *Area Under the Curve* (AUC) associated with the *Receiver Operating Characteristics* (ROC) curve.

In the case of a binary classification problem, the ROC curve plots Sensitivity $(\frac{TP}{TP+FN})$ against 1-Specificity $(1 - \frac{TN}{TN+FP})$ by moving the classification threshold between the two extremes of the classification boundaries. The AUC is simply the area that lies under the ROC curve, bounded from above to 1 and from below to 0; it has been described as *"closely*

related to the ranking quality of the classification"(Cortes & Mohri, 2003). In the case of a random classifier, the ROC curve will be a straight line connecting the origin to the point $(1, 1)$, with an AUC of exactly 0.5. Figure 11 shows a sample ROC curve in blue and its AUC of 0.795, highlighting also the ROC for a random classification in red.

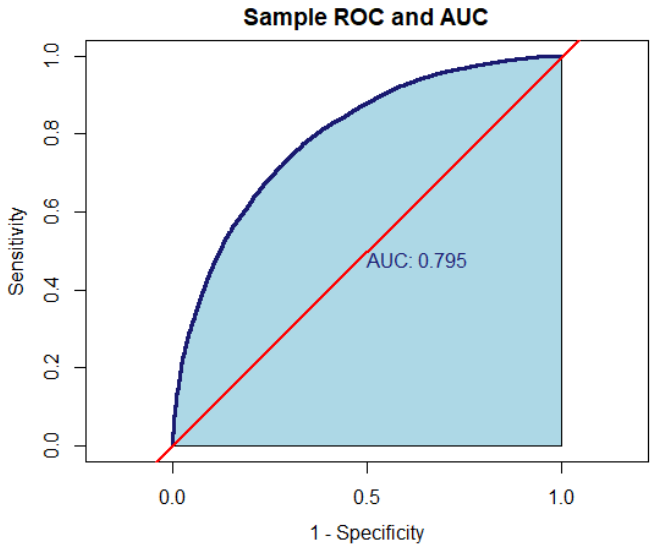

Figure 11: A Sample ROC curve and its AUC. Source: Self-Elaboration

The AUC method was chosen since it gives a more comprehensive view on classification models compared to confusion matrices and standard metrics due to its manipulation of the threshold and eye-catching nature.

4 ANALYSIS

In this chapter all of the procedures applied at all steps of the analysis will be presented and discussed in depth with a final presentation of the numerical results of the modelling phase.

4.1 Data Pre-Processing

As remarked in the previous section of this paper, data pre-processing is essential for any data analysis task. The data available for this study are a perfect example of data that would be nearly impossible to analyse without some specialized processing.

The aim of this phase is to get the data from their original form, which was extensively presented in Section 3.1, to a more manageable and compact structure which will facilitate the modelling process. The finalized dataset will be composed of one single observation for each distinct accident including some new engineered features that try to capture the most relevant peculiarities of each single case.

4.1.1 Data Loading and Handling of Missing Values

The first step towards the aforementioned goal was to load the 12 datasets and create a new categorical variable *month* in each of them representing the month in which the accident took place in order to be able keep track of the origin of each observation.

After this preliminary step, the datasets were merged and all of the non-confirmed crashes were discarded according to the *Confermato* variable leaving 69,260 confirmed cases. The handling of missing values was performed on variables that urgently needed it; for example the variable regarding the sex of the people involved presented various missing values in the cases in which a row represented a parked car, which of course has no sex attribute. In this case the value of **ND** was assigned to fix the problem. The structure of the dataset in which each row does not always represent the same type of entity created various situations in which similar problems occurred(e.g., *TipoPersona* and *TipoVeicolo*). All of them were fixed in a similar fashion, by inserting a new variable level which avoids missing values and makes the dataset more coherent.

The next important step was to delete the variables that presented an abnormal percentage of *NA* due to their excessively specific nature or lack of information. Table 2 shows the variables that were removed in this step and their associated percentage of *NA* values.

Variable Name	Percentage of *NA* values
STRADA2	64.08%
Strada02	22.45%
Chilometrica	58.37%
DaSpecificare	77.55%
Illuminazione	100%
DecedutoDopo	82.33%
CinturaCascoUtilizzato	40.91%
Airbag	56.56%

Table 2: Variables with notable *NA* percentages. Source: Self-Elaboration

Some other variables were also removed due to other specific reasons:

- **STRADA 1 & Localizzazione1**: indicated the name of the street in which the accident took place which is not beneficial for this particular study case.
- **TipoLesione**: did not present sufficient information on the injuries to transform the study in a multinomial classification problem and was therefore useless, being basically a duplicate of the response variable.
- **Deceduto**: for the purpose of this study the variable did not give any additional information compared to the *NUM_MORTI* variable.

The last step in this phase is to remove all of the accidents which presented any remaining *NA* values. Only 53 entities were removed from the dataset in this last process which is an infinitesimal portion of the initial number of rows in the dataset (0.08%) testifying the efficacy of the missing value handling performed.

4.1.2 Data Transformation and FE

The actual transformation of the dataset in the desired one-row-per-accident form will take place in this section. In order perform this task, a dataset specific iterative algorithm has been designed. The algorithm works by extracting a unique vector containing all of the protocols in the dataset: as explained in Section 3.1 the *Protocollo* variable acts as unique identifier of each distinct accident. Thanks to this new vector we can scan the dataset in an accident by accident fashion, analysing at each iteration all of the data-entries belonging to a single crash, regardless of what they represent (e.g., pedestrian, parked car, driver, etc.).

This approach allows for an in depth analysis of each case and for the creation of new variables that will summarize the most important characteristics of the accidents that would be otherwise lost when converting multiple rows into a single one. During this iterative process almost all of the FE variables are created with the aid of Regular Expression for string matching. The generated predictors in this section revolve around characteristics of the vehicles and people involved in the accidents and are described in the following list:

- **nummotoinvolved**: a numerical predictor indicating how many moving two-wheeled vehicles were involved in the crash. This aspect was deemed important also due to the results of previous research on the matter (M.-R. Lin & Kraus, 2008; Vajari et al., 2020).

- **numheavyinvolved**: a numerical predictor indicating how many moving heavy vehicles(e.g., buses, trucks, etc.) were involved in each crash.

- **numcarinvolved**: a numerical predictor indicating how many moving cars were involved in each crash.

- **NumMalesInvolved**: a numerical predictor indicating how many males were involved in each crash.

- **NumFemalesInvolved**: a numerical predictor indicating how many females were involved in each crash.

- **NumNDInvolved**: a numerical predictor indicating how many people with unknown sex were involved in each crash.

- **maledriver**: a boolean variable indicating whether one ore more of the drivers involved in the accident were males.

- **occupancyindex**: a continuos variable that was created with the aim of representing the number of people for each moving vehicle in the accident. Equation 5 shows the formula used to create this variable, *#MovingVehicles* is represented by the sum of the *nummotoinvolved*, *numheavyinvolved* and *numcarinvolved* variables.

$$\frac{\#MalesInvolved + \#FemalesInvolved + \#NDInvolved}{\#MovingVehicles} \tag{5}$$

- **timeofday**: a categorical variable deriving from the *DataOraIncidente* predictor. The resulting variable has 4 levels which were split based on the hour of the accident according to Table 3. This binning approach was applied to keep the most relevant information contained in the original predictor while avoiding excessive details. Some accidents which presented incoherent timestamps were removed in the process.

hour range	*timeofday* value
$0 \leq hr < 6$	night
$6 \leq hr < 12$	morning
$12 \leq hr < 18$	afternoon
$hr \geq 18$	evening

Table 3: Splits applied in the *timeofday* variable. Source: Self-Elaboration

- **Casualties**: The binary response variable of this analysis, taking values of *"No_casualty"* in the cases in which nobody was injured or died and *"Casualty"* otherwise.

After the transformation of the dataset we got a comprehensive data corpus of 26,243 different accidents.

4.1.3 Data Cleaning

The next step was to remove accidents that presented people with unknown medical records and accidents that did not take place on urban roads due to their excessive scarcity.

Binning was subsequently applied on some specific categorical variables:

- **FondoStradale**: presented 10 levels with one indicating dry road surface (the most prevalent with 22,101 occurrences) and the other 9 all indicating some variations of a slippery or wet surface. For this reason the variable was binned to create a binary predictor with levels *"Asciutto"* (dry) and *"Scivoloso"* (slippery). Figure 12 shows a visualization of the new binned variable.

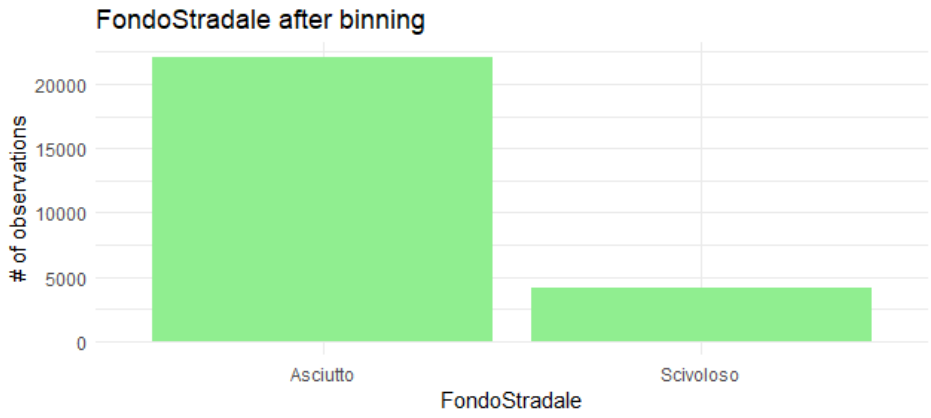

Figure 12: The binned *Fondostradale* variable. Source: Self-Elaboration

- **Pavimentazione**: presented 11 levels of which 10 them represented some very peculiar imperfection or characteristic of the road surface with a very low number of occurrences (cumulatively 939): for this reason the variable was binned in a similar way as in the case of *FondoStradale*. Figure 13 presents the bar-plot of the two remaining levels after binning which are *Asfaltata* ("paved") and *Con problematiche* ("presenting problems").

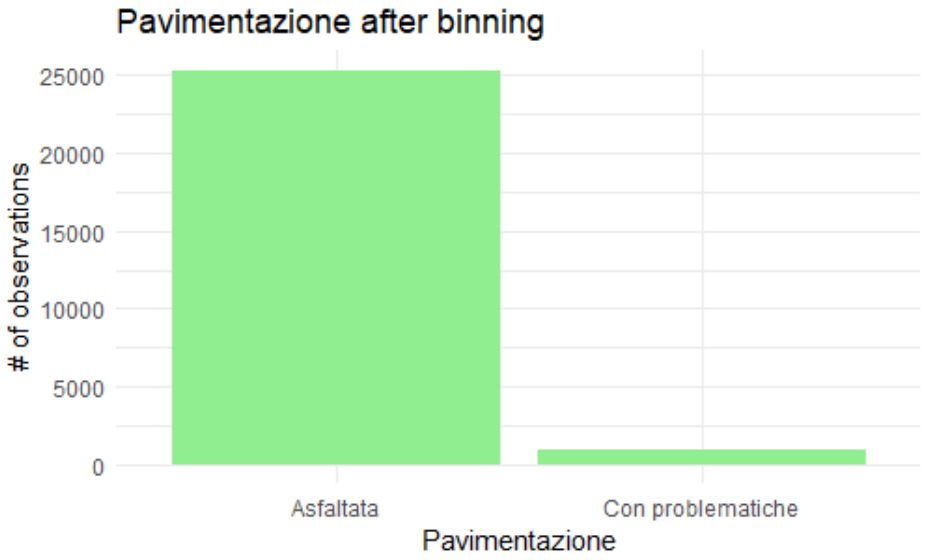

Figure 13: The binned *Pavimentazione* variable. Source: Self-Elaboration

- **NaturaIncidente**: presented 21 levels with some of them representing very similar accident dynamics; for example ”moving vehicle against parked vehicle” and ”moving vehicle against parked vehicles” are two distinct levels in the original data. The binning approach for this variable was aimed at reducing the total number of levels grouping together these very similar levels and also grouping all accidents in which the vehicle did not hit anything during the accident and was simply damaged by road characteristics or exited the road by accident. The final version of this variable presented 14 levels.

The last step in cleaning the data was to eliminate all observations that presented some very peculiar labels for some variables that were too sparse to be used in the modelling phase and to remove also all of the variables that did not make any sense in this new organization of the data (i.e., *TipoVeicolo*, *Progressivo*, *StatoVeicolo*, *NUM_RISERVATA*, *DataOraIncidente*, *Localizzazione1*, *NUM_MORTI*, *NUM_FERITI*,*NUM_ILLESI*, *Gruppo*, *Sesso*, *TipoPersona*).

4.1.4 Data Summary

Table 4 shows a brief schematic summary of the data corpus after *Pre-Processing* operations.

Variable	Symbol	Type	Description
Protocol	Protocollo	Numerical	Unique identifier of crashes
Type of accident	NaturaIncidente	Categorical	1, Collisione contro auto ferme; 2, Collisione contro auto in arresto; 3, Collisione contro auto in sosta; 4, Collisione contro oggetto fisso; 5, Incidente a solo; 6, Investimento di pedone; 7, Scontro frontale fra veicoli in marcia; 8, Scontro frontale/laterale DX fra veicoli in marcia; 9, Scontro frontale/laterale SX fra veicoli in marcia; 10, Scontro laterale fra veicoli in marcia; 11, Tamponamento; 12, Tamponamento Multiplo; 13, Veicolo in marcia che urta buche nella carreggiata; 14, Veicolo in marcia contro ostacolo accidentale.
Road characteristic	particolaritastrade	Categorical	1, Curva a visuale libera; 2, Curva senza visuale libera; 3, Incrocio; 4, Instersezione non regolata/non segnalata; 5, Intersezione semaforizzata; 6, Intersezione stradale segnalata; 7, Rettilineo; 8, Rotatoria
Type of runway	TipoStrada	Categorical	1, Una carreggiata a senso unico di marcia; 2, Una carreggiata a doppio senso; 3, Due carreggiate; 4, Più di due carreggiate
Road condition	FondoStradale	Categorical	1, Asciutto; 2, Scivoloso
Road pavement	Pavimentazione	Categorical	1, Asfaltata; 2, Con problematiche
Signage	Segnaletica	Categorical	1, Assente; 2, Orizzontale; 3, Verticale; 4, Verticale ed orizzontale
Weather	CondizioneAtmosferica	Categorical	1, Sereno; 2, Nuvoloso; 3, Pioggia in atto
Traffic	Traffico	Categorical	1, Scarso; 2, Normale; 3, Intenso
Visibility	Visibilita	Categorical	1, Insufficente; 2, Sufficiente; 3, Buona
Longitude	Longitude	Numerical	Min:8.539; 1stQu.:12.45; Median:12.494; Mean:12.496; 3rdQu.:12.55; Max:13.704
Latitude	Latitude	Numerical	Min:38.10; 1stQu.:41.86; Median:41.89; Mean:41.89; 3rdQu.:41.92; Max:45.70
Month	month	Categorical	1, Jan; 2, Feb; (...) 12, Dec
Number of two-wheeled	nummotoinvolved	Numerical	Min:0; 1stQu.:0; Median:0; Mean:0.2798; 3rdQu.:1; Max:3
Number of heavy v.	numheavyinvolved	Numerical	Min:0; 1stQu.:0; Median:0; Mean:0.05443; 3rdQu.:0; Max:2
Number of cars	numcarinvolved	Numerical	Min:0; 1stQu.:1; Median:1; Mean:1.448; 3rdQu.:2; Max:7
Number of other v.	numotherinvolved	Numerical	Min:0; 1stQu.:0; Median:0; Mean:0.006381; 3rdQu.:0; Max:1
Number of unknown people	NumNDInvolved	Numerical	Min:0; 1stQu.:0; Median:0; Mean:0.0486; 3rdQu.:0; Max:1
Number of males	NumMalesInvolved	Numerical	Min:0; 1stQu.:1; Median:2; Mean:1.619; 3rdQu.:2; Max:10
Number of females	NumFemalesInvolved	Numerical	Min:0; 1stQu.:0; Median:1; Mean:0.7747; 3rdQu.:1; Max:7
Male driver inv.	maledriver	Categorical	1, False; 2, True
Time of the day	timeofday	Categorical	1, morning; 2, afternoon; 3, evening; 4, night
Occupancy index	occupancyindex	Numerical	Min:1; 1stQu.:1; Median:1; Mean:1.384; 3rdQu.:1.5; Max:7
Casualties	Casualties	Categorical	Response Variable. 1, No_casualty; 2, Casualty

Table 4: Summary of the available data after *Pre-Processing*. Source: Self-Elaboration

4.2 Data Exploration

This section will be devoted to exploring the available data to get some initial insights in the characteristics of the observed phenomena.

4.2.1 Geographical Analysis

A geographical analysis was conducted in order to ensure that all of the analysed accidents belonged to the appropriate region. Figure 14 was realized thanks to the *maps* and *osmdata* packages for ***R*** and presents on the left all of the accidents appearing in the pre-processed dataset. It is apparent from the plot that some observations were erroneously inserted in the dataset since they are not even remotely close to the municipality of Rome. All of the 28 occurrences residing outside of the province of Rome were subsequently removed and the remaining crashes were plotted again on a smaller map.

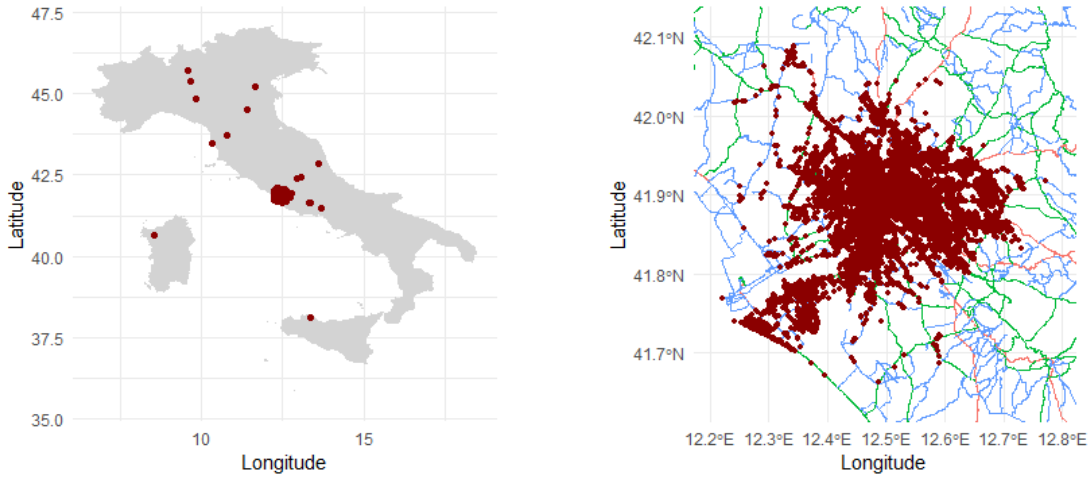

Figure 14: accidents before and after removal of outliers. Source: Self-Elaboration

4.2.2 Time Analysis

In order to understand whether the temporal aspects of the accidents could have an impact, both the *month* and *timeofday* variables were analysed deeply. Figure 15 shows a stacked

bar-plot of the number of accidents that took place during each month of the year. Each bar is subdivided into two colors: light-green, representing accidents with casualties and dark-green, representing accidents with no casualties. The percentages inside the bars represent how accidents were split during each single month. It is apparent from the bar-plot that months that included less accidents (e.g., January, March, etc.) tend to have also a lower casualty rate with the notable exception of August in which the number of accidents is quite low but the casualty rate is the highest in the dataset. The reason behind the low amount of accidents may very well be explained by the notorious decrease in Rome traffic during August, but the high casualty rate appears counter-intuitive. In general, the trend for casualty rate seems to follow the temperatures during the year, with warmer months presenting on average higher values. August is one of the hottest months of the year and also the one in which most of the people go for vacations: these factors may affect drivers' behaviour and focus on the road. In order to understand whether this factor is actually relevant the *month* variable will be kept in the study during modelling.

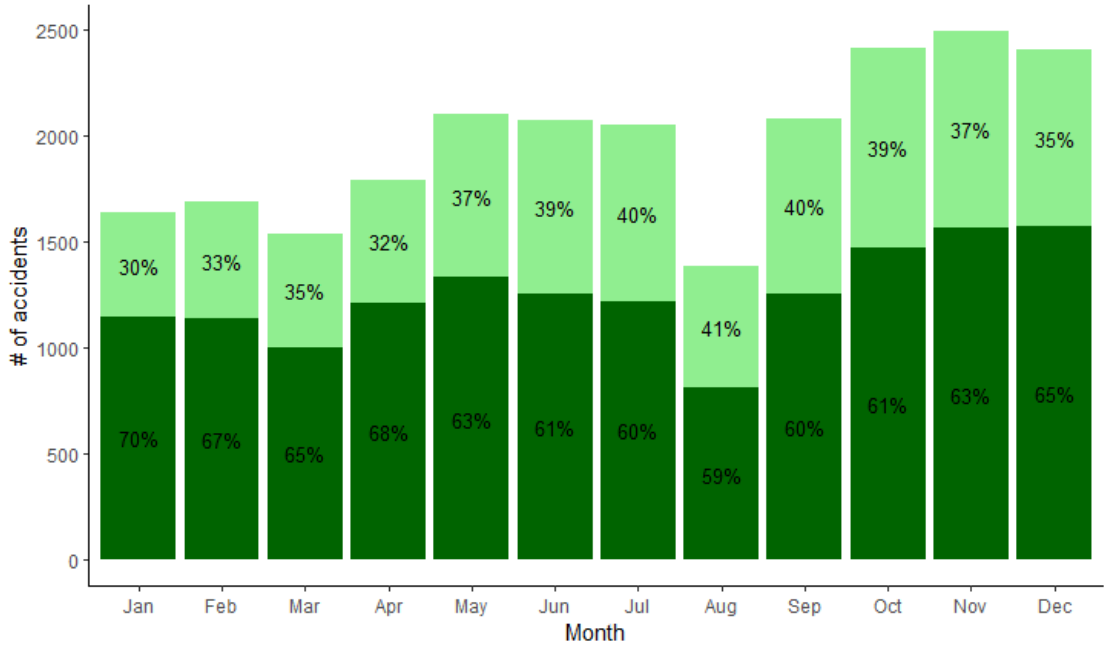

Figure 15: Number of accidents and casualty rates by months. Source: Self-Elaboration

A similar approach was applied for the *timeofday* variable in Figure 16, which presents

the number of accidents subdivided by the time of the day in which the crash took place, using the same colour scheme as Figure 15. The number of accidents follows the expected trend, peaking during the afternoon and dropping dramatically during the night: just like traffic levels in the city of Rome. Casualty rates, on the other hand, face a substantial increase in the later parts of the day, which is reasonable due to accumulated fatigue and worse lighting conditions when driving at night. This increase could also be due to possible differences in the emergency handling systems during the day and the night.

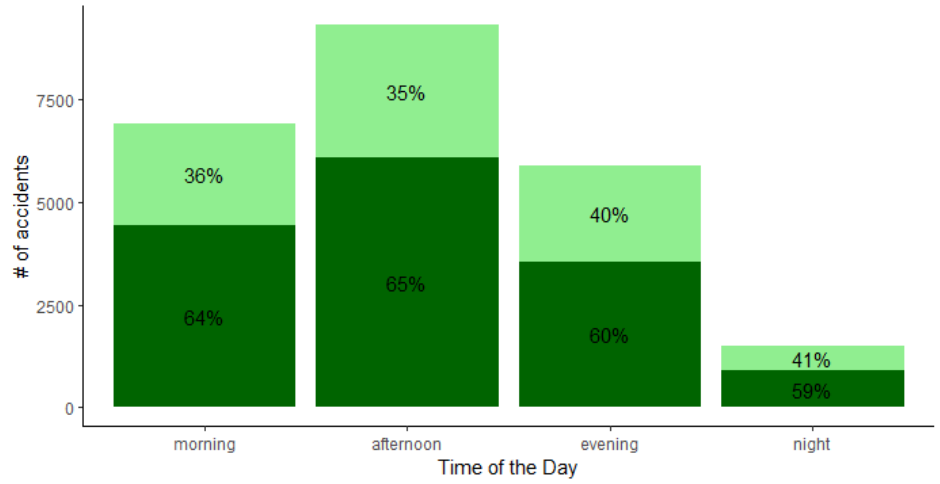

Figure 16: Number of accidents and casualty rates by time of the day. Source: Self-Elaboration

4.2.3 Gender Analysis

The gender of the people taking part in the accident was included in the dataset and, in order to understand whether there are actual differences between males and females during car crashes, an analysis of the new FE variables *numMalesinvolved*, *numFemalesinvolved*, *NumNDInvolved* and *maledriver* was conducted. Figure 17 highlights the enormous prevalence of males in the observed crashes: they compose comprehensively 66% of the people involved in car accidents (38,256) while females represent only the remaining 32% (18,311). This strong imbalance is coherent with the results of previous studies (Chang & Wang, 2006; C. Lin et al., 2020) and considering also the fact that in Italy 83.3% of the victims of car crashes in 2021

were men (ISTAT, 2022), this gender skewness may be very significant for the analysis.

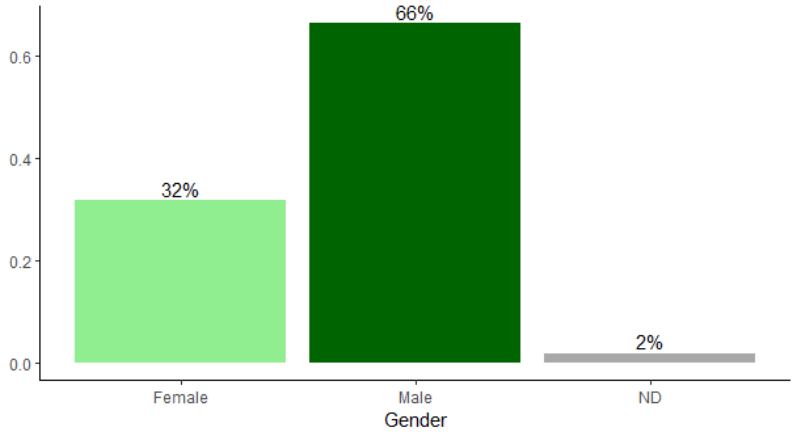

Figure 17: Percentages of males and females involved in crashes. Source: Self-Elaboration

In order to further explore the aforementioned dynamic, Figure 18 was plotted highlighting an evident co-occurrence between the presence of at least one male driver and the advent of casualties in a crash. This graphical intuition is also backed up by a Pearson's Chi-squared test with Yates' continuity correction presenting a *p-value* very close to 0.

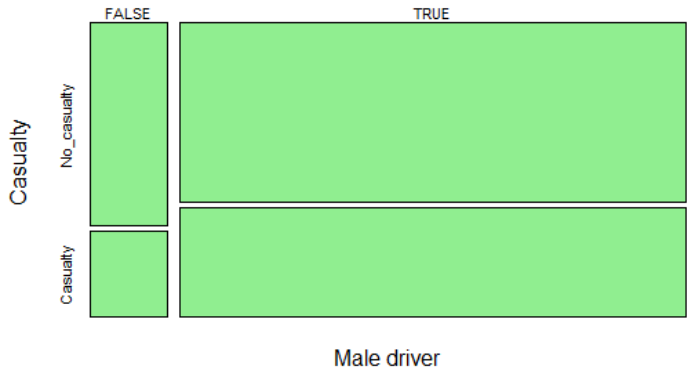

Figure 18: Mosaic plot of *maledriver* and *Casualties*. Source: Self-Elaboration

4.3 Modelling

In this section the training and validation of each of the utilized models will be presented. After some additional standard processing (e.g., standardization of numerical variables, one hot encoding of categorical variables, etc.), the comprehensive number of accidents was 22,653 which received a 80/20 split into training and testing set maintaining the same response class percentages.

4.3.1 KNN

KNN was the only non-tree based model employed in this study. Due to its lazy nature it was the fastest model in training, allowing for the CV of many possible values for the number of neighbours to consider in each prediction. Figure 19 shows the tested values for the k parameter highlighting the value of $k = 41$ as the best option with a CV score of 0.808 in the AUC regards. The performances are remarkable for such a simple model, however its notorious inefficiency in prediction and its high reliance on distances may lead to some degree of bias when facing new data. The best k value is relatively high which should counter to some degree this innate bias.

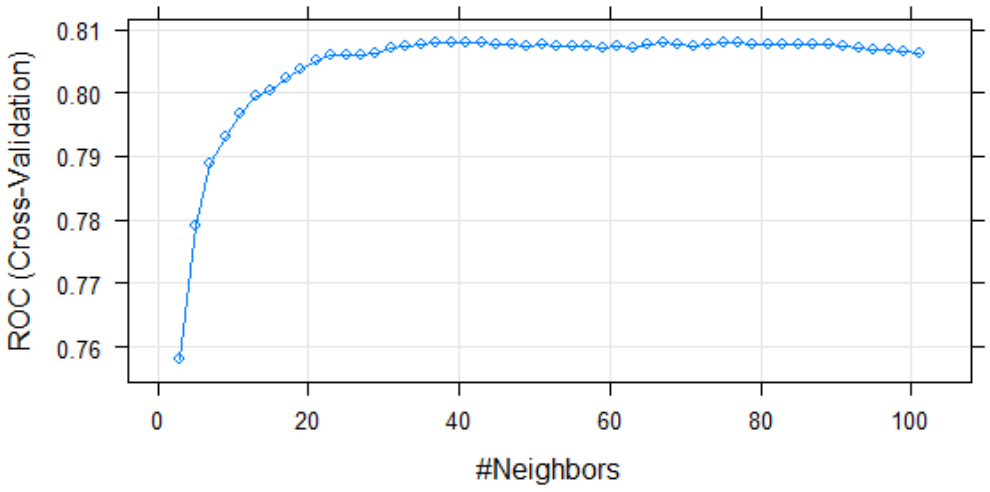

Figure 19: KNN model CV plot. Source: Self-Elaboration

4.3.2 CART

The first trained algorithm was the CART, which was implemented using the *rpart* caret model. The only hyper-parameter for this model is the so called complexity parameter (CP) which determines the entity of the pruning applied to the tree after training: a value of 1% in CP will prune the tree to find a subtree that is at most 1% worse in training performance in order to improve generalization. Figure 20 illustrates the tested values for CP and highlights how the best CV scores were achieved with a CP < 0.101.

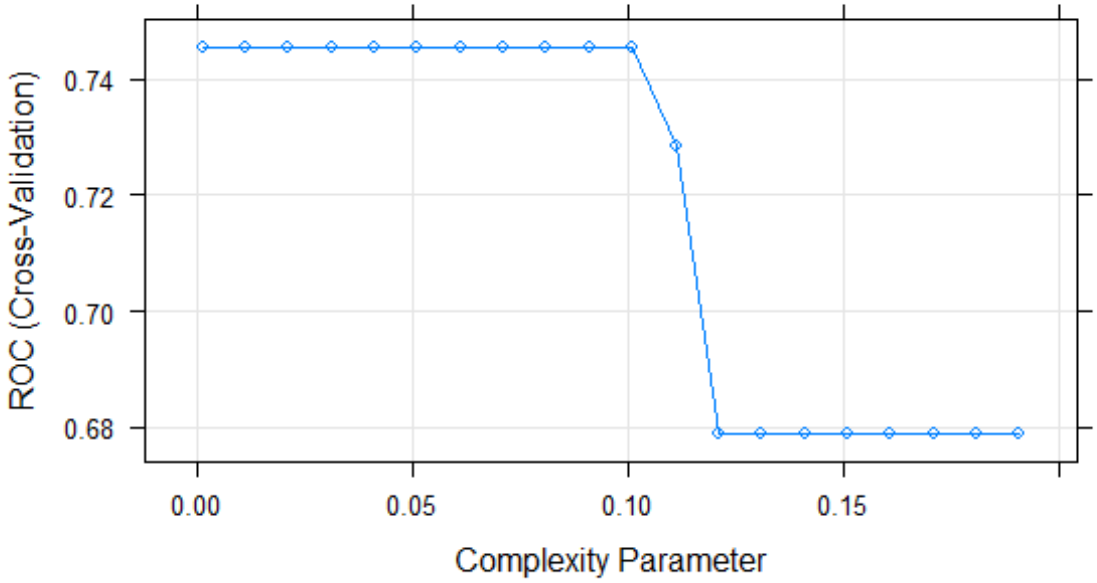

Figure 20: CART model CV plot. Source: Self-Elaboration

Given the generally simple interpretability of CART models, Figure 21 was used to visualize the final tree. The results are quite surprising: with the use of just two variables the model is able to get a decent CV score of 0.746 when considering AUC. It is apparent how pedestrians and motorcyclist incur in much greater risks when compared to other road users, also confirming the results of previous studies on the matter (M.-R. Lin & Kraus, 2008; Mokhtarimousavi et al., 2020; Vajari et al., 2020).

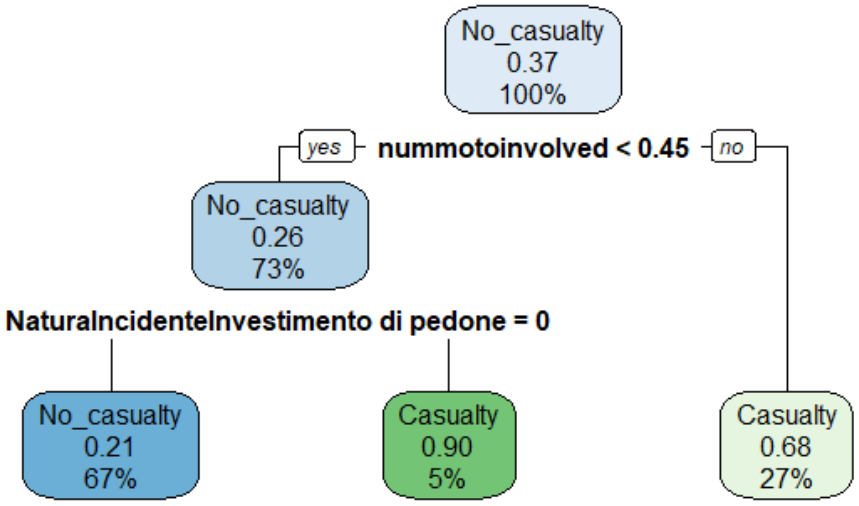

Figure 21: Final CART model. Source: Self-Elaboration

4.3.3 RF

The results of the previous section generated great expectations for more complex tree based models, especially for RF which are typically very efficient in surpassing simple CART models in performances when there a few very influential predictors in the dataset. Each tree will be trained on a different subset of variables hence the model will be able to analyse deeply the contribution of a wider variety of predictors. Figure 22 shows the number of randomly selected predictors in each single tree of the RF and their respective AUC: the best CV score was achieved with 5 variables for each tree and presented an AUC of 0.8113. This non-trivial improvement confirms the great capabilities of RF models.

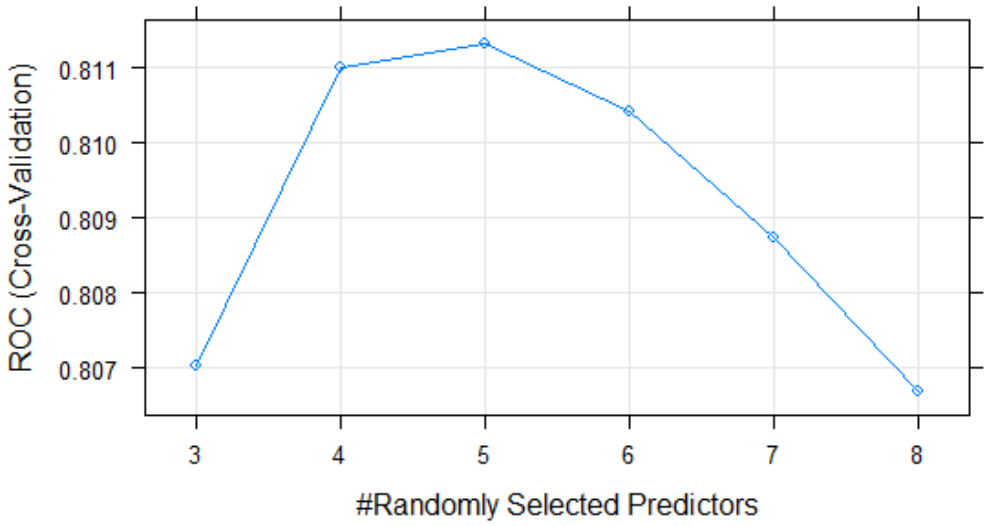

Figure 22: RF model CV plot. Source: Self-Elaboration

Figure 23 shows the variable importance plot for the RF model with its 9 most relevant predictors. The variables that were deemed important by the CART model were confirmed as important also by the RF; other predictors regarding the number of people in the accident and the time of the crash were also included.

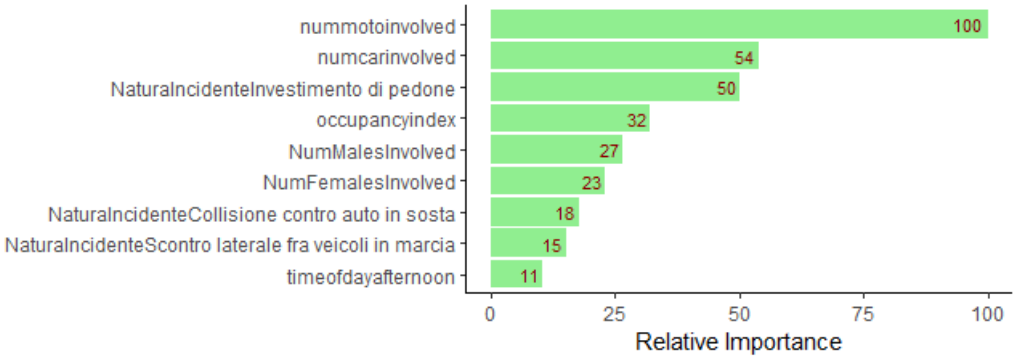

Figure 23: Variable importance plot for the RF model. Source: Self-Elaboration

4.3.4 XGBTree

Given its intrinsic additive structure, the XGBTree model was trained including only 20% of the total number of predictors for each boosting iteration in order to reduce correlation between individual trees. This approach allowed the algorithm to focus not only on the few most relevant features, like *nummotoinvolved*, but on a wider scheme just like in the RF model. Countless iterations of the algorithm were evaluated for the purpose of this study thanks to a personalized script which handles CV of parameters in a slightly less rigorous but much faster method. In said script the algorithm is first trained with some generic values for each of its 7 hyper-parameters and then an additive approach is applied to get to the final model. Instead of cross-validating all parameters at once, generating an extreme number of possible combinations and enormous running times, each parameter was cross-validated subsequently building on previously optimal values, progressively refining the algorithm. The optimization of the algorithm consisted in five rounds which are summarized in Table 5; it is important to notice how the number of boosting iterations was included in all rounds to enhance the flexibility of the algorithm.

Optimization Round	Cross-Validated Parameters
Round 1	*nrounds*; *eta*; *max_depth*
Round 2	*nrounds*; *max_depth*; *min_child_weight*
Round 3	*nrounds*; *colsample_bytree*; *subsample*
Round 4	*nrounds*; *gamma*
Round 5	*nrounds*; *eta*

Table 5: CV stages for the XGBTree model. Source: Self-Elaboration

After this semi-automated procedure, the parameters were analysed and other empirical tests were conducted in order to reach the final form of the XGBTree algorithm employed in this paper, which is presented in Table 6. This permutation of the model presents a CV score of AUC = 0.832, which is a non trivial improvement compared to the RF model.

Hyper-Parameter	Value
nrounds	5000
max_depth	3
eta	0.01
gamma	0.1
colsample_bytree	20%
min_child_weight	3
subsample	100%

Table 6: Final parameters of the XGBTree model. Source: Self-Elaboration

Just like in the case of RF, caret implemented XGBTree models come with an integrated feature importance metric which was plotted in Figure 24. The similarity with Figure 23 is evident, with many of the most relevant predictors being the same and in a similar order. This second plot confirms the paramount importance of the number of cars and two-wheeled vehicles as well as the fragility of pedestrians on the road.

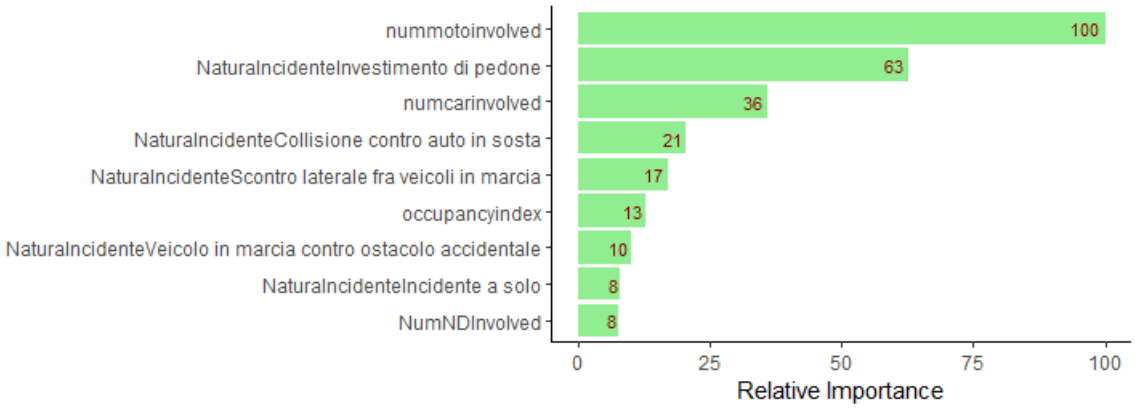

Figure 24: Variable importance plot for the XGBTree model. Source: Self-Elaboration

4.4 Performance Evaluation

After extensive model validation the performances of the employed models were the ones in the following table:

Model	Validation AUC
KNN	0.808
CART	0.746
RF	0.811
XGBTree	0.832

Table 7: Validation performances of the various models. Source: Self-Elaboration

The XGBTree model is clearly the superior, displaying a 2.6% increase in AUC compared to the second best model (RF) and a 11.5% increase with respect to the worst performing model (CART). Due to its superiority in validation, XGBTree was also used for making predictions on the testing set, composed of 4,530 previously unseen crash observations.

Figure 25 shows the achieved AUC of 0.842 in the testing set which surpasses the validation AUC by a non-trivial amount, confirming once again the great capabilities of XGBTree models in avoiding over-fitting, which facilitate the generalization to new data.

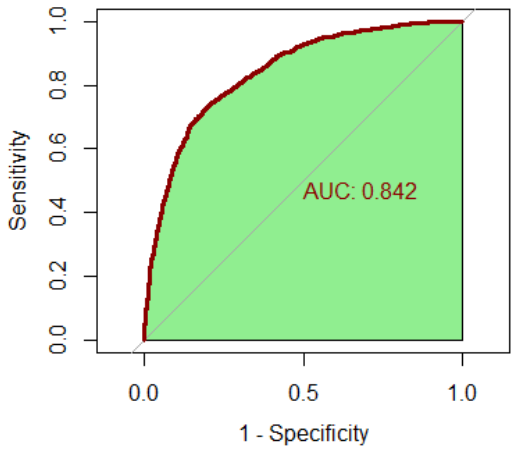

Figure 25: XGBTree model testing AUC. Source: Self-Elaboration

5 CONCLUSIONS

5.1 Results of the Analysis

The machine learning analysis of Rome road crashes happened in 2021 highlighted how tree based models, and in particular ensemble tree based models like RF and XGBTree, appear to be the best options when trying to model the severity of a car accident. This may be due to their intrinsic flexibility and innate capacity to be adapted to almost any data analysis task but it may also be a direct consequence of the way this kind of models are structured. The fact that ensemble tree based models are designed and optimized to be trained with a subset of the total number of predictors makes it so that they are able to comprehend even very complex phenomena where the variables come from different aspects of the problem and interact differently with one another. Car crashes are a perfect example of this kind of entities since their data include all of the aspects of the crash, from vehicle related features (e.g., vehicle type, age of the vehicle, etc.) to road specific features (e.g., type of road, speed limit on the road, etc.) and accident specific features (e.g., travelling speed, type of accident, people involved, etc.).

In the case of this study, considering the results from all of the models with a particular regard to the XGBTree one, which was the best performing, the most important risk factors in predicting the severity of a crash were the presence of two-wheeled vehicles and, to lesser extent, pedestrians. This result is completely in line with previous findings in the scientific literature and highlights the paramount importance of appropriate measures to safeguard the weakest road users. Other important characteristics for severity prediction were factors related to the entity of the accident, such as the number of cars involved, and factors related to the interaction between the number of people involved and the number of moving vehicles involved (i.e., *occupancyindex*). The type of the accident was also quite relevant with accidents that usually happen at higher speeds or normally involve more cars presenting an higher intrinsic risk.

The results of the analysis were comprehensively more than satisfactory from both a

performance and from an insight perspective, especially considering the questionable quality of the original input data and the computational limits of the employed machines.

5.2 Limitations of the Study

5.2.1 Data Limitations

The limitations of this paper reside mostly in the lack of more complete and coherent input data. Some of the variables that were present in the original dataset and could have been very interesting to analyse presented in fact some critical issues that precluded their study. This was the case for *illuminazione* (lighting condition), *CinturaCascoUtiltizzato* (safety belt and helmet use) and *Airbag* variables, which would seem, at least intuitively, very relevant for severity prediction but their prominent level of *NA* or *Not Specified* values prevented any further analysis on them.

Another limiting factor was the inability to extract more classification levels for crashes from the data. It was in fact impossible to determine or derive a proper severity scale for crashes given the initial data. This shortcoming transformed the analysis in a binomial classification which was definitely successful and coherent with previous research but which however will never be able to produce an understanding of the phenomena as deep as it could have been with more specific data.

The lack of some specific crash features also hindered the possible results of the analysis: the data was lacking any information on the people involved in the accident besides their gender and in particular there was no indication of the age of the participants in the accident, which has been discovered as a very influential factor in previous research (Al Mamlook et al., 2020).

5.2.2 Hardware Limitations

Hardware capabilities were also a limiting factor for this paper. Given the limited computational power of the employed machines, it was impossible to optimally fine tune some specific models (e.g, XGBTree) due to extreme running times of more than twelve hours in

some cases. Some complex and specialized methods such as kernel support vector machines and weighted KNN were inapplicable to the analysis due to repeated and unavoidable system crashes caused by lack of memory or cooling power.

5.3 Future Prospects

Future prospects for this analysis include a reiteration of the analysis process considering a wider variety and quantity of data. In addition it would be beneficial to analyse car accidents having access to a wider number of data sources: the collaborative effort of regulatory authorities, hospitals and local administrations would be needed for the creation of what could be defined as a complete crash dataset including all crash factors and medical records of the participants. Thanks to the interaction between different institutions it would be possible to perform an in depth specialized analysis, employing multinomial classification, in order to understand deeply what are the prominent risk factors for all categories of road users. As stated in Section 1.4, an in depth statistical and machine learning understanding of car accidents would provide enormous benefits for society, especially for the end users of the road system, which would be assisted by insurance companies and hospitals that are as prepared as possible in assessing risks and complications of each individual case.

REFERENCES

Al Mamlook, R. E., Abdulhameed, T. Z., Hasan, R., Al-Shaikhli, H. I., Mohammed, I., & Tabatabai, S. (2020). Utilizing machine learning models to predict the car crash injury severity among elderly drivers. *2020 IEEE international conference on electro information technology (EIT)*, 105–111.

AlMamlook, R. E., Kwayu, K. M., Alkasisbeh, M. R., & Frefer, A. A. (2019). Comparison of machine learning algorithms for predicting traffic accident severity. *2019 IEEE Jordan international joint conference on electrical engineering and information technology (JEEIT)*, 272–276.

Bagui, S., Nandi, D., Bagui, S., & White, R. J. (2021). Machine learning and deep learning for phishing email classification using one-hot encoding. *Journal of Computer Science*, *17*, 610–623.

Beshah, T., Ejigu, D., Abraham, A., Snasel, V., & Kromer, P. (2013). Mining pattern from road accident data: Role of road user's behaviour and implications for improving road safety. *International journal of tomography and simulation*, *22*(1), 73–86.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*, 197–227.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Burdett, B., Li, Z., Bill, A. R., & Noyce, D. A. (2015). Accuracy of injury severity ratings on police crash reports. *Transportation research record*, *2516*(1), 58–67.

Chang, L.-Y., & Chien, J.-T. (2013). Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model. *Safety science*, *51*(1), 17–22.

Chang, L.-Y., & Wang, H.-W. (2006). Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis & Prevention*, *38*(5), 1019–1027.

Cortes, C., & Mohri, M. (2003). Auc optimization vs. error rate minimization. *Advances in neural information processing systems*, *16*.

Duboue, P. (2020). *The art of feature engineering: Essentials for machine learning*. Cambridge University Press.

Evwiekpaefe, A., & Umar, S. M. (2021). Predicting road traffic crash severity in kaduna metropolis using some selected machine learning techniques. *Nigerian Journal of Technology*, *40*(5), 888–900.

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). Knn model-based approach in classification. *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*, 986–996.

Hastie, T., & Tibshirani, R. (1997). Classification by pairwise coupling. *Advances in neural information processing systems*, *10*.

Hosseinzadeh, A., Moeinaddini, A., & Ghasemzadeh, A. (2021). Investigating factors affecting severity of large truck-involved crashes: Comparison of the svm and random parameter logit model. *Journal of safety research*, *77*, 151–160.

Howard, J., & Bowles, M. (2012). The two most important algorithms in predictive modeling today. *Strata Conference presentation, February*, *28*.

INRIX. (2022). Global traffic score card. https://inrix.com/scorecard/

Iranitalab, A., & Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis & Prevention*, *108*, 27–36.

ISTAT. (2022). Report incidenti stradali 2021. https://www.istat.it/it/archivio/273324

ISTAT. (2023). Report incidenti stradali 2022. https://www.istat.it/it/archivio/286933

Laiou, A., Papadimitriou, E., Yannis, G., & Milotti, A. (2017). Road safety data and information availability and priorities in south-east european regions. *Transportation research procedia*, *25*, 3703–3714.

Lee, J., Yoon, T., Kwon, S., & Lee, J. (2019). Model evaluation for forecasting traffic accident severity in rainy seasons using machine learning algorithms: Seoul city study. *Applied Sciences*, *10*(1), 129.

Lin, C., Wu, D., Liu, H., Xia, X., & Bhattarai, N. (2020). Factor identification and prediction for teen driver crash severity using machine learning: A case study. *Applied Sciences*, *10*(5), 1675.

Lin, M.-R., & Kraus, J. F. (2008). Methodological issues in motorcycle injury epidemiology. *Accident Analysis & Prevention*, *40*(5), 1653–1660.

Loh, W.-Y. (2011). Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *1*(1), 14–23.

Mehra, N., & Gupta, S. (2013). Survey on multiclass classification methods. *International Journal of Computer Science and Information Technologies*, *4*(4), 572–576.

MIT. (2018, March). Classification. In *Machine Learning for Data Streams: with Practical Examples in MOA*. https://doi.org/10.7551/mitpress/10654.003.0011

Mokhtarimousavi, S., Anderson, J. C., Azizinamini, A., & Hadi, M. (2020). Factors affecting injury severity in vehicle-pedestrian crashes: A day-of-week analysis using random parameter ordered response models and artificial neural networks. *International journal of transportation science and technology*, *9*(2), 100–115.

Pillajo-Quijia, G., Arenas-Ramírez, B., González-Fernández, C., & Aparicio-Izquierdo, F. (2020). Influential factors on injury severity for drivers of light trucks and vans with machine learning methods. *Sustainability*, *12*(4), 1324.

Rifkin, R., & Klautau, A. (2004). In defense of one-vs-all classification. *The Journal of Machine Learning Research*, *5*, 101–141.

scikit-learn.org. (n.d.). https://scikit-learn.org/stable/_images/grid_search_cross_validation.png

Stevenson, M., Segui-Gomez, M., Lescohier, I., Di Scala, C., & McDonald-Smith, G. (2001). An overview of the injury severity score and the new injury severity score. *Injury Prevention*, *7*(1), 10–13.

Tang, J., Liang, J., Han, C., Li, Z., & Huang, H. (2019). Crash injury severity analysis using a two-layer stacking framework. *Accident Analysis & Prevention*, *122*, 226–238.

Tarko, A. P., Bar-Gera, H., Thomaz, J., & Issariyanukula, A. (2010). Model-based application of abbreviated injury scale to police-reported crash injuries. *Transportation research record*, *2148*(1), 59–68.

Tsui, K., So, F., Sze, N.-N., Wong, S., & Leung, T.-F. (2009). Misclassification of injury severity among road casualties in police reports. *Accident Analysis & Prevention*, *41*(1), 84–89.

UnitedNations. (2020). *A/res/74/299: Resolution adopted by the general assembly on 31 august 2020*. https://documents-dds-ny.un.org/doc/UNDOC/GEN/N20/226/30/PDF/N2022630.pdf?OpenElement

Vajari, M. A., Aghabayk, K., Sadeghian, M., & Shiwakoti, N. (2020). A multinomial logit model of motorcycle crash severity at australian intersections. *Journal of safety research*, *73*, 17–24.

Wang, X., & Kim, S. H. (2019). Prediction and factor identification for crash severity: Comparison of discrete choice and tree-based models. *Transportation research record*, *2673*(9), 640–653.

WHO. (2022). Road traffic injuries. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

Zajac, S. S., & Ivan, J. N. (2003). Factors influencing injury severity of motor vehicle–crossing pedestrian crashes in rural connecticut. *Accident Analysis & Prevention*, *35*(3), 369–379.

Zeng, G. (2014). A necessary condition for a good binning algorithm in credit scoring. *Applied Mathematical Sciences*, *8*(65), 3229–3242.