






## Article

# RoadLytics: Road Accidents Analytics Using Artificial Intelligence to Support Deaths' Prevention on Highways

Kelvin Rinaldi da Luz <sup>1</sup> , João Elison da Rosa Tavares <sup>1</sup> , Jorge Luis Victória Barbosa <sup>1</sup> , Daniel Hernández de la Iglesia <sup>2</sup>  and Valderi Reis Quietinho Leithardt <sup>3</sup> 

<sup>1</sup> Applied Computing Graduate Program (PPGCA), University of Vale do Rio dos Sinos (UNISINOS), Av. Unisinos 950, São Leopoldo, Rio Grande do Sul, Brazil; kelvin.rinaldi@hotmail.com, joaoer@unisinos.br, jbarbosa@unisinos.br

<sup>2</sup> Faculty of Informatics, Universidad Pontificia de Salamanca, C/Compañía 5, 37002 Salamanca, Spain; dhernandezde@upsa.es

<sup>3</sup> VALORIZA – Research Centre for Endogenous Resource Valorization, Polytechnic Institute of Portalegre, Portugal; valderi@ipportalegre.pt

**Abstract:** Daily thousands of people and goods move along Brazilian Federal highways. Traffic accidents are numerous on these highways and have a significant impact, whether on the economy or the health system. Identifying predictor variables, the probability of an event occurring and how to mitigate them are of paramount importance for the actions of the transit authorities that manage these roads. The main contribution of this study is the development of a predictive machine learning model which uses open data to show graphically the critical points in the highways. This model is fully reproducible and can be applied to any region worldwide helping to minimize the number of accidents and to prevent deaths by automotive collisions. For this study, 43 variables were analyzed supporting the identification of the causes of accidents with fatal victims on the main highways in the south of Brazil. RoadLytics is proposed as a supervised machine learning model, using the Random Forest algorithm to analyze about 33 thousand occurrences between 2017 and 2020. An exploratory analysis of the data was carried out to support the modeling and to facilitate data visualization. In this sense, heat maps were developed to support the analysis and identification of potential risk areas. The results show that BR386 highway registers the highest number of fatal occurrences, regardless of the season. Additionally, concerning the weather conditions, the analysis shows that 52% of accidents occurred in favorable conditions, such as clear skies, victimizing 501 people. The driver's lack of attention is the main reason for the accidents' occurrences. Applying the developed model, an accuracy of 77% was achieved for the classification of fatal accidents.

**Keywords:** Accidents, Data Analysis, Machine Learning, Transport



**Citation:** Da Luz, K.; Tavares, J.; Barbosa, J.; De la Iglesia, D.; Leithardt, V. RoadLytics: Road Accidents Analytics Using Artificial Intelligence to Support Deaths' Prevention on Highways. *Preprints* **2021**, *1*, 0. <https://doi.org/>

Received:

Accepted:

Published:

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## 1. Introduction

According with World Health Organization (WHO) [1], around of 1.35 million people died every year around the world as a result of road traffic crashes. These accidents penalize the countries with more than 3% of their Gross Domestic Product (GDP). Besides, road traffic injuries are the leading cause of death for children and young adults aged 5-29 years. Around the world the concern about traffic accidents has become increasingly greater as this phenomenon has emerged as a public health problem [2].

The problem related to road accidents in Brazil is a direct cause of the high costs demanded by the Unified Health System in Brazil, since it corresponds to a significant portion of care and prolonged treatments. Traffic accidents are also the cause of other costs with direct and indirect social and economic impacts [3].

Inada *et al.* [4] describe that even during the COVID-19 pandemic, Japan identified an increase in the number of deaths caused by traffic accidents. Empty roads possibly triggered speed-related traffic violations that caused fatal Motor Vehicle Collisions (MVCs). In the same way, in São Paulo, Brazil, this increase may be related to issues such as the low number of available professionals, medications, oxygen and Intensive Care Units (ICU)

beds. The number of injured motorcyclists also increased, which may be related to the delivery option that became common during the pandemic [5].

For a decrease in the volume of accidents, minimizing the impacts on the health system, it is necessary to understand the causes that lead to the occurrences. Currently, several data regarding accidents that occur on Brazilian Federal highways are recorded and made available publicly by the Federal Highway Police (FHP) [6].

The analysis of data from these accidents allows to extract information, understanding the reason and which variables that combined can impact the risk of an accident or fatality. However these datasets are voluminous and the analysis can be a human difficult task. In this case, mainly the visualization of correlations and implicit relationships between the variables involved can be performed using data mining techniques [7,8].

Data mining of open data can be applied through the use of Machine Learning (ML) techniques. Antonopoulos *et al.* [9] describe machine learning as a subarea of artificial intelligence whose adoption has grown exponentially in recent years. These are mathematical, statistical and computational algorithms that are capable of carrying out an inference process through learning based on examples [10]. The ML as well as Deep Learning (DL) techniques have been applied in a wide diverse areas, such as Healthcare [11–15], Business [16] and Agriculture [17,18]. However, there are still opportunities for application of ML and Open Data for traffic accidents prediction [19], mainly in the context of Smart Cities paradigm [20–23].

Machine Learning can be categorized mainly in two different ways, supervised and unsupervised learning. Supervised learning is the task of finding a role from labeled training data. The goal is to find the most optimized parameters from tests that can predict labels for new objects. In unsupervised learning there is usually less information about the object, being very difficult to carry out training using historical data. Therefore, when using unsupervised learning, non-labeled data is evaluated, which makes it difficult to identify patterns between independent variables for the model, making it necessary to group existing data so that we can then identify patterns. Among the algorithms for this type of learning, it can be cited K-Means, hierarchical clustering and artificial neural networks [24].

Random Forest (RF) is a supervised learning algorithm that is considered versatile due to the ability to be used in both classification and regression tasks. When used, the algorithm creates a combination of decision trees in order to obtain a prediction with greater accuracy and more stability [25]. Lin *et al.* [26] reinforce that RF algorithm should be used in different data analyzes, due to its origin and decision trees, which makes it easy to understand. The algorithm presents good results in different types of analysis and has implementations in the most diverse artificial intelligence libraries available.

The data analyzed during throughout this article are available in a public domain. Gewin [27] describes that the concept of open data can be applied for different sources of information and different topics, and any person or organization can publish data for the community. However, what stands out most is the data made available by the government, with information about budgets, transportation, culture, science, finance and climate. Veljković *et al.* [28] highlight two elements regarding open data, the first is the legal access to the data, the institution cannot create barriers for the data to be accessed by anyone who has an interest in the information. In addition, it is necessary that access is facilitated, in this way the data needs to be made available in a format that is easy to interpret and absorb, avoiding, for example, images and PDF files, and giving preference to data in bulk format.

The main contribution of this study is the development of a predictive machine learning model which uses open data to shows graphically the critical points in the highways. This model is fully reproducible and can be applied to any region worldwide helping to minimize the number of accidents and to prevent deaths by automotive collisions. For the experimentation proposes the model was applied to analyze the probability of the occurrence of fatal accidents on federal roads in the southern state of Brazil, Rio Grande do

Sul (RS), and to analyze the variables that may contribute to the occurrence of this type of event. This state was selected for the study due to the possibility of understanding patterns in the traffic of the region that is a major producer of soy [29] and rice [30] in the Brazilian territory and both products are exported to other countries, adding more complexity to the case study scenario.

This article is organized in the following way. Section 2 presents the related works. Section 3 describes the applied methodology, while Section 4 shows the exploratory analysis. Section 5 details the construction of the predictive algorithm and training applied. Finally, section 6 concludes the research with final considerations and future directions.

## 2. Related Works

This section presents related works selected from the search of terms "big data", "data mining", "traffic accidents" and "machine learning", in the MDPI<sup>1</sup>, IEEE Xplore<sup>2</sup>, Research Gate<sup>3</sup> and Scielo<sup>4</sup> repositories. At the end of this section, a summary of the comparison between the works is presented.

Barroso Junior *et al.* [31] analyzed the lethality of traffic accidents on Brazilian Federal highways in 2016, considering, in addition to the characteristics of the victims, information about the context in which these events occurred. During the development of the work, a binomial logistic regression model was used for data analysis, using the R language to adjust the model, which was subsequently subjected to adjustment through the Hosmer and Lemeshow test mechanism. Finally, the study found that, on average, the likelihood of an accident being lethal increases by 44% for males, pedestrians, locations in the northeast region, on Sundays and in rural areas.

Chang and Park [32] lead a study to identify possible fatigued drivers to find patterns related with crash accidents. The study considered GPS informations about the vehicles through a method based on the distribution of the driving duration and the boundary condition of the driving duration between fatigued and non-fatigued state. As a result, the authors identified that the fatigue data measured was a strong explanatory power with regard to the traffic accident rate, with a statistical correlation of 0.86 at least.

von Buxhoeveden and Becker [33] evaluated data on accidents involving trains, buses, ships, cars and planes, using different countries as sources, such as Switzerland, where they collected public transport data, Germany, where car accidents were collected and the United States in which air accidents were analyzed. The solution used during the development of the work was R, which served as a platform for the development of a web application for data exploration through graphics and heat maps as a final result, which can be used to cross-check information on public transport accidents and private, and may present points of greater occurrences and risks of future accidents.

Lamr [34] analyzed the data collected by Czech police searching for patterns in the accidents and generate information for warning the government agencies. During the development of the article the APRIORI algorithm was used for analyzed the data and gathering clusters in heat maps. An exploratory analysis was executed as well.

Wang *et al.* [35] worked on a neural network for analyze the traffic issues related with truck vehicles using its GPS signal. The work was developed in two phases, first the expansion phase where errors and omissions were treated, then the prediction phase, where the Long Short Term Memory (LSTM) and Gated Recursive Unit (GRU) neural network methods were applied to improve the accuracy. The data used during this work was from Zhengzhou city in China. As a result, the authors conclude that the LSTM approach was a better result than GRU with the accuracy.

Zhang and Hassan [36] worked on the evaluation of accidents in the Egyptian high-ways during the night, where the severity of the injuries cause by the crashes was evaluating

<sup>1</sup> <https://www.mdpi.com>

<sup>2</sup> <https://ieeexplore.ieee.org/Xplore/home.jsp>

<sup>3</sup> <https://www.researchgate.net/>

<sup>4</sup> <https://scielo.org/>

using a multinomial logit model and an exploratory analysis was made. As a result from the model, the authors conclude that the rainy conditions tend to increase the fatality rate during the night crashes, the authors also identified a relationship between male drivers and accidents related to high speed. The age factor is also a important factor, where the young drivers have a high probability of begin involved.

Mokoatle and Marivate [37] used road traffic data accident in Soshanguve, South Africa, to applied an exploratory analysis and a cluster analysis to identify the main reasons for the occurrences. As a result from the analysis, the authors identified that the most serious injuries are related with heavy vehicles, as trucks and buses.

Mazouri *et al.* [38] applied a different approach in the analysis of France road accidents, used the FP-growth algorithm combine with Spark framework to help in the identification and extraction of associations rules. This associations rules, after been identified can be helpful for the decisions makers to choose the best approach and strategy in the road administration.

The study developed by Chen [39] evaluated the public data regarding accidents in the city of Shanghai, China between the years 2015 to 2016 through the R, where he first performed the treatment of data, through the cleaning of non-useful data, data formatting and converting the accident location to latitude and longitudinal measurements facilitating the creation of heat maps to show the distribution of accidents across the city. The author developed three models to predict the probability of a certain accident category occurring, these models were validated through a dataset that was divided between training and testing. For the creation of the models, the linear discriminant analysis algorithms, random forest and decision trees were used. The discriminant linear analysis algorithm resulted the best accuracy of the model, random forest was better in the integration of different categories, and finally, decision trees obtained the best result in the kappa coefficient, which is used to measure the reliability of the model.

The last related work analysed was the article of Zůvala *et al.* [40]. In this article, the authors considered the data accident of Czech Republic using the Czech In-depth Accident Study (CzIDAS) and the data collected by the Police of the Czech Republic using a sample of the dataset of each source. The main objective was to generate information about the location and patterns of drivers, which was possible using both dataset, the secondary objective was validate both dataset, and the authors could conclude that they were comparable with each other, which can indicate an accurate collection of the information.

Table 1 presents a comparative analysis between the works related to this article. Each of selected article was analyzed considering the most relevant features for this study, such as the use of heat maps to facilitate the visualization of information, execution of Exploratory Data Analysis (EDA) over the analyzed dataset, if any machine learning technique is applied, in addition to having observed the application restriction techniques in the dataset and public access to this information. These items are marked as yes, when they are present, and as no, when they are not present.

**Table 1.** Comparative analysis between related works.

Related Work	Map Presentation	EDA	Machine Learning	Data Restriction	Open Data
Barroso Junior <i>et al.</i> [31]	No	Yes	Yes	No	Yes
Chang and Park [32]	Yes	Yes	No	No	Yes
von Buxhoeveden and Becker [33]	Yes	Yes	No	Yes	Yes
Lamr [34]	Yes	Yes	No	No	No
Wang <i>et al.</i> [35]	Yes	No	Yes	Yes	Yes
Zhang and Hassan [36]	No	Yes	No	No	Yes
Mokoatle and Marivate [37]	No	Yes	No	Yes	Yes
Mazouri <i>et al.</i> [38]	No	No	Yes	No	Yes
Chen [39]	Yes	No	Yes	Yes	Yes
Zůvala <i>et al.</i> [40]	Yes	Yes	No	Yes	Yes
RoadLytics	Yes	Yes	Yes	Yes	Yes

Based on the analysis of the related works, the identification of topics for the research emerged, such as the use of public data, which enables to continue new research from the same source of the data, which also guarantees authenticity in the analysis. Among the main points, the restriction of fatal data is one of improvement identified, which can present more detailed information about the occurrences. In addition, the use of heat map allows to present information in a visual way, applying data balancing techniques, aiming at minimize the discrepancy in the data analyzed.

### 3. Methods and Procedures

This section details the procedures adopted for conducting the research, from data collection to the restriction for further analysis and presentation of results.

#### 3.1. Work Delimitation

Traffic problems are part of people's daily lives and have a significant impact on the economy and public health. To conduct this study, the experiments used public accident data on RS federal highways, filtering between the years 2017 and 2020, these datasets are available by the FHP of Brazil [6]. Figure 1 shows a snippet of the dataset, which is fully available at FHP Open Data Repository<sup>5</sup>.

	id	type_accident	origin_accident	day_week	type_involved	age	latitude	longitude	track_type
1	8	Queda de ocupante de veículo	Fenômenos da Natureza	domingo	Condutor	19	-23,09880731	-52,38789369	Simple
2	9	colisão com objeto estático	Falta de Atenção à condução	domingo	Condutor	35	-27,8101	-48,6357	Dupla
3	11	Capotamento	Animais na Pista	domingo	Condutor	27	-23,36951985	309,93513107	Simple
4	11	Capotamento	Animais na Pista	domingo	Passageiro	27	-23,36951985	309,93513107	Simple
5	12	Tombamento	Avarias e/ou desgaste excessivo no pneu	domingo	Condutor	24	-16,27473677	-48,96908998	Dupla
6	13	saída de leito carroçável	Ingestão de Alcool	domingo	Condutor	57	-26,44675249	-49,20166969	Simple
7	14	colisão traseira	Falta de Atenção à condução	domingo	Condutor	35	-16,82489647	-49,53520775	Dupla
8	14	colisão traseira	Não guardar distância de segurança	domingo	Condutor	35	-16,82489647	-49,53520775	Dupla
9	14	colisão traseira	Falta de Atenção à condução	domingo	Condutor	35	-16,82489647	-49,53520775	Dupla
10	14	colisão traseira	Não guardar distância de segurança	domingo	Condutor	35	-16,82489647	-49,53520775	Dupla
11	15	Tombamento	Ingestão de Alcool	domingo	Condutor	49	-6,4622	-36,1899	Simple
12	16	colisão traseira	Ingestão de Alcool	domingo	Pedestre	45	-7,18668515	-48,22970957	Múltipla
13	16	colisão traseira	Não guardar distância de segurança	domingo	Pedestre	45	-7,18668515	-48,22970957	Múltipla
14	16	colisão traseira	Ingestão de Alcool	domingo	Pedestre	45	-7,18668515	-48,22970957	Múltipla
15	16	colisão traseira	Não guardar distância de segurança	domingo	Pedestre	45	-7,18668515	-48,22970957	Múltipla

Figure 1. Rows sample of traffic accidents' dataset

#### 3.2. Research Stages

Figure 2 shows the steps performed during this work, and then each step is described in detail.

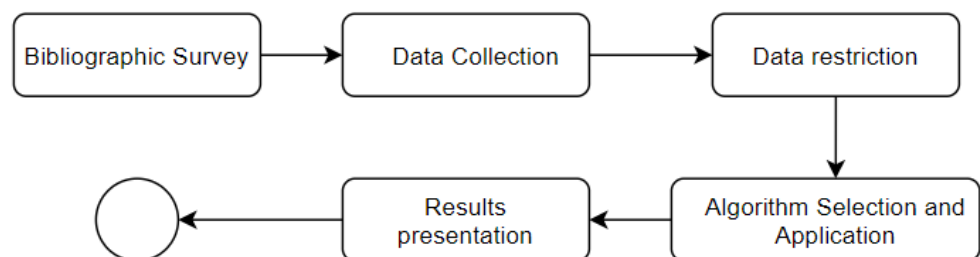


Figure 2. Flowchart of research steps

The following list details the steps performed in the presented pipeline. The items show the specification of actions presented in each step:

- *Bibliographic survey*: Selection of methods to be applied in the database selected for this research;
- *Data collection*: Load of Federal Highway Police records for the period from 2017 to 2020;
- *Data restriction*: In this step, cleaning was performed on erroneous data or that would not be used during the analysis, as described in section 3.4;

<sup>5</sup> <https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos>



- *Algorithm selection and application*: Evaluation and study of machine learning algorithms to perform the analysis of previously treated data;
- *Results presentation*: Result of tests performed and knowledge discoveries obtained through graphs, tables and heat maps, in addition to opportunities and recommendations for improvement for future analyzes.

### 3.3. Data Collection

Specifically, the data collection considered the public data between the years 2017 and 2020, since from the year 2017 the information provided contains more attributes regarding the occurrences, the data were grouped by person, cause and type of accident. After selecting the data, the use of R language by applying a filter to generate a new mass of data containing information only from RS/Brazil. This state was chosen for the relevance that the study can provide for the region that is a major producer of soy and rice in the Brazilian territory and where the research group is located.

### 3.4. Restriction of Data

The treatment of raw data<sup>6</sup> was carried out using the R language, with the data loading from CSV files. This step enables to remove inconsistent, duplicate or data that would not be useful during the analysis.

Firstly, the process considered the concatenation of files for the years 2017, 2018 and 2020 into a single dataset in R. Altogether, the original dataset had 37 variables. After the creation of the new data structure, a filter was applied to the *UF* variable, to return only the records related to the state of Rio Grande do Sul, which corresponds to the scope of this work.

In the *day\_week* field, which stores the day on which the occurrence was recorded, the values were transformed from text to a sequential integer, starting with sunday until saturday. A filter in the *gender* field was also used, to eliminate records where the value entered was "Not Informed" or "Ignored".

In the *latitude* and *longitude* fields, the conversion of the values that contained a comma to a point was performed, aiming at facilitating the identification of the coordinates of the occurrences.

The process considered the addition of a new column called *AgeRange*. This feature corresponds to the age group of the people involved in the accidents, where three intervals were defined, according to WHO (World Health Organization) definitions:

- Value 1: It corresponds to the young age group, people between 0 and 19 years old;
- Value 2: It corresponds to the adult age group, people between 20 and 64 years old;
- Value 3: It corresponds to the elderly age group, people aged 65 or over.

Through the *vehicle\_manufacture\_year* field, a filter proportionates the return of the occurrence records only when the date of manufacture of the vehicle is equal to or greater than 1960. This filter was applied because it was identified that vehicles with the year of manufacture had incomplete information, such as name and model.

For the types of vehicles involved in the accidents, the addition of a new field facilitate the classifications, therefore based on *type\_vehicle* field, the *group\_vehicle* field was created, respecting the rules described below:

- Passenger Transport: All vehicles of the automobile, van, utility, bus, minibuss, scooter, motorcycle, tricycle and moped were classified as means of transport for passengers;
- Cargo Transportation: All vehicles of the truck, pickup, trailer and semi-trailer type were classified as cargo transportation;
- Traction: All wheeled tractor and tractor truck vehicles were classified as traction transport.

<sup>6</sup> <https://git.io/JtEXK>

The added *health\_condition\_seq* field corresponds to a sequential according to the severity of the person involved in the accident. This field acts as a filter throughout the development of the research, respecting the following logic:

- Value 1: When the person's condition is classified as unharmed;
- Value 2: When the person's condition is classified as minor injury;
- Value 3: When the person's condition is classified as a serious injury;
- Value 4: When the person's condition is classified as death.

Two date fields were created, *month\_occurrence* and *year\_month\_occurrence*, to facilitate later filtering by dates, these fields were created from the extraction of the month and year in the numeric format of the field *data\_inversa*.

Finally, the variable *season* was created, so that it was possible to group the data by different seasons in order to analyze them. This field was created after analyzing the month of the accident.

Each accident may have one or more events associated with the same passenger, so to avoid duplication of information, only the first accident event was considered, which is marked as the main cause. A practical example of this scenario can be described as a person who loses control and hits the curb and dies. In the original dataset there are two records of the same accident, which may lead one to believe that two people died, when in fact it was just one. After the filter to return only the first event, we had a 44% reduction after this adjustment.

Originally the database had 101,293 records, after applying the above filters, remained a total of 43,945 records and 43 columns in the sanitized dataset.

#### 4. Exploratory Analysis Results

This section presents the results obtained through the exploratory analysis of the data, regarding recorded accidents, fatal cases and the heat maps.

##### 4.1. Accident Analysis

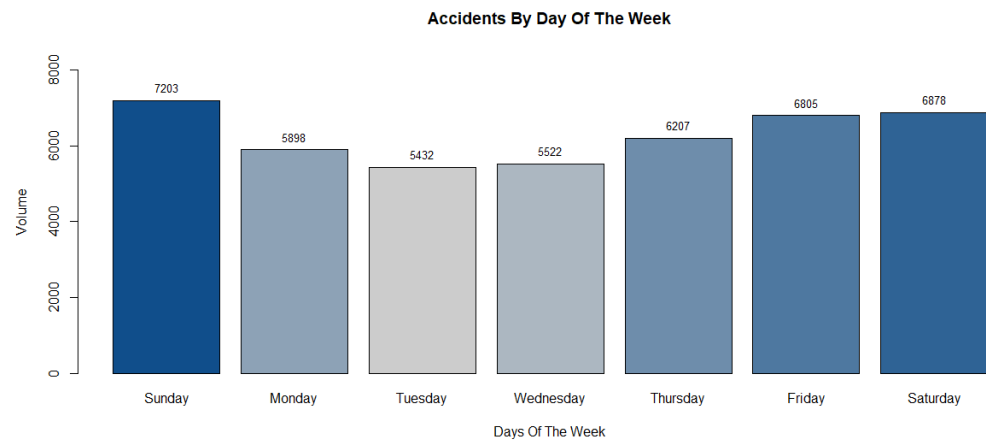
The graphics and observations<sup>7</sup> recorded in this subsection refer to the accident records in their entirety, regardless of whether it was fatal or not, which represents a total of 33,941 records. The colors of the graphs change according to the volume of occurrences, the dark blue tones, point to a greater volume of occurrences.

When grouping the data by gender, it is identified that practically three quarters of the occurrences are linked to male drivers or passengers, 73.6%.

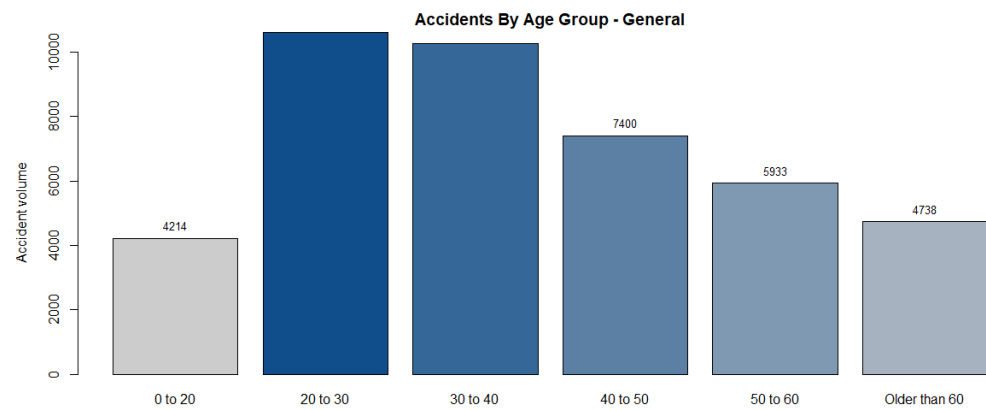
The evaluation by days of the week enables to identify from the data collected that on weekends there is an increase in the number of records, especially on Friday and Sunday, days commonly used for the displacement of many people during trips, as shown in Figure 3.

Figure 4 shows that among the main victims of accidents, the age groups between 20 and 40 years old, with the interval between 20 and 30 being where the rate is higher. This value may be related to the beginning of adulthood, where it is common to obtain a driver's license, we can also mention that it is common knowledge that in this age group the risks of accidents are greater, often caused by recklessness.

<sup>7</sup> <https://git.io/JtEXi>

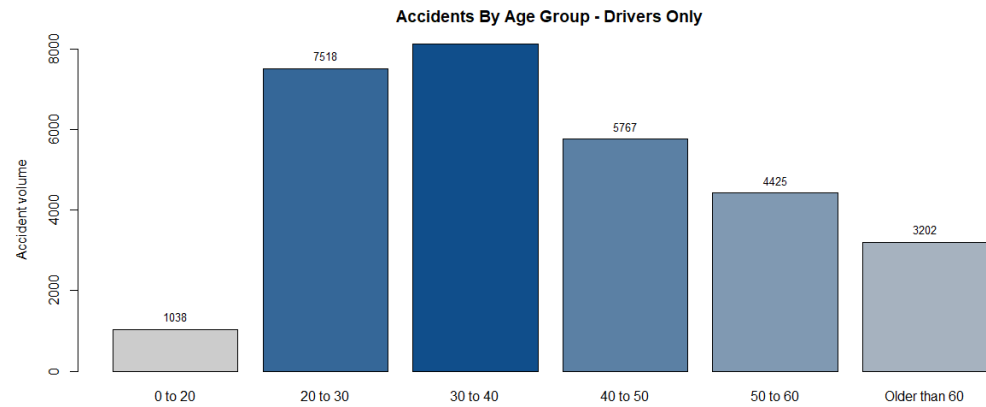


**Figure 3.** Number of accidents per day of the week



**Figure 4.** Number of accidents by age group

Figure 5 shows the data are presented in a similar way to the previous graph, but now listing only the cases in which the person involved in the occurrence was the driver. The behavior of the distribution of accidents remains the same, where for the age group of 20 to 40 years old there is the highest volume of occurrences.



**Figure 5.** Number of accidents by age group - Drivers only



4.2. Analysis of Fatal Occurrences

This section presents the events that resulted in fatalities. When analyzing the weather conditions related to fatal accidents, illustrated in Figure 6, which shows that more than half of the occurrences were during stable weather conditions. It should be noted that the information regarding the weather condition is described by the police officer who recorded the occurrence.

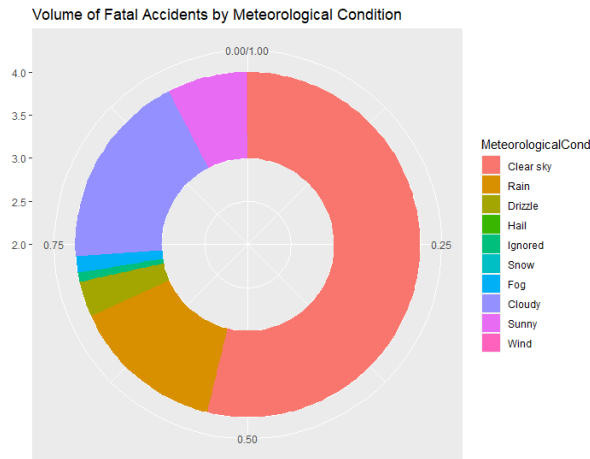


Figure 6. Distribution of fatal accidents by weather

Figure 7 presents all causes linked to the occurrences, highlighting the driver’s lack of attention in traffic, which caused the accident, followed by the driver’s disobedience to traffic rules. It is emphasized that one more cause may be linked to the same occurrence.

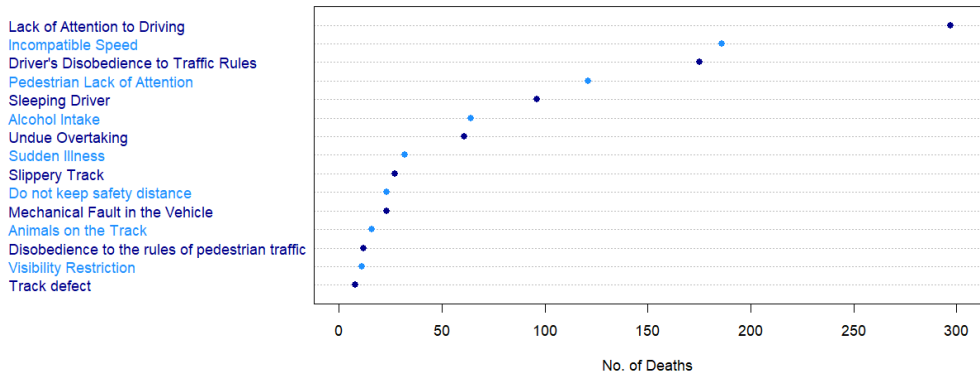
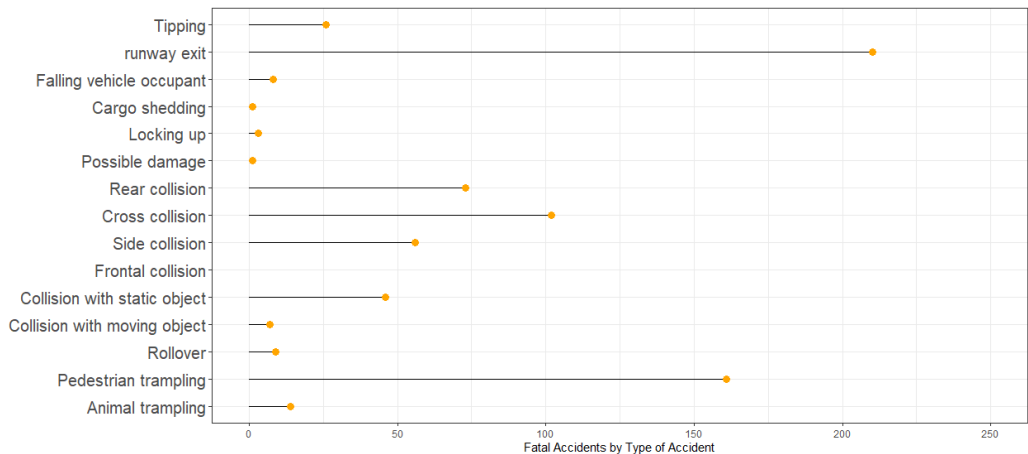


Figure 7. Fatal accidents due to the origin of the accident

Among the main types of accidents that led to death, it can be identified that collisions and the exit of the road are the types with the highest fatality rate, as shown in Figure 8.



**Figure 8.** Fatal accidents by type of accident

4.3. Heat Maps For Analysis of Fatal Events

Fatal accidents are grouped by the season in which they occurred. Using heat maps<sup>8</sup>. Figure 9 shows the distribution of fatal accidents by the state for each season of the year. It turns out, the distribution is similar in the 4 seasons. For each season a table was added with the number of accidents and the percentage for the 5 highways with the highest number of occurrences.

By analyzing in detail the fatal accidents that occurred in the summer, we can see that 173 occurrences are related to accidents with passenger vehicles, while the cargo transport is linked to 39 occurrences and the traction vehicle to 12. Table 1 shows the distribution of the 5 highways that most recorded fatalities in the summer.

Table 1: Summer Season - Highways with higher fatality rates

Federal Highway	Quantity	Percentage
BR386	48	21.15%
BR290	41	18.06%
BR116	31	14.10%
BR285	26	11.45%
BR158	15	6.61%

In the fall, again there is a greater representation of passenger vehicles in fatal accidents, a total of 189 cases, against 32 in cargo vehicles and 12 in traction vehicles. Among the highways with the highest number of fatalities in the fall, there is the distribution on the Table 2.

Table 2: Autumn Season - Highways with higher fatality rates

Federal Highway	Quantity	Percentage
BR386	44	18.18%
BR116	31	14.46%
BR290	29	12.81%
BR392	27	11.16%
BR285	27	11.16%

The analysis of the map of occurrences in winter shows that it is the most distributed throughout the state, with 194 occurrences with passenger vehicles, 33 with cargo vehicles,

<sup>8</sup> <https://git.io/JtEX1>

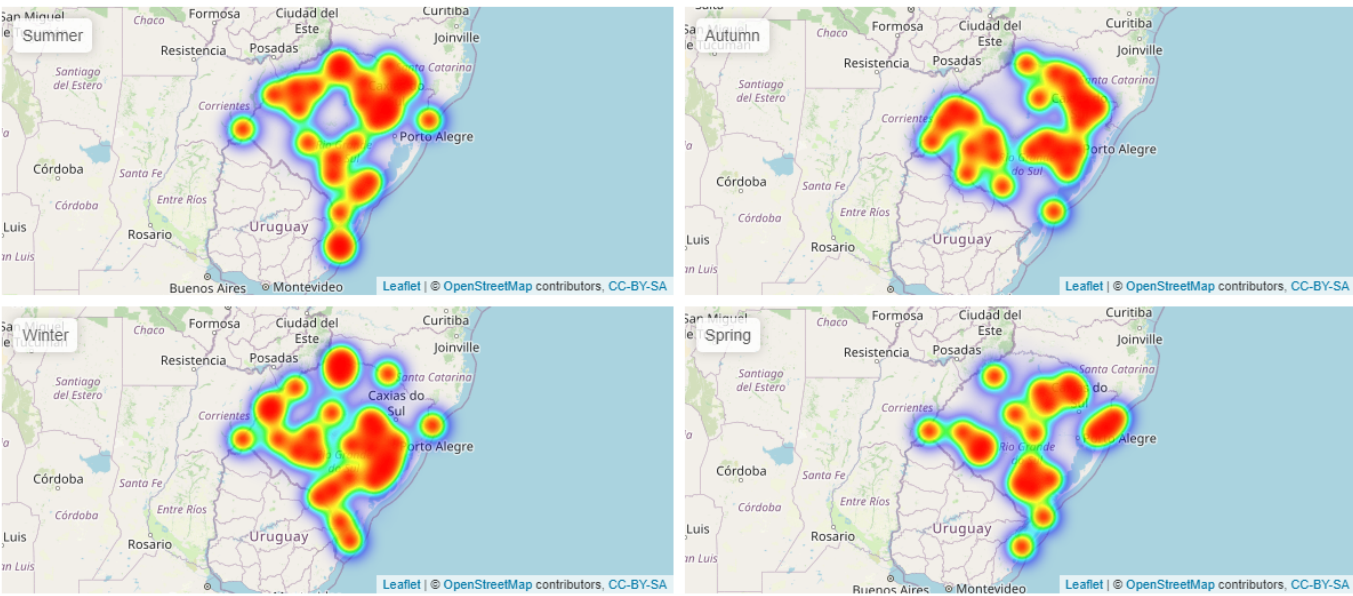


Figure 9. Heat maps of fatal accidents by season

18 with traction and one that was not registered. The distribution by the 5 main highways is presented in the Table 3.

Table 3: Winter Season - Highways with higher fatality rates

Federal Highway	Quantity	Percentage
BR386	54	21.51%
BR116	45	18.73%
BR290	34	13.55%
BR392	22	8.76%
BR285	20	7.97%

During the spring, there were 165 fatal accidents involving passenger vehicles, 43 with cargo vehicles, 8 with traction vehicles and 2 that were not recorded. Table 4 presents the highways with the highest record of fatalities.

Table 4: Spring Season - Highways with higher fatality rates

Federal Highway	Quantity	Percentage
BR386	50	22.12%
BR290	34	15.04%
BR392	31	13.72%
BR116	27	11.95%
BR158	19	8.41%

The heat maps of Figures 10, 11 and 12 show the distribution of fatal accidents by state, organized into 3 maps, one for each type of vehicle group. Here it is worth mentioning that a record was not categorized because it was listed as an ‘uninformed’ vehicle type. Figure 10 shows a distribution of occurrences throughout the RS for passenger vehicles category, which is explained by the higher number of accidents involving this category. For cargo-type vehicles, illustrated in Figure 11, it should be noted that the northern region of the state, where clearly the number of occurrences is greater. This category includes trucks, trucks, trailers and semi-trailers. Lastly, Figure 12 shows the distribution of fatal accidents involving vehicles catego-

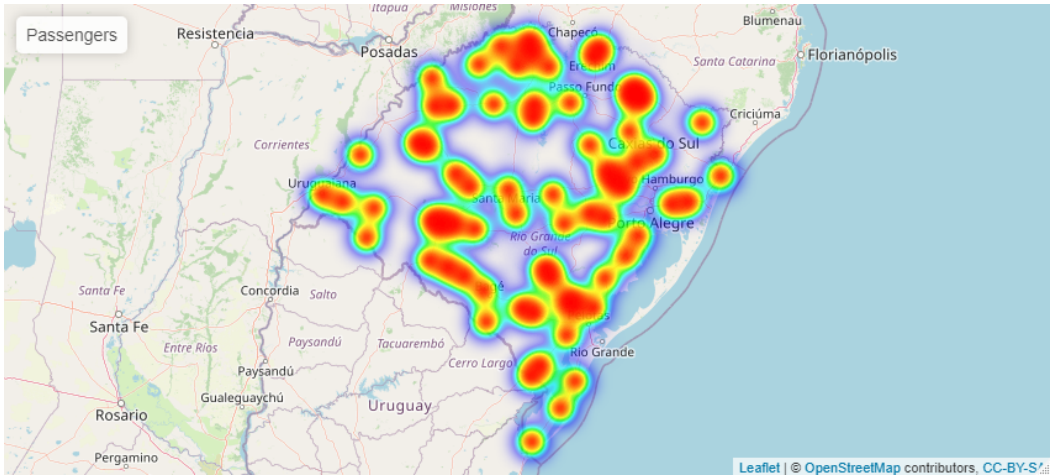


Figure 10. Heat map of fatal accidents in passenger vehicles

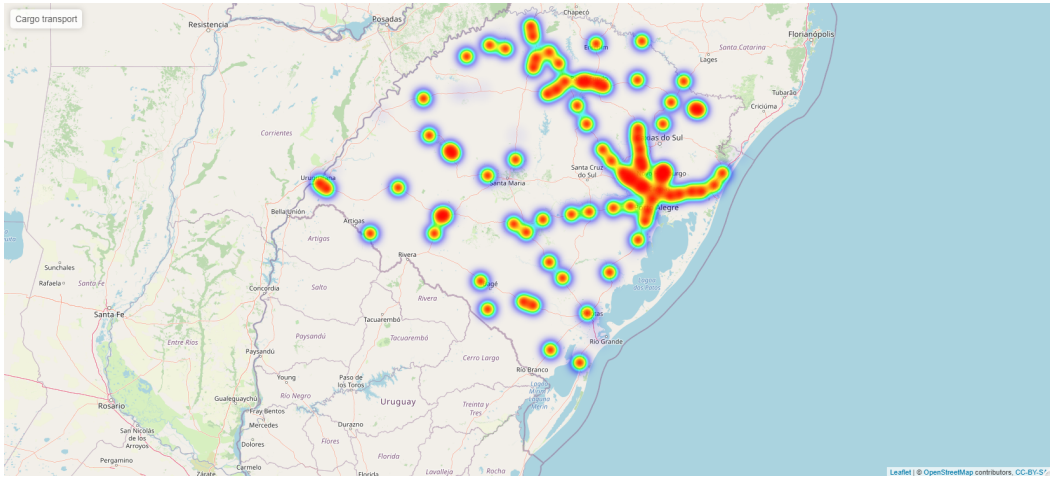


Figure 11. Heat map of fatal accidents in cargo vehicles

alized as being traction, they are, tractor and tractor truck, as can be seen in some areas of the state, as seen that these types of vehicles are often used only in rural areas.

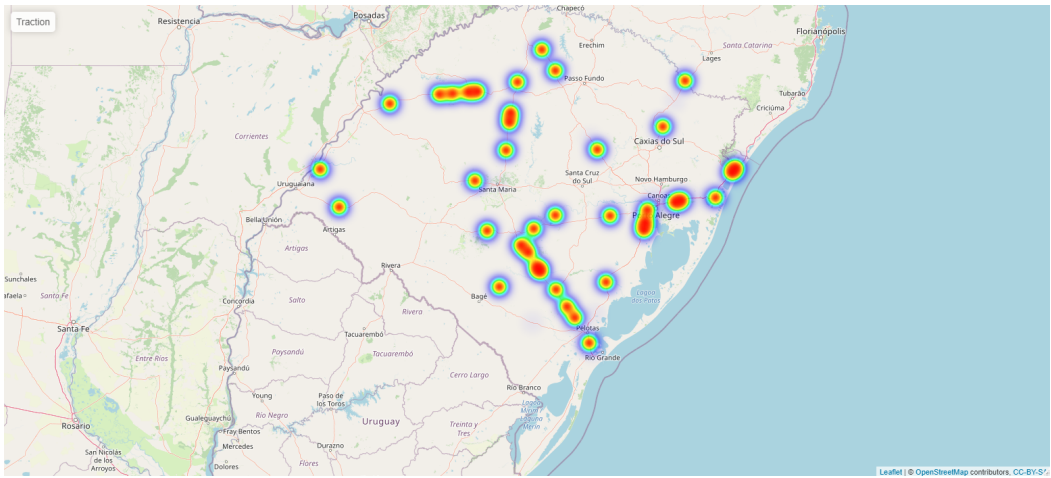


Figure 12. Heat map of fatal accidents in traction vehicles

## 5. RoadLytics Model

The predictive model that aims to classify death or survivor in an accident used the random forest algorithm, using the *type\_accident*, *origin\_accident*, *type\_involved*, *day\_phase*, *track\_layout*, *age*, *track\_type* *gender*, *soil\_use* and variables. Other algorithms were also evaluated, such as Naïve Bayes and decision tree, however when performing the initial validations on the dataset, it was decided to use the random forest because of the algorithm's own characteristics, such as for example, decreased risk of overfitting, since the algorithm works with multiple decision trees.

### 5.1. Data Balancing

When analyzing the original dataset, it was identified that most of the records were linked to the class of people involved who survived the accident, where only 921 died, so when using the dataset with any algorithm in order to create a predictive model, the accuracy will always be close to 100%, which can lead to a misinterpretation of the model when applied to new datasets.

Table 5: Original balance based on dead field

Class	Quantity	Proportion
0 (There was no death)	33020	97%
1 (death)	921	3%

Due to this problem, it was necessary to first rebalance the data, so that there were enough samples and with less discrepancy between the scenarios of people who survived and died. In order for a new dataset to be created, it was decided to select a set of survivor samples 3 times greater than that of deaths, this proportion was considered so that there would be no manipulation of the data in excess, but that it would be sufficient to perform algorithm validations, this technique is known as undersampling.

As indicated by Cartus *et al.* [41], undersampling involves randomly selecting examples from a majority class to delete from the dataset. This technique has the effect of reducing the number of examples from the majority class to the desired percentage of distribution.

Table 6: Rebalancing of data based on dead field

Class	Quantity	Proportion
0 (There was no death)	2763	75%
1 (death)	921	25%

The data created from the rebalancing were divided into two new datasets, one for training and one for testing. For this division to be possible, it was decided to divide 70% of the data for the training set, and the remaining 30% for the test set. In addition, an additional validation base was created, containing 1/20 of the data in the original dataset.

The model was created based on the training dataset, using as parameters of the random forest algorithm the value of 300 for *n\_tree*, which corresponds to the number of decision trees that must be created for analysis, this value was defined after the creation and evaluation of the model with different values, the value of 3 for the parameter *mtry*, which corresponds to the number of parameters that will be used to perform the division of the tree. The *strata* parameter was configured as the variable we are analyzing, in this case, the *mortos* variable, and the *samplesize* parameter was 100 sample of survivors and 50 deaths. These two parameters combined facilitate the classification of data created from a rebalancing of the original dataset. Finally, the *importance* and *prox* parameters were set to true, when enabling these two parameters we will extract the importance of each predictor variable for the model. Figure 13 shows a snippet of the source code, which is

fully available at GitHub repository<sup>9</sup>.

```
summary(training)

# Creating the predictive model with Random Forest
rf.model3 <- randomForest(mortos ~., data=training, ntree = 300, mtry = 3, importance = TRUE, prox=TRUE,
                          strata=training$mortos, sampsize=c(100,50))

rf.model3

# Applying the predictive model to the test dataframe
teste.pred1 <- predict(rf.model3, teste[, -1])

# Creating Confusion Matrix
matrizConfusao <- table(observed=teste$mortos, predicted=teste.pred1)
matrizConfusao

# Class error indicator for dead
matrizConfusao[2,1]/sum(matrizConfusao[2,])

# Visual Confusion Matrix 1
fourfoldplot(matrizConfusao, color = c("#CC6666", "#99CC99"),
              conf.level = 0, margin = 1, main = "Confusion Matrix - Test dataset")
```

Figure 13. Code sample of RoadLytics model

## 6. Application Results

In this section, the results achieved by applying the previously created model will be described.

### 6.1. Test Dataset Application

After applying the model to the test dataset, which has a total of 1658 records, the confusion matrix shown in Figure 14. As can be seen, the RoadLytics model was able to correctly classify 645 survivors and 232 fatal victims within a universe of 1,165 occurrences.

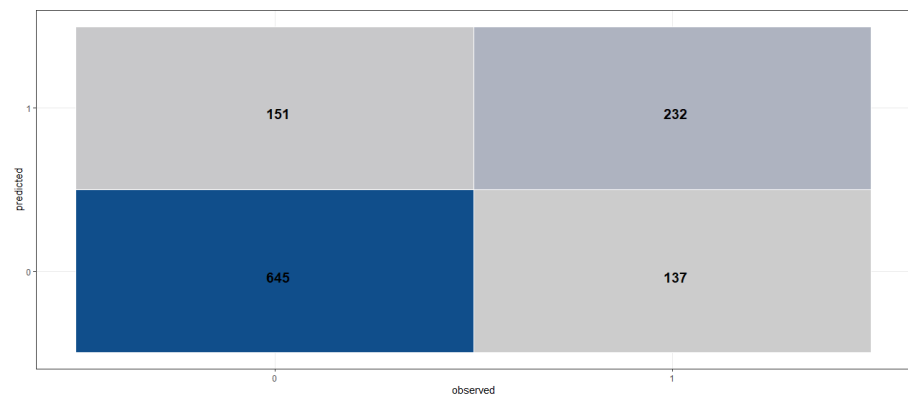


Figure 14. Confusion Matrix - Test dataset

The execution of the predictive model on the test dataset enables the collection of other metrics, like accuracy, which resulted in 0.77, this value represents the percentage of instances correctly classified by the model. The closer to 1, the better the model's accuracy. The specificity, which corresponds to the model's ability to identify negative instances was 0.80. Sensitivity, which has the ability to classify positive instances was 0.67. From these data, an F-Score of 0.61 is obtained, this metric corresponds to the model's assertiveness.

<sup>9</sup> <https://git.io/JLcWM>



6.2. Validation Dataset Application

Finally, the same model was applied to the validation dataset, which was extracted from the raw data, before any rebalancing technique has been applied. Through the confusion matrix presented in Figure 15, it can be noted that the model managed to classify 1745 survivors correctly, and 38 fatal victims.

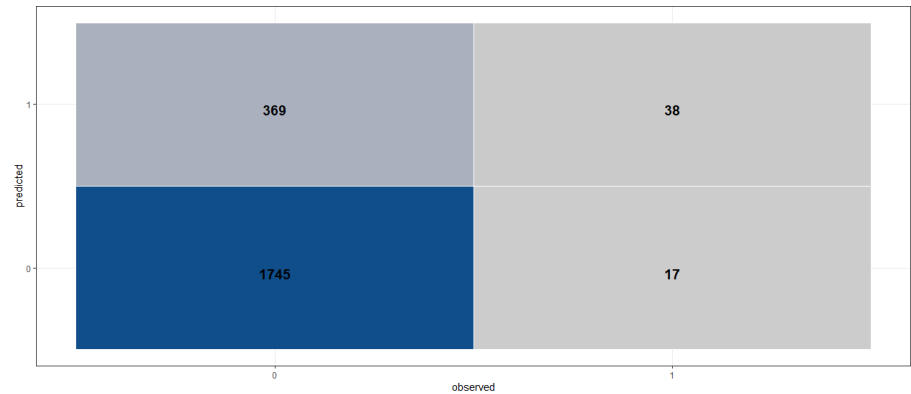


Figure 15. Confusion Matrix - Validation dataset

6.3. Importance of Predictor Variables

The creation of the model considered the analysis of the selected predictor variables. Figures 16 and 17 shows the gain of information through the use of each of the variables in the model.

This selection of variables succeeds the test execution with different combinations. The model combined these variables to reach an Out-Of-Bag (OOB) estimate of the error rate of 24,98%, this value represents the probability measure the random forest prediction error.

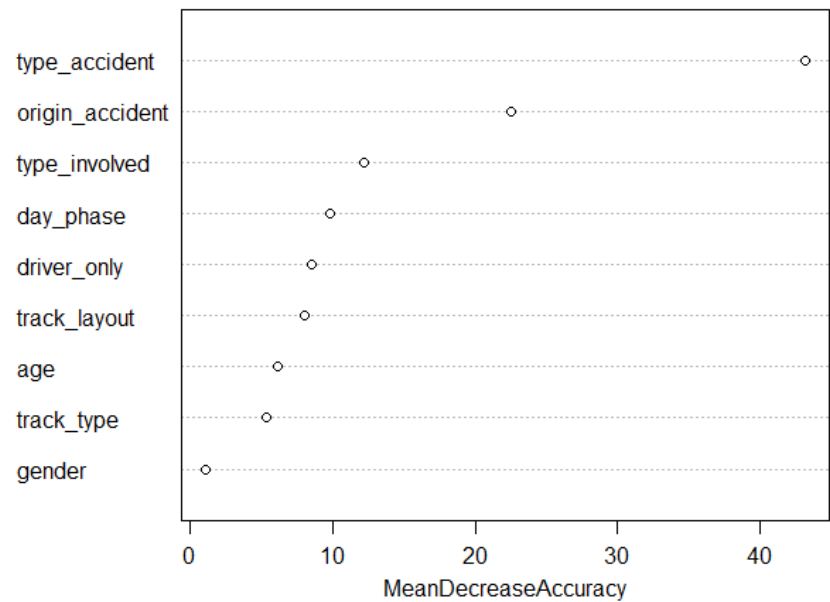


Figure 16. Predictor Variables - Mean Decrease Accuracy

Mean Decrease Accuracy shows how representative a variable is for the model. As we can see, we highlight the variable *type\_accident*, which leads us to the conclusion that without this predictive information, it would be difficult to obtain the result achieved. It is also worth highlighting the importance of the variables *origin\_accident* and *type\_involved*.

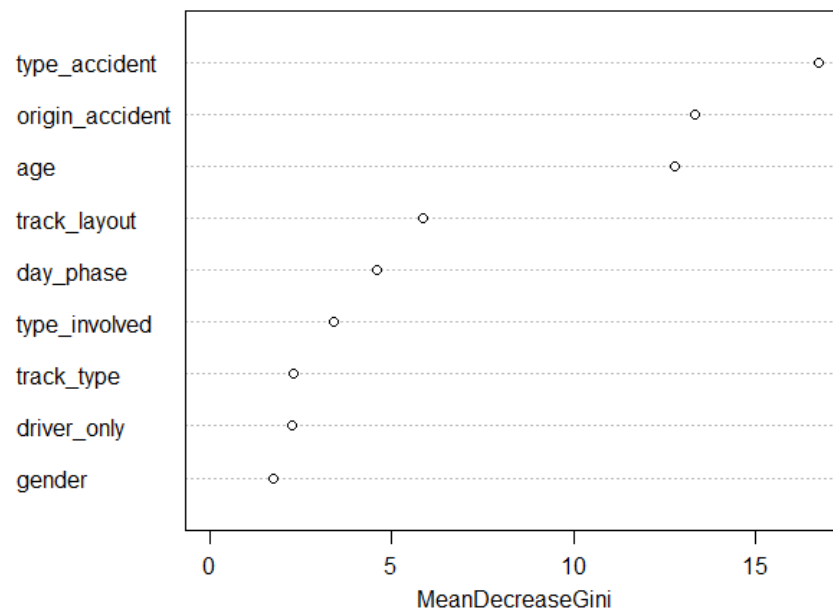


Figure 17. Predictor Variables - Mean Decrease Gini

When analyzing the Mean Decrease Gini values, the predictor variables that best obtain results are identified when they are used as dividing nodes by the decision trees. The variable *type\_accident* is highlighted, followed by *origin\_accident* and *age*.

## 7. Final Considerations

The present study aimed to collect, analyze and present information about traffic accidents that occurred on federal highways in the state of RS, between the years 2017 and 2020. The rationale for the development of this work was based on the need to analyze and understand the main factors that can explain the probability of accidents on the highways, especially fatal accidents. This model can be reused by related projects and can be applied to any region helping to minimize the number of accidents and to prevent deaths by automotive collisions.

This work proposed the predictive model created based on the random forest algorithm implemented by the programming language R, after collecting and pre-processing the data. The results obtained were presented in an exploratory and predictive way, making use of graphics and creation of heat maps to contribute to the visualization and understanding of accidents, as well as driver behavior, track and climatic conditions. Some variables were selected from the dataset to create the model, different combinations were used to arrive at the group of variables used to build the RoadLytics.

This study enables the analysis of days with the highest incidence of accidents, the profile of drivers, in the same way that the main causes of accidents and highways with the highest incidence can be assessed. The heat maps presented the distribution by season and type of vehicle, supporting the analysis with more technical elements. The predictive model showed satisfactory results, even with a 3 year limitation. The use of a longer period, a greater number of variables eligible to the predictive model, could contribute to obtain greater accuracy.

Studies of scenarios like this can contribute to public institutions, insurance companies and the population as a whole regarding the behavior of drivers and the main risks of

accidents on federal highways in the state of RS, and can serve as a basis for public policies aimed at mitigating the risk to the population and bring benefits in different areas, such as health, logistics and safety.

As future work, others machine learning algorithms can be evaluated, both supervised and unsupervised, in addition to the use of a longer period. According to the availability by the Federal Highway Police, the crossing data with the health open data available by the health ministry, called DATASUS [42] can be evaluated, providing extra information about deaths, adding new variables to perform the model training. Furthermore, the model can be deployed and made available as an App to help end-users to evaluate the highways risks including considering their profile [43], considering the data privacy of users [44]. More than that could be published as an open API to be integrated with map services helping users to identify critical points dynamically during their trips [45].

Finally, future work will explore the use of Context Histories [46–48] to organize the data, allowing pattern analysis [49], context prediction [50] and similarity analysis [51]. These strategies for handling context histories will improve the analysis of the data, mainly allowing the prediction and recommendation oriented to the safety of drivers on the highways.

#### **Author Contributions:**

Conceptualization, K.R.L. and J.E.R.T.; Investigation, K.R.L. and J.E.R.T.; Methodology, K.R.L., J.E.R.T. and J.L.V.B.; Software, K.R.L.; Project Administration, K.R.L. and J.E.R.T.; Supervision, J.L.V.B.; Validation, K.R.L., J.E.R.T. and J.L.V.B.; Writing—original draft, K.R.L. and J.E.R.T.; Writing—review and editing, J.L.V.B., V.R.Q.L. and D.H.I.; Financial, V.R.Q.L. and D.H.I. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by national funds through the Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) by the project UIDB/05064/2020 (VALORIZA – Research Centre for Endogenous Resource Valorization).

**Informed Consent Statement:** This research did not require ethical approval in accordance with the regulations of the University of Vale do Rio dos Sinos (UNISINOS). The data used in this study are public data, available in a government website for consultation.

**Acknowledgments:** The authors would like to thank the University of Vale do Rio dos Sinos (Unisinos), the Applied Computing Graduate Program (PPGCA), the Mobile Computing Laboratory (Mobilab), the Research Support Foundation of the State of Rio Grande do Sul (FAPERGS), the National Development Council Scientific and Technological (CNPq) and the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Code Funding 001.

**Conflicts of Interest:** The authors declare no conflict of interest.

#### **Abbreviations**

The following abbreviations are used in this manuscript:

BR116	Federal Brazilian Highway coded as 116
BR158	Federal Brazilian Highway coded as 158
BR285	Federal Brazilian Highway coded as 285
BR290	Federal Brazilian Highway coded as 290
BR386	Federal Brazilian Highway coded as 386
CSV	Comma Separated Value
DL	Deep Learning
FHP	Federal Highway Police
GDP	Gross Domestic Product
GPS	Global Positioning System
GRU	Gated Recursive Unit
ICU	Intensive Care Unit
LSTM	Long Short Term Memory
ML	Machine Learning
MVC	Motor Vehicle Collisions
OOB	Out-Of-Bag
PRF	Polícia Rodoviária Federal
RF	Random Forest
RS	Rio Grande do Sul
SUS	Sistema Único de Saúde
UHS	Unified Health System
WHO	World Health Organization

## References

1. WHO. Road traffic injuries. [www.who.int/news-room/fact-sheets/detail/road-traffic-injuries](http://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries), 2020. Access date: March 2021.
2. Gopalakrishnan, S. A public health perspective of road traffic accidents. *Journal of family medicine and primary care* **2012**, 1, 144–150. 24479025[pmid], doi:10.4103/2249-4863.104987.
3. Andrade, F.R.d.; Antunes, J.L.F. Trends in the number of traffic accident victims on Brazil's federal highways before and after the start of the Decade of Action for Road Safety. *Cadernos de Saúde Pública [online]* **2019**, 35. doi:doi.org/10.1590/0102-311X00250218.
4. Inada, H.; Ashraf, L.; Campbell, S. COVID-19 lockdown and fatal motor vehicle collisions due to speed-related traffic violations in Japan: a time-series study. *Injury Prevention* **2021**, 27, 98–100. doi:10.1136/injuryprev-2020-043947.
5. SP, I. Deaths by traffic cars accident growing in the São Paulo state. [www.infosiga.sp.gov.br/](http://www.infosiga.sp.gov.br/), 2021. Access date: March 2021.
6. GOV.BR. Open Data - Federal Highway Police. <http://portal.prf.gov.br/dados-abertos-acidentes>, 2020. Access date: February 2021.
7. Kavakiotis, I.; Tsave, O.; Salifoglou, A.; Maglaveras, N.; Vlahavas, I.; Chouvarda, I. Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal* **2017**, 15, 104–116. doi:doi.org/10.1016/j.csbj.2016.12.005.
8. Gaussmann, R.; Coelho, D.; Fernandes, A.; Crocker, P.; Leithardt, V.R.Q. Estimated Maintenance Costs of Brazilian Highways Using Machine Learning Algorithms. *Journal of Information Systems Engineering and Management* **2020**, 5. doi:10.29333/jisem/8427.
9. Antonopoulos, I.; Robu, V.; Couraud, B.; Kirli, D.; Norbu, S.; Kiprakis, A.; Flynn, D.; Elizondo-Gonzalez, S.; Wattam, S. Artificial intelligence and machine learning approaches to energy demand-side response: A systematic review. *Renewable and Sustainable Energy Reviews* **2020**, 130, 109899. doi:doi.org/10.1016/j.rser.2020.109899.
10. Benbarrad, T.; Salhaoui, M.; Kenitar, S.B.; Arioua, M. Intelligent Machine Vision Model for Defective Product Inspection Based on Machine Learning. *Journal of Sensor and Actuator Networks* **2021**, 10. doi:10.3390/jsan10010007.
11. da Rosa Tavares, J.E.; Victória Barbosa, J.L. Ubiquitous healthcare on smart environments: A systematic mapping study. *Journal of Ambient Intelligence and Smart Environments* **2020**, 12, 513–529. 6, doi:10.3233/AIS-200581.
12. da Rosa Tavares, J.E.; Victória Barbosa, J.L. Apollo SignSound: an intelligent system applied to ubiquitous healthcare of deaf people. *Journal of Reliable Intelligent Environments* **2021**, 7, 157–170. doi:10.1007/s40860-020-00119-w.

13. Machado, S.D.; Tavares, J.E.d.R.; Martins, M.G.; Barbosa, J.L.V.; González, G.V.; Leithardt, V.R.Q. Ambient Intelligence Based on IoT for Assisting People with Alzheimer's Disease Through Context Histories. *Electronics* **2021**, *10*. doi:10.3390/electronics10111260.
14. Bavaresco, R.; Barbosa, J.; Vianna, H.; Büttenbender, P.; Dias, L. Design and evaluation of a context-aware model based on psychophysiology. *Computer Methods and Programs in Biomedicine* **2020**, *189*, 105299. doi:doi.org/10.1016/j.cmpb.2019.105299.
15. Dias, L.P.S.; Barbosa, J.L.V.; Feijó, L.P.; Vianna, H.D. Development and testing of iAware model for ubiquitous care of patients with symptoms of stress, anxiety and depression. *Computer Methods and Programs in Biomedicine* **2020**, *187*, 105113. doi:doi.org/10.1016/j.cmpb.2019.105113.
16. Khan, M.A.; Saqib, S.; Alyas, T.; Ur Rehman, A.; Saeed, Y.; Zeb, A.; Zareei, M.; Mohamed, E.M. Effective Demand Forecasting Model Using Business Intelligence Empowered With Machine Learning. *IEEE Access* **2020**, *8*, 116013–116023. doi:10.1109/ACCESS.2020.3003790.
17. Helfer, G.A.; Victória Barbosa, J.L.; dos Santos, R.; da Costa, A.B. A computational model for soil fertility prediction in ubiquitous agriculture. *Computers and Electronics in Agriculture* **2020**, *175*, 105602. doi:doi.org/10.1016/j.compag.2020.105602.
18. Martini, B.G.; Helfer, G.A.; Barbosa, J.L.V.; Espinosa Modolo, R.C.; da Silva, M.R.; de Figueiredo, R.M.; Mendes, A.S.; Silva, L.A.; Leithardt, V.R.Q. IndoorPlant: A Model for Intelligent Services in Indoor Agriculture Based on Context Histories. *Sensors* **2021**, *21*. doi:10.3390/s21051631.
19. Thomas, R.W.; Vidal, J.M. Toward detecting accidents with already available passive traffic information. 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC), 2017, pp. 1–4. doi:10.1109/CCWC.2017.7868428.
20. Campolo, C.; Genovese, G.; Iera, A.; Molinaro, A. Virtualizing AI at the Distributed Edge towards Intelligent IoT Applications. *Journal of Sensor and Actuator Networks* **2021**, *10*. doi: 10.3390/jsan10010013.
21. Orrego, R.B.S.; Barbosa, J.L.V. A Model for Resource Management in Smart Cities Based on Crowdsourcing and Gamification. *International Journal of Universal Computer Science (J.UCS)* **2019**, *25*, 1018–1038. doi:10.3217/jucs-025-08-1018.
22. Sestrem Ochôa, I.; Reis Quietinho Leithardt, V.; Calbusch, L.; De Paz Santana, J.F.; Delcio Parreira, W.; Oriel Seman, L.; Albenes Zeferino, C. Performance and Security Evaluation on a Blockchain Architecture for License Plate Recognition Systems. *Applied Sciences* **2021**, *11*. doi: 10.3390/app11031255.
23. Rolim, C.O.; Rossetto, A.G.; Leithardt, V.R.; Borges, G.A.; Geyer, C.F.; dos Santos, T.F.; Souza, A.M. Situation awareness and computational intelligence in opportunistic networks to support the data transmission of urban sensing applications. *Computer Networks* **2016**, *111*, 55–70. Cyber-physical systems for Mobile Opportunistic Networking in Proximity (MNP), doi: https://doi.org/10.1016/j.comnet.2016.07.014.
24. Mahdavejad, M.S.; Rezvan, M.; Barekatain, M.; Adibi, P.; Barnaghi, P.; Sheth, A.P. Machine learning for internet of things data analysis: a survey. *Digital Communications and Networks* **2018**, *4*, 161–175. doi:doi.org/10.1016/j.dcan.2017.10.002.
25. Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32. doi:10.1023/A:1010933404324.
26. Lin, W.; Wu, Z.; Lin, L.; Wen, A.; Li, J. An Ensemble Random Forest Algorithm for Insurance Big Data Analysis. *IEEE Access* **2017**, *5*, 16568–16575. doi:10.1109/ACCESS.2017.2738069.
27. Gewin, V. Data sharing: An open mind on open data. *Nature* **2016**, *529*, 117–119. doi: 10.1038/nj7584-117a.
28. Veljković, N.; Bogdanović-Dinić, S.; Stoimenov, L. Benchmarking open government: An open data perspective. *Government Information Quarterly* **2014**, *31*, 278–290. doi: doi.org/10.1016/j.giq.2013.10.011.
29. Tridge. Brazil: Rio Grande do Sul soybean crop should hit record, says Emater. [www.tridge.com/news/rio-grande-do-sul-soybean-crop-should-hit-record-s](http://www.tridge.com/news/rio-grande-do-sul-soybean-crop-should-hit-record-s), 2021. Access date: April 2021.
30. Fageria, N.; Wander, A.; Silva, S. Rice (*Oryza sativa*) cultivation in Brazil. *Indian Journal of Agronomy* **2014**, *59*, 350–358. [www.indianjournals.com/ijor.aspx?target=ijor:ija&volume=59&issue=3&article=001](http://www.indianjournals.com/ijor.aspx?target=ijor:ija&volume=59&issue=3&article=001).
31. Barroso Junior, G.T.; Bertho, A.C.S.; Veiga, A.d.C. A letalidade dos acidentes de trânsito nas rodovias federais brasileiras. *Revista Brasileira de Estudos de População* **2019**, *36*, 1–22. doi: 10.20947/S0102-3098a0074.
32. Chang, H.; Park, D. Potentialities of Vehicle Trajectory Big Data for Monitoring Potentially Fatigued Drivers and Explaining Vehicle Crashes on Motorway Sections. *Sustainability* **2020**, *12*. doi:10.3390/su12155877.

33. von Buxhoeveden, G.; Becker, U. Comparison of traffic incident data in individual and public transport. 2016 3rd International Conference on Systems and Informatics (ICSAI), 2016, pp. 1067–1071. doi:10.1109/ICSAL.2016.7811109.
34. Lamr, M. Big Data and Its Usage in Systems of Early Warning of Traffic Accident Risks. 2018 Sixth International Conference on Enterprise Systems (ES), 2018, pp. 154–157. doi: 10.1109/ES.2018.00031.
35. Wang, S.; Zhao, J.; Shao, C.; Dong, C.; Yin, C. Truck Traffic Flow Prediction Based on LSTM and GRU Methods With Sampled GPS Data. *IEEE Access* **2020**, *8*, 208158–208169. doi: 10.1109/ACCESS.2020.3038788.
36. Zhang, K.; Hassan, M. Injury Severity Analysis of Nighttime Work Zone Crashes. 2019 5th International Conference on Transportation Information and Safety (ICTIS), 2019, pp. 1301–1308. doi:10.1109/ICTIS.2019.8883723.
37. Mokoatle, M.; Marivate, V. Collision Course: Challenges with Road Traffic Accident Data in South Africa. 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), 2018, pp. 1–6. doi:10.1109/ICABCD.2018.8465419.
38. Mazouri, F.Z.E.; Abounaima, M.C.; Najah, S.; Zenkour, K. Data mining for road accident analysis in a big data context. EAI, 2019. doi:10.4108/eai.24-4-2019.2284124.
39. Chen, C. Analysis and Forecast of Traffic Accident Big Data. *ITM Web Conf.* **2017**, *12*, 04029. doi:10.1051/itmconf/20171204029.
40. Zůvala, R.; Bucsuházy, K.; Valentová, V.; Frič, J. Representativeness of Czech In-Depth Accident Data. *CDV-Transport Research Center* **2021**, p. 12. doi:10.3390/safety7020040.
41. Cartus, A.R.; Bodnar, L.M.; Naimi, A.I. The Impact of Undersampling on the Predictive Performance of Logistic Regression and Machine Learning Algorithms: A Simulation Study. *Epidemiology* **2020**, *31*.
42. Ministry, H. Health Portal - SUS. <http://www2.datasus.gov.br>, 2021. Access date: April 2021.
43. Leithardt, V.; Santos, D.; Silva, L.; Viel, F.; Zeferino, C.; Silva, J. A Solution for Dynamic Management of User Profiles in IoT Environments. *IEEE Latin America Transactions* **2020**, *18*, 1193–1199. doi:10.1109/TLA.2020.9099759.
44. Lopes, H.; Pires, I.M.; Sánchez San Blas, H.; García-Ovejero, R.; Leithardt, V. PriADA: Management and Adaptation of Information Based on Data Privacy in Public Environments. *Computers* **2020**, *9*. doi:10.3390/computers9040077.
45. Tavares, J.; Barbosa, J.; Cardoso, I.; Costa, C.; Yamin, A.; Real, R. Hefestos: an intelligent system applied to ubiquitous accessibility. *Universal Access in the Information Society* **2016**, *15*, 589–607. doi:10.1007/s10209-015-0423-2.
46. Barbosa, J.; Tavares, J.; Cardoso, I.; Alves, B.; Martini, B. TrailCare: An indoor and outdoor Context-aware system to assist wheelchair users. *International Journal of Human-Computer Studies* **2018**, *116*, 1–14. doi:https://doi.org/10.1016/j.ijhcs.2018.04.001.
47. Aranda, J.A.S.; Bavaresco, R.S.; Carvalho, J.V.D.; Yamin, A.C.; Tavares, M.C.; Barbosa, J.L.V. A computational model for adaptive recording of vital signs through context histories. *Journal of Ambient Intelligence and Humanized Computing* **2021**. doi:10.1007/s12652-021-03126-8.
48. Rosa, J.H.; Barbosa, J.L.V.; Kich, M.; Brito, L. A Multi-Temporal Context-aware System for Competences Management. *International Journal of Artificial Intelligence in Education* **2015**, *25*, 455–492. doi:10.1007/s40593-015-0047-y.
49. Dupont, D.; Barbosa, J.L.V.; Alves, B.M. CHSPAM: a multi-domain model for sequential pattern discovery and monitoring in contexts histories. *Pattern Analysis and Applications* **2020**, *23*, 725–734. doi:10.1007/s10044-019-00829-9.
50. da Rosa, J.H.; Barbosa, J.L.; Ribeiro, G.D. ORACON: An adaptive model for context prediction. *Expert Systems with Applications* **2016**, *45*, 56–70. doi:doi.org/10.1016/j.eswa.2015.09.016.
51. Filippetto, A.S.; Lima, R.; Barbosa, J.L.V. A risk prediction model for software project management based on similarity analysis of context histories. *Information and Software Technology* **2021**, *131*, 106497. doi:doi.org/10.1016/j.infsof.2020.106497.



