Prediction of road accident risk through data mining methods

Academia Militar
Lisbon, Portugal
12721218david@gmail.com [1],
jose.silva@academiamilitar.pt [2]
David Dias [1], José Silva [2]

Instituto Superior Técnico
Lisbon, Portugal
alex@isr.tecnico.ulisboa.pt [3]
Alexandre José Malheiro Bernardino [3]

Abstract — **This work proposes a tool to assist policing guided by information, through a system for predicting the risk of road accidents. The system applies the various stages of the knowledge discovery process in databases in the *Guarda Nacional Republicana* (GNR) database, where several accident reports are found.**

In addition to accident reports, the GNR also provided data on administrative offenses that contain both the number of inspections carried out and the number of drivers with excessive alcohol consumption, speeding, among other offences. To complement the data provided by the GNR, other publicly available databases were explored, such as meteorological data and annual calendars with information on holidays and festivities. Both classic methods and deep learning methods were tested. The best results were obtained for the neural network algorithm.

Key words: risk prediction; road accidents; supervised classification; classical methods; deep neural networks.

I. Introduction

Road accidents cause several deaths per year and result in economic and physical damage to victims and the State. Prevention actions by the security forces have been focused on what has been called Information-Guided Policing. Since whenever there is an accident, the data related to it is stored in the GNR database, it creates a database in which it is possible to discover patterns and create knowledge. Data mining techniques have evolved and are being applied to more and more real-world problems. Data mining methods can be used in the provided database to extract knowledge that can somehow help guide policing and thus improve prevention techniques and awareness campaigns by security forces. The author, as a military electrical engineer for the GNR, with an interest in the area of machine learning, and concerned with National Security issues, sees in this topic a way to bring both interests together, increasing his technical knowledge.

The data provided by the GNR correspond to the years 2019 to 2021 in the district of Setúbal. In addition to being mostly categorical, which are usually harder to analyse, there are several incomplete and incorrect data, therefor making it necessary to aply techniques to solve those problems. In addition, it will be the first time that there will be aplied data mining algorithms to this dataset, so it is a possibility that we conclude that the error measures are too high for this tool to be viable and that another type of data is needed to achieve better performances.

A. Objectives

This work aims to develop a tool to aid information-guided policing in traffic policement. For this, several data mining algorithms that have already been shown to be successful when applied to different types of datasets will be tested. These tools will be applied to the GNR database, which contains several accident reports. To complement the data provided by the GNR, other publicly available databases will be explored, such as meteorological data and annual calendar.

II. Theoretical Framework

A. Knowledge Discovery in Databases

Current technology allows the storage of large and multiple databases. The analysis of these data is often useful; however, it is impractical without the aid of computational tools. From here came the Knowledge Discovery in Databases (KDD) process, represented in Fig. 1, which aims to identify valid and potentially useful patterns in data and information, in order to generate knowledge, using computational tools [1], [2], [3].

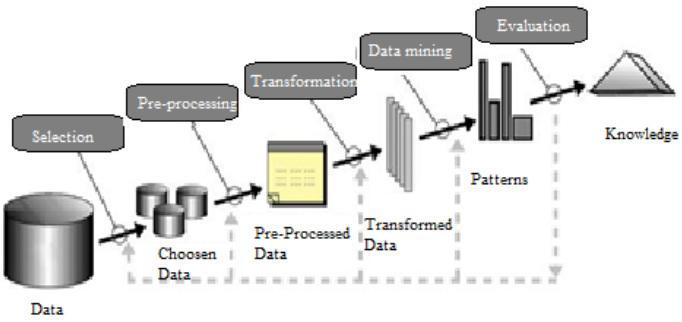

Figure 1 - Steps that make up the KDD process, adapted from [3]

Briefly, the meaning of each step can be given by:

1) **Data selection / Problem definition**: In this step, the domain of available data is defined, what information and data are relevant and what are the objectives of knowledge discovery [1]–[5]
2) **Pre-Processing**: This second stage aims to prepare the data for the algorithms of the next stage, namely, to perform data cleaning, data integration, data reduction and data transformation/normalization [1], [4] and [5].
3) **Data Mining**: Some authors refer to Data Mining and the Knowledge Discovery process in Databases as synonyms. However, Data Mining was considered as a step in this KDD process, as in [1]–[3], [5]. It is at

this stage that algorithms are applied to the data in search of knowledge, that is, to extract patterns in the data. The choice of algorithm to be applied depends on the type of task to be performed.

4) **Evaluation and representation of results**: In this step, the models obtained are interpreted and evaluation metrics are used to estimate the quality of the results. Finally, tools to visualize the data obtained as output must be used.

B. Data mining

Data mining is often divided into main groups. Supervised learning, unsupervised learning, reinforcement learning, among others. Supervised learning occurs when an algorithm learns from available data, where that data already has an associated output. For example, if the objective of a data mining problem is to predict the male or female gender through the image of a face, for this type of learning it would be necessary to have a set of faces already with the gender properly identified, so that the algorithm, through this set of images, could create a model to predict new images. It is important to distinguish regression problems where the data for which we want to predict the value are numerical values, from classification problems where the data are categorical values [4], [6]–[9].

In [7] 84 articles were analysed that discuss different techniques of supervised and unsupervised learning, in which the objective was to find a definition for the different terms and existing techniques. It was concluded that Decision Trees, Naive Bayes and Support Vector Machine are the most frequently used techniques in these 84 articles. Other most frequently used supervised learning algorithms are K-nearest neighbours (KNN) [6]–[8], [10] and artificial neural network (ANN) [8], [11], [12].

C. Selection of Attributes

By reviewing several articles like [13]–[20] it was possible to achieve Tab. 1, on the most important attributes for road traffic accidents and the three main groups in which they are separated.

Table 1 - Attributes considered important for traffic road accidents that were found in literature.

Weather conditions	Precipitation; temperature; wind force.
Human behavior	Seat belt use; cell phone use; alcohol consumption calendar
Road conditions	Road networks; luminosity; road identification; traffic volume.

For the selection of attributes, it is important to analyse the correlation between the different variables and the target variable. We often use Pearson's Correlation Coefficient to calculate the linear correlation between continuous numeric variables. However, we must use a different metric to calculate the correlation between categorical variables, as is the case with our dataset. Cramer's V correlation is used to calculate the correlation between nominal categorical variables with more than two (non-binary) values [21]. Because these algorithms are not as well-known, a more detailed review was made.

Due to the characteristics of our data, one of the metric for calculating correlations between attributes will be Cramer's V, defined as [22]:

$$\emptyset_c = \sqrt{\frac{X^2}{N(k-1)}} \quad (1)$$

Where:

- $\emptyset c$ is the value of V of Cramer;
- X^2 is the value of chi-squared;
- N is the number of samples;
- k is the number of categories of the variable with the smallest number of categories.

$$X^2 = \sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \quad (2)$$

Where:

- e_{ij} is the expected frequency value;
- o_{ij} is the observed frequency value of a combination of two values, one of variable i, the other of variable j;

$$e_{ij} = \frac{o_i . o_j}{N} \quad (3)$$

Where:

- e_{ij} is the value of the expected frequency of a combination of two values, one of variable i, the other of variable j;
- o_i is the marginal frequency of one of the values of variable i;
- o_j is the marginal frequency of one of the values of the variable j;
- N is the total number of samples.

The interpretation of how strong is the correlation between two nominal categorical variables from the values obtained by Cramer's V is given by Tab. 2.

Tabel 2 – Interpretation of V de Cramer, adapted from [23]

Values of Cramer's V, $\emptyset_c$	Interpretation
]0.25; 1.00]	Very Strong
]0.15; 0.25]	Strong
]0.10; 0.15]	Moderated
]0.05; 0.10]	Weak
[0; 0.05]	Very Weak

As for the correlation between nominal categorical and numerical categorical attributes, a good indicator is the Kruskal Wallis test. This aims to verify if there is a difference between several independent groups when these groups do not present a

normal distribution. In this case the groups do not need to have any type of distribution.

As explained, Cramer's V considers two variables to be independent if the value of the expected frequency is equal to the value of the observed frequency, meaning that the probability of two of the categories occurring is given by multiplying the probability of each one. In the Kruskal Wallis test, the meaning of variables being independent is that the sum of the classifications of all groups/categories tend to the same value [21], [23]–[25].

To be able to obtain a universal standard to eliminate attributes with low correlation values, it is important that all correlations calculated are comparable. If you cannot compare, you can use different valuation models. The Kruskal Wallis is equivalent to the chi-square also used in Cramer's V, so the values obtained can be compared between the two measures. The expression for the Kruskal Wallis test is given by [24], [25]:

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_i-\bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij}-\bar{r})^2} \quad (4)$$

- N is the number of total samples across all groups;
- g is the number of groups;
- n_i is the number of samples in group i ;
- r_{ij} is the rank value of sample j that belongs to group i;
- $\overline{r_i} = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$ is the mean value of the rank of all observations j in group i;
- $\bar{r} = \frac{1}{2}(N+1)$ is the average value of the sum of all classifications r_{ij}, that is, the expected value for the average of all groups.

Besides the correlation analysis two algorithms of selection of attributes were used, the Relief-based feature selection (RBA) and the Sequential Backward Selection (SBS).

The original RBA algorithm traverses m random training samples, where m is a user-defined parameter. At each cycle, Ri, one of the samples of m, is defined as the target sample. Through this target sample, the vector of characteristic weights, W, that contains the significance for each attribute, is updated based on the differences between the target sample and two closest neighbour samples. These two closest neighbour samples are the samples with the most similarity with the target sample, i.e. with the largest number of attributes with the same value, but while one corresponds to the closest sample with the same classification as the target sample, will call it C, other corresponds to the closest sample with a classification contrary to the sample target, will it F.

Features that have a different value between Ri and F support the hypothesis that they are features that add information regarding the target class, so that the quality estimate of the feature W[A] is increased. On the other hand, traits with differences between Ri and C provide evidence to the contrary, so that the quality estimate of trait W[A] is diminished. To address the regression problem since the target variable is the number of accidents, the algorithm used was the RBA variant with application in regression problems, which is given the name of RReliefF.

SBS is a variant of the Sequencial Forward Selection (SFS) algorithm. SFS starts from an empty set of features and gradually adds features selected by a performance measure, which measures how much each feature improves or worsens a mining method. At each iteration, the feature to be included in the feature set is selected from the features still available in the feature set.

D. Performance Measures

As a performance measure to evaluate the different mining algorithms, we chose to use the Mean Absolute Error, MAE, which is an error measure that sums the absolute error between the observations and the value obtained by the model. The Mean Squared Error was not used because the number of accidents has a lot of big outliers and the biggest errors happen for those values, so it would give a higher penalization by squaring that distance. The MAE is given by the following equation:

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|\overline{y_j} - y_j\right| \quad (5)$$

In addition to the MAE, it was also decided to use the Mean Absolute Percentage Error, which is given by the following alternative:

$$MAPE = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{A_t-F_t}{A_t}\right| \quad (6)$$

Where At is the real value and Ft is the value obtained through the model. Through this error model we can get an idea of the average error percentage that exists for each prediction.

Since the objective is to discover the accident risk and not to predict the exact value of accidents, the predicted values and the actual values will be grouped into three risk groups: low, medium, high. The choice of the interval in which the values of each of these groups are inserted will be made from the analysis of the accident frequency box diagram. After this grouping we can obtain a performance measure related to classification. We chose to use accuracy, which is given by:

$$Accuracy = \frac{vp+vn}{vp+fp+vn+fn} \quad (7)$$

Generally, accuracy measures the ratio of correct predictions to the total number of instances evaluated. vp is the number of true positives, Vn is the number of true negatives, fp is the number of false positives and fn is the number of false negatives [26].

III. RELATED WORK

The related works mentioned in this section are within the supervised classification and are divided into classical methods and deep learning methods.

A. Classic Approach

In 2016, Castro et al. [27], used a database of 451462 UK road accidents from 2010 to 2012, of which only 81690 of these accidents were included in the study. The WEKA tool was used, and 7 input variables were considered, which were: road type, light conditions, weather conditions, road surface conditions, vehicle maneuver, vehicle fuel type, the age of the vehicle and the severity of the accident. Three mining algorithms were used, which were: the Bayesian network, BayesNet, the decision tree algorithm and a neural network algorithm. All of them with an output variable with 3 possible values, which represent the severity of the accident (fatal, serious, or normal). The performance measure used was the accuracy and the severity of the accident was predicted by the 3 different algorithms with a very similar accuracy of around 72%. In the same year, Keshyap et al. [28] sought to find a link between road conditions and accident severity. Instead of decision trees, the Bayesian network algorithm was used, also through the WEKA software. In this work, 12 attributes have already been included, including: driver's condition, driver's experience, weather conditions, type of road, lighting conditions, vehicle conditions, type of vehicle included in the accident, type of animal, severity of the accident, seat belt use and location. A total of 31,698 accidents were analysed from questionnaires made to people who had suffered accidents, from 2003 to 2015. The use of attribute selection worsened the accuracy of the model and the best result obtained was 89% without any selection algorithm attributed. In 2019, Hussain et al. [29] carried out an evaluation of different classical data mining methods in road accidents, analysing literature similar to the one mentioned in the previous paragraph and reaching the conclusion that the most used algorithms and with greater accuracy are the Multi-layer Perceptron, the decision tree and Naive Bayes. In the same year, Kumeda et al. [20] applied 6 classic classification algorithms, such as Naive Bayes, Multi-layer Perceptron, Random Forest, among others, to find the factors that most influence the road accident.

Although the works mentioned above deal with classification rather than regression problems, they will be important to get an idea of the type of variables used. Most of the literature found with applications of classical methods did not aim to predict accident quantities, but to predict classes, such as accident severity, among others. All the works mentioned above applied classical mining techniques, most of them on a small-scale dataset (example: on one or a small number of roads) with a limited number of attributes.

B. Deep Learning Approach

Some more recent works have tried to face the problems in the analysis of road accidents by using Deep Learning Models. Chen et al. [30] used data from about 1.6 million GPS records and a history of accident records to build a model that relates human mobility to accident risk. In this way, the model evaluates the accident risk in real time through a classification of the accident risk for each zone of the map. Data such as the geo-location of the accident and the levels of human mobility in real time proved to be essential. The author also mentions that there are many factors that will lead to a traffic accident, such as driver behaviour, weather, and road conditions. But that, although some studies have focused on the connection between traffic accidents and these factors, it is very difficult to reveal the dynamic change in accident risk with these factors alone. In 2018, Yuan et al. [31] carried out a study in order to be able to predict the risk of an accident, according to the time, place and day. For this, a deep learning approach was used, which is based on the spatial and temporal heterogeneity of the data, a characteristic of road accidents. The study was carried out with data from the state of Iowa, in the United States of America and the sample contains 375,690 accidents from 2006 to 2014. External databases were added to this study, such as: data on traffic volume, road conditions, precipitation, and ambient temperature data from four different databases over 8 years. The algorithm used was an adaptation of the long and short memory convolutional neural network and a software was created that creates a predictive model for each region of the state, because it was concluded in this study that the main causes of accidents vary from region to region.

IV. Results and Discussion

In this chapter, the results produced by the methodology are presented and analysed. In tasks in which more than one technique was presented, these are compared to choose the technique that best suits the task in question.

A. Database

Regarding data selection, in Tab. 3 all selected attributes of accident reports are indicated. Due to time limitations for conducting interviews with specialists in the area, as mentioned in the theoretical framework, we chose to select the variables according to the study by J. Costa et. Al. [32].

Table 3 – Selected attributes of GNR Database

Attribute	Type of data
Identification of accident	Numerical
Data	Data
Hora	Hora
Type of local	Boolean
Localization	Accident - Localization
Type of accident	Accident – type of accident
Day of the week	Accident – day of the week
Holiday	Boolean
Alcohol	Numerical
Administrative offenses	Numerical
Weather Conditions	Accident – Weather conditions

Regarding the group of the time of the accident, the highest number of accidents occurs in the intervals between morning work, afternoon work and afternoon rush hours.

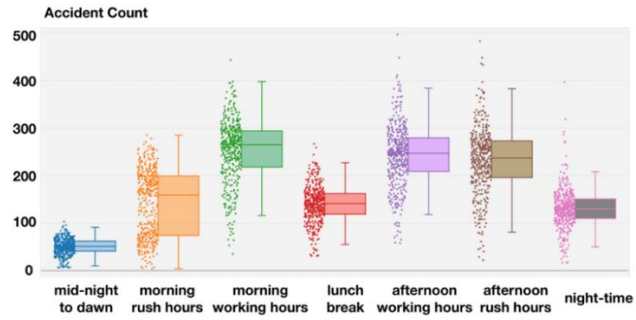

Figure 2 – Boxplots of the frequency of accidents occurred in different time intervals of Beijing. Taken from Ren et al. [25]

Comparing Fig. 2 and 3, we can verify that our data are consistent with the data found in the literature regarding the traffic hours with the highest frequency of accidents in Beijing.

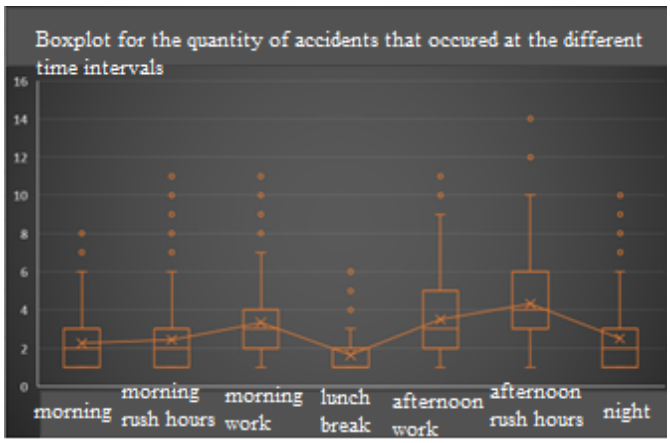

Figure 3 – Boxplots of the frequency of accidents occurred at the different time intervals in Setúbal, between 2019-2021.

The time interval for the accident clusters in Setúbal was based on the most common work and traffic schedules in Portugal, for Porto district [33], and it was assumed that the behaviour for Setúbal population is similar.

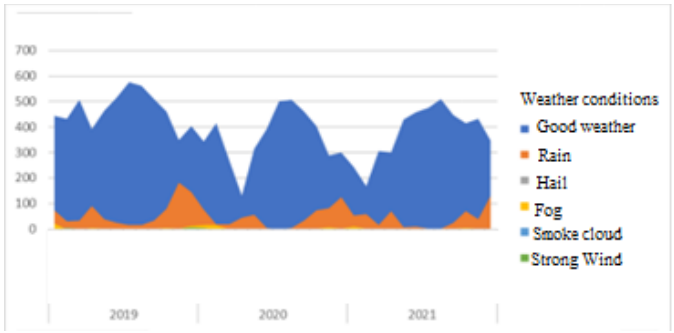

Figure 4 - Number of accidents grouped by month, year, and type of weather condition

Regarding atmospheric factors, it was possible to divide the accidents by the different atmospheric conditions in which they occurred. Considering the probability of raining as P(C) and the probability of having an accident as P(A), the graph in figure 4 gives us the probability of raining given that there was an accident, that is, P(C|A). It is intended to compare the probability of having an accident knowing that it rained P(A|C) with the probability of having an accident knowing that the weather is fine P(A|B). To make this comparison, the month of December was used as example. Thus, P(C|A) for the month of December is given by:

$$P(C|A) = \frac{144+126+132}{404+300+346} = 0.38 \quad (10)$$

Through the average number of rainy days for the month of December in Setúbal, we obtain that:

$$P(C) = \frac{8,5}{31} = 0,27 \quad (9)$$

By Bayes' theorem, it is obtained that:

$$P(A|C) = \frac{P(C|A).P(A)}{P(C)} \quad (10)$$

Using the same reasoning for good weather, it is concluded that the probability of an accident knowing that it is raining is greater than the probability of an accident, knowing that there is good weather:

$$P(A|B) = 0,85.P(A) < 1,4.P(A) = P(A|C) \quad (11)$$

Regarding the day of the week, the data were grouped by day of the week and by year, as can be seen in Fig. 5.

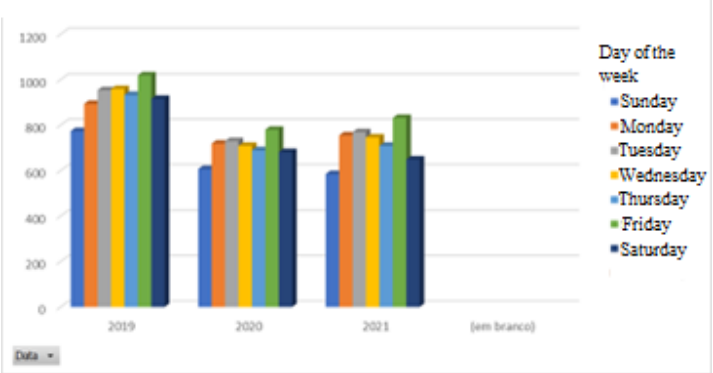

Figure 5 – Grouping of the number of accidents by day of the week and year

We can verify through the analysis of Fig. 5 the days of the week when the most accidents occur over the three years in the district of Setúbal. The day of the week with the most accidents was always Friday and with the fewest accidents, Saturday, and Sunday. Since Friday is the day of the week when there is often more congestion [33], this observation in Fig. 5 proves that the data are consistent, as it confirms that the volume of traffic is an influencing factor in the number of accidents.

B. Attributes selection

It was possible to obtain the different correlation values of Fig. 5 for pairs of nominal and numerical categorical variables (Kruskal Wallis test) and another for pairs of nominal

categorical variables with nominal categorical variables (Cramer's V).

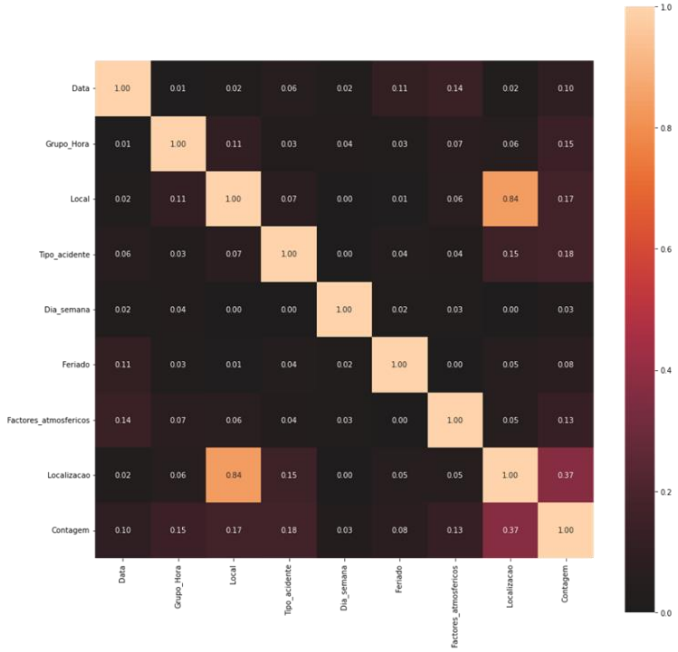

Figure 6 - Correlations of Cramer V and Kruskal Wallis test, depending on the type of pairs of variables

Observing Fig. 6 we can see that for the variable "*Contagem*", which represents the accident count, the variables with the highest correlation are the time of day, the type of place, the location, and the atmospheric factors. The type of accident, which represents the severity of the accident, was considered only to verify if there was a correlation between the severity of the accident and the number of accidents that occurred, which is confirmed. However, it is not useful for the predictive model, since it is not an information that can be obtained before the accident occurs.

The results obtained for RBA and SBS were analysed only for the highways because it was concluded that only for this location it was possible to achieve good results. Despite the results being different because of the different approaches of the algorithms, there were several attributes for which both algorithms agreed on the results that are represented in Tab. 4.

Table 4 - Relevance of characteristics for the creation of predictive models obtained with RBA and SBS, for incidents that occur on highways

Highways	Considered relevant by both algorithms	Considered irrelevant by both algorithms
RBA & SBS	'Rain', 'morning_work', afternoon_traffic', 'Friday', 'Saturday', 'August', 'February'	'sunday'

For the RBA algorithm, the RBA variant with application in regression problems was used, which is given the name of RReliefF. The number of random samples was 200, out of 1005 accidents on highways, and each sample was compared with its 4 nearest neighbours since the KNN algorithm obtained better results for 4 neighbours. As for the SBS algorithm, it uses a mining algorithm to measure performance by removing each of the features. The algorithm chosen was the one that obtained the best results for data mining, that is, the neural network. And the performance measure was the same as that used in the evaluation of mining algorithms, the MAPE. By iteratively removing each of the characteristics, we can see which ones improve or worsen the performance of the neural network.

The credibility of these algorithms is difficult to measure since these data are not synthetic and we do not know a priori which are the most important variables for each case. Although the algorithms measure the importance of each variable differently, there are several variables in which both algorithms agree that those variables influence or not accidents, as we can see in Tab. 4. From this it can be concluded that the variables that most influence the occurrence of accidents on highways, according to the algorithms used are: the atmospheric factor rain, the groups of hours from 10:00 to 12:29 and 17:00 to 19:59, the weekdays Friday and Saturday, and the months of August and February.

C. Data mining

Due to the importance of the location of accidents, after the initial experiences, it was decided to group the data by their location: Highways; National Roads or Itineraries; and Municipalities. To obtain an accident risk in addition, since it is not valued by the inclusion of values, it was decided to divide it into intervals that correspond to risk classes, determined by Tab. 5.

Table 5 – Intervals of number of accidents that correspond to the different classes of risk.

Classes	Intervals of number of accidents
Low Risk	<1.5
Medium Risk	>1.5; < 2.5
High risk	>2.5

This option was taken based on the box diagrams in Fig. 7. As can be seen, the variance of values is greater for the data set relating to highways and municipalities.

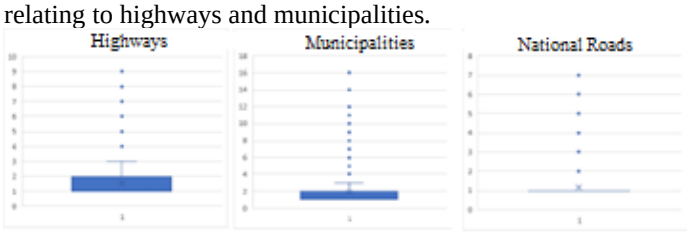

Figure 7 – Box plots for the values of the frequency of accidents in the different type of locations.

In the dataset of itineraries or national roads, for the chosen time interval, in most cases there is only 1 accident, so it would

be necessary to have a higher frequency of accidents for the model to be useful.

Table 6 – Summary of the best obtained results for the two experiences.

Summary of the best obtained results			
Algorithm (Regression w/ Neural Network)	**MAE (distance)**	**100-MAPE (%)**	**Precision**
General model	0.49	55.1%	88%
Highways (9,3% of total accidents)	0.57	56.4%	89%
Itineraries or National Roads (30% of total accidents)	0.55	50.6%	87%
Municipalities (60,7% of total accidents)	0.52	-4.3%	88%

For the model obtained for the municipalities, a very high error of approximately 104.3% is obtained (corresponding to -4.3% in Tab. 6). This can be explained by the high number of exceptions and the fact that these exceptions have very high values, which leads to higher percentage errors for higher values. As confirmation, when measuring the percentage error for values of the target variable equal to 1 and different from 1, was obtained an error of 30% and 197%, respectively.

Despite the limitations of the data, it was possible to obtain a good model for the highways. The highway, despite being the location with the lowest number of accidents for the district of Setúbal, is the location with the highest concentration of accidents per area when compared to municipalities and with the highest concentration of accidents per highway, when compared to the concentration of accidents on national routes or roads. The Highway is the type of location where there are more injuries and deaths by accident, which can be observed in Tab. 7.

Tabela 7 - Information related to the number of injured and dead by accident

Type of location	Percentage of accidents with injuries or deads	Nº of injured/deads per accident
Highway	25,1%	1,7
Municipalities	17,3%	1,2
Itineraries or National Roads	29,8%	1,42

The highway is also the location where it is possible for the GNR to carry out more effective inspections, since for the municipalities there are many roads and a vast area where accidents can occur, and there are no large concentrations of accidents on the national roads.

V. Conclusion

In this dissertation, the theme of road accident risk prediction through data mining methods was proposed. Data on accident reports were made available by the GNR, from accidents that occurred in Setúbal from 2019 to 2021. This work aims to create a model that can make a prediction with low error. The developed system consists of 3 main steps: (i) data selection and collection, (ii) pre-processing, (iii) mining algorithms.

Initially, through data analysis, it was possible to conclude that the highest concentration of accidents occurs during the time interval of 17:00-20:00. It was also possible to conclude that for our data, rain is the atmospheric factor with the highest probability of an accident. It is also concluded that the day of the week when more accidents occur is on Friday. Hence, it was proved that the data have some credibility as these conclusions are consistent with the literature.

Through an analysis of the correlation between the different variables, it was possible to conclude that the location of the accident is the variable that most influences the frequency of accidents. From this conclusion and the idea that for different types of locations there are different factors that influence the accident, these were grouped by the type of location they were in. For this reason, it was necessary to create different models for each set. In addition to the location, the correlation between variables also indicated other variables that most influence the frequency of accidents such as the time of day, the atmospheric factor and whether the accident occurred inside or outside a location. After dividing the data set into the three types of location (highways, national roads or itineraries and municipalities), it was possible, through the variable selection algorithms, to understand which variables most influence each type of location.

The data mining problem was approached as a regression problem since the target variable was the frequency of accidents in the defined time interval. The mining algorithms tested were KNN, simple linear regression, Lasso and Ridge, the decision tree for regression and the traditional neural network. Both for the initial dataset and for the datasets separated by location in the following sets: highways, national stays or itineraries and municipalities. The best result was obtained through the neural network. However, for each set, different models were obtained, with different architectures (number of nodes, training periods, etc). The best result was obtained for the highway dataset. The highway, despite being the location with the lowest number of accidents for the district of Setúbal, is the location with the highest concentration of accidents per area when compared to municipalities and the highest concentration of accidents per highway, when compared to the concentration of accidents in national routes or roads. In addition, the highway is the location where there are more injuries and deaths by accident. The highway is also the location where it is possible for the GNR to carry out more effective inspections, since for the municipalities there are a large number of roads and a vast area where accidents can occur, and there are no large concentrations of accidents on the routes. time. This allows the model to have more information to find the right parameters. The percentage of error in the regression was 44%, however it was not intended to obtain the exact number of accidents, so the results were grouped into 3 risk classes, according to the box diagram obtained for the frequency of accidents. Thus, an error percentage of only 11% was obtained.

VI. Future Work

As future work, the first step would be to improve data collection to ensure that the geo-location of accidents was acquired. In this way, it would be possible to opt for more complex approaches such as those mentioned in the deep learning approach studies. Another important variable to obtain would be the levels of human mobility, which would be possible to obtain through cooperation with applications such as Google Maps, Waze, or just through the speed at which Uber taxis or other companies travel, as was the case with some of the studies mentioned in the literature. Still in obtaining data, it would also be important to ensure their credibility by creating more mandatory parameters when the accident report is entered into the database by the military, and also making the military aware of the importance of entering data as coherently as possible. with the reality of the accident.

After that, it is also possible to create an application in which the military, through input data related to the month, day of the week, group time of day, whether it is a holiday or not, whether it is within a locality or not, the forecast weather for that day and place and, finally, the motorway that wants to predict the risk, can check the accident risk for this input set. Much of this data can be obtained automatically.

BIBLIOGRAPHIC REFERENCES

[1] R. Goldshmidt, E. Passos, E. Bezerra, "Data mining : conceitos, técnicas, algoritmos, orientações e aplicações", Elviser Ed. Rio de Janeiro, 2015.

[2] U. Fayyad, P. Smyth, G. Piatetsky-Shapiro, "Knowledge Discovery and Data Mining: Towards a Unifying Framework" American Association for Artificial Intelligence., 1996, pp. 82-88 [online] Acedido em: https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf

[3] T. Hendrickx, B. Cule, P. Meysman, S. Naulaerts, K. Laukens, and B. Goethals, "Mining association rules in graphs based on frequent cohesive itemsets," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9078, no. 3, pp. 637–648, 2015, doi: 10.1007/978-3-319-18032-8_50.

[4] S. Agarwal, *Data mining: Data mining concepts and techniques*, Elsevier I. Estados Unidos da America, 2014.

[5] S. Zhang, C. Zhang, and Q. Yang, "Data preparation for data mining," *Appl. Artif. Intell.*, vol. 17, no. 5–6, pp. 375–381, 2003, doi: 10.1080/713827180.

[6] L. Massaron, J. P. Mueller, Deep Learning for Dummies, New Jersey, For Dummies, 2019

[7] M. W. Berry, A. Mohamed, and B. W. Yap, *Supervised and Unsupervised Learning for Data Science*, no. January. 2020.

[8] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms: Second Edition*, Jonh Wiley., vol. 2. New Jersey, 2014.

[9] P. C. Sen, M. Hajra, and M. Ghosh, *Emerging Technology in Modelling and Graphics*, vol. 937. Springer Singapore, 2020.

[10] L. A. Belanche and F. F. González, "Review and Evaluation of Feature Selection Algorithms in Synthetic Problems," Universitat Politécnica de Catalunya, Barcelona, Spain, 2011, doi: https://doi.org/10.48550/arXiv.1101.2320, [Online]. Available: http://arxiv.org/abs/1101.2320

[11] R. Indrakumari, T. Poongodi, and K. Singh, *Introduction to Deep Learning*, Springer I. Croatia, 2021.

[12] L. Massaron, J. P. Mueller, Deep Learning for Dummies, New Jersey, For Dummies, 2019

[13] D. Eisenberg, "The mixed effects of precipitation on traffic crashes," Accid. Anal. Prev., vol. 36, no. 4, pp. 637–647, 2004, doi: 10.1016/S0001-4575(03)00085-X.

[14] R. B. Hayat *et al.*, "Explaining the road accident risk : Weather effects," *Accid. Anal. Prev.*, vol. 1, no. 60, pp. 456–465, 2013.

[15] J. D. Tamerius, X. Zhou, R. Mantilla, and T. Greenfield-Huitt, "Precipitation effects on motor vehicle crashes vary by space, time, and environmental conditions," *Weather. Clim. Soc.*, vol. 8, no. 4, pp. 399–407, 2016, doi: 10.1175/WCAS-D-16-0009.1.

[16] J. D. Febres, S. García-Herrero, S. Herrera, J. M. Gutiérrez, J. R. López-García, and M. A. Mariscal, "Influence of seat-belt use on the severity of injury in traffic accidents," *Eur. Transp. Res. Rev.*, vol. 12, no. 1, 2020, doi: 10.1186/s12544-020-0401-5.

[17] G. Musile, N. Pigaiani, D. Sorio, M. Colombari, F. Bortolotti, and F. Tagliaro, "Alcohol-associated traffic injuries in Verona territory: A nine-year survey," *Med. Sci. Law*, vol. 61, no. 1_suppl, pp. 7–13, 2021, doi: 10.1177/0025802420937577.

[18] Y. Song, S. Kou, and C. Wang, "Modeling crash severity by considering risk indicators of driver and roadway: A Bayesian network approach," *J. Safety Res.*, vol. 76, pp. 64–72, 2021, doi: 10.1016/j.jsr.2020.11.006.

[19] L. M. Martín-delosReyes, V. Martínez-Ruiz, M. Rivera-Izquierdo, E. Jiménez-Mejías, and P. Lardelli-Claret, "Is driving without a valid license associated with an increased risk of causing a road crash?," *Accid. Anal. Prev.*, vol. 149, no. November 2020, pp. 1–7, 2021, doi: 10.1016/j.aap.2020.105872.

[20] B. Kumeda, F. Zhang, F. Zhou, S. Hussain, A. Almasri, and M. Assefa, "Classification of road traffic accident data using machine learning Algorithms," in *2019 IEEE 11th International Conference on Communication Software and Networks, ICCSN 2019*, 2019, pp. 682–687, doi: 10.1109/ICCSN.2019.8905362.

[21] Z. Zhang, K. T. McDonnell, E. Zadok, and K. Mueller, "Visual correlation analysis of numerical and categorical data on the correlation map," *IEEE Trans. Vis. Comput. Graph.*, vol. 21, no. 2, pp. 289–303, 2015, doi: 10.1109/TVCG.2014.2350494.

[22] A. Bhattacharya and D. B. Dunson, "Simplex factor models for multivariate unordered categorical data," *J. Am. Stat. Assoc.*, vol. 107, no. 497, pp. 362–377, 2012, doi: 10.1080/01621459.2011.646934.

[23] H. Akoglu, "User's guide to correlation coefficients," *Turkish J. Emerg. Med.*, vol. 18, no. 3, pp. 91–93, 2018, doi: 10.1016/j.tjem.2018.08.001.

[24] S. Jun, "The Microbiome in Health and Disease", Volume 171 in the Progress in Molecular Biology and Translational Science, Elsevier Science, 2020, pp. 309-450.

[25] A. C. Leon, "Descriptive and Inferential Statistics" Compr. Clin. Psychol., New York, USA, vol. 3, pp. 243–285, 1998, doi: 10.1016/b0080-4270(73)00264-9.

[26] L. A. Belanche and F. F. González, "Review and Evaluation of Feature Selection Algorithms in Synthetic Problems," no. December 2013, 2011, [Online]. Available: http://arxiv.org/abs/1101.2320.

[27] Y. Castro and Y. J. Kim, "Data mining on road safety: Factor assessment on vehicle accidents using classification models," *Int. J. Crashworthiness*, vol. 21, no. 2, pp. 104–111, 2016, doi: 10.1080/13588265.2015.1122278.

[28] J. Kashyap, A. Chandra, and P. Singh, "Mining Road Traffic Accident Data to Improve Safety on Road-related Factors for Classification and Prediction of Accident Severity," *Int. Res. J. Eng. Technol.*, vol. 10, pp. 2395–56, 2016, [Online]. Available: https://www.irjet.net/archives/V3/i10/IRJET-V3I1041.pdf.

[29] S. Hussain, L. J. Muhammad, F. S. Ishaq, A. Yakubu, and I. A. Mohammed, "Performance evaluation of various data mining algorithms on road traffic accident dataset", Smart Innov. Syst. Technol., vol. 106, pp. 67–78, 2019, doi: 10.1007/978-981-13-1742-2_7.

[30] Q. Chen, X. Song, H. Yamada, and R. Shibasaki, "Learning deep representation from big and heterogeneous data for traffic accident inference," *30th AAAI Conf. Artif. Intell. AAAI 2016*, pp. 338–344, 2016.

[31] Z. Yuan, X. Zhou, and T. Yang, "Hetero-ConvLSTM: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 18, pp. 984–992, Jul. 2018, doi: 10.1145/3219819.3219922.

[32] J. Costa, E. Freitas, P. Pereira, M. Jacques, "Acidentes Rodoviários das Estradas Nacionais de Portugal", C-TAC - Comunicações a Conferências Nacionais, 2011. [Online]. URL: https://hdl.handle.net/1822/15483, Disponível em: https://repositorium.sdum.uminho.pt

[33] Seguropordias (2022). O congestionamento nas estradas da cidade do Porto. Disponível em: https://seguropordias.pt/blog/tr%C3%A2nsito-porto-portugal