IFCS 2022 — 17th Conference of the International Federation of Classification Societies

Promoters: MOPREVIS Modelação e Predição de Acidentes de Viação no Distrito de Setúbal; Universidade de Évora; GNR Guarda Nacional Republicana; FCT Fundação para a Ciência e a Tecnologia; República Portuguesa

Supports: IMT Instituto da Mobilidade e dos Transportes, I.P.; acp Automóvel Club de Portugal; Infraestruturas de Portugal; IPMA Instituto Português do Mar e da Atmosfera; ANSR Autoridade Nacional Segurança Rodoviária; waze *Portugal; altice

# Machine Learning approach to identify factors that influence accident severity

Daniel Santos [4,*], Vitor Nogueira [3,4,*], José Saias [3,4,*], Paulo Quaresma [3,4,*], Paulo Infante [1,2], Gonçalo Jacinto [1,2], Anabela Afonso [1,2], Leonor Rego [2], Pedro Nogueira [5,6], Marcelo Silva [5,6], Rosalina Pisco Costa [7,8], Patrícia Góis [9] and Paulo Rebelo Manuel [1]

[1] CIMA, IIFA, University of Évora, 7000-671 Évora, Portugal; pjsrm@uevora.pt
[2] Department of Matematics, ECT, University of Évora, 7000-671 Évora, Portugal; pinfante@uevora.pt; aafonso@uevora.pt; gjcj@uevora.pt; lrego@uevora.pt
[3] Algoritmi Research Centre, University of Évora, 7000-671 Évora, Portugal; vbn@uevora.pt (V.N.); pq@uevora.pt (P.Q.); jsaias@uevora.pt (J.S.)
[4] Department of Informatics, ECT, University of Évora, 7000-671 Évora, Portugal; dfsantos@uevora.pt
[5] ICT, IIFA, University of Évora, 7000-671 Évora, Portugal; pmn@uevora.pt (P.N.); marcelogs@uevora.pt (M.S.)
[6] Department of Geosciences, University of Évora, 7000-671 Évora, Portugal
[7] CICS.NOVA.UEVORA, IIFA, 7000-208 Évora, Portugal; rosalina@uevora.pt
[8] Department of Sociology, ECS, University of Évora, 7000-803 Évora, Portugal
[9] Department of Visual Arts and Design, EA, University of Évora, 7000-208 Évora, Portugal; pafg@uevora.pt
* These authors contributed equally to this work.

## INTRODUCTION

Since the twentieth century, road traffic accidents became a severe public health concern, with deaths and injuries posing a serious threat to world health and a negative influence on social and economic progress. One of the primary goals of accident data analysis is to determine the main factors that contribute to a traffic accident.

This study aims to create a Machine Learning approach capable of identifying the factors that influence accident severity (seriously injured/dead or lightly injured/no injured), supporting the analysis of accident data.

A seven-year traffic accident data set from 2016 to mid 2022 in the Portuguese district of Setúbal is used. Clustering, Random Forests and C5.0 rule models are some of the techniques used to select the most influential factors and represent them in rule sets.

Overview of the
RULE GENERATION APPROACH

**Data Processing**
Data treatment: Handling null values, encoding, etc...

**Clustering**
Creation of groups of data points based on the degree of similarity between them.

**Feature Selection**
To calculate the feature importances of each variable and select the most influential.

**C 5.0 Decision Tree**
C5.0 is a decision tree model capable of creating sets of rules to define the problem.

## MAIN GOALS

**Main goal:** Identification of main factors that influence accident severity.

**Data sources:** BEAV is a statistical notation tool that is filled by supervising authorities and is the main source of information about traffic accidents.

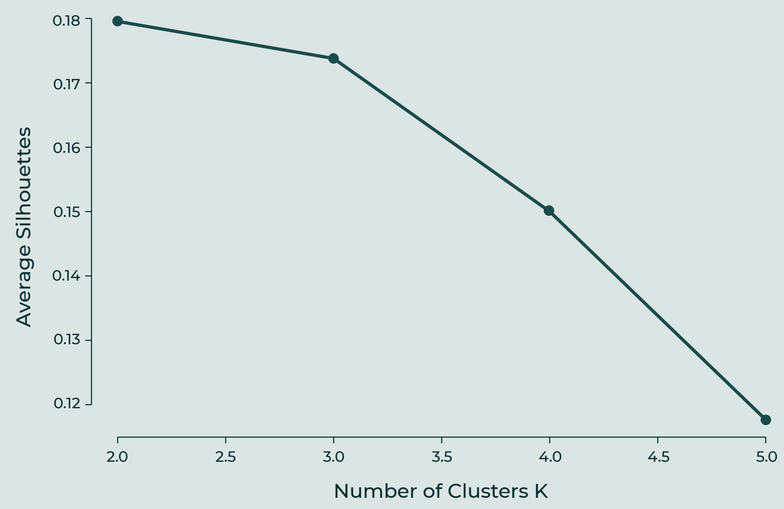
The data used for this specific study includes temporal and spatial variables as well as accident characteristics.


## CLUSTERING

Clustering or grouping of data is the creation of groups of data defined by their degree of similarity. The main objective of this step is to provide more homogeneity to the cluster data, which could allow the rule model to give more concise rules.

Both K-means and Agglomerative were applied to the data set and silhouette index was used to evaluate the resulting clusters, which ranges from [-1,1], with -1 indicating poor consistency within clusters and 1 indicating excellent consistency within clusters.

When both algorithms were applied to the data set and the silhouette index of the resulting groups was calculated, both approaches had a similar maximum value. Because none of the methods produced better indexes than the other in this circumstance, hierarchical clustering was selected. When deciding on the number of groups, it was discovered that the two-cluster models produced the best silhouette index results for both algorithms.




## RULE BASED MODEL

The algorithm used in this approach was the C5.0 algorithm, which creates decision trees and can also generate rules. The rule sets represent a simplified version of the important information found the in the decision trees. In this specific experiment we consider an accident with victims, any accident that results in injuries or death.

| Rule | Obs. N° | Error % | Class |
|---|---|---|---|
| Pedestrian running over in urban areas, no motorcycles involved | 460 | 15% | With victims |
| Motorcycles involved | 1575 | 37% | With victims |
| Pedestrian running over in rural areas | 263 | 16% | With victims |
| Collisions and crashes, no motorcycles involved | 18434 | 16% | No victims |

Table 2: Rule results of Cluster 1 (containing the majority of the data) for 2016-2019 data

| Rule | Obs. N° | Error % | Class |
|---|---|---|---|
| Motorcycles involved | 1133 | 24% | With victims |
| Pedestrian running over | 247 | 26% | With victims |
| Collisions and crashes, no motorcycles involved | 6030 | 9% | No victims |

Table 3: Rule results for Cluster 2 for 2016-2019 data

Although the rule sets are easy to understand, these rules represent only a portion of the dataset and they shouldn't be interpreted as a general rule of the entire data and should be tested for its veracity afterwards. As such, another experiment was conducted which skipped the clustering step. The following tables show the different rule sets for the pre-pandemic accidents and post-pandemic accidents.

| Rule | Obs. N° | Error % | Class |
|---|---|---|---|
| Pedestrian running over in urban areas | 681 | 17% | With victims |
| Pedestrian running over | 977 | 37% | With victims |
| Accidents with motorcycles, no light vehicles involved | 915 | 19% | With victims |
| Accidents with light vehicles, with a hit and run | 3803 | 4% | No victims |

Table 4: Some rule results for 2016 - 2019 data before pandemic

| Rule | Obs. N° | Error % | Class |
|---|---|---|---|
| Pedestrian running over in urban areas | 365 | 33% | With victims |
| Pedestrian running over and escape | 52 | 30% | With victims |
| Accidents with motorcycles and no escape | 1266 | 30% | With victims |
| Hit and run | 2502 | 6% | No victims |

Table 5: Some rule results for 2020 - mid 2022

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

## FEATURE SELECTION

Feature selection or variable selection consists in identifying the most important/discriminatory variables in order to simplify the models and eliminate non-impact variables. We use Random forests to rank the importance of variables. Based on the Mean Decrease Accuracy (MDA) and mean decrease in gini values, we select the most influential variables.

| Variables | Mean Decrese Accuracy | Mean Decrese in gini |
|---|---|---|
| Motorcycles | 140,08 | 1257,50 |
| Type of accident | 99,24 | 808,18 |
| Escape (hit and run) | 78,38 | 304,04 |
| Light vehicles | 77,22 | 848,28 |
| Damaged road | 62,91 | 151,20 |
| Heavy vehicle | 62,91 | 181,52 |
| Central separator | 62,91 | 217,14 |
| Urban / rural | 62,91 | 220,24 |
| Road type | 62,91 | 705,82 |

Table 1: Result with the most influential variables (only the 10 higher ranked)

## CONCLUSIONS

The feature selection gives us an initial idea of what factors are most important, and allows for the less relevant variables to be absent in the rule sets. The generated rule sets, although requiring a deeper examination of each rule, provides a good pointer as to what factors influence accident severity.

Overall the models point out that motorcycles or similar vehicles and pedestrian accidents are more likely to result in accidents with victims. An interesting find is that hit and runs have a lower chance of resulting in accidents with victims. Another interesting find is that, although post-pandemic accidents see an increase in pedestrian accidents, this increase mostly results in no victims.

## ACKNOWLEDGMENTS

This research was funded by the Portuguese funding agency, FCT – Fundação para a Ciência e a Tecnologia, grant number FCT DSAIPA / DS / 0090 / 2018.

## FUTURE WORK

With this contribution, a digital decision support tool will be developed to support GNR (National Republican Guard) of Setúbal to make more informed decisions regarding road accidents prevention.

## BIBLIOGRAPHY

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