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

Some factors that influence the nature of road traffic accidents

Paulo Infante [1, 2,*], Anabela Afonso [1, 2, *], Gonçalo Jacinto [1, 2, *], Leonor Rego [2, *], Pedro Nogueira [5, 6], Marcelo Silva [5, 6], Vitor Nogueira [3, 4], José Saias [3, 4], Paulo Quaresma [3, 4], Daniel Santos [4], Patrícia Góis [7] 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] 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

Between 2016 and 2019, there were 136654 accidents with victims in Portugal, resulting in 2466 deaths.

In this period, Setúbal was one of the districts with the highest number of accidents.

Using data collected by the GNR, and from the identification of municipalities with identical profiles, multinomial logistic regression and machine learning models, were used to identify some determinants for the nature of road accidents in the district of Setúbal.

Of the 28103 accidents resulted:

7536 LIGHT INJURIES

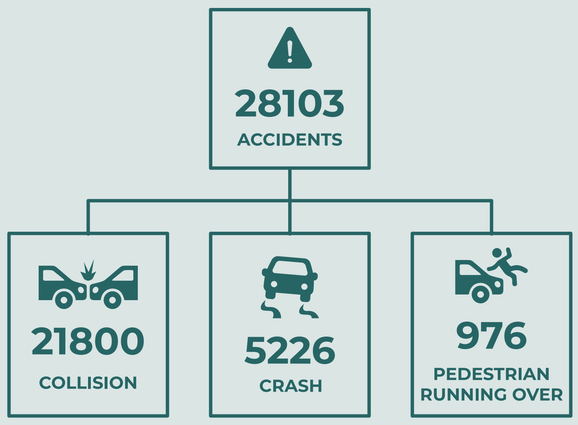
452 SERIOUS INJURIES

183 DEATHS

MAIN GOALS

1st: To identify the determinants for the nature of road accidents that occurred in the district of Setúbal.

2nd: To evaluate the performance of some machine learning classification models and compare them with the multinomial model.



DATA

· **Data type:** spatial, temporal, environmental, vehicles involved, actors, roads, others (typology, severity, traffic intensity, ...).

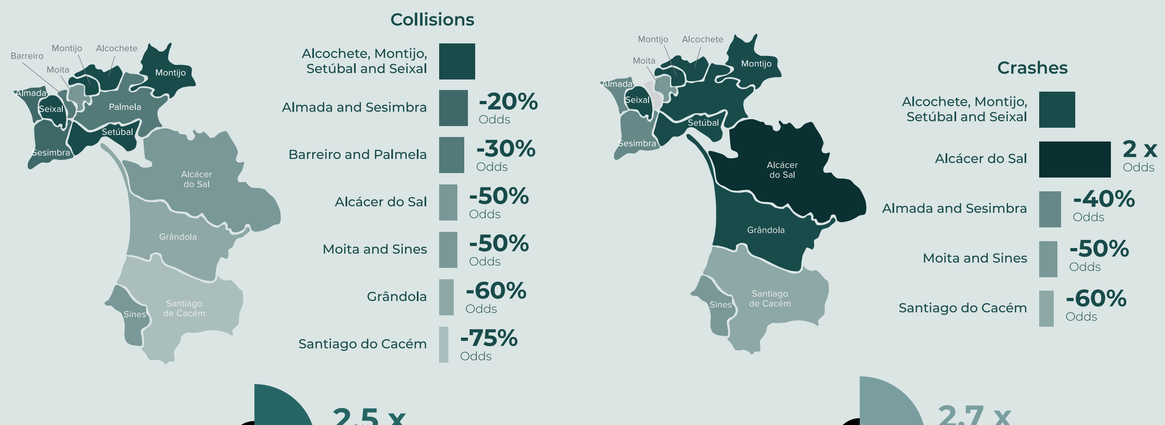
· **Data source:** GNR of Setúbal, Statistical Bulletin of Road Accidents (BEAV), ANSR, IPMA, IP, Waze Portugal, Altice, IMT.

· Data base: 947 variables

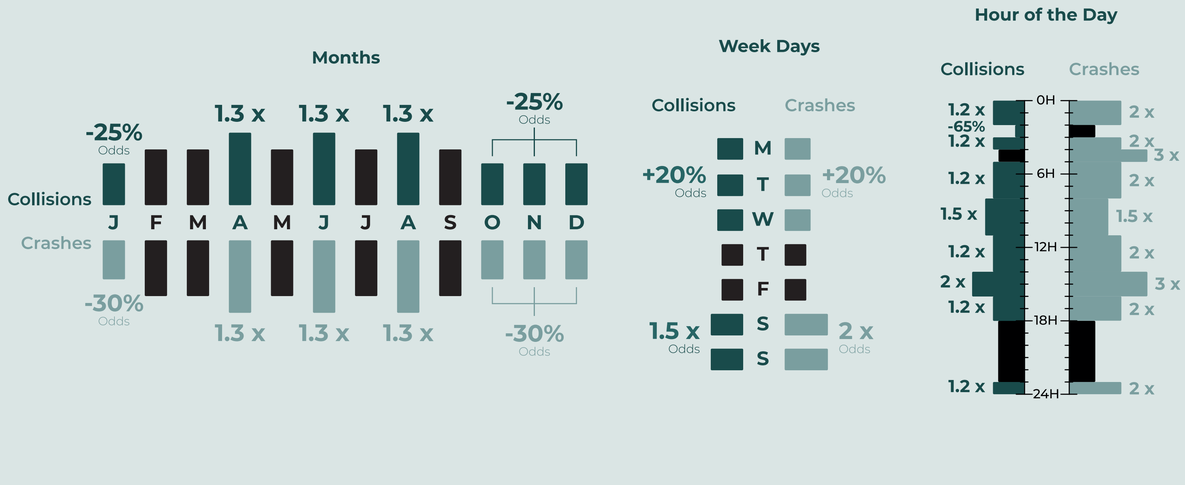
MULTINOMIAL MODEL

Response variable: Nature - Pedestrian running over (0) vs Collision (1) vs Crash(2)

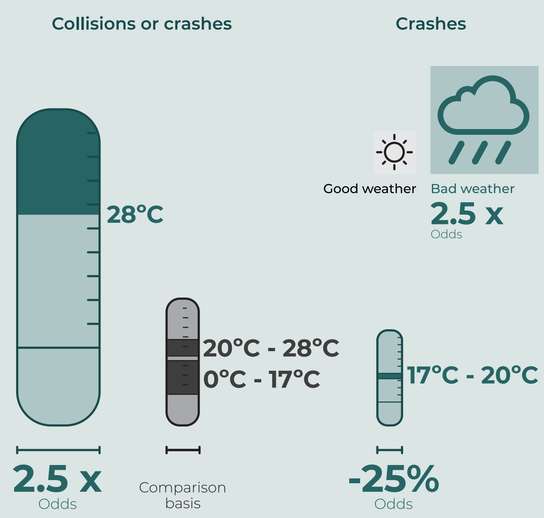
GEOGRAPHIC FACTORS



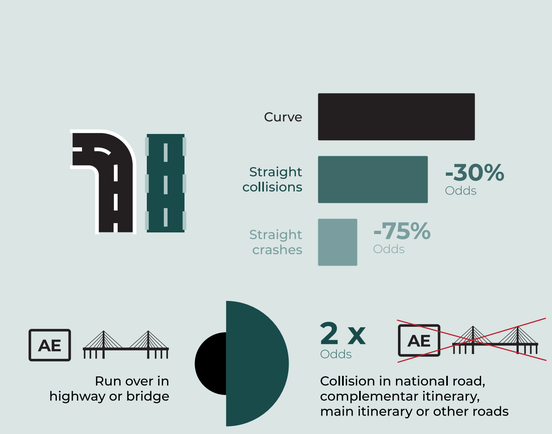
TIME FACTORS



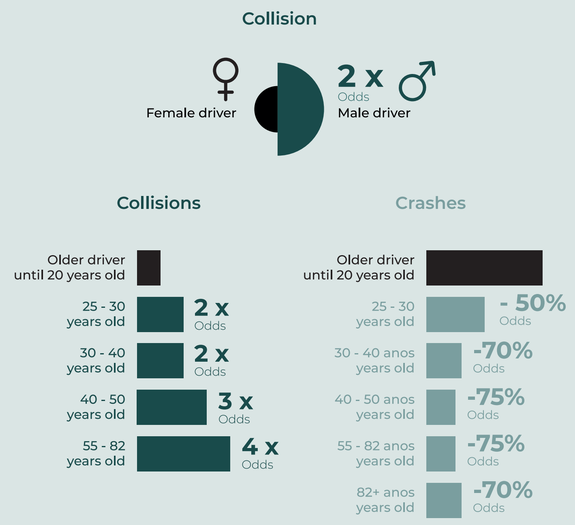
WEATHER FACTORS



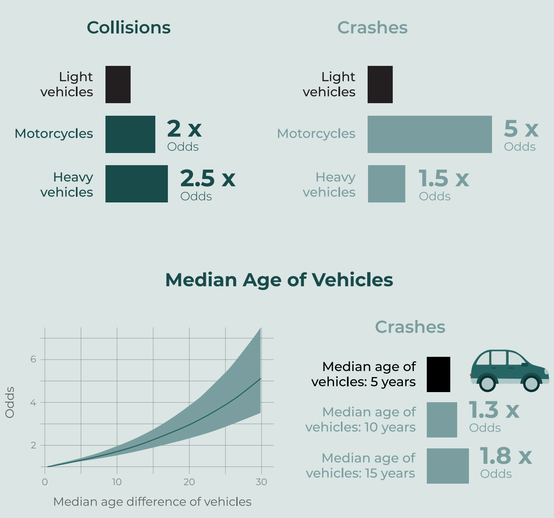
ROAD CHARACTERISTICS FACTOR



DRIVER FACTOR



VEHICLE FACTORS



MACHINE LEARNING

Goal: To evaluate the performance of some machine learning classification models and compare them with the multinomial model.

Machine learning algorithms:

· **Random Forest** (36 randomly elected variables at each split)

· **SVM** (tuning C parameter = 1)

· **Naive Bayes** (Constant Laplace smooth value)

· **C5.0** (rules model with 20 trials)

· **KNN** (K=5)

Performance measures relative to the set of variables used in the modeling of the multinomial model

Data base partition: 80% observations for training and 20% for testing;

Results presented for the same set of explanatory variables used in the multinomial model (71 predictions after using design variables);

Results were compared with and without SMOTE technique (used to correct data imbalance).

	With SMOTE						Without SMOTE					
	Multinomial	Random forest	SVM	C5.0	Naive-Bayes	KNN	Multinomial	Random forest	SVM	C5.0	Naive-Bayes	KNN
Accuracy	0,578	**0,900**	0,591	0,889	0,514	0,632	0,818	0,827	0,829	0,830	0,697	0,798
Sensitivity (Pedestrian running over)	0,143	**0,986**	0,125	0,986	0,347	0,840	0,000	0,005	0,000	0,022	0,255	0,000
Sensitivity (Collision)	0,705	**0,845**	0,741	0,817	0,738	0,504	0,964	0,957	0,981	0,963	0,793	0,985
Sensibility (Crash)	0,667	**0,913**	0,667	0,913	0,369	0,660	0,369	0,414	0,322	0,406	0,364	0,138
Especificity (Pedestrian running over)	0,965	**0,991**	0,964	0,988	0,831	0,836	1,000	0,999	1,000	0,999	0,879	1,000
Especificity (Collision)	0,592	**0,941**	0,601	0,941	0,533	0,821	0,324	0,364	0,280	0,362	0,501	0,129
Especificity (Crash)	0,751	**0,903**	0,763	0,889	0,881	0,784	0,961	0,955	0,979	0,961	0,904	0,981
Macro F1	0,524	**0,893**	0,512	0,871	0,512	0,680	-	0,457	-	0,519	0,469	0,511
MCC	0,316	**0,816**	0,305	0,780	0,270	0,501	0,391	0,336	0,240	0,351	0,298	0,286
Cohen Kappa	0,306	**0,814**	0,293	0,779	0,262	0,493	0,961	0,321	0,203	0,312	0,398	0,223

ACKNOWLEDGMENTS

This work is partially funded by National Funds through FCT – Fundação para a Ciência e Tecnologia within the scope of the MOPREVIS project "FCT DSAIPA/DS/0090/2018".

FUTURE WORK

Provide the Setúbal GNR with a digital tool to support decision-making, allowing for the optimization and management of resources for prevention.

BIBLIOGRAPHY

· Agresti, A. (2012). *Categorical Data Analysis*, 3rd edition, Wiley.
· Chawla, N. V., Bowyer, K W., Hall, L. O., *et al.* (2002). *SMOTE: Synthetic Minority Over-sampling Technique*, Journal of Artificial Intelligence Research, 16: 321-357.
· Kuhn, M., Johnson, k. (2013). *Applied Predictive Modeling*, Springer.
· Infante P, Jacinto G, Afonso A, *et al.* (2022). *Comparison of Statistical and Machine-Learning Models on Road Traffic Accident Severity Classification*. Computers. 11(5):80. https://doi.org/10.3390/computers11050080