SEVEN publicações acadêmicas
EM PARCERIA COM: IEMS UNIASSELVI

# Exploratory and predictive analysis of data on traffic accidents with victims treated at a university hospital

**Lino Marcos da Silva[1], Daniel Moreira Lopes[2], Marco Antonio de Jesus Saturnino[3], Márcio Ruan da Silva Carvalho[4].**

## ABSTRACT

Accidents involving motorcycles account for more than 75% of the number of traffic accidents in the country, causing great damage to society. In Petrolina-PE, in the São Francisco Valley region, according to data from the University Hospital of the Federal University of the São Francisco Valley (HU/Univasf), the number of patient admissions due to this type of accident is higher than the number of other types of accidents. An immediate consequence of this high rate is the overcrowding of hospital beds. However, the consequences go beyond the physical limitations of hospitals and reach other sectors. For example, the economic sector is directly affected by the negative effects of these results, due to the early removal or suppression of the workforce. In fact, it is estimated that, in 2017, the impact on the economy due to motorcycle accidents was approximately 200 billion reais. In addition, health authorities, when analyzing the effects of traffic accident rates on the health systems of large Brazilian cities, are unanimous in stating that this is a serious problem for hospital networks and consider that there is a global epidemic of traffic accidents, with motorcycles being the main villain.

**Keywords:** Traffic accidents, University hospital, Public health.

## INTRODUCTION

Accidents involving motorcycles account for more than 75% of the number of traffic accidents in the country, causing great damage to society [1, 2, 3]. In Petrolina-PE, in the São Francisco Valley region, according to data from the University Hospital of the Federal University of the São Francisco Valley (HU/Univasf), the number of patient admissions due to this type of accident is higher than the number of other types of accidents. An immediate consequence of this high rate is the overcrowding of hospital beds. However, the consequences go beyond the physical limitations of hospitals and reach other sectors. For example, the economic sector is directly affected by the negative effects of these results, due to the early removal or suppression of the workforce. In fact, it is estimated that, in 2017, the impact on the economy due to motorcycle accidents was approximately 200 billion reais [3]. In addition, health authorities, when analyzing the effects of traffic accident rates on the health systems of large Brazilian cities, are unanimous

[1] UNIVASF – BA
[2] UNIVASF – BA
[3] UNIVASF – BA
[4] UNIVASF – BA

in stating that this is a serious problem for hospital networks and consider that there is a global epidemic of traffic accidents, with motorcycles being the main villain [4].

In view of this, it is important to develop studies that enable a technical and scientific understanding of this problem, that help the planning of preventive and care activities and, above all, promote traffic education with a view to reducing the number of this type of accident. The problem of traffic accidents, particularly those involving motorcycles, has been studied under various aspects and in several countries [5, 6, 7, 8, 9, 10, 11].

Therefore, it is relevant to carry out studies on traffic accidents both at the national and local levels, under the most varied approaches and using the various methodological resources available. In this sense, it was proposed to carry out in this work the study of data related to traffic accidents involving motorcycles whose patients were treated at the HU-Univasf, using as methodology the collection, treatment and analysis of data, the development of exploratory and predictive mathematical models, using data science techniques.

## OBJECTIVE

The general objective of this study was to investigate the existence of relationships between sociodemographic factors of traffic accident victims, the severity of accidents and hospital evolution. The specific objectives were: to carry out an exploratory analysis of data on accidents involving land transport and with victims treated at the HU-Univasf; identify relationships between variables associated with traffic accidents and hospitalization; identify factors involved in the causes of accidents involving motorcyclists; and develop predictive mathematical models using accident variables.

## METHODOLOGY

The study was characterized as an exploratory research, with a quantitative approach and with the use of techniques and tools of descriptive and inferential statistics, Data Science and Artificial Intelligence. The dataset provided by the HU covered the period between 01/01/2018 and 12/31/2022 and contained 41 data out of a total of 40,345 victims.

Methodologically, an approach based on the fundamental principles of data science was adopted. These principles comprise a set of effective steps in the manipulation and analysis of large data sets, with the purpose of extracting relevant information to support decision-making processes. The techniques were implemented through the *sklearn*, *statsmodels* and *pycaret libraries*, the Python programming language and also through the *Orange software*.

Initially, the pre-processing of the data was carried out, where they were submitted to filtering, cleaning and selection procedures, with the objective of obtaining a set of complete and high-quality data.

This step included the treatment of missing values, the detection of outliers or anomalies, the conversion of categorical variables into numerical formats, as well as the normalization of data, among other pertinent procedures. At the end of this process, a dataset with 40,213 instances was obtained.

In order to gain a preliminary understanding of the dataset as well as the generation of *insights*, an exploratory analysis was performed. In a subsequent step, there was a selection of the variables that would compose the predictive models. Subsequently, machine learning techniques were used to generate models. As the last step in this process, evaluation metrics were calculated to quantify the performance of the models generated during data processing.

The machine learning technique employed in this work was classification, which consists of the use of algorithms called classifiers to identify characteristics of variables and label output data within a context, in distinct classes.

*Ensemble Learning* is a technique in the field of machine learning that involves combining multiple learning models to improve the accuracy and overall performance of prediction or classification. Rather than relying on just a single model, *Ensemble Learning* leverages the diversity of different models for more reliable and robust results.

The concept behind *Ensemble Learning* is based on the idea that different models can learn patterns and relationships in data in different ways. By combining these models, it is possible to take advantage of individual strengths and mitigate their weaknesses, resulting in a prediction that is generally more accurate and stable.

There are several approaches to *Ensemble Learning*, the most common of which are *Bagging* and *Boosting*. In the former, several identical models are trained on different datasets, generated by sampling with replacement of the original data. The results of these individual models are then combined, usually through voting or averaging, to form the final prediction. In the second, the technique consists of training the models sequentially. Each subsequent model is trained to correct the errors of the previous model, focusing on the examples that were misclassified. This allows the models to focus more on the hard cases and gradually improve their accuracy. In this work, one algorithm of each type was used.

## RESULTS

### EXPLORATORY ANALYSIS

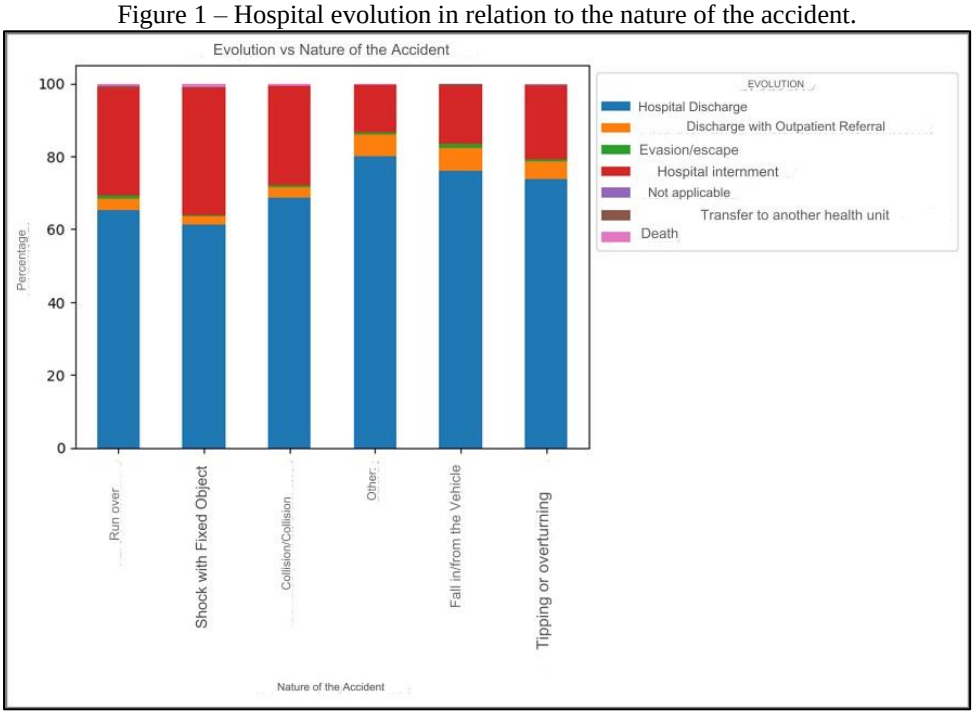
The results of the exploratory analysis of the data, among other findings, corroborated the information that the care of victims of accidents involving motorcycles is the main demand of the hospital, representing about 74% of the total. The sociodemographic profile of the victims revealed that 76.9% of the total number of patients treated were male; 52.5% were between 20 and 40 years old; and 94.8% of the victims were identified as brown.

Regarding accidents, it was observed that 73.7% involve motorcycles, 38.4% occur on Saturdays or Sundays; whereas about 65.0% occur in the afternoon or evening; whereas the most frequent injuries are fracture (21%), cut/laceration (11.8%) and traumatic brain injury (8.9%); whereas the most affected parts of the body are the lower limbs (33.9%), upper limbs (26.9%) and head (14.1%); that patients were discharged from the hospital in 71.5% of the cases, admitted to hospital in 22.9% and discharged with outpatient referral in 4.1%. Deaths accounted for about 0.5%. It is important to highlight that, in absolute terms, the number of hospitalizations and deaths corresponds, respectively, to about 9220 and 201, in the period analyzed.

Hospital evolution is a variable of interest, since it is responsible for indicating hospital overcrowding. In this sense, in this exploratory analysis, we sought to identify variables that were related to the victim's hospital evolution. The possible values for these variables are: hospital discharge, discharge with outpatient referral, evasion/escape, hospitalization, transfer to another health unit, and death, in addition to the option 'not applicable'.

The first variable analyzed in relation to hospital evolution was the Nature of the Accident, and the results are presented in **Figure 1**.

Figure 1 – Hospital evolution in relation to the nature of the accident.

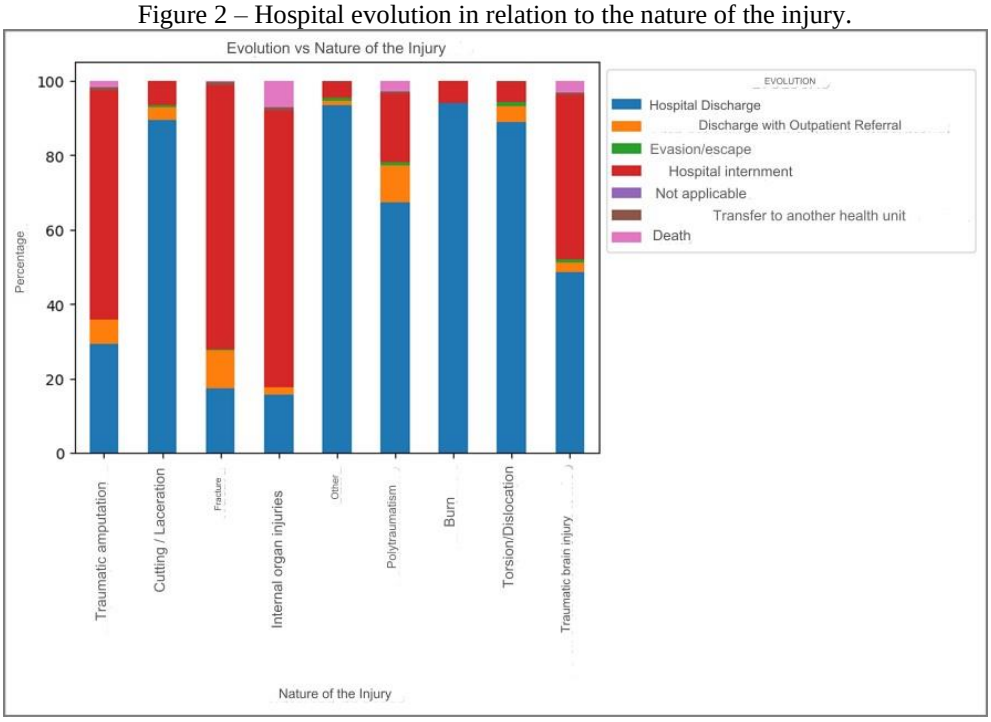



Source: Survey data

In this analysis, it was observed that the victims of accidents involving collision with a physical object, being run over and colliding/colliding were the ones who most evolved to hospitalization.

When comparing the **nature of the** injury suffered by the victim with his/her hospital evolution, it was observed that there was a greater evolution to hospitalization in the cases of victims who had traumatic amputation, fracture, internal organ injuries or traumatic brain injury, as shown in **Figure 2**. However, it is important to highlight that traumatic amputation and internal organ injuries accounted for less than 0.7% of the accidents. In relation to patients who suffered fractures, there was also a significant number of discharges with outpatient referral. In other words, in these cases, even after discharge, there is still a demand for hospital services.

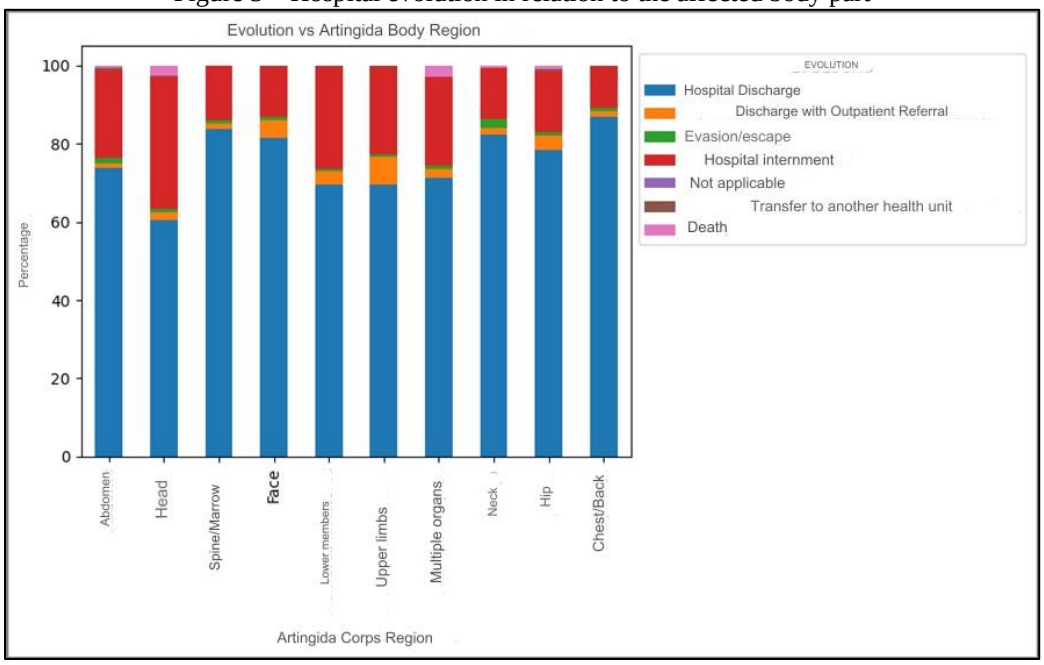
Figure 2 – Hospital evolution in relation to the nature of the injury.




Source: Survey data

Another variable that was analyzed in relation to hospital evolution was the Affected Body Region. For this variable, the results, which are presented in **Figure 3**, indicate that the highest rates of hospitalization occurred when the head and lower limbs were injured in the accident. It is important to note that these two regions of the body are the most affected in about 48% of accidents. It was also observed that the highest rates of hospital discharge with outpatient follow-up occurred in cases in which the face or upper limbs were affected.

Figure 3 – Hospital evolution in relation to the affected body part




Source: Survey data

Other variables, such as **Alcohol, Driving Conditions** and **Accident Occurrence Zone,** may also indicate an influence on the hospitalization of victims.

In fact, a higher rate of hospitalization was observed for victims who used alcoholic beverages, unlicensed beverages, and victims of accidents that occurred in rural areas.

On the other hand, the variables **Sex** and **Shift of the Accident** did not show significant differences in relation to hospital evolution, although accidents that occurred in the morning or with females had a slightly lower hospitalization rate.

Regarding the **Zone of Occurrence** of accidents, it was observed that, although most of the accidents occurred in the urban area (61.6%), the accidents occurred in the rural area resulted in a greater severity, since the hospitalization rate for this type of accident is approximately double the hospitalizations that occurred as a result of accidents in the urban area.

Considering that most of the accidents involved motorcyclists, when looking at the **Use of Helmets**, it was observed that, among the fatal victims of this type of accident, 63.3% were not wearing helmets; and that in 65.6% of the cases in which there was traumatic brain injury, the victim was not using this safety equipment. This data reinforces the perception that helmets play an important role in the conservation of life.

When the nature of the bodily injury was analyzed based on the use or not of the helmet, a higher number of traumatic brain injuries and cuts/lacerations were observed when the equipment was not being used. In addition, it was evidenced that the head and face were the most affected parts of the body in the cases of victims who were not wearing helmets.

PREDITIVE MODELS

Based on the results of the Exploratory Analysis, it was sought, through a Classification model, to categorize the **Hospital Evolution** of the Patient based on the **Nature of the Injury** suffered and the **Body Region Affected**.

Thus, for the construction of the model, the patient's Hospital Evolution was considered as the dependent variable, while the Affected Body Region and the Nature of the injury resulting from the accident were considered as independent variables.

To validate the choice of these variables, a correlation analysis was performed using Pearson's correlation coefficient between the independent variables. The calculated coefficient was 0.5, indicating a weak correlation according to the theory. Therefore, it is justifiable to incorporate both variables as independent.

To construct the classifier model, the dataset was initially divided into a training set and a test set, in the proportion 2/3 and 1/3, respectively.

Next, the training set was subjected to the following machine learning algorithms: *Random Forest* classifier, which is a *bagging* technique; and the *Gradient Boosting* classifier, which corresponds to a *boosting approach*.

To evaluate the classifiers, the metrics Accuracy, Recall, Precision and *F1-Score*, which are commonly used, were used. The results obtained are presented in **Table 1** and indicate that the Gradient Boosting classification model presented a superior performance, with an efficiency of around 85%.

Table 1 - Evaluation of the proposed model

| | **Accuracy** | **Recall** | **Precision** | **F1-score** |
|---|---|---|---|---|
| **Gradient Boosting** | 0,8512 | 0,8512 | 0,8227 | 0,8349 |
| **Random Forest** | 0,8403 | 0,8403 | 0,8031 | 0,8222 |

Source: Survey data

## FINAL THOUGHTS

In this study, based on an exploratory and predictive analysis carried out on a dataset of traffic accident victims treated at a university hospital, it was observed that some factors such as the area of occurrence of the accident, the nature of the accident, the driver's license condition, the consumption of alcoholic beverages, the nature of the injury and the region of the body affected may be related to the victim's hospital evolution.

However, it was observed that the variables indicating the type of injury suffered and the region of the body affected had a greater impact on the classification of hospital evolution. Thus, these two

variables were used in the development of a predictive model to predict the hospital evolution of each victim. From there, a classification model was created, which obtained an accuracy rate of 85%.

The results corroborate previous studies that indicated that victims of accidents involving motorcycles are the majority of hospital attendances and that the majority of victims are males aged between 20 and 40 years [12]. On the other hand, they pointed out different results in other aspects, such as the zone of greatest occurrence [12].

In addition, it was evidenced that, although most of the accidents occurred in the urban area, those that occurred in the rural area required more hospitalization, indicating the need for preventive actions directed to this segment of the population. It is important to highlight that Petrolina-PE and Juazeiro-BA are two municipalities with a great agricultural vocation and with a large part of the population living or working in rural areas.

For a better understanding of how sociodemographic factors influence the demand for hospital services, a more in-depth study is needed, including more data on victims and accidents.

Finally, it should be noted that this study serves as a starting point for understanding the profile of the victims treated at the HU-Univasf, the traffic accidents of which they were victims, as well as to map risk conditions that contribute to the severity of the accident, thus providing support for more assertive decision-making in the context of public health policies in the context of trauma involving road traffic accidents. In addition, other predictive models can be analyzed in order to obtain better results.

The researchers would like to thank the HU-Univasf, Facepe and Fapesb for their support in the development of this work.

## REFERENCES


Social Security (PREVIDÊNCIA SOCIAL). (2017). The impact of traffic accidents on social security. Epidemiological Bulletin 2. Available at: http://sa.previdencia.gov.br/site/2017/03/3%C2%B0-Quadrimestre-Boletim-2-Impacto-Acidentes-de-Tr%C3%A2nsito.pdf. Accessed March 29, 2023.

TV SENADO. (2022). Brazilian traffic: 45 thousand deaths and R$ 50 billion in economic losses. Available at: https://www12.senado.leg.br/tv/programas/em-discussao/2022/09/transito-brasileiro-45-mil-mortes-e-r-50-bilhoes-de-prejuizo-economico. Accessed March 29, 2023.

Accidents in traffic have a R$ 199 billion impact on the economy. (2018, August 14). Correio Braziliense. Available at: https://www.correiobraziliense.com.br/app/noticia/brasil/2018/05/14/interna-brasil,680658/acidentes-no-transito-tem-impacto-de-r-199-bi-na-economia.shtml. Accessed May 9, 2022.

Ramos, M. (2017). Traffic accidents have a direct impact on the public health network in Rio, says secretary. O GLOBO. Available at: https://oglobo.globo.com/rio/acidentes-de-transito-tem-impacto-direto-narede-de-saude-publica-do-rio-diz-secretario-21914148. Accessed May 9, 2022.

Rodrigues, C. L., Armond, J. E., Gorios, C., & Souza, P. C. (2014). Accidents involving motorcyclists and cyclists in the municipality of São Paulo: Characterization and trends. Revista Brasileira de Ortopedia, 49(6), 602–606. https://doi.org/10.1016/j.rboe.2014.09.003

Mendonça, M. F. S., Silva, A. P. S. C., & Castro, C. C. L. (2017). Spatial analysis of urban traffic accidents attended by mobile emergency care service: A spatial and temporal cut. Revista Brasileira de Epidemiologia, 20(4), 727-741. https://doi.org/10.1590/1980-5497201700040005

Nyakyi, V., Kuznetsov, D., & Nkansah-Gyekye, Y. (2014). Mathematical model to assess motorcycle accidents in Tanzania. Mathematical Theory and Modeling, 4(9), 25-34. https://doi.org/10.5539/mtm.v4n9p25

Miškinis, P., & Valuntaite, V. (2010). Mathematical simulation of the correlation between the frequency of road traffic accidents and driving experience. Transport, 25(3), 237-243. https://doi.org/10.3846/transport.2010.29

Lhueze, C. C. Onwrah, & Uchendu, O. (2018). Road traffic accidents prediction modelling: An analysis of Anambra State, Nigeria. Accident Analysis & Prevention, 112, 21-29. https://doi.org/10.1016/j.aap.2018.01.014

Haynes, S., Estin, P. C., Lazarevski, S., Soosay, M., & Kor, A. H.-L. (2019). Data analytics: Factors of traffic accidents in the UK. In The 10th IEEE International Conference on Dependable Systems, Services and Technologies, DESSERT 2019, 5-7 June, Leeds, United Kingdom. https://doi.org/10.1109/DESSERT.2019.00017

Nour, M., Naseer, A., Alkazemi, B., & Jamil, M. A. (2020). Road traffic accidents injury data analytics. International Journal of Advanced Computer Science and Applications (IJACSA), 11(12). https://doi.org/10.14569/IJACSA.2020.0111275

Fernandes, F. E. C. V., Melo, R. A., Araújo, F. S. A., Borges, F. K. B., Holanda, O. Q., & Campos, M. E. A. (2019). Motorcycle accidents and factors associated with driver's license status. Archives of Health Sciences, 26(2), 130-135. https://doi.org/10.5935/2237-2202.20190016