A machine learning approach for pedestrian accident prediction model

Monica Meocci[1]*, Andrea Paliotto[2], Francesca La Torre[2], Alessandro Terrosi[2] & Giulia Manetti[3].
*lead presenter
[1]monica.meocci@unifi.it, University of Florence, Italy
[2]University of Florence, Italy
[3] Systra SWS, Italy

Introduction

Pedestrians represent one of the most vulnerable road user classes due to the difference in speed with vehicular flow and the absence of body protection during a collision. Therefore, a collision between a vehicle and a pedestrian too often results in severe injury or fatality.

There are many initiatives such as Vision Zero which aim to reduce accident rates, but the efforts made to date are not enough to be in line with the targets defined at national and international levels.

In Italy, in 2022, 485 pedestrians died in road accidents compared to the total number of deaths in road accidents equal to 3159 (15.35% of road deaths).

In this context, the Accident Prediction Models (APMs) represent one of the best tools available that allow road engineers and/or national Road Authorities (RAs) to relate the number and severity of accidents expected at a specific site with the geometrical and functional characteristics of the road.

The tool is spread worldwide for accidents involving vehicles but only in recent years, it has also been used for accident analysis on Vulnerable Road Users (VRUs) such as pedestrians.

The Highway Safety Manual represents one of the most used tools to assess road safety, but the accident prediction model for pedestrians gives the fatalities as a percentage of the accidents involving vehicles, and the main factor that affects the model cannot be transferred to the Italian reality. The literature provides other models but for specific sites and are difficult/not transferable to different realities.

Recently, the topic has been further investigated in terms of the variables which affect the phenomenon and the formulation of pedestrian accident prediction models. The advent of modern Artificial Intelligence (AI) techniques has made it possible to simplify some aspects of safety analysis and obtain results which, in addition to providing predictions on the number of fatalities, give us some information concerning how the different variables considered influence the topic evaluated (correlations, weight in the prediction, etc.).

In this context, the research conducted aims both to create a predictive model for pedestrian accidents along road segments based on AI algorithms and to define the variable most relevant for car-pedestrian accidents both in terms of risk increase and reduction of accident rates. The model provided is referred to 150 road sections within the city of Florence.

Research Methodology

The steps summarising the model construction of the pedestrian accident model for road sections in Florence are:

- identification of the variables affecting pedestrian accidents (e.g., number of crashes, pedestrian flows, site characteristics, etc.) All the variables are considered in the 5 years 2014-2018 according to the accident database availability;
- identification of the 150 urban road sections. All the sections are selected with similar lengths to limit the section length effect in crash occurrence. The variable is considered to change linearly with the number of crashes;
- data collection and database creation: all the road sections are analysed, and the information is collected in the database;
- due to the small number of data, the topic is evaluated as a binary classification problem such as whether an accident occurred/will occur: yes or not. The decision is consequent to the limited number of sites with a number of crashes greater than 1;
- choice of the mathematical model used as classifiers. Three different Machine Learning algorithms are used: logistic regression, decision trees and random forest;
- correlation analysis of the different chosen variables and evaluation of the model parameters;
- evaluation of the model performance with reference to accuracy, precision and recall metrics.

Two classes of independent variables are considered: binary (e.g., parking presence on the roadside, bus stop presence, crosswalks presence, etc.) and continuous (e.g., carriageway length, number of lanes, sidewalks length, traffic values, etc.).

Results

Among the three algorithms used the one that performs best is the Random Forest algorithm.
Concerning the values of Precision and Recall, a difference is observed between the value concerning the sites characterized by one accident (or more) and those with the absence of an accident. The ability of the model to predict the right result is higher in the site without accident in all algorithms used. This can be probably a consequence of the high number of labels equal to 0 compared to those with labels equal to 1, and therefore it is likely to believe that the model could have minor overfitting issues due to the high number of variables considered compared to the dataset size.
The balanced accuracy obtained on the training set and test set is equal to about 0.60 and 0.65 respectively with the maximum value of 0.68 obtained using the Random Forest algorithm. Therefore, the results are consistent.
The correlation analysis conducted in the variables used in the model highlighted the presence of correlated features, therefore the analysis is repeated to define a simplified model, excluding some correlated features. All three algorithms used gave results consistent with the first analysis conducted and the values of the balanced accuracy obtained are slightly higher than those obtained in the first analysis, confirming that a reduction in the number of variables reduces the risk of model overfitting. The most relevant variables in accident occurrence are carriageway length, pedestrian flow values and the number of points of interest in the road section considered.

Conclusions

The research conducted proposes an interesting and powerful methodology to estimate the number of accidents in different road sections considering both the road geometry and the site characteristics. The results obtained showed a good ability of the model to predict the risk of an accident occurrence. The procedure used allows us also to evaluate the influence of the considered variables in vehicle-pedestrian accident occurrence. The dataset size represents one of the most important variables to obtain reliable results, especially when the phenomenon needs to be described by a high number of features.