sustainability

MDPI

Article

A Hybrid-Stratified Approach for the Identification of Pedestrian Crash Scenarios: The Effect of Demographic Vulnerability and Spatial-Temporal Shifts in the Pre- and Post-COVID-19 Period in Italy (2010–2023)

Giuseppe Cappelli [1,2,*], Sofia Nardoianni [1,2], Mauro D'Apuzzo [1,2,*] and Vittorio Nicolosi [3]

[1] Department of Civil and Mechanical Engineering, University of Cassino and Southern Lazio, Via G. Di Biasio 43, 03043 Cassino, Italy; sofia.nardoianni@unicas.it

[2] European University of Technology EUt+, European Union, B-1049 Brussels, Belgium

[3] Department of Enterprise Engineering "Mario Lucertini", University of Rome Tor Vergata, Via del Politecnico 1, 00133 Rome, Italy; nicolosi@uniroma2.it

* Correspondence: giuseppe.cappelli1@unicas.it (G.C.); dapuzzo@unicas.it (M.D.)

Abstract

Pedestrian safety represents a critical priority for the development of sustainable urban mobility systems. This study proposes an innovative methodological framework integrating supervised and unsupervised learning techniques with econometric modeling to identify and interpret risk scenarios. Using the Italian national dataset from 2010 to 2023, an XGBoost model has been initially trained and tested. Then, SHapley Additive exPlanations (SHAPs) have been applied to highlight contributing factors. Using the resulting SHAP values, a K-Means clustering algorithm was finally employed to segment crashes into homogeneous clusters. For each cluster, a Generalized Linear Mixed Model incorporating geographic random intercepts and temporal random slopes was calibrated. Through this hybrid-stratified approach, three risk scenarios have been identified, primarily driven by demographic vulnerability. For elderly pedestrians, the involvement of heavy vehicles nearly doubles the odds of a fatal outcome. Crash dynamics varied significantly: heavy vehicles and speeding nearly double the fatality risk for elderly pedestrians; nighttime represents a severe hazard for adults (OR = 3.87) and youths (OR = 7.99), with the latter also highly penalized by unsafe road behaviors (OR = 3.12). From a spatio-temporal perspective, random effects revealed that the Islands (Sicily and Sardinia) are the most critical macro-areas (+55.2% baseline risk for adults) and the North-West the safest. Furthermore, the COVID-19 pandemic mitigated fatal risk for young pedestrians nationwide, had a neutral impact on the elderly, and for adults was protective in Southern regions but corresponded to higher odds of mortality in the North, reflecting altered traffic dynamics.

Keywords: pedestrian safety; machine learning; econometric model; XGBoost; GLMM Model; clustering; K-Means

Check for updates

Academic Editors: Marco Guerrieri and Larysa Neduzha

Received: 18 June 2026
Revised: 30 July 2026
Accepted: 31 July 2026
Published: 4 August 2026

1. Introduction

Road safety remains a critical societal issue worldwide, responsible for approximately 1.19 million fatalities annually. Recognizing the urgency of this crisis, the European Union established the Road Safety Policy Framework 2021–2030 [1], which aims to halve road deaths and serious injuries by 2030 compared to 2019 levels. Although the Italian National Road Safety Plan adopted this identical objective, current outcomes fall significantly short [2]. In 2024, Italy recorded 3030 road fatalities, 31% higher than the target. In response to this challenging scenario, United Nations 2030 Agenda recognized road safety as a key prerequisite to promote Sustainable Development, directly supporting key Sustainable Development Goals (SDGs): halving global road deaths and injuries (SDG 3.6); developing resilient and sustainable infrastructure (SDG 9.1); ensuring safe, accessible, and sustainable transport systems that prioritize VRUs (SDG 11.2); and providing universal access to safe public spaces (SDG 11.7) [3].

Pedestrians have been considered the most vulnerable road users since, due to the disparity in physical characteristics compared to vehicles, they are subject to much more serious or fatal injuries when crashes occur [4]. For decades, the analysis of risk factors associated with crash severity has been carried out using econometric models such as Multinomial and Ordered Logit models [5–10]. With the advent of more advanced Machine Learning models, researchers' attention has shifted toward these approaches [4,11–26]. Unlike the previous models, however, a common historical criticism concerns their limited interpretability. Consequently, the need emerged to employ additional techniques, such as SHAP analysis, to make these models interpretable and unveil their black-box nature. The justified use of Machine Learning has mainly been driven by the ability of these models to capture the strong nonlinear relationships between the dependent and independent variables. Recently, a dual trend has emerged among scholars: on the one hand, a renewed shift toward more complex econometric models, such as Mixed Logit Models with heterogeneity in mean and variance [27]; on the other hand, the development of hybrid models that combine the main strengths of both econometric and Machine Learning approaches [20,21].

The state of the art shows that XGBoost is among the best-performing models for predicting the severity of pedestrian crashes, due to its scalability and efficiency in handling large amounts of data [28]. In a recent study of Wu et al. [18], the authors compared the performances of XGBoost and Random Forest (RF) in variable selection. The results show that XGBoost has been proven as more reliable than RF for pedestrian crash data: the most critical risk factors that have been found are posted speed limit, visibility conditions, pre-crash circumstances, pedestrian location, distraction and alcohol impairment. In another study from the same authors [29], a small dataset has been used to create injury severity models. Despite this limitation in dataset size, the XGBoost reaches quite good performances, highlighting that nighttime hour crashes show a higher probability of fatal outcomes. The scenario that has an average probability of fatal crashes occurs at intersections during weekend nights with adverse weather conditions. Another common problem that emerges from pedestrian injury severity models is related to the imbalances of datasets and the poor interpretability of Machine Learning Algorithms. In the study of Hasanat et al. [30], the authors have introduced an XGBoost with loss function balancing and undersampling techniques such as RDPVR (Random Data Partitioning with Voting Rule). Unlike other studies, researchers avoided synthetic oversampling techniques, such as SMOTE (Synthetic Minority Over-sampling Technique) [31], to preserve the integrity of real behavioral patterns, and also used SHAP to interpret the results of models. High vehicle speed, poor lighting, heavy vehicle involvement, and driver age emerged as the strongest risk factors of fatal outcomes. SHAP analysis revealed a decrease in the risk of serious crashes in the period 2020–2023 compared to 2014–2016, suggesting a measurable impact of recent safety policies in Jordan.

Other studies try to highlight the most important features related to pedestrian injury severity. In the study of Chang et al. [32], the authors aimed to predict the effects of the built environment on pedestrian fatalities, overcoming the limitations of typical aggregate data. The application of XGBoost and SHAP shows that the most critical factors for mortality risk were high residential density, road segment betweenness, and network accessibility.

https://doi.org/10.3390/su18157911

Other studies have focused on the effectiveness of rapid flashing light indicators (RRFBs) on the severity of pedestrian crashes [4]. The authors, using XGBoost, SHAP, and econometric models, found that installing RRFBs significantly reduces the severity of injuries in nighttime and pedestrian crashes. XGBoost showed an accuracy of 97% for night crashes, much higher than the statistical model (73.8%). Still comparing the results of XGBoost, LightGBM and CatBoost and other econometric models (together with SHAP), it was found that CatBoost achieved the best accuracy (95.72%) compared to the other models tested. Critical factors for serious/fatal crashes during nighttime hours at intersections include pedestrian/driver intoxication, sudden crossing behavior, and inadequate lighting [33]. In line with what has been stated above, the study by Wisutwattanasak et al. [34] investigated the temporal instability of crashes in Thailand. The authors promote a comparison between SVM, Gradient Boosting, AdaBoost, XGBoost and CatBoost. Again, Catboost, unlike previous studies, was found to be the most accurate model. The SHAP analysis instead highlighted how mortality is significantly higher during the evening rush hour (06:00 pm–09:00 pm). Other scholars, for example, have introduced recent optimizations of XGBoost. Chen et al. [35], proposed an MSCPO-XGBoost (XGBoost optimized via an algorithm inspired by crested porcupine behavior) with the application of SMOTE to balance gravity levels and SHAP for interpretability. The MSCPO-XGBoost model outperformed SVM, Random Forest, and CNN with an accuracy of 83.57% and an AUC of 92.82%. The main predictors were found to be engine displacement, vehicle mass, and driver age.

Another aspect that has recently received particular attention in the scientific community is the role played by the age of the pedestrian. In a study that analyzed factors influencing injury severity for older pedestrians (over 65 years) [14], comparing the performance of XGBoost (while using SHAP to visualize the marginal contributions of variables with multinomial logistic regression (MNL)), the authors found that in addition to the better performance of XGBoost, vehicle speed emerged as the most critical factor for elderly pedestrians, followed by lighting conditions (high risk at dawn, between 5 and 7 am). Similar results have been recorded in a study conducted in Seoul [36]. The objective was to discover how aggressive driving behaviors influence crashes involving elderly pedestrians. Using XGBoost and SHAP, the authors found how aggressive driving behaviors (rapid acceleration/deceleration, speeding) are dominant factors in both the frequency and mortality of older pedestrians. Other studies analyzing the injury severity of elderly pedestrians have highlighted the effectiveness of tree models in managing data complexity (particularly XGBoost) and confirm the crucial role of speed limits and driver behavior in protecting the elderly [37].

It is also worth highlighting how, in some recent studies on pedestrian analysis in national contexts, there is a tendency to identify geographic factors. In a study conducted in the United States, the authors integrated Emerging Hotspot Analysis and XGBoost using partial dependence plots to visualize the effects of the variables on pedestrian crash data in North Caroline [38]. In this study, the authors focused exclusively on a single state to understand the role of the variables analyzed in that particular context. The analysis shows that agricultural land development, highways (Interstate/US routes), road piracy, drunk driving (driver or pedestrian), and urban areas are the main risk factors. Still in the US context, Sulle et al. [39] identify the key factors of pedestrian fatalities by comparing the five US states with the highest risk with the five with the lowest risk to highlight the different geographical contexts. Age has emerged as the most impactful predictor: increasing age significantly increases the risk of fatality, especially in high-risk states. Alcohol and drug use are strong predictors everywhere, but their impact is even more pronounced in low-risk states. Nighttime crashes show higher SHAP values, indicating a much greater likelihood of fatal outcomes due to poor visibility. Furthermore, distraction (both driver and pedestrian, the latter often linked to smartphone use) is a critical factor in all the regions analyzed. Therefore, this study aims to provide a two-fold contribution. It introduces a novel methodological framework that integrates Machine Learning (XGBoost), SHAP values, unsupervised clustering, and Generalized Linear Mixed Models (GLMMs) to objectively identify and evaluate crash risk scenarios. Then, it applies this framework to a highly relevant and understudied national case study: the Italian context serves as an ideal testing ground for this methodology.

Further findings emerging from the analysis of the recent literature highlight how the COVID-19 pandemic has introduced profound temporal instability, making predictive models in the pre-pandemic period less accurate for the current context [40]. Although the overall volume of crashes decreased during lockdowns, the severity of injuries often increased, a phenomenon attributed to riskier driving behavior, such as higher speeds [41,42]. In the study by Tamakloe et al. [43], the authors note that back injuries are positively correlated with fatal outcomes after the pandemic, while they showed an opposite effect (risk reduction) at intersections before the pandemic. After the pandemic, the presence of pedestrian crossings reduced the likelihood of fatality more effectively at intersections than in other road segments.

The literature analyzed focuses mainly on the classification and evaluation of the importance of individual risk factors in isolation. Although some studies mention specific interactions (e.g., between lighting and pedestrian position or between speed and road type) or a priori defined risk scenarios [18,33,34], a structured methodology to uniquely define the different risk scenarios is lacking. Unlike existing studies, this study introduces a Hybrid-Stratified methodological approach for identifying risk scenarios, combining supervised and unsupervised algorithms with econometric models. Specifically, starting from the critical role of demographic differences highlighted in the previous literature, this approach stratifies the analysis through distinct pedestrian age clusters (e.g., young, adults, and elderly). This allows us not only to predict risk within specific demographic groups, but to map complex scenarios to support more targeted and proactive road safety interventions across the Italian territory. To highlight this contribution, the overall methodology will be described in the following Section 2, while the results will be shown in Section 3. Finally, in Section 4, the main results will be discussed.

2. Methodology

As stated in the Section 1, this paper aims to analyze pedestrian crash risk scenarios in Italy, from 2010 to 2023. A hybrid-stratified approach is proposed by leveraging the most implemented models in the traffic safety field (see Figure 1).

In this context, hybrid is intended as a methodology that combines methods from different analytical fields (Machine Learning and Econometrics); while stratified means that the risk scenario analysis is performed on homogeneous sub-populations. However, unlike traditional ex ante stratification based on raw data, this approach employs a data-driven, post hoc segmentation based on SHAP values to objectively define these homogeneous groups. In particular, Supervised, Unsupervised, and Econometric Models were implemented with the aim of identifying the riskiest scenarios in Italy. The hybrid-stratified methodology starts with the creation of Crash Injury Severity Models, with the help of XGBoost, and then, with SHAP analysis, the relationships between dependent and independent variables are discovered. By collecting all SHAP values, a K-Means Algorithm was applied to these values with the aim of identifying clusters/scenarios. Once the most frequent clusters were identified, for each of them, a GLMM Model was estimated to understand the relationships between pedestrian crash injury severity and contributing factors within that scenario. In the following subsections, the main models that constitute this hybrid-stratified approach will be shown in detail.

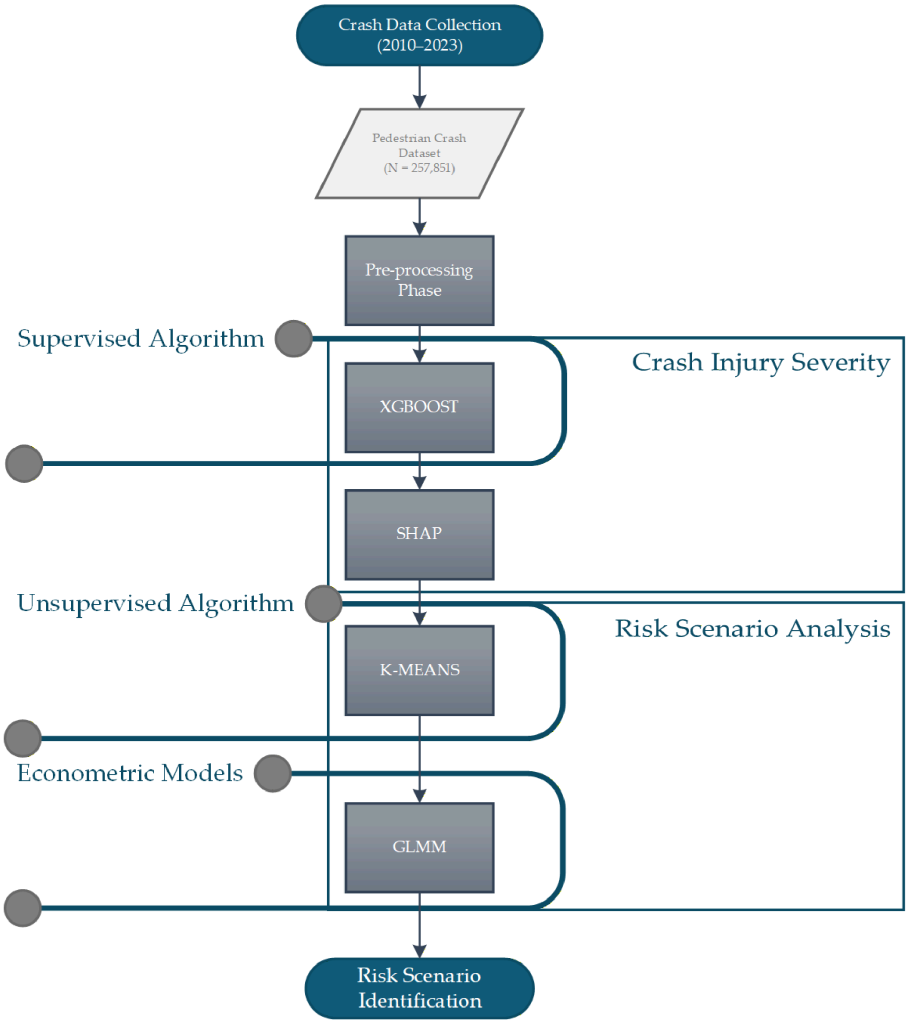


Figure 1. Flow chart of the Risk Scenario Identification Methodology.

2.1. Data Description

The dataset used for the analysis of risk scenarios in Italy covers 14 full years, spanning from 2010 to 2023. The utilized data were obtained from the Italian National Institute of Statistics (ISTAT) [44]. A total of 257,851 records concerning pedestrian-involved crashes were extracted from the database. Figure 2 displays the chronological trend of the analyzed crashes. As can be observed, prior to 2020, the annual average of pedestrian crashes stood at 19,334. In addition to the drastic and significant drop recorded in 2020 (where total pedestrian crashes fell to 12,714), a gradual recovery followed the pandemic crisis, although without fully reverting to the pre-pandemic scenario (recording 18,481 pedestrian crashes in 2023). A significant aspect worth highlighting is that while the average fatality rate (the ratio of fatalities to total pedestrian crashes) sits at 2.78% on average, this proportion rose to 3.04% in 2020.

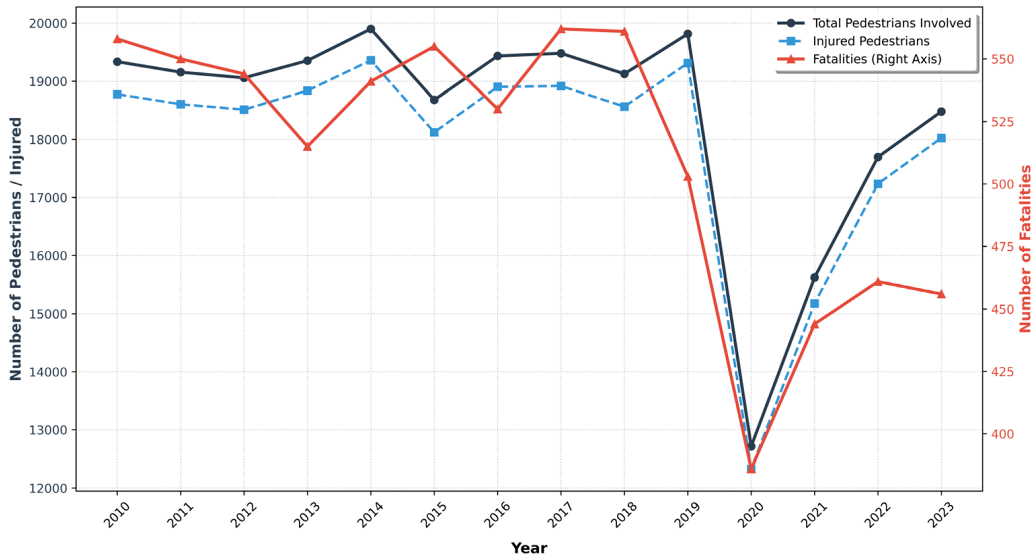


Figure 2. Historical trend of pedestrian crashes in Italy (2010–2023).

2.2. XGBoost Algorithm

The XGBoost Algorithm was first introduced in 2016 by Chen and Guestrin [25], and it is widely recognized for its scalability and speed, but also for its own capability of handling numerical and categorical data. In addition, it can handle nonlinearity within a more interpretable structured framework [45]. The algorithm is an extension of gradient boosting methods [46]. The boosting technique consists of simple models that are recursively added to the previous ones, while in gradient boosting, new models are added in order to predict the residuals of previous models. An optimized version of gradient boosting is XGBoost, which introduces some penalties to reduce tree complexity and avoid overfitting. It is structured as an ensemble and supervised model based on a decision and classification tree (CART). Despite other Machine Learning methods and Neural Networks, it works on an additive sequential criterion, and this means that the outcomes depend on previous results on other trees. The XGBoost Algorithm does not build independent trees (as in Random Forest) but trains each new tree f_t with the specific task of correcting residual errors made by the set of previous trees. The final prediction $\hat{y}_i^{(t)}$ for the observation x_i at iteration t could be obtained from the following Equation (1):

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) \tag{1}$$

where:

- t is the number of iterations;
- k is the number of generated trees;
- x_i is the current observation;
- $f_k(x_i)$ is the tree function;
- $\hat{y}_i^{(t)}$ is the overall prediction.

The same Equation (1) could also be written as follows (Equation (2)):

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{2}$$

where:

- $\hat{y}_i^{(t-1)}$ is the prediction at $t-1$ iteration;
- $f_t(x_i)$ is the tree function at t iteration.

A peculiar aspect of XGBoost lies in its penalized objective function, which penalizes the model's predictive ability based on the structural complexity of the trees to prevent overfitting. The objective function to be minimized at iteration t is defined as (Equation (3)):

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{t} \Omega(f_k) \tag{3}$$

where:

- n is the total number of observations in the dataset;
- $\mathcal{L}^{(t)}$ is the penalized objective function at iteration t;
- $l(y_i, \hat{y}_i)$ is the loss function;
- $\Omega(f_k)$ is the regulation term.

The regulation term is defined as follows (Equation (4)):

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{4}$$

where:

- j is the leaf score;
- T is the total number of leaf nodes;
- w_j is the leaf weight;
- γ is the penalty parameter for reducing the complexity of trees;
- λ is the regularization parameter for the L_2 norm.

2.3. SHapley Additive exPlanations (SHAPs) Algorithm

One main criticism of Supervised and Unsupervised Machine Learning Algorithms is their poor interpretability, often referred to as "black box" nature. This issue is related to the fact that there is no explainable mathematical structure that defines the method itself, so relationships between independent variables and outcomes are somehow difficult to discover, as well as the inner prediction process. To overcome this limitation, the SHAP method has been widely used both for regression and classification tasks [15,20,32]. The SHAP method is based on game theory: every feature in the model is considered a player [47,48]. By using Shapley values, the marginal contribution of features to the model output is evaluated. The aim of SHAP analysis is to find a simple linear model that is able to explain the prediction of the complex one (Equation (5)):

$$g(z') = \phi_0 + \textstyle\sum_{i=1}^{m} \phi_i {z'}_i \tag{5}$$

where:

- $g(z')$ is the explanation model;
- ${z'}_i$ is a representation of the input features x;
- m is the number of features;
- ϕ_0 represents the average prediction of the model;
- ϕ_i is the Shapley value for feature i.

The Shapley value could be evaluated as follows (Equation (6)):

$$\Phi_i = \sum_{S \subseteq \{1,\ldots,m\} \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{m!} [f_x(S \cup \{i\}) - f_x(S)] \tag{6}$$

where:

- $|S|$ is the number of elements in the subset S;
- $f_x(S)$ is the model prediction;
- $f_x(S \cup \{i\})$ is the model prediction after adding feature i.

2.4. K-Means Algorithm

The fundamental step of this methodology lies in the application of the K-Means algorithm not on the original feature space, but on the SHAP value space [48]. While conventional clustering identifies groups of observations with similar characteristics, clustering on SHAP values identifies groups of observations that share the same risk factors. This way of clustering allows us to solve the problem of the unit of measurement of the different variables, also taking into account the nonlinear interactions between the quantities first highlighted with XGBoost and then made explainable with SHAP.

The K-Means algorithm is one of the most widely used unsupervised clustering techniques in data analysis [49,50]. Unlike predictive models, the goal of K-Means is not to estimate a target variable, but rather to identify latent structures within the dataset by grouping the n observations into K distinct clusters based on their similarity. This data partition into K distinct clusters could be obtained by minimizing the intra-group variance, known as Within-Cluster Sum of Squares (WCSS) (Equation (7)):

$$J(C,\mu) = \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \mu_k\|^2 \quad (7)$$

where:

- C is the set of clusters $\{C_1, C_2, \ldots, C_k\}$;
- μ is the set of cluster centroids $\{\mu_1,\ \mu_2, \ldots, \mu_k\}$;
- x_i is the observation in the dataset;
- $\|x_i - \mu_k\|^2$ is the squared Euclidean distance of the dataset observation and the cluster centroid.

As previously stated, in the standard formulation (Equation (7)), K-Means works on the feature space. If these values are replaced with SHAP values ϕ_i, the objective function becomes the minimization of the variance of the predictive contributions, i.e., how much the single factor weighed on the severity of the crash (Equation (8)):

$$J(C,\mu) = \sum_{k=1}^{K} \sum_{i \in C_k} \|\phi_i - \mu_k\|^2 \quad (8)$$

where:

- C is the set of clusters $\{C_1, C_2, \ldots, C_k\}$;
- μ is the set of cluster centroids $\{\mu_1,\ \mu_2, \ldots, \mu_k\}$;
- ϕ_i is the SHAP value;
- $\|\phi_i - \mu_k\|^2$ is the squared Euclidean distance of SHAP value and the cluster centroid.

2.5. Generalized Linear Mixed Model (GLMM)

The Generalized Linear Model (GLM) is an extension of traditional linear regression (LR) [51]. With this model, it is possible to relax the normality and constant error-variance assumptions, providing a more flexible tool for predicting both discrete and continuous variables [52]. To account for spatial unobserved heterogeneity and potential correlation within geographical groups, a Generalized Linear Mixed Model (GLMM) specification with Random Intercept and Slope was adopted.

Given the binary nature of the severity in this particular study, the model could be simplified as follows (Equations (9) and (10)) [53,54]:

$$p_{ij} = P\left(Y_{ij} = 1 \middle| x_{ij},\ \mathrm{T}_{ij}, \gamma_{0j}, \gamma_{1j}\right) \tag{9}$$

$$logit(p_{ij}) = ln\left(\frac{p_{ij}}{1 - p_{ij}}\right) = x_{ij}\beta + \gamma_{0j} + \gamma_{1j}\mathrm{T}_{ij} \tag{10}$$

where:

- Y_{ij} is the binary response variable related to observation i and group j, in particular, $Y \in \{0, 1\}$;
- p_{ij} is the conditional probability that an event happens;
- x_{ij} is the vector containing the explanatory variables;
- β is the vector of estimates;
- $x_{ij}\beta$ is the Fixed Effect term;
- γ_{0j} is the Random Intercept for macro-area group j, representing the baseline geographic risk during the pre-COVID period;
- γ_{1j} is the Random Slope for macro-area group j, capturing the area-specific deviation in risk due to the pandemic;
- T_{ij} is a dummy variable equal to 1 if the observation belongs to the post-2020 period, and 0 if it belongs to the pre-2020 baseline.

The random effects γ_{0j} and γ_{1j} are assumed to follow a bivariate normal distribution with mean zero and an unstructured variance-covariance matrix Σ:

$$\begin{pmatrix} \gamma_{0j} \\ \gamma_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma\right) \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix}\right) \tag{11}$$

By extracting the exponential function from Equation (10), the Odds could be expressed as follows (Equation (12)):

$$Odds_{ij} = \frac{p_{ij}}{1 - p_{ij}} = \left(e^{x_{ij}\beta}\right)\left(e^{\gamma_{0j}}\right)\left(e^{\gamma_{1j}\mathrm{T}_{ij}}\right) \tag{12}$$

where:

- $e^{x_{ij}\beta}$ represents the Baseline Odds;
- $e^{\gamma_{0j}}$ is the Geographic Risk Multiplier, defining the structural safety baseline of the j macro-area during the pre-pandemic period;
- $e^{\gamma_{1j}\mathrm{T}_{ij}}$ is the temporal COVID-19 multiplier: it functions as an area-dependent temporal multiplier that quantifies how the COVID-19 pandemic amplified or mitigated the baseline severity in that specific macro-area.

3. Results

As previously shown in the Section 2, the Hybrid-Stratified Approach provides that the first step in defining risk scenarios is the development of a severity model. For this purpose, an XGBoost classifier has been trained and tested, and its predictive performance has been evaluated on these two sets of observations (see Table 1).

Table 1. Main performances of the XGBoost Model.

Metrics	Training Set	Test Set
Accuracy	0.9426	0.9391
Specificity (Injured)	0.9597	0.9579
Sensitivity/Recall (Fatal)	0.3417	0.2820
Precision (Injured)	0.9808	0.9790
F1 score (Injured)	0.9701	0.9682
F1 score (Fatal)	0.2485	0.2049
F1 score (macro)	0.6093	0.5866
Balanced Accuracy	0.6507	0.6199

It is worth noting that no balancing techniques (e.g., oversampling and undersampling) have been applied to address class imbalance between fatal and injured observations. Some class weights have been used to prioritize fatal crashes over injuries, without adding and generating synthetic observations. As shown in Table 1, the model achieves a global accuracy of 0.9391 and a balanced accuracy of 0.6199, demonstrating some robustness to class imbalance in the pedestrian crash dataset. By comparing the model's performance with the Confusion Matrices in Figure 3, the proposed model shows limited ability to identify fatal observations (404 out of 1433 actual fatalities), with a Sensitivity/Recall value for the Fatal class of 0.2820. The F1 score on the minority class is 0.2049 in the testing phase.

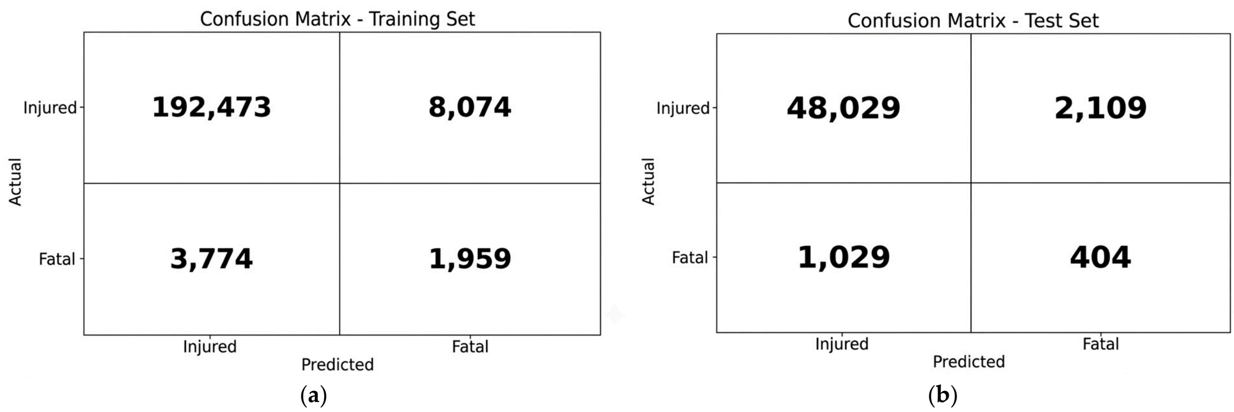


Figure 3. Confusion matrices for the XGBoost Algorithm evaluated on (**a**) training and (**b**) test sets.

To discover the relationship founded by XGBoost, SHAP analysis was performed. A global SHAP summary plot is proposed, revealing the importance and the impacts of risk factors (see Figure 4).

The proposed analysis identifies pedestrian demographic vulnerability as one of the most important contributing factors to severe pedestrian crashes. According to SHAP results, elderly pedestrians (defined as pedestrians aged 65 or older) are strongly associated with increased pedestrian crash severity. Red dots indicate high SHAP values. When these points are on the right side of the SHAP graph, it means that higher values of that variable are associated with greater severity in pedestrian crashes. This suggests a proportional relationship between the independent and dependent variables. On the contrary, young pedestrians (under 25 years old) are less likely to sustain high levels of crash injury severity when they are involved in crashes. Another pedestrian demographic characteristic that plays a relevant role is pedestrian gender. As it is possible to observe in Figure 4, female pedestrians are less likely to suffer fatal injuries compared to males. Beyond demographics, the model also captures other critical scenarios. High values for variables such as *Area_Rural*, *Hour_Evening*, and *Vehicle_Type_Heavy_vehicle* exhibit positive SHAP contributions, confirming that high-speed contexts, poor visibility during nighttime hours, and high kinetic energy associated with heavy vehicles are strong pedestrian risk factors. *Area_Urban* presents a positive SHAP trend, acting as a protective factor likely due to lower impact speeds. It is also worth noting that the presence of intersections and one-way carriageways is a mitigating factor for pedestrian injury severity. Finally, the *Macro_Area_Nord-Ovest* feature predominantly clusters on the negative side of the SHAP axis, confirming its role as a mitigating geographic factor.

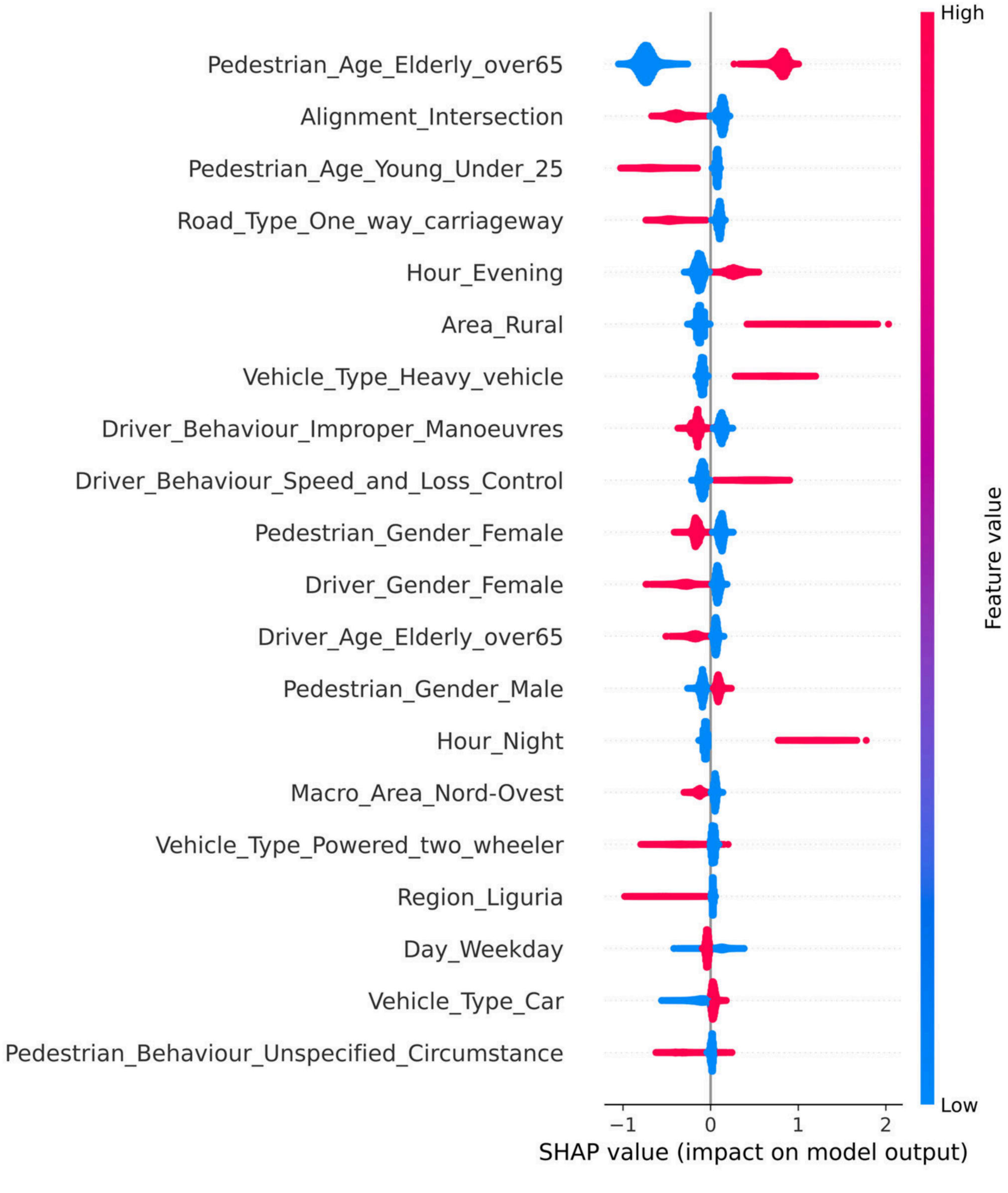


Figure 4. Graphical results of SHAP analysis.

Applying the K-Means algorithm to SHAP values from the XGBoost analysis segmented the entire set of observations into three pedestrian risk scenarios. The algorithm clusters observations based on pedestrian demographic profiles, particularly pedestrian age, rather than other environmental or dynamic crash factors (such as day, time, vehicle type, and so on). This result shows that age is the most significant predictor of pedestrian crash injury severity. Because age variables dominate the SHAP feature space, it is expected that K-Means clustering in this space will primarily reproduce age-based partitions. Table 2 summarizes the cluster characteristics and scenarios. It is worth noting that clusters are numbered and classified according to the average fatality rate (fatal crashes divided by the total number of observations in that cluster/risk scenario).

Table 2. Cluster characteristics.

Cluster/Scenario	Fatality Rate [%]	Cluster Size	Demographics	Other Characteristics
0 (Elderly)	5.23	79,554 (33.35%)	Elderly >65 (100%), Female (54.37%)	Urban (96.37%), Clear Weather (83.08%), Car (75.99%), Weekday (80.50%)
1 (Adults)	1.80	113,541 (47.60%)	Adults 25–65 (97.96%), Female (53.96%)	Urban (94.29%), Clear Weather (80.50%), Car (73.71%), Weekday (80.56%)
2 (Young)	0.90	45,433 (19.05%)	Young <25 (100%), Male (51.61%)	Urban (95.50%), Clear Weather (81.89%), Car (75.48%), Weekday (79.95%)

Regardless of the demographic differences among scenarios, the vast majority of crashes occurred under standard, everyday conditions: in urban environments (ranging from 94.29% to 96.37%), during clear weather (over 80%), on weekdays (approximately 80%), and involving vehicle car drivers (around 75%). Since environmental characteristics are almost the same across all pedestrian crashes in the three scenarios, the fatality rate shows that demographic characteristics again play a main role. Cluster 2 (young pedestrians under 25 years) could be considered the baseline scenario, characterized by high physical resilience, with a mortality rate of only 0.90%. Cluster 1 (adult pedestrians aged 25–65 years) represents the median scenario in terms of fatality rate (1.80%), but accounts for the largest share of total pedestrian crashes (47.60%). Conversely, Cluster 0 reveals a higher vulnerability of elderly pedestrians (over 65 years) than others. While experiencing the same baseline conditions, their mortality rate spikes to 5.23%. This indicates that an elderly pedestrian faces a lethal risk factor over six times higher than a young pedestrian under the same conditions.

After the identification of the most frequent scenarios by using the K-Means Algorithm, a stratified modeling approach was implemented. To choose the number of clusters, the trade-off between Inertia and Silhouette Score was analyzed (see Table 3). The value K = 2 has the highest silhouette score (0.3994), indicating more separated clusters, but provides a less detailed representation of the data structure. As the number of clusters increases, Inertia progressively decreases, signaling better internal group adaptation, while the Silhouette Score tends to decrease. The choice of K = 3, therefore, represents a good compromise between cluster compactness and separation capacity: compared to K = 2, a significant reduction in Inertia is obtained (from 197,110.75 to 177,024.03) while still maintaining an acceptable silhouette value (0.2713). Higher values of K instead produce more limited improvements than Inertia in the face of lower separation quality. While K = 2 yielded the highest Silhouette score, it oversimplified the crash scenarios. K = 3 was selected as the optimal compromise: it provided a substantial reduction in Inertia (elbow) and, crucially, generated the most distinct behavioral profiles from an interpretability standpoint. Although K = 4 showed a marginally higher Silhouette score than K = 3 (0.283 vs. 0.271), it resulted in fragmented sub-clusters lacking practical domain significance, reinforcing K = 3 as the most actionable choice. Three separate GLMMs have been calibrated on each cluster, with the aim of highlighting contributing factors related to that specific scenario. The goodness-of-fit and predictive performances of the three discrete models are summarized in Table 4.

As is possible to observe, despite class imbalances, the models show moderate predictive performance, with Balanced Accuracies peaking at 0.7363 for the adult cluster. The F1 score and the McFadden Pseudo R2 indicate a mixed trend. F1 score is higher for elderly and adult than for young pedestrians, while McFadden Pseudo R2 is lower for elderly than for adult and young pedestrians. The different vulnerabilities among the elderly could explain this latest trend. This means that the elderly could be involved in non-serious crashes, but, due to their higher vulnerability, they could suffer more fatal consequences.

Table 3. Results in terms of Inertia and Silhouette Score for K-Means clustering.

K	Inertia	Silhouette Score
2	197,110.75	0.3994
3	177,024.03	0.2713
4	160,739.66	0.2826
5	155,630.22	0.2100

Table 4. Performances of GLMM.

Metrics	Elderly	Adults	Young
McFadden Pseudo R2	0.0878	0.2339	0.1942
Nakagawa R2 (Marginal)	0.1417	0.2731	0.2496
Nakagawa R2 (Conditional)	0.1469	0.2856	0.3771
F1-Score	0.1543	0.0988	0.0412
Balanced Accuracy	0.6341	0.7363	0.6871
AIC	132,676.4	164,280.5	69,563.4

Finally, the model mathematically validates the spatial/geographic hypothesis introduced in this study. For all three clusters, the Nakagawa Conditional R^2 (which accounts for both fixed factors and the geographic random effect) is strictly greater than the Nakagawa Marginal R^2 (which accounts only for the fixed crash characteristics). This confirms that including the spatial random intercept (Macro_Area) and Random Slope (Year) captures unobserved heterogeneity and systematically improves goodness-of-fit, demonstrating that geography is a key feature of pedestrian crashes. As shown in Tables 5–7, geography interacts strongly with pedestrian demographics, especially age.

Table 5. Geographic random effect for Cluster 0 (elderly over 65 years) in the pre-COVID-19 and post-COVID-19 period.

Macro Area	Random Intercept (Std. Err.)	Geographic Risk OR Multiplier (Pre-COVID) [95% CI]	Random Slope C_{ij} (Std. Err.)	Geographic Risk OR Multiplier (Post-COVID) [95% CI]
Islands	0.2354 (0.0362)	1.2654 [1.1787–1.3585]	−0.0319 (0.0405)	0.9686 [0.8946–1.0487]
North-Est	0.0853 (0.0270)	1.0891 [1.0329–1.1484]	−0.0635 (0.0307)	0.9384 [0.8837–0.9966]
South	−0.0936 (0.0288)	0.9106 [0.8606–0.9635]	0.0998 (0.0329)	1.1049 [1.0359–1.1786]
Centre	−0.0314 (0.0223)	0.9691 [0.9277–1.0123]	−0.0377 (0.0255)	0.9630 [0.9160–1.0125]
North-West	−0.2275 (0.0231)	0.7965 [0.7613–0.8334]	−0.0156 (0.0260)	0.9845 [0.9356–1.0360]

Integrating random effects into the GLMM model allowed us to isolate the influence of geographic factors and assess the heterogeneous impact of the COVID-19 pandemic on the three demographic clusters (see Figure 5). These models provide useful explanatory insights into relative risk factors rather than robust predictive classification. Analyzing the basic pre-pandemic risk, a marked territorial gap is observed, albeit with different magnitudes between age groups. The Islands remain the most critical area for all categories (+26% for the elderly, +55% for adults, and 163% for young people). Young people (Cluster 2) were the most exposed group nationwide before 2020, highlighting a systematically higher risk across Italy than the average. At the other extreme, the North-West area emerges as the safest macro-area overall, consistently recording the lowest intercept values (down to an OR of 0.57 for adults).

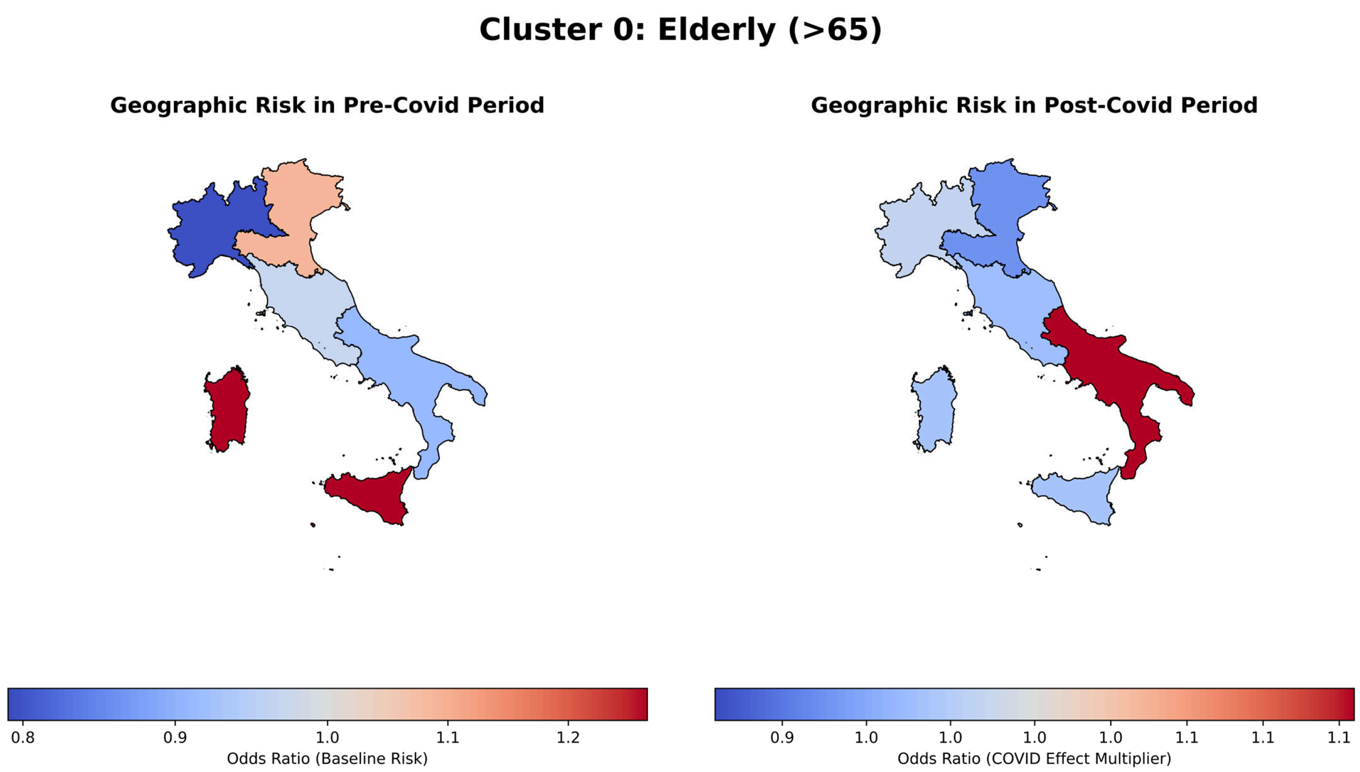
Table 6. Geographic random effect for Cluster 1 (adults between 25 and 65 years) in the pre-COVID-19 and post-COVID-19 period.

Macro Area	Random Intercept (Std. Err.)	Geographic Risk OR Multiplier (Pre-COVID) [95% CI]	Random Slope C_{ij} (Std. Err.)	Geographic Risk OR Multiplier (Post-COVID) [95% CI]
Island	0.4373 (0.0367)	1.5486 [1.4410–1.6641]	−0.1823 (0.0411)	0.8333 [0.7689–0.9031]
South	0.3169 (0.0264)	1.3728 [1.3036–1.4458]	−0.2255 (0.0307)	0.7981 [0.7515–0.8476]
Centre	−0.0293 (0.0206)	0.9711 [0.9326–1.0112]	0.0841 (0.0239)	1.0878 [1.0380–1.1400]
North-Est	−0.0413 (0.0274)	0.9596 [0.9094–1.0125]	0.0895 (0.0311)	1.0936 [1.0290–1.1622]
North-West	−0.5600 (0.0237)	0.5712 [0.5453–0.5983]	0.1888 (0.0266)	1.2078 [1.1464–1.2725]

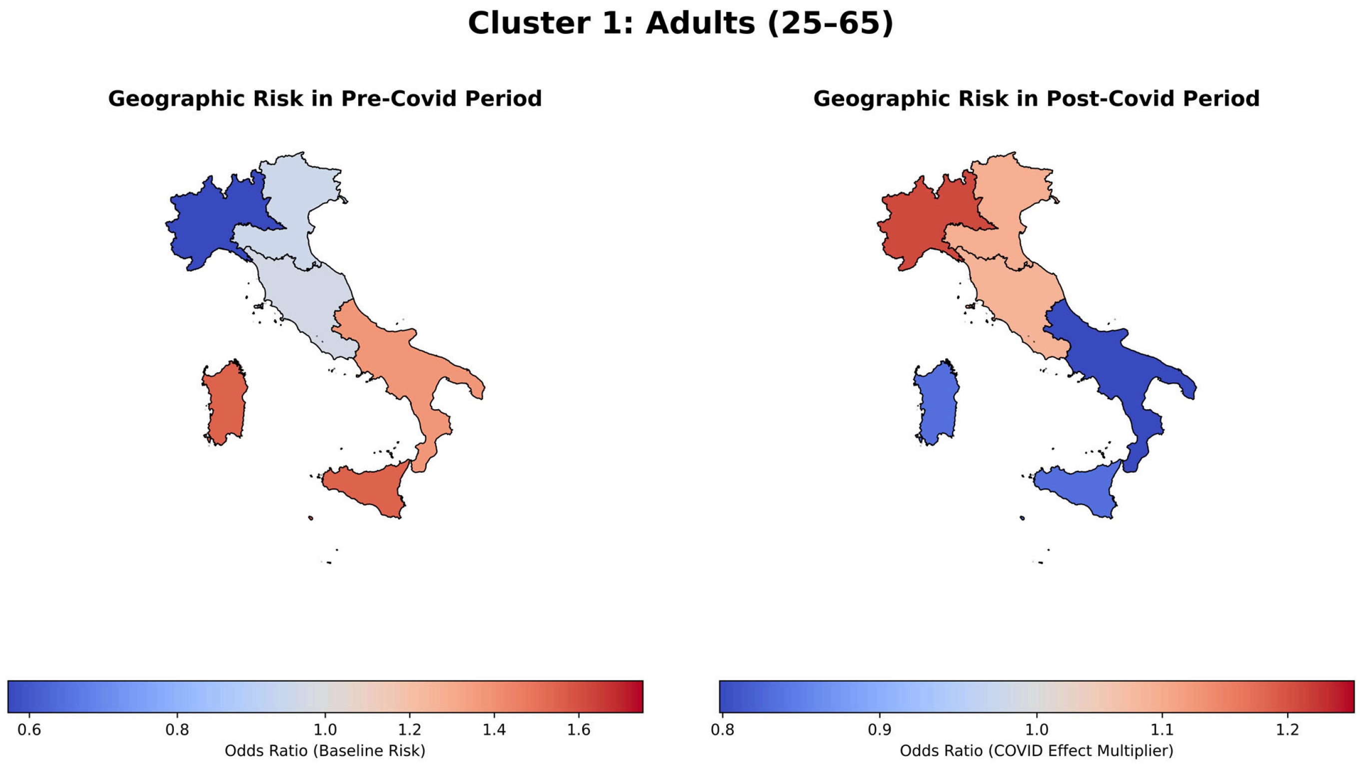
Table 7. Geographic random effect for Cluster 2 (youth under 25 years) in the pre-COVID-19 and post-COVID-19 period.

Macro Area	Random Intercept (Std. Err.)	Geographic Risk OR Multiplier (Pre-COVID) [95% CI]	Random Slope C_{ij} (Std. Err.)	Geographic Risk OR Multiplier (Post-COVID) [95% CI]
Island	0.9693 (0.0611)	2.6361 [2.3385–2.9717]	−0.9657 (0.0715)	0.3807 [0.3309–0.4380]
North-Est	0.9164 (0.0381)	2.5003 [2.3202–2.6944]	−0.9679 (0.0449)	0.3799 [0.3479–0.4148]
South	0.5106 (0.0462)	1.6663 [1.5220–1.8243]	−0.6544 (0.0534)	0.5198 [0.4681–0.5771]
Centre	0.4683 (0.0399)	1.5973 [1.4772–1.7272]	−0.3464 (0.0448)	0.7072 [0.6478–0.7721]
North-West	0.3662 (0.0352)	1.4423 [1.3461–1.5452]	−0.7540 (0.0404)	0.4705 [0.4346–0.5093]

The introduction of random gradients to account for the post-2020 temporal effect reveals how the pandemic has altered road crash dynamics. For young people, COVID-19 was associated with lower odds of severe outcomes everywhere, dramatically reducing the likelihood of a fatal outcome across regions (with reductions ranging from 29% in the Center to nearly 62% in the Islands). For the adult cluster (Cluster 1), in the southern areas (Islands and South), mobility restrictions were associated with lower odds of severe outcomes as well (OR = 0.83 and 0.80); conversely, in the northern areas (North-West, North-East, and Center), post-2020 mobility and traffic restrictions have increased mortality (up to an OR of 1.21 in the Northwest). Finally, for the elderly (Cluster 0), the pandemic impact appears more nuanced and neutral (values ranging from 0.93 to 1.10), suggesting that, for this category, intrinsic vulnerability is the most important factor, regardless of lockdown-induced restrictions.

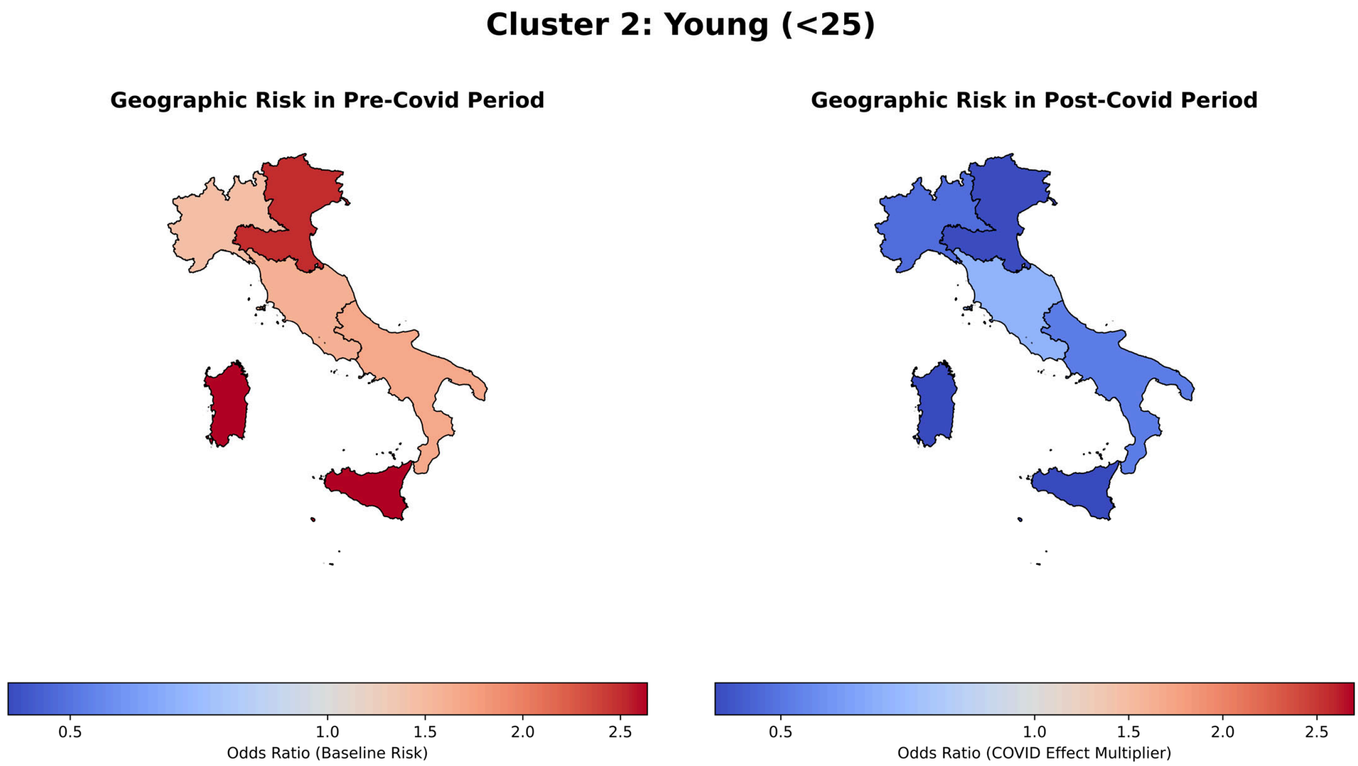


(a)



(b)

Figure 5. *Cont.*



(**c**)

Figure 5. Graphical representation of the geographic random effect in the pre- and post-COVID-19 period: (**a**) Cluster 0 (elderly, over 65 years); (**b**) Cluster 1 (adults, age between 25 and 65 years); (**c**) Cluster 2 (youths, under 25 years).

According to the specification of GLMM Models, comparing Odds Ratios (ORs) across demographic scenarios shows that the magnitude of risk factors depends heavily on the pedestrian's physical vulnerability. As shown in Tables 8–10, the involvement of heavy vehicles in crashes is a risk factor, but its magnitude varies across clusters (the OR for the elderly is 1.86; for adults, 2.48; and for the young, 3.04). Driver Speeding and Loss of Control increases the odds of a fatal outcome by roughly 2 times for the elderly (OR = 2.12) but reaches higher values for adults (OR = 2.99) and youths (OR = 2.68). Differences in pedestrian vulnerability could explain these ORs: elderly pedestrians can suffer fatal consequences even at low speeds, thereby reducing the overall statistical weight of speeding compared to younger individuals. It is worth highlighting that only variables that did not show quasi-separation were retained in the final models. For instance, the variable *AreaUrban*, despite accounting for over 94% of observations across all clusters, was excluded from the elderly and young models due to quasi-complete separation and sparse cell counts when cross-tabulated with fatal outcomes.

The analysis results also show that nighttime driving (Hour_Night) is a severe risk factor for young pedestrians (OR = 7.99) and adults (OR = 3.87), likely reflecting the combination of high speeds, poor visibility, and potential psychomotor impairment in drivers. If a crash occurs in an urban area, the probability of fatal outcomes for adults decreases (OR = 0.18) due to lower vehicular speeds, traffic calming measures, and immediate access to trauma care compared to rural settings. Behavioral variables exhibit highly divergent effects across scenarios. Unsafe Position or Movement by the pedestrian acts as a strong risk factor for the young cluster (OR = 3.07) and adults (OR = 1.08), while it appears slightly protective for the elderly (OR = 0.82). This could be attributed to the fact that elderly pedestrians generally move more slowly, allowing drivers more reaction time to brake, even if they move in an unsafe way. Across all scenarios, being a Male Pedestrian is a risk factor (ORs range from 1.61 to 1.74). It is also worth noting the mitigating effect when micromobility users are involved in a crash. However, these consistently low odds ratios should not be automatically interpreted as evidence that micromobility is protective. This result likely reflects lower kinetic impact energies than with heavy vehicles, as well as distinct reporting patterns in which only specific types of micromobility crashes are recorded. Furthermore, without exposure measures or distinctions between shared spaces and dedicated infrastructure, these estimates remain context-dependent.

Table 8. GLMM description for Cluster 0 (elderly over 65 years).

Variable	Estimate	Std. Error	z-Value	Pr (>∣z∣)	OR	95% CI
(Intercept)	−0.4616	0.0925	−4.9923	0.0000	0.6303	[0.53–0.76]
DayWeekend	0.1004	0.0164	6.1282	0.0000	1.1056	[1.07–1.14]
Road_TypeOne_way_carriageway	−0.7174	0.0474	−15.1242	0.0000	0.4880	[0.44–0.54]
Road_TypeTwo_carriageways	0.0124	0.0501	0.2468	0.8051	1.0124	[0.92–1.12]
Road_TypeTwo_way_carriageway	0.0210	0.0455	0.4624	0.6438	1.0213	[0.93–1.12]
AlignmentIntersection	−0.6164	0.0314	−19.6353	0.0000	0.5399	[0.51–0.57]
AlignmentOther	0.3666	0.0553	6.6293	0.0000	1.4428	[1.29–1.61]
AlignmentRoad_Segment	−0.1578	0.0298	−5.2944	0.0000	0.8540	[0.81–0.91]
Road_Surface_ConditionWet	0.0374	0.0250	1.4962	0.1346	1.0381	[0.99–1.09]
Traffic_SignageSignage_Present	0.1321	0.0229	5.7585	0.0000	1.1413	[1.09–1.19]
WeatherClear_Weather	0.0265	0.0238	1.1106	0.2667	1.0268	[0.98–1.08]
Vehicle_TypeHeavy_vehicle	0.6207	0.0218	28.4594	0.0000	1.8602	[1.78–1.94]
Vehicle_TypeMicro_Mobility	−1.4664	0.0656	−22.3425	0.0000	0.2308	[0.20–0.26]
Vehicle_TypeOther	−0.2010	0.0726	−2.7692	0.0056	0.8179	[0.71–0.94]
Vehicle_TypePowered_two_wheeler	−0.2549	0.0224	−11.4036	0.0000	0.7750	[0.74–0.81]
Driver_BehaviourRegular_Driving	0.4485	0.0251	17.8832	0.0000	1.5660	[1.49–1.64]
Driver_BehaviourSpeed_and_Loss_Control	0.7535	0.0213	35.3925	0.0000	2.1244	[2.04–2.21]
Driver_BehaviourUnspecified_Circumstance	0.5588	0.0854	6.5436	0.0000	1.7485	[1.48–2.07]
Pedestrian_BehaviourUnsafe_Crossing_or_Emergence	−0.0251	0.0231	−1.0872	0.2770	0.9752	[0.93–1.02]
Pedestrian_BehaviourUnsafe_Position_or_Movement	−0.1885	0.0283	−6.6593	0.0000	0.8282	[0.78–0.88]
Pedestrian_BehaviourUnspecified_Circumstance	−0.5934	0.0869	−6.8287	0.0000	0.5524	[0.47–0.66]
Driver_AgeElderly_over65	−0.1675	0.0151	−11.0956	0.0000	0.8457	[0.82–0.87]
Driver_AgeNot_applicable	−0.4888	0.0548	−8.9241	0.0000	0.6133	[0.55–0.68]
Driver_AgeYoung_Under_25	0.0890	0.0234	3.8105	0.0001	1.0931	[1.04–1.14]
Driver_GenderMale	0.3574	0.0165	21.7236	0.0000	1.4296	[1.38–1.48]
HourEvening	0.2432	0.0179	13.5872	0.0000	1.2753	[1.23–1.32]
HourMorning	−0.0903	0.0155	−5.8144	0.0000	0.9137	[0.89–0.94]
YearBefore_2020	0.1050	0.0338	3.1076	0.0019	1.1107	[1.04–1.19]
SeasonSpring	−0.2254	0.0202	−11.1665	0.0000	0.7982	[0.77–0.83]
SeasonWinter	−0.0720	0.0168	−4.2754	0.0000	0.9305	[0.90–0.96]
Pedestrian_GenderMale	0.5567	0.0132	42.1644	0.0000	1.7449	[1.70–1.79]

Table 9. GLMM description for Cluster 1 (adults between 25 and 65 years).

Variable	Estimate	Std. Error	z-Value	Pr (>∣z∣)	OR	95% CI
(Intercept)	0.4759	0.1468	3.2426	0.0012	1.6094	[1.21–2.15]
DayWeekend	0.2451	0.0145	16.8469	0.0000	1.2777	[1.24–1.31]
AreaUrban	−1.7213	0.0196	−88.0146	0.0000	0.1788	[0.17–0.19]
Road_TypeOne_way_carriageway	−1.1277	0.0355	−31.7817	0.0000	0.3238	[0.30–0.35]
Road_TypeTwo_carriageways	−0.1738	0.0367	−4.7351	0.0000	0.8405	[0.78–0.90]
Road_TypeTwo_way_carriageway	−0.4652	0.0329	−14.1446	0.0000	0.6280	[0.59–0.67]
AlignmentIntersection	−0.6543	0.0295	−22.1453	0.0000	0.5198	[0.49–0.55]
AlignmentOther	0.5343	0.0515	10.3799	0.0000	1.7062	[1.54–1.89]
AlignmentRoad_Segment	0.1101	0.0271	4.0664	0.0000	1.1164	[1.06–1.18]
Road_Surface_ConditionWet	0.1227	0.0222	5.5254	0.0000	1.1305	[1.08–1.18]
Traffic_SignageSignage_Present	0.1812	0.0213	8.5016	0.0000	1.1986	[1.15–1.25]
WeatherClear_Weather	0.1195	0.0216	5.5384	0.0000	1.1269	[1.08–1.18]
Vehicle_TypeHeavy_vehicle	0.9106	0.0204	44.6351	0.0000	2.4857	[2.39–2.59]
Vehicle_TypeMicro_Mobility	−1.7295	0.0718	−24.0722	0.0000	0.1774	[0.15–0.20]
Vehicle_TypeOther	−0.0658	0.0505	−1.3023	0.1928	0.9363	[0.85–1.03]

https://doi.org/10.3390/su18157911

Table 9. *Cont.*

Variable	Estimate	Std. Error	z-Value	Pr (>|z|)	OR	95% CI
Vehicle_TypePowered_two_wheeler	−0.7288	0.0208	−35.0405	0.0000	0.4825	[0.46–0.50]
Driver_BehaviourRegular_Driving	0.4864	0.0235	20.7213	0.0000	1.6265	[1.55–1.70]
Driver_BehaviourSpeed_and_Loss_Control	1.0973	0.0183	60.0176	0.0000	2.9960	[2.89–3.11]
Driver_BehaviourUnspecified_Circumstance	0.9699	0.0544	17.8446	0.0000	2.6377	[2.37–2.93]
Pedestrian_BehaviourUnsafe_Crossing_or_Emergence	−0.0698	0.0225	−3.1069	0.0019	0.9326	[0.89–0.97]
Pedestrian_BehaviourUnsafe_Position_or_Movement	0.0756	0.0246	3.0741	0.0021	1.0785	[1.03–1.13]
Pedestrian_BehaviourUnspecified_Circumstance	−0.6817	0.0553	−12.3377	0.0000	0.5058	[0.45–0.56]
Driver_AgeElderly_over65	−0.2333	0.0150	−15.5378	0.0000	0.7919	[0.77–0.82]
Driver_AgeNot_applicable	−0.0669	0.0377	−1.7745	0.0760	0.9353	[0.87–1.01]
Driver_AgeYoung_Under_25	0.2395	0.0198	12.0934	0.0000	1.2706	[1.22–1.32]
Driver_GenderMale	0.4866	0.0161	30.2331	0.0000	1.6268	[1.58–1.68]
HourEvening	0.5036	0.0152	33.1956	0.0000	1.6546	[1.61–1.70]
HourMorning	−0.1121	0.0163	−6.8716	0.0000	0.8940	[0.87–0.92]
HourNight	1.3542	0.0291	46.5722	0.0000	3.8738	[3.66–4.10]
YearBefore_2020	0.0570	0.0676	0.8436	0.3989	1.0587	[0.93–1.21]
SeasonSpring	−0.2415	0.0191	−12.6676	0.0000	0.7854	[0.76–0.82]
SeasonWinter	0.0054	0.0158	0.3419	0.7324	1.0054	[0.97–1.04]
Pedestrian_GenderMale	0.4784	0.0124	38.6937	0.0000	1.6135	[1.57–1.65]

Table 10. GLMM description for Cluster 2 (youths under 25 years old).

Variable	Estimate	Std. Error	z-Value	Pr (>|z|)	OR	95% CI
(Intercept)	−2.6030	0.6505	−4.0016	0.0001	0.0741	[0.02–0.26]
DayWeekend	0.5037	0.0216	23.3261	0.0000	1.6548	[1.59–1.73]
Road_TypeOne_way_carriageway	−0.3033	0.0615	−4.9360	0.0000	0.7383	[0.65–0.83]
Road_TypeTwo_carriageways	0.3642	0.0645	5.6465	0.0000	1.4393	[1.27–1.63]
Road_TypeTwo_way_carriageway	0.4047	0.0578	6.9960	0.0000	1.4989	[1.34–1.68]
AlignmentIntersection	−0.6523	0.0504	−12.9356	0.0000	0.5208	[0.47–0.57]
AlignmentOther	1.2871	0.0776	16.5953	0.0000	3.6222	[3.11–4.22]
AlignmentRoad_Segment	−0.0193	0.0465	−0.4152	0.6780	0.9809	[0.90–1.07]
Road_Surface_ConditionWet	−0.4512	0.0363	−12.4215	0.0000	0.6369	[0.59–0.68]
Traffic_SignageSignage_Present	−0.1538	0.0310	−4.9614	0.0000	0.8574	[0.81–0.91]
WeatherClear_Weather	−0.0341	0.0339	−1.0060	0.3144	0.9665	[0.90–1.03]
Vehicle_TypeHeavy_vehicle	1.1120	0.0329	33.8468	0.0000	3.0404	[2.85–3.24]
Vehicle_TypeMicro_Mobility	−1.7743	0.1178	−15.0679	0.0000	0.1696	[0.13–0.21]
Vehicle_TypeOther	0.0111	0.0989	0.1123	0.9106	1.0112	[0.83–1.23]
Vehicle_TypePowered_two_wheeler	−0.8924	0.0337	−26.4992	0.0000	0.4097	[0.38–0.44]
Driver_BehaviourRegular_Driving	−0.3920	0.0345	−11.3602	0.0000	0.6757	[0.63–0.72]
Driver_BehaviourSpeed_and_Loss_Control	0.9865	0.0291	33.8597	0.0000	2.6818	[2.53–2.84]
Driver_BehaviourUnspecified_Circumstance	1.1759	0.0818	14.3745	0.0000	3.2411	[2.76–3.80]
Pedestrian_BehaviourUnsafe_Crossing_or_Emergence	0.6759	0.0332	20.3845	0.0000	1.9658	[1.84–2.10]
Pedestrian_BehaviourUnsafe_Position_or_Movement	1.1223	0.0399	28.0948	0.0000	3.0719	[2.84–3.32]
Pedestrian_BehaviourUnspecified_Circumstance	−0.9071	0.0836	−10.8454	0.0000	0.4037	[0.34–0.48]
Driver_AgeElderly_over65	−0.1028	0.0231	−4.4492	0.0000	0.9023	[0.86–0.94]
Driver_AgeNot_applicable	−0.5088	0.0775	−6.5684	0.0000	0.6012	[0.52–0.70]
Driver_AgeYoung_Under_25	0.2237	0.0282	7.9356	0.0000	1.2507	[1.18–1.32]
Driver_GenderMale	0.7850	0.0250	31.3648	0.0000	2.1925	[2.09–2.30]
HourEvening	0.7087	0.0227	31.2885	0.0000	2.0313	[1.94–2.12]
HourMorning	0.2705	0.0256	10.5856	0.0000	1.3106	[1.25–1.38]
HourNight	2.0786	0.0412	50.4233	0.0000	7.9931	[7.37–8.67]
YearBefore_2020	0.8265	0.7253	1.1396	0.2545	2.2853	[0.55–9.47]
SeasonSpring	−0.0399	0.0279	−1.4297	0.1528	0.9608	[0.91–1.01]
SeasonWinter	0.0498	0.0245	2.0331	0.0420	1.0511	[1.00–1.10]
Pedestrian_GenderMale	0.4170	0.0192	21.7739	0.0000	1.5174	[1.46–1.58]

Another objective of this study is to assess the temporal shift in crash severity induced by the pandemic. Looking at the fixed effects, the Year_Before_2020 variable indicates an overall trend in which the pre-pandemic period shows higher baseline odds of fatal outcomes for the elderly (OR = 1.11, statistically significant) and young pedestrians (OR = 2.28). However, the lack of statistical significance of this fixed parameter in the adult and young models suggests that the post-COVID-19 alteration in mobility is not a uniform national trend, but rather a geographically distributed effect, which is far better captured by the area-specific Random Slopes introduced in the mixed structure. Another explanation could be the severe strain on the healthcare system in northern Italy during the initial pandemic waves (affecting the speed of emergency care) and the geographical transit constraints in the Islands (helicopter/ferry delays).

4. Conclusions

Based on the literature review, in recent years the research community's attention towards pedestrian road safety has focused heavily on developing severity models using Machine Learning algorithms. To unveil the black-box nature of these algorithms, result interpretability techniques such as SHAP have been implemented. As highlighted in the literature, the recent trend involves both the use of hybrid methodologies that simultaneously leverage the advantages of Machine Learning and Econometric Models and a return to high-performing, complex econometric models. This paper attempts to position itself between these two trends by adopting a hybrid-stratified approach. The aim is not merely to identify the risk factors associated with pedestrian crash injury severity, but rather to identify specific risk scenarios in Italy during the 2010–2023 period. In this study, supervised and unsupervised algorithms, along with econometric models, were integrated to create a reliable methodology for identifying pedestrian risk scenarios. Specifically, XGBoost was integrated with SHAP, and the resulting SHAP values were used as inputs to a K-Means clustering algorithm. On these clusters, GLMM models have been calibrated.

Through this hybrid-stratified approach, three risk scenarios have been identified, primarily driven by demographic vulnerability. The first risk scenario involves elderly pedestrians (Cluster 0). This cluster has a higher mortality rate than the others (5.23%), confirming the previous literature findings of severe physical frailty in this user group [37]. The other two identified risk scenarios predominantly involve adult (Cluster 1) and young (Cluster 2) pedestrians, with mortality rates of 1.80% and 0.90%, respectively.

Focusing on Cluster 0 (elderly) and the model in Table 8, results show that when elderly pedestrians are involved in a crash, the involvement of heavy vehicles nearly doubles the odds of a fatal outcome (OR = 1.86), as does speeding-related driver behavior (OR = 2.12). Furthermore, impacts are significantly more critical during evening hours (OR = 1.27) and on weekends (OR = 1.10), whereas intersections (OR = 0.54) and one-way streets (OR = 0.49) significantly reduce the risk.

Regarding Cluster 1 (adult pedestrians), this scenario exhibits strong sensitivity to environmental conditions and driving dynamics (Table 9). Nighttime represents the primary hazard, with the risk increasing (OR = 3.87) compared to daytime. Loss of vehicle control due to high speeds has a profound impact on crash severity (OR = 2.99). Unlike the other risk scenarios, strong spatial variability is observed here. Being located in the Islands increases the baseline risk by 55.2% compared to the national average, while in the North-West, the risk decreases by 42.9%. If the crash occurs in an urban area, the odds of a fatal outcome are 82.2% lower than in rural areas.

For Cluster 2, which involves young pedestrians, crash severity is strictly linked to high kinetic energy contexts and poor visibility (Table 10). During nighttime hours, the risk of a fatal crash is over 8 times higher than during the daytime (OR = 7.99), and a significant multiplier is observed for impacts involving heavy vehicles (OR = 3.04). Unlike the elderly, unsafe road occupation by young pedestrians drastically increases the risk of death (OR = 3.07).

Across all scenarios, being a male pedestrian or driver systematically increases the odds of a fatal outcome. From a modeling perspective, the inclusion of the geographic random intercept improved the explanatory power of the models (Conditional R^2 > Marginal R^2), demonstrating that geographic location in Italy acts as an intrinsic risk factor. Geography in Italy exhibits a clear polarization between the North-West and the Islands. A crash occurring in the Islands has pre-pandemic fatal odds substantially higher than the national average (up to +54% for adults). In contrast, in the North-West, the risk drastically decreases (up to −43%). Adults emerged as the category most sensitive to territorial differences. In contrast, geographic variance is less decisive for the elderly, confirming that their physical frailty overrides the protective or penalizing effects of the local context.

Among the innovative aspects introduced in this study, it is worth noting the proposal of an ad hoc hybrid-stratified methodology for identifying and analyzing risk scenarios, which effectively combines the strengths of state-of-the-art models. While the existing literature specifically focuses on highly vulnerable users, such as the elderly, this study also sought to comprehensively understand the risk factors associated with crashes involving adult and young pedestrians. Additionally, an effort was made to investigate two aspects of risk scenarios previously unaddressed in the literature: the localized impact of the COVID-19 pandemic and the structural role of geographic variables. A key strength of the proposed hybrid-stratified methodology is its high transferability. Because of the analytical pipeline, this methodological framework can be readily transferred and applied to analyze road safety databases in other national and international contexts.

Despite these innovations, certain limitations must be acknowledged. The first concerns the stratification and sequential nature of the methodology: a future objective is to overcome this fragmentation and develop a single, integrated model capable of autonomously identifying the various risk scenarios. Another limitation concerns the dataset, highlighting the need to integrate the most recent data (from 2024 onwards) and additional external information through data fusion. Due to the natural rarity of fatal crashes, the models exhibit low overall predictive metrics (such as F1-score). Synthetic oversampling techniques (e.g., SMOTE) have been deliberately avoided to prevent biasing the true underlying Odds Ratios with artificial data, relying instead on class weights. Consequently, these models serve strictly for explanatory inference of risk multipliers (the main novelty of the methodology itself), rather than for deterministic prediction. A significant limitation of this study is the absence of exposure data, such as traffic volumes, pedestrian mobility intensity, or changes in the composition of pedestrian activity; therefore, our findings must be interpreted as temporal associations rather than direct causal effects. As for future developments, beyond addressing the aforementioned limitations, the use of higher-performing predictive models and more refined techniques for handling class imbalances represent viable and promising paths forward.

This approach, leveraging both supervised and unsupervised learning techniques, overcomes the limitations of standard descriptive analyses, providing policymakers with data-driven tools to implement targeted and proactive mitigation strategies, fundamental for reducing pedestrian fatalities in road crashes.

Author Contributions: Conceptualization, G.C., S.N. and M.D.; methodology, G.C., S.N. and M.D.; software, G.C.; validation, G.C.; formal analysis, G.C.; investigation, G.C.; resources, G.C., S.N. and M.D.; data curation, G.C.; writing—original draft preparation, G.C.; writing—review and editing, G.C., S.N. and M.D.; visualization, G.C.; supervision, M.D. and V.N.; project administration, M.D. and V.N.; funding acquisition, G.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets presented in this article are not readily available because of confidentiality restrictions. Requests to access the datasets should be directed to the Italian Institute of Statistics (ISTAT).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study, in the collection, analysis, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

References

1. Directorate-General for Mobility and Transport (European Commission). *Next Steps Towards 'Vision Zero': EU Road Safety Policy Framework 2021 2030*; Publications Office of the European Union: Luxembourg, 2020; ISBN 978-92-76-13219-6.
2. Ministero Delle Infrastrutture e Dei Trasporti. Piano Nazionale Sicurezza Stradale 2030. Available online: https://www.mit.gov.it/node/15908 (accessed on 3 June 2026).
3. Department of Economic and Social Affairs. Transforming Our World: The 2030 Agenda for Sustainable Development. Available online: https://sdgs.un.org/2030agenda (accessed on 31 March 2026).
4. Goswamy, A.; Abdel-Aty, M.; Islam, Z. Factors Affecting Injury Severity at Pedestrian Crossing Locations with Rectangular RAPID Flashing Beacons (RRFB) Using XGBoost and Random Parameters Discrete Outcome Models. *Accid. Anal. Prev.* **2023**, *181*, 106937. [CrossRef] [PubMed]
5. Chen, Z.; Fan, W. Modeling Pedestrian Injury Severity in Pedestrian-Vehicle Crashes in Rural and Urban Areas: Mixed Logit Model Approach. *Transp. Res. Rec.* **2019**, *2673*, 1023–1034. [CrossRef]
6. Alogaili, A.; Mannering, F. Differences between Day and Night Pedestrian-Injury Severities: Accounting for Temporal and Unobserved Effects in Prediction. *Anal. Methods Accid. Res.* **2022**, *33*, 100201. [CrossRef]
7. Zamani, A.; Behnood, A.; Davoodi, S.R. Temporal Stability of Pedestrian Injury Severity in Pedestrian-Vehicle Crashes: New Insights from Random Parameter Logit Model with Heterogeneity in Means and Variances. *Anal. Methods Accid. Res.* **2021**, *32*, 100184. [CrossRef]
8. Nishimoto, T.; Kubota, K.; Ponte, G. A Pedestrian Serious Injury Risk Prediction Method Based on Posted Speed Limit. *Accid. Anal. Prev.* **2019**, *129*, 84–93. [CrossRef] [PubMed]
9. Salon, D.; McIntyre, A. Determinants of Pedestrian and Bicyclist Crash Severity by Party at Fault in San Francisco, CA. *Accid. Anal. Prev.* **2018**, *110*, 149–160. [CrossRef] [PubMed]
10. Sasidharan, L.; Wu, K.-F.; Menendez, M. Exploring the Application of Latent Class Cluster Analysis for Investigating Pedestrian Crash Injury Severities in Switzerland. *Accid. Anal. Prev.* **2015**, *85*, 219–228. [CrossRef] [PubMed]
11. Zhu, C.; Brown, C.T.; Dadashova, B.; Ye, X.; Sohrabi, S.; Potts, I. Investigation on the Driver-Victim Pairs in Pedestrian and Bicyclist Crashes by Latent Class Clustering and Random Forest Algorithm. *Accid. Anal. Prev.* **2023**, *182*, 106964. [CrossRef] [PubMed]
12. Haleem, K.; Alluri, P.; Gan, A. Analyzing Pedestrian Crash Injury Severity at Signalized and Non-Signalized Locations. *Accid. Anal. Prev.* **2015**, *81*, 14–23. [CrossRef] [PubMed]
13. Usman, S.M.; Khattak, A.J. Beyond the Conventional: Exploring Pedestrian Safety on Interstates with Bayesian and Machine Learning Models. *J. Saf. Res.* **2025**, *94*, 141–156. [CrossRef] [PubMed]
14. Guo, M.; Yuan, Z.; Janson, B.; Peng, Y.; Yang, Y.; Wang, W. Older Pedestrian Traffic Crashes Severity Analysis Based on an Emerging Machine Learning XGBoost. *Sustainability* **2021**, *13*, 926. [CrossRef]
15. Li, D.; Ranjitkar, P.; Zhao, Y.; Yi, H.; Rashidi, S. Analyzing Pedestrian Crash Injury Severity under Different Weather Conditions. *Traffic Inj. Prev.* **2017**, *18*, 427–430. [CrossRef] [PubMed]
16. Wang, Z.; Guo, H.; Zhang, C.; Hu, Z.; Zhou, F.; Sun, Z.; Sherony, R.; Bao, S. Investigating Pedestrian Crash Injury Patterns: A Comparative Study of Children and Non-Children. *Accid. Anal. Prev.* **2025**, *222*, 108223. [CrossRef] [PubMed]
17. Zhao, B.; Zuniga-Garcia, N.; Xing, L.; Kockelman, K.M. Predicting Pedestrian Crash Occurrence and Injury Severity in Texas Using Tree-Based Machine Learning Models. *Transp. Plan. Technol.* **2024**, *47*, 1205–1226.
18. Wu, Z.; Misra, A.; Bao, S. Modeling Pedestrian Injury Severity: A Case Study of Using Extreme Gradient Boosting vs Random Forest in Feature Selection. *Transp. Res. Rec.* **2024**, *2678*, 1–11. [CrossRef]
19. Rella Riccardi, M.; Mauriello, F.; Sarkar, S.; Galante, F.; Scarano, A.; Montella, A. Parametric and Non-Parametric Analyses for Pedestrian Crash Severity Prediction in Great Britain. *Sustainability* **2022**, *14*, 3188. [CrossRef]
20. Barua, S.; Chakraborty, R.; Islam, M.; Das, S. A Data-Driven Approach to Child Pedestrian Crash Analysis Using Dimension Reduction, Clustering, and Explainable AI. *Accid. Anal. Prev.* **2025**, *222*, 108229. [CrossRef] [PubMed]
21. Sun, Z.; Wang, D.; Gu, X.; Abdel-Aty, M.; Xing, Y.; Wang, J.; Lu, H.; Chen, Y. A Hybrid Approach of Random Forest and Random Parameters Logit Model of Injury Severity Modeling of Vulnerable Road Users Involved Crashes. *Accid. Anal. Prev.* **2023**, *192*, 107235. [CrossRef] [PubMed]
22. Rehman, Z.U.; Chaozhe, J.; Jamal, A.; Almarhabi, Y.; Wang, C.; Alam, E.; Islam, M.K. Unveiling Pedestrian Injury Severity Risk Factors: A Comparative Analysis of Machine Learning and Deep Learning Methods. *Transp. Eng.* **2025**, *21*, 100352. [CrossRef]

23. Akter, R.; Susilawati, S.; Zubair, H.; Chor, W.T. Analyzing Feature Importance for Older Pedestrian Crash Severity: A Comparative Study of DNN Models, Emphasizing Road and Vehicle Types with SHAP Interpretation. *Multimodal Transp.* **2025**, *4*, 100203. [CrossRef]
24. Khan, M.N.; Das, S.; Liu, J. Predicting Pedestrian-Involved Crash Severity Using Inception-v3 Deep Learning Model. *Accid. Anal. Prev.* **2024**, *197*, 107457. [CrossRef] [PubMed]
25. Kopsacheilis, A.; Politis, I. Exploring Factors Influencing Pedestrian Accident Severity: A Multi-Source Study in the City of Berlin. *Eur. Transp. Res. Rev.* **2024**, *16*, 63. [CrossRef]
26. Antariksa, G.; Tamakloe, R.; Liu, J.; Das, S. Automated and Explainable Artificial Intelligence to Enhance Prediction of Pedestrian Injury Severity. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 5568–5584. [CrossRef]
27. Hossain, A.; Sun, X.; Das, S.; Jafari, M.; Rahman, A. Investigating Pedestrian-Vehicle Crashes on Interstate Highways: Applying Random Parameter Binary Logit Model with Heterogeneity in Means. *Accid. Anal. Prev.* **2024**, *199*, 107503. [CrossRef] [PubMed]
28. Chen, T.; Guestrin, C. Xgboost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.
29. Wu, S.; Yuan, Q.; Yan, Z.; Xu, Q. Analyzing Accident Injury Severity via an Extreme Gradient Boosting (XGBoost) Model. *J. Adv. Transp.* **2021**, *2021*, 3771640. [CrossRef]
30. Al Hasanat, H.A.; Hassanat, A.B.; Alharasees, O.; Tarawneh, A.S.; Altarawneh, G.A.; Alhasanat, L.A.; Begović, M. Interpretable Machine Learning for Imbalanced Pedestrian Injury Severity Prediction in Urban Jordan. *J. King Saud Univ.–Eng. Sci.* **2025**, *37*, 39. [CrossRef]
31. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
32. Chang, I.; Park, H.; Hong, E.; Lee, J.; Kwon, N. Predicting Effects of Built Environment on Fatal Pedestrian Accidents at Location-Specific Level: Application of XGBoost and SHAP. *Accid. Anal. Prev.* **2022**, *166*, 106545. [CrossRef] [PubMed]
33. Usman, S.M.; Khattak, A.J. Applying Artificial Intelligence Techniques to Examine Nighttime Pedestrian Crash Injury Severity at Intersections. *J. Transp. Saf. Secur.* **2025**, *18*, 86–132. [CrossRef]
34. Wisutwattanasak, P.; Se, C.; Sum, S.; Champahom, T.; Ratanavaraha, V.; Jomnonkwao, S. Temporal Instability of Highway Pedestrian Crash Severity: Comparative Analysis of Machine Learning Models. *Transp. Res. Interdiscip. Perspect.* **2026**, *37*, 101935. [CrossRef]
35. Chen, F.; Liu, X.Q.; Yang, J.J.; Liu, X.K.; Ma, J.H.; Chen, J.; Xiao, H.Y. Traffic Accident Severity Prediction Based on an Enhanced MSCPO-XGBoost Hybrid Model. *Sci. Rep.* **2025**, *15*, 25729. [CrossRef] [PubMed]
36. Kim, M.; Kim, D.; Shim, J. The Association between Aggressive Driving Behaviors and Elderly Pedestrian Traffic Accidents: The Application of Explainable Artificial Intelligence (XAI). *Appl. Sci.* **2025**, *15*, 1741. [CrossRef]
37. Wang, H.; Liang, G. Analysis of Injury Severity in Elderly Pedestrian Traffic Accidents Based on XGBoost. *Appl. Sci.* **2025**, *15*, 9909. [CrossRef]
38. Li, Y.; Fan, W.; Song, L.; Liu, S. Combining Emerging Hotspots Analysis with XGBoost for Modeling Pedestrian Injuries in Pedestrian-Vehicle Crashes: A Case Study of North Carolina. *J. Transp. Saf. Secur.* **2023**, *15*, 1203–1225. [CrossRef]
39. Sulle, M.; Mwakalonge, J.; Comert, G.; Siuhi, S.; Gyimah, N.K. Unraveling Pedestrian Fatality Patterns: A Comparative Study with Explainable Ai. *Transp. Res. Interdiscip. Perspect.* **2026**, *36*, 101856. [CrossRef]
40. Ali, Y. Disentangling the Effects of COVID-19 on Pedestrian Injury Severity in Motorcycle-Related Crashes: Evidence from Mixed Logit Modelling. *J. Transp. Health* **2026**, *46*, 102232. [CrossRef]
41. Barbour, N.; Abdel-Aty, M.; Yang, S.; Mannering, F. Pedestrian Injury Severities Resulting from Vehicle/Pedestrian Intersection Crashes: An Assessment of COVID-Contributing Temporal Shifts. *Anal. Methods Accid. Res.* **2024**, *43*, 100334. [CrossRef]
42. Krizsik, N.; Pauer, G. Impact of Covid-19 on Pedestrian Safety. *Transp. Res. Part F Traffic Psychol. Behav.* **2023**, *98*, 216–225. [CrossRef]
43. Tamakloe, R.; Zhang, K.; Kim, I. Temporal Instability of the Determinants of Fatal/Severe Elderly Pedestrian Injury Outcomes in Intersections and Non-Intersections before, during, and after the COVID-19 Pandemic. *Accid. Anal. Prev.* **2024**, *205*, 107676. [CrossRef] [PubMed]
44. ISTAT Crash Data. Available online: https://www.istat.it/tag/incidenti/ (accessed on 20 January 2026).
45. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; MIT Press: Cambridge, MA, USA, 2012; ISBN 0-262-30432-5.
46. Friedman, J.; Hastie, T.; Tibshirani, R. Additive Logistic Regression: A Statistical View of Boosting (with Discussion and a Rejoinder by the Authors). *Ann. Stat.* **2000**, *28*, 337–407. [CrossRef]
47. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4768–4777. [CrossRef]
48. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef] [PubMed]

49. McQueen, J.B. Some Methods of Classification and Analysis of Multivariate Observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*; University of California Press: Berkeley, CA, USA, 1967; pp. 281–297.
50. Lloyd, S. Least Squares Quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137. [CrossRef]
51. McCullagh, P. *Generalized Linear Models*; Routledge: Abingdon-on-Thames, UK, 2019; ISBN 0-203-75373-9.
52. Dean, C.B.; Nielsen, J.D. Generalized Linear Mixed Models: A Review and Some Extensions. *Lifetime Data Anal.* **2007**, *13*, 497–512. [CrossRef] [PubMed]
53. Bates, D.; Mächler, M.; Bolker, B.; Walker, S. Fitting Linear Mixed-Effects Models Using Lme4. *J. Stat. Softw.* **2015**, *67*, 1–48.
54. Bolker, B.M. Linear and Generalized Linear Mixed Models. *Ecol. Stat. Contemp. Theory Appl.* **2015**, *2015*, 309–333. [CrossRef]