Street Art & Urban Creativity | Vol. 11, No. 7, 2025 | ISSN 2183-9956
Street Art & Urban Creativity. Scientific Journal
https://doi.org/10.62161/sauc.v11.5977

SPATIAL PEDESTRIAN SAFETY IN RIYADH SCHOOL ZONES: A DATA-DRIVEN APPROACH

Evaluating Spatial Pedestrian Safety in Riyadh's School Zones Using Multiple Linear Regression and Machine Learning: A Data-Driven Approach

ALA HUSNI ALSOUD [1], AHMAD H. ALOMARI *[2], MOAMAR QRARAH [3], ZAKI ABU AHMAD [4]

[1] Executive Director, Infrastructure Department, Consolidated Consultants Group (CCG), Riyadh, Saudi Arabia. Email: a.alsoud@group-cc.com

[2*] Corresponding Author, Traffic and Transportation Expert, Consolidated Consultants Group (CCG). Email: a.alomari@group-cc.com, and Professor of Civil Engineering, Yarmouk University, Irbid, Jordan. Email: alomarish@yu.edu.jo

[3] Senior Roads & Bridges Engineer, Project Manager, Infrastructure Department, Consolidated Consultants Group (CCG), Riyadh, Saudi Arabia. Email: M.Qrarah@group-cc.com

[4] Senior Traffic and Transportation Engineer, Project Manager, Infrastructure Department, Consolidated Consultants Group (CCG), Riyadh, Saudi Arabia. Email: z.abuahmad@group-cc.com

KEYWORDS

School Zone
Traffic Safety
Pedestrian
Spatial
Regression
Machine Learning
Artificial Intelligence

ABSTRACT

This study comprehensively analyzes pedestrian-runover crash density (the number of crashes per square kilometer of district area) in Riyadh's school zones, employing advanced Artificial Intelligence (AI) techniques, including Multiple Linear Regression (MLR) and Machine Learning (ML), to enhance urban efficiency, quality of life, and resilience. Data were collected from 884 school zones distributed across Riyadh, encompassing diverse infrastructural, socioeconomic, and demographic contexts. The optimized MLR model identified significant predictors, including district road lengths, average crash severity (EPDO), average income, population density, and transit stop availability, which collectively explained approximately 65% of the crash-density variability. The Random Forest ML model further improved predictive accuracy ($R^2 \approx 0.88$), revealing complex, nonlinear interactions among key variables, including traffic volume, speed limits, lane counts, crosswalk availability, and student population. Integrating traditional regression with cutting-edge ML methodologies, this research provides actionable insights for policymakers, urban planners, and engineers, enabling targeted, data-driven interventions to enhance pedestrian safety and promote sustainable, smart urban mobility in Riyadh's school zones.

Received: 03/ 08 / 2025
Accepted: 01 / 09 / 2025

1. Introduction

Road traffic injuries are among the leading causes of death and disability worldwide, affecting millions each year. According to the World Health Organization (WHO, 2023a), road crashes cause 1.19 million fatalities and over 50 million injuries each year. The high number of deaths among vulnerable road users, such as cyclists, motorcyclists, and pedestrians, demonstrates the necessity for specific intervention strategies (WHO, 2023a). Children between 5- and 19-years old face the highest risk among vulnerable road users. The leading cause of death for children in this age group is crashes, and school zones present additional dangers because of high pedestrian activity and limited traffic understanding among young students (UNICEF, 2022).

Research shows that children between 5- and 19-years old face a high risk of traffic injuries, especially when walking to and from school (DiMaggio & Li, 2013). The Safe Routes to School (SRTS) programs in the United States, Canada, and Europe have shown that traffic calming measures, designated pedestrian crossings, and lower speed limits are effective in reducing these risks (Lee et al., 2024). However, many developing countries continue to struggle with implementing safety measures in school zones, which results in a higher number of deaths and injuries (Ehsani et al., 2023). These challenges need to be addressed through predictive modeling, data-driven decision-making, and systematic interventions that are tailored to the local urban environment.

The last ten years have brought substantial progress to Saudi Arabia's road safety measures because the country reduced traffic fatalities from 28 to 18.5 deaths per 100,000 people between 2016 and 2021, due to enhanced enforcement and better speed monitoring and road infrastructure (WHO, 2023b). The combination of high vehicle dependency and inadequate pedestrian infrastructure with traffic congestion in urban school zones creates significant safety risks for pedestrians (Alharbi et al., 2024). The national Vision 2030 initiative works to improve road safety through advanced transportation technology and infrastructure development, yet there are still unknown factors about how particular school zone features influence crash risk.

The immediate danger of accidents and deaths from school zone crashes creates additional public health and urban mobility challenges. Research shows that dangerous school travel environments discourage children from walking or biking to school, which leads to increased obesity rates and reduced physical activity (University of Cambridge, 2019). The safety of school environments represents both a transportation priority and a public health necessity. This study analyzes school zone crash risks in Riyadh, Saudi Arabia, to provide evidence-based findings about school environment traffic crash factors, which support data-driven safety decisions and targeted interventions.

This research evaluates the Pedestrian-Runover Crash Density (number of crashes per square kilometer of district area) in school zones of Riyadh, Saudi Arabia, using multiple linear regression (MLR) and machine learning (ML) methods. The analysis investigates a set of variables, including road geometry, school attributes, surrounding buffer zones, and district-wide characteristics, to determine their effect on pedestrian crash frequency. The study employs MLR and ML as analytical frameworks to discover statistically significant predictors of pedestrian crash density to establish data-driven knowledge about safety risk factors in school zones.

The following sections outline the structure of this paper: *Section 2* reviews existing research on school zone traffic safety from worldwide and regional viewpoints, together with statistical modeling approaches for pedestrian crash risk assessment. *Section 3* explains the study area, data sources, and analytical framework, which uses MLR and ML regression methods to study pedestrian-runover crash density and its related influencing factors. *Section 4* presents the results of exploratory data analysis, correlation assessments, and regression modeling, which show the statistical significance of different predictors, the performance of the optimized MLR model, and the results of the Machine Learning analysis. The final section of this paper (*Section 5*) summarizes

the main findings and provides recommendations for improving pedestrian safety in Riyadh's school zones.

2. Literature Review

Ensuring traffic safety in school zones is a global priority due to the heightened vulnerability of children during school commutes. Extensive research has explored the factors influencing school zone crashes, including roadway design, driver behavior, enforcement policies, and socioeconomic disparities, to understand their impact on pedestrian safety. While studies from developed countries emphasize the effectiveness of structured interventions—such as Safe Routes to School (SRTS) programs, traffic calming measures, and automated enforcement—research from developing nations highlights challenges like weak enforcement, high vehicle dependency, and inadequate pedestrian infrastructure, which contribute to disproportionately higher risks. With the growing application of statistical modeling techniques, particularly MLR, along with spatial analytics and predictive modeling, recent studies have sought to quantify pedestrian crash risks and identify high-risk zones. This literature review synthesizes existing research by examining key predictors of pedestrian-runover crash density, evaluating intervention strategies, and addressing gaps in current methodologies, particularly in the context of school zones in developing urban environments like Riyadh.

2.1. Traffic Safety in School Zones: Global Overview

The safety of traffic zones around schools continues to be a worldwide priority because children face the highest risk as road users when traveling to and from school. Kingham et al. (2011) in New Zealand observed that improved safety measures led to a decrease in road traffic crashes, but school travel times remained dangerous, thus requiring specific safety measures. The combination of weak enforcement together with high traffic congestion and insufficient pedestrian infrastructure in school zones creates severe traffic safety challenges according to research conducted in India and Cameroon (Lordswill et al., 2024; Tetali et al., 2016). Research shows that pedestrian safety measures, including speed enforcement and designated crossing zones, effectively decrease accidents in wealthy nations, yet enforcement and infrastructure limitations hinder progress in lower-income and middle-income countries (Bahrami et al., 2024; Rothman et al., 2017a;). The significant difference between developed and developing regions demonstrates the immediate requirement for universal safety solutions that unite infrastructure development with policy execution and community participation to protect school zones.

2.2. Key Risk Factors Affecting School Zone Safety

The physical design of roadways, along with their intersections, determines how well students will be protected from risk. According to Zhao et al. (2015), research has proven that poor traffic control devices, combined with insufficient crosswalks and fast-moving roadways, create hazardous conditions for pedestrians near schools. Canada and South Korea conducted research which demonstrates that wide roads combined with uncontrolled crossings and complicated intersections lead to higher pedestrian injuries yet raised crosswalks and pedestrian signal controls decrease the risk of injury (Lee et al., 2016; Rothman et al., 2015). The risk of traffic crashes increases in commercial areas with high population density surrounding schools, while mixed-use neighborhoods with strong pedestrian infrastructure result in reduced injury rates (Yu & Zhu, 2016). According to Oh & Kim et al. (2025), school zones represent lower risk areas for pedestrians than commercial and mixed-use destinations that exist after school hours.

School zone crashes primarily stem from the unsafe driving practices of drivers who speed while their attention is divided between driving and performing traffic violations. The research by Flanagan and Morgan (2023) demonstrates that dangerous driving practices, including double parking, sudden stopping, and unsafe drop-offs, lead to traffic congestion, which makes

walking dangerous for pedestrians. Hu et al. (2025) analyzed Chinese school zone accidents during peak morning hours and determined that the highest risk areas existed in upstream and outside lanes when students gathered and during peak times before shifting to middle lanes as dissipation occurred, with lane changes being the most dangerous safety threat. The observational research by Tetali et al. (2016) and Ivan et al. (2019) showed that children from lower to middle-income communities display riskier pedestrian behavior since their neighborhoods lack adequate safety features.

Socioeconomic factors have a substantial impact on the safety measures implemented in school zones. Research by Farid et al. (2024) and Rothman et al. (2017b) shows that American and Canadian schools located in low-income neighborhoods face elevated crash statistics because their infrastructure is substandard, enforcement is weak, and crossing options are limited. The regulation of traffic enforcement differs widely across countries, as some implement automated speed cameras, while others face weak enforcement mechanisms (Eun, 2023). Research conducted in Saudi Arabia and the Philippines demonstrates that school zone safety improves when policy-based interventions combine more substantial penalties with pedestrian infrastructure investments (Alharbi et al., 2024; Regidor et al., 2023).

2.3. School Safety Interventions and Their Effectiveness

The combination of speed humps with raised crosswalks and flashing beacons proves effective in lowering school zone crash incidents. Research indicates that Safe Routes to School (SRTS) programs in the United States decrease pedestrian injury rates through their implementation of traffic calming measures and improved crossing infrastructure (DiMaggio & Li, 2013; Lee et al., 2024). Predictive modeling and machine learning (ML) technologies are now used to evaluate crash risks while automated speed cameras decrease speeding violations in countries with strict enforcement systems (Eun, 2023Zhang et al., 2024).

The supervised ML models of crosswalk-use decisions in dense urban corridors of Dhaka, Bangladesh, demonstrated how infrastructure design and traffic operations create unsafe crossing conditions, which Star Rating for Schools (SR4S) audits can use to determine appropriate countermeasures near schools (Sakib et al., 2024). Another study on Dhaka schoolchildren's street crossing behavior demonstrated that educational programs combined with enforcement measures and traffic speed control can help minimize dangerous street crossing behavior (Basunia et al., 2025).

The International Road Assessment Programme (iRAP) protocols use ML and computer-vision technology to bring risk assessment capabilities to Low- and Middle-Income Countries (LMIC), which enables authorities to conduct school-area risk assessments at scale and select the most important locations for SR4S upgrades to deliver proven safety treatments (iRAP, 2025).

The "AI&Me Safe School Zones" program in Vietnam uses artificial intelligence (AI) and ML to analyze thousands of schools through satellite and Street-View imagery for identifying dangerous frontage roads, which then directs funding to implement site-specific engineering solutions for school zone improvements in limited resource cities (AIP Foundation, 2025).

The essential role of education and awareness programs emerges through research, which demonstrates how student safety training and parental involvement create safer school travel conditions (Eun, 2023; Flanagan & Morgan, 2023).

2.4. Gaps in Existing Research and Future Directions

The research on school zone safety has produced many findings, but scientists still need to understand how urban infrastructure interacts with demographic factors and policy interventions to affect pedestrian crash risks in rapidly developing cities such as Riyadh. Most of the existing research depends on traditional statistical methods with linear assumptions while failing to explore advanced predictive modeling approaches that detect complex nonlinear

relationships. The existing research lacks comprehensive analytical frameworks that integrate traffic flow with road geometry, land use, enforcement measures, and socioeconomic conditions.

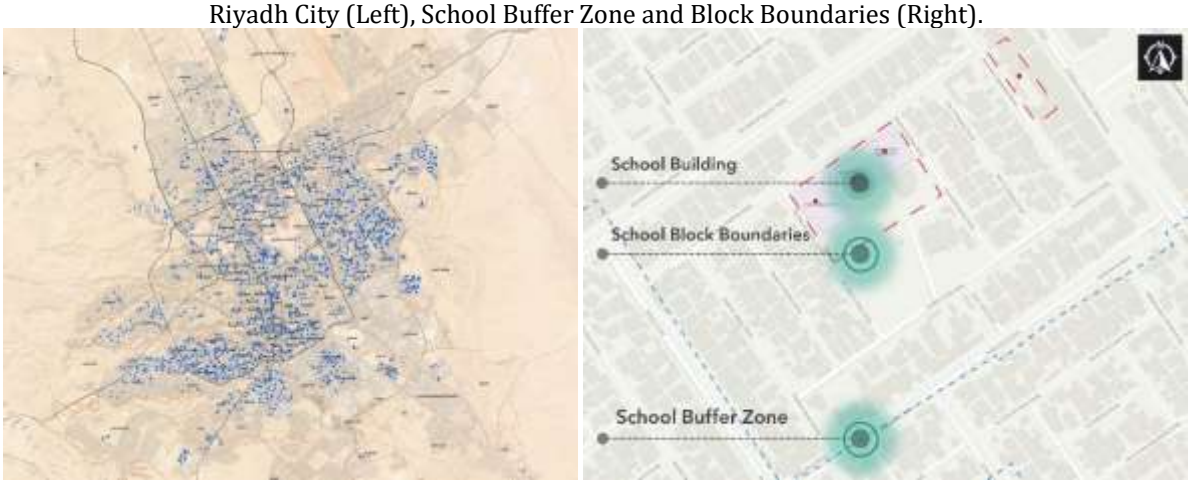
The research fills existing knowledge gaps through the combination of Multiple Linear Regression (MLR) modeling with Machine Learning (ML) regression techniques that use Random Forest modeling to boost prediction accuracy and reveal nonlinear relationships between predictors. The predictive power of ML techniques exceeds MLR, while MLR provides interpretable statistical results that help identify risk factors. The study combines spatial analysis with both MLR and ML techniques to create a framework that produces evidence-based, actionable recommendations for specific pedestrian safety interventions in Riyadh's school zones.

3. Methodology

Riyadh, the capital of Saudi Arabia, serves as the political, economic, and cultural hub of the nation. With a population of 8,591,748 in 2022, it is the largest city in the country, characterized by rapid urbanization, extensive road networks, and a strong reliance on private vehicle transportation (Datasaudi, 2025). Under Vision 2030, the Saudi government has been actively working to improve road safety and sustainable urban mobility, recognizing traffic congestion and pedestrian safety as key challenges. The educational infrastructure of Riyadh consists of 1,430,320 students and 105,942 teachers, who are distributed across its schools (Datasaudi, 2025). The average student enrollment per school reaches 240, while the average number of teachers per school amounts to 17.8, which demonstrates the high density of schools and the need for specific traffic safety measures around these facilities (Datasaudi, 2025). The rising student enrollment and expanding urban areas make it essential to analyze school zone traffic risks in Riyadh to create data-based safety solutions.

This study investigates 884 school zones located throughout all districts and municipalities of Riyadh City, as shown in Figure 1. The chosen schools represent various socioeconomic and infrastructural settings, providing complete spatial coverage of the city. The extensive geographic area enables researchers to study pedestrian-runover crash density while examining roadway features, urban development patterns, and traffic condition variations. This research combines data from multiple school environments to establish city-wide school zone safety insights, which generate evidence-based policy recommendations for addressing both specific risk factors and general urban mobility issues.

Figure 1. Map of Selected School Zones and Their Distribution Across All Districts and Municipalities of Riyadh City (Left), School Buffer Zone and Block Boundaries (Right).



Source: Own elaboration, 2025.

School zones in Riyadh represent high-risk areas for pedestrian safety, mainly due to the significant interaction between vehicles and schoolchildren during peak commute hours. School

buffer zones have been delineated to address these risks, incorporating a combination of early warning zones and speed reduction zones. The analysis considers relevant spatial data within 200 and 300 meters of school boundaries (**Figure 1**) to assess contributing factors such as road geometry, pedestrian infrastructure, and traffic control measures. The database was developed and organized using ArcGIS Pro (Esri, 2025). The dataset was characterized by attributes at four levels:

- **School Level:** Number of classes, students, teachers, and administrators.
- **District Level:** Road lengths, school density, land use (commercial/mixed-use area), pedestrian-runover crash density, average deaths/injuries, average income, population density, and population 6-18 years percentage.
- **School Buffer Zone Level:** Traffic flow, vehicle speed, intersections, and transit stop availability.
- **School Adjacent Block Level:** Number of lanes and average speed.

Table 1 provides a detailed list of the studied variables, categorizing them according to their relevance to school zones, district-level influences, and school characteristics.

Table 1. Summary of Studied Variables.

Category	Variable Name		Description
School Characteristics	1	No. of Classes	The total number of classrooms in the school.
	2	No. of Students	The total student population enrolled at the school.
	3	No. of Teachers	The total number of teachers employed at the school.
	4	No. of Administrators	The total number of school administrative staff.
District Characteristics	5	Road Lengths (Km/District Area)	The total length of roads in the district (km) per District Area (km^2)
	6	School Density (School/District Area)	The number of schools per District Area (km^2)
	7	Commercial/Mixed-Use Land Use Area (%)	The percentage of district land allocated for commercial or mixed-use purposes = (Total Area of Commercial/Mixed-Use / Total Area) x 100 %
	8	Average Deaths/Injuries (EPDO/ km^2)	The average number of traffic-related deaths and injuries in the district (Years: 2018 – 2023) = [12x (no. of deaths) + 5 x (no. of injuries)] / District Area
	9	Average Income (SAR/Capita per Year)	The average income level of residents in the district.
	10	Population Density (Capita / km^2)	The number of residents per unit area in the district = Population (capita) / Area (km^2)
	11	Population 6-18 Years (%)	The percentage of the district's population aged 6–18 years = Population in the Age Group (6-18) Years (Capita) ÷ Total Municipality Population (Capita)
	12	Pedestrian-Runover Crash Density (Crash/ km^2)	The number of pedestrian runover crashes per district area (Years: 2018 – 2023) = Pedestrian Runover Crashes / Area (km^2)
	13	Buffer Average Speed (km/hour)	The average vehicle speed (km/hour) within the school buffer zone.

School Buffer Zone Characteristics	14	Buffer Average Flow (Vehicle/hour)	The average traffic volume per hour within the school buffer zone.
	15	Number of Intersections	The number of intersections within the school buffer zone.
	16	Number of Transit Stops	The number of public transit stops within the school buffer zone.
Adjacent Block Street Characteristics	17	Average Number of Lanes	The average number of lanes in streets adjacent to the school.
	18	Average Speed (km/hour)	The average speed of vehicles on streets adjacent to the school.

Source: Own elaboration, 2025.

This research implements MLR modeling to perform a detailed analysis of Pedestrian-Runover Crash Density (number of crashes per km^2 of district area) in school zones across Riyadh, Saudi Arabia. This research model uses quantitative methods to examine the relationships between roadway characteristics, demographic attributes, and infrastructure elements to determine their effects on pedestrian-related traffic incidents.

The statistical modeling technique MLR enables researchers to establish relationships between dependent variables and multiple independent variables according to Kutner et al. (2005). The technique demonstrates exceptional effectiveness in determining how multiple factors affect outcomes, making it suitable for traffic safety research involving various environmental, infrastructural, and demographic elements that lead to crashes (Washington et al., 2020). MLR provides strong interpretability together with flexibility and continuous and categorical predictor handling capabilities, which enable researchers to assess variable effects on pedestrian-runover crash density through a clear statistical framework (Washington et al., 2020). The statistical inference capabilities of MLR allow researchers to determine which factors are statistically significant and how they affect crash frequency (Montgomery et al., 2021). This study utilizes MLR to analyze relationships between various characteristics and variables with pedestrian crash density because this approach delivers an explainable model which guides policy recommendations and urban planning strategies for improving school zone safety.

Mathematically, an MLR model is expressed as (Montgomery, 2017; Alomari et al., 2016):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \quad (1)$$

Where:

- **Y** is the dependent variable (Pedestrian-Runover Crash Density in this study).
- $\mathbf{X_1, X_2, \ldots, X_k}$ are the independent (predictor) variables, representing various factors influencing pedestrian crashes.
- $\boldsymbol{\beta}$ is the intercept, representing the baseline pedestrian crash density when all predictor variables are zero.
- $\boldsymbol{\beta_1, \beta_2, \ldots, \beta_k}$ are the regression coefficients, which indicate the expected change in **Y** for a one-unit change in the corresponding $\mathbf{X_j}$ variable, holding all other variables constant.
- $\boldsymbol{\varepsilon}$ is the error term, capturing unobserved factors and random variability in the data.

The first step of MLR required Variance Inflation Factor (VIF) analysis to check for multicollinearity between predictors. The analysis eliminated all variables that showed VIF values above 10 to reduce multicollinearity. The evaluation of remaining variables occurred through two criteria, which included statistical significance at $p < 0.05$ and their capacity to improve model interpretability and their influence on Akaike Information Criterion (AIC) and Bayesian

Information Criterion (BIC) values. The model selection process involved removing non-significant or redundant predictors until the model fit no longer improved through additional deletions, resulting in the retention of only independent and meaningful predictors. The optimization process improved the stability and reproducibility and reduced the complexity of the MLR model.

The Multiple Linear Regression (MLR) results received additional support through the implementation of machine learning (ML) regression models, which included k-Nearest Neighbors (KNN), Decision Tree, and Random Forest regressors. The KNN method uses observed values from the k most similar school zones in feature space to predict crash density (James et al., 2021). Decision Trees use feature thresholds to divide data into smaller homogeneous groups, which lead to outcome predictions (Mienye & Jere, 2024).

For ML models, all available predictors were initially considered. The model development process included a 10-fold cross-validation procedure, which helped to achieve strong performance and prevent overfitting during both model development and hyperparameter optimization. The dataset received a random partition that split it into ten equal parts for each iteration, with nine parts dedicated to training and one part used for validation. The experiment ran ten times with performance results from each fold combined to obtain an unbiased prediction accuracy measurement. The cross-validation framework employed grid search to determine the optimal set of hyperparameters (e.g., the number of neighbors k in KNN, maximum tree depth, and the number of estimators in Random Forest).

The Random Forest model served as the main ML model because it provided the best predictive results according to Probst et al. (2020). The Random Forest regressor generated feature importance rankings, which helped identify the most critical variables while maintaining interpretability and achieving good predictive results and efficient computation.

Random Forest creates multiple decision trees that receive training from bootstrapped data samples and random subsets of predictor variables. The ensemble averaging method produces a stable prediction by combining predictions from all trees, which leads to better accuracy and reduced overfitting. The Random Forest prediction ($\hat{Y}$) can be expressed as (Pedregosa et al., 2011):

$$\hat{Y} = \frac{1}{B}\sum_{b=1}^{B} f_b(X) \quad (2)$$

The prediction output from the b^{th} decision tree is denoted by $f_b(X)$ while B represents the total number of trees in the ensemble. Random Forests excel at detecting intricate, nonlinear patterns between features, which proves essential for understanding pedestrian crash risks across different urban settings. The study selected Random Forest as its primary ML method because of its strong predictive power, its ability to provide interpretable results through feature importance measures, and its resistance to noise.

4. Analysis & Results

The analysis of pedestrian-runover crash density in Riyadh's school zones relies on MLR and ML modeling in this section. The analysis starts with Exploratory Data Analysis (EDA) to present essential statistics while showing distribution patterns and detecting relationships between variables. The analysis results guide the selection of variables and transformation methods for predictive modeling. The study implements MLR to establish relationships between roadway characteristics, demographics, school zone features, and traffic conditions that affect pedestrian crash frequency.

The following section evaluates model performance and variable significance while using Variance Inflation Factor (VIF) analysis to address multicollinearity for statistical validity. The optimization process involves selecting variables again while removing redundant elements to enhance both predictive accuracy and efficiency. The comparison between the original

MLR model and its optimized version demonstrates better model robustness together with improved interpretability.

The predictive performance of the MLR model is validated through visualization techniques and statistical metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) to compare actual versus predicted crash densities. The obtained insights enable the identification of dangerous school zones, which supports evidence-based policy development for specific pedestrian safety measures.

4.1. Exploratory Data Analysis (EDA)

The exploratory data analysis (EDA) included complete statistical summaries of data characteristics, distribution analysis, and correlation identification for key variables. The EDA results led to the selection of appropriate variables for feature engineering through the identification of relevant predictors for pedestrian-runover crash density prediction.

This method enabled the selection of important predictors, which improved the accuracy and effectiveness of the resulting models to support proactive decision-making for targeted traffic safety interventions. Table 2 contains summary statistics that describe the main characteristics of the dataset (884 school zones), which provides critical information about the central tendencies, variability, and distribution of variables used in predictive modeling.

Table 2. Summary Statistics.

		Mean	Std	Min	0.25	0.5	0.75	Max
1	School Number of Classes	25.10	16.0769	1.0	13	20	33	99.0
2	School Number of Students	610.0	519.5201	0.0	235	457.5	854.25	2960.0
3	School Number of Teachers	43.97	29.9148	0.0	23	33.5	58	174.0
4	School Number of administrators	15.07	18.2166	0.0	2	6	20	80.0
5	District Road Lengths (Km/District Area)	22.50	7.2965	2.36	19.9	22.92	25.32	56.68
6	District School Density (School/District Area km^2)	7.24	4.3320	0.0666	3.12	7.68	10.59	22.62
7	District Commercial-Mixed-Use Land Use Area (%)	19.21	11.8551	0.0005	8.80	18.56	28.1	84.31
8	District Average Deaths/Injuries (EPDO/ District Area km^2)	21.89	12.5263	0.0	14.655	19.58	29.7	91.12
9	District Average Income (SAR/Capita per Year)	91461.9	24337.81	24000.0	81600	93600	104400	248400.0
10	District Population Density (Capita / District Area km^2)	10719.7	7685.17	31.26	6232.74	9727.7	3025.54	49093.2
11	District Population 6-18 Yrs (%)	17.79	8.0170	1.77	13.04	17.61	21.14	76.57
12	District Pedestrian-Runover Crash Density (Crash/ District Area km^2)	1.81	1.5082	0.0	0.79	1.44	2.29	11.66
13	Buffer Average Speed (km/hour)	43.189	2.3989	40.0	41.50	42.63	44.38	55.17
14	Buffer Average Flow (Vehicle/hour)	1619.09	1329.06	500.0	595.59	1278.79	1867.95	8730.41
15	Buffer Number of Intersections	41.344	15.9936	8.0	32	39	47	138.0
16	Buffer Number of Transit Stops	1.068	1.4324	0.0	0	0	2	7.0

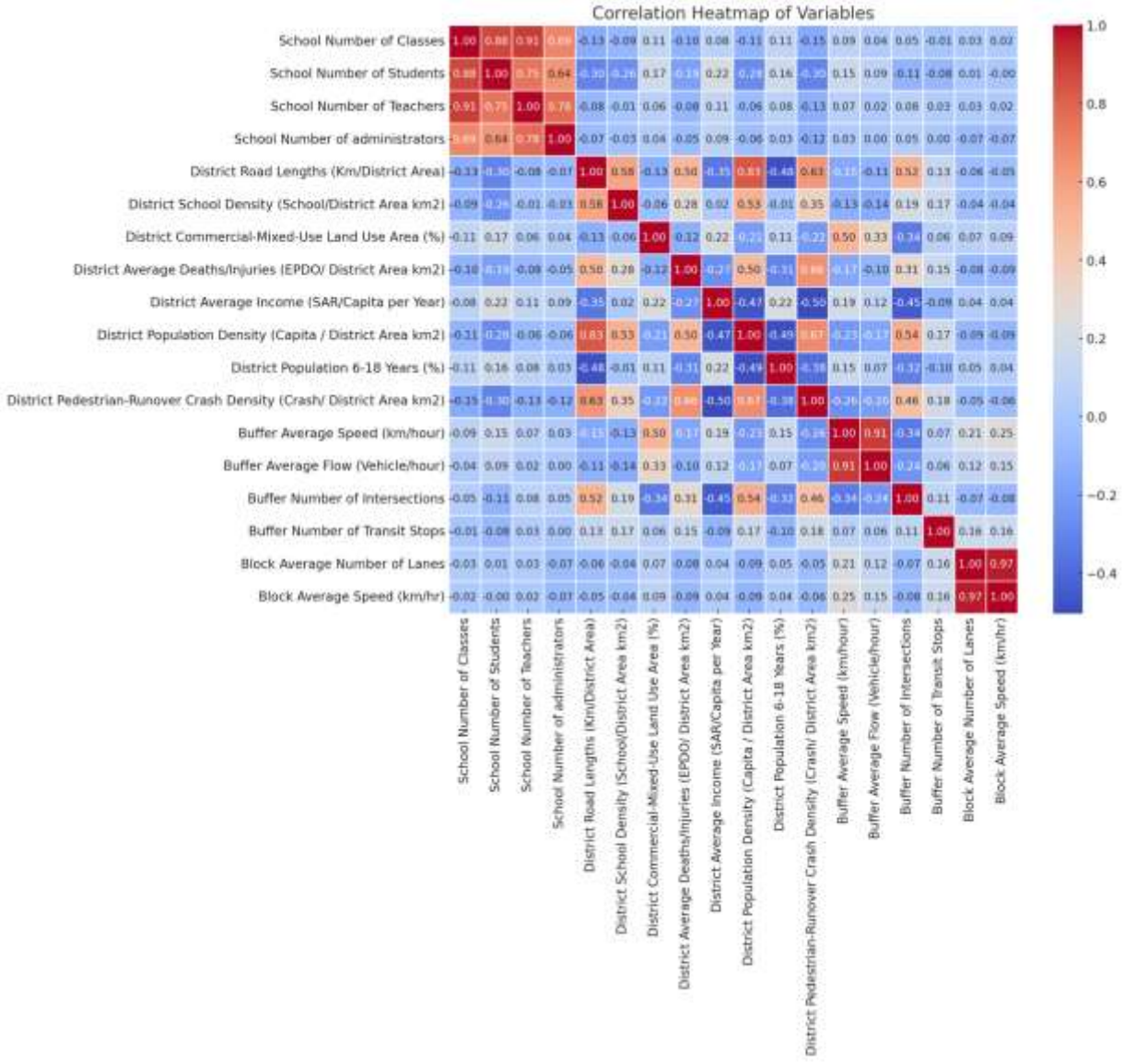
17	Block Average Number of Lanes	2.300	0.6053	2.0	2	2	2.5	8.0
18	Block Average Speed (km/hr)	41.711	3.6713	40.0	40	40	42.42	80.0

Source: Own elaboration, 2025.

The summary statistics table reveals critical insights into the characteristics of school zones and their surrounding environments in Riyadh. Notably, pedestrian-runover crash density varies significantly across districts, ranging from 0 to approximately 11.664 crashes/km^2, with an average of about 1.810 crashes/km^2, indicating substantial variability in crash frequency. Similarly, the severity indicator (EPDO) shows considerable variation, averaging around 21.889 EPDO/km^2 with a maximum reaching 91.120, reflecting high variability in the severity of pedestrian crashes across school zones. Additionally, the data highlights substantial differences in school-related attributes (e.g., student numbers, class sizes, teacher counts) and surrounding traffic conditions (e.g., average flow, intersection density). These observed variabilities underline the importance of tailored, context-specific interventions to effectively enhance pedestrian safety around schools in diverse urban settings.

Correlation analysis is a critical component of data preprocessing that enhances predictive modeling by identifying relationships among variables influencing pedestrian and traffic safety in school zones. The correlation heatmap below (Figure 2) illustrates relationships among key numerical variables in the dataset, providing insights into which factors most significantly influence pedestrian-runover crash density and severity (EPDO) in school zones.

Figure 2. Correlation Matrix Heatmap for All Indicators.



Source: Own elaboration, 2025.

The correlation heatmap provides valuable insights into the relationships between different variables affecting pedestrian-runover crash density in Riyadh's school zones. One of the most significant observations is the strong positive correlation between District Pedestrian-Runover Crash Density and District Average Deaths/Injuries (EPDO per km^2) (**r ≈ 0.66**). This suggests that areas with high pedestrian crash densities also experience more severe traffic incidents. Additionally, pedestrian-runover crash density exhibits a notable positive correlation with District Population Density (**r ≈ 0.67**) and District School Density (**r ≈ 0.35**), indicating that highly populated areas and those with more schools per unit area tend to have a higher frequency of pedestrian crashes. This is expected as increased foot traffic, particularly from students and young pedestrians, naturally increases the likelihood of pedestrian-related crashes. Lastly, Buffer Number of Intersections (**r ≈ 0.46**) and District Road Lenths (**r ≈ 0.63**) is positively correlated with pedestrian crashes, meaning that areas with more intersections and roads tend to witness more pedestrian-runover incidents. This is logical, as intersections/roads increase pedestrian exposure to vehicles, particularly if pedestrian crossings and safety measures are inadequate.

On the other hand, there are some noteworthy negative correlations. For instance, District Average Income has a moderate negative correlation with Pedestrian-Runover Crash Density (r ≈

-0.50), implying that wealthier areas tend to experience fewer pedestrian crashes. This could be attributed to enhanced infrastructure, improved pedestrian facilities, or driving behaviors. Another interesting negative correlation is between Block Average Speed and District School Density ($r \approx -0.26$), suggesting that areas with higher concentrations of schools tend to have lower average speeds, likely due to traffic calming measures or increased congestion.

Furthermore, Block Average Number of Lanes and Average Speed ($r \approx -0.05$ and -0,06, weak correlation) does not seem to have a substantial impact on pedestrian crashes, which might indicate that the number of lanes alone is not a defining factor in pedestrian safety, but rather a combination of lane width, traffic volume, and other pedestrian infrastructure. These insights can inform urban planners and policymakers about key risk factors and guide the implementation of targeted pedestrian safety interventions.

4.2. Multiple Linear Regression (MLR) Model

To predict pedestrian-runover crash density in school zones across Riyadh, an MLR model was developed. The model examines how various factors—such as road network characteristics, school density, land use patterns, and traffic conditions—contribute to the frequency of pedestrian-related crashes. The dependent variable in the analysis is the pedestrian-runover crash density (crashes per km^2 of district area), while the independent variables include school-related attributes, district level features, school buffer zone characteristics, and school adjacent block level factors. This approach provides a baseline model for understanding pedestrian crash risks and serves as a foundation for more advanced predictive models.

The MLR model demonstrates strong explanatory power, with a coefficient of determination (R-squared) value of 0.653, indicating that 65.3% of the variance in pedestrian-runover crash density can be explained by the selected independent variables. The adjusted R-squared value of 0.646 further confirms that the predictors meaningfully contribute to explaining the dependent variable. Additionally, the model yields an F-statistic of 95.68 ($p < 0.001$), confirming that the overall model is statistically significant. Table 3 summarizes the model's performance metrics.

Table 3. Model Performance & Fit.

Metric	R-squared	Adjusted R-squared	F-statistic	Prob (F-statistic)	Log-Likelihood	AIC	BIC
Value	0.653	0.646	95.68	<0.001	-1149.8	2336.0	2422.0

Source: Own elaboration, 2025.

The independent variables in the model have varying degrees of significance in predicting pedestrian-runover crash density. Table 4 presents the coefficient values, statistical significance (p-values), and confidence intervals for all variables.

Table 4. Variables & Their Influence.

	Variable	Coeff.	p-value	Lower 95% CI	Upper 95% CI	Interpretation
1	Constant	1.7867	0.291	-1.533	5.107	The baseline crash density when all predictors are zero.
2	School Number of Classes	0.0115	0.088	-0.002	0.025	More classes may slightly increase pedestrian crashes, though not strongly significant.
3	School Number of Students	-0.0003	**0.041**	-0.001	-1.29e-05	More students are associated with a slightly lower crash density.
4	School Number of Teachers	-0.0022	0.463	-0.008	0.004	No significant impact on pedestrian crashes.

5	School Number of Administrators	-0.0037	0.184	-0.009	0.002	No significant impact on pedestrian crashes.
6	District Road Lengths (Km/District Area)	0.0264	**0.003**	0.009	0.043	More road length is associated with higher pedestrian crash density.
7	District School Density (School/District Area km^2)	0.0170	0.120	-0.004	0.038	More schools per area might increase crash density but is not statistically significant.
8	District Commercial-Mixed-Use Land Use Area (%)	-0.0056	0.088	-0.012	0.001	More commercial land use may slightly reduce crash density.
9	District Average Deaths/Injuries (EPDO per km^2)	0.0467	**<0.001**	0.041	0.052	Higher crash severity correlates with more pedestrian crashes.
10	District Average Income (SAR/Capita per Year)	-1.359e-05	**<0.001**	-1.69e-05	-1.03e-05	Higher-income areas tend to have fewer pedestrian crashes.
11	District Population Density (Capita / District Area km^2)	3.287e-05	**<0.001**	1.65e-05	4.92e-05	Higher population density increases pedestrian crashes.
12	District Population 6-18 Years (%)	-0.0091	0.066	-0.019	0.001	No significant impact on pedestrian crashes.
13	Buffer Average Speed (km/hour)	0.0158	0.678	-0.059	0.091	Speed does not significantly impact pedestrian crash density.
14	Buffer Average Flow (Vehicle/hour)	-8.636e-05	0.154	-0.000	3.25e-05	Traffic flow does not significantly impact pedestrian crash density.
15	Buffer Number of Intersections	0.0014	0.576	-0.004	0.007	More intersections do not significantly impact pedestrian crash density.
16	Buffer Number of Transit Stops	0.0438	0.051	-0.000	0.088	More transit stops slightly increase pedestrian crashes.
17	Block Average Number of Lanes	0.2758	0.168	-0.117	0.669	No significant impact on pedestrian crashes.
18	Block Average Speed (km/hr)	-0.0429	0.199	-0.108	0.023	No significant impact on pedestrian crashes.

Source: Own elaboration, 2025.

The results indicate that several factors significantly influence pedestrian-runover crash density in school zones. The most impactful predictors ($p < 0.05$) include District Average Deaths/Injuries, District Roads Length, District Population Density, and District Average Income, where higher crash severity, longer road networks, and denser populations contribute to increased crash density, while higher-income areas are associated with lower crash risks. Conversely, other variables, such as Buffer Speed, Number of Lanes, and Number of Intersections, do not show statistically significant effects ($p > 0.05$), suggesting that these infrastructure characteristics may not directly determine pedestrian crash risks. School-related variables exhibit mixed results, with School Number of Students being slightly negatively correlated with crash density, possibly due to enhanced safety measures in high-student-density areas, while School Number of Classes shows a weak positive association, indicating that more dispersed classrooms may increase pedestrian exposure. These findings provide valuable insights into the key factors contributing to pedestrian crash risks and highlight the need for targeted interventions to improve school zone safety.

Multicollinearity was assessed using the Variance Inflation Factor (VIF) to determine if independent variables were highly correlated, potentially affecting model reliability. The VIF

results highlight that some predictors exhibit multicollinearity concerns, particularly school-related variables and road network attributes. High VIF values (>10) suggest a strong correlation between these variables, as shown in Table 5.

Table 5. Multicollinearity Considerations and Check.

Variable	VIF
School Number of Classes	**12.95**
School Number of Teachers	8.62
School Number of Students	6.70
Buffer Average Speed	9.12
Buffer Average Flow	7.10
Block Average Number of Lanes	**16.07**
Block Average Speed	**16.43**
All other variables	< 5.0

Source: Own elaboration, 2025.

Since several Variance Inflation Factor (VIF) values exceeded 5, signaling potential multicollinearity among independent variables, all variables with VIF above 10 were systematically removed to enhance model stability and reliability. This refinement ensures that each predictor contributes independently to the regression model, reducing redundancy and improving interpretability. Table 6 presents the Optimized MLR Model Summary, highlighting the refined coefficients and statistical significance of the remaining variables. Additionally, Table 7 provides a comparative analysis of the Original vs. Optimized MLR Model, demonstrating improvements in model efficiency, multicollinearity reduction, and overall predictive robustness.

Table 6. Optimized MLR Model Summary

	Variable	Coeff.	Std. Err.	t	P>|t|	0.025	0.975
1	**Constant**	0.751	1.476	0.509	0.611	-2.146	3.648
2	**School Number of Students**	-0.0001	0	-1.143	0.253	0	8.35E-05
3	**School Number of Teachers**	0.0017	0.002	0.824	0.41	-0.002	0.006
4	**School Number of Administrators**	-0.005	0.003	-1.85	0.065	-0.01	0
5	**District Road Lengths (Km/District Area)**	0.0259	0.009	2.965	**0.003**	0.009	0.043
6	**District School Density (School/District Area km²)**	0.0194	0.011	1.791	0.074	-0.002	0.041
7	**District Commercial-Mixed-Use Land Use Area (%)**	-0.0054	0.003	-1.67	0.095	-0.012	0.001
8	**District Average Deaths/Injuries (EPDO per km²)**	0.0467	0.003	16.276	**<0.001**	0.041	0.052
9	**District Average Income (SAR/Capita per Year)**	-1.43E-05	1.62E-06	-8.843	**<0.001**	-1.75E-05	-1.11E-05
10	**District Population Density (Capita / District Area km²)**	3.30E-05	8.31E-06	3.97	**<0.001**	1.67E-05	4.93E-05
11	**District Population 6-18 Years (%)**	-0.009	0.005	-1.806	0.071	-0.019	0.001
12	**Buffer Average Speed (km/hour)**	0.0147	0.037	0.4	0.689	-0.057	0.087
13	**Buffer Average Flow (Vehicle/hour)**	-8.90E-05	5.93E-05	-1.501	0.134	0	2.74E-05
14	**Buffer Number of Intersections**	0.0016	0.003	0.629	0.53	-0.003	0.007
15	**Buffer Number of Transit Stops**	0.0442	0.022	1.996	**0.046**	0.001	0.088

Source: Own elaboration, 2025.

Table 7. Comparison of Original vs Optimized MLR Model.

Metric	Original Model	Optimized Model
R-squared	0.65255507	0.65061962
Adjusted R-squared	0.64573456	0.64499093
F-statistic	95.6753733	115.590029
Prob (F-statistic)	2.99E-185	2.74E-187
AIC	2335.65255	2334.56324
BIC	2421.77277	2406.3301

Source: Own elaboration, 2025.

The optimized MLR model maintains a strong predictive power while addressing multicollinearity issues. The R-squared value remains nearly unchanged (0.653 → 0.651), indicating that the model continues to explain approximately 65% of the variance in pedestrian-runover crash density. Similarly, the Adjusted R-squared value remains stable (0.646 → 0.645), confirming that removing of high VIF variables did not significantly impact the explanatory capability of the model. This stability suggests that the refined model still captures the key relationships between independent variables and pedestrian crashes while eliminating redundancy.

In terms of model efficiency, the F-statistic increased from 95.68 to 115.59, indicating a more substantial statistical significance and improved efficiency in explaining the dependent variable. Additionally, the Akaike Information Criterion (AIC) decreased slightly (2335.65 → 2334.56), suggesting that the optimized model achieves better performance with fewer predictors, making it more parsimonious. Notably, the p-value for the F-statistic remains highly significant ($p < 0.001$), reinforcing the overall reliability and robustness of the optimized model.

The most significant improvement occurs through the reduction of multicollinearity. The Variance Inflation Factor (VIF) values are now below 5, which confirms that multicollinearity concerns have been successfully mitigated. The improvement strengthens both the stability and reliability of estimated coefficients because each predictor now makes independent contributions to the model without excessive correlation. The model becomes more interpretable and statistically valid after removing School Number of Classes, Block Average Number of Lanes, and Block Average Speed, which results in a more reliable tool for identifying high-risk pedestrian crash zones.

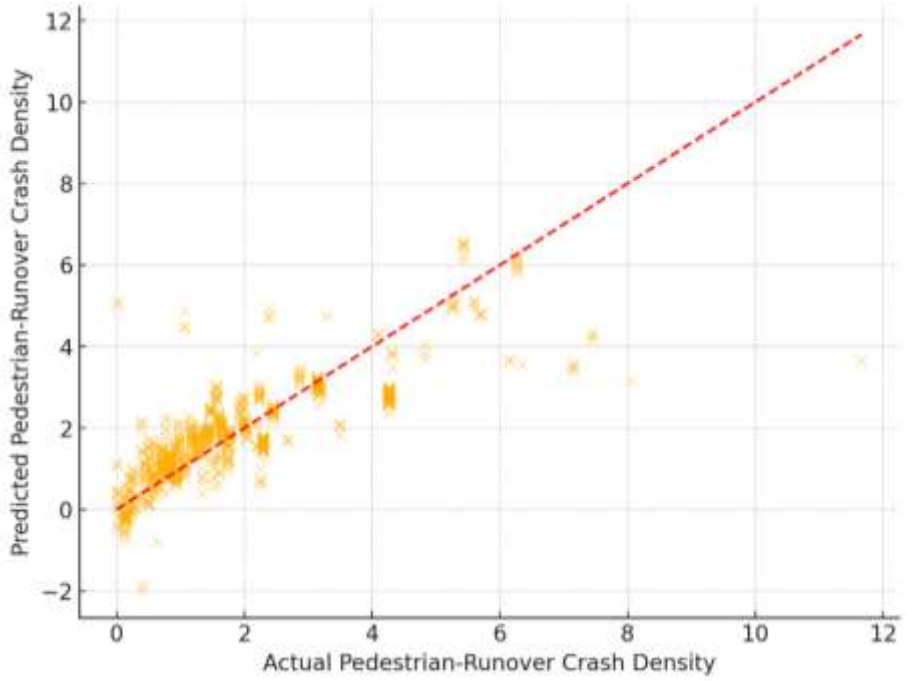
4.2.1. Evaluation and Validation of the Optimized MLR Model

The evaluation of regression model performance requires a comparison between actual and predicted values. The comparison between actual and predicted values helps determine the ability of the optimized MLR model to generalize to real-world pedestrian-runover crash density predictions. The model demonstrates an effective understanding of independent variables and pedestrian crashes when actual values match predicted values. The model demonstrates overfitting, underfitting, or missing predictor variables that affect pedestrian crash density when there are significant deviations between actual and predicted values. The actual vs. predicted value visualization enables both quantitative and intuitive assessment of model reliability and accuracy, which guides future model improvements and real-world deployments.

The scatter plot of Actual vs. Predicted pedestrian-runover crash density (**Figure 3**) visually represents the relationship between observed crash densities and model-generated predictions. Ideally, if the model were perfect, all data points would lie exactly on the red dashed 1:1 reference line, meaning the predicted values would match the actual values perfectly. In this plot, most data points cluster closely around this ideal line, demonstrating that the model makes reasonable and accurate predictions. However, some deviations are present, particularly at higher crash densities, indicating that while the model captures general trends effectively, it may have some limitations

in extreme cases. These insights help assess the consistency of the model's predictions across different districts, reinforcing its usability in prioritizing high-risk pedestrian crash zones.

Figure 3. Actual Vs. Predicted Values (Optimized MLR Model).



Source: Own elaboration, 2025.

The optimized model evaluation metrics further quantify the predictive performance of the model, as shown in Table 8. The Mean Absolute Error (MAE) of 0.59 indicates that, on average, the predicted pedestrian-runover crash density deviates from actual values by approximately 0.59 crashes per km^2, providing an easy-to-interpret measure of error. The Mean Squared Error (MSE) of 0.79 emphasizes the impact of larger errors by squaring them, while the Root Mean Squared Error (RMSE) of 0.89 converts this squared error back to the original scale of the dependent variable, making it a useful benchmark for understanding the expected error magnitude. The R-squared (R^2) value of 0.651 suggests that the model explains approximately 65% of the variability in pedestrian-runover crash density, demonstrating a strong predictive capability while leaving room for further improvements.

Table 8. Optimized Model Evaluation Metrics.

Metric	Value
Mean Absolute Error (MAE)	0.5945
Mean Squared Error (MSE)	0.7938
Root Mean Squared Error (RMSE)	0.8909
R-squared (R^2)	0.6506

Source: Own elaboration, 2025.

The optimized MLR model operates successfully as a deployed system, which has proven its predictive accuracy through visual analysis and performance evaluation. The Actual vs. Predicted scatter plot demonstrates that the model matches real-world pedestrian crash densities because its clusters around the 1:1 reference line, which indicates strong predictive accuracy. The model exhibits minor deviations at higher crash density levels, yet it maintains its ability to detect risky school zones. The performance metrics demonstrate the model's effectiveness through its MAE (0.59), which indicates low average error, and its RMSE (0.89), which provides an understandable error measurement. The R^2 value of 0.651 shows that the model explains

a significant amount of variance in pedestrian crash density, which makes it useful for urban planners and traffic safety analysts.

4.2.2. Discussion

The results of this study align with international research on what makes pedestrian zones safer. This analysis supports Rothman et al. (2015) in Toronto and Ivan et al. (2019) in Bucharest because it shows that pedestrian-run-over crash density depends on population density and roadway infrastructure, including road length and average speed. The research confirms that increased pedestrian crash risks appear in urban areas with dense populations and extensive road networks. The study confirms socioeconomic effects on pedestrian safety, which matches Rothman et al.'s (2017b) findings about how socioeconomic inequalities determine Toronto school-aged child pedestrian injury rates. The steady relationship between income levels and pedestrian safety outcomes demonstrates why socioeconomic factors need to be incorporated into school zone safety improvement plans.

The current study differs from previous research regarding specific roadway characteristic effects. The study found no statistical significance in Riyadh regarding traffic speed and road geometry effects on pedestrian crashes, although Zhao et al. (2015) and Forward et al. (2025) demonstrated their importance in their research. The differences between this research and previous studies likely result from Riyadh-specific urban planning regulations, infrastructure standards, traffic enforcement methods, and local urban planning policies. The studies conducted by Yu and Zhu (2016) and Bahrami et al. (2024) emphasize traffic congestion and intersection complexity as essential predictors, but this study showed that these factors did not affect pedestrian crash density in Riyadh. The findings demonstrate the need to develop locally specific interventions instead of using generic solutions that originate from different areas.

The optimized MLR model established in this research produces strong interpretable results that match previous studies. The model supports policy strategies to improve pedestrian facilities, speed enforcement, and crossing enhancement programs that effectively reduce school area crash risks.

4.3. Machine Learning Analysis

In this section, we propose a regression-based machine learning (ML) model to predict continuous pedestrian crash density in each school zone rather than grouping them into broad risk levels. This way, authorities can assign priorities to the risk areas according to the magnitude of risk and then allocate their safety resources (Shuai & Kwon, 2025). The use of such predictive modeling has been supported by the latest research, for instance, a nationwide study in the U.S. employed a Random Forest regressor to predict crash counts at locations with high accuracy and useful information for preventive road safety interventions (Yamarthi et al., 2025). Our analysis follows this approach by examining various ML regression models to select the best one for estimating school-zone crash density.

Some of the regression algorithms that have been used extensively in transportation safety research include k-Nearest Neighbors (KNN) regressor, Decision Tree, and Random Forest. These models range from simple instance-based prediction (KNN) to single-tree decision rules and ensemble tree-based methods. Random Forest ensemble method was chosen because it has been shown to be effective in handling non-linear relationships and interactions in the data and has been found to produce better predictions than traditional statistical models (Donnell et al., 2020). The available dataset of school zones was used for training and validation of the models, and each zone was characterized by its features (e.g., traffic volume, road attributes, and environmental factors) and the observed pedestrian crash density. Hyperparameter tuning was done for each model (e.g., the number of neighbours k in KNN, maximum tree depth, etc.) using cross-validation to ensure that the models were compared fairly. The performance of the models was evaluated using the coefficient of determination (R^2) and error metrics such as

Root Mean Squared Error on a held-out test set. The model that performed best in the test set – the Random Forest regressor in this case – was chosen for further analysis of its predictions and the underlying factors.

4.3.1 Methods

For the regression task, the dataset was structured such that each entry corresponds to a specific school zone, described by various predictive features and the target variable of pedestrian crash density (e.g., crashes per unit area or some standardized exposure). Key predictor variables included traffic exposure metrics (such as average daily traffic volume), road geometric features (number of lanes, presence of crosswalks, speed limit, etc.), and surrounding environmental characteristics (land use, student population, etc.). All features were normalized or encoded appropriately for use in the ML models. We split the data into training and testing sets to enable an unbiased evaluation of each model's predictive performance.

We trained three ML regressors on the training data by using (1) k-Nearest Neighbors (KNN) to predict crash density from similar school zones in the feature space, (2) Decision Tree to learn a hierarchy of if-then rules for data partitioning, and (3) Random Forest as an ensemble of many decision trees that votes on the prediction. Model hyperparameters were optimized using grid search and cross-validation on the training set. The number of neighbors k in KNN was tuned, and the depth and number of trees in the Random Forest were adjusted to balance bias and variance. We also employed techniques such as early stopping or pruning (for trees) and set aside a portion of training data for validation to prevent overfitting. The performance of each final model was then assessed on the test set using R^2 (which indicates the proportion of variance in crash density explained) and error metrics like Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE).

Table 9 summarizes the performance of the three regression models on the test dataset. The Random Forest regressor achieved the highest accuracy with an R^2 of about 0.88 and the lowest prediction error (RMSE ≈ 0.56). The Decision Tree model showed moderate performance (R^2 ~0.68), while the KNN regressor trailed with R^2 around 0.56. This ranking is consistent with expectations and literature, as ensemble tree methods tend to outperform simpler models for crash prediction tasks (Donnell et al., 2020). Given its superior performance, we focus our subsequent results on the Random Forest model.

Table 9. Test set performance of different regression models for predicting pedestrian crash density in school zones.

Regression Model	R^2 (Test)	RMSE (Test)
k-Nearest Neighbors	0.56	1.01
Decision Tree	0.68	0.85
Random Forest	**0.88**	**0.56**

Source: Own elaboration, 2025.

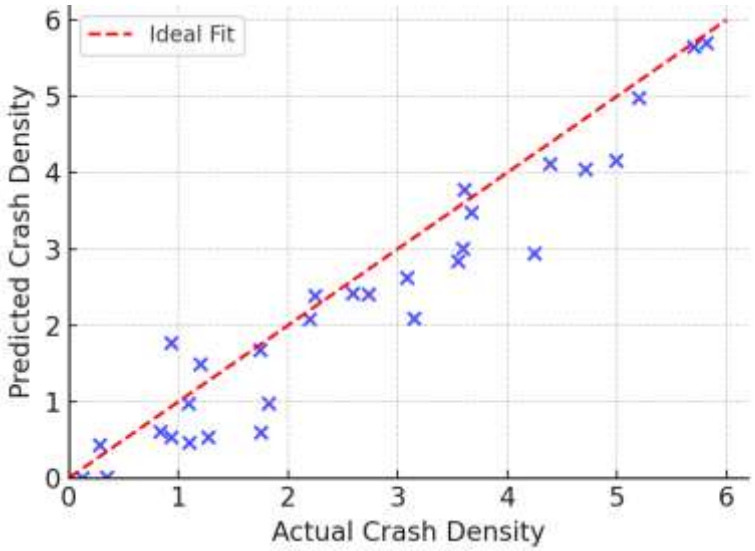
4.3.2 Regression Results

The Random Forest regression model, having the best performance, was used to predict pedestrian crash densities for all school zones in the test set (**Figure 4**). Overall, the model's predictions show a strong agreement with observed values. The coefficient of determination R^2 ≈ 0.88 suggests that roughly 88% of the variability in crash density across school zones is explained by the model – a notably high proportion for crash data, which are often noisy. In practical terms, this level of accuracy means the model can reliably distinguish higher-risk zones (with elevated crash densities) from lower-risk ones in a continuous spectrum, providing a nuanced risk estimate for each location.

Each blue "×" represents a school zone; points lying on the red dashed line indicate a perfect prediction (Predicted = Actual). The clustering of points around the diagonal line reflects the model's strong performance – most zones' predicted crash density is close to the observed value.

A few outliers can be seen where the prediction error is larger (points farther from the line). However, no systematic bias is evident (the errors are relatively symmetrically scattered around the line). This suggests that the Random Forest did not consistently over-predict or under-predict across the range of crash densities but rather captured the trend well while occasional deviations likely correspond to zones with unique, hard-to-predict conditions.

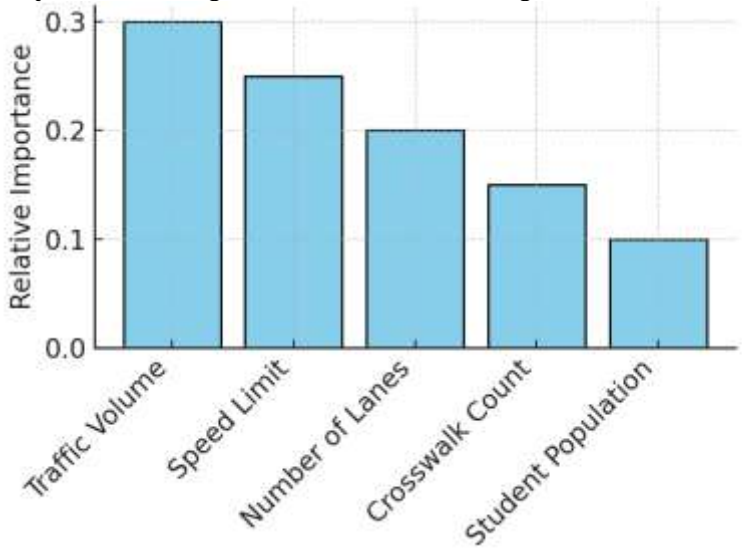
Figure 4. Predicted vs. actual pedestrian crash density for each school zone in the test set, using the Random Forest model.



Source: Own elaboration, 2025.

Beyond overall accuracy, we examined the Random Forest model to identify which features were most influential in predicting crash density. The model's ensemble of trees enables the computation of feature importance values, indicating the extent to which each predictor contributes to reducing prediction error. Figure 5 presents the top-ranking features by their relative importance. The horizontal axis shows the relative importance (contribution to reducing prediction error) of the top five features. Traffic Volume (vehicle flow through the school zone) is the strongest predictor, reflecting the increased crash likelihood with more vehicle exposure (Turner et al., 2017). Speed Limit is the second most important factor – higher speed limits generally correlate with both higher crash risk and severity, as drivers have less reaction time and longer stopping distances (Shuai & Kwon, 2025). The number of Lanes also has a substantial influence; wider roads (with more lanes) encourage higher speeds and create longer crossing distances for pedestrians, elevating risk. The Crosswalk Count in a zone shows moderate importance, suggesting that while crosswalks are safety features, their presence is associated with areas of concentrated pedestrian movement (and thus crashes, if not adequately safeguarded). The student population of the school contributes as well, aligning with the notion that zones with more students (and presumably more walkers) experience higher pedestrian exposure and thus more potential crash incidents.

Figure 5. Feature importance ranking from the Random Forest regression model for crash density prediction.



Source: Own elaboration, 2025.

As expected, measures of traffic exposure and roadway characteristics dominate; traffic volume (vehicle flow through the zone) is the single most important factor, followed by speed limit and number of lanes. These factors intuitively align with safety mechanisms – higher traffic volumes and multi-lane, high-speed roads tend to increase the potential conflict points and severity for pedestrians, leading to higher crash occurrence rates (Turner et al., 2017).

The model also identified noteworthy contributions from features such as the crosswalk count in the zone and the student population (or enrollment size) of the school. A greater number of marked crosswalks could either indicate a response to pedestrian demand (hence inherently higher exposure) or potentially mitigate risk – the model's inclusion of this feature suggests zones with more crossings have distinct crash patterns. Meanwhile, larger student populations may increase pedestrian activity around schools, thereby elevating exposure risk, which is consistent with findings that areas with more students or schools tend to have higher pedestrian crash frequencies (Moradi et. al, 2016).

4.3.3 Discussion

Our regression analysis demonstrates that ML models can effectively predict the continuous crash risk in school zones, providing a granular risk metric for each location. The Random Forest model significantly outperformed the simpler KNN and single decision tree models, reinforcing the benefit of ensemble techniques in capturing the complex, nonlinear relationships inherent in crash data (Donnell et al., 2020). This finding aligns with other research in road safety prediction, where advanced models, such as Random Forest and gradient-boosted trees, have achieved high accuracy in forecasting crash frequencies (Shuai & Kwon, 2025). Notably, the Random Forest's accuracy (R^2 ~0.88) in our study is on par with the performance reported in similar crash prediction efforts (typically R^2 in the 0.8–0.9 range) (Shuai & Kwon, 2025), indicating that our approach is competitive with the state-of-the-art despite focusing on a specific context (school zones).

From a practical standpoint, predicting crash density provides more detailed insights than simple risk classification. Quantifying crash risk allows decision-makers to prioritize zones more effectively —for example, allocating more resources to zones predicted to have significantly higher crash densities. Analysis of feature importance further informs targeted interventions; our findings highlight traffic volume and speed limits, suggesting strategies like traffic calming, speed enforcement, or rerouting traffic (Turner et al., 2017). Road design factors, such as lane count and the presence of crosswalks, also underscore the potential effectiveness of engineering measures

(e.g., road diets and pedestrian refuges). Additionally, the size of the student population indicates the need for enhanced supervision and educational measures at larger schools.

Regarding limitations and future work, the Random Forest model, though highly accurate, remains a "black box," limiting direct interpretability. We addressed this partly through feature importance but recommend applying advanced interpretability methods (e.g., SHAP values, partial dependence plots) to further elucidate nonlinear effects. Incorporating additional variables—such as detailed built-environment features, driver behaviors, or real-time traffic data—may improve predictions. Future research could explore models such as XGBoost or neural networks, although substantial improvements beyond Random Forest are unlikely. Finally, despite strong statistical performance, field validation through on-site assessments is crucial to confirm practical safety improvements.

5. Conclusions

The findings of this study demonstrate that pedestrian-runover crash density in Riyadh's school zones is significantly influenced by district-level demographic and infrastructural characteristics, including population density, road network length, and socioeconomic conditions. The MLR analysis revealed that increased road network density and higher population density correlate significantly with elevated crash frequencies. In contrast, higher average district income is associated with reduced pedestrian crash risks. The ML analysis using the Random Forest model further enhanced predictive accuracy, effectively identifying critical factors such as traffic volume, speed limits, lane counts, crosswalk availability, and student population. By integrating traditional statistical methods and cutting-edge AI techniques, this study provides robust, data-driven insights that support smarter urban interventions, aligning with the goals of enhanced urban efficiency, quality of life, and resiliency.

Despite strong analytical outcomes, this study has limitations. The reliance on historical crash data, potentially subject to underreporting and inaccuracies, is a notable constraint. Additionally, given Riyadh's unique urban and socioeconomic context, careful consideration is required when generalizing these findings to other urban environments.

Future research should leverage advanced AI tools, digital twin simulators, and detailed behavioral datasets—including real-time pedestrian and driver behaviors, enforcement effectiveness, and quality of pedestrian infrastructure—to further refine predictive accuracy. Incorporating spatial-temporal dynamics through advanced geospatial modeling can further optimize predictive performance and intervention effectiveness.

Policymakers, urban planners, and engineers can utilize these AI-enhanced insights to implement targeted and equitable safety interventions. City managers should focus on the high-risk school zones that the Random Forest model has identified for immediate engineering improvements to minimize crashes at a reasonable cost. Municipal GIS dashboards receive model-generated risk mapping data, which enables officials to allocate budgets and schedule maintenance based on data analysis. The implementation of automated enforcement systems, such as speed cameras or radar feedback signs, should be established in school zones with low speed limits that are located in areas with high population density and severe crash records. The deployment of Safe Routes to School campaigns and crossing-guard programs in areas with high student populations requires active collaboration with local education authorities. The system requires periodic model updates with new crash and traffic data to remain effective in evolving urban environments, as part of Riyadh's Vision 2030 smart-city development. Ultimately, this study's approach supports the smart city objective of creating safer, resilient, and data-responsive school environments in Riyadh and comparable urban contexts.

6. Acknowledgements

The present study arises within the framework of the Riyadh Municipality's project, 'Improvement of Traffic Safety around Schools and Mosques in Riyadh (Phase-1)'.

7. Conflict of Interest Statement

The authors confirm there are no conflicts of interest regarding the publication of this article.

8. Research Ethics Statement

This research complies with the Committee on Publication Ethics (COPE) best-practice guidelines for responsible research and publication.

References

AIP Foundation. (2025). Smart Solutions for Safer Schools: Vietnam Leads with AI-Driven Road Safety Big Data Screening across the country. https://www.aip-foundation.org/smart-solutions-for-safer-schools-vietnam-leads-with-ai-driven-road-safety-big-data-screening-across-the-country/

Alharbi, R., Alghamdi, A., Al-Jafar, R., Almuwallad, A., & Chowdhury, S. (2024). *Identifying the key characteristics, trends, and seasonality of pedestrian traffic injury at a major trauma center in Saudi Arabia: a registry-based retrospective cohort study, 2017–2022*. BMC emergency medicine, 24(1), 135. https://doi.org/10.1186/s12873-024-01051-5

Alomari, A. H., Al-Deek, H., Sandt, A., Rogers Jr, J. H., & Hussain, O. (2016). *Regional evaluation of bus rapid transit with and without transit signal priority. Transportation Research Record*, *2554*(1), 46-59. https://doi.org/10.3141/2554-06

Bahrami, V., Lavrenz, S., & Ahmed, M. M. (2024). *Severity Analysis of Pedestrian and Bike Crashes in School Buffer Zones*. Transportation Research Record, 03611981241297682. https://doi.org/10.1177/03611981241297682

Basunia, A., Anchal, T. H., Tasnim, J., Ahmed, N., Mahzabeen, T. Z., & Rifaat, S. M. (2025). Exploring factors influencing jaywalking to promote safe and active travel to school among Dhaka city adolescent students. Journal of Road Safety, 36(2), 47-62. https://doi.org/10.33492/JRS-D-25-1-2470690

Datasaudi. (2025). *A unified platform to present and analyze the latest economic and social data for the Kingdom.* Ministry of Economy & Planning. Kingdom of Saudi Arabia. https://datasaudi.sa/en

DiMaggio, C., & Li, G. (2013). *Effectiveness of a safe routes to school program in preventing school-aged pedestrian injury. Pediatrics, 131*(2), 290-296. https://doi.org/10.1542/peds.2012-2182

Donnell, E. T., Hanks, E., Porter, R. J., Cook, L., Srinivasan, R., Li, F., ... & Eccles, K. A. (2020). *The Development of Crash Modification Factors: Highway Safety Statistical Paper Synthesis*. No. FHWA-HRT-20-069. United States. FHWA, Federal Highway Administration.. https://www.fhwa.dot.gov/publications/research/safety/20069/20069.pdf

Ehsani, J. P., Michael, J. P., & MacKenzie, E. J. (2023). *The future of road safety: challenges and opportunities.* The Milbank Quarterly, 101(Suppl 1), 613. https://doi.org/10.1111/1468-0009.12644

Esri. (2025). *ArcGIS Pro, The world's leading desktop GIS software*. https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview

Eun, S. J. (2023). *Effects of tougher school zone laws on road traffic safety in school zones for children in South Korea.* Journal of Transport & Health, 32, 101687. https://doi.org/10.1016/j.jth.2023.101687

Farid, A., Lin, E., & Pande, A. (2024). *Analysis of School Zone Crash Severities with an Equity Lens: A Random Parameters Modeling Approach.* Transportation Research Record, 03611981241295716. https://doi.org/10.1177/03611981241295716

Flanagan, R. and Morgan, R. (2023). *Improving traffic safety during arrival and dismissal for students at the Quinsigamond School.* Project Report. Worcester Polytechnic Institute, MA, United States. https://digital.wpi.edu/downloads/rr172143p

Forward, S., Henriksson, P., Silvano, A. P., Miyoba, T., Sinkala, S., Mawele, S., & Mwamba, D. (2025). *Increasing traffic safety at schools in Zambia: a before and after study.* Reg. No., VTI: 2022/0296-8.3 https://urn.kb.se/resolve?urn=urn:nbn:se:vti:diva-21493

Hu, X., Deng, H., Liu, H., Zhou, J., Liang, H., Chen, L., & Zhang, L. (2025). *Assessment of the collision risk on the road around schools during morning peak period.* Accident Analysis & Prevention, 210, 107854. https://doi.org/10.1016/j.aap.2024.107854

iRAP. (2025). What is AiRAP. https://irap.org/project/ai-rap/

Ivan, K., Benedek, J., & Ciobanu, S. M. (2019). *School-aged pedestrian–vehicle crash vulnerability.* Sustainability, 11(4), 1214. https://doi.org/10.3390/su11041214

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R*. (2nd ed.). Springer.

Kingham, S., Sabel, C. E., & Bartie, P. (2011). *The impact of the 'school run'on road traffic accidents: A spatio-temporal analysis.* Journal of transport geography, 19(4), 705-711. https://doi.org/10.1016/j.jtrangeo.2010.08.011

Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2005). *Applied Linear Statistical Models.* McGraw-Hill.

Lee, G., Park, Y., Kim, J., & Cho, G. H. (2016). *Association between intersection characteristics and perceived crash risk among school-aged children.* Accident Analysis & Prevention, 97, 111-121. https://doi.org/10.1016/j.aap.2016.09.001

Lee, I. J., Sagar, S., Agarwal, N., Srinivasan, S., & Steiner, R. (2024). *Data-Driven Approach to Develop a Master Plan to Prioritize Schools for the Safe Routes to School Program.* Transportation Research Record, 03611981241250019. https://doi.org/10.1177/03611981241250019

Lordswill, N. T., Jean-Francois, W., Fondzenyuy, S. K., Feudjio Tezong, S. L., Ndonue, A. R., Usami, D. S., & Persia, L. (2024). *Assessment and countermeasures selection for safer roads to schools in the city of Yaoundé: progressive evaluation using surveys and iRAP methodology.* Transportation Research Procedia, 1-8. AIIT 4th International Conference on Transport Infrastructure and Systems (TIS ROMA 2024), 19th - 20th September 2024, Rome Italy. https://iris.uniroma1.it/bitstream/11573/1722401/1/Ndingwan_Assessment-and-countermeasures-selection_2024.pdf

Mienye, I. D., & Jere, N. (2024). *A survey of decision trees: Concepts, algorithms, and applications*. IEEE access. https://doi.org/10.1109/ACCESS.2024.3416838

Montgomery, D. C. (2017). *Design and analysis of experiments.* John wiley & sons.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis.* John Wiley & Sons.

Moradi, A., Soori, H., Kavousi, A., Eshghabadi, F., & Jamshidi, E. (2016). Spatial factors affecting the frequency of pedestrian traffic crashes: A systematic review. Archives of trauma research, 5(4), e30796. https://doi.org/10.5812/atr.30796

Oh, J., & Kim, J. (2025). *Potential risk factors of child pedestrian crashes after-school hours in Seoul, Korea.* Journal of Transport Geography, 123, 104084. https://doi.org/10.1016/j.jtrangeo.2024.104084

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). *Scikit-learn: Machine learning in Python*. the Journal of machine Learning research, 12, 2825-2830.

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2020). *Hyperparameters and tuning strategies for Random Forest*. WIREs Data Mining and Knowledge Discovery, 10(3), e1371. https://doi.org/10.1002/widm.1371

Regidor, J. R. F., Kamid, S. A., Latonero, G. S. D., Abao, N. S. A., & Sigua, R. D. (2023). *Evaluation and Improvement of Road Safety In the Vicinity of Schools.* Philippine Institute of Civil Engineers 2023 National Convention. https://www.researchgate.net/publication/378077600_Evaluation_and_Improvement_of_Road_Safety_in_the_Vicinity_of_Schools

Rothman, L., Buliung, R., Howard, A., Macarthur, C., & Macpherson, A. (2017a). *The school environment and student car drop-off at elementary schools.* Travel Behaviour and Society, 9, 50-57. https://doi.org/10.1016/j.tbs.2017.03.001

Rothman, L., Howard, A., Buliung, R., Macarthur, C., Richmond, S. A., & Macpherson, A. (2017b). *School environments and social risk factors for child pedestrian-motor vehicle collisions: A case-control study.* Accident Analysis & Prevention, 98, 252-258. https://doi.org/10.1016/j.aap.2016.10.017

Rothman, L., Macarthur, C., To, T., Buliung, R., & Howard, A. (2015). *Motor vehicle-pedestrian collisions and walking to school: the role of the built environment.* Pediatrics, 133(5), 776-784. https://doi.org/10.1542/peds.2013-2317

Sakib, N., Paul, T., Ahmed, M. T., Al Momin, K., & Barua, S. (2024). Investigating factors influencing pedestrian crosswalk usage behavior in Dhaka city using supervised machine learning techniques. Multimodal Transportation, 3(1), 100108. https://doi.org/10.1016/j.multra.2023.100108

Shuai, Z., & Kwon, T. J. (2025). *Analyzing Winter Crash Dynamics Using Spatial Analysis and Crash Frequency Prediction Models with SHAP Interpretability.* Future Transportation, 5(1), 17. https://doi.org/10.3390/futuretransp5010017

Tetali, S., Edwards, P., Murthy, G. V. S., & Roberts, I. (2016). *Road traffic injuries to children during the school commute in Hyderabad, India: cross-sectional survey.* Injury prevention, 22(3), 171-175. https://doi.org/10.1136/injuryprev-2015-041854

Turner, S., Sener, I. N., Martin, M. E., Das, S., Hampshire, R. C., Fitzpatrick, K., ... & Wijesundera, R. K. (2017). *Synthesis of methods for estimating pedestrian and bicyclist exposure to risk at areawide levels and on specific transportation facilities*. No. FHWA-SA-17-041. United States. Department of Transportation. Federal Highway Administration. Office of Safety. https://highways.dot.gov/sites/fhwa.dot.gov/files/2022-06/fhwasa17014.pdf

UNICEF. (2022). *Technical guidance for child and adolescent road safety.* New York: United Nations Children's Fund.

University of Cambridge. (2019). *Children who walk to school less likely to be overweight or obese.* ScienceDaily. ScienceDaily, 21 May 2019. www.sciencedaily.com/releases/2019/05/190521101344.htm.

Washington, S., Karlaftis, M. G., Mannering, F., & Anastasopoulos, P. (2020). *Statistical and econometric methods for transportation data analysis.* Chapman and Hall/CRC.

WHO, World Health Organization. (2023a). *Global status report on road safety 2023.* Geneva: WHO Press. https://www.who.int/teams/social-determinants-of-health/safety-and-mobility/global-status-report-on-road-safety-2023

WHO, World Health Organization. (2023b). *Reducing Road Crash Deaths in the Kingdom of Saudi Arabia.* https://www.who.int/news/item/20-06-2023-reducing-road-crash-deaths-in-the-kingdom-of-saudi-arabia

Yamarthi, D., Raman, H., Parvin, S. (2025). *United States Road Accident Prediction using Machine Learning Algorithms.* arXiv:2505.06246v1 28 Apr 2025. https://arxiv.org/pdf/2505.06246v1

Yu, C. Y., & Zhu, X. (2016). *Planning for safe schools: Impacts of school siting and surrounding environments on traffic safety.* Journal of Planning Education and Research, 36(4), 476-486. https://doi.org/10.1177/0739456X15616460

Zhang, K., Tamakloe, R., Cao, M., & Kim, I. (2024). *Exploring fatal/severe pedestrian injury crash frequency at school zone crash hotspots: using interpretable machine learning to assess the micro-level street environment.* Journal of Transport Geography, 121, 104034. https://doi.org/10.1016/j.jtrangeo.2024.104034

Zhao, X., Li, J., Ding, H., Zhang, G., & Rong, J. (2015). *A generic approach for examining the effectiveness of traffic control devices in school zones.* Accident Analysis & Prevention, 82, 134-142. https://doi.org/10.1016/j.aap.2015.05.021