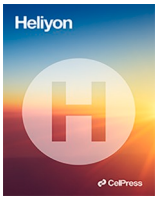
Heliyon 10 (2024) e33346

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Analyzing the effects of streetscape and land use on urban accidents and predicting future accidents by using machine learning algorithms (case study: Mashhad)

Seyed Amir Mohammad Bagheri [a,*], Barat Mojaradi [a], Neda Kamboozia [a], Mohsen Faizi [b]

[a] School of Civil Engineering, Iran University of Science and Technology, Iran
[b] School of Architecture and Environmental Design, Iran University of Science and Technology, Iran

ARTICLE INFO

Keywords:
Land Use
Streetscape
Machine Learning
Urban Accident Modeling
Safety

ABSTRACT

In general, land use and layout of streets can have a significant impact on the behavior of drivers and pedestrians. In particular, streetscape has often been overlooked that recognizing the role of streetscape on street accident in urban areas is important. The aim of this research is to investigate the influence of streetscape and land use on urban accidents that occurred in Mashhad between the years 2017 and 2021. To achieve this objective, the study focused on analyzing accidents in three different urban zones. It also considered the land use types adjacent to both closed and open streets, including residential, commercial, and mixed land uses. The research employed various surveys to gather the necessary data and insights related to the targeted areas. Statistics on accident in three zones show that among the mentioned land uses, commercial areas have experienced the highest number of accidents, with their share being approximately three times that of accidents in residential areas. Additionally, 75 % of all accidents took place in areas with open streetscape, whereas accidents in areas with enclosed view accounted for one third of the number of accidents in open streetscape areas. In this research, analysis and modeling were conducted using machine learning algorithms implemented in the Python programming language. Several models were employed, and the best models were selected based on their performance and accuracy, which include Random Forest Regression (RFR), Multilayer Neural Network Perceptron Regression (MLP) and Extreme Boost Gradient Regression (XGBoost). The accuracy of the machine learning models which successfully predicted future outcomes was as follows: Random Forest Regression (RFR) achieved 85 % accuracy, Extreme Boost Gradient Regression (XGBoost) achieved 81 % accuracy, and finally, Neural Network Multilayer Perceptron Regression (MLP) achieved 75 % accuracy.

1. Introduction

Throughout human existence, transportation has consistently remained one of the most essential and fundamental needs. With the evolving course of human life, the establishment of cities, and the emergence of longer distances, the need for transportation has

* Corresponding author.
E-mail address: bagheri_sa@civileng.iust.ac.ir (S.A.M. Bagheri).

https://doi.org/10.1016/j.heliyon.2024.e33346
Received 13 October 2023; Received in revised form 19 April 2024; Accepted 19 June 2024
Available online 22 June 2024

become increasingly pronounced. As cities expanded in size and population, the significance of transportation intensified. Particularly with the rise of densely populated urban centers and the spatial diversification of cities, this phenomenon has garnered considerable attention within the transportation system. Additionally, transportation has consistently played a crucial role in shaping economic representations of geographical space. Technological advancements have significantly transformed the characteristics of transportation throughout various historical periods [1].

Accidents indeed result in substantial financial and human losses within any society, and without proper planning, these losses can escalate. An inadequate transportation network leads to wastage of time and energy, air pollution, and a high incidence of accidents and traffic-related incidents. These accidents not only cause a loss of capital for countries but also impose substantial costs and damages on societies, particularly affecting families. Each year, tens of millions of people suffer injuries, resulting in significant societal burdens. Furthermore, these accidents tragically lead to the loss of approximately 1.3 million lives annually [2]. The occurrence of accidents is influenced by multiple factors, including road conditions, environmental factors, human conditions, and vehicle characteristics. Analyzing and predicting accidents is a complex task due to the interplay of these various factors.

The main objective of this research is to examine and analyze the impact of streetscape and land use on urban accidents. To achieve this goal, the study employs machine learning algorithms as a tool for analysis and modeling in Python. In this research, the study focuses on collecting statistics and information related to urban accidents in the city of Mashhad. The goal is to examine the influence of various factors within different urban contexts through analysis and modeling techniques. By conducting this analysis, the research aims to provide appropriate safety solutions based on the results obtained. The findings of the research will be valuable for engineers and planners, as they will gain insights into the factors contributing to accidents and identify high-risk areas. Armed with this knowledge, engineers and planners can make informed decisions and implement changes and improvements in safety planning. By addressing these factors and high-risk areas, they can work towards enhancing road safety and reducing the likelihood of accidents in Mashhad.

Indeed, the research focusing on urban accidents adds an additional layer of complexity to the study of this issue. Urban areas often present unique challenges due to higher population densities, mixed land uses, increased traffic volumes, and more complex road networks. Understanding and mitigating accident risks in such environments require specific attention. Furthermore, the impact of land use and streetscape on accidents is an area that has received relatively less attention from researchers in the field of road safety. While factors like road conditions, human behavior, and vehicle characteristics have been extensively studied, the specific influence of land use and streetscape on accident occurrences has often been overlooked. This study represents the first instance of such research being conducted using data from Iran. Analyzing the relationship between land use and streetscape with accidents is crucial for a more comprehensive understanding of road safety. Factors such as the presence of commercial activities, residential density, pedestrian infrastructure, and visibility can significantly impact accident rates. Identifying these relationships and their implications can help enhance urban planning decisions, road design strategies, and targeted interventions to reduce accidents in urban areas. By focusing on these aspects, this research contributes to filling the gap in the studies and provides valuable insights into the interplay between land use, streetscape, and accident occurrences in urban environments.

The research findings suggest that accidents in commercial areas have a higher likelihood of resulting in severe injuries compared to other land use categories. Additionally, the research highlights that accidents occurring in open streetscapes have a greater impact on injury severity compared to accidents in closed streetscapes.

The following section provides a summary of the review of previous literature on the effects of land use and streetscape on urban accidents. Section 3 presents the methodology employed in the study, while section 4 offers a detailed description of the data used. The results and discussions are described in Section 5, and, finally, Section 6 concludes the article.

2. Literature review

Numerous studies have been conducted to predict and model traffic accidents and their outcomes. However, the focus on the impact of land use and streetscape has been relatively limited in these investigations. Understanding the influence of these factors can be crucial in mitigating the risks of potential accidents and reducing the severity of their consequences. By gaining insights into the relationship between land use, streetscape, and accidents, effective measures can be implemented to enhance road safety and prevent accidents.

Land use has a profound impact on various aspects of urban traffic, including the attracted traffic direction, traffic flow ratio, and travel patterns, all of which are essential factors related to public traffic demand. To ensure the efficient operation of urban traffic, it is imperative to engage in rational planning of urban land [3]. Road infrastructure undoubtedly plays a crucial role in determining the risk of pedestrian injury or fatality. However, it's essential to recognize that the type of land use in an urban area is equally significant. High-density commercial centers, for instance, can exert a dominant influence on pedestrian density and the likelihood of pedestrian injury [4]. Mukherjee et al. conducted research in India in 2017, utilizing logistic regression and examining various types of areas to investigate pedestrian safety. Their findings revealed that commercial areas have the most significant impact on pedestrian safety. Specifically, they observed that a 1 % increase in the area dedicated to commercial use results in a 3 % increase in fatal pedestrian accidents. As a result, the researchers suggested implementing safety improvement measures, such as providing guardrails along footpaths to facilitate safe crossings at specific locations [5]. According to the research conducted by Yu in 2015, it was concluded that the density of sidewalks and the presence of commercial land use carry the highest risk of accidents involving pedestrians [6].

According to a 2018 study by Mukoko et al. who investigated bicycle accidents, which were divided into 17 areas such as residential, commercial, industrial, administrative, recreational, and parking, they finally concluded that the commercial type had the greatest effect on accidents [7]. In 2017, Sayed et al. investigated 134 traffic analysis zones (TAZ) in Vancouver, Canada, using

generalized linear model and full Bayesian analyses, and concluded that cyclist accidents decrease with the increase in the density of recreational and residential areas. On the contrary, increasing the density of commercial areas causes an increase in accidents. The result regarding the impact of recreational areas on cyclist accidents is logical, as these areas often offer dedicated, continuous routes for active transportation users, reducing the likelihood of collisions between vulnerable travelers and vehicles. The negative correlation between residential area density and cyclist accidents can be attributed to the implementation of traffic reduction measures aimed at promoting active transportation and limiting motorized traffic in residential neighborhoods. Measures like speed bumps, diverters, and traffic calming measures help slow down vehicles and control their movement in residential areas. On the other hand, the association between commercial areas and cyclist safety can be explained by the higher activities occurring in side streets, which may increase the risk of cyclist collisions with motorized traffic [8].

Streetscape refers to the natural and built texture of the street, encompassing the quality of street design and its visual impact. It is characterized by two types: enclosed and open streetscapes. The streetscape can be examined from various aspects, including urban furniture, flooring, signs, and lighting. For the purpose of this research, the focus is on the trees and buildings surrounding the street [9]. Greater open space on the sides of roads allows drivers to regain control of a stray vehicle before colliding with a fixed object. However, implementing such schemes may lead to increased speeding and riskier driving behavior, ultimately impacting traffic safety adversely. This becomes particularly problematic on urban arterials, which must balance heavy vehicular traffic with the vulnerability of non-vehicular users to high-speed crashes. Moreover, urban arteries often have limited space for additional free areas on the roadside due to dense land uses and existing multimodal infrastructures, such as sidewalks and bike lanes [10–12]. Thus, ensuring traffic safety along urban arteries relies heavily on promoting moderate speeds and discouraging risky behaviors among drivers. Instead of viewing dense and complex urban environments as barriers to safety, it is crucial to explore how they shape streetscapes that inherently exist in urban settings. Ultimately, enhancing safety can be achieved by encouraging responsible driver behavior [13].

An alternative framework that establishes a connection between roadside design and traffic safety, potentially more suitable for urban environments, involves more enclosed environmental designs. Enclosed streetscape includes minimal building setbacks, narrow spaces between buildings, and street tree canopies, aligning with livability goals that promote walkability, vibrant public spaces, and urban green space [14]. In such situations, drivers tend to be more aware of potential risks and exhibit less risky behaviors when their environment is more constrained and offers fewer design adaptations [15]. Consequently, drivers adapt their behavior to respond to environmental conditions, resulting in lower driving speeds, increased capacity to react to unexpected events, and reduced severity of accidents when they occur. This framework is particularly relevant for urban environments, where complex traffic patterns and a variety of road users make speed a major contributing factor to crash severity [16,17].

This study stands apart from previous research because it examines the combined impact of land use and streetscape on accidents. By analyzing accident data from urban roads in Mashhad over a five-year period (2017–2021) using machine learning techniques, the study aims to shed light on the factors influencing accidents. Given the alarming rate of road accidents in Iran, there is a pressing need for effective measures to mitigate the severity of accidents and protect all road users. The specific focus on environmental conditions, including streetscape, in this research is a unique aspect that makes a significant contribution to understanding road safety in Iran, as previous studies have paid less attention to this aspect.

3. Methodology

Implementing machine learning involves developing a model that is trained on specific training data to learn patterns, relationships, and underlying structures. Through this training process, the model gains the ability to analyze and process new or unseen data, enabling it to generate predictions or make informed decisions. The field of machine learning encompasses a wide range of models that have been extensively explored and studied. These models vary in their algorithms, architectures, and learning approaches. The models used in this research are as follows.

3.1. Random forest regression (RFR)

The first algorithm for random decision forests was introduced by Tin Kam Ho in 1995, utilizing the random subspace method. Random Forests, also known as Random Decision Forests, is an ensemble learning technique employed for tasks such as classification and regression. It involves constructing a substantial number of decision trees during the training phase. In classification tasks, the random forest algorithm determines the final class prediction based on the majority decision among the individual trees. Each tree provides its own prediction, and the class with the highest number of votes is selected as the final output. For regression tasks, the random forest algorithm computes the mean or average of the predictions generated by the individual trees. This aggregated result represents the regression output of the random forest model [18,19]. The algorithm was indeed developed by Leo Breiman, who introduced the concept of "Breiman's bag" and introduced random forests in an article. In this article, a method is described for constructing a forest of unrelated trees using a technique such as CART (Classification and Regression Trees), combined with random node optimization and bagging techniques [20].

Indeed, random forest regression is an ensemble learning regression model that is built upon the decision tree algorithm. The guiding technique employed by the random forest regression algorithm is called bagging. This technique involves creating subsets of the data to build individual regression trees, starting with tree 1. The process is then repeated by sampling data with replacement from the original dataset to construct subsequent trees, such as tree 2, tree 3, and so on, until the desired number of trees, denoted as tree n, is created. The iterative nature of this process is illustrated in Fig. 1 [21]. The objective of the random forest algorithm is to analyze and comprehend the relationship between a dependent variable and a set of predictor variables. In the decision tree nodes, variable

selection is carried out on small random subsets of predictor variables, and the best split of predictors is used to divide the node. The individual trees in the forest are then averaged to generate output probabilities and produce a final model that is robust and reliable. The random forest model has key hyperparameters that can be tuned to optimize its performance. These include the number of forest trees (n_estimators), which determines the number of trees in the ensemble. The maximum number of features (max_features) is used to determine the subset of features considered for each split. Additionally, the maximum depth of the tree (max_depth) indicates the depth level of the decision tree, limiting its complexity and controlling overfitting [22].

3.2. Multilayer perceptron (MLP) regressor

Artificial Neural Networks (ANN) are computational systems that draw inspiration from the biological neural networks found in the human brain. These systems are designed to learn and perform tasks without being explicitly programmed with specific rules. Instead, they learn from examples and are capable of carrying out operations similar to those performed by natural nervous systems [23]. Artificial neural networks indeed consist of interconnected units or nodes known as artificial neurons, which are designed to model the behavior of neurons in a biological brain. These artificial neurons are connected through connections, similar to synapses in a biological brain, allowing the transmission of information from one neuron to another. In an artificial neural network, an artificial neuron receives input signals and processes them using a specific activation function. The processed signal is then passed on to other connected neurons, allowing information to flow through the network. This interconnected structure enables the network to learn and make predictions based on the patterns and relationships within the data. Indeed, the connections between artificial neurons in an artificial neural network are referred to as edges. Each edge is associated with a weight that determines the strength of the connection between the connected neurons. During the learning process, these weights are adjusted to optimize the performance of the neural network. The weight increases or decreases signal strength in a connection. Artificial neurons are typically organized into layers within the neural network. The input layer receives the external inputs or features, and subsequent layers, known as hidden layers, perform transformations on these inputs. The final layer, called the output layer, generates the network's final predictions or outputs. In Fig. 2, which depicts the artificial neural network symbolically, the circles represent the neurons present in the three layers: input, output, and hidden layers. The input layer is denoted by X, and the output layer is denoted by Y [24,25].

The multi-layer perceptron (MLP) neural network is indeed a widely used feedforward neural network model. Each layer, except for the input layer, comprises neurons that utilize non-linear activation functions. In MLP, the learning process is typically carried out through a supervised learning technique called backpropagation. This technique enables the network to iteratively update its parameters to minimize the difference between its predictions and the desired outputs. One of the key advantages of MLP is its ability to approximate continuous nonlinear systems. Even with just one hidden layer and an arbitrary finite activation function, the network has the capacity to approximate complex relationships and patterns within the data [26]. Regression models are well-suited for investigating relationships between dependent and independent variables, particularly in cases involving small sample sizes based on least squares fit. However, it is important to note that regression models assume certain conditions, such as the independence of variable components, which may not always align with the actual situation. The multilayer perceptron (MLP) formulation of regression aims to achieve the highest multiple correlation coefficient (R) while minimizing the standard deviation. By optimizing these measures, the MLP model can identify the statistically significant variables in the system [27].

3.3. eXtreme Gradient Boosting (XGBoost) regression

The XGBoost algorithm is a powerful machine learning technique that enhances the predictive performance of weak models, often based on decision trees. It is widely recognized as one of the top supervised learning algorithms. XGBoost follows a flow that includes

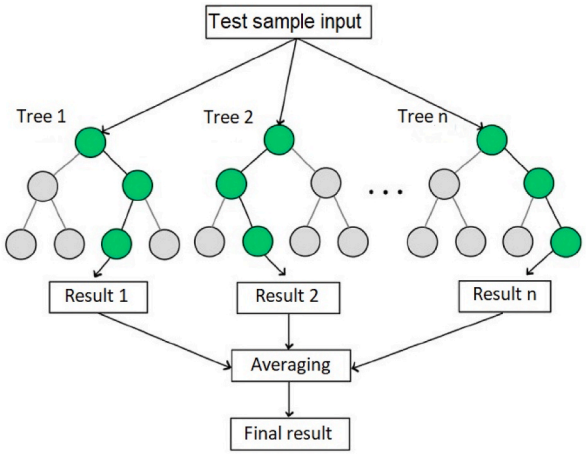


Fig. 1. Random forest regression model.

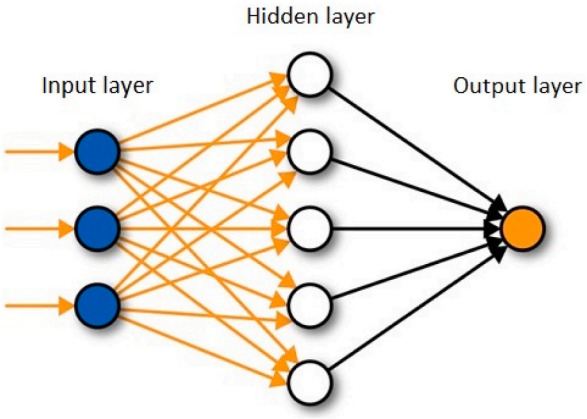


Fig. 2. Artificial neural networks model.

an objective function and base learners, similar to Gradient Boosting. The algorithm constructs the model in a stepwise manner, iteratively improving its performance by minimizing a loss function that measures the discrepancy between predicted and actual values. This algorithm employs ensemble learning, which involves combining multiple base learners to make predictions. A regressor is a model that fits a function using provided features and predicts unknown output values. XGBoost models are highly regarded for their ability to address complex classification and prediction tasks [28,29].

XGBoost is a powerful framework utilized for supervised learning tasks, where the model is trained on multiple features (x_1, x_2, x_3, …) to predict a specific target variable. One of the notable advantages of XGBoost is its scalability and efficient resource utilization compared to other predictive models, making it suitable for various scenarios. The algorithm incorporates parallel and distributed computing, which accelerates the model training process and enables rapid exploration of different model configurations. Additionally, XGBoost incorporates regularization techniques to prevent overfitting and provides robust handling of diverse patterns and data distributions. These features contribute to the versatility and effectiveness of XGBoost in handling complex predictive tasks [30,31].

3.4. Evaluation

Until now, various methods have been used to determine the relative importance and contribution of input variables to the output of models [32]. In this research, sensitivity analysis based on Pearson correlation coefficient was employed to determine the effect of input variables on the dependent variable [33]. To evaluate the models, the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) are utilized, where y_i is the real value and y_i' is the predicted value; $\overline{y}$ is the average of real values, $\overline{y'}$ is the average of predicted values, and N is the number of examined data.

$$R^2 = \frac{\left(\sum_{i=1}^{N}(y_i - \overline{y})(y_i' - \overline{y'})\right)^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2 (y_i' - \overline{y'})^2} \tag{1}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - y_i')^2}{N}} \tag{2}$$

$$MAE = \frac{\sum_{i=1}^{N}\left|(y_i - y_i')\right|}{N} \tag{3}$$

4. Data description

Information serves as the foundation for analysis and modeling processes. Gathering accurate and relevant information is crucial in order to conduct precise and realistic analyses. In the context of accidents, various types of information are essential, including the accident type, road characteristics, environmental factors, and more. The information on annual accidents from 2017 to 2021 and three specific regions within the thirteen regions of Mashhad was obtained through the Transportation Studies Office of the Transport and Traffic Organization of Mashhad Municipality. This office serves as a reliable source of data for transportation-related studies and provides valuable insights into the accidents that occurred in Mashhad during the specified time period. Mashhad is a major city situated in northeastern Iran and serves as the capital of Khorasan Razavi province. Mashhad covers an expansive area of 351 square kilometers, making it the second largest city in Iran, second only to Tehran [34]. In terms of population, Mashhad is also the second most populous city in Iran, with a population of 3,062,242 according to the general population and housing census of 2016. On a global scale, it ranks as the 95th most populous city in the world.

By referencing the data provided by the Transportation Studies Office, the research has access to accurate and official accident

information, which forms the basis for conducting in-depth analysis and modeling. This data enables researchers to examine the trends, patterns, and characteristics of accidents within the specified regions and time frame. A comprehensive dataset was created by combining information on the characteristics of accidents and the injured occupants. The dataset contains 1601 records, representing a total of 7543 accidents that occurred over a period of five years. These accidents took place across 322 different streets.

In the modeling process, it is essential to distinguish between dependent and independent variables. In this study, the dependent variable is represented by the total number of accidents, encompassing both injury and fatality incidents. The data on accidents were presented in a combined form, without disaggregation. The reason for this amalgamation is that in accidents resulting in injuries, there exists the possibility of subsequent fatalities after the injured individuals have been transported to the hospital. Distinguishing between these two groups becomes practically challenging, leading to potential errors and decreased model accuracy. Table 1 provides additional information regarding the dependent variables used in the analysis and modeling process. To facilitate the utilization of these statistical data, it is crucial to label and encode the variables appropriately. This involves converting nominal variables into binary format using one-hot-encoding methodology [35]. This conversion ensures compatibility with regression models and avoids any issues during the modeling process (see Table 2).

The variables in this study consist of both nominal and quantitative variables. The nominal variables include land use, which is categorized into residential, commercial, and other uses (administrative, recreational, and educational). Additionally, streetscape is considered, with two types: open view and closed view, based on the cross-section ratio (cross-section is the ratio of the height adjacent to the width of the street). Furthermore, urban areas are classified into three zones: area 1, area 9, and area 11. Type of urban roads is divided into arterial, collector, and local thoroughfares, each of which is totally 73, 117 and 28 km long respectively. Median type is also included, curb, green space, and, non-physical barriers. Direction of movement of the street is denoted as one-way or two-way. In addition to these nominal variables, other quantitative variables are considered, such as the number of lanes, street length, street width, height of both sides of the street, median width, zone population, year of accident, and the number of accidents.

5. Results and discussion

After pre-processing and data preparation, the data is split into two categories: training and testing datasets, where the share of training data is 80 % and testing is 20 %. The algorithms are implemented in Python using the scikit-learn library. In the training phase of the random forest regression algorithm, each decision tree is grown by creating binary splits on the input features that best isolate the target variable. During this step, two parameters of estimators and random state are determined. The estimators parameter controls the number of decision trees to be included in the random forest set, where n_estimators = 100 means the random forest will consist of 100 decision trees. The random state parameter, random_state = 99, is an arbitrary value used to set the random seed and ensure the reproducibility of the results when there is a need to repeat and compare the results. The built model was evaluated with an R^2 of 84 % on the testing data and 86 % on the training data. The performance diagram of the random forest regression model is shown in Fig. 3. The training stage of the neural network multilayer perceptron regression model involves an iterative optimization process called backpropagation. It commences with random initial weights and biases, and the network is trained on a labeled dataset comprising

Table 1
Summary of variables.

Variables	mean	std	min	max
Median Width	2.60	2.01	0	9
Number of Lane	2.86	0.58	1	4
Length (m)	677.03	280.53	120	1530
Width (m)	9.45	2.02	3	14
Height (m)	9.16	2.83	3	18
Year	1398	1.41	1396	1400
Population	234934.47	68737.92	170282	361703
Land_Use_commercial (1: yes; 0: no)	0.47	0.50	0	1
Land_Use_other (1: yes; 0: no)	0.11	0.31	0	1
Land_Use_residential (1: yes; 0: no)	0.42	0.49	0	1
Zone_1 (1: yes; 0: no)	0.36	0.48	0	1
Zone_9 (1: yes; 0: no)	0.26	0.44	0	1
Zone_11 (1: yes; 0: no)	0.38	0.49	0	1
Streetscape_enclosed (1: yes; 0: no)	0.40	0.49	0	1
Streetscape_open (1: yes; 0: no)	0.60	0.49	0	1
Thoroughfare Type_arterial	0.32	0.47	0	1
Thoroughfare Type_collector	0.55	0.50	0	1
Thoroughfare Type_local	0.13	0.34	0	1
Dominant Adjacent Space_both (1: yes; 0: no)	0.52	0.50	0	1
Dominant Adjacent Space_building (1: yes; 0: no)	0.27	0.44	0	1
Dominant Adjacent Space_tree (1: yes; 0: no)	0.20	0.40	0	1
Median Type_curb (1: yes; 0: no)	0.07	0.26	0	1
Median Type_green space (1: yes; 0: no)	0.70	0.46	0	1
Median Type_none (1: yes; 0: no)	0.22	0.42	0	1
Direction_go (1: yes; 0: no)	0.53	0.50	0	1
Direction_return (1: yes; 0: no)	0.47	0.50	0	1

Table 2
The results of three regression models.

Model	R^2		RMSE		MAE	
	Training	Testing	Training	Testing	Training	Testing
Random Forest Regression (RFR)	0.86	0.84	2.53	2.98	1.43	1.71
Multilayer Perceptron (MLP) Regressor	0.81	0.80	2.95	3.35	1.85	2.14
eXtreme Gradient Boosting (XGBoost) Regression	0.83	0.80	2.76	3.35	1.78	2.16

input features and corresponding target values. Throughout training, the network compares its predicted values with the actual target values and adjusts the weights using gradient descent to minimize the difference between them. This iterative process continues until the network performance is optimized. The model consists of two hidden layers, each containing 50 neurons. Additionally, the Relu activation function (activation = 'relu') is utilized to introduce non-linearity, and the Adam solver (solver = 'adam') is employed for optimization. Other parameters, such as alpha (alpha = 0.001), the maximum number of iterations (max_iter = 500), and the learning rate (learning_rate = 0.001), are also specified. The built model was evaluated with an R^2 of 80 % on the testing data and 81 % on the training data. The performance diagram of the neural network multilayer perceptron regression model is depicted in Fig. 4. The XGBoost regression algorithm is implemented using the xgboost library in Python. This algorithm utilizes a set of decision trees, where each subsequent tree is trained to correct the errors of the previous trees. This iterative process aims to minimize a specified loss function that measures the error between the predicted and actual values. During the training phase, several parameters are set, including the number of estimators or reinforcement rounds (n_estimators = 100), the random seed for repeatability (random_state = 99), the learning rate for each reinforcement iteration (learning_rate = 0.1), and the maximum depth of each tree (max_depth = 6), which determines the complexity of the patterns. The built XGBoost regression model was evaluated with an R^2 of 80 % on the testing data and 83 % on the training data and the performance diagram of the model is depicted in Fig. 5. According to Fig. 6, the importance of each independent variable on the target variable is presented, demonstrating the influence of these variables in predicting the accidents. The figure clearly indicates that land use has a greater impact on accidents compared to streetscape.

In this research, the focus was on investigating three types of land use in urban streets: residential, commercial, and other uses. As depicted in Fig. 7, commercial land uses exhibited the highest number of accidents, surpassing 4800 incidents, followed by residential land uses with a notable difference. The accidents occurring on commercial land uses were nearly three times more than those in residential land uses. The research findings suggest that accidents in commercial areas have a higher likelihood of resulting in injuries compared to residential and other land use categories. This can be attributed to factors such as increased traffic volumes and attracted trips associated with commercial activities.

Additionally, the focus was on investigating the type of urban streetscapes, specifically open and enclosed streetscapes. As depicted in Fig. 8, open streetscapes recorded 5678 accidents, while enclosed streetscapes had 1865 accidents. Notably, 75 % of the accidents occurred in open streetscapes, indicating that for every accident in enclosed streetscapes, three accidents occurred in open streetscapes. Thus, the research highlights that accidents occurring in open streetscapes have a greater impact on injuries compared to accidents in enclosed streetscapes. Indeed, significant impact of the environmental characteristics such as visibility, layout, and design of open streetscapes contribute to inducing effect of streetscapes on road users like higher speeds, increased risk-taking behavior, and

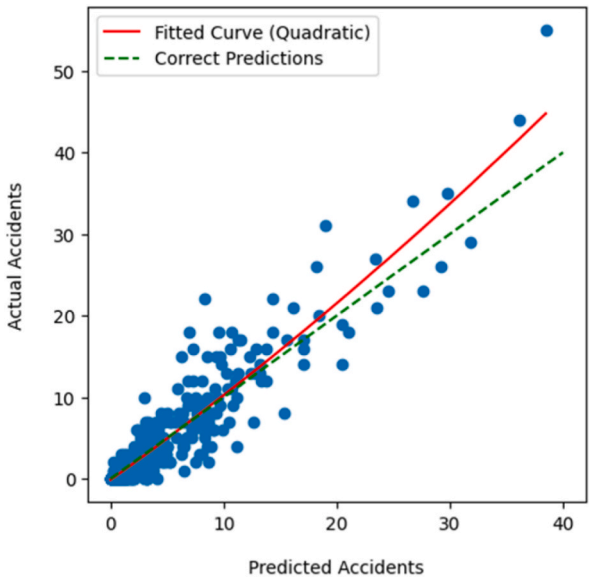


Fig. 3. The performance graph of random forest regression model.

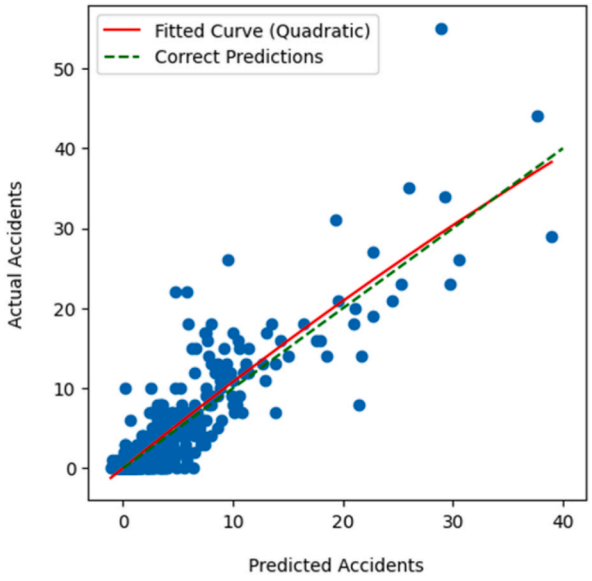


Fig. 4. The performance graph of neural network multilayer perceptron regression model.

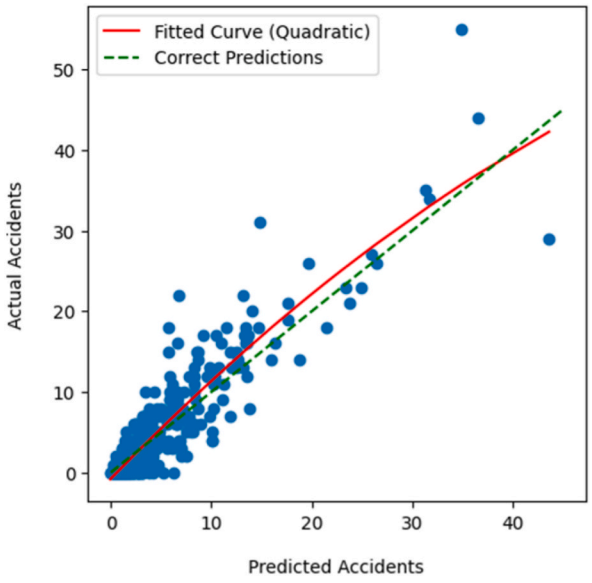


Fig. 5. The performance graph of XGBoost regression model.

reduced reaction times, leading to higher severity of injuries in accidents within these environments.

6. Conclusion

In this research, the focus was on analyzing and modeling the urban accidents in Mashhad, specifically in zones 1, 9, and 11. The study examined the influence of street landscape and land use on these accidents. To accomplish this, machine learning algorithms in Python programming language were employed. Based on the accuracy of the results, the best models for predicting and analyzing the data were identified. These models include Random Forest Regression (RFR), Multilayer Neural Network Perceptron Regression (MLP), and Extreme Gradient Boosting Regression (XGBoost). These models demonstrated superior performance in capturing the relationships between variables and accurately predicting accident outcomes. By utilizing these machine learning models, the research aimed to provide valuable insights into the impact of street landscape and land use on accident occurrences.

Through the analysis and modeling process using the mentioned algorithms, the research found that the Random Forest Regression (RFR) model exhibited the highest accuracy among all the models used in this study. The RFR model achieved a remarkable accuracy of

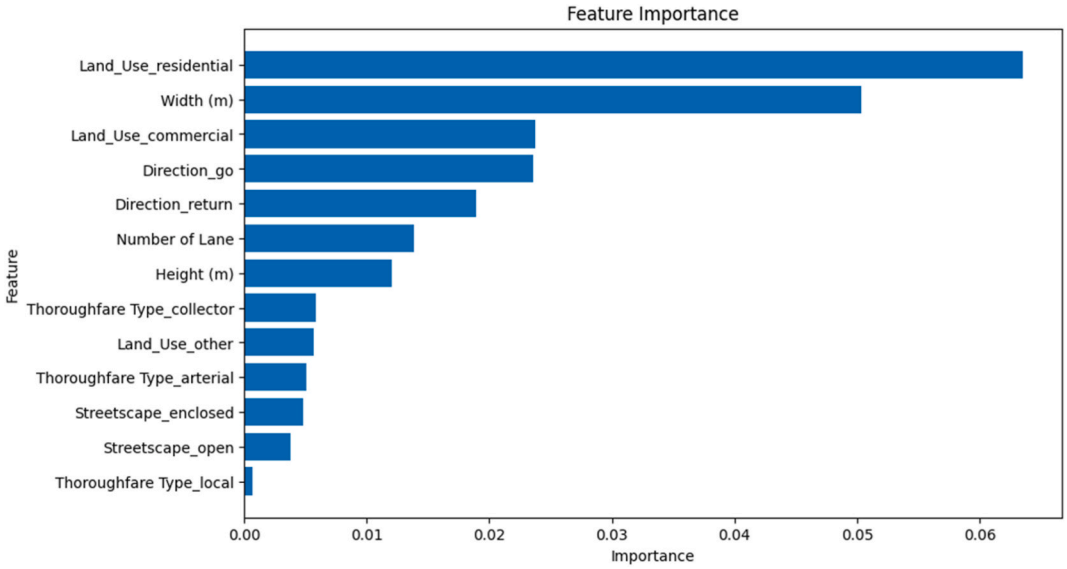


Fig. 6. Feature importance diagram of independent variables.

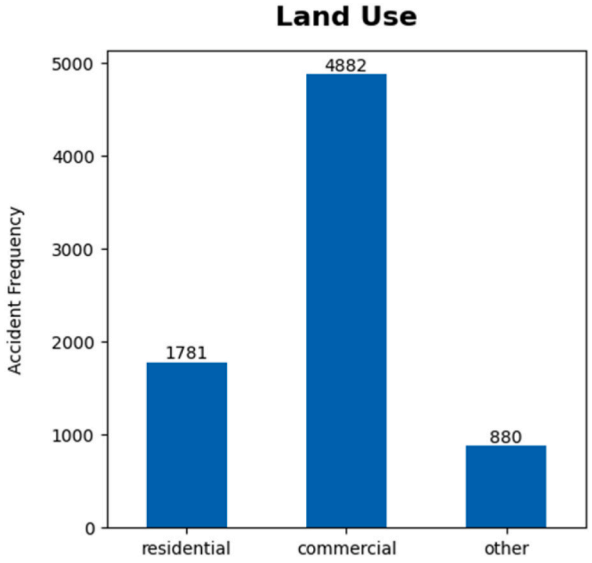


Fig. 7. Land use effect on accidents chart.

96 % on the training data and 85 % on the experimental data. Following RFR, the Extreme Gradient Boosting Regression (XGBoost) model demonstrated an accuracy of 92 % on the training data and 81 % on the test data. Finally, the Neural Network Multilayer Perceptron (MLP) regression model achieved an accuracy of 87 % on the training data and 75 % on the experimental data. These accuracy metrics highlight the performance of each model in accurately predicting the outcomes based on the given data. The high accuracy of the RFR model on both the training and experimental data suggests its robustness and suitability for this research. However, the XGBoost and MLP models also exhibited respectable accuracy rates, showcasing their effectiveness in capturing the underlying patterns and relationships within the data. These results provide valuable insights into the predictive capabilities of these models and their potential usefulness in understanding and addressing urban accidents in the studied areas.

The study aimed to investigate the impact of streetscape and land use on accident occurrences. Three types of land uses, namely residential, commercial, and other, were considered, along with two types of streetscapes were taken into account, namely open and closed. Statistical analysis conducted in this research revealed that the highest number of accidents occurred in commercial areas, accounting for approximately two-thirds of all accidents. Residential areas followed with a quarter of the accidents, while other land uses accounted for 12 % of the accidents. The significantly higher number of accidents in commercial areas compared to residential areas can be attributed to the higher volume of trips and greater population density in these commercial zones. These findings

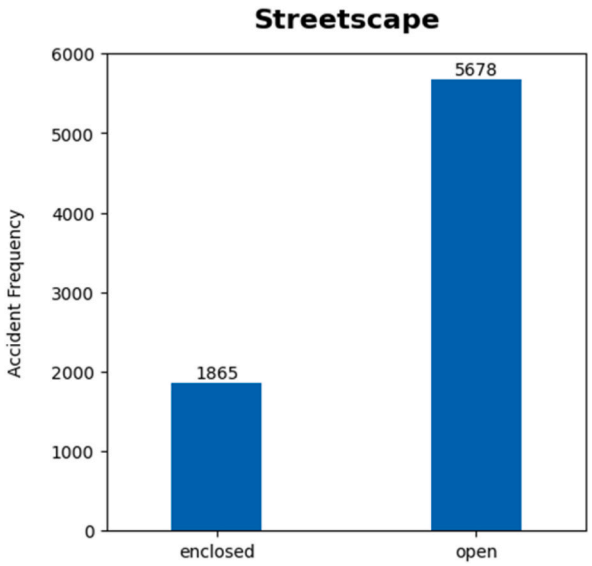


Fig. 8. Streetscape effect on accidents chart.

emphasize the importance of considering the characteristics of land use and its impact on accident occurrences. It suggests that strategies and measures to enhance safety should specifically target commercial areas due to their higher accident rates. The research findings indicate that accidents occurring in an open streetscape account for 75 % of all accidents, which is three times higher than the number of accidents in an enclosed streetscape. This statistic suggests that the characteristics of the environment and its influence on road users have a significant and undeniable impact on accident occurrences. The visibility and design of the open landscape appear to play a crucial role in accident patterns. Moreover, through examining the independent variables and their influence on predicting accidents, the study reveals that the effect of land use is significantly more substantial than the impact of streetscape. This finding highlights the significance of considering land use factors, such as residential or commercial areas, when assessing accident risk and implementing preventive measures. These findings emphasize the importance of incorporating both the characteristics of the physical environment and land use considerations in efforts to enhance road safety. The findings can inform urban planners, policymakers, and relevant stakeholders in developing effective strategies and interventions to mitigate accidents and improve road safety of transportation systems.

Data availability statement

Data will be made available on request.

CRediT authorship contribution statement

Seyed Amir Mohammad Bagheri: Writing – original draft, Writing – review & editing, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Barat Mojaradi:** Writing – review & editing, Validation, Supervision. **Neda Kamboozia:** Writing – review & editing, Data curation, Validation, Supervision. **Mohsen Faizi:** Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Jean-Paul Rodrigue, Claude Comtois, Brain Slack, The Geography of Transport Systems, 2006.
[2] World Health Organization, Global Status Report on Road Safety - Time for Action, 2021.
[3] Tianqi Zhang, Lishan Sun, Liya Yao, Jian Rong, Impact analysis of land use on traffic Congestion using real-time traffic and POI, J. Adv. Transport. 2017 (2017) 1–8, https://doi.org/10.1155/2017/7164790.
[4] G. Cho, D.A. Rodríguez, A.J. Khattak, The role of the built environment in explaining relationships between perceived and actual pedestrian and bicyclist safety, Accid. Anal. Prev. 41 (4) (2009) 692–702.
[5] Dipanjan Mukherjee, Sudeshna Mitra, Impact of road infrastructure land Use and traffic operational characteristics on pedestrian fatality risk: a case study of Kolkata, India, Transportation in Developing Economies 5 (2019), https://doi.org/10.1007/s40890-019-0077-5.
[6] Chia-Yuan Yu, Built environmental designs in promoting pedestrian safety, Sustainability 7 (2015) 9444–9460, https://doi.org/10.3390/su7079444.

[7] K. Kanya, a Mukoko, Srinivas S. Pulugurtha, Examining the Influence of Network, Land Use, and Demographic Characteristics to Estimate the Number of Bicycle Vehicle Crashes on Urban Roads: A Case Study of North Carolina, USA, 2018.
[8] Ahmed Osama, Tarek Sayed, Evaluating the impact of Socioeconomics, land use, built environment, and road facility on cyclist safety, Transport. Res. Rec.: J. Transport. Res. Board 2659 (2017) 33–42, https://doi.org/10.3141/2659-04.
[9] Chester Harvey, Lisa Aultman-Hall, Urban streetscape design and crash severity, Transport. Res. Rec.: J. Transport. Res. Board 2500 (2015) 1–8, https://doi.org/10.3141/2500-01.
[10] P.E. Gårder, The impact of speed and other variables on pedestrian safety in Maine, Accid. Anal. Prev. 36 (4) (2004) 533–542.
[11] Eric Dumbaugh, Wenhao Li, Designing for the safety of pedestrians, cyclists, and motorists in urban environments, J. Am. Plann. Assoc. 77 (2011) 69–88, https://doi.org/10.1080/01944363.2011.536101.
[12] S. Ukkusuri, L.F. Miranda-Moreno, G. Ramadurai, J. Isa-Tavarez, The role of built environment on pedestrian crash frequency, Saf. Sci. 50 (4) (2012) 1141–1151.
[13] E. Dumbaugh, Y. Zhang, The relationship between community design and crashes involving older drivers and pedestrians, J. Plann. Educ. Res. 33 (1) (2013) 83–95.
[14] R. Ewing, S. Handy, Measuring the unmeasurable: urban design qualities related to walkability, J. Urban Des. 14 (1) (2009) 65–84.
[15] E. Dumbaugh, Design of safe urban roadsides: an empirical analysis, in: Transportation Research Record: Journal of the Transportation Research Board, No. 1961, Transportation Research Board of the National Academies, Washington, D.C., 2006, pp. 74–82.
[16] C.S. Hanson, R.B. Noland, C. Brown, The severity of pedestrian crashes: an analysis using google street view imagery, J. Transport Geogr. 33 (2013) 42–53.
[17] A.V. Moudon, L. Lin, J. Jiao, P. Hurvitz, P. Reeves, The risk of pedestrian injury and fatality in collisions with motor vehicles, a social ecological study of state routes and city streets in king county, Washington, Accid. Anal. Prev. 43 (1) (2011) 11–24.
[18] Tin Kam Ho, Random decision forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995, 1995, pp. 278–282. Archived from the original (PDF) on 17 April 2016. (Accessed 5 June 2016).
[19] T.K. Ho, "The random subspace method for constructing decision forests" (PDF), IEEE Trans. Pattern Anal. Mach. Intell. 20 (8) (1998) 832–844, https://doi.org/10.1109/34.709601.S2CID206420153.
[20] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, https://doi.org/10.1023/A:1010933404324. Bibcode:2001MachL..45....5B.
[21] Seungmi Kwak, Jaehwang Kim, Hongsheng Ding, Xuesong Xu, Ruirun Chen, Jingjie Guo, Hengzhi Fu, Machine learning prediction of the mechanical properties of γ-TiAl alloys produced using random forest regression model, J. Mater. Res. Technol. 18 (2022), https://doi.org/10.1016/j.jmrt.2022.02.108.
[22] Yashon Ouma, Ditiro Moalafhi, George Anderson, Boipuso Nkwae, Phillimon Odirile, Bhagabat Parida, Jiaguo Qi, Dam water level prediction using vector AutoRegression, random forest regression and MLP-ANN models based on land-use and climate factors, Sustainability 14 (2022) 14934, https://doi.org/10.3390/su142214934.
[23] G. Carleo, M. Troyer, Solving the quantum many-body problem with artificial neural networks, Science 355 (6325) (2017) 602–606.
[24] R. Radman, S. Alimohammadi, E. Jabari, Comparison of classic models and Artificial Neural Network in prediction of river flow. CONFERENCE OF WATER RESOURCES MANAGEMENT OF IRAN, 2003.
[25] D. Zhang, L. Zeng, K. Cao, M. Wang, S. Peng, Y. Zhang, W. Zhao, All spin artificial neural networks based on compound spintronic synapse and neuron, IEEE transactions on biomedical circuits and systems 10 (4) (2016) 828–836.
[26] Y.O. Ouma, C.O. Okuku, E.N. Njau, Use of artificial neural networks and multiple linear regression model for the prediction of dissolved oxygen in rivers: case study of hydrographic basin of River Nyando, Kenya, Complexity 2020 (2020) 9570789.
[27] S.K. Golfinopoulos, G.B. Arhonditsis, Multiple regression models: a methodology for evaluating trihalomethane concentrations in drinking water from raw water characteristics, Chemosphere 47 (9) (2002) 1007–1018.
[28] Avanija Jangaraj, Gurram Sunitha, Reddy Madhavi, Padmavathi Kora, R. Hitesh, Sai Associate, Prediction of house price using XGBoost regression algorithm, Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12 (2021) 2151–2155.
[29] T. Chen, C. Guestrin, XGBoost, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 16 (2016).
[30] M. Gumus, M.S. Kiran, Crude oil price forecasting using XGBoost, in: 2017 International Conference on Computer Science and Engineering, UBMK), 2017.
[31] Kusal Kankanamge, Yasiru Witharanage, Chanaka Withanage, Malsha Hansini, Damindu Lakmal, Uthayasanker Thayasivam, Taxi Trip Travel Time Prediction with Isolated XGBoost Regression, 2019, pp. 54–59, https://doi.org/10.1109/MERCon.2019.8818915.
[32] S. Singh, S. Jain, A. B′ardossy, Training of artificial neural networks using information-rich data, Hydrology 1 (1) (2014) 40–62.
[[33]] Yashon Ouma, Clinton Okuku, Evalyne Njau, Use of artificial neural networks and multiple linear regression model for the prediction of dissolved oxygen in rivers: case study of hydrographic basin of river nyando, Kenya, Complexity 2020 (2020).
[34] Statistics, Research, and Strategic Studies Office, Retrieved from, Department of Deputy Planning and Development, Mashhad Municipality, 2021, https://planning.mashhad.ir/.
[35] A.Y. Hussein, P. Falcarin, A.T. Sadiq, Enhancement performance of random forest algorithm via one hot encoding for IoT IDS, Period. Eng. Nat. Sci. 9 (3) (2021) 579–591.