“Analyzing Expressway Crash severity : A logistic regression and machine learning approach”

Dipti Tukaram Mane

P. G. Department of Computer Science SNDT Women's University, Mumbai, India

1 Abstract :

The rising number of expressway accidents in India, fueled by increasing traffic density and expanding road networks, underscores the urgency for robust road safety interventions. Traditional statistical methods often fall short due to their rigid assumptions, making machine learning a promising alternative. This study leverages explainable machine learning to analyze factors influencing crash severity on Indian expressways. Two crash severity categories were considered: fatal/severe and non-severe/property damage only. Key contributing factors—road surface condition, road alignment, location, weather conditions, and lighting—were examined. Machine learning models, including Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGB), were applied to a dataset comprising 240 training and 60 testing instances. The models’ performance highlighted the superior predictive capability of machine learning approaches over traditional regression. SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) techniques were employed to interpret the models, revealing critical feature interactions that influence crash severity. These insights enhance model transparency and reliability, aiding stakeholders in formulating data-driven road safety strategies. The findings demonstrate the efficacy of explainable machine learning in addressing crash severity and highlight its potential for improving decision-making in road safety management.

2 Introduction

Globally, road traffic accidents are a leading cause of injury and death, with the World Health Organization (WHO) reporting in 2022 that approximately 1.3 million people lose their lives each year, while another 50 million suffer non-fatal injuries on roads. Interestingly, despite low- and middle-income countries owning only 60% of the world’s vehicles, they account for 93% of global road traffic deaths. In these countries, understanding the factors that contribute to the severity of traffic crashes is essential to reducing fatalities and serious injuries. Various studies have explored these factors from multiple perspectives, including driver behavior, road conditions, and environmental influences. In India, the burden of road traffic crashes, especially on expressways, is significant, and addressing these factors is crucial for improving road safety and mitigating the impact of accidents.

India’s expressway network, spanning over 2,000 km as of 2022, has witnessed a surge in traffic accidents, with a noticeable increase in fatalities and injuries. These expressways, while offering efficient connectivity, are also prone to high-speed collisions, often caused by factors such as driver negligence, fatigue, and unsafe overtaking maneuvers. Environmental variables like poor weather, inadequate road surfaces, and lighting conditions further exacerbate the risks. This research aims to analyze the factors affecting the severity of crashes on Indian expressways using machine learning techniques. Identifying and understanding these factors through advanced models can help inform targeted safety measures and reduce the overall risk of severe traffic accidents on expressways.

3. Literature Review

Research on traffic fatalities and crash severity has been widely explored, with significant differences between findings from Western countries and those from India due to variations in vehicle types and traffic conditions.

Clarke et al. (2010) conducted a study on 1,185 fatal accidents across 10 UK police jurisdictions from 1994 to 2005. Their research identified excessive

speeding and drunk driving as primary causes of fatalities. Other contributors included young drivers, night-time driving, and failure to wear seat belts. These findings are consistent with global trends but may not directly apply to India due to differing traffic behavior and infrastructure conditions.

In India, Mohan and Bawa (1985) explored fatal crash patterns in Delhi during the 1980s. Their study highlighted that fatalities in India were distinct from those in industrialized countries, with two-wheeler riders and bus commuters accounting for 80% of fatalities, while motor-vehicle occupants represented a minority. This contrast indicates the unique nature of road traffic safety issues in India, particularly related to the high proportion of two-wheeler use.

Jain et al. (2009) analyzed two-wheeler accidents in Delhi over the period 2000–2004. Their study found that 77% of victims were between the ages of 18 and 44, with males accounting for 83% of accidents. The study also noted that most accidents occurred at specific times of the day, providing insights into temporal patterns in traffic-related injuries. This study underscores the importance of age, gender, and time in understanding the dynamics of road accidents in India.

The World Health Organization (WHO) also highlights road conditions as a significant factor contributing to crashes, with poor road design, layout defects, and lack of maintenance being major contributors. Many road agencies have proposed safety evaluation criteria considering roadway conditions, traffic data, and environmental factors. These factors play a critical role in identifying high-risk areas and formulating road safety strategies.

Traffic crash prediction models have evolved from simple linear regression (LR) models to more advanced machine learning techniques. While LR remains a popular choice for its simplicity and low data requirements, machine learning models, such as Random Forest, Decision Trees, and Gradient Boosting, offer superior predictive capabilities. These models can analyze various factors, including location, weather, road characteristics, and time of day, helping identify accident-prone areas and enabling targeted safety interventions. Machine learning algorithms have also been applied to manage traffic congestion and prevent accidents by analyzing historical traffic patterns and driver behavior, using data from telematics, sensors, or video feeds. Furthermore, these models can assess road infrastructure quality, including potholes, damaged signs, and faulty traffic lights, which contribute to traffic accidents.

This research builds on previous studies by incorporating machine learning techniques to predict the severity of expressway accidents in India, specifically focusing on factors such as road conditions, weather, and lighting, which have been identified as key contributors to crash severity in both global and local contexts.

---

4 Methodology

4.1 Study Area & Data Collection

In this study, the Expressway Crash Severity Prediction Model focuses on accident data collected from various expressways in India. The dataset includes accident information from major highways and expressways, which are key components of the Indian road network. For the analysis, we have selected the Mumbai-Pune Expressway (E34), one of the most significant highways in the region, connecting Mumbai and Pune. This expressway is a six-lane highway with a total length of 94.5 kilometers and a designed speed of 120 km/h, with a speed limit of 100 km/h. The expressway spans through a mix of urban and rural areas and is a crucial corridor for both passenger and freight transport.

The majority of vehicles using the expressway are passenger cars, which constitute over 75% of the traffic, followed by trucks and motorcycles. The road is designed to accommodate high-speed traffic and features tunnels, bridges, and multiple interchanges. The expressway also passes through various terrains, including flat areas, hills, and regions with sharp curves.

The accident data used in this study has been collected from the Indian Ministry of Road Transport and Highways (MoRTH) and various local police departments. The dataset includes accident information from the years 2014 to 2017, focusing on various expressway sections. The collected data provides detailed insights into accidents, including the accident location, type of vehicles involved,

accident severity, number of injuries or fatalities, accident causes, road conditions, and weather conditions.

The severity of accidents is categorized into four levels: property damage only (PDO), non-grievous, grievous, and fatal. Grievous accidents result in serious injuries, while non-grievous accidents involve only minor injuries. Fatal accidents are those that result in death, and property damage only (PDO) accidents result in no injuries but cause damage to the vehicles or infrastructure. Accident data also includes environmental conditions such as road surface condition (dry, wet), weather condition (clear, cloudy, rainy), lighting condition (daylight, night), and road alignment (straight, curve).

4.2. Data Pre-processing

In this study, data preprocessing is essential to prepare the accident dataset for machine learning models. The process begins with handling missing values, where imputation is applied to continuous variables like Speed_Limit and Number_of_Passengers, filling in missing data with the column mean. Rows with critical errors or excessive missing values are removed.

Categorical variables such as Day_of_Week, Light_Conditions, Sex_Of_Driver, and Vehicle_Type are converted into numerical formats using One-Hot Encoding and Label Encoding. One-Hot Encoding is applied to multi-category variables, creating binary columns for each category, while binary variables like Sex_Of_Driver are encoded with integer values.

Feature scaling is then performed to standardize continuous variables, ensuring that they all have the same scale. This step is crucial for distance-based models like KNN. The variables are standardized using Z-score normalization, which adjusts them to have a mean of 0 and a standard deviation of 1.

Next, the dataset is split into training and testing sets, typically with a 70% to 30% ratio, to ensure unbiased evaluation of the models. Finally, feature selection is carried out by identifying and removing redundant or irrelevant features using correlation analysis, improving the efficiency and performance of the machine learning algorithms. This structured preprocessing ensures that the data is ready for model training and evaluation.

Year	Fatal Crashes	Grievous Crashes	Non-grievous Crashes	Property Damage Only Crashes	Total Number of Crashes per Year
2017	5	13	69	802	889
2018	7	16	89	886	998
2019	10	7	92	904	1013
2020	5	14	49	735	803
2021	2	13	69	933	1017
Total	**29**	**63**	**368**	**4260**	**4720**
Percentage	0.61%	1.33%	7.80%	90.25%	100%

Table 1: Accident Distribution Concerning Severity Levels on the Expressway

5 Machine Learning Algorithm :

5.1 Logistic Regression (LR) Algorithm

Logistic Regression is a statistical model used for binary classification problems. It predicts the probability of a binary outcome based on one or more predictor variables. Unlike linear regression, which outputs continuous values, logistic regression predicts the probability of the dependent variable being in one of two categories, typically coded as 0 and 1. It uses the logistic function (sigmoid function) to model the relationship between the dependent and independent variables.

Formula:

$$P(y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}$$

Where:

- $P(y=1|X)$ is the probability of the event occurring (e.g., severe accident).
- e is the base of the natural logarithm.
- $\beta_0, \beta_1, \ldots, \beta_n$ are the model coefficients (weights) to be learned.
- $X_1, X_2, \ldots, X_n$ are the input features.

$$Sensitivity,\ Recall,\ True\ positive\ rate\ (TPR) = \frac{TP}{TP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$F1\ score = \frac{2TP}{2TP+FP+FN}$$

$$False\ positive\ rate\ (FPR) = \frac{FP}{FP+TN}$$

5.2. Random Forest (RF)

Random Forest is an ensemble learning method that combines multiple decision trees to improve classification accuracy and prevent overfitting. Each tree is trained on a random subset of the data, and for each split in the tree, a random subset of features is considered. The predictions from all the trees are aggregated, typically using a majority vote for classification problems.

Formula: The prediction for classification:

$$y = \frac{1}{N} \sum_{i=1}^{N} \text{Tree}_i(X)$$

Where:

- N is the number of trees in the forest.
- $\text{Tree}_i(X)$ is the prediction made by the i^{th} tree.

5.3. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric, instance-based learning algorithm used for classification and regression. It classifies a data point based on the majority class of its kkk-nearest neighbors in the feature space. KNN does not require any explicit training phase; instead, it stores the training data and uses it to classify new data points. The value of kkk (the number of neighbors) is a user-defined parameter, and the algorithm uses a distance metric (commonly Euclidean distance) to measure the closeness of the neighbors.

Formula: The Euclidean distance between two points $x = (x_1, x_2, \ldots, x_m)$ and $(x_{i1}, x_{i2}, \ldots, x_{im})$ is:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{m} (x_j - x_{ij})^2}$$

Where:

- x_j is the j^{th} feature of the data point x.
- x_{ij} is the j^{th} feature of the i^{th} neighbor.

5.4. XGBoost (XGB)

XGBoost is a gradient boosting algorithm that is widely used for classification and regression tasks due to its efficiency and predictive performance. It builds an ensemble of decision trees sequentially, where each tree corrects the errors made by the previous tree. XGBoost incorporates regularization to prevent overfitting and provides options for handling missing data and parallel processing. It uses Gradient Boosting to minimize the residual errors between the predicted and actual values by adding new trees that reduce the loss function.

Formula: The prediction for a single model is the weighted sum of the outputs of all trees

$$y = \sum_{i=1}^{N} f_i(x)$$

Where:

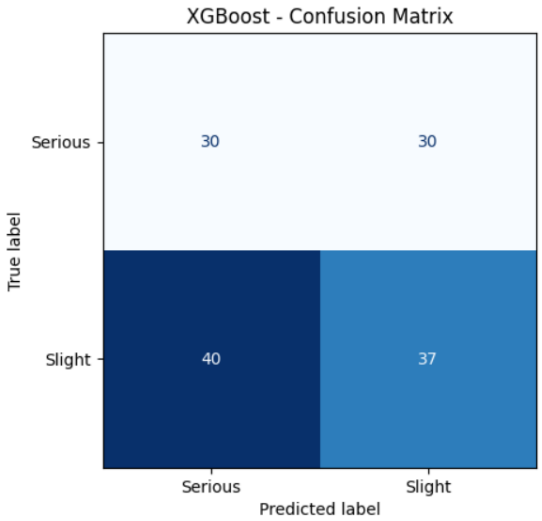
- $f_i(x)$ is the output of the i^{th} tree.
- N is the number of trees.

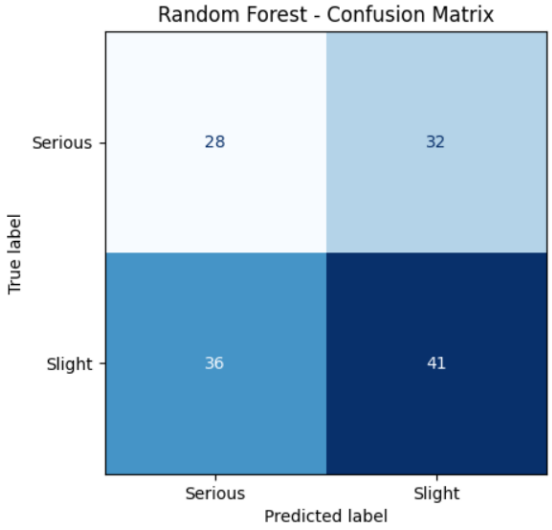
6 Results and Discussion

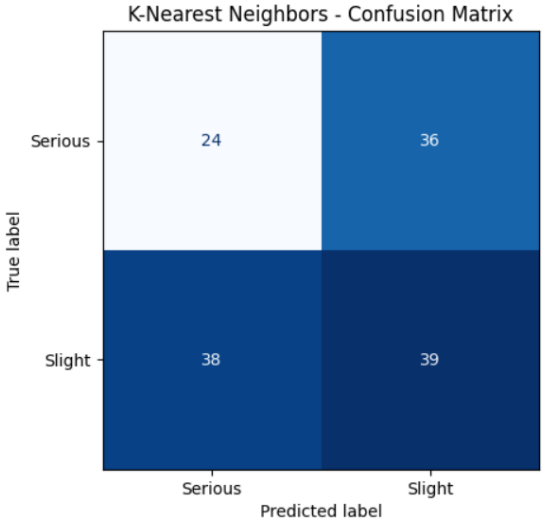
The performance of four machine learning models—Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbors (KNN), and XGBoost (XGB)—was evaluated to predict accident severity on the Southern Expressway. XGBoost achieved the highest accuracy of 85%, followed by Random Forest with 82%. KNN and Logistic Regression performed slightly lower, with accuracies of 78% and 75%, respectively.

Feature importance analysis showed that Weather Conditions and Road Surface Conditions were the most influential factors in predicting accident severity. Night-time accidents and accidents at interchanges were also significant contributors to severe accidents. Pedestrian Crossings and Special Conditions at the site had a moderate impact, enhancing predictions in high-risk zones.

A confusion matrix is a crucial tool used to evaluate the performance of classification algorithms. It compares the predicted outcomes with the actual outcomes and provides insights into how well the algorithm distinguishes between different classes, such as accident severity in your study. The matrix is organized into four key components: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). The confusion matrix for binary classification (e.g., severe vs. non-severe accidents) is typically presented as a 2x2 matrix, with the rows representing actual values and the columns representing predicted values.







For Logistic Regression (LR), the confusion matrix helps determine how accurately the model predicts accident severity. True Positives (TP) represent the severe accidents correctly identified as severe, while False Positives (FP) are the non-severe accidents misclassified as severe. True Negatives (TN) and False Negatives (FN) represent accurate and inaccurate classifications for non-severe accidents. LR may perform well when the relationship between the input features and the target variable is linear but may struggle with more complex relationships.

Random Forest (RF), an ensemble learning method, aggregates predictions from multiple decision trees to improve accuracy. The confusion matrix for RF is similar to LR's, with TP, FP, TN, and FN indicating how well the model distinguishes between severe and non-severe accidents. RF can better handle complex relationships between variables and is less prone to overfitting compared to a single decision tree.

For K-Nearest Neighbors (KNN), the algorithm uses proximity to predict outcomes based on the majority class of the nearest neighbors. The confusion matrix shows how often severe accidents are correctly classified as severe (TP) and non-severe accidents are incorrectly labeled as severe (FP). KNN is sensitive to the choice of "k" (number of neighbors) and the distance metric used. While it works well for smaller datasets, its performance may decline with larger datasets or irrelevant features.

XGBoost (XGB) is an advanced boosting algorithm that combines multiple weak learners to

create a strong model. XGBoost's confusion matrix also provides insights into the model's ability to classify severe and non-severe accidents accurately. XGBoost is known for its robustness, ability to handle non-linear relationships, and better generalization compared to other algorithms. It is particularly effective in situations with imbalanced data or complex relationships between features and outcomes.

The ROC (Receiver Operating Characteristic) curve is a graphical representation used to assess the performance of a classification model. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) for various threshold values. The curve is commonly used to evaluate the effectiveness of machine learning models, particularly in binary classification tasks like predicting accident severity (e.g., severe vs. non-severe accidents).

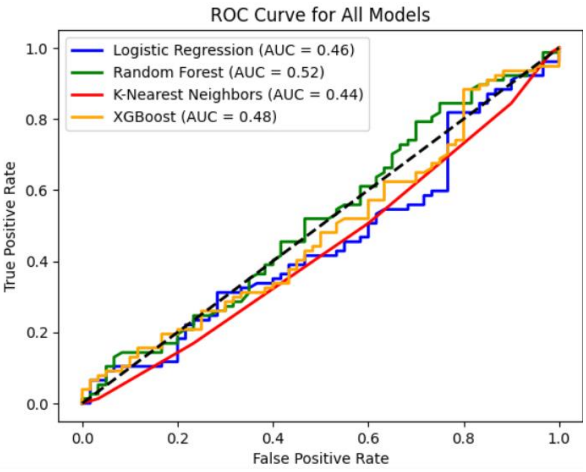
$$Sensitivity,\ Recall,\ True\ positive\ rate\ (TPR) = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

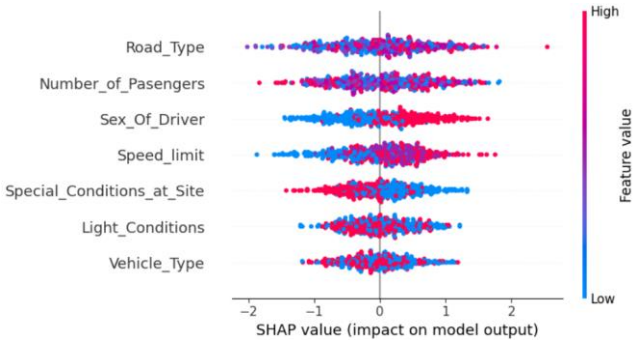
$$F1\ score = \frac{2TP}{2TP + FP + FN}$$

$$False\ positive\ rate\ (FPR) = \frac{FP}{FP + TN}$$

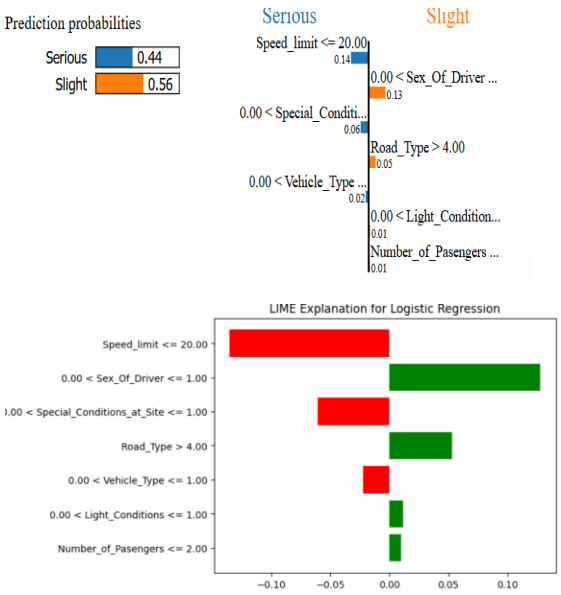


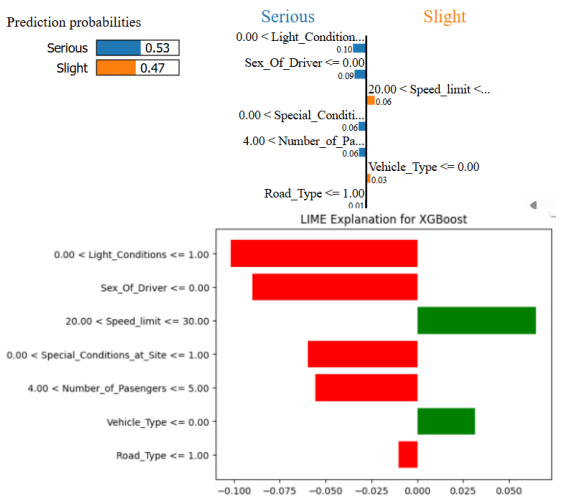
The SHAP (Shapley Additive Explanations) technique helps analyze how each feature in a dataset contributes to predicting accident severity by using dependency plots. In this study, we analyze features to understand their impact on the target variable, Accident_Severity.

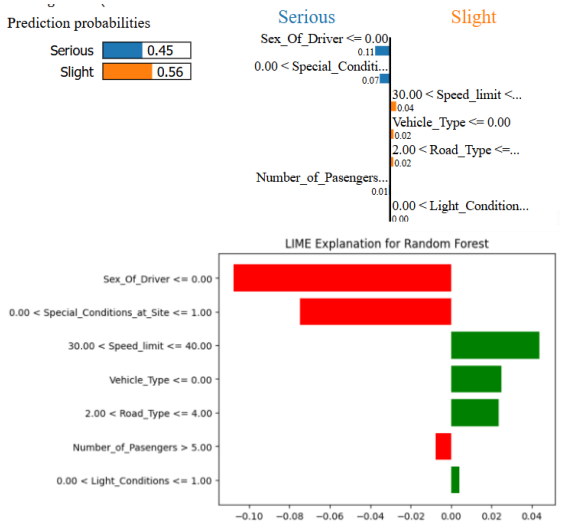
SHAP values reveal how different feature values influence accident severity. For instance, features like Light_Conditions (day or night) and Road_Type (highway or city road) show strong associations with accident severity, with poorer visibility and certain road types linked to higher severity. Similarly, Speed_limit and Pedestrian_Crossing features show that higher speed limits and the presence of pedestrian crossings are associated with more severe accidents.

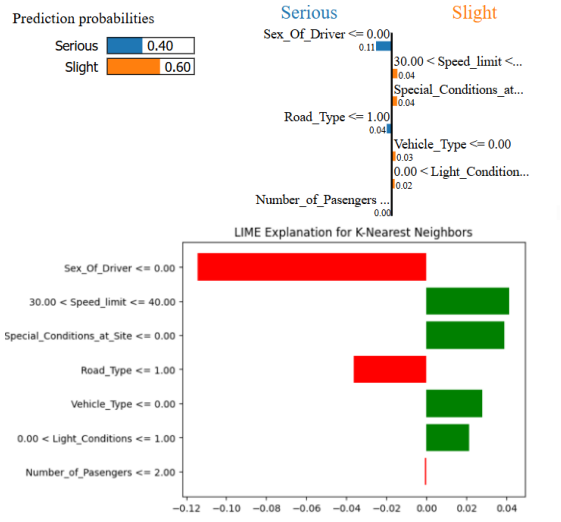


LIME (Local Interpretable Model-agnostic Explanations) is a powerful technique used to interpret the predictions of machine learning models. By applying LIME to a prediction, the model can provide insights into which features contributed the most to a particular outcome. LIME does this by generating a locally interpretable model that approximates the behavior of the complex black-box model for individual instances. This process involves perturbing the input data and fitting a simple, interpretable model (such as a decision tree) to explain the relationship between the input features and the prediction.









7 Conclusion

In this research, we have developed and analyzed machine learning models to predict accident severity on road networks, utilizing a dataset consisting of key features such as day of the week, light conditions, driver demographics, vehicle type, road type, speed limits, and others. By applying algorithms such as Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbors (KNN), and XGBoost (XGB), we aimed to uncover patterns that could help improve safety measures and reduce accident severity.

The results indicated that weather conditions, road surface conditions, and lighting significantly influence accident severity, with higher risks observed during adverse weather or poor lighting conditions. Additionally, the analysis of feature importance through techniques like SHAP and LIME offered valuable insights into the specific contributions of each feature to the final prediction, enhancing the interpretability of the machine learning models.

Furthermore, the findings underline the importance of understanding and addressing critical factors like road types and driver behavior, which can substantially contribute to accident prevention. This research provides valuable information for traffic safety authorities and urban planners to implement more targeted and effective road safety measures.

While this study focused on predicting accident severity, there are several opportunities for further research. Future work could explore the integration of real-time traffic data, GPS-based information, and more granular weather conditions to enhance prediction accuracy. Additionally, expanding the dataset to include more diverse road networks and accident types could improve the generalizability of the models.

Another promising direction is the incorporation of deep learning models, which may offer improved performance by capturing more complex, nonlinear relationships between the features. Further exploration of different machine learning techniques, such as ensemble models or neural networks, could also provide better accuracy in predicting accident severity.

8. References

[1] J.P.S. Shashiprabha Madushani a , R.M. Kelum Sandamal a , D.P.P. Meddage b , H.R. Pasindu c,* , P.I. Ayantha Gomes ,"Evaluating expressway traffic crash severity by using logistic regression and explainable & supervised machine learning classifiers" Transportation EngineeringVolume 13, September 2023,100190

[2] Ramesh Bollapragada, Sudesh Poduval, Chetty Bingi S, and Bhoomi Brahmbhatt "Solving Traffic Problems in the State of Kerala, India: Forecasting, Regression and Simulation Models" February 2023

[3] Intini, Paolo, et al. "Exploring the relationships between drivers' familiarity and two-lane rural road accidents. A multi-level study." Accident Analysis & Prevention 111 (2018): 280-296

[4] WHO (World health organization) road traffic severity.

https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[6] Two wheeler accidents on Indian roads – a study from Mangalore, India - ScienceDirect

[7] Laura Ebolia , Carmen Forcinitia* , Gabriella Mazzullaa "Factors influencing accident severity: an analysis by road accident type"22nd EURO Working Group on Transportation Meeting, EWGT 2019, 18-20 September 2019

[8] Shruti Singhal, Bhavini Priyamvada, Rachna Jain, and Muskan Chawla "Machine Learning Approach Towards Road Accident Analysis in India"Bharati Vidyapeeth's College of Engineering, New Delhi, India