# Enhancing Road Safety: Data-Driven Insights and Predictive Modeling for Car Accident Prevention


## ABSTRACT

Road safety remains a critical concern worldwide, with traffic accidents causing significant human and economic losses annually. Leveraging data-driven insights and predictive modeling offers a promising avenue for mitigating these risks. This paper presents a comprehensive approach to enhancing road safety using a dataset encompassing 46 features across 46 states in the US, spanning from February 2016 to March 2023. Through rigorous data preprocessing, including univariate analysis for outlier detection and removal, feature engineering, and handling class imbalance with Random Under Sampling, I built robust predictive models using Random Forest, Decision Tree, and Gradient Boosting algorithms. The models were evaluated using a suite of metrics, including accuracy, precision, recall, and F1 score, supplemented by detailed classification reports. The findings underscore the potential of predictive modeling in proactively identifying high-risk scenarios and informing targeted interventions for car accident prevention.


## INTRODUCTION

Road safety is an urgent and universal concern, with traffic accidents posing a substantial threat to public health and economic stability. According to the World Health Organization, road traffic accidents are among the leading causes of death globally, necessitating innovative solutions to enhance road safety measures. In recent years, the advent of big data and advanced analytical techniques has opened new frontiers in The ability to understand and mitigate these risks.

This study aims to harness the power of data-driven insights and predictive modeling to advance road safety. The dataset comprises an extensive collection of records from 46 states in the United States, spanning from February 2016 to March 2023, and includes 46 features capturing various aspects of road safety. This rich dataset provides a robust foundation for developing predictive models capable of identifying and anticipating car accidents.

My approach begins with thorough data preprocessing, including univariate analysis to detect and remove outliers, ensuring the integrity of The dataset. I also address class imbalance—a common challenge in accident prediction—using Random Under Sampling to create a balanced dataset that enhances model performance. Feature engineering plays a pivotal role in The methodology, with the creation of new features such as rush hour indicators and accident duration, which provide deeper insights into the factors influencing road safety.

I employ advanced machine learning algorithm such Random Forest, Decision Tree, KNN, Naïve Bayes, Gradient Boosting, Ada boost, —to build The predictive models. To optimize the performance of these models, I perform hyperparameter tuning and utilize grid search to identify the best parameters for each algorithm. This ensures that The models are finely tuned to achieve optimal results. The models are rigorously evaluated using a comprehensive set of metrics such as accuracy, precision, recall, and F1 score. In addition, I generate detailed classification reports to thoroughly assess performance across different classes, providing a clear view of how well The models perform under various conditions.
The findings demonstrate the significant potential of predictive modeling in enhancing road safety. By proactively identifying high-risk scenarios and providing actionable insights, these models can inform targeted interventions and policy decisions aimed at reducing the incidence

and severity of car accidents. This paper contributes to the growing body of research on data-driven approaches to road safety, offering valuable perspectives on the application of predictive analytics in this critical domain.

## LITERATURE REVIEW

[1] The 2015 WHO report indicates that globally, road traffic deaths have plateaued around 1.25 million annually, with low-income countries experiencing the highest fatality rates. While there have been improvements in road safety legislation and vehicle safety, progress is deemed too slow to achieve ambitious goals like halving road traffic deaths by 2020. The report highlights the need for urgent action to address gaps in infrastructure, enforcement, and public awareness campaigns.

[2] This paper investigates a method for analyzing the distribution of accident severities on highways. Traditionally, transportation agencies focus on accident frequency, but this study argues for considering the full spectrum of severities (property damage only, possible injury, injury, fatality).

[3] This study investigates the factors that influence the severity of accidents on the German Autobahn (highway) in the state of North Rhine-Westphalia. The researchers analyzed data from 2009 to 2011 and considered factors such as traffic information, road conditions, types of accidents, speed limits, driver demographics, and accident location. Their findings align with similar studies conducted elsewhere, suggesting that accidents are generally less severe during daylight hours and at interchanges or construction sites. Conversely, accidents involving collisions with roadside objects, pedestrians, motorcycles, or occurring in bad weather conditions tend to be more serious.

[4] The study by Leo Breiman (2001) explores Random Forests, machine learning method that combines multiple decision trees. Random Forests can outperform single decision trees by reducing variance and overfitting. Breiman also establishes the theoretical foundation for Random Forests, including how their error on unseen data converges as the number of trees increases. Finally, the study highlights how Random Forests can be used to assess the relative importance of features in a dataset.

[5] This methodology review surveys statistical and machine-learning approaches for predicting clearance times of road incidents. It synthesizes methods used to analyze incident data and forecast clearance times, highlighting advancements and challenges in improving incident management and traffic flow through predictive modeling.

[6] This research report investigates strategies to reduce motorcycle accidents in Malaysia through exposure control methods. It examines how adjusting exposure variables like traffic volume and road conditions can mitigate accident risks and improve road safety outcomes, contributing to targeted interventions in motorcycle accident prevention.

[7] Published in Lancet, this article discusses disparities in road-traffic injuries globally and proposes strategies for addressing this significant public health issue. It examines socio-economic factors, legislative measures, and public health interventions aimed at reducing the burden of road-traffic injuries and improving outcomes for affected populations.
[8] This paper presents RFCNN, a model combining machine learning and deep learning techniques for predicting traffic accident severity. It explores decision-level fusion strategies to

enhance predictive accuracy, demonstrating advancements in integrating heterogeneous data sources for improved accident severity assessment.

[9] Presented at JEEIT, this study compares various machine learning algorithms for predicting traffic accident severity. It evaluates algorithmic performance metrics and discusses the strengths and limitations of different approaches, offering insights into selecting optimal models for accident severity prediction tasks.

[10] This article discusses knowledge discovery from road traffic accident data in Ethiopia, focusing on data quality assessment, ensembling techniques, and trend analysis for enhancing road safety. It explores methods to leverage data insights for informed policy-making and intervention strategies in Ethiopian road safety management.

[11] This methodology review article provides a comprehensive overview of statistical and machine-learning methods used for predicting clearance times of road incidents. It examines the methodologies employed, their strengths, and limitations, offering a critical synthesis of approaches to improve incident management and traffic flow prediction.

[12] This empirical analysis investigates the factors influencing highway accident severities using the mixed logit model. It explores driver, vehicle, and environmental variables to understand their impact on accident outcomes, contributing valuable insights into mitigating accident severity through targeted interventions and policy measures.

## METHODOLOGY:

**Data Preprocessing:** To ensure data integrity, I meticulously examined the dataset for duplicate entries. The thorough inspection confirmed that there were no duplicate records, establishing a clean and reliable dataset as the foundation for The analysis. I engaged in dimensionality reduction to streamline The dataset and focus on the most relevant features. Several features were deemed unnecessary and removed to simplify the dataset and enhance computational efficiency. These included End_Time, Weather_Timestamp, Street and others deemed redundant. Handling missing values involved removing features like End_Lat and End_Lng due to significant missing data, eliminating null values in Precipitation and Wind Chill to ensure data accuracy, and employing interpolation techniques to restore data completeness, thereby enhancing the integrity of The dataset for further analysis and modeling.

**Feature Engineering:** To enhance the value of The dataset and gain deeper insights, I performed comprehensive feature engineering. Temporal Features were derived from the Start_Time column, extracting the year, month, day, and hour of each accident. These features enable us to identify detailed patterns and trends in accident occurrences over time, providing a richer temporal context for The analysis. I also created the Accident_Type feature by categorizing each event based on its text description into groups like 'Accident', 'Crash', 'Incident', or 'Other'. This classification offers a nuanced understanding of the nature of accidents, facilitating more targeted analysis. In terms of weather and temperature, I introduced Temperature_Category and Weather_Bin. These new features categorize temperatures into 'Cold', 'Moderate', or 'Warm', and weather conditions into 'Clear', 'Cloudy', or 'Rainy', respectively. This simplification aids in efficiently analyzing the impact of weather and temperature on accident occurrences. Additionally, I calculated Accident_Duration by measuring the time difference between the Start_Time and End_Time of each accident. This feature provides crucial insights into the duration of accidents, helping to understand their timelines and

patterns. Finally, I added the Is_Rush_Hour feature to classify accident occurrences into rush hour periods. This helps us analyze whether accidents are more frequent during peak traffic times, offering valuable information on the influence of traffic congestion on accidents. These engineered features significantly enrich The dataset, enabling a more robust analysis and better-informed modeling.

**EXPLORATORY DATA ANALYSIS:**

The exploratory data analysis focuses on traffic accidents across different states in the USA. The goal is to uncover patterns and insights that can help understand the distribution and frequency of accidents across various regions and times. I employ both visual and statistical methods to dive deep into the data, offering a comprehensive look at how and where these accidents occur.

**Geographic Analysis of Accidents:**

Scatter Plot Analysis: I begin The analysis with a scatter plot, where each blue dot represents an individual accident's location based on latitude and longitude coordinates. The scatter plot (Fig 1.1) reveals dense clustering along the east and west coasts and in major urban areas, suggesting these regions experience higher traffic volumes and consequently more accidents.

State-wise Accident Quantification: I quantified the number of accidents by state and identified that California leads the USA in the number of traffic accidents, followed closely by Texas and Florida, while other states also report significant but comparatively lower accident rates.

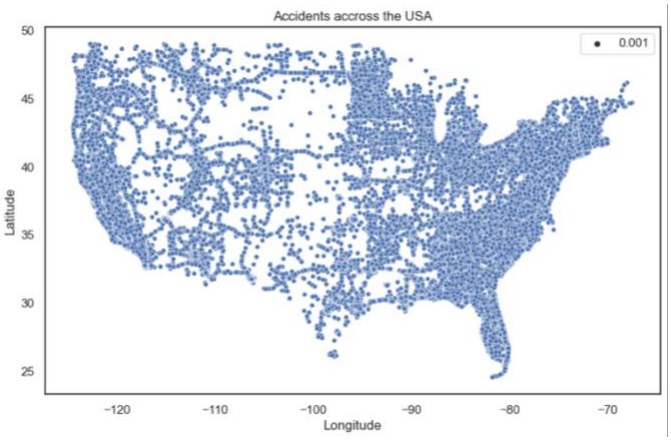


Fig 1.1

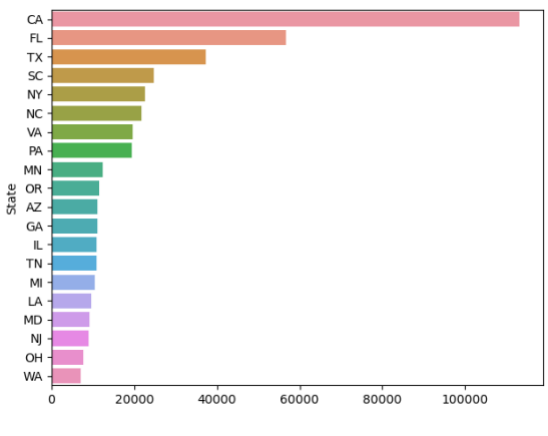


Fig 1.2

The interactive Tableau dashboard (Fig 1.3) uncovers detailed weekly and daily trends in traffic accidents. Fridays experience the highest number of accidents, suggesting increased travel or end-of-week activities, while the weekend shows a significant decline, likely due to reduced traffic volumes or different travel behaviors. Daily analysis reveals that on weekdays, accidents peak during morning and evening rush hours, with a midday dip around lunchtime. Conversely, weekend accidents are more evenly spread throughout the day, possibly reflecting later start times, leisure activities, or weekend outings. In a state-specific analysis focusing on Illinois, weekday patterns align with the national trends, but weekend accidents in Illinois are evenly distributed throughout the day, indicating unique local traffic behaviors. This comprehensive analysis highlights how both time of week and state-specific factors influence traffic accident patterns. I also examined Illinois separately in The interactive dashboard and it

showed weekday accident trends in Illinois align with national data, showing morning and evening peaks. On weekends, however, accidents in Illinois are more evenly distributed across the day, differing from the national pattern.

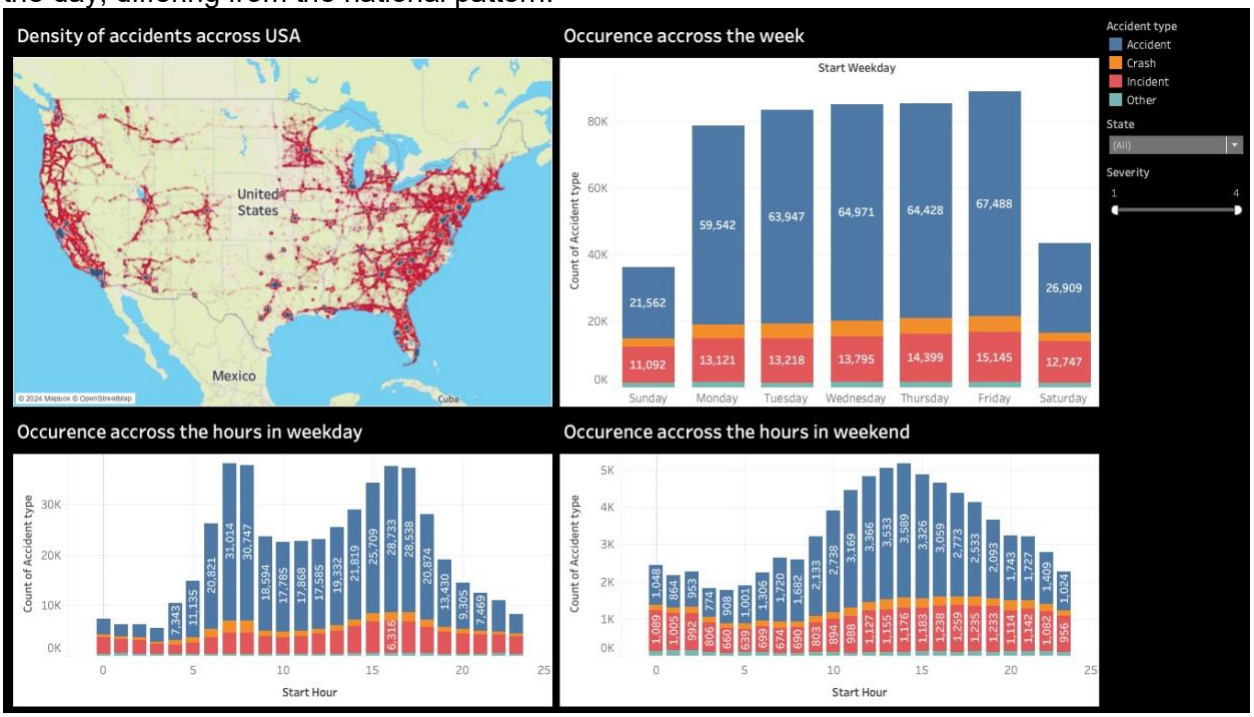


Fig 1.3

**Impact of Traffic Control Features and Weather Conditions**

Traffic Control Features: The analysis of traffic control features (Fig 1.4) showed that the majority of accidents occur near traffic signals, followed by crossings and junctions. This suggests that areas with these traffic control features are more prone to accidents, possibly due to increased vehicle interactions and decision points.

Weather Conditions: Comparing weather conditions (Fig1.5) at the time of accidents revealed that most accidents occur under fair and clear conditions, followed by cloudy conditions. This indicates that while poor weather can increase accident risks, most accidents happen under seemingly normal weather conditions, possibly due to higher traffic volumes or a false sense of security among drivers.

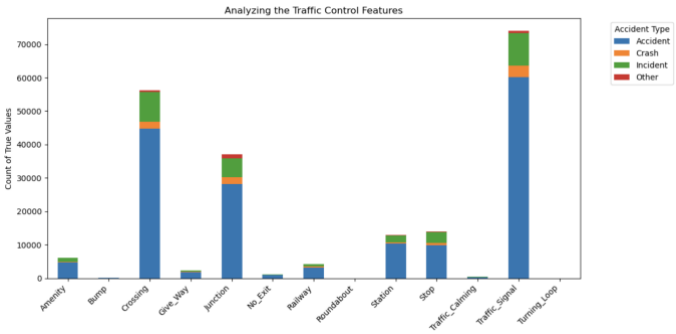


Fig 1.4

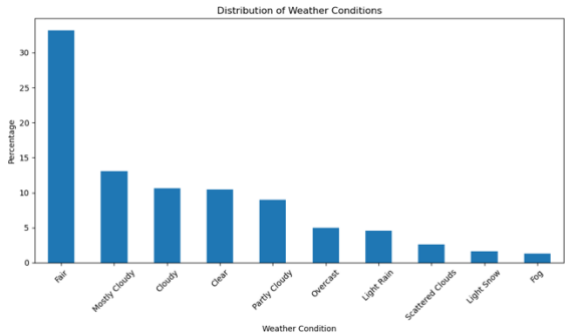


Fig 1.5

**Statistical Summary and Outlier Analysis**

Distribution and Outliers:
I performed a statistical summary to understand the distribution of various metrics, such as wind chill temperature. By comparing the mean, maximum, and standard deviation, I identified potential outliers in the dataset. These outliers were further analyzed and removed to ensure a cleaner and more accurate dataset for The analysis.

From the EDA, I got insights into the spatial and temporal distribution of traffic accidents across the USA. Key findings highlight the significance of traffic volume, control features, and daily/weekly patterns in influencing accident occurrences. By understanding these factors, I can better target safety measures and policies to reduce traffic accidents and improve road safety.

**MODELS:**

To prepare the dataset for modeling, I first encoded categorical variables using Label Encoding. This step was crucial to transform categorical data into a numerical format that could be processed by The machine learning algorithms. I identified columns containing boolean, object, or categorical data types and applied Label Encoding to them, ensuring that each categorical variable was appropriately converted into numeric form.

Following encoding, I conducted a Variable Inflation Factor (VIF) analysis to assess multicollinearity among the numerical features. VIF quantifies how much a feature is correlated with other features. High VIF values indicate high multicollinearity, which can lead to inflated standard errors and unreliable estimates in regression models. The VIF analysis identified several features exhibiting high multicollinearity within The dataset. Notably, Pressure(in) showed a VIF of 165.44, while Start_Lat and Temperature(F) had VIFs of 67.96 and 54.82, respectively. Additionally, Visibility(mi) and Humidity(%) demonstrated VIFs of 18.39 and 13.65, respectively. These findings highlighted significant correlations among these variables, suggesting potential issues that could affect model performance. I carefully considered these results during The feature selection and model evaluation stages to mitigate the impact of multicollinearity and ensure the robustness of The analytical approach.

And I developed and evaluated machine learning models: Random Forest, Decision Tree, and K-Nearest Neighbors (KNN), Naïve Bayes, Gradient boosting, Ada boost. Each model was assessed based on its ability to predict the severity of traffic accidents, with performance metrics including accuracy, precision, recall, and F1 score. And the models results are tabulated in the Table 1.

**Decision Tree and Random Forest Classifiers:**
The Decision Tree and Random Forest models exhibited notable improvements following hyperparameter tuning. The Decision Tree model, initially achieving an accuracy of 62% with default settings, significantly enhanced its performance to 86% after tuning. This improvement was mirrored in precision, recall, and F1-score metrics, underscoring the effectiveness of optimizing parameters such as criterion, max_depth, and min_samples_split. Similarly, the Random Forest model, starting at 83% accuracy, improved to 85% post-tuning. This model benefited from optimizing parameters like n_estimators, max_depth, and bootstrap, which helped mitigate overfitting and enhance overall predictive accuracy.

**KNN Classifier:**
The KNN model, while demonstrating lower performance initially with a 40% accuracy using default parameters, showed improvement to 48% accuracy following hyperparameter tuning.

Tuning parameters such as n_neighbors and metric (e.g., euclidean, manhattan) contributed to better capturing the local structure of data points, thereby improving its predictive capabilities.

**Naive Bayes Classifier:**
The Naive Bayes classifier, known for its simplicity and assumption of independence among features, exhibited significant improvement through hyperparameter tuning. Starting with a low accuracy of 24% using default parameters, tuning parameters related to distribution assumptions (e.g., GaussianNB, MultinomialNB) boosted accuracy to an impressive 86%. This demonstrates the critical role of parameter selection in optimizing performance for probabilistic classifiers.

**XGBoost and AdaBoost Classifiers:**
The XGBoost and AdaBoost models, both ensemble methods, showed consistent performance with default parameters and marginal improvement following tuning. The XGBoost model achieved 73% accuracy post-tuning, slightly up from 72% initially, reflecting the robustness of boosting algorithms in handling complex datasets. Similarly, the AdaBoost model, starting with 64% accuracy, improved to 86% after tuning. Adjusting parameters such as learning_rate and base_estimator (e.g., DecisionTreeClassifier) played a pivotal role in enhancing predictive power by iteratively correcting classification errors.

| Model | Default Accuracy | Tuned Accuracy | Default Precision | Tuned Precision | Default Recall | Tuned Recall | Default F1-score | Tuned F1-score |
|---|---|---|---|---|---|---|---|---|
| **Decision Tree** | 0.62 | 0.86 | 0.62 | 0.85 | 0.63 | 0.86 | 0.62 | 0.85 |
| **Random Forest** | 0.83 | 0.85 | 0.8 | 0.84 | 0.83 | 0.83 | 0.79 | 0.84 |
| **KNN** | 0.4 | 0.48 | 0.42 | 0.49 | 0.39 | 0.5 | 0.4 | 0.48 |
| **Naive Bayes** | 0.24 | 0.86 | 0.06 | 0.85 | 0.24 | 0.86 | 0.09 | 0.85 |
| **XGBoost** | 0.72 | 0.73 | 0.72 | 0.73 | 0.72 | 0.72 | 0.71 | 0.72 |
| **AdaBoost** | 0.64 | 0.86 | 0.64 | 0.85 | 0.63 | 0.86 | 0.64 | 0.85 |

Table 1. Comparison of model results

**CONCLUSION:**

In conclusion, hyperparameter tuning proved instrumental in optimizing the predictive performance of each machine learning model evaluated for traffic accident prediction. Decision Tree, Random Forest, and AdaBoost models particularly stood out, achieving significant accuracy improvements post-tuning. The results underscore the importance of selecting appropriate parameters tailored to each model's algorithmic characteristics and dataset complexities. Moving forward, further exploration into feature engineering and dataset refinement could potentially yield additional enhancements in predictive accuracy and robustness across these models.

## AUTHOR CONTRIBUTIONS

**Anbu Suriya Kumar** – Took charge of sourcing the dataset, finalizing the project scope, performing feature engineering, creating dashboards, contributing significantly to data cleaning, conducting literature review, building models, and writing the report sections on models, results, and conclusions.

**Snehaa Durairaj** – Conducted thorough literature reviews, performed exploratory data analysis (EDA), handled encoding tasks, conducted VIF tests, contributed to data cleaning efforts, and authored the abstract, introduction, and literature review sections of the report.

## REFERENCES

[1] World Health Organization, Global Status Report on Road Safety 2015, World Health Org., Geneva, Switzerland, 2015.

[2] J. C. Milton, V. N. Shankar, and F. L. Mannering, ‘‘Highway accident severities and the mixed logit model: An exploratory empirical analysis,’’ Accident Anal. Prevention, vol. 40, no. 1, pp. 260–266, 2008.

[3] H. Manner and L. Wünsch-Ziegler, ‘‘Analyzing the severity of accidents on the German autobahn,’’ Accident Anal. Prevention, vol. 57, pp. 40–48, Aug. 2013.

[4] L. Breiman, ‘‘Random forests,’’ Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.
R. E. Schapire, ‘‘A brief introduction to boosting,’’ in Proc. IJCAI, vol. 99, 1999, pp. 1401–1406.

[5] J. Tang, L. Zheng, C. Han, W. Yin, Y. Zhang, Y. Zou, and H. Huang, ‘‘Statistical and machine-learning methods for clearance time prediction of road incidents: A methodology review,’’ Anal. Methods Accident Res., vol. 27, Sep. 2020, Art. no. 100123.

[6] M. N. Ghani, A. Zainuddin, R. R. Umar, and H. Hussain, ‘‘Use of exposure control methods to tackle motorcycle accidents in Malaysia research report 3/98,’’ Road Saf. Res. Centre, Serdang, Malaysia, Tech. Rep., 1998.

[7] S. Ameratunga, M. Hijar, and R. Norton, ‘‘Road-traffic injuries: Confronting disparities to address a global-health problem,’’ Lancet, vol. 367, no. 9521, pp. 1533–1540, May 2006.

[8] M. Manzoor *et al.*, "RFCNN: Traffic Accident Severity Prediction Based on Decision Level Fusion of Machine and Deep Learning Model," in *IEEE Access*, vol. 9, pp. 128359-128371, 2021

[9] R. E. AlMamlook, K. M. Kwayu, M. R. Alkasisbeh and A. A. Frefer, "Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity," 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, 2019, pp. 272-276

[10] Beshah, Tibebe & Dedefa, Dejene & Abraham, Ajith & Krömer, Pavel. (2012). Knowledge Discovery from Road Traffic Accident Data in Ethiopia: Data Quality, Ensembling and Trend Analysis for Improving Road Safety. Neural Network World. 22. 10.14311/NNW.2012.22.013.

[11] Jinjun Tang, Lanlan Zheng, Chunyang Han, Weiqi Yin, Yue Zhang, Yajie Zou, Helai Huang,

Statistical and machine-learning methods for clearance time prediction of road incidents: A methodology review, Analytic Methods in Accident Research, Volume 27, 2020, 100123, ISSN 2213-6657

[12] J. C. Milton, V. N. Shankar and F. L. Mannering, "Highway accident severities and the mixed logit model: An exploratory empirical analysis", *Accident Anal. Prevention*, vol. 40, no. 1, pp. 260-266, 2008.