2022 International Conference for Advancement in Technology (ICONAT)
Goa, India. Jan 21-22, 2022

Data Visualization and Machine Learning Approach for Analyzing Severity of Road Accidents

Rasika Vijithasena
Department of Information Technology
University of Moratuwa
Sri Lanka
rasikav@uom.lk

Wikasitha Herath
Department of Computer Science & Engineering
University of Moratuwa
Sri Lanka
wikasitha.20@cse.mrt.ac.lk

***Abstract*— Road accidents are a major critical problem in recent days, which cause death, severe injuries, disabilities to human lives, and substantial economic losses worldwide. A detailed analysis of the factors responsible for the accidents is required to reduce the accident rate. In this study, a descriptive analysis has been performed in-depth to identify the significant factors influencing accident severity. Machine learning techniques have been applied to predict the severity of accidents considering location, time, infrastructure, and environmental conditions-related factors that may cause road accidents. Here the findings of this study emphasize that the mediate severity of accidents have a high frequency of occurring rather than severities of very low and high risk. Moreover, the factors like infrastructure, day of the week, and weather conditions influence the severity of accidents differently. The Random Forest algorithm took the best performance with 97.2% high accuracy in predicting the severity of road traffic accidents. These results can guide relevant parties to identify dangerous situations and take the necessary actions to improve road safety by reducing accidents.**

Keywords*— *Data visualization, Descriptive analytics, Machine Learning, Road accidents

I. INTRODUCTION

Road accidents are increased worldwide with the rapid growth of population and vehicles. The WHO records show that approximately 1.3 million people die due to road accidents every year, and 20-50 million people suffer from non-fatal injuries. Most countries have to face a 3% loss on their Gross Domestic Product due to traffic accidents [1]. Hence, it is required to get necessary safety action plans to reduce and prevent road accidents and their severity.

Since this is a critical global problem, authorities should pay attention to improving road safety with achieving minimum accident occurrences. Using historical data of road traffic accidents, researchers can figure out the influencing factors in such accidents, which play a key role in finding ideal solutions to address this issue in the future.

Data Visualization is used to detect data insights such as patterns, trends, and outliers by using charts, graphs, and maps like visual representations. It can be used to perform descriptive analysis on large datasets [2]. Machine learning is an artificial intelligence technique that can be used to implement models through past data. Predictions can be made using a Machine Learning model without being explicitly programmed. Machine learning and different statistical approaches have been applied to analyze road accidents datasets in past research studies, and valuable information has been extracted regarding road accidents.

A. Motivation

Nowadays road accidents are rising rapidly everywhere, and it badly affects human lives and the economy of countries. For that, there is an existing research gap which is needed more research studies to identify factors and reasons for rapidly growing accident occurrences and their severity levels. Those can be useful for governments to make their safety plans. The significance of this research study is specially to focus on infrastructure factors, different weather conditions, accident locations towards on having road accidents and its severity.

B. Objectives

The main objective of this research is to identify the influencing factors for determining the severity of road accidents and develop a predictive model for forecasting the severity of accidents considering different features. Further, this study has been conducted to achieve the following specific objectives.

(1) Identifying the influencing weather conditions, infrastructure, and environmental factors which influence accident severity.

(2) Implementing a machine learning classification model for predicting accident severity level using most influencing features.

This paper is organized as follows. The related work section contains the literature about road accident analysis conducted as research works. Section 3 describes the data collection, data pre-processing, model creation, and model testing processes. The results of descriptive analysis and evaluation results of classification algorithms are presented in section 4. Finally, section 5 discusses the conclusion and future works of this study.

II. RELATED WORK

Many research studies have been undertaken to address various aspects of road accidents, with the majority of them focused on machine learning approaches for traffic accident analysis.

Road accidents can be occurred due to the driver's emotions such as happiness, sadness, and anger. And also, weather conditions, road traffic, road conditions, driver's health, and speed can contribute to traffic accidents [3]. S. Vasavi has extracted the hidden patterns in the accidents

2022 International Conference for Advancement in Technology (ICONAT) | 978-1-6654-2577-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICONAT53423.2022.9726042

which cause the accidents using machine learning techniques such as clustering and association rule mining. According to the results, the number of accidents is higher on weekends and cold nights.

For discovering meaningful patterns and trends among the traffic accident dataset, data visualization techniques can be used. Feng et al. [4] have identified accident-causing attributes using data visualization techniques and developed a LSTM model using time series and deep learning techniques to forecast the accidents that can occur in the future. According to their analysis, more accidents had occurred on normal days rather than snowing or heavy rainy days with high winds. Babic and Zuskáþová [5] have applied descriptive and predictive analytics on a historical road accident data sample from the UK from 2005 to 2015. According to their results, a smaller number of accidents occurred on Sunday. But accidents that occur on Sundays have the highest probability of fatal consequences. And males are facing accidents when compared with females.

Human factors (age, gender, behavioral and physical characteristics of driver), environmental factors (weather, lighting condition), and vehicle-related factors (type of vehicle, braking system, safety features, tire quality) are influencing the severity of road accidents. Musa et al. [6] have conducted a study to evaluate the factors influencing the severity of road accidents in Malaysia. An ordered logistic regression model was developed, considering accident severity as the dependent variable and the risk factors of accident severity as the independent variable. As per the result from the analysis, the likelihood of the more severe accident severity due to the poor horizontal alignment was around 0.4 times lower when compared to the absence of horizontal alignment. It has been determined that the current findings may aid local governments in taking proactive measures to prevent serious traffic accidents on road segments with standard horizontal alignment.

Isaac and Alice [7] proposed an ordered logistic regression model using 1989 to 2019 accident data from the Motor Traffic and Transport Department database to identify factors that have a significant impact on road accident severity in Ghana. According to the results, day of week, road type, the nature of the vehicle, average speed level, and location (rural or urban) contribute to accident severity. Danthanarayana and Mallikahewa [8] conducted research to investigate the enduring factors that influenced road accidents from 1997 to 2017 in Sri Lanka. According to the results, increasing road lengths and the number of vehicles could increase severe road accidents. Chukwutoo et al. [9] conducted a study to analyze road accidents in Nigeria, developing ARIMA and ARIMAX models for predicting the frequency of accidents occurring. ARIMAX model outperformed ARIMA. The findings of the study show that all contributing factors incorporating human, vehicle, and related environmental factors have a significant impact on accident occurrence.

Bahiru et al. [10] conducted a study to identify influencing factors for road accidents and predict the severity of traffic accidents using data mining techniques. Decision Tree (ID3, J48, and CART) and Naive Bayes models were developed to predict the severity of accidents. According to the experimental results J48 classifier outperforms the other classifiers by providing the highest accuracy results. The study revealed that speed limit, weather condition, lane number, lighting condition, and accident time are the most influential factors in road accidents. Gender, age, accident location, and vehicle type, on the other hand, have less of an impact on the severity of a road accident.

III. METHODOLOGY

A. Data Collection

A countrywide car accident dataset was collected from February 2016 to December 2019, covering 49 states of the USA to analyze the road accident severity. This dataset has been taken from public free Kaggle datasets [11]. It contains around 3.0 million road accident records. These traffic data were captured from different sources such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road networks.

Some of the selected features are included in the following.

Traffic attributes:

- Source: The source of the accident report
- TMC: Traffic Message Channel (TMC) code which provides a more informative description of the accident.

Weather conditions:

- Temperature(F)
- Humidity (%)
- Pressure(in)
- Visibility(mi)
- Weather condition
- Wind speed(mph)
- Wind direction
- Wind Chill(F)
- Weather timestamp: Timestamp of weather record
- Sunrise_sunset: Sunset or sunrise

Infrastructure factors:

- Bump
- Crossing
- Junction
- Traffic Signal
- Railway
- Roundabout
- Station
- Give_Way: Availability of the give-way sign nearby
- No_Exit: Availability of no exit sign nearby
- Stop: the presence of stop signs nearby
- Traffic_Calming: Availability of traffic calming means nearby
- Turning_loop: Availability of turning loop in a nearby location

Location and time:

- Number
- Street
- Side
- City
- Country
- State
- Zipcode
- Time zone
- Start_Time: Accident start time
- End_Time: Accident end time
- Start_Lat: Latitude of the accident start place
- Start_Lng: Longitude of the accident start place
- End_Lat: Latitude of the accident end place
- End_Lng: Longitude of the accident end place
- Distance: The length of the road that affected by the accident
- Airport_code

- Astronomical_Twilight: Period of the day considering astronomical Twilight

Class variable: Severity - Represent the severity of the accident, four severity levels (1-4), where 1 indicates the least impact on traffic

B. Data Preprocessing

Data pre-processing is primarily concerned with the consistency and completeness of the data and its accuracy in the machine learning environment. It transforms original data into efficient and usable formats. To achieve better accuracy, incomplete values, null fields, and various repetitive values need to be handled in the data cleaning step. And data transformation and data reduction steps need to be carried out further [12]. Hence dataset is thoroughly screened, and the following steps were performed on the dataset.

- Remove columns that contained a huge number of null values (End_Lat, End_Lng, Precipitation(in), Number, Wind_Chill(F))
- Replace rest of missing values with the mean value of the column
- Remove columns that contains descriptive information (description)
- Remove columns that do not carry information about the severity of the accident (ID, Source)
- Feature addition: year, month, day, weekday, hour, and minute features were created using start_time
- Normalization: To improve the performance of our models, normalized the values of the continuous features. (Temperature, humidity)
- Feature encoding: Encode the categorical variables such as side, city, wind direction

C. Descriptive Data Analysis

Descriptive analytics involves analyzing the past to examine what has happened and provide insights into approaching the future [13]. The main objective of descriptive analytics is to reveal the reasons behind the success or failure that occurred in the past.

The road accident severity analysis has been carried out considering the available features of the dataset. Severity distribution analysis and severity-wise analysis based on the day of the week, weather conditions, infrastructure factors (near bump, near traffic signal, near the junction, near rounder bound) have been done using data visualization techniques.

The correlation coefficient of the variables is calculated to identify the influence of each feature on the dependent variable [14].

D. Model Creation

Predictive analysis analyses the past data patterns and trends and provides future speculations using them. Predictive analysis uses statistics, machine learning, data mining, etc., in predicting future trends and outcomes [15]. Machine learning techniques have become popular in applying predictive analytics due to their exceptional usage on a large scale. Therefore, machine learning and artificial intelligence algorithms can be used to optimize and find new statistical patterns in road traffic accident analysis.

1) Logistic regression

Logistic regression is a statistical model that uses a logistic function to model a binary dependent variable in its most basic form. It can be binomial, ordinal, or multinominal. Binomial logistic regression applies to the scenarios where the dependent variable has two possible values such as 0 and 1. Multinominal can apply to scenarios where the dependent variable has three or more values [16].

2) K-Nearest Neighbors

The K-Nearest Neighbors (KNN) algorithm is a supervised machine learning algorithm that is simple and non-parametric. It can be used to implement classification and regression models. Since it's easy to implement, it is mostly used in the industry. But it has the disadvantage of becoming noticeably slower as the amount of data in use grows [17].

3) Random Forest

Random forest is a machine learning algorithm that consists of many decision trees. Bagging or bootstrap aggregation is used to train the 'forest' created by the random forest algorithm. For the classification and regression problems, the Random Forest technique can be used. [18]

4) Decision Tree

The decision tree is a machine learning technique, mostly used for implementing classification and prediction models in the form of a tree structure [19]. A test on an attribute is represented by each internal node. A class label is represented by its leaf nodes. The outcome of the test is represented by each branch. The root node is the decision tree's topmost node and represents the best predictor node.

E. Model Evaluation

Model evaluation is one of the steps of machine learning model implementation. In this step, different machine learning models are evaluated using evaluation metrics such as accuracy, RMSE, precision, and the best model is selected considering model performance [20]. In this study, prediction models are evaluated using various measurements such as accuracy, precision, recall, and f1-score. The best model which provides the highest accuracy was selected as the best model.

IV. Results and Discussion

The road accident dataset includes around 3 million records from different states of the United States. Different descriptive-analytical steps have been performed to analyze the dataset from different perspectives considering the accident's severity.

The distribution of the severity values can be seen in figure 1. Severity 2 type accidents count to more than 12000, which have the highest frequency and lowest shown by severity type 1 accident making the severity three accidents to the second highest frequency. Severity 4, which has the highest risk, has a low count below 5000 and medium risky severity is of the highest frequency.

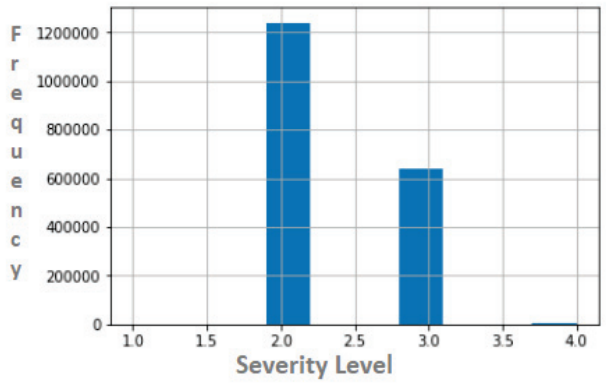

Fig. 1. Data Distribution among severity types

The analysis in Figure 2 indicates the first five days of the week has four times higher accidents count than the other two days, in each severity type. Weekend shows low risk in happening accidents rather than weekdays. And severity-wise, the charts are merely similar with small variations.

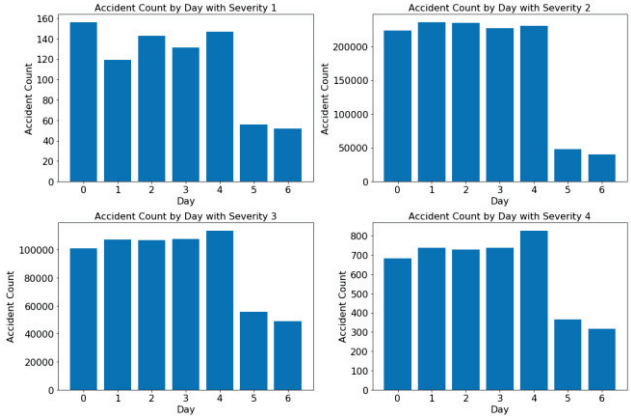

Fig. 2. Severity wise accident distribution based on day of the week

When considering the weather conditions, most severity two and severity 3 type accidents have occurred when the weather is clear (Figure 3).

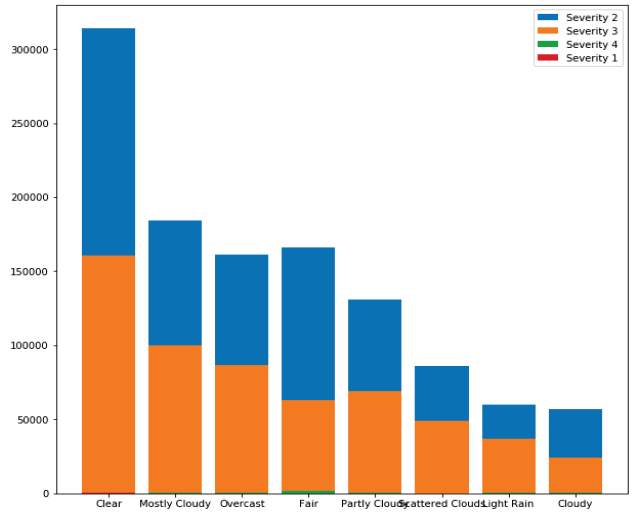

Fig. 3. Severity Distribution based on weather

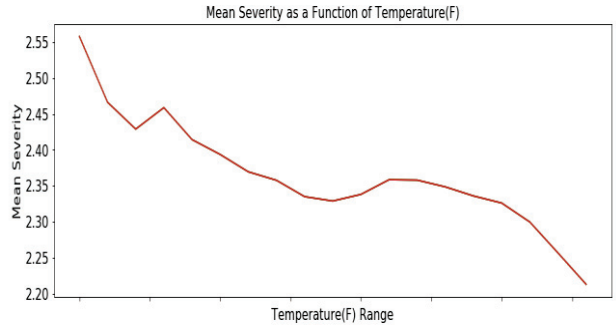

Fig. 4. Mean severity based on temperature

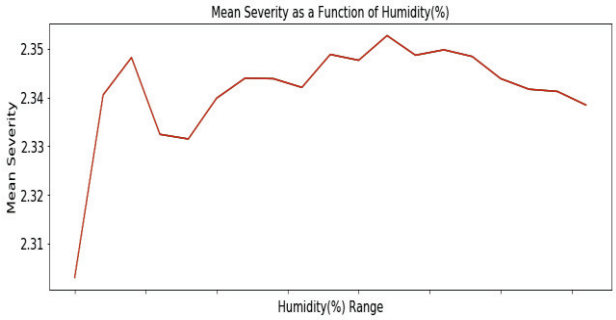

Fig. 5. Mean severity based on humidity

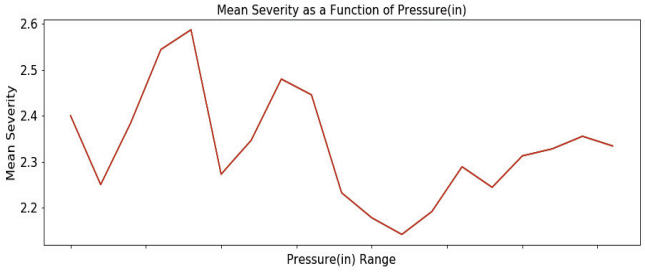

Fig. 6. Mean severity based on pressure

Temperature, Humidity, and Pressure have been plotted against mean severity, as shown in figures 4, 5, and 6, respectively. Severity increases as a function of humidity with fluctuations. Pressure and temperature are inversely proportional to mean severity.

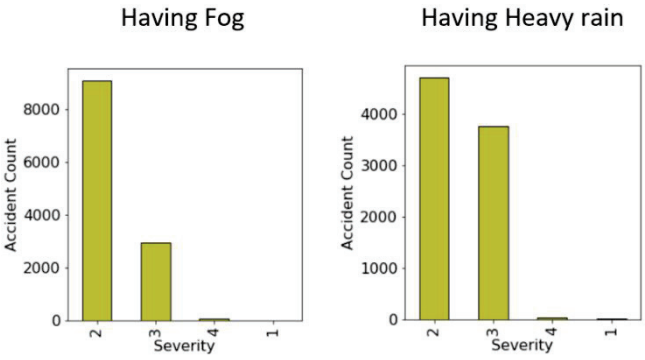

Fig. 7. Severity distribution based on different weather conditions

Different considerable weather conditions are separately analyzed, namely fog and heavy rain. As figure 7 represents, severity 2 type accidents have occurred mostly under these weather conditions.

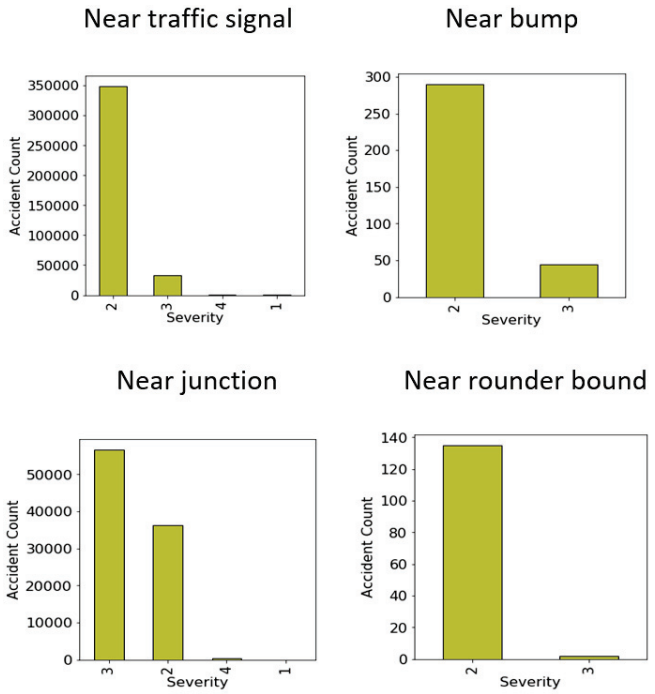

Fig. 8. Severity distribution based on different infrastructure conditions

As per the graphs shown in figure 8, severity 2 shows high frequency in the other three types of infrastructure except for near junctions, which indicates the highest frequency in severity 3. Values show the high number of severity two accidents near traffic signals and the low number of counts near rounder bound. Analysis implies the highest risk areas near traffic signals and junctions.

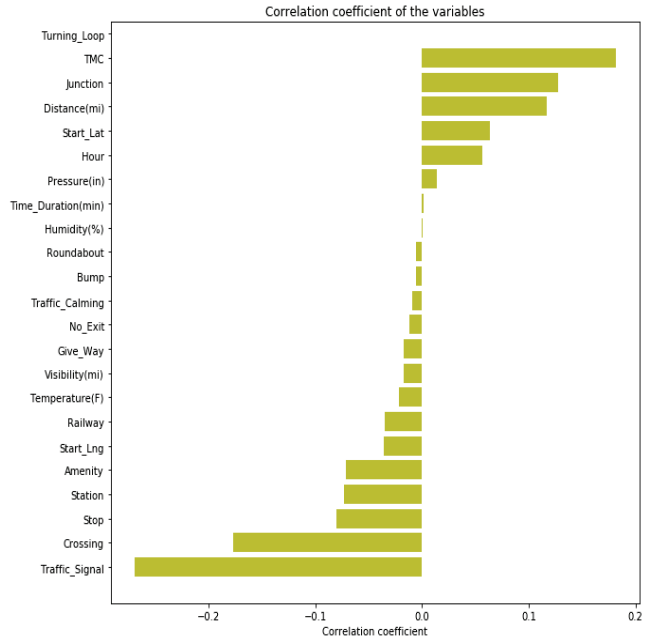

Fig. 9. The correlation coefficient of the features

The correlation of the target variable with the given set of variables is low overall (figure 9). There are some variables with no correlation, like Humidity and Time duration, which implies that those factors have no impact on severity count.

TABLE I. CORRELATION COEFFICIENT VALUES

Column labels	Correlation values
Traffic signal	-0.268792
Crossing	-0.176923
Stop	-0.080385
Station	-0.072823
Amenity	-0.071612
Hour	0.056853
Start_Lat	0.063626
Distance	0.116984
Junction	0.127309
TMC	0.181313

Table 1 represents the correlation values greater than 0.05 and -0.05. The highest negative correlation among all factors, -0.268792 is represented by Traffic signal indicates inversely low proportional impact where severity decreases when Traffic Signal increases and positive correlation, 0.181313 is shown by TMC where TMC increases the severity will also increase.

TABLE II. ACCURACY RESULTS

Model	Accuracy
Log Regression	0.955
KNN	0.938
Decision Tree	0.962
Random Forest	0.972

TABLE III. PRECISION, RECALL, F1-SCORE VALUES

Model	Severity	Precision	Recall	f1-score
Logistic Regression	1	0.0	0.0	0.0
	2	0.97	0.99	0.98
	3	0.75	0.56	0.64
	4	0.0	0.0	0.0
KNN	1	0.0	0.0	0.0
	2	0.94	1.00	0.97
	3	0.78	0.19	0.31
	4	0.0	0.0	0.0
Decision tree	1	0.0	0.0	0.0
	2	0.97	0.98	0.98
	3	0.76	0.67	0.71
	4	1.00	0.33	0.5
Random forest	1	0.0	0.0	0.0
	2	0.98	0.99	0.98
	3	0.88	0.68	0.77
	4	0.0	0.0	0.0

According to the accuracy results (table 2), the model created using the Random Forest method has provided the highest accuracy (97%) than other models. From table 3, the

Random Forest model represents the overall best evaluation results for precision, recall, and f1-score measures as well.

V. Conclusion

Roads are the main means of transport all over the world which in turn leads to a substantial number of accidents. This paper has analyzed almost all three million data indicating various road accident types and casualty severity. The key findings from our study indicate that the mediate severity of accidents has a high frequency of occurring rather than severities of very low and high risks, and factors like weather, day of the week, and infrastructure factors have various effects on these accidents. Random Forest algorithm obtained the best performance in predicting these road traffic accidents. Relevant responsible authorities can take unnecessary safety actions to reduce and minimum accidents by considering this information. Moreover, this research can be improved more by focusing on worldwide accident records.

REFERENCES

[1] "Road traffic injuries." https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries (accessed Nov. 03, 2021).

[2] "What is data visualization and why is it important?," *OctopusBI*, Mar. 02, 2021. https://octopusbi.com/what-is-data-visualization-and-why-is-it-important/ (accessed Nov. 03, 2021).

[3] S. Vasavi, "Extracting Hidden Patterns Within Road Accident Data Using Machine Learning Techniques," in *Information and Communication Technology*, Singapore, 2018, pp. 13–22. doi: 10.1007/978-981-10-5508-9_2.

[4] "Towards Big Data Analytics and Mining for UK Traffic Accident Analysis, Visualization & Prediction | Proceedings of the 2020 12th International Conference on Machine Learning and Computing." https://dl.acm.org/doi/abs/10.1145/3383972.3384034?casa_token=W9agbUitjPYAAAAA:nO-Ir-uwc-9RTqZ9BhNVnsFw7FHxd4byt0e8csGhQjLu0t5o4MeEELexyPS-FIJIURqaB64pHE8- (accessed Nov. 03, 2021).

[5] F. Babič and K. Zuskáčová, "Descriptive and predictive mining on road accidents data," in *2016 IEEE 14th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, Jan. 2016, pp. 87–92. doi: 10.1109/SAMI.2016.7422987.

[6] M. F. Musa, S. A. Hassan, and N. Mashros, "The impact of roadway conditions towards accident severity on federal roads in Malaysia," *PLOS ONE*, vol. 15, no. 7, p. e0235564, Jul. 2020, doi: 10.1371/journal.pone.0235564.

[7] I. O. Asare and A. C. Mensah, "Crash severity modelling using ordinal logistic regression approach," *Int. J. Inj. Contr. Saf. Promot.*, vol. 27, no. 4, pp. 412–419, Oct. 2020, doi: 10.1080/17457300.2020.1790615.

[8] C. T. Danthanarayana and S. N. Mallikahewa, "An Analysis of the Enduring Factors of Road Traffic Accidents in Sri Lanka," *Sri Lanka Journal of Economic Research,* vol. 8(2), pp. 39-50, March 2021, doi: http://doi.org/10.4038/sljer.v8i2.136.

[9] C. C. Ihueze and U. O. Onwurah, "Road traffic accidents prediction modelling: An analysis of Anambra State, Nigeria," *Accid. Anal. Prev.*, vol. 112, pp. 21–29, Mar. 2018, doi: 10.1016/j.aap.2017.12.016.

[10] T. K. Bahiru, D. Kumar Singh, and E. A. Tessfaw, "Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity," in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Apr. 2018, pp. 1655–1660. doi: 10.1109/ICICCT.2018.8473265.

[11] "US Accidents (updated)." https://kaggle.com/sobhanmoosavi/us-accidents (accessed Nov. 04, 2021).

[12] M. Kuhn and K. Johnson, "Data Pre-processing," in *Applied Predictive Modeling*, M. Kuhn and K. Johnson, Eds. New York, NY: Springer, 2013, pp. 27–59. doi: 10.1007/978-1-4614-6849-3_3.

[13] "Definition of Descriptive Analytics - Gartner Information Technology Glossary," *Gartner*. https://www.gartner.com/en/information-technology/glossary/descriptive-analytics (accessed Nov. 04, 2021).

[14] "What Is the Correlation Coefficient?," *Investopedia*. https://www.investopedia.com/terms/c/correlationcoefficient.asp (accessed Nov. 04, 2021).

[15] "Predictive analytics," *Wikipedia*. Oct. 29, 2021. Accessed: Nov. 04, 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Predictive_analytics&oldid=1052544105

[16] "Logistic regression," *Wikipedia*. Oct. 29, 2021. Accessed: Nov. 04, 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Logistic_regression&oldid=1052545558

[17] "*k*-nearest neighbors algorithm," *Wikipedia*. Oct. 24, 2021. Accessed: Nov. 04, 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=K-nearest_neighbors_algorithm&oldid=1051590352

[18] "Introduction to Random Forest in Machine Learning | Engineering Education (EngEd) Program | Section." https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/ (accessed Nov. 04, 2021).

[19] "Decision Tree." https://www.saedsayad.com/decision_tree.htm (accessed Oct. 16, 2021).

[20] "Model Evaluation Techniques|Machine Learning Model Evaluation," *Analytics Vidhya*, May 06, 2021. https://www.analyticsvidhya.com/blog/2021/05/machine-learning-model-evaluation/ (accessed Nov. 04, 2021).