A Framework for Analyzing Road Accidents Using Machine Learning Paradigms

Shweta1, J Yadav2, K Batra3 and A K Goel4
1,2,3 Faculty of Engineering and Technology, SGT University, Haryana, India
4 Galgotias University, Greater Noida, India.

E-mail: shweta7084@gmail.com

Abstract. Road Safety is a matter of great concern throughout the world. As number of casualties is increasing more than 4% annually in all age groups. It has been predicted that due to road accidents causality rate will grow around 8% till 2030. It's entirely admissible and saddening to let citizens get killed in road accidents. As a result, to handle this sort of situation, an in-depth analysis is required. The Data of Road accidents are very heterogeneous in nature so analysis of such type of data is tricky. Segmentation is the main task for analyzing such data. So, K-means clustering method is mainly used for it as proposed in the research work. Second task of this model is to extract the data, images and hidden patterns by using Supervised Machine Learning algorithm that will help to form the policies for the prevention from road accidents. The combination of segmentation machine learning algorithm produces meaningful information.

Keywords: Machine Learning; Clustering; K-means Clustering; Feature Selection; Road Accidents

1. Introduction
In modern times, roads and transport has become an indispensable part of human’s everyday life. Researchers are continuously working to reduce road accidents by using technology. Many factors are responsible for traffic accidents. The cost of these accidents has great impact on Society and Financial eco system of any country [22][23]. According to the report published by World Health Organization [21]

- Approximately 13.5 lakhs causalities every year in road traffic crashes,
- More than 50% of all road traffic deaths are of pedestrians, cyclists, and motorcyclists.
- Death due to Road traffic injuries are more in children and adults between age group 5-29 years.

Road accidents data are the essential measure of safety with the assistance of which we will establish the size and therefore the nature of road safety problems. Therefore, accidents database is the key to think about the management of road safety. The major thing that helps us to find the factors of road accidents is the data set. The country data are not accurate either they are lack of useful information or they have no strong format system. For the analysis, the data set must be accurate. It has been found that common causes of accidents are road way conditions, weather conditions, lack of traffic indication, vehicle problem, driver’s behavior etc. Researchers have used various techniques to analyze the problems like Neural Networks, Fuzzy logic, Data mining and Machine Learning [4][9][12].

Different parameters have different effect on road accidents. The most difficult thing for the analysis is the dissimilarity. Therefore, the segmentation is required. We can measure the dissimilarities with
help of analysis of the given information and finding the connection between these dissimilarities that can help us to find the hidden pattern required for analysis of the road accidents [5][6].

In this research paper, the researchers have used the data mining techniques [6]. Firstly, they have pre-processed the data to find the locations with different frequencies of road accident and then analyzed these data to find the factors that are impacting on that locations. After that, using the K- means clustering algorithm, they divided the accident locations into 3 categories based on frequency count – high, moderate and low. Then, they have used the association rule to find out the connection between the different attributes with the help of which they exhibited that different locations have different accident frequencies. And with this approach they were able to find some hidden information which can be used for prevention of road accidents [8][10]. In this paper, the researchers have proposed the vision-based method to analysis the road traffic data with the help of which they can learn the traffic pattern that help them to extract the velocity of vehicles and its distances. They have used here the PPNN i.e., Parzen Probabilistic Neural Network for machine learning. They concluded that proposed approach shows approximately 85% accuracy in detecting special situation. In this research paper, the researchers have analyzed the situation of occurrence of road accident with the help of machine learning algorithms. They have used different algorithms of machine learning like CART, Naive bayes, ROC value etc. They have concluded that by applying the CART algorithm they achieved 81.5% of accuracy.

In this research paper, the researchers have done analysis to identify the strength of road accidents in Bangladesh by using machine learning algorithms. Here, they had tried to find out the factors which have impact on road accidents and using those factors they tried to give us some beneficial suggestions regarding road accidents. Here, they have used the four supervised machine learning techniques namely Decision Tree, K-Nearest Neighbour (KNN), Naive Bayes and AdaBoost. They have used these algorithms to classify the severity of accidents into four categories that are Fatal, Grievous, Simple Injury and Motor Collision. Amongst these algorithms, AdaBoost algorithm achieved best performance.

Machine learning algorithms have been used to find the hidden pattern and extract the useful information. Clustering can help for grouping and segmentation of road accidents data to find the beneficial suggestions [19].

2. Methodology
The major types of machine learning algorithms are supervised learning, semi- supervised, unsupervised learning and reinforcement learning [8]. Among these approaches unsupervised learning has been utilized in this research paper. Here, the three popular machine learning algorithms have been used namely, Clustering, K-means Clustering and Feature Selection. Figure 1. It represents the working processes of proposed model.

![Figure 1. The flowchart for proposed model](image-url)
2.1. Dataset Preparation

Extensive and accurate data records of road accidents are the foremost important to analyze performance by using machine learning algorithms. But getting an accurate and an ideal data set isn't a simple task, it is quite challenging. Therefore, to process the information supported the necessity we are using two following steps. They are:

2.1.1. Data collection

The most important aspect for any data analysis is the data. And collection of correct data is essential. Special attention is needed for analyzing and understanding the content and structure of the data.

The dataset used is collected from the reports of City of Toronto Police Open Data Portal which represents the Killed or Seriously Injured (KSI) traffic accidents. It contains the information for all traffic accidents reported between 2007 and 2017. The information contained is about time, location and incident type with distinct aspects of traffic conditions at the moment of occurrence of incident.

2.1.2. Data Pre-processing

Data preprocessing is the important step. It is the technique of data mining. Here, we will try to remove the dimensionality of the data by encoding mechanism. The following table 1 contains the attributes name and its description that are used in this research.

| Attribute    | Attribute Description                             |
|--------------|--------------------------------------------------|
| ACCNUM       | The total number of accidents.                   |
| YEAR         | The year in which accidents took place.          |
| MONTH        | Month in which the accidents took place.         |
| DAY          | Day on which the accidents took place            |
| HOUR         | Hour in which the Accident (24hrs)              |
| MINUTES      | Minute in which the Accidents took place.        |
| WEEKDAY      | Weekday on which accidents took place (0 is Monday) |
| LATITUDE     | Latitude                                         |
| LONGITUDE    | Longitude                                        |
| Ward_Name    | City Ward                                        |
| Ward ID      | City Ward ID                                     |
| Hood_Name    | Neighborhood Name                                |

And many more attributes used in the dataset. There are following steps taken in data pre-processing. They are:

a) Data cleaning:

Data Cleaning is the process of clean or removing the noisy, incomplete and unwanted data. It helps us to modify and remove the data which is not useful or relevant for our analysis. This data is not important or helpful for analyzing because it may provide inaccurate result or it can hinder the process. There are different methods for cleaning the data and it depends upon how it is sorted [9].

Here, the data has been cleaned by replacing blank value by NA and dropping columns with large amounts of missing value and by changing its data types like from int to string as per the analysis.

b) Feature Selection:
Working with sizable number of features can affect the performance. Also, it's going to contain the risk of over fitting with the number of features. Therefore, for relevant or appropriate outcomes the feature selection technique is used. Feature selection is basically used to reduce the dimensionality. It allows machine learning algorithms to train the dataset faster.

3. Experimental Setup
In this section, we are going to explain how the dataset has been pre-processed and the algorithm is used to get the result.
In this paper, we are using the data set of the Killed or Seriously Injured (KSI) traffic accident reports from the City of Toronto between 2007 and 2017.

3.1. Data cleaning:
Usually the dataset contains noisy, missing data values, which is going to affect the quality of the result. So, this step is to pre-process the raw set of data to improve the clustering process. Here, data has been cleaned by replacing blank value by NA and dropping columns with large amounts of missing value and by changing its data type.

3.2. Data exploration:
Here we explore the data by analyzing the accident numbers against month and year. Total number of incidents has reduced slightly over the years.

![Figure 2. Number of incidents happened over ten years (2007 to 2017).](image)

![Figure 3. Number of incidents happened month wise.](image)
From the data above, accidents happened more from June to October.

![Graph showing ACCLASS vs Fatal injury](image)

**Figure 4.** ACCLASS vs Fatal injury

Now we will look for the target column FATAL, therefore, we will analyze the ACCLASS vs Fatal injury. ACCLASS is column which is classified into 3 categories while fatal only show whether it is fatal or not.

3.3. **Data Modeling:**
The data is analyzed on the basis of categories like District category, Light Category, Visibility category and Road Surface Condition Category for applying the cluster analysis.

a) **Cluster Analysis using K-means Clustering:**
k-means clustering algorithm is the unsupervised machine learning algorithm that groups the data items into k cluster. It helps to solve the clustering problems in easy way. It helps us to group the data on the basis of similarity as well as keeping it different as far as possible.

b) **Feature Selection:**
The main aim to implement Feature selection and cluster analysis was to minimize dimension and optimize for best modeling score. Feature selection applied on columns such as year, month, accnum, hoodID or name, minute, hour and weekday have major implications to modeling score and relation to fatal injuries. But accnum, hour and minute was not selected for final modeling as these attributes are less significant in reality.

4. **Results & Conclusion**
In this modelling, several graphs were used to explore relations inside the dataset. Column 'FATAL' was chosen as target output. Here we have taken different attributes for analysis and comparing the situations in which

a) Region-wise, Toronto east region was reported to be the region with most accidents ending up in fatal injury as showed in figure.
b) The primary cause of accident over the years has been automobiles involving aggressive driving and pedestrians.

c) AG_Driv (Aggressive and Distracted Driving) is one of the major reasons to give rise to accidents (62.9%), speeding accounts for 21.4%, red-light accounts for 10.4% and Alcohol contributes to 5.3%.
From k-means cluster method, 4 clusters were used to map the hoodname for final modeling to present which region result in more fatal injury and what is the cause behind it. In general, the results are not very significant that can demonstrate big reversal of the cause. But relative more secure zone or risky zone can be distinguished.

In most cases fatal injuries happens due to harsh driving, inattentiveness and speeding. But road with low light or dark majorly contributes in causing issues compared to other regions.

Reference

[1] Hébert A, Guédon T 2019 High-resolution road vehicle collision prediction for the city of montreal 2019 IEEE International Conference on Big Data (Los Angeles, CA; USA) V3 pp. 1804-1813.
[2] Arvindhan M 2019 Clustering algorithm networks test cost sensitive for specialist divisions ICTACT Journal on image and video processing 2008-2102.
[3] Kumeda B, Zhang F 2019 Classification of road traffic accident data using machine learning algorithms 2019 IEEE 11th ICSVV pp. 682-687.
[4] Bruehning E, One D & Pastor C 2005 30 years in-depth accident studies for improving traffic safety Zeitschrift für Verkehrssicherheit 175-181.
[5] Paratasarathy G 2019 Using hybrid data mining algorithm for analysing road accidents data set 2019 3rd ICCT 7-13.
[6] Sakhare A, Kasbe, P 2017 A review on road accident data analysis using data mining techniques ICICIECS (Coimbatore: IEEE) pp. 1-5.
[7] Krause S, Busch F 2019 New insights into road accident analysis through the use of text mining methods 2019 6th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS) pp. 1-6.
[8] Chong M., Abraham A., Paprzycki M., Traffic Accident Data Mining Using Machine Learning Paradigms, Fourth International Conference on Intelligent Systems Design and Applications (ISDA’04), Hungary, ISBN 9637154302, pp. 415-420, 2004.
[9] Pereira F, Rodrigues F 2013 Text analysis in incident duration prediction Transportation Research Part C: Emerging Technologies Volume 37 pp. 177-192.
[10] Pawlowski W, Lasota D, Goniewicz M, et al. The effect of ethyl alcohol upon pedestrian trauma sustained in traffic crashes. International journal of environmental research and public health, (April 2019), 16(8): 1471. DOI = https://doi.org/10.3390/ijerph16081471.
[11] AlMamlook R E, Kwayu K M 2019 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 pp. 272-276.
[12] Kassu, A.; Anderson, M. Determinants of severe injury and fatal traffic accidents on urban and rural highways. Int. J. Traffic Transp. Eng. 2018, 8, 294–308.
[13] Beshah T, Ejigu D 2012 Learning the classification of traffic accident types 2012 Fourth International Conference on Intelligent Networking and Collaborative Systems Bucharest: IEEE pp. 463-468.
[14] Dadashova, B.; Ramirez, B.A.; McWilliams, J.M.; Izquierdo, F.A. The identification of patterns of interurban road accident frequency and severity using road geometry and traffic indicators. Transp. Res. Procedia 2016, 14, 4122–4129. [CrossRef]
[15] 2020, January 20 world health organization. Retrieved march 31, 2020, from who.int: https://www.who.int/gho/road_safety/mortality/number_text/en/.
[16] Bülbul H I, Kaya T & Tulgar Y 2016 Analysis for status of the road accident occurrence and determination of the risk of accident by machine learning in Istanbul 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA) Anaheim, CA pp. 426-430.
[17] Elahi M M L, Yasir R, Synus M A, Nine M S Q Z, Hossain I & Ahmed N 2014 Computer vision based road traffic accident and anomaly detection in the context of Bangladesh 2014 International Conference on Informatics, Electronics & Vision (ICIEV) Dhaka pp. 1-6.
[18] Kumar S, Toshniwal D 2016 A data mining approach to characterize road accident locations Journal of Modern Transportation 24 62–72.
[19] Labib M F, Rifat A S, Hossain M M, Das A K and Nawrine F 2019 Road accident analysis and prediction of accident severity by using machine learning in Bangladesh 2019 7th International Conference on Smart Computing & Communications (ICSCC), Sarawak, Malaysia pp. 1-5.
[20] Ling H 2014 Computational Intelligence in Industrial Application. (CIIA 2014), Singapore.
[21] WHO/T. Pietrasik February 7, 2020 https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.
[22] Chong M, Abraham A, Paprzycki M 2004 Traffic accident data mining using machine learning paradigms Fourth International Conference on Intelligent Systems Design and Applications (ISDA’04), Hungary, ISBN 9637154302, pp. 415-420.
[23] Chong M, Abraham A, Paprzycki M 2004 Traffic accident analysis using decision trees and neural networks IADIS International Conference on Applied Computing, Portugal, IADIS Press, NunoGuimarães and Pedro Isilas (Eds.), ISBN-9729894736, Volume 2, pp. 39-42.