International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 01 | Jan 2021 www.irjet.net p-ISSN: 2395-0072

Road Accident and Emergency Management: A Data Analytics Approach

Bhavani Shanmugam[1], Mafas Raheem[2], Nowshath K Batcha[3]

[1,2,3]*School of Computing, Asia Pacific University of Technology & Innovation, Kuala Lumpur, Malaysia*

Abstract - *Finding preventable reasons for road accidents to improve emergency services is in demand these days. Academics and road safety authorities have raised concerns on the contributing factors in this regard. This paper investigated the road accidents and its severity via descriptive and predictive analytical approach to draft valid pieces of recommendations in the domain of emergency management. From the findings of the descriptive analytics, it could be noted that the factors such as accident clearance duration, accident occurring time zones, accident-prone states, proximity to traffic objects and days of the week which get more accidents seem crucial for the authorities who involve in the emergency management concerning the road accidents. Further, the predictive model named Random Forest could be effectively used in predicting the severity of road accidents with 94% accuracy.*

Key Words: **road accident, emergency management, accident analytics, predictive modelling, model tuning**

1. INTRODUCTION

A circumstance that puts health, life, property or environment at immediate danger is an emergency. People happen to be involved in various emergencies such as road accidents, fire, robbery, earthquakes, droughts, tornadoes and many more daily. Most emergencies necessitate immediate means to prevent the worst scenario, but mitigation might not be possible in some circumstances and authorities may only be able to provide emergency care aftermath. People tend to panic when they are involved in an emergency, and mostly fail to take the necessary actions where the role of the Emergency Medical Services (EMS) is understood as crucial.

Emergency management comprises five main phases like prevention, preparedness, response, recovery and mitigation [1]. Prevention focuses on the preventive measures designed to avoid the occurrence of a disaster. In this research paper, the prevention of road is primarily focused to minimize the risk of loss of life and injury caused by road accidents. Preparedness is a continuous cycle of planning, organizing, and training, equipping and taking corrective action to respond to any emergencies. The response phase focuses on the reaction after an emergency followed by the recovery phase which is carried out right after the threat to human life has diminished. The final phase of emergency management is the mitigation phase which consists of structural and non-structural steps taken to restrict the aftereffect of emergencies.

Every year, there is an increasing number of vehicles on the road thus the road cannot accommodate such high volume and this causes congestion. Besides, the migration of people towards urban also contributes as it increases the number of vehicles on the city roads. Furthermore, improper town planning also makes the problem worse as the projects did not have much long term visions on the town & country planning especially in designing the road.

1.1 Problem Statement

Road accidents usually cause high mortality, severe injuries, and considerably high economic losses. 12% of the total casualties & diseases and high rates of unintentional injuries are caused by the road accidents [2]. Annually, 1.2 million deaths and more than 50 million injuries occur due to road accidents [2]. Most of the road accidents occur due to the geographic, demographic, environmental and individual factors. Hence, the prevention and control of context-specific accidents are very crucial where access to the information about road accidents in a given context is significant.

Studies concluded that accident hazards are considerably amplified during snowy and icy road conditions. According to [3] the accident hazard was four times greater for snowy and icy roads compared to normal roads. It was also concluded that 5–20% (depending on the month) of the accidents were recorded during rail fall [4]. The corresponding risk for fatal accidents was fivefold for slushy roads and the number of accidents per vehicle mileage occurring in specific conditions is also important to understand the severity of risk [3].

High-quality insights are needed to find preventable causes of road accidents to improve emergency services. Concerns have been raised by academics and road safety authorities over the reliability of police-reported contributing factors but there has been little or no empirical attempt to investigate this issue [5].

1.2 Aim

This research aims to bid recommendations concerning emergency management via descriptive and predictive analytics.

1.3 Objectives

- To perform descriptive analytics to identify and derive the prime factors affecting road accidents.

- To build a predictive model which could predict the severity of road accidents.
- To draft valid recommendations that provide suitable information to the relevant emergency services.

2. LITERATURE REVIEW

2.1 Introduction

Emergency management can be defined as "A discipline that deals with risk and risk avoidance" [6]. Natural hazards are a threat as they possess a threat to human and animal lives, agriculture and infrastructures such as hurricanes, floods, storms, earthquakes, road crashes and many more. Implementing preventive measures would help to reduce the severity of the consequences of a catastrophe and considered as good emergency management. Strong preparedness for emergencies helps to relieve some of the instability caused by the unforeseen tragedy [6].

The study conducted by [7] explores risk factors that contribute to serious road accidents, which would be critical for the prevention of accidents. Tanzania is among the countries with high rates of road crashes where the study was to determine the pattern, associated factors and management of road injury patients in Tanzania [8]. The methods used in this research is a cross-sectional study of victims involved in motor crashes in Tanzania. The outcomes of this research were collected from factors such as demographic profile, circumstances of the injury, nature of injury and factors associated with mortality [8].

Further, the United States of America is also listed as one of the highest road accidents occurring countries where Massachusetts is the state which recorded with the highest road accidents in the recent past [9].

2.2 Data Analytics in Road Accidents

Road accidents are one of the largest national level hazardous in the world. The burden of road accident casualties and damages is commonly higher in developing countries than developed countries where exceptions could be applicable too. Many factors (driver, environment, vehicle, etc.) are related to accidents, some of those factors are more important in determining the accident severity than others [10]. Factors such as weather condition, type of vehicle, drivers' behaviour and personal characteristics like age and gender of the driver play a major role in determining the severity of the accident [11].

An enhanced machine learning model is always critical to investigate the complex relationships between roads and highways, traffic, elements of the ecosystem and crashes [12]. Including the MVNB regression layer in the supervised tuning, the proposed method further accounted for differential propagation patterns in crashes through accident frequency and provide superior crash forecasts. The results imply that the proposed solution is a reasonable alternative for crash forecasts and the overall precision of the forecast calculated by RMSD can be increased by 84.58% and 158.27% relative to the deep learning model without the regression layer and the SVM model, respectively [12].

In a research by [10], the classification algorithm named Random forest was used to reveal relevant patterns and for predicting the type of accident severity into 3 categories as fatal, serious and slight. The best accuracy scores of the Random Forest model was obtained as 78.9%, 62.5% and 49.8% on Hybrid, oversampling and under sampling data sets respectively. The Random Forest model outperformed the other predictive models by selecting the significant attributes to build the model [10]. In contrast, a Naive Bayes model was build using the attributes such as Time, Day, Vehicle type, Sex Type, Weather, Type of Road, and Road Surface. Prediction Model predicted accident severity based on attributes which are Time, Day, Vehicle type, Sex Type, Weather, Type of Road, and Road Surface although with an accuracy of 39.49% [11].

3. METHODOLOGY

3.1 Knowledge discovery in databases

Knowledge discovery in databases (KDD) is a primitive data mining methodology useful to reveal valuable information from a dataset. This methodology involves data preparation, data collection, data cleaning and integration of previous knowledge about datasets and analysis of appropriate solutions from the obtained findings [14]. The KDD process is collaborative and iterative, requiring multiple steps for the user to make choices [15].

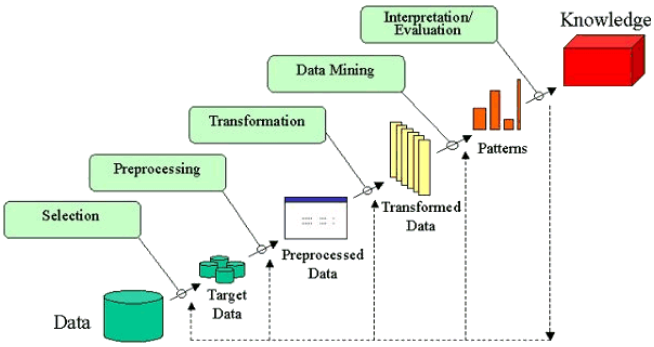

Fig-1: KDD [14]

The road accidents and emergency management has been taken as the domains in this study where a clear understanding of the domain is crucial to recognize the problem background in depth. As stated earlier, the problem is the increasing number of accidents yearly in the USA and find ways to take precautions by identifying the factors affecting the severity of accidents.

3.2 Data

The US accident dataset obtained from Kaggle was used for this study. The dataset consists of 3513616 records with a total of 49 columns and the data was found until June 2020. The dataset contained many errors, missing values, and inaccurate data. This is a very usual problem in any dataset and various techniques to be implemented for preprocessing. Based on the objective of this study, the accident severity

attribute was chosen as the target variable and the suitable input variables were selected to perform both descriptive and predictive analytics.

3.3 Machine Learning Algorithms

The most suitable machine learning algorithms were selected such as Logistic Regression, Gaussian Naïve Bayes, Decision Tree, Random Forest, Support Vector Classifier, and Gradient Boosting to build different predictive models. The section was made based on the literature and the suitability according to the type of the target variable.

3.4 Evaluation measures

Evaluation metrics distinguish the adaptive and non-adaptive machine learning models based on the performance of the algorithm that generates the model results [16]. The predictive validity of the models can be boosted by testing it with various performance evaluation metrics before being implemented in real life [16]. Many problems may occur when a predictive model is deployed before being properly evaluated using many different metrics such as leading to weak predictions that would not be able to provide any benefit to the researcher. The more accurate predictive model was chosen based on the evaluation measures such as accuracy as the target variable is supposed to be balanced before the model building process.

4. DATA ANALYSIS

Data Analysis is the method by which mathematical or logical methods are used to explain and demonstrate, compress, recapture and analyze data. This chapter includes data cleaning, data visualization, and data modelling based on the dataset. Initially, data cleaning was carried out to prepare the dataset ready for further analysis. Next, data visualization was done to the cleaned dataset to get the information and insights at one glance and finally the data modelling was performed.

4.1 Data Preprocessing

The latest dataset of road accidents in the USA was chosen to conduct this study. Preprocessing is the most crucial task as a dataset with missing values and outliers will not be able to generate an accurate result and it may mislead the decision-makers. The dataset was found with different types of noise and cleaned using the suitable preprocessing techniques such as imputing the missing values, converting the attribute format, dropping the variable which is not suitable, handling outliers and encoding categorical variables.

4.2 Descriptive Analytics

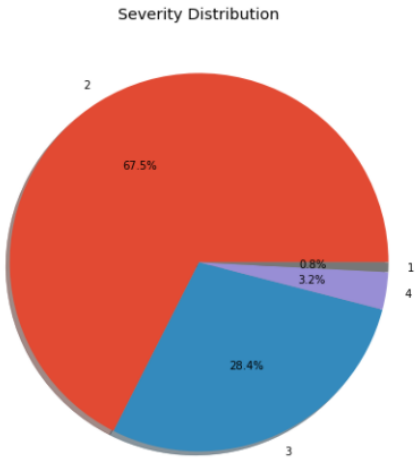

Fig-2: Severity of the accidents

The Fig-2 displays the percentage of the severity of the accidents which is ranked 1 to 4. The severity rank 1 indicates the least impacts of the accident where there will with a short delay as a result of the accident whereas rank 4 indicates a severe impact which is delayed for a longer period. Fig-3 shows the time taken to clear an accident where it displays that most of the accidents were cleared within 30 minutes.

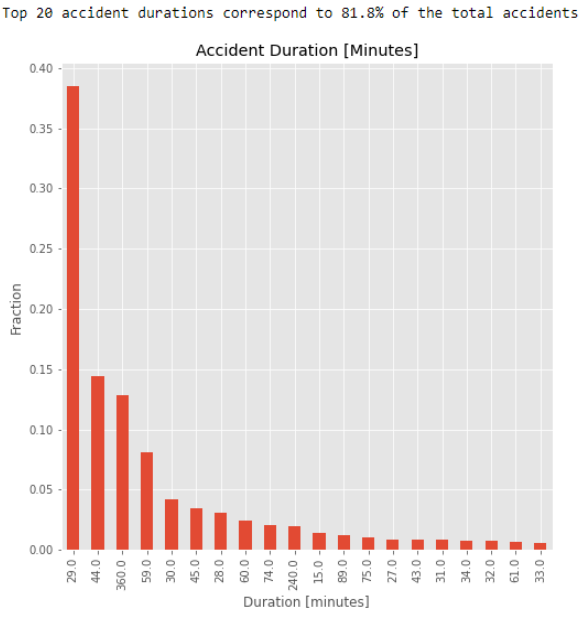

Fig-3: Time to clear the accidents

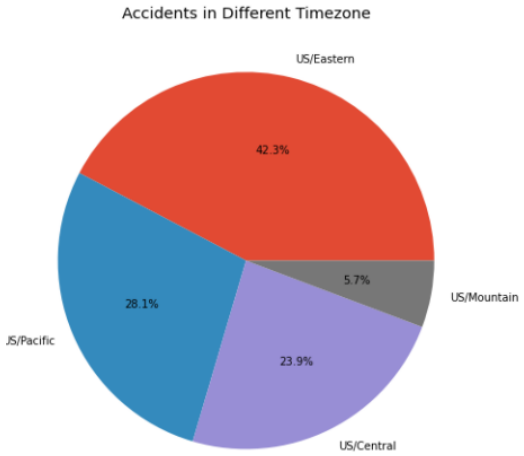

Fig-4: Accidents occurring in different time zones

Fig-4 shows the percentage of accidents occurring in different time zones of the USA namely Pacific, Mountain, Central and Eastern as we move from west to the east region of the country. The figure depicts that a higher percentage of accidents occur in the US Eastern time zone whereas the US Mountain registered the lowest percentage of accidents. States like Massachusetts, Pennsylvania and South Carolina fall in the Eastern time zone.

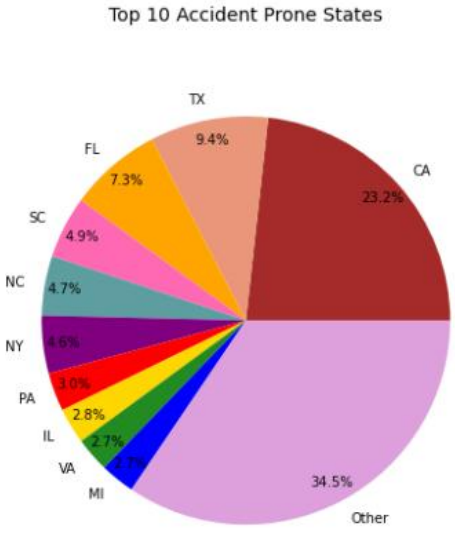

Fig-5: Accidents prone states

The Fig-5 displays the top 10 accident-prone states in the USA as per the dataset. Based on the figure, California (CA) is the most accident-prone state with 22.4 % of accidents followed by Texas (TX) and Florida (FL) with the percentage of 8.3% and 6.3% respectively. Utah (UT) ranked as the least accident-prone state in the US with 2.6 % of accidents.

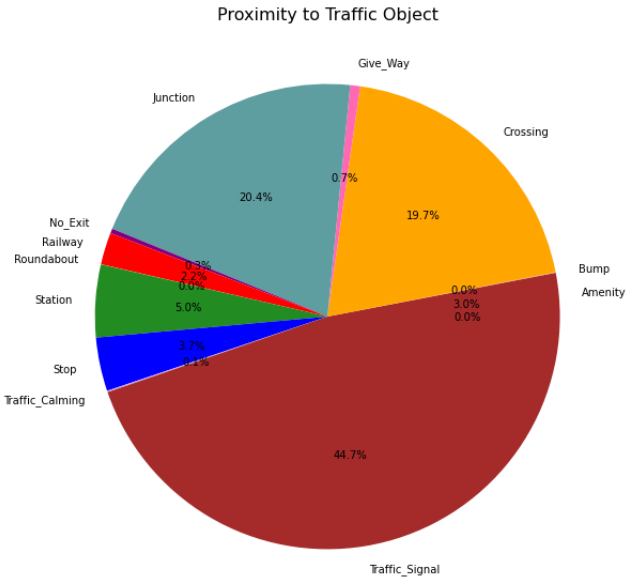

Fig-6: Proximity of Objects

Fig-6 shows the proximity to traffic objects. The factor signal shows the highest proximity to object with the percentage of 45.1% followed by crossing and junction with an equal percentage of proximity 19.6%. The calming and bump recorded the least proximity to objects with a percentage of 0.01%. Roundabout does not show any proximity to objects.

Fig-7 shows the number of accidents on different days of a week. The number of accidents recorded on Tuesday, Friday and Wednesday were approximately the same which is more than 160000 and comparatively the number of accidents recorded on Sunday was the lowest which is less than 60000. The overall trend shows the number of accidents recorded is higher on weekdays compared to weekends.

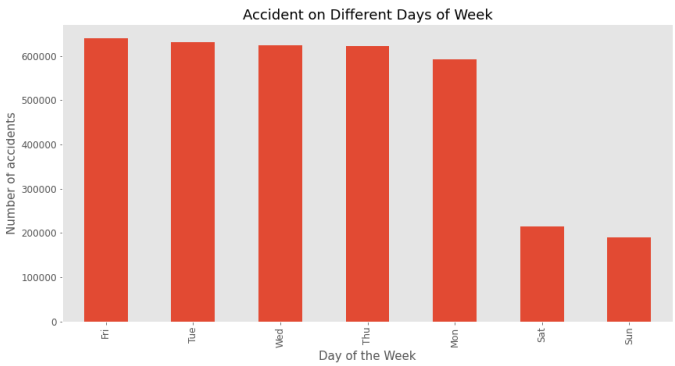

Fig-7: Number of accidents on different days

From the descriptive analytics, it can be understood that the factors such as accident clearance duration, accident occurring time zones, accident-prone states, proximity to traffic objects and days of the week which get more accidents seem crucial for the authorities who involve in the emergency management on the road accidents. These factors better are taken into consideration and set proper measures in aligning it to prevent major losses from road accidents.

4.3 Predictive Analytics

A predictive model suitable to predict the severity of the road accidents based on the set of inputs would always be useful to the relevant emergency management authorities. A series of predictive machine learning models were constructed and a robust predictive model was selected based on the accuracy values. However, the total number of records were not taken into consideration due to the limitation in the personal computational resources. In this line, the data were filtered based on a state named Massachusetts as the state is known for the most number of road accidents as per the recent survey [9].

After completing all the required preprocessing and filtering, the class balancing was done as the target variable was found imbalance.

Class Balancing

As shown in Fig-8, the distribution across the known classes is biased or skewed and may vary from a slight bias to a severe imbalance. A predictive model building using imbalanced data may pose a challenge as most of the machine learning algorithms are designed with the assumption of an equal class.

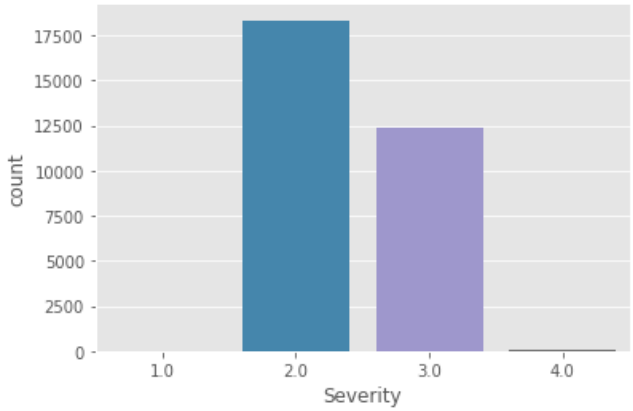

Fig-8: Class distribution before class balancing

An oversampling technique using SMOTE (Synthetic Minority Over-sampling Technique) was applied to perform class balancing where the minority class examples were balanced with the majority class as shown in Fig-9.

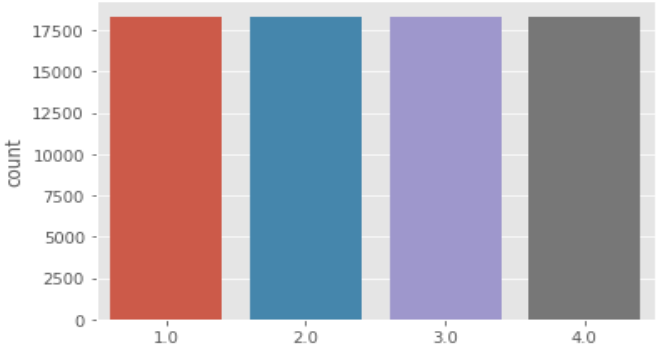

Fig-9: Class distribution after class balancing

Data split

The data was split into a train (80%) and test (20%) as splitting the data is an important task in any predictive model construction process.

Optimisation/Model tuning

The hyper-parameters of the machine learning algorithms are the model's external structure which plays a major role in adjusting the behaviour of a predictive machine learning model (Paul, 2018). The model parameters are set manually and obtained from the model construction and used to evaluate the performance of the models.

The hyper-parameters of the selected machine learning algorithms as listed below were tuned accordingly:

- Logistic Regression: Random State.
- Gaussian Naïve Bayes: var_smoothing.
- Decision Tree: max_depth of the tree and criterion either 'gini' or 'entropy'.
- Random Forest: n_estimators which are the number of trees you want to build before taking the maximum voting or averages of predictions.
- Support Vector Machine: kernel, gamma and C
- Gradient Boosting: n_estimators and learning_rate

The above-stated hyper-parameters were tuned using the grid search method along with cross-validation. Grid search is a hyperparameter tuning technique with in-built cross-validation that systematically builds and tests a model for every algorithm parameter variation defined in the grid [16].

4.4 Predictive models

A certain number of predictive models were built using more suitable machine learning algorithms. The models were trained and tuned using the Grid Search Cross-Validation technique to build the model with the more profitable hyper-parameters. The selected hyper-parameters for the respective models are tabulated in Table-1.

Table-1: Selected Hyper-Parameters

Model	Hyper-Parameters	Type
Logistic Regression	multi_class = 'ovr' solver = 'liblinear' random_state = 0	Traditional Machine Learning Algorithms
Gaussian Naïve Bayes	var_smoothing = 3.511e-07	
Decision Tree	criterion = 'entropy' max_depth = 12	
Support Vector Machine	C = 100 Gamma = 0.001 Kernel = 'rbf'	
Random Forest	n_estimators = 100	Machine Learning Algo
Gradient Boosting	learning_rate = 0.3 n_estimators = 100	

The predictive models were tunes/optimized according to the tabulated hype-parameters and the accuracy scores were recorded and compared as shown in Fig-10.

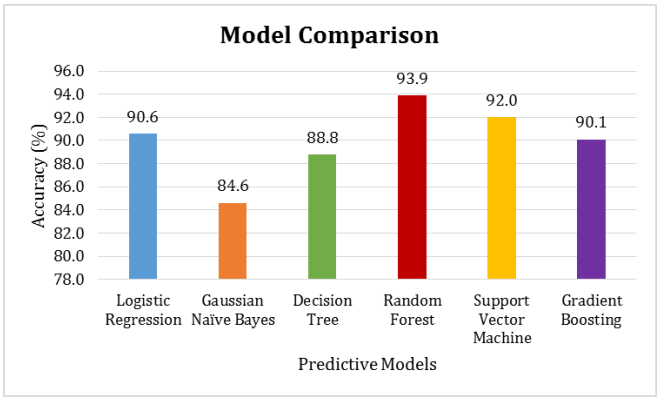

Fig-10: Model comparison using accuracy scores

According to Fig-10, Random Forest was outperforming than the other models based on the accuracy scores which is 93.9%. Further, the Random Forest model was also been used to detect the important features which could be used to build a better model as the initial number of features were 49 and 33 at last after completing all the preprocessing activities. The important features are shown in Fig-11.

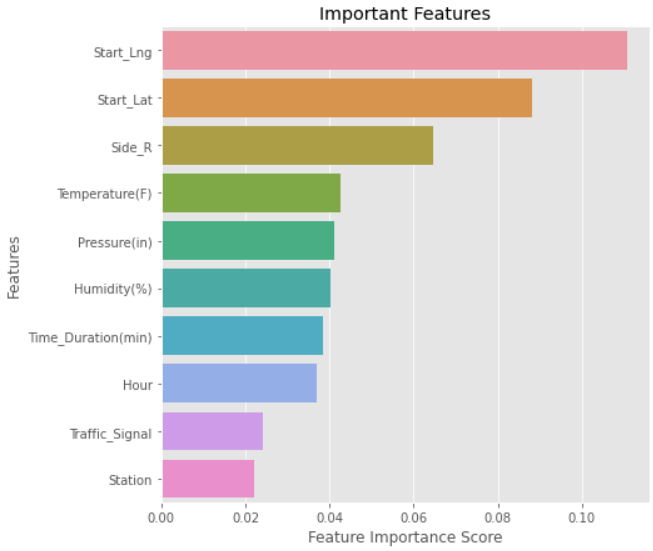

Fig-11: Important Features (top 10)

5. RECOMMENDATIONS

Emergency Medical Services (EMS) is a system that provides emergency medical care. The EMS is activated by incidents that cause serious illness or injuries and focuses completely on the medical care of the disaster victims. The main purpose of EMS is to provide immediate medical care to the needy and to be able to provide a better quality of life. Multiple people and agencies are involved in this system of coordinated response.

The factors such as accident clearance duration, accident occurring time zones, accident-prone states, proximity to traffic objects and days of the week which get more accidents seem crucial for the authorities who involve in the emergency management with regard to the road accidents. The relevant authorities need to take note to implement suitable measures such as educate the drivers on the safety aspects, deploy more emergency workers on the specific days, assign dedicated & rapid communication services to get the reporting of the accident. Also, considering the factors listed by the Random Forest model will be more useful in setting a comprehensive plan to reduce the number of accidents and to recover the victims too.

Accordingly, an emergency response time is the amount of time that the EMS team takes to arrive at the scene of an incident from the time that the emergency response system was activated. Response time is usually used as a proxy for the effectiveness and efficiency of an emergency response program. A long response time might result in increased damage, a higher likelihood of fatalities and distress to those involved.

Further, the Emergency Medical Services comprises 6 general principles such as early detection, early reporting, early response, good on-scene care, care in transit and transfer to definitive care. Early detection of the incident is very important as it is the first step to recognize that there's a problem and it is required to seek help. Early detection of the incident leads to early reporting. To make a report of an emergency the public has to be aware of the emergency number of their country. When the EMS team is well aware of the situation and patient's location, they will be able to head over quickly to the emergency spot to give medical assistance as soon as possible.

The next principle, good on-scene care makes sure that the EMS team which responded has to be able to provide proper first aid to prevent further damage from developing. Care in transit enables the EMS team to transport the patient to the nearest hospital and then the patient is handed over to the emergency facility of the hospital.

6. CONCLUSIONS

As the aim of this research is to bid recommendations concerning emergency management via descriptive and predictive analytics, explicit pieces of recommendation were drafted and presented via this study. It could be noted that a list of factors identified from the findings (descriptive and predictive) was crucial when implementing accident prevention plans in future. Further, the predictive model named Random Forest could be effectively used in predicting the severity of road accidents with 94% accuracy which is far better model than the ones found in the literature.

More features can be added and new approaches can be implemented as future work. Instead of using historical data, time series analysis can be used to get predictions for road accidents. Deep learning architecture can also be used to build more robust predictive models and to get more accurate results. Many key aspects have been captured and would be good use in future implementation as per the real-world demand.

REFERENCES

[1] Bexar.org, "The Five Phases of Emergency Management," [online] Available at: https://www.bexar.org/694/Five-Phases, 2020. [Accessed 14 Oct. 2020].

[2] M. Parvareh, A. Karimi, S. Rezaei, A. Woldemichael, S. Nili, B. Nouri and N. E. Nasab, "Assessment and prediction of road accident injuries trend using time-series models in Kurdistan," Burns Trauma, Mar. 2018, doi: 10.1186/s41038-018-0111-6.

[3] R. Salli, M. Lintusaari, H. Tiikkaja and M. Pöllänen, "Keliolosuhteet ja henkilöautoliikenteen riskit [Wintertime road conditions and accident risks in passenger car traffic]," Tampere University of Technology, Department of Business Information Management and Logistics, Apr. 2008.

[4] F. Malin, I. Norros, & S. Innamma, "Accident risk of road and weather conditions on different road types," Accident Analysis & Prevention, 122. 2008, pp. 181-188.

[5] J. J. Rolison, S. Regev, S. Moutari, A. Feeney, "What are the factors that contribute to road accidents? An assessment of law enforcement views, ordinary drivers' opinions, and road accident records," Accident Analysis & Prevention, 115, 2018, pp. 11-24.

[6] G. Haddow, J. Bullock and D. P. Coppola, "Introduction to emergency management," 5th Edition. Waltham, MA: Butterworth-Heinemann, 2013.

[7] G. Liu, S. Chen, Z. Zeng, H. Cui, Y. Fang, D. Gu, Z. Yin and Z. Wang, "Risk factors for extremely serious road accidents: Results from national Road Accident Statistical Annual Report of China," Plos One, [online] 1(null), 2019, pp. 2-11. Available at: https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0201587&type=printable [Accessed 13 Oct. 2020].

[8] R. Boniface, L. Museru, O. Kiloloma and V. Munthali, "Factors associated with road traffic injuries in Tanzania," 2019, [online] Panafrican-med-journal.com. Available at: http://www.panafrican-med-journal.com/content/article/23/46/full/ [Accessed 13 Oct. 2020].

[9] Insurify Insights, "Eyes on the Road: States with the Most Car Accidents," [online] Available at: https://insurify.com/insights/states-with-most-car-accidents-2020/, 2020, [Accessed 13 December 2020].

[10] S. Ramya, S. K. Reshma, V. Manogna, Y. Saroja, Gandhi and Gaurav, "Accident Severity Prediction Using Data Mining Methods," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 2019, pp. 528-536.

[11] W. Budiawan, S. Saptadi, Sriyanto, C. Tjioe, and T. Phommachak, "Traffic Accident Severity Prediction Using Naive Bayes Algorithm - A Case Study of Semarang Toll Road," IOP Conference Series: Materials Science and Engineering, 598, 2019.

[12] C. Dong, C. Shao and J. Li, Z. Xiong, "An Improved deep learning model for traffic crash prediction," Journal of Advanced Transportation, 2018, 10.1155/2018/3869106.

[13] U. Fayyad, "Knowledge discovery in databases: An overview," International Conference on Inductive Logic Programming, 2005, pp. 1-16.

[14] R. J. Brachman, and T. Anand, "The process of knowledge discovery in databases. Advances in Knowledge Discovery and Data Mining," American Association for Artificial Intelligence, 1996, pp. 37–57.

[15] B. Singh, "Evaluation Metrics for Machine Learning Models," 2020, [online] Available at: https://heartbeat.fritz.ai/evaluation-metrics-for-machine-learning-models-d42138496366 [Accessed 3 November 2020].