INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

ROAD ACCIDENT PREDICTION USING MACHINE LEARNING ALGORITHM

K.Venkat Rao, CH.Damodaram, M.Baji babu,T.Bhavani,V.Lokeshwari
Associate Professor, Priyadarshini Institute of Technology & Science, AP, India.
Under Graduate, Priyadarshini Institute of Technology & Science, AP, India.

Abstract:

The traffic has been transformed into the difficult structure in points of designing and managing by the reason of increasing number of vehicles. This situation has discovered road accidents problem, influenced public health and country economy and done the studies on solution of the problem. Large calibrated data agglomerations have increased by the reasons of the technological improvements and data storage with low cost. Arising the need of accession to information from this large calibrated data obtained the corner stone of the data mining. In this study, assignment of the most compatible machine learning classification techniques for road accidents estimation by data mining has been intended.

1 INTRODUCTION

Road accidents have proved to be one of the leading causes of severe injury and has been on the increase over the years. With almost double the number of vehicles on the road compared to a few years ago, road accidents have been at an all time high; thus taking a huge toll on health, finance and property. Although various laws and safety measures have come into effect, there is always a probability of an accident occurring due to a variety of reasons. Driver neglect, driver recklessness, road conditions, weather conditions, driving skill and a number of other factors influence the safety of both the vehicle and the surroundings.Road accident reports in the UK suggest that driver error has been the leading cause of vehicle collision, with the driver failing to look at his surroundings properly. Driver misjudging distance and speed of both same side and oncoming traffic has found to be a close second cause of accidents with about 80% of these collisions occurring on the same side of the road. Driving with poor maneuvering skills, low visibility, loss of control and driving on slippery surfaces also majorly contributed to the occurence of these accidents. With close to about 50000 cases having been reported in the year 2018, a vast majority of these accidents could have been avoided if the driver took the required precautions while on the road.

One of the most complicated and difficult daily needs is overland transportation. In India, more than 150,000 people are killed each year in traffic accidents. That's about 400 fatalities a day and far higher than developed auto markets like the US, which in 2016 logged about 40,000. Every year over 1 million vehicles are added to traffic averagely. 1.2 million People have died and over 50 million people have been injured in road accidents in the world every year. Studies on traffic have executed that road accidents and death- laceration ratio will increase.

Design and control of traffic by advanced systems come in view as the important need. Assumption on the risks in traffic and the regulations and interventions in the end of these assumptions will reduce the road accidents. An assumption system which will be prepared with available data and new risks will be advantageous. Data mining concept had been come up with by increasing and storage of data in the digital stage. Data mining involves the studies which will discover information from systematic and purposeful data structures obtained from disordered and meaningless data. Machine learning which is sub-branch of artificial intelligence supplies learning of computer taking advantage of data warehouses. Assumption abilities of computer systems have advanced in the event of machine learning. Utilization of machine learning is a widespread and functional method for taking authentic decisions by using information from data and use statistical method.

The costs of fatalities and injuries due to traffic accidents have a great impact on the society. In recent years, researchers have paid increasing attention to determining factors that significantly affect severity of driver injuries caused by traffic accidents [29][30].

There are several approaches that researchers have employed to study this problem. These include neural network, nesting logic formulation, log-linear model, fuzzy ART maps and so on. Applying data mining techniques to model traffic accident data records can help to understand the characteristics of drivers' behaviour, roadway condition and weather condition that were causally connected with different injury severity. This can help decision makers to formulate better traffic safety control policies. Roh et al. [22] illustrated how statistical methods based on directed graphs, constructed over data for the recent period, may be useful in modelling traffic fatalities by comparing models specified using directed graphs to a model, based on out-of-sample forecasts, originally developed by Peltzman [23]. The directed graphs model outperformed Peltzman's model in root mean squared forecast error. Ossenbruggen et al. [24] used a logistic regression model to identify statistically significant factors that predict the probabilities of crashes and injury crashes aiming at using these models to perform a risk assessment of a given region.

These models were functions of factors that describe a site by its land use activity, roadside design, use of traffic control devices and traffic exposure. Their study illustrated that village sites are less hazardous than residential and shopping sites. Abdalla et al. [25] studied the relationship between casualty frequencies and the distance of the accidents from the zones of residence. As might have been anticipated, the casualty frequencies were higher nearer to the zones of residence, possibly due to higher exposure. The study revealed that the casualty rates amongst residents from areas classified as relatively deprived were significantly higher than those from relatively affluent areas. Miaou et al. [26] studied the statistical properties of four regression models: two conventional linear regression models and two Poisson regression models in terms of their ability to model vehicle accidents and highway geometric design relationships. Roadway and truck accident data from the Highway Safety Information System (HSIS) have been employed to illustrate the use and the limitations of these models. It was demonstrated that the conventional linear regression models lack the distributional property to describe adequately random, discrete, nonnegative, and typically sporadic vehicle accident events on the road. The Poisson regression models, on the other hand, possess most of the desirable statistical properties in developing the relationships. Abdelwahab et al. studied the 1997 accident data for the Central Florida area [2]. The analysis focused on vehicle accidents that occurred at signalized intersections. The injury severity was divided into three classes: no injury, possible injury and disabling injury.

2 RELATED WORK:

Sachin Kumar et al. [1] , used data mining techniques to identify the locations where high frequency accidents are occurred and then analayze them to identify the factors that have an effect on road accidents at that locations. The first task is to divide the accident location into k groups using the k-means clustering algorithm based on road accident frequency counts. Then, association rule mining algorithm applied in order to find out the relationship between distinct attributes which are in accident data set and according to that know the characteristics of locations.

S. Shanthi et al. [2] proposed data mining classification technology based on gender classification, in which RndTree and C4.S use AdaBoost Meta classifier to provide high-precision results. From the Critical Analysis Reporting Environment (CARE) system provided by the Fatal Analysis Reporting System (FARS) used by the training data set.

Tessa K. Anderson et al. [3] proposed a method of identifying high-density accident hotspots, which creates a clustering technique that determines that stochastic indices are more likely to exist in some clusters, and can therefore be compared in time and space. The kernel density estimation tool enables the visualization and manipulation of density-based events as a whole, which in turn is used to create the basic spatial unit of the hotspot clustering method.

The severity of damage occurring during a traffic accident is replicated using the performance of various machine learning paradigms, such as neural networks trained using hybrid learning methods, support vector machines, decision trees, and concurrent mixed models involving decision trees and neural networks. The experimental results show that the hybrid decision tree neural network method is better than the single method in machine learning paradigms.

There have been works in the prediction of accident severity that have used algorithms such as Random Forest, Naive Bayes, linear regression and other methods to predict the severity of accidents. These methods of road traffic accidents have played a major role in setting up precautionary measures along areas that have been classified as danger zones or potential accident sites.

Road Accident Prediction has been done in various countries using a number of algorithms but one of the biggest issues is the fact that there lies a data imbalance. As all the data collected is of the occurrence of an accident but no record of the absence of an accident. Therefore various methods have been used to perform negative sampling. Another issue is that it is difficult to perform road accident analysis for larger areas. All papers have utilised datasets consisting of only a small area or restricted themselves to a few road segments. Accident Risk Prediction based on Driving behaviour Feature using XGboost and Cart uses various parameters of driving behavior and are evaluated using which key features depending on correlation to the occurence of the accident is selected. This ensures that only the required features based on contribution to the accident plays a role in prediction and leaves out the redundant measures that have an indirect role to play in the collision.

Using XGBoost to predict the crash using characteristics of collision, time of the accident and the location of the accident and environmental factors showed to have the most accurate results. For usage of Naive Bayes algorithm it was found that grouping of characteristics into elements such as vehicles, road, human and environment helped get a more accurate result.

They compared the performance of Multi-layered Perceptron (MLP) and Fuzzy ARTMAP, and found that the MLP classification accuracy is higher than the Fuzzy ARTMAP. Levenberg-Marquardt algorithm was used for the MLP training and achieved 65.6 and 60.4 percent classification accuracy for the training and testing phases, respectively. The Fuzzy ARTMAP achieved a classification accuracy of 56.1 percent. Yang et al. used neural network approach to detect safer driving patterns that have less chances of causing death and injury when a car crash occurs. They performed the Cramer's V Coefficient test to identify significant variables that cause injury to reduce the dimensions of the data. Then, they applied data transformation method with a frequency-based scheme to transform categorical codes into numerical values. They used the Critical Analysis Reporting Environment (CARE) system, which was developed at the University of Alabama, using a Backpropagation (BP) neural network. They used the 1997 Alabama interstate alcohol-related data, and further studied the weights on the trained network to obtain a set of controllable cause variables that are likely causing the injury during a crash. The target variable in their study had two classes: injury and non-injury, in which injury class included fatalities. They found that by controlling a single variable (such as the driving speed, or the light conditions) they potentially could reduce fatalities and injuries by up to 40%. Sohn et al. applied data fusion, ensemble and clustering to improve the accuracy of individual classifiers for two categories of severity (bodily injury and property damage) of road traffic accidents.

The individual classifiers used were neural network and decision tree. They applied a clustering algorithm to the dataset to divide it into subsets, and then used each subset of data to train the classifiers. They found that classification based on clustering works better if the variation in observations is relatively large as in Korean road traffic accident data. Mussone et al. used neural networks to analyze vehicle accident that occurred at intersections in Milan, Italy. They chose feed-forward MLP using BP learning. The model had 10 input nodes for eight variables (day or night, traffic flows circulating in the intersection, number of virtual conflict points, number of real conflict points, type of intersection, accident type, road surface condition, and weather conditions). The output node was called an accident index and was calculated as the ratio between the number of accidents for a given intersection and the number of accidents at the most dangerous intersection. Results showed that the highest accident index for running over of pedestrian occurs at non-signalized intersections at nighttime. Dia et al. used real-world data for developing a multi-layered MLP neural network freeway incident detection model. They compared the performance of the neural network model and the incident detection model in operation on Melbourne's freeways. Results showed that neural network model could provide faster and more reliable incident detection over the model that was in operation. They also found that failure to provide speed data at a station could significantly deteriorate model performance within that section of the freeway. Shankar et al. applied a nested logic formulation for estimating accident severity likelihood conditioned on the occurrence of an accident.

They found that there is a greater probability of evident injury or disabling injury/fatality relative to no evident injury if at least one driver did not use a restraint system at the time of the accident. Kim et al. developed a log-linear model to clarify the role of driver characteristics and behaviors in the causal sequence leading to more severe injuries. They found that alcohol or drug use and lack of seat belt use greatly increase the odds of more severe crashes and injuries. Abdel-Aty et al. used the Fatality Analysis Reporting System (FARS) crash databases covering the period of 1975-2000 to analyze the effect of the increasing number of Light Truck Vehicle (LTV) registrations on fatal angle collision trends in the US [1]. They investigated the number of annual fatalities that resulted from angle collisions as well as collision configuration.

Time series modeling results showed that fatalities in angle collisions will increase in the next 10 years, and that they are affected by the expected overall increase of the percentage of LTVs in traffic. Bedard et al. applied a multivariate logistic regression to determine the independent contribution of driver, crash, and vehicle characteristics to drivers' fatality risk [3]. They found that increasing seatbelt use, reducing speed, and reducing the number and severity of driver-side impacts might prevent fatalities. Evanco conducted a multivariate population-based statistical analysis to determine the relationship between fatalities and accident notification times [6]. The analysis demonstrated that accident notification time is an important determinant of the number of fatalities for accidents on rural roadways. Ossiander et al. used Poisson regression to analyze the association between the fatal crash rate (fatal crashes per vehicle mile traveled) and the speed limit increase [13]. They found that the speed limit increase was associated with a higher fatal crash rate and more deaths on freeways in Washington State. Finally, researchers studied the relationship between drivers' age, gender, vehicle mass, impact speed or driving speed measure with fatalities and the results of their work can be found in. This paper investigates application of neural networks, decision trees and a hybrid combination of decision tree and neural network to build models that could predict injury severity. The remaining parts of the paper are organized as follows. In Section 2, more details about the problem and the pre-processing of data to be used are presented, followed, in Section 3, by a short description the different machine learning paradigms used. Performance analysis is presented in Section 4 and finally some discussions and conclusions are given towards the end.

3 METHODOLOGY:

Dataset Description:

Data Cleaning and Data Transformation

After we have selected the dataset. The next step is to clean the data and transform it into the desired format as it is possible the dataset we use may be of different format. It is also possible that we may use multiple datasets from different sources which may be in different file formats. So to use them we need to convert them into the format we want to or the type that type prediction system supports. The reason behind this step is that it is possible that the data set contains the constraints which are not needed by the prediction system and including them makes the system complicated and may extend the processing time. Another reason behind data cleaning is the dataset may contain null value and garbage values too. So the solution to this issue is when the data is transformed the garbage values are replaced. There are many methods to perform that.

Data Processing and Algorithm Implementation

After the data is been cleaned and transformed it's ready to process further. After the data has been cleaned and we have taken the required constraints. We divide the whole dataset int o the two parts that can be either 70-30 or 80-20. The larger portion of the data is for the processing. The algorithm is applied on that part of data. Which helps the algorithm to learn on its own and make prediction for the future data or the unknown data. The algorithm is executed in which we take only the required constraints from the cleaned data. The output of the algorithm is in 'yes' and 'no'. It gives the error rate and the success rate..

Separating Features: In this following step we are going to separate the features which we take to train the model by giving the target value i.e. 1/0 for the particular of features. This dataset are many columns and rows and all numbers of null values will be fulfill in forward fill method and also use the classification algorithm entire dataset. In that classification algorithm we will use Logistic Regression Algorithm The logistic algorithm will make the prediction in terms of percentage, to find accuracy level in percentage and Error percentages. This Algorithm is only for the yes and no type of result or successful and unsuccessful. The equation for combinations of all 15 input variables. The classification algorithm of the entire dataset. In the Road Accident prediction final result is to find the percentage of accident in particular area. Having lower number of features helps the algorithm to converge faster and increases accuracy. In the Road Accident prediction final result is to find the percentage of accident in particular area. Then we apply logistic regression on these features and obtain the least error.

Normalization: Normalization is a very important step while we are dealing with the large values in the features as the higher bit integers will cost high computational power and time. To achieve the efficiency in computation we are going to normalize the data values.

Training and test data: Training data is passed to the MLP classifier to train the model. Test data is used to test the trained model whether it is making correct predictions or not. Past research focused mainly on distinguishing between no-injury and injury (including fatality) classes. We extended the research to possible injury, non-incapacitating injury, incapacitating injury, and fatal injury classes. Our experiments showed that the model for fatal and non-fatal injury performed better than other classes. The ability of predicting fatal and non-fatal injury is very important since drivers' fatality has the highest cost to society economically and socially. It is well known that one of the very important factors causing different injury level is the actual speed that the vehicle was going when the accident happened. Unfortunately, our dataset doesn't provide enough information on the actual speed since speed for 67.68% of the data records' was unknown. If the speed was available, it is extremely likely that it could have helped to improve the performance of models studied in this paper.

4 RESULTS

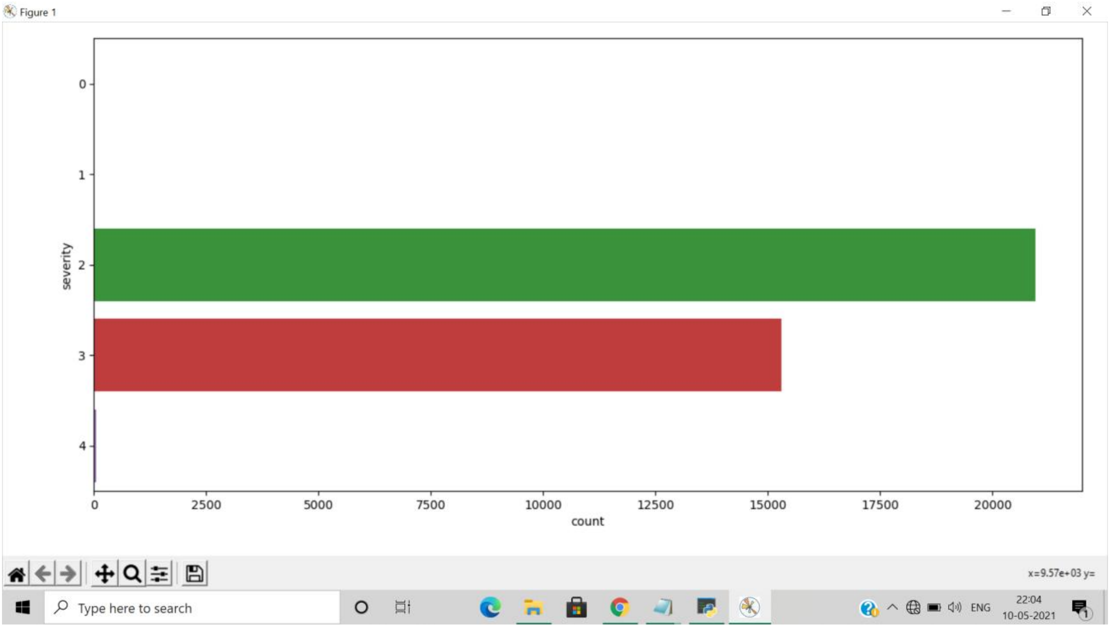

Fig1: Severity of the accidents

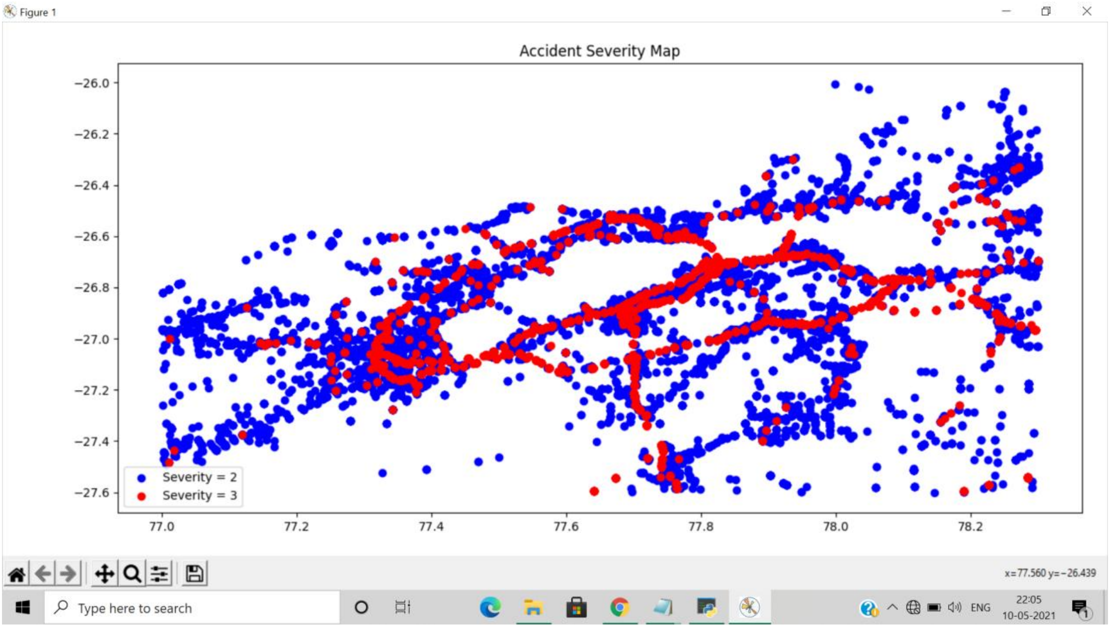

Fig2 : Location wise accidents of the data

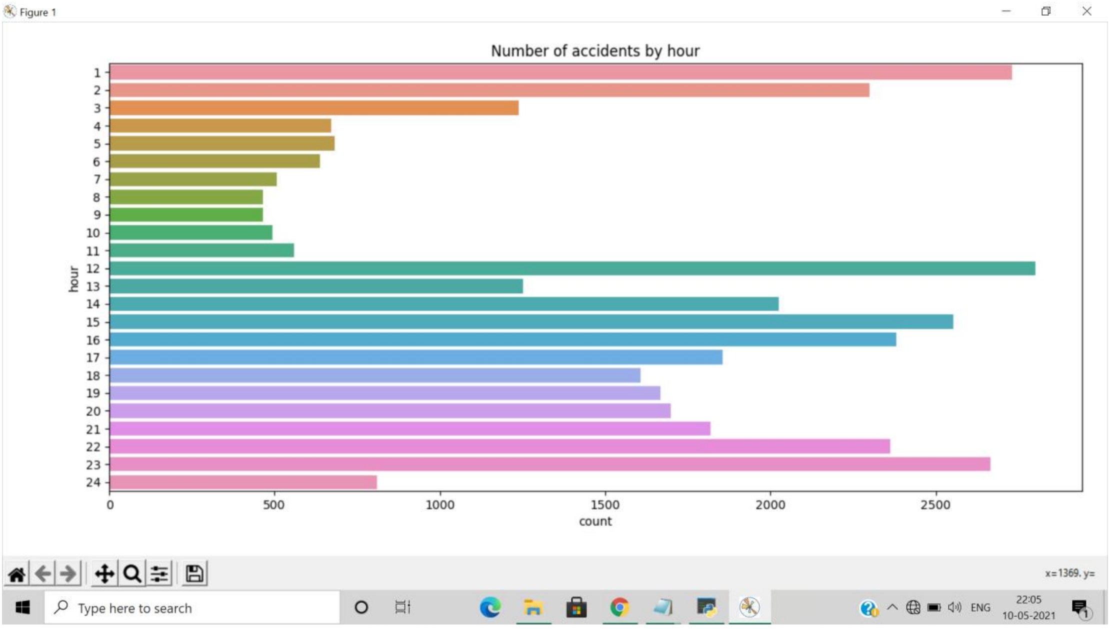

Fig 3: Accidents by hour

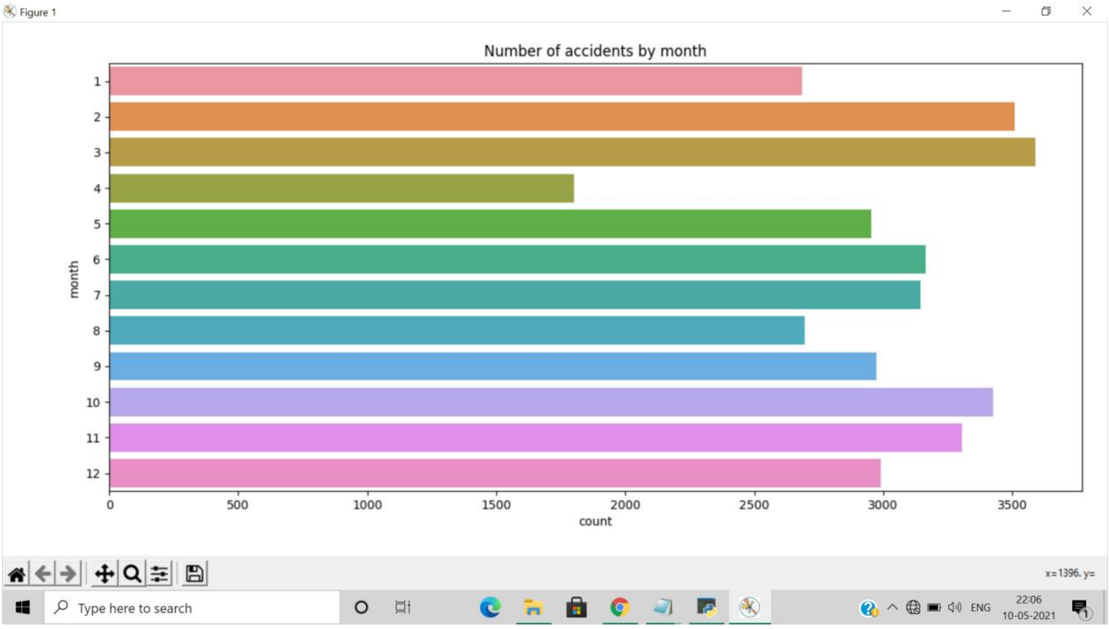

Fig 5: Accidents by Month

5 CONCLUSION:

This paper provides a way to analyse the severity of road accidents and the factors that lead to them. It was observed that factors such as lighting conditions had a high effect on the severity of an accident. Factors like lighting and conditions can be improved upon to make roads safer which can then lead to lower rates of road accidents. Providing a database which contains such a large variety of data such as three classes of accident severity (slight, severe and fatal) and light conditions and details about the police officers at the scene, can be further analysed to provide useful insights and contribute to road safety. Although the occurrence of an accident cannot be controlled, the analysis of this data can enable the government and its citizens to take precautionary steps towards keeping themselves safer.

6 REFERENCE:

[1] Abdel-Aty, M., and Abdelwahab, H., Analysis and Prediction of Traffic Fatalities Resulting From Angle Collisions Including the Effect of Vehicles' Configuration and Compatibility. Accident Analysis and Prevention, 2003.

[2] Abdelwahab, H. T. and Abdel-Aty, M. A., Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections. Transportation Research Record 1746, Paper No. 01-2234.

[3] Bedard, M., Guyatt, G. H., Stones, M. J., & Hireds, J. P., The Independent Contribution of Driver, Crash, and Vehicle Characteristics to Driver Fatalities. Accident analysis and Prevention, Vol. 34, pp. 717-727, 2002.

[4] Buzeman, D. G., Viano, D. C., & Lovsund, P., Car Occupant Safety in Frontal Crashes: A Parameter Study of Vehicle Mass, Impact Speed, and Inherent Vehicle Protection. Accident Analysis and Prevention, Vol. 30, No. 6, pp. 713-722, 1998.

[5] Dia, H., & Rose, G., Development and Evaluation of Neural Network Freeway Incident Detection Models Using Field Data. Transportation Research C, Vol. 5, No. 5, 1997, pp. 313-331.

[6] Evanco, W. M., The Potential Impact of Rural Mayday Systems on Vehicular Crash Fatalities. Accident Analysis and Prevention, Vol. 31, 1999, pp. 455-462.

[7] Hand, D., Mannila, H., & Smyth, P., Principles of Data Mining. The MIT Press, 2001.

[8] Kim, K., Nitz, L., Richardson, J., & Li, L., Personal and Behavioral Predictors of Automobile Crash an Injury Severity. Accident Analysis and Prevention, Vol. 27, No. 4, 1995, pp. 469-481.

[9] Kweon, Y. J., & Kockelman, D. M., Overall Injury Risk to Different Drivers: Combining Exposure, Frequency, and Severity Models. Accident Analysis and Prevention, Vol. 35, 2003, pp. 441-450.

[10] Martin, P. G., Crandall, J. R., & Pilkey, W. D., Injury Trends of Passenger Car Drivers In the USA. Accident Analysis and Prevention, Vol. 32, 2000, pp. 541-557.

[11] Mayhew, D. R., Ferguson, S. A., Desmond, K. J., & Simpson, G. M., Trends In Fatal Crashes Involving Female Drivers, 1975-1998. Accident Analysis and Prevention, Vol. 35, 2003, pp. 407-415.

[12] Mussone, L., Ferrari, A., & Oneta, M., An analysis of urban collisions using an artificial intelligence model. Accident Analysis and Prevention, Vol. 31, 1999, pp. 705-718..

[13] https://web.stanford.edu/class/cs231a/prev_projects_2016/output%20(1).pdf

[14] Prasadu Peddi (2019), "Data Pull out and facts unearthing in biological Databases", International Journal of Techno-Engineering, Vol. 11, issue 1, pp: 25-32.