Accident Risk Prediction based on Machine Learning

Nanditha B, Nidhi Prabhu, Prakruthi M R, Pramatha Nadig H R
Dept. of CSE, Global Academy of Technology, Bangalore - 98, India.
Dr Kavitha K S
Professor, Dept. of CSE, Global Academy of Technology, Bangalore – 98, India.

***Abstract*— As per the Indian Government data approximately 1.51 lakh people lost their life in road accidents in 2019, which means every year approximately 2,32,140 people die in India due to road accident. Indian roads witnessed more accidents on bright sunny days as compared to rainy or foggy days in 2019 according to Indian Government data. The main objective of this paper proposes to decrease the number of road accidents caused due to weather conditions. This system analyzes the amount of risk in terms of percentage present in that particular area based on the present weather conditions and previously collected dataset. This system detects current weather conditions using google weather API. Also the system uses logistic regression algorithm to determine the amount of risk in terms of percentage present in the particular area. The logistic regression algorithm shows a better performance with 90 percent of accuracy.**

Keywords—road accidents, deaths, risk factors, google weather API, logistic regression.

1. INTRODUCTION

Road accident cost India 3-5% of gross domestic product every year and are avoidable if India could improve its road and city planning, and enforce traffic laws properly. 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 also discovered road accidents problem, influenced public health and country economy and done the studies on solution of the problem. 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 experience. Machine learning is able to attain, extract information from data and use statistical method. Design and control of traffic by advanced systems come in view as the important need. Assumption of 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.

1.1 Objectives

The main objectives of the road accident prediction system are:
Analyze the previously occurred accidents in the locality which will help us to determine the most accident prone areas.

- The road accident predictions are made in terms of percentage.
- The predictions are made based on constraints like latitude and longitude, road class, speed limit, weather conditions etc.

1.2 Literature Review

In previous theory, [1] reviewed about road traffic accident which has become the reason due to increase in the number of vehicles of the city which is also caused due to the increase in the population of the people using more number of vehicles. They concluded the factors such as types of vehicles, age of the driver, age of the vehicle, weather condition, road structure and so on, So as per the studies they have added up the prediction which is based on the above factors. On the other hand [2], the data was analyzed using data mining and machine learning techniques which focused on identifying factors that affected the cause of the accident.

According to the paper, some of the internal factors are related to the driver and some external factors that affected are adverse weather condition such as fog, snowfall, rainfall, which caused partial visibility and which would have become difficulty for the drivers. Hence the above study used the technique of regression for a large set of data to identify the reasons of the road accidents which was implemented in the paper. This has helped the government regarding the road safety policy.

Additionally [3] described about the death and injuries caused by the road accident and the severity factors due to which the accident occurred. According to world health organization there were 20 to 50 million people who suffered from injuries and death. This paper aimed at overviewing the factors which was influenced the road traffic accident severity, and also highlights the techniques that are used such as logistic regression, power model etc.. From the analysis the severity of the accident was also from the factors such as speed of the vehicle, human characteristics, vehicle type, vehicle condition, etc.

The accident risk factors are categorized into four division they are: human, road, vehicle, environment. Regarding the factors as stated above many countries are unable to provide adequate information. Therefore very least number of studies proved and accomplished covering all the four factors. To demonstrate the four factors various statistical techniques and methods have been employed from the reviewed paper. Further more from [4], it is evident that due to more number of accidents many people are getting killed by road accidents.

To handle this situation analysis was made. This paper on the whole analysis traffic accidents and the intensity of the accidents using machine learning approaches. KNN, Naive Bayes, Ada Boost algorithm have been used to give out the best performance. It has been essential to control and arrange traffic by using advanced systems to reduce the number of accidents. From the studies using the above algorithms the severity of the road accidents was determined.

In addition to the above studies, [5] states that the improved deep learning model has been proposed to explore the interactions among road ways, environmental elements, traffic crashes and traffic. The proposed model has 2 modules: an unsupervised modules which identifies the functional network between explanatory variables and feature representation. On the other hand supervised module performs traffic crash prediction. Multivariate negative binomial model was being used for embedding supervised module as a regression layer. The proposed model was a superior alternative for traffic crash predictions and the average

accuracy was improved by 84.5% and 158.2% compared to deep learning model without the regression layer.

This study represented an innovative method for traffic crash prediction. It also demonstrated a novel deep learning technique embedded within a multivariate regression mode which was used to identify relationship between examined variables and traffic crashed. The results showed the feature learning module recognized information between input variables and output feature representation.

1.3 Problem Statement

Road Traffic Accidents (RTA) are a major cause of death globally leading to many deaths and to a great extent of injuries every year. Transport authorities world wide have been striving to implement strategies to minimize RTA's. This however is a difficult task- despite the adoption of various regulations and safety measure, RTA's have not decreased significantly. This failure partially stems from the difficulty in predicting when and where RTA's may occur.

1.4 Methodology

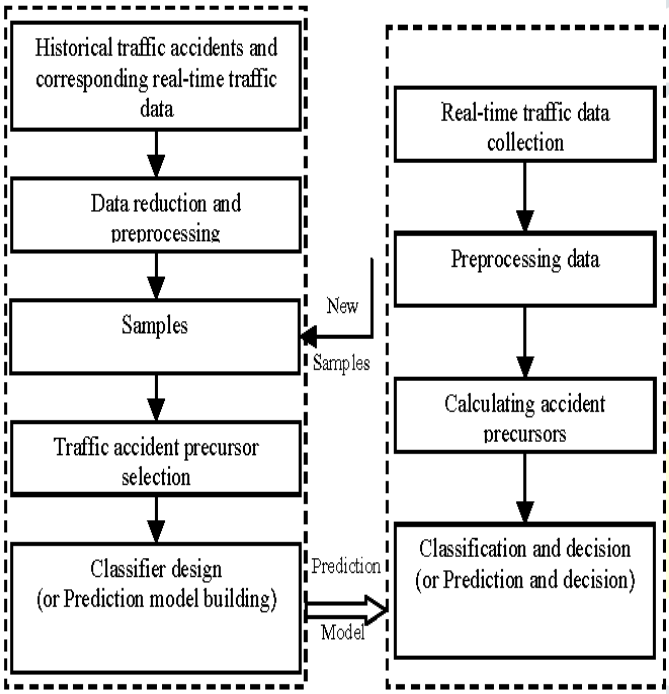

Figure 1: Proposed Architecture

The above proposed model collects the previously occurred traffic accidents and their corresponding real time traffic data of a particular location based on factors like weather conditions, road conditions etc. After these data are collected, the noise information is removed in data reduction and preprocessing step in order to reduce time and the sample data would be prepared.

At this stage the prepared sample will also be classified into train set and test dataset. After the sample is collected the traffic accident precursor selection used to select a specific location and predict the amount of risk present in that specific location. The classifier is designed for making the predictions of the given specific location depending upon the weather conditions of the location and based on the historical dataset. To make predictions the system will learn from the trained dataset of a particular location and then the test dataset predicts from the given trained dataset.

The 2nd stage of the proposed architecture and represents the traffic real time data which the test dataset and as shown in the proposed architecture, the test dataset can be entered by the user input also which would be pre-processed and then based on the given location the test dataset is compared with the trained dataset and the prediction about the risk percentage would be made possible.

2. Dataset description

The dataset has been obtained from Dark Sky API. The dataset contains attributes such as latitude and longitude, road class, temperature, visibility, speed limit etc.. The Dark Sky API allows to look up the weather anywhere on the globe, returning current weather condition. The below table represents the attributes which are used in the proposed project and their description respectively.

Table 1: Attributes and their description

ATTRIBUTES	DESCRIPTION
Day of the week	Represents the particular day of the week on which the accident took place.
Number of days	The total number of days on which the accidents took place.
Road class	Specifies the type of road.
Road number	Represents the address of the road.
Speed limit	Specifies the speed limit during the accident.
Year and Date	Specifies the year and date.
Pressure	Specifies the value of the air pressure.
Weather summary	Gives the summary of the weather condition such as cloudy, sunny, windy etc.
Visibility	Is a measure of distance at which an object or light can be clearly discerned.
Wind bearing	Indicates the direction towards which the object is moving, it is measured in degrees which vary from 0 to 360 degrees.
Wind gust	Is a brief increase in the speed of wind and measured using anemometer
Wind speed	Is a fundamental atmospheric quantity caused by air moving from high to low pressure and is measured in terms of miles per hour.

3. Algorithm

In the proposed system the main criteria is to calculate the amount of risk present in a specific area based on the real time traffic data and historical dataset in terms of percentage. The algorithm that is used in the model to calculate the risk of a specific location in terms of the percentage criteria is logistic regression algorithm. Logistic regression is a regression analysis and dependent upon the variables in binary numbers that is (0's and 1's). Enhancing further the percentage criteria can also be covered using logistic regression.

All regression analysis, the logistic regression is a prediction analysis. Logistic regression is used to find the details about the data and to graphically explain the relationship between dependent binary variables and more nominal, ordinal, interval and independent variables. Using the knowledge of sigmoid function and decision boundary we can write a prediction function. Prediction function in logistic regression returns the probability of the observation being positive or true or 1. We can call this

class as 1

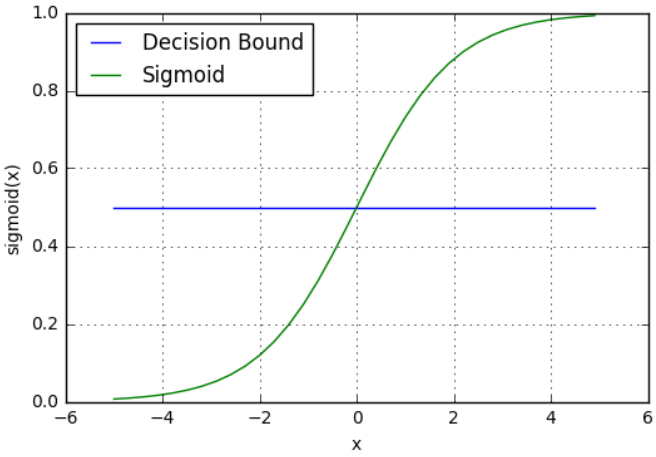

Figure 2: Sigmoid Curve

Transformation of the output can be done using the sigmoid function to return a probability value between 0 or 1.

$$P(\text{class} = 0/1) = 1/1+e^{-(b0+b1*x1+b2*x2\ldots.+bn*xn)}$$

If P>= 0.5, then class 1

Else class 0

4. Module split-up

Data Source:

The previous studies used variety of statistical techniques which applied to the research such as logistic regression, multiple logistic regression, ordered logistic regression, generalized ordered logistic regression, generalize ordered logistic model and binary logistic model. The different techniques of approach can either provide advantages or disadvantages.

Many researchers used logistic regression frequently as a tool to study risk factors that associated with accident severity. The most common factors used in logistic regression found similarly in many literature review papers were: gender; unfit safety status of vehicle; streetlight; bad visibility; driving during weekend; speed limit; traffic volume during non-peak time; and weather condition.

Data Analysis:

Logistic regression model has been used widely for road safety research to find the best fit of model. In the researches logistic regression model technique were used, which was the most frequent model found on literature review papers. Logistic regression is used to obtain odds ratio in the presence of more than one explanatory variable. It is very similar to multiple linear regression except logistic regression is binomial.

5. RESULT

Datasets are the set of values which are used in the project along with their attributes to determine the probability of the accident that might occur based on the prediction made. An example of the dataset is given below in the table along with their respective values which are used in the project.

Sl.No	longitude	latitude	D.O.W	Hour	Spd.Limit
1)	-0.1855	51.48325	5	7	30
2)	-0.1973	51.50888	4	8	30
3)	-0.17345	51.48199	7	23	30
4)	-0.20968	51.51672	7	10	30

Table 2: Datasets

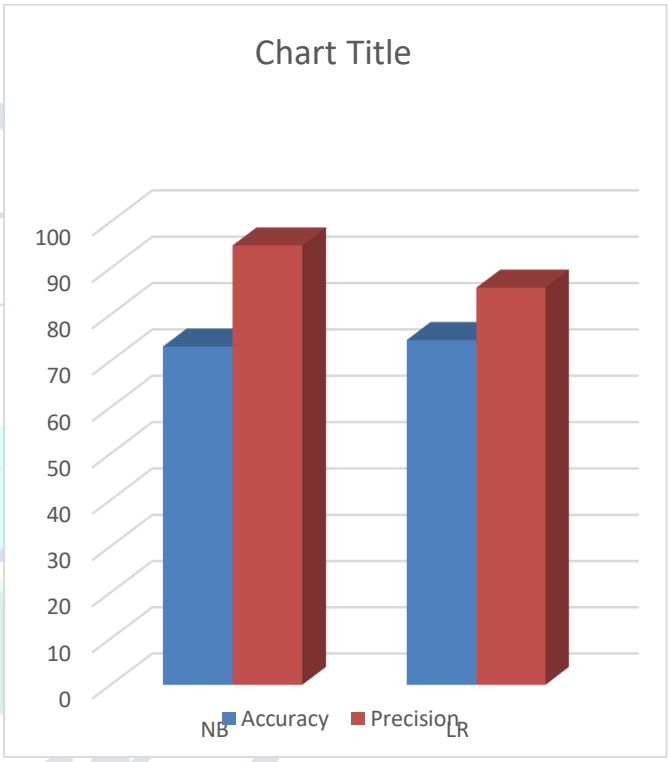

Figure 3: Comparison of data accuracy and precision

The above graph represents the bar graphs through which the suitable algorithms are used to calculate the accuracy of the road accident risk prediction. Here [2] reported Naïve Bayes algorithm in the above graph based on its calculation gives accuracy of 73.1% with 94.9% precision in calculating the amount of risk of accident present in a particular area. Whereas on the other hand logistic regression algorithm gives the accuracy of 74.5% with 85.8% precision in calculating the amount of risk of accident present in an area. And the calculation are based on some of the following attributes as shown above in the table 2.

6. CONCLUSION

Road accidents are caused by various factors. Few factors which affect road accidents are road class, weather condition, speed limit, wind speed etc. Thus we have come up with an application which provides an efficient prediction of road accidents which are caused by the factors as mentioned above. This work can also be implemented in google maps as a future work.

7. FUTURE WORK

Dataset could be enriched with more predictors such as population density, traffic volume, number of shops, number of tourist spots etc. More past data could be included in the model. Explore different combinations of parameters for optimization. Weighting method could be used on multiple accident points, e.g. assigning heavier weight to them. In this way, a single cluster will have different probabilities which is more informative.

REFERENCES

[1] Vipul Rana1, Hemant Joshi2, "Road Accident Prediction using Machine Learning Algorithm", 2019 International Research Journal of Engineering and Technology (IRJET).

[2] Shristi Sonal and Saumya Suman, "A Framework for Analysis of Road Accidents" , Proceedings of 2018 International Conference on Emerging Trends and Innovations in Engineering and Technological Research (ICETIETR).

[3] Alysaa Ditcharoen , Bunna Chhour ,"Road Traffic Accidents Severity Factors: A Review Paper", 2018 5th International Conference on Business and Industrial Research (ICBIR), Bangkok, Thailand.

[4] Md. Farhan Labib, Ahmed Sady Rifat," Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh ", 2019 7th International Conference on Smart Computing & Communications (ICSCC).

[5] Chunjiao Dong,1,2,3 Chunfu Shao,1,2 Juan Li,1 and Zhihua Xiong 1,"An Improved Deep Learning Model for Traffic Crash Prediction" , 2018 Journal of Advanced Transportation.