
Road accidents prediction in India

PROJECT REPORT

Submitted in partial fulfillment of the requirements for the award of the degree

Of

**-MASTER IN COMPUTER APPLICATIONS-
(2019-2022)**

In

-INFORMATION TECHNOLOGY-

By

**Prachi Tomar
04304092019**

*Guided
by*

**Mr. Gaurav Indra
Assistant Professor**

Department of IT , IGDTUW



**INDIRA GANDHI DELHI TECHNICAL UNIVERSITY
FOR WOMEN
NEW DELHI – 110006**

MAY 2022

CERTIFICATE

This is to certify that the project entitled “**Road accidents prediction in India**” is being submitted at IGDTUW, Delhi for the award of **Masters in Computer Applications in Information Technology** degree. It contains the record of bonafide work carried out by **Prachi Tomar 04304092019** under my supervision and guidance. It is further certified that the work presented here has reached the standard of MCA and to the best of my knowledge has not been submitted anywhere else for the award of any other degree or diploma.



Mr.Gaurav Indra

Department of IT
IGDTUW

Date: 22 May 2022

Place: NEW DELHI

UNDERTAKING REGARDING ANTI-PLAGIARISM

I, ***Prachi tomar*** hereby, declare that the material/ content presented in the report are free from plagiarism and is properly cited and written in my own words. In case, plagiarism is detected at any stage, I shall be solely responsible for it. A copy of the Plagiarism Report is also enclosed.*

Prachi Tomar

(04304092019)

ACKNOWLEDGEMENT

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the completion of my project. All that I have done is only due to such supervision and assistance and I would not forget to thank them.

I respect and thank our respected faculty , Mr.Gaurav Indra, for providing me an opportunity to do the project work and giving us all support and guidance which made me complete the project duly. I am extremely thankful to mam for providing such a nice support and guidance, although she had busy schedule.

I am thankful to and fortunate enough to get constant encouragement, support and guidance from all Teaching staffs of esteemed University, which helped us in successfully completing our project work.

Prachi Tomar

(04304092019)

DECLARATION

I, **Prachi Tomar** solemnly declare that the project report, **ROAD ACCIDENTS PREDICTION IN INDIA**, is based on my own work carried out during the course of our study under the supervision of **Mr.Gaurav Indra,department of IT ,IGDTUW**. I assert the statements made and conclusions drawn are an outcome of my research work. I further certify that:

- I. The work contained in the report is original and has been done by me under the supervision of my supervisor.
- II. The work has not been submitted to any other Institution for any other degree/diploma/certificate in this university or any other University of India or abroad.
- III. We have followed the guidelines provided by the university in writing the report.
- IV. Whenever we have used materials (text, data, theoretical analysis/equations, codes/program, figures, tables, pictures, text etc.) from other sources, we have given due credit to them in the report and have also given their details in the references.

Prachi Tomar

(04304092019)

LIST OF ABBREVIATIONS

This section includes the list of abbreviations we used in the document. The table includes the abbreviation & full form of the respective abbreviation.

Abbreviation	Description
GDP	Gross Domestic Product
ITF	International Transport Forum
OECD	The Organisation for Economic Co-operation and Development

TABLE OF FIGURES

This section includes the list of figures we used in the document. The table includes name of figure and page number in sequence on which the figure is located.

S.NO	FIGURE NAME	PAGE NO.
1.	Flowchart for Project	17
2.	Time Slot Distribution using Pi chart	20
3.	State wise Analysis Graph	21
4.	Year wise Analysis Graph	22
5.	Top 5 State Analysis	23
6.	Least 5 State Analysis	24
7.	Union Territories wise Analysis	25
8.	Region wise Analysis	26
9.	Time slot wise Analysis	26
10.	Other Countries Analysis	27
11.	Graphical Representation (Linear Regression)	28
12.	Conclusion (Linear Regression)	29
13.	Graphical Representation (Gaussian Distribution)	30

LIST OF TABLES

This section includes the list of tables used in the document. The table includes the name of the table and the page number in sequence of the location of the table.

S.NO.	TABLE NAME	PAGE NO.
1.	Table 1: Data set by Ministry of Road Transport and Highways	18
2.	Table 2: Data set by International Transport Forum	18
3.	Table 3: Description of first data set	18
4.	Table 4: Description of second data set	19

ABSTRACT

Road accidents in India touched an all- time high in the year 2018, reporting almost 1.5 lakhs deaths, and almost 4,70,000 suffered- injuries. This costs India approximately 3-5 percent of its GDP share and therefore needs immediate addressing. In order to reduce the number of road accidents, it is important to study the data in detail and come up with effective solutions. Data analysis lets us predict the possible number of accidents that can take place in future.

Image detection and classification is used to detect potholes- on roads which is another major cause of road accidents.

INDEX

Certificate	1
Undertaking regarding anti plagiarism	2
Acknowledgement	3
Declaration	4
List of Abbreviations	5
Table of Figures.....	6
List of Tables	7
Abstract	8

1. Introduction

1.1. Literature review	#
1.2. Scope of the training/internship/project.....	#
1.3. Problem statement.....	#

2. Related Work

3. Proposed Methodology

3.1. Dataset Description	8
3.2. Attribute Description.....	9

4. Visualization

4.1. Time Slot Distribution.....	11
4.2. State Wise Analysis	12
4.3. Year Wise Analysis	13
4.4. Top 5 States Analysis	14
4.5. Least 5 States Analysis	14
4.6. Union Territory Wise Analysis	15
4.7. Region Wise Analysis	15
4.8. Time Slot Wise Analysis	16
4.9. Other Countries analysis	16

5. Algorithms

5.1. Linear Regression17

5.1.1. Description 17

5.1.2. Graphical Representation 17

5.1.3. Conclusion 18

5.2 Gaussian Distribution18

5.1.1 Description 18

5.1.2 Graphical Representation 19

5.1.3 Conclusion.....

6. Conclusion..... #

7. Future Scope#

8. Bibliography #

Chapter 1: Introduction

Citing the pace of the world today, road transport is the most extensively used mode of transport. According to reports, India reported approximately 151 thousand deaths due to road accidents alone in 2018, highest since the year 2005. Over 70 percent of the casualties reported are young Indians aged between 18-45 years. According to the global report by WHO, India accounts for almost 11 percent of road related deaths in the world. India ranks first in the number of road accidents in the world followed by China and the US.

There is one death in every four minutes in India and approximately 16 children die on roads everyday. Ranking second highest in the population and being a developing country itself, it is important for India to formulate measures to reduce the number of accidents. Almost 3-5 percent of GDP is invested in road accidents in India every year, which if taken care can increase India's GDP by approximately 7 percent. Therefore, road safety is the major cause of concern for the government and people themselves.

While the problem is well understood and reported, changes in policies and regulations require a detailed study of the data and causes. Data depicting accidents at different time periods, in different zones and different states allows scope for data analysis that can help produce results and roll out road traffic regulatory rules. Analysis lets us make predictions on the number of accidents that can take place in future. This information can be used to make arrangements to reduce casualties.

National Crime Reports Bureau, Ministry of Road Transport and Highway, Law commission of India, Global status report on road safety 2013 provided state-wise analysis of road accidents in India. Reports show that Delhi tops the charts with almost 5 deaths daily. One serious road accident occurs in every minute and 16 die on Indian roads every hour. Such data is useful to do targeted analysis and predictions.

There are various reasons leading to road accidents, drunk driving, over speeding, poor enforcement and poor city planning to name a few. While the causes can be numerous, some change in rules and regulations, better driver training, etc can help reduce the count by many-folds. The Supreme Court of India termed 15,000 deaths due to potholes as “unacceptable”. This is an alarming situation and negligence on the part of authorities in maintaining the roads. The number is greater than the deaths caused by terrorist attacks or on borders. Image classification of roads to detect potholes is helpful in maintenance and repair and therefore avoid such uninvited, easily avoidable deaths.

We investigate the problem of improper measures taken to reduce the accidents by analysing the data collected over 15 years (2001-2014). Careful study and visualization show where the problem needs to be addressed immediately and effectively. We implemented linear regression on our data to predict the number of accidents in any one of the time slots based on the number of accidents in other time slots. Gaussian distribution is applied to predict the number of accidents which can take place in any state at a given time in future based on previously recorded data. Convolutional Neural Network (CNN), a type of Artificial Neural Network is used for image recognition and classification to detect potholes.

1.1 Literature review

A review of existing literature provided insight into the dimensions of research done on road accidents across the globe. Extensive studies have been conducted relating to the demographic aspects of road accidents and infrastructure analysis in developed countries. Very little research has been done on the above in developing countries. Both descriptive and predictive research had been done and a review of the same had been carried out.

Rune (2006) suggests that the influence of a number of important Risk factors on road accidents can be described in terms of a few highly General “laws of accident causation”. These laws determine the shape of the Relationship between the risk factors and the occurrence of accidents.

Instances of all these laws as well as a discussion of how these laws can be Tested empirically were documented in this paper.

Chen (2007) developed a causal model to study the effect of mobile phone on driving safety. A relationship did exist between the use of mobile phones and driving safety suggesting an overall prohibition in mobile phone use while driving. Johnson et al (2004) studied the rate of usage of cell phones during day and night conditions along-with other demographic characteristics. They found that the most frequent distraction to a driver was the usage of cell phone. However the usage was less on weekends, or at night, or when the driver was exceeding the speed limit or had a passenger in the car. Rajalin et al (2005) conducted a study on the extent of awareness of the new law in Finland on using hands-free kits while driving. The survey determined the usage levels of mobiles with both handheld sets and with hands-free kits, after the legislation came into force.

Organisation of report :

- Visualisation displaying data to show state-wise, region-wise, time-slot wise distribution of road accidents.
- Comparison of India to other countries on the total number of accidents every 10,00,000 individuals.
- Linear regression to predict the number of accidents in any one of the time slots based on the number of accidents in other time slots.
- Gaussian distribution to predict the number of accidents which can take place in any state at a given time in future based on previously recorded data.
- Convolutional Neural Network (CNN), a type of Artificial Neural Network is used for image recognition and classification to detect potholes.

Conclusion : Successfully classified the images and achieved an average training accuracy of 96.76 percent and average validation accuracy of 84.12 percent.

1.3 Problem Statement

Road accidents account for a good share of the country's GDP. It is also the prime source of transportation for daily commuters which include pedestrians, cyclists, public and private transport users. Road accidents in India are major cause of deaths, injuries, fatalities every year. Therefore, its a major and growing health burden on Indian economy. Also traffic accidents cause physical, financial and mental stress on everyone directly and indirectly involved.

1.4 Objective

Predict the number of accidents in a time slot based on seven other time slots using linear regression and also predicting that how many accidents are- possible in given time slot in given current year using past years" data.

Visualize number of accidents for each state by year, change in percentage of accidents over the years, number of accidents for each state in different time slots and number of accidents in day and night using various charts and plots.

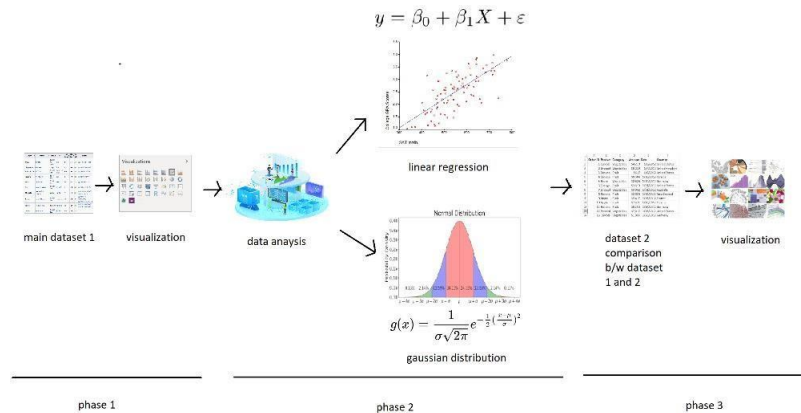
Classify the image of a road as whether the road have potholes or not normal using CNN.

Chapter 2: Related Work

Some works which related to the project are as follows : dueblein et al. [4] proposed a methodology to predict road accidents and the level of injury caused by the accidents. Yannis et al. [5] showed the effect of weather conditions on the road accidents and how dangerous an accidents can be due to weather conditions

Chapter 3: Proposed Methodology

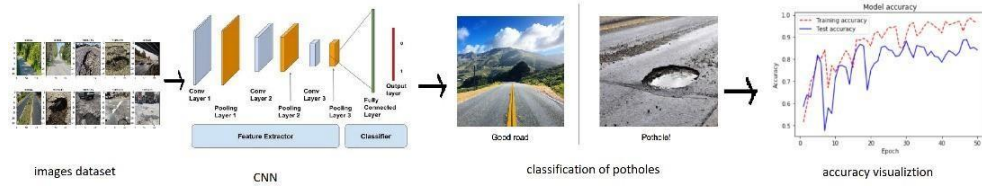
The flow chart for the whole project working is shown below:



Flow Chart Showing the methodology for Prediction and Visualization

The figure is divided into 3 phases, where in the first phase our main accident dataset undergoes various visualizations to understand nuances of the dataset. In phase 2 dataset is analysed through two algorithms linear regression and Gaussian distribution. In linear regression dataset values (from 0 to 21 hrs.) are taken as input and output the possible no. of accidents can occur in 21-24 hr. In Gaussian distribution algorithm, whole dataset act as an input and output is no. of accidents in next upcoming year.

In phase 3 the main data set is compared with another dataset of other countries and further different nuances are captured as visualizations.



Flow Chart Showing the methodology for Classification

Different road images dataset is given as input in Convolutional Neural Network (CNN), in this an image goes under 4 layers: Convolutional layer, ReLU layer, pooling, and fully connected layer. It outputs the classification of image as potholes or good road. After classification, the accuracy of the model is captured through graphs.

3.1 Dataset Description

We have used three datasets in the whole project.

The first dataset described in Table 3.2 data was collected from Ministry of Road Transport and Highways, and was provided in kaggle.[3] Table 3.3 is an example describing the dataset through counts of some key entities involved in the dataset. Every dataset also comprises of data attributes. Table 3.5 describes attributes of data. All the attributes in the dataset is unlabeled.

Table 3.1 Details of the dataset.

Details	Count
Number of instances	490
Number of attribute	11

The second dataset described in Table 3.3 data is collected from ITF(International Transport Forum) Transport Statistics and was provided on OECD data.[1]

Table 3.2 Details of the dataset.

Details	Count
Number of instances	722
Number of attribute	7

The third dataset contains two folders - normal and potholes. "Normal" contains images of smooth roads and "Potholes" contains images of roads with potholes in them. The images are collected from internet and was provided on kaggle.

Table 3.3 Details of the dataset.

Details	Count
Number of images in potholes	352
Number of images in normal	329

3.2 Attribute Description

The description of the first dataset attributes are shown in table 3.4 .

Table 3.4 Details of Data Attributes.

Data Attributes	Brief Explanation
STATE/UT	The state or union territory of India
YEAR	year of observation(2001-2014)
0-3 hrs. (Night)	Number of accidents in this time slot
3-6 hrs. (Night)	Number of accidents in this time slot
6-9 hrs (Day)	Number of accidents in this time slot
9-12 hrs (Day)	Number of accidents in this time slot
12-15 hrs (Day)	Number of accidents in this time slot
15-18 hrs (Day)	Number of accidents in this time slot
18-21 hrs (Night)	Number of accidents in this time slot
21-24 hrs (Night)	Number of accidents in this time slot
Total	Total number of accidents in that year

The description of second dataset's attributes are shown in table 3.5 .There were total seven attributes in the dataset but only five are used in the project. [-0.05in]. **Table 3.5** Details of Data Attributes.

Data Attributes	Brief Explanation
Location	The country where the accident occurred
Indicator	Indicates the place of accident
Subject	Indicates the severity of the accident
Time	Indicates the year of accident
Value	Number of accidents per 1000000 individual

The third dataset contains images with .jpg extension. There are two folders of images having more than 300 images the names of folders are "potholes" and "normal".

Chapter 4: Visualization

The information which can be observed using the datasets are shown below using graphs and charts.

4.1 Time Slot Distribution

Time slot distribution of all accidents in India(2001-14)

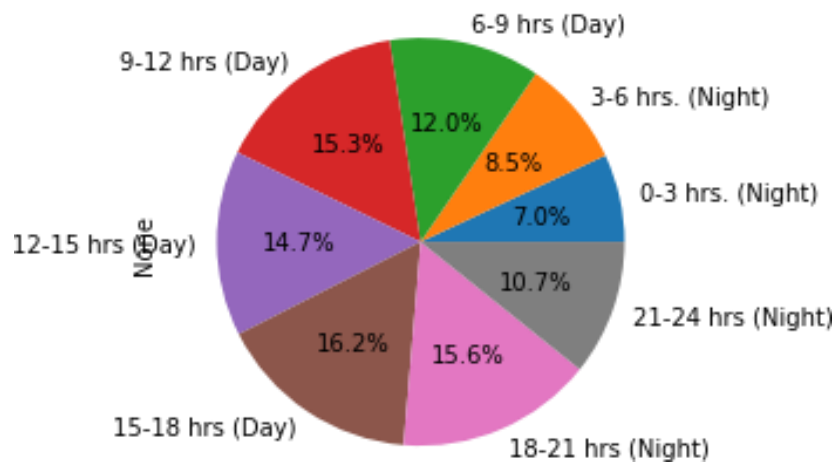


Fig 4.1 Distribution Using Pie Chart

fig 3.1 shows time slot distribution (hourly basis)for both day night using a pie chart. That is the percentage of accidents based on hourly slot. maximum accidents occur during the day in between 15:00 and 18:00. minimum accidents occur during the night in between 00:00 and 03:00.

4.2 State Wise Analysis

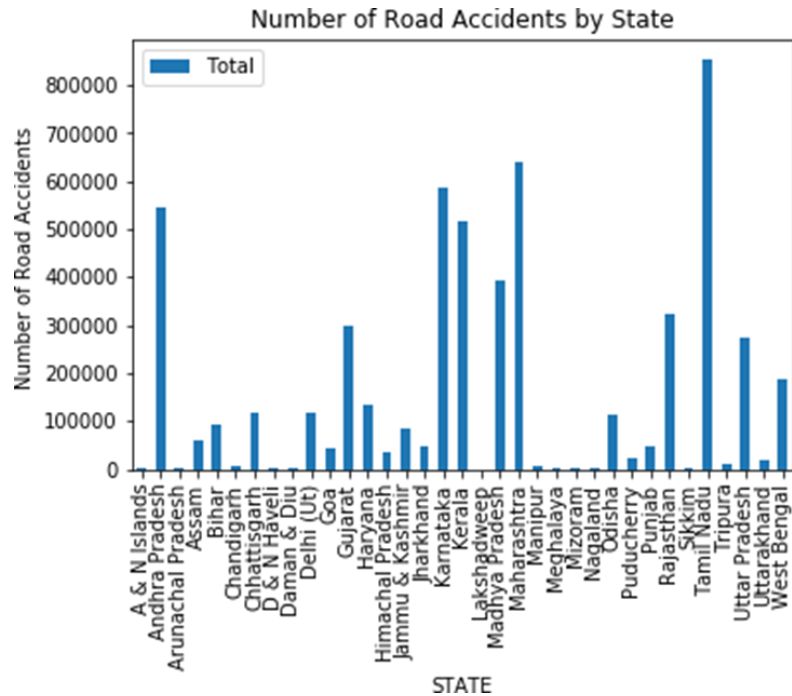


Fig 4.2 State Wise analysis Graph

fig 4 .2 shows the bar graph between states on X-axis and number of road accidents on Y-axis. The number of accidents is the total number of accidents observed year wise. conclusion- the state with least number of accidents is The state with highest number of accidents is Tamil Nadu.

4.3 Year Wise Analysis



Fig 4.3 Yearly accidents

fig 4.3 shows the bar graph between Year on X-axis and number of accidents on Y-axis. conclusion- least accidents in the year 2001 and highest number of accidents in the year 2014. we can clearly draw inference that number of accidents are increasing year by year.

4.4 Top 5 States Analysis

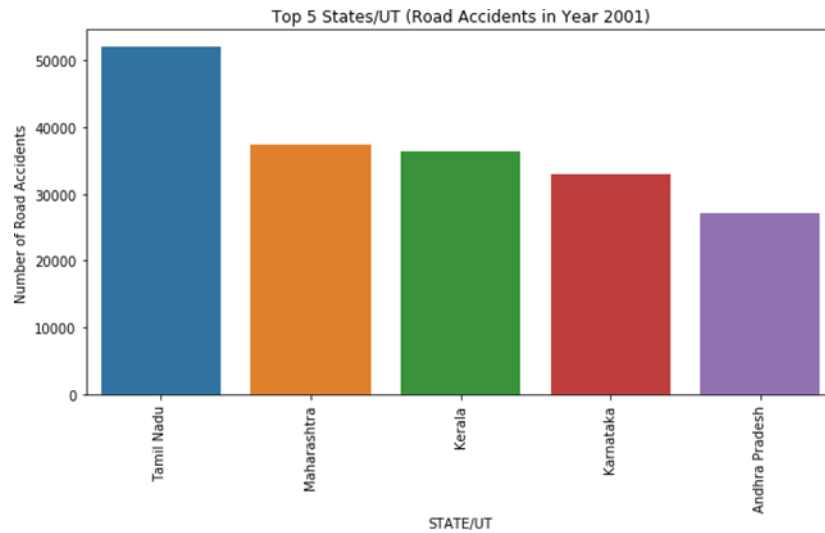


Fig 4.4 States having Maximum Number of Accidents

Fig 4.4 shows the bar graph between top 5 states in terms of road accidents and number of road accidents happening on Y-axis. Tamil Nadu is the state with highest number of road accidents. Andhra Pradesh is the state with least number of road accidents among the top 5 states. Also we can infer, all the 5 states are southern states of India. Hence, a higher risk in south indian states.

4.5 Least 5 States Analysis

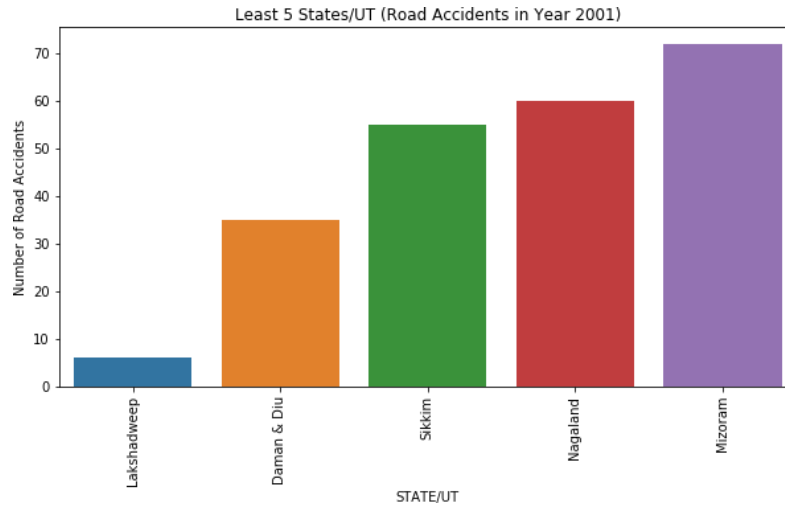


Fig 4.5 States having Minimum number of Accidents

The above figure shows the bar graph between top 5 states/UTs with least number of accidents . Lakshadweep has least number of accidents. Mizoram has highest number of accidents among these. we can also infer that the north eastern states have least accidents among all.

4.6 Union Territory Wise Analysis

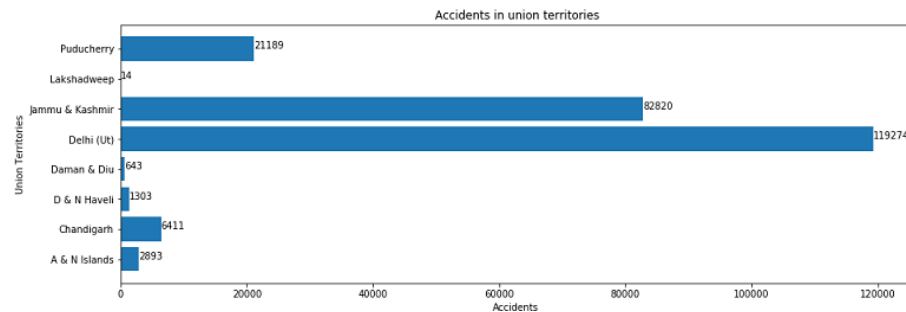


Fig 4.6 Accidents in Union Territories

fig 4.6 shows the bar graph between union territories on X-axis and number of accidents on Y-axis. Delhi has maximum number of accidents lakshadweep has least number of accidents.

4.7 Region Wise Analysis

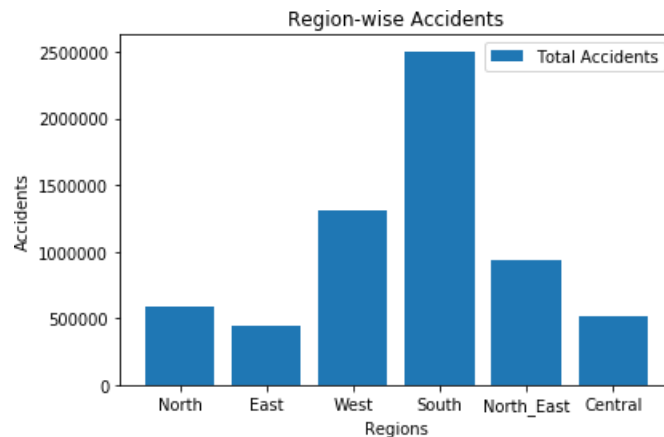


Fig 4.7 Region Wise Bar Graph

fig 4.7 shows the bar graph between Regions of Indian state on X-axis and accidents on Y-axis. south India has the highest occurrence of accidents East India has the lowest

4.8 Time Slot Wise Analysis

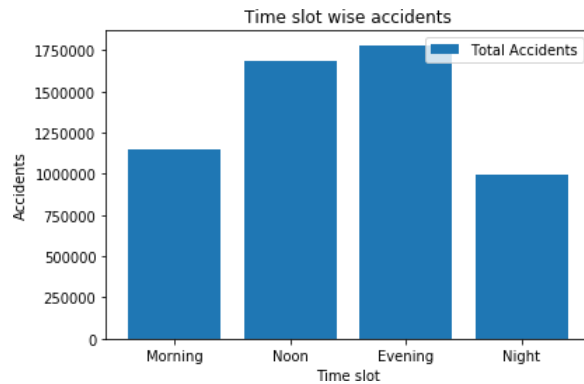


Fig 4.8 Time Wise Bar Graph

fig 4.8 shows the bar graph between the time slot (morning, noon, evening, night) on X-axis and number of accidents on Y-axis. maximum accidents occur during evening time slot and minimum accidents occur at night.

4.9 Other Countries analysis

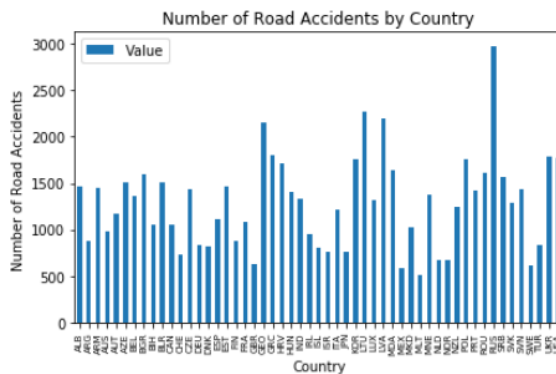


Fig 4.9 Country Wise Analysis

fig 3.9 shows the bar graph with number of road accidents on y-axis and name of country on the x-axis. By using this graph we can compare india with other countries in terms of Number of accidents. We can see that Malta has least number of accidents and Russia has highest number of accidents.

Chapter 5 : Algorithms

There are total three algorithm used in this project. They are explained below and the inputs outputs are also specified..

5.1 Linear Regression

5.1.1 Description

This algorithm is based on supervised learning. It is used for finding the relationship between variables and forecasting. Linear regression performs the task to predict a dependent variable (y) based on a given independent variable (x). So, this algorithm finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression. In the figure below, X (input) are the time slots from 0-21 hrs and Y (output) is the time slot 21-24 hrs.

Equation :

$$Y = a_0 + a_1X_1 + a_2X_2$$

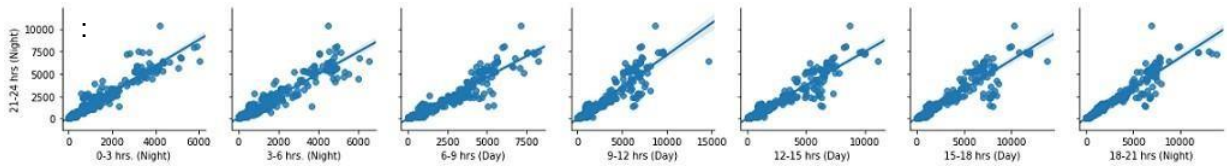
Our case is of multiple linear regression as we have multiple time slot values (independent variables) and based on number of accidents in all these seven time slots we are predicting the number of accidents in 8th time slot (dependent variable).

Equation:

$$Y = a_0 + \sum_{i=1}^7 a_i X_i$$

where $a_0 = \text{intercept}$ and $a_i = \text{regression coefficient/slope}$

5.1.1 Graphical Representation



5.1.2 Conclusion

Co-efficient of determination , that is, score is 0.9831726728479063 slope :

0.95377072 for 0-3 hrs(night)

-0.05388771 for 3-6 hrs(night)

-0.12513327 for 6-9 hrs(day)

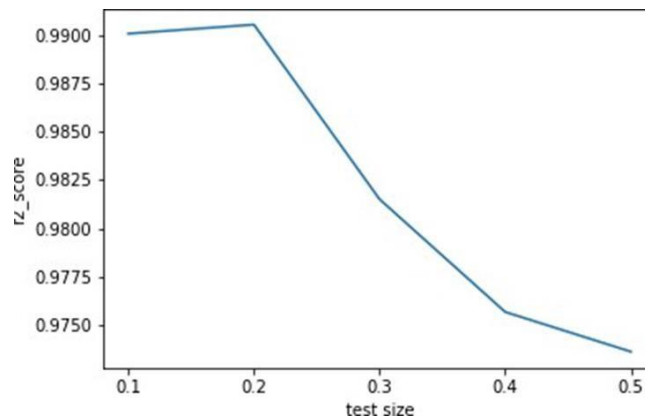
-0.23734929 for 9-12 hrs(day)

0.09933494 for 12-15 hrs(day)

0.31747461 for 15-18 hrs(day)

0.20402073 for 18-21 hrs(night)

Intercept is -3.5794359619339957



Conclusion: r2 score falls as test size increases

5.2 Gaussian Distribution

5.2.1 Description

This algorithm is also called the Gaussian distribution or the bell curve distribution.

The distribution can be defined using two parameters:

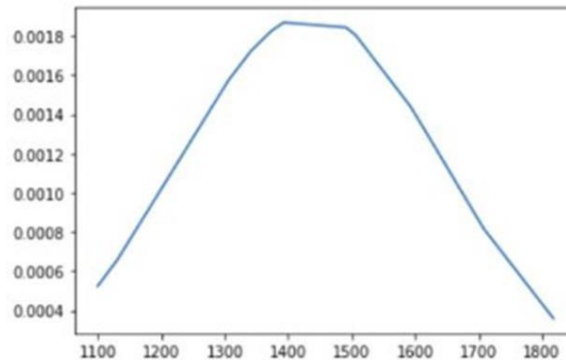
Mean (μ): The expected value.

Variance (σ^2): The spread from mean.

Standard Deviation (σ): The average spread from the mean.

Motivation is from [2]. A distribution with a mean of zero and a standard deviation of 1 is called a standard normal distribution, and often data is reduced or “standardized” to this for analysis for an easy interpretation and comparison. By using Gaussian distribution we have predicted the number of accidents which can take place in a given state of India at a given time in the year 2015. In the below shown graph we have shown the probability distribution and given state “Delhi (UT)” and time “18-21 hrs (night)” as input and bell shaped graph is obtained.

5.3 Graphical Representation



5.3.1 Conclusion

The Number of average accident which can occur in 18-21 hrs in 2015 are 1436.

CHAPTER 6: CONCLUSION

1. Maximum accidents occur during the day in between 15:00 and 18:00. Minimum accidents occur during the night in between 00:00 and 03:00.
2. r^2 score falls as test size increases
3. The Number of average accident which can occur in 18-21 hrs in 2015 are 1436.
4. Tamil Nadu is the state with highest number of road accidents. Andhra Pradesh is the state with least number of road accidents

CHAPTER 7: FUTURE SCOPE

It is important for India to formulate measures to reduce the number of accidents. Almost 3-5 percent of GDP is invested in road accidents in India every year, which if taken care can increase India's GDP by approximately 7 percent. Therefore, road safety is the major cause of concern for the government and people themselves.

While the problem is well understood and reported, changes in policies and regulations require a detailed study of the data and causes. Data depicting accidents at different time periods, in different zones and different states allows scope for data analysis that can help produce results and roll out road traffic regulatory rules. Analysis lets us make predictions on the number of accidents that can take place in future. This information can be used to make arrangements to reduce casualties.

BIBLIOGRAPHY

- [1] <https://data.oecd.org/transport/road-accidents.htm>
- [2] <https://machinelearningmastery.com/continuous-probability-distributions-for-machine-lear>
- [3] Dataset is taken from, <https://www.kaggle.com/vikasds101/road-accident-state-time>
- [4] Deublein, M., Schubert, M., Adey, B.T., Köhler, J., Faber, M.H.: Prediction of road accidents: A bayesian hierarchical approach. Accident Analysis & Prevention 51, 274–291 (2013)
- [5] Yannis, G., Karlaftis, M.G.: Weather effects on daily traffic accidents and fatalities: a time series count data approach. In: Proceedings of the 89th Annual Meeting of the Transportation Research Board. vol. 10, p. 14 (2010)
- [6] ACT, (2008). National Road Safety Action Plan 2009 and 2010. Australian Transport Council, Canberra.
- [7] Ardekani, S., Hauer, E., Jamei, B., (2000). Traffic impact models. In: Gartner, N., Messer, C.J., Rathi, A.K. (Eds.), Traffic Flow Theory: A State of the Art Report. Federal Highway Administration, Washington.
- [8] Barnes, G.R., Thompson, K.B., (2006). A longitudinal analysis of the effect of bicycle facilities on commute mode share. In: Transportation Research Board 85th Annual Meeting, Transportation Research Board, Washington, DC.

- [9] BISCG(2008). Bengaluru Injury Surveillance Collaborators Group. Bengaluru Injury Surveillance Programme: a feasibility study. National Institute of Mental Health and Neuro Sciences, Bangalore. Publication No. 68, 2008.
- [10] Brüde, U., Larsson, J., (1993). Models for predicting accidents at junctions where pedestrians and cyclists are involved. How well do they fit? Accident Analysis and Prevention 25 (5), 499–509.
- [11] Crandall JR, Bhalla KS, Madeley NJ (2002). Designing road vehicles for pedestrian protection. British Medical Journal, 2002, 324: 1145–1148.
- [12] De Hartog, J.J., Boogaard, H., Nijland, H., Hoek, G.,(2010). Do the health benefits of cycling outweigh the risks? Environmental Health Perspective, 118 (8),1109–1116.
- [13] Damsere-Derry J, et al(2010). Pedestrians,, injury patterns in Ghana. Accident Analysis & Prevention, 2010, 42: 1080–1088.
- [14] Elvik, R., Bjørnskau, T., (2005). How accurately does the public perceive differences in transport risks? An exploratory analysis of scales representing perceived risk. Accident Analysis and Prevention 37 (6), 1005–1011.
- [15] Gallagher, Rob, The Rickshaws of Bangladesh. Dhaka: University Press Limited, 1992.
- [16] Garder, P., Leden, L., Pulkkinen, U., (1998). Measuring the safety effect of raised bicycle crossings using a new research methodology. Transportation Research Record, 1636, 64–70.
- [17] Gommers, M.J.P.F.,Bovy, P.H.L., (1987). Routekeuzegedrag en netwerkgebruik (RouteChoice and Network Use). Delft University of Technology.

[18] Heinen, E., Van Wee, B., Maat, K., (2010). Commuting by bicycle: an overview of the literature. *Transport Reviews* 30 (1), 59–96.

[19] Houtenbos, M., (2008). *Expecting the Unexpected: A Study of Interactive Driving Behaviour at Intersections*. Institute for Road Safety Research, Leidschendam.

[20] Gururaj G(2006). *Road traffic Injury Prevention in Bangalore, India*. National Institute of Mental Health and Neuro Sciences, Bangalore. Publication No. 56.

