Project Presentation

Road accidents prediction in India

Under Supervision of:

Ms. Monalisa

Professor

Department of Information Technology

Department of Information Technology

Indira Gandhi Delhi Technical University for Women

Kashmere Gate,New Delhi 110006

December 2020

Submitted By:

Anjali Deswal: 00904092018

Prabhleen Kaur: 02404092018

Himanshi Sharma: 04304092018

Neha Verma: 05304092018

Contents

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.

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.

Project's contribution and 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.

1.1 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, it's a major and growing health burden on Indian economy. Also traffic accidents cause physical, financial and mental stress on everyone directly and indirectly involved.

1.2 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

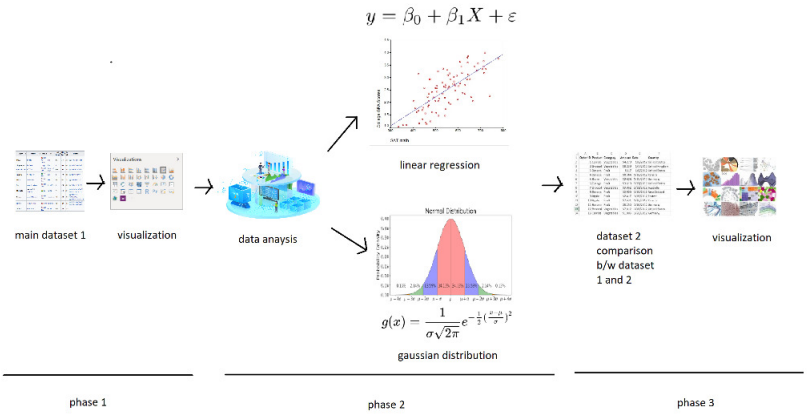
Related Work

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

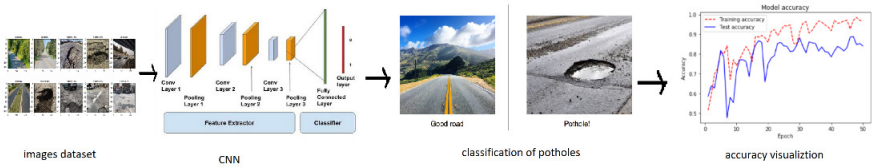
Proposed Methodology

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 analyses 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 hrs. 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 in 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 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 form internet and was provided on kaggel.

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's 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 occured
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

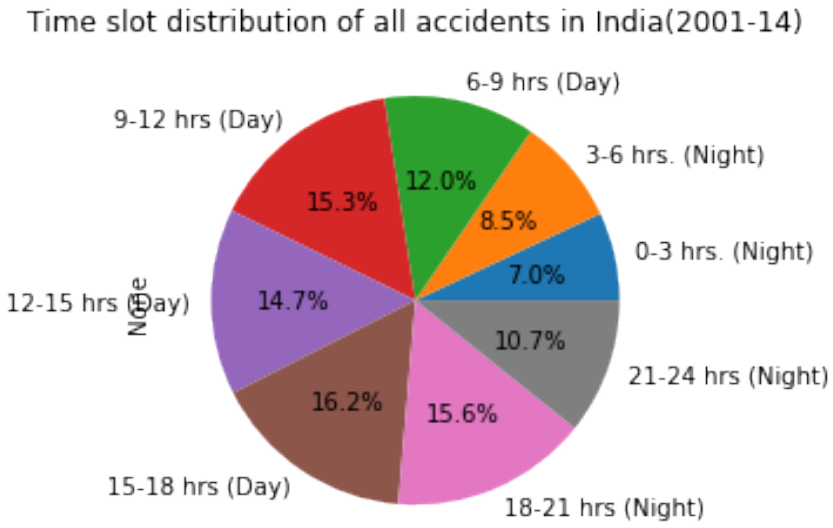


Fig 4.1 Distribution Using Pie Chart

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

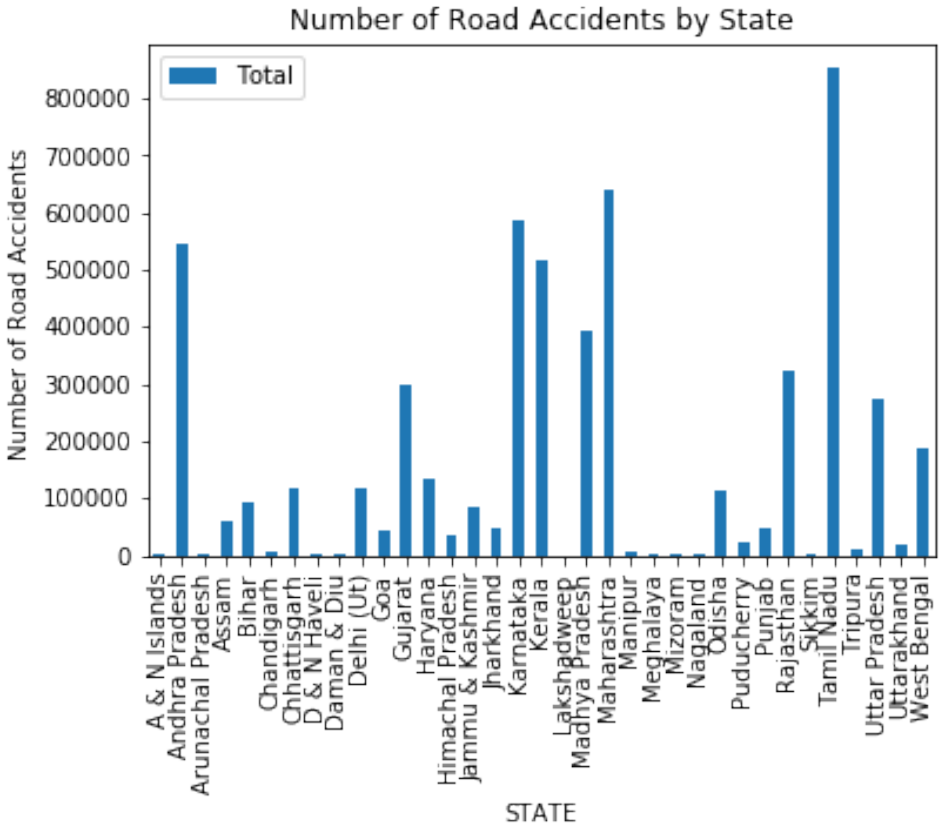


Fig 4.2 State Wise analysis Graph

fig 3.2 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

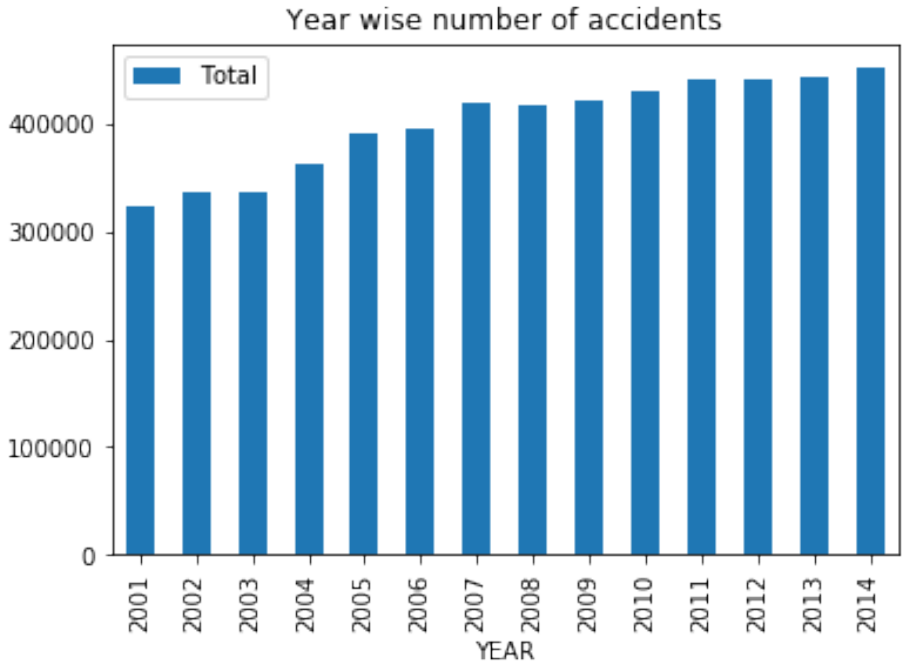


Fig 4.3 Yearly accidents

fig 3.3 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

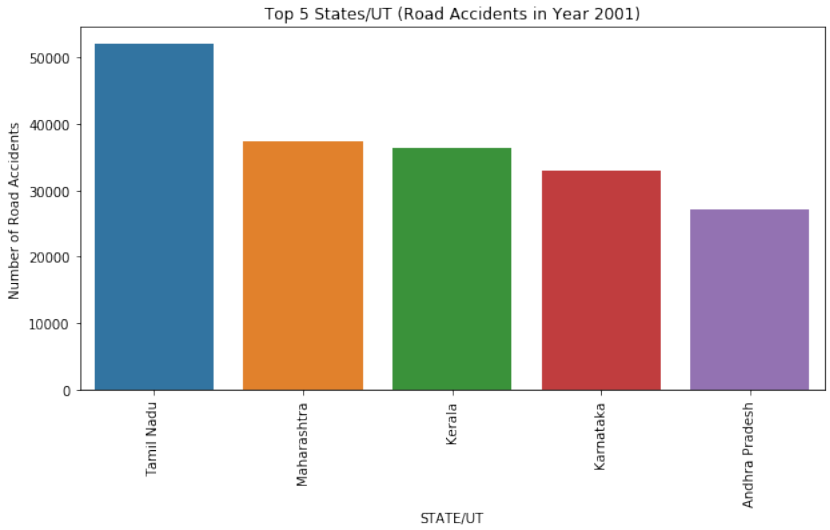


Fig 4.4 States having Maximum Number of Accidents

Fig 3.4 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

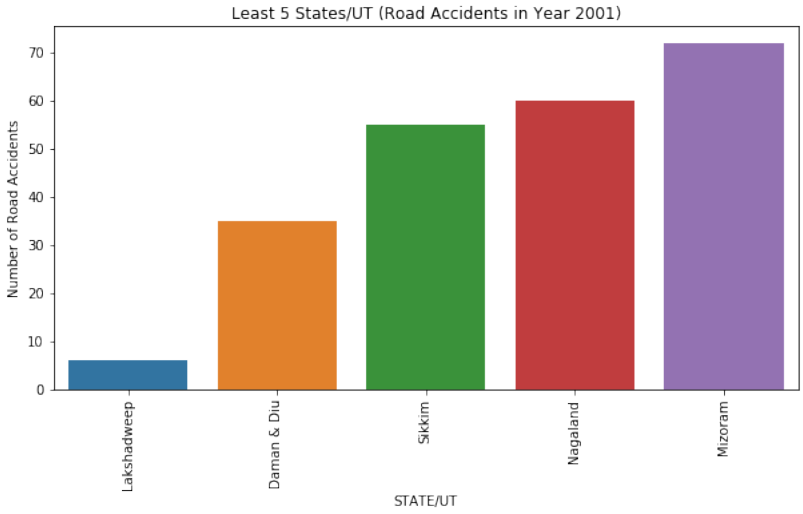


Fig 4.5 States having Minimum number of Accidents

fig 3.5 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

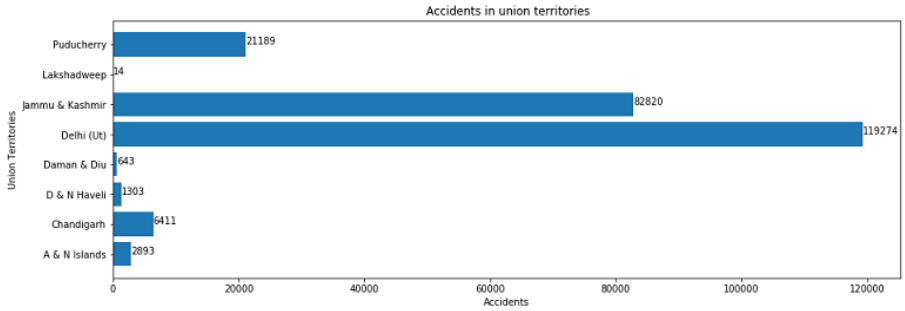


Fig 4.6 Accidents in Union Territories

fig 3.6 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

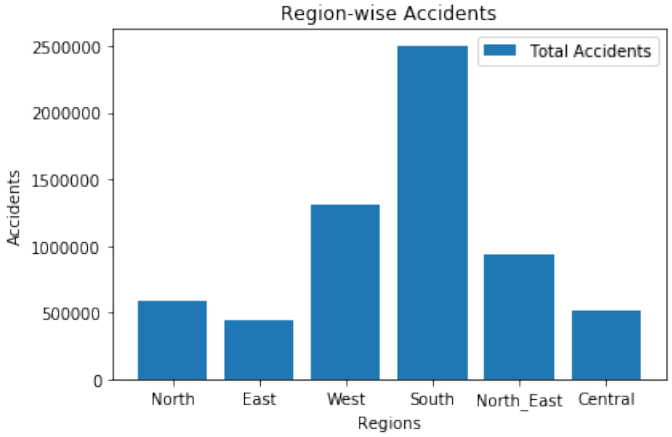


Fig 4.7 Region Wise Bar Graph

fig 3.7 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 occurrence of states.

4.8 Time Slot Wise Analysis

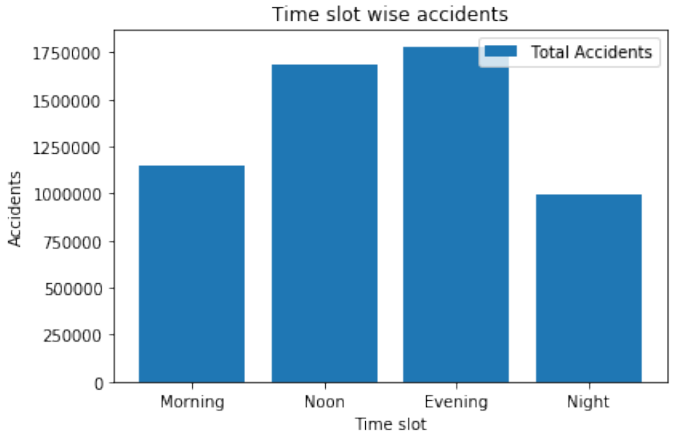


Fig 4.8 Time Wise Bar Graph

fig 3.8 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

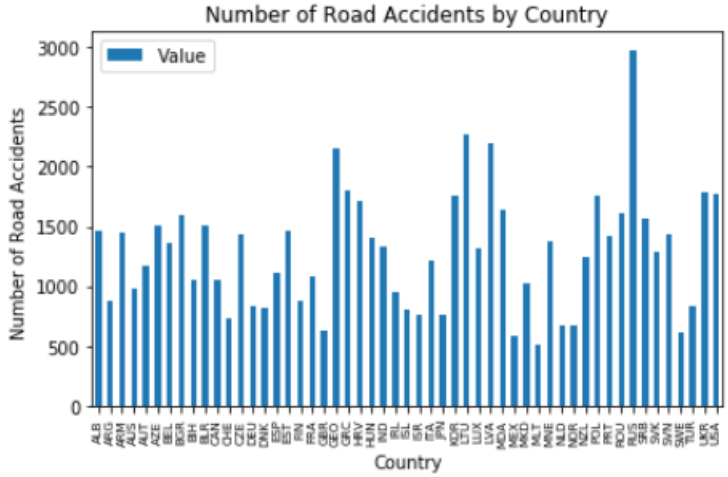


Fig 4.9 Country Wise Analysis

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 :

$$Y = a_0 + a_1 X_1 + a_2 X_2$$

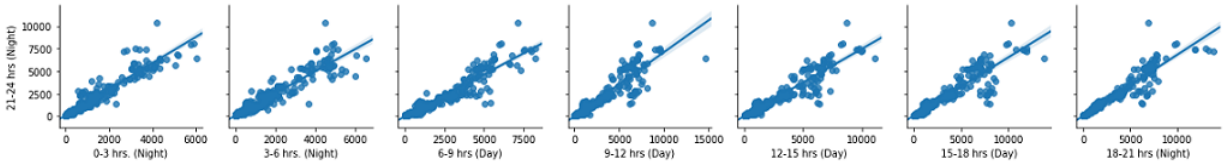
Our case is of multiple linear regression as we have multiple time slot values(independent variables) and based on number of accidents in all these seven time slots we are predicting the number of accidents in 8th time slot(dependent variable).
Equation:

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

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

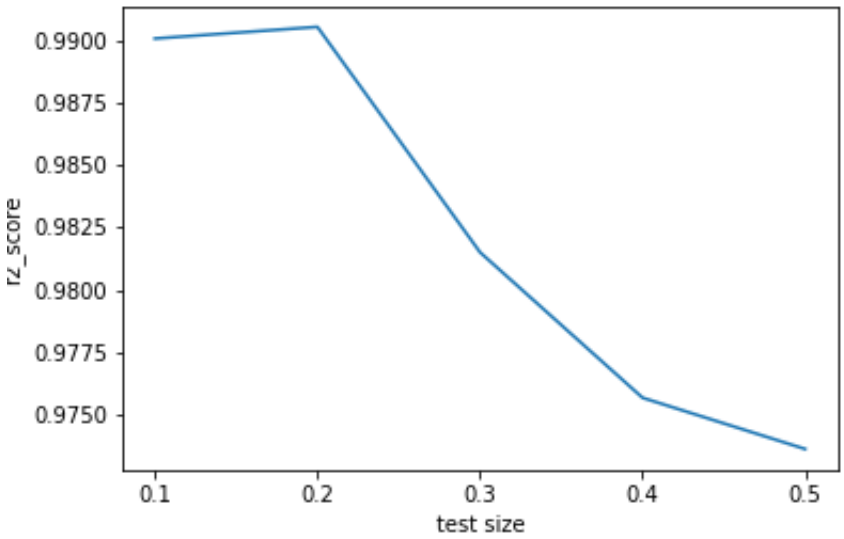
5.1.2 Graphical Representation

:



5.1.3 conclusion

: coefficient 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

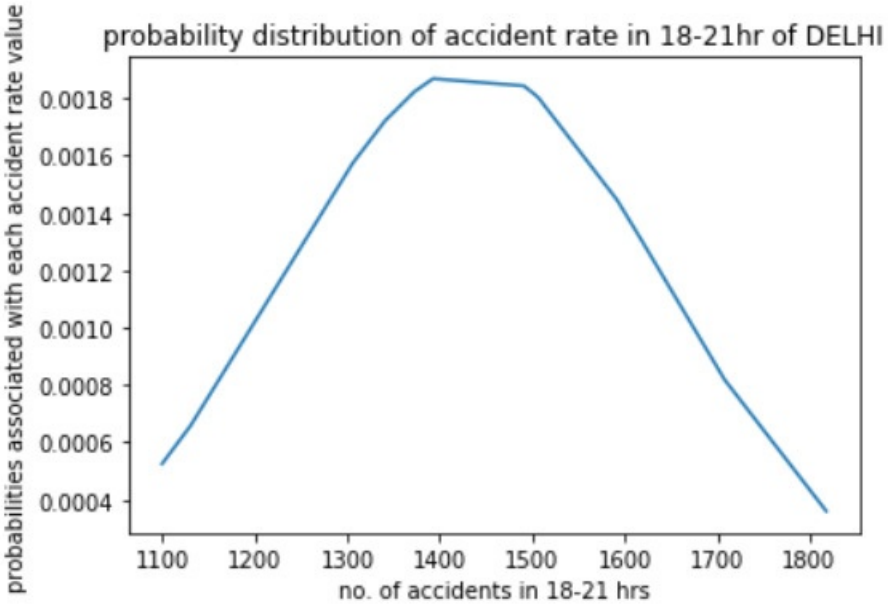
5.2.1 Description

This algorithm is also called the Gaussian distribution or the bell curve distribution.
The distribution can be defined using two parameters:
Mean (mu): The expected value.
Variance (sigma*sigma): The spread from mean.
Standard Deviation (sigma): The average spread from the mean.
Motivation is from [2]. A distribution with a mean 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.

5.2.2 Graphical Representation

:



5.2.3 Conclusion

The Number of average accidents which can occur in 18-21 hrs(Night) in 2015 Delhi(UT) is: 1436

5.3 Convolutional neural network

5.3.1 Description

Convolutional Neural Network is a type of Artificial Neural Network which is used in recognition and classification of images. It basically has four layers – Convolutional layer, ReLU layer, pooling and fully connected layer.
The architecture of the CNN model is discussed below-
1) We developed a sequential model, where layers are connected sequentially to each other.

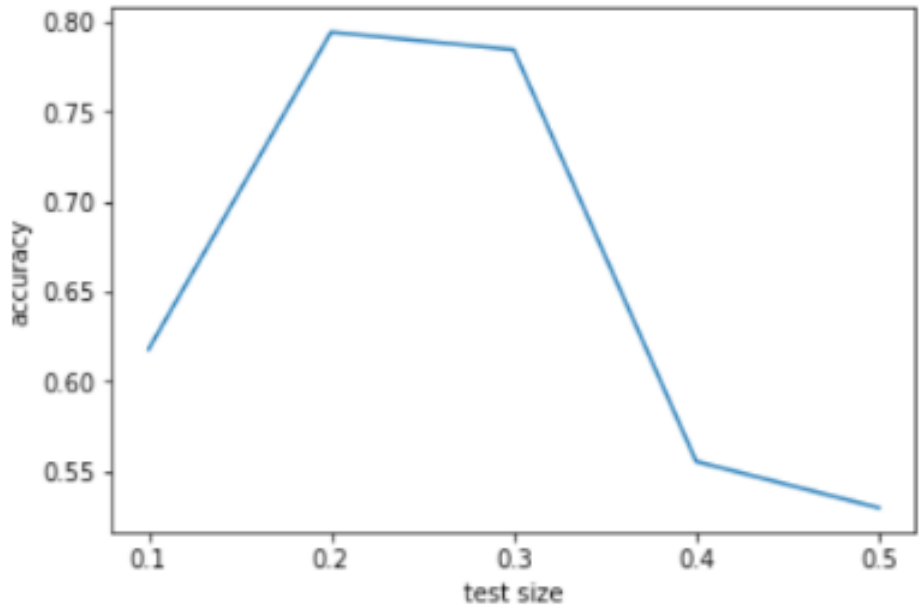
2) Input is passed to a series of convolution layers and ReLU activations. These layers help in the extraction of some features from the image. We used ReLU layer to remove all the negative values that we got from the output of convolutional layer.
3) Each convolution layer is followed by a max pooling layer which helps in reducing the dimension of input further.
4) Then we used Flatten layer this is the final layer where actual classification takes place.so here we take our filtered and shrink images and we put them into a single list.
5) Then we used Dropout for randomly discarding the neurons to avoid overfitting.
6) Finally, the output of the previous layer enters as an input to the dense layer with one neuron that finally classifies the input as 0 or 1.
7) The model uses categorical cross-entropy as the loss function which is a logarithmic loss function.
8) Adam optimizer with various parameters like learning rate has been used for optimization.

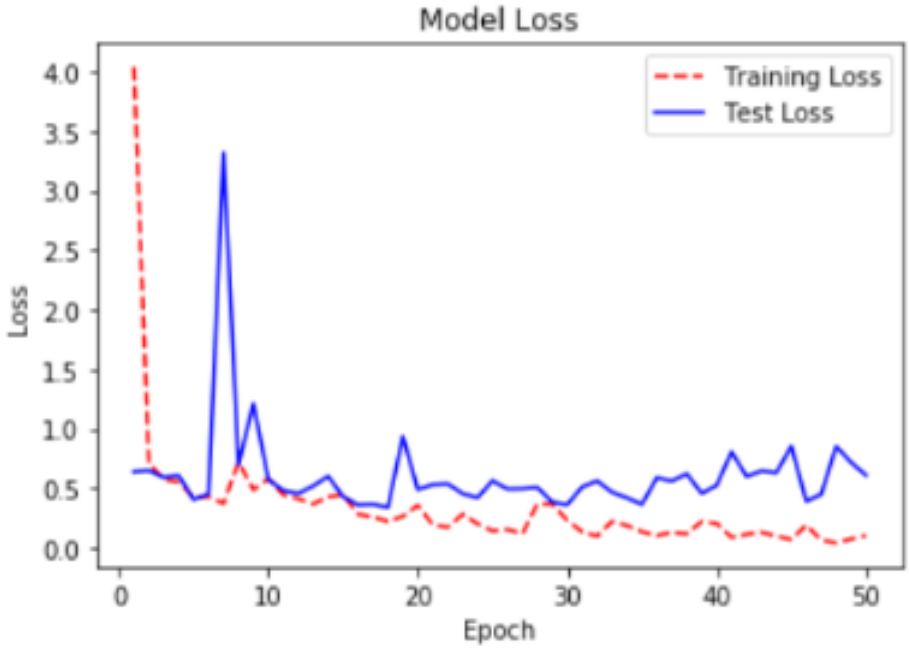
The parameters taken during the experiment are given below –

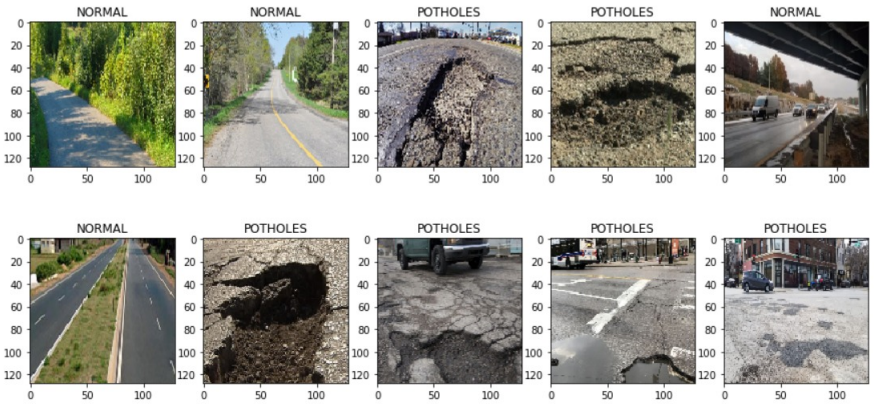
- Train-test split: 75:25
- Image size: 128*128
- Total categories:2
- Total images:681
- Number of epochs:50
- Batch size:12
- Learning rate:0.001
- Activation: ReLU for convolutional layer, softmax for Dense layer
- Loss function: categorical cross entropy.

5.3.2 Graphical Representation

:







5.3.3 Conclusion

From this experiment, we have successfully classified the images and achieved an average training accuracy of 96.76 percent and average validation accuracy of 84.12 percent.

Bibliography

[1] https://data.oecd.org/transport/road-accidents.htm

[2] https://machinelearningmastery.com/continuous-probability-distributions-for-machine-learn

[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)