

BUSINESS INTELLIGENCE PROJECT

Road accidents prediction in India

UNDER SUPERVISION OF:

Mr. Rishabh kaushal

ASSISTANT PROFESSOR

DEPARTMENT OF INFORMATION TECHNOLOGY



Department of Information Technology

Indira Gandhi Delhi Technical University for Women

Kashmere Gate, New Delhi 110006

March 2020

SUBMITTED BY:

TEAM 3

MUSKAAN MITTAL: 00304092018

ANJALI DESWAL: 00904092018

PRABHLEEN KAUR: 02404092018

HIMANSHI SHARMA: 04304092018

VASUNDHRA BHARDWAJ: 05004092018

NEHA VERMA: 05304092018

Contents

1 Introduction

1.1 PROBLEM STATEMENT

1.2 OBJECTIVE

2 Proposed Methodology

2.1 DATASET DESCRIPTION

2.2 ATTRIBUTE DESCRIPTION

3 Visualization

3.1 TIME SLOT DISTRIBUTION

3.2 STATE WISE ANALYSIS

3.3 YEAR WISE ANALYSIS

3.4 TOP 5 STATES/UTs ANALYSIS

3.5 LEAST 5 STATES/UTs

3.6 UNION TERRITORY WISE ANALYSIS

3.7 REGION WISE ANALYSIS

3.8 TIME SLOT ANALYSIS

4 Algorithms

4.1 LINEAR REGRESSION

4.2 GAUSSIAN DISTRIBUTION

1 Introduction

In today's world road and transport has become an integral part of every human being. Every body is a road user in one shape or the other. Road Transport is considered to be one of the most cost effective and preferred mode of transport, both for freight and passengers, keeping in view its level of penetration into populated areas. The typical road users include pedestrians, cyclists, motorists, vehicle passengers and passengers of on-road public transport. Road safety is a necessary measure that needs to be taken for the safety of all road users.

1.1 Problem Statement

Road transport is the dominant mode of transport in India, in terms of traffic share and in terms of contribution to the national economy. Road accidents in India are major source 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 effects for everyone 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.

2 Proposed Methodology

2.1 Dataset Description

The data was collected from Ministry of Road Transport and Highways, and was provided in kaggle. Table 1 is an example describing the dataset through counts of some key entities involved in the dataset. Every dataset also comprises of data

Details	Count
Number of instances	490
Number of attribute	11

Table 1. Details of the dataset.

attributes. Table 2 describes attributes of data. All the attributes in the dataset is unlabeled.

2.2 Attribute Description

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

Table 2. Details of Data Attributes.

3 visualization

3.1 Time slot distribution

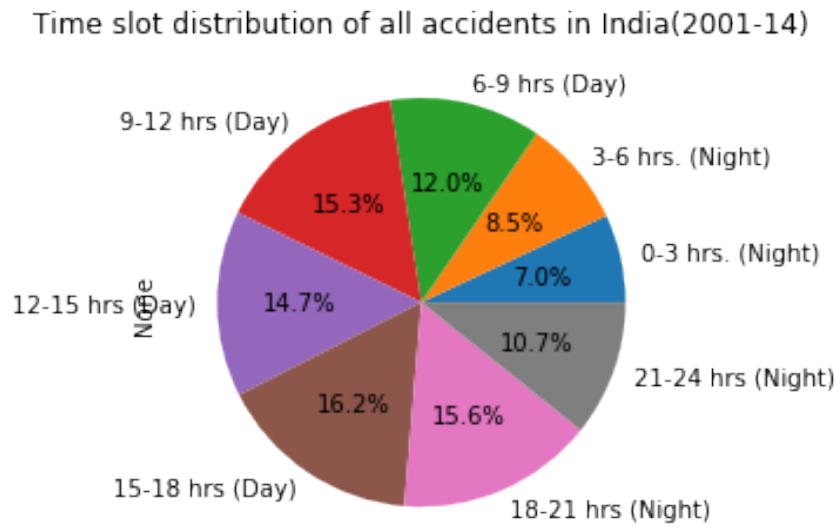


fig 3.1

fig 3.1 shows time slot distribution (hourly basis) for both day and 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.

3.2 state wise analysis

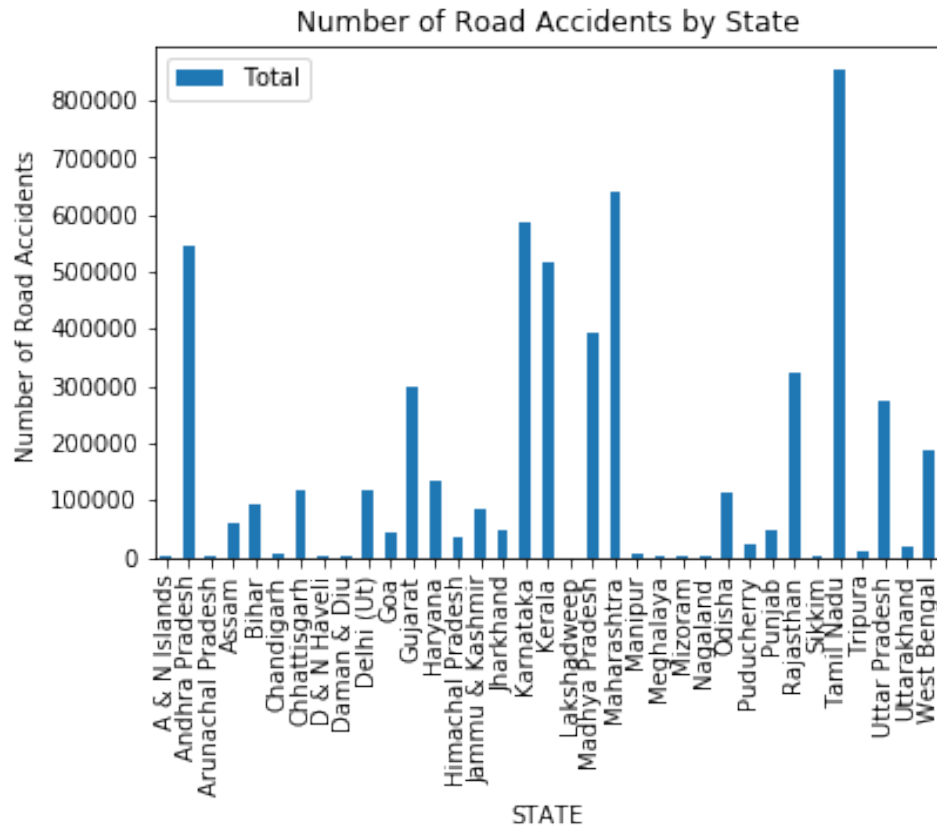


fig 3.2

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.

3.3 year wise analysis



fig 3.3

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.

3.4 top 5 states analysis

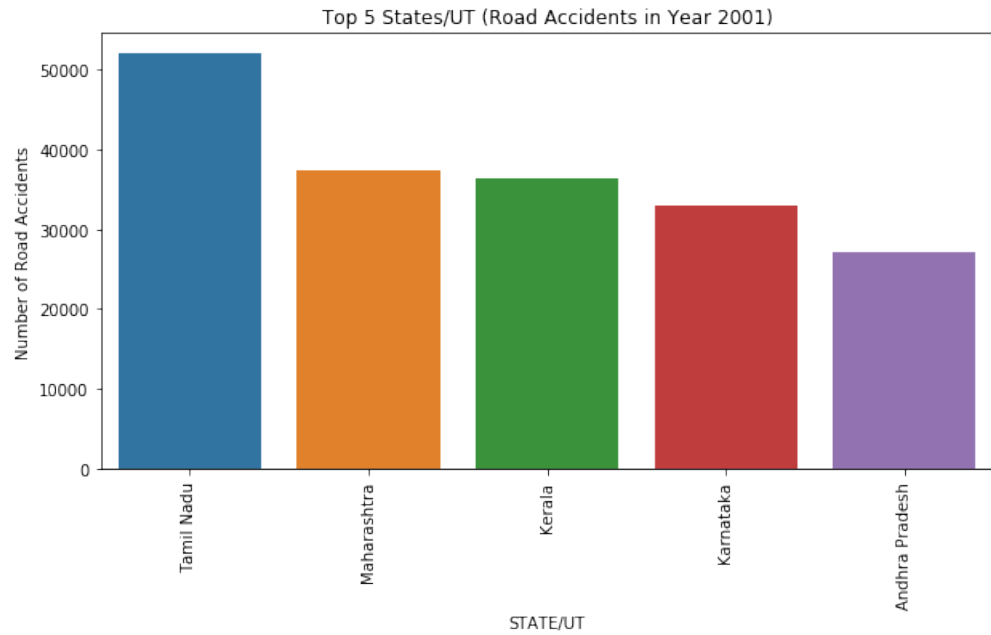


fig 3.4

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.

3.5 least 5 states analysis

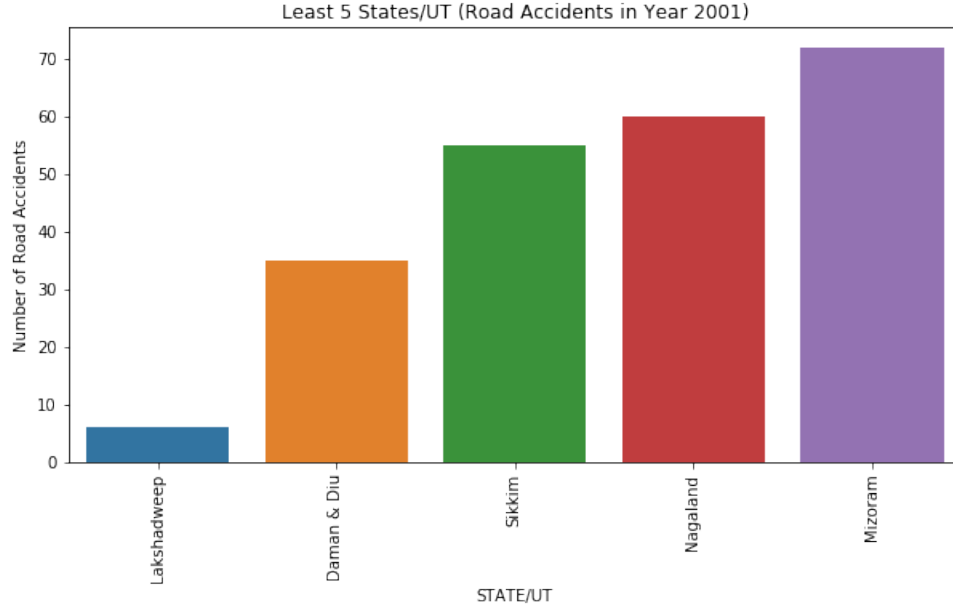


fig 3.5

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.

3.6 union territory wise analysis

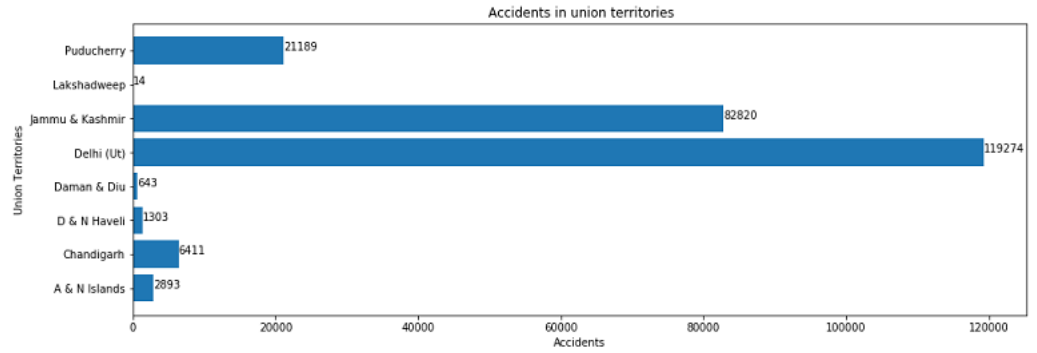


fig 3.6

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.

3.7 region wise analysis

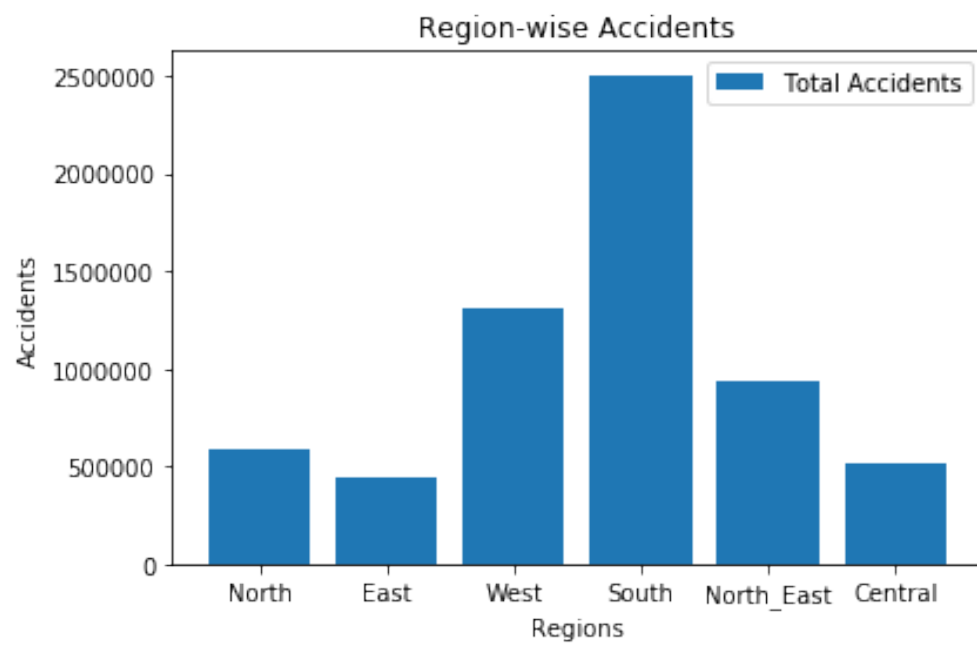


fig 3.7

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.

3.8 time slot wise analysis



fig 3.8

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.

3.9 Other Countries analysis

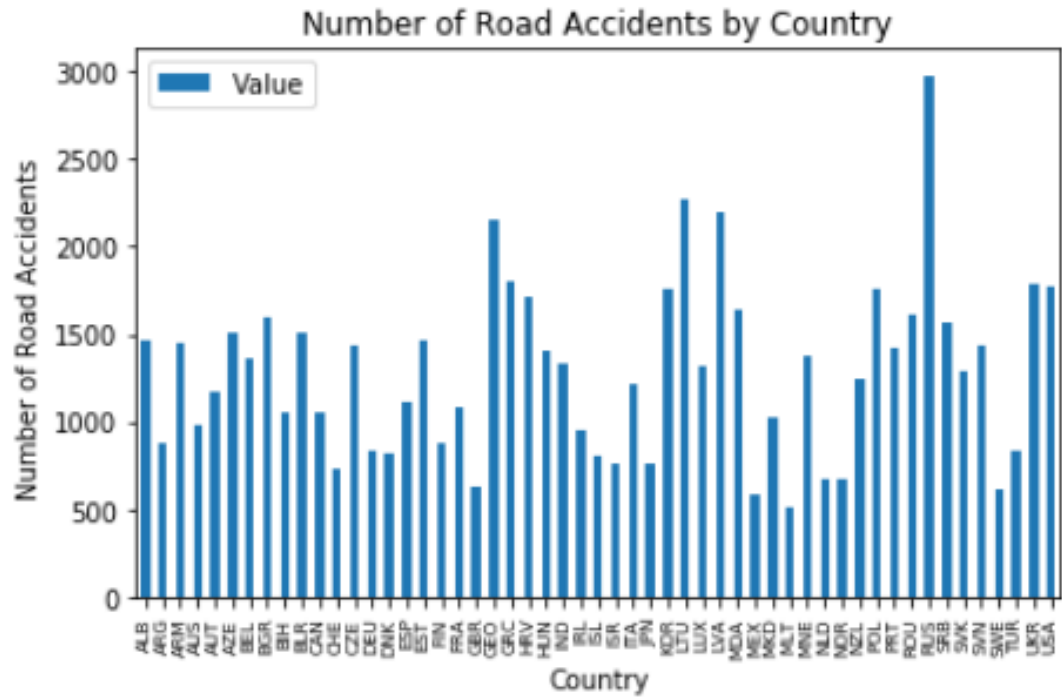


fig 3.9

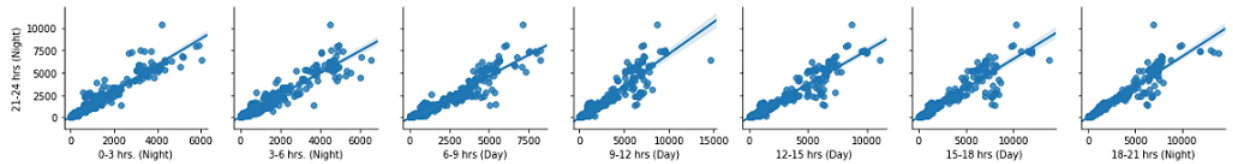
fig 3.9 shows the bar graph with number of road accidents on y-axis and name of country on the x-axis. By using this graph we can compare india with other countries in terms of Number of accidents. We can see that Malta has least number of accidents and Russia has highest number of accidents.

4 Algorithms

4.1 Linear Regression

description Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique 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. The regression line is the best fit line for our model.

Graphical Representation :



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)

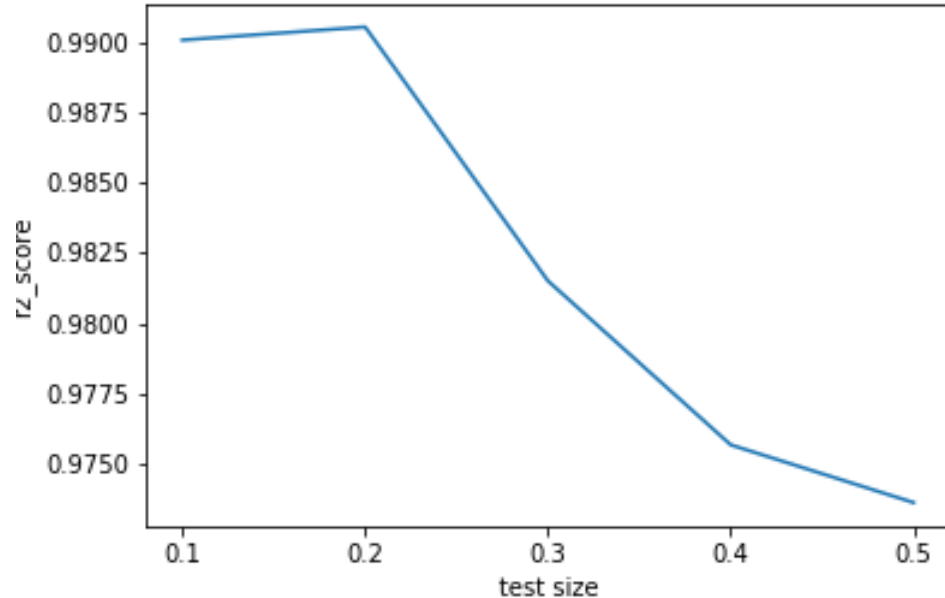
-0.23734929 for 9-12 hrs(day)

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

4.2 Gaussian Distribution

description The normal distribution is also called the Gaussian distribution or the bell curve distribution.

Continuous probability distributions are encountered in machine learning, most notably in the distribution of numerical input and output variables for models and in the distribution of errors made by models. Knowledge of the normal continuous probability distribution is also required more generally in the density and parameter estimation performed by many machine learning models. continuous probability distributions play an important role in applied machine learning and there are a few distributions that a practitioner must know about.

The distribution covers the probability of real-valued events from many different problem domains, making it a common and well-known distribution.

The distribution can be defined using two parameters:

Mean (μ): The expected value.

Variance (σ^2): The spread from mean.

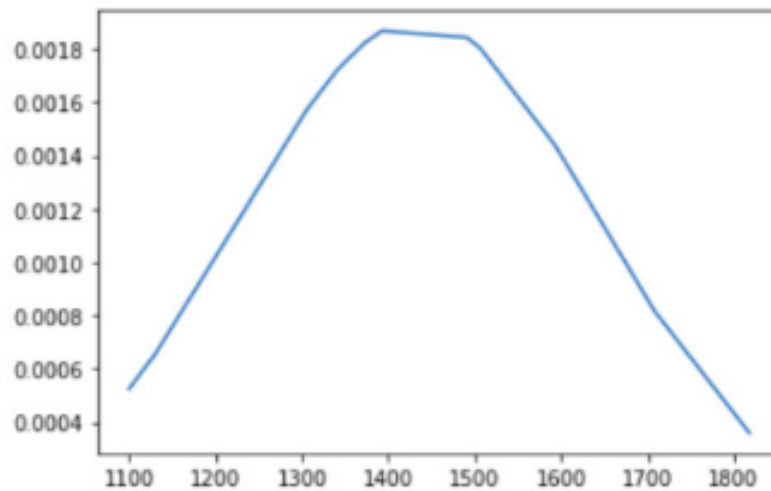
Standard Deviation (σ): The average spread from the mean.

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 ease of 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.

Graphical Representation :



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

References

1. <https://machinelearningmastery.com/continuous-probability-distributions-for-machine-learning/>
2. Dataset is taken from, <https://www.kaggle.com/vikasds101/road-accident-state-time>

[1] [2]