M S Ramaiah Institute of Technology
(An Autonomous Institute, Affiliated to VTU)

MSR nagar, MSRIT post, Bangalore-54

A Dissertation Report on

Social Impact of Efficient Data Analytics to Prevent the Rate of Accidents

Submitted by

Abhinav Shikhar	1MS12CS002
Aishwarya Kamraj	1MS12CS007
Gowri Ramaprasad	1MS12CS035
Naveen JS	1MS13CS416

in partial fulfillment for the award of the degree of

Bachelor of Engineering in Computer Science & Engineering

>

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

M.S.RAMAIAH INSTITUTE OF TECHNOLOGY
(Autonomous Institute, Affiliated to VTU)
BANGALORE-560054
www.msrit.edu, **May 2015**

Abstract

Utilizing Data analytics to minimize the rate of accidents in an area is the goal of our project. Past research has indicated that examining accidents individually is far less effective than examining a cluster of them that occurred in an area over time and follow a pattern. Our aim is to study the data supplied on the accident and investigation reports, which is tabulated into groups or categories. Though the similarities in any two accidents are not very evident, data collected over years and grouped does follow a pattern. Our aim is to utilize analytics techniques to uncover these patterns and draw helpful conclusions which would bring about some solutions to current crash related problems.

We have collected a particular state's accident reports of over 1, 00,000 and stored it in a database. After collection, data has been normalized for analysis. Each row in the data represents an accident, and each column is a characteristic of the event. A year worth of data from the database were loaded in R system to discover and visualize correlations and patterns in data and summarize key findings suggest possible solutions. R Charts have provided us with a way to identify patterns in data and visualize them in interactive graphs.

Contents

1 INTRODUCTION

General Introduction:

Vehicular data is a crucial in examining the cause and prevention of accidents, and analyzing the data for policy making and governance of road traffic. However, several challenges are presented in this analysis due to the dearth of accident data analysis and, examination of the cause and effects of this data. Therefore, researchers use common data mining and prediction models on the wealth of accident related information available. Performing trend analysis and recognizing particular patterns might evolve the responsiveness to accidents in metropolitan areas. Taking into account, the effects of accidents, this paper aims at predicting the nature of the future accidents. However, several challenges are presented in this analysis due to the dearth of accident data analysis and examination of the cause and effects of this data. Therefore, researchers use common data mining and prediction models on the wealth of accident related information available. Performing trend analysis and recognizing particular patterns might evolve the responsiveness to accidents in metropolitan areas. The dataset has close to 4.8 million data points and consists of attributes such as the city names, the type of accident, the condition of light, severity, speed zone, whether it was caused due to the consumption of alcohol, whether it is a hit and run case, whether it is attended by the police and so on. The idea is to use statistical analysis using algorithms such as k-means and decision trees to arrive at a substantial value of information. In this study, we aim to find common causes of accidents by mining the past accident data. The approach to this study places a focus on both quantitative and qualitative research.
A case-by-case approach attempts to examine the accidents as isolated events and examines the possible causes of the accident. This, approach is useful in collecting the data and creating the database and metadata. The second approach called as Statistical approach is used to analyze and observe patterns and common trends in accidents. In our research, we aim to focus heavily on the second approach since we are examining pre-recorded data. Statistical Methods is dealing with the analysis of causes of accidents in cities. Since, the causes and empirical analysis can lead to prevention of accidents; the concept of accident involvement risk suggests itself as a methodical framework for empirical accident causation studies.

Statement of the Problem:

The International Traffic Safety Data & Analysis Group (IRTAD) conducts surveys generating massive datasets with millions of data points. The rich dataset contains detailed information of approximately 4.8 million accidents occurred in a span of 5 years in Australia. It consists of the city names, the type of accident, condition of light, severity, speed zone, consumption of alcohol etc. The paper focuses on the application of data analytics to minimize the rate of accidents by studying, testing and training the past data. The objective of the project is to predict certain dependent data by using the independent factors on which it depends, uncover the pattern followed by the data collected over years and to draw helpful conclusions which would bring about some solutions to current crash related problems. Using some of the relevant attributes, demographic are plotted to bring out better insights.

In this study, we aim to find common causes of accidents by mining the accident data. The approach to this study places a focus on both quantitative and qualitative research. The following, are the common ways of approaching accident data:

- • Case-by-case approach: Accident causes attributed to registered accidents and road users involved by expert judgment.
- • Statistical approach: Accident causes as risk factors for accident involvement

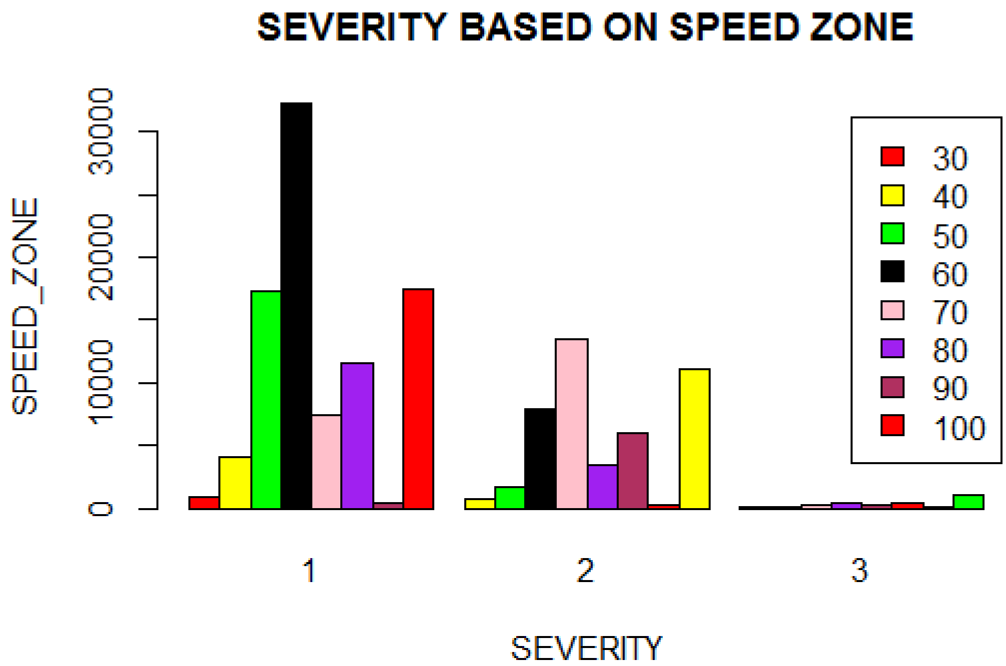
A case-by-case approach attempts to examine the accidents as isolated events and examines the possible causes of the accident. This, approach is useful in collecting the data and creating the database and metadata. The second approach is used to analyze and observe patterns and common trends in accidents.

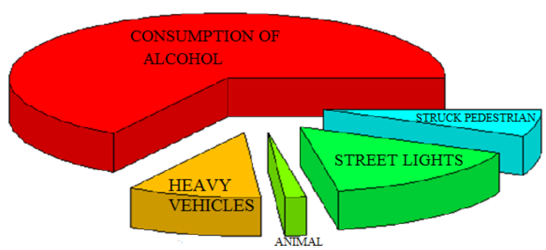
In our research, we aim to focus heavily on the second approach since we are examining pre-recoded data. "Statistical Methods "is dealing with the analysis of causes of accidents in cities. Since, the causes and empirical analysis can lead to prevention of accidents; the concept of accident involvement risk suggests itself as a methodical framework for empirical accident causation studies.

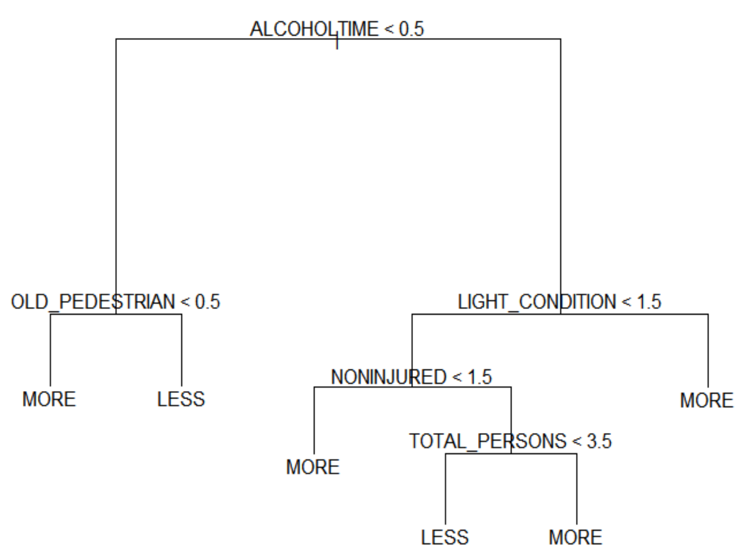
Objectives of the project:
The general objective was to observe a common order of events which lead to severe accidents and analysis pertaining to high-density or "accident prone" areas. We further want to examine the levels of severity in an accident to examine how accidents can be mitigated. We also examine, all the parties affected by the project such as pedestrians, pillion riders, car drivers and wildlife. We also examine the extent to which each party is affected and whether there is a correlation between cause and effect in the cases of their accidents. The intent of this paper is to examine and ponder on the solution to accidents and accident response in urban areas. We do so using common statistics and analytical tools. The R language has been used to mine the data for statistics. R is an open source tool and is highly extensible. The prepackaged library is readily available.
Modules have been created to deal with the huge datasets and to bring out unique insights into the results. The module involving the determination of the severity of the accidents, on the basis of certain factors such as the city, consumption of alcohol, speed zone, the condition of the light etc. The above prediction is done using Decision Tree analysis which generates a decision tree as a result of prediction. Because it's a programmable environment that uses command-line scripting, you can store a series of complex data-analysis steps in R.

Project deliverables:
The deliverables for each module would be its output in the form of info graphics such as

1. Bar Graphs
2. Facets
3. Pie Charts
4. Scatter Diagrams
5. Decision Trees (using Classification)


SEVERITY BASED ON SPEED ZONE
SPEED_ZONE
0
10000
20000
30000
1
2
3
SEVERITY
30
40
50
60
70
80
90
100


CONSUMPTION OF ALCOHOL
STRUCK PEDESTRIAN
STREET LIGHTS
HEAVY VEHICLES
ANIMAL


ALCOHOLTIME < 0.5
OLD_PEDESTRIAN < 0.5
LIGHT_CONDITION < 1.5
MORE
LESS
NONINJURED < 1.5
MORE
MORE
TOTAL_PERSONS < 3.5
LESS
MORE

Current Scope:

The analysis of the accident data is performed quantitatively and qualitatively. The sequences of conditions which lead to severe accidents are examined using the decision trees. Also, regions of high density accidents, the severity of the accident and the attributing factors were examined using the k-means clustering algorithm. The project in its early iteration can glean insight into locations which are accident prone and the causes for it. The secondary, goal of this research is to examine, how wildlife gets affected by road traffic. This predictive model can be implemented in several mobile applications which can minimize risk of accidents. It also attempts to research the possible causes of accident and the possible solutions to them. Further work can be done to build upon the insights that have been uncovered in this paper.

Future Scope:

This system aims to improve the approach to solving the method of collection and analysis of accident related information. This research attempts to reduce the average of time required by an ambulance to arrive at the spot of the accident and turnaround time with further data analysis. This predictive model can be implemented in several mobile applications which can minimize risk of accidents. This data can be used to implement protective measure to ensure road-safety of elderly citizens, wildlife and school children etc. We believe that this data can also be used, to build hospitals in locations so as to optimize response time and ensure timely arrival of ambulances, provide cab service to the people when they have consumed alcohol etc. The government can use this analysis to install more street lights along the positions where accidents are occurring in higher numbers during dawn, dusk or dark nights. This data can be used by policy-makers to design policies which ensure road safety. The optimal speed limit, alcohol consumption patterns and time of injury could be decided based on the analysis. This data can be used, by the department of forestry, wildlife etc., for deeper analysis and scrutiny.

2. PROJECT ORGANIZATION

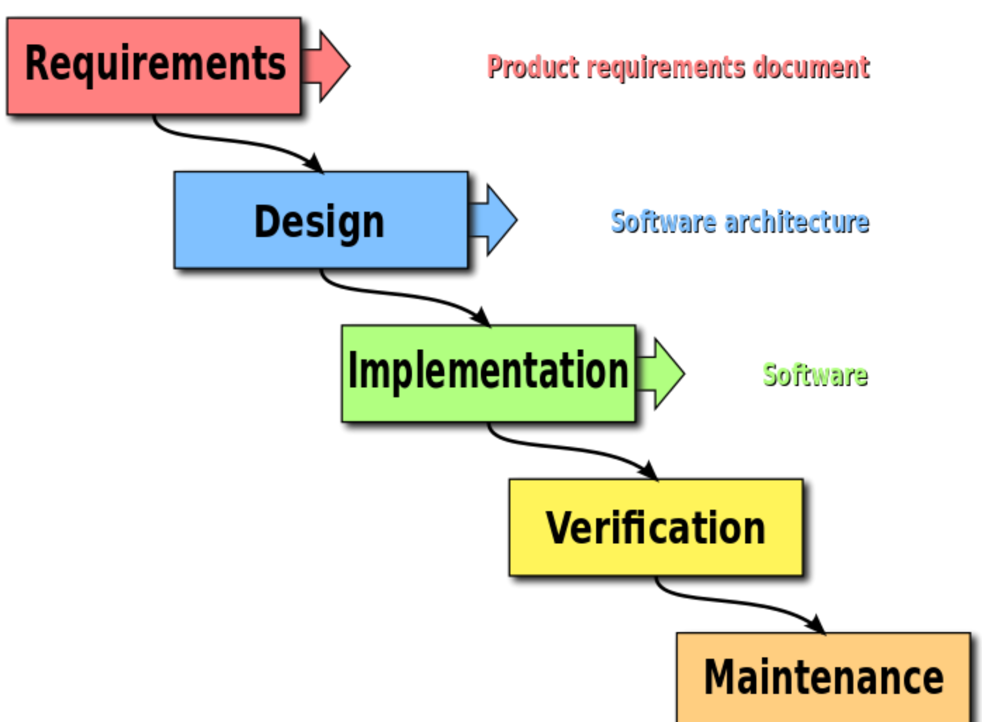
2.1. **Software Process Models** : In our work we used the simplified version of the waterfall model as our software process model and then proceeded to use peer programming.

WATERFALL MODEL:
The **waterfall model** is a sequential design process, used in software development processes, in which progress is seen as flowing steadily downwards (like a waterfall) through the phases of conception, initiation, analysis, design, construction, testing, production/implémentation and maintenance.

The waterfall development model originates in the manufacturing and construction industries: highly structured physical environments in which after-the-fact changes are prohibitively costly, if not impossible. Since no formal software development methodologies existed at the time, this hardware-oriented model was simply adapted for software development.

PAIR PROGRAMMING:
Pair programming (sometimes referred to as peer programming) is an agile software development technique in which two programmers work as a pair together on one workstation. One, the driver, writes code while the other, the observer, pointer or navigator, reviews each line of code as it is typed in. The two programmers switch roles frequently.
While reviewing, the observer also considers the "strategic" direction of the work, coming up with ideas for improvements and likely future problems to address. This frees the driver to focus all of his or her attention on the "tactical" aspects of completing the current task, using the observer as a safety net and guide.

2.2. **Roles and Responsibilities :** The roles were assigned on the basis of the waterfall model, and weekly meetings were held to address the challenges and issues faced in the project. Routine reports were generated and discussed with our guide to solve the issues.



3. LITERATURE SURVEY

3.1 Introduction : It is noteworthy to mention the sparse availability of accident data in major indian cities however, the literature available on accident data was organized in this manner:

- ➡ **Camera matching**: Camera matching uses accident scene photos that show various points of evidence. The technique uses CAD software to create a 3-dimensional model of the accident site and roadway surface. All survey data and photos are then imported into a three dimensional software package like 3D Studio Max. A virtual camera can be then be positioned relative to the 3D roadway surface. Physical evidence is then mapped from the photos onto the 3D roadway to create a three dimensional accident scene drawing.
- ➡ **Photogrammetric**: Photogrammetric is used to determine the three-dimensional geometry of an object on the accident scene from the original two dimensional photos. The photographs can be used to extract evidence that may be lost after the accident is cleared. Photographs from several viewpoints are imported into software like PhotoModeler. The forensic engineer can then choose points common to each photo. The software will calculate the location of each point in a three dimensional coordinate system.
- ➡ **Rectification**: Photographic rectification is also used to analyze evidence that may not have been measured at the accident scene. Two dimensional rectification transforms a single photograph into a top-down view. Software like PC-Rect can be used to rectify a digital photograph.

3.2 Main Body:

The common Accident analysis methods involve the following:

- Accident
- Debugging
- Failure mode and effects analysis
- Forensic engineering
- Forensic science
- Why-Because Analysis
- AcciMap Analysis

The timeline of accident data analysis is as follows:

- **1990's :** Nicholas Faith wrote a paper titled "Black Box: Why Air safety is a no accident zone" to examine and set the standards of safety in the aviation sector.
- **2000's :** A book titled "Enhanced Occupational Safety and Health" which attempted to examine the correlation between safety and road accident data was published. It proved insightful to policy makers in setting road safety measures.

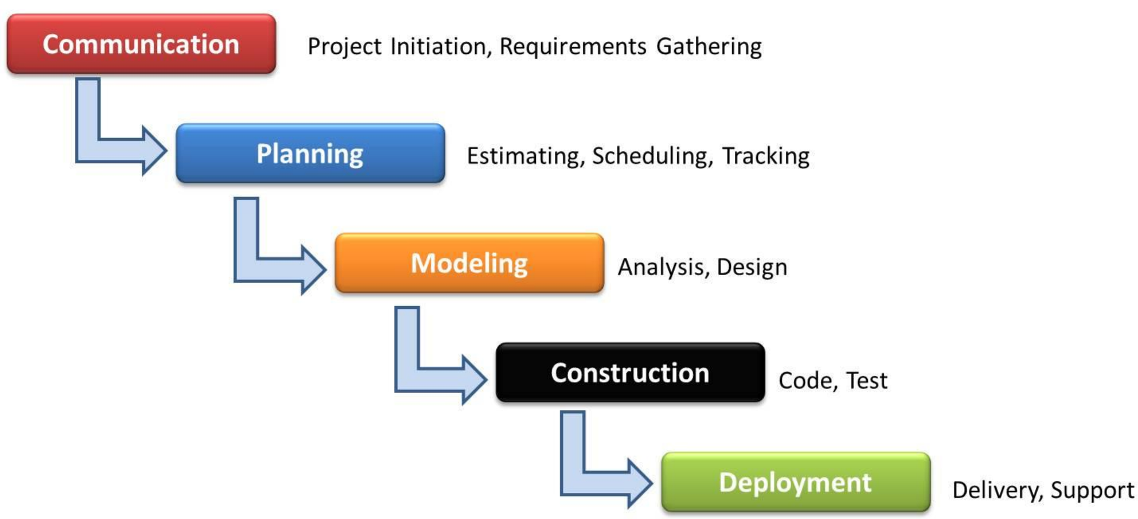
- **Late 2000's:** A survey was conducted by the US Department of Labor, Safety and Health Administration titled "accident investigation" to examine road accidents in the United States of America.

3.3 Conclusion of Survey: We thus, conclude that the literature survey, indicates a lack of accident data and prediction in India, and has tremendous potential for growth and further investigation.

4. SOFTWARE REQUIREMENT SPECIFICATIONS

Analysis gathers the requirements for the system. This stage includes a detailed study of the needs of the organization. Design focuses on high level design like, what programs are needed and how are they going to interact, low-level design (how the individual programs are going to work), interface design (what are the interfaces going to look like) and data design (what data will be required).

The main output of this phase is software specifications, which is the detail statement of the system function in order to achieve the objectives. Figure 2 shows the requirement analysis process.



The Waterfall Model: A Traditional Approach of SDLC

a. Identify the requirement

In order to get the information sources about the system requirement, several approaches are used to get user requirement as stated below such as interview, observation, document study and discussion.

b. Requirement Analysis

In this phase special tools and techniques help to make requirement determinations. Such tool used was the data flow diagram (DFD) to chart the input, process and the output of the system.

During this phase, the system analyst also analyzes the structured decisions made. Structured decisions were those for which the conditions, condition alternatives, actions, and action rules had been determined.

The output of this phase must be presented and documented in the simple way. Analysis approach is based on programming concept which called object oriented. Generally, this type of system model is assumed as the abstract of the developed system. Therefore, all the entities involve in the system must be presented in a model form.

c. Determination of Requirement

The next phase that the analyst enters is that of determining information requirements for the particular users involved. Among the tools used to define information requirements in the organizations are sampling and investigating hard data, interviewing, questionnaires, observing decision maker’s behavior and even prototyping. In this phase, the details of current system function are needed.

d. Requirement Specification

Requirement specification helps to analyze the requirement in details, so the specification and the determination is synchronized. Basically, requirement specification is presented in System Model that developed in requirement analysis process.

1. External Interface Requirements
    a. User Interfaces

        We are using the R Console which is user friendly.

    b. Software Interfaces

        We require a tool to analyze the collected data, currently using R Software. R includes virtually every data manipulation, statistical model, and chart that the modern data scientist could ever need. Representing complex data with charts and graphs is an essential part of the data analysis process, and R goes far beyond the traditional bar chart and line plot. Instead of using point-and-click menus or inflexible "black-box" procedures, R is a programming language designed expressly for data analysis. R scripts are easily automated, promoting both reproducible research and production deployments.

2. Functional Requirements

   The analysis project has no functional requirements.

3. Software System Attributes

   R system :

   a. is Reliable
   b. Known for Availability
   c. Provides Security
   d. Provides Portability
   e. Assures Maintainability
   f. Takes care of Performance

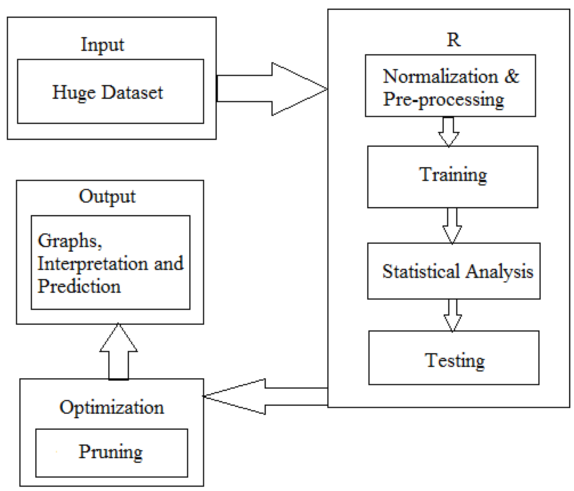
4. Database Requirement

   For accurate results in the final predictions and reports, the data must be sufficient (at least 1,00,000 tuples).

   - Dataset Identification
   - Understanding the Attributes of Data
   - Identifying the subset of attributes required for the task identified.
   - Dimensionality, Reduction,Normalization & Preprocessing of Data

5. DESIGN

The dataset is in .csv format and is read in MS-Excel, which is the input here. The R system is used to perform normalizing, pre-processing, distribution and other major analytic data steps. We then prepared a hypothetical analytic model plan using sample data, which is referred to testing in the system architecture. Some inferences are drawn from the results obtained from the testing data and are further generalized for the testing data. The statistical analysis are first performed on the training data and then on the testing data. The optimization is performed later in order to minimize the deviation of the obtained results. This deviation is measured in the terms of the mean values. After pruning of decision trees is performed, the optimal prediction and the decision tree is generated.



The project has many number of modules. Each of the graphs we are generating using R can be considered to be separate modules.

```
##
library(ggplot2)
library(MASS)
library(plotrix)

#read data
s10 <- read.csv(file = "C:/Users/sony/Desktop/R/PBL.csv")
# load the MASS package
#school = s10$ALCOHOL_RELATED  # the painter schools
school=s10$LGA_NAME
school.freq = table(school)   # apply the table function
#Then we apply the barplot function to produce its bar
graph.
colors = c("red", "yellow", "green", "violet","orange",
"blue", "pink", "cyan")
barplot(school.freq, col=colors)
#barplot(school.freq)          # apply the barplot function
title("CITY Vs Number of accidents")
```

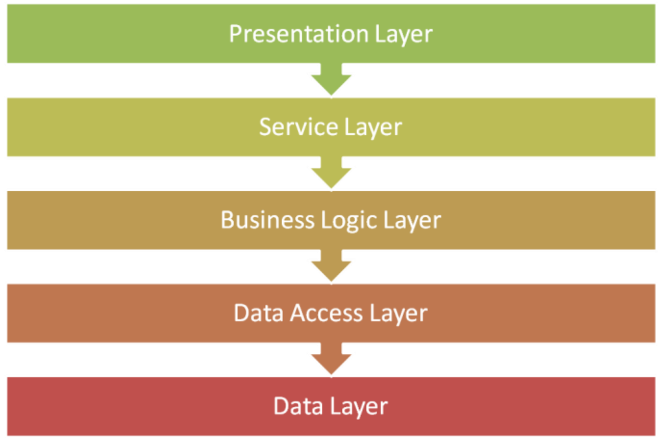
```
#read data
s10 <- read.csv(file = "C:/Users/sony/Desktop/R/PBL.csv")
LIGHT_CONDITION.freq=with(s10,table(LIGHT_CONDITION))
pie(LIGHT_CONDITION.freq)
pie(LIGHT_CONDITION.freq,col=c("orange","yellow","blue","gr
een","violet","brown","purple","pink"))
pie(LIGHT_CONDITION.freq,col=rainbow(8),radius=1,labels=nam
es(LIGHT_CONDITION.freq))

lbls=paste("\n\n",names(LIGHT_CONDITION.freq),"\n\n",LIGHT_
CONDITION.freq,sep="")
pie(LIGHT_CONDITION.freq,col=rainbow(8),radius=1,labels=lbl
s)
pie3D(LIGHT_CONDITION.freq,col=rainbow(8),explode=0.3,,labe
lcex=0.8)

#radial.pie(LIGHT_CONDITION.freq,col=rainbow(8),labels=lbls
,show.grid.labels=0)
title(" The number of accidents occured in different phases
of the day\n 1- Dark with no street lights\t 2-Dusk \t 3-
Dawn \t 4-Unknown \t 5- Day \n ")



#
library(ggplot2)
library(MASS)
library(plotrix)

#read data
s10 <- read.csv(file = "C:/Users/sony/Desktop/R/PBL.csv")
# load the MASS package
school = s10$OLD  # the painter schools
school.freq = table(school)   # apply the table function
#Then we apply the barplot function to produce its bar
graph.

colors = c("red", "yellow", "green", "violet","orange",
"blue", "pink", "cyan")
barplot(school.freq, col=colors)
#barplot(school.freq)         # apply the barplot function
title("CITY Vs Number of OLD PEOPLE")

school1 = s10$YOUNG  # the painter schools
school1.freq = table(school1)   # apply the table function
#Then we apply the barplot function to produce its bar
graph.

colors = c("red", "yellow", "green", "violet","orange",
"blue", "pink", "cyan")
barplot(school1.freq, col=colors)
#barplot(school1.freq)         # apply the barplot function
title("CITY Vs Number of YOUNG PEOPLE")


##
```

```
library(ggplot2)
library(MASS)
library(plotrix)

#read data
s10 <- read.csv(file = "C:/Users/sony/Desktop/R/PBL.csv")
# load the MASS package
#school = s10$HIT_AND_RUN  # the painter schools
#school.freq = table(school)   # apply the table function
#Then we apply the barplot function to produce its bar
graph.

#colors = c("red", "yellow", "green", "violet","orange",
"blue", "pink", "cyan")
#barplot(school.freq,  col=colors)
#barplot(school.freq)        # apply the barplot function
#title("CITY Vs Number of accidents")

counts <- table(s10$HIT_AND_RUN)
barplot(counts, main="number of hit aND RUN CASES",
xlab="COUNT")

#read data
s10 <- read.csv(file = "C:/Users/sony/Desktop/R/PBL.csv")
LONGITUDE.freq=with(s10,table(LONGITUDE))
#pie(LONGITUDE.freq)
#pie(LONGITUDE.freq,col=c("orange","yellow","blue","green",
"violet","brown","purple","pink"))
#pie(LONGITUDE.freq,col=rainbow(8),radius=1,labels=names(LO
NGITUDE.freq))

#lbls=paste("\n\n",names(LONGITUDE.freq),"\n\n",LONGITUDE.f
req,sep="")
#pie(LONGITUDE.freq,col=rainbow(8),radius=1,labels=lbls)
pie3D(LONGITUDE.freq,col=rainbow(8),explode=0.3,,labelcex=0
.8)

#radial.pie(LONGITUDE.freq,col=rainbow(8),labels=lbls,show.
grid.labels=0)
title(" The number of accident and the latitude \n ")


#read data
s10 <- read.csv(file = "C:/Users/sony/Desktop/R/PBL.csv")
LATITUDE.freq=with(s10,table(LATITUDE))
#pie(LATITUDE.freq)
#pie(LATITUDE.freq,col=c("orange","yellow","blue","green","
violet","brown","purple","pink"))
#pie(LATITUDE.freq,col=rainbow(8),radius=1,labels=names(LAT
ITUDE.freq))

#lbls=paste("\n\n",names(LATITUDE.freq),"\n\n",LATITUDE.fre
q,sep="")
#pie(LATITUDE.freq,col=rainbow(8),radius=1,labels=lbls)
pie3D(LATITUDE.freq,col=rainbow(8),explode=0.3,,labelcex=0.
8)
```

```
#radial.pie(LATITUDE.freq,col=rainbow(8),labels=lbls,show.g
rid.labels=0)
title(" The number of accidents and the latitude \n ")



data <- read.csv(file = "C:/Users/sony/Desktop/R/PBL.csv")
plot(x=data$LGA_NAME,  y=data$ALCOHOL_RELATED,  xlab  =  "SL
(mm)", ylab = "BD (mm)", pch=data$ALCOHOL_RELATED)


axis(side=1,                lwd=3,                xpd=TRUE,
at=c(min(data$LGA_NAME):max(data$LGA_NAME)))
axis(side=2,                lwd=3,                xpd=TRUE,
at=c(min(data$ALCOHOL_RELATED):max(data$ALCOHOL_RELATED)))


#read data
s10 <- read.csv(file = "C:/Users/sony/Desktop/R/PBL.csv")

SPEED_ZONE.freq=with(s10,table(SPEED_ZONE))
pie(SPEED_ZONE.freq)
pie(SPEED_ZONE.freq,col=c("orange","yellow","blue","green",
"violet","brown","purple","pink"))
pie(SPEED_ZONE.freq,col=rainbow(8),radius=1,labels=names(SP
EED_ZONE.freq))

lbls=paste("    \n         ",names(SPEED_ZONE.freq),"    \n
",SPEED_ZONE.freq,sep="")
pie(SPEED_ZONE.freq,col=rainbow(8),radius=1,labels=lbls)
pie3D(SPEED_ZONE.freq,col=rainbow(8),explode=0.3,,labelcex=
0.8)

#radial.pie(SPEED_ZONE.freq,col=rainbow(8),labels=lbls,show
.grid.labels=0)
title("  \n\n  The   number   of   accidents   and   the   speed
limits(in km/hr) \n\n ")



#read data
s10 <- read.csv(file = "C:/Users/sony/Desktop/R/PBL.csv")
DAY_OF_WEEK.freq=with(s10,table(DAY_OF_WEEK))
#pie(DAY_OF_WEEK.freq)
#pie(DAY_OF_WEEK.freq,col=c("orange"))
#pie(DAY_OF_WEEK.freq,col=rainbow(8),radius=1,labels=names(
DAY_OF_WEEK.freq))
#lbls=paste("\n\n",names(DAY_OF_WEEK.freq),"\n\n",DAY_OF_WE
EK.freq,sep="")
#pie(DAY_OF_WEEK.freq,col=rainbow(8),radius=1,labels=lbls)
#pie3D(DAY_OF_WEEK.freq,col=rainbow(8),explode=0.3,,labelce
x=0.8)

radial.pie(DAY_OF_WEEK.freq,col=rainbow(8),,)
title(" The  number  of  accidents  occured  on  different  days
of  the  week  \n  MONDAY  (DARK  BLUE)  TO  SUNDAY(LEFT  OF
MONDAY)")
```

2. Architecture Design

- The architectural design is the design of the entire software system; it gives a high-level overview of the software system, it provides information on the decomposition of the system into modules (classes), dependencies between modules, hierarchy and partitioning of the software modules.



3. A **data access layer** (**DAL**) in computer software, is a layer of a computer program which provides simplified access to data stored in persistent storage of some kind, such as an entity-relational database. This acronym is prevalently used in Microsoft ASP.NET environments.

4. For example, the DAL might return a reference to an object (in terms of object-oriented programming) complete with its attributes instead of a row of fields from a database table. This allows the client (or user) modules to be created with a higher level of abstraction. This kind of model could be implemented by creating a class of data access methods that directly reference a corresponding set of database stored procedures. Another implementation could potentially retrieve or write records to or from a file system. The DAL hides this complexity of the underlying data store from the external world.

5. For example, instead of using commands such as *insert*, *delete*, and *update* to access a specific table in a database, a class and a few stored procedures could be created in the database. The procedures would be called from a method inside the class, which would return an object containing the requested values. Or, the insert, delete and update commands could be executed within simple functions.

6. IMPLEMENTATION

After identifying the problem and the need for a system design, we have aggregated enough data to draft an analytic plan and reviewed it among peers for the correctness of said data. Followed by this was preparing the data for analysis in R system, which included Normalizing, pre-processing, distribution and other major analytic data steps. We then prepared a hypothetical analytic model plan using sample data that can be refined for actual data analysis.

Our final step in design was identifying the appropriate algorithms such as K-means, decision trees on raw data, that would help create infographics or do statistical analysis to achieve our initial hypothetical goals. Using R-Console, we generate visual images like pie charts, bar graphs etc. Some of the R packages which we need to implement our project are GDATA, CRAN, MASS, GGPLOT2, PLOTRIX etc.

In decision analysis a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.

A decision tree consists of 3 types of nodes:

- Decision nodes - commonly represented by squares
- Chance nodes - represented by circles
- End nodes - represented by triangles

Decision trees are commonly used in operations research and operations management. If in practice decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as a best choice model or online selection model algorithm. Another use of decision trees is as a descriptive means for calculating conditional probabilities. Decision trees, influence diagrams, utility functions, and other decision analysis tools and methods are taught to undergraduate students in schools of business, health economics, and public health, and are examples of operations research or management science methods.

Advantages and disadvantages

- Among decision support tools, decision trees (and influence diagrams) have several advantages. Decision trees:
- Are simple to understand and interpret. People are able to understand decision tree models after a brief explanation.
- Have value even with little hard data. Important insights can be generated based on experts describing a situation (its alternatives, probabilities, and costs) and their preferences for outcomes.
- Allow the addition of new possible scenarios
- Help determine worst, best and expected values for different scenarios
- Use a white box model. If a given result is provided by a model.
- Can be combined with other decision techniques.

Disadvantages of decision trees:

- For data including categorical variables with different number of levels, information gain in decision trees are biased in favor of those attributes with more levels.
- Calculations can get very complex particularly if many values are uncertain and/or if many outcomes are linked.

k-means clustering is a method of vector quantization, originally from signal processing, that is popular forcluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

The problem is computationally difficult (NP-hard); however, there are efficient heuristic algorithms that are commonly employed and converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Additionally, they both use cluster centers to model the data; however, k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

The algorithm has a loose relationship to the k-nearest neighbor classifier, a popular machine learning technique for classification that is often confused with k-means because of the k in the name. One can apply the 1-nearest neighbor classifier on the cluster centers obtained by k-means to classify new data into the existing clusters. This is known as nearest centroid classifier or Rocchio algorithm.

Description

Given a set of observations (x1, x2, …, xn), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k (≤ n) sets S = {S1, S2, …, Sk} so as to minimize the within-cluster sum of squares (WCSS) (sum of distance functions of each point in the cluster to the K center). In other words, its objective is to find:

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2$$

where μi is the mean of points in Si.

The decision tree algorithm was used to discern the cause of accidents in metropolitan areas. The alcohol time was used as the root, and light condition and age of pedestrian were used as the right and left nodes, and the number of injuries was the child of the light condition. We were able to discern, that there was a high incidence of injuries when a particular path was high-lighted. When the person was under the influence of alcohol and the light conditions were low, the number of deaths averaged at 3.5. Furthermore, the age of the pedestrian played an important role in the cause of injuries.

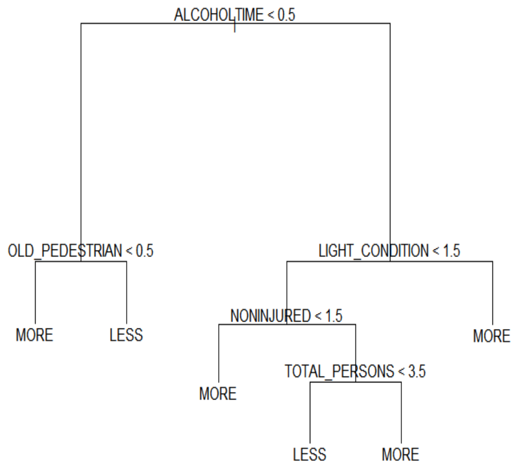
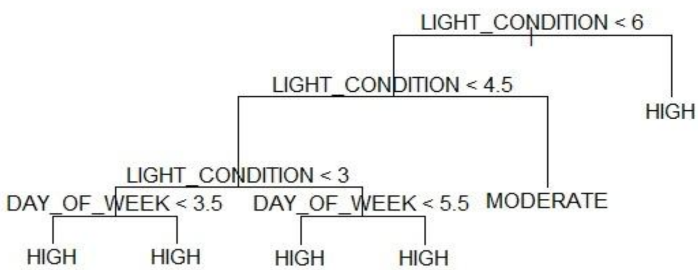


Figure 1: Decision Tree based on the time of Alcohol ingestion, Age of Pedestrians and the Light conditions.

When investigated further, the light condition based decision tree indicates the number of accidents according to the light condition. It was observed that the optimal light condition was during dawn and night time with street lights, where accidents were of a moderate quantity. However, the number of accidents during total darkness and harsh sun were higher than the average.



.

Figure 2: Decision Tree based on Light conditions and average accidents, and day of week.

The k-means algorithm was used to determine high-density accident zones in the country. The x-axis represents twenty-five cities and the y-axis counted the total number of persons who were in accidents. It was concluded that the region of cities between (20-25) experienced a higher density of accidents. The seed for the k-means algorithm in this zone was (purple dot city). The k-means algorithm was conclusive in determining the accidents density however; this method did not describe the quality or severity of the accident. Further analysis on the severity of the accident was required.

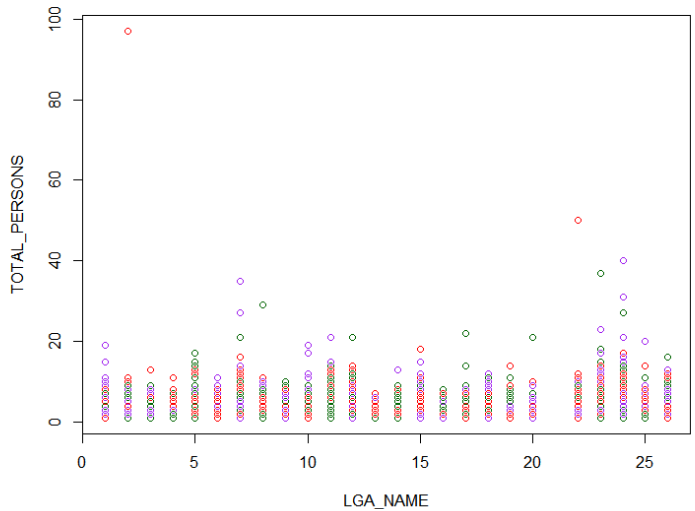


Figure 3 K-mean Clustering with X-axis indicating the city and the Y-axis indicating the number of accidents

In the severity analysis, we examined three factors: Severity of accidents based on alcohol consumption, Severity of accidents based on light condition and Severity of accidents based on the speed-zone. The results gave a deeper insight into the factors that caused accidents. The severity of alcohol related accidents averaged around 40,000 and it drastically reduced when there was no alcohol consumed by the injured party. Subsequently, accidents were higher and more severe, when there was pitch-darkness and street lights were absent (Greater than 40,000). In stark contrast, night-drives with street lights reduced the number of severe accidents to less than 10,000. Lastly, the number of accidents were higher when the speed limit was around 60 kmph and reduced to less than 10,000 when the speed was less than 40 kmph. An aberrant observation was the number of accident when the speed limit was 30kmph. There were, an increasing severity when the number of deaths were more than 20,000.

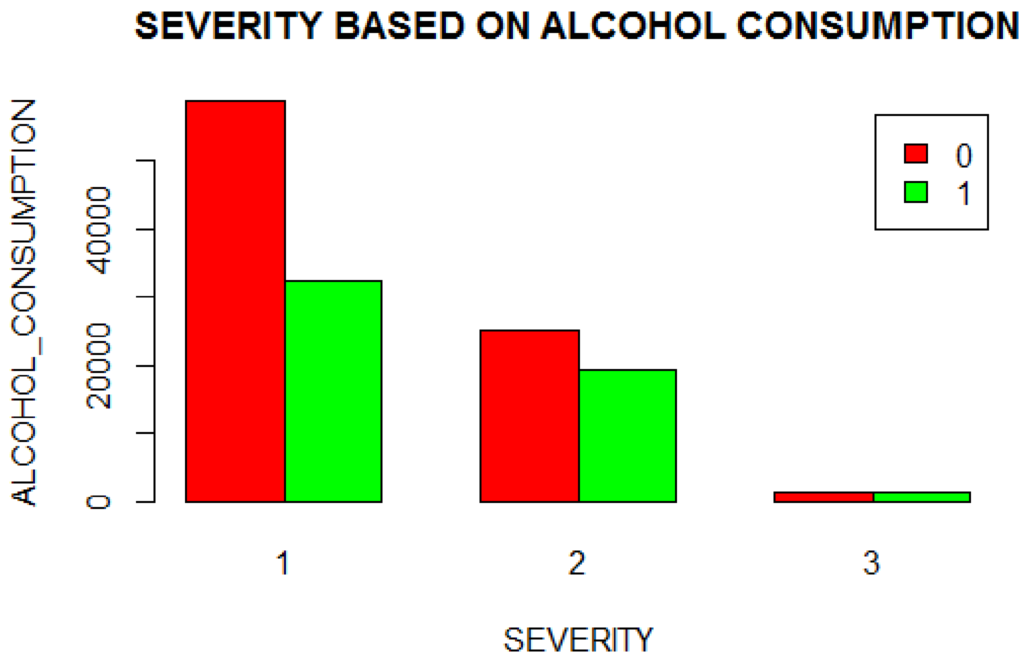


Figure 4: Severity analysis 1

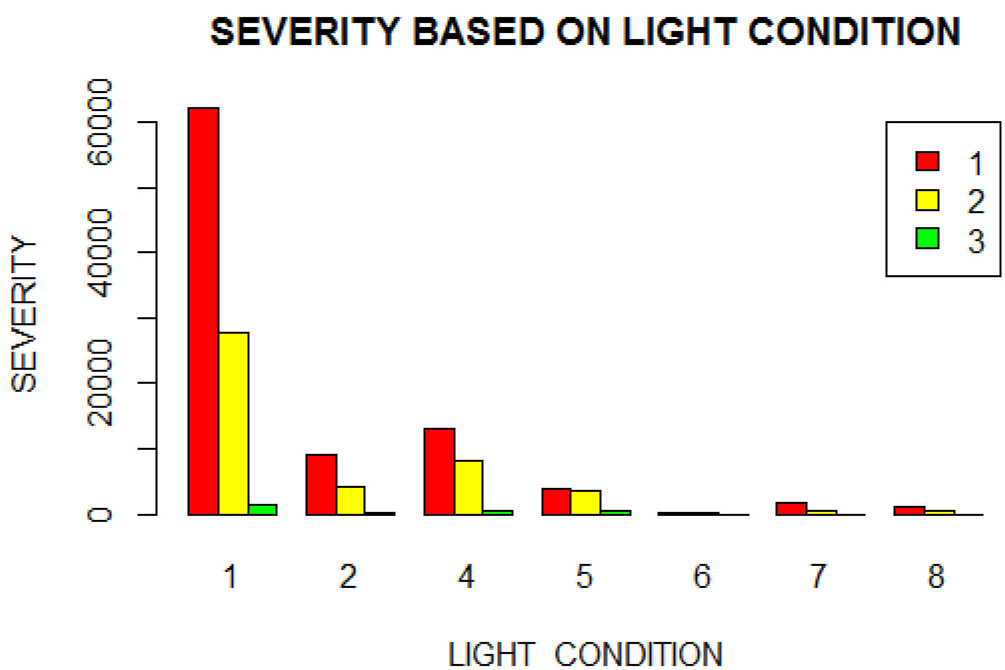
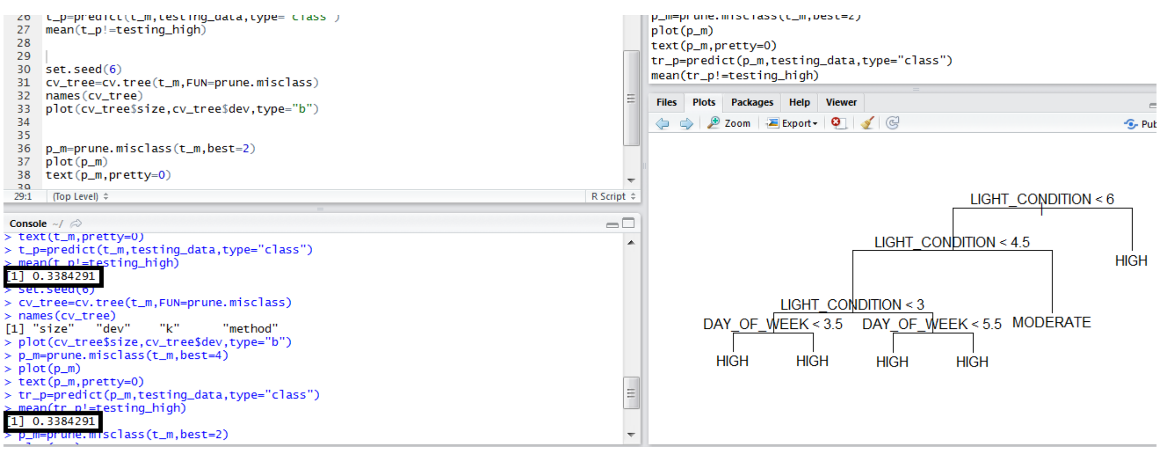


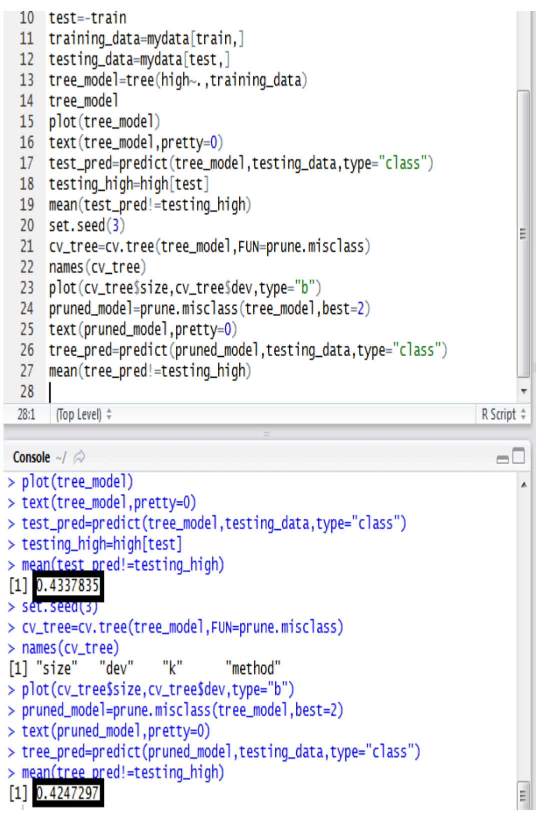
Figure 5: Severity Analysis 2

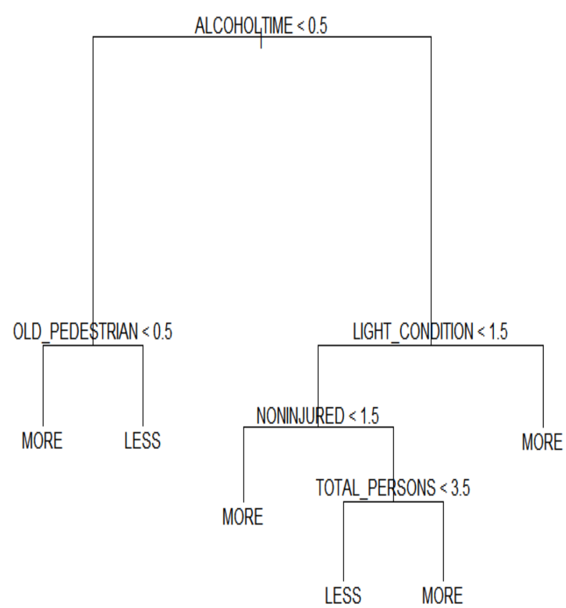
7. CODE AND TESTING

```
library(ISLR)
library(tree)

mydata<-read.csv("C:/Users/sony/Desktop/PBL Data/data_now.csv")
attach(mydata)
head(mydata)

range(SEVERITY)
high=ifelse(SEVERITY>=2,"MODERATE","HIGH")
mydata=data.frame(mydata,high)
names(mydata)
mydata=mydata[,-4]
names(mydata)
set.seed(3)
train=sample(1:nrow(mydata),nrow(mydata)/1000)
test=-train
training_data=mydata[train,]
testing_data=mydata[test,]
testing_high=high[test]


t_m=tree(high~.,training_data)
plot(t_m)
text(t_m,pretty=0)

t_p=predict(t_m,testing_data,type="class")
mean(t_p!=testing_high)


set.seed(6)
cv_tree=cv.tree(t_m,FUN=prune.misclass)
names(cv_tree)
plot(cv_tree$size,cv_tree$dev,type="b")


p_m=prune.misclass(t_m,best=2)
plot(p_m)
text(p_m,pretty=0)

tr_p=predict(p_m,testing_data,type="class")
mean(tr_p!=testing_high)
```



```
library(ISLR)
library(tree)
mydata=read.csv("C:/Users/Naveen/Desktop/final.csv")
high=ifelse(mydata$DAY_OF_WEEK>=4,"MORE","LESS")
mydata=data.frame(mydata,high)
names(mydata)
mydata=mydata[-3]
set.seed(5)
train=sample(1:nrow(mydata),nrow(mydata)/265)
test=-train
training_data=mydata[train,]
testing_data=mydata[test,]
tree_model=tree(high~.,training_data)
tree_model
plot(tree_model)
text(tree_model,pretty=0)
test_pred=predict(tree_model,testing_data,type="class")
testing_high=high[test]
mean(test_pred!=testing_high)
set.seed(3)
cv_tree=cv.tree(tree_model,FUN=prune.misclass)
names(cv_tree)
plot(cv_tree$size,cv_tree$dev,type="b")
pruned_model=prune.misclass(tree_model,best=2)
text(pruned_model,pretty=0)
tree_pred=predict(pruned_model,testing_data,type="class")
mean(tree_pred!=testing_high)
```





The decision tree algorithm makes use of entropy which is calculated using a formula, specified earlier. Using the value of the entropy, we find out the largest information gain value. The information gain represents nothing but the extent to which the prediction of the event happens correctly using the method or algorithm used for training the machine. The comparison of the values of the means of the deviations for the module C, when the pruning was done and when it was done is shown in Table 1.

Module	**Number of Data points**	**Pruning Done**	**Mean of Deviations**
C	4.28 million	No	0.4337385
C	4.28 million	Yes	0.4247297

Initially, when there was no pruning introduced to the decision tree, the mean of the deviations was almost 0.433 as in Fig 10.

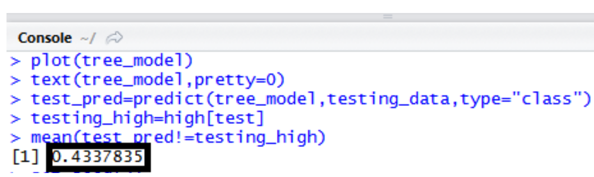


Fig 10. Before pruning

The pruning basically shows us a graph using which we can decide the numbers of levels to be used for the decision making. Usually, the number of levels for the pruning lies in the middle of the range. Here, we use the value of 2 for the pruning, which means that it would generate a new tree based on the new level value.

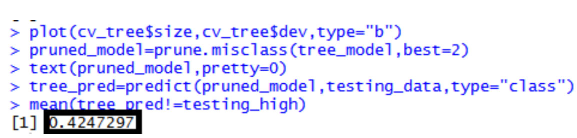


Fig 11. After pruning

After pruning is performed, the value of mean of the deviations has fallen down to 0.424. Pruning causes a reduction in the value of the mean of the deviations , which means that it adds accuracy to the prediction and makes the model fit to the data better.

8. CONCLUSION & SCOPE FOR FUTURE WORK

Through our research on accident statistics in major cities we were able to form conclusions about the patterns. With this body of work, we were thus able to analyze the accident data quantitatively and qualitatively. Using, the decision trees, we were able to examine the sequence of conditions, which lead to severe accidents. The clustering algorithm was able to indicate regions of high density accidents and the severity of the accident and the attributing factors were examined

This system aims to improve the approach to solving the method of collection and analysis of accident related information. This research attempts to reduce the average of time required by an ambulance to arrive at the spot of the accident and turnaround time with further data analysis. It is also attempts to research the possible causes of accident and the possible solutions to them. The secondary, goal of this research is to examine, how wildlife gets affected by road traffic. This predictive model can be implemented in several mobile applications which can minimize risk of accidents.

We believe that project and data analysis we have performed has the ability to significantly impact the society around us. The project in its early iteration can glean insight into locations which are accident prone and the causes for it. This data can be used by policy-makers to design policies which ensure road safety. The optimal speed limit, alcohol consumption patterns and time of injury are few of the attributes which we have examined. However, we believe that this data can also be used, to build hospitals in locations so as to optimize response time and ensure timely arrival of ambulances. Furthermore, our findings and analytics model can be incorporated in popular mobile applications which can design measure, to reduce accidents. Our secondary, goal was to discern the population which was most affected by road accidents. Thus, this data can be used to implement protective measure to ensure road-safety of elderly citizens, wildlife and school children etc., This data can be used, by the department of forestry, wildlife etc., for deeper analysis and scrutiny.

9. REFERENCES

1. Praticò Filippo.G (Reggio Calabria Mediterranean University), Vaiana Rosolino (Researcher, DIPITER Department Of Territorial Planning, University of Calabria-Arcavacata Campus (Cosenza)), Accident Data Analysis: An experimental Investigation for a rural, multilane, median separated, Italian Road.
2. Heinz Hautzinger, Claus Pastor, Manfred Pfeiffer, Jochen Schmidt, Analysis Methods for Accident and Injury Risk Studies.
3. Chen Lei, Xu Nuo, Analyzing Method of Traffic Accident Causation through Experts Method and Statistical Analysis .
4. A.K. Jain (Michigan State University),M.N. Murty (Indian Institute of Science) and P.J. Flynn (The Ohio State University), Data Clustering: A Review:
5. Nishant Mathur, Sumit Kumar, Santosh Kumar, and Rajni Jindal, The Base Strategy for ID3 Algorithm of Data Mining Using Havrda and Charvat Entropy Based on Decision Tree.
6. A Programming Environment for Data Analysis and Graphics Version 3.2.2 (2015-08-14), An Introduction to R Notes on R.
7. S S Dimov, and C D Nguyen (Manufacturing Engineering Centre, Cardiff University, Cardiff, UK), Selection of K in K-means clustering D T Pharm.
8. Kardi Teknomo,PhD, K-Means Clustering Tutorial.
9. MaxCameron Monash (University Accident Research Centre), Accident Data Analysis To Develop Target Groups Countermeasures Volume 1 :Methods and Conclusions.
10. MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297. MR 0214227. Zbl 0214.46201. Retrieved 2009-04-07.
11. Lloyd, S. P. (1957). "Least square quantization in PCM". Bell Telephone Laboratories Paper. Published in journal much later: Lloyd., S. P. (1982). "Least squares quantization in PCM" (PDF). IEEE Transactions on Information Theory 28 (2): 129–137. doi:10.1109/TIT.1982.1056489. Retrieved 2009-04-15.
12. Quinlan, J. R. (1987). "Simplifying decision trees". International Journal of Man Machine Studies 27 (3): 221. doi:10.1016/S0020-7373(87)80053-6.R.
13. Quinlan, "Learning efficient classification procedures", Machine Learning: an artificial intelligence approach, Michalski, Carbonell & Mitchell (eds.), Morgan Kaufmann, 1983, p. 463-482.
14. Utgoff, P. E. (1989). Incremental induction of decision trees. Machine learning, 4(2), 161-186.
15. Deng,H.; Runger, G.; Tuv, E. (2011). Bias of importance measures for multi-valued attributes and solutions. Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN).
16. http://www.ijarcsse.com/docs/papers/Volume_3/6_June2013/V3I6-0454.pdf
17. http://www.ijiee.org/papers/93-I118.pdf
18. http://research.ijcaonline.org/volume80/number7/pxc3891742.pdf
19. https://www.edx.org/course/introduction-r-programming-microsoft-dat204x-0
20. https://cran.r-project.org/doc/manuals/R-intro.pdf
21. https://www.cs.umd.edu/~mount/Projects/KMeans/pami02.pdf
22. https://www.ee.columbia.edu/~dpwe/papers/PhamDN05-kmeans.pdf