

**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

<b>Abhinav Shikhar</b>	1MS12CS002
<b>Aishwarya Kamraj</b>	1MS12CS007
<b>Gowri Ramaprasad</b>	1MS12CS035
<b>Naveen JS</b>	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](http://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

<i>Declaration</i>	<i>i</i>
<i>Acknowledgements</i>	<i>ii</i>
<i>Abstract</i>	<i>iii</i>
<i>List of Figures</i>	
<i>List of Tables</i>	
<b>1</b>	<b>INTRODUCTION</b>
1.1	General Introduction
1.2	Statement of the Problem
1.3	Objectives of the project
1.4	Project deliverables
1.5	Current Scope
1.6	Future Scope
<b>2</b>	<b>PROJECT ORGANIZATION</b>
2.1	Software Process Models
2.2	Roles and Responsibilities
<b>3</b>	<b>LITERATURE SURVEY</b>
3.1	...Introduction
3.2	...Main Body
3.3	Conclusion of Survey
<b>4</b>	<b>SOFTWARE REQUIREMENT SPECIFICATIONS</b>
4.1	External Interface Requirements
5.2.1	User Interfaces
5.2.2	Software Interfaces
4.2	Functional Requirements
4.3	Software System Attributes
4.3.1	Reliability
4.3.2	Availability
4.3.3	Security
4.3.4	Portability
4.3.5	Maintainability
4.3.6	Performance
4.4	Database Requirement

<b>5</b>	<b>DESIGN</b>
5.1	Introduction
5.2	Architecture Design
5.3	Graphical User Interface
5.4	Metric calculation
<b>6</b>	<b>IMPLEMENTATION</b>
6.1	Tools Introduction
6.2	Technology Introduction
6.3	Overall view of the project in terms of implementation
6.4	Explanation of Algorithm and how it is been implemented
6.5	Information about the implementation of Modules
<b>7</b>	<b>TESTING</b>
7.1	Results and Snapshots
<b>8</b>	<b>CONCLUSION &amp; SCOPE FOR FUTURE WORK</b>
<b>9</b>	<b>REFERENCES</b>

# 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-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.

### **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.

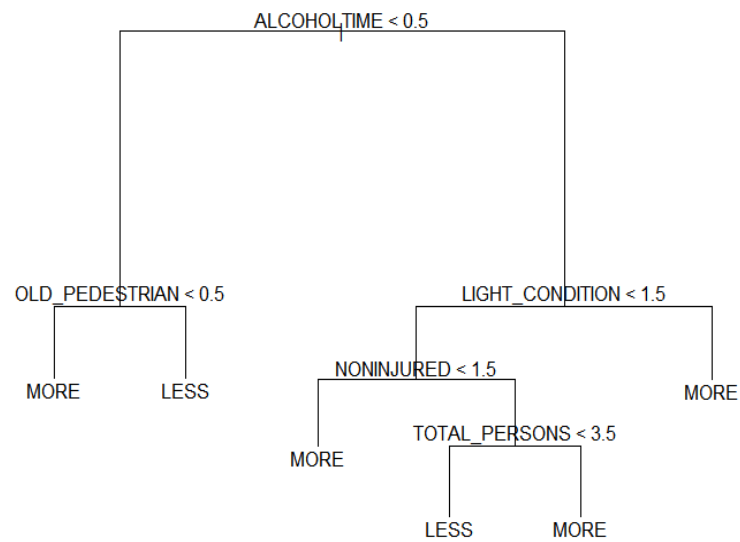
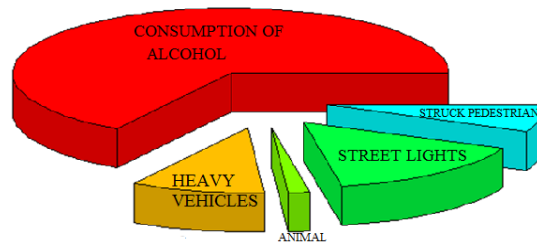
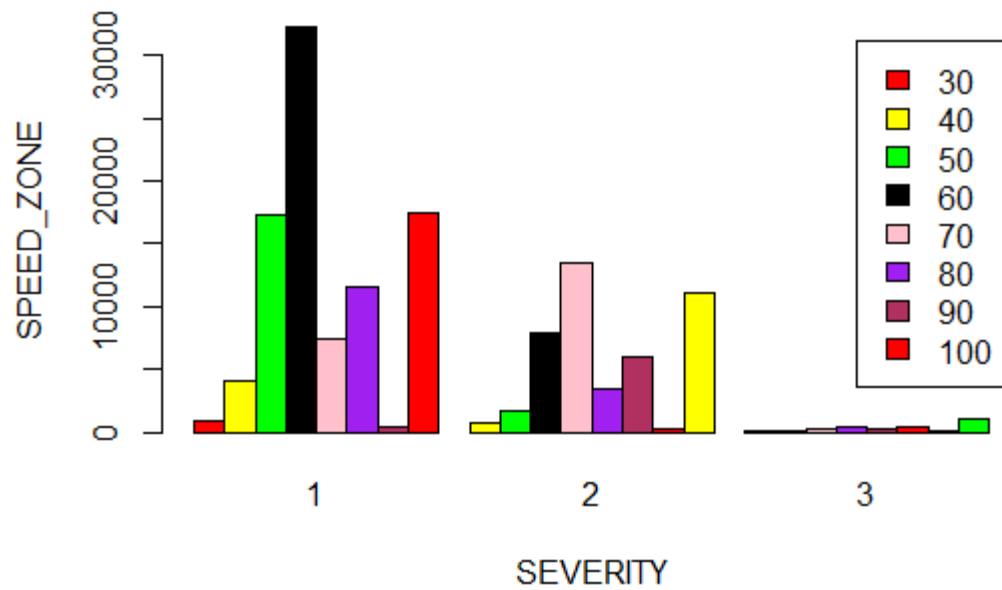
Modules have been created to deal with the huge datasets and to bring out unique insights into the results. The module involving the determination of the severity of the accidents, on the basis of certain factors such as the city, consumption of alcohol, speed zone, the condition of the light etc. The 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



**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.

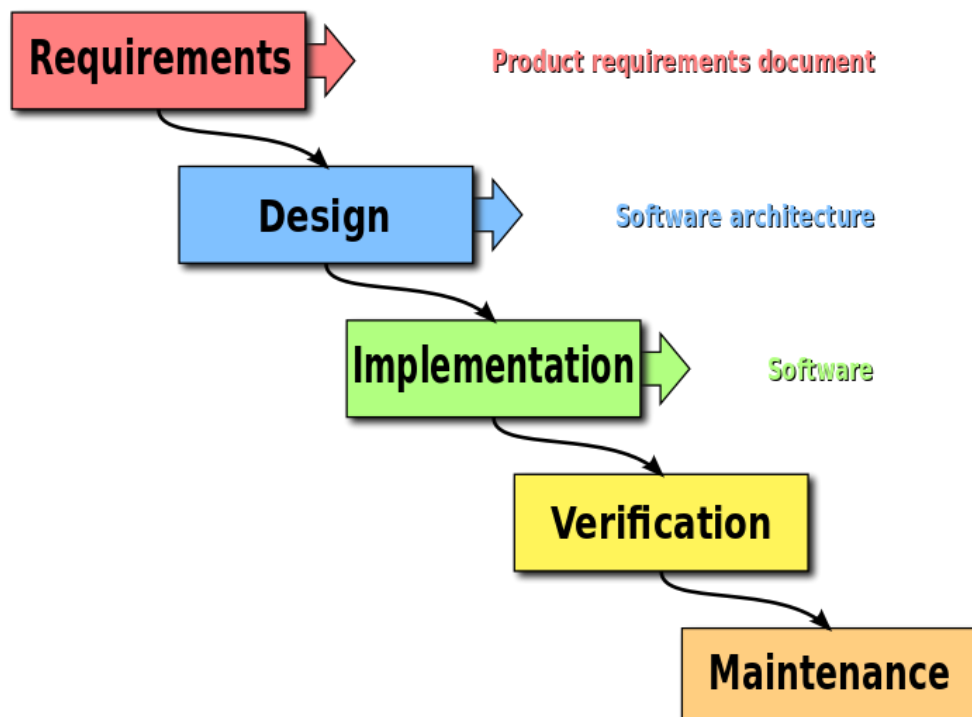
The waterfall development model originates in the manufacturing and construction industries: highly structured physical environments in which after-the-fact changes are prohibitively costly, if not impossible. Since no formal software development methodologies existed at the time, this hardware-oriented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

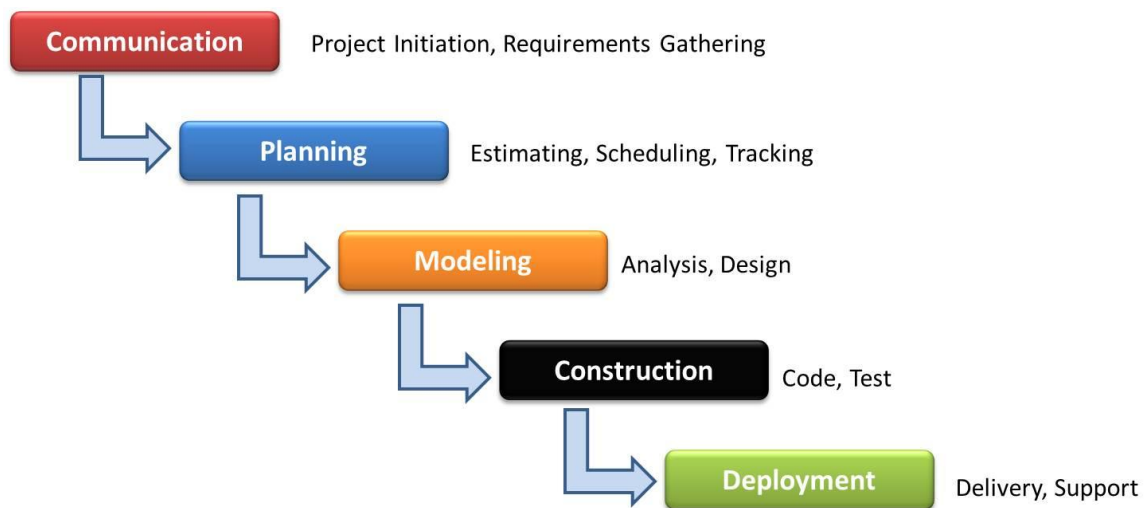
- **Late 2000's:** A survey was conducted by the US Department of Labor, Safety and Health Administration titled “accident investigation” to examine road accidents in the United States of America.

**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

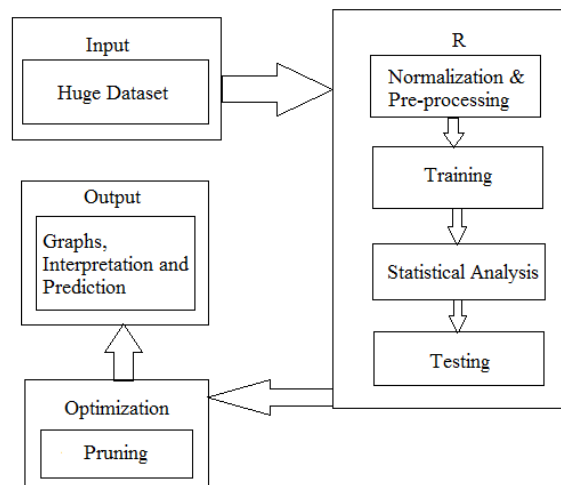
4. Database Requirement

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","green",
"violet","brown","purple","pink"))
pie(LIGHT_CONDITION.freq,col=rainbow(8),radius=1,labels=names(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=lbls)
pie3D(LIGHT_CONDITION.freq,col=rainbow(8),explode=0.3,,labelcex=0.8)

#radial.pie(LIGHT_CONDITION.freq,col=rainbow(8),labels=lbls,show.grid.labels=0)
title(" The number of accidents occurred 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(LONGITUDE.freq))

#lbls=paste("\n\n",names(LONGITUDE.freq),"\n\n",LONGITUDE.freq,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(LATITUDE.freq))

#lbls=paste("\n\n",names(LATITUDE.freq),"\n\n",LATITUDE.freq,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)
```

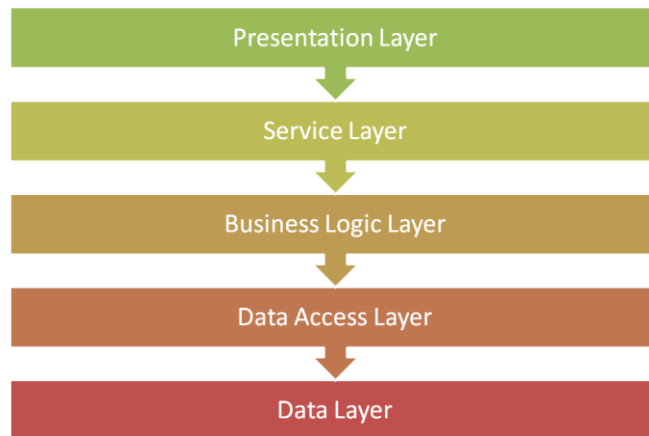
```
#radial.pie(SPEED_ZONE.freq,col=rainbow(8),labels=lbls,show
.grid.labels=0)
title(" \n\n The number of accidents and the speed
limits(in km/hr) \n\n ")
```

```
#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 occurred 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, GGLOT2, 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 for cluster 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 ( $x_1, x_2, \dots, x_n$ ), where each observation is a  $d$ -dimensional real vector, k-means clustering aims to partition the  $n$  observations into  $k$  ( $\leq n$ ) sets  $S = \{S_1, S_2, \dots, S_k\}$  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:

$$\arg \min_S \sum_{i=1}^k \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \mu_i\|^2$$

where  $\mu_i$  is the mean of points in  $S_i$ .

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.

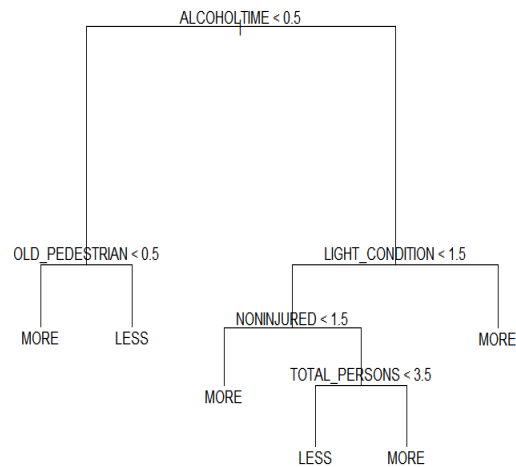


Figure 1: Decision Tree based on the time of Alcohol ingestion, Age of Pedestrians and the Light conditions.

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.

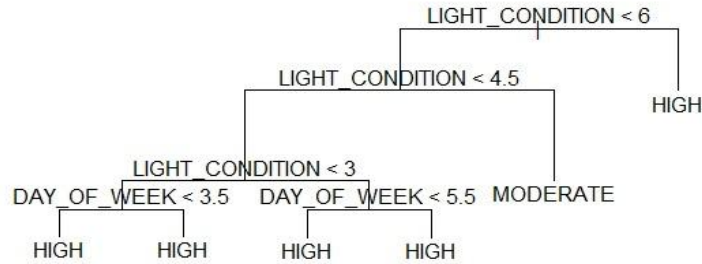


Figure 2: Decision Tree based on Light conditions and average accidents, and day of week.

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.

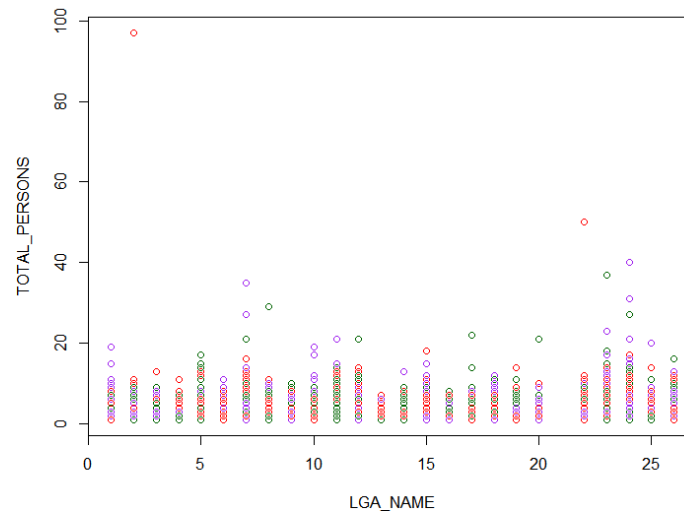


Figure 3 K-mean Clustering with X-axis indicating the city and the Y-axis indicating the number of accidents

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.

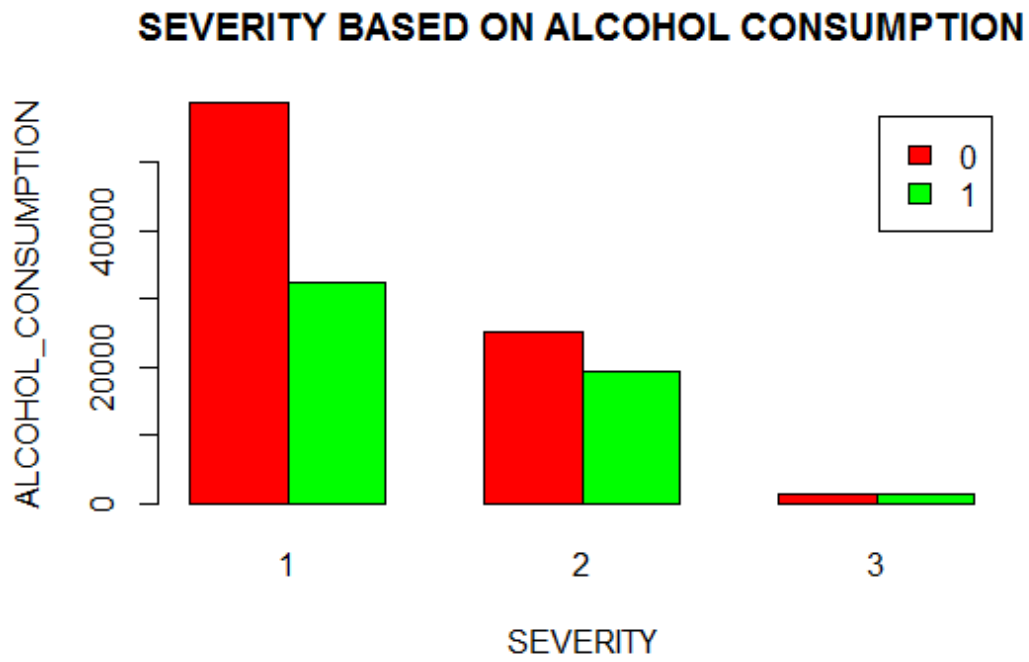


Figure 4: Severity analysis 1

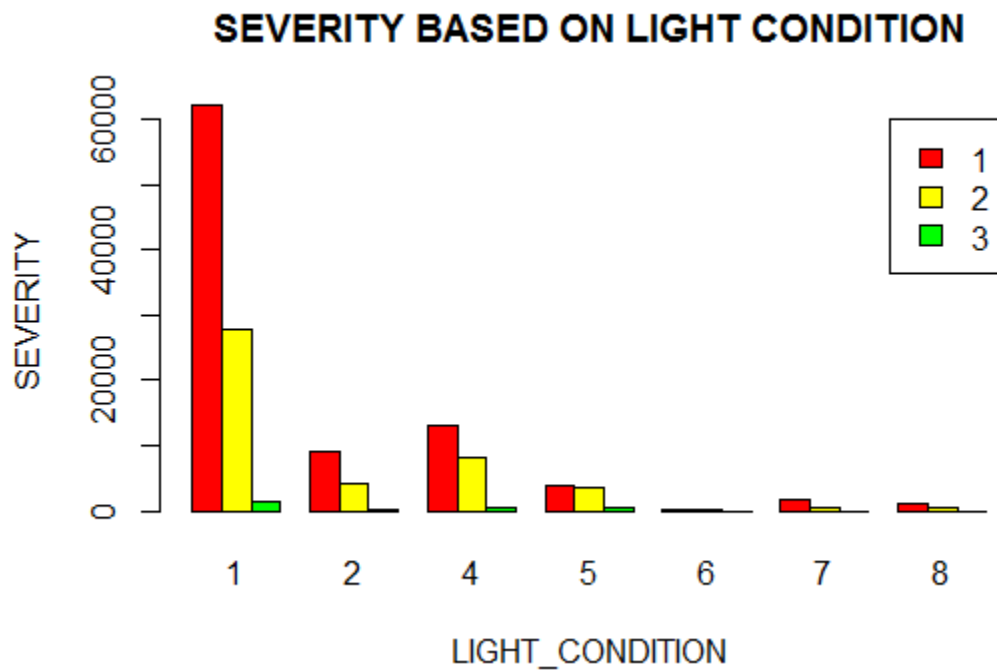


Figure 5: Severity Analysis 2

## 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)
```

```
20 t_p=predict(t_m,testing_data,type="class")
21 mean(t_p!=testing_high)
22
23
24
25
26
27
28
29
30
31 set.seed(6)
32 cv_tree=cv.tree(t_m,FUN=prune.misclass)
33 names(cv_tree)
34 plot(cv_tree$size,cv_tree$dev,type="b")
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2590
2591
2592
2593
2594
2595
2596
2597
2598
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
261
```



```

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)

```

```

10 test=-train
11 training_data=mydata[train,]
12 testing_data=mydata[test,]
13 tree_model=tree(high~.,training_data)
14 tree_model
15 plot(tree_model)
16 text(tree_model,pretty=0)
17 test_pred=predict(tree_model,testing_data,type="class")
18 testing_high=high[test]
19 mean(test_pred!=testing_high)
20 set.seed(3)
21 cv_tree=cv.tree(tree_model,FUN=prune.misclass)
22 names(cv_tree)
23 plot(cv_tree$size,cv_tree$dev,type="b")
24 pruned_model=prune.misclass(tree_model,best=2)
25 text(pruned_model,pretty=0)
26 tree_pred=predict(pruned_model,testing_data,type="class")
27 mean(tree_pred!=testing_high)
28 |

```

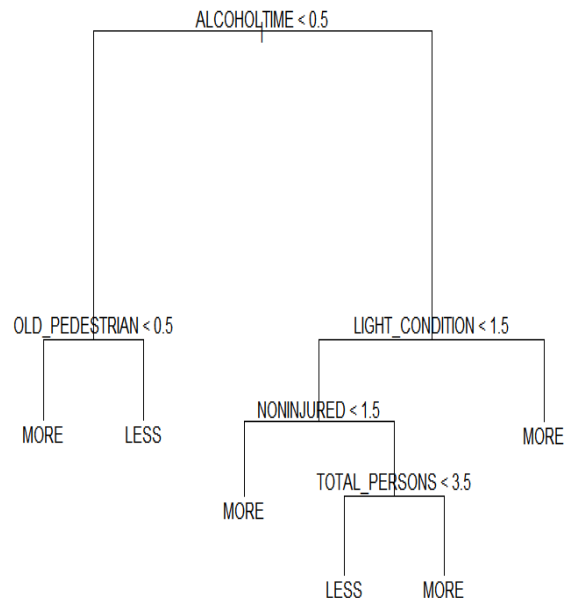
28:1 (Top Level) R Script

Console

```

> 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)
[1] 0.4337835
> set.seed(3)
> cv_tree=cv.tree(tree_model,FUN=prune.misclass)
> names(cv_tree)
[1] "size" "dev" "k" "method"
> 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)
[1] 0.4247297

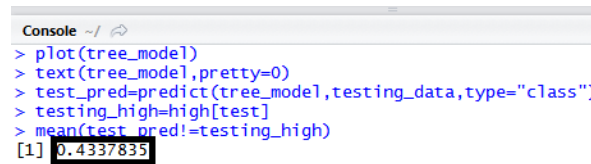
```



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 not 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.



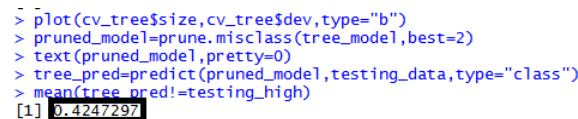
```

Console ~/ / 
> 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)
[1] 0.4337385

```

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.



```

> 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)
[1] 0.4247297

```

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. Zbl0214.46201. 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.ijarcse.com/docs/papers/Volume\\_3/6\\_June2013/V3I6-0454.pdf](http://www.ijarcse.com/docs/papers/Volume_3/6_June2013/V3I6-0454.pdf)
17. <http://www.ijee.org/papers/93-II18.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>