

**SCHOOL OF SCIENCE AND INFORMATION SCIENCES**

**BSC IN COMPUTER SCIENCE**

**COM 423 COMPUTER SCIENCE PROJECT 2**

**PREDICTION OF ROAD FATALIY USING UN/SUPERVISED LEARNING TECHNIQUES**

**OMBATI KERUBO DIANA**

**BS02/1002/2014**

**A project report submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science.**

# DECLARATION

I hereby declare that the research project presented in this report is my original work and has not been presented in any other institution. Due and clear reference is made to the works of other researchers that have informed this project.

Signature\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Moses Nooseli

The Research Project has been submitted in partial fulfillment of the Requirements of the Degree of Science in Information Science at Maasai Mara with my approval as the University Supervisor.

**Signature\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

# DEDICATION

To

All my teachers who have taught and counseled me over the years,

My father and mother,

Above all to God

# ACKNOWLEDGEMENT

The counsel and advice of my supervisor Mr. Moses Nooseli is highly appreciated. I also appreciate the useful advice I got from my evaluation panel members and fellow classmates as a whole. Above all to God be the glory.

**ABSTARCT**

Road accidents are the primary concern. With the urbanization process around the globe, traffic accidents have undergone a rapid growth in recent decades, causing significant life and property losses. Predicting traffic accidents is a crucial problem to improving transportation and public safety as well as safe routing. However, the problem is also challenging due to the imbalanced classes, clustering, and the associative relationship between dependent and independent variables. Some previous research on traffic accident prediction conducted by researchers simply applied classical prediction models on limited data without addressing the above challenges properly, thus leading to unsatisfactory performance.

These days the traffic has been immensely widespread due to the fact that production and usage of vehicles are drastically improved. The dataset taken is mainly focusing on the fatality of unpredictable circumstances that happen on roads. Dataset consists of the attributes like time period at which accident occurred, location points, number of police personnel present, weather conditions etc.

The algorithm chosen takes specific attributes of dataset as input and produce clusters based on traits considered. Based on visualization results, we project the measures that are to be taken to overcome the problem. With this we can assess the various variables and find out measure to be taken to curb road accidents.

This paper, through a case study, presents our explorations on effective techniques to address the above challenges for better prediction results.

Basically we group the fatality rate of an accident into three classes (low, medium and high), Clusters that are formed using K-means, and expectation maximization algorithms are then analyzed to discover relationship using the rpart based decision tree for classification. Results showed that the selected machine learning techniques are able to extract hidden patterns from the data. Density histograms are used for accident data visualization.

**Keywords:** Clustering, Rpart classification, Rattle

**TABLE OF CONTENTS**

CONTENTS

[DECLARATION 2](#_Toc514073402)

[DEDICATION 3](#_Toc514073403)

[ACKNOWLEDGEMENT 4](#_Toc514073404)

[**CHAPTER ONE** 6](#_Toc514073405)

[**1.0** **INTRODUCTION** 6](#_Toc514073406)

[**1.1** **PROBLEM STATEMENT** 7](#_Toc514073407)

[**1.2** **OBJECTIVES** 7](#_Toc514073408)

[**1.3 SCOPE** 8](#_Toc514073409)

[**1.4JUSTIFICATION** 8](#_Toc514073410)

[**3.0.4 Hardware Required** 10](#_Toc514073411)

[**1.7. BUDGET** 11](#_Toc514073412)

[**BACKGROUND STUDY** 12](#_Toc514073413)

[**CHAPTER 2** 13](#_Toc514073414)

[**LITERATURE REVIEW** 13](#_Toc514073415)

[**INTRODUCTION** 13](#_Toc514073416)

[**MODELS FOR MACHINE LEARNING** 13](#_Toc514073417)

[**CHAPTER THREE** 17](#_Toc514073418)

[**METHODOLOGY** 17](#_Toc514073419)

[**2 DELIVERABLES** 20](#_Toc514073420)

[**DEVELOPMENT METHODOLOGY** 21](#_Toc514073421)

[**i)** **Data Collection Methods** 22](#_Toc514073422)

[**ii)** **Data Analysis** 22](#_Toc514073423)

[**CHAPTER FOUR** 23](#_Toc514073424)

[**SYSTEM ANALYSIS AND REQUIREMENT MODELLING** 23](#_Toc514073425)

[**Phase1: Preprocessing** 23](#_Toc514073426)

[**Phase2: k-means clustering** 23](#_Toc514073427)

[**Phase3: Rpart algorithm** 24](#_Toc514073428)

[**CHAPTER FIVE** 25](#_Toc514073429)

[**SYSTEM DESIGN** 25](#_Toc514073430)

[**INTRODUCTION** 25](#_Toc514073431)

[**RESULTS** 26](#_Toc514073432)

[**CONCLUSION** 33](#_Toc514073433)

### **CHAPTER ONE**

# **INTRODUCTION**

Data mining is defined as a non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. Indeed, it is a vital part of business analytics and the most important trends in information technology. It involves many common classes of tasks (clustering, classification algorithm etc.) which are designed for knowledge discovery.

Due to the outrageous events of the accidents on the road, this has become a major challenge to reduce this issue. Many of the solutions with the diversified algorithms which have different views on the problem and different analysis are shown. But the time and again the issue of this is inevitable. The proposal of this project is to provide a solution for the conflict of road accidents its adversities. This implementation is conducted by the inspiration from the road traffic and accidents that occur around. Data mining uses many different techniques and algorithms to discover the relationship in large amount of data. It is considered one of the most important tool in information technology in the previous decades.

The ability to predict future accidents (e.g., where, when, or how) is thus very useful not only to public safety stakeholders (e.g., police) but also transportation administrators and individual travelers. A potential application of such technique would be real-time safe route recommendation for drivers. With the rapid development of data collection techniques and the availability of big datasets in recent years, predicting traffic accidents has become more realistic. Detailed weather condition, light conditions, and speed limit among other variables could give enough information to assess the fatality rate.

However, this problem is very challenging due to a few issues. Class imbalance. Traffic accidents are rare incidents. If we construct class labels based on accident vs. no-accident for each road, the classes will be severely imbalanced. Spatial heterogeneity, i.e., the prediction model parameters may vary from place to place. For example, factors causing traffic accidents in large cities with dense population and lower speed limits might be very different from those in rural areas with low population density but high speed limit. A global model might not be very accurate everywhere. The relationship between environmental factors and accidents might be complex and non-linear. Simple linear models might not achieve good performance.

**We highlight our contribution as follows:**

* First we collect and analyze the data to remove noise from the dataset
* Then perform clustering on the selected data to find the available clusters that satisfy the dataset.
* Classification is done on the dataset to specify which classes cause a higher rate of fatality compared to other variables which have a low impact to the rate of fatality.

## **PROBLEM STATEMENT**

The purpose of this investigation is to reduce the number of road accidents in Kenya by finding risks and circumstances which can be shown to be regular contributing factors to road accidents. If the major contributors can be established, these can be published to make people more aware of when they are potentially at risk of an accident, and allow them to avoid these risks

when possible (Gupta, AUGUST, 2017). We therefore intend to us the data mining software Rattle and Microsoft Access or Microsoft Excel to interrogate the dataset and extract any patterns and constraints which can be used to establish the main contributing factors to road accidents. We plan to find attributes which have high relevance to road accidents and then construct association rules with which we can achieve a high level of support and therefore have a high level of confidence.

## **OBJECTIVES**

#### **General**

To develop a model using decision tree algorithm rpart that shows how various variables affect the fatality rate of an accident.

#### **Specific**

1. To review related systems and models for predicting the likelihood of a driver causing an accident,
2. To develop an algorithm for predicting the likelihood of a driver causing an accident.
3. To find out how use of seat belts determines the severity of a road traffic accident.
4. To assess how drunk-driving determines severity of a road traffic accident.
5. To scrutinize how incorporation of roll over protection bars on the vehicle determines the severity of a road traffic accident.
6. To test and validate the developed algorithm

## **1.3 SCOPE**

This research limits itself to machine learning since it is a current application of artificial intelligence based around the idea that should really just be able to give machines access to data and let them learn for themselves. The research also is limited to prediction as machine learning here will be used to train an existing data to make prediction outcomes in future.

## **1.4JUSTIFICATION**

Over many years, road accident data has been collected with most relevant data has collected thousands of datasets. These sets, if analyzed can be used in marketing or even in predicting the right targets for different outcomes. Without an accurate model for prediction, one cannot know the targeted outcome on the developed mode. Other reasons are;

* **Performance-** the use of more than one algorithm in prediction will help to cater for required results. And provide the best algorithm that can yield better results.
* **Time-** the Quality model can minimize the time required to process large amounts of data in a given dataset.

**PROJECT SCHEDULE**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | FROM | TO | JAN | | February | | MARCH | | APRIL | | MAY | |
| TASK |  |  | 1-15 | 16-28 | 1-15 | 16-28 | 1-30 |  | 1-30 |  | 1-17 |  |
| SYSTEM PLANNING AND SELECTION | 1/01/18 | 15/1/18 |  |  |  |  |  |  |  |  |  |  |
| PROPOSAL WRITING | 16/1/18 | 28/1/18 |  |  |  |  |  |  |  |  |  |  |
| SYSTEMS ANALYSIS AND DESIGN | 1/2/18 | 15/2/18 |  |  |  |  |  |  |  |  |  |  |
| SYSTEM IMPLEMENTATION | 16/2/18 | 30/03/18 |  |  |  |  |  |  |  |  |  |  |
| DOCUMENTATION | 01/4/18 | 30/04/18 |  |  |  |  |  |  |  |  |  |  |
| DELIVERY | 1/5/18 | 15/5/18 |  |  |  |  |  |  |  |  |  |  |

**1.6 RESOURCE REQUIRED**

1. Minimum requirements

**System Software**

Windows version i.e. 8.1, 10, 7. Windows version is preferred for this project because it simple to understand to everyone bearing in mind that this system is going to be used by novice computer users.

**Application software**

Bootstrap- bootstrap is preferred to css because it is very responsible when viewing. Weather in computer or mobile phone the viewing is the same

Rstudio for training model using the data set.

R programming language will help in training the model.

SHINY- for developing user interface

Internet browser: Mozilla Firefox, internet explorer, google chrome, safari, opera- any browser can work here. This browser can be used as an optional for shiny

## **3.0.4 Hardware Required**

1.6 GHz processor Pentium 111- since this project is a heavy one such kind of processor is required for the system to work with the preferred speed.

4GB RAM- any RAM below this might make the computer hung.

4 GB free hard disk space- this is for storing the work that will be going on. For example, storing data in the database.

A PC /laptop, GSM modem. Modem is required for the internet. Any other means of internet connection it’s allowed.

**On the client side**

A PC/ laptop running windows versions connected to the internet,

**Supported platform**

Personal computer running windows versions like 10, 8.1, or 7.

### **1.7. BUDGET**

|  |  |
| --- | --- |
| NAME | ESTIMATED COST |
| HP laptop Intel Core i3, 500 gb hard disk, 4 gb RAM | Kshs. 32,000 |
| Modem | Kshs. 2500 |
| Hosting fee | Kshs 5000 per annum |
| TOTAL | Kshs. 39500 |

Table 2.0 budget

# **BACKGROUND STUDY**

Road and traffic accidents are uncertain and unpredictable incidents and their analysis requires the knowledge of the factors affecting them. Road and traffic accidents are defined by a set of variables which are mostly of discrete nature. The major problem in the analysis of accident data is its heterogeneous nature. Thus heterogeneity must be considered during analysis of the data otherwise, some relationship between the data may remain hidden. Although, researchers used segmentation of the data to reduce this heterogeneity using some measures such as expert knowledge, but there is no guarantee that this will lead to an optimal segmentation which consists of homogeneous groups of road accidents. Therefore, cluster analysis can assist the segmentation of road accidents.

Cluster analysis which is an important data mining technique can be used as a preliminary task to achieve various goals. Cluster analysis is used to categorize the accident data into different categories and further analyzed cluster results using Negative Binomial (NB) to identify the impact of driver age on road accidents. Ma and Kockelman used clustering as their first step to group the data into different segments and further they used Probit model to identify relationship between different accident characteristics. Poisson models [5, 6] and Negative binomial (NB) models [7, 8, 9] have been used extensively to identify the relationship between traffic accidents and the causative factors. (Tiwari, 2017)

In 2004 the Government revamped the road transport and PSV regulations by developing very stringent rules which cost the law breakers hefty fines. These rules were famously known as the ‘Michuki’ rules in reference to the then minister for transport the late Honorable John Michuki. PSV industry was highly affected as the crew members were mandated to wear special uniform, vehicles to be fitted with speed governors and functional seat belts.

Most PSVs complied across the country and during this period the number of accidents significantly reduced. The Government of Kenya through the department of transport established structures to enforce and monitor compliance with the road traffic accidents.

### **CHAPTER 2**

## **LITERATURE REVIEW**

# **INTRODUCTION**

Determinants of severity of road traffic accidents were discussed in detail using the views of other researchers and authors. Various theories and concepts on causes of the road traffic accident were discussed by large and far.

Most accidents are dependent on the area (rural or urban), and the type of street (intersection or highway)

# **MODELS FOR MACHINE LEARNING**

According to IBM data scientist M.Tim Jones machine learning fall roughly into three categories: supervised, unsupervised and reinforcement. Supervised learning involves feedback to indicate when a prediction is right or wrong whereas unsupervised learning involves no response. The reinforcement learning is similar to supervised learning in that it receives feedback, but it’s not necessarily for each input or state.

Since accident prediction is a supervised learning we will discuss it in details;

**2.2.1** **SUPERVISED LEARNING**

**Input**

**Error**

**Critic**

**Output**

Fig 2 Learning model for supervised learning

In supervised learning a data set includes its desired outputs such that a function can calculate an error for a given prediction. The supervision come when a prediction is made and an error produced (actual vs desired) to alter the function and the mapping.

Supervised learning is the simplest of the learning models to understand. Learning in the supervised model entails creating a function that can be trained by using a training data set, then applied to unseen data to meet some predictive performance. The goal is to build the function so that it generalizes well over data it has never seen.

**2.2.2 BUILDING AND TESTING A MAPPING MODEL IN SUPERVISED LEARNING**

You build and test a mapping function with supervised learning in two phases. In the first phase, you segment a data set into two types of samples: training data and test data. Both the training and testing data contain a test vector (inputs) and one or more known desired output. You train the mapping function with the training data set until it meets some level of performance. In the context of supervised learning, this occurs with each training sample where you use the error (actual vs desired output) to alter the mapping function. In the next phase, you test the trained mapping function against the test data. The test data represents data that has not been used for training and provides a good measure for how well the mapping function generalizes to unseen data.

Numerous algorithms exist under supervised learning. And this case decision tree is used.

**2.2.3 DECISION TREE**

According to IBM the international technology industry, a decision tree is a supervised method of classification. Algorithms of this variety creates trees that predict the result of an input vector based on decision rules inferred from the features present in the data. Decision trees are useful because they are easy to visualize so you can understand the factors that lead to a result.

Two types of models exist for decision trees:

* Classification trees
* Regression trees

**Classification trees**

Where the target variable is discrete value and the leaves represent class labels.

**Regression trees**

The target variable can take continues values.

You use a data set to train the tree, which builds a model from the data. Decision tree presentation is useful when you want to see how attributes in the data can split, or partition, the whole dataset into subsets relevant to the problem.

**RPART DECISION TREE ALGORITHM**

Classification models anticipate clear cut class names and expectation models foresee consistent esteemed capacities. The classification model consists of 2 steps **building a model** and **using that mode**l for classification. In the first step those arrangement calculations build the classifier. The classifier may be assembled from those preparing set committed up of database tuples. In the second step it builds the classification tree depending on the classification rules or classifier. Depending on the classification tree you can predict whether certain event can happen or not. Predictions through classification is accurate because it builds the tree depending on the target classes of the target variable. The efficiency of the tree depends mainly on the root node we have chosen.

### **CHAPTER THREE**

# **METHODOLOGY**

The model development methodology is a collection of techniques used to collects facts and data, tools to analyze data and processes, tools to implement and test the system and the documentation that enable developer in their effort to implement a new system. There is different methodology adapted to implement this new model over a given time.

**WATERFALL METHODOLOGY**

The waterfall model is an approach in software development that describes a linear and sequential development method. The waterfall model consists of five to seven phases, which structures the development of this model. Once a phase is complete, the next development step follows and the results of the previous phase flow into the next phase. Each phase is defined by different tasks and objectives, whereby the entirety of the phases describes the life cycle of the software up to its delivery.

Research Problem and Goal

Interaction/Redevelopment of the model

Documentation

Installation

Maintenance

**STEP 1: REQUIREMENT GATHERING AND ANALYSIS**

**IDENTIFY AND PREPARE DATA**

Predicting gender on twitter will help especially in targeting the right audience in the market. Based on the research question, translation of the research objective into analytic terms that is building a predictive model. Selecting the best data for target modelling requires understanding of the market. So, in this stage the best type of data to use is demographic of gender and the source is Kaggle.com which gives out the general data. Also, this step, data is prepared into right format for analysis and the data needed here is book related data from the general data. By getting this, the data is loaded into excel and then search for “book” to narrow down the data. Cleaning the data is done and defining the variables.

**PHASE 2: DESIGN**

**ANALYSE/TRANSFORM DATA**

Once data is in right shape and perform

* univariate analysis: checking the distribution of each of the variables and features
* multivariate analyses: checking relationships with other variables and with dependent variables

Based on the model (predictive model), transforming the variables will use classification approach

**PHASE 3: TESTING**

Checking if there is a problem in the designed model, based on the functional and non-functional requirement. This stage will involve the client and his/her technician activity, to ensure good satisfaction from the client. If there happen to be any problem, must be turned back to design stage and coding and testing will be done again.

**DEVELOP/TRAIN MODEL**

Based on the defined objectives of the research (supervised) selecting one of or combinations of modeling techniques which is support vector machine for classification and prediction. Validation and assumption of SVM is done in this stage Validate the assumptions of the chosen algorithm. Checking for Multicollinearity and Redundancies of Independent Variables (Features). Developing/Training Model on Training Sample, which is 70% of the data (21,051 in total). Checking Model performance by using ROC.

**VALIDATE/TEST MODEL**

Score and Predicting using Test Sample. Check for the robustness and stability of the model and Checking the Model Performance using ROC.

**PHASE 4: INSTALLATION**

The last product which in this case is a model, will be needed to be installed at the client’s system of which his/her system will be tested so that there is no problem when the client is using the system in future. In short, the product is handed over to the client.

**PHASE 5: MAINTAINANCE**

Some support regarding the software have to be provided to the client and in case the same client needs enhancement to the current software, all the process need to be started from the requirements stage.

### **2 DELIVERABLES**

Deliverables are those tasks that the software is supposed to perform. The proposed system will have the following deliverables.

#### **Phase one**.

Come up with a dataset and clean it to ensure no null values and duplicates of any kind

#### **Phase two.**

Convert the clean data set from categorical to factorial. Divide the dataset into two. 70% Training, 15%validation and 15% testing

**Phase three*.***

Develop the model now. Developing the model is divided into three phases. Part one is Train the dataset to capture a certain trend in the data. For example using R

Model<- rpart (medata~., Training, method=”class”)

Part two is to test the trained model

Model<- predict (model, Testing type= “class”

The third part is to view the result

View (model)

Plot (model)

#### **Phase four**

Develop a user interface platform using shiny.

#### **Phase five**

View the dataset in shiny and see it summary

Upload the trained model in the shiny platform

Perform the prediction by checking the required variables from the check boxws..

Perform the prediction

## **DEVELOPMENT METHODOLOGY**

In this particular research, the technique employed is the Software Development Life cycle

(SDLC). The software development life cycle is the process that documents software evolution

slowly, from the time the idea of the software is conceived to the realization of the final product.

This involves; specification of the software’s expected inputs, processes and outputs; a design

layout of the software’s major components; a validation tool for the software developed; and an

evolution mechanism to track new changes or versions of the software.

This is based on the waterfall model which is a sequential design process often used in software

development process. The development is seen as owing downwards steadily. In the model, one

phase has to be finalized before the process can progress to the next phase.

Analysis

Planning

Maintenance

Design

Implementation

### **Data Collection Methods**

In order to achieve the objectives of the study, both primary and secondary data was used.

Secondary data used in this study included road traffic offence records and accident records from

Kenya Police and NTSA, as well as road traffic offence manuals and other documentations

relevant to the study (Service, 2017).

The researcher used both questionnaires (for NTSA officials and traffic police officers) and

interviews (for matatu drivers) in collecting primary data. This data was useful in understanding

the underlying factors associated with road crashes, determining violations considered high

determinants to road accidents, feasibility of the system, as well as testing whether the system

meets the users’ desired functionalities.

### **Data Analysis**

Data obtained from the field in raw form is difficult to interpret; such data must be cleaned, coded, keypunched into computer and analyzed. It is from the results of such analysis that researchers are able to make sense of the data (Caliendo, 2007)

During data analysis, relationships or differences supporting or conflicting with original or new

hypotheses should be subjected to statistical tests of significance to determine with what validity

data can be said to indicate any conclusions. This research utilized both qualitative and quantitative methods of data analysis. SPSS-a statistical package and excel sheets were used in data analysis.

Reliability of the data was measured against past research data on accident records as well as database records for integrity measurement. The viability of the data is guaranteed by the validation rules in the system. There are input masks to ensure that the data taken for different tests is only the desired set of information needed.

### **CHAPTER FOUR**

## **SYSTEM ANALYSIS AND REQUIREMENT MODELLING**

# **Phase1: Preprocessing**

Before applying data analytic methods one should preprocess the data for efficient results. In our project NA values are deleted and normalization is done in order to remove the outliers in the dataset. Missing values can be deleted by using a command in R Tool. Normalization is done using (a+(X-A) (b-a))/ (B-A) where X=dataset A=min in set B=max in set a=min in range b=max in range.

# **Phase2: k-means clustering**

K-Means clustering is an unsupervised learning technique in which whole dataset is segmented into k clusters groups in which every perception has a place with the group with the closest mean, serving as a model of the group. K-Means is simple and fast to implement by finding the distances and means among the elements and clusters. Distances can be measured by using different techniques such as Euclidean distance measure etc. In K-means distance are measured using Euclidean distance. But the disadvantage of this algorithm is it can’t detect outliers where outliers also play an important role for detecting the behavior.

1. Select ‘k’ randomly where k indicates the number of clusters.
2. Distance between each point and to cluster center is calculated.
3. Place the data item in the respective cluster whose distance is minimum to that cluster.
4. Repeat step 3 until no data item is present in original dataset
5. Calculate the cluster centers
6. Find the distance between each and cluster head, place the data item in the respective cluster with minimum distance from the point.
7. Repeat step 6 until no data item moves among clusters.

# **Phase3: Rpart algorithm**

It a famous data Mining technique which is used for finding predictive models. It builds the regression and classification models. A dataset is broken into smaller and smaller subsets and associated decision tree is incrementally developed. Decision nodes and leaf nodes are the final results. Decision tree can build for both categorical and numerical values. The topmost node in the tree is the root node.

**CONTEXTUAL DIAGRAM FOR THE CURRENT SYTEM**

New traffic data

Existing traffic data

Data processing

Prediction Report

Clustering

Prediction Algorithm

### **CHAPTER FIVE**

## **SYSTEM DESIGN**

## **INTRODUCTION**

Design is the process of applying various principles and techniques in order to define a process or a system in adequate detail for it to be physically realized. There are certain items such as modules, relationship among modules, data structures, relationship between the data structures and algorithms for implementation that must be designed in this phase. During system development, design is the first step into the development phase.

The design stage is expected to deliver outlines of different technical answers that meet the expectations of system analysis and requirement modeling stage.

These solutions include:

1. A minimum-cost solution. This just does the job and nothing more.

2. A medium –cost solution. This is convenient to users and does the job well. It may have additional features which the client did not ask for but the developer thinks they will be needed from experience.

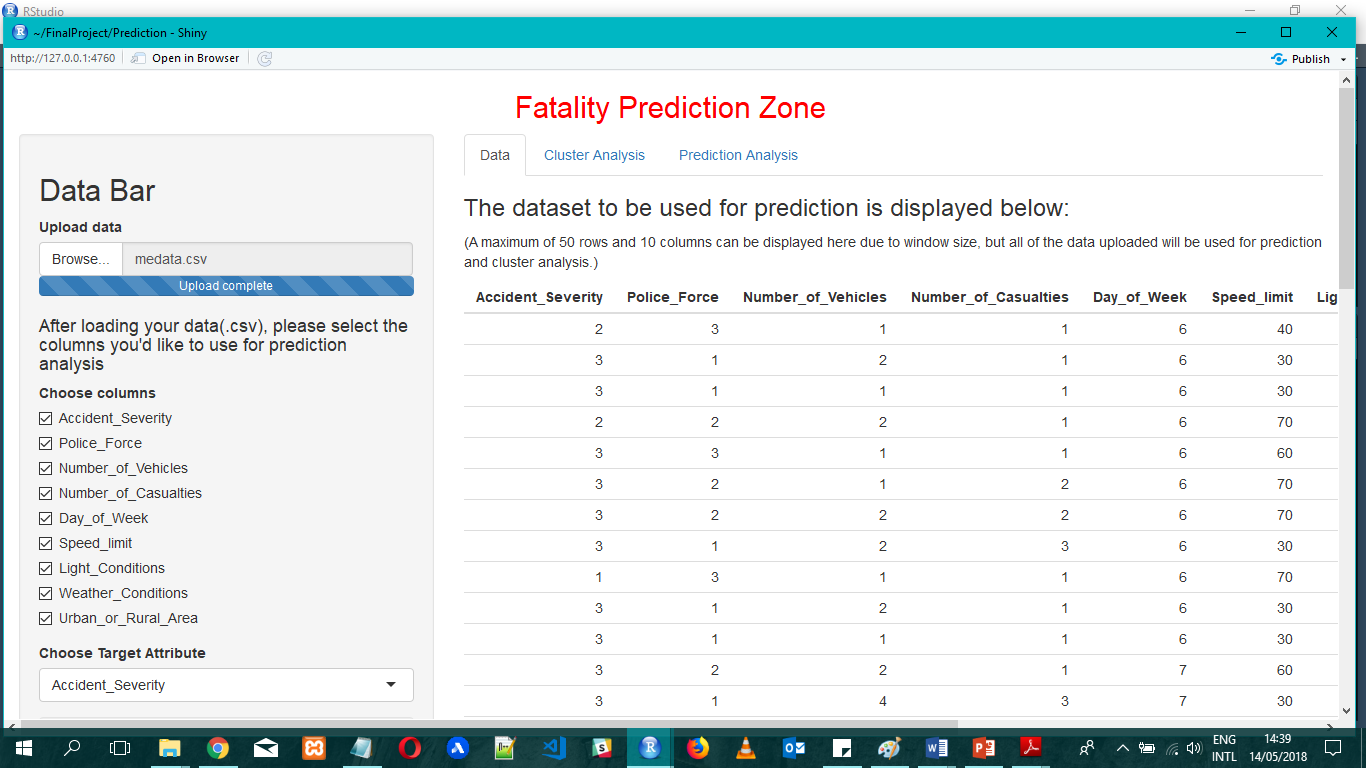
3. A high- cost solution. This includes anything that the client needs.

In our design phase we did the following:

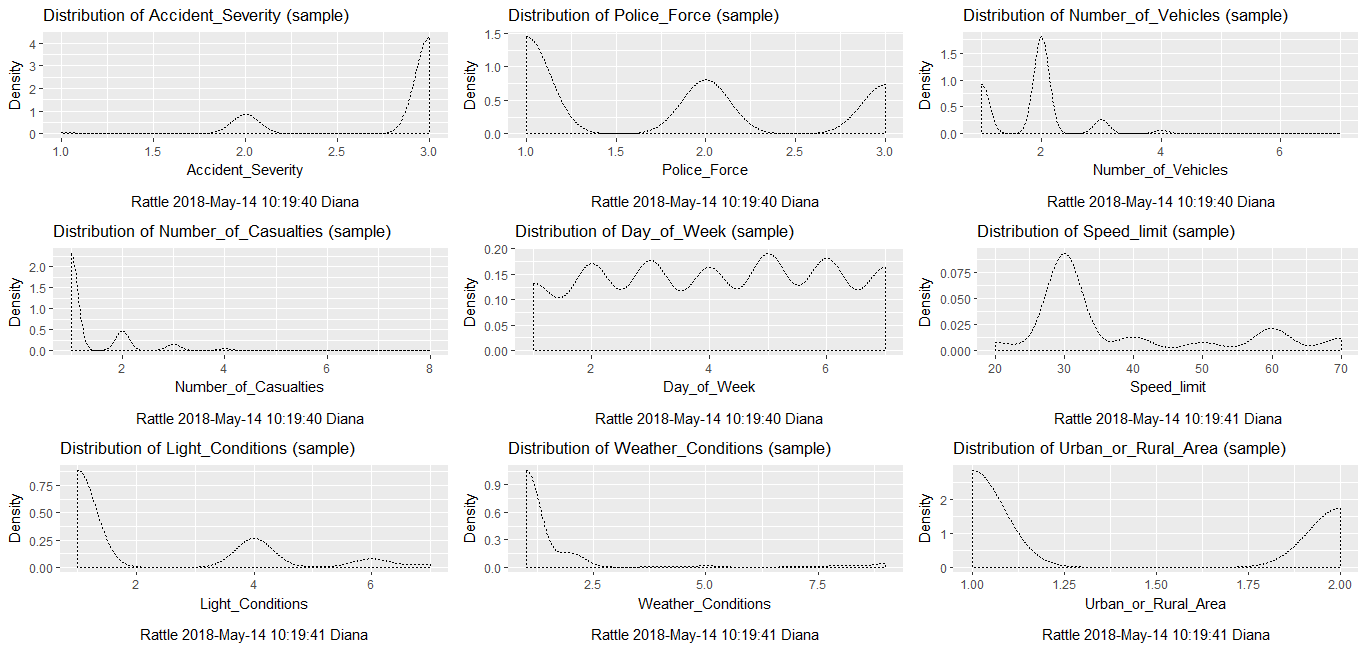
* Organized the system into modules
* Organized sub-modules for each module
* Allocated tasks to processors
* Choose an approach to manage data store
* Handled access to global resources
* Choose implemented logic

## **RESULTS**

**UPLOAD THE DATASET**



**VISUALISATION OF THE VARIABLES**

****

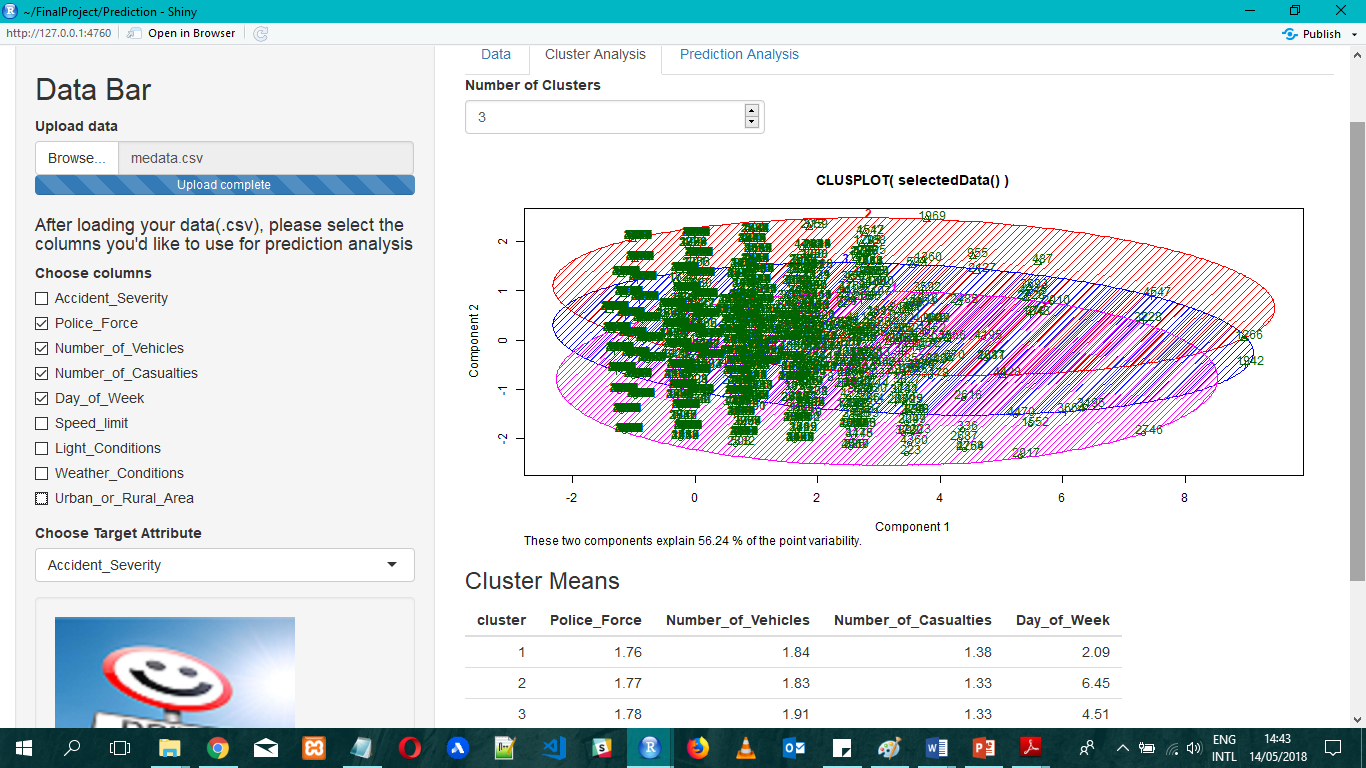
**PERFOMING CLUSTERING**

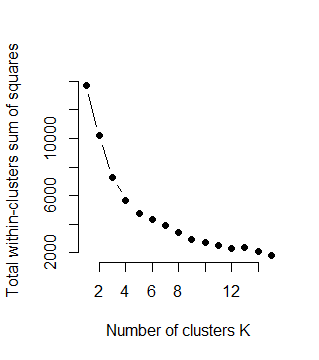
For the dataset first missing values are removed and normalization is done using commands. For the processed dataset clustering is done. In this case we have took 3 columns.(number of vehicles, number of causalities, day of week) and done into 3 clusters. The points which are outside or on the boundaries of clusters represent the clusters. For the dataset first missing values are removed and normalization is done using commands. For the processed dataset clustering is done. In this case we have took 3 columns.(number of vehicles, number of

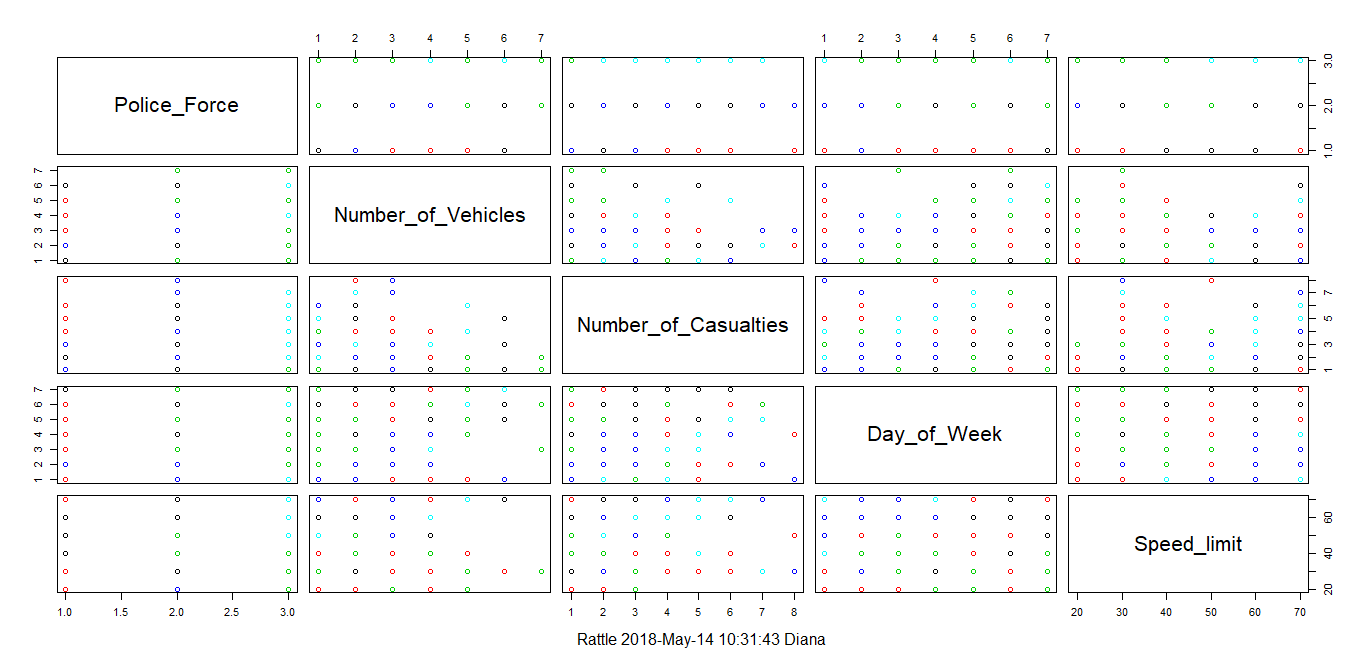
causalities, day of week) and done into 3 clusters. The points which are outside or on the boundaries of

clusters represent the clusters.



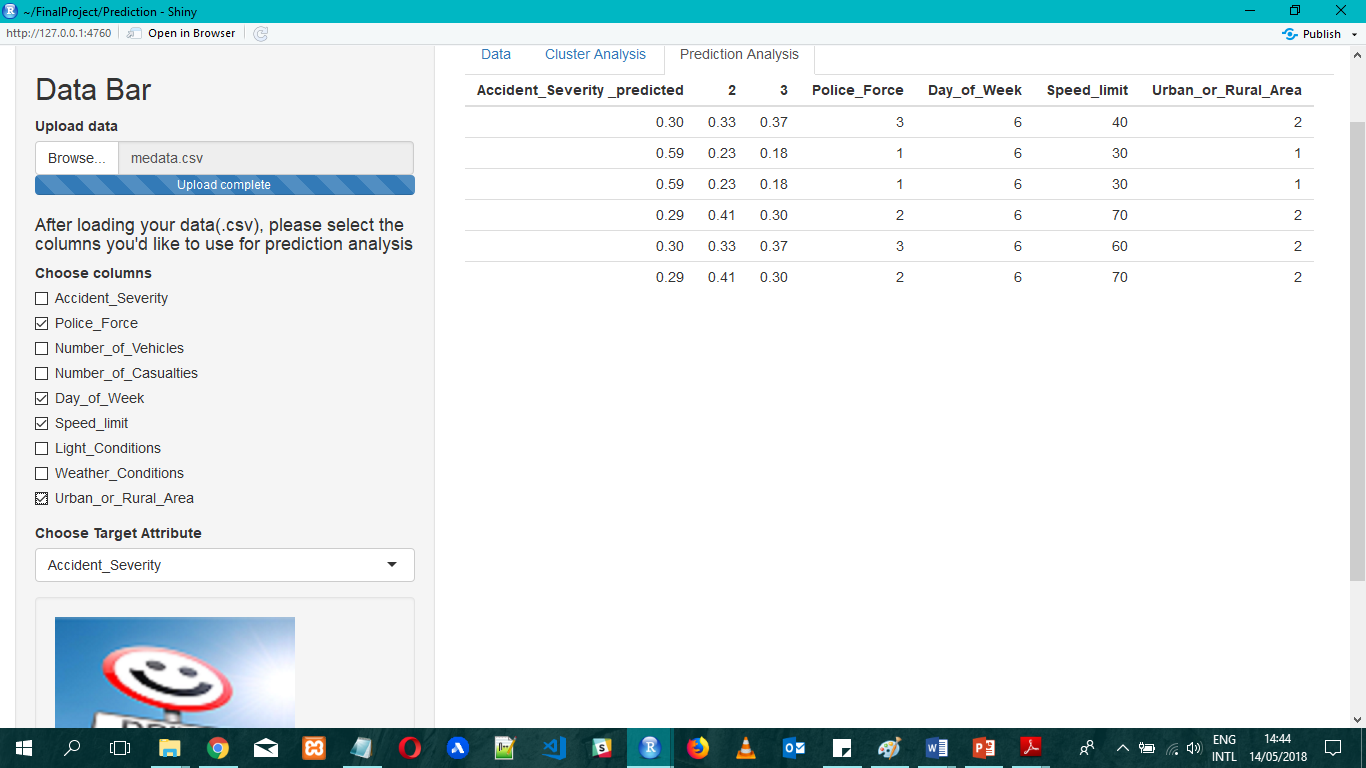


find number of clusters

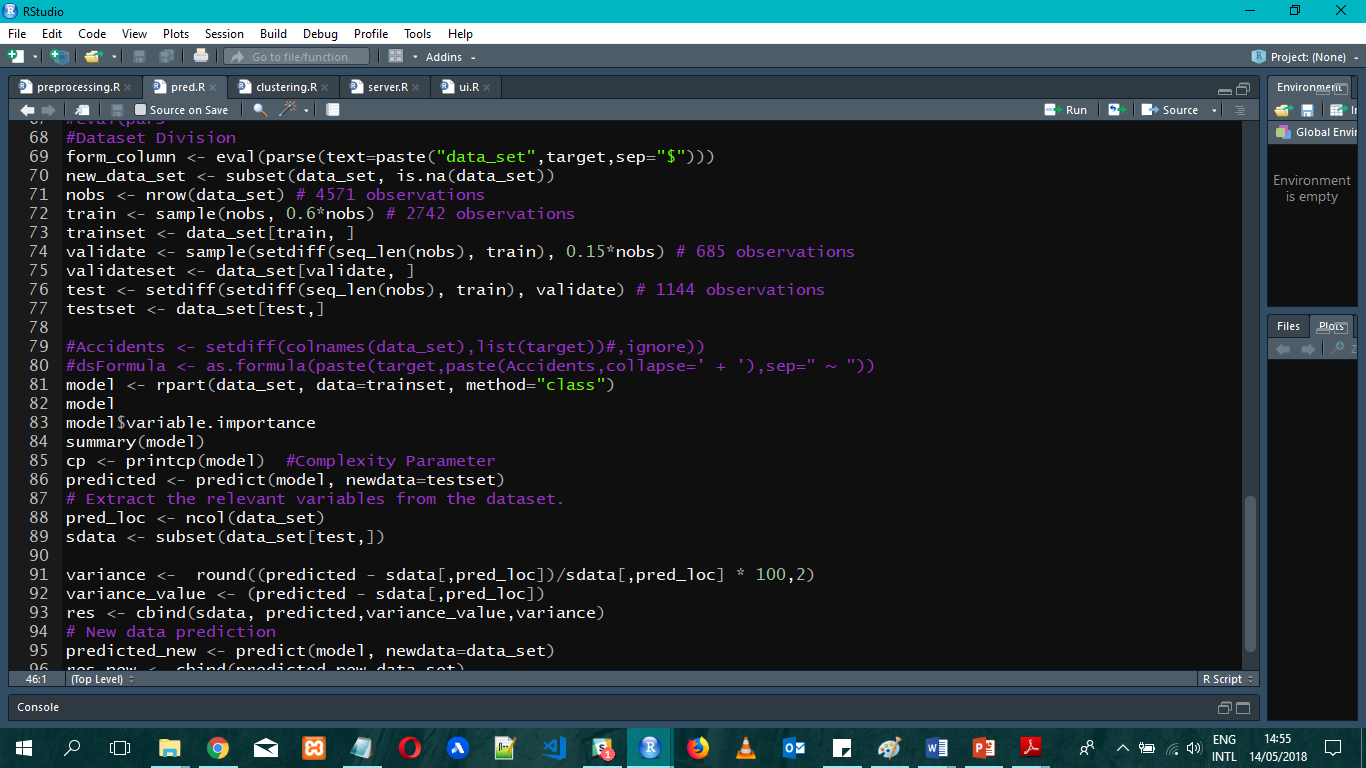


**PREDICTION**

From the values we can predict what the accident severity is when there was a certain police force. The following figure shows the probability of police force when the accident severity. Here we see the result shows many attributes involved. Therefore if all the attribute values are met, there will be high probability of accidents.



**PROGRAM CODE FOR THE MODEL CREATION**



### **CONCLUSION**

**Future scope:**

The analysis says that, if the accident severity is more than it means the police force at that place is less. Therefore there is a need to engage police force at places of need. With proper police force and speed limits accidents happening can be reduced which leads to reduction in the amount used to spend on causalities. The decision trees says that on Tuesday more accidents are happening, therefore on Tuesdays the traffic lanes have to be diverted to some other lanes at regular intravels to reduce accidents. Here we have implemented it based on only some attributes, but if this is implemented based on real geographical areas related data, we can find the location (latitude,longitude) exactly where an action is supposed to be taken. With this results and conclusions we can end saying that if proper measures are taken and implemented really, we can see an accident free

society.

* Build a live app for predicting the likelihood of road accident based on live data
* Find ways to account for randomness

# References

A. Savasere, E. O. (1995). *“An efficient algorithm for mining association rules in large databases,”.* Zurich, Switzerland: in Proceedings of the 21th International Conference on Very Large Database (VLDB '95), pp. 432–443.