2021

Navigation guidance to pedestrians' road accident safety in Kenya using Bayesian analysis.

Kuria, Paul Munene
School of Computing and Engineering Sciences
Strathmore University

NAVIGATION GUIDANCE TO PEDESTRIANS ROAD ACCIDENT SAFETY IN KENYA USING BAYESIAN ANALYSIS

Paul Munene Kuria

Submitted in partial fulfillment of the requirements for the Degree of Master of Science in Information Technology (MSc.IT) at Strathmore University

School of Computing and Engineering Sciences, Strathmore University, Nairobi, Kenya.

September, 2021

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted inany previous application for the award of a degree in any University. Except where states otherwiseby reference or acknowledgment, the work presented is entirely my own. I hereby give consent for my thesis, if accepted, to be available for photocopying and understand that any reference to or quotation from my thesis will receive an acknowledgement.

Paul Munene

Kuria

20/09/2021

Registration Number: 124686

SUPERVISOR'S DECLARATION

I affirm that the work reported in this thesis was carried out by the candidate under my supervisionand has been submitted with my approval as the university supervisor.

Signature…….  …… Date……20th September 2021….

Name of supervisor: Prof. Ismail Ateya
School of Computing and Engineering Sciences,Strathmore University.

Abstract

In most countries, policy makers find the safety of pedestrians to be a major concern because many pedestrians are involved in road traffic accidents. Roughly, 1.35 million people's lives every year are shuttered because of road traffic accident, among these fatalities pedestrians are most vulnerable road users given that the road infrastructure is designed with less or no consideration to safety of a pedestrian. Therefore, the purpose of this research was to propose a navigation aid system that informs pedestrians about the safety status of a given road route in Kenya. Experimental research was used because it facilitates the manipulation of variables which is essential in this study. In designing the system, the research implemented the Agile development methodology. The data used was from the secondary references which included traffic accident fatalities records from NTSA. The data was used to train a Bayesian Linear Regression model, which predicts the pedestrian's road accident fatality likelihood in a given road route. The model's results, posterior distribution likelihood, are presented in terms of color coding on the Google Maps routes via a mobile application, where red indicates high danger, red for high-medium, brown for medium, yellow for low-medium, green for low. The study found that pedestrians constitute 54.14% of total road accidents fatalities and pedestrian's risk of being involved in a fatal road accident is influenced by following factors, population increase with dependency of region urbanization, the time of the day with 2000hrs-21000hrs posing highest risk, tendency of using same road route and gender in which a male's risk is 73% higher in comparison to female's risk. These findings were used in fine tuning the Bayesian model and enriching the system's web portal reports. The implication of the research is that a pedestrian is able to know the safety status of a given road route in respect to road accidents via a mobile application enabling him/her to choose among the available routes and taking the necessary safety precautions based on the safety status of the chosen route.

Acknowledgements

I want to express gratitude to my university supervisor Prof. Ismail Ateya and Co-ordinator Dr. Vincent Omwenga for taking their precious time to guide me through the entire thesis. Additionally, I would like to acknowledge and appreciate Strathmore University fraternity especially my fellow colleagues who have always inspired me.

Table of Contents

Abbreviations and Acronyms

AI	:	Artificial Intelligent.
API	:	Application Programming Interface.
GPS	:	Global Positioning System.
HDI	:	Highest Density Interval.
MAPE	:	Mean Absolute Percentage Error.
NoSQL	:	Non-Structured Query Language.
NTSA	:	National Transport and Safety Authority.
OLS	:	Ordinary Least Squares.
WHO	:	World Health Organization.
UK	:	Unique Key.
FK	:	Foreign Key.
PK	:	Primary Key.

Definition of terms

Pedestrians	Is a person travelling on foot, whether walking or running (Hezaveh, 2018).
Bayesian analysis	Refers to reliability reallocation across likelihood which starts with a prior point of view in every likelihood followed by collection of data and re-allocation of reliability across the likelihood leading to a posterior point of view in every likelihood, (Zong, 2019).
Bayesian network	A graphical representation of random variables with their dependencies, (Ying, 2017).
Machine learning model	This is a file that has been well trained to recognize specific types of patterns, (Habibullah, 2019).
Google Maps	It is a Web-based service that gives detailed information about geographical regions around the world, (Rahmi, 2017).
Agile software development	Refers to methodologies of software development that are centered around the idea of iterative development, (Dhir, 2019).

List of Tables

List of Figures

List of Equations

Chapter 1 : Introduction

1.1 Background of the Study

Policymakers find pedestrian safety to be a major concern in most countries. This is because many pedestrians are involved in road traffic accidents. During an accident both direct and indirect effects occur. Whereby the direct effects include injuries, death and damage to properties whilst the indirect effects include; resentment because of expectation or rather fear of being involved in a road traffic accident, (Jamroz et al., 2015). Jamroz notes that these indirect effects have attracted a lot of attention recently. Therefore, some concepts like subjective danger and amenity are used for this. Currently, the quantification of these concepts cannot be done and so only if the influence of feeling unsafe and the traffic behavior is known that's when one may look for some indicators of these concepts. The direct effects of accidents mainly injury and damage are well known which is called objective unsafety. This kind of data can be used as well to stipulate the traffic risk that is the probability of a pedestrian fatality as a result of traffic accidents.

The World Health Organization (WHO) report 2018 says that roughly 1.35 million people's lives are shortened or ruined because of road accidents yearly. About 20-50 million non-fatal casualties agonize with so many suffering from disability caused by the injuries. In spite of developing countries having 60% of the vehicles in the world, 93% of the world's fatalities on the roads happen in those countries. Road traffic accidents have a great impact on the global economy which costs many countries about 3% of their gross domestic product (GDP). The WHO also noted that in Kenya every year between 3, 000-13, 000 individuals die in road accidents. In which the road users who include pedestrians, motorcyclists and cyclists are the most vulnerable. In addition, one-third of the fatalities are formed by passengers who die due to risky means of public transportation. In the year 2020, 1108 pedestrians died as a result of road accidents, the number is slightly higher when compared to 2019 where 1137 pedestrians lost their lives by the same date. A total of 3114 people died due to road accidents in 2020 compared to 2942 who lost their lives in 2019, (NTSA, 2020).

Automatic route navigation systems have become an essential tool in modern society for people who wish to navigate in physical spaces, especially within unknown areas. Before the

advent of Web-based technologies, a common approach in deciding which route to take was to refer to a physical map before or during the travel and the entire trip would generally need to be pre-planned by the travelers. Automatic route navigation systems had begun to emerge and became more popular after the growth of mobile technology and improvements in digital mapping techniques. Currently there are a number of services designed to help users navigate effectively which have been made available to the public. Google Maps in particular, which was established in 2005, has grown to become one of the most popular navigation services, helping millions of users find the most cost-effective path towards their desired destinations through driving, public transportation or walking, (Siriaraya et al., 2020)

Earlier research into route navigation systems tended to focus on such utilitarian aspects of travelling. These navigation systems were generally designed to provide routes which allow users to travel from one point to another while minimizing costs in time, distance and resources. When calculating the cost of a route, dynamic factors such as congestion, delays in public transportation and the weather were later taken into consideration. More recent studies have seen the development of navigation systems which go beyond recommending routes purely based on their utilitarian qualities. Instead of recommending routes which only minimize the economical costs of travelling, researchers have been examining algorithms to recommend paths that maximize the enjoyment and positive experiences of users as they travel from one point to another based on the hedonic attributes within the routes (i.e., the aesthetic qualities of a given path or the presence of scenic or fashionable locations). Increasing attention is also being given to recommend routes which emphasize on the safety and well-being of users as they travel, such as routes which have less probability of users encountering crime, accidents and pollution, (Gavalas et al., 2017).

Several studies have been carried out to propose road accident prediction and analysis models, each one of them framed within the socioeconomic, cultural and development conditions of the country where it was proposed, and therefore, making evident the difficulty of proposing a single predictive or analytical model that works in all contexts. The main areas of interest that these models' study are; i) detection of problematic areas for circulation; ii) real time detection of traffic incidents; iii) road accident forecasting and iv) prediction of the severity of the consequences suffered by being involved in a road accident. Current advances in analytic algorithms and machine learning methods allow researchers to propose models of complex issues in the study and prediction of road accidents, such as the mining of heterogeneous data sources,

the process and classification of real time data flows, the need to process multi-dimensional data having strong prediction capabilities, and to study, analysis and model the real-time information provided by on-road equipment such as loop inductors, cameras and radar equipment. Therefore, the study of road accident prediction is a field of relevant and current scientific knowledge, open to innovation in the research of algorithms and data analysis techniques that respond to the challenge of generating a more secure mobility environment, which considers the particularities of each country or region, i.e., traffic composition, weather conditions, roads conditions, and demography, (Gutierrez-Osorio & Pedraza, 2020).

Road traffic accident analysis has drawn significant attention to the researchers recently. As it is used in determining the factors that cause traffic accidents even though most of the methods used for research depend on statistical data or by performing basic surveys derived from interviews and questionnaires which do not give a better and impeccable solution. The main challenge is that studying some features based on subject behaviors in traffic accidents is a bit challenging when using some of these research methods. In addition, road traffic accidents are very unpredictable which makes direct observation a bit difficult making it hard to get 100% accurate data and so application of improved methods which can yield better analysis results is really required. Among the highly advanced AI disciplines there is machine learning which has many developed methods that can be used to examine this sector (Habibullah, 2019). Among these Machine Learning methods Bayesian analysis is one of the methods that can be used to give better outcomes. The aim of the study was to examine the occurrence of road crashes as well as predict the likelihood of a pedestrian being involved in a road accident based on traffic data scored against the road routes by use of advanced machine learning techniques. The study is focused on performing traffic accident analysis, by applying Bayesian analysis.

1.2 Statement of the Problem

According to WHO (2018), Kenya is ranked as a lower middle-income country that has experienced rising numbers of road accidents in the last decades, this is because of urbanization in the country. Road infrastructure development in Kenya is still lagging as well as policy implementation challenges in adhering to the international safety standards. NTSA (2020), which is a report published on yearly basis by NTSA, an organization in charge of the transport sector in

Kenya notes that; road crashes have claimed the lives of 3,114 persons compared to a similar period in 2019 where 2,942 lives were lost.

Countries that are becoming motorized and which have a high number of people walking and bicycling have a big problem of collision between pedestrians and bicyclists or motor vehicles. Pedestrians are often known as the vulnerable road users since in collision with a motor vehicle their difference in mass and weight as well as absence of a protective structure or cover makes their injury susceptible. Protection of pedestrians is very challenging as road infrastructure has been purposefully constructed for motor vehicles, with minimal attention to people walking and they may wish to walk on roads or alongside the roads, or may want to cross them, or change direction at intersections, (Das, 2020).

It is of great importance to take full advantage of the traffic accident statistics and mine potential information so as to provide a basis for the analysis of accident mechanisms and the improvement of road safety. Bayesian network is one of the effective methods in the field of machine learning to express uncertainty analysis and probability reasoning of a system. It can exploit the dependence relationships based on local conditions in a model to conduct bidirectional uncertainty investigation for prediction, classification, and diagnostic analyses, (Zou & Yue, 2017). A navigation guidance system that utilizes the Bayesian network modeling to inform on route safety status in respect to road accidents is therefore important as it can help pedestrians in choosing among the available routes and taking necessary precautions based on the safety status of the route chosen.

1.3 Research Objectives

1.3.1 General Objective

Providing the pedestrians and traffic officials with valuable information on how safe a route is in respect to pedestrian road accidents using Bayesian analysis in predicting the likelihood of road accidents in a given route.

1.3.2 Specific Objectives

i. To examine factors contributing to road accidents involving pedestrians in Kenya.
ii. To examine pedestrian's movement patterns with respect to road accidents.

iii. To review related systems and models for predicting the likelihood of a road accident.
iv. To propose a navigation aid system for predicting the safety status of a given road route in respect to pedestrian road accidents by leveraging Bayesian network modeling.
v. To test the developed system.

1.4 Research Questions

i. What are the factors contributing to pedestrian's road accidents in Kenya?
ii. What are pedestrian movement patterns in regions where there is a high number of pedestrian road accidents?
iii. What are the existing systems and models for predicting the likelihood of a road accident occurring?
iv. How can Bayesian network modeling be used to propose a navigation aid system that predicts the safety status of a given road route in respect to pedestrian road accidents?
v. How can the proposed navigation aid system be tested?

1.5 Justification

The navigation aid system proposed in this research informs various stakeholders who include but not limited to pedestrians and traffic officials about the safety status of a given road route in Kenya. The system has a mobile application that aids pedestrians in choosing among available route options to use based on their safety status and taking necessary safety precautions based on the safety status of the chosen road route. Traffic officials using the system's web portal reports are informed on the routes to pay more attention to and plan on how to effectively direct traffic and enforce laws and safety guidelines. The system's web portal reports also provide traffic administrators with a globally overview of road accident safety status of the roads within the scope, aiding in effective resources distribution that focus on improving road safety.

1.6 Scope of the study

This research focused on roads within the environs of Nairobi, Kiambu and Nakuru. Road routes to consider, Nakuru-Eldoret highway, Mombasa Road, Nairobi's Waiyaki Way, Limuru Road, Kiambu Road and Thika Superhighway. In addition, the study majorly focused on road

accidents reports on injuries and fatalities data collected by NTSA, over a period of 3 years 2018-2020.

1.7 Limitation of the Study

The limitations of this research are as follows. First, generalization on the basis of scope considered against the entire road infrastructure in Kenya. Secondly, lack of access to road accident fatality autopsy reports limits the data enrichment. Finally, the lack of access to live and historical Global Positioning System (GPS) location for pedestrians' limits application of segmentation and personalization.

Chapter 2 : Literature Review

2.1 Introduction

In the new era, we cannot deny the fact that the transportation system greatly improves people's standard of living and contributes to a country's economic growth. It allows individuals to move more easily from one place to another to achieve their tasks across the world carrying along their goods. Transportation is naturally deemed as a representation of self-image and social well-being of a country, (Ashraf, 2019). Ashraf notes that Gross Domestic Product (GDP) made from transportation is frequently known to estimate the contribution of transportation in the economy of a nation. With the increasing stiff competition to supply sumptuous traveling service and human population the transportation infrastructure has continuously become complicated and larger. This concludes that the transportation system is a critical component of human advancement which also has potential if mismanaged to turn into a harmful component.

According to WHO (2018), the number of road traffic fatalities is increasingly rising reaching about 1.35 million in 2016. There exists a significant relationship between a country's level of income and the risk of road accident fatalities, for low-income nations the average rate of 27.5 fatalities per 100, 000 human population which is greater than 3 times in comparison to high-income nations with average rate of 8.3 fatalities per 100, 000 human population. As illustrated in Figure 2.1, the onus of road accident fatalities is significantly disproportionate among middle- and low-income nations in correlation to motor vehicles in operation and population size. Despite the fact that low-income nations account for 1% of the global motor vehicles, 13% of the global road accident fatalities occur in these nations. According to NTSA (2020) in Kenya 390 pedestrians have died on the road since the year began. This is according to the latest survey results by the National Transport and Safety Authority dated April 23, a fatality summary breakdown shown in Figure 2.2. Pedestrians form 38.16% of the total fatalities, an indication that in a given road accident the likelihood of a pedestrian being fatally involved is high.

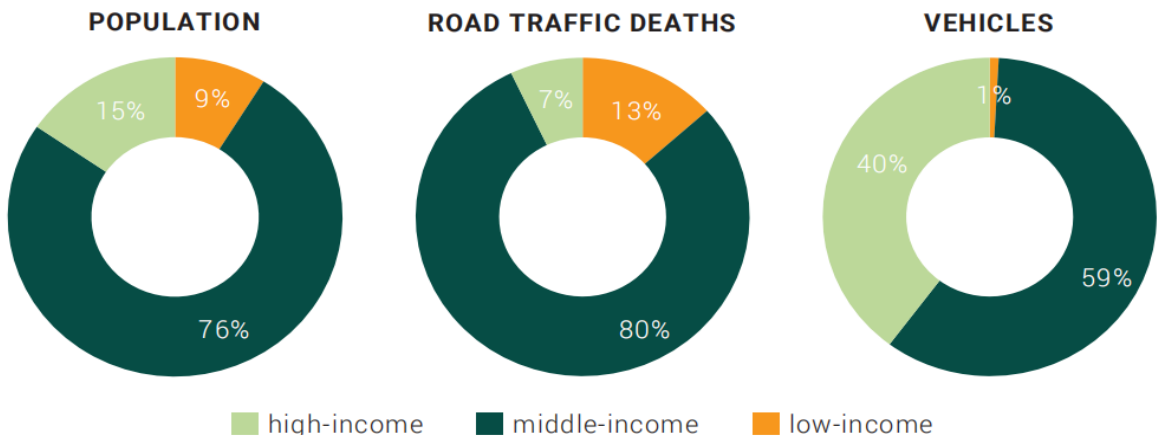


Figure *2.1* Percentage of registered motor vehicles, population and road accident fatalities by nation income classification, 2016 (WHO, 2018).

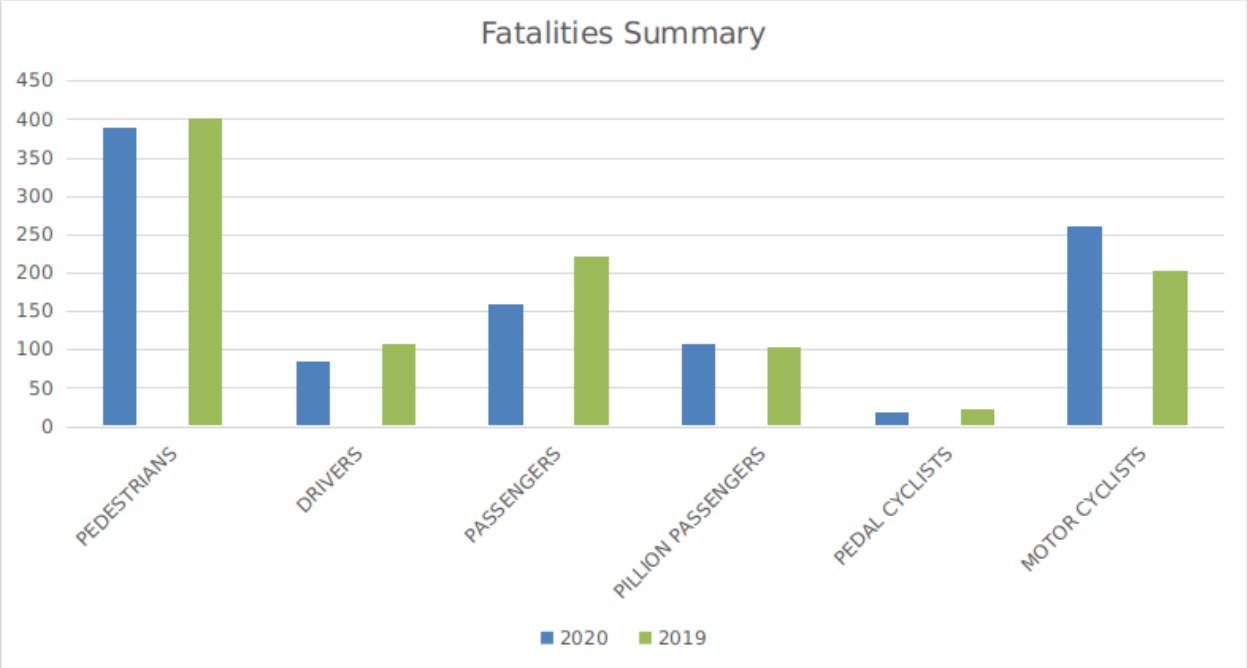


Figure *2.2* Fatalities Summary in Kenya, April 23, 2020.

2.2 Empirical Literature

2.2.1 Factors Contributing to Roads Accidents in Kenya

Being involved in a car accident causes some unwanted consequences which includes serious injuries, slight injuries, property damage and loss of life. In Africa factors contributing to

road accidents can be categorized into human, mechanical and environmental factors. According to Deme (2019) the following were identified as causes of road accidents; speeding, distracted driving, disregarding red lights and stop signs, reckless driving such as drunk driving, rain, improper turns, night driving, teenage drivers, recklessly changing of lanes, design defects, tailgating, wrong-way driving, driving on snow and ice, fog, potholes, road rage, deadly curves, street racing, drowsy driving, animal crossing and tire blowouts. Identifying factors that cause road accidents was one of the primary steps used to stop the problem that causes huge losses of life, injuries and property damage.

In Kenya the most factors contributing to road traffic accidents were caused by human errors, poorly maintained roads and defective vehicles which accounted for 59.6%, 19.5%, 29.9% respectively. Precursor factors related to road traffic accidents are overloading, over-speeding, and laxed policing (Deme, 2019). Deme noted that there is a multifactorial characteristic in causing road traffic crashes in Nairobi County and can be classified into vehicle, driver and roadway factors. Driver factors relate to all direct factors resulting in an accident occurring that can be affiliated to the driver. He also noted that speeding was a major cause of accidents and there is need for policy to be put in place towards addressing driver's behavior this is because it exposes the driver and other road users to accidents causing injuries and fatalities.

Manyara (2016) also noted that pedestrians, drivers, motorcyclists and vehicle defects are the most common causes of road accidents even though road defects and passengers do not significantly cause the accidents. Manyara concluded that the prime cause of road accidents was human error factor. The driver-related accident cases were caused by speeding, overtaking carelessly, misjudging clearance, cutting in and pulling from near side, which were found to be the top five causes of driver-related cases. He also stated that the Spearman Correlation Coefficient showed that there is a significant direct proportionality between the road accidents and drivers' behavior. Additionally, he also revealed that incidence of road traffic accidents was majorly associated with the driving behavior such as speeding. Finally, he concluded that drivers' personal characteristics influence the occurrence of road accidents. Contrary, his study identified that road traffic accidents in Kenya are influenced by road surface conditions.

According to NTSA (2020) the leading cause of road accidents losing control followed by failing to keep left, overtaking improperly, misjudging clearance, error of judgement, excessive speeding, stepping/walking running off vehicle, crossing road in non-designated area,

walking/standing in road, vehicle turning without due care, swerving, brake failure, drunk driving, stationary vehicle dangerous parking, falling from moving vehicle, boarding or alighting from a vehicle, failure of wheel-tires/wheels, slipping or falling and lane changing, as shown in the Figure 2.3.

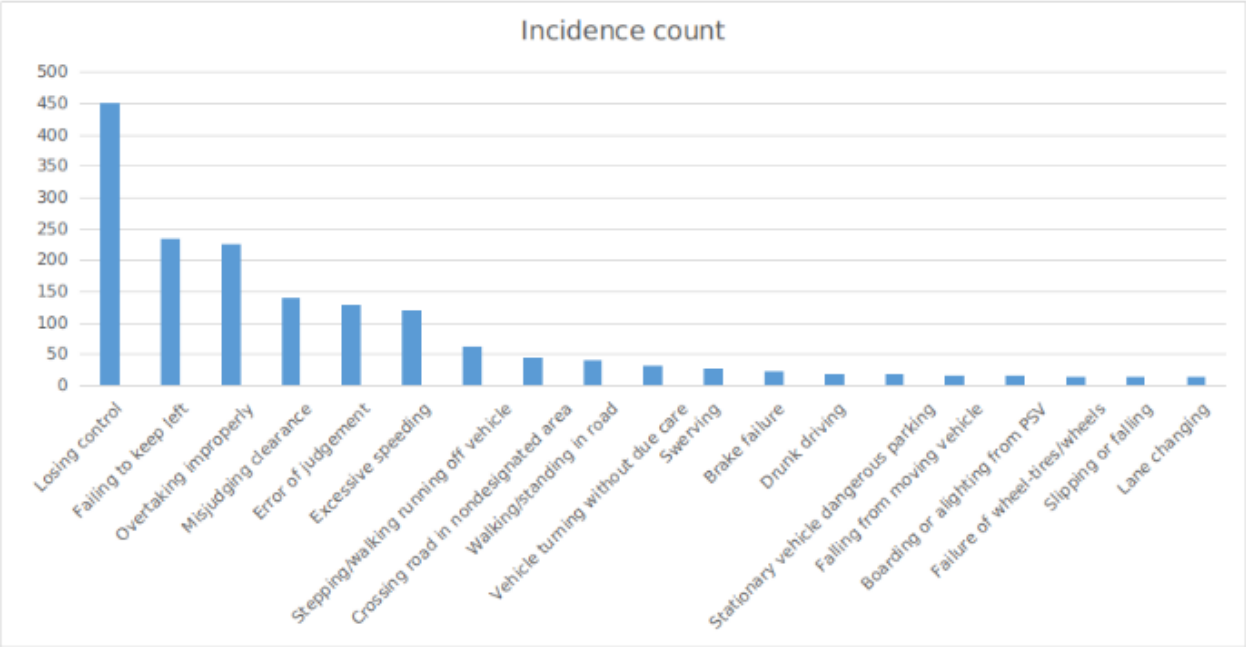


Figure 2.3 Incidence Count of Road Accidents Causes in Kenya.

2.2.2 Pedestrians Behavior and Movement

A major contributing factor to traffic accidents is alcohol-related impairment, however, impairment of pedestrians as a crash characteristic has a few studies focused on (Das, 2020). Human behavior and judgment are adversely affected by alcohol adversely, and it is one of the contributing factors in road accidents (Hezaveh, 2018). According to Das, walking is outlined as a healthy mode of traveling as it is associated with numerous health advantages. It imposes very few inauspicious effects to the transport infrastructure. Nevertheless, among the road users, pedestrians are the most unsafe mostly because of inadequate protection and poor transport infrastructure design, resulting in increased road accident risk of pedestrians. The patterns of unique certainty with high risks makes pedestrian road accidents an important traffic safety concern. According to the Hezaveh (2018) study, roughly 22% of the pedestrian road accident

fatalities were intoxicated with alcohol. His analysis indicated that walking while drunk, road accidents that occurred during the nights formed 83%, during the weekends was at 54%, occurred at mid-block sections was at 69% and areas with no traffic control was at 85%.

2.3 Framework and models

2.3.1 Linear Regression

Schroeder (2016) says that the basic and the most used type of predictive analysis is linear regression which focuses on examining two features: first, if a collection of independent variables helps in forecasting a target variable. Second, which variables are significantly independent of the target variable, and in which manner do these variables indicate by sign and magnitude of the beta evaluations of the target variable influence. These regression approximations are utilized in explaining the correlation between one or more predictor variables and one response variable. Schroeder notes that there exist three important uses for regression analysis which includes forecasting an effect, determining the strength predictor variables and trend forecasting.

Plotting and drawing a best line of fit of observations on variables x and y in a sample comprises linear regression analysis. This best line of fit is chosen such that the summation of the residuals is at minimum, where residuals refers to the shortest distance of each observation to the line of fit. Figure 2.4. This line is defined by equation 2.1

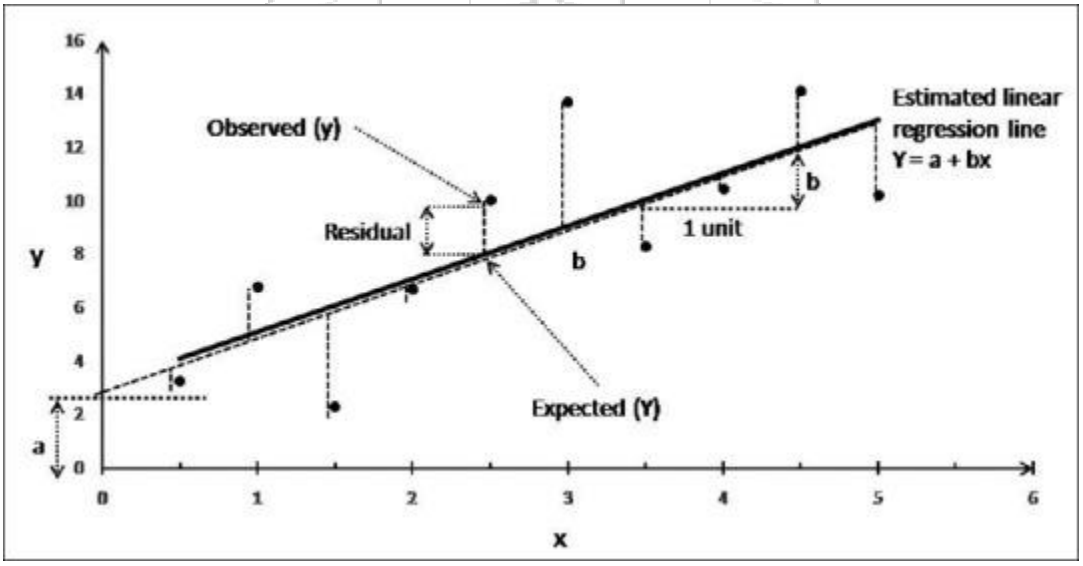


Figure 2.4 The Least Square Line, (Schroeder, 2016).

$Y = a + bx$

Equation 2.1 Least square line

Where:

x - the known value of observation

Y - the response value of y (target variable) for a observation

a - the constant of Y when x is 0

b - the gradient of the approximation.

Communication measures represent the impact of one unit increase of x on y which is also known as coefficient of regression or gradient, (Hayes,2017). However, linear regression analysis makes several assumptions; first, x and y have a linear relationship. Secondly, the observations in a sample should be independent of each other therefore if on any individual observation the data includes more than one observation this method should not be used. Furthermore, the data should not be having outlier values since these values may produce a false correlation in the data. Failure to meet these presumptions analysis of linear regression becomes ambiguous, (Schroeder,2016). Due to these limitations, linear regression is not suitable for predicting pedestrian road accidents but lays a baseline of why Bayesian analysis is best suited for this.

2.3.2 Bayesian Analysis

Reliability Re-allocation across likelihood is known as Bayesian analysis. It starts with a prior point of view in every likelihood followed by collection of data and re-allocation of reliability across the likelihood leading to a posterior point of view in every likelihood, (Zong, 2019). Bayes' rule as shown in equation 2.4, is used in evaluating the exact credibility re-allocation across parameter values, the specifications of which the evaluation uncertainty is directly shown by the posterior distribution extension. When posterior distribution extends across a wide scope of parameter values, this indicates high uncertainty because of things like having a small data set, but when the posterior distribution extends across a narrow scope of parameter values, this indicates high certainty due to things like having a large date set. Highest density interval (HDI) of 95% is the best way to outline the uncertainty, which holds the highest likelihood parameter values and that range the 95 % most likely values. This shows that parameter values within the HDI have high reliability, (Castillo, 2017). Bayesian analysis has been proved to be able to examine the multidimensional nature of crashes due to its strong ability in structure learning. It has been widely used in crash causation analysis and certificated to have satisfactory prediction accuracy, (Zong et

al., 2019), this makes it the most suitable approach for predicting pedestrian road accidents. Bayesian analysis entails the following algorithms; Likelihood, Parameters, Prior Distribution, Posterior Distribution, Bayesian Linear Regression.

$$Posterior = \frac{Likelihood * Prior}{Normalization}$$

Equation 2.2 Bayes theorem.

2.3.2.1 Likelihood

The likelihood maps each conjecture onto the number of ways the data can occur given the possibilities, which is achieved via plausibility function. In this stage, assumptions such as independence between observations can be made (Castillo, 2017). The gathered data y affects the resulting posterior distribution through the likelihood function r(y|θ), where θ is the parameters in the likelihood function. Castillo shows that given parameters for the chosen likelihood function, the probability for each occasion can be estimated.

2.3.2.2 Parameters

Often likelihood functions can have input parameters which can be adjusted to tune the model, and the different conjectures or explanations of the data are represented by these input parameters. Sometimes these input parameters have to be estimated, and this makes it difficult to distinguish them from data when performing Bayesian statistics. The traditional aspect of this is to view data as measured and known, while parameters are unknown and have to be estimated (Castillo, 2017).

2.3.2.3 Prior Distribution

The model has to be given an initial guess of the possible values for the parameters that are going to be estimated. This set of plausibility is the probability distribution for the parameters also called priors. Priors are assumptions, information that is known about the parameters and the prior distributions are used to give a reasonable range for the parameters (Grande, 2017). Depending on how the prior distribution is assigned, it varies in how much information it contains and how much impact it will have on the posterior distribution. A flat, or non-informative prior distribution can be used when there is a will to let the data speak for itself (Grande, 2017). Weakly informative prior, or regularizing prior, can be compared to the non-Bayesian statistical procedure with

penalized likelihood. When the model fits the training data too well there is a risk of overfitting which can be corrected by regularizing prior so the model will not become too excited. The prior basically tells the model to be skeptical. This weakly informative prior has to be tuned, because if the prior is too skeptic it will not allow any information and will miss out on important features in the data (Grande, 2017).

2.3.2.4 Posterior Distribution

For every unique combination of the observations, likelihood, parameters and prior, there is a set of estimates known as the posterior distribution. These estimates are the relative probability of the different values of the parameter's conditional on the given data, (Castillo, 2017).

2.3.2.5 Bayesian Linear Regression

In the Bayesian viewpoint, linear regression formulation is by use of likelihood distributions instead of observations estimates (Baldwin, 2017). The response y is presumed to be derived from a likelihood distribution. Bayesian Linear Regression model that has the response selected from a normal distribution is given by equation 2.2

y~ N (β^T x, σ^2 I)

Equation 2.3 Bayesian linear regression.

Where, y is the outcome of a normal distribution which is given by a variance and average. The average for linear regression is given by the product of weight matrix and forecaster matrix while the variance is given by the product of square of the standard deviation and the identity matrix. This is because the model undergoes a multi-dimensional formulation, (Baldwin, 2017).

The focus of Bayesian Linear Regression is to evaluate the posterior distribution for the model parameters. Both model parameters and response generated are presumed to come from a likelihood distribution, (Bourdache, 2019). The model parameters' posterior probability is dependent upon the training inputs and outputs as shown in Equation 2.3:

$$p(\beta|y,x) = \frac{p(y|\beta,x) \times p(\beta|x)}{p(y|x)}$$

Equation 2.4 Posterior probability distribution.

Where,

P (β|y, X) - is the model parameters' posterior probability distribution

P (y|β, X) - is the likelihood of the data

P (β|x) - is the prior probability of the parameters

P(y|x) - normalization constant

In contrast to Ordinary least squares (OLS) regression, Bourdache notes that the model parameters' posterior distribution is proportionate to the multiplication of the data likelihood and the parameters' prior probability. This bears the two core advantages of Bayesian Linear Regression. First, prior: if there exists field knowledge, or a clue for what the model parameters values should be, they can form part of modeling, far from the frequentist approximation where assumption about the parameters is entirely based on the data. Lastly, posterior: having the model possible parameters' distribution depending on prior and data set aids in estimating model uncertainty. Where posterior distribution spread out shows fewer data points exists, likelihood clears the prior as the amount of data points increases.

2.4 Designs and Architecture

2.4.1 Road Accident Severity Modeling with Bayesian Network

The Bayesian network is widely used for expert knowledge encoding with uncertainty in expert systems. Areas of applications of Bayesian network include decision support systems, document classification, information retrieval, data fusion, image processing, and medicine (Ying, 2017). Ying describes the Bayesian network as a graphical illustration of unsystematic variables with their corresponding dependencies. Figure 2.5 illustrates a basic form of a Bayesian network.

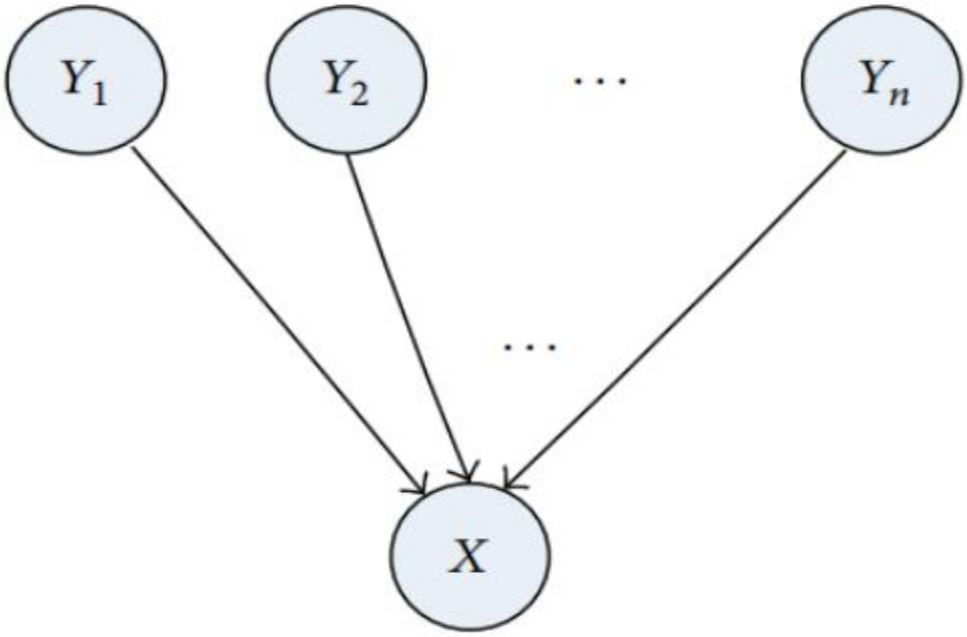


Figure 2.5 A simple Bayesian network, (Ying, 2017).

Finding the most probable network structure is computationally impractical due to the variable's exponential nature. Ying proposed the K2 algorithm, which is among the most common methods for Bayesian network structure learning. Apart from the basic Bayesian hypothesis, the presumptions used in the K2 algorithm include that variables are ordered and a prior of all structures are equally probable. This algorithm takes into consideration the ordering of each given node and ascertains parents' nodes namely; At first, presume that everyone node is an orphan then consequently where the resulting structure score can be increased, parents for some nodes are added. It tries to determine parents for every node up until no addition of parents can result in increasing the structure score, making the final structure to be with the highest score. The K2 algorithm and the Full Bayes Net Toolbox (BNT), a directed graphical modelling package supported by MATLAB, are used to model the severity prediction Bayesian network structure. In reference to the modeled structure, Bayesian estimation is employed to learn the parameters. Dirichlet distribution is assumed for all the variables' prior distributions, which allows closed-form solution for prediction and closed form for parameters posterior distribution. The Full-BNT is utilized to achieve the Bayesian estimation algorithm, (Zong, 2019). The parent nodes influence is focused while these nodes collect the indirect nodes effects and convey them to the child nodes. In examining models' accuracy, Hit ratio and Mean Absolute Percentage Error (MAPE) indicators

are used. MAPE, which computes the mean percentage dissimilarity between response values and actual ones, (Zong, 2019).

The fatality forecasting model has a MAPE value of 0.0226 and 100% Hit ratio, both of which indicates a reliable model precision, (Ying, 2017). In line with the developed structure Vehicle condition (Vc), Number of injuries (Noi), and Location-Regular road section (L-Rrs) are Number of fatalities (Nof)'s parent nodes. The results from estimation shows that as the condition of the vehicle gets worse the likelihood of a fatal road accident incident increases. In addition, a larger number of fatalities is correlated with a larger count of injuries. When road accidents occur, more fatalities are observed at a normal road section compared to an abnormal section or road intersection. The rationale may be that at intersections or abnormal sections of the road, the implicated vehicle generally reduces speed when driving through them. The forecasting model of injury bears 100% Hit ratio with 0.0013 MAPE, which stipulates a reliable model precision. The number of injuries in a road accident is directly impacted by the two parent nodes, Bus or truck involved (Bti) and Vc. The results from estimation indicates that the bus or truck implicated in accidents tend to cause additional injuries. Additionally, there exists a positive correlation between the condition of the vehicle and the number of casualties. The property damage forecasting model has a 100% Hit ratio with 0.0019 MAPE, which indicates a high model precision. Vc and L-Rrs, parents nodes have absolute influence on property destruction. The outcomes conclude that poor vehicle maintenance condition is directly proportional to property damage. In addition, large amounts of property damage are observed when accidents occur at irregular sections of road or intersection. Considering the combined influence of L-Rrs on Nof and Property damage (Pd), it is concluded that the road accidents occurring at regular road sections incline to sequel in high fatalities numbers with slight amounts of property destruction, (Ying, 2017).

2.4.2 Road Accident Severity Modeling with Regression

Logistic Regression and the Ordered Probit models are the most frequently utilized regression models based on road accident analysis, (Schroeder, 2016). Logistic Regression models fail to take into consideration the ordering of the target variable since Pd and Noi substitutes are ordered and also have irrelevant alternatives independence problems. The Ordered Probit model is utilized in predicting Pd and Noi. For Nof prediction the Binary Logit which is among the Logistic Regression modeling, is adopted, which has two discrete alternatives, (Ying, 2019).

Binary Logit model is among the Binomial choice models, which is frequently used in modeling of discrete choice. As per to the random utility theory (Schroeder, 2016), given by the following equation 2.5

$$Uin = Vin + \varepsilon in$$

Equation 2.5 Random utility.

Where,

i - is use of alternative (i=1 or 2 for Nof = 0 or Nof ≥ 1)

n - is the number of accidents

Vin - indicates the deterministic component of Uin

εin - is the random component of U.

The Ordered Probit, which presumes a standard normal distribution, is the frequently used ordered multiple choice model (Schroeder, 2016).

2.4.3 Estimation Modelling Results.

The Binary Logit and Ordered Probit models are evaluated by using logistic and probit procedures, (Ying, 2019), and the outcomes are shown in Table 2.1.

Table 2.1 Estimation of Binary Logit and Ordered Probit Models, *(Ying, 2019)*.

Variables	Fatality forecasting model		Injury forecasting model		Property damage forecasting model	
	Coef.	Z-stat.	Coef.	Z-stat	Coef.	Z-stat
Constant	-2.57	-12.53				
Mi	0.44	3.36	-0.14	-2.50	-0.09	-1.85
Bi	1.11	8.82	-0.30	-4.99		
Wc	-0.27	-1.66	0.23	2.70	-0.12	-1.85
Tod	-0.44	-4.03			-0.15	-3.29
Vd					0.10	4.88
Pc			-0.38	-5.24		
Tsc	0.22	2.06	-0.07	-1.57	0.11	2.76
L-Mvl			-0.08	-1.57		
L-C	-0.52	-1.39			-0.40	-3.51
L-Rrs	0.73	5.98			0.72	4.07
L-I			0.23	4.94	0.20	1.10
$\alpha 1$			-1.50		0.80	
$\alpha 2$			1.47		2.77	
MAPE	0.0530		0.0425		0.0698	
Hit ratio (%)	84.65		80.20		60.23	

The estimation outcomes indicate that, in contrast with the Regression model, the Bayesian network emerges better fitting for road accident severity forecasting. Ying study outcomes can be employed in forecasting road accident severity and determining the key factors that influence road accident severity. In contrasting the Bayesian network, parameters and the structure of the Bayesian network changes when new records reach 5% and 10% of the actual cases number,

respectively. Ying concluded that by differentiating the test outcomes of Hit ratio and MAPE with respect to the three severity indicators predictions, the goodness of Bayesian network relevance exceeds the Regression model. This concludes that the Bayesian network is a better fit than Regression models in road accident severity prediction, Figure 2.6.

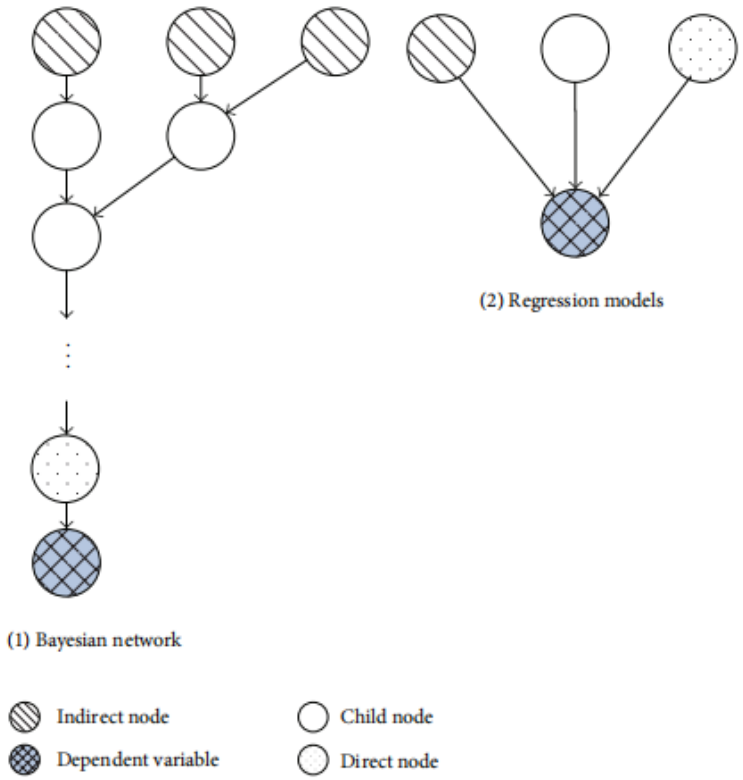


Figure 2.6 Comparison of Regression and Bayesian network models, (Ying, 2019).

2.5 Related Applications / Solutions

2.5.1 Driving Risk Status Prediction

Yan et al (2017) proposed a system to predict driver risk status using Bayesian networks and logistic regression. Their study notes that the ability to determine driving risk status acts as a key role for reducing the number of traffic accidents. In order to extract the key factors that significantly impacted driving risk status, Bayesian network (BN) was applied. The key factors identified based on driving simulation experiments were sex, environment, driver state, experience and vehicle state. Then, a logistic regression was utilized to develop the driving risk status prediction model, and to evaluate the model performance the receiver operating characteristic

curve was used. The area under the curve was 0.903, an indication that the model developed was both practical and adaptable.

2.5.2 Causation Analysis of Road Accidents

Zou & Yue (2017) proposed a system for causation analysis of road accidents using Bayesian network approach. Bayesian network model was arrived at by taking Adelaide Central Business District (CBD) in South Australia as a case, the Bayesian network structure was developed by integrating K2 algorithm with experts' knowledge, and Expectation-Maximization algorithm for processing missing data was selected to conduct the parameter learning in Netica, which is Bayesian network development software. Then posterior probability reasoning, inferential analysis and the most probable explanation were carried out. The results indicated that the model could effectively explore the complex logical relation in road accidents and convey the uncertain association among related variables.

2.5.3 Google Maps API

Google map supports API ability to build a web or mobile based application. Developers are able to integrate Google Maps with mobile apps via android operating system thus enabling the users to access google maps features such as location display, showing different road routes with names on the map, as shown in Figure 2.7. Google map API are so helpful in making a data distribution mapping since Android supports real-time process of determining coordinates by using the GPS technology. A Google Map API key is needed to amalgamate maps in Android apps (Rahmi, 2017). When pedestrians use the mapping and navigation applications in their smartphones so as to be able to determine and use safe paths in cities. The visibility of walkable paths in geographical maps can be strongly associated with their safety. In regards to that, google maps can be referred to a crucial user interface for understanding the safest routes for pedestrians avoiding traffic risks such as accidents (Kapenekakis, 2017).

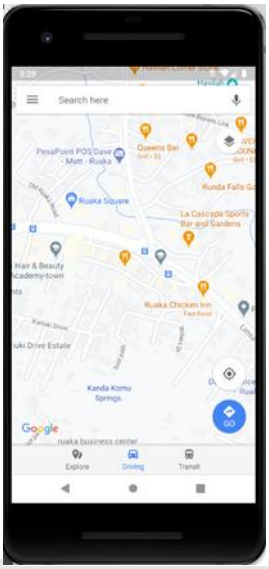


Figure 2.7 Google Map, showing different road routes with names on the map.

2.6 Research Gap

Although very many researchers have done research on predicting road accidents likelihood with key considerations being the driving behavior, there is little research which has been done around pedestrians' road safety in relation to vehicle-pedestrian road traffic accidents.

2.7 Conceptual Framework

The first process entails data extraction of existing and new traffic offences such as inappropriate overtaking, speeding, drunk driving among others, and accident fatalities. The data is then cleaned to handle missing values, transformed to the required data types numerical and categorical and modeled into documents for storage in a NoSQL database. Followed by data visualization in the database producing visualization reports which aid to deriving insights, identifying and handling outliers and discovering patterns. Feature engineering is informed from the visualization results and the new features identified if any, this aids in fine tuning the prediction algorithm in determining the probability of having an accident in a given road route. Based on the road accident prediction against available road routes the mapping is done and presented on a google map, Figure 2.6.

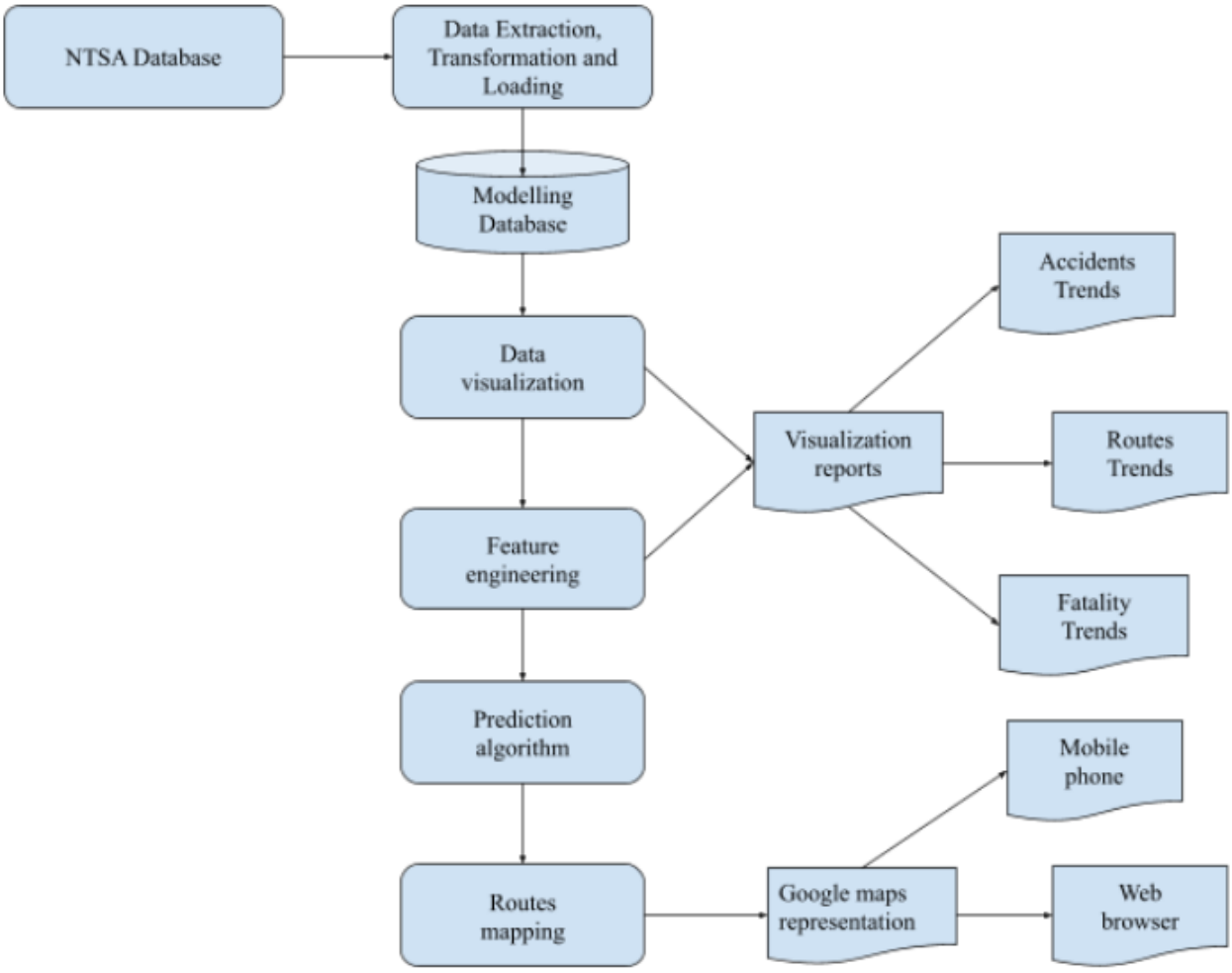


Figure 2.8 Conceptual Framework.

2.8 Operationalization of Variables

The breakdown of the navigation aid model variables is as shown in Table 2.2.

Table 2.2 Navigation aid model variable

Concept	Variables	Indicator
Road accident	• Time • County • Road • Victim • Number of fatalities • Cause • Age -	• Time of accident • County in which the accident happened • Road route on which the accident happened • The type of victim involved in an accident • Counts of fatalities in an accident • Indicates the cause of the accident • The age of the victim
Pedestrian Navigation safety	• Current location • Route fatality likelihood	• GPS coordinate • Route color

Chapter 3 : Research Methodology

3.1 Overview

This chapter introduces research strategy and the technique applied. In more details, the research design, the research method and approach, the data collection method, data analysis, research reliability and ethical considerations as well as results utilization and dissemination of the project are outlined in this section.

3.2 Research Design

Leavy (2017) noted that research design is a major methodological thrust of the research work, being the specific and distinctive approach that is usually used to provide answers for the study research inquiries. The main aim of the research guided by the objectives and research questions influence the type of research design to be used. Bloomfield (2019) noted that the motive of the research design is to guide the study in examining the research problem hence refining the validity of the study. So as to decide which type of research design is suitable for the study, it is important to put into consideration a number of factors such as the person or object of data collection this is the unit of analysis, the focus of research (orientation or action) and the time dimension. This study used experimental research because it facilitates the manipulation of variables which is essential in this study, (Kalu, 2017). For instance, road traffic fatalities related with road crashes in Kenya and current strategies of determining crash prone road routes as well as the movement behavior of the pedestrians in relation to occurrence of road accidents.

In designing the system, the research implemented Agile development methodology. Agile software development refers to methodologies of software development that are placed around the idea of iterative development. The solutions and its requirements unravel through cooperation between self-organizing cross-functional teams. Dhir (2019) says that Agile development is better to use because it allows the users to deliver values faster and easily with greater quality and predictability as well as greater aptitude to respond to change. This research employed SCRUM technique Figure 3.1, an agile process which enables users to focus on delivering values very fast.

SCRUM continuously examines actual working software thus emphasizing teamwork, iterative progress and accountability in relation to a well-defined goal.

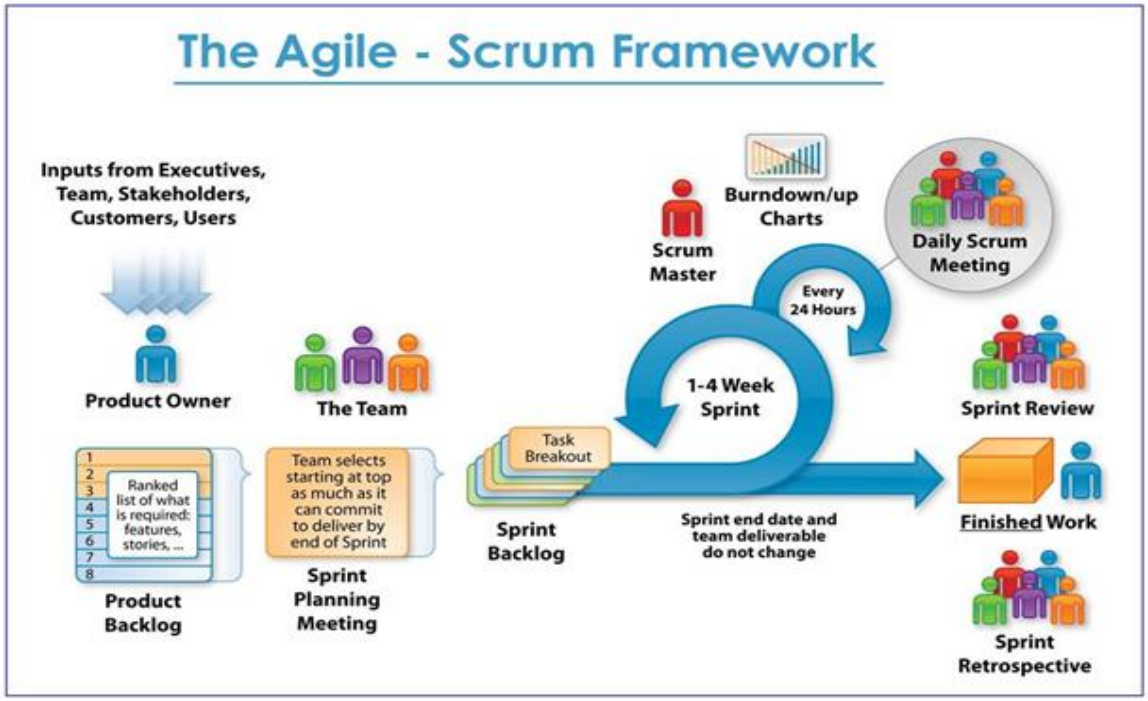


Figure 3.1 The Agile - Scrum Framework, (Jha, 2016).

The Scrum Framework is able to deal with the fact that there is a likelihood for requirements changing or mostly being unknown at the start of the project (Jha, 2016). Moreover, Jha notes that SCRUM handles the complexity in work by ensuring that the information is transparent so that people can easily inspect and be able to adapt based on the current conditions instead of the predicted conditions. This enables the teams to address the common challenges of a waterfall development process which includes chaos resulting from the constantly changing requirements, time underestimation, cost and resources, inaccurate progress reporting as well as compromises on software quality. The research design will have the following style:

3.2.1 Acquisition of Data

The data included secondary sources of data from NTSA and other documentation relevant to the study. The secondary data collected included the inputs recorded when a road accident occurs such as time, county, road name, type of victim, number of fatalities, cause and age. These

values were used as inputs for the Bayesian model for the purposes of training, testing and validation of the model.

3.2.2 Sample Split

The sample split aided in identifying the type of data that was included for each dataset by identification of the data that was used to train the model and the data used to test the model. Where the test dataset entails the sample of data used to fit the model while the training dataset entails the sample of data used to provide an unbiased evaluation of the final model fit on the training dataset.

3.2.3 Model Training

This involved the provision of inputs to the proposed Bayesian model that was used for training and processing on the provided input data and then getting the anticipated output. The training data was fed into the Bayesian inference algorithm. Several iterations were taken during the training process which were aimed at reducing the error rate and fine tuning the model parameters. Each dataset that was collected was partitioned into training set and test set in order to determine the likelihood of a pedestrian road accident in a given road route of malaria, where the likelihood is the posterior distribution from the Bayesian model.

3.2.4 System RESTful Application Programming Interface

In order to allow seamless communication between the system database, the Bayesian model, the web portal and the mobile application RESTful Application programming interface (API) was used. In order to query the Bayesian model, the model query function was exposed as an API endpoint. The data analysis and model prediction outcomes stored in the database were exposed as API endpoints in order to be easily accessed by the web portal. The mobile application was able to make a request to the Bayesian model while recording the request at database level via the API endpoints.

3.2.5 Mobile Application Development

Mobile application development involved the use of Android programming language in designing and implementing the application interface. The interface allows the pedestrian to view the safety status of the available routes via coloring on a Google Map, get a quick tour of how the application works and give feedback on the service experience. Network libraries and dependencies were used to facilitate the network calls to the backend APIs endpoints. Global positioning system (GPS) dependencies were included to enable the application to get the last known location of the user.

3.2.6 Web Portal Development

The system's web portal was developed using Hypertext Markup Language (HTML), Cascading Style Sheets (CSS) and JavaScript. HTML enabled the implementation of the web portal layout detailing the necessary road accidents reports, while CSS enhanced the appearance of the layout and JavaScript enabled the handling of the network calls to the APIs endpoints for system login and reports.

3.2.7 System Testing

System testing involved running through test scenarios, where each scenario had a pass or a failure. Any test scenario that failed, which implied the failure to satisfy set conditions for acceptance criteria, was picked as a system bug and fixed. All the set system test scenarios were executed with a pass rate of 100%, the test scenarios were categorized into two, model and mobile application test cases. Model test cases included; use of test data on the model and use of new observation on the model, while Mobile application test cases included; location access permission request, location access status, display of last known location marker, display of the safety status of the available routes and capturing user feedback.

3.3 Target Population

According to Taherdoost (2016), population refers to a group of individuals with similar characteristics that attract attention to the researcher. In this study the target population was made up of accident data from road routes; Nakuru-Eldoret highway, Mombasa Road, Nairobi's Waiyaki Way, Limuru Road, Kiambu Road and Thika Superhighway, as collected by NTSA.

3.4 Sampling Design

Purposive sampling was conducted in this research. This method is categorized as non-probability sampling techniques where samples are selected on the basis of their knowledge, relationships and expertise regarding the research subject, (Etikan, 2016). In this study, the route sample data was selected according to the road accidents that involving pedestrians. Cochran's formula was used to calculate the sample size of this research, (Khoshi, 2018). Where by the representative sample for proportion was calculated as follows;

$$n_0 = \frac{z^2 pq}{e^2}$$

Equation 3.1 Cochran's formula

where,

n_o = Sample size

Z = The selected critical value of desired confidence level (1.96 for 95% confidence level)

P = The estimated population is assumed to be 0.5 (50%)

q = 1-p

e = Desired level of precision should be within ±5%

If the population is finite, Cochran also mentioned that the sample size can be reduced because in proportion a very large population provides more information than a smaller population. He therefore proposed a correction formula to calculate the final sample size as shown below;

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}}$$

Equation 3.2 Cochran's correction formula

Here,

n_o = Sample size derived from the above equation and,

N = Population size.

3.4 Data Collection

For the objectives of this thesis, data form secondary sources were used. Secondary data used includes road accident records from NTSA and other documentations relevant to the study. These data try to inform on the underlying factors associated with road crashes, determining which road routes are more prone to road accidents and how pedestrians' movements relate with the road

accidents observed. The minimum number of observations required using Cochran's formula is 385, more than 1000 observations were collected.

3.5 Data Analysis

The identified software used for data analysis was Pandas package, a Python library, which is a very powerful package for data analysis. It is created on the top of the NumPy package, (Kabita, 2019). Plotting works from Matplotlib and machine learning algorithms in Scikit-learn. Kabita indicates that Pandas supports Exploratory Data Analysis (EDA) which is an approach used in summarizing the data by taking their main characteristics and visualizing it with proper representations. EDA entails, first, data description, which shows the number of rows and columns in a data sets, data types, missing data, preview and statistical summary for numerical observations which indicates the mean, median, quartiles, standard deviation and most frequent observation. Secondly, data cleaning, which involves handling of missing data, invalid data types and incorrect values. Thirdly, data visualization, shows data distributions by use of bar charts, line graph, pie chart, histograms, box plots among others. Finally, calculating and visualizing correlation between variables using heat map. Secondary data was formatted among the file formats supported by Pandas package, followed by loading the data into Pandas package then EDA was performed.

3.6 Research Reliability

3.6.1 Data Reliability

Data reliability is a measure of the stability or consistency of the results gathered, (Etikan, 2016). Etikan argues that it reflects consistency and replicability over time. The output data of the model was tested on several occasions on the training phase of the agile development. In return, by doing so there was reduced error rate and the inputs for each variable at this point was near optimal. The measures of performance were computed for the evaluation of the model.

3.6.2 Data Validity

The variable inputs that were used for the model were used to observe the likelihood of road accidents occurring in a given route and the output was compared to the actual road accidents data collected. Data validity was done in order to confirm that the data used for the study is true representation of the given inputs and that the given inputs have been used in the model.

3.7 Ethical Considerations

Use of secondary data is in itself, a highly ethical practice Jol (2016) argued that it maximizes the value of any (public) investment in data collection, reduces the burden on respondents, ensures replicability of study findings and therefore, greater transparency of research procedures and integrity of research work. But the value of secondary data is only fully realized if these benefits outweigh the risks, notably in terms of re-identification of individuals and disclosure of sensitive information. For this to happen, use of secondary data in this research must meet some key ethical conditions as follows; firstly, data must be de-identified before release to the researcher, secondly consent of study subjects can be reasonably presumed, thirdly outcomes of the analysis must not allow re-identifying participants and lastly use of the data must not result in any damage or distress.

3.8 Utilization of Results

The output of this study is being used by pedestrians to be aware of the road route safety status and make decisions based on available alternate safer road routes or the necessary safety precautions to undertake. For the traffic officers the output is used to drive insights on traffic resources distributions flagging routes which require more attention/resources and how to effectively optimize available resources. Finally, for the transport ministry the output can be used to inform decisions in road infrastructure design and development.

3.9 Dissemination of Results

Pedestrians can access the output of this study via a stand-alone mobile application that displays a Google Map with available road routes having safety status indication. For the

traffic officers and transport ministry the output is accessed via a web browser portal that displays graphical and tabular representation of the road routes safety analysis.

Chapter 4 : System Analysis, Design and Architecture

4.1 Introduction

System design is the phase that bridges the gap between problem domain and the existing system in a manageable way, (Oyelere et al., 2018). This phase focuses on the solution domain, i.e. “how to implement?”, it is where the system requirements specifications are converted into a format that can be implemented and decides how the system will operate. In this phase, the complex activity of system development is divided into several smaller sub-activities, which coordinate with each other to achieve the main objective of system development. This chapter details these sub-activities.

4.2 System Analysis

System analysis deals with planning the development of information systems through understanding and specifying in detail what a system should do and how the components of the system should be implemented and work together, (Oyelere et al., 2018). This approach made it easier to perform requirement gathering.

4.2.1 Requirements gathering

The data was downloaded from NTSA online database, the body in charge of safety and transportation oversight in Kenya. NTSA has two main types of data available to the public as a transparency and sensitization of safety: Daily reports and fatal accident reports covering the entire country. Data collection involved downloading all entries availed in the website entry by entry. Daily reports show the fatalities and injuries categorized as: pedestrians, passengers, drivers, pillion passengers (motorcycle passengers), pedal cyclist and motor-bike cyclist. In fatal reports, the agency avails fatality incidences citing the place (County), time, details of the victim and category. Fatal accident report shows time, Base/Subbase, County, Road, Place, motor vehicle (MV) Involved, Brief Accident Details, (Name of Victim), gender, age, cause code, victim, and no. of fields as entry columns. Of this, we are concerned with time, County, brief accident details and victim. The data covers the entire country subdivided into 47 regions (counties), from which Nairobi, Nakuru and Kiambu counties data was selected to meet the study scope.

Data in the fatal accident report is as shown in Figure 4.1 based on time of day and Figure 4.2 based on county. From the figure, accidents are prevalent in rush hours (from 1600 hrs to 2100 hrs). In terms of county, Nairobi County has the highest incidences followed by Nakuru and Kiambu county. Figure 4.3 shows that pedestrians form 54.14% of the accidents fatalities and Figure 4.4 indicates that most of the accident’s fatalities happen on a weekday, Monday having the top percentage of 21.30%.

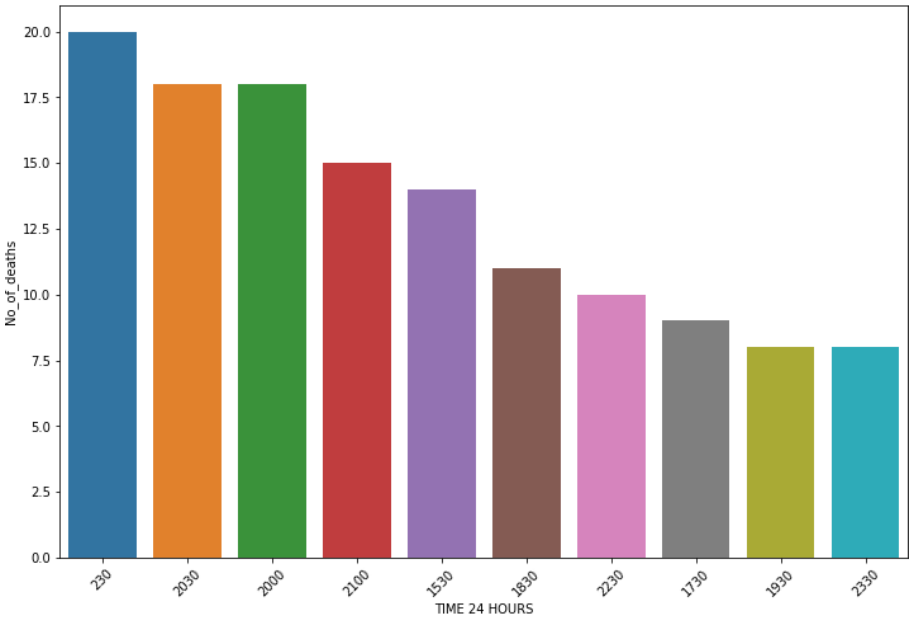


Figure *4.1* Hourly distribution of accidents.

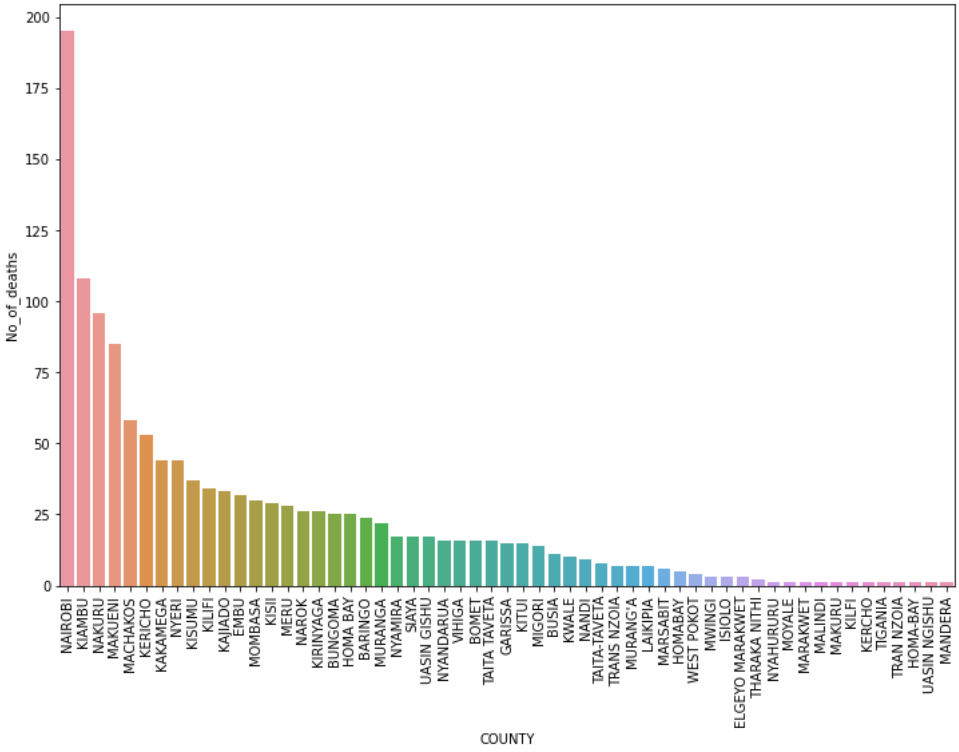


Figure *4.2* County distribution of accidents ranked with prevalence.

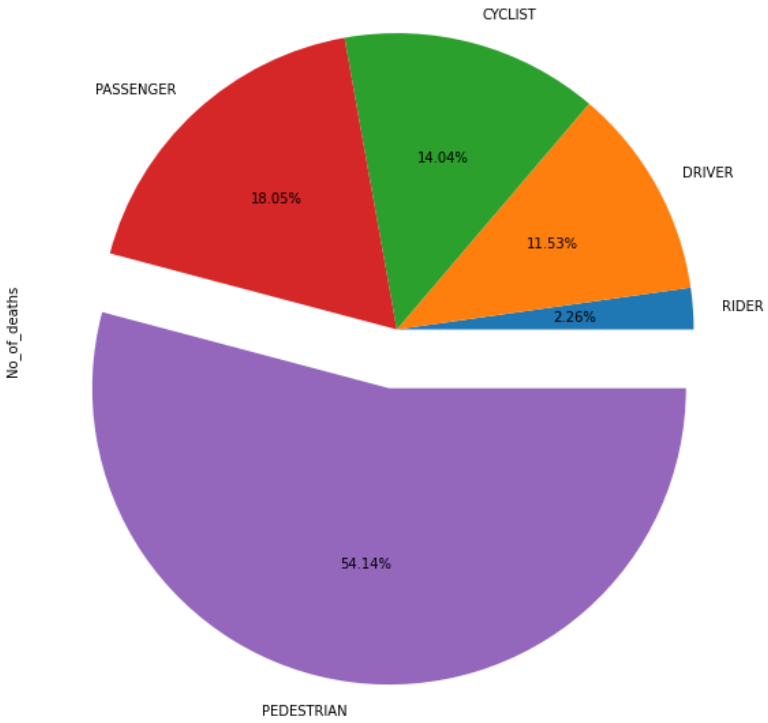


Figure *4.3* Accidents victims summary by percentage.

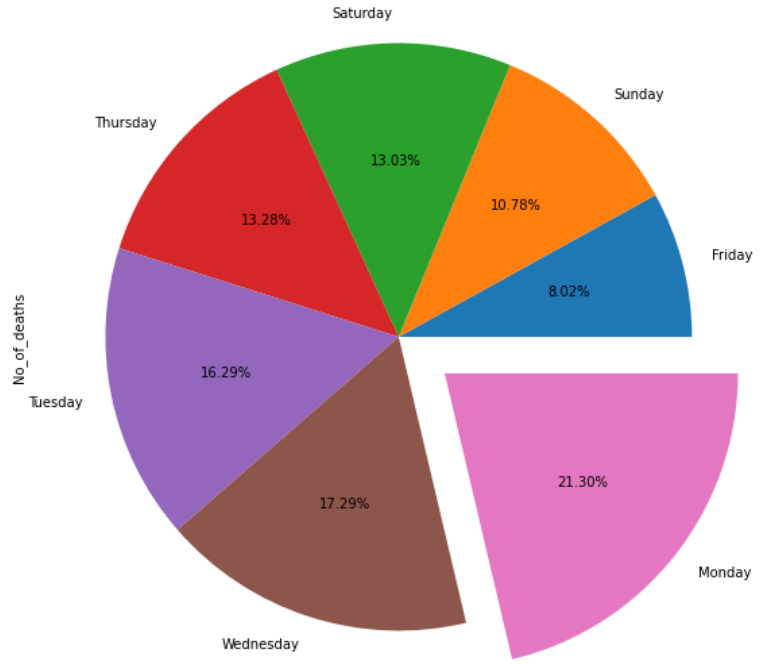


Figure *4.4* Accidents fatalities distribution per week day by percentage.

4.2.2 Functional Requirements

i. A user should be able to download the android application through either an application store or similar service on the mobile phone. The phone should have an android OS version of either Kitkat (4.4), Lollipop (5.0), Marshmallow (6.0), Nougat (7.0), Oreo (8.0) or Pie (9.0).
ii. When a new/updated version or release of the software is released, the user should be notified on next system use. The download of the new release should be done through the mobile phone in the same way as downloading the mobile application.
iii. The android application should allow the user to enable gps location.
iv. The android application should display all the routes available with their safety status coloring where dark red indicates almost certain (0.91-1.0), red for likely (0.9-0.71), brown for possible (0.7-0.41), yellow for unlikely (0.4-0.21), green for rare (0.2-0.0).
v. The system should allow the user to give feedback.
vi. The web portal should allow the user to login and manage the account.
vii. The web portal should allow the user to visualize and pull route-based road accidents reports.

4.2.3 Non-functional Requirements

i. Performance and scalability: Performance defines how fast a software system or its particular piece responds to certain users' actions under certain workload. In most cases, this metric explains how much a user must wait before the target operation happens (the page renders, a transaction is processed, etc.) given the overall number of users at the moment, (Eckhardt et al, 2016). Eckhardt indicates that scalability assesses the highest workloads under which the system met the performance requirements. The system uses scalable Application Programming Interfaces (APIs) endpoints which are able to support 500 requests per second, over an LTE connection.
ii. Portability and compatibility: Eckhardt states that portability defines how a system or its element can be launched on one environment or another. It usually includes hardware, software, or other usage platform specification. Compatibility defines how a system can co-exist with another system in the same environment. For instance, software installed on an operating system must be compatible with its firewall or antivirus protection. The user android app must support phones running on android OS versions; Kitkat (4.4), Lollipop (5.0), Marshmallow (6.0), Nougat (7.0), Oreo (8.0) and Pie (9.0). The system's backend services support orchestration via docker engine which enables the services to be platform independent.
iii. Reliability and Availability: Eckhardt argues that reliability specifies how likely the system or its element would run without a failure for a given period of time under predefined conditions. There are three ways to measure it; probability percentage, the number of critical failures and mean times between failures. The web dashboard and the mobile app must be available to users 98 percent of the time every month. The system uses

logs for monitoring, the logs capture warning and errors information which can be used to detect and debug an issue.

iv. Security: assures that all data inside the system or its part is protected against malware attacks or unauthorized access. The web dashboard supports SSL and access levels authorization. The system uses self-signed SSL certificates and basic authentication while accessing the APIs endpoints.

v. Usability: addresses a simple question: How hard is it to use the product? Eckhardt indicates that it can be evaluated within the following five dimensions; First, learnability, how fast is it for users to complete the main actions once they see the interface? Secondly, efficiency, how quickly users can reach their goals? Thirdly, memorability, can users return to the interface after some time and start efficiently working with it right away? Fourth, errors, how often do users make mistakes? And lastly, satisfaction, is the design pleasant to use? The mobile application includes the help menu which shows a quick tour of how the application works.

4.3 System Architecture

The architectural design of the proposed system is as shown on Figure 4.5. The system requires location input from the pedestrian, and displays the safety status of the available routes based on the specified destination. The system allows traffic officials to access summary and trends reports of road accidents by road routes. The system components include the following, firstly a mobile android app that allows the pedestrian to share their curet location and then road routes within the scope near the pedestrian are displayed with their safety coloring, and the app allows the pedestrian to given feedback on experience and improvement comments. Secondly, server which hosts the data services such as scoring service which is used to calculate road route accident frequency, prediction service which is the Bayesian Linear Regression model used to score the road accident fatality of a given road route, report service is used to aggregate the necessary reports which are accessed via the web portal, feedback service is used to process the feedback given by the user and user management service is used to process user data for users accessing the system via the web portal. Thirdly, Web portal which displays the summary and trend reports of road accidents fatalities data and enables an admin user to manage users accessing the portal. Finally, Database forms the last component which is used to store data that includes but not limited to users, road route reports, configurations, routes, accidents fatalities, model weights, feedback and devices.

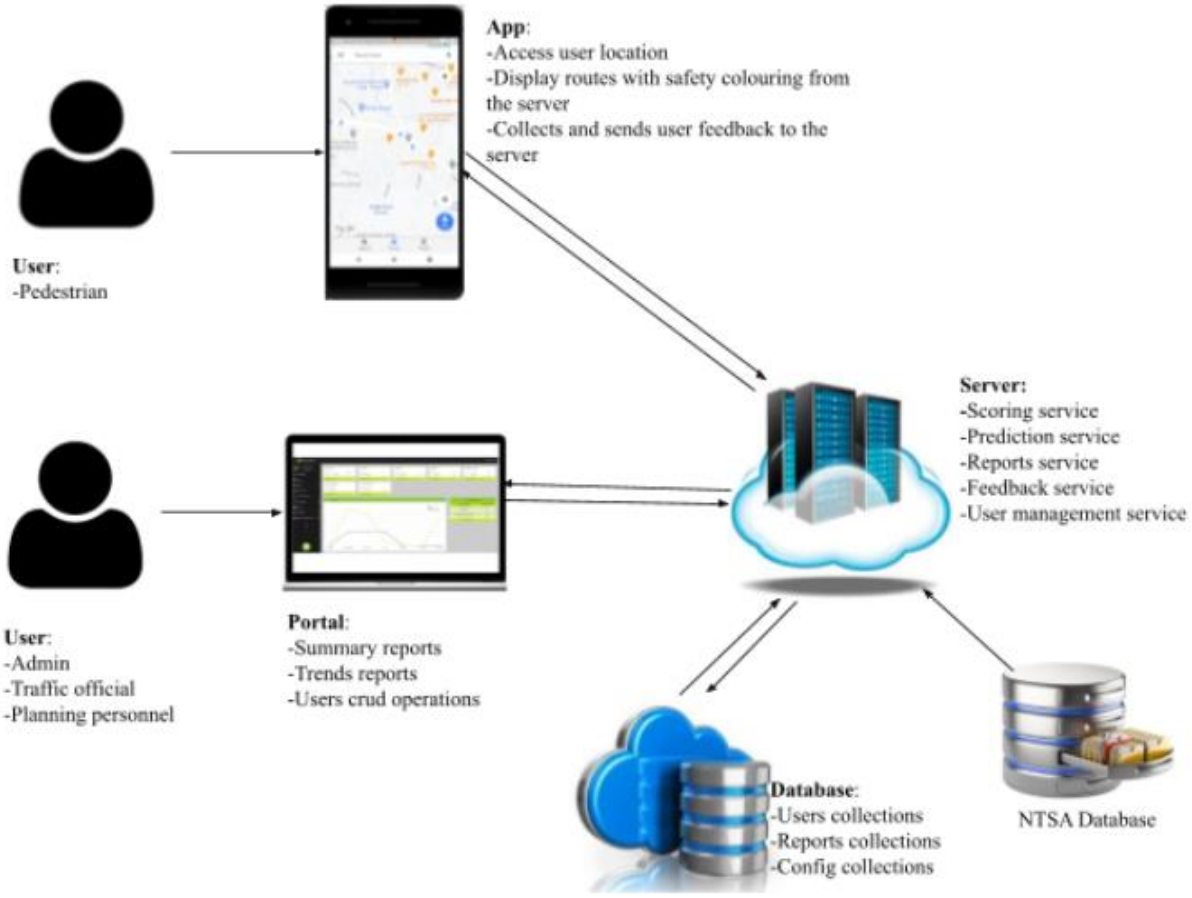


Figure *4.5* System Architecture Diagram.

4.4 System Design

Systems design is the process of defining elements of a system like modules, architecture, components and their interfaces and data for a system based on the specified requirements, (Shylesh, 2017). Shylesh states that it is the process of defining, developing and designing systems which satisfies the specific needs and requirements of a business or organization. A systemic approach is required for a coherent and well-running system. Bottom-Up or Top-Down approach is required to take into account all related variables of the system. A designer uses the modelling languages to express the information and knowledge in a structure of system that is defined by a consistent set of rules and definitions. The designs can be defined in graphical or textual modelling languages. The system design methodology used is object-oriented systems analysis and design which facilitates logical, rapid, and thorough methods for creating new systems responsive to a changing business landscape.

4.4.1 Use case diagram

Figure 4.6, illustrates the use case diagram of the system. The main actors are the pedestrian, system admin and traffic/planning officer. Each actor interacts with the system based on use case scenario, firstly, for a pedestrian the scenarios include Location capture and Route information. Location capture includes the app download/update use case, this is to indicate that in order to access the last known location of a pedestrian, the pedestrian is required to have downloaded the android app. Route information use case which entails the display of the road route safety coloring on a Google map includes the Route safety prediction use case which provides the road routes predictions results from the Bayesian Linear Regression model. Route information use case also extends user feedback use case which allows the pedestrian or traffic officer to give feedback on app/system experience. Secondly, for a traffic or/and planning officer the scenarios include Login use case where the user provides login credentials in order to access the system and Report generation use case entails the user been able to access road fatalities summary reports, alerts reports based on changes of observed in new data added if any there be and trends reports that shows the patterns of road accidents fatalities. Finally for a system admin the scenario includes user management which entails creation of users' accounts, deactivation and reactivation of the users' accounts.

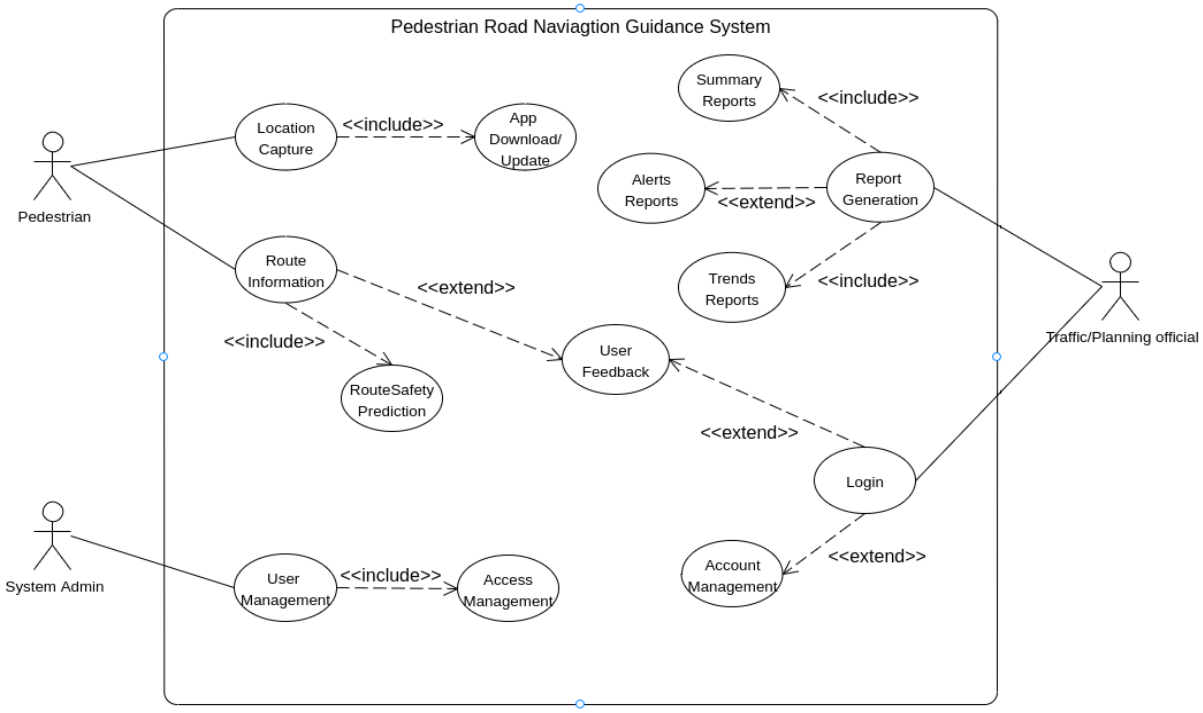


Figure *4.6* Use case Diagram.

4.4.2 Sequence diagrams

System Sequence Diagram for the android mobile app is as shown in Figure 4.7, illustrating the interaction between the pedestrian, and android mobile application. The interactions are initiated by the pedestrian after downloading the android mobile app. The app requests the pedestrian to allow location access via a prompt, once the pedestrian agrees to location access the app displays the safety of the available routes. Based on the user experience the pedestrian has the option of giving feedback after which the app indicates the status of feedback submission.

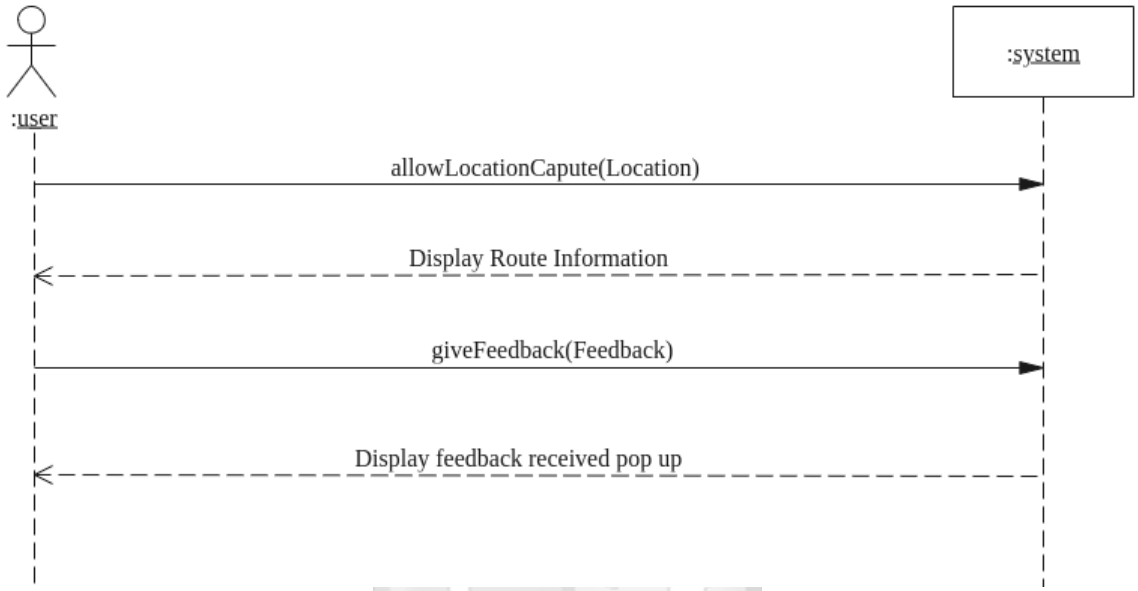


Figure *4.7* Android App System Sequence Diagram.

System Sequence Diagram for web portal application is as shown in Figure 4.8, illustrating the interaction between the user, traffic/planning officer, and the system, web portal. The interactions are initiated by the traffic/planning officer by accessing the web portal where the user is requested to provide login credentials. Once the user provides the login credentials the system responses by showing authentication status where if successful the user is redirected to the reports dashboard. From the reports dashboard the user can request; summary reports where the system responses with graphical representation of the summaries, trends reports where the system responses with graphical representations of the trends, and alerts reports where the system responses with tabular representation of the observation beyond a given threshold. Also, from the reports dashboard the user can provide feedback on user experience and the system responds with feedback received acknowledgement.

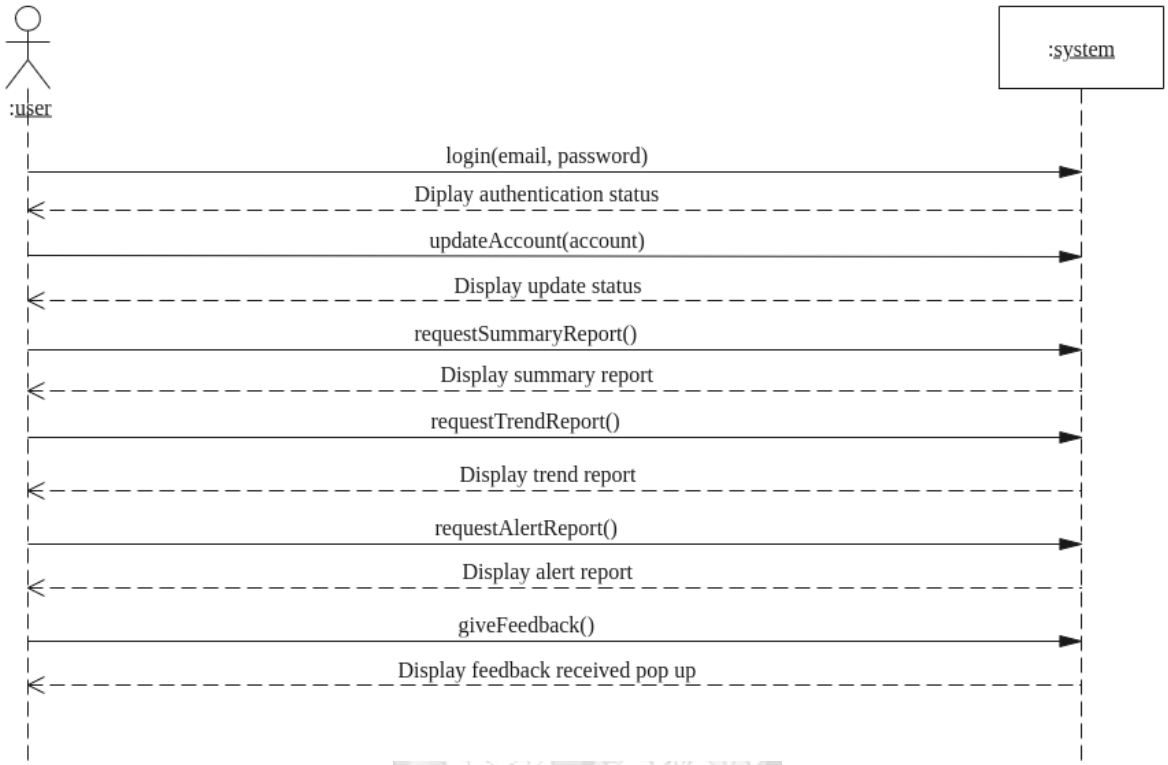


Figure *4.8* Web Portal System Sequence Diagram.

Sequence Diagram is as shown in Figure 4.9, illustrates the exchange and flow of messages among the system objects arranged in time sequence.

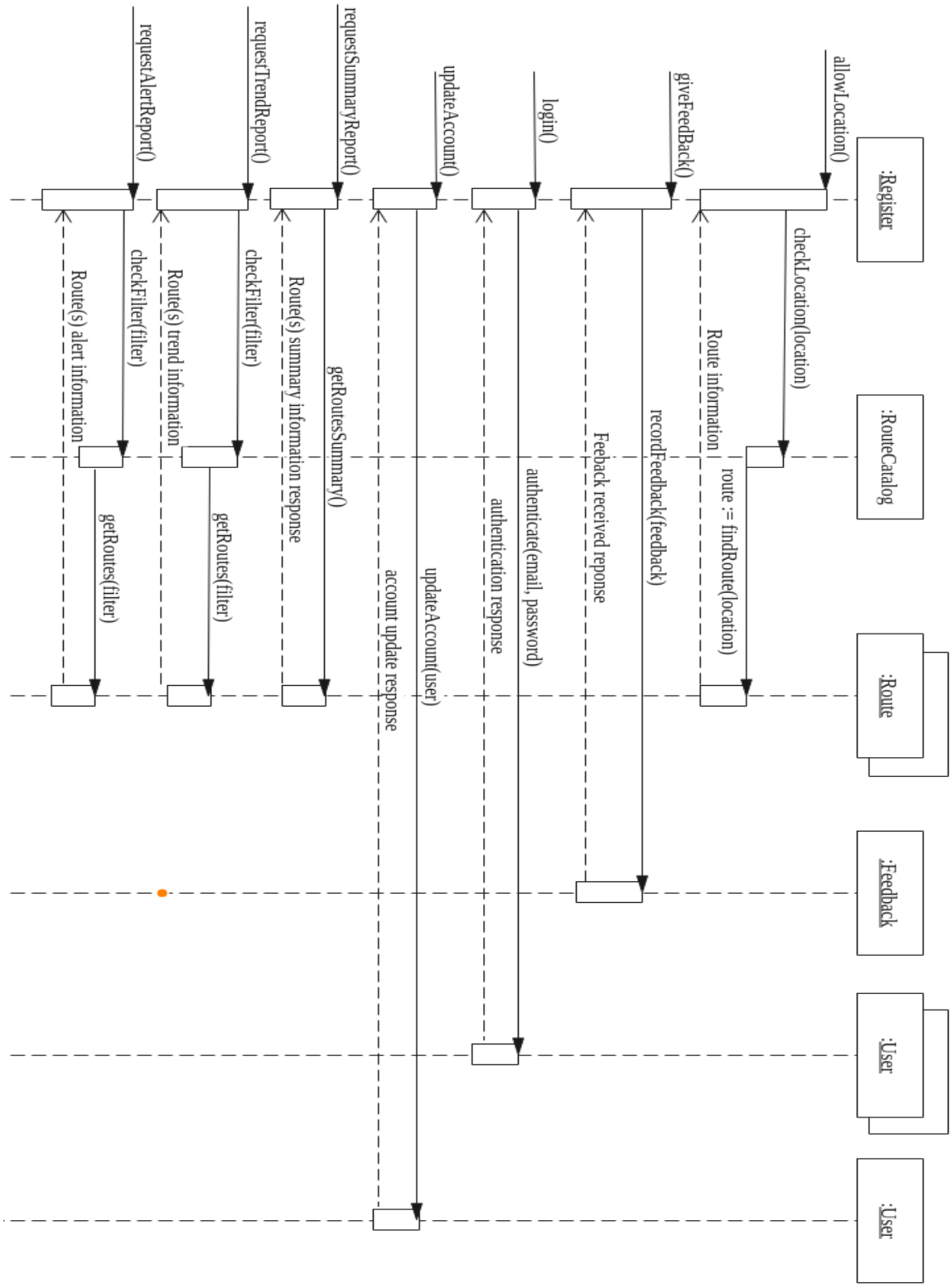


Figure *4.9* Sequence Diagram.

4.4.3 Entity Relationship Diagram

Figure 4.10, illustrates the relationships of entity sets stores in the system database, where entity in this context is an object, a data component. The entities include Route which has Accident(s) associated with, a many to many relationships, and contained in RouteCatalog associated with one-to-many relationship. RouteCatlog has ModelWeight association with many to one relationship, and housed in Register. Register has a one-to-many relationship with RouteCatalog, which also houses Device with a one-to-one relationship. Feedback is given by either the Device or User, with Feedback having a many to one relationship with both Device and User.

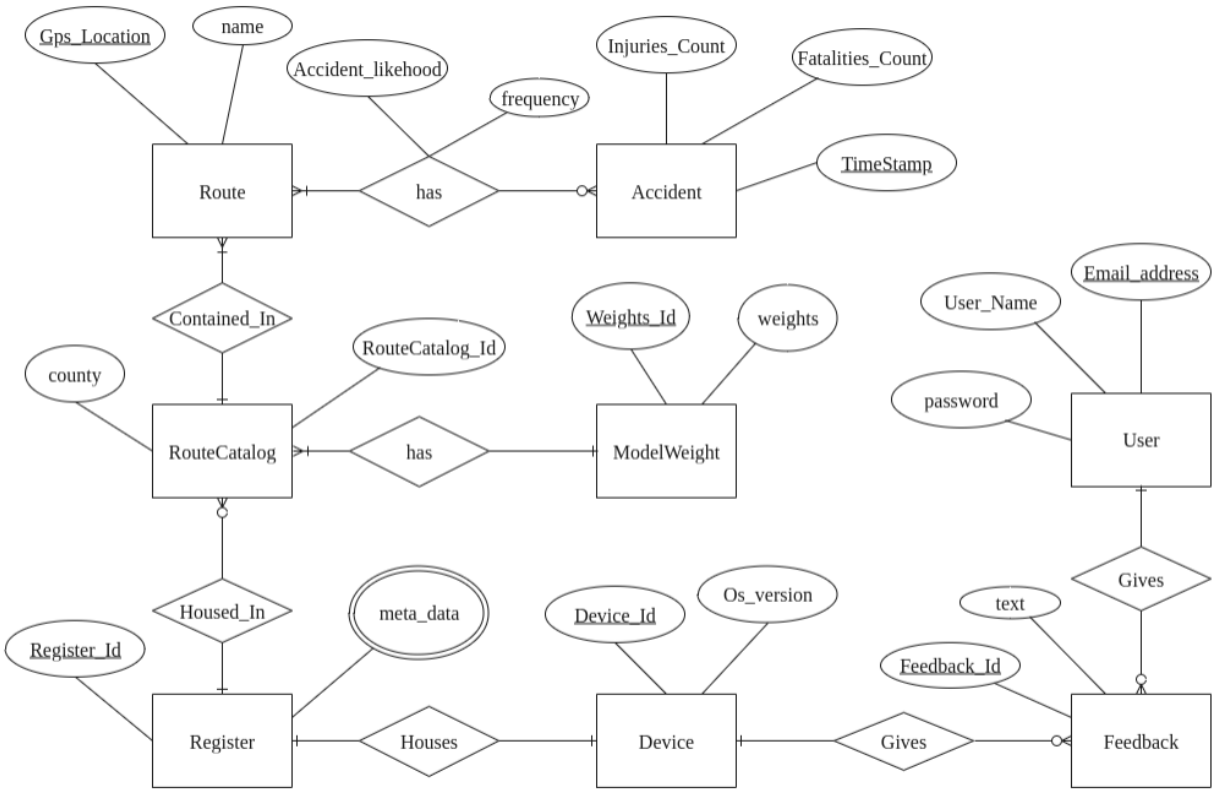


Figure *4.10* Entity Relation Diagram.

4.4.3.1 Database Schema

Table 4.1, illustrates the organization of data as a blueprint of how the system database was constructed. The tables include Route, which stores location and the name of the available routes within the scope. RouteSafety stores route's accident likelihood and accident occurrence frequency count. Accident stores road accident victim, the cause of the accident, age of the victim, injuries count of the victims and fatalities count of the victims. RouteCatalog stores town and county references while ModelWeight stores models weights and accuracy. Register stores user's request state, category and accompanying meta data necessary for restoration in case of failure. Device stores the android device information; device ID and operating system version. Feedback stores

user's feedback in form of text. Last but not the least User table stores user details for users accessing the system via web portal, these details include; user name, email address, encrypted password, access level and active status.

Table 4.1 Database Schema

Route

Id(UK)	route_catalog_id(FK)	gps_location(PK)	name

RouteSafety

Id(UK)	route_id(FK)	accident_id(FK)	accident_likelihood	frequency

Accident

Id(UK)	timestamp(PK)	victim	cause	age	injuries_count	fatalities_count

RouteCatalog

Id(UK)	register_id(FK)	model_weight_id(FK)	town(PK)	county

ModelWeight

Id(UK)	weights	accuracy

Register

Id(UK)	status	category	meta_data

Device

Id(UK)	register_id(FK)	os_version

Feedback

Id(UK)	sender_id(FK)	type	text

User

Id(UK)	user_name(UK)	email_address(PK)	password	access_level	status

4.4.4 Class Diagram

Figure 4.11, illustrates the system's classes, their attributes, operations (or methods) and the relationship among object, where an object in this context is an instance of a class. Register operations includes processing system requests and retrieving register records. Register captures

RouteCatalog whose operation is retrieving route catalog records. RouteCatolog contains ModelWeight and Route. ModelWeight depends on knowing Accident. Route describes RouteSafety which uses and depends on Accident.

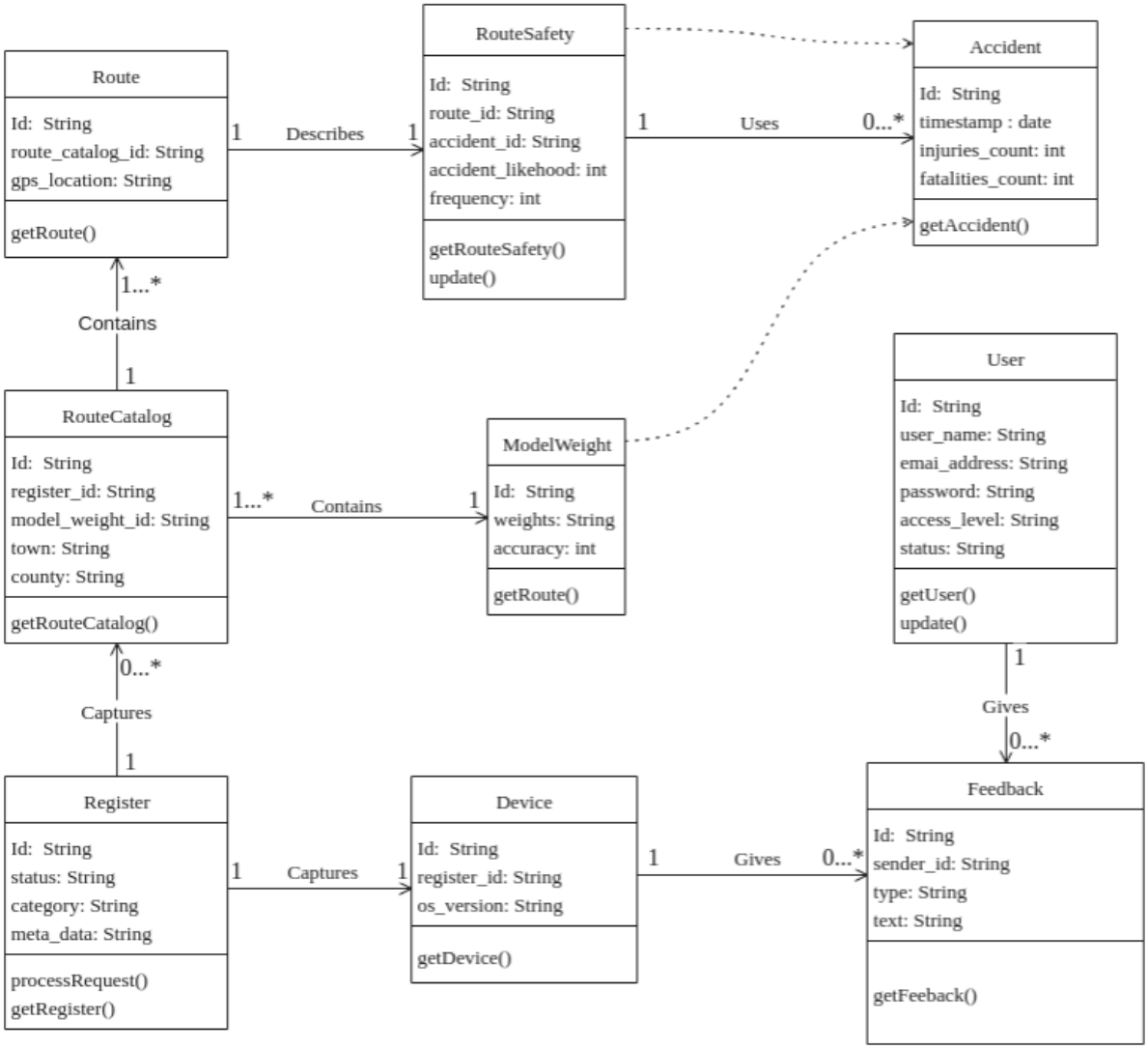


Figure *4.11* Class Diagram.

4.4.5 Wireframes of the system

The wireframe for android mobile app is as shown in Figure 4.12. The wireframe illustrates the flow of displays when a user is interacting with the mobile app. The first step shows the display prompt which requests the user to allow location access, if the user allows the location access the display that follows is the second step which entails the display of available routes with their accompanying safety status. The third step entails the display of feedback entry where the user can give feedback on user experience and if the user gives feedback the final step which is the display of feedback submission prompt status is reached.

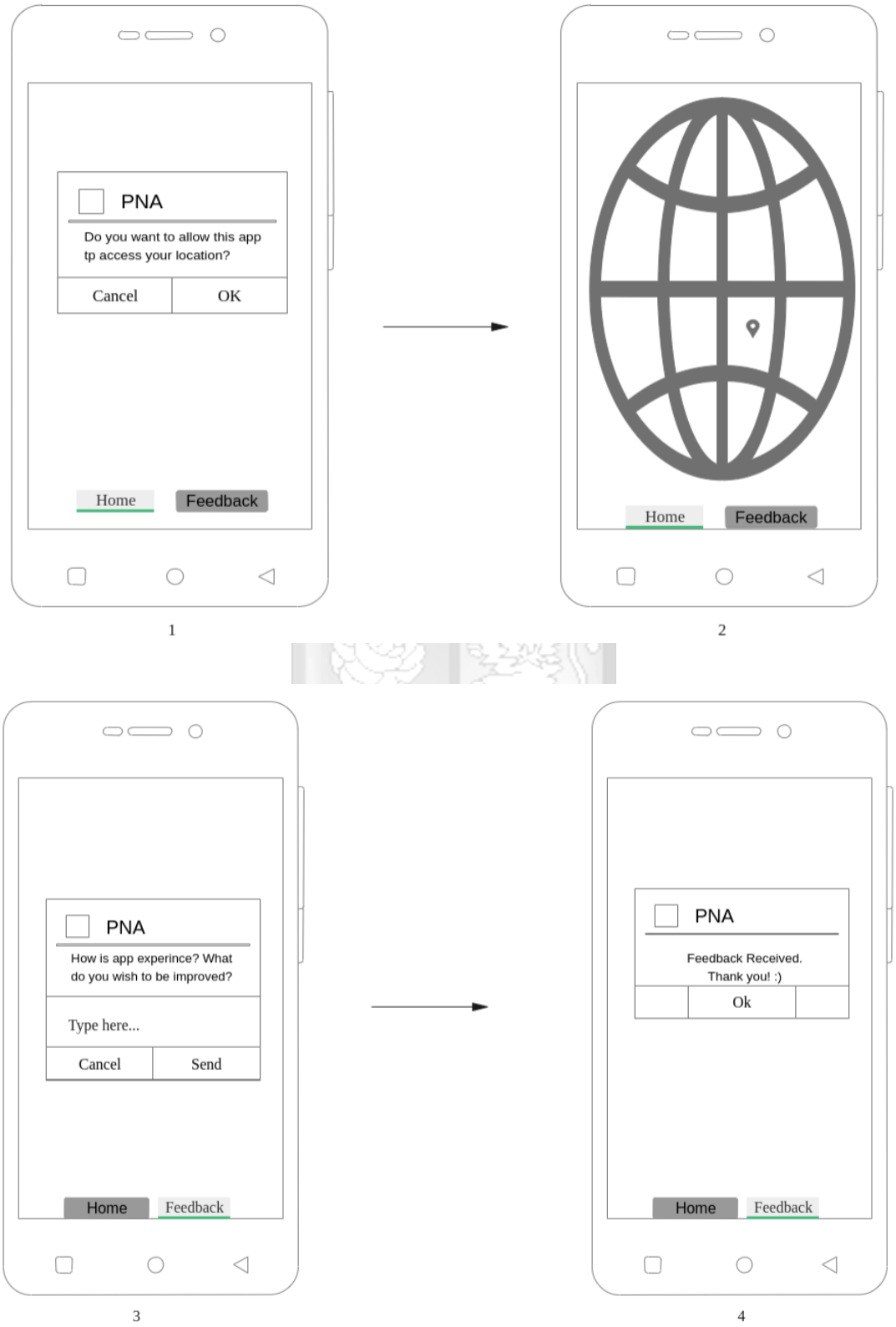


Figure *4.12* Android Mobile App Wireframe.

Chapter 5 : System Implementation and Testing

5.1 Introduction

In Software Engineering, the implementation is regarded as one of many phases of the software development process. After the requirement phase was complete, the next phase of the software development was the implementation of the actual product in a way all requirements identified in the previous phase are best met.

5.2 System Implementation

The entire project development (including project tasks and activities) was solely dependent on the Agile Methodology. This project management technique was carefully and predominantly used for an effective planning of the project work with the use of user stories, increments and different tasks for each use-cases. An agile planning tool called Trello board was chosen for planning and guiding the project in an iterative approach. In the Trello board, the product backlog was created with the series of user-stories and requirements identified in the requirements phase of the development. Each user story or requirement consisted of several tasks where each task had to be fully completed in order to successfully implement the features associated with the certain requirement. The implementation underwent three incremental phases, which are described below;

5.2.1 First Increment

The first increment mainly focused on setting up the resources such as Trello board, version controller and Bayesian linear regression model to get started with the development work. After the creation of the system project board on Trello all tasks performed for this increment were added and moved. In implementing the Bayesian linear regression model following core steps were taken. First, getting data from NTSA online database, the data period chosen was 2018-2020. Secondly, data cleaning which entailed removing the data points with missing fatalities data, converting the date column to timestamp data type, filling the missing time value with the time average and standardizing the road route naming. Thirdly, data visualization which entailed visualizing county against fatalities, weekday against fatalities, time against fatalities, victims against fatalities, weekday and time. Visualization enabled further cleaning of data and feature engineering new columns such as weekday extracted from the date column and hour of the day extracted from the time column.

Fourthly, data formatting which entailed one-hot encoding of categorical variables using the Pandas library encoding function, finding the top six correlated variables to the number of fatalities using the Pandas correlation function and sort function to arrange the results in descending order, and splitting the data into train and test subsets using the model selection module from the sklearn library, as shown in Figure 5.1. Fifthly, defining the model context by specifying

the necessary parameters using Bayesian inference function with prior distribution taken as normal distribution, and then training the model as shown in Figure 5.2. Finally, the model evaluation function was created in order to track the model performance against the chosen metrics, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), as extra fine tuning is done and more data added. The model evaluation function entails comparing the model's estimate predictions against the actual values from the training dataset, the error is calculated as the difference between estimate predictions and the actual values, MAE is then calculated as the absolute mean of the error while RMSE is calculated as square root of the mean of error squared. The model's MAE and RMSE are then added into a comparison dataframe in order to compare against the baseline MAE and RMSE, as shown in Figure 5.3. The baseline MAE and RMSE are calculated using the median values in place of estimates, having a baseline enabled to track model improvements and fine-tuning opportunities and also to avoid under and over fitting the model.

```
# Takes in a dataframe, finds the most correlated variables with the
# No_of_deaths and returns training and testing datasets
def format_data(df):
    # Targets are final No_of_deaths
    labels = df['No_of_deaths']

    # One-Hot Encoding of Categorical Variables
    df = pd.get_dummies(df)

    # Find correlations with the No_of_deaths
    most_correlated = df.corr().abs()['No_of_deaths'].sort_values(ascending=False)

    # Maintain the top 6 most correlation features with No_of_deaths
    most_correlated = most_correlated[:7]

    df = df.loc[:, most_correlated.index]

    # Split into training/testing sets with 25% split
    X_train, X_test, y_train, y_test = train_test_split(df, labels,
                                                        test_size = 0.25,
                                                        random_state=42)


    return X_train, X_test, y_train, y_test
```

Figure 5.1 Data encoding and splitting.

```
# Context for the model
with pm.Model() as normal_model:

    # The prior for the model parameters will be a normal distribution
    family = pm.glm.families.Normal()

    # Creating the model requires a formula and data (and optionally a family)
    pm.GLM.from_formula(formula, data = X_train, family = family)

    # Perform Markov Chain Monte Carlo sampling
    normal_trace = pm.sample(draws=2000, chains = 2, tune = 500, cores=-1)
```

Figure 5.2 Bayesian Linear Regression Model definition and training.

```python
# Evalute the Bayesian LR model trace and compare against baseline
def evaluate_trace(trace, X_train, X_test, y_train, y_test, model_results):

    # Dictionary of all sampled values for each parameter
    var_dict = {}
    for variable in trace.varnames:
        var_dict[variable] = trace[variable]

    # Results into a dataframe
    var_weights = pd.DataFrame(var_dict)

    # Means for all the weights
    var_means = var_weights.mean(axis=0)

    # Create an intercept column
    X_test['Intercept'] = 1

    # Align names of the test observations and means
    names = X_test.columns[1:]
    X_test = X_test.loc[:, names]
    var_means = var_means[names]

    # Calculate estimate for each test observation using the average weights
    results = pd.DataFrame(index = X_test.index, columns = ['estimate'])

    for row in X_test.iterrows():
        results.loc[row[0], 'estimate'] = np.dot(np.array(var_means), np.array(row[1]))

    # Metrics
    actual = np.array(y_test)
    errors = results['estimate'] - actual
    mae = np.mean(abs(errors))
    rmse = np.sqrt(np.mean(errors ** 2))

    print('Model  MAE: {:.4f}\nModel RMSE: {:.4f}'.format(mae, rmse))

    # Add the results to the comparison dataframe
    model_results.loc['Bayesian LR', :] = [mae, rmse]
```

Figure 5.3 Model evaluation function with MAE and RMSE metrics.

5.2.2 Second Increment

The second increment mainly focused on implementing the database, implementing the data access service with RESTful Application Programming Interface (API) and developing the android application. Firstly, setting up the database was the first task in order to store model results against the routes available and feedback given by users. Secondly, the tasks that followed were focused on processing the data stored in the database in order to have the data easily accessible to integrating systems such as the mobile android application and the web portal. Data accessibility to integrating systems was achieved by exposing RESTful APIs and using swagger for APIs documentation. The tasks under exposing data were; expose manage user service endpoints, expose accidents service endpoints, expose route safety service endpoints, expose dump accidents

reports service endpoints and expose routes service endpoints. The endpoints exposed were secured via basic authentication security and documented via swagger API documentation plugin. Finally, the remaining tasks under this iteration were focused on developing the first android app prototype, these tasks included; design the home fragment, design feedback fragment, implement the navigation between the home and feedback fragment, implement map display on the home fragment, implement last location capture via mobile device Global Positioning System (GPS), implement route path display on the map and implement the feedback capture. The final outlook of the app is as shown in Figure 5.4 which indicates the home fragment layout and Figure 5.5 which indicates the feedback layout.

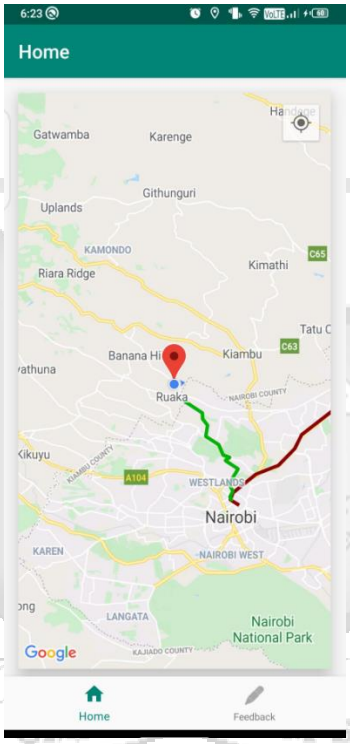


Figure 5.4 Android App Home Fragment Layout.

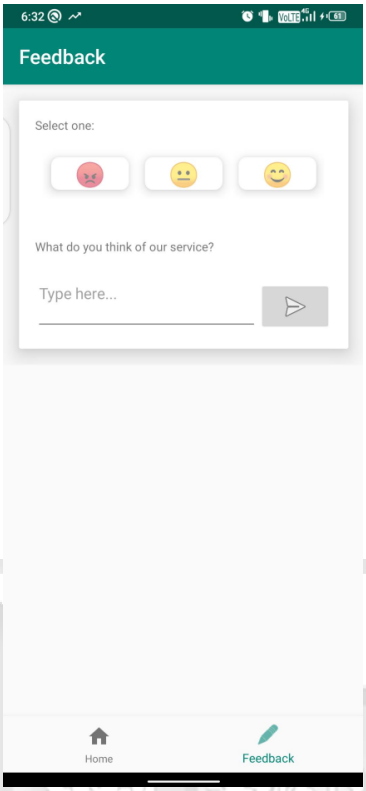


Figure 5.5 Android App Feedback Fragment Layout.

5.2.3 Third Increment

The third increment mainly focused on designing the web portal; login layout, reset password layout, home page layout and reports layout, and connecting/consuming the RESTful APIs. Designing login entailed capturing the user credentials which are the user email address and password, thus the design has two input fields for email and password, and in case the user forgets his/her password the design offers a link to reset password. Designing reset password entailed capturing user email address and offering a link to navigate back to the login page as. Designing the home page entailed displaying graphically the reports of total number of recorded road accident fatalities, the percentage of victims; pedestrians, passengers and drivers, the trends of road accidents fatalities with time and the summary of the victims.

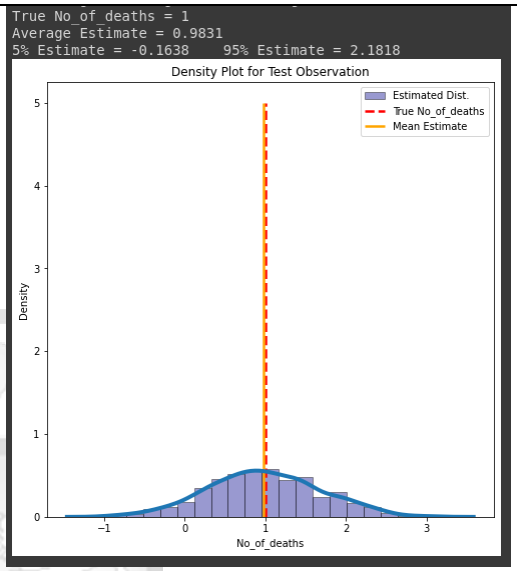
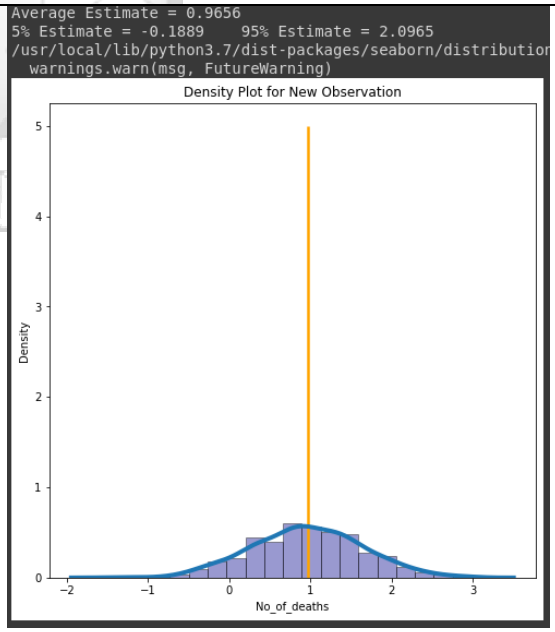
5.3 System Testing

5.2.1 Bayesian Linear Regression Model Testing

Table 5.1 illustrates the test case scenarios; use of test data on the model and use of new observation data on the model that were carried out for the Bayesian Linear Regression model. The parameters supplied to the model, so as to predict the pedestrian road accident fatality likelihood, based on the results after model training were; day of the week, hour of the day, road route name and gender. The model output is a posterior distribution of the number of pedestrian

fatalities likelihood, the values considered from this distribution are the average estimate and the estimated value range which is taken from the estimated value at 5th percentile of the estimates to the estimated value at the 95th percentile of the estimates.

Table 5.1 Bayesian Linear Regression Model Test Scenarios.

Test Scenario	Expected Output	Screenshot
Use of Test data on the model	Output should show: - Average Estimated value - The true value of test data - 5% and 95% estimations - Optional distribution diagram of estimations	 ***Figure 5.6 Sample output of the model on test observation.***
Use of New observation data	Output should show: - Average Estimated value - 5% and 95% estimations - Optional distribution diagram of estimations	 ***Figure 5.7 Sample output of the model on new observation.***

5.2.2 Route Safety Mobile Application Testing

Table 5.2 illustrates the part A of the test scenarios carried out of the route safety mobile application, scenarios include Location access permission request and location access status. Table 5.3 illustrates part B of test scenarios; display of last known location maker and display of the safety status of the available road routes. Road route safety status is displayed as a color code, the mobile phone application captures timestamp and nearest road name based on the last known GPS location. A confirmation is made that the mobile application makes a network call to the route safety service endpoint, which responds with colour code assigned after the service queries the model API endpoint to get the estimates. Based on the analysis performed on the data collected, the gender parameter was set to male as a default since males constitute 85% of the pedestrian road accident fatalities. Finally, Table 5.4 illustrates the test scenario for capturing user feedback. A confirmation is made that the mobile application makes a network call to feedback service API endpoint which responds with a feedback saving status.

Table 5.2 Route Safety Mobile Application Test Scenarios.

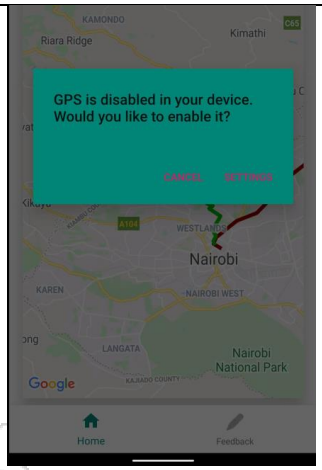
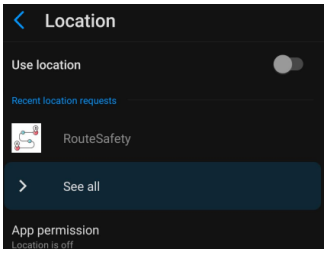
Test Scenario	Expected output	Screenshot
Location Access Permission Request	When the app does not have access to device location permission, the user should be prompted to allow.	 ***Figure 5.8 Location access permission request prompt.***
Location access status	When the mobile location is disabled prompt the user to enable location access: - If the user clicks on the settings, he/she is redirected to GPS enabling settings within the mobile device. - If the user cancels the prompt the whole application closes.	 ***Figure 5.9 Prompt when location access status is disable.***  ***Figure 5.10 Enable location access settings.***

Table 5.3 Route Safety Mobile Application Test Scenarios.

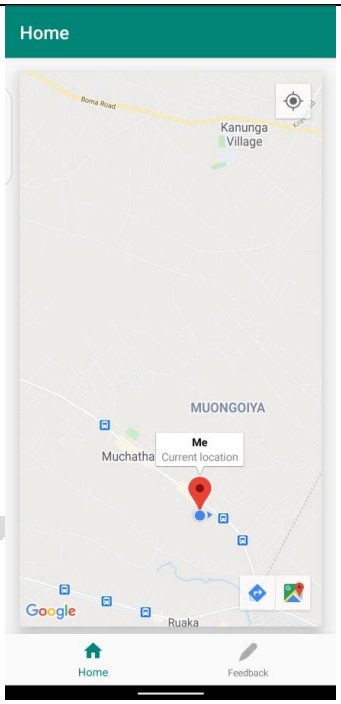
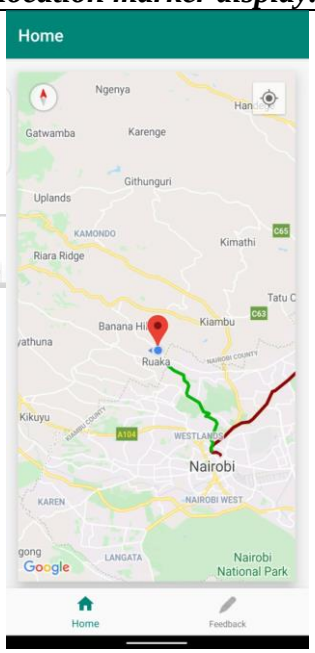
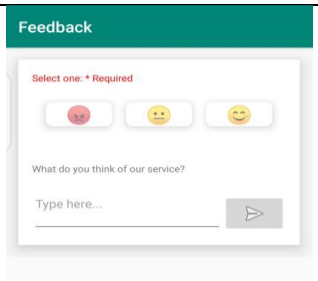
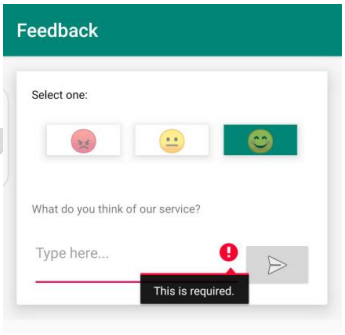
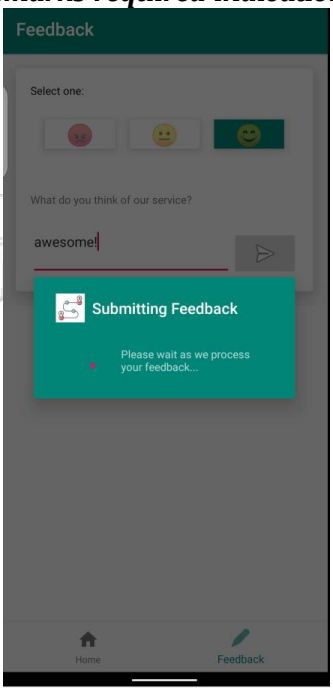
Display of last known location marker	When location access permission is granted and enabled the user should be able to see the last known location marker as per the location accessed	 ***Figure 5.11 Last known location marker display.***
Display of the safety status of available routes	When location access permission is granted and enabled the user should be able to see the safety status of the routes if any available within the covered scope and a distance radius of 3km. Safety status is indicated by route coloring where color represents the likelihood of a fatal accident occurring. Color code is as follows; dark red indicates high, red for high-medium, brown for medium, yellow for low-medium, green for low. The road routes captured under this test were Limuru Road and Thika Road.	 ***Figure 5.12 Route's safety display with coloring.***

Table 5.4 Route Safety Mobile Application Test Scenarios.

Capture user feedback	When the user navigates to feedback layout, he/she should be able to provide feedback by selection one emoji expression and giving some few remarks. - If the user gives the remarks without selecting emoji expressions there is an indication of required emoji expression selection - If user skips giving remarks there is indication that remarks are required. - If all required fields are met, when the user clicks on the send button there is a prompt showing submission progress.	 ***Figure 5.13 Emoji expression required indication.***  ***Figure 5.14 Feedback remarks required indication.***  ***Figure 5.15 Submitting feedback prompt.***

Chapter 6 : Discussion

6.1 Overview

This chapter is devoted to providing a discussion of the findings, when set against the existing literature as discussed in chapter two. The overall aim of this study was to propose a system that provides the pedestrians with valuable information on how safe a road route is in respect to road accidents using Bayesian analysis in predicting the likelihood of road accidents in a given route.

6.2 Identified factors contributing to road accidents in Kenya

This study found that the top four factors contributing to road accidents in Kenyan roads were losing control, failing to keep left and overtaking improperly. These factors were evident from analyzing the accident description column of the dataset used. The findings agree with NTSA (2020) where losing control was the leading cause followed by failing to keep left, overtaking improperly, misjudging clearance, error of judgement, excessive speeding, stepping/walking running off vehicle, crossing road in non-designated area, walking/standing in road, vehicle turning without due care, swerving, brake failure, drunk driving, stationary vehicle dangerous parking, falling from moving vehicle, boarding or alighting from a vehicle, failure of wheel-tires/wheels, slipping or falling and lane changing. These identified factors are classified as human error which agrees with Deme (2019) where he notes that most factors contributing to road traffic accidents in Kenya were mainly caused by human errors at 59.6% percentage.

6.3 Pedestrian's movement patterns identified

The research revealed that pedestrian road accident fatalities are high in highly populated counties, with Nairobi as the top county. Time of the day also correlates with the number of pedestrians fatalities in a road accident, with late evening hours that is 2000hrs to 2100hrs and after midnight hours that is 0100hrs to 0300hrs having the highest count of fatalities, this finding agrees with Hezaveh (2018) whose analysis indicates that 83% of road accidents occurred during the night. The analysis revealed that routes prone to road accidents portrays the same pattern of accidents fatalities with time, indicating that pedestrians have the tendency of using the same route even though the route is prone to road accidents, this finding agrees with NTSA (2020) where the pedestrians are identified as the major victims in fatal road accidents. Finally, it was noted that gender also plays part in the number of pedestrian road accident fatalities with percentage distribution of 85% and 12% for male and female respectively, this is to say a male is seven times more likely to be involved in a fatal road accident than a female, this agrees with Lotfi et al. (2019) analysis that male constitute 79.02% and female 20.98% of the road accidents fatalities.

6.4 Related systems and models identified

The research revealed that majority of the systems in use for road accident analysis, which includes but not limited to predicting the likelihood of a driver causing an accident, forecasting road accident fatality, forecasting property damage and injury forecasting, utilizes Logistic Regression and Ordered Probit models. These models require a lot of data in ordered to achieve a significant score in prediction, this agrees with the same observations made by Schroeder (2016). It was noted that where there is less data available, which are observations of the phenomenon under study, Bayesian Linear Regression can be used in modeling road accident fatality. In the case where there is less the Bayesian Linear Regression performs better than Logistic Regression and Ordered Probit. The finding agrees with Bourdache (2019) whose analysis indicates that where posterior distribution spread out shows fewer data points exists and likelihood clears the prior as the amount of data points increases and also agrees with Ying (2019) conclusion that notes by differentiating the test outcomes of Hit ratio and MAPE with respect to the three severity indicators predictions, the goodness of Bayesian network relevance exceeds the Logistic Regression model.

6.5 Outline of the proposed navigation aid system

As discussed above on the identified systems that are currently in use in different countries, they utilize Logistic Regression and Ordered Probit models which are data intensive. For this study the data collected on road accidents in Kenya was not sufficient to model using these widely used approaches. The approach adopted which works best on less data points is Bayesian Linear Regression, which was developed as per the following steps; data formatting with hot encoding categorical variables, finding top correlated variables, splitting the data into train and test subsets, defining model context and specifying performance metrics. This agrees with development approach used by Ying (2019) in developing the injury forecasting model. In developing the mobile android application, the research shows the approach used in customizing the Google Map APIs, which agrees with Rahmi (2017) approach of integrating Google Maps with mobile apps via android operating system.

The proposed system allows a pedestrian to access the road route safety information via an android mobile application which displays the safety status of the available road routes dependent on the current or the last known location of the pedestrian. The route safety is indicated by coloring where; dark red indicates high, red for high-medium, brown for medium, yellow for low-medium, green for low. The proposed system makes it easier for the pedestrians in Kenya to access the crucial navigation information which is not the case in comparison to other related systems mentioned here which only provide the results to the traffic officials and these related systems cannot be used under Kenyan context due the road accident data limitation.

Chapter 7 : Conclusion and Recommendation

7.1 Conclusions

The main factor contributing to road accidents in Kenya was identified to be driver error. The driver error in this case included; losing control, failing to keep left and overtaking improperly. Losing control can be caused by; rainy weather which makes roads slick, poor road conditions such as potholes or gravel which increases the likelihood of skids, speeding which makes it harder to control a vehicle, and mechanical failures due to poor vehicle maintenance condition. Failing to keep left and overtaking improperly describes the driver behavior in failing to abide by traffic rules.

From the findings, road accidents are seemingly frequent in regions with high population and at part, by urbanization and increased motorization in the region. Nairobi, Nakuru and Kiambu counties with a population of 4.4m, 2.2m, 2.4m respectively as per 2019 Kenya population Census were leading with highest reported fatal accidents. Urbanization in Kenya was at 27.51 percent in 2019 compared to 23.18% in 2009, which is an indication of continuous growth of the share of the urban population. Number of registered vehicles in Kenya have continuously increased, in 2018 the percentage increase was at 72.67% in an interval of a decade.

The research revealed that pedestrian's risk of being involved in a fatal road accident in Kenya increases as population increases with dependency of region urbanization, where a pedestrian in an urban area is considered to have a higher risk of being involved in a fatal road accident compared to a pedestrian in a rural area. Generally, victims in fatal road accidents pedestrians constitute 54.14%, passengers 18.05% and drivers 11.53%. Pedestrians who travel during the late evening hours 2000hrs to 2100hrs are more prone to fatal road accidents given that the route being used has a significant road accident frequency. Pedestrians have the tendency of using the same road route regardless of the road accidents frequency in the given road route. Pedestrian's gender affects the risk of a pedestrian being involved in a fatal road accident, in which a male risk is higher by 73% in comparison to female's risk, meaning a male is seven times more likely to die in a road accident as compared to a female.

The study revealed that Ordered Probit and Logistic Regression models are the major available accident-prediction systems in use. These systems have been used consistently to help many countries across the world to forecast fatalities, injuries and property damage in their traffic systems. Systems for traffic monitoring and management have also been assisted by the fact that many of the stakeholders in the transport sector in these countries have been directly involved in the road safety campaigns, thus willing to support the existing systems. However, these systems don't give attention to pedestrians who are major victims of fatal road accidents. Pedestrians remain not aware of the road route safety in respect to road accidents, this points out the gap the study tries to achieve.

Using Bayesian Linear Regression modelling was well suited since there were few accidents fatalities data points. Using the mobile android application developed a pedestrian can be able to view the road route safety status via color coding where dark red indicates high, red for

high-medium, brown for medium, yellow for low-medium, green for low. For a pedestrian to access the proposed system he/she is required to download the android application from the app store. After app installation the pedestrian will be requested to allow location access, this in order to access the last known location of the pedestrian. After location access the app displays the available road routes with safety coloring, from this information the pedestrian can be able to make an informed navigation decision taking necessary precautions where necessary. The system also allows the pedestrians to give feedback in order to have improved future interactions.

7.2 Limitations

The study faced a data limitation challenge in that there was scarcity of accident-related information from the available data. Retrieved data seemed to be have entry errors with age data having 48.6% missing values, road routes naming did not share a common standard naming approach. The data was victim-oriented even though lacking some victim's information such as victim's intoxication state, victim's disable state and victim's occupation. The challenge posed with kind of data, it is hard to make inferences on prevailing conditions or accident analysis. Details like weather conditions, road intersection and type, road condition, road pedestrian signs, vehicle maintenance state, vehicle category (type), etc. if included could aid in further fine tuning of the model.

7.3 Recommendations

The following recommendations would aid in improving the accuracy of predicting and informing on road route safety in Kenya's roads. Firstly, the age of the road accident victim(s) should be marked as a requirement when filing an accident, this is to aid in data enrichment. Secondly, under influence of substance should be marked as a requirement when filing an accident, this is to aid in data enrichment. Thirdly, the condition of the road should be marked as a requirement when filing an accident, this is to aid in data enrichment. Fourthly, the GPS location should be marked as a requirement when filing an accident, this is to aid in segmenting safety status of a given road route. Finally, for pedestrians without access to smartphones the road route with high likelihood of fatal accidents can have LED billboards with route safety information as pulled from the system proposed in this research.

7.4 Future Work

Below are areas identified for further research; firstly, there is need to research on the digital recording of road accidents reports. This research would aid in proposing approaches to reduce the missing values, ensure standard formatting and central access of data. Secondly, research on pedestrian's occupations in relation to being prone to road accidents would be necessary in order to build more on the project proposed in this research. Finally, research on traffic resources distribution in respect to road accidents would give a deeper understanding of

road accidents trends by trying to show any relation whose outcomes would be used with the outcomes of this research to identify patterns and areas of optimization if any there be.

References

Abd Halim, M., Foozy, C.F.M, Rahmi, l. and Mustapha, A.,. (2018). A review of live survey application: SurveyMonkey and SurveyGizmo. *JOIV: International Journal on Informatics Visualization, 2*(4-2), pp. 309-312.

Arifin, W. N and Zahiruddin, W. M.,. (2017). Sample size calculation in animals studies using resource equation approach. *The Malaysian Journal of medical sciences: MJMS*, 24(5), p. 101.

Ashraf, l., Hur, S., Shafiq, M. and Park, Y.,. (2019). Catastrophic factors involved in road accidents: Underlying causes and descriptive analysis. *PloS one*, 14(10).

Baldwin S. A. and Larson, M.J. (2017). An introduction to using Bayesian linear regression with clinical data. 98, pp. 58-75.

Bentley, F.R., Daskalova, N. aand White, B.,. (2017, May). Comparing the reliability of Amazon Mechanical Turk and SurveyMonkey to traditional market research surveys. *In Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, (pp. 1092-1099).

Bloomfield, J. and Fisher, M.J.,. (2019). Quantitative research design. *Jornal of the Australasian Rehabilitation Nurses Association*, 22(2), p.27.

Bourdache, N., Perny and Spanjaard, O.,. (2019, August). Incremental elicitation of rank-depedent aggregation function based on Bayesian linear regression.

Castillo, E., Grande, Z., Mora, E., Lo, H.K. and Xu, X.,. (2017). Complexity reduction and sensitivity analysis in road probabilistic safety assessment Bayesian network models. *Computer-Aided Civil and Infrastructure Engineering*, 32(7), pp. 546-561.

Castillo, E., Grande, Z., Mora, E., Xu, X. and Lo, H.K.,. (2017). Proactive, backward analysis and learning in road probabilistic Bayesian network models. *Computer Aided Civil and Infrastructure Engineering*, pp. 32(10), pp. 820 - 835.

Deme, D.,. (2019). Review on factors causing road traffic accident in Africa. *Journal of architecture and construction*, 2(3), pp.41-49.

Dhir, S.m Kumar, D. and Singh, V.B.,. (2019). Success and failure factors that impact on project implemetation using agile software development methodology. *In Software Engineering*, pp. (647-654). Springer, Singapore.

Eckhardt, J., Vogelsang, A. and Fernández, D.M.,. (2016). Are" non-functional" requirements really non-functional? an investigation of non-functional requirements in practice. *In Proceedings of the 38th International Conference on Software Engineering*, 832-842.

Etikan, l., Musa, S.A. and Alkassim, R.S.,. (2016). Comparison of convenience sampling and purposive sampling. *American journal of theoretical and applied statistics*, 5(1), pp. 1-4.

Gavalas, D., Kasapakis, V., Konstantopoulos, C., Pantziou, G., & Vathis, N. (2017). Scenic route planning for tourists. *Personal and Ubiquitous Computing*, 137-155.

Grande, Z., Castillo, E., Mora, E. and Lo, H.K.,. (2017). Highway and road probabilistic safety assessment on Bayesian network models. *Computer Aided Civil and Infrastructure Engineering*, pp. 32(5), pp. 379-396.

Gutierrez-Osorio, C., & Pedraza, C. (2020). Modern data sources and techniques for analysis and forecast of road accidents. *Journal of traffic and transportation engineering*.

Habibullah, K. M, A. Alam, S. Saha, A. Amin and A. K. Das,. (2019). A driver -Centric Carpooling Optimal Route-Finding Model Using Heuristic Multi-Objective Search. *A 2019 4th International Conference on Computer and Communication Systems (ICCCS).* Singapore.

Hayes, A.F. and Montoya, A.k.,. (2017). A tutorial on testing, visualizing and probing an interaction involving a multicategorical variable in linear regression analysis. *Communication methods and measures*, 11(1), pp. 1-30.

Hezaveh, A.M. and Cherry, C.R.,. (2018). Walking under the influence of the alcohol: A case study of pedestrian crashes in Tennessee. *Accident Analysis & Prevention*, 121, pp.64-70.

Jamroz, K., M. Budzynski., S. Gaca, M. Kiec, et al.,. (2015). Methodologies for systematic studies of pedestrian behavior and pedestrian-driver relations. *Secretariat of the National Road Safety Council*.

JHA, M.M., Vilardell, R.M.F. and Narayan, J.,. (2016, August). Scaling agile scrum software development: providing agility and quality to platform development by reducing time in market. *In 2016 IEEE 11TH international conference on global software engeneering (ICGSE)*, pp. (84-88). IEEE.

Jol, G. a. (2016). Ethical considerations of secondary data use: What about informed consent? *Dutch Journal of Applied Linguistics, 5*(2), 180-195.

Kabita Sahoo, A. K. (2019, October). Exploratory Data Analysis using Python. *International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8*(12), pp. 2278-3075.

Kalu, F.A. and Bwalya, J.C.,. (2017). What makes qualitative research a good research? An exploratory analysis of critical elements. *International Journal of Social Science Research*, 5(2), pp. 43-56.

Koshi, A., Gooshki, H.S and Mahmoudi, N.,. (2018). The data on the effective qualifications of teacher in medical sciences: An application of combined fuzzy AHP and fuzzy TOPSIS methods. *Data in brief*, 21, pp. 2689 - 2693.

Leavy, P. (2017). *Research design: Quantitative, qualitative, mixed methods, arts-based, and community-based participatory research approaches.* . Guilford Publications.

Lotfi, S., Honarvar, A. R., & Gholamzadeh, S. (2019). Analysis and identification of the hidden relationships between effective factors in the mortality rate caused by road accidents: a case study of Fars Province, Iran. *Chinese Journal of Traumatology*, 233-239.

Manyara, C. (2016). Combating road traffic accidents in Kenya: A challenge for an emerging economy. *In Kenya After 50*, PP. 101-122.

NTSA. (2020, March 01). *National Transport and Safety Authority.* From Online: http://www.ntsa.go.ke/index.php?option=com_content&view=article&id=237

Oyelere, S.S.,Suhonen, J.,Wajiga, G.M. and Sutinen, E.,. (2018). Design, development, and evaluation of a mobile learning application for computing education. *Education and Information Technologies*, 467-495.

Santos, K.O.B., Carvalho, F.M. and Araujo, T.M.D.,. (2016). Internal consistency of the self-reporting questionnaire-20 in occupational groups. *Revista de saude publica*, 50, p.6.

Schroeder, L.D., Sjoquist, D.L. and Stephan, P.E.,. (2016). *Understanding regression analysis: An introductory guide* (Vol. 57). Sage Publications.

Shylesh, S. (2017). A study of software development life cycle process models. *In National Conference on Reinventing Opportunities in Management, IT, and Social Sciences.*, 534-541.

Simonsohn, U., Simmons, J.P. and Nelson, L.D.,. (2019). Specification curve: Descriptive and inferential statistics on all reasonable specifications. *Available at SSRN 2694998.*

Siriaraya, P., Wang, Y., Zhang, Y., Wakamiya, S., Jeszenszky, P., Kawai, Y., & Jatowt, A. (2020). Beyond the Shortest Route: A Survey on Quality-Aware Route Navigation for Pedestrians. *IEEE Access*, 135569-135590.

Taherdoost, H. (2016, April 10). Sampling methods in research methodology; how to choose a sampling technique for research. *How to Choose a Sampling Technique for Research*.

Wang, H., Gu, Y. and Kamijo, S.,. (2017). May. Pedestrian positioning in urban city with the aid of Google maps street view. *In 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)* (pp. 456-459.). IEEE.

World Health Organization. (2018). *Global status report on road safety*. From Online: https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/

Yan, L., Huang, Z., Zhang, Y., Zhang, L., and Ran, B. (2017). Driving risk status prediction using Bayesian networks and logistic regression. *Iet intelligent transport systems*, 431-439.

Ying, G.S., Maguire, M.G., Glynn, R. and Rosner, B.,. (2017). Tutorial on biostatistics: linear regression analysis of continuous correlated eye data. *Ophthalmic epidemiology*, 24(2), pp. 130-140.

Zong, F., Chen, X., Tang, J., Yu, P. and Wu, T. (2019). Analyzing traffic crash severity with combination of information entropy and Bayesian network. *IEEE Access*, 7, pp. 63288 - 63302.

Zou, X., and Yue, W. (2017). A bayesian network approach to causation analysis of road accidents using netica. *Journal of advanced transportation*.

Appendices

Appendix A: Budget

	Description of Work	**Cost Ksh**
Phase One	Proposal and Thesis documents	2,000/-
Phase Two	System Analysis, Design and Architecture	30,000/-
Phase Three	Implementation	20,000/-
Phase Four	Testing	10,000/-
Phase Five	Deployment	15,000/-
Phase Six	Maintenance	10,000/- per month
	Total	**87,000/-**

Appendix B: Ouriginal Report

Curiginal

Document Information

Analyzed document	NAVIGATION GUIDANCE TO PEDESTRIANS ROAD ACCIDENT SAFETY IN KENYA USING BAYESIAN ANALYSIS.pdf (D103346776)
Submitted	4/30/2021 2:48:00 AM
Submitted by	
Submitter email	Paul.kuria@strathmore.edu
Similarity	3%
Analysis address	library.strath@analysis.urkund.com

Appendix C: Research overall design and flow processes