



**University of Nairobi**

**Predict Spread of Infectious Diseases Across Regions Via Transport  
Systems in Kenya Using Machine Learning Models: Case of Truck Crews  
and Bus Passenger Movement**

**Fredrick Onyango**

**P52/35154/2019**

MSc Computational Intelligence

**Supervised by Prof. Peter Waiganjo Wagacha**


**24<sup>th</sup> November 2023**

*A project proposal submitted in partial fulfillment of the requirement for the degree of  
Master of Science in Computational Intelligence in the Department of Computing and  
Informatics of the University of Nairobi*

**DECLARATION**

**STUDENT**

This is my original research project, and it has not been submitted to any other university for an academic award.

Sign...  ..... Date... 28th November 2023 .....

FREDRICK ONYANGO

P52/35154/2019

**SUPERVISOR**

As the designated University supervisor, I have approved the submission of this project.

Sign...  ..... Date... 24<sup>th</sup> November 2023 .....

PROF. PETER WAGACHA WAIGANJO

PROJECT SUPERVISOR

## **ABSTRACT**

The recurrent outbreak of infectious diseases has led to a global health catastrophe with substantial socioeconomic interruptions, morbidity, and deaths. The coronavirus pandemic began in December 2019 in Wuhan, China, and has since transmission throughout the world. It is the most recent outbreak of a viral illness. Understanding and predicting the factors that contribute to COVID-19 spread is critical for providing insights to public health decision-makers in order to inform interventions. The spread of COVID-19 could be propagated by human mobility. People travel for several reasons such as business, education, tours, and work. Buses and matatus are the common modes of transport in Nairobi which contributes to 41.99 % according to KIHBS 2015/2016. The immediate drastic measures taken by most governments is to restrict movements between countries, counties, and towns which has a significant impact on the economy. Developing accurate machine learning models to demonstrate the relationship between the movement of people and the spread of COVID-19 is needed to help come up with measures to reduce the spread and prepare surveillance for future occurrences. Predicting the spread of COVID-19 is currently being done using various machine learning and epidemiological. However, disease modelling based on mobility patterns remains complex due to the siloed data hence the need for innovative methods to enhance prediction accuracy. This research used Logistic Regression and K-Nearest Neighbor machine learning algorithms to forecast the transmission of COVID-19 using modelled data. The training and testing data came from a variety of sources, including truck traffic obtained from NCTTCA reports containing average daily weighted traffic captured at weighbridges. The bus transportation data was modeled using the linear regression trip generation model and parameters from the NUTRANS report (The study on Master Plan for Urban Transport in The Nairobi Metropolitan Area in The Republic of Kenya), with the base year derived from socioeconomic attributes in the 2019 census report. The projections have been made based on the population growth rate derived from the world's development indicators. COVID-19 data was derived from the World Health Organization. Logistic regression demonstrated a testing accuracy score of 0.64 while K-Nearest Neighbor achieved a testing accuracy score of 0.643 using the combined truck and bus daily traffic data. Daily bus data demonstrated an accuracy score of 1 while daily truck data demonstrated an accuracy score of 0.83. We conclude that

using modeled data retrieved from various sources, we can forecast the spread of viral illnesses through human mobility patterns.

**Keywords:** Machine Learning, Prediction, Mobility Patterns, COVID-19

## Table of Contents

<b>DECLARATION</b> .....	<b><i>i</i></b>
<b>ABSTRACT</b> .....	<b><i>ii</i></b>
<b>Table of Contents</b> .....	<b><i>iv</i></b>
<b>DEFINITION OF IMPORTANT TERMS</b> .....	<b><i>ix</i></b>
<b>ACKNOWLEDGEMENTS</b> .....	<b><i>x</i></b>
<b>CHAPTER ONE: INTRODUCTION</b> .....	<b><i>2</i></b>
<b>1.1. Background</b> .....	<b><i>2</i></b>
<b>1.2. Problem statement and goal</b> .....	<b><i>5</i></b>
<b>1.3. Objectives of the study</b> .....	<b><i>5</i></b>
<b>1.3.1. General Objective</b> .....	<b><i>5</i></b>
<b>1.3.2. Specific Objectives</b> .....	<b><i>5</i></b>
<b>1.4. Research Questions</b> .....	<b><i>5</i></b>
<b>1.5. Scope of the study</b> .....	<b><i>6</i></b>
<b>1.6. Research outcomes and their significance to key audiences</b> .....	<b><i>6</i></b>
<b>1.7. Study assumptions and limitations</b> .....	<b><i>6</i></b>
<b>CHAPTER TWO: REVIEW OF LITERATURE</b> .....	<b><i>8</i></b>
<b>2.1. Spread of Infectious diseases</b> .....	<b><i>8</i></b>
<b>2.2. Epidemiological models are used in predicting the transmission of communicable diseases</b> .....	<b><i>8</i></b>
<b>2.3. Machine Learning Models for Transmissible Diseases Prediction</b> .....	<b><i>9</i></b>
<b>2.4. Major transportation corridors that aid in the transmission of viral illnesses</b> .....	<b><i>9</i></b>
<b>2.5. Artificial Intelligence Frameworks used to forecast the spread of COVID-19</b> .....	<b><i>10</i></b>
<b>2.6. Research Gap</b> .....	<b><i>11</i></b>
<b>2.7. Conceptual Architecture</b> .....	<b><i>11</i></b>
<b>CHAPTER THREE: RESEARCH METHODOLOGY</b> .....	<b><i>13</i></b>
<b>3.1 Design of Research</b> .....	<b><i>13</i></b>
<b>3.2 The locality of the project and Beneficiaries of the project</b> .....	<b><i>13</i></b>
<b>3.3 Human transport-based models</b> .....	<b><i>14</i></b>
<b>3.4 Epidemiological techniques utilized in forecasting the spread of contagious diseases</b> .....	<b><i>14</i></b>

3.5	Determine the models based on machine learning that were utilized to make the prediction.....	14
3.6	Determine the predictive performance of machine learning-based models. ....	14
3.7	Data analysis methods .....	15
3.8	Ethical clearance considerations .....	16
<b>CHAPTER FOUR: RESULT AND DISCUSSION.....</b>		<b>17</b>
4.1	<b>DESIGN AND ANALYSIS .....</b>	<b>17</b>
4.1.1	Diseases Caused by Communicable Agents .....	17
4.1.2	COVID-19 Spread per County .....	18
4.1.3	Trip Generation .....	18
4.1.4	Gravity Model for inter-county interaction.....	24
4.1.5	Modal Split .....	25
4.1.6	Buses and Matatus Transport.....	28
4.1.6.1	Primary data for bus and matatus – Questionnaire Feedback .....	28
4.1.6.1	Number of people traveling based on population growth rate.....	29
4.1.7	Trucks Transport.....	30
4.1.7.1	Truck Data Sources .....	31
4.1.7.2	Smoothing on an Exponential Scale .....	32
4.1.7.2.1	Assumptions of Exponential Smoothing .....	33
4.2	<b>Epidemiological and Machine Learning Models.....</b>	<b>34</b>
4.2.1	SVM, SSL and DNN.....	34
4.3	<b>IMPLEMENTATION .....</b>	<b>34</b>
4.3.1	A comparative study of predictive machine learning models to forecast COVID-19 spread....	35
4.3.1.1	Linear Regression .....	35
4.3.1.2	Support Vector Machine (SVM).....	36
4.3.1.3	Random Forest .....	37
4.3.1.4	K-Nearest Neighbors.....	37
4.3.1.5	Stochastic Gradient Descent.....	37
4.4	<b>RESULTS .....</b>	<b>42</b>
4.4.1	Models Results Evaluation Metrics .....	42
4.4.1.1	F1 Score .....	43
4.4.1.2	Mean Absolute Error (MAE).....	43

4.4.1.3 Mean Squared Error (MSE) .....	44
4.4.2 Model Evaluation Results for Combine Truck and Bus Traffic Monthly Data .....	44
4.4.3 Model Evaluation Results for Truck Traffic Daily Data .....	45
4.4.4 Model Evaluation Results for Bus Traffic Daily Data .....	45
4.5 DISCUSSION.....	46
4.7 FUTURE WORK.....	47
<b>REFERENCES .....</b>	<b>48</b>
<b>APPENDICES .....</b>	<b>54</b>
<i>Appendix A - Data Sources Summary .....</i>	<i>54</i>
<i>Appendix B - Kenya expressways, main roads, and streets network Map .....</i>	<i>55</i>
<i>Appendix C - COVID-19 cases in Kenya for the top ten counties with high numbers of infections Interview Questionnaires.....</i>	<i>56</i>
<i>Appendix D - Ethical clearance consideration/permit .....</i>	<i>57</i>
<i>Appendix E - Interview Questionnaires Conducted at Various Bus Station in Nairobi .....</i>	<i>58</i>
<i>Appendix F - Bus Volumes and frequencies (Research Feedback ) .....</i>	<i>62</i>
<i>Appendix G - The total number of COVID-19 cases reported monthly in Kenya .....</i>	<i>64</i>
<i>Appendix H - Truck-based person traffic volumes. ....</i>	<i>66</i>
<i>Appendix I - Truck-based person traffic volumes exponential smoothing. ....</i>	<i>67</i>
<i>Appendix J - Truck Traffic person volume and bus monthly data and total number of COVID-19 Cases reported.....</i>	<i>70</i>

## LIST OF FIGURES

Figure 2. 1 Conceptual Architecture to forecast transmission of infectious diseases.....	12
Figure 3. 1 CRISP-DM .....	15
Figure 4. 1 Modeling the number of people who traveled by bus and matatu. ....	30
Figure 4. 2 The Northern Corridor Road Network .....	31
Figure 4. 3 truck-based person traffic volumes .....	32
Figure 4. 4 Forecasting the truck-person volume traffic graph. ....	33
Figure 4. 5 Compare the accuracy score for KNN using combined truck and bus data.....	45
Figure 4. 6 Compare the accuracy score for KNN using combined truck data .....	45
Figure 4. 7 Compare the accuracy score for KNN using combined bus data. ....	46

## LIST OF TABLES

Table 4. 1 The socioeconomic attributes of traffic analysis zones. ....	20
Table 4. 2 Trip Production for the traffic analysis zones.....	22
Table 4. 3 Trip Attractions for the traffic analysis zones. ....	23
Table 4. 4 Total Number of Trips for the traffic analysis zones in 2019.....	24
Table 4. 5 How people travel from work.....	27
Table 4. 6 Total Number of People who Traveled via matatus and buses in 2019. ....	28
Table 4. 7 People Travelling by bus and matatus based on population growth rate.....	29



## ACRONYMS

<b>AIDS</b>	Acquired immunodeficiency syndrome
<b>ANFIS</b>	Adaptive network-based fuzzy inference system
<b>ARIMA</b>	Autoregressive Integrated Moving Average Model
<b>CNN</b>	Convolutional neural network
<b>CNN-LSTM</b>	Convolutional neural network-long short-term memory
<b>COVID-19</b>	Coronavirus disease
<b>CRISP-DM</b>	Cross Industry Standard Process for Data Mining
<b>DNN</b>	Deep Neural Network
<b>DSPM</b>	Deep Sequential Prediction Mode
<b>EVD</b>	Ebola Virus Disease
<b>JICA</b>	Japan International Cooperation Agency
<b>HIV</b>	Human Immunodeficiency Virus
<b>KIHBS</b>	Kenya Integrated Household Budget Survey
<b>KNBS</b>	Kenya National Bureau of Statistics
<b>KNN</b>	K- Nearest Neighbor
<b>LSTM</b>	Long short-term memory
<b>LR</b>	Linear Regression
<b>NCTTCA</b>	Northern Corridor Transit and Transport Coordination Authority
<b>NN</b>	Neural Network
<b>NRM</b>	Non-parametric regression model
<b>NUTRANS</b>	Nairobi Urban Transport Study
<b>MAE</b>	Mean Absolute Error
<b>MAPE</b>	Mean Absolute Percentage Error

<b>MLP-ICA</b>	Multi-layered perceptron-imperialist competitive algorithm
<b>MSE</b>	Mean Squared Error
<b>RMSE</b>	Root Mean Squared Error
<b>ROC</b>	Receiver Operating Characteristic
<b>SACCO</b>	Savings and Credit Cooperative Organization
<b>SDGs</b>	Sustainable Development Goal
<b>SEIR</b>	Susceptible(S), exposed(E), infected(I) and recovered or removed(R)
<b>SGR</b>	Standard Gauge Railway
<b>SSL</b>	Semi-Supervised Learning
<b>SVM</b>	Support Vector Machine
<b>SVR</b>	Support Vector Regression
<b>TAZ</b>	Traffic Analysis Zones
<b>WHO</b>	World Health Organization

## **DEFINITION OF IMPORTANT TERMS**

**INFECTIOUS:** Highly spread from one organism to another

**EPIDEMIOLOGICAL:** It is a branch of medicine that deals with the spread and control of disease

## **ACKNOWLEDGEMENTS**

I would like to thank my supervisor, Professor Peter Wagacha Waiganjo, for his guidance and unwavering support throughout this process. I would not have embarked on this journey without the guidance and expertise of my supervisor.

I would also like to thank Professor Daniel Orwa, Mr. James Gachanja, and Mr. Ngaruiya Eliud for their contributions during the journey's conceptualization stage. I also want to thank the panel for their patience and guidance.

Finally, I am grateful to my family, particularly my spouse, children, parents, and colleagues, for their support and motivation throughout this journey.

## **CHAPTER ONE: INTRODUCTION**

### **1.1. Background**

The universe has been battling recurring trends of contagious diseases for years. These diseases have caused a worrying trend globally in both developed and underdeveloped countries due to their impact on health and the economy. Previous epidemics and pandemics have indicated that the possibility of these diseases reoccurring or new strains occurring is extremely high due to the interdependence of the organism within the ecosystem and human behavior. The possibility of reducing the spread of infectious diseases through predictions puts hope to many nations as this will enable the implementation of control and preventive measures in advance (Sabin et al., 2020).

On March 23, 2014, The World Health Organization reported Ebola Virus in the rural region of Southeastern Guinea. The transmission of the virus to other countries and cities was attributed to weak surveillance systems, poor health infrastructure, and travel across the border. The first case confirmed in the United States during the 2014-2016 epidemic was travel related. The victim was a man traveling from West Africa to Dallas, Texas. The health workers who nursed him in Dallas also tested positive for the virus (“The Ebola Outbreak,” 2014).

Acute respiratory illness caused by a novel coronavirus is a transmissible illness that has spread globally. COVID-19's emergence is similar to that of the severe acute respiratory syndrome coronavirus in 2002 and the Middle East respiratory syndrome coronavirus in 2012. Independent researchers have linked novel coronavirus to  $\beta$ -coronavirus with the corresponding genetic of coronavirus discovered in bats, suggesting that bats may be the primary host (Guo et al., 2020).

The newly discovered coronavirus spreads through small droplets emitted by infected individuals. The first case was reported in December 2019 in Wuhan, China. Since then, the virus has spread to several continents and countries around the world. Human mobility is thought to be the primary cause of this spread. Air and water transport have made significant contributions to inter-country spread, while bus, truck, and railway transport have made significant contributions to corridor spread.

The first case of the virus in Kenya was reported by the Minister of Health on March 11, 2020, when a student returned from the United States via London. The disease has since spread to all counties in Kenya. According to the World Health Organization, COVID-19 had already spread to 343,955 people in Kenya as of September 7, 2023, with 5,689 deaths. The virus has infected 770,563,467 people worldwide, with 6,957,216 deaths confirmed (WHO Coronavirus (COVID-19) Dashboard, 2023).

The transport network could be the main contributor to the spread of infectious diseases because it interconnects the world. The main modes of transport are the air, rail, water, and road. Infectious diseases change their status from epidemic to pandemic within a short timespan due to the world's interconnection and human mobility (Fang et al., 2020). It is possible for an individual who is a virus carrier to move from one region to another. The major transport system in Kenya is passenger buses and shuttles. It is considered that infectious diseases could easily spread during the journey from one person to another. Some passengers highlight along the way and others board the buses along the way. (Chinazzi et al., 2020) stated in the research they conducted that they wanted to understand the implications of travel bans on the transmission of the COVID-19 outbreak by using a worldwide metapopulation illness model of transmission and Chinese epidemiological data. They stated that the travel restriction implemented in Wuhan only delayed the spread of the virus rather than controlling it.

Clark et al., (2020) indicated that the rising infections and fatalities statistics have forced authorities to research a more accurate model to use in forecasting the spread. The research outcome will help authorities adopt preventive measures to lower disease spread to less than a third of the world's population. This approach target to prevent the possibility of acute illnesses to population with pre-existing health conditions. In every five individuals worldwide, two have pre-existing health conditions making them vulnerable to other infections.

A more accurate model is suggested to help the governments make critical decisions that will help save lives and at the same time save the economy considering that the transport sector in Kenya is a major informal employer. The government of Kenya's main focus is to acquire a more accurate prediction model that will aid in reducing the impact of the virus preventing it from overwhelming the already limited health system in the country as witnessed in other nations with advanced healthcare systems such as China, United States of America, Italy,

Brazil, South Africa and lately India which has been overwhelmed by the Indian variant according to BBC news report ("*India records 300,000 COVID deaths as pandemic rages*", 2021).

WHO have indicated that human health is a fundamental factor for a productive nation. Human health factor has been tested with the outbreak of coronavirus disease triggering global warnings. The virus has so far claimed many lives due to a lack of information to help guide on the probable mitigation measures to help avoid burdening the limited health systems. Even the most developed nations have been struggling to manage this virus due to its unprecedented spread through human movements. Human mobility is fundamental to building a strong economy, people travel for several reasons ranging from business, adventure, academics, and jobs. Most governments are looking for insights on how human mobility contributes to the spread of contagious diseases. These intuitions would guide in building control measures to ensure people travel without spreading diseases.

Acquiring accurate insights would help in decision-making and controlling the spread of infectious diseases in the corridors without interfering with human mobility. Multiple epidemiological and machine learning models have been suggested and used to help predict infectious diseases spread in some counties. The SEIR model was first used in 1969 by a mathematician. The model was also used in United States to predict interventions for influenza pandemic in 2009. The government of China also used SEIR in Wuhan to predict the spread of coronavirus to guide on interventions. Coronavirus trends can be predicted using logistic and machine learning models (Wang et al., 2020).

Brand et al., (2020) also developed a modeling framework to simulate the virus transmission in Kenya focusing on the epidemiological characteristics. The modeling framework known as KenyaCov was used to forecast coronavirus spread within different regions and age groups.

Many developed countries such as China, United States of America, Russia, and India have developed machine learning models to predict spread of infectious diseases. We have used existing machine learning models in determining how human mobility contributes to the transmission of contagious illness in Kenya. This study also focus on mobility via buses and truck. The data collected was used to generate predictive models for emerging infectious disease transmission that guide in developing measures to help control the spread.

## **1.2. Problem statement and goal**

Movement of people across the country is considered important because it is instigated by numerous factors such as work, studies, businesses, and leisure which all contribute to building a stable economy. Contagious diseases spread is associated with human mobility leading to stringent measures put in place by the government of Kenya. There is need to develop an accurate machine learning model to predict how transport systems/mobility patterns contribute contagious diseases spread. The model will provide accurate insights to public health decision-makers to inform interventions that balance between allowing people to move and still control the spread of infectious diseases.

## **1.3. Objectives of the study**

### **1.3.1. General Objective**

The goal of this study was to build a machine learning framework to predict contagious disease transmission through Kenyan transportation systems.

### **1.3.2. Specific Objectives**

1. Create human transmission of infectious disease networks using transportation systems, taking into account origin, destination, and routes.
2. To review epidemiological and machine learning models for predicting the transmission of infectious diseases.
3. To develop predictive models using machine learning and evaluate their accuracy.

## **1.4. Research Questions**

1. What are the human transmission networks of infectious diseases in Kenya?
2. How do transportation systems contribute to the transmission of contagious diseases, taking into account origin, route, and destination?
3. How do epidemiological models forecast contagious illness spread through transportation systems?
4. How are machine learning algorithms utilized to forecast the transmission of contagious illnesses?
5. How accurately do machine learning algorithms forecast the propagation of contagious ailments?

### **1.5. Scope of the study**

The research was carried out in Kenya, with interviews conducted at the bus SACCO offices in Nairobi with bus and shuttle employees and passengers. The study also examined NCTTCA truck transport data. The study focused on the transport and health ministries in both the national and county governments, as they were involved in the development and implementation of surveillance measures to combat the transmission of COVID-19.

### **1.6. Research outcomes and their significance to key audiences**

The study's goal was to create, test, and refine a machine learning-based model to predict how human mobility contributes to the spread of infectious diseases along two major Kenyan transportation corridors. These findings were intended to aid in surveillance and the implementation of control measures in Kenya. This research also helps to achieve the Sustainable Development Goals (SDGs), specifically SDG number three in the Kenyan context, which targets at ensuring a healthy life and advocate for the overall well-being of its citizens. The research contributes to the former government's Big Four health agenda. It also aims to contribute to the current Kenya Kwanza government agenda number eighteen on healthcare, which is aimed at recovering from COVID-19 and preparing for future health emergencies.

### **1.7. Study assumptions and limitations**

The study assumes the following:

1. Infectious diseases spread through human mobility by taking into consideration that if one infected person boards public transport, other passengers may be infected during the journey.
2. Transmission is from the major cities to other regions not considering the transmission could also be from the small towns to major cities.
3. Trucks and buses have similarities in the distance they cover and the number of stopovers. The buses have high occupancy.
4. The NUTRANS report (The Research on a Master Plan for the urban Transportation in The Nairobi Metropolitan Area in the Republic of Kenya) utilized when this study was conducted in the urban area of Nairobi Metropolitan (NMS). The study assumes that the subject counties are only populated by urban residents.



The study has the following limitations and challenges:

1. Infectious diseases spread are propagated through bus, shuttle, and truck transport systems with exclusion to other modes of transport such as private vehicles, planes, and trains plying the same routes with passengers on board.
2. Missing data. Northern Corridor data has been updated until December 2021.
3. Siloed data. Only the NCTTCA has access to weighbridge data.

## **CHAPTER TWO: REVIEW OF LITERATURE**

This chapter contributes to the theoretical examination, conceptual architecture, and related literature.

### **2.1. Spread of Infectious diseases**

Commonly referred to as communicable diseases, they are ailments caused by an array of organisms, including bacteria, fungi, parasites, and viruses. These can reside inside or outside one's body and can lead to illnesses that can be transmitted to others. Some of these include Coronaviruses, which affect both animals and humans (Infectious Diseases - Symptoms and Causes, 2021). They contagious illness are common cold, severe Middle East Respiratory Syndrome, severe acute respiratory syndrome, and the Novel Coronavirus. Ebola is caused by contacting the genus Ebolavirus. The virus was first detected in 1976 near the Ebola River in what is now the Democratic Republic of Congo. Flu is also a contagious viral infection that spreads from one person to another through cough and sneezing. Other common infectious diseases in Kenya are polio which invades the nervous system and cholera which causes severe watery diarrhea leading to dehydration and death if not treated (Infectious Diseases - Symptoms and Causes, 2021).

The spread of infectious diseases is highly attributed to human activities. Changing human behavior, disease surveillance and control measures are the practical solutions to controlling human-to-human transmission (Sabin et al., 2020b).

### **2.2. Epidemiological models are used in predicting the transmission of communicable diseases.**

Researchers have been studying how human mobility affects the spread of COVID-19 and other infectious diseases. (Brand et al., 2020) created the KenyaCov modeling framework for predicting the propagation of the virus across different regions and age groups. The model's goal was to simulate virus transmission in Kenya while focusing on epidemiological attributes.

Odhiambo et al. (2020) proposed virus modeling in Kenya by employing a generalized linear regression approach to determine a straight-line connection between indicators of risk. They also proposed using the Compound Poisson Regression to predict confirmed virus outbreaks through utilization of data from the Ministry of Health's website. They suggested that models

that incorporate non-linearity theory can accurately improve the instantaneous spread of disease.

### **2.3. Machine Learning Models for Transmissible Diseases Prediction.**

The growing availability of electronic health data is encouraging the use of machine learning models in the fight against transmissible illnesses. Machine learning techniques are being considered by experts in epidemiology for predicting the propagation of viral illnesses and prevention strategies (Use of Artificial Intelligence in Infectious Diseases, 2020).

Ming et al. (2016) found that the stochastic model predicted transmissible illness outbreaks with high accuracy. The model was evaluated on surveillance data, providing an improved awareness of the dynamics of viral infection patterns in an array of individuals.

Ayris et al. (2022) proposed using data from input time series to train a deep sequential prediction model (DSPM) and a non-parametric regression model (NRM) to learn distinctive characteristics in order to accurately predict the transmission of COVID-19. They used an openly accessible dataset assembled by John Hopkins University in 2020. NRM had a low MAE and error rate of 0.6%, whereas DSPM had a lower error rate for cases per country.

Xu et al. (2022) created a deep-learning framework to forecast COVID-19 transmission in Brazil, India, and Russia. To forecast the spread of COVID-19, they developed Long short-term memory (LSTM), Convolutional neural network (CNN), and Convolutional neural network-long short-term memory (CNN-LSTM).

Garg et al. (2022b) proposed a model that uses population density and economic conditions to improve its performance. The model demonstrated an average accuracy on MAPE (Mean Absolute Percentage Error) score of between 65% to 85%.

### **2.4. Major transportation corridors that aid in the transmission of viral illnesses.**

Researchers have been attempting to forecast how human mobility affects the transmission of viral illnesses around the world. In their article Hybrid deep learning-based epidemic prediction framework of COVID-19, using South Korea as a case study, (Rahmadani & Lee, 2020) stated that infection is primarily caused by droplets from infected people, and thus human mobility must be considered in the epidemic analysis.

Modeling human movement actions to the enormous scale transmission of viral illnesses demonstrated that human mobility and behavior among individuals play a significant role in viral illness propagation (Meloni et al., 2011). Early travel precautions are regarded as critical in the fight against the spread of transmissible illnesses (Grépin et al., 2021).

In their research on the effects of travel bans on the transmission of COVID-19, Chu et al., (2021) stated that travel barriers would reduce the virus's global spread. They also suggest that other strategies such as isolation, social distancing, and hand washing would aid in averting the transmission of the virus in the community.

Roads are the primary mode of transportation in Kenya. The Kenya map (refer to Appendix B, Kenya expressways, main roads, and streets network map) shows that the entire country is linked by major roads that run from major cities to towns (Kenya Map | Map of Kenya, 2021).

## **2.5. Artificial Intelligence Frameworks used to forecast the spread of COVID-19.**

COVID-19 spread can be forecast through machine learning algorithms (Ogundokun & Awotunde, 2020). As stated by (Sufian et al., 2020) in their article "A Survey on Deep Transfer Learning to Edge Computing for Mitigating the COVID-19 Pandemic," authorities must plan ahead of time to mitigate future pandemics in order to avoid the negative impact on their economy, global health, and education. The analytical approach could be expanded to create a more precise model for particular countries. It is suggested that adding more variables to the model will provide greater depth in analyzing virus spread (Gupta et al., 2020).

Roy et al., (2020) investigated the global effects of pandemics using a machine-learning technique. (Tomar & Gupta, 2020) proposed data-driven machine learning models such as long short-term memory and fitting curves to forecast the number of COVID-19 incidents in India. The model forecasted thirty days ahead, taking into account preventive measures such as lockdown and social isolation. According to (Surya, 2018b), the use of artificial intelligence and Machine Learning techniques have yielded some results in terms of controlling the transmission of viral illnesses and providing new containment insights. He also claims that a large amount of health data is essential for utilizing Machine Learning algorithms to identify and evaluate transmissible diseases.

Pandemic through machine learning and cloud computing, adding other indicators to the regression model such as age distribution, population density, individual and community movements, and level of healthcare available would improve the accuracy of predictions. Tuli et al. (2020). Using the COVID-19 Kaggle data, (Sujantha, Chatterjee, & Hassanien, 2020) proposed a model to help anticipate the propagation of COVID-19 in India through multilayer perceptron, linear regression, and vector autoregression.

Ardabili et al.,(2020) indicated that machine learning models are effective in modeling disease outbreaks due to the high uncertainty and limited data. According to their findings, accurately predicting the virus outbreak using machine learning is critical in providing insights into the transmission of viral illnesses.

Using machine learning approaches, this study aimed to show how to accurately predict how human mobility via truck and bus transport contributes to the transmission of viral illnesses.

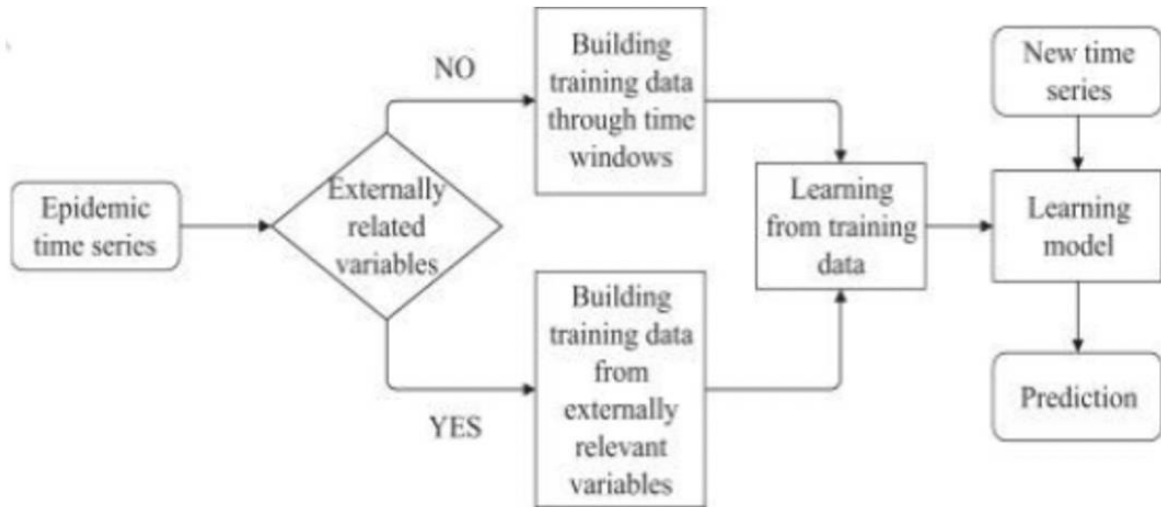
## **2.6. Research Gap**

The need for an accurate machine learning model that can predict how human mobility via commuter buses, shuttles, and truck transport contributes to the spread of infectious diseases in Kenya is growing. The forecast insights will allow the government to implement surveillance and control to avert the transmission of contagious diseases without enacting measures that restrict human mobility, thereby saving the economy while protecting human lives.

## **2.7. Conceptual Architecture**

The diagram below depicts how the machine learning frameworks are utilized to forecast the transmission of viral illnesses. The conceptual architecture is based on research on how government institutions may utilize AI and machine learning to detect the spread of transmissible illnesses (Surya, 2018).

According to the architecture, the model receives data from a variety of sources, including the National Transport Safety Authority, Kenya bus Sacco Associations, COVID-19 cases, and simulation. The model validates the data's relationship. Externally relevant variables provide the training data, while variables that are unrelated provide training data via time windows. The model learns from training data and makes projections through test data (Figure 2.1).



**Figure 2. 1 Conceptual Architecture to forecast transmission of infectious diseases.**

## **CHAPTER THREE: RESEARCH METHODOLOGY**

This chapter describes the methodology that was used in this research. The methods to be used to answer the research questions are presented in this chapter. It also discusses ethical concerns.

### **3.1 Design of Research**

The study employed quantitative methods, which entails the collection of numerical data in order to forecast the spread of infectious diseases through human mobility . A quantitative strategy focuses on variables that can be measured numerically and analyzed statistically. It uses numerical data collection and analysis to describe, explain, predict, or control variables and events of interest (Mills & Gay, 2016). A quantitative study aids in the description of the current situation and the establishment of relationships between variables. This approach focuses on providing a definitive description of the event under investigation (Creswell, 2005).

The approach used a nonexperimental research design. A nonexperimental research design is a technique that has no variable manipulation. The variables are measured naturally as they occur (Mertler, 2019). The focus of nonexperimental research design is descriptive research and correlational research. The correlational research will be used due to the need to investigate the relationship between the spread of infectious diseases and human mobility. The relationship means an individual's status on one variable tends to reflect on another variable (Mertler, 2019). Understanding the relationship between the variables will aid in forecasting the future based on what is currently known. The goal of this study is to observe the spread of COVID-19 through human mobility as it occurs in the world, with no attempts at controlling the variables.

### **3.2 The locality of the project and Beneficiaries of the project**

The data-driven forecasting was applied to the daily transport data. The data was acquired through questionnaires administered to the bus and shuttle employees and passengers. The main beneficiaries of this project if adopted by the government will be the Ministry of Health and the citizens of Kenya. The project will give insights into predicting the spread of infectious diseases and future reoccurrences.

### **3.3 Human transport-based models**

A desk review was used to identify existing human transport-based models. In addition, the epidemiological methods used to predict how human mobility contributes to the spread of infectious diseases will be examined. The country was divided into major transportation hubs that were linked by major highways. The daily movement of people between the subpopulations connected them.

The bus and shuttle transportation data was gathered from the public transportation systems that ply Kenya's major highways on a daily basis. The information was gathered from the bus and shuttle Sacco offices, the National Transport Safety Authority, and major hotels along the highway where people stop to rest. The main indicators will be the number of buses and shuttles that travel the route, the frequency of the buses, and the onboard passenger capacity.

### **3.4 Epidemiological techniques utilized in forecasting the spread of contagious diseases.**

The epidemiological techniques utilized for forecasting the transmission of viral illnesses were reviewed using a desk review.

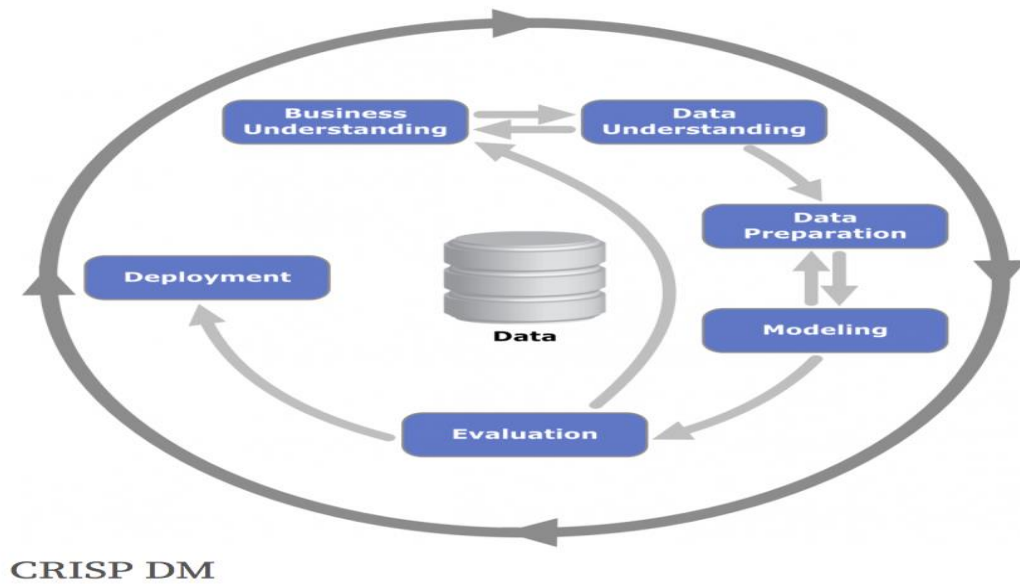
### **3.5 Determine the models based on machine learning that were utilized to make the prediction.**

A desk review was conducted to identify machine learning algorithms that were previously utilized for forecasting COVID-19 spread. The technique of simulation will additionally be applied to generate data.

### **3.6 Determine the predictive performance of machine learning-based models.**

The Cross-Industry Standard Process for Data Mining (CRISP-DM) guided the development, implementation, and evaluation of the machine learning-based model (Hotz, 2023). Business understanding would focus on providing an understanding of the research objectives and constraints, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment (Figure 3.1) are the six steps that guided this study.





**Figure 3. 1 CRISP-DM**

### 3.7 Data analysis methods

The data collected was manipulated, transformed, and visualized to get meaningful insights from the output. Machine Learning algorithm will be used to analyze the data because it constitutes model-building computerization for data analysis. Linear regression model was adopted for data analysis. The linear regression model relies on the assumption of a linear association between the input variables (x) and the solution, which is the outcome that is expected for the provided set of input values (y). Each input value is assigned one scale factor by the equation for linearity, which is symbolized by the uppercase Greek letter Beta(B). Another coefficient, an intercept, is added to provide greater flexibility and bias ("*Linear regression for machine learning,*" 2020).

$$Y = B_0 + B_1 * x$$

The data will be prepared using the below heuristics:

**Linear Assumption** – The linearity of the association between input and outcome is assumed. The data may be transformed to make the relationship linear.

**Remove Noise** – Linear Regression works well with clean data. The need to have the output data cleaned from outliers.

**Rescale Inputs** –The data will also be rescaled using standardization and normalization.

### **3.8 Ethical clearance considerations**

A set of rules, regulations, laws, and ethical standards regulate research involving people's health data. The government establishes legal standards that apply to everyone in a state or country. Ethical standards, on the other hand, do not always have a legal basis and are viewed as voluntary and individual activities based on one's perception of what they consider to be acceptable or unacceptable (*An Overview of the Legal and Ethical Issues in Healthcare, 2020*).

The people who participated received assurance that the research was intended for educational purposes and that complete privacy would be maintained. To avoid tracing the data back to the respondents, the data was coded before being captured on the computer.

The research permission letters were obtained from the Chair of the Department of Computing and Informatics at The University of Nairobi (refer to Appendix E, ethical clearance consideration).

## **CHAPTER FOUR: RESULT AND DISCUSSION**

### **4.1 DESIGN AND ANALYSIS**

#### **4.1.1 Diseases Caused by Communicable Agents**

Communicable illnesses are diseases caused by viruses and microorganisms which propagate effortlessly from person to person through interaction with contaminated body fluids, waste, surfaces, insect bites, or air. HIV (Human Immunodeficiency Virus), measles, salmonella, Hepatitis A, B, and C, tuberculosis, and malaria are among the transmissible illnesses prevalent in Kenya.

In Kenya, the most prevalent causes of death are tuberculosis, HIV/AIDS, and malaria. HIV/AIDS has 37.8 million individuals living with it, newly infected individuals are 1.7 million, and those killed are 770,000 worldwide. In Kenya, there were a total of 1.3 million adults living with HIV, and 139,000 children, with twenty-five thousand deaths in 2018. Malaria continues to contribute to global health problems regardless of the fact that it is treatable and can be prevented. In the same year, children under the age of five accounted for 67 percent of malaria deaths globally. More than half of the inhabitants of Kenya is at risk of malaria, with fourteen million people living in regions where it is endemic and another seventeen million living in seasonal malaria areas. Tuberculosis is also a leading transmissible killer. Globally, ten million individuals were infected with tuberculosis in the year 2018, with a total of 1.5 million dying from the disease. Kenya is ranked fifteenth out of the twenty-two nations with the most Tuberculosis cases. Tuberculosis infections are estimated to be four hundred and twenty-six per hundred thousand individuals nationwide (Communicable Diseases, 2020).

Cholera is an infectious disease common in informal settlements. Kenya has encountered widespread cholera outbreaks since 1997. The Cholera outbreak in Kenya in the year 2015 caused a public health threat. The spread of cholera is linked to open waste disposal, the use of water that is polluted, and impoverished conditions. In Kenya, the average population practices open defecation at 12%, with other areas recording 95%. Thirty-seven percent of Kenyans remain without the ability to obtain clean water to consume, potentially resulting in outbreaks and the transmission of cholera (Pan African Medical Journal, 2017).

COVID-19 is a viral infection that was discovered in Kenya on March 12, 2020. The disease has since spread to all the counties in Kenya. The disease was declared a pandemic on March 11th, 2020, due to its widespread. The world is currently focused on fighting and controlling COVID-19 spread and related deaths. This research focuses on COVID-19 transmissible diseases, how they spread from person to person, and potential control measures (WHO Director-General's Opening Remarks at the COVID-19 Media Briefing - 11 March 2020, 2020).

#### **4.1.2 COVID-19 Spread per County**

Since the first case was reported in Kenya, the number of COVID-19 cases has been increasing. According to health data released by the government of Kenya on March 31, 2022. Nairobi, Mombasa, Kiambu, Nakuru, Uasin Gishu, Machakos, Kisumu, Kajiado, Kilifi, and Busia are the top ten counties with the highest number of COVID-19 cases (refer to Appendix C, COVID-19 cases in Kenya for the top ten counties with a high number of infections).

Nairobi has the largest number of COVID-19 cases, according to reports. The report by Kenya National Bureau of Statistics, data on the most populous counties in Kenya in 2019, Nairobi had the highest population of 4.34 million people, followed by Kiambu with 2.4 million people.

Kiambu County to the west and north, Machakos County to the south, and Kajiado County to the east are also among the top ten counties with the largest number of COVID-19. Because of their close proximity to Nairobi County, Nakuru, Kisumu, Mombasa, and Busia have also been identified as counties with a high number of COVID-19 cases. COVID-19 Cases in Kenya stands at 343,955 with 5,689 deaths (see Appendix G for the total number of COVID-19 cases reported monthly in Kenya and Appendix A for a summary of data sources).

#### **4.1.3 Trip Generation**

The goal of trip generation is to forecast the number of excursions in each traffic analysis zone. The model employs socioeconomic data to calculate the volume of trips generated and drawn to Traffic Analysis Zones (Japan International Cooperation Agency (JICA), 2006, p. 12-5). Based on present growth rates, population data is forecast into the future. According to the model, the overall number of trips generated is symmetrical to the total the number of trips

attracted. According to the basic travel requirements, trips generated at one point are attracted to another point to complete the journey cycle (Bates, 2000). Regression models are used in trip generation.

The following equation represents the trip production model:

$$T_i[k]=f(X^k[C_i^k..]) \quad (4.1)$$

Where (k) represents a population segmentation, (i) represents the origin, ( $X^k$ ) represents a vector of characteristics for segmentation (k), and ( $C_i^k$ ) represents the composite cost of traveling from the origin (Bates, 2000). The model is calibrated using data from the population's distribution by socioeconomic classification (Kenya National Bureau of Statistics, 2020).

Each Traffic Analysis Zone's trip generation was divided into four groups based on purpose: expeditions to work, home, school, and other excursions such as shopping and leisure. The research utilized the linear regression generation of trips model and NUTRANS report parameters. The base year for this study was derived from the 2019 census report's socioeconomic attributes.

According to the National Census, 2019 volumes IV and III, the factors that influence people's movements are the working population, the population seeking work, and the population aged 3 years and above attending school (Table 4.1).

**Table 4. 1 The socioeconomic attributes of traffic analysis zones.**

<b>Counties</b>	<b>Age 5 years and above</b>	<b>Working population (Seeking work + Working)</b>	<b>Other Population (Persons outside the Labor force and note stated)</b>	<b>Population Aged 5 years and above attending school</b>
<b>Nairobi</b>	3,814,871	2,234,599	1,580,272	1,244,416
<b>Kajiado</b>	951,420	532,840	418,580	366,914
<b>Kiambu</b>	2,125,316	1,282,211	843,105	785,800
<b>Kisumu</b>	1,004,086	485,155	518,931	471,248
<b>Machakos</b>	1,267,555	747,139	570,416	450,160
<b>Nakuru</b>	1,870,222	1,004,508	865,714	791,613
<b>Mombasa</b>	1,043,603	565,381	478,222	348,469
<b>Busia</b>	773,860	373,732	400,128	388,496
<b>Kilifi</b>	1,240,674	611,631	629,043	574,829
<b>Uasin Gishu</b>	1,012,918	517,705	495,213	454,517

**Population Aged 5 years and above attending school** = Population aged 3 years and above attending school – (population aged between 3 - 5 attending school) + Population Aged 5 years.

$$\text{Nairobi} = 1,358,304 - 203,802 + 89,914 = \mathbf{1,244,416}$$

$$\text{Kajiado} = 389,090 - 53,326 + 31,150 = \mathbf{366,914}$$

$$\text{Kiambu} = 778,586 - 41,876 + 49,040 = \mathbf{785,800}$$

$$\text{Kisumu} = 504,812 - 61,951 + 28,387 = \mathbf{471,248}$$

$$\text{Machakos} = 505,677 - 58,851 + 3,334 = \mathbf{450,160}$$

$$\text{Nakuru} = 843,280 - 104,542 + 52,875 = \mathbf{791,613}$$

$$\text{Mombasa} = 378,423 - 57,828 + 27,874 = \mathbf{348,469}$$

$$\text{Busia} = 405,072 - 40,625 + 24,049 = \mathbf{388,496}$$

$$\text{Kilifi} = 605,776 - 74,467 + 43,520 = \mathbf{574,829}$$

$$\text{Uasin Gishu} = 480,402 - 53,145 + 27,260 = \mathbf{454,517}$$

### **Data Generation Assumptions**

1. The NUTRANS report was conducted for the Nairobi Metropolitan Area (Nairobi, Kiambu, Machakos, and Kajjido) which is considered urban. The presumption is that the counties with the largest number of COVID-19 cases, which include Nakuru, Mombasa, Kilifi, Kisumu, Uasin Gishu, and Busia, are also urban.
2. The data used to derive the parameter in this study has not been segregated into urban and rural with the assumption that these counties are urban.

#### **4.1.3.1 Trip Productions**

The method below was used to calculate trip production (Table 4.2).

$$\begin{aligned} \text{Home trip produced} &= (\text{population age 5 and above} * 0.1018) + (\text{workers at office base} * 1.0279) \\ &+ (\text{student at enrolment base} * 1.4670) \end{aligned} \quad (5.1)$$

$$\text{Work trip produced} = (\text{population age 5 and above} * 0.5226). \quad (5.2)$$

$$\text{School trip produced} = (\text{population age 5 and above} * 0.2131). \quad (5.3)$$

$$\begin{aligned} \text{Other trip produced} &= (\text{population age 5 and above} * 0.3102) + (\text{Workers at office} \\ &\text{base} * 0.1888) \end{aligned} \quad (5.4)$$

**Table 4. 2 Trip Production for the traffic analysis zones.**

<b>TAZ</b>	<b>Home trip</b>	<b>Work trip</b>	<b>School Trip</b>	<b>Other trip</b>	<b>Total</b>
<b>Nairobi</b>	4510856.45	1993651.58	265185.05	1605265.28	<b>8,374,958.36</b>
<b>Kajiado</b>	1182823.63	191749.26	78189.37	395730.68	<b>1,848,492.94</b>
<b>Kiambu</b>	1867816.59	1110690.14	452904.84	901354.46	<b>4,332,766.03</b>
<b>Kisumu</b>	1292227.59	524735.34	213970.73	403064.74	<b>2,433,998.4</b>
<b>Machakos</b>	1557406	662424.24	270116.97	534255.40	<b>3,024,202.61</b>
<b>Nakuru</b>	2384218.64	977378.02	398544.31	769793.98	<b>4,529,934.95</b>
<b>Mombasa</b>	1198597.94	545386.93	222391.80	430469.58	<b>2,396,846.25</b>
<b>Busia</b>	1032861.70	404419.24	164909.57	310611.97	<b>1,912,802.48</b>
<b>Kilifi</b>	1598270.26	648376.23	264387.63	500333.01	<b>3,011,367.13</b>
<b>Uasin Gishu</b>	1302040.46	529350.95	215852.83	411949.87	<b>2,459,194.11</b>
<b>Total</b>	<b>17,927,119</b>	<b>7,588,161.93</b>	<b>2,546,453.1</b>	<b>6,262,828.97</b>	<b>34,324,563</b>

#### **4.1.3.2 Trip Attractions**

The method below was used to calculate trip attraction (Table 4.3).

$$\text{Home trip attracted} = (\text{population age 5 and above} * 1.0317) \quad (5.5)$$

$$\text{Work trip attracted} = (\text{workers at office base} * 1.0165) \quad (5.6)$$

$$\text{School trip attracted} = (\text{student at enrolment base} * 0.9476) \quad (5.7)$$

$$\text{Others trip attracted} = (\text{population age 5 and above} * 0.2109) + (\text{workers at office base} * 0.3341) \quad (5.8)$$



**Table 4. 3 Trip Attractions for the traffic analysis zones.**

<b>TAZ</b>	<b>Home trip</b>	<b>Work trip</b>	<b>School Trip</b>	<b>Other trip</b>	<b>Total</b>
<b>Nairobi</b>	3935802.41	2271469.88	1497465.75	1551135.82	<b>9,255,873.86</b>
<b>Kajiado</b>	981580.01	541631.86	347687.71	378676.32	<b>2,249,575.9</b>
<b>Kiambu</b>	2192688.52	1303367.48	744624.08	876615.84	<b>5,117,295.92</b>
<b>Kisumu</b>	1035915.53	493160.06	446554.60	373852.02	<b>2,349,482.21</b>
<b>Machakos</b>	1307736.49	759466.79	426571.82	516946.49	<b>3,010,721.59</b>
<b>Nakuru</b>	1929508.04	1021082.38	750132.48	730035.94	<b>4,430,758.84</b>
<b>Mombasa</b>	1076685.22	574709.79	339209.22	408989.67	<b>2,399,593.9</b>
<b>Busia</b>	798391.36	379898.58	368138.81	288070.94	<b>1,834,499.69</b>
<b>Kilifi</b>	1280003.37	621722.91	544707.96	466004.06	<b>2,912,438.3</b>
<b>Uasin Gishu</b>	1045027.50	526247.13	430700.31	386589.65	<b>1,448,039.59</b>
<b>Total</b>	<b>15,583,338.45</b>	<b>8,492,756.86</b>	<b>5,895,792.74</b>	<b>5,976,916.75</b>	<b>35,948,804.8</b>

The base year used in this study was derived from 2019 socio-economic attributes in the 2019 census report (refer to Appendix A, data sources summary). The total number of trips generated for the traffic analysis zones is a combination of trips produced and trips attracted (Table 4.4).

**Table 4. 4 Total Number of Trips for the traffic analysis zones in 2019.**

<b>TAZ</b>	<b>Trips Produced</b>	<b>Trips Attracted</b>	<b>Total Number of Trips</b>
<b>Nairobi</b>	4,510,856.45	1,993,651.58	<b>6,504,508.03</b>
<b>Kajiado</b>	1182823.63	191749.26	<b>1,374,572.89</b>
<b>Kiambu</b>	1867816.59	1110690.14	<b>2,978,506.73</b>
<b>Kisumu</b>	1292227.59	524735.34	<b>1,816,962.93</b>
<b>Machakos</b>	1557406	662424.24	<b>2,219,830.24</b>
<b>Nakuru</b>	2,384,218.64	977378.02	<b>3,361,596.66</b>
<b>Mombasa</b>	1198597.94	545386.93	<b>1,743,984.87</b>
<b>Busia</b>	1032861.70	404419.24	<b>1,437,280.94</b>
<b>Kilifi</b>	1598270.26	648376.23	<b>2,246,646.49</b>
<b>Uasin Gishu</b>	1302040.46	529350.95	<b>1,831,391.41</b>
<b>Total</b>	<b>17,927,119.26</b>	<b>7115861.21</b>	<b>25515281.19</b>

#### **4.1.4 Gravity Model for inter-county interaction**

Tobler’s third law of geography states that all places are related to each other, and they interact, but places that are near to each other interact more than places that are far apart. A methodology for adjusting an aggregate gravity framework on a measure-specific network of roads is introduced in a study on the combined gravity framework for inter-urban spatial relationships at various levels conducted in three separate parts of the western part of the United States. They aimed to create a model that could be utilized on the road network to derive origin-destination information based on the flow of traffic. The model is comprised of two parts, the first component calculates movement along a grid section by incorporating the gravitational inputs of all Origin-Destination pairs while pursuing the shortest route possible. The second component focuses on the socioeconomic aspects of traffic patterns, with a recurring feature for each Origin-Destination pair (Thompson et al., 2019).

The interaction between two places  $i$  and  $j$  in the people and goods gravity model is assumed to be the product of two concepts that stand for each area independently and an interaction term. The gravity model is written as follows:

$$T_{ij} = A(i) B(j)F(d_{ij}). \quad (1)$$

Where  $T_{ij}$  is the number of interactions between areas  $i$  and  $j$ .  $A$  and  $B$  are initial and goal functions, respectively, while  $F$  is the reducing function of its arguments and  $d_{ij}$  is the distance between locations  $i$  and  $j$ . According to the law, regions that are far apart interact less than regions that are close to each other.

The gravity model has been used to select the counties that interact more with Nairobi. Figure 4.1 shows counties with more than 5,000 cases of COVID-19 Nairobi, Kiambu, Machakos, Kajiado, Mombasa, Nakuru, Uasin Gishu, Busia, and Kilifi. Nairobi is Kenya's capital city. The international airport connects it to every country on the planet. The first confirmed manifestation of COVID-19 in Kenya was reported on March 12, 2020, in Nairobi. The passenger arrived in Kenya via the Jomo Kenyatta International Airport, according to the Ministry of Health. The COVID-19 cases data from the Ministry of Health indicate that Nairobi has the highest cases of COVID-19 due to several factors such as the population, proximity, and socio-economic factors.

According to data from the Kenya National Bureau of Statistics' 2019 Kenya Population and Housing Census, Nairobi had the largest number of people with 4,397,077, which was followed by Kiambu with 2,417,735 and Nakuru with 2,162,202. Nairobi has a high population density. It covers 703.9 square kilometers of land and has a population density of 6,247 people per square kilometer (Trizer, 2019). Nairobi is also home to Kibera, Africa's largest slum. Kibera is home to about 250,000 residents who are spread out over a 2.5-kilometer space (Kibera: A Look Inside Africa's Largest Slum, 2020). Seventy percent of Kenya's urban residents lives in Mombasa, Kiambu, Kisumu, Nairobi, Machakos, Uasin Gishu, Nakuku, and Kajiado (Trizer, 2020). In this study, Nairobi is considered the origin while other counties such as Kiambu, Kisumu, Machakos, Kajiado, Busia, Nakuru, and Kilifi are the destination.

#### **4.1.5 Modal Split**

COVID-19 is extremely contagious and spreads from person to person due to proximity. Because people can easily interact while traveling, transportation is regarded as a medium for the spread of diseases. Kenya's transportation network includes road, air, maritime, rail, and waterways within the country. The road network is the primary means of access for urban

areas, accounting for 93% of all freight and passenger traffic in the entire country. The old meter gauge and standard gauge rail have limited use compared to the road network due to the limited coverage and connectivity (Kenya National Highways Authority, 2023).

The survey conducted by KIHBS 2015/16 on how people travel to work shows that walking takes the larger share in all the counties except in Nairobi County where matatus and buses take 41.99 percent compared to walking at 36.86 percent. Walking is considered a healthy mode of transport because it has less exposure to the spread of infectious diseases. Private vehicles and employer-provided transport are considered not to expose people to infectious diseases because most people who travel using private vehicles and employer-provided transport take much time together at work and home. Transport using matatus and buses is considered to have the highest exposure because of congestion, proximity, and the possibility of all passengers being strangers to each other (Table 4.5).

**Table 4. 5 How people travel from work**

<b>Counties</b>	<b>Walk</b>	<b>Bicycle &amp; Motor bike</b>	<b>Tuk-Tuk</b>	<b>Commuter Train</b>	<b>Matatus &amp; Buses</b>	<b>Employer-provided</b>	<b>Private vehicles</b>	<b>Other</b>	<b>Not Applicable</b>
<b>Nairobi</b>	36.86	2.53	0.00	0.09	41.99	2.74	4.35	0.00	11.44
<b>Kajiado</b>	52.01	3.76	0.20	0.09	23.46	0.04	3.70	0.73	15.99
<b>Kiambu</b>	56.26	2.37	0.29	0.00	24.75	3.70	3.63	0.00	9.01
<b>Kisumu</b>	37.54	12.49	5.20	0.00	11.52	0.23	1.00	0.00	32.01
<b>Machakos</b>	51.24	7.16	0.00	0.00	4.25	2.09	1.92	0.22	33.13
<b>Nakuru</b>	58.27	4.94	0.27	0.00	8.85	5.48	0.95	0.00	21.23
<b>Mombasa</b>	47.33	6.09	2.47	0.25	30.36	1.79	1.76	0.27	9.69
<b>Busia</b>	71.61	9.46	0.00	0.00	0.49	0.46	0.45	0.84	16.69
<b>Kilifi</b>	49.40	8.3	0.19	0.00	9.13	0.38	0.33	0.40	31.86
<b>Uasin Gishu</b>	62.18	7.04	0.00	0.00	16.69	2.15	2.05	0.25	9.64

The number of people who traveled using matatus and buses in 2019 has been projected based on the total number of trips generated in the ten counties with the highest number of COVID-19 Cases (Table 4.6).

**Table 4. 6 Total Number of People who Traveled via matatus and buses in 2019.**

<b>Counties</b>	<b>Matatus &amp; Buses</b>	<b>Total Number of Trips</b>	<b>Total Number of People Who Used Matatus and Buses in 2019</b>
<b>Nairobi</b>	41.99	6,504,508.03	2,731,242.92
<b>Kajiado</b>	23.46	1,374,572.89	322,474.80
<b>Kiambu</b>	24.75	2,978,506.73	737,180.42
<b>Kisumu</b>	11.52	1,816,962.93	209,314.13
<b>Machakos</b>	4.25	2,219,830.24	94,342.79
<b>Nakuru</b>	8.85	3,361,596.66	297,501.30
<b>Mombasa</b>	30.36	1,743,984.87	529,473.81
<b>Busia</b>	0.49	1,437,280.94	7,042.68
<b>Kilifi</b>	9.13	2,246,646.49	205,118.82
<b>Uasin Gishu</b>	16.69	1,831,391.41	305,659.19
<b>Total</b>		<b>25515281.19</b>	<b>5,439,350.86</b>

#### **4.1.6 Buses and Matatus Transport**

Secondary and primary data were used to obtain bus travel information. The Kenya Economic Survey secondary data on departing visitors by country of residence and purpose was used to estimate traffic volume in Kenya. Primary data was gathered from Nairobi bus transportation providers to capture trip routes, destinations, frequency of trips, and passenger volumes (refer to Appendix E, interviews questionnaires conducted at various Nairobi bus stations).

##### **4.1.6.1 Primary data for bus and matatus – Questionnaire Feedback**

A survey conducted on July 28<sup>th</sup>, 2021, indicated that government measures to reduce the public transport capacity led to all buses and matatus carrying 60 percent of the total capacity. The government later allowed public transport to carry full capacity (refer to Appendix F, bus volumes and frequencies - research feedback).

#### 4.1.6.1 Number of people traveling based on population growth rate.

Kenya's population was 47,564,296 according to census results from the Kenya National Bureau of Statistics (KNBS). The World Bank data on the annual percentage of population growth indicate that this population represented a 1.978 percent growth from 2019 (World Bank Open Data, Population growth annual percentage - Kenya). The population of Kenya grew by 2.01 percent in 2020 and 1.943 percent in 2021 (Table 4.7).

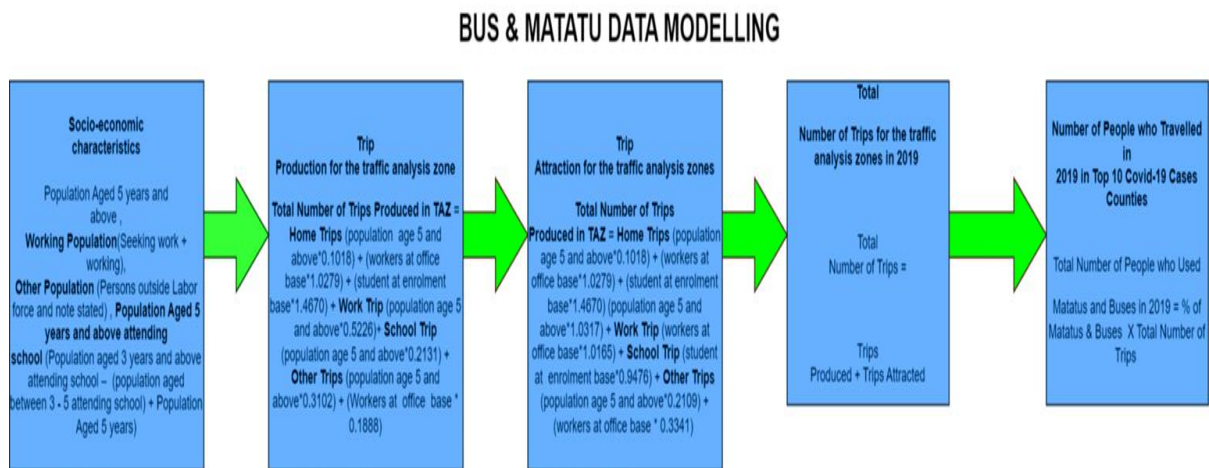
**Table 4. 7 People Travelling by bus and matatus based on population growth rate.**

<b>Year</b>	<b>Growth Rate</b>	<b>Urban Population</b>	<b>Total Population</b>	<b>People traveling in Traffic Analysis Zones using bus and matatu (Yearly)</b>	<b>100% Occupancy People traveling in Traffic Analysis Zones based on population growth rate (Monthly Average Data)</b>	<b>Sixty percent Passenger Occupancy Rate (Per Month)</b>
2021	1.943%		<b>49,463,088</b>	<b>5,656,492.70</b>	<b>471,374.39</b>	<b>282,824.63</b>
2020	2.01%	14,975,059	<b>48,520,338</b>	<b>5,548,681.81</b>	<b>462,390.15</b>	<b>277,434.09</b>
2019	1.978%	14,362,838	<b>47,564,296</b>	<b>5,439,350.86</b>	<b>453,279.24</b>	<b>271,967.54</b>
2018	2.033%	13,771,805	<b>46,623,474</b>	<b>5,331,760.50</b>	<b>444,313.38</b>	<b>266,588.03</b>
2017	2.176%	13,201,347	<b>45,675,619</b>	<b>5,223,365.81</b>	<b>435,280.48</b>	<b>261,168.29</b>

The government of Kenya in July 2020 reduced the capacity of the public vehicles from full capacity to 60%. This directive was given in the protocol for public road operations adopted

in preparation lifting of COVID-19 restrictions. The information from the interviews conducted with the bus companies also indicated the COVID-19 restrictions limited them to 60% capacity. The ban on night travel was also raised by the bus employee respondents as a factor contributing to the reduced number of passengers.

Modeling the number of people who traveled by bus and matatu involved trip production and trip attraction for the traffic analysis zones based on National Census, 2019 volumes IV and III (Figure 4.1).



**Figure 4. 1 Modeling the number of people who traveled by bus and matatu.**

#### 4.1.7 Trucks Transport

Trucks are the primary mode of conveyance for goods and services. During contagious disease pandemics, the trucks transport crucial supplies, exposing the truck crew to infections and the possibility of spreading the disease over a large geographical area. According to the study, SARS-CoV-2 immunoglobulin G seroprevalence is 42.3% much greater within truck drivers and their assistance in Kenya than among medical professionals and individuals who donate blood. Truck drivers are thought to interact with a wide range of social and professional individuals, putting themselves in danger of COVID-19 infection. Because of their role in maintaining the supply chain, they may also contribute to the spread of infectious diseases (Kagucia et al., 2021).



The origin and destinations form the Traffic Analysis Zones (TAZ). Trucks are considered for the long-distance spread of infectious diseases. The trucks use the major highways to access the border points before getting into the neighboring countries.

The map below depicts the Northern Corridor Road system, which is utilized for shipping cargo from Mombasa, Kenya to various East African countries via the Malaba - Busia borders (Northern Corridor Transit and Transport Coordination Authority: Northern Corridor Maps, 2021).

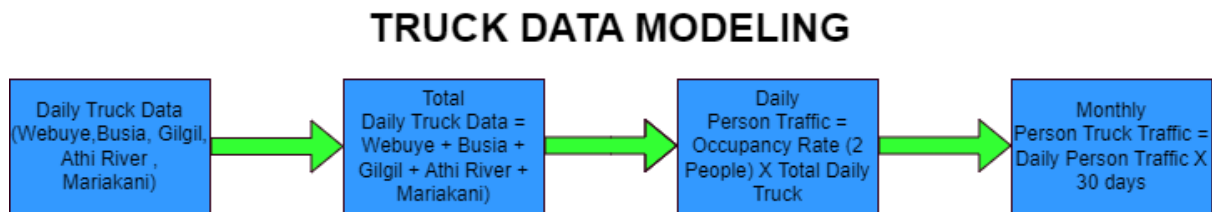


**Figure 4. 2 The Northern Corridor Road Network**

#### **4.1.7.1 Truck Data Sources**

The truck-based data on traffic came from NCTTCA reports, which contain information on average daily weighted traffic capture at weighbridges. There are nine fixed weighbridges in Athi-River, Mariakani, Webuye, Gilgil, Busia, Mtwapa, and Bondo, the first five of which are located along the Northern Corridor. To reduce congestion, the Kenya National Highway Authority has installed fully automated weighbridges. The analysis used data for Webuye, Gilgil, Athi River, Mariakani and Busia weigh bridges to capture Uganda-bound traffic

volumes. The person traffic was calculated using an occupancy rate of two people per truck (Figure 4.3) (refer to Appendix H, truck-based person traffic volumes). Data collection for Tanzania-bound traffic volumes is being collected. Trip frequency was also obtained from NCTTCA reports. Each month, there were three roundtrips to Uganda. The distance that exists between regions contributes to the volume of return trips. The start of Standard Gauge Railway (SGR) operations is expected to reduce the proportion of truck traffic. statistical data, indicate that total SGR production rates have been increasing. For the period January 2018 to February 2019, the total twenty-foot equivalent units shifted by SGR was 323,158. Truck freight service providers, on the other hand, have resorted to last-mile connections.



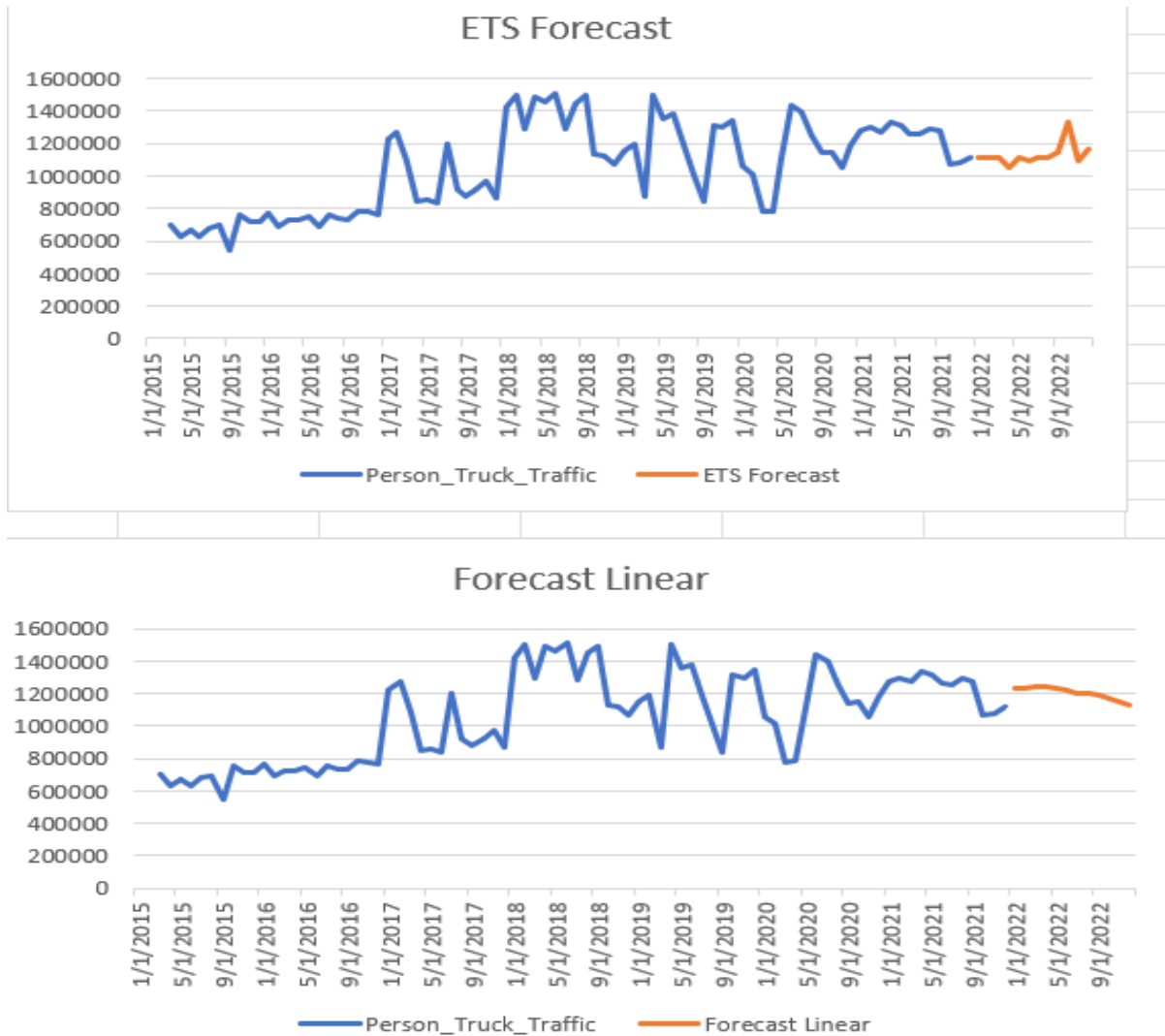
**Figure 4. 3 truck-based person traffic volumes**

#### 4.1.7.2 Smoothing on an Exponential Scale

Exponential smoothing was proposed in the late 1950s and has been utilized in numerous fruitful predictions. Predictions generated through the exponential smoothing method are weighted averages of old historical data, with the weights decaying rapidly as the data points get older. For a wide range of time series, exponential smoothing produces consistent forecasts.

The trend-corrected exponential smoothing (Holt's) method was utilized in the evaluation to forecast truck traffic volumes. As discussed in Arroyo et al (2007), forecasts based on this method are reliable.

The predicted daily truck traffic data increased in linear form based on the forecast (Figure 4.4) (refer to Appendix I, truck-based person traffic volumes exponential smoothing).



**Figure 4. 4 Forecasting the truck-person volume traffic graph.**

**4.1.7.2.1 Assumptions of Exponential Smoothing**

1. Exponential smoothing has been used to generate the missing truck data with the assumption that the truck traffic would progress in a forecast manner.
2. The assumption is that the worst-case scenario of curfew, closing of hot spot areas, and vaccination will be factored in by the lower bound.

## **4.2 Epidemiological and Machine Learning Models**

### **4.2.1 SVM, SSL and DNN**

To forecast the future, predictive models are used. Most events and world phenomena repeat themselves, making it possible to foresee their occurrence. Machine learning and epidemiological models may forecast disease outbreaks and spread. To validate whether new data could be used for forecasting disease outbreaks, researchers conducted study on transmissible illness outbreak forecasting using news articles and models powered by machine learning. The Support Vector Machine (SVM), Semi-Supervised Learning (SSL), and Deep Neural Network, or DNN, models were utilized in this research. The frameworks were utilized for analyzing knowledge contained in content on websites and detect emerging transmission patterns.

Three models were compared to see which one was the most accurate. The study's findings revealed that SSL performed the best, with an average precision of 0.838 and 0.834. SSL also performed best, with average ROC (Receiver Operating Characteristic) values of 0.791 and 0.805 and average F1 scores of 0.832 and 0.802. SVM and DNN also performed well, with an average accuracy of 07 and an F1 score of 0.75. According to the study, enhancing the training data to at least a year's worth of data would enhance the accuracy of predictions. The study also left room for future research into how, rather than using feature selection, useful data accumulation can help improve prediction (Kim, 2021).

## **4.3 IMPLEMENTATION**

The main means of transport that connects Nairobi, which is considered as the Origin to other counties along the Northern Corridor are Trucks, Matatus, and Buses. The counties that border Kenya and other two East African countries Uganda and Tanzania are Busia (Busia Boarder), Malaba, Kajiado (Namanga Boarder), Isibania Boarder, and Lungalunga Boarder. Kisumu and Nakuru counties also have high COVID-19 cases because they are the stopping points for those people using the Northern Corridor.

The monthly data was compiled from COVID-19 cases revealed every day. The data spans from the 12th of March in 2020, when the very first case in Kenya was reported, to the 7th of September 2023. The NCTTCA reports, which include data on daily average weighed traffic

capture at weighbridges, were also used to generate the monthly truck data. The truck data was also combined to generate the monthly data (see Appendix J for truck traffic person volume and bus data, as well as the total number of COVID-19 cases reported).

### 4.3.1 A comparative study of predictive machine learning models to forecast COVID-19 spread.

#### 4.3.1.1 Linear Regression

The linear regression technique is a supervised machine learning algorithm. It simulates a goal prediction value that is based on independent variables. The model discovers associations among variables, which are then used to make forecasts for the future. The variable that is independent  $x$  (input) is used to predict the outcome variable  $y$  (output). The model employs the hypothesis function shown below.

$$y = \theta_1 + \theta_2 \cdot x \quad \text{equation 1}$$

Where  $x$  represents the input training data,  $Y$  represents the data labels (supervised learning),  $\theta_1$  represents the intercept, and  $\theta_2$  represents the  $x$  coefficient.

Finding the line with the greatest accuracy to forecast  $y$  for a given  $x$  is how models are trained. The most optimal regression match is identified by the best  $\theta_1$  and  $\theta_2$  values. The goal of linear regression is to find the greatest accurate regression line. To allow the model to minimize error, the variance between its real  $y$  value and what was anticipated should be as small as possible. The Root Mean Squared Error between the anticipated value  $y$  and the real value  $y$  is the Cost Function ( $J$ ) for Linear Regression. The equations below are used to calculate the Cost Function.

$$\text{minimize } \frac{1}{n} \sum_{i=1}^n (\text{pred}_i - y_i)^2 \quad \text{Equation 2.}$$

$$J = \frac{1}{n} \sum_{i=1}^n (\text{pred}_i - y_i)^2 \quad \text{Equation 3.}$$

Gradient Descent is utilized to lower the Cost Function (minimize the RMSE value) while adjusting the  $\theta_1$  and  $\theta_2$  values in the model in order to attain the most accurate line (GeeksforGeeks, 2018).

#### 4.3.1.2 Support Vector Machine (SVM)

Support Vector Machine is a type of supervised learning derived machine learning algorithm. It can manage both regression and classification problems. The algorithm plots each data point as a point in n-dimensional space. The number of features is represented by the n value. Ray (2021) defines each feature as a specific coordinate. An SVM classifier is used to find the hyperplane that separates the classes during classification. The goal of SVM is to maximize the distance between the data points and the hyperplane. The loss function used to maximize the margin is hinge loss (Gandhi, 2018).

$$c(x, y, f(x)) = \begin{cases} 0, & \text{if } y * f(x) \geq 1 \\ 1 - y * f(x), & \text{else} \end{cases} \quad c(x, y, f(x)) = (1 - y * f(x))_+$$

Hinge loss function (function on left can be represented as a function on the right)

Eq. 3

When the actual and predicted values are the same, the SVM cost is zero. If the actual and predicted values differ, the loss value is computed. To balance the loss and margin maximization, the value for regularization is added to the cost function. The cost of the function with the value to regulate is shown in the equation below.

$$\min_w \lambda \| w \|^2 + \sum_{i=1}^n (1 - y_i \langle x_i, w \rangle)_+$$

Loss function for SVM

Eq. 4

The gradient is determined using the derivatives that are partial of the weights. The gradient is then used to update the weights. The equation below shows the gradient equation.

$$\frac{\delta}{\delta w_k} \lambda \| w \|^2 = 2\lambda w_k$$

$$\frac{\delta}{\delta w_k} (1 - y_i \langle x_i, w \rangle)_+ = \begin{cases} 0, & \text{if } y_i \langle x_i, w \rangle \geq 1 \\ -y_i x_{ik}, & \text{else} \end{cases}$$

Gradients

Eq. 5

SVM that correctly predicts the class will only require a gradient update from the regularization parameter as shown in the equation below.

$$w = w - \alpha \cdot (2\lambda w)$$

Gradient Update — No misclassification

Eq. 6

SVM that fails to predict the class data point must include the loss as well as the parameter to regulate in order to perform a slope update, as shown in the equation below (Gandhi, 2018).

$$w = w + \alpha \cdot (y_i \cdot x_i - 2\lambda w)$$

Gradient Update — Misclassification

Eq. 7

### 4.3.1.3 Random Forest

Random Forest is a supervised machine learning-derived machine learning algorithm. It can manage problems with classification as well as regression. The algorithm is made up of many decision trees that work together as an ensemble. Ensemble learning is a technique that combines multiple classifiers to find the best solutions to complex problems (Yiu, 2021). Each decision tree divides the class prediction and assigns supports to each one. Because the trees protect each other from errors, the class with the most supports is considered the model's prediction. Bagging, a collaborative meta-algorithm utilized to improve the accuracy of algorithms for machine learning, is used to train the technique known as Random Forest. The algorithm also improves its accuracy by averaging or meaning the output of various trees (Introduction to Random Forest in Machine Learning, 2020).

### 4.3.1.4 K-Nearest Neighbors

K-Nearest Neighbors is a controlled machine learning technique. It can manage problems with classification and regression. The model implies that identical things exist nearby (Harrison, 2019).

### 4.3.1.5 Stochastic Gradient Descent

The stochastic gradient descent technique is an artificial intelligence algorithm for learning that finds the value of the model that best match the relationship between actual and predicted outputs. Instead of using the entire dataset for each iteration, the model chooses a few samples

at random. Instead of finding the gradient of the cost function for the whole collection, the model finds the slope of the cost function for each class at each iteration. The SGD algorithm is depicted in the equation below (GeeksforGeeks, 2021).

$$\begin{aligned} &\text{for } i \text{ in } \tilde{\text{range}}(m) : \\ &\theta_j = \theta_j - \alpha (\tilde{y}^i - y^i) x_j^i \end{aligned}$$

Eq. 8

### 4.3.2 Cross Validation

Validation is the process of comparing the expected against the actual output and determining whether the results quantify the hypothesized relationship between the variables. The cross-validation is a data sampling retakes approach utilized to assess forecasting model generalization and prevent overfitting from occurring. It may also refer to evaluating machine learning models by training them on a subset of the dataset and then assessing their accuracy on the remaining data subset. A cross-validation technique has been used to compare different supervised machine learning model prediction accuracy.

#### Importing the necessary libraries..

Import Pandas and Numpy for a table data and quantitative computing, respectively.

```
import pandas as pd.
```

```
import numpy as np.
```

#### Determine which columns to use from the dataset.

```
numeric_cols = columns_to_use = ['Number of People Travelling buy bus',  
Person_Truck_Traffic', 'COVID-19 Cases']
```

#### Importing the data set to the Jupyter Notebook

```
raw_data = pd.read_csv('Predict_Spread_of_COVID-19_Combined_Data.csv', usecols =  
numeric_cols)
```

#### To understand the dataset by generating high-level information using

```
raw_data.info()
```



```
In [34]: #Understand dataset by generating more information about the dataset
raw_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 3 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Number of People Travelling buy bus    60 non-null     float64
1   Person_Truck_Traffic                    98 non-null     float64
2   Covid-19 Cases                          38 non-null     float64
dtypes: float64(3)
memory usage: 2.5 KB
```

**View the top five rows of the dataset using.**

`raw_data.head()`

```
In [33]: #view the first 5 rows of the dataset
raw_data.head()
```

Out[33]:

	Number of People Travelling buy bus	Person_Truck_Traffic	Covid-19 Cases
0	NaN	NaN	NaN
1	NaN	NaN	NaN
2	NaN	23300.0	NaN
3	NaN	20914.0	NaN
4	NaN	22290.0	NaN

**Check if we have any missing data.**

`raw_data.isnull().mean()` is used to get the average of the missing values in each column.

```
In [5]: raw_data.isnull().mean()
```

```
Out[5]: Number of People Travelling buy bus    0.40
Person_Truck_Traffic                          0.02
Covid-19 Cases                                0.62
dtype: float64
```

### 4.3.2.1 Data Clean Up

**Create an object of simple Imputer.**

`imputer = SimpleImputer(missing_values = np. nan, strategy = 'mean')` replaces missing values using the mean of the column.

**Fit the imputer on the data using.**

```
imputer.fit(raw_data)
```

```
SimpleImputer()
```

**View what values the imputer has fit on the data by providing a mean.**

```
imputer.statistics_
```

---

```
In [8]: imputer.statistics_
```

```
Out[8]: array([ 37777.294      , 41713.83846939, 211363.89473684])
```

**Verify if the imputer has given us the correct data.**

```
raw_data.mean()
```

---

```
In [9]: raw_data.mean()
```

```
Out[9]: Number of People Travelling buy bus    37777.294000  
Person_Truck_Traffic                        41713.838469  
Covid-19 Cases                              211363.894737  
dtype: float64
```

**Input the values to the data by using transform.**

```
raw_data = imputer.transform(raw_data)
```

**Convert the data into a dataframe for easy interpretation.**

```
raw_data_1 = pd.DataFrame(raw_data, columns = numeric_cols)
```

**Confirm if we still have any missing data.**

```
raw_data_1.isnull().mean()
```

---

```
In [12]: raw_data_1.isnull().mean()
```

```
Out[12]: Number of People Travelling buy bus    0.0  
Person_Truck_Traffic                        0.0  
Covid-19 Cases                              0.0  
dtype: float64
```

---

To view the data in the pandas data frame

`raw_data_1`

```
In [13]: raw_data_1
```

Out[13]:

	Number of People Travelling buy bus	Person_Truck_Traffic	Covid-19 Cases
0	37777	41713	211363
1	37777	41713	211363
2	37777	23300	211363
3	37777	20914	211363
4	37777	22290	211363
...	...	...	...
95	37777	58017	342470
96	37777	57977	342817
97	37777	58534	342919
98	37777	59168	342976
99	37777	59623	342992

100 rows × 3 columns

**Combine two Dataset columns to get the total number of people traveling by bus and truck.**

```
raw_data_1['Total_Number_Of_Travellers'] = raw_data_1.apply(lambda x: x['Number of People Travelling buy bus'] + x['Person_Truck_Traffic'], axis=1)
```

```
print (raw_data_1)
```

```
raw_data_1['Total_Number_Of_Travellers'] = raw_data_1.apply(lambda x: x['Number of People Travelling buy bus'] + x['Person_Truck_Traffic'])
print (raw_data_1)
```

	Number of People Travelling buy bus	Person_Truck_Traffic	Covid-19 Cases
0	37777	41713	211363
1	37777	41713	211363
2	37777	23300	211363
3	37777	20914	211363
4	37777	22290	211363
..	...	...	...
95	37777	58017	342470
96	37777	57977	342817
97	37777	58534	342919
98	37777	59168	342976
99	37777	59623	342992

	Total_Number_Of_Travellers
0	79490
1	79490
2	61077
3	58691
4	60067
..	...
95	95794
96	95754
97	96311
98	96945
99	97400

[100 rows x 4 columns]

## Building a machine Learning Models

The data is separated into two arrays: x, which includes data utilized to make forecasts, and y, which contains data to be predicted. We divided the data into training and test sets using Scikit-learn. The training data constitutes 40% of the total data set.

```
x_train, x_test, y_train, y_test = train_test_split(raw_data_1,y, test_size=0.40)
```

Import all the models for accuracy comparison.

```
#Importing Machine Learning models
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
```

## 4.4 RESULTS

### 4.4.1 Models Results Evaluation Metrics

Model metrics for assessment are critical in this study for identifying the best model utilized to accurately forecast the propagation of COVID-19. Unsupervised and supervised learning are supported by the Sklearn library. It also provides the machine learning model's score by specifying the accuracy and errors.

#### 4.4.1.1 F1 Score

The F1 score is the average of recall and precision. It assesses model accuracy on a scale of 0 to 1. The score indicates how many instances were correctly classified. An elevated rating for F1 suggests that the predicted outcome is very good (Mishra, 2021).

$$F1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}}$$

F1 Score

Precision can be described as the total number of accurate outcomes that are favorable divided by the number of favorable outcomes forecast by the classification algorithm.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

Precision

Recall is referred to as the quantity of accurate outcomes divided by the total quantity that are relevant.

$$Precision = \frac{TruePositives}{TruePositives + FalseNegatives}$$

Recall

#### 4.4.1.2 Mean Absolute Error (MAE)

The average variance between the actual and predicted values is referred to as the MAE. This metric indicates how far the actuals differ from the predicted. Models with a lower MAE are more accurate (Mishra, 2021).

$$MeanAbsoluteError = \frac{1}{N} \sum_{j=1}^N |y_j - \hat{y}_j|$$

#### 4.4.1.3 Mean Squared Error (MSE)

The average square of the variance between the actual and predicted values is the MSE. This metric aids in calculating the gradient. Models with a lower MSE are more accurate (Mishra, 2021).

$$MeanSquaredError = \frac{1}{N} \sum_{j=1}^N (y_j - \hat{y}_j)^2$$

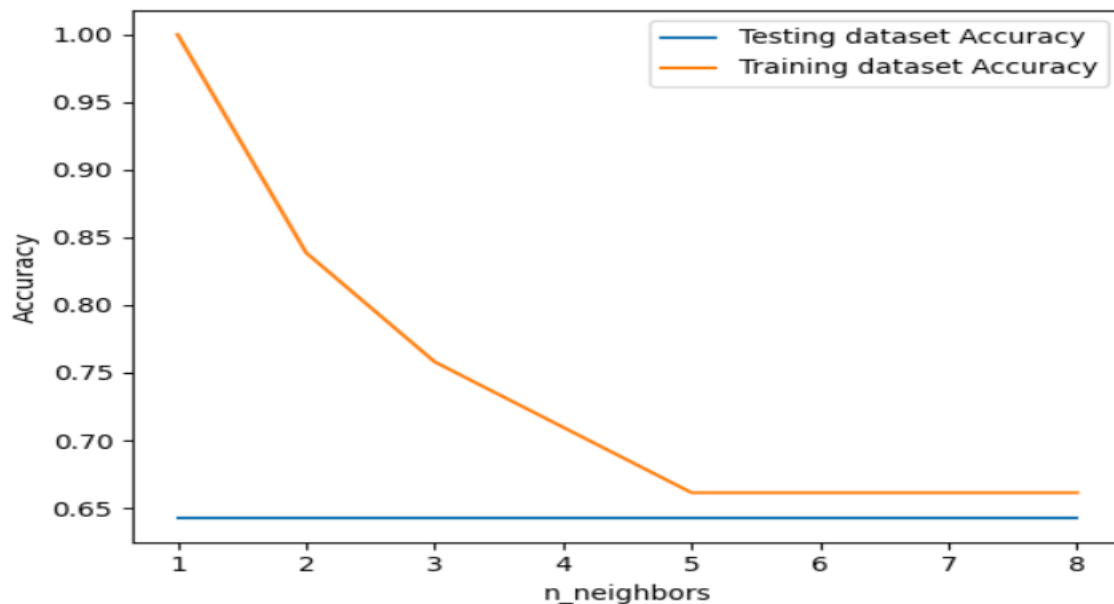
Mean Squared Error

#### 4.4.2 Model Evaluation Results for Combine Truck and Bus Traffic Monthly Data

The F1 score for K-Nearest Neighbor is 0.643, with a MAE of 4151 and a MSE of 1.08. In comparison to KNN, Logistic Regression has a slightly poor F1 score of 0.64. The MAE for logistic regression is 3967, while the MSE is 1.11.

	Model	F1_score	MAE	MSE
0	LogReg	0.640	3967.071400	1.112190e+08
1	KNN	0.643	4151.857143	1.087495e+08

K-Nearest Neighbor has a higher accuracy score where neighbors are equal to 3 and a low accuracy score where neighbors are equal to 9.



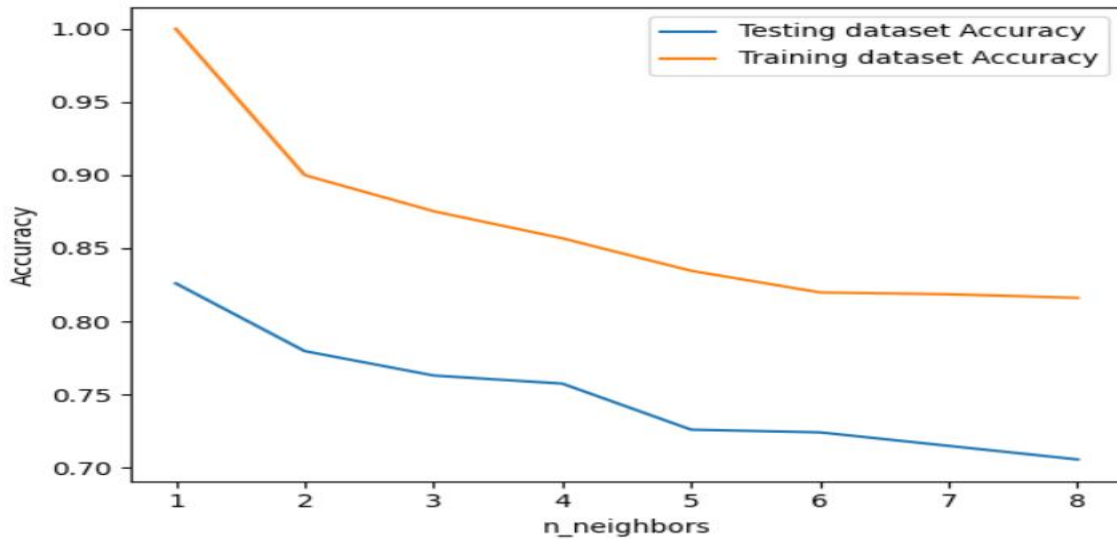
**Figure 4. 5 Compare the accuracy score for KNN using combined truck and bus data**

**4.4.3 Model Evaluation Results for Truck Traffic Daily Data**

The F1 score for K-Nearest Neighbor is 0.83, with a MAE of 4 and a MSE of 176.7. In comparison to KNN, Logistic Regression has a slightly poor F1 score of 0.56. The MAE for logistic regression is 32.9, while the MSE is 3117.88.

	Model	F1_score	MAE	MSE
0	LogReg	0.56	32.914800	3117.877800
1	KNN	0.83	4.038889	176.653704

K-Nearest Neighbor has a higher accuracy score where neighbors are equal to 1 and a low accuracy score where neighbors are equal to 9.



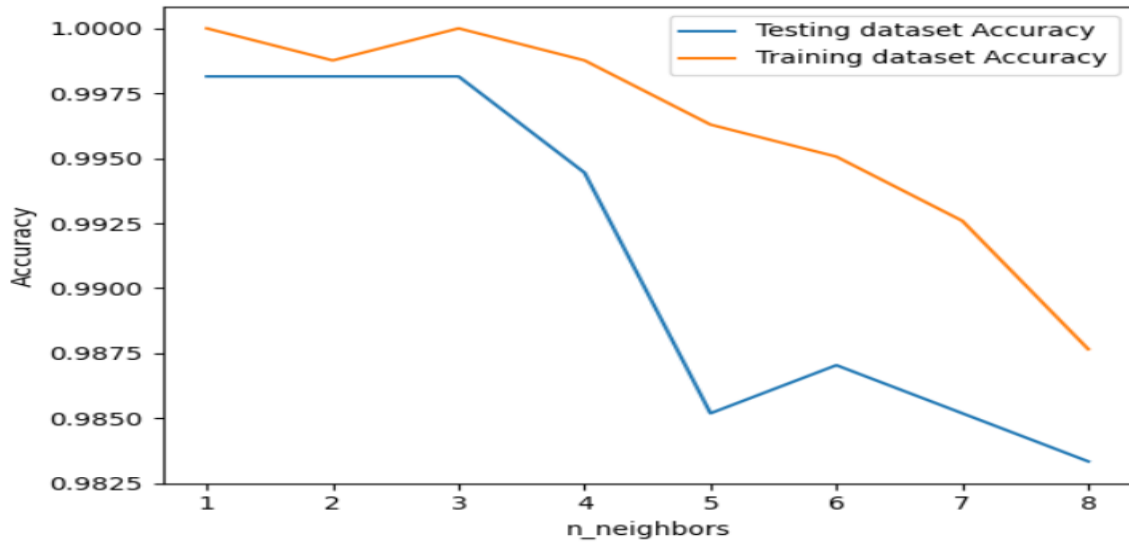
**Figure 4. 6 Compare the accuracy score for KNN using combined truck data**

**4.4.4 Model Evaluation Results for Bus Traffic Daily Data**

The F1 score for K-Nearest Neighbor is 1, with a MAE of 0.56 and a MSE of 168. In comparison to KNN, Logistic Regression has a poor F1 score of 0.46. The MAE for logistic regression is 160.56, while the MSE is 48168.

	Model	F1_score	MAE	MSE
0	LogReg	0.46	160.557400	48168.31300
1	KNN	1.00	0.557407	167.77963

K-Nearest Neighbor has a higher accuracy score where neighbors are equal to 2 and a low accuracy score where neighbors are equal to 9.



**Figure 4. 7 Compare the accuracy score for KNN using combined bus data.**

#### 4.5 DISCUSSION

It is important to use machine learning models to identify how human mobility patterns contribute to contagious illnesses spread. We have demonstrated that existing data and models from various sources can be used to forecast the spread of transmittable diseases. The discrepancy in available data resulted in lower accuracy when predicting using the monthly truck and bus traffic using the K-nearest neighbor model, with an accuracy of 0.64. Predictions based on daily traffic for the KNN model gave an accuracy of only 0.83, while daily bus data gave an accuracy of one. This disease modeling approach and prediction using K-Nearest Neighbor can be adopted for other diseases that spread by human movement.

#### 4.6 CONCLUSION

In this research, we have analyzed how human mobility contributes to the spread of infectious diseases considering that the disease can move from one region to another when an infected person travels from one place to another. Truck transport interconnects the country to other countries through the border points while most of the population travels using bus transport.



We examined epidemiological models such as KenyaCov, which are used to anticipate virus distribution in distinct locations and age groups by simulating virus transmission with an emphasis on epidemiological characteristics. We also looked at machine learning models like the Stochastic model, which predicted disease outbreaks with high accuracy.

We developed two models, KNN and Logistic Regression, to see which one can learn on a given subset of data and has the highest prediction accuracy score. The truck data was modelled using NCTTCA reports, which provide average daily weighted traffic collection at five weighbridges. Bus statistics for the Nairobi Metropolitan Area was developed using Trip Generation, which forecasted trips in each traffic analysis zone by applying socioeconomic data to calculate trips attracted and generated to Traffic Analysis Zones. COVID-19 data was generated using global COVID-19 data. In estimating the transmission of COVID-19 via transportation systems, KNN showed a higher accuracy of one for bus data and 0.643 for combined daily truck and bus data. KNN also demonstrated an accuracy of 0.83 for truck data.

This study has confirmed that it is possible to access and model data from various sources. The modeled data can be used to make accurate predictions using machine learning models.

#### **4.7 FUTURE WORK**

The dataset was derived from the NUTRANS study, which was conducted for the urban Nairobi Metropolitan Area. The ten counties studied in this research include both rural and urban areas. It is necessary to narrow the research and demonstrate how the model works in rural and urban areas separately.

It is also necessary for demonstrating the efficacy of the KNN model in forecasting transmission of diseases through human movement. KNN models can also be used to show the propagation of illnesses to nearby cities and regions.

## REFERENCES

- An overview of the legal and ethical issues in healthcare. (2020). StPatrick. <https://www.st-patricks.ac.uk/blog/posts/2019/october/an-overview-of-the-legal-and-ethical-issues-in-healthcare/>
- Ardabili, S. F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A. R., Reuter, U., Rabczuk, T., & Atkinson, P. M. (2020). COVID-19 Outbreak Prediction with Machine Learning. <https://ssrn.com/abstract=3580188>
- Ayris, D., Imtiaz, M., Horbury, K., Williams, B., Blackney, M., See, C. S. H., & Shah, S. a. A. (2022). Novel deep learning approach to model and predict the spread of COVID-19. *Intelligent Systems With Applications*, 14, 200068. <https://doi.org/10.1016/j.iswa.2022.200068>
- Bates, J. (2000). HISTORY OF DEMAND MODELING. IN: HANDBOOK OF TRANSPORT MODELLING. <https://www.semanticscholar.org/paper/HISTORY-OF-DEMAND-MODELING.-IN%3A-HANDBOOK-OF-Bates/22a2edad2cfbca97cd997b8c18be2587eed498c>
- Brand, S., Aziza, R., Kombe, I., Agoti, C., Hilton, J., Rock, K., Parisi, A., Nokes, D. J., Keeling, M., & Barasa, E. (2020). Forecasting the scale of the COVID-19 epidemic in Kenya. <https://doi.org/10.1101/2020.04.09.20059865>
- Clark, A., Jit, M., Warren-Gash, C., Guthrie, B., Wang, H., Mercer, S. W., Sanderson, C., McKee, M., Troeger, C., Ong, K., Checchi, F., Perel, P., Joseph, S., Gibbs, H., Banerjee, A., Eggo, R. M., Nightingale, E., O'Reilly, K., Jombart, T., . . . Jarvis, C. I. (2020). Global, regional, and national estimates of the population at increased risk of severe COVID-19 due to underlying health conditions in 2020: a modelling study. *The Lancet Global Health*, 8(8), e1003–e1017. [https://doi.org/10.1016/s2214-109x\(20\)30264-3](https://doi.org/10.1016/s2214-109x(20)30264-3)
- Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., Pastore Y Piontti, A., Mu, K., Rossi, L., Sun, K., Viboud, C., Xiong, X., Yu, H., Halloran, M. E., Longini, I. M., & Vespignani, A. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science*, 368(6489), 395–400. <https://doi.org/10.1126/science.aba9757>
- Chu, A. M., Tiwari, A., Chan, J. N., & So, M. K. (2021). Are travel restrictions helpful to control the global COVID-19 outbreak? *Travel Medicine and Infectious Disease*, 41, 102021. <https://doi.org/10.1016/j.tmaid.2021.102021>
- Communicable diseases*. (2020, November 4). Amref Health Africa in Kenya. <https://amref.org/kenya/our-work/pillar-2-innovative-health-services-solutions/communicable-diseases/>
- Creswell, J. W. (2005). Educational research planning, conducting, and evaluating quantitative and qualitative research. Upper Saddle River, NJ Pearson. - References - Scientific

research publishing. SCIRP Open  
Access. [https://www.scirp.org/\(S\(351jmbntvnsjt1aadkposzje\)\)/reference/ReferencesPapers.aspx?ReferenceID=554097](https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=554097)

Diffusion of Innovation Theory. (2019, September 9). Diffusion of Innovation Theory. <https://sphweb.bumc.bu.edu/otlt/MPH-Modules/SB/BehavioralChangeTheories/BehavioralChangeTheories4.html>

Education, I. C. (2021, August 3). *Neural Networks*. IBM. <https://www.ibm.com/cloud/learn/neural-networks>

Exponential smoothing explained: <https://otexts.com/fpp2/expsmooth.html>; Arroyo, J., San Roque, A. M., Maté, C., & Sarabia, A. (2007). Exponential smoothing methods for interval time series. In *Proceedings of the 1st European Symposium on Time Series Prediction* (pp. 231-240).

Fang, H., Wang, L., & Yang, Y. (2020). Human Mobility Restrictions and the Spread of the Novel Coronavirus (2019-nCoV) in China. *SSRN Electronic Journal*. Published. <https://doi.org/10.2139/ssrn.3756202>

Firda, R., & Hyunsoo, Lee. (2020). Hybrid Deep Learning-Based Epidemic Prediction Framework of COVID-19: South Korea Case

Gandhi, R. (2018, July 5). *Support Vector Machine — Introduction to Machine Learning Algorithms*. Medium. <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>

Garg, S., Kumar, S., & Muhuri, P. K. (2022). A novel approach for COVID-19 Infection forecasting based on multi-source deep transfer learning. *Computers in Biology and Medicine*, 149, 105915. <https://doi.org/10.1016/j.compbiomed.2022.105915>

GeeksforGeeks. (2021, September 13). *ML | Stochastic Gradient Descent (SGD)*. <https://www.geeksforgeeks.org/ml-stochastic-gradient-descent-sgd/>

Grépin, K. A., Ho, T. L., Liu, Z., Marion, S., Piper, J., Worsnop, C. Z., & Lee, K. (2021). Evidence of the effectiveness of travel-related measures during the early phase of the COVID-19 pandemic: a rapid systematic review. *BMJ Global Health*, 6(3), e004537. <https://doi.org/10.1136/bmjgh-2020-004537>

Guo, Y. R., Cao, Q. D., Hong, Z. S., Tan, Y. Y., Chen, S. D., Jin, H. J., Tan, K. S., Wang, D. Y., & Yan, Y. (2020). The origin, transmission, and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak – an update on the status. *Military Medical Research*, 7(1). <https://doi.org/10.1186/s40779-020-00240-0>

Gupta, A., Gharehgozli, A., & Nazarian, D. (2020). Developing a Machine Learning Framework to Determine the Spread of COVID-19. <https://ssrn.com/abstract=3635211>

- Harrison, O. *Machine Learning Basics with the K-Nearest Neighbors Algorithm*. Medium. <https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761> (Accessed 2019, July 14).
- Hotz, N. (2023, January 19). What is CRISP DM? - Data Science Process Alliance. Data Science Process Alliance. <https://www.datascience-pm.com/crisp-dm-2/>
- How government can use AI and ML to identify spreading infectious diseases. (2018). ResearchGate. [https://www.researchgate.net/publication/349310427\\_How\\_government\\_can\\_use\\_AI\\_and\\_ML\\_to\\_identify\\_spreading\\_infectious\\_diseases](https://www.researchgate.net/publication/349310427_How_government_can_use_AI_and_ML_to_identify_spreading_infectious_diseases) (Accessed 2019, July 14).
- Infectious diseases - Symptoms and causes. (2021, April 7). Mayo Clinic. <https://www.mayoclinic.org/diseases-conditions/infectious-diseases/symptoms-causes/syc-20351173>
- Infectious Diseases –. National Foundation for Infectious Diseases. Retrieved June 5, 2021, from <https://www.nfid.org/infectious-diseases/>
- India records 300,000 COVID deaths as pandemic rages. (2021, May 24). BBC News. <https://www.bbc.com/news/world-asia-57224565>
- Introduction to Random Forest in Machine Learning*. (2020, December 11). Engineering Education (EngEd) Program | Section. <https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/>
- Japan International Cooperation Agency (JICA). (2006). The Study on Master Plan for Urban Transport in The Nairobi Metropolitan Area in The Republic of Kenya (No. SD JR 06-041). [https://openjicareport.jica.go.jp/pdf/11823093\\_01.pdf](https://openjicareport.jica.go.jp/pdf/11823093_01.pdf)
- Kagucia, E. W., Gitonga, J. N., Kalu, C., Ochomo, E., Ochieng, B., Kuya, N., Karani, A., Nyagwange, J., Karia, B., Mugo, D., Karanja, H. K., Tuju, J., Mutiso, A., Maroko, H., Okubi, L., Maitha, E., Ajuck, H., Mukabi, D., Moracha, W., . . . Scott, J. A. G. (2021). Anti-Severe Acute Respiratory Syndrome Coronavirus 2 Immunoglobulin G Antibody Seroprevalence Among Truck Drivers and Assistants in Kenya. *Open Forum Infectious Diseases*, 8(7). <https://doi.org/10.1093/ofid/ofab314>
- Kavadi, D. P., Patan, R., Ramachandran, M., & Gandomi, A. H. (2020). Partial derivative Nonlinear Global Pandemic Machine Learning prediction of COVID-19. *Chaos, Solitons and Fractals*, 139. <https://doi.org/10.1016/j.chaos.2020.110056>
- Kenya Map | Map of Kenya. (2021). Maps of the World. <https://www.mapsofworld.com/kenya/>
- Kenya National Bureau of Statistics. (2020, February 21). *2019 Kenya Population and Housing Census Volume IV: Distribution of Population by Socio-Economic*

- Characteristics - Kenya National Bureau of Statistics.*  
<https://www.knbs.or.ke/?wpdmpro=2019-kenya-population-and-housing-census-volume-iv-distribution-of-population-by-socio-economic-characteristics>
- Kenya National Highways Authority. (2023, March 1). Home - Kenya National Highways Authority. Kenya National Highways Authority - Kenya National Highways Authority. [https://www.kenha.co.ke/index.php?option=com\\_content&view=article&id=37&Itemid=108&limitstart=6](https://www.kenha.co.ke/index.php?option=com_content&view=article&id=37&Itemid=108&limitstart=6)
- Kenya Population (2021) - Worldometer. (2019). Kenya Population (2021) - Worldometer. <https://www.worldometers.info/world-population/kenya-population/>
- Kibera: a look inside Africa's largest slum. (2020, November 26). Concern Worldwide. <https://www.concern.org.uk/news/kibera-look-inside-africas-largest-slum>
- Kim, J. (2021, February 24). *Infectious disease outbreak prediction using*. . . Scientific Reports. [https://www.nature.com/articles/s41598-021-83926-2?error=cookies\\_not\\_supported&code=7852f483-3dfb-4d97-87d3-d9bb99943c43](https://www.nature.com/articles/s41598-021-83926-2?error=cookies_not_supported&code=7852f483-3dfb-4d97-87d3-d9bb99943c43)
- Linear regression for machine learning*. (2020, August 14). Machine Learning Mastery. <https://machinelearningmastery.com/linear-regression-for-machine-learning/>
- Meloni, S., Perra, N., Arenas, A., Gómez, S., Moreno, Y., & Vespignani, A. (2011). Modeling human mobility responses to the large-scale spreading of infectious diseases. *Scientific Reports*, 1(1). <https://doi.org/10.1038/srep00062>
- Mertler, C. A. (2019). *Action research - International student edition: Improving schools and empowering educators*.
- Mills, G. E., & Gay, L. R. (2016) *Education research: Competencies for analysis and applications*. London, England: Pearson Education. (2018). *Journal of Applied Learning and Teaching*, 1(2). <https://doi.org/10.37074/jalt.2018.1.2.14>
- Ming, R. X., Liu, J., Cheung, W. K. W., & Wan, X. (2016). Stochastic modeling of infectious diseases for heterogeneous populations. *Infectious Diseases of Poverty*, 5(1). <https://doi.org/10.1186/s40249-016-0199-5>
- Mishra, A. (2021, December 29). Metrics to Evaluate your Machine Learning Algorithm. Medium. <https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234>
- Northern Corridor Transit and Transport Coordination Authority: Northern Corridor Maps*. (n.d.). Northern Corridor Maps. Retrieved September 9, 2021, from <http://www.ttcanc.org/maps.php>
- Ogundokun, R. O., & Awotunde, J. B. (2020). MACHINE LEARNING PREDICTION FOR COVID-19 PANDEMIC IN INDIA 1, \*. <https://doi.org/10.1101/2020.05.20.20107847>
- Odhiambo, J. O., Ngare, P., Weke, P., & Otieno, R. O. (2020). Modeling of COVID-19 Transmission in Kenya Using Compound Poisson Regression Model. *Journal of*

- Advances in Mathematics and Computer Science, 35(2), 101–111.  
<https://doi.org/10.9734/jamcs/2020/v35i230252>
- Pan African Medical Journal. (2017, March 10). *Factors associated with cholera in Kenya, 2008–2013*. Copyright PAMJ. <https://panafrican-med-journal.com/content/article/28/101/full/>
- Pinter, G., Felde, I., Mosavi, A., Ghamisi, P., & Gloaguen, R. (n.d.). COVID-19 Pandemic Prediction for Hungary; a Hybrid Machine Learning Approach. <https://ssrn.com/abstract=3590821>
- Ray, S. (2021, August 26). *SVM / Support Vector Machine Algorithm in Machine Learning*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/>
- Ritchie, H. (2020, March 5). *Coronavirus Pandemic (COVID-19) - Statistics and Research*. Our World in Data. <https://ourworldindata.org/coronavirus>
- Roy, A. N., Jose, J., Gautam, N., Nathalia, D., & Suresh, A. (2020). Prediction and Spread Visualization of the COVID-19 Pandemic using Machine Learning. <https://doi.org/10.20944/preprints202005.0147.v1>
- Sabin, N. S., Calliope, A. S., Simpson, S. V., Arima, H., Ito, H., Nishimura, T., & Yamamoto, T. (2020). Implications of human activities for (re)emerging infectious diseases, including COVID-19. *Journal of Physiological Anthropology*, 39(1). <https://doi.org/10.1186/s40101-020-00239-5>
- Sufian, A., Ghosh, A., Sadiq, A. S., & Smarandache, F. (2020). A Survey on Deep Transfer Learning to Edge Computing for Mitigating the COVID-19 Pandemic: DTL-EC. *Journal of Systems Architecture*, 108. <https://doi.org/10.1016/j.sysarc.2020.101830>
- Sujath, R., Chatterjee, J. & Hassaniien, A. (2020, July 27). A machine learning forecasting model for the COVID-19 pandemic in India. *Stochastic Environmental Research and Risk Assessment*. [https://link.springer.com/article/10.1007/s00477-020-01827-8?error=cookies\\_not\\_supported&code=6222a72a-072c-4e8b-9035-80c6d2e2e97e](https://link.springer.com/article/10.1007/s00477-020-01827-8?error=cookies_not_supported&code=6222a72a-072c-4e8b-9035-80c6d2e2e97e)
- Surya, L. (2018). How government can use AI and ML to identify spreading infectious diseases. ResearchGate. <https://doi.org/10.1729/Journal.25743>
- The Ebola outbreak. (2014). *The Pharmaceutical Journal*. Published. <https://doi.org/10.1211/pj.2014.20066306>
- Thompson, C., Saxberg, K., Lega, J., Tong, D., & Brown, H. (2019). A cumulative gravity model for inter-urban spatial interaction at different scales. *Journal of Transport Geography*, 79, 102461. <https://doi.org/10.1016/j.jtrangeo.2019.102461>
- Trizer, M. (2019, November 4). *2019 Kenya Population and Housing Census Results*. Kenya National Bureau of Statistics. <https://www.knbs.or.ke/?p=5621>

- Tuli, S., Tuli, S., Tuli, R., & Gill, S. (2020, September 1). Predicting the growth and trend of the COVID-19 pandemic using machine learning and cloud computing. ScienceDirect. <https://www.sciencedirect.com/science/article/abs/pii/S254266052030055X>
- Use of artificial intelligence in infectious diseases.* (2020). PubMed Central (PMC). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7153335/>
- Wang, P., Zheng, X., Li, J., & Zhu, B. (2020). Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. *Chaos, Solitons & Fractals*, 139, 110058. <https://doi.org/10.1016/j.chaos.2020.110058>
- WHO Coronavirus (COVID-19) Dashboard. (2021). With Vaccination Data. <https://COVID19.who.int/>
- WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020.* (2020, March 11). World Health Organisation. <https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-COVID-19---11-march-2020>
- World Bank Open Data. (n.d.-c). World Bank Open Data. <https://data.worldbank.org/indicator/SP.POP.GROW?locations=KE>
- Xu, L., Magar, R., & Farimani, A. B. (2022). Forecasting COVID-19 new cases using deep learning methods. *Computers in Biology and Medicine*, 144, 105342. <https://doi.org/10.1016/j.combiomed.2022.105342>
- Yiu, T. (2021, September 29). *Understanding Random Forest - Towards Data Science.* Medium. <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>

## APPENDICES

### Appendix A - Data Sources Summary

Data Source	Link	Usage in the report
World Health Organization COVID-19 Global Data	<a href="https://COVID19.who.int/WHO-COVID-19-global-data.csv">https://COVID19.who.int/WHO-COVID-19-global-data.csv</a>	Determine the number of COVID-19 infections in Kenya
COVID-19 Cases reported in Kenya	<a href="https://COVID19.who.int/region/afro/country/ke">https://COVID19.who.int/region/afro/country/ke</a>	Determine the number of COVID-19 Cases in Kenya
National Census, 2019 volumes IV and III	<a href="https://www.knbs.or.ke/?wpdmp=2019-kenya-population-and-housing-census-volume-iv-distribution-of-population-by-socio-economic-characteristics">https://www.knbs.or.ke/?wpdmp=2019-kenya-population-and-housing-census-volume-iv-distribution-of-population-by-socio-economic-characteristics</a>	The report has been used to generate below data: <ol style="list-style-type: none"> <li>1. Socio-economic characteristics for traffic analysis zones.</li> <li>2. Population Age 5 years and above</li> <li>3. Working population</li> <li>4. Population attending school</li> </ol>
World Development Indicators	<a href="https://databank.worldbank.org/reports.aspx?source=2&amp;series=SP.POP.GROW&amp;country=KEN#">https://databank.worldbank.org/reports.aspx?source=2&amp;series=SP.POP.GROW&amp;country=KEN#</a>	Kenyan population growth rate
Kenya Integrated Household Budget Survey (KIHBS) 2015/16	<a href="https://www.knbs.or.ke/2015-16-kenya-integrated-household-budget-survey-kihbs-progress-report-october-2015/?option=com_phocadownload&amp;view=category&amp;download=797:2015-16-kenya-integrated-household-budget-survey-kihbs-progress-report-october-2015&amp;id=129:2015-16-kenya-integrated-household-budget-survey-kihbs&amp;Itemid=599">https://www.knbs.or.ke/2015-16-kenya-integrated-household-budget-survey-kihbs-progress-report-october-2015/?option=com_phocadownload&amp;view=category&amp;download=797:2015-16-kenya-integrated-household-budget-survey-kihbs-progress-report-october-2015&amp;id=129:2015-16-kenya-integrated-household-budget-survey-kihbs&amp;Itemid=599</a>	Determine which mode of transport every household uses.
Northern Corridor Transit and Transport Coordination Authority (NCTTCA) reports	<a href="https://top.ttcanc.org/indicators/indicator?show=19">https://top.ttcanc.org/indicators/indicator?show=19</a>	Daily Average weighted traffic captured at weighbridges
Health and Mobility Research Questionnaire		Questionnaires were administered to Bus employees and passengers.

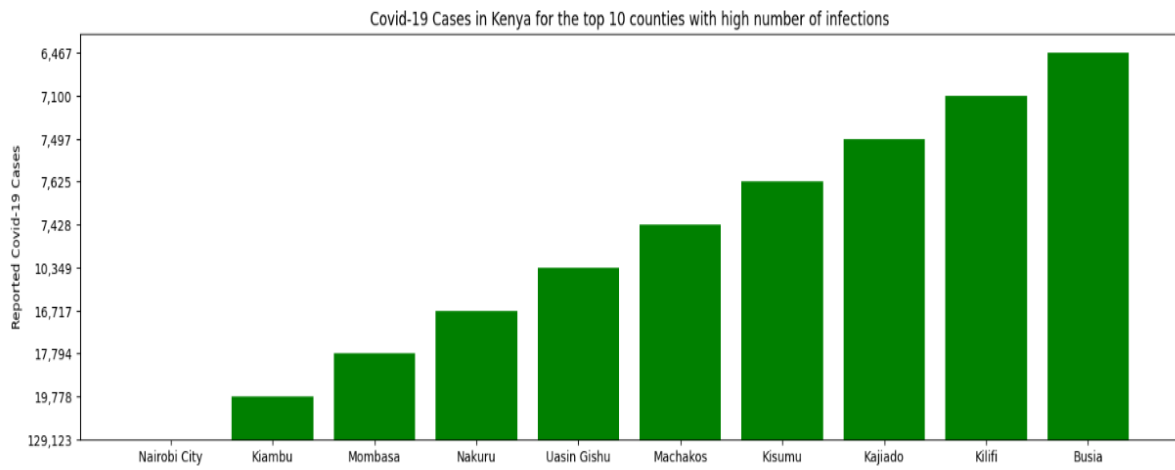


**Appendix B - Kenya expressways, main roads, and streets network Map**



Source: <https://www.mapsofworld.com/kenya/>

## Appendix C - COVID-19 cases in Kenya for the top ten counties with high numbers of infections Interview Questionnaires



**Appendix D - Ethical clearance consideration/permit**

Letter from the University of Nairobi – Director School of Computing and Informatics



**UNIVERSITY OF NAIROBI  
COLLEGE OF BIOLOGICAL AND PHYSICAL SCIENCES  
SCHOOL OF COMPUTING AND INFORMATICS**

Telephone: 4447870/4446543/4444919  
Telegrams: "Varsity" Nairobi  
Telefax: +254-20-4447870  
Email: [director-sci@uonbi.ac.ke](mailto:director-sci@uonbi.ac.ke)

P. O. Box 30197  
00100 GPO  
Nairobi, Kenya

Our Ref: UON/CBPS/MS/CI/2019

21<sup>st</sup> June 2021

**TO WHOM IT MAY CONCERN**

Dear Sir/Madam

**RE: RESEARCH PERMIT – FREDRICK ONYANGO – REG. NO. P52/35154/2019**

The above named is a bona fide student pursuing an MSc course in Computational Intelligence at the School of Computing and Informatics, University of Nairobi. He is currently carrying out his research on the project entitled "***A Study on Predicting the Spread of Infectious Disease in Kenya via Transport System using Machine Learning Models. Case study of Covid-19.***"

The project involves gathering relevant information from various institutions and he has informed the office that he would wish to carry his research in your organization.

We would be grateful if you could assist Mr. Onyango as she gathers data for his research. If you have any queries about the exercise please do not hesitate to contact us.

Yours sincerely

A handwritten signature in black ink, appearing to read 'ROO'.

**PROF. ROBERT O. OBOKO  
DIRECTOR  
SCHOOL OF COMPUTING AND INFORMATICS**

**School of Computing & Informatics  
University of NAIROBI  
P. O. Box 30197  
NAIROBI**

ROO/jsn

## **Appendix E - Interview Questionnaires Conducted at Various Bus Station in Nairobi**

### **Health and Mobility Research Questionnaire**

I am a student taking a master's in computational Intelligence at the University of Nairobi. I am seeking information from bus company employees and passengers about the health measures put in place during the journey.

Giving me just 3 minutes of your time can help. You could be providing the information needed to make positive changes in health. All the responses you provide will be anonymous, private, and confidential. I kindly ask you to answer as honestly as possible.

### **Inclusion Criteria**

Passengers traveling within the Kenyan Corridors.

### **BUS EMPLOYEE**

1. Which bus Sacco do you work for?
2. What is the origin and destination of your buses?
3. Which routes do your buses ply?
4. Do you have any stopping or dropping points along the way? Yes NO
5. Do you have passengers boarding and dropping on the way? Yes No
6. What is the full bus capacity? .....
7. What capacity do you carry during COVID-19?
8. What capacity did you use to carry before COVID-19?
9. How many trips do you make per day? One trip Two trips
10. What is the duration of the journey?
11. What is the general age range of your passengers?
  - Under 5
  - 5 – 20 years old
  - 21 – 40 years old
  - 41 – 50 years old
  - Above 50

- Not prefer to say
12. Which gender most prefers bus travel?
- Male
  - Female
  - Not prefer to say.
13. What is the reason for travel for most passengers?
- Business
  - Study
  - Tourist
  - Work
  - Prefer not to say.
14. On a scale of 0 -10, how many people do you approximate to travel for business purposes?
15. On a scale of 0 -10, how many people do you expect to travel for Leisure/ tour?
16. On a scale of 0 -10, how many people do you approximate to travel for work?
17. On a scale of 0 -10, how many people do you expect to travel to study?
18. Which health measures have been put in place to control the spread of diseases during the journey?
19. Have you ever experienced passengers getting sick on the way? Yes No
20. What are some of the actions you take when a passenger gets sick on the way?

**Thank you.**

## Health and Mobility Research Questionnaire

I am a student taking a master's in computational Intelligence at the University of Nairobi. I am seeking information from bus company employees and passengers about the health measures put in place during the journey.

Giving me just 3 minutes of your time can help. You could be providing the information needed to make positive changes in health. All the responses you provide will be anonymous, private, and confidential. I kindly ask you to answer as honestly as possible.

### Inclusion Criteria

Passengers traveling within the Kenyan Corridors.

### PASSENGERS QUESTIONNAIRE

1. Which bus company are you planning to travel with?.....
2. Are you traveling to which county? .....
3. Why do you prefer traveling via bus? .....
4. What is the reason for your journey? .....
5. Out of 10, how many passengers do you think travel for business purposes?  
.....
6. Out of 10, how many passengers do you think travel for work?  
.....
7. Out of 10, how many passengers do you think travel for leisure or tour?  
.....
8. Out of 10, how many passengers do you think travel for studies?  
.....
9. Out of 10, how many passengers do you think travel for other purposes?  
.....
10. How often do you travel by bus? .....
11. Have you ever encountered passengers boarding or dropping along the way? A) Yes  
b) No
12. Do you fear contacting COVID-19 during the journey?  
.....

13. What are some of the COVID-19 health measures do you take as a passenger?

.....

14. Have you ever fallen sick during the journey? a) Yes b) No

If yes, what did you do to manage the illness? .....

15. Have you ever encountered a passenger who falls sick during the journey?

.....

16. What did you do like passengers to manage the situation?

.....

17. How old are you?

- a. Under 18
- b. 19 - 35
- c. Over 50
- d. Prefer Not to say.

18. Are you traveling alone or with family?

- a. Alone
- a. With Family
- b. Prefer Not to say.

**Thank you.**

**Appendix F - Bus Volumes and frequencies (Research Feedback )**

Bus Company	Routes	Bus Capacity	Number of Passenger Before COVID-19	Number of Passenger During COVID-19	Frequency before COVID (Daily) to & from	Frequency during COVID (Daily)	COVID-19 Measures
Tahameed Coach	Tanzania Uganda Route A ; Nairobi - Nakuru - Kisumu - Busia - Kampala  Route B ; Nairobi - Mombasa - Lungalunga - Busia	49	49	28	3	1	Social Distancing, Sanitizing, Masks always, COVID-19 Certificate
Mash	Uganda (Kampala)  Route; Nairobi - Nakuru - Kisumu - Busia - Kampala	40	40	30	3	1	Social Distancing, Sanitizing, Masks always, COVID-19 Certificate, Yellow Fever
Dreamliner	Uganda (Kampala)	42	42	28	3		
Easy Coach	Uganda (Kampala) Corridor A ; Nairobi - Nakuru - Kericho - Kisumu- Busia - Kampala Corridor B;	38	38	23	4	0	Social Distancing, Sanitizing, Masks always, COVID-19 Certificate, Yellow Fever, Temperature check,



	Nairobi - Nakuru - Kapsabet - Kaimos - Mumia - Bungoma - Malaba -						No talking, No exchange of things during the journey, no snacks previously offered
Modern Coast	Uganda (Busia Border) Nairobi - Nakuru - Kisumu - Busia - Kampala  Tanzania (Manza) Nairobi - narok - Kisii - Isibania - Manza	45	45	24		1	Social Distancing, Sanitizing, Masks always, COVID-19 Certificate,
Bus Car	Kampala	40	40	20	3	1	
Dar Lux	Tanzania (Dar el Salaam)  Route; Nairobi -	53	53				
Dar Express	Tanzania (Dar el Salaam)	53	53		3	1	
<b>Total</b>							

**Appendix G - The total number of COVID-19 cases reported monthly in Kenya**

<b>Date Reported(Month)</b>	<b>New Monthly Cases</b>	<b>Cumulative Cases</b>
1/31/2020	0	0
2/29/2020	0	0
3/31/2020	50	50
4/30/2020	334	384
5/31/2020	1504	1888
6/30/2020	4302	6190
7/31/2020	13723	19913
8/31/2020	14144	34057
9/30/2020	4321	38378
10/31/2020	15419	53797
11/30/2020	29519	83316
12/31/2020	12935	96251
1/31/2021	4424	100675
2/28/2021	4973	105648
3/31/2021	26998	132646
4/30/2021	26175	158821
5/31/2021	11826	170647
6/30/2021	12956	183603
7/31/2021	18351	201954
8/31/2021	33344	235298
9/30/2021	13876	249174
10/31/2021	4119	253293
11/30/2021	1686	254979
12/31/2021	37258	292237
1/31/2022	29098	321335
2/28/2022	1595	322930
3/31/2022	472	323402
4/30/2022	358	323760
5/31/2022	1099	324859
6/30/2022	8431	333290
7/31/2022	4309	337599
8/31/2022	571	338170
9/30/2022	254	338424
10/31/2022	689	339113
11/30/2022	2460	341573
12/31/2022	897	342470
1/31/2023	347	342817
2/28/2023	102	342919
3/31/2023	57	342976
4/30/2023	84	343060

5/31/2023	252	343312
6/30/2023	474	343786
7/31/2023	132	343918
8/31/2023	37	343955

Data Source : <https://COVID19.who.int/WHO-COVID-19-global-data.csv>

## Appendix H - Truck-based person traffic volumes.

Date	Number_Of_Trucks_Mariakani	Number_Of_Trucks_Athi_River	Number_Of_Trucks_Gilgil	Number_Of_Trucks_Webuye	Number_Of_Trucks_Busia	Total_Trucks_Per_Month	Occupancy_Rate	Person_Traffic
01/01/2015								
01/02/2015								
01/03/2015	2484	5228	2474	962	502	11650	2	23300
01/04/2015	2412	4510	2293	809	433	10457	2	20914
01/05/2015	2608	4824	2392	916	405	11145	2	22290
01/06/2015	2535	4161	2390	940	401	10427	2	20854
01/07/2015	2659	4764	2563	935	428	11349	2	22698
01/08/2015	2705	4912	2363	1051	576	11607	2	23214
01/09/2015	2636	3433	1576	980	501	9126	2	18252
01/10/2015	2689	6017	2602	941	421	12670	2	25340
01/11/2015	2424	5749	2509	958	289	11929	2	23858
01/12/2015	2507	5421	2650	958	420	11956	2	23912
01/01/2016	2371	6349	2691	970	442	12823	2	25646
01/02/2016	2354	5325	2555	853	399	11486	2	22972
01/03/2016	2507	5583	2655	956	432	12133	2	24266
01/04/2016	2542	5466	2778	944	420	12150	2	24300
01/05/2016	2610	5632	2762	1013	422	12439	2	24878
01/06/2016	2445	5359	2271	967	470	11512	2	23024
01/07/2016	2739	5855	2625	958	454	12631	2	25262
01/08/2016	2327	5709	2805	957	471	12269	2	24538
01/09/2016	2620	5400	2769	977	436	12202	2	24404
01/10/2016	2687	5897	2998	1025	427	13034	2	26068
01/11/2016	2539	5683	3372	940	435	12969	2	25938
01/12/2016	2259	5100	4257	706	440	12762	2	25524
01/01/2017	4043	9619	4641	1671	446	20420	2	40840
01/02/2017	5312	9687	4297	1502	448	21246	2	42492
01/03/2017	2162	10861	2999	1670	516	18208	2	36416
01/04/2017	2374	5355	4462	1432	535	14158	2	28316
01/05/2017	2434	5471	4479	1304	592	14280	2	28560
01/06/2017	2191	5052	4396	1659	611	13909	2	27818
01/07/2017	2419	10540	4782	1587	646	19974	2	39948
01/08/2017	1927	7440	4198	1282	470	15317	2	30634
01/09/2017	2086	5880	4384	1709	584	14643	2	29286
01/10/2017	4885	3448	4689	1774	504	15300	2	30600
01/11/2017	4248	4674	4925	1817	552	16216	2	32432
01/12/2017	4128	4953	3195	1542	619	14437	2	28874
01/01/2018	2110	11755	6586	2575	697	23723	2	47446
01/02/2018	4973	10949	6186	2300	596	25004	2	50008
01/03/2018	4846	8698	4940	2426	650	21560	2	43120
01/04/2018	5085	10212	6376	2511	610	24794	2	49588
01/05/2018	4987	9868	6186	2627	659	24327	2	48654
01/06/2018	5319	9979	6586	2699	596	25179	2	50358
01/07/2018	4452	6973	6698	2636	680	21439	2	42878
01/08/2018	4572	9951	6456	2413	751	24143	2	48286
01/09/2018	7220	8153	6296	2520	703	24892	2	49784
01/10/2018	4932	10407		2669	854	18862	2	37724
01/11/2018	5154	10290		2439	791	18674	2	37348
01/12/2018	5272	9073		2701	794	17840	2	35680
01/01/2019	2445	9356	4479	2412	535	19227	2	38454
01/02/2019	1811	11789	4206	1664	406	19876	2	39752
01/03/2019	2093	9528		2519	422	14562	2	29124
01/04/2019	5329	12207	4434	2364	696	25030	2	50060
01/05/2019	5641	9654	4228	2334	694	22551	2	45102
01/06/2019	5540	10603	3923	2252	677	22995	2	45990
01/07/2019	2739	10709	4396	1444	611	19899	2	39798
01/08/2019	2327	5880	6697	1555	454	16913	2	33826
01/09/2019	2007	3548	6285	1545	646	14031	2	28062
01/10/2019	2687	10228	6537	1774	680	21906	2	43812
01/11/2019	2539	10805	6102	1613	576	21635	2	43270
01/12/2019	2259	11755	6102	1817	471	22404	2	44808

**Appendix I - Truck-based person traffic volumes exponential smoothing.**

Date	Person_Truck_Traffic	Forecast Linear	ETS Forecast	Confidence	Upper Bound	Lower Bound
1/1/2015						
2/1/2015						
3/1/2015	699000					
4/1/2015	627420					
5/1/2015	668700					
6/1/2015	625620					
7/1/2015	680940					
8/1/2015	696420					
9/1/2015	547560					
10/1/2015	760200					
11/1/2015	715740					
12/1/2015	717360					
1/1/2016	769380					
2/1/2016	689160					
3/1/2016	727980					
4/1/2016	729000					
5/1/2016	746340					
6/1/2016	690720					
7/1/2016	757860					
8/1/2016	736140					
9/1/2016	732120					
10/1/2016	782040					
11/1/2016	778140					
12/1/2016	765720					
1/1/2017	1225200					
2/1/2017	1274760					
3/1/2017	1092480					
4/1/2017	849480					
5/1/2017	856800					
6/1/2017	834540					
7/1/2017	1198440					
8/1/2017	919020					
9/1/2017	878580					
10/1/2017	918000					
11/1/2017	972960					
12/1/2017	866220					
1/1/2018	1423380					
2/1/2018	1500240					
3/1/2018	1293600					
4/1/2018	1487640					
5/1/2018	1459620					
6/1/2018	1510740					
7/1/2018	1286340					
8/1/2018	1448580					
9/1/2018	1493520					
10/1/2018	1131720					
11/1/2018	1120440					

12/1/2018	1070400					
1/1/2019	1153620					
2/1/2019	1192560					
3/1/2019	873720					
4/1/2019	1501800					
5/1/2019	1353060					
6/1/2019	1379700					
7/1/2019	1193940					
8/1/2019	1014780					
9/1/2019	841860					
10/1/2019	1314360					
11/1/2019	1298100					
12/1/2019	1344240					
1/1/2020	1060680					
2/1/2020	1012320					
3/1/2020	780960					
4/1/2020	785160					
5/1/2020	1103100					
6/1/2020	1437780					
7/1/2020	1394880					
8/1/2020	1249440					
9/1/2020	1142940					
10/1/2020	1149780					
11/1/2020	1052940					

12/1/2020	1182600					
1/1/2021	1275720					
2/1/2021	1297440					
3/1/2021	1271460					
4/1/2021	1332780					
5/1/2021	1311540					
6/1/2021	1260120					
7/1/2021	1257420					
8/1/2021	1295040					
9/1/2021	1276740					
10/1/2021	1071180					
11/1/2021	1081800					
12/1/2021	1114680					
1/1/2022		1230326.918	1110898.797	371256.162	1482154.959	739642.6353
2/1/2022		1235562.045	1112507.883	468631.753	1581139.636	643876.13
3/1/2022		1243162.152	1115586.057	543836.788	1659422.845	571749.2686
4/1/2022		1244244.506	1047392.018	547665.597	1595057.615	499726.4203
5/1/2022		1233938.409	1112666.056	474751.487	1587417.543	637914.5685
6/1/2022		1221633.168	1096041.397	534452.005	1630493.402	561589.392
7/1/2022		1205504.635	1109917.492	782538.972	1892456.464	327378.52
8/1/2022		1206210.496	1118346.182	841519.527	1959865.709	276826.6548
9/1/2022		1190548.747	1148833.239	779690.193	1928523.432	369143.0461
10/1/2022		1169383.846	1333992.579	594306.208	1928298.787	739686.3713
11/1/2022		1146906.797	1097832.959	617261.633	1715094.593	480571.3261

12/1/2022		1124488.298	1167407.865	639528.523	1806936.387	527879.3417
1/1/2023		1090068.288	1063579.721	615255.76	1678835.482	448323.9607
2/1/2023		1095091.85	1043448.704	607368.494	1650817.198	436080.21
3/1/2023		1109100.077	992615.0625	804628.736	1797243.798	187986.3269
4/1/2023		1107326.02	1068718.242	645926.918	1714645.16	422791.3234
5/1/2023		1125408.574	1411294.598	706310.084	2117604.683	704984.5141
6/1/2023		1144925.398	1275246.177	740189.067	2015435.244	535057.1107
7/1/2023		1174906.715	1216969.297	736641.47	1953610.767	480327.8272
8/1/2023		1184146.757	1239952.966	757938.174	1997891.139	482014.7923
9/1/2023		1216687.114	1060019.352	451804.283	1511823.636	608215.0691
10/1/2023		1262364.625	1288179.89	377132.441	1665312.331	911047.4493
11/1/2023		1263125.166	1240944.302	490595.334	1731539.636	750348.9683
12/1/2023		1261378.538	1474910.331	472489.799	1947400.13	1002420.532
1/1/2024		1249799.981	1509922.824	496867.906	2006790.73	1013054.917

**Appendix J - Truck Traffic person volume and bus monthly data and total number of COVID-19 Cases reported**

Date	Number of People Travelling buy bus	Person Truck Traffic Monthly	Spread of COVID-19
1/31/2015	416334.56		
2/28/2015	416334.56		
3/31/2015	416334.56	699000	
4/30/2015	416334.56	627420	
5/31/2015	416334.56	668700	
6/30/2015	416334.56	625620	
7/31/2015	416334.56	680940	
8/31/2015	416334.56	696420	
9/30/2015	416334.56	547560	
10/31/2015	416334.56	760200	
11/30/2015	416334.56	715740	
12/31/2015	416334.56	717360	
1/31/2016	425704.31	769380	
2/29/2016	425704.31	689160	
3/31/2016	425704.31	727980	
4/30/2016	425704.31	729000	
5/31/2016	425704.31	746340	
6/30/2016	425704.31	690720	
7/31/2016	425704.31	757860	
8/31/2016	425704.31	736140	
9/30/2016	425704.31	732120	
10/31/2016	425704.31	782040	
11/30/2016	425704.31	778140	
12/31/2016	425704.31	765720	
1/31/2017	435280.48	1225200	
2/28/2017	435280.48	1274760	
3/31/2017	435280.48	1092480	
4/30/2017	435280.48	849480	
5/31/2017	435280.48	856800	
6/30/2017	435280.48	834540	
7/31/2017	435280.48	1198440	
8/31/2017	435280.48	919020	
9/30/2017	435280.48	878580	
10/31/2017	435280.48	918000	
11/30/2017	435280.48	972960	
12/31/2017	435280.48	866220	
1/31/2018	444313.38	1423380	
2/28/2018	444313.38	1500240	



3/31/2018	444313.38	1293600	
4/30/2018	444313.38	1487640	
5/31/2018	444313.38	1459620	
6/30/2018	444313.38	1510740	
7/31/2018	444313.38	1286340	
8/31/2018	444313.38	1448580	
9/30/2018	444313.38	1493520	
10/31/2018	444313.38	1131720	
11/30/2018	444313.38	1120440	
12/31/2018	444313.38	1070400	
1/31/2019	453279.24	1153620	
2/28/2019	453279.24	1192560	
3/31/2019	453279.24	873720	
4/30/2019	453279.24	1501800	
5/31/2019	453279.24	1353060	
6/30/2019	453279.24	1379700	
7/31/2019	453279.24	1193940	
8/31/2019	453279.24	1014780	
9/30/2019	453279.24	841860	
10/31/2019	453279.24	1314360	
11/30/2019	453279.24	1298100	
12/31/2019	453279.24	1344240	
1/31/2020	462390.15	1060680	
2/29/2020	462390.15	1012320	
3/31/2020	462390.15	780960	50
4/30/2020	462390.15	785160	384
5/31/2020	462390.15	1103100	1888
6/30/2020	462390.15	1437780	6190
7/31/2020	462390.15	1394880	19913
8/31/2020	462390.15	1249440	34057
9/30/2020	462390.15	1142940	38378
10/31/2020	462390.15	1149780	53797
11/30/2020	462390.15	1052940	83316
12/31/2020	462390.15	1182600	96251
1/31/2021	471374.39	1275720	100675
2/28/2021	471374.39	1297440	105648
3/31/2021	471374.39	1271460	132646
4/30/2021	471374.39	1332780	158821
5/31/2021	471374.39	1311540	170647
6/30/2021	471374.39	1260120	183603
7/31/2021	471374.39	1257420	201954
8/31/2021	471374.39	1295040	235298
9/30/2021	471374.39	1276740	249174

10/31/2021	471374.39	1071180	253293
11/30/2021	471374.39	1081800	254979
12/31/2021	471374.39	1114680	292237
1/31/2022	480377.64		321335
2/28/2022	480377.64		322930
3/31/2022	480377.64		323402
4/30/2022	480377.64		323760
5/31/2022	480377.64		324859
6/30/2022	480377.64		333290
7/31/2022	480377.64		337599
8/31/2022	480377.64		338170
9/30/2022	480377.64		338424
10/31/2021	480377.64		339113
11/30/2021	480377.64		341573
12/31/2021	480377.64		342470
1/31/2023			342817
2/28/2023			342919
3/31/2023			342976
4/30/2023			343060
5/31/2023			343312
6/30/2023			343786
7/31/2023			343918
8/31/2023			343955