UNIVERSITY OF NAIROBI

DEPARTMENT OF COMPUTER SCIENCE

FACULTY OF SCIENCE AND TECHNOLOGY

NEURAL NETWORK - BASED PREDICTIVE ANALYTICS OF COVID-19 IN THE HEALTH SECTOR FOR DECISION SUPPORT IN KENYA

BY

MWAURA JUDITH WAIRIMU

THIS PROJECT REPORT IS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF MASTER OF SCIENCE IN COMPUTATIONAL INTELLIGENCE OF THE UNIVERSITY OF NAIROBI

JUNE 2022

DECLARATION

Researcher's Declaration

This project report is my original work and has not been presented in any other institution for the purpose of any academic award.

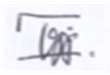
SIGNATURE:  Date: 24/06/2022

Mwaura Judith Wairimu

Registration Number: P52/38827/2020

Supervisor's Approval

This project report has been submitted in partial fulfilment of the requirements for the Degree of Master of Science in Computational Intelligence of the University of Nairobi with my approval as the University Supervisor.

SIGNATURE:  Date: 24/06/2022

Prof. Elisha T. O. Opiyo

Faculty of Science and Technology

University of Nairobi

DEDICATION

This work is dedicated to my son, Ethan James Mwaura.

ACKNOWLEDGEMENT

I wish to thank the Almighty God for giving me strength and guidance in this project. Special thanks to my supervisor Professor Opiyo and the University of Nairobi staff especially those at the Faculty of Science and Technology for their great support and good relation all through the period as a student at this university. The contribution from my mother, Mary Wairimu, my father Peter Mwaura, my uncle Samuel Kariuki, and friends towards this project is recognized and highly appreciated. Thank you all.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

PHEIC	Public Health Emergency of International Concern
MERS	Middle East Respiratory Syndrome
SARS	Severe Acute Respiratory Syndrome Coronavirus
WHO	World Health Organisation
CDC	Centres for Disease Control and Prevention
LMIC	Low-to-Middle-Income Countries
PPE	Personal Protective Equipment
HDU	High Dependency Units
SII	Serum Institute of India
IPC	Infection Prevention and Control
HDFS	Hadoop Distributed File System
MIT	Massachusetts Institute of Technology
IEEE	Institute of Electrical and Electronics Engineers
D3	Data-Driven Document
MLP	Multilayer Perceptron
LR	Logistic Regression
KNN	K-Nearest Neighbour
DT	Decision Trees
ANN	Artificial Neural Networks
MLPNN	Multilayer Perceptron Neural Network
PPA	Prey-Predator Algorithm
DSV	Data Summarization and Visualization
HTML	Hypertext Markup Language
PC	Personal Computer
FFNN	Feed-Forward Neural Network
CSS	Cascading Style Sheets
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
ARIMA	AutoRegressive Integrated Moving Average
XAMPP	Cross-platform, Apache, MySQl, PHP and Perl
URL	Uniform Resource Locator

DEFINITION OF TERMS

Model – in machine learning, a model refers to the conceptual structure developed once an algorithm is fed with sample input and corresponding output data for training.

Adoption – means to choose as a standard or requirement in a course.

Visualization – a technique for creating diagrams, animations, or images with an aim of communicating a message.

Predictive Analytics – is a technology that is used to predict future events / data based on present and past data.

Incubation Period – the period between when one becomes infected with something and when the symptoms start to show.

Latent Period – the period between when an individual becomes infected with something and when he or she becomes infectious

Tipping point – the point beyond which a situation or process experiences an unavoidable effect or change.

ABSTRACT

COVID-19 spread rapidly across the globe and claimed many lives. As a result, it was declared a pandemic. Kenya, like most countries was negatively affected by the disease. In the face of COVID-19, predictive analytics plays a crucial role by providing the opportunity to proactively combat and curb the spread of the virus. This can be achieved by ensuring that hospitals are equipped with enough health resources such as ICU beds and ventilators. Statistical models and regression have been adopted in recent studies to try and predict the spread of the virus. Visualization dashboards have also been created, but they do not inform decision making. In this study, the development of a predictive system will aid in forecasting the spread of the virus. There is currently adequate data on COVID-19 cases since its onset. Visual representations of data on the other hand, are always easy to comprehend. This study is therefore focused on using neural networks as a machine learning algorithm to create a predictive model. Open-source data on COVID-19 confirmed cases in Kenya was collected, pre-processed, and used to train the model. MSE, MAE and RMSE were used as performance metrics to test the model which was then used to predict the expected number of COVID-19 cases within the next 60 days. These predictions were visualized onto Tableau dashboard against the COVID-19 infection cases and Health Resources available to contain the spread of the virus, hence providing an avenue for health resource planning and decision making. The study adopted LSTM neural network algorithm for univariate time series forecasting. Univariate time-series forecasting proved to be the best approach since the virus is quite unpredictable due to the low correlation between the multiple parameters and variables associated with it. The study also established that, merging the powerful aspects of visualization and predictive analytics enhances decision support. The system was also presented to the users upon completion and an acceptance response of 95% was received.

Keywords:

COVID-19 pandemic, Neural Networks, Predictive Analytics, Visualization, Decision Support, Healthcare resource management.

CHAPTER ONE: INTRODUCTION

1.1 Background

The first case of COVID-19 outbreak was reported in Wuhan, China, on December 31, 2019. On January 13, 2020, Thailand reported the first incident outside of China. At the end of January 2020, the World Health Organization consequently declared the outbreak of COVID-19 as a Public Health Emergency of International Concern (PHEIC). A month later, there were over 76,000 confirmed cases of COVID-19 globally. The World Health Organization subsequently declared COVID-19 a pandemic on March 11, 2020 (Hui et al., 2020). In Kenya, the Ministry of Health gave an account of the first COVID-19 case on March 13, 2020.

The outbreak of an infectious disease refers to its occurrence within a geographical area or community when it is not expected (Toppenberg-Pejcic et al., 2019). Infectious diseases are characterised by their rapid spread which endangers the health of many people, hence necessitating immediate action to prevent the disease spread at the community level (Lin et al., 2016). COVID-19 disease is caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-Cov-2), initially referred to as the 2019 novel Coronavirus (2019-nCoV). Along with the Middle East Respiratory Syndrome (MERS), the virus is the seventh member of the coronavirus family that can infect humans. Fever, coughing, shortness of breath, and diarrhoea are all signs of COVID-19. COVID-19 can predispose an individual to pneumonia and even cause death in severe cases. The incubation period of the disease can last up to two weeks and is infectious during the latent period and can be transmitted from one person to another through respiratory droplets and close contact (Hui et al., 2020).

Most countries implemented both containment and mitigation measures as response strategies to the virus. The main aim being to delay major surges of patients, reduce the demand for hospital beds and protect the individuals who face a high risk of infection such as the elderly as well as those with comorbidities. Activities to accomplish the goal depend on the country's risk assessment. National response strategies have been adopted in the face of the pandemic. These include self-isolation or quarantine, contact tracing, banning of public meetings on a large-scale, promoting public health measures such as handwashing, wearing of surgical masks and social distancing. In most countries, the healthcare systems have also been prepared for a surge of patients in critical conditions who require isolated rooms with oxygen, and mechanical ventilators (Bedford et al., 2020).

(Barasa et al., 2020) conducted a study to assess how many cases the health sector, both public and private hospitals, can handle during the pandemic. They assumed that, based on modelled estimates for Kenya, two percent of the population become symptomatic with SARS-Cov-2. The health system’s surge capacity was determined by taking into consideration three transmission curve scenarios: six, twelve, and eighteen months. Four main measures of hospital surge capacity were calculated country-wise and within the 47-county governments: 1) Surge capacity for ICU beds 2) Surge capacity for hospital beds 3) The ICU bed tipping point, and 4) The hospital bed tipping point. The study revealed that ICU beds, ventilators and oxygen inhibit the Kenyan hospital’s ability to deal with the increasing COVID-19 cases. It also revealed that, the availability of these resources in the face of COVID-19 vary across the Kenyan counties indicating discrepancies in access to healthcare resources for the COVID-19 patients.

In global health, more data does not always imply more evidence-based programs and policies. There is a paucity of knowledge on how to facilitate data-driven decision-making successfully.Therefore, there is a paramount need to present data in a form that can be easily understood and interpreted (Rodríguez et al., 2017). Adequate data related to COVID-19 is now available from the infection, recovery, and vaccination cases. This data makes other data such as resources available in controlling and suppressing the spread of the virus for instance the number of ventilators and ICU beds available, crucial.

Techniques for data visualization are gradually being adopted in global health. Dashboards have been adopted by Low-to-Middle-Income Countries (LMIC) and by development partners to identify factors of great interest to the countries. Effective data tracking, analysis and visualization helps to capture and represent data for decision-making purposes. It is possible to accurately interpret data and use it wisely by adopting a well-designed data visualization dashboard (Aung et al., 2019). Predicting the spread of the pandemic is very crucial. It aids in understanding the reasoning behind important national decisions and ensures efficient resource and utility management. (Ghosh, 2020). Predicting the trends in the spread of the virus versus the potential healthcare resource use under various scenarios, aids in preparing the healthcare system for an increase or decrease in COVID-19 patients through proper resource planning.

1.2 Statement of the Problem

COVID-19 has gained its popularity through the high infection and fatality rates. As a health concern, it brings along the need to be prepared in terms of the availability of health resources. Failing to understand the fluctuations in the number of infection cases during different periods makes it difficult to plan for health resources. A lack of readiness to combat the spread of the virus will lead to many fatalities which adversely affects the country's economy. Therefore, it is very crucial to ensure that the patterns and trends in the spread of the virus are well forecasted. Machine learning based algorithms for COVID-19 prediction have been developed in other nations. However, the models are specifically designed to suit the needs of those countries and cannot be readily replicated to model in the Kenyan setting. Machine Learning algorithms, especially neural networks require large amounts of data to train (Kumar, 2017). However, at the onset of the epidemic, there was little historical COVID-19 data that could be used to train the models. This made it difficult for researchers and scholars to predict the trends in the spread of the virus (Feng et al., 2021). Since the onset of COVID-19 pandemic, the daily recorded cases has risen hence providing adequate amounts of data. This therefore poses a requirement for machine learning models that can reliably predict the spread of the virus.

In Kenya, we have statistical models developed to aid in forecasting the spread of the virus. However, according to Shinde et al (2020), statistical models are not suited for large amounts of data. Hence, with the increasing amounts of COVID-19 data, deep learning models are more preferred since they provide better optimization. Current visualization dashboards purely convert the numeric representations of COVID-19 cases to a visual format. However, the visualized data cannot be comprehensively relied upon to inform decision-making. Kenya has also previously faced a scenario whereby, the COVID-19 cases rose drastically, yet the health resources available were not adequate to take care of all patients. Therefore, there is need to adopt machine learning techniques (Neural Networks) to carry out predictive analytics and visualize the forecasts alongside the health care resources (ventilators and ICU beds). This provides a platform for decision support that can be used in making informed data-driven decisions.

1.3 General Objective

The general objective of this study is to use Artificial Neural Networks (ANN) in carrying out predictive analytics to aid decision related issues in healthcare resource planning, with the aim of containing and mitigating the spread of the virus.

1.4 Objectives

1. To review and evaluate the COVID-19 predictive systems currently in use.
2. To develop a predictive machine learning model using neural networks to forecast the spread of the COVID-19 virus.
3. To design and develop a visualization dashboard for the predicted data, COVID-19 cases and health resources set aside to contain the virus.
4. To evaluate the performance of the model and functionality of the dashboard.

1.5 Research Questions

1. How can you determine the best algorithm to carry out prediction?
2. How can one forecast the spread of COVID-19 virus?
3. Which is the best way to represent data in a form that can be easily understood and interpreted?
4. How can you determine if the visualized data is accurate based on the predictions?

1.6 Significance of the Study

In this day and age of big data, data visualization and predictive analytics have been widely adopted. Large corporations, such as Amazon and Apple, rely heavily on data visualization and predictive analytics. The two technologies are important in simplifying and justifying business decisions, as well as in maximizing operational efforts. The healthcare sector has also adopted the creation of predictive models using machine learning algorithms to forecast the nature of diseases against predisposed individuals and to predict diseases / conditions based on symptoms presented by the patients. The predictions are always accurate and efficient if the models are well trained on vast and different data types. This study seeks to create a predictive machine learning model using deep learning, that is trained using COVID-19 data, to predict the trends in the spread of the virus. The forecasted data is visualized against the COVID-19 cases and health resources available. This aids in determining whether the cases are expected to decrease or increase and whether the resources available are enough to combat the spread. Being able to understand when the infection rates are expected to be higher will enhance decision support towards the health sector, both public and private, the Government and the Ministry of Health. It enables them to prepare for a surge by carrying out health care resource planning. A forecast showing an increment in the infection cases for example, requires additional resources to be procured, hence proactively combating the spread of the virus. Having such information province-wise in a country like Kenya helps to identify the risky zones that should be avoided and enables the Government to set across measures that limit movement into and out of the risky zones, hence containing the spread of the virus.

1.7 Scope of the Study

The study focused on building a predictive model using Neural Networks with the aim of containing and mitigating COVID-19 spread. The research analysed the COVID-19 cases within Kenya against the resources available in terms of ICU beds and ventilators. It therefore seeks to predict the spread of the virus across the Kenyan provinces to enhance decision making in terms of resource planning.

CHAPTER TWO: LITERATURE REVIEW

2.1 COVID-19 Situation in the World

COVID-19 started in Wuhan, China in December 2019 as an illness with symptoms similar to those of pneumonia and rapidly spread worldwide (Hui et al., 2020). The virus was proclaimed a pandemic in March 2020, and with the rapid increase in the infection and death cases to parts of Asia and Europe at the time, COVID-19 was declared a pandemic that must be stopped. According to the Joint Mission Team of China and the World Health Organization, the pandemic epidemic spread swiftly between January 10 and 12, 2020. The cases peaked and plateaued by late-January 2020 following the rigorous containment and quarantine efforts in the country (Rabi et al., 2020).

The different factors that influence the spread of the virus must be immediately addressed to flatten the epidemiological curve. It is crucial for the disease to be detected early by carrying out random tests on individuals, isolating those infected, and preparing the health sector to handle the rapidly rising infection cases. Despite rigorous mitigation efforts, the increasing prevalence and severity of the disease inundated even the most developed countries' healthcare resources. Healthcare systems in countries like Italy, the United Kingdom (UK), and the United States (USA) were severely strained. This is owing to a dearth of medical professionals, a paucity of mechanical ventilators in intensive care units, and a shortage of personal protective equipment (PPE), all of which jeopardize public health outcomes (Cavallo et al., 2020).

Viruses are continually changing owing to mutation, according to the Centres for Disease Control and Prevention, and new virus variations are likely to emerge. The World Health Organization publicized a new Greek alphabet-based system for naming SARS-COV-2 variants to prevent the reference of the variants based on the countries they were initially detected. The variants have widely spread to most parts of Canada. B.1.1.7 (Alpha) was first discovered in the United States in June 2020, but its origin is the United Kingdom. It greatly contributed to the third wave in Canada at the start of 2021. Alpha is 50% more contagious compared to earlier strains. In May 2021, the B.1.351 (Beta) variant was first detected in South Africa. This was at a time when the country was experiencing the second wave and it led to increased hospitalizations and deaths. The P.1 (Gamma) variant was discovered in Brazil in November 2020, while the B.1.617.2 (Delta) variant was discovered in India in October 2020. Delta variant is highly transmissible and is widely spreading in Canada compared to Alpha (Duong, 2021). In comparison to other disease variants, COVID-19 variants are highly contagious. Increased caseloads will lead to a strain in healthcare resources, causing more hospitalizations and deaths.

The United States recorded the highest number of COVID-19 fatalities in North America, with 110220 deaths according to the World Health Organisation as of June 2020. Mexico came in second with 13 699, and Canada came in third with 7800. Brazil had the most fatalities in South America, with 36455 deaths, followed by Peru with 5465 and Chile with 2264. South Africa recorded the most fatalities on the African continent, with 1080, followed by Algeria (715), and Nigeria (361). The countries with the highest recorded fatalities in Europe were the United Kingdom (40597 deaths), Italy (33964 deaths), and France (29149 deaths). Iran had the most fatalities in Asia, with 5957, China followed by 4643, then India (1074). Australia had the most number of death cases in their Oceania region, with 102, New Zealand followed with 22 and Guam with 5 fatalities (Sanyaolu et al., 2021).

In June 2021, according to WHO, the total number of deaths per week continued to decline. COVID-19 cases remain high with over 370000 newly reported cases daily across the globe. The aggregate number of cases reported across the globe was more than 180 million and the death cases was close to 4 million. In the same month, the African region recorded a 33% increase in incident cases and 42% increase in death cases, which was very high. The Eastern Mediterranean and European regions also observed an increase in the total number of cases reported for the month. Except for the African region, all other regions reported a decrease in the number of deaths in the previous week (Sanyaolu et al., 2021).

2.2 COVID-19 Situation in Kenya

The Kenyan Ministry of Health confirmed its first COVID-19 infection case on March 13, 2020. On April 6, 2020, the Kenyan Government closed international borders, places of worship, restaurants, bars, and nightclubs as a proactive measure against the spread of COVID-19. Meetings, social gatherings were banned all over the country and a dusk to dawn curfew imposed. Movement restrictions were imposed in Nairobi and Mombasa, which were at the time considered as hotspot zones. Individuals who tested positive and those exposed were quarantined at designated facilities. The Ministry of Health also enforced precautionary measures such as maintaining social distance, wearing of personal protective equipment such as surgical masks, washing of hands frequently or sanitizing and isolation of those who have tested positive, to help in containing and mitigating the spread of the virus (Muhula et al., 2021).

One of the main concerns from the unmitigated spread was the Kenyan health system's limited surge capacity (Barasa et al., 2020). The group of individuals among the Kenyan population regarded as highly vulnerable to the virus was another major concern. By Polymerase Chain Reaction (PCR), only a few Kenyans tested positive for SARS-CoV-2 into June 2020 or fell severely ill or died with COVID-19 as the established cause. In a day, July 24, 2020, Kenya recorded 796 new cases. This was the highest single-day cases ever reported. The high infection rate among private health-care workers is attributed to a lack of adherence to Infection Prevention and Control (IPC) guidelines and an insufficient supply of personal protective equipment (PPEs) (Macharia et al., 2020).

On March 3, 2021, Kenya received the first batch of AstraZeneca Vaccine manufactured by the Serum Institute of India (SII). The elderly (58 years and above), teachers, health workers and security officers were prioritized in the vaccination. Total doses administered amounted to 1477916 which accounted for 45% females and 55% males. A recent statistic by the Ministry of Health released on July 1, 2021, indicated that 1127 patients had been admitted in various health facilities countrywide. There were 5722 patients in the Home Based Isolation and Care program, 130 in the ICU, 39 on ventilatory support, 53 on supplemental oxygen, and 38 under observation.168 patients were on separate supplemental oxygen, 153 in general wards and 15 in High Dependency Units (HDU) (Macharia et al., 2020).

2.3 COVID-19 Situation in Kenyan Counties

The first COVID-19 cases, according to the Kenyan Ministry of Health, were reported in Nairobi and Mombasa counties. The pattern has recently changed since June 17, 2021, when the Government of Kenya made several changes to the measures against the COVID-19 pandemic. 13 counties in the Nyanza, Western and Rift Valley regions of the nation were identified as hotspot zones for the pandemic. Movement in and out of the 13 counties was restricted and the curfew within the counties set from 7pm to 4am daily.

2.4 Data Analytics

The process of analytics aids in extracting patterns that were hidden and unknown as well as useful unknown relationships. The advanced analytical processes that are widely used are association rules, clustering, and regression. Different analytical approaches are chosen based on the different types of data. Below are the different types of Data Analytics used today.

2.4.1 Predictive Analytics

Predictive analytics is used in determining patterns in previous data and providing the possible solutions in each situation. Predictive analysis examines both current and historical data to forecast what is likely to happen in the future and provides the probabilities. It is used in Big Data in forecasting unavailable data. It is the most widely used analytical method for data on sales leads, social media, and customer relationship management (Pritee Chunarkar Patil, 2018).

2.4.2 Prescriptive Analytics

Prescriptive analytics seeks to identify actions and suggests next steps. It provides a focused response to the situation. Prescriptive data analysis goes beyond predictive analytics as it aims at providing multiple actions and the corresponding likely outcomes. Organizations do not favour this type of analytics, but the data it produces can be impressive if used correctly (Pritee Chunarkar Patil, 2018).

2.4.3 Diagnostic Analytics

Diagnostic analysis examines past data to determine why something occurred. It uncovers hidden patterns that are essential in the root cause determination and identification of any factors that directly or indirectly lead to the result. Diagnostic analysis is commonly used in social media to analyse the number of posts, shares, and so on (Pritee Chunarkar Patil, 2018).

2.4.4 Descriptive Analytics

Descriptive analysis is also referred to as data mining. It is responsible for converting large data sets to small bytes of data and operates in real time. It determines what happened or what is happening. Descriptive analysis is used to uncover patterns and relationships and is mostly characterised by visualizations (Pritee Chunarkar Patil, 2018).

2.5 Data Visualization

Visualization comes from the Latin word “visualis” which refers to the visual results of well observed and analysed physical occurrences or digital information or physical events that are imagined, observed, and presented. Visualization, like volume, velocity, variety, value, and veracity, has emerged as a key characteristic of big data. Visualization has been a main topic of discussion in conferences, symposiums and workshops that have been conducted by academic institutions on various aspects of information processing. The Institute of Electrical and Electronics Engineers (IEEE) hosted the first visualization conference in 1990 (Mashkoor & Ahamad, 2017).

Visual transmission of information is very ancient and goes way back to the ancient carving of stones, Egyptian hieroglyphs, Greek geometry, revolutionary painting techniques by Leonardo da Vinci and technical drawings used in engineering (Institute of Information Technology of Azerbaijan National Academy of Sciences et al., 2018). An American psychologist, (Friendly, 2008) structures the history of data visualization into eight stages. Until the seventeenth century, we had early maps and diagrams. Measurements and theories came in between

1600 and 1699 and thereafter, the new graphic forms followed between 1700 and 1799. The fourth stage was 1800-1850 where the modern graphics began. The golden age of statistical graphics followed between 180 and 1900, then thereafter came the dark ages, 1900-1950. Between 1950 and 1975, data visualization experienced a renaissance. From 1975 to date, we have interactive and dynamic data visualization.

(Gokhale & Mahajan, 2020) further explains data visualization technology as a combination of data visualization tools and techniques. While data visualization tools deal with what data is to be visualized and how that data is to be processed further for visualization, data visualization techniques deal with how that processed data is to be visualized and in what format. The visualization tool lets the user choose which data to convert into a graphic while the technique allows the user to choose the chart types and how the data will be displayed. Data presentation and data exploration are two important goals of data visualization technology. Data presentation is concerned with merely displaying facts and figures that the user is already familiar with while data exploration is concerned with uncovering hidden insights from large datasets.

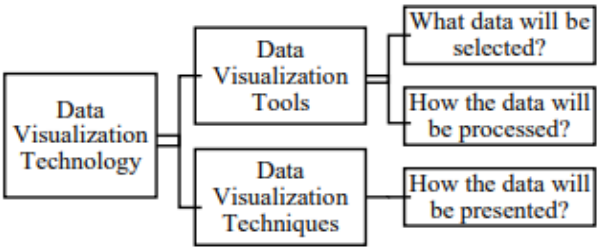


Figure 2.1: Data Visualization Technology integrates Tools and Techniques
Source: (Gokhale & Mahajan, 2020)

2.5.1 Data Visualization Tools

Digital data visualization tools provide a fantastic opportunity to investigate the value of digital big data. Big data visualization provides numerous benefits, including improved business decision making, new insights, new business opportunities, and increased profitability. (Gokhale & Mahajan, 2020). Different visualization software differ in terms of license; whether they are open source or not, whether they are structured as a dynamic web-tool that can be updated in real-time or whether they are designed as a post-processing software, the presence or absence of analytics, and the types of connectors used to input data files (Petrovich, 2020).

2.5.1.1 Tableau

Tableau is one of the best existing data visualization tools, with a professional platform used by many companies and industries. It is highly developed and specialized in business intelligence, with a user-friendly interactive interface that is also suitable for machine learning and Big Data. It is also used by the Italian ISTAT13, among others. Tableau enables extremely detailed data descriptions, as well as the use of analytics and server solutions without the need for on-site installation. It works well with MySQL, Hadoop, AWS, SAP, and Teradata. Tableau also has the capability of integrating machine learning (Kemal, 2019). An example of an application is provided below:

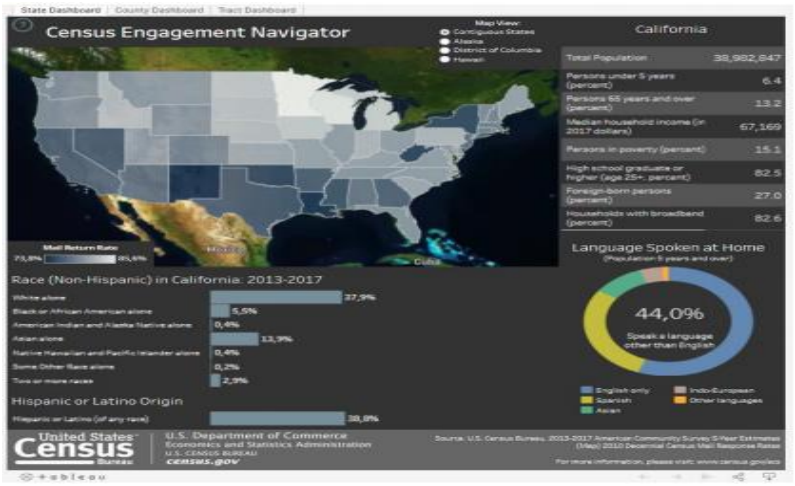


Figure 2.2: A dashboard with Tableau

Source: (Kemal, 2019)

2.5.1.2 D3.js

D3 is an abbreviation for Data-Driven Document, which combines extremely powerful visualization components with data-driven manipulation methods DOM (Document Object Model). D3 is a JavaScript library that adheres to standards of the web, allowing one to take advantage of all the capabilities of modern browsers without being any confinement to a specific proprietary framework. It makes use of JavaScript, HTML, SVG, and CSS. It is a highly technical and developer-oriented tool with a large gallery containing hundreds of charts, maps, and interactive and innovative diagrams that come with examples and are ready to be reused in a variety of contexts (Bostock et al., 2011).

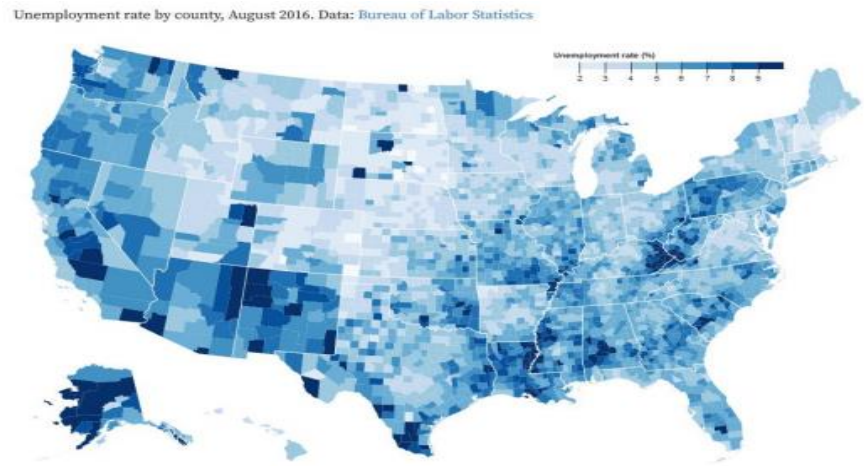


Figure 2.3: A visualization with D3.js

Source: (Bostock et al., 2011)

2.5.1.3 Dash-Plotly

Dash is a Python-based web application framework. It's based on Flask, Plotly.js, and React.js. It is intended for the development of data visualization apps, particularly those with interactive graphics, flexible and customizable interfaces. It is especially suited to those who work with data in Python. Dash is a very simple application to use, as it already includes the technologies and protocols needed to build an interactive web application. It is an open-

source library distributed under the Massachusetts Institute of Technology (MIT) license and includes a dashboard design (Petrovich, 2020). Below is an example.

Figure 2.4: A Visualization with Dash-Plotly

Source: (Petrovich, 2020)

2.5.1.4 Grafana

Grafana is a free and open-source software designed to serve as a "monitoring tool" for internet infrastructures and software analysis. Its applications have expanded to include industrial sensor monitoring, home automation, weather applications, and process control. It can communicate with databases such as Graphite, Elasticsearch, PostgreSQL, Cloudwatch, Prometheus, InfluxDB, and others. Grafana is a software that is quick and easy to use, particularly for the creation of dashboards. It can take advantage of SQL's high performance for data queries. However, it has a limited number of charts and customization options (Petrovich, 2020). An example of an application is depicted in the figure below:

Figure 2.5: A Visualization with Grafana

Source: (Petrovich, 2020)

2.6 Predictive Analytics

Predictive Analytics has greatly transformed and changed organizations' expectations and business strategies in the past century. The importance of predictive analysis stems from its ability to recommend the best future planning by combining data on who, what, where, and when to analyse why and how. Predictive analytics is the process of predicting what will happen in the future by using historical data, machine learning, and artificial intelligence in the organization (Henrys, 2021). Predictive analytics uses analytical techniques and historical data to forecast future outcomes. The analytical techniques include machine learning and statistical modelling. According to Hair (2007), predictive analytics forecasts future outcomes based on previously established relationships between explanatory and criterion variables. The goals of predictive analytics are to generate relevant and actionable information, deliver better results, make more intelligent decisions, and predict future events by analyzing large amounts of data for volume, veracity, speed, variety, and value. The prediction's accuracy is dependent on measuring the correct variable. There are numerous methods for developing prediction models, with each suiting a specific scenario and needs. In general, more data and a simple model work well for predicting future outcomes. The three key requirements for successful predictive analytics strategies are data, statistics, and assumptions (Rajni & Malaya, 2015).

2.6.1 Predictive Analytics Models

Predictive analysis approaches and techniques are classified as statistical and machine learning techniques. Some of the most popular Predictive Analytics Models are Linear regression model, Logistic regression, Forecast Model, Clustering Model, Classification Model, Regression techniques, Discrete choice models and Time Series Model (Henrys, 2021). Some of the models are discussed in detail below:

Logistic Regression: The logistic regression model is used to forecast the presence or absence of a categorical event. It is more useful while analyzing categorical dependent or binary variables compared to count data (Alt et al., 2001). The Negative Binomial Model falls under Logistic Regression and is befitting for analysing count data that has a skewed distribution. The model has been used to investigate asthma determinants based on how long an individual stays in hospital. It has been used to compare multiple statistical predictive models in forecasting the number of asthmatic patient admissions on a daily basis, in London (Soyiri et al., 2012). Logistic regression models can be difficult to build since they necessitate the presence of domain expertise as well as an insight of statistical approaches such as odds ratios, statistical significance, interactions, and confounding (Ayer et al., 2010)

Linear Regression: Linear regression is the most used model due to its ease of interpretation. The model provides reasonably accurate results and has numerous applications in modelling trends and seasonality. Linear Regression derives estimates through the ordinary least squares method, that assumes, for the dependent variable, the independent variables or regressors do not have an error. Therefore, it is crucial to have the error component taken into consideration during modelling, to account for the problem. Linear regression also requires a lot of data on all attributes to estimate the parameters (Cook & Weisberg, 1982).

Time Series Regression: The Time Series Regression model has a wide application and forecasting capabilities. It has the potential to be more consistent than basic linear regressions and provides outputs that are easy to interpret (Chatfield, 2003). Adequate data is necessary in Time series models for reliable forecasting (Soyiri et al., 2012).

Quantile Regression: Quantile regression is good at forecasting unusual occurrences. It allows for the modelling and forecasting of extreme distributions of data. As a result, they are more useful than linear regression, which forecasts based on the overall average distribution. One significant weakness of the approach is that it only deals with pertinent or distinct categories of data, which means that some details which may affect impact the accuracy and statistical power of the analysis is lost. (Soyiri et al., 2012).

Artificial Neural Network (ANN): ANN is a black box modelling technique that has been shown to produce more reliable results than the traditional causal approach. (Breiman, 2001). By automatically adjusting to changes in time series based on experiment design, ANN can model complex and, arbitrary systems. (Jones et al., 2008). The model's main disadvantage is that it is hard to comprehend and has very few available statistical software packages, unlike other approaches discussed above (McKendry, 2002). Neural networks are made up of computing units (artificial neurons) that are linked together so that each neuron can send and receive signals from and from the others. Models of distributed nonlinear systems include neural networks. These adaptable nonlinear models can discover hidden patterns in data. They gain knowledge through learning because they are designed to function similarly to the human brain.

Advantages of Neural Networks

There are several advantages to using a neural network. Among these are the potential to give valid predictions upon training. This ability is maintained even when confronted with noisy data. Non-linear mappers are also used by neural networks. Because of this property, neural networks are an excellent candidate for solving a wide range of non-linear problems (Kajitani et al., 2005).

Time Series Prediction Using Feed-Forward Neural Network Models

Many artificial neural network architectures have been investigated to solve the time series problem. These architectures include multilayer Feedforward neural networks (FFNN), recurrent networks, and radial basis functions (RBF). There is great inspiration behind using FFNN to predict time series. In forecasting, FFNN outperforms the chaos model, but both outperform the traditional random walk model. FFNN seems to be appropriate for time series prediction with low signal-to-noise ratios if sufficient data is present and appropriate data transformation techniques are used. (Kajitani et al., 2005).

Back Propagation Networks

A back propagation network employs a feed forward topology, supervised learning, and a back propagation algorithm. In terms of the necessary requirements for computational training, it is a powerful but costly model. The “Sliding window” technique can be used to forecast time series using a back propagation network. In this case, the neural network can be given a set period, and the desired output is the function at the next time. A back propagation network's training process is controlled by learning rate and momentum. The learning rate determines if the neural network will make any significant changes after each learning trial. Momentum is used to control potential weight oscillations (S.-C. Wang, 2003).

2.6.2 Predictive Analytics Modelling Process

The below steps are undertaken during the Predictive Modelling Process. They can be divided into three principal categories: data preparation, building a prediction model and deployment (Henrys, 2021).

1. Understanding the Objective of the Business

2. Defining the Modelling Goals
3. Selecting/Getting Data
4. Preparing the Data
5. Analysing and Transforming the Variables
6. Random Sampling (Train and Test)
7. Model Selection and Develop Models (Training)
8. Test the model
9. Validate the model
10. Finding the right Modelling
11. Optimization of the model
12. Deployment and scaling

2.6.2.1 Data Preparation

Data preparation is a self-service process which involves the transformation of disparate, raw, and jumbled data into a clear and consistent picture. It includes activities such as research, cleansing, transformation, organization, and data collection. This process is very critical but takes time. In order to have data ready for analysis, the data teams can devote up to 80% of their time to combine data from various sources into a single file, then convert the raw data into high-quality results (Henrys, 2021).

2.6.2.2 Building the Prediction Model

The category involves building a prediction model that predicts the future by learning from historical and current data. The type of algorithm in developing the predictive model is entirely determined by the type of data and the end goal. The predictive model should undergo testing and validation before it can be used (Henrys, 2021).

2.6.2.3 Deployment and Scaling

This category involves releasing the model to the organisation for all decision-makers and users to access, after all prior steps have been completed (Henrys, 2021).

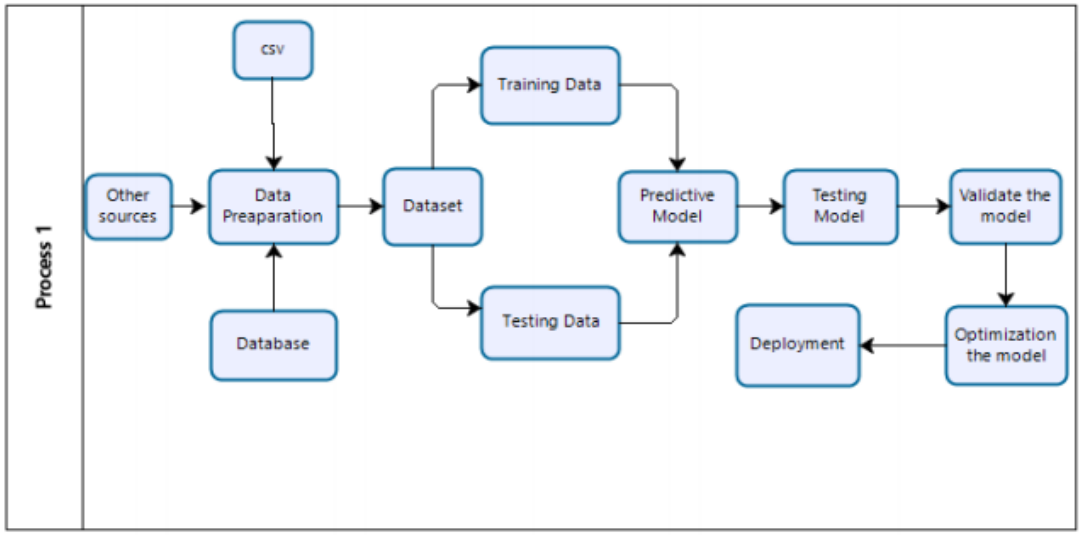


Figure 2.6: The steps involved in building Predictive Analytics Models

Source: (Henrys, 2021)

2.6.3 Predictive Analytics Tools

There are numerous software tools vailed to carry out predictive analytics. However, one can also choose to write their own code instead of using the software. Python and R programming are the best languages used in the domain of predictive analytics. An advantage of programming is that there are no monthly or annual license fees which is not the case with generic software tools (Henrys, 2021). Below are some Generic software tools used for carrying out predictive analytics:

Product	**Best For**
SAP Analytics Cloud	The overall best Predictive Analytics solution
SAS Advanced Analytics	The best business intelligence tool suited for all enterprises
RapidMiner	The top best open-source predictive analytics software
Alteryx	Best predictive analytics tool for team collaboration
Emcien	Top predictive analytics tools for marketing
Ibi WebFOCUS	Good predictive analytics for beginners
IBM SPSS	Good predictive analytics for researchers

Table 2.1: Well-known Predictive Analytics Tools

Source: (Henrys, 2021)

2.7 Related Works

(Mondal et al., 2020) used data analytics to understand the spread of an infectious disease in China using an open dataset published by Johns Hopkins University. Data visualizations were presented in the form of tables, bar charts, pie charts, and bubble graphs. The study also predicted future trends in the virus's spread by simulating the global rise of the number of infection cases. Python programming language was used to compare the performance of linear regression and polynomial regression algorithms. The highest values of R2 and adjusted R2 were obtained from the model built using the polynomial regression algorithm. The below graph depicts the actual confirmed cases and predictions made using the polynomial regression algorithm:

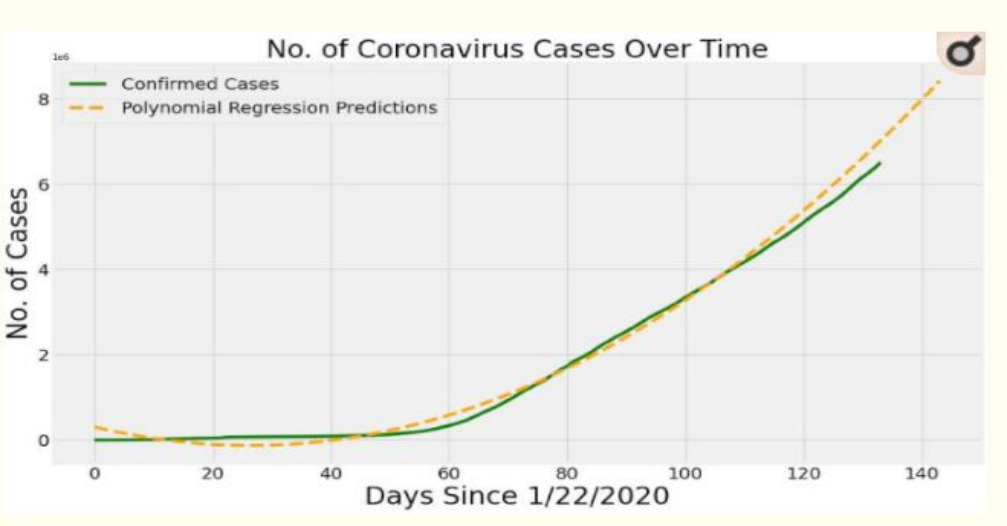


Figure 2.7: Prediction Using Polynomial Regression
Source: (Mondal et al., 2020)

The study also used classification to diagnose COVID-19 patients in Brazil. They used classification algorithms such as the Multilayer Perceptron (MLP), XGBoost, Logistic Regression (LR), K-Nearest Neighbour and Decision Trees (DT) to compare their performances. Hospital Israelita Albert Einstein located in Brazil provided 5644 samples of COVID-19 test data to predict the likelihood of a sample testing positive for the virus. Tables, bar charts and pie charts are used to visualize data in this study. Below is the performance comparison of the classifiers:

Classifier	Precision	Recall	F1 Score	AUC value	Testing Accuracy
MLP	93%	93%	93%	96.70%	93.13%
LR	92%	93%	92%	96.60%	92.12%
XGBoost	92%	92%	92%	96.30%	91.57%
KNN	89%	89%	89%	93.70%	88.91%
DT	87%	87%	87%	94.40%	86.71%

Figure 2.8: Performance Comparison of the classifiers
Source: (Mondal et al., 2020)

From the study, the best classifiers based on the accuracy levels were MLP, XGBoost, and LR, which predicted the COVID-19 patients more reliably. However, the study focused on a small dataset. It is therefore important to validate the effectiveness of the classifiers and regression algorithms on a larger dataset to conclude which models give the most accurate predictions. The study also suggests building models that can forecast the effects of various vaccinations and treatments performed on COVID-19 patients and one that can predict how the virus spreads to identify possible risky zones that should be avoided.

(Yadav, 2020) applied data analytics to a dataset of COVID-19 cases in India obtained from Kaggle using machine learning methods. The study focuses on predicting the rates of COVID-19 infection for a duration of seven consecutive days. The aim is to enhance decision making by the Indian Government and doctors. MATLAB software is used to carry out the data analysis and regression analysis using six different models: quadratic

polynomial, Third degree polynomial, Fourth degree polynomial, Fifth degree polynomial, Sixth degree polynomial and the Exponential polynomial. The researcher chooses a training dataset consisting of confirmed cases between March 1, 2020, and April 11, 2020, and a test dataset consisting of confirmed cases recorded between April 12, 2020, and April 19, 2020. The R2, adjusted R2, Sum Square of Errors (SSE) and the Degree of Freedom for Error (DFE) are the metrics used in measuring the performance of each model. Below are the results obtained based on the study:

S. no.	Model	SSE	R^2	DFE	Adjusted R^2
1	Exponential polynomial	845,000	0.9951	40	0.9950
2	Quadratic polynomial	9,209,100	0.9463	39	0.9436
3	Third degree polynomial	6,466,400	0.9962	38	0.9959
4	Fourth degree polynomial	2,777,700	0.9984	37	0.9982
5	fifth degree polynomial	2,426,900	0.9986	36	0.9984
6	Sixth degree polynomial	1,656,800	0.9990	35	0.9989

Figure 2.9: Performance Comparison of the models
Source: (Yadav, 2020)

Based on the prediction outcome and above metrics, the Sixth-degree polynomial model performs best by giving predictions that are very close to the actual results as shown below:

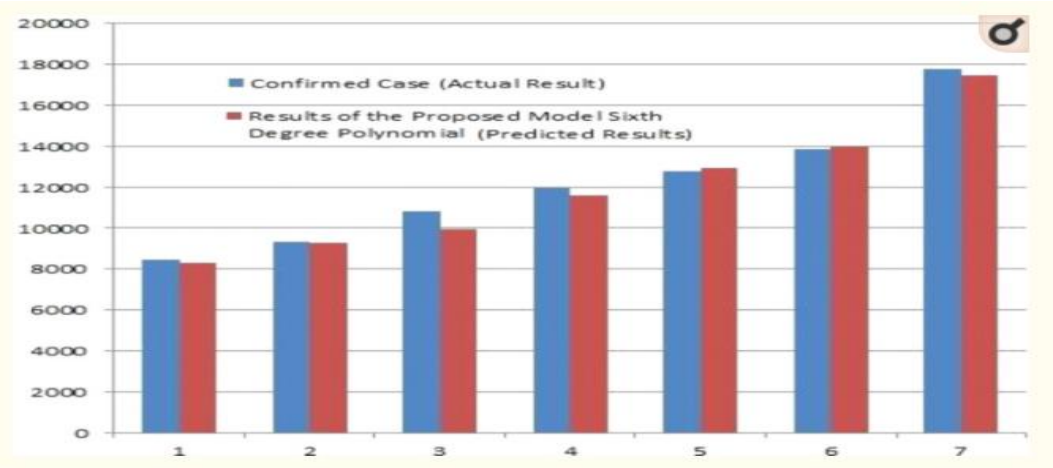


Figure 2.10: Actual Confirmed Cases Against Predicted Results by Sixth Degree Polynomial
Source: (Yadav, 2020)

The training set of the model is very small as it consists of data from cases reported within six months. It therefore needs to be trained with a larger dataset to confirm its effectiveness and reliability in prediction. The researcher gives an insight into developing a regression analysis model based on ANN that can acquire data at set intervals to predict the number of infection cases recorded weekly and bi-weekly.

Artificial Neural Networks (ANN) are mostly used in carrying out time series forecasting. Raw data can be fed into ANN-based techniques and automatically identify the required feature representation. Moreover, ANNs provide reliable results based on performance, accuracy, latency, speed, convergence, and size (L. Wang et al., 2018). (N. Hamadneh et al., 2021) propose an ANN model using a multilayer perceptron neural network (MLPNN) and a prey-predator algorithm (PPA) resulting to a hybrid model (MLPNN-PPA). The model is used to forecast COVID-19 confirmed and recovered cases in Saudi Arabia, using data between January and September

2020. PPA is a metaheuristic algorithm which determines the best values for the model parameters to be used on MLPNN, to improve its performance. The MLPNN uses one hidden layer with ten hidden neurons, and the sigmoid as the activation function. The study used RMSE, and correlation coefficient (R) as performance metrics. An RMSE value of 0.13 and R value of 0.93 were obtained. Based on the test data, the MLPNN-PPA model outperforms all others in forecasting the infected and recovered cases in Saudi Arabia. The model predicted an increase in the number of infected persons with recoveries between 2000 and 4000 in a day. However, the model was trained on a dataset obtained within a short period of time.

(Nyoni, 2020) uses the multi-layer perceptron neural network to predict the number of COVID-19 cases in Morocco. The data used is between March 2, 2020, and October 31, 2020. The study employs the ANN (12, 12, 1) model and uses the hyperbolic tangent function as the activation function. The data used in the study is derived from the online database of USA's Johns Hopkins University. An MSE of 30758.263463 and an MAE of 117.385530 were obtained upon testing. Upon prediction, the study showed that, COVID-19 cases will rise from November1, 2020 with 3661 cases till December 11, 2020, with 4305 cases after which an equilibrium daily case volume will be obtained for the rest of the days. In this study as well, data used to train the model was obtained within a short period of time.

(Ghosh, 2020) presents a mathematical model to forecast the COVID-19 spread in India. The study suggests a 'change-factor' based mathematical model for the prediction, with data from multiple sources. The term 'change-factor' is defined as "a measure of the daily increase in relation to the previous N days". The proposed system depicted an accuracy of 90.36% in January which increased to 96.67% in April. The accuracy of the model increases as time goes by. It incorporates the preceding day's data into the trendline for the subsequent days, allowing it to self-correct. As much as the model predicts better with increase in data and as days go by, it is built to specifically suit the needs of India and requires modifications to the model to forecast COVID-19 cases in other countries.

According to N. Hamadneh et al (2021) there are published studies on COVID-19 that focus on the computational, mathematical, and statistical aspects of different viruses. Through mathematical models, different models are used to study the dynamics of COVID-19. The Susceptible Infectious Recovered (SIR) model is mostly used. A system of time-dependent differential equations is used to describe the epidemic's growth. (Zakary et al., 2016) adopted a similar model to assess the significance of HIV/AIDS awareness programs and travel restrictions as preventative measures of the outbreaks (Godio et al., 2020) utilized the Susceptible-Exposed-Infectious-Removed (SEIR) model to study the COVID-19 outbreak, case study of Italy. They used the Particle Swarm Optimization (PSO) to determine the parameters of the model through a stochastic approach. The main aim being improvement of the predictions' reliability within a term of 30 days. The researchers compared their results against data and predictions made for Spain and South Korea.

(Mbogo & Orwa, 2021) employ an SEIHCRD mathematical model with the reported COVID-19 cases in Kenya to determine the variation in the transmission of the disease over time. According to Singh & Kumar Bajpai (2020), the SEIHCRD model is divided into seven state compartments: S-susceptible, E-exposed, I-infected, H-hospitalized, C-critical, R-recovered, and D-deceased or death. The model illustrates the transition from state to state beginning from the susceptible class (S), which represents people with no history of the disease infection. Once a person gets infected, they become symptomatic and transition to compartment E, exposed. If the exposed

develop symptoms, they become infected and move to compartment I. The infected either quarantine themselves at home or go to the hospital to get medical attention. Those who get hospitalized are categorized under compartment H. If hospitalized, the situation can either worsen and the person moved to ICU class (C) or recover and transition to compartment R. The final compartment represents the death class (D), those who succumb to the disease. The study uses it to develop a system of non-linear ordinary differential equations to model the patterns in the disease outbreak and spread. This enhances an understanding of the trends in the spread It calculates the rate of reproduction of the disease using the next generation matrix approach with the aim of facilitating the investigation on the factors causing the infection. Moreover, the model also examines the effect of COVID-19 mass testing as well as self-initiated behavioural change. From the results, both play a major role in eliminating and lowering the spread of COVID-19. The model uses data on the estimated number of new infections cases, total number of infections recorded so far, total number of people in hospitals, those in ICU, those recovered and the number of death cases to predict the number of expected COVID-19 cases in the next 14 days.

(Odhiambo et al., 2020) present their work on Stochastic modelling and prediction of COVID-19 cases in Kenya using R statistical program. The stochastic model comprises of several methods which are combined. The conceptual framework explains the three stochastic states of the COVID-19 infection for any given population: susceptible, infectious and recovery. The Susceptible state (S) contains individuals who are predisposed to the virus. The Infectious state (I) contains those infected by the virus while the Recoveries state (R) comprises of those who have either recovered or succumbed. An assumption is made that an individual moves into another state within the duration of time that the individual remains infected. The number of confirmed COVID-19 infections in Kenya is then modelled using a Compound Poisson model as a Compound Poisson Process. The parameters derived from the model together with data on the confirmed COVID-19 cases, the number of individuals who were in contact with those infected and the number of flights recorded on a daily basis from countries with the infection are used to carry out a multiple linear regression to forecast COVID-19 infection cases in Kenya.

(Ahmed et al., 2020) investigate the COVID-19 Data Summarization and Visualization (DSV) tool's data collection design, development, and implementation Go.Data. The tool aided in bridging the gap on data collection that existed between the beginning of the disease outbreak and the beginning of collection of data regarding it. In WHO African regions, Go.Data is a highly-values electronic platform. The movement restrictions and quarantine within the continent's various geographical areas create a void in the collection and management of COVID-19 data using Go.Data. In addition, the study compares, and contrasts Go.Data to other existing WHO electronic platforms within Africa. The DSV tool is used in 18 countries. Its main advantage is that it is available in multiple languages.

(Jayesh & Sreedharan, 2020) conduct an exploratory visual data analysis on COVID-19 cases in Kerala, India. The analysis is carried out in three phases: between January 30 and March 9, 2020; between March 10 and May 8, 2020; and between May 9 and May 31, 2020. The analysis of COVID-19 is carried out using MATLAB software, which includes the creation of bar graphs, time-series line plots, pie charts, and parallel plots. From the study understanding the plots requires knowledge of mathematics. Nevertheless, parallel coordinate plots are superior to other graphs due to their ability to compare multiple variables at once. In the study, the analysis of new confirmed cases, deaths, recovered cases, testing and welfare measures are all considered to get a clear picture

of this pandemic outbreak. This study mainly involves analysing the death cases, confirmed cases, and new cases from the COVID-19 tests conducted. The quarantine and hospital admission report data is visualized. However, the study does not involve any prediction, as no prediction models are mentioned.

(Gao et al., 2020) conduct research on visualizing the spread and expansion of coronavirus disease using area cartograms. The study combines six circular cartograms. These cartograms depict COVID-19 infection cases in China on January 24th, January 26th, January 28th, January 30th, February 1st, and February 3rd. On these days, confirmed cases from the previous 48 hours are considered. These cartograms represent each province of China. Their size varies with the number of confirmed cases. The spatial and temporal information of the data can be conveyed in an efficient and effective manner using this design. However, the study does not involve any sort of prediction nor adoption of any prediction models.

2.8 Research Gap

According to Poletto et al (2020), there have been over thirty thousand academic publications with the title COVID-19 on Google Scholar, but less than 2% of the papers state from the title, that they carry out predictive analytics or predictive modelling. It is therefore necessary to create more models that predict COVID-19 cases and identify which algorithms perform best for different scenarios. From the studies, models have been created to perform prediction in other countries but cannot be readily replicated to perform prediction in Kenya. The statistical models created are good while handling small amounts of data, but with vast amounts of data, comes the need to use deep learning models (Shinde et al., 2020). The visualization platforms created are purely for visualizing COVID-19 cases. Therefore, this study aims at using Neural Networks to forecast the virus' spread and visualize the forecasted data, COVID-19 cases, and health resources available to combat the spread of the disease. It merges the prediction using ML models and visualization using dashboards which will enhance decision support.

2.9 Conceptual Model

The researcher has critically analysed the existing models for predicting COVID-19 cases. From the analysis, the researcher has identified a need to create a model with the current COVID-19 large datasets using Deep Learning. The raw data will be cleaned by filling in any missing data. The cleaned data will be integrated into one platform and visualized. It will also be split into the training and testing set. The training set will be passed through the ANN algorithm to develop the predictive model. The testing set will be used to make predictions hence testing the model to determine its efficiency. These predictions will also be visualized. The visualized integrated data and predictions will help uncover any hidden patterns, enhance an understanding of the data. This knowledge will be used by the Kenyan government and the Ministry of Health to make data driven decisions regarding the health resources in terms of the ICU beds and ventilators.

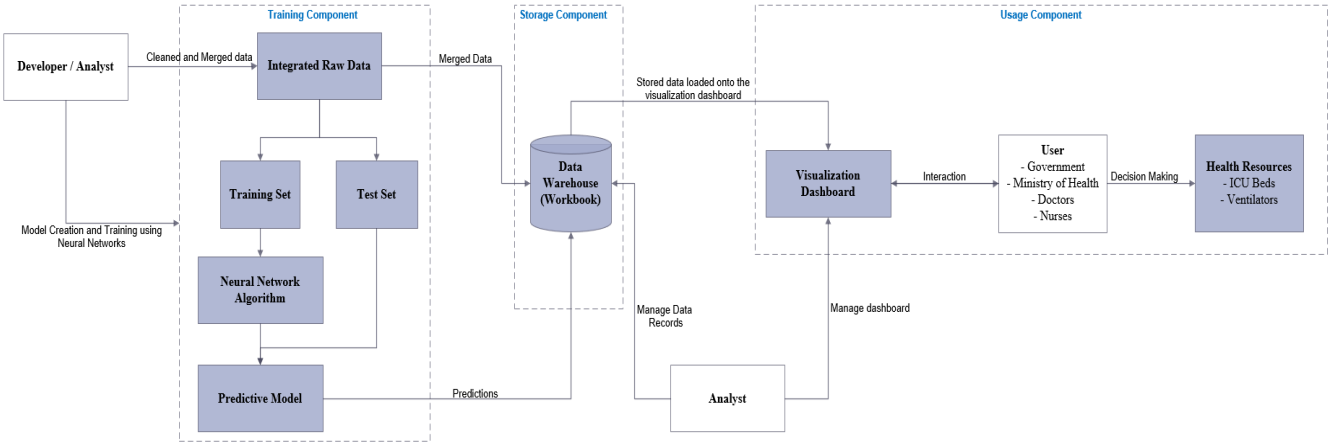


Figure 2.11: Conceptual Model

Source: Author

The diagram below depicts a high-level architecture of the dashboard, the front end and back end and the languages to be used in the development:

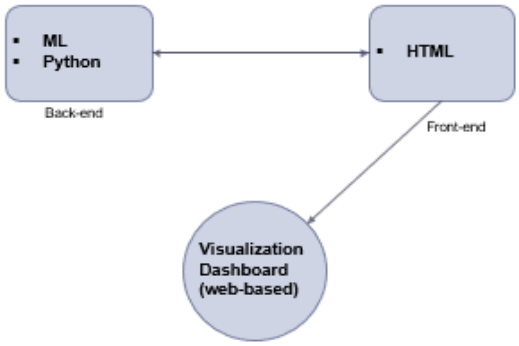


Figure 2.12: Dashboard Architecture

Source Author

CHAPTER THREE: RESEARCH METHODOLOGY

3.1 Overview

Research methodology describes the step-by-step procedures that the researcher followed in the execution of the project to achieve the research objectives. It includes how the researcher carried out analysis, design, development and testing of the predictive model and visualization dashboard.

3.2 Research Design

Both qualitative and quantitative research approaches were used in the study. Qualitative research was suitable as it mainly helped to gather the system requirements from its proposed users while quantitative research was suitable during the collection and preparation of the COVID-19 data, which is in numeric form, to be used in developing and testing the prototype. The study involved a descriptive research design that was based on the Waterfall Model. The research design was suitable as this study aims to use COVID-19 infection cases recorded over time as a univariate variable to predict number of COVID-19 cases in the next 60 days. This enhances decision support in terms of the health resources, which is the goal of the study. The Waterfall model was suitable for this research due to the two codes / platforms that were independently developed before integration. Moreover, the requirements of the study were well defined beforehand which allowed for unit testing and for tests to be performed at the end of each stage before progressing to the next. This gave the researcher surety when it came to the development the dashboard since the predictive model was independently tested before being subjected to perform predictions which would be visualized at a later stage.

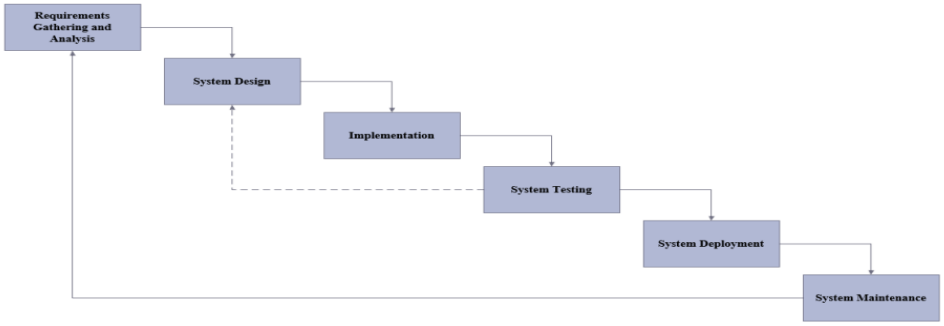


Figure 3.1: Waterfall Model

This research involved four major phases that are addressed at the different stages of the Waterfall Model approach:

1. Review and evaluate the COVID-19 predictive systems currently in use.
2. Develop a predictive machine learning model using neural networks to forecast the patterns in the spread of the COVID-19 virus.
3. Design and develop a visualization dashboard for the predicted data, COVID-19 cases and health resources set aside to contain the virus.
4. Evaluate the performance of the model and functionality of the dashboard.

Each of the phases were further broken down into specific research tasks which address specific areas of the research objective.

3.2.1 Requirements Gathering

This research followed a qualitative approach; therefore, this phase involved gathering the information needed to support the development of the system. This was done by conducting a survey on the existing predictive analytics systems developed in the phase of COVID-19, the underlying algorithms used, performance of the systems and recommendations given by the developers. The same method was used to collect data on the visualization tools used to visualize COVID-19 data, their effectiveness, and the limitations of each. Below is a breakdown of the steps followed and the outcome of each:

a) Identify the existing COVID-19 predictive systems and the countries in which they have been used.
   This was achieved by conducting a comprehensive survey on the existing literature which allows the researcher to identify the currently deployed machine learning predictive systems and in which countries they are used. The researcher was also able to understand the challenges faced during the development and deployment of the systems based on the conclusions and recommendations given on the studies. This paved way to the subsequent task.

b) Identify the algorithms applied in the development of the systems.
   This was achieved by conducting a survey on the existing literature which enabled the researcher to identify the underlying algorithms used in the systems, hence paving way to the next task.

c) Identify the best predictive systems based on the performance of the underlying algorithms
   This was achieved by assessing the performance measures recorded in the studies of the various deployed machine learning algorithms, the advantages, and disadvantages they present. This task was also exploratory as it required conducting a survey on the existing literature.

d) Identify the existing visualization tools that can integrate predictive machine learning models
   This was achieved by conducting a survey on the existing literature regarding the current visualization tools, their strengths, weaknesses, and ability to integrate the aspect of machine learning for visualizing predictions.

The output from above tasks informed the researcher on the predictive systems currently available, where they are deployed, the underlying machine learning algorithms and which algorithms to adopt for performance comparison with the Neural Networks, based on those that performed best in the studies. These steps were systematically undertaken under the literature review chapter of this study. They enabled the researcher to meet the first objective.

Study Population

The research involved 18 respondents. They comprised of eight individuals from the Ministry of Health offices, five from a major public hospital within Nairobi County and another five from a private hospital within the same county. They were either doctors, nurses, health workers or advisers of the government regarding the COVID-19 pandemic. The researcher paid a visit to the workplaces of the respondents. The respondents were suitable for the study as they are the individuals who make decisions regarding the availability of health resources. Whether to buy more ventilators and ICU beds, or not.

Sources of Data

This study used open-source COVID-19 data provided by the Kenya Ministry of Health to train the model. The COVID-19 data on infection cases was appropriate for this study at its aim is to predict the number of infection

cases in the days to come, to prepare in terms of the availability of health resources. Since the data is in read-only format and can be acquired from different pdf files, it was integrated into one excel file. Data simulation was carried out where data was missing as a pre-processing step. The researcher collected COVID-19 data from when the first case was identified, on March 12, 2020, till December 12, 2021. This was a quantitative approach. The information gathered from the surveys, interviews and data collected aided in justifying the proposed algorithm and visualization platform for this study as well as gather any requirements of the system from the proposed users. This bit of analysis was qualitative.

Requirements Analysis

After conducting the requirements gathering, the information gathered, and data collected was analysed during the requirements analysis. The literature review on the predictive analytics systems developed and underlying algorithms as well as visualization tools present enabled the researcher to identify the best algorithm to use in this study (ANN). The researcher also identified the best visualization tool to use based on the simplicity with which visualizations can be developed and understood (Tableau).

The researcher analysed the feedback from the interviews by noting down key points that gave an understanding of the expectations of the users. This aided in deciding on how best to meet those requirements. They also enabled the researcher to decide on what other system requirements or features to include during deployment.

The data collected was critically analysed to ensure that it suits the objective of the research. The researcher used open-source data on COVID-19 cases, the timelines the cases were recorded, the situation in terms of lockdowns and curfews and data on the health resources (ICU beds and ventilators) which are what the study aims at determining.

3.2.2 System Design

During this phase of the waterfall model, the researcher studied the requirements from the previous phase. The researcher then designed the system based on those requirements with the aim of integrating the powerful aspect of predictive analytics with visualization. This enabled the researcher to determine which programming languages, hardware, and software to adopt for the system.

3.2.2.1 System Requirements

The researcher identified the software and hardware tools to use based on the feedback from the respondents as well as from the survey conducted on the published works regarding predictive analytics systems and visualization tools.

Software Requirements:

1. Juypter Notebook – Integrated Development Environment used to develop the predictive model.
2. Python – programming language for coding.
3. Libraries – Numpy, Keras and Tensorflow. Keras and Tensorflow required as they are used for developing deep learning models. Numpy facilitated mathematical operations on data.
4. Tableau – visualization tool used to represent patterns in data.
5. ANN – deep learning algorithm used in training the predictive model.
6. HTML – the web language used to design the Tableau web platform.
7. CSS – to allow formatting of the web platform

8. XAMPP - allows for building WordPress site offline on a local web server on the PC

Hardware Requirements:

The system was developed on a 64-bit windows PC running on 2.70GHz processor speed with internet connectivity.

3.2.3 Software Implementation

The researcher developed the system with the aim of integrating the powerful aspect of predictive analytics with visualization tools. This phase involved developing the predictive model and visualization dashboard. The predictive model was individually tested prior to integrating the predictions with the visualization tool. Below are the two tasks that were conducted under this phase and the underlying steps for each:

3.2.3.1 Development of a predictive machine learning model using Neural Networks

This entailed creating and training a machine learning model using past data. This is the model that the researcher used to carry out predictions which is the general objective of this study.

a) Collect the COVID-19 data to be used to train the model

   The researcher gathered open-source training data provided by the Kenya Ministry of Health. The data was used to train the model in the next task of this phase. The data comprised of COVID-19 confirmed cases in Kenya across the 47 counties. The cases were then combined to represent the eight provinces in Kenya. Data on resources available (ICU beds and ventilators), lockdown and curfews were also collected.

b) Train the predictive model using Neural Networks

   The data collected was then integrated from the different sources, compiled, and saved in one excel spreadsheet. The researcher used Juypter notebook to clean the data by filling in missing values before feeding it to the algorithm. The data was then split into training and testing set. The model was created using LSTM neural networks algorithm and python programming language. Different hidden layers were used, and the performance evaluated (hyper-parameter testing) to determine the best model with the optimal network layer size and number of hidden units. The result of this task was a model trained using past data. The neural network is preferred at this because neural networks are effective at making predictions due to the hidden layers which makes the predictions accurate.

c) Test the model performance

   The researcher tested the trained model. The model was tested and evaluated at this stage by calculating the MAE, RMSE and MSE. The performance was be compared against the training accuracy to ensure that the model did not overfit. The metrics allowed the researcher to determine which model performs best and gives valid predictions.

d) Create a predictive model using Multiple Linear Regression, Bayesian Polynomial Regression, Time Series Forecasting using Feed Forward Neural Networks and Traditional ARIMA algorithms and test their performance and compare against LSTM neural networks

   The researcher then adopted other machine learning algorithms. The algorithms were trained using the same dataset collected in the first task of this phase. The researcher tested the performance of the models using the same metrics adopted in testing the LSTM Neural Networks model. This task allowed the researcher to explore more algorithms and get to compare their performances.

The output of the above tasks was a trained and tested machine learning model, hence meeting objective two.

3.2.3.2 Design and development of a visualization dashboard

This entailed designing and developing a visualization dashboard that used to visualize the COVID-19 data and predictions. This platform is user friendly and will be used by the key stakeholders in making data-driven decisions.

a) Design the dashboard and import the data collected to it.
   This step involved designing the dashboard using Tableau. The researcher read through Tableau tutorials and findings from the interviews to identify a suitable workbook and visualization for this research.
b) Import data to the visualization tool
   The data collected in the previous task was imported to the visualization tool (Tableau). Data on health resources (ICU beds and ventilators) was also visualized per county.
c) Test the performance of the visualization dashboard
   The researcher selected different features from the dashboard view to check whether the visualizations change as expected.

The output of the tasks enabled the researcher to satisfactorily meet objective three.

3.2.4 Software Testing and Evaluation

This phase involved importing the predictions to the visualization software. The researcher also checked whether the manipulations of data within Tableau gave the expected outcome. The graphs update in real-time when different variables / features are selected. The live evaluated dashboard with the predictions was then transferred to a Tableau web page. This was done by connecting the Tableau desktop server to an online Tableau account. The web page was manipulated using HTML language on Brackets text and code editor to give the platform a name relating to the objective of this research. The link to the online page was then shared with the respondents engaged during the requirements gathering phase to evaluate the system. Feedback from the individuals was analysed by the researcher and any modifications made to suit their requirements and views. This enabled the researcher to meet the last objective of the study.

3.3 Ethical Considerations

The research has applied specific ethical guidelines to ensure all the research participants were not at any time during the research exposed to any form of harm that may ruin their reputation or careers. The study protected the identity of all the participants by ensuring that they maintain as confidential any information given to them orally or in writing. The researcher also ensured that all work quoted in the thesis is credited using APA 7th edition in-text citations and references. The researcher was honest regarding the purpose of the study and maintained ethical behaviour to promote a sense of trust between the participants and the researcher during the requirements gathering process.

CHAPTER FOUR: SYSTEM ANALYSIS AND DESIGN

4.1 Introduction

This section captures the requirements necessary for the effective functioning of the system, the tools used, and outlines a detailed design of the system. It will also analyse and evaluate the system to determine whether it meets the user's requirements and objectives of the study.

4.2 System Analysis

4.2.1 Requirements Gathering

The researcher paid a visit to the Ministry of Health offices, one major public and one private hospital within Nairobi County to conduct interviews. The interviews with the users of the proposed system allowed the researcher to identify the requirements and necessary features that should be included to the system. The researcher carried out content analysis to identify some of the key questions that included the nature of the expected dashboard and features needed within the dashboard, number of days to predict into the future. Most of the responses to the queries included dashboards that are easy to interpret, a forecast for a minimum of 50 days to allow for proper preparation in terms of health resources in each of the provinces and a visual indication of whether the resources available are adequate based on the predictions made. For this, a suggestion was made for the dashboard to display a green zone and a red zone. The Red zone (shaded in a red colour) would indicate a danger point whereby, the resources available are not adequate whereas, the green zone (shaded in a green colour) would represent a scenario whereby, resources are adequate to contain and mitigate the spread of the virus.

A survey approach was used to understand the algorithms that have been adopted in carrying out prediction, their advantages, and disadvantages. Some of these such as Multiple Linear Regression, Bayesian Polynomial Regression, Time Series Forecasting using Feed Forward Neural Networks and Traditional ARIMA algorithm were adopted for comparison with LSTM Time Series Forecasting. This was a key step in the research since it made it easier to identify an appropriate algorithm for the model development. The survey also allowed the researcher to identify the best type of graphs to use in the representation of the data on the dashboard. The requirements are summarized in the table below:

Key Concerns from the Users	**Recommendations/Suggestions from the Users**
Number of days to predict into the future	A minimum of 50-day forecasts to allow for proper planning in terms of additional resources that may need to be procured or re-allocation of existing resources.
Complexity of the dashboards	Simple graphs that are easy to interpret.
How to establish the health resources required, from the dashboards	Proposal to have the dashboard to display a green zone and a red zone. The Red zone (shaded in a red colour) would indicate a danger point whereby, the resources available are not adequate whereas, the green zone (shaded in a green colour) would represent a scenario

	whereby resources available are enough to contain and mitigate the spread of the virus.
Interactive capabilities of the dashboards	A dashboard that incorporates the aspect of filtering especially if multiple variables will be visualized within the same dashboard. The possibility to get the exact number of cases predicted or recorded on a particular day by simply hovering the cursor across the graphs.
Knowledge on how to use the system and interpret the dashboards.	Proposal to have the analyst inform the users on how to interpret the visualizations and how to interact with the system.
How long it will take for a new forecast to be made	The dashboards to be updated with new forecasts when five days minimum are remaining in the existing forecast.

Table 4.1: Requirements Gathered from the users
Source: Author

4.2.2 Requirements Specification

The target users of this system are doctors, nurses, and the Ministry of Health on behalf of the Government of Kenya. The proposed system architecture is comprised of a training component which outputs the predictive model which is created using python programming language on Juypter notebook IDE. A storage component, a data warehouse, which is an excel workbook connected to Tableau desktop that stores all the coronavirus cases as well as prediction for visualization. Thirdly, it comprises of a Usage component which is the visualization dashboard. The dashboard is created on Tableau desktop. The Tableau desktop server is then connected to an online Tableau account. The online Tableau account generates scripts that are used to import the dashboards to a web page using html. All these components help in the generation of a visualization dashboard that gives forecasts to aid in decision related issues in healthcare resource planning which is the general objective of the study. The system comprises of both functional and non-functional requirements. The non-functional requirements are a set of criteria used to evaluate a system's specific operation. Therefore, a software system's quality attribute is defined by a non-functional need. A functional requirement on the other hand, defines a system or one of its components and it specifies the tasks that the program or system must complete.

4.2.2.1 Functional Requirements

1. The number of cases dynamically display on the dashboard when a user hovers the cursor across the visualizations.
2. The user should have the capability to filter the different views of the graph (actual cases, predictions, green and red zones, ventilator thresholds).
3. The visualizations should be easy to interpret.
4. The site should only be availed to the stakeholders (health workers, doctors, nurses, and the Ministry of health to avoid unnecessary panic to the members of the public.

5. The users of the system should be informed about the key attributes on the dashboards.
6. The users should be able to always download dashboards from the site.

4.2.2.2 Non-functional Requirements

1. The site should be always accessible.
2. Each dashboard should not take more than ten seconds to load and fully display.
3. The performance of the dashboard should not deteriorate irrespective of the number of simultaneous users.
4. The analyst should regularly make updates in case new ICU beds or ventilators have been procured
5. Regular updates upon procurement of additional ICU beds or ventilators should be done and reflected on the dashboards.
6. New predictions should be made when five days are remaining to exhaust the forecast period.

4.2.3 Feasibility Study

4.2.3.1 Technical Feasibility

The researcher carried out technical feasibility and it was evident that the project was feasible with minimum risk as outlined below:

Technology required	**Current availability**	**Risk**	**Action**
Machine Learning Algorithms	Available	None	N/A
Visualization	Available (Open source)	None	N/A
Deep Learning libraries	Available (Open source)	None	N/A
Access to COVID-19 data	Available (Open source)	None	N/A
Programming language	Available (Python)	None	N/A
Technical manpower	Available	None	N/A

Table 4.2: Technical Feasibility

Source: Author

4.2.3.2 Economic Feasibility

There were no significant implications on cost since all the required technologies were open source. The only cost would be incurred during the project phase where the developer expects a salary payment which in this case is a non-factor given the nature of the project. Moreover, although there were no costs in terms of salary, there was an identified opportunity cost as the time spent on the project could have been directed to other activities.

4.3 System Design

The researcher designed the system to combine the powerful aspect of predictive analytics with visualization in the face of coronavirus. The main aim behind the platform is to enhance decision making in the health care sector with respect to resources available (ICU beds and ventilators), with the aim of containing and mitigating the spread of the virus. The design considered capabilities such as viewing and downloading web-based coronavirus prediction visualizations of the various provinces within Kenya.

4.3.1 Components of the Designed System

The developed prototype is comprised of three main components, the input, process, and output. The inputs comprise of all the data collected that is divided into five categories; the COVID-19 Infection Cases, Recoveries, Vaccinations, Death Cases and Health Resources which are mainly the ICU beds and ventilators. All these were pre-processed by performing data integration, data cleaning and data transformation. Data on COVID-19 infection cases was used as a univariate variable to perform time series forecasting. It was divided into the training and testing set.

The training set was fed into the input layer of the LSTM neural network architecture. This marked the beginning of the processing phase which is the second component of the system. The neural network is comprised of three LSTM layers, a dropout and a dense layer which feeds into the output layer. The training occurred for a total of 70 epochs to generate the predictive model. The testing set was used to evaluate the predictive model by determining how well it predicts. The predictions together with the remaining four categories of the data collected were then passed on for visualization. Data visualization began by importing the files containing the data to Tableau desktop. Tableau was used to create visual representations of the data using graphs. Thresholds on the number of ventilators available and reference lines to indicate the availability of ICU beds were incorporated to the visualization dashboards. The visualizations on predictions were developed for each of the eight provinces. Tableau server was used to connect to and save the dashboards on online Tableau public. Visualizations on Tableau public autogenerate an embedded code that can be used to export the dashboards to a site designed using HTML and CSS. Once this was completed, Apache on XAMPP was used to create a localhost to host the site locally on the personal computer. This activity marked the end of the processing component of the system.

The output component is comprised of the web-based interactive visualization dashboard that is meant to be used by the doctors, nurses, and the Ministry of Health on behalf of the Government of Kenya. It gives them the capability to view and download visualizations from the dashboard. Below is a detailed architecture of the system:

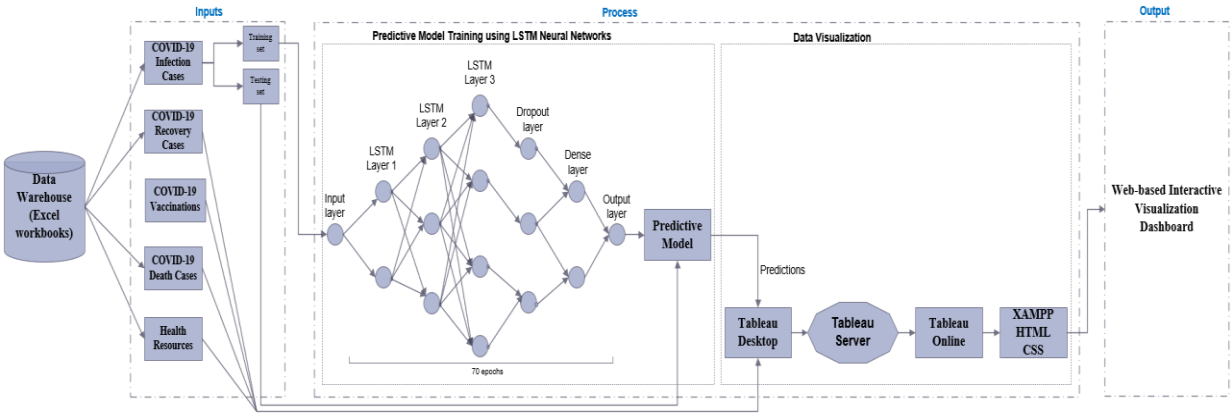


Figure 4.1: Low Level System Design

Source: Author

The system is comprised of logical flow of events as shown below:

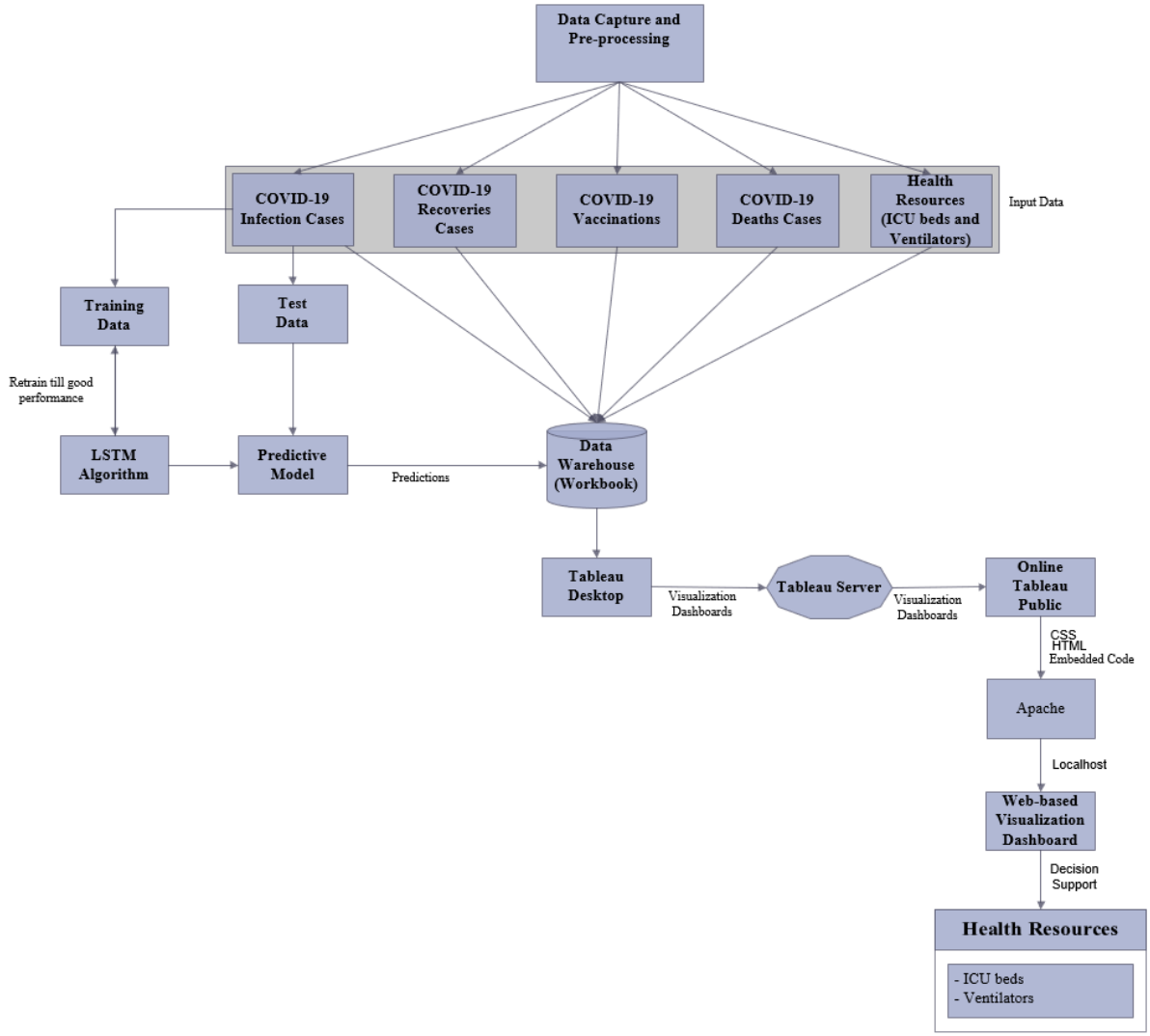


Figure 4.2: Logic flow of events in designing the system

Source: Author

CHAPTER FIVE: SYSTEM IMPLEMENTATION, TESTING AND EVALUATION

5.1 Introduction

This section captures the tasks and steps involved in the development phase of the project; the performance metrics used in testing it as well as approaches used in its evaluation.

5.2 System Implementation

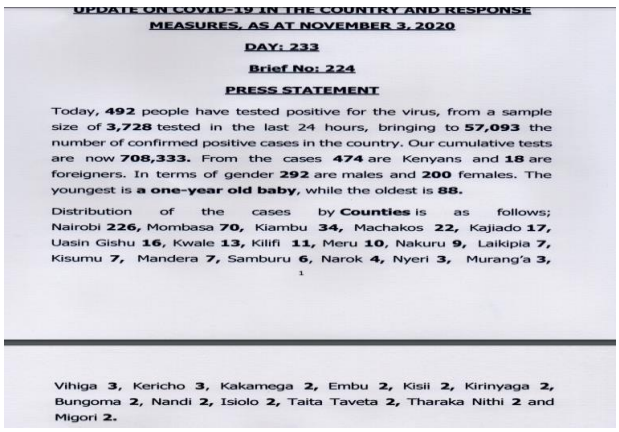
The implementation phase of this project comprises of two major tasks: the development of the predictive model and development of the visualization dashboard.

5.2.1 Development of the Predictive Model

The researcher chose Juypter IDE and python programming language as tools for the development of the predictive model. Guido van Rossum created the Python programming language in 1985. It is a high-level programming language with a simple programming syntax and has very short lines of code, making it simple to learn. Furthermore, Python has a large number of user friendly and less complicated libraries and frameworks. (Dhiman, 2020). For this instance, the scikit learn libraries came in handy in the development of the predictive model due to the various algorithms that the researcher worked with. These are the main reasons that the researcher opted for it. Juypter Notebook integrates well with python language. Moreover, it allows for line-by-line execution of the code making it easy to debug. Installation of new libraries is also quite simple with Juypter Notebook as the pip commands can be run on the notebook itself. This ease of use and integration with Python language and its libraries led the researcher to settle for Juypter Notebook as the Integrated Development Environment.

5.2.1.1 Data Extraction

Data extraction refers to the process of retrieving data that has been saved to a certain location, to have it replicated to a destination file for further processing. The researcher used the Press Statements on COVID-19 uploaded to the Ministry of Health site. This data is in pdf format and contains daily updates on the COVID-19 infection cases in each county, cumulative recoveries, cumulative deaths, and cumulative vaccinations in Kenya. The researcher accessed each pdf document from the one dated March 12, 2020, which is when the first coronavirus case was reported in Kenya till December 12, 2021. The cases recorded each day were then integrated to an excel sheet (.xlsx file) for data pre-processing. Below is the raw data in pdf format:



UPDATE ON COVID-19 IN THE COUNTRY AND RESPONSE MEASURES, AS AT NOVEMBER 3, 2020

DAY: 233

Brief No: 224

PRESS STATEMENT

Today, **492** people have tested positive for the virus, from a sample size of **3,728** tested in the last 24 hours, bringing to **57,093** the number of confirmed positive cases in the country. Our cumulative tests are now **708,333.** From the cases **474** are Kenyans and **18** are foreigners. In terms of gender **292** are males and **200** females. The youngest is **a one-year old baby**, while the oldest is **88.**

Distribution of the cases by **Counties** is as follows; Nairobi **226,** Mombasa **70,** Kiambu **34,** Machakos **22,** Kajiado **17,** Uasin Gishu **16,** Kwale **13,** Kilifi **11,** Meru **10,** Nakuru **9,** Laikipia **7,** Kisumu **7,** Mandera **7,** Samburu **6,** Narok **4,** Nyeri **3,** Murang'a **3,**

1

Vihiga **3,** Kericho **3,** Kakamega **2,** Embu **2,** Kisii **2,** Kirinyaga **2,** Bungoma **2,** Nandi **2,** Isiolo **2,** Taita Taveta **2,** Tharaka Nithi **2** and Migori **2.**

Figure 5.1: COVID-19 data in pdf format before collection

Below is data integrated from different pdfs and represented on excel:

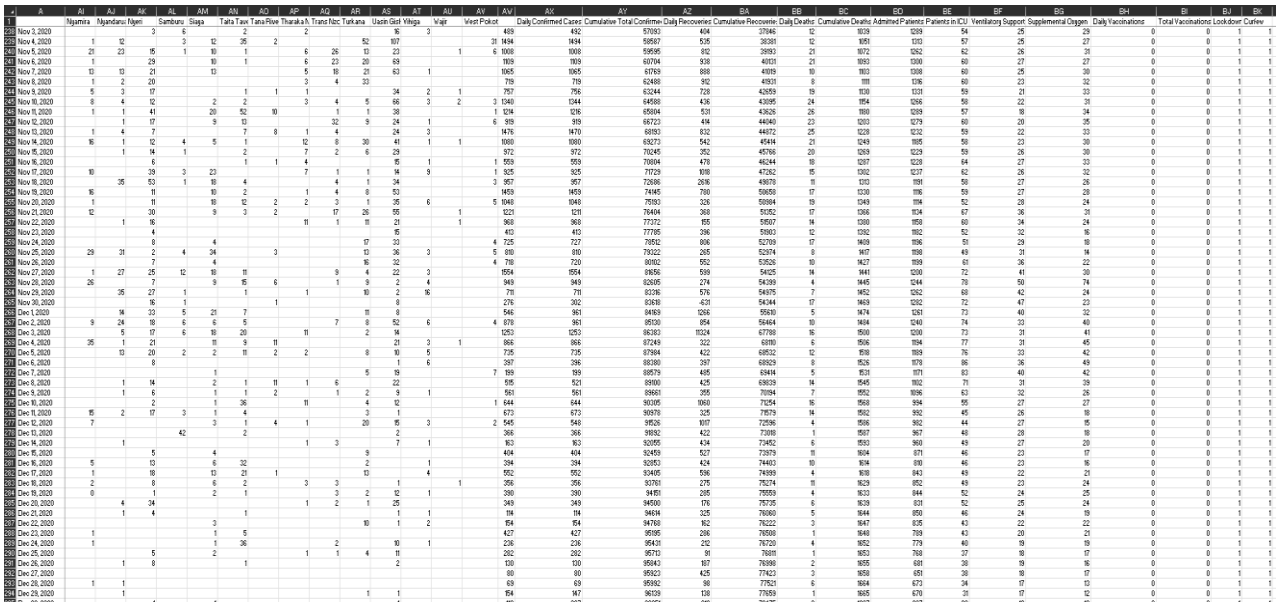


Figure 5.2: COVID-19 data in .xlsx format after collection

Open-source data on COVID-19 resources available in the Kenyan counties was also collected and extracted to an excel sheet:

COUNTY	ICU Beds total	Beds in Use	Available ICU Beds	No of functional ventilators	Ventilators in use	Available Ventilators
Nairobi	276	166	110	167	80	87
Uasin Gishu	59	41	18	31	15	16
Mombasa	45	28	7	17	12	5
Bungoma	24	10	14	8	7	1
Kisumu	23	12	11	11	4	7
Nakuru	20	12	8	8	0	8
Kiambu	16	13	3	4	0	4
Nyeri	15	13	2	3	2	1
Kisii	14	13	1	5	1	4
Kakamega	9	9	0	0	0	0

Figure 5.3: Online data on ICU beds and Ventilators in Kenyan Counties

Data on Health resources (ICU beds and Ventilators) collected and saved on an excel sheet:

	A	B	C	D	E	F	G
1	County	Total ICU beds	Total ICU beds in use	Available ICU beds	No of functional ventilators	Ventilators in use	Available ventilators
2	Nairobi City	276	166	110	167	80	87
3	Uasin Gishu	59	41	18	31	15	16
4	Mombasa	45	28	7	17	12	5
5	Bungoma	24	10	14	8	7	1
6	Kisumu	23	12	11	11	4	7
7	Nakuru	20	12	8	8	0	8
8	Kiambu	16	13	3	4	0	4
9	Nyeri	15	13	2	3	2	1
10	Kisii	14	13	1	5	1	4
11	Kakamega	9	9	0	0	0	0
12	Garissa	6	4	2	0	0	0
13	Kwale	6	3	3	0	0	0
14	Narok	6	0	6	0	0	0
15	Embu	18	5	0	0	0	0
16	Kericho	5	5	0	0	0	0
17	Machakos	5	5	0	0	0	0
18	Kirinyaga	4	4	0	0	0	0
19	Makueni	4	2	2	0	0	0
20	Murangá	3	3	0	0	0	0
21	Meru	2	2	0	2	2	0
22	Baringo	0	0	0	0	0	0
23	Bomet	0	0	0	0	0	0
24	Busia	0	0	0	0	0	0
25	Elgeyo Marakwet	0	0	0	0	0	0
26	Homa Bay	0	0	0	0	0	0
27	Isiolo	0	0	0	0	0	0
28	Kajiado	0	0	0	0	0	0
29	Kilifi	0	0	0	0	0	0
30	Kitui	0	0	0	0	0	0
31	Laikipia	0	0	0	0	0	0
32	Siaya	0	0	0	0	0	0
33	Turkana	0	0	0	0	0	0
34	Migori	0	0	0	0	0	0
35	Nandi	0	0	0	0	0	0
36	Trans Nzoia	0	0	0	0	0	0
37	Nyandarua	0	0	0	0	0	0
38	Taita Taveta	0	0	0	0	0	0
39	Nyamira	0	0	0	0	0	0
40	Vihiga	0	0	0	0	0	0
41	Lamu	0	0	0	0	0	0
42	Marsabit	0	0	0	0	0	0
43	Tharaka-Nithi	0	0	0	0	0	0
44	West Pokot	0	0	0	0	0	0
45	Mandera	0	0	0	0	0	0
46	Tana River	0	0	0	0	0	0
47	Samburu	0	0	0	0	0	0
48	Wajir	0	0	0	0	0	0

Figure 5.4: Data on ICU beds and Ventilators in Kenyan Counties in .xlsx format after collection

5.2.1.2 Data Pre-processing

It is critical to extract useful knowledge from data to make sound decisions. Not all data, however, is always ready for processing. Data preparation is a critical step in preparing data for final processing because it may contain noise, missing values, redundant properties, and so on. (Qamar & Raza, 2020). Three steps were involved in data pre-processing.

a) Data Integration

Data integration is the process of combining data from various data sources, such as databases and files, into a coherent dataset. (Malley et al., 2016). COVID-19 data from different pdf files was integrated into a single excel sheet with all the necessary data needed for the creation of the system to meet the objectives of the study.

b) Data Cleaning

Data in the real world is typically "messy," as it may be incomplete, noisy, or inconsistent. (Malley et al., 2016). The integrated COVID-19 data contained missing values which were handled in two ways:

i. Data simulation

Data simulation involves generating data values where they may be missing to obtain a complete and consistent dataset. A complete dataset is necessary to ensure that the model trains on enough data for it to give valid predictions. The COVID-19 data contained missing values on several columns of the excel sheet upon data integration. They included the total number of recoveries, the number of patients in the ICU, the number of patients under ventilatory support and those under supplemental oxygen between March 12, 2020, and October 18, 2020. These were filled by calculating the probability, mean and standard deviation. This was done by using the

NORM.INV(RAND(), mean, standard deviation) function on excel. This enabled the researcher to generate a set of random numbers to fill in the blanks between the March and October. The generated values were rounded off to whole numbers and re-arranged in ascending order. This is mainly because, after a study of the integrated data, the number of cases for each of the columns recorded an increase as the days went by. For one day missing data, the researcher filled that by calculating the mean between the value recorded for the previous day and that for the day after. Below is a snip showing this:

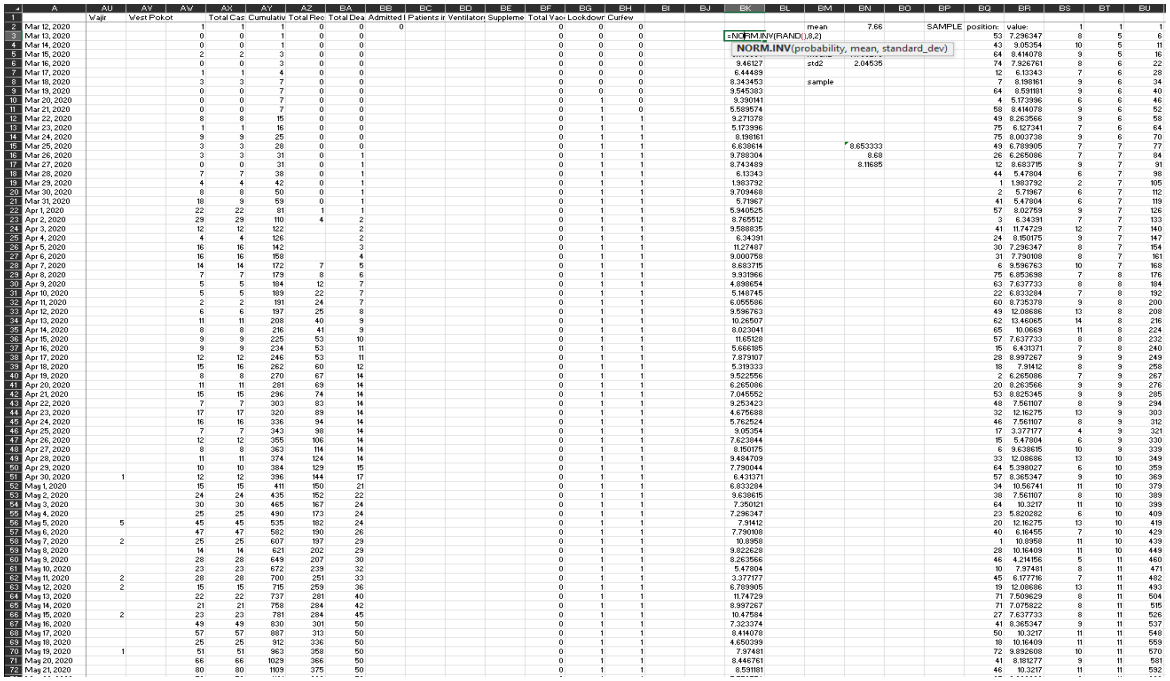


Figure 5.5: The data simulation process

ii. Replacing nan values with zero

This is a step necessary for the proper training of the predictive model to get valid and reliable predictions. This was done from the Juypter notebook IDE using the data frame method **fillna().** Zero was used to fill the blanks where no case was reported as a pre-processing step. Below is a snip of the dataset with nan values:

```
df_confirmed = pd.read_csv('corona_cases_per_province1.csv')
df_confirmed.head()
```

	Province	12/03/2020	13/03/2020	14/03/2020	15/03/2020	16/03/2020	17/03/2020	18/03/2020	19/03/2020	20/03/2020	...	04/12/2021	05/12/2021	06/12/2021	07/
0	Rift Valley	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	23	6	2	
1	Western	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	2	1	0	
2	Eastern	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	0	2	4	
3	North Eastern	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	0	0	0	
4	Nyanza	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	1	0	0	

Figure 5.6: Dataset with nan values

Below is a snip of the dataset with the nan values replaced with zero:

```
df_confirmed = pd.read_csv('corona_cases_per_province1.csv')
df_confirmed.fillna(value=0, inplace=True)
df_confirmed.head()
```

	Province	12/03/2020	13/03/2020	14/03/2020	15/03/2020	16/03/2020	17/03/2020	18/03/2020	19/03/2020	20/03/2020	...	04/12/2021	05/12/2021	06/12/2021	07/
0	Rift Valley	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	23	6	2	
1	Western	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	2	1	0	
2	Eastern	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0	2	4	
3	North Eastern	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0	0	0	
4	Nyanza	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1	0	0	

Figure 5.7: Dataset after replacing nan values with zero

c) Data Transformation

This is the third pre-processing step that was undertaken. The aim of data transformation is to change the data values to a format, scale or unit that is more appropriate for the intended analysis. Two tasks were performed as approaches to data transformation.

i. Normalization

Normalization is a data preparation technique that is frequently used in machine learning. Normalization is the process of converting the values of numeric columns in a dataset to a similar scale without changing the ranges of values. The researcher adopted the MinMaxScaler() function from Sklearn library to normalize the data prior to training. Scaling the data also helps in removing any skewness in the data for proper training. Below is a snip showing the data in form of an array, after normalization. All the values are normalized to a standard scale as shown below:

```
In [22]: train_scaled
array([[0.11922141],
       [0.14111922],
       [0.14355231],
       [0.18004866],
       [0.14355231],
       [0.09975669],
       [0.09002433],
       [0.18491484],
       [0.05596107],
       [0.13868613],
       [0.16058394],
       [0.23114355],
       [0.24087591],
       [0.22627737],
       [0.19708029],
       [0.08029197],
       [0.28467153],
       [0.35766423],
       [0.26034063],
```

Figure 5.8: Normalized data

The curfew and lockdown data were also normalized by replacing yes with 1s and no with 0s.

ii. Aggregation

Aggregation involves combining two or more values of the same attribute into one value. The researcher collected COVID-19 infection data and data on ICU beds and ventilators available for each of the counties. COVID-19 cases of counties within the same province were then combined and the columns renamed to the respective provinces. Therefore, predictive models were created for each of the eight provinces to perform predictions for each of them. Below is the initial data for each of the counties before aggregation:

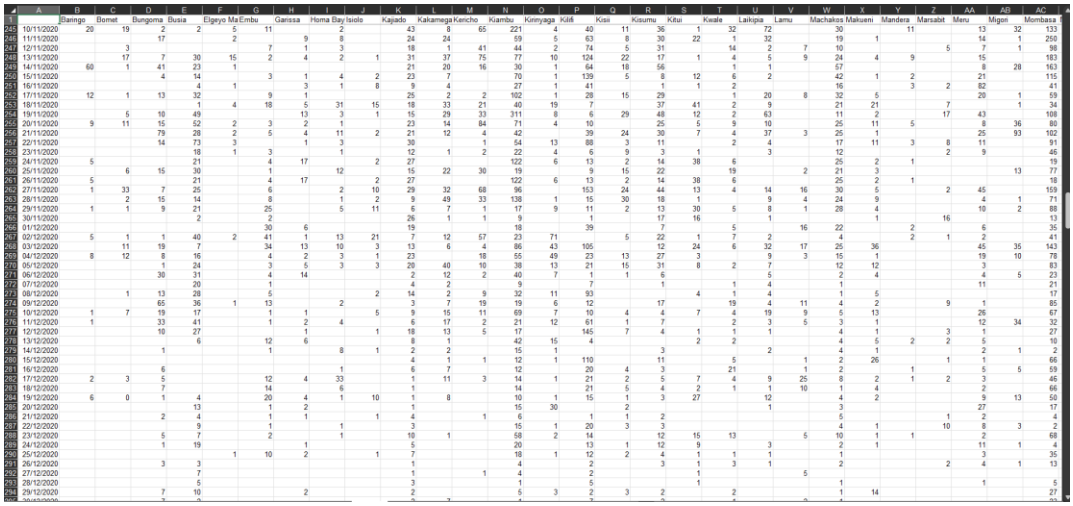


Figure 5.9: Data on COVID-19 infection cases recorded per county

Below is the corresponding data aggregated to represent the eight provinces:

	Dates	Rift_Valley	Western	Eastern	North Eastern	Nyanza	Central	Coast	Nairobi
245	10/11/2020	385	15	58	13	91	249	207	322
246	11/11/2020	116	41	56	9	68	109	376	439
247	12/11/2020	218	2	29	1	49	69	206	345
248	13/11/2020	210	77	48	13	42	102	335	649
249	14/11/2020	304	85	77	1	123	45	229	216
250	15/11/2020	97	25	88	3	17	102	262	378
251	16/11/2020	28	9	113	6	2	37	86	278
252	17/11/2020	119	56	73	1	78	144	96	358
253	18/11/2020	146	34	123	5	87	147	47	368
254	19/11/2020	229	88	87	13	106	346	118	472
255	20/11/2020	207	87	54	7	81	91	113	408
256	21/11/2020	214	119	65	5	179	76	153	410
257	22/11/2020	99	87	58	5	17	100	181	421
258	23/11/2020	34	19	27		13	39	52	229
259	24/11/2020	93	21	71	18	20	158	38	306
260	25/11/2020	157	70	25		125	71	110	252
261	26/11/2020	90	21	71	18	20	156	37	305
262	27/11/2020	235	67	111		89	163	343	546
263	28/11/2020	70	82	48		85	153	111	400
264	29/11/2020	109	53	109		22	122	106	190
265	30/11/2020	46	3	35		17	26	15	134
266	01/12/2020	68		58	8	28	70	102	212
267	02/12/2020	285	59	70	3	55	150	53	203
268	03/12/2020	120	32	178	13	75	218	291	326
269	04/12/2020	167	27	43	3	99	130	124	273
270	05/12/2020	150	70	43	5	51	90	119	207
271	06/12/2020	26	79	14	14	12	74	24	154
272	07/12/2020	44	22	13		2	12	29	77
273	08/12/2020	78	43	18		2	60	123	191
274	09/12/2020	50	109	29		20	34	130	189
275	10/12/2020	97	51	68	1	9	134	126	158
276	11/12/2020	132	91	17	2	62	54	104	211
277	12/12/2020	102	53	12	1	21	20	178	158
278	13/12/2020	78	7	30	8		83	18	142
279	14/12/2020	21	4	10		12	20	2	94
280	15/12/2020	37	1	30		15	21	182	118
281	16/12/2020	21	14	7	1	24	27	133	167
282	17/12/2020	57	20	34	5	54	42	118	222
283	18/12/2020	15	7	26	1	23	23	100	161
284	19/12/2020	88	14	72	4	20	23	66	103
285	20/12/2020	38	13	32	2	2	89	17	156
286	21/12/2020	7	7	10	1	3	12	6	68
287	22/12/2020	17	11	24		13	17	22	50

Figure 5.10: Aggregated data on COVID-19 infection cases per province

Below is the data on health resources available in each of the counties before aggregation:

	A	B	C	D	E	F	G
1	County	Total ICU beds	Total ICU beds in use	Available ICU beds	No of functional ventilators	Ventilators in use	Available ventilators
2	Nairobi City	276	166	110	167	80	87
3	Uasin Gishu	59	41	18	31	15	16
4	Mombasa	45	28	7	17	12	5
5	Bungoma	24	10	14	8	7	1
6	Kisumu	23	12	11	11	4	7
7	Nakuru	20	12	8	8	0	8
8	Kiambu	16	13	3	4	0	4
9	Nyeri	15	13	2	3	2	1
10	Kisii	14	13	1	5	1	4
11	Kakamega	9	9	0	0	0	0
12	Garissa	6	4	2	0	0	0
13	Kwale	6	3	3	0	0	0
14	Narok	6	0	6	0	0	0
15	Embu	18	5	0	0	0	0
16	Kericho	5	5	0	0	0	0
17	Machakos	5	5	0	0	0	0
18	Kirinyaga	4	4	0	0	0	0
19	Makueni	4	2	2	0	0	0
20	Murangá	3	3	0	0	0	0
21	Meru	2	2	0	2	2	0
22	Baringo	0	0	0	0	0	0
23	Bomet	0	0	0	0	0	0
24	Busia	0	0	0	0	0	0
25	Elgeyo Marakwet	0	0	0	0	0	0
26	Homa Bay	0	0	0	0	0	0
27	Isiolo	0	0	0	0	0	0
28	Kajiado	0	0	0	0	0	0
29	Kilifi	0	0	0	0	0	0
30	Kitui	0	0	0	0	0	0
31	Laikipia	0	0	0	0	0	0
32	Siaya	0	0	0	0	0	0
33	Turkana	0	0	0	0	0	0
34	Migori	0	0	0	0	0	0
35	Nandi	0	0	0	0	0	0
36	Trans Nzoia	0	0	0	0	0	0
37	Nyandarua	0	0	0	0	0	0
38	Taita Taveta	0	0	0	0	0	0
39	Nyamira	0	0	0	0	0	0
40	Vihiga	0	0	0	0	0	0
41	Lamu	0	0	0	0	0	0
42	Marsabit	0	0	0	0	0	0
43	Tharaka-Nithi	0	0	0	0	0	0
44	West Pokot	0	0	0	0	0	0
45	Mandera	0	0	0	0	0	0
46	Tana River	0	0	0	0	0	0
47	Samburu	0	0	0	0	0	0
48	Wajir	0	0	0	0	0	0

Figure 5.11: Data on COVID-19 health resources per county

Below is the aggregated data on the health resources in each of the provinces:

	A	B	C
1	Province	Tota ICU beds	Total functional ventilators
2	Nairobi	276	167
3	Rift Valley	90	39
4	Coast	51	17
5	Western	33	8
6	Nyanza	37	16
7	Central	38	7
8	North East	6	0
9	Eastern	29	2

Figure 5.12: Aggregated data on COVID-19 health resources per province

5.2.1.3 Model Training

To train a model to carry out prediction of COVID-19 cases in the future, three major steps are involved. The first is to collect the sample data, which is comprehensively discussed in this study under Data extraction. The second involves learning the model. This involves taking some data with known relationships between the variables and creating a model for those relationships. The trained model is expected to give outputs based on the unseen inputs, making predictions which is the third step in training a model. The accuracy of the predictions is based upon how well the model has been trained.

The researcher adopted several algorithms in the study. The different algorithms allowed the researcher to settle on the best based on the one with the least error values of MSE, MAE and RMSE. These include Multiple Linear Regression using Neural Networks, Linear Regression, Decision Tree, and Random Forests. Those that were Time Series based include Bayesian Polynomial Regression, Traditional ARIMA, LSTM Networks and FFNN.

Multiple Linear Regression Using Neural Networks

This is a statistical strategy that predicts the result of a response variable by combining numerous explanatory variables. The researcher used data on COVID-19, Recoveries, Deaths, Hospital admissions, Vaccinations, Lockdown, Curfew, Patients in the ICU, Patients on Ventilatory support and Patients on Supplemental Oxygen to predict the number of COVID-19 infection cases. A heatmap correlation matrix was used to determine the relationship between each of the variables. Below is its output:

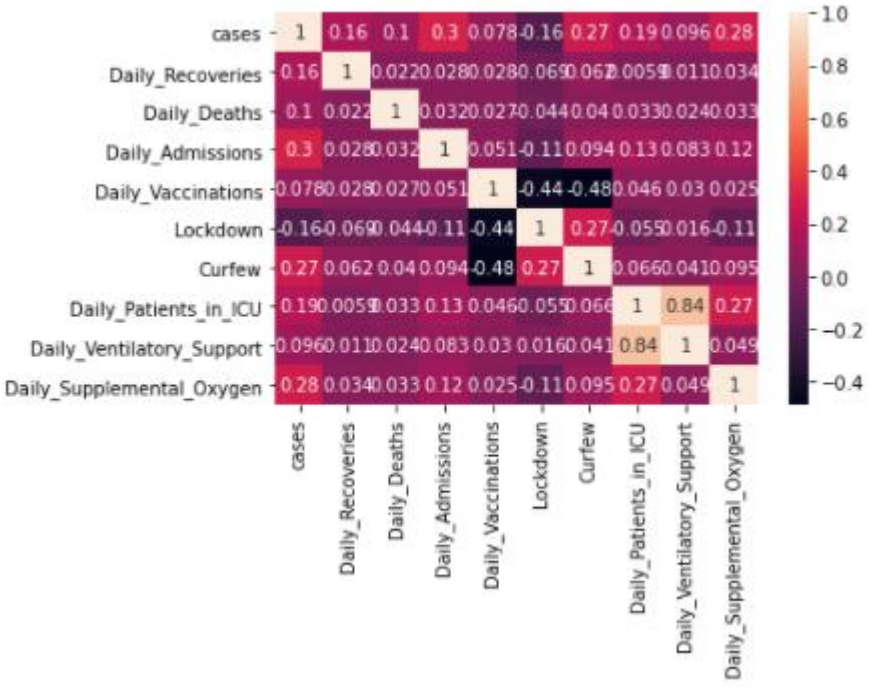


Figure 5.13: Correlation Matrix for Multiple Linear Regression

The correlation matrix unravels that there is no strong relationship between the number of cases recorded and the rest of the variables relating to the coronavirus dataset. The highest level of confidence is 0.28 observed between the daily number of recorded cases and the daily number of patients getting supplemental oxygen, which is relatively low. This is one of the reasons why the researcher ended up dropping Multiple Linear Regression as an approach prediction. The researcher split the data into 80% for the training set and 20% for the testing set and trained the model with 13 hidden layers and a dropout layer. The dropout layer in neural networks helps to avoid overfitting. Below is a graph showing the training and validation losses:

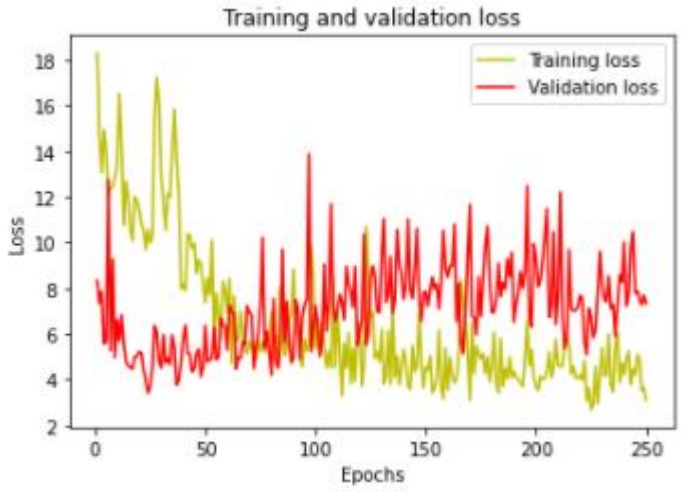


Figure 5.14: Training and Validation loss from Multiple Linear Regression

The validation loss is higher than the training loss even after adding a dropout layer in the neural network, an indication of overfitting hence the generated model is not generalised enough to work on real life datasets. This is the second reason that led to the researcher dropping the model. Moreover, the MSE, MAE and RMSE values recorded were relatively high. Below is a graph showing the test data against the predictions made. The blue dots represent the test data while the red dots represent the predictions. Despite the presence of a linear increment in the number of COVID-19 cases on the test data, the predicted cases are scattered on the graph and only two instances are correctly predicted, which accounts for 0.02 accuracy in the predictions. This is the third reason that led to the researcher dropping this statistical approach to prediction.

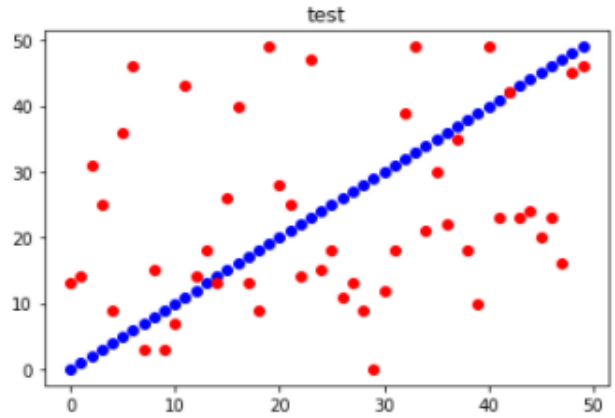


Figure 5.15: Real vs Predicted values for Multiple Linear Regression

Linear Regression, Decision Tree, and Random Forest

The researcher adopted the three regressors to predict the number of COVID-19 infection cases. A split of 80% training data and 20% testing data was used. The researcher considered daily number of recorded cases and the daily number of patients getting supplemental oxygen as variables for the predictive model. For each, high values of MSE, MAE and RMSE were recorded. Moreover, a graph of the test values against those predicted had four or less instances of valid predictions which accounts for 0.03 accuracy in the predictions.

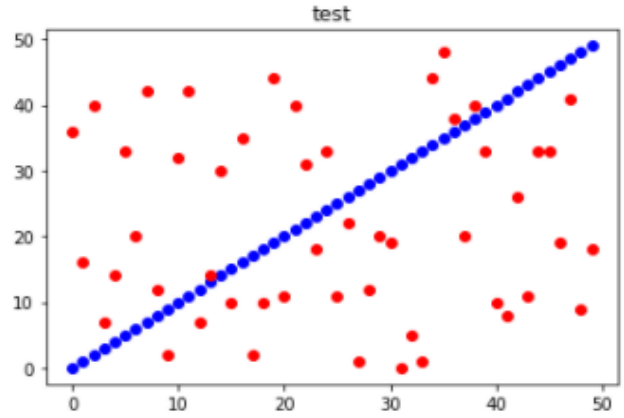


Figure 5.16: Real vs Predicted values for Linear Regression

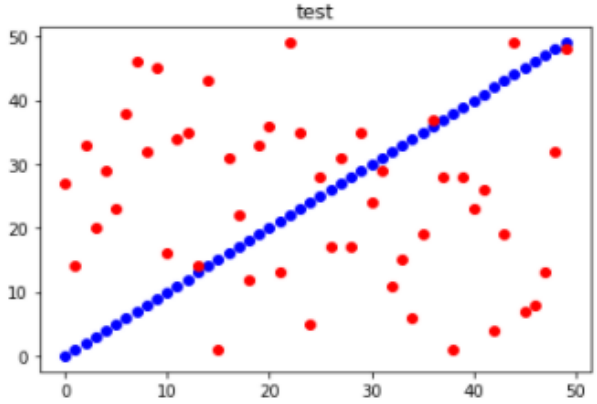


Figure 5.17: Real vs Predicted values for Decision Tree

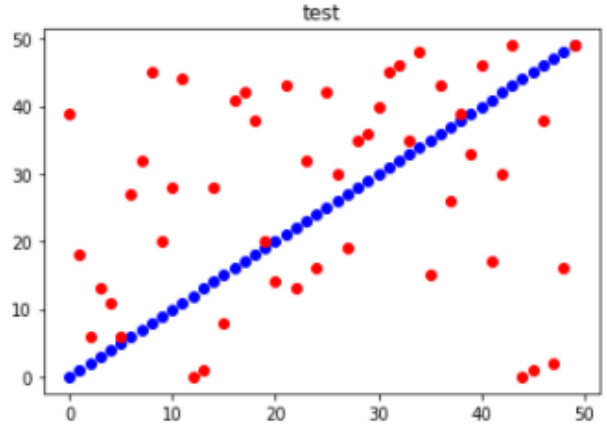


Figure 5.18: Real vs Predicted values for Random Forest

After the discovery that there is no strong relationship between the coronavirus infection cases and the rest of the variables, forecasting based on the multiple parameters is challenging. Below is a plot of the number of COVID-19 infection cases reported in each of the provinces over a set period.

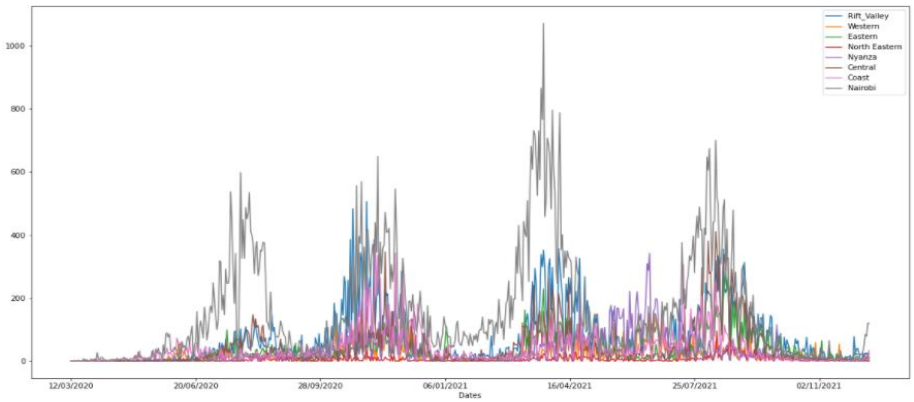


Figure 5.19: COVID-19 cases recorded per province over time.

A time series is a set of observations arranged chronologically. Therefore, time series analysis involves plotting and identifying patterns within the time series period and time series forecasting is the process of using past observations to predict what will happen in the future if the same patterns remain. Therefore, the direction in which the data is changing can be identified with good forecasting. The graph above depicts some seasonality in the number of positive cases recorded over time. Seasonality is one of the features necessary to allow for time series prediction. Therefore, with the seasonality in data and lack of strong relationship between the variables, the researcher opted to consider the number of cases as a single variant (univariant) for time series forecasting. Time series algorithms adopted in this study include Traditional ARIMA, Bayesian Polynomial Regression, FFNN and LSTM.

Traditional ARIMA

Traditional Autoregressive Integrated Moving Average model is well known to for time series forecasting. It is a combination of two models: the AR (AutoRegressive) and MA (Moving Average). ARIMA model requires for the data to be stationary. Therefore, upon Data pre-processing, the researcher started by doing a stationarity check on the data. Stationarity means that the data should have a constant mean and variance between the values of the variables and autocovariance that doesn't depend on time. Stationarity check was done by plotting the rolling statistics as below:

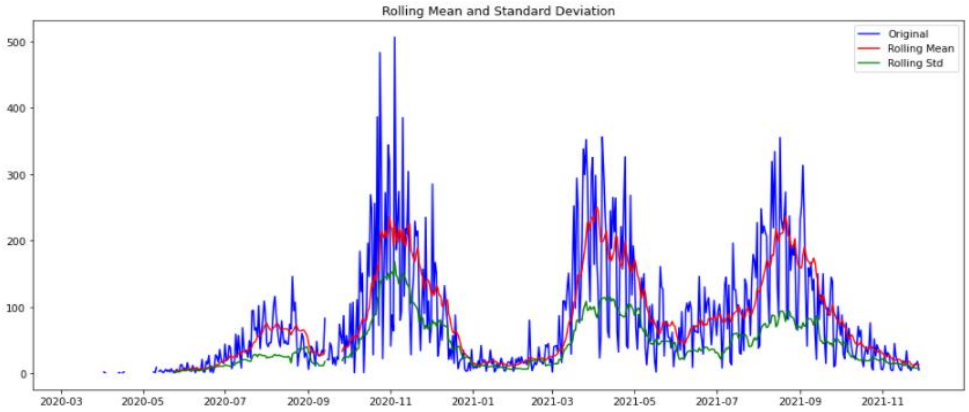


Figure 5.20: Rolling statistics of the dataset

From the graph, it is evident that the data is not stationary. The data was made stationary by carrying out a log transformation. This transformation was used to create the predictive model. Once the predictions were made, the values were converted back to the original state by doing an inverse to the log. The model gave an RMSE value of nan and below is the prediction graph. The nan RMSE value made it difficult for the researcher to evaluate the model, hence dropping it.

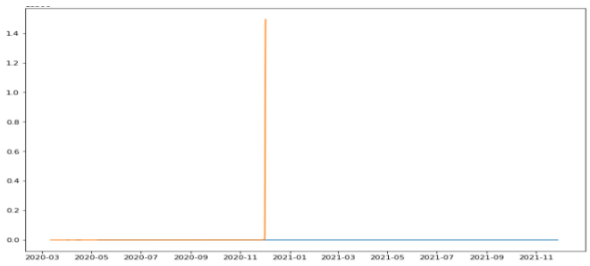


Figure 5.21: Real vs Predicted values for Traditional ARIMA

Bayesian Polynomial Regression

This algorithm is also applied in time series forecasting. The researcher imported the polynomial features from the Sklearn library for model training. The model was trained with different values for cross validations and iterations. The lowest MSE, MAE and RMSE values were attained with 10 cross validations and 1000 iterations. However, generally, the values were still high. The corresponding graph on test data and predictions is as below:

Figure 5.22: Real vs Predicted values for Bayesian Polynomial Regression

With the aim to attain very low error values for MSE, MAE and RMSE, the researcher opted to adopt deep learning algorithms for time series prediction. Deep learning neural networks, in general, can learn arbitrary complex input-output mappings automatically and handle many inputs and outputs. Neural nets have successfully solved loan appraisal, signature identification, time series forecasting, classification analysis, and numerous other difficult pattern recognition problems. While directly writing down a set of rules for such pattern recognition tasks is typically difficult or impossible, a neural network can be trained to create a solution using raw data (Tang & Fishwick, 1993). Moreover, many research publications have successfully utilized deep learning approaches to time series forecasting.

Feed Forward Neural Networks

This is one of the most popular neural net paradigms. The neurons are usually arranged in layers. The data was split into a third for testing and two thirds for training. The model was trained with different number of layers ranging from three to ten and 100, 200 and 250 epochs. There was a great improvement in the MSE, MAE and RMSE values with seven layers and 200 epochs. Moreover, the model did not show any overfitting as it was the case with multiple linear regression.

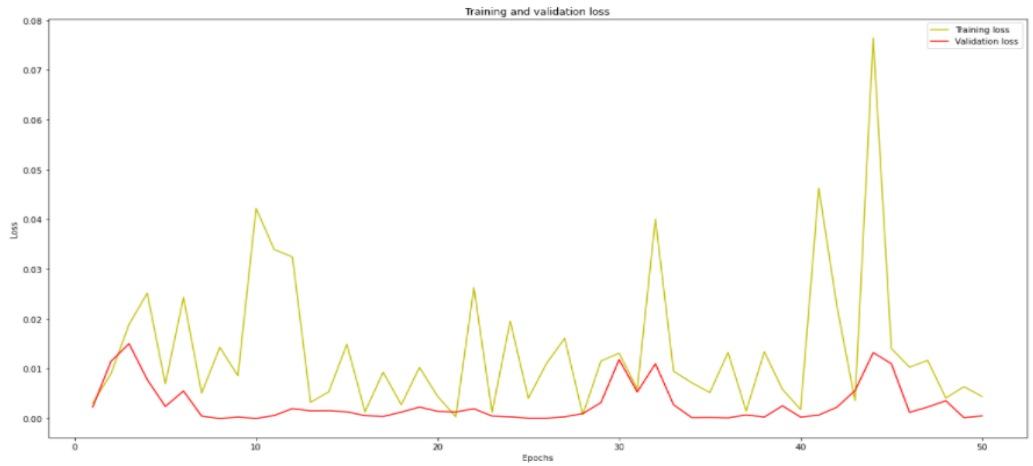


Figure 5.23: Training and Validation loss from FFNN

The training and validation losses are relatively low and tended to converge at certain points. The validation loss was also lower than the training loss. This was a great improvement in the performance compared to the prior models. A plot of the predictions and test data was also good.

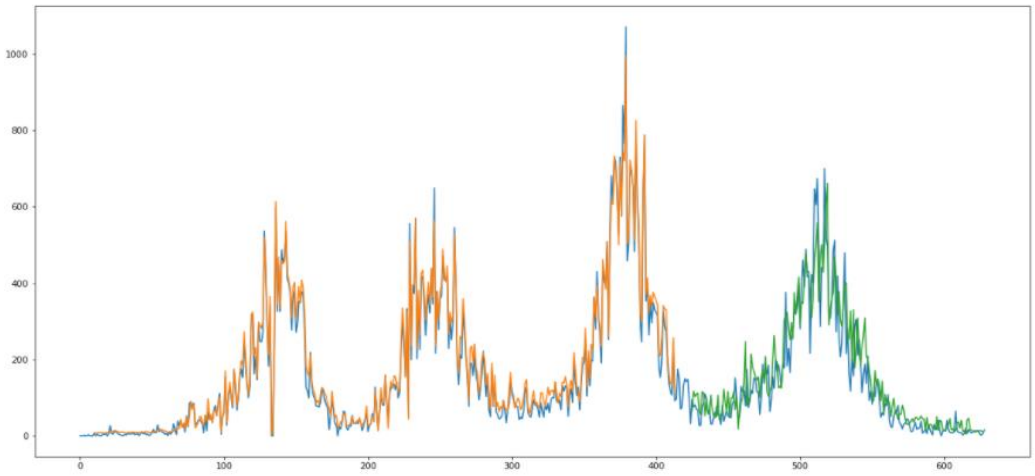


Figure 5.24: Real vs Predicted values for FFNN

However, the researcher still aimed to attain lower error values and retain a good fit of the predictions against the test data.

Long Short-Term Memory Networks

LSTM networks are well-known for time-series forecasting. This was based on only one attribute, which is the number of COVID-19 infection cases recorded over time, hence a univariate variable in the training of the model. This data is used to create the predictive model using LSTM networks to predict the number of infection cases expected to be recorded in the next 60 days. During the training, researcher obtained the lowest error values of 0.0784, 0.026 and 0.028 for MSE, MAE and RMSE respectively. This was with 70 epochs with 10 steps per epoch, three hidden layers with 256, 128 and 64 neurons respectively, and two dense layers. Moreover, a good fit of predictions vs real values graph was obtained as well. The data was split into 10% for testing and 90% for training. The model from this algorithm gave the expected results and performed best. Below are the loss and prediction graphs:

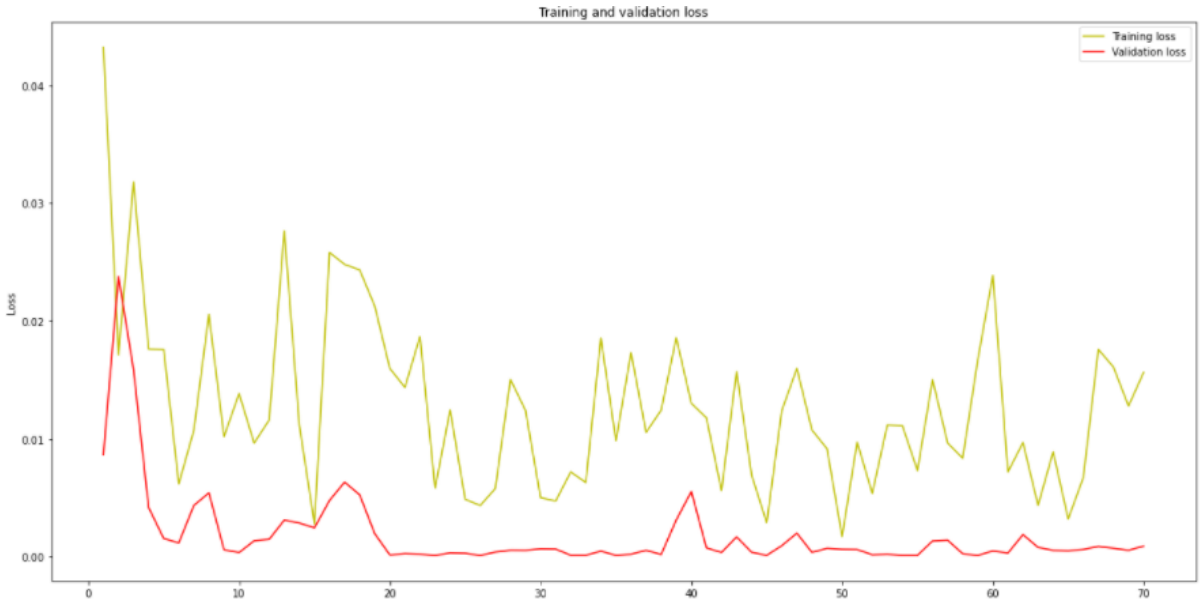


Figure 5.25: Training and Validation loss from LSTM

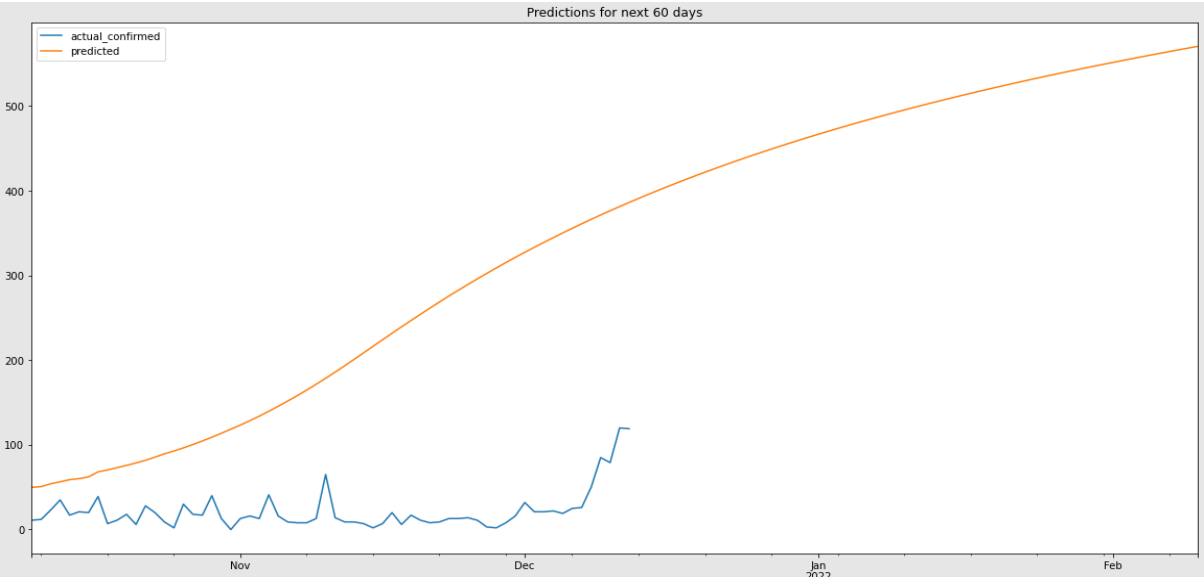


Figure 5.26: Real vs Predicted values for LSTM

5.2.2 Development of the Visualization Dashboard

A dashboard is a graphical representation of data that is used to monitor conditions and/or aid comprehension. (Farkas et al., 2020). The aim of this study is to visualize the predictions of COVID-19 infection cases against the health care resources available in terms of ICU beds and ventilators. This will aid in decision making by determining whether there is a need to procure more resources or not. During the development of dashboards, there are two major design perspectives to consider. Dashboards are a visual genre that provide a visual representation of data in the form of a tiled layout of large numbers or simple charts. The functional genre, on the other hand, provides an interactive visual representation that allows for real-time monitoring of dynamically updating data. There are three major factors that influence the design perspective that should be chosen. (Sarikaya et al., 2019).

1. Purpose of the dashboard
2. Target audience
3. Expected visual features and interactivity

Purpose of the dashboard

A dashboard's visual design and functional affordances are determined by its intended purpose. A dashboard can be used to provide decision support or to facilitate communication and learning. This research is geared toward decision support. As a result, we can have either operational dashboards that represent the present and recent past in terms of easily measurable indicators that can be linked to the entities responsible, or strategic dashboards that present a wider view of actionability, integrating a collection of high-level variables to guide decision-making over a longer time frame. As a result, operational dashboards are updated in real time, whereas strategic dashboards are updated over short or medium time periods. (Sarikaya et al., 2019). This study focuses on creating strategic dashboards that give predictions on COVID-19 cases over a period and are updated a few days towards the end of the predicted periods.

Target audience

The visual and functional components of a dashboard should reflect the intended audience, their domain and visualization experience as well as how they relate with the data. Depending on how the dashboard is intended to circulate, the audience can be either be categorised as being public, social, organisational, or individual. This study mainly targets individuals in the health sector including the workers in the Ministry of Health and the government of Kenya responsible for making decisions on availability and utilisation of health resources. This therefore captures a social audience whereby access to the dashboard is controlled and only availed to relevant individuals who are the doctors, nurses, Ministry of Health, and the Government. The visualization literacy is also a factor to consider on the target audience. The researcher adopted basic visualizations in this study to ensure that the information conveyed in them is easily comprehensible. Therefore, the researcher chose to use area graphs to represent the data (Sarikaya et al., 2019).

Expected visual features and interactivity

Different dashboards offer varying levels of interactivity. These include construction and composition, in which users can adjust the views, multipage, in which the dashboards support tabbed layouts, and interactive interfaces, in which users perform filtering within the views and interactive interfaces, in which cross-highlighting on

particular data items is possible, or in which traditional interactive components such as dropdowns are present (Sarikaya et al., 2019). The researcher has an interactive dashboard in this study.

In addition to the three factors, some additional semantics such as the dashboard being updateable is considered in the study. Once the datasets connected to the dashboard are updated, the visualizations also update accordingly. The researcher considered above features in the development of the visualization dashboard.

5.2.2.1 Installation of Tableau Desktop

Tableau is an easy-to-use tool that connects users to a variety of data sources and allows them to create data visualizations using a simple drag-and-drop interface to create charts, maps, dashboards, and stories (Martinez et al., 2016). The researcher started by downloading and installing Tableau Desktop Personal version 21.4 (64-bit). An email address was required for setting up an account. To obtain a 1-year free product key with support for students and instructors in academic programs, the researcher submitted proof papers by email. The Tableau Personal version provides additional features compared to Tableau Public, such as the capability to connect to Tableau server and access to Tableau tutorials.

5.2.2.2 Connecting to Sample Data

Sample data refers to the data collected during and after the development of the predictive model. These are COVID-19 infection cases in each of the eight provinces, the predictions for each of the provinces, Cumulative COVID-19 infection cases recorded nationally, the death cases, recoveries, vaccinations, and resources available to contain and mitigate the spread of the virus. The data was stored in an excel file. The data on it was extracted to Tableau by connecting the specific excel files to Tableau Desktop via the Connections section on Tableau start page. The attached datasets appear in the upper left corner, and the sheets are listed at the bottom of the datasets. If the predictions for each of the provinces are selected and dragged to the 'Drag sheets here' section, the prediction sheets will be added to the section, and the data will be displayed below. To merge the confirmed cases recorded in each province with the corresponding predictions, the predictions are dragged to 'Union' as shown below:

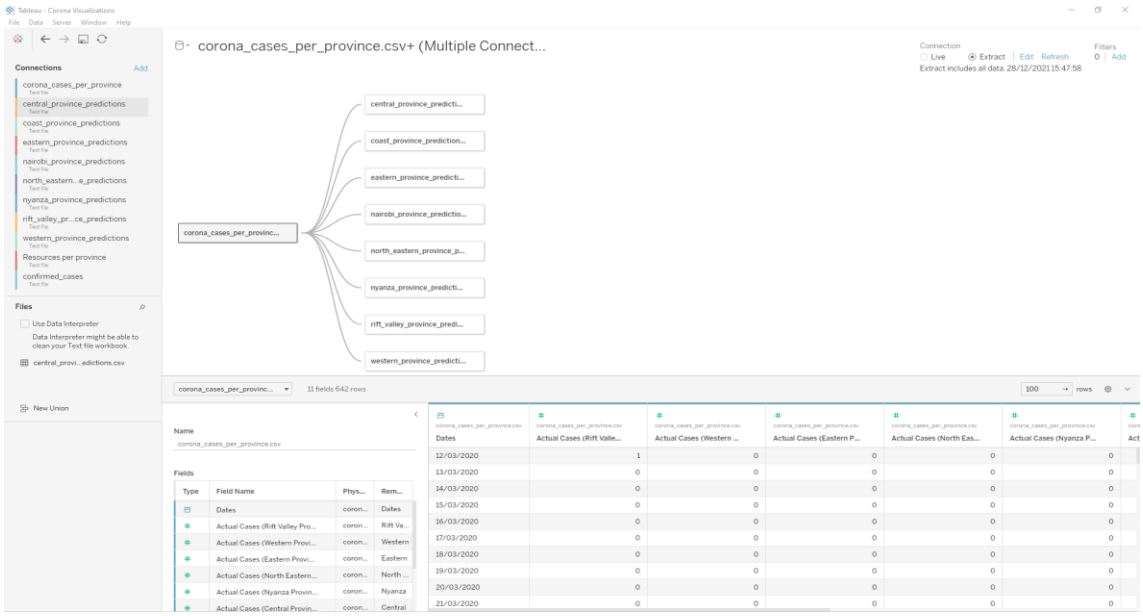


Figure 5.27: Importing data to Tableau

5.2.2.3 Creating a Worksheet

The researcher then moved on to create a new worksheet by clicking next to the data source tab at the bottom. The Tableau workspace consists of the following functions and items:

1. Menus: files, data, worksheets, dashboards, stories, etc.
2. Toolbar: undo, redo, swap rows and columns, show me, etc.
3. Data Pane: Tableau assigns data fields of the connected dataset in the left section. 'Formatting' and 'Analytics' panes are also available. - 'Dimensions' in the upper section include discrete categorical data such as type or date. - 'Measures' in the lower section contain continuous data.
4. Shelves and Cards: page shelf, filter shelf, mark cards, and top space with row and column shelves.
5. View: The right space to create a view (table, graph) by adding fields.
6. Sheet Tabs: Create or move to created worksheets, dashboards, and stories.

Separate worksheets were created for the eight provinces, each capturing the infection cases since the onset of the virus in Kenya and a prediction for the next 60 days. Below are the steps that were followed for each of them:

1. Drag the 'Dates' column under the corona cases per province excel and drop it under 'Columns' on Tableau.
2. Drag the 'Actual Cases (Nairobi Province)' column under the corona cases per province excel and drop it under 'Rows' on Tableau.
3. Select the type of graph to use. In this study, the researcher chose area graphs as they are easy to interpret and represent the data in subject effectively.
4. Drag the 'Dates' column under the Nairobi province predictions excel and drop it under 'Columns' on Tableau.
5. Drag the 'Predicted Cases (Nairobi Province)' column under the central province predictions excel and drop it under 'Rows' on Tableau.
6. Format both Dates under Columns on Tableau to show in 'DD/MM/YY'
7. Format both Actual and Predicted Cases under Rows on Tableau to display as Continuous variables and the Measure as a Sum.
8. Select 'Dual Axis' along the y-axis of the plot to have both the actual cases figures and predictions share the same y-axis calibration
9. Select 'Dual Axis' along the x-axis of the plot to have both the dates of the confirmed cases and those of the predictions appear on the same x-axis.
10. Synchronize the x and y axis for the four inputs to share the same graph for better visualization and graph interpretation.
11. Set the Measure Names as Actual Cases and Predicted Cases.
12. Add a reference line to set thresholds for COVID-19 resources present to curb the spread of the virus for that specific province. A red zone is set to serve as an alert whenever the COVID-19 prediction cases made are more than the ICU beds available while a green zone is set to represent a section where the ICU beds are enough to cater for the cases predicted.
13. Add a constant line to represent the availability of Ventilators in the province. If the number of predicted cases goes above the Ventilators threshold, that means that ventilators are not enough in case the infected individuals need ventilator support while if below, they are enough.

The above steps were followed to visualize the COVID-19 infection cases and corresponding predictions for each of the provinces. Below is a sample output for Nairobi province:

Figure 5.28: Visualization of Confirmed Cases, Predictions and determinants for ICU beds and Ventilator availability on Tableau Worksheet

Worksheets with visualizations on the national COVID-19 resources (ICU beds and Ventilators), Cumulative Infection, Recoveries, Death, and Vaccination cases since the onset of the virus, were also created to understand the general trend in each of them and discover any relationships between them. Below is a visualization of COVID-19 resources using bar graphs:

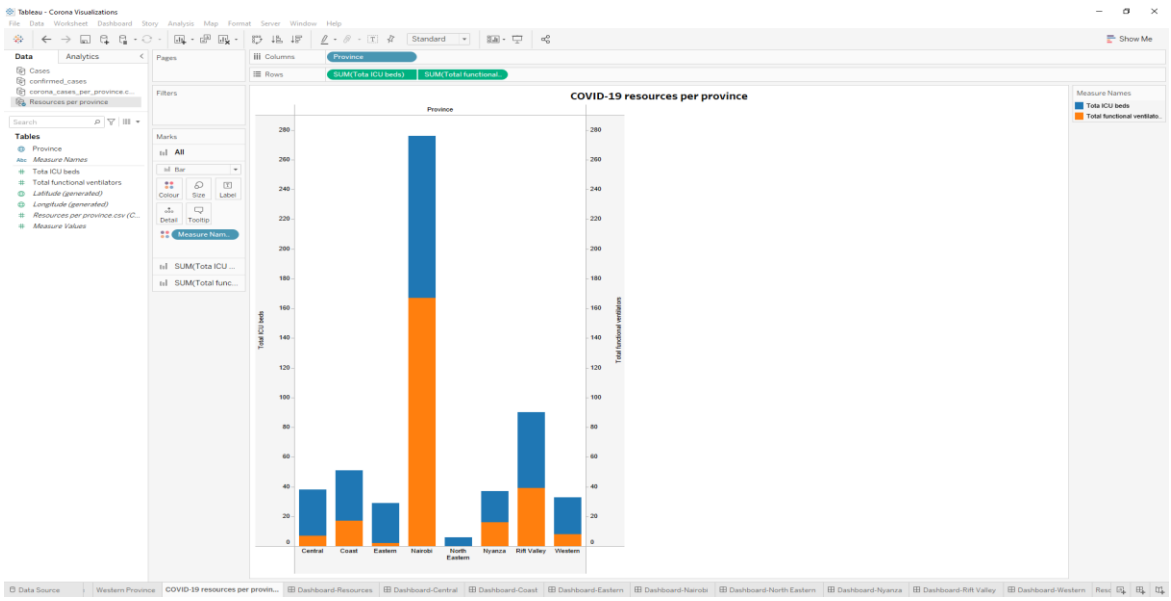


Figure 5.29: COVID-19 Resources per province

Below is a visualization of Infection and recovery cases using area graphs:

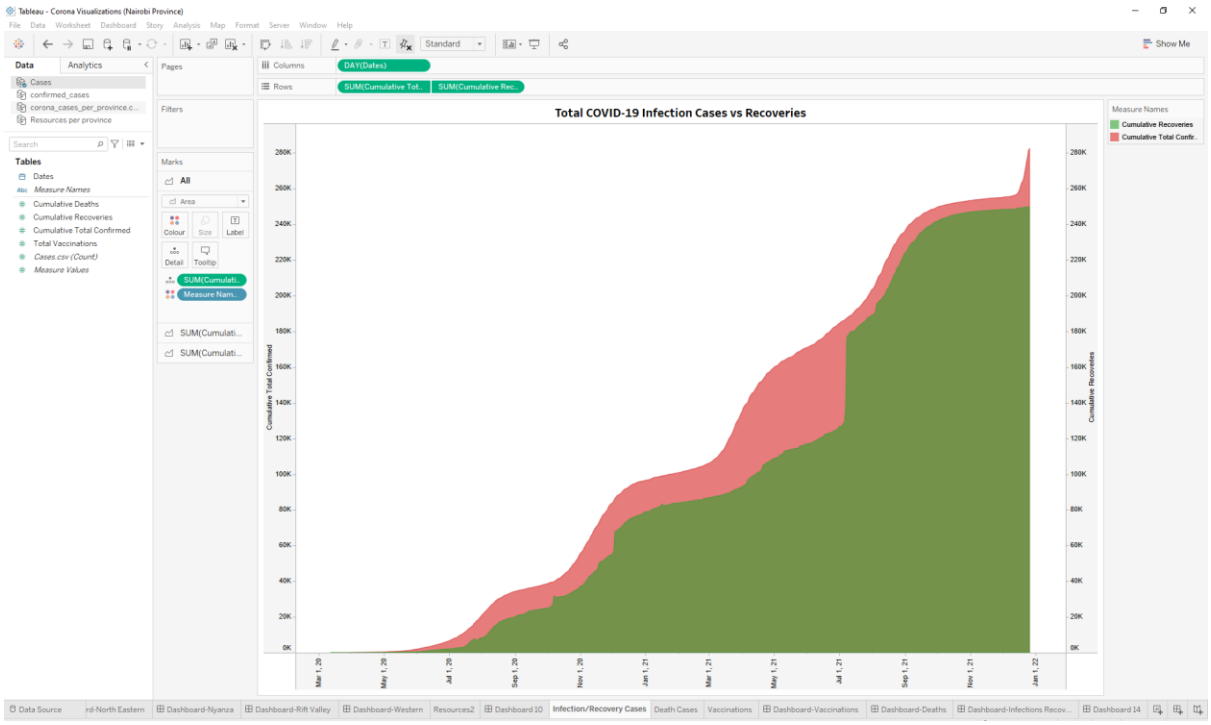


Figure 5.30: Cumulative Infection Cases vs Recoveries on Tableau Worksheet

Below is a visualization of Death cases using area graphs:

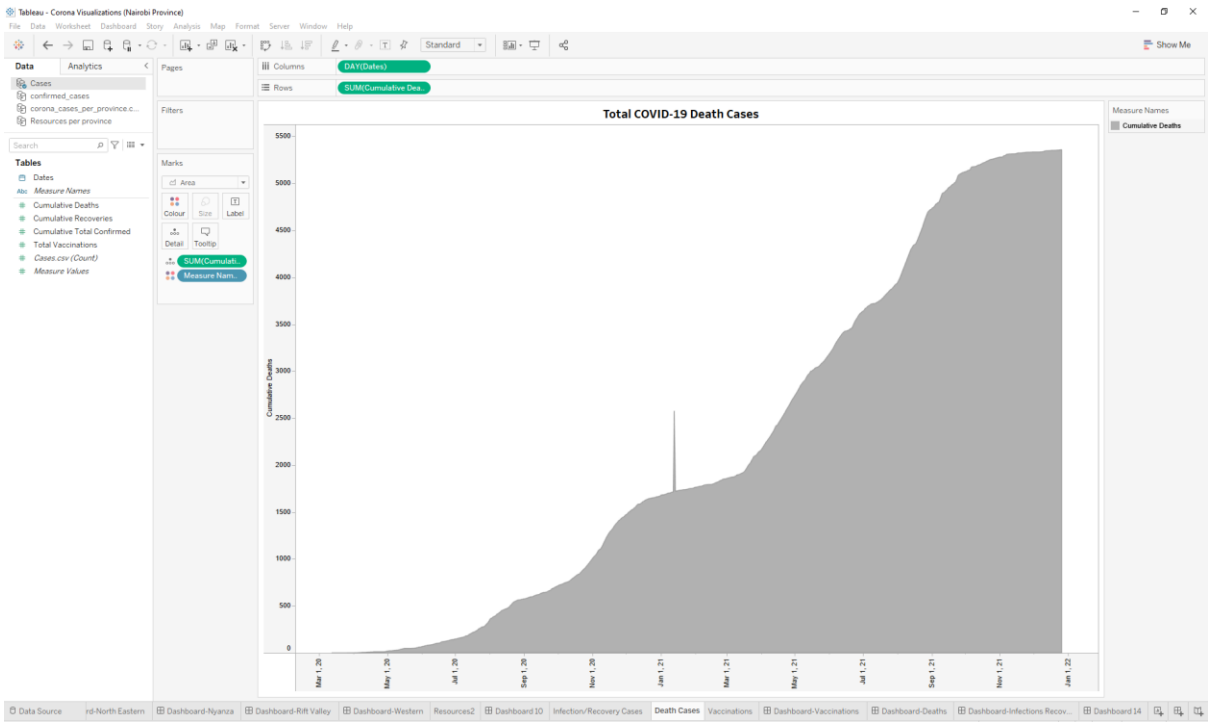


Figure 5.31: Cumulative Death Cases on Tableau Worksheet

Below is a visualization of Vaccinations administered using area graphs:

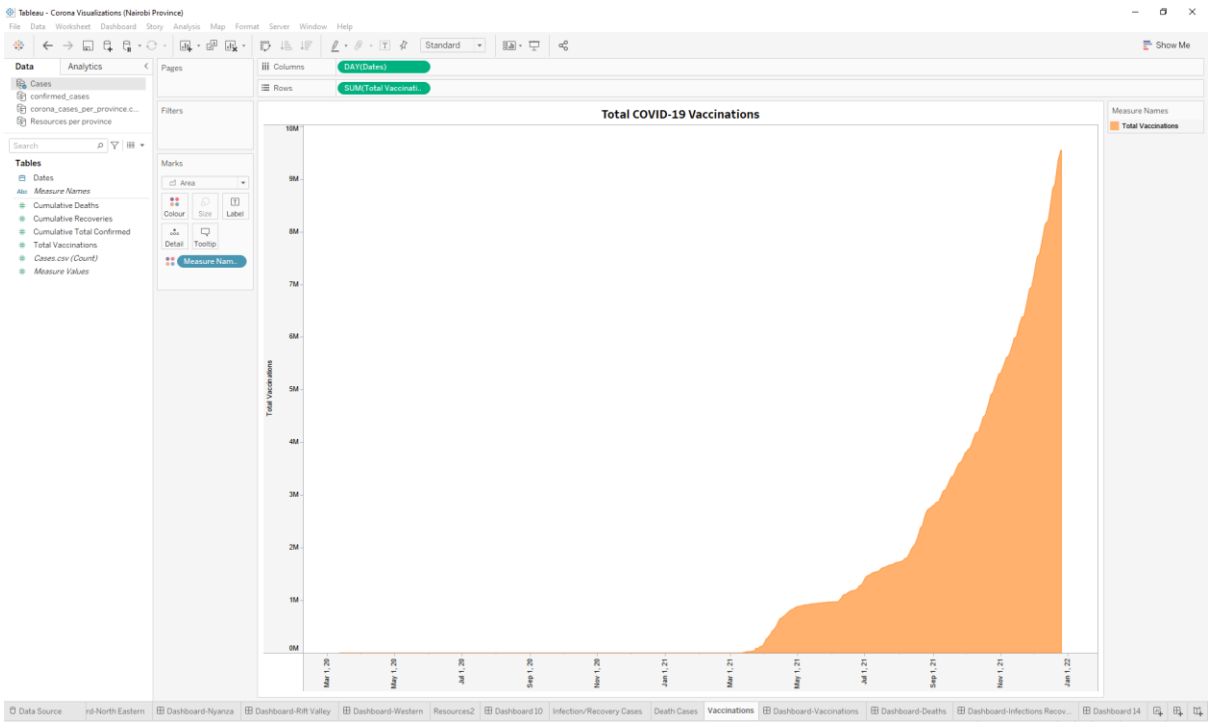


Figure 5.32: Cumulative Vaccinations Administered represented on Tableau Worksheet

5.2.2.4 Creating a Dashboard on Tableau Desktop

The researcher then created the dashboards corresponding to each of the worksheets discussed. This was done by clicking on the dashboard menu on the top left corner of Tableau Desktop. This opened a blank view with the writings 'Drop sheets here' as shown below:

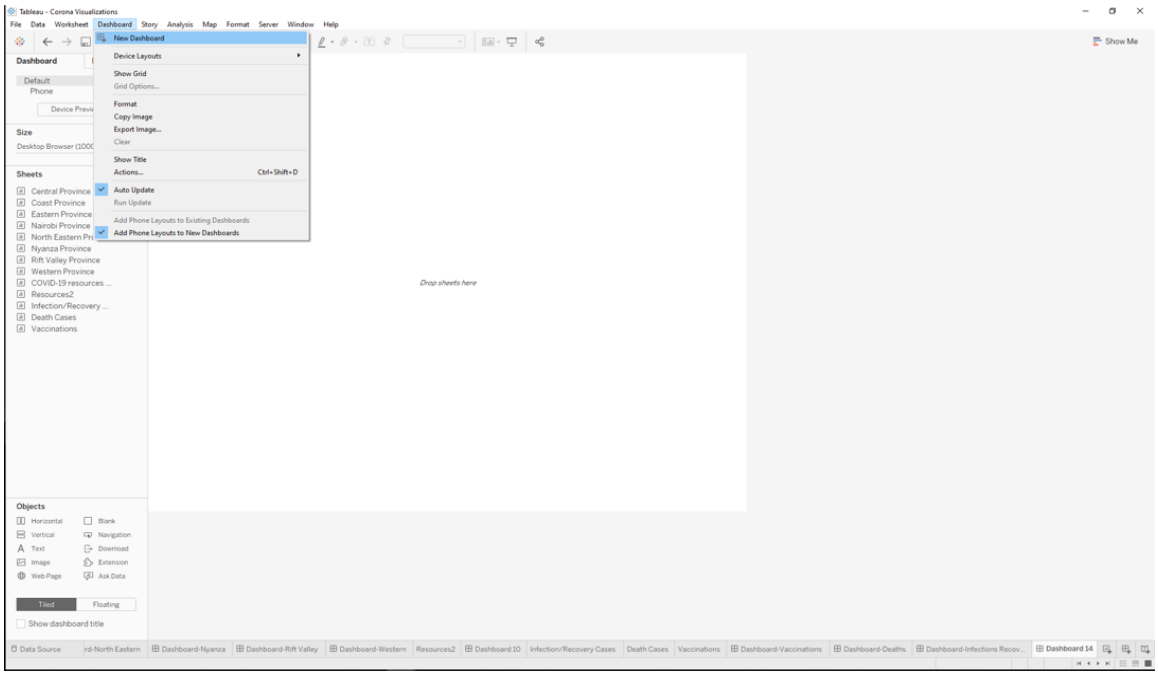
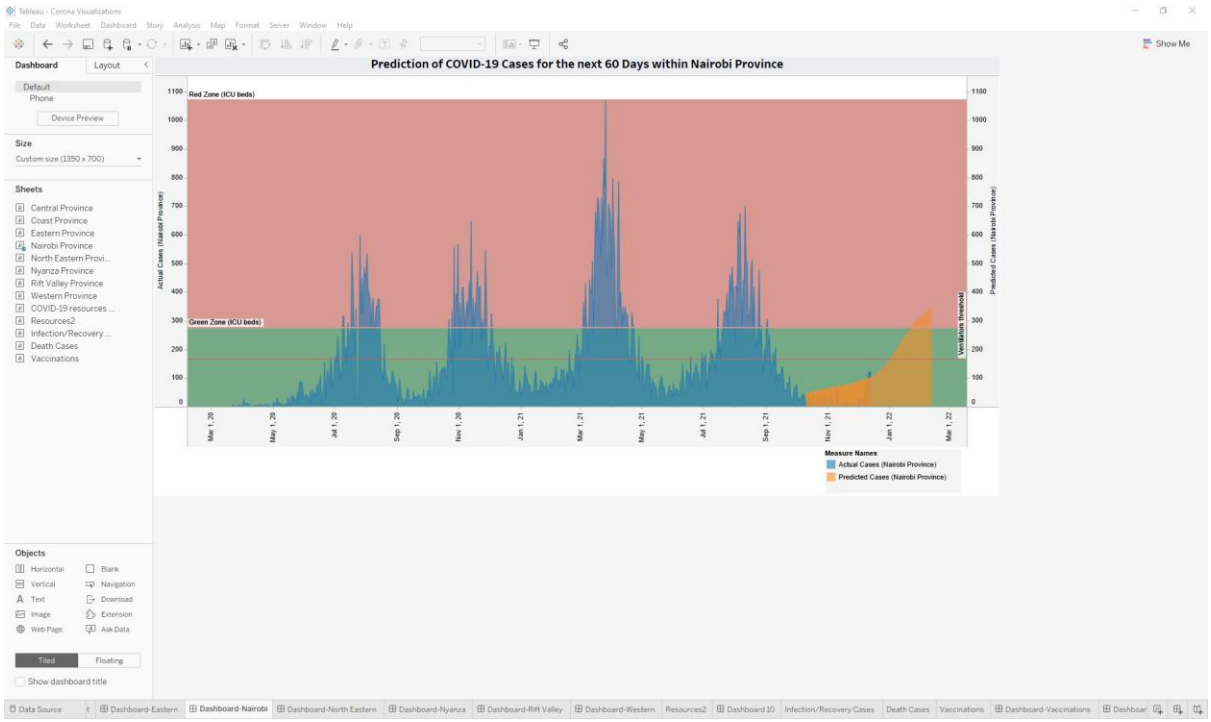


Figure 5.33: Creating a Dashboard on Tableau Desktop

At the left of the blank dashboard, we have worksheets. The worksheets correspond to each of those created in the previous section. Dragging the sheet names 'Nairobi Province' allowed the researcher to import the visualization under the Worksheet to the Dashboard. The researcher then adjusted the size of the dashboard and set it to a custom size of 1350 pixels for the width and 700 pixels for the height. Below is the output:



The same step was followed for the rest of the worksheets created, to get their corresponding dashboards.

Figure 5.34: Visualization of Confirmed Cases and Predictions on Tableau Dashboard

5.2.2.5 Transferring the Visualizations on Tableau Desktop to Tableau Public Online

Step 1: The researcher started by creating an account on the online Tableau public using an email address different from the one used to download and install Tableau Desktop. Using a similar account would now allow the researcher to save the visualizations online.

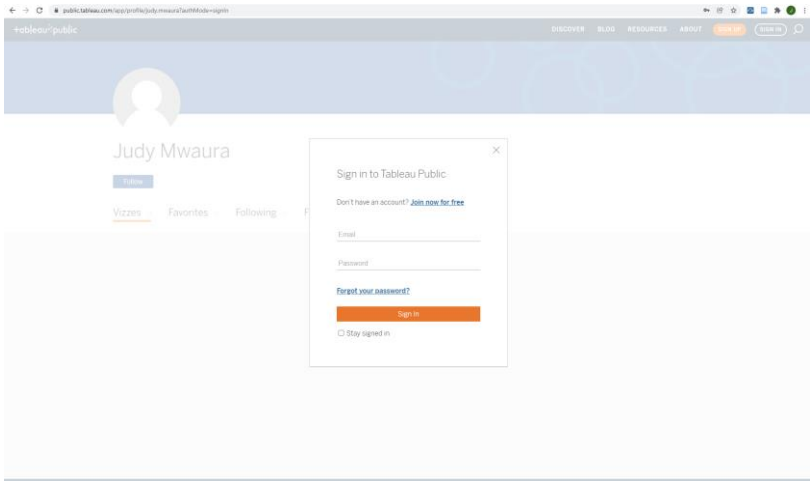


Figure 5.35: Creating an account on Tableau Public

Step 2: The researcher then updated the visualization permissions under account settings by disabling downloads and copies as well as hiding published Visualizations on Profile as this is sensitive information.

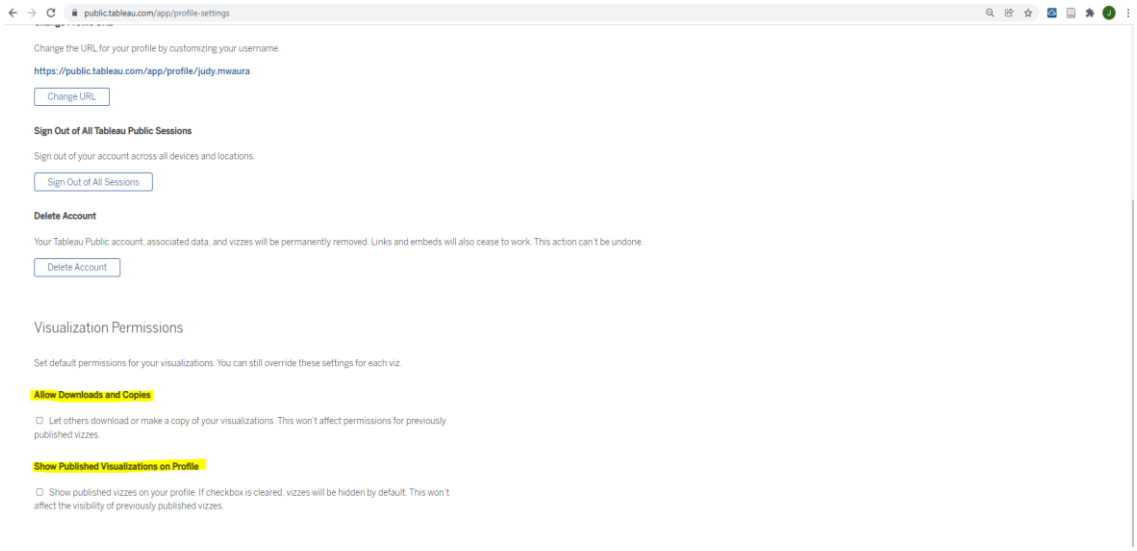


Figure 5.36: Setting Visualizations Permissions

Step 3: On the Menu on the left side of the desktop dashboard, the researcher clicked on Server >> Tableau Public >> Save to Tableau Public As to save the visualizations online as shown below:

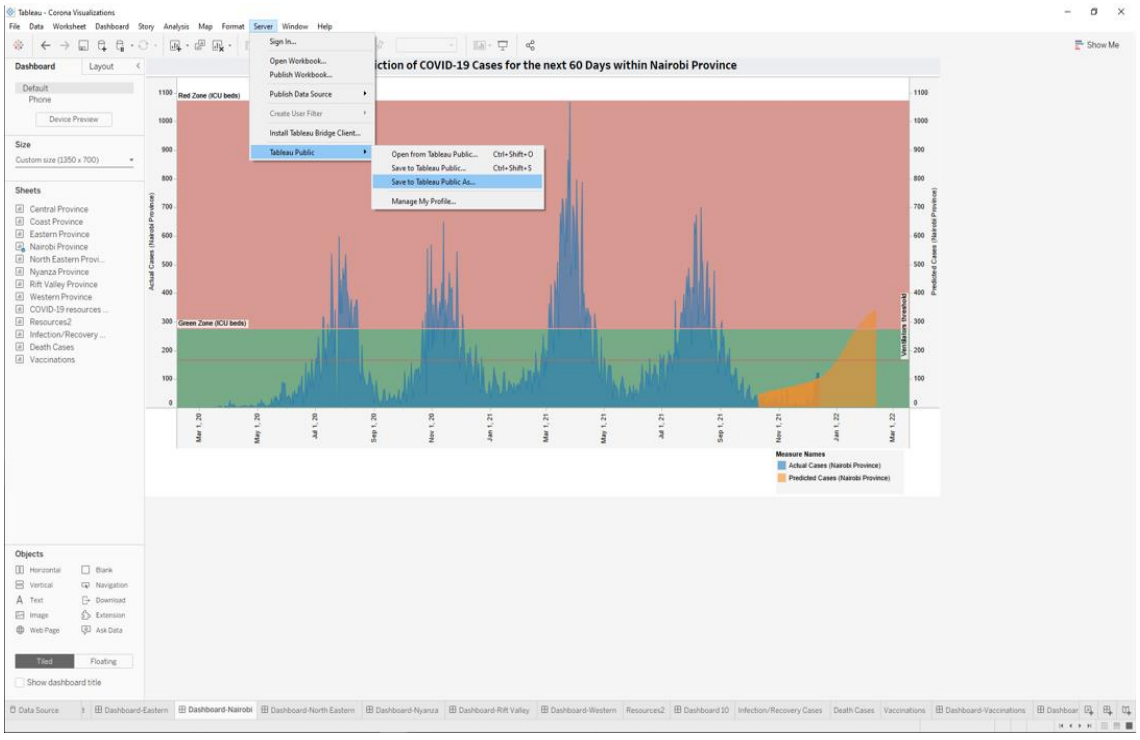


Figure 5.37: Transferring the Visualizations on Tableau Desktop to Tableau Public Online

Step 4: Upon completion of Step 3, the researcher was prompted to sign into Tableau Public using the newly created account as shown below:

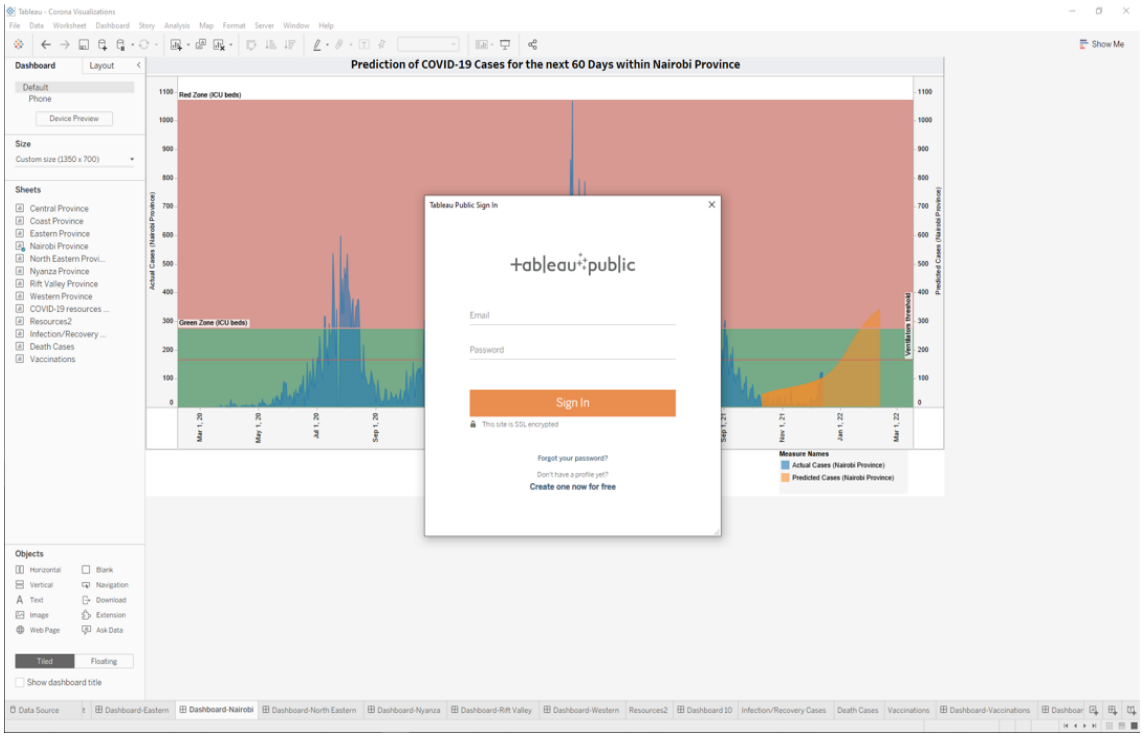


Figure 5.38: Signing into Tableau Public from Tableau Desktop

Step 5: Upon signing in, a prompt to rename the dashboard online popped. The researcher chose to retain the name used on the desktop dashboard:

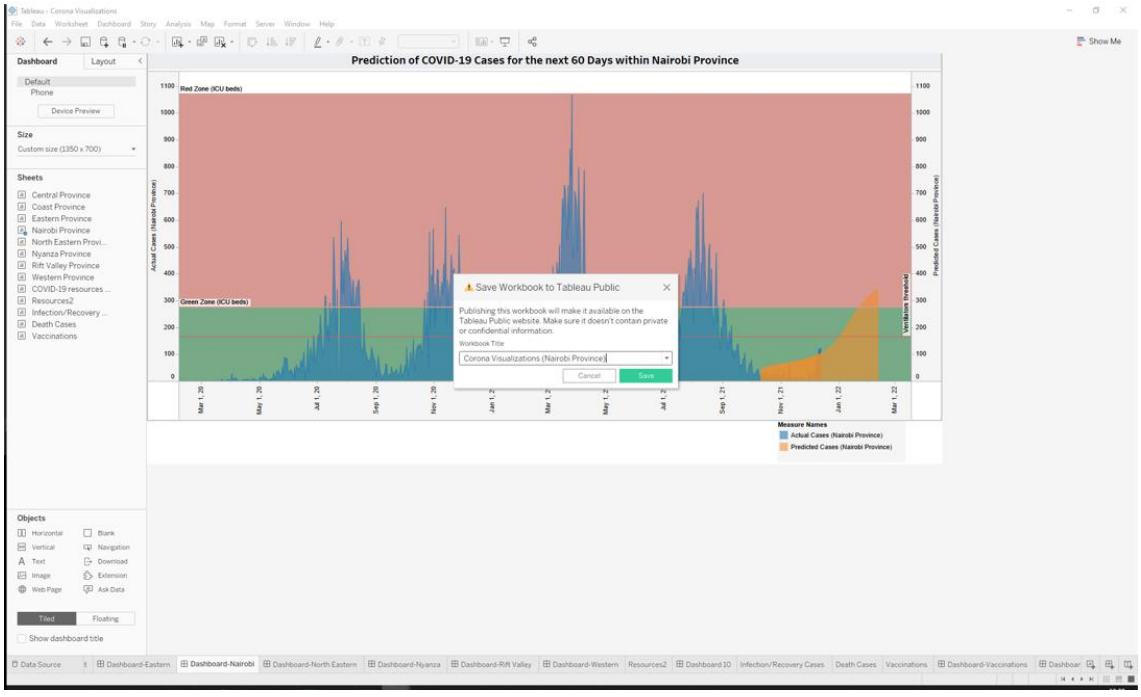


Figure 5.39: Prompt to rename the dashboard

Step 6: Once the researcher clicked on 'Save' on the figure above, the same dashboard opened on Tableau Public online as shown below:

Figure 5.40: Dashboard view on Tableau Public online

The six steps were followed for the rest of the visualizations to have all on online Tableau Public as shown below:

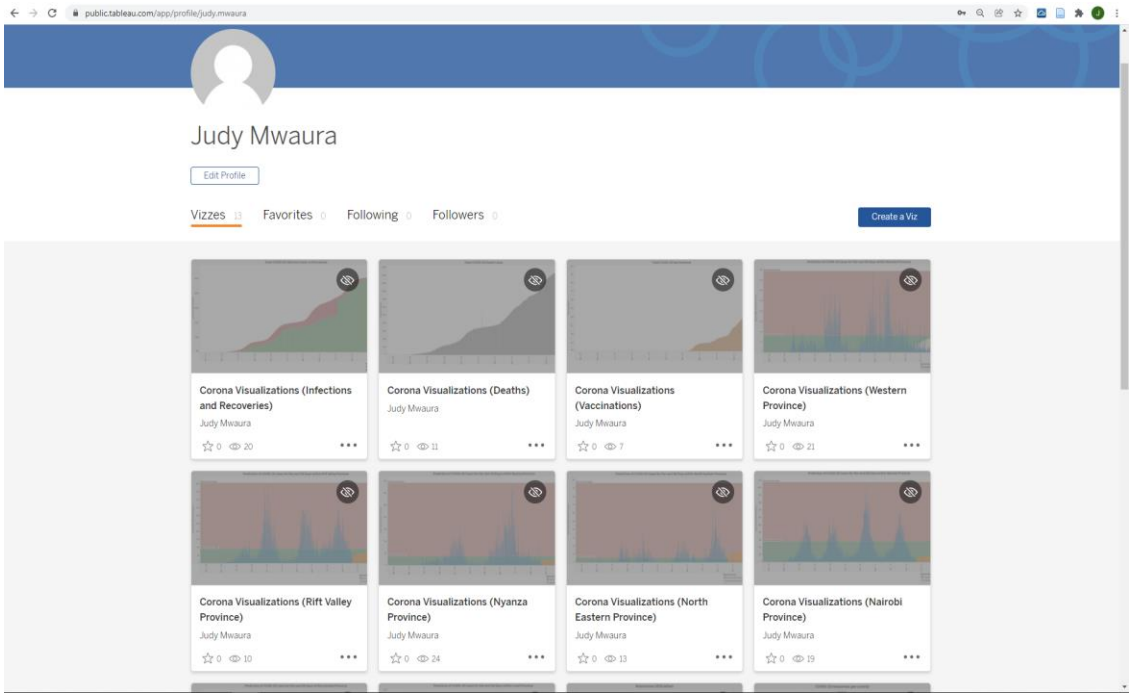


Figure 5.41: Visualizations transferred to Online Tableau Public

5.2.2.6 Creating a web-based Visualization dashboard

The researcher adopted HTML and CSS to design a website on which to transfer the visualizations using Brackets text and code editor. The website acts as a central location which the users can access the proposed system that meets the requirements and specifications set. This is an interactive site as it allows the users to modify the views by permitting filtering within them and cross-highlighting on specific data items.

Step 1: The researcher developed different .html files for each of the dashboards created. A code derived from the visualizations on online Tableau Public was embedded into the body of the html files for the respective dashboards. Below is the code:

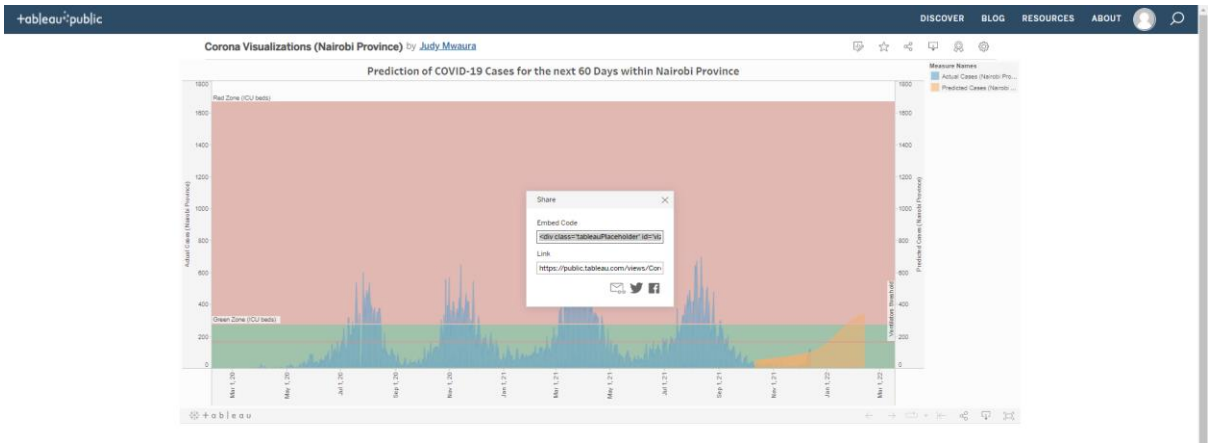


Figure 5.42: Code to embed onto HTML

Step 2: The link embedded to the .html file. It consists of a div class and JavaScript that help in the formatting of the visualization dashboard on the web. As shown on the right side of the figure below, the researcher created different .hml files for each of the visualization dashboards present. The index .html file was used to link all of them together.

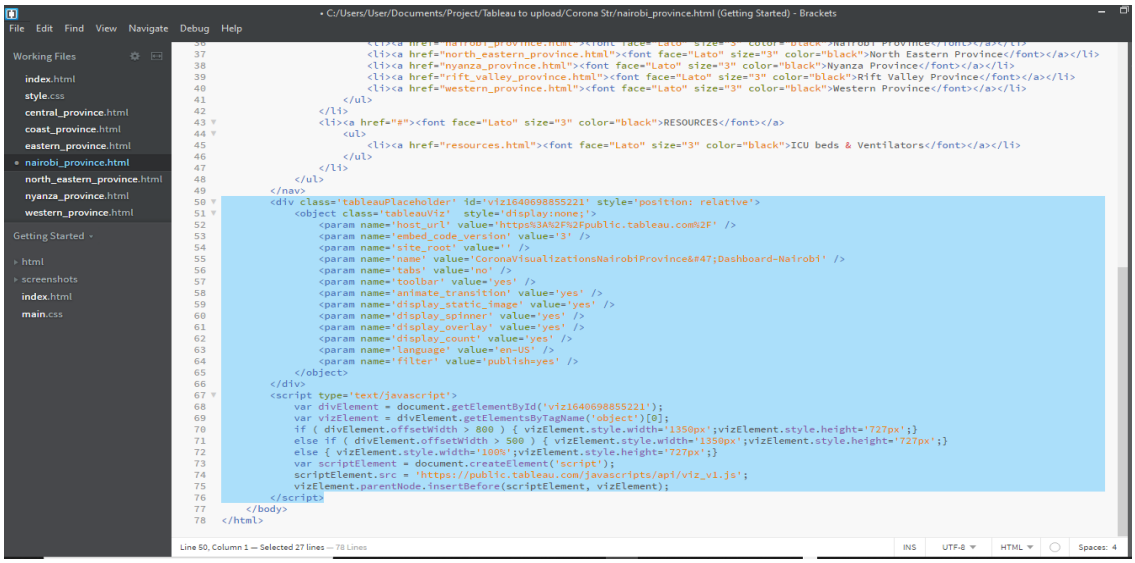


Figure 5.43: Embedded code to .html file

Step 3: The .css file was used to format the web page and position the visualization dashboards, drop down menus, titles, and background images accordingly.

Figure 5.44: Home Page

Corresponding visualization dashboard on Figure 5.36 on online Tableau Public, displayed on the web page:

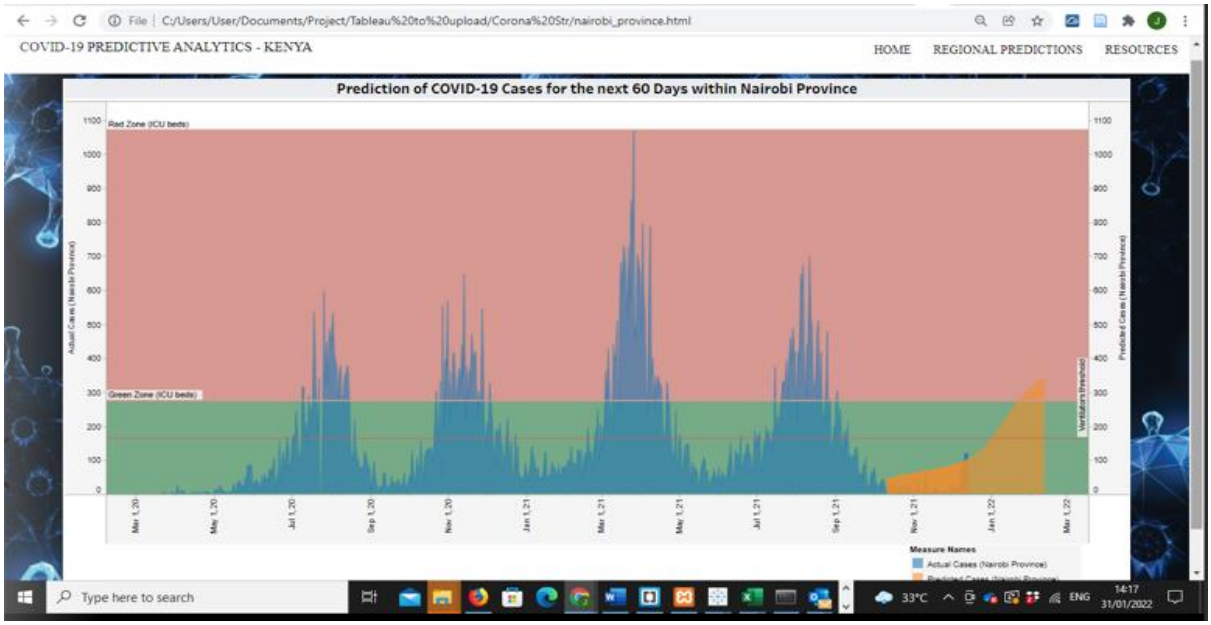


Figure 5.45: Dashboard view on the web page

Step 4: The researcher finally created a localhost for the web page by downloading and installing XAMPP which allows for building WordPress site offline on a local web server on the PC. XAMPP was installed on local drive E as opposed to C because there was another application running and using server port 80. This would conflict with XAMPP as the Apache server uses the same port. The Apache module was started, and the researcher could

view the web page using local host/corona/index.html URL to access the Home Page. Below figures represent XAMPP and access to the home page after setting up a localhost.

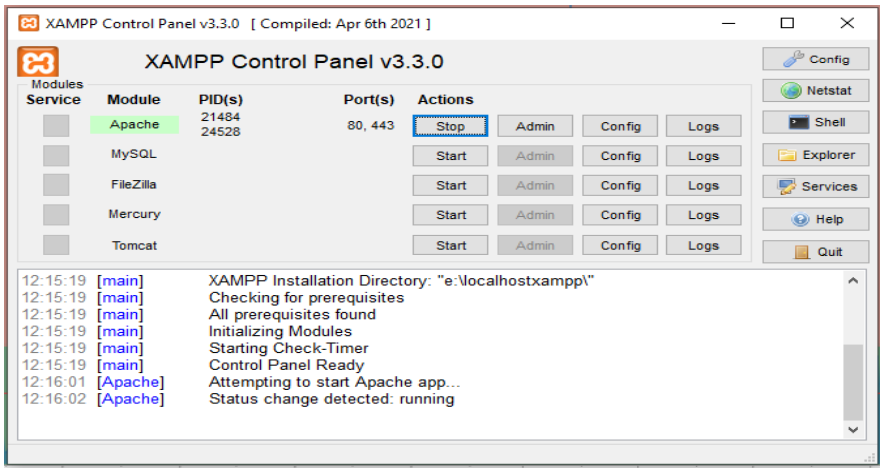


Figure 5.46: Started Apache module on XAMPP

Figure 5.47: Localhost on local web server on the computer

5.3 System Testing

System testing is a phase within the Waterfall Model. The researcher carried out unit testing on the predictive model independent of the visualization dashboard. Testing is done to ensure that all units are operating well, no bugs are present, and all requirements are met.

5.3.1 Performance of the Predictive Models

MAE and RMSE are two commonly used measures for evaluating prediction model performance. Despite the fact that both MAE and RMSE are commonly used, there is no evidence that one is preferable over the other. Since MAE is easier to calculate, it is less ambiguous than RMSE. As a result, several studies decided to employ MAE

rather than RMSE. RMSE penalizes variation by giving more weight to errors with greater absolute values than to errors with lower absolute values. (Choy et al., 2021).

Many systems use the root mean square error (RMSE) as a standard evaluation metric to calculate the differences between expected and actual values. MAE, on the other hand, stands for the Mean Absolute Value of Errors. According to all of the studies reviewed, MAE and RMSE are the most commonly used evaluation measures for prediction model performance (W. Wang & Lu, 2018). Therefore, in this study, the researcher used MAE and RMSE as evaluation tools to assess the performance of the time-series algorithms. The Mean Squared Error (MSE) is derived by calculating the square of the RMSE.

Hyperparameter tuning was carried out for each of the algorithms to attain the lowest error values as possible. This was done by adjusting different parameters such as the number of hidden layers as well as number of neurons within each layer, adding or omitting the dropout layer, adjusting the number of epochs, steps per epoch, and adjusting the batch size. Hyperparameter tuning was done while creating a model to predict COVID-19 cases within Nairobi province. This is because data on Nairobi province was adequate. The parameters that gave the best values for MSE, MAE and RMSE were retained to create models for the rest of the provinces, for each of the algorithms. Below are the best MSE, MAE and RMSE values recorded for each of the algorithms.

Machine Learning Algorithms	**MSE**	**MAE**	**RMSE**
Multiple Linear Regression **Parameters:** Dropout (0.5), Dense(units=1) optimizer = ‘adam’, epochs = 250 batch_size = 64	260.28	3.68	16.13
Linear Regression	149.92	2.94	12.24
Decision Tree	227.75	3.37	15.09
Random Forest	227.50	2.93	15.08
ARIMA **Parameters:** Order = (4,1,3)	nan	nan	nan
Bayesian Polynomial Regression **Parameters:** ‘tol’: tol, 'alpha_1': alpha_1, 'alpha_2' : alpha_2, 'lambda_1': lambda_1, 'lambda_2' : lambda_2 'normalize' : normalize cv = 10 n_jobs = -1	83326.82	249.86	288.66

n_iter = 1000			
FFNN **Parameters:** Dropout (0.5), Dense(units=2) Optimizer = 'adam', epochs = 200 batch_size = 64	1959.14	30.68	44.26
LSTM **Parameters:** Dense(units=1) Optimizer = 'adam', epochs = 70 Steps_per_epoch = 10 batch_size = 16	0.0784	0.26	0.28

Table 5.1: MSE, MAE and RMSE values for the algorithms

Source: Author

From the table, it is evident that, LSTM gave the lowest error values and hence performed best. This is the algorithm that the researcher adopted for the predictive model. This allowed the researcher to satisfactorily achieve the second objective of this study.

5.3.2 Functionality of the Visualization Dashboard

The researcher performed tests on the visualization dashboard by ensuring that the data graphed corresponds to the specific imported data from each excel. This was done by cross-checking the naming of each excels and ensuring that they are the same ones that were imported to the dashboard. This was done upon completion of the visualization dashboard on Tableau desktop. The researcher also checked the web-based dashboard to ensure that it met three main factors discussed earlier in the development of the dashboard.

Purpose of the dashboard – the final output is a visualization dashboard that performs predictive analytics and uses the constant and reference lines to inform on the status of the resources. This will be used to drive decision making on whether to procure additional health resources or if the ones available are enough.

Target audience – the target audience are individuals whose expertise lie towards the health sector. Therefore, the researcher used area graphs which is a basic visualization that can be easily interpreted and understood.

Expected visual features and interactivity – the user can modify the views by performing filtering within the views. Filters can be applied on the Measure names to either observe the Actual infection cases or the Predictions. Interactive interfaces are also present whereby the user can either highlight the cases in the red zone or those in the green zone. Moreover, there are also interactive components such as dropdowns whereby the user has the option to choose the regional prediction to observe, whether to view the infection, deaths, recoveries, or vaccination cases. The user can also hover the mouse or pointer across the graphs to get number of cases recorded and the day they were recorded.

Below a filter has been done to show the predicted cases, the cases under the green zone have been highlighted and the day and cases recorded where we have the pointer are showing:

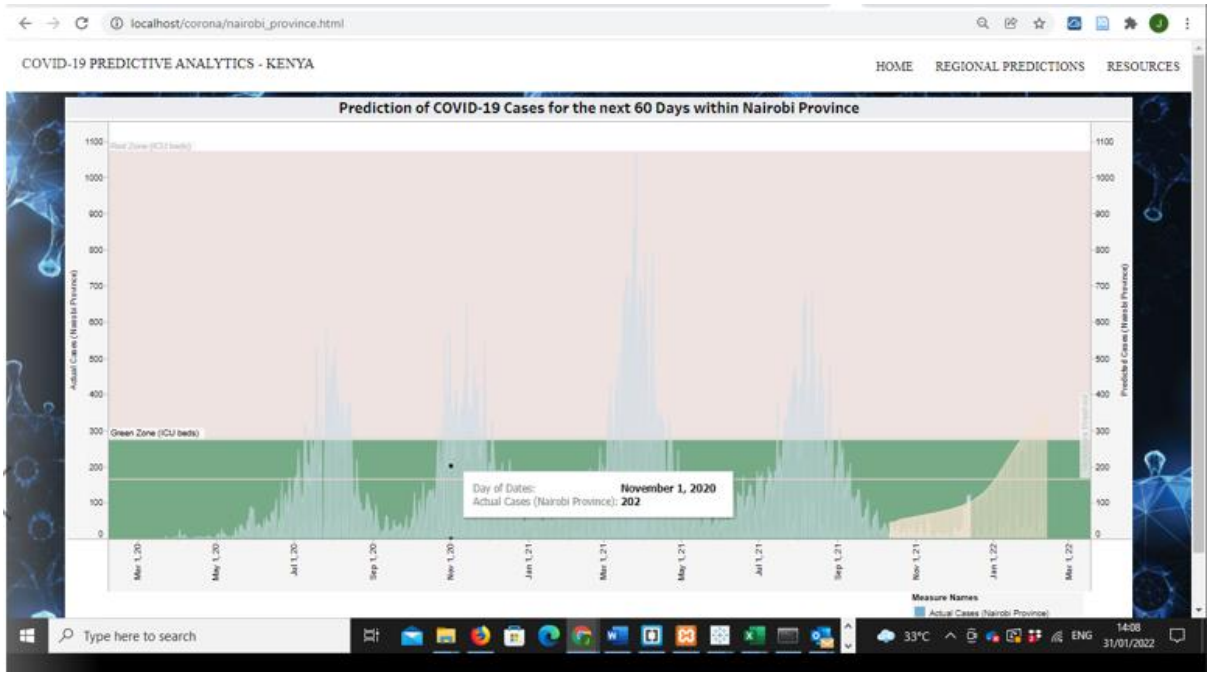


Figure 5.48: Highlighting and Filtering

Below is an example of a dropdown which is an interactive feature:

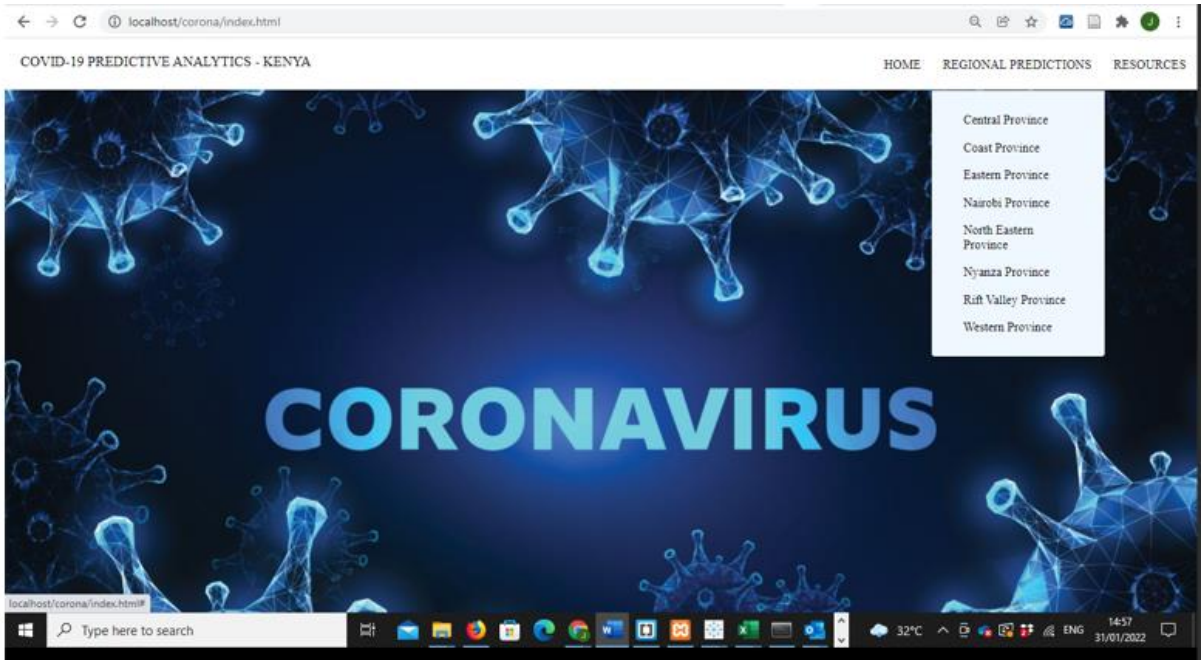


Figure 5.49: Dropdowns

5.4 System Evaluation

The researcher evaluated the system based on the research objectives of this study. Satisfactorily meeting each objective allows the researcher to meet the general objective of the study which is to use Artificial Neural Networks (ANN) in carrying out predictive analytics to aid in decision related issues in healthcare resource planning, with the aim of containing and mitigating the spread of the virus.

Reviewing and Evaluating the COVID-19 predictive systems currently in use.

This objective was met during the literature review phase by analysing the predictive systems developed in the face of COVID-19 and evaluating their performances, advantages, and disadvantages. This objective also gave the researcher the determination to adopt Neural networks which had barely been used in previous studies due to the limited COVID-19 data that was available at the time. The predictive systems developed mainly use statistical models and regression to forecast. While evaluating the predictive systems, the researcher noted that, most of these platforms were purely for prediction and visualizations were incorporated in the IDEs. On the other hand, visualization dashboards created for COVID-19 data do not incorporate the prediction bit. It is basically a transformation of the number of COVID-19 infection cases to a visual representation.

Development of the predictive machine learning model using neural networks to forecast the spread of the COVID-19 virus.

The researcher adopted Multiple Neural Regression using Neural Networks, LSTMs and FFNN during the development of the predictive machine learning model. Each of these are deep learning algorithms. For performance comparison, the researcher adopted other algorithms such as Bayesian Polynomial Regression, Traditional ARIMA, Linear Regression, Decision Tree, and Random Forests, which are not deep learning algorithms, for performance comparison. From the System Testing phase of this study, Table 5.1, it is evident that the model developed using LSTM networks outperformed all the rest based on the low MSE, MAE and RMSE values observed. Upon performing hyper-parameter tuning, this model was adopted to perform predictions for the next 60 days in each of the eight provinces. The predictions were then saved on a different excel file, ready for visualization.

Design and development of the visualization dashboards for the predicted data, COVID-19 cases and health resources set aside to contain the virus.

Having the predicted data after achieving the second objective and COVID-19 cases and health resources after completing the data extraction and pre-processing steps under the development of the predictive model, the researcher designed and developed the visualization dashboards for each of them using Tableau Desktop application. The researcher then created a web page for the visualizations. This allowed for the users to access them from a central location.

Evaluation of the model performance and functionality of the dashboards.

The researcher evaluated the performance of the model by comparing its performance against the rest of the models developed using other algorithms based on the MSE, MAE and RMSE performance metrics. Predictions were also done on the test data. Both the test data and predictions were visualized. The predictions and test data values should be close to if not the same. If the graph on test data shows an increment in the COVID-19 cases, the graph on predicted data should record a similar trend. That was noted on the graphs for the predictions for the provinces:

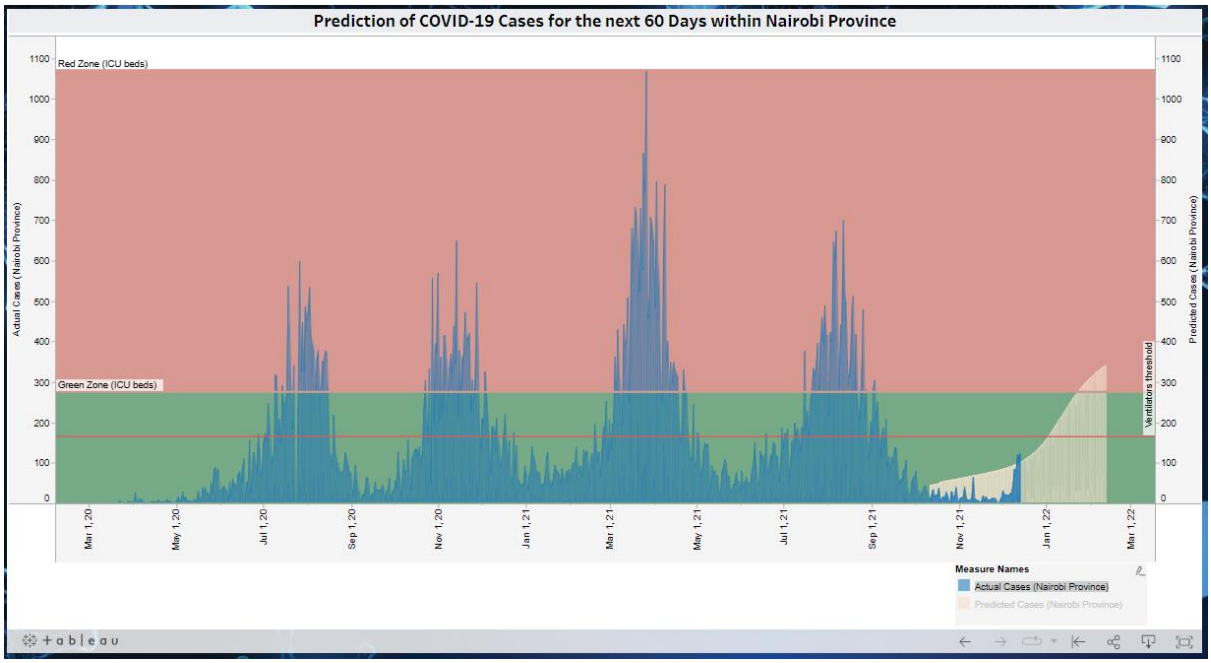


Figure 5.50: Actual Test data against Predictions

The researcher presented the system to the users of the system to evaluate. The users commended its simplicity as it only provides information that is useful to them, the background chosen, which majority advised is very appropriate for this study. They also appreciated the interactive capabilities of the system (Dropdowns, Highlights, Filtering) and the graphs used that are easy to interpret. Most of all, majority were very pleased that the researcher incorporated the requirements they had provided during the initial stages of this study, regarding the Red and Green Zones for the ICU beds, and the thresholds for the ventilators. This was a key requirement to enhance decision support by ensuring that any dangers can be noted first-hand by simply viewing the dashboards. The below table gives a summary of the feedback obtained from the users during evaluation. The feedbacks are also analysed and compared against the recommendations that had been given during the requirements analysis phase.

Key features and concerns regarding the system	**Evaluation of the system by users based on the Recommendations from the Requirements Gathering Phase.**
Number of days to predict into the future	The researcher carried out 60-days forecast for each of the provinces. This satisfied the users of the system since during the requirements gathering phase, majority had proposed a minimum of 50-days forecast to allow for enough time to plan to procure or reallocate resources. The users appreciated this.
Complexity of the dashboards	The users of the system had proposed simple graphs that are easy to interpret. The researcher used area graphs that are simple to understand. This was appreciated by the users.
How to establish the health resources required, from the dashboards	The researcher incorporated red and green zones on the dashboards as per the suggestion gives by the users during the requirements gathering phase. Any forecasted cases that enter the red zone is an indication that more resources need to be

	procured for the specific provinces while the forecasted cases that remain within the green zone give an indication that the health resources available (ICU beds and ventilators) are enough for the infected individuals to get appropriate health care. This was highly commended by the users as it enhances decision support by ensuring that any dangers can be noted first-hand by simply viewing the prediction dashboards.
Interactive capabilities of the dashboards	This attribute was highly appreciated by the users. The researcher incorporated Dropdowns to allow the user to choose the predictions to view out of the eight provinces and easily manoeuvre across the site. Highlights allowed the users to view the exact number of cases predicted or recorded on a particular day by simply hovering the cursor across the graphs and since the dashboards contain graphs of different variable, the researcher incorporated filtering capabilities as initially recommended by the users during the requirements gathering phase.
Knowledge on how to use the system and interpret the dashboards.	The researcher arranged for a session to present the system to the users and inform them of its capabilities and how to interpret the visualizations as proposed during the requirements gathering phase. This was well received by the users who even proposed to have regular sessions to address any queries they have till they fully familiarize themselves with it, and to also convey any other suggestions to improve and enhance the system.
How long it will take for a new forecast to be made	The researcher informed the users that the dashboards would be updated with new forecasts when a minimum of five days are remaining in the existing forecast as they had requested. This was well perceived.
Ability to download visualization dashboards from the site	The researcher added this feature, though it was not one of those proposed during requirements gathering. Nonetheless, the users appreciated it as it gave them the ability to save them and compare the predictions against future ones.

Table 5.2: System Evaluation

Source: Author

5.5 Summary of Results

1. The results obtained were expected and contributed to the general objective of the study.
2. COVID-19 is a little more unpredictable hence forecasting is a bit challenging. It has multiple parameters and variables, yet the variables have very low correlation between them.
3. The different algorithms adopted in the study were those implemented in other studies. Of these, LSTM Neural Networks performed best for univariate time series prediction.
4. Tableau visualization tool is a very powerful and easy to use tool.
5. Predictive analytics and visualization are both very powerful when it comes to data representation and decision making.

5.6 Interpretation and Discussion of Findings

This section discusses the results obtained. It compares the performance of the main algorithm adopted in this study (ANN) to those in previous studies, assesses the significance of those results and discusses new insights in line with the approaches used in solving the problem of the study.

COVID-19 is a little more unpredictable hence forecasting is a bit challenging. From other studies, prediction of COVID-19 cases has been done through statistical models such as SEIR and SEIHCRD, through regression and ANN algorithms. All these algorithms were time series – based. This is because COVID-19 has multiple parameters and variables, yet the variables have very low correlation between them. This is proven when the researcher adopted Multiple Linear Regression and used the correlation matrix to determine which variables have a high correlation. It showed that, there is no strong relationship between the number of cases recorded and the rest of the variables relating to the coronavirus dataset. Based on this, univariate time series forecasting proved to be the best approach.

LSTM neural network algorithm gives the best performance while carrying out univariate time series forecasting – as discussed in Chapter four of this study, the researcher was able to adopt different algorithms that had been previously used in other studies reviewed in Chapter 2 to predict the spread of COVID-19 cases. The performance of the algorithms was compared against LSTM neural networks. LSTM gave the best performance (lowest MSE, MAE and RMSE values). From a study by (N. Hamadneh et al., 2021), who adopted ANN using a multilayer perceptron neural network (MLPNN) and a prey-predator algorithm (PPA) resulting to a hybrid model (MLPNN-PPA), RMSE was used to test the performance of the predictive model. A value of 0.13 was obtained which is relatively low meaning that the model was reliable. It is slightly lower than the 0.28 obtained in this study. From another study that employed ANN, MLPNN, (Nyoni, 2020), the metrics used to test the model were MSE and MAE. The values obtained for MAE and MSE were 117.385530 and 30758.263463 respectively, which are considerably higher than those obtained in this study. As much this the study was directly aimed towards adoption of Neural Networks for prediction, the researcher was determined to establish the best algorithm by carrying out performance comparison.

Tableau visualization tool is very powerful yet easy to learn and use. The researcher had explored different visualization tools for the study during the literature review phase. However, Tableau proved to be the simplest given the limited amount of time availed for this study. The procedures followed were simple while the output was tremendous.

Merging predictive analytics and visualization enhances decision support. Upon reviewing the algorithms and statistical models used in forecasting COVID-19 as well as the visualization tools adopted in the creation of dashboards upon the onset of the virus, the researcher discovered that predictive analytics and visualization are both very powerful when it comes to data representation and decision making. However, none of the studies had adopted both. The researcher was able to address this research gap by carrying out predictive analytics (predicting the spread of COVID-19 cases for the next 60 days) and presenting the output on a visualization dashboard. The same dashboard gave a visual interpretation of the health resources available and their adequacy in taking care of COVID-19 patients. During the evaluation of the prototype by the users, they were able to determine with ease where and when additional health resources would be required for each of the provinces, meaning that, besides meeting the general objective of the study, merging visualization and predictive analytics enhances decision support.

5.7 Assumptions

The researcher made the below assumptions during the research and development of the prototype:

1. All COVID-19 infection cases require the use of ventilators and ICU beds for recovery.
2. The number of ICU beds and ventilators remain a constant on the dashboard, for the period the predictions have been made.
3. Health resources are planned for province-wise (their procurement and distribution)

CHAPTER SIX: CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

COVID-19 has spread over the world since its first official appearance in China, in December 2019. To date, it has claimed thousands of lives due to its high spread and transmission rate. Though the virus is still claiming lives in Kenya, the situation has however improved in the recent past due to several steps taken such as enforcing social distancing, frequent washing of hands and sanitization, and wearing of personal protective equipment such as surgical masks. Although vaccines have been developed to prevent individuals from contracting the virus, different strains of COVID-19 continue to evolve and outweigh the strength of the vaccines as those vaccinated still face a chance to get infected. COVID-19 still has no cure, and the existing treatments are only for the symptoms. Severe cases of COVID-19 require ICU beds and ventilator support. From the research, Kenya has previously faced a risk whereby, these resources were not enough to take care of the COVID-19 patients requiring them. Despite the shortage, the resources are also not evenly distributed across the provinces. Hence, forecasting the number of infection cases is crucial to avoid a similar occurrence by allowing for pro-active monitoring of infection cases against the health resources available.

The research was able to adopt different algorithms for prediction and compare their performances based on the MSE, MAE and RMSE values. LSTM neural networks performed best and was used to develop the predictive model. Tableau on the other hand, was the best in visualizing due to its simplicity. Tableau was used to visualize the number of infection cases recorded over time, as well as the predictions made. The health resources available were also visualized on the same platform. A red zone was used to represent a section where ICU beds are not adequate while a green zone represented a section where the beds are adequate. Therefore, any predictions falling on the red zone are meant to act as an indicator of shortage of the resources and have the health workers prepare in advance. Those that fall within the green zone have enough resources for care. A constant line was used for ventilators whereby, any cases above the constant line indicate more ventilators should be procured. From the review of existing literature, the researcher had discovered that, different platforms for carrying out predictive analytics of COVID-19 exist. Visualization dashboards for performing pure visualization of COVID-19 data also exist. However, there is no platform that combines the powerful capabilities of visualization with predictive analytics. Therefore, this research bridged that gap by providing a system/platform that performs predictions on COVID-19 infection cases (predictive analytics) and represents them in form of interactive graphs on visualization dashboards. Moreover, it visualizes the adequacy of health resources available within the same dashboard, to enable the health workers to make decisions and carry out health resource planning for the predicted infection cases, hence providing a platform for decision support.

Different forms of coronavirus epidemics are expected to continue evolving in the future. As a result, ongoing research is essential in the study of current and future coronaviruses as well as their spread. These should be compared against the capability of the country/provinces to handle these cases with the health resources available.

6.2 Challenges and Limitations

The following challenges were faced during the study:

1. Data complexity – the open-source data on COVID-19 infection cases was unstructured. It therefore required a lot of manual effort and time to extract into a structured form and pre-process it by cleaning and aggregating to use it to train the predictive model.
2. Time constraints– The amount of time available to perform the study was limited by the deadline. Most of the time was spent in collecting the COVID-19 data where the researcher had to read through each figure and transfer that on an excel file. Development of the predictive models using different algorithms for performance comparison was also time-consuming, especially when the researcher had to carry out hyperparameter tuning to find the parameters that gave the best results (lowest MAE, MSE and RMSE values).

6.4 Contributions to the Study

The study helped the researchers to determine the relationships between the different variables relating to COVID-19 and if they directly influence the spread of the virus. This further led to the identification of the best algorithm suitable for univariate time series forecasting. The research has offered a platform for decision support in the health sector using a visualization dashboard that is easy to understand and interpret. Individuals working in the heath sector will be able to proactively prepare to handle any infection cases in the future days. This has been made possible by merging predictive analytics with visualization. Therefore, a repeat of a situation whereby the number of health resources are not adequate to take care of COVID-19 patients requiring them, will be avoided.

6.5 Recommendations for Future Studies

The researchers recommend that this study be further developed to automate the collection of COVID-19 cases and health resources as opposed to doing it manually as this will save on the time and shorten the processes undertaken in the development of the prototype. Moreover, the research can be expounded based on the assumption made that all infection cases will require ICU beds and ventilators. This can be expounded to determine how many of the cases predicted will not require ICU beds and ventilator support, how many will require and determine if the resources available are adequate.

REFERENCES

Ahmed, K., Bukhari, M. A., Mlanda, T., Kimenyi, J. P., Wallace, P., Okot Lukoya, C., Hamblion, E. L., & Impouma, B. (2020). Novel Approach to Support Rapid Data Collection, Management, and Visualization During the COVID-19 Outbreak Response in the World Health Organization African Region: Development of a Data Summarization and Visualization Tool. *JMIR Public Health and Surveillance*, *6*(4), e20355. Retrieved May 31, 2022, from https://doi.org/10.2196/20355

Alt, J. E., King, G., & Signorino, C. S. (2001). Aggregation Among Binary, Count, and Duration Models: Estimating the Same Quantities from Different Levels of Data. *Political Analysis*, *9*(1), 21–44. Retrieved May 31, 2022, from https://doi.org/10.1093/oxfordjournals.pan.a004863

Aung, T., Niyeha, D., & Heidkamp, R. (2019). Leveraging data visualization to improve the use of data for global health decision-making. *Journal of Global Health*, *9*(2), 020319. Retrieved May 31, 2022, from https://doi.org/10.7189/jogh.09.020319

Ayer, T., Chhatwal, J., Alagoz, O., Kahn, C. E., Woods, R. W., & Burnside, E. S. (2010). Comparison of Logistic Regression and Artificial Neural Network Models in Breast Cancer Risk Estimation. *RadioGraphics*, *30*(1), 13–22. Retrieved May 31, 2022, from https://doi.org/10.1148/rg.301095057

Barasa, E. W., Ouma, P. O., & Okiro, E. A. (2020). Assessing the hospital surge capacity of the Kenyan health system in the face of the COVID-19 pandemic. *PLOS ONE*, *15*(7), e0236308. Retrieved May 31, 2022, from https://doi.org/10.1371/journal.pone.0236308

Bedford, J., Enria, D., Giesecke, J., Heymann, D. L., Ihekweazu, C., Kobinger, G., Lane, H. C., Memish, Z., Oh, M., Sall, A. A., Schuchat, A., Ungchusak, K., & Wieler, L. H. (2020). COVID-19: Towards controlling of a pandemic. *The Lancet*, *395*(10229), 1015–1018. Retrieved May 31, 2022, from https://doi.org/10.1016/S0140-6736(20)30673-5

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D^3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2301–2309. Retrieved May 31, 2022, from https://doi.org/10.1109/TVCG.2011.185

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3). Retrieved May 31, 2022, from https://doi.org/10.1214/ss/1009213726

Cavallo, J. J., Donoho, D. A., & Forman, H. P. (2020). Hospital Capacity and Operations in the Coronavirus Disease 2019 (COVID-19) Pandemic—Planning for the Nth Patient. *JAMA Health Forum*, *1*(3), e200345. Retrieved May 31, 2022, from https://doi.org/10.1001/jamahealthforum.2020.0345

Chatfield, C. (2003). *The Analysis of Time Series* (0 ed.). Chapman and Hall/CRC. Retrieved May 31, 2022, from https://doi.org/10.4324/9780203491683

Choy, Y.-T., Hoo, M., & Khor, K. C. (2021). *Price Prediction Using Time-Series Algorithms for Stocks Listed on Bursa Malaysia* (p. 5). Retrieved May 31, 2022, from https://doi.org/10.1109/AiDAS53897.2021.9574445

Cook, R. D., & Weisberg, S. (1982). Criticism and Influence Analysis in Regression. *Sociological Methodology*, *13*, 313. Retrieved May 31, 2022, from https://doi.org/10.2307/270724

Dhiman, A. (2020). *Why Is Python Programming Language Gaining So Much Popularity*.

Duong, D. (2021). Alpha, Beta, Delta, Gamma: What's important to know about SARS-CoV-2 variants of concern? *Canadian Medical Association Journal*, *193*(27), E1059–E1060. Retrieved May 31, 2022, from https://doi.org/10.1503/cmaj.1095949

Farkas, C., Iclanzan, D., Oltean-Péter, B., & Vekov, G. (2020). Comparing epidemiological models with the help of visualization dashboards. *Acta Universitatis Sapientiae, Informatica*, *12*, 260–282. Retrieved May 31, 2022, from https://doi.org/10.2478/ausi-2020-0016

Feng, S., Feng, Z., Ling, C., Chang, C., & Feng, Z. (2021). Prediction of the COVID-19 epidemic trends based on SEIR and AI models. *PLOS ONE*, *16*(1), e0245101. Retrieved May 31, 2022, from https://doi.org/10.1371/journal.pone.0245101

Friendly, M. (2008). A Brief History of Data Visualization. In C. Chen, W. Härdle, & A. Unwin, *Handbook of Data Visualization* (pp. 15–56). Springer Berlin Heidelberg. Retrieved May 31, 2022, from https://doi.org/10.1007/978-3-540-33037-0_2

Gao, P., Zhang, H., Wu, Z., & Wang, J. (2020). Visualising the expansion and spread of coronavirus disease 2019 by cartograms. *Environment and Planning A: Economy and Space*, *52*(4), 698–701. Retrieved May 31, 2022, from https://doi.org/10.1177/0308518X20910162

Ghosh, S. (2020). Predictive model with analysis of the initial spread of COVID-19 in India. *International Journal of Medical Informatics*, *143*, 104262. Retrieved May 31, 2022, from https://doi.org/10.1016/j.ijmedinf.2020.104262

Godio, A., Pace, F., & Vergnano, A. (2020). SEIR Modeling of the Italian Epidemic of SARS-CoV-2 Using Computational Swarm Intelligence. *International Journal of Environmental Research and Public Health*, *17*(10), 3535. Retrieved May 31, 2022, from https://doi.org/10.3390/ijerph17103535

Gokhale, L., & Mahajan, K. (2020). *Comparative Study of Data Visualization Tools*. *IX*, 6017–6021.

Hair, J. F. (2007). Knowledge creation in marketing: The role of predictive analytics. *European Business Review*, *19*(4), 303–315. Retrieved May 31, 2022, from https://doi.org/10.1108/09555340710760134

Henrys, K. (2021). ROLE OF PREDICTIVE ANALYTICS IN BUSINESS. *SSRN Electronic Journal*. Retrieved May 31, 2022, from https://doi.org/10.2139/ssrn.3829621

Hui, D. S., I Azhar, E., Madani, T. A., Ntoumi, F., Kock, R., Dar, O., Ippolito, G., Mchugh, T. D., Memish, Z. A., Drosten, C., Zumla, A., & Petersen, E. (2020). The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—The latest 2019 novel coronavirus outbreak in Wuhan, China. *International Journal of Infectious Diseases*, *91*, 264–266. Retrieved May 31, 2022, from https://doi.org/10.1016/j.ijid.2020.01.009

Institute of Information Technology of Azerbaijan National Academy of Sciences, Hajirahimova, M., Ismayilova, M., & Institute of Information Technology of Azerbaijan National Academy of Sciences. (2018). BIG DATA VISUALIZATION: EXISTING APPROACHES AND PROBLEMS. *Problems of Information Technology*, *09*(1), 65–74. Retrieved May 31, 2022, from https://doi.org/10.25045/jpit.v09.i1.07

Jayesh, S., & Sreedharan, S. (2020). *Covid-19 Cases in India: A Visual Exploratory Data Analysis Model* [Preprint]. Health Informatics. Retrieved May 31, 2022, from https://doi.org/10.1101/2020.09.11.20193029

Jones, S. S., Thomas, A., Evans, R. S., Welch, S. J., Haug, P. J., & Snow, G. L. (2008). Forecasting Daily Patient Volumes in the Emergency Department. *Academic Emergency Medicine*, *15*(2), 159–170. Retrieved May 31, 2022, from https://doi.org/10.1111/j.1553-2712.2007.00032.x

Kajitani, Y., Hipel, K. W., & Mcleod, A. I. (2005). Forecasting nonlinear time series with feed-forward neural networks: A case study of Canadian lynx data. *Journal of Forecasting*, *24*(2), 105–117. Retrieved May 31, 2022, from https://doi.org/10.1002/for.940

Kemal, M. (2019). *Data Visualization Tools In Action Choosing a Visualization Software*. Retrieved May 31, 2022, from https://doi.org/10.13140/RG.2.2.11690.26560

Kumar, R. (2017). Machine Learning—Basics. In R. Kumar, *Machine Learning and Cognition in Enterprises* (pp. 51–64). Apress. Retrieved May 31, 2022, from https://doi.org/10.1007/978-1-4842-3069-5_4

Lin, L., McCloud, R. F., Bigman, C. A., & Viswanath, K. (2016). Tuning in and catching on? Examining the relationship between pandemic communication and awareness and knowledge of MERS in the USA. *Journal of Public Health*, fdw028. Retrieved May 31, 2022, from https://doi.org/10.1093/pubmed/fdw028

Macharia, P. M., Joseph, N. K., & Okiro, E. A. (2020). A vulnerability index for COVID-19: Spatial analysis at the subnational level in Kenya. *BMJ Global Health*, *5*(8), e003014. Retrieved May 31, 2022, from https://doi.org/10.1136/bmjgh-2020-003014

Malley, B., Ramazzotti, D., & Wu, J. (2016). Data Pre-processing. In *Secondary Analysis of Electronic Health Records* (pp. 115–141). Retrieved May 31, 2022, from https://doi.org/10.1007/978-3-319-43742-2_12

Martinez, R., Ordunez, P., Soliz, P. N., & Ballesteros, M. F. (2016). Data visualisation in surveillance for injury prevention and control: Conceptual bases and case studies. *Injury Prevention: Journal of the International Society for Child and Adolescent Injury Prevention*, *22 Suppl 1*, i27-33. Retrieved May 31, 2022, from https://doi.org/10.1136/injuryprev-2015-041812

Mashkoor, & Ahamad, M. V. (2017). *Visualization, Security and Privacy Challenges of Big Data*. https://www.semanticscholar.org/paper/Visualization%2C-Security-and-Privacy-Challenges-of-Mashkoor-Ahamad/853fcf118dad632e531e4dc504f0dbbe1e9ce0d9

Mbogo, R. W., & Orwa, T. O. (2021). SARS-COV-2 outbreak and control in Kenya—Mathematical model analysis. *Infectious Disease Modelling*, *6*, 370–380. Retrieved May 31, 2022, from https://doi.org/10.1016/j.idm.2021.01.009

McKendry, I. G. (2002). Evaluation of Artificial Neural Networks for Fine Particulate Pollution (PM_{10} and $PM_{2.5}$) Forecasting. *Journal of the Air & Waste Management Association*, *52*(9), 1096–1101. Retrieved May 31, 2022, from https://doi.org/10.1080/10473289.2002.10470836

Mondal, M. R. H., Bharati, S., Podder, P., & Podder, P. (2020). Data analytics for novel coronavirus disease. *Informatics in Medicine Unlocked*, *20*, 100374. Retrieved May 31, 2022, from https://doi.org/10.1016/j.imu.2020.100374

Muhula, S., Opanga, Y., Oramisi, V., Ngugi, C., Ngunu, C., Carter, J., Marita, E., Osur, J., & Memiah, P. (2021). Impact of the First Wave of the COVID-19 Pandemic on HIV/AIDS Programming in Kenya: Evidence from Kibera Informal Settlement and COVID-19 Hotspot Counties. *International Journal of Environmental Research and Public Health*, *18*(11), 6009. Retrieved May 31, 2022, from https://doi.org/10.3390/ijerph18116009

N. Hamadneh, N., A. Khan, W., Ashraf, W., H. Atawneh, S., Khan, I., & N. Hamadneh, B. (2021). Artificial Neural Networks for Prediction of Covid-19 in Saudi Arabia. *Computers, Materials & Continua*, *66*(3), 2787–2796. Retrieved May 31, 2022, from https://doi.org/10.32604/cmc.2021.013228

Nyoni, T. (2020). *PREDICTING COVID-19 CASES IN MOROCCO USING ARTIFICIAL NEURAL NETWORKS*.

Odhiambo, J. O., Okungu, J. O., & Mutuura, C. G. (2020). Stochastic Modeling and Prediction of the COVID-19 Spread in Kenya. *Engineering Mathematics*, *4*(2), 31. Retrieved May 31, 2022, from https://doi.org/10.11648/j.engmath.20200402.12

Petrovich, C. (2020). *Data visualization tools for web applications—A survey*. Retrieved May 31, 2022, from https://doi.org/10.13140/RG.2.2.27342.89920

Poletto, C., Scarpino, S. V., & Volz, E. M. (2020). Applications of predictive modelling early in the COVID-19 epidemic. *The Lancet Digital Health*, *2*(10), e498–e499. Retrieved May 31, 2022, from https://doi.org/10.1016/S2589-7500(20)30196-5

Pritee Chunarkar Patil, A. B. (2018). Big data analytics. *Open Access Journal of Science*, *Volume 2*(Issue 5). Retrieved May 31, 2022, from https://doi.org/10.15406/oajs.2018.02.00095

Qamar, U., & Raza, M. S. (2020). *Data Preprocessing* (pp. 63–86). Retrieved May 31, 2022, from https://doi.org/10.1007/978-981-15-6133-7_4

Rabi, F. A., Al Zoubi, M. S., Kasasbeh, G. A., Salameh, D. M., & Al-Nasser, A. D. (2020). SARS-CoV-2 and Coronavirus Disease 2019: What We Know So Far. *Pathogens*, *9*(3), 231. Retrieved May 31, 2022, from https://doi.org/10.3390/pathogens9030231

Rajni, J., & Malaya, D. B. (2015). Predictive Analytics in a Higher Education Context. *IT Professional*, *17*(4), 24–33. Retrieved May 31, 2022, from https://doi.org/10.1109/MITP.2015.68

Rodríguez, D. C., Hoe, C., Dale, E. M., Rahman, M. H., Akhter, S., Hafeez, A., Irava, W., Rajbangshi, P., Roman, T., Ţîrdea, M., Yamout, R., & Peters, D. H. (2017). Assessing the capacity of ministries of health to use research in decision-making: Conceptual framework and tool. *Health Research Policy and Systems*, *15*(1), 65. Retrieved May 31, 2022, from https://doi.org/10.1186/s12961-017-0227-3

Sanyaolu, A., Okorie, C., Hosein, Z., Patidar, R., Desai, P., Prakash, S., Jaferi, U., Mangat, J., & Marinkovic, A. (2021). Global Pandemicity of COVID-19: Situation Report as of June 9, 2020. *Infectious Diseases: Research and Treatment*, *14*, 117863372199126. Retrieved May 31, 2022, from https://doi.org/10.1177/1178633721991260

Sarikaya, A., Correll, M., Bartram, L., Tory, M. K., & Fisher, D. (2019). What Do We Talk About When We Talk About Dashboards? *IEEE Transactions on Visualization and Computer Graphics*. Retrieved May 31, 2022, from https://doi.org/10.1109/TVCG.2018.2864903

Singh, A., & Kumar Bajpai, M. (2020). SEIHCRD Model for COVID-19 Spread Scenarios, Disease Predictions and Estimates the Basic Reproduction Number, Case Fatality Rate, Hospital, and ICU Beds Requirement. *Computer Modeling in Engineering & Sciences*, *125*(3), 991–1031. Retrieved May 31, 2022, from https://doi.org/10.32604/cmes.2020.012503

Soyiri, I. N., Soyiri, I. N., & Reidpath. (2012). Evolving forecasting classifications and applications in health forecasting. *International Journal of General Medicine*, 381. Retrieved May 31, 2022, from https://doi.org/10.2147/IJGM.S31079

Tang, Z., & Fishwick, P. (1993). Feedforward Neural Nets as Models for Time Series Forecasting. *INFORMS Journal on Computing*, *5*, 374–385. Retrieved May 31, 2022, from https://doi.org/10.1287/ijoc.5.4.374

Toppenberg-Pejcic, D., Noyes, J., Allen, T., Alexander, N., Vanderford, M., & Gamhewage, G. (2019). Emergency Risk Communication: Lessons Learned from a Rapid Review of Recent Gray Literature on Ebola, Zika, and Yellow Fever. *Health Communication*, *34*(4), 437–455. Retrieved May 31, 2022, from https://doi.org/10.1080/10410236.2017.1405488

Wang, L., Wang, Z., Qu, H., & Liu, S. (2018). Optimal Forecast Combination Based on Neural Networks for Time Series Forecasting. *Applied Soft Computing*, *66*, 1–17. Retrieved May 31, 2022, from https://doi.org/10.1016/j.asoc.2018.02.004

Wang, S.-C. (2003). Artificial Neural Network. In S.-C. Wang, *Interdisciplinary Computing in Java Programming* (pp. 81–100). Springer US. Retrieved May 31, 2022, from https://doi.org/10.1007/978-1-4615-0377-4_5

Wang, W., & Lu, Y. (2018). Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model. *IOP Conference Series: Materials Science and Engineering*, *324*, 012049. Retrieved May 31, 2022, from https://doi.org/10.1088/1757-899X/324/1/012049

Yadav, R. S. (2020). Data analysis of COVID-2019 epidemic using machine learning methods: A case study of India. *International Journal of Information Technology*, *12*(4), 1321–1330. Retrieved May 31, 2022, from https://doi.org/10.1007/s41870-020-00484-y

Zakary, O., Larrache, A., Rachik, M., & Elmouki, I. (2016). Effect of awareness programs and travel-blocking operations in the control of HIV/AIDS outbreaks: A multi-domains SIR model. *Advances in Difference Equations*, *2016*(1), 169. Retrieved May 31, 2022, from https://doi.org/10.1186/s13662-016-0900-9