KENYATTA UNIVERSITY

SCHOOL OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF COMPUTING AND INFORMATION TECHNOLOGY

SCO400 PROJECT PROPOSAL

PROJECT TITLE: **COVID-19 CASELOAD AND MORTALITY ANALYSIS AND PREDICTION WITH MACHINE LEARNING**

submitted by

NAME: **IAN MOSES NJARI**

REG. NO: **J17/0803/2017**

EMAIL: **iannjari@gmail.com**

on

March 16th,2021

SUPERVISOR: **MR. JOSEPH MURIUKI MWANGI, M.Sc.**

This project proposal is submitted in partial fulfillment of requirements for a Bachelor's of Science (Computer Science) at Kenyatta University.

Table of Contents

CHAPTER ONE – INTRODUCTION:

1.1 Background of the Study

Coronavirus disease (COVID-19) is a respiratory infection caused by a virus known as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). On 30th January 2020, the director-general of the World Health Organization (WHO) declared the COVID-19 outbreak a public-health emergency of international concern. Six weeks later, the outbreak was categorized as a pandemic. COVID-19 has already caused 110,770,701 cases globally with 2,453,381 deaths (as of 20th February 2021).

The COVID-19 emergency is occurring in a much more digitized and connected world. Since the start of the pandemic in early 2019, data on the caseload, mortality and recoveries has been collected globally and unified together by various governments, institutions and organizations including the WHO.

This data is only meaningful only if it is processed correctly and transformed into some useful information. This information is very helpful in planning strategies and various decision-making tasks. In this project, the analysis is performed on the global COVID-19 available data to get some meaningful insights about the disease.

1.2 Problem Statement

Challenges with tracking these cases have arisen because of minimalized access to open-source solutions that;

a) Avail analysis tools for tracking the cases globally,

b) Statistically project the caseload per country/region,

c) Track the state of vaccinations globally.

Where these tools exist, they are only usable by end-users on the web as their source code is not freely available to the public, which would make it easier to deploy modified versions of the same to achieve more specific organizational goals especially for smaller organizations that may not have the financial power and technical pool to come up with similar solutions.

1.3 Objectives

The objectives of the project are to;

- Analyze the state of tracking applications and availability of open-source code for their implementation,
- Design an open-source solution which is easy to deploy and use,
- Create an interactive dashboard app tracking the number of cases, deaths, and recoveries by country and region,
- Build machine learning model, implemented in Python, that predicts the cases for the next 21 days in a country /region selected by the user,
- Create a daily report generation engine that shows the state of vaccinations worldwide.

1.4 Scope and Limitation of the Project

This project will include problem research and analysis, solution design and development project artifacts. It will also the documentation of the final project artifacts as well as defence at a panel.

Project deliverables will include;

- Problem analysis document
- Solution design
- Dashboard app
- Prediction model
- Report generator
- Documentation

This project will not include auxiliary services such as automatic report generation, email sending and data analysis.

1.5 Justification

The project is justifiable because, it involves application of fields learnt throughout the four-year program including but not limited to;

- Probability and statistics
- Intellectual property rights and policies
- Object Oriented programming
- System Analysis and Design
- Research methods and technical writing
- AI
- Software project management

The project also includes extensive learning and research in the area of machine learning which is time consuming.

CHAPTER TWO (Literature review)

2.1 Abstract/Introduction

Since the start of the COVID-19 pandemic in early 2020, data on the caseload has been collected globally and unified together by various governments, institutions and organizations. Challenges with tracking have arisen because of lack of open-source resources with effective and easy to use analysis tools for tracking the cases globally, tools for statistically projecting the caseload per country/region and comparison data for vaccination rates versus caseloads.

According to World Health Organization. (n.d.), there are over 110,224,709 confirmed cases as of 20th February 2021.

This paper aims to review the state of data collection, analysis, presentation and utilization by various entities worldwide.

2.2 Existing research in this field.

This data is only meaningful only if it is processed correctly and transformed into some useful information. This information is very helpful in planning strategies and various decision-making tasks (Reeve, 2013, pp. 113). In this project, the analysis is performed on the global COVID-19 available data to get some meaningful insights about the disease.

To better understand and alleviate the COVID-19 pandemic, many papers and preprints have been published online in the last few months.

As the leaders in the war against the novel coronavirus, the World Health Organization (WHO) and Center for Disease Control and Prevention (CDC) have in liaison with governments, health institutions and academia to come up with reporting protocols for various data concerning the outbreak which has helped to significantly reduce risks from the spread of COVID-19 outbreak. For example, as a search engine giant, Google launched a COVID-19 portal (www.google.com/covid19), where we can find useful information, such as coronavirus map, latest statistics, and most common questions on COVID-19.

Others include, the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University- which this project will utilize as a data source.

Big data associated with intelligent AI-based tools can build complex simulation models using coronavirus data streams for outbreak estimation (Pham et al., 2020, p. 130835). This is aiding health agencies in monitoring the coronavirus spread.

Models are being helpful as tools to make informed guesses about the disease and its future spread. Curve-fitting/extrapolation models infer trends about an epidemic in a given location by looking at the current status and then applying a mathematical approximation of the likely future epidemic path, which is drawn from experiences in other locations and/or assumptions about the population, transmission, and public health policies in place. These, as (MICHAUD et al., 2020) opinioned, are used together with SEIR/SIR models and Agent-based models for epidemiological modelling.

The administration of vaccines worldwide is another interesting field. The biggest vaccination campaign in history is underway. More than 199 million doses have been administered across 87 countries, according to (Randall et al., n.d.). The latest rate was roughly 6.50 million doses a day. In the U.S., more Americans have now received at least one dose than have tested positive for the virus since the pandemic began. So far, 60.5 million doses have been given, according to a state-by-state tally. In the last week, an average of 1.49 million doses per day were administered. How does this statistic compare to others?

2.3 Research gap to be addressed

Providing accurate information related to COVID-19 is an essential public service. For example, CovidGraph. It is an open-source graph database that brings together information on COVID-19 from different sources.

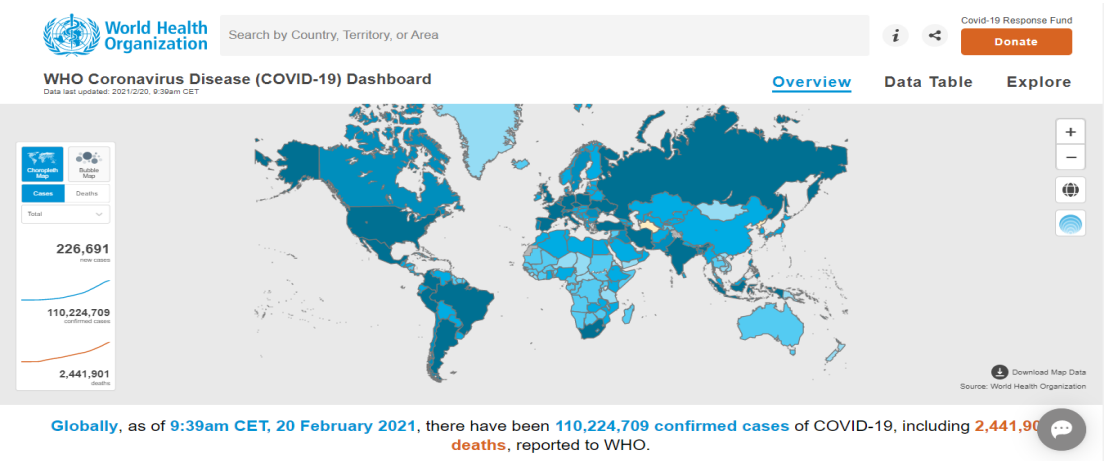
The power of graph data [distributed via an open-source management system] is that it can pull together disparate datasets from medical practitioners, public health officials and other scientific publications into one central view. People can then make connections between all facts. This is useful when looking for future long-term solutions." CovidGraph helped institutions like the Canadian government integrate data from multiple departments and facilities. The need for data is paramount. This isn't a matter of using data to sell ads, it's a matter of using data to save lives.

There however remains a desert of available easy to deploy solutions that do not use a distributed system (more prone to failure). This project uses a single repository as a source of data.

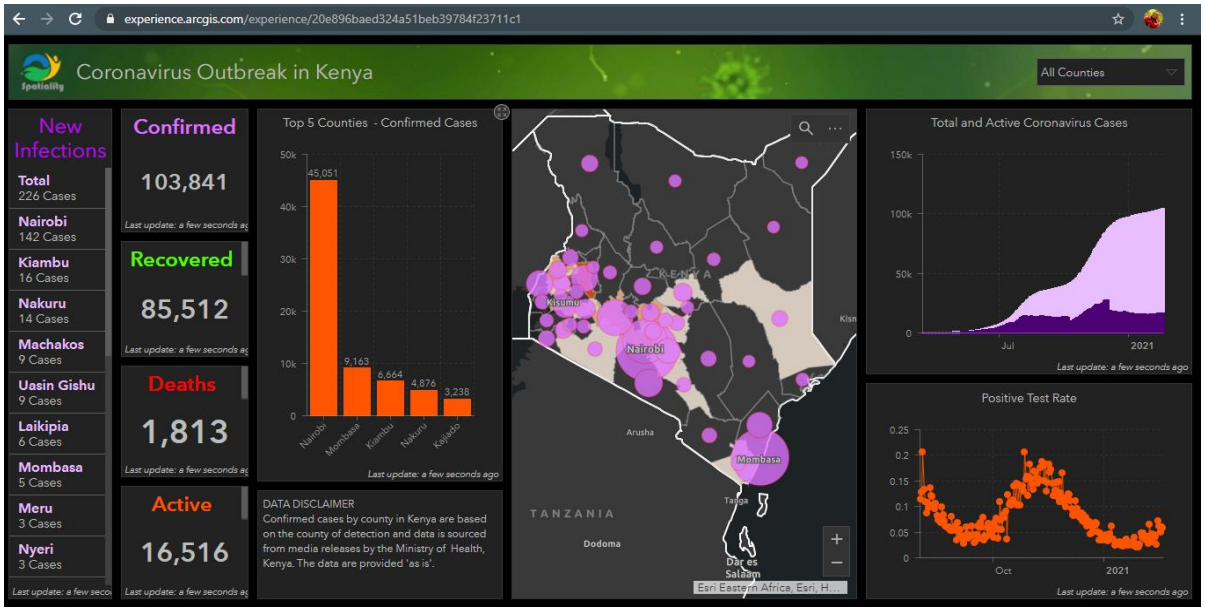
2.4 Case studies of Similar systems

Similar systems to what is being considered include;

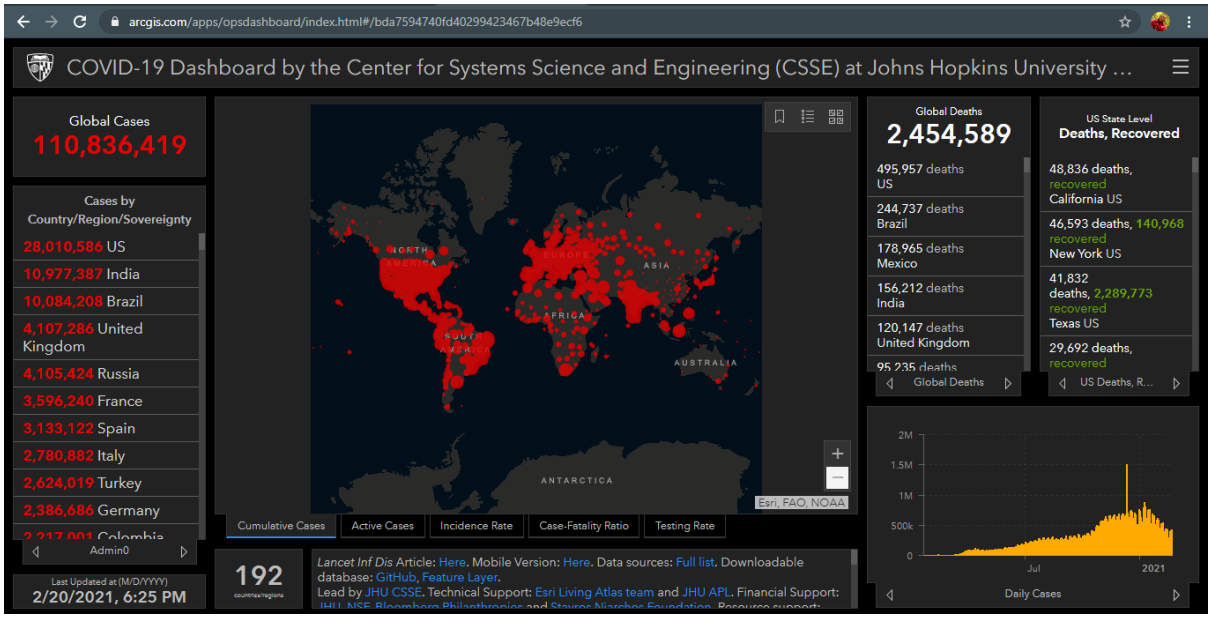
WHO Coronavirus Disease (COVID-19) Dashboard



COVID-19 Dashboard Kenya – ArcGIS



COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)



The proposed project artefacts take inspiration from the case studies and adds more functionality.

CHAPTER THREE (Methodology)

3.1 System Development Life Cycle

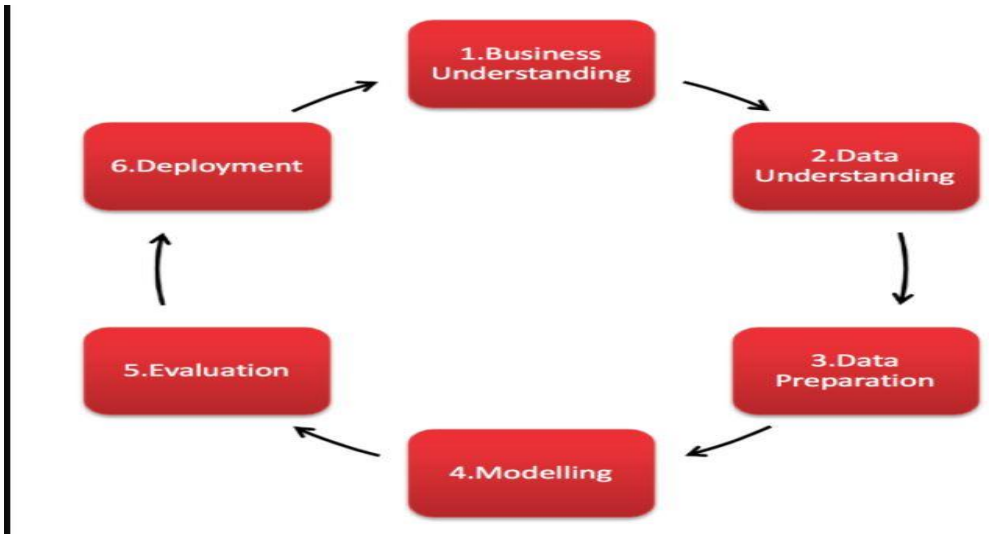
The **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (*CRISP-DM*) is a process methodology with six phases that naturally describes the **data science life cycle**. It has a set of steps to help plan, organize, and implement the data science (or machine learning) project. The CRISP-DM methodology provides a structured approach to planning a data science project (Smart Vision Europe, 2020). It is a robust and well-proven methodology.

This methodology was chosen because of its well-defined processes that fit well into the data science pipeline of the project.

The steps will be as follows;

- Business Understanding
- Data Understanding
- Data Preparation (will include the Exploratory Data Analysis Phase)
- Modelling
- Evaluation (Testing)
- Deployment

The sequence of the phases is not strict and moving back and forth between different phases as it is always required.



3.2 Project model/framework

3.2.1 Data collection and analysis techniques

The data will be sourced from a daily – updated COVID-19 Data Repository by the Centre for Systems Science and Engineering (CSSE) at Johns Hopkins University via GitHub under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

The data will be pulled from the GitHub repository through an API call.

3.2.2 Implementation Tools

Programming Languages;

- Python (Libraries:
  - Pandas,
  - Scikit-learn,
  - Prophet,
  - Plotly Express,
  - Dash,
  - FPDF)
- CSS- For styling the Plotly Dash HTML components.

GIT and GitHub- For version control

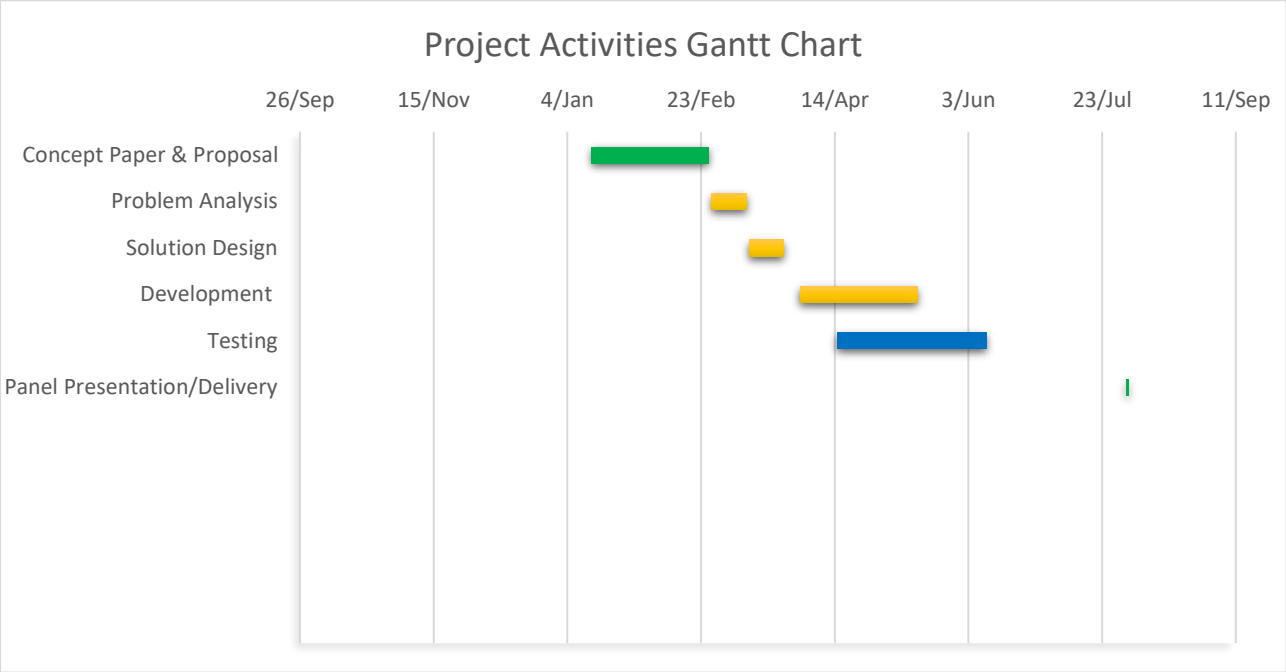
Development Environments;

- Jupyter Notebooks- For doing Exploratory Data Analysis and de-bugging.
- Visual Studio Code – For writing CSS and Python scripts.

- Microsoft Azure – For deployment of the three components of the final product to the cloud.

3.3 Schedule of activities

The schedule of activities will be as follows;



3.4 Project Budget

The project shall have the following cost projections in KShs;

ITEM	UNIT COST	QTY	TOTAL COST
Wi-Fi Charges	3000	8	24000
Azure Cloud	2000	1	2000
Transport	300	20	6000
Printing Costs	200	3	600

On-demand Courses	3000	1	3000

Total Cost = **35,600.00**

References

1. World Health Organization. (n.d.). *WHO Coronavirus Disease (COVID-19) Dashboard*. Retrieved February 20, 2021, from https://covid19.who.int/
2. Pham, Q.-V., Nguyen, D. C., Huynh-The, T., Hwang, W.-J., & Pathirana, P. N. (2020). Artificial Intelligence (AI) and Big Data for Coronavirus (COVID-19) Pandemic: A Survey on the State-of-the-Arts. *IEEE Access*, *8*, 130820–130839. https://doi.org/10.1109/access.2020.3009328
3. MICHAUD, J., KATES, Dr. J., & LEVITT, L. (2020, April 16). COVID-19 Models: Can They Tell Us What We Want to Know? KFF. https://www.kff.org/policy-watch/covid-19-models/
4. Randall, T., Sam, C., Tartar, A., Murray, P., & Cannon, C. (n.d.). *More Than 199 Million Shots Given: Covid-19 Tracker*. Bloomberg. Retrieved February 20, 2021, from https://www.bloomberg.com/graphics/covid-vaccine-tracker-global-distribution/
5. Smart Vision Europe. (2020, June 17). *Crisp DM methodology*. https://www.sv-europe.com/crisp-dm-methodology/
6. Jones, C. (2017). *Software Methodologies*. Amsterdam University Press.
7. Reeve, A. (2013). *Managing Data in Motion: Data Integration Best Practice Techniques and Technologies (The Morgan Kaufmann Series on Business Intelligence)* (1st ed.). Morgan Kaufmann.