CHAPTER THREE

RESEARCH METHODOLOGY AND SYSTEM ANALYSIS

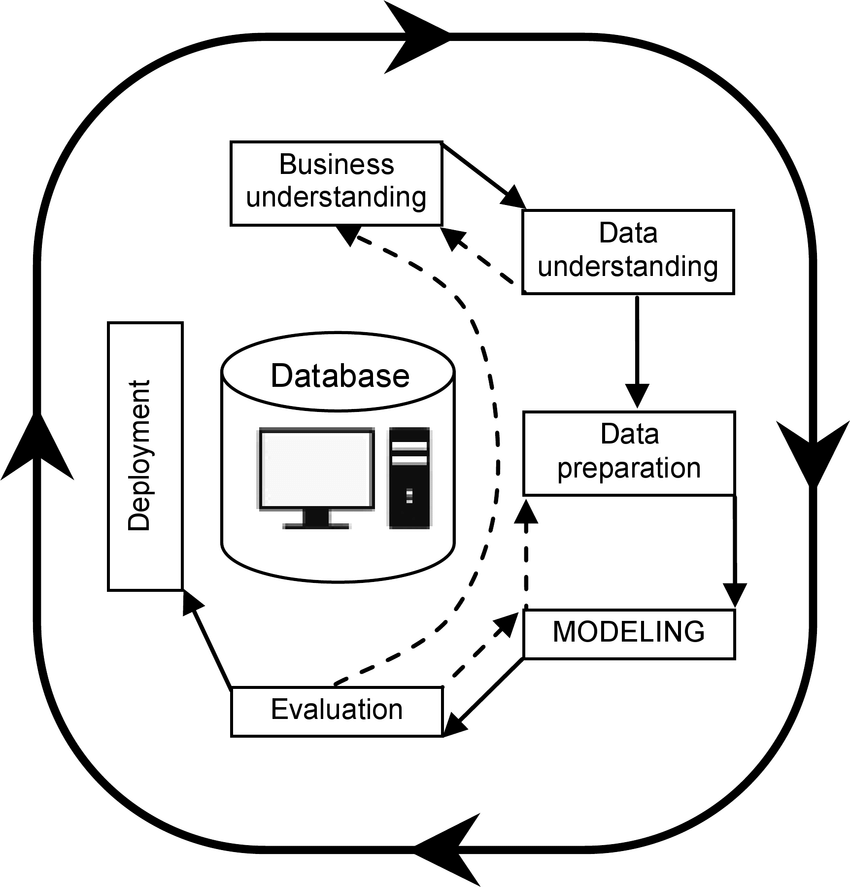
* 1. Introduction

Here in this chapter, a detailed evolutionary or investigation of the present system is carried out to help give us an in-depth knowledge of the present system so as enable us to identify the problems presented by this. For a proper understanding of the present system we decided to carry out this research by extensively studying the whole from the point of sourcing information, its arrangement, documentation storage, and distribution of appropriate quarters.

3.2 Research Methodology

The methodology is a formalized approach to the system development lifecycle (SDLC). It is also part of any analysis or research that is used to find out the type of data that is to be maintained, how it is found and how it is recorded. for this research, the cross-industry standard process for data mining (CRISP-DM) will be adopted.

CRISP-DM methodology is a standard cyclic framework used for the analysis of all data mining, predictive analysis, and data science projects (Pamanalytics, 2020). It helps one to plan, organize, and implement a data science project without stress. CRISP- MD methodology has six phases that naturally describe the data science life cycle as shown in figure 25 below. They have cycle iterations with the following steps: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.



**Figure 3.1: CRISP-DM Diagram of Data Science Life Cycle(Researchgate)**

The methodology flow begins with a proper understanding of the success criteria of the problem intending solving. The problem needs must be well-defined to know what the project can accomplish. The next step is to ascertain the type of data that is needed. Secondary data collection techniques to extract data from the official website of the Nigeria Center for Disease and Control. The site will be explored and searched for related data.

Data preparation refers to all the activities required to construct the final dataset out of the initial raw data. It involves the pre-processing of the dataset by cleaning the dataset and also checking for missing values and replacing them. About 85% of the success of a data science project is centered on the data preparation process. This phase takes the longest time because involves the following steps: selection of data, cleaning the data, constructing the data, integrating the data, and formatting the data. The next thing to do is feature extraction from the dataset which involves extracting useful information from the dataset. The dataset is classified into attributes and targets to suit the purpose of the research. The dimensionality reduction approach of feature extraction methods is studied in detail and computed in this proposed approach to make comparisons.

After data preparation comes modeling. A particular model is adopted that will fit into the context of the project. This phase is divided into four steps, namely:

1. Selection of modeling techniques.
2. Generate a test design based on the model technique by splitting the into training, test, and validation set.
3. Build the model(s) which can be executed by a few lines of code.
4. Iteratively assess the models until the best is selected.

During the evaluation, different classification algorithms will then be developed to train on the dataset and the one with the best accuracy will use to predict the system. The final phase is dependent on the nature of the requirements. It defines how stakeholders of the new system will access the result. The result can be as simple as generating a report or as complex as implementing a data mining process repeatedly in an organization.

3.3 Benefits of the CRISP-DM methodology

1. CRISP-DM is well-suited for machine learning projects
2. It has a uniform standard template that is made up of six phases thereby making it very simple.
3. A loose CRISP-DM is flexible as it adopts a principle that attracts the benefits of other methodologies like agile and waterfall.
4. This method is cost-effective as it integrates several processes.
5. CRISP-DM can be implemented in any data science project irrespective of its domain.
6. CRISP-DM is a cross-industry platform as such it is generally accepted.

3.4 Data Gathering Techniques

Based on the nature of this research, the method of data gathering technique that was adopted is secondary data. This involves the use of data that already exists from established sources. The research adopted secondary data because of the unavailability of covid-9 records which can be used for the development of the model. Therefore, online data was the source of our data. The data for the analysis was scrapped from the official website of the Nigeria Center for Disease and Control using the python beautiful soup module which is the only source to get a suitable dataset for the analysis. For this analysis, quantitative data was adopted because the dataset for the analysis comprises numerical details and information.

3.5 Analysis of the Existing System

The existing system of information gathering, tracking, and retrieving in Nigeria for the management of the COVID-19 outbreak is handled manually and unprofessionally. There was no platform for people to report suspected persons with COVID-19. That is to say, reporting of cases of COVID-19 is done mainly in isolation centers, and hence most times law enforcement agencies are not able to identify the person that entered the country either because of a bribe or no proper technology to do so. In transportation companies in Nigeria today, people still travel from state to state without proper medical checks for COVID-19, and even if the medical checks are done, fast reporting or call to the COVID-19 emergency response team are not fast contacted, therefore, the new system is designed to fill in this gap by providing a platform for private, public organizations, people to provide information on suspected persons with COVID-19 after medical examinations conducted. Furthermore, because the recording of infected or suspected persons of COVID-19 is done manually made accurate data storing, retrieving, tracking, and updating is difficult and time wastage sometimes leads to data loss. "There are no how restrictions to accurate data to researchers and technologist will not affect the rate of technological advancement or products development that could assist the medical personnel's and government fight the coronavirus". It was because of access to this information that made china design numerous technological devices that helped manage the disease in their country and other countries to determine if they are going for a total lockdown or partial lockdown based on the perdition of the system. Therefore, the researchers urge the Nigerian government to build a data bank or warehouse (if not in existence) and make it open for researchers and technologists to use in building software and hardware or techniques that could help in managing this deadly disease.

Furthermore, there are quite a several machine learning analyses that have been performed on covid-19 cases in Nigeria but all these models are based on mathematical modeling and also no predicting model was developed using the Nigeria covid-19 dataset from the official website. There is always a need for the development of a model capable of predicting the incidence of covid-19 based on the information gathered, this is what led to the development of the new system.

3.6 Analysis of the new system

The new system is an enhancement of the existing system. As such, all the features of the existing system were enhanced and improved upon by using a different and more robust algorithm in the analysing covid-19 dataset obtained from the official website of NCDC. The system was modeled to extract a dataset from the website and use the extracted dataset to build a model capable of analyzing the covid-19 cases in Nigeria. The process involves in developing the new system starts with the extraction of data from the site, till the final prediction was performed. These processes include data scrapping, data collection, Data exploration, model selection, model training, model evaluation, and finally model deployment.

3.6.1 Data Scrapping and Data Collection

Data collection is a crucial step in the machine learning process. Machine learning algorithms require large amounts of data to learn and make predictions. This data can come from a variety of sources, such as text files, databases, or web scraping.

The data for this analysis was automatically extracted from websites. It involves making HTTP requests to a website's server, downloading the HTML of the web page, and then parsing that HTML to extract the data you're interested in.

Python is a popular language for web scraping, with libraries like Beautiful Soup and Scrapy making it easy to extract information from HTML and XML documents. For this research, the extraction was done using a python library called Beautifulsoup.

3.6.2 Data Exploration

This stage involves analyzing and understanding a dataset before building a machine-learning model. It is an important step in the machine learning process as it helps to identify patterns, relationships, and trends in the data. This information can then be used to inform the model selection and feature engineering process. For this research, the following common techniques are used during the exploration stage using various libraries in python such as pandas, NumPy, and matplotlib. These libraries provide functions to perform data exploration and visualization.

1. Visualizing the data using plots and charts, such as histograms, scatter plots, and heatmaps, to gain insights into the distribution, relationship, and outliers of the features.
2. Summarizing the data by calculating statistics, such as mean, median, and standard deviation, to get a general understanding of the data.
3. Identifying correlations between features to understand how they relate to one another.
4. Checking for missing or duplicate values in the data, as these can impact the performance of the model.

3.6.3 Model Selection and Training

Model selection and training are two important steps in the machine-learning process.

Model selection refers to the process of choosing the appropriate algorithm or architecture for the problem at hand. This decision is based on the characteristics of the data and the problem to be solved. For this analysis, the polynomial regression model is considered for the problem because of the distributed nature of the dataset. Linear regression couldn't fit perfectly into the dataset because the parameter is not linearly proportional to each other. There are various libraries in Python used in achieving this such as sci-kit-learn. After the selection of the appropriate algorithm, the next stage is the training of the model. Training a machine learning model involves feeding it a large amount of data and adjusting the model's parameters so that it can make accurate predictions on new data. The process of training a model is also known as fitting the model. The first step of the training was splitting the data into testing and training data. The training data is used to optimize the model's parameters by minimizing a loss function, which measures the difference between the model's predictions and the true values of the output variable. This process is typically done using an optimization algorithm such as gradient descent.

3.6.4 Model Evaluation

Evaluating a polynomial regression model is similar to evaluating any other type of regression model. For this polynomial regression, we use visual evaluation methods. We plotted the predicted values against the actual values and observe how closely the predictions match the actual values. We can also plot the residuals, which are the differences between the predicted and actual values, and check if there is any pattern in the residuals, indicating a problem with the model.

Another way uses to evaluate a polynomial regression model is by comparing it with other models such as a linear regression model. This was done by comparing the performance of the models using the evaluation metrics mentioned above.

3.7 Data Flow Diagram (DFD) of the New System

The Data Flow Diagram of the new system is shown below

Feature extraction

Polynomial Regression

Result

ML Algorithms

Data classification

Machine learning model

dataset

Data pre-processing

Figure 3.2 Data flow diagram for the proposed system

### 3.8 Proposed System Design

The design of this model was done using object–oriented modeling. The user actions, system actions, class diagrams, and sequence diagrams used to explain the flow of the algorithm design from data dataset preparation to the final model design are explained in this chapter.

#### 3.8.1Interface Design

The Graphical User Interface enables the attraction of users to the system. The part of the model, in which the user interacts, is the graphical user interface generated by the Jupiter Notebook which is viewed through the browser. Other visible entities in the model are the plots and graphs generated by matplotlib. The final deployed application is based on a web browser for viewing and input into the system.

3.8.2 Sequence Diagram

This is a UML tool that helps to know the flow of work to be done and how to accomplish it. The dataset is processed to remove inconsistencies in the data after which the data is trained using a machine learning algorithm on the train data and the performance of the algorithm is tested using the test data to generate the result of the prediction.

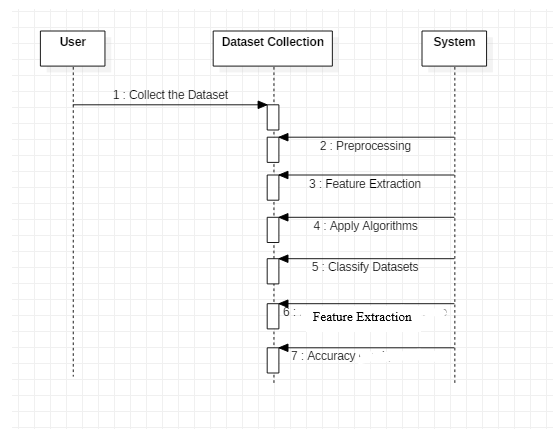


Figure 3.3: Sequence Diagram

### 3.8.3 Activity Diagram

An activity diagram is a flow chart to represent the flow from one event to another. The event can be described as an operation of the system; therefore, the control flow is drawn from one operation to another. This flow can be parallel, sequential, branched, or concurrent. The flow of information for the model is described below

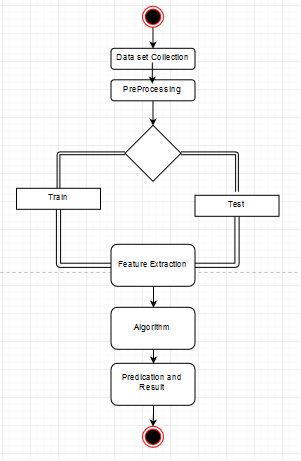


Figure 3.4: Activity Diagram

The activity diagram above illustrate the process involves in the development of the proposed system.