University of Ibadan
Journal of Science and Logics in ICT Research

Prediction of the Effectiveness of Government Measures towards Covid-19 Using Multiple Linear Regression Analysis

[1]Fagbuagun, O. A., [2]Folorunsho, O., [3]Nwankwo, O. and [4]Akinpelu, S. A.

[1] Department of Computer Science, Federal University Oye-Ekiti: abayomi.fagbuagun@fuoye.edu.ng
[2] Department of Computer Science, Federal University Oye-Ekiti: olaiya.folorunsho@fuoye.edu.ng
[3] Department of Computer Science, Novena University Ogume, Delta State: tuk2obinna@gmail.com
[4] Department of Computer Science, Federal University Oye-Ekiti: Samson.akinpelu@fuoye.edu.ng

Abstract
Several millions of people around the globe have been affected by the emergency of Coronavirus disease 2019 (COVID-19) pandemic. This menace has caused a geometrical mortality rate and stressed medical facilities in many counties. The lack of immediate treatment for the disease propelled the government of various countries to put in place some control measures to contain the rapid spread of the disease. Some of the measures taken include: staying at home orders, restricting movements, closing schools and workplaces, etc. This paper aims at determining the efficacy of these measures towards the containment of Covid-19. The Oxford Covid-19 Government Response Tracker (OxCGRT) dataset was used for this research. The data sets consist of daily entries of covid-19 cases in countries and various governments' active Covid-19 control measures/policies. Each policy indicator is given an ordinal score to denote its stringency. The data were analysed to gain insight into feature relationships and trends. Pearson's correlation coefficient (PCC) was used to select four features that contributed the most to the response variable. The training was carried out on 80% of the dataset using the python scikit-learn implementation of the Linear Regression algorithm, while testing was carried out on the remaining 20%. The model was trained on two features: The Containment Health Index which aggregates the other features, and the total number of Covid-19 cases. It achieved a coefficient of determination (r-squared) score of 0.09.

Keywords: *Containment Health Index, Machine Learning, Regression, Pandemic.*

1. Introduction

The outbreak of Severe Acute Respiratory Syndrome, also known as COVID-19 in Wuhan, China, was first reported in the year 2019, and it is a fatal pandemic in that it claimed millions of lives worldwide. According to [1], it is a shocking pandemic for humans due to the high rate of fatality recorded from the disease. Consequently, Covid-19 was declared the number six public health emergency that is of concern to the international community [2]. The disease outbreak was a major health risk to humans due to the rapid mode of transmission of the disease. The disease has spread across international borders within a very short time interval. Therefore, every country put a controlled and proactive response to contain the spread of the surge to the barest minimum. The National Health Commission of China disclosed that among the confirmed cases of Covid-19, the mortality rate was put at 2.1% in February [3]. The mortality rate among patients outside China was 0.2% [3], while within Chinese hospitals was about 11% and 15% [4, 5]. For instance, the statistical record from the Ministry of Health in Ethiopia indicated that the first case of Covid-19 in March 2020 [6] and that by March 2021, above 284.8 million instances of the disease recorded with 5.4 million deaths globally [7]. The rapid increase in the number of confirmed cases and the mortality rate worldwide has led the government of each country of the world to put in place some control measures to prevent the rapid spread of the disease.

In Nigeria, for example, some of the control measures include contact tracing immediately an index case was established in February 2020 [8]. Other policy measures include lockdown order, travel ban, the introduction of overnight curfew, physical distancing and mandatory wearing of face masks, and the closing of public transport air, land, and sea transports. Other control measures include a socio-economic policy such

Fagbuagun, O. A., Folorunsho, O., Nwankwo, O. and Akinpelu, S. A. (2021). Prediction of the Effectiveness of Government Measures towards Covid-19 Using Multiple Linear Regression Analysis. *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR),* Vol. 7 No. 1, pp. 1-9

as food assistance, economic stimulus, stimulus packages, basket fund, etc [8].

Despite all these controls measures, it has been discovered that the spread of the deadly COVID-19 continued even though at a much reduced rate in some countries of the world, while in some others, it is hard to tell whether a relationship exists between the spread of Covid-19 and the government measures against the disease in countries of the world. The focus of this project work is to predict the efficacy of governments' measures towards the control of COVID-19 pandemic using the technique of machine learning.

Machine Learning

Machine learning allows the users of a system to input a great deal of data into a system by using powerful computer algorithms. It enables the computer to analyse the data and make data-driven suggestions or recommendations, or decisions that are only the result of the data that was input into the system. In computational science, machine learning focuses on analysing and interpreting patterns in data, thereby enabling reasoning, learning, and decision-making entirely without human interaction or influence [9]. In machine learning, a huge amount of data is needed for the algorithm to work on to make a system do analysis. Data-driven recommendations are made while decisions based on the input data are reported. In any machine learning task, the most crucial task is classification [10] and classification has applications in various machine learning domains. With machine learning, systems can learn and improve from experience gained from training without being necessarily programmed. This is often called the most popular technology common to the industrial revolution [10, 11]. According to [12, 13], the data characteristics, the nature of data, and the learning algorithm's performance determine how efficient a machine learning solution to a problem will be. Furthermore, to model systems driven by data, learning algorithms exist, clustering, feature extraction and dimensionality reduction, learning by association, feature engineering, and reinforcement learning as tools to enhance the ability of model systems to make inferences from data.

Regression Analysis

According to [14], regression analysis is one of the several machine learning methods used to estimate the relationship that exists between a dependent variable and one or more independent variables. Regression is a way of making data-driven decisions to draw conclusions from existing data patterns. In [15] regression analysis was described as a method of mathematically sorting out which of the independent variables does have an impact on the dependent variable. In this research work, the independent variables used are the various government control measures. The output is whether relationships exist between them or not as they apply to countries of the world in the containment of Covid-19. The most commonly used types of regression analysis are linear regression, ridge regression, and polynomial regression. For the purpose of this research work, linear regression is adopted, which can usually be simple regression or multiple linear regression.

Linear regression model is a mathematical description of the relationship that exists between dependent variable and the independent variable. This type of regression can be modelled as presented in Equation (1).

$$Y = a + bX + err \qquad (1)$$

where Y represents the dependent variable, X represents the independent variable, a and b represents intercept and slope respectively. Err is the error term also known as the residual which is the distance between each observation and the line of best fit. It can also be stated that the error is the difference between the regression prediction and the actual observation.

The multiple linear regression model of regression is similar to the one presented in Equation (1) but differs because multiple independent variables are involved in the model. This model can be expressed mathematically, as illustrated in Equation 2.

$$Y = a + bX_i + cX_2 + dX_3 + err \qquad (2)$$

where Y is the variable, X_1, X_2 and X_3 are the independent variables, a is the intercept while b, c, and d, are the slopes. The coefficients of regression *b, c,* and *d* indicate how the dependent variable is affected by a change in an

independent variable when all other independent variables are kept constant.

2. Research Methodology

The research design consists of the following steps as it is implemented on using the Python scikit-learn implementation of the Linear Regression algorithm. The first stage is the preprocessing of the data. The preprocessing tasks include removing all measures from the data that cannot contribute to the containment of covid-19. The total data used for the study was 112,102, which contains the daily cases of covid-19 from the World Health Organization's website. They record new cases, total cases, and new deaths and total deaths statistics for each country affected by covid-19.

2.1 The Dataset

Two datasets were used in this study; one consists of government control measures while the other consists of the daily records of Covid-19 cases and deaths for each country. The government measures dataset used was obtained from the Oxford Covid-19 Government Response Tracker (OxCGRT) dataset. The dataset consists of daily entries of countries and the active Covid-19 control measures/policies taken by their governments. Each policy indicator is given an ordinal score to denote its stringency. Each measure is accompanied by a flag column, which indicates whether the measure was applied generally or to a specific area. The dataset also consists of policy indices calculated by aggregating the sub-index scores given to the individual control measures. The indices of the individual component indicators is described in Equation 3,

$$index = \frac{1}{k}\sum_{j=1}^{k} I_j \tag{3}$$

where k is the number of component indicators in an index and I_j is the sub-index score for an individual indicator.

All the indices use ordinal indicators where policies are ranked on a simple numerical scale. Five non-ordinal indicators – E3, E4, H4, H5 and M1 were recorded, although these records were not utilized in the index calculations. Some indicators – C1-C7, E1, H1, and H4 – have an additional binary flag variable that can be either 0 or 1. For C1-C7, H1 and H5 this corresponds to the geographic scope of the policy. For E1, this flag variable corresponds to the sectorial scope of income support. For H5, this flag variable corresponds to whether the individual or government is funding the vaccination. Because different indicators (j) have different maximum values (N_j) in their ordinal scales, and only some have flag variables, each sub-index score is calculated separately. The different indicators are presented in Table 2.

Table 1: Indices and their ordinal values

Index name	*k*	C1	C2	C3	C4	C5	C6	C7	C8	E 1	E 2	E 3	E 4	H 1	H 2	H 3	H 4	H 5	M1
Govern-ment response index	15	x	x	x	X	x	x	x	x	X	x			x	x	x	x	x	
Contain ment and health index	13	x	x	x	X	x	x	x	x					x	x	x	x	x	
Stringen cy index	9	x	x	x	X	x	x	x	x					x					
Governm ent response index	15	x	x	x	X	x	x	x	x	X	x			x	x	x	x	x	

Table 2: Indicators and their values

Indicator	Name	Max. Value (N_j)	Flag? (F_j)
C1	School_closing	3_(0, 1, 2, 3)	yes=1
C2	Workplace_closing	3_(0, 1, 2, 3)	yes=1
C3	Cancel public_events	2_(0, 1, 2)	yes=1
C4	Restrictions on public gatherings	4_(0, 1, 2, 3, 4)	yes=1
C5	Close public_transport	2_(0, 1, 2)	yes=1
C6	Stay_at_home requirements	3_(0, 1, 2, 3)	yes=1
C7	Restrictions on internal_movement	2_(0, 1, 2)	yes=1
C8	International travel_controls	4_(0, 1, 2, 3, 4)	no=0
E1	Income support	2_(0, 1, 2)	yes=1
E2	Debt / control relief	2_(0, 1, 2)	no=0
H1	Public information campaigns	2_(0, 1, 2)	yes=1
H2	Testing policy	3_(0, 1, 2, 3)	no=0
H3	Contact tracing	2_(0, 1, 2)	no=0
H4	Facial coverings	4_(1, 0, 1, 2, 3, 4)	yes=1
H5	Vaccination policy	5_(1, 0, 1, 2, 3, 4, 5)	yes=1

Each sub-index score (I) for any given indicator (j) on any given day (t), is calculated by the Equation:

$$I_{j,t} = 100\frac{v_{j,t} - 0.5(f_j - f_{j,t})}{N_j} \qquad (4)$$

Where: N_j *is the* maximum value of the indicator, F_j indicates whether an indicator has a flag (F_j=1 if the indicator has a flag variable, or 0 if the indicator does not have a flag variable), $v_{j,t}$ *is* the recorded policy value on the ordinal scale, $f_{j,t}$ *is the* recorded binary flag for that indicator. The different ordinal scales are normalised to produce a sub-index score between 0 and 100 where each full point on the ordinal scale is equally spaced. For indicators that do have a flag variable, if this flag is recorded as 0 then this is treated as a half-step between ordinal values. The features are coded as shown in Table 3.

Table 3: Feature coding

ID	Name	Description	Measurement	Coding
C1	C1_School _closing	Closure of schools and universities' record	Ordinal scale	0 – no closing of schools 1 - closing of all school recommended or all schools open with alterations resulting in significant differences compared to non-Covid-19 operations 2 - some levels / categories required to close 3 – all levels closed down Blank – no_ ata
	C1_Flag		Binary flag for geographic scope	0 -targeted 1-general Blank – no_data
C2	C2_Workpla ce__closing	Closure of workplaces record	Ordinal scale	0 - no measures 1 - recommend closing (or recommend work from home) or all businesses open with alterations resulting in significant differences compared to non-Covid-19 operation 2 - require closing (or work from home) for some sectors or categories of workers 3 - require closing (or work from home) for all-but-essential workplaces (eg grocery stores, doctors) Blank – no_data

	C2_Flag		Binary flag for geographic scope	0 - targeted 1- general Blank - no data
C3	C3_Cancel_public_events	Public events cancellation record	Ordinal scale	0 - no measures 1 - recommend cancelling 2 - require cancelling Blank - no data
	C3_Flag		Binary flag for geographic scope	0 - targeted 1- general Blank – no_data
C4	C4_Restrictions_on_gatherings	Record on limitation of gatherings	Ordinal scale	0 - no restrictions 1 – limitation set on large gathering (no gathering of more than 1000 people allowed) 2 - gathering limited to between 101 to 1000 people 3 - gathering limited to between 11 to100 people 4 - gathering limited to a maximum of 10 people Blank - no data
	C4_Flag		Binary flag for geographic scope	0 - targeted 1- general Blank - no data
C5	C5_Close_public transport	Record of closure of public transportation	Ordinal scale	0 - no measures 1 - recommend closing (or significantly reduce volume/route/means of transport available) 2 - require closing (or prohibit most citizens from using it) Blank - no_data
	C5_Flag		Binary flag for geographic scope	0 - targeted 1- general Blank - no_data
C6	C6_Stay_at_home requirements	Record orders to "shelter-in-place" and otherwise confined to the home	Ordinal scale	0 – when there is no measure in place 1 - recommend not leaving house 2 - require not leaving house with exceptions for daily exercise, grocery shopping, and 'essential' trips 3 - require not leaving house with minimal exceptions (eg allowed to leave once a week, or only one person can leave at a time, etc) Blank - no_data
	C6_Flag		Binary flag for geographic scope	0 - targeted 1- general Blank - no_data
C7	C7_Restrictions on internal movement	Record of restrictions on inter-cities/regions	Ordinal scale	0 - no measures 1 - recommend not to travel between regions/cities 2 - internal movement restrictions in place Blank - no_data
	C7_Flag		Binary flag for geographic scope	0 - targeted 1- general Blank - no data
C8	C8_International_travel controls	International travel restriction's record	Ordinal scale	0 - no restrictions 1 - screening arrivals 2 - quarantine arrivals from some or all regions 3 - ban arrivals from some regions 4 - ban on all regions or total border closure Blank - no_data
H1	H1_Public_information campaigns	Record of public info campaigns presence	Ordinal scale	0 - no Covid-19 public information campaign 1 - public officials urging caution about Covid-19 2- coordinated public information campaign (eg across traditional and social media) Blank - no_data

	H1_Flag		Binary flag for geographic scope	0 - targeted 1- general Blank - no data
H 2	H2_Testing _policy	Record of government policy on those who have access to testing	Ordinal scale	0 - no testing policy 1 - only those who both (a) have symptoms AND (b) meet specific criteria (eg key workers, admitted to hospital, came into contact with a known case, returned from overseas) 2 - testing of anyone showing Covid-19 symptoms 3 - open public testing (eg "drive through" testing available to asymptomatic people) Blank no_data
H 3	H3_Contact _tracing	Record of government policy on contact tracing after tested positive	Ordinal scale	0 - no contact tracing 1 - limited contact tracing; not done for all cases 2 - comprehensive contact tracing; done for all identified cases
H 4	H4_Facial_ Coverings	Record policies on the use of facial coverings outside the home	Ordinal scale	0 - No enforcement of usage of facial covering 1 – Enforcement 2 - Required in some specified shared/public spaces outside the home with other people present, or some situations when social distancing not possible 3 - Required when outside home where others are present or where social distancing might not possible 4 - Required only outside homes at all times
	H4_Flag		Binary flag for geographic scope	0 - targeted 1- general Blank - no_data

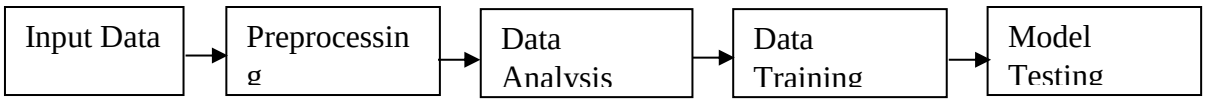

Figure 1: System Architecture

The second dataset is the daily cases dataset which was obtained from the World Health Organization's Website. It records new cases, total cases, new deaths and total deaths statistics for each country.

3. System Architecture

The architecture of the system is presented in Figure 1.

3.1 Preprocessing Since the dataset includes measures such as income support, debt relief and other economic and miscellaneous measures that do not contribute to containment, all measures that do not belong to the containment and health category were removed. Sub-national records such as those belonging to states and regions were also removed. New cases and total cases data were joined to the government measures dataset based on the date column on both datasets. A new column was then formed by calculating the percentage of the new cases based on the total number of cases using Equation 5.

$$P = \left(\frac{N}{T} \right) \times 100 \qquad (5)$$

where P is the percentage increase in cases, n is the new cases recorded, and T is the total cases recorded. The new feature was calculated as a percentage due to the high range of the data. The *Containment Health Index* feature and the *Number of Cases* feature were used as the training features while the calculated *Percentage Increase* served as the target feature. Other operations carried out on the data

are normalisation by scaling and filling missing values with the mode for each feature.

3.2 Data Analysis

The data was analysed to gain insight into feature relationships and trends in the data. The Pandas data analysis library was used for analysis, while matplotlib was used to visualise the results. Pearson's correlation coefficient (PCC) selected four features that contributed the most to the response variable. PCC finds the covariance of a single feature with the response variable and then divides it by the product of their standard deviation. The PCC is defined as:

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma x \, . \sigma y} \qquad (6)$$

Where *cov* is the covariance, σ_x is the standard deviation of x, σ_y is the standard deviation of y. This process revealed that the top four features are: Facial covering requirements, Testing policy, Limitation of social gatherings, and Public information campaign.

3.3 Model Training and Testing

The training was carried out on 80% of the dataset using the python scikit-learn implementation of the Linear Regression algorithm while testing was carried out on the remaining 20% of the dataset. The coefficient of determination (r-squared) score was recorded. The model was trained on two features: The Containment Health Index, which aggregates the other features, and the total number of Covid-19 cases. An interface presented in Figure 2 was designed to allow a user to easily input the features (government measures) by using dropdown selection menus and to generate the predicted efficacy by pressing a button.

4. Result and Analysis

The model was trained on two features: The Containment Health Index which aggregates the other features, and the total number of Covid-19 cases. It achieved a coefficient of determination (r-squared) score of 0.09. The scatter plot of the Containment Health Index against the target feature explains this low score as shown in Figure 3.

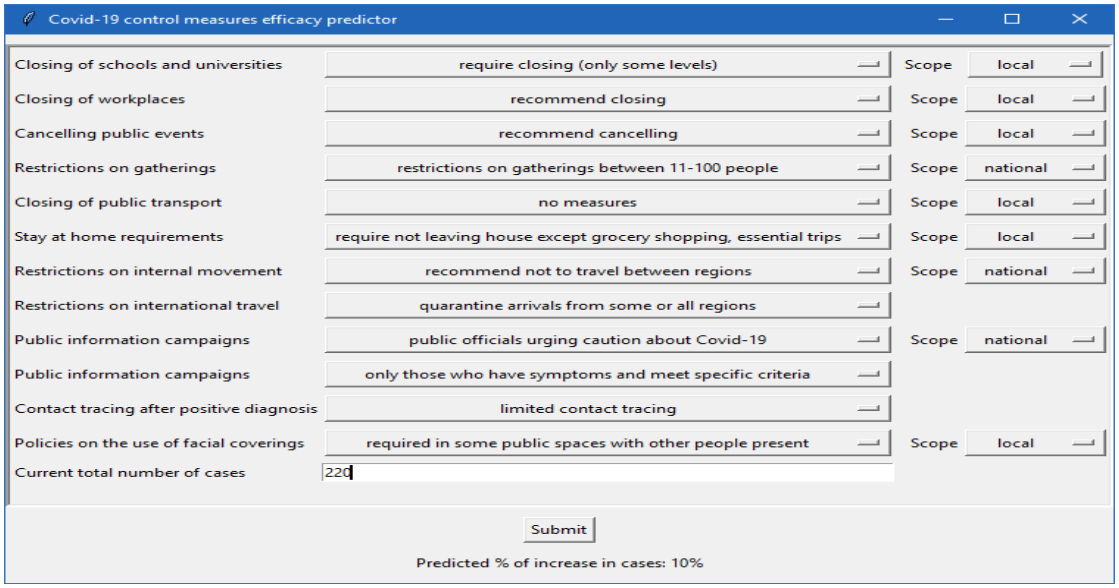

Figure 2: User Interface for the System.

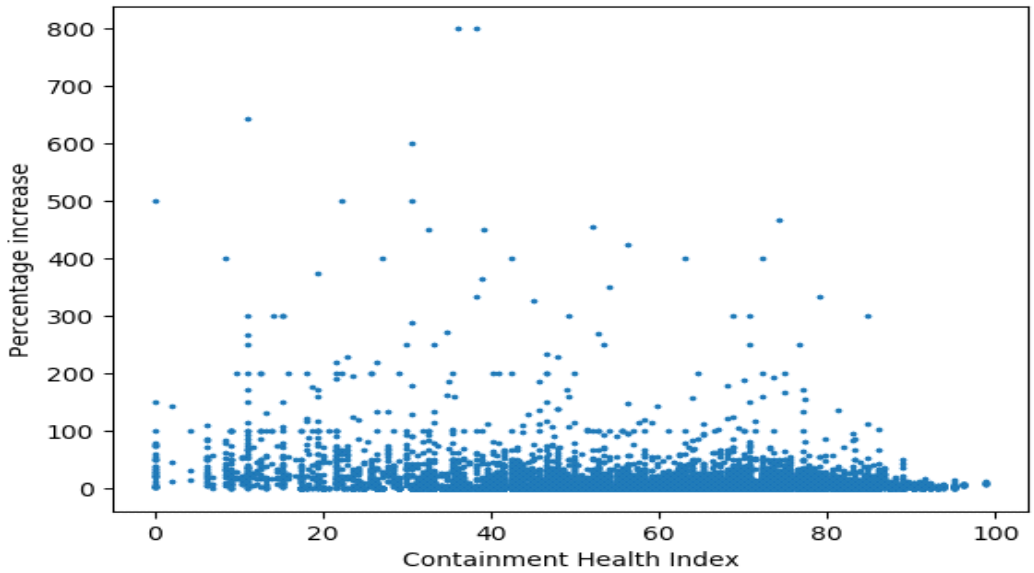

Figure 3: Percentage increase and containment index.

It can be seen from the plot that there is little correlation between the two features. This can be due to differences in levels of compliance to government policies in different countries, the state of countries' health sectors before the pandemic, variations in case reporting policies and many other factors. The study revealed little correlation between the containment indices and the rate of increase in covid-19 cases.

5. Conclusion and Recommendation

The whole world is facing the wave of COVID-19. There is a need to take a necessary precaution and put the spread of the pandemic at the barest minimum. Therefore, this article proposed utilising multiple linear regression analysis to predict the effectiveness of Government measures towards containment of the virus. A limitation in the approach is that factors such as socio-political differences, cultural differences and developmental differences vary among the countries considered. Another factor that vary among the countries is the level of support given to the vulnerable groups to assist them in complying with the measures aimed at controlling and containing the spread of the disease. The variation in these factors may have led to the result obtained. It is recommended that the research be carried out for individual countries of the world in order to know the efficacy of government control measures towards the containment of Covid-19.

References

[1] Agegnehu, B., Abera, M., Azene, T., Behailu, T. (2021). Adherence with COVID-19 Preventive Measures and Associated Factors Among Residents of Derashe District, Southern Ethiopia. This article was published in the following Dove Press journal: Patient Preference and Adherence downloaded from https://www.dovepress.com/ by 197.156.93.185 on 14-May-2021.

[2] Shewasinad, Y. S., Asefa, K. K., Mekonnen, A. G., Gemeda, B. N., Shiferaw, W. S., Aynalem Y. A, *et al.* (2021). Predictors of adherence to COVID-19 prevention measure among communities in North Shoa Zone, Ethiopia based on health belief model: A cross-sectional study. PLoS ONE 16(1): e0246006. pmid:33481962

[3] NHS press conference (2020). National Health Commission (NHC) of the People's Republic of China.http://www.nhc.gov.cn/xcs/xwbd/202002/235990d202056cfcb202043f202004a202070d20200 7f209703b202113c202000.shtml.

[4] Huang, C., Wang, X., Li, L., Ren, J., Zhao, Y., & Hu, *et al* (2020). Clinical features of patients infected with 2019 novel Coronavirus in Wuhan, China. Lancet 2020.

[5] Xiong, Q., Xu, M., Li, J., Liu, Y., Zhang, J., Xu, Y., & Dong, W. (2021). Clinical sequelae of COVID-19 survivors in Wuhan, China: a single-centre longitudinal study. *Clinical Microbiology and Infection*, *27*(1), 89-95.

[6] World Health Organization (2020). First case of Covid-19 Confirmed in Ethiopia. Retrieved from https://www.afro.who.int/news/first-case-covid-19-confirmed-ethiopia

[7] World meter (2021). COVID-19 update, Coronavirus Update (Live): 119,723,984 Cases and 2,653,796 Deaths from COVID-19 Virus Pandemic—World meter (worldometers.info) <https://www.worldometers.info/coronavirus/>

[8] Report (2020). Nigeria's Policy Response to Covid-19. Retrieved from https://centerforpolicyimpact.org/wp-content/uploads/sites/18/2020/06/Nigeria-National-Response-to-COVID19_FINAL.pdf.

[9] Ruchi, B., Jagdeep, S., & Ranjodh, K. (2019). Machine Learning and its Applications: A Review. *Journal of Applied Science and Computations*, Vol. 6, Issue 6, pp. 1392- 1398.

[10] Sarker, I. H., Hoque, M. M., MdK Uddin, & Tawfeeq A. (2020). Mobile data science and intelligent apps: concepts, AI-based modeling and research directions. Mobile Network Applications, pages 1–19.

[11] Sarker I.H., Kayes, A.S.M., Badsha, S., Alqahtani, H., Watters, P., & Ng A. (2020). Cybersecurity Data Science: An overview from machine learning perspective. J Big Data. 7(1):1–29.

[12] Han, J., Pei J, & Yin, Y. (2000). Mining Frequent Patterns without candidate Generation. In: ACM Sigmod Record, ACM. 2000;29: 1–12.

[13] Witten I.H., & Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann; 2005.

[14] Sarker I. H. (2021). Machine Learning: Algorithms, Real World Applications and Research directions. SN Computer Science (2021) 2:160

[15] Gallo, A. (2015). A Refresher on Regression Analysis. An online material retrieved on the 20th April, 2021 from https://hbr.org/2015/11/a-refresher-on-regression-analysis.