



# COVID-19 Risk Factors, Economic Factors, and Epidemiological Factors nexus on Economic Impact: Machine Learning and Structural Equation Modelling Approaches

David Opeoluwa Oyewola<sup>a</sup>, Emmanuel Gbenga Dada<sup>b,\*</sup>, Ndunagu Juliana Ngozi<sup>c</sup>, Terang A. U.<sup>a</sup>, Akinwumi S. A.<sup>a</sup>

<sup>a</sup>*Department of Mathematics and Computer Science, Federal University Kashere, Gombe, Nigeria*

<sup>b</sup>*Department of Mathematical Sciences, University of Maiduguri, Maiduguri, Nigeria*

<sup>c</sup>*Department of Computer Sciences, National Open University of Nigeria, Nigeria*

## Abstract

Since the declaration of COVID-19 as a global pandemic, it has been transmitted to more than 200 nations of the world. The harmful impact of the pandemic on the economy of nations is far greater than anything suffered in almost a century. The main objective of this paper is to apply Structural Equation Modeling (SEM) and Machine Learning (ML) to determine the relationships among COVID-19 risk factors, epidemiology factors and economic factors. Structural equation modeling is a statistical technique for calculating and evaluating the relationships of manifest and latent variables. It explores the causal relationship between variables and at the same time taking measurement error into account. Bagging (BAG), Boosting (BST), Support Vector Machine (SVM), Decision Tree (DT) and Random Forest (RF) Machine Learning techniques was applied to predict the impact of COVID-19 risk factors. Data from patients who came into contact with coronavirus disease were collected from Kaggle database between 23 January 2020 and 24 June 2020. Results indicate that COVID-19 risk factors have negative effects on epidemiology factors. It also has negative effects on economic factors.

DOI:10.46481/jnsps.2021.173

**Keywords:** COVID-19, Structural Equation Modelling, Latent variables, Random forest, Boosting.

## Article History :

Received: 14 March 2021

Received in revised form: 27 May 2021

Accepted for publication: 11 September 2021

Published: 29 November 2021

©2021 Journal of the Nigerian Society of Physical Sciences. All rights reserved.  
Communicated by: B. J. Falaye

## 1. Introduction

The negative impact of COVID-19 is felt by everyone in one way or another. The pandemic has created a situation whereby some people are more likely to experience severe illness because they have medical conditions that increase their risk. These are commonly called risk factors. Examples include age, race, gender, poverty and overcrowding, certain occupations and pregnancy [1]. Epidemiologic factors are definable entities that have

the potential to bring about a change in a health condition or other defined outcome [2], while macroeconomic factors are a trend or condition that comes from or applies to a broad aspect of an economy rather than a certain population. Common macroeconomic factors include gross domestic product, the rate of employment, phase of business cycle, rate of inflation and money supply [3].

From the time when it was first discovered in Wuhan in December 2019, the 2019 novel coronavirus also known as COVID-19 has quickly transmit to all regions, metropolises, and au-

\*Corresponding author tel. no:

tonomous provinces in China and has infected many countries in Asia, Europe, Oceanic, North America, South America, and Africa [4]. This novel virus has posed a serious challenge to preventing the spread of the deadly virus disease in many countries and regions and has had a great impact on economic, financial, commercial, and social development [5]. During the peak of this pandemic, most cities in different nations of the world have embraced closed management techniques which make businesses to operate on their own with little or no influence from the outside world. The purpose of this is to prevent further transmission of the virus and to lower the likelihood of more patients being infected [6, 7]. However, in the past few months, due to the unproductivity of most business and industrial activities in nations of the world, the great number of poor people have been confined to their house, and this has resulted to several social, economic and financial problems. It has also had an enormous impact on the economic and financial development of nations of the world [7].

COVID-19 is hypothetically a single-stranded enclosed viruses with an RNA with size of around 26 – 32 kilobyte [8]. World Health Organization (WHO) declared coronavirus as global public health emergency on 30 January 2020. This is because of to the sudden eruption of respiratory disorder. The novel coronavirus was classified by WHO as Severe acute respiratory syndrome coronavirus 2 (SARS-COV-2) and was termed the coronavirus disease 2019 (COVID-19) [9]. Respiratory symptoms, fever, dry cough, fatigue, sputum production, shortness of breath, sore throat, headache, myalgia or arthralgia, chills, nausea or vomiting, nasal congestion, diarrhea, hemoptysis and conjunctiva congestion are common symptoms of such infection. In critical cases of COVID-19 disease, the symptoms can lead to kidney failure, death and severe acute respiratory syndrome [10].

In the light of the above development, the appearance of the novel COVID-19 disease had elicited immense concerns on the science and art of preventing disease among the populace. This is also proved to have declining universal socioeconomic effects in due course. If the pandemic is left unchecked to continue wreaking havoc without any vigorous, reliable and sustainable effort or policy to improve health of the populace, then many economies around the globe will witness more reduced economic activities, and many will get poorer than before [11]. The economic impact of COVID-19 on the nations of the world cannot be overemphasized. Industrial plants and business facilities have been collapsed in a number of affected nations [11]. Also, the delivery of goods and services through a transnational corporations' worldwide network has been interrupted. For instance, the worsening universal economic impact of the COVID-19 endemic, and the feud between Saudi Arabia and Russia have made Brent crude prices to sell lower \$22 per barrel. This happens to be the least selling price since 2003 [12]. With the impending economic downturn as a consequence of the endemic, the situation can only be salvaged if adequate measures are put in place [12]. For people living in underdeveloped and developing countries with densely-populated houses, limited hygiene, and unavailability of funds to ease avoiding contacts with people, the needy are at higher risk of getting in-

fectured. Furthermore, the world is at the risk of seeing more people fall below the poverty line as a result of the high cost of medical treatment, increased economic shock, financial crisis, and increased number of deaths. As these viruses transcend borders, the global impacts will keep on spreading. It has been reported that about 94% of businesses across the globe have been negatively affected, and are now experiencing COVID-19 interruptions [12]. While it is expected that the COVID-19 threat will sooner or later disappear just like as the Ebola, Zika, and SARS viruses that have plagued the nations of the world in the last few years. Nevertheless, social-economic impact will still linger even after the virus is gone.

Machine learning (ML) algorithms have been applied to solve problems in different domain by analyzing and interpreting large quantity of data. Machine Learning has aided in the detection and identification of diseases, as well as drug discovery, medical imaging, smart health records, radiotherapy, robotic surgery and pharmaceutical development. Many researchers use machine learning algorithms to solve economic and financial problems.

Supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning are all forms of machine learning. How well a machine-learning system works depends on the type of data it uses, how well the learning algorithms function. Structural Equation Modelling (SEM) is a sophisticated multivariate analytic approach commonly utilized in social sciences. It may be used for a variety of purposes, ranging from determining basic connections between variables to doing more complicated studies of measurement equivalency for simpler notions. A number of SEM analytical methods are combined. These include comparisons of variance between and within groups, which are usually linked to ANOVA. It also incorporates path analysis, which involves solving equations that describe the influence of one or more variables on others in order to evaluate the strength of their connection. As a result, path analysis illustrates the predicted causal links between the variables being investigated.

Since the beginning of the pandemic, some studies have been conducted to provide better understanding of the COVID-19 factors that are making negative impact on the global economy. In this paper ML algorithms and Structural Equation Modelling (SEM) was applied to predict the impact of COVID-19 pandemic on global economy. ML algorithms have proved over the years that they are very efficient and robust algorithm that successfully cope with huge data. They can therefore be used to prudently predict the impact of COVID-19 risk factors on economic factors. This paper analyzes the correlation among COVID-19 risk factors, economic factors, and epidemiology factors and their impact on the COVID-19 crisis. Having an understanding of this will further help in policy formulation that will assist in mitigating against the effect of the pandemic. Moreover, it has the potential to positively impact labor productivity and economic growth of any nation. The major contributions of this paper are:

1. A survey of different ML algorithms and SEM that have been applied to predicting COVID-19 risk factors, eco-

conomic factors, and epidemiology factors was presented.

2. Application of machine learning techniques and SEM on epidemiological data of COVID-19 infection cases in South Korea was discussed.
3. Performance of the all ML algorithms and SEM was evaluated using different performance metrics.

The rest of this paper is organized as follows: Section 2 is the literature review. Section 3 discusses the materials and methods used in this work as well as our employed performance measurements. The results and the discussion are presented in section 3 and section 4 is the conclusion.

## 2. Literature Review

COVID-19 has captured the attention of researchers around the world. In this section, we present reviews of the impacts of COVID-19 on business, economies, transportation and so on. The coronavirus outbreaks have spread to 215 countries worldwide with a total of 43,824,534 cases, 1,165,290 deaths and 32,205,492 recovered from these diseases by 27 October 2020 [13]. The pandemic has tremendous consequences in the economy, although the exact magnitude of the effect is still uncertain. Many countries around the world had already implemented partial or absolute lockdowns [14]. Governments take emergency steps to control the epidemic, such as social distancing, quarantine and care of the reported cases to control the illness on one side.

The authors [15] suggested image processing of time series crude oil price through the integration of Directed Acyclic Graph to Convolutional Neural Network (DAG). The results indicated that combining DAG with CNN increases forecast accuracy by 14.18%. and it was established that COVID-19 negatively impacts the Nigerian crude oil price, suggesting a decreasing trend in the crude oil prices. The effects of financial markets on the COVID-19 pandemic was discussed in a study by [16]. During the period of 22 January 2020 through 17 April 2020, the researcher uses COVID-19 confirmed cases, death and market prices of stocks from 64 countries. The outcome was that the stock markets have a negative reaction to growth in COVID-19 cases. In other words, stock market decreased as the number of reported cases grew. In addition, they find proactive response on financial markets with the rise in the numbers of suspected cases as opposed to the increase in deaths. The developing countries were severely impacted by the COVID-19. The authors [17] mentioned the effect of COVID-19 on transportation in Lagos, Nigeria. This study was based on an email and social media administered to the residents of the Lagos state from 18<sup>th</sup> to 24<sup>th</sup> May 2020, to assess the effects of COVID-19 on transport in Lagos. Findings have shown that the transportation has been negatively affected by the pandemic. There is also a positive association between COVID-19 and transportation with its effect on people's economic, social and religious practices.

The estimated rate of deaths in India from SARS-COV-2 for 6 weeks from day 0 to 100 on 14 March 2020 was predicted by Ghosal et al [18] using multiple and linear regression. Findings indicate that week 6 death counts are not statistically significant

while week 5 death count is statistically significant. For sixty-four days, two months and three days the author [19] compiled data from the Nigeria Center for disease control. They employed three different linear regression such as quadratic, cubic, and quartic. The result shows that quartic linear regression model with an autocorrelated error of order one performed better. Sharif et al [20] examined the linkage within spread of COVID-19, oil price volatility shock, the stock market, geopolitical risks as well as US economic policy. In their application, they show the unprecedented effect of COVID-19 on geopolitics, economic policy instability and the market volatility of lower rate bands, as well as oil price shocks, through the wavelet dependent Granger causality tests. The findings show that the COVID-19 has a considerably higher impact on the geopolitical risk than on the US financial uncertainty.

The authors in [21] examined the effect of COVID-19 on the correlation between crude oil and agriculture. The cross-correlation between the crude oil of Brent and agriculture future of London Sugar, London Wheat, USA cotton and USA orange juice have been studied using a multifaceted cross-relationship analysis. The results demonstrated the strongest link between Brent Crude oil and London Sugar future and three other future agricultural markets. They also investigated the impact of COVID-19 on cross-correlations between crude oil and agriculture. The findings exhibited the greatest cross correlation between Brent Crude Oil and the London Sugar future market among other three future agriculture markets. The results showed that COVID-19 persistence has become stronger and the correlation between the multifractal relations between the crude oil and the sugar future markets are strongest. Overall analysis show that COVID-19 has a strong impact on the correlation between the crude oil and selected future agriculture market and multifractal properties.

In this analysis, the authors in [22] employed four different machine learning techniques such as linear regression (LR), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), and exponential smoothing (ES) to predict the threatening factors of COVID-19. The study considered three types of model such as the number of newly infected cases, the number of deaths, and the number of recoveries in the next 10 days. The results showed that the ES performed better than the remaining three machine techniques. Six machine learning techniques such as decision tree, support vector machine, Naive Bayes, logistic regression, random forest, and K-nearest neighbor algorithms were used by [23]. The model predicted that COVID-19 patients would recover from the virus for minimum and maximum days, that the patients with high levels of risk would not recover from the COVID-19 pandemic, the patients with potential recovery and those likely to recover quickly from COVID-19. The results show that decision tree performs better than other algorithms.

The spread of SARS appears to be influenced by the weather. During the first 16 weeks of the pandemic, [24] conducted a worldwide scale research that included 134,871 virologic climatic demographic data from 209 nations. The relationship between COVID 19, population density, and climate was studied using Structural Equation Modeling (SEM). The findings of

the study revealed that the spread of COVID 19 is influenced by both climate and population density. The author [25] investigated the factors that might impact public perceptions of Indonesia's Pembatasan Sosial Berskala Besar (PSBB). Partial Least Squares and structural equation models were used and data were collected from 856 respondents across Indonesia's provinces. The advantages of the PSBB, positive perception, negative perception, threatening perceptions of COVID-19 and attitude toward the PSBB policy were all utilized to evaluate these policies. More than half of the attitudes toward PSBB policy implementation can be explained by the model, which takes into account perceived advantages, negative and positive views, and the danger posed by COVID-19.

The authors in [26] used structural equation modeling to examine the relationship between socio-demographic variables (gender, age, level of education, place of residence, and employment status) and COVID-19 preventive behaviour and the threat appraisal of COVID-19, fear of COVID-19, trust in COVID-19 information sources, and COVID-19 conspiracy beliefs. COVID-19 assessment of threat, confidence in COVID-19 information sources, and fear of COVID-19 are all significant predictors of COVID-19 preventative activity, according to the results. COVID-19 conspiracy theories have a negative correlation with threat assessment and trust in COVID-19 information sources. COVID-19 danger assessment has an important and direct role in explaining COVID-19 phobia. A study conducted by [27] used Structural Equation Modeling to predict how work-life balance will be affected by factors such as their own health and emotional well-being as well as their current relationship status as well as their location of employment. Findings showed that elements including physical and mental health, activities, relationship status, and place of employment directly affected the work-life balance. There was a notable gender gap among dentists, with far fewer women than men. Structural equation modeling was used by Franzen et al. [28] to simulate a group of young individuals who were polled shortly after the conclusion of Switzerland's initial lockdown. To find out why and how much they helped in averting the pandemic by following the advice to stay at home as much as possible. They believe that people who believe they are at danger, or who have relatives in the risk category, are more likely to follow safety precautions than people who do not believe they are at risk. Coronavirus social separation procedures were well followed during the first shutdown, according to research. Young individuals felt the virus posed little personal risk, but society as a whole was at risk. Furthermore, the findings show that support for preventative measures is the most significant factor in fostering collaboration in the effort to contain the spread of COVID-19.

### 3. Methodology

#### 3.1. Description of Dataset

The dataset used in this research comprises of epidemiological data of COVID-19 infection cases in South Korea which were obtained from Kaggle database [29] and macro-economic data was obtained from yahoo finance [30]. The dataset is

made up of data from 23/01/2020 to 24/06/2020 recorded daily, patient\_id, sex, age, country, province, city, infected\_by, contact\_number, symptom\_onset\_date, confirmed\_date, released\_date and state which consists of released, deceased and isolated. In this study, due to the nature of the dataset, we extracted sex, age, state, confirmed\_date, released date while the macro-economic dataset obtained from yahoo finance which consists of South Korea exchange rate (KR), Jakarta Composite Index (JK), KOSPI Composite Index (KS) as shown in Table I. It displays sex (male=1, female=2), Age (numeric), state (released=1, deceased=2, isolated=3), DR is obtained from subtracting released\_date from confirmed\_date. Table I shows that sex, KS, JK and KR is negative moderately skewed while Age, state and DR is positively skewed. The data distribution of Sex, Age, State, JK is platykurtic since the data distribution is less than 3 while DR, KS and KR is leptokurtic since the data distribution is greater than 3. Sex and State have a very low variance while Age, DR, KS, JK, KR have a very high variance. This is an indication that the data points are widely spaced from each other and this may also result in high degree of error. With this intuitive knowledge at hand, Table II is the variance reduction of dataset. Fig. 1 shows a scatter plot of matrices, with bivariate scatter plots below the diagonal, histograms on the diagonal and Pearson correlation above the diagonal. A scatter plot shows a relationship sex, age, state, DR, JK, KS and KR. The Pearson correlation coefficients of the data shows a positive relationship between sex and age, negative relationship between state and DR, positive relationship between KS and JK, JK and KR and KS and KR. Also, the diagonal of Fig.1 shows the histogram of all the observed variables. There is more female than male in the sex variable, there is more recorded cases of COVID-19 within the age of 20s than the age of 90s. In the state variable, there is more released of patient of COVID-19 than death. DR, JK, KS and KR are numeric values.

#### 3.2. Structural Equation Modelling

Structural equation modeling (SEM) is also known as the causality model or covariance structural model. It is a method for defining, estimating and evaluating the causality model [31]. It comprises a range of statistical analytical methods including confirmatory factor analysis, variance, covariances, regression and latent growth curve. It is a very broad and linear method of statistical modeling that tests hypotheses according to theories. This model was first presented by Wright [32-33]. SEM equation are split into measurement model and structural equation model. Measurement model primarily tests the correlation between latent and significant variables while structural equation model tests mainly causality among the latent variables. The key characteristics of scientific research are the estimation, relativizing variables and disclosure of causality [34]. Moreover, observable or manifest variables such as sex, age, state, DR, JK, KS and KR can be calculated while latent variables such as COVID-19 risk factors, epidemiological factors and macro-economic factors cannot be directly measured as shown in Fig. 2. In such cases, regression equalities should be defined which demonstrate how endogenous and exogenous structures are related and which benefits from a statistical technique, which has

Table 1. Variance Reduction of Dataset

Variables	Number of Sample	Minimum	Median	Mean	Variance	Skewness	Kurtosis
Sex	3782	0.00	2.00	1.54	0.25	-0.19	1.05
Age	3782	0.00	40.00	40.38	408.03	0.31	2.33
State	3782	1.00	1.00	1.66	0.87	0.72	1.55
DR	3782	1.00	1.00	10.93	204.23	1.50	5.20
KS	3782	0.00	1938	1815.04	163848.07	-1.18	5.50
JK	3782	0.00	4091	3767.94	3153274.83	-0.24	1.84
KR	3782	0.00	1110	1086.97	19496.45	-3.29	29.63

Table 2. Variance Reduction of Dataset

Variables	Number of Sample	Minimum	Median	Mean	Variance	Skewness	Kurtosis
Sex	3782	0.00	2.00	1.54	0.25	-0.19	1.05
Age	3782	0.00	4.00	4.04	4.08	0.31	2.33
State	3782	1.00	1.00	1.66	0.87	0.72	1.55
DR	3782	0.10	0.10	1.09	2.04	1.50	5.20
KS	3782	0.00	1.93	1.82	0.16	-1.18	5.50
JK	3782	0.00	4.09	3.76	3.15	-0.24	1.84
KR	3782	0.00	1.11	1.08	0.02	-3.29	29.63

a broad range of applications to combine measurement principles such as SEM [35]. The risk factors of COVID-19 include age, race/ethnicity, gender, some medical conditions, use of certain drugs, poverty and crowding, certain occupations and pregnancy [36].

Due to the sparsity of data of COVID-19 risk factors, we considered only the sex (Sex) and age (Age) of epidemiology of South Korea COVID-19. Epidemiology factors consists of state (State) of COVID-19 patients and duration (DR) is obtained from subtracting released\_date from confirmed\_date of epidemiology of South Korea COVID-19. COVID-19 has catastrophically affected economy. The survival of economy greatly relies on the crude oil price and other macro-economic factors [37]. In this study, we considered three macro-economic factors such as South Korea exchange rate (KR), Jakarta Composite Index (JK) and KOSPI Composite Index (KS). Fig. 3 is the schematic diagram of structural equation modeling of impact of COVID-19 risk factors on epidemiology and economic factors. Circles are displayed as Latent variables; Squares are displayed as Manifest, measured or observed variables; Arrows displayed the paths from latent variables to observed variables; Residuals and variances are indicated as double headed arrows. The latent variables are  $I$ ,  $Y$ ,  $Z$  while the observed variables are  $sex$ ,  $age$ ,  $state$ ,  $DR$ ,  $KR$ ,  $JK$ ,  $KR$ . The paths from latent variables to observed variable are  $\lambda_1$  to  $\lambda_6$ .  $\varphi_1$  to  $\varphi_3$  double headed curve arrows are path from each latent variables while  $\theta_1$  to  $\theta_{10}$  are the double headed curve arrows for both observed and latent variables. The structural equations of impact of COVID-19 risk factors on epidemiology and economic factors can be re-

resented as:

$$I = \alpha + sex\beta_1 + age\beta_2 + \xi_1 \quad (1)$$

$$Y = \rho + state\beta_3 + DR\beta_4 + \xi_2 \quad (2)$$

$$Z = \gamma + KR\beta_4 + JK\beta_5 + KS\beta_6 + \xi_3 \quad (3)$$

$$I = Y\eta + Z\pi + \xi_4 \quad (4)$$

Where  $I$  is the COVID-19 risk factors,  $Y$  is the Epidemiology factors,  $Z$  is the economic factors,  $\alpha$ ,  $\rho$ ,  $\gamma$  are the intercept,  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$ ,  $\beta_4$ ,  $\beta_5$ ,  $\beta_6$  are the predictor observable variable,  $\xi_1$ ,  $\xi_2$ ,  $\xi_3$ ,  $\xi_4$  are the residual error,  $\eta$ ,  $\pi$  are latent predicted variable.

### 3.3. Machine Learning

#### 3.3.1. Bagging (BAG)

Bagging which is an acronym for Bootstrap Aggregating is a Parallel ensemble technique. It provides a way to decrease the variance of prediction model throughout the training phase by producing extra data. This is accomplished through arbitrary sampling and substitution from the original data. Decisions made by multiple learners can be integrated into a single prediction. In the case of classification, it is clearly a vote to combine these decisions. Models of bagging bear the same weight as good models of bagging because an executive can use a collection of expert advice based on their previous right predictions to achieve other outcomes. It is considered right which one gets more votes than other groups. If more votes are expected, they are reliable because more votes are present [38]. BAG is used in this paper because of its capacity to minimize the variance of

Table 3. Correlation Coefficient of Observed Variables

	Sex	Age	State	DR	KS	JK	KR
Sex	1.00	0.12	-0.03	0.03	0.02	0.02	0.01
Age	0.12	1.00	0.08	-0.01	0.08	0.11	0.08
State	-0.03	0.08	1.00	-0.28	0.11	-0.01	0.12
DR	0.03	-0.01	-0.28	1.00	0.20	0.25	-0.07
KS	0.02	0.08	0.11	0.20	1.00	0.76	0.15
JK	0.02	0.11	-0.01	0.25	0.76	1.00	0.23
KR	0.01	0.08	0.12	-0.07	0.15	0.23	1.00

Table 4. Summary of Impact of COVID-19 risk factor (I) on Epidemiology factor (Y) and Economic factor (Z).

Latent Variables	Estimate	Std.Err	Z-value	p-value
$I \sim \text{sex}$	1.00			
Age	21.84	16.62	1.32	0.19
$Y \sim \text{state}$	1.00			
DR	-68.89	199.89	-0.35	0.7
$Z \sim \text{KS}$	1.00			
JK	5.60	0.21	26.23	0.0
KR	0.10	0.01	13.98	0.0
Covariances				
$I \sim Y$	-0.00	0.00	-0.32	0.75
$I \sim Z$	0.00	0.00	1.31	0.19
$Y \sim Z$	-0.00	0.01	-0.34	0.73
Variance				
Sex	0.24	0.01	34.89	0.00
Age	1.46	1.98	0.74	0.46
State	0.86	0.03	34.17	0.00
DR	-23.53	73.83	-0.32	0.75
KS	0.07	0.00	17.30	0.00
JK	0.08	0.11	0.73	0.47
KR	0.02	0.00	43.29	0.00
$I$	0.01	0.00	1.27	0.20
$Y$	0.01	0.02	0.34	0.73
$Z$	0.09	0.01	19.94	0.00

a decision tree classifier. BAG enables a trade-off balance between variance and bias by reducing the variance and carefully adjusts the prediction to an estimated result. The mathematical equation that depicts the parameters used in bagging is in equation 5.

$$H(d_i, c_j) = \sum_{m=1}^M \alpha_m H_m(d_i, c_j) \tag{5}$$

Where  $H_m$  is the weak classifiers,  $d_i$  is classified to the classes  $c_j$  and  $\alpha_m$  is the constant parameter.

### 3.3.2. Boosting (BST)

Boosting is a successive ensemble technique that decreases bias error and produces outstanding prediction models. The word 'Boosting' describes a set of methods that transforms a poor learner into a strong learner. Stochastic Gradient Boosting (BST) method is a hybrid of boosting and bagging proposed by Friedman [39]. BST is a set of learning algorithm with a combination of boosting and decision tree, which classifies the value

of all trees by weighing all trees. The new model is constructed along the path of gradient descent of the loss function of the previous tree. It is important to note that the loss function between classification and actual function is reduced by the training function of the classification function [40]. BST technique was selected for use in this paper because the algorithm continues its iteration until a learner with superior results compared to a random guess is achieved. BST approach therefore helps in increasing the capability of machine learning and improving prediction accuracy.

The mathematics equation of the loss function is given in equation 6 and 7:

$$\rho(y_k, F_k(x)) = \sum_{k=0}^K y_k \log \left[ \frac{e^{F_k(x)}}{\sum_{k=1}^K e^{F_k(x)}} \right] \tag{6}$$

$$\widehat{y}_k = - \left[ \frac{\partial \rho(Y_k, F_k(x))}{\partial F_k(x)} \right] = y_k - P_k(x) \tag{7}$$

Where  $y$  is the output variable,  $x$  is the input variables,  $k$  is the

Table 5. Parameters Estimate of Structural Equation Modeling

Parameter	Value
$\lambda_1$	0.15
$\lambda_2$	0.80
$\lambda_3$	0.08
$\lambda_4$	-3.54
$\lambda_5$	0.77
$\lambda_6$	0.99
$\lambda_7$	0.23
$\theta_1$	0.98
$\theta_2$	0.36
$\theta_3$	0.96
$\theta_4$	-11.52
$\theta_5$	0.40
$\theta_6$	0.03
$\theta_7$	0.95
$\theta_8$	1.00
$\theta_9$	1.00
$\theta_{10}$	1.00
$\varphi_1$	-0.01
$\varphi_2$	-0.07
$\varphi_3$	0.14

Table 6. Performance evaluation of Machine Learning

Algorithm	RMSE	MSE	MASE
BAG	0.0541	0.0029	0.7758
BST	0.0512	0.0026	0.7685
SVM	0.0584	0.0034	0.8954
RF	0.0003	0.0187	0.2229
DT	0.0033	0.0577	0.8902

number of classes,  $P_k(x)$  is the probability.

### 3.3.3. Support Vector Machine (SVM)

SVM procedure categorizes both linear and non-linear data. SVM uses a non-linear mapping to transform the training set to a high level. In this new dimension, SVM explores the ideal linear hyperplane separation as a decision limit by which the tuples of a class of one class are split from another. There are two class data that can be separated by a hyperplane with the proper, non-linear upper dimensional mapping. In contrast to the other approaches, hyperplanes are highly robust for overfitting [41]. SVM is considered in this research due to its ability to handle numerous continuous and categorical variable. SVM technique is used in this paper because it has a low bias and a high variance, nevertheless, the trade-off can be modified by tuning the C parameter, which determines the number of infractions of the border permitted in training data, raising the bias while reducing the variance. Equations 8, 9 and 10 for SVM are stated below:

$$w^T \cdot x + b = 1 \tag{8}$$

$$w^T \cdot x + b = -1 \tag{9}$$

The set of inequalities can be combined to form:

$$y[w^T \cdot x + b] \geq 1 \tag{10}$$

The equation can be formulated as a minimization problem given in equations 11, 12 and 13:

$$\min_{w, \tau} J(w, \tau) = \frac{1}{2} w^T w + c \sum_{i=1}^N \tau \tag{11}$$

Subject  $y[w^T \cdot x + b] \geq 1$  to Lagrangian function, we then have

$$L(w, b, \alpha, \beta) = J(w, \tau) - \sum_{i=1}^N \alpha (y [w^T \cdot x + b] - 1 + \tau) - \sum_{i=1}^N \beta \tau \tag{12}$$

The optimal point of the Lagrangian function is given as

$$\max_{w, \beta} \min_{w, b, \tau} L(w, b, \alpha, \beta) \tag{13}$$

Differentiate (12), we obtain

$$\frac{\partial L}{\partial w} = 0, \quad w = \sum_{i=1}^N \alpha y x \tag{14}$$

$$\frac{\partial L}{\partial b} = 0, \quad \sum_{i=1}^N \alpha y \tag{15}$$

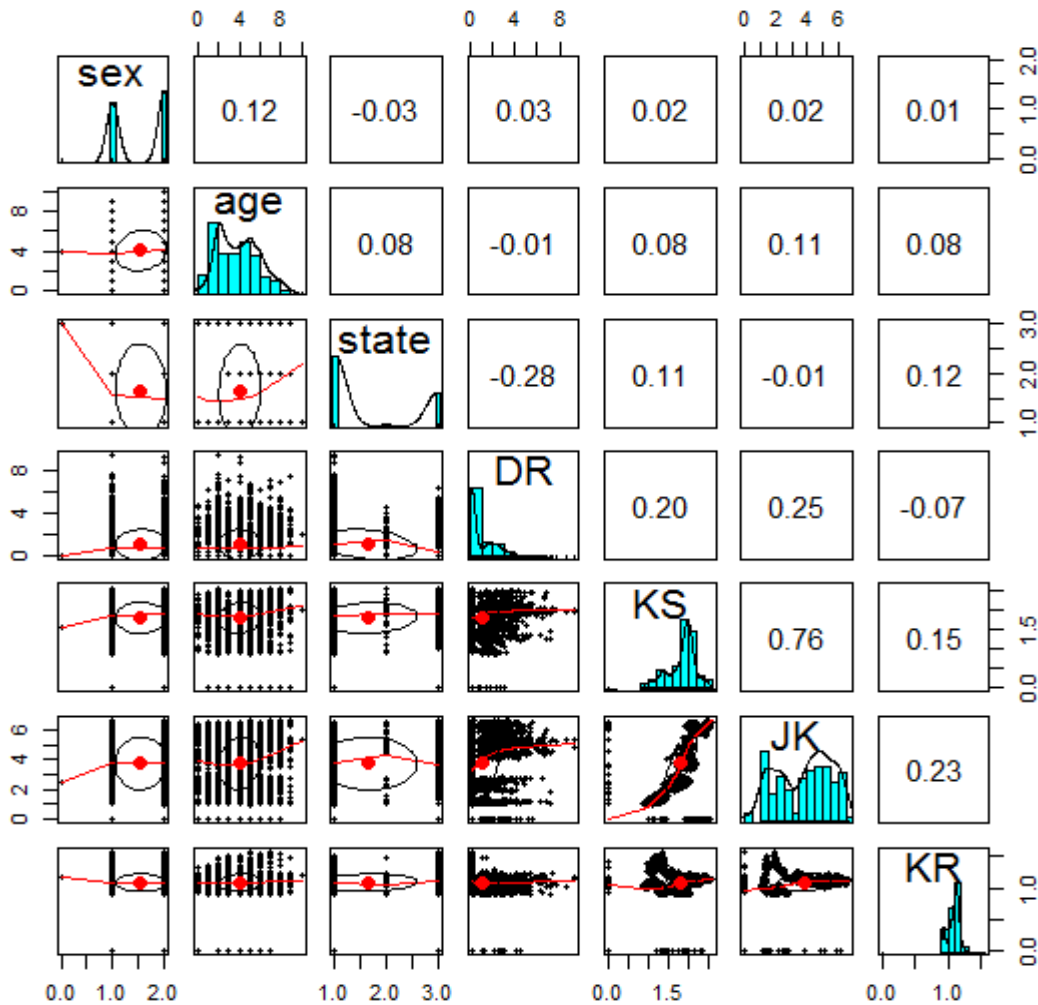


Figure 1. Pairwise Scatter plots of the dataset

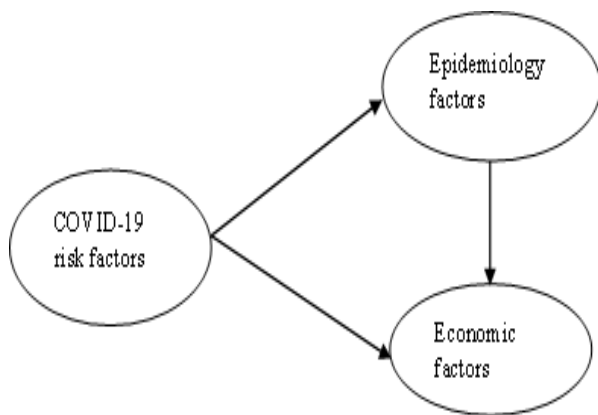


Figure 2. Impacts of COVID-19 risk factors with respect to epidemiology and economic factor

$$\frac{\partial L}{\partial \tau} = 0 \tag{16}$$

The quadratic programming problem will be form by substitut-

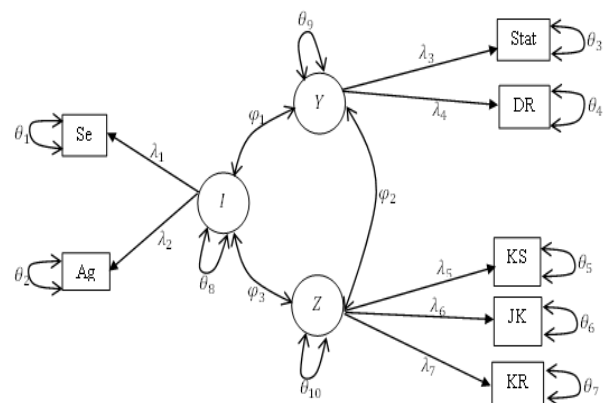


Figure 3. Schematic diagram of Structural Equation Modeling of Impacts of COVID-19 risk factors on epidemiology and economic factors. Circles are displayed as Latent variables; Squares are displayed as manifest, measured or observed variables; Arrows are displayed as paths from latent variables to observed variables; Residuals and variances are indicated as double headed arrows.

ing (14), (15) and (16) to equation (12), we then have

$$\min_{\alpha} \gamma(\alpha) = \sum_{i=1}^N \alpha - \frac{1}{2} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (17)$$

Where  $K(x_i, x_j)$  is the kernel function,  $w$  is the weight,  $x$  is the input,  $b$  is the bias,  $\alpha, \beta$  are the positive real constants.

### 3.3.4. Random Forest (RF)

Random forest (RF) is a decision-making ensemble classifier that has various types of trees. An arbitrary sequence of features at each node is used to evaluate the division to create an individual decision tree. Each tree is based on the individual values of a random variable. We are able to shape an RF using bagging along with the selection of the random attribute, using the CART method, in order to increase the trees. RF uses the random linear combination of the input attributes. The sub-cluster of features is not chosen randomly, but new attributes are created, which reflect a linear combination of existing features [42]. RF model will assist in the construction of numerous decision trees and their merging to produce a more accurate and reliable prediction. RF is an ensemble learning technique that applies the idea of Bagging. It provides a compromise between variance and bias by lowering the variance and judiciously fine-tunes the prediction to a desired result.

### 3.3.5. Decision Tree (DT)

Decision trees (DT) are an easy model which classify by dividing training data into pieces and mainly holding the result of each part [43]. It is a natural non-parametric supervised learning model, also called Classification and Regression Tree (CART) which produces accurate classifications with easily understood regulations. DT Models transparency makes them highly relevant for economic and financial purposes. In addition, continuous and discrete data can be dealt with using DT. Our choice of DT model in this work is based on its ability to fit the training data flawlessly fine.

## 3.4. Performance Evaluation

Three measures such as Root Mean Square Error (RMSE), Mean Square Error (MSE) and Mean Absolute Scaled Error (MASE) are used to calculate the prediction efficiency of impacts of COVID-19 risks factors with respect to epidemiology factors and economic factors.

### 3.4.1. Root Mean Square Error (RMSE)

RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{n} \sum_{n=1}^n (I_n - \hat{I}_n)^2} \quad (18)$$

Mean Square Error (MSE)

MSE is defined as:

$$MSE = \frac{1}{n} \sum_{n=1}^n (I_n - \hat{I}_n)^2 \quad (19)$$

Mean Absolute Scaled Error (MASE)

MASE is defined as:

$$MASE = \frac{1}{n} \sum_{n=1}^n \frac{|I_n - \hat{I}_n|}{\frac{1}{n-m} \sum_{n=m+1}^n |I_n - \hat{I}_{n-m}|} \quad (20)$$

Where  $I_n$  is the real COVID-19 risk factors,  $\hat{I}_n$  is the predicted values and  $m$  is the seasonal period of  $I_n$ .

## 4. Result and Discussion

This section presents the experimental results of structural equation modeling and machine learning techniques such as Bagging (BAG), Stochastic Gradient Boosting (BST), Support Vector Machine (SVM), Random Forest (RF), Decision Tree for predicting impacts of COVID-19 risk factors with respect to epidemiology and economic factors. We compared the performances of the algorithms under consideration using Root Mean Square Error (RMSE), Mean Square Error (MSE) and Mean Absolute Scaled Error (MASE) to discern which is more accurate in predicting the impacts of COVID-19 risk factors. As stated earlier, we used data from the Kaggle database for COVID-19 infection cases in South Korea.

Statistical correlation analysis was used to determine the strength of the association between observed variables. Table IV is the result of correlation coefficients of observed variables. A correlation coefficient of 0.12 was noted between the sex and age of COVID-19 risk factors, this indicates a weak positive linear relationship between them. A correlation coefficient of -0.28 was noted between state and DR of epidemiology factors, this indicates a weak negative linear relationship. There is a strong positive relationship between KS and JK while between JK and KR there is a weak positive relationship between them. However, KS and KR also show a weak positive linear relationship of the three economic factors such as JK, KS and KR.

Table V displays the estimate of latent and observe variable, covariance and variance of the COVID-19 risk factor ( $I$ ), epidemiology ( $Y$ ) and economic factor ( $Z$ ). The estimate, standard error, z-value and p-value are also shown in the table. The estimate of latent variable  $I$  and the observed variables such as sex and age are estimated as 21.84 with a standard error estimate of 16.62, z-value is 1.52 but it was observed that there is no statistically significant linear dependence of the mean of  $I$  with respect to age. This means that no effect was observed. Also, the estimate of latent variable  $Y$  and the observed variable such as state of COVID-19 (state) and duration of COVID-19 (DR) are estimated as -68.89 with a standard error estimate of 199.89, z-value is -0.35 but it is insignificant at the 0.05 level. There is statistically significant linear dependence of the mean of latent variable  $Z$  with respect to JK and KR. The covariance result of latent variable indicates that the covariance between  $I$ ,  $Y$  and  $Y, Z$  is approximately -0.00, which indicates that the relationship is negative while the covariance result of  $I, Z$  is approximately 0.00, which indicate that the relationship is positive. The small variance of sex, age, state, JK, KS, KR,  $I$ ,  $Y$ ,  $Z$  it shows that the data points appear to be very similar to average and to one another, while DR with high variance shows that the

data points are very widespread from the average and from each other.

Table VI is the Parameter estimate of Structural Equation Modeling. There is a positive relationship between latent and observed variable as shown in  $\lambda_1, \lambda_2, \lambda_3, \lambda_5, \lambda_6, \lambda_7$  except in  $\lambda_4$  which shows a negative relationship between them. The parameter estimates of  $\theta_1, \theta_2, \theta_3, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9, \theta_{10}$  also shows a positive relationship between them except in  $\theta_4$  which shows a negative relationship.  $\varphi_1, \varphi_2, \varphi_3$  is the impact of COVID-19 risk factors ( $I$ ) with respect to epidemiology factors ( $Y$ ) and economic factors ( $Z$ ). There is a negative effect of COVID-19 risk factors ( $I$ ) on epidemiology factors ( $Y$ ), negative effect also obtained on epidemiology factors ( $Y$ ) and economic factors ( $Z$ ) but positive effect was obtained from COVID-19 risk factors ( $I$ ) and economic factors ( $Z$ ). Table VII report RMSE, MSE and MASE of the extracted predicted value of COVID-19 risk factors. RF outperform the rest of the algorithms with smaller accuracy result than other methods, which means that the approaches are more effective than others. The findings show that RF perform well.

## 5. Conclusion

Structural equation modeling (SEM) has provided a means to understand direct impact of COVID-19 risk factors with respect to epidemiology factors and economic factors. Latent variable SEM has provided the necessary tools for developing several equations to describe the COVID-19 behavioural impact framework. It has the potential to quantify and test the relationships between latent and observed variables. They measure the uniformity and plausibility of the assumed model in relation to the findings observed. Furthermore, a researcher can examine both direct and mediate relationships. Findings indicate that COVID-19 risk factors have negative effects on epidemiology factors. It also has negative effects on economic factors.

The result indicates that there is no statistically significant linear dependence of the mean of COVID-19 with respect to Age. This means that no effect was observed. Also, the estimate of latent variable epidemiology and the observed variable such as state of COVID-19 and duration of COVID-19 is insignificant at the 0.05 level. Also, there is a negative effect of COVID-19 risk factors on epidemiology factors, negative effect also obtained on epidemiology factors and economic factors but positive effect was obtained from COVID-19 risk factors and economic factors.

Future research areas will consider impacts of COVID-19 on other factors such as environmental factors, socioeconomic factors, educational factors and so on.

## References

- [1] E.J. Williamson, A.J. Walker, K. Bhaskaran, S. Bacon, C. Bates, C.E. Morton, H.J. Curtis, A. Mehrkar, D. Evans, P. Inglesby, J. Cockburn, "Factors associated with COVID-19-related death using Open Safely", *Nature*, **584** (2020) 430.
- [2] X. Yewwei, W. Zaisheng, L. Huipeng, M. Gifty, W. Dan, T. Weiming, "Epidemiologic, clinical and laboratory findings of the COVID-19 in the

- current pandemic: systematic review and meta-analysis", *BMC Infectious Diseases* (2020) 1.
- [3] F. Zakaria, A. F. Filali, "The COVID-19: macroeconomics scenario and role of containment in Morocco", *One Health*, **10** (2020) 100152.
- [4] Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K.S. Leung, E.H. Lau, J.Y. Wong, X. Xing, "Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia", *N. Engl. J. Med.*, **382** (2020) 1199.
- [5] S. Roush, H. Fast, C.E. Miner, H. Vins, L. Baldy, R. McNall, S. Kang, V. Vund, "National Center for Immunization and Respiratory Diseases (NCIRD) Support for Modernization of the Nationally Notifiable Diseases Surveillance System (NNDSS) to Strengthen Public Health Surveillance Infrastructure in the US. In 2019", *CSTE Annual Conference*. CSTE
- [6] S.A. Ekanem, E.P.K. Imarenezor, C.P. Kolisah, "An Essencist Evaluation of Socio-Economic Impacts of Coronavirus Disease (COVID-19) Pandemic in Nigeria". *Mediterranean Journal of Social Sciences*, **11**(2020), 70.
- [7] A. Obioma, A.A. Reuben, A.B. Elekwachi, "Potential Impact of COVID-19 Pandemic on the Socio-Economic Situations in Nigeria: A Huge Public Health Risk of unprecedented Concern", *J Qual Healthcare Eco.*, **3** (2020) 000175.
- [8] L.L. Ren, Y.M. Wang, Z.Q. Wu, Z.C. Xiang, L. Guo, T. Xu, Y.Z. Jiang, Y. Xiong, Y.J. Li, X.W. Li, H. Li, "Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study", *Chinese medical journal*, **133** (2020) 1015
- [9] WHO. Novel Coronavirus–China: <https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/>. Accessed: 20 October, 2020.
- [10] ational Center for Immunization and Respiratory Diseases (NCIRD) DoVD. Coronavirus Disease 2019 (COVID-19) Situation Summary: Centers for Disease Control and Prevention. <https://www.cdc.gov/coronavirus/2019-nCoV/summary.html>. Accessed: 15 November, 2020.
- [11] From pandemic to poverty: Nigeria's future with COVID-19. (May 2020). Nairametrics. Retrieved from <https://nairametrics.com/2020/05/17/from-pandemic-to-poverty-nigerias-future-with-COVID-19/>. Accessed: 16 November, 2020.
- [12] COVID-19: A Business Impact Series. <https://home.kpmg/ng/en/home/insights/2020/04/COVID-19-a-business-impact-series.html>. Accessed 16 November, 2020.
- [13] Coronavirus: <https://www.worldometers.info/coronavirus/> Accessed: 27, October 2020.
- [14] K.B. Ajide, R.L. Ibrahim, O.Y. Alimi, "Estimating the impacts of lockdown on COVID-19 cases in Nigeria. *Transportation Research Interdisciplinary Perspectives*", **7** (2020) 100217.
- [15] D. O. Oyewola, A. F. Augustine, E. G. Dada, A. Ibrahim, "Predicting Impact of COVID-19 on Crude Oil Price Image with Directed Acyclic Graph Deep Convolutional Neural Network", *Journal of Robotics and Control*, **2** (2020) 103-109.
- [16] B.N. Ashraf, "Stock markets' reaction to COVID-19: Cases or fatalities?", *Research in International Business and Finance*, **54** (2020) 101249.
- [17] E. Mogaji, "Impact of COVID-19 on transportation in Lagos, Nigeria", *Transportation Research Interdisciplinary Perspectives*, **6** (2020) 100154.
- [18] S. Ghosal, S. Sengupta, M. Majumder, B. Sinha, "Linear Regression Analysis to Predict the number of deaths in India due to SARS-COV-2 at 6 weeks from day 0 to 100 cases March 14<sup>th</sup> 2020, *Diabetes & Metabolic Syndrome*", *Clinical Research & Reviews*, **14** (2020) 311-315.
- [19] K. Ayinde, F. A. Lukman, I. Rauf, O. O. Alabi, C. E. Okon, O. E. Ayinde, "Modeling Nigerian COVID-19 cases: A comparative analysis of models and estimators", *Chaos Solitons and Fractals*, **138** (2020) 109911.
- [20] A. Sharif, C. Aloui, L. Yarovaya, "COVID-19 pandemic, oil prices, stock market, geopolitical risk and policy uncertainty nexus in the US economy: Fresh evidence from the wavelet-based approach", *International Review of Financial Analysis*, **70** (2020) 101496.
- [21] J. Wang, W. Shao, J. Kim, "Analysis of the impact of COVID-19 on the correlations between crude oil and agricultural futures", *Chaos, Solitons and Fractals*, **136** (2020) 109896.
- [22] F. Rustam, A. A. Reshi, A. Mehmood, S. Ullah, B. W. On, W. Aslam, G. S. Choi, "COVID-19 Future Forecasting Using Supervised Machine Learning Models", *IEEE*, **8** (2020) 101489-101499.
- [23] L. J. Muhammad, M. M. Islam, S. S. Usman, S. I. Ayon, "Predictive Data

- Mining Models for Novel Coronavirus (COVID-19) Infected Patients' Recovery", *SN Computer Science* (2020) 1-7.
- [24] A. Spad, F. A. Tucci, A. Ummano, P. P. Ciavarella et al., "Structural equation modeling to shed light on the controversial role of climate on the spread of SARS-CoV-2", *Scientific Reports*, **11** (2020) 8358.
- [25] S.G. Purnama, D. Susanna, "Attitude to COVID-19 Prevention with Large-Scale Social Restrictions (PSBB) in Indonesia: Partial Least Squares Structural Equation Modeling" *Front. Public Health*, **8** (2020) 570394. doi: 10.3389/fpubh.2020.570394.
- [26] S. Šuri, K. Martinsone, V. Perepjolkina, J. Kolesnikova, U. Vainik, A. Ruža, J. Vrublevska, D. Smirnova, K.N. Fountoulakis, E. Rancans, "Factors Related to COVID-19 Preventive Behaviors: A Structural Equation Model", *Front. Psychol.*, **12** (2021) 676521. doi: 10.3389/fpsyg.2021.676521.
- [27] S. Pai, V. Patil, R. Kamath, M. Mahendra, D.K. Singhal, V. Bhat, "Work-life balance amongst dental professionals during the COVID-19 pandemic—A structural equation modelling approach", *PLoS ONE*, **16** (2021): e0256663. <https://doi.org/10.1371/journal.pone.0256663>
- [28] A. Franzen, F. Wohner, "Coronavirus risk perception and compliance with social distancing measures in a sample of young adults: Evidence from Switzerland", *PLoS ONE*, **16** (2021):e0247447. <https://doi.org/10.1371/journal.pone.0247447>
- [29] Kaggle: <https://www.kaggle.com/kimjihoo/coronavirusdataset>. Accessed: 18, September, 2020.
- [30] Yahoo Finance: <https://finance.yahoo.com/>. Accessed: 19, September, 2020.
- [31] Y. Liping, C. Yuqing, P. Yuntao, W. Yishan, "Research on the evaluation of academic journals based on structural equation modeling, *Journal of Informetrics*", **3** (2019) 304–311.
- [32] S. Wright, "Correlation and Causation", *Journal of Agricultural Research* **20** (1921) 557.
- [33] S. Wright. S (1934), "The method of path coefficients", *Annals of Mathematical Statistics* **5** (1934) 161.
- [34] S. Kocakaya, F. Kocakaya, "A Structural Equation Modeling on Factors of How Experienced Teachers Affects the Students Science and Mathematics Achievements", *Education Research International* (2014) 1.
- [35] J.H. Hair, R. L. Tatham, R. E. Anderson, "Multivariate Data Analysis", Prentice Hall International, New York, NY, USA, 5<sup>th</sup> edition, 1998.
- [36] COVID-19 Risk Factors: <https://www.cdc.gov/coronavirus/2019-ncov/COVID-data/investigations-discovery/assessing-risk-factors.html>. Accessed: 23 October, 2020.
- [37] D. O. Oyewola, A. F. Augustine, E. G. Dada, A. Ibrahim, "Predicting Impact of COVID-19 on Crude Oil Price Image with Directed Acyclic Graph Deep Convolutional Neural Network", *Journal of Robotics and Control (JRC)*, **2** (2020) 103-109.
- [38] D. O. Oyewola, E. G. Dada, O. T. Omotehinwa, I.A. Ibrahim, "Comparative Analysis of Linear, Non Linear and Ensemble Machine Learning Algorithms for Credit Worthiness of Consumers", *Computational Intelligence & Wireless Sensor Networks*, **1** (2019) 1-11.
- [39] J. H. Friedman, "Stochastic gradient boosting", *Computational Statistics & Data Analysis*, **38** (2002) 367–378.
- [40] Y. Shin, "Application of Stochastic Gradient Boosting Approach to Early Prediction of Safety Accidents at Construction Site", *Advances in Civil Engineering* (2019) 1-9.
- [41] S. Kim, J. Choi, "An SVM-based high-quality article classifier for systematic reviews", *Journal of Biomedical Informatics*, **47** (2014) 153.
- [42] R. Katuwal, P.N Suganthan, L. Zhang, "Heterogeneous Oblique Random Forest", *Pattern Recognition*, **99** (2019) 107078.
- [43] S. Sivakumar, S. Venkataraman, R. Selvaraj, "Predictive Modeling of Student Dropout Indicators in Educational Data Mining using Improved Decision Tree", *Indian Journal of Science and Technology*, **9** (2016) 1.