ISSN: 3092-8729 | e-ISSN: 3092-8737

Ajayi Crowther Journal of Pure and Applied Sciences (ACJPAS)
https://acjpas.acu.edu.ng

Ajayi Crowther J. Pure Appl. Sci. 2023, 2(1), pp. 24-31
https://doi.org/10.56534/acjpas.v2i1.77

Article

The Prediction Rate of COVID-19 using Random Forest Approach

[1]Ebenezer Olukunle Oyebode, [2]Olufemi Olayanju Awodoye, [1]Funmilola Wumi Ipeayeda

[1]Department Computer Science, Faculty of Natural Sciences, Ajayi Crowther University, Oyo, Nigeria,
eo.oyebode@acu.edu.ng (E.O.O.); fw.ipeayeda@acu.edu.ng (F.W.I.)
[2]Department of Computer Engineering, Faculty of Engineering, Ajayi Crowther University, Oyo, Nigeria,
oo.awodoye@acu.edu.ng (O.O.A.)

* Correspondence: (E.O.O.) eo.oyebode@acu.edu.ng; +234 803 248 1502

Article history; Received: Aug. 2, 2022; Revised: Oct. 1, 2022; Accepted: Oct. 5, 2022; Published: Jan 20, 2023.

Abstract

COVID-19 is a pandemic disease that claimed a lot of human lives and caused economic setbacks among other problems. Its effects became more negative and prolonged partly because of poor prediction rates that could give priviledge information for preparation against the disease. A dataset containing several information inform of records of suspected number of cases, deaths, location etc of 338 records were gathered from ECDPC. Random forest tree model was setup in Phython using ANACONDA. The random forest model setup was trained to predict the number of cases and likely number of deaths resulting from such cases within 1-10 days and in weeks. It was observed that the model accuracy in the prediction for the number of cases in days was 99.2% while that of the number of cases in weeks is 99.94%.

Keywords: Pandemic, model, random forest, dataset

1. Introduction

COVID-19 is a pandemic disease that has a lot of negative implications on human life. It affected economic activities and caused sorrow. It cut across all continents of the world and has tremendous negative influence to both developed and underdeveloped world. COVID-19 started on 31st Dec. 2019 in Wuhan city, Hubei province, China and reported as unknown form of Pnemonia with trace of the virus found at lower respiratory tract of Wuhan patients [2]. Within a short period it spread to other continents.

It was established that COVID-19 spread mechanism was through close contact between people and contaminated surfaces. Corona virus has been linked with extreme outbreaks because of its high transnational spread, increase trade relationships that marked globalization among other factors [11]. With globalization and inter-country relationship that facilitates traveling from country to country and place to place, the spread of COVID-19 operated at high exponential rate. Many symptoms that characterize infected patients include respiratory symptoms, breathing difficulties, cough, fever, shortness of breath etc. It can also result to very serious health conditions that could trigger pneumonia, severe acute respiratory syndrome, kidney failure and sudden death. For patients with disease such as hypertension, heart diseases, renal infection etc before infected with COVID-19 their health status became more critical. COVID-19 has negative effect on stock market due to lockdowns in many parts of the world [12]. It compelled government of various countries to spend heavily on stimulus packages to cushion the negative impact of the COVID-19 pandemic on households, health care, manufacturing and servicing industries [10].

Quarantine for affected patients, major lock down etc were part of the measures adopted to salvage the situation [3]. With such a pandemic disease, medical facilities and techniques were over stressed. This is because poor predictions were made, coupled with inadequate preparations and as a result many lives were lost. Various studies have shown the relationship between mortality and statistics of diagnosed COVID-19 cases [7]. It is important to verify the spread pattern of COVID-19 in different continents and to find out the prediction mechanism that is efficient. The prediction can help in the aspect of planning of resources, determine government policy and produce positive effects in the health sector and society at large.

A study on age-structured effects on social distancing on Chinese and other parts of the world revealed that short term lock-down will not be sufficient to control COVID-19 but prolonged lock down. Various studies involving statistical techniques such as time series models, multivariate linear regression, grey forecasting models, back propagation ANN, Recurrent ANN have been widely deployed [7]. It has been stated that quantitative approaches alone seem not to be adequate to determine the randomness of COVID-19 as it is difficult to generalize models. Some of the models used attempted to estimate the frequency prevalence and mortality levels of COVID-19 mostly in China, Italy and India. Some of the study include the use of adaptive Neurofuzzy Inference System (ANFIS) in combination with improved Flower Pollination Algorithm.

To mention a few, attempt was also made on using machine learning to predict end times of COVID-19 ranging from diagnosis to prediction. Susceptible-Infected Removed (SIR) model was used to predict the end time of the second wave of COVID-19 in different states in India. The medical science world and other professional disciplines that are involved in detail analysis including quantitative reasoning such as prediction of time series analysis etc. have been making efforts to overcome COVID-19 [3]. The time series parameters represent the major factors that affect the prediction accuracy of infectious diseases.

Many studies embarked on several forms of prediction in other to gain understanding on COVID-19. Such study attempted to predict the pattern of COVID-19 and linked the prediction to source of COVID-19 using the Center of Infection Mass (CoIM). The CoIM was defined using eq(1)

$$(X,Y)_{CoIM} = \left(\frac{\sum_{i\in n} x_i.m_i}{\sum_{i\in n} m_i}, \frac{\sum_{i\in n} y_i.m_i}{\sum_{i\in n} m_i}\right) \qquad \dots (1)$$

$(x_i, y_i)\ i \in (1..n)$ represents the cartesian coordinate system of the locations involved. m_i represents mass which is analogue to confirmed cases c_i. The model used an assumption derived from source of the COVID-19 and aggregated data of other countries up to countries being considered. When many countries are involved, the model may become more complex and accuracy can be affected.

Singh [1], proposed an attempt to use SVM for time series using world's total population of known cases was. The SVM optimization eq(2)

$$h = \min_{w,c} \frac{1}{2}\|w\|^2 \text{ subject to } y_i(mx + c)\text{-}1,\ i = 1\dots z \qquad \dots (2)$$

SVM uses a technique of upper limit for error by minimizing the boundary distance between the training data and hyperplane opposed to traditional ways of reducing observational testing errors. Some other approaches included predicting COVID-19 cases in different regions using ensembles [4]. The model used ensemble method of regression learners to predict incidence of COVID-19 in different regions using seven variables. The study used the COVID-19 results of first 14 days to predict the results of next 14 days. The output of the model gave global picture of likely estimates of cases of COVID-19 across different continents. Study on duration of hospitalization for COVID-19 was among related predictions. Ebinger *et al.* [6] applied 3 machine learning algorithms to predict recovery time of COVID-19 for patients. Such studies were meant to support the management of healthcare facilities for COVID-19. Other dimensions of the study on COVID-19 in Europe also focused on knowing the transmissibility and mortality rate of COVID-19. Yuan *et al.*

[5] setup two models that estimated real values base on exponential growth rate and dependent method. The RO package of an R-language coded statistical package was used in the study.

In this study, the use of Random forest algorithm has been proposed. The random forest algorithm has been applied in various ways and good results have been obtained. Random forest is a form of supervised classification and regression machine learning technique. The Random Forest represents an ensemble of decision trees. The Random Forest algorithm uses extra randomness when growing trees [9]. Random forest algorithm can solve regression problems by applying the mean squared error (MSE) to how data branches from each node

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \qquad \ldots (3)$$

Where N represents the number of data points, x_i represents the value returned by the model and
y_i represents the actual value for data point i.

Instead of searching for the very best feature among the subset of features, random forest can average the outcomes of multiple decision trees from subsets of a dataset in order to improve its predictive accuracy. It uses projections from selected tress to compute results. It can also compute the estimate of the feature's importance by obtaining the average depth at which it appears across all the trees in use that represents the forest. Random forest approach can be applied to identify or gain quick understanding of important features of a dataset. The opportunity to predict to the nearest accuracy the spread of disease may mitigate/reduce the inherent trade-off between life-safe and economy. It can positively influence the design of economic policy and boost welfarism.

2. Materials and Methods

The dataset preparation was done by cleaning the dataset. The first step in the cleaning process is the extraction of dataset that belongs to the case study which is Nigeria. It was followed by removal of unwanted fields such as geold, country territory and continent. It was then followed with steps to generate relevant statistics for the dataset. Partitioning the dataset into training and testing also was done in the ratio 80:20. The selected data include the data for Africa countries from which Nigeria was identified. The dataset represented contains dateRep, day, month, year, recorded cases, deaths, countries and Territories, geoId and country territoryCode [13]. The experimentation platform is Intel Core™ i3 CPU@2.3GHz 6GB of RAM running 64 bits of MS Windows. The preprocessing and model construction has been implemented in Phython using ANACONDA.

The model setup preparation began with invoking set of objects required for building the model including its parameters, setting up data structures that can handle multiple data sets and allow necessary computation without generating errors, specifying various constants. It also include given specification for the dimension of amount of data that will be used for training the model and the actual amount for testing the model. Training and retraining were carried out and actual performance measures were used as a stopping criteria.

A random forest is an ensemble of decision trees and so in depth understanding of decision tree is a part of the random forest model. A decision tree starts with a feature as a node at the top of the tree which splits or branches out into different results/outcomes [14]. The criteria for splitting can be based on the information gain. In Python environment, Scikit learns using the information gain through the Gini impurity. The gini impurity is a measure of the probability that reflects if a sample chosen randomly in a node may be incorrectly labeled using the distribution of samples within the node. Feature with the lowest gini impurity is used as the root of the decision tree. The algorithm tries to split on the feature that results in the lowest gini impurity. Using this, a decision tree can generate nodes containing a high proportion of samples from a single class by obtaining values from set of features that divide the data into classes [15]. In the case of random forest, the

majority output is a combination of results from the decision tress. The random forest algorithm is shown in algorithm 1.

Random forest algorithm1

Input: Dataset of Covid-19 for a country, maximum number of decision trees, performance measure criteria

Ouput: prediction output

Step1: Selection of n number of random records
Step2: Set loop to create n number of decision trees
Step3: generate n decision trees
Step4: end loop
Step5: Combine results from each decision tree and compare with accuracy measure
Step6: if result in step5 is less than expected accuracy goto step2 else proceed to load the testing data and generate results.

3. Results and Discussion

The Python environment is shown in Fig. 1. The plot for Nigeria containing the cases discovered is shown in Fig. 2. The plot showed the alarming rate of cases discovered in a pattern that is exponential. The death pattern that resulted from cases discovered is also shown in Fig. 3. This can be further seen in the rate of change of the cases.

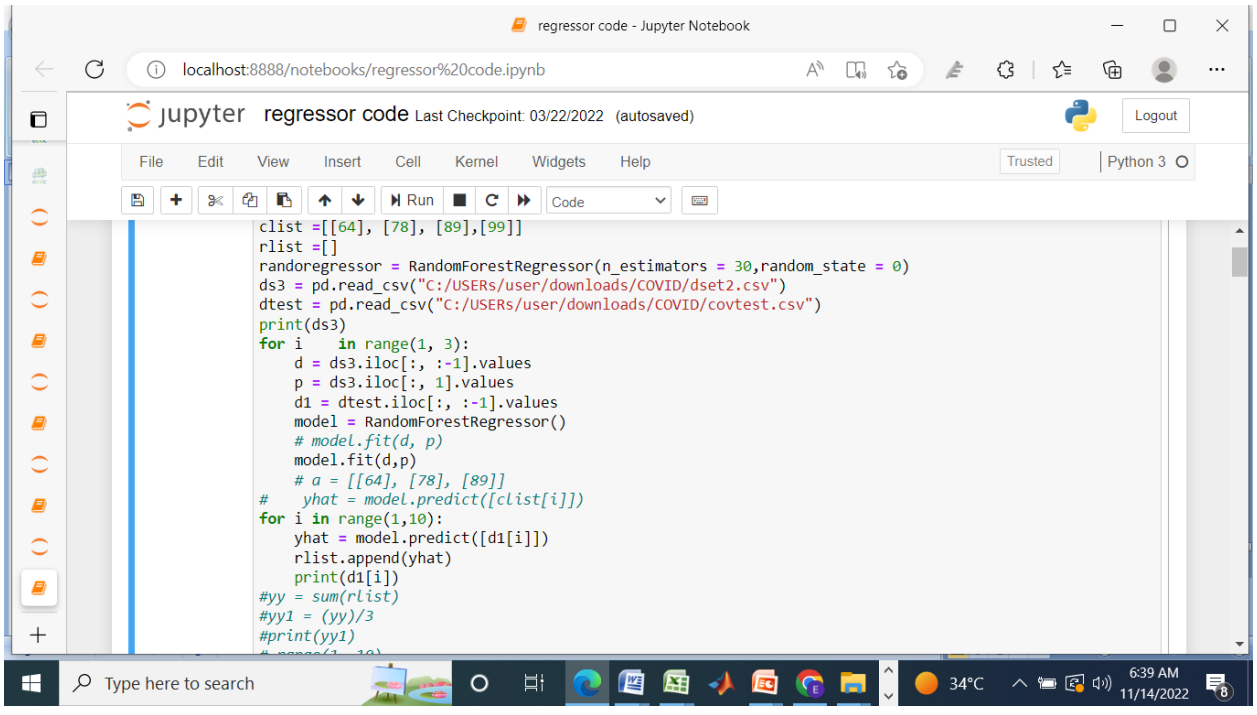

Figure 1: The Python environment

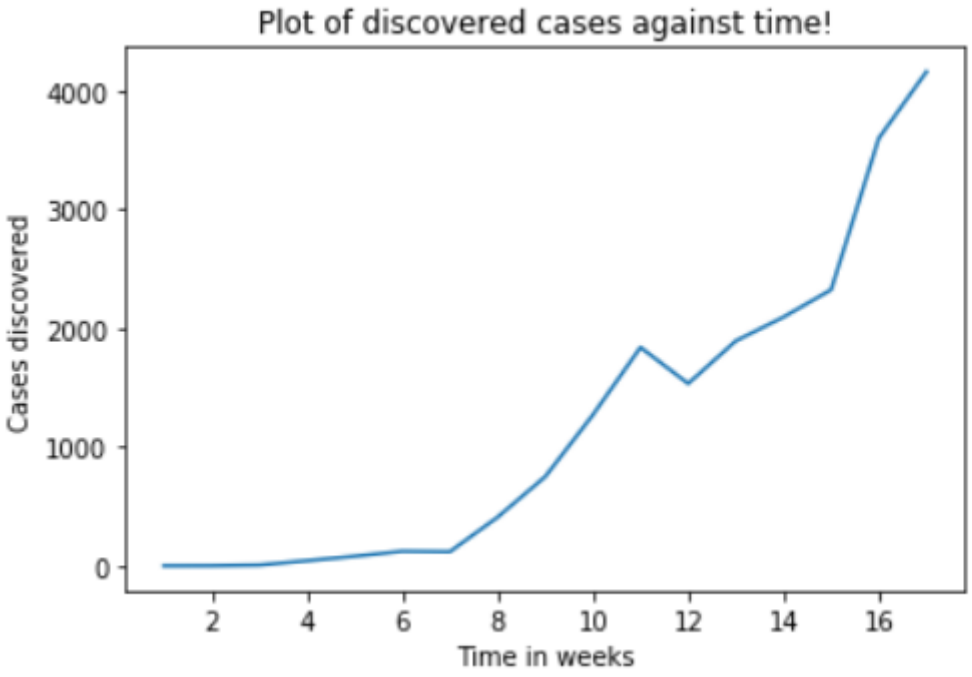

Figure 2: Rate of discovering COVID-19 cases in Nigeria

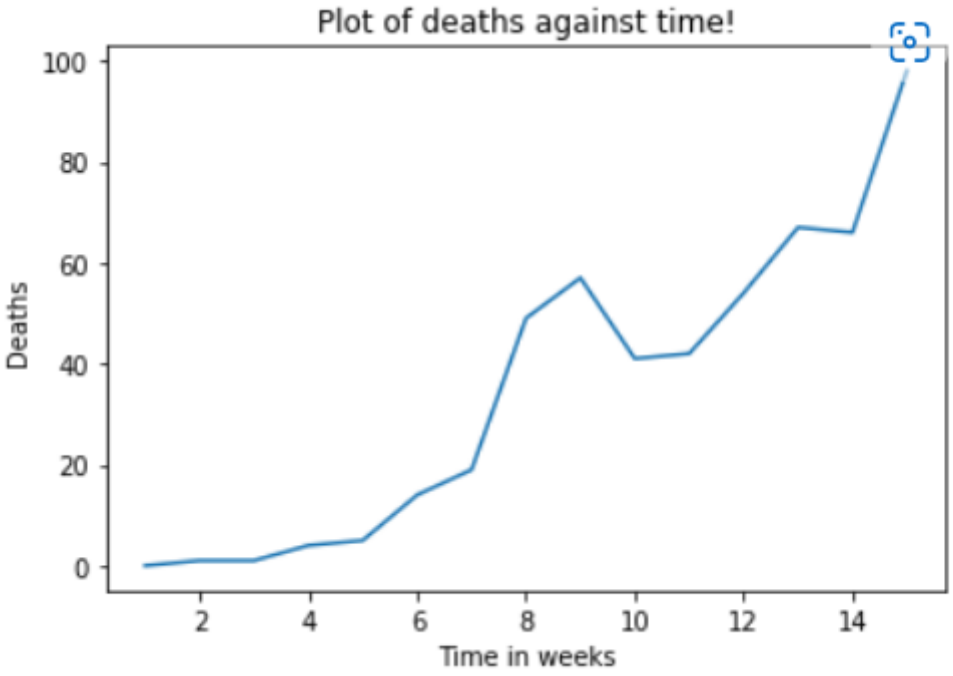

Figure 3: Death Rate of COVID-19 cases in Nigeria

After the data preparation has been done, the random forest model was setup and the dataset was separated to training and testing datasets. The model parameter such as the number of estimators were obtained. In the training period, the number of estimator was made to vary from 20-40. That is the number of decision trees to be considered. It was observed that the number of estimator at 35 gave most accurate training result. The trained model was then used to predict on short term bases in days and medium terms in weeks. The predictions were made on the basis of likely number of cases in days and likely number of deaths that could result from such cases. Fig.3 and Fig.4

represent the results obtained for number of cases in days and number of deaths in days from the model.

The percentage error was calculated using eq(5).

$$\%\ error = \frac{(A_v - O_v)}{A_v} \quad \ldots (4)$$

Where A_v represents the actual observed value and O_v represents the predicted value.

Figure 4: Prediction Rate of COVID-19 cases in Days

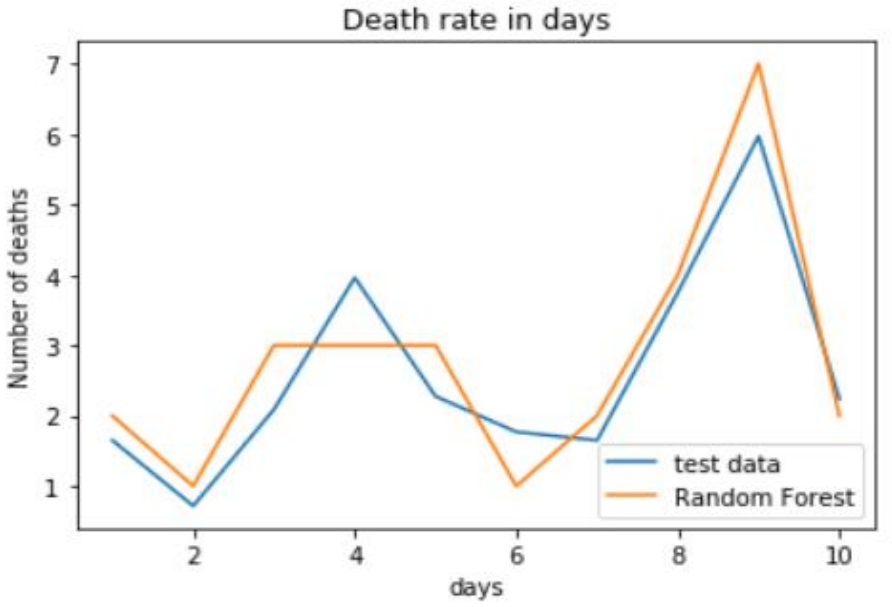

Figure 5: Prediction Rate of COVID-19 deaths in Weeks

The trained model was also used to make prediction for cases in weeks and likely number of deaths in weeks. Fig.5 and Fig.6 represent the results obtained for number of cases in days and number of deaths in days from the model.

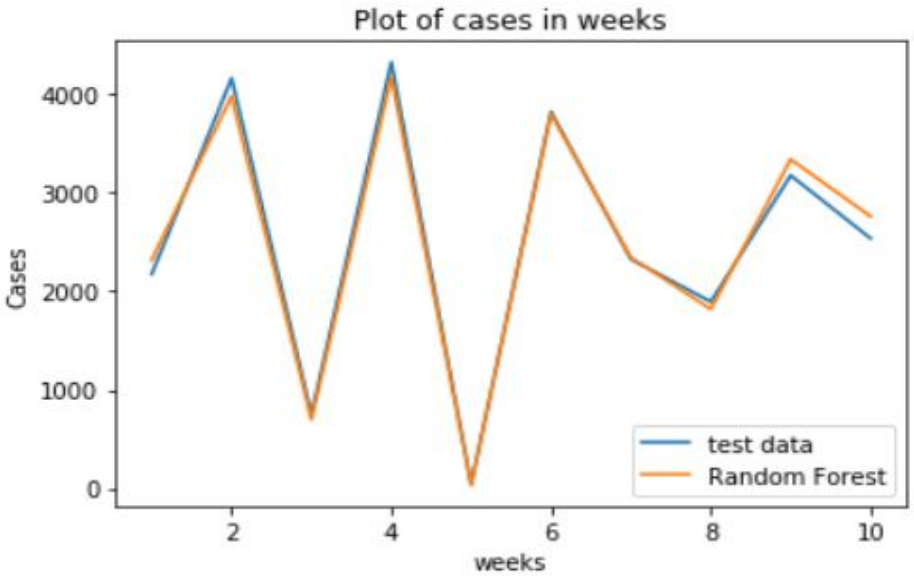

Figure 6: Prediction Rate of COVID-19 deaths in Weeks

Figure 7: Prediction Rate of COVID-19 deaths in Weeks

The performance of the model shows some resemblance in the pattern when compared to known results. The degree of accuracy of the model however varies as the prediction changes from days to weeks. It was observed that model accuracy in the prediction for the number of cases in days was 99.2% while that of the number of cases in weeks was 99.94%.

4. Conclusions

The effort to know ahead the likely number of cases of patients and number of deaths were attempted using the random forest model which has capabilities to engage multiple decision trees so as to predict results. The model is capable of long term forecast. However due to the nature of COVID-19 data available, the short term prediction which has higher accuracy is worth using for planning and other related purposes.

Funding: Not applicable.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Singh, V. **(2020)**. Prediction of COVID-19 corona virus pandemic based on time series data using support vector machine. *Journal of Discrete Mathematical Sciences and Cryptography* 23(8): 1583–1597. **doi:** https://doi.org/10.1080/09720529.2020.1784535.
2. Khan, M.A., Alhaisoni, M., Tariq, U., Hussain, N., Majid, A., Damaševi˘cius, R. and Maskeliunas, R. **(2021)**. COVID-19 Case Recognition from Chest CT Images by Deep Learning, Entropy-Controlled Firefly Optimization, and Parallel Feature Fusion. *Sensors 21*(21): 1-19. **doi:** https://doi.org/10.3390/s21217286.
3. Alassafi, M.O., Jarrah, M. and Alotaibi, R. **(2022)**. Time series prediction of COVID-19 based on deep learning. *Neurocomputing* 468: 335–344. **doi:** https://doi.org/10.1016/j.neucom.2021.10.035.
4. Ahouz, F. and Golabpour, A. **(2021)**. Predicting the incidence of COVID-19 using data mining. *BMC Public Health* 21: 1-12. https://doi.org/10.1186/s12889-021-11058-3.
5. Yuan J., Li M., Gang L. and Lu Z. K. **(2020)**. Monitoring transmissibility and mortality of COVID-19 in Europe. *International Journal of Infectious Diseases* 95: 311–315. **doi:** https://doi.org/10.1016/j.ijid.2020.03.050.
6. Ebinger, J., Wells, M., Ouyang, D., Davis, T. Kaufman, N. Cheng, S. and Chugh S. **(2021)**. A Machine Learning Algorithm Predicts Duration of hospitalization in COVID-19 patients. *Intelligence-Based Medicine* 5: 1-5. **doi:** https://doi.org/10.1016/j.ibmed.2021.100035.
7. Liu, K., Chen, Y., Lin R. and Han, K. **(2020)**. Clinical features of covid-19 in elderly patients: A comparison with young and middle-aged patients, *Journal of Infection*. 80(6): 14-18. **doi:** http://doi.org/10.1016/j.jinf.2020.03.005.
8. Hengl T., Nussbaum M., Wright M. N., Gerard B.M. Heuvelink G. B. M. and Graeler B. **(2018)**. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables *PeerJ* 6(8):e5518. **doi:** https://doi.org/10.7717/peerj.5518.
9. Basu S., Karl Kumbier K., Brown J. B. and Yu B. **(2018)**. Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences of the United States of America* 115(8): 1943–1948. **doi:** https://doi.org/10.1073/pnas.1711236115.
10. Tang B., Brayazzi L. N., Li Q. Tang S. Xiao Y. Wu J. **(2019)**. An updated estimation of the risk of transmission of the novel coronavirus. *Infectious Disease Modelling* 5: 248-255. **doi:** https://doi.org/10.1016/j.idm.2020.02.001.
11. Wang, C., Horby, P. W. , Hayden, F. G., and Gao, G. F. **(2020).** A novel coronavirus outbreak of global health concern. *The Lancet* 395 (10223): 470- 473. **doi:** http://doi.org/10.1016/S0140-6736(20)30185-9.
12. Su, K., Xu, L. Li, G., Ruan, X., Li, X., Deng, P., Li, X., Li. Q., Chen, X. and Xiong, Y., *et al.* **(2019)**. Forecasting inuenza activity using self-adaptive ai model and multi-source data in chongqing, china. *EBioMedicine* 47: 284-292. **doi:** http://doi.org/10.1016/j.ebiom.2019.08.024.
13. https://opendata.ecdc.europa.eu/covid19/casedistribution/csv/data.csv
14. Bala, S. V. **(2012).** Development of Data Clustering Algorithm for predicting Heart. *IJCA* 48(7): 8-13.
15. Pattekari, S. A. and Parveen, A. **(2012).** Prediction system for heart disease using naïve bayes. *International Journal of Advanced Computer and Mathematical Sciences* 3(3): 290-294.

Cite article as:

Oyebode E. O., Awodoye O.O., Ipeayeda F.W. **(2023).** The Prediction Rate of COVID-19 using Random Forest Approach. *Ajayi Crowther J. Pure Appl. Sci.* 2(1): 24-31. https://doi.org/10.56534/acjpas.v2i1.77