Forecasting of Covid-19 Using Time Series Regression Models

AKRAM M. RADWAN
Deanship of Information Technology
University College of Applied Sciences
Gaza, Palestine
aradwan@ucas.edu.ps

2021 Palestinian International Conference on Information and Communication Technology (PICICT) | 978-1-6654-3651-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/PICICT53635.2021.00014

***Abstract*— The novel coronavirus (COVID-19) pandemic is a major global health threat that is spreading very fast around the world. In the current study, we present a new forecasting model to estimate the number of confirmed cases of COVID-19 in the next two weeks based on the previously confirmed cases recorded for 62 countries around the world. The cumulative cases of these countries represents about 96% of the total global up to the date of data gathering. Seven regression models have been used for two rounds of predictions based on the data collected between February 21, 2020 and August 31, 2020. We selected five feature sets using various feature-engineering methods to convert time series problem into a supervised learning problem and then build regression models. The performance of the models was evaluated using Root Mean Squared Log Error (RMSLE). The findings show a good performance and reduce the error about 70%. In particular, XGB and LGBM models have demonstrated their efficiency over other models.**

Keywords*— *COVID-19, Forecasting, Predictive Analytics, Machine Learning, Regression, Time Series.

I. INTRODUCTION

The new coronavirus disease (COVID-19) is a virus infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV2). This epidemic is spreading very quickly all over the world and has affected in more than 200 countries. The World Health Organization (WHO) declared it as a global pandemics. According to WHO, globally 25,484,767 confirmed cases have been recorded and 850,535 deaths have been recorded till the end of August 2020 [1]. The spread of Covid-19 is very dangerous, requires more strict policies, and plans that aid in the healthcare service preparation, which have already been implemented in many countries around the world. Thus, it is very vital to forecast the confirmed cases in the upcoming days to support prevention of outbreak of Covid-19 pandemic and to prepare against possible threats.

In the last year, numerous studies have addressed forecasting the number of confirmed cases of Covid-19. Various mathematical methods, time series models and machine learning techniques have been proposed to estimate the future trend of pandemic COVID-19 [2,3]. A few examples of these methods are Multiple Linear Regression [4], Bayesian Network, Auto-Regressive Integrated Moving Average (ARIMA)[5], Deep learning via Long Short-Term Memory (LSTM)[6], SEIR model [7], Adaptive Neuro-Fuzzy Inference System (ANFIS)[6], and Simulation models [8].

There exist a large number of evidences where regression algorithms have proven to give efficient predictions the COVID-19 prevalence in many countries [2,4]. The prediction based on regression methods have many approaches.

Many researches aimed to predict the prevalence of COVID-19 in one country or union of territories [6,9-11], however, our study handling the estimation of the confirmed cases in the most affected countries worldwide.

The paper is structured as follows. Section II describes features sets that extracted for time series data. Section III presents data and the models used in this study. Next Section IV covers experimental results and performance evaluation. Finally, the conclusions are summarized in Section V.

II. FEATURES ENGINEERING

Feature extraction is the most critical step in designing an algorithm in order to achieve good performance. We have used feature engineering to transform a time series raw data into a supervised learning dataset for machine learning algorithms. It is one of the most effective ways to improve predictive models performance. The process takes in one or more existing columns of raw data and converts it into many columns of new features. Extracting useful information can helping with time series data forecasting [12]. From melting data, we can generate a number of various time series features that can be useful to predict future value based on these features.

A. Lag Features

When we try to predict the confirmed cases for a country, the previous day's cases is significant to make a prediction. In other words, the value at time *t* is affected by the value at time *t-1*. The past time series values are known as lags, so *t-1* is lag1, *t-2* is lag2, and so on. We created lag features for three days.

B. Difference feature

This diff feature computes the difference between the confirmed cases in the previous day and the day before, i.e., Diff1 = $P_{t-1} - P_{t-2}$.

978-1-6654-3651-9/21/$31.00 ©2021 IEEE
DOI 10.1109/PICICT53635.2021.00014

C. Rolling Window Features

The rolling window feature for time series calculates some statistical and aggregate functions based on past values. The size of the rolling window *m* is defined as the time frame, which in our case is the number of days. We created rolling window mean of size 3 at day t-1, denoted by M_{t-1}.

D. Z-Scores scaling

Z-score is linearly transformed data value having a mean of zero and a standard deviation of 1. Z-score, or standard score, is used for standardizing scores on the same scale by dividing a score's deviation by the standard deviation in a dataset and is given by the formula: $Z_x = \frac{X_i - mean(X)}{Std(X)}$. A z-score can be zero, positive or negative. A negative score indicates a value less than the mean, and a positive score indicates a value greater than the mean. The standardization of the time series for each country reduces the differences between the confirmed cases. In our model, we computed the z-score value at day t-1, denoted by Z_{t-1}.

E. Rate of Change feature

Rate of change is the percentage change between the current and a prior values, and is given by formula (1):

$$Rch_{t-1} = \frac{P_{t-1}}{P_{t-2}} - 1 \quad (1)$$

where P_{t-1} and P_{t-2} are the number of confirmed cases in day t-1 and the day before t-2. When apply this feature, all rows in a dataset with zero cases were deleted and the number of samples was reduced.

F. Rank

Rank feature gets the data frame by ascending order with maximum rank value, and equal values have the same rank. In our model, we computed the rank value at day t, denoted by $Rank_t$.

G. Cumulative maximum

Cmax feature finds the cumulative maximum value along confirmed cases from day 1 to day t-1. $Cmax_{t-1} = \max\{P_1, P_2, \dots, P_{t-1}\}$.

H. Entropy

The concept of entropy in information theory measures the amount of uncertainty of a random variable X. The entropy in terms of X, with possible outcomes $x_1, x_2, \dots, x_n$ is defined as:

$$H(X) = -\sum_{i=1}^{n} p(x_i) log_2\, p(x_i) \quad (2)$$

where $p(x_i)$ is simply the frequentist probability of an element/class 'i' in our data. When apply the entropy feature, all rows in a dataset with zero cases were deleted and number of samples was reduced.

For this study, five feature sets are designed, and described in Table I.

TABLE I. DESCRIPTIONS OF FEATURE SETS

Feature Set	Lag-1	Diff-1	Lag-3	Diff-3	M_3	Rch_{t-1}	Cmax	$Rank_t$	Z_{t-1}	Entropy
FS1	√	√								
FS2			√	√						
FS3	√	√			√	√				
FS4	√	√					√	√		
FS5	√	√							√	√

III. EXPERIMENTAL DESIGN AND METHODS

This section presents the description of the used data, regression models with their parameter settings and the performance measures.

A. Dataset

The data includes time series data tracking the number of people affected by Coronavirus (COVID-19) worldwide. The employed dataset contains data on COVID-19 including new daily-confirmed cases and it covers the period 21th February 2020 to 31th August 2020. Data are categorized by country named conforming to the WHO. It covers 62 countries around the world that are the most affected countries worldwide. The cumulative Covid-19 cases of these countries represents 96% of the total global up to the date of data collection. The dataset used in this work was collected from the repository of John Hopkins University Center for Systems Science and Engineering (CSSE) [13]. We evaluated the performance of the presented method using two datasets of daily Covid-19 confirmed cases. The first one is called DS1; it starts from February 21 and continues until June 25, 2020. Whereas, the second one is called DS2; it starts from June 26 to August 31, 2020. The raw training data consists of samples; each records daily-confirmed cases for 126 days in DS1 and 67 days in DS2 for each of the country as shown in Table II and Table III.

When we try to fit a regression model for each country, we face a problem due to having only 126 in DS1 and 67 in DS2 data entries for each model, which is small and not enough to get good results. To encounter this problem, we have used melting data, which converts wide-format data with several measurement columns into long-format with much more rows. In this case, each row becomes: Country, Day, Confirmed cases and we have 7874 rows (in DS1) and 3960 rows (in DS2) to train the models.

TABLE II. SAMPLE OF THE DATASET DS1.

Country	Feb 21	Feb 22	Feb 23	…	June 23	June 24	June 25
Afghanistan	0	0	0	…	324	159	535
Algeria	0	0	0	…	156	172	197
Argentina	0	0	0	…	2272	2648	2606
⋮	⋮	⋮	⋮	…	⋮	⋮	⋮
Italy	17	42	93	…	113	577	296
Japan	11	17	25	…	59	84	92
Kazakhstan	0	0	0	…	534	520	465
South Korea	100	229	169	…	51	28	39
⋮	⋮	⋮	⋮	…	⋮	⋮	⋮
UK	0	0	0	…	921	655	1118
Ukraine	0	0	0	…	845	951	1002
US	2	0	0	…	35189	34836	39972

TABLE III. SAMPLE OF THE DATASET DS2.

Country	June 26	June 27	June 28	…	Aug 29	Aug 30	Aug 31
Afghanistan	276	165	351	…	3	19	3
Algeria	240	283	305	…	378	365	348
Argentina	2886	2401	2189	…	9230	7187	9309
⋮	⋮	⋮	⋮	…	⋮	⋮	⋮
Italy	255	175	174	…	1444	1365	996
Japan	107	92	112	…	854	605	438
Kazakhstan	569	0	1008	…	126	111	77
South Korea	51	62	42	…	299	248	235
⋮	⋮	⋮	⋮	…	⋮	⋮	⋮
UK	1381	634	407	…	1110	1752	1415
Ukraine	1121	957	924	…	2579	2179	2202
US	45255	42705	39605	…	47153	35337	34156

B. *Regression Models*

Regression models are statistical sets of processes that are used to estimate or predict the target or dependent variable based on one or more independent variables. In this section, a brief of popular prediction algorithms are described, which are employed in the data analysis and experimental results.

1) *Decision Tree (DT)*

DT solves the regression problem by transforming the data into tree representation. Each internal node of the tree denotes an attribute or feature and each leaf node denotes a class label. While DT requires less effort for data preparation during pre-processing, it often involves higher time to train the model.

2) *Random Forest (RF)*

RF is a bagging ensemble models that combines the prediction of multiple decision trees to create a more accurate final prediction. The final prediction is computed by taking the mean of the individual decision-tree predictions. RF is a fast and robust learning method able to deal with the randomness of the time series [14].

3) *Gradient Boosting Regression (GBR)*

GBR is a type of ensemble where additional trees are added at each stage to compensate the shortcoming of the existing weak learners. These models are generally employed where features are too heterogeneous. Gradient Boosting model is more robust to outliers than boosting algorithm [15]. In our model of Gradient Boosting Regressor we have used Huber loss function in loss function parameter.

4) *Extreme Gradient Boosting (XGB)*

XGB is a tree-based model. It stacks many trees, each new tree attempting to reduce the error of the preceding ensemble. The main goal is to develop a strong predictor by combining many weak predictors. XGB is one of the most powerful regression algorithms with high speed and performance [16]. It runs more than ten times faster than existing popular solutions on a single machine. XGB is an efficient and scalable implementation of GBR. Moreover, it is feasible to train on large datasets. XGB can also be used for time series prediction.

5) *Light Gradient Boosting Machine (LGBM)*

LGBM is a gradient boosting framework based on decision tree algorithm. LGBM has faster training speed with lower memory usage compare to XGB [17]. Moreover, it can handle the large size of data and support GPU learning. Even though both XGB and LGBM models follow Gradient Boosting, XGB grows tree level-wise and LightGBM grows tree leaf-wise.

6) *Support Vector Regression (SVR)*

This model works similarly to SVM (Support Vector Machine), but is adapted to handle regression. SVR uses kernel function to calculate the similarity between two data points when dealing with non-linear problem. SVR involves two parameters that should be tuned for the model to perform well; the regularization parameter (referred to C) and the error sensitivity parameter (referred to ε) [18].

7) *Stacking-Ensemble learning (SEL)*

Stacking Generalization is an ensemble learning technique to combine multiple regression models (base-learners) via a meta-regressor. The individual regression models are trained based on the complete training set; then, the meta-regressor is fitted based on the outputs of the individual regression models in the ensemble [19]. The main advantage of the SEL is that this technique can improve the accuracy and reduce error variance. For this study, we trained a stacking-ensemble model using DT and RF as base-learners and as the LGBM meta-regressor.

This study aims to assess the ability of the regression models with different feature sets to forecast the

confirmed COVID-19 cases by comparing their performances. Fine tuning predictive model hyper-parameters is a crucial step to find the best fit parameters that improve accuracy of the forecasted results. The choice of inappropriate parameters' values may result in a poor performance. The parameters' setting for the models used in our study is listed in Table IV. For stacking method, no need to tune the parameters since this method is a combinations of the best regressions.

For each feature set, seven regression models were explored and their relative performances were compared.

TABLE IV. PARAMETERS' SETTING

Algorithm	Parameters Setting
DT	max_depth=5
RF	n_estimators=1000, n_jobs=-1, random_state=0
GBR	n_estimators =300, max_depth= 4, min_samples_split= 2, learning_rate= 0.01
XGB	learning_rate=0.1, base_score=0.5, max_depth=3, min_child_weight=2, n_estimators=200
LGBM	num_leaves=10, learning_rate=0.1, n_estimators=100, reg_lambda=0.30
SVR	C=3.0, ε =0.2

C. *Evaluation Criteria*

To check the performance of the seven models used in this study, we used Root Mean Squared Logarithmic Error (RMSLE), which is computed as follows:

$$\text{RMSLE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(\hat{y}_i + 1) - (\log(y_i + 1))^2} \quad (3)$$

where N is the number of data observations, $\hat{y}_i$ is the predicted count, and y_i is the actual count. RMSLE is less sensitive to outliers than other metrics [20]. RMSLE is preferable when there is a wide range in the target variables and targets having exponential growth, such as population counts. Also it is appropriate when we care about percentage errors rather than the absolute value of errors.

All experiments are implemented using python and its libraries such as scikit-learn, numpy and Pandas [21].

IV. RESULTS AND DISCUSSION

In this section, we described the performed experiments and discussed the obtained results. Our prediction results were obtained using seven regression models. Experiments were conducted over two rounds of forecasts where the first round was made for the two weeks from 12/06/2020 to 25/06/2020 based the data points available from 21/02/2020 to 11/06/2020 (training data from dataset DS1). Once more data became available, the second round of forecast was done for another two weeks from 18/08/2020 up end of Aug 2020 based on the actual data from 26/06/2020 to 17/08/2020 (training data from dataset DS2).

First, we set a baseline prediction score, it is 0.755. Now, for a predictive model has a RMSLE below than the baseline, it is good.

Table V shows the RMSLE scores obtained by all regression models on DS1 using selected feature sets explained earlier in Section II, and the best performance is shaded with grey. Overall, all models demonstrated good performance. From Table V, the XGB achieved the best performance using all feature sets, except for FS4, the Stacking-Ensemble outperformed.

TABLE V. COMPARISON OF THE RMSLE SCORES OBTAINED ON DS1 WITH FIVE FEATURE SETS.

Algorithm	FS1	FS2	FS3	FS4	FS5
DT	0.656	0.564	0.539	0.431	0.373
RF	0.642	0.588	0.601	0.266	0.240
GBR	0.556	0.532	0.515	0.459	0.437
XGB	0.517	0.513	0.494	0.390	0.215
LGBM	0.520	0.514	0.509	0.276	0.232
SVR	0.796	0.649	0.651	0.759	0.583
Stacking-Ensemble	0.706	0.637	0.666	0.265	0.240

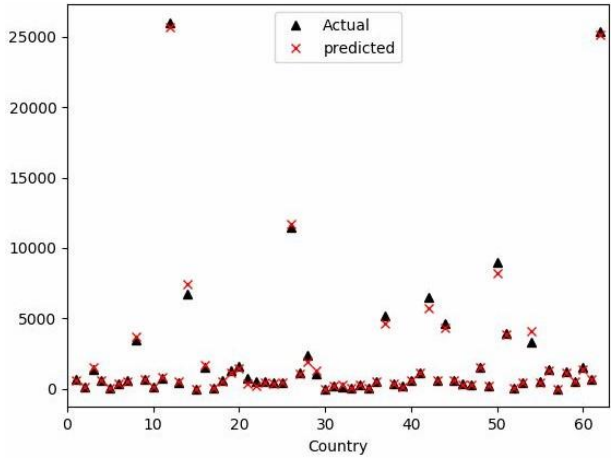
Table VI shows the RMSLE scores obtained by all regression models on DS2 with five feature sets. One can easily note that the LGBM achieved the best performance using the first three feature sets while XGB outperformed for FS4 and FS5.

TABLE VI. COMPARISON OF THE RMSLE SCORES OBTAINED ON DS2 WITH FIVE FEATURE SETS.

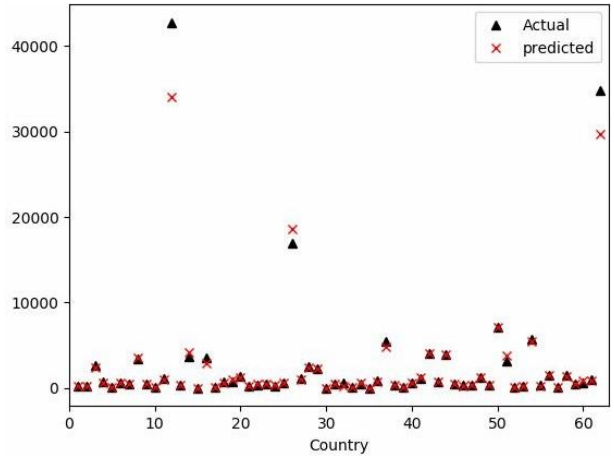
Algorithm	FS1	FS2	FS3	FS4	FS5
DT	0.673	0.785	0.664	0.816	0.464
RF	0.729	0.680	0.686	0.528	0.256
GBR	0.644	0.647	0.629	0.635	0.433
XGB	0.653	0.655	0.720	0.526	0.239
LGBM	0.642	0.622	0.624	0.555	0.275
SVR	0.766	0.772	0.765	0.824	0.658
Stacking-Ensemble	0.743	0.731	0.846	0.550	0.274

In addition, it is easy to observe that using FS5, all models achieved the best prediction results in two round experiments. FS5 includes z-score and entropy features besides Lag1 and Diff1 features. Note that adding z-score and entropy to FS1 (includes only lag1 and diff1) can produce 50% improvements or more in prediction results.

Since the best scores was achieved by XGB, it used to predict the COVID-19 daily-confirmed cases for the first day and last day of the 2-weeks prediction. Figures 1 graphically compares the actual value with predicted value of confirmed cases using XGB model with FS5 for June 12, 2020 and June 25, 2020 respectively, where the x-axis represents the country, and the y-axis represents the corresponding daily-confirmed cases: actual (black) versus predicted (red). Figure shown that the dots representing the actual and predicted observation are very close. These findings confirmed the results that we have shown earlier in Table V.



(a)



(b)

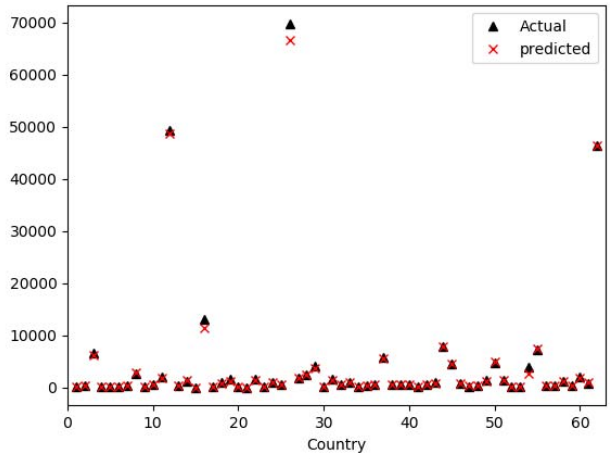
Figure 1. Actual and predicted confirmed cases recorded for 62 countries. (a) For June 12, 2020, and (b) For June 25, 2020.

Similarly, for the second round experiment, the comparison between the actual and the predicted values of confirmed cases using XGB model with FS5 for August 18, 2020 and August 31, 2020 is shown in figure 2. In the figure, the dots representing the actual and predicted values are close. However, we can observe that the XGB model gives some prediction values that are slightly far from the actual value in last day figure of two rounds, see 1(b) and figure 2(b).

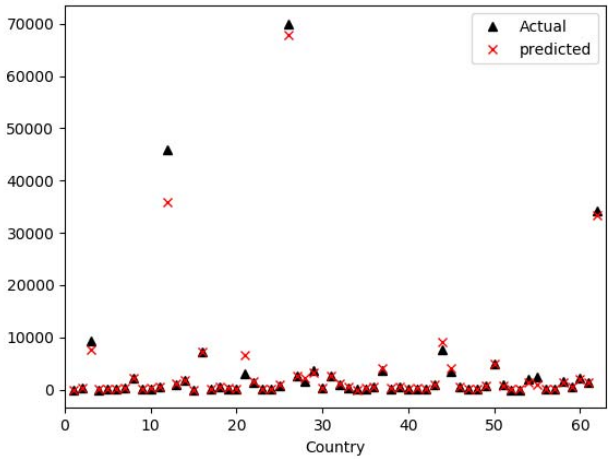
V. CONCLUSION

In this paper, we have conducted a two-round study of COVID-19 confirmed cases in the most affected countries worldwide. The most prominent techniques of regression models were used to analyze and predict the daily cases. The regression models are trained on five feature sets, which are extracted through the original feature set. We have analyzed epidemic data made available by the CSSE within about seven months to forecast the number of confirmed cases of COVID-19 for the next two weeks based on data available within enough time period before. The performance of the models was assessed using RMSLE and achieved 0.215 and 0.239 for XGB model on DS1 and DS2, respectively. Remember that our baseline was 0.755, XGB improved results to be 0.215 and 0.239, which means a 68-71% error reduction. The results also show that XGB and LGBM models are appropriate for predicting the prevalence of COVID-19 in the future.

There are some limitations in the forecasted numbers of COVID-19 cases. First, Some countries have missing values for some days, so datasets record 0 values for these days and add missing cases for this day to next day. Therefore, results for these countries are not accurate. As we know training models with datasets of poor quality will guide to misleading results. Second, a prediction model of Covid-19 relies on past behavior. So the existence of outliers and noise in the data makes it hard to accurately predict the number of cases. Therefore, it is necessary to use noise filters to reduce noise's effects. But this is missing in our study.



(a)



(b)

Figure 2. Actual and predicted confirmed cases recorded for 62 countries. (a) For August 18, 2020, and (b) For August 31, 2020.

ACKNOWLEDGMENTS

The author would like to thank the John Hopkins University for making the updated data on the number of infected cases of COVID-19 available to the public.

REFERENCES

[1] WHO Organization. Novel Coronavirus (2019-nCoV) 2020. Available online: https://www.who.int/ (Accessed on 1 Sep. 2020).

[2] Ahmad, A., Garhwal, S., Ray, S.K. et al. (2020) "The Number of Confirmed Cases of Covid-19 by using Machine Learning: Methods and Challenges". *Archives of Computational Methods in Engineering*, 28:2645–2653 https://doi.org/10.1007/s11831-020-09472-8.

[3] Vytla, V., Ramakuri, S. K., Peddi, A., . et al. (2021) "Mathematical models for predicting COVID-19 pandemic: a review". *Journal of Physics*: Conference Series, 1797 (1).

[4] Rath S., Tripathy A., Tripathy A.R. (2020) "Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model". *Diabetes Metab Syndr Clin Res Rev.*14(5):1467–1474. doi: 10.1016/j.dsx.2020.07.045.

[5] Hernandez-Matamoros, A., Fujita, H., Hayashi, T., & Perez-Meana, H. (2020). "Forecasting of COVID19 per regions using ARIMA models and polynomial functions". *Applied soft computing*, 96, 106610. https://doi.org/10.1016/j.asoc.2020.106610.

[6] Chowdhury, A.A., Hasan, K.T. & Hoque, K.K.S. (2021) "Analysis and Prediction of COVID-19 Pandemic in Bangladesh by Using ANFIS and LSTM Network". *Cognitive Computation*, 13, 761–770. https://doi.org/10.1007/s12559-021-09859-0

[7] Feng S., Feng Z., Ling C., Chang C., Feng Z. (2021) "Prediction of the COVID-19 epidemic trends based on SEIR and AI models". *PLoSONE*, 16(1). https://doi.org/10.1371/journal.pone.0245101.

[8] Hassanat, A.B.; Mnasri, S.; Aseeri, M.A., ., et al. (2021) "A Simulation Model for Forecasting COVID-19 Pandemic Spread: Analytical Results Based on the Current Saudi COVID-19 Data". *Sustainability*, 13, 4888. https://doi.org/10.3390/su13094888.

[9] Al-qaness, M.A.A.; Ewees, A.A.; Fan, H.; Abd El Aziz, M. (2020) "Optimization Method for Forecasting Confirmed Cases of COVID-19 in China". *Clinical. Medicine*, 9, 674; doi:10.3390/jcm9030674.

[10] Samson, T. K., Ogunlaran, O. M., & Raimi, O. M. (2020). "A Predictive Model for Confirmed Cases of COVID-19 in Nigeria". *European Journal of Applied Sciences*, 8 (4):1–10.

[11] Ribeiro, M.H.D.M., Da Silva, R.G., Mariani, V.C. and Coelho, L.S. (2020) "Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil". *Chaos, Solitons and Fractals*, 135, 1-10.

[12] Basic Feature Engineering With Time Series Data in Python, Available online: https://machinelearningmastery.com/basic-feature-engineering-time-series-data-python/ (Accessed on 25 October 2020).

[13] Novel coronavirus (Covid-19) cases, provided by Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, USA. Available online: https ://github.com/CSSEG ISand Data/COVID-19. (Accessed on 1 October 2020).

[14] Breiman L.(2001) "Random forests". Machine Learning, 45 (1): 5-32.

[15] Gumaei, A., Al-Rakhami, M., Al Rahhal, M. M., et al. (2021) "Prediction of COVID-19 confirmed cases using gradient boosting regression method". *Computers, Materials and Continua*, 66(1).

[16] Chen, T. and Guestrin, C. (2016) "Xgboost: A scalable tree boosting system". In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 785–794, New York, NY, USA. ACM

[17] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. (2017) "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". *Advances in Neural Information Processing Systems*, 30, 3146–54.

[18] Drucker H , Burges CJC , Kaufman L , Smola AJ , Vapnik V . (1997) "Support Vector Regression Machines". In *Mozer MC, Jordan MI*, Petsche T, editors. Advances in neural information processing systems 9. MIT Press, 155–61 .

[19] Ribeiro MHDM, Coelho LdS. (2020) "Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series". *Applied soft computing*, 86(105837). doi: 10.1016/j.asoc.2019.105837.

[20] What's the Difference Between RMSE and RMSLE?. Available online: https://medium.com/analytics-vidhya/root-mean-square-log-error-rmse-vs-rmlse-935c6cc1802a. (Accessed on 1 Nov. 2021)

[21] Bloice M.D., Holzinger A. (2016) "A Tutorial on Machine Learning and Data Science Tools with Python". In: *Holzinger A. (eds) Machine Learning for Health Informatics*. Lecture Notes in Computer Science, vol 9605. Springer, Cham. https://doi.org/10.1007/978-3-319-50478-0_22.