Forecasting COVID-19 Confirmed Cases Using Time Series Analysis

التنبؤ بعدد الحالات المؤكدة لكوفيد - 19 باستخدام تحليل السلاسل الزمنية

Akram Mohammed Radwan
Assistant Professor\ University College of Applied Sciences\ Palestine
aradwan@ucas.edu.ps

أكرم محمد رضوان
أستاذ مساعد/ الكلية الجامعية للعلوم التطبيقية / فلسطين

Received: 10/ 1/ 2022, **Accepted:** 14/ 11/ 2022
DOI: 10.33977/2106-000-006-002
http://journals.qou.edu/index.php/PJTAS

تاريخ الاستلام: 10/ 1/ 2022م، **تاريخ القبول:** 14/ 11/ 2022م.
E - ISSN: 2521 - 411X
P - ISSN: 2520 - 7431

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 represent about 95% of the total global up to the date of data gathering. Seven regression models have been used for three rounds of predictions based on the data collected between February 21, 2020 and December 29, 2020. A number of different time series features have generated using feature-engineering methods to convert a time series forecast into a supervised learning problem and then build regression models. The performance of the models was evaluated using root mean squared log error, root mean squared error, mean absolute error, mean absolute percentage error, coefficient of determination and running time. The findings show a good performance and can reduce the error about 72% with a high coefficient of R2 = 0.990. In particular, XGB and Random Forest models have demonstrated their efficiency over other models.

Keywords: COVID-19, predictive analytics, machine learning, regression, time series.

الملخص:

تعد جائحة كورونا (COVID-19) تهديدًا صحياً عالمياً رئيسياً انتشر بسرعة كبيرة في جميع أنحاء العالم. في الدراسة الحالية، قدمنا نموذجاً للتنبؤ لتقدير عدد الحالات المؤكدة لكوفيد- 19 في الأسبوعين المقبلين بناءً على أعداد الحالات المؤكدة مسبقاً التي سجلت في 62 دولة حول العالم. تمثل الحالات التراكمية لتلك الدول حوالي 95 % من الإجمالي العالمي حتى تاريخ جمع البيانات. تم استخدام سبع خوارزميات انحدار لثلاث جولات من التنبؤات بناءً على البيانات التي تم جمعها في الفترة ما بين 21 فبراير 2020 و29 ديسمبر 2020. تم استخراج عدد من ميزات السلاسل الزمنية باستخدام أساليب هندسة الميزات لتحويل التنبؤ بالسلاسل الزمنية إلى مسألة تعلم آلي خاضع للإشراف، ثم باستخدام الميزات الأكثر أهمية تم بناء نماذج الانحدار لعمل التنبؤ المطلوب. تم تقييم أداء النماذج باستخدام المقاييس التالية: جذر متوسط الخطأ اللوغاريتمي التربيعي، جذر متوسط الخطأ التربيعي، متوسط الخطأ المطلق، متوسط نسبة الخطأ المطلق، معامل التحديد (مربع معامل الارتباط) ووقت التنفيذ. أظهرت نتائج هذا البحث أن أداء نماذج الانحدار كانت جيدة، واستطاعت تقليل الخطأ بنسبة 72 % مع معامل تحديد عالٍ R2 وصل لـ 990. على وجه الخصوص، أظهرت كلّ من خوارزميات XGB و Random Forest كفاءة أعلى في الأداء مقارنة مع الخوارزميات الأخرى.

الكلمات المفتاحية: كوفيد- 19، التحليل التنبؤي، التعلم الآلي، الانحدار، السلاسل الزمنية.

1. Introduction

The new coronavirus disease (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV2). This epidemic is spreading very quickly all over the world, affecting more than 200 countries. The World Health Organization (WHO) declared it as a global pandemic. According to WHO, 82,679,350 confirmed cases have been recorded and 1,872,994 deaths have been reported till the end of December 2020[(1)] globally. The spread of Covid-19 is very dangerous

requiring more strict policies, and plans that aid 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 the prevention of outbreak of Covid-19 pandemic and to prepare against possible threats (Oliveira and Moral, 2021).

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 (ML) techniques have been proposed to estimate the future trend of pandemic Covid-19 (Ahmad et al., 2020) and (Vytla et al., 2021). A few examples of these methods are Multiple Linear Regression (Rath et al., 2020), Bayesian Network, Auto-Regressive Integrated Moving Average (ARIMA)

(Hernandez-Matamoros et al., 2020), Deep learning via Long Short-Term Memory (LSTM) (Chowdhury et al., 2021), SEIR model (Feng et al., 2021), Adaptive Neuro-Fuzzy Inference System (ANFIS) (Chowdhury et al., 2021), and Simulation models (Hassanat et al., 2021).

Time series forecasting is a method to predict future values based on previously observed values using temporal features. This method has been studied widely in Covid-19 forecasting. Using historical time series data, we can forecast the number of daily new confirmed cases in the next days (Oliveira & Moral, 2021; Ahmad et al., 2020).

There exist a large number of evidences where regression algorithms have proven to give efficient predictions the Covid-19 prevalence in many countries (Ahmad et al., 2020; Rath et al., 2020). The prediction based on regression methods has many approaches.

Many researches aimed to predict the prevalence of Covid-19 in one country or union of territories (Chowdhury et al., 2021; Al-Qaness et al. (2020; Samson et al., 2020; Ribeiro et al., 2020), however, our study handles the estimation of the confirmed cases in the most affected countries worldwide.

The paper is structured as follows. Section 2 provides a literature review. Section 3 describes feature sets, which generated from time series data. In section 4, we present data and the models used in this study. Next Section 5 covers experimental results and performance evaluation. Finally, the conclusions are summarized in Section 5.

2. Literature Review

During the last two years, several research papers were published that tackled the applications of machine learning techniques for the prediction of Covid-19. In this section, we briefly review state-of-the-art that are relevant to our work.

Pandey et al. (2020) used linear and polynomial regression to predict the number of confirmed cases in India. They use data from January 30, 2022 to March 25. 2020 as the training data and predict the number of COVID-19 cases for next two weeks. The performance of the model was evaluated using RMSLE and achieved 1.75. Gu et al. (2020) applied cubic regression equations which use the number of days as the input variable to predict the confirmed cases in the whole of China except Hubei based on the existing data. Multiple Regression Analysis Models are suitable to fit the model and to predict the COVID-19 epidemic. Another research on the forecasting COVID-19 has been conducted by Rath et al. (2020). The authors used multiple linear regression to predict the next number daily active cases during the second week of August.

Gecili, Ziady and Szczesniak (2021) presented four different time series models for forecasting the numbers of confirmed cases, deaths and recoveries of COVID-19 for both the USA and Italy based on the daily reported data covered the period from February 22, 2020 until April 29, 2020. The performance of these models was evaluated using mean absolute error (MAE) and mean absolute percentage error (MAPE). The ARIMA model, is useful and powerful in time series analysis, was the most consistent across the other models and it had smaller prediction errors and narrower prediction intervals.

The study of Wang et al. (2022) proposed the ARIMA, SARIMA and Prophet models to predict daily new cases and cumulative confirmed cases in the USA, Brazil and India over the next 30 days based on the time series data from May 1, 2020 to November 30, 2021. The performance of different models was evaluated by using the root mean square error (RMSE), MAE and MAPE. The experimental results showed that the Prophet's model is more suitable for daily new cases of the USA with large fluctuations and has its unique advantages compared with ARIMA model, which is better for predicting of new cumulative cases in Brazil and India.

The authors in (Ribeiro et al., 2020) implemented six ML models to forecast with 1, 3 and 6 days ahead the COVID-19 cumulative confirmed cases of the most affected states of Brazil. They are ARIMA, ridge regression (RR), cubist (CUBIST), random forest (RF), SVR and stacking-ensemble learning (SEL) model. In the SEL approach, the CUBIST regression, RF, RIDGE, and SVR models are adopted as a base-

learners and Gaussian process (GP) as a meta-regressor. Their developed models can generate accurate prediction with a reasonable error.

Al-Qaness et al. (2020) combined ANFIS with an enhanced Flower Pollination Algorithm (FPA) and Salp Swarm Algorithm (SSA) to optimize the parameters of the model. The enhanced model is used to estimate the number of confirmed cases of COVID-19 in the upcoming ten days based on the previously confirmed cases recorded in China. Their approach showed better performance in terms of MAPE, Root Mean Squared Relative Error (RMSRE), coefficient of determination (R2), and computing time.

The COVID19 data is time-series data that compiles the number of confirmed cases where the cases are increasing over time until it arrives at a certain peak curve. Deep learning techniques such as LSTM can handle the nonlinearity and complexity of COVID-19 time-series data. The researchers in (Chowdhury et al, 2021) have used ANFIS and LSTM to predict the newly infected cases in Bangladesh. They showed that LSTM works better on a scenario based model for Bangladesh with MAPE= 4.51, RMSE= 6.55 and Correlation Coefficient= 0.75. Generally, LSTM is preferable in predicting the long term where ARIMA is for the short term (Kibria et al., 2022).

3. Features Engineering

Feature extraction is the most critical step in designing an algorithm in order to achieve good performance (Abuzir et al., 2021). 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 help with time series data forecasting(2). 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 are significant to make a prediction. In other words, the value at day t is affected by the value at day *t-1*. The past time series values are known as lags, so for *t-1* is lag_{t-1} for *t-2* is , 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.,$D_{t-1}=X_{t-1}-X_{t-2}$. We created diff feature features for three days.

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}, would be mean (X_{t-1},X_{t-2} ,X_{t-3}).

D. Z-Scores scaling

The Z-score is linearly transformed data value having a mean of zero and a standard deviation of one. 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 by the formula:

$$Z_t = \frac{X_t - mean(X)}{Std(X)} \quad (1)$$

where be a set of confirmed cases for a country and X_t is confirmed cases at day t. 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. Rank

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

F. Cumulative maximum

Cmax feature finds the cumulative maximum

valu $Cmax_{t-1} = \max\{X_1, X_2, \dots, X_{t-1}\}$ day t-1.

G. 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 po:

$$H(X) = -\sum_{i=1}^{n} p(X_i) log_2\ p(X_i) \qquad (2)$$

where $p(X_i)$ is simply the frequentist probability of a confirmed cases for a country at day *i*. When applies the entropy feature, all rows in a dataset with zero cases were deleted and the number of samples was reduced.

For this study, feature set is designed to include the following extracted features: lag_{t-1}, lag_{t-2}, lag_{t-3}, D_{t-1}, D_{t-2}, D_{t-3}, M_{t-1}, Z_{t-1}, $Rank_t$, $Cmax_{t-1}$ and $H(X_{t-1})$.

4. Experimental Design and Methods

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

4.1. Dataset

The data includes time series data tracking the number of people affected by Coronavirus worldwide. The employed dataset contains data on Covid-19 including new daily-confirmed cases and it covers the period 21th February 2020 to 29th December 2020. Data is 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 represent 95% of the total global up to the date of data collection. The data used in this work were collected from the repository of the John Hopkins University Center for Systems Science and Engineering (CSSE)[3]. We evaluated the performance of the presented method using three datasets of daily Covid-19 confirmed cases. The first one is called DS1; it starts from February 21 and continues until June 25, 2020. The second one is called DS2; it starts from June 26 to August 31, 2020 whereas, the third is called DS3; it starts from the first of September to December 29, 2020. The raw data consists of samples; each records daily-confirmed cases for 126 days in DS1, 67 days in DS2 and 120 days in DS3 for each country. Table 1 depicts a sample of dataset DS2.

When we try to fit a regression model for each country, we faced a problem due to having a little data entries (number of days in a dataset), 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 samples (in DS1), 3960 samples (in DS2) and 7259 samples (in DS3) to train and test the models.

4.2. 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. It is widely used when both dependent and independent variables are linearly or non-linearly related, and the target variable has a set of continuous values. 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.

Table 1.

Sample of the dataset DS2.

Country	June 26	June 27	..	Aug 30	Aug 31
Afghanistan	276	165	..	19	3
Algeria	240	283	..	365	348
Argentina	2886	2401	..	7187	9309
⋮	⋮	⋮	..	⋮	⋮
Italy	255	175	..	1365	996
Japan	107	92	..	605	438
Kazakhstan	569	0	..	111	77
South Korea	51	62	..	248	235
⋮	⋮	⋮	..	⋮	⋮
UK	1381	634	..	1752	1415
Ukraine	1121	957	..	2179	2202
US	45255	42705	..	35337	34156

2) Random Forest (RF)

RF is a bagging ensemble model 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 (Breiman, 2001).

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 (Gumaei et al., 2021). 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 (Chen, 2016). 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 a decision tree algorithm. LGBM has faster training speed with lower memory usage compared to XGB (Ke et al., 2017). 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 LGBM 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 the 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) (Drucker et al, 1997).

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 (Ribeiro & Coelho, 2020). The main advantage of the SEL is that this approach can improve the accuracy and additionally 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 a selected feature set to forecast the confirmed Covid-19 cases by comparing their performances. Fine tuning predictive model hyperparameters 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 2. For SEL method, no need to tune the parameters since this method is a combination of the best regressions.

Table 2.

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=300
LGBM	num_leaves=10, learning_rate=0.1, n_estimators=100, reg_lambda=0.30
SVR	C=5.0, ε =0.2

4.3. Evaluation Criteria

To check the performance of th seven models used in this study, we use the following statistical measures:

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

where N is the number of data observations, is the actual count and is the predicted count

- $$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (4)$$

- $$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (5)$$

$$\text{MAPE} = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (6)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \quad (7)$$

The lowest value of RMSLE, RMSE, MAE, and MAPE refers to the best model while the higher value R2 indicates better correlation for the model.

RMSLE is less sensitive to outliers than other metrics[4]. RMSLE is preferable when there is a wide range in the target variables and targets having exponential growth, such as population counts. Therefore, we can rely more heavily on this metric.

4.4. Research Methodology

In Figure 1, the procedure of COVID-19 forecasting has been shown. In the first step, the real data of COVID-19 are collected for the analysis. After collecting the raw data, we use melt method to change a data-frame from wide to long format and then obtain a time series of data to work with. After that we extract a time series feature set from melting data, as explained in Section 2, and it is used as inputs of each model. Then, the resulting data is divided into a training and test sets. We assign the last two weeks of data as a test set in order to evaluate regression models. The comparative analysis of seven regression models with optimal hyper-parameters are done and the best prediction model is identified based on the prediction results. The performance of the models will be verified by comparing the predicted data with real data via different statistical measures, including RMSE, MAE, MAPE, RMSLE and R2.

All experiments are implemented using python and its libraries such as scikit-learn, numpy and pandas. We have performed our experiments in Intel Core i7 CPU clocked at 2.00 GHz, 16 GB RAM (Raschka, Patterson & Nolet, 2020).

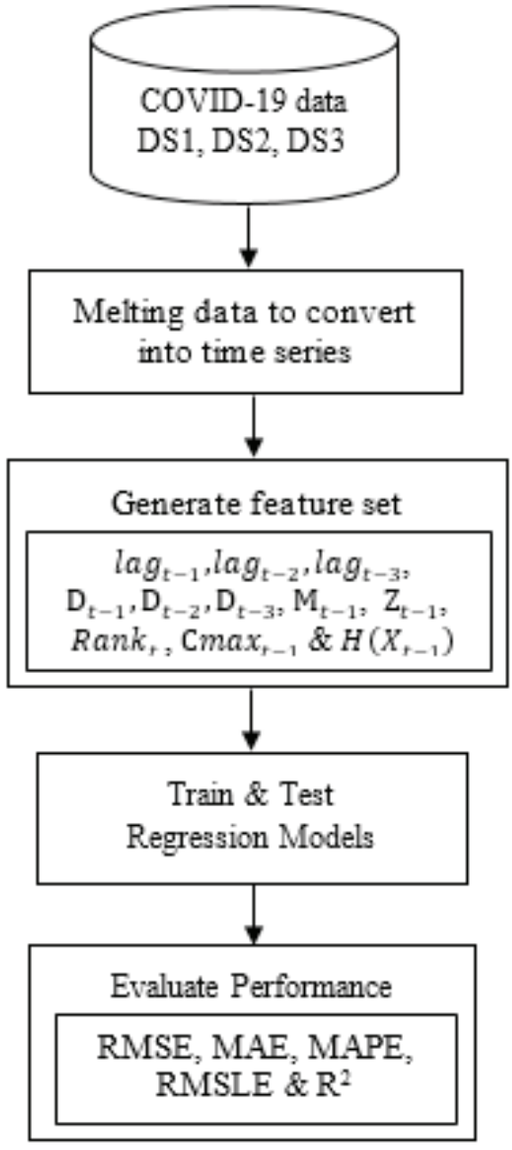


Figure.1

Proposed framework for forecasting

5. 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 three 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 for dataset DS1). The second round of the forecast was made 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 for dataset DS2). Once more data became available, the third round was done for another two weeks from 16/12/2020 to 29/12/2020 based on the actual data from 01/09/2020 to 15/12/2020 (training data for dataset DS3). Raw melting data and extracted feature set, explained in Section 2, are used as inputs of each regression model.

The baseline prediction algorithm is used as the criterion by which all other regression models can be compared. The first row of experiments table results shows the baseline scores. Thus, if a model achieves a predictive score below the baseline, it is good. Whereas model with higher R^2 value indicates a good model.

Table 3 shows the predictive scores obtained by all regression models on three datasets, and the best performance is shaded with grey. Overall, all models except SVR demonstrated good performance. It can be concluded from Table 3, the RF and XGB outperformed the compared models in all measures. The results indicate the XGB is ranked first in terms of RMSLE and MAPE whereas the RF is ranked first in terms of RMSE and MAE. They are almost equal in term of R^2. However, the time computation (in seconds) undertaken by XGB is less than RF. Moreover, the scores produced by LGBM and SEL relatively well and they are ranked third after RF and XGB. Note that LGBM achieved optimal run time because it speeds up the training process and is almost seven times faster than XGB[(2)]. Thus, it is a much better technique of handling large datasets.

The value of R^2 indicates the correlation between the prediction obtained by the regression model and the actual Covid-19 confirmed cases. From Table 3, the high values of R^2, which are .961, .990 and .978 on DS1, DS2 and DS3 respectively, indicate the fitting goodness for predicting confirmed cases.

Table 4 demonstrates the average predictive scores achieved by baseline and the first ranked model in terms of RMSLE, RMSE, MAE and MAPE for three datasets. Based on these averages, we estimate error reductions, which are high in terms of RMSLE and MAPE; good in terms of RMSE and MAE.

Table 3.

Comparison results obtained by regression models on three datasets.

Dataset	Algorithm	RMSLE	RMSE	MAE	MAPE	R^2	Time (s)
DS1	Baseline	0.682	1947	483	78.96	0.831	--
	DT	0.366	1936	588	30.46	0.855	0.41
	RF	0.199	975	255	13.88	0.961	235.79
	GBR	0.275	2381	637	20.94	0.828	87.51
	XGB	0.159	1094	286	11.35	0.960	10.65
	LGBM	0.175	1367	342	12.56	0.938	1.55
	SVR	0.557	3159	825	91.68	0.694	21.40
	SEL	0.200	1104	301	14.32	0.956	209.79
DS2	Baseline	0.669	2587	737	43.26	0.925	--
	DT	0.421	1960	640	36.41	0.968	0.28
	RF	0.237	1014	324	16.30	0.989	123.99
	GBR	0.285	3246	871	22.62	0.918	48.86
	XGB	0.167	1093	331	11.48	0.990	6.24
	LGBM	0.232	1459	422	14.32	0.983	1.40
	SVR	0.659	4436	1193	94.39	0.819	4.93
	SEL	0.248	1194	356	16.64	0.986	124.23
DS3	Baseline	0.572	6295	1827	29.72	0.871	--
	DT	0.405	7353	2280	35.46	0.873	0.44
	RF	0.198	2846	931	14.13	0.978	254.23
	GBR	0.315	10915	2568	26.50	0.830	95.83
	XGB	0.158	3291	981	12.41	0.975	11.92
	LGBM	0.178	4616	1278	14.24	0.964	1.80
	SVR	0.690	13284	3323	102.77	0.668	18.97
	SEL	0.193	4419	1231	13.98	0.946	232.19

Table 4.

Error reduction.

Algorithm	RMSLE	RMSE	MAE	MAPE
Baseline	0.641	3610	1016	152
XGB	0.175			38.26
RF		1612	503	
Error reduction	72.7%	55.3%	50.5%	74.8%

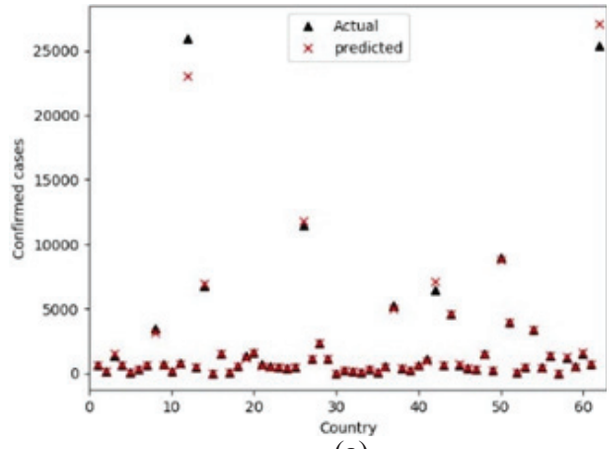
From the results, we can observe that the quality of our results is better than that appeared in recent previous studies. It can be easily seen that the RF model in our work outperformed the XGB and other models in Larabi-Marie-Sainte et al. (2022) for the four datasets (KSA, Brazil, Spain, and the US) in terms of RMSE and MAE. By comparing our estimates with those in Rguibi et al. (2022), we find that the prediction models in our study are better than ARIMA and LSTM models that used to predict the confirmed cases of Covid-19 in the upcoming two months in Morocco based on MAE and MAPE evaluation metrics. However, LSTM model showed lower prediction error values in term of RMSE and RMSLE. In such work, average values were RMSE = 795.3 and RMSLE = 0.00394. Another point to be highlighted is the comparison to the results obtained in the work of Wang et al. (2022), which tried to predict daily new cases in USA, Brazil and India over the next 30 days. We note that the prediction accuracies of our XGB and RF models are higher than that produced by SARIMA and Prophet models.

Since the best score was achieved by XGB, it was used to predict the Covid-19 daily-confirmed cases for the first day of the 2-weeks prediction. Figure 2 graphically compares the actual value with the predicted value of confirmed cases using XGB model for June 12, 2020, August 18, 2020 and December 16, 2020 respectively, where the x-axis represents the country, and the y-axis represents the corresponding daily-confirmed cases: actual (black) versus predicted (red). Figures 2 (a), (b) and (c) show that the dots representing the actual and predicted points are very close. However, we can observe that the XGB model gives some prediction values that are slightly far from the actual value, especially in DS1 and DS3, figure 2(a) and (c).

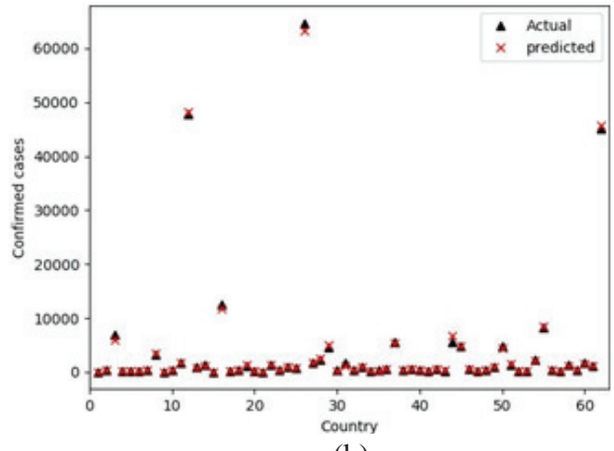
6. Conclusion

In this paper, we have conducted a three-round study of COVID-19 confirmed cases in the most affected countries worldwide. The most prominent techniques of regression models are used to analyze and predict the daily cases. The regression models are trained on a time series features, which are extracted through the original data. We have analyzed epidemic data made available by the CSSE within ten 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 experimental results show that XGB and RF models produced good scores in terms of the five measures over three rounds and they may be appropriate for predicting the prevalence of COVID-19 in the future. Our analysis can help in understanding the trends of the pandemic outbreak.

There are some limitations in the forecasted numbers of COVID-19 cases. First, some countries have missing values for some days, so raw data record 0 values for these days and add missing cases for this day to next days. Therefore, the results for these countries are not accurate. Second, the prediction models rely on past behavior. Therefore, the existence of outliers and noise in the data make 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)

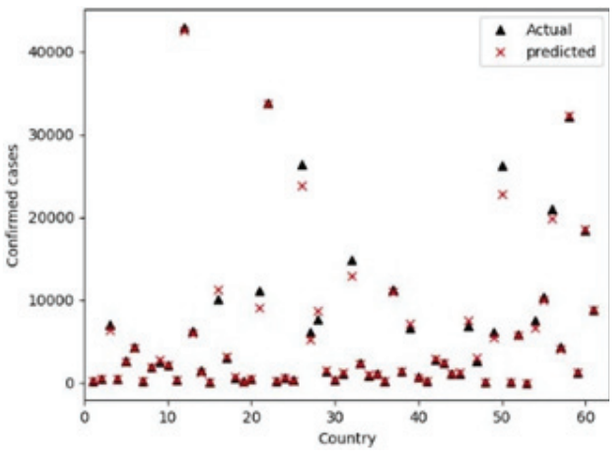


Figure 2.

Actual and predicted cases recorded for 62 countries in (a) June 12, 2020 (b) August 18, 2020 and (c) December 16, 2020

REFERENCES

- Abuzir, Y. S., Amro, I. Y., Tork, B., & Issa, M. (2021). A Prediction Model of Newly Admitted Students in the Level Exam Using Data Mining. Palestinian Journal of Technology and Applied Sciences (PJTAS), No 4, 23-35.
- Ahmad, A., Garhwal, S., Ray, S. K., Kumar, G., Malebary, S. J., & Barukab, O. M. (2021). The number of confirmed cases of covid-19 by using machine learning: Methods and challenges. Archives of Computational Methods in Engineering, 28(4), 2645-2653.
- Al-Qaness, M. A., Ewees, A. A., Fan, H., & Abd El Aziz, M. (2020). Optimization method for forecasting confirmed cases of COVID-19 in China. Journal of Clinical Medicine, 9(3), 674.
- Breiman, L. (2001). Random forests. machine learning, vol. 45, 5-32.
- Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
- 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(3), 761-770.
- Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1996). Support vector regression machines. Advances in neural information processing systems, 9.
- Feng, S., Feng, Z., Ling, C., Chang, C., & Feng, Z. (2021). Prediction of the COVID-19 epidemic trends based on SEIR and AI models. PLoS One, 16(1), e0245101.
- Gecili, E., Ziady, A., & Szczesniak, R. D. (2021). Forecasting COVID-19 confirmed cases, deaths and recoveries: revisiting established time series modeling through novel applications for the USA and Italy. PloS one, 16(1), e0244173.
- Gu, C., Zhu, J., Sun, Y., Zhou, K., & Gu, J. (2020). The inflection point about COVID-19 may have passed. Science bulletin, 65(11), 865-867.
- Gumaei, A., Al-Rakhami, M., Al Rahhal, M. M., Albogamy, F. R., Al Maghayreh, E., & AlSalman, H. (2021). Prediction of COVID-19 confirmed cases using gradient boosting regression method. Comput Mater Continua, 66, 315-329.
- Hassanat, A. B., Mnasri, S., Aseeri, M. A., Alhazmi, K., Cheikhrouhou, O., Altarawneh, G., ... & Almoamari, H. (2021). A simulation model for forecasting covid-19 pandemic spread: Analytical results based on the current saudi covid-19 data. Sustainability, 13(9), 4888.
- 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.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.
- Kibria, H. B., Jyoti, O., & Matin, A. (2022). Forecasting the spread of the third wave of COVID-19 pandemic using time series analysis in Bangladesh. Informatics in Medicine Unlocked, 28, 100815.
- Larabi-Marie-Sainte, S., Alhalawani, S., Shaheen, S., Almustafa, K. M., Saba, T., Khan, F. N., & Rehman, A. (2022). Forecasting COVID19 parameters using time-series: KSA, USA, Spain, and Brazil comparative Case study. Heliyon, e09578.
- Oliveira, T. D. P., & Moral, R. D. A. (2021). Global short-term forecasting of COVID-19 cases. Scientific reports, 11(1), 1-9.
- Raschka, S., Patterson, J., & Nolet, C. (2020). Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. Information, 11(4), 193.
- 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 & Metabolic Syndrome: Clinical Research & Reviews, 14(5), 1467-1474.
- Ribeiro, M. H. D. M., & dos Santos Coelho, L. (2020). Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. Applied Soft Computing, 86, 105837.
- Ribeiro, M. H. D. M., da Silva, R. G., Mariani, V. C., & dos Santos Coelho, L. (2020). Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. Chaos, Solitons & Fractals, 135, 109853.
- Rguibi, M. A., Moussa, N., Madani, A., Aaroud, A., & Zine-Dine, K. (2022). Forecasting covid-19 transmission with arima and lstm techniques in

morocco. SN Computer Science, 3(2), 1-14.

- Samson, T. K., Ogunlaran, O. M., & Raimi, M. O. (2020). A Predictive Model for Confirmed Cases of COVID-19 in Nigeria. TK Samson, OM Ogunlaran, OM Raimi (2020), 1-10.
- Vytla, V., Ramakuri, S. K., Peddi, A., Srinivas, K. K., & Ragav, N. N. (2021, February). Mathematical models for predicting COVID-19 pandemic: a review. In Journal of Physics: Conference Series (Vol. 1797, No. 1, p. 012009). IOP Publishing.
- Wang, Y., Yan, Z., Wang, D., Yang, M., Li, Z., Gong, X., ... & Wang, Y. (2022). Prediction and analysis of COVID-19 daily new cases and cumulative cases: times series forecasting and machine learning models. BMC Infectious Diseases, 22(1), 1-12.

Endnote

1. https://covid19.who.int

2. https://machinelearningmastery.com/basic-feature-engineering-time-series-data-python

3. https://github.com/CSSEG ISand Data/COVID-19

4. https://medium.com/analytics-vidhya/root-mean-square-log-error-rmse-vs-rmlse-935c6cc1802a

5. https://neptune.ai/blog/xgboost-vs-lightgbm