Machine learning and automatic ARIMA/Prophet models-based forecasting of COVID-19: methodology, evaluation, and case study in SAARC countries

Iqra Sardar1 · Muhammad Azeem Akbar2 · Víctor Leiva3 · Ahmed Alsanad4 · Pradeep Mishra5

Accepted: 26 August 2022 / Published online: 5 October 2022 © The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract
Machine learning (ML) has proved to be a prominent study field while solving complex real-world problems. The whole globe has suffered and continues suffering from Coronavirus disease 2019 (COVID-19), and its projections need to be forecasted. In this article, we propose and derive an autoregressive modeling framework based on ML and statistical methods to predict confirmed cases of COVID-19 in the South Asian Association for Regional Cooperation (SAARC) countries. Automatic forecasting models based on autoregressive integrated moving average (ARIMA) and Prophet time series structures, as well as extreme gradient boosting, generalized linear model elastic net (GLMNet), and random forest ML techniques, are introduced and applied to COVID-19 data from the SAARC countries. Different forecasting models are compared by means of selection criteria. By using evaluation metrics, the best and suitable models are selected. Results prove that the ARIMA model is found to be suitable and ideal for forecasting confirmed infected cases of COVID-19 in these countries. For the confirmed cases in Afghanistan, Bangladesh, India, Maldives, and Sri Lanka, the ARIMA model is superior to the other models. In Bhutan, the Prophet time series model is appropriate for predicting such cases. The GLMNet model is more accurate than other time-series models for Nepal and Pakistan. The random forest model is excluded from forecasting because of its poor fit.

Keywords Artificial intelligence · Facebook Prophet algorithm · GLM · R software · SARS-CoV-2 · South Asian Association for Regional Cooperation countries · Time-series models

1 Introduction
Machine learning (ML) has solved many complex real-world problems over the last decade in diverse fields as business, climatology, healthcare, and robotics (Kelleher et al. 2020; Akbar et al. 2022). ML algorithms are trial-and-error-based methods rather than conventional algorithms that follow programming commands based on decision statements (Makridakis et al. 2018; Sardar et al. 2021). Forecasting/prediction is the most significant part of ML (Bontempi et al. 2012), and multiple ML algorithms have been used to forecast diseases, stock markets, and weather (Mahdi et al. 2021; Bustos et al. 2022; Chaouch et al. 2022; Ma et al. 2022). Regression and neural network models have been employed to predict a patient’s future condition for a specific disease (Harrell et al. 1985). Different studies have been performed for diseases’ prediction, such as breast cancer (Asri et al. 2016), cardiovascular...
(Anderson et al. 1991), and coronary artery (Lapuerta et al. 1995) conditions, via ML techniques.

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), also known as COVID-19, was discovered in December 2019, in Wuhan, China, and declared by The World Health Organization (WHO) as a pandemic in March 2020 (WHO 2020a, b). Different types of SARS-CoV-2 have been identified (Alkadya et al. 2022). COVID-19 has revolutionized the world’s population, as well as the economy (Chahuan-Jimenez et al. 2021) and finance (Liu et al. 2021), forcing us to a new manner of life that will keep its mark on society forever (Mahdi et al. 2021).

The spread of COVID-19 was so rapid that almost all countries imposed either partial or complete lockdown in affected areas to curb its spread (Ospina et al. 2022). Precautionary measures imposed by different governments are directing their masses to follow the standard operating procedure and control the spread of SARS-CoV-2 (Jerez-Lillo et al. 2022). Several studies based on ML tools have tried to predict COVID-19 confirmed cases (Petropoulos and Makridakis 2020) and COVID-19 outbreaks (Grasselli et al. 2020), as well as improving its clinical diagnostic (Bustos et al. 2022), which help in decision-making.

The South Asian Association for Regional Cooperation (SAARC) was established in 1985, and it is composed by Afghanistan, Bangladesh, Bhutan, India, the Maldives, Nepal, Pakistan, and Sri Lanka. COVID-19 has disturbed the economy of SAARC countries as well (De la Fuente-Mella et al. 2021). The confirmed cases in overall SAARC countries are 16,870,105, with 39,053 as deaths till 19 July 2020 (WHO 2020a, b; Zhao et al. 2022). India is the largest country of the SAARC affected from COVID-19, with cases rising to 1.1 million, which is second only to the US and with a recovery rate of 63.75%. In Latin America, the deaths have been high as well (Martin-Barreiro et al. 2021), particularly in Brazil (Ospina et al. 2022) and Chile (Jerez-Lillo et al. 2022).

The COVID-19 infection rate in SAARC countries is approximately 11.78% of the world population. India has developed medical facilities to fight against it. However, despite lockdowns, three million spiked active cases were reported in India. Comparatively, Bangladesh and Pakistan have crossed two lacs of cases of COVID-19. Active cases in Pakistan are gradually decreasing, and now it is 53,652 cases with less testing facilities. The death rate in Bangladesh is 6.61%. No casualty was found to date due to COVID-19 in Afghanistan and Bhutan. The death rate is below 1% in the Maldives and Sri Lanka.

Time-series models are widely studied and applied (Brockwell and Davis 1991; Leiva et al. 2021), especially in epidemic disease prediction (Ospina et al. 2022; Jerez-Lillo et al. 2022). Several researchers are studying different dimensions of the epidemic and trying to find the results to help humanity. In Fanelli and Piazza (2020), a temporal dynamic was considered for predicting COVID-19 in China, France, and Italy. In Chimmula and Zhang (2020), the COVID-19 outbreak in Canada was forecasted by employing a long short-term memory network. In Malavika et al. (2021), COVID-19 was predicted in India based on a SIR model (Fierro et al. 2015; Rangasamy et al. 2022). In Petropoulos and Makridakis (2020) and Alzahrani et al. (2020), exponential smoothing and autoregressive integrated moving average (ARIMA) models were used for forecasting COVID-19.

Currently, ML is a helpful technique for forecasting and researchers are employing it in different fields of science. In Rustam et al. (2020), ML models were compared with varying forecasting models based on COVID-19 data, finding that the ML model performance was better than other models. For example, in Sujath et al. (2020), ML techniques were utilized for predicting COVID-19 data in India. The employment of ML tools for forecasting COVID-19 have been very helpful in controlling this epidemic. To the best of our knowledge, no studies based on ML techniques and automatic ARIMA/Prophet models for forecasting COVID-19 cases on SAARC countries have been conducted.

The specific contributions that this investigation aims to make are to:

1. Propose a methodology for ML-based forecasting of COVID-19 data. The proposed methodology is represented in Fig. 1 to state the problem setting.
2. Derive an autoregressive modeling framework for forecasting COVID-19 data in a region composed of countries with different realities.
3. Compare different forecasting models by means of selection criteria.
4. Choose the best forecasting model for SAARC countries based on COVID-19 data by using evaluation metrics.
5. Forecast the COVID-19 confirmed cases in SAARC countries.
6. Analyze the impact of forecasted COVID-19 confirmed cases on these countries.

Therefore, the objectives of this study are related to each of the contributions mentioned in (1)–(6). All the calculations and graphical plots were performed with the R software (R Development Core Team 2020). The plots were produced with ggplot2 of R by using a tool of interactive forecast visualization. This is a wrapper for plot_time_series() that generates an interactive(plotly) or static (ggplot2) plot with the forecasted data. The plots of time series were constructed with interactive plotting for one or more time series. This is a workhorse time-series plotting function that generates interactive plots.
plotly plots, consolidates 20+ lines of ggplot code, and scales well to many time series.

The remainder of this article is organized as follows. Section 2 discusses the methodology, including data source, automatic and ML models, as well as the evaluation metrics used in the present investigation. In Sect. 3, we report the results of our study and provide a discussion of these results. In Sect. 4, conclusions and future work suggestions are stated.

2 Methodology

In this section, the data source, methods, and evaluation metrics are provided.

2.1 Data and their sources

The COVID-19 data set employed in the present investigation was sourced from “Our World in Data” (Ritchie et al. 2020) using the website:

https://ourworldindata.org/coronavirus;

which was accessed on 14 May 2022. In this study, we have considered only the variable “number of confirmed COVID-19 cases” for SAARC countries and its corresponding time-series data with its metadata related to date and location. The data were collected from 31 January 2020 to 19 July 2020 from the repository:

github.com/owid/covid-19-data/tree/master/public/data,

where other COVID-19 variables can be obtained as well. Specifically, the data set used in our analysis was secured from:

github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data.csv

which was accessed on 14 May 2022. The original data source (raw data) on confirmed COVID-19 cases and deaths for all countries correspond to the COVID-19 data repository of the “Center for Systems Science and Engineering (CSSE)” at Johns Hopkins University (www.jhu.edu) with link given by:

github.com/CSSEGISandData/COVID-19.

The full COVID-19 data set is a collection of the COVID-19 data maintained by Our World in Data (ourworldindata.org), which is updated daily and includes data on confirmed cases, deaths, hospitalizations, and testing.

2.2 Time-series models

In this study, we use automatic forecasting models based on: (i) ARIMA and Prophet time-series frameworks; as well as (ii) extreme gradient boosting (XGBoost), generalized linear model elastic net (GLMNet), and random forest ML methods. We employ these models to analyze the COVID-19 data from the SAARC countries mentioned in Sect. 2.1. Next, we describe each of these methods.

Automatic ARIMA models: A statistical analysis model known as ARIMA employs time-series data to help researchers better understand a data collection or to forecast future trends. If a statistical model forecasts future values using data from the past, it is said to be autoregressive. These models (Hyndman and Khandakar 2008) combine unit root tests and minimize the Akaike information criterion, AIC in short (Ventura et al. 2019). Note that the maximum likelihood method can be used to estimate the ARIMA parameters. Indeed, the R software estimates the ARIMA parameters with this method so that we use it. When estimating the ARIMA parameters, the maximum likelihood method is like the least squares method. Note that the maximum likelihood estimators are more efficient than other estimators.

The ARIMA acronym is formed by the terms AR, I, and MA, where the notation ARIMA(p, d, q) is often utilized. In this notation, the acronym AR(p) is related to the autoregressive part of the model, and the parameter p specifies how many lagged series should be used to forecast future periods (Hyndman et al. 2006). The acronym I(d) is associated with the integrated part of the
model, and the parameter \( d \) indicates how many differences are needed to make the series stationary, that is, \( I(d) \) is the differencing component or number of differences. Then, the acronym \( MA(q) \) is the moving average part of the model, where \( q \) is the number of lag elements in the prediction equation for the forecast error. An ARIMA(\( p, d, q \)) model is stated as

\[
\phi(B)(1-B)^d)Y(t) = \delta + \theta(B)e(t),
\]

(1)

where \( \phi(B) \) and \( \theta(B) \) are polynomials of orders \( p \) and \( q \), respectively, assuming that \( \phi(B) \) and \( \theta(B) \) have no roots for \( |B|<1 \) to ensure causality and invertibility, whereas \( \delta \) is a constant. In the expression defined in (1), \( B \) is a backshift operator; \( Y(t) \) is the response variable; and \( e(t) \) is a white noise both at time \( t \), with \( e(t) \sim N(0, \sigma^2) \), that is, the white noise follows a normal (Gaussian) distribution with zero mean and variance \( \sigma^2 \). Note that there is an indirect polynomial of order \( d \) in the forecasting if \( \delta \neq 0 \) (Brockwell and Davis 1991). An approach for automatic ARIMA modeling is stated in Algorithm 1.

![Algorithm 1 Automatic ARIMA modeling](image)

where \( Y(t) \) stated in (2) is the response variable to be predicted at time \( t \); \( g(t) \) is a trend function employed to analyze time-series non-periodic changes; \( s(t) \) is a periodic or seasonality term reflecting the change of a week or a year; \( h(t) \) is the influence of occasional days or holidays; ; and \( e(t) \) is an error term which is assumed to be normally distributed such as in (1). Here, we consider the non-periodic changes of time series. As mentioned, the Prophet model handles the outliers of time series as well as the missing values. This model can automatically forecast the future trend, with this trend \( g(t) \) stated in (2) being described by saturating growth and a piecewise linear models. The growth model is established by a logistic regression formulated as

\[
g(t) = \frac{K}{1 + e^{-(c+\beta(t)\gamma)\cdot(t-(u+\beta(t)\varphi))}},
\]

(3)

where \( e \) is the exponential or Euler function. In the model presented in (3), the elements \( K \), \( c \), and \( u \), corresponding to the curve maximum value or carrying capacity \( (K) \), logistic growth rate or steepness of the curve \( (c) \), and sigmoid point or offset parameter \( (u) \), are not constant, because they depend on time \( t \), so that the formulation stated in (3) becomes as

\[
g(t) = \frac{K(t)}{1 + e^{-(c(t)+\beta(t)\gamma(t)-(u(t)+\beta(t)\varphi))}}.
\]

(4)

When one finds the change point of the time series, the trend changes at this point, and then the model expressed in (4) is now given by

\[
g(t) = \frac{K(t)}{1 + e^{-(c(t)+\beta(t)\gamma(t)-(u(t)+\beta(t)\varphi))}}.
\]

(5)

where \( \beta(t) = (\beta_1(t), \ldots, \beta_s(t))^\top \), \( \gamma = (\gamma_1, \ldots, \gamma_s)^\top \), and \( \varphi = (\varphi_1, \ldots, \varphi_s)^\top \), with \( (\cdot)^\top \) denoting the transpose of a matrix. For a piecewise linear function, the model defined in (5) is established as

\[
g(t) = (c(t)+\beta(t)\gamma(t)) t + (u(t)+\beta(t)\varphi(t)),
\]

where \( c \) is the growth rate, \( u \) denotes the offset parameter, \( \gamma \) has the rate adjustments, and \( \varphi \) is set to make the function continuous; for more details, see Taylor and Letham (2018).

2.3 Machine learning models

ML algorithms are a branch of computer science that are trained from past data. The algorithm selects a suitable model to be calibrated according to data characteristics and forecasts future values. In this ML framework, the algorithm takes a data set with input instances and a regressor to train the model. The trained model generates a prediction for an unpredicted test data set (Talabis et al. 2015). The ML methods use regression and classification algorithms.

Several ML models, as logistic regression, neural networks, or support vector machines, have been utilized to analyze and forecast COVID-19 (Bustos et al. 2022). We have found that some ML models predict better when compared to others of this type. We apply ML models to our COVID-19 data set for predicting the effects of SARS-CoV-2 in SAARC countries and to global level. Three ML methods are employed in the study of COVID-19 prediction. These methods correspond to GLMNet, random forest, and XGBoost.
GLMNet: The elastic net is a particular case of the shrinkage method, which holds both ridge and least absolute shrinkage and selection operator (LASSO) regressions. The GLMNet property has the ability to handle \( n < p \) problems (Zou and Hastie 2005). The elastic net penalty term is given by

\[
P_c(a) = \sum_{j=1}^{p} (0.5(1-\gamma)x_j^2 + \gamma|x_j|),
\]

in the objective function (Friedman et al. 2005) stated as

\[
\min_{(x_0, x) \in \mathbb{R}^{p+1}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (Y_i - x_0 - z_i^T x)^2 + \tau P_c(x) \right\},
\]

where now \( \gamma \) is a parameter that controls the shrinkage type affecting the estimation process; \( \tau \) is a penalty parameter that controls the shrinkage amount; \( z_i \in \mathbb{R}^p \) are the values of the covariate vector \( Z \in \mathbb{R}^p \); \( x_0 \) is the regression intercept; \( x = (x_1, \ldots, x_p) \) is the vector of regression coefficients associated with the values of the covariate vector, with \( x_j \) being the \( x_j \)-th element of the vector \( x \); \( p \) is the number of covariates; and \( n \) is the size of the sample data set.

Note that we are assuming a linear regression structure given by \( E(Y|Z = z) = x_0 + z^T x \), where “E” denotes the expected value of \( Y \) conditional to \( Z = z \). The glmnet package of the R software (R Development Core Team 2020) offers various kinds of regression methods for both variable selection and prediction techniques when \( n < p \), depending on the data and problem under consideration. If \( \gamma = 1 \), the LASSO defaulting option of the glmnet package is obtained with a penalty parameter \( l_1 \) and carries out both parameter shrinkage and variable selection. If \( \gamma = 0 \), it gives a ridge regression with a penalty parameter \( l_2 \). We choose \( 0 < \gamma < 1 \) to optimize the elastic net. Efficiently, this shrinks certain coefficients and makes some equal to zero for sparse selection. The model structure needs variable selection to form a subset of predictors. The elastic net uses the \( p >> n \) problem approach, which implies that the number of parameters is much greater than the size of the sample utilized in the modeling. The elastic net is suitable when the variables form groups that contain highly correlated covariates. The variable selection is combined with the model structure to help in increasing the accuracy. When a group of covariates is highly correlated, the model automatically selects those with greater predictive power.

Random forest: To build a set of decision trees, two methods are combined: bootstrap aggregating and random subspace, with the decision tree set being categorized. Random forest comprises many decision tree classifiers, whose classification of the decision tree obtains the output groups. In the single decision tree, two random selection processes are used in the algorithm. The first one is the random selection of training samples, and the second one is a random selection of sample features. Then, the decision trees are built. Classification results are finalized by the weight voting method. Random forest is a joint classifier that builds independent and non-matching decision trees depending on randomization. It can be defined as \( g(y, \varphi_l) \), for \( l \in \{1, \ldots, L\} \), where now \( \varphi_l \) is the mutual independent random vector parameter, and \( y \) are data input values (Provost et al. 2016). Every decision tree uses a random vector of parameters, features of samples selected randomly, and a subset of sample data chosen randomly as a training set. The structure of the random forest is stated in Algorithm 2 (Bradter et al. 2013).
**Algorithm 2** Structure of the random forest

1. Select \( l \) decision trees, \( n \) training data that each tree links, and \( N \) bootstrap samples.
2. Form \( l \) training groups with \( N \) bootstrap samples.
3. Collect \( m \) dissection features randomly on a single node from \( M \) sample features, for \( m << M \).
4. Calculate the best dissection feature according to \( m \) features for every decision tree node.
5. Construct each decision tree entirely increasing without trimming.
6. Classify data from the testing set with the random forest model formed by various decision trees.
7. Evaluate the adequacy of the random forest model.
8. Use the random forest model if this is adequate.

**XGBoost:** The extreme gradient boosting model (Tianqi C, Guestrin 2016) is an ensemble ML system for tree boosting. The XGBoost model is used for problems related to regression trees and classification based on the boosting technique. At each step, the XGBoost model produces a weak learner and cumulates it into the model. The XGBoost model is based on the gradient direction of a loss function called gradient boosting machine (Jerome and Friedman 2002). The difference between random forest and gradient boosting machine is that decision trees are built independently in a random forest, while the another method adds a new tree to counterpart previously built ones (Luckner et al. 2017). The XGBoost is a method where new models are built to predict prior models’ residuals and further make predictions. The XGBoost model is established as

\[
F = \lambda + \omega,
\]

where \( \lambda \) is the loss function that controls the predictive power and \( \omega \) is the regularization part of controlling ease and over-fitting.

### 2.4 Model evaluation metrics

Evaluation metrics are used to measure the quality of statistical models. Evaluation metrics tell us that our model is operating correctly. There are numerous kinds of evaluation metrics to test a model. The metrics employed in this investigation are defined next. The mean absolute error is established as

\[
\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |Y_i - F_i|,
\]

which indicates the difference between the observed \( (Y_i) \) and forecasted \( (F_i) \) values for case \( i \). The MAE is obtained from the data set by averaging the absolute difference.

The mean absolute percentage error is stated by

\[
\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{Y_i - F_i}{Y_i} \right| \times 100\%,
\]

which measures the size of error in percentage form. The mean absolute scaled error (Hyndman and Khandakar 2008) is formulated as

\[
\text{MASE} = \frac{\text{MAE}}{\text{MAE}_{\text{naive-sample}}} = \frac{\frac{1}{n} \sum_{i=2}^{n} |Y_i - F_i|}{(T-1) \sum_{t=2}^{T} |Y_t - Y_{t-1}|},
\]

which is used as a measure of the accuracy of forecasts, with the denominator \( \text{MAE}_{\text{naive-sample}} \) being the mean absolute error of the one-step “naive forecast method” on the training set (here defined as \( t \in \{1, \ldots, T\} \)), that utilizes the true value from the prior period as the forecast: \( F_t = Y_{t-1} \). The symmetric mean absolute percentage error is expressed by

\[
\text{SMAPE} = \frac{1}{n} \left( \sum_{i=1}^{n} \left| \frac{Y_i - F_i}{\frac{|Y_i| + |F_i|}{2}} \right| \right) \times 100\%,
\]

which is based on the percentage error. The root of the mean square error is given by

\[
\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - F_i)^2}{n}},
\]

which corresponds to the square root of the mean square error. A well-known indicator of adequacy for statistical models is the determination coefficient that is defined in our case as

\[
\text{R-square} = 1 - \frac{\sum_{i=1}^{n} (Y_i - F_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2},
\]

which tells us how well the fit values are compared to the true values, where \( \bar{Y} \) is the mean of true values. After the training model is obtained, we check the model goodness of fit. The values between zero and one are interpreted as percentages. Note that 0% shows that the response variable is not explained around its mean based on the covariates, whereas 100% indicates that the model describes the
response variability around its mean perfectly. As the value of R-square of the trained model increases, the model better fits the data set.

3 Results and discussion

In this section, we use the automatic and ML models to forecast the COVID-19 confirmed cases in SAARC countries. The forecasting is done by employing the best models that are suitable for this framework. Firstly, the data set is preprocessed. Then, we split the data into a set to train the models and another set (10 days) to test the models. ML models as GLMNet, XGBoost, and automatic models, as ARIMA and Prophet, are used. These models are trained for confirmed cases and evaluated on the mentioned metrics. The flowchart of the proposed methodology is shown in Fig. 2.

To compare the models, we evaluate their performance based on the metrics mentioned (MAE, MAPE, MASE,
We consider the confirmed COVID-19 cases of SAARC countries. Up to 19 July 2020, confirmed cases in the world were 14,443,127, and in SAARC countries 1,605,501; of which 1,033,502 are recovered cases; 36,270 are deaths; and 535,728 are active cases. The status of COVID-19 in SAARC countries is presented in Table 1, which includes the information of confirmed cases, deaths, recovered cases, and active cases till 19 July 2020. India has the highest number of COVID-19 active cases with 26,348 deaths. In Sri Lanka, the Maldives, and Nepal, the death shares below 50, whereas 204,276 patients were recovered from COVID-19 in Pakistan. The number of actives cases is 90,265 in Bangladesh.

According to the objective of our investigation, we compare forecasting models based on selection criteria. Best models are selected based on MAE, MAPE, MASE, SMAPE, RMSE, and R-square. The performance of the ARIMA model is best in most countries; see Table 2.

![Fig. 3 Time-series plot of confirmed COVID-19 cases of (a) global; and (b) SAARC countries](image)

Figures 5, 6, 7, 8, 9, 10, 11, 12 and 13 show the forecasted plot for all models used in this study. Description of best models and forecast values are provided in Table 2. For SAARC countries, the ARIMA(1, 2, 0)(1, 0, 2) model with drift \((p, d, q)(P, D, Q)_m(D = 0)\) is suitable for forecasting. Based on this, cases reach 1,315,265 on July 2020 (Table 3). The residual values of the best prediction model were the smallest ones, and it shows that the model is the best fit for the data; see Table 4. Note that the random forest model is excluded from forecasting during the analysis because of its poor fit. Model accuracy values are provided for an individual member of SAARC countries shown in Tables 4, 5, 6, 7, 8, 9, 10 and 11.

According to the ARIMA model, in Afghanistan, confirmed cases are greater than in other models. Table 4 reports that the values of model selection are better in the ARIMA(0, 2, 3) model. Performance of time-series models of Bangladesh COVID-19 confirmed cases is provided in Table 5, where the ARIMA(0, 2, 1)(2, 0, 0) model with drift is detected as the best one. The predicted values of

| SAARC countries | Confirmed cases | Recovered cases | Deaths | Active cases |
|-----------------|----------------|----------------|--------|--------------|
| India           | 107,878        | 677,856        | 26,838 | 374,088      |
| Pakistan        | 263,496        | 204,276        | 5568   | 53,652       |
| Bangladesh      | 204,525        | 11,164         | 2618   | 90,265       |
| Afghanistan     | 35,475         | 23,634         | 118    | 10,660       |
| Nepal           | 1750           | 11,637         | 40     | 5825         |
| Maldives        | 2930           | 2354           | 15     | 561          |
| Sri Lanka       | 2704           | 2023           | 1      | 670          |
| Bhutan          | 87             | 80             | 0      | 7            |

Source: Worldometer: www.worldometers.info/coronavirus/#countries; accessed on 10 July 2020
Fig. 4 Time-series plot of COVID-19 confirmed cases per month of (a) Afghanistan; (b) Bangladesh; (c) Bhutan; (d) India; (e) Maldives; (f) Nepal; (g) Pakistan; and (h) Sri Lanka
cases in Bangladesh are 181,051, which are very near to the true value of 181,129 on 12 July 2020 (Zhao et al. 2022).

The Prophet time series model is suitable for prediction in Bhutan; see Table 6. The model predicted value is almost the same as the true value of Bhutan. The values of RMSE, MAE, MAPE were the smallest ones in Bhutan. Still, in this country, the number of confirmed cases is 87 with no deaths due to COVID-19, which shows less spread.

For India, the forecasting in terms of model accuracy indicates that the ARIMA model is better compared to Prophet, XGBoost, and GLMNet methods, which is reported in Table 7. The prediction values for the confirmed cases are 811,178.8, which is very close to the true values. Medical facilities were improved in the last four months, and the recovery rate is greater than 60%.

Table 8 reports that the COVID-19 cases in the Maldives require stationary differencing. The prediction value of 12 July 2020 is like the true value. After Bhutan, the Maldives is having less COVID-19 confirmed cases among other SAARC countries. For the Bhutan model, residual values were less than 1%. In the Nepal data set, the GLMNet ML model is more accurate as compared to time series models: see Table 9. The values of RMSE (834.10) and MAE (641) are minimum in all compared models.
Fig. 7 Plot of forecasting for confirmed COVID-19 cases in Bangladesh

Fig. 8 Plot of forecasting for confirmed COVID-19 cases in Bhutan

Fig. 9 Plot of forecasting for confirmed COVID-19 cases in India

Fig. 10 Plot of forecasting for confirmed COVID-19 cases in the Maldives

Fig. 11 Plot of forecasting for confirmed COVID-19 cases in Nepal

Fig. 12 Plot of forecasting for confirmed COVID-19 cases in Pakistan
Table 10 provides the model accuracy of Pakistan about the confirmed COVID-19 cases. Like Nepal, in Pakistan also, the GLMNet model performance is better as compared to time-series models. The values of the random forest method are very different from other models.

The forecasted values of the GLMNet method in Pakistan show the COVID-19 confirmed cases are 250,528 on 12 July 2020. Still, according to its spread, Pakistan needs to increase the testing facilities, as the recovery rate is 97.35%. Table 11 reports that the ARIMA model accuracy is much better in Sri Lanka. The residual values are small and show that the ARIMA model is the best fitted.

Fig. 13 Plot of forecasting for confirmed COVID-19 cases in Sri Lanka

Table 3 Accuracy results of confirmed COVID-19 cases in SAARC countries

| Model              | MAE   | MAPE | MASE | SMAPE | RMSE  | R-square |
|--------------------|-------|------|------|-------|-------|----------|
| ARIMA(1,2,0)(1,0,2) | 9935.48 | 0.80 | 0.3  | 0.81  | 11,541.89 | 1.00     |
| Prophet            | 13,140.0 | 1.05 | 0.4  | 1.05  | 16,953.09 | 1.00     |
| GLMNet             | 91,372.40 | 7.44 | 2.89 | 7.78  | 100,978.10 | 0.99     |
| Random forest      | 589,028.17 | 48.94 | 18.65 | 65.35 | 601,966.30 | 0.86     |
| XGBoost            | 13,136.24 | 1.05 | 0.4  | 1.05  | 16,879.37 | 1.00     |

Table 4 Accuracy results of confirmed COVID-19 cases in Afghanistan

| Model              | MAE   | MAPE | MASE | SMAPE | RMSE  | R-square |
|--------------------|-------|------|------|-------|-------|----------|
| ARIMA(0,2,3)       | 123.36 | 0.37 | 0.46 | 0.37  | 157.59 | 0.98     |
| Prophet            | 547.86 | 1.6  | 2.03 | 1.57  | 862.79 | 0.94     |
| GLMNet             | 1261.75 | 3.76 | 4.68 | 3.84  | 1323.77 | 0.97     |
| Random forest      | 10,011.0 | 29.9 | 37.09 | 35.46 | 10,239.54 | 0.55     |
| XGBoost            | 620.17 | 1.8  | 2.30 | 1.79  | 918.94 | 0.93     |

Table 5 Accuracy results of COVID-19 data in Bangladesh

| Model              | MAE   | MAPE | MASE | SMAPE | RMSE  | R-square |
|--------------------|-------|------|------|-------|-------|----------|
| ARIMA(0,2,1)(2,0,0) | 358.69 | 0.20 | 0.31 | 0.20  | 397.33 | 1.00     |
| Prophet            | 3384.84 | 1.90 | 0.42 | 1.92  | 3772.35 | 0.99     |
| GLMNet             | 6507.96 | 3.67 | 2.89 | 3.74  | 6622.81 | 1.00     |
| Random forest      | 67,951.44 | 38.4 | 18.65 | 47.56 | 68,036.40 | 0.65     |
| XGBoost            | 3445.9 | 1.93 | 0.42 | 1.96  | 3825.89 | 0.98     |

Table 6 Accuracy results of confirmed COVID-19 cases in Bhutan

| Model              | MAE   | MAPE | MASE | SMAPE | RMSE  | R-square |
|--------------------|-------|------|------|-------|-------|----------|
| ARIMA(0,2,1)(1,0,0) | 1.18  | 1.44 | 0.56 | 1.45  | 1.23  | 0.78     |
| Prophet            | 0.35  | 0.43 | 0.23 | 0.43  | 0.38  | 0.94     |
| GLMNet             | 0.70  | 0.85 | 1.89 | 0.86  | 0.74  | 0.81     |
| Random forest      | 27.44 | 33.47 | 17.7 | 40.20 | 27.45 | 0.53     |
| XGBoost            | 0.54  | 0.65 | 1.23 | 0.66  | 0.57  | 0.69     |
Conclusion

In this article, we have proposed and derived an autoregressive modeling framework based on machine learning and statistical methods to predict the number of COVID-19 cases in the South Asian Association for Regional Cooperation countries. We have considered automatic forecasting models based on autoregressive integrated moving average and Prophet time series structures, as statistical methods. In the case of machine learning, we have employed extreme gradient boosting, generalized linear model elastic net, and random forest techniques.

The dataset was based on time series data from 31 January 2020 to 19 July 2020, and was used for upcoming forecasting. Different forecasting models were compared utilizing several evaluation metrics, based on MAE, MAPE, MASE, SMAPE, RMSE, and R-square values.

| Model | MAE   | MAPE  | MASE  | SMAPE | RMSE  | R-square |
|-------|-------|-------|-------|-------|-------|-----------|
| ARIMA(0,2) | 2.25  | 0.70  | 2.28  | 20,776.11 | 1.00   |
| Prophet  | 6.00  | 1.90  | 6.30  | 61,119.96 | 0.94   |
| GLMNet   | 10.2  | 3.15  | 10.90 | 89,010.92 | 0.99   |
| Random forest | 53.35 | 15.94 | 73.46 | 406,965.77 | 0.84   |
| XGBoost  | 6.00  | 1.90  | 6.30  | 61,076.94 | 0.94   |

Table 7: Accuracy results of confirmed COVID-19 cases in India

| Model | MAE   | MAPE  | MASE  | SMAPE | RMSE  | R-square |
|-------|-------|-------|-------|-------|-------|-----------|
| ARIMA(0,2) | 47,200.57 | 6.00 | 1.90 | 6.30 | 61,119.96 | 0.94   |
| Prophet  | 78,360.15 | 10.2 | 3.15 | 10.90 | 89,010.92 | 0.99   |
| GLMNet   | 396,671.74 | 53.35 | 15.94 | 73.46 | 406,965.77 | 0.84   |
| Random forest | 47,178.79 | 6.00 | 1.90 | 6.30 | 61,076.94 | 0.94   |
| XGBoost  | 47,200.57 | 6.00 | 1.90 | 6.30 | 61,119.96 | 0.94   |

Table 8: Accuracy results of confirmed COVID-19 cases in the Maldives

| Model | MAE   | MAPE  | MASE  | SMAPE | RMSE  | R-square |
|-------|-------|-------|-------|-------|-------|-----------|
| ARIMA(1,2,2)  | 1327.80 | 8.1   | 5.43  | 7.64  | 1658.72 | 0.93   |
| Prophet  | 940.6  | 5.76  | 3.85  | 5.52  | 1151.74 | 0.95   |
| GLMNet   | 641.05 | 3.9   | 2.62  | 3.79  | 834.10  | 0.94   |
| Random forest | 7357.50 | 46.08 | 30.10 | 59.95 | 7390.10 | 0.79   |
| XGBoost  | 848.74 | 5.18  | 3.47  | 4.99  | 1062.04 | 0.95   |

Table 9: Accuracy results of confirmed COVID-19 cases in Nepal

| Model | MAE   | MAPE  | MASE  | SMAPE | RMSE  | R-square |
|-------|-------|-------|-------|-------|-------|-----------|
| ARIMA(3,2,2) | 4361.95 | 1.80  | 1.46  | 1.78  | 5375.71 | 1.00   |
| Prophet  | 21,688.37 | 8.93 | 7.24  | 8.3   | 28,595.20 | 0.96   |
| GLMNet   | 1354.18 | 0.56  | 0.45  | 0.56  | 1676.14 | 1.00   |
| Random forest | 104,982.54 | 44.29 | 35.03 | 57.28 | 106,490.52 | 0.74   |
| XGBoost  | 21,830.37 | 8.99 | 7.28  | 8.36  | 28,686.44 | 0.96   |

Table 10: Accuracy results of confirmed COVID-19 cases in Pakistan

| Model | MAE   | MAPE  | MASE  | SMAPE | RMSE  | R-square |
|-------|-------|-------|-------|-------|-------|-----------|
| ARIMA(0,2,2) | 101.0  | 4.18  | 2.04  | 4.44  | 173.65 | 0.70   |
| Prophet  | 711.7  | 30.88 | 14.39 | 41.55 | 924.08 | 0.84   |
| GLMNet   | 196.30 | 8.37  | 3.97  | 9.1   | 280.20 | 0.69   |
| Random forest | 426.34 | 19.03 | 8.6   | 21.30 | 462.16 | 0.74   |
| XGBoost  | 699.0  | 30.30 | 14.14 | 40.66 | 912.89 | 0.85   |

Table 11: Accuracy results of confirmed COVID-19 cases in Sri Lanka
Results have proven that the ARIMA model is found to be suitable and ideal for forecasting COVID-19 cases in the mentioned countries. The random forest model did not fit to our dataset well so that it was excluded from the forecasting purpose. Model accuracy values were studied for individual countries. In Afghanistan, Bangladesh, India, the Maldives, and Sri Lanka, the ARIMA model was superior to other models. In Bhutan, the Prophet time series model was suitable for projection purposes. Still, in Bhutan, the total number of confirmed cases was 87, with no deaths. In Nepal and Pakistan, the generalized linear model elastic net model was more accurate than other models. This forecast study can help authorities to take timely actions and to control COVID-19 spread.

We are planning as a future research will explore the projection methodology and update the dataset using more accurate and suitable machine learning methods as well as statistical models for prediction.

Acknowledgements The authors would also like to thank the reviewers for their constructive comments which led to improve the presentation of this article.

Funding The authors also thank the financial support provided by the Deanship of Scientific Research, King Saud University. This research was partially funded by FONDECYT grant number 1200525 (V. Leiva) from the National Agency for Research and Development (ANID) of the Chilean government under the Ministry of Science, Technology, Knowledge, and Innovation.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest regarding the present study.

References

Akbar MA, Leiva V, Rafi S, Qadri SF, Mahmood S, Alsanad A (2022) Towards roadmap to implement blockchain in healthcare systems based on a maturity model. J Softw Evol Process. https://doi.org/10.1002/smr.2500

Alkadya W, ElBahnsy K, Leiva V, Gad W (2022) Classifying COVID-19 based on amino acids encoding with machine learning algorithms. Chemom Intell Lab Syst 224:104535

Alzahrani SI, Aljamaan IA, Al-Fakih EA (2020) Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions. J Infect Public Health 13:914–919

Anderson KM, Odell PM, Wilson PW, Kannel WB (1991) Cardiovascular disease risk probes. Am Heart J 121:293–298

Asri H, Mousannif H, Moatasim HA, Noel T (2016) Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Comput Sci 83:1064–1069

Bontempi G, Taieb SB, LeBorgne YA (2012) Machine learning strategies for time-series forecasting. Springer, New York, pp 62–77

Bradter U, Kunin WE, Altringham JD, Thom TJ, Benton TG (2013) Identifying appropriate spatial scales of predictors in species distribution models with the random forest algorithm. Methods Ecol Evol 4:167–174

Brockwell P, Davis R (1991) Time series: theory and methods. Springer, New York

Bustos N, Tello M, Droppelmann G, Garcia N, Fejoo F, Leiva V (2022) Machine learning techniques as an efficient alternative diagnostic tool for COVID-19 cases. Signa Vitae 18:23–33

Chahuan-Jimenez K, Rubilar R, de la Fuente-Mella H, Leiva V (2021) Breakpoint analysis for the COVID-19 pandemic and its effect on the stock markets. Entropy 23:100

Chaouch H, Charfeddine S, Aoun SB, Jerbi H, Leiva V (2022) Multiscale monitoring using machine learning methods: new methodology and an industrial application to a photovoltaic system. Mathematics 10:890

Chimmula VKR, Zhang L (2020) Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos Solitons Fractals 134:109864

De la Fuente-Mella H, Rubilar R, Chahuan-Jimenez K, Leiva V (2021) Modeling COVID-19 cases statistically and evaluating their effect on the economy of countries. Mathematics 9:1558

Fanelli D, Piazza F (2020) Analysis and forecast of COVID-19 spreading in China, Italy, and France. Chaos Solitons Fractals 134:109761

Fierro R, Leiva V, Balakrishnan N (2015) Statistical inference on a stochastic epidemic model. Commun Stat Simul Comput 44:2297–2314

Friedman J, Hastie T, Tibshirani R (2005) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33(1):1–22

Grasselli G, Pesenti A, Cecconi M (2020) Critical care utilization for the COVID-19 outbreak in Lombardy, Italy: early experience and forecast during an emergency response. J Am Med Assoc 323:1545

Harrell JFE, Lee KL, Matchar DB, Reichert TA (1985) Regression models for prognostic prediction: advantages, problems, and suggested solutions. Cancer Treat Rep 69:1071–1077

Hyndman RJ, Rob J, Anne BK (2006) Another look at measures of forecast accuracy. Int J Forecast 22:679–688

Hyndman RJ, Khandakar Y (2008) Automatic time series forecasting: the forecast package for R. J Stat Softw 27(3):1–22

Jerez-Lillo N, Lagos-Alvarez B, Muñoz-Gutiérrez J, Figueroa-Zuñiga J, Leiva V (2022) A statistical analysis for the epidemiological surveillance of COVID-19 in Chile. Signa Vitae 18:19–30

Jerome H, Friedman (2002) Stochastic gradient boosting. Comput Stat Data Anal 4:367–378

Kelleher JD, Mac Namee B, D’arcy A (2020) Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT Press, Cambridge

Lapuerta P, Azen SP, Labree L (1995) Use of neural networks in predicting the risk of coronary artery disease. Comput Biomed Res 28:38–52

Leiva V, Saulo H, Souza R, Aykroyd RG, Vila R (2021) A new BISARMA time series model for forecasting mortality using weather and particulate matter data. J Forecast 40:346–364

Liu Y, Mao C, Leiva V, Liu S, Silva Neto WA (2021) Asymmetric autoregressive models: statistical aspects and a financial application under COVID-19 pandemic. J Appl Stat 49:1323–1347

Ma L, Zhang Y, Leiva V, Liu S, Ma T (2022) A new clustering algorithm based on a radar scanning strategy with applications to machine learning data. Expert Syst Appl 191:116143

Mahdi E, Leiva V, Mara’Beh S, Martin-Barreiro C (2021) A new approach to predicting cryptocurrency returns based on the gold prices with support vector machines during the COVID-19 pandemic using sensor-related data. Sensors 21:6319

Malavika B, Marimuthu S, Joy M, Nadaraj A, Asirvatham ES, Jeyaseelan L (2021) Forecasting COVID-19 epidemic in India.
and high incidence states using SIR and logistic growth models. Clin Epidemiol Glob Health 9:26–33
Makridakis S, Spiliotis E, Assimakopoulos V (2018) Statistical and machine learning forecasting methods: concerns and ways forward. PLoS ONE 13:3
Luckner M, Topolski B, Mazurek M (2017) Application of XGBoost algorithm in fingerprinting localisation task. In: Saed K, Homenda W, Chaki R (eds) Computer information systems and industrial management. Lecture Notes in Computer Science, vol 10244. Springer, Cham, pp 661–671. https://doi.org/10.1007/978-3-319-59105-6-57
Martin-Barreiro C, Ramirez-Figueroa JA, Cabezas X, Leiva V, Galindo-Villardón MP (2021) Disjoint and functional principal component analysis for infected cases and deaths due to COVID-19 in South American countries with sensor-related data. Sensors 21:4094
Ospina R, Leite A, Ferraz C, Magalhaes A, Leiva V (2022) Data-driven tools for assessing and combating COVID-19 out-breaks based on analytics and statistical methods in Brazil. Signa Vitae 18:18–32
Petropoulos F, Makridakis S (2020) Forecasting the novel coronavirus COVID-19. PLoS ONE 15:e0231236
Prophet. Automatic forecasting procedure. (2020). https://github.com/facebook/prophet
Provost F, Hibert C, Malet JP, Stumpf A, Doubre C (2016) Automatic classification of endogenous seismic sources within a landslide body using random forest algorithm. EGU Gen Assem Conf Abstr 18:15705
R Development Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
Rangasamy M, Chesneau C, Martin-Barreiro C, Leiva V (2022) On a novel dynamics of SEIR epidemic models with a potential application to COVID-19. Symmetry 14:1436
Ritchie H, Mathieu E, Rodés-Guírao L, Appel C, Giattino C, Ortiz-Ospina E, Hasell J, Macdonald B, Beltekian D, Roser M (2020) Coronavirus pandemic (COVID-19). Our World in Data. https://ourworldindata.org/coronavirus
Rustam F, Reshi AA, Mehmoord A, Ullah SB, Aslam W et al (2020) COVID-19 future forecasting using supervised machine learning models. IEEE Access 8:101489–101499
Sardar I, Karakaya K, Makarovskikh T, Abotaleb M, Garda S, Mishra P (2021) Machine learning-based COVID-19 forecasting: impact on Pakistan stock exchange. Int J Agric Stat Sci 17:53–61
Sujath R, Chatterjee JM, Hassanien AE (2020) A machine learning forecasting model for COVID-19 pandemic in India. Stoch Env Res Risk Assess 4:959–972
Talabis MRM, McPherson R, Miyamoto I, Martin JL, Kaye D (2015) Analytics dened. In: Information security analytics: finding security insights, patterns and anomalies in big data. Syngress Books, Elsevier, Amsterdam, pp 1–12. https://doi.org/10.1016/B978-0-12-800207-0.00001-0
Taylor S, Letham B (2018) Forecasting at scale. Am Stat 72:37–45
Tianqi C, Guestrin C (2016) XGBoost: a scalable tree boosting system. arXiv:1603.02754
Ventura M, Saulo H, Leiva V, Monsueto S (2019) Log-symmetric regression models: Information criteria, application to movie business and industry data with economic implications. Appl Stoch Model Bus Ind 35:963–977
WHO (2020a) Naming the coronavirus disease (COVID-19) and the virus that causes it. World Health Organization, Geneva, Switzerland
WHO (2020b) Coronavirus disease 2019 (COVID-19). World Health Organization. Situation Report, 167, Geneva, Switzerland
Zhao J, Han M, Wang Z, Wan B (2022) Autoregressive count data modeling on mobility patterns to predict cases of COVID-19 infection. In: Stochastic environmental research and risk assessment (in press). https://doi.org/10.1007/s00477-022-02255-6
Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc B 67:301–320

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.