**THE COVID-19 CUMULATIVE CONFIRMED CASES FORECASTING IN SHORT-TERM MULTI-DAYS AHEAD PERSPECTIVE: THE BRAZILIAN CASE**

**ABSTRACT**

The new Coronavirus (COVID-19) is an emerging disease responsible for infecting millions of people since the first notification until nowadays. Developing efficient short-term forecasting models allows knowing the number of future cases. In this context, it is possible to develop strategic planning in the public health system to avoid deaths. In this paper, autoregressive integrated moving average, cubist, random forest (RF), ridge regression (RIDGE), support vector regression (SVR), and stacking-ensemble learning are evaluated in the task of forecasting one, three, and six-days ahead the COVID-19 cumulative confirmed cases in ten Brazilian states with a high daily incidence. In the stacking approach, the cubist, RF, RIDGE, and SVR models are adopted as base-learners, and gaussian process (GP) as meta-learner. The models' effectiveness is evaluated based on the improvement index, mean absolute error, and symmetric mean absolute percentage error criteria. In most of the cases, the GP reach a better performance regarding adopted criteria than compared models, as well as SVR regression, both with a linear kernel. In a general context, the developed models can generate accurate forecasting, achieving forecasting errors in a range of 0.97\% - 4.25\%, 1.02\% - 6.54\%, and 0.95\% - 6.90\% in one, three, and six-days ahead forecasting. The ranking of models in all scenarios is SVR, stacking ensemble, CUBIST, RIDGE, ARIMA, and RF models. The use of evaluated models is recommended to forecasting and monitor the ongoing growth of COVID-19 cases, once these models can assist the managers in the decision-making.

Keywords: COVID-19 \sep Machine learning \sep Statistical models \sep Forecasting \sep Decision-making \sep Time-series

**INTRODUCTION**

The new Coronavirus (COVID-19) is an emerging disease responsible for infecting millions of people and killing thousands worldwide since the first notification until nowadays, according to the World Health Organization (WHO) \cite{whoCOVID19,SOHRABI202071}. Also according to WHO, Brazil registered 40.581 confirmed cases until April 22th 2020, holding the 12th position in the world ranking in the number of confirmed cases of COVID-19, and 2nd position in the Americas (behind the United States only).

Due to the impacts of the COVID-19 pandemic in people’s lives and the world’s economy, the governments and population are most concerned with (i) when the COVID-19 outbreak will peak; (ii) how long the outbreak will last and (iii) how many people will eventually be infected \cite{zhang2020predicting}. Further, Boccaletti et al. \cite{boccaletti2020modeling} have identified at least three scientific communities that may cooperate in the effort to deal with the current pandemic: (1) the community of applied mathematicians, virologists and epidemiologists, developing sophisticated diffusion models to the specific properties of a given pathogen; (2) the community of complex systems scientists who study the spread of infections using compartmental models, using methods and principles from statistical mechanics and nonlinear dynamics; and (3) the community of scientists who incorporate artificial intelligence (AI) and deep learning to produce accurate predictive models. Also, different studies are evaluating the impacts of COVID-19 on society, whether through predictions of future cases, as well as variables capable of helping to understand the spread of this disease \cite{becerra2020forecasting, fanelli2020analysis,FONG2020106282,ROOSA2020256,EFFENBERGER2020}.

Moreover, epidemiological time series forecasting plays an important role in health public system, once it allows the managers to develop strategic planning to avoid possible epidemics. Forecasting diseases as accurate as possible is important due to their impact on the public health system. To ensure this accuracy, AI models have been widely used to forecast epidemiological time series over the years \cite{davis2019genetic, ribeiro2019forecasting,CHAKRABORTY2019121266}. Moreover, in the AI context, Vaishya et al. \cite{VAISHYA2020337} presented a review of trends in COVID-19 data analysis.

Regarding this context, the objective of this paper is to explore and comparing the predictive capacity of machine learning regression models and statistical models, in the task of forecasting one, three, and six-days ahead COVID-19 cumulative cases, in the Brazilian context. In this respect, datasets of ten Brazilian states with a high incidence of COVID-19 are adopted to evaluates the forecasting efficiency through of the autoregressive integrated moving average (ARIMA), cubist regression (CUBIST), random forest (RF), ridge regression (RIDGE), and support vector regression (SVR), and stacking-ensemble learning models. In the stacking modeling, which is an effective ensemble learning approach \cite{ribeiro2020ensemble, moreno2019very}, Cubist, RF, RIDGE, and SVR are used as base-learners (weak models), and gaussian process (GP) as meta-learner (strong model). The out-of-sample forecasting accuracy of each model is compared by the improvement percentage index (IP), mean absolute errors (MAE), and symmetric mean absolute percentage error (SMAPE).

The contributions of this paper can be summarized as follows:(i) The first contribution is related to the presentation of a novel analysis of the forecast model for cumulative confirmed cases of COVID-19 in Brazil, whose accuracy of the models assists governors in decision-making to contain the pandemic and strategies concerning the health system; (ii) As the second contribution, we can highlight the use of heterogeneous machine learning models, as well as the stacking-ensemble learning approach to forecast the Brazilian cumulative confirmed cases of COVID-19, (iii) Last, this paper evaluates models forecasting in a multi-day ahead forecasting strategy. The forecasting time horizons are the interval of one, three, and six-days ahead. This range of the forecasting time horizon allows us to verify the effectiveness of the predicting models in different scenarios, helping in future strategies in fighting COVID-19.

The remainder of this paper is organized as follows: Section~\ref{material} a brief description of the dataset adopted in this paper is given. The forecasting models applied in this study are described in Section~\ref{sub:BL}. Section~\ref{MET} details the procedures applied in the research methodology. Results obtained and related discussion about models’ forecasting performance are given on Section~\ref{RES}. Finally, Section~\ref{CONC} concludes this work with final considerations and some directions for future research proposals.

**MATERIAL AND METHODS**

This section presents the description of the material analyzed (Section \ref{material}) as well the models definition applied in this paper (Section \ref{sub:BL}).

**DATASET DESCRIPTION**

The collected dataset refers to the cumulative confirmed cases of COVID-19 that occurred in Brazil. The dataset was collected from an API \cite{brasilio} that retrieves the daily information about COVID-19 cases from all 27 Brazilian State Health Offices, gather them, and make it a publicly available. Among the 27 federative units (26 states and one federal district), ten states with high incidence of COVID-19 cases were chosen: Amazona (AM), Bahia (BA), Cear\'a (CE), Minas Gerais (MG), Paran\'a (PR), Rio de Janeiro (RJ), Rio Grande do Norte (RN), Rio Grande do Sul (RS), Santa Catarina (SC), and S\~ao Paulo (SP). The measurement period of each state varies, once each state counts since the day of its first case until the day of the last report. The cumulative confirmed cases and deaths of each state, as well as the dates of first and last reports, are illustrated in Table~\ref{tab:reports}. Further, a heatmap of the cumulative confirmed cases is presented in Figure~\ref{fig:heatmap}.

**TABLE 1**

**FIGURE 1**

**METHODS**

This section describes a brief of each model employed in the data analysis.

\begin{itemize}

\item ARIMA is a Box \& Jenkins modelling usually employed to deal with non-stationary time series. In fact, the ARIMA model is full specified by autoregressive (\textit{p}), different degrees of trend differences (\textit{d}), and moving average operators (\textit{q}). These parameters are used do define the model order, and usually defined by grid-search, as well as by autocorrelation and partial autocorrelation function. In this context, the model is described as ARIMA\textit{(p,d,q)} \cite{box2015time}.

\item CUBIST is a rule-based model, which performs predictions following the regression of trees principle \cite{Quinlan1993}. Through the use of a committee of the rules, and using the neighborhood concept similar to $k$-nearest-neighbor modeling, the final forecasting is obtained.

\item GP is composed of a set of random variables Gaussian distributed and fully specified by its mean and covariance (kernel) function \citep{Rasmussen2004}. In this paper, the GP with a linear kernel is adopted.

\item RIDGE is a regularized regression approach \cite{HOERL1970} which employs a penalization term in the ordinary least squares algorithm. It is an effective tool, once it reduces the bias of parameter estimates by controlling the standard errors. Moreover, the model can deal with inputs multicollinearity problem.

\item RF is a bagging ensemble-based model, which combines the bagging advantages characterized by the creation of multiple samples, with refitting through of the bootstrap technique, from the same set of data, and random selection of predictors to compose each node of the decision tree \cite{breiman2001random}. RF is a fast and robust supervised learning method able to deal with the randomness of the time series. Furthermore, it is interesting because, in addition to being an ensemble approach, only the number of predictors for each node needs to be tuned.

\item SVR consists in determining support vectors (points) close to a hyperplane that maximizes the margin between two-point classes obtained from the difference between the target value and a threshold. To deal with non-linear problems SVR takes into account kernel functions, which calculates the similarity between two observations. In this paper, the linear kernel is adopted. The main advantages of the use of SVR lies in its capacity to capture the predictor non-linearities and then use it to improve the forecasting cases. In the same direction, it is advantageous to employ this perspective in the case studies adopted, since the samples are small \cite{Drucker1997}.

\item Stacked Generalization or stacking is an ensemble-based approach \cite{wolpert1992stacked} which combines through a meta-learner the predictions of a set of weak models (base-learners) to obtain a stronger learner. This approach usually operates into two levels, where in the first level the base-learners are trained and its predictions are obtained. In the next stage, a meta-learner uses, as inputs, the predictions of the previous level in the training phase. The stacking predictions are obtained from meta-learner predictions. The main advantage of the stacking ensemble is that this approach can be improving the accuracy and additionally reducing error variance \cite{ribeiro2020ensemble}.

\end{itemize}

**PROPOSED MODELS FRAMEWORK FOR SHORT-TERM FORECASTING COVID-19 CASES**

This section describes the main steps adopted in the data analysis adopted by CUBIST, RF, RIDGE, SVR, and stacking models. Also, the ARIMA modeling is described.

\textbf{Step 1}: \label{step1} Firstly, the raw data is split into training and test datasets. The test dataset is composed of six least observations, and the training dataset by the remain samples \cite{ribeiro2020ensemble}. The training data are centered by its mean value and divided by its standard deviation. To develops multi-step ahead COVID-19 cases forecasting, recursive strategy is employed \cite{RODRIGUESMORENO2020112869}. In this aspect, one model is fitted for one-step ahead forecasting. In the following, the recursive strategy uses the forecasting value as an input for the same model to forecast the next step ahead, continuing in this manner until reaching the desirable horizon. In fact, in this respect, the training structure adopted in this paper is stated as follows,

**EQ 01**

\noindent in which $f$ is a function related to the adopted model in the training stage, $y\_{(t+1)}$ is the COVID-19 case one-day ahead, $n\_y = 5$ are the past confirmed cases, $\epsilon$ is the random error. In this paper, the aim is to obtain the cases up to $H$ next days, especially up to 1 (OSA), 3 (TSA), and 6-days ahead (SDA), respectively. The following structures are considered,

**EQ 02**

\noindent where $\hat{y}\_{(t+h)}$ is the forecast value at time $t$ and forecast horizon up to $h$, $y\_{t+h-n\_y}$ and $\hat{y}\_{t+h-n\_y}$ are the previously observed and forecast cases lags in $n\_y = 5$ days. The $n\_y$ value is chosen through grid-search with purpose to capture the best data behavior.

\textbf{Step 2}: In the stacking modeling, the base-learners CUBIST, RF, RIDGE, SVR are trained and its forecasting are used as inputs for meta-learner GP. In the training stage, leave-one-out cross-validation with a time slice is adopted \cite{ribeiro2020ensemble}. Finally, the out-of-sample forecasts are computed. These approaches are developed using the \texttt{caret} package \cite{caret2017}. The ARIMA modeling is performed through the use of \texttt{forecast} package \cite{forecastpackage,Rob2008} with use of \texttt{auto.arima} function. To define the ARIMA order, grid-search is adopted, and the most suitable order is that reach a lower Akaike and Bayesian Akaike criteria information. Both analyses are developed using \texttt{R} software \cite{R}. All hyperparameters employed in this study are presented in Table \ref{tab:hyper} in \ref{apendixB}.

\textbf{Step 3}: To evaluate the effectiveness of adopted models, from obtained forecasts out-of-sample (test set), performance IP, MAE, and sMAPE criteria are computed as defined in \eqref{eq:criteria}.

**EQ 03**

\noindent where $n$ is the number of observations, $y\_i$ and $\hat{y}\_{i}$ are the \textit{i}-th observed and predicted values, respectively. Also, the $M\_c$ and $M\_b$ represent the performance measure of compared and best models, respectively.

**RESULTS**

This section describes the results of the developed experiments in forecasts out-of-sample (test set). First, Section \ref{sec:E1} compares the results of evaluated models over ten datasets and three forecasting horizons adopted. In this respect, in Table \ref{tab:performancemeasure} in \ref{apendixA}, the best results regarding accuracy are presented in bold. Additionally, Figures \ref{fig:performance1} and \ref{fig:performance3} illustrate the relation between observed and predicted values achieved by models with best set of performance measures depicted in Table \ref{tab:performancemeasure}, as well box-plots for out-of-sample errors in Figure \ref{fig:Boxplot}.

**PERFORMANCE MEASURES FOR COMPARED MODELS**

In this section, the main results achieved by the best model regarding MAE and sMAPE criteria are presented for each state.

\begin{itemize}

\item AM: In this state, CUBIST, RIDGE, and SVR approaches could be considered to forecasting COVID-19 cases. In fact, in the context of ODA and TDA, CUBIST outperforms compared models, while for SDA, RIDGE achieves better accuracy regarding MAE and sMAPE than others. The improvement in the MAE for ODA and TDA achieved by CUBIST ranges between 6.58\% - 92.77\%, and 11.39\% - 88.54\%, respectively. Through sMAPE analysis, the RIDGE model outperforms other models, and this criterion is reduced in the range of 16.46\% - 91.88\%, for SDA horizon.

\item BA, MG, RS, and SP: For these set of states, in all forecasting windows, the SVR approach achieved better accuracy than other models, for both MAE and sMAPE criteria in the multi-days ahead forecasting task of the confirmed number of COVID-19. In fact, the improvement in MAPE is ranged in 6.62\% - 94.43\%, 1.55\% - 94.23\%, and 38.69\% - 94.69\%, in ODA, TDA, and SDA forecasting horizons. Moreover, the same behavior is observed when the improvement in the sMAPE criterion is obtained.

\item CE and RN: In the CE context, the CUBIST model have a better performance in the forecasting out-of-sample than other models. In this aspect, the for MAE criterion, the improvement is ranged between 7.26\% - 88.85\%, 13.15\% - 79.78\%, and 14.43\% - 89.62\%, for ODA, TDA, and SDA time windows, respectively. In its turn, for sMAPE, the improvement on ODA, TDA, and SDA horizons is 6.52\% - 90.15\%, 14.06\% - 83.35\%, and 17.59\% - 91.29\%, respectively. The SVR has results similar to CUBIST model. Considering the RN state, the same analysis is developed for ODA, and TDA horizons. The exception to the SDA horizon, in which the RIDGE model has better effectiveness in the MAE and sMAPE criteria than remain models.

\item PR, RJ, and SC: For these states, localized into the south (PR and SC) and southeast (RJ) of Brazil, the most appropriate approach to forecast COVID-19 is the stacking ensemble. It overcomes the drawback of single models and achieves the best accuracy than other models. In fact, for the three states, the improvement in MAE and sMAPE are ranged between 14.01\% - 95.41\%, and 17.48\% - 95.41\%, for ODA horizon. The improvement in the order forecasting horizons presents the same behavior of ODA, with the greatest magnitude of improvement for TDA and SDA.

\end{itemize}

\textbf{Remark:} In this experiment, 180 scenarios (10 datasets, 3 forecasting horizons, and 6 models) were evaluated for the task of forecasting COVID-19 cases. In a overview, the best models for each state have sMAPE ranged between 0.97\% - 4.25\%, 1.02\% - 6.54\%, and 0.95\% - 6.90\% to ODA, TDA, and SDA forecasting. The ranking of models in all scenarios is SVR, stacking ensemble, CUBIST, RIDGE, ARIMA, and RF models. In contrast to finds of \cite{BENVENUTO2020105340}, for the datasets evaluated in this paper, ARIMA modeling was not effective. From a broader perspective, the efficiency of SVR is due to its ability to deal with small sizes, while the stacking ensemble combines the advantages of several single models to learn the data behavior and obtain forecasts similar to observed values. On the other hand, the difficulty of the RF model to forecasting COVID-19 cases could be attributed to the fact that this approach requires more observations to effectively learn the data pattern.

According to the information depicted in Figures \ref{fig:performance1} and \ref{fig:performance3} it is possible to identify that the behavior of the data is learned by the evaluated models, which can forecasting compatible cases with the observed values. The good performance obtained in the training phase persists in the test stage. In the Figures \ref{fig:pred\_AM}, \ref{fig:pred\_MG}, and \ref{fig:pred\_RN} the models, CUBIST and SVR models presented difficulties to capture the variability of the first observations. However, from a certain moment on, the data pattern was learned. The sample sizes are reduced for all states, which justifies the difficulties of the models to learn the datasets' behavior.

**FIGURE 2**

**FIGURE 3**

Figure \ref{fig:Boxplot} shows the box-plots of out-of-sample forecasting errors in the SDA horizon for each model and dataset used. This horizon is chosen to analysis due to the recursive strategy adopted, once the errors increase according to the growth of the forecasting horizon. The box diagram depicts the variation of absolute errors for each model, which reflects the stability of each model. In this context, the dots out of boxes are considered outliers errors, and the black dot inside of the box is the MAE for each model.

**FIGURE 4**

Through the box-plot analysis, boxes with lower size indicate models with lower variation in the errors, and the results presented in Table \ref{tab:performancemeasure} are corroborated by the depicted in Figure \ref{fig:Boxplot}. Models with lower errors also reach better stability, which means that the most suitable modeling for each state can maintain a learning pattern, achieving homogeneous prediction errors.

**CONCLUSION AND FUTURE RESEARCH**

In this paper, six machine learning approaches named CUBIST, RF, RIDGE, SVR, and stacking ensemble, as well as ARIMA statistical model were employed in the task of forecasting one, three, and six-days ahead the COVID-19 cumulative confirmed cases in ten Brazilian states with a high daily incidence. The COVID-19 cumulative confirmed cases for AM, BA, CE, MG, PR, RJ, RN, RS, SC, and SP states were used. The IP, MAE, and sMAPE criteria were adopted to evaluate the performance of the compared approaches. Moreover, the stability of out-of-sample errors was evaluated through box-plots.

In respect of obtained results, it is possible to infer that SVR and stacking-ensemble learning model are suitable tools to forecast COVID-19 cases for most of the adopted states, once that these approaches were able to capt the non-linearities inherent to the evaluated epidemiological time series. Also, CUBIST and RIDGE models deserve attention for the development of this task. Therefore, the ranking of models in all scenarios is SVR, stacking ensemble, CUBIST, RIDGE, ARIMA, and RF models.

For future works, it is intended (i) to adopt deep learning approaches combined to stacking ensemble, (ii) to employ copulas functions for data augmentation and deal with for small samples, (iii) to use multi-objective optimization to tune hyperparameters of adopted models, (iv) to select stacking models using optimization approaches, (v) to adopt set of features which can help to explain the future cases o COVID-19.