DOI: https://doi.org/10.36489/saudecoletiva.2021v11i67p6851-6862

Forecasting COVID-19 cases in santa catarina cities with recurrent neural networks

Predicción de casos de COVID-19 en ciudades de santa catarina utilizando redes neurales recurrentes

Predição de casos de COVID-19 nos municípios de santa catarina utilizando redes neurais recorrentes

ABSTRACT
Objective: evaluate the forecasting of COVID-19 daily incidence in the cities of Santa Catarina, through a machine learning algorithm, within a time horizon of 14 days. Method: a recurrent neural network was applied to model a regression problem with a predictive purpose, using a retrospective longitudinal epidemiological study of COVID-19 cases in the analyzed cities. Results: the data model obtained with a machine learning algorithm presented an RMSE of 20.74, less than the baseline established through a persistence model. Conclusion: from the result achieved by the data model, it follows that the Artificial Intelligence tools used in the research are important instruments to face the COVID-19 pandemic, providing the management improvement of health resources, which need a suitable allocation for an adequate sanitary response to the disease progress.
DESCRIPTORS: Artificial Intelligence; Machine Learning; Coronavirus Infections; Public Health Surveillance; Epidemiology.

RESUMEN
Objetivo: el objetivo es evaluar la predicción de la incidencia diaria de COVID-19 en los municipios de Santa Catarina, mediante la ejecución de un algoritmo de aprendizaje automático, en un horizonte temporal de 14 días. Método: se utilizó una red neuronal recurrente para modelar un problema de regresión con propósito predictivo, mediante un estudio epidemiológico longitudinal retrospectivo de la incidencia de COVID-19 en los municipios analizados. Resultados: el modelo de datos obtenido mediante el algoritmo de aprendizaje automático presentó un RMSE de 20,74, inferior a la línea de base establecida mediante un modelo de persistencia. Conclusión: con el resultado alcanzado por el modelo de datos, se concluye que el uso de herramientas de Inteligencia Artificial permite obtener un instrumento importante para enfrentar la pandemia COVID-19, proporcionando una mejora en la gestión de los recursos de salud, los cuales deben ser adecuadamente asignados para una adecuada respuesta sanitaria al avance de la enfermedad.
DESCRIPTORES: Inteligencia Artificial; Aprendizaje Automático; Infecciones por Coronavirus; Vigilancia en Salud Pública; Epidemiología.

RESUMO
Objetivo: objetiva-se avaliar a predição da incidência diária de COVID-19 nos municípios de Santa Catarina, através de um algoritmo de aprendizagem de máquina, em um horizonte de tempo de 14 dias. Método: uma rede neural recorrente foi utilizada para modelar um problema de regressão com propósito preditivo, através de um estudo epidemiológico longitudinal retrospectivo da incidência de COVID-19 nos municípios analisados. Resultados: o modelo de dados obtido através do algoritmo de aprendizagem de máquina apresentou um RMSE de 20,74, menor do que a linha de base estabelecida através de um modelo de persistência. Conclusão: com o resultado alcançado pelo modelo de dados, conclui-se que emprego de ferramentas de Inteligência Artificial permite a obtenção de um importante instrumento para o enfrentamento da pandemia de COVID-19, proporcionando um aperfeiçoamento no gerenciamento de recursos de saúde, que precisam ser adequadamente alocados para uma resposta sanitária adequada ao avanço da doença.
DESCRITORES: Inteligência Artificial; Aprendizado de Máquina; Infecções por Coronavirus; Vigilância em Saúde Pública; Epidemiologia.

RECEIVED ON: 03/02/2021 **APPROVED ON:** 03/10/2021

Leonardo Silva Vianna
Dentist. Master by the Postgraduate Program in Health Informatics at the Federal University of Santa Catarina - PEN/ UFSC. Professor, Department of Dentistry, Centro Universitário Avantis - UNIAVAN. Balneário Camboriú (SC).
ORCID: 0000-0003-4947-0938

Juliano de Amorim Busana
Nurse. Doctoral student of the Postgraduate Program in Nursing at the Federal University of Santa Catarina - PEN/ UFSC. Professor at the Nursing Department at Centro Universitário Avantis - UNIAVAN. Balneário Camboriú (SC).
ORCID: 0000-0001-7004-2917

INTRODUCTION

Throughout human history, it has been possible to observe the existence of several pandemics, characterized by the confirmation of a rapid spread of a disease in a large part of the globe. Pandemics threaten the health of the entire population, forcing the State and companies to mobilize, imbued with the mission of mitigating the loss of human life and economic losses.[1] Among the many overwhelming pandemics that humanity has faced are the bubonic plague of the Middle Ages, the Spanish flu of the 20th century and the swine flu in the last decade. Even after SARS-CoV-1 in 2002 plagued the world with thousands of deaths, no vaccine has been developed, nor have drugs been created against the coronavirus. And as described by Paumgartten et al., [2] "The COVID-19 pandemic can be anything but unpredictable."

Although the current pandemic officially started in China in December 2019, in Brazil, the first official case was notified in March 2020. The delay made it possible for the Brazilian State to implement preventive policies, in view of the analysis of the development of the disease in other countries.[3] However, health managers, both state and local, have employed disparate strategies, some of which are in line with the recommendations of WHO and experts in the field and other strategies that are out of line with such guidelines. For the adequate coping with the COVID-19 pandemic, it is necessary to process the existing data on the disease, enabling its adequate epidemiological analysis and subsequent sharing of information. In this way, Health Surveillance systems can be improved to allow real-time monitoring of the disease. [4]

The evaluation of the dynamics of transmission of infectious diseases, using mathematical models obtained by data processing, can support the proper planning of health intervention strategies. The predictions made through the models allow the adequate allocation of resources, avoiding redundancy in the distribution of health professionals, as well as they can improve the effectiveness of vaccination campaigns. [5]. Segundo Bontempi et al., [6] the data processing, performed through machine learning algorithms, allows the prediction of the occurrence through the analysis of historical data, enabling the understanding of events in different areas of knowledge. And in order to make this possible, the algorithms are used to solve nonlinear and nonparametric problems, which classical statistical tests cannot solve. The prediction of epidemiological information improves interventions in the peaks of disease occurrence, which can be obtained through the application of machine learning algorithms for the analysis of complex temporal patterns. [7]

Throughout human history, it has been possible to observe the existence of several pandemics, characterized by the confirmation of a rapid spread of a disease in a large part of the globe.

Thus, the objective of the present study is to evaluate the prediction of the daily incidence of COVID-19 in the municipalities of Santa Catarina, through the execution of a data modeling with a machine learning algorithm, in a time horizon of 14 days. Determining an appropriate approach for building a data model allows obtaining an instrument that can improve the capacity of public managers and administrators of health institutions, for the management of health resources in facing the pandemic.

METHOD

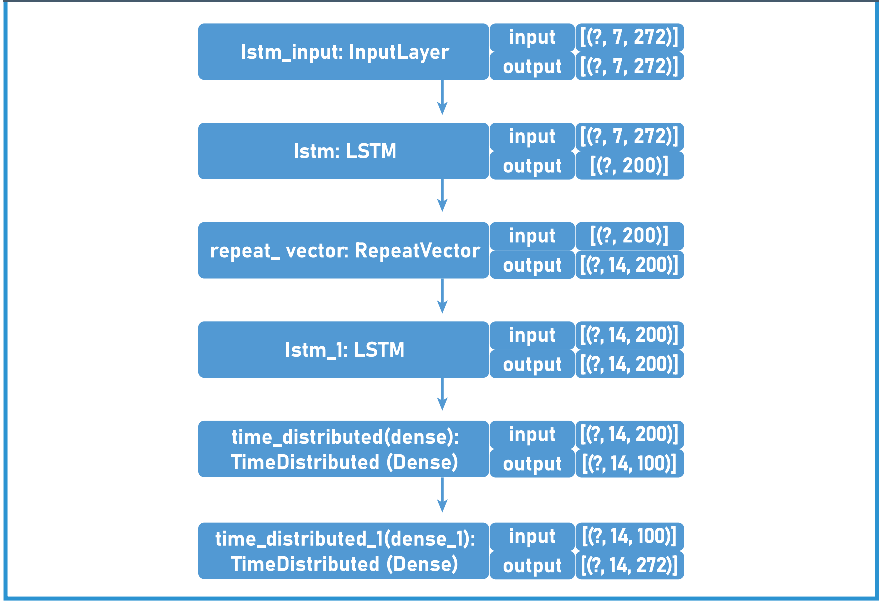
From the research objective, a longitudinal epidemiological retrospective study of the incidence of COVID-19 in the municipalities of the State of Santa Catarina was delimited. The period for the dissemination of the disease in the territory, considered in the survey, started on March 9th, 2020 (date of registration of the first tests with a positive result) until January 24th, 2021. This time interval was determined in order to adjust weekly, totaling 46 weeks.

The case records of COVID-19 were obtained from the data repository maintained by the State of Santa Catarina. On January 26th, 2021, the file was downloaded from the Open Data Portal, which, according to the information presented on the website itself, is an official open government data platform, with the purpose of improving transparency and social control, integrating information from the e-SUS Epidemiological Surveillance system and the Influenza Epidemiological Surveillance Information System (SIVEP Gripe - Sistema de Informação de Vigilância Epidemiológica da Gripe). There was also information that the file in the standard comma separated values (CSV) had already been processed and the information available was anonymized. [8] Thus, in order to allow its wide and unrestricted dissemination, the information contained in the records made it impossible for any type of patient identification. The obtained file contained 564.163 records, with information on patients' gender, age and municipality of residence, including clinical data and symptoms presented, data on the diagnostic method, among others. Considering the information in the records, the only inclusion criterion adopted was the date of the test result with confirmation of a positive case of the disease, according to the period considered in the research. Considering the municipality of residence informed in the existing records in the data set, the records with patients residing in other states of the country were adopted as exclusion criteria.

In the preparation (or pre-processing) stage, the data were tabulated in order to represent the incidence of COVID-19 cases in each municipality, by date of the result of the disease confirmation test. The preparation of the data produced a time series with daily periodicity, in the different municipalities of Santa Catarina. Starting from the premise of the existence of an epidemiological interrelation, all municipalities were arranged in a table (or data set), but in different columns, in order to allow an integrated data modeling that considered this connection.

With the tabulated data set, an autocorrelation analysis of the time series data was also performed using full and partial autocorrelation functions. The heuristic analysis of all this information made it possible to determine the phases following the preparation of the data. To allow an adequate evaluation of the data model obtained, 30% of the time series was segregated as a set of test data. In time series modeling, the relationship between the instances needs to be preserved by maintaining the sequence of data disposition; thus, the test set referred to the latest data available in the tabulation previously made. The test data were not used in the modeling processes, remaining reserved only for validation of the models obtained.

Graph 1. Representation of the recurrent neural network used for data modeling.



Source: Elaborated by the authors, 2021.

Then, data modeling and model evaluation were performed using the walk-forward validation method, using the sliding window process with the size of the data entry window (steps in) defined in 7 days. A grid search method was applied to adjust the hyperparameters of the learning algorithm used - epochs= 300, batch= 8 and nodes= 200 (except for the last intermediate layer that had half the amount of artificial neurons). The model's output data were predictions over the 14-day time horizon, considering the dependent variables used to analyze the quality of the data model through an evaluation metric.

The data modeling processes were performed using the Keras [9] library, executed over a TensorFlow API, [10] in a Python programming environment. All stages of the research, including the pre-processing of data, the construction of graphs and the analyzes applied to the models obtained, were also performed in the same programming environment. The source code was based on a computer program registered with the National Institute of Industrial Property (INPI), under registration BR n ° 512021000139-7.[11]

Data modeling was performed using a recurrent neural network, using multiple Long Short-Term Memory (LSTM) layers, which have the ability to process multivariate data sets. This configuration consisted of an encoder-decoder model - composed of two submodels: an encoder and a decoder -, proposed by Malhotra et al. [12] with some modifications suggested by Brownlee. [13] The recurrent neural network applied in the data processing is represented in Graph 1. To initialize the weight matrices, the function developed by Glorot [14] was applied, with the exception of bias units, in which random functions were applied. Both functions were configured to obtain a matrix of weights with a normal distribution.

The root mean squared error (RMSE) assessment metric - in Portuguese, square root of the mean square error - was used to evaluate the models obtained by comparing the output data of the processing and the existing real data, using the formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} \square\, (p_t - a_t)^2}$$

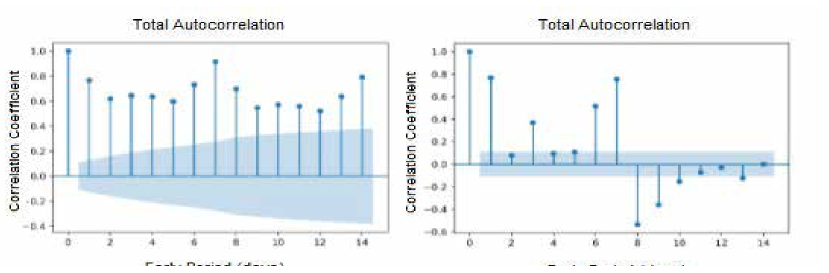
where, pt is the predicted value and at is the current value, at a given time t, for a data set of size n.

The data predicted by the model were compared with the real data through a graph, considering the set of samples obtained by the walk-forward validation method, on each day of the prediction horizon. The results obtained were compared with a persistence model - which considers the value of the position t+1 equal to that of the position t -, which is usually used as a baseline for the performance of data models built by machine learning algorithms.

RESULTS

With the tabulation of the file collected in the repository of the State of Santa Catarina, the functions of full and partial autocorrelation confirmed the existence of a relationship in the temporal sequence, which allowed to infer about the possibility of predicting the epidemiological progression of the disease. The application of the total autocorrelation function demonstrated statistically significant results (95% confidence interval) in all 14 previous days, with the highest result on the seventh day. The partial autocorrelation showed significant results in different periods of time, being one of the most relevant ones that also occurred in 7 previous days; which, in the end, was established as the best data entry window for the model. Graph 2 presents the results of the application of the partial and total autocorrelation functions.

Graph 2. Result of the total and partial autocorrelation functions, in the tabulated data set.



Source: Elaborated by the authors, 2021.

Table 1.Statistical summary of the results of the data model, obtained by the machine learning algorithm, and of the persistence model.

	DATA MODEL	PERSISTENCE MODEL
σ	20,74	24,76
σ	5,16	6,85
min	14,39	11,76
Q1	16,70	19,51
Q2	19,06	24,38
Q3	24,30	28,10
max	31,74	43,94

Source: Elaborated by the authors, 2021.

According to the research configuration, the recurrent LSTM neural network employed produced a 14-day prediction sequence for each 7-day data entry window. In this way, the RMSE was calculated in each of the output sequences, between the data predicted by the model and the actual data from the test data set. Considering all the modeling cycles performed by the walk-forward validation process, an average RMSE of 20,74 was obtained. The persistence model, built to enable comparison by establishing a baseline, obtained an average RMSE of 24,76. Table 1 presents the summary statistical information of the results of both models, also showing results less dispersed in the data model obtained (standard deviation of 5,16) than in the persistence model (standard deviation of 6,85).

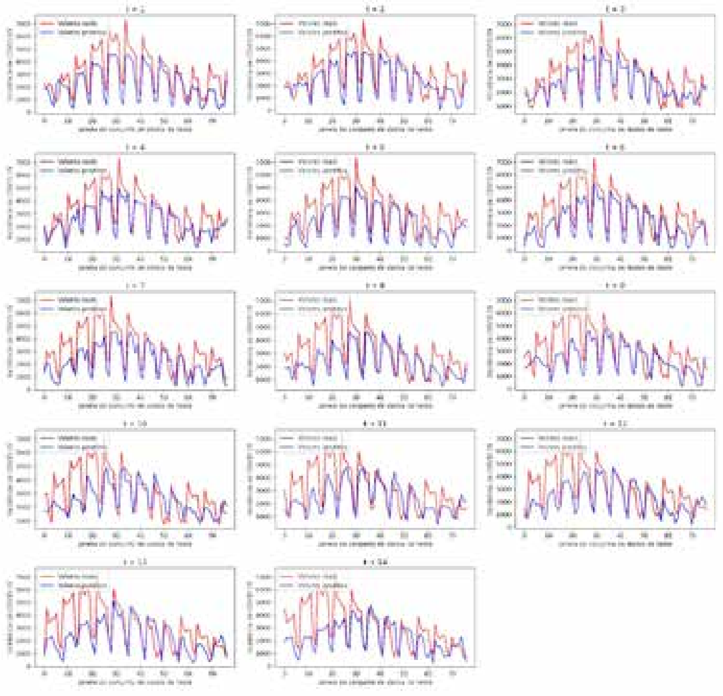
It is also possible to analyze the results by observing Graph 3, which presents the data outputs of the data model, in all windows produced by the walk-forward validation process in the test data set. Each of the 14 graphs represents all the consolidated results for a given period of the prediction time horizon (t+1, t+2..., t+14). It is important to note that the individual graphs do not express a temporal sequence, but rather samples of consolidated data; however, this format makes it possible to compare the actual values of the test data set and the values predicted by the data model. Analyzing the results displayed in the graph, it is possible to notice a smaller discrepancy between the actual values and predicted values in the time period of one day (t+1) than the graph of the last day of the time horizon (t+14).

DISCUSSION

Data tabulated in time series are used in different areas of knowledge, but widely used in epidemiological studies. Time series can be defined as a sequence of data recorded over a regular period of time, such as, for example, the incidence of a disease.[15] When arranged in Cartesian charts, time series analysis is traditionally used in epidemiological studies, being also a provider of important information and allowing the construction of useful inferences. But when it is necessary to establish interrelationships between different variables, other methods for the disclosure of this intrinsic knowledge must be applied. In this context, machine learning algorithms, based on the foundations of Artificial Intelligence, stand out when compared to the heuristic method.

Considering the limited period of time analyzed in the research, restricted to the recent months of the COVID-19 pandemic, it was necessary to consider some concepts that would provide adequate modeling of the data used in the research. As reported by Mussumeci and Coelho, (16) the use of multivariate data sets - in this case, characterized by an epidemiological series in different municipalities - allows us to explore the epidemiological concept of similarity. Consequently, the reported limitation was offset by the breadth of the analyzed municipalities and the approach applied to the data pre-processing processes. In another aspect, the premise of the existence of an epidemiological interrelation was decisive for the use of recurrent neural networks used to model the data in this research, as they allow the data to be entered in the appropriate format to model the epidemiological similarity, as can be seen. observed in the input (input) of the first layer of the recurrent neural network presented in Graph 1 (with 272 municipalities that had reported cases of COVID-19).

Graph 3. Actual values (red) and predicted values (blue) of the daily incidence of COVID-19, in different periods of the prediction time horizon.



Source: Elaborated by the authors, 2021.

Furthermore, recurrent neural networks have the ability to store the representation of their previous connections, applying the LSTM model developed by Hochreiter and Schmidhuber. [17] This architecture allows to model time intervals through a constant error flow in the calculation units. Networks with LSTM layers provide the learning of long-term correlations in a sequence, showing better results than conventional neural networks. Thus, mainly in the modeling of complex multivariate sequences, LSTM networks present more accurate results. [12] And although the results presented consider the entire data set, the models obtained performed predictions of the cases of COVID-19 in each municipality, separately. Consequently, the data model can be applied for individualized prediction, even if a longer period of time is used. This aspect should be considered important, as this research can produce a useful tool in the planning of health intervention during the COVID-19 pandemic.

In relation to the results obtained, the tendency to increase the RMSE is an expected behavior due to the accumulation of error over the prediction horizon, as the model generates predictions that are more distant from the data entry window. The use of recurrent neural networks sought to minimize this problem, through the constant flow of error calculation in the LSTM layers. In another aspect, the lack of a seasonal cycle in the evolution of the RMSE, over the prediction horizon, allows us to state that it was possible to model the weekly seasonality of the analyzed time series.

The use of machine learning algorithms mostly disregards the need for decomposing time series, which help to model these types of data; however, future studies that apply the process may improve the results of this research.

An inherent limitation in the design of the proposed research was the absence of segregation of records, considering information such as, for example, age, gender, and the presence of comorbidities; or, in other respects, the occurrence of hospitalization, including in an intensive care unit, type of diagnostic exam performed, among other information contained in the records obtained. Segregation can lead to a better analysis of existing biases in the analyzed data and be addressed in future epidemiological studies that aim to analyze these characteristics.

CONCLUSION

The result achieved by the data model, obtained through the application of a machine learning algorithm, presented an average error less than the baseline established through a persistence model. Consequently, it is concluded that the data modeling performed allowed the prediction of the daily incidence of COVID-19 in Santa Catarina municipalities, within a 14-day prediction time horizon. Considering all the reported aspects, the evaluation of the prediction of COVID-19 cases made it possible to infer that the use of Artificial Intelligence tools provides the obtaining of an important instrument to face the COVID-19 pandemic. Its application can improve the capacity of public managers and administrators of health institutions in the management of resources, which need to be appropriately allocated for an adequate sanitary response to the progress of the disease. ■

REFERENCES

1. Freire-Silva J, dos Santos FH, Candeias ALB, Pinho MAB, Oliveira BRB. A utilização do planejamento territorial no combate da COVID-19: considerações sobre a situação dos leitos nos municípios de Pernambuco, Brasil. Vigil sanit debate. 2020;8(2):16-27. doi:10.22239/2317-269x.01546

2. Paumgartten FJR, Delgado IF, Rocha PL, de Oliveira ACAX. Drug repurposing clinical trials in the search for life-saving COVID-19 therapies. Vigil sanit debate. 2020;8(2):39-53. doi:10.22239/2317-269x.01596

3. Ministério da Saúde. O que é coronavírus? (Covid-19). Brasília, DF: Ministério da Saúde; 2020[acesso 16 ago 2020]. Disponível em: https://coronavirus.saude.gov.br

4. Lopes-Júnior LC. A Saúde Coletiva no epicentro da pandemia de COVID-19 no Sistema Único de Saúde. Saúde Colet. (Barueri). 2020;10(56):3080-9. doi:10.36489/saudecoletiva.2020v10i56p3080-3089

5. Bistrian DA, Dimitriu G, Navon IM. Processing epidemiological data using dynamic mode decomposition method. In: AIP Conference Proceedings; 2019 Jun 20-25; Albena, Bulgaria; Maryland: AIP Publishing LLC; 2019. doi:10.1063/1.5130825

6. Bontempi G, Taieb SB, Le Borgne Y. Machine learning strategies for time series forecasting. In: 2nd European Summer School Business Intelligence, eBISS; 2012 Oct 15-21; Brussels, Belgium; New York: Springer; 2012. doi:10.1007/978-3-642-36318-4_3

7. Wu Y, Yang Y, Nishiura H, Saitoh M. Deep learning for epidemiological predictions. In: 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018. 2018 Jul 8-12; Ann Arbor, United States. New York: Association for Computing Machinery. doi:10.1145/3209978.3210077

8. Controladoria-Geral do Estado de Santa Catarina. Portal de Dados Abertos do Estado de Santa Catarina. COVID-19 - Casos Confirmados. 2020[acesso em 14 ago]. Disponível em: http://dados.sc.gov.br/dataset/covid-19-dados-anonimizados-de-casos-confirmados

9. Chollet F, et al. Keras. 2015[acesso 20 mar 2020]. Disponível em: https://keras.io

10. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-scale machine learning on heterogeneous systems. [acesso 14 ago 2020]. Disponível em: https://www.tensorflow.org

11. Vianna, LS. Sistema de predição de casos de COVID-19 nos municípios de Santa Catarina. BR n° 512021000139-7. Concessão: 25 de fev. 2021.

12. Malhotra P, Vig L, Shroff G, Agarwal P. Long short term memory networks for anomaly detection in time series. In: 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2015; 2015 Apr 22-24; Bruges, Belgium.

13. Brownlee, J. Deep learning for time series forecasting: Predict the future with MLPs, CNNs and LSTMs in Python. Vermont: Machine Learning Mastery; 2018.

14. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings; 2010 May 13-15; Sardinia, Italy.

15. Bhaskaran K, Gasparrini A, Hajat S, Smeeth L, Armstrong B. Time series regression studies in environmental epidemiology. Int. J. Epidemiol. 2013;42(4):1187-95. doi:10.1093/ije/dyt092

16. Mussumeci E, Coelho FC. Machine-learning forecasting for Dengue epidemics-Comparing LSTM, Random Forest and Lasso regression. medRxiv. 2020[acesso em 20 mar 2020]. Disponível em: https://www.medrxiv.org/content/10.1101/2020.01.23.20018556v1. doi:10.1101/2020.01.23.20018556

17. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8): 1735-80. doi:10.1162/neco.1997.9.8.1735