 doi:10.20944/preprints202009.0228.v1

Type: Journal Article

Forecasting COVID 19 Confirmed Cases Using Machine Learning: the Case of America

Mario Jojoa[1,*], Begoña Garcia-Zapirain[2]

[1] e-VIDA Laboratory University of Deusto, Bilbao, 48007, Spain
[2]e-VIDA Laboratory University of Deusto, Bilbao, 48007, Spain
[*]Corresponding Author: Mario Jojoa. Email: mariojojoa@deusto.es
Received: 08 September 2020; Accepted: XX Month 202X

Abstract: This paper presents a Multilayer Perceptron and Support Vector Machine algorithms approach to predict the number of COVID19 infections in different countries of America. It intends to serve as a tool for decision-making and tackling the pandemic that the world is currently facing. The models were trained and tested using open data from the European Union repository where a time series of confirmed contagious cases was modeled until May 25, 2020. The hyperparameters as number of neurons per layer were set up using a tabu list algorithm. The countries selected to carry out the study were Brazil, Chile, Colombia, Mexico, Peru and the United States. The metrics used are Pearson's correlation coefficient (CP), Mean Absolute Error (MAE), and Mean Percentage Error (MPE). For the testing stage we obtained the following results: Brazil, CP=0.65, MAE=2508 and MPE=17%; Chile, CP=0.64, MAE=504, MPE=16%; Colombia, CP=0.83, MAE=76, MPE=9%; Mexico, CP=0.77, MAE=231, MPE=9%; Peru, CP=0.76, MAE=686, MPE=18% and the United States of America, CP=0.93, MAE=799, MPE=4%. This resulted in powerful machine learning tools although it is necessary to use specific algorithms depending on the data and the stage of the country's pandemic.

Keywords: Multilayer Perceptron, Support Vector Machine, COVID19, SarsCov2, Forecasting, Machine Learning, Public Health, Pandemic.

1 Introduction

The pandemic caused by the SARS-COV2 [1] virus is a concern for governments across the world. By studying its behavior, we can save millions of lives as authorities can use this information to prepare and face the crisis. We know that the effect of the pandemic caused by this virus is a result of the weakness of health systems in countries worldwide. Everyday thousands of COVID19 [2] patients arrive needing immediate assistance, which is often not available. Hence, it is necessary to build a technology tool that allows to predict the number of infected people and anticipate the worst scenarios, among which we can mention closure of the economy and collapse of the health system. On the other hand, the impact depends on the characteristics of the affected population, such as their social contacts, personal economic and educational levels, and the government resources to face the crisis [3]. These different factors mean that there is no single contagion model since information is scarce and the cost of measurement is still expensive for developing economies. The latter directly justifies our work because it takes advantage of the data and information from previous experiences to accurately predict the number of infections as a time series before reaching the peak of the curve that models the behavior of the infection. We therefore propose a machine learning tool as an inexpensive and reliable prediction solution that allows decision-making in a fast and timely manner.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 September 2020 doi:10.20944/preprints202009.0228.v1

In this work we propose the use of a Multilayer Perceptron and a Vector Support Machine [4] to predict the time series of the number of infected cases per day as an intelligent tool to make decisions in terms of public health strategies in countries of America [5]. These models were trained with data previously collected for each country and were tested for the last ten days until 25th of May 2020. The adjustment of the hyperparameters was carried out in a different manner for each algorithm and country. The following sections of this paper are: 2) the Materials and Methods section which describes how we built and trained the models and adjusted the hyperparameters. 3) the Results section which shows the comparison of the performance of the models for each country and the box plot charts to show how the distribution of the predicted data corresponds with the testing data. In 4) the Discussion section, the related work is described and compared with state of the art, exploring whether it is aligned with our proposal, and in section 5 the conclusions are presented.

2 Materials and Methods

We propose a method based on machine learning to make the prediction of possible confirmed cases of Covid19 in six different American countries, which could be used for planning in the containment stage of the pandemic. We propose two classic regression models [6] to perform this task and we compare their performance. The proposed models are Multilayer Perceptron and Support Vector Machine. It is most important to highlight that we used these machine learning algorithms because the amount of available data is small [7], on average 78 registers per country, which correspond to the accumulated confirmed cases day by day during the evolution of the pandemic from the start date of measurement to May 25. The data used for this work are public and were downloaded from the open data portal of the European Union [8].

2.1 Data Description

The COVID19 database was downloaded on May 25th and was composed of 11 fields: 1. dateRep, 2. day, 3. month, 4. year, 5. cases, 6. deaths, 7. CountriesAndTerritories, 8. geold, 9. CountryterritoryCode, 10. PopData2018 and 11. ContinentExp. To carry out our experiments we used fields 1, 3, 5 and 7, enabling us to obtain the time series evolution of SarsCov2 confirmed cases for the countries under study. The full database contains 19,248 records.

Table 1. Data description

Field	Description
dateRep	Corresponds to the registration date
day	Indicates the corresponding day within the sequence
month	Indicates the month of registration
year	Indicates the year of registration
cases	Indicates the number of confirmed cases of COVID19
deaths	Indicates the number of deaths from COVID19
CountriesAndTerritories	Indicates the country and territory of the data source
geold	Geo-referencing identifier of the territory
CountryterritoryCode	Identifier of the country and territory source of the data
PopData2018	Population in 2018
ContinentExp	Continent Identifier

2.2 Proposed models

To carry out the COVID19 forecast we proposed two machine learning regression algorithms:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 September 2020 doi:10.20944/preprints202009.0228.v1

Multilayer Perceptron MLP and Support Vector Machine SVM. They were trained and tested using the same data and their hyperparameters were adjusted using state of the art information and hill climbing tabu list optimization algorithm.

2.2.1 Multilayer Perceptron

We built a feed forward backpropagation network with two hidden layers, five neurons in the input layer and one neuron in the output. The number of n neurons in the first hidden layer and the number of m neurons in the second hidden layer are hyperparameters which were selected using an optimization algorithm. The Figure 1 shows the general structure of the network used [9].

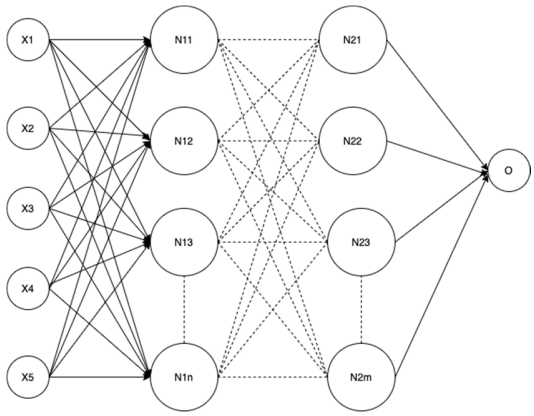

Figure 1. Multilayer Perceptron structure

We can describe this functional structure using a feedforward propagation formula as shown in Eq. (1) and Eq. (2) equations

$$o_1 = Relu({w_1}^T x) \quad (1)$$

$$o = Relu({w_2}^T o_1) \quad (2)$$

where w_1 and w_2 are the weights matrix of the layers respectively and their dimensions depend on the n and m values. On the other hand, we used Relu activation function in the two layers. All weights in the structure were adjusted using an optimization (ADAM) algorithm to minimize the cost function.

2.2.2 Support Vector Machine

We proposed the support vector machine as a model of regression, which can be found in the literature as Support Vector Regression. This model based its performance on the use of support vectors of the data to trace a multiple hyperplane to obtain zones where the data correspond. The Figure 2 shows the hyperplane examples for this application with two features with a linear kernel [10].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 September 2020 doi:10.20944/preprints202009.0228.v1

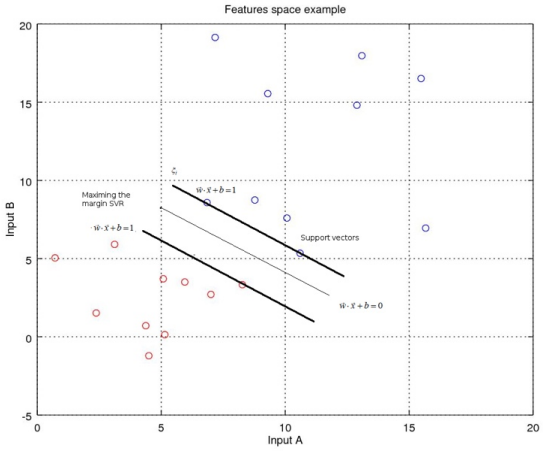

Figure 2. Support Vector Machine

The Support Vector Machine bases its functioning on achieving the maximum margin between the hyperplanes built using the vectors from the data, which is the reason for the name of the model. On the other hand, the problem is reduced to an optimization problem where Eq. (3) is the objective function and Eq. (4) are the constraints.

$$\min \frac{1}{2}||w||^2 + L_1|L_2 \tag{3}$$

$$y_i(wx_i + b) > 1 - \zeta_i \; \forall \; x_i; \; \zeta_i > 0 \tag{4}$$

2.2.3 Adjusting the hyperparameters

For the Multilayer Perceptron, we adjusted the hyperparameters in two ways: general hyperparameters were adjusted for all data on all countries and specific hyperparameters were adjusted for data on a specific country. The first ones were static and were adjusted based on state of the art [11]. The second ones were adjusted using hill climbing – tabu list algorithm [4-12]. Table 2 shows the hyperparameters for this machine learning structure.

Table 2. Hyperparameters adjusted for the MLP model

General Hyperparameters	Value	Specific adjusted Hyperparameters	Value
Training Algorithm	ADAM	Neurons Layer 1 (n)	Range [1 - 20]
Regularization	L2	Neurons Layer 2 (m)	Range [0 - 20]
Initital learning Rate	0.01	Number of Layers	Range [1 - 2]
Activation Function	RELU		

The values n and m were adjusted using hill climbing – tabu list optimization algorithm [13]. As we can see, Figure 3 shows the convergence of the algorithm and the number of steps taken to get the best value in terms of the metric MAE [14]. We selected this metric to be optimized, because the deviation of infected cases per day is the most relevant.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 September 2020 doi:10.20944/preprints202009.0228.v1

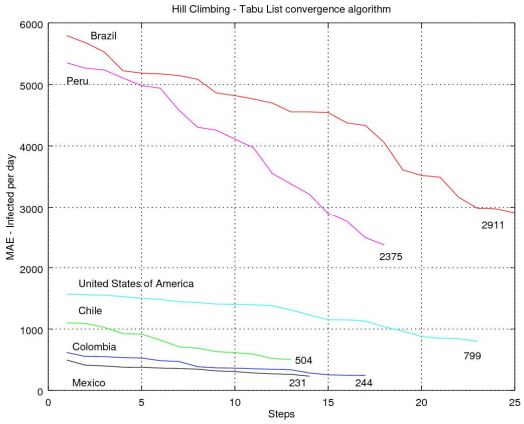

Figure 3. Convergence hill climbing – tabu list algorithm

In the case of the Support Vector Machine, we adjusted the hyperparameters based on the methods proposed in [15]. Table 3 shows the hyperparameters evaluated for each combination using the grid method.

Table 3. Hyperparameters and values for SVM model

General Hyperparameters	Values
Kernel	[Linear, Polynomial, RBF]
Regularization	[L1, L2]
Gamma	[0.0, 0.25, 0.5, 0.75, 1]

2.2.4 Training and testing the models

To carry out the training and testing of the proposed regression models, we formed matrices respectively, in such a way that the rows correspond to the last 5 data of the time series and the columns to the progress, day by day, of the confirmed cases series of COVID-19. In order to do so, we created a sliding window with a length of 5s and took the last 5 data as inputs and the current data as target. Thus, we repeated this task until we reached a point of division between the training and test data. The percentages we used were 80% and 20% respectively. The Figure 4 describes this process.

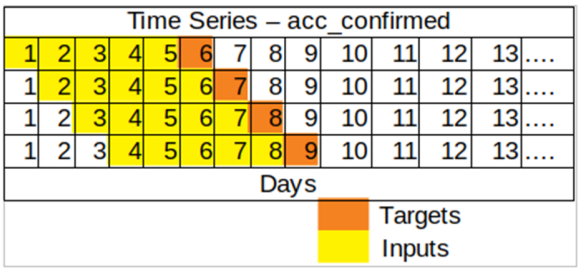

Figure 4. Training and Testing Matrix Conformation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 September 2020 doi:10.20944/preprints202009.0228.v1

2.2.5 Proposed System

The proposed system consists mainly of five parts: the inputs, the predictor module, the accumulation stage, the performance measurement stage and the automatic adjustment and initialization stage. The Figure 5 shows the model of the proposed solution.

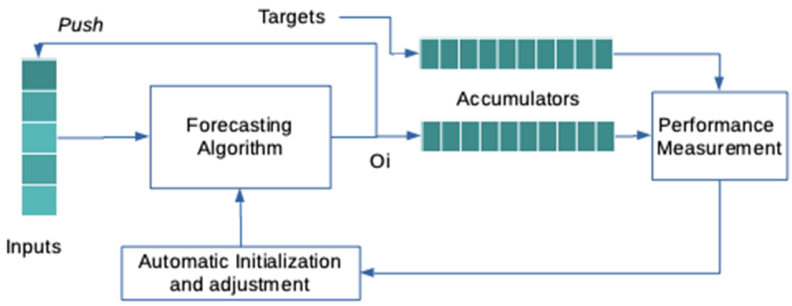

Figure 5. Proposed system for forecasting confirmed Covid 19 cases.

Inputs: In this stage we organized the inputs in such a way that the prediction was made with the amount of people infected in the last 5 days of the time series for each country. It is very important to note that the inputs are moved and discarded as the output is fed back as a new entry, thus predicting a sequence of values. This is how we forecasted the time series.

Forecasting Algorithm: We selected the prediction algorithm based on its performance. For the proposed system the two possible algorithms are the Multi-layer Perceptron and the Vector Support Machine. These regression models of the time series are appropriate for this application and each one was chosen based on the best performance for each country, since it was not possible to obtain a best-performance general regression model for all countries.

Accumulators: In order to measure the performance of the applied algorithms, we accumulated the prediction of 10 consecutive days and compared them with the corresponding values contained in the open data repository of the European Union from May 16 to May 25, 2020.

Performance Measurement: At this stage, we measured the performance of the two proposed algorithms for the same data set. We used the 3 evaluation metrics described in the next section. The values obtained here were used as an input for the hyperparameter optimization algorithm. In addition, these performance measures were the selection criteria for the model to be used for a specific country data set.

Automatic Initialization and adjustment: It was important that the models to be used were correctly tuned. To achieve this, we carried out an optimization algorithm that automatically initializes the optimal hyperparameter values. This ensures that the selected model is the best one to perform the task of regressing the number of SarsCov-2 infections. For this work, the selected optimization algorithm is hill climbing - tabu list. On the other hand, the random weights are initialized under a normal distribution of mean 0 and variance 1.

2.3 Performance Metrics

To measure the forecasting performance of our proposed models we used the metrics presented in Eq.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 September 2020 doi:10.20944/preprints202009.0228.v1

(5) Pearson Correlation Coefficient, Eq. (6) Mean Absolute Error and Eq. (7) Mean Percentage Error.

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{5}$$

$$MAE = \frac{\sum_i^n |e_i|}{n} \tag{6}$$

$$MPE = \frac{100}{n} \sum_i^n \frac{|real - predicted|}{real} \tag{7}$$

3. Results

The results obtained in the adjustment, training and testing stages of the proposed models are shown below. Table 4 shows the adjusted hyperparameters and their values for each of the models. On the other hand, in Table 5 we highlight in bold the best performances obtained for each country, showing better results for Chile, Mexico and the USA using the Multilayer Perceptron model, and for Brazil, Colombia and Peru with the Support Vector Machine model. The largest Mean Percentage Error was found for Peru.

Table 4. Multilayer Perceptron & Support Vector Machine used hyperparameters.

Country	Multilayer Perceptron (Neurons per Layer) Hyperparameters	Training Start Point Day	Support Vector Machine Hyperparameters
Brazil	[18:19]	20	Kernel = Linear Gamma = 0.25 Regularization L2
Chile	[17:18]	25	
Colombia	[18:19]	25	
Mexico	[18:19]	20	
Peru	[19:19]	20	
USA	[19:19]	20	

Table 5. Performance of the algorithms in forecasting confirmed cases of COVID19

	Multilayer Perceptron			Support Vector Machine		
Country	Pearson Correlation Coefficient	Mean Absolute Error (Infected per day)	Mean Percentage Relative Error	Pearson Correlation Coefficient	Mean Absolute Error (Infected per day)	Mean Percentage Relative Error
Brazil	0.44	2911	0.20	**0.65**	**2508**	**0.17**
Chile	**0.64**	**504**	**0.16**	0.44	741	0.23
Colombia	0.79	244	0.32	**0.83**	**76**	**0.09**
Mexico	**0.77**	**231**	**0.09**	0.44	550	0.18
Peru	0.52	2375	0.60	**0.76**	**686**	**0.18**
USA	**0.93**	**799**	**0.04**	0.18	4092	0.17

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 September 2020 doi:10.20944/preprints202009.0228.v1

In the Figures 6 to 11, we show the box plot of the real values vs predicted values for each country. These charts indicate the correspondence between the distributions, giving us an idea of the good performance of each model. All distributions fit and are coupled in a look-acceptable manner; the largest deviation was found in the Peru chart.

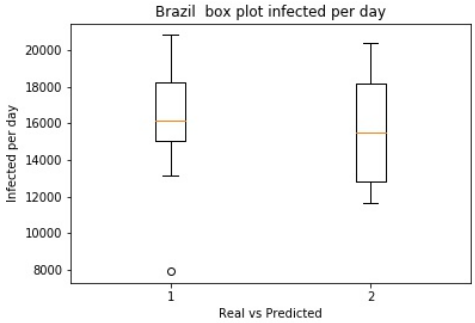

Figure 6. Box-plot chart real vs predicted quantity of infected cases in Brazil

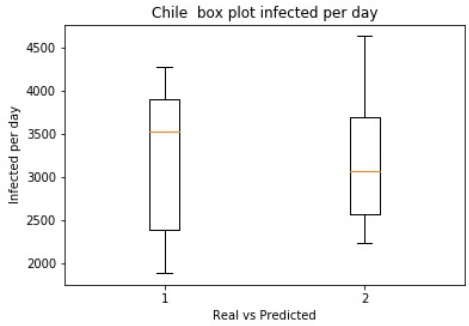

Figure 7. Box plot chart real vs predicted quantity of infected cases in Chile

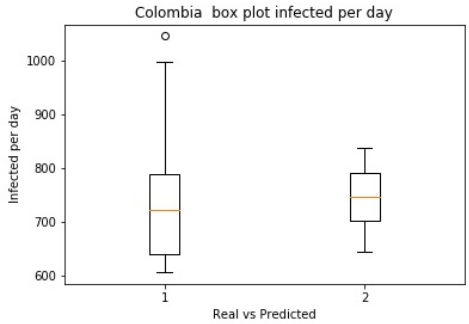

Figure 8. Box plot chart real vs predicted quantity of infected cases in Colombia

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 September 2020 doi:10.20944/preprints202009.0228.v1

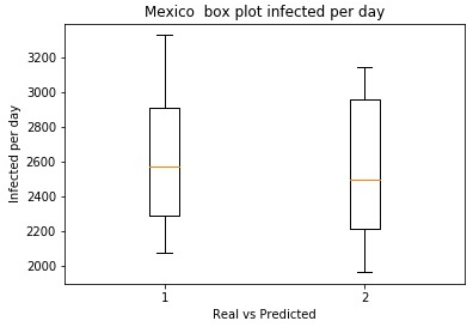

Figure 9. Box plot chart real vs predicted quantity of infected cases in Mexico

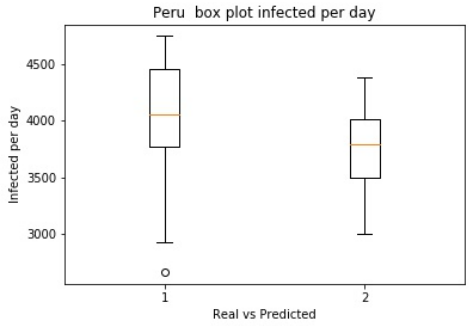

Figure 10. Box plot chart real vs predicted quantity of infected cases in Peru

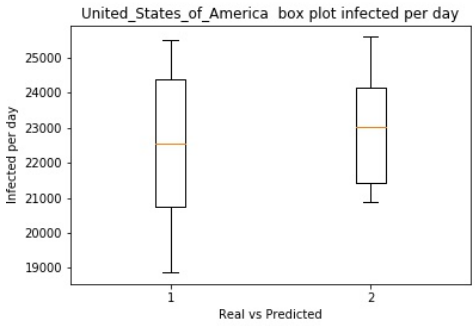

Figure 11. Box plot chart real vs predicted quantity of infected cases in the United States

4. Discussion

The proposed machine learning methods enable us to forecast the confirmed Covid19 cases with outstanding performance as shown in Table 2. However, determining which is the best model to use depends on the behavior of the data in each country, as for some cases it is better to use the Multilayer Perceptron while for others the Support Vector Machine. In this sense, this work contributes to demonstrating that the use of different regression models based on machine learning allows to obtain a reliable and adaptable tool for predicting the behavior of the pandemic in each country. In the same way, it contributes to demonstrating the need to fine tune the hyperparameters of the models to obtain the best performance sought as shown in Table 1. Similarly, we were able to verify that the distributions of the predicted data versus the real data are

 doi:10.20944/preprints202009.0228.v1

similar, as observed in figures 1 to 9. On the other hand, figure 10 verifies the technique presented in [4] applied to the prediction of SarsCov2 infected cases in different countries of Latin America. On the other hand, in [16] we find the use of clustering and autoencoders techniques to forecast infected cases in China, which enables us highlight the joint use of artificial intelligence algorithms as a powerful tool in this type of application for health care and decision-making. It differs in the methods used as we apply supervised learning in all cases.

In [17] [18] and [19] classical statistical-mathematical models such as ARIMA and optimization heuristics such as the flower pollen algorithm and Swarm particles algorithm are used for pandemic forecasting in Iran, China and globally respectively. These works are similar to ours as they seek the prediction of the time series of confirmed cases in the growth curve of the pandemic. However, we observe a difference as the authors did not use machine learning algorithms to carry out this task. In [20] the application of autoregressive neural networks for the prediction of SarsCov2 confirmed cases in Egypt is presented. This work is very similar to ours as the authors use one of the models which we propose in our work and they apply it almost identically. However, the authors did not adjust the hyperparameters of the MLP structure using an optimization algorithm such as hill climbing - tabu list. On the other hand, they do not compare the performance with another machine learning method although they compare with a statistical method such as ARIMA.

These works lead us to consider the main limitation of our research: the small amount of data available and used, which did not allow us to train more sophisticated regression models based on Deep Learning that could have a better performance. In this sense, it is possible that, as the infectious phenomenon evolves, the models that we propose could require a re-entrenchment and a new adjustment of hyperparameters to follow the trend of the behavior curve. Finally, due to the characteristics of the pandemic, we found the need to include other variables related to the characteristics of each country in order to improve the performance of the models used. We highlight variables such as gross domestic product and population density to contribute significantly to forecast the number of confirmed cases of Covid19. In the same way, it is very important to note that we did not find related works in state of the art that forecast the confirmed cases of Covid19 on the American continent.

5. Conclusion

Public health requires efficient planning in resources administration, which are limited in developing countries. Having prediction tools allows efficient management of government resources to face the problems caused by the COVID19 pandemic. In this way, machine learning-based prediction systems are fundamental tools in decision-making to save thousands of lives. For this reason, our model contributes significantly to developing tools to build plans aimed at facing the world's current public health problem.

This research work demonstrates that an MLP can predict better when an optimization algorithm to determine the hyperparameter (number of layers and number of neurons per layer) is applied, as hill climbing tabu list algorithm. However, a global minimum [21] was not found for all cases, making it necessary to use a Support Vector Machine when MLP did not perform well. In general, the comparison of the resulting evaluation metrics shows evidence that the proposed methods could be used as a forecasting tool in the COVID19 pandemic before reaching the peak where the data quickly increase.

With the proposed optimization algorithms, improvements were achieved on the performance metrics for the MLP model of Chile, Mexico and the USA. The Pearson correlation coefficient shows that the trend of the data continues and MAE and MPE show a good performance. On the other hand, SVM performs well in the same metrics for Brazil, Colombia and Peru. We can infer that it is important to constantly use different machine learning models to reach the global minima for each data case.

References

[1] YANG, Xiaobo, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. The Lancet Respiratory Medicine, 2020.

[2] WORLD HEALTH ORGANIZATION, et al. Coronavirus disease 2019 (COVID-19): situation report, 85. 2020.

[3] ATKESON, Andrew. What will be the economic impact of COVID-19 in the US? Rough estimates of disease scenarios. National Bureau of Economic Research, 2020.

[4] CABALLERO, Liesle; JOJOA, Mario; PERCYBROOKS, Winston S. Optimized neural networks in industrial data analysis. SN Applied Sciences, 2020, vol. 2, no 2, p. 300.

[5] MOONEY, Stephen J.; PEJAVER, Vikas. Big data in public health: terminology, machine learning, and privacy. Annual review of public health, 2018, vol. 39, p. 95-112.

[6] ARNAU, Jaume; BALLUERKA, Nekane. Análisis de datos longitudinales y de curvas de crecimiento. Enfoque clásico y propuestas actuales. Psicothema, 2004, vol. 16, no 1, p. 156-162.

[7] GUNST, Richard F. Regression analysis and its application: a data-oriented approach. Routledge, 2018.

[8] Open data repository of the European Union. https://data.europa.eu/euodp/es/data/. Downloaded 25 of May 2020.

[9] HAYKIN, Simon. Neural networks: a comprehensive foundation. Prentice Hall PTR, 1994.

[10] SHAWE-TAYLOR, John, et al. Kernel methods for pattern analysis. Cambridge university press, 2004.

[11] DUAN, Kaibo; KEERTHI, S. Sathiya; POO, Aun Neow. Evaluation of simple performance measures for tuning SVM hyperparameters. Neurocomputing, 2003, vol. 51, p. 41-59.

[12] SELMAN, Bart; GOMES, Carla P. Hill-climbing Search. Encyclopedia of cognitive science, 2006.

[13] CHELOUAH, Rachid; SIARRY, Patrick. Tabu search applied to global optimization. European journal of operational research, 2000, vol. 123, no 2, p. 256-270.

[14] WILLMOTT, Cort J.; MATSUURA, Kenji. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research, 2005, vol. 30, no 1, p. 79-82.

[15] HENAO GIRALDO, Ricardo. Selección de hiperparámetros en máquinas de soporte vectorial. Departamento de Ingeniería Eléctrica, Electrónica y Computación, 2004.

[16] HU, Zixin, et al. Artificial intelligence forecasting of covid-19 in china. arXiv preprint arXiv:2002.07112, 2020.

[17] MOFTAKHAR, Leila; SEIF, Mozhgan. The Exponentially Increasing Rate of Patients Infected with COVID-19 in Iran. Archives of Iranian medicine, 2020, vol. 23, no 4, p. 235-238.

[18] AL-QANESS, Mohammed AA, et al. Optimization method for forecasting confirmed cases of COVID-19 in China. Journal of Clinical Medicine, 2020, vol. 9, no 3, p. 674.

[19] BENVENUTO, Domenico, et al. Application of the ARIMA model on the COVID-2019 epidemic dataset. Data in brief, 2020, p. 105340.

[20] SABA, Amal I.; ELSHEIKH, Ammar H. Forecasting the prevalence of COVID-19 outbreak in Egypt using nonlinear autoregressive artificial neural networks. Process Safety and Environmental Protection, 2020.

[21] HAEFFELE, Benjamin D.; VIDAL, René. Global optimality in neural network training. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017. p. 7331-7339.

Acknowledgement: Acknowledgement to University of Deusto.

Funding Statement: "The author(s) received no specific funding for this study."

Conflicts of Interest: "The authors declare that they have no conflicts of interest to report regarding the present study."