Article

Predicting COVID-19 cases, deaths and recoveries using machine learning methods

Mohamed Lounis[1],* and Farhan Mohammad Khan[2]

[1] Department of Agro-veterinary Science, Faculty of Natural and Life Sciences, University of Ziane Achour, BP 3117, Road of Moudjbara, Djelfa 17000, Algeria.

[2] Department of Civil Engineering, BITS Pilani, Pilani Campus, India.

* Correspondence: lounisvet@gmail.com

Abstract: In the presented work we applied three machine learning techniques to forecast and predict COVID-19 cases, deaths ad recoveries numbers in Algeria for the next six months using data from February 25th, 2020 to April 26th , 2021. These models are represented by the Gaussian process regression (GPR), the support vector machine (SVM) and the decision tree (DT). The plotting results and parameters evaluation pointed out that the Gaussian Process Regression (GPR) has the best performance. Prediction with this model showed that the number of cases, deaths and recoveries will increase in the next months Algeria recording a peak in the month of August and the curve will tend to decrease later.

1. Introduction

Seventeen months after its emergence, the coronavirus disease 2019 (COVID-19) continues its propagation affecting more than 165 million patients leading to more than 3.4 million deaths surpassing all expectations. Algeria has seen its former case emerged on February, 25th 2020. After multiple facets of the epidemiological curve, the number of cases has attained 125,896 subjects. This number seems to be lower than the number of cases reported in the bordered countries like Morocco (515,758 cases) and Tunisia (329,925 cases). With 3,395 deaths and 87,746 recovered persons these numbers determine until now a fatality rate of 2.7 % and a cured rate of 69.7% respectively [1]. To understand the epidemiological traits of this disease and to predict its evolution and its probable end-point multiple approaches have been used in Algeria and around the world. These approaches varied from epidemiological and mathematical/statistical to deep learning/machine learning models [2]. In this way, machine learning models are of great importance [3]. These tools which have proved their role in different complicated problems in different field in the last years including health, agriculture, engineering, sport, climate and robotics [4] have been widely used in the current context of COVID-19 [5–8].

Among these models we can find auto regressive integrated moving average (ARIMA) models [5], BSTS (Bayesian structural time series) [4], simple RNN (recurrent neural network) [7], artificial neural network (ANN) [8], long-short term memory (LSTM) [9], linear regression [10], adaptive neurofuzzy inference system (ANFIS) [11], least absolute shrinkage and selection operator (LASSO) regression [12], CUBIST (cubist regression) [13], Gaussian process regression (GPR) [14], exponential smoothing (ES) [15], random forest (RF) [8,13,16], ridge regression (RIDGE) [13], support vector machine (SVM) [8,13], Naïve bayes (NB) [8], decision tree (DT) [8], box-jenkins method [17], variational auto encoder (VAE) [7,10], gated recurrent units (GRU) [7,9] and multi-layer perceptron (MLP), models [18].

2. Related works

Multiple researchers have been carried out using machine learning methods in the actual COVID-19 context. Below are cited some conducted researches using Gaussian process regression (GPR), support vector machine (SVM) and decision tree (DT):

After analyzing historical COVID-19 data, Velásquez and Lara [14] forecasted COVID-19 affection with reduced-space Gaussian process regression associated to chaotic dynamical systems using obtained information of the two first months (January 21, to April 12, 2020). Their work demonstrated the usefulness of the Gaussian models in the COVID-19 infection prediction.

In their study, Ribeiro *et al.*, [13] set as objectives the evaluation of the performance of multiple models like autoregressive integrated moving average (ARIMA), cubist regression (CUBIST), random forest (RF), RIDGE regression, support vector regression (SVR), and stacking-ensemble learning in a COVID-19 cases short projection of 1, 3 and 6 days of the ten most affected states in Brazil. The performance evaluation has given the following classification: SVR, stacking-ensemble learning, ARIMA, CUBIST, RIDGE, and RF models.

In a comparison made by Ball [14], the support vector machines (SVM) has demonstrated higher performance than linear regression, multi-layer perceptron, random forest models in predicting COVID-19 trend in USA, Germany and the global. In Mexico logistic regression, decision tree, support vector machine, naive Bayes, and artificial neutral network to study COVID-19 cases by Muhammed *et al.*, [8]. The researchers observed that decision tree, support vector machine and Naïve bayes model have the highest accuracy (94.99%), sensitivity (93.34%) and specificity (94.30%) respectively.

Daniyal *et al.*, [19] in Pakistan, compared the performance of three regression models including linear, logarithmic, and quadratic in modeling of COVID-19 deaths using data of about 5 months. Later, they deduced that the rate of mortality will decrease by the end of October as shown by the quadratic regression model which has shown the best performance.

Prediction of COVID-19 mortality in Korea was the main objective of the study of An *et al.*, [12]. The study begun by testing the least absolute shrinkage and selection operator (LASSO), linear support vector machine (SVM), SVM with radial basis function kernel, random forest (RF), and k-nearest neighbors. As a result, LASSO and linear SVM has shown high sensitivities (90.7% and 92.0%, respectively) and specificities (91.4% and 91.8%, respectively). In the same country, Das *et al.*, [20] predicted mortality in 3,524 COVID-19 patients using five machine learning models (logistic regression, support vector machine, K nearest neighbor, random forest and gradient boosting). The logistic regression model was proposed as an open-source online prediction tool for decision-making due to its high performances.

3. Methodology

In this paper, COVID-19 time series data available till 26th April 2021 in Algeria were used for a projection of daily cases, deaths and recoveries for the next six months using three machine learning techniques that are Gaussian process regression (GPR), support vector machine (SVM) and decision tree (DT). Data regarding the number of cases reported in Algeria, were extracted from Worldometer. The COVID-19 curve evolution is shown in Figure 1.
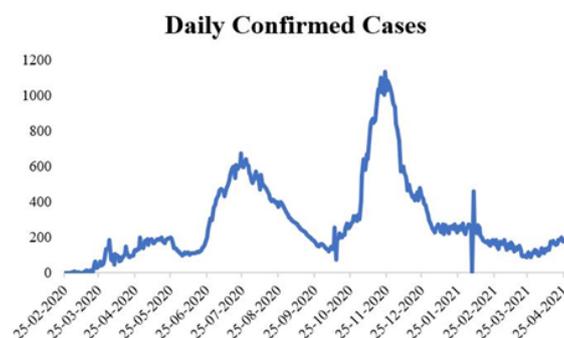
Figure 1. COVID-19 curve evolution in Algeria.

4. Results and discussion

In the current wok, three machine learning approaches were applied to predict the number of COVID-19 cases, deaths and cured persons in Algeria. We first evaluated the forecast performance of these models by the estimation of parameters like the root mean square error (RMSE), the mean square error (MSE), the mean absolute error (MAE), and the coefficient of determination (R2) values for COVID-19 daily cases. Results

Table 1. Performance parameters of the three model.

Model parameters	GPR	Quadratic SVM	DT
RMSE	31.126	42.485	37.93
R-squared	0.98	0.97	0.97
MSE	968.83	1804.9	1438.7
MAE	18.334	26.268	22.996

showed that if the three models have shown acceptable performances (Table 1), the GPR model was the most efficient showing an RMSE of 31.126 and an R2 of 0.98. These parameters were calculated by comparing actual/predicted cases after a 10-fold cross-validation. Figures 2, 3 and 4 showed response plots the three models GPR, SVM, and DT respectively. Figures 5, 6 and 7 present the predicted/observed pattern of each model.
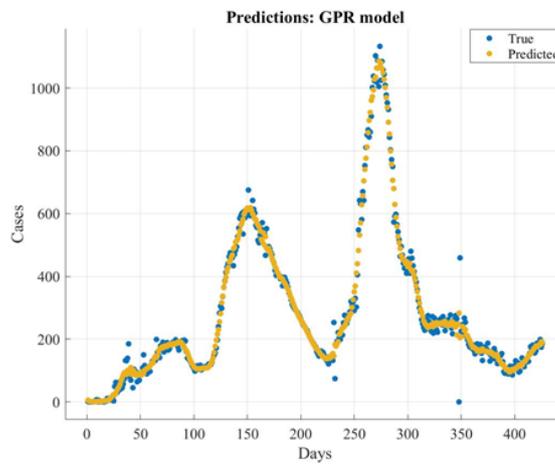
Figure 2. Plotting response of the GPR model
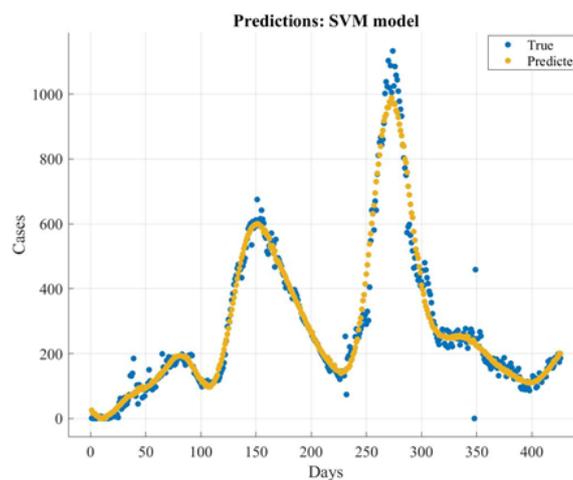
Figure 3. Plotting response of the SVM model.
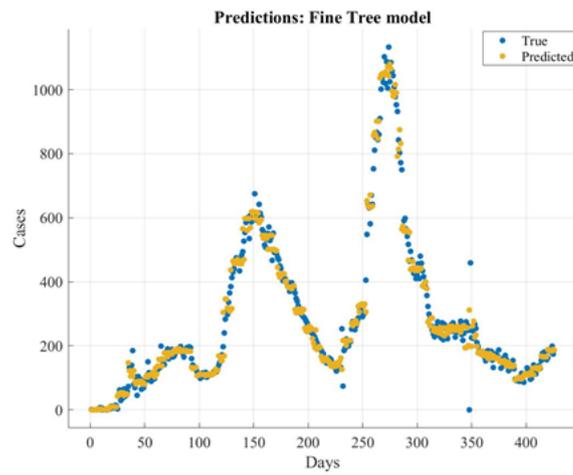
Figure 4. Plotting response of the DT model.
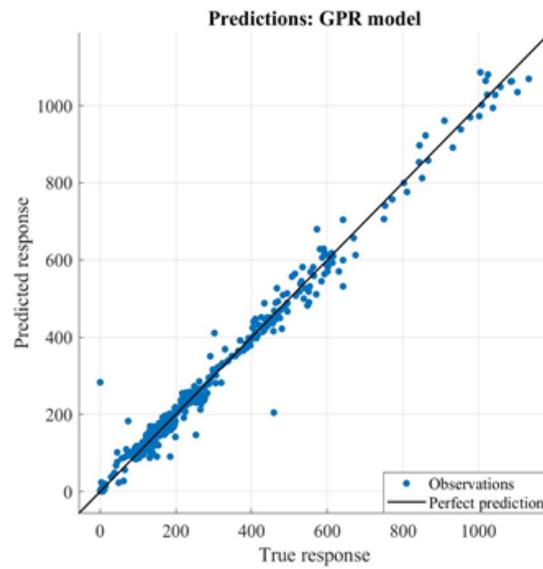
Figure 5. Predicted/Observed cases pattern of the GPR model.
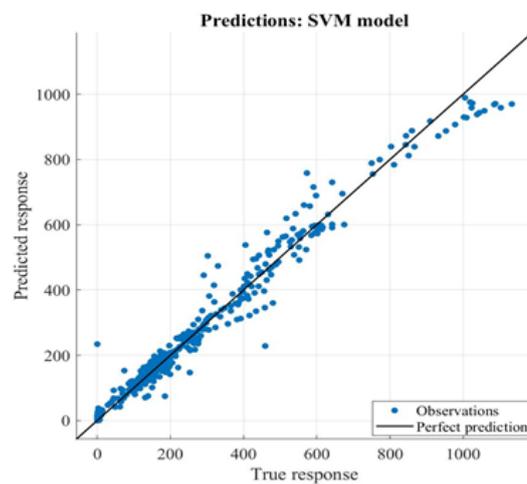
Figure 6. Predicted/Observed cases pattern of the SVM model.
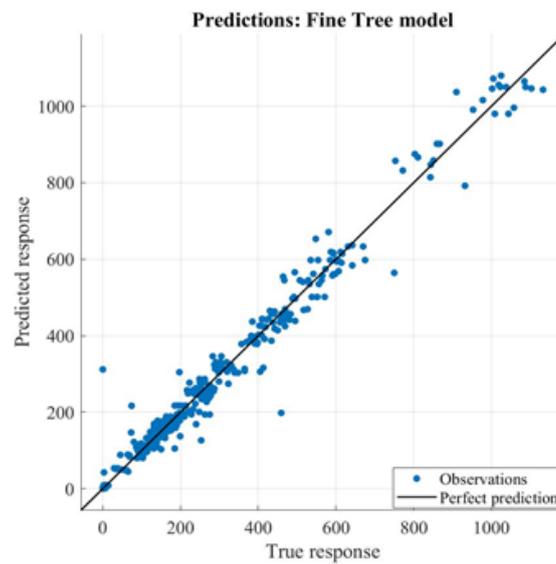
Figure 7. Predicted/Observed cases pattern of the DT model.

We then, used available data till 26th April 2021, of daily confirmed, recovered, and deceased cases of COVID-19 cases in Algeria and forecasted them using the three models for the next six months. Predicted daily new cases, recovered and dead persons are shown in Figures 8, 9 and 10 respectively. Results showed that confirmed cases will increase in the next months and will start their declining from the first week of October according to the GPR model. The number of recoveries (Figure 9) and deaths (Figure 10) follow generally the same evolutionary curve.
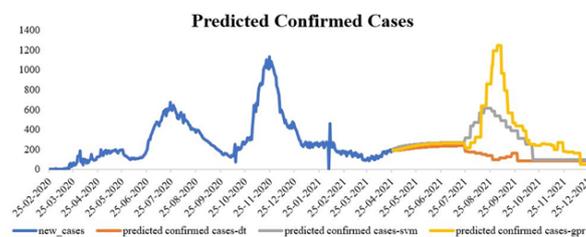
Figure 8. Forecast of Daily Confirmed Cases using machine learning methods.
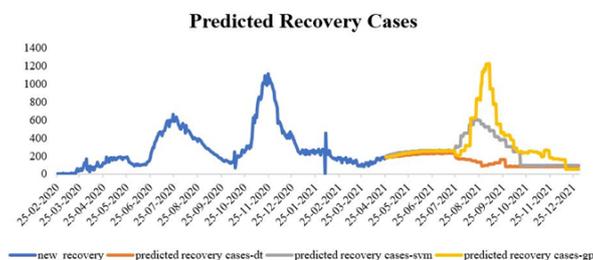
Figure 9. Forecast of Daily Recovery Cases using machine learning methods.
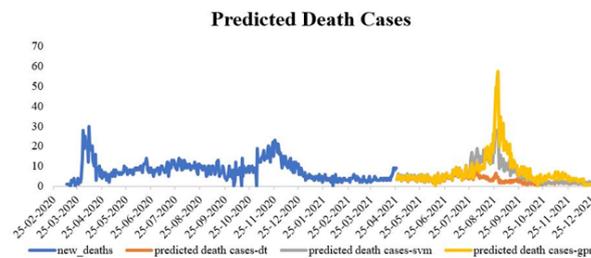
Figure 10. Forecast of Daily Deceased Cases using machine learning methods.

It is to mention that these projections were done without considering the effect of preventive measures which are considered to be the same in the next months. Prediction performance could be ameliorated if their effect will be added. The performance of our models has shown a high value for the coefficient of determination of the three models used in this study. As a comparison we can show that our models have better performances in term of R2 than other models like ARIMA (0.95) [21] and ANFIS (0.956) [22]. Other models like MPL-ICA (0.9971) [22], logistic regression (0.996) [23] and lasso regression (1.0) [24] have demonstrated higher performances.

5. Conclusion

As a conclusion, we used in this work three machine learning methods to forecast COVID-19 cases, deaths and recoveries in Algeria. Results showed that all models showed acceptable performances and the exponential GPR network was the most efficient. Prediction with this model showed that Algeria will probably recorded a third wave with a peak in the month of August 2021.

Author Contributions: All authors contributed equally to the writing of this paper. All authors read and approved the final manuscript.

Conflicts of Interest: "The authors declare no conflict of interest."

References

[1]   Coronavirus updates Live 2020. https://www.worldmeters.info/coronavirus/ Accessed May 20, 2021.

[2]   Kotwal, A., Yadav, A. K., Yadav, J., Kotwal, J., & Khune, S. (2020). Predictive models of COVID-19 in India: a rapid review. *Medical Journal Armed Forces India, 76*(4), 377-386.

[3]   Kwekha-Rashid, A. S., Abduljabbar, H. N., & Alhayani, B. (2021). Coronavirus disease (COVID-19) cases analysis using machine-learning applications. *Applied Nanoscience*, 1-13.

[4]   Mojjada, R. K., Yadav, A., Prabhu, A. V., & Natarajan, Y. (2020). Machine learning models for covid-19 future forecasting. *Materials Today: Proceedings*, https://doi.org/10.1016/j.matpr.2020.10.962.

[5]   Khan, F., Saeed, A., & Ali, S. (2020). Modelling and forecasting of new cases, deaths and recover cases of COVID-19 by using vector autoregressive model in Pakistan. *Chaos, Solitons & Fractals, 140*, 110189, https://doi.org/10.1016/j.chaos.2020.110189.

[6]   Feroze, N. (2020). Forecasting the patterns of COVID-19 and causal impacts of lockdown in top five affected countries using Bayesian structural time series models. *Chaos, Solitons & Fractals, 140*, 110196, https://doi.org/10.1016/j.chaos.2020.110196.

[7]   Zeroual, A., Harrou, F., Dairi, A., & Sun, Y. (2020). Deep learning methods for forecasting COVID-19 time-Series data: A comparative study. *Chaos, Solitons & Fractals, 140*, 110121, https://doi.org/10.1016/j.chaos.2020.110121.

[8]   Muhammad, L. J., Algehyne, E. A., Usman, S. S., Ahmad, A., Chakraborty, C., & Mohammed, I. A. (2021). Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. *SN computer science, 2*(1), 1-13.

[9]   Shahid, F., Zameer, A., & Muneeb, M. (2020). Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos, Solitons & Fractals, 140*, 110212, https://doi.org/10.1016/j.chaos.2020.110212.

[10]   Rustam, F., Reshi, A. A., Mehmood, A., Ullah, S., On, B. W., Aslam, W., & Choi, G. S. (2020). COVID-19 future forecasting using supervised machine learning models. *IEEE Access, 8*, 101489-101499.

[11]   Al-Qaness, M. A., Ewees, A. A., Fan, H., & Abd El Aziz, M. (2020). Optimization method for forecasting confirmed cases of COVID-19 in China. *Journal of Clinical Medicine, 9*(3), 674, https://doi.org/10.3390/jcm9030674.

[12] An, C., Lim, H., Kim, D. W., Chang, J. H., Choi, Y. J., & Kim, S. W. (2020). Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study. *Scientific reports, 10*(1), 1-11.

[13] Ribeiro, M. H. D. M., da Silva, R. G., Mariani, V. C., & dos Santos Coelho, L. (2020). Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. *Chaos, Solitons & Fractals, 135*, 109853, https://doi.org/10.1016/j.chaos.2020.109853.

[14] Velásquez, R. M. A., & Lara, J. V. M. (2020). Forecast and evaluation of COVID-19 spreading in USA with reduced-space Gaussian process regression. *Chaos, Solitons & Fractals, 136*, 109924, https://doi.org/10.1016/j.chaos.2020.109924.

[15] Shastri, S., Singh, K., Kumar, S., Kour, P., & Mansotra, V. (2021). Deep-LSTM ensemble framework to forecast Covid-19: an insight to the global pandemic. *International Journal of Information Technology, 13*, 1291Ű1301.

[16] Balli, S. (2021). Data analysis of Covid-19 pandemic and short-term cumulative case forecasting using machine learning time series methods. *Chaos, Solitons & Fractals, 142*, 110512, https://doi.org/10.1016/j.chaos.2020.110512.

[17] Das, R. C. (2020). Forecasting incidences of COVID-19 using Box-Jenkins method for the period July 12-Septembert 11, 2020: A study on highly affected countries. *Chaos, Solitons & Fractals, 140*, 110248, https://doi.org/10.1016/j.chaos.2020.110248.

[18] Kafieh, R., Arian, R., Saeedizadeh, N., Minaee, S., Amini, Z., Yadav, S. K., & Javanmard, S. H. (2020). COVID-19 in Iran: a deeper look into the future. *MedRxiv*, https://doi.org/10.1101/2020.04.24.20078477 .

[19] Daniyal, M., Ogundokun, R. O., Abid, K., Khan, M. D., & Ogundokun, O. E. (2020). Predictive modeling of COVID-19 death cases in Pakistan. *Infectious Disease Modelling, 5*, 897-904.

[20] Das, A. K., Mishra, S., & Gopalan, S. S. (2020). Predicting CoVID-19 community mortality risk using machine learning and development of an online prognostic tool. *PeerJ, 8*, e10083, https://doi.org/10.7717/peerj.10083.

[21] Khan, F. M., & Gupta, R. (2020). ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India. *Journal of Safety Science and Resilience, 1*(1), 12-18.

[22] Pinter, G., Felde, I., Mosavi, A., Ghamisi, P., & Gloaguen, R. (2020). COVID-19 pandemic prediction for Hungary; a hybrid machine learning approach. *Mathematics, 8*(6), 890, https://doi.org/10.3390/math8060890.

[23] Batista, M. (2020). Estimation of the final size of the COVID-19 epidemic. *MedRxiv,* https://doi.org/10.1101/2020.02.16.20023606.

[24] Onovo, A., Atobatele, A., Kalaiwo, A., Obanubi, C., James, E., Gado, P., & Russell, M. (2020). *Using supervised machine learning and empirical Bayesian kriging to reveal correlates and patterns of COVID-19 disease outbreak in sub-Saharan Africa: exploratory data analysis.* Available at SSRN: https://ssrn.com/abstract=3580721 or http://dx.doi.org/10.2139/ssrn.3580721.