CHAPTER 15

PREDICTING COVID-19 SPREAD USING MACHINE LEARNING ALGORITHMS

Berislav Žmuk[1], Hrvoje Jošić[2]

Abstract

COVID-19 was declared as a world health emergency in January 2020. Since then it has affected all aspects of our lives. Countries closed their borders, put their population in self-quarantine and closed businesses and schools. As of March 31, the infection sickened more than 770,000 people all around the world with thousands of fatalities. There is a little consensus how long will the infection last and what the number of infeced will be. Therefore, it is essential to implement suitable methods for COVID-19 spread prediction which is the goal of this paper. An open source machine learning software Weka and its algorithms (Linear regression, Gaussian Processes, SMOreg and neural network Multilayer Perceptron) have been used to predict the number of cases and fatalities of COVID-19 disease for 10 days in the future. The accuracy of the forecasts is measured using MAPE and RMSE error metrics. The results of the analysis have shown that Weka and its algorithms can be successfully used for prediction of COVID-19 spread in the world. The results of the analysis indicated that the Gaussian processes and Multilayer perceptron neural network are the most precise algorithms for the prediction of new and total cases of COVID-19 disease on a global scale and on an individual country level. The values of MAPE criterion for 12 selected countries, in majority of cases, have shown a highly accurate or good forecasting ability. The results obtained from this analysis can be important for global community and especially for economic and health policy makers in order to guide COVID-19 surveillance and implement public health policy measures.

Key words: COVID-19, machine learning, disease spread prediction.

JEL classification: C53, I19

1. Introduction

The World Health Organization (WHO) first declared COVID-19 a world health emergency in January 2020, Congressional Research Service (2020). Since then the virus has affected all aspects of our lives. According to Boissay and Rungcharoenkitkul (2020) the Covid-19 pandemic is not only the most serious global health crisis since the 1918 Great Influenza (Spanish flu), but is set to become one of the most economically costly pandemics in recent history.

1 Assistant Professor, University of Zagreb, Faculty of Economics and Business in Zagreb, J. F. Kennedy Square 6, 10 000 Zagreb, Croatia. Department of statistics. Phone: +385 1 238 3372. E-mail: bzmuk@efzg.hr.

2 Assistant Professor, University of Zagreb, Faculty of Economics and Business in Zagreb, J. F. Kennedy Square 6, 10 000 Zagreb, Croatia. Department of international economics. Phone: +385 1 238 3350. E-mail: hjosic@efzg.hr.

Countries closed their borders, put their population in self-quarantine and closed businesses and schools. As of March 31, the infection sickened more than 770,000 people all around the world with thousands of fatalities. There is little consensus how long will infection last and what the number of cases of infection overall will be. Therefore, it is essential to implement suitable methods for COVID-19 spread prediction. For that purpose the use of artificial intelligence and machine learning algorithms can be very usefull. Few authors have investigated the use of artificial antelligence in Covid-19 forecasting, Hu et al (2020) in the case of China, Mosavi et al (2020) investigated the Covid-19 outbreak prediction with machine learning while Luo (2020) essayed on predictive monitoring of Covid-19.

Goal of this paper is predict the COVID-19 spread using an open source machine learning software Weka and its algorithms (Linear regression, Gaussian Processes, SMOreg and neural network Multilayer Perceptron), University of Waikato (2020). WEKA has been used for time series forecasting and financial stock market prediction, Kannan et al (2010), Kulkarni and More (2016). Other applications of WEKA related to disease prediction are for dengue disease prediction, Shakil, Anis and Alam (2015), classification and prediction of diabetics, Selvi et al (2018) and for heart diseases prediction, Saad Mohamed (2020). In the paper the machine learning algorithms and MAPE and RMSE metrics for measuring the accuracy of forecasts will be used. It is expected that WEKA's machine learning tools will have good to excellent predictive ability in forecasting total cases and fatalities due to COVID-19 infection.

Paper is structured in five chapters. After the introduction, short literature review elaborates on the use of artificial intelligence, machine learning specifically in COVID-19 spread prediction. In the methodology and data section descriptive statistics of data are presented and methodology of the paper is explained. In the results and discussion section the main results of the analysis are displayed. Firstly on a global level of 69 countries in the World and after that on an individual countries level detaily analysing 12 countries with the most cases of COVID-19 infection. Final chapter presents concluding remarks.

2. Short literature review

In this chapter short literature review will be presented elaborating on the use of artificial intelligence and machine learning in COVID-19 spread prediction. Among the wide range of machine learning models investigated, multi-layered perceptron and adaptive network-based fuzzy inference system showed most promising results in predicting the COVID-19 outbreak, Mosavi et al (2020). During the COVID-19 pandemic various models and methods have been developed and adopted to forecast infection cases and deaths with some of them influencing the policies in some countries, Luo (2020). The prediction of future is uncertain by nature and no model or data can accurately predict the pandemic spread. Hu et al (2020) developed a modified stacked auto-encoder for

modeling and forecasting the transmission dynamics of the Covid-19 epidemics applied on the confirmed cases of disease across China. The accuracy of the artificial intelligence based methods for forecasting the trajectory of Covid-19 was high. It was predicted that the epidemics of Covid-19 will be over by the middle of April. Gu et al predicted the trend of the COVID-19 epidemic in the whole of China except Hubei based on the existing dana using five mathematical models for regression and simulation (cubic, quadratic, exponential, logarithmic and power models). They found that the inflection point about the COVID-19 may have passed. Zhang, Renjun and Lin (2020) predicted turning points, duration and attack rate of COVID-19 outbreaks in major Western countries by employing a segmented Poisson model. The analysis allowed them to identify and predict the turning point, spread, the duration and the final size of the outbreak of COVID-19 in countries studied. Li et al (2020) undergo propagation analysis and prediction of the COVID-19. It was found that imposing controls would have important impact on the epidemic. Big data integration and analytics played a key role to successfully prevent COVID-19 hospital outbreaks in Taiwan, Chen, F-M. et al. (2020). Cássaro and Pires (2020) tackled the question can the occurrence of COVID-19 cases be predicted. Simulations based on a simple model of growth and initial COVID-19 cases (for 1st 2nd and 3rd weeks) showed that it is almost impossible to predict how the pandemic will evolve. In this paper the open source machine learning software Weka and its algorithms will be used to predict the COVID-19 spread for 69 countries in the world and 12 chosen countries in details. Weka has already been used for for dengue disease prediction, Shakil, Anis and Alam (2015), classification and prediction of diabetics, Selvi et al (2018) and for heart diseases prediction, Saad Mohamed (2020).

3. Methodology and data

In this paper the COVID-19 spread in 69 countries in the world will be inspected by using Weka machine learning algorithms. Data which will be observed are the new cases and deaths, total cases and total deaths due to the COVID-19 disease. The daily values of these four variables will be observed in the period from December 31, 2019 to March 31, 2020. The countries which will be inspected are the ones that had 100 or more cases of COVID-19 infection and 20 or more deaths in the observed period. The list of the observed countries is the following: Albania, Algeria, Andorra, Argentina, Australia, Austria, Bangladesh, Belgium, Bosnia and Herzegovina, Brazil, Bulgaria, Burkina Faso, Canada, Chile, China, Colombia, Croatia, Czechia, Democratic Republic of the Congo, Denmark, Dominican Republic, Ecuador, Egypt, Estonia, Finland, France, Germany, Greece, Honduras, Hungary, India, Indonesia, Iran, Iraq, Ireland, Israel, Italy, Japan, Luxembourg, Malaysia, Mexico, Moldova, Morocco, Netherlands, North Macedonia, Norway, Pakistan, Panama, Peru, Philippines, Poland, Portugal, Puerto Rico, Romania, Russia, San Marino, Saudi Arabia, Serbia, Slovenia, South Korea, Spain, Sweden, Switzerland, Thailand, Tunisia, Turkey,

Ukraine, United Kingdom and United States of America. As a data source the EU Open Data Portal (2020) was used.

In order to forecast values of the four observed variables for each country separately and for the world overall, four forecasting approaches or algorithms are used: Gaussian processes, Linear regression, Multilayer perceptron and SMOreg. More information about WEKA's forecasting algorithms can be found in Popescu et al (2009), Rasmussen and Williams (2006) and Smola and Schölkopf (2004). For all four forecasting approaches default settings in Weka, statistical software which was used to calculate the forecasts, will be applied. In this way the results and forecasting levels of precision between countries can be directly compared. The actual data from December 31, 2019 to March 31, 2020 were used to forecast values of the four observed variables for 10 days in the future. More precisely, the forecasts for the period from April 1 to April 10, 2020 were calculated. After that the forecast values from those 10 days are compared with the actual values and the differences are measured by calculating selected forecasting errors. Weka's interface offers various statistical error measures for evaluation of aforementioned algorithms. In order to determine forecast precision levels two forecast errors were used: mean absolute percentage error (MAPE) and root mean squared error (RMSE). They were chosen as the most common error measures used for performance analysis. The formulas for calculation of MAPE and RMSE metrics are displayed in equations 1 and 2.

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{\bar{Y}_i - Y_i}{\bar{Y}_i}\right| \qquad (1)$$

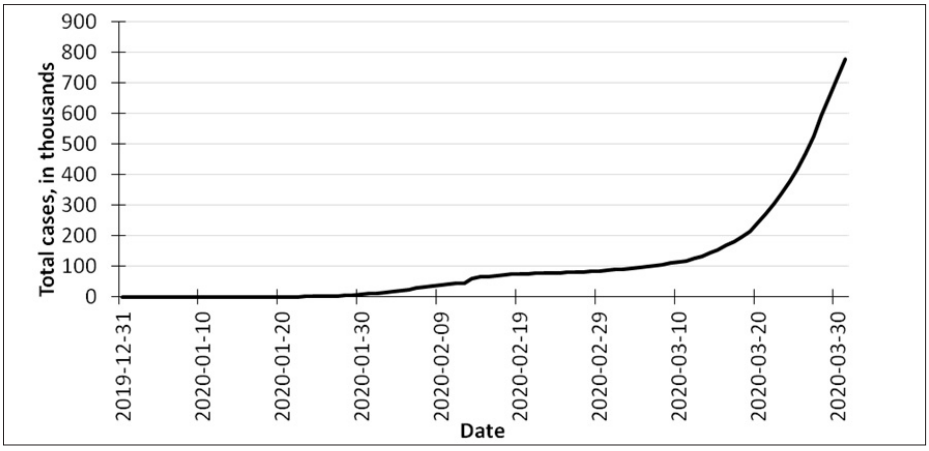
$$RMSE = \sum\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\bar{Y}_i - Y_i)^2} \qquad (2)$$

where $\bar{Y}_i$ is the predicted value and Y_i is the observed value for the number of observations.

There were only 27 cases of COVID-19 in the World on December 31, 2019. All 27 cases were confirmed in China. Three months later (March 31, 2020) the total number of COVID-19 cases of disease increased to 777,133. The total number of COVID-19 cases in the World in the period from December 31, 2019 to March 31, 2020 is shown in the Figure 1.

According to the Figure 1, three phases in development of total number of COVID-19 cases can be detected in the observed period. In the period from December 31, 2019 to January 23, 2020 there was a slight increase in the total number of cases in the World. From January 23, 2020 positive linear trend is present until March 10, 2020. In period from March 10, 2020 to March 20, 2020, slight positive linear trend rapidly increased its slope. Because of that the slope of linear trend is much higher for period after March 20, 2020 than in the period before.

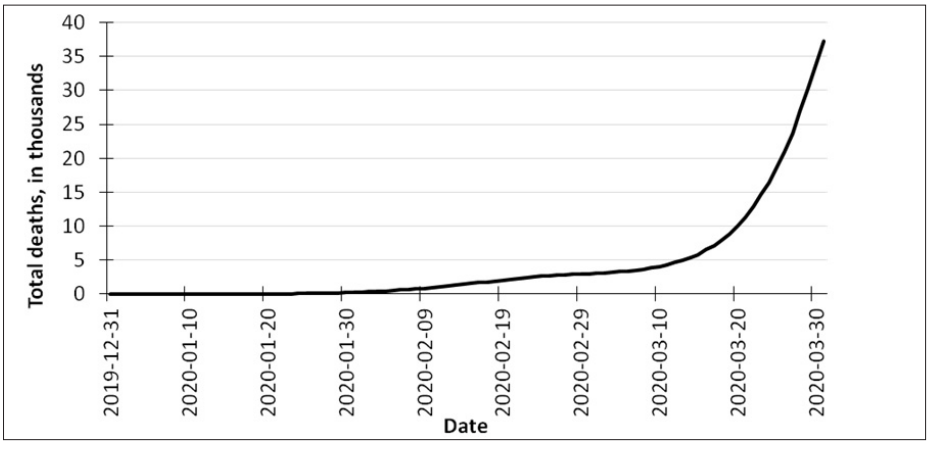
Figure 1: Total number of COVID-19 cases of disease in the period from December 31, 2019 to March 31, 2020



Source: EU Open Data Portal (2020), authors.

The number of total deaths due to COVID-19 in the World in the period from December 31, 2019 to March 31, 2020 is presented in Figure 2. The shape of curve for total deaths due to COVID-19 (Figure 2) follows almost perfectly the shape of curve for total cases (Figure 1). Again, after March 10, 2020 the huge increase in the daily number of deaths due to the COVID-19 appeared.

Figure 2: Total deaths in the World in the period from December 31, 2019 to March 31, 2020



Source: EU Open Data Portal (2020), authors.

Table 1: Top 10 countries and the World according to the total cases of the COVID-19, situation on March 31, 2020

Country	Total cases	Share of cases in total cases in the World (%)	Share of cases in total population (%)
United States of America	164,620	21.2%	0.050%
Italy	101,739	13.1%	0.168%
Spain	85,195	11.0%	0.182%
China	82,241	10.6%	0.006%
Germany	61,913	8.0%	0.075%
France	44,550	5.7%	0.067%
Iran	41,495	5.3%	0.051%
United Kingdom	22,141	2.8%	0.033%
Switzerland	15,412	2.0%	0.181%
Belgium	11,899	1.5%	0.104%
World	777,133	100.0%	0.010%

Note: population data related to 2018.

Source: EU Open Data Portal (2020), authors.

In Table 1 the top 10 countries in the World according to the number of total cases on March 31, 2020 are listed. The country with the most cases is the United States of America in which the 21.2% of total cases of infection in the World appeared. In the top four countries (the United States of America, Italy, Spain, China) more than 50% (to be precise 55.8%) of total cases in the World was found. If the share of cases in total population is observed, from the 10 listed countries, the highest shares have Spain (0.182%), Switzerland (0.181%) and Italy (0.168%).

Table 2: Top 10 countries and the World according to the total deaths due to the COVID-19, March 31, 2020

Country	Total deaths	Share of deaths in total cases (%)	Share of deaths in total population (%)
Italy	11,591	11.39%	0.0192%
Spain	7,340	8.62%	0.0157%
China	3,309	4.02%	0.0002%
United States of America	3,170	1.93%	0.0010%
France	3,024	6.79%	0.0045%
Iran	2,757	6.64%	0.0034%
United Kingdom	1,408	6.36%	0.0021%
Netherlands	864	7.35%	0.0050%
Germany	583	0.94%	0.0007%
Belgium	513	4.31%	0.0045%
World	37,272	4.80%	0.0005%

Note: population data related to 2018.

Source: EU Open Data Portal (2020), authors.

Table 2 is focused on the number of total deaths due to the COVID-19 on March 31, 2020. According to the data from Table 2 the most deaths due to the COVID-19 appeared in Italy (11,591 deaths). About 31% of deaths in the total number of deaths due to COVID-19 happened in Italy. In addition, among 10 listed countries, Italy has the highest share of deaths in total cases of 11.39%.

4. Results and discussion

In Table 3 the best forecasting approaches, among the four observed, for forecasting new cases, new deaths, total cases and total deaths for 69 countries and the World as a whole is shown.

Table 3: The number of cases for the best forecasting approach according to MAPE and RMSE criteria, daily data from December 31, 2019 to March 31, 2020, forecasting period is from April 1 to April 10, 2020, sample of 69 countries and the World as a whole

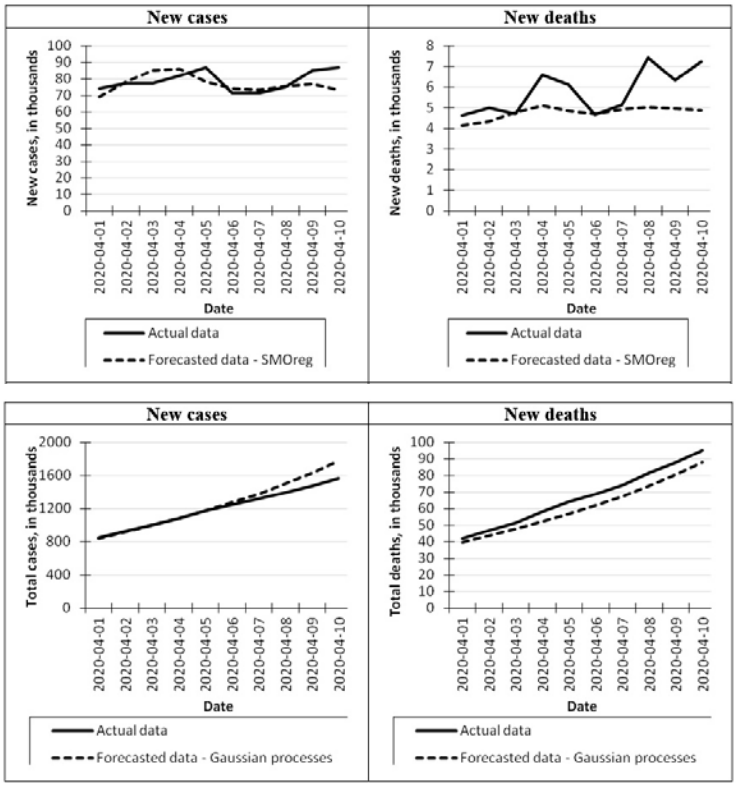
Variable	Forecasting approach	Forecasting error	
		Mean absolute percentage error (MAPE)	Root mean squared error (RMSE)
New cases	Gaussian processes	18	24
	Linear regression	5	3
	Multilayer perceptron	38	36
	SMOreg	9	7
	Total	70	70
New deaths	Gaussian processes	23	30
	Linear regression	1	1
	Multilayer perceptron	41	33
	SMOreg	5	6
	Total	70	70
Total cases	Gaussian processes	34	32
	Linear regression	4	4
	Multilayer perceptron	17	22
	SMOreg	15	12
	Total	70	70
Total deaths	Gaussian processes	38	41
	Linear regression	4	3
	Multilayer perceptron	13	16
	SMOreg	15	10
	Total	70	70

Note: population data related to 2018.

Source: EU Open Data Portal (2020), authors.

The best forecasting approach is selected by applying the two criteria separately. Obviously, in dependence which criteria is observed, minimum mean absolute percentage error or minimum root mean squared error, the best forecasting approach could be different. However, some general conclusions can be brought. When individual changes are observed, like new cases and new deaths, it has been shown that in the most cases the most successful forecasting approach was multilayer perceptron (according to both forecasting error criteria). After multilayer perceptron, very successful in forecasting new cases and new deaths was the Gaussian processes forecasting approach. In addition, the Gaussian processes forecasting process was the most precise in the most cases when the number of total cases and total deaths is observed. The linear regression forecasting approach has shown to be the most precise in the least number of cases. If the mean absolute percentage error criterion is observed, to forecast new cases and new deaths for the World the most precise forecast approach turned out to be SMOreg. On the other hand, to forecast total cases and total deaths for the World the most precise forecast approach is Gaussian processes. In Figure 3 the actual values of the observed variables and forecasted values by using the most precise forecasting approaches (SMOreg or Gaussian processes) for the World is displayed.

Figure 3: Actual and forecasted values of the observed variables for the World in the period from April 1, 2020 to April 10, 2020



Source: EU Open Data Portal (2020), authors.

According to Figure 3 it can be concluded that chosen forecasting approaches can quite precisely forecast values of the observed variables for the World. However, the new cases and the total cases seem to be better forecasted than the new deaths and the total deaths in period from April 1, 2020 to April 10, 2020. Namely, the forecasting approaches underestimated number of the new deaths and the total deaths.

Table 4: The best forecasting approach for selected 12 countries, according to MAPE, forecasting conducted based on daily actual data from December 31, 2019 to March 31, 2020, forecasting period from April 1 to April 10, 2020

Country	Total cases		Total deaths	
	Forecasting approach	MAPE	Forecasting approach	MAPE
Belgium	Multilayer perceptron	10%	Gaussian processes	18%
China	Multilayer perceptron	31%	Multilayer perceptron	1%
Croatia	Multilayer perceptron	8%	Multilayer perceptron	42%
France	Gaussian processes	10%	Gaussian processes	18%
Germany	Gaussian processes	53%	Gaussian processes	6%
Iran	Linear regression	8%	Linear regression	6%
Italy	Gaussian processes	18%	Gaussian processes	10%
Netherlands	Gaussian processes	37%	Gaussian processes	5%
Spain	Gaussian processes	37%	Gaussian processes	7%
Switzerland	SMOreg	12%	Multilayer perceptron	41%
United Kingdom	SMOreg	48%	SMOreg	12%
United States of America	Gaussian processes	20%	Gaussian processes	12%

Source: EU Open Data Portal (2020), authors.

In Table 4 the best forecasting approaches according the MAPE criterion for 12 countries for total cases and total deaths variables are shown. The 12 countries are selected from the list of 10 countries with the most total cases and the list of the 10 countries with the most total deaths. Because 9 countries can be found on both lists (see Table 1 and Table 2), overall 11 countries are observed from those two lists. In addition, Croatia is observed as well. In Table 4, next to the most precise forecasting approach, the values of MAPE indicator are given. It can be noticed that most MAPE values have acceptable low level of errors. However, it can be found some MAPE values with value higher than 40%. Therefore, the highly accurate forecasting (the value of MAPE lower than 10) was found for the variable total cases in Croatia using Multilayer perceptron algorithm and Iran using Linear regression. The interpretation of MAPE values according to the range of observed errors can be found in Lewis (1982). For the variable total deaths the highly accurate forecasting (the value of MAPE lower than 10) was evident in China, Germany, Iran, Netherlands and Spain.

On the other side, inaccurate forecasting (the value of MAPE higher than 50) was found only in Germany for total cases variable and Gaussian processes algorithm. In Figure A1 in Appendix the actual and forecasted values of total cases and total deaths for the 12 observed countries are presented. Forecasted values are calculated by using forecasting approach which resulted in the lowest MAPE value in the forecasting period from April 1, 2020 to April 10, 2020. Again, it can be noticed that the Gaussian processes and Multilayer perceptron neural network are the most successful forecasting algorithms for the prediction of COVID-19 spread.

5. Conclusions

Goal of this paper was predict the COVID-19 spread using an open source machine learning software Weka and its algorithms (Gaussian processes, Linear regression, Multilayer perceptron and SMOreg). The precision of aforementioned algorithms was observed using MAPE and RMSE error criterions. Results of the analysis have shown WEKA and its algorithms can be used for prediction of COVID-19 spread in the World. The results of the analysis indicated that Gaussian processes and Multilayer perceptron neural network were the most precise algorithms for the prediction of new and total cases of COVID-19 disease. The values of MAPE criterion for 12 selected countries, in majority of cases, have shown a highly accurate or good forecasting ability.

Limitations of the paper are related to the observation using sample of 69 countries. The whole population of countries could not be included due to insufficient number of confirmed cases and fatalities due to COVID-19 disease. Also, the reported number of cases of disease and fatalities could be lower than the real number due to lack of medical equipment, insufficient testing with many suspected cases remained to be confirmed because some people showing symptoms may not even be counted as suspected cases yet. Recommendations for future research in this important field of artificial intelligence and machine learning application in COVID-19 spread prediction can relate to prediction of COVID-19 spread in other time period and analysing in details an individual country cases. The results obtained from this analysis can be important for global community and especially for economic and health policy makers.

References

Boissay, F., and Rungcharoenkitkul, P. (2020) “Macroeconomic effects of Covid-19: an early review”, BIS Bulletin, No.7. Available at: <https://www.bis.org/publ/bisbull07.pdf> [Accessed: May 1, 2020]

Cássaro, F.A.M., Pires, L. F.(2020) „Can we predict the occurrence of COVID-19 cases? Considerations using a simple model of growth“, Science of the Total Environment, 728 (2020), 138834.

Chen, F-M. et al. (2020) „Big data integration and analytics to prevent a potential hospital outbreak of COVID-19 in Taiwan“, Journal of Microbiology, Immunology and Infection, https://doi.org/10.1016/j.jmii.2020.04.010.

Congressional Research Service (2020) “Global Economic Effects of COVID-19”, Available at: https://fas.org/sgp/crs/row/R46270.pdf [Accessed: May 1, 2020]

EU Open Data Portal (2020). COVID-19 Coronavirus data [online]. Available at: <https://data.europa.eu/euodp/en/data/dataset/covid-19-coronavirus-data/resource/55e8f966-d5c8-438e-85bc-c7a5a26f4863> [Accessed: May 1, 2020]

Gu, C. et al. (2020) „The inflection point about COVID-19 may have passed“, Science Bulletin, https://doi.org/10.1016/j.scib.2020.02.025.

Hu, Z. et al. (2020) “Artificial Intelligence Forecasting of Covid-19 in China”. Available at:https://www.researchgate.net/publication/339324015_Artificial_Intelligence_Forecasting_of_Covid-19_in_China [Accessed: May 1, 2020]

Kannan, K. S. et al. (2010) “Financial Stock Market Forecast using Data Mining Techniques”, Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol. I, IMECS 2010, March 17-19, 2010, Hong Kong. Available at: <http://www.iaeng.org/publication/IMECS2010/IMECS2010_pp555-559.pdf> [Accessed: May 1, 2020]

Kulkarni, A. D. & More, A. (2016) “Formulation of a Prediction Index with the Help of WEKA Tool for Guiding the Stock Market Investors”, Oriental journal of Computer science & technology, Vol.9, No.3, pp. 212-225.

Lewis, C. D. (1982) “Industrial and Business Forecasting Methods”, Butterworths Publishing, London, 40.

Li, L. et al (2020) „Propagation analysis and prediction of the COVID-19“, Infectious Disease Modelling, Vol. 5, (2020) 282e292.

Luo, J. (2020) “Predictive monitoring of Covid-19”, Available at: <https://ddi.sutd.edu.sg/> [Accessed: May 1, 2020]

Mosavi, A. et al (2020) “Covid-19 Outbreak Prediction with Machine Learning”. Available at: <https://www.researchgate.net/publication/340782507_COVID-19_Outbreak_Prediction_with_Machine_Learning> [Accessed: May 1, 2020]

Popescu, M. et al. (2009) “Multilayer Perceptron and Neural Networks”, WSEAS Transactions on Circuits and Systems, Vol. 8, No. 7, pp. 579-588.

Rasmussen, C. E., Williams, C. K. I. (2006) “Gaussian Processes for Machine Learning”. MIT Press.

Saad Mohamed, T. (2020) “Heart Diseases Prediction Using WEKA”, Journal of Baghdad College of Economic Sciences Issue, No. 58, doi: 10.13140/RG.2.2.15160.08960.

Selvi, K. et al. (2018) “Classification and Prediction of Diabetics using Weka and Hive Tool”, International Journal of Advance Engineering and Research Development, Vol. 5, Issue 4, pp. 105-111.

Shakil, K. A., Anis, S., Alam, M. (2015) “Dengue disease prediction using Weka data mining tool”, *ArXiv, abs/1502.05167*.

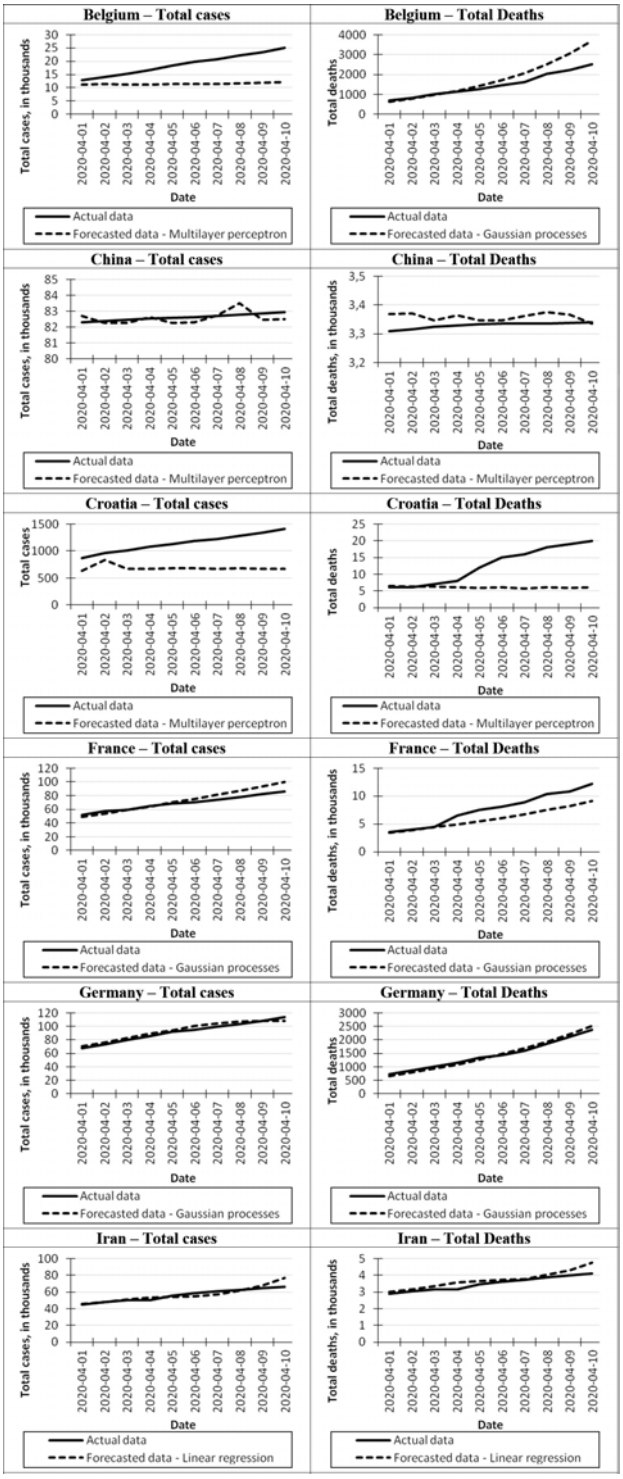
Smola, A. J., Schölkopf, B. (2004) “A tutorial on support vector regression”, Statistics and Computing, Vol. 14, pp. 199-222.

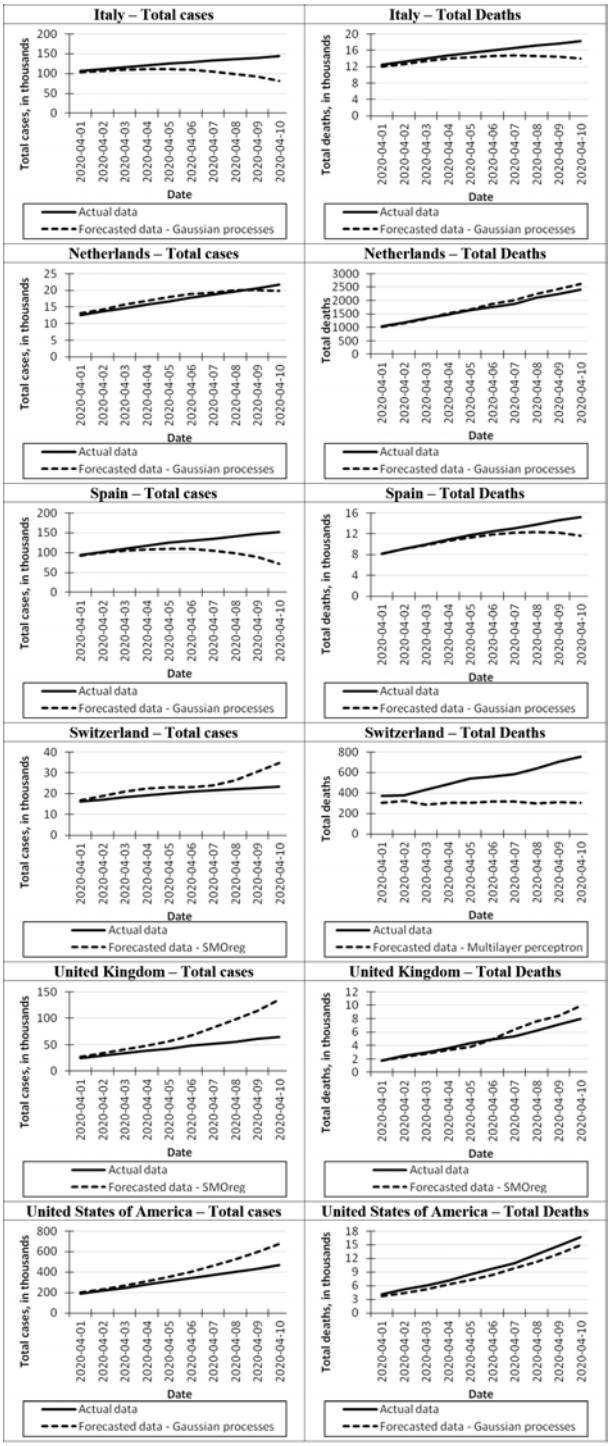
University of Waikato (2020) “WEKA: The workbench for machine learning”. Available at: <https://www.cs.waikato.ac.nz/ml/weka/> [Accessed: May 1, 2020]

Zhang, X., Renjun, M., Lin W. (2020) “Predicting turning point, duration and attack rate of COVID-19 outbreaks in major Western countries”, Chaos, Solitons and Fractals, 135, 109829.

APPENDIX

Figure A1: Actual and forecasted values of the observed variables for the selected 12 countries in the period from April 1, 2020 to April 10, 2020





Source: EU Open Data Portal (2020), authors.