**A machine learning methodology for forecasting of the COVID-19 cases in India**

***R. Sujatha***

***Associate Professor, Vellore Institute of Technology, Vellore, India***

***Email:*** [***sujarada@gmail.com***](mailto:sujarada@gmail.com)

***Jyotir Moy Chatterjee***

***Assistant Professor, Lord Buddha Education Foundation, Kathmandu, Nepal***

***Email:*** [***jyotirchatterjee@gmail.com***](mailto:jyotirchatterjee@gmail.com)

***Aboul Ella Hassanien***

***Faculty of Computers and Artificial Intellgence, Cairo University & Scientific Research Group in Egypt (SRGE)***

***Email: aboitcairo@gmail.com***

**Abstract:** Coronavirus disease (COVID-19) is an inflammation disease from a new virus. The disease causes respiratory ailment (like influenza) with manifestations, for example, cold, cough & fever, and in progressively serious cases, the problem in breathing. COVID-2019 has been perceived as a worldwide pandemic and a few examinations are being led utilizing different numerical models to anticipate the likely advancement of this pestilence. These numerical models dependent on different factors and investigations are dependent upon potential inclination. Here, we propose a basic model that could be helpful to foresee the spread of COVID-2019. We performed linear regression, Multilayer perceptron and Vector autoregression model for expectation on the COVID-19 kaggle information to anticipate the epidemiological pattern of the disease and rate of COVID-2019 cases in India. Predicted the possible trends of COVID-19 impacts in India based on data collected from Kaggle. With the prevailing data about confirmed, recovered and death across India for over the time duration helps in predicting and forecasting the near future. For additional examination or future point of view, case definition and information assortment must be kept up continuously.

**Keywords:** COVID-19, Prediction, Linear Regression (LR), Multilayer Perceptron (MLP) & Vector Autoregression (VAR).

1. **Introduction**

As of date confirmed COVID-19 cases across the globe are 1,498,833 [29] and mortality approximately 5.8%. Gradually the mortality rate is increasing and it’s an alarming factor for the whole world. Transmission is categorized into 4 stages based on the mode of spread and time. Every nation imposed different methodologies starting from staying in-home, using masks, travel restrictions, avoiding social gatherings, frequently washing hands and sanitizing the places often in the case of a common effort to combat the outbreak of this disease. Many countries imposed a lockdown state that prevents the movement of the citizens unnecessarily. Due to this social distancing factor and movement restrictions, the wellbeing and economy of the various nations are being under jeopardy. GPD of the entire world dropped drastically. When the person is found infected, he is isolated and treatment is given for recovery. But based on the severity it will cause death and also people left with a higher level of depression.

In India, the outbreak of coronavirus as disturbed the functioning of life as a whole. all were pushed to stay back to safeguard from the dreadful transmission. In the initial stages, the confirmed cases are those returned from oversees followed by transmission via local transmission. More caution is given to elderly and immunity less people. The demographic of the infected people in India indicates that 39 years is the median. Comparatively people between 21 to 40 years are being affected more. The everyday predominance information of COVID-2019 from January 22, 2020, to April 6, 2020, was gathered from the website of Kaggle [26] and Weka 3.8.4 [27] & Orange [28] is utilized to decipher the information. LR, MLP, and VAR are applied on the Kaggle dataset having 76 instances for anticipating the future effects of COVID-19 pandemic in India. Forecasting is the need of an hour that helps to device a better strategy to tackle this crucial hour across the globe because of this infectious disease. As mentioned by the visual capitalist, the human race as crossed several outbreaks because of the several microbes that were invisible and invincible. COVID-19 is the current threat in the highly sophisticated 21st century. Figure 1 is a snapshot of the visual capitalist [12].

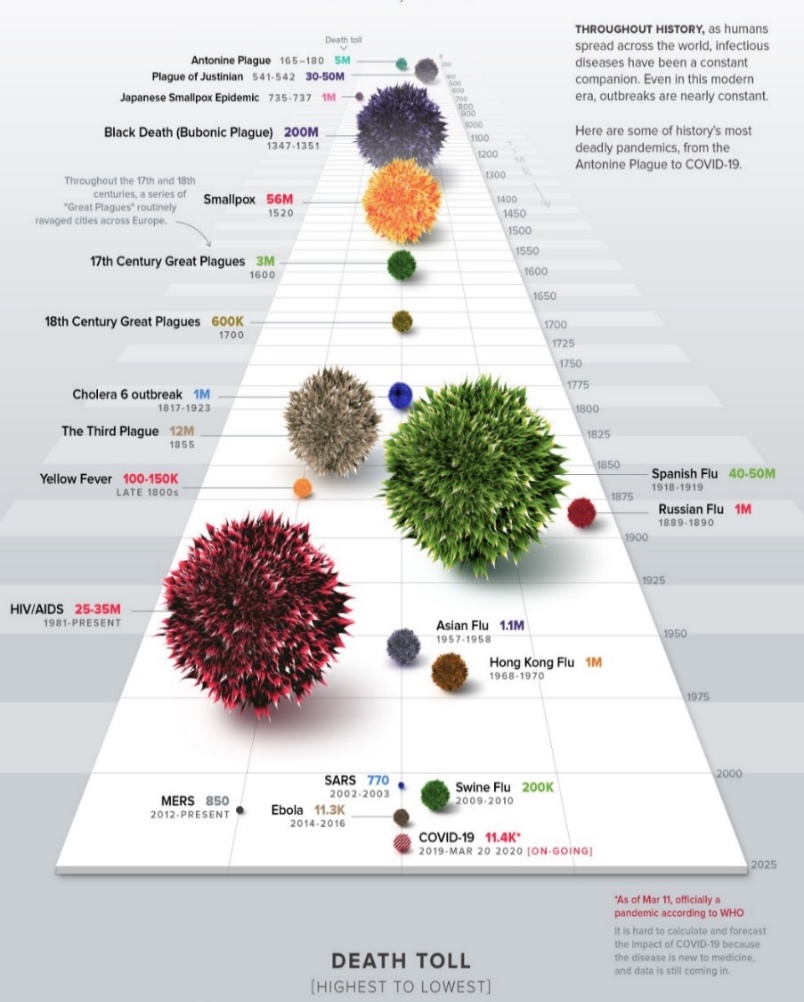


Fig. 1. History of pandemic [12]

1. **Methods and Materials**

In statistics, **Linear Regression (LR)** is a direct way to deal with demonstrating the connection between a dependent variable and at least one independent variable. LR was the main kind of regression analysis to be concentrated thoroughly and to be utilized widely in useful applications [1]. LR shows the connection between two variables by fitting a straight condition to based information. One variable is viewed as an independent and the other is viewed as a dependent. An LR line has a condition of the structure:

(1)

here X is the independent & Y is the dependent variable. The slope of the line is b & a is the intercept (the value of y when x = 0) [2].

**A multilayer perceptron (MLP)** is a type of feedforward artificial neural network (FANN). The term MLP is utilized vaguely, now and then freely to indicate any FANN, now and then carefully to allude to systems made out of various layers of perceptron (with threshold activation) [3]. An MLP is a perceptron that is generally used for complex issues. The formula for MLP [4] is:

(2)

here **w** is for the vector of weights, **x** is for the vector of inputs, **b** is for bias & phi are the non-linear activation function (Neural Network Concepts, n.d.)

**Vector Autoregression (VAR)** is a prediction calculation which is utilized when at least two-time series impact one another, i.e., the connection between the time arrangements included is bi-directional [5]. The formula for MLP is:

(3)

where α is the intercept, a constant & β1, β2 till βp are the coefficients of the lags of Y till order p.

Order ‘p’ means, up to p-lags of Y is utilized & they are the predictors in the equation. The εt is the error considered as white noise.

1. **Experimental Results**

The structure of data based on date, confirmed, recovered and death are shown in Figure 2 with the boxplots and its very clear that several cases are in so primitive stages. As mentioned by WHO, right now India is in the second phase indicating very few cases and forecast of this same is the potential work that is required at this juncture [6-8].

|  |
| --- |
|  |
|  |
|  |
|  |

Fig. 2. Boxplot of INDIA COVID-19

Sieve diagram provides the visualization of the dataset along with that showing the sieve rank. Figure 3 illustrates attributes that have a strong relationship with the dark shades. The interestingness of the pair of attributes is represented via this contingency table. It's a very graphical way of frequency visualization.

|  |  |
| --- | --- |
|  |  |

Fig. 3. Sieve Diagram for INDIA COVID-19

|  |  |
| --- | --- |
|  |  |

Fig. 4. Pearson and Spearman correlation

Correlation plays a great role in finding the dependency among the features of the dataset. Our dataset revolves around the confirmed, recovered and death of cases because of the COVID-19 outbreak over the time frame of around 2 months in India. From the Spearman correlation, it's very evident that based on progressive of the day (date) the possibility of getting prone to sickness is very high and that is given with the +0.949 correlation value. Figure 4 provides a glance at the correlation between Pearson and the spearman process. Appreciably the date attribute is holding a higher level of importance and that’s is reason globally the measures have been taken for social distancing [9-11]. Normally the spread happens just in contact with the person by a handshake is the big brother in case of COVID-19. Correlation provides the signal about the impact and necessary countermeasures to be taken into consideration. Across the globe, leaders of the nation are carrying out various trial and error methods to combat the seriousness of the disease.

Forecasting gives pertinent and consistent input about the past, present and future happenings with certain statistical and scientific approaches. Helps in string decision making in all perspectives. Broadly classified into qualitative and quantitative approaches. Steps involved in forecasting is the deciding factor of the task. Initial understanding of the problem with complete analysis, making a strong foundation, collecting data based on the previous two steps followed by future estimation. Comparison between actual and estimated with followup actions. Various applications like economic & sales prediction, budget, census & stock market analysis, yield projections & many more fields. The medical field also a potential area to deploy the forecast and predication to serve the number of people in need [13-16]. Our work carried out with linear regression, multilayer perceptron, and VAR model over the time series dataset to provide the forecast.

VAR model is a more suitable analysis model in the multivariate time series. It helps in inferencing and analysis of policy. It is used more in a practical forecasting scenario but it is hading superior forecasting performance. Technically narrating about the VAR, it is an m-equation, m-variable model in which individual variable explains on its own based on current, past values. Various parameters of VAR begins with maximum auto-regression order. Various information criteria that help in optimize autoregressive order are Akaike's information criterion (AIC), Bayesian information criterion (BIC), Hann-Quinn & Final prediction error (FPE). By adding and varying trends from constant, linear and quadratic with forecast steps ahead & confidence intervals [17-23]. The formula for calculating AIC, BIC & Hic is as follows:

(4)

(5)

(6)

Where n is the number of attributes in the system, X is the sample size, and is an estimate of the covariance matrix.

1. **Sitational Forcast in India**

{\displaystyle \Sigma }In our forecast work the maximum auto-regression order of 6 followed by an average of information criterion is used for visualization. The trend of constant, linear and quadratic along with 10 steps ahead and 95% confidence interval is introduced [24-25].

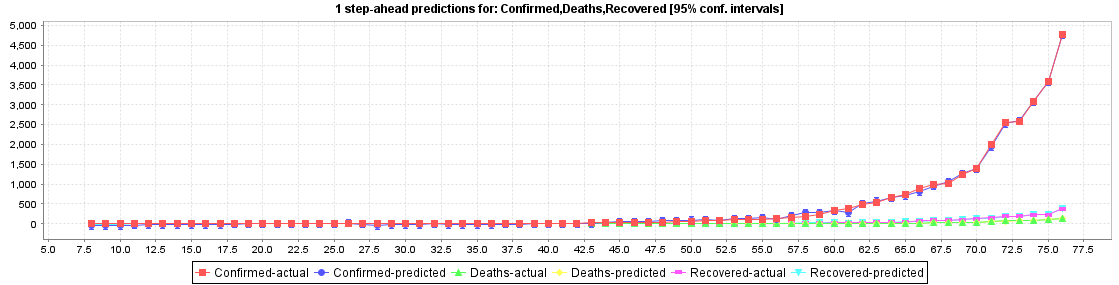


Fig. 5. Prediction of Confirmed cases, Deaths & Recovered case of COVID-19 in India

Figure 5 shows the COVID-19 predicted confirmed cases; death cases & recovered cases based on actual confirmed, death & recovered data with a 95% confidence interval with LR.

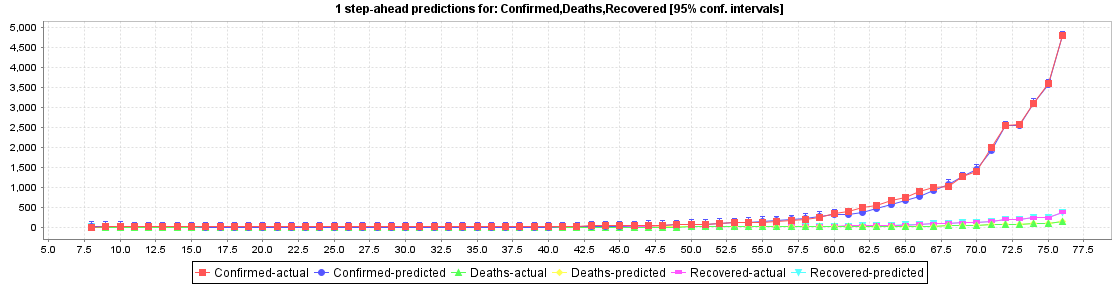


Fig. 6. Prediction of Confirmed cases, Deaths & Recovered case of COVID-19 in India

Figure 6 shows the COVID-19 predicted confirmed cases; death cases & recovered cases based on actual confirmed, death & recovered data with a 95% confidence interval with MLP.

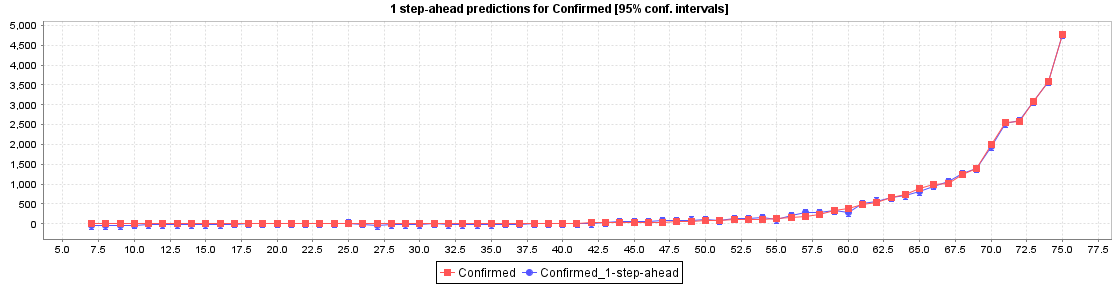


Fig. 7. Prediction of Confirmed cases, Deaths & Recovered case of COVID-19 in India

Figure 7 shows the predicted confirmed cases based on the actual confirmed case data with a 95% confidence interval with LR.

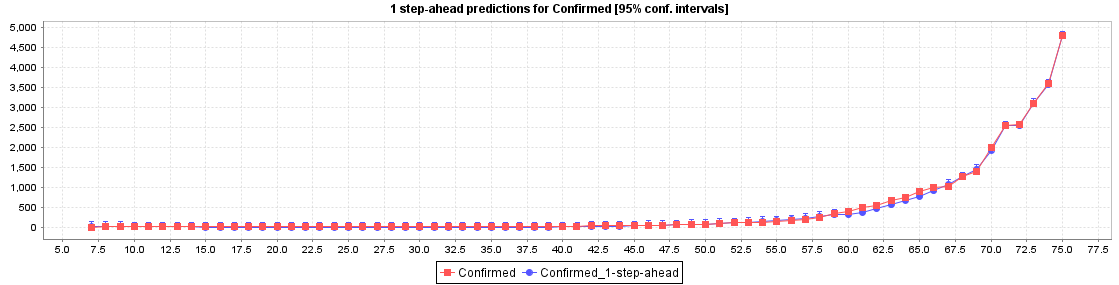


Fig. 8. Prediction of Confirmed cases, Deaths & Recovered case of COVID-19 in India

Figure 8 shows the predicted confirmed cases based on the actual confirmed case data with a 95% confidence interval with MLP.

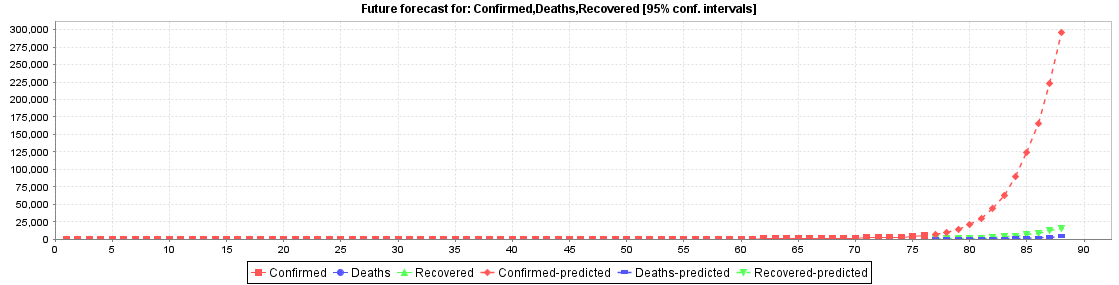


Fig. 9. Prediction of Confirmed cases, Deaths & Recovered case of COVID-19 in India

Figure 9 predicts the impacts of COVID-19 based on the actual data of confirmed, death & recovered cases with LR.

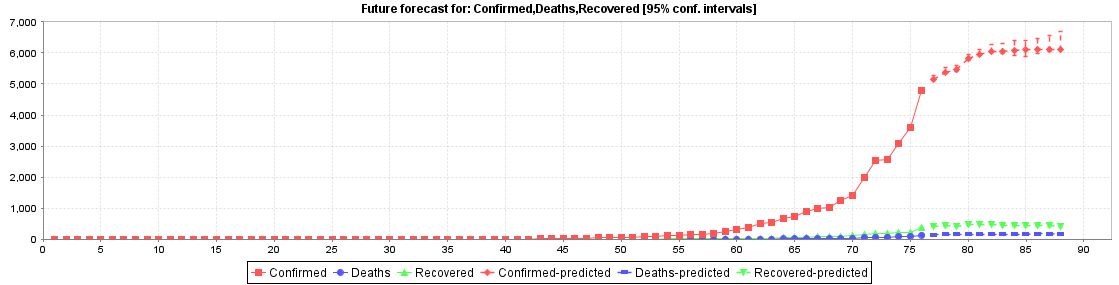


Fig. 10. Prediction of Confirmed cases, Deaths & Recovered case of COVID-19 in India

Figure 10 predicts the impacts of COVID-19 based on the actual data of confirmed, death & recovered cases with MLP.

Table 1. Prediction based on LR

|  |  |  |  |
| --- | --- | --- | --- |
| No of Day | Confirmed | Deaths | Recovered |
| 77\* | 7437.4294 | 184.796 | 484.8578 |
| 78\* | 9935.9238 | 226.1435 | 631.3209 |
| 79\* | 14510.5043 | 279.2629 | 869.2275 |
| 80\* | 20962.0472 | 347.6713 | 1190.1726 |
| 81\* | 29690.4997 | 448.2866 | 1697.5058 |
| 82\* | 44802.8252 | 655.4222 | 2455.8102 |
| 83\* | 63111.0739 | 879.288 | 3330.8007 |
| 84\* | 89304.5137 | 1241.1112 | 4756.0492 |
| 85\* | 124382.5201 | 1678.9316 | 6565.868 |
| 86\* | 165330.8889 | 2190.4666 | 8964.5985 |
| 87\* | 222928.4303 | 3003.2659 | 12259.1426 |
| 88\* | 296072.0567 | 4094.9869 | 16136.5664 |

We have given data of cases till the 76th day i.e. 6th April 2020. Table 1 shows the predicted values of cases (confirmed, death, recovered) by using LR from the 77th day i.e. 7th April 2020 for the next 10 days, i.e. 18th April 2020.

Table 2. Prediction based on MLP

|  |  |  |  |
| --- | --- | --- | --- |
| No of Day | Confirmed | Deaths | Recovered |
| 77\* | 5156.0153 | 149.2294 | 402.7235 |
| 78\* | 5374.8583 | 155.9816 | 430.1433 |
| 79\* | 5449.6853 | 157.3286 | 427.3714 |
| 80\* | 5815.5954 | 166.0541 | 461.858 |
| 81\* | 5957.2631 | 169.6262 | 480.9021 |
| 82\* | 6039.2467 | 170.65 | 484.1291 |
| 83\* | 6049.9342 | 170.7711 | 449.5301 |
| 84\* | 6070.8264 | 170.8924 | 431.3646 |
| 85\* | 6099.3429 | 171.3597 | 442.9075 |
| 86\* | 6120.7886 | 171.6982 | 455.8906 |
| 87\* | 6115.4564 | 171.5553 | 437.826 |
| 88\* | 6101.6622 | 171.215 | 419.0384 |

We have given data of cases till the 76th day i.e. 6th April 2020. Table 2 shows the predicted values of cases (confirmed, death, recovered) by using MLP from the 77th day i.e. 7th April 2020 for the next 10 days, i.e. 18th April 2020.

1. **Conclusion**

Information and communication technology help in the decision-making process based on the past data with the data analytics and data mining perspective. The size of data available is huge and gathering information and getting an interesting pattern out of the cumulated data is a challenging task. With the prevailing data about confirmed, recovered and death across India for over the time duration helps in predicting and forecasting the near future. The correctness of the model could be increased by introducing related attributes like several hospitals, the immune system of the infected person, age of the patient, gender of the patient, steps taken to combat the proliferation of the virus and so on to make it completely informative. As of now, it's very prudent that yards to carry needs to be stringent and vigil in nature to handle this crucial situation by social distancing, lockdown, curfew, quarantine and isolation to prevent the transmission. By seeing the predicted values and matching with cases from John Hopkins University [29] data we can conclude that the MLP method is giving good prediction results than that of the LR method using WEKA.

**References**

1. Yan, X., & Su, X. (2009). *Linear regression analysis: theory and computing*. World Scientific.
2. <http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm>
3. <https://en.wikipedia.org/wiki/Multilayer_perceptron>
4. <https://missinglink.ai/guides/neural-network-concepts/perceptrons-and-multi-layer-perceptrons-the-artificial-neuron-at-the-core-of-deep-learning/>
5. <https://www.machinelearningplus.com/time-series/vector-autoregression-examples-python/>
6. Tareen, A. D. K., Nadeem, M. S. A., Kearfott, K. J., Abbas, K., Khawaja, M. A., & Rafique, M. (2019). Descriptive analysis and earthquake prediction using boxplot interpretation of soil radon time-series data. *Applied radiation and isotopes*, *154*, 108861.
7. Ferreira, J. E. V., Pinheiro, M. T. S., dos Santos, W. R. S., & da Silva Maia, R. (2016). Graphical representation of chemical periodicity of main elements through boxplot. *Educación química*, *27*(3), 209-216.
8. Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational statistics & data analysis*, *52*(12), 5186-5201.
9. Mu, Y., Liu, X., & Wang, L. (2018). A Pearson’s correlation coefficient-based decision tree and its parallel implementation. *Information Sciences*, *435*, 40-58.
10. Zhou, H., Deng, Z., Xia, Y., & Fu, M. (2016). A new sampling method in particle filter based on the Pearson correlation coefficient. *Neurocomputing*, *216*, 208-215.
11. Gautheir, T. D. (2001). Detecting trends using Spearman's rank correlation coefficient. *Environmental Forensics*, *2*(4), 359-362.
12. <https://www.visualcapitalist.com/history-of-pandemics-deadliest/>
13. Hajirahimi, Z., & Khashei, M. (2019). Hybrid structures in time series modeling and forecasting: A review. *Engineering Applications of Artificial Intelligence*, *86*, 83-106.
14. Yamana, T. K., & Shaman, J. (2019). A framework for evaluating the effects of observational type and quality on vector-borne disease forecast. *Epidemics*, 100359.
15. Monteiro, C., Ramirez-Rosado, I. J., Fernandez-Jimenez, L. A., & Ribeiro, M. (2018). New probabilistic price forecasting models: Application to the Iberian electricity market. *International Journal of Electrical Power & Energy Systems*, *103*, 483-496.
16. Potočnik, P., Šilc, J., & Papa, G. (2019). A comparison of models for forecasting the residential natural gas demand of an urban area. *Energy*, *167*, 511-522.
17. Billio, M., Casarin, R., & Rossini, L. (2019). Bayesian nonparametric sparse VAR models. *Journal of Econometrics*, *212*(1), 97-115.
18. Chiu, C. W. J., Mumtaz, H., & Pinter, G. (2017). Forecasting with VAR models: Fat tails and stochastic volatility. *International Journal of Forecasting*, *33*(4), 1124-1143.
19. Portet, S. (2020). A primer on the model selection using the Akaike Information Criterion. *Infectious Disease Modelling*.
20. Snipes, M., & Taylor, D. C. (2014). Model selection and Akaike Information Criteria: An example from wine ratings and prices. *Wine Economics and Policy*, *3*(1), 3-9.
21. Dirick, L., Claeskens, G., & Baesens, B. (2015). An Akaike information criterion for multiple event mixture cure models. *European Journal of Operational Research*, *241*(2), 449-457.
22. Zanini, A., & Woodbury, A. D. (2016). Contaminant source reconstruction by empirical Bayes and Akaike's Bayesian Information Criterion. *Journal of contaminant hydrology*, *185*, 74-86.
23. Zhang, P., & Krieger, A. M. (1993). Appropriate penalties in the final prediction error criterion: a decision-theoretic approach. *Statistics & probability letters*, *18*(3), 169-177.
24. Efendi, R., Arbaiy, N., & Deris, M. M. (2018). A new procedure in stock market forecasting based on a fuzzy random auto-regression time series model. *Information Sciences*, *441*, 113-132.
25. Tapia, J. A., Salvador, B., & Rodríguez, J. M. (2020). Data envelopment analysis with estimated output data: Confidence intervals efficiency. *Measurement*, *152*, 107364.
26. <https://www.kaggle.com/imdevskp/corona-virus-report/data>
27. <https://sourceforge.net/projects/weka/>
28. <https://orange.biolab.si/>
29. <https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6>