MATHEMATICAL MODELING OF COVID-19 SPREAD USING GENETIC PROGRAMMING ALGORITHM

Leo Benolić[1], Anđela Blagojević[1,2], Tijana Šušteršič[1,2], Zlatan Car[3] and Nenad Filipović[1,2]

[1] Bioengineering Research and Development Centre (BioIRC), Prvoslava Stojanovica 6, 34000 Kragujevac, Serbia
[2] Faculty of Engineering, University of Kragujevac, Sestre Janjić 6, 34000 Kragujevac
[3] Faculty of Engineering, University of Rijeka, Vukovarska 58, 51000 Rijeka, Croatia
e-mail: leo.benolic@kg.ac.rs, andjela.blagojevic@kg.ac.rs, tijanas@kg.ac.rs, car@riteh.hr, fica@kg.ac.rs

Abstract:

This paper analyses the possibilities of using Machine learning to develop a forecasting model for COVID-19 with a publicly available dataset from the Johns Hopkins University COVID-19 Data Repository with the addition of the percentage of each variant from the GISAID Variant database. The Genetic programming (GP) symbolic regressor algorithm is used for the estimation of new confirmed cases, hospitalized cases, cases in intensive care units (ICUs), and the number of deaths. This metaheuristics method algorithm is made from a dataset for Austria and its neighboring countries the Czech Republic, Slovenia, and Slovakia. Machine learning was performed twice to create individual models for each country, but the second time the process covered all countries at once as a multi-country model. Variance-based sensitivity analysis was initiated using the obtained mathematical models. This analysis showed us on which input variables the output of the obtained models is sensitive, like in case of how much each covid variant affects the spreading of the virus or the number of deaths. Individual short-term models show very high R2 scores, while long-term predictions have lower R2 scores. The multi-country model achieved inferior results as additional valuables needed to be added in order to obtain better results.

Keywords: artificial intelligence, COVID-19, genetic programming, mathematical prediction models, variants

1. Introduction

The COVID-19 pandemic started spreading in 2019 in Wuhan, China, and within 3 months, it began to spread to every province of mainland China, and eventually it reached the other 27 states [1]. The virus comes from the coronavirus family of diseases that come from bats, it was renamed to COVID-19 to distinguish it from the coronavirus Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS). It is worth mentioning that SARS and MERS share 79% and 50% of the genome sequence respectively [2]. Due to the characteristics of the virus, after more than two years, we have over 500 000 000 cumulative confirmed cases of corona [1]. The last discovered variant is omicron, which has higher transmissibility but decreased level of severity and mortality. In Europe, most countries have 20000-50000 cumulative confirmed cases per million, which means that almost half of the people were naturally exposed to the virus and 75% of the vaccinated people actually had the virus[3]. After the Omicron variant subsided, many of EU countries lifted the Covid measures.. Although the interest inthe coronavirus is currently dropping, collected data of the virus should

be used as much as possible so that we can better prepare for similar threats in the future. Creating an epidemiological model might be a good idea as well.

This paper investigates the possibility of using the machine learning technique, more precisely, the symbolic regression genetic programming GP algorithm. This research is based on the COVID-19 dataset from Austria and its neighboring countries, the Czech Republic, Hungary, Slovakia, and Slovenia. Using time-series variables as input data for obtaining a predictive model of the future state in the form of a mathematical equation of new confirmed cases, hospitalized, cases in intensive care unit, and deaths. The goal is to obtain satisfactory accuracy for a longer period of the time that could be used to plan lockdowns and increase the capacity of covid hospitals. Also, the goal is to analyze the importance of input model variables such as the percentage of each covid variant.

2. Materials and methods

In order to facilitate and accelerate machine learning, characteristics of the spread and behavior of viruses over time are necessary. A paper by Wang, F. et al. (2020) has defined the timeline of covid cases in the first month after its discovery, and this will be useful for adjusting the input to facilitate and speed up the machine learning [4]. For ML, due to its standardized form , the dataset will be used from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [3]. Also, the "GISAID EpiCoV" database, where they tested the percentage of each variant. It was necessary to insert the percentage of each variant, and an input is needed to increase the accuracy of the model because each variant has its own characteristics transmissibility, hospitalization, and mortality rates[5].

The GP algorithm, is selected for the ML learning method, as the GP algorithm is a metaheuristic method inspired by Charles Darwin's theory of natural evolution [6].

The GP symbolic regressor is able to create a symbolic mathematical function that best describes the given data. The mathematical function is written in the form of a tree where the functions are nodes and the sheets are variables or constants. The nodes and leaves are primarily obtained randomly, they are altered by the process of mutation reproduction and crossover. After the implementation of the genetic operation, the offspring population is evaluated to assess the quality of the results and to select the best results that will participate in the next iteration of the genetic algorithm. What is phenomenal about the GP is that the result is universal, generally understandable and can be easily transferred to another program/environment.

2. Results and discussion

Figure 1 shows the real number of deceased cases (blue curve) and the estimation of the deceased cases (the orange curve) for a period of 14 days. Both curves are similar in nature, where the number of deceased cases (blue curve) has more fluctuation than the estimation curve (orange curve), thus the algorithm precisely estimates the number of deceased cases.

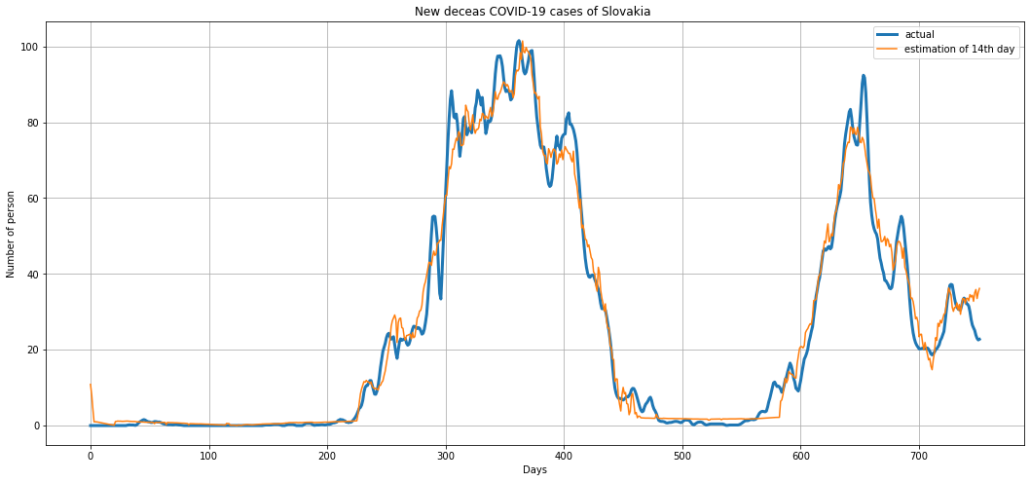

Fig. 1. Estimation of new deceased cases for Slovakia 14th day

The importance of the input variables of the model will be analyzed on the obtained mathematical expression using the sobol sensitivity analysis package from sib.lib in Python. The program will show which input variables most influence the output of the obtained mathematical model. It will also show which input variables have the greatest impact on the output of the obtained mathematical model. The Sensitivity of the model (Figure 1) to omicron, delta, alpha and other variants. (Other variants include all the variants at the beginning of the pandemic until the moment the scientists began testing an separating them) It is as follows: the mean sobol index for other variant percentage is 0.48534, followed by delta (B.1.617.2) with 0.04045 and alpha (B.1.1.7) with 0.00057 while the omicron does not affect the output and its sobol index is 0. Ther Sobol index which depends on the number of newly infected cases is shown in Figure 2.

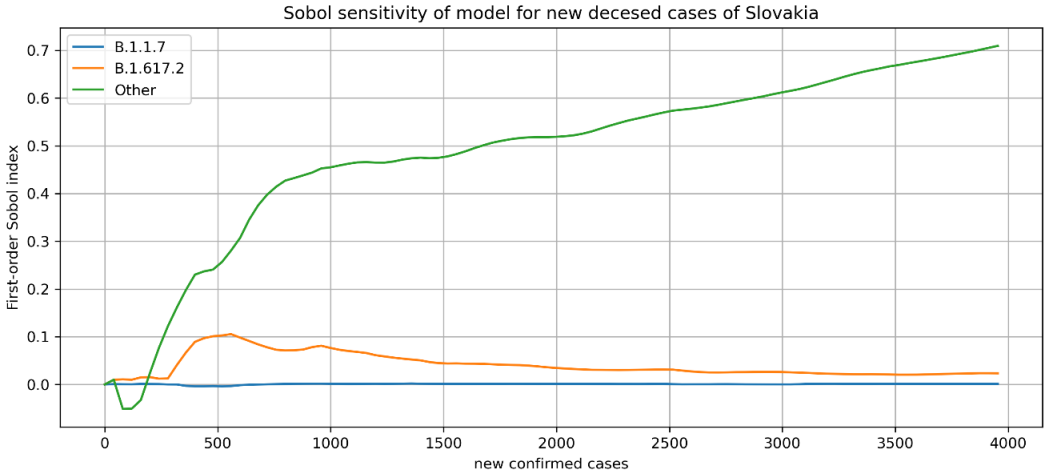

Fig. 2. Sobol sensitivity of estimation model for new deceased cases for Slovakia

The result is understandable. The omicron variant does not affect the output because it has a lower mortality rate, but it is quite strange in the cases of delta and alpha due to the higher mortality rate. The results show the highest correlation with other variants probably due to excess mortality in the beginning when the countries were not ready and they did not know how to approach the patients. In these cases, it was assumed that mortality related to the other variant.

3. Conclusions

In this paper, a GP algorithm was used to develop a model for estimating new confirmed COVID cases, hospitalized cases, intensive care unit cases, and deaths using data available online. Each individual country model has reached high accuracy with high Coefficient determination. In the case of multi-country model solutions, they have inferior results compared to when we looked at the countries individually. This is especially evident for Slovenia, which shows unsatisfactory results compared to the other countries. The reason for such results is probably due to the fact that the countries differ in demographics and other factors, also, the data used to get such results was quite limited.

Acknowledgment: This research is supported by the project that has received funding from the European Union's Horizon 2020 research and innovation programmes under grant agreement No 952603 (SGABU project). This article reflects only the author's view. The Commission is not responsible for any use that may be made of the information it contains. T. Šušteršič also acknowledges the support from L'OREAL-UNESCO "For Women in Science" National Fellowship program Serbia (2021 Fellows).

References

[1] WHO, 'COVID-19 Situation reports', [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports Accessed 30/03/2022.
[2] Hu, B., Guo, H., Zhou, P., & Shi, Z. L. (2021). Characteristics of SARS-CoV-2 and COVID-19. Nature Reviews Microbiology, 19(3), 141-154.
[3] COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, https://github.com/CSSEGISandData/COVID-19 Accessed 30/03/2022.
[4] Wang, F., Qu, M., Zhou, X., Zhao, K., Lai, C., Tang, Q., & Liu, L. (2020). The timeline and risk factors of clinical progression of COVID-19 in Shenzhen, China. *Journal of Translational Medicine*, *18*(1), 1-11.
[5] GISAID, https://www.gisaid.org/hcov19-variants/ Accessed 30/03/2022
[6] De Jong, K. (1988). Learning with genetic algorithms: An overview. Machine learning, 3(2), 121-138.