Data Science Driven Analysis of the Impact of COVID-19

Șerban Gabriel TIMOFTE[1], Ioana Ruxandra STROE[2], Daniel TAȘCU[3],
Raluca Cristina TERTEȘ[4], Radu Ioan MOGOȘ
[1,2,3,4,5] The Bucharest University of Economic Studies, Romania
[1]EBR Net
timoftesrbn@gmail.com, stroeioana22@stud.ase.ro, tascudaniel22@stud.ase.ro,
tertescristina22@stud.ase.ro, mogos.radu@gmail.com

In the context of a pandemic that emerged with lightning speed, data science has become a cornerstone for governments decision-making processes. By analyzing numerous centralized databases, researchers have been able to identify trends, the spread of the virus, and run artificial intelligence (AI) simulations to anticipate crucial points of the COVID-19 pandemic. Data warehouses created during this period offer real-time monitoring of the global effects of the virus. The health databases are already common in national systems, but their usefulness rises above storing medical histories. The cross-disciplinary nature of the COVID-19 pandemic accentuates the need for collaboration between doctors, medical specialists and data analysts, data engineers, and Artificial Intelligence engineers. This article provides a comprehensive overview of how databases and data warehouses can offer different scenarios for citizens and health specialists alike.

Keywords: *Data Science, Data Warehouses, COVID-19, Data Analytics*

1 Introduction

Data Science and the health informatics industry are rapidly evolving in the previous years. Health informatics could be described ”as an evolving scientific discipline” [1], serving the numerous interest of the specialists - from the research of rare or dangerous diseases (cancer, Parkinson, Crohn) to the prevention of global medical hazards.

Data Science Driven Analysis represents more than running different algorithms on databases. The most important part of a Data Science project is to use proper data sets with real information. Cleaning data represents the process of identifying and removing errors ”in order to improve data quality” [2]. However, the process is time-consuming and resource-consuming due to the appearance of ”Big Data”. This phenomenon led to the appearance of real-time data analysis. In the health system, Big Data is represented by the digital medical records of the patient.

Artificial Intelligence technology was used during the pandemic years, by official entities. For example, World Health Organization (WHO) created a project called ”WHO Coronavirus Dashboard” which is empowered by artificial intelligence software to provide real-time insights around the globe [3]. European Center for Disease Prevention also created during the pandemic a series of AI models to predict the infectious disease. Their model is based on assumptions and it analyses different databases provided by different entities (authorities, hospitals and laboratories). The ECDC’s models are focused on how people could interact throughout the day. Also, this is the main reason why are necessary more databases provided by different entities [4]

Access to real information to train the AI models or to create custom dashboards about the spreading of the COVID-19 virus is a problem. Over the internet there are many open-source databases with wrong data and the government does not offer databases in different formats (csv, sql or others). The access is restricted to manual database input. These pandemic problems highlighted the importance of international collaboration to create relational databases with real information.

2. Methods

In our research, we set out to identify how Data Science could impact directly the management of the COVID-19 pandemic and how accessible is this kind of technology. We measured the impact by analyzing the most well-known scenarios about the most affected sections of the population using R and datasets with records around the world and by implementing AI predictions using the SARIMAX model. Pursuing the same concept, we created a Tableau dashboard using real-time synchronization with the database.

Our project is created using the methodologies belonging to multiple sciences: database development, data analytics & science, statistics and machine learning. All of these are "under the hat" of quantitative measures in the digital age. In the following sections, we used open-source databases implemented with PostgreSQL (PSQL), taking care of the data cleaning and normalization process.

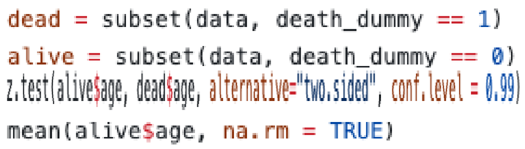
The universality of AI technology is demonstrated by world industries. Nowadays, car manufacturers, marketing companies, health equipment, financial technology (FinTech) and many others. The key factor is the knowledge transfer from human to machine. Having a foundation built by human and the capacity of storing large amounts of information, the AI could support people in their activities as an assistant researcher. The machine is able to "memories" significantly more knowledge than a human being, but the human brain is still the cornerstone of society.

An interesting example of implementing AI in a "less popular segment" is presented by doctors Diana Hintea, James Brusey and Elena Gaura in an article regarding the implementation of different AI models to estimate the cabin occupant equivalent temperature [5].

3. Z Tests

For statistical hypothesis analysis, we used the Z test, managing data for all the countries. The Z test result is calculated by dividing the difference between two sample means by the standard error of the difference. This type of hypothesis testing is a standard way of decision-making or a way to analyze the impact of the virus [7]. We consider the null hypothesis (H0) and the HA which rejects the null hypothesis. Also, our risk for this experiment is noted with L.

The first statistical test we implemented examined the hypothesis that ”Older people died after the infection with SARS-CoV-2”. All the data are stored in a numeric vector.



```
dead = subset(data, death_dummy == 1)
alive = subset(data, death_dummy == 0)
z.test(alive$age, dead$age, alternative="two.sided", conf.level = 0.99)
mean(alive$age, na.rm = TRUE)
```

Fig. 1. Subsets of Test Number 1

In our test, the H0 is represented by ”older people died after SARS-CoV-2 infection”. In this case, the alternative hypothesis is ”there is no connection between age and mortality after SARS-CoV-2 infection”. We used the ”two-sided” alternative because the scope of the test is to identify if the scenario is true in the provided context.

```
z.test(cured$age, deaths$age,
alternative="two.sided", conf.level =
0.99)
```

Fig. 2. Z Test Number 1 in R

The resulting interval is [-25.52122; -15.50661]. Even though the distance between the ends of the interval is far apart, they represent a shred of evidence to confirm the hypothesis. The p-value helps the researcher to determine the statistical significance of the test. In our case, the value of it is 2.2e-16, significantly above 0.05 - the reference value (the alpha from the z.tests function body).

```
z.test(men$deaths_dummy,women$deaths_d
    ummy, alternative="two.sided",
          conf.level = 0.99)
```

Fig. 3. Z Test Number 2 in R

For the second test, the resulting interval is [0,8;8,8]. It shows the risk of the men dying after the infection. The practical interpretation is that men have a 0,8% to 8% higher risk to die after infection. In this case, the p-value is significantly higher than in the previous test. The value is 0.002, but is still lower than the reference value. However, the mortality of the men is 8,5%, while the mortality of women is 3,7%.

These two tests confirm both null hypotheses and we can conclude that the risk is a variable that depends very much on gender and age. Also, in our analysis there is not recorded the impact of the medical records of the patients from the previous years.

4. Summary of the database using Jupyter Notebook

Our objective in Jupyter Notebook was to discover trends of the virus spreads and to view its geographical expansion.

There are two default manners of analyzing trends evolution according to A. Morgan, A. Amed, D. George and M.Hallett:

- ”*Overall positive change* (higher highs and higher lows - expressed as value increase) +1”
- ”*Overall negative change* (lower highs and lower lows - expressed as value decrease) -1” [8]

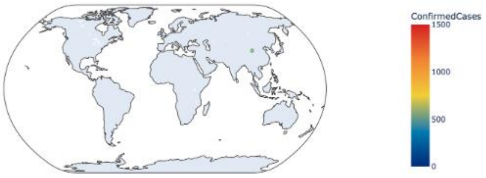


Fig. 4. Spread map in T0

Considering January 2020 as being the start moment of the pandemic period in the world, we have in the Fig. 4 the map of COVID-19 spread in the debut of the pandemic. The main variable of the targeted trend is the number of confirmed cases.

The second and the third key moments visible on the world map created in Jupyter Notebook are 03/03 and 03/29. From these two pictures, we can observe that the spread direction is from East (E) to West (W) and in less than a month, all the Europe’s countries jumped over the superior limit of our analytics.

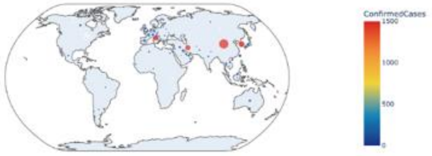


Fig. 5. Spread map in T1 - 03/03

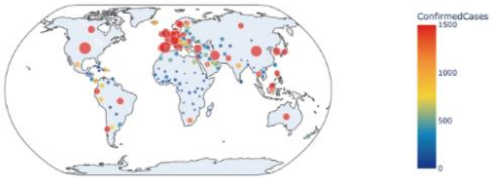


Fig. 6. Spread map in T2 - 03/29

In the Fig. 6, the trend is confirmed spreading the virus all around the globe.

In the generated Jupyter Notebook with the evolution of COVID-19 in the world, the user could easily observe that a crowded area with a high number of people in one square kilometer is affected first.

These observations are obvious looking at a video or trying to remember the first part of 2020. However, these videos and analyses would not have been possible without Big Data or Health Informatics. The speed of development is a high advantage in crisis situations when the authority's reaction should be almost immediate. The data analyses role is to manage all the data flow in real-time to help in the management of the decisions.

A clear illustration of the importance of these technologies in a crisis situation is represented in Fig. 7. To be able to visualize the map from Fig. 4, Fig. 5 and Fig. 6 could prevent the authorities and the population that the situation represented in Fig. 7 is imminent.

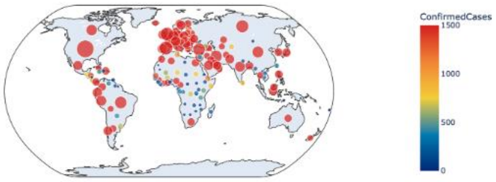


Fig. 7. Spread map in T3 - 05/05

5. Time series

”Time series forecasting models are used to predict the futuristic outcomes based on historical information” [9].

We applied the SARIMAX model to the official datasets of governments about COVID-19 confirmed cases.

Seasonal AutoRegressive Integrated Moving Average with exogenous variables (SARIMAX) is a time-series model based on the popular AutoRegressive Integrated Moving Average (ARIMA) model. The difference is that in the SARIMAX model, the seasonality and exogenous variables are analyzed, too. ”It is interesting to think that all exogenous factors are still technically indirectly modeled in the historical model forecast” [10]. For our analysis, there are four main categories of external data with a significant impact on the prediction accuracy:

- *Geographic & Demographic distribution*
- *The economic power of the state*
- *Urban mobility*
- *Vaccination rate* (this one is not possible to be applied in our research, because at the moment of data recording, there was no vaccine)

In order to apply the SARIMAX model to the COVID-19 dataset, an important step before is a stationary check. The mathematical formula for this verification is that standard deviation is not dependent on time. This is the equivalent of the idea that a shift in the series does not create any change in the shape of its distribution. [11]

The Seasonality component of SARIMAX models refers to events that are not consistent during the time and appears just in shorter periods. Viktor Mehandzhiyski has a practical example of this phenomenon: the” Jingle Bells” song [12].

In the SARIMAX model, there is used a combination of autoregression with seasonality. The actual output is the dependence of the past values of the variables and the moving average (dependence of the variable on forecast error) [13].

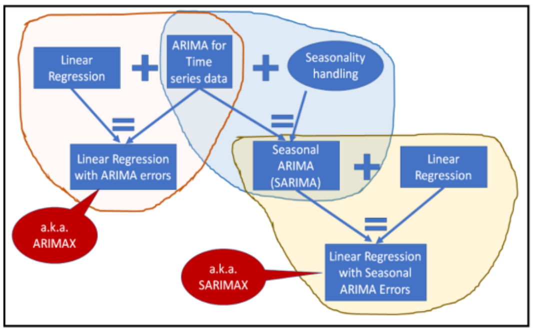


Fig. 8. SARIMAX model components - Time series Analysis, Regression and Forecasting with Python - https://timeseriesreasoning.com/contents/regression-with-arima-errors-model/

In order to apply the SARIMAX model to the COVID-19 dataset, an important step before is a stationary check. The mathematical formula for this verification is that standard deviation is not dependent on time. This is the equivalent of the idea that a shift in the series does not create any change in the shape of its distribution. [11]

The Seasonality component of SARIMAX models refers to events that are not consistent during the time and appears just in shorter periods. Viktor Mehandzhiyski has a practical example of this phenomenon: the” Jingle Bells” song [12].

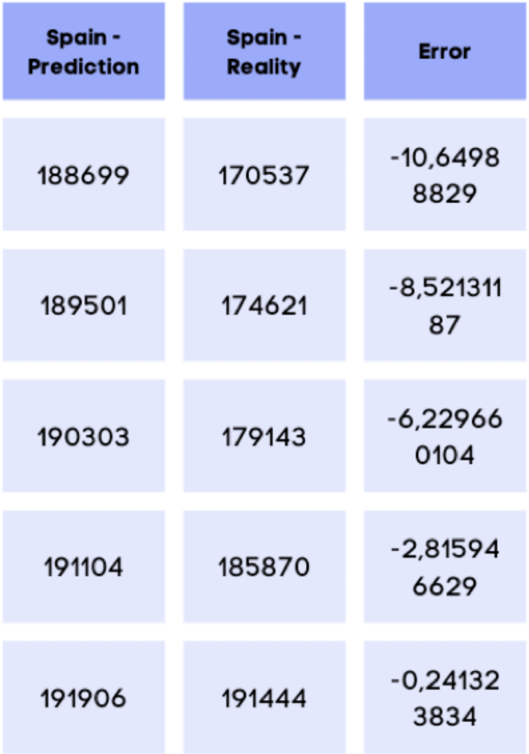
In the SARIMAX model, there is used a combination of autoregression with seasonality. The actual output is the dependence of the past values of the variables and the moving average (dependence of the variable on forecast error) [13].

The most important principle in which the SARIMAX model works is that training data prepare the algorithm for the prediction. However, the accuracy of the model is measured using the computation with the real situations.

After we trained the model with data fromthe first three months of the COVID-19 pandemic in Romania and Spain, we collected the next results (Tabel. 1 for Spain and Tabel. 10 for Romania). The datasets were provided as excels from the official websites of the Ministry of Health in these two countries.

Every row in these tables represents the simulation in a day. There is a prediction on 5 days, storing data from the last three months (90 days). The report is 1/18 (simulated days / analyzed days).

Table 1. Romania's prediction

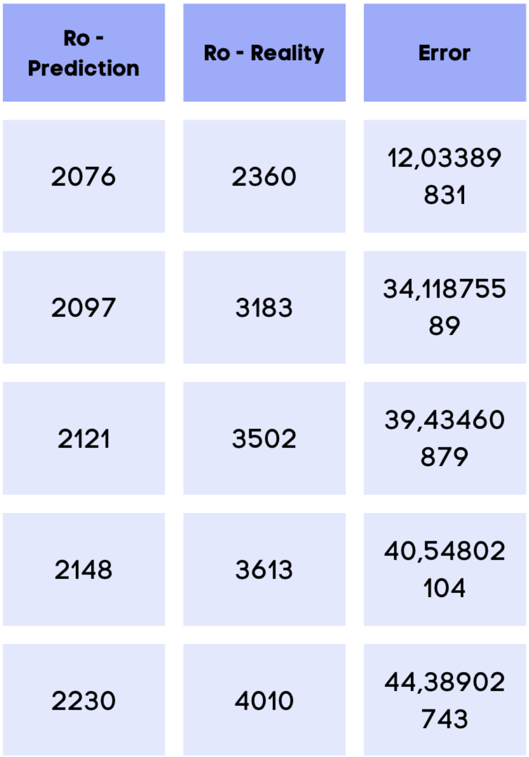


Spain - Prediction	Spain - Reality	Error
188699	170537	-10,64988829
189501	174621	-8,52131187
190303	179143	-6,229660104
191104	185870	-2,815946629
191906	191444	-0,241323834

Intriguingly, the first error values considered in the module are closely (10,65 for Spain and 12,03 for Romania). Looking for other simulations, we observed that the first value of the error in the module is situated between 10% and 15% (Germany 12,43%, Italy 9,8%, France 14,74%). This fact could indicate a standard error of the model caused by the lack of a larger dataset. Also, SARIMAX is a model that takes into consideration historical data and this is another factor or this first error.

I considered these two countries in my example in order to show the importance of a proper dataset when we are running AI-driven simulations.

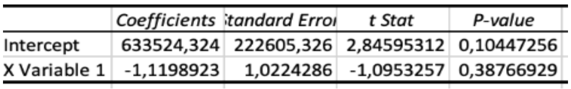
Table 2. Spain's prediction



Ro - Prediction	Ro - Reality	Error
2076	2360	12,03389831
2097	3183	34,11875589
2121	3502	39,43460879
2148	3613	40,54802104
2230	4010	44,38902743

Spain offered a higher database and the testing ration of the local authorities is higher. In the

Spain scenario, SARIMAX identified correctly the increasing trend of confirmed Covid-19 cases. Looking at the Romanian's prediction, the trend is also correctly identified (in the context of the global raising of confirmed cases). However, the speed of raising is totally wrong. After, we started with an error of 10,6498%, on the last simulated day, Spain had the module value of error around 0,2413. In Romania, the first error is 12,033%, but the last one is 44,389%. These differences are caused by improper datasets. However, the speed of raising is totally wrong. After, we started with an error of 10,6498%, on the last simulated day, Spain had the module value of error around 0,2413. In Romania, the first error is 12,033%, but the last one is 44,389%. These differences are caused by improper datasets.



	Coefficients	;tandard Erroı	t Stat	P-value
Intercept	633524,324	222605,326	2,84595312	0,10447256
X Variable 1	-1,1198923	1,0224286	-1,0953257	0,38766929

Fig. 8. Error trend

This shows a significant problem of AI-driven simulations. If the researcher and/or the authorities could not provide real data, the output will contain important mistakes. "If the input data is flawed or
inaccurate, then any results will be similarly problematic. The quality of data used for machine learning is critical for both the accuracy and reliability of the algorithmic output” [14].
To support the importance of transparency in publishing realistic data, we forward the example of China to analyze. The error is constantly -98%.

6. Vaccine results using regression

The regression model is the cornerstone of any prediction model. The SARIMAX model has an autoregression component included. Regression represents the statistical method to identify the correlation between two variables (the independent and the dependent variable) [15, 16]. The objective of this part of the research is to identify the connection between vaccinations and confirmed cases in a time frame of three months.
The analysis dataset incorporates all the available data on the official websites created by Romanian Government. X is called the dependent variable and in our case, it represents the number of confirmed cases per month starting from July 2021. Y (independent variable) means the number of first doses administrated by Romanians starting from May 2021. The correlation is powerful and it is a negative one. The regression index (r) is 0,6123. According to the coefficient of Intercept and X Variable 1, one more vaccinated person could lead to a decrease in confirmed cases with 1,11.

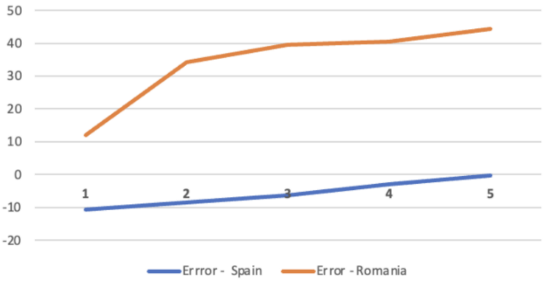


Fig. 9. Regression table (vaccine-confirmed cases)

This result led us to an intermediate conclusion - the vaccine has a small role in stopping the virus spread in Romania, based on the available data.
The resulting regression equation is:

Y = 633524,324 - 1,11989 * X

The Significant F indicator shows us the relevance of the test. In our case, it is approximately 0,3 and shows us that there are not enough data in the set. We can use our results, taking into consideration that are not accurate enough to reflect totally the truth [17].

The second dataset contains the number of vaccines and the number of deaths caused by coronavirus. The timeframe is similar, comparing the number of vaccines from May and the number of deaths in July. The correlation is weaker this time with r = 0,469 (weighted negative connection), but it is still relevant for the macro-analyze. According to the regression coefficients (Intercept and X Variable 1), one more vaccinated person could disease the number of deaths after two months with 33,46 people.
In this second test, the Significant F has still a high value (0,4). The available data are still not enough in Romania to generalize risk about the phenomenon under investigation.
The resulting regression equation is:

Y = 597826,2 - 33,461592 * X

Connecting these two results, it is clear that the variables have a connection and the vaccine directly affects both the confirmed cases and the number of deaths. However, this analysis raised yet another question - for one more vaccinated person, the other 33 people do not die because they get infected with a lighter form or not at all?

7. Results

The theories about how age and gender could impact the mortality of coronavirus are true. Both indicators have an important role in treatment.

Data Analytics and Machine Learning are important and efficient tools for risk management. Looking at the provided graphs and simulations, central authorities could prepare the next steps to prevent overcrowding of the health system and reduce mortality. Also, the data analytics field could impact the transparency of the information regarding to impact of coronavirus. Tableau and PowerBI are two important tools for synthesizing big data sets based on geographical locations or the number of cases. These summaries and highlights will support the press and local authorities in their management decisions.

The Machine Learning SARIMAX model is also an important way to predict the evolution of the pandemic in a country if a proper database is provided.

The vaccine could reduce the number of deaths in a medium-term (at least 2 months or more) and could decrease the number of confirmed cases.

8. Conclusions

It is clear that Data Science could positively impact the management of crisis situations (in our case - COVID-19 pandemic). By introducing AI models and data analysis techniques, the responsible authorities could increase their efficiency. „AI technologies offer significant opportunities to enhance COVID-19 management and control. Machine learning models and natural language processing (NLP) techniques are capable of detecting, diagnosing, and predicting COVID-19, which helps public health officials and medical professionals make informed decisions and allocate resources accordingly" [18].

The most significant hindrance is that numerous countries have not published enough data. Most developed countries like France, Spain, the USA or the UK uploaded a significant dataset about confirmed cases and the number of vaccines. In their case, a scientific research is relevant and could support the decisions of governments. We submitted the analysis of Romania, where access to data represented a problem during the pandemic years [19].

Data Science is a pillar of support for Health Informatics and, taking into consideration the best practices, it could positively affect the development of health system around the world!

References

[1] E. Hovenga, M. Kidd, S. Garde, C. Hullin Lucay Cossion, *Health Informatics*, IOS Press, Amsterdam, Berlin, Tokyo, Washington DC, 2010

[2] E. Rahm, H. Hai Do, *Data Cleaning: Problems and Current Approaches*, Bulletin of the Technical Committee on Data Engineering, IEE Computer Society, 2008.

[3] World Health Organization, WHO Coronavirus (Covid-19) Dashboard, 2020

[4] European Center for Disease Prevention, Monitoring and reporting data and trends, 2020

[5] Diana Hintea, James Brusey, Elena Gaura, *A Study on Several Machine Learning Methods for Estimating Cabin Occupant Equivalent Temperature*, Coventry University, 2020

[6] O. J. Watson, *Global impact of the first year of Covid-19 vaccination: a mathematical modelling study*, The Lancet Infection Diseases, 2022

[7] C. Osborn, *Statistical Applications for Health Information Management, Second Edition*, American Health Information Management Association, 2006

[8] A. Morgan, A. Amed, D. George, M. Hallet, *Mastering Spark for Data Science*,

[9] N. Kumar, S. Susan, *Covid-19 Pandemic Prediction using Time Series Forecasting Models*, IEEE Explore, 2020

[10] B. Artley, *Time Series Forecasting with ARIMA, SARIMA and SARIMAX: A deep-dive on the gold standard of time series forecasting*, TowardsDataScience, 2022

[11] T. Khachatryan, *Time Series Forecasting with SARIMAX*, Geometrein.medium, 2022

[12] V. Mehandzhiyski, *What Is a SARIMAX Model*, 365 DataScience, 2023

[13] Box, G. Jenkins, G. Reinsel, G. C. Ljung, *Time series analysis: forecasting and control*, John Wiley & Sons, 2015

[14] E. Topol, *The importance of data quality for machine learning in healthcare*, International Journal of Epidemiology, 2020

[15] D. Montogomery, E. Peck, G. Vining, *Introduction to Linear Regression Analysis*, Wiley, 2012

[16] P. Patel, *Introduction to Quantitative Methods*, Harvard University, 2009

[17] D. Paulson, *Handbook of Regression and Modeling*, Chapman & Hall/CRC, 2006

[18] M. Rahimi, M. Shamsi, A. Rajabi, *Artificial Intelligence applications in managing and controlling Covid-19 pandemic*, Journal of Cellular and Molecular no. 3, 2021

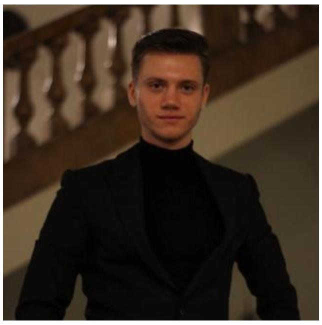
[19] S. Dascalu, O. Geambașu, C. Raiu, D. Azoicai, E. Popovici, C. Apetrei, *COVID-19 in Romania: What Went Wrong?*, Frontiers, 2021

[20] R. Klement, H. Walach, *SEIR Models in the light of Critical Realism - A critique of exaggerated claims about the effectiveness of Covid-19 vaccinations*, Futures 148, 2023

Serban-Gabriel TIMOFTE is student at the Faculty of Cybernetics, Statistics and Economic Informatics of the Academy of Economic Studies from 2022, after he studies one year of Computer Science at Coventry University. He worked for international organizations in Romania: Endava, Digital Nation and UK company RWSBC. At present, he works for EBR Net SRL in Bucharest as a software developer and is responsible for the finance department of the start-up. The main focus is to implement AI-related technologies in different business models to anticipate production results and decrease the pollution level of the business.

Ioana Ruxandra STROE graduated from Carol 1 National College in Craiova. Currently, Ruxandra is a first-year student at the Bucharest Academy of Economic Studies, specializing in Economic Informatics. During high school, she was involved in numerous volunteering activities both within her school and outside of it. For example, she helped renovate the school library, one of the oldest in the country, and contributed to the preservation of numerous books and paintings dating back hundreds of years. Additionally, she participated in various volunteering activities organized by JCI, one of the most well-known NGOs dedicated to young people in Craiova.

Cristina Raluca TERTEȘ graduated from Mihai Viteazul National College in Bucharest. Currently, Raluca is a first-year student at the Bucharest University of Economic Studies (ASE), enrolled in the Faculty of Cybernetics, Economic Informatics, and Statistics. Her specialization lies in Economic Informatics, highlighting her interest in applying technology within the economic and business context. Moreover, during high school, she actively participated in a variety of volunteering activities organized by Proedus, the largest volunteer program in Bucharest.



Daniel-Valentin TAȘCU is a first-year student at the Faculty of Cybernetics, Statistics, and Economic Informatics in Bucharest, is driven by a strong desire for knowledge. He excels in his studies and actively volunteers with the student organization, "Business Organization for Students." His dedication to continuous learning and selfless service fuels his ambition to make a positive impact in his field.

Radu-Ioan MOGOȘ is a senior lecturer/associate professor within the Department of Economic IT and Cybernetics at Bucharest University of Economic Studies (ASE). He is teaching disciplines like Programming fundamentals, Programming Techniques and Algorithms, Artificial intelligence Economic Information Systems, Evolutionary Programming and Genetic Algorithms. He is also researcher in the field of applied computer science, being the author or co-author to several books and articles that were published in national and international journals and conference proceedings. He had postdoctoral studies at ASE during 2014-2015 and had defended his PhD thesis on 2011 at ASE. He has a Master Degree in Business Relations and Communication (ASE, 2007) and a master degree in English Language Education and Research Communication for Business and Economics - EDURES (ASE, 2014). He is also member of the Romania Project Management Association, being during the time member in several major projects.