European Research Studies Journal
Volume XXIII, Special Issue 3, 2020

pp. 556-566

The Use of Decision Trees for Analysis of the Potential Determinants for the Incidence of Deaths and Cases of Coronavirus (Covid-19) in Different Countries

Submitted 28/10/20, 1st revision 14/11/20, 2nd revision 05/12/20, accepted 28/12/20

Joanna Wyrobek[1]

Abstract:

Purpose: *The objective of this paper is to identify the key economic factors determining the intensity of COVID virus infections and deaths in various countries.*
Design/Methodology/Approach: *The publication uses the methods of k-means clustering, k-nearest neighbors algorithm, DBSCan algorithm to divide countries into different groups in terms of the level of disease and death from COVID. The decision tree analysis provided potential determinants for the severity of the pandemic in different regions of the world. We analyzed 211 countries. As potential determinants, we examined the following factors: average temperature, average precipitation, GDP per capita, population density (in 2018), hospital beds per 1.000 citizens, doctors per 1.000 citizens, share of people aged 65 years or above, pollution (based on the PM2.5 indicator from The World Bank), total tests per 1.000 citizens and health expenditure as a percentage of GDP.*
Findings: *Our analysis revealed that the COVID pandemic intensity in analyzed countries depends on the number of doctors, population density, average temperature, total tests per 1.000 citizens, and GDP per capita by using data from the World Bank.*
Practical Implications: *Research suggests that the efficient healthcare system supports immunological response of population to the COVID virus. Another critical factor is the density of population. These two factors proved to play a critical role in determining the level of the COVID cases in deaths in various countries. Nevertheless, countries with the lowest GDP per capita had very low levels of COVID cases, which suggests that either they do not recognize COVID patients correctly or they will experience the pandemic with time delay.*
Originality/value: *Economic factors proved to be good predictors of the COVID virus development in the analyzed countries, however, contrary to expectations, countries with low GDP per capita so far suffered the least from the COVID virus pandemic.*

Keywords: *COVID, data mining, economic determinants.*

JEL classification: *C58, G01.*
Paper Type: *Research study.*

Acknowledgement: *The publication is financed by the Ministry of Science and Higher Education of Poland within —Regional Initiative of Excellence Programme for 2019-2022. Project no.: 021/RID/2018/19, total financing 11 897 131,40 PLN.*

[1] *Prof. Economics/Finance Department, School of Economics, Finance and Law, Cracow University of Economics, Poland, e-mail: wyrobekj@uek.krakow.pl;*

1. Introduction

The research aimed to investigate the extent to which the severity of COVID-19 depends on the state of the healthcare system and GDP per capita in individual countries. To verify the above hypothesis, various measures of the healthcare system and national wealth were used, together with a number of tests in respective countries. The research also considered weather differences that could affect the time at which the residents of different countries remained outside their household.

2. Literature review

Previous studies claimed that the incidence and number of COVID depend not only on a given person's immune factors but also on economic factors. Qui *et al.* (2020) suggest that the government's response has a proactive reduction effect on the COVID incidence rates. Lebni *et al.* (2020) mention the importance of the government's financial support that may keep people stay at home and not go to crowded places. One example is paid sick leave to keep contagious workers at home (Barmby and Larguem, 2009; Pichler and Ziebarth 2017). Political decisions and economic stabilization are mentioned by Zhang *et al.* (2019), Adda (2016), Sukharev (2020), Jain and Singh (2020). Apart from such factors as age, gender, coexistence, ethnicity, and obesity, Goutte *et al.* (2020) also lists population density, unemployment and poverty rates, a lack of formal education, and housing situation.

Sarmadi *et al.* (2020) supplement this list with the prevailing climate (warm temperature) and GDP per capita. Reduction of infection rate with an increase in the maximum temperature was found by Behnood *et al.* (2020).

Singu *et al.* (2020), Hawkins *et al.* (2020), and Varkey *et al.* (2020) suggest the influence of systemic social inequalities and differences in socioeconomic status (SES). High incidence of deaths from the COVID virus among individuals with income below median are described by Seligman *et al.* (2021). Seligman *et al.* also found high incidence of deaths among individuals of non-white ethnicity, less than high school level of education, and veterans. Baumer *et al.* (2020) add poor living conditions and neighborhood deprivation because they inflict psychosocial stress of individuals and cause immune system disfunction.

According to Milusheva (2020), limiting travel from places with high incidence is essential. The theory of the impact of travel and per capita income on COVID and the mean age of mortality is also suggested by Gangemi *et al.* (2020). Gangemi even measures the number of trips per capita as an essential determinant of periodic COVID.

In turn, Bhuiyan *et al.* (2020) point to the deaths resulting from isolation and mental problems related to the pandemic as indirectly but unambiguously caused by the COVID-19 pandemic.

3. Methodology

The research employed data from the World Bank regarding income, quality of health care, and the average temperature in the studied period (i.e., October 2020). To determine the severity of the situation in different countries, we used two figures - the number of illness cases per 100,000 inhabitants and the number of deaths from COVID per 100 inhabitants. These data are derived from the WHO (World Health Organization) database. Countries were grouped into clusters using three algorithms - k-means, hierarchy, and the DBScan algorithm. Based on the number of cases and deaths, these algorithms grouped 211 countries in the world into 4 clusters (such a number of clusters minimized the variance within the clusters). To assign an interpretation to the obtained results, we built a decision tree for each of the obtained clusters based on the Gini criterion. The algorithm automatically selected the most favorable split criterion (Gini coefficient). The depth of the decision trees was limited to four levels, as the aim was to distinguish the key factors allowing for assigning countries to different clusters.

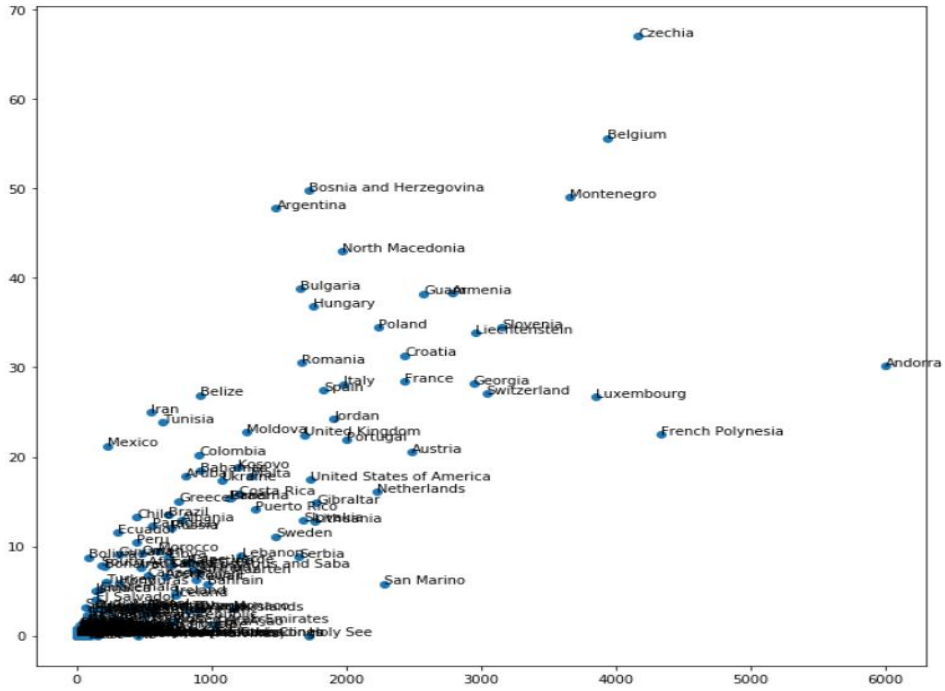
4. Research Results and Discussion

Figure 1 shows the concentration of different countries in terms of solely the number of COVID cases and deaths in October 2020. We assumed that in October 2020, the pandemic had already spread globally, and medical services had already gained the ability to identify the affected people. Figure 1 suggests that there are several levels of severity of the COVID pandemic in various countries. To define the exact number of clusters, we analyzed the decrease in variance in the clusters, which determined the selection of four clusters.

As it can be seen in the presented maps (Figures 2, 3, and 4), regardless of the selected clustering algorithm, the lowest level of cases and deaths in October 2020 occurred in China, Canada, most of Africa, South Asia, and Australia (class 0). The group of countries with a relatively good situation included the countries of Russia and South America (class 2). The third group with a somewhat challenging situation included Germany, Iceland, Ireland, Kuwait, United Arab Emirates, Azerbaijan, Cyprus, Denmark, Latvia, San Marino (class 3). Finally, the greatest severity of the virus was observed in the remaining European countries and the United States (class 1).

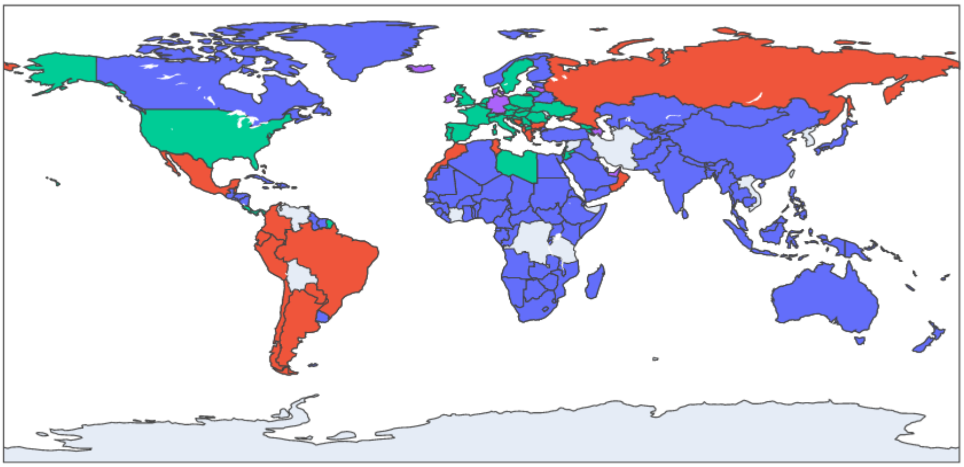
However, clustering does not answer the question regarding the factors that determined the varying scale of COVID cases between countries, at least in October. Assuming that the incidence is not declining as a third wave of the virus is expected, we trained decision trees to find the main determinants of differentiation between countries.

Figure 1. *The number of cases (x-axis) per 100.000 citizens and the number of deaths per 100.000 citizens in October 2020*



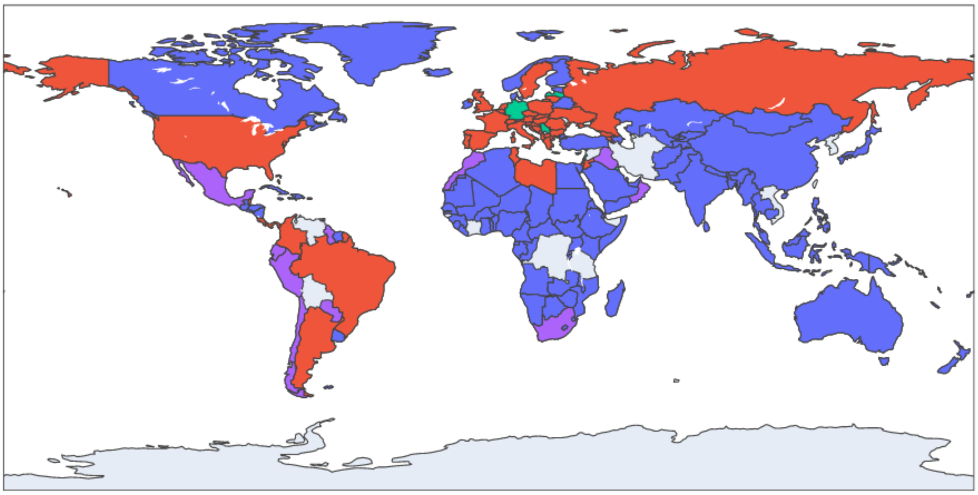
Source: *Own study.*

Figure 2. *K-means clustering based on the number of deaths and cases per 100.000 citizens, October 2020*



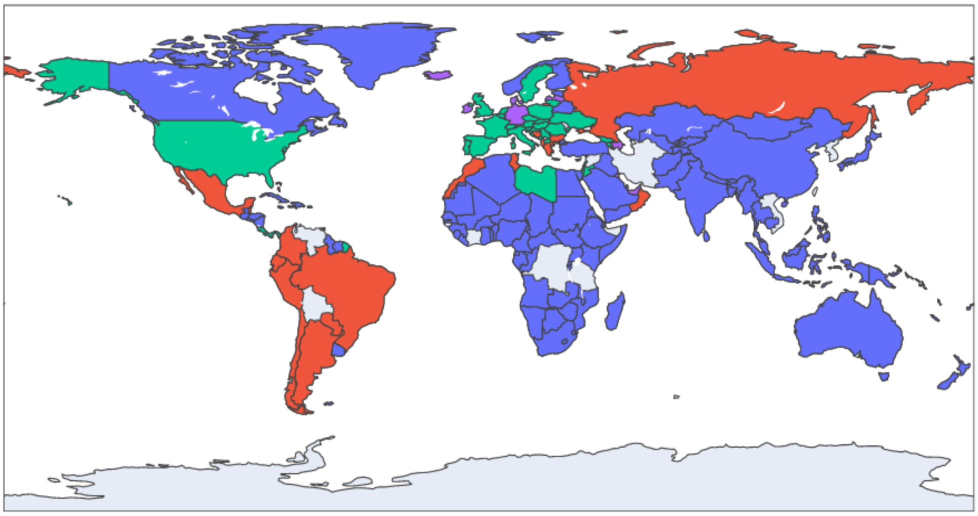
Source: *Own study.*

Figure 3. *Ward's hierarchical clustering based on the number of deaths and cases per 100,000 citizens, October 2020*



Source: *Own study.*

Figure 4. *Density-Based Spatial Clustering of Applications with Noise (epsilon = 0.55, min samples = 3) of countries based on the number of cases and deaths in October 2020*

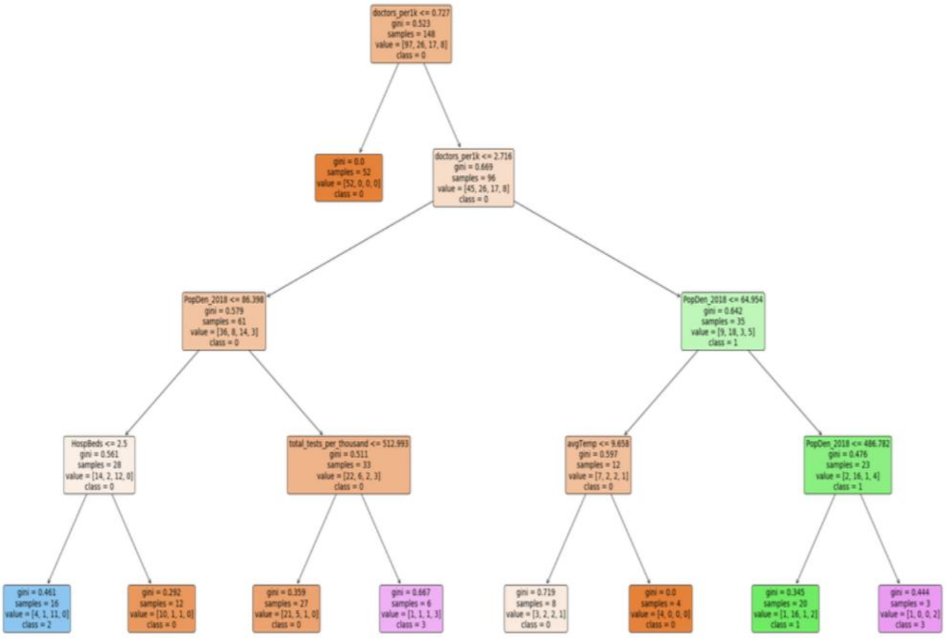


Source: *Own study.*

The results of DT algorithm are shown graphically in Figures 5, 6, and 7. As demonstrated in the constructed trees, the essential criterion determining the allocation to a given cluster was usually the number of tests performed, and doctors per 1,000 citizens. These two features were closely followed by population density. An additional criterion of the division was the number of beds per capita (it lowered the number of cases), the average temperature (cold weather reduced the number of disease cases). Countries with a very low GDP per capita (below USD 5,000) observed fewer COVID cases than the remaining nations (and it also shows that countries with very high GDP per capital have more cases than other countries (Qiu *et al.*, 2020). The age above 65 years is not visible in the presented figures, but it was significant for the trees of depths with 5 levels or more.

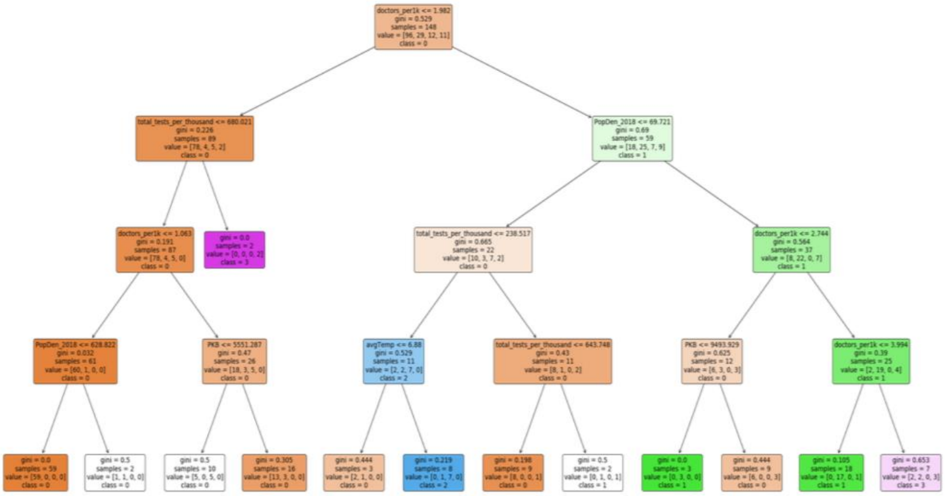
Figures 5-7 show that the most interesting cluster seemed to include Germany, Denmark, Iceland and the United Arab Emirates. Countries in this group were characterized by a high per capita income, an excellent condition of medical services, a high population density, and many tests performed for the COVID virus. Despite the very high number of tests per 100,000 inhabitants, countries in this group had a much lower number of cases than in other European countries. The number of deaths per 100,000 inhabitants was also lower.

Figure 5. *Decision tree depicting determinants of classifying observations into clusters created with the K-means algorithm*



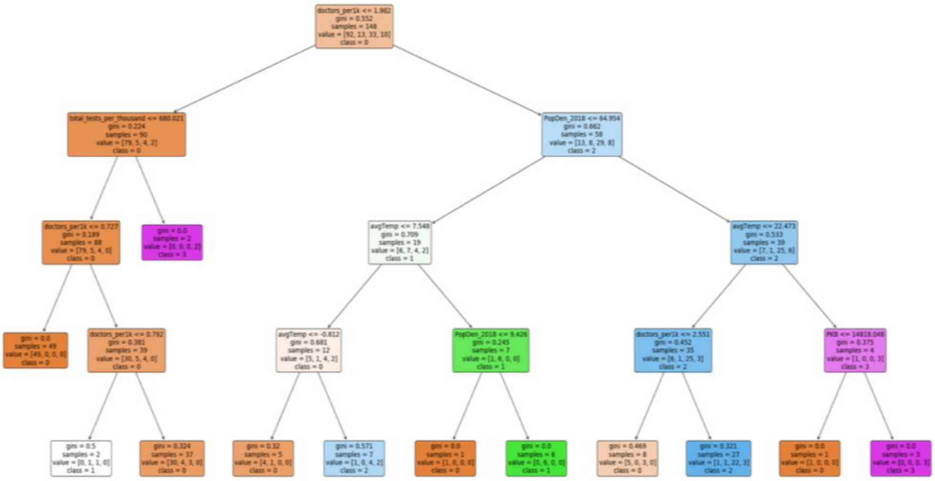
Source: *Own study.*

Figure 6. *Decision tree depicting the determinants for classifying observations into clusters created with the hierarchization algorithm*


doctors_per1k <= 1.982 gini = 0.529 samples = 148 value = [96, 29, 12, 11] class = 0
total_tests_per_thousand <= 680.021 gini = 0.226 samples = 89 value = [78, 4, 5, 2] class = 0
PopDen_2018 <= 69.721 gini = 0.69 samples = 59 value = [18, 25, 7, 9] class = 1
doctors_per1k <= 1.063 gini = 0.191 samples = 87 value = [78, 4, 5, 0] class = 0
gini = 0.0 samples = 2 value = [0, 0, 0, 2] class = 3
total_tests_per_thousand <= 238.517 gini = 0.665 samples = 22 value = [10, 3, 7, 2] class = 0
doctors_per1k <= 2.744 gini = 0.564 samples = 37 value = [8, 22, 0, 7] class = 1
PopDen_2018 <= 628.822 gini = 0.032 samples = 61 value = [60, 1, 0, 0] class = 0
PKB <= 5551.287 gini = 0.47 samples = 26 value = [18, 3, 5, 0] class = 0
avgTemp <= 6.88 gini = 0.529 samples = 11 value = [2, 2, 7, 0] class = 2
total_tests_per_thousand <= 643.748 gini = 0.43 samples = 11 value = [8, 1, 0, 2] class = 0
PKB <= 9493.929 gini = 0.625 samples = 12 value = [6, 3, 0, 3] class = 0
doctors_per1k <= 3.994 gini = 0.39 samples = 25 value = [2, 19, 0, 4] class = 1
gini = 0.0 samples = 59 value = [59, 0, 0, 0] class = 0
gini = 0.5 samples = 2 value = [1, 1, 0, 0] class = 0
gini = 0.5 samples = 10 value = [5, 0, 5, 0] class = 0
gini = 0.305 samples = 16 value = [13, 3, 0, 0] class = 0
gini = 0.444 samples = 3 value = [2, 1, 0, 0] class = 0
gini = 0.219 samples = 8 value = [0, 1, 7, 0] class = 2
gini = 0.198 samples = 9 value = [8, 0, 0, 1] class = 0
gini = 0.5 samples = 2 value = [0, 1, 0, 1] class = 1
gini = 0.0 samples = 3 value = [0, 3, 0, 0] class = 1
gini = 0.444 samples = 9 value = [6, 0, 0, 3] class = 0
gini = 0.105 samples = 18 value = [0, 17, 0, 1] class = 1
gini = 0.653 samples = 7 value = [2, 2, 0, 3] class = 3

Source: *Own study.*

Figure 7. *Decision tree depicting the determinants for classifying observations into clusters created with the DBScan algorithm*


doctors_per1k <= 1.982 gini = 0.552 samples = 148 value = [92, 13, 33, 10] class = 0
total_tests_per_thousand <= 680.021 gini = 0.224 samples = 90 value = [79, 5, 4, 2] class = 0
PopDen_2018 <= 64.954 gini = 0.662 samples = 58 value = [13, 8, 29, 8] class = 2
doctors_per1k <= 0.727 gini = 0.189 samples = 88 value = [79, 5, 4, 0] class = 0
gini = 0.0 samples = 2 value = [0, 0, 0, 2] class = 3
avgTemp <= 7.548 gini = 0.709 samples = 19 value = [6, 7, 4, 2] class = 1
avgTemp <= 22.473 gini = 0.533 samples = 39 value = [7, 1, 25, 6] class = 2
gini = 0.0 samples = 49 value = [49, 0, 0, 0] class = 0
doctors_per1k <= 0.792 gini = 0.381 samples = 39 value = [30, 5, 4, 0] class = 0
avgTemp <= -0.812 gini = 0.681 samples = 12 value = [5, 1, 4, 2] class = 0
PopDen_2018 <= 9.426 gini = 0.245 samples = 7 value = [1, 6, 0, 0] class = 1
doctors_per1k <= 2.551 gini = 0.452 samples = 35 value = [6, 1, 25, 3] class = 2
PKB <= 14818.048 gini = 0.375 samples = 4 value = [1, 0, 0, 3] class = 3
gini = 0.5 samples = 2 value = [0, 1, 1, 0] class = 1
gini = 0.324 samples = 37 value = [30, 4, 3, 0] class = 0
gini = 0.32 samples = 5 value = [4, 1, 0, 0] class = 0
gini = 0.571 samples = 7 value = [1, 0, 4, 2] class = 2
gini = 0.0 samples = 1 value = [1, 0, 0, 0] class = 0
gini = 0.0 samples = 6 value = [0, 6, 0, 0] class = 1
gini = 0.469 samples = 8 value = [5, 0, 3, 0] class = 0
gini = 0.321 samples = 27 value = [1, 1, 22, 3] class = 2
gini = 0.0 samples = 1 value = [1, 0, 0, 0] class = 0
gini = 0.0 samples = 3 value = [0, 0, 0, 3] class = 3

Source: *Own study.*

5. Summary and Concluding Comments

The analysis of COVID incidence must take into account considerable differences in the level of available tests. Therefore, the outcomes of our research should be treated with great caution as preliminary results. Nevertheless, we have seen a confirmation of the hypothesis that the countries' economic situation had an impact on the course of the COVID pandemic. Those with a large number of doctors per 1,000 inhabitants and a large number of hospital beds were facing a completely different situation than countries with a low number of doctors and beds. The differences are evident in the number of deaths. Countries with poor medical conditions have worse rates of disease to deaths compared to those with a more favorable healthcare setting.

On the other hand, at least over the period considered, the most impoverished countries had the fewest cases per 100,000 inhabitants and the lowest deaths. It is difficult to say whether this is due to the delay with which the virus would reach these countries, the lack of medical care, or other reasons presently unknown – one of the potential causes might be population density. The worst situation in the analyzed period was in Europe and in the USA, which can be attributed (in Europe) to a high population density, increased mobility of the population, and tests, which are the primary diagnostic method. Therefore, although the number of beds and doctors reduces the mortality in disease, a high GDP per capita is positively related to both the number of deaths and COVID cases.

The obtained results suggest several possible determinants of the intensity of the COVID pandemic. It does not seem that temperature directly influences the level of illness and death, but rather, we should suspect the activity of people outside their household and the time spent in public space, especially as a positive relationship between population density and the intensity of the pandemic is confirmed. Thus, the lockdown, despite its inconvenience and negative impact on the GDP, should reduce the incidence rate. The number of doctors per 1,000 inhabitants is of great importance, which confirms the significant of isolating sick people and appropriate treatment. The level of GDP had little impact on the severity of the pandemic, but very low-income countries has few cases and deaths - it is difficult to determine whether this was due to a low mobility or the lack of an adequate number of tests and doctors to diagnose patients. On the other hand, the spending of the GDP on medical care in a given country is irrelevant, which may result from different purchasing power parity and effectiveness of using funds (and equal access to medical care). Similarly, the level of air pollution turned out to be negligible.

Summarizing the obtained results, analyzing the maps shown in the publication it can be concluded that countries with a similar geographical location had a similar pandemic situation but with a significant impact on the quality of health care and the number of tests performed. This proves the considerable influence of economic and social factors (medical care), which have considerable power on the course of the pandemic.

It is worth paying attention to the limitations of the conducted research. Due to the use of temperature in the study, the study covered data only for one month of October 2020. In the future, the analysis will be extended to the remaining months of the year to check whether the season, rainfall, and temperature could have had a significant impact on the pandemic (due to the amount of time spent outside the household). However, it is worth emphasizing that the most critical determinants were the number of tests and the quality of the healthcare system, while climatic factors were of secondary importance.

References:

Adda, J. 2016. Economic Activity and the Spread of Viral Diseases: Evidence from High-Frequency Data. Q. J. Econ., 131(2), 891-941. DOI: 10.1093/qje/qjw005.

Barmby, T., Larguem, M. 2009. Coughs and sneezes spread diseases: an empirical study of absenteeism and infectious illness. J Heal. Econ., 28(5), 1012-1017. DOI: doi:10.1016/j.jhealeco.2009.06.006.

Baumer, Y., Farmer, N., Premeaux, T.A., Wallen, G.R., Powell-Wiley, T.M. 2020. Health Disparities in COVID-19: Addressing the Role of Social Determinants of Health in Immune System Dysfunction to Turn the Tide. Front. Public Heal., 8, 1-10. DOI: 10.3389/fpubh.2020.559312.

Behnood, A., Mohammadi Golafshani, E., Hosseini, S.M. Determinants of the infection rate of the COVID-19 in the U.S. using ANFIS and virus optimization algorithm (VOA). Chaos, Solitons and Fractals, 139, 110051. DOI: https://doi.org/10.1016/j.chaos.2020.110051.

Bhuiyan, A.K.M.I., Sakib, N., Pakpour, A.H., Griffiths, M.D., Mamun, M.A. 2020. COVID-19-Related Suicides in Bangladesh Due to Lockdown and Economic Factors: Case Study Evidence from Media Reports. Int. J. Ment. Health Addict. DOI: 10.1007/s11469-020-00307-y.

Gangemi, S., Billeci, L. Tonacci, A. 2020. Rich at risk: Socioeconomic drivers of COVID-19 pandemic spread. Clin. Mol. Allergy, 18(1), 10-12. DOI: 10.1186/s12948-020-00127-4.

Goutte, S., Péran, T., Porcher, T. 2020. The role of economic structural factors in determining pandemic mortality rates: Evidence from the COVID-19 outbreak in France. Res. Int. Bus. Finance, 54, 101281. DOI: 10.1016/j.ribaf.2020.101281.

Hawkins, R.B. Charles, V., Mehaffey, J.H. 2020. Socioeconomic status and COVID-19–related cases and fatalities. Public Health, 189, 129-134. DOI: 10.1016/j.puhe.2020.09.016.

Jain, V., Singh, L. 2020. Global Spread and Socio-Economic Determinants of COVID-19 Pandemic. Seoul Journal of Economics, 33(4), 561-600.

Lebni, J.Y., Abbas, J., Moradi, F., Salahshoor, M.R., Chaboksavar, F., Irandoost, S.F., Nezhaddadgar, N., Ziapour, A. 2020. How the COVID-19 pandemic effected economic, social, political, and cultural factors: A lesson from Iran. Int. J. Soc. Psychiatry, 0020764020939984. DOI: 10.1177/0020764020939984.

Liu, S., Zhang, W., Yinhe, X., Feng, B. 2019. The Effectiveness of China’s Renewable Energy Policy: An Empirical Evaluation of Wind Power Based on the Framework of Renewable Energy Law and Its Accompanying Policies, Emerg. Mark. Financ. Trade. DOI: https://doi.org/10.1080/1540496X.2019.1628016.

Milusheva, S. 2020. Managing the spread of disease with mobile phone data. J. Dev. Econ., 147, 102559. DOI: https://doi.org/10.1016/j.jdeveco.2020.102559.

Pichler, S., Ziebarth, N. 2017. The pros and cons of sick pay schemes: Testing for contagious presenteeism and noncontagious absenteeism behavior, J. Public Econ., 156(c), 14-33. At: https://econpapers.repec.org/RePEc:eee:pubeco:v:156:y:2017:i:c:p:14-33.

Qiu, Y., Chen, X., Shi, W. 2020. Impacts of Social and Economic Factors on the

Transmission of Coronavirus Disease 2019 (COVID-19) in China. medRxiv, 2020.03.13.20035238. DOI: 10.1101/2020.03.13.20035238.

Sarmadi, M., Marufi, N. Moghaddam, V.K. 2020. Association of COVID-19 global distribution and environmental and demographic factors: An updated three-month study, Environ. Res., 188, 109748. DOI: https://doi.org/10.1016/j.envres.2020.109748.

Seligman, B., Ferranna, M., Bloom, D.E. 2021. Social Determinants of Mortality from COVID-19: A Simulation Study Using NHANES. PLOS Med., 1-13, DOI: 10.1371/journal.pmed.1003490.

Singu, S., Acharya, A., Challagundla, K., Byrareddy, S.N. 2020. Impact of Social Determinants of Health on the Emerging COVID-19 Pandemic in the United States. Front. public Heal., 8, 406. DOI: 10.3389/fpubh.2020.00406.

Sukharev, O.S. 2020. Economic crisis as a consequence COVID-19 virus attack: risk and damage assessment. Quant. Financ. Econ., 4(2), 274-293. DOI: 10.3934/qfe.2020013.

Varkey, R.S., Joy, J., Sarmah, G., Panda, P.K. 2020. Socioeconomic determinants of COVID-19 in Asian countries: An empirical analysis. J. Public Aff., 1-10. DOI: 10.1002/pa.2532.

Appendix: Confusion matrices for decision trees trained in the publication

Figure 8. *Confusion matrix for the decision tree estimated for classification with the K-means algorithm, the training sample*

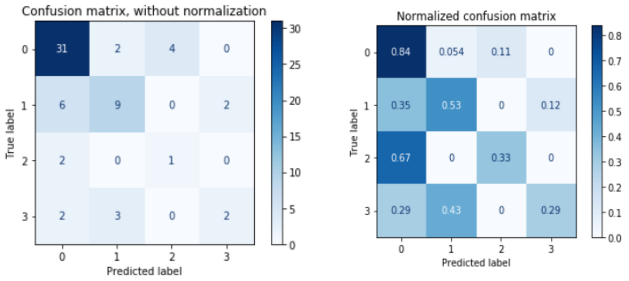


Figure 9. *Confusion matrix for the decision tree estimated for classification with the K-means algorithm, validation sample*

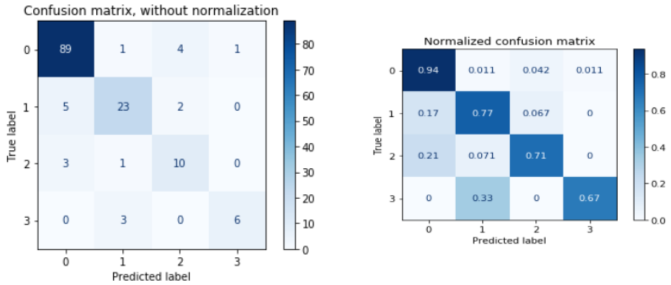


Figure 10. *Confusion matrix for the decision tree estimated for classification with the hierarchization algorithm, the training sample*

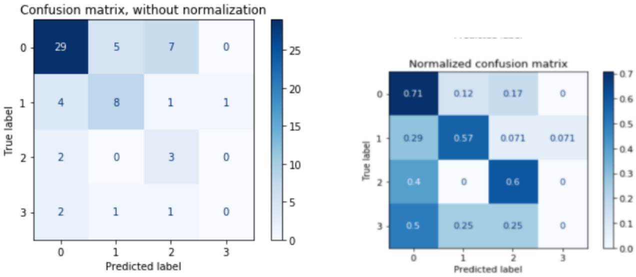


Figure 11. *Confusion matrix for the decision tree estimated for classification with the hierarchization algorithm, validation sample*

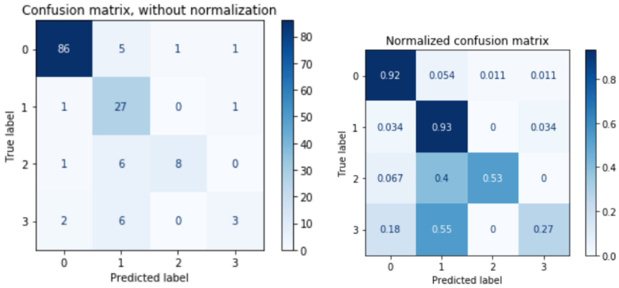


Figure 12*. Confusion matrix for the decision tree estimated for classification with the DBScan algorithm, the training sample*

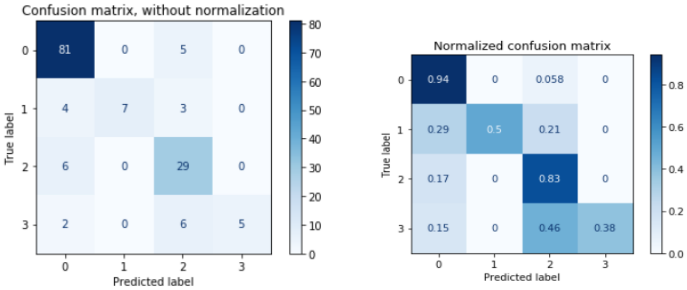
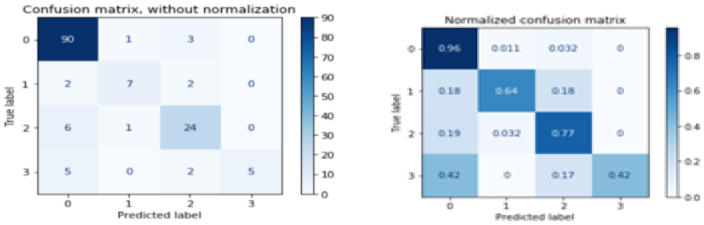


Figure 13. *Confusion matrix for the decision tree estimated for classification with the DBScan algorithm, validation sample*


Confusion matrix, without normalization
True label
Predicted label
Normalized confusion matrix