Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 4840488, 6 pages
https://doi.org/10.1155/2021/4840488

Research Article

Machine Learning Model and Statistical Methods for COVID-19 Evolution Prediction

M. D. Alsulami,[1] **Hanaa Abu-Zinadah**,[2] **and Anwar Hassan Ibrahim**[3]

[1]*University of Jeddah, College of Sciences and Arts at Alkamil, Department of Mathematics, Jeddah, Saudi Arabia*
[2]*University of Jeddah, College of Science, Department of Statistics, Jeddah, Saudi Arabia*
[3]*Qassim University, College of Engineering, Department of Electrical Engineering, Qassim, Saudi Arabia*

Correspondence should be addressed to Hanaa Abu-Zinadah; hhabuznadah@uj.edu.sa

Received 5 October 2021; Revised 20 October 2021; Accepted 18 November 2021; Published 15 December 2021

Academic Editor: Rashid A Saeed

Copyright © 2021 M. D. Alsulami et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In this paper, we discuss the statistical processing of COVID-19 data. COVID-19 was initially recognized in Wuhan, China, on December 31, 2019. It then spread to other parts of the world, so it became known as a pandemic. It has received interest due to its sudden emergence as a deadly human pathogen. The effect is not only confined to morbidity and mortality but also extends to social and economic consequences. Statistical analysis is required to measure the damage done to humans and take the necessary measures to limit this damage. The objective of the work was to examine the effects of various factors on the deaths due to COVID-19. To achieve this goal, we applied a logistic regression (LR) model, as a statistical method, and a decision tree model, as a machine learning method, to model the deaths due to COVID-19 in France, Germany, Italy, and Spain. The predictive abilities of these two models were compared. The overall accuracies of the decision tree and LR were 94.1% and 93.9%, respectively. It was also observed that countries with high population densities tended to have more cases than those with smaller population densities. There were more female deaths than male deaths in the United Kingdom, and more deaths occurred for those aged 65 years and older. The data were collected from the World Health Organization's official website from January 11, 2020, to May 29, 2020. The results obtained were in agreement with the previous results obtained by others.

1. Introduction

The spread of the COVID-19 virus was first confirmed in Italy on January 31, 2020, after Chinese tourists that visited tested positive in Rome [1]. After seven days, a male tourist from Italy returned home from Wuhan, China. He was hospitalized and proved to be the third case in Italy [2]. On February 21, 2020, more cases were detected, beginning with 16 proven cases in Lombardy [3]. The following day, an additional 60 cases and the first death were reported [4]. In early March, COVID-19 moved across Italy [5]. As of July 19, 2020, there were 12,440 active cases in Italy. During the peak of the pandemic, the number of active cases in Italy was among the highest in the world [6]. There were 244,434 confirmed cases and 35,045 deaths, an average of 578 deaths per million inhabitants [7], while there were 196,949 cases of recovery or dismissal. By July 19, Italy tested about 3,741,000 residents [8].

The pandemic has caused significant damage to the Italian economy. The tourism, residential, and food service sectors were the most severely affected by the travel restrictions from foreign countries to Italy. Moreover, a national closure was imposed by the government on March 8, 2020 [9]. By April, the Minister of Finance, Roberto Gualtieri, expected the gross domestic product (GDP) to drop by 6% for 2020. The first five proven cases in France were people that arrived from China [10, 11].

Table 1: Summary of the model.

Conditions	Growing technique	CHAID
	Dependent variable	Death
	Independent variables	New_cases, domestic general government health expenditure per capita, PPP (present international $), joblessness, youth total (% of the total workforce aged 15-24) (modeled ILO estimate), real GDP growth (annual percent change), projected old-age dependency ratio per 100 persons
	Validation	None
	Maximum depth	3
	Lowest number of cases in parent node	100
	Lowest number of cases in child node	50
Results	Independent variables included	New cases, projected old-age dependency ratio per 100 persons
	No. of nodes	8
	No. of terminal nodes	5
	Depth	2

On January 28, a tourist from China was hospitalized in Paris but died on February 14. This was the first death from COVID-19 in France and outside of Asia [12, 13]. The main factor in the spread of COVID-19 across the metropolitan area was the yearly gathering of the Christian Open-Door Church on February 17–24, in Mulhouse, which was attended by nearly 2,500 people. About 50% of the attendees were thought to be infected with COVID-19 [14]. On June 21, there were 29,640 deaths, 160,377 confirmed cases, and 74,372 cases in which the person recovered after staying in hospitals in France.

A group of epidemiologists from France reported that less than 5% of France's population (about 2.8 million residents) could test positive for COVID-19, which is considered a high percentage in Île-de-France and Alsace [15]. France has faced a major recession due to the impact of the COVID-19 pandemic, which has affected the country's full production capacity, decreased global demand, and raised concerns about the availability of the raw materials. As a result, the country's manufacturing and other industrial sectors have temporarily ceased their industrial operations [16]. The first positive patient was reported near Munich, Bavaria, in Germany on January 27, 2020 [17].

Most of the cases arose in January and early February from the same auto parts manufacturer, and these were reported to be the earliest cases. Several cases of the Italian outbreak were discovered in Baden-Wurttemberg on February 25 and 26. A large group was associated with a carnival in Heinsberg, North Rhine-Westphalia, with the first reported death on March 9, 2020 [18, 19]. Some groups appeared across Heinsberg as well as China, Iran, and Italy [20]. By July 18, 2020, the Robert Koch Institute (RKI) officially reported 202,572 cases, 9,162 deaths, and about 197,200 recoveries [21]. On June 10, 2020, the RKI reported that the total confirmed number of cases was 184,861, of which 41.2% were in the age group of 35 to 59 years, i.e., called the youth age. The number of deaths reached 8,729, of which 85.3% were in the older age group (70 to 99 years) [22].

Table 2: Classification.

Observed	Predicted		
	No death	Death	Percent correct
No death	133	15	89.9%
Death	14	330	95.9%
Overall percentage	29.9%	70.1%	94.1%
Growing method: CHAID Dependent variable: death			

By January 31, 2020, the virus was first 1reported in Spain when a traveler from Germany with SARS-CoV-2 was diagnosed in La Gomera, Canary Islands [23]. Post-ad hoc genetic analysis revealed that about 15 strains of the virus were brought in, and infection started from the community by mid-February [24]. On March 13, 2020, 1,531 confirmed cases and 37 deaths were reported in the country. By July 17, 2020, 260,255 cases were confirmed and 28,420 deaths were reported [25].

Eleven vital blood indices were extracted using the random forest (RF) method to design an assistant discrimination tool [26]. This method yielded accuracies of 96.97% and 97.95% for the test and cross-validation sets, respectively. A convolutional neural network (CNN) was employed for feature extraction, and long short-term memory was used for the classification of patients based on X-ray images [27].

2. Methods

In this study, data were collected from the WHO database for four European countries. The data for COVID-19 in each country from January 11, 2020, to May 29, 2020, were

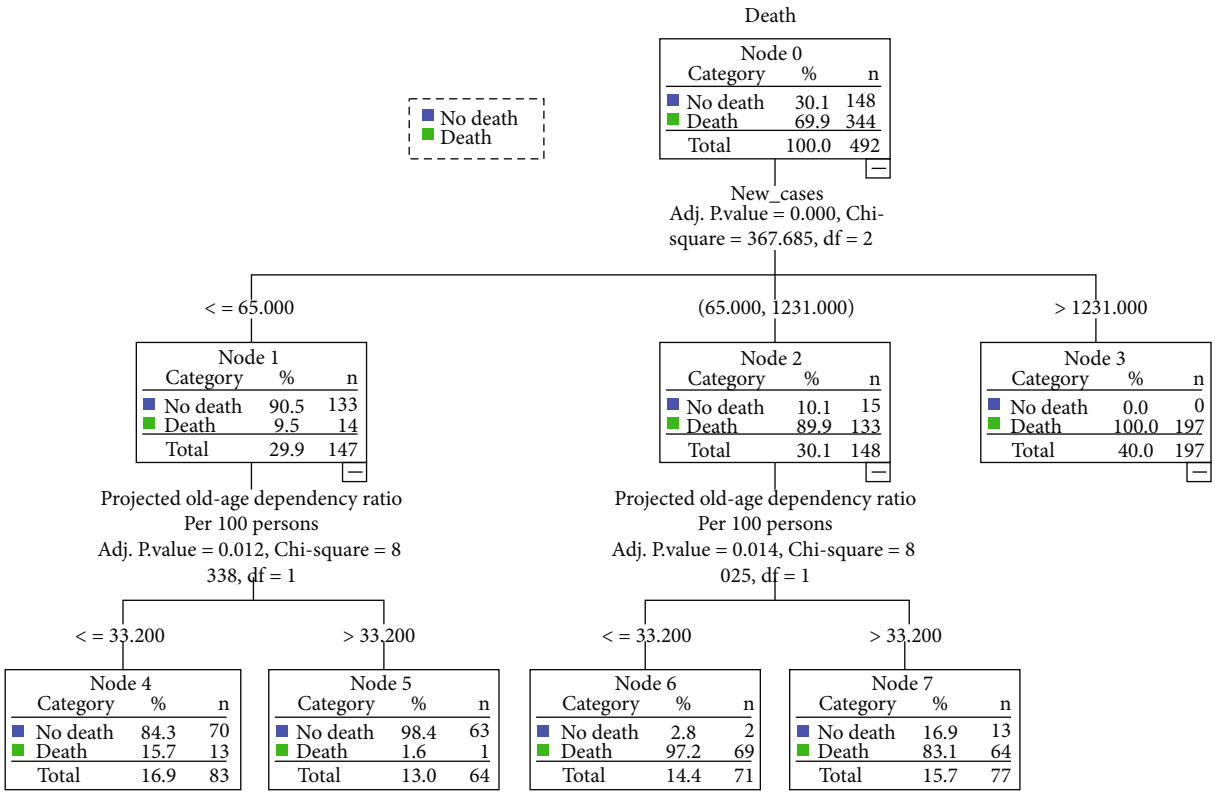

Figure 1: The tree diagram with 7 nodes.

correlated to other attributes: the real GDP growth (annual percentage change), national health expenditure per capita (current international $), and projected old-age dependency ratio per 100 persons. The total unemployment of the youth (percentage of the total labor force of ages 15–24, modeled International Labor Office (ILO) estimate) was estimated, and new cases were reported. Machine learning [26–28] and logistic regression (LR) models were employed.

Table 3: Summary of case processing.

Unweighted Cases[a]	*N*	Percent
Chosen cases		
Included in analysis	492	100.0
Missed cases	0	0.0
Total	492	100.0
Unselected cases	0	0.0
Total	492	100.0

3. Decision Tree

(i) The decision tree classifier is a supervised machine learning method [29, 30]. To develop a model, researchers input training data corresponding to correct output labels. The model was learned from the patterns in the training data [31, 32]. After this, data that the model had not encountered yet were input to determine how the model performed [33]. The decision tree model included three kinds of components [34, 35] as follows:

(a) Nodes represent decisions over the values of certain features

(b) Edges represent answers from nodes and are used to build connections to subsequent nodes

(c) Leaf nodes represent exit points for the result of the decision tree

4. Logistic Regression

The LR contains the linear regression equation within a sigmoid function [30, 36–41].

The formula of the LR takes the following form:

$$f(z) = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)\}}. \quad (1)$$

A sigmoid function is employed to map the values from a large range to the range of 0 to 1.

Table 4: The variables selected in the model.

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	New_cases	0.008	0.001	70.065	1	0.000	1.008
	Constant	-1.699	0.213	63.896	1	0.000	0.183
Step 2[b]	New_cases	0.009	0.001	65.844	1	0.000	1.009
	National health expenditure per capita, PPP (current international $)	-0.001	0.000	6.211	1	0.013	0.999
	Constant	0.130	0.736	0.031	1	0.860	1.139
Step 3[c]	New_cases	0.009	0.001	65.992	1	0.000	1.009
	Joblessness, youth total (% of total work force aged 15-24) (modeled ILO estimate)	-0.269	0.132	4.134	1	0.042	0.764
	National health expenditure per capita, PPP (current international $)	-0.002	0.001	7.218	1	0.007	0.998
	Constant	6.746	3.332	4.099	1	0.043	850.538
Step 4[d]	New_cases	0.010	0.001	61.487	1	0.000	1.010
	Real GDP growth (annual percent change)	1.372	0.488	7.905	1	0.005	3.945
	Joblessness, youth total (% of total work force aged 15-24) (modeled ILO estimate)	-0.491	0.159	9.497	1	0.002	0.612
	National health expenditure per capita, PPP (current international $)	-0.004	0.001	13.509	1	0.000	0.996
	Constant	26.598	7.962	11.160	1	0.001	355839592019.724

Table 5: Omnibus tests of model coefficients.

		Chi-square	df	Sig.
Step 1	Step	392.944	1	0.000
	Block	392.944	1	0.000
	Model	392.944	1	0.000
Step 2	Step	6.563	1	0.010
	Block	399.508	2	0.000
	Model	399.508	2	0.000
Step 3	Step	4.267	1	0.039
	Block	403.775	3	0.000
	Model	403.775	3	0.000
Step 4	Step	8.740	1	0.003
	Block	412.515	4	0.000
	Model	412.515	4	0.000

5. Data Analysis

5.1. Decision Tree. The growth method and the dependent and independent variables of the model are summarized in Table 1.

Table 2 is the classification table that shows that 94.1% of the training samples were classified correctly.

Figure 1 shows the tree diagram with 7 nodes.

5.2. Logistic Regression. The LR sought to predict deaths based on the following factors:

(1) x_1: projected old-age dependency ratio per 100 persons

(2) x_2: real GDP growth (annual percent change)

(3) x_3: joblessness, youth total (percentage of the total workforce aged 15–24, modeled ILO estimate)

(4) x_4: national health expenditure per capita

(5) x_5: new cases

The regression equation for the LR is as follows:

$$P(1) = \frac{e^{Y'}}{\left(1 + e^{Y'}\right)}. \quad (2)$$

As shown in Table 3, the analysis included 492 samples, with no missing data.

The analysis procedure was as follows:

(i) Step 1. Variable(s) input: new cases

(ii) Step 2. Variable(s) input: national health expenditure per capita (current international $)

(iii) Step 3. Variable(s) input: joblessness, youth total (percentage of the total workforce aged 15–24, modeled ILO estimate)

(iv) Step 4. Variable(s) input: real GDP growth (annual percent change)

Table 4 shows the variables selected in the model and their statistical significance.

Table 6: Classification table.

Observed			Predicted		
			Death		Percentage correct
			No death	Death	
Step 1	Death	No death	143	5	96.6
		Death	23	321	93.3
	Overall percentage				94.3
Step 2	Death	No death	144	4	97.3
		Death	23	321	93.3
	Overall percentage				94.5
Step 3	Death	No death	143	5	96.6
		Death	25	319	92.7
	Overall percentage				93.9
Step 4	Death	No death	144	4	97.3
		Death	26	318	92.4
	Overall percentage				93.9

Table 5 shows the omnibus test results of the model coefficients based on the chi-squared test. When the p value was <0.001, the null hypothesis was rejected. This finding suggests that the model best fits the data.

Table 6 shows that the overall percentage of correct classification was 93.9%.

6. Conclusion

The spread of the COVID-19 pandemic in most countries has threatened people and the economy. Therefore, this paper is aimed at evaluating the application of a machine learning model and a statistical model, namely, the decision tree and LR, to study the effects of various factors on the deaths due to COVID-19. This provides some statistical indicators about COVID-19. We determined that the decision tree performed better than LR. The overall accuracies were 94.1% and 93.9% for the decision tree and LR models, respectively, as shown in Tables 2 and 6. In addition, the results show that the areas with larger populations tended to have more cases than those with smaller populations. The number of deaths of females was greater than that of males in the UK, and it was greater for those aged 65 years and older.

Data Availability

Data are however available from the authors upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest to report regarding the present study.

Acknowledgments

We thank LetPub (http://www.letpub.com/) for its linguistic assistance during the preparation of this manuscript.

References

[1] https://www.corriere.it/cronache/20_gennaio_30/coronavirus-italia.

[2] E. Anzolin and A. Amante, "First Italian dies of coronavirus as outbreak flares in north, healthcare & pharma," 2020, https://www.reuters.com/article/us-china-health-italy-idUSKBN20F0UI.

[3] https://www.reuters.com/article/us-china-health-italy/coronavirus-outbreak-grows-in-northern-italy.

[4] https://www.corriere.it/cronache/20_febbraio_22/coronavirus-italia-nuovi-contagi-lombardia-veneto.

[5] "Coronavirus Colpite tutte le regioni. La Protezione civile: ecco i numeri aggiornati," https://www.avvenire.it/attualita/pagine/coronavirus-aggiornamento-5-marzo-2020.

[6] "COVID-19 dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University," 2020, https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6.

[7] "COVID-19 coronavirus pandemic," 2020, https://www.worldometers.info/coronavirus/#countries.

[8] "PCM-DPC dati forniti dal Ministero della Salute," 2020, http://www.salute.gov.it/imgs/C_17_notizie_4623_0_file.pdf.

[9] "Nothings less than a catasrophe': Venice left high and dry by coronavirus," 2020, https://www.theguardian.com/travel/2020/mar/17/nothing-less-than-a-catastrophe-venice-left-high-and-dry-by-coronavirus.

[10] E. Jacob, *Coronavirus: trois premiers cas confirmés en France, deux d'entre eux vont bien*, 2020, https://www.lefigaro.fr/sciences/coronavirus-trois-premiers-cas-confirmes-en-france-20200124.

[11] E. Surveill, *COVID-19 Is an Emerging, Rapidly Evolving Situation*, 2020, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7029452/.

[12] "Wuhan virus: France confirms fourth case of coronavirus in elderly Chinese tourist," 2020, https://www.straitstimes.com/world/europe/france-confirms-fourth-case-of-coronavirus-in-elderly-inese-tourist.

[13] "Coronavirus: first death confirmed in Europe 202," https://www.bbc.com/news/world-europe-51514837.

[14] "COVID-19-France," 2020, https://dashboard.covid19.data.gouv.fr/.

[15] https://www.lemonde.fr/sciences/article/2020/05/13/en-france-le-covid-19.

[16] "Insights on the impact of COVID-19 on the French economy - players profiled include AXA, Air France-KLM & Auchan Retail among others - ResearchAndMarkets.com," 2020, https://www.businesswire.com/news/home/20200421005461/en/Insights-Impact-COVID-19-French-Economy—Players.

[17] "Corona-Pandemie 900 Milliarden gegen die Angst 2020," https://www.spiegel.de/consent.

[18] https://www.kreis-heinsberg.de/aktuelles/aktuelles/?pid=5136.

[19] "Coronavirus: Immer mehr Infektionen in Deutchland," 2020, https://www.abendblatt.de/vermischtes/article228637475/Coronavirus-Corona-News-Live-Ticker-Covid-19.

[20] "COVID-19: Fallzahlen in Deutschland und weltweit," 2020, https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Fallzahlen.html.

[21] R. Koch-Institut, *COVID-19-Dashboard*, 2020, https://experience.arcgis.com/experience/478220a4c454480e823b17327b2bf1d4.

[22] "COVID-19 in Germany," 2020, https://www.rki.de/EN/Home/homepage_node.html.

[23] https://elpais.com/sociedad/2020/01/31/actualidad/1580509404_469734.html.

[24] "Sanidad confirma en La Gomera el primer caso de coronavirus en España," 2020, https://elpais.com/ciencia/2020-04-22/el-analisis-genetico-sugiere-que-el-coronavirus-ya.circulaba-por-espana-a-mediados-de-febrero.html.

[25] https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov-China/documentos.

[26] J. Wu, P. Zhang, L. Zhang et al., "Rapid and accurate identifcation of COVID-19 infection through machine learning based on clinical available blood test results," medRxiv, 2020.

[27] I. Abdullahi Baba, H. Ahmad, M. D. Alsulami, K. M. Abualnaja, and M. Altanji, "A mathematical model to study resistance and non-resistance strains of influenza," *Results in Physics*, vol. 26, article 104390, 2021.

[28] M. D. Alsulami, "Stochastic modeling of infectious disease," *International Journal of Innovation in Science and Mathematics*, vol. 8, no. 6, pp. 262–269, 2020.

[29] S. J. Almalki, T. A. Abushal, M. D. Alsulami, and G. A. Abd-Elmougod, "Analysis of type-II censored competing risks' data under reduced new modified Weibull distribution," *Complexity*, vol. 2021, Article ID 9932840, 13 pages, 2021.

[30] M. D. Alsulami, "Assorting faces by singular value decomposition," *IOSR Journal of Mathematics*, vol. 16, no. 6, pp. 01–05, 2020.

[31] S. M. Abo-Dahab, M. Ragab, A. A. Elhag, and S. Abdel-Khalek, "Free convection effect on oscillatory flow using artificial neural networks and statistical techniques," *Alexandria Engineering Journal*, vol. 59, no. 5, pp. 3599–3608, 2020.

[32] S. Abdel-khalek, A. Alhag, M. Ragab, S. M. Abo-Dahab, A. Algarni, and H. Ahmad, "Atomic Fisher information and entanglement forecasting for quantum system based on artificial neural network and time series model," *International Journal of Quantum Chemistry*, vol. 121, 2021.

[33] M. Zabirul Islam, M. Milon Islam, and A. Asraf, "A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images," *Informatics in Medicine Unlocked*, vol. 20, article 100412, 2020.

[34] S. L. Dharmapuri, P. K. Dandamudi, V. M. Botcha, and B. P. Kolla, "Detecting central nervous system disorder using machine learning technique (XGB classifier)," *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 4, pp. 1142–1147, 2020.

[35] K. B. Prakash, S. S. Imambi, M. Ismail, T. P. Kumar, and Y. V. R. Naga Pawan, "Analysis, prediction and evaluation of COVID-19 datasets using machine learning algorithms," *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 5, 2020.

[36] M. Gal-Or, J. H. May, and W. E. Spangler, "Using decision tree models and diversity measures in the selection of ensemble classification models," in *Multiple Classifier Systems. MCS 2005*, N. C. Oza, R. Polikar, J. Kittler, and F. Roli, Eds., vol. 3541 of Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2005.

[37] M. D. Guerrero, L. M. Vanderloo, R. E. Rhodes, G. Faulkner, S. A. Moore, and M. S. Tremblay, "Canadian children's and youth's adherence to the 24-h movement guidelines during the COVID-19 pandemic: a decision tree analysis," *Journal of Sport and Health Science*, vol. 9, no. 4, pp. 313–321, 2020.

[38] A. A. Elhag and H. Abu-Zinadah, "Forecasting under applying machine learning and statistical models," *Thermal Science*, vol. 24, Suppl. 1, pp. 131–137, 2020.

[39] H. Abu-Zinadah and A. Binkhamis, "Goodness-of-fit tests for the beta Gompertz distribution," *Thermal Science*, vol. 24, Suppl. 1, pp. 69–81, 2020.

[40] H. Abu-Zinadah and A. Binkhamis, "Lifetime competing risks data from Lomax distribution in the presence of accelerates life-testing model with type-I censoring," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 3, pp. 2873–2883, 2020.

[41] M. D. Alsulami, "Computational mathematical techniques model for investment strategies," *Applied Mathematical Sciences*, vol. 15, no. 1, pp. 47–55, 2021.

Copyright © 2021 M. D. Alsulami et al. This work is licensed under http://creativecommons.org/licenses/by/4.0/(the “License”). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.