Comparative study on COVID-19 and Analysing the prediction of Death, Recovery and confirmed Cases Using Rapid Minor Tool

Fakhra Akhtar

Faculty of Computer Science an Information Technology, RIPHA University Lahore, Pakistan

30th-June-2020

hanaali829@gmail.com

Muhammad Tauseef Hanif

Faculty of Computer Science COMSATS University Islamabad, Lahore-Campus

tauseefhanif05@gmail.com

Faizan Ahmed Khan

Faculty of Computer Science COMSATS University Islamabad, Lahore-Campus

geofaizan.khan@gmail.com

Asad Imran Malik

Superior University

malik_asad75@hotmail.com

Abstract— Coronavirus (COVID-19) is caused by SARS-COV2 and it represents the possibility of a serious infection relating to respiratory and various health issues and it has caused a major outbreak worldwide. It was originated from Wuhan city of China and its victims were those who were exposed to the wet animal market, its Person-to-person transmission led to the isolation of patients that were Quartined for treatments. In this paper the author has predicted the number of deaths, recoveries and confirmed cases from all over the world, Using various machine learning algorithms and rapid minor tool. The dataset used in this paper relates to coronavirus COVID-19 The algorithms used in this paper are Random Forest, Decision Tree and Gradual boost tree. Data set regarding COVID-19 is collected from Cagle. Which contains deaths, recoveries and confirmed cases from all over the world.

The results show the fact that the death ratio is less which justifies the fact that patients experience symptoms relating to COVID-19 mostly recover within a specific period and by adapting isolation and precautionary measures.

Keywords—COVID-19, Rapid minor, Kaggle

*****

INTRODUCTION

Coronavirus (COVID-19) is a severe infectious disease caused by the respiratory syndrome. This infection was first distinguished in December 2019 in Wuhan, the capital of China's Hubei region, and caused a significant coronavirus flare-up around the world. According to most recent reports In May 2020, more than 3.34 million cases have been accounted for across 187 nations and domains, bringing about more than 238,000 deaths. The World Health Organization took genuine and broad careful steps to diminish the individual-to-individual transmission of COVID-19 to control the present episode. Individuals, particularly youth tainted with the COVID-19 infection experience respiratory ailment and recoup without requiring unique treatment as a result of the solid resistant framework, while Older individuals having clinical history identifying with cardiovascular malady, diabetes, respiratory illness, and disease are bound to endure more. This sickness affected China alone as well as caused a significant flare-up worldwide and neighbouring nations too. In Pakistan, the principal instance of COVID-19 was affirmed and announced by the Ministry of Health on February 26, 2020, in Karachi.

Correspondingly another case affirmed by the Pakistan Federal Ministry of Health in Islamabad. Inside fifteen days, the quantity of all-out affirmed positive cases came to twenty out of 471 speculated cases with the most noteworthy numbers in the Sindh territory alone. The entirety of the affirmed cases had travel history from Iran, Syria and London. What's more, inside recent months the circumstance gained out of power bringing about the update of different vital plans and careful steps identifying with wellbeing and the executives. One ought to have appropriate mindfulness in regards to the counteraction of COVID-19 infection and its indications. Prudent steps ought to be received by Protecting yourself as well as other people from disease by washing your hands or utilizing a liquor based medicinally affirmed hand sanitisers and by covering your face with exceptional therapeutically endorsed masks. The first phase of the paper explains about literature review regarding the defined dataset i.e. CVOID-19, it specifies all the previous research work on coronavirus and its impact on the worldwide countries. The second phase deals with the methodology data analysis tool and the classification of algorithms. The third part discusses the research results based on the classification algorithms used in Rapid Miner. The fourth phase deals with Discussion, conclusion and references. The results are discussed in detail at the end of the paper in the form of various graphs and tables.

LITERATURE REVIEW

(1) Peng, L., at. el. (2020). describe in their research that the outbreak of coronavirus (COVID-19) has attracted worldwide consideration. They proposed a generalized SEIR model to analyze this epidemic outbreak. They gathered the data based on the public data of the National Health Commission of China from Jan. 20th to Feb. 9th, 2020. The findings of their research pointed out that the situation in Wuhan city is still very severe. More effective policies and more efforts on medical care and clinical research are required. **(2)** Bi, Q., at. el (2020). Describe in their research that the rapid spread of SARS-CoV-2 in Wuhan encouraged sharp surveillance in Shenzhen and elsewhere in China. They identified 391 cases from January 14 to February 12, 2020, and 1286 close associates. Their findings suggest that children are at similar risk of infection as the general population, though less likely to have severe symptoms; hence should be considered in analyses of transmission and control. **(3)**Schwartz, D. A. (2020). Highlight the fact that COVID-19, has rapidly spread across the globe creating a massive public health problem. This research analyzes literature describing 38 pregnant women with COVID-19 and their newborns in China to assess the effects on mothers and infants. Based on laboratory data, and the transmissibility of the virus from mother to fetus. The findings revealed that unlike coronavirus infections of pregnant women did not lead to maternal deaths. And no confirmed cases were reported.**(4)** Pedersen, M. G., & Meneghini, M. (2020). In their research described Italy as the most affected Western country, with more than 41.000 confirmed cases, For this purpose, they purposed COVID-19 dynamics with a SIQR model, They estimated model parameters by fitting model expressions to predict the number of unidentified positive confirmed cases. Their results predicted that recent drastic restrictions have reduced virus spreading modestly but insufficiently to halt the epidemic. They concluded that extreme social distancing is needed to contain the disease because of the large amount of undetected, infectious individuals in the absence of social-wide testing for SARS-nCov2.

METHODOLOGY

The dataset used in this paper relates to coronavirus COVID-19 The algorithms used in this paper are Random Forest, Decision Tree and Gradial boost tree. Data set regarding COVID-19 is collected from Kaggle. Which contains deaths, recoveries and confirmed cases from all over the world. Machine learning algorithms are applied to the dataset to predict the ratio of death recover and confirmed cases.

I. DEFINITIONS OF ALGORITHMS

Random Forest Algorithm

Random forests or random decision forests are considered as a learning strategy for arrangement, relapse and different errands that work by developing a mass of choice trees at preparing time (5) It alternatively delivers two extra snippets of data: a proportion of the significance of the indicator factors, and a proportion of the inside structure of the information. The necessary calculations are carried out tree by tree as the random forest is constructed.

Decision Tree

A decision tree partitions the input space of a data set into mutually exclusive regions, each of which is assigned a label, a value to characterize its data points. The choice tree system is straightforward and we can follow a tree structure effectively to perceive how the choice is made [6]. A choice tree is a tree structure comprising of inner and outer hubs associated by branches. An inside hub is a dynamic unit that assesses a choice capacity to figure out which kid hub to visit straight away. The outside hub, then again, has no kid hubs and is related with a mark or worth that portrays the given information that prompts it's being visited. [7]. The pruned choice tree that is utilized for characterization objects is known as the grouping tree.

Gradial Boost Tree

(GBDT) [8] is a widely-used machine learning algorithm, due to its efficiency, accuracy, and interpretability. GB DT accomplishes cutting edge exhibitions in many AI assignments, for example, multi-class order it has many viable usages, for example, XGBoost and pGBRT. Albeit many designing improvements have been received in these executions, the proficiency and adaptability are as yet unsuitable when the component measurement is high and information size is enormous. A significant explanation is that for each component, they have to examine all the information occasions to gauge the data addition of all conceivable split focuses, which is very tedious.

A. Data

The Dataset regarding COVID-19 is collected from Kaggle, This dataset contains deaths, recoveries and confirmed cases from all over the world. Machine learning algorithms are applied to the dataset and predict the ratio of death cases, recover case and confirmed cases

Rapid Minor

Rapid Miner is a data science software platform developed by the company of the same name that provides an integrated environment for data preparation, machine learning, and predictive analytics. It is utilized for business and business applications just as for examining, instruction, preparing, quick prototyping, and application advancement Rapid Miner (RM) is a domain for AI and information mining forms [10]. It is open-source, free task executed in Java. It speaks to another way to deal with configuration even convoluted issues a secluded administrator idea which permits the plan of complex settled administrator chains for countless learning issues.".

RESULTS

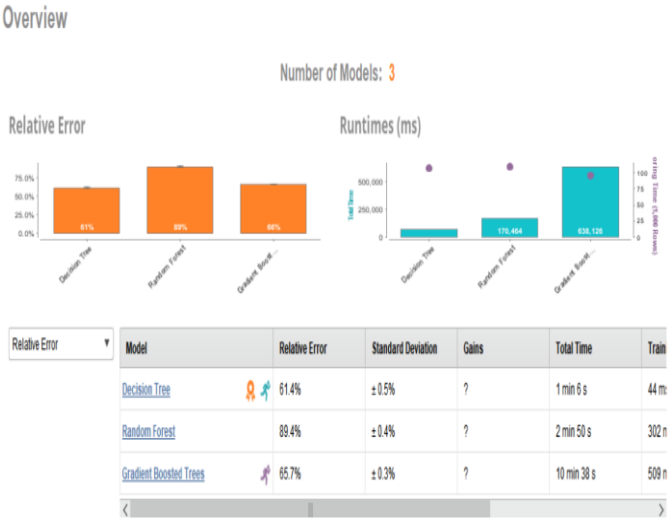
The rapid minor tool was used to analyze the results the data contained 10182 states and eight attributes having 19928 rows. The maximum number of confirmed cases reported were 299691 whereas the number of deaths reported was 27682. The number of recovered cases were 132929 the results showed that the death ratio is less which justifies the fact that patients experience symptoms relating to COVID-19 mostly recover within a specific period after staying in isolation and by taking precautionary measures. The results are shown in the form of tables and graphs.

Table 1 analysis of the number of cases relating to the corona virus

Cases	Min	max	average
Confirmed	0	299691	3550.325
deaths	0	27682	219.624
recovered	0	132929	927.808

Table 1 represents the number of cases of COVID-19 reported in the period from 22nd January 2020 to 29th April 2020. There are 10182 states and eight attributes in the data having 19928 rows. The maximum number of confirmed cases reported were 299691 whereas the number of deaths reported was 27682. The number of recovered cases were 132929 the results show the fact that the death ratio is less which justifies the fact that patients experience symptoms relating to COVID-19 mostly recover within a specific period.

Figure 1 Relative error Interpretation



Model	Relative Error	Standard Deviation	Gains	Total Time	Train
Decision Tree	61.4%	± 0.5%	?	1 min 6 s	44 m
Random Forest	89.4%	± 0.4%	?	2 min 50 s	302 n
Gradient Boosted Trees	65.7%	± 0.3%	?	10 min 38 s	509 n

The following figure shows the results of the application of relative error on all three classifiers i.e. Decision tree, Random forest and gradient boosted trees. The highest percentage of error was recorded on random forest i.e. 89.4% whereas the lowest percentage was of decision tree i.e. 61.4% Gradient boosted trees result in response time was high i.e. 10 min 38s as compared to other two algorithms.

Random forest

In random forest classifier following results and graphs will be interpreted weightage will define the number of cases every quarter, performance analysis will be used to calculate root mean square error absolute error, relative error and correlation the data will be visualized in the form of prediction charts and tables which will show observations based on confirmed and recovered cases of COVID-19

Table 2 Weightage

	Recovered	confirmed	Quarter 1	Quarter 2	Quarter 3
RaRRndfrs t for	0.655	0.308	0.050	0.018	0.191

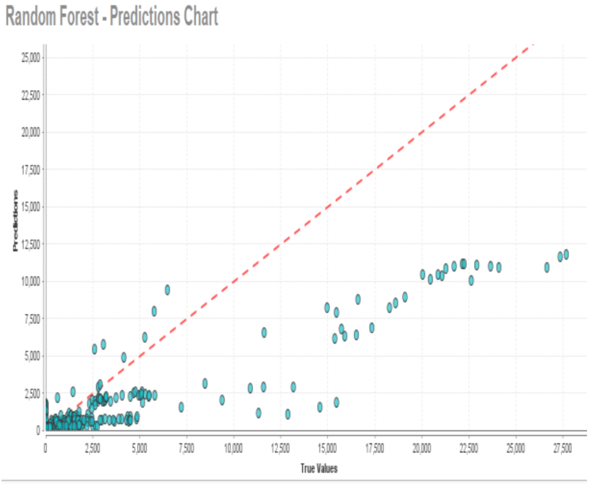
In table 2 there were 0.050 cases as compared to 2nd and 3rd quarter whereas the number of recovered cases were 0.655

Table 2(a) Performances

	root mean squared error	Absolute error	Relative error lenient	Squared error	Crelation
Random forest	939.438	230.364	89.39	904659.825	0.913

The results of table 2(a) show absolute error and root mean squared error the correlation of random forest algorithm is 0.913

Figure 2 Prediction chart



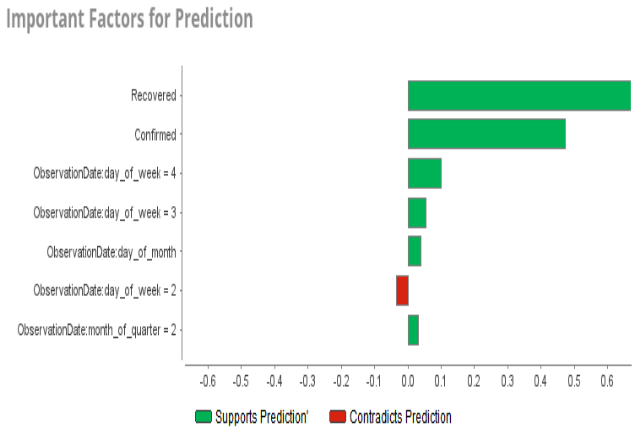
Prediction chart shows how much better a machine learning model performs as compared with a random guess. It also shows you the point at which the predictions become less useful.

Simulator prediction

	Simulator prediction
Random forest	135.750

Figure 2 predict and visualize the results of the random forest algorithms. The number of simulator prediction is 135.750 whereas the highest results are recorded in the favour of support predictor which show the results of recovered and confirmed cases.

Figure 2 (b)



The following figure shows the number of recovered and confirmed cases based on the results of the Random Forest predictions

Decision tree

In decision tree classifier following results and graphs will be interpreted weightage will define the number of cases every quarter, performance analysis will be used to calculate root mean square error absolute error, relative error and correlation the data will be visualized in the form of prediction charts and tables which will show observations based on confirmed and recovered cases of COVID-19

Table 3 Weightage

	Recovered	confirmed	Quarter 1	Quarter 2	Quarter 3
DeciDtree	0.055	0.638	0.000	0.014	0.132

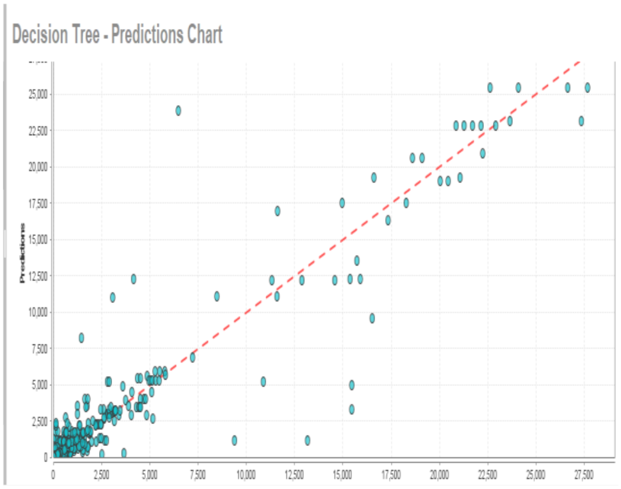
In table 3 there were 0.000 cases in first quarter whereas in 2nd and 3rd quarter 0.014 and 0.132 were reported whereas the number of recovered cases were 0.055

Table 3 (a) Performances

	root mean squared error	**Absolute error**	**Relative error lenient**	**Squared error**	**Correlation**
ddtree	388.308	53.804	61.44	168781.316	0.968

The results of table 3(a) show absolute error and root mean squared error the correlation of decision tree algorithm is 0.968

Figure 3 Prediction chart



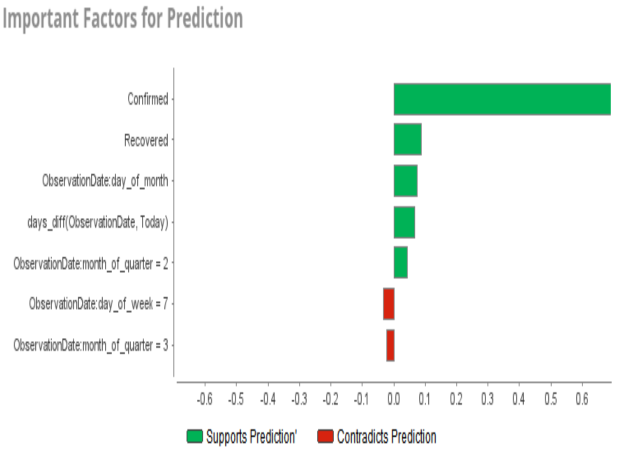
Prediction chart shows how much better an AI model proceeds as contrasted and an arbitrary supposition. It additionally shows you where the expectations become less valuable.

Simulator prediction

	Simulator prediction
Decision tree	107.073

Figure 3 predict and visualize the results of the decision tree algorithms. The number of simulator prediction is 107.073 whereas the highest number of confirmed cases were recorded.

Figure 3 (b)



The following figure shows the number of recovered and confirmed cases based on the results of the Decision tree predictions

Gradient boost tree

In gradient boost tree classifier following results and graphs will be interpreted weightage will define the number of cases quarterly, performance analysis will be used to calculate root mean square error absolute error, relative error and correlation the data will be visualized in the form of prediction charts and tables which will show observations based on confirmed and recovered cases of COVID-19

Table 4 Weightage

	Recovered	confirmed	Quarter 1	Quarter 2	Quarter 3
grbtre	0.064	0.731	0.000	0.051	0.182

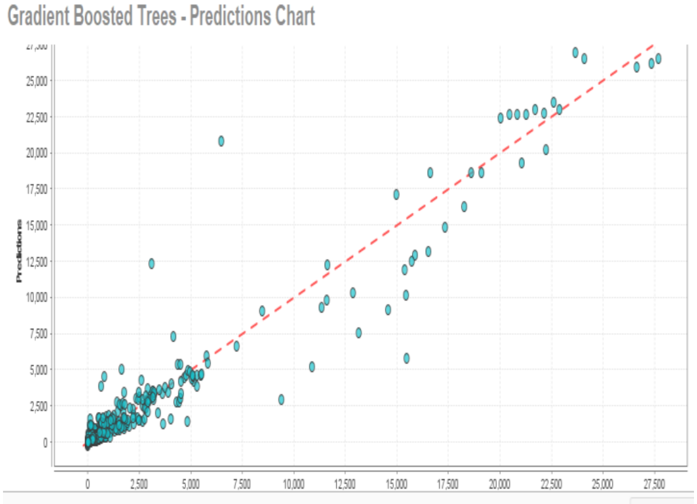
In table 4 there were 0.000 cases in first quarter whereas in 2nd and 3rd quarter 0.051 and 0.182 cases were reported whereas the number of recovered cases were 0.64.

Table 4 (a) Performances

	root mean squared error	Absolute error	Relative error lenient	Squared error	Correlation
Gragdtre	367.222	54.362	65.65	159992.062	0.961

The results of table 4(a) show absolute error and root mean squared error the correlation of gradient decision tree algorithm is 0.961

Figure 4 Prediction chart



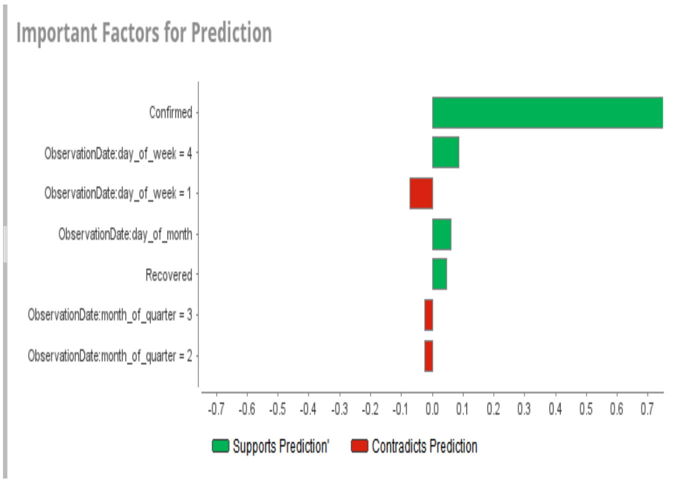
Prediction chart shows how much better an AI model proceeds as contrasted and an arbitrary supposition. It additionally shows you where the expectations become less valuable.

Simulator prediction

	Simulator prediction
Gradient boost tree	90.753

Figure 4 predict and visualize the results of the gradient boosted tree algorithms. The number of simulator prediction is 90.753 whereas the highest of confirmed cases were recorded

Figure 4 (b)



The following figure shows the number of recovered and confirmed cases based on the results of the gradient boost tree predictions

DISCUSSION

Based on the assumptions of results it was discovered that there were 10182 states and eight attributes in the data having 19928 rows. The maximum number of COVID-19 confirmed cases reported were 299691 whereas the number of deaths reported was 27682. The number of recovered cases were 132929 which justifies the fact that the death ratio is less and patients experience symptoms relating to COVID-19 mostly recover within a specific period. The highest percentage of error was recorded on random forest i.e. 89.4% whereas the lowest percentage was of decision tree i.e. 61.4% Gradient boosted trees result in response time was high i.e. 10 min 38s as compared to other two algorithms. Performance analysis was used to calculate root mean square error absolute error, relative error and correlation Prediction chart was used to show how much better a machine learning model performs as compared with a random guess.

CONCLUSION AND RECOMMENDATION

The results show the fact that the death ratio is less which justifies the fact that patients experience symptoms relating to COVID-19 mostly recover within a specific period and by adapting isolation and precautionary measures. This research recommends that we can predict the cases or number of deaths by defining certain algorithms and classifiers and based on their results we can achieve our research objectives. COVID-19 impacted a lot of panics worldwide many developed and underdeveloped countries suffered a huge loss relating to the economy, inflation and unemployment, COVID-19 has changed the mindset of the people and have raised awareness to the people so that they can mentally prepare themselves for the outbreak and develop strategies for their better future.

REFERENCES

[1] Peng, L., Yang, W., Zhang, D., Zhuge, C., & Hong, L. (2020). Epidemic analysis of COVID-19 in China by dynamical modeling. arXiv preprint arXiv:2002.06563.

[2] Bi, Q., Wu, Y., Mei, S., Ye, C., Zou, X., Zhang, Z., ... & Gao, W. (2020). Epidemiology and Transmission of COVID-19 in Shenzhen China: Analysis of 391 cases and 1,286 of their close contacts. MedRxiv.

[3] Schwartz, D. A. (2020). An analysis of 38 pregnant women with COVID-19, their newborn infants, and maternal-fetal transmission of SARS-CoV-2: maternal coronavirus infections and pregnancy outcomes. Archives of pathology & laboratory medicine.

[4] A. (2020). An analysis of 38 pregnant women with COVID-19, their newborn infants, and maternal-fetal transmission of SARS-CoV-2: maternal coronavirus infections and pregnancy outcomes. Archives of pathology & laboratory medicine.

[5] Pedersen, M. G., & Meneghini, M. (2020). Quantifying undetected COVID-19 cases and effects of containment measures in Italy. ResearchGate Preprint (online 21 March 2020) DOI, 10.

[6] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

[7] J.S R Jang (1993). ANFIS Adaptive Network Based Fuzzy inference System. IEEE Transaction on Systems, Man and Cybernetics. Vol.23, no3, pp 665-685

[8] Mansour Y (1997). Pessimistic decision tree pruning based on tree size. In Press of Proc. 14th International Conference on Machine Learning. Pp.195-201.

[9] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In Advances in neural information processing systems (pp. 3146-3154).

[10] Graczyk, M., Lasota, T., & Trawiński, B. (2009, October). Comparative analysis of premises valuation models using KEEL, RapidMiner, and WEKA. In International conference on computational collective intelligence (pp. 800-812). Springer, Berlin, Heidelberg.

[11] Covid, C. D. C., & Team, R. (2020). Severe outcomes among patients with coronavirus disease 2019 (COVID-19)—United States, February 12–March 16, 2020. *MMWR Morb Mortal Wkly Rep*, *69*(12), 343-346.

[12] Bai, Y., Yao, L., Wei, T., Tian, F., Jin, D. Y., Chen, L., & Wang, M. (2020). Presumed asymptomatic carrier transmission of COVID-19. *Jama*, *323*(14), 1406-1407.

[13] Zu, Z. Y., Jiang, M. D., Xu, P. P., Chen, W., Ni, Q. Q., Lu, G. M., & Zhang, L. J. (2020). Coronavirus disease 2019 (COVID-19): a perspective from China. *Radiology*, 200490.

[14] Remuzzi, A., & Remuzzi, G. (2020). COVID-19 and Italy: what next?. *The Lancet*.

[15] Liu, W., Zhang, Q., Chen, J., Xiang, R., Song, H., Shu, S., ... & Wu, P. (2020). Detection of Covid-19 in children in early January 2020 in Wuhan, China. *New England Journal of Medicine*, *382*(14), 1370-1371.

[16] Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*, *20*(5), 533-534.

[17] World Health Organization. (2020). Coronavirus disease 2019 (COVID-19): situation report, 88.

[18] Liu, Yang, Li-Meng Yan, Lagen Wan, Tian-Xin Xiang, Aiping Le, Jia-Ming Liu, Malik Peiris, Leo LM Poon, and Wei Zhang. "Viral dynamics in mild and severe cases of COVID-19." *The Lancet Infectious Diseases* (2020).

[19] Dong, Yuanyuan, Xi Mo, Yabin Hu, Xin Qi, Fan Jiang, Zhongyi Jiang, and Shilu Tong. "Epidemiology of COVID-19 among children in China." *Pediatrics* 145, no. 6 (2020).

[20] Le, T. T., Andreadakis, Z., Kumar, A., Roman, R. G., Tollefsen, S., Saville, M., & Mayhew, S. (2020). The COVID-19 vaccine development landscape. *Nat Rev Drug Discov*, *19*(5), 305-306.

[21] Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, *21*(3), 660-674.

[22] Freund, Y., & Mason, L. (1999, June). The alternating decision tree learning algorithm. In *icml* (Vol. 99, pp. 124-133).

[23] Utgoff, P. E., Berkman, N. C., & Clouse, J. A. (1997). Decision tree induction based on efficient tree restructuring. *Machine Learning*, *29*(1), 5-44.

[24] Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, *28*(1), 71-72.