ANALYSIS OF COVID-19 IN THE UNITED STATES USING MACHINE LEARNING

James G. Koomson

School of Business and Technology, Marymount University, Arlington, Virginia, USA

Abstract

The unprecedented outbreak of COVID-19 also known as the coronavirus has caused a pandemic like none ever seen before this century. Its impact has been massive on a global level. The deadly virus has commanded nations around the world to increase their efforts to fight against the spread of the virus after the stress it has put on resources. With the number of new cases increasing day by day around the world, the objective of this paper is to contribute towards the analysis of the virus by leveraging machine learning models to understand its behavior and predict future patterns in the United States (US) based on data obtained from the COVID-19 Tracking Project.

KEYWORDS

COVID-19, machine learning, deep learning, prediction, pandemic, linear regression, correlation, SARS-COV-2, Coronavirus Outbreak

1. INTRODUCTION

Originating in Wuhan, China around December 2019, the SARS-CoV-2 coronavirus disease (COVID-19) is a new strain of coronavirus that had previously not been identified in humans [1]. The SARS-CoV-2 belongs to the same coronavirus group (Betacoronavirus) as SARS and MERS viruses that caused two of the more severe epidemics in recent years. As with SARS and MERS, this new coronavirus, 2019-nCoV, is believed to be of zoonotic origin, but may also be transmitted through the respiratory tract, by direct contact, and possibly via patients' excreta which may contain the living virus [2].

To date, at the rate of the virus spread globally it has become a serious infectious disease affecting human health worldwide [3]. In this unprecedented time of self-isolation and stay-at-home orders, Americans were online - working, watching, communicating, and yes, searching for medical information online and answers to the questions during the coronavirus outbreak [5]. In the absence of specific therapeutic drugs or vaccines for the 2019 novel coronavirus disease (COVID-19), it is essential to detect the diseases at an early stage and immediately isolate the infected person from the healthy population. The key features of COVID-19 are respiratory symptoms with a fever and cough. Like all new infections, understanding COVID-19 is important and changes rapidly. In the United States, the CDC is proactively monitoring the virus and taking measures like providing guidance for healthcare workers and issuing travel recommendations [1]. To this end, the objective of this paper is to conduct analysis utilizing machine learning predictions, correlation, and linear regression of COVID-19 dataset metrics that are reflective of the state numbers within the United States to provide greater insight into the virus that is plague citizens throughout the world.

DOI:10.5121/mlaij.2021.8101

1.1. Existing Work

During the COVID-19 pandemic, there has been interest by researchers in analyzing COVID-19 datasets utilizing machine learning to gain a deeper understanding of the trends and patterns caused by the virus on the human population. There has been a sizable amount of literature that has been published on the utilization of machine linear regression models, and others. Some existing works will be stated in this section. The machine learning algorithm model known as Support Vector Regression was utilized in literature to do a comparative analysis of COVID-19 results between different countries such as China, Italy, South Korea, and others [6]. For the analysis of COVID-19 several machine learning algorithms have been used for regression analysis like linear and polynomial for the country of India for prediction purposes [7]. Existing research works attempt to analyze datasets for respective countries around the world, however, no literature exists that I'm aware of, which leverages machine learning models to predict survival rates and illustrate the correlation of variables, and comparative analysis based on the COVID-19 dataset between states within the United States of America.

2. METHODOLOGY

The approach to this paper was to examine the COVID-19 dataset from September 27, 2020, obtained from the COVID-19 Tracking Project at www.covidtracking.com. Since the virus spreads majorly through human contact, it was essential to consider only the dataset reflective of COVID-19 numbers at the state level to ensure high-level analysis is achieved. Needing to convey a high-level analysis of the metrics, consideration was given to the dataset which has the number of tests conducted in each state at the daily level in the United States. The Python Scikit-Learn library for machine learning models was implemented for the analysis of the dataset. The machine learning models leveraged within Python for analysis of the coronavirus dataset were correlation, and linear regression models respectively. Correlation is a statistical technique that determines how one variables moves/changes in relation to the other variables. This gives a sense of the idea of the relationship between the two variables [4]. It is a bi-variate analysis measure that describes the association between different variables. It is useful to express one subject in terms of its relationship with others. Linear Regression is also utilized to search for linear relationships between variables/features within the COVID-19 dataset. The algorithm uses linear regression for prediction and could work with weighted instances. This method of regression is simple and provides an adequate and interpretable description of how the input affects the output [8]. This model was considered due to the suspicion that some features within the dataset influence others. The dataset was split and trained with 20% of the data utilized for testing, and 80% of data utilized for training. The objective was to return four datasets, and X set (split between training and test set) and returns the dependent variable y_train and test as well which validates results. In training the model, the X_train and the correct values in y_train were passed. Once the model was trained and fitted it was used to predict the survival rate of COVID-19 patients based on the test data. The mean squared error was also calculated to determine how close a regression line is to a set of points which is important to remove any negative sign as the goal is to find the average of set errors.

3. RESULTS AND DISCUSSION

The experimental observational results are presented in detail in table and figure formats respectively from the machine learning models previously stated. Table.1 represents the sum, mean, and median values of positive COVID-19 cases in the United States. The results graph in Figure 2. Represents the cumulative number of patient hospitalizations in the United States and shows the rise in number from the start of the pandemic in March 2020 continuing into April

2020. In Figure 3., the graph represents the trending numbers in the rise of deaths as a result of COVID-19 from the start of the pandemic in March 2020 continuing into April 2020. Fig. The results from Table 2., represent the top 10 states with the highest death amounts as a result of COVID-19. The total amount of positives cases is also displayed, as well as the death rate for the state. In reviewing the metrics of this table, it appears that the state of New York had at the time the most deaths, the highest positive cases, and the highest death rate due to COVID-19 of .04. Comparatively, the state of Georgia had the lowest number of COIVD-19 deaths relatively speaking, the lowest amount of positives cases, however, it had the same death rate at .04 as the state of New York. Table 2., represents 16 variables and shows the correlation table for positive, negative, hospitalized currently, hospitalized cumulatively, in ICU currently, in ICU cumulatively, recovered, deaths, hospitalized, total testing including pending, and total tests in the United States. It is observed from the table that there is a strong correlation between positive cases and deaths with a coefficient value of 0.96. It is observed that there is also a strong positive correlation between patients who are hospitalized and those that have recovered from COVID-19 with a coefficient value of 0.97. There are also many other notable variable correlations on as seen in Table 2. Figure 4., results highlight the top ten states with the highest amount of positive COVID-19 cases from the start of the pandemic in March 2020 continuing into April 2020. Figure 5., the results graph highlights the top ten states with the lowest amount of positive COVID-19 cases from the start of the pandemic in March 2020 continuing into April 2020. The prediction plot in Figure 6., shows the survival rate of patients that test positive for COVID-19 and also illustrates that as X increases, the amount recovered (Y) also goes up. Therefore, it is observed that this graph is linear. The mean squared error is calculated to be 8.3597.0638019641, which is due to several states reporting no cases at the start of the pandemic.

Table 1. Sum, Mean, and Median Values of Positive COVID-19 Cases

Sum, Mean, and Median Values of Positive COVID-19 Cases	
Data type	**Result**
Sum of Positive COVID-19 cases	3168431.0
Mean of Positive COVID-19 cases	1739.940142778693
Median of Positive COVID-19 cases	108.0

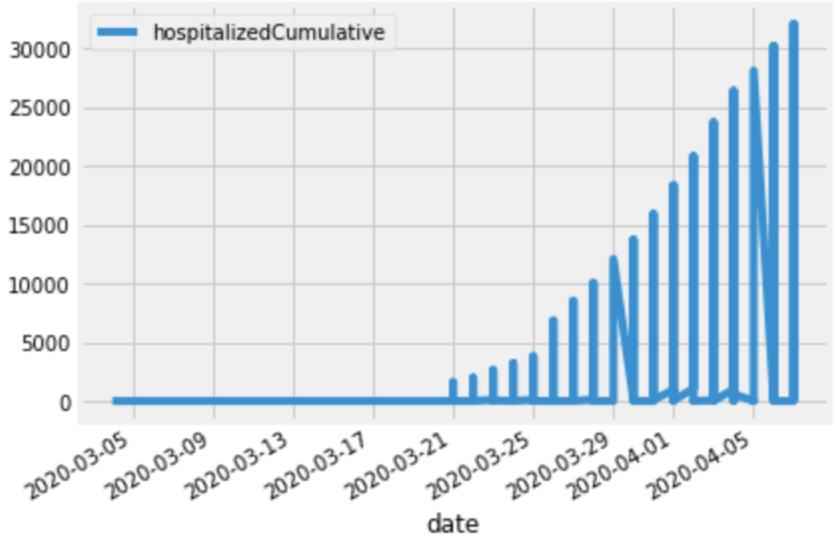


Figure 1. Total patient hospitalizations from COVID-19 cases

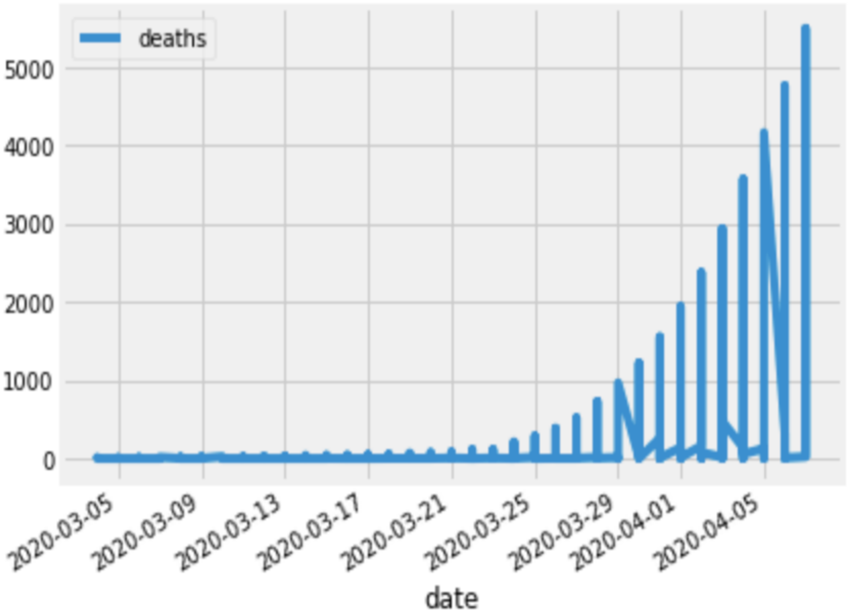
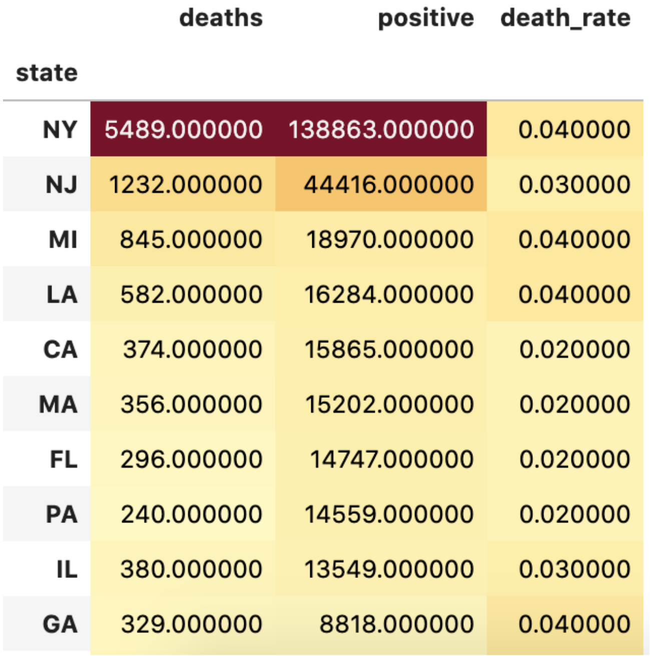


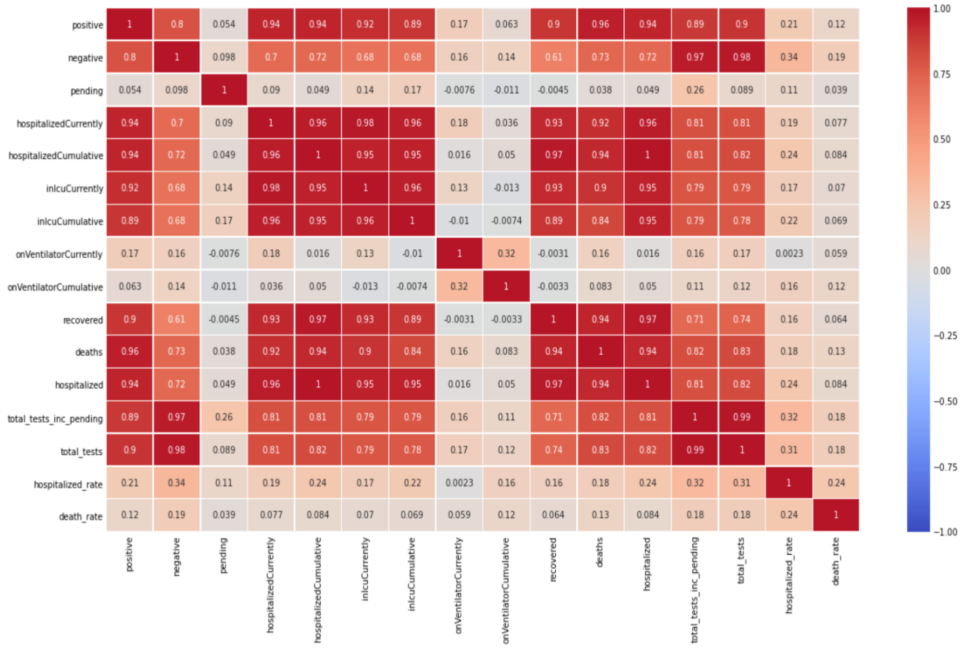
Figure 2. Trending rise of deaths from COVID-19 cases

Table 2. Top 10 States with the highest deaths from COVID-19



	deaths	positive	death_rate
state			
NY	5489.000000	138863.000000	0.040000
NJ	1232.000000	44416.000000	0.030000
MI	845.000000	18970.000000	0.040000
LA	582.000000	16284.000000	0.040000
CA	374.000000	15865.000000	0.020000
MA	356.000000	15202.000000	0.020000
FL	296.000000	14747.000000	0.020000
PA	240.000000	14559.000000	0.020000
IL	380.000000	13549.000000	0.030000
GA	329.000000	8818.000000	0.040000

Table 3. Correlation Table



	positive	negative	pending	hospitalizedCurrently	hospitalizedCumulative	inIcuCurrently	inIcuCumulative	onVentilatorCurrently	onVentilatorCumulative	recovered	deaths	hospitalized	total_tests_inc_pending	total_tests	hospitalized_rate	death_rate
positive	1	0.8	0.054	0.94	0.94	0.92	0.89	0.17	0.063	0.9	0.96	0.94	0.89	0.9	0.21	0.12
negative	0.8	1	0.098	0.7	0.72	0.68	0.68	0.16	0.14	0.61	0.73	0.72	0.97	0.98	0.34	0.19
pending	0.054	0.098	1	0.09	0.049	0.14	0.17	-0.0076	-0.011	-0.0045	0.038	0.049	0.26	0.089	0.11	0.039
hospitalizedCurrently	0.94	0.7	0.09	1	0.96	0.98	0.96	0.18	0.036	0.93	0.92	0.96	0.81	0.81	0.19	0.077
hospitalizedCumulative	0.94	0.72	0.049	0.96	1	0.95	0.95	0.016	0.05	0.97	0.94	1	0.81	0.82	0.24	0.084
inIcuCurrently	0.92	0.68	0.14	0.98	0.95	1	0.96	0.13	-0.013	0.93	0.9	0.95	0.79	0.79	0.17	0.07
inIcuCumulative	0.89	0.68	0.17	0.96	0.95	0.96	1	-0.01	-0.0074	0.89	0.84	0.95	0.79	0.78	0.22	0.069
onVentilatorCurrently	0.17	0.16	-0.0076	0.18	0.016	0.13	-0.01	1	0.32	-0.0031	0.16	0.016	0.16	0.17	0.0023	0.059
onVentilatorCumulative	0.063	0.14	-0.011	0.036	0.05	-0.013	-0.0074	0.32	1	-0.0033	0.083	0.05	0.11	0.12	0.16	0.12
recovered	0.9	0.61	-0.0045	0.93	0.97	0.93	0.89	-0.0031	-0.0033	1	0.94	0.97	0.71	0.74	0.16	0.064
deaths	0.96	0.73	0.038	0.92	0.94	0.9	0.84	0.16	0.083	0.94	1	0.94	0.82	0.83	0.18	0.13
hospitalized	0.94	0.72	0.049	0.96	1	0.95	0.95	0.016	0.05	0.97	0.94	1	0.81	0.82	0.24	0.084
total_tests_inc_pending	0.89	0.97	0.26	0.81	0.81	0.79	0.79	0.16	0.11	0.71	0.82	0.81	1	0.99	0.32	0.18
total_tests	0.9	0.98	0.089	0.81	0.82	0.79	0.78	0.17	0.12	0.74	0.83	0.82	0.99	1	0.31	0.18
hospitalized_rate	0.21	0.34	0.11	0.19	0.24	0.17	0.22	0.0023	0.16	0.16	0.18	0.24	0.32	0.31	1	0.24
death_rate	0.12	0.19	0.039	0.077	0.084	0.07	0.069	0.059	0.12	0.064	0.13	0.084	0.18	0.18	0.24	1

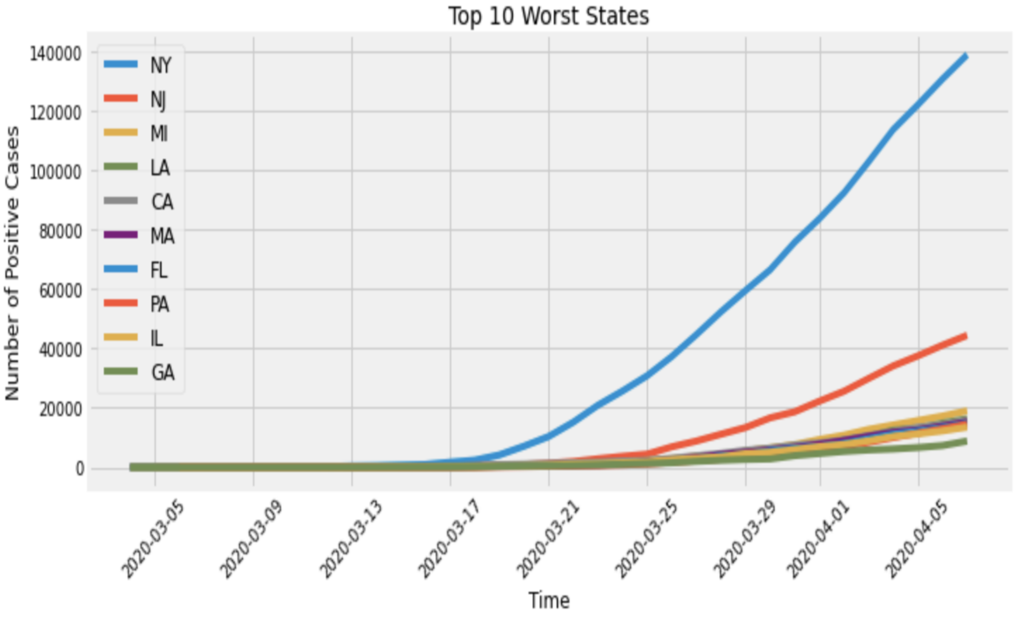


Figure 3. Top 10 States with the highest COVID-19 positive cases

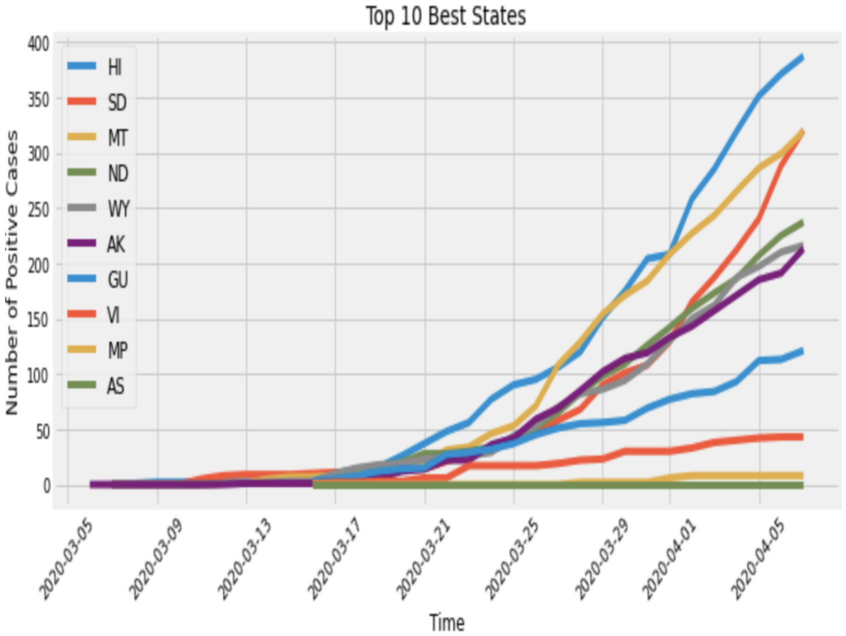
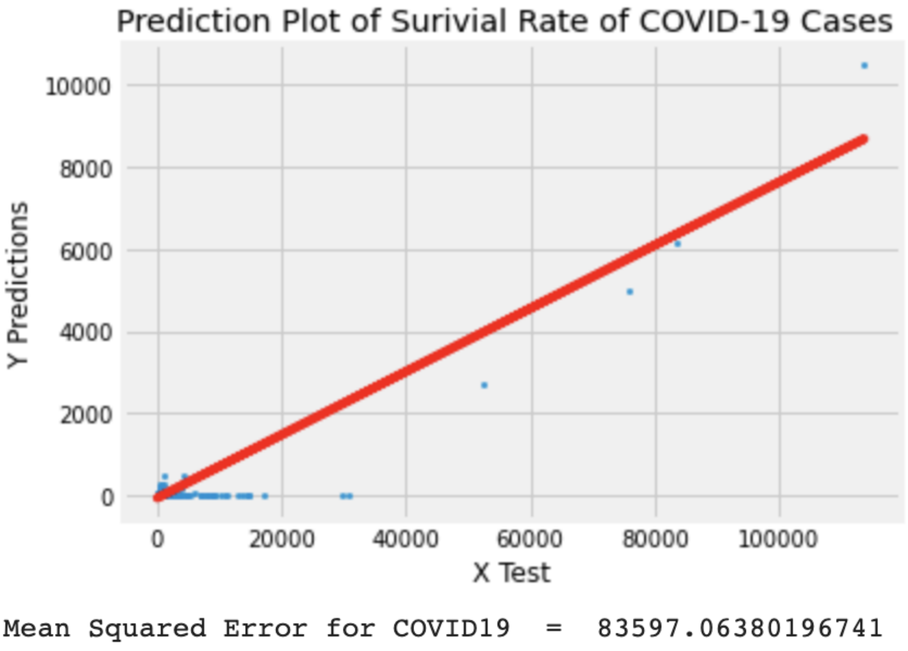


Figure 5. Top 10 States with the lowest COVID-19 positive cases



```
Mean Squared Error for COVID19  =  83597.06380196741
```

Figure 6. Prediction of Survival Rate of COVID-19 Cases

4. CONCLUSION

With the COVID-19 cases and deaths raising around the world and particularly in the United States, the pandemic has taken a serious toll on people from all walks of life. Researchers are still learning more about this virus daily. Although machine learning consists of a full range of models, the models chosen for this research showed promising results. In this paper, at a state level, key variables were correlated, and linear regression algorithms were leveraged successfully which yielded a linear prediction of the survival rate of patients who have tested positive for COVID-19. The results of this research, also suggest that machine learning is a powerful tool where the correlation of variables and predictions amongst other things can be successfully realized through machine learning models. With COVID-19 ushering in many unknowns due to its highly complex nature, this research could aid states to be prepared for what will happen in the future by receiving a better understanding of the virus that has plagued millions of citizens. Future works from this study could explore COVID-19 datasets comparatively between counties or parishes within a given state in the US using other machine learning models.

ACKNOWLEDGMENTS

I would like to extend special thanks to Dr. Nathan Green at Marymount University in Arlington, VA, USA for his support and appreciated advice during the process of leveraging machine learning to analyze the COVID-19 dataset for the United States.

REFERENCES

[1] Mayo Clinic. (2020). Coronavirus disease 2019 (COVID-19). Available at: https://www.mayoclinic.org/diseases-conditions/coronavirus/symptoms-causes/syc-20479963

[2] H.-W. Zhang, J. Yu, H. J. Xu et al., "Coronavirus international public health emergencies: implications for radiology management," Academic Radiology, vol. 27, no. 4, pp. 463–467, 2020.

[3] C. Li, Y. Yang, and L. Ren, "Genetic evolution analysis of 2019 novel coronavirus and coronavirus from other species," Infection, Genetics and Evolution, vol. 82, p. 104285, 2020.

[4] A. Upadhyay. (5 June 2019). "What Is Correlation in Machine Learning?". Available at: https://medium.com/analytics-vidhya/what-is-correlation-4fe0c6fbed47

[5] D. Wyld. (May 2020). "An Analysis of Consumer Interest Level Data for Online Health Information in the Early Stages of the Covid-19 Pandemic". Available at: https://aircconline.com/ijmit/V12N2/12220ijmit01.pdf

[6] M. Yadav, M. Perumal, and M. Srinivas, "Analysis on novel coronavirus (COVID-19) using machine learning methods," Chaos, Solitons & Fractals, vol. 139, p. 110050, 2020.

[7] P. Raji and G. R. Lakshmi (October 2020) "Covid-19 pandemic Analysis using Regression". Available: https://www.researchgate.net/publication/346204738_Covid-19_pandemic_Analysis_using_Regression.

[8] L. McClendon and N. Meghanathan, "Using Machine Learning Algorithms to Analyze Crime Data," vol. 2, pp. 1–12, Mar. 2015, doi: 10.5121/mlaij.2015.2101.

AUTHOR

James Koomson, Doctoral Student in Cybersecurity at Marymount University