COVID-19 Predictions

DATA 621 Final Project

David Moste, Sadia Perveen, and Vanita Thompson

Abstract

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. COVID-19 is caused by SARS- CoV-2. It has resulted in a worldwide pandemic that has put adults at risk of serious illness or death. The COVID-19 pandemic has challenged our worldwide healthcare resources. We have proposed analytical model to predict future COVID-19 cases using global COVID-19 data. The proposed model will work with pervious confirmed cases, deaths reported, recovered cases and daily cases. The proposed model can help us better prepare our future healthcare needs.

Key Words

COVID-19, predictive model, infectious disease, future cases

Introduction

The World Health Organization declared COVID-19 outbreak an epidemic in March of 2020. COVID-19 is a respiratory disease caused by SARS-CoV-2. It causes illness similar to a cold, SARS (severe acute respiratory syndrome) and MERS (Middle East respiratory syndrome). The virus is thought to spread mainly from person to person through respiratory droplets produced when an infected person coughs, sneezes, or talks. Although some individuals will not have any symptoms, others may have mild to severe symptoms. Individuals with underlying conditions and adults over the age of 65 are at higher risk of severe illness.

The COVID-19 pandemic continues to pose threat to the wellbeing of individuals worldwide. As of May 21, 2021 there have been over 165,158,000 confirmed cases of COVID-19, including over 3,425,000 deaths, reported to WHO. As the number of COVID-19 cases grow, it continues to challenge our healthcare system. “International hospitals and healthcare facilities are facing catastrophic financial challenges related to the COVID-19 pandemic… Overall, a lack of preparedness was a major contributor to the struggles experienced by healthcare facilities around the world. Items such as personal protective equipment (PPE) for healthcare workers, hospital equipment, sanitizing supplies, toilet paper, and water were in short supply. These deficiencies were exposed by COVID-19 and have prompted healthcare organizations around the world to invent new essential plans for pandemic preparedness.”(Kaye, A. D., Okeagu)

In this paper, we propose an analytical model to predict future COVID-19 cases using global COVID-19 data. The proposed model will work with pervious confirmed cases, deaths reported, recovered cases and daily cases. The proposed model can help us better prepare our future healthcare needs. The unpredictive needs of the pandemic has resulted in a personal and economical loss. “The American Hospital Association estimates a financial impact of $202.6 billion in lost revenue for America's hospitals and healthcare systems, or an average of $50.7 billion per month.” (Kaye, A. D., Okeagu)

Literature Review

Discuss how other researchers have addressed similar problems, what their achievements are, and what the advantage and drawbacks of each reviewed approach are. Explain how your investigation is similar or different to the state-of-the- art. Please cite the relevant papers where appropriate.

In Wuhan City, Hubei Province, China an outbreak of the coronavirus disease occurred. The World Health Organization(WHO) declared the outbreak as a public health emergency of international concern on January 30, 2020. Due to the recent nature of the pandemic, it must be noted that limited research is available. As the COVID-19 virus is still unknown to the world several studies have attempted to predict future COVID-19 outcomes.

On approach used in predicting COVID-19 cases is based on symptoms. Zoabi, Y., Deri-Rozov, S. & Shomron, N. established a machine-learning approach that model predicted COVID-19 test results with high accuracy using only eight binary features: sex, age ≥60 years, known contact with an infected individual, and the appearance of five initial clinical symptoms. The purpose of their study was to prioritize the use of testing resources on individuals who have a higher chance of testing positive for COVID-19 based on the model. The research found that "fever and cough were key to predicting contraction of the disease. As expected, close contact with an individual confirmed to have COVID-19 was also an important feature, thus corroborating the disease's high transmissibility15 and highlighting the importance of social distancing." As noted in the research, there were several drawbacks. "We relied on the data reported by the Israeli Ministry of Health, which has limitations, biases and missing information regarding some of the features. For example, for patients labeled as having had contact with a person confirmed to have COVID-19, additional information such as the duration and location (indoors/outdoors) of the contact was not available." A similar research done by Dr. Pallavi Mirajkar1, Dr. Rupali Dahake concluded that an analytical model can be employed to predict outbreak spreading trend are at high risk of developing Covid complications.

To better prepare our healthcare industry it would be beneficial to predict future COVID-19 cases. Hongwei Zhao et al. "describe a new approach that forecasts the number of incident cases in the near future given past occurrences using only a small number of assumptions" Using a Poisson distribution for the daily incidence number, and a gamma distribution for the series interval they modeled the observed incidence cases. They then estimated the effective reproduction number assuming its value stays constant during a short time interval and draw future incidence cases from their posterior distributions. Their model was focused on COVID-19 data available on Texas. Hongwei Zhao et al. stated "Our method produces reasonably accurate results when the effective reproduction number is distributed similarly in the future as in the past. Large deviations from the predicted results can imply that a change in policy or some other factors have occurred that have dramatically altered the disease transmission over time." Some of the drawbacks of the study included complexity inherent in how data are collected. Hongwei Zhao et al. stated "Some major complexities of the data include: policies about testing algorithms (e.g. which suspect cases are tested); if screenings or surveillance is conducted, which diagnostic test is acceptable or required for reporting; accessibility and availability of testing; administrative issues such as reporting requirements, procedures, and infrastructure." In our model we focus on global COVID-19 data. The proposed model will work with pervious confirmed cases, deaths reported, recovered cases and daily cases.

Methodology

Data Description: Our data was obtained from the data repository for the 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). The data contains worldwide daily covid-19 confirmed cases, deaths reported, and recovered cases by location. We also have additional latest data that gave us overall confirmed cases, deaths reported, recovered cases by location. At last, we have data on incident rate and case to fatality ratio by location.

Data Cleaning and Preparation: All the data was first pulled into the workbook. We prepared our data by first removing the individual dates and then summing our confirmed cases, deaths reported, and recovered cases. Our data showed 165,531,431 confirmed cases, 3,430,326 deaths reported and 101,821,428 cases recovered worldwide. We did the same for each country and built tables and CMAPS to better visualize our data.

Exploratory Data Analysis: After preparing our data we built tables to show number of Confirmed Cases, Number of Deaths, Number of Recoveries, Number of Active Cases, and Mortality Rate. Each table was accompanied by a cmap that highlighted extreme situations. We followed this by making a couple charts (bar and pie) to show the most infected countries and provinces. We finished our EDA by graphing COVID-19 deaths, cases, and recoveries over time.

Our EDA shows that when looking at country level, the United States has the highest number of confirmed cases, with India having the second highest. When we look at the Province/State Name level we see that Maharashtra, India has the highest number of confirmed cases. When viewing our pie chart we see that the United States, India and Brazil combined account for more than half of worldwide cases. At last when looking at states within the United States we see that California has the highest number of confirmed cases.

Experimentation and Results

Model Building: To see if we could predict future confirmed cases, we built two models: our first model was a linear regression and our second model was an SVM.

Using the sci-kit learn package, we built a linear regression model after splitting our data into training and test sets. Before running our regression model, we transformed the data into a form that would allow for a second order polynomial regression, as we noticed that the data had a clear curve to is and was not exactly linear. For the first 400 days after January 1, 2020 the model appears to have pretty good agreement with the data.

Using the same package, we also built a polynomial SVM regression model to fit our data. This model used our original data rather than the polynomial transformation. It ended up producing a nearly identical model to our linear regression.

Results: We ended up with two models that predict future cases of COVID-19 around the world. Both models indicate a continued increase in cases, which could be refined with more data.

Any increase is unlikely to meet the expectations of our model as people become vaccinated and transmission slows over time.

Discussion and Conclusions

In conclusion we found continued increase in cases worldwide. Although our data was very intensive it had some limitations. For one the method in collecting the data is not concise. Data was collaborated from several different sources which can result in higher chance of error. Additionally, worldwide there is no clear guidance on collecting COVID-19 data. There is still a lot more research and guidance needed on COVID-19.

References

Kaye, A. D., Okeagu, C. N., Pham, A. D., Silva, R. A., Hurley, J. J., Arron, B. L., Sarfraz, N., Lee, H. N., Ghali, G. E., Gamble, J. W., Liu, H., Urman, R. D., & Cornett, E. M. (2020). Economic impact of COVID-19 pandemic on healthcare facilities and systems: International perspectives. Best Practice & Research. Clinical Anaesthesiology, Advance online publication. https://doi.org/10.1016/j.bpa.2020.11.009

Pallavi Mirajkar, Rupali Dahake,, "Predictive System of COVID -19 Using Response Based Analytical Model", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456- 3307, Volume 7 Issue 2, pp. 05-10, March-April 2021.

Zhao H, Merchant NN, McNulty A, Radcliff TA, Cote MJ, et al. (2021) COVID-19: Short term prediction model using daily incidence data. PLOS ONE 16(4): e0250110.https://doi.org/10.1371/journal.pone.0250110

Zoabi, Y., Deri-Rozov, S. & Shomron, N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj Digit. Med. 4, 3 (2021). https://doi.org/10.1038/s41746-020-00372-6

Appendices

Supplemental tables and/or figures (attached in our notebook).

Python programming code (attached as a notebook).