A Project report

on

(Data-analysis-on-COVID-19 Vaccinations)

Submitted By

Gali Yaswanth(20bcs046)

Ameyotosh Michael Roy(20bcs010)

Under the guidance of

Dr. Uma Sheshadri

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY

DHARWAD

Introduction

COVID-19 is an infectious disease caused by severe acute respiratory syndrome virus 2 (SARS-CoV2). The COVID-19 virus has been declared a pandemic by the World Health Organization (WHO) with more than ten million cases and 503,862 deaths across the world as per WHO statistics of 30 June 2020. It is the third epidemic caused by coronaviruses in the twenty-first century. The serious acute respiratory syndrome (SARS) epidemic emerged in China in 2003 and spread to several countries. Middle East respiratory syndrome virus (MERS) emerged in 2012 in Saudi Arabia and spread to 27 countries. The disease has become pandemic, affecting almost all nations of the world, and has caused enormous economic, the social, and psychological burdens on countries. The clinical picture of COVID-19 differs from the previous pandemics by a larger proportion of mild cases that may remain active in society and facilitate the spread of the virus. Even though the scale of previous epidemics has been considerably smaller, the literature from heavily affected areas provides valuable information on patient flow dynamics in the face of an epidemic.

The spectrum of the disease ranges from asymptomatic to severe, sometimes requiring prolonged treatment in the intensive care unit. COVID-19 puts serious strain on the ICU and inpatient capacity of healthcare systems and has been mitigated by suspending non-urgent care. Hygiene and educational campaign programs have been identified to be potent public health interventions that can curtail the spread of ics by a larger proportion of mild cases who may remain active in the society and facilitate the spread of the virus. During the current coronavirus pandemic, monitoring the evolution of COVID-19 cases is of utmost importance for the authorities to make informed policy decisions (e.g., lock-downs), and to raise awareness in the general public for taking appropriate public health measures.

At the time of the pandemic outbreak, a lack of laboratory tests, materials, and human resources implied that the evolution of officially confirmed cases did not represent the total number of cases. Even now, there are significant differences across countries in terms of the availability of tests. For this reason, given the rapid progression of the pandemic, in some cases health authorities are forced to make important decisions based on sub-optimal data. For this reason, alternatives to testing that can be rapidly deployed are likely to help authorities, as well as the general population, to better understand the progress of a pandemic, particularly at its early stages or in low-income countries, where massive testing is unfeasible.

Results and discussion

Thorough data exploration was done on the taken datasets. For interactive plots, we have used Plotly and its sub-components Plotly graphs, Plotly express, and for other figures, we have explicitly used seaborn package and matplotlib. The inference was done on some of the selected countries like the USA, Canada, India, Germany, and the European region.

Key Findings and insights:-

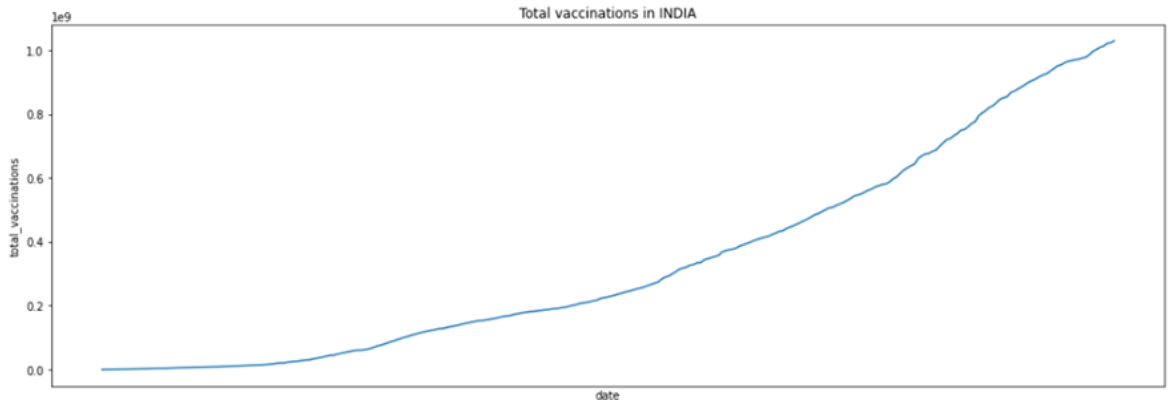
1.)The most vaccinated country is China owing to its large population

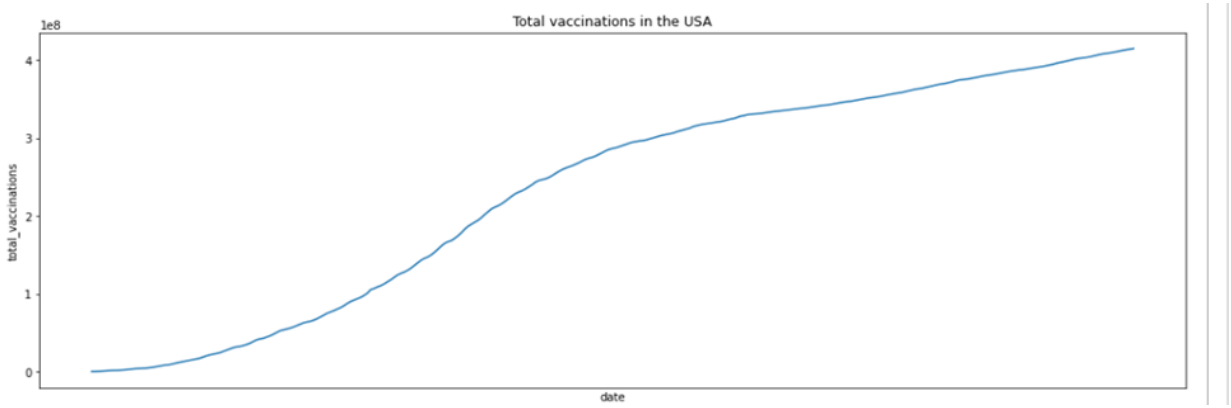
2.)Most of the African countries are least vaccinated.

3.) The Highest percentage of people vaccinated are in the Australian and American region

4.)The BoiNtech vaccine is the most used vaccine and highest used in the European region

5.)Although the USA was most affected with Covid-19, the vaccination of people increased dramatically when compared to India whose progress is slow and steadily increasing.





Model

The model used for time series forecasting is Prophet which was created by the FaceBook research team and is commonly known as fbprophet.

The key features of Prophet are:-

1.)Fully automatic

2.)Accurate and Fast

3.)Easy tuning of hyperparameters

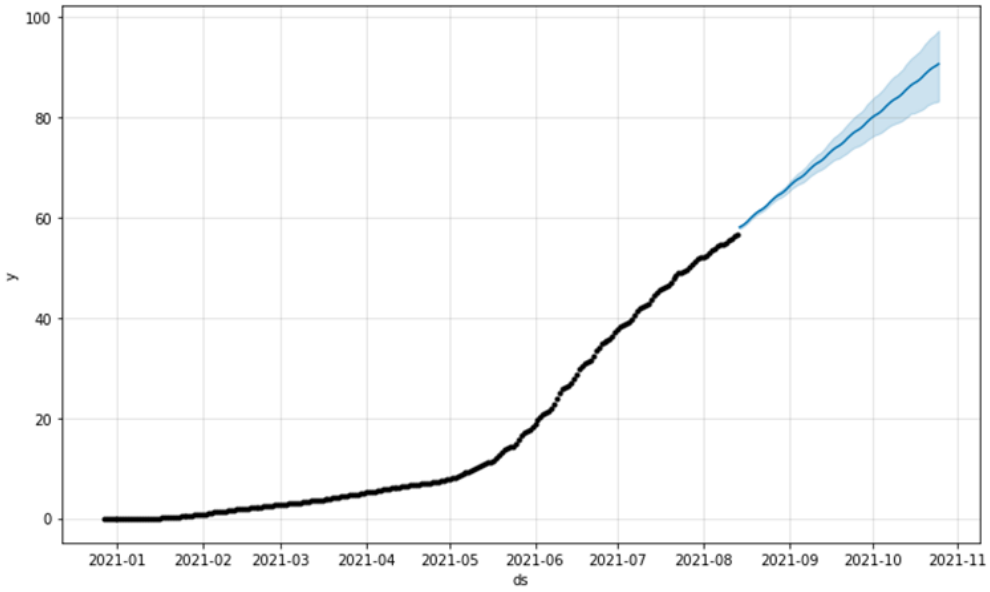
The main reason for using Prophet is that it is most useful and efficient for univariate analysis.

From the dataset, the target column is people vaccinated per hundred which is used to train the prophet model. Now, prophet expects the data to be in a certain format that the date column should be indexed and renamed as ‘ds’ and the target column must be renamed as ‘y’ for the model to identify.

The model was trained on Germany’s data collected from the original datasets

which had data from '2021-02-22' to'2021-10-26'

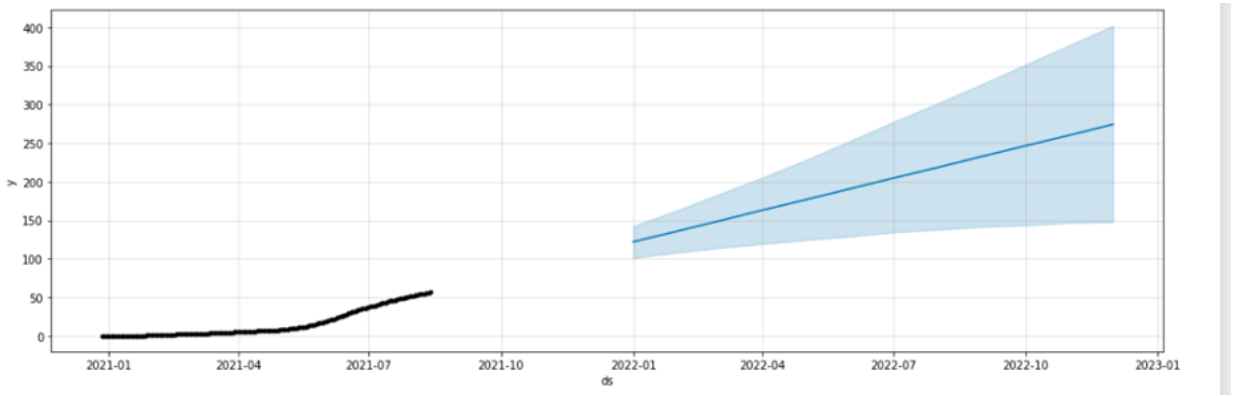
The predictions on the test set are as shown below:



When compared with the target variable from the test dataset the **error (MSE)** was found out to be **1.11** which is neither too large nor low.

Often, In Time Series Forecasting we need to predict not only for validation or test set but for DateTime in the future which is not present in the dataset. So, we create a DateTime data frame of future dates that are in the 2022 year and predict those dates.

The predictions for the 2022 year are as follows:-



From the above forecast, we can predict that by the end of April the people in Germany will be vaccinated completely given that our model had only 300 data points for Germany so we can't expect our model to be highly accurate and thus we can add a buffer of additional two months, Thus by end of May at most the people in Germany will be completely vaccinated.

According to current statistics in Germany, approximately 72% of people are vaccinated with 1st dose and 58% of people are vaccinated completely

Conclusion

Covid-19 has hit us hard and on top of that, each and every sector was severely affected by this pandemic. The only cure for this pandemic is vaccination and it is mandatory for you and your loved ones. Thus, we have carried out data analysis on vaccination data of each and every country and found out some of the meaningful insights which help us to understand the vaccination programs conducted all around the globe. For data visualization, we have used Plotly

excessively along with seaborn and matplotlib. For Time Series Forecasting, the machine learning used is Prophet as it works efficiently for univariate analysis. The model was trained on German vaccination data extracted from datasets. As predicted by the model, all the people in Germany will be vaccinated completely by at most May 2022.

References

1.) Time Series Forecasting Using Facebook FbProphet

2.)https://facebook.github.io/prophet/

3.)https://www.kaggle.com/gpreda/covid-world-vaccination-progress

4.)https://www.kaggle.com/desalegngeb/plotly-guide-customize-for-better-visualizations

5.)https://www.kaggle.com/gpreda/covid-19-vaccination-progress