MACHINE LEARNING PROJECT PROPOSAL

Understanding the increase of COVID-19 Recovery, Confirmed and Death Cases and future prediction in WORLD

**Group members:**

Tejasvita Pandey

Laxmi Vishwakarma

Chetan Sharma

Anuj Soni

**TABLE OF CONTENT**

[**PROJECT SUMMARY**](#_ynx7urofd5fd) **3**

[**OBJECTIVES OF PROJECT**](#_e8juxfc2mxhn) **3**

[**INPUT REQUIRED FOR PROJECT**](#_nrg97wv26u97) **3**

[**EXPECTED OUTPUT AND OUTCOME OF THE PROPOSAL**](#_q4vlaewi2ciz) **4**

[**STATE OF ART**](#_hy6rbsjb2rr3) **5**

[**ORIGIN OF THE PROPOSAL**](#_rdksln7cyeg) **6**

[**SCIENTIFIC RATIONALE AND IMPORTANCE OF THE PROPOSED WORK**](#_8p3ovy8ppwhg) **6**

[**METHODOLOGY**](#_6gosxse6g1dq) **7**

[**PROVIDE LINK FOR DATASETS RELATED TO PROJECT**](#_quk9mcckui8r) **8**

[**BIBLIOGRAPHY**](#_piw3puyocrqr) **8**

# PROJECT SUMMARY

In this project, we have proposed methods for predicting Covid-19 in patients using machine learning techniques. The machine learning techniques that were used include Time series forecasting. We hope that our forecasts will be a useful tool for governments and individuals towards making decisions and taking the appropriate actions to contain the spreading of the virus to the degree possible. Regardless of what one’s beliefs are, we believe that their associated uncertainty can and should be an integral part of the decision-making process, especially in high risk cases. Apart from the significant public health concerns, the dangers imposed on global supply chains and the economy as a whole are also considerable. Risk-averse people can focus on the worst-case-scenarios and act accordingly. Deciding to discard any formal, statistical forecasts and acting conservatively, still implies an underlying forecasting process, even if this process is not formalised.

# OBJECTIVES OF PROJECT

Our goal is to analyze the covid-19 data in programming languages Python using Prediction algorithms based on time series forecasting; and how the governments of different countries can use that prediction for making effective decisions and future strategies. The outcomes should be in graphical form so that anyone can easily understand the trends. Assuming that the data used is reliable and that the future will continue to follow the past pattern of the disease, our forecasts suggest a continuing increase in the confirmed, death and recovery COVID-19 cases with sizable associated uncertainty.

# INPUT REQUIRED FOR PROJECT

Input data is available as csv files in the Johns Hopkins Github repository. Data is extracted from google sheets associated and made available on kaggle. The dataset has daily level information on the number of affected cases, deaths and recovery from 2019 novel coronavirus. Please note that this is a time series data and so the number of cases on any given day is the cumulative number.

The data is available from 22 jan, 2020.

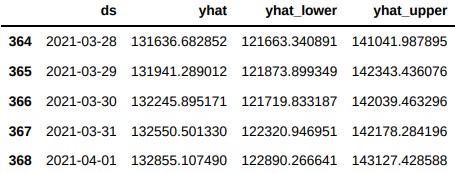
Column Description

* Sno - serial number
* ObservationDate - Date of the observation in MM/DD/YYYY
* Province/State - Province or state of the observation
* Country/Region - Country of observation
* Last Update - Time in UTC at which the row is updated for the given province or country.
* Confirmed - Cumulative number of deaths till that date
* Recovered - Cumulative number of recovered cases till that date

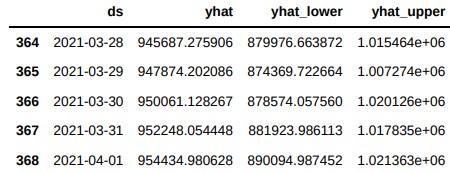
# EXPECTED OUTPUT AND OUTCOME OF THE PROPOSAL

Conclusion of our results is presented in Table below. These results were obtained by applying prophet statistic forecasting. The forecast method will allot each row a predicted value which it means “yhat”. If you pass in historical dates, it'll provide an in-sample fit. The forecast object here could be a new dataframe that has a column “yhat” with the forecast, furthermore as columns for components and uncertainty intervals “yhat-lower, yhat-upper”.

Last five data of Recovered cases forecasting table



Last five data of Confirmed cases forecasting table



Last five data of Death cases forecasting table



# STATE OF ART

The purpose of the project is to focus on the cumulative daily figures aggregated globally of the three main variables of interest: confirmed, death, and recovery cases. Our analysis depicts how these variables interact and essentially predict the spreading over worldwide.In order to answer these questions, we use times series machine learning technique implemented by Fbprophet.The approach we used is divided into following phases:

* Identify determinant and acquire data: Identification of influencing factors and delivery of historical data.
* Prepare and clean data: Preprocessing and transforming downloaded data into machine learning readable format.
* Train model: Fbprophet fit method is used to train processed data to analyze pattern and forecast future.
* Plotting: Plotting is one of the essential parts of the project, we are plotting the result in a graphical manner so it is easily understandable.

# ORIGIN OF THE PROPOSAL

From the World Health Organization - On 31 December 2019, WHO was alerted to many cases of pneumonia in Wuhan City, Hubei Province of China. The virus failed to match the other known virus. This raised concern because when a pandemic is new, we don't know the way it affects people. So day-to-day information on the affected people can give some interesting insights when it's made available to the broader data science community. John Hopkins University has made a superb dashboard using the affected cases data. Data is extracted from the google sheets associated.

# SCIENTIFIC RATIONALE AND IMPORTANCE OF THE PROPOSED WORK

Question arises What will be the worldwide impact of the novel coronavirus? Answering this question requires forecasting the spread of confirmed cases moreover as analysis of the number of deaths and recoveries. Forecasting however, requires sufficient historical data. At the same time, no forecasting is precise because the long run repeats itself hardly ever within the same course of action as the past.Moreover, forecasts are influenced by the reliability of the information, vested interests, and what variables are being predicted. Also, psychological factors play a giant role in how people perceive and react to the danger from the disease and thus the fear that it should affect them personally. This manifesto introduces an objective approach to predicting the continuation of the COVID-19 employing a straightforward, but powerful method to do so. The risks are removed from symmetric as underestimating its spread style of an outbreak and not doing enough to contain it's far more severe than overspending and being over careful when it'll not be needed. This proposal describes the timeline of a live forecasting exercise with massive potential implications for planning and decision making and provides objective forecasts for the confirmed, death and recovered cases of COVID-19.

# METHODOLOGY

The methodology of our project is based on Time Series Forecasting. Time Series is a group of data points that are collected at a constant interval of time. These are analyzed to determine the long term trend so as to forecast the future or perform some other form of analysis.There are many ways to implement time series, for this project we used FbProphet particularly.

The code architecture is as follows:

* Perform data scraping to get the raw data
* Import necessary libraries(pandas, fbprophet, matplotlib, statsmodels)
* Load in raw data
* Pre-process data(feature scaling, time formatting)
* Training data with FbProphet model
* Predicting data for next 1 year(period=365)
* Plotting forecasting with matplotlib and plotly

Math of Prophet

Prophet is a machine learning technique for forecasting time series data based on an additive(preservative) model where non linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. The Prophet fits several linear and nonlinear functions of time as integrant.

In its simplest form:

y*(t)* = g*(t)* + s*(t)* + h*(t)* + *e(t)*

where:

*g(t) -* Trend models non-periodic changes(i.e. Growth over time)

s(t) - Seasonality presents periodic changes(i.e. Weekly, monthly, yearly)

h(t) - Ties in effects of holidays

e(t) - Covers idiosyncratic changes not accommodated by the model

In other words, the procedure’s equation can be written as:

y(t) = piecewise\_trend(t)+

seasonality(t)+

holiday\_effects(t)+

i.i.d. noise

# PROVIDE LINK FOR DATASETS RELATED TO PROJECT

<https://docs.google.com/spreadsheets/d/e/2PACX-1vQU0SIALScXx8VXDX7yKNKWWPKE1YjFlWc6VTEVSN45CklWWf-uWmprQIyLtoPDA18tX9cFDr-aQ9S6/pubhtml>

(No. of Total Confirmed, Total Recovered and Total Deceased cases of CoronaVirus)

# BIBLIOGRAPHY

* <https://github.com/CSSEGISandData/COVID-19> (COVID-19 Data Repository by the center for System Science and Engineering (CSSE) at John Hopkins University)

Last Accessed: 23/04/2020

* <https://facebook.github.io/prophet> (FB-Prophet Documentation)
* [https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6](https://facebook.github.io/prophet) (Visual DashBoard-CSSE)
* <https://github.com/Covid-19> Project link